University of Cambridge
Department of Chemistry

Energy Landscapes for Protein Folding

This dissertation is submitted to the University of Cambridge
for the degree of Doctor of Philosophy

Jerelle Aurelia Joseph
Churchill College

September 2018

Energy Landscapes for Protein Folding

Jerelle Aurelia Joseph

Abstract

Proteins are involved in numerous functions in the human body, including chemical transport, molecular recognition, and catalysis. To perform their function most proteins must adopt a specific structure (often referred to as the folded structure). A microscopic description of folding is an important prerequisite for elucidating the underlying basis of protein misfolding and rational drug design. However, protein folding occurs on heterogeneous length and time scales, presenting a grand challenge to both experiments and simulations. In computer simulations, challenges are generally mitigated by adopting coarse-grained descriptions of the physical environment, employing enhanced sampling strategies, and improving computing code and hardware. While significant advances have been made in these areas, for numerous systems a large spatiotemporal gap between experiment and simulations still exists, due to the limited time and length scales achieved by simulation, and the inability of many experimental techniques to probe fast motions and short distances.

In this thesis, kinetic transition networks (KTNs) are constructed for various protein folding systems, via approaches based on the potential energy landscape (PEL) framework. By applying geometry optimisation techniques, the PEL is discretised into stationary points (i.e. low-energy minima and the transition states that connect them). Essentially, minima characterise the low-lying regions of the PEL (thermodynamics) and transition states encode the motion between these regions (dynamics). Principles from statistical mechanics and unimolecular rate theory may then be employed to derive free energy surfaces and folding rates, respectively, from the KTN. Furthermore, the PEL framework can take advantage of parallel and distributed computing, since stationary points from separate simulations can be easily integrated into one KTN. Moreover, the use of geometry optimisation facilitates greater conformational sampling than conventional techniques based on molecular dynamics. Accordingly, this framework presents an appealing means of probing complex processes, such as protein folding. In this dissertation, we demonstrate the application of state-of-the-art theory, combining PEL analysis and KTNs to three diverse protein systems.

First, to improve the efficiency of protein folding simulations, the intrinsic rigidity of proteins is exploited by implementing a local rigid body (LRB) approach. The LRB approach effectively integrates out irrelevant degrees of freedom from the geometry optimisation procedure and further accelerates conformational sampling. The effects of this approach on the underlying PEL are analysed in a systematic fashion for a model protein (tryptophan zipper 1). We demonstrate that conservative local rigidification can reproduce the thermodynamic and dynamic properties for the model protein.

Next, the PEL framework is employed to model large-scale conformational changes in proteins, which have conventionally been difficult to probe *in silico*. Methods based on geometry optimisation have proved useful in overcoming the broken ergodicity issue, which is associated with proteins that switch morphology. The latest PEL-based approaches are utilised to investigate the most extreme case of fold-switching found in the literature: the α-helical hairpin to β-barrel transition of the C-terminal domain of RfaH, a bacterial transcription factor. PEL techniques are employed to construct the free energy landscape (FEL) for the refolding process and to discover mechanistic details of the transition at an atomistic level.

The final part of the thesis focuses on modelling intrinsically disordered proteins (IDPs). Due to their inherent structural plasticity, IDPs are generally difficult to characterise, both experimentally and via simulations. An approach for studying IDPs within the PEL framework is implemented and tested with various contemporary potential energy functions. The cytoplasmic tail of the human cluster of differentiation 4 (CD4), implicated in HIV-1 infection, is characterised. Metastable states identified on the FEL help to unify, and are consistent with, several earlier predictions.

Declaration

The work described in this dissertation was carried out by the author in the Department of Chemistry at the University of Cambridge between October 2014 and August 2018. The contents are the original work of the author except where otherwise indicated and contain nothing that is the outcome of collaboration. The contents have not previously or concurrently been submitted for any other degree or qualification at the University of Cambridge or another institution. The number of words does not exceed 60000.

Jerelle Aurelia Joseph
September 2018

Acknowledgements

Thanks to my supervisor David Wales FRS for his support and guidance throughout my PhD. It has been a great privilege to work with such a brilliant scientist, whose dedication to his students is unmatched. Thanks to David's family for always being so welcoming, and for the lovely holiday dinners. Thanks to all the members of the Wales Group for their scientific support and camaraderie over the years. Special thanks to: Kyle Sutherland-Cash for introducing me to the group software; Chris Whittleston for his assistance during the first year of my PhD; Joanne Carr for her useful insight; Konstantin Röder and Sam Niblett for their useful suggestions; Rosie Mantell for assisting me with computer-related queries; Dmitri Schebarchov and Chris Forman for useful discussions; Debayan Chakraborty for his mentorship and friendship, and for his invaluable suggestions, many of which helped shape this thesis.

Thanks to the Department of Chemistry and the Cambridge Theoretical Chemistry Group for creating a good space for research and collaboration. Special thanks to: Daan Frenkel ForMemRS for his scientific support; Catherine Pitt and the entire IT support team for their assistance; Sue Harding for helping me with registration, and for being so caring; Lisa Masters for always being so helpful, and for making my time in the department a pleasant one; the Downing Lunch Crew for the laughs, random discussions, and advice; Room 360 members, both past and present, including Gül Güryel, Stefano Martiniani, Boris Fačkovec, Alex Cumberworth, Tom Sayer, Michael Willatt, and Myra Biedermann.

Thanks to the Gates Cambridge Trust for funding this research, and for their commitment to all scholars. Thanks to my Gates Family, and everyone who served on the Gates Council over the years; they truly made my time in Cambridge an unforgettable one, and I am honoured to have met so many passionate and inspiring people.

Thanks to everyone at Churchill College for making the college feel like home. Special thanks to: Rebecca Sawalmeh for her unparalleled dedication to the graduate students; the Porters and the college nurse Jo Livingstone for their kindness and assistance; Shelley Surtees for her support.

Thanks to the Huntingdon Road Surgery, particularly David Hayton and Tony Flinn. Thanks to Kevin Hunt of the Spinal Physiotherapy and Sports Medicine Clinic for assisting me during my leg injury. Thanks to the entire Addenbrooke's Hospital staff, particularly Michelle Fynes, Saikat Banerjee, and the wonderful nurses at the treatment centre.

Thanks to Sean McDowell for encouraging me to pursue a PhD at Cambridge. Thanks to Jeanese Badenock and Arl Anselm for their assistance with my Cambridge application. Thanks to Janet Caroo for her useful tips on surviving the cold weather, and for her kindness. Thanks to Verieux Mourillon for his mentorship.

Thanks to all my friends for keeping me sane, and for making me smile during the most difficult times of the my life. Special thanks to: Hansini Mallikarachchi, Cherysa Anselm, Olapeju Oyeyemi, Tricia Lewis, Maxine Welsh, Tsehai Grell, Giselle Volney, Natalie Rebeyev, Michelle Teplensky, Veselina Petrova, Chiara Avancini, Tetlanyo Lekalake, Dipti Jasrasaria, Thomas Aquilina, and Randall Martyr. Special thanks to Brooke Husic for her unwavering support throughout my PhD, and for her excellent feedback during the preparation of this dissertation. I am so blessed to have a friend like her.

Thanks to my family members for always being there for me. Special thanks to my EU-based family, particularly Ingrid Thomas, Jerome and Maxine Caprice, and my dear cousins Monique Caprice and Kimani Mourillon. Thanks to my in-laws Vishwanauth and Aheelia Singh for treating me like their own. Thanks to my grandparents John and Isadora Caprice for believing in my dreams. Thanks to Vishal Singh, Janice Sharplis, Elka Giraudel, Jovian Winston, and Melissa Didier for their love and support.

Thanks to my aunt Catherine Hypolite for her love and prayers, and for nurturing throughout my life. Thanks to my sisters Joycelyn, Jael and Aurelle Joseph for always cheering me on. They believe in me more than I believe in myself, and I love them so much.

Thanks to my amazing husband Vishi Singh for his love, for sacrificing his own opportunities to support me on this journey, and for always motivating me. I am so blessed to have him in my life.

Dedicated to the memories of my loving parents,
Aurelius and Victoria Joseph. To my father, for teaching me independence of thought and action, and instilling in me the value of education. To my mother, for her unconditional love, and for all the sacrifices she made to give me the opportunities she never had. I love and miss you two dearly.

Acronyms and Abbreviations

AFM	atomic force microscopy
AMBER	assisted model building with energy refinement
ANS	8-anilinonaphthalene-1-sulfonic acid
BE-META	bias-exchange metadynamics
BH	basin-hopping
BHPT	basin-hopping parallel tempering
CD	circular dichroism
CD4	cluster of differentiation 4
$\mathrm{CD4}_{\mathrm{RP}}$	cluster of differentiation 4 receptor peptide
CFA	Coulomb field approximation
CG	conjugate gradient
CHARMM	chemistry at Harvard macromolecular mechanics
COSY	correlation spectroscopy
CTD	carboxy terminal domain
DNEB	doubly-nudged elastic band
DPS	discrete path sampling
DSSP	define secondary structure of proteins
FCS	fluorescence correlation spectroscopy
FE	free energy
FEL	free energy landscape
FES	free energy surface
FRET	Förster resonance energy transfer

FTIR	Fourier transform infrared
GB	generalised Born
GROMOS	Groningen molecular simulation
HEF	hybrid eigenvector-following
HIV-1	human immunodeficiency virus type 1
^{1}H-NMR	hydrogen nuclear magnetic resonance
HSA	harmonic superposition approximation
IAEDANS	5-((((2-iodoacetyl)amino)ethyl)amino)naphthalene-1-sulfonic acid
IAF	iodoacetamidofluorescein
IDP	intrinsically disordered protein
IDR	intrinsically disordered region
IPolQ	implicitly polarised charge
IR	infrared
KTN	kinetic transition network
L-BFGS	limited-memory Broyden-Fletcher-Goldfarb-Shanno
LPS	lipopolysaccharide
LRB	local rigid body
MC	Monte Carlo
MD	molecular dynamics
MFPT	mean first passage time
MHC-II	Major Histocompatability Complex class II
MIN	minimum
mRNA	messenger ribonucleic acid
MSM	Markov state model
Nef	negative factor
NGT	new graph transformation
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy

NTD	amino terminal domain
OPLS	optimised potentials for liquid simulations
ops	operon polarity suppressor
PB	Poisson-Boltzmann
PDB	protein data bank
PE	potential energy
PEL	potential energy landscape
PES	potential energy surface
PT	parallel tempering
PUMA	p53 upregulated modulator of apoptosis
QCI	quasi-continuous interpolation
RDC	residual dipolar coupling
REM	replica exchange method
REMD	replica exchange molecular dynamics
RET	replica-exchange-with-tunnelling
RHP	random heteropolymer
RNAP	ribonucleic acid polymerase
RPHSA	reaction path Hamiltonian superposition approach
SAXS	small angle X-ray scattering
SD	steepest-descent
rms	root-mean-square
rmsd	root-mean-square deviation
TCR	T cell receptor
TIS	transition interface sampling
TPS	transition path sampling
tREMD	temperature replica exchange molecular dynamics
TS	transition state
TSG	transition state guess
TST	transition state theory

TZ1	tryptophan zipper (trpzip) 1
Vpu	viral protein U
WHAM	weighted histogram analysis method

Publications

Chapter 3

J. A. Joseph, C. S. Whittleston and D. J. Wales. *Structure, Thermodynamics, and Folding Pathways for a Tryptophan Zipper as a Function of Local Rigidification.* Journal of Chemical Theory and Computation **2016**, 12 (12), 6109–6117.

Chapter 4

J. A. Joseph, D. Chakraborty, and D. J. Wales. *Energy Landscape for Fold-Switching in Regulatory Protein RfaH* (submitted).

Chapter 5

J. A. Joseph and D. J. Wales. *Intrinsically Disordered Landscapes for Human CD4 Receptor Peptide* (submitted).

Other Publication(s)

I have also contributed to the following publication during my PhD:
J. A. Joseph, K. Röder, D. Chakraborty, R. G. Mantell and D. J. Wales. *Exploring Biomolecular Energy landscapes.* Chemical Communications **2017**.

Contents

1

Introduction

"There is pleasure in recognising old things from a new viewpoint."

– Richard Feynman

The almost 70-year inquiry into protein structure and dynamics has, at every juncture, compelled researchers to confront the problem from a new viewpoint. Before 1951, it was believed by many that proteins were amorphous entities. That year Sanger and Tuppy demonstrated that proteins were built from amino acids (§ 1.1). In 1958, electron density maps from X-ray crystallography studies strongly suggested that proteins formed well-defined structures (§ 1.1). These two discoveries, in particular, raised several questions, arguably the most important: how is protein structure determined from the amino acid sequence? Four years later, Anfinsen's work on renaturation of ribonuclease (§ 1.2.1) marked a major turning point, and it was at that juncture that the 'protein folding problem' was born. Subsequent paradoxical discussions by Levinthal (§ 1.2.2) on protein folding would fuel the field for at least two decades, as researchers sought to propose models (§ 1.2.3) that could explain how proteins fold. Towards the end of the 1980s it became apparent that no single model could fully account for protein folding. In 1987, principles learnt from spin glasses were applied to study protein folding phenomena, and a new holistic view of folding emerged: the energy landscape perspective (§ 1.2.4).

Experimental work on protein folding (§ 1.3) has presented a grand challenge to scientists, fostering creativity and advancement of new technologies in the field, which have led to important insights. On another front, computer "simulation studies have augmented and directed development of the modern landscape perspective of protein folding"[1] by revealing intricate details, particularly those not amenable to experiment (§ 1.4). Indeed, the first molecular dynamics simulations in 1977 challenged the view of proteins as static structures and recast them as dynamic entities (§ 1.4.1). Computer investigations, however, have presented their own challenges; some of which have necessitated the development of enhanced sampling techniques (§ 1.4.2), to achieve time scales trivially probed by experiments. Orthogonal viewpoints and techniques have materialised, each with varying degrees of success. Certainly, the field of protein folding has benefited, and continues to benefit, from researchers approaching the problem with a different lens. This dissertation follows in the same vein.

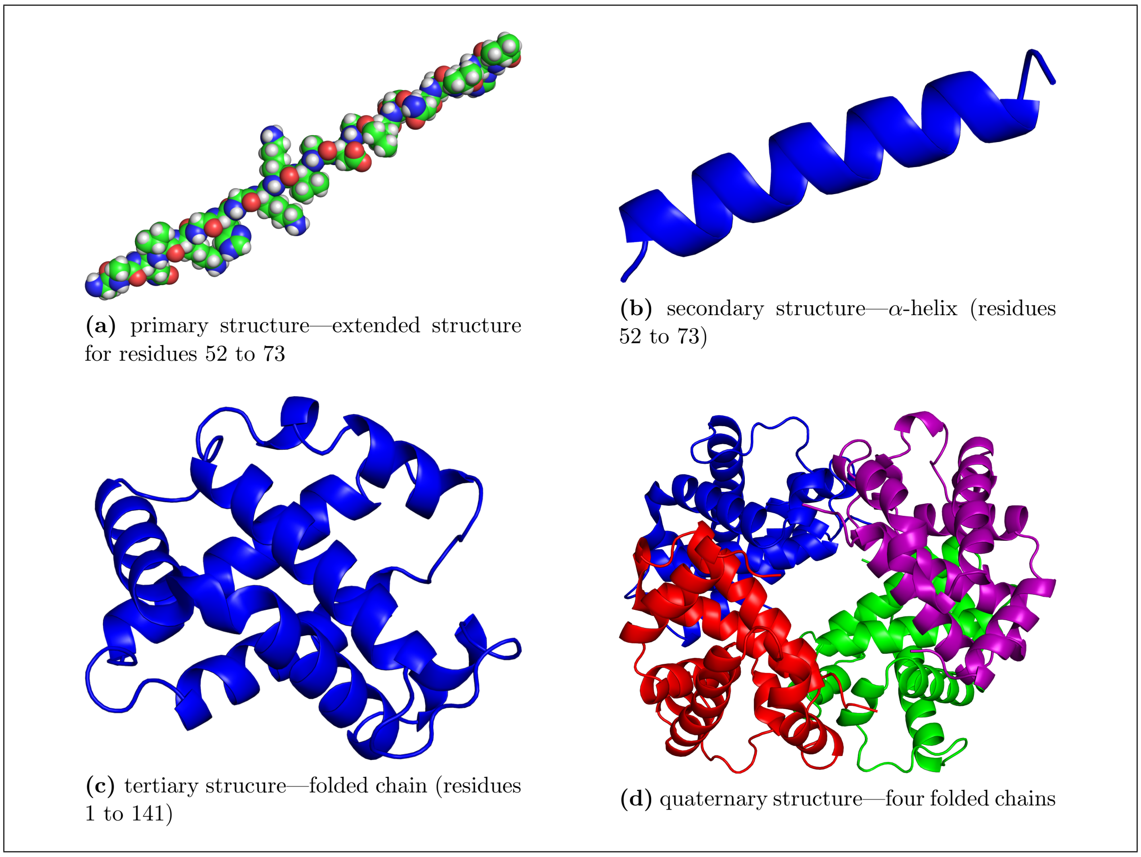
1.1 Fundamental Levels of Protein Structure

In 1951, Sanger and Tuppy's seminal work on the sequencing of insulin[2,3] transformed our understanding of protein structure: from seemingly amorphous entities to highly ordered unbranched biopolymers. At the basic level, proteins are synthesised from amino acid monomers, each monomer consisting of an amino group (NH_2), a carboxyl group (CO_2H), a unique side-chain (R) (which may be acidic, basic, polar or non-polar) and a hydrogen atom coordinated to a C_α atom.

During protein synthesis, amino acids are linked via peptide bonds (OC–NH) to yield polypeptide chains. The sequence of amino acids constitutes the primary structure of a protein (Figure 1.1a). In the cell, polypeptide chains are assembled on the ribosomes from the amino end (N-terminal) to the carboxyl end (C-terminal) of the amino acids. This arrangement results in a net dipole along the main polypeptide chain (backbone) and encourages the formation of intramolecular hydrogen-bonds ($NH \cdots OC$).

Steric interactions of the side-chains play a crucial role in modulating the periodicity of hydrogen-bonding along the backbone, and ultimately dictate which secondary structural elements are formed. In naturally occurring proteins, α-helices (Figure 1.1b) and β-sheets are the most common types of secondary structures.[4] Hydrogen-bonding between –NH and –CO groups four residues[*] apart give rise to α-helices. Whereas, β-sheets develop when hydrogen-bonds form between –NH and –CO groups in adjacent segments of a polypeptide chain, and are characterised as either antiparallel or parallel β-sheets, depending on how the segments are aligned.

[*] amino acids within polypeptide chains



(a) primary structure—extended structure for residues 52 to 73

(b) secondary structure—α-helix (residues 52 to 73)

(c) tertiary strucure—folded chain (residues 1 to 141)

(d) quaternary structure—four folded chains

Figure 1.1: Levels of protein structure depicted using haemoglobin.

Another way of distinguishing protein secondary structure is based on the dihedral angles along the backbone; specifically, the ϕ and ψ angles that describe the rotation about the N–C$_\alpha$ and C$_\alpha$–C bonds, respectively (Figure 1.2). Whereas the peptide bonds are restricted in the trans configuration,[†][5] notable variation is observed in ϕ and ψ angles. Ramachandran and co-workers investigated the distribution of ϕ and ψ angles[6] for highly resolved protein structures and found that not all combinations of ϕ and ψ angles are allowed. Furthermore, distinct regions of high density were identified for α-helices and β-sheets. Thus the well-known Ramachandran plot is characterised by the two-dimensional ϕ verses ψ space.

[†] with the exception of proline, the C$_\alpha$–C–N–C$_\alpha$ dihedral, ω, $\approx 180^\circ$

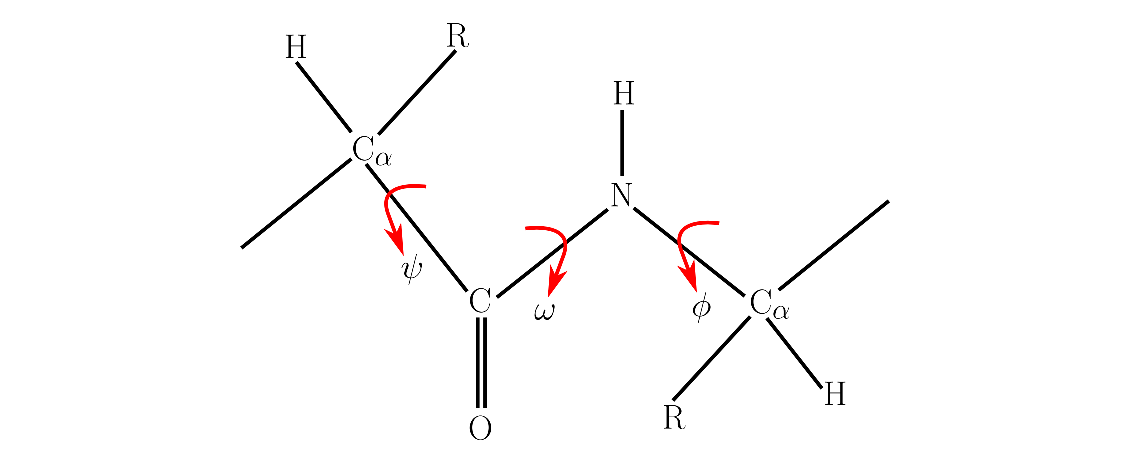


Figure 1.2: Protein backbone dihedral angles. ψ, ω, and ϕ angles describe the rotation about the C_α–C, C–N, and N–C_α bonds, respectively.

A given polypeptide chain may consist of several secondary structural elements along its length. Intramolecular hydrogen-bonding, hydrophobic, and electrostatic interactions, or the formation of disulfide bonds (via cysteine residues), between complementary side-chains, often lead to more compact structures—defining the tertiary level of protein structure (Figure 1.1c). The first three-dimensional protein structure was solved for myglobin,[7] which consists of a single polypeptide chain folded into a compact bundle of eight α-helices connected by loops. Later, Perutz and colleagues demonstrated that the structural hierarchy of proteins may extend even further to the quaternary level, with the discovery of the crystal structure of haemoglobin.[8] Haemoglobin is characterised by four folded polypeptide chains (subunits) linked via non-covalent intermolecular interactions, forming a distinct tetrahedral arrangement (Figure 1.1d).

Discovery of these intricate protein structures, and others, raised the fundamental question: what rules govern protein folding? In the following sections, various theories for protein folding are summarised.

1.2 Protein Folding Theories

1.2.1 Thermodynamic hypothesis

During the second half of the 20th century, scientists sought to discover the underpinnings of protein folding. The first piece of critical information came from the lab of Anfinsen. At the time, Anfinsen and colleagues were conducting *in vitro* studies on the refolding of ribonuclease.[9] This protein, which contains four disulphide bonds, was first treated with a chemical denaturant (a cocktail of urea and mercaptoethanol or thioglycolic acid), in order to break the disulphide bonds, while keeping the other covalent bonds intact. The denatured (unfolded) protein contained no evidence of its native structure (i.e. the characteristic fold determined by experiment). However, when the denaturant was removed the protein spontaneously refolded—reforming its native disulphide bonds, despite the numerous alternative ways that eight –SH groups could potentially bond.

Anfinsen hypothesised, "The three-dimensional structure of a native protein in its normal physiological milieu is the one in which the Gibbs free energy of the whole system is its lowest,"[10,11] (the Thermodynamic Hypothesis). It follows from the Thermodynamic Hypothesis that the native conformation of the protein is determined entirely by its amino acid sequence in thermodynamic equilibrium.

1.2.2 Levinthal's paradox

Given that protein structure is governed by the amino acid sequence, Levinthal outlined how improbable it would be for a protein (even with a few hundred amino acids) to fold to its free energy minimum, by sampling all possible amino acid conformations. He explained further that even if the dimensionality of the problem was reduced (e.g. by considering only the backbone and side-chain rotations), the time taken to fold by a random search would be unrealistically long.[12–14] This dilemma is commonly cited in the literature as Levinthal's Paradox.

Levinthal's thoughts on protein folding propelled subsequent discussions on the theory—many of the models developed in the next three decades would aim to reconcile Levinthal's paradox. The basic idea was that there must be some simplified mechanism that explains how proteins fold.

1.2.3 Models of protein folding

Sequential and nucleation models were among the first postulates for how proteins fold. The premise of these theories was that since the volume of the configuration space is so large, there must be some initial event or unique sequence of events that leads to folding, thus reducing the subsequent number of possibilities. Levinthal considered the case for specific folding pathways, which might guide the protein from the denatured state to the folded state.[13] In the Cooperative Sequential Model,[15] it was argued that, in addition to following a unique pathway (through successive intermediates), protein folding is initiated by a nucleation event. A similar model was proposed by Wetlaufer—the Nucleation Model[16]—in which the rate of folding depends on the formation of an initial nucleus (a small localised region). He further suggested that the nucleus grows by either adding neighbouring residues (fast kinetics) or distal residues (slow kinetics).

In subsequent nucleation-based models, namely the Stage-wise Mechanism[17] and the Cluster Model,[18] it was proposed that there exist multiple nucleation sites (so-called "centres of crystallisation" or clusters) along the polypeptide chain, which eventually merge or collapse to produce the native structure.

Around the same time, Karplus and Weaver presented arguments for the Diffusion–Collision Model.[19,20] In this model, protein folding begins with the formation of transient segments of secondary structure (microdomains). The authors rationalised that since microdomains are made up of a small number of amino acids, the protein could efficiently sample all the available conformations for a given microdomain, thereby avoiding Levinthal's paradox. Once formed, microdomains were supposed to move diffusely and collide with each other. The rate-limiting step was attributed to the formation of microdomain intermediates—produced when collisions lead to coalescence. Unlike classical nucleation models, which were primarily qualitative, the Diffusion–Collision model provided recipes for extracting quantitative information about the folding process; for example, formulations for the folding rate (or rate of coalescence) were derived based on the physical properties of the microdomains.[20]

In the Noninteracting Local Structure Model,[21] a statistical mechanical approach to the protein folding problem was taken. A local structure is defined as a continuous segment of the polypeptide chain that adopted an equivalent conformation in the native structure. An important element of the model is that the interactions between local structures are assumed to be negligible. Additionally, the free energy of a given local structure is estimated from the atomic coordinates of the native configuration, thus providing a means of computing the partition function. This simplified model was used to demonstrate how protein folding might proceed by first forming local structures, which subsequently grew or merged (Growth–Merge Model)[21] to yield the native structure.

A year later, Kim and Baldwin described the Framework Model[22] for protein folding. Experimental evidence at the time suggested the existence of folding intermediates that contained significant secondary structure. Consequently, the authors argued that during folding, hydrogen-bonded secondary structure is formed first, followed by tertiary interactions (Figure 1.3). Hence, in the Framework Model, it was suggested that stable secondary structure formed independently of tertiary structure.

Dill proposed that protein folding is driven by the association of hydrophobic residues to avoid contact with the solvent—the Hydrophobic Collapse Model.[23] The protein would undergo rapid collapse around the hydrophobic side-chains and then fold slowly, from the compact intermediate to the native state (Figure 1.3). In this model, the intermediate contained very little secondary structure, and was therefore in direct contrast to the Framework Model.[22]

Finally, in the Nucleation–Condensation Model,[24] strong arguments for two-state folding (i.e. lack of folding intermediates) were presented. Unlike earlier nucleation models, in which the nucleus was defined as a small incipient localised region, in this model the nucleus was assumed to be large and diffuse, and emerged in the transition state. Thus, the nucleus would correspond to the best-formed interactions in the transition state, and be stabilised by both local (between neighbouring residues) and long-range (between distal residues in sequence) interactions. Moreover, in nucleation–condensation, secondary and tertiary interactions occur simultaneously, and folding can therefore be described as a two-state process (Figure 1.3).

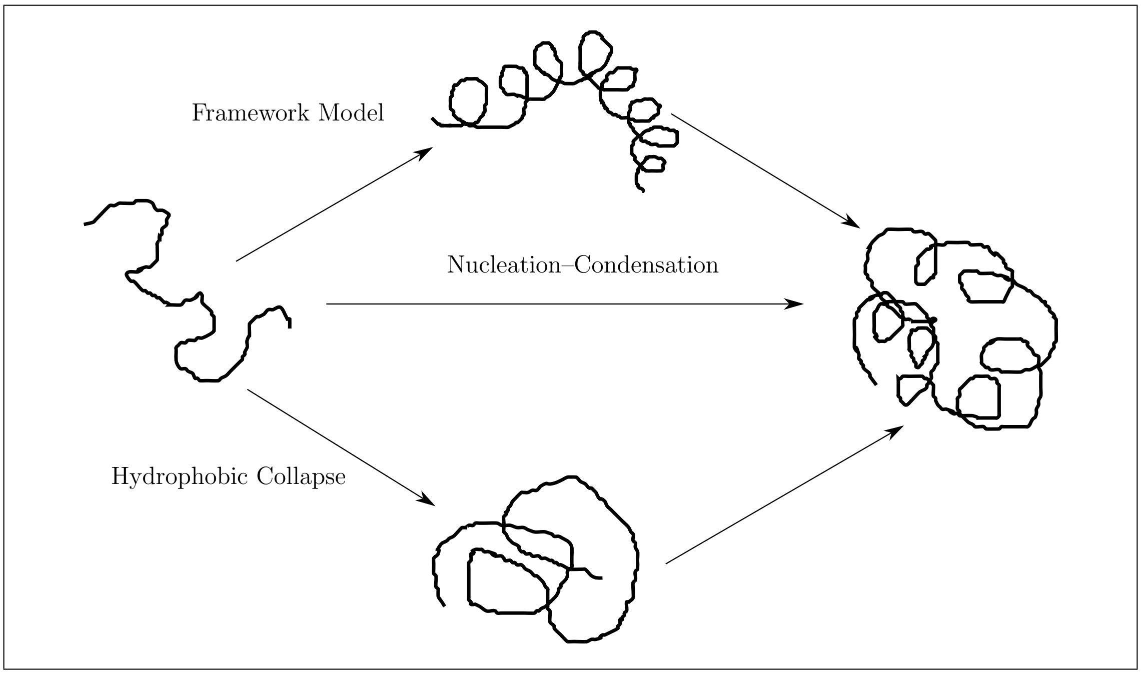


Figure 1.3: Illustration of the Framework, Hydrophobic Collapse and Nucleation–Condensation models of protein folding. Refer to the text for description.

The preceding models, though not an exhaustive list, give an idea of how thought has evolved in the field of protein folding since Anfinsen's experiments. Although the initial impetus to develop such models was to resolve Levinthal's paradox, many of the later theories sought to account for experimental observations (§ 1.3). However, as more experimental evidence became available in support of both two-state and sequential folding, many models became inadequate or obsolete. To obtain a unified theory for protein folding, a completely different view of the problem became imperative. In the next section, the energy landscape perspective for protein folding is outlined.

1.2.4 Energy landscape perspective

Each residue in a polypeptide chain can adopt many different stable configurations. Therefore, compared to an ordinary chemical reaction, the number of degrees of freedom in protein folding is considerably greater. Consequently, the reactant (denatured protein) and product (native fold) in the folding reaction are distinctively heterogeneous, to various extents; unlike typical chemical reactions, in which these states are generally homogeneous. Accordingly, the folding reaction is itself extremely heterogeneous, and a suitable model for protein folding should explicitly account for this heterogeneity.

In the Energy Landscape Model,[25–28] the organisation of the protein free energy landscape is considered in terms of ensembles of structures, wherein the structures in a given ensemble have similar conformations and energies. Therefore, protein folding can be regarded as a progressive organisation of ensembles, and statistical mechanics may be applied to study this process. The three key postulates of the model, which help explain protein folding and avoid Levinthal's paradox, are: under physiological conditions (1) the free energy of the folded state is lower than the denatured state, (2) all states are not equally probable, and (3) the energy landscape is inherently biased towards the native state. These assumptions are discussed further below.

Entropy and enthalpy (i.e. potential energy), which together encode the free energy, are the main parameters of the energy landscape model. In the unfolded ensemble, interactions are predominantly weak and of the same order of magnitude. Hence, diverse conformations with comparable energies are accessible, and the entropy of the denatured state is high. Conversely, in the folded state interactions are highly stabilising, and there are smaller fluctuations in structure, corresponding to reduced configurational entropy upon folding. Therefore, the potential energy gradient from the denatured to the native state favours folding, whereas the entropy gradient change opposes folding. Thus, protein folding is achieved by a delicate balance of these energy terms[‡]; in which the rate of decrease in the potential energy

[‡] The balance of energy terms must also include changes in solvent entropy and enthalpy during folding.

exceeds the reduction in the entropy, as more compact states are formed.

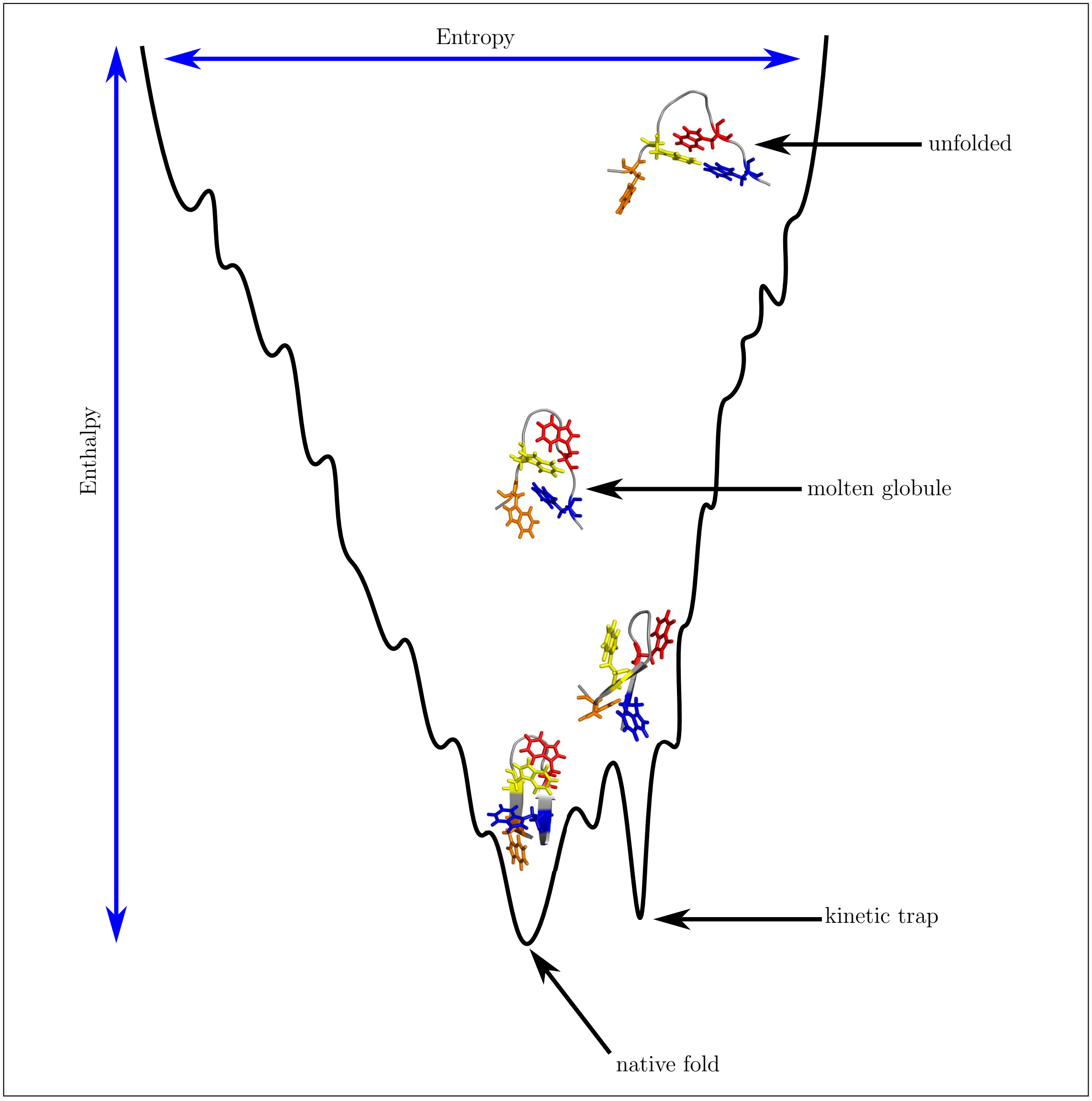


Figure 1.4: A schematic funnelled landscape for a model protein. The width of the landscape corresponds to entropy, and the height represents enthalpy (potential energy). The folded protein resides at the the bottom of the landscape. Folding proceeds from the unfolded protein, in the high energy regions of the landscape. Hydrophobic collapse leads to the formation of a molten globular (compact) state, which comprise some native structure. Occasionally the protein may encounter misfolded states en route to the native state. These states are generally separated from the rest of the landscape by high energy barriers, and so act as kinetic traps in the folding process.

In the denatured state, where the conformational entropy is high, the energy landscape is relatively flat. As the protein folds, formation of native contacts lead to a large thermodynamic barrier to unfolding. This free energy barrier limits the number of accessible conformations and generally guides the search downhill, in the direction of the free energy gradient. As a result, the protein never has to search the entire conformational space (as in Levinthal's Paradox); instead, the search is guided towards the native state. In this scenario the energy landscape is globally funnelled,[29,30] with the native state residing at the bottom (Figure 1.4).

Onuchic and colleagues[27] explained that a funnelled energy landscape is expected for naturally occurring proteins, due to evolution. In contrast, if one were to assemble a random polypeptide chain, efficient folding would be very unlikely. This is because random heteropolymers (RHPs) contain equally stabilising (local) and destabilising (non-local) contacts on average; the system is energetically frustrated.[25] However, in natural proteins strong interactions are primarily native contacts and very few interactions oppose folding; so there is minimal energetic frustration.[25,26]

Within the Energy Landscape description, the mechanism for folding depends on the underlying topology of the landscape. In general, the free energy landscape is not globally smooth; rather it supports many local minima, due to the interplay of entropic and enthalpic terms.[27] The dynamics between various states will depend on the barriers and the overall funnelled organisation. The barrier heights in various parts of the landscape dictate whether intermediates accumulate during folding.[28] Additionally, formation of non-native contacts (increased frustration) may lead to misfolded states, analogous to deep wells on the landscape, which act as kinetic traps, slowing down folding.[29,30] Due to the structural heterogeneity, it is expected that multiple pathways lead to the folded state, each becoming increasingly more distinct as the native state is approached and conformational entropy is minimised.

Since the amino acid sequence encodes the various interactions, from which the energy landscape emerges, variation in the precise folding events is expected from one sequence to the next. For example, a uniform attraction of hydrophobic residues may favour rapid collapse of the unfolded state into a compact globule (folding intermediate) that slowly rearranges into the native state. In another scenario, local interactions along the chain may facilitate transient secondary structure formation in the denatured state and eventual coalescence to give the native fold. Each process will have associated energetic and kinetic barriers, and the gradient of the energy will depend on the relative stability of the denatured and folded states. However, the underlying rules governing folding are the same, and the energy landscape model can be adopted to interpret the various folding scenarios, and to account for the ensuing dynamics.[31]

Ultimately, a global view of the energy landscape is the key ingredient for elucidating protein folding. Over the last few decades much effort has been expended in probing the underlying energy landscape of proteins, both experimentally and via computer simulations. Advances in these areas are discussed in the following sections.

1.3 Experimental Techniques

1.3.1 Protein folding initiation methods

Protein folding is usually initiated from the denatured state. This process involves perturbing the prevailing conditions to produce a non-equilibrium ensemble, which can then relax to a new equilibrium state. Rapid mixing-based methods, namely stopped-flow[32–36] and quenched-flow,[37–39] are historically the most common techniques employed to trigger protein folding. Generally, a denaturing agent (e.g. guanidine hydrochloride or urea) is first used to unfold the protein. The protein-denaturant solution is then diluted by rapidly mixing a buffer that favours folding. Once the solutions are mixed the folding process can be probed. Rapid-mixing techniques are appealing, since no chemical changes need to be made to the protein under investigation. However, these methods suffer from limited time resolution due to the inherent dead times of the mixing apparatus (on the order of milliseconds), which generally exceed the time scales of the fastest folding events (on the order of nanoseconds to microseconds; Figure 1.5). Alternative variations, such as the continuous-flow technique,[40–42] offer a slight improvement in the time resolution, with dead times in the microsecond regime. Accordingly, faster folding events, which occur in the early stages of folding, cannot be probed using these approaches. Nonetheless, rapid mixing-based methods have been instrumental in probing folding intermediates, particularly molten globules, as well as providing evidence in support of early secondary structure formation during folding.[33–36,38,39,41]

Laser-induced temperature jump from a cold-denatured state is the most widely used triggering method to study protein folding events on the sub-millisecond time scale.[43–53] Significant lowering of the temperature below physiological values leads to protein unfolding.[54–56] Hence, refolding may be initiated by a rapid jump in temperature from a cold-denatured state. Short laser pulses (picosecond/nanosecond) can be used to excite the infrared (IR) vibrational modes of water, which generally relax on the picosecond time scale, and produce a ultra-fast jump in temperature. The rapid temperature jump destabilises the denatured state and the protein subsequently refolds. Fast folding events, particularly formation of secondary structures, can then be probed. Studies employing laser-induced temperature jump techniques were among the first to provide time-resolved structural dynamics for the helix-coil transition[46] and β-hairpin formation.[45]

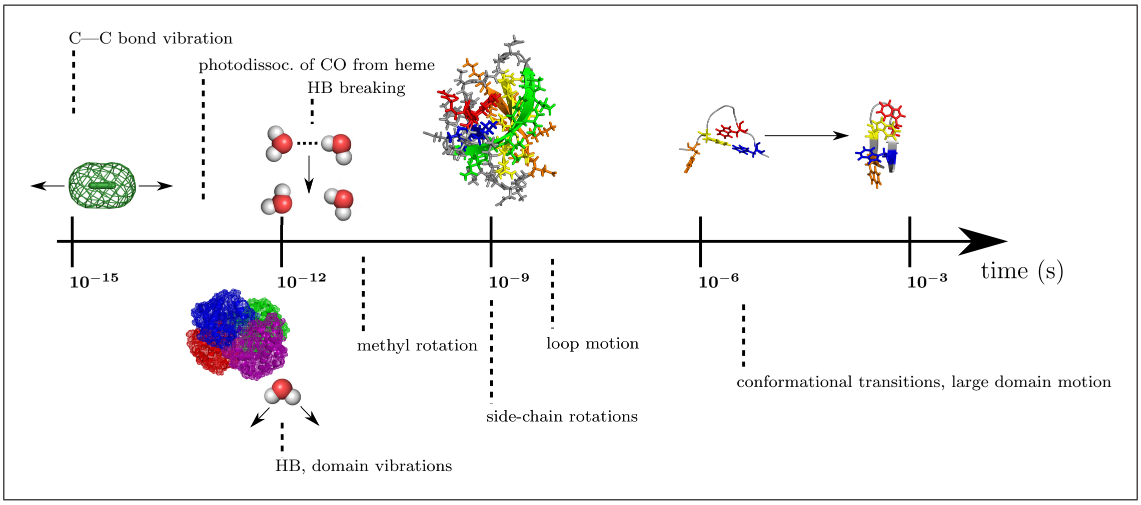


Figure 1.5: Hierarchy of time scales for protein motions.

Alternatively, rapid changes in pressure may be used to induce protein folding (or unfolding).[57–60] Generally, proteins denature under high pressures and may relax to the native state following a negative jump in pressure. Pressure perturbations can significantly alter the rate constant for folding; thus, reductions in the folding rate, via appropriate pressure jumps, can be employed to stabilise folding intermediates and characterise them.[61,62]

Another means of initiating protein folding is via rapid electron transfer. This method is particularly useful in studies involving redox-active proteins, such as cytochrome *c*.[63,64] For cytochrome *c*, the reduced state (Fe^{2+}) is more stable towards unfolding than the oxidised form (Fe^{3+}). Hence, rapid injection of electrons into the unfolded oxidised protein can be employed to trigger folding.

Photo-induced ligand dissociation may also be utilised to trigger protein folding. Carbon monoxide (CO) is known to bind to the haem group of proteins such as myoglobin and cytochrome *c* and lead to unfolding. Rapid photolysis of the CO ligand causes the proteins to refold.[65,66] Since the dissociation can occur on the sub-picosecond time scale, this technique is very useful in probing fast folding events. A more generally applicable approach is to engineer a photo-trigger into the protein, which can stabilise the unfolded state and then be photo-cleaved to initiate folding, irreversibly.[67,68] Conversely, reversible folding is achieved by using photo-switches that initiate folding (or unfolding) via photo-induced isomerisation.[69,70]

Finally, mechanical force can also be employed to control protein folding. Atomic force microscopy (AFM)[71,72] and optical tweezers[73,74] facilitate the single-molecule protein folding studies, which are generally not possible using other techniques. Accordingly, invaluable insight into the protein folding landscape can be attained using these methods.

1.3.2 Structural and kinetic probes

Pioneering work by Kendrew and colleagues in deciphering the structure of myoglobin via X-ray crystallography[7] ushered in a new era of protein discovery. Although the spacial resolution of the initial X-ray crystal structure was low (approximately 6 Å), the details provided suggested that an intricate connection existed between protein structure and function. Subsequent X-ray structures for haemoglobin,[8] lysozyme,[75] ribonuclease,[76,77] among other proteins, further elucidated this connection. To date, X-ray crystallography has been the most extensively used technique to determine protein structure;[78] accounting for over 80% of protein entries in the Protein Data Bank (PDB). The main challenge of this technique is in obtaining single crystals for X-ray diffraction. Specifically, membrane proteins, multi-domain proteins that consist of flexible linkers, and intrinsically disordered proteins (i.e. proteins that lack well-defined native structures) are often difficult to crystallise, prohibiting characterisation by traditional X-ray diffraction. Another drawback of this technique is that X-ray structures are at best static representations of proteins and do not explicitly capture the inherent structural heterogeneity. As an extension to this method, small angle X-ray scattering (SAXS), has been successfully applied to provide time-resolved structural data for proteins in solution, albeit at lower spacial resolutions.[79–83]

Nuclear magnetic resonance (NMR) spectroscopy is one of the leading technologies for probing protein folding.[84–87] In particular, NMR parameters such as chemical shifts, scalar coupling constants, residual dipolar couplings (RDCs) and nuclear Overhauser effects (NOEs) provide ensemble averages for protein structure and dynamics. Backbone chemical shifts can distinguish between α-helices and β-strands[88] and line broadening of NMR peaks (e.g. in 1D ^{1}H-NMR) can be used to monitor exchange rates between ensembles.[89] The extent of line broadening depends on the ratio of the difference in the resonance frequencies of native and denatured spins and the rate of exchange between the two states. For large ratios, peaks are not averaged and appear as separate lines, while small ratios lead to complete averaging and one line is observed in the NMR spectrum. In the intermediate regime, significant line broadening occurs due to incomplete averaging. Line shapes can then be fitted to deduce the exchange rate constants (with μs to s accessible time scales).[89] Relaxation dispersion NMR techniques,[90] which employ multidimensional NMR,[91] modulate exchange effects, thus altering the sharpness of NMR peaks. These procedures are commonly used to probe protein dynamics in the μs to ms regime.[86]

Scalar coupling constants and ^{1}H-^{1}H cross peaks (COSY and NOESY; correlation and NOE spectroscopy, respectively) provide local distance constraints (for bonded atoms or nuclei within 5 Å) and are particularly useful for structural characterisation of the folded state.[85,92] NMR spectrum for partially folded states suffer from poor dispersion of ^{1}H and ^{13}C resonances (peaks overlap severely), NOEs weaken; therefore, fewer NMR restraints are available for structural characterisation.[93] Two-dimensional ^{1}H-^{15}N NMR correlation spectroscopy techniques provide reasonably dispersed spectra for these states, and ^{15}N relaxation data can be used to quantify backbone motion.[93] Amide hydrogen-exchange techniques coupled with NMR are particularly useful for probing partially folded states.[94] Exchange rates are inferred from changes in peak intensities and the fraction of time for which protons were protected from exchange (by participating in hydrogen-bonding) can be determined. These techniques are often combined with rapid initiation methods (discussed in § 1.3.1) to probe folding intermediates and mechanisms.[37–39,60,95–97]

RDCs are useful for the determination of backbone conformations (helical axis/backbone curvature) or relative orientations of multi-domains in larger proteins.[98,99] This parameter quantifies the relative alignment of internuclear bonds with the external magnetic field and, therefore, provides information on the relative orientation of the residues, irrespective of spacial distance. RDCs have also been used to probe the structure and dynamics of unfolded[100] and intrinsically disordered[101] proteins.

Another useful technique for probing both equilibrium and kinetic aspects of protein folding (and unfolding) in solution is circular dichroism (CD).[102–106] When chiral molecules, such as proteins, interact with circularly polarised light, the differential absorption of the left and right light components at various wavelengths can be used to decipher structural propensities. Absorption (usually by peptide bonds) in far-UV CD studies (180 nm to 240 nm) is used to estimate secondary structure content; whereas absorption (by aromatic side-chains or disulphide bonds) in near-UV CD experiments (250 nm to 290 nm) gives information about tertiary structure.[103,104] Changes in CD spectra can be used to monitor structural changes during folding and unfolding, and therefore provide key insight into folding mechanisms and stability, respectively.[107] CD techniques were among the first used to detect folding intermediates or partially folded proteins (molten globule state).[108] In many cases, the intermediates exhibited large CD signatures in the far-UV regime, indicative of the presence of some secondary structure, and negligible CD signals in the near-UV region of the spectrum. This technique has also been employed to quantify structural changes in related proteins (e.g. mutants),[109,110] and follow conformational transitions in proteins that are prone to misfolding and aggregation,[111] such as amylin[112–114] and β-amyloid,[115–117] which are implicated in diabetes and Alzheimer's disease, respectively. Changes in absorption (or more specifically, ellipticity) in CD spectra can be directly correlated with changes in equilibrium concentrations, which can be used to compute thermodynamic parameters; while a CD spectrometer can be attached to a suitable rapid initiation apparatus to monitor dynamics.

Infrared (IR) spectroscopy is another well-established technique for investigating protein structure and dynamics.[118,119] In particular, Fourier transform infrared (FTIR) spectroscopy has been extensively used to obtain time-resolved data for protein folding, with moderate effort and at high temporal resolutions ($< 1\,\mu$s).[120–122] Since proteins generally absorb IR radiation throughout their structure, elaborate labelling procedures are not required. However, this property can also hinder structural characterisation, due to the existence of numerous overlapping absorption bands. In general, the IR spectrum of proteins contain nine characteristic bands.[119] The amide I band, which is largely attributed to the C–O stretch, is one of the most prominent, occurring around 1700–1600 cm^{-1}. The amide I band is highly sensitive to hydrogen-bonding patterns and small changes in backbone geometry. Hence, the specific position of this band is strongly correlated with secondary structure; making it a good diagnostic when coupled with empirical fitting techniques. The relative intensities of absorption bands in the amide I band region can also be used to estimate overall secondary structure composition. While complete structure prediction may be a challenge, techniques such as difference IR spectroscopy are very useful in detecting conformational changes, and are often employed to probe reaction sites (e.g. in enzyme-substrate interactions), folding intermediates and protein flexibility in general.[118,119] Like NMR-based studies, band broadening effects provide information on structural flexibility; with more flexible structures giving rise to broader absorption bands.[119] Two-dimensional infrared spectroscopy (2D IR) offers significant improvements for probing protein conformational dynamics on the sub-picosecond time scale, since vibrational modes are extremely sensitive.[123] Along with FTIR,[124] this technique has been instrumental in probing protein misfolding and aggregation; providing key insight into the evolution of various neurological disorders.[125] As a complementary technique to IR spectroscopy, Raman spectroscopy also provides highly sensitive signatures for secondary and tertiary structure of proteins.[126,127]

Fluorescence is the last major experimental probe presented in this section. Broadly, fluorescence techniques for protein folding either exploit native or non-native fluorescence.[128,129] In native (intrinsic) fluorescent experiments, residues, such as tryptophan, tyrosine and phenylalaline, that emit fluorescence are used as the probes.[130,131] In non-native (extrinsic) fluorescence investigations fluorescent dyes (e.g. ANS) are used to follow the folding process.[132] In general, the quantum yields of fluorescent probes are extremely sensitive to the local environment (e.g. degree of solvent exposure) and mobility (e.g. movement of side-chains). Additionally, many probes emit fluorescence on the nanosecond time scale; therefore, high time resolution kinetic data can be obtained. These features, along with high signal-to-noise ratios, make fluorescence-based approaches very appealing for monitoring protein conformational dynamics.

Typically, fluorescence signals of buried residues are blue-shifted; whereas, the signals are red-shifted when the residues are exposed to solvent.[128,129] This property has been used in folding/unfolding studies to detect various equilibrium or intermediate states. Fluorescent dyes may bind transiently at different stages of the folding process, which effectively modulates signal intensities. In classical fluorescent studies, detailed structural characterisation is inhibited due to the inherent local nature of signals. Hence, these studies are most meaningful when performed in conjunction with other techniques, such as FTIR, CD and NMR. Alternative fluorescence approaches, namely Förster resonance energy transfer (FRET)[133] and fluorescence correlation spectroscopy (FCS),[134] provide more direct spacial information. In FRET studies, coupling effects between donor and acceptor fluorescent pairs are used to characterise protein folding events. Generally, energy transfer between FRET pairs has a $1/r^6$ dependence (where r is the distance between the probes); thus carefully chosen pairs (e.g. IAEDANS and IAF) can be used to probe local structure and degree of compaction. FRET techniques have also been instrumental in single-molecule protein folding studies; where the structural heterogeneity along folding/unfolding pathways can be explored.[135–137] Finally, in FCS, fluctuations in signal intensities can be correlated with conformational dynamics and used to compute relaxation constants.

1.4 Computer Simulations

Protein folding simulations can probe atomistic details of the folding process not amenable to most experimental techniques; in particular, high spacial (distance) and temporal (fastest motions) resolutions are achievable. The accuracy of protein simulations depends largely on the form and parameters of the energy functions used to represent the protein and the surrounding environment, and the methods used to sample the conformational space. These factors are intrinsically linked to the available computing hardware and, unlike experiments, generally limit the length scales and duration of computer investigations. In this section, an overview of protein simulation techniques is presented: from classical techniques to more enhanced methods for exploring the high-dimensional folding space.

1.4.1 Classical protein simulation techniques

Conventional protein simulations focus primarily on refining experimental structures, where model coordinates are derived mainly from X-ray crystallography studies.[138–140] The earliest study employed coordinate fitting procedures (model building) to refine atomic coordinates.[138] Subsequent refinement procedures sought to minimise the potential energy of the system with respect to the Cartesian coordinates (energy refinement). In one treatment, Levitt and Lifson[139] defined the potential energy as a function of the bond lengths, bond angles, dihedral angles, and non-bonded pairs, along with a constraint term, which ensured that the deviations of the atomic coordinates from the experimental ones were kept to a minimum. Equilibrium bond lengths and angles were obtained from X-ray structures of small molecules, and torsional parameters were taken from the Ramachandran plot. These types of coordinate refinement procedures were used to optimise the geometries of single-domain globular proteins[139–141] and to compute the conformational preferences of side-chains.[142,143] Alternative energy minimisation procedures perturbed internal coordinates (dihedral angles) to search for lower energy structures.[144] The researchers found that the potential energy of the lowest energy structures was about 40 kcal/mol lower than the highest energy ones.[144]

These initial studies shed light on the complexity of the protein conformational space and revealed that, even in the vicinity of the native state, proteins exhibited significant conformational heterogeneity (‘multiple-minima problem’).[145] To achieve better conformational sampling, Levitt and Warshel introduced a simplified representation for proteins.[146] In their model, each residue in the protein was represented by the C_α atom and the centroid of the side-chain. Representations such as this served to reduce the number of degrees of freedom, so a larger region of conformational space could be explored. The key assumption was that a separation of time scales of protein motions (short-range verses long-range) existed, which permitted a time-averaging of the short-range motion to be adopted without significantly altering the main features of the folding process. They proposed that in the early stages of folding long-range forces played a central role in restricting the conformational space by directing protein collapse.[146]

Monte Carlo (MC) methods were later developed in an attempt to overcome the multiple-minima problem, and have been some of the most widely used approaches for protein structure prediction.[147] Though structure prediction techniques do not probe protein folding directly, they can provide valuable information about the topography of the folding space. In a typical MC protein simulation, the energy of the starting structure is calculated, followed by a random perturbation of the coordinates to give a new configuration. A Metropolis condition[148] is often introduced next, which serves to bias the search towards the low-lying regions of the energy landscape. Additionally, in classical MC only small step sizes in configurational space are allowed, otherwise all steps would be rejected. This technical aspect makes convergence of thermodynamic parameters for complex systems, such as proteins, nearly impossible.

Exploration of conformational space with MC methods has also been used extensively in conjunction with lattice models.[149–151] Although overly simplified, lattice models of proteins were invaluable in shaping the energy landscape view of protein folding (§ 1.2.4). In 1994, Hao and Scheraga[152] suggested that the folding (or unfolding) of small proteins may involve a first-order transition; wherein at the transition temperature the prominent structures correspond to the native state and the unfolded protein, with a negligible population of partially folded states. That same year, Socci and Onuchic[153] used lattice models to classify polypeptides as either good- or non-folding sequences. They demonstrated that good folders are ones in which the folding temperature exceeds the glass transition temperate[§] ($T_f > T_g$), whereas non-folding sequences become kinetically trapped before complete folding is achieved ($T_f < T_g$; native state is kinetically inaccessible).

Molecular dynamics (MD)[154] is undeniably the most extensively used technique to simulate proteins.[155–157] In classical MD, the dynamics are simulated by solving Newton's equations of motion numerically.[158] Initial values are assigned for the position (usually experimentally derived atomic coordinates) and the velocity (often randomised) of the protein, and the potential energy is modelled by an empirical potential. At fixed time intervals (time steps) the values are updated, until the final simulation time is reached. Unlike MC methods, MD simulations are inherently time-dependant. Furthermore, the time step is usually constrained by the fastest motion in the system and so fully atomistic simulations, of even small proteins, require significant computation time, as for standard MC.

Hence, due to limited computing power, classical MD techniques [158–160] are often

[§] The glass transition temperature, T_g, was defined as the temperature at which the relaxation time of the system exceeds the observation time.

only able to probe protein motion in the neighbourhood of the native state, and are largely employed to refine X-ray/NMR structures. In 1981, Northrup *et al.* reported that B factors (a measure of the spread of electron density of atoms) measured in X-ray crystallography studies showed good agreement with the mean-square fluctuations of atoms in MD simulations.[160] The stability of hydrogen-bonds was also directly related to proton exchange rates in NMR studies, and the role of hydrogen-bonding in the folded state was probed at high spacial resolutions.[159] Additionally, to achieve greater computing efficiency, NOE distances were used as constraints for structure prediction from extended states.[161]

1.4.2 Enhanced sampling of protein conformational space

The barriers associated with folding are generally high[¶] and so in a typical simulation the frequency at which the system acquires sufficient energy to overcome those barriers will be low. Hence, folding can be regarded as a "rare event", and it is common for classical simulations to become trapped in a local region of configuration space. Therefore, to achieve improved sampling of the conformational space numerous MC/MD based algorithms and protocols have been proposed over the past few decades.

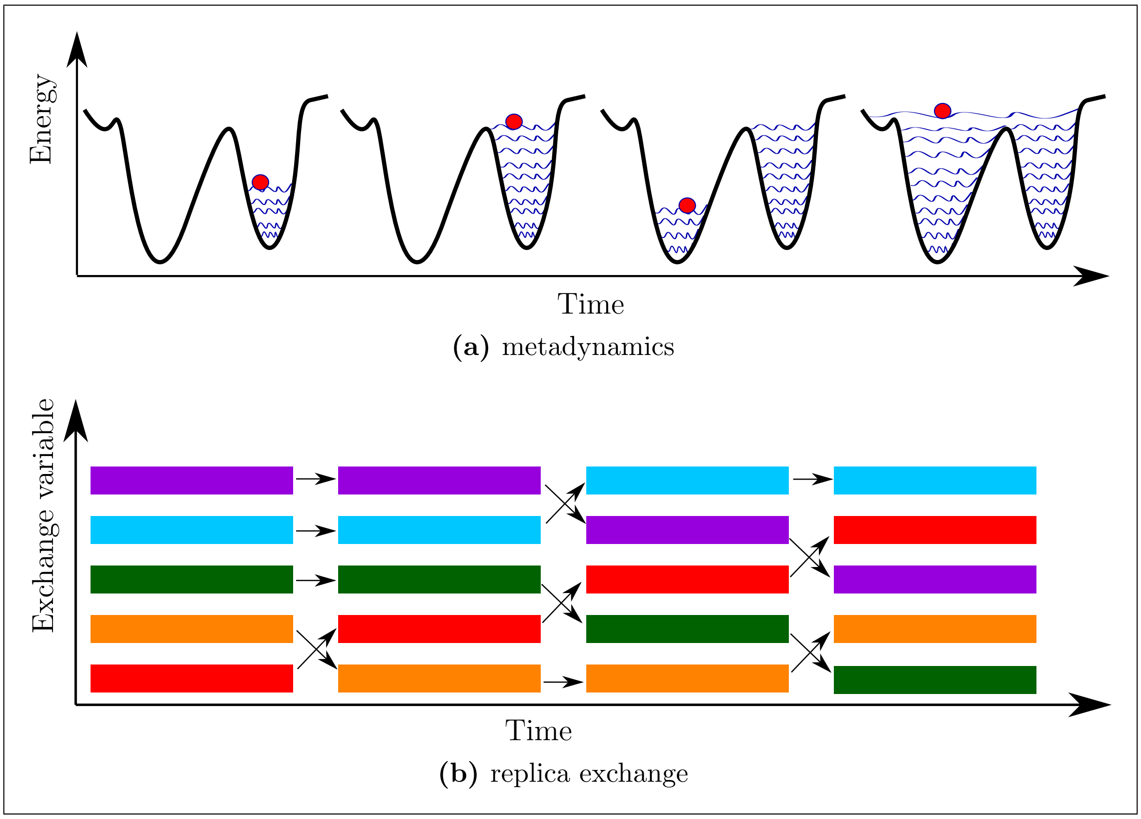
Elevated temperatures were used to drive unfolding and probe the corresponding structure and dynamics; for example, important intermediates in the helix-coil transition were characterised via high temperature MD. In similar studies, pressure variations, low pH and denaturants were used to unfold proteins and to study the effects of solvents on protein denaturation.[162,163] Such investigations were not possible via experiment, and MD simulations were able to probe, in atomistic detail, the disruption of hydrogen-bonds in proteins by water during denaturation.[164] However, some researchers were sceptical about the validity of these simulations; specifically, there were queries on the extent to which the principle of microscopic reversibility would hold under non-equilibrium conditions and on how much information about the reverse process (folding) could be inferred from unfolding studies.[1] It was suggested that, while the fine details may vary, unfolding at moderately elevated temperatures proceeded via similar intermediates (in reverse order) to folding under native conditions.[1]

In the 1980s protein simulations that employed "umbrella" potentials began to emerge.[165,166] In umbrella sampling, a reaction coordinate (or collective variable) that partitions the configuration space is selected, and a biasing potential (based on the collective variable) is added to the 'true' potential energy function, to direct

[¶] Barriers are high in comparison to k_BT; the energy of random thermal fluctuations.

sampling in specific regions. An early study exploited this technique to probe slow conformational changes in proteins, specifically ring flipping.[165] Free energy surfaces based on the biased simulations (potential of mean force) can be derived using the weighted histogram analysis method (WHAM).[167] During the 1990s Brooks and colleagues[168–172] published several papers on an analogous biased-sampling technique, which they used to study small peptides in solution. They reported that the surrounding solvent weakened exposed hydrogen-bonds in helices and loops, whereas hydrogen-bonds in β-sheets were relatively stronger, about 5 kT. Other related approaches, which probe the free energy via biased potentials, include metadynamics[173–175] (Figure 1.6a), steered MD,[176,177] and targeted MD.[178] The latter two methods may be useful for studying protein-ligand binding (and unbinding) and the effects of mechanical force on protein structure, similar to AFM experiments. Alternatively, energy ranges may be used to restrict the sampling region, as in the multicanonical[179,180] and Wang-Landau[181] approaches.

Methods that exploit *a priori* or time-dependent reaction coordinates to evolve the system may introduce systematic errors in protein simulations, as it is generally exceedingly difficult to define these coordinates. One must, therefore, be cautious when interpreting the results of such studies. Instead of biasing the sampling with a reaction coordinate, the phase space may be partitioned using some order parameter (which characterises different states), and unbiased simulations can be performed between states. This idea is at the heart of methods such as milestoning,[182,183] transition path sampling (TPS)[184–186] and its derivative, transition interface sampling (TIS).[187] These methods are capable of yielding accurate folding times and mechanisms; however, they can be computationally costly and so are generally limited to study smaller proteins. Bolhuis[188] applied TPS to investigate the order of folding events for the β-hairpin of protein G (16 residues) and found that, consistent with experiment, the peptide folded by hydrophobic collapse, followed by the formation of secondary contacts. The unfolding time, probed by TIS, was also in reasonable agreement with that found in fluorescence studies.[188]



(a) metadynamics

(b) replica exchange

Figure 1.6: Two common enhanced sampling techniques used to probe protein folding. (a) Metadynamics: commonly described as 'filling the energy landscape with sand'. The system evolves based on collective variables—used to construct a Gaussian, which is added to the true potential. This formulation discourages resampling of previously explored regions. (b) Replica exchange: multiple copies of the system are simulated simultaneously (e.g. at different temperatures). At regular intervals, exchanges between adjacent replicas are attempted.

Replica exchange methods (REMs), specifically parallel tempering (PT) approaches, are perhaps the most commonly used techniques for achieving enhanced sampling of proteins.[189,190] In PT, multiple independent copies of the system are simulated at different temperatures via MC[189] or MD sampling.[190] Exchanges between adjacent replicas are attempted after some fixed interval (MC steps or MD time steps), with exchange probabilities usually based on the Metropolis criterion (Figure 1.6b). Alternatively, the temperatures of adjacent copies may be swapped. This protocol facilitates the escape of the system from low-lying regions of phase space, and, in theory, all barriers can be surmounted via exchange with higher energy replicas. Replica exchange MD (REMD),[190] in particular, has been applied to study a wide range of protein systems, including: small peptides in explicit/implicit solvent,[191–193] amyloid-forming proteins[194,195] and chaperones.[196] REMs also take advantage of parallel computing and have even been tailored to proceed on distributed computing, as in multiplexed-REMD.[197] The main drawback of REMs is that these procedures lead to reduced kinetic data, due to the inherent exchange of trajectories.

In recent years, construction of kinetic transition networks (KTNs) from MD or MC simulations has been attempted and is capable, in principle, of preserving the observable features of protein folding. Instead of running one long simulation, the goal here is to build statistically robust models from numerous independent simulations. One way of analysing data from MD simulations is by building Markov state models (MSMs).[198–201] MD trajectories are first decomposed into states, and a transition probability matrix is then built for the coarse-grained system. The transition matrix is derived based on the observed transitions between the states, where the transitions are assumed to be history-independent (Markovian). Hence, the resulting MSM encodes both the thermodynamics and dynamics of the folding process.[202–205] This approach capitalises on distributed computing platforms, such as Folding@home pioneered by Shirts and Pande.[206] Since MSM analyses do not require that simulations are performed as a single long trajectory, greater simulation duration can be obtained from simulations performed in parallel.

MSMs have suggested a hub-like character of the native state, wherein the folded protein is generally accessible via multiple paths from the heterogeneous unfolded state.[207] Importantly, it was shown that non-native states play a crucial role in slowing down the folding rate, as previously reported in experimental studies.[207] Concurrently, Voelz *et al.*[208] presented folding pathways for a millisecond-folder modelled atomistically with implicit solvent. Notably, the estimated folding time of 1 ms was in good agreement with the experimental one, 1.5 ms. Later, folding simulations of an 80-residue protein was conducted at the atomic level with explicit solvent representation.[209] This study further highlighted the power of the MSM technique in integrating large amounts of MD data, to yield thermodynamic and dynamic insights, on time and length scales much longer than the conventional nanosecond–millisecond simulation limit.

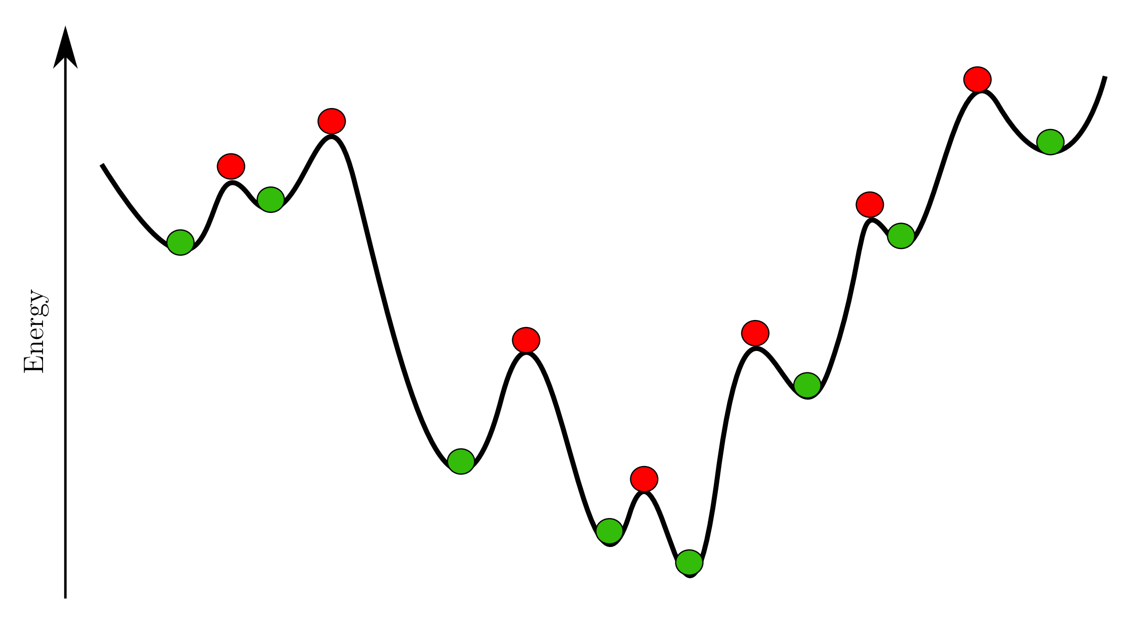


Figure 1.7: The potential energy landscape approach. Using geometry optimisation techniques, the PEL is discretised into stationary points: minima (green circles) and the transition states that connect them (red circles).

1.5 Thesis Overview

In this thesis, kinetic transition networks (KTNs) are constructed for protein folding, via approaches based on the potential energy landscape (PEL).[210–212] By applying geometry optimisation techniques, the PEL is discretised into stationary points (i.e. minima and the transition states that connect them; Figure 1.7). Essentially, minima characterise the low-lying regions of the PEL (thermodynamics) and transition states encode the motion between these regions (dynamics). Principles from statistical mechanics and transition state theory (unimolecular rate theory) may be conveniently employed to derive free energy landscapes and folding rates, respectively, from the resulting potential energy transition network.[212] Furthermore, the PEL framework can take advantage of parallel and distributed computing, since stationary points from separate simulations can be easily integrated into one KTN. Moreover, the use of geometry optimisation facilitates greater conformational sampling than conventional techniques. Accordingly, this framework presents an appealing means of probing complex processes, such as protein folding.

The remaining parts of this thesis are organised as follows:

- In § 2 the main theories and techniques employed to construct potential energy landscapes are summarised.

- § 3 explores the intrinsic rigidity of proteins and describes how a local rigid body (LRB) approach may be adopted within the PEL framework to probe protein folding. The LRB approach effectively integrates out irrelevant degrees of freedom from the geometry optimisation procedure and further accelerates conformational sampling. The effects of this approach on the underlying PEL are analysed in a systematic fashion, for a model protein (tryptophan zipper 1).[213]

- § 4 is concerned with fold-switching in proteins. Such large-scale conformational changes are generally difficult to probe computationally. Methods based on geometry optimisation have proved useful in overcoming the broken ergodicity issue, which is associated with proteins that undergo large-scale conformational changes. In this chapter, the latest PEL-based approaches are utilised to investigate the most extreme case of fold-switching found in the literature: the α-helical hairpin to β-barrel transition of RfaH, a bacterial regulatory protein.[214,215]

- § 5 focusses on modelling intrinsically disordered proteins (IDPs). Due to their inherent structural plasticity, IDPs are generally difficult to characterise, both experimentally and via simulations. An updated approach for studying IDPs within the PEL framework is presented and tested with various contemporary potential energy functions. The cytoplasmic tail of human cluster of differentiation 4 (CD4),[216] implicated in HIV-1 infection, is investigated in this work.

- Finally, § 6 highlights the key findings of the dissertation and future research directions are proposed.

2

Methods

2.1 The Potential Energy Surface

The time-independent, non-relativistic Schrödinger equation for a system of N nuclei and n electrons is given by:[217]

$$\hat{H}\psi(\mathbf{x},\mathbf{X}) = E\psi(\mathbf{x},\mathbf{X}), \tag{2.1}$$

where $\hat{H}$ is the molecular Hamiltonian, $\psi(\mathbf{x},\mathbf{X})$ is the molecular wave function in terms of the set of electronic and nuclear coordinates (i.e. $\mathbf{x}$ and $\mathbf{X}$, respectively), and E is the total energy. Expressing the Hamiltonian in terms of the various kinetic and potential energy operators gives:

$$\begin{aligned} \hat{H} &= \hat{T} + \hat{V} \\ &= \hat{T}_N + \hat{T}_n + \hat{V}_{NN} + \hat{V}_{Nn} + \hat{V}_{nn}, \end{aligned} \tag{2.2}$$

corresponding to the nuclear kinetic energy and electronic kinetic energy operators, and the three pair-wise interaction energy operators. Explicitly, $\hat{H}$ has the form:

$$\begin{aligned} \hat{H} = &-\sum_{I}^{N}\frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_{i}^{n}\frac{\hbar^2}{2m_i}\nabla_i^2 \\ &+\sum_{I}^{N}\sum_{J>I}^{N}\frac{Z_I Z_J e^2}{|\mathbf{X}_J - \mathbf{X}_I|4\pi\varepsilon_0} - \sum_{I}^{N}\sum_{i}^{n}\frac{Z_I e^2}{|\mathbf{x}_i - \mathbf{X}_I|4\pi\varepsilon_0} \\ &+\sum_{i}^{n}\sum_{j>i}^{n}\frac{e^2}{|\mathbf{x}_j - \mathbf{x}_i|4\pi\varepsilon_0}. \end{aligned} \tag{2.3}$$

In equation (2.3), M_I and m_i represent the mass of nucleus I (with charge Z_I) and electron i (with charge e), respectively. Since the mass of a proton is approximately 1836 times greater than the mass of an electron, $M_I >> m_i$, Born and Oppenheimer rationalised that classically the nuclei are fixed.[218] Accordingly, for fixed nuclear coordinates the wave function becomes:

$$\psi(\mathbf{x}, \mathbf{X}) = \psi_{elec}(\mathbf{x}; \mathbf{X})\psi_{nuc}(\mathbf{X}), \tag{2.4}$$

where $\psi_{elec}(\mathbf{x}; \mathbf{X})$ represents the electronic wave function (which is evaluated at fixed nuclear positions), and $\psi_{nuc}(\mathbf{X})$ corresponds to the nuclear wave function. Substituting equations (2.4) and (2.2) in (2.1), and expanding for $\hat{T}_N$ gives:

$$\begin{aligned} \hat{H}\psi_{elec}(\mathbf{x}; \mathbf{X})\psi_{nuc}(\mathbf{X}) = &- \sum_{I}^{N} \frac{\hbar^2}{2M_I}\psi_{elec}(\mathbf{x}; \mathbf{X})\nabla_I^2\psi_{nuc}(\mathbf{X}) \\ &- \sum_{I}^{N} \frac{\hbar^2}{2M_I}[2\nabla_I\psi_{elec}(\mathbf{x}; \mathbf{X})\nabla_I\psi_{nuc}(\mathbf{X}) + \psi_{nuc}(\mathbf{X})\nabla_I^2\psi_{elec}(\mathbf{x}; \mathbf{X})] \\ &+ [\hat{T}_n + \hat{V}_{NN} + \hat{V}_{Nn} + \hat{V}_{nn}]\psi_{elec}(\mathbf{x}; \mathbf{X})\psi_{nuc}(\mathbf{X}) \end{aligned} \tag{2.5}$$

Since the electronic wave function only depends parametrically on $\mathbf{X}$, the derivatives of $\psi_{elec}(\mathbf{x}; \mathbf{X})$ w.r.t nuclear coordinates can be ignored. Additionally, $\hat{T}_n$ has no $\mathbf{X}$ dependence. These approximations allow us to simplify equation (2.5) as:

$$\begin{aligned} \hat{H}\psi_{elec}(\mathbf{x}; \mathbf{X})\psi_{nuc}(\mathbf{X}) &= \psi_{elec}(\mathbf{x}; \mathbf{X})(\hat{T}_N\psi_{nuc}(\mathbf{X})) + \psi_{nuc}(\mathbf{X})(\hat{T}_n\psi_{elec}(\mathbf{x}; \mathbf{X})) \\ &+ \psi_{elec}(\mathbf{x}; \mathbf{X})(\hat{V}_{NN}\psi_{nuc}(\mathbf{X})) + \psi_{nuc}(\mathbf{X})(\hat{V}_{Nn}\psi_{elec}(\mathbf{x}; \mathbf{X})) \\ &+ \psi_{nuc}(\mathbf{X})(\hat{V}_{nn}\psi_{elec}(\mathbf{x}; \mathbf{X})) \\ &= E\psi_{elec}(\mathbf{x}; \mathbf{X})\psi_{nuc}(\mathbf{X}) \end{aligned} \tag{2.6}$$

We can therefore define the electronic Schrödinger equation as:

$$\begin{aligned} \hat{H}_{elec}\psi_{elec}(\mathbf{x}; \mathbf{X}) &= \hat{T}_n\psi_{elec}(\mathbf{x}; \mathbf{X}) + \hat{V}_{Nn}\psi_{elec}(\mathbf{x}; \mathbf{X}) \\ &+ \hat{V}_{nn}\psi_{elec}(\mathbf{x}; \mathbf{X}) \\ &= E_{elec}(\mathbf{X})\psi_{elec}(\mathbf{x}; \mathbf{X}). \end{aligned} \tag{2.7}$$

In equation (2.7) the electronic energy $E_{elec}(\mathbf{X})$ is a function of the nuclear positions, i.e. the set of electronic energies (of different electronic states) for fixed nuclear configurations. Hence, for a range of nuclear configurations, $E_{elec}(\mathbf{X})$ maps out an electronic potential energy surface; nuclear motion is studied on a surface provided by the electronic Schrödinger equation.

Putting equations (2.6) and (2.7) together gives:

$$\begin{aligned}\hat{H}\psi_{elec}(\mathbf{x};\mathbf{X})\psi_{nuc}(\mathbf{X}) &= (\hat{H}_{elec}\psi_{elec}(\mathbf{x};\mathbf{X}))\psi_{nuc}(\mathbf{X}) + \psi_{elec}(\mathbf{x};\mathbf{X})(\hat{T}_N\psi_{nuc}(\mathbf{X})) \\ &\quad + \psi_{elec}(\mathbf{x};\mathbf{X})(\hat{V}_{NN}\psi_{nuc}(\mathbf{X})) \\ &= E\psi_{elec}(\mathbf{x};\mathbf{X})\psi_{nuc}(\mathbf{X}). \qquad (2.8)\end{aligned}$$

It follows that:

$$\begin{aligned}\hat{H}\psi_{nuc}(\mathbf{X}) &= E_{elec}(\mathbf{X})\psi_{nuc}(\mathbf{X}) + \hat{T}_N\psi_{nuc}(\mathbf{X}) + \hat{V}_{NN}\psi_{nuc}(\mathbf{X}) \\ &= [\hat{T}_N + \hat{V}_{NN} + E_{elec}(\mathbf{X})]\psi_{nuc}(\mathbf{X}) \\ &= [\hat{T}_N + V(\mathbf{X})]\psi_{nuc}(\mathbf{X}) \\ &= E\psi_{nuc}(\mathbf{X}), \qquad (2.9)\end{aligned}$$

where $V(\mathbf{X})$ is the effective potential that describes the variation of the electronic energy as a function of nuclear coordinates, i.e. the potential energy surface (PES). In this work, PES was computed for systems in their electronic ground state. For complex systems such as proteins, $V(\mathbf{X})$ is a function of many internal coordinates; an N atom protein, in general, requires $3N-6$ vibrational coordinates.[*] Hence, the PES, in terms of these coordinates, is a multidimensional $(3N-6+1)$ hypersurface.[210]

The extrema (stationary points; $\nabla V(\mathbf{X}) = 0$) of this multidimensional space are generally the main points of interest; specifically, minima (stationary points with all their non-zero Hessian eigenvalues positive) and transition states (stationary points with a single negative Hessian eigenvalue). By definition, minima lie at the bottom of wells on the hypersurface, and transition states lie at the well boundaries, encoding the structure and dynamics of the system, respectively. The remainder of this chapter is mainly concerned with tools and approaches employed to probe the discretised PES.

2.2 Potential Energy Functions

In the previous section we showed how the concept of a potential energy surface develops from the Born-Oppenheimer approximation.[218] To construct a PES, we must first define an appropriate functional form for $V(\mathbf{X})$ (i.e. a potential energy function). Typically, potential energy functions are defined as a sum of independent

[*] In the absence of external forces, the PES is invariant to translation and rotation of the entire molecule.

or coupled analytical equations that describe the molecular forces. For proteins, energy functions are generally expressed in terms of nuclear coordinates, and the physical constants (i.e. parameters) are obtained from *ab initio* computations or experimental techniques such as spectroscopy.

A given functional form and parameter set together comprise a force field. Common force fields used in computer simulations of proteins include: AMBER (Assisted Model Building with Energy Refinement),[219–221] CHARMM (Chemistry at Harvard Macromolecular Mechanics),[222,223] OPLS (Optimised Potentials for Liquid Simulations),[224] and GROMOS (GROningen MOlecular Simulation).[225]

In this work, AMBER force fields were used to simulate protein structure and dynamics. Weiner and Kollman[219] first presented a description of the AMBER potential, wherein all atoms are represented explicitly. The full energy function is given by:

$$\begin{aligned} V_{\text{AMBER}} = \sum_{\textit{bonds}} k_r(r-r_{eq})^2 \quad &+ \sum_{\textit{angles}} k_\theta(\theta-\theta_{eq})^2 + \sum_{\textit{dihedrals}} \frac{V_n}{2}[1+\cos(n\theta-\gamma)] \\ &+ \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}\left|\frac{q_i q_j}{\varepsilon r_{ij}}\right|. \end{aligned} \tag{2.10}$$

The first two terms in equation (2.10) represent the bond and angle contributions to the potential energy, respectively. These terms are harmonic potentials centred on equilibrium values (i.e. r_{eq} and θ_{eq}); accordingly, the AMBER potential does not allow for bond breaking. The third term in equation (2.10) encompasses the torsional strain in the molecule, and the dihedral parameters (such as rotational energy barriers, V_n) are derived from empirical fitting procedures. The final two terms in the AMBER potential are the non-bonded terms—van der Waals and electrostatic energies, respectively. The van der Waals term includes repulsion (the $1/R_{ij}^{12}$ part) and dispersion attractions (the $1/R_{ij}^{6}$ part). The electrostatic energy is represented by a Coulomb potential, which sums over pairwise atomic charges; atomic charges are fixed, and so the potential does not account for polarisation effects explicitly.

All-atom AMBER force fields are advantageous since they allow for the computation of ^{13}C and ^{1}H-NMR properties. Moreover, steric effects due to hydrogen atoms and hydrogen-bonding, which play crucial roles in directing protein structure, can be accounted for in simulations. In this work, several modern all-atom AMBER force fields were used namely: ff99SB,[226] ff99SB-ILDN,[227] ff14ipq,[228] and ff14SB.[229]

2.3 Solvent Models

The structure and dynamics of proteins are significantly influenced by their solvent environment, and to be physically meaningful *in silico* experiments must account for solvent effects. Explicit representation of solvent significantly increases the degrees of freedom, leading to longer computer simulation times; thus modelling fully solvated proteins on biologically relevant time scales has been an ongoing challenge. Accordingly, implicit solvent models, such as generalised Born (GB) solvation methods, have been developed to address this issue.

In GB models the free energy of solvation is summarised as:[230]

$$\Delta G_{sol} = \Delta G_{cav} + \Delta G_{vdW} + \Delta G_{pol}, \tag{2.11}$$

where ΔG_{cav} is the cavity free energy, which accounts for the disruption of solvent-solvent interactions to accommodate solute molecules, and ΔG_{vdW} represents the favorable van der Waals interactions between solute and solvent. These non-polar terms vary linearly with the solvent-accessible surface area (S) and can be evaluated together as:

$$\Delta G_{cav} + \Delta G_{vdW} = \sum_{k=1}^{N} \sigma_k \mathrm{S}_k; \tag{2.12}$$

σ_k is the atomic solvation coefficient, which is derived empirically.[231,232] The last term in equation (2.11) refers to the electrostatic free energy, and is a measure of the dielectric response of the solvent to the solute charge distribution. Different solvent models are generally distinguished by the method used to assess the polar component of solvation free energy. A numerical solution can be obtained for ΔG_{pol} by employing the Poisson-Boltzmann (PB) equation, where the electrostatic potential is modelled by the dielectric function and the charge distribution of the molecule.[233] This approach requires that the potential be evaluated for every structural change in the molecule, rendering such calculations computationally costly for most biological systems of interest.

Unlike the PB method, which employs an iterative procedure, the GB treatment assumes a solvent-induced field energy produced solely by the solute charges:[234]

$$\Delta G_{pol} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right) \sum_{i,j} \frac{q_i q_j}{f_{ij}}, \tag{2.13}$$

where ε_p and ε_w are the dielectric constants of the solute and solvent respectively, and q_i, q_j are the partial charges of the solute. The f_{ij} term is the effective molecular

Born radius taken as:

$$f_{ij} = [r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)]^{\frac{1}{2}}, \tag{2.14}$$

where r_{ij} is the distance between atoms i and j, and R_i is the effective Born radius of atom i,[234] which encompasses both the inherent atomic radius, as well as that arising from the influence of its neighbouring atoms. An analytical solution for R_i is then obtained by applying the Coulomb field approximation (CFA), where it is assumed that as the solvation proceeds the dielectric displacement of atom i is Coulombic in nature. The effective Born radius then becomes

$$\frac{1}{R_i} = \frac{1}{a_i} - \frac{1}{4\pi} \int_{solute,\, r>a_i} \frac{1}{r^4} dV; \tag{2.15}$$

a_i is the atomic radius of atom i, which is at the centre of a cavity over which the integral is evaluated.[235] GB solvent models differ in their approximation of the integral in equation (2.15). For instance, in the model developed by Hawkins, Cramer and Truhlar ($\mathrm{GB}^{\mathrm{HCT}}$)[236] the integral is evaluated over the van der Waals (vdW) spheres of the solute atoms, which tend to underestimate the cavity size (molecular volume). This model was improved by Onufriev, Bashford and Case ($\mathrm{GB}^{\mathrm{OBC}}$),[237] who rescaled the effective radii to account for the degree of atomic burial within the cavity, and so provide a better approximation of the molecular volume. Recently, further improvements have been reported for the GB-Neck2 parameter set.[238]

2.4 Structure Prediction by Basin-Hopping Global Optimisation

In the realm of computational biology, global optimisation mainly involves finding the global minimum of the potential energy function for the system of interest. Low-lying minima on the PES often form stable ensembles on the free energy surface (FES) computed at low temperatures; hence, locating these structures is an important area of research. However, even for a short peptide sequence, the PES supports numerous local minima, many of which may be separated by high energy barriers. The task then becomes finding an appropriate search algorithm that can predict the global minimum structure in an efficient and unbiased manner.

The basin-hopping global optimisation procedure[239,240] has been utilised in this work to locate the global minimum. This method employs a hypersurface deformation, without changing the global minimum of the potential energy surface. Each configuration on the PES can be represented by a unique $3N$-dimensional vector $\mathbf{X}$, where N is the number of atoms, and the energy corresponding to $\mathbf{X}$ is given by $V(\mathbf{X})$. Accordingly, the energy obtained by a minimisation on the PES starting from $\mathbf{X}$ can be taken as $\min\{V(\mathbf{X})\}$.

By performing energy minimisations for every point on the PES, the transformed PES is obtained: $\widetilde{V}(\mathbf{X}) = \min\{V(\mathbf{X})\}$. The PES is now represented by 'basins of attraction'[241] of discrete energies; each hosting the configurations whose minimisation led to a particular stationary point (local minimum). Basin-hopping effectively removes all transition state regions, as shown in Figure 2.1, and thus results in accelerated motion on the PES.

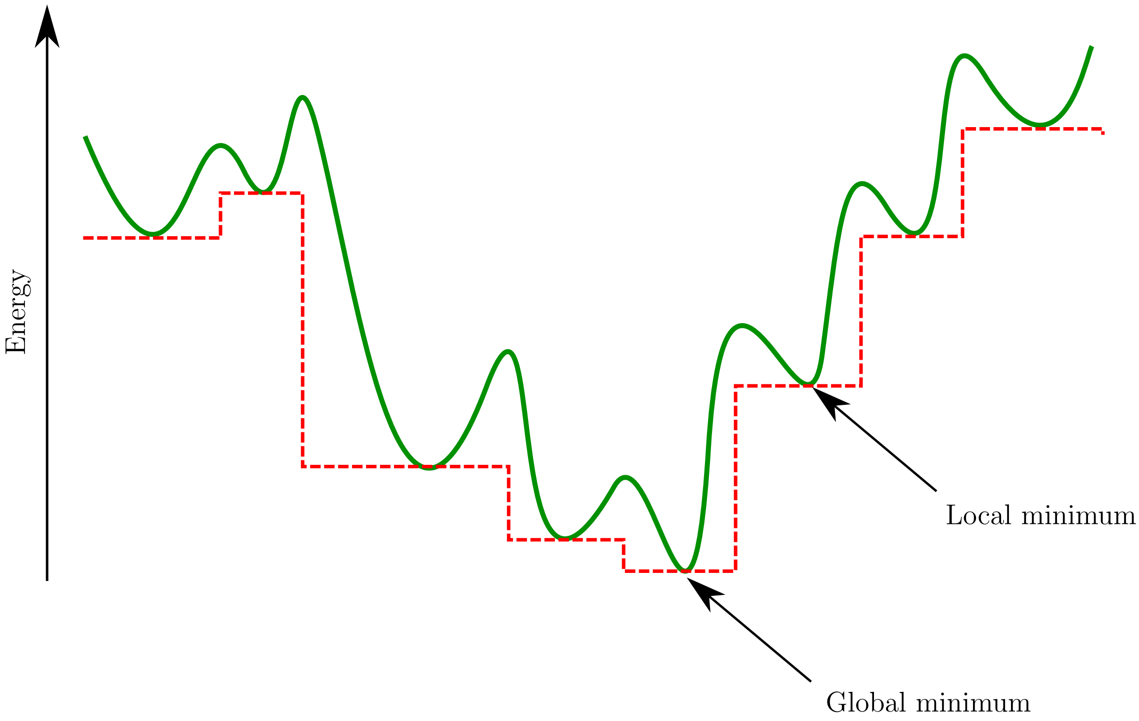


Figure 2.1: Illustration of the energy landscape transformation. The green curve is the original surface, and the red curve represents the transformed surface. Each local minimum on the original surface $V(\mathbf{X})$ corresponds to a 'basin of attraction' on the transformed surface $\widetilde{V}(\mathbf{X})$.

The main basin-hopping algorithm is summarised below:

Geometry perturbation and energy minimisation

- Let $\boldsymbol{c}_i$ represent the initial configuration of the system whose global minimum is to be determined, and V_i represent its energy.
- An appropriate step-taking routine (e.g. a randomised Cartesian, angular, or

dihedral move) is then employed to perturb $\boldsymbol{c}_i$.

- The perturbation is then followed by an energy minimisation to yield a new configuration $\boldsymbol{c}_n$ with energy V_n.

Acceptance or rejection of a step

- A step is accepted if the energy of the new configuration V_n is less than that of the initial one V_i: $V_n < V_i$.
- A step is also accepted if $V_n > V_i$ and satisfies the Metropolis criterion:[148] $Ran(0 \to 1) < \exp[-(V_n - V_i)/k_BT]$; where $Ran(0 \to 1)$ is a uniformly generated random number between zero and one.
- If the step is accepted, the next perturbation is applied to $\boldsymbol{c}_n$.
- Otherwise, the step is rejected and another perturbation is applied to the initial configuration $\boldsymbol{c}_i$.

In the present work, energy minimisations were performed using the limited-memory BFGS (L-BFGS) algorithm,[242,243] named for Broyden,[244] Fletcher,[245] Goldfarb,[246] and Shanno.[247] This algorithm is well suited for large-scale problems, since the user is able to control the amount of storage required for approximation of the Hessian matrix. The basin-hopping procedure has been implemented in the GMIN software,[248] which was used in this thesis.

2.5 Building Kinetic Transition Networks from Discrete Path Sampling

In discrete path sampling[249–251] (DPS) the aim is to determine the kinetics of a system from a collection of pathways, connecting reactant (e.g. an unfolded protein) to product (e.g. a folded protein). Wales[249] describes a discrete path as a sequence of local minima on the potential energy surface (PES) and the transition states that directly connect them. Recall that, within the potential energy landscape framework, minima are classified as stationary points with all their non-zero Hessian eigenvalues positive, and transition states are defined as stationary points with a single negative Hessian eigenvalue.

2.5.1 Finding an initial path

The first step in DPS is to construct an initial path from the reactant (A) to the product (B). Appropriate initial endpoints (three-dimensional coordinates) must be chosen to represent the reactant and product. This is the only stage in DPS where any prior knowledge is required. Once suitable representative structures are selected, a cycle of events, discussed below, then proceeds to connect them.

Doubly-nudged elastic band method

Transition state guesses are first obtained using the doubly-nudged[252] elastic band[253,254] (DNEB) procedure. A double-ended interpolation between A and B produces an intermediate set of images $[\mathbf{X}_1, \mathbf{X}_2 ... \mathbf{X}_M]$, where $\mathbf{X}_i$ 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,

$$E_{spr} = \frac{1}{2} k_{spr} \sum_{i=1}^{M+1} |\mathbf{X}_i - \mathbf{X}_{i-1}|^2, \tag{2.16}$$

where k_{spr} is the force constant of the spring, M is the number of intermediate images and $\mathbf{X}_0$ and $\mathbf{X}_{M+1}$ represent $3N$-dimensional vectors of the coordinates of the endpoints. Additionally, each image also has a true potential, denoted $V(\mathbf{X}_i)$. An important element of the nudged elastic band approach is the derivation of the elastic band gradient, which involves a projection of the forces due to the spring potential and the true potential. This projection 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.[254] The gradient of the doubly-nudged elastic band[252] is taken as:

$$\mathbf{g}_{DNEB} = \mathbf{g}^{\perp} + \mathbf{g}_{spr}^{\parallel} + \mathbf{g}_{spr}^{*}, \tag{2.17}$$

where $\mathbf{g}^{\perp}$ is the component of the true gradient perpendicular to the path, and $\mathbf{g}_{spr}^{\parallel}$ is the component of the spring gradient parallel to the path. $\mathbf{g}_{spr}^{*}$ is given by:

$$\mathbf{g}_{spr}^{*} = \mathbf{g}_{spr}^{\perp} - (\mathbf{g}_{spr}^{\perp} \cdot \hat{\mathbf{g}}^{\perp})\hat{\mathbf{g}}^{\perp}, \tag{2.18}$$

with $\hat{\mathbf{g}}^{\perp}$ representing a unit vector along the component of the true gradient perpendicular to the path.

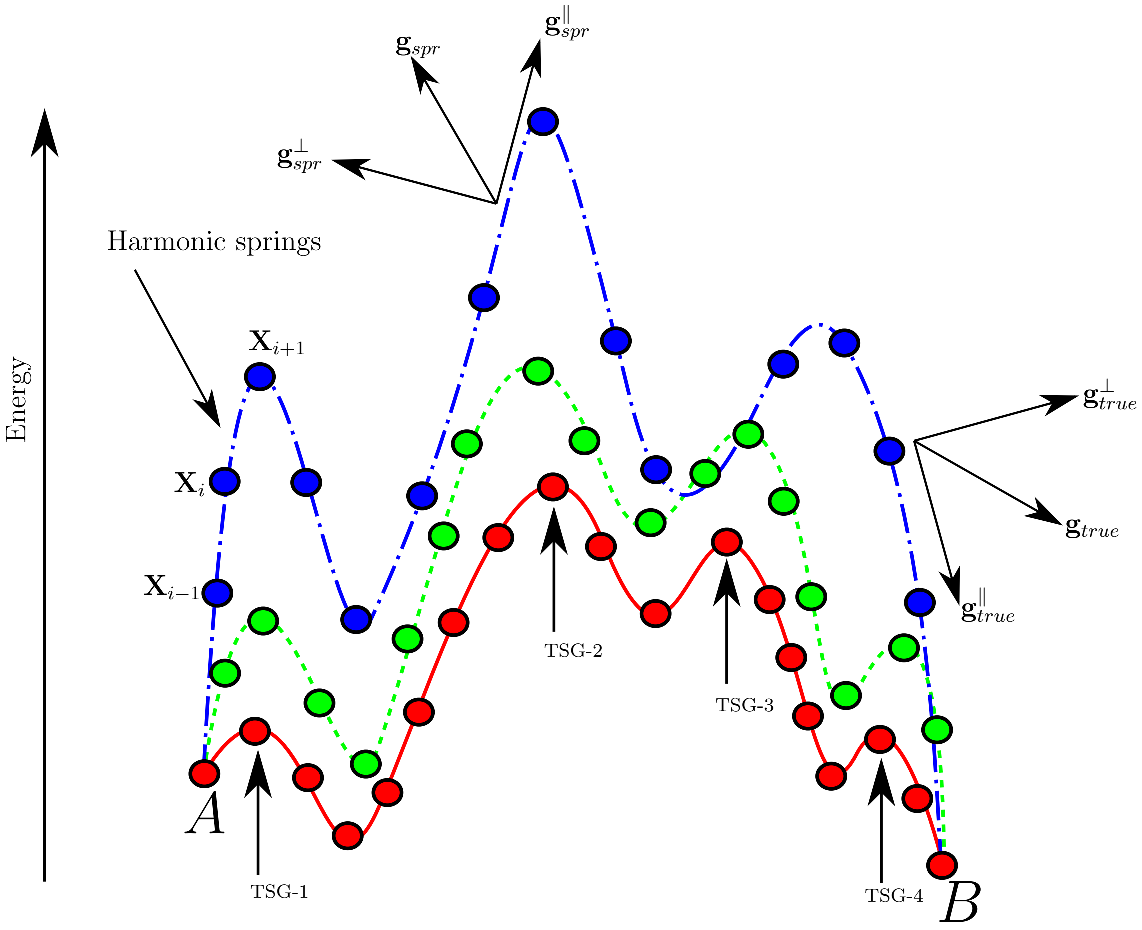


Figure 2.2: Summary of the doubly-nudged elastic-band procedure, which is employed to obtain transition state candidates in DPS. The blue curve represents an initial interpolation between A and B. The elastic-band energy is iteratively minimised, until the gradient falls below a user-defined threshold. The red curve corresponds to the converged elastic-band; each transition state guess (TSG) is indicated on the curve. These TSGs can then be refined further using methods such as hybrid eigenvector-following (discussed in the text).

The set of paired images is then relaxed by L-BFGS minimisations (Figure 2.2); $\mathbf{g}_{DNEB}$ serves to stabilise the path during minimisations. Without the formulation in equation (2.18), the spring gradient would prevent the elastic band from forming a curved energy path (due to 'corner-cutting'), and the true gradient would cause images to slide away from transition state regions towards proximate minima. The objective is to optimise the elastic band until a connected path between A and B is obtained. However, approximate transition states found during DNEB calculations must first be converged more tightly via hybrid eigenvector-following (HEF).[255,256]

Hybrid eigenvector-following

This is a single-ended eigenvector-following procedure for locating transition states on the potential energy surface.[255,256] In hybrid eigenvector-following (HEF) the smallest non-zero Hessian eigenvalue (λ_{min}) and the corresponding eigenvector ($\mathbf{x}_{min}$) are used for uphill searches, and minimisation (e.g. using the L-BFGS algorithm) is performed in the tangent space until a transition state is found.[256] The smallest non-zero eigenvalue can be found using the Rayleigh-Ritz ratio:[255]

$$\lambda(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{H} \mathbf{x}}{\mathbf{x}^2}, \tag{2.19}$$

where $\mathbf{x}$ represents the displacement from the current configuration $\mathbf{X}$, and $\mathbf{x}^T$ is its transpose. To avoid explicit computation of the Hessian $\mathbf{H}$, $\lambda(\mathbf{x})$ is estimated from the numerical second derivative of the energy:

$$\lambda(\mathbf{x}) \approx \frac{V(\mathbf{X}+\mathbf{x}) + V(\mathbf{X}-\mathbf{x}) - 2V(\mathbf{X})}{\mathbf{x}^2}, \tag{2.20}$$

where $V(\mathbf{X})$ is the energy of configuration $\mathbf{X}$, and $\mathbf{x}$ is a very small change ($|\mathbf{x}| \ll 1$) in its geometry. Hence, the minimum value of $\lambda(\mathbf{x})$ can be obtained by differentiating equation (2.20):

$$\frac{\partial \lambda(\mathbf{x})}{\partial \mathbf{x}} = \frac{\nabla V(\mathbf{X}+\mathbf{x}) - \nabla V(\mathbf{X}-\mathbf{x}) - 2\lambda(\mathbf{x})\mathbf{x}}{\mathbf{x}^2}; \tag{2.21}$$

which can be minimised using the L-BFGS algorithm to give λ_{min}, and thus $\mathbf{x}_{min}$.[†]

However, for numerically large values of $V(\mathbf{X})$, equation (2.21) may suffer from loss of precision due to roundoff error. Ergo, an alternative formulation has been proposed for estimating λ_{min}:[257]

$$\lambda(\mathbf{x}) \approx \frac{\{\nabla V(\mathbf{X}+\mathbf{x}) - \nabla V(\mathbf{X}-\mathbf{x})\} \cdot \mathbf{x}}{2\mathbf{x}^2}. \tag{2.22}$$

In each HEF optimisation cycle, $\lambda(\mathbf{x})$ is computed, until the root-mean-square deviation falls below a user-defined threshold. An uphill step is then taken in the direction specified by $\mathbf{x}$, and L-BFGS minimisations are performed in the orthogonal subspace. Through successive iterations,[‡] tightly converged transition states are located on the PES. These transition states are then connected to minima by

[†] Overall rotational and translational degrees of freedom are projected out to avoid convergence of $\mathbf{x}$ to one of the corresponding eigenvectors.

[‡] HEF iterations continue until the overlap of $\mathbf{x}$ between the current and previous cycle exceeds 0.999.

following their approximate steepest-descent paths parallel and anti-parallel to the corresponding eigenvector.

Dijkstra's shortest path

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 to be connected to yield the shortest path. Carr *et al.*[258] employed Dijkstra's algorithm,[259] which selects minima based on a minimised Euclidean distance metric.

In a database of stationary points the total set of minima can be described using a complete graph $G(M, D)$; where M represents all minima and D all the edges between them. Edges exist between all minima in the database, and the edge weight between arbitrary minima x and y is taken as a function of the Euclidean distance:

$$w(x,y) = \begin{cases} 0, & \text{if } x \text{ and } y \text{ are connected by a single transition state,} \\ \infty, & \text{if } n(x,y) = n_{max}, \\ f(d(x,y)) & \text{otherwise.} \end{cases} \tag{2.23}$$

In equation (2.23), $n(x, y)$ is the number of connection attempts between minima x and y, n_{max} is the maximum connection attempts (set by the user), and $d(x, y)$ is their minimised Euclidean distance.[§] The above metric is used at the beginning of each DNEB/HEF cycle to select appropriate minima in the existing database for connection attempts. This process is repeated until there are no missing connections along the path—i.e. $w(A, B) = 0$.

2.5.2 Refining the stationary point database

The initial path found between A and B is usually long with many high barriers, particularly for endpoints distant in configuration space. The objective then is to grow and refine the stationary point database by making more connections between minima. At any point in DPS, the fastest path ($B \leftarrow A$) is taken as the path 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; see § 2.7.2). Once the fastest path is identified, the energy barriers are

[§] The minimised Euclidean distance is computed with respect to overall rotation, translation and permutation of identical atoms.

calculated, as well as the distances between minima along the path that are not connected by a single transition state. This path is then perturbed in various ways in search of paths that are more kinetically relevant.

SHORTCUT

This scheme chooses minima from the current 'fastest' path that are separated by at least a minimum number of transition states (steps).[257,260] The minima are then connected using the procedures discussed in § 2.5.1. The SHORTCUT procedure usually decreases the total number of steps on the path and leads to an increase in the rate constant.

SHORTCUT BARRIER

Alternatively, the SHORTCUT BARRIER method[258,260] 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, thus improving the rate constant.

UNTRAP

While the SHORTCUT and SHORTCUT BARRIER approaches improve the rate constant, these procedures may also introduce kinetic traps on the PES. These traps take the form of low-lying minima separated from the product minimum by high barriers. Most traps are artificial and are due to incomplete sampling. To find low-barrier paths for these minima the UNTRAP scheme[260] 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 at 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.

2.6 Calculating Thermodynamic Properties

The canonical partition function, $Z(T)$, is computed as a sum of contributions from the basins of attraction of different local minima in our stationary point database:[210]

$$Z(T) = \sum_i Z_i(T); \tag{2.24}$$

where $Z_i(T)$ is the partition function for the catchment basin of minimum i. The catchment basin of i represents all the configurations whose minimisations led to i. It follows that the equilibrium occupation probability of minimum i is given by:

$$p_i^{eq} = \frac{Z_i(T)}{Z(T)}. \tag{2.25}$$

A harmonic approximation is used to estimate the vibrational partition function of each minimum i:[210,261]

$$Z_i(T) = \frac{n_i e^{-\beta V_i}}{(\beta h \bar{\nu}_i)^{\kappa}}; \tag{2.26}$$

where n_i is a factor that ensures that Z_i accounts for all nonsuperimposible permutation isomers of i. For a system of N_A atoms of A, N_B atoms of B, and so on, $n_i = 2N_A!N_B!N_C!\cdots/o_i$; where o_i is the point group order.[210] $\beta = 1/k_BT$ (k_B is the Boltzmann constant), V_i is the potential energy of minimum i, h is Plank's constant, and $\kappa = 3N - 6$ (total number of vibrational degrees of freedom).[210]

The $\bar{\nu}_i$ term in equation (2.26) represents the geometric mean vibrational frequency, $\bar{\nu}_i = [\prod_{\alpha=1}^{\kappa} \nu_\alpha(i)]^{1/\kappa}$; where the vibrational frequencies of each minimum, $\nu_\alpha(i)$, are computed via normal mode analysis.[210,261]

Equilibrium statistical mechanics can then be employed to estimate the free energy of minimum i,¶

$$F_i(T) = -k_BT \ln Z_i(T), \tag{2.27}$$

the total energy in the canonical ensemble,

$$\langle E \rangle = \kappa k_B T + \frac{1}{Z(T)} \sum_i Z_i V_i, \tag{2.28}$$

and the canonical heat capacity at constant volume (the derivative of $\langle E \rangle$ with respect to T),[210]

$$C_v(T) = \kappa k_B - \frac{(Z_1(T))^2}{k_BT^2(Z_0(T))^2} + \frac{Z_2(T)}{k_BT^2 Z_0(T)}; \tag{2.29}$$

where

$$Z_p(T) = \sum_i Z_i(V_i)^p. \tag{2.30}$$

¶The free energies of transition states are computed in a similar fashion, except the imaginary vibrational frequency is omitted from $\bar{\nu}_i$.

2.7 Calculating Rate Constants

2.7.1 Unimolecular rate constants

Transition state theory (TST) can be used to estimate the unimolecular canonical rate constant, $k_i^\dagger(T)$, out of minimum i through transition state † at temperature T:

$$k_i^\dagger(T) = \frac{k_B T Z^\dagger(T)}{h Z_i(T)} e^{-\Delta V / k_B T}, \tag{2.31}$$

where $Z^\dagger(T)$ and $Z_i(T)$ are the partition functions for † and i, respectively; ΔV is the potential energy difference between † and i. For a transition from minimum i to j, the total rate constant $k_{ij}(T)$ may be computed by summing $k_i^\dagger(T)$ for all intervening transition states. This procedure provides an upper-bound to the true rate constant; since it does not correct for recrossing events (certain transition states may be encountered more than once), which reduce the overall rate.

2.7.2 Steady-state rate constant

The TST unimolecular rate constants for elementary transitions can be used to compute rate constants k_{AB} and k_{BA} between reactant A and product B states. In an optimised DPS stationary point database A and B may be connected by an infinite number of discrete paths. The aim here is to locate the most kinetically relevant paths connecting A and B; which may be achieved by taking a weighed sum all discrete paths, based on the equilibrium occupation probabilities of the states on those paths.

At equilibrium:

$$\frac{p_a(t)}{p_A(t)} = \frac{p_a^{eq}(t)}{p_A^{eq}(t)} \quad \text{and} \quad \frac{p_b(t)}{p_B(t)} = \frac{p_b^{eq}(t)}{p_B^{eq}(t)}, \tag{2.32}$$

where $p_a(t)$ and $p_b(t)$ represent the occupation probabilities of a and b minima at time t within states A and B, respectively. If we assume that the dynamics is Markovian (memoryless), then the change in the occupation probability of minimum a, for example, can be expressed as:[249]

$$\frac{dp_a(t)}{dt} = \sum_{b \neq a} [k_{ab} p_b(t) - k_{ba} p_a(t)], \tag{2.33}$$

where k_{ab} is the unimolecular rate constant for transitions from b to a, with $b \neq a$ indicating that the sums are over geometrically distinct minima.[210]

From equation (2.33), it follows that:[249]

$$\frac{dp_A(t)}{dt} = k_{AB}p_B(t) - k_{BA}p_A(t) \quad \text{and} \quad \frac{dp_B(t)}{dt} = k_{BA}p_A(t) - k_{AB}p_B(t), \tag{2.34}$$

where

$$k_{AB} = \frac{1}{p_B^{eq}} \sum_{a \in A} \sum_{b \in B} k_{ab} p_b^{eq} \quad \text{and} \quad k_{BA} = \frac{1}{p_A^{eq}} \sum_{a \in A} \sum_{b \in B} k_{ba} p_a^{eq}. \tag{2.35}$$

In many cases it may be difficult to classify all minima in the stationary point database as belonging to set A or B. It therefore becomes necessary to define a third set I for all intervening minima. To recover two-state rate constants, we must assume that the time evolution of the occupation probability of each minimum i in I is negligible:[249]

$$\frac{dp_i(t)}{dt} = \sum_j k_{ij}p_j(t) - p_i(t)\sum_i k_{ji} \approx 0; \tag{2.36}$$

thus,

$$p_i(t) = \frac{\sum\limits_j k_{ij}p_j(t)}{\sum\limits_i k_{ji}}. \tag{2.37}$$

Accordingly, k_{AB} and k_{BA}, within the steady-state approximation, can be expressed as:[249]

$$k_{AB}^{SS} = \frac{1}{p_B^{eq}} \sum_{a \leftarrow b} \frac{k_{ai_1}k_{i_1i_2}...k_{i_nb}p_b^{eq}}{\sum\limits_{j_1} k_{j_1i_1} \sum\limits_{j_2} k_{j_2i_2} ... \sum\limits_{j_n} k_{j_ni_n}} \tag{2.38}$$

and

$$k_{BA}^{SS} = \frac{1}{p_A^{eq}} \sum_{b \leftarrow a} \frac{k_{bi_1}k_{i_1i_2}...k_{i_na}p_a^{eq}}{\sum\limits_{j_1} k_{j_1i_1} \sum\limits_{j_2} k_{j_2i_2} ... \sum\limits_{j_n} k_{j_ni_n}}, \tag{2.39}$$

respectively. The fastest path between A and B is then appropriately defined as the discrete path making the largest contribution to k_{BA}^{SS}, and can be extracted from the kinetic transition network using Dijkstra's shortest path algorithm.

2.7.3 New graph transformation rate constant

The new graph transformation (NGT) procedure[262] provides a robust formalism for extracting two-state rate constants (k_{AB} and k_{BA}) from kinetic transition networks.

For a given minimum a the mean waiting time in that minimum (i.e. time for escape from a) can be defined as:

$$\tau_a = \frac{1}{\sum\limits_\beta k_{\beta a}}, \tag{2.40}$$

and the transition probability from a to b can be estimated as:

$$P_{ba} = k_{ba}\tau_a. \tag{2.41}$$

Thus, the transition probability from a to the set B is given by:

$$P_{Ba} = \sum_{b\in B} P_{ba}. \tag{2.42}$$

When the NGT scheme is applied, minima in the database are first assigned to one of three sets: reactant (A), product (B), or intervening (I). Minima in I are then progressively removed and the transition probabilities and waiting times are renormalised to leave the average mean first passage time (MFPT)[‖] unchanged.[263]

The overall rate constants are then given by:

$$k_{AB}^{NGT} = \frac{1}{p_B^{eq}} \sum_{b\in B} \frac{P'_{Ab} p_b^{eq}}{\tau'_b} \quad \text{and} \quad k_{Ba}^{NGT} = \frac{1}{p_A^{eq}} \sum_{a\in A} \frac{P'_{BA} p_a^{eq}}{\tau'_a}, \tag{2.43}$$

where the prime symbol is used to indicate that the terms have been renormalised. The NGT method is advantageous in that it avoids the steady-state assumption, and can be implemented efficiently to obtain kinetics at different temperatures. Rate constants reported in this dissertation were computed via the NGT procedure.

2.7.4 Free energy regrouping

At equilibrium, protein reactant (A) and product (B) states correspond to ensembles of unfolded and folded structures, respectively. To appropriately define these ensembles for our kinetic transition networks, we employed an iterative regrouping scheme[264] (REGROUPFREE in PATHSAMPLE), which lumps minima into free energy groups based on the relative free energy barriers.

In this scheme, the free energy of a group of minima I is given by:

$$F_I(T) = -k_B T \sum_{i\in I} \ln Z_i(T), \tag{2.44}$$

and the free energy of the group of transition states connecting I to J is taken as:

$$F_{IJ}(T) = -k_B T \sum_{(i,j)^\dagger} \ln Z^\dagger(T), \tag{2.45}$$

[‖] The simulation time for a path from A to B can be estimated as a sum of the waiting times of states on the path. The weighted sum of simulation times for all discrete paths connecting A to B is the MFPT.

where $Z^{\dagger}(T)$ is the canonical partition function of the transition state connecting i and j.[264]

Unless otherwise stated, steady-state and NGT rate constants reported in this thesis were computed by first lumping minima and transition states into their respective free energy groups, based on a pre-defined free energy threshold for the barrier heights.

2.8 Representing Energy Landscapes as Disconnectivity Graphs

Originally introduced by Becker and Karplus,[265] disconnectivity graphs have played a pivotal role in conceptualising energy landscapes.[266–268] Figure 2.3 illustrates how a disconnectivity graph (red curve) 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 energy of that conformer. At the energy threshold E_2 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, at this threshold the two sets of minima are still disjoint, since the transition state connecting them lies above the threshold. When the threshold is high enough (i.e. at E_3) the two sets of minima merge. Since the energy spacing (ΔE) determines how the analysis is performed, the graph is most meaningful when the thresholds are spaced at suitable regular intervals. For instance, if the spacing is too small, it may be difficult to obtain any useful information from the analysis, and if it is too large, the graph may be misleading (in that, minima separated by particularly high barriers may be grouped together.

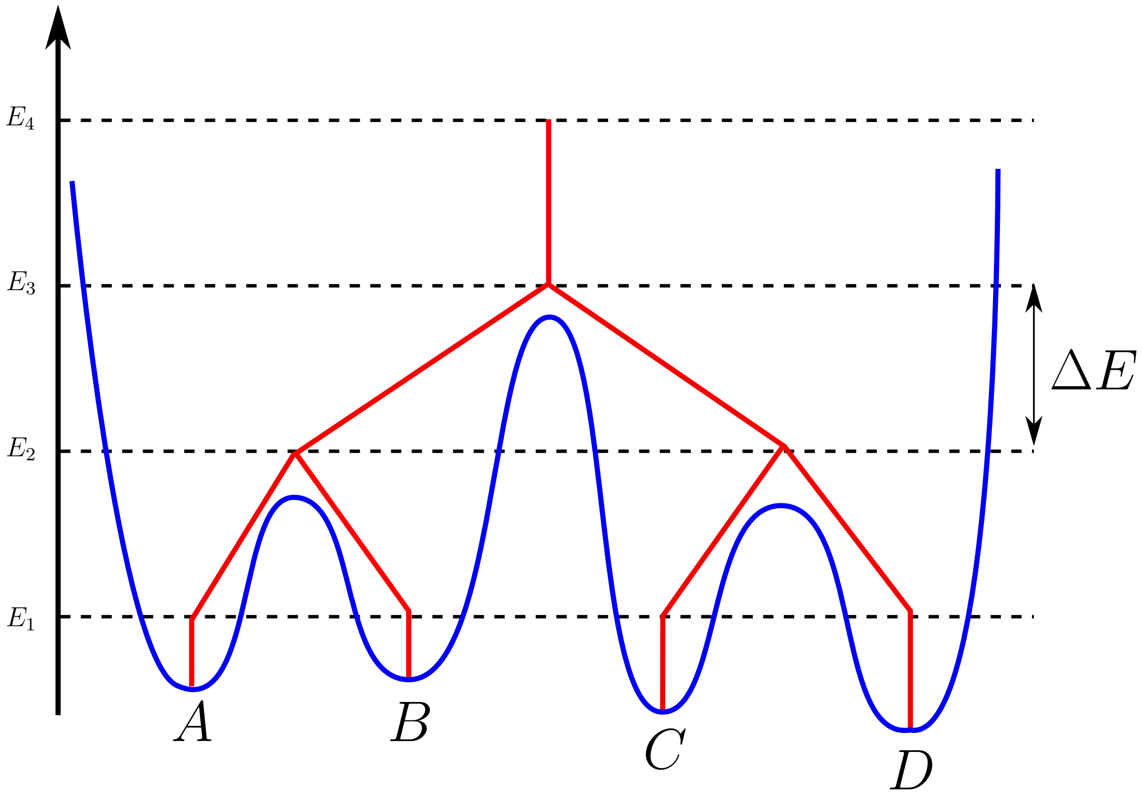


Figure 2.3: Construction of a disconnectivity graph from a database of stationary points. Minima are labelled A–D on the energy surface (blue). In the disconnectivity graph (red) each local minimum is represented by a vertical line, starting at the energy of that minimum. At a given energy threshold, $E + n\Delta E$, minima connected by transition states that lie below the threshold are grouped into disjoint sets.

The final analysis is a graph that resembles a tree of some sort; for example, the 'palm tree' disconnectivity graph, depicted in Figure 2.4, is characterised by a well-defined global minimum with the minima leading towards it separated by relatively small downhill barriers.[266] These graphs may aid in identifying putative 'kinetic traps' in the landscape,[269] which may take the form of low-lying minima of comparable energy to the global minimum, but separated by particularly high energy barriers.

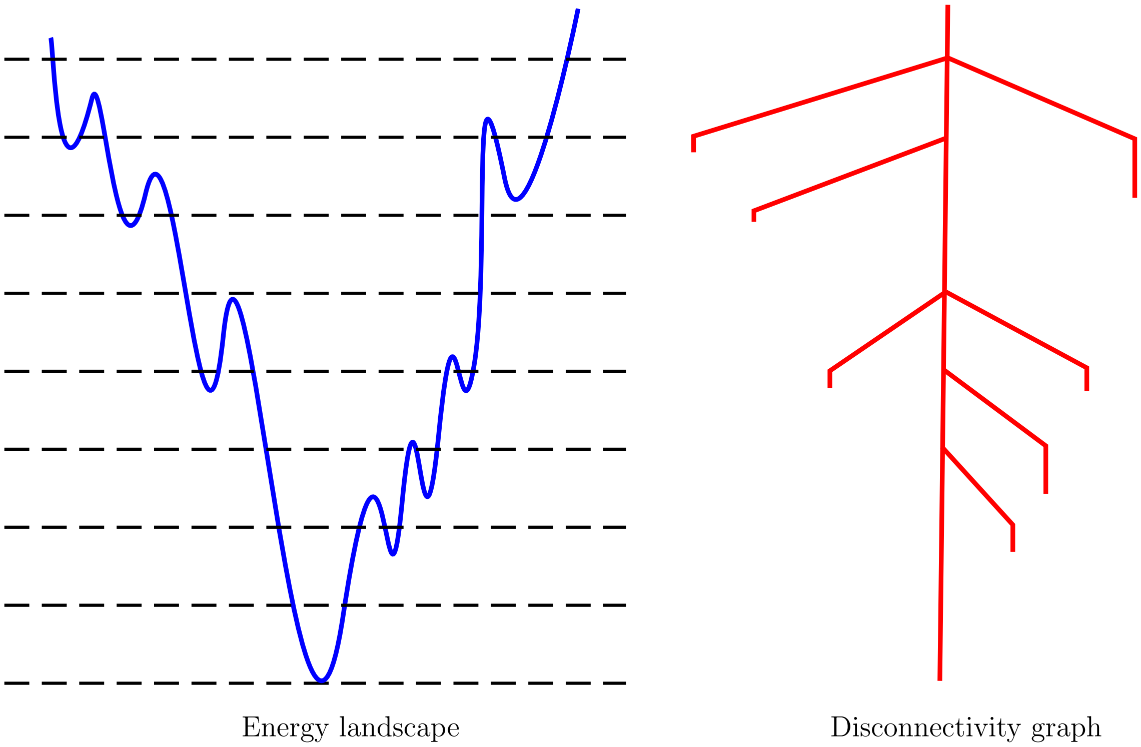


Figure 2.4: An energy landscape and the corresponding disconnectivity graph. The disconnectivity graph has a palm tree-like topology.

Free energy disconnectivity graphs can be constructed from their corresponding potential energy graphs by employing the superposition approach and the harmonic vibrational densities of states, at a chosen temperature.[261,270–272] In approximating the free energy using disconnecting graphs, these graphs can also be coloured based on some order parameter (e.g. number of hydrogen-bonds, radius of gyration etc.) to give further insight into the underlying thermodynamics and kinetics of the system.

3

Protein Folding as a Function of Local Rigidification

3.1 Introduction

Computer simulations continue to improve our understanding of protein folding.[273–275] However, the interplay of hierarchical length and time scales poses a significant challenge to *in silico* investigations. With standard techniques, conformational dynamics of proteins can only be probed over relatively short time scales, which do not capture important biological processes. Accordingly, advancements in computing code and hardware,[276–278] sampling techniques,[190,201] and energy functions[227,279,280] have been actively pursued, to achieve longer spatiotemporal scales.[203,281,282] Alternatively, some of the complexity may be mitigated by developing approaches that reduce the number of degrees of freedom.[283–285]

Coarse-graining involves reducing the degree of detail used to describe a system. Numerous coarse-grained models have been proposed and implemented for biomolecules, with varying levels of success.[21,286–293] In one approach, amino acid side-chains and α-helices are represented as spheres and cylinders, respectively;[286] in elastic network models[287,294] amino acid residues are reduced to beads interacting via interresidue potentials. Structure-based potentials, such as Gō models,[21,295] 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 penalised. Over the last decade, much effort has been expended on deriving multiscale procedures[296–298] for simulating biomolecules. These methods aim to capitalise 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 parametrised and optimised. 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,[299–302] to address some of the inherent difficulties in modelling proteins. This framework has been benchmarked for structure prediction of model peptides using all-atom potentials[301] and, in the current contribution, we extend it to explore the global thermodynamics and mechanics of peptide folding. Local rigidification exploits the separation of time scales[283–285,303,304] between low frequency vibrational modes and localised, fast vibrations, which suggests that specific regions 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 landscape is significantly reduced, resulting in substantial computational speedup.[301] 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 proteins.

3.2 Methods

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 stabilisation, and a distinct funnelled energy landscape.[45,251] Therefore, β-hairpins are good candidates for benchmarking new protein folding simulation methods.

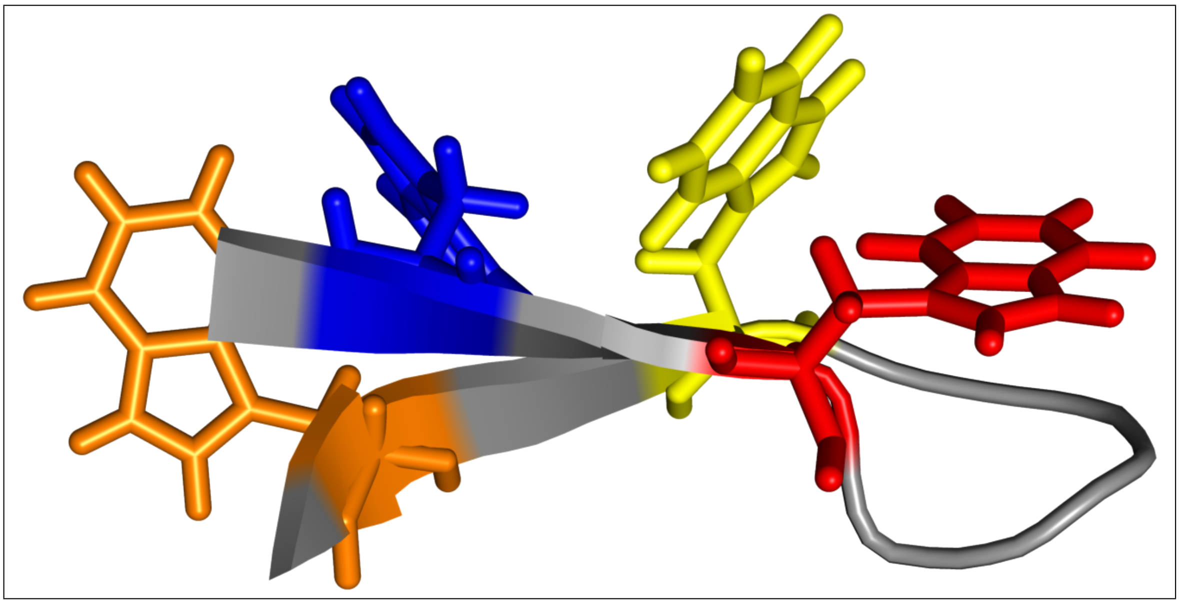


Figure 3.1: NMR structure for the tryptophan zipper 1 (TZ1; PDB code: 1LE0) showing the characteristic stacking of indole rings.

In this work we focus on the tryptophan zipper 1 (TZ1). TZ1 is one of the family of 12-residue β-hairpins designed by Cochran and coworkers.[213] 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 time scale,[49] making them accessible in fully atomistic simulations.

The NMR structure for TZ1 is shown in Figure 3.1. 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 sequence: SWTWEGNKWTWK). TZ1 was represented by the AMBER ff99SB[226] potential energy function and the GB$^{\rm OBC}$ solvation potential.[237] 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.[305] No periodic boundary conditions were imposed on the system, and no cutoffs were set for non-bonded interactions. For calculation of effective atomic Born radii a cutoff of 25 Å was used. The AMBER potential was symmetrised, as described by Małolepsza *et al.*,[306] so that interconvertible permutational isomers have the same energy.

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,[307] structure prediction of organic compounds[308,309] and water clusters,[300,302,310] protein-protein docking,[311,312] and self-assembly of virus capsids.[299,313]

Definitions

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

$$\mathbf{X}_I = \frac{1}{n_I} \sum_{i \in I}^{n_I} \mathbf{x}_i, \tag{3.1}$$

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:[299–302]

$$\mathbf{p}_I = \theta_I \hat{\mathbf{p}}_I, \tag{3.2}$$

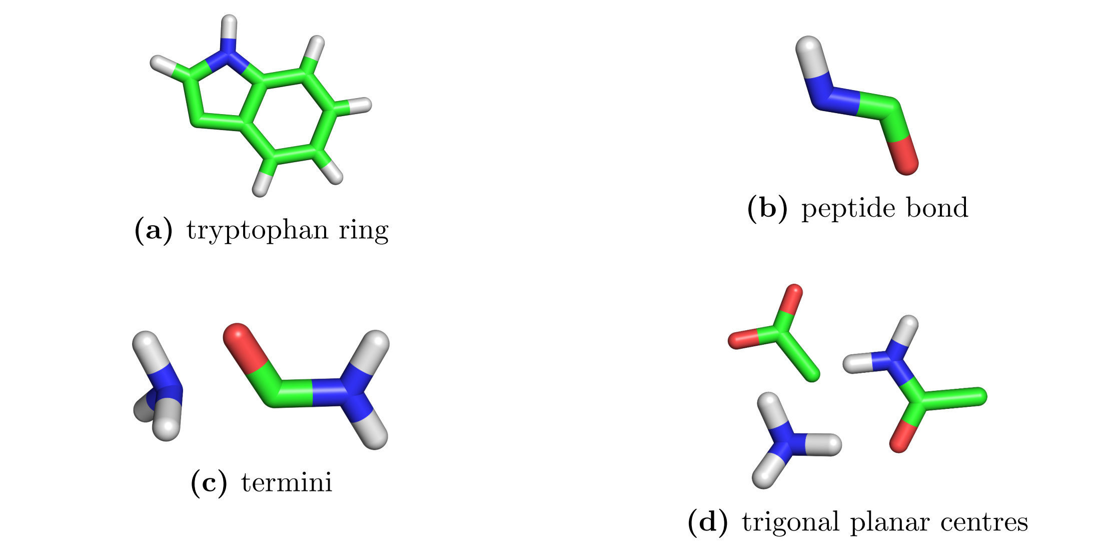
where $\mathbf{p}_I$ is a rotation vector, characterising the angle θ_I and axis $\hat{\mathbf{p}}_I$ of rotation.[299,300] Rigid body reference coordinates are usually obtained from the global minimum of the potential energy landscape, corresponding to the unconstrained representation.[301]

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 coordinates of the n^{th} free atom; N is the number of LRBs and $\{\mathbf{X}_N, \mathbf{p}_N\}$ are the rigid body coordinates of the N^{th} 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 ($\mathbf{R}_I$),[314,315] which can be applied to the reference structure of the rigid body ($\mathbf{x}^0_{i \in I}$) to obtain the atomistic coordinates:

$$\mathbf{x}_{i \in I} = \mathbf{X}_I + \mathbf{R}_I \mathbf{x}^0_{i \in I}. \tag{3.3}$$

Groupings and schemes

Suitable LRB groupings can be suggested from principal component analysis,[316,317] approaches developed from graph theory,[318–320] or some other metric. In this study, the LRB groupings for TZ1 were adopted from previous work;[301] namely tryptophan rings, peptide bonds, termini and trigonal planar centres (Figure 3.2).



(a) tryptophan ring

(b) peptide bond

(c) termini

(d) trigonal planar centres

Figure 3.2: Local rigid bodies considered for tryptophan zipper 1.

These groupings were used to define several local rigidification schemes, outlined in Figure 3.3. 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 (around 27 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 is more aggressive—with about 60 percent of the atoms grouped as LRBs ($25 \times 6 + 89 \times 3 = 417$ degrees of freedom).

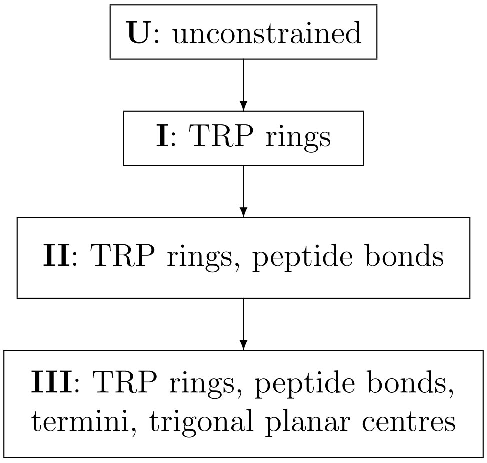

U: unconstrained
I: TRP rings
II: TRP rings, peptide bonds
III: TRP rings, peptide bonds, termini, trigonal planar centres

Figure 3.3: 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.

Potential energy landscapes with LRBs

The local rigidification was applied within the framework of potential energy landscape theory.[210] Conceptually, the potential energy landscape (PEL) supports the local minima and the transition states that connect them (§ 2.1). These stationary points constitute a kinetic transition network (KTN; § 2.5), from which the global thermodynamics (§ 2.6) and kinetics (§ 2.7) may be extracted. The complexity of the PEL increases as the system size grows. Hence, a LRB formalism becomes appealing; since this approach effectively reduces the number of stationary points on the PEL, leading to increased computational efficiency.

Energy minimisations

Energy minimisations were performed using a customised L-BFGS algorithm,[242,243] in the mixed coordinate space. This approach has the advantage of reducing the number of minimisation steps required for convergence.[301] Rühle *et al.*[302] have developed a method for computing the energy gradients with respect to generalised coordinates (mixed or atomistic), hence providing a convenient means of measuring convergence, which is invariant to coordinate transformations, as it should be.[302]

Building kinetic transition networks

Appropriate initial endpoints for the reactant (A) and product (B) were first chosen. Here, a denatured conformation (obtained from an MD simulation at 330 K), with a high occupation probability in the vicinity of the experimental melting temperature,[213] was selected as the reactant. The product was represented by the global minimum of the potential energy landscape (obtained by basin-hopping global optimisation; § 2.4)[210,240,248] 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[252] elastic band[253,254] (DNEB) procedure was then used to locate transition state candidates, which were converged further using hybrid eigenvector-following (HEF).[255,256] 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 generalised coordinate space.[302] Iterative DNEB/HEF searches[249–251] eventually provided a global survey of the potential energy landscape. All these procedures are implemented in the OPTIM[321] and PATHSAMPLE[322] programs, which are available for use under the GNU General Public License.

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 (§ 2.6). Equilibrium statistical mechanics was then used to estimate free energies, as well as heat capacities, from the molecular partition function (§ 2.6). Vibrational frequencies were computed using normal mode analysis, and within the local rigid body framework these are evaluated for the generalised coordinates by including the appropriate metric tensor.[302] Additionally, we can adapt the normal mode analysis to scale favourably with system size, by utilising 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 anharmonic effects become more significant, corrections are needed. These can be added by employing methods such as the reaction path Hamiltonian superposition approach (RPHSA).[261] Nonetheless, a consistent use of the harmonic superposition approximation (HSA) here is sufficient for comparing the global thermodynamics within the various LRB schemes.

Depicting energy landscapes

Disconnectivity graphs[265,266] were used to visualise the energy landscapes (§ 2.8).

3.3 Results and Discussion

We begin by characterising 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.

3.3.1 Potential energy landscapes

Figure 3.4 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 PEL, and is coloured based on the value of two order parameters, L and S. The structural order parameter L, defined by Snow *et al.*[49] 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.[49] 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).

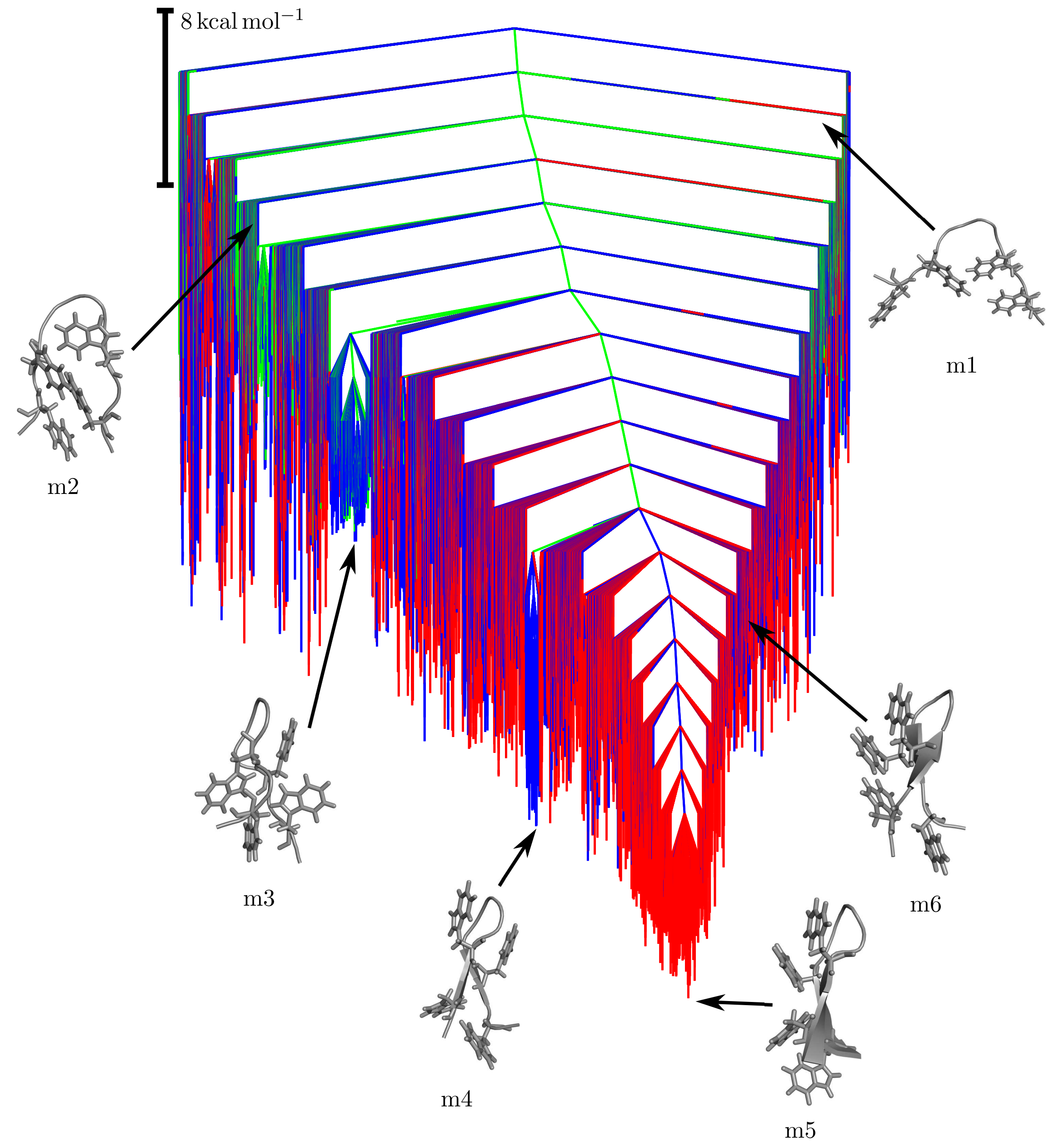
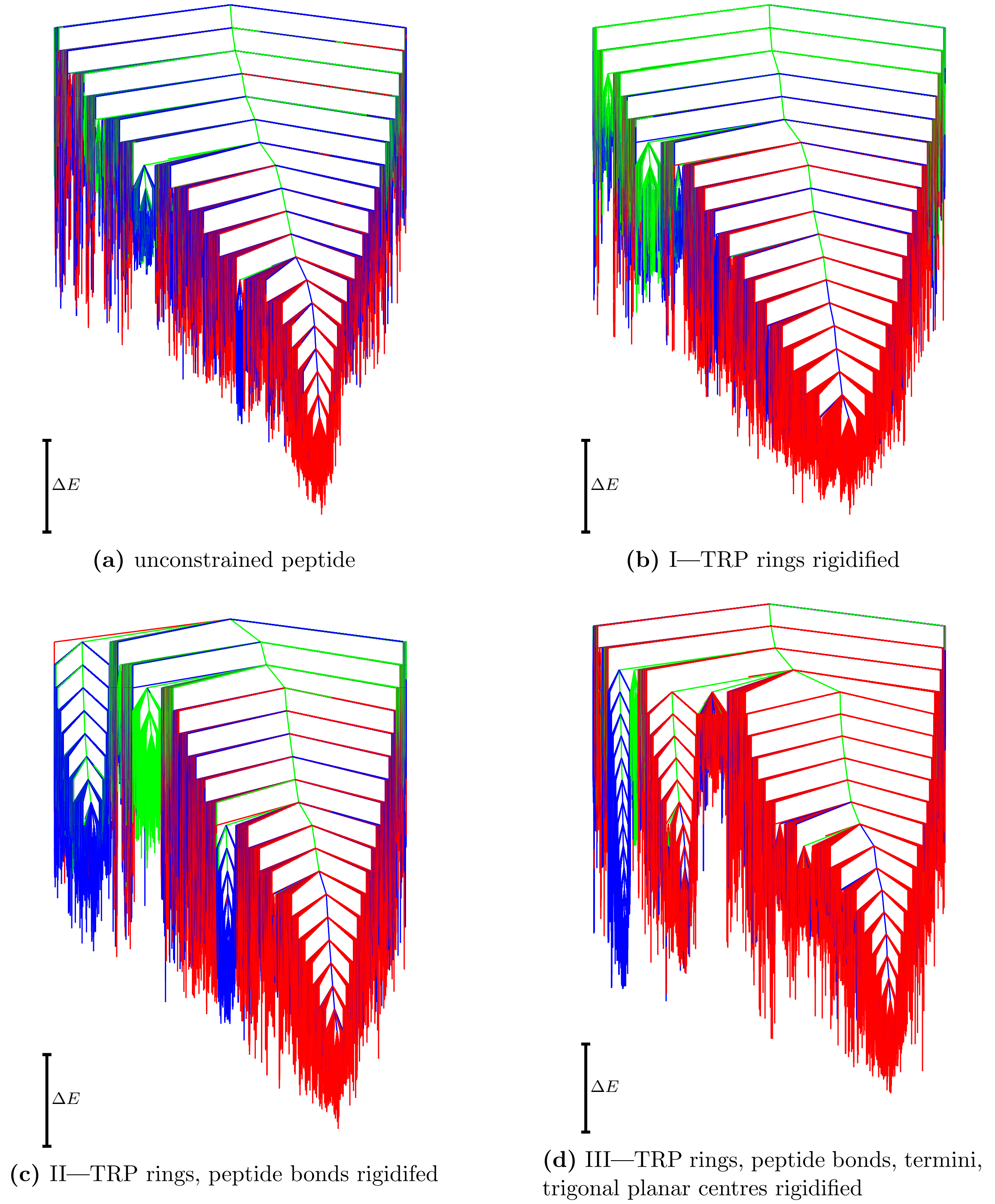


Figure 3.4: Potential energy disconnectivity graph for the unconstrained TZ1 peptide. The branches are coloured based on order parameters L (the sum of the four inner native hydrogen-bond lengths and the distances between the CD2 atoms of the three TRP pairs) and S (the orientation of the TRP rings—refer to text for description). The three main morphologies are: red denoted F1 ($L < 60$ Å, S value $= +1$), blue denoted F2 ($L < 60$ Å, S value $= -1$), green denoted F3 (all other minima).

The L and S values were together used to visualise the organisation 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 characterised as competing structures, which lead to topological frustration. Yang and Gruebele demonstrated that such structures act as kinetic traps,[323] 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 one KTN. Minima and transition states on the unconstrained landscape were also reoptimised 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 analysed using the same metrics as described in Figure 3.4.



(a) unconstrained peptide

(b) I—TRP rings rigidified

(c) II—TRP rings, peptide bonds rigidifed

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

Figure 3.5: Potential energy disconnectivity graphs for TZ1 at different levels of local rigidification. The branches are coloured based on order parameters L and S, as in Figure 3.4. The three main PE conformational groups are: red—F1 ($L < 60\,\text{Å}$, S value $= +1$), blue—F2 ($L < 60\,\text{Å}$, S value $= -1$), green—F3 (all other minima), as described in the text.

Comparing the disconnectivity graphs in Figure 3.5, depicting the PE landscapes of TZ1 from the unconstrained representation up to aggressive local rigidification, reveals several trends:

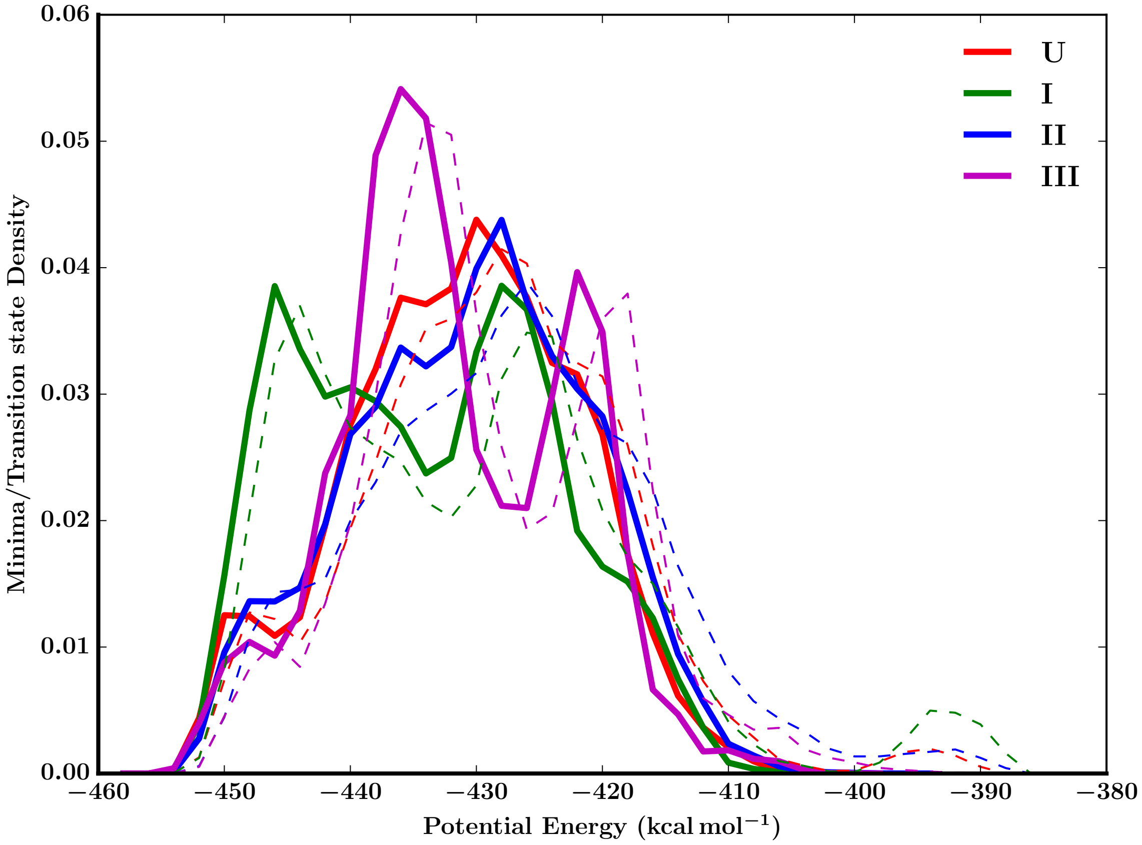
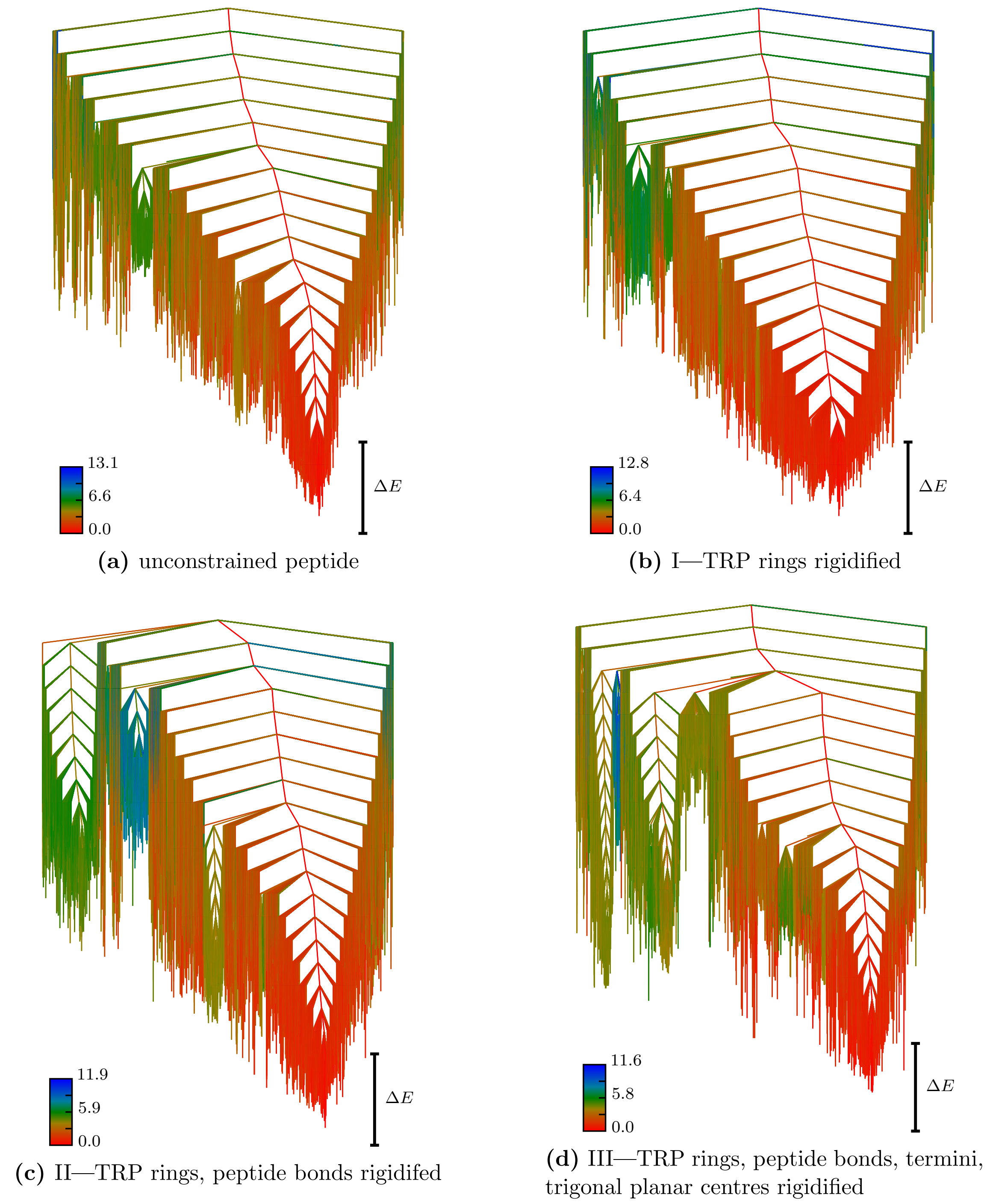


Figure 3.6: Distribution of the total energies of minima (solid lines) and transition states (dashed lines) on the potential energy landscapes of TZ1 at different levels of local rigidification: U—unconstrained (no local rigid bodies), I—TRP rings, II—TRP rings, peptide bonds, III—TRP rings, peptide bonds, termini, trigonal planar centres.

Potential energy range

The PE range for all four graphs is similar (Figure 3.6), with a difference of approximately 64 kcal mol^{-1} between the highest and lowest transition states. 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^{-1} respectively, while the corresponding transition states on the most rigidified landscape lie at −388.3 and −452.7 kcal mol^{-1}. 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^{-1}, while local minima on the PE landscape for schemes I, II and III cover a range of 51, 57 and 54 kcal mol^{-1}, respectively.



(a) unconstrained peptide

(b) I—TRP rings rigidified

(c) II—TRP rings, peptide bonds rigidifed

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

Figure 3.7: Potential energy disconnectivity graph for TZ1 ($\Delta E = 8\,\mathrm{kcal\,mol^{-1}}$) at different levels of local rigidification. The branches are coloured based on overall geometric root-mean-square deviation with reference to the global PE minimum.

Structural heterogeneity

A diverse collection of local minima, with varying geometric root-mean-square deviations from the global minimum, is identified in each scheme (Figure 3.7). The three PE groups identified for the unconstrained potential energy landscape are also present on the locally rigidified landscapes. Hence, we find that upon reoptimisation 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,[301] 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.

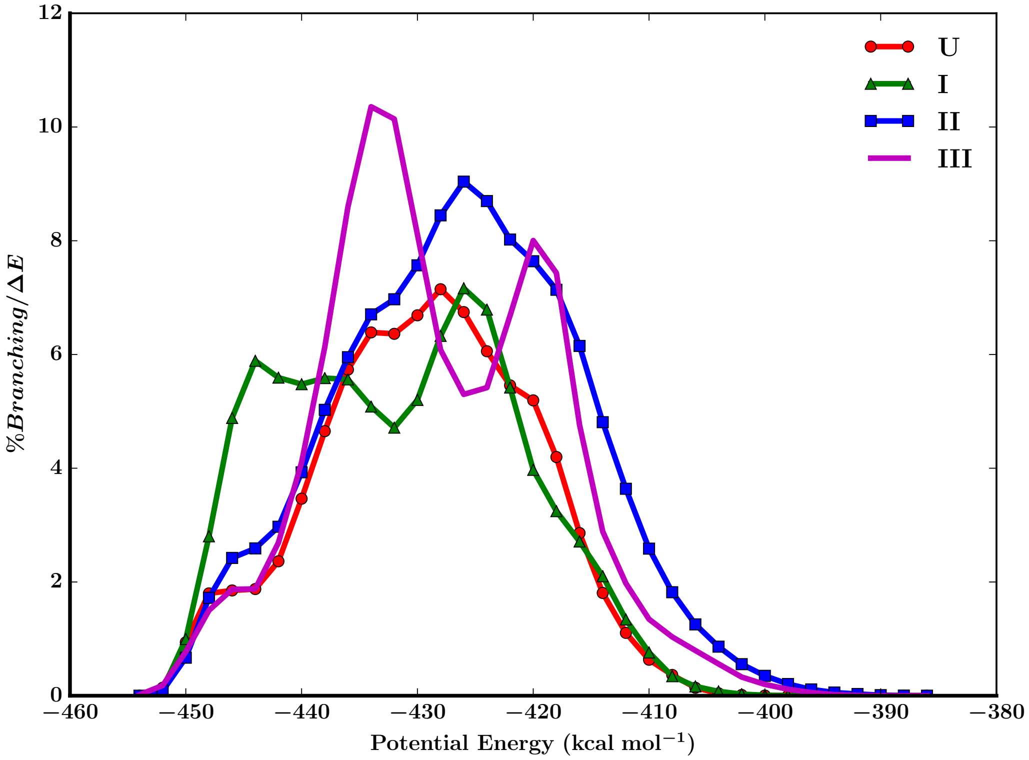


Figure 3.8: Roughness of the potential energy landscape of TZ1 corresponding to the unconstrained representation (U), and locally rigidified representations (I—TRP rings, II—TRP rings, peptide bonds, III—TRP rings, peptide bonds, termini, trigonal planar centres). Here, the landscape roughness is defined as the variation of the roughness density with energy; where the roughness density is taken as the quotient of the percentage of minima that branch off at a particular energy level and the energy threshold, ΔE, used for the superbasin analysis. In the plots $\Delta E = 2\,\text{kcal}\,\text{mol}^{-1}$.

Landscape roughness

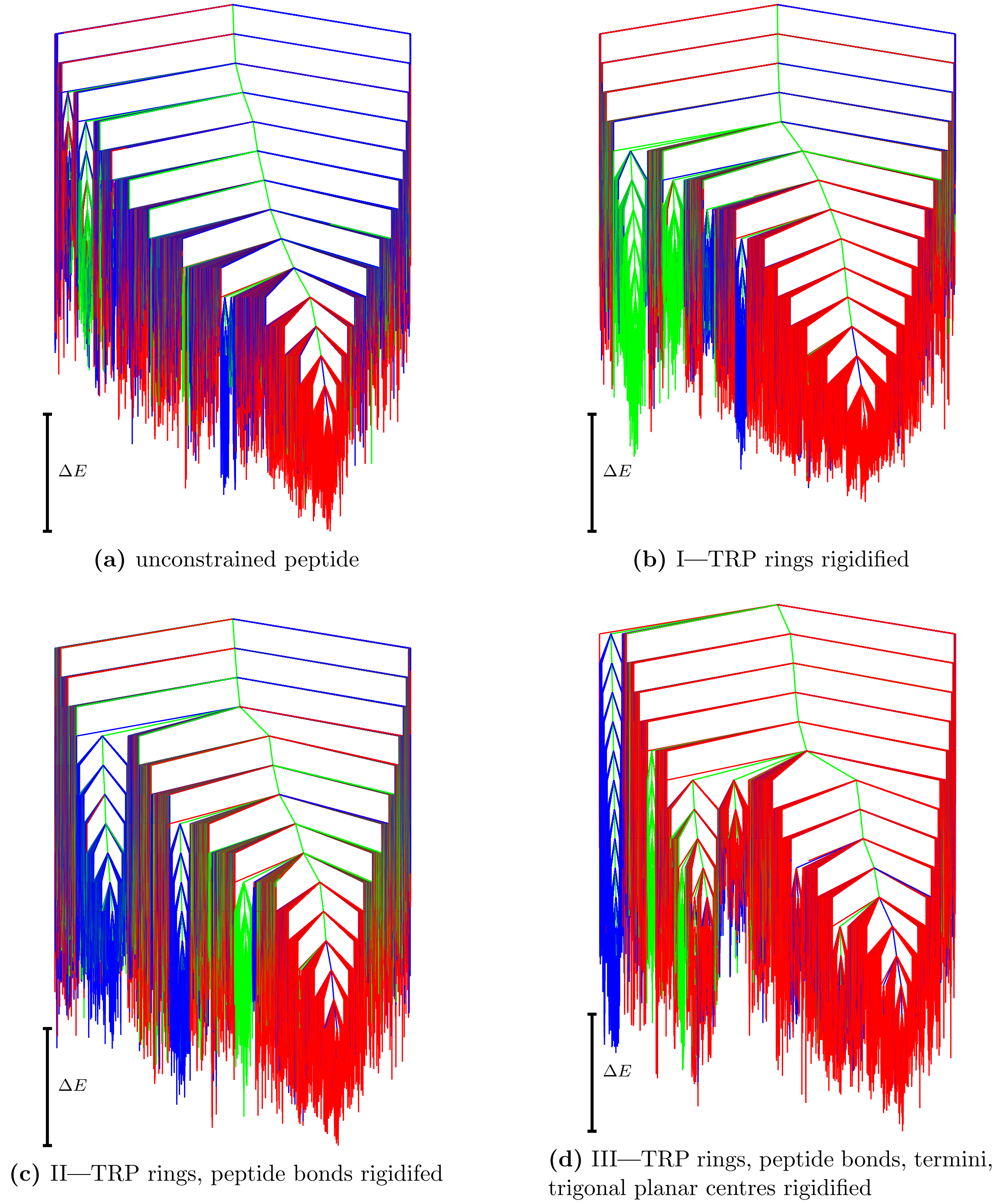
Levy and Becker presented an account of how disconnectivity graphs may be used to assess energy landscape roughness.[324] In their treatment, the roughness density is taken as the quotient of the percentage of minima that branch off a given energy level and the energy threshold used for the superbasin analysis. We computed this property for our disconnectivity graphs (Figure 3.8). On the unconstrained landscape and the locally rigidified landscapes corresponding to schemes I and II, the maximum roughness occurs around 30 kcal mol^{-1} above the global minimum. The overall landscape 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^{-1} above the global minimum are in F1). For scheme III, maximum landscape roughness occurs closer to the global minimum (about 20 kcal mol^{-1} 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 3.5d).

Additional trends

In addition to the potential energy landscapes depicted in Figure 3.5, landscapes were computed for several other local rigidification schemes (not shown), namely: IV—TRP rings, termini, V—peptide bonds, VI—TRP rings, trigonal planar centres, VII—TRP rings, peptide bonds, termini, VIII—peptide bonds, termini, trigonal planar centres, IX—entire side-chains. The trends observed for schemes I to III are also apparent when we include the additional schemes in our analysis; for example, on moving from scheme V $\rightarrow$ II $\rightarrow$ VII and from scheme VIII $\rightarrow$ III, the landscapes become evidently more frustrated, as larger sets of the protein are constrained. We also evaluated the effects of different types of local rigid bodies; for instance, comparison of schemes II, IV, and VI revealed that rigidification of trigonal planar centres results in the greatest increase in the landscape frustration, while rigidification of termini is least significant. The most aggressive local rigidification (i.e. scheme IX) failed to yield connected pathways in an efficient manner; thus, for TZ1 the flexibility of the side-chains is required for cooperative folding.



(a) unconstrained peptide

(b) I—TRP rings rigidified

(c) II—TRP rings, peptide bonds rigidifed

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

Figure 3.9: Free energy disconnectivity graph for TZ1 ($\Delta E = 8\,\mathrm{kcal\,mol^{-1}}$) computed at 298 K 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 orientation of the TRP rings—refer to text for description). The three main morphologies are: red—F1 ($L < 60$ Å, S value $= +1$), blue—F2 ($L < 60$ Å, S value $= -1$), green—F3 (all other minima).

3.3.2 Thermodynamics of folding

The free energy landscape (FEL),[267,268] 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 PELs, although there is some difference in the ordering of minima when entropy is considered (Figure 3.9). 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 3.10).

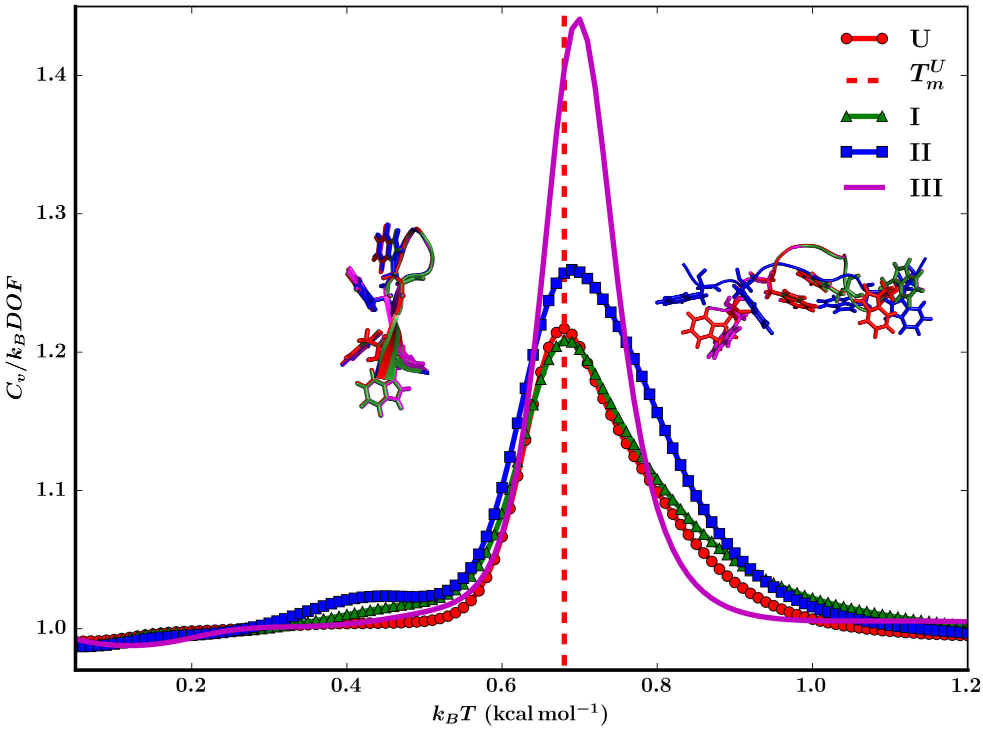


Figure 3.10: Constant volume heat capacity curves for TZ1 at various levels of local rigidification: unconstrained—no local rigid bodies; I—TRP rings, II—TRP rings, peptide bonds, III—TRP rings, peptide bonds, termini, trigonal planar centres treated as rigid bodies. The heat capacities are divided by 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 landscape, computed at low (0.48 kcal mol^{-1}) and high (0.88 kcal mol^{-1}) temperatures, are superimposed on the plot; Key: red (U), green (I), blue (II), magenta (III).

The melting temperature (T_m) is an important thermodynamic property for proteins, as it is often used as a 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 (Figure 3.11), which translates to preservation of the main basins of attraction and phase volumes on the energy landscape when local rigidification is applied.

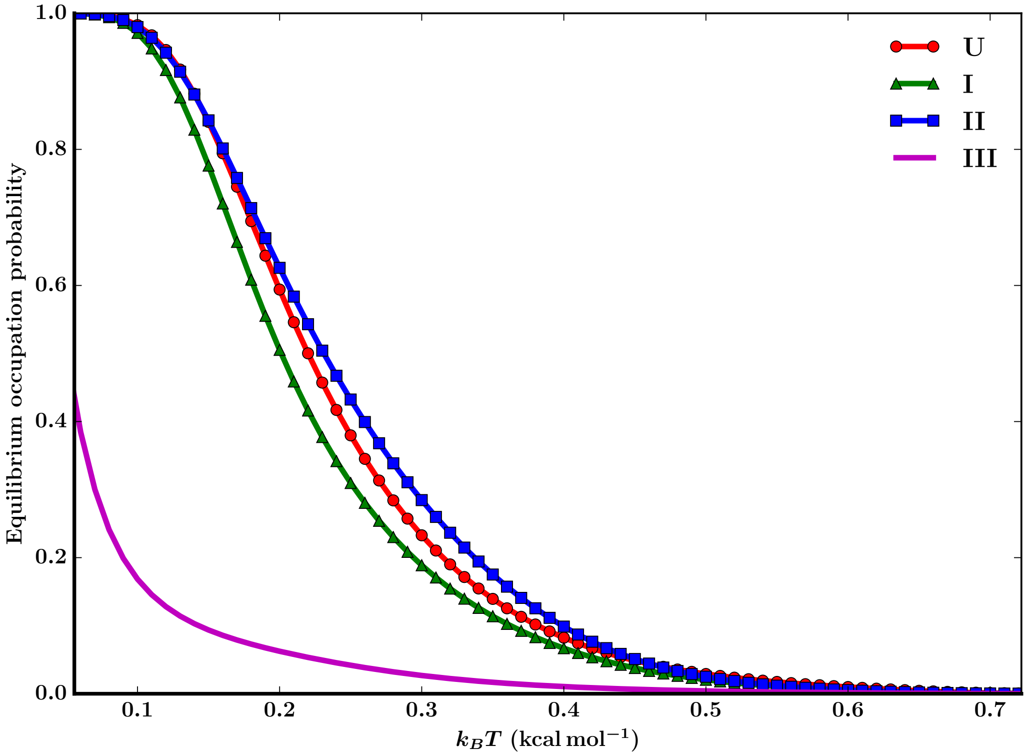
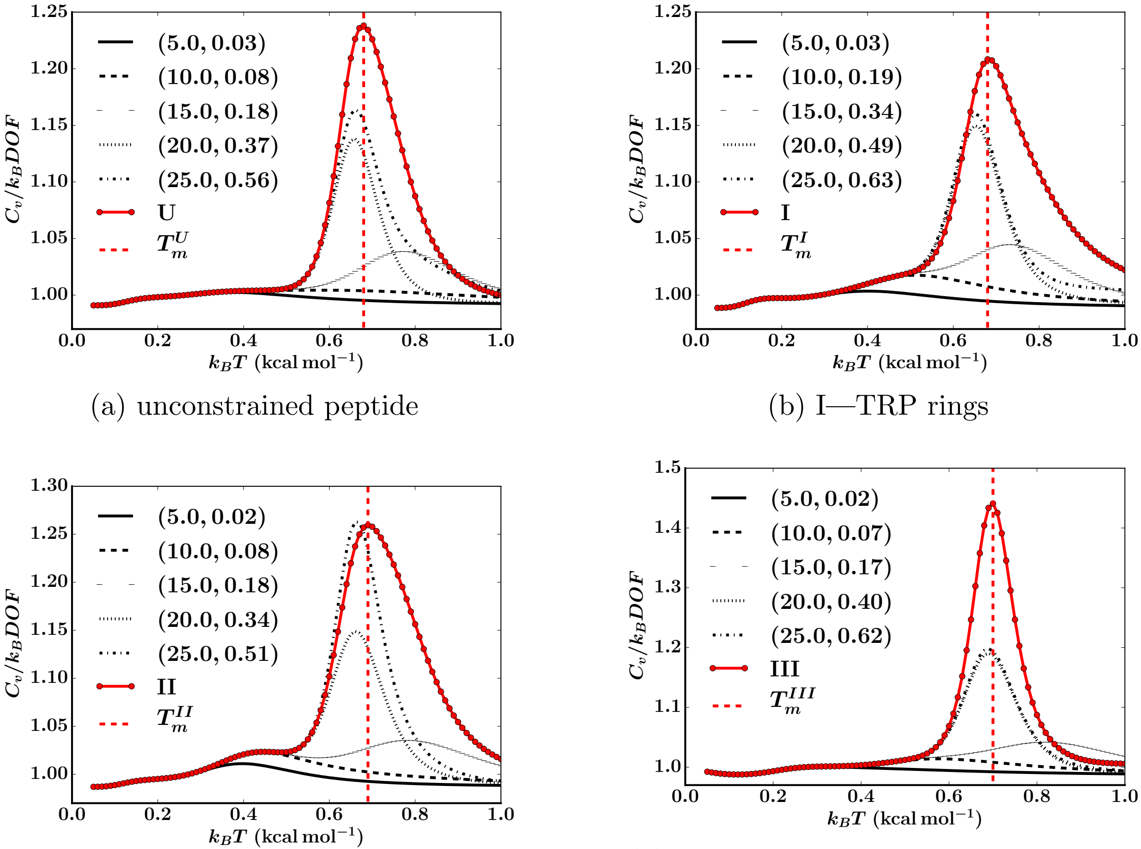


Figure 3.11: Variation in the equilibrium occupation probabilities of the PE global minimum at different levels of local rigidification: U—unconstrained (no local rigid bodies), I—TRP rings, II—TRP rings, peptide bonds, III—TRP rings, peptide bonds, termini, trigonal planar centres.

For the unconstrained peptide (red curve in Figure 3.10), the melting transition is calculated at a temperature equivalent to 0.68 kcal mol^{-1} (experimental value = 0.64 kcal mol^{-1}).[213] The heat capacity curve for scheme I is qualitatively similar to that of the unconstrained peptide, and the melting temperature is accurately predicted (Figure 3.10). A small positive offset in T_m from the reference value was observed for schemes II ($T_m = 0.69$ kcal mol^{-1}) and III ($T_m = 0.70$ kcal mol^{-1}). 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 stabilise 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 (Figure 3.10), implying that the important basins that govern the phase transition are retained.

In the unconstrained landscape and the landscapes corresponding to schemes I and II, the PE global minimum dominates the thermodynamic properties at low temperatures, Figure 3.11. However, at these temperatures the equilibrium occupation probability of the PE global minimum is notably lower in scheme III than in the other schemes; implying that other states make significant contributions to the thermodynamics as well.

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. The heat capacity curves were evaluated as a function of all the minima in the database lying below a given energy threshold (Figure 3.12). 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.



(a) unconstrained peptide

(b) I—TRP rings

(c) II—TRP rings, peptide bonds

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

Figure 3.12: Convergence of the heat capacity of TZ1 computed using the harmonic superposition approach. Each C_v curve is calculated for minima below a given energy threshold. The threshold and the corresponding fraction of minima are indicated in the legend. As a reference, the heat capacity curves for each scheme, as well as the position of the melting peak T_m^X, is also shown on the plots.

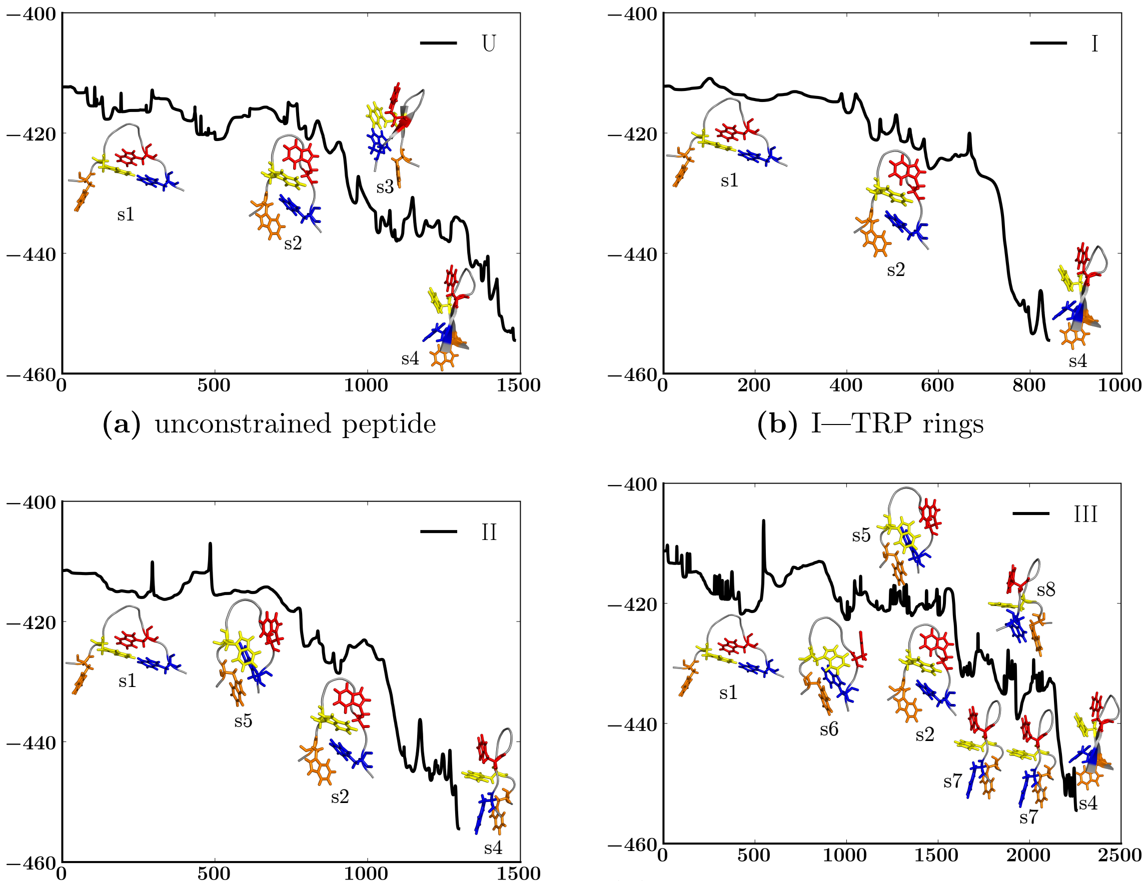
The global minimum of the FEL was computed for each local rigidification scheme at temperatures below and after the melting transition (Figure 3.10). At 0.48 kcal mol^{-1}, the overall geometric rmsd values of the FE global minimum for schemes I, II, II with respect to the unconstrained peptide are 0.47, 0.60, 0.67 Å, respectively. The corresponding deviations at 0.88 kcal mol^{-1} 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.

3.3.3 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 ($A \to B$) 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; see § 2.7.2).[249–251] The main conformational states encountered on each path were then identified by employing the density–based clustering algorithm[325] available within AMBER tools;[326] this approach essentially defines an average structure for different sections of the path.

Figure 3.13a 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 possesses 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[327] and follows the order of TZ folding events determined by temperature jump fluorescence .[49]

On the conservatively rigidified landscape (Figure 3.13b), 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 3.13c) 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 3.13: Variation of the total potential energy (kcal mol^{-1}) 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 3.13d), 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 3.5d, 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,[269] where the number of rearrangements on the fastest path from a given local minimum to the global PE minimum is computed. The distributions for the number of rearrangements can then be used to analyse 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 3.14), 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.

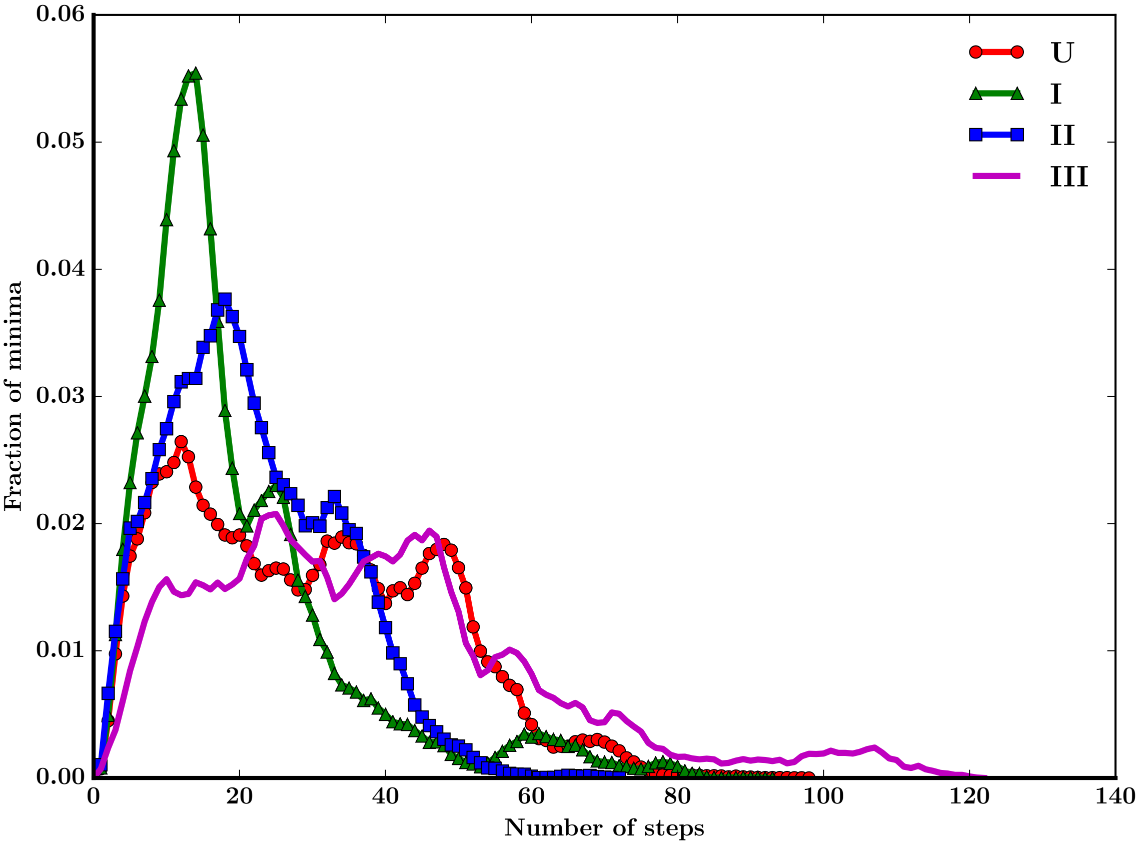


Figure 3.14: Distribution of the number of steps (transition states) on the fastest paths from a given minimum to the global minimum for TZ1 at different levels of local rigidification.

3.4 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 characterised 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.[213] For the unconstrained peptide, schemes I and II (TRP rings, peptide bonds), the folding mechanism corresponds to an initial hydrophobic collapse and subsequent zipping.[49,327] 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.[210] However, local rigidification reduces the conformational search space, by constraining degrees of freedom that fluctuate on a much faster time scale 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 was 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 time scale 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.

4

Energy Landscape for Fold-Switching in RfaH-CTD

4.1 Introduction

Over the last two decades there has been increasing evidence of fold-switching, where certain proteins are capable of adopting distinct, stable folds in a reversible fashion.[328–335] These proteins, commonly referred to as metamorphic,[332] extend the classical view of protein conformational dynamics, beyond movements of loop regions and side-chains, to large-scale rearrangements at the level of secondary structure. For example, human chemokine lymphotactin (Ltn) exists as two distinct conformations: a monomeric form (Ltn10), consisting of a three-stranded β-sheet and an α-helix, and an all-β-sheet dimeric form (Ltn40).[333] Under physiological conditions, the two conformers are in equilibrium, and bind to different molecular partners. Other well-known examples of metamorphic proteins include mitotic arrest deficiency 2 (Mad2) protein[329] and chloride intracellular channel 1 (CLIC1) protein.[328]

Perhaps the most dramatic example of protein conformational switching has been reported for RfaH (162 amino acids)[*].[336] RfaH is a regulatory protein found in *Escherichia coli* (*E.coli*) and *Salmonella*,[337] and is known to increase the expression of genes in operons containing an operon polarity suppressor (*ops*) site (a short, well-conserved DNA sequence)[†].[337–339] RfaH comprises two domains connected by a flexible linker: an N-terminal domain (NTD) and a C-terminal domain (CTD).[214] In the domain-closed state, the CTD adopts an α-helical hairpin fold, and binds tightly to the NTD.[214,215] When the transcribing RNA polymerase (RNAP) pauses at the

[*] RfaH is named after the *rfaH* gene that encodes it.

[†] Operon polarity encompasses the decreased expression of genes in an operon. Hence, an operon polarity suppressor site acts to counteract this effect.

ops site, interactions between RNAP, the *ops* site, and RfaH lead to domain separation.[215] At this stage, RfaH-NTD binds to RNAP in a clamp-like fashion, modifying RNAP into a pause-resistant state, and ensuring that synthesis of messenger RNA (mRNA) is complete without pausing or premature termination. Accordingly, the main purpose of the CTD in the domain-closed state is to mask the RNAP binding site of RfaH-NTD (a hydrophobic cavity);[215] thus, RfaH-CTD serves as a regulator of transcription, and effectively restricts RfaH to operons containing an *ops* site.

Upon domain separation, the CTD of RfaH undergoes a dramatic conformational transition: the α-helical hairpin refolds into a five-stranded β-barrel scaffold (i.e. an all-$\alpha \rightarrow$ all-β transition).[215] RfaH-CTD, in the β-barrel conformation, then binds to ribosomal protein S10, thereby recruiting the ribosomal 30S subunit to the nascent mRNA, significantly promoting translation.[215] Hence, for RfaH-CTD the same amino acid sequence folds into two distinct conformations with two distinct functions, constituting a special type of metamorphic system known as a transformer protein[340] (Figure 4.1).

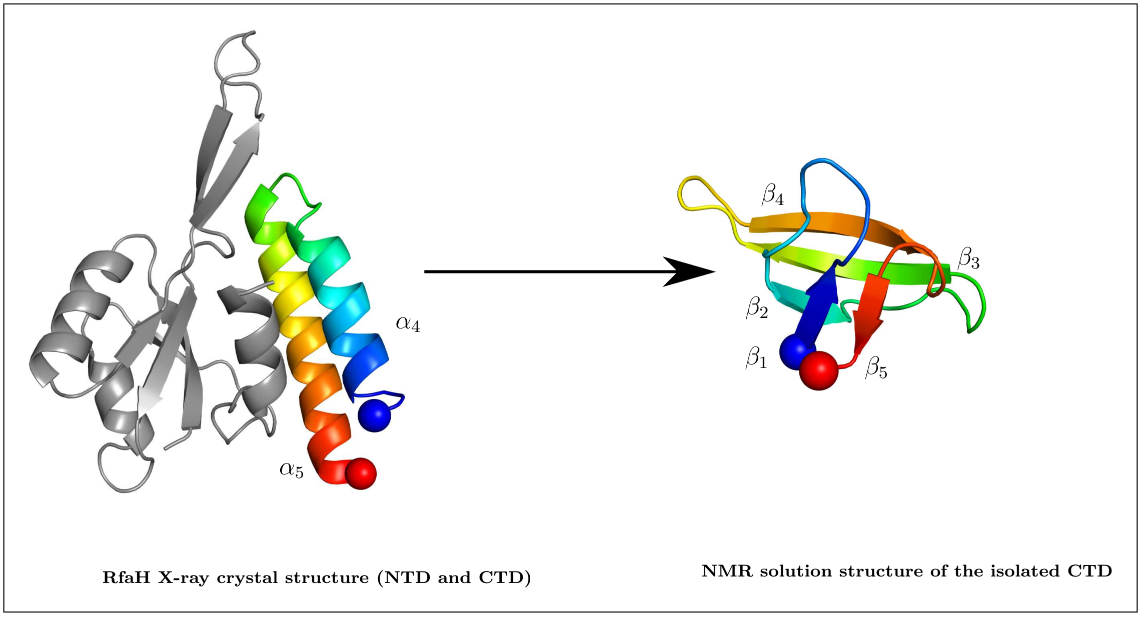

α_4
α_5
β_4
β_3
β_2
β_1
β_5
RfaH X-ray crystal structure (NTD and CTD)
NMR solution structure of the isolated CTD

Figure 4.1: X-ray crystallography structure of RfaH (residues 1–100 and 115–156), and NMR solution structure of the C-terminal domain of RfaH (residues 115–162). Upon domain separation, the CTD of RfaH transforms from an α-helical hairpin (α_4 = 115–130; α_5 = 135–155) to a five-stranded β-barrel scaffold: β_5 (158–160), β_1 (115–118), β_2 (127–130), β_3 (138–144) and β_4 (149–155). The N-terminus and C-terminus are highlighted with blue and red spheres, respectively.

The all-$\alpha \rightarrow$ all-β transition of RfaH-CTD is interesting for several reasons. Firstly, the genes in RfaH-regulated operons encode several bacterial virulence factors, including lipopolysaccharide (LPS) core, exopolysaccharide and haemolysin toxin, and the action of RfaH increases the expression of these factors, which are otherwise poorly transcribed (due to the large percentage of rare codons) and translated (due to a lack of canonical ribosomal recruitment sites). Hence, RfaH-CTD represents a good model for understanding gene regulation of these operons, which may be shared by other regulation factors. Secondly, the rules governing the refolding of RfaH-CTD may also be implicated in protein misfolding, so elucidating the mechanism for the large-scale structural transition of RfaH-CTD may aid in protein engineering and drug design.

While several experimental studies have been successful in characterising the domain-closed and domain-opened states of RfaH, the details of the refolding process have been inherently difficult to probe. NMR studies of the full-length RfaH are complicated by precipitation of the hydrophobic NTD once the protein dissociates from RNAP, or once domain dissociation is initiated *in vitro*.[215] Additionally, NMR shifts for the isolated CTD strictly mirror those of the β-barrel conformer,[215] and conversion back to the α-helical structure is not observed. This effect is largely due to the fact that contacts with the NTD are critical for inducing refolding to the α-helical state.[341] Therefore, several groups have implemented computer simulation techniques to analyse the refolding process.

Unfortunately, large-scale structural rearrangements generally occur on relatively long time scales, and so are difficult to simulate in an efficient manner via standard techniques. The refolding of RfaH-CTD has been probed using several computational approaches, including replica exchange molecular dynamics (REMD),[342] construction of Markov state models (MSMs),[343] and replica-exchange-with-tunnelling (RET).[344]

REMD has been used to investigate the refolding of the isolated RfaH-CTD in implicit solvent.[342] A free energy surface was constructed by projecting the replicas simulated at 310 K onto the root-mean-square deviation (rmsd) from the all-α state and the end-to-end distance. The structural transition was reported to proceed via a completely unfolded state, and the simulation yielded a relatively flat all-β-sheet structure compared to the barrel-like scaffold obtained in the NMR experiments.[215] Li *et al.*[343] constructed an MSM for RfaH-CTD from numerous MD trajectories. Based on the final MSM, they concluded that the conversion process could occur via heterogeneous routes, and postulated that the underlying energy landscape for refolding was ‘rough’, which we interpret in terms of competing low energy structures separated by high barriers.

Recently, Bernhardt and Hansmann applied RET to decipher the refolding mechanism for RfaH-CTD.[344] In RET,[345,346] replicas evolve in the microcanonical ensemble for a short period, and are then provisionally exchanged, while simultaneously rescaling their velocities to ensure that the total energy before and after exchange is invariant. The replicas are then allowed to evolve again at constant energy, and the final structures are accepted or rejected based on a Metropolis criterion. This procedure ultimately may lead to improved acceptance probabilities compared to the standard REMD procedure. Using RET, a significant free energy barrier (approximately 10 RT) separating the all-α and all-β states of RfaH-CTD was identified, and the transition was reported to occur via a disordered conformer.[344]

In the present work, the potential energy landscape (PEL) framework and kinetic transition network (KTN) analysis are combined to probe the refolding of RfaH-CTD. In particular, discrete path sampling (DPS)[249–251] is used to construct the PEL (which encompasses low-lying minima and the corresponding transition states that connect them) for the structural transition at full atomistic resolution. The free energy landscape (FEL) for RfaH-CTD is then derived from the PEL avoiding low-dimensional projections, and mechanistic details of the refolding process are outlined. We find that the free energy landscape of isolated RfaH-CTD at 310 K is multifunnelled. Consistent with previous NMR studies, the β-barrel state is more stable than the α-helical hairpin ensemble. The structural transition occurs via a compact coil-like intermediate, and complete loss α-helical character.

4.2 Methods

Preparation of initial structures

The crystal structure of RfaH (residues 1–100 and 115–156) and the NMR solution structure of the isolated C-terminal region of RfaH (residues 97–162) were obtained from the protein data bank via the PDB accession codes 2OUG[214] and 2LCL,[215] respectively. Residues 115–162 of the NMR structure were selected as the initial all-β conformer. The terminal six residues (157–162) were added to residues 115–156 of the crystal structure using PyMOL,[347] and the resulting structure represented the initial all-α conformer for the MD simulations.

Explicit solvent MD

The atomic interactions were modelled using the AMBER ff99SB-ILDN[227] parameter set. The initial all-α and all-β structures were first minimised in vacuum for 10000 steps. Each structure was then solvated using TIP3P water[307] in a truncated octahedron, with the box edges restricted to a minimum distance 10 Å from the protein. The solvated systems were then minimised for a further 10000 steps, and a restraining force of 100 kcal mol^{-1} Å^{-2} was applied to each protein structure. They were then heated from 0 to 300 K over 20 ps, with a weak restraint of 10 kcal mol^{-1} Å^{-2} on the protein molecule. The restraints were subsequently removed and each system was equilibrated in the NPT ensemble (pressure = 1 atm; temperature = 300 K) for 5 ns, followed by 2 ns of constant volume MD. Finally, a 300 ns production run was performed at 300 K in the canonical ensemble. For all MD runs the temperature was regulated using a Langevin thermostat with a collision frequency of 1 ps. All bonds involving hydrogen were constrained using SHAKE, permitting a time step of 2 fs. Structures were saved every 10 ps for further analysis.

Preliminary analysis of the MD trajectories revealed that the all-β conformer sampled the native basin throughout the simulation (with relatively small deviations from the NMR topology). The structure with the lowest energy was selected as the starting geometry for discrete path sampling (DPS). For the all-α conformer, significant structural fluctuations were observed on the simulation time scale. Hence, additional equilibration and production (50 ns) runs were conducted for the α-helical hairpin, using backbone dihedral angle restraints, based on the crystal structure (i.e. for residues 115–156). The lowest energy α-helical conformers from both sets of MD runs were chosen as endpoints for DPS.

Construction of the potential energy landscape with DPS

To improve the efficiency of DPS, a generalised Born implicit solvent, GB-Neck2,[238] was used, with a cutoff 25 Å for evaluation of the Born radius. A salt concentration of 0.1 M was maintained, and the ff99SB-ILDN force field was also properly symmetrised, using the method suggested by Małolepsza *et al.*[306]

DPS[249–251] was performed using the OPTIM[321] and PATHSAMPLE[322] programs, with a GPU interface for OPTIM to accelerate sampling.[348] Firstly, we obtained paths connecting the two α-helical conformers to the β-barrel scaffold (suggested by the MD simulations). Once two endpoints were chosen, a structural alignment was performed, which minimises the distance between the endpoints based on overall rotation, translation and permutation of identical atoms. The next step involves interpolation between the aligned configurations.

Since the conformational transition from the α-helical hairpin to the β-barrel scaffold is expected to be complex, RfaH-CTD is a good test system for the enhanced quasi-continuous interpolation (QCI)[349] scheme. Here, an auxiliary potential is used to derive a set of discrete images between two endpoints. The auxiliary potential contains constraint and repulsive terms for bonded and non-bonded atoms, respectively, and, in the latest scheme, sequence information from the AMBER topology file is employed. The new QCI routine also includes harmonic springs between images, and cis-trans peptide bond constraints. These improvements together minimise the likelihood of chain-crossings and cis-trans isomerism, which are undesirable consequences of (linear) interpolation techniques, especially for distant conformations.

The auxiliary potential is set up for the aligned endpoints, and discrete images between these two starting configurations are built by adding one atom (or residue) at a time. Before another atom (or residue) is added, the potential is minimised using an L-BFGS minimiser,[242,243] with a predefined root-mean-square (rms) gradient condition. This procedure is repeated until the full set of intermediate configurations is obtained for all atoms. The minimised images were then used to seed a double-nudged[252] elastic band[253,254] (DNEB; § 2.5.1) computation, which yields transition state guesses that are then tightly converged using hybrid eigenvector-following (HEF; § 2.5.1).[255,256]

For a given set of RfaH-CTD endpoints one QCI cycle was performed in the first connection attempt. DNEB–HEF cycles were then used for subsequent connection attempts. After each cycle, pairs of minima for connection were selected using a modified Dijkstra algorithm.[258] To locate an initial path, this process was performed in parallel using the PATHSAMPLE program. The number of minima pairs to connect per cycle was defined *a priori*. The 'best path' between the two main endpoints was then computed using the Dijkstra missing connection algorithm (§ 2.5.1).[258] Unconnected minima on the best path (see § 2.5.1) were then chosen for QCI–DNEB–HEF[‡]/DNEB–HEF computations. Once the connection runs for minima pairs were completed, the new minima and transition states were added to the existing database of stationary points.

To compute the best path we must keep track of all distances between minima in the database; it is therefore important that the stationary point database does not grow too quickly before an initial path is found. Accordingly, two key strategies were employed in this work: (i) For each unconnected pair, an individual best path was computed and connection attempts between minima separated by the largest gaps were prioritised. (ii) On completion of connection runs, only stationary points on the individual best paths were added to the main database. Connection cycles were repeated until an initial path was found.

The initial connected database was then refined using the SHORTCUT[257,258,260] and UNTRAP[260] procedures in PATHSAMPLE, and the progress was monitored by

[‡] QCI was only used if the minimal aligned distance between a given minima pair exceeded 50 Å.

checking for convergence of the $\alpha \rightarrow \beta$ rate constant and the heat capacity curve.

Derivation of the free energy landscape

The free energy landscape for RfaH was computed at 300 and 310 K using the harmonic superposition approximation (HSA).[261] A recursive regrouping procedure[264] was employed to cluster minima and transition states in the kinetic transition network (KTN) into free energy groups, based on a free energy threshold. The structural rearrangement pathways were computed for the regrouped KTN using Dijkstra's shortest path algorithm with suitable edge weights.[258]

Computation of structural order parameters

Secondary structure analysis was performed using the DSSP algorithm.[350] The mass-weighted geometric root-mean-square deviation (rmsd) from selected minima/free energy groups, and the radius of gyration (R_g) of free energy groups were computed using the CPPTRAJ program in the AMBER tools package.[326] The CPPTRAJ software was also used to compute the total number of hydrogen-bonds in the various RfaH-CTD conformational states, with hydrogen-bond distance and angle cutoffs of 3.5 Å and 150°, respectively.

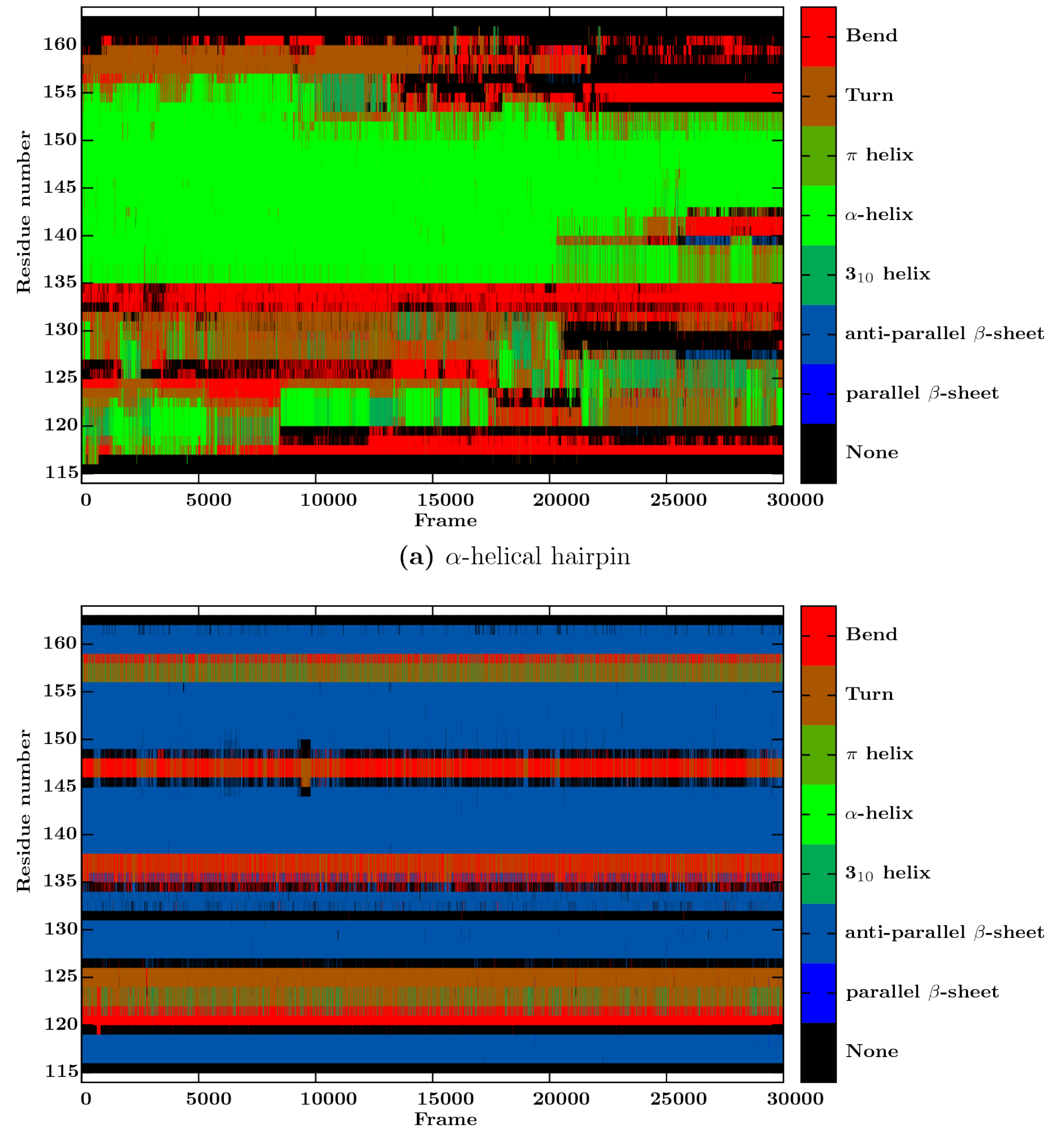
Depiction of energy landscapes

The computed potential and free energy landscapes were visualised using disconnectivity graphs.[265,266]

4.3 Results and Discussion

4.3.1 MD simulations for the α-helical and β-sheet conformers

In the domain-closed X-ray crystal structure the C-terminal domain of RfaH assumes an α-helical hairpin conformation with two antiparallel α-helices, and an intervening turn region. When domain separation is triggered, the CTD is known to refold to a β-barrel scaffold, with five antiparallel β-strands. Short molecular dynamics simulations (300 ns) were used to probe the short time stability of the two extreme RfaH-CTD forms.



(a) α-helical hairpin

(b) β-barrel

Figure 4.2: Secondary structure assignments for configurations along MD trajectories. The MD simulations (300 ns) for RfaH-CTD α-helical hairpin and β-barrel conformers were computed at 300 K in the NVT ensemble with explicit solvent.

The simulation initiated from the α-helical conformer (Figure 4.2a) shows that α_4 (residues 115–140; see Figure 4.1) has a higher propensity for helical unwinding than α_5 (residues 135–155); α_4 is partially unfolded throughout the entire production run, while α_5 maintains most of its α-helical structure. These findings agree well with previous work,[342,343] in which α_4 was reported to be less stable than α_5 for the isolated CTD. Several authors[341,351] suggest that interdomain contacts between the

NTD and the CTD are responsible for maintaining the stability of the α-helical form of RfaH-CTD, and when these contacts are disrupted the probability of forming the β-sheet analogue increases.

Figure 4.2b reveals that the β-barrel form of isolated RfaH-CTD is quite stable on the short MD simulation time scale. Throughout the MD run, the β-strands remain intact; with β_2 (127–130), β_3 (138–144) and β_4 (149–155) closely matching the NMR solution structure,[215] and β_1 (116–119) and β_5 (159–161) starting one residue later. An additional short β-strand (residues 132–133) was predicted between β_2 and β_3. A previous study also found that these residues had a tendency to adopt β-sheet structure, specifically, predicting that β_2 extended from residues 127 to 134.[342]

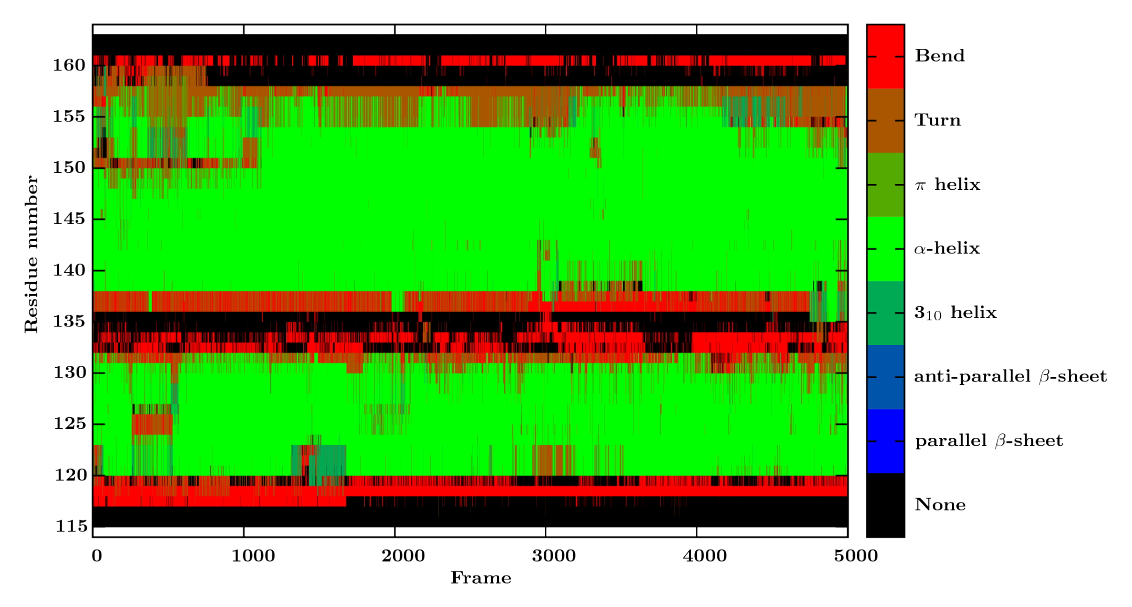


Figure 4.3: Secondary structure assignments for configurations along an MD trajectory. The MD simulation (50 ns) for RfaH-CTD α-helical hairpin was computed at 300 K in the NVT ensemble with explicit solvent. Backbone dihedral angle restraints for residues 115–156 were employed throughout.

The MD β-barrel and the partially unfolded α-helical structure are likely to be important conformers on the potential energy landscape for RfaH-CTD. They were therefore chosen as endpoints for discrete path sampling. However, since we are mainly interested in probing the refolding process from the α-helical hairpin form, we performed further structural refinement of the crystal structure with dihedral angle restraints (Figure 4.2). The refined structure was also used as an endpoint for DPS. These initial DPS endpoints closely resemble the structures depicted in Figure 4.5a.

4.3.2 Effects of QCI parameters on optimised paths

Computation of initial pathways between selected endpoints represents one of the main challenges in DPS. An interpolation procedure is first used to predict intervening structures between a given pair of conformers, which are then optimised to yield transition states and corresponding local minima, as described in § 4.2. For conformers close in configuration space, an initial linear interpolation scheme is generally sufficient; however, such schemes perform poorly for distant minima.

The quasi-continuous interpolation (QCI) scheme[349,352] has recently been shown to yield kinetically relevant paths for several large-scale rearrangements.[349,353] It allows the user to control several parameters; including the total number of images (i.e. intervening geometries; $N_{\text{max}}^{\text{images}}$), the cutoff distance for activating repulsive terms between unconstrained atoms (r_{rep}), the force constant for harmonic springs connecting images (k_{spr}), and the method used for growing the images (e.g. atom-by-atom, residue-by-residue), among others.

QCI parameters	**Int-I**	**Int-II**
r_{rep} (Å)	8.0	6.0
$k_{\text{spr}}^{\text{images}}$	10.0	10.0
$N_{\text{max}}^{\text{images}}$	200	50
method	add residue	add residue

Table 4.1: Comparison of selected QCI parameters for two different interpolations.

Table 4.1 compares some QCI parameters for two different interpolations from the RfaH-CTD lowest MD α conformer to the lowest MD β structure. In the first interpolation (Int-I) more images were used ($N_{\text{max}}^{\text{images}} = 200$) and a slightly larger repulsive cutoff distance was employed ($r_{\text{rep}} = 8.0$) than in Int-II. In both schemes the same value was set for the spring force constant, and images were constructed by adding one residue at a time.

The resulting optimised initial path corresponding to each QCI interpolation scheme is depicted in Figure 4.4. Int-II leads to a significantly shorter path connecting the α and β conformers than the final path obtained using Int-I. In the latter case, the protein becomes kinetically trapped, over about 1000 steps, before finally folding downhill towards the β-sheet structure. In this case, it seems that having a large number of images is actually less efficient. Interestingly, when the two paths were merged into one KTN, the longer path was no longer kinetically competitive. Hence, it is beneficial to include initial paths corresponding to different QCI interpolations (for the same set of endpoints) in the KTN, to increase the likelihood of finding the most biologically relevant paths.

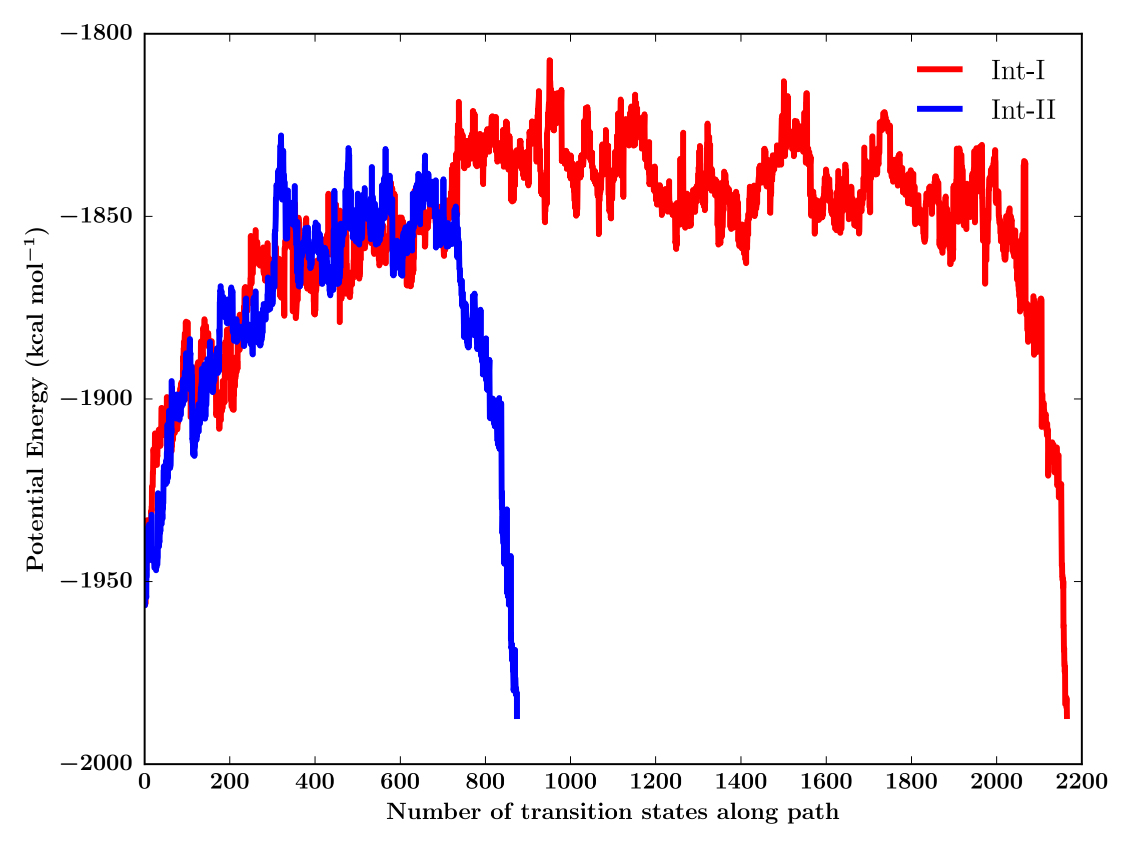
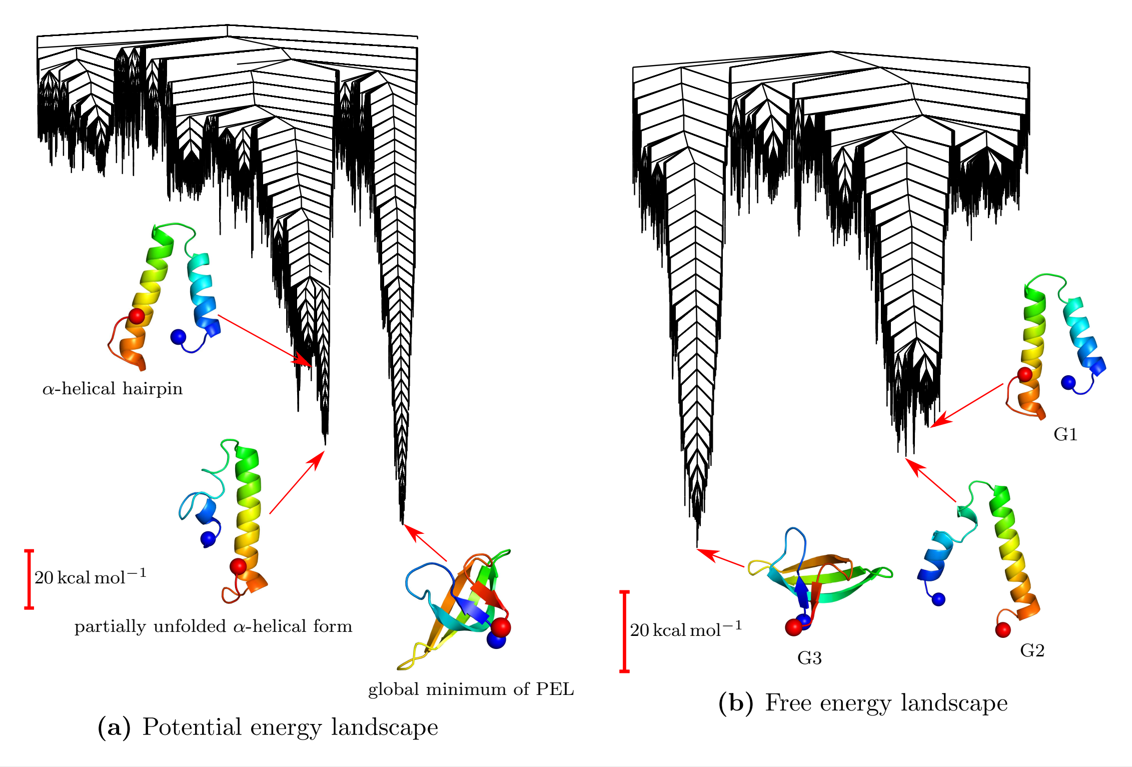


Figure 4.4: Optimised initial paths corresponding to two different QCI interpolations from the lowest-α to lowest-β conformer.

4.3.3 Potential and free energy landscapes

The PEL for the isolated RfaH-CTD is shown in Figure 4.5a; there are two prominent deep funnels. The major funnel, which includes the all-β conformer, is notably lower in energy than the one corresponding to the all-α structure, and contains the global minimum. The partially unfolded α-helical conformer is enthalpically more favourable than the α-helical hairpin form.



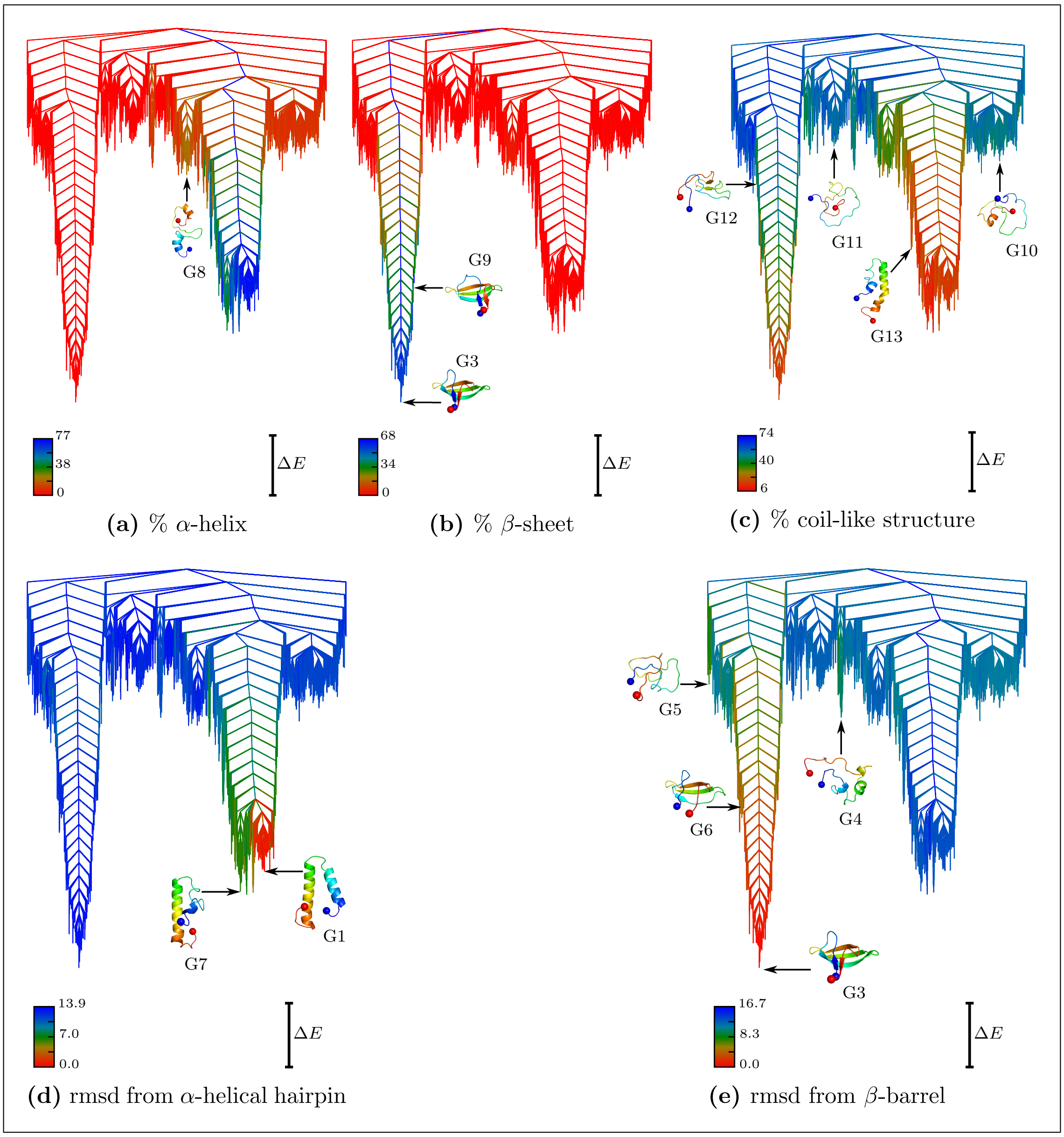
(a) Potential energy landscape

(b) Free energy landscape

Figure 4.5: Disconnectivity graphs for the isolated RfaH-CTD, in terms of (a) potential and (b) free energies. In (a) the lowest energy α-helical conformer (partially unfolded), the α-helical conformer with maximum helical content (hairpin), and the global minimum of the PEL are all superimposed on the graph. The free energies were computed at 310 K with minima and transition states regrouped based on an energy threshold of 5 kcal mol^{-1}. Representative structures for selected free energy groups (G1 to G3) are also shown.

The FEL was computed from the PEL at 300 K (not shown) and 310 K (Figure 4.5b). These two temperatures were chosen to allow for direct comparison with previous simulations (MSM construction at 300 K;[343] replica exchange approaches at 310 K[342,344]) and the original NMR experiment (at 310 K).[215] There was no significant difference between the two the landscapes, and so further analysis refers to the FEL at 310 K. Each branch on the free energy disconnectivity graph corresponds to a free energy group. The topology of the global free energy minimum, G3, is consistent with the NMR solution structure for the isolated RfaH-CTD (all-atom geometric rmsd = 1.57 Å); however, β_2 is visibly longer compared to the experimental structure. From the FEL, it is evident that the β-barrel scaffold is the most stable conformer for the isolated CTD of RfaH. In addition, the partially unfolded α-helical state, G2, is slightly more stable than the analogue with both helices intact, G1. Combined with the MD results, these results suggest that upon domain separation α_4 quickly loses some of its helical character, and G2 is an important intermediate in the refolding process of RfaH-CTD.

Since the barriers on the FEL for the isolated domain are particularly high, we infer that, in the absence of the appropriate molecular partner, the refolding process is likely to be slow. In fact, Burmann *et al.* probed the refolding process, by engineering a cleavage site into the linker region between the two domains, and reported that β-sheet structure was only detected 42 hours after incubation.[215]



(a) % α-helix

(b) % β-sheet

(c) % coil-like structure

(d) rmsd from α-helical hairpin

(e) rmsd from β-barrel

Figure 4.6: Free energy disconnectivity graphs ($\Delta E = 20\,\text{kcal}\,\text{mol}^{-1}$) for RfaH-CTD computed at 310 K with a regrouping threshold of $5\,\text{kcal}\,\text{mol}^{-1}$. The landscape is reproduced for several structural order parameters, and representative structures for selected free energy groups are highlighted.

4.3.4 Conformational states on the FEL

To gain better insight into the various conformational states on the FEL, the free energy disconnectivity graph was coloured based on several structural order parameters. Secondary structure analysis was performed for each free energy group, and these results are summarised in Figures 4.6a–c. Considerable variation in α-helical and β-sheet content is observed in Figures 4.6a and b. The G1 ensemble displays about 77% α-helical content, while ensembles in the high energy regions of the FEL and in the neighbourhood of the global FE minimum (G3) generally show negligible α-helical character. Maximum β-sheet content was observed for G3 (68%), while ensembles in the intermediate regions of the FEL contain some degree of α-helical or β-sheet content. Significant coil–like structure (i.e. lack of regular secondary structure) was observed for many ensembles in the high energy region of the landscape (e.g. G11 in Figure 4.6c).

The free energy disconnectivity graphs are also depicted in terms of the all-atom geometric rmsd from G1 (Figure 4.6d) and G3 (Figure 4.6e). These graphs further highlight the inherent structural heterogeneity of the states on the FEL. The principal funnel corresponding to ensembles with high α-helical content separates into two main sub-funnels: ensembles closely related to the hairpin state (G1) and those with α_4 partially unfolded (e.g. G7). The ensembles gradually deviate from G1 on traversing the landscape towards G3. A similar trend is observed from G3 towards G1.

Based on these results, we infer that on moving from the α-helical hairpin ensemble to the β-barrel state, RfaH-CTD gradually loses α-helical character, and the structural conversion occurs via an essentially unstructured intermediate.

4.3.5 Mechanism for fold-switching in RfaH-CTD

A detailed description of the refolding process can be obtained by examining the pathway between the all-α and all-β conformations that makes the largest contribution to the rate constant. For RfaH-CTD, the two forms were again defined as states by lumping stationary points into free energy groups, using recursive regrouping[264] with an energy threshold of 11 kcal mol^{-1}.[§] The fastest pathway between selected states was then extracted by employing Dijkstra's shortest path algorithm on the clustered stationary point database with appropriate edge weights.[258]

[§] Regrouping thresholds for which the rate constant is converged give constant results. However, if the threshold is too small excessive detail may be retained, and analysing the mechanism may prove difficult.

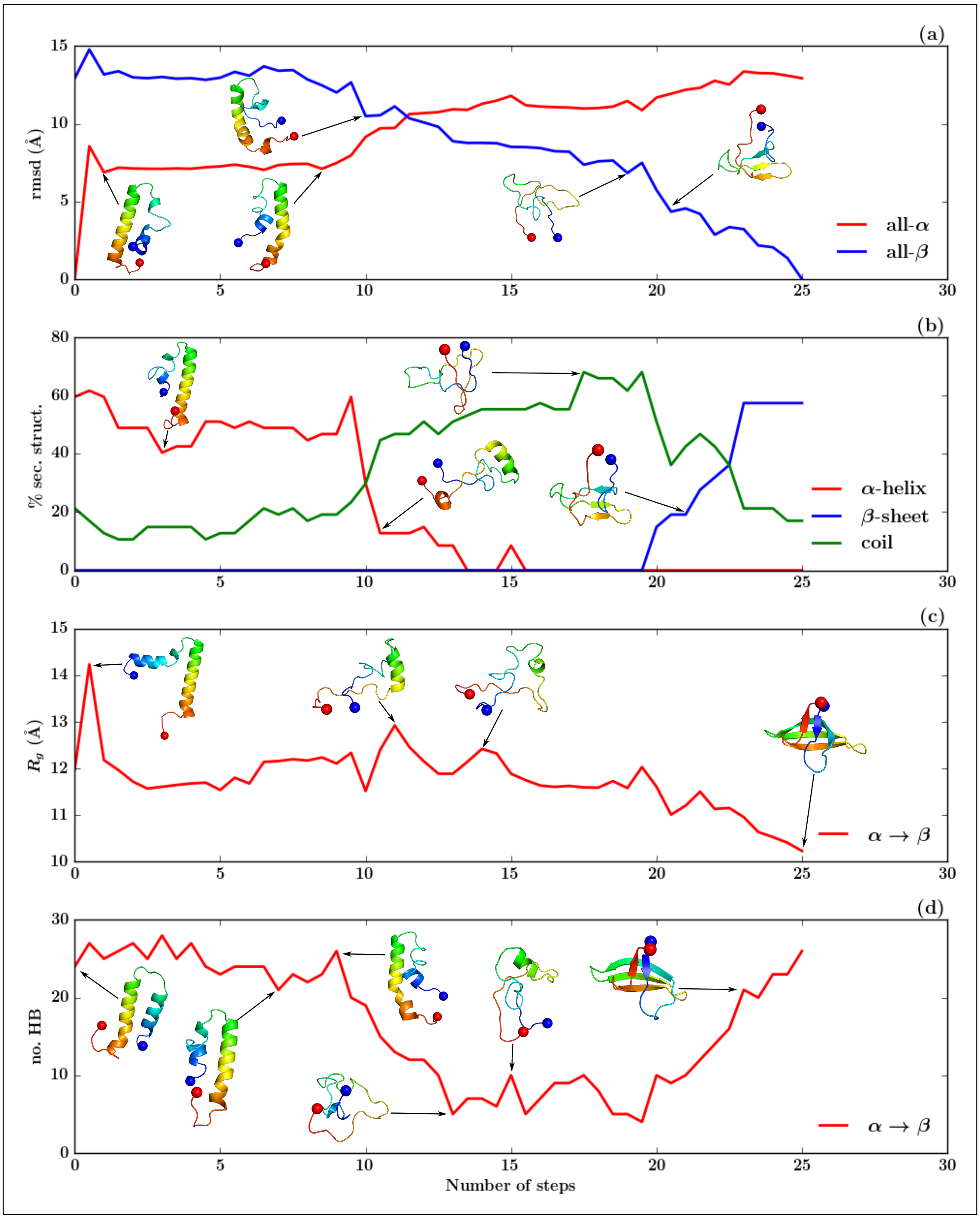


Figure 4.7: Evolution of selected order parameters on the fastest folding pathway from the α-helical hairpin to β-barrel state of RfaH-CTD: (a) all-atom geometric rmsd from all-α or all-β state; (b) secondary structure content; (c) radius of gyration; (d) total number of hydrogen-bonds. Representative structures for some states along the path are shown. The number of steps corresponds to the number of transition states along the path.

Figure 4.7 shows the variation in several structural order parameters along the fastest pathway. A significant deviation from the initial α-helical hairpin coincides with helical unwinding of α_4 (Figure 4.7a); GLU124 to THR131 unfolds and a short turn develops (GLN127 to ALA128). Two other groups also reported that unwinding of α_4 marked the first stage of the structural transition.[342,343] Geometric rms deviations from the all-α state oscillate at around 7 Å for about eight steps; α_4 continues to shorten, while α_5 generally remains intact. The protein then passes through an 'unstructured' intermediate (at step 19; Figure 4.7a), and then the configurations progressively become more β-like. Li *et al.* also observed a high population of compact coil-like states in their MSM for RfaH-CTD.[343]

The α-helical content decreases sharply at step ten of the folding transition (Figure 4.7b). At this stage, a transition state develops with low helical content in α_5, only maintaining helical structure from ALA137 to LEU142. From steps 14 to 19 the protein contains negligible α-helix or β-sheet structure; in that part of path states display maximum coil-like structure (notice the green curve in Figure 4.7b). The protein only adopts β-sheet-like structure in the latter segment of the path, as the canonical β-strands begin to form.

The radius of gyration (R_g), which was taken as the average mass-weighted squared distances of all atoms from protein centre of mass, is another useful order parameter for monitoring structural changes during the rearrangement process. For most of the refolding process R_g is about 12 Å, suggesting that the protein remains relatively compact during the transition (Figure 4.7c). Notably, in the early stages of folding, a significant increase in R_g is observed; a transition state forms with the two α-helices orientated roughly orthogonal to each other. This state is strikingly similar to the one located on the free energy surface of RfaH-CTD by GC *et al.*[342] On further investigation, it seems that this state forms due to breakage of a hydrogen-bond between THR119:HG1 (in α_4) and GLU149:O (in α_5), which causes the two helices to separate. However, new hydrogen-bonds are formed; for example, a short turn simultaneously forms in α_4, perhaps to accommodate the nearby bulky phenylalanine residues (PHE126, PHE130).

The variation in the number of hydrogen-bonds along the path was also examined. Figure 4.7d reveals that the protein does maintain some degree of hydrogen-bonding throughout refolding. For instance, between steps 13 to 19 (when the coil-like character is at a maximum) there is still some hydrogen-bonding due to turns (e.g. PRO133 to ASP134) and 3_{10} helices (e.g. GLY125 to GLN127). The intermediates in that region also contain a significant amount of bends (loops), which lead to compact morphologies. The hydrogen-bond pattern then increases steadily as the β-strands nucleate to yield the all-β state.

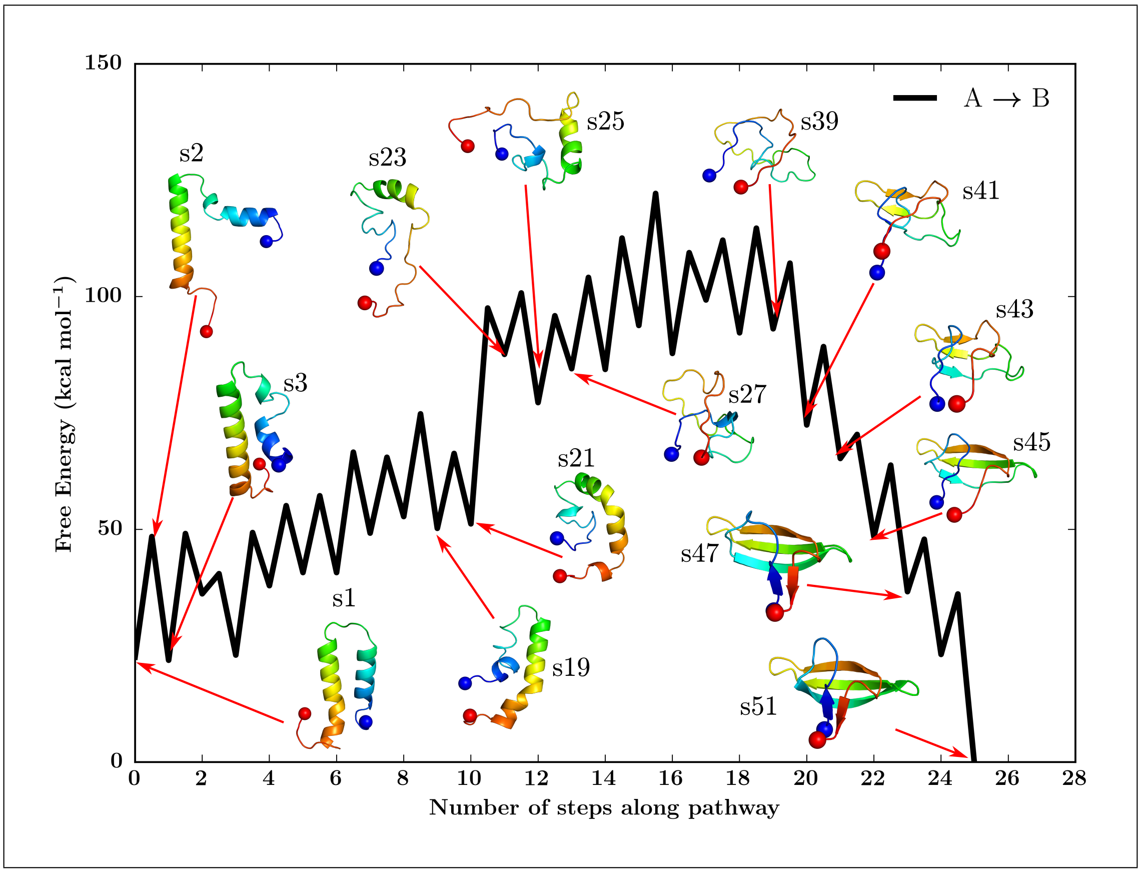


Figure 4.8: Pathway for the α-helical hairpin $\rightarrow$ β-barrel structural rearrangement of RfaH-CTD. Stationary points in the kinetic transition network were regrouped based on a threshold of 11 kcal mol^{-1}. Representative structures of selected states are superimposed on the path. States are numbered based on their positions along the path: s1 corresponds to the α-helical hairpin ensemble, and s51 represents the β-barrel state.

Finally, the refolding pathway of RfaH-CTD is presented in Figure 4.8 in terms of free energies. The structural transition occurs in three main stages:

1. The formation of a kink (short turn) in the neighbourhood of the bulky phenylalanine residues initiates the refolding process (s3). α_4 gradually shortens in the direction of the N-terminus (s19). Loss of α-helical character in α_4 then accommodates expansion of α_5 (s21)—starting from the C-terminus.

2. Helical unwinding eventually leads to the formation of a compact intermediate (s23), which includes residual α-helical character (ALA137 to LEU142). The formation of this intermediate is preceded by a major free energy barrier in the refolding process, which may therefore be classified as rate-limiting. Once unfolding of α_5 is complete, the C-terminal part of the protein crosses over the N-terminus, yielding a compact coil-like state (s27). Small conformational

changes lead to the formation of s39, which exhibits a β-barrel-like topology, with complete loss of α-helical character.

3. Once s39 forms, nucleation of the β-strands commences. β_3 (LEU142 to ASN144) and β_4 (GLU149 to LYS151) begin to develop first (s41), followed by nucleation of β_2 (s43). Strands 1 and 5 form last—completing the β-scaffold (s51).

4.4 Conclusions

Large-scale conformational changes in proteins are relatively difficult to probe. Such structural transformations may lead to the exposure of hydrophobic residues, resulting in aggregation *in vitro*, impeding experimental characterisation. Additionally, one metamorphic partner may be more stable than the other, and so probing the reverse process, at physically relevant temperatures, may be a challenge.

Computer simulations of fold-switching can therefore play an important role in improving our understanding of these processes, and aid in the design of novel protein-based architectures. However, *in silico* studies of systems undergoing large-scale changes have their own challenges. In particular, these processes often occur on long time scales, and the morphologies of interest may be separated by substantial free energy barriers. To circumvent these issues various sampling and data analysis strategies have been adopted.

In the present work, methods based on geometry optimisation were employed to characterise energy landscapes for the all-α to all-β transition of RfaH-CTD. The new quasi-continuous interpolation scheme[349] was employed to obtain initial guesses for putative structures on the refolding path; together with other discrete path sampling strategies, a kinetic transition network for RfaH-CTD was constructed consisting of stationary points on the potential energy landscape.

The free energy landscape for RfaH-CTD was computed at 310 K within the harmonic superposition approximation. The landscape is characteristically multifunnelled, and, consistent with experiment,[215] the β-barrel scaffold is the favoured conformer. The proposed mechanism for the structural transition is in good agreement with previous work,[343] and some of the important structural ensembles identified in this study have been found in REMD simulations[342] and in MSM constructions.[343] New details for the refolding process have been provided in the present work.

The ability of our approach to preserve the full atomistic resolution should aid in deriving design principles for protein fold–switching. It would therefore be of interest to extend this work to other transformer proteins (as they become available) and related systems.

5

Intrinsically Disordered Landscapes for Human CD4 Receptor Peptide

5.1 Introduction

In the mid-1990's, the suggestion that many functional proteins were natively disordered was met with much scepticism.[354,355] This notion was in direct conflict with the linear sequence–structure–function paradigm that had guided molecular biology for over half a century. Since then, overwhelming evidence from experiment[356–360] and bioinformatics surveys of entire genome sequences[361,362] have driven a paradigm shift, and the abundance and importance of intrinsically disordered proteins (IDPs) or regions (IDRs) are now widely recognised. It has been estimated that about 50% of mammalian proteins contain contiguous disordered regions (>30 residues long) and about 25% are fully disordered.[363] Disordered sequences are generally characterised by a low content of bulky hydrophobic amino acids, a high proportion of polar and charged amino acids, and an overall low complexity*.[357,364,365] Due to their composition, IDPs usually do not collapse to form hydrophobic cores, or fold into stable three-dimensional structures. [354,360] Instead, under physiological conditions, they display a high degree of conformational flexibility, in which several states rapidly interconvert.[366,367] Accordingly, IDPs exhibit diverse functional modes: in some cases, their functions may depend directly on the disordered state,[358,368] while, in other instances, they undergo induced folding upon interaction with a molecular partner.[356,369,370]

These characteristics confer several functional advantages to IDPs, such as the

[*] Low complexity sequences contain multiple repeats of single amino acids or amino acid motifs.

ability to regulate distance and orientation of protein domains, bind to targets with high specificity and low affinity, interact with multiple targets, and undergo post-translational modifications. IDPs therefore play a central role in many fundamental processes, including transcription regulation, translation, cell signalling and phosphorylation.[360,368,371,372] On the other hand, intrinsic disorder in proteins "can also have a biological cost in terms of the promotion and proliferation of protein folding diseases".[360] Specifically, the altered expression of IDPs has been linked to several diseases,[373] including cancer, diabetes, neurodegenerative diseases, and cardiovascular diseases.[374–379] Additionally, structural disorder has been associated with the action of pathogens;[354] for example, certain viruses may mimic IDRs and interfere with their regulation inside the cell.[380]

Thus, characterisation of IDPs is an important research area, to elucidate their functions and to identify potential therapeutic targets. Experimental characterisation of IDPs via conventional techniques has proved to be challenging, largely due to the fact that these methods were originally optimised for folded proteins. In particular, due to their dynamic nature, the electron density of IDRs is often absent from X-ray diffraction maps.[358] Additionally, in standard NMR studies, IDPs usually form aggregates at the required experimental concentrations, interconversion of conformational states leads to line-broadening effects, and peaks corresponding to disordered proteins are often poorly dispersed (see also § 1.3.2).[381] Techniques such as near and far-UV circular dichroism (CD) have been employed to distinguish between folded and disordered proteins. Since intrinsic disorder is commonly confined to a small region of the protein, CD methods, which lack residue-specific information, need to be combined with other techniques to properly characterise IDRs.[358] Mutidimensional NMR approaches,[85,381] SAXS,[83,382] single-molecule FRET[383] and AFM[384] have all proved useful in providing structural and dynamical information for disordered proteins. In one study, multidimensional NMR was used in conjunction with isotopic labelling to probe IDPs *in vivo*.[385] Other hybrid approaches, such as protease digestion with mass spectrometry,[386] have also been employed to identify IDPs, albeit less effectively than NMR-based techniques.

Molecular simulations, which can provide high resolution structural and kinetic information for IDP ensembles, have complemented experimental studies.[387–390] In particular, these simulations can provide predictions for experiment, or can be used in conjunction with experiments to aid interpretation.[391,392] The quality of such predictions and interpretations depends primarily on the accuracy of protein force fields and on the efficiency of the sampling strategy employed. As for conventional experimental techniques, protein force fields were originally developed for well-folded globular proteins; thus, they may display secondary structure biases.[226,393] Hence, is its necessary to develop force fields that can achieve a good balance among secondary structures. To this end, several standard force fields have been modified and benchmarked for IDPs.[227,394–400] While most modifications have led to improved performance, several studies have yielded inconsistent results, and there is currently no general consensus on the best force field for IDPs.

Since the conformational space of disordered proteins is significantly more complex and heterogeneous than for globular proteins, enhanced sampling for IDPs is critical.[401–403] Accordingly, various enhanced sampling schemes and algorithms have been employed, such as temperature replica exchange molecular dynamics (tREMD),[404] bias-exchange metadynamics (BE-META),[405–408] and Markov state models (MSMs).[409–414] In replica exchange protocols the protein undergoes a random walk, where information is exchanged between replicas periodically.[190] Metadynamics based approaches may use experimental data to guide sampling (i.e. need biased collective coordinates to be defined),[415] whereas in MSMs the goal is to probe long time dynamics by constructing kinetic models from numerous (mainly short) unbiased MD simulations.[205] Several other sampling approaches exist,[401–403] including some that utilise reweighing techniques.[392]

In the present work, the cytoplasmic tail of human cluster of differentiation 4 (CD4), an IDR linked to HIV-1 infection, is investigated computationally.[216] CD4 is a glycoprotein (gp), mainly expressed on the surface of regulatory T cells (a subclass of white blood cells).[416] The glycoprotein contains 433 residues and is made up of three regions: an extracellular N-terminal region (371 residues), a transmembrane helix (24 residues) and a cytoplasmic C-terminal domain (38 residues), Figure 5.1.[216,417,418] The T cell receptors (TCR) are responsible for recognising antigen peptides bound to Major Histocompatability Complex class II (MHC-II) molecules. During this recognition process, CD4 functions as a coreceptor, whereby the extracellular region binds to a region of MHC-II.[419,420] Additionally, the cytoplasmic tail of CD4 interacts with a lymphocyte-specific protein kinase ($\mathrm{p56^{lck}}$), by jointly coordinating $\mathrm{Zn^{2+}}$ via a pair of cysteines on each molecule (residues 420 and 422 in CD4).[421] Ultimately, the interaction between CD4 and the TCR complex recruits $\mathrm{p56^{lck}}$ to TCR (which induces phosphorylation of TCR-associated molecules),[422,423] and is a central upstream event in TCR signal transduction and the immune response.

CD4 also acts as a primary receptor of human immunodeficiency virus type 1 (HIV-1).[424–427] Specifically, entry of HIV-1 into the T cells is initiated by binding of the glycoprotein gp120 (on the HIV-1 envelope) to the extracellular region of CD4.[428]

This initial interaction eventually leads to fusion of the viral and cell membranes and subsequent infection. Once the cell has become infected, the presence of CD4 is known to interfere with the viral life cycle;[429,430] for example, CD4 may inhibit the release of nascent viruses.[431,432] However, HIV-1 has evolved to produce viral proteins, Nef (negative factor) and Vpu (viral protein U), which physically bind to and downregulate CD4 in T cells, ensuring viral proliferation.[433,434]

Co-immunoprecipitaion and mutational analyses suggest that residues 402–419 of CD4, located in the membrane-proximal region of the cytoplasmic tail, are necessary for HIV-1 viral protein-induced downregulation.[435–438] Moreover, a putative amphipathic α-helix in that region was found to be responsible for binding of both Nef and Vpu to CD4.[437,439–442] The CD4 receptor peptide (residues 403–419) was characterised by CD and NMR spectroscopy and it was reported that residues 403–412 formed an α-helix, with an equilibrium population of about 25%.[443] In a recent simulation study, employing REMD and MSM building techniques, it was demonstrated that the free energy landscape (FEL) for the receptor peptide (residues 402–419) was mainly flat—a characteristic feature of IDPs.[444]

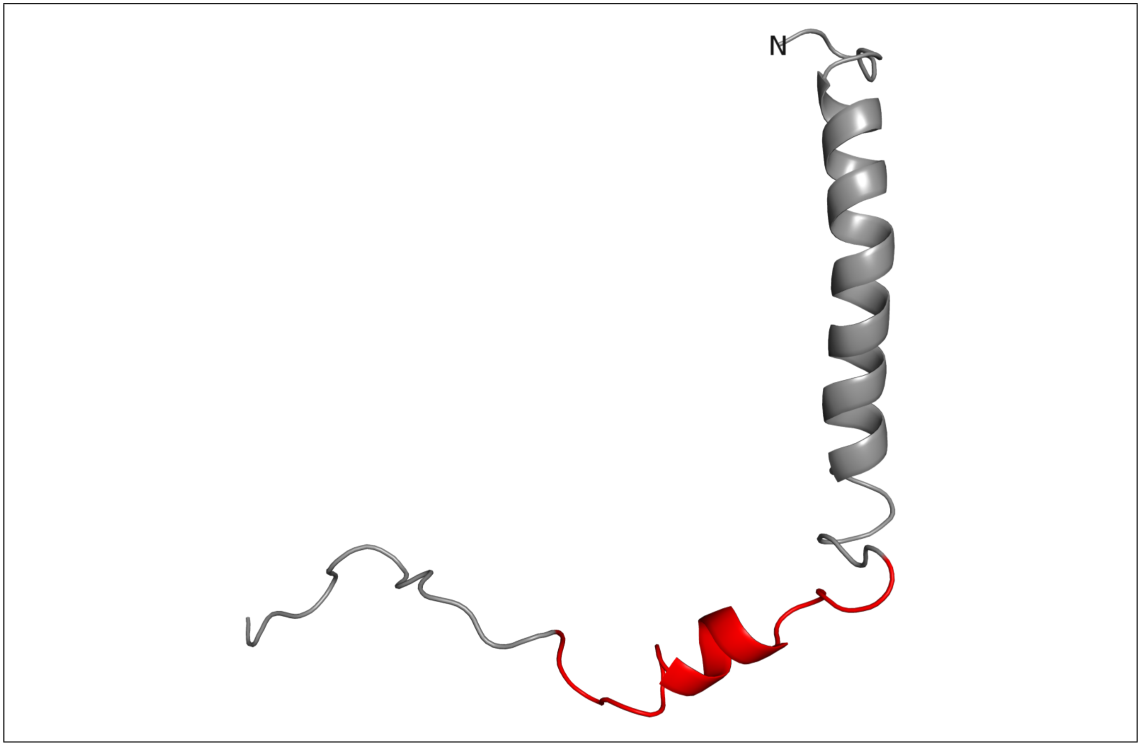


Figure 5.1: NMR solution structure for the transmembrane and cytoplasmic domains of human CD4 (PDB code: 2KLU).[418] The CD4 receptor peptide (402–419) is highlighted in red.

Here, an alternative approach for modelling intrinsic disorder in proteins, based on the potential energy landscape (PEL) framework,[210–212] is presented. Specifically, geometry optimisation-based approaches, namely basin-hopping parallel tempering (BHPT)[445] and discrete path sampling (DPS),[249–251] are utilised to map out the PEL and FEL for the CD4 receptor peptide ($\mathrm{CD4_{RP}}$). Recently, Chebaro *et al.*[446] used REMD and DPS to probe the FEL for an IDR of the p53 upregulated modulator of apoptosis (PUMA) protein, and found that the potential energy landscape was inherently multifunnelled. In earlier work, BHPT was employed to predict favourable oligomers of amyloid β-peptide, $\mathrm{A}\beta_{1-42}$,[445] implicated in Alzheimer's disease.

The BHPT–DPS approach combines structure prediction and network building for efficient sampling of IDPs. Using these tools, benchmarks are presented for $\mathrm{CD4_{RP}}$ ensembles generated with various state-of-the-art AMBER force fields (ff99SB-ILDN,[227] ff14ipq,[228] ff14SB[229]). AMBER force fields are widely used to model protein folding and there is an ongoing need to examine their performance, particularly for IDPs. Since the PEL framework offers an integrated approach for probing structure, dynamics and thermodynamics, along with powerful landscape visualisation techniques,[210] it can provide systematic comparisons of IDP ensembles generated by different force fields. We find that ff99SB-ILDN achieves a better balance of helical and random-coil structure for $\mathrm{CD4_{RP}}$ than the more recent ff14ipq and ff14SB parametrisations (§ 5.3.3 and 5.3.4). The free energy landscape for $\mathrm{CD4_{RP}}$ was computed for ff99SB-ILDN, and various metastable states were identified (§ 5.3.5). These results unify previous conflicting experimental findings for $\mathrm{CD4_{RP}}$ and account for the rich functional repertoire of this intrinsically disordered region. Finally, we discuss the biological implications of our results relating to HIV-1 infection and opportunities for rational drug design (§ 5.4).

5.2 Methods

Preparation of initial structures

The NMR solution structures for human CD4 (403–419) were obtained from the Protein Data Bank (PDB ID: 1WBR). All 32 models were extracted, and the missing arginine residue (402) was added using PyMOL.[347] The N- and C-terminal ends of the peptide were capped with an acetyl group and methylamide, respectively.

Force fields and solvent models

The performance of three AMBER force fields was tested in this study. AMBER ff99SB-ILDN[227] was developed to improve the side-chain torsion potentials of isoleucine, leucine, aspartate and asparagine in AMBER ff99SB.[226] Improvements of side-chain torsion potentials for most amino acids, as well as backbone ϕ and ψ dihedral parameters, were later implemented as AMBER ff14SB.[229] Hence, these two force fields derive from the same parent (AMBER ff94).[221] In contrast, AMBER ff14ipq[228] represents a full rederivation of torsion parameters, coupled with the implicitly polarised charge (IPolQ) model[447] for approximating protein partial charges in the condensed phase.

Simulations of CD4$_{\mathrm{RP}}$ (402–419) using the above force fields were performed with both explicit and implicit solvent models. For the explicit solvent investigations, ff99SB-ILDN and ff14SB force fields were paired with TIP3P,[307] and ff14ipq was paired with TIP4P-Ew.[448] The explicit solvent models were chosen to be consistent with the models used to parametrise the respective force fields.

All implicit solvent simulations were conducted by coupling the force fields with GB-Neck2[238] (igb=8 in Sander). Maffucci and Contini[400] reported that GB-Neck2 was able to compensate for slight helical biases displayed by the ff99SB series. The ff99SB-ILDN/GB-Neck2 combination was also able to discriminate helices from IDPs and reasonably predicted β-hairpins in their study.[400] An effective salt concentration of 0.1 M was maintained to provide mobile counterions in solution, and a cutoff of 25 Å was used for computation of the effective Born radius. The AMBER force fields were also correctly symmetrised,[306] to ensure that the energies of permutational isomers were identical.

Structural refinement via MD

For each force field, the 32 initial structures were first minimised in vacuum for 8000 steepest-descent (SD) steps, followed by 2000 steps using the conjugate gradient (CG) method. The minimised structures were then solvated in an octahedral box, with a minimum distance of 12 Å between the peptide and the box edge. For each system, four Cl^{-} ions were added to neutralise the charges on the peptide. The solvated structures were then minimised for 10000 steps (8000 SD; 2000 CG) with a cutoff of 10 Å and periodic boundary conditions for computing non-bonded interactions. A force constant of $100\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$ was used to restrain the peptide. The restraint was subsequently removed and the systems were allowed to relax for a further 10000 steps, using the same minimisation protocol as before.

Each system was then heated to 300 K for 20 ps, and the temperature was regulated using a Langevin thermostat with a collision frequency of 1 ps. A small restraint of 10 kcal mol^{-1} Å^{-2} was imposed on the peptide, and bonds involving hydrogen were constrained using the SHAKE algorithm, permitting an integration time step of 2 fs. The peptide restraints were removed and 5 ns of constant pressure (1 atm) MD was carried out; followed by 2 ns of MD in the canonical ensemble. Finally, 100 ns of MD in the NVT ensemble at 300 K was performed for each system (i.e. an aggregate production time of 3.2 μs for each force field). Frames along the trajectories were saved every 10 ps. The lowest energy conformer from each MD simulation (32 conformers per force field) was then selected to initiate structure prediction runs via basin-hopping parallel-tempering (BHPT).

Structure prediction via BHPT

In BHPT, multiple basin-hopping (BH)[239,240] runs for replicas at different temperatures are performed simultaneously.[445] As with other replica exchange methods, the lower temperature limit for BHPT is usually chosen as the temperature at which physical observables are to be evaluated, while the high temperature limit is selected to permit crossing of the highest energy barriers on the landscape. After a given number of BH steps, replicas at adjacent temperatures may be exchanged, based on a Metropolis criterion for the energies of the replicas.

In this work, the goal is not to locate the global potential energy minimum, since this structure for IDPs is unlikely to dominate the equilibrium properties. Instead, BHPT was utilised to explore the low energy regions of the PEL. The exchange of replicas in the BHPT procedure avoids kinetic trapping and facilitates accelerated exploration and sampling of the conformational space.

For each CD4$_{\mathrm{RP}}$ conformer, BHPT runs were conducted using 16 replicas at temperatures exponentially distributed between 300 and 550 K using an implicit solvent representation. Each BHPT run consisted of 5000 BH steps and exchanges among neighbouring replicas were attempted every 10 steps. To explore the conformational space, random displacements of Cartesian coordinates were used (with a maximum step size of 1 Å) and step sizes were adjusted to achieve an acceptance probability of 20%. The most stable conformer at 300 K from each BHPT run was chosen as a starting point for construction of the PEL via discrete path sampling (DPS).

Construction of potential energy landscapes via DPS

Initial discrete paths between pairs of CD4$_{\mathrm{RP}}$ conformers were first constructed. The starting conformer pairs (endpoints) were aligned using the LPERMDIST procedure in OPTIM.[321] This procedure effectively computes the minimised distance between endpoints with respect to translation, rotation and permutation. The updated quasi-continuous interpolation (QCI) scheme[349] was then used to predict configurations between the two endpoints. The interpolated geometries were then fed to the doubly-nudged[252] elastic band[253,254] (DNEB) routine, providing a discrete set of images between selected endpoints. Maxima along this profile were then taken as initial transition state guesses. These candidates were then tightly converged using the hybrid eigenvector-following (HEF) procedure.[255,256] Refined transition states were then connected to minima by following steepest-descent paths parallel and anti-parallel to the unique downhill direction. Minima were converged using a modified limited-memory Broyden-Fletcher-Golgfarb-Shano (L-BFGS) algorithm[242,243] with a convergence condition of 10^{-7} kcal mol^{-1} Å^{-1} for the root-mean-square gradient. The initial paths were then optimised using the SHORTCUT[257,258,260] and UNTRAP[260] schemes in PATHSAMPLE[322] to produce the final stationary point databases (kinetic transition networks). These procedures are summarised in sections 2.5.1 and 2.5.2.

Estimation of free energies and conversion pathways

The free energy landscape (FEL) was computed at 300 K using the harmonic superposition approximation (HSA; § 2.6).[261] Minima and transition states were first clustered into free energy groups using a recursive lumping procedure (REGROUPFREE in PATHSAMPLE; § 2.7.4).[264] The conversion pathways between conformers were obtained by applying Dijkstra's shortest path algorithm[258] to the clustered kinetic transition network (KTN).

Secondary structure analysis, computation of NMR shifts and coupling constants

Secondary structure analysis of minima in each transition network was conducted using the DSSP program[350] and NMR shifts were computed with ShiftX.[449] To determine overall NMR shifts, weighted sums, based on the equilibrium occupation probabilities, were calculated at 300 K for minima in the KTN.

Three-bond J_{HNH_α} coupling constants were computed using the Karplus equation:

$$^3J_{HNH_\alpha}(\phi) = Acos^2(\phi) + Bcos(\phi) + C, \tag{5.1}$$

with $A = 9.5$, $B = -1.4$ and $C = 0.3$, as proposed by Brüschweiler and Case.[450]

Visualisation of energy landscapes

Finally, the potential energy landscapes (and free energy surfaces) were visualised using disconnectivity graphs.[265,266]

5.3 Results and Discussion

5.3.1 Low-lying conformers on the potential energy landscape

The NMR solution structures for CD4 (residues 403–419) are depicted in Figure 5.2a. The NMR-averaged ensemble displays notable structural variation, particularly in the C-terminal region, and the conformers show some tendency to form α-helices between residues 403–412, as reported earlier by Willbold and Rösch.[443] Each of the 32 structures in the NMR ensemble were simulated via MD and subsequently BHPT using the three AMBER force fields. Explicit solvent MD was used to probe regions of local stability on the potential energy landscape of $\mathrm{CD4_{RP}}$. Seeding BHPT from MD trajectories is not strictly required, since BHPT runs can be initiated from the prepared NMR structures themselves. However, starting from locally stable regions on the PEL meant that fewer BHPT steps were required to satisfactorily explore low-lying wells on the PEL. On average, the potential energy of conformers improved by about $8\,\mathrm{kcal\,mol^{-1}}$ after BHPT, and the final 32 conformers covered a range of less than $20\,\mathrm{kcal\,mol^{-1}}$ for each of the three force fields.

The BHPT predicted conformers from each force field are shown in Figure 5.2. It is evident that ff99SB-ILDN and ff14ipq show greater structural variation than ff14SB. The conformers for the first two force fields resemble molten globules, in which there is partial formation of helical structure. Conversely, the final ensemble for ff14SB reveals significant helical character, where the $\mathrm{CD4_{RP}}$ forms linear helices. On average, contiguous α-helices nine, seven, and eleven residues long were obtained for ff99SB-ILDN, ff14ipq, and ff14SB, respectively. Notably, consistent with the NMR experiment, no antiparallel β-strands were seen.

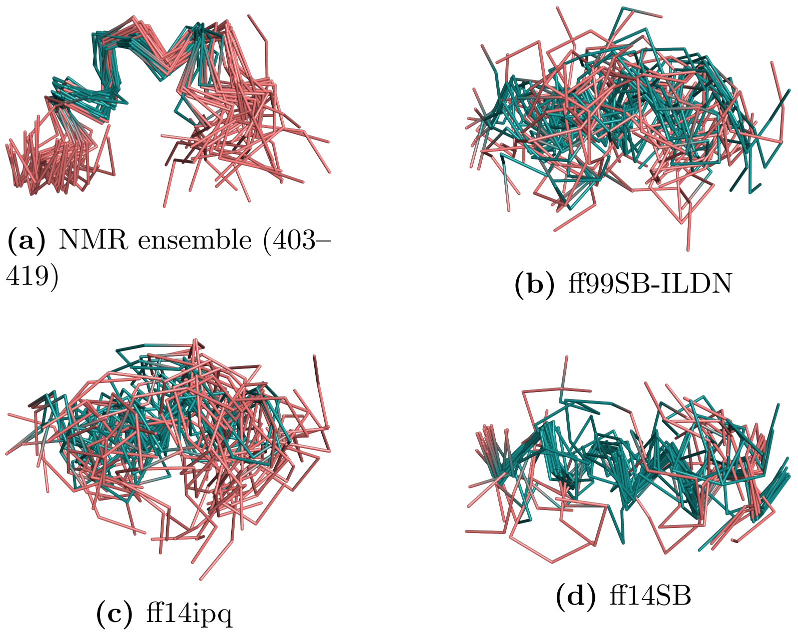
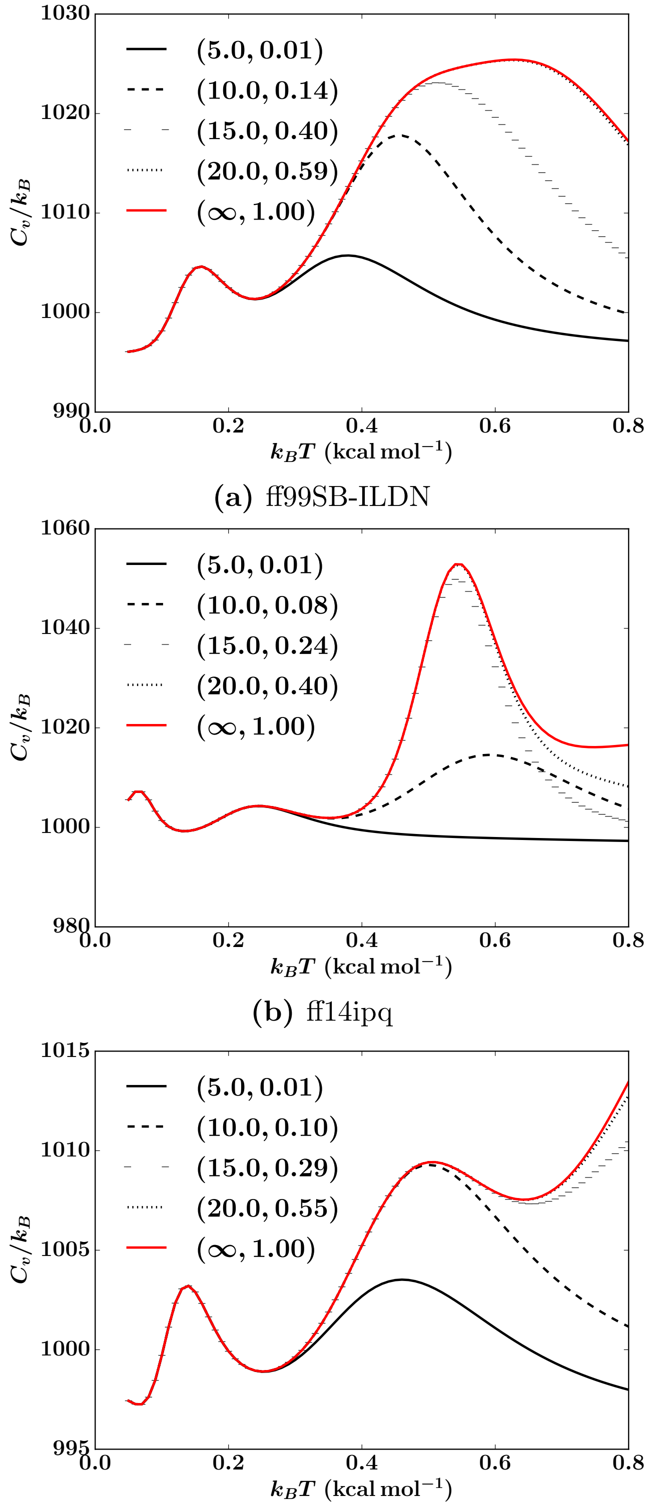


Figure 5.2: Structural ensembles for the human CD4 receptor peptide: (a) 32 NMR solution structures (residues 403–419); (b)–(d) conformers (residues 402–419) obtained after BHPT, using various AMBER force fields. The conformers were clustered using PyMOL and were oriented with the N-terminal region on the left. The colouring scheme is: teal (helices), salmon (loops).

5.3.2 Convergence of stationary point databases

The conformers in Figure 5.2b–5.2d were used to construct potential energy landscapes for $\mathrm{CD4_{RP}}$ via discrete path sampling. To test for convergence of the respective stationary point databases, the constant volume heat capacity (C_v) curves were computed for subsets of the local minima. In particular, C_v curves were derived for minima within a given threshold above the global potential energy minimum (Figure 5.3). At low temperatures, we expect that features in the heat capacity may be attributed to a small subset of local minima. Hence, this approach should be robust in testing for convergence of the databases and should be able to reveal any deficiencies in sampling. For the three force fields tested, the low-temperature features in the heat capacity are converged when about 50% of all minima in the corresponding stationary point database are included. A summary of the number of local minima and transition states in the final database for each force field is given in Table 5.1. In the next section, the properties of the resulting potential energy landscapes for $\mathrm{CD4_{RP}}$ are compared systematically.



(a) ff99SB-ILDN

(b) ff14ipq

(c) ff14SB

Figure 5.3: Constant volume heat capacity curves (C_v) for the human CD4 receptor peptide. The curves were computed using the harmonic superposition approximation for the peptide simulated by various AMBER force fields. Minima within a specified energy threshold of the global potential energy minimum were incrementally included in the calculation of C_v. The energy thresholds (kcal mol^{-1}) and the corresponding fraction of minima in the database are provided for each plot in parentheses.

AMBER force field	No. of MIN	No. of TS
99SB-ILDN[227]	47434	51675
14ipq[228]	53926	58008
14SB[229]	67276	68689

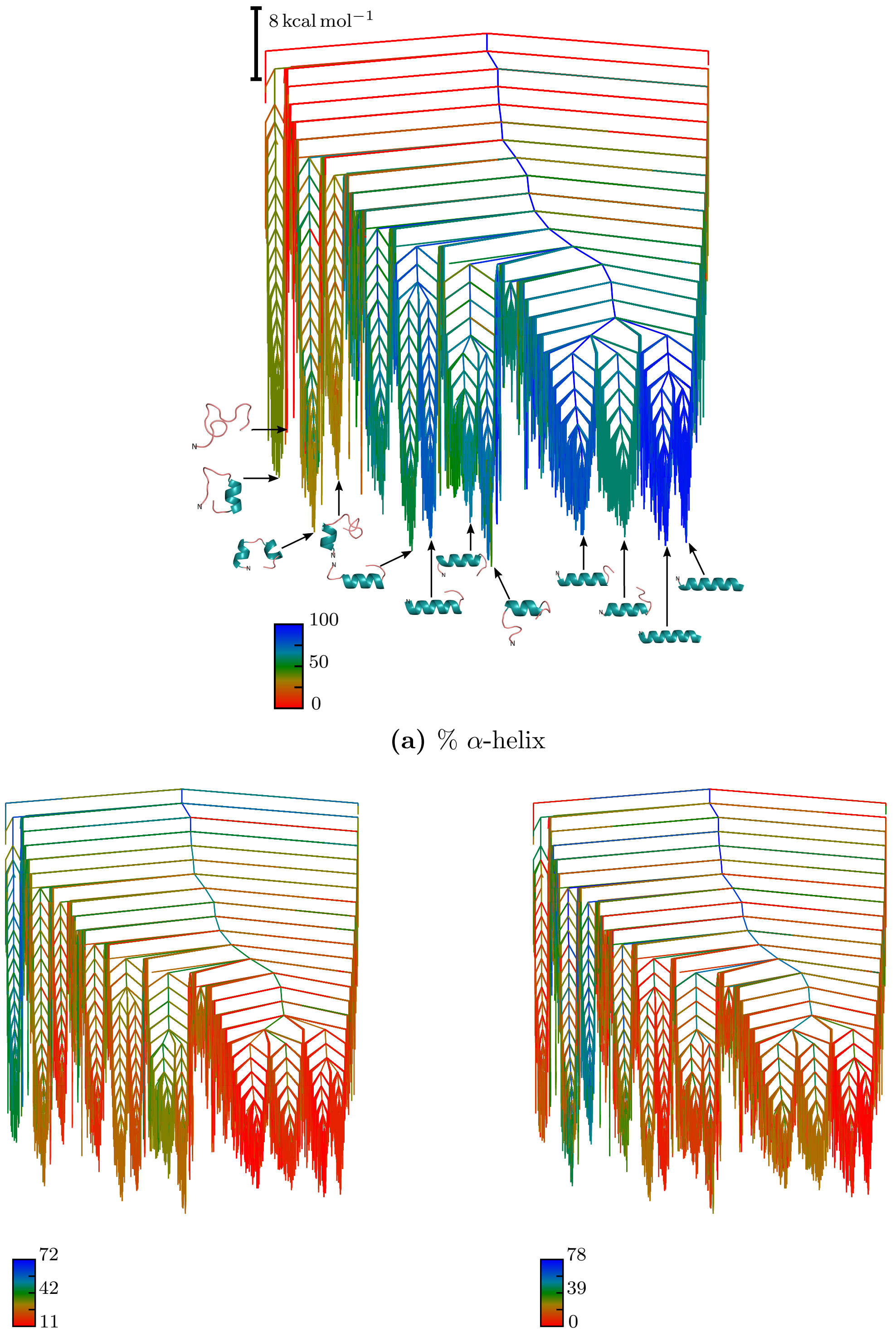
Table 5.1: Number of minima (MIN) and transition states (TS) in the DPS databases for the CD4 receptor peptide modelled by various AMBER force fields

5.3.3 Characterisation of potential energy landscapes

The potential energy landscape for $\mathrm{CD4_{RP}}$ modelled by the ff99SB-ILDN force field is presented in Figure 5.4; it is distinctively multifunnelled, with several prominent low-lying funnels. It was previously postulated that such multifunnelled potential energy landscapes are characteristic features of intrinsically disordered proteins.[446]

The disconnectivity graphs in Figure 5.4 are coloured based on the secondary structure content; specifically, the percent α-helical, random-coil, and turn structure are shown, since these features were most common in the computed KTNs. The fractional α-helicity is best able to classify the funnels on the landscape (Figure 5.4a). Many low energy conformers contain linear α-helices, differing mainly in the length of the helical segment and orientation of side-chain groups. The shortest α-helices in that region are about six residues long (approx. 1.7 turns), and the longest consists of 18 residues. Slightly higher in energy are conformers that display helical propensities in the N- and C-terminal segments of the peptide. There are also a number of conformers that lack α-helical structure and are quite low in energy.

Chebaro *et al.* found that for the PUMA peptide the contiguous α-helix was enthalpically unstable and that low-lying minima were relatively unstructured.[446] The organisation of the PEL for the PUMA peptide strongly suggested that the interaction of this IDP with molecular partners would mainly involve an induced-fit type mechanism. As mentioned above, the longest contiguous α-helix (residues 402–419) in $\mathrm{CD4_{RP}}$ corresponds to a low energy conformer on the PEL. This result suggests that interactions involving direct binding to $\mathrm{CD4_{RP}}$ are most likely to occur via conformational selection.



(a) % α-helix

(b) % random-coil

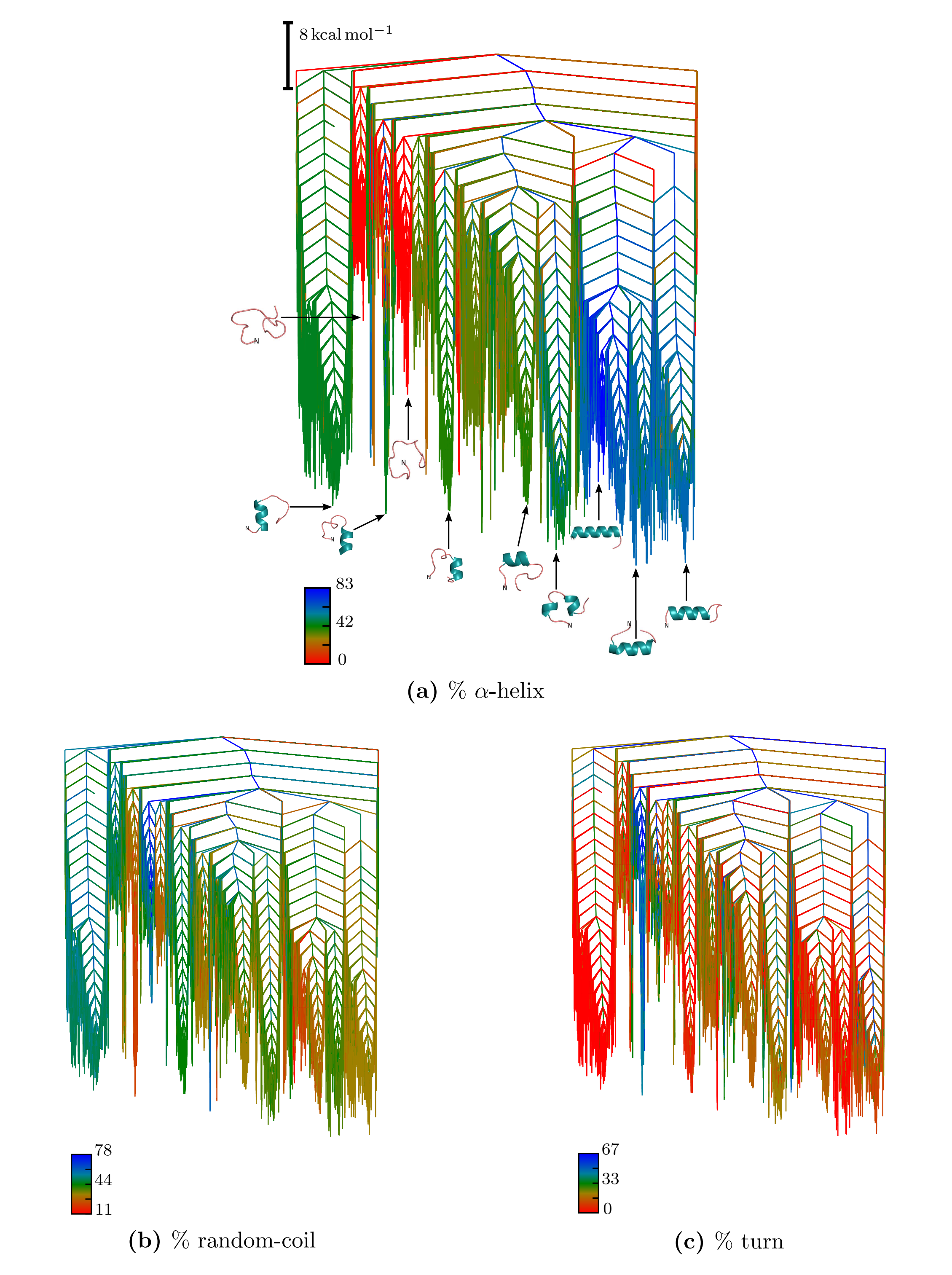
(c) % turn

Figure 5.4: Potential energy landscape for CD4$_{\mathrm{RP}}$ modelled by the AMBER 99SB-ILDN force field. In the disconnectivity graphs, minima (branches) are coloured based on secondary structure content. Selected structures are superimposed on the graph and the α-helical segments are coloured in teal.

The potential energy barriers between conformers exceeding 50% α-helical content are much smaller than those separating conformers with residual helical content. The latter conformers generally lie higher in energy and display greater coil-like and turn secondary structure. Very few β-strand conformers were identified for the ff99SB-ILDN force field. For this representation, the conformers generally show α-helical propensities, and only a few of them are largely unstructured.

Figure 5.5 illustrates the potential energy landscape for $\mathrm{CD4_{RP}}$ computed with the first-generation ff14ipq force field. As for ff99SB-ILDN, the landscape is intrinsically disordered, with multiple prominent funnels. However, the relative potential energy barriers between the various low-lying conformers are notably larger than those obtained with ff99SB-ILDN. The conformers for ff14ipq are also more heterogeneous than those located with ff99SB-ILDN. The predicted linear α-helical motifs in $\mathrm{CD4_{RP}}$ are also shorter, by about two residues, for this force field. As before, conformers lacking significant helical structure, characterised by turns (Figure 5.5c) and coil-like (Figure 5.5b) secondary structure, generally reside in the high energy regions of the landscape. Additionally, a greater proportion of minima contain random-coil character than observed for ff99SB-ILDN.

Finally, the landscape for $\mathrm{CD4_{RP}}$ was probed using the ff14SB force field; the topology is evidently less multifunnelled than for the former two force fields (Figure 5.6). Additionally, the landscape is dominated by linear α-helical conformers. Secondary structure characterisation also reveals regions with substantial coil-like structure (Figure 5.6b), separated from the main part of the landscape by particularly high potential energy barriers. However, even in these regions, conformers still exhibit some helical character. Similar regions were also identified on the ff99SB-ILDN and ff14ipq landscapes, and the entropic barrier between them and the other parts of the PEL is therefore likely to be physically realistic.



(a) % α-helix

(b) % random-coil

(c) % turn

Figure 5.5: Potential energy landscape for CD4$_{\mathrm{RP}}$ modelled by the AMBER 14ipq force field. In the disconnectivity graphs, minima (branches) are coloured based on secondary structure content. Selected structures are superimposed on the graph in (a) and the α-helical segments are coloured in teal.

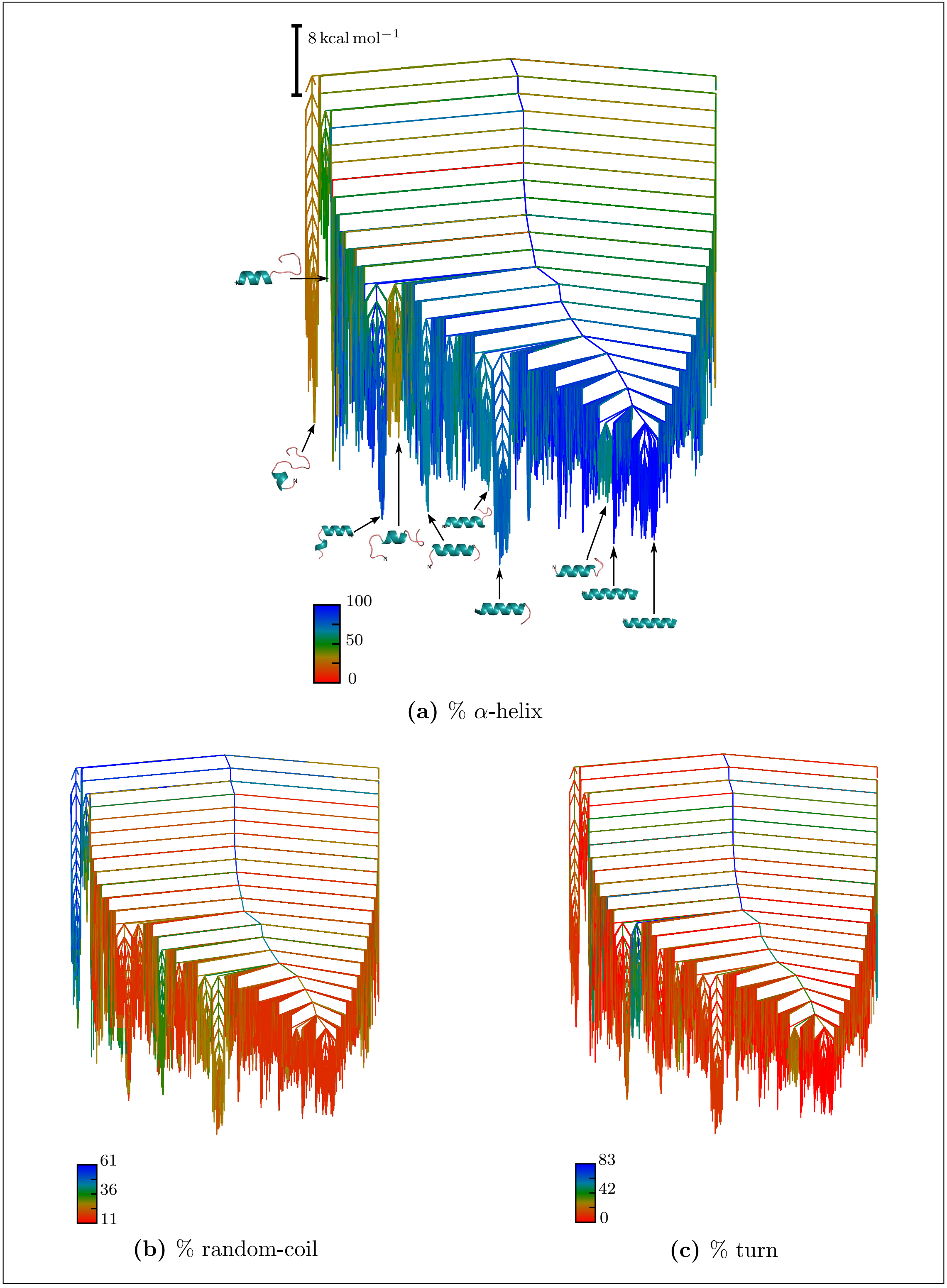
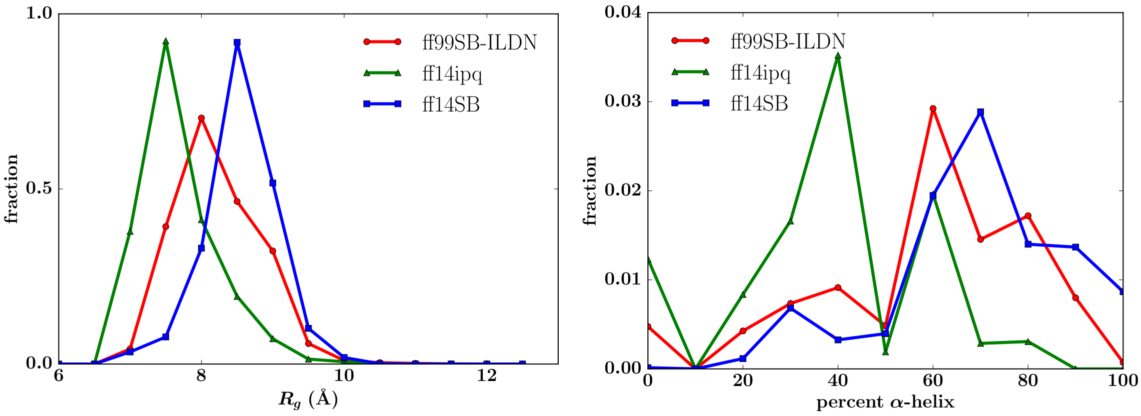


Figure 5.6: Potential energy landscape for CD4$_{\mathrm{RP}}$ modelled by the AMBER 14SB force field. In the disconnectivity graphs, minima (branches) are coloured based on secondary structure content. Selected structures are superimposed on the graph in (a) and the α-helical segments are coloured in teal.

The distribution of the radius of gyration (R_g) for the local minima of CD4$_{\mathrm{RP}}$ is shown in Figure 5.7a. For all the force fields, there is a single peak in the R_g distribution. Conformers generated with ff14SB are most extended (i.e., corresponding to linear helices), while those predicted with ff14ipq are most compact (consistent with a greater presence of molten-globule like structures on the landscape). A slightly broader R_g distribution is obtained for ff99SB-ILDN, which peaks between the two newer AMBER force fields.

Figure 5.7b summarises the distribution of fractional α-helicity. All force fields show multiple peaks in the distribution (at a bin width of 10%), with the major peaks for ff99SB-ILDN and ff14SB at about 60–70% α-helicity. The distribution for ff14ipq is shifted to the left; in general, ff14ipq predicts peptides with fewer helical residues than the other two force fields. Additionally, on inspecting the left tails of the distributions, it is apparent that only a few conformers located with f14SB lack α-helical character. The distributions for ff99SB-ILDN and ff14SB are somewhat similar, although the former parametrisation predicts a greater fraction of structures with low α-helical content.



(a) **(b)**

Figure 5.7: Distribution of radius of gyration (R_g) and fractional α-helicity for local minima of CD4$_{\mathrm{RP}}$ generated with various AMBER force fields, as indicated.

5.3.4 NMR shifts and coupling constants

Predicted NMR shifts are a useful tool for validating simulation data by comparison with experiment. The computed HN NMR shifts for human CD4 (residues 403–419) are presented in Figure 5.8. The agreement between the experimental and calculated

NMR shifts was assessed by computing

$$\chi^2_\sigma = \frac{1}{N}\sum_{i=1}^{N}\frac{(\delta_{i,cal}-\delta_{i,exp})^2}{\sigma^2_\delta}, \tag{5.2}$$

where i is the residue number, N is the number of residues, $\delta_{i,cal}$ and $\delta_{i,exp}$ are the calculated (as described in § 5.2) and experimental chemical shifts, respectively, and σ^2_δ is the uncertainty ($\sigma_\delta = 0.49$ ppm for HN using SHIFTX). By definition, $\chi^2_\sigma > 1$ indicates a significant difference between computed and experimental structures.[451] The χ^2_σ values for the HN NMR shifts for ff99SB-ILDN, ff14ipq and ff14SB structures are 0.28, 0.55 and 0.42, respectively. Based on this metric, the HN shifts for the ensembles generated with all three force fields agree quite well with the experimental findings.[443] The main deviations from the NMR experiments occur in the C-terminal region of the peptide (residues 417–419); the *de novo* approach employed in this work generally yields more ordered C-termini than found in experiment. In the N-terminal part of the peptide, ff14SB predicted more structured helices, while ff14ipq predicts more disorder than experiment. However, the overall HN NMR shifts, particularly those computed for ff99SB-ILDN, closely match the literature values.

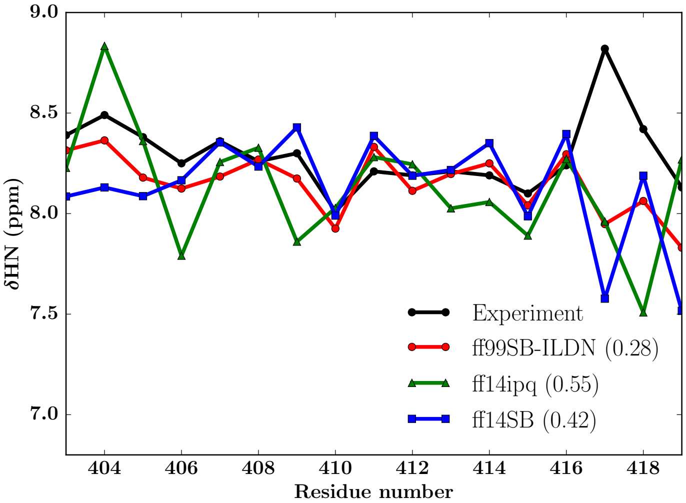


Figure 5.8: Comparison of HN NMR shift for CD4$_{RP}$ ensembles generated with various AMBER force fields and experiment. The χ^2_σ uncertainty, which measures the agreement between computed and experimental shifts, is given in parentheses.

In addition to NMR shifts, average $^3J_{HNH_\alpha}$ values (weighted by occupation probabilities at 300 K) were computed for residues 403–419, and the root-mean-square deviation (rmsd) of the calculated and experimental values was determined. For ff99SB-ILDN, ff14ipq and ff14SB rms deviations of 1.40, 1.68 and 1.50 Hz were obtained, respectively. Overall, ff99SB-ILDN performs best for CD4$_{RP}$, in terms of reproducing the HN NMR shifts and J coupling constants. This force field was therefore selected for further analysis, specifically for probing the free energy surface.

5.3.5 Free energy landscape for CD4$_{\mathrm{RP}}$

The free energy landscape for CD4$_{\mathrm{RP}}$ was produced by clustering minima and transition states in the ff99SB-ILDN-derived KTN, based on the relative free energy barriers. To cluster the stationary points, the lowest energy minima with highest fractional random-coil and α-helical structure were taken as the starting configurations for A (reactant) and B (product), respectively. Using a recursive regrouping scheme, A and B were then expanded as free energy groups, based on a predefined free energy threshold (ΔF) at 300 K. The rate constant for the $A \rightarrow B$ conversion was derived by employing the new graph transformation formulation (see § 2.7.3 for details).[262] To determine an appropriate value for ΔF, the procedure was repeated for different values in the range 0 to 20 kcal mol^{-1}, and the smallest threshold (in this case, 5.5 kcal mol^{-1}) for which the rate constant converged was chosen for probing the free energy surface.

The resulting free energy landscape is depicted as a disconnectivity graph in Figure 5.9, and representative structures for various free energy groups are superimposed on the graph. The computed FEL is similar to the corresponding PEL in Figure 5.4. The lowest free energy group at 300 K (g1) contains a contiguous α-helix extending from residues ALA404 to SER415. Other free energy groups, of comparable energies, also contain linear α-helical motifs: g2 (406–415), g4 (403–419), g5 (402–415).

A putative amphipathic α-helix between residues 402–419 of the cytoplasmic tail of human CD4 may be necessary for interaction with both Nef and Vpu—HIV-1 accessory proteins.[437,439–442] Willbold and Rösch predicted an α-helix from residues 403–412 for CD4$_{\mathrm{RP}}$.[443] Putative α-helices extending from residues 402–417,[216] 406–415[452] and 404–413[417,418] have also been reported for the cytoplasmic tail of human CD4. On the computed free energy surface, these motifs are most similar to those found in g4, g2 and g9, respectively, which are all low-lying states. These results suggest that CD4$_{\mathrm{RP}}$ is capable of adopting a wide range of helical motifs of varying lengths, which extend from different residues; however, the precise α-helical structure observed is most likely dependent on the prevailing conditions or the resolution of the experiment. The results here help to unify previous experimental work, and explain the ability of CD4$_{\mathrm{RP}}$ to bind different molecular partners. Additionally, Wittlich and collegues reported that the amphipathic helix was stable even at

45°C.[418] The free energy landscape was also examined at 318 K (not shown) and is in agreement with those findings.

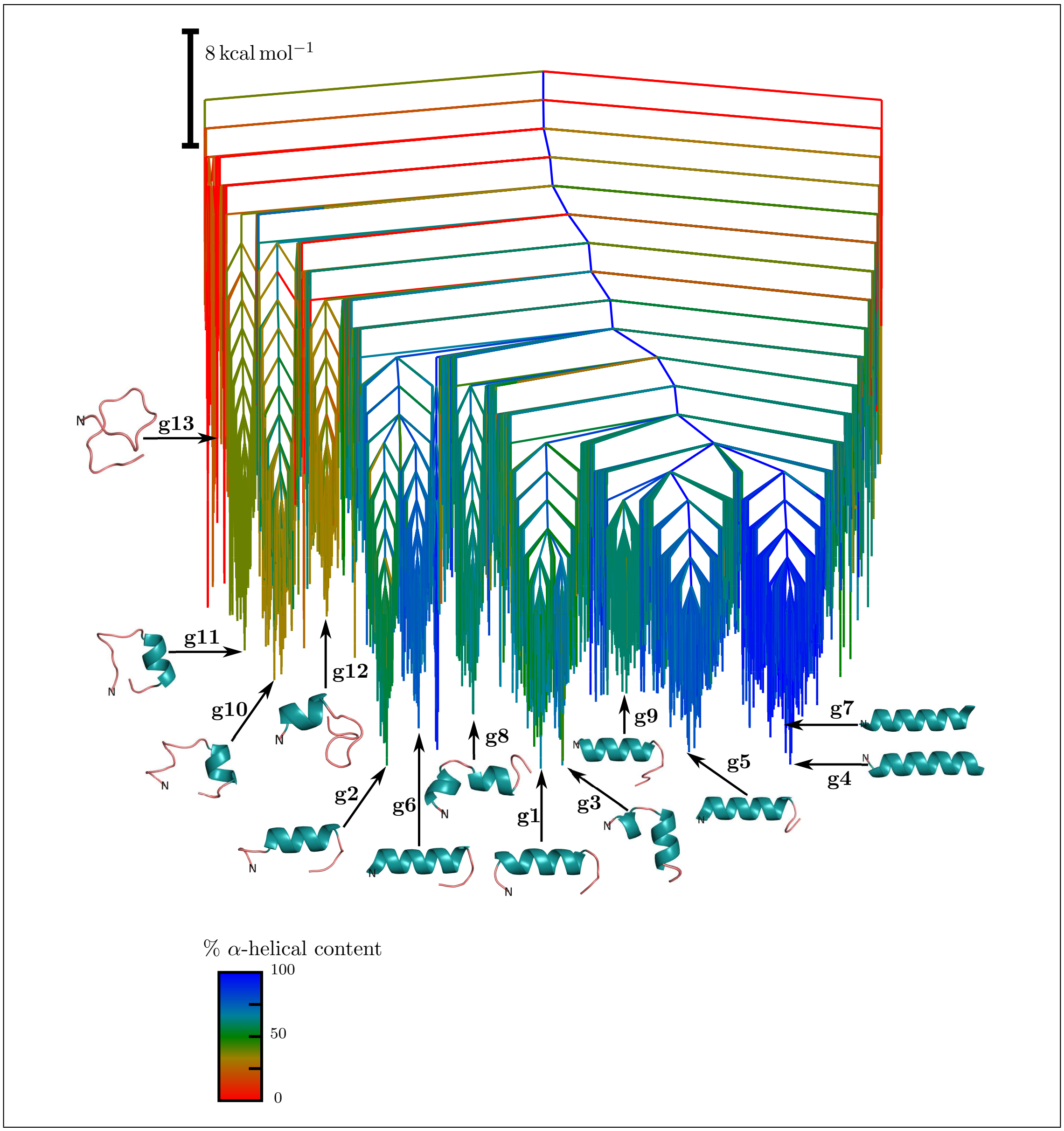


Figure 5.9: Free energy disconnectivity graph for CD4$_{\mathrm{RP}}$ modelled by the ff99SB-ILDN force field. The free energy groups (branches) are coloured according to the α-helical content. Representative structures of selected free energy groups are superimposed on the graph and the α-helical motifs are coloured in teal.

The relatively high free energy barriers between the various α-helical states on the FEL also suggest that conversion between them is likely to occur on long time scales. Furthermore, fully unstructured states are generally higher in energy on the FEL and it is unlikely that CD4$_{\mathrm{RP}}$ is unstructured under physiological conditions.

Moreover, the intrinsic disorder in this peptide may be accurately described as a tendency to adopt various ordered states encompassing α-helical scaffolds. The dominance of these α-helical scaffolds in the low-lying regions of the FEL suggests that interaction with molecular partners may occur via conformational selection.

5.3.6 Folding mechanism for CD4_{RP}

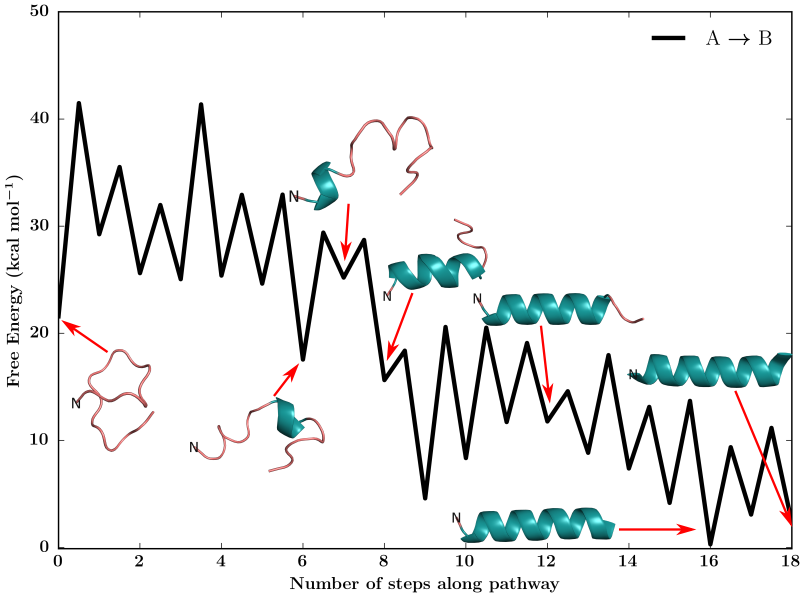


Figure 5.10: Folding pathway from a fully unstructured (A) state to a state containing a contiguous α-helical structure 18 residues long (B) on the FEL for CD4_{RP}. The number of steps corresponds to the number of transition states on the discrete path.

To gain further insight into the folding dynamics of CD4_{RP}, the folding pathway from a fully unstructured state (g13) to a state containing the longest contiguous α-helix (g7; 18 residues long) was examined[†] (Figure 5.10). Folding from the 'random-coil-like' state occurs in a cooperative fashion; i.e. there is a gradual decrease in the free energy of the states encountered as the peptide folds. This result agrees well with the findings of Ahalawat and co-workers,[444] who also reported that the formation of the helix began at residues 407–410.[444] For the pathway depicted in Figure 5.10, the helix initiates from a similar region (410–413). The N-terminal portion of the helix forms first and several states on the pathway (between steps 8–12) exhibit α-helical residues for the membrane proximal region (403–413) with an unstructured C-terminus. Interestingly, the helical motif in that segment of the path is almost

[†] These free energy groups correspond to the final free energy groups for the initial A and B conformers selected by defining the reactant and product via regrouping.

identical in length and position with the original experimental prediction (403–412). It appears that this scaffold is an important structure for $\mathrm{CD4}_{\mathrm{RP}}$ function.

Finally, pathways and corresponding rate constants for conversions between various states on the FEL were examined. Generally, the rate for folding from states with linear helical motifs (e.g. g4, g9) to g1 is four to six orders of magnitude faster than from states without helices (e.g. g10, g11, g13). However, the rate constant for interconversions of various ordered helical scaffolds (e.g. g7 → g1) is of the order of $10^{-5}\,\mathrm{s}^{-1}$, suggesting that these transformations may occur slowly. The free energy barriers for such interconversions may be lowered in the presence of a molecular partner. Alternatively, different prevailing conditions in the cell may favour one state over the another, explaining why several groups have reported differing lengths for linear helical motifs for $\mathrm{CD4}_{\mathrm{RP}}$.

5.4 Conclusions

Binding of HIV-1 glycoprotein gp120 to the extracellular domain of CD4 leads to exposure of epitopes, which can be targeted by antibodies.[453] However, the presence of HIV-1 accessory proteins, Vpu and Nef, indirectly reduces the exposure of these epitopes by downregulating CD4 from the surface of infected cells.[454,455] Several approaches for HIV-1 antiviral drug design have focused on inducing exposure of HIV-1 epitopes, in the absence of CD4, to promote antibody-mediated responses.[456–459] These approaches adopt the conventional structure-based drug design model, which is based on targeting well-defined regions in proteins.[460]

Recently, some researchers suggested that an alternative route for drug development could be based on designing drugs that can mimic the interaction of the cytoplasmic tail of CD4 with HIV-1 accessory proteins, in an antagonistic fashion.[459,461] Similar design principles have already been adopted in cancer therapies, where small molecules that mimic intrinsically disordered regions have been used to inhibit critical protein-protein interactions.[462,463] However, the development of CD4 cytoplasmic tail-mimicking molecules has been less prolific, because the CD4–Vpu/Nef interaction site has been difficult to characterise.

In the present work, the free energy landscape for the CD4 receptor peptide has been characterised by employing geometry optimisation-based approaches. The metastable states we have identified help to unify, and are consistent with, several earlier predictions. The conformers identified herein may be used as starting points in docking studies with HIV-1 accessory proteins, to better probe the recognition process. Additionally, the strategies presented in this work may be used to characterise the full-length cytoplasmic domain. These investigations may prove useful, not only in achieving thorough characterisation of the reaction interface, but may also represent a critical step towards developing viable CD4 receptor peptide mimetics.

6

Conclusions and Outlook

In 1977, McCammon and Karplus published their seminal paper, entitled "Dynamics of folded proteins".[158] Their molecular dynamics simulation of bovine pancreatic trypsin inhibitor (a folded globular protein) would be the first of its kind and was only about nine picoseconds long. Today, similar proteins are simulated on supercomputers, such as Anton,[464] in the millisecond regime. Advancements of computing hardware and algorithms have been accompanied by critical theoretical discussions, improvements in physical models and development of new sampling techniques—all of which have broadened our understanding of proteins and protein folding. Notably, investigations of disordered and misfolded proteins and their role in disease have been significantly aided by *in silico* approaches.

However, new discoveries have presented new challenges: experimental characterisation of larger protein structures (beyond the conventional 25 kDa limit)[465] necessitates greater computational efficiency. Large-scale structural changes (e.g. in proteins that switch folds) represent kinetic bottlenecks, and models traditionally optimised for folded globular proteins are generally insufficient for describing disordered proteins.

In this thesis, we sought to address some of the aforementioned challenges by adopting approaches based on the computational potential energy landscape (PEL) methodology. In this framework, the PEL is discretised into stationary points (minima and transition states) via geometry optimisation techniques. Thermodynamic and kinetic information is then extracted from the resulting kinetic transition networks (KTNs) using established methods from equilibrium statistical mechanics and unimolecular rate theory.

First, we demonstrated how a local rigid body (LRB) framework may be implemented to probe protein folding (§ 3). This framework had previously been used to improve computational efficiency in global optimisation. In this work, we investigated how the properties of the underlying PEL of tryptophan zipper 1 varied as a function of local rigidification. We found that conservative local rigidification was able to reproduce the thermodynamic and kinetic properties of the model protein, as well as the mechanistic details of folding. However, a more aggressive local rigidification led to undesirable features, such as increased frustration in the landscape and lengthening of putative folding pathways. Accordingly, we propose that within the context of protein folding local rigid bodies must be carefully chosen.

Next, we mapped out the energy landscape for the C-terminal domain (CTD) of bacterial regulatory protein RfaH (§ 4), which undergoes a dramatic structural rearrangement from an α-helical hairpin to the β-barrel scaffold. Large-scale structural transitions, such as those observed in RfaH-CTD, are generally not amenable to conventional molecular dynamics simulations. Additionally, within the PEL framework, interpolation issues have impeded studies of such transitions. In this work, we adopted a new quasi-continuous interpolation scheme, and constructed KTNs for the refolding process. Our computed free energy landscape for RfaH-CTD at 310 K is multifunnelled, and the predicted free energy ensembles are in good agreement with experiment and other simulation studies. We found that the structural rearrangement (from the α-helical conformer to the β-sheet) proceeded via a completely unstructured state.

Finally, in § 5, the human CD4 receptor peptide ($\mathrm{CD4_{RP}}$), implicated in HIV-1, was characterised via basin-hopping parallel tempering (BHPT) and discrete path sampling (DPS). In this study, we also investigated the effects of three state-of-the-art AMBER forcefields on the energy landscape. Through comparison with experiment, our study revealed that AMBER ff99SB-ILDN is best suited to model the intrinsically disordered protein (IDP). Furthermore, we were able to rationalise why several previous studies had reported seemingly conflicting results for the CD4 cytoplasmic tail. Our work therefore helps to unify prior findings, as well as identify possible starting points for investigating the reaction interface between CD4 and HIV-1 accessory proteins.

Some avenues for future studies emerge from the results in this thesis. These include, but are not limited to:

- Extending the LRB framework to probe pocket dynamics in enzymes and other receptor proteins.[466] Hybrid approaches[467] can be employed, in which the binding pocket is represented in full atomistic detail, and the rest of the protein is treated as a larger rigid body. We anticipate that these studies may achieve significant gains in computing efficiency, while preserving the properties of interest.

- Employing the updated PEL approaches in the design of artificial proteins and synthetic foldamers.[468,469] For instance, the design of novel cyclic peptides with diverse therapeutic capabilities is an emerging field,[470] and the computational PEL framework may prove useful in this context.

- Utilising the approaches presented in this dissertation to investigate other important IDPs,[372] and in the optimisation of corresponding protein forcefields. Construction of KTNs via DPS may emerge as an important step towards characterisation of the reaction interfaces between IDPs and their molecular partners.

Throughout this thesis, we employed implicit solvation to permit efficient energy landscape construction. However, it would be worthwhile to optimise the approaches described therein for explicit solvent and ion representation, which should provide a better approximation of the physical environment. Overall, the PEL framework presents many opportunities for probing protein folding, by providing powerful procedures for obtaining thermodynamic and kinetic information. Furthermore, the derivation of free energy landscapes does not require the use of reaction coordinates (or collective variables), which are generally difficult to define *a priori*, and may lead to overly simplified representations of the complex folding subspace.

It is indeed an exciting time to be studying proteins and protein folding. I envision a future in which the strengths of different approaches are integrated to yield robust hybrid formalisms, capable of replicating conditions inside the cell.

"A combination of these theoretical approaches with the interpretation of related experiments will provide a unified description of motions in proteins."

– McCammon, Karplus (1977)

References

[1] J.-E. Shea and C. L. Brooks III, *Annu. Rev. Phys. Chem.*, **2001**, 52, 499–535.

[2] F. Sanger and H. Tuppy, *Biochem. J.*, **1951**, 49, 463.

[3] F. Sanger and H. Tuppy, *Biochem. J.*, **1951**, 49, 481.

[4] L. Pauling, R. B. Corey and H. R. Branson, *Proc. Natl. Acad. Sci. USA*, **1951**, 37, 205–211.

[5] R. B. Corey and L. Pauling, *Proc. R. Soc. London, Ser. B*, **1953**, 141, 10–20.

[6] G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.*, **1963**, 7, 95–99.

[7] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff and D. C. Phillips, *Nature*, **1958**, 181, 662–666.

[8] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will and A. C. T. North, *Nature*, **1960**, 185, 416–422.

[9] E. Haber and C. B. Anfinsen, *J. Biol. Chem.*, **1962**, 237, 1839–1844.

[10] C. J. Epstein, R. F. Goldberger and C. B. Anfinsen, in *Cold Spring Harb. Symp. Quant. Biol.*, volume 28, Cold Spring Harbor Laboratory Press, pp. 439–449.

[11] C. B. Anfinsen, *Science*, **1973**, 181, 223–230.

[12] C. Levinthal, *Molecular model-building by computer*, WH Freeman and Company, **1966**.

[13] C. Levinthal, *J. Chem. Phys.*, **1968**, 65, 44–45.

[14] C. Levinthal, in *Mossbauer Spectrosc. Biol. Syst.*, University of Illinois Press Allerton House, Monticello, Illinois, pp. 22–24.

[15] T. Y. Tsong, R. L. Baldwin, P. McPhie and E. L. Elson, *J. Mol. Biol.*, **1972**, 63, 453–469.

[16] D. B. Wetlaufer, *Proc. Natl. Acad. Sci. USA*, **1973**, 70, 697–701.

[17] O. B. Ptitsyn and A. A. Rashin, *Dokl. Akad. Nauk SSSR*, **1973**, 213, 473–475.

[18] M. I. Kanehisa and T. Y. Tsong, *J. Mol. Biol.*, **1978**, 124, 177–194.

[19] M. Karplus and D. L. Weaver, *Nature*, **1976**, 260, 404.

[20] M. Karplus and D. L. Weaver, *Protein Sci.*, **1994**, 3, 650–668.

[21] N. Go and H. Abe, *Biopolymers*, **1981**, 20, 991–1011.

[22] P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.*, **1982**, 51, 459–489.

[23] K. A. Dill, *Biochemistry*, **1985**, 24, 1501–1509.

[24] A. R. Fersht, *Curr. Opin. Struct. Biol.*, **1997**, 7, 3–9.

[25] J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, **1987**, 84, 7524–7528.

[26] J. D. Bryngelson and P. G. Wolynes, *J. Phys. Chem.*, **1989**, 93, 6902–6915.

[27] J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, *Annu. Rev. Phys. Chem.*, **1997**, 48, 545–600.

[28] C. M. Dobson, A. Šali and M. Karplus, *Angew. Chem., Int. Ed.*, **1998**, 37, 868–893.

[29] P. E. Leopold, M. Montal and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA*, **1992**, 89, 8721–8725.

[30] J. D. Bryngelson, J. N. Onuchic, N. D. Socci and P. G. Wolynes, *Proteins: Struct., Funct., Bioinf.*, **1995**, 21, 167–195.

[31] J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.*, **2004**, 14, 70–75.

[32] A. M. Labhardt, *Proc. Natl. Acad. Sci. USA*, **1984**, 81, 7674–7678.

[33] M. Ikeguchi, K. Kuwajima, M. Mitani and S. Sugai, *Biochemistry*, **1986**, 25, 6965–6972.

[34] R. I. Gilmanshin and O. B. Ptitsyn, *FEBS Lett.*, **1987**, 223, 327–329.

[35] K. Kuwajima, H. Yamaya, S. Miwa, S. Sugai and T. Nagamura, *FEBS Lett.*, **1987**, 221, 115–118.

[36] K. Kuwajima, A. Sakuraoka, S. Fueki, M. Yoneyama and S. Sugai, *Biochemistry*, **1988**, 27, 7419–7428.

[37] H. Roder and K. Wüthrich, *Proteins: Struct., Funct., Bioinf.*, **1986**, 1, 34–42.

[38] H. Roder, G. A. Elöve and S. W. Englander, *Nature*, **1988**, 335, 700.

[39] J. B. Udgaonkar and R. L. Baldwin, *Nature*, **1988**, 335, 694.

[40] S. Takahashi, S.-R. Yeh, T. K. Das, C.-K. Chan, D. S. Gottfried and D. L. Rousseau, *Nat. Struct. Mol. Biol.*, **1997**, 4, 44.

[41] C.-K. Chan, Y. Hu, S. Takahashi, D. L. Rousseau, W. A. Eaton and J. Hofrichter, *Proc. Natl. Acad. Sci. USA*, **1997**, 94, 1779–1784.

[42] M. C. R. Shastry, S. D. Luck and H. Roder, *Biophys. J.*, **1998**, 74, 2714–2721.

[43] C. M. Phillips, Y. Mizutani and R. M. Hochstrasser, *Proc. Natl. Acad. Sci. USA*, **1995**, 92, 7292–7296.

[44] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff and R. B. Dyer, *Biochemistry*, **1996**, 35, 691–697.

[45] V. Muñoz, P. A. Thompson, J. Hofrichter and W. A. Eaton, *Nature*, **1997**, 390, 196.

[46] P. A. Thompson, W. A. Eaton and J. Hofrichter, *Biochemistry*, **1997**, 36, 9200–9210.

[47] R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff and R. B. Dyer, *Proc. Natl. Acad. Sci. USA*, **1997**, 94, 3709–3713.

[48] U. Mayor, C. M. Johnson, V. Daggett and A. R. Fersht, *Proc. Natl. Acad. Sci. USA*, **2000**, 97, 13518–13522.

[49] C. D. Snow, L. Qiu, D. Du, F. Gai, S. J. Hagen and V. S. Pande, *Proc. Natl. Acad. Sci. USA*, **2004**, 101, 4077–4082.

[50] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton and J. Hofrichter, *J. Mol. Biol.*, **2006**, 359, 546–553.

[51] J. Chen, D. L. Rempel and M. L. Gross, *J. Am. Chem. Soc.*, **2010**, 132, 15502–15504.

[52] C. M. Davis, S. Xiao, D. P. Raleigh and R. B. Dyer, *J. Am. Chem. Soc.*, **2012**, 134, 14476–14482.

[53] X. Zhang and A. Tokmakoff, *Biophys. J.*, **2018**, 114, 209a.

[54] P. L. Privalov, *Crit. Rev. Biochem. Mol. Biol.*, **1990**, 25, 281–306.

[55] J. Zhang, X. Peng, A. Jonas and J. Jonas, *Biochemistry*, **1995**, 34, 8631–8641.

[56] J. Sabelko, J. Ervin and M. Gruebele, *J. Phys. Chem. B*, **1998**, 102, 1806–1819.

[57] K. M. Pryse, T. G. Bruckman, B. W. Maxfield and E. L. Elson, *Biochemistry*, **1992**, 31, 5127–5136.

[58] G. Desai, G. Panick, M. Zein, R. Winter and C. A. Royer, *J. Mol. Biol.*, **1999**, 288, 461–475.

[59] J. Woenckhaus, R. Köhling, P. Thiyagarajan, K. C. Littrell, S. Seifert, C. A. Royer and R. Winter, *Biophys. J.*, **2001**, 80, 1518–1523.

[60] T. R. Alderson, C. Charlier, D. A. Torchia, P. Anfinrud and A. Bax, *J. Am. Chem. Soc.*, **2017**, 139, 11036–11039.

[61] X. Peng, J. Jonas and J. L. Silva, *Proc. Natl. Acad. Sci. USA*, **1993**, 90, 1776–1780.

[62] M. W. Lassalle, H. Yamada and K. Akasaka, *J. Mol. Biol.*, **2000**, 298, 293–302.

[63] T. Pascher, J. P. Chesick, J. R. Winkler and H. B. Gray, *Science*, **1996**, 271, 1558–1560.

[64] E. Chen, P. Wittung-Stafshede and D. S. Kliger, *J. Am. Chem. Soc.*, **1999**, 121, 3811–3817.

[65] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. Iben, T. B. Sauke, E. Shyamsunder and R. D. Young, *Proc. Natl. Acad. Sci. USA*, **1985**, 82, 5000–5004.

[66] C. M. Jones, E. R. Henry, Y. Hu, C.-K. Chan, S. D. Luck, A. Bhuyan, H. Roder, J. Hofrichter and W. A. Eaton, *Proc. Natl. Acad. Sci. USA*, **1993**, 90, 11860–11864.

[67] K. C. Hansen, R. S. Rock, R. W. Larsen and S. I. Chan, *J. Am. Chem. Soc.*, **2000**, 122, 11567–11568.

[68] L. Redecke, S. Binder, M. I. Y. Elmallah, R. Broadbent, C. Tilkorn, B. Schulz, P. May, A. Goos, A. Eich, M. Rübhausen and C. Betzel, *Free Radic. Biol. Med.*, **2009**, 46, 1353–1361.

[69] T. E. Schrader, W. J. Schreier, T. Cordes, F. O. Koller, G. Babitzki, R. Denschlag, C. Renner, M. Löweneck, S.-L. Dong, L. Moroder, P. Tavan and W. Zinth, *Proc. Natl. Acad. Sci. USA*, **2007**, 104, 15729–15734.

[70] A. A. Beharry and G. A. Woolley, *Chem. Soc. Rev.*, **2011**, 40, 4422–4437.

[71] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez and H. E. Gaub, *Science*, **1997**, 276, 1109–1112.

[72] J. M. Fernandez and H. Li, *Science*, **2004**, 303, 1674–1678.

[73] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier and C. Bustamante, *Science*, **1997**, 276, 1112–1116.

[74] E. A. Shank, C. Cecconi, J. W. Dill, S. Marqusee and C. Bustamante, *Nature*, **2010**, 465, 637.

[75] C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips and V. R. Sarma, *Nature*, **1965**, 206, 757.

[76] G. Kartha, J. Bello and D. Harker, *Nature*, **1967**, 213, 862.

[77] H. W. Wyckoff, K. D. Hardman, N. M. Allewell, T. Inagami, L. N. Johnson and F. M. Richards, *J. Biol. Chem.*, **1967**, 242, 3984–3988.

[78] Y. Shi, *Cell*, **2014**, 159, 995–1014.

[79] D. J. Segel, A. Bachmann, J. Hofrichter, K. O. Hodgson, S. Doniach and T. Kiefhaber, *J. Mol. Biol.*, **1999**, 288, 489–499.

[80] K. W. Plaxco, I. S. Millett, D. J. Segel, S. Doniach and D. Baker, *Nat. Struct. Mol. Biol.*, **1999**, 6, 554.

[81] L. Pollack, M. W. Tate, A. C. Finnefrock, C. Kalidas, S. Trotter, N. C. Darnton, L. Lurio, R. H. Austin, C. A. Batt, S. M. Gruner and S. G. J. Mochrie, *Phys. Rev. Lett.*, **2001**, 86, 4962.

[82] H. D. T. Mertens and D. I. Svergun, *J. Struct. Biol.*, **2010**, 172, 128–141.

[83] A. G. Kikhney and D. I. Svergun, *FEBS Lett.*, **2015**, 589, 2570–2577.

[84] K. Wuthrich, *NMR of proteins and nucleic acids*, Wiley, **1986**.

[85] H. J. Dyson and P. E. Wright, *Chem. Rev.*, **2004**, 104, 3607–3622.

[86] A. Mittermaier and L. E. Kay, *Science*, **2006**, 312, 224–228.

[87] A. E. Smith, Z. Zhang, G. J. Pielak and C. Li, *Curr. Opin. Struct. Biol.*, **2015**, 30, 7–16.

[88] D. S. Wishart and B. D. Sykes, in *Methods Enzymol.*, volume 239, Elsevier, **1994**, pp. 363–392.

[89] J. K. Myers and T. G. Oas, *Annu. Rev. Biochem.*, **2002**, 71, 783–815.

[90] P. Neudecker, P. Lundström and L. E. Kay, *Biophys. J.*, **2009**, 96, 2045–2054.

[91] K. H. Gardner and L. E. Kay, *Annu. Rev. Biophys. Biomol. Struct.*, **1998**, 27, 357–406.

[92] D. Marion and K. Wüthrich, *Biochem. Biophys. Res. Commun.*, **1983**, 113, 967–974.

[93] D. R. Shortle, *Curr. Opin. Struct. Biol.*, **1996**, 6, 24–30.

[94] S. W. Englander and L. Mayne, *Annu. Rev. Biophys. Biomol. Struct.*, **1992**, 21, 243–265.

[95] S. T. Gladwin and P. A. Evans, *Fold. Des.*, **1996**, 1, 407–417.

[96] K. Kuwata, R. Shastry, H. Cheng, M. Hoshino, C. A. Batt, Y. Goto and H. Roder, *Nat. Struct. Mol. Biol.*, **2001**, 8, 151.

[97] C. Nishimura, H. J. Dyson and P. E. Wright, *J. Mol. Biol.*, **2002**, 322, 483–489.

[98] N. Tjandra and A. Bax, *Science*, **1997**, 278, 1111–1114.

[99] J.-C. Hus, D. Marion and M. Blackledge, *J. Am. Chem. Soc.*, **2001**, 123, 1541–1542.

[100] R. Mohana-Borges, N. K. Goto, G. J. A. Kroon, H. J. Dyson and P. E. Wright, *J. Mol. Biol.*, **2004**, 340, 1131–1142.

[101] M. R. Jensen, R. W. H. Ruigrok and M. Blackledge, *Curr. Opin. Struct. Biol.*, **2013**, 23, 426–435.

[102] N. J. Greenfield and G. D. Fasman, *Biochemistry*, **1969**, 8, 4108–4116.

[103] S. M. Kelly and N. C. Price, *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.*, **1997**, 1338, 161–185.

[104] N. J. Greenfield, *TrAC, Trends Anal. Chem.*, **1999**, 18, 236–244.

[105] S. M. Kelly, T. J. Jess and N. C. Price, *Biochim. Biophys. Acta, Proteins Proteomics*, **2005**, 1751, 119–139.

[106] L. Whitmore and B. A. Wallace, *Biopolymers*, **2008**, 89, 392–400.

[107] N. J. Greenfield, *Nat. Protoc.*, **2006**, 1, 2876.

[108] S. W. Provencher and J. Gloeckner, *Biochemistry*, **1981**, 20, 33–37.

[109] S. Vuilleumier, J. Sancho, R. Loewenthal and A. R. Fersht, *Biochemistry*, **1993**, 32, 10303–10313.

[110] N. Sreerama, M. C. Manning, M. E. Powers, J.-X. Zhang, D. P. Goldenberg and R. W. Woody, *Biochemistry*, **1999**, 38, 10814–10822.

[111] S. Benjwal, S. Verma, K.-H. Röhm and O. Gursky, *Protein Sci.*, **2006**, 15, 635–639.

[112] L. R. McLean and A. Balasubramaniam, *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.*, **1992**, 1122, 317–320.

[113] C. Goldsbury, K. Goldie, J. Pellaud, J. Seelig, P. Frey, S. A. Müller, J. Kistler, G. J. S. Cooper and U. Aebi, *J. Struct. Biol.*, **2000**, 130, 352–362.

[114] D. Kurouski, R. K. Dukor, X. Lu, L. A. Nafie and I. K. Lednev, *Biophys. J.*, **2012**, 103, 522–531.

[115] C. J. Barrow, A. Yasuda, P. T. M. Kenny and M. G. Zagorski, *J. Mol. Biol.*, **1992**, 225, 1075–1093.

[116] E. Terzi, G. Hoelzemann and J. Seelig, *Biochemistry*, **1994**, 33, 1345–1350.

[117] P. K. Mandal and J. W. Pettegrew, *Neurochem. Res.*, **2004**, 29, 2267–2272.

[118] R. B. Dyer, F. Gai, W. H. Woodruff, R. Gilmanshin and R. H. Callender, *Acc. Chem. Res.*, **1998**, 31, 709–716.

[119] A. Barth, *Biochim. Biophys. Acta, Bioenerg.*, **2007**, 1767, 1073–1101.

[120] W. K. Surewicz, H. H. Mantsch and D. Chapman, *Biochemistry*, **1993**, 32, 389–394.

[121] J. Kong and S. Yu, *Acta Biochim. Biophys. Sin.*, **2007**, 39, 549–559.

[122] H. Yang, S. Yang, J. Kong, A. Dong and S. Yu, *Nat. Protoc.*, **2015**, 10, 382.

[123] A. Ghosh, J. S. Ostrander and M. T. Zanni, *Chem. Rev.*, **2017**, 117, 10726–10759.

[124] J. Seo, W. Hoffmann, S. Warnke, X. Huang, S. Gewinner, W. Schöllkopf, M. T. Bowers, G. von Helden and K. Pagel, *Nat. Chem.*, **2017**, 9, 39.

[125] S. J. Roeters, A. Iyer, G. Pletikapić, V. Kogan, V. Subramaniam and S. Woutersen, *Sci. Rep.*, **2017**, 7, 41051.

[126] G. J. Thomas Jr, *Annu. Rev. Biophys. Biomol. Struct.*, **1999**, 28, 1–27.

[127] R. Tuma, *J. Raman Spectrosc.*, **2005**, 36, 307–319.

[128] J. M. Beechem and L. Brand, *Annu. Rev. Biochem.*, **1985**, 54, 43–71.

[129] C. A. Royer, *Chem. Rev.*, **2006**, 106, 1769–1784.

[130] S. Khorasanizadeh, I. D. Peters and H. Roder, *Nat. Struct. Mol. Biol.*, **1996**, 3, 193.

[131] L. Qiu, S. A. Pabit, A. E. Roitberg and S. J. Hagen, *J. Am. Chem. Soc.*, **2002**, 124, 12952–12953.

[132] A. Hawe, M. Sutter and W. Jiskoot, *Pharm. Res.*, **2008**, 25, 1487–1499.

[133] R. M. Clegg, *Curr. Opin. Biotechnol.*, **1995**, 6, 103–110.

[134] E. Haustein and P. Schwille, *Methods*, **2003**, 29, 153–166.

[135] X. Michalet, S. Weiss and M. Jäger, *Chem. Rev.*, **2006**, 106, 1785–1813.

[136] B. Schuler and W. A. Eaton, *Curr. Opin. Struct. Biol.*, **2008**, 18, 16–26.

[137] H. S. Chung, K. McHale, J. M. Louis and W. A. Eaton, *Science*, **2012**, 335, 981–984.

[138] R. Diamond, *Acta Crystallogr.*, **1966**, 21, 253–266.

[139] M. Levitt and S. Lifson, *J. Mol. Biol.*, **1969**, 46, 269–279.

[140] M. Levitt, *J. Mol. Biol.*, **1974**, 82, 393–420.

[141] B. R. Gelin and M. Karplus, *Proc. Natl. Acad. Sci. USA*, **1977**, 74, 801–805.

[142] B. R. Gelin and M. Karplus, *Proc. Natl. Acad. Sci. USA*, **1975**, 72, 2002–2006.

[143] R. Hetzel, K. Wüthrich, J. Deisenhofer and R. Huber, *Biophys. Struct. Mech.*, **1976**, 2, 159–180.

[144] K. D. Gibson and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **1969**, 63, 9–15.

[145] G. Némethy and H. A. Scheraga, *Q. Rev. Biophys.*, **1977**, 10, 239–352.

[146] M. Levitt and A. Warshel, *Nature*, **1975**, 253, 694.

[147] U. H. E. Hansmann and Y. Okamoto, *Curr. Opin. Struct. Biol.*, **1999**, 9, 177–183.

[148] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, **1953**, 21, 1087–1092.

[149] J. Skolnick and A. Kolinski, *Science*, **1990**, 250, 1121–1125.

[150] E. Shakhnovich, G. Farztdinov, A. M. Gutin and M. Karplus, *Phys. Rev. Lett.*, **1991**, 67, 1665.

[151] A. Kolinski and J. Skolnick, *Proteins: Struct., Funct., Bioinf.*, **1994**, 18, 338–352.

[152] M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **1994**, 98, 4940–4948.

[153] N. D. Socci and J. N. Onuchic, *J. Chem. Phys.*, **1994**, 101, 1519–1528.

[154] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, volume 1, Elsevier, **2001**.

[155] M. Karplus and J. A. McCammon, *Nat. Struct. Mol. Biol.*, **2002**, 9, 646.

[156] T. Hansson, C. Oostenbrink and W. van Gunsteren, *Curr. Opin. Struct. Biol.*, **2002**, 12, 190–196.

[157] M. Karplus and J. Kuriyan, *Proc. Natl. Acad. Sci. USA*, **2005**, 102, 6679–6685.

[158] J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, **1977**, 267, 585–590.

[159] M. Levitt, *Nature*, **1981**, 294, 379.

[160] S. H. Northrup, M. R. Pear, J. D. Morgan, J. A. McCammon and M. Karplus, *J. Mol. Biol.*, **1981**, 153, 1087–1109.

[161] M. Karplus and G. A. Petsko, *Nature*, **1990**, 347, 631.

[162] V. Daggett and M. Levitt, *Curr. Opin. Struct. Biol.*, **1994**, 4, 291–295.

[163] A. Caflisch and M. Karplus, *Proc. Natl. Acad. Sci. USA*, **1994**, 91, 1746–1750.

[164] M. Karplus and A. Šali, *Curr. Opin. Struct. Biol.*, **1995**, 5, 58–73.

[165] S. H. Northrup, M. R. Pear, C.-Y. Lee, J. A. McCammon and M. Karplus, *Proc. Natl. Acad. Sci. USA*, **1982**, 79, 4035–4039.

[166] A. Warshel, *Proc. Natl. Acad. Sci. USA*, **1984**, 81, 444–448.

[167] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *J. Comput. Chem.*, **1992**, 13, 1011–1021.

[168] D. J. Tobias, J. E. Mertz and C. L. Brooks III, *Biochemistry*, **1991**, 30, 6054–6058.

[169] D. J. Tobias and C. L. Brooks III, *Biochemistry*, **1991**, 30, 6059–6070.

[170] D. J. Tobias, S. F. Sneddon and C. L. Brooks III, *J. Mol. Biol.*, **1992**, 227, 1244–1252.

[171] C. L. Brooks III and L. Nilsson, *J. Am. Chem. Soc.*, **1993**, 115, 11034–11035.

[172] W. S. Young and C. L. Brooks III, *J. Mol. Biol.*, **1996**, 259, 560–572.

[173] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. USA*, **2002**, 99, 12562–12566.

[174] A. Laio and F. L. Gervasio, *Reports Prog. Phys.*, **2008**, 71, 126601.

[175] F. Baftizadeh, P. Cossio, F. Pietrucci and A. Laio, *Curr. Phys. Chem.*, **2012**, 2, 79–91.

[176] M. Balsera, S. Stepaniants, S. Izrailev, Y. Oono and K. Schulten, *Biophys. J.*, **1997**, 73, 1281–1287.

[177] J. R. Gullingsrud, R. Braun and K. Schulten, *J. Comput. Phys.*, **1999**, 151, 190–211.

[178] J. Schlitter, M. Engels and P. Krüger, *J. Mol. Graph.*, **1994**, 12, 84–89.

[179] B. A. Berg and T. Neuhaus, *Phys. Lett. B*, **1991**, 267, 249–253.

[180] N. Nakajima, H. Nakamura and A. Kidera, *J. Phys. Chem. B*, **1997**, 101, 817–824.

[181] F. Wang and D. P. Landau, *Phys. Rev. Lett.*, **2001**, 86, 2050.

[182] A. K. Faradjian and R. Elber, *J. Chem. Phys.*, **2004**, 120, 10880–10889.

[183] A. M. A. West, R. Elber and D. Shalloway, *J. Chem. Phys.*, **2007**, 126, 04B608.

[184] C. Dellago, P. G. Bolhuis, F. S. Csajka and D. Chandler, *J. Chem. Phys.*, **1998**, 108, 1964–1977.

[185] P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, *Annu. Rev. Phys. Chem.*, **2002**, 53, 291–318.

[186] C. Dellago and P. G. Bolhuis, in *Adv. Comput. Simul. Approaches Soft Matter Sci. III*, Springer, **2009**, pp. 167–233.

[187] D. Moroni, T. S. van Erp and P. G. Bolhuis, *Physica A*, **2004**, 340, 395–401.

[188] P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA*, **2003**, 100, 12129–12134.

[189] U. H. E. Hansmann, *Chem. Phys. Lett.*, **1997**, 281, 140–150.

[190] Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, **1999**, 314, 141–151.

[191] K. Y. Sanbonmatsu and A. E. Garcia, *Proteins: Struct., Funct., Bioinf.*, **2002**, 46, 225–234.

[192] A. K. Felts, Y. Harano, E. Gallicchio and R. M. Levy, *Proteins: Struct., Funct., Bioinf.*, **2004**, 56, 310–321.

[193] P. H. Nguyen, G. Stock, E. Mittag, C. Hu and M. S. Li, *Proteins: Struct., Funct., Bioinf.*, **2005**, 61, 795–808.

[194] M. Cecchini, F. Rao, M. Seeber and A. Caflisch, *J. Chem. Phys.*, **2004**, 121, 10748–10756.

[195] A. Baumketner and J.-E. Shea, *J. Mol. Biol.*, **2007**, 366, 275–285.

[196] S. Patel, E. Vierling and F. Tama, *Biophys. J.*, **2014**, 106, 2644–2655.

[197] Y. M. Rhee and V. S. Pande, *Biophys. J.*, **2003**, 84, 775–786.

[198] C. Schütte, A. Fischer, W. Huisinga and P. Deuflhard, *J. Comput. Phys.*, **1999**, 151, 146–168.

[199] N. Singhal, C. D. Snow and V. S. Pande, *J. Chem. Phys.*, **2004**, 121, 415–425.

[200] W. C. Swope, J. W. Pitera and F. Suits, *J. Phys. Chem. B*, **2004**, 108, 6571–6581.

[201] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte and F. Noé, *J. Chem. Phys.*, **2011**, 134.

[202] V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, **2010**, 52, 99–105.

[203] T. J. Lane, D. Shukla, K. A. Beauchamp and V. S. Pande, *Curr. Opin. Struct. Biol.*, **2013**, 23, 58–65.

[204] J. D. Chodera and F. Noé, *Curr. Opin. Struct. Biol.*, **2014**, 25, 135–144.

[205] B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.*, **2018**, 140, 2386–2396.

[206] M. Shirts and V. S. Pande, *Science*, **2000**, 290, 1903–1904.

[207] G. R. Bowman and V. S. Pande, *Proc. Natl. Acad. Sci. USA*, **2010**, 107, 10890–10895.

[208] V. A. Voelz, G. R. Bowman, K. Beauchamp and V. S. Pande, *J. Am. Chem. Soc.*, **2010**, 132, 1526–1528.

[209] G. R. Bowman, V. A. Voelz and V. S. Pande, *J. Am. Chem. Soc.*, **2011**, 133, 664–667.

[210] D. J. Wales, *Energy landscapes: Applications to clusters, biomolecules and glasses*, Cambridge University Press, Cambridge, U.K., **2003**.

[211] D. J. Wales, *Phys. Biol.*, **2005**, 2, S86.

[212] D. J. Wales, *Curr. Opin. Struct. Biol.*, **2010**, 20, 3–10.

[213] A. G. Cochran, N. J. Skelton and M. A. Starovasnik, *Proc. Natl. Acad. Sci. USA*, **2001**, 98, 5578–5583.

[214] G. A. Belogurov, M. N. Vassylyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassylyev and I. Artsimovitch, *Mol. Cell*, **2007**, 26, 117–129.

[215] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch and P. Rösch, *Cell*, **2012**, 150, 291–303.

[216] V. Wray, D. Mertins, M. Kiess, P. Henklein, W. Trowitzsch-Kienast and U. Schubert, *Biochemistry*, **1998**, 37, 8527–8538.

[217] N. S. Ostlund and A. Szabo, *Modern Quantum Chemistry: Introduction to advanced electronic structure theory*, Dover Publications Inc New edition edn, **1996**.

[218] M. Born and R. Oppenheimer, *Ann. Phys.*, **1927**, 389, 457–484.

[219] P. K. Weiner and P. A. Kollman, *J. Comput. Chem.*, **1981**, 2, 287–303.

[220] S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case, *J. Comput. Chem.*, **1986**, 7, 230–252.

[221] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **1995**, 117, 5179–5197.

[222] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, **1983**, 4, 187–217.

[223] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, **1998**, 102, 3586–3616.

[224] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, **1988**, 110, 1657–1666.

[225] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger and W. F. van Gunsteren, *J. Phys. Chem. A*, **1999**, 103, 3596–3607.

[226] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, **2006**, 65, 712–725.

[227] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.*, **2010**, 78, 1950–1958.

[228] D. S. Cerutti, W. C. Swope, J. E. Rice and D. A. Case, *J. Chem. Theory Comput.*, **2014**, 10, 4515–4534.

[229] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, **2015**, 11, 3696–3713.

[230] S. R. Edinger, C. Cortis, P. S. Shenkin and R. A. Friesner, *J. Phys. Chem. B*, **1997**, 101, 1190–1197.

[231] D. Eisenberg and A. D. McLachlan, *Nature*, **1986**, 319, 199–203.

[232] T. Ooi, M. Oobatake, G. Nemethy and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **1987**, 84, 3086–3090.

[233] J. Tomasi and M. Persico, *Chem. Rev.*, **1994**, 94, 2027–2094.

[234] W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.*, **1990**, 112, 6127–6129.

[235] D. Bashford and D. A. Case, *Annu. Rev. Phys. Chem.*, **2000**, 51, 129–152.

[236] G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem.*, **1996**, 100, 19824–19839.

[237] A. Onufriev, D. Bashford and D. A. Case, *Proteins: Struct., Funct., Bioinf.*, **2004**, 55, 383–394.

[238] H. Nguyen, D. R. Roe and C. Simmerling, *J. Chem. Theory Comput.*, **2013**, 9, 2020–2034.

[239] Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **1987**, 84, 6611–6615.

[240] D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, **1997**, 101, 5111–5116.

[241] D. J. Wales, *J. Chem. Soc. Faraday Trans.*, **1992**, 88, 653–657.

[242] J. Nocedal, *Math. Comput.*, **1980**, 35, 773–782.

[243] D. C. Liu and J. Nocedal, *Math. Program.*, **1989**, 45, 503–528.

[244] C. G. Broyden, *IMA J. Appl. Math.*, **1970**, 6, 76–90.

[245] R. Fletcher, *Comput. J.*, **1970**, 13, 317–322.

[246] D. Goldfarb, *Math. Comput.*, **1970**, 24, 23–26.

[247] D. F. Shanno, *Math. Comput.*, **1970**, 24, 647–656.

[248] D. J. Wales, GMIN: A program for finding global minima and calculating thermodynamic properties, http://www-wales.ch.cam.ac.uk/GMIN.

[249] D. J. Wales, *Mol. Phys.*, **2002**, 100, 3285–3305.

[250] D. J. Wales, *Mol. Phys.*, **2004**, 102, 891–908.

[251] D. A. Evans and D. J. Wales, *J. Chem. Phys.*, **2004**, 121, 1080–1090.

[252] S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, **2004**, 120, 2082–2094.

[253] G. Henkelman and H. Jónsson, *J. Chem. Phys.*, **1999**, 111, 7010–7022.

[254] G. Henkelman and H. Jónsson, *J. Chem. Phys.*, **2000**, 113, 9978–9985.

[255] L. J. Munro and D. J. Wales, *Phys. Rev. B*, **1999**, 59, 3969.

[256] Y. Kumeda, D. J. Wales and L. J. Munro, *Chem. Phys. Lett.*, **2001**, 341, 185–194.

[257] D. J. Wales, J. M. Carr, M. Khalili, V. K. de Souza, B. Strodel and C. S. Whittleston, in *Proteins Energy, Heat Signal Flow*, Computation in chemistry, CRC Press, **2009**, p. 315.

[258] J. M. Carr, S. A. Trygubenko and D. J. Wales, *J. Chem. Phys.*, **2005**, 122, 234903.

[259] E. W. Dijkstra, *Numer. Math.*, **1959**, 1, 269–271.

[260] B. Strodel, C. S. Whittleston and D. J. Wales, *J. Am. Chem. Soc.*, **2007**, 129, 16005–16014.

[261] B. Strodel and D. J. Wales, *Chem. Phys. Lett.*, **2008**, 466, 105–115.

[262] D. J. Wales, *J. Chem. Phys.*, **2009**, 130, 204111.

[263] D. J. Wales, *Int. Rev. Phys. Chem.*, **2006**, 25, 237–282.

[264] J. M. Carr and D. J. Wales, *J. Phys. Chem. B*, **2008**, 112, 8760–8769.

[265] O. M. Becker and M. Karplus, *J. Chem. Phys.*, **1997**, 106, 1495–1517.

[266] D. J. Wales, M. A. Miller and T. R. Walsh, *Nature*, **1998**, 394, 758–760.

[267] S. V. Krivov and M. Karplus, *J. Chem. Phys.*, **2002**, 117, 10894–10903.

[268] D. A. Evans and D. J. Wales, *J. Chem. Phys.*, **2003**, 118, 3891–3897.

[269] M. A. Miller and D. J. Wales, *J. Chem. Phys.*, **1999**, 111, 6610–6616.

[270] C. S. Whittleston, *Energy landscapes of biological systems*, Ph.D. thesis, University of Cambridge, **2012**.

[271] D. Chakraborty, R. Collepardo-Guevara and D. J. Wales, *J. Am. Chem. Soc.*, **2014**, 136, 18052–18061.

[272] K. Röder and D. J. Wales, *J. Am. Chem. Soc.*, **2018**, 140, 4018–4027.

[273] K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl, *Annu. Rev. Biophys.*, **2008**, 37, 289.

[274] K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, *Science*, **2011**, 334, 517–520.

[275] K. A. Dill and J. L. MacCallum, *Science*, **2012**, 338, 1042–1046.

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

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

[278] W. Jiang, J. C. Phillips, L. Huang, M. Fajer, Y. Meng, J. C. Gumbart, Y. Luo, K. Schulten and B. Roux, *Comput. Phys. Commun.*, **2014**, 185, 908–916.

[279] S. Piana, K. Lindorff-Larsen and D. E. Shaw, *Biophys. J.*, **2011**, 100, L47–L49.

[280] M. J. Robertson, J. Tirado-Rives and W. L. Jorgensen, *J. Chem. Theory Comput.*, **2015**, 11, 3499–3509.

[281] R. B. Best, *Curr. Opin. Struct. Biol.*, **2012**, 22, 52–61.

[282] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu and D. E. Shaw, *Annu. Rev. Biophys.*, **2012**, 41, 429–452.

[283] C. Clementi, *Curr. Opin. Struct. Biol.*, **2008**, 18, 10–15.

[284] M. G. Saunders and G. A. Voth, *Annu. Rev. Biophys.*, **2013**, 42, 73–93.

[285] 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.

[286] A. Warshel and M. Levitt, *J. Mol. Biol.*, **1976**, 106, 421–437.

[287] I. Bahar, A. R. Atilgan and B. Erman, *Fold Des.*, **1997**, 2, 173–181.

[288] P. Derreumaux, *J. Chem. Phys.*, **1999**, 111, 2301–2310.

[289] N.-V. Buchete, J. E. Straub and D. Thirumalai, *J. Chem. Phys.*, **2003**, 118, 7658–7671.

[290] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura and D. Baker, in *Numer. Comput. Methods, Part D*, *Methods in Enzymology*, volume 383, Academic Press, San Diego, CA, **2004**, pp. 66–93.

[291] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S.-J. Marrink, *J. Chem. Theory Comput.*, **2008**, 4, 819–834.

[292] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, **2013**, 9, 687–697.

[293] S. Abeln, M. Vendruscolo, C. M. Dobson and D. Frenkel, *PLoS One*, **2014**, 9, 1–8.

[294] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin and I. Bahar, *Biophys. J.*, **2001**, 80, 505–515.

[295] H. Abe and N. Go, *Biopolymers*, **1981**, 20, 1013–1031.

[296] A. Ahmed and H. Gohlke, *Proteins: Struct., Funct., Bioinf.*, **2006**, 63, 1038–1051.

[297] A. P. Heath, L. E. Kavraki and C. Clementi, *Proteins: Struct., Funct., Bioinf.*, **2007**, 68, 646–661.

[298] R. D. Hills Jr, L. Lu and G. A. Voth, *PLoS Comput. Biol.*, **2010**, 6, 1–12.

[299] D. J. Wales, *Phil. Trans. R. Soc. Lond. A*, **2005**, 363, 357–377.

[300] D. Chakrabarti and D. J. Wales, *Phys. Chem. Chem. Phys.*, **2009**, 11, 1970–1976.

[301] H. Kusumaatmaja, C. S. Whittleston and D. J. Wales, *J. Chem. Theory Comput.*, **2012**, 8, 5159–5165.

[302] V. Rühle, H. Kusumaatmaja, D. Chakrabarti and D. J. Wales, *J. Chem. Theory Comput.*, **2013**, 9, 4026–4034.

[303] H. Gohlke and M. F. Thorpe, *Biophys. J.*, **2006**, 91, 2115–2120.

[304] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner and G. A. Voth, *Biophys. J.*, **2008**, 95, 5073–5083.

[305] J. Srinivasan, M. W. Trevathan, P. Beroza and D. A. Case, *Theor. Chem. Acc.*, **1999**, 101, 426–434.

[306] E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. N. Fejer and D. J. Wales, *J. Comput. Chem.*, **2010**, 31, 1402–1409.

[307] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, **1983**, 79, 926–935.

[308] S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, **2010**, 12, 8478–8490.

[309] A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, C. C. Pantelides, S. L. Price, P. T. A. Galek, G. M. Day and A. J. Cruz-Cabeza, *Int. J. Pharm.*, **2011**, 418, 168–178.

[310] D. J. Wales and I. Ohmine, *J. Chem. Phys.*, **1993**, 98, 7257–7268.

[311] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker, *J. Mol. Biol.*, **2003**, 331, 281–299.

[312] I. A. Vakser, *Biophys. J.*, **2014**, 107, 1785–1793.

[313] B. M. Hespenheide, D. J. Jacobs and M. F. Thorpe, *J. Phys. Condens. Matter*, **2004**, 16, S5055.

[314] R. W. Brockett, in *Math. theory networks Syst.*, Springer, pp. 120–129.

[315] R. M. Murray, Z. Li, S. S. Sastry and S. S. Sastry, *A mathematical introduction to robotic manipulation*, CRC press, **1994**.

[316] A. Kitao and N. Go, *Curr. Opin. Struct. Biol.*, **1999**, 9, 164–169.

[317] O. F. Lange and H. Grubmüller, *J. Phys. Chem. B*, **2006**, 110, 22842–22852.

[318] D. J. Jacobs, A. J. Rader, L. A. Kuhn and M. F. Thorpe, *Proteins: Struct., Funct., Bioinf.*, **2001**, 44, 150–165.

[319] M. F. Thorpe, M. Lei, A. J. Rader, D. J. Jacobs and L. A. Kuhn, *J. Mol. Graph.*, **2001**, 19, 60–69.

[320] J. R. Costa and S. N. Yaliraki, *J. Phys. Chem. B*, **2006**, 110, 18981–18988.

[321] D. J. Wales, OPTIM: A program for optimising geometries and calculating pathways, http://www-wales.ch.cam.ac.uk/OPTIM.

[322] 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.

[323] W. Y. Yang and M. Gruebele, *J. Am. Chem. Soc.*, **2004**, 126, 7758–7759.

[324] Y. Levy and O. M. Becker, *Phys. Rev. Lett.*, **1998**, 81, 1126–1129.

[325] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, in *Kdd*, volume 96, pp. 226–231.

[326] D. R. Roe and T. E. Cheatham III, *J. Chem. Theory Comput.*, **2013**, 9, 3084–3095.

[327] A. R. Dinner, T. Lazaridis and M. Karplus, *Proc. Natl. Acad. Sci. USA*, **1999**, 96, 9068–9073.

[328] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin and R. Tonini, *J. Biol. Chem.*, **2004**, 279, 9298–9305.

[329] X. Luo, Z. Tang, G. Xia, K. Wassmann, T. Matsumoto, J. Rizo and H. Yu, *Nat. Struct. Mol. Biol.*, **2004**, 11, 338.

[330] A. Andreeva and A. G. Murzin, *Curr. Opin. Struct. Biol.*, **2006**, 16, 399–408.

[331] S. Meier and S. Özbek, *BioEssays*, **2007**, 29, 1095–1104.

[332] A. G. Murzin, *Science*, **2008**, 320, 1725–1726.

[333] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron and B. F. Volkman, *Proc. Natl. Acad. Sci. USA*, **2008**, 105, 5057–5062.

[334] N. Tokuriki and D. S. Tawfik, *Science*, **2009**, 324, 203–207.

[335] P. N. Bryan and J. Orban, *Curr. Opin. Struct. Biol.*, **2010**, 20, 482–488.

[336] M. J. A. Bailey, V. Koronakis, T. Schmoll and C. Hughes, *Mol. Microbiol.*, **1992**, 6, 1003–1012.

[337] M. J. A. Bailey, C. Hughes and V. Koronakis, *Mol. Microbiol.*, **1997**, 26, 845–851.

[338] M. J. A. Bailey, C. Hughes and V. Koronakis, *Mol. Gen. Genet.*, **2000**, 262, 1052–1059.

[339] G. A. Belogurov, R. A. Mooney, V. Svetlov, R. Landick and I. Artsimovitch, *EMBO J.*, **2009**, 28, 112–122.

[340] S. H. Knauer, I. Artsimovitch and P. Rösch, *Cell Cycle*, **2012**, 11, 4289–4290.

[341] S. K. Tomar, S. H. Knauer, M. NandyMazumdar, P. Rösch and I. Artsimovitch, *Nucleic Acids Res.*, **2013**, 41, 10077–10085.

[342] J. B. GC, Y. R. Bhandari, B. S. Gerstman and P. P. Chapagain, *J. Phys. Chem. B*, **2014**, 118, 5101–5108.

[343] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng and H. Jiang, *J. Chem. Theory Comput.*, **2014**, 10, 2255–2264.

[344] N. A. Bernhardt and U. H. E. Hansmann, *J. Phys. Chem. B*, **2018**.

[345] F. Yaar, N. A. Bernhardt and U. H. E. Hansmann, *J. Chem. Phys.*, **2015**, 143, 224102.

[346] N. A. Bernhardt, W. Xi, W. Wang and U. H. E. Hansmann, *J. Chem. Theory Comput.*, **2016**, 12, 5656–5666.

[347] The PyMOL molecular graphics system, version 1.8, **2015**.

[348] R. G. Mantell, C. E. Pitt and D. J. Wales, *J. Chem. Theory Comput.*, **2016**, 12, 6182–6191.

[349] K. Röder, D. J. Wales, K. Roder and D. J. Wales, *J. Chem. Theory Comput.*, **2018**, 14, 4271–4278.

[350] W. Kabsch and C. Sander, *Biopolymers*, **1983**, 22, 2577–2637.

[351] C. A. Ramírez-Sarmiento, J. K. Noel, S. L. Valenzuela and I. Artsimovitch, *PLoS Comput. Biol.*, **2015**, 11, e1004379.

[352] D. J. Wales and J. M. Carr, *J. Chem. Theory Comput.*, **2012**, 8, 5020–5034.

[353] D. Chakraborty and D. J. Wales, *Phys. Chem. Chem. Phys.*, **2017**, 19, 878–892.

[354] P. Tompa, *Trends Biochem. Sci.*, **2012**, 37, 509–516.

[355] F.-X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P. B. Crowley, L. Gierasch and G. J. Pielak, *Chem. Rev.*, **2014**, 114, 6661–6714.

[356] P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, **1999**, 293, 321–331.

[357] V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins: Struct., Funct., Bioinf.*, **2000**, 41, 415–427.

[358] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff and K. W. Hipps, *J. Mol. Graph. Model.*, **2001**, 19, 26–59.

[359] P. Tompa, *Trends Biochem. Sci.*, **2002**, 27, 527–533.

[360] H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Biol.*, **2005**, 6, 197.

[361] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner and C. J. Brown, *Genome Informatics*, **2000**, 11, 161–171.

[362] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, **2004**, 337, 635–645.

[363] A. K. Dunker, I. Silman, V. N. Uversky and J. L. Sussman, *Curr. Opin. Struct. Biol.*, **2008**, 18, 756–764.

[364] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins: Struct., Funct., Bioinf.*, **2001**, 42, 38–48.

[365] S. Vucetic, C. J. Brown, A. K. Dunker and Z. Obradovic, *Proteins: Struct., Funct., Bioinf.*, **2003**, 52, 573–584.

[366] V. N. Uversky and A. K. Dunker, *Biochim. Biophys. Acta, Proteins Proteomics*, **2010**, 1804, 1231–1264.

[367] V. N. Uversky, *Biochim. Biophys. Acta, Proteins Proteomics*, **2013**, 1834, 932–951.

[368] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradović, *Biochemistry*, **2002**, 41, 6573–6582.

[369] A. P. Demchenko, *J. Mol. Recognit.*, **2001**, 14, 42–61.

[370] H. J. Dyson and P. E. Wright, *Curr. Opin. Struct. Biol.*, **2002**, 12, 54–60.

[371] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic and A. K. Dunker, *Nucleic Acids Res.*, **2004**, 32, 1037–1049.

[372] P. E. Wright and H. J. Dyson, *Nat. Rev. Mol. Cell Biol.*, **2015**, 16, 18.

[373] M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, **2011**, 21, 432–440.

[374] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway and P. T. Lansbury, *Biochemistry*, **1996**, 35, 13709–13715.

[375] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović and A. K. Dunker, *J. Mol. Biol.*, **2002**, 323, 573–584.

[376] J. M. R. Baker, R. P. Hudson, V. Kanelis, W.-Y. Choy, P. H. Thibodeau, P. J. Thomas and J. D. Forman-Kay, *Nat. Struct. Mol. Biol.*, **2007**, 14, 738.

[377] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge and A. R. Fersht, *Proc. Natl. Acad. Sci. USA*, **2008**, 105, 5762–5767.

[378] V. N. Uversky, C. J. Oldfield and A. K. Dunker, *Annu. Rev. Biophys.*, **2008**, 37, 215–246.

[379] M. D. Mukrasch, S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow and M. Zweckstetter, *PLoS Biol.*, **2009**, 7, e1000034.

[380] N. E. Davey, G. Travé and T. J. Gibson, *Trends Biochem. Sci.*, **2011**, 36, 159–169.

[381] H. J. Dyson and P. E. Wright, *Methods Enzymol.*, **2001**, 339, 258–270.

[382] D. Eliezer, *Curr. Opin. Struct. Biol.*, **2009**, 19, 23–30.

[383] A. C. M. Ferreon, C. R. Moran, Y. Gambin and A. A. Deniz, *Methods Enzymol.*, **2010**, 472, 179–204.

[384] A. Miyagi, Y. Tsunaka, T. Uchihashi, K. Mayanagi, S. Hirose, K. Morikawa and T. Ando, *ChemPhysChem*, **2008**, 9, 1859–1866.

[385] J.-F. Bodart, J.-M. Wieruszeski, L. Amniai, A. Leroy, I. Landrieu, A. Rousseau-Lescuyer, J.-P. Vilain and G. Lippens, *J. Magn. Reson.*, **2008**, 192, 252–257.

[386] L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith and E. J. Ackerman, *Protein Sci.*, **2001**, 10, 560–571.

[387] S. Rauscher and R. Pomès, *Biochem. Cell Biol.*, **2010**, 88, 269–290.

[388] C. M. Baker and R. B. Best, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **2014**, 4, 182–198.

[389] Z. A. Levine and J.-E. Shea, *Curr. Opin. Struct. Biol.*, **2017**, 43, 95–103.

[390] R. B. Best, *Curr. Opin. Struct. Biol.*, **2017**, 42, 147–154.

[391] M. Vendruscolo, *Curr. Opin. Struct. Biol.*, **2007**, 17, 15–20.

[392] M. Bonomi, G. T. Heller, C. Camilloni and M. Vendruscolo, *Curr. Opin. Struct. Biol.*, **2017**, 42, 106–116.

[393] P. H. Nguyen, M. S. Li and P. Derreumaux, *Phys. Chem. Chem. Phys.*, **2011**, 13, 9778–9788.

[394] R. B. Best and G. Hummer, *J. Phys. Chem. B*, **2009**, 113, 9004–9015.

[395] R. B. Best and J. Mittal, *J. Phys. Chem. B*, **2010**, 114, 14916–14923.

[396] W. Wang, W. Ye, C. Jiang, R. Luo and H. Chen, *Chem. Biol. Drug Des.*, **2014**, 84, 253–269.

[397] F. Palazzesi, M. K. Prakash, M. Bonomi and A. Barducci, *J. Chem. Theory Comput.*, **2015**, 11, 2–7.

[398] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot and H. Grubmüller, *J. Chem. Theory Comput.*, **2015**, 11, 5513–5524.

[399] M. D. Smith, J. S. Rao, E. Segelken and L. Cruz, *J. Chem. Inf. Model.*, **2015**, 55, 2587–2595.

[400] I. Maffucci and A. Contini, *J. Chem. Theory Comput.*, **2016**, 12, 714–727.

[401] C. K. Fisher and C. M. Stultz, *Curr. Opin. Struct. Biol.*, **2011**, 21, 426–431.

[402] M. Schor, A. S. J. S. Mey and C. E. MacPhee, *Biophys. Rev.*, **2016**, 8, 429–439.

[403] S.-H. Chong, P. Chatterjee and S. Ham, *Annu. Rev. Phys. Chem.*, **2017**, 68, 117–134.

[404] K. Ostermeir and M. Zacharias, *Biochim. Biophys. Acta, Proteins Proteomics*, **2013**, 1834, 847–853.

[405] J. Michel and R. Cuchillo, *PLoS One*, **2012**, 7, e41070.

[406] F. Baftizadeh, F. Pietrucci, X. Biarnés and A. Laio, *Phys. Rev. Lett.*, **2013**, 110, 168103.

[407] D. Granata, F. Baftizadeh, J. Habchi, C. Galvagnion, A. De Simone, C. Camilloni, A. Laio and M. Vendruscolo, *Sci. Rep.*, **2015**, 5, 15449.

[408] G. H. Zerze, C. M. Miller, D. Granata and J. Mittal, *J. Chem. Theory Comput.*, **2015**, 11, 2776–2782.

[409] Y.-S. Lin, G. R. Bowman, K. A. Beauchamp and V. S. Pande, *Biophys. J.*, **2012**, 102, 315–324.

[410] Q. Qiao, G. R. Bowman and X. Huang, *J. Am. Chem. Soc.*, **2013**, 135, 16092–16101.

[411] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, *J. Chem. Phys.*, **2013**, 139, 07B604_1.

[412] N. Stanley, S. Esteban-Martín and G. De Fabritiis, *Nat. Commun.*, **2014**, 5, 5272.

[413] A. Garcia-Pino, *Biophys. J.*, **2017**, 112, 317a–318a.

[414] J.-M. Choi, J. Wang, A. S. Holehouse, S. Alberti, A. A. Hyman and R. V. Pappu, *Biophys. J.*, **2018**, 114, 561a.

[415] S. Piana and A. Laio, *J. Phys. Chem. B*, **2007**, 111, 4553–4559.

[416] J. Zhu and W. E. Paul, *Blood*, **2008**, 112, 1557–1569.

[417] M. Wittlich, B. W. Koenig, S. Hoffmann and D. Willbold, *Biochim. Biophys. Acta, Biomembr*, **2007**, 1768, 2949–2960.

[418] M. Wittlich, P. Thiagarajan, B. W. Koenig, R. Hartmann and D. Willbold, *Biochim. Biophys. Acta, Biomembr*, **2010**, 1798, 122–127.

[419] C. Doyle and J. L. Strominger, *Nature*, **1987**, 330, 256.

[420] B. P. Sleckman, A. Peterson, W. K. Jones, J. A. Foran, J. L. Greenstein, B. Seed and S. J. Burakoff, *Nature*, **1987**, 328, 351.

[421] J. M. Turner, M. H. Brodsky, B. A. Irving, S. D. Levin, R. M. Perlmutter and D. R. Littman, *Cell*, **1990**, 60, 755–765.

[422] D. B. Straus and A. Weiss, *Cell*, **1992**, 70, 585–593.

[423] A. C. Chan, M. Iwashima, C. W. Turck and A. Weiss, *Cell*, **1992**, 71, 649–662.

[424] A. G. Dalgleish, P. C. L. Beverley, P. R. Clapham, D. H. Crawford, M. F. Greaves and R. A. Weiss, *Nature*, **1984**, 312, 763.

[425] D. Klatzmann, E. Champagne, S. Chamaret, J. Gruest, D. Guetard, T. Hercend, J.-C. Gluckman and L. Montagnier, *Nature*, **1984**, 312, 767.

[426] P. J. Maddon, A. G. Dalgleish, J. S. McDougal, P. R. Clapham, R. A. Weiss and R. Axel, *Cell*, **1986**, 47, 333–348.

[427] Q. J. Sattentau, A. G. Dalgleish, R. A. Weiss and P. C. Beverley, *Science*, **1986**, 234, 1120–1123.

[428] J. S. McDougal, M. S. Kennedy, J. M. Sligh, S. P. Cort, A. Mawle and J. K. Nicholson, *Science*, **1986**, 231, 382–385.

[429] H. L. Robinson and D. M. Zinkus, *J. Virol.*, **1990**, 64, 4836–4841.

[430] C. D. Pauza, J. E. Galindo and D. D. Richman, *J. Exp. Med.*, **1990**, 172, 1035–1042.

[431] T. M. Ross, A. E. Oran and B. R. Cullen, *Curr. Biol.*, **1999**, 9, 613–621.

[432] K. Levesque, Y.-S. Zhao and É. A. Cohen, *J. Biol. Chem.*, **2003**, 278, 28346–28353.

[433] J. V. Garcia and A. D. Miller, *Nature*, **1991**, 350, 508.

[434] R. L. Willey, F. Maldarelli, M. A. Martin and K. Strebel, *J. Virol.*, **1992**, 66, 7193–7200.

[435] M. Y. Chen, F. Maldarelli, M. K. Karczewski, R. L. Willey and K. Strebel, *J. Virol.*, **1993**, 67, 3877–3884.

[436] M. J. Vincent, N. U. Raja and M. A. Jabbar, *J. Virol.*, **1993**, 67, 5538–5549.

[437] C. Aiken, J. Konner, N. R. Landau, M. E. Lenburg and D. Trono, *Cell*, **1994**, 76, 853–864.

[438] S. J. Anderson, M. Lenburg, N. R. Landau and J. V. Garcia, *J. Virol.*, **1994**, 68, 3092–3101.

[439] X.-J. Yao, J. Friborg, F. Checroune, S. Gratton, F. Boisvert, R. P. Sékaly and E. A. Cohen, *Virology*, **1995**, 209, 615–623.

[440] E. Tiganos, X.-J. Yao, J. Friborg, N. Daniel and E. A. Cohen, *J. Virol.*, **1997**, 71, 4452–4460.

[441] A. Preusser, L. Briese and D. Willbold, *Biochem. Biophys. Res. Commun.*, **2002**, 292, 734–740.

[442] S. K. Singh, L. Möckel, P. Thiagarajan-Rosenkranz, M. Wittlich, D. Willbold and B. W. Koenig, *FEBS J.*, **2012**, 279, 3705–3714.

[443] D. Willbold and P. Rösch, *J. Biomed. Sci.*, **1996**, 3, 435–441.

[444] N. Ahalawat, S. Arora and R. K. Murarka, *J. Phys. Chem. B*, **2015**, 119, 11229–11242.

[445] B. Strodel, J. W. L. Lee, C. S. Whittleston and D. J. Wales, *J. Am. Chem. Soc.*, **2010**, 132, 13300–13312.

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

[447] D. S. Cerutti, J. E. Rice, W. C. Swope and D. A. Case, *J. Phys. Chem. B*, **2013**, 117, 2328–2338.

[448] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, **2004**, 120, 9665–9678.

[449] S. Neal, A. M. Nip, H. Zhang and D. S. Wishart, *J. Biomol. NMR*, **2003**, 26, 215–240.

[450] R. Brüschweiler and D. A. Case, *J. Am. Chem. Soc.*, **1994**, 116, 11199–11200.

[451] K. A. Ball, D. E. Wemmer and T. Head-Gordon, *J. Phys. Chem. B*, **2014**, 118, 6405–6416.

[452] P. W. Kim, Z.-Y. J. Sun, S. C. Blacklow, G. Wagner and M. J. Eck, *Science*, **2003**, 301, 1725–1728.

[453] M. Veillette, A. Désormeaux, H. Medjahed, N.-E. Gharsallah, M. Coutu, J. Baalwa, Y. Guan, G. Lewis, G. Ferrari and B. H. Hahn, *J. Virol.*, **2014**, 88, 2633–2644.

[454] J. F. Arias, L. N. Heyer, B. von Bredow, K. L. Weisgrau, B. Moldt, D. R. Burton, E. G. Rakasz and D. T. Evans, *Proc. Natl. Acad. Sci. USA*, **2014**, 111, 6425–6430.

[455] M. Veillette, M. Coutu, J. Richard, L.-A. Batraville, O. Dagher, N. Bernard, C. Tremblay, D. E. Kaufmann, M. Roger and A. Finzi, *J. Virol.*, **2015**, 89, 545–551.

[456] L. Martin, F. Stricher, D. Missé, F. Sironi, M. Pugnière, P. Barthe, R. Prado-Gotor, I. Freulon, X. Magne and C. Roumestand, *Nat. Biotechnol.*, **2003**, 21, 71.

[457] H. Haim, Z. Si, N. Madani, L. Wang, J. R. Courter, A. Princiotto, A. Kassa, M. DeGrace, K. McGee-Estrada and M. Mefford, *PLoS Pathog.*, **2009**, 5, e1000360.

[458] F. Baleux, L. Loureiro-Morais, Y. Hersant, P. Clayette, F. Arenzana-Seisdedos, D. Bonnaffé and H. Lortat-Jacob, *Nat. Chem. Biol.*, **2009**, 5, 743.

[459] J. Richard, M. Veillette, N. Brassard, S. S. Iyer, M. Roger, L. Martin, M. Pazgier, A. Schön, E. Freire and J.-P. Routy, *Proc. Natl. Acad. Sci. USA*, **2015**, 112, E2687–E2694.

[460] Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y.-Y. J. Van, P. Romero, M. S. Cortese, V. N. Uversky and A. K. Dunker, *Trends Biotechnol.*, **2006**, 24, 435–442.

[461] F. Curreli, Y. Do Kwon, H. Zhang, Y. Yang, D. Scacalossi, P. D. Kwong and A. K. Debnath, *Antimicrob. Agents Chemother.*, **2014**, 58, 5478–5491.

[462] M. Arkin, *Curr. Opin. Chem. Biol.*, **2005**, 9, 317–324.

[463] D. C. Fry and L. T. Vassilev, *J. Mol. Med.*, **2005**, 83, 955–963.

[464] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan and B. Towles, in *Proc. Conf. High Perform. Comput. Networking, Storage Anal.*, pp. 1–11.

[465] D. P. Frueh, A. C. Goodrich, S. H. Mishra and S. R. Nichols, *Curr. Opin. Struct. Biol.*, **2013**, 23, 734–739.

[466] A. Stank, D. B. Kokh, J. C. Fuller and R. C. Wade, *Acc. Chem. Res.*, **2016**, 49, 809–815.

[467] B. K. Ho, D. Perahia and A. M. Buckle, *Curr. Opin. Struct. Biol.*, **2012**, 22, 386–393.

[468] C. M. Goodman, S. Choi, S. Shandler and W. F. DeGrado, *Nat. Chem. Biol.*, **2007**, 3, 252.

[469] G. Guichard and I. Huc, *Chem. Commun.*, **2011**, 47, 5933–5941.

[470] S. H. Joo, *Biomol. Ther. (Seoul).*, **2012**, 20, 19.