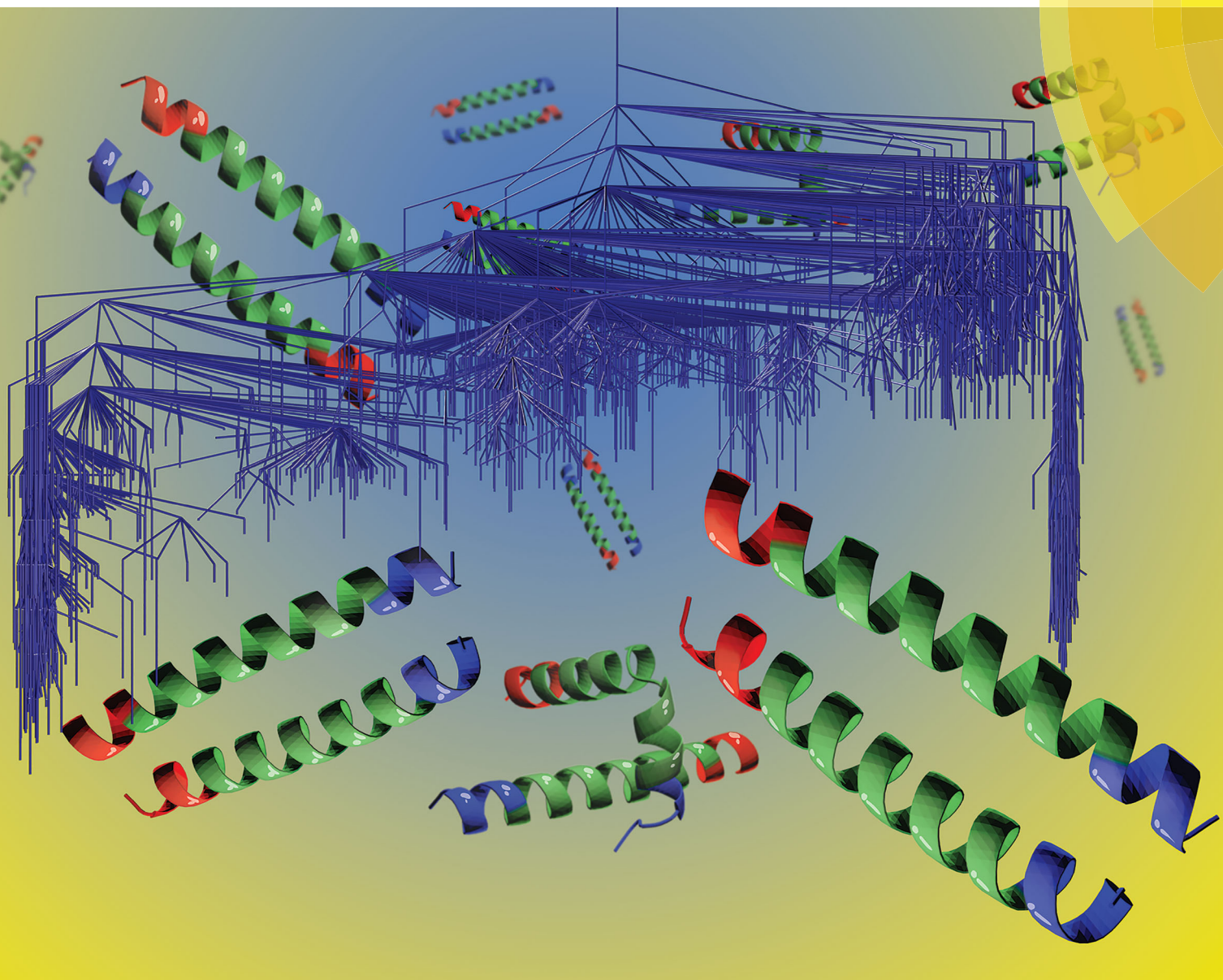


# ChemComm

Chemical Communications

rsc.li/chemcomm



ISSN 1359-7345



ROYAL SOCIETY  
OF CHEMISTRY

FEATURE ARTICLE






David J. Wales *et al.*

Exploring biomolecular energy landscapes



## Exploring biomolecular energy landscapes

Cite this: *Chem. Commun.*, 2017, 53, 6974

Jerelle A. Joseph, <sup>a</sup> Konstantin Röder, <sup>a</sup> Debayan Chakraborty, <sup>ab</sup>  
 Rosemary G. Mantell <sup>a</sup> and David J. Wales <sup>\*a</sup>

The potential energy landscape perspective provides both a conceptual and a computational framework for predicting, understanding and designing molecular properties. In this Feature Article, we highlight some recent advances that greatly facilitate structure prediction and analysis of global thermodynamics and kinetics in proteins and nucleic acids. The geometry optimisation procedures, on which these calculations are based, can be accelerated significantly using local rigidification of selected degrees of freedom, and through implementations on graphics processing units. Results of progressive local rigidification are first summarised for trpzip1, including a systematic analysis of the heat capacity and rearrangement rates. Benchmarks for all the essential optimisation procedures are then provided for a variety of proteins. Applications are then illustrated from a study of how mutation affects the energy landscape for a coiled-coil protein, and for transitions in helix morphology for a DNA duplex. Both systems exhibit an intrinsically multifunnel landscape, with the potential to act as biomolecular switches.

Received 29th March 2017,  
 Accepted 2nd May 2017

DOI: 10.1039/c7cc02413d

[rsc.li/chemcomm](http://rsc.li/chemcomm)

### 1 Introduction

The structure–function paradigm continues to play a central role in advancing our understanding of molecular biology. With the advent of modern spectroscopy and computer simulation techniques, it is now possible to study biomolecules at very fine spatial and temporal resolutions to obtain new insight into the structure–function relationship. It is becoming increasingly clear that evolution has endowed biomolecules with a certain

degree of flexibility, which allows them to not only adopt stable structures, but also switch between different conformations over a hierarchy of timescales, to perform different functions.<sup>1,2</sup> These results highlight the importance of a dynamical perspective. The notion that biomolecules are ‘static sculptures’ is not sufficient to describe their impressive functional capabilities.<sup>3</sup> A deeper understanding of the intimate connections between structure, dynamics and function at the molecular level is therefore necessary. In this Feature Article we illustrate how these connections can be addressed in the framework of potential energy landscape theory.<sup>4</sup>

There are actually two aspects to the protein folding problem, which have been extensively discussed in previous reviews.<sup>5–9</sup> The first one concerns structure prediction: given an amino

<sup>a</sup> Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK. E-mail: [djw34@cam.ac.uk](mailto:djw34@cam.ac.uk)

<sup>b</sup> Department of Chemistry, The University of Texas at Austin, 24th Street Stop A5300, Austin, TX 78712, USA



**Jerelle A. Joseph**

*Jerelle A. Joseph was born on the Caribbean island of Dominica. She received her BSc and MPhil degrees from the University of the West Indies (Cave Hill Campus, Barbados) in 2012 and 2014 respectively. She is currently pursuing a PhD in Chemistry as a Gates Scholar at the University of Cambridge. Her research focuses on efficient sampling of biomolecules and large scale structural changes in proteins.*



**Konstantin Röder**

*Konstantin Röder was born in Jena, Germany. He received his MSci and BA degrees in 2015 from the University of Cambridge. He is currently pursuing a PhD at the same institution. His research focuses on multifunnel energy landscapes of biomolecular systems, and the effect of mutations on these landscapes.*



acid sequence, what three-dimensional structures will a biomolecule populate at thermodynamic equilibrium to fulfil a certain biological function? The second aspect concerns folding mechanisms and kinetics, which describe how the system achieves its functional three-dimensional structure from a relatively unstructured configuration by navigating along folding routes on the underlying landscape, or switches between different metastable conformations that dominate the equilibrium. Many evolved proteins conform to the 'principle of minimum frustration', corresponding to landscapes where kinetic traps are absent, and fold on physiological timescales. This scenario led to the funnel hypothesis, that the native state lies at the bottom of a relatively smooth free energy landscape.<sup>5,6</sup> However, some proteins have evolved to exhibit multifunctional character, with an underlying multifunnelled landscape, which can be tuned

via mutations,<sup>10</sup> ligand-binding,<sup>11,12</sup> or even subtle changes in environmental conditions.<sup>13</sup> An analogous situation exists for nucleic acids, especially RNA.<sup>14,15</sup> Although protein and RNA folding share several common features, significant differences arise due to the polyelectrolyte nature of RNA. Furthermore, due to the lower stability gap between native and non-native states, structure prediction as well as studies of folding mechanisms and kinetics become even more challenging for RNA.<sup>16,17</sup>

Various experiments<sup>18–20</sup> and numerical simulations<sup>21–23</sup> have succeeded in extracting specific features of biomolecular energy landscapes. Nonetheless, determining the global topography of the landscape, and predicting or interpreting emergent properties have proved more difficult. In this context, single molecule pulling experiments are an attractive approach.<sup>24–26</sup> Recently, several studies<sup>27,28</sup> have employed non-equilibrium pulling experiments to reconstruct landscape profiles, using an extended version of Jarzynski's equality<sup>29</sup> suggested by Szabo and Hummer.<sup>30,31</sup> The specific challenges encountered in computational studies of biomolecular structure and pathways are determined by the complexity of the underlying potential energy landscape. Most biological systems exhibit relatively large barriers between alternative competing morphologies, leading to broken ergodicity. In such situations, brute-force simulations have limited applicability, because the equivalence between time and ensemble averages breaks down over short timescales. Hence, the development of enhanced sampling techniques is essential. Many of these techniques primarily aim at predicting thermodynamic properties, and exploit multiple coupled simulations at different temperatures<sup>32–34</sup> or energies,<sup>35</sup> predefined reaction coordinates,<sup>36–38</sup> or hypersurface deformations<sup>39–42</sup> to improve conformational sampling. Obtaining kinetic information, and identifying key intermediates along the multitude of folding pathways, is more difficult. Although techniques based on systematic sampling of reactive trajectories, such as transition path sampling,<sup>43,44</sup> transition interface sampling,<sup>45</sup>



**Debayan Chakraborty**

*Debayan Chakraborty was born in Kolkata, India. He received his BSc degree from St. Stephens College, University of Delhi, in 2010, and MSc degree from the University of Oxford in 2011. He carried out his doctoral research in the area of Theoretical Chemistry from 2012 to 2016, under the supervision of Prof. David J. Wales at the University of Cambridge. Currently he is a postdoctoral research scholar at the University of Texas at Austin.*

*His research interests include the study of protein and nucleic acid folding exploiting ideas rooted in energy landscape theory, and coarse-grained modelling of biomolecules.*



**Rosemary G. Mantell**

*Rosemary G. Mantell received her BSc in Chemistry in 2012 from the University of Bristol. She then went on to complete an MPhil in Scientific Computing at the University of Cambridge. She continued her graduate studies at the same institution under the direction of Professor David J. Wales. Her PhD has so far focused on the acceleration of computational energy landscape methods using Graphics Processing Units (GPUs).*



**David J. Wales**

*David J. Wales received his BA, PhD, and ScD degrees from Cambridge University in 1985, 1988 and 2004. He spent 1989 as a Lindemann Trust Fellow at the University of Chicago, working with Prof. R. S. Berry. He was a Research Fellow at Downing College Cambridge in 1990, a Lloyd's of London Tercentenary Fellow in 1991, and a Royal Society University Research Fellow from 1991 to 1998. He was awarded the Meldola Medal and*

*Prize by the Royal Society of Chemistry in 1992, and the Tilden Prize in 2015. In 1998 he was appointed to a Lectureship in Cambridge and is now Professor of Chemical Physics and Chair of the Theory Group in the Chemistry Department. He was elected as a Fellow of the Royal Society in 2016.*



milestoning,<sup>46,47</sup> and forward flux sampling,<sup>48</sup> can provide kinetic and mechanistic insight, they can be computationally intensive.

The construction of kinetic transition networks provides a complementary way to study biomolecular energy landscapes.<sup>49–51</sup>

We have actively developed methods based on the potential energy landscape perspective, which provides a convenient framework for building and analysing transition networks.<sup>4,50</sup> In this approach, the landscape is coarse-grained into a set of interconnected stationary points [minima (M) and transition states (TS)]. This simplification is particularly attractive because stationary points can be located using geometry optimisation, largely independent of energy barriers. The global thermodynamics is dictated by the configuration space associated with the potential energy minima. The connections between different regions of the landscape are defined in terms of M–TS–M triples, which encode the kinetics.<sup>4</sup> Observable properties are extracted from the databases of stationary points using established tools of statistical mechanics and unimolecular rate theory.<sup>52–55</sup> It is also possible to analyse the landscape in terms of selected order parameters and lumping schemes to identify states or ensembles, commensurate with the experimental definition.<sup>56</sup> Methods that employ explicit dynamics to construct transition networks, mainly in the context of Markov State Models (MSM),<sup>49,51,57</sup> are again complementary to methodology based on geometry optimisation.

The potential energy landscape approach has been used to study a range of biomolecular processes. Recent applications include folding of RNA hairpins,<sup>58</sup> the origin of heterogeneity in the intrinsically disordered PUMA peptide,<sup>59</sup> ring puckering in collagen,<sup>60</sup> the effect of mutation on the influenza A virus,<sup>61</sup> and the transformation of the landscape for a model protein by a static pulling force.<sup>62</sup> The purpose of the present contribution is to briefly summarise the potential energy landscape framework, and discuss recent advances that have made new applications possible.

Biological systems of practical interest can often exceed 100 000 atoms in size. The number of degrees of freedom poses a serious challenge to current simulation techniques, and even for a biomolecule of moderate size with around 1000 atoms, attaining true equilibrium is time consuming. To circumvent this problem, a common approach is to coarse grain the atomistic representation of the molecule.<sup>63–65</sup> However, such reductionist schemes may not be able to represent key dynamical features, and can result in an artificially smooth landscape. Here we present two strategies within the potential energy landscape perspective that have proved effective in structure prediction, and obtaining thermodynamic and kinetic insight at the all-atom level. In Section 3, we review the key findings of a general local rigid body approach from a recent study of the trpzip1 peptide. On the technological side, implementing geometry optimisation routines for GPU hardware has proved effective, as summarised in Section 4. We then discuss two recent case studies involving biological conformational switches. The first example analyses how mutations reshape the underlying landscape of a coiled-coil protein, and the second case study shows how the potential

energy landscape framework can be used to probe transitions between different helical morphologies of a DNA duplex. Both examples exhibit multifunnel energy landscapes, which we associate with intrinsic multifunctional behaviour.<sup>59</sup>

## 2 Methods

The potential energy landscape is characterised by stationary points: local minima and transition states. A stationary point is a local minimum if all the non-zero normal mode frequencies are real, whereas transition states are classified geometrically as stationary points with one imaginary normal mode frequency.<sup>66</sup> In the following sections we provide a brief summary of the approaches employed to locate stationary points and construct kinetic transition networks; specifically, basin-hopping global optimisation<sup>67–70</sup> and discrete path sampling.<sup>71–73</sup> We then explain how the free energy surface may be computed from the underlying potential energy landscape using the superposition approach.<sup>4,74</sup> Lastly, we discuss the use of disconnectivity graphs<sup>75–78</sup> to visualise the landscape. The more recent developments, in terms of local rigidification and implementation of geometry optimisation techniques on GPU hardware, are described in more detail in the following sections.

### 2.1 Structure prediction by basin-hopping global optimisation

The basin-hopping global optimisation procedure<sup>67,68</sup> has been successfully applied to a wide range of systems, spanning atomic clusters,<sup>68</sup> glass formers,<sup>79</sup> and biomolecules.<sup>80</sup> This method employs a hypersurface deformation, but does not change the global minimum of the potential energy surface (PES). Each configuration on the PES can be represented by a  $3N$ -dimensional vector  $\mathbf{X}$ , where  $N$  is the number of atoms and the energy corresponding to  $\mathbf{X}$  is given by  $E(\mathbf{X})$ . The energy obtained by a minimisation starting from  $\mathbf{X}$  is written as  $\min\{E(\mathbf{X})\}$ . In the present work energy minimisation was achieved using the limited-memory BFGS (L-BFGS)<sup>81,82</sup> algorithm, which is well-suited for large-scale problems, since the user is able to control the amount of storage required. The PES is the union of the ‘basins of attraction’<sup>83,84</sup> of all the local minima. This procedure effectively removes all downhill barriers between connected minima. Basin-hopping global optimisation has been implemented in the GMIN program,<sup>85</sup> which is available for use under the GNU General Public License.

### 2.2 Kinetic transition networks from discrete path sampling

In discrete path sampling<sup>71–73</sup> (DPS) the aim is to determine the kinetics of a system from a collection of transition pathways, connecting reactant (*e.g.* a denatured configuration) and product (*e.g.* the native structure) states. A discrete path is defined as a sequence of local minima on the potential energy surface (PES) and the transition states that directly connect them.

We start by constructing an initial path from the reactant (A) to the product (B).<sup>86–88</sup> Transition state candidates are first



located using the doubly-nudged elastic band (DNEB) procedure.<sup>86</sup> A double-ended interpolation between A and B produces an intermediate set of images  $[X_1, X_2 \dots X_M]$ , where  $X_i$  represents the Cartesian coordinates of the  $i$ th image. Next, harmonic springs are used to connect equivalent atoms in adjacent images, resulting in a spring potential. Components of the spring gradient and the gradient of the true potential are then utilised in the derivation of the elastic band gradient. This gradient prevents interference of the spring potential (which affects convergence of images) and the true potential (which affects the spacing of images) and gives the band its ‘nudging’ properties.<sup>89</sup> The complete set of images is then relaxed by L-BFGS minimisation, using a weak convergence condition, which focuses on the local maxima.<sup>86</sup>

Approximate transition states (the images corresponding to maxima in the DNEB energy profile) are then converged more tightly by hybrid eigenvector-following (HEF).<sup>90,91</sup> This is a single-ended procedure for locating transition states, where a single starting configuration is considered.<sup>90,91</sup> In contrast, for double-ended procedures, such as the DNEB method, two initial endpoint geometries are needed. In HEF, only one Hessian eigenvector ( $e_{\min}$ ) and the corresponding eigenvalue ( $\lambda_{\min}$ ) are used for uphill searches and a minimisation procedure, such as the L-BFGS algorithm, is used for optimisation in all other directions.<sup>91</sup> The smallest non-zero eigenvalue can be found using the Rayleigh–Ritz ratio  $\lambda(\mathbf{x}) = \frac{\mathbf{x}'\mathbf{H}\mathbf{x}}{\mathbf{x}^2}$ , where  $\mathbf{x}$  represents a small perturbation in the current geometry. To avoid explicit computation of the Hessian ( $\mathbf{H}$ ),  $\lambda(\mathbf{x})$  is estimated from the numerical second derivative of the energy. Through successive iterations, tightly converged transition states are characterised. These transition states are then connected to minima by following approximate steepest-descent paths parallel and antiparallel to the eigenvector corresponding to the unique negative eigenvalue.

The minima and transition states found during DNEB/HEF searches form a database of stationary points. Recall that the goal is to find a connected path between the reactant (A) and product (B). Before each new DNEB/HEF cycle, a metric is needed to determine which minima in the database are most appropriate for subsequent connection attempts. A modified version<sup>92</sup> of Dijkstra’s algorithm,<sup>93</sup> which selects pairs based on a minimised Euclidean distance metric, is used for this purpose. In a database of stationary points the total set of minima can be described using a complete graph  $G(M,D)$ ; where the nodes  $M$  represent minima and the edges  $D$  represent the transition states. The edge weight  $w(x,y)$  between arbitrary minima  $x$  and  $y$  is set to zero if the minima are connected by a single transition state. If the number of connection attempts between minima  $x$  and  $y$  reaches the maximum value (set by the user), the edge-weight becomes infinite. Otherwise,  $w(x,y)$  is expressed as a function of the Euclidean distance. At the beginning of each DNEB/HEF cycle, pairs of minima for connection attempts are prioritised based on these weights. This process is repeated until there are no missing connections along the path.

The doubly-nudged elastic band procedure, hybrid eigenvector-following, and the shortest path algorithm are implemented in the OPTIM code.<sup>94</sup> Parallel OPTIM runs are organised by the PATHSAMPLE program.<sup>95</sup>

### 2.3 Optimisation of stationary point databases

The initial path found between reactant and product states is usually long with many high barriers, particularly for states distant in configuration space. The objective is then to grow the stationary point database and locate kinetically relevant paths. At any point, the fastest path ( $B \leftarrow A$ ) is taken as the one making the largest contribution to the steady-state rate constant,  $k_{BA}^{SS}$ , which can be defined as a sum over all discrete paths with the steady-state approximation for intervening minima.<sup>71,72</sup> Once the fastest path is identified, it can then be used in various ways to search for new paths.

The SHORTCUT scheme<sup>96,97</sup> chooses pairs of minima from the current ‘fastest’ path that are separated by a minimum number of transition states (steps) and attempts to connect them using the procedures discussed above. The SHORTCUT procedure usually decreases the total number of steps on the path and leads to a significant increase in the overall rate constant. Alternatively, the SHORTCUT BARRIER scheme<sup>92,96</sup> selects minima on either side of the largest barriers on the current path, up to a maximum number of steps apart. Additional connection attempts between these minima may find paths avoiding such high barriers, again improving the rate constant.

However, these procedures may also introduce kinetic traps, in the form of low-lying minima separated from the product minimum by high barriers. Most of these traps are artificial and are due to insufficient sampling. To find low-barrier paths for these minima the UNTRAP scheme<sup>96</sup> is used. Candidate minima for ‘untrapping’ are chosen based on the ratio of the potential energy barrier and potential energy difference from the product (B). Hence, minima with low potential energies connected by high barriers are most likely to be chosen. Connection attempts between these minima and the product minimum then proceed in search of better paths. Local free energies can also be used.

### 2.4 Computation of free energies

Thermodynamic properties can be estimated directly from the underlying potential energy landscape using the superposition approach. At a given temperature, the canonical partition function  $Z(T)$  is written as a sum of contributions from the basins of attraction of the local minima:<sup>4,52,74,98–101</sup>

$$Z(T) = \sum_i Z_i(T). \quad (1)$$

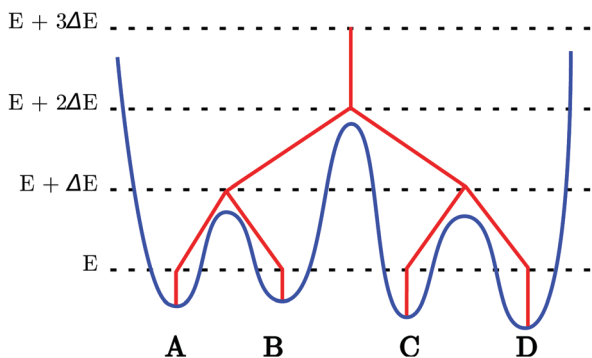
A harmonic approximation is often used to estimate the vibrational partition function of each local minimum,

$$Z_i(T) = \frac{n_i \exp(-\beta V_i)}{(\beta \hbar \bar{\nu}_i)^\kappa}, \quad (2)$$

where  $n_i$  is the number of distinct permutational isomers of minimum  $i$  with potential energy  $V_i$ ,  $\beta = 1/k_B T$ ,  $k_B$  is the Boltzmann constant,  $\bar{\nu}_i$  is the geometric mean vibrational frequency and  $\kappa$  is the number of vibrational degrees of freedom.<sup>4</sup> The free energy of each minimum is then:

$$F_i(T) = -k_B T \ln Z_i(T). \quad (3)$$





**Fig. 1** Construction of a disconnectivity graph from a database of stationary points. Minima are labelled A–D. In the disconnectivity graph (red) each local minimum is represented by a vertical line, starting at the energy of that minimum. At a given energy threshold,  $E + n\Delta E$ , minima connected by transition states that lie below the threshold are grouped into disjoint sets.

The harmonic approximation has also been used as a starting point for superposition calculations that include quantum corrections and anharmonicity.<sup>99,102–107</sup>

### 2.5 Visualisation of energy landscapes: disconnectivity graphs

Originally introduced by Becker and Karplus,<sup>75</sup> disconnectivity graphs have played a pivotal role in conceptualising energy landscapes.<sup>76–78</sup> Fig. 1 illustrates how a disconnectivity graph (red lines) may be constructed from a database of minima and the transition states that connect them (blue curve). The energy is represented on the vertical axis of the graph, while the horizontal axis can be arbitrary or may represent an order parameter. In the disconnectivity graph, a vertical line is drawn from each minimum (A–D), beginning at the potential or free energy of that state. At the energy threshold  $E + \Delta E$  minima A and B are grouped together, since the transition state connecting them lies below the threshold, and similarly for minima C and D. However, the two sets are disjoint, as the transition state connecting them lies above this threshold. When the threshold is high enough the two sets of minima merge. Since the energy spacing ( $\Delta E$ ) determines how the analysis is performed, the graph is most meaningful when the thresholds are spaced at suitable regular intervals. This analysis provides useful information on relative barriers separating minima in different regions of the landscape.

## 3 Systematic local rigidification for trpzip1

Bimolecular processes, such as protein folding, frequently exhibit a separation of timescales between low frequency modes and localised fast vibrations.<sup>109–113</sup> For instance, changes in ring planarity occur orders of magnitude faster than the formation of salt bridges. We can exploit this disparity in timescales to improve computational efficiency, by effectively factorising out the irrelevant degrees of freedom using a local rigid body (LRB) framework.<sup>114–117</sup> This approach is similar in spirit to the FEG-RBD procedure,<sup>118</sup> where rigid body dynamics is employed to compute free energy gradients in constrained MD simulations.

In the local rigidification scheme employed here, groups of atoms, whose relative coordinates vary on timescales significantly shorter than the process of interest, are treated as rigid units.<sup>115–117</sup> Each non-linear rigid unit, regardless of size, then has six degrees of freedom: three translational and three rotational. Since interactions between sites within each local rigid body are not required, this formulation results in computational savings in evaluating the energy and gradient. Additionally, fewer geometry optimisation steps are generally needed, due to the dimensionality reduction of the conformational search space. Hence, significant speedups in the mean first encounter time for global optimisation are possible.<sup>116</sup> Moreover, a clear mapping between the minima found on locally rigidified and unconstrained potential energy landscapes can be maintained.<sup>116</sup>

In a recent article,<sup>108</sup> we investigated the systematic effects of local rigidification on the structure, thermodynamics and kinetics of trpzip1.<sup>119</sup> The tryptophan side-chain rings, peptide bonds, trigonal planar centres and termini were grouped as rigid bodies, and, based on these sets, we formulated three LRB schemes: I – rings, II – rings and peptide bonds, III – rings, peptide bonds, termini, trigonal planar centres treated as local rigid bodies. Trpzip1 was represented by the AMBER99SB forcefield,<sup>120</sup> and discrete path sampling<sup>71,72</sup> was used to analyse the potential energy landscape for each LRB scheme, as well as the unconstrained (U) peptide.

The main conclusions are summarised in Fig. 2 and 3. The potential energy range and structural heterogeneity of the locally rigidified landscapes are consistent with the unconstrained landscape (Fig. 2). However, there is a significant increase in the number of prominent subfunnels in the disconnectivity graph for the most constrained case (Fig. 2d). This result suggests that excessive rigidification can reduce the conformational flexibility of the protein in an unphysical manner, leading to artificially high barriers on the potential energy landscape.

Nonetheless, the free energy global minimum (particularly at low temperatures) and the melting temperature for the locally rigidified implementations were in good agreement with the unconstrained benchmarks (Fig. 3a). To assess the effects of local rigidification on the folding mechanism, we extracted the fastest folding path<sup>71–73</sup> from each transition network (Fig. 3b). The folding pathways for the unconstrained peptide, schemes I and II, were found to be consistent with an initial hydrophobic collapse ( $s1 \rightarrow s2$ ), followed by subsequent zipping ( $s2 \rightarrow s4$ ). However, for LRB scheme III a significant lengthening of the folding pathway was observed, which is consistent with excessive reduction in conformational flexibility.

These results suggest that local rigidification should be useful for analysis of thermodynamics and folding mechanisms of biomolecules. Moreover, the results provide some practical guidelines for systematically eliminating unnecessary degrees of freedom, and demonstrate how too much rigidification can significantly perturb pathways and kinetics. An appropriate choice of local rigidification (based on the level of resolution required, and the timescale to be probed) can preserve the underlying properties, and provide useful gains in computational efficiency.



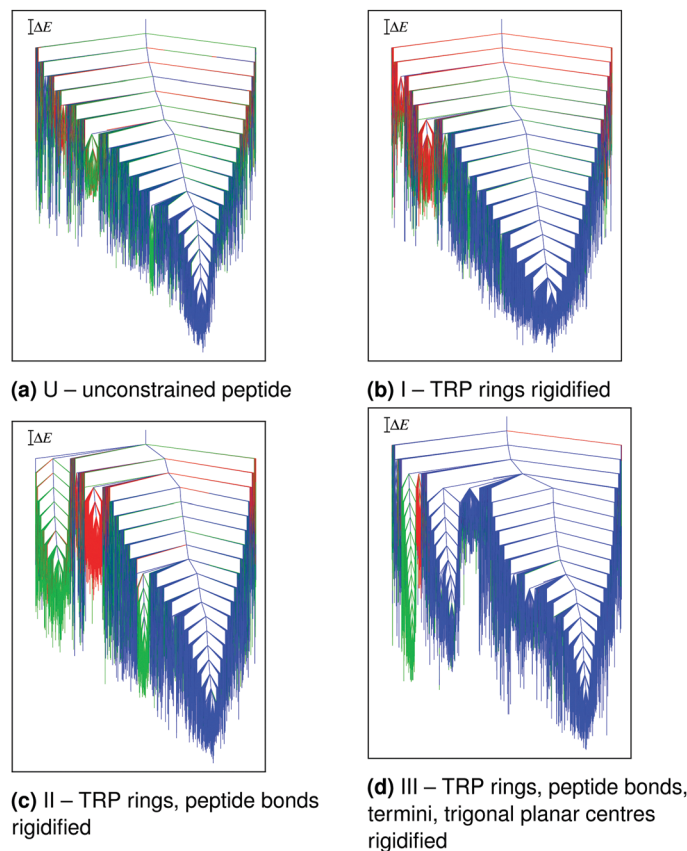


Fig. 2 Potential energy disconnectivity graphs for TZ1 ( $\Delta E = 2 \text{ kcal mol}^{-1}$ ) at different levels of local rigidification. The branches are coloured based on order parameters  $L$  (the sum of the four inner native hydrogen-bond lengths and the distances between the CD2 atoms of the three TRP pairs) and  $S$  (the relative orientation of the TRP rings). The three main morphologies are: blue ( $L < 60 \text{ \AA}$ ,  $S$ -value = +1), green ( $L < 60 \text{ \AA}$ ,  $S$ -value = -1), and red (all other minima). Adapted from ref. 108.

The number of stationary points on the energy landscape increases exponentially with system size,<sup>98,121</sup> but the mean first encounter time for the global minimum in basin-hopping global optimisation for atomic clusters seems to scale roughly as the cube of the number of atoms.<sup>68</sup> Hence, removing irrelevant degrees of freedom will be even more beneficial if the size scaling is steeper, which may be the case for larger proteins.

## 4 GPU-accelerated geometry optimisation

Basin-hopping global optimisation,<sup>67,68</sup> the doubly-nudged<sup>86</sup> elastic band<sup>87,89</sup> method (DNEB), hybrid eigenvector-following (HEF)<sup>90</sup> and the local rigid body framework<sup>116,117</sup> have all been adapted to run on graphics processing units (GPUs).<sup>122</sup> GPU hardware is well suited to massively parallel computations, as a greater number of transistors are devoted to data processing than to data caching and flow control, relative to a CPU.<sup>123</sup>

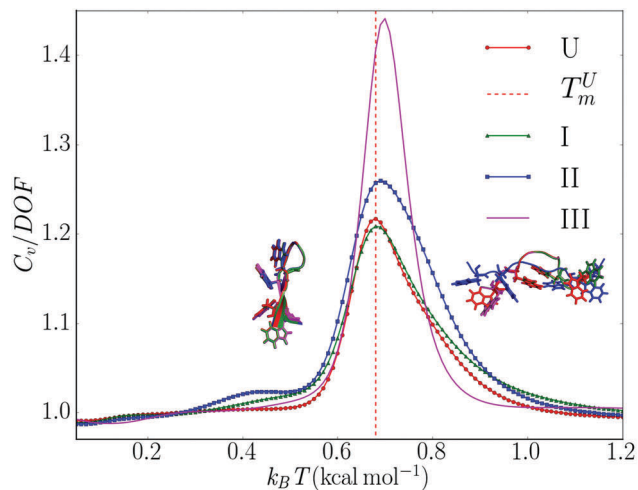
The most time-consuming component in all of the computational energy landscapes approaches is the calculation of the potential energy and gradient. For biomolecules, the task of adapting a force field for GPU was facilitated by the release of a GPU-accelerated version of the AMBER potential.<sup>124</sup> The generalised

Born (GB) implicit solvent potential from AMBER 12 was interfaced with all the energy landscapes codes. The DPDP precision model was used, in which contributions to forces and their accumulation are both performed in double precision.

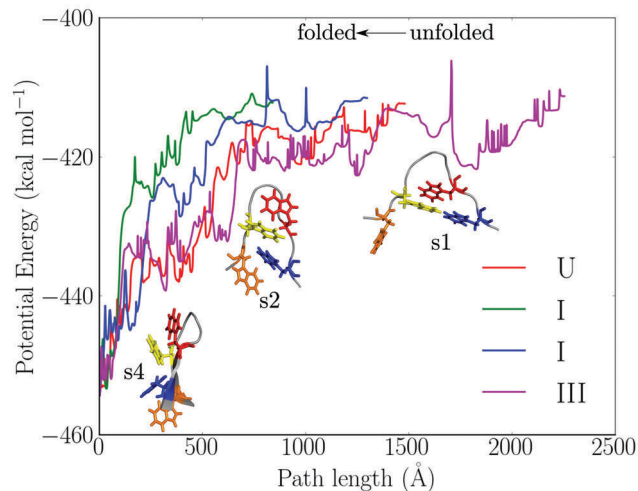
Basin-hopping global optimisation and transition state determination employ several variants of the L-BFGS algorithm. In each case, the whole algorithm was ported to the GPU, and compared to an implementation with just the potential on GPU. Our GPU implementation of L-BFGS was based on the code of Wetzl and Taubmann,<sup>125,126</sup> with various modifications applied to make it as similar as possible to our CPU code. The coordinate transformation and gradient projection functions from our local rigid body framework were also ported to GPU. Extensive use was made of the cuBLAS library<sup>127</sup> in this work, alongside custom CUDA kernels.

Tests were performed for eight different system sizes ranging from 81 to 22 811 atoms. A subset of these results is presented here for the trimeric haemagglutinin (HA) glycoprotein of the influenza A(H1N1) virus<sup>128</sup> ('full trimer', 22 811 atoms), a monomeric version of HA ('full monomer', 7585 atoms), and finally a truncated version of this monomer ('truncated monomer', 3522 atoms). These systems are shown in Fig. 4. L-BFGS was benchmarked in isolation as a proxy for full basin-hopping global optimisation. The change in coordinates and the change





(a) Constant volume heat capacity curves.



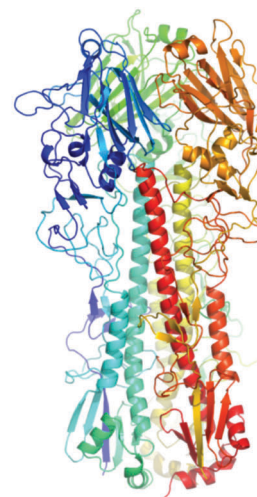
(b) Individual fastest folding pathways.

**Fig. 3** Summary of thermodynamic and mechanistic properties for trpzp1, as a function of local rigidification: U – no local rigid bodies; I – TRP rings, II – TRP rings and peptide bonds, III – TRP rings, peptide bonds, termini, trigonal planar centres treated as rigid bodies. (a) Constant volume heat capacity curves: the heat capacities are divided by the appropriate total number of degrees of freedom (DOF) and the melting temperature of the unconstrained peptide,  $T_m^U$ , is indicated. The free energy global minima at 0.48 and 0.88 kcal mol<sup>-1</sup> are superimposed on the plot; key: red (U), green (I), blue (II), magenta (III). (b) Variation of the total potential energy (kcal mol<sup>-1</sup>) with the integrated path length (Å) for the fastest folding path from the denatured trpzp1 peptide to the global minimum. Some morphologies encountered along the paths are shown. Adapted from ref. 108.



(a) HA truncated monomer (3522 atoms)

(b) HA monomer (7585 atoms)



(c) HA trimer (22811 atoms)

**Fig. 4** Structures of the biomolecules used in the tests of GPU geometry optimisation procedures. Adapted from ref. 122.

in gradient for the last  $m$  steps are used in calculating the step direction for minimisation.<sup>129</sup> The history size,  $m$ , affects the accuracy of the L-BFGS step direction and the overall time taken for minimisation. Results for L-BFGS are shown in Table 1 for history sizes 4 and 1000. Significant speed ups are obtained of up to two orders of magnitude. A large history size is more favourable for bigger systems with the whole L-BFGS algorithm on GPU, whereas a smaller history size gives optimal performance for smaller systems; in contrast, for CPUs a larger history size is always more favourable. We extended our analysis to find

the optimal history size and GPU implementation for the full HA trimer. The fastest average minimisation time obtained was 1788 s for a history size of 75 with the whole L-BFGS algorithm on GPU (implementation 1).

Tests were also performed to compare transition state searches on GPU and CPU. Interfacing the DNEB procedure with the GPU-accelerated potential gave a 178 times speed up compared to CPU for the full HA trimer with a small history size of 4.



**Table 1** L-BFGS benchmarking for HA with history sizes 4 and 1000 using two GPU and one CPU implementation<sup>a</sup>

History size	System	Average minimisation time for GPU implementation 1/seconds	Average minimisation time for GPU implementation 2/seconds	Average minimisation time for CPU implementation/seconds	Time for CPU implementation/time for GPU implementation 1	Time for CPU implementation/time for GPU implementation 2
4	Truncated monomer	40.1	39.4	5937.7	148.1	150.8
	Full monomer	489.9	492.1	86552.8	176.7	175.9
	Full trimer	2546.9	2527.1	517567.7	203.2	204.8
1000	Truncated monomer	382.1	239.6	4561.2	11.9	19.0
	Full monomer	1249.5	1609.0	48754.5	39.0	30.3
	Full trimer	2392.5	4747.1	337077.0	140.9	71.0

<sup>a</sup> GPU implementation 1 has the entire L-BFGS routine on GPU, including the potential calculation. GPU implementation 2 has just the potential calculation on GPU. See the original paper for full descriptions of all the parameters.<sup>122</sup>

**Table 2** Hybrid eigenvector-following benchmarking for HA with history sizes 4 and 1000 using GPU and CPU<sup>a</sup>

History size	System	Average time for GPU implementation ( $m = 4$ )/seconds	Average time for CPU implementation ( $m = 4$ )/seconds	Time for CPU implementation/time for GPU implementation ( $m = 4$ )
4	Truncated monomer	25.7	3529.1	137.4
	Full monomer	197.1	36779.8	186.6
	Full trimer	1047.6	176550.7	168.5
1000	Truncated monomer	186.3	3160.9	17.0
	Full monomer	669.6	32046.2	47.9
	Full trimer	1225.4	169234.0	138.1

<sup>a</sup> The GPU implementation has the entire Rayleigh–Ritz L-BFGS and L-BFGS routines with gradient projection on GPU, including the potential calculation. See the original paper for full descriptions of all the parameters.<sup>122</sup>

Results for hybrid eigenvector-following with HA are shown in Table 2. Again, significant improvements are obtained.

Local rigidification is now supported on GPU in conjunction with all of these methods in the GMIN<sup>85</sup> and OPTIM<sup>94</sup> programs. The GPU implementations of the coordinate transformations and gradient projection were found to incur negligible computational overhead in each L-BFGS procedure.

## 5 Multifunnel energy landscapes for proteins

The question that originally motivated interest in the energy landscapes of proteins is how a sequence of amino acids achieves its native fold.<sup>130</sup> From the application of statistical mechanics to various model systems the existence of intrinsic properties of protein sequences emerged, which lead to a preference for particular folded states. This idea produced the notion of a funnelled free energy landscape containing the native state as the global minimum.<sup>9,131,132</sup>

A single funnel on the free energy landscape leads to a single dominant ensemble corresponding to the native state, yet for many biological systems more than one stable fold is observed, *e.g.* activated states or misfolded states. In many cases these additional stable states require a significant rearrangement of the original folded structure. The resulting alternative morphologies are unlikely to exist within a single funnel on the energy landscape. As a result we observe multifunnel energy landscapes that contain at least two competing morphologies. One important

group of multifunnel landscapes are associated with proteins that exhibit stable misfolded structures. Such structures alter the functionality of the protein, and hence they may be pathological. Particularly important examples are amyloid forming proteins, such as A $\beta$ <sub>1–42</sub>. Another set of proteins likely to exhibit multifunnel landscapes are those that show activated and inactive states, such as kinases. Apart from these important biological and medically relevant proteins, artificial systems used in nanotechnology, such as peptide-based molecular switches, may also require multifunnel landscapes to represent alternative states of an engineered system.

An interesting question for multifunnel landscapes involves the changes caused by mutations, which affect the observable properties. Analysing these changes will lead to a better understanding of biological processes, such as activation and disease, as well as design principles for functional peptides. Previous studies of mutations suggest that naturally occurring proteins possess an inherent stability towards mutations.<sup>134</sup> Hence, if a protein still folds after mutation, the folded structure may closely resemble the wild type native state. However, it has also been noted that this stability is likely to be marginal, and may be overcome.<sup>135–137</sup>

Recently, we have studied a system that exemplifies these features.<sup>133</sup> The coiled-coil protein, GCN4-pLI, experimentally exhibits exclusively parallel assemblies of  $\alpha$ -helices. A mutation of the parent sequence, namely E20S, leads to the observation of both parallel and antiparallel states.<sup>138</sup> Other studies have considered a variety of similar coiled-coil systems, and concluded that there is competition between different oligomer sizes,



as well as between parallel and antiparallel alignments.<sup>139–142</sup> A multifunnel landscape was suggested as one explanation for this behaviour.<sup>143</sup> Our calculations used the properly symmetrised<sup>144,145</sup> AMBER ff99SB force field,<sup>120,146–148</sup> with an implicit generalised Born solvation model<sup>149,150</sup> using infinite interaction cutoffs and the Debye–Hückel approximation for salt (0.1 M).<sup>151</sup>

The energy landscapes for the parent sequence and the E20S mutant dimers resemble each other, as shown in Fig. 5, in that the main topography is maintained. There are two distinct funnels for parallel (red) and antiparallel (blue) structures. However, a key difference is found in the intermediate region between these funnels. The E20S mutant supports stable intermediate structures, which show a kink in one of the helices (one example is shown in Fig. 5). These low energy structures facilitate interconversion at lower temperatures, which will not occur for the parent sequence, a feature that leads to signatures in the calculated heat capacity curves. The structural basis for this effect is the change in interaction patterns caused by the mutation. When the energy landscapes are analysed regarding the distribution of the local minima with respect to potential energy, and the logarithm of the product of their normal mode frequencies,<sup>133</sup> we find that the E20S mutant exhibits a bimodal distribution, corresponding to the parallel and the antiparallel funnels in the landscape. In contrast, the parent sequence only exhibits unimodal distributions, suggesting that the energy landscape is based on an underlying single funnel.

Further insight can be obtained by decomposing the contributions of local minima to the heat capacity features, using the temperature gradient of the occupation probabilities.<sup>152</sup> For the E20S mutant three peaks in the heat capacity curve can be associated with three distinct transitions (see Fig. 6): (i) from the global minimum to other low-lying minima in the same funnel, (ii) from minima in the parallel funnel to minima in the intermediate region, and (iii) a melting peak. The second transition is the key to the thermal accessibility of both parallel and antiparallel states at ambient conditions.<sup>133</sup> In contrast, the parent sequence only shows transitions within the same funnel and a single melting peak.

Moving to more disruptive mutations, the energy landscapes of three further mutants, E20P, E20A and E20G, all exhibit distinct features (Fig. 7). While all three still support stable dimers, both antiparallel and parallel, a number of significant changes occur. Firstly, the energy differences between funnels shift as the dimers are destabilised. Additionally, for E20A and E20G, we see new funnels appearing. Furthermore, the number of intermediate structures increases and they show a larger variety of conformations. While the E20S mutant exhibits one particular interconversion mechanism, the larger perturbations of the bonding pattern lead to coiling and helix kinking in many places.

These results suggest a number of important conclusions. First of all, the effects of mutations can be analysed using energy landscape methods, leading to a detailed understanding of energetic balances, structural and mechanistic differences, and changes of associated observables, such as the heat capacity curves. Moreover, protein sequences may be mutated *in silico* to

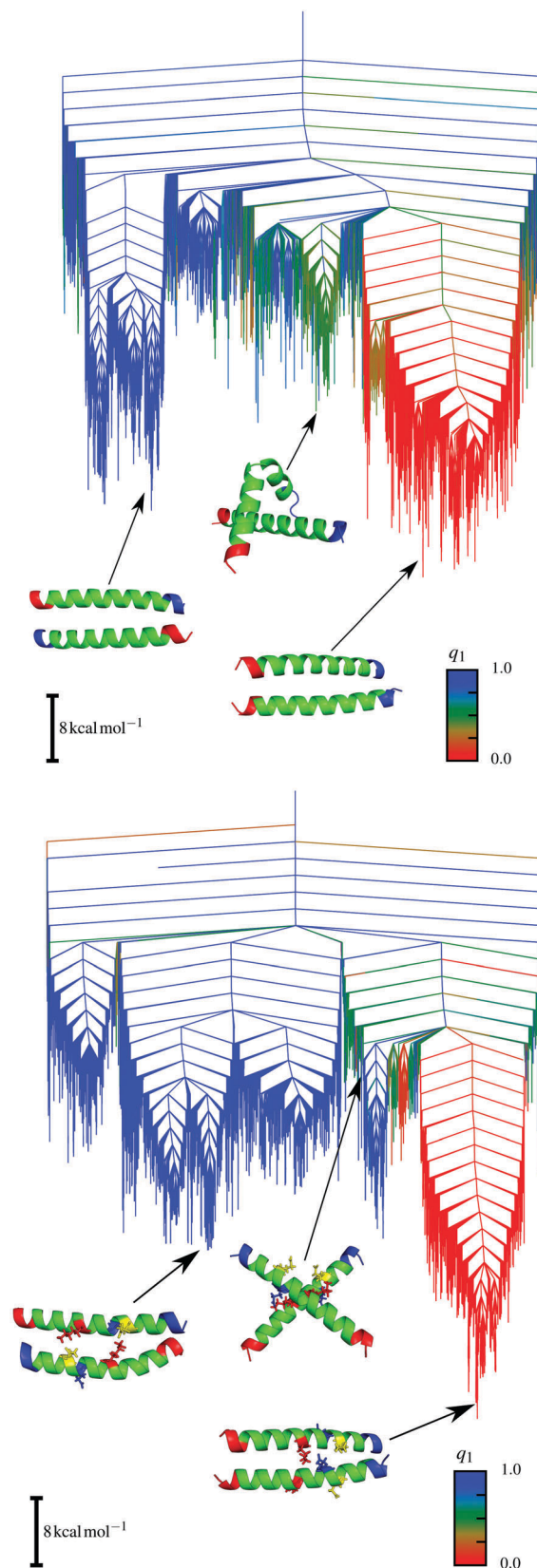
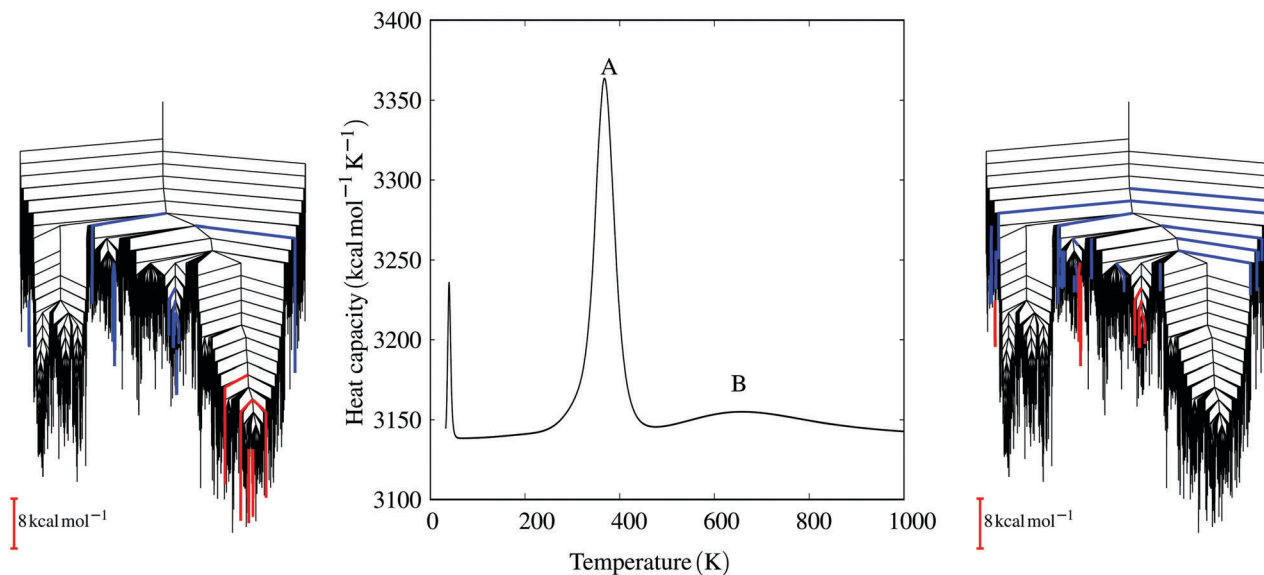
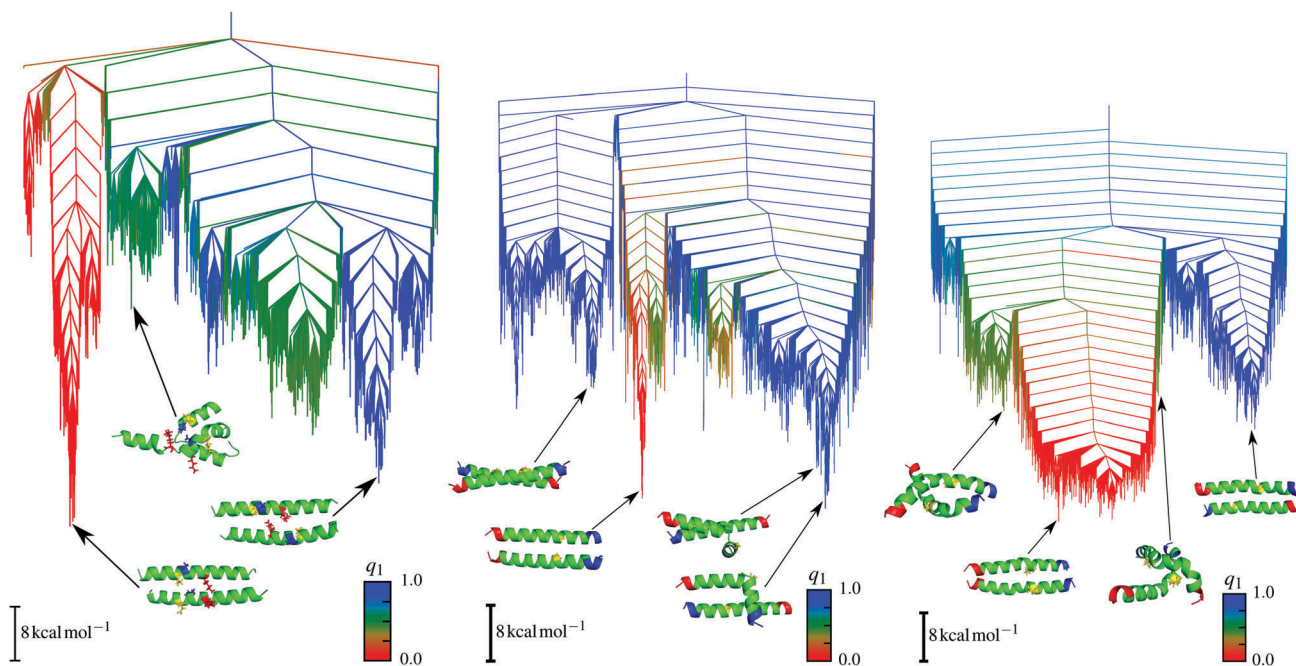


Fig. 5 Disconnectivity graphs for the E20S mutant (top) and the parent sequence (bottom). The order parameter  $q_1$  represents the angle between the two helices and is 0 for parallel and 1 for antiparallel alignments. The separate funnels are clearly visible. Adapted from ref. 133.





**Fig. 6** The heat capacity curve for the E20S mutant shows three distinct peaks. The central peak (A) corresponds to the transition from the parallel funnel to intermediate states. The disconnection graph (left) highlights the minima involved, colouring branches red and blue for decreasing and increasing occupation probabilities, respectively, for the states that make the largest contributions to the heat capacity peak.<sup>152</sup> The highest temperature peak (B) corresponds to the melting peak and a disconnection graph with the key minima coloured in the same way is shown on the right. These results exploit a formulation of the heat capacity written in terms of occupation probability temperature gradients,<sup>152</sup> which enables the contribution of each minimum to be quantified. Adapted from ref. 133.



**Fig. 7** Disconnection graphs for E20A (left), E20G (middle) and E20P (right) mutants. The order parameter  $q_1$  represents the angle between the two helices and is 0 for parallel and 1 for antiparallel alignments. Adapted from ref. 133.

use energy landscape methods as a predictive tool in engineering and *de novo* design. Our results also reveal the inherent complexity of multifunnel landscapes and the nature of changes associated with mutations. Last but not least, even the relatively simple coiled-coil system highlights the marginal stability of these energy landscapes towards mutations.

## 6 Helical transitions in a DNA duplex

The B  $\rightarrow$  Z transition is one of the slowest conformational changes that has been characterised in biomolecules. ZDNA exhibits a left-handed helical structure,<sup>153–155</sup> in contrast to the well known right-handed helices of BDNA and ADNA. This left-handed helix



has zig-zag phosphate backbone with an alternating pattern of *syn* and *anti* glycosidic torsions.<sup>154–156</sup> The discovery that ZDNA has important cellular functions<sup>156–161</sup> makes the complex rearrangements required to interconvert alternative helices particularly interesting. Experimentally, BDNA may convert to ZDNA at high salt concentrations,<sup>153,156,162,163</sup> through negative supercoiling,<sup>164–166</sup> methylation,<sup>167–169</sup> and binding to protein ligands.<sup>170,171</sup> Various suggestions have been proposed for the mechanism,<sup>172</sup> ranging from cooperative schemes,<sup>173</sup> pathways involving base-pair opening,<sup>154</sup> ADNA intermediates,<sup>174</sup> stretched intermediates,<sup>175,176</sup> and a zipper model, which involves stepwise propagation of a B-Z junction.<sup>177,178</sup>

Although the helical transitions are relatively slow, this does not present a problem for methodology based on geometry optimisation. In fact, the transition state theory formulation we employ to calculate the individual minimum-to-minimum rate constants using harmonic vibrational densities of states will probably be most reliable for pathways with higher overall barriers. We investigated the landscape and helix interconversion kinetics for a DNA hexamer duplex with the sequence (CGCGCG)<sub>2</sub>, using a properly symmetrised<sup>144,145</sup> version of the AMBER99bsc0 force-field,<sup>179</sup> and  $\epsilon$  and  $\zeta$  torsional corrections<sup>180</sup> with implicit solvent treated by a generalised Born model,<sup>149,150</sup> and an effective salt concentration of 1.0 M from the Debye–Hückel approximation.<sup>151</sup> Full details are available in the original report,<sup>181</sup> including comparisons with molecular dynamics runs using explicit solvent, which provide a check on the stability of the predicted structures on short timescales.

The resulting free energy landscape at 298 K is illustrated in Fig. 8, where minima separated by free energy barriers less than 5 kcal mol<sup>-1</sup> have been recursively regrouped.<sup>56,182</sup> There are two principal funnels, which can be largely assigned to left- and right-handed helices. The global free energy minimum under these conditions corresponds to an ensemble of BDNA configurations, including structures with terminal bases frayed out, stacked on top of the base-pairing partner, and *trans* Watson–Crick/sugar edge hydrogen-bonding patterns, in addition to the canonical form. Helices exhibiting B-Z junctions and ADNA structures are also present in this funnel, and are generally connected to BDNA minima *via* low downhill barriers. The B-Z junctions have the first two base-pairs in a ZDNA-like conformation, and the last three in a BDNA-like conformation, with the G3 and C10 bases at the boundary flipped out.

The funnel containing ZDNA is more structurally homogeneous, and our results support the suggestion that BDNA is more flexible. This funnel also contains conformations that we have described as left-handed BDNA.<sup>181</sup> At high energy we see stretched DNA structures, where the helix is significantly unwound. These minima are different from the SDNA state that forms during overstretching, which is thought to be a ladder-like structure, with Watson–Crick base-pairs maintained throughout.<sup>183,184</sup> The stretched forms that we have characterised have no such pairs.

A mechanistic analysis of the landscape reveals two distinct pathways (Fig. 9) for the B → Z transformation:<sup>181</sup> one involving a B-Z junction and the other sampling stretched intermediates. The stretched minima are accessed *via* helix unwinding, loss of

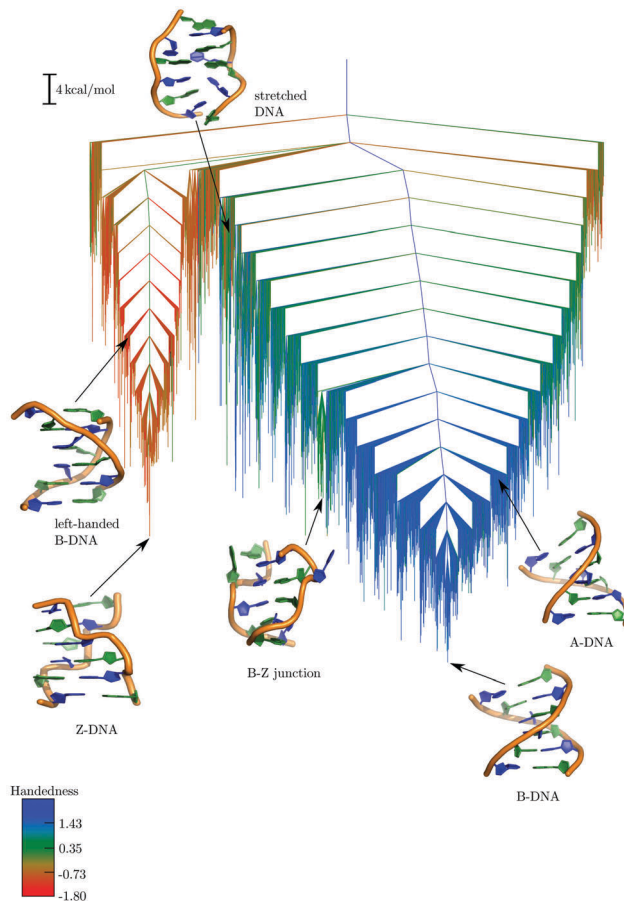


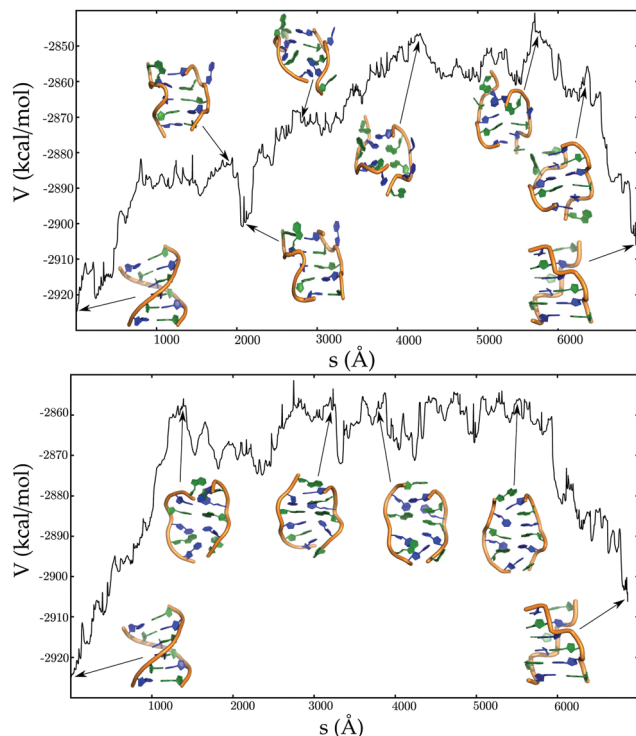
Fig. 8 Free energy landscape computed at 298 K using a regrouping threshold<sup>182</sup> of 5 kcal mol<sup>-1</sup>.<sup>181</sup> The branches are coloured according to the handedness (*H*) of the DNA structure, with positive values representing right-handed conformations and negative values representing left-handed conformations.<sup>185</sup> Some representative structures from the different conformational ensembles are also shown. Adapted from ref. 181.

Watson–Crick base-pairs, with the central guanines flipping from *anti* to *syn* conformations before the guanines near the helix termini. This ordering is reversed for the BZ-junction mechanism, which is predicted to be more favourable at 298 K, since the enthalpy barrier is lower.<sup>181</sup> However, stretched intermediates will become more competitive at higher temperatures.

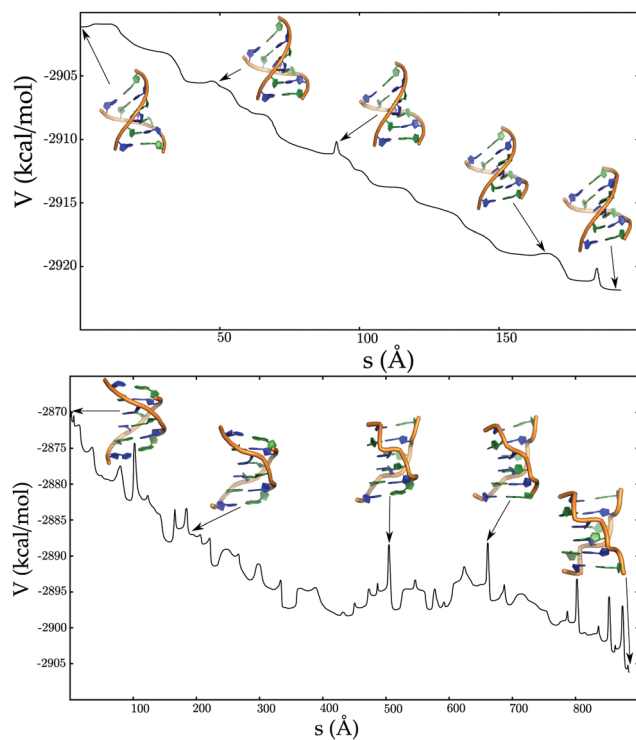
In contrast to the B → Z transition pathways, the A → B transformation is essentially downhill (Fig. 10), consistent with previous molecular dynamics simulations.<sup>186,187</sup> The small barriers encountered along the pathway correspond to repuckering of the ribose sugars, which in turn are coupled to global changes in the helix morphology. The key features of the Calladine–Drew model are correctly captured in our proposed mechanism.<sup>188</sup>

A switch in the chain topology is required for the transformation from the left-handed BDNA structure to ZDNA (Fig. 10). This transition does not require large-scale deformations of the DNA duplex, and is largely driven by concerted motions involving only a few nucleobases. The *anti* to *syn* torsional flips of the guanine bases, which are necessary to reach the ZDNA state, introduce kinetic bottlenecks in the pathway.





**Fig. 9** Potential energy ( $V$ ) as a function of the integrated path length ( $s$ ) for pathways corresponding to the BDNA to ZDNA transition, via BZ junction (top panel), and stretched intermediates (bottom panel). Adapted from ref. 181.



**Fig. 10** Potential energy ( $V$ ) as a function of the integrated path length ( $s$ ) for pathways corresponding to ADNA to BDNA transition (top panel), left-handed BDNA to ZDNA transition (bottom panel). Adapted from ref. 181.

Due to the intrinsic coupling between many degrees of freedom, probing helical transitions in a DNA duplex is a challenging task. This case study therefore provided an ideal testing ground for the potential energy landscape framework. Our simulations captured several aspects of the models proposed for helical transformations in DNA, and revealed some additional complexity, which would be difficult to capture in approaches that require the definition of a reaction coordinate.

## 7 Conclusions

The computational procedures associated with the potential energy landscapes framework are largely complementary to conventional molecular dynamics and Monte Carlo schemes. Coarse-graining the landscape, in terms of local minima and the transition states that connect them, enables us to predict structure, dynamics, and thermodynamic properties using established methods and approximations of statistical mechanics and unimolecular rate theory. In this Feature Article we have highlighted the gains in efficiency resulting from implementation of all the key geometry optimisation techniques on GPU, and from a general local rigid body formulation. Recent results for systematic analysis of these schemes have been summarised, and specific applications to protein and nucleic acid systems provide examples of interesting test cases.

Accelerating the sampling and exploration of the potential energy landscape has many benefits, especially in terms of treating larger systems. Initial surveys involving global optimisation to find favourable conformations are also a stringent test of the empirical force field. Identifying unphysical features can inform future force field development and improvement, which provides the foundations on which analysis of thermodynamic properties and kinetics depends.

Many new applications can be envisaged for biomolecules, including more coarse-grained models of mesoscopic systems. The two applications we have summarised for this report show how new insight into mechanism and design of new materials might arise. Both the coiled-coil protein and the DNA duplex landscapes exhibit multifunnel organisation, with competing low energy morphologies separated by relatively high barriers. The potential to tune such landscapes through mutations, applied fields, salt concentration, temperature, and specific ligands, raises the possibility of new design principles for multifunctional devices. The ability to explore such designs through theory and simulation, in tandem with feedback from experiment, could significantly facilitate progress in these endeavours.

## Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (RGM and KR), the Cambridge Commonwealth, European and International Trust (DC), and the Gates Cambridge Trust (JAJ).



## References

- K. Henzler-Wildman and D. Kern, *Nature*, 2007, **450**, 964–972.
- A. M. Mustoe, C. L. Brooks and H. M. Al-Hashimi, *Annu. Rev. Biochem.*, 2014, **83**, 441–466.
- R. Nussinov and P. G. Wolynes, *Phys. Chem. Chem. Phys.*, 2014, **16**, 6321–6322.
- D. J. Wales, *Energy Landscapes*, Cambridge University Press, 2003.
- J. N. Onuchic and P. G. Wolynes, *Annu. Rev. Phys. Chem.*, 1997, **48**, 545–600.
- J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.*, 2004, **14**, 70–75.
- K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl, *Annu. Rev. Biophys.*, 2008, **37**, 289–316.
- J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins*, 1995, **21**, 167–195.
- P. G. Wolynes, J. Onuchic and D. Thirumalai, *Science*, 1995, **267**, 1619–1620.
- H. Nguyen, M. Jäger, A. Moretto, M. Gruebele and J. W. Kelly, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 3948–3953.
- K. Okazaki and S. Takada, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 11182–11187.
- K. Osazaki, N. Koga, S. Takada, J. Onuchic and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 11844–11849.
- L. Cruzeiro, *J. Phys. Org. Chem.*, 2008, **21**, 549–554.
- P. Brion and E. Westhof, *Annu. Rev. Biophys. Biomol. Struct.*, 1997, **26**, 113–137.
- D. Thirumalai and C. Hyeon, *Non-Protein Coding RNAs*, Springer, Berlin, Heidelberg, 2009, vol. 13, pp. 27–47.
- D. Thirumalai and S. A. Woodson, *Acc. Chem. Res.*, 1996, **29**, 433–439.
- D. Thirumalai, D. K. Klimov and S. A. Woodson, *Theor. Chem. Acc.*, 1997, **96**, 14–22.
- K. Truex, H. S. Chung, J. M. Louis and W. A. Eaton, *Phys. Rev. Lett.*, 2015, **115**, 018101.
- H. Dietz and M. Reif, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 16192–16197.
- H. Ma, D. J. Proctor, E. Kierzek, R. Kierzek, P. C. Bevilacqua and M. Gruebele, *J. Am. Chem. Soc.*, 2006, **128**, 1523–1530.
- H. A. Scheraga, *Annu. Rev. Phys. Chem.*, 2007, **58**, 57–83.
- J. K. Noel and P. Whitford, *Isr. J. Chem.*, 2014, **54**, 1093–1107.
- R. B. Best, *Curr. Opin. Struct. Biol.*, 2012, **22**, 52–61.
- A. N. Gupta, A. Vincent, K. Neupane, H. Yu, F. Wang and M. T. Woodside, *Nat. Phys.*, 2011, **7**, 631–634.
- M. T. Woodside and S. M. Block, *Annu. Rev. Biophys.*, 2014, **43**, 19–39.
- K. Neupane, D. A. N. Foster, D. R. Dee, H. Yu, F. Wang and M. T. Woodside, *Science*, 2016, **352**, 239–242.
- M. C. Engel, D. B. Ritchie, D. A. N. Foster, K. S. D. Beach and M. T. Woodside, *Phys. Rev. Lett.*, 2014, **113**, 238104.
- A. P. Manuel, J. Lambert and M. T. Woodside, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 7183–7188.
- C. Jarzynski, *Phys. Rev. Lett.*, 1997, **78**, 2690–2693.
- G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 3658–3661.
- G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 21441–21446.
- R. H. Swendsen and J.-S. Wang, *Phys. Rev. Lett.*, 1986, **57**, 2607–2609.
- D. J. Earl and M. W. Deem, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3910–3916.
- Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
- F. Wang and D. P. Landau, *Phys. Rev. Lett.*, 2001, **86**, 2050–2053.
- G. M. Torrie and J. P. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.
- J. Kästner, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 932–942.
- A. Barducci, M. Bonomi and M. Parrinello, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**, 826–843.
- A. F. Voter, *Phys. Rev. Lett.*, 1997, **78**, 3908–3911.
- A. F. Voter, *J. Chem. Phys.*, 1997, **106**, 4665–4677.
- D. Hamelberg, J. Mongan and J. A. McCammon, *J. Chem. Phys.*, 2004, **120**, 11919–11929.
- C. E. Y. Miao, F. Feixas and J. A. McCammon, *J. Comput. Chem.*, 2015, **36**, 1536–1549.
- P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, *Annu. Rev. Phys. Chem.*, 2002, **53**, 291–318.
- C. Dellago, P. G. Bolhuis and D. Chandler, *J. Chem. Phys.*, 1998, **108**, 9236–9245.
- T. S. van Erp, D. Moroni and P. G. Bolhuis, *J. Chem. Phys.*, 2003, **118**, 7762–7774.
- A. K. Faradjian and R. Elber, *J. Chem. Phys.*, 2004, **120**, 10880–10889.
- E. Vanden-Eijnden, M. Venturoli, G. Ciccotti and R. Elber, *J. Chem. Phys.*, 2008, **129**, 174102.
- R. J. Allen, C. Valeriani and P. R. ten Wolde, *J. Phys.: Condens. Matter*, 2009, **21**, 463102.
- F. Noé and S. Fischer, *Curr. Opin. Struct. Biol.*, 2008, **18**, 154–162.
- D. J. Wales, *Curr. Opin. Struct. Biol.*, 2010, **20**, 3–10.
- V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, **52**, 99–105.
- M. R. Hoare and J. J. McInnes, *Faraday Discuss. Chem. Soc.*, 1976, **61**, 12–24.
- H. Pelzer and E. Wigner, *Z. Phys. Chem. B*, 1932, **15**, 445.
- H. J. Eyring, *J. Chem. Phys.*, 1935, **3**, 107–115.
- M. G. Evans and M. Polyani, *Trans. Faraday Soc.*, 1935, **31**, 875–894.
- D. J. Wales and P. Salamon, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 617–622.
- J. D. Chodera and F. Noé, *Curr. Opin. Struct. Biol.*, 2014, **25**, 135–144.
- D. Chakraborty, R. Collepardo-Guevara and D. J. Wales, *J. Am. Chem. Soc.*, 2014, **136**, 18052–18061.
- Y. Chebaro, A. J. Ballard, D. Chakraborty and D. J. Wales, *Sci. Rep.*, 2015, **5**, 10386.
- W. Y. Chow, D. Bihan, C. J. Forman, D. A. Slatter, D. G. Reid, D. J. Wales, R. W. Farnsdale and M. J. Duer, *Sci. Rep.*, 2015, **5**, 12556.
- C. Shang, C. S. Whittleston, K. H. Sutherland-Cash and D. J. Wales, *J. Chem. Theory Comput.*, 2015, **11**, 2307–2314.
- D. J. Wales and T. Head-Gordon, *J. Phys. Chem. B*, 2012, **116**, 8394–8411.
- J. J. de Pablo, *Annu. Rev. Phys. Chem.*, 2011, **62**, 555–574.
- D. A. Potoyan, A. Savelyev and G. A. Papoian, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, **3**, 69–83.
- W. G. Noid, *J. Chem. Phys.*, 2013, **139**, 090901.
- J. Murrell and K. J. Laidler, *Trans. Faraday Soc.*, 1968, **64**, 371–377.
- Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 6611–6615.
- D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
- D. J. Wales and H. A. Scheraga, *Science*, 1999, **285**, 1368–1372.
- K. Sutherland-Cash, D. J. Wales and D. Chakraborty, *Chem. Phys. Lett.*, 2015, **625**, 1–4.
- D. J. Wales, *Mol. Phys.*, 2002, **100**, 3285–3305.
- D. J. Wales, *Mol. Phys.*, 2004, **102**, 891–908.
- D. A. Evans and D. J. Wales, *J. Chem. Phys.*, 2004, **121**, 1080–1090.
- B. Strodel and D. J. Wales, *Chem. Phys. Lett.*, 2008, **466**, 105–115.
- O. M. Becker and M. Karplus, *J. Chem. Phys.*, 1997, **106**, 1495–1517.
- D. J. Wales, M. A. Miller and T. R. Walsh, *Nature*, 1998, **394**, 758–760.
- S. V. Krivov and M. Karplus, *J. Chem. Phys.*, 2002, **117**, 10894–10903.
- D. A. Evans and D. J. Wales, *J. Chem. Phys.*, 2003, **118**, 3891–3897.
- J. P. K. Doye, D. J. Wales, F. H. M. Zetterling and M. Dzugasov, *J. Chem. Phys.*, 2003, **118**, 2792–2799.
- M. C. Prentiss, D. J. Wales and P. G. Wolynes, *J. Chem. Phys.*, 2008, **128**, 225106.
- J. Nocedal, *Math. Comput.*, 1980, **35**, 773–782.
- D. C. Liu and J. Nocedal, *Math. Program.*, 1989, **45**, 503–528.
- P. G. Mezey, *Potential Energy Hypersurfaces*, Elsevier, Amsterdam, 1987.
- D. J. Wales, *J. Chem. Soc., Faraday Trans.*, 1992, **88**, 653–657.
- D. J. Wales, GMIN: A program for finding global minima and calculating thermodynamic properties, <http://www-wales.ch.cam.ac.uk/GMIN/>, accessed Sep 29, 2016.
- S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, 2004, **120**, 2082–2094.
- D. Sheppard, R. Terrell and G. Henkelman, *J. Chem. Phys.*, 2008, **128**, 134106.
- Y. Zeng, P. Xiao and G. Henkelman, *J. Chem. Phys.*, 2014, **140**, 044115.
- G. Henkelman and H. Jónsson, *J. Chem. Phys.*, 2000, **113**, 9978–9985.



- 90 L. J. Munro and D. J. Wales, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 3969–3980.
- 91 Y. Kumeda, D. J. Wales and L. J. Munro, *Chem. Phys. Lett.*, 2001, **341**, 185–194.
- 92 J. M. Carr, S. A. Tryubenko and D. J. Wales, *J. Chem. Phys.*, 2005, **122**, 234903.
- 93 E. W. Dijkstra, *Numer. Math.*, 1959, **1**, 269–271.
- 94 D. J. Wales, OPTIM: A program for optimising geometries and calculating pathways, <http://www-wales.ch.cam.ac.uk/OPTIM/>, accessed Sep 29, 2016.
- 95 D. J. Wales, PATHSAMPLE: A driver for OPTIM to create stationary point databases using discrete path sampling and perform kinetic analysis, <http://www-wales.ch.cam.ac.uk/PATHSAMPLE/>, accessed Sep 29, 2016.
- 96 B. Strodel, C. S. Whittleston and D. J. Wales, *J. Am. Chem. Soc.*, 2007, **129**, 16005–16014.
- 97 D. J. Wales, J. M. Carr, M. Khalili, V. K. de Souza, B. Strodel and C. S. Whittleston, *Proteins: Energy, Heat and Signal Flow*, CRC Press, 2009, p. 315.
- 98 F. H. Stillinger and T. A. Weber, *Science*, 1984, **225**, 983–989.
- 99 D. J. Wales, *Mol. Phys.*, 1993, **78**, 151–171.
- 100 F. H. Stillinger, *Science*, 1995, **267**, 1935–1939.
- 101 V. A. Sharapov, D. Meluzzi and V. A. Mandelshtam, *Phys. Rev. Lett.*, 2007, **98**, 105701.
- 102 S. F. Chekmarev and I. H. Umirzakov, *Z. Phys. D*, 1993, **26**, 373–376.
- 103 J. P. K. Doye and D. J. Wales, *J. Chem. Phys.*, 1995, **102**, 9659–9672.
- 104 J. P. K. Doye and D. J. Wales, *J. Chem. Phys.*, 1995, **102**, 9673–9688.
- 105 F. Calvo, J. P. K. Doye and D. J. Wales, *J. Chem. Phys.*, 2001, **115**, 9627–9636.
- 106 F. Calvo, J. P. K. Doye and D. J. Wales, *J. Chem. Phys.*, 2001, **114**, 7312–7329.
- 107 I. Georgescu and V. A. Mandelshtam, *J. Chem. Phys.*, 2012, **137**, 144106.
- 108 J. A. Joseph, C. S. Whittleston and D. J. Wales, *J. Chem. Theory Comput.*, 2016, **12**, 6109–6117.
- 109 H. Gohlke and M. F. Thorpe, *Biophys. J.*, 2006, **91**, 2115–2120.
- 110 Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfandtner and G. Voth, *Biophys. J.*, 2008, **95**, 5073–5083.
- 111 C. Clementi, *Curr. Opin. Struct. Biol.*, 2008, **18**, 10–15.
- 112 M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.*, 2013, **42**, 73–93.
- 113 H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole and S. J. Marrink, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 225–248.
- 114 D. J. Wales, *Philos. Trans. R. Soc. A*, 2005, **363**, 357–377.
- 115 D. Chakrabarti and D. J. Wales, *Phys. Chem. Chem. Phys.*, 2009, **11**, 1970–1976.
- 116 H. Kusumaatmaja, C. S. Whittleston and D. J. Wales, *J. Chem. Theory Comput.*, 2012, **8**, 5159–5165.
- 117 V. Rühle, H. Kusumaatmaja, D. Chakrabarti and D. J. Wales, *J. Chem. Theory Comput.*, 2013, **9**, 4026–4034.
- 118 P. Tao, A. J. Sodt, Y. Shao, G. Köniig and B. Brooks, *J. Chem. Theory Comput.*, 2014, **10**, 4198–4207.
- 119 A. G. Cochran, N. J. Skelton and M. A. Starovansnik, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 5578–5583.
- 120 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, 2006, **65**, 712–725.
- 121 D. J. Wales and J. P. K. Doye, *J. Chem. Phys.*, 2003, **119**, 12409–12416.
- 122 R. G. Mantell, C. E. Pitt and D. J. Wales, *J. Chem. Theory Comput.*, 2016, **12**, 6182–6191.
- 123 CUDA C Programming Guide, <http://docs.nvidia.com/cuda/cuda-c-programming-guide/>, accessed Sep 21, 2016.
- 124 A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2012, **8**, 1542–1555.
- 125 CudaLBFGS, <https://github.com/jwetzl/CudaLBFGS>, accessed Oct 1, 2013.
- 126 J. Wetzl, O. Taubmann, S. Haase, T. Köhler, M. Kraus and J. Hornegger, *Bildverarbeitung für die Medizin 2013: Algorithmen – Systeme – Anwendungen. Proceedings des Workshops vom 3. bis 5. März 2013 in Heidelberg*, Springer, Berlin, Heidelberg, 2013, pp. 21–26.
- 127 cuBLAS, <https://developer.nvidia.com/cublas>, accessed Sep 21, 2016.
- 128 S. Chutinimitkul, S. Herfst, J. Steel, A. C. Lowen, J. Ye, D. van Riel, E. J. Schrauwen, T. M. Bestebroer, B. Koel, D. F. Burke, K. H. Sutherland-Cash, C. S. Whittleston, C. A. Russell, D. J. Wales, D. J. Smith, M. Jonges, A. Meijer, M. Koopmans, G. F. Rimmelzwaan, T. Kuiken, A. D. Osterhaus, A. Garcia-Sastre, D. R. Perez and R. A. Fouchier, *J. Virol.*, 2010, **84**, 11802–11813.
- 129 J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2nd edn, 2006, ch. 7, pp. 176–180.
- 130 M. Karplus, *Folding Des.*, 1997, **2**, S69–S75.
- 131 J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 1987, **84**, 7524–7528.
- 132 P. E. Leopold, M. Montal and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 8721–8725.
- 133 K. Röder and D. J. Wales, *J. Chem. Theory Comput.*, 2017, **13**, 1468–1477.
- 134 E. D. Nelson and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 10682–10686.
- 135 B. V. Reddy, S. Datta and S. Tiwari, *Protein Eng.*, 1998, **11**, 1137–1145.
- 136 D. M. Taverna and R. A. Goldstein, *J. Mol. Biol.*, 2002, **315**, 479–484.
- 137 D. M. Taverna and R. A. Goldstein, *Proteins*, 2002, **46**, 105–109.
- 138 M. K. Yadav, L. J. Leman, D. J. Price, C. L. Brooks, C. D. Stout and M. R. Ghadiri, *Biochemistry*, 2006, **45**, 4463–4473.
- 139 Y. Deng, J. Liu, Q. Zheng, Q. Li, N. R. Kallenbach and M. Lu, *Chem. Biol.*, 2008, **15**, 908–919.
- 140 Y. Gambin, A. Schug, E. A. Lemke, J. J. Lavinder, A. C. M. Ferreon, T. J. Magliery, J. N. Onuchic and A. A. Deniz, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 10153–10158.
- 141 J. Ramos and T. Lazaridis, *Protein Sci.*, 2011, **20**, 1845–1855.
- 142 S. Rämisch, R. Lizatović and I. André, *Proteins*, 2015, **83**, 235–247.
- 143 G. Grigoryan and A. Keating, *Curr. Opin. Struct. Biol.*, 2008, **18**, 477–483.
- 144 E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. N. Fejer and D. J. Wales, *J. Comput. Chem.*, 2010, **31**, 1402–1409.
- 145 E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. N. Fejer and D. J. Wales, *J. Comput. Chem.*, 2012, **33**, 2209.
- 146 D. A. Case, T. Darden, T. E. Cheatham III, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. Götz, I. Kolossváry, K. Wong, F. Paesani, J. Vanicek, R. Wolf, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko and P. Kollman, *AMBER12*, 2012.
- 147 S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case, *J. Comput. Chem.*, 1986, **7**, 230–252.
- 148 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, *Comput. Phys. Commun.*, 1995, **91**, 1–41.
- 149 A. Onufriev, D. Bashford and D. A. Case, *J. Phys. Chem. B*, 2000, **104**, 3712–3720.
- 150 A. Onufriev, D. Bashford and D. A. Case, *Proteins*, 2004, **55**, 383–394.
- 151 J. Srinivasan, M. W. Trevathan, P. Beroza and D. A. Case, *Theor. Chem. Acc.*, 1999, **101**, 426–434.
- 152 D. J. Wales, *Phys. Rev. E*, 2017, **95**, 030105.
- 153 F. M. Pohl and T. M. Jovin, *J. Mol. Biol.*, 1972, **67**, 375–396.
- 154 A. H.-J. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel and A. Rich, *Nature*, 1979, **282**, 680–686.
- 155 T. J. Thamann, R. C. Lord, A. H. Wang and A. Rich, *Nucleic Acids Res.*, 1981, **9**, 5443–5457.
- 156 A. Rich, A. Nordheim and A. H. J. Wang, *Annu. Rev. Biochem.*, 1984, **53**, 791–846.
- 157 A. Rich and S. Zhang, *Nat. Rev. Genet.*, 2003, **4**, 566–572.
- 158 A. Herbert and A. Rich, *Struct. Funct. Genomics*, Springer, Netherlands, Dordrecht, 1999, pp. 53–72.
- 159 A. Jaworski, W. Zacharias, W.-T. Hsieh, J. A. Blahó, J. E. Larson and R. D. Wells, *Gene*, 1988, **74**, 215–220.
- 160 A. Jaworski, W. T. Hsieh, J. A. Blahó, J. E. Larson and R. D. Wells, *Science*, 1987, **238**, 773–777.
- 161 R. Liu, H. Liu, X. Chen, M. Kirby, P. O. Brown and K. Zhao, *Cell*, 2001, **106**, 309–318.
- 162 M. Guéron, J. Demaret and M. Filoche, *Biophys. J.*, 2000, **78**, 1070–1083.
- 163 T. M. Jovin, D. M. Soumpasis and L. P. McIntosh, *Annu. Rev. Phys. Chem.*, 1987, **38**, 521–558.



- 164 F. Azorin, A. Nordheim and A. Rich, *EMBO J.*, 1983, **2**, 649–655.
- 165 A. Nordheim and A. Rich, *Proc. Natl. Acad. Sci. U. S. A.*, 1983, **80**, 1821–1825.
- 166 A. Nordheim, E. M. Lafer, L. J. Peck, J. C. Wang, B. D. Stollar and A. Rich, *Cell*, 1982, **31**, 309–318.
- 167 W. Zacharias, A. Jaworski and R. D. Wells, *J. Bacteriol.*, 1990, **172**, 3278–3283.
- 168 M. Behe and G. Felsenfeld, *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 1619–1623.
- 169 Y. Xu, R. Ikeda and H. Sugiyama, *J. Am. Chem. Soc.*, 2003, **125**, 13519–13524.
- 170 H. Takeuchi, N. Hanamura and I. Harada, *J. Mol. Biol.*, 1994, **236**, 610–617.
- 171 T. Schwartz, J. Behlke, K. Lowenhaupt, U. Heinemann and A. Rich, *Nat. Struct. Mol. Biol.*, 2001, **8**, 761–765.
- 172 M. A. Fuertes, V. Cepeda, C. Alonso and J. M. Pérez, *Chem. Rev.*, 2006, **106**, 2045–2064.
- 173 S. Harvey, *Nucleic Acids Res.*, 1983, **11**, 4867–4868.
- 174 W. Saenger and U. Heinemann, *FEBS Lett.*, 1989, **257**, 223–227.
- 175 W. Lim and Y. P. Feng, *Biopolymers*, 2005, **78**, 107–120.
- 176 W. Lim and Y. P. Feng, *Biophys. J.*, 2005, **88**, 1593–1607.
- 177 L. J. Peck and J. Wang, *Proc. Natl. Acad. Sci. U. S. A.*, 1983, **80**, 6206–6210.
- 178 P. S. Ho, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**, 9549–9553.
- 179 A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817–3829.
- 180 M. Zgarbová, F. J. Luque, J. Šponer, T. E. Cheatham, M. Otyepka and P. Jurečka, *J. Chem. Theory Comput.*, 2013, **9**, 2339–2354.
- 181 D. Chakraborty and D. J. Wales, *Phys. Chem. Chem. Phys.*, 2017, **19**, 878–892.
- 182 J. M. Carr and D. J. Wales, *J. Phys. Chem. B*, 2008, **112**, 8760–8769.
- 183 P. Cluzel, A. Lebrun, C. Heller, J. L. Vivoy and F. Caron, *Science*, 1996, **271**, 792–794.
- 184 N. Bosaeus, A. H. El-Sagheer, T. Brown, S. B. Smith, B. Åkerman, C. Bustamante and B. Nordén, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 15179–15184.
- 185 M. Moradi, V. Babin, C. Roland and C. Sagui, *Nucleic Acids Res.*, 2013, **41**, 33–43.
- 186 L. Yang and B. M. Pettitt, *J. Phys. Chem.*, 1996, **100**, 2564–2566.
- 187 T. E. Cheatham and P. A. Kollman, *Structure*, 1997, **5**, 1297–1311.
- 188 C. R. Calladine and H. R. Drew, *J. Mol. Biol.*, 1984, **178**, 773–782.

