Structure, Thermodynamics and Folding Pathways for a Tryptophan Zipper as a Function of Local Rigidification

Jerelle A. Joseph, Chris S. Whittleston, and David J. Wales*

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

E-mail: dw34@cam.ac.uk

Abstract

We investigate how the underlying potential energy landscape for a tryptophan zipper changes as indole rings, peptide bonds, termini and trigonal planar centres are systematically grouped into local rigid bodies. The local rigid body framework results in a substantial computational speedup by effectively reducing the total number of degrees of freedom. Benchmarks are presented for the thermodynamics and folding mechanism. In general, the melting transition, as well as the precise sequence of folding events, is accurately reproduced with conservative local rigidification. However, aggressive rigidification leads to increased topological frustration and a concomitant slowing down of the global kinetics. Our results suggest that an optimal choice of local rigidification, and perhaps a hierarchical approach, could be very useful for investigating complex pathways in biomolecules.

1 Introduction

Computer simulations continue to improve our understanding of protein folding.[1–3] However, the interplay of hierarchical length- and timescales poses a significant challenge to in silico investigations. With standard techniques, conformational dynamics of proteins can only be probed over relatively short timescales, which do not capture important biological processes. Accordingly, advancements in computing code and hardware,[4–6] sampling techniques[7,8] and energy functions[9–11] have been actively pursued, to achieve longer spatio-temporal scales.[12–14] Alternatively, some of the complexity may be mitigated by developing approaches that reduce the number of degrees of freedom.[15–17]

Coarse-graining involves reducing the degree of detail used to describe a system. Numerous coarse-grained (CG) models have been proposed and implemented for biomolecules, with varying levels of success.[18–26] In one approach, amino acid side-chains and α-helices are represented as spheres and cylinders respectively;[18] in elastic network models[20,27] amino acid residues are reduced to beads interacting via inter-residue potentials. Structure-based potentials, such as Gō models,[19,28] lead to smoother landscapes, which may assist structure prediction. In these models, native-like structures are faithfully represented, while competing structures on the protein energy landscapes are penalized. Over the last decade, much effort has been expended on deriving multiscale procedures[29–31] for simulating biomolecules. These methods aim to capitalize on both the efficiency of coarse-graining and the detail present in fully atomistic computations. However, multiscale procedures rely on extensive statistical analysis and structural data obtained from ab initio computations and experiments; hence, success is based on the extent to which the models have been parametrized and optimized. Consequently, these approaches can be quite system specific and transferability between unrelated structures may be an issue.

Here we adopt a different route, based on the local rigid body (LRB) framework,[32–35] to address some of the inherent difficulties in modeling biomolecules. This framework has been benchmarked for structure prediction of model peptides using all-atom potentials[34] and, in

the current contribution, we extend it to explore the global thermodynamics and mechanics of peptide folding. Local rigidification exploits the separation of timescales[15–17,36,37] between low frequency modes and localized, fast vibrations, which suggests that specific units within the protein can be described as rigid bodies. As a result of rigidification, the number of stationary points (minima and transition states) on the potential energy surface is significantly reduced, resulting in substantial computational speedup.[34] Despite the reduction in the total number of degrees of freedom, local rigidification preserves the full atomistic resolution, and thus the resulting interatomic interactions. Hence it might be viewed as a coarse-graining of the energy landscape, rather than the potential energy function.

In the present work, we provide systematic benchmarks for tryptophan zipper 1 at different levels of local rigidification. Our results indicate that a suitable choice of local rigidification can capture the underlying physics of protein folding, and faithfully represent the global features of the energy landscape — preserving key aspects of an unconstrained description of the protein. We believe that this framework will present new opportunities for exploring the structure, dynamics and thermodynamics of biomolecules.

2 Methodology

Deciphering the folding pathway for large proteins necessitates a detailed understanding of how elementary structures, such as β-hairpins, are formed. The β-hairpin is the simplest β-structural element, composed of two hydrogen-bonded antiparallel strands connected by a short turn. Many of the fundamental characteristics of protein folding are represented in β-hairpin formation, such as hydrogen-bond and hydrophobic core stabilization, and a distinct funneled energy landscape.[38,39] Therefore, β-hairpins are good candidates for benchmarking new protein folding simulation methods.

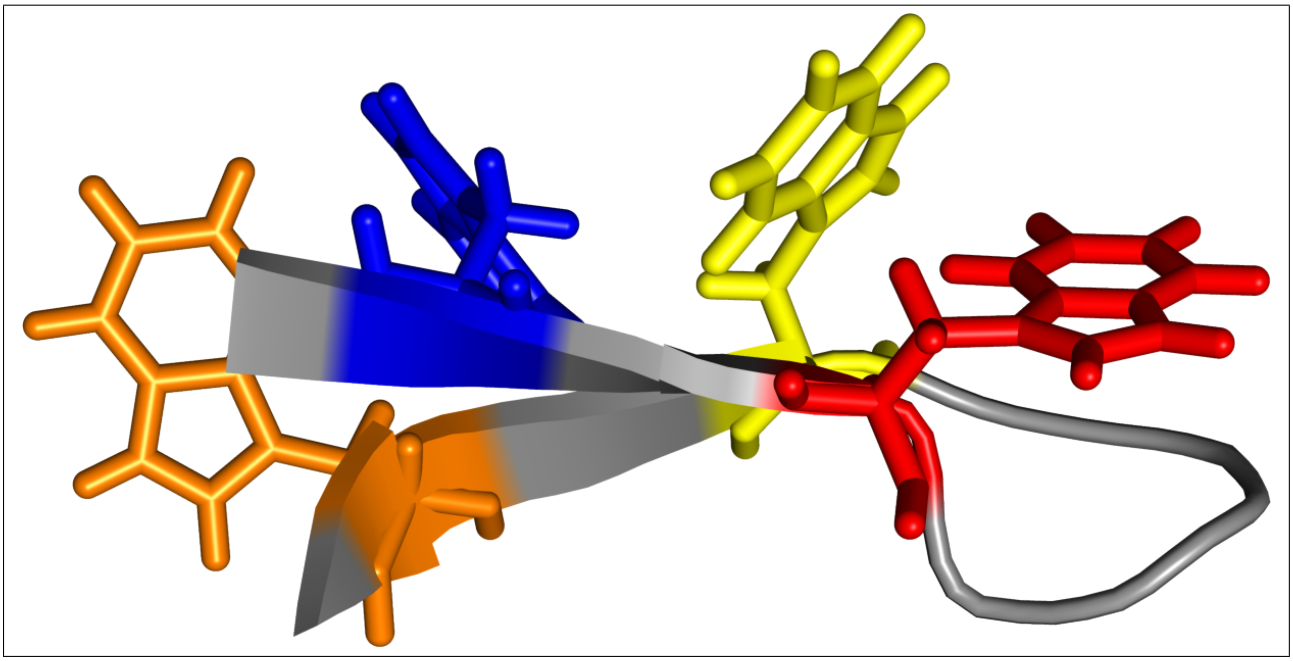


Figure 1: NMR structure for the tryptophan zipper 1 (TZ1; PDB code: 1LE0) showing the characteristic stacking of indole rings.

In this study we focus on the tryptophan zipper 1 (TZ1). TZ1 is one of the family of 12-residue β-hairpins designed by Cochran and coworkers.[40] The peptides are monomeric and adopt a well-defined tertiary structure with a unique structural motif termed a 'trpzip': cross-strand tryptophan residues interlock in a zipper-like fashion, resulting in a stable native state. In addition to their small size, the peptides fold on the microsecond timescale,[41] making them accessible in fully atomistic simulations.

The NMR structure for TZ1 is shown in Figure 1. It has a type II′ turn (turn sequence EGNK) flanked on either side by the WTW triad, and terminated by serine and lysine residues. TZ1 was represented by the AMBER99SB[42] potential energy function and the GB^{OBC} solvation potential.[43] We employ an implicit solvent representation to avoid convolution with explicit solvent degrees to freedom, which would make some of our conclusions less definitive. Since the peptide is charged, a salt concentration of 0.1 M was maintained to represent mobile counterions in solution.[44] No periodic boundary conditions were imposed on the system, and no cutoffs were set for non-bonding interactions. For calculation of effec-

tive atomic Born radii a cutoff of 25 Å was used. The AMBER potential was symmetrized, as described by Malolepsza et al.,[45] so that interconvertible permutational isomers have the same energy.

2.1 Local Rigid Body Framework

Local rigidification involves grouping sets of atoms into rigid units, each with six remaining degrees of freedom: three translations and three rotations. Rigid body representations have been exploited in many areas, including molecular dynamics simulations with explicit water,[46] structure prediction of organic compounds[47,48] and water clusters,[33,35,49] protein-protein docking[50,51] and self-assembly of virus capsids.[32,52]

2.1.1 Definitions

In the present work, rigid body translational degrees of freedom ($\mathbf{X}_I$) are defined by Cartesian coordinates of the centre of geometry,

$$\mathbf{X}_I = \frac{1}{n_I} \sum_{i \in I}^{n_I} \mathbf{x}_i, \tag{1}$$

where the number of atoms in rigid body, I, is given by n_I. The orientation of a local rigid body, relative to a fixed reference structure, is described using angle-axis variables:[32–35]

$$\mathbf{p}_I = \theta_I \hat{\mathbf{p}}_I, \tag{2}$$

where $\mathbf{p}_I$ is a rotation vector, characterizing the angle, θ_I, and axis, $\hat{\mathbf{p}}_I$, of rotation.[32,33] Rigid body reference coordinates are usually obtained from the global minimum of the potential energy surface, corresponding to the unconstrained representation.[34]

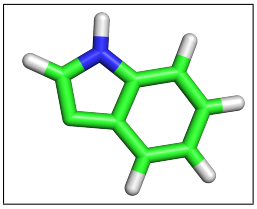
Using the local rigid body (LRB) approach, the coordinate space for the peptide was redefined in terms of mixed (atomistic and rigid body) coordinates, $\{\mathbf{x}_1, .., \mathbf{x}_n, \mathbf{X}_1, .., \mathbf{X}_N, \mathbf{p}_1, ..., \mathbf{p}_N\}$; n is the number of unconstrained atoms in the peptide and $\mathbf{x}_n$ represents the atomistic co-

ordinates of the nth free atom; N is the number of LRBs and $\{\mathbf{X}_N, \mathbf{p}_N\}$ are the rigid body coordinates of the Nth rigid body. This implementation leaves the potential energy function unchanged, although there is no need to include terms corresponding to sites in the same rigid body. To compute the potential energy of the system using an all–atom force field, we must be able to map the rigid body coordinates to the atomistic ones. Accordingly, the rotation vector $\mathbf{p}_I$ is used to construct a rotation matrix ($\mathbf{R}_I$),[53] which can be applied to the reference structure of the rigid body ($\mathbf{x}^0_{i\in I}$) to obtain the atomistic coordinates:

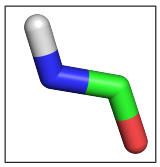
$$\mathbf{x}_{i\in I} = \mathbf{X}_I + \mathbf{R}_I \mathbf{x}^0_{i\in I}. \tag{3}$$

2.1.2 Groupings and Schemes

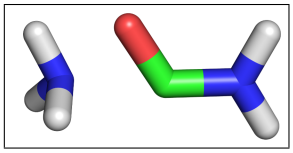
Suitable LRB groupings can be suggested from principal component analysis,[54,55] approaches developed from graph theory,[56,57] or some other metric. In this study, the LRB groupings for TZ1 were adopted from previous work;[34] namely, tryptophan rings, peptide bonds, termini and trigonal planar centres (Figure 2).



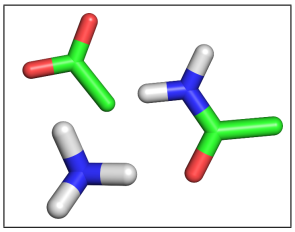
(a) tryptophan ring



(b) peptide bond



(c) termini



(d) trigonal planar centres

Figure 2: Local rigid bodies considered for tryptophan zipper 1.

These groupings were used to define several local rigidification schemes, outlined in Figure 3. The TZ1 model peptide contains 220 atoms, and the number of degrees of freedom for the unconstrained representation is therefore 660. In scheme I, aromatic rings in tryptophan residues were grouped as LRBs; the benzene and pyrrole components in each indole ring were treated separately to allow for slight bending motions. Hence, each peptide in this

scheme contains eight LRBs ($\approx$ 20 percent of the atoms) and 160 unconstrained atoms ($8 \times 6 + 160 \times 3 = 528$ degrees of freedom). Thus, scheme I represents conservative local rigidification, since only a small percentage of atoms were constrained. Conversely, scheme III represents a more aggressive scheme — with about 60 percent of the atoms grouped as LRBs ($25 \times 6 + 89 \times 3 = 417$ degrees of freedom).

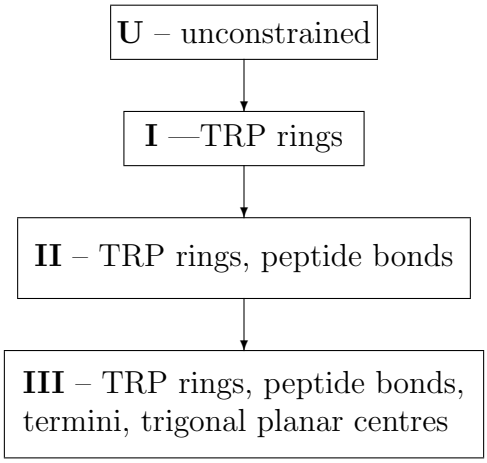


Figure 3: Systematic application of local rigidification for trptophan zipper 1. For **U** no local rigid bodies were used; for schemes **I** to **III**, increasingly larger subsets of the peptide were locally rigidified.

2.2 Potential Energy Landscapes with LRBs

The local rigidification was applied within the framework of potential energy landscape theory.[58] Conceptually, the potential energy surface (PES) supports the local minima and the transition states that connect them. Local minima are defined as stationary points where all the non-zero normal mode frequencies are real, while transition states are defined as stationary points with one imaginary normal mode frequency.[59] These stationary points constitute a kinetic transition network (KTN), from which the global thermodynamics and kinetics may be extracted. The complexity of the PES increases as the system size grows. Hence, a LRB formalism becomes appealing; since this approach effectively reduces the number of stationary points on the PES, leading to increased computational efficiency.

2.2.1 Energy Minimizations

Energy minimizations were performed using a customized L-BFGS algorithm,[60,61] in the mixed coordinate space. This approach has the advantage of reducing the number of minimization steps required for convergence.[34] Rühle et al.[35] have developed a method for computing the energy gradients with respect to generalized coordinates (mixed or atomistic), hence providing a convenient means of measuring convergence, which is invariant to coordinate transformations, as it should be.[35]

2.2.2 Building Kinetic Transition Networks

Appropriate initial endpoints for the reactant (A) and product (B) were first chosen. Here, a denatured peptide (obtained from an MD simulation at 330 K), with a high occupation probability in the vicinity of the experimental melting temperature,[40] was selected as the reactant. The product was represented by the global minimum of the potential energy surface (obtained by basin-hopping global optimization)[58,62,63] corresponding to the unconstrained peptide.

Once the endpoints were selected, the LRB scheme provided the rigid body groupings for the endpoints, which were then represented using mixed coordinates. The doubly-nudged[64] elastic band[65,66] (DNEB) procedure was then used to locate transition state candidates, which were converged further using hybrid eigenvector-following (HEF).[67,68] Transition states were subsequently connected to minima by following approximate steepest-descent paths parallel and antiparallel to the unique downhill direction. Both the DNEB and transition state refinement methods have been reformulated for use in the generalized coordinate space.[35] Iterative DNEB/HEF searches[39,69,70] eventually provided a global survey of the potential energy surface. All these procedures are implemented in the OPTIM[71] and PATHSAMPLE[72] programs, which are available for use under the GNU General Public License.

2.2.3 Depicting Potential Energy Landscapes

Disconnectivity graphs[73,74] were used to visualize the potential energy landscapes. At a given energy threshold, minima are grouped into disjoint sets ('superbasins'), where members of can interconvert without exceeding the threshold. Hence, in disconnectivity graphs, 'state–to–state' transitions of reaction pathways are replaced by 'basin-to-basin' transitions.

2.3 Thermodynamic Calculations

The partition function for the model peptide, $Z(T)$, was computed as a sum of contributions from the basins of attraction of local minima, $\sum_\alpha Z_\alpha(T)$, in the stationary point database. A harmonic approximation was used to estimate the vibrational partition function of each minimum,[75]

$$Z_\alpha(T) = \frac{n_\alpha \exp(-V_\alpha/k_B T)}{(h\bar{\nu}_\alpha/k_B T)^\kappa}; \tag{4}$$

V_α is the potential energy of minimum α, n_α is the number of distinct permutational isomers of α, $\bar{\nu}_\alpha$ is the geometric mean vibrational frequency and κ is the number of vibrational degrees of freedom.[58,75] Equilibrium statistical mechanics was then used to estimate the free energy, as well as the heat capacities, from the molecular partition function. Vibrational frequencies were computed using normal mode analysis, and within the local rigid body framework these are evaluated for the generalized coordinates by including the appropriate metric tensor.[35] Additionally, we can adapt the normal mode analysis to scale favorably with system size, by utilizing a sparse Hessian approach for larger biomolecules.

Generally, the harmonic approximation holds at low temperatures, and reliable estimates of the density of states of low-lying minima can be obtained. However, at higher temperatures, where vibrational modes are softer and anharmomic effects become more significant, corrections are needed. These can be added by employing methods such as the reaction path Hamiltonian superposition approach (RPHSA).[75] Nonetheless, we reckon that a consistent use of the HSA here is sufficient for comparing the global thermodynamics within the various

LRB schemes.

3 Results and Discussion

We begin by characterizing the unconstrained TZ1 peptide. Locally rigidified potential energy landscapes are then constructed, and their resulting topological properties are compared to those of the unconstrained representation. Next, the effects of local rigidification on the thermodynamic properties of TZ1 are assessed further by systematically evaluating the heat capacity corresponding to the various TZ1 models. Finally, we discuss how the predicted folding pathways are affected by local rigidification.

3.1 Potential Energy Landscapes

Figure 4 illustrates the potential energy landscape corresponding to the unconstrained TZ1 peptide. The landscape exhibits a prominent funnel-like bias towards the global minimum. Each branch on the potential energy (PE) disconnectivity graph represents a minimum on the PES and is colored based on the value of two order parameters, L and S. The structural order parameter L, defined by Snow et al.[41] in a previous study on the kinetics of tryptophan zippers, represents the sum of the inner native hydrogen-bond lengths and the distances between adjacent TRP rings.[41] L therefore measures the degree of compaction and can be used to distinguish between compact and extended/denatured peptides. We also define an order parameter S, which describes the orientation of the TRP rings with respect to the TZ1 backbone. Two dihedral angles $d1$ (TRP4:CZ2–TRP9:CA–TRP4:CA–TRP9:CZ2) and $d2$ (TRP2:CZ2–TRP11:CA–TRP2:CA–TRP11:CZ2) were computed and, based on the sign of these angles, S was assigned a value of either $+1$ ($d1$, $d2$ positive) or -1 ($d1$ or $d2$ negative). This order parameter was mainly used to identify folded/partially folded states on the TZ1 landscape with indole rings exhibiting non-native stacking (i.e S–value of -1 for rings on opposite faces of the hairpin or with reversed stacking compared to the native

arrangement).

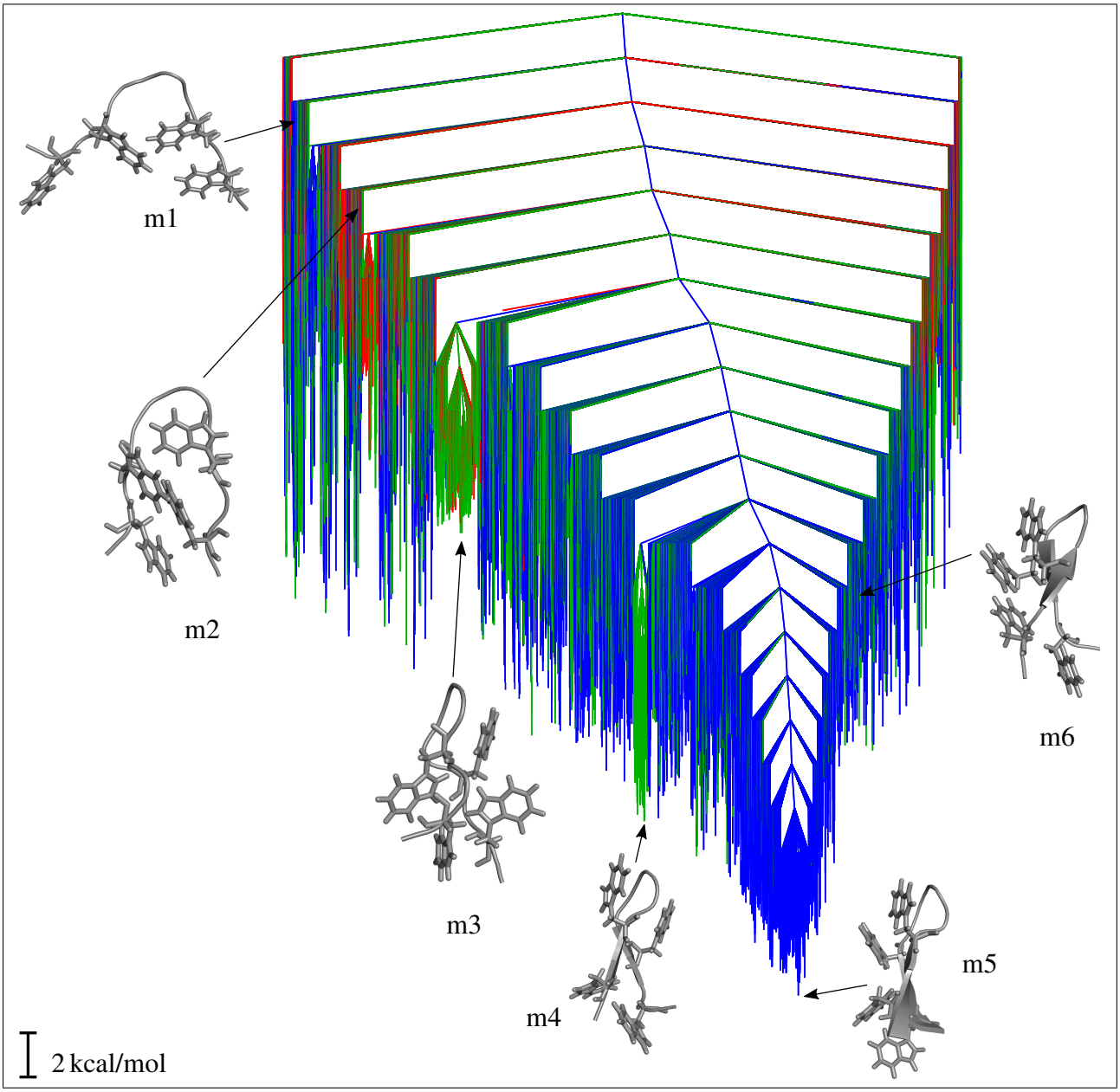
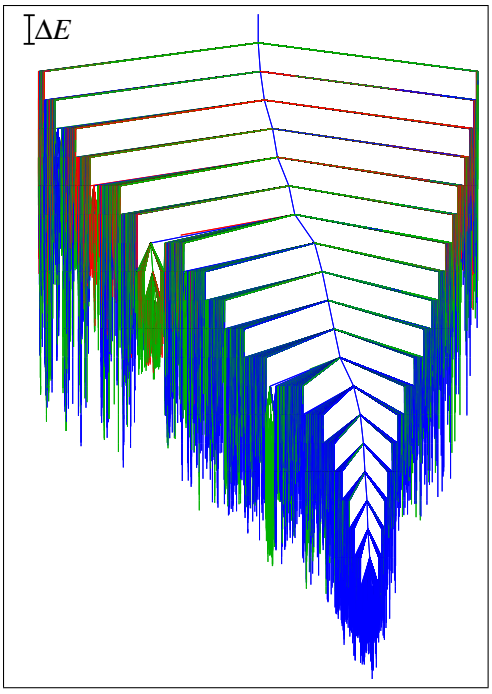


Figure 4: Potential energy disconnectivity graph for the unconstrained TZ1 peptide ($\Delta E =$ 2 kcal/mol). The branches are colored based on order parameters L (the sum of the four inner native hydrogen-bond lengths and the distances between the CD2 atoms of the three TRP pairs) and S (the orientation of the TRP rings — refer to text for description). The three main morphologies are: blue denoted F1 ($L < 60$ Å, S–value $= +1$), green denoted F2 ($L < 60$ Å, S–value $= -1$), red denoted F3 (all other minima).

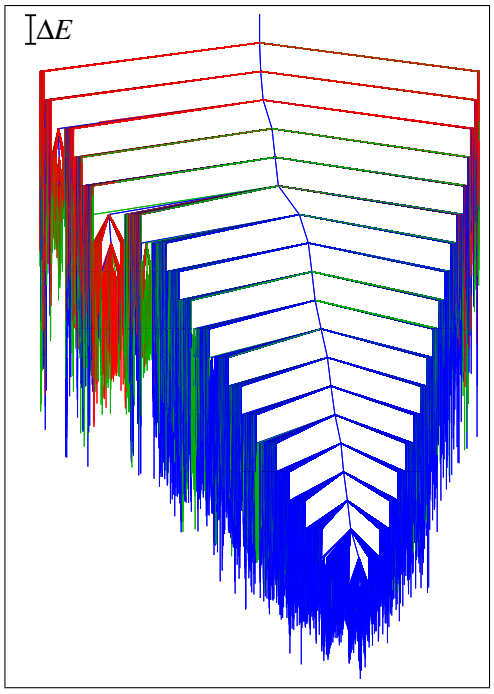
The L and S values were together used to visualize the organization of different minima

on the PE landscape. Three interspersed groups of minima were identified in the graph: F1 consists of structures with partial or complete hairpin architectures, with all TRP rings oriented on one face of the hairpin (m2, m6, m5). Several minima in F1 have all four inner native hydrogen-bonds intact; these structures constitute the bottom of the major funnel and include the global minimum (m5). F2 corresponds to conformational ensembles exhibiting some hairpin structure, but with indole rings lying on both faces (m3, m4). These hairpins can be characterized as competing structures which lead to topological frustration. Yang and Gruebele demonstrated that such structures act as kinetic traps,[76] since the reorientation of TRP rings requires that existing hydrogen-bonds must be broken and then reformed. These processes are generally associated with high energy barriers. Consequently, several hairpins in F2 are arranged in distinct subfunnels on the landscape. The final group, F3, consists of structures with residual β-hairpin content and minimal native contacts. Members of this group are located in the higher potential energy regions, where most denatured peptides reside (m1).

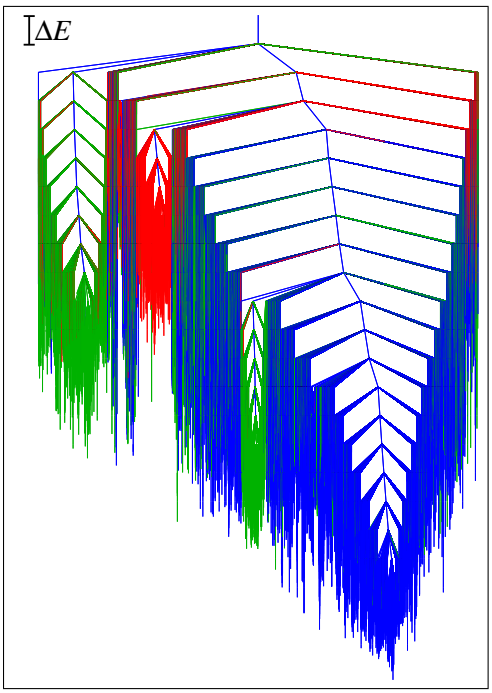
In addition to the main end points (m1 and m5), structures in each of the PE groups described above provided useful targets for building KTNs with local rigidification. Accordingly, initial folding paths, starting from the unfolded peptide and selected structures in each of the PE groups, were constructed within each of the LRB schemes. At each level of local rigidification, the resulting pathways were combined to yield a stationary point database. Minima and transition states on the unconstrained landscape were also re-optimized at the appropriate level of local rigidification and added to the corresponding database. Upon convergence of the folding rate constants, each stationary point database was analyzed using the same metrics as described in Figure 4.



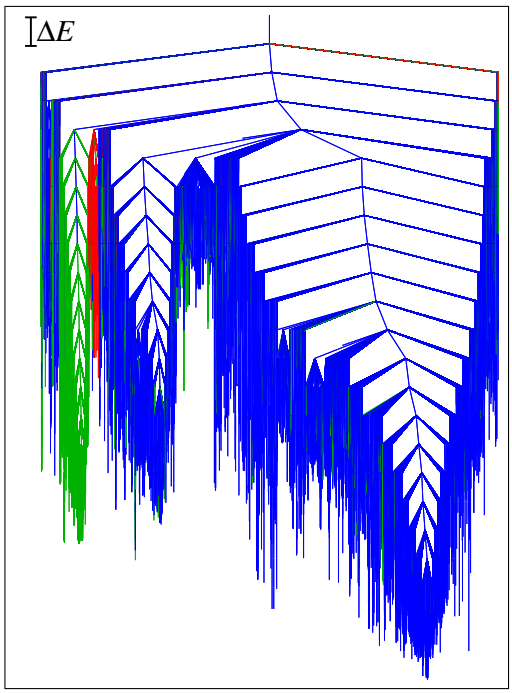
(a) unconstrained peptide



(b) I – TRP rings rigidified



(c) II – TRP rings, peptide bonds rigidified



(d) III – TRP rings, peptide bonds, termini, trigonal planar centres rigidified

Figure 5: Potential energy disconnectivity graphs for TZ1 ($\Delta E = 2\,\text{kcal/mol}$) at different levels of local rigidification. The branches are colored based on order parameters L and S, as in Figure 4. The three main PE conformational groups are: blue — F1 ($L < 60\,\text{Å}$, S–value $= +1$), green — F2 ($L < 60\,\text{Å}$, S–value $= -1$), red — F3 (all other minima), as described in the text.

Comparing the disconnectivity graphs in Figure 5, depicting the PE landscapes of TZ1 from the unconstrained representation up to aggressive local rigidification, reveals several systematic trends:

- Potential energy range — the PE range for all four graphs is similar, with a difference of approximately 64 kcal/mol between the highest and lowest transition states (Supporting Information: Figure 2). Local rigidification does, however, lead to a slight increase in barrier heights. For example, the highest and lowest transition states on the unconstrained landscape lie at −390.0 and −453.8 kcal/mol respectively, while the corresponding transition states on the most rigidified landscape lie at −388.3 and −452.7 kcal/mol. The range of energies covered by local minima on the various landscapes is comparable; on the unconstrained landscape the PE range is 50 kcal/mol, while local minima on the PE landscape for schemes I, II and III cover a range of 51, 57 and 54 kcal/mol, respectively.

- Structural heterogeneity – a diverse collection of local minima, with varying geometric rms deviations from the global minimum (Supporting Information: Figure 3), is identified in each scheme. The three PE groups identified for the unconstrained potential energy landscape are also present on the locally rigidified landscapes. Hence, we find that upon reoptimization most local minima on the unconstrained landscape are recovered on the rigidified landscapes, and the structural heterogeneity of the folding subspace is largely preserved with local rigidification. This result supports previous findings,[34] where a strong correlation was found between unconstrained and locally rigidified local minima for TZ1. This correlation is very important if the approach is to be useful.

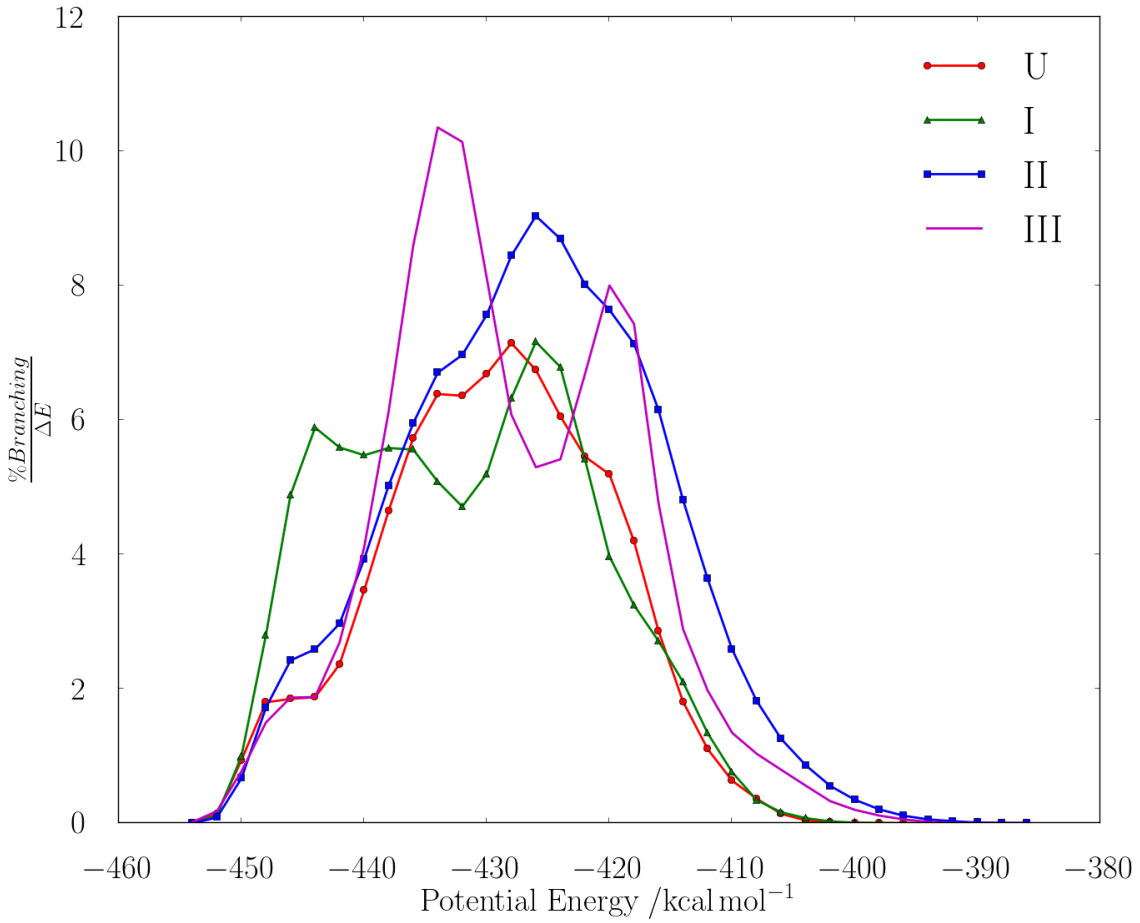


Figure 6: Surface roughness of the potential energy landscape of TZ1 ($\Delta E = 2$ kcal/mol) corresponding to the unconstrained representation (**U**), and locally rigidified representations (**I** – TRP rings, **II** – TRP rings and peptide bonds, **III** – TRP rings, peptide bonds, termini, trigonal planar centres). The surface roughness is the variation with energy in the roughness density, defined as the quotient of the percentage of minima that branch off at a particular energy level and the threshold, ΔE, used for the superbasin analysis.

- Surface roughness – Levy and Becker presented an account of how disconnectivity graphs may be used to assess surface roughness for energy landscapes.[77] In their treatment, the roughness density is taken as the quotient of the percentage of minima that branch off a given energy level and the energy threshold used for the superbasin analysis. We computed this property for our disconnectivity graphs (Figure 6). On the unconstrained landscape and the locally rigidified landscapes corresponding to schemes I and II, the maximum roughness occurs around 30 kcal/mol above the global minimum.

The overall surface roughness for scheme II is comparable to the reference landscape; however, there is a significant increase in the roughness density in the lower energy region of the disconnectivity graph when only TRP rings are locally rigidified. Conservative local rigidification creates a small initial bias to the folded state, which leads to increased sampling of native-like conformations (most minima around 10 kcal/mol above the global minimum are in F1). For scheme III, maximum surface roughness occurs closer to the global minimum (about 20 kcal/mol above) and the overall roughness is somewhat greater than that observed for the other schemes.

- Overall connectivity — as larger subsets of TZ1 are locally rigidified, the number of prominent subfunnels in the landscape generally increases. The inherent reduction in local flexibility, which is associated with the LRB framework, leads to decreased connectivity among structurally dissimilar minima. With aggressive local rigidification, scheme III, the extensive reduction in local flexibility results in increased frustration in the landscape and a dramatic change in the connectivity of basins within the F1 group (Figure 5d).

3.2 Thermodynamics of Folding

The free energy (FE) landscape,[78,79] computed at 298 K using harmonic vibrational densities of states, for the unconstrained and locally rigidified systems reveals similar trends to those observed for the PE surfaces, although there is some difference in the ordering of minima when entropy is considered (Supporting Information: Figure 4). Here we are considering free energies for individual potential energy minima, without further regrouping. To gain further insight into the effects of local rigidification on the folding thermodynamics of TZ1, we evaluate the heat capacity and compare the predicted melting temperature of TZ1 within the various LRB schemes (Figure 7).

The melting temperature (T_m) is an important thermodynamic property for proteins, as it is often used as measure of protein stability. Hence a good model should aim to reproduce

T_m. The temperature dependent equilibrium occupation probabilities of the folded and unfolded ensembles should then also be reasonably well reproduced, which translates to preservation of the main basins of attraction and phase volumes on the energy landscape when local rigidification is applied.

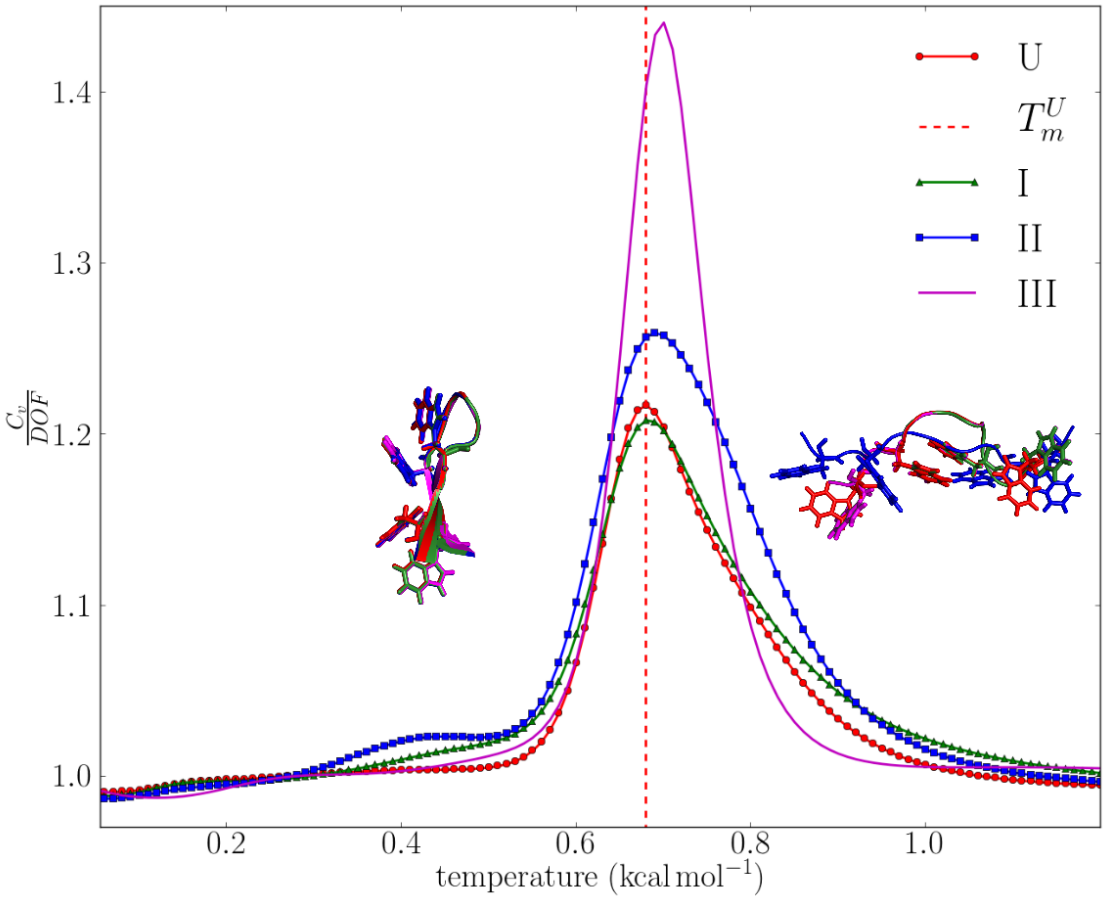


Figure 7: Constant volume heat capacity curves for TZ1 at various levels of local rigidification: unconstrained — no local rigid bodies; I — TRP rings, II — TRP rings and peptide bonds, III — TRP rings, peptide bonds, termini, trigonal planar centres treated as rigid bodies. The heat capacities are divided by the appropriate total number of degrees of freedom (DOF) and the melting temperature of the unconstrained peptide, T_m^U, is indicated. The global minimum structures of the free energy surface, computed at low (0.48 kcal/mol) and high (0.88 kcal/mol) temperatures, are superimposed on the plot; Key: red (U), green (I), blue (II), magenta (III).

For the unconstrained peptide, the melting transition is calculated at a temperature equivalent to 0.68 kcal/mol (experimental value = 0.64 kcal/mol).[40] The heat capacity curve

for scheme I is qualitatively similar to that of the unconstrained peptide and the melting temperature is accurately predicted. A small positive offset in T_m from the reference value was observed for schemes II ($T_m = 0.69\,\text{kcal/mol}$) and III ($T_m = 0.70\,\text{kcal/mol}$). These shifts in T_m suggest that local rigidification may lead to a small underestimation of the landscape entropy; hence slightly higher temperatures are needed to stabilize the unfolded state. However, this effect is minimal, and the T_m for schemes I to III roughly coincides with that of the unconstrained landscape, implying that the important basins that govern the phase transition are retained.

We also assessed the convergence of the heat capacity for the individual landscapes, to ensure that the trends observed were not artifacts of incomplete sampling (Supporting Information: Figure 5). The heat capacity curves were evaluated as a function of all the minima in the database lying below a given energy threshold. For all schemes approximately 40% of the minima are sufficient to provide a good estimate of the melting peak and T_m. Therefore, we are confident that the observable features are well converged.

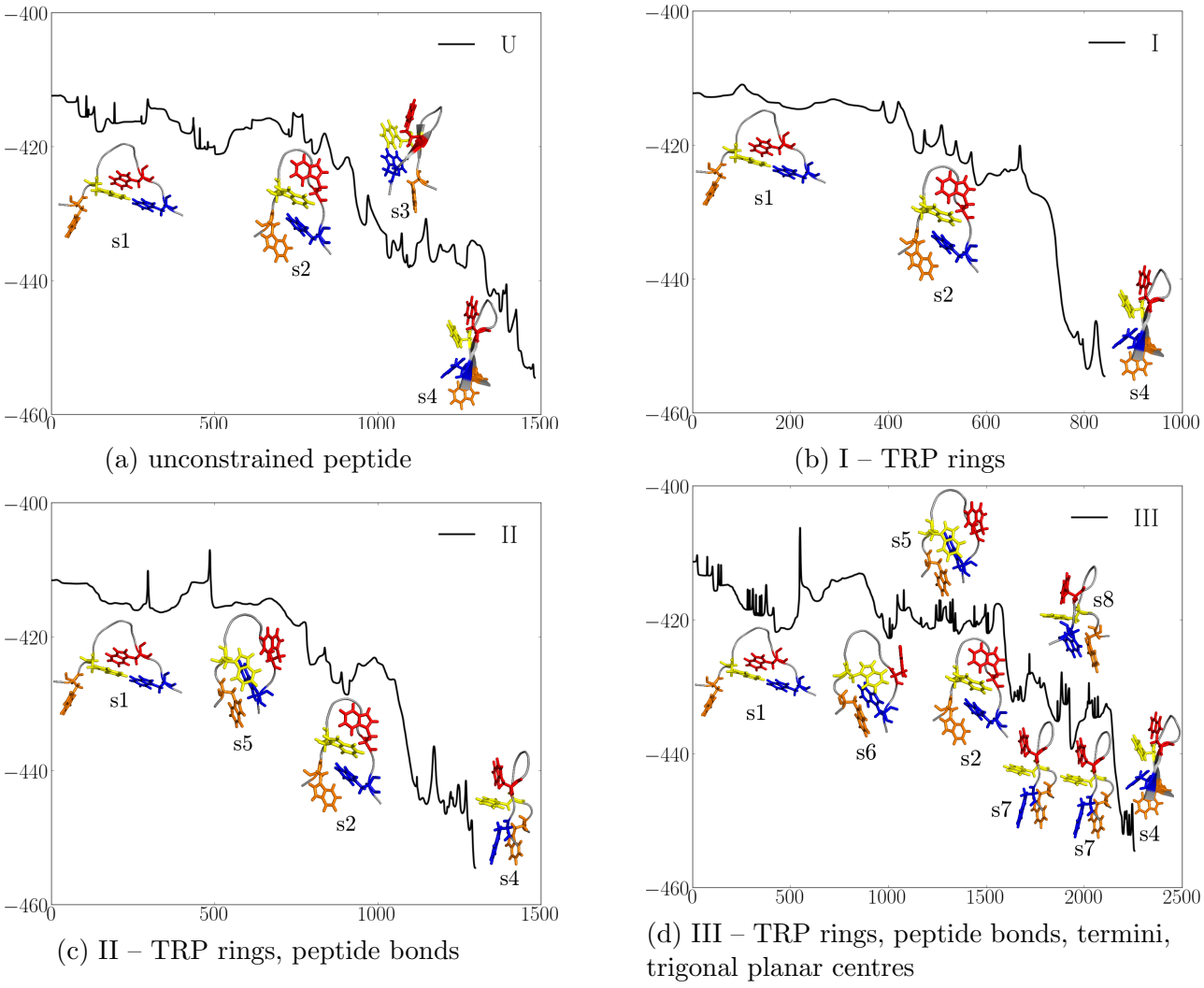
The global minimum of the FE landscape was computed for each local rigidification scheme at temperatures below and after the melting transition (Figure 7). At 0.48 kcal/mol, the overall geometric rmsd values of the FE global minimum for schemes I, II, II with respect to the unconstrained peptide are 0.47, 0.60, 0.67 Å, respectively. The corresponding deviations at 0.88 kcal/mol are 3.01, 5.79, 3.00 Å. As expected, there is greater structural variation among the FE global minima at higher temperatures, due to entropic factors. However, in general, qualitatively similar minima are responsible for the melting transition on the unconstrained and locally rigidified landscapes. In addition, the good agreement between the different FE global minima, especially at low temperatures, demonstrates the validity of local rigidification in structure prediction.

3.3 Folding Mechanism

To evaluate the effects of local rigidification on the folding pathways, we compare the individual fastest paths from the denatured state to the PE global minimum for each TZ1 model. The fastest path ($A \rightarrow B$) is the one that makes the largest contribution to the steady–state rate constant, k_{BA}^{SS} (the sum over all discrete paths with the steady–state approximation for intervening minima).[39,69,70] The main conformational states encountered on each path were then identified by employing the density–based clustering algorithm[80] available within AMBER tools;[81] this approach essentially defines an average structure for different sections of the path. Figure 8a illustrates the fastest folding pathway corresponding to the unconstrained representation of TZ1.

The unfolded state (s1) undergoes initial hydrophobic collapse to yield a compact intermediate (s2), which possess a native-like face-to-face stacking of the TRP4 and TRP9 indole rings. In the next phase of folding, the zipping process commences with the formation of some inner native hydrogen-bonds. The TRP2 and TRP11 residues of the frayed–like intermediate (s3) then rotate to complete the 'trpzip' and the final inner native hydrogen-bonds form, tethering the ends of the hairpin. This mechanism agrees with the hydrophobic–collapse model for β–hairpin formation proposed by Karplus and coworkers[82] and follows the order of TZ folding events determined by temperature jump fluorescence.[41]

On the conservatively rigidified landscape (Figure 8b), the first stage of folding is consistent with the unconstrained counterpart. However, the s3–intermediate is not encountered; rather, in one phase the inner hydrogen-bonds form, concurrently zipping the hairpin. As a result, the number of transition states on this pathway (16) is significantly less than on the reference folding path (32). Further local rigidification (scheme II, Figure 8c) leads to an increase in the relative PE barriers traversed in the early stages of folding, and a short-lived intermediate (s5) is encountered prior to forming the compact state (s2). The last phase of folding is comparable to that of scheme I. This path is comparable in length (27 transition states) to the unconstrained folding pathway.



(a) unconstrained peptide

(b) I – TRP rings

(c) II – TRP rings, peptide bonds

(d) III – TRP rings, peptide bonds, termini, trigonal planar centres

Figure 8: Variation of the total potential energy (kcal/mol) with the integrated path length (Å) for the fastest folding path from the denatured TZ1 peptide to the global minimum. The major conformational ensembles encountered along each path are shown.

With aggressive local rigidification (Figure 8d), there is substantial lengthening of the folding pathway and the number of transition states (63) encountered doubles relative to the unconstrained pathway. A significant reduction in the local flexibility of the peptide results in the formation of many unfavorable non–native contacts, increasing the PE barriers along the path. Moreover, the peptide revisits the same average structure twice (s7), as it tries to locate the native state. These results support the the observations in Figure 5d, where the landscape is noticeably more frustrated.

Finally we comment on how the folding kinetics may be affected by local rigidification. Here we adapt the procedure outlined in a previous study,[83] where the number of rearrangements on the fastest path from a given local minimum to the global minimum is computed. The distributions for the number of rearrangements can then be used to analyze the structure-seeking properties of the peptide within the various schemes. For schemes I and II the distribution is narrower than for the reference (Figure 9), indicating that there is a general acceleration in the folding dynamics when local rigidification is applied. However, for the most rigidified system, scheme III, a broader distribution is obtained and the major mode at 10–20 steps vanishes. This level of local rigidification may be too aggressive for correctly describing the folding kinetics of TZ1, since the folding is hindered by the significant loss in local flexibility.

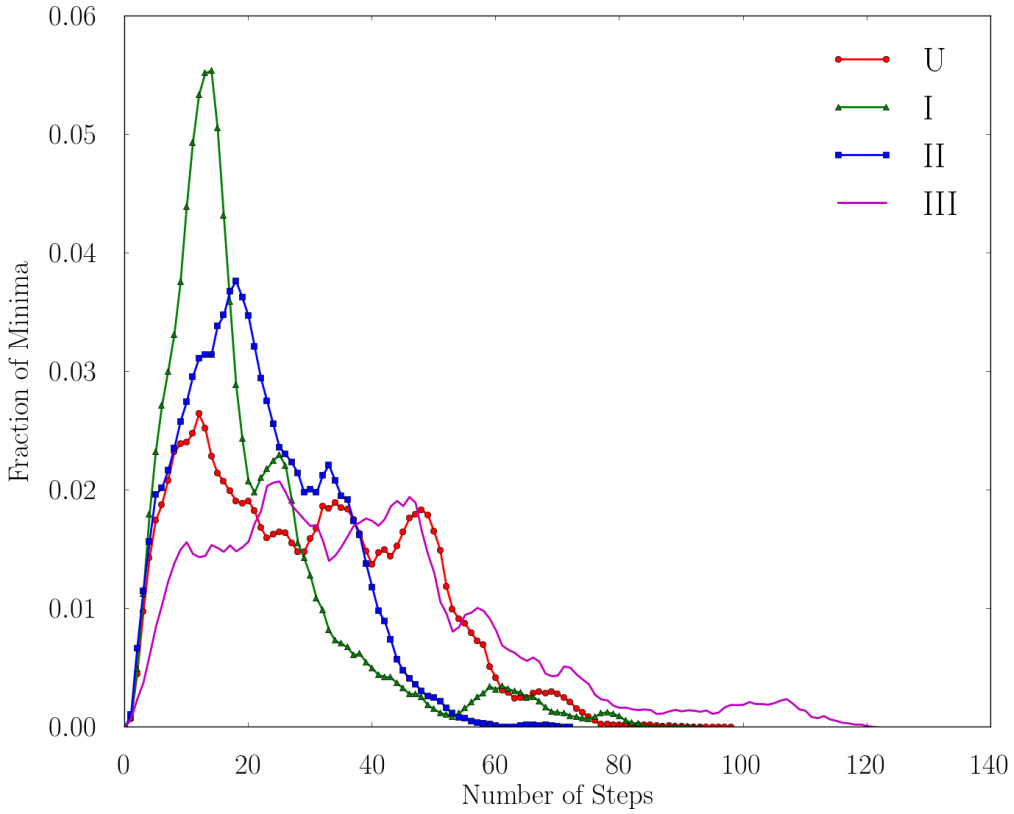


Figure 9: Distribution of the number of steps (transition states) on the fastest paths from a given minimum to the global minimum for TZ1 at different levels of local rigidification.

4 Conclusions

We have investigated how the underlying potential energy landscape for the TZ1 peptide is affected by local rigidification. The atoms associated with various functional components of TZ1 were systematically grouped into local rigid bodies and the corresponding landscape was characterized using the discrete path sampling approach. The predicted melting temperatures corresponding to the unconstrained representation and local rigid body schemes I (TRP rings) to III (TRP rings, peptide bonds, trigonal planar centres and termini) are reasonably consistent and in agreement with experiment.[40] For the unconstrained peptide, schemes I and II (TRP rings, peptide bonds), the folding mechanism corresponds to an initial hydrophobic collapse and subsequent zipping.[41,82] However, for the most rigidified system (scheme III), the peptide visits several structural ensembles that do not appear on the unconstrained pathway.

These results support the hypothesis that a subset of relevant degrees of freedom are sufficient to describe protein folding pathways. However, the local rigid body scheme must be judiciously chosen to preserve the observable properties of interest. Moreover, a representation that reproduces the folding thermodynamics does not necessarily reproduce the mechanism, which tends to be more sensitive to changes in local flexibility of the peptide. The LRB framework does not alter the atomistic resolution of the peptide, so greater accuracy for of the properties of interest (such as the folding pathways) may be conveniently obtained by relaxing the rigidified systems to their unconstrained counterparts.

The number of minima on the potential energy landscape scales with system size in a roughly exponential fashion. However, local rigidification reduces the conformational search space, by constraining degrees of freedom that fluctuate on a much faster timescale than the process of interest, decreasing the number of irrelevant minima significantly. Additionally, since the degrees of freedom within each local rigid body are frozen, corresponding terms in the potential energy function need not be calculated. In previous work, this formulation has been shown to result in a significant reduction in the computational effort required to

locate local and global minima. We anticipate that computational gains will be even more impressive for larger proteins, where regions might be locally rigidified depending on the timescale to be probed (for example, in the study of drug/ligand binding, pocket dynamics). Lastly, since the local rigid bodies implemented in this work constitute the basic building blocks of proteins, this approach is likely to be transferable between different systems.

Acknowledgement

The authors thank Dr D. Chakraborty, Dr J. M. Carr, Dr D. Schebarchov and B. E. Husic for their useful insights and suggestions. J. A. J acknowledges financial support from the Gates Cambridge Trust.

Supporting Information Available

The variation in the equilibrium occupation probabilities of the PE global minimum at the different levels of local rigidification was investigated. Additionally, we studied the variation in the distribution of the total energies of minima and transtition states with local rigidification. The PE landscapes were also compared by constructing disconnectivity graphs and coloring the branches based on the overall geometric rmsd values from the global minimum. This treatment was used to assess the structural diversity. Approximate free energy surfaces were constructed from the PE surfaces using the harmonic superposition approximation[75] to estimate the density of states. Lastly, the convergence of the heat capacity for the individual LRB schemes was tested, by investigating the variation of the C_v curve as a function of the number of minima included in the sums. This method provided a robust measure of how thoroughly the configuration space was sampled for each scheme. This information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289.

(2) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

(3) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.

(4) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. *Commun. ACM* **2008**, *51*, 91–97.

(5) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

(6) Jiang, W.; Phillips, J. C.; Huang, L.; Fajer, M.; Meng, Y.; Gumbart, J. C.; Luo, Y.; Schulten, K.; Roux, B. *Comput. Phys. Commun.* **2014**, *185*, 908–916.

(7) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(8) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schtte, C.; No, F. *J. Chem. Phys.* **2011**, *134*.

(9) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(10) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47–L49.

(11) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2015**, *11*, 3499–3509.

(12) Best, R. B. *Curr. Opin. Struct. Biol.* **2012**, *22*, 52 – 61.

(13) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

(14) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58 – 65.

(15) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(16) Saunders, M. G.; Voth, G. A. *Annu. Rev. Biophys.* **2013**, *42*, 73–93.

(17) Inglfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. *Wiley Interdiscip Rev Comput Mol Sci* **2014**, *4*, 225–248.

(18) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *106*, 421–437.

(19) G, N.; Abe, H. *Biopolymers* **1981**, *20*, 991–1011.

(20) Bahar, I.; Atilgan, A. R.; Erman, B. *Fold Des.* **1997**, *2*, 173–181.

(21) Derreumaux, P. *J. Chem. Phys.* **1999**, *111*.

(22) Buchete, N.-V.; Straub, J. E.; Thirumalai, D. *J. Chem. Phys.* **2003**, *118*.

(23) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Numerical Computer Methods, Part D*; Methods in Enzymology; Academic Press: San Diego, CA, 2004; Vol. 383; pp 66–93.

(24) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(25) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2013**, *9*, 687–697.

(26) Abeln, S.; Vendruscolo, M.; Dobson, C. M.; Frenkel, D. *PLoS ONE* **2014**, *9*, 1–8.

(27) Atilgan, A.; Durell, S.; Jernigan, R.; Demirel, M.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505–515.

(28) Abe, H.; G, N. *Biopolymers* **1981**, *20*, 1013–1031.

(29) Ahmed, A.; Gohlke, H. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 1038–1051.

(30) Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646–661.

(31) Hills, R. D., Jr; Lu, L.; Voth, G. A. *PLoS Comput. Biol.* **2010**, *6*, 1–12.

(32) Wales, D. J. *Phil. Trans. R. Soc. Lond. A* **2005**, *363*, 357–377.

(33) Chakrabarti, D.; Wales, D. J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1970–1976.

(34) Kusumaatmaja, H.; Whittleston, C. S.; Wales, D. J. *J. Chem. Theory Comput.* **2012**, *8*, 5159–5165.

(35) Rühle, V.; Kusumaatmaja, H.; Chakrabarti, D.; Wales, D. J. *J. Chem. Theory Comput.* **2013**, *9*, 4026–4034.

(36) Gohlke, H.; Thorpe, M. F. *Biophys. J.* **2006**, *91*, 2115–2120.

(37) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073–5083.

(38) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.

(39) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080–1090.

(40) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578–5583.

(41) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A* **2004**, *101*, 4077–4082.

(42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

(43) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383–394.

(44) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theor. Chem. Acc.* **1999**, *101*, 426–434.

(45) Małolepsza, E.; Strodel, B.; Khalili, M.; Trygubenko, S.; Fejer, S. N.; Wales, D. J. *J. Comput. Chem.* **2010**, *31*, 1402–1409.

(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*.

(47) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

(48) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J. *Int. J. Pharm.* **2011**, *418*, 168 – 178.

(49) Wales, D. J.; Ohmine, I. *J. Chem. Phys.* **1993**, *98*, 7257–7268.

(50) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003**, *331*, 281–299.

(51) Vakser, I. A. *Biophys. J.* **2014**, *107*, 1785–1793.

(52) Hespenheide, B. M.; Jacobs, D. J.; Thorpe, M. F. *J. Phys. Condens. Matter* **2004**, *16*, S5055.

(53) Belongie, S. Rodrigues' Rotation Formula. From MathWorld, A Wolfram Web Resource. http://mathworld.wolfram.com/RodriguesRotationFormula.html (accessed Sep 29, 2016).

(54) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169.

(55) Lange, O. F.; Grubmüller, H. *J. Phys. Chem. B* **2006**, *110*, 22842–22852.

(56) Jacobs, D. J.; Rader, A.; Kuhn, L. A.; Thorpe, M. *Proteins: Struct., Funct., Bioinf.* **2001**, *44*, 150–165.

(57) Thorpe, M.; Lei, M.; Rader, A.; Jacobs, D. J.; Kuhn, L. A. *J. Mol. Graph.* **2001**, *19*, 60–69.

(58) Wales, D. J. *Energy landscapes: Applications to clusters, biomolecules and glasses*; Cambridge University Press: Cambridge, U.K., 2003.

(59) Murrell, J.; Laidler, K. J. *Trans. Faraday Soc.* **1968**, *64*, 371–377.

(60) Nocedal, J. *Math. Comput.* **1980**, *35*, 773–782.

(61) Liu, D. C.; Nocedal, J. *Math. Program.* **1989**, *45*, 503–528.

(62) Wales, D. J.; Doye, J. P. K. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(63) Wales, D. J. GMIN: A program for finding global minima and calculating thermodynamic properties. http://www-wales.ch.cam.ac.uk/GMIN/ (accessed Sep 29, 2016).

(64) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

(65) Henkelman, G.; Jnsson, H. *J. Chem. Phys.* **1999**, *111*.

(66) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(67) Munro, L. J.; Wales, D. J. *Phys. Rev. B* **1999**, *59*, 3969.

(68) Kumeda, Y.; Wales, D. J.; Munro, L. J. *Chem. Phys. Lett.* **2001**, *341*, 185–194.

(69) Wales, D. J. *Mol. Phys.* **2002**, *100*, 3285–3305.

(70) Wales, D. J. *Mol. Phys.* **2004**, *102*, 891–908.

(71) Wales, D. J. OPTIM: A program for optimising geometries and calculating pathways. http://www-wales.ch.cam.ac.uk/OPTIM/ (accessed Sep 29, 2016).

(72) Wales, D. J. 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).

(73) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(74) Wales, D. J.; Miller, M. A.; Walsh, T. R. *Nature* **1998**, *394*, 758–760.

(75) Strodel, B.; Wales, D. J. *Chem. Phys. Lett.* **2008**, *466*, 105–115.

(76) Yang, W. Y.; Gruebele, M. *J. Am. Chem. Soc.* **2004**, *126*, 7758–7759.

(77) Levy, Y.; Becker, O. M. *Phys. Rev. Lett.* **1998**, *81*, 1126–1129.

(78) Krivov, S. V.; Karplus, M. *J. Chem. Phys.* **2002**, *117*.

(79) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2003**, *118*.

(80) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996; pp 226–231.

(81) Roe, D. R.; Thomas E. Cheatham, I. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

(82) Aaron R. Dinner, M. K., Themis Lazaridis *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068–9073.

(83) Miller, M. A.; Wales, D. J. *J. Chem. Phys.* **1999**, *111*, 6610–6616.

For Table of Contents Only

Title of Paper: Structure, Thermodynamics and Folding Pathways for a Tryptophan Zipper as a Function of Local Rigidification

Authors: Jerelle A. Joseph, Chris S. Whittleston, and David J. Wales