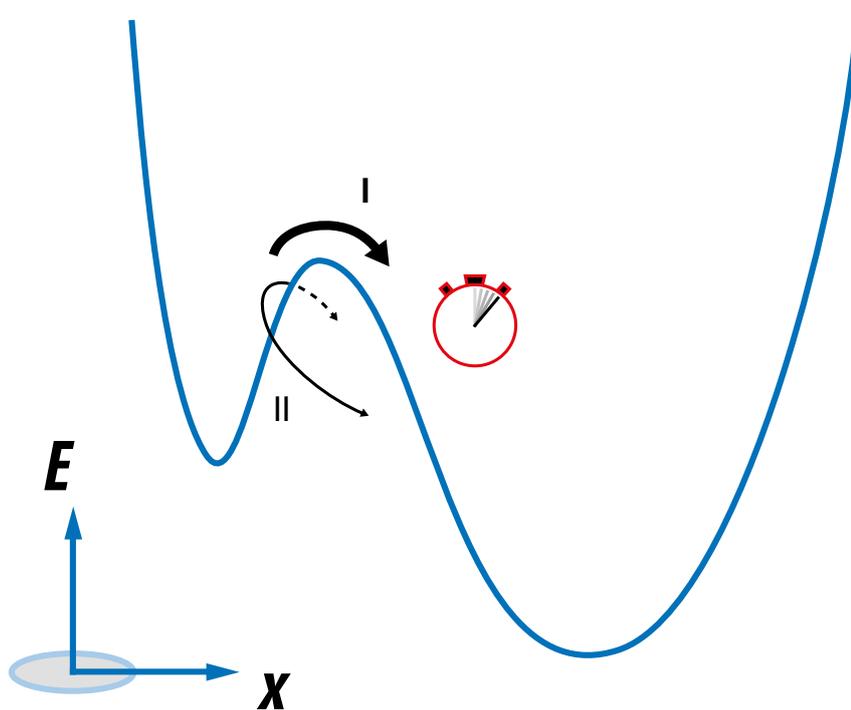


ALEXANDER MEHLICH

TRANSITION PATHS OF PROTEIN-FOLDING  
PROBED WITH OPTICAL TWEEZERS

Establishing transition path analysis techniques in single-molecule  
force spectroscopy based on simulation and experiments performed  
on natural and artificial proteins

DISSERTATION AT THE PHYSICS DEPARTMENT E22  
TECHNISCHE UNIVERSITÄT MÜNCHEN





TECHNISCHE UNIVERSITÄT MÜNCHEN  
LEHRSTUHL FÜR BIOPHYSIK E22

## TRANSITION PATHS OF PROTEIN-FOLDING PROBED WITH OPTICAL TWEEZERS

Establishing transition path analysis techniques in single-molecule force spectroscopy  
based on simulation and experiments performed on natural and artificial proteins

ALEXANDER MEHLICH

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung  
des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Martin Zacharias

Prüfer der Dissertation: 1. Prof. Dr. Matthias Rief  
2. Prof. Dr. Hendrik Dietz

Die Dissertation wurde am 18.10.2017 bei der Technischen Universität München eingereicht und durch die  
Fakultät für Physik am 05.12.2017 angenommen.

Alexander Mehlich: *Transition paths of protein-folding probed with optical tweezers*, Establishing transition path analysis techniques in single-molecule force spectroscopy based on simulation and experiments performed on natural and artificial proteins, © October 2017

All that is gold does not glitter,  
not all those who wander are lost;  
the old that is strong does not wither,  
deep roots are not reached by the frost.

— J.R.R. Tolkien

Exponential dependencies are nature's ace in the hole  
to make miracles become reality.

— A. M.



## ABSTRACT

---

Proteins are essential for life. To be able to perform their tasks, proteins need to fold into their functional form. From a physicist's perspective, nature solves this 'folding problem' by providing a multidimensional energy landscape which efficiently guides a loose peptide towards a distinct three-dimensional structure which is solely predefined by its unique amino acid sequence. A powerful method to directly study the folding mechanics of proteins is single-molecule force spectroscopy, which is used in this thesis. With a variety of sophisticated analysis tools it is possible to derive transition state positions and barrier heights up to entire one-dimensional projections of the folding energy landscapes of proteins from single-molecule trajectories. Recent technological advances to improve temporal and spatial resolution have opened doors towards directly accessing protein folding transition paths. The establishment of appropriate transition path analysis techniques and their correct interpretation was one of the main objectives of this work.

To this end, transition path analysis techniques were introduced and thoroughly tested using two-dimensional Brownian dynamics simulations. Using this, a very high inter-dependence between the friction involved in protein diffusion and the friction inherent to the detection system was revealed. Owing to this inter-dependence, previously described methods which aim to localise transition states from transition path ensembles or reconstruct barrier heights from committers typically fail to directly provide correct results. Conversely, the strong friction-dependence of detected transition path ensembles have enabled the introduction of a new, merely diffusion-based method of internal protein friction determination. This new method was successfully applied to experimental data of the three-helix bundle protein R15 and revealed the lowest internal friction directly reported from optical tweezers experiments.

The second main objective of this work was the comparison between the folding mechanics of natural and artificial proteins. One artificial protein, a labile re-designed version of the Ferredoxin-like fold, was found to be a relatively simple two-state folder. By contrast, another *de novo* designed protein, the Rossmann fold, was revealed to have an extremely rough energy landscape as opposed to the naturally occurring 'ideal' two-state folder R15. This high energy landscape roughness of the Rossmann fold was characterized by misfolds, multi-pathway folding, a greater effective roughness of  $\epsilon_{\text{rms}} = 2.4 k_{\text{B}}T$  with respect to R15, a local roughness of up to  $10 k_{\text{B}}T$ , two orders of magnitude longer transition path time averages  $\langle \tau_{\text{TP}} \rangle$  and, hence, slower folding. Various mutants showed the enforced key lock mechanism of the C-terminal  $\alpha$ -helix and its potential bending around position P86 to be the main cause of the folding problems of the Rossmann fold.

The results presented in this work pave the way to a more detailed understanding of protein folding mechanics from a transition path perspective.

## ZUSAMMENFASSUNG

---

Proteine sind lebenswichtig. Um ihre biologischen Aufgaben erfüllen zu können, müssen Proteine ihre funktionsfähige Form durch die sogenannte Faltung annehmen. Aus physikalischer Sicht löst die Natur dieses 'Faltungsproblem' mit Hilfe einer mehrdimensionalen Energielandschaft, in der ein loses Peptid in seine dreidimensionale Struktur überführt wird, die ausschließlich durch seine einzigartige Aminosäuresequenz vorbestimmt ist. Eine sehr mächtige Methode zur direkten Untersuchung der Faltungsmechanik von Proteinen ist die Einzelmolekül-Kraftspektroskopie, die auch in dieser Arbeit zum Einsatz kommt. Mit einer Vielzahl anspruchsvoller Methoden zur Auswertung von Einzelmolekül-Messungen wurde es möglich, die Position und Energiebarrierenhöhe von Übergangszuständen bis hin zu vollständigen Projektionen eindimensionaler Energielandschaften der Proteinfaltung zu bestimmen. Die jüngsten technologischen Fortschritte in Bezug auf die zeitliche und räumliche Auflösung haben nun auch die Türen zur direkten Detektion von Übergangspfaden der Proteinfaltung geöffnet. Die Einführung passender Analysemethoden von Übergangspfaden und deren korrekte Interpretation stellen wesentliche Ziele dieser Arbeit dar.

Um dies zu erreichen, wurden Analysemethoden von Übergangspfaden eingeführt und im Rahmen von zweidimensionalen Brownschen Bewegungssimulationen umfassend getestet. Auf diese Art wurde eine starke Abhängigkeit zwischen der Reibung der eigentlichen Proteinfaltung und der Reibung im Messsystem festgestellt. Aufgrund dieser Abhängigkeit scheitern Methoden, die auf der Grundlage von Ensembles aus Übergangspfaden die Position von Übergangszuständen feststellen oder die Höhe von Energiebarrieren rekonstruieren sollen, oft daran, auf direkte Weise korrekte Ergebnisse zu liefern. Im Gegensatz dazu konnte die starke Reibungsabhängigkeit der Ensembles aus Übergangspfaden dazu genutzt werden, eine neue, ausschließlich auf Diffusion basierte Methode zur Bestimmung der inneren Proteinreibung zu entwickeln. Diese neue Methode wurde erfolgreich am dreifachen Helixbündel-Protein R15 angewandt und offenbarte dabei die bisher geringste Reibung, die direkt von Messungen mit optischen Fallen berichtet wurden.

Ein weiteres wichtiges Ziel dieser Arbeit ist es, die Faltungsmechanik von natürlichen und künstlichen Proteinen miteinander zu vergleichen. Ein künstliches Protein, welches ein explizit mechanisch labiles Design des Ferredoxin-Faltungsmotivs repräsentiert, wies eine 'einfache' Faltung auf. Ein weiteres *de novo* Protein, das sogenannte Rossmann-Faltungsmotiv, wies hingegen eine extrem raue Energielandschaft auf, die im völligen Gegensatz zu der 'idealen' Zweizustandsfaltung des natürlichen Proteins R15 steht. Zusammengefasst zeichnet sich die starke Rauheit des Rossmann-Proteins durch eindeutige Hinweise auf Fehlfaltungen, mehrere Faltungspfade, eine gegenüber R15 um  $\epsilon_{\text{rms}} = 2.4 k_{\text{B}}T$  erhöhte effektive Reibung, eine lokale Rauheit von bis zu  $10 k_{\text{B}}T$  und um zwei Größenordnungen längere mittlere Übergangspfadzeiten  $\langle \tau_{\text{TP}} \rangle$

und damit eine entsprechend langsamere Faltung, aus. Verschiedene Mutanten weisen darauf hin, dass der verstärkte Verriegelungsmechanismus der C-terminalen  $\alpha$ -Helix und deren Verkrümmung im Bereich der Prolinposition P86 für die Probleme bei der Faltung des Rossmann Proteins verantwortlich sein könnten.

Die Ergebnisse dieser Arbeit ebnen den Weg in Richtung eines besseren Verständnisses der Proteinfaltungsmechanik sowohl im Allgemeinen als auch im Besonderen aus der Sicht von Übergangspfaden.



## PUBLICATIONS

---

While the focus of this thesis lies on unpublished scientific results, a few of the concepts and figures which emerged along with this work can be found in the following publications:

- RINGER P, WEIßL A, COST A.-L., FREIKAMP A, SABASS B, MEHLICH A, TRAMIER M, RIEF M, GRASHOFF C (2017) Multiplexing molecular tension sensors reveals piconewton force gradient across talin-1. *Nature Methods*
- FREIKAMP A, MEHLICH A, KLINGNER C, GRASHOFF C (2016) Investigating piconewton forces in cells by FRET-based molecular force microscopy. *Journal of Structural Biology*
- AUSTEN K<sup>1</sup>, RINGER P<sup>1</sup>, MEHLICH A<sup>1</sup>, CHORSTK-GRASHOFF A, KLUGER C, KLINGNER C, SABASS B, ZENT R, RIEF M, GRASHOFF C (2015) Extra-cellular rigidity sensing by talin isoform-specific mechanical linkages. *Nature Cell Biology*
- MEHLICH A<sup>1</sup>, AUSTEN K<sup>1</sup>, RINGER P<sup>1</sup>, RIEF M, GRASHOFF C (2015) Evaluation of molecular tension sensors using single-molecule force spectroscopy and live cell FRET imaging. *Nature Protocol Exchange*<sup>2</sup>
- FANG J, MEHLICH A, KOGA N, HUANG J, KOGA R, GAO X, HU C, JIN C, RIEF M, KAST J, BAKER D, LI H (2013) Forced protein unfolding leads to highly elastic and tough protein hydrogels. *Nature Communications*
- VON HANSEN Y, MEHLICH A, PELZ B, RIEF M, NETZ R (2012) Auto- and cross-power spectral analysis of dual trap optical tweezer experiments using Bayesian inference. *Review of Scientific Instruments*

---

<sup>1</sup> These authors contributed equally.

<sup>2</sup> Non-peer reviewed.



# CONTENTS

---

<b>i</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>1</b>	<b>INTRODUCTION AND MOTIVATION</b>	<b>3</b>
1.1	Solved, or not solved? - the protein-folding problem	3
1.2	The energy landscape perspective	5
1.2.1	Transition paths - distilled essence of reactions	6
1.3	From protein prediction towards protein design	8
1.3.1	Two designed ideal protein structures - Ferredoxin-like fold and Rossmann fold	9
1.4	Thesis outline	10
<b>ii</b>	<b>METHODS, THEORIES, AND DATA ANALYSIS</b>	<b>13</b>
<b>2</b>	<b>EXPERIMENTAL SETUP AND IMPLEMENTATION</b>	<b>15</b>
2.1	How to apply force to a single protein molecule	15
2.1.1	Optical tweezers setup	16
2.1.2	Experimental geometry: the dumbbell assay	17
2.2	Building protein-DNA conjugates	19
2.2.1	Protein-oligonucleotide construct purification	20
2.2.2	Maleimide reactivity test	21
2.2.3	Bioorthogonal chemical coupling alternatives	24
2.3	Experimental modes of force probing	25
2.4	From transitions towards transition paths	28
<b>3</b>	<b>THEORY AND ANALYSIS OF SINGLE MOLECULE EXPERIMENTS</b>	<b>31</b>
3.1	Polymer mechanics	31
3.1.1	Contour length transformation	32
3.2	Free energy calculation	33
3.2.1	Force-dependent probabilities	35
3.3	Fitting equilibrium unfolding/folding transitions	35
3.4	Extracting force-dependent transition rates	36
3.4.1	Rates from force distribution histograms	37
3.4.2	Constructed equilibrium: Oberbarnscheidt's method	37
3.4.3	Rates from lifetime distributions: Hidden Markov model analysis of equilibrium trajectories	38
3.4.4	Pushing limits: autocorrelation analysis	40
3.5	Kinetic models	40
3.5.1	Blind spot at zero force: Arrhenius equation	41
3.5.2	The power of force: Zhurkov-Bell model	41
3.5.3	Force-induced barrier: Schlierf-Berkemeier model	42
3.5.4	Getting in shape: using Kramers theory	44
3.6	Force distribution histograms	46
3.6.1	Elastic compliance: loading rate determination	47
3.7	Deconvolution of probability distributions	49
3.7.1	Setting up the point spread function	50

3.8	Transition Path Analysis	51
3.8.1	Path ensembles and commitment probabilities	52
3.8.2	Committer-based barrier shape reconstruction	54
3.8.3	Bayesian path statistics localise transition states	55
3.8.4	Transition path times and kinetic rates	57
3.8.5	Extracting transition paths	59
iii	<b>SIMULATION RESULTS</b>	63
4	<b>LANGEVIN DYNAMICS SIMULATIONS</b>	65
4.1	Simulation principle	65
4.2	Studies on one-dimensional rough energy landscapes	68
4.2.1	Predicting transition path times and kinetic rates	68
4.2.2	Diffusion in a rough energy landscape	71
4.2.3	Energy landscape reconstruction from transition path times reflecting local roughness	74
4.2.4	A fingerprint for roughness: predominant transition state switching induced by force	79
4.3	Energetically coupled two-dimensional diffusion	81
4.3.1	Deconvolution as a link between dimensions	81
4.3.2	The impact of anisotropic friction on rates, transition path times, and transition path ensembles	84
4.3.3	Transition averages at experimental resolution: seeking traces of protein folding	92
4.3.4	Transition path analysis in force spectroscopy	100
4.4	Summary and Outlook	109
iv	<b>EXPERIMENTAL RESULTS</b>	113
5	<b>FERREDOXIN-LIKE FOLD: MECHANICS OF A TWO-STATE FOLDER</b>	115
5.1	A re-designed ideal protein for tissue engineering	115
5.2	From simple towards complex two-state mechanics	117
5.2.1	Slowing down: approaching equilibrium	117
5.2.2	Varying distance: shifting equilibrium populations	119
5.2.3	How well do different analysis methods agree?	121
5.2.4	Deviations from simple two-state mechanics at low forces	123
5.3	Discussion	125
5.3.1	Force-induced transition state movement: confirming Hammond behaviour by deconvolution	125
5.3.2	Case study FLred: a comparison between the predictive capabilities of different kinetic models	128
5.3.3	Qualitative energy landscape reconstruction	131
5.3.4	Mechanical parameters of designed and natural proteins: revealing a key role of the transition state?	135
5.4	Summary and Outlook	138
5.4.1	Successful protein design verification	138
5.4.2	Application as an <i>in vivo</i> force sensor	139
5.4.3	Force spectroscopy - a test bench for protein design, folding concepts, and models	141

6	ROSSMANN FOLD: AN ARTIFICIAL PROTEIN IN TROUBLE	143
6.1	Design meets evolution: Rossmann vs. R15	143
6.2	A multidimensional and rough energy landscape – a two-state perspective	144
6.2.1	Multiple fingerprints and multi-modal unfolding	144
6.2.2	Split and kinked rate plots	149
6.2.3	Transition averages and prolonged transition path times	152
6.2.4	Deconvolution and commitment probabilities	156
6.2.5	Transition path ensembles and probabilities	161
6.3	Mutants	165
6.3.1	P86A and P86S: proline-free mutants reveal that proline serves as a stabilizing design element	165
6.3.2	S49Cc: directed unfolding of the C-terminal half of ROSS uncovers an energetic imbalance	169
6.3.3	S49Cn and S85Cn: tracing another intermediate	177
6.4	Extended discussion	181
6.4.1	Effective roughness and friction under force	181
6.4.2	Transition-path-assisted protein friction determination	184
6.4.3	Localization and structural interpretation of high-energy intermediates of the Rossmann fold	188
6.4.4	Transition-path-assisted energy landscape reconstruction	192
6.5	Summary and Outlook	195
6.5.1	Folding reliability: evolution outperforms design by means of reduced complexity	196
6.5.2	Transition paths - high potential with imposed limits	202
V	APPENDIX	209
A	MATERIALS AND METHODS	211
A.1	Protein sequences	211
A.1.1	Ferredoxin-like fold	211
A.1.2	Rossmann fold	212
A.1.3	R15	214
A.1.4	GB1	215
A.1.5	Tension sensors: Ypet-HP35(st)-mCherry	215
A.1.6	Anti-parallel heterodimeric coiled coil: U2LZ	217
A.1.7	Homodimeric coiled coils from kinesin-1	217
A.2	Molecular cloning	217
A.2.1	Functional regions and domains of kinesin-1	217
A.2.2	Design and cloning of coiled coil unzipping constructs	219
A.3	Protocols	224
A.3.1	Protein expression and purification	224
A.3.2	Oligonucleotide attachment details	227
A.3.3	SEC in series with Ni-NTA	234
A.3.4	DNA-handle PCR	238
A.3.5	Sample preparation and measurement at the trap	240
	BIBLIOGRAPHY	247

## LIST OF FIGURES

---

Figure 1	Illustration of a protein folding energy landscape.	6
Figure 2	General transition path definition in a simple double-well potential.	7
Figure 3	From protein prediction towards protein design.	8
Figure 4	NMR structures of the Ferredoxin-like fold and the Rossmann fold.	9
Figure 5	Schematic of the optical tweezers setup.	16
Figure 6	Schematic of the dumbbell configuration.	18
Figure 7	SEC purification of an oligonucleotide attachment.	21
Figure 8	Maleimide oligonucleotide reactivity test with DTT.	22
Figure 9	Screening maleimide-thiol reaction parameters.	23
Figure 10	Force-extension trace measured in constant velocity mode.	25
Figure 11	Time traces measured in constant distance mode.	27
Figure 12	Transition paths from highly resolved transitions.	29
Figure 13	Key parameters characterizing folding energy landscapes.	43
Figure 14	Three kinetic models for fitting force-dependent rates.	44
Figure 15	Force distribution histograms and elastic compliance contributions.	48
Figure 16	Illustration of the transition path ensemble and the commitment probability.	53
Figure 17	Commitment probability calculation and committor-based barrier shape reconstruction.	55
Figure 18	Transition path probability calculation and transition state localization.	56
Figure 19	Transition path time averages and distributions.	58
Figure 20	Transition path extraction from simulated and experimental trajectories.	60
Figure 21	2D Langevin dynamics simulation principle.	67
Figure 22	Analysis of transition path times and rates from a 1D diffusion simulation.	69
Figure 23	A thousandfold increase in transition path times induced by energy landscape roughness.	72
Figure 24	Barrier height dependence of dwell times for a given transition state position.	74
Figure 25	Energy landscape reconstruction from transition path times reflecting local roughness.	76
Figure 26	Force induced transition state switching as a fingerprint for roughness.	80
Figure 27	Boltzmann inversion and deconvolution confirm an energetically consistent 2D diffusion simulation.	83
Figure 28	Speed limit protein folding under tension.	85

Figure 29	Friction-dependence of transition path times, rates, and path ensembles of protein folding probed by force. 87
Figure 30	Friction acts as a switch between preferential transition paths. 91
Figure 31	Averaged transitions of the deflection signal with and without inherent protein folding at the 'speed limit'. 94
Figure 32	Averaged transitions of recorded bead dynamics reveal inherent protein folding time scales. 95
Figure 33	Averaged transitions reveal basic features of protein folding energy landscapes. 98
Figure 34	Committer-based barrier reconstruction I: case of a symmetric protein folding energy landscape. 101
Figure 35	Transition path probability calculation I: case of a symmetric protein folding energy landscape. 103
Figure 36	Deconvolution of an asymmetric protein folding energy landscape. 106
Figure 37	Committer-based barrier reconstruction II: case of an asymmetric protein folding energy landscape. 107
Figure 38	Transition path probability calculation II: case of an asymmetric protein folding energy landscape. 108
Figure 39	Positions of the 16 point mutations of FLred. 116
Figure 40	Force-extension traces of FLred at different pulling speeds. 117
Figure 41	Energy landscape parameter reconstruction from data obtained in constant velocity mode. 119
Figure 42	Energy landscape parameter reconstruction from data obtained in constant distance mode. 120
Figure 43	Deviations from simple two-state mechanics at low forces. 124
Figure 44	Transition state movement revealed by deconvolution and energy landscape transformation. 126
Figure 45	The two potential folding units of FLred. 132
Figure 46	Energy landscape reconstruction scenario for FLred. 133
Figure 47	Force sensor performance prediction for FLred. 140
Figure 48	Affecting rates by shifting the transition state position. 141
Figure 49	Crystal and solution NMR structures of the spectrin domain R15 and the 2x2 Rossmann fold. 143
Figure 50	Force spectroscopic fingerprints of ROSS and R15. 145
Figure 51	ROSS has a multi-modal unfolding force distribution. 146
Figure 52	ROSS has split and kinked rate plots. 150
Figure 53	Transition averages of R15 and ROSS. 153
Figure 54	Individual equilibrium transitions and transition path time distributions of R15 and ROSS. 154
Figure 55	Committer-based barrier reconstruction for R15. 157
Figure 56	Committer-based barrier reconstruction for ROSS. 158
Figure 57	R15 and ROSS point towards two fundamental weaknesses of deconvolution. 160
Figure 58	Transition path probability calculation for R15 and ROSS. 162

Figure 59	Proline-free mutants of ROSS by substitution: P86A and P86S. <a href="#">165</a>
Figure 60	Directed unfolding of the C-terminal half of ROSS: S49Cc. <a href="#">170</a>
Figure 61	Detailed characterization of the on-pathway intermediate of S49Cc by constant distance measurements. <a href="#">172</a>
Figure 62	Why $\beta$ -sheet swapping might cause 'native-like' fingerprints. <a href="#">175</a>
Figure 63	Directed N-terminal unfolding of ROSS: S49Cn and S85Cn. <a href="#">178</a>
Figure 64	Effective energy landscape roughness derived from transition path times measured under force. <a href="#">182</a>
Figure 65	Transition-path-assisted friction coefficient determination of R15. <a href="#">185</a>
Figure 66	Localization and structural interpretation of high-energy intermediates of ROSS. <a href="#">189</a>
Figure 67	Transition-path-assisted energy landscape reconstruction for ROSS at 9.5 pN. <a href="#">193</a>
Figure 68	Reaction schemes for R15 and ROSS. <a href="#">197</a>
Figure 69	A 'speed limit' reference and candidates for sequence-resolved friction maps. <a href="#">205</a>
Figure 70	Force-dependent transition path times, histograms, and ensembles. <a href="#">207</a>
Figure 71	Testing an engineered cystein crosslink on an SDS-gel. <a href="#">228</a>
Figure 72	SEC purification of a thiol oligo attachment. <a href="#">230</a>
Figure 73	Thiol oligo reactivity test with BM(PEG) <sub>2</sub> . <a href="#">231</a>
Figure 74	Protein-oligo construct test on an agarose gel. <a href="#">234</a>
Figure 75	SEC column characterization: resolution and time shifts. <a href="#">235</a>
Figure 76	Contamination of maleimide oligos revealed by SEC. <a href="#">236</a>
Figure 77	Improved purification of an oligo attachment reaction by an additional Ni-NTA column. <a href="#">237</a>

## LIST OF TABLES

---

Table 1	Rate constant calculations: theory versus simulation. <a href="#">70</a>
Table 2	Detection limit for protein folding transitions. <a href="#">96</a>
Table 3	Contour gains and folding free energies of FLred from different analysis methods and modes. <a href="#">121</a>
Table 4	Kinetic parameters of FLred from measurements in constant velocity and constant distance mode. <a href="#">122</a>
Table 5	Force-induced transition state movement of FLred. <a href="#">127</a>
Table 6	Energy landscape parameter prediction capabilities of different kinetic models when applied to FLred data. <a href="#">129</a>
Table 7	Mechanical features of designed and natural proteins. <a href="#">136</a>

Table 8	Kinetic ZB-model fit parameters of R15 and ROSS based on force distribution histograms. <a href="#">148</a>	
Table 9	Kinetic SB-model fit parameters of R15 and ROSS based on OBS and HMM rates. <a href="#">151</a>	
Table 10	Contour length gains and folding free energies of R15 and ROSS. <a href="#">152</a>	
Table 11	Transition path time averages and bead relaxation times of R15 and ROSS. <a href="#">155</a>	
Table 12	Contour length gains, relative fingerprint occurrences, and folding free energies of ROSS, P86A, and P86S. <a href="#">166</a>	
Table 13	Comparing the kinetics of ROSS and the proline-free mutants P86A and P86S. <a href="#">167</a>	
Table 14	Contour length gains and relative fingerprint occurrences of S49Cc compared to ROSS. <a href="#">171</a>	
Table 15	Comparing the kinetics of S49Cc and ROSS. <a href="#">171</a>	
Table 16	SB-model fit parameters and folding free energies of S49Cc and its obligatory on-pathway intermediate. <a href="#">173</a>	
Table 17	Contour length gains and unfolding characteristics of S49Cn and S85Cn. <a href="#">179</a>	
Table 18	Theoretical speed limit versus experiment: protein friction coefficients implied by transition path times. <a href="#">183</a>	
Table 19	Predominant on-pathway intermediate positions of ROSS. <a href="#">190</a>	
Table 20	Reaction premix for DNA-handle PCR. <a href="#">239</a>	
Table 21	Temperature step protocol for DNA-handle PCR. <a href="#">239</a>	
Table 22	Final measurement mix example. <a href="#">242</a>	

## ACRONYMS

---

<b>AFM</b>	Atomic Force Microscope
<b>BM(PEG)<sub>2</sub></b>	1,8-Bismaleimido-diethyleneglycol
<b>BSA</b>	Bovine Serum Albumin
<b>DTDP</b>	2,2'-Dithiodipyridine
<b>DTT</b>	Dithiothreitol
<b>DNA</b>	Deoxyribonucleic Acid
<b>FRET</b>	Förster Resonance Energy Transfer
<b>FWHM</b>	Full Width at Half Maximum
<b>GFP</b>	Green Fluorescent Protein

**GODCAT** Glucose/Glucose-Oxidase/Catalase oxygen scavenging system  
**HMM** Hidden Markov model  
**MD** Molecular dynamics  
**NMR** Nuclear Magnetic Resonance  
**PBS** Phosphate Buffered Saline  
**PCR** Polymerase Chain Reaction  
**PSF** Point Spread Function  
**RT** Room Temperature  
**SEC** Size Exclusion Chromatography  
**TCEP** Tris(2-carboxyethyl)phosphine  
**TPT** Transition Path Theory  
**TST** Transition State Theory  
**WHAM** Weighted Histogram Analysis Method  
**WLC** Worm-like Chain  
**eWLC** extensible Worm-like Chain  
**YFP** Yellow Fluorescent Protein

## Part I

### INTRODUCTION

Picking up the golden thread. A problem, a concept, and design.  
And last but not least: a plan of action.



## INTRODUCTION AND MOTIVATION

---

Proteins are among the smallest machines on Earth. Only a few nanometres in size, proteins make and keep us alive. To perform a multitude of tasks like, e. g., digesting food, transporting oxygen in our blood stream, or contracting muscles for the movement of our body, proteins need to fold into their specific shape. Over millions of years, this folding process has been optimized by evolution to guarantee proper function. However, misfolding and, hence, a loss of function can still occur which, in the worst case, may lead to a number of neurodegenerative and other diseases [174]. With the protein folding process being at the core of life, its fundamental understanding is the main motivation for the protein mechanical studies in this thesis.

How is it experimentally possible to access protein folding? Among the vast number of different experimental approaches which exist to investigate the properties of proteins, one can discern single-molecule techniques from ensemble based methods. With respect to ensemble measurements, experiments on a single molecule level, while often times being more elaborate, typically have the advantage of providing much more detailed information. A very prominent single-molecule technique are Förster Resonance Energy Transfer (FRET) experiments which allow distances within protein structures to be resolved as well as detecting conformational changes under zero-force conditions [44, 96, 198, 200]. Another very powerful method which directly probes the folding mechanics of individual molecules is single-molecule force spectroscopy. While this started out as an Atomic Force Microscope (AFM) based technique, optical tweezers are now also well-established as a particularly useful tool to study protein folding in the lower piconewton (pN) regime [83, 181, 213, 242, 250, 260]. All experimental results presented in this work have been performed on a custom-built high-resolution dual-beam optical tweezers setup.

### 1.1 SOLVED, OR NOT SOLVED? - THE PROTEIN-FOLDING PROBLEM

In 1961, Anfinsen and co-workers found that proteins can reversibly fold in solution without need for an extra cellular machinery [2]. Back then, the first published globular protein structure, myoglobin, had a surprisingly complex shape which was lacking any of the expected symmetry and regularity [114]. Together, these two findings launched the protein-folding problem which can be roughly encircled by the following three questions [62, 205]: First, what general physical rules govern the folding of a disordered one-dimensional amino acid chain into a functional three-dimensional structure? Second, by which folding mechanism can proteins fold so fast into their distinct shape despite an overwhelmingly large number of possible conformations? And

third, is it possible to predict the native structure that an unknown protein will adopt merely by knowing its amino acid sequence?

*From Levinthal to Shakespeare*

An illustration which addresses the second question from above and which is often used to describe the astonishing complexity of protein folding is 'Levinthal's paradox'. The paradox states that it would take an enormously long time for proteins to fold if they had to find their native configuration by a random search among all possible configurations. Yet it is known that proteins can fold within time scales reaching down to a few microseconds [45, 130]. In a typical example of Levinthal's paradox, each bond connecting two amino acids can have, e. g., three degrees of freedom, so that a protein comprising 101 amino acids could exist in  $3^{100}$  possible configurations. Even if the protein could sample through all configurations by testing one configuration every attosecond, one quintillionth of a second, this process would last for about  $10^{22}$  years – more than ten orders of magnitude longer than the universe has existed for [227].

A way to resolve the paradox is pointed out by Dawkins, an evolutionary biologist, in the context of a slightly different question [53]: How long would it take a randomly typing monkey to produce Hamlet's remark 'Methinks it is like a weasel'? Including spaces, this statement is 28 characters long with each character having 27 possible 'configurations', 26 letters and a space. In total, this yields up to  $28^{27}$  necessary key strokes to 'produce' the original Hamlet. Naturally, a monkey will not be able to 'try' out one configuration every attosecond. Nevertheless, Dawkins observed that if the monkey cannot change those letters that are already correctly in place, Hamlet's phrase can be realized by a random search in only a few thousand key strokes [263]. Whether it is writing Hamlet or folding proteins, the key to speeding things up is a biased search as opposed to a completely random walk through configurational phase space.

*Solved parts of a growing puzzle*

Which one of the three questions encircling the protein-folding problem have already been answered? The first question asks for general folding rules which corresponds to a search for forces that drive a protein into its three-dimensional structure. It is known that among the contributing factors are: hydrogen bonds which lead to the formation of  $\alpha$ -helices and  $\beta$ -sheets, close-ranged van der Waals interactions which govern tight protein packing, backbone angle preferences similar to the three degrees of freedom of the connecting bonds between amino acids from above, longer-range electrostatic interactions which cause amino acids to attract or repel each other because of different net charges, hydrophobic interactions which are also referred to as the 'hydrophobic effect' where hydrophobic amino acids are buried in a protein's core and 'shielded' from solution by polar amino acids on a protein's

surface, and, finally, chain entropy which acts against folding by entropically favouring any random coil configuration over a distinctly folded structure.

Levinthal's paradox illustrates the extent of the second question, which addresses the speed of protein folding. A solution to the paradox is given by the commonly accepted picture of funnel-shaped energy landscapes which provide the necessary bias as they guide unfolded proteins energetically downhill towards their native structure. Since detailed studies of energy landscapes which characterize the rate mechanism of protein folding mark an essential part of this thesis, a few more thoughts on their concept and implications are presented in [Section 1.2](#).

Concerning sequence-based protein structure prediction, which is referred to by the third and last question, scientists have improved the accuracy of their prediction algorithms over the past decades. Since 1994, advances in prediction are directly mirrored in the results of the biannual competition known as the Critical Assessment of protein Structure Prediction (CASP) [151]. In 2010, the best predictions for about 10% of all CASP target proteins reached accuracies which are high enough to initiate drug discovery programs that need structural errors to be less than 2 to 3 Å [62]. Now, the remaining challenge is to achieve excellent predictions for the other 90%.

## 1.2 THE ENERGY LANDSCAPE PERSPECTIVE

Inspired by statistical mechanics, energy landscapes provide a conceptual framework for describing protein folding mechanisms as a thermodynamically driven process of populating a diversity of possible states with a certain probability [162, 240]. While protein folding was first believed to be limited to very specific pathways, by now a more general picture of a funnel-shaped energy landscape has evolved, see [Figure 1](#). This generalized view expands the pathway concept of sequential events towards the funnel concept of parallel events which ultimately lead to the same native structure. Common features among a multitude of possible pathways can accumulate into ensembles of configurations which are accessed by experimentalists as transition states, intermediates, and on- or off-pathway events [61].

Which general folding mechanisms are known? Proteins seem to fold in units of secondary structure. Growth into more global structures appears to be preceded by local structure formation. Along with structural growth, proteins gain stability. This divide-and-conquer, local-to-global process is one of the factors that enables proteins to fold so fast [62, 241].

As suggested by the funnel-shaped energy landscape in [Figure 1](#), protein-folding is a highly multidimensional problem. However, in force spectroscopic experiments, we are limited to measuring only one reaction coordinate in space. This coordinate is the extension of unfolded polypeptide along the direction of the acting force. The second important dimension in our experiments is time. From measuring force-extension relations over time, it is possible to recognize and characterize complex protein folding networks involving multiple pathways and intermediates [189, 197, 216].

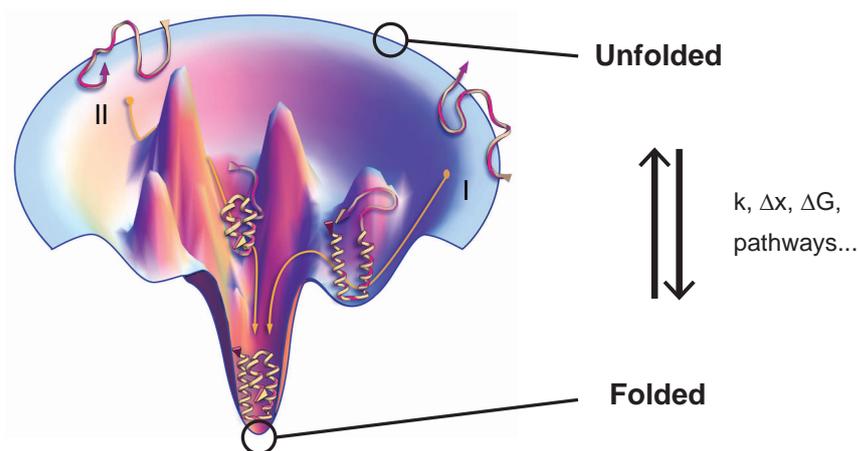


Figure 1: Illustration of a funnel-shaped protein folding energy landscape adapted from [62]. Conformational entropy is indicated radially, energy and similarity to the native structure by depth.

Proteins find their native configuration by following the principle of minimal frustration which may be compared to avoiding high energy barriers and running preferentially downhill towards the bottom of a funnelled landscape [32]. In this context, slowed folding kinetics are attributed to frustrated proteins. A way to describe frustration is by assuming higher internal protein friction which corresponds to a smaller diffusion coefficient and, hence, slower diffusion in an energy landscape [238]. Alternatively, frustration is associated with rough energy landscapes where proteins need to pass through many ‘bumps’ before reaching their native state [262]. To resolve how these two different perspectives relate to each other, this work addresses the following general question:

- **What does energy landscape roughness mean?**

A general strength of the energy landscape perspective is its ability to provide a condensed and demonstrative overall picture. The experimental determination of essential energy landscape parameters such as transition state positions  $\Delta x$  along the reaction coordinate  $x$  and their corresponding energy barrier heights  $\Delta G$  further allows the modelling of specific protein folding kinetics and provides deeper insights into general aspects of folding.

### 1.2.1 *Transition paths - distilled essence of reactions*

As illustrated in Figure 2, transition paths are defined as those pieces of trajectories by which the actual event of barrier crossing occurs. In their role as reactive trajectories, transition paths are immediate witnesses of a reaction and bear essential information on the underlying reaction mechanism.

The ‘speed limit’ for protein folding was both experimentally and theoretically estimated to be  $\sim 1 \mu\text{s}$  [92, 119]. Hence, the ultimate prerequisite for a

direct observation of transition paths in single-molecule experiments is a high enough temporal resolution. By now, state-of-the-art optical tweezers setups have reached a temporal resolution of the order of  $10\ \mu\text{s}$  which puts the investigation of protein folding transition paths within reach [261]. And indeed, right after protein transition path times were reported from single-molecule FRET experiments [45], results from single-molecule force spectroscopy followed, at first only for Deoxyribonucleic Acid (DNA) [154], and very recently also for proteins [156, 157].

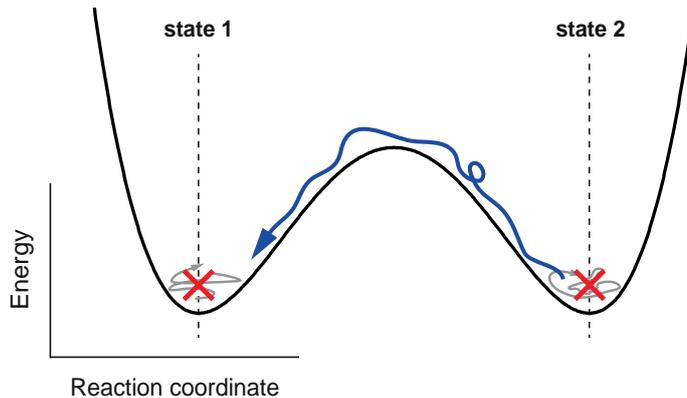


Figure 2: General transition path definition in a simple double-well potential.

In this work, transition paths are extracted from equilibrium and non-equilibrium protein folding trajectories. In addition to their duration, transition paths also provide spatial information. Ideas for how this information can be used to learn something about the underlying reaction mechanisms can be found in Transition Path Theory (TPT) [236, 237]. However, one also has to be aware of the fact that measured transition paths are altered by the measurement itself. With respect to the applied force spectroscopic approach used in this work, effects arising from the additional friction inherent to the detection system need to be addressed. The following key question expresses the need for applicable transition path analysis techniques and their comprehensive understanding when being applied to force spectroscopic data:

- **How and what can we learn from studying transition paths?**

To answer this question, this work complements the well-established toolbox of models based on Transition State Theory (TST), which dates back to Eyring and Polyani [73, 74], with analysis methods based on TPT by following thoughts presented in [101]. After testing the performance of a set of transition path analysis tools on simulated force spectroscopic experiments, these tools are applied to experimental data for the first time.

Furthermore, some of the findings in this thesis were independently confirmed by a series of recent publications which, however, do not provide the same level of insight as the results presented here [143, 155, 157].

## 1.3 FROM PROTEIN PREDICTION TOWARDS PROTEIN DESIGN

Protein design can be considered as the ‘inverted’ protein-folding problem [164]. Instead of trying to predict the structure for a given amino acid sequence, protein design looks for an amino acid sequence which folds into a given protein structure<sup>1</sup>, see Figure 3. This way, scientists already have created folds with topologies non-existent in nature, enhanced or inhibited enzyme activity, created new enzymes with new functions, and improved drug development [47, 100, 244].

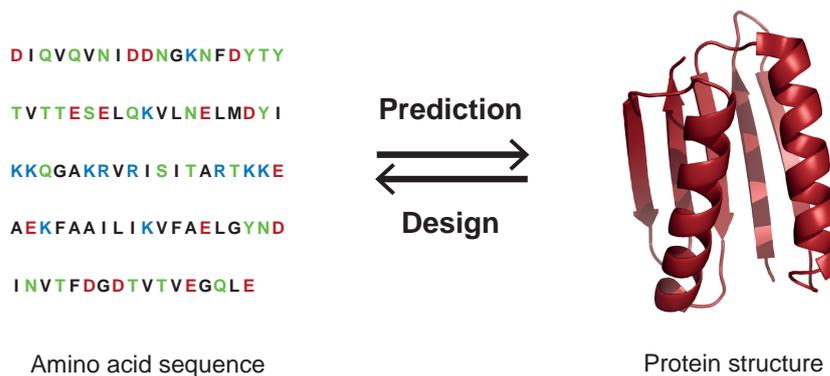


Figure 3: From protein prediction towards protein design. The designed 93 amino acid long sequence on the left correctly folds into the protein structure on the right which served as a template for design. This so-called Top7 protein (pdbID: 1QYS) has a topology not found in nature [121].

After choosing a topology, protein design turns into an optimization problem. The basis for the optimization is provided by energy functions which describe the forces that drive an amino acid sequence into a folded three-dimensional structure<sup>2</sup>. With the aim being to find the amino acid sequence that is lowest in energy out of all possible amino acid sequences, optimization is done by sampling alternative backbone and side-chain conformations. Here, the main challenge is the sampling of the backbone conformational space<sup>3</sup> which involves high numerical cost.

One of the most successful prediction and design algorithms is ROSETTA. The way how this algorithm works can be learned from a protein folding game called ‘Foldit’<sup>4</sup> [48]. A particular strength of the algorithm is a clever combination of available experimental data and heavy computing [126, 205].

Since both success and failure in protein design provide insight into the complex mechanisms governing protein folding, all necessary steps along the road towards a full structural and functional control of proteins are promising not only in view of the advances which were and will be brought to medical research by the protein design approach.

<sup>1</sup> Protein design inverts question number three which encircles the protein-folding problem.

<sup>2</sup> See question one which encircles the protein-folding problem and its answer in Section 1.1.

<sup>3</sup> Here, we are reminded of Levinthal’s paradox and question number two.

<sup>4</sup> <https://fold.it/portal/> - check it out! By solving puzzles for science, you can actively contribute to gaining more insight into the complex mechanisms which govern protein folding.

### 1.3.1 Two designed ideal protein structures - Ferredoxin-like fold and Rossmann fold

While Top7, shown in [Figure 3](#), is the first designed fold with a topology which does not exist in nature [\[121\]](#), the Ferredoxin-like fold (FL) and the 2x2 Rossmann fold (ROSS) shown in [Figure 4](#) are the first fully designed proteins which represent a topology that is abundantly found in nature [\[115\]](#). Being designed as ideal proteins using the ROSETTA algorithm, FL and ROSS are intended to serve as building blocks for much larger proteins. While ROSS can provide an ATP hydrolysis site and be a minimal core for designed enzymes, FL is a typical fold that can serve as a binding pocket for metals and be involved in electron transport [\[118, 178, 190\]](#).

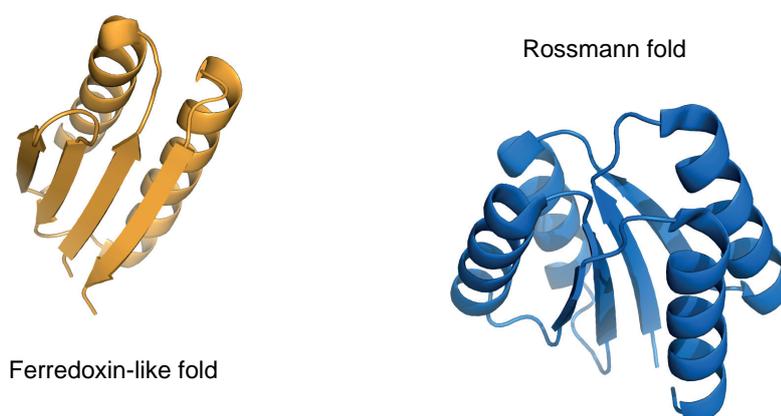


Figure 4: NMR structures of the Ferredoxin-like fold and the Rossmann fold (pdbID: 2KL8 and 2LV8). Both proteins are designed ideal structures having a topology also found in nature [\[115\]](#).

As mentioned above, protein design aims at finding optimal amino acid sequences which fold into a desired shape. A successful design is then judged based on the stability of the resulting protein as well as its similarity to the original template which is evaluated by comparison to crystal or Nuclear Magnetic Resonance (NMR) structures. However, in terms of actual folding pathways which can be probed by using single-molecule force spectroscopy, no comparison has ever been made between natural and designed, i. e., artificial proteins. A central question which will be addressed in this work by detailed mechanical studies of FL and ROSS therefore is:

- **Are natural and artificial proteins mechanically different?**

The answer to this question has important implications on the design of protein structures especially if pharmaceutical applications are intended. In such a case, fast folding and the avoidance of kinetic traps have to be of higher priority than mere overall stability.

#### 1.4 THESIS OUTLINE

This thesis is divided into five parts. In the first part, I propose and justify protein mechanical studies as a direct way to look for answers to the protein-folding problem. Within this question, three different aspects are of central interest: energy landscape roughness, the analysis of transition paths, and the comparability of natural and artificial proteins.

The second part introduces the applied experimental method of single-molecule force probing and its implementation, see [Chapter 2](#). Here, results from a newly developed reaction parameter screening for protein-DNA conjugation are presented in more detail in [Section 2.2](#). In [Chapter 3](#), all theoretical models and analytical methods necessary for data analysis are explained. After giving a broad overview of all applied and well-established analysis techniques, in [Section 3.8](#), this toolbox of techniques is expanded towards transition path analysis methods. Importantly, in [Section 3.8.4](#), the barrier-shape-dependence ( $\Delta G, \Delta x$ ) of transition path time averages  $\langle \tau_{TP} \rangle$  is elaborated and found to match experimental data.

The simulation results presented in the third part of this work establish a sound basis for the evaluation and interpretation of experimental data. In [Chapter 4](#), two main foci lie on studying effects arising from energy landscape roughness and implications of the difference in friction involved in actual protein folding and its detection. Results include the necessity of sequential transition path time calculations to better model asymmetric potentials; a numerical confirmation of Zwanzig's roughness concept; the development of a dwell-time-based reconstruction technique for rough energy landscapes; the finding that force-induced transition state switches have a similar yet stronger impact on transition rates than the Hammond effect; the influence of friction on measured kinetics; the detection limit for direct traces of protein folding transitions, and, finally, the potential and limits of transition path analysis in force spectroscopy.

In the fourth part, two chapters are dedicated to experimental results and their detailed discussion. [Chapter 5](#) can be considered as a showcase for almost all state-of-the-art analysis techniques in single-molecule force spectroscopy. Besides the verification of an intentional labile re-design of the Ferredoxin-like fold, in [Section 5.2.3](#), the robustness and interconvertibility of different analysis approaches and measurement techniques are also confirmed. Furthermore, force-dependent transition state movement reveals the principle need to account for Hammond behaviour, see [Section 5.3.1](#). Moreover, the role of the transition state position with respect to a proteins function is discussed in [Section 5.4.3](#). A general trend between transition state positions and combined rates is described for a whole set of proteins in [Section 5.3.4](#). [Chapter 6](#) is a mechanical study which compares the artificial 2x2 Rossmann fold (ROSS) to a natural protein, namely the spectrin domain R15 (R15). A special feature of this chapter is the first application of transition path analysis tools to analyse experimental force spectroscopic data ([Section 6.2](#)). While R15 appears to be an ideal two-state folder, ROSS is shown to have

an extremely rough energy landscape. In [Section 6.4](#), this roughness is revealed to be imposed by three on-pathway intermediates. Various mutants and the realization of different pulling geometries are used to decipher the multidimensional folding mechanism of ROSS, see [Section 6.3](#). Furthermore, a transition-path-assisted protein friction determination based on the transition path ensemble deformation is introduced in [Section 6.4.2](#).

In the last part, the appendix, I provide additional information which was found to be very important in performing optical tweezers experiments successfully. Notably, details on the design and molecular cloning of coiled-coil unzipping constructs have been written down. These details include the choice of successful crosslinking positions, multiple sequence alignment for cysteine-free mutant design, silent mutations, and the design of a cloning vector with tryptophan-enhanced protein absorption. Furthermore, all important protocols from protein expression and purification up to sample preparation and measurement at the trap are provided with troubleshooting sections.

Owing to the novel nature of the applied analytical methods and the obtained results of this work, care has been taken to provide all necessary details and information to enable the reader to reproduce the latter and to understand all crucial lines of thought.



## Part II

### METHODS, THEORIES, AND DATA ANALYSIS

How we do what. What and how we evaluate. Which theories are used to conquer new territory.



EXPERIMENTAL SETUP AND IMPLEMENTATION

---

Probing the mechanics of a protein by single-molecule force spectroscopy is an experimentally challenging task as it involves the precise manipulation and detection of protein fluctuations on the nanometer scale. This chapter introduces the setup and assay which were applied to perform the protein mechanical studies underlying this thesis. Important technical details and experimental results crucial for assay development are included here. A key method, namely the construction of protein-DNA conjugates, is highlighted. For a first impression, data samples of measured protein mechanics are given along with a brief overview of their standard evaluation methods. The chapter closes with a set of data featuring experimental fingerprints of transition paths of protein folding which constitute a central topic of this work.

## 2.1 HOW TO APPLY FORCE TO A SINGLE PROTEIN MOLECULE

The possibility to trap particles using radiation pressure was first shown by Arthur Ashkin in the years 1969/70. In the late '80s he again reported the first single-beam trap for micron-sized dielectric particles as well as successful trapping and manipulation of single cells [5, 6, 7]. In the 1990s scientists started using optical tweezers for mechanical studies of single molecules, e. g., the movement of molecular motors, mechanical stretching of DNA, and the folding/unfolding of single proteins [14, 79, 113, 223, 235]. Since then, optical tweezers have become a well elaborated technique for performing force spectroscopic measurements at unprecedented resolution. Within a force range between 0.5 and 100 pN, protein conformational changes can be discerned at sub-nanometer resolution and characteristic timescales down to tens of microseconds can be resolved [38, 261]. This makes optical tweezers a highly valuable supplement with respect to alternative force spectroscopic techniques like the AFM or magnetic tweezers [181, 218], especially in the low force regime relevant for many cellular processes on the single-molecule level.

*Principle of optical tweezers*

The principle of optical tweezers is based on the interaction of electro-magnetic waves, e. g., the light of a laser beam, with dielectric and, hence, refracting particles like glass beads. To trap a bead in a focused laser beam the *scattering force* and the *gradient force* have to be accounted for. While the latter pulls the bead into the region of highest laser intensity, i. e., into the laser focus, the scattering force pushes the bead along the direction of beam propagation. In brief, if the gradient force can compensate the scattering force, the bead remains trapped slightly behind the laser focus. The forces acting on the bead



quadrant photodiodes (QPDs) owing to their better temporal resolution [166]. Last but not least, an additional fluorescence laser and a fluorescence detection are incorporated to distinguish between fluorescently labelled and unlabelled beads. By this means, the distinction between differently functionalized and at the same time differently labelled glass beads needed to establish the measurement configuration shown in Figure 6 is substantially facilitated.

More details concerning design considerations or the calibration procedure can be found here [34, 146], further characterization or serial numbers of individual parts here [166, 230].

#### *Operational settings and measurement conditions*

The stiffness calibration of the traps was performed using the method described in [225]. Trapping strengths were adapted to experimental demands<sup>2</sup> and ranged from 0.15 to 0.45 pN/nm for each individual trap with an estimated calibration error of 10%.

All measurements were performed in standard Phosphate Buffered Saline (PBS), pH 7.4, at a Room Temperature (RT) of about  $23 \pm 1$  °C ( $\approx 296$  K). However, the temperature at the position of the tethered protein also depends on laser-induced heating [168]. To reach a trapping strength of 0.3 pN/nm, an overall laser power of about 1.2 W was used yielding approximately 303 K at the measurement spot. This was independently verified by comparison to the temperature implied by the viscosity of the medium surrounding the trapped beads as provided by the calibration procedure [257].

Depending on the required temporal resolution, data were acquired at frequencies ranging from 20 to 200 kHz. Prior to recording, signals were filtered with a hardware low-pass filter set to a 3dB-frequency equal to half the acquisition frequency. Crosstalk due to the imperfect behaviour of the polarization beam splitter cubes and objectives [8] as well as crosstalk due to the proximity of the two trapped beads was subtracted after recording [82]. Final data analysis was performed on the differential signal for a better signal-to-noise ratio [150].

#### *2.1.2 Experimental geometry: the dumbbell assay*

The described dual-trap optical tweezers setup enables us to trap two individual beads, to measure the forces acting on each bead, and to change the inter-trap distance. Hence, by tethering an individual protein between the beads, single-molecule mechanical measurements can be performed. The tether is realized with two DNA-handles each connecting one specific site of the protein to one bead. This results in a dumbbell-like shape composed of a stretched bead-DNA-protein-DNA-bead construct as shown in Figure 6. Furthermore, Figure 6 also illustrates the three signals which are recorded to measure the mechanical properties of the tethered protein: the inter-trap distance

<sup>2</sup> Note that lower trap stiffnesses can improve the signal-to-noise ratio. For an explanatory illustration see FIGURE 5 in [141].

$d$  and the deflection of each bead out of its respective trap center labelled  $x_1$  and  $x_2$ . With  $R$  being the beads' radius, the extension of the stretched tether comprising both DNA-handles and the protein equals:

$$x_{\text{tether}} = d - 2 \cdot R - x_1 - x_2. \quad (1)$$

With the calibrated trap stiffnesses  $k_1$  and  $k_2$ , the force acting on the system equals<sup>3</sup>  $F = k_1 \cdot x_1 = k_2 \cdot x_2 = k_{\text{eff}} \cdot (x_1 + x_2)$ , where  $k_{\text{eff}} = (1/k_1 + 1/k_2)^{-1}$  is the effective spring constant.

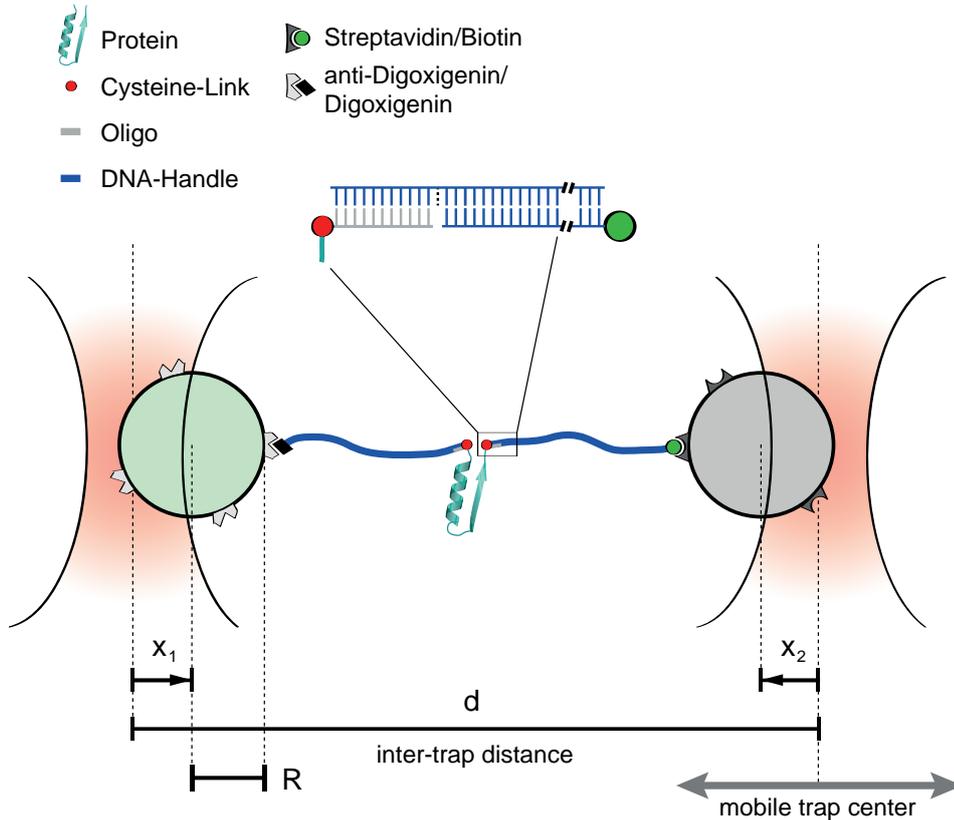


Figure 6: Schematic of the dumbbell configuration. Fluorescent labelling of the beads is illustrated by their green (labelled) versus gray (unlabelled) shade.

By carrying either a biotin or a digoxigenin label on one end, DNA-handles bind to dielectric beads that have been functionalized with streptavidin or anti-digoxigenin, respectively. The other end of the DNA-handles comprises a single-stranded (ss) overhang which is complementary to a ss-DNA oligonucleotide that has been covalently attached to the protein using cysteine-based chemistry. The two types of DNA-handles were produced by Polymerase Chain Reaction (PCR), for details see Section A.3.4. In brief, while sense primers carry either three biotin or three digoxigenin sites, anti-sense primers have an abasic site in the middle, where the polymerase falls off which, in turn, leads to the ss-overhang. The hybridization of a DNA-handle to an oligonucleotide bound to the protein is highlighted in the inset of Figure 6.

<sup>3</sup> Despite their opposing directions in Figure 6, here  $x_1$  and  $x_2$  are both considered positive for a force stretching the tether.

Beads were functionalized following an in-house protocol. The applied chemistry is based on EDC/NHS<sup>4</sup> crosslinking of the carboxylated silica beads with the primary amines of either streptavidin or digoxigenin-antibodies. In case of the latter additional TMR<sup>5</sup>-Bovine Serum Albumin is added as a fluorescent label for distinction between the two bead types.

Even though polystyrene beads have a higher refractive index allowing for higher trapping strengths to be reached, silica beads were found to offer improved performance in the dumbbell assay. The major reasons are less sample heating [168] and lowered damage to the sample due to a smaller amount of free oxygen radicals being produced [123]. Another reason is the apparently<sup>6</sup> improved force-clamping behaviour found for silica beads. When polystyrene beads were used, the observed persistence length of DNA-handles was only 10 nm [83, 146]. This value increased to more than 20 nm when silica beads were used under nearly<sup>7</sup> the same experimental conditions [216]. The size of the beads was chosen to be 1  $\mu\text{m}$  as a trade-off between handling and a fast system response. While smaller beads have a lower hydrodynamic drag with faster response times, they become more and more difficult to be seen in the brightfield and more difficult to handle during functionalization.

To reduce the potential damage induced by oxygen radicals a Glucose/Glucose-Oxidase/Catalase oxygen scavenging system (GODCAT) was used. However, a recent finding indicates significant nuclease contamination in many commercially available catalase stocks. Hence, it is very important to ensure the use of clean catalase [204]. In the long run, the protocatechuic acid/protocatechuate-3,4-dioxygenase oxygen scavenging system (PCA/PCD) may be a promising alternative to GODCAT [1, 123]. A detailed step by step protocol for the preparation of sample chambers, including the addition of oxygen scavenger, can be found in Section A.3.5.

## 2.2 BUILDING PROTEIN-DNA CONJUGATES

Originally, entire DNA-handles were directly attached to the protein to form the complete protein-DNA conjugates necessary to establish the dumbbell configuration shown in Figure 6. However, after the first DNA-handle is bound, it sterically shields the protein against the attachment of the second handle [39]. To avoid this, the preparation of protein-DNA conjugates was split up into two steps. First, short oligonucleotides (oligos) were attached to the protein forming the protein-oligonucleotide construct, and second, much larger DNA-handles with complementary ss-overhangs hybridized to the already bound oligos. While diffusing much faster, short oligos are expected to be less repulsive as well, resulting in higher yields for protein-oligonucleotide constructs and subsequent entire protein-DNA conjugates [82].

<sup>4</sup> 1-Ethyl-3-[3-Dimethylaminopropyl]Carbodiimide / N-HydroxySuccinimide

<sup>5</sup> Tetramethylrhodamine

<sup>6</sup> Until now, there is only the conjecture that polystyrene beads and/or their surface might be somehow slightly deformable; this was not further investigated.

<sup>7</sup> In between, the optical tweezers setup has been recalibrated.

The assembly of stable and pure protein-oligonucleotide (protein-oligo) constructs optimised for optical trapping setups is key to successful experiments. In the following, the principle of protein-oligo construct purification is briefly discussed. Further a reactivity test for maleimide oligos is introduced. More details on the optimization of maleimide as well as thiol oligo attachments including, e. g., protein activation, removal of oligo contamination, or the necessary characterization of the purification system, are provided in the protocol section of [Appendix A](#).

### 2.2.1 Protein-oligonucleotide construct purification

Within the frame of this work, either thiol or maleimide oligonucleotides were attached to thiol groups of cysteines engineered into the protein sequence at the two positions, where the external force will be applied. For the attachment reaction, protein and either thiol or maleimide oligos are mixed in a 1:1 ratio between oligos and cysteines, for details see [Section A.3.2](#). Products of an attachment reaction are purified by Size Exclusion Chromatography (SEC) to separate unreacted oligos, unreacted protein, and protein with only one attached oligo from the desired protein-oligo construct with two oligos.

[Figure 7A](#) shows a typical chromatogram generated during purification of a protein + maleimide oligo reaction mixture in comparison to reference runs of the individual reactants, namely pure protein and pure oligo(s). Recorded signals reflect the absorption of the eluate leaving the chromatography column at the indicated time. To facilitate the comparison of different runs, detected absorption signals were 'normalized' by setting their enclosed area<sup>8</sup> equal to one resulting in some kind of relative absorption. [Figure 7C](#) presents the corresponding 260 nm/280 nm absorbance ratios which reveal high sample purity in the main peaks of the reactants. While the ratio of the purified attachment reaction confirms that the absorbance signal of oligo containing products is dominated by the oligos, the dip around 35 minutes originates from unreacted protein. Owing to the dominant oligo signal, the difference of normalized absorption signals of attachment reaction and pure oligos helps to identify the peak of the desired protein-oligo construct as shown in [Figure 7B](#).

In case of very small proteins, a precise and well characterized purification system is crucial especially for getting rid of dimerized oligos which drastically corrupt measurement efficiency. The necessary measures taken to circumvent this problem are presented in [Section A.3.3](#).

In order to provide a well-defined environment for the oligo attachment and to improve protein solubility, additional proteins like ubiquitin can be genetically inserted at the N- and/or C-terminus of the protein under study. For proteins with contiguous termini, e. g., coiled coils, the Ferredoxin-like fold, or the Rossmann fold, additional proteins are used to spatially separate the oligo attachment positions and prevent them from (intra-)crosslinking.

<sup>8</sup> For SEC parameters as indicated in [Figure 7](#), the area enclosed by the recorded signals was typically evaluated between 10 and 40 minutes.

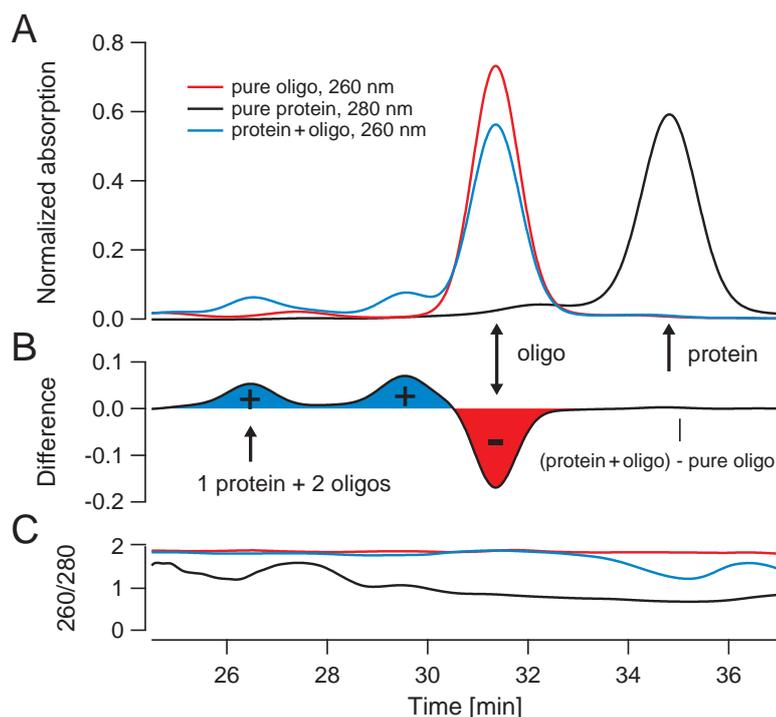


Figure 7: SEC purification of an oligonucleotide attachment. SEC parameters: Superdex 200 column, 0.5 mL/min, 100  $\mu$ L-loop, PBS. (A) Superimposed 'normalized' SEC runs show pure protein, pure oligos, and an oligo attachment reaction. (B) The difference between attachment reaction and pure oligos visualizes the elution of individual (+)-marked products. The first peak corresponds to protein with two bound oligos, the second to protein with one oligo. (C) 260 nm/280 nm ratio of each SEC run. The pure protein peak approaches a typical value around 0.5. The peaks of the other SEC runs show a ratio of more than 1.8, indicating that absorption is dominated by the contribution of DNA.

An important requirement for suitable 'spacer' proteins is a high mechanical stability to avoid interference with the actual mechanical measurements. While ubiquitins are known to fulfil this requirement [83, 108, 188, 216, 261], in this thesis it was shown that GB1 is less favourable as a spacer protein. Even though ubiquitin and GB1 share a common fold and GB1 unfolds at slightly higher forces in mechanical AFM-studies [35, 195], it undergoes slow equilibrium fluctuations on a time scale of many seconds up to minutes at a constant force bias of about 10 pN, see Figure 69A.

### 2.2.2 Maleimide reactivity test

The use of maleimide oligos for an oligo attachment has as a couple of advantages over thiol oligos. Besides their faster reaction with the thiols of the protein's cysteines, maleimide oligos do not require activation or ethanol precipitation as opposed to thiol oligos. Further, protein activation based on Tris(2-carboxyethyl)phosphine (TCEP) is easier than the activation based on

2,2'-Dithiodipyridine (DTDP) solved in dimethyl sulfoxide (DMSO) where a risk of precipitation followed by protein loss has been observed. Most importantly, the maleimide oligo attachment reaction will not be completely inhibited with TCEP being present [84]. As the presence of reducing agent prevents neighbouring cysteines from crosslinking, the necessity of additional spacer proteins for proteins with contiguous termini can be circumvented.

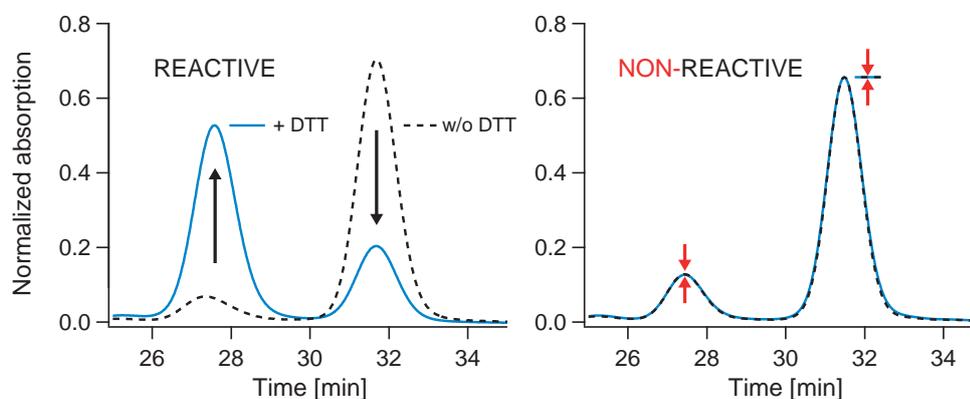


Figure 8: Maleimide oligonucleotide reactivity test with DTT as an idealized protein mimic. SEC parameters: Superdex 200 column, 0.5 mL/min, 100  $\mu$ L-loop, PBS. Shown signals reflect the relative absorption at a wavelength of 260 nm.

However, the basic requirement for the maleimide oligo attachment to work is, of course, reactivity of the maleimide itself. As this was not guaranteed by the supplier, a reactivity test has been developed where Dithiothreitol (DTT) is used as an idealized protein mimic. When added to maleimide oligos in a 2:1 ratio between oligos and DTT, the two thiol groups of the DTT can each react with one oligo resulting in dimerized oligos. If the oligos are reactive, a clear gain in the amount of dimerized oligos after addition of DTT can be visualized by SEC while non-reactive oligos appear indifferent as shown in Figure 8. An alternative to SEC is to determine the remaining reactive thiols after incubation of (excess) DTT with maleimide using Ellman's reagent [69]. For both tests, the limited half-life of DTT needs to be accounted for [211].

#### Screening maleimide-thiol reaction conditions

The developed reactivity test allows the screening of various parameters which influence the efficiency of maleimide-thiol reactions. An overview of representative data is given in Figure 9.

**Time.** Increasing the time<sup>9</sup> from a few up to 30 minutes increases the amount of desired product, i. e., the height of the oligo dimer peak as shown in Figure 9A. A series of measurements comparing 1, 2, 4 and 24 hours lasting reactions yielded no further product increase (data not shown). For proteins 2 h are typically enough, see Section A.3.2.

<sup>9</sup> Indicated reaction times start right with initial mixing in a minimal reaction volume and end with a quick dilution instantly followed by the actual SEC run. Until not being separated by size, DTT and maleimide oligos can continue to react during the initial phase of the SEC run. By this means, indicated reaction times are slightly underestimated.

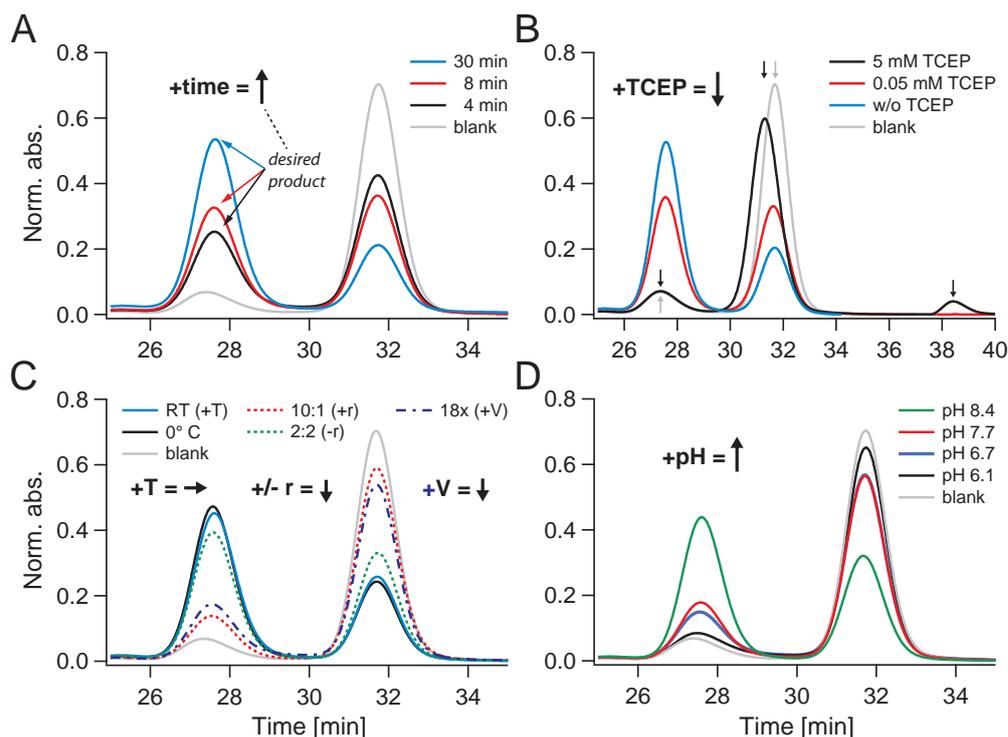


Figure 9: Screening maleimide-thiol reaction parameters. **SEC** parameters: Superdex 200 column, 0.5 mL/min, 100  $\mu$ L-loop, **PBS**. Signals reflect the relative absorption at a wavelength of 260 nm. If not stated differently, two **PBS** pH 6.7 solutions containing 0.1 mM maleimide oligos and 0.1 mM **DTT** were mixed at a ratio of 2:1 and reacted for 30 min at **RT** prior to final dilution for **SEC** purification. (A) Increasing reaction times from 4 to 30 min increases the desired product amount. (B) Adding different amounts of **TCEP**. (C) Increasing the reaction volume by immediate 18x dilution corresponds lowering concentrations. The effect of temperature and reactant ratio on final product amounts are monitored as well. (D) Screening the pH. Runs in (A) and (B) use the same oligo batch. (C) and (D), each use different batches.

**TCEP.** As seen in Figure 9B, addition of similar molar amounts of **TCEP** as compared to the reactants shows a reduction of reaction efficiency as expected in [84]. High excess of **TCEP** completely inhibits the reaction and reveals that **TCEP** appears to interact with maleimide. Similar to the finding of high molecular weight **TCEP** after reaction with maleimide in [206], here, a shift of the monomer oligo peak towards higher molecular weight is observed and indicated by two small parallel arrows. The small arrow pointing at a peak around 39 minutes highlights the characteristic elution volume of **TCEP**.

**pH.** Screening of the pH confirms high thiol reactivity at higher pH, see Figure 9D. To maintain specificity of maleimide reactions with thiols, the pH needs to be kept within the pH range of 6.5 - 7.5 [97].

**Temperature.** Working on either ice or at **RT** does not affect efficiency under the indicated reaction conditions, see Figure 9C. This is another advantage of maleimide over thiol oligos.

**Volume.** Increasing the reaction volume by immediate dilution of an otherwise standard reaction corresponds to using lower concentrations. As a result, fewer oligo dimers are detected after 30 minutes.

**Ratio.** To realize as many oligo dimers as possible, the optimal mixing ratio between maleimide oligos and DTT is 2:1 since a single DTT carries two thiols. As expected, deviations from the optimal ratio reduce reaction efficiency.

### 2.2.3 Bioorthogonal chemical coupling alternatives

As an alternative to thiol or maleimide oligos, azide oligos can be used in a copper-free click chemistry approach where the azide modified oligos attach to dibenzocyclooctyne-maleimide bound to the protein's cysteines [11, 257]. This approach still utilises the cysteines within the protein, as is the case for the thiol and maleimide oligos. To be able to use cysteines within the protein sequence for different purposes like, e. g., fluorescent labelling with dyes, other oligo attachment approaches are needed.

Here, three considerable examples have been selected from the vast field of bioorthogonal chemistry, namely the HaloTag, bacterial sortase, and unnatural amino acids. A general overview of the field can be found here [207].

The HaloTag covalently and irreversibly binds a synthetic ligand consisting of a reactive linker and, in principle, any functional group. If the functional group is used to bind to an oligo, the oligo can be specifically bound to the binding pocket of the HaloTag through the reactive linker [133]. Being a protein fusion tag, the HaloTag offers an easy implementation by genetic fusion to the N- and C-termini of the protein of interest. A mechanical AFM-study of the HaloTag protein suggests that it is stable enough to not unfold at forces typical for equilibrium measurements with optical tweezers [171].

A second approach uses sortases. These enzymes anchor cell surface proteins to the cell wall. In case of *Staphylococcus aureus* sortase A, a short C-terminal recognition motif (LPXTG) on the target protein is cleaved, followed by the formation of a covalent amide bond with the pentaglycine (GGGGG) cross-bridge in the cell wall [173]. It has been shown that this mechanism can be exploited for site-specific N- and C-terminal labelling of a single protein by using sortases of different specificity [3]. Hence, this approach may be even used to attach oligos with different sequences.

Finally, the expansion of the genetic code has enabled the incorporation of unnatural amino acids with novel structural, chemical, and physical properties into the sequence of proteins [234]. This way, the click chemistry approach with azide modified oligos can be used without depending on cysteines. So far, up to three distinct labels could be attached to a single protein using incorporated unnatural amino acids [129]. Besides their application in labelling, artificial amino acids have been successfully used for photo-induced intercrosslinking of protein substructures, which can be useful for the structural interpretation of unfolding pathways [80].

## 2.3 EXPERIMENTAL MODES OF FORCE PROBING

Prior to performing force spectroscopic measurements, the dumbbell configuration shown in [Figure 6](#) needs to be established. As this comprises the experimentally most important step, detailed instructions on successful tether formation including sample preparation, measurement mixes and troubleshooting sections, are given in [Section A.3.5](#).

*Constant velocity mode*

Each force spectroscopic measurement starts in constant velocity mode. In this mode, the mobile trap is repeatedly moved away from and back towards the fixed trap. This repeated movement is performed at constant velocities typically ranging from 10 to 1000 nm/s. After probing for a successful formation of the dumbbell configuration, constant velocity mode measurements are used to screen typical force ranges of unfolding and refolding of the protein under study.

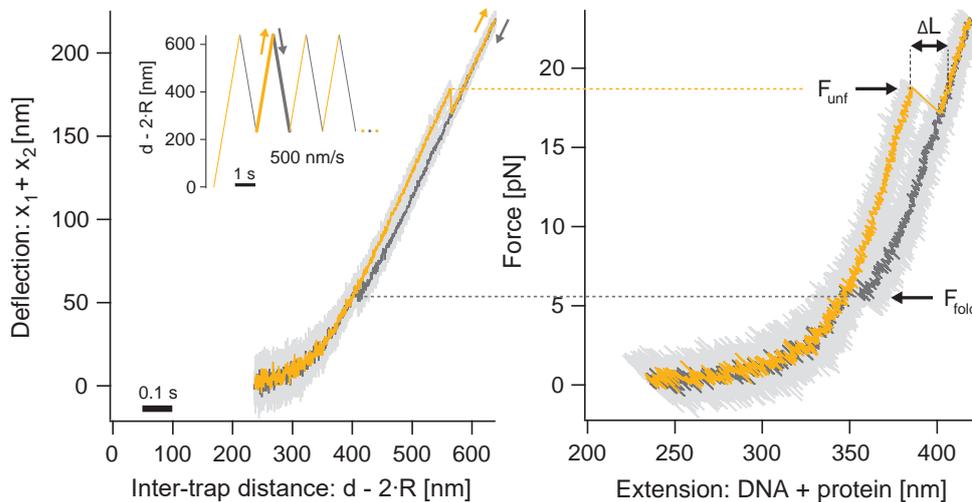


Figure 10: A force-extension trace measured in constant velocity mode. Intense colours reflect moving-average filtered traces based on the full resolution data shown in light gray. **Left:** Overall deflection  $x_1 + x_2$  out of the trap centres versus inter-trap distance  $d$  minus the two beads' radii  $2 \cdot R$ . The shown trace comprises an entire stretch-and-relax cycle (orange and dark gray) corresponding to the highlighted cycle in the inset. *Inset:* Inter-trap distance cycles over time at 500 nm/s. **Right:** Measured signals converted into a force ( $F$ ) versus extension ( $x_{\text{tether}}$ ) trace, where  $F = k_{\text{eff}} \cdot (x_1 + x_2)$  and  $x_{\text{tether}} = d - 2 \cdot R - x_1 - x_2$ , see [Figure 6](#).

A representative trace of a so called stretch-and-relax cycle is shown in [Figure 10](#). The given force versus extension relation of the entire tether is derived from the recorded inter-trap distance and bead deflection signals. When the inter-trap distance is increased, the ends of the entire tether are pulled apart and the force acting along the direction of movement rises. This leads to an increasing mechanical destabilization of the tethered protein which finally re-

sults in its unfolding. Each unfolding event is characterized by the force  $F_{\text{unf}}$  of its occurrence and attended by a characteristic contour-length increase  $\Delta L$ . As the latter allows the beads to relax towards their respective trap centres, a sudden drop in force indicates unfolding. After reaching a maximal inter-trap distance, the mobile trap is moved back towards the fixed trap which relaxes the force acting on the tether. If the unfolded protein is able to refold against an acting force, a sudden increase of force due to a reduced contour-length indicates refolding.

The contour-length changes upon protein unfolding and refolding are determined by mechanical fits, see [Section 3.1](#). These fits allow a transformation of the force-extension trace into the protein's contour-length space to be performed, where even very short-lived intermediate states within unfolding or refolding transitions can be identified [12]. Finally, the detected length changes can be assigned to the unfolding of a certain number of amino acids. This permits a structural interpretation of the observed (protein) states with respect to the protein's solved crystal or NMR structure [59, 60].

After a tether has formed, hundreds of stretch-and-relax cycles can be performed on a single protein molecule. The obtained force distributions for unfolding (u) and (re-)folding (f) depend on the respective transition state position  $\Delta x_{\text{u/f}}$  and the intrinsic rate constants  $k_0^{\text{u/f}}$  and, hence, reveal information about the underlying energy landscape [103, 196]. Based on the obtained rate constants, barrier heights can be estimated according to the Arrhenius equation [4]. By varying the pulling velocity, the loading rate dependence of the average unfolding and refolding forces can be examined to gain further insight into the energy landscape shape and dimensionality [194, 197]. Despite the non-equilibrium nature of constant velocity mode experiments, there are methods to assess equilibrium free energies  $\Delta G_0$  from their data by using, e. g., Jarzynski's equality or the Crooks fluctuation theorem [51, 112]. Furthermore, an extension of Jarzynski's equality has been shown to reconstruct entire energy landscape profiles [91, 102]. Alternatively, if both rate constants are known, estimates for  $\Delta G_0$  can be obtained by applying the principle of detailed balance [99]. Moreover, if stretch-and-relax cycles are considered as multiple succeeding measurements at a locally constant force bias, even force-dependent rates  $k^{\text{u/f}}(F)$  can be calculated from long enough series of cycles [161]. In [Chapter 3](#) some of the aforementioned concepts are presented in more detail.

#### *Constant distance (passive) mode*

When both traps are kept at constant distance, no additional external work is performed on the (established dumbbell) system, which, hence, is in thermal equilibrium. By tuning the distance, a desired constant force bias can be applied to the tethered protein. Provided that the folding barrier is low enough to allow unfolding and refolding to occur at similar forces, conformational equilibrium fluctuations between the folded and unfolded states of a protein can be traced over time.

Typical time traces revealing conformational fluctuations of a protein are shown in Figure 11. As these fluctuations usually manifest themselves in changes of the protein's end-to-end distance, they can be observed as abrupt changes in the overall bead deflection signal similar to the unfolding and re-folding events in the above described stretch-and-relax cycles. By this means, each populated conformational state automatically has its individual force bias. This passive experimental approach yields superior temporal as well as spatial resolution as opposed to techniques where the same force bias for all occurring states is achieved either by an active regulation of the inter-trap distance, or by measuring the deflection within the flat (beyond linear) region of the trapping potential of one of the two traps (which is set to a much lower trapping strength than the other) [89, 124].

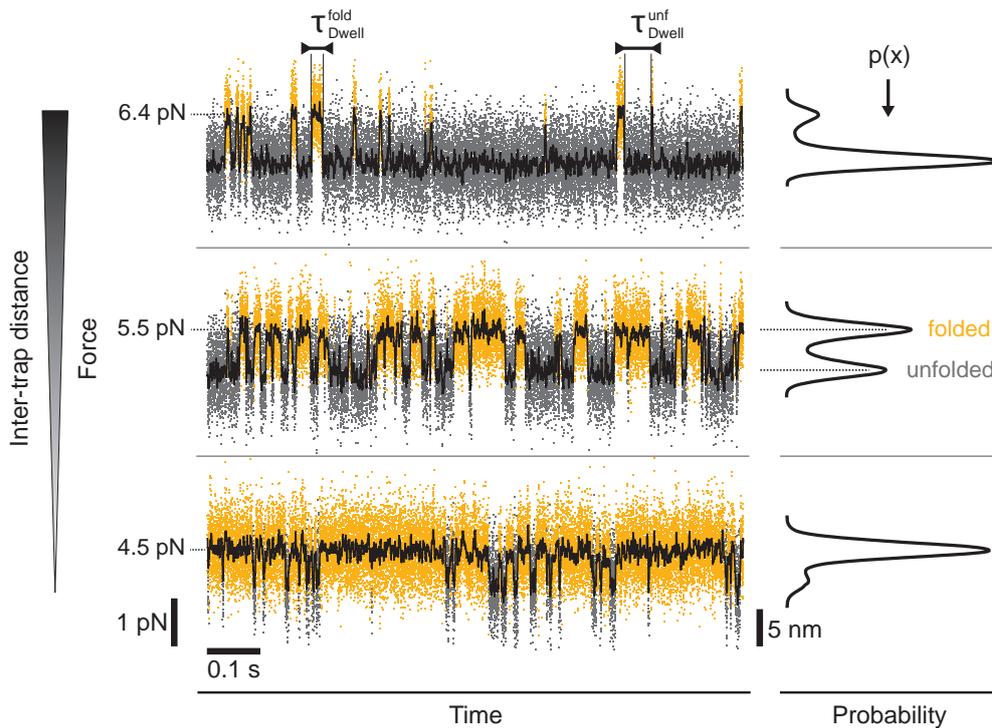


Figure 11: Time traces measured in constant distance mode. The population probabilities of the folded and unfolded states shift with force.

In particular, the time traces presented in Figure 11 show the fluctuations of a two-state folder at three different pretensions. To assign a state to each data point, a Hidden Markov model (HMM) is applied to each trace (see Section 3.4.3 for details) yielding states indicated by different colours [175, 216]. From each assignment, the dwell times  $\tau_{\text{Dwell}}^i$  spent in state  $i$  at the given force bias can be readily collected to obtain a dwell time distribution for this state. While the upper level with higher force represents the folded state (orange), the lower level represents the unfolded (gray), more elongated state. By increasing the force bias, the population shifts from predominantly folded

to predominantly unfolded. This is directly reflected in the change in the corresponding probability distributions<sup>10</sup>  $p(x)$ .

Stepping through the force range where equilibrium fluctuations can be observed, the dwell time distributions and state occupancies can be used to derive force-dependent microscopic kinetic rates  $k^{ij/j^i}(F)$  as well as equilibrium free energies  $\Delta G_0^{ij}$  between two states  $i$  and  $j$ . Different models ranging from extensions of Bell's expression to more elaborate models based on Kramers' theory can be used to fit the chevron plots of rates versus applied force [15, 116]. These fits reveal transition state positions, extrapolated zero-force rate constants, and even energy barrier heights which are key features related to the energy landscape describing the folding mechanism of the protein under study [67, 83, 196, 247].

However, irrespective of the model being used, the effective diffusion constant or roughness parameter necessary to further quantify the energy landscape remains elusive without additional assumptions or information beyond the force-dependence of rates. One way of gaining further information is the deconvolution of the probability distributions from constant distance data to subtract all additional thermal and mechanical contributions not originating from the protein. This has been successfully used to reconstruct the entire energy landscape shape including widths and curvatures along with energy barrier heights for DNA-hairpins as well as coiled coil proteins [83, 98, 243]. Recently, after having been introduced for testing the reaction coordinate quality, the calculation of the splitting probability has been shown to recover the approximate shape and height of the transition barrier directly from constant distance data without deconvolution [42, 143].

After the mechanics of a protein have been measured under a certain set of conditions, it seems straightforward to simply change these conditions to find out more about the protein under study. Possible ways of doing so are given by varying salt or nucleotide concentrations, or by adding ligands, isomerases or chaperones into the surrounding solution. From the observation of altered states and stabilities induced by the new conditions, conclusions on affinities and interactions between multiple partners can be drawn [12, 167, 188, 213, 215]. In all cases, the ability to measure under a stable, constant force bias over tens of minutes at very high spatial and temporal resolution is key.

#### 2.4 FROM TRANSITIONS TOWARDS TRANSITION PATHS

A transition path is the interconnection between two states and can be regarded as the transition event itself. Usually, the transition between different conformational states of a protein is so fast that its signature within force spectroscopic data would essentially only reflect the damped relaxation of the trapped beads as a response to the conformational change of the investigated protein. Nevertheless, by looking at transitions of a protein at very high temporal resolution, we found transition path times of about a millisec-

<sup>10</sup> The probability distributions in Figure 11 correspond to the black time traces - a moving-average filtered version of the coloured 30 kHz data with a window size of 1.7 ms.

ond under a force of around 10 pN, where relaxation times of only some tens of microseconds are expected based on reported values from autocorrelation analysis [261].

Examples of transition paths observed in measurements performed in either constant distance or constant velocity mode are shown in Figure 12. The long lasting transition path times  $\tau_{TP}$  indicate that the studied protein's unfolding and refolding is slowed down due to the diffusion across multiple barriers in a rough energy landscape [249].

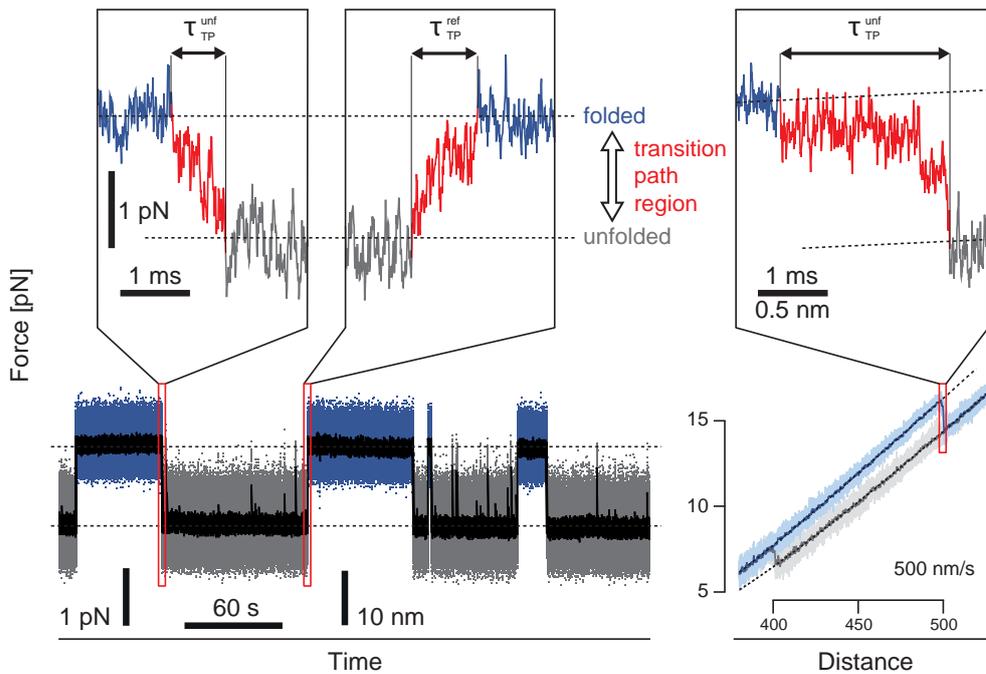


Figure 12: Transition paths from highly resolved transitions. On the left and right are zooms into data recorded in constant distance and constant velocity mode. Even for data recorded at 500 nm/s the force bias is almost constant while passing the transition path region.

As transition paths reflect the system's response during the event of barrier crossing(s), they must contain all the information necessary to characterize the underlying energy landscape. From this point of view, a method called transition path sampling has been developed for simulation studies to reduce sampling down to the mere events of barrier crossing making long lasting simulations of equilibrium fluctuations obsolete [28, 55]. From the transition path ensemble comprising many individual transition paths, the transition state position and rate coefficients can be derived [20, 101]. Just recently, the transition path analysis approach has been reported to identify the transition state position for DNA-hairpin folding/unfolding [155]. An introduction to transition path theory and an overview of transition path analysis techniques is given in Section 3.8. Based on Langevin dynamics simulations, the potential and limitations of these techniques, when being applied to force spectroscopic data, are discussed in Section 4.3.4. Finally, the results presented in Chapter 6 provide the first-time application of transition path analysis tools

to experimental data including transition-path-assisted energy landscape reconstruction and a model-free diffusion-based protein friction determination.

## THEORY AND ANALYSIS OF SINGLE MOLECULE EXPERIMENTS

---

To study the mechanics of an individual protein a comprehensive understanding of the thermodynamics and kinetics of single molecule reactions is necessary. A highly recommendable introduction into the basic principles governing single molecule reactions probed by force can be found here [224]. More detailed explanations and derivations directly related to models and methods developed or adapted and used at the biophysics chair in Garching can be retrieved from numerous theses, especially [19, 29, 193, 212, 257]. This chapter gives a compact overview of the analysis tools on hand for single-molecule force spectroscopy data. In Section 3.8, this 'classic' toolbox is expanded towards transition path analysis techniques.

### 3.1 POLYMER MECHANICS

When a protein loses its native form upon forced unfolding, the force response of the remaining unstructured polypeptide can be described with the Worm-like Chain (WLC) model [117]. An extrapolation formula suitable for fitting force-extension traces of unfolded proteins is given by [33]:

$$F_{\text{WLC}}(x_P, [L_P, p_P, T]) = \frac{k_B T}{p_P} \left[ \frac{x_P}{L_P} + \frac{1}{4 \left(1 - \frac{x_P}{L_P}\right)^2} - \frac{1}{4} \right]. \quad (2)$$

Here,  $x_P$  is the extension of the unfolded protein defined as its end-to-end distance in space. The contour length  $L_P$  is the actual length of the backbone. The persistence length  $p_P$  is a measure for the distance over which the polypeptide's backbone orientation is correlated. Finally,  $T$  is the temperature and  $k_B$  the Boltzmann constant.

In addition to the protein under study the trap constructs comprise DNA-handles as well. Their force response can be reproduced with an extensible Worm-like Chain (eWLC) model [235]:

$$F_{\text{eWLC}}(x_D, [L_D, p_D, T, K]) = \frac{k_B T}{p_D} \left[ \frac{x_D}{L_D} + \frac{1}{4 \left(1 - \frac{x_D}{L_D} + \frac{F}{K}\right)^2} - \frac{1}{4} - \frac{F}{K} \right]. \quad (3)$$

With respect to the WLC model, here, the stretch modulus  $K$  is introduced to account for the additional extensibility of DNA.

*Contour length gains from force spectroscopy and crystal structures*

Within the time resolution of our experiments the entire trap construct comprising both DNA-handles and the protein can be assumed to experience the same force, i. e.,  $F_{\text{eWLC}} = F_{\text{WLC}} = F$ . The overall measured extension of the tether  $x_{\text{tether}}$  can then be written as the sum of the individual extensions:

$$x_{\text{tether}} = x_{\text{D}}(F) + x_{\text{P}}(F) \cdot P_{\text{unf}}(F). \quad (4)$$

Here the extension of both DNA-handles has been summed up into  $x_{\text{D}}(F)$  which represents the inverse of Equation 3. The protein contributes  $x_{\text{P}}(F)$ , the inverse of Equation 2, to the overall extension upon unfolding as indicated by multiplication with the respective probability  $P_{\text{unf}}(F)$ . Typically, all DNA parameters can be retrieved from a fit to the force-extension trace, where the protein is still folded. Unfolding of the protein triggers a sudden gain in contour length of the tethered construct, the extension of which is then fitted with a WLC model with a fixed value for the protein's persistence length  $p_{\text{P}}$  in series with an eWLC.

The choice of an appropriate persistence length is not obvious. So far, values ranging from 0.35 up to 0.9 nm have been reported [60, 166]. A possible way to obtain a reasonable persistence length is by reducing the recorded force-distance data to the mere contribution of the protein which can then be directly fitted with the WLC model [166, 257]. Note that the persistence length  $p_{\text{P}}$  directly affects the contour length gains obtained from WLC fits as well as the resulting average contour length gain per amino acid  $d_{\text{aa}}$  (see below) [257]. Likewise, the equilibrium energy calculations presented in Section 3.2 are influenced by the choice of protein persistence length.

For a structural interpretation of measured contour length gains, a comparison to values calculated from the protein's crystal structure is key. Expected contour length gains are calculated as follows:

$$\Delta L_{\text{P}}^{\text{calc}} = N_{\text{aa}}^{\text{unf}} \cdot d_{\text{aa}} - d_{\text{init}}^{\text{fold}} + d_{\text{remain}}^{\text{fold}}. \quad (5)$$

The number  $N_{\text{aa}}^{\text{unf}}$  of amino acids involved in unfolding is multiplied with an average contour length per amino acid  $d_{\text{aa}}$  which is assumed to equal 3.65 Å [59, 60]. The resulting value is reduced by the initial distance  $d_{\text{init}}^{\text{fold}}$  between the two contact points of acting force before unfolding. If a part of the protein still remains folded after an observed contour length gain, the associated spatial distance  $d_{\text{remain}}^{\text{fold}}$  along the axis of acting force needs to be added. The distances  $d_{\text{init}}^{\text{fold}}$  and  $d_{\text{remain}}^{\text{fold}}$  are retrieved from the respective PDB structures<sup>1</sup>.

### 3.1.1 Contour length transformation

Usually contour length gains obtained from WLC fits are averaged or plotted in a histogram and fitted with a Gaussian distribution to identify the contour

<sup>1</sup> Distances within PDB structures were measured using PyMOL software which can be found under: <https://www.pymol.org/>.

length of an individual state from force-extension traces. A much more precise and advantageous method to find the contour length, especially for short lived and rarely populated intermediate states, is by building accumulated histograms of all unfolding or folding transitions which include all available data points. To create this quasi non-equilibrium transition path ensemble<sup>2</sup> from transitions occurring at different forces, the force-dependence of the extension needs to be overcome by transforming the force-extension traces into force-independent contour space. This contour length transformation will be outlined below. So far, this method has been proven useful for the investigation of short-lived intermediates upon the unfolding of, e. g., knotted proteins or DnaK [12, 257].

The relation between an acting force and a protein's extension  $x_P$  can also be interpreted as a dependence of the force on the relative extension  $X_P = \frac{x_P}{L_P}$ . Hence, by replacing the prefactor  $\frac{k_B T}{p_P}$  with  $Z$ , Equation 2 can be rewritten as:

$$F(X_P) = Z \left[ X_P + \frac{1}{4(1-X_P)^2} - \frac{1}{4} \right]. \quad (6)$$

After using  $Y = \frac{F}{Z}$  and regrouping the equation, a cubic expression in  $X_P$  is obtained:

$$X_P^3 - \left( 2 + \left( Y + \frac{1}{4} \right) \right) X_P^2 + \left( 1 + 2 \left( Y + \frac{1}{4} \right) \right) X_P - Y = 0. \quad (7)$$

With further simplifications and substitutions, Equation 7 can be transformed into a reduced cubic equation which can be solved using the Cardano formula as shown in detail in [19]. The analytical solution for  $X_P(F)$  is used to calculate the contour length of an unfolded protein with an extension  $x_P$  at force  $F$  according to:

$$L_P(F, x_P) = \frac{x_P}{X_P(F)}. \quad (8)$$

Again, an appropriate value for the persistence length  $p_P$  has to be chosen. The inverse of Equation 2 indicated as  $x_P(F)$  in Equation 4 is, in principle, simply a conversion of Equation 8. In practice,  $x_P = x_{\text{tether}} - x_D(F)$  where  $x_{\text{tether}}$  is measured according to Equation 1 and  $x_D(F)$  is implied by  $X_D(F)$  derived from Equation 3 using the same formalism as indicated for  $X_P(F)$ .

The contour length transformation formalism can, of course, also be directly applied to constant distance traces.

### 3.2 FREE ENERGY CALCULATION

To calculate the free energy of the entire dumbbell system comprising two trapped beads, two DNA-handles and a protein, it is sufficient to treat the system as is if it only consisted of one bead, one handle, and one protein [82]. While the dual-trap can be described by a single effective trap with an

<sup>2</sup> The transition path ensemble is introduced in Section 3.8.1.

effective spring constant  $k_{\text{eff}} = (1/k_1 + 1/k_2)^{-1}$  and a bead displacement of  $x = x_1 + x_2$ , the two DNA-handles are readily described by a single handle with twice the contour length. Consequently, the energy stored in the dumbbell system consists of the Hookean energy stored in the displacement of the effective trap  $G_{\text{bead}}$ , the mainly entropic energies stored in stretching the DNA-handle  $G_D$  as well as unfolded protein  $G_P$ , and the free energy of folded protein  $G_0$ . When the protein is in state  $i$  with the entire system being at a force  $F_i$  all contributions to the system's free energy can be summed up:

$$G^i(F_i) = G_0^i + G_{\text{bead}}^i(F_i) + G_D^i(F_i) + G_P^i(F_i) = G_0^i + G_{\text{pot}}^i(F_i). \quad (9)$$

As an intrinsic property, the free energy  $G_0^i$  of the folded part of the protein in state  $i$ , is force-independent. All other contributions including the stretching of the unfolded part of the protein, which is on hand in state  $i$ , depend on the force  $F_i$ . In Equation 9, these contributions are summarized in an overall potential energy  $G_{\text{pot}}^i(F_i)$  stored within the dumbbell configuration upon stretching. In detail, these contributions are<sup>3</sup>

$$G_{\text{bead}}^i(F_i) = \int_0^{x_i(F_i)} dx F_{\text{spring}}(x) = \frac{1}{2} k_{\text{eff}} \cdot x_i^2(F_i) = \frac{1}{2} F_i \cdot x_i(F_i), \quad (10)$$

$$G_D^i(F_i) = \int_0^{x_D(F_i)} dx F_{\text{eWLC}}(x), \quad (11)$$

$$G_P^i(F_i) = \int_0^{x_P^i(F_i)} dx F_{\text{WLC}}(x). \quad (12)$$

When a protein undergoes a conformational change from state  $i$  to state  $j$  the accompanying contour length change causes a change in acting force from  $F_i$  to  $F_j$ . The resulting free energy difference is:

$$\Delta G^{ij}(F_i, F_j) = G^j(F_j) - G^i(F_i) = \Delta G_0^{ij} + \Delta G_{\text{pot}}^{ij}(F_i, F_j). \quad (13)$$

The different elastic contributions to  $\Delta G_{\text{pot}}^{ij}(F_i, F_j)$  are nicely illustrated in [19, 193]. While taking care to choose the right bounds for integration to account for positive and negative energetic contributions, the resulting energy difference can be shown to equal the area enclosed by corresponding stretch-and-relax cycles [19]. Under force these contributions build up an additional mechanical barrier for folding as well as unfolding. The force-dependent composition of this barrier is explicitly plotted in [193], which allows, e. g., the discussion of the influence of effective spring constants on observed rates.

<sup>3</sup> In practice, the polymers' stretching energies are calculated from the inverse force functions yielding, e. g., in case of Equation 11:  $G_D(F) = F \cdot x_D(F) - \int_0^F x_D(F') dF'$ .

### 3.2.1 Force-dependent probabilities

If the dumbbell system is kept at equilibrium over a certain amount of time during a constant distance measurement, probabilities of being in different states identified at different bead positions  $x$  can be derived. The ratio of their Boltzmann distributions (see Equation 17) allows the force-dependent probabilities to be directly related to the free energy differences between the respective states:

$$\frac{P_j(x_j, F_j)}{P_i(x_i, F_i)} = \exp\left(-\frac{\Delta G^{ij}(F_i, F_j)}{k_B T}\right). \quad (14)$$

By exploiting the fact that the sum over all probabilities equals 1 and by making use of Equation 13 and Equation 14 we can express the probability of finding the system in state  $i$  in terms of free energy differences ( $\beta^{-1} = k_B T$ ):

$$P_i(F_i) = \frac{1}{1 + \sum_{j \neq i} \exp\left(-\beta \left(\Delta G_0^{ij} + \Delta G_{\text{pot}}^{ij}(F_i, F_j)\right)\right)}. \quad (15)$$

By this means, as soon as the probabilities are known along with all elastic energy contributions already introduced above, Equation 15 can be used as a fit function to derive the folding free energy values  $\Delta G_0^{ij}$ . Typically, the state probabilities are calculated from the sum of all lifetimes spent in a state identified by HMM analysis (see Section 3.4.3) divided by the entire length of the respective constant distance trace. Errors accounting for the finite measurement time are estimated as the standard deviation of probabilities obtained from multiple Monte Carlo simulations performed under the assumption of the measured transition rate constants.

## 3.3 FITTING EQUILIBRIUM UNFOLDING/FOLDING TRANSITIONS

Already early force-spectroscopic measurements using the AFM technique revealed fast equilibrium unfolding/folding transitions to show up as characteristic kinks or humps within stretch-and-relax cycles [182]. Here, a model for fitting such recorded traces by linking them to the entire system's energy function is briefly outlined in accord with [9, 261].

The force at inter-trap distance  $d$  equals the effective spring constant  $k_{\text{eff}}$  multiplied with the measured mean value  $\langle x \rangle$  of the overall bead deflection signal. This mean value can be rewritten according to its definition as first moment where  $p(x)$  is the distribution of  $x$ , see also Equation 44:

$$F(d) = k_{\text{eff}} \cdot \langle x \rangle = k_{\text{eff}} \cdot \int_{-\infty}^{+\infty} dx x \cdot p(x). \quad (16)$$

While considering  $x$  as a continuum of states,  $p(x)$  is linked to the entire system's energy function  $H(x)$  by a Boltzmann distribution where states of lower energy have a higher probability of being occupied:

$$p(x) = \frac{1}{Z} \cdot e^{-\beta \cdot H(x)}. \quad (17)$$

Here,  $Z = \int_{-\infty}^{+\infty} dx e^{-\beta \cdot H(x)}$  is the canonical partition function comprising all accessible states. Now only an appropriate energy function or ‘Hamiltonian’ describing the entire system still needs to be formulated. Without loss of generality it is assumed that the recorded equilibrium transition reflects the fast and independent unfolding/refolding of individual subunits each contributing an equal contour gain upon unfolding while storing the same fraction of the overall folding free energy when folded. Given that there are  $N$  such subunits involved in the measured mean deflection signal  $\langle x \rangle$ , the Hamiltonian  $H(x, i)$  of each of the  $N+1$  possibly populated states  $i \in \{0, 1, \dots, N\}$  reads:

$$H(x, i) = \left(1 - \frac{i}{N}\right) \cdot \Delta G_0 + G_{\text{bead}}(x) + G_D(F(x)) + \frac{i}{N} \cdot G_P(F(x)). \quad (18)$$

All four energetic contributions correspond to the ones presented in [Equation 9](#)<sup>4</sup>. The last term reflects a fraction of the maximal stretching energy  $G_P$  which is reached when all subunits are unfolded in state  $i = N$ . The calculation of the last two summands implies the integration along the entire stretched tether which is explicitly given by  $\int_0^{d-2R-x} F_{\text{tether}}(x') dx'$  where  $F_{\text{tether}}$  is the inverse of [Equation 4](#) with  $P_{\text{unf}} = 1$ . After summing up all states’ energy contributions by setting  $e^{-\beta \cdot H(x)} = \sum_{i=0}^N e^{-\beta \cdot H(x, i)}$  in [Equation 17](#) a general expression for a force-distance relation under equilibrium conditions described by [Equation 16](#) can be derived:

$$F(d, [x, N]) = \frac{k_{\text{eff}}}{\int_{-\infty}^{+\infty} dx \sum_{i=0}^N e^{-\beta \cdot H(x, i)}} \cdot \int_{-\infty}^{+\infty} dx x \cdot \sum_{i=0}^N e^{-\beta \cdot H(x, i)}. \quad (19)$$

[Equation 19](#) allows the direct fitting of stretch-and-relax cycles. Only the overall folding free energy  $\Delta G_0$  and the total contour length gain  $\Delta L_{\text{tot}}$  included in  $G_P$  are used as free fit parameters. All other parameters including the number of independent subunits  $N$  are kept constant for each individual fit. Typically,  $N$  is iteratively increased until a satisfying goodness of fit judged from the residuals is reached.

A characteristic feature of equilibrium unfolding/folding transitions is to not show any hysteresis within experimental resolution. This allows the  $\Delta G_0$  to be directly read off from the area enclosed by the stretch-and-relax cycle and the mechanical fit to the fully unfolded construct, see [\[81\]](#).

An alternative way to fit equilibrium transitions by using the mid-force  $F_{1/2}$  and total contour length gain  $\Delta L_{\text{tot}}$  as free fit parameters was recently presented and used to characterize the intrinsically-disordered protein  $\alpha$ -synuclein and the prion protein PrP [\[209, 249\]](#).

### 3.4 EXTRACTING FORCE-DEPENDENT TRANSITION RATES

Force-dependent transition rates can be retrieved from constant velocity as well as constant distance measurements in different ways, three of which are presented here.

<sup>4</sup> Here, the  $\Delta$  in front of  $G_0$  is added for conformity with [\[261\]](#).

### 3.4.1 Rates from force distribution histograms

Force distribution histograms  $p(F)$  directly relate to their corresponding rates  $k(F)$  according to [19, 65]:

$$k(F) = \frac{p(F) \cdot \dot{F}}{1 - \int p(F) dF}. \quad (20)$$

A formula to convert an unfolding force histogram directly into a respective rate versus force plot is, hence, given by [67]:

$$k_{\text{unf}}(F_l) = \frac{h_l \cdot \dot{F}(F_l)}{\left(\frac{h_l}{2} + \sum_{i=l+1}^N h_i\right) \cdot \Delta F}. \quad (21)$$

The unfolding rate at force  $F_l = F_0 + (l - \frac{1}{2}) \cdot \Delta F$ , where  $l = 1, 2, \dots$ , is calculated from a histogram with  $N$  bins of width  $\Delta F$  starting at  $F_0$  and ending at  $F_N = F_0 + N \cdot \Delta F$ . The parameter  $h_l$  is the normalized height of the  $l^{\text{th}}$  bin which is calculated from the number of counts  $C_l$  in this bin according to  $h_l = C_l / (N_{\text{tot}} \cdot \Delta F)$  with  $N_{\text{tot}}$  being the total number of counts.  $\dot{F}$  denotes the force-dependent loading rate which is further discussed in Section 3.6.1. For the derivation of refolding rates the effective change of direction of integration over  $p(F)$  needs to be considered, see Section 3.6.

### 3.4.2 Constructed equilibrium: Oberbarnscheidt's method

A constant velocity experiment can be divided into a sequence of short segments with nearly constant acting force. Due to the assumed Markovian nature of a protein's conformational transitions, segments of same acting force from multiple stretch-and-relax cycles can be stitched together into a virtual quasi-equilibrium measurement obeying the same Poisson statistics that each individual segment obeys. The transition rate  $k(F)$  at force  $F$  is then calculated from the ratio between the number of observed transitions  $N(F)$  and the entire time spent at this force given by the number  $M(F)$  of jointed segments of duration  $\Delta t$  [161]:

$$k(F) = \frac{N(F)}{M(F) \cdot \Delta t}. \quad (22)$$

Typically, the considered force range of the stretch-and-relax cycles is binned into small sections and the time spent in each bin is calculated according to the modelled force-distance relation and the employed velocity. By this means, the rate calculation becomes independent from the loading rate  $\dot{F}$  and therefore is also referred to as a model free rate determination.

As opposed to typical constant distance measurements, here, the force range of observable transitions can be expanded by increasing the velocity. This makes the described method a valuable supplement for force-dependent rate constant determination. However, the most effective way to increase statistics beyond equilibrium are jump experiments [189].

Errors for obtained rates are estimated through division of the respective rate by the square root of the number of events  $N$  from which the rate is calculated. A more sophisticated error estimate based on the likelihood function of the rate's calculation can be found in [19].

In congruence with its author Leoni Oberbarnscheidt the presented method is herein after referred to as OBS method [161].

### 3.4.3 Rates from lifetime distributions: Hidden Markov model analysis of equilibrium trajectories

The purpose of using a Hidden Markov Model (HMM) analysis is to unravel a stochastically 'well-behaved' interconnected network of states 'hidden' behind experimental data typically showing a poor signal-to-noise ratio. The model describing the network of states is assumed to comprise a memory-less continuous-time Markov process. This in turn allows the use of probabilistic arguments to infer a distinct sequence of states  $\mathbf{S} = \{s_0, s_1, s_2, \dots, s_t, \dots, s_T\}$  from a sequence of observables  $\mathbf{O} = \{o_0, o_1, o_2, \dots, o_t, \dots, o_T\}$  which represent recorded snapshots of the hidden network. A HMM is characterized by a set of emission probabilities  $E_i(o_t)$  reflecting the probability to make an observation  $o_t$  while the system is in state  $i$ . Hence, these probabilities relate measured data points to hidden states. The second important ingredient to parametrize a HMM is the transition matrix  $T_{ij}$  which specifies the transition probability from state  $i$  to state  $j$  at each recorded time step. Being a  $N \times N$  matrix,  $T_{ij}$  specifies the interconnectivity of all  $N$  states accessible to the hidden network. Together, the emission and transition probabilities constitute the model parameter set  $\mathcal{M} = \{E_i(o_t), T_{ij}\}$ .

While one of their early applications was to infer words from recorded speech within speech recognition software [175, 176], over the past years HMMs have become a versatile tool to analyse single-molecule data [18, 145, 216]. A detailed description of the HMM analysis applied within this thesis to infer protein conformational states from bead deflection signals measured with an optical trap setup can be found here [212, 214]. An example of a carried out state assignment is given in Figure 11. In the following, the main steps of a HMM analysis are presented in brief.

#### *HMM analysis steps in a nutshell*

As a preparative step, the input data, i.e., the raw difference signal of the two bead deflections, is further binned to ease numerical calculations. Now, at first, the HMM needs to be initialized. As a good first guess, the emission probabilities which reflect the deflection distribution or Point Spread Function (PSF) of each supposed state are represented by Gaussians. Adequate parameters for the positions and widths of the Gaussians are either guessed by eye or retrieved from a multiple Gaussian fit to a histogram of the deflection signal. Furthermore, the transition probability matrix  $T_{ij}$  is set up and normalized such that the sum of all transition probabilities out of a state equals 1. After initialization, an iterative state assignment process is started

by estimating state probabilities for each data point. Therefore, a forward-backward algorithm calculates, once forward and once backward in time<sup>5</sup>, the probability of producing the sequence of recorded data points preceding the currently analysed data point at which the system is assumed to be in state  $i$ . From the product of the forward and backward probabilities a corresponding posterior probability can be formulated for each data point to be in either one of all possible states. The most probable state being assigned to an analysed data point is revealed by the maximal posterior probability. To this end all probability calculations are based on a fixed parameter set  $\mathcal{M}$ . However, after a state assignment through a run of the forward-backward algorithm, the Baum-Welch algorithm can be used to optimize the parameter set, in particular the emission probabilities. In order to do so, a new parameter set  $\mathcal{M}^*$  is derived directly from the resulted state assignment and its maximal forward and backward probabilities. It can be shown that by this means the overall likelihood to produce the observed trajectory  $\mathbf{O} = \{o_0, o_1, o_2, \dots, o_t, \dots, o_T\}$  is maximized given  $\mathcal{M}^*$  [13, 214]. Hence, together, the forward-backward algorithm and the Baum-Welch algorithm constitute a maximum likelihood method. The iterative state assignment and parameter optimization proceeds until the assignment converges.

One way to supervise the performance of the HMM is by comparing the self-optimized emission probabilities with expected PSFs, also see Section 3.7.1. Furthermore, the quality of the assignment can be judged by the goodness of single-exponential fits to extracted lifetime distributions. Sometimes, the goodness of single-exponential fits can be enhanced by manual fine-tuning of individual entries of the transitions matrix  $T_{ij}$ .

Note that the performance of the Viterbi algorithm which is commonly used in HMM analysis was found to be very sensitive to a good initial guess for  $T_{ij}$  while not being indispensable for a good analysis of our trapping data [214, 229].

#### *Accounting for missed events upon transition rate calculation*

The above HMM based state assignment automatically yields a number of dwell or lifetimes  $\tau(F_i)$  spent in state  $i$  at force  $F_i$ . The corresponding normalized integrated lifetime histograms are fit by a single exponential function using the following formula [214]:

$$p(t) = \frac{e^{-k_{\text{off}} \cdot t} - e^{-k_{\text{off}} \cdot \tau_{\text{min}}}}{e^{-k_{\text{off}} \cdot \tau_{\text{max}}} - e^{-k_{\text{off}} \cdot \tau_{\text{min}}}}. \quad (23)$$

Here,  $k_{\text{off}}$  is a fit parameter representing the inverse lifetime  $\tau$  and, hence, the off-rate out of a state. The fixed values for  $\tau_{\text{min}}$  and  $\tau_{\text{max}}$  are chosen to account for experimental limitations. While  $\tau_{\text{max}}$  is set equal to the full duration of the analysed trace to respect the finite measurement time, by  $\tau_{\text{min}}$  the

<sup>5</sup> The forward-backward algorithm exploits the time-invariance of a Markov process to improve calculation accuracy while concurrently allowing to calculate probabilities for data points at time 'zero' lacking a preceding sequence 'forward' in time.

possibility of missed events due to a limited resolution is considered. The latter depends on the sampling frequency which defines the absolute minimum, the response time of the entire measurement system, and the signal-to-noise ratio, whereof the last two improve with rising force. Finally, the transition rate from state  $i$  to state  $j$  is given by [214]:

$$k^{ij}(F_{(i)}) = \frac{k_{\text{off}}^i}{1 + \sum_{k \neq i} \frac{N^{ik}}{N^{ij}}}. \quad (24)$$

The sum considers the that there may be more but one transition possible out of state  $i$ . The value  $N^{ij}$  reflects the number of detected transitions from state  $i$  to state  $j$ .

Error bars given for calculated transition rates represent  $1\sigma$  estimators [216], more refined error estimates can be calculated according to [19, 43].

#### 3.4.4 Pushing limits: autocorrelation analysis

If the equilibrium fluctuations of a protein are so fast and small that a state identification within the measured difference signal is not even possible with HMM analysis, there still remains another way to extract kinetic information: autocorrelation analysis. The force-dependent decay of the autocorrelation (AC) function of a recorded constant distance trace can be fitted by a double exponential function yielding two rate constants provided by the decay times  $\tau_D^{\text{AC}}$  and  $\tau_P^{\text{AC}}$  characteristic for the DNA-handles and the protein [261]. The inverse of  $\tau_P^{\text{AC}}$  reflects the combined or effective rate  $\lambda$  of protein folding and unfolding – here denoted as  $k^{ij}$  and  $k^{ji}$  – at the measured average force  $F$ :

$$(\tau_P^{\text{AC}})^{-1} = \lambda(F) = k^{ij}(F) + k^{ji}(F). \quad (25)$$

### 3.5 KINETIC MODELS

In order to interpret the information on a protein's folding mechanism provided by experimentally determined force-dependent transition rates, kinetic models characterizing this force-dependence are indispensable. Besides yielding the crucial transition state position at  $\Delta x$ , kinetic models are especially useful for extrapolating rates from within the accessible force range towards the zero-force rate constants  $k_0^{ij}$ . In addition to giving a more natural, i. e., force-free picture of the folding mechanism, zero-force rate constants serve as an important consistency check when being compared to values from experiments which use other denaturants than force to trigger protein unfolding and folding. However one should keep in mind that a mechanically triggered and spatially constrained pathway as observed in force spectroscopy must not follow the same kinetics as it would in the absence of force. The following section introduces different kinetic models in view of their meaning and applicability.

### 3.5.1 *Blind spot at zero force: Arrhenius equation*

In force-spectroscopy transitions of a protein between different conformational states are typically described as transitions between local minima of an energy landscape where the folding free energy is plotted against a reaction coordinate  $x$ , see [Figure 13](#). In our case this reaction coordinate is the measured change in extension of the molecule under study along the direction of the acting force. Albeit, this inferred one-dimensional energy landscape has to be considered as a simplified projection of a complex and multidimensional protein folding process.

While the probabilities of finding a system in either state  $i$  or  $j$  are fully defined by their folding free energy difference  $\Delta G_0^{ij}$  (see, e. g., [Equation 15](#)), the actual transition rate between two states crucially depends on the height of the transition barrier separating them. This barrier, which is also referred to as the activation free energy barrier, comprises a local energy maximum located at the position of the transition state (TS) and has a height of  $\Delta G_0^{iTS}$  or  $\Delta G_0^{jTS}$  with respect to the transition's initial state. After reaching the transition state further progress towards the final state is energetically favoured and, hence, not rate limiting anymore. With no force involved as indicated by the index '0', the transition rate  $k_0^{ij}$  from state  $i$  to state  $j$  decreases exponentially with  $\Delta G_0^{iTS}$  and is given by [4]:

$$k_0^{ij} = k_A \cdot \exp\left(-\frac{\Delta G_0^{iTS}}{k_B T}\right). \quad (26)$$

The pre-exponential factor  $k_A$ , as further elaborated in [Section 3.5.4](#), is usually interpreted as a measure of the system's fluctuation frequencies which for their part reflect a kind of energy landscape roughness [40, 106]. Therefore, intuitively the transition rate can be interpreted as an attempt frequency inherent to the initial state times the probability of reaching the transition state. The range of  $k_A$  has been reported to cover  $10^3 - 10^9 \text{ s}^{-1}$  [23, 201], but even  $10^{11} - 10^{13} \text{ s}^{-1}$  is used [148]. By this means [Equation 26](#) can be applied to estimate force-free energy barrier heights from zero-force extrapolated rate constants only under the assumption of a reasonable value for  $k_A$  [59, 83].

Even though [Equation 26](#) is commonly known as Arrhenius equation, Svante Arrhenius himself quoted it from Van't Hoff who already stated the temperature-dependence of the rate constant  $k$  of a chemical reaction to be given by [228]:  $d/dT \cdot \log(k) = A/T^2 + B$ .

### 3.5.2 *The power of force: Zhurkov-Bell model*

The Zhurkov-Bell model, henceforth abbreviated as ZB-model, was first developed to describe the strength of solids prior being applied to the fracture kinetics of polymers [255, 256]. Since its follow-up application for the characterization of cell to cell adhesion the ZB-model became more commonly known as Bell model. In this model, the force-dependence of the transition rate is calculated from the effect of an externally applied force  $F$  on the rate

limiting barrier height at the position  $\Delta x$  of the transition state according to [15]:

$$k^{ij}(F) = k_A \cdot \exp\left(-\frac{\Delta G_0^{iTS} - F \cdot \Delta x}{k_B T}\right) = k_0^{ij} \cdot \exp\left(\frac{F \cdot \Delta x}{k_B T}\right). \quad (27)$$

While an applied force effectively tilts a protein's energy landscape as shown in [Figure 13](#), the expression  $F \cdot \Delta x$  corresponds to the work that is performed on or rather by a protein to either extend or compress it towards the transition state. Depending on the corresponding sign<sup>6</sup> of  $\Delta x$ , the force-free transition barrier  $\Delta G_0^{iTS}$  is either reduced or increased by the external work which, in turn, affects the transition rate. [Equation 27](#) is used to fit rate versus force plots as shown, e. g., in [Figure 43C](#). The fit parameters  $k_0^{ij}$  and  $\Delta x$  reflect the intercept with the rate axis at zero force and the exponential slope of the fit.

Even though the simplicity of the ZB-model is very appealing it comes along with a number of limitations when being applied to data from the dual-optical trap assay. The overall work performed to reach the transition state is assumed to increase linearly with force. This neglects all nonlinear contributions from the measurement apparatus including [DNA](#)-handles as well as the unfolded polypeptide. Furthermore, the fit parameters  $k_0^{ij}$  and  $\Delta x$  are considered to be constant at all forces. The pre-exponential factor  $k_A$  included within  $k_0^{ij}$  may be force-dependent though. And while the extensibility of a folded protein is very small due to the strong but short-ranged interactions keeping it in the folded state, the extensibility of the polypeptide chain of an unfolded protein is highly force-dependent as implied by the [WLC](#) model. By this means especially for refolding transitions a force-dependence of  $\Delta x$  needs to be accounted for.

Note that the value  $\Delta x$  reflects a length change due to the stretching of the entire measurement system. To not mix up the  $\Delta x$  values obtained from different kinetic models, the transition state positions extracted with the ZB-model for a transition ending in state  $i$  will be indexed as  $\Delta x_i^{ZB}$ .

### 3.5.3 Force-induced barrier: Schlierf-Berkemeier model

In contrast to the ZB-model, the Schlierf-Berkemeier (SB) model accounts for all energetic contributions of the measurement system which add up to a more refined force-dependent decrease or increase of the transition barrier. After having been developed to model protein refolding observed at the [AFM](#), the SB-model was readily adapted for optical tweezers experiments; its general formula reads [83, 196]:

$$k^{ij}(F) = k_0^{ij} \cdot \exp\left(-\frac{\Delta G_{\text{pot}}^{iTS}(F, F_{TS}, [\Delta x])}{k_B T}\right). \quad (28)$$

The force-induced barrier  $\Delta G_{\text{pot}}^{iTS}$  represents the part of the equilibrium free energy difference between state  $i$  and the transition state TS which comprises

<sup>6</sup> In this work, all  $\Delta x$  values are given as positive values the true sign of which can be deduced from the slope of the corresponding logarithmic rate versus force plots.

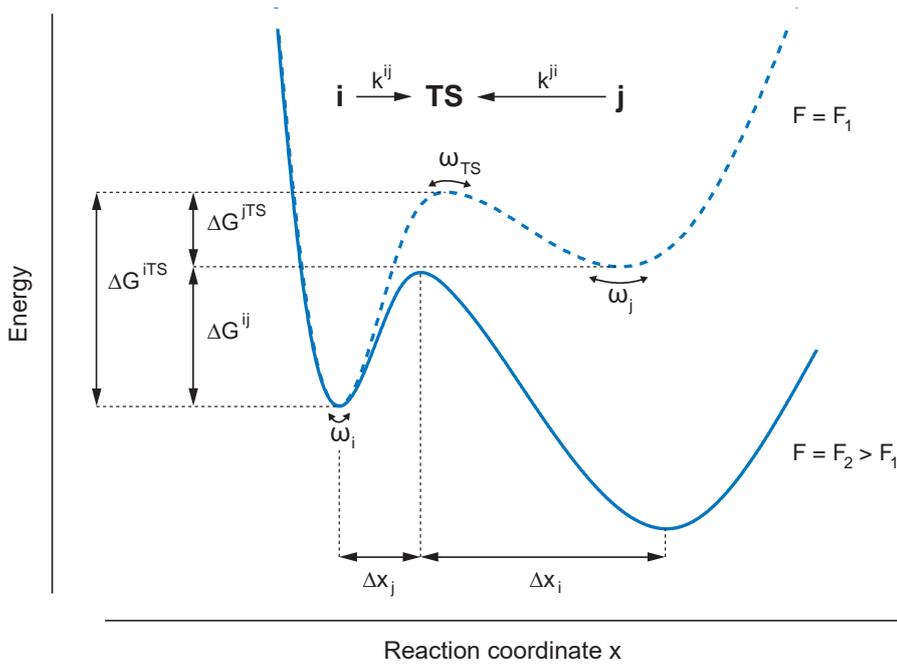


Figure 13: Key parameters characterizing a folding energy landscape tilted by an external force. Transitions between the two states  $i$  and  $j$  are controlled by a barrier having its maximum at the position of the transition state TS. Marked parameters are explained in the text.

the individual contributions of the beads, DNA-handles, and unfolded protein in accord with Equation 13. The only two fit parameters in Equation 28 are the zero-force rate constant  $k_0^{ij}$  given by Equation 26 and the transition state position  $\Delta x$ . The force-dependence of the latter is calculated from a WLC model with contour length  $\Delta L_p^{iTS}$  to describe the transition state position in contour space. The force  $F_{TS}$  acting on the protein at the transition state is obtained from the measured force  $F$  and the currently chosen  $\Delta L_p^{iTS}$ . All other necessary parameters are known from mechanical fits and the calibration of the traps.

Due to the included force-dependence of the transition state position  $\Delta x$ , SB-model fits show a curved shape with a slope equal to zero at zero force, see, e. g., Figure 41A. This is a direct consequence of the force-dependent extension of the transition state position modeled by a WLC: at zero force the average end-to-end distance of an unstructured polypeptide equals zero. The curvature resulting from the force-dependence of  $\Delta x$  also affects the zero-force extrapolated rate constants. The good agreement between zero-force free energies  $\Delta G_0^{ij}$  obtained from probabilities measured under force (see Equation 15) and from extrapolated zero-force rate constants (see Equation 32) proves the curvatures to be meaningful, see Section 5.2.3. Another proof of principle for the SB-model is given if  $\Delta L_p^{ij} = \Delta L_p^{iTS} + \Delta L_p^{jTS}$ , which means that the sum of the absolute values of the transition state position given in contour length sum up to the total contour length gain observed for the corresponding transition in force-extension or constant distance traces.

All values declared as  $\Delta x_i$  without an upper index stem from SB-model fits and are given in contour length of the protein.

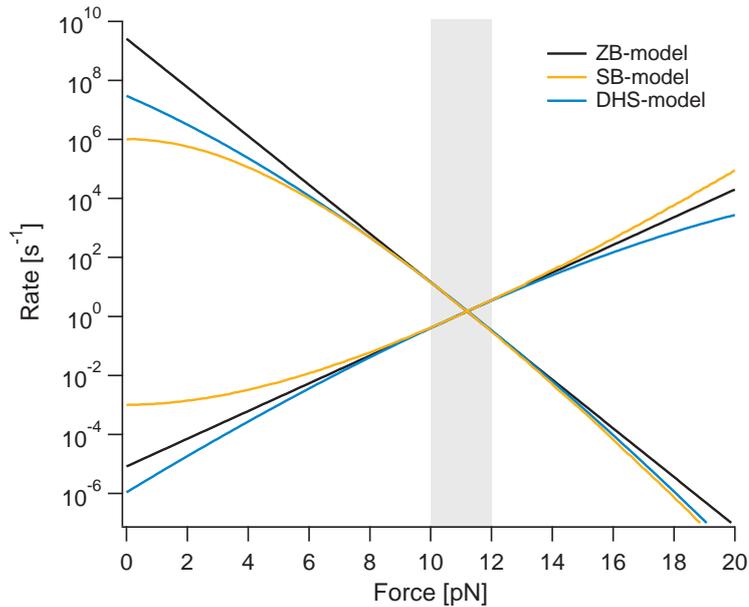


Figure 14: Three kinetic models for fitting force-dependent rates. The SB-model fits were fixed at  $k_0^f = 10^6 \text{ s}^{-1}$ ,  $k_0^u = 10^{-3} \text{ s}^{-1}$ ,  $\Delta x_u = 9 \text{ nm}$ , and  $\Delta x_f = 15 \text{ nm}$ . The ZB- and DHS-models were fitted to match the SB-model plots within the gray shaded area. Notably,  $\Delta G_0^{ij} \sim 21 \text{ k}_B\text{T}$  implied for the SB-model by  $\ln(k_0^u/k_0^f)$ , see Equation 32, can also be reproduced from the zero-force barrier height difference supplied by the DHS-model.

#### 3.5.4 Getting in shape: using Kramers theory

Kramers developed his rate-theory for particles moving in a potential as described by a Langevin equation, see Equation 59. This equation considers a detailed energy landscape shape  $G(x)$  as the potential (see Figure 13 for an example), a stochastically acting thermal force representing Brownian motion, and damping characterized by the friction coefficient  $\gamma$  which influences the speed  $\dot{x} = v$  of particle diffusion. To calculate the rate  $k_0^{ij}$  over a barrier separating two states  $i$  and  $j$  in a double-well potential as sketched in Figure 13, Kramers derived the stationary probability current  $j^{\text{TS}}$  at the transition barrier from a constructed probability density function  $\rho(x, v)$  that solves the corresponding stationary Fokker-Planck equation around the barrier top while meeting two further requirements – basically a Boltzmann distributed particle population within the energy well around state  $i$  and a particle sink beyond the transition state. The ratio of the resulting stationary current and the population  $n_i$  of particles in state  $i$  directly yield the transition rate according to  $k_0^{ij} = j^{\text{TS}}/n_i$ , where  $j^{\text{TS}} = \int_{-\infty}^{+\infty} dv v \rho(x_{\text{TS}}, v)$  and  $n_i = \int_{i\text{-well}} dx dv \rho(x, v)$  are given by  $\rho(x, v)$ . The final result describes the thermally assisted escape rate for a particle over a barrier for moderate up to high friction. The proba-

bly most widely appreciated and simple appearing result is obtained in the Smoluchowski limit, i. e., the overdamped case where  $\gamma \gg \omega_{\text{TS}}$  [95, 116]:

$$k_0^{ij} = \frac{\omega_i \cdot \omega_{\text{TS}}}{2\pi \cdot \gamma} \cdot \exp\left(-\frac{\Delta G_0^{\text{iTS}}}{k_{\text{B}}T}\right). \quad (29)$$

Here,  $\omega_i$  and  $\omega_{\text{TS}}$  represent a direct link to the energy landscape and describe the oscillation frequencies around the minimum of state  $i$  and around the barrier top at the transition state. These frequencies are further defined as the absolute values of the square root of the respective local curvatures of the (harmonic) potential  $G(x)$ . Next, a more general expression for a particle's escape rate is given in terms of the mean first-passage time [95, 99]:

$$\tau_{\text{mfpt}} = \left(k_0^{ij}\right)^{-1} = \frac{\gamma}{k_{\text{B}}T} \cdot \int_{x_i}^{x_{\text{TS}}} e^{-\beta G(x')} dx' \cdot \int_x^{x_{\text{TS}}} e^{\beta G(x')} dx'. \quad (30)$$

The integrals around the transition state position as well as over the entire potential well of state  $i$  stem from the above sketched formalism to calculate the escape rate.

Within the framework of Kramers formalism, Dudko, Hummer, and Szabo derived a kinetic model, in the following referred to as DHS-model, to describe the force-dependence of transition rates. In analogy to the emergence of the ZB-model out of the Arrhenius equation, see Equation 27, here, the effect of an external force on the transition rate is introduced by setting  $G(x) = G_0(x) - F \cdot x$  in Equation 30 which tilts the zero-force energy landscape  $G_0(x)$ . The integrals inferred by Equation 30 for the expression  $k^{ij(F)}/k_0^{ij}$  can be evaluated analytically for different shapes of  $G_0(x)$  that are distinguished by the scaling factor  $\nu$  within the unified form of the DHS-model [65]:

$$k^{ij(F)} = k_0^{ij} \cdot \left(1 - \frac{F \cdot \Delta x}{\Delta G_0^{\text{iTS}}} \cdot \nu\right)^{\frac{1}{\nu}-1} \cdot \exp\left\{\frac{\Delta G_0^{\text{iTS}}}{k_{\text{B}}T} \left[1 - \left(1 - \frac{F \cdot \Delta x}{\Delta G_0^{\text{iTS}}} \cdot \nu\right)^{\frac{1}{\nu}}\right]\right\}. \quad (31)$$

By setting  $\nu = 2/3$  or  $1/2$  the DHS-model assumes either a linear-cubic potential or a potential with a cusp-like barrier as zero-force energy landscape. For  $\nu = 1$ , the DHS-model converts into the ZB-model. Above the critical force, where  $F_c \cdot \Delta x = \Delta G_0^{\text{iTS}}$ , the DHS-model breaks down as a consequence of the high-barrier limit at which Kramers theory holds. This limit requires<sup>7</sup>  $\Delta G_0^{\text{iTS}} \gg 1 k_{\text{B}}T$  to assure that only the transition barrier is rate limiting and no other local roughness of the order of thermal noise. Without a doubt, a unique feature of the DHS-model is to provide the zero-force barrier height  $\Delta G_0^{\text{iTS}}$  as a fit parameter which is owed to the consideration of the full energy landscape shape  $G(x)$  in Kramers formalism. Another noteworthy side effect is that the curvatures of the force-dependent rate plots show expected Hammond behaviour for crossing the transition barrier in both directions, i. e., the more the reaction product prevails the closer 'comes' the transition state towards

<sup>7</sup> For experimental data analyzed with Kramers rate-theory  $\Delta G_0^{\text{iTS}} > 2 k_{\text{B}}T$  and  $\Delta G_0^{\text{iTS}} \geq 3 k_{\text{B}}T$  have been postulated [40, 44].

the reactant in terms of a reaction coordinate [94]. An illustration of this effect can be found on the left of [Figure 26](#) in [Section 4.2.4](#). However, as for the ZB-model, the energetic contribution of the external force enters the DHS-model only linearly. To overcome this drawback it may be an intriguing idea to replace the term  $F \cdot \Delta x$  in [Equation 31](#) by  $\Delta G_{\text{pot}}^{\text{ITS}}(F, F_{\text{TS}}, [\Delta x])$  from the SB-model. Transition state positions extracted with the DHS-model for transitions ending in state  $i$  are declared as  $\Delta x_i^{\text{DHS}}$ .

For the first time, [Equation 29](#) breaks down the pre-exponential factor  $k_A$  from [Equation 26](#) into distinct contributions. One can see that  $k_A$  is not necessarily the same for  $k_0^{ij}$  and  $k_0^{ji}$  as in each case only the reactant's side up to the transition state is considered relevant for particle escape. As their experimental determination is quite challenging, involved oscillation frequencies are often assumed to be about the same for data interpretation [44, 86, 120, 201, 247] unless they are directly assessed by, e. g., deconvolution of the underlying energy landscape [154]. Attempts to resolve the diffusion constant<sup>8</sup>  $D$  characteristic for protein folding have led to different observations including an expected temperature dependence [106, 159], coordinate-dependent diffusion [21, 40, 132], and a potentially implied effective overall energy landscape roughness<sup>9</sup> ranging from 2 to 5  $k_B T$  which may even contain local traps of up to  $\sim 20 k_B T$  [148, 154, 249].

#### Detailed balance

If two states  $i$  and  $j$  are in equilibrium with a folding free energy difference  $\Delta G^{ij}$  their equilibrium rate constant  $K_{\text{eq}}$  is given by:

$$K_{\text{eq}} = \frac{k^{ij}}{k^{ji}} = \exp\left(-\frac{\Delta G^{ij}}{k_B T}\right). \quad (32)$$

As a consequence of the principle of detailed balance, [Equation 32](#) implies that at a given folding free energy difference the rate of the forward direction of a reaction predetermines the rate of the backward direction and vice versa. This result directly follows from the equilibrium condition  $P_i \cdot k_{ij} = P_j \cdot k_{ji}$  in connection with [Equation 14](#). Alternatively it can be derived from, e. g., [Equation 26](#) and the relation  $\Delta G^{i\text{TS}} - \Delta G^{j\text{TS}} = \Delta G^{ij}$  as illustrated in [Figure 13](#). Note that [Equation 32](#) assumes the same pre-exponential factor  $k_A$  for both directions of the equilibrated reaction, in particular  $\omega_i = \omega_j$  with respect to [Equation 29](#), which neglects potential differences in energy landscape shape which may even extend towards differences in local friction.

### 3.6 FORCE DISTRIBUTION HISTOGRAMS

Force distribution histograms are a classical way to accumulate and present unfolding and refolding forces collected from multiple stretch-and-relax cy-

<sup>8</sup> The diffusion constant  $D$  is directly linked to the friction coefficient through the Einstein relation [68]:  $D \cdot \gamma = k_B T$ .

<sup>9</sup> The principle of an effective energy landscape roughness in terms of an energy  $\epsilon$  is introduced in [Section 4.2.2](#).

cles performed on a protein in constant velocity mode. By their nature, constant velocity experiments impose a time-dependent force  $F(t)$  on the protein under study. A force distribution histogram is, hence, a product of a transition rate which increases in time and a likelihood of survival in the original state that decreases in time [71, 72]:

$$p(F) = \frac{k(F)}{|\dot{F}|} \cdot \exp\left(-\int_{F_{\text{start}}}^F \frac{k(F')}{F'} dF'\right). \quad (33)$$

Here, the absolute value compensates for a negative loading rate  $\dot{F}$  experienced during relaxation. The loading rate itself corresponds to an effective harmonic spring  $\kappa_{\text{eff}}$  times the velocity  $v$ , where  $\kappa_{\text{eff}}$  incorporates all compliances of the entire system as elaborated in Section 3.6.1 [10]:

$$\dot{F} = \kappa_{\text{eff}} \cdot v. \quad (34)$$

Using Equation 33 and Equation 34 in conjunction with the ZB-model for the transition rate's force-dependence given in Equation 27 we arrive at a formula for the probability distribution of unfolding forces by integrating from  $F_{\text{start}} = F(t=0) = 0$  up to  $F$ :

$$p_{\text{unf}}^{\text{ZB}}(F) = \frac{k_0^u}{\kappa_{\text{eff}} \cdot v} \cdot \exp\left\{\frac{F \cdot \Delta x}{k_B T} - \frac{k_0^u \cdot k_B T}{\kappa_{\text{eff}} \cdot v \cdot \Delta x} \cdot \left[\exp\left(\frac{F \cdot \Delta x}{k_B T}\right) - 1\right]\right\}. \quad (35)$$

The two free fit parameters  $k_0^u$  and  $\Delta x$  are directly passed over from the ZB-model<sup>10</sup>. For refolding during relaxation  $\Delta x$  and  $v$  have negative values. Furthermore, in a relaxation experiment integration starts at a higher force as opposed to where it ends. Choosing  $F_{\text{start}}$  high enough for refolding to be impossible right at the start of a relaxation cycle effectively cancels out the very last contribution in Equation 35 as given by the extreme case of  $e^{-\infty}$  equals zero. The typical shapes of unfolding and refolding force distributions are illustrated in Figure 15A.

In view of the generality of Equation 33 any of the kinetic models introduced in Section 3.5 can be used to derive fitting functions for experimentally determined force distributions [65, 169, 196].

### 3.6.1 Elastic compliance: loading rate determination

To calculate the loading rate given by Equation 34 it is crucial to be able to evaluate the total system's compliance represented by  $\kappa_{\text{eff}}$  at all forces. In the dumbbell configuration  $\kappa_{\text{eff}}$  can be split up into three contributions:

$$(\kappa_{\text{eff}})^{-1} = \frac{1}{\kappa_{\text{eff}}} + \frac{1}{k_D} + \frac{1}{k_P}. \quad (36)$$

The three contributions comprise the effective spring constant  $\kappa_{\text{eff}}$  of the two optical traps as well as effective spring constants of the DNA-handles  $k_D$  and

<sup>10</sup> Note that the single index  $u$  is a simplified representation of the two indices  $ij$  in  $k_0^{ij}$  from Equation 27. Here,  $u$  represents a simple two-state transition ending in the unfolded state.

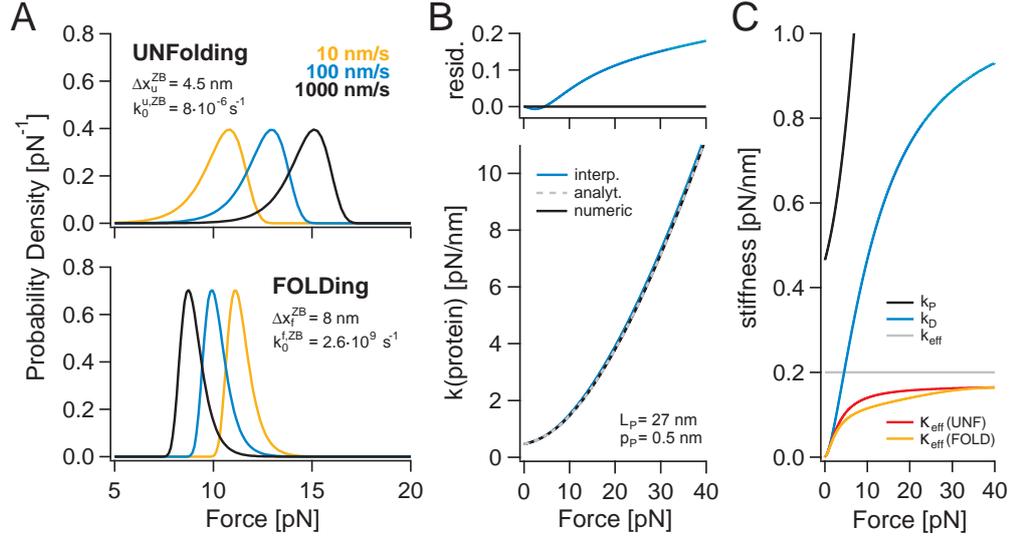


Figure 15: Force distribution histograms and elastic compliance contributions. (A) Unfolding and folding force probability distributions at different constant velocities according to the ZB-model and Equation 33. (B) Force-dependence of  $k_P$  according to Equation 37, Equation 38, and numerical differentiation of a WLC. (C) All individual contributions to the system's compliance in a typical optical trapping experiment. The eWLC parameters are  $p_D = 25$  nm,  $L_D = 370$  nm, and  $K = 400$  pN. The effective trap stiffness was set to  $k_{\text{eff}} = 0.2$  pN/nm.

unfolded protein  $k_P$ . While  $k_{\text{eff}}$  is constant within the chosen force range, both  $k_D$  and  $k_P$  are force-dependent. Since  $k_P$  represents unfolded protein, an analytic expression for its force-dependent elastic contribution can be directly obtained from the local slope of the corresponding WLC model:

$$k_P = \frac{\partial F_{\text{WLC}}}{\partial x_P} = \frac{k_B T}{p_P} \left[ \frac{1}{L_P} + \frac{1}{2L_P \cdot \left(1 - \frac{x_D}{L_D}\right)^3} \right]. \quad (37)$$

To be able to express  $k_P$  directly in terms of the acting force  $F$ , an alternative interpolation formula has been proposed [67]:

$$k_P \approx \frac{3 + 5\beta \cdot F \cdot p_P + 8(\beta \cdot F \cdot p_P)^{5/2}}{2\beta \cdot L_P \cdot p_P \cdot (1 + \beta \cdot F \cdot p_P)}. \quad (38)$$

Figure 15B shows a comparison between the force-dependence of  $k_P$  for a protein with  $L_P = 27$  nm and  $p_P = 0.5$  nm inferred by Equation 37, the interpolation formula in Equation 38, and a direct numerical differentiation of the corresponding WLC model. From the residuals with respect to the analytic expression in Equation 37 one can see that the interpolation formula models  $k_P$  reasonably well but a direct calculation from the corresponding WLC model should be preferred. Values for  $k_D$  are henceforth also derived from their related eWLC models. Figure 15C shows the force-dependent effective stiffness  $k_{\text{eff}}$  experienced at the protein's DNA-handle attachment positions

within a stretched dumbbell configuration. From the protein's point of view the effective spring constant  $k_{\text{eff}}$  of the two traps is drastically reduced at forces below 10 pN, while at higher forces an almost constant reduction of about 20 % can be observed. Taken together this highlights the importance of a correct loading rate determination.

Within the limits of our optical trapping experiments the above made considerations concerning the loading rate determination are reasonable. However, in some cases a further refinement of the loading rate may be considered necessary [136, 226].

### 3.7 DECONVOLUTION OF PROBABILITY DISTRIBUTIONS

A very powerful method to fully reconstruct a protein's folding energy landscape is given by deconvolution. If the equilibrium probability distribution  $P_P(x)$  of a fluctuating protein is known, the underlying energy landscape  $G(x)$  is simply retrieved by a Boltzmann inversion:

$$G(x) = -k_B T \cdot \ln(P_P(x)) + \text{const.} \quad (39)$$

However, the measured probability distribution  $P_{\text{msmt}}(x)$  of the fluctuations of a protein recorded in a constant distance optical trapping experiment is only a blurred version of  $P_P(x)$  due to the coupling of thermal noise to the entire measurement system. Mathematically,  $P_{\text{msmt}}(x)$  can be interpreted as a convolution of the probability distribution  $P_P(x)$  with the probability distributions of the fluctuations of the trapped beads and DNA-handles represented by  $\Psi_x(x)$ , i. e., the point spread function (PSF):

$$P_{\text{msmt}}(x) = \Psi_x(x) \otimes P_P(x) = \int_{-\infty}^{+\infty} dx' \Psi_x(x - x') P_P(x'). \quad (40)$$

The index  $x$  within  $\Psi_x(x)$  indicates that the PSF itself depends on the measured extension  $x$  through the involved acting force, also see Section 3.7.1. To remove the additional contributions of the beads and DNA-handles included in  $P_{\text{msmt}}(x)$ , a constrained nonlinear iterative deconvolution can be performed to reconstitute  $P_P(x)$  [110]:

$$P_P^{n+1}(x) = P_P^n(x) + r(P_P^n(x)) \cdot [P_P^0(x) - \Psi_x(x) \otimes P_P^n(x)]. \quad (41)$$

Here, the relaxation function  $r(P_P^n(x)) = r_0 \cdot (1 - 2|P_P^n(x) - 1/2|)$  constrains the resulting probabilities between 0 and 1 to remain within physical boundaries. Reported values for the amplitude  $r_0$  which controls the speed of convergence are, e. g., 1 and 2. The iteration can be started by using  $P_P^0(x) = P_{\text{msmt}}(x)$ , alternatively an educated guess based on complementary data providing, e. g., the transition state position, is equally justifiable. To reach convergence a few hundred up to tens of thousands of iterations can be necessary [83, 243, 247].

An alternative approach turns the iterative deconvolution procedure for reconstructing  $P_P(x)$  out of  $P_{\text{msmt}}(x)$  into an optimization problem where, in principle, the convolved form of an estimated deconvolved energy landscape

$\hat{G}(x)$  is fitted to  $P_{\text{msmt}}(x)$  [98]. A numerical implementation of this approach foresees the minimization of the following function [177]:

$$\mathcal{O} = \langle |\ln(\hat{P}_P(x) \otimes \Psi_x(x)) - \ln(P_{\text{msmt}}(x))| \rangle_x + \lambda \cdot \sum_i [\beta \cdot \partial_x^2 \hat{G}(x)|_{\hat{x}_i}]^2. \quad (42)$$

As the estimated probability distribution of the protein is given by  $\hat{P}_P(x) = \exp(-\beta \cdot \hat{G}(x))$ , the only unknown in this equation is  $\hat{G}(x)$ .  $\hat{G}(x)$  itself is represented by a cubic spline which interconnects a number of equally spaced base points denoted by  $\hat{x}_i$ . While the first part of Equation 42 represents an average over all  $x$  values, the second part is a penalty function to ensure that solutions found for  $\hat{G}(x)$  are smooth. Typical values for  $\lambda$  are in the order of  $10^{-3} \text{ nm}^2$  and smaller. Minimization of Equation 42 is initiated using  $\hat{P}_P(x) = P_{\text{msmt}}(x)$  and performed by an algorithm which varies the interpolation points in order to find the best estimate for  $\hat{G}(x)$ . A quasi-Newton optimizer with Moré-Hebden steps has been shown to decently fulfill this task while an interesting alternative may also be provided by simulated annealing [56, 177]. An example showing both the convolved energy landscape inferred by  $P_{\text{msmt}}(x)$  as well as its deconvolved version obtained after optimization of Equation 42 is given in Figure 27C of Section 4.3.1.

After a change of variables  $\hat{G}_P(x \rightarrow L_P)$  which involves a contour length transformation, see Section 3.1.1, the deconvolved energy landscape  $\hat{G}_P(L_P, F_{\text{msmt}})$  observed at pretension  $F_{\text{msmt}}$  can be transformed to any other pretension using Equation 13, see, e. g., Figure 44A in Section 5.3.1. After transformation to a common pretension, multiple energy landscapes can be combined using the Weighted Histogram Analysis Method (WHAM) to enhance resolution [77, 98, 122]. Local stabilities are directly provided by the energy landscapes' derivatives  $\frac{\partial \hat{G}_P(L_P)}{\partial L_P}$ . Errors for deconvolved energy landscapes can be estimated by bootstrapping [177].

Even though deconvolution provides a very useful tool for energy landscape reconstruction there certainly are limitations. To obtain good and meaningful results the protein's true energy landscape must be reducible to the single dimension of our force-spectroscopic measurement. Only in that case the measured changes in protein extension can be considered as a good reaction coordinate given the spatial and temporal resolution are high enough.

There are other methods of energy landscape reconstruction which exist besides the deconvolution of equilibrium data. Recently, a proposed reconstruction from non-equilibrium constant velocity measurements was tested on experimental data and verified by direct comparison to deconvolution [91, 102]. Another approach suggests to apply deconvolution directly to non-equilibrium data [104]. A way of reconstructing the shapes of energy barriers by the analysis of transition paths is presented in Section 3.8.

### 3.7.1 Setting up the point spread function

The key ingredient of any deconvolution is the PSF which ideally enables us to remove all interfering noise. It has been shown that the distribution of the

detected position noise of beads in a stretched dumbbell configuration can be described by a skewed Gaussian function  $\Gamma(x, \mu, \sigma, \gamma, i)$ , where  $\mu$  is the mean,  $\sigma$  the standard deviation, and  $\gamma$  the skewness of the distribution of deflection values measured while the protein is in state  $i$  [216]. It has been further elaborated that the PSF needs to be adapted to the change of acting average force as bead deflection changes in constant distance experiments operated in passive mode [83]. To do so,  $\Psi_x(x)$  is derived for each deflection value by a linear interpolation between  $\mu$ ,  $\sigma$ , and  $\gamma$  of the point spread functions at the two extremal deflections marking the edges of  $P_{\text{msmt}}(x)$ :

$$\begin{aligned}\Psi_{x_{\text{max}}}(x) &= \Gamma(x, \mu, \sigma, \gamma, i) \\ \Psi_{x_{\text{min}}}(x) &= \Gamma(x, \mu, \sigma, \gamma, j).\end{aligned}\quad (43)$$

In case of transitions recorded from a simple two-state folder, the states  $i$  and  $j$  correspond to the folded (closer to  $x_{\text{max}}$ ) and unfolded state (closer to  $x_{\text{min}}$ ). By inserting the definition of a Boltzmann distribution, see Equation 17, into the definition of the moments of a distribution, it is possible to directly link the moments to the mechanical energy  $G_{\text{pot}}^i(F_i(x))$  stored in the stretched dumbbell with the protein in state  $i$ . The  $n$ -th moment of a distribution  $p(x)$  about the origin is then given by [261]:

$$\mu_n = \langle x^n \rangle = \int_{-\infty}^{+\infty} dx x^n \cdot p(x) = \frac{\int_{-\infty}^{+\infty} dx x^n \cdot \exp(-\beta \cdot G_{\text{pot}}^i)}{\int_{-\infty}^{+\infty} dx \exp(-\beta \cdot G_{\text{pot}}^i)}. \quad (44)$$

The equations for calculating  $G_{\text{pot}}^i(F_i(x))$  are implied by Equation 9 and Equation 13 for which all necessary parameters can be experimentally assessed. As  $\mu$ ,  $\sigma$ , and  $\gamma$  can be directly calculated from their corresponding moments, Equation 43 in combination with Equation 44 pave the way towards the calculation of  $\Psi_x(x)$  for all deflection values. In detail, with the mean being the first (raw) moment it is directly given by  $\mu = \mu_1$ , the standard deviation being the positive square root of the variance, the second central moment, reads  $\sigma = \sqrt{\mu_2 - \mu_1^2}$ , and the skewness equals  $\gamma = (\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3)/\sigma^3$ .

Note that the distribution moments calculated using Equation 44 are only valid at infinite bandwidth [261]. Calibration factors accounting for the finite experimental bandwidth can be derived by comparing the values inferred by Equation 44 to those actually measured for dimerized oligos [177].

Instead of a linear interpolation between two extremal point spread functions one might think of explicitly calculating  $\Psi_x(x)$  at all bead positions to account for the nonlinearity of  $G_{\text{pot}}^i(F_i(x))$ . However, besides being numerically more time-consuming, this idea would require further assumptions about how to transition between protein states. This could be feasible by considering additional information on individual state distributions and their overlap with others.

### 3.8 TRANSITION PATH ANALYSIS

Inasmuch as transition paths are immediate witnesses of a reaction, they also bear essential information on the underlying reaction mechanism. A frame-

work for the statistical analysis of a representative ensemble of transition paths is provided by transition path theory (TPT) [236, 237]. To date, this theory was mainly used to interpret rare transition events acquired from molecular simulations especially by using the transition path sampling method [28, 55]. Recent advances in single-molecule force spectroscopy, however, allow the recording and extraction of experimental fingerprints of transition paths, see, e.g., Figure 12, which now call for an appropriate analysis. After introducing the key ingredients of TPT, its relationship with experimental time traces is presented followed by emerging applications for experimental data.

### 3.8.1 Path ensembles and commitment probabilities

Let  $x(\mathcal{T}) \equiv \{x_0, x_{\Delta t}, x_{2 \cdot \Delta t}, \dots, x_{\mathcal{T}}\}$  be a random trajectory of duration  $\mathcal{T}$  through a system's phase space represented by an ordered sequence of states separated by a small time increment  $\Delta t$ . If considering a Markovian process the dynamical path probability of the entire path  $x(\mathcal{T})$  can be derived from the product of short-time probabilities associated with the evolution of state  $x_t$  into state  $x_{t+\Delta t}$  [55, 172]:

$$\mathcal{P}[x(\mathcal{T})] = \rho(x_0) \cdot \prod_{i=0}^{\mathcal{T}/\Delta t - 1} p(x_{i \cdot \Delta t} \rightarrow x_{(i+1) \cdot \Delta t}). \quad (45)$$

Here,  $\rho(x_0)$  denotes the probability distribution of the initial state  $x_0$  which, for instance, can be represented by the canonical (Boltzmann) distribution attributed to the system's Hamiltonian  $H(x)$ , i.e.,  $\rho(x_0) \propto e^{-\beta \cdot H(x_0)}$ . For a system evolving according to a Langevin equation, see Equation 59, the single step transition probability can be expressed as  $p(x_t \rightarrow x_{t+\Delta t}) = \sqrt{2\pi \cdot \sigma^2}^{-1} \cdot \exp(-(\Delta x)^2/2\sigma^2)$  in the high friction limit. This Gaussian distribution reflects the statistics of the random displacement  $\Delta x$  resulting from a thermally driven Gaussian random force, see Section 4.1 for more details. Having a zero mean, the width for the random displacement depends on the time step and the temperature,  $\sigma^2 = 2(k_B T/\gamma) \cdot \Delta t$ . Next, the random path ensemble represented by Equation 45 is restricted to merely reactive trajectories [55]:

$$\mathcal{P}_{ij}[x(\mathcal{T})] \equiv \frac{1}{Z_{ij}(\mathcal{T})} \cdot h_i(x_0) \cdot \mathcal{P}[x(\mathcal{T})] \cdot h_j(x_{\mathcal{T}}), \quad (46)$$

where the characteristic function  $h_i(x)$  forces the path to start in state  $i$ , and  $h_j(x)$  constrains the path to end in state  $j$  according to:

$$h_{i,j} = \begin{cases} 1 & \text{if } x \in i, j \\ 0 & \text{if } x \notin i, j. \end{cases} \quad (47)$$

The normalization factor  $Z_{ij}(\mathcal{T})$  has the form of a partition function and is defined as  $Z_{ij}(\mathcal{T}) \equiv \int \mathcal{D}x(\mathcal{T}) h_i(x_0) \cdot \mathcal{P}[x(\mathcal{T})] \cdot h_j(x_{\mathcal{T}})$ , where  $\int \mathcal{D}x(\mathcal{T}) \equiv \int \dots \int dx_0 dx_{\Delta t} dx_{2 \cdot \Delta t} \dots dx_{\mathcal{T}}$  indicates a summation over all pathways  $x(\mathcal{T})$ . Taken

together, the probability functional given by Equation 46 is a statistical description of all reactive pathways of duration  $\mathcal{T}$  connecting states  $i$  and  $j$ . This set of paths where each path has its own statistical weight proportional to  $\mathcal{P}[x(\mathcal{T})]$  is termed the transition path ensemble.

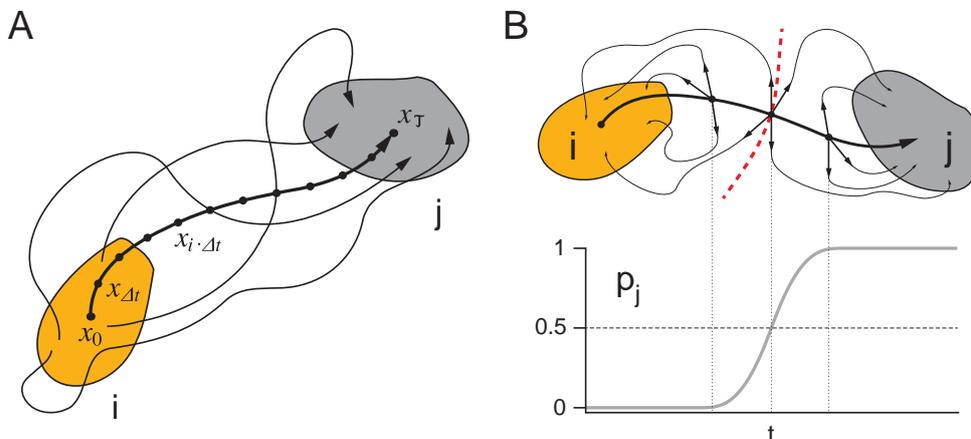


Figure 16: Illustration of the transition path ensemble and the commitment probability. (A) A transition path connects stable regions  $i$  and  $j$  (thick solid line). The transition path ensemble consists of all pathways interconnecting these two regions; examples are indicated (thin lines). (B) The committor  $p_j(r, t_s)$  for a time slice  $t$  along a transition path (thick solid line, top panel) reflects the fraction of pathways initiated from a configuration  $r$  that reach stable region  $j$  within time  $t_s$ . Transition states lie along the separatrix (red dashed line), where  $p_j \approx 0.5 \approx p_i$ .

The commitment probability<sup>11</sup>, in short committor, plays a central role in the analysis of transition paths as it allows the definition and the study of the ensemble of transition states positioned on the separatrix - a surface dividing basins of attraction of stable states. In principle, the committor calculates the relaxation probability of a system into a final state, here  $i$ , within a short time  $t_s$  given that the system is in the initial configuration  $r$  [28]:

$$p_i(r, t_s) \equiv \frac{\int \mathcal{D}x(t_s) \mathcal{P}[x(t_s)] \delta(r_0 - r) \cdot h_i(x_{t_s})}{\int \mathcal{D}x(t_s) \mathcal{P}[x(t_s)] \delta(r_0 - r)}. \quad (48)$$

The delta function on the right hand side assures that only traces  $x(t_s)$  which start in configuration  $r_0 = r$  contribute to  $p_i(r, t_s)$ . The denominator effectively represents the equilibrium probability distribution for configuration  $r$  and is used for normalization. For a two-state system with stable states  $i$  and  $j$  the condition  $p_i(r, t_s) = p_j(r, t_s) = 0.5$  defines the configuration  $r$  as a transition state. Owing to their relation to the transition state ensemble which comprises all configurations critical for a reaction, committor distributions which peak around a value of  $p_{i/j} = 0.5$  serve as proof for a good choice of reaction coordinate.

<sup>11</sup> The commitment probability is also known as the splitting probability. In the context of protein folding the terms  $p_{\text{fold}}$  and  $p_{\text{unfold}}$  are typically used [42, 143].

Finally the whole introduced set of constrained and unconstrained path ensembles and characteristic functions can be used to calculate reaction rate constants as well as free energies within the TPT framework [26, 28, 237].

### 3.8.1.1 Committor calculation from equilibrium trajectories

The experimentally recorded equilibrium fluctuations of a protein's extension are nothing but a series of concatenated forward and backward transition paths which are only constrained in time by the duration of the experiment and not by default as in Equation 48. Therefore, under the assumption of overdamped Langevin dynamics with constant diffusion coefficient  $D$ , it is also possible to estimate the commitment towards state  $i$  for a given reaction coordinate value  $x$  from an observed equilibrium trajectory  $X(t)$ ,  $t \in [0, \mathcal{T}]$ <sup>12</sup>, according to [42]:

$$\hat{p}_i(x) = \frac{\int_0^{\mathcal{T}} dt \delta(x - X(t)) \cdot c_{x_i}(t)}{\int_0^{\mathcal{T}} dt \delta(x - X(t))}. \quad (49)$$

Here, the hitting function  $c_{x_i}(t)$  does the characteristic function's job by controlling whether  $X(t)$  hits an absorbing boundary at  $x_i$  (representing the location of state  $i$  along the reaction coordinate) before  $x_j$  (representing the location of another state  $j$ ) immediately following time  $t$  in which case it assumes unity, and zero otherwise. In other words, as indicated by the delta functions, the denominator counts all crossings of  $X(t)$  at position  $x$  while the numerator counts only those crossings which are followed by a crossing at position  $x_i$  prior to a crossing at position  $x_j$ . Numerically this is accomplished by filling up histogram bins for an appropriate number of positions  $x$  between  $x_i$  and  $x_j$ . Figure 17 visualizes what is effectively calculated by Equation 49.

### 3.8.2 Committor-based barrier shape reconstruction

Consider a double-well potential  $G(x)$  with absorbing boundaries  $x_i$  and  $x_j$  placed in the vicinity of the two minima representing states  $i$  and  $j$  as depicted in Figure 17A. Then, for a diffusive process governed by overdamped Langevin dynamics, the probability of first encountering  $x_i$  before  $x_j$  starting from  $x \in [x_i, x_j]$  is given by [42, 179]:

$$p_i(x) = \frac{\int_x^{x_j} dx' D(x')^{-1} \cdot e^{\beta \cdot G(x')}}{\int_{x_i}^{x_j} dx' D(x')^{-1} \cdot e^{\beta \cdot G(x')}}. \quad (50)$$

Here, the denominator represents the equilibrium distribution  $p_{\text{eq}}(x)$  of all data points within the two boundaries. A potential dependence of the diffusion constant  $D$  on the position  $x$  is indicated. Figure 17C illustrates the distributions described by the integrals in Equation 50.

<sup>12</sup> Here,  $t$  is a running variable and no maximal duration value as in Equation 48.

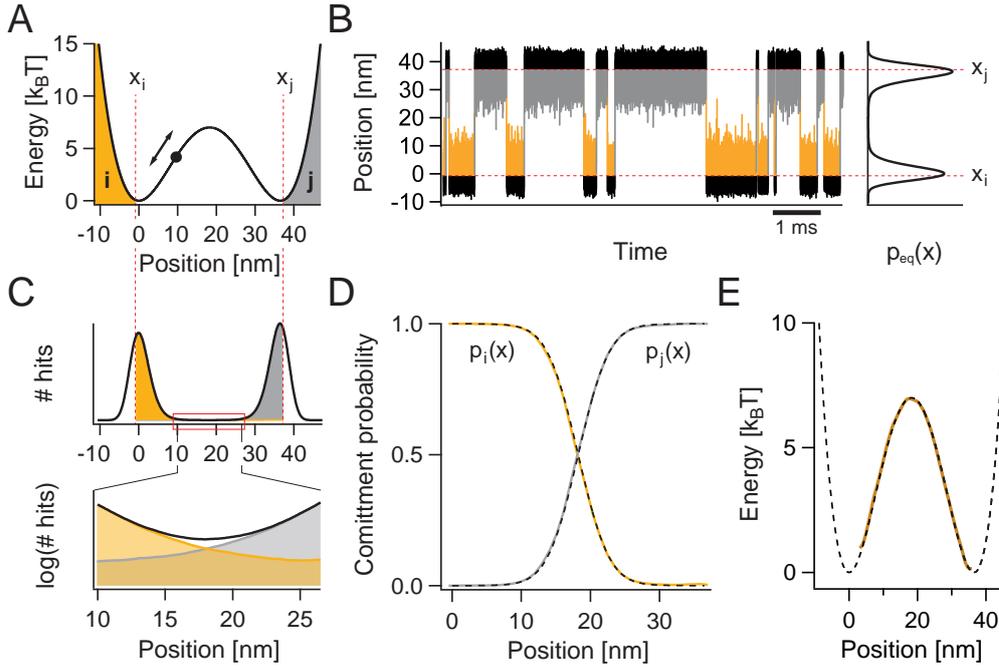


Figure 17: Commitment probability calculation and committor-based barrier shape reconstruction. (A) Potential  $G(x)$  used for a 1D diffusion Monte Carlo simulation. Dashed lines indicate positions of absorbing boundaries. (B) Representative stretch of simulated time trace with positional distribution to the left. Colour-coded commitment to state  $i$  (orange) or  $j$  (gray) forward in time. (C) ‘Split’ probability distributions (upper panel) and a zoom into the critical transition region (lower panel). (D) Resulting commitment probabilities calculated using Equation 49. Dashed lines represent committors directly calculated from the underlying potential. (E) Reconstructed barrier shape (dark orange) according to Equation 51.

By inversion of Equation 50 a direct relation between committors and their determining energy profile  $G(x)$  can be established [143]:

$$G(x) = \beta^{-1} \cdot \ln \left( [D(x)] \cdot \left| \frac{dp_i(x)}{dx} \right| \right). \quad (51)$$

As Equation 51 only holds to within a constant offset,  $D(x)$  may be omitted if it is assumed to have the same value at all positions  $x$ . In that case,  $\ln(D)$  can be separated as another additional constant offset which does not affect the overall barrier shape. As shown in Figure 17E, Equation 51 allows the shape of the barrier of  $G(x)$  to be reconstructed from the corresponding committors.

### 3.8.3 Bayesian path statistics localise transition states

Transition path segments isolated from equilibrium trajectories can be used to estimate the transition path (TP) ensemble  $p(x|TP)$ , a conditional phase-space density, from histograms of points  $x$  along the isolated segments, see Figure 18B. As individual path lengths are variable in terms of their duration,

each path enters the transition path ensemble with a relative weight  $1/\tau_{TP}$  to ensure detailed balance. By applying a Bayesian relation between equilibrium and transition path ensembles, the conditional probability of being on a transition path given that the system is at position  $x$  is given by [101]:

$$p(TP|x) = \frac{p(x|TP) \cdot p(TP)}{p_{eq}(x)}. \quad (52)$$

Here,  $p_{eq}(x)$  represents the equilibrium probability distribution of the entire trajectory and  $p(TP)$  indicates the fraction of time spent on transition paths within the trajectory. While  $p(TP)$  serves as a normalizing factor,  $p(x|TP)$  and  $p_{eq}$  are normalized probability distributions the ratio of which 'measures' the significance of  $x$  in transitions between stable states.

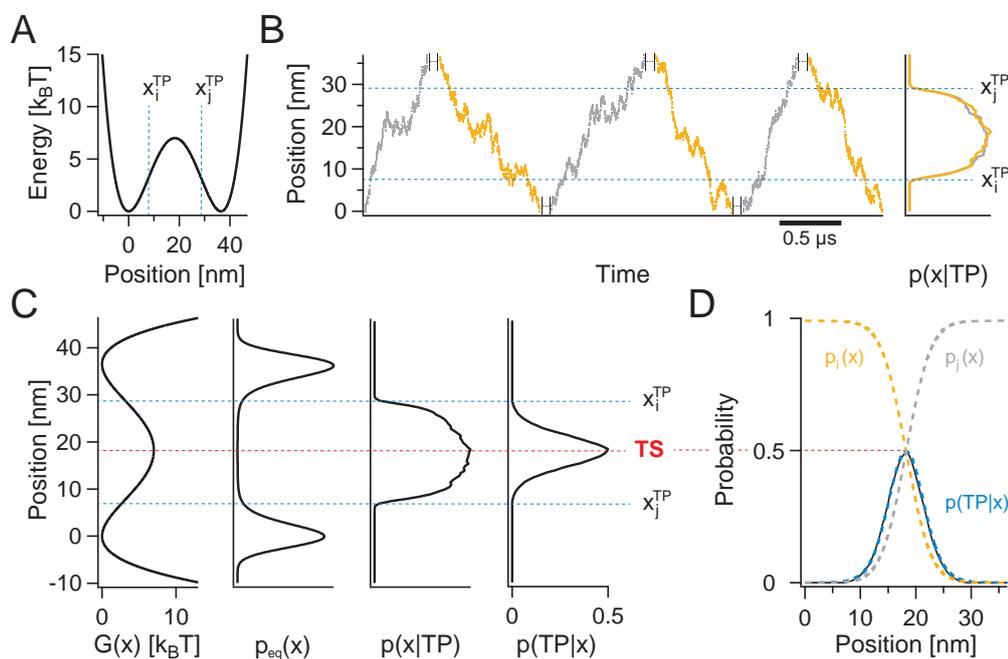


Figure 18: Transition path probability calculation and transition state localization. (A) Potential  $G(x)$  used for a 1D diffusion Monte Carlo simulation. Dashed lines confine the transition path region to within the inflection points of  $G(x)$ . (B) Isolated forward (orange) and backward (gray) transition path segments. Right panel: transition path ensemble distributions of forward and backward paths perfectly overlap. (C) From left to right: potential  $G(x)$  used to produce a simulated trajectory yielding the equilibrium distribution  $p_{eq}(x)$ , properly weighted transition path ensemble  $p(x|TP)$  of all forward and backward transitions, and transition path probability  $p(TP|x)$ . (D) Dashed lines represent comittors and the transition path probability directly calculated from  $G(x)$ . The black solid line shows  $p(TP|x)$  from the right of (C).

As can be seen in Figure 18C, the transition path probability  $p(TP|x)$  is sharply peaked at the transition state position where it ideally reaches a value of 0.5 in case where  $x$  is a good reaction coordinate for capturing the dynamics of a two-state system.

Furthermore, the transition path probability directly relates to the product of the system's commitment probabilities according to [101]:

$$p(\text{TP}|x) = 2 p_i(x) \cdot p_j(x) = 2 p_i(x) \cdot [1 - p_i(x)], \quad (53)$$

where the factor of two arises from the fact that a forward and backward path can always be swapped [27]. The correlations implied by Equation 53 are illustrated in Figure 18D. Subsequently, the committor-based barrier shape reconstruction introduced in Section 3.8.2 can be achieved by calculating  $p(\text{TP}|x)$  as well.

#### 3.8.4 Transition path times and kinetic rates

The average transition path time  $\langle \tau_{\text{TP}} \rangle$  which is necessary for crossing a harmonic barrier by one-dimensional diffusion is approximately<sup>13</sup> given by [44, 101, 138]:

$$\langle \tau_{\text{TP}} \rangle \approx \frac{\ln(2e^{\gamma_E} \cdot \beta \cdot \Delta G^{\text{TS}})}{\beta \cdot D \cdot \omega_{\text{TS}}^2} = \frac{\gamma_E + \ln(2\beta \cdot \Delta G^{\text{TS}})}{\omega_{\kappa}}. \quad (54)$$

Here,  $\gamma_E \approx 0.577\dots$  is Euler's constant,  $\omega_{\text{TS}}^2$  is the curvature at the barrier top, and  $\beta \cdot D \cdot \omega_{\text{TS}}^2 = \omega_{\kappa}$ . Equation 54 is valid in the high-barrier limit, i. e.,  $\Delta G^{\text{TS}} > 2 k_{\text{B}}T$ , and assumes a constant diffusion coefficient  $D$ .

The right-hand side of Equation 54 implies that  $\langle \tau_{\text{TP}} \rangle$  is composed of two contributions, where  $\gamma_E/\omega_{\kappa}$  appears to be a mere diffusive term. However, it can be shown that  $\omega_{\kappa}$  depends on the barrier height through  $\omega_{\text{TS}}$ . Due to  $\omega_{\text{TS}}^2 = |\partial^2 G(x)/\partial x^2|$  at the barrier top, it is possible to express  $\omega_{\text{TS}}^2$  in mere terms of  $\Delta G^{\text{TS}}$  and  $\Delta x^{\text{TS}}$  for different types of potentials  $G(x)$  yielding the following universal dependence:

$$\omega_{\text{TS}}^2 = \kappa_{\star} \cdot \frac{\Delta G^{\text{TS}}}{(\Delta x^{\text{TS}})^2}, \quad (55)$$

where  $\kappa_{\star}$  corresponds to either  $\kappa_{\text{harm}} = 4$ ,  $\kappa_{\text{sin}} = \pi^2/2$ , or  $\kappa_{\text{cubic}} = 6$  for a harmonic, sinusoidal, or linear-cubic potential<sup>14</sup>, respectively. Hence, Equation 55 implies that  $\omega_{\kappa}$  changes proportionally to changes of the barrier height  $\Delta G^{\text{TS}}$  if the distance between the minima of the double-well potential is kept constant, see inset of Figure 19A. Respecting this interdependence, the relation of  $\langle \tau_{\text{TP}} \rangle$  to characteristic parameters of  $G(x)$  can be deduced from Equation 54. For the case of a harmonic potential, Figure 19A and Figure 19C illustrate<sup>15</sup>

13 The underlying analytic expression reads  $\langle \tau_{\text{TP}} \rangle = D^{-1} \cdot \int_{x_i}^{x_j} e^{-\beta \cdot G(x)} \cdot p_i(x) \cdot p_j(x) dx \cdot \int_{x_i}^{x_j} e^{\beta \cdot G(x)} dx$ , where  $G(x)$  is the free energy as the function of the reaction coordinate  $x$  and  $p_i(x)$  and  $p_j(x)$  are defined by Equation 50.

14 The potentials used to derive Equation 55 are constructed such that they provide a well-defined  $\Delta x^{\text{TS}}$  and  $\Delta G^{\text{TS}}$ . In detail,  $G_{\text{harm}}^{\text{well}}(x) = 2\Delta G^{\text{TS}} \cdot (x/\Delta x^{\text{TS}})^2$  in combination with a turned and shifted  $G_{\text{harm}}^{\text{barrier}}(x)$ ,  $G_{\text{sin}}(x) = \frac{1}{2}\Delta G^{\text{TS}} \cdot \cos(\pi \cdot (x/\Delta x^{\text{TS}}))$ , and  $G_{\text{cubic}}(x) = \frac{3}{2}\Delta G^{\text{TS}} \cdot (x/\Delta x^{\text{TS}}) - 2\Delta G^{\text{TS}} \cdot (x/\Delta x^{\text{TS}})^3$ .

15 Parameters used for the calculations in Figure 19 were chosen such that  $\langle \tau_{\text{TP}} \rangle$  reaches the expected speed limit of about 1  $\mu\text{s}$  for a protein comprising 100 amino acids [119].

the inferred dependence on the barrier height  $\Delta G^{\text{TS}}$  and the position of the transition state  $\Delta x^{\text{TS}}$  while the respective other parameter is kept constant. Both dependencies were tested and confirmed by one-dimensional diffusion simulations as described in Section 4.1, data not shown.

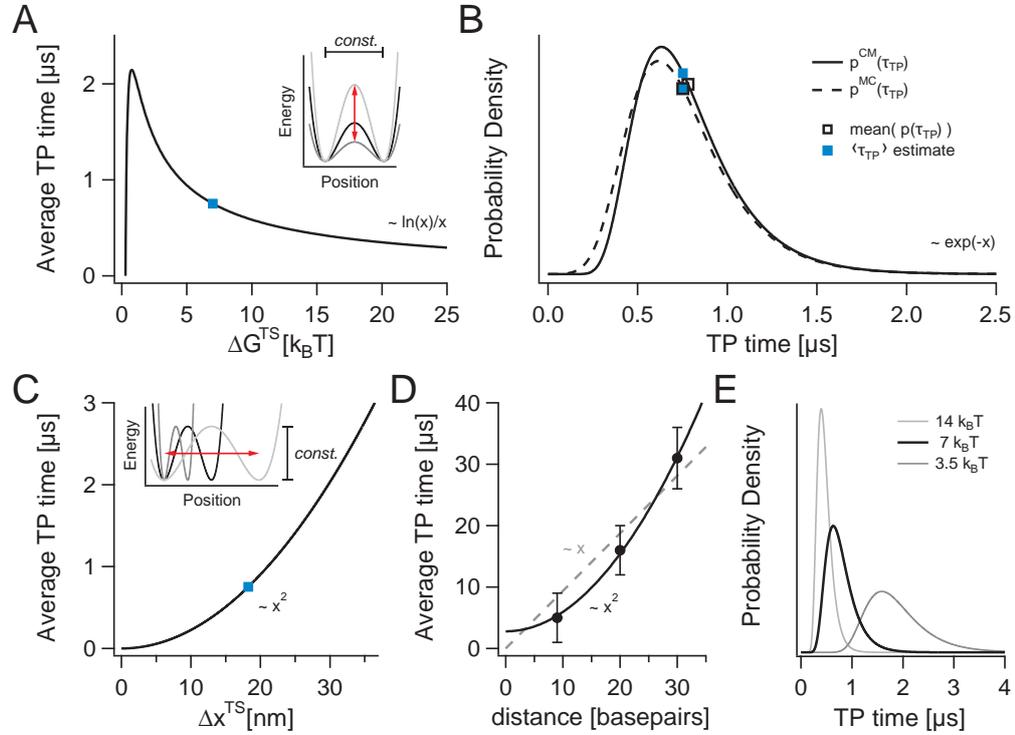


Figure 19: Transition path time averages and distributions. Constant parameters are  $\Delta G^{\text{TS}} = 7 k_B T$ ,  $\Delta x^{\text{TS}} = 36.5/2 \text{ nm}$ , a friction coefficient of  $\gamma = 4 \times 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , and thereby  $D = k_B T / \gamma = 10^8 \text{ nm}^2 \cdot \text{s}^{-1}$  at 300 K. (A) Barrier height-dependence of  $\langle \tau_{\text{TP}} \rangle$  for a fixed transition state position. (B) Probability distributions of  $\tau_{\text{TP}}$  according to Equation 56 and Equation 57 with  $\kappa = 1 k_B T$ . The mean values of the distributions agree well with  $\langle \tau_{\text{TP}} \rangle$  estimated by Equation 54. (C) Dependence of  $\langle \tau_{\text{TP}} \rangle$  on the transition state position at constant harmonic barrier height. (D) Experimentally determined stem-length dependence of transition path times for DNA hairpin unfolding fitted by a quadratic function. Data adapted from [154]. (E) Barrier height-dependence of  $p^{\text{CM}}(\tau_{\text{TP}})$  for a fixed transition state position.

In contrast to the strong force-dependence of kinetic rates which typically change over several orders of magnitude (see Section 3.5), the transition path time shows a much weaker dependence on both the barrier height and the transition state position. An increasing  $\Delta G^{\text{TS}}$  leads to a shorter  $\langle \tau_{\text{TP}} \rangle$  as a result of confining the probability distribution of transition path times to those from transition paths with non-vanishing dynamical path probabilities. This confinement can be recognized in Figure 19E. As expected,  $\langle \tau_{\text{TP}} \rangle$  becomes longer by increasing the distance between the two stable configurations and, hence, by increasing  $\Delta x^{\text{TS}}$ , see inset of Figure 19C.

It is remarkable that in agreement with the quadratic dependence on the transition state position, the experimentally determined stem-length dependence of transition path times of DNA hairpin unfolding is very well fitted by a quadratic function, too, see [Figure 19D](#). To further investigate this correlation, a corresponding analysis of protein coiled-coil constructs with different transition state positions can be considered based on constructs and construct design introduced in [Section A.1](#) and [Section A.2.2](#).

An approximate expression for the distribution of transition path times is given by [\[41, 251\]](#):

$$p^{\text{CM}}(\tau_{\text{TP}}) \approx \frac{\omega_{\kappa} \sqrt{\beta \cdot \Delta G^{\text{TS}}}}{1 - \text{erf}\left(\sqrt{\beta \cdot \Delta G^{\text{TS}}}\right)} \cdot \frac{\exp\left[-\beta \cdot \Delta G^{\text{TS}} \cdot \coth\left(\omega_{\kappa} \cdot \tau_{\text{TP}}/2\right)\right]}{\sinh\left(\omega_{\kappa} \cdot \tau_{\text{TP}}/2\right) \cdot \sqrt{2\pi} \cdot \sinh\left(\omega_{\kappa} \cdot \tau_{\text{TP}}\right)}, \quad (56)$$

where  $\text{erf}(\dots)$  is the error function and the superscript refers to Chaudhury and Makarov (CM) who presented the explicit form of [Equation 56](#). Another formula to calculate the probability distribution of transition path times has been introduced by Malinin and Chernyak (MC) [\[138\]](#):

$$p^{\text{MC}}(\tau_{\text{TP}}) = (2\omega_{\kappa} \cdot \Delta G^{\text{TS}}/\kappa) \cdot \exp\left\{-\omega_{\kappa} \cdot \tau_{\text{TP}} - (2\Delta G^{\text{TS}}/\kappa) \cdot e^{-\omega_{\kappa} \cdot \tau_{\text{TP}}}\right\}. \quad (57)$$

Here,  $\kappa$  describes the noise strength<sup>16</sup> in units of  $k_{\text{B}}T$ . Apart from different variables, [Equation 57](#) is of the same form as for a distribution of refolding forces, see [Section 3.6](#). The typical shape of transition path time distributions is shown in [Figure 19B](#).

Finally, the average time needed to transit between two states  $i$  and  $j$  relates to the corresponding transition rates according to [\[101\]](#):

$$k^{ij} \approx \frac{p(\text{TP})}{2 \langle \tau_{\text{TP}} \rangle \cdot P_i}. \quad (58)$$

While  $p(\text{TP})$  is the fraction of time spent on transition paths,  $P_i$  is the equilibrium probability of being in state  $i$ . By this means, [Equation 58](#) requires detailed knowledge of the entire equilibrium trajectory.

### 3.8.5 Extracting transition paths

In order to extract transition paths from an equilibrium trajectory it is necessary to first choose the transition path boundaries. A transition path is then typically defined as the shortest interconnection between the chosen boundaries where after crossing the first boundary the second boundary must be reached prior to recrossing of the first one.

[Figure 20A](#) shows the effect of different boundary positions  $x_i^{\text{TP}}$  and  $x_j^{\text{TP}}$  on the transition path ensemble and probability extracted from a one-dimensional diffusion Monte Carlo simulation (for simulation details see [Section 4.1](#)). Even

<sup>16</sup> For  $\kappa = 1 k_{\text{B}}T$ , the average transition path time derived from [Equation 57](#) exactly matches [Equation 54](#) [\[138\]](#).

though the transition path ensembles  $p(x|TP)$  look quite different, the transition path probabilities  $p(TP|x)$  calculated according to Equation 52 perfectly match each other. This confirms the robustness of Equation 52 which allows the transition path boundaries to be flexibly chosen within the range of a vanishing transition path probability. Basically, this is possible because the normalizing factor  $p(TP)$  compensates for the altered height of the normalized transition path ensemble along with the altered transition path times  $\tau_{TP}$  which in turn have to be interpreted with care.

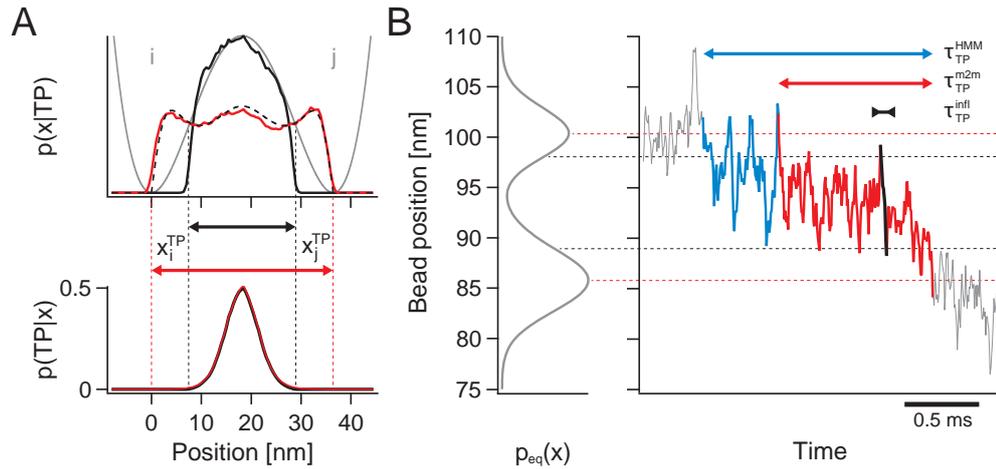


Figure 20: Transition path extraction from simulated and experimental trajectories. (A) Transition path ensembles (upper panel) and probabilities (lower panel) for differently chosen transition path boundaries between either the minima (red) or the inflection points (black) of the underlying potential (gray) of a 1D diffusion simulation. The theoretically expected transition path ensemble is shown as a dashed line. (B) The equilibrium distribution of a measured protein folding trajectory (left) is used to locate transition path boundaries either at the maxima or the inflection points indicated by red or black dashed lines. For an individual transition (right) the read-out of transition path times between maxima  $\tau_{TP}^{m2m}$  or inflection points  $\tau_{TP}^{infl}$  differs drastically. The transition path time suggested by HMM analysis assuming multiple very short-lived states along the transition path is marked in blue.

Other than for simulations, the original folding energy landscape underlying experimentally recorded protein fluctuations is hidden behind additional thermal noise coupling to the measurement system, see Section 3.7. On this account, the entire region between maxima of the equilibrium distribution of the differential bead position signal has to be considered to extract transition paths as exemplified in Figure 20B. This assures the transition state position to be included which is an issue especially with regard to protein folding energy landscapes potentially being asymmetric. Furthermore, this way the transition path involves the crossing of the full barrier height which becomes relevant for transition path time calculations as introduced in Section 3.8.4 and confirmed in Section 4.2.1.

Besides an unknown exact energy landscape shape, another issue of experimental data is their limited resolution. Alongside strong filtering which in turn leads to a reduction of spatial information, one can consider to treat transition paths as a series of adjoining short-lived intermediate states. A corresponding [HMM](#) analysis will then allow the differentiation between thermal noise along the transition path and actual transition path boundary crossing. This method is particularly useful when the lifetimes of the two considered stable states are much longer than the lifetimes of potential high-energy intermediates identified along their interconnecting transition paths, see [Section 6.2.3](#).



## Part III

### SIMULATION RESULTS

Binary-supported tests of theories and hypotheses.



## LANGEVIN DYNAMICS SIMULATIONS

---

Numerical simulations are a powerful tool for testing theoretical models and hypothesis derived from experimental observations. One particular strength of simulations lies within providing the environment for and the control over a completely self-contained deterministic system. In the following, the principle of the simulation of our entire force-spectroscopic measurement system is briefly introduced. Thereafter, simulations are performed to test different formulas, to investigate different concepts of roughness, and to analyse the response of the measurement system to different underlying energy landscapes of protein folding with a particular focus on friction.

### 4.1 SIMULATION PRINCIPLE

The Brownian motion of a particle in an external potential  $G(x)$  can be described by the following Langevin equation [17]:

$$m \cdot \ddot{x}(t) + \gamma \cdot \dot{x}(t) + \partial_x G(x(t)) = \sqrt{2k_B T \cdot \gamma} \cdot \zeta(t). \quad (59)$$

The three terms on the left describe forces acting on the particle which arise from inertia inferred by the mass  $m$  of the particle, friction induced by the surrounding medium characterized by the friction coefficient  $\gamma$ , and the local potential gradient. The term on the right of Equation 59 represents a Gaussian random force uncorrelated in time and satisfying the fluctuation dissipation theorem. Therefore,  $\zeta(t)$  is an uncorrelated (white) noise function with mean value  $\langle \zeta(t) \rangle = 0$  and variance  $\langle \zeta(t)\zeta(t') \rangle = \delta(t - t')$  [142]. In the limit of high friction which is also referred to as the overdamped case, the inertial term in Equation 59 can be omitted. Now, after transiting from continuous time to discrete time steps by replacing  $\dot{x} = \frac{\Delta x}{\Delta t}$ , where  $x$  still implies  $x(t)$ , rewriting of Equation 59 leads to<sup>1</sup>:

$$\Delta x = \frac{\Delta t}{\gamma} \cdot \left( -\partial_x G(x) \cdot k_B T + \sqrt{2k_B T \cdot \gamma / \Delta t} \cdot \zeta(t) \right). \quad (60)$$

After choosing an initial position  $x_{\text{init}}$  at time  $t = 0$ , the simulation of a trajectory of a diffusing particle is performed by iteratively evaluating Equation 60 at each time step  $\Delta t$  and adding the resulting  $\Delta x$  to the current position such that  $x(t + \Delta t) = x(t) + \Delta x$ . Replacing  $G(x)$  with  $H(x, i)$  from Equation 18 turns Equation 60 into an equation suitable to simulate the differential signal of our optical trapping experiments where the currently populated state  $i$  of the protein is decided upon by a Monte Carlo step based on predefined values  $k^{ij}$  [212].

<sup>1</sup> The additional factor of  $k_B T$  is applied in units of  $\text{pN} \cdot \text{nm}$  to compensate for  $G(x)$  being given in units of  $k_B T$ . The amplitude of the random force is additionally divided by  $\sqrt{\Delta t}$  to assure an effective variance of  $\sigma^2 = 2(k_B T / \gamma) \cdot \Delta t$  for the thermally driven random displacement.

A purely mechanical and, hence, more general approach for setting up a simulation of the entire force-spectroscopic measurement system is achieved by moving from discrete protein states  $i$  towards a continuous protein energy landscape where the contour length  $L_P$  of unfolded polypeptide serves as reaction coordinate. By this means and in analogy to Equation 18, a two-dimensional Hamiltonian  $H(x, L_P)$  is introduced to describe the full energy function of the system when held at constant trap separation  $d$  [177]:

$$H(x, L_P) = G_0(L_P) + \frac{1}{2}k_{\text{eff}} \cdot x^2 + \int_0^{d-2R-x} F_{\text{tether}}(x_{\text{tether}}, L_P) dx_{\text{tether}}. \quad (61)$$

The coordinate  $x$  again comprises the sum of the two bead deflection signals  $x_1 + x_2$ . The extension  $x_{\text{tether}}$  of DNA-handles and unfolded protein is given by Equation 1 and Equation 4, where  $x_P(F) \cdot P_{\text{unf}}(F)$  is replaced by  $x_P(F)$  alone and defined by  $L_P$ . The acting force  $F_{\text{tether}}$  is given by the corresponding inverse of Equation 4. The most central ingredient of the Hamiltonian is the zero-force energy landscape  $G_0(L_P)$  which is tilted by the force acting at inter-trap distance  $d$ . To simulate the time evolution of the system described by  $H(x, L_P)$ , the current values for  $x$  and  $L_P$  are updated according to their stochastic equations of motion at each time step  $\Delta t$ :

$$\begin{aligned} \Delta x &= \frac{\Delta t}{\gamma_B} \cdot \left( -\partial_x H(x, L_P) \cdot k_B T + \sqrt{2k_B T \cdot \gamma_B / \Delta t} \cdot \zeta(t) \right) \\ \Delta L_P &= \frac{\Delta t}{\gamma_P} \cdot \left( -\partial_{L_P} H(x, L_P) \cdot k_B T + \sqrt{2k_B T \cdot \gamma_P / \Delta t} \cdot \zeta(t) \right). \end{aligned} \quad (62)$$

With the beads being the largest components of the system, the friction along the measurement coordinate  $x$  is dominated by Stokes friction. Hence, for the sake of simplicity,  $\gamma_B = 6\pi \cdot \eta \cdot R$  is used within the simulations, where  $R = 500 \text{ nm}$  for the bead radius and  $\eta = 10^{-9} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-2}$  for the viscosity. The friction coefficient  $\gamma_P$  in the protein potential is usually adapted such that experimental rates are recovered by the simulation. At room temperature it can attain<sup>2</sup> values of  $\gamma_P \sim 10^{-8} - 10^{-3} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The time step  $\Delta t$  between simulated data points crucially depends on the smallest friction coefficient within the system and needs to be chosen such that the underlying energy landscape is fully sampled by the simulation. Here,  $\Delta t = 10^{-9} \text{ s}$  is used if not stated differently. As experimentally recorded time traces typically have much bigger time steps  $\Delta t_{\text{record}}$  due to slower sampling, only each  $(\Delta t_{\text{record}} / \Delta t)$ -th simulated data point is saved to realize time traces directly comparable to experiments.

In Figure 21 a simulation example of a constant distance experiment is illustrated by showing the two-dimensional energy landscape defined by Equation 61 as well as the resulting time traces calculated according to Equation 62. In the upper panel of Figure 21B black arrows indicate lost transitions when sampling is reduced from full resolution at 1 GHz (black) down to 200 kHz

<sup>2</sup> Reported values for intrachain diffusion in unfolded proteins typically reach  $D_P \sim 10^7 - 10^8 \text{ nm}^2 \cdot \text{s}^{-1}$ . However,  $D_P$  can drop down to  $10^3 - 10^5 \text{ nm}^2 \cdot \text{s}^{-1}$  in case of particularly slow diffusion or misfolding transitions [92, 153, 231, 249]. Friction and diffusion are linked over temperature:  $D_P \cdot \gamma_P = k_B T$ .

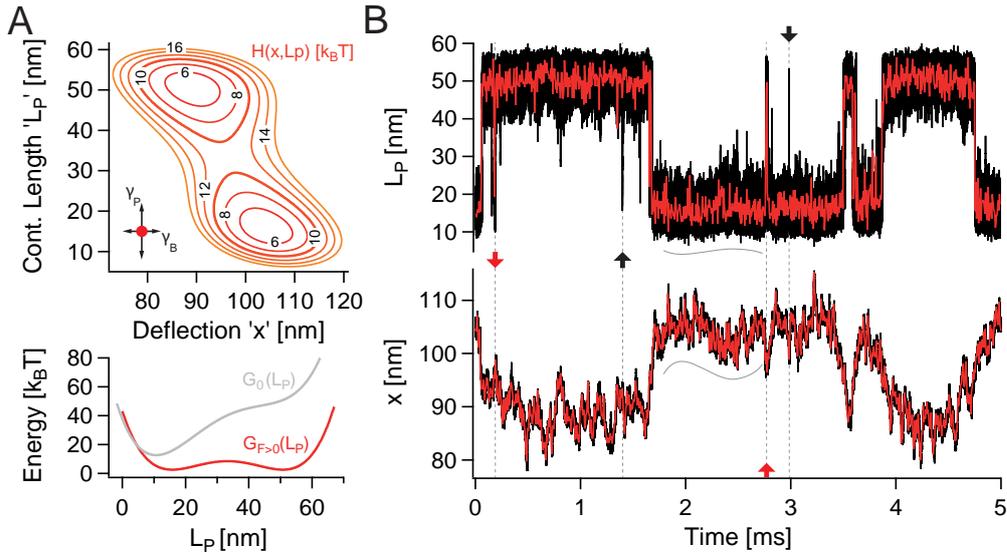


Figure 21: Hamiltonian and time traces of an energetically coupled two-dimensional Langevin dynamics simulation. Parameters are  $T = 300\text{ K}$ ,  $\gamma_B = 9.4 \cdot 10^{-6}\text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ ,  $\gamma_P = 4 \cdot 10^{-8}\text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , and  $\Delta t = 10^{-9}\text{ s}$ . (A) Upper panel: Hamiltonian calculated according to Equation 61 for a two-state system under force. Lower panel: the underlying zero-force protein energy landscape  $G_0(L_P)$  (gray) is tilted by force (red). Parameters: inter-trap distance  $d = 479\text{ nm}$ ,  $k_{\text{eff}} = 0.11\text{ pN/nm}$ ,  $p_P = 0.7\text{ nm}$ ,  $L_D \sim 370\text{ nm}$ ,  $p_D \sim 50\text{ nm}$ ,  $K \sim 200\text{ pN}$ , and  $T = 300\text{ K}$ . (B) Simulated time traces for protein diffusion (top) and bead displacement (bottom) based on  $H(x, L_P)$  shown in (A). Traces are sampled at  $1\text{ GHz}$  (black, full resolution) and  $200\text{ kHz}$  (red, maximal experimental resolution). Arrows indicate lost transitions (black) and missed events (red) due to differences in sampling and friction.

(red) pointing out the need for a high enough resolution. In contrast, owing to much higher involved friction, a resolution of  $200\text{ kHz}$  appears sufficient to fully recover the bead displacement shown in the lower panel. However, due to higher friction the effect of missed events needs to be considered (red arrows). Going further one can already slightly see how the slower bead diffusion affects the diffusion in the protein potential by comparing the outer border of the time trace of the protein to the average bead position (see solid gray lines to guide the eye). For even higher differences in friction, this can induce drastic shifts in transition rates as the assumption of a constant force bias breaks down.

As a final remark it needs to be noted that by merely following the coordinates of bead deflection and diffusion in a protein potential the presented simulation only provides a minimal representation of our entire measurement system. One central assumption is that the DNA-handles and unfolded polypeptide are in equilibrium with the beads at all times. This is legitimate as the beads are the system's slowest component [142]. Further, by introducing the differential signal as one coordinate which combines both beads, the

latter need to be assumed close to equilibrium, too. Under force, this assumption can be considered valid for time scales down to  $\sim 10 \mu\text{s}$  [261].

## 4.2 STUDIES ON ONE-DIMENSIONAL ROUGH ENERGY LANDSCAPES

All one-dimensional diffusion simulations performed in this section are based on Equation 60. At first, equations to estimate transition path times, their distribution, and their connection to transition rates are tested. Next, the predicted effect of an average overall energy landscape roughness on transition path times is investigated. Thereafter, a method for energy landscape reconstruction from transition path times resulting from local roughnesses represented by high-energy intermediates is introduced. Finally, the effect of force-induced transition state switching on overall transition rates is studied using the example of an energy landscape with a single high-energy intermediate.

### 4.2.1 Predicting transition path times and kinetic rates

To test the qualitative significance of simulation results, here, the transition path times and their distribution, and kinetic transition rates extracted from a simulated time trajectories are compared to their theoretical predictions. The relevant data is retrieved from the same trajectory already used in Section 3.8 to exemplify and confirm more quantitative transition path analysis methods including the calculation of committers and transition path probabilities.

The average transition path time out of 871 simulated forward and backward transitions between the two minima of the double-well potential<sup>3</sup> depicted in Figure 22A is  $\langle \tau_{\text{TP}}^{\text{sim}} \rangle = 0.66 \pm 0.21 \mu\text{s}$ . This value compares very well to a theoretical value of  $\langle \tau_{\text{TP}} \rangle = 0.75 \mu\text{s}$  predicted by Equation 54. The distribution of transition path times from the simulated trajectory shown in Figure 22B reveals a tendency towards shorter transition path times, too. However, the distribution obtained from the simulation compares reasonably well to the distribution expected according to Equation 56. In contrast, the inset of Figure 22B reveals that the distribution of transition path times extracted for transitions merely between the inflection points of the underlying potential are way too short to match with the theory.

A possible explanation for the slightly shorter transition path times provided by the simulation could be a lack of resolution. However, no noticeable change in the average extracted transition path time is obtained at an increased resolution of  $\Delta t = 10^{-10} \text{s}$  (data not shown). Another more tangible reason for the systematic tendency towards shorter transition path times is given by the actual energy landscape shape though. As can be seen in Figure 22A, the curvatures at the minima are significantly higher than the curvature at the barrier top. Yet, only the latter is considered when estimating the average transition path time according to Equation 54 which, by this means, appears to be a simplified estimate considering  $\omega_{\text{TS}} = \omega_i = \omega_j$ . An-

<sup>3</sup> The potential  $G(x) = \Delta G^{\text{TS}} \cdot \left( (x/\Delta x^{\text{TS}})^2 - 1 \right)^2$  is adapted from [101].

other simulation performed in a double-well potential which complies with this simplification indeed provides an average transition path time of  $\langle \tau_{\text{TP}}^{\text{sim}2} \rangle = 0.69 \pm 0.23 \mu\text{s}$  which deviates from the theoretical value by less than 10% (data not shown). Hence, for a more accurate prediction, transition path times should be calculated under consideration of the entire barrier shape. The simplest way to do so is by splitting up the barrier into four path segments delimited by the two energy minima, the maximum, and the inflection points in between. Now, by treating the energy minima as inverted barrier tops, each path segment can be attributed a curvature  $\omega_{\text{seg}}$  and a barrier height  $\Delta G_{\text{seg}}$  which equals twice the maximal energy difference along the respective path segment. The time provided by a path segment to the overall transition path time for crossing the entire barrier is then approximately given by one quarter of the time resulting from feeding  $\Delta G^{\text{seg}}$  and  $\omega_{\text{seg}}$  into Equation 54. The thereupon improved predictive capabilities of Equation 54 were confirmed by one-dimensional diffusion simulations as described in Section 4.1 (data not shown). Note that the briefly outlined sequential transition path time calculation is particularly useful to get more accurate results for asymmetric energy landscapes like, e. g., in Section 6.4.1.

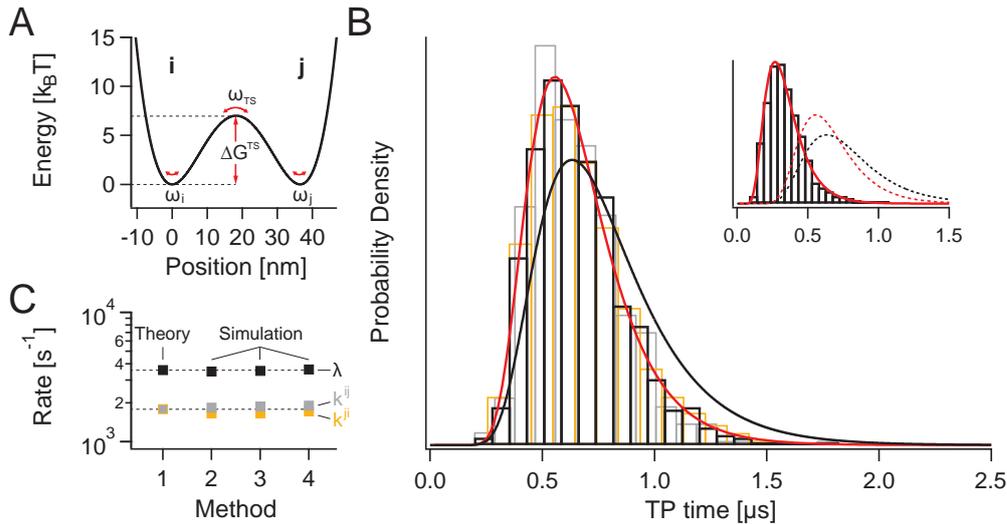


Figure 22: Analysis of transition path times and rates from a 1D diffusion simulation performed at  $T = 300 \text{ K}$  with  $\gamma = 4 \times 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The data set comprises a 0.5 s lasting time trajectory with 1 GHz resolution; also see Figure 17 and Figure 18. (A) Underlying double-well potential  $G(x)$  with  $\Delta G^{\text{TS}} = 7 k_{\text{B}}T$  separating states  $i$  and  $j$ . (B) Transition path time histogram of all (black),  $i$  to  $j$  (gray), and  $j$  to  $i$  (orange) transitions measured for crossing the entire barrier between both minima. Expected (black) and fit (red) distributions are calculated according to Equation 56. *Inset*: Fit transition path time histogram of all transitions between the inflection points of  $G(x)$ . Dashed lines refer to the curves from the main figure. (C) Graphical representation of the results listed in Table 1 with the methods being numbered according to their row of entry.

The red line in [Figure 22B](#) shows a perfectly matching fit to the extracted transition path time distribution based on [Equation 56](#). As expected, due to shorter transition path times, the fit yields an overestimated barrier of  $\Delta G^{\text{TS}} = 10.7 \pm 0.9 k_{\text{B}}T$  and a curvature increased by a factor of 1.3. In contrast, if the curvature is kept fixed as the value implied by the underlying potential, the energy barrier is underestimated as  $\Delta G^{\text{TS}} = 5 \pm 0.4 k_{\text{B}}T$ . In combination with extracting transition path times from transitions covering only the distance between the inflection points of the potential, fixing the value for the curvature at the barrier top leads towards a substantial underestimation of the barrier height ending up in  $\Delta G^{\text{TS}} = 1 \pm 0.2 k_{\text{B}}T$ . Within limits, this connection may partially explain the large discrepancy between barrier heights obtained either by deconvolution or from transition path time distributions of transitions covering the middle half of the total distance between the two main energy minima reported in [156]. Considering the weak dependence of  $\langle \tau_{\text{TP}} \rangle$  on  $\Delta G^{\text{TS}}$  as depicted in [Figure 19A](#), deducing barrier heights from transition path times involving mere barrier crossing is highly error-prone by default.

METHOD	$k^{ij}$ [s <sup>-1</sup> ]	$k^{ji}$ [s <sup>-1</sup> ]	$\lambda$ [s <sup>-1</sup> ]	REFERENCE
Kramers theory	1785	1785	3570	<a href="#">Equation 29</a>
Counting transitions	1831	1653	3484	see text for $k_{\#}^{ij}$
HMM	1873	1652	3525	<a href="#">Equation 24</a>
$\langle \tau_{\text{TP}}^{\text{sim}} \rangle$	1899	1714	3613	<a href="#">Equation 58</a>

Table 1: Rate constant calculations: comparing theory with simulation. Calculations are based on indicated references. The evaluated data set is the one described in [Figure 22](#), as well as [Figure 17](#) and [Figure 18](#).

Based on Kramers rate-theory in the high friction limit represented by [Equation 29](#), the transition rates between states  $i$  and  $j$  shown in [Figure 22A](#) are expected to be  $k^{ij} = k^{ji} = 1785 \text{ s}^{-1}$ . Together, both rates yield a combined rate of  $\lambda = k^{ij} + k^{ji} = 3570 \text{ s}^{-1}$ . To extract the transition rates from the simulated trajectory for comparison, three different methods are applied. The most direct method is simply based on counting the number of observed transitions per second  $\#t^{ij}$  from initial state  $i$  into final state  $j$  with respect to the fraction of time spent in the initial state. For a two-state system this is simply given by  $k_{\#}^{ij} = \#t^{ij}/P_i$ , where  $P_i$  is the equilibrium probability of being in state  $i$ . The two other methods for extracting rates are by performing a HMM analysis (see [Section 3.4.3](#)) and by using the information provided by transition paths in combination with the entire equilibrium trajectory (see [Section 3.8.4](#)). All results are listed in [Table 1](#) and illustrated in [Figure 22C](#). With respect to the predicted values, all extracted rates are faster for  $k^{ij}$  and slower for  $k^{ji}$ . This is simply caused by a slight asymmetry within the state occupancies which is owed to the limited duration of the simulation where  $P_i \sim 0.47$  and  $P_j \sim 0.53$ ,

see [Figure 17B](#). The combined rate resulting from counting the transitions slightly underestimates the predicted value. This deviation can be attributed to the limited resolution and is considered within the [HMM](#) analysis (see [Equation 24](#) in combination with [Equation 23](#)) which yields a combined rate very close to the theoretical value. The combined rate retrieved from the average transition path time ( $\langle \tau_{\text{TP}}^{\text{sim}} \rangle$ ) slightly overestimates the theoretical value, which, to some extent, may be caused by the underestimated transition path time average, see [Equation 58](#). All in all, as all deviations are smaller than 2.5 %, theory and simulation are found to match each other perfectly. Most importantly, by this means, Kramers rate-theory allows very precise estimates to be made not only for transition rates but also for expected average dwell times  $\tau^i = (k^i)^{-1}$ .

Note that during the simulation, as expected, starting from the energy minimum of either state  $i$  or state  $j$ , the transition state position at the barrier top is reached about twice as often as entire transitions to the respective other state actually occur.

#### 4.2.2 Diffusion in a rough energy landscape

Slowed down diffusion in a potential is often put down to an increased overall energy landscape roughness typically conceived as being composed of many small potential barriers randomly distributed along the reaction pathway [[106](#), [132](#), [159](#), [238](#), [249](#)]. A simple relation introduced by Zwanzig can be used to quantify roughness through its effect on observed diffusion [[262](#)]:

$$D^* = D \cdot \exp\left(-\left[\frac{\epsilon_{\text{rms}}}{k_{\text{B}}T}\right]^2\right). \quad (63)$$

Here,  $D^*$  is the effective diffusion coefficient in an effective smooth potential  $G(x)$  which replaces the original diffusion coefficient  $D$  in the originally rough potential  $G(x) + G_{\text{rough}}(x)$ . The parameter  $\epsilon_{\text{rms}}$  is the root-mean-squared roughness,  $\epsilon_{\text{rms}}^2 = \langle G_{\text{rough}}^2 \rangle$ , which reflects the typical amplitude of the perturbation superimposed on  $G(x)$ . For the case of a mere sinusoidal perturbation with amplitude  $\epsilon_{\pi}$ , Zwanzig introduced the less frequently used relation  $D^* = D \cdot \exp(-2\epsilon_{\pi}/k_{\text{B}}T)$  indicating Arrhenius behaviour valid specifically for low temperatures, where  $\epsilon_{\pi}/k_{\text{B}}T$  is very large. [Figure 23A](#) illustrates smooth potentials to which an either quasi-random or a simple periodic roughness have been added<sup>4</sup>. Next to its typical amplitude, an energy landscape roughness is characterized by its typical length scale  $\Delta x^{\text{r}}$  over which the amplitude averages out to zero. An important key for [Equation 63](#) to be valid is that  $\Delta x^{\text{r}}$  is small enough to be well separated from any other motion in the potential. Zwanzig chose  $\Delta x^{\text{r}}$  to be of the order of 0.1 in units of  $x$ .

<sup>4</sup> The quasi-random roughness  $G_{\text{rough}}^{\text{rdm}}(x) = \epsilon_{\text{rms}} \cdot [\cos(0.01 \cdot 167x) + \sin(0.01 \cdot 73x)]$  is built on the basis of the potential used in [[262](#)] with the additional factor 0.01 corresponding to  $\Delta x^{\text{r}}/\Delta x_{\text{appl}}^{\text{r}}$  (see text and caption of [Figure 23](#)). The simple periodic roughness is constructed according to  $G_{\text{rough}}^{2\pi}(x) = \epsilon_{\pi} \cdot \sin(0.01 \cdot 73x - \pi)$ . To match with typical scales in contour length space of protein folding, here,  $x$  assumes 'nm'-units.

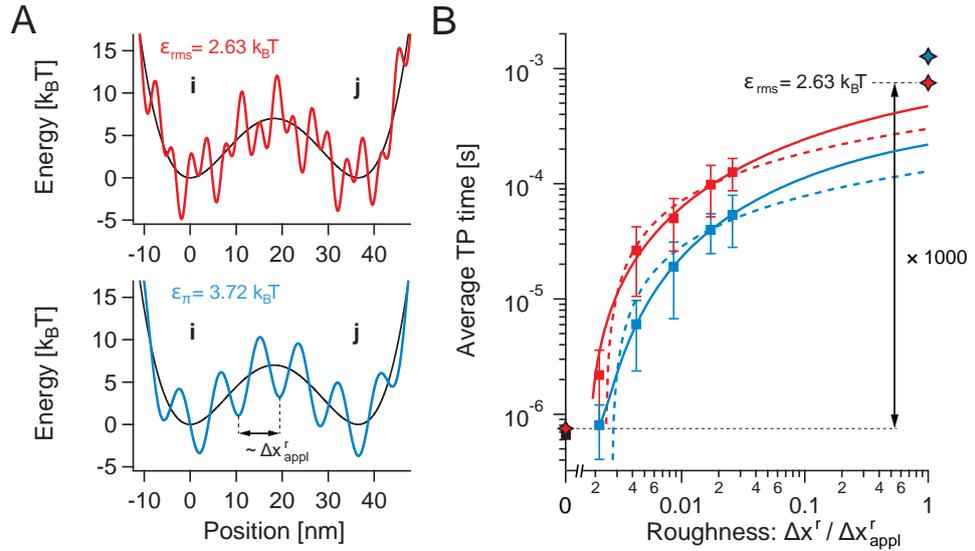


Figure 23: A thousandfold increase in transition path times induced by energy landscape roughness. Simulation parameters are  $T = 300 \text{ K}$ ,  $\gamma = 4 \times 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , and  $\Delta t = 10^{-9} - 10^{-11} \text{ s}$ . (A) A quasi-random (red) and a simple periodic (blue) roughness superimposed on a smooth energy landscape (black). Both types have a root-mean-squared roughness equal to  $2.63 k_B T$  and a typical length scale  $\Delta x_{\text{appl}}^r \sim 10 \text{ nm}$ . (B) Transition path times from diffusion in energy landscapes with different degrees of random (red squares) and simple periodic (blue squares) roughness scaled according to  $\Delta x^r / \Delta x_{\text{appl}}^r$  with  $\Delta x^r = 0.1 \text{ nm}$ . Error bars represent the standard deviation. Red stars at zero and 'full' roughness are predictions based on Equation 54 in combination with Equation 63. The black square is the simulation result for the smooth energy landscape (see Section 4.2.1). The blue star is predicted based on  $D^* = D \cdot \exp(-2\epsilon_{\pi}/k_B T)$ ,  $\epsilon_{\pi} = 3.72 k_B T$ . Thick lines are logarithmic fits of  $x$  (solid) and  $1/x$  (dashed) to guide the eye.

Since hopping between the many minima of a rough energy landscape involves many additional dwell times along the reaction pathway, it appears to be straightforward that the effective diffusion in a rough potential slows down as implied by Equation 63. However, this is not obvious right away. In fact, if the time needed to cross over a rough energy barrier is considered to be composed of the sum of all individual dwell and transition path times inferred by the roughness, one arrives at a paradox: by introducing more and more dwells and barriers, the mere sum of all dwell and transition path times approaches zero. This follows directly from the dependence of dwell and transition path times on the inverse of the curvatures of the underlying energy landscape (see the inverse of Kramers rate equation, Equation 29, for the expected dwell time average and Equation 54 for  $\langle \tau_{\text{TP}} \rangle$ ). The curvatures in turn are proportional to the inverse of the squared distance of the transition state  $(\Delta x^{\text{TS}})^2$  to which dwell and transition path times thus directly relate (see Figure 19C in Section 3.8.4). If keeping this in mind, doubling the amount of dwells within a fixed overall distance automatically divides all  $\Delta x^{\text{TS}}$  in

half which then leads to all dwell and transition path times being multiplied with an extra one half squared (barrier heights are kept constant). By this means, the expected effect of roughness has to arise rather from revisiting the same dwells for multiple times than from merely summing up all dwell and transition path times along the reaction pathway.

Assuming that Equation 63 is true means that a root-mean-squared roughness of  $\epsilon_{\text{rms}} = 2.63 k_{\text{B}}T$  leads to a thousandfold decrease of the effective diffusion coefficient. In contrast, as a consequence of its inverse dependence on the diffusion coefficient, the average transition path time  $\langle\tau_{\text{TP}}\rangle$  is expected to increase accordingly, see Equation 54. The simulation results shown in Figure 23B prove this conclusion as well as Equation 63 to be right. By decreasing the typical length scale  $\Delta x_{\text{appl}}^{\text{r}}$  of the applied roughness to approach a value of  $\Delta x^{\text{r}} = 0.1 \text{ nm}$  necessary for Equation 63 to be valid, the predicted increase of  $\langle\tau_{\text{TP}}\rangle$  over several orders of magnitude is reproduced<sup>5</sup>. Furthermore the importance of a good separation of length scales for diffusion versus the overall motion between the indicated states  $i$  and  $j$  is well illustrated. In the end, this separation effectively hides the global barrier shape away from the diffusing particle such that vanishing dynamical path probabilities are bygone and longer and longer overall transition path times become possible. The shorter transition path times observed for the simple periodic roughness confirm this picture. Here, an even smaller  $\Delta x^{\text{r}}$  would be necessary to satisfy the theoretic prediction. In return, the larger variety of energy well depths and barrier heights in the potential with a quasi-random roughness obviously helps to hide the global barrier shape away. In principle, the high-resolution simulation data provides all necessary information to further analyse and quantify the interplay between smaller and larger energy well depths with respect to revisiting frequencies of individual minima and the resulting change of the effective diffusion coefficient.

Hence, an ideal general roughness is comprised of many potential minima of varying depth where the frequency of revisiting deeper wells is promoted by more shallow wells around them. If so, it seems legitimate to alternatively consider roughness to be composed of only a few and, hence, experimentally easier to access high-energy intermediates which represent locally merged deeper wells and their shallow promoters. Starting from this simplified view on roughness, it is possible to extract key parameters of folding energy landscapes from transition path times as will be shown in Section 4.2.3. Moreover, this view allows to overcome two limitations of the general overall roughness picture introduced by Zwanzig, namely the restriction to one dimension as well as the assumption of an overall constant diffusion coefficient. While off-pathway intermediates allow the introduction of multidimensionality, position-dependent well depths of the high-energy intermediates can compensate for a position-dependent diffusion coefficient.

<sup>5</sup> Simulations further approaching  $\Delta x_{\text{appl}}^{\text{r}} = \Delta x^{\text{r}} = 0.1 \text{ nm}$  were not performed due to high numeric cost.

### 4.2.3 Energy landscape reconstruction from transition path times reflecting local roughness

When additional energy minima are located along a transition path connecting two stable states, the dwell times emerging from passing through those minima contribute a certain amount of time to the detectable overall transition path time necessary to complete a transition between the two stable states. In case of only a few but deep additional minima, the distribution of collected transition path times basically reflects a combined dwell time distribution. This can be exploited to estimate the barrier heights determinant for the dwell times if complementary information on the positions of relevant minima and transition states is available.

A simple example of an additional energy minimum is depicted in [Figure 24A](#) where all transitions between states  $i$  and  $m$  pass through an intermediate  $j$ . The dwell time  $\tau_{\text{DW}}^j$  spent in the intermediate state corresponds to the inverse off-rate  $(k_{\text{off}}^j)^{-1} = (k^{ji} + k^{jm})^{-1}$ , which due to the symmetry in the shown example can be alternatively written as  $(k_{\text{off}}^j)^{-1} = (2 \cdot k^{ji})^{-1} = (2 \cdot k^{jm})^{-1}$ . The inverse of Kramers rate equation, [Equation 29](#), then yields  $\tau_{\text{DW}}^j = 16.7 \mu\text{s}$  based on the curvatures provided by the detailed energy landscape in [Figure 24A](#).

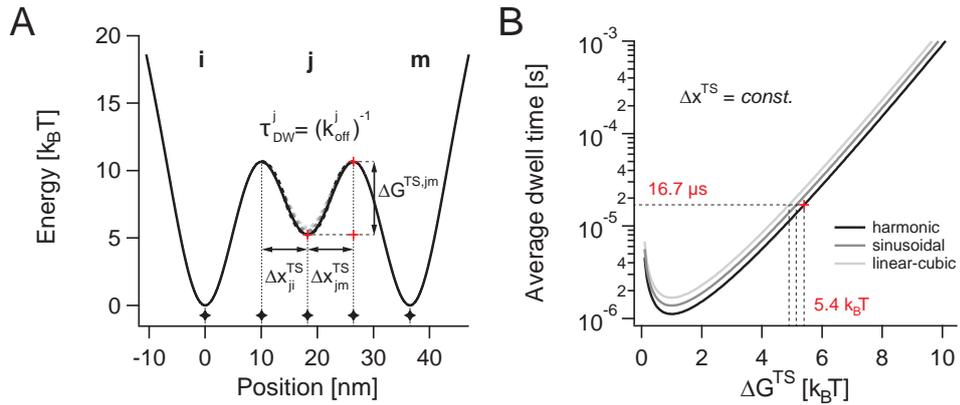


Figure 24: Barrier height dependence of dwell times for a given transition state position. Calculation parameters are  $T = 300 \text{ K}$  and  $\gamma = 4 \times 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . (A) Energy landscape with three local minima representing states  $i$ ,  $j$ , and  $m$ . Thick dashed lines are reconstructions for  $\tau_{\text{DW}}^j = 16.7 \mu\text{s}$ . (B) Barrier height dependence of  $\tau_{\text{DW}}^j$  calculated from the inverse of Kramers rate equation, [Equation 29](#), for a transition state position fixed at  $\Delta x^{\text{TS}} = \Delta x_{jm}^{\text{TS}}$  as marked by black stars in (A). The barrier shapes are calculated for either harmonic, sinusoidal, or linear-cubic potentials.

Now, what if only the average dwell time  $\tau_{\text{DW}}^j$  and the transition state positions  $\Delta x^{\text{TS}}$  with respect to the intermediate  $j$  were known? In this case, the determinant barrier height  $\Delta G^{\text{TS}}$  can still be estimated by making use of the fact that the curvatures of an energy landscape  $G(x)$  can be expressed in mere terms of  $\Delta x^{\text{TS}}$  and  $\Delta G^{\text{TS}}$  as indicated by [Equation 55](#). By this means and

for constant temperature  $T$  and friction coefficient  $\gamma$ , Kramers rate equation can be rewritten in mere terms of  $\Delta x^{\text{TS}}$  and  $\Delta G^{\text{TS}}$ , too. [Figure 24B](#) shows the expected average dwell time  $\tau_{\text{DW}}^j$  as a function of the barrier height for a symmetric barrier<sup>6</sup> with a fixed transition state position. Within the high-barrier limit such a plot allows a numerical estimate for an unknown barrier height to be directly read off as the value corresponding to the known dwell time.

For the theoretically expected dwell time of  $16.7 \mu\text{s}$  in the case of the energy landscape depicted in [Figure 24A](#), the plot shown in [Figure 24B](#) indicates a harmonic barrier height of  $\Delta G_{\text{harm}}^{\text{TS},j^{\text{m}}} = 5.38 k_{\text{B}}T$ . As the original value is  $\Delta G^{\text{TS},j^{\text{m}}} = 5.41 k_{\text{B}}T$ , the thick dashed black line in [Figure 24A](#), which shows the reconstructed harmonic energy well of the intermediate  $j$ , almost perfectly overlaps with the original landscape. Owing to the strong barrier height dependence,  $\tau_{\text{DW}} \propto \exp(\Delta G^{\text{TS}})/\Delta G^{\text{TS}}$ , even a rough estimate of an average dwell time will already yield a good guess for the determinant barrier height. If, for example, the estimated average dwell time is wrong by a factor of 2, the derived energy barrier height is wrong by not more than  $0.7 k_{\text{B}}T$  as  $\Delta G^{\text{TS}} \propto \ln(\tau_{\text{DW}})$  for high enough barriers.

#### 4.2.3.1 Testing reconstruction capabilities on simulated time trajectories

[Figure 25](#) illustrates how the above outlined method of a dwell-time-based energy landscape reconstruction works under various conditions. The left column shows the potentials used to perform one-dimensional diffusion simulations. While [Figure 25A](#) shows a three-state potential<sup>7</sup> with, in terms of their energy level, two evenly high, and, in terms of their position, equidistant transition states, in case of [Figure 25D](#) a three-state potential with one predominant transition state is depicted. The four-state potential shown in [Figure 25G](#) has two additional minima located before and after the main transition state.

The central column in [Figure 25](#) displays distributions of transition path times extracted for transitions<sup>8</sup> between the two outermost minima of the respective potentials on the left all of which include diffusion through interjacent minima. Since these additional minima are quite deep, their dwell times add up to the overall transition path times as their main contribution. By this means, each transition path time distribution can be alternatively interpreted as a dwell time distribution. As such, the dwell time distributions can be fit to extract the corresponding average dwell times. As indicated in each graph, all fit results yield dwell times  $\tau^{\text{fit}}$  which are in very good agreement with theoretical predictions. In [Figure 25B](#) and [Figure 25E](#) single exponential fits perfectly match the overall transition path time distributions which confirms their interpretation as dwell time distributions of the corresponding single additional minimum along the overall transition path. In case of the distribution

<sup>6</sup> Here, for simplification  $\omega_i = \omega_{\text{TS}}$  is assumed in [Equation 29](#).

<sup>7</sup> The potentials in [Figure 24A](#) and [Figure 25A](#) are the same.

<sup>8</sup> The distributions in [Figure 25B](#), [Figure 25E](#), and [Figure 25H](#) are based on  $N = 898$ ,  $N = 2242$ , and  $N = 1257$  transitions, respectively.

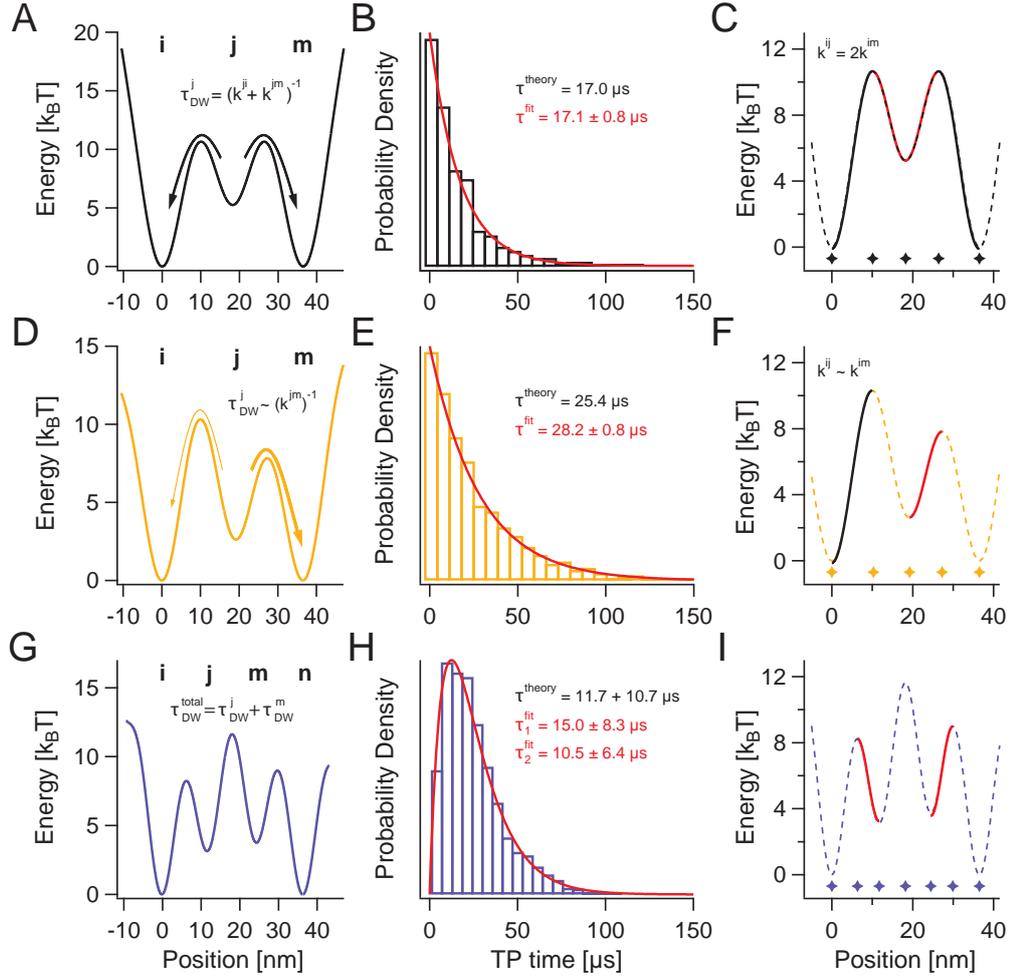


Figure 25: Energy landscape reconstruction from transition path times reflecting local roughness. Simulation parameters:  $T = 300 \text{ K}$ ,  $\gamma = 4 \times 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , and  $\Delta t = 10^{-8} \text{ s}$ . (A), (D), (G) Potentials  $G(x)$  underlying the simulation. (B), (E), (H) Red lines are fits to extracted transition path time distributions which reflect high-energy intermediate dwell times. (C), (F), (I) Partially reconstructed energy landscape profiles. See text for more details.

shown in Figure 25B, the indicated theoretic value  $\tau^{\text{theory}} = 17.0 \mu\text{s}$  comprises a dwell time contribution of  $\tau_{\text{DW}}^j = 16.7 \mu\text{s}$ , as derived above, and a mere transition path time contribution of  $\langle \tau_{\text{TP}}^{i,m} \rangle = 0.3 \mu\text{s}$  for crossing all barriers between states  $i$  and  $m$ . The latter value is derived by applying Equation 54 and summing up all included path segments. The dwell time distribution obtained in Figure 25H is not single exponential due to two substantial additional minima contributing to the overall transition path time. Dwell time distributions exhibiting two determinant rates can be fit according to [184]:

$$p(\tau_{1+2}) = \frac{k_1 \cdot k_2}{k_1 - k_2} \cdot (e^{-k_2 \cdot \tau_{1+2}} - e^{-k_1 \cdot \tau_{1+2}}). \quad (64)$$

Here,  $k_1$  and  $k_2$  represent off-rates out of the two intermediate states of the potential drawn in Figure 25G while  $\tau_{1+2}$  reflects their combined dwell time.

To assure that Equation 64 is applicable, the central out of the three relevant transition states has to be either first or second in the hierarchy. This way the probability to pass through an intermediate for multiple times during a transition is minimized. The excellent agreement between the measured distribution and its fit in Figure 25H again confirms the overall picture of additional dwell times being the main contribution to the extracted transition path times. Note that even though all obtained transition path time averages show quite large standard deviation errors due to the non-Gaussian nature of all three transition path time distributions, they are still in very good agreement with the theoretic values: for Figure 25B, Figure 25E, and Figure 25H the average values are  $18.7 \pm 19.7 \mu\text{s}$ ,  $27.7 \pm 26.4 \mu\text{s}$ , and  $25.0 \pm 17.7 \mu\text{s}$ , respectively.

Finally, the resulting reconstructed energy landscape profiles are shown in the right column of Figure 25. Red barriers are derived following the above described dwell-time-based reconstruction method using the fit results obtained from the extracted transition path time distributions. Black barriers are reconstructions based on the observed effective overall transition rates between the outer minima. In combination with Kramers rate equation expressed in terms of  $\Delta x^{\text{TS}}$  and  $\Delta G^{\text{TS}}$ , the overall rates can reveal at least one barrier height in case of an outer minimum being located directly next to the predominant transition state. For all reconstructions it is assumed that all important landmarks of the underlying energy landscape, i. e., all positions of relevant minima and transition states, are known. The same holds for the relative height of the transition states. The claim of  $k^{ij} = 2 \cdot k^{im}$  in Figure 25C directly follows from symmetry: a particle reaching intermediate  $j$  starting from state  $i$  has a fifty percent chance of reaching back to state  $i$  and a fifty percent chance of going further to state  $m$ . In Figure 25F the situation is different. Here, a particle reaching intermediate  $j$  starting from state  $i$  is most likely proceeding towards state  $m$ . Note that except for the completely symmetric energy landscape in Figure 25C, the relative heights between the minima are not exactly known. To better visualize the good reconstruction capabilities, the reconstructed barriers were intentionally aligned with the transition states of the original energy landscapes shown as dashed lines. Taken together, the presented simulation results confirm the robustness of the introduced energy landscape reconstruction based on dwell time estimates.

#### 4.2.3.2 Road map towards an application to experimental data

Good knowledge of all relevant landmark positions  $\Delta x_{**}^{\text{TS}}$  is an important prerequisite for the dwell-time-based reconstruction of localised energy landscape roughness along a transition path. Further, the relative heights among the transition states are of interest, too. Finally, an appropriate choice for the friction coefficient  $\gamma$  and, hence, the diffusion coefficient  $D$ , in combination with meaningful curvatures  $\omega_*$  is necessary.

To localise all populated states and intermediates, histograms of contour length gains obtained from WLC fits to force-extension traces are good first indications. Under certain circumstances, accumulated histograms of entire transitions may yield an improved resolution especially in case of short-lived

and rarely populated intermediates, also see [Section 3.1.1](#). Histograms can be accumulated either over an entire force range or for a constant force level. In the latter case, deconvolution can be helpful to further increase resolution. Deconvolution is also the method of choice to unravel intermediate states from entire equilibrium trajectories, see [Section 3.7](#), provided an overlap of multiple pathways can be excluded.

Transition state positions are typically provided either by fits to force distribution histograms or by kinetic model fits to extracted force-dependent transition rates, see [Section 3.4](#), [Section 3.5](#), and [Section 3.6](#). In case of transition rates involving the crossing over multiple transition states, striking deviations from kinetic models are expected over larger force ranges due to the force-induced changes of the transition states' predominance as outlined in [Section 4.2.4](#). A correct interpretation of these changes allows a good guess for the positions of the different involved transition states to be made. Here, the force-dependent determination of transition path probabilities may also be of interest for localizing the predominant transition state position, see [Section 3.8.3](#) and [Section 4.3.4](#).

Concerning the curvatures at the bottom of energy wells and at the top of energy barriers, the simplest way of their determination is through direct deconvolution of equilibrium trajectories [154]. If deconvolution is not applicable due to the multidimensionality of the underlying energy landscape or poor resolution, it may yet provide useful information on curvatures of predominantly populated states. As a good first approximation, the curvature at a suspected transition state position may simply be chosen to be the same as the curvature of a neighbouring state [44, 247]. For a further refinement, the curvatures of all other states and barriers may be calculated based on a linear interpolation with respect to their relative positions, which, in principle, corresponds to the concept of the linearly interpolated point spread function used for improved deconvolution results, see [Section 3.7.1](#).

The determination of an appropriate diffusion coefficient  $D$  can be viewed as the calculation of a correction term. In case of the entire folding energy landscape shape being already known by deconvolution,  $D$  is typically chosen such that experimentally observed force-dependent rates are in agreement with complementary diffusion simulations [143, 177]. This basically means that the diffusion coefficient is adapted to match Kramers rate-theory for already given curvatures and energy barrier heights. In such a scenario, a possible procedure is to numerically tilt the deconvolved energy landscape to zero force and then derive  $D$  inferred by [Equation 29](#) upon using the zero-force extrapolated folding rate provided by a kinetic model, see [Section 5.3.1](#). Further, if a protein's transition path time average is known, an alternative way to derive  $D$  is given by using [Equation 54](#) which directly relates transition path times to barrier heights and curvatures, see also [Section 6.4.1](#). However, if detailed information on the energy landscape shape is not available,  $D$  has to be adjusted to conform with force-dependent rates dominated by single barriers similar to the situation depicted for  $k_{im}$  in [Figure 25F](#), as well as ob-

served transition state switches discussed in [Section 4.2.4](#), and detectable free energy differences according to [Section 3.2](#).

The application of the transition-path-assisted reconstruction of locally concentrated energy landscape roughness to experimental data holds an interesting advantage over the analysis of entire equilibrium fluctuations. By focussing on transition paths, the analysis is reduced down to ‘productive’ events while fading out all potentially confusing side reactions. This way, yet barely noticeable on-pathway intermediates may be identified and eventually even further characterized up to unprecedented resolution. A detailed transition-path-assisted energy landscape reconstruction based on experimental data is presented in [Section 6.4.4](#).

#### 4.2.4 *A fingerprint for roughness: predominant transition state switching induced by force*

The slope of a rate plot directly relates to the transition state position  $\Delta x$  of the observed reaction, see [Section 3.5](#). Hence, a change in slope within a rate plot implies a reaction coordinate-dependent change of the transition state position. The occurrence of drastic changes in slope, i. e., kinked rate plots, are typically explained by either switches between different pathways or by switching of the predominant transition state along a reaction pathway with multiple sequential barriers [[131](#), [191](#)]. In the following, the scenario of a transition state switch inferred by two sequential barriers flanking an on-pathway high-energy intermediate is investigated.

[Figure 26A](#) depicts a simple symmetric two-state energy landscape  $G_0(x)$  (thick black line) and linearly tilted versions of it calculated according to  $G_{F^*}(x) = G_0(x) - F^* \cdot x$ . The constant  $F^*$  is given in units of  $k_B T \cdot \text{nm}^{-1}$  and denotes the direction and extent of tilt by sign and size. Within each of the shown potentials, 5 minute long diffusion trajectories are simulated. From each trajectory transition rates  $k_{\text{forward}}$  and  $k_{\text{backward}}$  are extracted using [HMM](#) analysis, see [Section 3.4.3](#). The results are plotted as empty circles in the rate versus tilting force  $F^*$  plot shown in [Figure 26B](#). Drawn in dashed and solid lines are rate predictions by the ZB- and DHS-model based on [Equation 27](#) and [Equation 31](#), respectively. Necessary model parameters are derived from  $G_0(x)$ , namely  $\Delta x = 18.5 \text{ nm}$ ,  $\Delta G_0^{\text{TS}} = 7 k_B T$ , and  $\nu = 2/3$ . The average value of  $k_{\text{forward}}$  and  $k_{\text{backward}}$  at zero tilting force  $F^*$  is chosen as (pseudo) zero-force rate constant  $k_0 = 6.9 \text{ s}^{-1}$  which is in excellent agreement with Kramers theory that predicts  $k_0 = 7.1 \text{ s}^{-1}$  according to [Equation 30](#). In contrast to [Figure 26A](#), [Figure 26C](#) shows a symmetric energy landscape with two sequential barriers flanking an additional high-energy intermediate (thick black line). In the following, this landscape will be referred to as complex two-state rather than three-state in accordance with [[131](#)]. [Figure 26D](#) presents the overall transition rates extracted from simulations within the potentials of [Figure 26C](#). Here, dashed and solid lines are rate predictions by the ZB- and DHS-model calculated under the assumption of an effective transition state right in the middle between the two outer states. The respective model parameters are

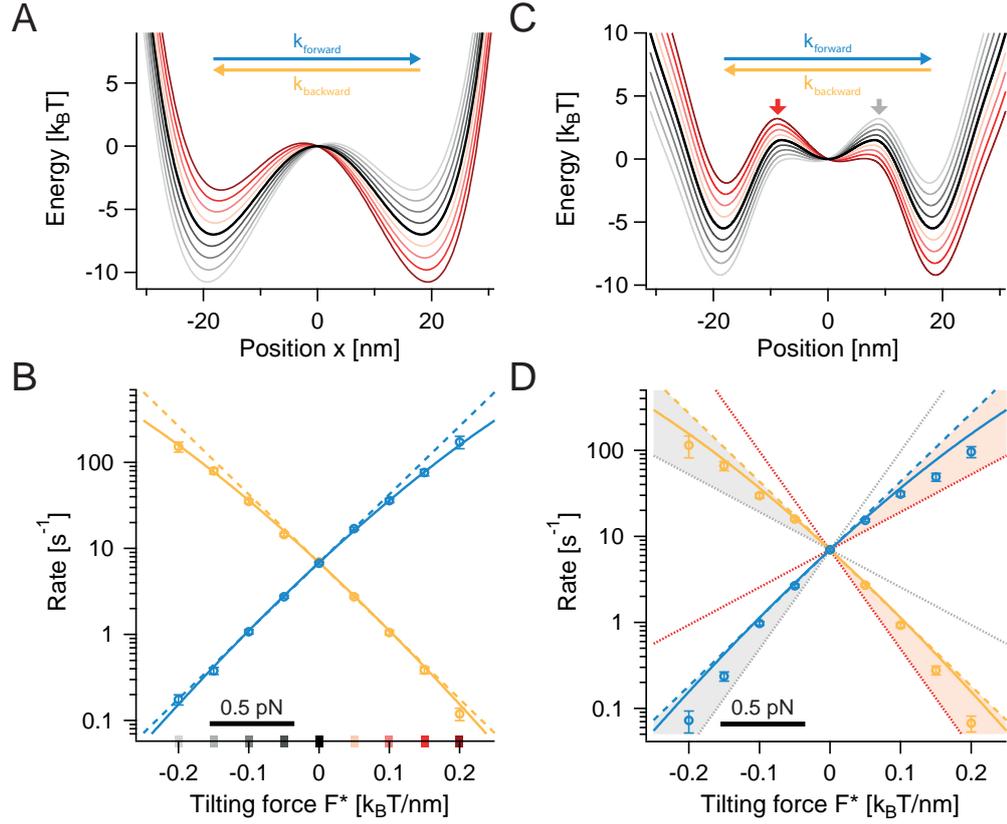


Figure 26: Force induced transition state switching as a fingerprint for roughness. Simulation parameters are  $T = 300$  K,  $\gamma = 10^{-5}$  pN  $\cdot$  s  $\cdot$  nm $^{-1}$ ,  $\Delta t = 10^{-6}$  s, and  $\Delta t_{\text{record}} = 5 \times 10^{-6}$  s. (A) Simple symmetric two-state energy landscape  $G_0(x)$  (thick black line) and tilted versions of it. (B) Transition rates versus tilting force  $F^*$  (empty circles) obtained from 5 min diffusion trajectories in the potentials shown in (A) using HMM analysis. Colours refer to (A). Dashed and solid lines are rate model predictions based on  $G_0(x)$ . (C) Complex symmetric two-state energy landscape and tilted versions of it. Short arrows highlight the predominant transition state switch. (D) Overall transition rates versus tilting force. See text for more details.

$\Delta x = 18.25$  nm,  $\Delta G_0^{\text{TS}} = 7 k_B T$ ,  $\nu = 2/3$ , and  $k_0 = 7.0 \text{ s}^{-1}$ . Additional dotted lines are ZB-model predictions reflecting the expected rate dependence in case only one of the two sequential barriers existed. Red dotted lines consider the left barrier marked by a short red arrow in Figure 26C and use  $\Delta x_{\text{forward}} = 10.1$  nm and  $\Delta x_{\text{backward}} = 26.4$  nm, where necessarily  $\Delta x_{\text{forward}} + \Delta x_{\text{backward}} = 2 \cdot \Delta x$ . Gray dotted lines consider the opposite configuration.

The rate plots presented in Figure 26 disclose a number of important results. As expected, in case of zero tilting force there is  $k_{\text{forward}} = k_{\text{backward}}$  due to the symmetry of both simple and complex two-state energy landscapes. Even though being a trivial result, this is another rudimentary proof of the simulation's integrity. Furthermore, the transition rates in Figure 26B obtained upon linearly tilting a simple two-state energy landscape feature a moderately curved shape in accord with expected Hammond behaviour [94]. The

causative transition state movement inferred by tilting  $G_0(x)$  can be seen in [Figure 26A](#); for a more elaborate illustration of the Hammond effect see [78], for an explanation directly referring to the force-spectroscopic framework see [254]. Since the DHS-model (solid lines) accounts for Hammond behaviour, see [Section 3.5.4](#), its prediction clearly outperforms the simpler ZB-model (dashed lines). This section's key result, however, is highlighted within the red and gray shaded areas of [Figure 26D](#). The more the complex two-state energy landscape is being tilted away from its symmetric configuration, the higher the deviation of the extracted overall transition rates from simple two-state mechanics. This deviation is caused by a switch in predominant transition state which is highlighted by short arrows in [Figure 26C](#) and manifesting itself as a prominent kink within the rate plots shown in [Figure 26D](#). The kink is located at zero tilting force where the two sequential barriers equally contribute to the overall transition. Within the force range covered by the simulations one can see that in both directions of tilt the three outermost extracted rates are parallel to the ZB-model predictions which assume only one of the barriers to be predominant (dotted lines). In summary, the conducted simulations prove that sequential barriers along a reaction pathway can indeed explain kinked rate plots.

Since force-induced pathway switches as well as switching of the predominant transition state have already been suggested for protein folding [83, 197], the results from this section are of high interest. In [Chapter 5](#), experimental results obtained from force-spectroscopic experiments performed on a small artificial protein give experimental evidence for Hammond behaviour to be relevant for protein folding, see [Section 5.3.1](#). Further, the data also suggest a force-induced transition state switch due to a prominent kink within the force-dependent rates, see [Section 5.2.4](#).

### 4.3 ENERGETICALLY COUPLED TWO-DIMENSIONAL DIFFUSION

The simulations performed in this section are based on [Equation 62](#) which allows time traces conforming with experimental measurement conditions to be deduced from traces with much higher resolution. By this means, the simulations are intended to serve as numerical benchmark experiments. After an initial consistency check, the impact of the difference in friction inherent to the actual protein folding process with respect to the friction involved in the detecting system is analysed. Thereafter, detectable traces of protein folding within transition averages are sought. Finally, the transition path analysis tools introduced in [Section 3.8](#) are tested for their applicability and potential benefit in force spectroscopy.

#### 4.3.1 Deconvolution as a link between dimensions

As described in [Section 3.7](#), deconvolution can serve as a tool to reconstruct protein folding energy landscapes out of the equilibrium fluctuations measured in a force-spectroscopic experiment. Here, deconvolution is used as a

consistency check of the two-dimensional diffusion simulation introduced in [Section 4.1](#) by testing whether the zero-force protein folding energy landscape serving as the simulation's input can be reconstructed from the simulated measurement signal.

In [Figure 27](#) the energetically coupled protein and bead diffusion trajectories provided by a 2 s long two-dimensional diffusion simulation are analysed in matters of the energy landscapes derived from the corresponding probability distributions. Each plot in [Figure 27](#) includes a red dashed line which is basically a projected average of the two-dimensional energy function  $H(x, L_p)$  defined in [Equation 61](#) on one of its two axes. While  $x$  reflects the bead deflection signal which is accessible to the experimentalist,  $L_p$  holds the information to which extent the protein under investigation is actually folded in measures of unfolded contour length. The black lines in [Figure 27A](#) and [Figure 27B](#) show the Boltzmann inverted probability distributions<sup>9</sup> of the bead and protein diffusion trajectories. [Figure 27C](#) presents the protein folding energy landscape (dark blue line) resulting from deconvolving the probability distribution of the bead trajectory by minimizing [Equation 42](#). This protein folding energy landscape is then further transformed into contour space as shown in [Figure 27D](#), also see [Section 3.1.1](#). Finally, the deconvolved and transformed protein folding energy landscape is further transformed to zero acting force based on [Equation 13](#). The result is depicted by the light blue line in [Figure 27E](#). The black dashed line illustrates  $G_0(L_p)$  which serves as input for the simulation, the dark blue and the red dashed line in [Figure 27E](#) are the same energy landscapes as in [Figure 27D](#). The insets in [Figure 27](#) illustrate a direct (single arrow) or indirect (dashed line plus arrow) origin of the probability distributions which are used to derive the depicted energy landscapes.

Red dashed lines in [Figure 27](#) reflect the energy landscape shape expected for the equilibrium described by  $H(x, L_p)$  along the indicated reaction coordinate. The expectations are perfectly met by the Boltzmann inverted equilibrium probability distributions of the simulated bead and protein diffusion signals shown in [Figure 27A](#) and [Figure 27B](#). Though already having the correct barrier height, the deconvolved protein energy landscape presented in [Figure 27C](#) has a striking asymmetric shape. Generally speaking, this shape is a result of the interplay between non-linear effects inherent to the measurement system, a non-constant acting force, as well as a position-dependent point-spread function. However, after a contour length transformation, the deconvolved energy landscape perfectly complies with the expected shape as can be seen in [Figure 27D](#). This, in turn, confirms the asymmetric shape in [Figure 27C](#) to be correct. Convincing proof for the two-dimensional simulation to be fully consistent is given in [Figure 27E](#). Here, the deconvolved protein folding energy landscape is transformed back to zero force (light blue line) reproducing the zero-force energy landscape input  $G_0(L_p)$  of the simulation (black dashed line) very well.

<sup>9</sup> Here, Boltzmann inversion means solving [Equation 17](#) for  $H(x)$ .

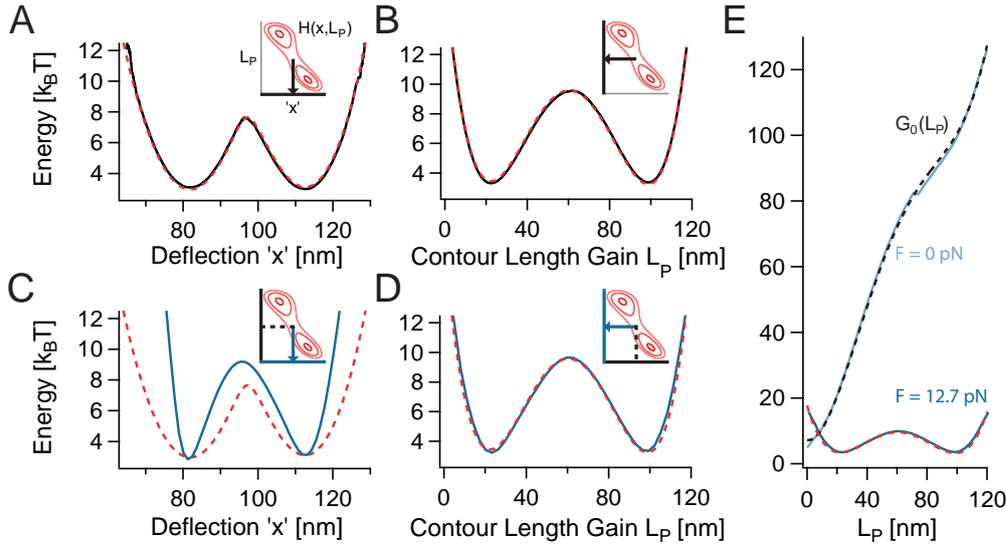


Figure 27: Boltzmann inversion and deconvolution confirm an energetically consistent two-dimensional diffusion simulation. Simulation parameters are  $T = 300$  K,  $\gamma_B = 9.4 \cdot 10^{-6}$  pN  $\cdot$  s  $\cdot$  nm $^{-1}$ ,  $\gamma_P = 4 \cdot 10^{-8}$  pN  $\cdot$  s  $\cdot$  nm $^{-1}$ , and  $\Delta t = 10^{-9}$  s. (A) - (E) Red dashed lines are projections of  $H(x, L_p)$  on the respective coordinate. Calculation parameters: inter-trap distance  $d = 494.4$  nm,  $k_{\text{eff}} = 0.11$  pN/nm,  $p_P = 0.7$  nm,  $L_D \sim 370$  nm,  $p_D \sim 50$  nm,  $K \sim 200$  pN, and  $T = 300$  K. (A) Boltzmann inverted probability distribution (black) of the bead deflection signal. (B) Boltzmann inversion (black) of the protein diffusion signal. (C) Protein folding energy landscape (blue) deconvolved from the bead deflection signal. (D) Contour length transformation (blue) of the deconvolved protein folding landscape shown in (C). (E) Zero-force transformation (light blue) of the protein folding energy landscape (blue) and the input  $G_0(L_p)$  (black dashed line) used in the simulation.

The correct calculation of  $H(x, L_p)$  basically implies a convolution of all relevant signals contributing to the modelled force-spectroscopic experiment. Hence, from a different perspective, the simulation-based analysis illustrated in Figure 27 reveals excellent performance of the method applied for deconvolution. In this context, however, it needs to be remarked that even though the correct point-spread function is exactly known due to  $H(x, L_p)$ , the deconvolution result is quite sensitive to an appropriate choice of optimization parameters including the number of interpolation points as well as the prefactor  $\lambda$  of the penalty function<sup>10</sup>, see Section 3.7 and Equation 42.

Note that the friction coefficients  $\gamma_B$  and  $\gamma_P$  have no effect on  $H(x, L_p)$ , see Equation 61. Hence, the shape of the equilibrium probability distributions obtained from simulated trajectories will not be affected by any change in friction as long as the trajectories are long enough to sample equilibrium. As a consequence, energy landscapes obtained by deconvolution are insensitive towards changes in friction.

<sup>10</sup> The deconvolved energy landscape shown in Figure 27C was obtained by using 11 base points and  $\lambda = 10^{-2}$  nm $^2$ .

#### 4.3.2 *The impact of anisotropic friction on rates, transition path times, and transition path ensembles*

Since there is no way to directly measure the internal friction of a protein, its determination is a very challenging task for experimentalists. In this section, exact knowledge of the friction involved in a simulated protein diffusion process is utilized to investigate effects arising from the difference in friction between actual protein folding and the system used to detect the protein folding transitions. After starting off with example traces for speed limit protein folding under force, an extensive study of transition rates and transition path times in dependence of the friction coefficient  $\gamma_P$  is presented. Further, two-dimensional representations of transition path ensembles give insight into a transition process altered by friction.

##### 4.3.2.1 *'Watching' speed limit protein folding under force*

The friction coefficient of  $\gamma_P = 4 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  which is used to characterize protein diffusion in [Section 4.3.1](#) conforms with the highest reported values for intrachain diffusion in unfolded proteins, see [Section 4.1](#). Hence, this choice of  $\gamma_P$  can be expected to reflect protein folding at the speed limit. An example of a protein unfolding and refolding process at the speed limit while the protein is being held under tension is extracted from the simulation performed in [Section 4.3.1](#) and presented in [Figure 28](#). The folding transitions are shown at full resolution, i. e., a sampling frequency of 1 GHz which allows detailed insight into the transition paths. In case of the protein diffusion trace, transition paths are marked in blue. The same stretches are also marked within the bead diffusion trace. Coloured dashed lines in [Figure 28](#) indicate the location of the minima of both folded and unfolded states while black dashed lines refer to the respective transition state position.

An eye-catching result provided by [Figure 28](#) is the much higher noise level within the protein diffusion trajectory as compared to the bead trajectory. This is a direct consequence of the over 200 times smaller friction coefficient  $\gamma_P$  with respect to  $\gamma_B$  which results in larger steps of  $\Delta L_P$  per iteration of the diffusion simulation, see [Equation 62](#)<sup>11</sup>. However, even though the friction coefficient  $\gamma_P$  is very small, the transition paths (marked in blue) of protein diffusion last way too long to reflect folding at the speed limit. Given the energy landscape shown on the upper right of [Figure 28](#), a transition path time of about 3  $\mu\text{s}$  is expected<sup>12</sup> according to [Equation 54](#). Being a few tens of microseconds long, the transition paths of protein diffusion provided by the simulation clearly stand in conflict with this expectation. Another striking observation are multiple distinct barrier crossings within a single transition

<sup>11</sup> Note that the local slopes of the energy landscapes characterizing protein and bead diffusion, i. e.,  $\partial_{L_P} H(x, L_P)$  and  $\partial_x H(x, L_P)$  in [Equation 62](#), are within the same order of magnitude.

<sup>12</sup> The protein folding energy landscape displays  $\Delta L_P = 76 \text{ nm}$  upon a transition. This corresponds to  $\sim 200$  amino acids being involved in a folding process, see [Section 3.1](#) for details. Since the speed limit for folding  $N$  residues is estimated to be  $N/100 \mu\text{s}$ , see [\[119\]](#), here we reach at 2  $\mu\text{s}$ . The good match confirms a reasonable model system.

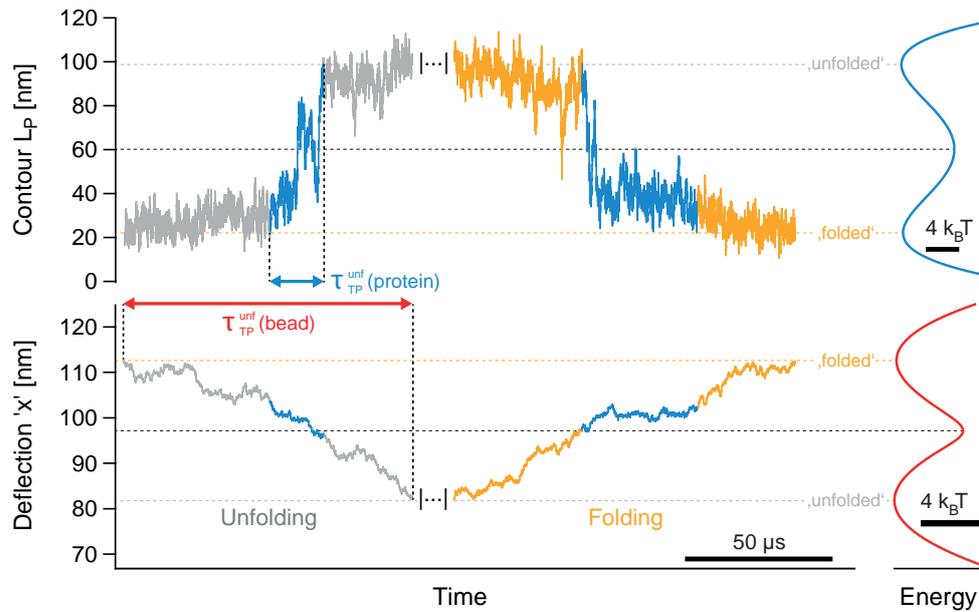


Figure 28: An example of speed limit protein folding under tension. Simulation parameters are the same as in Figure 27. Upper panel: Stretches of protein unfolding and refolding in contour space. The characteristic protein folding energy landscape is shown on the right. Transition paths are highlighted in blue. Lower panel: Simultaneous stretches of the bead diffusion trajectory. Blue stretches correspond to those in the upper panel.

path of protein diffusion. Compared to the protein trajectory, the bead signal complies much better with the expectations. Note that the shown transition of the bead trajectory represent entire transition paths interconnecting the folded and unfolded states. By this means, the transition path time detected by the beads is obviously much longer than the time needed by the protein as indicated by the double arrows in Figure 28. Further, note that the protein folding transitions seem to occur only during a phase where the bead diffuses around the transition state position of its convolved energy landscape while having a slight bias towards higher forces, i. e., towards the folded state. This observation suggests that in the given scenario the protein folding rate is under strong influence of bead diffusion.

The above critical look at Figure 28 suggests that the anisotropic friction involved in force-spectroscopic measurements has a crucial impact on rates and transition path times. In the following, a detailed analysis of friction-dependent kinetics aims at finding the detection limits inherent to force-spectroscopic systems.

#### 4.3.2.2 Quantitative analysis of protein and bead diffusion trajectories

Figure 29 presents the results of a detailed evaluation of transition path times and transition rates in dependence of the friction coefficient  $\gamma_P$ . While changing  $\gamma_P$  within the simulation, the energy function  $H(x, L_P)$  and the friction coefficient  $\gamma_B$  of the beads are kept constant. To provide good statistics at rea-

sonable numeric cost, the time steps of the simulation were increased along with  $\gamma_P$  up to a maximal value of  $\Delta t = 10^{-6}$  s. The overall duration of the simulated trajectories ranges from 2 up to 1000 s. Transition path time averages extracted from protein and bead diffusion trajectories are shown in [Figure 29A](#) as filled blue and red circles. Interconnecting lines are there to guide the eye. Dashed lines indicate transition path times expected according to [Equation 54](#) based on parameters of the corresponding energy landscapes shown on the right of [Figure 28](#). [Figure 29B](#) displays combined rates determined by counting successful transitions<sup>13</sup>, see [Section 4.2.1](#) for details. Here, dashed lines correspond to predictions by Kramers rate-theory, see [Equation 29](#). The four stars at the bottom of [Figure 29B](#) refer to values of  $\gamma_P$  which were derived from force-spectroscopic experiments. In detail, these values refer to folding of the native prion protein (I), DNA unzipping (II), coiled coil unzipping (III) [143, 177, 247], and folding of the re-designed Ferredoxin-like fold (IV). While the first three values are taken from literature, the last one is derived in this thesis, see [Section 5.3.1](#). The right of [Figure 29](#) depicts transition path ensembles obtained for three different values of  $\gamma_P$ . Here, the ensembles comprise unfolding as well as refolding transitions since no clear differences were seen upon distinction. Dashed lines indicate the expected transition path ensemble distributions. To calculate the expected distributions, at first committers are derived from the known energy profile by solving [Equation 51](#) for  $p_i(x)$ . After rescaling  $p_i(x)$  to its correct range, the transition path probability  $p(\text{TP}|x)$  can be derived according to [Equation 53](#). Finally, the transition path ensemble is given by [Equation 52](#) after solving for  $p(x|\text{TP})$  followed by renormalization.

There are two important questions to be raised for the discussion of the results from [Figure 29](#). First: Can the experimentalist see what happens? And second: Can this, what is happening, be modelled? The first question addresses the fact that the bead deflection signal reflects the experimentalist's measurement of the actual protein folding process. The second question puts the predictive capabilities of the applied theoretical models to test.

In [Figure 29A](#), the data points for speed limit folding confirm the observations already made in [Figure 28](#). The average transition path time for protein folding is an order of magnitude too long with respect to its prediction. In contrast to that, the average transition path time of the bead signal matches its prediction perfectly while being another order of magnitude longer than the actual protein transition path time average. Due to the much higher friction coefficient of the bead in case of speed limit folding, the bead effectively slows down the protein's transition path time. This situation is just the other way around in case where the protein experiences much higher friction than the bead, see the data points obtained for  $\gamma_P = 4 \cdot 10^{-4}$  pN · s · nm<sup>-1</sup>. Right in between, the intersection of the transition path time predictions is an outstanding 'sweet' spot marked with an empty black circle in [Figure 29A](#). Even though the friction of the bead is slightly higher at this intersection ( $\gamma_B =$

<sup>13</sup> Here, this technique can be successfully applied since  $H(x, L_P)$  is chosen such that the state levels of the bead trajectory are well separated to avoid counting thermal artefacts. Otherwise, HMM analysis should be preferred.

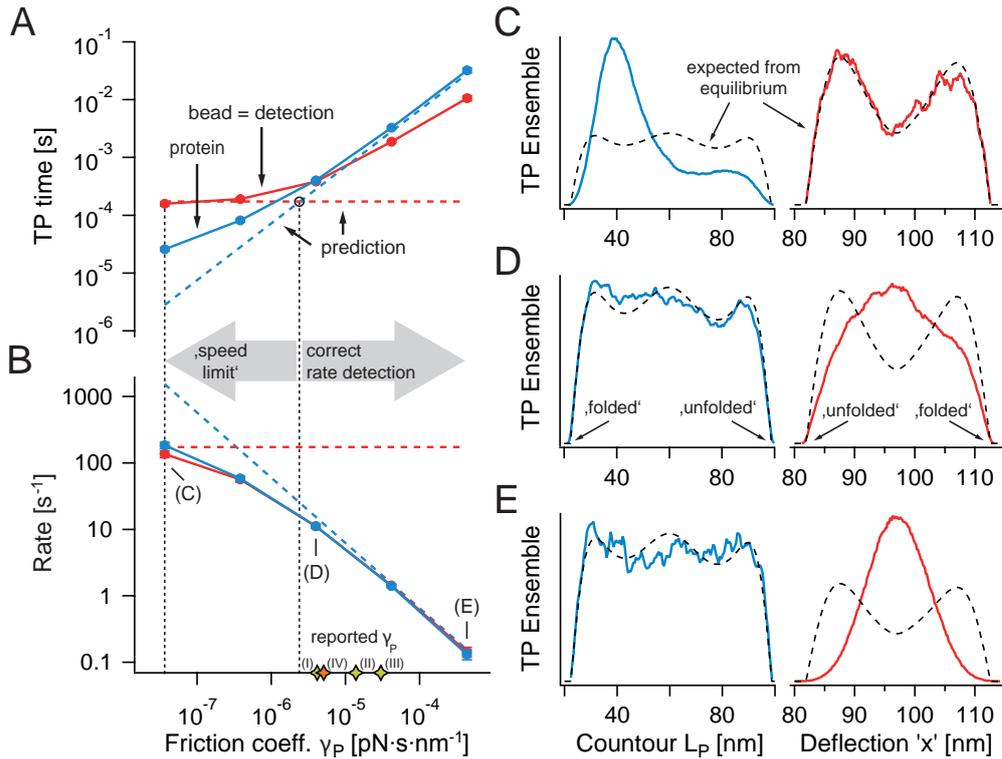


Figure 29: Friction-dependence of transition path times, rates, and path ensembles of protein folding probed by force. Simulation parameters are the same as in Figure 27 except for the indicated increase of  $\gamma_P$ . (A) Dependence of transition path times on  $\gamma_P$  while  $\gamma_B$  is kept constant. Connected circles are data provided by the simulation for protein (blue) and bead (red) diffusion. Dashed lines are predictions according to Equation 54. (B) Friction-dependence of rates. Dashed lines are predicted by Kramers theory, see Equation 29. (C) Transition path ensembles in contour and deflection space obtained for  $\gamma_P = 4 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . Dashed lines are ensembles expected from the equilibrium distributions. (D) and (E) Same as in (C) but with  $\gamma_P = 4 \cdot 10^{-6}$  and  $4 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ .

$9.4 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  at all times), the friction coefficients are effectively the same as right here, differences in barrier height and distance compensate for each other. Importantly, close to this spot even the extracted transition path times agree well with each other and their prediction.

When looking at the combined transition rates in Figure 29B the situation seems better than for transition path times. Here, rates obtained for bead and protein diffusion almost perfectly match each other at all times. This means that the actual protein transition rates are reasonably well detected. However, when moving away from the above mentioned 'sweet' spot towards smaller protein friction coefficients, protein folding rates are increasingly underestimated with respect to their prediction. A simple explanation for this observation is directly suggested by the overall shape of the rate versus protein friction plot with respect to predicted rates: The slowest component within a force-spectroscopic measurement system acts as a low-pass filter. Hence, in

case the friction coefficient of the protein is effectively larger than the one of the bead, very good agreement between actual protein folding rates, their detection and their prediction can be expected. An interesting quantity which indirectly confirms the suppression of faster transition rates by the system's slower component is the factor between the number of times the faster component reaches the transition state out of a state's minimum and the number of successful transitions. As already noted at the end of [Section 4.2.1](#), a factor of 2 is expected for unimpeded diffusion. Here, in case of speed limit folding, the protein reaches the transition state about 7 times more often than an actual transition occurs.

The transition path ensembles shown on the right of [Figure 29](#) confirm the conclusions already drawn from transition path times and rates. In addition, the extracted ensembles directly visualize the friction-induced distortion of equilibria extracted from long-term fluctuations during the comparatively short time scales of transition paths. [Figure 29C](#) depicts the speed limit folding scenario. Again, the transition path ensemble of the slower component, i. e., the bead, is in accord with the theoretically expected shape<sup>14</sup>. In contrast to bead diffusion, the ensemble obtained from protein diffusion is strongly deformed with respect to its prediction and shows a much higher probability on the side of the more folded state. This, in turn, leads to a shift of the transition path probability of protein diffusion (not shown) into the direction of this state. Note that this shift is in accord with the bias of the bead transition stretches during the protein transitions shown in [Figure 28](#). In the opposite scenario presented in [Figure 29E](#) where protein friction is much higher than the friction of the bead, the protein transition path ensemble matches its prediction very well. Here, the ensemble of the bead is completely inverted with respect to its expected shape. The most interesting transition path ensembles are obtained close to the 'sweet' spot with effectively almost isotropic friction, see [Figure 29D](#). While the protein's ensemble matches its prediction almost perfectly, the ensemble of the bead already strongly deviates from the expected shape. This has to do with the preference of the minimal energy gradient transition path for crossing the higher protein folding barrier in this scenario, see discussion of [Figure 30](#) below. The result that the bead's transition path ensemble already deviates from its prediction in case of similar friction coefficients allows an important conclusion. A deviation of the bead ensemble from its prediction is direct evidence for another process being the driving force behind observed transitions which dominates over mere bead diffusion. In other words, measured transitions indeed reflect protein folding.

To summarize the results obtained from [Figure 29](#), here, the two questions raised at the beginning of the results' discussion are answered. In general, a correct detection of the transition path times of a protein is impossible. The only exception is distinguished at the 'sweet' spot with an effectively

<sup>14</sup> Note that the theoretically expected transition path ensembles for protein and bead diffusion are quite different in shape even though both are derived from simple two-state energy landscapes. Basically, this is owed to the relatively higher curvature at the barrier top with respect to the state's minima in deflection space as opposed to contour space. By this means, the probability to diffuse around the barrier top while being on a transition path is reduced.

isotropic friction where actual protein folding, its detection, and its prediction all agree with each other. In contrast to transition path times, protein folding rates are correctly detected at all times. However, since the slower component of the force-spectroscopic system acts as a low-pass filter, a rate detection of protein folding in conformity with its prediction is only possible in case the protein experiences higher friction than the bead. Taken together, the predictive capabilities of one-dimensional diffusion models when being applied to a two-dimensional diffusion problem are limited. The significance of measured transition path ensembles with respect to the protein folding process is analysed in detail in [Section 4.3.4](#).

All reported values for  $\gamma_P$  which are shown at the bottom of [Figure 29B](#) were derived from optical tweezers experiments. Despite differences in instrumentation, applied models, and referred coordinate space, all friction coefficients are higher but altogether close to the indicated case of isotropic friction (the 'sweet' spot). A discussion of whether this is mere coincidence, a systematic deviation, or a universal property goes beyond the scope of this work. In view of the above outlined limited applicability of one-dimensional diffusion models typically used to derive values for  $\gamma_P$ , the reported values can be seen as meaningful correction factors while their absolute validity should be handled with care. Note that for all values  $\gamma_P < \gamma_B$  the underestimation of the protein folding rate automatically leads to an overestimation of the friction coefficient  $\gamma_P$  while relevant barrier heights, if derived by deconvolution, are insensitive towards  $\gamma_P$ , see [Section 4.3.1](#).

The elastic compliance involved in force-spectroscopic measurements has direct influence on the detection limits. By changing, e. g., the trap stiffness, the distances  $\Delta x^{\text{TS}}$  for barrier crossing in deflection space will change accordingly. As suggested by [Equation 55](#), a change in  $\Delta x^{\text{TS}}$  affects the curvature at the barrier top which, in turn, has strong influence on the transition path time average given by [Equation 54](#). By this means, changing the systems compliance should allow the 'sweet' spot marked in [Figure 29A](#) to be moved. An alongside determination of the friction coefficient  $\gamma_P$  may answer the question whether or whether not the obtained result is independent from the measurement apparatus. However, note that changing the compliance on the detection's side also affects the barrier height in deflection space which can act against expected changes in curvature.

All elaborated results and conclusions presented in this section stand in strong agreement with recent publications dedicated to the numerical investigation of effects arising from anisotropic diffusion [[16](#), [49](#)]. Note that these publications introduce and confirm an analytic expression based on Langer theory, a multidimensional generalization of Kramers rate-theory, which can be used to characterize effective rates [[95](#), [125](#)]. Next, two-dimensional transition path ensembles are evaluated for a better understanding of how anisotropic friction induces effective rates and transition path times which stand in conflict with expectations derived from equilibrium distributions.

#### 4.3.2.3 Anisotropic friction acts as a switch between preferential transition paths

Thin lines in [Figure 30](#) are lines of equal potential representing the Hamiltonian  $H(x, L_P)$  which is used throughout this section. Thick dashed lines display minimal energy gradient paths for barrier crossing with preferred directionality. These paths are directly calculated from  $H(x, L_P)$  and indicate the energy minima within slices of either constant deflection (red) or constant contour length (black). Numbers in the top right corner of each plot specify the friction coefficient  $\gamma_P$  in units of  $\text{pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The background of [Figure 30A](#) shows the equilibrium probability distribution (different shades of earth color) of the entire two-dimensional simulation for speed limit folding where  $\gamma_P = 4 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . Thick lines refer to the maximal probability of the distribution along slices of either constant deflection (green) or constant contour length (light blue). [Figure 30B](#) presents the two-dimensional ensemble of all transition paths interconnecting the minima of  $H(x, L_P)$  according to diffusion in deflection space. Red thin dashed lines mark the relevant boundaries. The whole transition stretches shown in [Figure 28](#) are the one-dimensional representation of two examples which belong to the ensemble in [Figure 30B](#). The projection of this ensemble into mere deflection space can be seen on the right of [Figure 29C](#). The ensemble of transition paths interconnecting the minima of  $H(x, L_P)$  according to diffusion in contour space is displayed in [Figure 30C](#). [Figure 30D](#) and [Figure 30E](#) show the same as [Figure 30B](#) and [Figure 30C](#) but for different values of  $\gamma_P$ . For simplicity, all plots aside from [Figure 30A](#) show only the maximal probability within slices of the coordinate being determinant for the presented ensemble. Further, all distributions in [Figure 30](#) are not normalized.

With their preferred direction of propagation being in parallel to either deflection or contour space, thick dashed lines in [Figure 30](#) proceed along a minimal local energy gradient between the two minima of  $H(x, L_P)$ . By this means, thick dashed lines represent the energetically most favourable transition path which is expected for propagation along an axis with its associated friction coefficient. For the equilibrium probability distribution obtained from diffusion in the potential of  $H(x, L_P)$ , the probability to propagate along the thick dashed lines is maximal for each preferred direction, see [Figure 30A](#). Note that this is independent of the choice of  $\gamma_P$  because the latter does not affect the shape of a well sampled equilibrium distribution.

In case of the transition path ensemble which is composed of transitions in deflection space shown in [Figure 30B](#), the maximal probability in slices of constant deflection (green) excellently matches its prediction. This is not the case for the maximal probability in slices of constant contour length (light blue) derived from the transition path ensemble comprising transitions in contour space presented in [Figure 30C](#). Here, black arrows indicate the strong deviation from the prediction (black thick dashed line). Since [Figure 30B](#) and [Figure 30C](#) reflect transition path ensembles obtained for speed limit folding where, within this study, the anisotropy in friction is highest, they suit very well to directly illustrate effects arising from anisotropic friction. Simply speaking, anisotropic friction implies a clash of time scales. While protein dif-

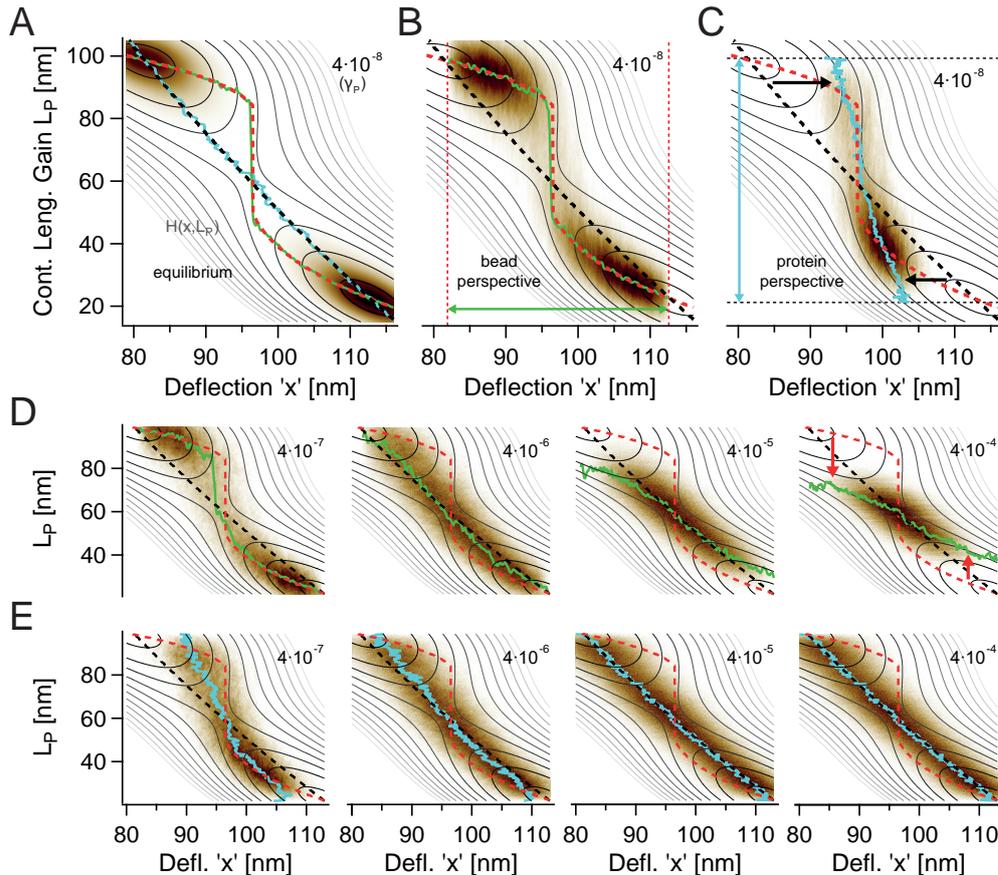


Figure 30: Friction acts as a switch between preferential transition paths. (A) Lines of equal potential represent  $H(x, L_P)$ . Shades of earth colour show an equilibrium probability distribution. Thick dashed lines mark preferential transition paths with respect to contour (black) and deflection (red) space. Green and light blue lines are preferential paths extracted from the distribution. (B) and (C) Transition path ensembles of full transitions in deflection and contour space. (D) and (E) Same as (B) and (C) with varied  $\gamma_P$ . See text for more details.

fusion is very fast in case of speed limit folding, bead diffusion is unable to follow the protein's fast movement. Hence, a full transition in contour space has to occur within a relatively small bead displacement. Energetically this is only feasible in case the bead already diffuses around the transition state region. The result can directly be seen in Figure 30C where the most probable transition path for protein diffusion is completely off its – with respect to equilibrium – energetically favoured path ending up in transitions almost orthogonal to deflection space. The two-dimensional distribution of this ensemble displays the same asymmetry already seen on the left of Figure 29C with the same bias towards the folded state in both contour and deflection space, for the latter see Figure 28. Figure 30C offers a graphical explanation for this bias which has to do with energetic considerations. The spatial separation between the two energetically favourable transition paths is smaller on the

more folded side (with higher deflection values) than on the more unfolded side. Further, the effective spatial orientation of both paths is more resembling to one another. This, in turn, reduces effects arising from anisotropic friction and allows an overall longer time to be spent on the more folded side during protein folding transitions. In the end, the asymmetry of the ensemble in [Figure 30C](#) reflects the asymmetry of  $H(x, L_P)$  which is directly visible on closer inspection of its two-dimensional representation. One reason for this asymmetry is the more compact nature of the folded state.

By stepwise increasing  $\gamma_P$ , in [Figure 30D](#) and [Figure 30E](#) the situation discussed for [Figure 30B](#) and [Figure 30C](#) changes towards the other extreme where the bead can diffuse much faster than the protein. Red arrows in the outermost right plot of [Figure 30D](#) indicate the deviation of the most probable transition path in deflection space from its energetically favoured one. Again, transitions are almost orthogonal to the axis representing the coordinate with higher friction. The only configuration where transitions in deflection space completely capture protein folding transitions is approximately given for  $\gamma_P = 4 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  which is closest to the 'sweet' spot with effectively isotropic friction. It is remarkable that the transition path ensembles based on transitions in deflection space continuously change along with  $\gamma_P$  while contour space ensembles remain unchanged as soon as protein friction dominates the transitions. Basically, this has to do with the higher barrier for protein folding transitions in contour space.

In summary, [Figure 30](#) illustrates how anisotropic friction shifts the most probable transition paths away from their energetically favoured route. These shifts, in turn, directly explain the aforementioned conflicting transition path times and rates shown in [Figure 29](#). The degree of shifting stands in direct relation to the difference in friction as well as the difference in barrier height.

Finally, the finding of a continuously changing transition path ensemble in deflection space in response to changes of the friction coefficient  $\gamma_P$  suggests a new approach to measure protein friction. In this approach, the deviation of a measured transition path ensemble from its theoretically expected shape serves as a direct measure of protein friction. If all relevant mechanical parameters as well as the deconvolved protein folding energy landscape of a force-spectroscopic experiment are known, the Hamiltonian  $H(x, L_P)$  describing the entire system can be calculated. Based on  $H(x, L_P)$ , a numerical determination of  $\gamma_P$  is straight forward. The first-time application of this approach to experimental data is presented in [Section 6.4.2](#), see also [Figure 65](#).

#### 4.3.3 *Transition averages at experimental resolution: seeking traces of protein folding*

As elaborated in [Section 4.3.2](#), transitions which are detected in a force-spectroscopic experiment rarely fully capture entire protein folding transitions. Here, transition averages from bead diffusion are analysed at experimental resolution to find out down to which time scale the measurement system is sensitive enough to directly reveal traces of protein folding within detected

transitions. Thereafter, the effect of basic protein folding energy landscape features on the shape of detected averaged transitions is investigated. Features comprise the transition state position as well as on-pathway intermediates.

#### 4.3.3.1 *Dynamic fingerprints: introducing transition averages*

Before being averaged, transitions are detected with the help of HMM analysis and aligned in two different ways. One way, is to use the location of the instant transitions within the HMM state classification trace for alignment. Alternatively, transitions are aligned at their starting point which is located right before the last crossing of the initial state level. Henceforth, HMM-centred transition averages are referred to as 'averaged transitions', while the other averages are termed 'bead relaxation curves'. Both types of transition averages can be seen in Figure 31. Dashed vertical lines indicate the positions of alignment, dashed horizontal lines represent the state levels.

The averaged transitions shown in Figure 31A are based on over 3000 transitions extracted from a one-dimensional diffusion simulation in the potential depicted in Figure 27A with  $\gamma = \gamma_B = 9.4 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The time-reversed representation of one of the two transition directions reveals that the averaged transitions are exactly the same for both directions. This is expected, since the same barrier is being crossed independent of direction. The black thick dashed line is a sigmoidal fit to the averaged transitions according to the following adapted equation of the logistic function [248]:

$$X(t) = \frac{A}{1 + e^{-\alpha \cdot (t-t_0)}} + X_0. \quad (65)$$

Here,  $X$  is the measured deflection at time  $t$ ,  $A$  is a scale factor,  $X_0$  is an offset,  $t_0$  is the time right at the middle of the averaged transition, and  $\alpha$  reflects the slope of the transition. While  $A$  adapts to the distance between the two state levels,  $X_0$  equals to their mean value. Further, averaged transitions are aligned such that  $t_0 = 0$ . Judging from the quality of the fit in Figure 31A, a sigmoidal shape characterizes the averaged transitions reasonably well.

Figure 31B presents averaged transitions based on about 100 transitions extracted from the deflection signal of the two-dimensional diffusion simulation characterized in Figure 27. By this means, transitions belonging to Figure 31B effectively pass through the same potential which is used for one-dimensional diffusion in Figure 31A. However, in Figure 31B additional speed limit protein folding transitions are present. While the red line is a sigmoidal fit to the data, the black dashed line displays the fit from Figure 31A representing one-dimensional bead diffusion unaltered by protein folding.

The full and empty circles shown in Figure 31C are bead relaxation curves calculated from the same transitions as in Figure 31B. Red lines are single exponential fits to the data. Again, black dashed lines refer to transitions from one-dimensional diffusion. Even though the single exponential fits capture the trend within the data fairly well, the two-dimensional bead relaxation curves reveal deviations from an ideal exponential decay. These deviations are similarly observed for one-dimensional diffusion (data not shown).

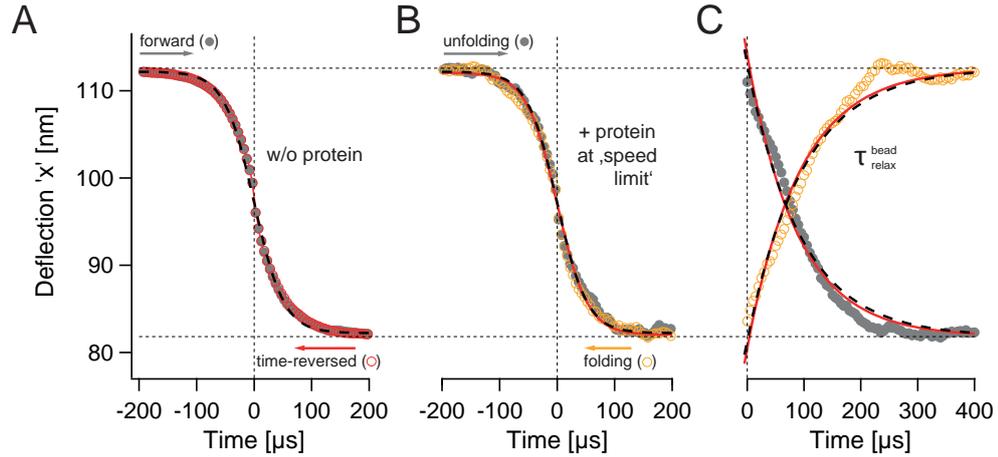


Figure 31: Averaged transitions of the deflection signal with and without inherent speed limit protein folding. (A) Averaged transitions (full and empty circles) from one-dimensional diffusion in the bead's effective energy landscape shown in Figure 27A using  $\gamma = \gamma_B = 9.4 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  while sampling at 200 kHz. The dashed black line is a sigmoidal fit to the data, see Equation 65. (B) Same simulation as in Figure 27 sampled at 200 kHz. Averaged transitions detected upon two-dimensional diffusion involving speed limit protein folding. The red line fits the data, the black dashed line refers to (A). (C) Bead relaxation curves from the same data set as used in (B). Red lines are single exponential fits to the data, black dashed lines refer to the data set represented in (A).

The key result of Figure 31 is directly put across by nearly identical fits for one- and two-dimensional diffusion: for the given set of parameters, measurable transitions do not give a hint at traces of protein folding even if the former are caused by the latter. In other words, speed limit protein folding transitions are invisible to the detection system. Note that the higher noise level within the averaged transitions from two-dimensional diffusion has nothing to do with protein folding but is solely a result of less statistics, see above. Note as well that besides the relaxation curves also the averaged transitions show a faint but orderly aberration from the shape of their sigmoidal fit. To perform a quantitative analysis of transition averages, the systematic deviations within the data require adjustments of the fit functions in order to extract more accurate characteristic lifetimes.

#### 4.3.3.2 In search of the detection limit for protein folding transitions

Now, the detection limit for direct traces of protein folding is determined based on simulations. Therefore, the friction coefficient  $\gamma_P$  is stepwise increased to slow down protein diffusion in contour space. In Figure 32, the thereupon obtained transition averages in deflection space (empty circles) are directly compared to the corresponding fit from one-dimensional diffusion (black thick dashed lines). Here, sigmoidal and exponential fits (solid lines) are modified to adapt to the systematic deviations seen in Figure 31. The

sigmoidal fits in Figure 32A are complemented by a second sigmoid which allows the averaged transitions to be split up into two regimes. One regime characterizes the steep slope around the alignment position  $t_0$ , while the other one fits the exponential decay towards each state level more accurately. Concerning the exponential decay of the bead relaxation curves in Figure 32B, a combination of three time shifted exponentials was found to reproduce the data best. To combine two time shifted decays, a third exponential with opposite amplitude is needed to compensate for the overlap of the other two around time zero. Coloured dashed lines in Figure 32B are simple exponential fits.

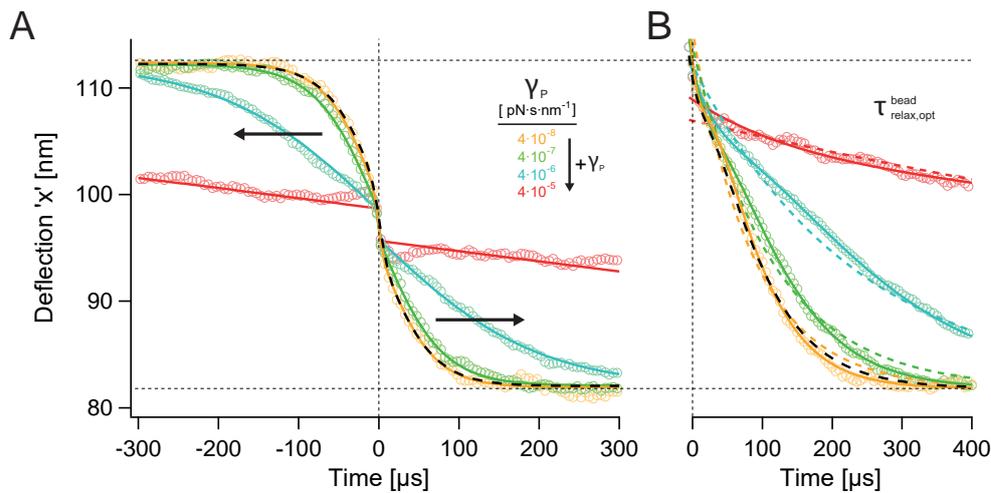


Figure 32: Averaged transitions of recorded bead dynamics reveal the inherent protein folding time scale. Simulation parameters are the same as in Figure 31 except for the indicated change of  $\gamma_P$ . (A) Empty circles are averaged transitions obtained upon varying  $\gamma_P$ . Solid lines are double sigmoidal fits. The black thick dashed line is a fit referring to one-dimensional diffusion. (B) Bead relaxation curves (empty circles) fit by time shifted exponentials (solid lines) for different values of  $\gamma_P$ . Coloured dashed lines are single exponential fits. Data and color code are the same as in (A). The black dashed line refers to one-dimensional diffusion.

For a friction coefficient of  $\gamma_P = 4 \cdot 10^{-7}$  pN · s · nm<sup>-1</sup> which is ten times higher than for speed limit protein folding, the transition averages in Figure 32 already reveal a noticeable deviation from unperturbed one-dimensional diffusion, i. e., from the black dashed lines. Along with increasing friction, deviations become stronger and stronger. Hence, above a critical value for  $\gamma_P$ , transition averages can give direct proof for protein folding. Further, the distinct deviation from unperturbed bead diffusion represents a fingerprint of the protein's characteristic folding time scale, while no detectable deviation indicates, at least, an upper limit of the folding time. Thanks to their modification, all fits in Figure 32 reproduce the data very well. In Figure 32B, the time shifted exponential fits (solid lines) clearly outperform their single exponential alternatives (short dashed lines). A selection of the fit results is presented and discussed next.

**Table 2** lists average transition path and relaxation times of transitions extracted from one- and two-dimensional diffusion simulations. The relaxation times  $\tau_{\text{relax}}^{\text{bead}}$  and  $\tau_{\alpha}^{\text{bead}}$  are provided by single exponential and sigmoidal fits to transition averages. In case of the sigmoid, relaxation times are calculated according to  $\tau_{\alpha}^{\text{bead}} = 2/\alpha$  in units of seconds [s], where  $\alpha$  is the slope defined by [Equation 65](#). Relaxation times marked with the index ‘opt’ are the longest relaxation times provided by the customized exponential and sigmoidal fits, see above. By this means,  $\tau_{\text{relax,opt}}^{\text{bead}}$  and  $\tau_{\alpha,\text{opt}}^{\text{bead}}$  characterize the slowest exponential decay found within a transition average. Values in brackets are theoretical predictions based on [Equation 54](#) with the addition that curvatures at the barrier top as well as at the bottom of the potential wells are proportionally considered, see discussion in [Section 4.2.1](#).

$\gamma_B$	$\gamma_P$	$\langle \tau_{\text{TP}}^{\text{prot}} \rangle$	$\langle \tau_{\text{TP}}^{\text{bead}} \rangle$	$\tau_{\text{relax}}^{\text{bead}}$	$\tau_{\alpha}^{\text{bead}}$	$\tau_{\text{relax,opt}}^{\text{bead}}$	$\tau_{\alpha,\text{opt}}^{\text{bead}}$
[pN · s · nm <sup>-1</sup> ]				[μs]			
$9.4 \cdot 10^{-6}$	-	-	162 (173)	<b>96</b>	<b>52</b>	<b>69</b>	<b>69</b>
-	$4 \cdot 10^{-8}$	3.2 (2.8)	-	-	-	-	-
$9.4 \cdot 10^{-6}$	$4 \cdot 10^{-8}$	26	158	<b>91</b>	<b>53</b>	<b>58</b>	<b>66</b>
$9.4 \cdot 10^{-6}$	$4 \cdot 10^{-7}$	82	191	<b>115</b>	<b>70</b>	<b>73</b>	<b>77</b>
$9.4 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	400	383	232	172	153	197
$9.4 \cdot 10^{-6}$	$4 \cdot 10^{-5}$	3263	1876	1617	1248	1048	1400

Table 2: Detection limit for protein folding transitions. Transition path and relaxation times from one- and two-dimensional diffusion simulations. Above a critical value for  $\gamma_P$ , the measurement system’s characteristic relaxation time is slowed down by protein diffusion. Most of the presented results are extracted from data presented in [Figure 29](#), [Figure 31](#), and [Figure 32](#).

For one-dimensional diffusion, transition path time averages listed in [Table 2](#) agree very well with their theoretical predictions confirming, once again, integrity of the simulations. Concerning bead relaxation times, there is a pronounced difference between  $\tau_{\text{relax}}^{\text{bead}}$  and  $\tau_{\alpha}^{\text{bead}}$  especially for fast transitions where  $\langle \tau_{\text{TP}}^{\text{bead}} \rangle$  is shorter than 200 μs. This difference is a consequence of the discrepancies between the transition averages and their simple fit functions visible in [Figure 31](#): while simple sigmoidal fits underestimate the relaxation times, single exponentials clearly tend to overestimate them. In contrast,  $\tau_{\text{relax,opt}}^{\text{bead}}$  and  $\tau_{\alpha,\text{opt}}^{\text{bead}}$  match each other much better. The relaxation times obtained for one-dimensional bead diffusion listed in the first row do almost not change in case of speed limit protein folding presented in row number three. Compared to speed limit folding in row three, substantially longer relaxation times in row four clearly detect the increase in protein friction  $\gamma_P$ . So what does this particular case reveal about the general detection limit for direct protein folding transitions? To answer this question, it suffices to include the theoretically expected transition path times for protein folding into

the picture. As indicated in row two, protein folding transitions are expected to last  $2.8 \mu\text{s}$  in case of unperturbed speed limit protein folding. Hence, in row four, an average protein folding time of  $28 \mu\text{s}$  is theoretically expected as  $\gamma_P$  is ten times higher. This leads to the plausible result that traces of protein folding transitions can be directly detected when unperturbed protein folding times are of the same order of magnitude (or longer) as the (slowest) relaxation time of unperturbed bead diffusion. Here, this critical relaxation time lies between 60 and  $70 \mu\text{s}$ . In summary, the detection limit for direct protein folding transitions is predetermined by the response time of the detection system.

Concerning the superior quality of the customized fits, of course, fitting becomes easier by introducing more fit parameters. The important advantage of the expanded fits is that by entirely modelling the data they provide a more accurate relaxation time of the slowest system's component which is determinant for the detection limit. Importantly, the relaxation times  $\tau_{\text{relax,opt}}^{\text{bead}}$  and  $\tau_{\alpha,\text{opt}}^{\text{bead}}$  are in good agreement with results provided by single exponentials which are locally fit to the transitions within their last third before reaching the final state.

The apparent step right in the middle of all averaged transitions looks very much like an alignment artefact. Instead of using the HMM state classification trace, one might think of fitting the logistic function to each transition and use the provided  $t_0$  for alignment, see [248]. However, fitting a sigmoid to individual transitions already anticipates a symmetric shape for their average which, in general, is not the case, see below. The problem whether the deviations observed for bead relaxation curves are a mere (alignment) artefact or the result of, say, two time shifted exponential decays reflecting diffusion within a potential well and diffusion over a barrier and into a well, is beyond the scope of this work.

By comparing experimentally determined transition path time averages to their theoretical (= 'protein-free') predictions in Figure 53 of Section 6.2.3, both faint as well as pronounced traces of protein folding are revealed.

#### 4.3.3.3 Averaged transitions reveal basic protein folding energy landscape features

Typically, protein folding energy landscapes are not symmetric. Therefore, the influence of asymmetric potentials in contour space on the shape of transition averages measured in deflection space are qualitatively studied in the last part of this section. Figure 33 illustrates the effect of either a non-centred on-pathway high-energy intermediate or an asymmetric predominant transition state (TS). The upper panel of Figure 33A shows two protein folding energy landscapes, one with (solid line) and one without intermediate (dashed line). Below, their corresponding bead diffusion potentials are displayed. Note that the (almost) symmetric potentials represented by the black dashed lines are the ones which are also used for the other two-dimensional diffusion studies in this section. Full and empty circles in Figure 33B display averaged transitions detected upon diffusion in the energy landscape with intermediate. The dashed sigmoid is a fit to averaged transitions obtained from diffusion in the

energy landscape without intermediate. The same fit and associated data are depicted in Figure 31B. Red lines in Figure 33C are single exponential fits to the bead relaxation curves obtained for diffusion influenced by the additional intermediate. Again, thick dashed lines refer to diffusion without intermediate; associated data are fit and shown in Figure 31C. In an analogous manner, Figure 33D - Figure 33F demonstrate the impact of an asymmetric transition state position. Note that, in Figure 33D, the red protein folding energy landscape shape is designed such that forward and backward transition rates are about equal according to Kramers rate equation, see Equation 29.

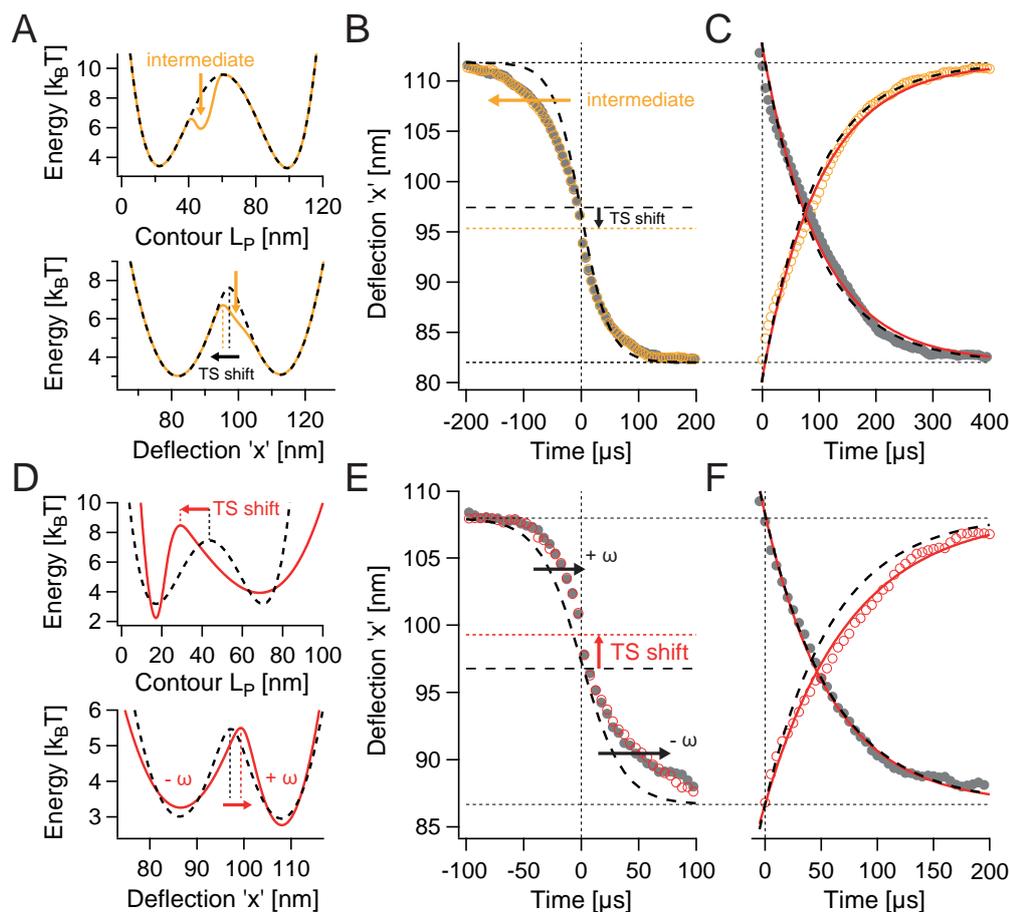


Figure 33: Averaged transitions reveal basic features of protein folding energy landscapes. Same simulation parameters as in Figure 31B. (A) - (C) An additional on-pathway intermediate deforms detected averaged transitions. Thick black dashed lines in (B) and (C) are reference fits from averaged transitions detected upon diffusion in a potential without intermediate. (D) - (F) Effect of an asymmetric protein folding potential on detected transition averages. See text for more details.

The additional intermediate of the protein folding energy landscape shown in Figure 33A has a considerable impact on the barrier shape of the convolved bead diffusion potential. This asymmetric deformation leads to a deviation of the averaged transitions from a symmetric shape represented by the sigmoid in Figure 33B. As indicated by the orange arrow, detected diffusion from or

into the folded state, i. e., the upper state level, is effectively slowed down due to the intermediate. Further, the transition state shift, which is pointed out in the lower panel of [Figure 33A](#), can also be seen within the averaged transitions: their alignment position is clearly off the middle between the two state levels. In [Figure 33C](#), the red fits to the bead relaxation curves are almost identical to the ones obtained for the symmetric folding energy landscape (black dashed lines). This indirectly suggests that bead relaxation times are rather dominated by the shape of energy wells than by the shape of the crossed barrier since, regardless of the presence of the intermediate, in [Figure 33A](#), all energy wells have almost identical shapes within their respective coordinate space. On the other hand, a direct comparison between [Figure 33C](#) and [Figure 31C](#) may lead to believe that relaxation times are slightly prolonged due to the deformation of the barrier caused by the additional intermediate.

As one can see in [Figure 33D](#), a strongly asymmetric protein folding energy landscape causes a corresponding asymmetry in deflection space. While the barrier shape around the shifted transition state position still remains rather symmetric, now, the energy wells of the convolved bead potential show a strong asymmetry which manifests itself in a gain ( $+\omega$ ) or loss ( $-\omega$ ) in curvature. Again, the alignment position of the averaged transitions in [Figure 33E](#) reflects the shifted transition state position in deflection space. In addition, as indicated by black arrows, diffusion out of or into the upper state level is accelerated, while diffusion into or out of the lower state level is slowed down. A comparison between the exponential fits in [Figure 33F](#) reveals an asymmetric deviation: while bead relaxation curves from the upper towards the lower state level do not indicate any deviation from the reference fit obtained from the symmetric energy landscape (black dashed line), the inverse transition seems to be slowed down. In a certain way, this observation suggests that relaxation times for barrier crossing are rather dominated by the time needed for reaching the barrier than for leaving it.

The results in [Figure 33](#) represent the speed limit protein folding case. Hence, detected averaged transitions and bead relaxation curves are not additionally deformed by protein diffusion, see [Figure 31](#). From this follows that all deviations from a symmetric shape basically reflect the asymmetric features of the convolved bead potential. In case of averaged transitions, slowed down or accelerated diffusion can be directly explained by the relative changes in deflection space inferred by the asymmetric protein folding energy landscapes. Therefore relevant are changes in barrier height  $\Delta G^{\text{TS}}$ , barrier position  $\Delta x^{\text{TS}}$ , as well as *all* involved curvatures, see [Figure 19](#) in [Section 3.8.4](#) and the discussion in [Section 4.2.1](#). Under consideration of the named parameters, the slight prolongation of relaxation times observed in [Figure 33C](#) appears to be reasonable. The different relaxation times in [Figure 33F](#) suggest a direction-dependence which stands in conflict with the perfect overlap of the respective averaged transitions. Further, an equilibrium process should not yield direction-dependent times for barrier crossing. In view of these contradictions, a more in-depth analysis will be necessary to understand the ori-

gin of the apparent direction-dependence and to rule out artefacts potentially arising from a lack of statistics or resolution.

In summary, in case of speed limit protein folding, detected averaged transitions reflect basic features of the convolved bead potential. One feature comprises the transition state position in deflection space which is marked by the inflection points of the averaged transitions and located right at their alignment position. This result is particularly relevant as it reveals that symmetric sigmoidal fits introduced in [247] are really useful only in the exceptional case of perfectly symmetric energy landscapes.

Another important result is the fact that both the additional intermediate and the shifted predominant transition state in contour space have an overall similar effect on the detected transitions. This means that as long as the protein friction coefficient is unknown, asymmetric averaged transitions need careful interpretation. In addition to asymmetries in the protein folding energy landscape, non-constant diffusion coefficients may cause asymmetric averaged transitions as well.

The barrier determining the intermediate's dwell time in [Figure 33A](#) is so small that Kramers rate-theory in the high-barrier limit can only provide an upper limit, see [Figure 24B](#). In the analysed case, i. e., for  $\gamma_P = 4 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , the expected upper limit of the intermediate's dwell time is as short as  $1.4 \mu\text{s}$  and thus far below the detection limit, see above. In this context, the mere ability to detect the deformation of the convolved bead potential caused by the intermediate is already remarkable.

In principle, averaged transitions can be interpreted as temporal committers or splitting probabilities. Hence, the inflection points of averaged transitions mark a (time-dependent) transition probability of 0.5 as they coincide with the transition state position. From this perspective, it will be interesting to see what effects on the shape of asymmetric averaged transitions arise upon an increase of protein friction.

#### 4.3.4 *Transition path analysis in force spectroscopy*

In this section, two-dimensional diffusion simulations are used as a test bed for transition path analysis tools introduced in [Section 3.8](#). Since these tools have not been applied to experimental force-spectroscopic data so far, here, their applicability and informative value are analysed under consideration of a whole set of different boundary conditions. These conditions comprise two protein folding energy landscapes with either a central or an asymmetric transition state position. In addition, three different protein friction coefficients are considered to cover cases of speed limit protein folding with dominating bead friction, effective isotropic friction, and predominating protein friction, see [Section 4.3.2](#) for details on the differentiation. The analysis focuses on the committor-based barrier shape reconstruction, see [Section 3.8.2](#), as well as the transition state localization based on the calculation of transition path probabilities, see [Section 3.8.3](#).

Full lines in Figure 34A represent commitment probabilities derived from deflection signals ‘measured’ for diffusion in the same two-dimensional energy landscape but with different protein friction coefficient values  $\gamma_P$ . Dashed lines are theoretically expected committors derived from the potentials shown in Figure 34B. Details on the derivation of committors from barrier shapes are given in Section 4.3.2. The inset in Figure 34A is a zoom into the intersect region of the commitment probabilities. While the red dashed line in Figure 34B reflects the convolved bead potential expected from the Hamiltonian  $H(x, L_P)$ , the dark blue dashed line represents the deconvolved bead potential. Note that these potentials are the same as the ones shown in Figure 27C. Full lines in Figure 34B are committor-based barrier reconstructions according to Equation 51. All curves in Figure 34C are transition path probabilities directly calculated from the committors shown in Figure 34A according to Equation 53.

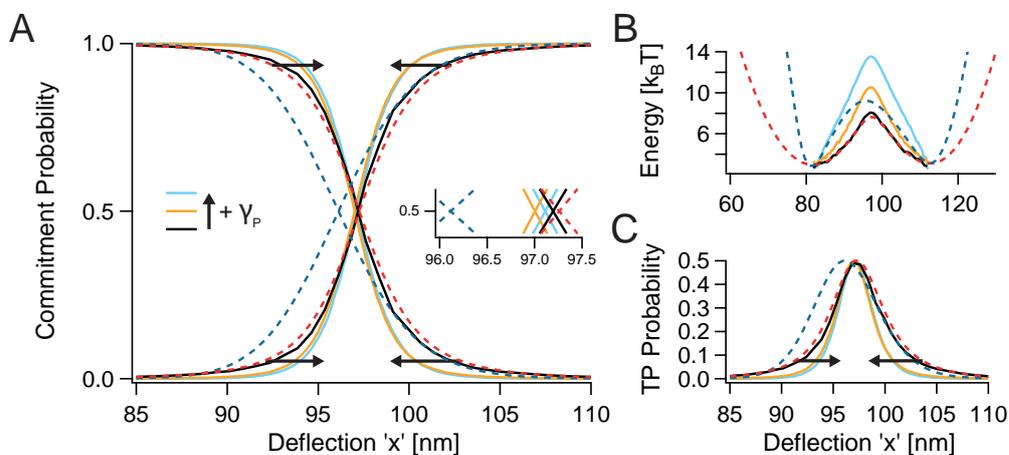


Figure 34: Commitment probability calculation and barrier reconstruction for a symmetric protein folding energy landscape. Same simulation as in Figure 27 sampled at 200 kHz with  $\gamma_P = 4 \cdot 10^{-8}$  (black lines),  $4 \cdot 10^{-6}$  (orange), and  $4 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  (light blue). (A) Full lines represent commitment probabilities derived from equilibrium trajectories according to Equation 49. Dashed lines are theoretically expected committors based on the dashed potentials shown in (B). (B) Committor-based barrier shape reconstruction based on Equation 51. Dashed lines represent the convolved (red) and deconvolved (dark blue) bead potential. (C) Transition path probabilities derived from committors in (A) based on Equation 53.

In case of speed limit protein folding, commitment probabilities in Figure 34A (black lines) are very close to the shape expected for unaltered bead diffusion (dashed red lines). When protein friction is increased, commitment probabilities become considerably steeper around the intersect region as indicated by black arrows. At the same time, only a marginal difference between isotropic (orange) and high protein friction (light blue) can be seen. The inset highlights that all committors intersect very close to the intersection derived from the convolved bead potential. Figure 34B shows that the speed limit folding scenario yields an almost perfect barrier reconstruction of the convolved

bead potential. Further, reconstructed barrier heights increase along with protein friction. Independent of differences in friction, all reconstructed barrier shapes have more resemblance to the convolved than to the deconvolved barrier. In case of isotropic friction (orange), the reconstructed barrier height is about as high as the deconvolved barrier (dashed dark blue line). The transition path probabilities in Figure 34C are solely a different representation of their corresponding committors and become more pointed with increasing protein friction. As for the committors, the deformation of the transition path probabilities is slightly asymmetric with a tendency towards accounting for the shape expected for the deconvolved bead potential.

The subsequent interpretation of evaluated simulation data shown in Figure 34 is based on the demand that committors, reconstructed barriers, and transition path probabilities match the respective curves derived from the deconvolved energy landscape (dashed dark blue lines) as the latter is expected to represent protein diffusion in deflection space. From this point of view, the results in Figure 34 lead to the following conclusions. As illustrated in Figure 34B, an exact barrier reconstruction in terms of height and shape is impossible. At least, in case of effective isotropic friction, the reconstructed barrier (orange) yields an acceptable estimate of the protein folding barrier height. Altogether, the committor-based barrier reconstruction technique appears to be quite sensitive and, hence, not very robust. This can be deduced from the fact that committors for isotropic and high protein friction (orange and light blue) are almost identical while their inferred barriers are substantially different in barrier height. Note that the similarity between the committors stands in agreement with the similarity between the respective transition path ensembles of protein diffusion shown on the left of Figure 29D and Figure 29E. According to Section 3.8.1, a commitment probability of 0.5 indicates a transition state configuration. By this means, each intersection in the inset of Figure 34A marks a presumable transition state position. The zoom into the intersect region of the committors reveals a relatively bad committor-based transition state localization for protein diffusion. While the best result is obtained for isotropic friction, all intersections are considerably closer to the transition state position of the convolved bead potential than to the position expected for protein diffusion. However, all intersections show the right tendency and deviate by less than 1.1 nm from the expected position which corresponds to less than 5 % of the total distance between the two state levels.

Figure 35 presents transition path probabilities derived from transition path ensembles according to Equation 52. The figure is subdivided into two main columns and three rows. While the left column compares data from speed limit folding simulations (black lines) to predictions (dashed lines), the right column compares data recorded for different protein friction coefficients  $\gamma_p$ . Note that predictions are independent of  $\gamma_p$ . Full lines in the top row are equilibrium distributions of entire trajectories in deflection space. The middle row displays transition path ensembles which include all detected unfolding and refolding transitions. Emerging transition path probabilities are depicted in the bottom row. While results represented by full lines are derived from

trajectories sampled at 200 kHz, dotted lines in the middle row are derived from the same trajectories at full 1 GHz bandwidth<sup>15</sup>. In accordance with Figure 34, thick dashed red and dark blue lines reflect theoretically expected shapes based on the convolved and deconvolved bead potentials. Transition path probabilities shown in Figure 35C and Figure 35F are based on the measured probability distribution. In contrast, probabilities shown in Figure 35G are obtained upon using the deconvolved probability distribution as  $p_{\text{eq}}(x)$  in Equation 52. The inset number in Figure 35G reflects the evaluated additional (re-)normalization factor<sup>16</sup> which has not been applied upon calculation.

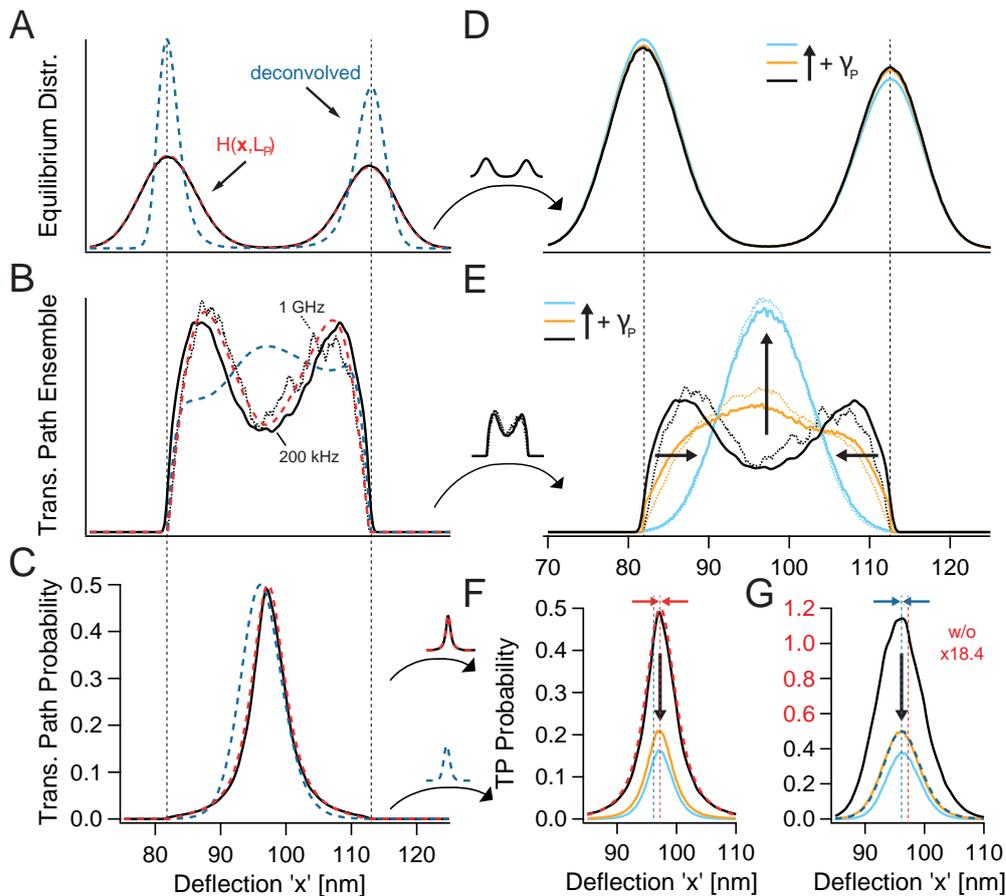


Figure 35: Transition path probability calculation for a symmetric protein folding energy landscape. Simulations and color code as in Figure 34. (A) Equilibrium distribution  $p_{\text{eq}}(x)$  of the bead deflection signal (black line). Dashed lines represent the deconvolved (dark blue) and convolved (red) distribution. (B) Extracted transition path ensemble  $p(x|TP)$  (black line). (C) Transition path probability  $p(TP|x)$  according to Equation 52 (black line). (D) - (F) Same as (A) - (C) but with varied  $\gamma_p$ . (G) Same as (F) but using the deconvolved distribution as  $p_{\text{eq}}(x)$  in Equation 52. See text for more details.

<sup>15</sup> Dotted lines in the middle row are the same transition path ensembles as in Figure 29.

<sup>16</sup> The additional normalization factor is defined as the integral over the ratio between deconvolved and convolved probability distributions, see supporting information for [155].

Independent of the protein friction coefficient  $\gamma_p$ , all measured equilibrium distributions in the top row of [Figure 35](#) are about the same and, hence, match the convolved distribution (red dashed line in [Figure 35A](#)) expected from the simulation's Hamiltonian  $H(x, L_p)$ . In contrast, transition path ensembles in the middle row show high sensitivity towards changes in  $\gamma_p$ , for a detailed discussion see [Section 4.3.2](#). Here, one can also see that detected ensembles are slightly broader<sup>17</sup> when sampled at experimental resolution. Despite huge differences between the transition path ensembles, all transition path probabilities in the bottom row have very similar shapes which basically differ only by amplitude. Transition path probabilities in [Figure 35C](#) and [Figure 35F](#) peak at the transition state position of the convolved bead potential. If the deconvolved equilibrium distribution is used for their calculation as in [Figure 35G](#), probabilities peak at the transition state position of the deconvolved bead potential. Independent of their calculation, probabilities get smaller with increasing protein friction. In case of speed limit protein folding, directly calculated transition path probabilities almost perfectly match the shape expected for the convolved bead potential, see black line in either [Figure 35C](#) or [Figure 35F](#). Probabilities based on the deconvolved equilibrium distribution perfectly match the shape derived from the deconvolved bead potential in case of effective isotropic friction, see orange line in [Figure 35G](#).

A surprising result in [Figure 35](#) is provided by the transition path probabilities derived upon using the deconvolved probability distribution as  $p_{\text{eq}}(x)$  in [Equation 52](#). Even without multiplication with the additional normalization factor, the probability for speed limit folding, see black line in [Figure 35G](#), exceeds the maximally possible value of 0.5. Furthermore, all probabilities in [Figure 35G](#) would exceed a meaningful maximum after multiplication with the corresponding normalization factor of 18.4. A possible explanation may be a literal mistake in the supporting information of [155]. If using  $\int_{x_1}^{x_2} p_{\text{deconvolved}}(x) dx / \int_{x_1}^{x_2} p_{\text{eq}}(x) dx$  instead of  $\int_{x_1}^{x_2} p_{\text{decon}}(x) / p_{\text{eq}}(x) dx$  to quantify the fraction of statistical weight in the transition region, here, a factor of  $\sim 1$  instead of  $\gg 1$  would be reached.

Transition path probabilities are maximal at the transition state position. When using [Equation 52](#) to calculate transition path probabilities, the location of the maximum strongly depends on the shape of  $p_{\text{eq}}(x)$ : the smaller  $p_{\text{eq}}(x)$ , the higher the weighting of the measured transition path ensemble at the respective position. As probability distributions have a minimum at their corresponding transition state position, this position gets the highest additional weight. The peak positions of differently calculated probabilities in [Figure 35](#) confirm this interdependence: after division of the ensembles by  $p_{\text{eq}}(x)$ , resulting probability maxima are close to the transition state position inferred by the equilibrium distribution. This suggests that an exact transition state localization is barely possible when calculations are merely based on measured probability distributions which typically differ from deconvolved ones.

<sup>17</sup> The slight broadening of transition path ensembles sampled at experimental resolution may be caused by a less accurate identification of the beginning and end of transition paths which, in turn, results in a tendency towards more additional points in the outer transition path regions.

Compared to the comittor-based probabilities in [Figure 34C](#), transition path probabilities derived from transition path ensembles bear more information as they are not set to a maximal value of 0.5 by default. Independent of their calculation, all transition path probabilities get more narrow with increasing protein friction. This narrowing reflects an effective reduction of the region in deflection space which is relevant for barrier crossing. By this means, narrowed transition path probabilities are a directly measurable effect of the friction-induced switch between preferential transition paths which is discussed in [Section 4.3.2](#). Importantly, this noticeable and verified relation between the shape of a transition path probability and the relevance of the respective coordinate in terms of reactivity confirms the potential of the former as an indicator of reaction coordinate quality. However, in combination with the findings from [Section 4.3.2](#), results in [Figure 35](#) reveal that transition path probabilities alone are likely to be misinterpreted when reaction coordinate quality is tested. The directly calculated probability in [Figure 35C](#) almost reaches a maximal value of 0.5 which identifies the deflection signal as a very good reaction coordinate, see [Section 3.8.3](#). While in case of speed limit protein folding this is true for bead diffusion, for protein diffusion this is not, see [Figure 30B](#) and [Figure 30C](#). Orange lines in [Figure 35F](#) and [Figure 35G](#) represent the effective isotropic friction case where the deflection signal captures protein diffusion best according to [Section 4.3.2](#). This is confirmed by the probability based on the (typically not available) deconvolved equilibrium distribution and its maximal value of 0.5 in [Figure 35G](#). Opposed to that, the directly calculated probability in [Figure 35F](#) only has a maximal value of 0.2 suggesting that the deflection signal is only a mediocre reaction coordinate. Taken together, these examples already illustrate that the use of transition path probabilities as a reliable test for reaction coordinate quality is limited in case of two-dimensional diffusion with anisotropic friction.

For completeness it should be noted that there is another parameter which is rarely directly mentioned but nevertheless crucial for transition path probability calculations and their interpretation:  $p(\text{TP})$ . While representing the fraction of time spent on transition paths,  $p(\text{TP})$  is directly proportional to the probability height according to [Equation 52](#). This makes  $p(\text{TP})$  co-determinant for reaction coordinate quality tests. Arising from the frequency and duration of transition paths,  $p(\text{TP})$  depends on detected transition path times and rates. By this means, the height of transition path probabilities also depends on the position chosen for the borders of the transition path region and on the quality of spatial and temporal resolution.

Next, an asymmetric protein folding energy landscape is used to test the performance of different transition path analysis tools. [Figure 36](#) shows an analysis of this landscape in the style of [Figure 27](#). In [Figure 36A](#) and [Figure 36B](#), Boltzmann inverted probability distributions (black lines) derived from simulated trajectories perfectly match the shapes expected from  $H(x, L_P)$  (dashed red lines). The dark blue line in [Figure 36A](#) represents the deconvolved energy landscape which, after transformation, also matches the shape expected in contour space quite well, see [Figure 36B](#). The zero-force transfor-

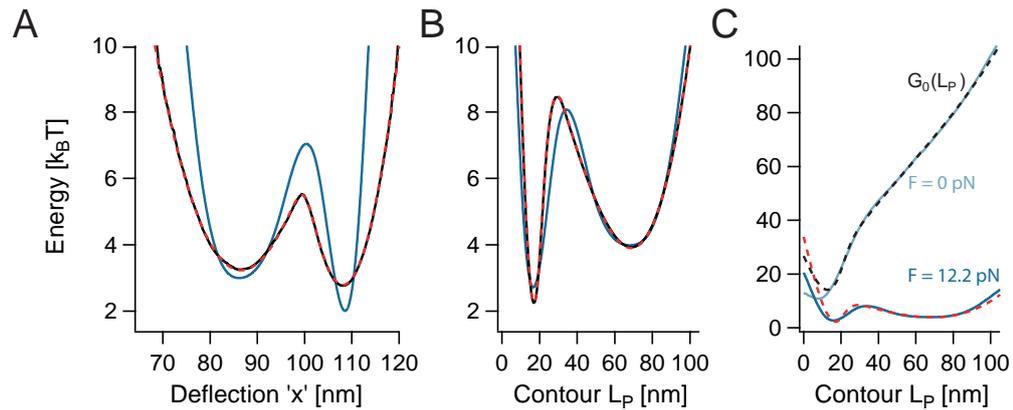


Figure 36: Deconvolution of an asymmetric protein folding energy landscape. (A) - (C) Boltzmann inversion and deconvolution of a two-dimensional diffusion simulation with an asymmetric protein folding energy landscape input. Except for  $G_0(L_P)$ , all simulation parameters are the same as in Figure 27. Figures are kept in the style of Figure 27C - Figure 27E. Dark blue lines are based on deconvolution. Additional black lines are Boltzmann inverted probability distributions of the respective individual simulation trajectories.

mation in Figure 36C confirms a good reconstruction of the simulation input  $G_0(L_P)$ . Since deconvolution represents the most solid energy landscape reconstruction technique on hand, in the following, the obtained deconvolved energy landscape is used as reference for retrievable information on protein diffusion concealed within the analysed deflection signals.

Figure 37 presents the exact same evaluation as Figure 34 with the only difference being the underlying asymmetric protein folding energy landscape. As one can see, the results provide the exact same overall picture. Concerning the transition state localization, tendencies towards the transition state of the deconvolved energy landscape are stronger, see inset of Figure 37A. In case of effective isotropic friction (orange lines), committors even almost intersect at the transition state position of the deconvolved energy landscape (dashed dark blue lines). As already seen in Figure 34A, the intersection moves back towards the transition state of the convolved energy landscape in the high protein friction case (light blue). This can be explained by an increased relative diffusional freedom of the bead which allows the transition barrier in deflection space to be probed more often than the protein gets close to its transition state in contour space.

In Figure 38, transition path probability calculations are presented in the same way as in Figure 35. Again, results from diffusion in an asymmetric protein folding energy landscape correspond to those obtained from diffusion in the symmetric case. In case of isotropic friction, the transition path probability in Figure 38F (orange line) even shows a marginal tendency towards the transition state position of the deconvolved potential (dark blue vertical dashed line). The maximal probability height, however, is already smaller than for the symmetric case presented in Figure 35F and only marginally higher than

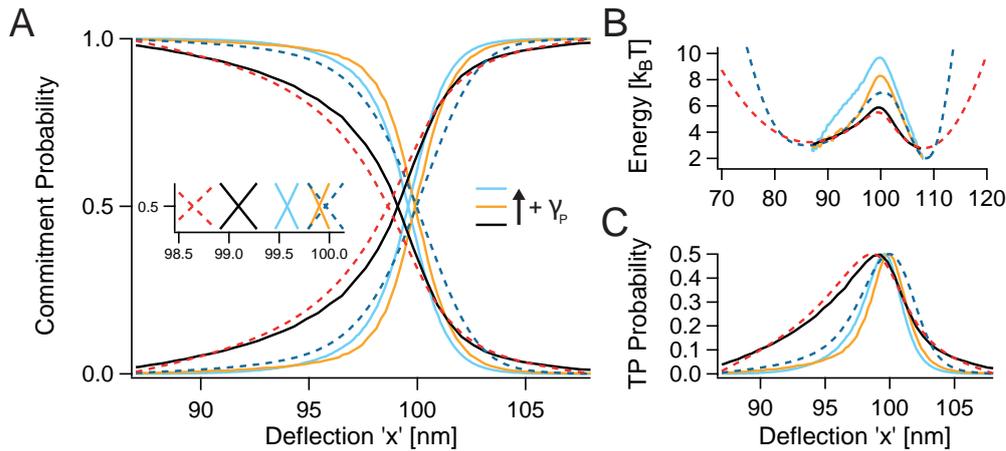


Figure 37: Commitment probability calculation and barrier reconstruction for an asymmetric protein folding energy landscape. (A) - (C) Arrangement, color code for  $\gamma_P$ , and sampling are the same as in Figure 34.

for high protein friction (light blue). A possible explanation could be a loss in spatial resolution due to the smaller distance between the two state levels in case of the asymmetric protein folding energy landscape. Further, the condition of effective isotropic friction has been adopted from the symmetric case which, hence, is only a rough estimate. A two-dimensional analysis as in Section 4.3.2 may help to better understand these details.

In summary, with respect to expected barrier reconstructions and transition state localizations, tested transition path analysis tools have shown poor performance within the frame of anisotropic two-dimensional diffusion. While best results are obtained in case of almost effectively isotropic friction, expectations risen by the excellent performance shown for one-dimensional diffusion in Section 3.8 are not met. However, all expectations are based on a deconvolved energy landscape reflecting transitions in equilibrium. In case of anisotropic friction, results in Section 4.3.2 clearly show that transitions do not match expectations from equilibrium. From this perspective, the apparent poor performance of transition path analysis tools is direct proof of the fact that detected bead diffusion is altered by the tethered protein. Narrowed and reduced transition path probabilities directly reflect the gain in protein friction which leads to transitions truly guided by protein diffusion. Along these lines, interpretations of maximal transition path probability values in terms of reaction coordinate quality need to be done with care. Another important finding of this section is that transition state localizations based on Equation 52 are strongly biased towards the minimum of the applied equilibrium probability distribution. Further, the high sensitivity of transition path ensembles towards protein friction advises their explicit consideration when analysing transition paths.

The results obtained in this section are highly relevant with respect to a series of recent publications, in particular [42, 143, 155, 156, 157]. In addition to two-dimensional diffusion and one-dimensional position-dependent friction

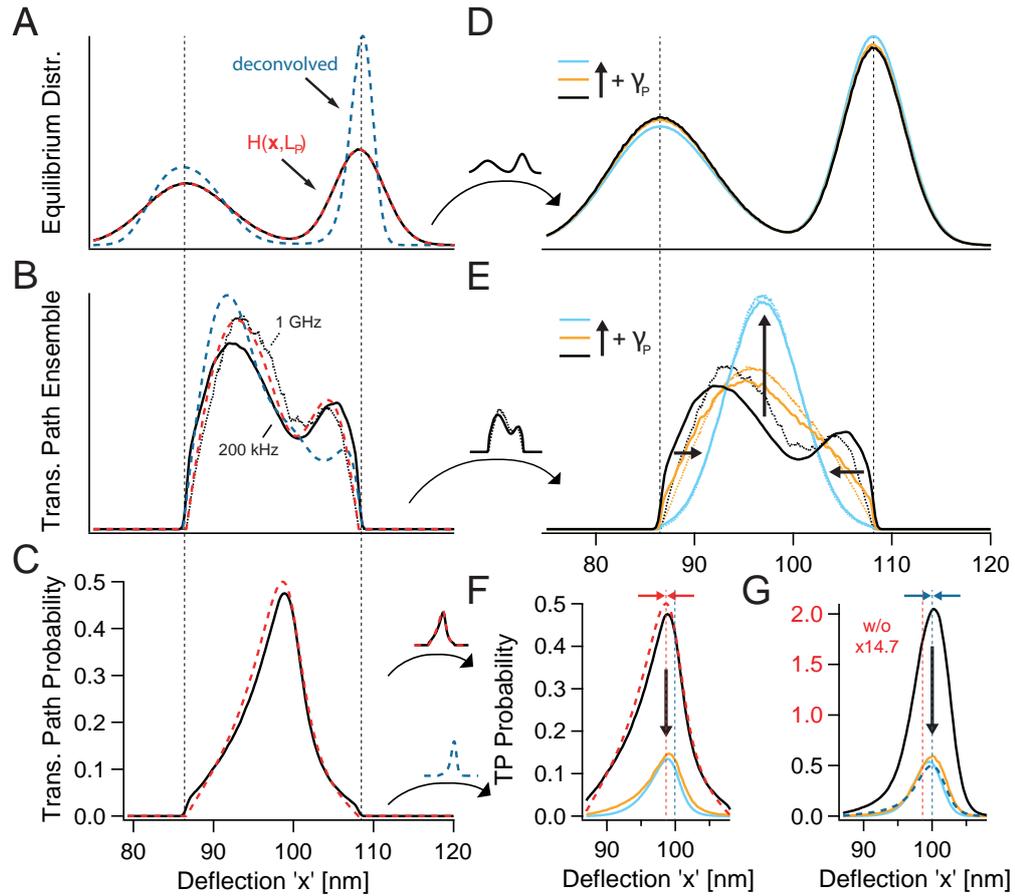


Figure 38: Transition path probability calculation for an asymmetric protein folding energy landscape and the effect of friction. Same simulations and color code as in Figure 37. (A) - (G) Arrangement as in Figure 35.

considered in [42], here, the impact of anisotropic friction on commitment probabilities has been investigated for the first time. In [143], the comittor-based barrier reconstruction technique was introduced for force-spectroscopic experiments where  $\gamma_B \approx \gamma_P$ . This basically represents the isotropic friction scenario of this work which is the exceptional case yielding acceptable results. And, finally, the supposedly improved transition path probability calculation which involves the deconvolved probability distribution as well as an additional (re-)normalization factor, see [155], has been revealed to simply yield the transition state position implied by the deconvolved distribution. Given that the deconvolved energy landscape must be already known in advance, this technique does not provide new insight with respect to the transition state. In addition, even without applying the (re-)normalization factor, a significant overestimation of resulting probabilities has been observed, see Figure 35G and Figure 38G, which results in meaningless probability values. Note that in [155], transition paths are restricted to the region between the inflection points of the convolved bead potential. This has remarkable effects on transition path ensembles  $p(x|TP)$  and transition path times which further

affect  $p(\text{TP})$ , see [Section 3.8.5](#) and [Section 4.2.1](#). According to the results from this section, reported probabilities of almost 0.5 correspond to the speed limit protein folding scenario. In this scenario, deflection signals are typically not a truly good reaction coordinate, see [Section 4.3.2](#).

#### 4.4 SUMMARY AND OUTLOOK

##### *Evidence for self-contained simulations*

The one-dimensional diffusion simulation performed in [Section 4.2.1](#) shows that [Equation 54](#) is a good approximation for expected average transition path times  $\langle\tau_{\text{TP}}\rangle$ . In case of identical curvatures at the barrier top and the bottom of the energy wells, discrepancies between theory and simulation are smaller than 10%. If curvatures differ, transition path times should be calculated as a sum of correspondingly weighted path segments where energy wells are treated like inverted barriers starting from the respective inflection points. Furthermore, rates  $k^{\text{ij}}$  extracted from the simulation confirm the predictive capabilities of Kramers rate equation, see [Table 1](#). Here, theory and simulation differ by less than 2.5%. Importantly, the good agreement between theory and simulation proves that simulation time steps  $\Delta t$  are chosen such that the discretization in [Equation 60](#) and [Equation 62](#) is fine enough to fully capture the modelled diffusive processes.

In [Section 4.3.1](#), the consistency of the two-dimensional diffusion simulation is confirmed by a correct reconstruction of the simulation input  $G_0(L_P)$  from the output of the simulation. The reconstruction itself is based on deconvolution of the simulated measurement signal.

##### *Different types of roughness and their characterization*

[Section 4.2.2](#) presents the first numerical verification of the roughness concept introduced by Zwanzig. Inspired by the principle of frequently revisited deeper minima which lead to increased overall transition path times, a simplified concept for roughness represented by only a few high-energy intermediates has been proposed. However, since Zwanzig's roughness is characterized by a relatively small typical length scale  $\Delta x^r$  compared to the spatial dimension of the overall barrier, [Equation 63](#) does not apply to the simplified concept.

A method to reconstruct the energy well depth of high-energy on-pathway intermediates based on measured transition path times is introduced in [Section 4.2.3](#). Direct application to different scenarios including one or two intermediates confirmed the robustness of the method, see [Figure 25](#).

Simulations performed in [Section 4.2.4](#) confirm Hammond behaviour and prove that kinked chevron plots can indeed be attributed to a rough energy landscape comprised of high-energy intermediates, see [Figure 26](#).

*Dynamic fingerprints of protein folding*

Two types of transition averages are introduced in [Section 4.3.3](#). These averages can be modelled with sigmoidal or exponential fits to extract characteristic lifetimes inherent to the transitions. Transition averages deviating from the shape expected for unperturbed one-dimensional bead diffusion are direct proof for traces of protein folding. The detection limit for protein folding transitions is determined by the fastest possible response of the measurement system which is defined by the typical relaxation time for unperturbed bead diffusion. By this means, apparently unperturbed transition averages automatically yield an upper limit for protein folding times. Further, [HMM-centred](#) averaged transitions reveal features of the convolved bead potential which, in turn, reflects basic properties of the underlying protein folding energy landscape.

*Anisotropic friction is a troublemaker*

The friction inherent to a force-spectroscopic system which is used to detect protein folding is typically different from the friction involved in the actual protein folding process. Consequences arising from this anisotropy with respect to rates, transition path times, and transition path ensembles are studied in [Section 4.3.2](#). In general, a correct detection or prediction of the transition path times of a protein is not possible. The only exception is observed close to the 'sweet' spot standing out due to an effectively isotropic friction. In contrast to transition path times, protein folding rates are correctly detected at all times. However, since the slowest component of a force-spectroscopic system acts as a low-pass filter, detected protein rates only conform with their prediction if the protein experiences higher friction than the bead. Two-dimensional representations of transition path ensembles in [Figure 30](#) illustrate how anisotropic friction shifts transition paths away from their energetically favoured routes. The degree of shifting stands in direct relation to differences in friction and barrier height. Altogether, this leads to one central finding: transitions exposed to anisotropic friction do not meet expectations derived from equilibrium. And further, the predictive capabilities of one-dimensional diffusion models are strongly limited in case of two-dimensional diffusion with anisotropic friction.

In [Section 4.3.4](#) two-dimensional diffusion simulations are used as a test bed for transition path analysis tools. At first sight, the tools fail to meet expectations risen by the excellent performance shown for one-dimensional diffusion in [Section 3.8](#). Only in case of isotropic friction, techniques for reconstructing barrier shapes or localising transition states yield acceptable results. At second sight, transition path analysis tools are very sensitive indicators for deviations from unperturbed bead diffusion and, hence, give proof for the presence of protein folding transitions. While reaction coordinate quality tests do indeed give an idea of the extent to which the measured deflection signal overlaps with the signal of interest, i. e., actual protein diffusion, a transition

path probability of 0.5 does not automatically mean that the measurement coordinate is a good representation of the signal of interest.

*Ideas for improvement and potential applications*

This chapter proves that simulations are a mighty tool for testing theories and evaluation techniques for their applicability and informative value within a controlled environment. To further improve precision of the two-dimensional diffusion simulation especially with respect to speed limit protein folding one can consider to split the deflection coordinate ' $x$ ' up into the two contributions  $x_1$  and  $x_2$  of a dual-trap setup. Besides eventually yielding an even more realistic picture, the influence of differences in trapping strength could be studied.

Since detected transition path ensembles are found to continuously change along with the protein friction coefficient  $\gamma_P$  they could be used as a measure of the latter. A successful attempt to do so is presented in [Section 6.4.2](#). Moreover, in this context, a variation of trapping strength or linker parameters may be used for a controlled manipulation of the system's Hamiltonian  $H(x, L_P)$ . The calculation of expected two-dimensional off-equilibrium transition paths based on the shape of  $H(x, L_P)$  and the friction coefficients  $\gamma_B$  and  $\gamma_P$  is the final missing link for bringing into play an entire two-dimensional theoretical model to study measured transition paths.



## Part IV

### EXPERIMENTAL RESULTS

The experimentalist's fun-part - a two-course menu: How simple things can be useful. How complicated simple can be.



## FERREDOXIN-LIKE FOLD: MECHANICS OF A TWO-STATE FOLDER

---

This chapter introduces results from single-molecule force spectroscopic measurements performed on an artificial protein that was designed to be mechanically labile. Along with the mechanical characterization, the measurement modes presented in [Section 2.3](#) and the evaluation of obtained data based on models described in [Chapter 3](#), are showcased. Data from both measurement modes independently confirm the desired low stability of the re-designed Ferredoxin-like fold. Further, indications found for a deviation from simple two-state mechanics are discussed, which brings the crucial role of the transition state position into focus.

### 5.1 A RE-DESIGNED IDEAL PROTEIN FOR TISSUE ENGINEERING

Owing to their potential applicability in tissue engineering, protein-based hydrogels have begun to attract considerable attention over the past years [135]. Inspired by biological materials like, e. g., muscles, the idea behind incorporating proteins into a hydrogel is the incorporation of sacrificial bonds to increase the overall toughness of the material [181, 208]. Simply speaking, when a protein-based hydrogel is stretched, the acting forces induce unfolding of the incorporated proteins which results in a massive dissipation of energy [220]. Via such a mechanism, the overall hydrogel can be prevented from being destroyed too easily. Furthermore, after the unfolded proteins refold upon relaxation, all initial mechanical properties of the hydrogel are restored.

A crucial step in engineering highly elastic and tough protein hydrogels is, of course, to find suitable proteins. Since, as a response to an externally induced overall deformation, hydrogels are expected to exert forces of only a few picoNewtons (pN) on the individual incorporated proteins, one important requirement for suitable proteins is a rather low mechanical stability [76]. If the proteins were too stable, they would not unfold upon stretching of the hydrogel and, hence, no energy could be dissipated to prevent breakage. A search for potential candidates was conducted with the help of protein design techniques. Protein design is aimed at finding an optimal amino acid sequence which folds into a desired, unique protein structure. Along with the optimization for attaining the desired structure, the absolute value of the folding free energy of the protein is usually maximized to energetically favour one distinct fold, see also [Section 1.3](#). This, however, interferes with the required low mechanical stability. The candidate that was found to meet all of the aforementioned requirements, is a re-designed ('red') Ferredoxin-like fold, hereafter referred to as FLred.

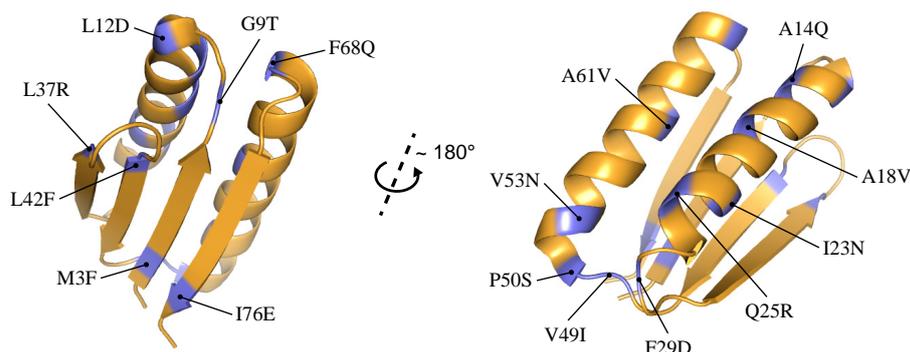


Figure 39: Positions of the 16 point mutations of the re-designed Ferredoxin-like fold (FLred) assigned to the structure of the original ideal fold (FL) (pdbID: 2KL8). The left and right show the same structure viewed from opposite sides.

In principle, FLred is a ‘negative’ design of a previously designed ideal protein which was optimized towards assuming the classical Ferredoxin-like fold (FL) [76, 115]. The re-design comprises 16 point mutations within the sequence of FL; exact sequences are given in Section A.1. In Figure 39 the mutations and their positions are indicated within FL’s solution NMR structure. Even though no crystal or NMR structure has been determined for FLred, its tertiary structure is predicted to be the same as for FL. This appears reasonable regarding almost 80% sequence identity as well as the fact that all mutations are chosen such that they still fulfil the basic design rules that were used to create FL. In contrast to the expected same shape, bulk measurements reveal an enormous drop in folding free energy from 9.1 kcal/mol for FL down to 3.8 kcal/mol for FLred<sup>1</sup>. To understand, how this dramatic drop could be achieved to reach the desired low mechanical stability for FLred, one has to look at the point mutations in more detail. In brief, the mutations reflect a weakening of the hydrophobic effect and an increase of solvent contact area, both leading to destabilization. This can be deduced from the fact that almost all mutations replace hydrophobic side-chains with bulkier, polar or even charged side-chains. As the mutations are located especially at transitions between secondary structural motifs and within surface exposed loops, negative effects on the original fold are minimised.

Whether or not FLred really assumes the predicted Ferredoxin-like fold remains speculative as long as no crystal or NMR structure has been determined. In contrast, the mere fact of being mechanically less stable by more than a factor of 2 as compared to the original ideal protein gives rise to another concern: undesired additional energy minima with a depth similar to the global energy minimum. To rule out the possibility of kinetic traps, wrong

<sup>1</sup> At a temperature of 303 K, the values given in kcal/mol correspond to 15.1 and 6.3 k<sub>B</sub>T for FL and FLred, respectively.

folds, or off-pathway intermediates arising, a mechanical characterization on a single molecular level is needed to address the following questions:

- Does FLred completely fold?
- Is there a simple and reliable unfolding/folding mechanism?
- How fast can FLred – if at all – refold against force to restore the gel’s mechanical properties?
- Can the folding free energy from bulk be confirmed?

The data from single-molecule measurements obtained to answer these and more questions are presented in the next section.

## 5.2 FROM SIMPLE TOWARDS COMPLEX TWO-STATE MECHANICS

### 5.2.1 *Slowing down: approaching equilibrium*

The mechanical characterization of a new protein starts in constant velocity mode to look for the protein’s fingerprint which comprises typical unfolding and refolding pattern(s). Figure 40 shows typical fingerprints of FLred recorded at different pulling velocities.

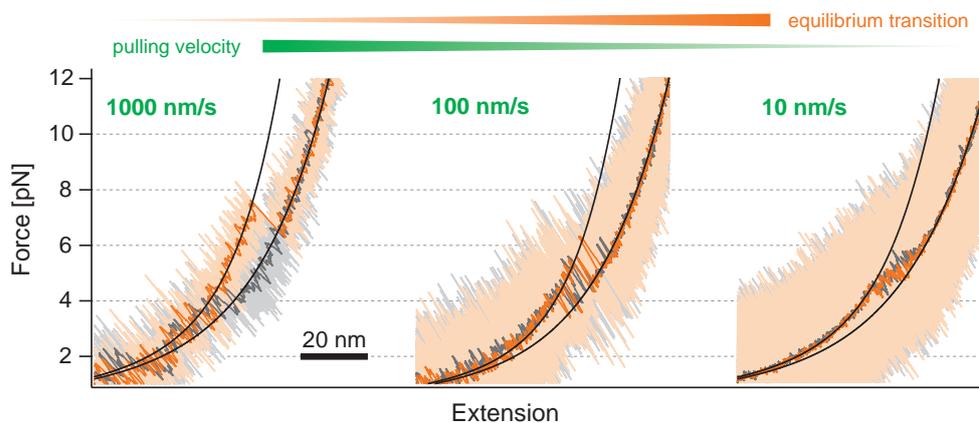


Figure 40: Typical force-extension traces of FLred recorded at 1000, 100, and 10 nm/s pulling speed. Stretch cycles are shown in orange, relax cycles in gray. Light colours show full resolution data, dark colours a smoothed version. Black lines are mechanical fits to the folded and unfolded state.

By decreasing the velocity from 1000 down to 10 nm/s, one can observe that the unfolding/refolding approaches a hump-like transition similar to what has been reported for the ultra-fast folding villin headpiece [9, 261]. At 100 nm/s one can repeatedly observe unfolding followed by refolding and vice versa within both the stretch and relax cycle. This indicates that FLred is able to refold against force at a rate of at least some tens per second. Furthermore, the fingerprints reveal that the transition mid-force, at which the

protein spends half the time in the folded and the other half in the unfolded state, is positioned at about 5 pN.

To address the question whether FLred is completely folded, we compare measured contour length gains with the value expected upon unfolding of the fully folded protein. In the solution NMR structure of the original ideal protein the N- to C-terminal distance of the folded structure, which comprises 76 amino acids, equals 0.49 nm. Given these numbers, we end up with an expected contour length gain of  $L_p^{\text{calc}} = 27.25$  nm according to Equation 5. This value compares remarkably well with the average measured contour length gain<sup>2</sup> of  $L_p^{\text{CV}} = 27.1 \pm 1.5$  nm. The corresponding histogram of contour length gains is shown in Figure 41C. Thus, our measurements clearly suggest that FLred is completely folded. Note that the contour length gains were obtained using a persistence length of  $p_p = 0.5$  nm to model the unfolded polypeptide chain.

After collecting unfolding and refolding forces from multiple stretch-and-relax cycles, the built-up statistics allow the extraction of force-dependent unfolding and refolding rates according to Oberbarnscheidt's method presented in Section 3.4.2. To do so, data were collected and pooled from cycles performed at 100, 500, and 1000 nm/s. Note that it is important to read off the *initial*, i. e., the first unfolding event within a stretch as well as the first refolding event within a relax cycle to create meaningful statistics. The resulting force-dependent rates are given in Figure 41A. The intersection where unfolding and refolding rates equal each other is confirmed to lie at 5.1 pN. Here, both rates equal about  $40 \text{ s}^{-1}$ . From SB-model fits to the force-dependent rates, the transition state position is determined to be at  $\Delta x_u = 10.4 \pm 1.7$  nm for unfolding and  $\Delta x_f = 16.1 \pm 1.1$  nm for refolding. Both lengths add up to  $26.5 \pm 2.0$  nm which, within error, is in perfect agreement with the contour length gain expected and observed for FLred. This suggests that the reaction coordinate of folding and unfolding of FLred is very well described by our measurement coordinate of changing bead deflection. Note that this is not necessarily the case for any protein. The deflection signal reflects contour length changes of the investigated protein as a response to an acting force. The value of this acting force is measured along the axis of two attachment points. How the acting force vector 'splits up' into forces 'felt' by individual parts of the protein remains speculative. Here, in case of a two-state folder with the attachment points being at N- and C-termini, the reaction coordinate and the measurement coordinate can be expected to nicely overlap.

To directly assess the folding free energy of FLred we can make use of its close-to-equilibrium unfolding/refolding transition at very slow pulling velocities where stretch-and-relax cycles average out into hump-like transitions which perfectly superimpose. By applying the fit described in Section 3.3 to multiple slowly pulled cycles we obtain  $\Delta G_0^{\text{EQ}} = 5.7 \pm 0.6 k_B T$ . A representative trace and fit are shown in Figure 41B. Indirectly, the folding free energy

<sup>2</sup> Here, the average contour gain is deduced from WLC-Fits to  $N = 125$  force-extension traces from six molecules. A persistence length of  $p_p = 0.5$  nm was chosen for the unfolded polypeptide chain. The average DNA-handle parameters are  $p_D = 24$  nm,  $L_D = 369$  nm, and  $K = 550$  pN.

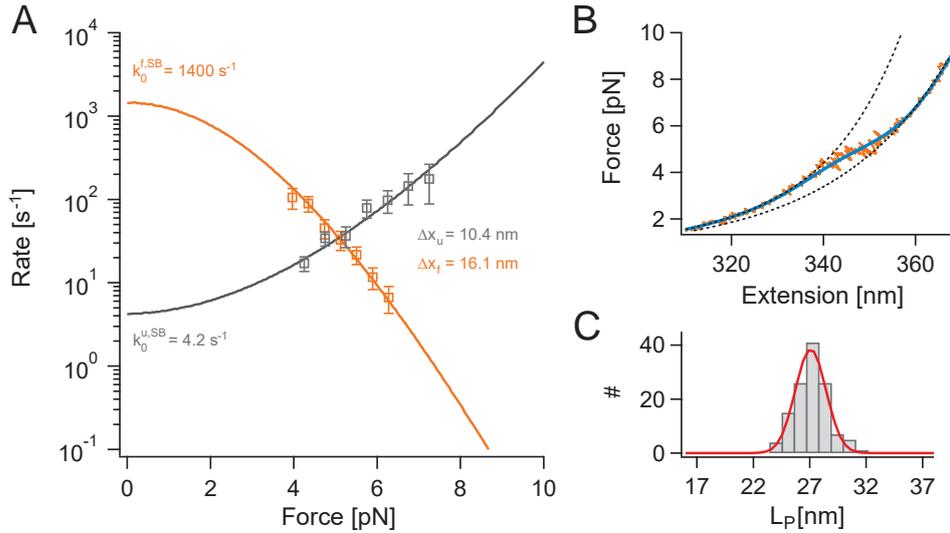


Figure 41: Energy landscape parameter reconstruction from data obtained in constant velocity mode. (A) Force-dependent rates (empty squares) derived with the OBS method and SB-Model fits (solid lines). (B) Equilibrium transition fit (blue line) to a slow 10 nm/s stretch cycle (orange). Dashed lines are mechanical fits to the folded and unfolded state. (C) Histogram of contour length gains and Gaussian fit.

can be estimated by exploiting the principle of detailed balance. Provided that the zero-force extrapolation of the SB-model applied to the force-dependent rates is correct, the logarithm of the rate constants' ratio should yield the desired folding free energy, see Equation 32. Here, we get  $\Delta G_0^{k_0, CV} = 5.8 \pm 0.6 k_B T$ , which is in perfect agreement with the previous result from the equilibrium transition fit.

Within all stretch-and-relax cycles, neither unfolding events with a strikingly different contour length, nor an accumulation of successive unexpectedly low unfolding forces could be observed. From this point of view, wrong folds or kinetic switches can be excluded. As a result FLred appears to be a perfect two-state folder.

### 5.2.2 Varying distance: shifting equilibrium populations

The constant velocity data suggest equilibrium fluctuations of FLred around 5 pN. Examples of time traces showing these fluctuations at constant trap distance are given in Figure 11. The time traces' state assignment is done by using HMM under the assumption of simple two-state behaviour. Examples of extracted integrated lifetime histograms at different force biases are shown in Figure 42A. The histograms are well fitted by single exponentials which is strong evidence for the assumed underlying two-state folding mechanism.

The rates derived from the exponential fits are plotted against force<sup>3</sup> in [Figure 42B](#). In [Figure 42C](#), the corresponding state occupancies are shown.

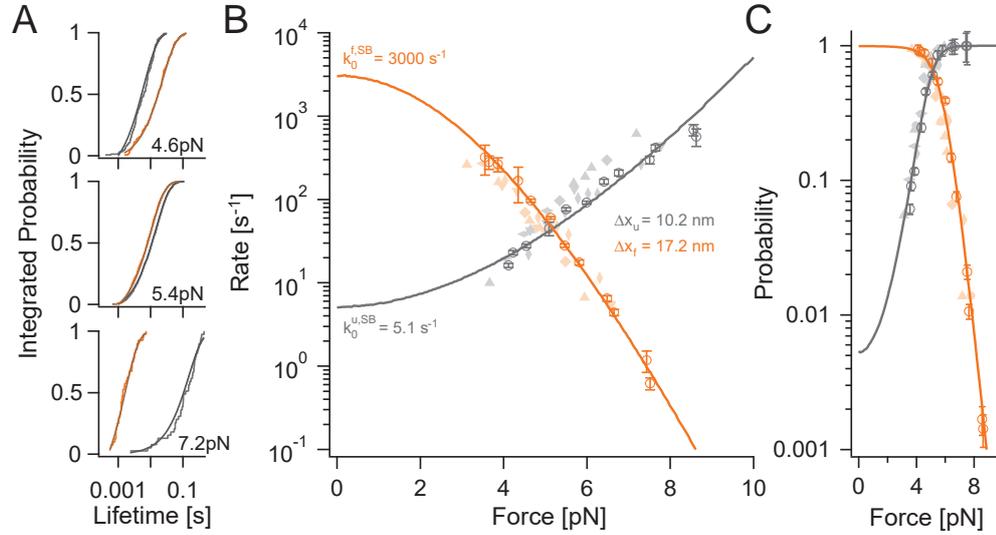


Figure 42: Energy landscape parameter reconstruction from data obtained in constant distance mode. (A) Integrated lifetime histograms (lighter colours) and single-exponential fits (darker colours) at different force biases. (B) Force-dependent rates (empty circles) of one single molecule. Lines are SB-Model fits. (C) State occupancy probabilities (empty circles) and fits (lines). For comparison, (B) and (C) are complemented with rates and probabilities from further molecules (differently shaped points in light colours).

The force-dependent change in deflection of each time trace upon unfolding and refolding can be transformed into a contour length change, see [Section 3.1.1](#). The average contour length gain from all constant distance traces which contribute to [Figure 42B](#) is  $L_p^{CD} = 27.2 \pm 1.1$  nm. With respect to the above constant velocity data, this result independently confirms complete folding of FLred.

By fitting the SB-model to force-dependent rates, the transition state position can be determined. The sum of  $\Delta x_u = 10.2 \pm 0.2$  nm and  $\Delta x_f = 17.2 \pm 0.5$  nm from the particular fits shown in [Figure 42B](#) perfectly matches all measured as well as calculated values for the entire protein contour length. At the intersection of the two exemplary SB-model fits, here we find a common rate of  $45 \text{ s}^{-1}$  at 5.2 pN.

There are several ways to determine the folding free energy of a protein from constant distance data. As suggested by [Equation 13](#) in [Section 3.2](#), folding free energies can be calculated for each time trace individually. Here, the average from all analysed trajectories is  $\Delta G_0^{CD,indiv} = 5.8 k_B T$ . Further, the folding free energy can be obtained from different molecules by global fits to their state occupancies according to [Equation 15](#). This way, an average of  $\Delta G_0^{CD,glob} = 5.5 k_B T$  is reached. Finally, considering detailed balance, the zero-force rate

<sup>3</sup> The forces indicated in [Figure 42](#) refer to the force acting on the state out of which the final state of a transition is reached.

constants resulting from fitting the SB-model to the individual molecule highlighted in Figure 42B return  $\Delta G_0^{k_0,CD} = 6.4 k_B T$ . For all calculated energies, here, the estimated error is 10 %. Due to the best statistics among all derived energies,  $\Delta G_0^{CD,indiv} = 5.8 \pm 0.6 k_B T$  is further considered as the final experimentally determined folding free energy value of FLred.

### 5.2.3 How well do different analysis methods agree?

‘There is more than one way to skin a cat.’, says a common proverb. In the context of collecting and analysing force spectroscopic data of protein folding this proverb holds as well. The special properties of the mechanically labile FLred with a fast combined rate  $\lambda_{mid}$  around the mid-force allows us to compare results from non-equilibrium constant velocity experiments with constant distance measurements at equilibrium within a *common* force range. This overlapping force range manifests itself in the near-equilibrium fast unfolding/refolding transition of FLred at very slow pulling velocities.

Table 3 lists contour length gains and folding free energies calculated from data obtained in either constant velocity or constant distance mode using different fitting procedures. While contour gains are crucial for a structural interpretation of observed unfolding and refolding events, folding free energies are useful to compare single-molecule results with bulk measurements. The corresponding discussion of measured and expected values can be found in Section 5.2.3.1.

METHOD	Text reference	$L_p$ [nm]	$\Delta G_0$ [ $k_B T$ ]
Const. Velocity	$L_p^{CV}$ & $\Delta G_0^{Crooks}$	$27.1 \pm 1.5$	n.d.
Const. Velocity	$L_p^{EQ}$ & $\Delta G_0^{EQ}$	$27.4 \pm 0.7$	$5.7 \pm 0.6$
Const. Distance	$L_p^{CD}$ & $\Delta G_0^{CD,indiv glob}$	$27.2 \pm 1.1$	$5.8   5.5 \pm 0.6$
Reference values	$L_p^{calc}$ & $\Delta G_0^{bulk}$	27.25	6.3

Table 3: Contour gains and folding free energies from different analysis methods and modes.  $L_p^{CV}$  is derived from WLC-fits to the folded and unfolded state.  $\Delta G_0^{Crooks}$  was not determined but is included for the sake of completeness. Values in the second row stem from equilibrium unfolding/folding transition fits to 10 nm/s traces.  $\Delta G_0^{CD}$  is calculated from individual constant distance trajectories and, in case of enough data points per molecule, from global probability fits.  $L_p^{CD}$  is derived through contour length transformation.

One can recognize strong agreement between the contour length gains and folding free energies listed in Table 3 even though different fitting methods as well as measurement modes were used to obtain these values. In Table 4 kinetic parameters resulting from SB-model fits to force-dependent rates are compared to each other. Here, the rates were either extracted from constant velocity data based on Oberbarnscheidt’s method or from constant distance

data using [HMM](#) analysis. The total accord between the kinetic parameters based on data from equilibrium fluctuations and non-equilibrium pulling experiments emphasizes the robustness especially of Oberbarnscheidt’s method. Taken together, the excellent agreement between different measurement modes and analysis methods confirms the possibility of using and combining all of them especially in cases where force ranges do not overlap and, hence, one or the other method may not be applicable.

METHOD	$\log_{10} k_0^f$ [s <sup>-1</sup> ]	$\log_{10} k_0^u$ [s <sup>-1</sup> ]	$\Delta x_f$ [nm]	$\Delta x_u$ [nm]	$\Delta G_0^{k_0}$ [k <sub>B</sub> T]
OBS rates (CV)	3.2 ± 0.1	0.6 ± 0.2	16.1 ± 1.1	10.4 ± 1.7	5.8 ± 0.6
HMM rates (CD)	3.5 ± 0.1	0.7 ± 0.1	17.2 ± 0.5	10.2 ± 0.2	6.4 ± 0.6

Table 4: Kinetic parameters from measurements in constant velocity (CV) and constant distance (CD) mode. Given values result from fitting the SB-model to force-dependent rates extracted from either CV data using the OBS method or from CD data using [HMM](#) analysis.

For completeness, [Table 4](#) also includes the  $\Delta G_0^{k_0}$  values which are the free energy values derived from the zero-force extrapolated rate constants based on the principle of detailed balance, see [Equation 32](#). The perfect agreement with all other values obtained is very good evidence for a consistent data set and analysis.

### 5.2.3.1 Comparing single-molecule results to calculated values and bulk measurements

While the measured contour gains match the value calculated from the solution [NMR](#) structure of FL perfectly, the single-molecule folding free energies slightly deviate from bulk measurements. Even though the bulk value of  $\Delta G_0^{\text{bulk}} = 6.3 \text{ k}_B\text{T}$  is recovered within error, most single-molecule results show a tendency of being a bit too small, see [Table 3](#) and [Table 4](#) in [Section 5.2.3](#). A simple explanation could be a possible destabilization of FLred through the presence of an additional GB1 protein at the N-terminus of the entire protein construct that was used for the mechanical measurements at the optical trap. In contrast, this additional GB1 was not yet included in the construct used for chemical denaturation experiments. For a subdomain of calmodulin an energetic destabilization of 7% has been reported due to N- and C-terminal ubiquitins [[212](#)]. As ubiquitins share a same fold with GB1, here we might observe a similar effect for FLred.

Another reason for a lower folding free energy value from single-molecule mechanical measurements as compared to bulk may be rooted in the persistence length  $p_p = 0.5 \text{ nm}$  which was chosen to model the unfolded polypeptide chain. A bigger persistence length would result in a bigger folding free energy along with a shorter overall contour length gain. While the perfect match between measured and calculated contour lengths argues against a

bigger persistence length, we have to consider that this may also be misleading since the calculation of an expected contour length, see Equation 5, is based on an assumed average contour length per amino acid  $d_{aa}$  which in turn is affected by the choice of  $p_p$ . Due to the very small folding free energy we cannot completely exclude that a small fraction of FLred potentially unfolds at very low forces below our detection limit. If that was the case, a shorter measurable contour length gain would be the consequence. Additionally, as long as no crystal or NMR structure of FLred exists, the assumption of completely attaining the Ferredoxin-like shape of FL is only based on predictions. A solution to finding the correct  $p_p$ -value could be to extract the mere protein contour gain contribution from the difference between folded and unfolded parts along stretches of overlapping forces within force-distance traces as described in [257]. Unfortunately the very narrow force range for overlaps at small forces does not allow a clear statement on which persistence length fits the data best to be made. However, a tendency towards a slightly higher persistence length of about  $p_p = 0.6$  nm seems possible (data not shown).

#### 5.2.4 Deviations from simple two-state mechanics at low forces

The mechanical characterization of FLred in Section 5.2.1 and Section 5.2.2 clearly revealed the behaviour of a simple two-state folder. However, for the derivation of force-dependent rates from constant velocity measurements, all unfolding and refolding events that occurred below 4 pN were yet not considered. The data shown in Figure 43 by contrast, include the entire available data set. As can be seen in Figure 43A, the additional points below 4 pN mark a considerable deviation from the SB-model, which was only fitted to rates at forces above 4 pN yielding the same result as already shown in Figure 41A.

On the left of Figure 43C unfolding and refolding rates are fitted separately for forces below and above 4.5 pN using the ZB-model. For lower forces, the transition state position is located at  $\Delta x_{u,low}^{ZB} = 7.6 \pm 1.0$  nm and  $\Delta x_{f,low}^{ZB} = 2.2 \pm 0.8$  nm. For higher forces, the transition state appears much closer to the native state at  $\Delta x_{u,high}^{ZB} = 3.4 \pm 0.4$  nm and  $\Delta x_{f,high}^{ZB} = 5.8 \pm 1.2$  nm. Within each force range, the data are fitted very well by the ZB-model. However, the strong divergence between both force ranges implies a force-dependent change of the predominant barrier position.

Figure 43B shows ZB-model-based fits to unfolding force histograms retrieved from data recorded at pulling speeds of 100, 500, and 1000 nm/s. As the unfolding forces increase along with the pulling velocity one can observe the same transition from a more distant towards a much closer transition state with respect to the native conformation. This confirms the indication of a force-dependent change of the predominant barrier position along the unfolding pathway. In principle, unfolding or refolding force histograms are nothing but a different representation of force-dependent rates. Figure 43D shows force-dependent unfolding rates directly derived from unfolding force histograms according to Equation 21. As expected, the rates scatter around the ZB-model fits obtained from the corresponding OBS rates. This agree-

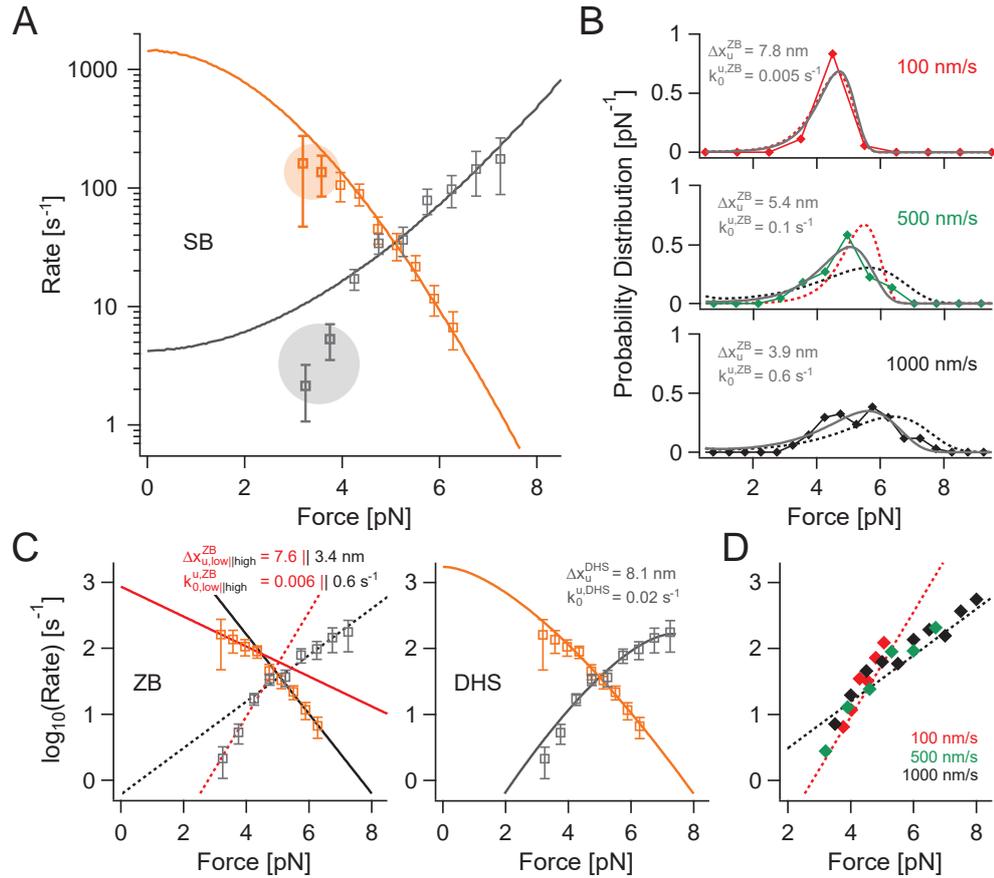


Figure 43: Deviations from simple two-state mechanics at low forces. (A) OBS rates including forces below 4 pN. Lines are SB-model fits to data above 4 pN. (B) Unfolding force distributions fitted by Equation 35 (gray lines). Dotted lines are calculated distributions inferred by the ZB-model fits in (C). (C) **Left:** Full and dashed lines are ZB-model fits at lower (red) and higher (black) forces. **Right:** Lines are DHS-model fits to the upper force range. (D) Rates derived from unfolding force histograms according to Equation 21. Dashed lines are ZB-model fits from (C).

ment is again proof for a consistent data set and an example for the many interchangeable ways of analysing force spectroscopic data.

Finally, also the DHS-model is applied<sup>4</sup> to fit unfolding and refolding rates as shown on the right of Figure 43C. When the upper force range of measured unfolding rates is considered, the DHS-model yields  $\Delta x_u^{DHS} = 8.1 \pm 0.9$  nm,  $k_0^{u,DHS} = 0.02 \pm 0.02$   $s^{-1}$ , and  $\Delta G_{0,TS}^{u,DHS} = 10.5 \pm 1.6$   $k_B T$ . While the transition state is located very close to  $\Delta x_{u,low}^{ZB} = 7.6$  nm obtained with the ZB-model at forces below 4.5 pN, the zero-force extrapolated rate constant lies between the two ZB-model values for lower and higher forces. Fitting the upper force range of folding rates with the DHS-model results in  $\Delta x_f^{DHS} = 2.2 \pm 0.2$  nm,  $k_0^{f,DHS} = 1700 \pm 500$   $s^{-1}$ , and  $\Delta G_{0,TS}^{f,DHS} = 0.3 \pm 1.4$   $k_B T$ . Again, the transition

<sup>4</sup> Here, the scaling factor in Equation 31 is set to  $\nu = 2/3$  which assumes an underlying potential that contains linear and cubic terms.

state position for folding matches the low force value  $\Delta x_{f,\text{low}}^{\text{ZB}} = 2.2 \text{ nm}$  of the ZB-model. Further, the obtained zero-force rate constant for folding corresponds well to the SB-model fit:  $k_0^{f,\text{SB}} = 1400 \pm 400 \text{ s}^{-1}$ . However, a barrier height of less than  $3 k_{\text{B}}T$  indicates that the DHS-model is applied beyond its validity.

Compared to the SB- and ZB-model fits, here, the curvature of the DHS-model fit captures the overall tendencies within the data best. This suggests that the force-dependent change of the predominant barrier position to some extent may be owed to Hammond behaviour. Albeit, none of the three models is able to reproduce the kinetics of FLred over the entire force range by means of a single two-state model fit.

### 5.3 DISCUSSION

#### 5.3.1 Force-induced transition state movement: confirming Hammond behaviour by deconvolution

In [Figure 43](#), the best fit over the entire force range of unfolding rates of FLred is provided by the DHS-model. The curvature of the fit reflects Hammond behaviour: shifting a reaction towards the product side is accompanied by a transition state movement towards the reactants side [94]. Hence, for unfolding in a force spectroscopic assay: with rising force, the transition state moves towards the native state which increasingly resembles the unfolded denatured state by being more and more destabilized [107]. In contrast to the DHS-model fits for unfolding, corresponding SB-model fits show an opposite curvature. Here, the transition state position is fixed to a certain contour length, adapting a force-dependent extension which increases with force according to a WLC-model, see [Section 3.5.3](#). In case of refolding, both the curvatures of the DHS-model and the SB-model, are in accord with Hammond behaviour: the lower the force the more the native configuration is stabilized and easier 'to reach' from the unfolded state, which is reflected by a transition state movement towards the unfolded state. This is one reason for the good agreement between the zero-force extrapolated folding rates provided from both models. However, in case of the SB-model, the apparent accord with Hammond behaviour is simply owed to the fact that with decreasing force the extension of the WLC-modelled but otherwise fixed  $\Delta x_f$ , decreases, too.

In this section, deconvolution is used to analyse to which extent Hammond behaviour is involved in the measured mechanics of FLred. [Figure 44A](#) shows an example of an energy landscape deconvolved from equilibrium fluctuations of FLred recorded at a pretension of 6.4 pN. The transformation of this energy landscape to lower as well as higher pretensions reveals a considerable force-induced transition state movement as expected for Hammond behaviour. The quality of the transformation is approved by very good agreement between transformed landscapes and directly deconvolved ones as depicted in [Figure 44B](#). Importantly, this approval additionally confirms the revealed transition state movement to at least be valid within the entire range

accessible to equilibrium fluctuation measurements, i. e., within 4 to 8 pN. Further, the zero-force extrapolated energy landscape shows a folding free energy of  $\Delta G_{0^*} = 6.0 k_B T$  which perfectly complies with all previous results listed in Table 3. This is another proof for a meaningful deconvolution and energy landscape transformation.

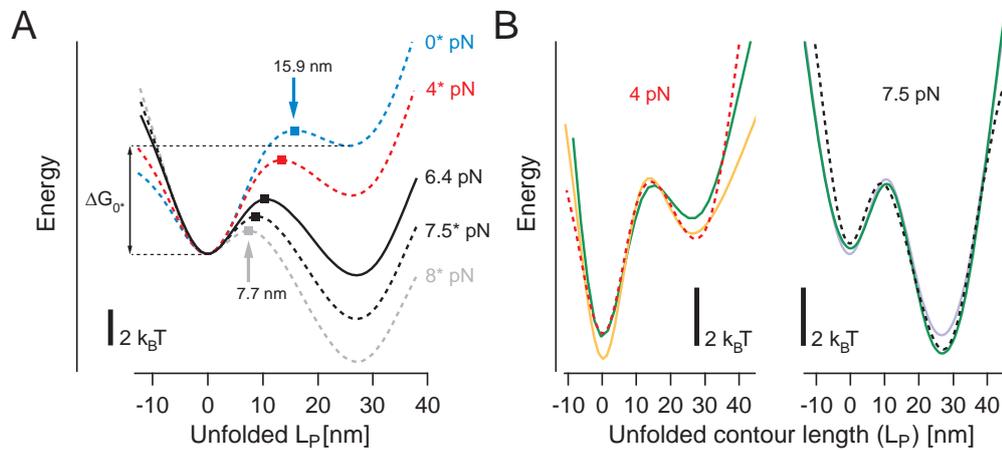


Figure 44: Transition state movement revealed by deconvolution and energy landscape transformation. (A) Energy landscape of FLred deconvoluted at 6.4 pN pretension on the folded state (solid line) and transformed to other starred pretensions. The zero-force energy landscape yields  $\Delta G_{0^*} = 6.0 k_B T$ . (B) Comparison of transformed landscapes (dashed lines) from (A) to landscapes directly deconvoluted from traces measured at the indicated pretension. The good agreement proves the quality of energy landscape transformation along with the deduced transition state movement.

At this point, the results obtained from deconvolution and energy landscape transformation clearly indicate the presence of Hammond behaviour as something that needs to be considered when analysing the mechanics of proteins. In case of FLred, the small size and the low folding free energy in combination with an almost vanishing folding barrier at zero force, enhance force-induced transition state movement within a relatively small force range. Whether Hammond behaviour can explain the kinked chevron plots of FLred and the implications on the prediction capabilities of different kinetic models are discussed after the following passage as well as in Section 5.3.2.

Given the shape of the zero-force extrapolated energy landscape and the extrapolated zero-force rate constant, it is possible to deduce the internal friction coefficient of FLred under force-free conditions. After solving Equation 29 from Kramers rate-theory for  $\gamma$  we reach at  $\gamma_{\text{FLred}} = 5.1 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . For the calculation  $k_0^{f, \text{SB}} = 3000 \pm 400 \text{ s}^{-1}$  was used as obtained by fitting the SB-model to constant distance data, see Table 4 and Figure 42B. The resulting friction coefficient compares very well with values reported for other constructs, see Figure 29B in Section 4.3.2, suggesting the friction for FLred to be on the rather low side. However, this result needs to be handled with care since its derivation is sensitive to the model used for the zero-force rate ex-

trapolation, the considered coordinate space (here unfolded contour length), as well as additional friction involved in the measurement, see [Section 4.3.2](#).

*Can Hammond behaviour explain the kinked chevron plots of FLred?*

The transition state movement revealed by the transformed energy landscapes in [Figure 44A](#) starts from an unfolding barrier position of  $\Delta x_{u,opN}^{decon} = 15.9$  nm at zero force. With increasing force, the distance of the unfolding barrier decreases according to Hammond behaviour. At 4 pN, the transition state is only  $\Delta x_{u,4pN}^{decon} = 13.5$  nm away from the native state. This distance in unfolded contour length corresponds to about half the entire contour length gain expected for FLred and, hence, indicates a perfectly symmetric energy landscape in terms of transition state position. A further increase of force up to 8 pN reduces the distance between the native and the transition state down to  $\Delta x_{u,8pN}^{decon} = 7.7$  nm of unfolded contour length. By this means, the unfolding barrier distance decreases down to about half its value upon an increase of force up to 8 pN.

To judge, whether the transition state movement uncovered by deconvolution can explain the kinked chevron plots of FLred shown in [Figure 43](#), here, the ratio between unfolding and folding barrier distances  $\Delta x_u/\Delta x_f$  is introduced as a measure of the relative transition state position. This value allows a simplified comparison between barrier positions retrieved in different ways.

	METHOD					
	Deconvolution			SB-model	ZB-model	
Force [pN]	0	4	8	4 - 8	3 - 4.5	4.5 - 8
$\Delta x_u/\Delta x_f$	<b>1.4</b>	1.0	0.4	0.6	<b>3.5</b>	0.6

Table 5: The relative transition state position of FLred depends on force. Values listed for deconvolution and the SB-model are based on time trajectories and fits to [HMM](#) rates from constant distance data, see [Section 5.2.2](#). ZB-model values are deduced from fits to OBS rates extracted from constant velocity data, see [Section 5.2.4](#).

In [Table 5](#), the force-dependence of the relative transition state position is specified for different analysis methods and data pools. For deconvolution, a decreasing ratio along with an increasing force directly reflects the observed force-induced Hammond behaviour. The value of  $\Delta x_u/\Delta x_f = 1.0$  at 4 pN indicates the above mentioned symmetry where  $\Delta x_u \sim \Delta x_f$ . The values obtained from the SB- and ZB-model are given for their respective fitting range. A value of 0.6 supplied by fitting the SB-model to [HMM](#) rates extracted from constant distance data measured at forces ranging from 4 to 8 pN lies right in between the values returned by deconvolution at the edge of this force range. This result indicates two things. First, as the transition state positions extracted from deconvolution and the SB-model are both measured in changes of unfolded contour length, the dataset is self-consistent. Second,

over a certain force range, the SB-model obviously provides an averaged transition state position. The excellent agreement between the relative transition state positions indicated by the SB-model and the ZB-model within their overlapping force ranges confirms the suitability of the barrier distance ratio to compare transition state positions supplied by different models. Finally, there is an eye-catching discrepancy between the relative transition state position indicated by the deconvolved energy landscape that has been transformed to zero force and the relative position obtained from fitting the ZB-model to the lower force range of experimentally determined force-dependent rates. This deviation is strong evidence for the change in predominant barrier position, as implied by the kinked chevron plots in [Figure 43](#), to not be reproducible by mere Hammond behaviour.

Taken together, studying the force-dependence of the relative transition state position of FLred revealed the following. For forces above 4 pN, all methods are in agreement with the occurrence of force-induced Hammond behaviour. However, mere Hammond behaviour is not enough to explain the mechanics observed at forces below 4 pN. On this account, a more pronounced effect has to be considered as the reason for the kinked chevron plots of FLred which is further elaborated in [Section 5.3.3](#).

### 5.3.2 *Case study FLred: a comparison between the predictive capabilities of different kinetic models*

The mechanics of FLred reflect a simple two-state folder with superimposed Hammond behaviour at forces above 4 pN. For forces below 4 pN, a deviation from this simple two-state mechanism is observed. In this part, influences of these conditions on the quality of energy landscape parameter predictions by different kinetic models are investigated. Among the compared kinetic models are the ZB-, DHS-, and SB-model, which are all introduced in detail in [Section 3.5](#). The main differences between these models consist of the referenced reaction coordinate space, the considered force-induced effects on the returned transition state position, and the treatment of energetic contributions involved in the modelled kinetics.

In brief, the ZB-model assumes a fixed transition state position within a force-independent reaction coordinate space which basically reflects absolute changes within the measured deflection signal. The same reaction coordinate space is used within the frame of the DHS-model. However, the value of the transition state position returned by the DHS-model refers to a fixed zero-force position which changes according to force-induced Hammond behaviour. The SB-model assumes a fixed transition state position in unfolded contour length space. As the unfolded contour length is force-dependent, it is modelled with a [WLC](#) which effectively serves as an interconversion between the measured deflection signal and energetically relevant distances. By this means, the SB-model rescales the reaction coordinate space to adapt to experimental conditions, i. e., to force-dependent distances of the deflection signal, while the DHS- and the ZB-model assume a fixed and therewith

force-independent reaction coordinate scale. Concerning energetic contributions, the ZB- and the DHS-model consider an overall energy which changes linearly with force. In addition, the DHS-model includes a fixed zero-force transition barrier height which changes in response to an applied force. In contrast to the ZB- and DHS-model, the SB-model accounts for all linear and non-linear energetic contributions of the entire measurement system including beads, DNA-handles, and unfolded protein.

Table 6 gives an overview of the results<sup>5</sup> obtained from all SB-, ZB-, and DHS-model fits shown in Figure 43. Among the listed parameters are the relative transition state position  $\Delta x_u/\Delta x_f$  reflecting some sort of spatial resolution, the sum of the transition state distances  $\Delta x_u + \Delta x_f$  as an overall reaction coordinate space control, and the folding free energy  $\Delta G_0^{k_0}$  resulting from applying the principle of detailed balance to the zero-force extrapolated rate constants, see also Equation 32. The latter value can be considered as an indirect control parameter of a correct treatment of the involved energies.

	MODEL			
	ZB	DHS	ZB	SB
Force [pN]	3 - 4.5	4 - 8	4.5 - 8	4 - 8
$\Delta x_u/\Delta x_f$	3.5	<b>3.7</b>	<b>0.6</b>	<b>0.7</b>
$\Delta x_u + \Delta x_f$ [nm]	9.8	10.3	9.2	26.5
$\Delta G_0^{k_0}$ [k <sub>B</sub> T]	11.9	<b>11.3</b>	<b>10.8</b>	<b>5.8</b>

Table 6: Energy landscape parameter prediction capabilities of different kinetic models when applied to FLred data. All listed values are based on fits to OBS rates, see Section 5.2.1 and Section 5.2.4. When the principle of detailed balance is applied, only the SB-model yields a reasonable folding free energy.

When merely judging the quality of the fits in Figure 43 by eye, all three models nicely approximate the force-dependent rates of FLred for forces above 4 pN. The main reason for that is the restricted force range which is too small to allow pronounced curvatures to be observed within the experimentally determined rate plots. Hence, even the above confirmed force-induced Hammond behaviour, which results in a transition state movement of about 20 % with respect to the entire reaction coordinate space, does only marginally affect the curvature of the chevron plots of FLred between 4 and 8 pN. However, despite an overall good approximation of the data through all three models, the interpretation of their fit results presented in Table 6 reveals two striking contradictions. First, the relative transition state position provided by the DHS-model strongly deviates from the corresponding results of the ZB- and the SB-model. And second, the folding free energies derived from

<sup>5</sup> The zero-force extrapolated rates of the ZB-model used to calculate the  $\Delta G_0^{k_0}$  values in Table 6, are:  $k_{0,low}^{f,ZB} = 871 \pm 94 s^{-1}$ ,  $k_{0,low}^{u,ZB} = 0.006 \pm 0.005 s^{-1}$ , and  $k_{0,high}^{f,ZB} = 30000 \pm 14000 s^{-1}$ ,  $k_{0,high}^{u,ZB} = 0.6 \pm 0.3 s^{-1}$ . The corresponding fits are shown in Figure 43C.

the zero-force extrapolated rate constants of the ZB- and the DHS-model are about twice as big as the energy calculated for the SB-model which, in turn, is already known to be correct, see [Section 5.2.3](#).

As a consequence of neglecting force-induced Hammond behaviour, the SB- and ZB-model provide an averaged transition state position for forces between 4 and 8 pN, also see [Table 5](#). In contrast, the DHS-model is conform with Hammond behaviour and its provided transition state position refers to a zero-force extrapolated energy landscape. This is the simple reason why in [Table 6](#), the relative transition state position  $\Delta x_u/\Delta x_f$  from the DHS-model does not match the corresponding values from the other models which directly refer to the fitted force range. Yet, the barrier distance ratio of 3.7 supplied by the DHS-model is also in disagreement with the relative transition state position of 1.4 indicated by deconvolved energy landscapes transformed from within the fitted force range to zero force, see [Table 5](#). This suggests a zero-force extrapolation failure of the DHS-model. Obviously, for the given dataset of FLred, the inherent strength of the DHS-model, which consists of extrapolating force-dependent mechanics towards a force-free energy landscape, becomes a weakness. The deviation from a simple two-state folding mechanism at lower forces clearly means that FLred fails to meet the requirement of two-state behaviour at all forces, necessary for applying the DHS-model.

The sum of the transition state distances  $\Delta x_u + \Delta x_f$  is nearly the same for all fits obtained with the ZB- and the DHS-model even including different force ranges. This confirms a common and apparently force-independent reaction coordinate space for both models. For the SB-model, the sum of the transition state distances confirms the fully unfolded contour length of FLred to be the relevant reaction coordinate space for the conformational mechanics.

Finally, the strong deviation of the  $\Delta G_0^{k_0}$  values provided by the ZB- and DHS-model is evidence for their failure of an energetically adequate extrapolation<sup>6</sup> of the zero-force rate constants. Especially at forces below 10 pN, the overall measurement system's compliance consists of strongly non-linear and force-dependent contributions as illustrated in [Figure 15C](#) of [Section 3.6.1](#). Besides the wrong assumption of mere linear energetic contributions, the assumed force-independent reaction coordinate space is an oversimplification of the force-dependent distances measured within the deflection signals, particularly at very low forces.

In summary, out of all three models, the SB-model performs best in terms of its predictive capabilities. As already confirmed by deconvolution, the SB- and the ZB-model yield a correct relative transition state position average valid for within their fitted force range. Further, only the SB-model yields zero-force extrapolated rate constants which are conform with the principle of detailed balance. Concerning the DHS-model, in literature it was shown to successfully reproduce force-dependent unfolding rates of [DNA](#) hairpins

<sup>6</sup> The difference between the zero-force unfolding and folding barriers supplied by the DHS-model implies a folding free energy difference of  $\Delta G_{0,TS}^{u,DHS} - \Delta G_{0,TS}^{f,DHS} = 10.2 k_B T$ . This value is about consistent with the wrong energy suggested by the zero-force rate constants.

as well as proteins [66, 249]. However, as the results from the DHS-model refer to an experimentally inaccessible energy landscape at zero force, they give no direct insight into the fitted force range and, hence, need to be interpreted with care. As shown for the case of the yet rather simple FLred, any slight deviation from perfect two-state mechanics lets the DHS-model fail to make predictions that conform with any of the other applied models and methods. All in all, to characterize and understand the mechanics of a protein within an experimentally accessible force range, the results obtained with the SB-model are suited best and appear to not be negatively affected by Hammond behaviour.

### 5.3.3 Qualitative energy landscape reconstruction

As stated above, the observed deviation of FLred mechanics from simple two-state behaviour at low forces cannot be explained by mere Hammond behaviour. Then, if the unfolding and refolding of FLred is considered as a process in a strictly one-dimensional energy landscape with conserved end-to-end and force vector directions, this deviation must point at a force-induced switch in predominant transition barrier analogous to simulations in Section 4.2.4 as well as considerations in [83, 197]. A switch in transition barriers, in turn, implies the existence of an additional transiently populated intermediate state along the reaction coordinate. Interestingly, high-energy intermediates implied by curved chevron plots have already been reported for Ferredoxin-like folds [131, 163, 191]. In a recent computational study the loop length connecting the N- and C-terminal beta sheets with the alpha helices has been identified to determine the presence or absence of intermediate states. A longer loop length was found to reduce the overall cooperativity between the two ascertained folding units of FLred designated in Figure 45B and, hence, to favour sequential folding [219].

The qualitative energy landscape reconstruction, which characterizes mechanical features of FLred, is essentially done in two steps. At first, transition state positions are deduced from kinetic information. And second, a potential intermediate state is assigned upon structural considerations. From the relative transition state positions supplied by the ZB- and SB-model, the first transition state TS<sub>1</sub>, which dominates at higher forces, is located in contour space about 10 nm away from the native state and about 17 nm away from the unfolded state. At lower forces the second transition state TS<sub>2</sub> dominates unfolding and refolding. Assuming the sum of transition state distances to recover the entire contour length of FLred as is the case at higher forces, TS<sub>2</sub> is about 21 nm away from the folded state when considering the ZB-model<sup>7</sup>. The search for a possible candidate for an intermediate is illustrated in Fig-

<sup>7</sup> Note that the SB-model also yields a reasonable fit for refolding at low forces (fit not shown). The fit result indicates TS<sub>2</sub> to be positioned at  $\Delta x_{f,low}^{SB} = 7.8 \pm 1.3$  nm away from the unfolded state; the zero-force refolding rate constant is  $k_{0,low}^{f,SB} = 321 \pm 93$  s<sup>-1</sup>. This implies a distance of about 19.5 nm between TS<sub>2</sub> and the native state, which is in good agreement with the 21 nm retrieved from the relative transition state position supplied by the ZB-model.

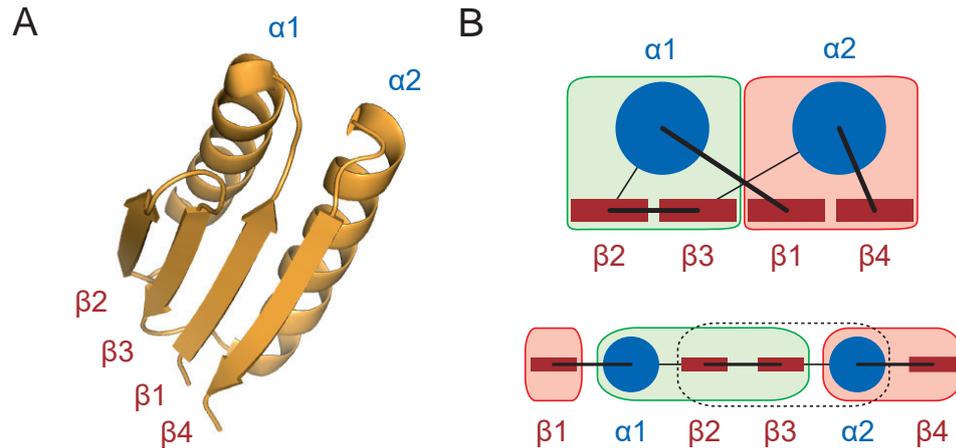


Figure 45: The two potential folding units of FLred. (A) Secondary structure elements of FLred assigned to the solution NMR structure of FL. (B) Schematic representation of the folded FLred structure (top) and the unfolded polypeptide chain (bottom). Reported folding units are highlighted. The folding unit  $\alpha 1\beta 2\beta 3$  should fold more readily out of a stretched conformation while, eventually, competing with the non-native conformation  $\beta 2\beta 3\alpha 2$ .

ure 46B, where one structural element of FLred is unfolded after the other. Here, only more reasonable configurations were allowed and projected onto their expected gain in contour space. This is why, e. g., unfolding is only considered to start from the C-terminal beta strand  $\beta 4$  located at the outer edge of the beta-sheet, as the more likely initiation of forced unfolding compared to ripping out the N-terminal beta strand  $\beta 1$  out of its central beta-sheet position. The most promising candidate for representing the intermediate state I<sub>12</sub> is the configuration where the central two beta strands  $\beta 2$  and  $\beta 3$  and the N-terminal alpha helix  $\alpha 1$  are folded. Together,  $\alpha 1\beta 2\beta 3$  form a folding unit that has been identified to be one of the two folding cores of Ferredoxin-like folds, where the second core consists of  $\beta 1\alpha 2\beta 4$  [219], see Figure 45B.

The positions of TS<sub>1</sub>, TS<sub>2</sub>, and of potential intermediates, are indicated within the schematic energy landscapes depicted in Figure 46A. The landscapes drawn with solid lines illustrate the force-induced switch in the main barrier position. While at forces below 4.5 pN TS<sub>2</sub> dominates (red), TS<sub>1</sub> becomes the prevailing transition state at higher forces (black). The energy well depth of the intermediate is as yet only putative. Gray dashed lines are energy landscapes transformed to zero and 8 pN after deconvolution, see Figure 44A, where transition state movement is merely based on Hammond behaviour and clearly does not recover the transition between TS<sub>1</sub> and TS<sub>2</sub>.

Along with the projection of possible unfolded configurations of FLred into the energy landscape picture, the folding pathway of FLred under force can be hypothesized (unfolding follows the reverse course of events). First, the folding core  $\alpha 1\beta 2\beta 3$  attempts to form in the transition state region around TS<sub>2</sub>. The intermediate state region I<sub>1</sub> is reached upon successful formation of the first folding core. In the transition state region around TS<sub>1</sub> the second

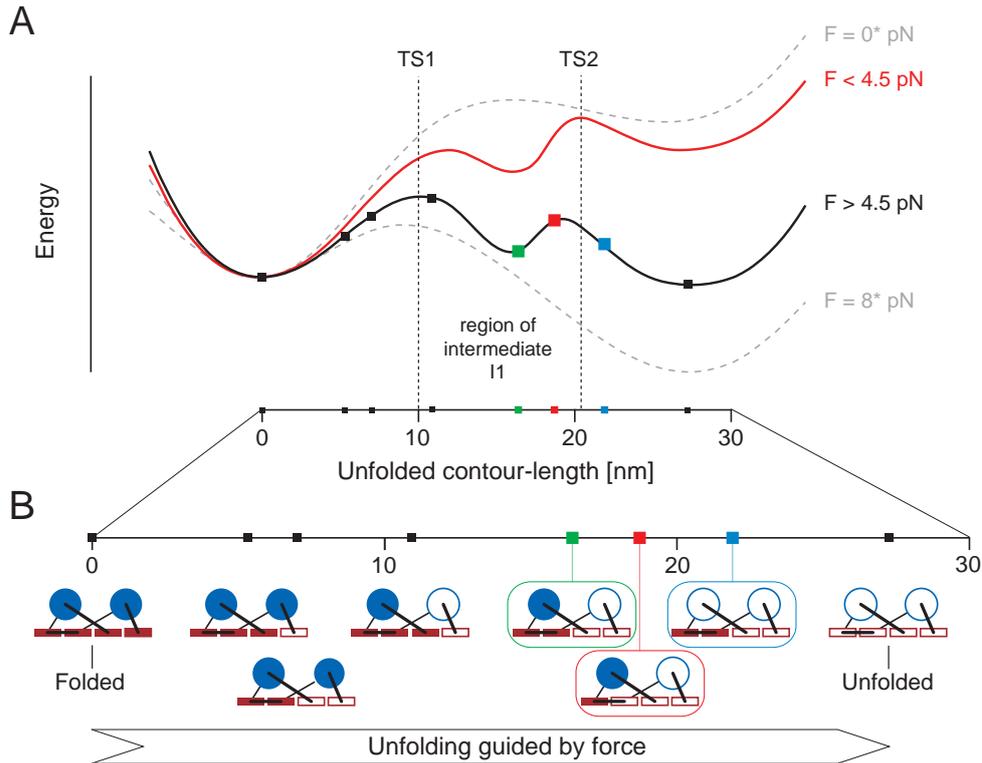


Figure 46: Qualitative energy landscape reconstruction for FLred. (A) Energy landscape of FLred before and after a force-induced transition state switch. The relative energy level of the transiently populated intermediate state  $I_1$  is hypothesized. For reference, gray dashed lines refer to the deconvolved and transformed energy landscape already shown in Figure 44A. (B) Force as a guide through conformational space. Different conformations of FLred are arranged along the axis of their expected unfolded contour-length.

folding core folds and FLred ends up in its final folded configuration. In this rather sequential view, force restricts cooperativity through confinement of configurational space which is accessible to the stretched polypeptide chain. Out of the two identified folding cores  $\alpha 1\beta 2\beta 3$  and  $\beta 1\alpha 2\beta 4$ , it appears reasonable to assume that the first one is more readily formed under a residual force as the involved structural elements follow one after the other within the protein's sequence, as indicated at the bottom of Figure 45B. Prior to formation of the second folding core, the N- and C-terminal ends need to get close enough to each other.

#### *Further considerations and perspectives*

Data below 4 pN have a rather poor resolution which makes the correct detection of unfolding and refolding forces difficult. To exclude systematic errors, improved statistics will strengthen the argument for a switch in transition barriers. Furthermore, data recorded at higher temporal resolution and

lower trap stiffnesses<sup>8</sup> than the up to now used 30 kHz bandwidth and 0.3 - 0.4 pN/nm of individual trapping strength may allow a more detailed analysis of individual unfolding and folding transitions. So far, only unfolding events at higher forces, where the folded and unfolded state separate more clearly, seem to directly give slight hints of a transiently populated intermediate state (data not shown). At higher forces, the deconvolution of some equilibrium traces indicate a subtle intermediate state located around the proposed position for the formation of the folding unit  $\alpha 1\beta 2\beta 3$  (data not shown). These yet rather vague observations support the occurrence of a force-induced cooperativity breakdown between the two folding units which form FLred.

In the context of protein design, an intentional transiently populated intermediate state can be interpreted as a guide through conformational space. Besides helping to avoid undesirable misfolding, an intermediate enables a protein to refold against higher forces as well [196]. In other words, the observed switch in barrier position may point at a roughened energy landscape where sequential folding allows higher mechanical stability to be reached at the cost of overall cooperativity. However, beyond the scenario of a one-dimensional energy landscape governed by sequential barrier transitions, alternative scenarios such as multiple paths or ground state effects need to be accounted for [192, 197, 246]. For this purpose complementary Monte Carlo simulations based on mere kinetic models or the diffusion in an underlying energy landscape as described in Chapter 4, can be a very helpful control.

An alternative explanation for the kinked chevron plots of FLred may be a competing pathway which introduces a second dimension into the energy landscape picture. As suggested at the bottom of Figure 45B, the folding unit  $\alpha 1\beta 2\beta 3$  may compete with the non-native conformation  $\beta 2\beta 3\alpha 2$ . If this conformation forms a kinetic trap, i. e., an off-pathway intermediate ( $I_{\text{off}}$ ), with an increasing probability of being reached at lower forces with respect to the obligatory on-pathway intermediate ( $I_{12} = I_{\text{on}}$ ), it could trigger a reduced effectiveness of folding of FLred at low forces. The simplest kinetic scheme to describe such a situation is:  $F \rightleftharpoons I_{\text{on}} \rightleftharpoons U \rightleftharpoons I_{\text{off}}$ .

One key to understanding the origin of experimentally observed transition state movement might be the *intrinsic dynamics* of the transition state position. If a transition state is very close to the native state, i. e.,  $\Delta x_{\text{u}}$  is very small, then any force-induced transition state movement by tilting the folding energy landscape can be expected to have a much smaller effect than for a large  $\Delta x_{\text{u}}$ . Compared to natural proteins like, e. g., titin (Ig4, Ig8, or I27), GB1, ubiquitin, Green Fluorescent Protein (GFP), or domain 4 from filamin (ddFLN4), with reported  $\Delta x_{\text{u}}^{\text{ZB}}$  values far below 1 nm, any  $\Delta x_{\text{u}}$  found for FLred is very big [36, 37, 59, 181, 194, 195]. However, this can only be a lead as the reported values all stem from AFM-experiments where the measured  $\Delta x_{\text{u}}$  is affected

<sup>8</sup> Under force, a lower trapping strength results in a larger separation of the average dwell levels of the measured deflection signal. Therefore lower trap stiffnesses may increase spatial resolution especially at higher forces.

by a much higher tether stiffness as compared to experiments with optical tweezers.

### 5.3.4 Mechanical parameters of designed and natural proteins: revealing a key role of the transition state?

How do the mechanical properties of the designed protein FLred compare with natural proteins 'optimized' by evolution? To better judge the measured folding free energy and rates under force, we want to compare two characteristic values of FLred with values obtained for natural proteins: the combined rate  $\lambda_{\text{mid}}$  measured at the mid-force  $F_{\text{mid}}$  and the average folding free energy per amino acid  $\Delta G_0/\text{aa}$ . The average folding free energy allows an approximate comparison of energies stored within protein folds of different size. The combined rate is the sum of unfolding and refolding rates  $\lambda_{\text{mid}} = k_{\text{mid}}^{\text{u}} + k_{\text{mid}}^{\text{f}}$  measured at the mid-force. As  $F_{\text{mid}}$  marks the unique spot where the rate plots intersect due to equal unfolding and refolding rates,  $\lambda_{\text{mid}}$  can be a useful parameter to compare rates of different proteins under force especially around equilibrium. [Table 7](#) lists the parameters  $F_{\text{mid}}$ ,  $\lambda_{\text{mid}}$ , and  $\Delta G_0/\text{aa}$  for multiple proteins. To complete the picture, the zero-force folding rate  $k_0^{\text{f}}$ , the total number of amino acids (aa), and the transition state position ratio<sup>9</sup>  $\Delta x_{\text{u}}/\Delta x_{\text{f}}$  are included as well. This last parameter is a measure of the symmetry of the underlying energy landscape.

All proteins listed in [Table 7](#) have been mechanically characterized using optical tweezers. The data shown for the domain 20 of human filamin A in either *cis* or *trans* configuration (FLNa20<sup>cis/trans</sup>)<sup>10</sup> are derived from [187, 188]. Values for the spectrin domain R15 originate from the mechanical characterization presented in [Chapter 6](#). The range of values covering the correct folding and unfolding transitions of two EF hands into a subdomain of calmodulin (CaM)<sup>11</sup> at 10 mM Ca<sup>2+</sup>, namely  $F_{1234} \rightleftharpoons F_{12}$ ,  $F_{1234} \rightleftharpoons F_{34}$ ,  $F_{12} \rightleftharpoons \text{U}$ , and  $F_{34} \rightleftharpoons \text{U}$ , were taken from [212, 213]. The mechanical parameters for LZ26<sup>11,12</sup> were extracted for the transition between two intermediates ( $\text{I}_1 \rightleftharpoons \text{I}_2$ ) of a GCN4 leucine zipper construct investigated in [83]. Data for the wild-type as well as a stable variant of the villin headpiece (HP35<sup>WT/ST</sup>)<sup>12</sup> were selected from [261]. Among all listed proteins, only FLNa20 consists solely of beta sheets. While FLred contains a mix of beta sheets and alpha helices, all other proteins only comprise alpha helices. Most of the proteins have a globu-

<sup>9</sup> The transition state position ratio is reminiscent of Leffler's proportionality constant indicating resemblance of the transition state to either product or reagent. Applied to the given case it would correspond to  $\alpha_{\text{u}} = \Delta x_{\text{u}}/(\Delta x_{\text{u}} + \Delta x_{\text{f}})$  and range from 0 to 1 [107, 128]. However, here,  $\Delta x_{\text{u}}/\Delta x_{\text{f}}$  was chosen to indicate perfect symmetry when a value equal to 1 is reached.

<sup>10</sup> For FLNa20 the specified number of amino acids only comprises the part involved in folding. The N-terminal beta sheet which would bind to the preceding domain 19 and the follow-up loop region need to be disregarded as can be guessed from [187].

<sup>11</sup> For CaM folding and unfolding of a single subdomain is considered while the second subdomain is either correctly folded or unfolded. Hence, the amount of involved amino acids equals about half the entire amino acid sequence of CaM.

<sup>12</sup> Including HP35<sup>ST</sup>, all transitions in [Table 7](#) are assumed to be two-state upon fitting the respective folding and unfolding rates.

PROTEIN	$\log_{10} k_0^f$ [s <sup>-1</sup> ]	$\frac{\Delta x_u}{\Delta x_f}$	$\lambda_{\text{mid}}$ [s <sup>-1</sup> ]	$F_{\text{mid}}$ [pN]	$\frac{\Delta G_0}{\text{aa}}$ [k <sub>B</sub> T]	No. aa
FLNa20 <sup>cis</sup>	2.6	0.12	$\sim 10^{-1}$	7	0.19	57
R15	3.8	0.12	$\sim 10^{-1}$	8	0.15	106
CaM <sup>domains</sup>	5.0 - - 5.8	0.10 - - 0.24	$10^0$ - - $10^1$	9 - - 10	0.20 - - 0.28	74
FLNa20 <sup>trans</sup>	2.0	0.39	$\sim 10^1$	3	0.06	57
<b>FLred</b>	<b>3.5</b>	<b>0.59</b>	$\sim 10^2$	<b>5</b>	<b>0.08</b>	<b>76</b>
LZ26 <sup>I1I2</sup>	7.8	1.05	$\sim 10^3$	14	0.41	58
HP35 <sup>WT</sup>	5.4	0.83	$\sim 10^4$	7	0.14	35
HP35 <sup>ST</sup>	6.1	0.65	$\sim 10^4$	10	0.28	35

Table 7: Mechanical features of the designed FLred and natural proteins. Listed are the zero-force folding rate constants  $k_0^f$ , the transition state position ratio  $\Delta x_u/\Delta x_f$ , the combined rate  $\lambda_{\text{mid}}$  at the mid-force  $F_{\text{mid}}$ , the average folding free energy per amino acid  $\Delta G_0/\text{aa}$ , and the total number of amino acids contributing to the folding/unfolding of different proteins. Details of proteins and listed values can be found in the text.

lar structure. Only R15 and LZ26<sup>I1I2</sup> make an exception as they have a rather stretched conformation by folding into a three-helix bundle and a coiled coil, respectively.

The entries in Table 7 are ordered from top to bottom by ascending  $\lambda_{\text{mid}}$ . Within the entire range covering six orders of magnitude, the combined rate of about  $10^2 \text{ s}^{-1}$  for FLred is positioned on the faster side. As an increase of protein size, i. e., a higher number of amino acids, may be considered to slow down  $\lambda_{\text{mid}}$ , FLred is surprisingly fast. This is favoured by the low energy stored within the re-designed fold which results in a smaller barrier having to be crossed at equilibrium and therefore enables faster transition rates. In addition, FLred has no off-pathway intermediates which slow down the transition rates as, e. g., in case of CaM.

The average folding free energy per amino acid  $\Delta G_0/\text{aa}$  of FLred is of the same order as for the highly destabilized FLNa20<sup>trans</sup>, which indeed underlines a successful labile design. Furthermore,  $\Delta G_0/\text{aa}$  can be seen to increase along with the mid-force. This can be expected as the energy landscape of a protein with a very high folding free energy needs to be tilted much more until an equilibrium between the folded and the unfolded state is reached. Interestingly, in case of FLNa20<sup>trans</sup> the reduced  $\Delta G_0/\text{aa}$  value as compared to FLNa20<sup>cis</sup> also results in a much faster  $\lambda_{\text{mid}}$ . The particularly high  $\Delta G_0/\text{aa}$  value of  $0.41 \text{ k}_B\text{T}$  for LZ26<sup>I1I2</sup> is a speciality of coiled coils which can store up to  $4.42 \text{ k}_B\text{T}$  within a single helix turn<sup>13</sup> theoretically allowing a maximal value of  $\Delta G_0/\text{aa} = 0.63 \text{ k}_B\text{T}$  [30].

<sup>13</sup> As coiled coils are dimers, the formation of a 3.5 amino acid long helix turn comprises 7 aa.

The extrapolated rate constants  $k_0^f$  for folding at zero force almost cover eight orders of magnitude, which is basically the entire range of what has been reported in literature so far. By taking a closer look at [Table 7](#) one can see that  $k_0^f$  has a general tendency to increase along with  $F_{\text{mid}}$ . If all  $\Delta x_u$ ,  $\Delta x_f$ , and  $\lambda_{\text{mid}}$  were the same, this observation would, of course, be trivial. But even given the large deviations, the zero-force rate constants more or less directly relate to the mid-force.

It is a striking observation that the transition state position ratio<sup>14</sup>  $\Delta x_u/\Delta x_f$  increases along with  $\lambda_{\text{mid}}$ . From this finding it appears as if a symmetric energy landscape favours faster rates within equilibrium fluctuations under force. Especially at higher forces, a rather central transition state or at least similar  $\Delta x_u/\Delta x_f$  values for folding and unfolding seem to enable equilibrium fluctuations to occur in the first place. There are two further examples that corroborate this picture. The  $\alpha$ -helical linker of the C-terminal immunoglobulin domain My12 of myomesin ( $\alpha^{12}$ ), which was mechanically investigated with an AFM, shows a  $\lambda_{\text{mid}}$  value of  $10^2 \text{ s}^{-1}$  at a mid-force of 40 pN and a transition state position ratio  $\Delta x_u/\Delta x_f$  of 0.94. The 14 amino acids involved in the reversible unfolding of  $\alpha^{12}$  under load store a folding free energy of  $16 \text{ k}_B\text{T}$ , which implies an extraordinarily high  $\Delta G_0/\text{aa}$  value of  $1.14 \text{ k}_B\text{T}$  [18, 19]. However, this value originates from the special pulling geometry where the main part of My12 adjacent to  $\alpha^{12}$  is not being pulled upon while still providing a stabilizing interface. A similar stabilization can be considered for the mechanical characterization of a fast-folding subpart of the ATP-lid in the active pulling geometry of the thermophilic variant of adenylate kinase (thADK42 – 144). At a mid-force of about 17 pN the combined rate  $\lambda_{\text{mid}}$  is of the order of  $10^3 \text{ s}^{-1}$  for thADK42 – 144 [166]. With a  $\Delta x_u/\Delta x_f$  value<sup>15</sup> of 0.75 we again find a quite central transition state position along with fast equilibrium transitions at a comparably high force.

As all of the aforementioned proteins are very different from each other, the observed relations between combined rate and transition state position as well as zero-force folding rate, mid-force and stored energy may be a mere coincidence. However, precisely because of their observation despite huge differences, these relations might be universally valid.

In summary, the comparison of the characteristic values  $\lambda_{\text{mid}}$  and  $\Delta G_0/\text{aa}$  of the designed protein FLred with the corresponding values of natural proteins reveals the following. First, the folding free energy of FLred is indeed very small. And, second, the combined rate at the mid-force is comparably fast. The latter is clearly a result of the small folding free energy and, additionally, seems to be favoured by a rather symmetric energy landscape. Taken together this speaks for a mechanically labile but sound design.

14 Note that all data represented in [Table 7](#) were extracted from fitting the SB-model to experimentally determined force-dependent rates. The only exception is CaM, where  $\Delta x_u$  was determined using the Zhurkov-Bell model. Hence, the given range for  $\Delta x_u/\Delta x_f$  must be considered as a lower estimate for the two domains of calmodulin.

15 By way of personal communication from Benjamin Pelz:  $\Delta x_u = 2.1 \text{ nm}$ ,  $\Delta x_f = 2.8 \text{ nm}$ .

## 5.4 SUMMARY AND OUTLOOK

### 5.4.1 *Successful protein design verification*

The single-molecule mechanical characterization leaves no doubt about FLred being a successful re-design of the original FL. The 16 point mutations on FL, which make up 21 % of the entire protein sequence, lead to an almost three times smaller folding free energy of about  $5.8 k_B T$  for FLred versus  $15.1 k_B T$  for FL. This result also confirmed the folding free energy obtained from bulk measurements. Furthermore, the measurements clearly indicate that FLred comprises an entirely folded protein structure which is even able to completely refold against small residual forces. At a mid-force of about 5 pN, the unfolding and refolding rate both equal to about  $40 s^{-1}$ . Based on Kramers rate-theory, the zero-force folding rate constant of  $k_0^{f,SB} = 3000 s^{-1}$  in combination with the zero-force extrapolated deconvolved energy landscape of FLred implies a comparably low internal friction coefficient of  $\gamma_{FLred} = 5.1 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . Taken together, FLred meets all requirements as a candidate for being applied in a protein-based hydrogel as formulated in [Section 5.1](#). And indeed, hydrogels containing chemically cross-linked multimers of FLred were shown to be highly stretchable and able to fully recover their massive hysteresis [76].

The folding of FLred, thought previously to be a simple two-state mechanism was revealed to be of a more complex nature. It is not yet entirely clear whether a very broad transition state ensemble or two distinct transition states are the reason for the strong deviation from simple two-state behaviour at forces below 4 pN. Transition state movement in accord with Hammond behaviour was reproduced by deconvolution but could not entirely explain the kinked chevron plots. Therefore, under the assumption of a discrete force-induced transition state switch, a transiently populated on-pathway intermediate was hypothesized comprising the folding unit  $\alpha 1 \beta 2 \beta 3$ . Such an intermediate suggests a force-induced cooperativity breakdown between the two folding units forming FLred. From a protein design perspective, including a transiently populated, relatively confined intermediate state appears reasonable as stepwise or sequential folding of a protein can help to avoid misfolded states as well as to refold against force.

Improved statistics and data recorded at higher resolution along with complementary Monte Carlo simulations will help to unravel the true nature behind the complex two-state folding of FLred. Especially higher resolution data may enable alternative measures of transition state movement like, e. g., force-dependent splitting probabilities or even more advanced transition path analysis methods introduced in [Section 3.8](#).

Finally, an investigation of the much more stable ideal Ferredoxin-like fold (FL) could be of interest. Here, several scenarios are possible. Eventually, an overly transition state movement will not be directly detectable for FL as force ranges for unfolding and refolding might not sufficiently overlap due to potentially higher transition barriers. In that case, a sum of transition state

positions which does not match the full contour length gain might still indicate transition state movement. Another option is that the hypothesized intermediate shows up as expected. However, due to a much higher folding free energy being involved, it may even be possible to detect more than just one intermediate including misfolded states from non-native pathways.

#### 5.4.2 Application as an *in vivo* force sensor

By using genetically encoded tension sensors it has become possible to measure forces *in vivo*, i.e., within living cells. The working principle of these tension sensors is the combination of distance-dependent FRET, which can be observed from the outside of cells, with the force-dependent extension of polymers by placing the latter as a force-sensitive element between a pair of fluorophores that can undergo efficient FRET [88]. For a correct interpretation of FRET-signals obtained from the force sensing modules after insertion into cells, it is crucial to have a comprehensive understanding of the mechanical behaviour of the force-sensitive element. Therefore, a single-molecule mechanical characterization of isolated force sensing modules appears to be the best and most direct approach.

Within the framework of this thesis a model was developed to calculate an expected FRET-force relation for *in vivo* force sensing modules based on force-dependent probability distributions extracted from single-molecule mechanical measurements. In brief, for a given force, the model assigns a FRET-efficiency value to each populated state based on the FRET-distance<sup>16</sup> of the fluorophores  $E(x)$  and the inverted force-extension relation  $x_i(F)$  of each individual state  $i$ . Then, all states' contributions are weighed with their respective force-dependent probability  $p_i(F)$  of being populated at the given force and added up to an expected overall FRET-efficiency. Hence, in case of a two-state folder, the FRET-force relation  $E(F)$  simply reads [9]:

$$E(F) = E(x_i(F)) \cdot p_i(F) + E(x_j(F)) \cdot p_j(F). \quad (66)$$

The model was first applied to rigorously mechanically characterized tension sensors based on the villin headpiece peptide, namely HP35-TS and HP35st-TS (here, TS stands for tension sensor; for details on sensor design, see Section A.1.5). The predicted difference in force sensitivity was indirectly confirmed by subsequent *in vivo* measurements, which also marked an important step ahead in expanding the accessible force range for tension sensors [9]. Due to the model's very general approach, in principle, a probability distribution measured for any kind of protein can be readily converted into a FRET-force relation. However, among other aspects, the timescales for reaching the equilibrium described by the probability distributions at a given force need to be accounted for. A general list of requirements that should be met by a protein serving as a force-sensitive element as well as considerations on how to expand the current toolbox of single-molecule calibrated FRET-based tension sensors can be found here [81].

<sup>16</sup> The FRET-distance relation is calculated after the inverse 6<sup>th</sup> power law of FRET [111].

The near-equilibrium transition of FLred at very slow pulling speed shown in Figure 40 is strongly reminiscent of the humplike transitions reported for HP35-TS and HP35st-TS [9]. The mechanical properties of the HP35-based sensors can be characterized by their folding free energies of  $\Delta G_0^{\text{HP35}} = 5.6 \pm 1 \text{ k}_B\text{T}$  and  $\Delta G_0^{\text{HP35st}} = 10.4 \pm 1 \text{ k}_B\text{T}$  accompanied by total contour length gains of  $L_p^{\text{HP35}} = 10.7 \pm 0.4 \text{ nm}$  and  $L_p^{\text{HP35st}} = 10.9 \pm 0.4 \text{ nm}$ . At the transition mid-points located at the forces  $F_{\text{mid}}^{\text{HP35}} = 7.4 \pm 0.5 \text{ pN}$  and  $F_{\text{mid}}^{\text{HP35st}} = 10.6 \pm 0.4 \text{ pN}$ , combined rates of  $\lambda_{\text{mid}}^{\text{HP35}} = 1.7 \times 10^4 \text{ s}^{-1}$  and  $\lambda_{\text{mid}}^{\text{HP35st}} = 2.1 \times 10^4 \text{ s}^{-1}$  were extracted by autocorrelation analysis analogous to [261]. Even though the combined rate of FLred at its transition mid-point at 5 pN is about 200 times slower than the ones of the HP35-based sensors, FLred still equilibrates within just a few tens of milliseconds. Given an almost three times longer contour length gain of  $L_p^{\text{FLred}} = 27.1 \pm 1.5 \text{ nm}$  along with an only slightly bigger folding free energy of  $\Delta G_0^{\text{FLred}} = 5.8 \pm 0.6 \text{ k}_B\text{T}$  as compared to HP35-TS, FLred can be expected to yield a much sharper force sensitivity within the lower range of HP35-TS.

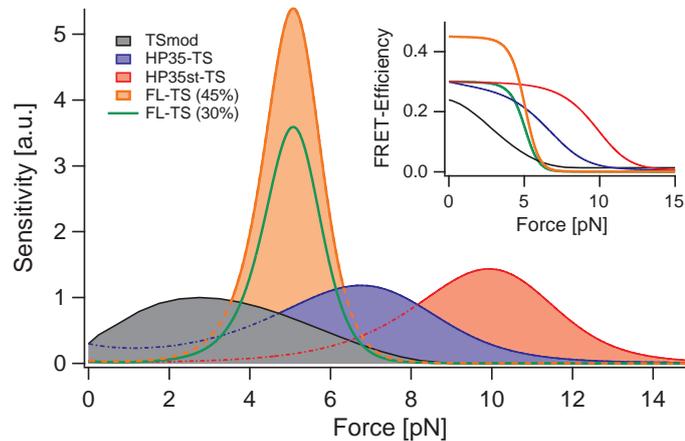


Figure 47: Force sensor performance of FL-TS with FLred as force-sensing element compared to TSmod, HP35-TS and HP35st-TS. The illustrated renormalized sensor sensitivities reflect the absolute value of the local FRET-force relation slopes from the inset; a fixed normalization factor was introduced to set the maximum for TSmod equal to 1. *Inset*: Same color code as for main figure. Calculated FRET-force relations of FL-TS with either 30 (green) or 45% (orange) starting FRET-efficiency. Data for the other curves are taken from [9, 88]. Above 9 pN values for TSmod are set constant.

Figure 47 shows force sensitivity plots for different sensors including curves describing the expected performance of FLred as a force-sensing module within FL-TS. The sensitivities represent the renormalized absolute values of the derivative of their corresponding FRET-force relations. The calculation of the FRET-force relation for FL-TS is based on the probability distributions inferred by a two-state model which fits slowly pulled force-extension traces of FLred very well, see Figure 41B. This does not contradict the rather complex two-state nature of FLred, as the deviations from simple two-state behaviour only arise at forces below 4 pN, which is already beyond the 'active' sensor

region ranging from 4 to 6 pN. As FL-TS is assumed to have the same fluorophores as the HP35-based sensors, the same Förster distance of 5.8 nm as well as starting FRET-efficiency of 30% is considered for the relevant part of the FRET-distance relation. However, owing to the shorter N- to C-terminal distance of FLred as compared to HP35, a maximum of up to 45% of starting FRET-efficiency could be expected in the case where there is no other steric hindrance. When comparing all sensitivity plots with each other, FL-TS clearly shows outperforming precision in measuring forces around 5 pN and thereby blends in perfectly between the well-established TSmod and the recently developed HP35-TS. Hence, FLred appears to be a promising candidate not only for being applied in protein-based hydrogels but also within force-sensing modules. Note that this hypothesis has recently been verified by a collaborative publication [185].

#### 5.4.3 Force spectroscopy - a test bench for protein design, folding concepts, and models

The mechanical characterization of FLred has proven single-molecule force spectroscopy to be a valuable tool for testing principles used to design artificial proteins. The observed folding properties could be reasonably explained by the physical concept of an underlying folding energy landscape. To interpret the information extracted from force spectroscopic data, suitable theoretical models are needed. Thanks to the still rather simple mechanics of FLred, a whole arsenal of models and their fit functions could be tested and compared to each other. The overlapping force range equally accessed by non-equilibrium and equilibrium measurement techniques turned out to be especially useful to confirm excellent agreement between data obtained and analysed in different ways.

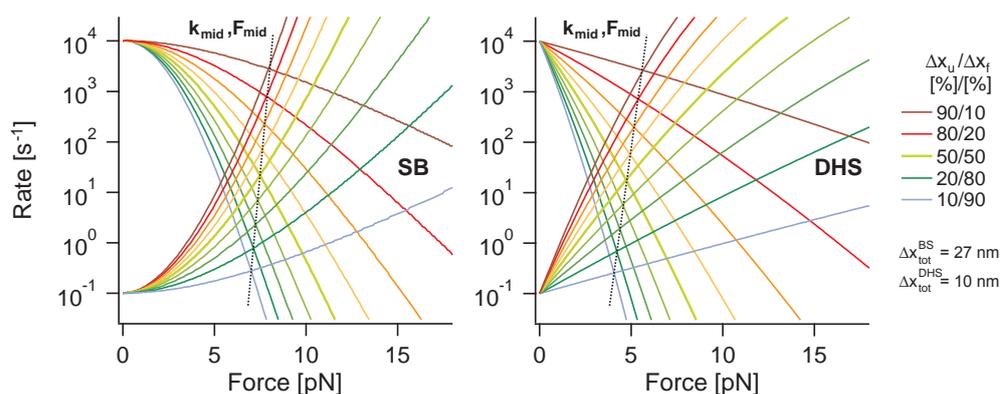


Figure 48: Affecting rates by shifting the transition state position. For both, the SB- and the DHS-model, equilibrium rates  $k_{\text{mid}}$  are strongly affected by the transition state position ratio  $\Delta x_{\text{u}}/\Delta x_{\text{f}}$ , while  $F_{\text{mid}}$  almost remains constant. Zero-force rate constants are kept constant yielding a folding free energy of  $\Delta G_0^{k_0} = 11.5 k_{\text{B}}T$ . Further, for the DHS-model, the barrier heights at the transition state position are fixed to  $\Delta G_0^{\text{u,DHS}} = 16.5 k_{\text{B}}T$  and  $\Delta G_0^{\text{f,DHS}} = 5 k_{\text{B}}T$  in accord with  $\Delta G_0^{k_0}$ .

When comparing the mechanical properties of FLred to naturally occurring proteins, FLred was found to have an extremely low folding free energy. In addition, similarly to other proteins with a rather central transition state position, the combined rate  $\lambda_{\text{mid}}$  of FLred at the transition mid-point is comparably fast, see Table 7. Besides a possible hint towards an energy landscape of a downhill folder<sup>17</sup>, the transition state position can drastically affect the combined rates at the transition mid-point no matter which rate model is used. This is exemplified in Figure 48, which also reveals that at a fixed transition state ratio  $\Delta x_{\text{u}}/\Delta x_{\text{f}}$  the  $\lambda_{\text{mid}}$  values are nearly the same for both the SB- as well as the DHS-model. The difference in force range for  $F_{\text{mid}}$  can be attributed to the difference in slopes, i. e., not perfectly matched  $\Delta x_{\text{tot}}$  values.

While the observed result can be expected from the SB-model, obtaining the very same result with the DHS-model is remarkable. Here, the question arises on how much the model is undermined by keeping the zero-rate constants fixed. Apart from the diffusion constant, the zero-force rate constants also comprise the curvatures of the underlying energy landscape which, in principle, need to change upon transition state movement.

A possible interpretation of the interplay between transition state position and force-dependent rates can be the following. Proteins, which are intended to mechanically withstand high forces, should rather have a transition state very close to the folded state along with a high unfolding barrier. This way, the effect of an external force on the transition state position and the barrier height for unfolding is minimized. However, in case the protein does unfold or 'rupture', refolding under load is not very likely either. In contrast, if a protein is involved in mechanical processes, where it is important to maintain a folded structure under mechanical load like, e. g., in case of the villin headpiece or coiled coil proteins, the transition state may preferably be located more centrally between the folded and unfolded states. This way, much faster refolding under load is possible to keep up proper functioning. Finally, if the transition state gets close to the unfolded state, the intended protein's task should be a permanent fluctuation as unfolding and folding occur at the fastest possible rates. In summary, the transition state position can potentially reveal whether a protein's task is to withstand forces, to maintain a folded structure under load, or to fluctuate.

<sup>17</sup> It seems intuitively clear that the energy barriers resulting from tilting the energy landscape towards the transition mid-point will be as small as possible for a downhill folder regardless of the location of the transition state.

## ROSSMANN FOLD: AN ARTIFICIAL PROTEIN IN TROUBLE

---

The basis of this chapter is a comparative mechanical study of an artificial and a natural protein. While the artificial protein is specifically designed to have a high folding free energy by a computer algorithm, the natural protein is optimized towards reliable folding by evolution. Besides other aspects, the results of the single-molecule force spectroscopic measurements reveal that high folding free energies come at the cost of cooperativity and an increased energy landscape roughness. A special feature of this chapter is the first-time application of transition path analysis tools to experimental data.

### 6.1 DESIGN MEETS EVOLUTION: ROSSMANN VS. R15

The two proteins of the comparative mechanical study are a designed 2x2 Rossmann fold (ROSS) and the naturally occurring spectrin domain R15 (R15). While being similar in size, the proteins have very different shapes as can be seen upon inspection of their structures in [Figure 49](#). ROSS, a designed representative of the Rossmann fold, comprises four  $\beta\alpha$ -repeats forming a parallel  $\beta$ -sheet sandwiched between two  $\alpha$ -helices on each side. Taken as a whole, ROSS is a densely packed globular protein. R15, in contrast, has a rather extended shape. Consisting of three  $\alpha$ -helices wrapped around each other, R15 represents a fold termed helix bundle.

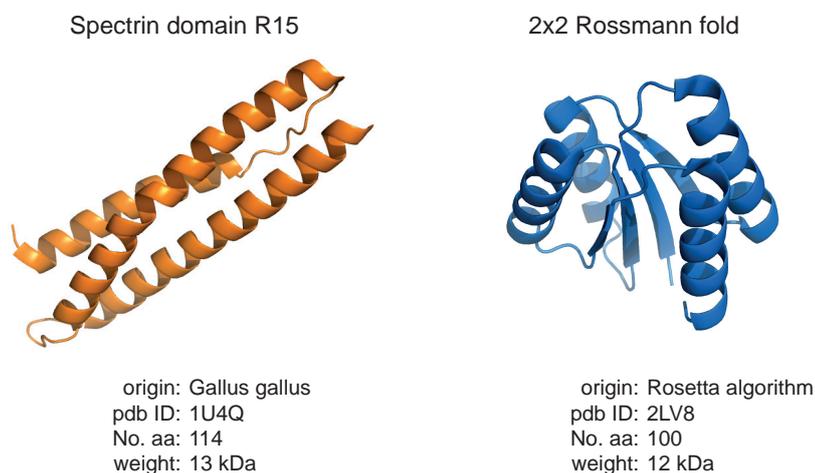


Figure 49: Crystal and solution NMR structures of the spectrin domain R15 and the 2x2 Rossmann fold.

Besides their difference in shape, ROSS and R15 also differ in terms of their reported folding free energies being 14.9 kcal/mol for ROSS and 6.8 kcal/mol

for R15<sup>1</sup> [115, 203]. By this means, in terms of folding free energy, ROSS is more than twice as stable as R15. This appears reasonable since ROSS has a comparably smaller solvent-exposed surface area due to its compact structure. Further, this compactness also implies a larger stabilizing interface area which allows formation of a relatively higher number of native contacts. The folding free energy per amino acid values  $\Delta G_0/\text{aa}$  of ROSS and R15 set them in relation to other mechanically characterized proteins, see Table 7. A value of 0.25 for ROSS compares well with individual domains of calmodulin or the artificially stabilized villin headpiece. In case of R15, a value of 0.10<sup>2</sup> relates best to the wild type of villin or the artificially destabilized FLred. As R15 is part of the cytoskeleton, one of its biologically relevant functions is to maintain the shape and structure of a cell. Owing to that function, R15 can be expected to show considerable mechanical stability despite its relatively small folding free energy.

Note that with respect to the Rossmann fold there exists a structurally very similar fold known as the P-loop<sup>3</sup> or Walker fold [232]. The only difference between their tertiary structures are the swapped two  $\beta$ -strands in the middle of the central  $\beta$ -sheet. Interestingly, although being designed according to the same design rules of one and the same protein design study, the P-loop has a folding free energy of 4.8 kcal/mol which is 3 times less as compared to ROSS.

## 6.2 A MULTIDIMENSIONAL AND ROUGH ENERGY LANDSCAPE – A TWO-STATE PERSPECTIVE

This section introduces the results of a study which compares the mechanical behaviour of the natural protein R15 to the designed protein ROSS. To facilitate comparison, all figures are typically arranged in the style of Figure 49: data referring to R15 are coloured in brick red and positioned next to the respective results obtained for ROSS which are coloured in dark blue. If not stated differently, transitions between fully folded and entirely unfolded protein states are evaluated from a two-state perspective which enables uniform data processing.

### 6.2.1 Multiple fingerprints and multi-modal unfolding

In the following we start off with constant velocity measurement results which reveal that R15 behaves like a perfect two-state folder while ROSS has a rough and multidimensional folding energy landscape.

Figure 50 displays typical force spectroscopic fingerprints, i. e., stretch-and-relax cycles of R15 and ROSS recorded at 500 nm/s. While there is only one

<sup>1</sup> At a temperature of 298 K, the values given in kcal/mol correspond to 25.2 and 11.5 k<sub>B</sub>T for ROSS and R15, respectively.

<sup>2</sup> Note that the indicated value of 0.10 is derived from literature while the value of 0.15 in Table 7 is based on measurements performed in this thesis.

<sup>3</sup> P-loop stands for phosphate-binding loop.

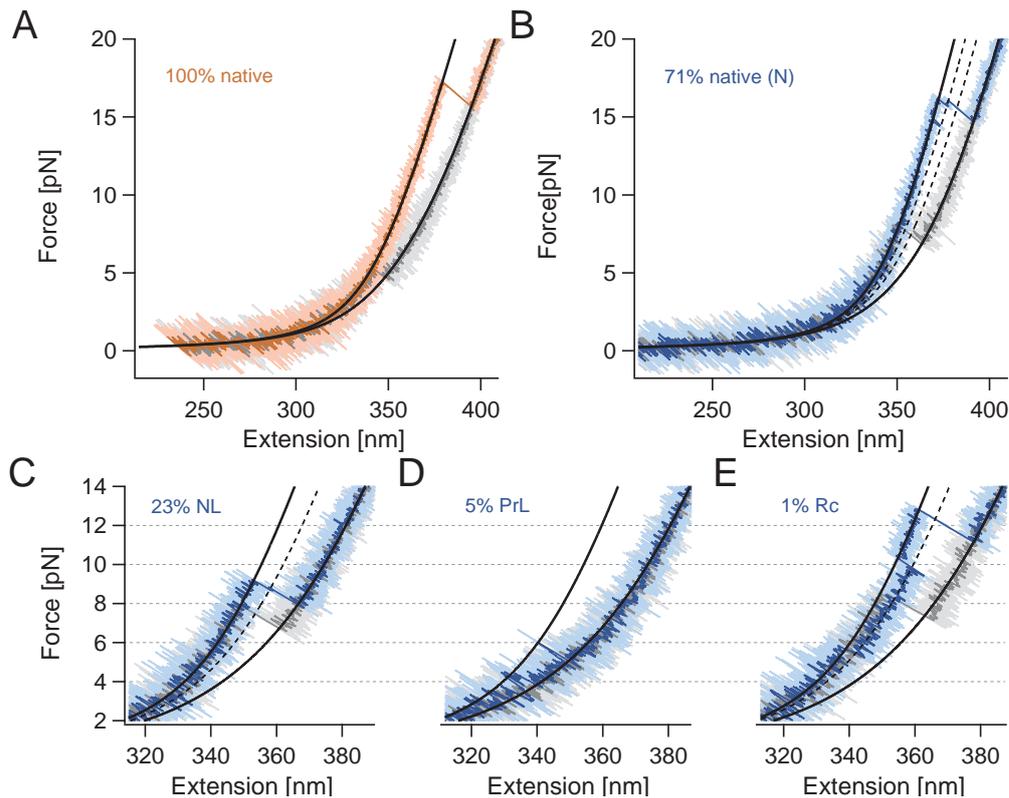


Figure 50: Force spectroscopic fingerprints of R15 and ROSS recorded at 500 nm/s. Percentages indicate the relative occurrence of the respective fingerprint. (A) Typical force-extension trace of R15 showing unfolding during stretching (brick red) and refolding during relaxation (grey). Black lines are *WLC*-fits to the folded and fully unfolded state. (B) Force-extension trace of ROSS representing the predominating ‘native’ unfolding and refolding pattern. (C) Representative ‘native-like’ (NL) stretch-and-relax cycle of ROSS. (D) ‘Proline-like’ (PrL) unfolding and refolding pattern of ROSS. (E) Typical ‘rescue’ (Rc) pattern of ROSS.

characteristic ‘native’ fingerprint for R15, see Figure 50A, at least four different fingerprints can be discerned for ROSS, see Figure 50B - Figure 50E. According to the predominating ‘native’ fingerprint of the artificial Rossmann fold, unfolding typically occurs around 15 pN via two short-lived on-pathway intermediates, see dashed *WLC*-fits in Figure 50B. In addition, the native (N) fingerprint often features pre-flipping into the first on-pathway intermediate prior final unfolding. During relaxation, usually partial refolding attempts can be observed until ROSS finally refolds at forces around 6 pN. Second in observed frequency is the ‘native-like’ (NL) fingerprint of ROSS which typically unfolds at forces around 10 pN upon passing through one short-lived on-pathway intermediate. Next in relative occurrence is the ‘proline-like’ (PrL) fingerprint shown in Figure 50D which represents a very unstable configuration that already unfolds below 5 pN. Most importantly, the PrL only appears in a series of multiple stretch-and-relax cycles which, along with the low stability, makes it reminiscent of the proline-switch observed for filamin A [189].

As ROSS indeed has a proline at position P86, see [Section A.1.2](#) for the exact protein sequence and [Figure 59A](#) for the position within the solution NMR structure, the proline-switch hypothesis is tested with proline-free mutants in [Section 6.3.1](#). The fourth identified fingerprint for ROSS comprises a 'rescue' (Rc) pattern within the stretch cycle: after partial unfolding at relatively low forces, ROSS refolds against higher force into the (presumably) native configuration before complete unfolding, see [Figure 50E](#).

Since PrL and Rc are each clearly distinguishable from all other fingerprints, their relative occurrence of 5 % and 1 %, see [Figure 50](#), can be directly derived by merely counting them among all observed stretch-and-relax cycles performed at 500 nm/s. The remaining 94 % comprise the N and NL fingerprints which are more difficult to discern due to their overlapping force ranges. In short and in anticipation of the following paragraphs, the relative amount between N and NL is obtained by fitting two proportionate distributions to the unfolding force histogram of the remaining 94 % of complete unfolding events, see [Figure 51B](#).

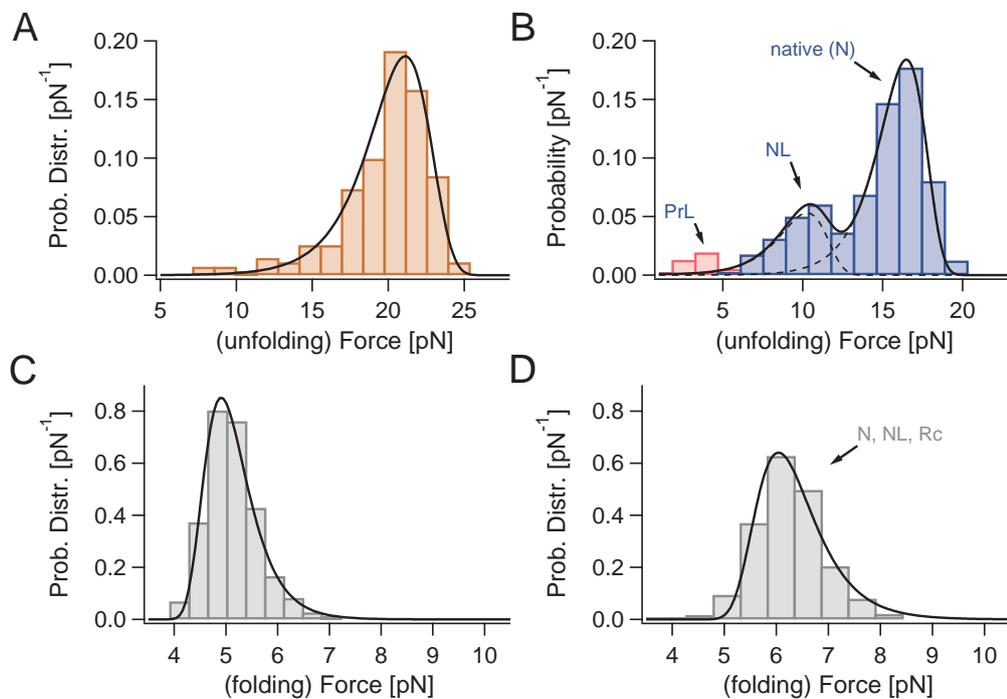


Figure 51: ROSS has a multi-modal unfolding force distribution. Displayed distributions stem from stretch-and-relax cycles performed at 500 nm/s. (A) Unfolding force histogram of R15 (brick red bars) fitted based on [Equation 35](#) (black line). (B) The unfolding force histogram of ROSS (dark blue and pink bars) has three distinct peaks. The black line fits two proportionate force distributions (dashed lines) to the dark blue histogram. (C) and (D) Refolding force histograms (grey bars) and fits for R15 and ROSS.

The force distribution histograms in [Figure 51](#) totally comply with the above made classification of fingerprints. In case of R15, both the unfolding and refolding force distributions in [Figure 51A](#) and [Figure 51C](#) are single-

peaked and can be well fitted according to [Equation 33](#) which is based on a two-state model. Further, the force distribution for unfolding is broader than for refolding which indicates a transition state position relatively close to the folded state. All in all, these data support that R15 is a worthy representative of naturally occurring proteins which fold in a simple two-state manner and have only one single typical fingerprint.

In contrast to both single-peaked distributions of R15, the unfolding force distribution of ROSS shows three distinct peaks which can be attributed to the three fingerprints N, NL, and PrL, see [Figure 51B](#). Due to the unique unfolding pattern of the PrL fingerprint, its force distribution can be clearly identified and is therefore given its own pink color while still being part of the overall unfolding force distribution of ROSS. Unfolding forces observed after a ‘rescue’ transition in Rc patterns intrinsically do not fully sample the unfolding force distribution and are therefore not included in the histogram of [Figure 51B](#) to not obscure the statistics of ‘clean’ N and NL fingerprints. By this means, the double-peaked dark blue part of the unfolding force histogram representing  $p_{\text{unf}}^{\text{N+NL}}(F)$  of ROSS can be used to derive the relative amount between N and NL unfolding events. To do so, two independent distributions are simultaneously fit<sup>4</sup> to the histogram according to

$$p_{\text{unf}}^{\text{N+NL}}(F) = (1 - \varphi_{\text{NL}}) \cdot p_{\text{unf}}^{\text{N}}(F) + \varphi_{\text{NL}} \cdot p_{\text{unf}}^{\text{NL}}(F), \quad (67)$$

where  $\varphi_{\text{NL}}$  is the fraction of NL unfolding and  $p_{\text{unf}}^{\text{N}}(F)$  and  $p_{\text{unf}}^{\text{NL}}(F)$  are calculated according to [Equation 35](#). In [Figure 51B](#), the black bi-modal fit yields  $\varphi_{\text{NL}} = 0.25$  which in turn corresponds to 25 % of the 94 % of N and NL among all observed fingerprints of ROSS. An important conclusion which can be directly drawn upon seeing the multi-modal unfolding force histogram is that the mechanics of the artificial Rossmann fold go beyond a simple single-path picture [64]. In other words, the observed multi-modal force distribution is a direct signature of multi-pathway conformational transitions which determine the mechanics of ROSS [170]. When moving over to the folding force distribution of ROSS, multi-modality seems to be gone, see [Figure 51D](#). However, the shown distribution only comprises the majority of refolding events of N, NL and Rc fingerprints which are marked by a relatively distinct transition as the ones shown in [Figure 50](#). For simplicity, a variety of ‘creeping’ refolding transitions were not taken into account to which PrL refolding belongs, too.

[Table 8](#) lists all parameters which characterize the fits to the force distributions of R15 and ROSS in [Figure 51](#). When comparing R15 to the predominating native fingerprint of ROSS, there are no drastic differences. However, upon closer inspection, the transition state of ROSS appears to be more central and, hence, further away from the folded state as compared to R15. Further, an asymmetric divergence of the probability maxima positions of the two proteins stands out: while R15 withstands pulling up to higher forces than ROSS, ROSS is able to refold against higher forces than R15. In addition to a shifted transition state position, the ability to refold against higher forces may also

<sup>4</sup> Note that for actual fitting  $p_{\text{unf}}^{\text{N+NL}}(F)$  is renormalized to have an area of one on its own.

be a result of the higher folding free energy of ROSS which is intended by design. The higher resistance of R15 against unfolding can be interpreted as a proof for use-oriented natural design by evolution. Despite its more than two times lower folding free energy, R15 withstands higher mechanical stress during stretch-and-relax cycles than ROSS which makes R15 a better mechanical stabilizer. Besides being a three helix bundle which involves shearing all three helices against each other to induce unfolding upon stretching, an important key to the performance of R15 appears to be its high cooperativity which is expressed by its simple two-state behaviour. Finally, a comparison between the characteristic parameters of N and NL unfolding kinetics reveals that NL is basically a less stable version of ROSS which has an almost two orders of magnitude higher zero-force unfolding rate constant according to the ZB-model. However, as NL contributes about the same overall contour length gain as N upon unfolding, both N and NL are difficult to be distinguished especially within the overlapping force range of their unfolding force distributions.

PROTEIN	$\log_{10} k_0^{f,ZB}$ [s <sup>-1</sup> ]	$\log_{10} k_0^{u,ZB}$ [s <sup>-1</sup> ]	$\Delta x_f^{ZB}$ [nm]	$\Delta x_u^{ZB}$ [nm]	$F_{max}^{fold}$ [pN]	$F_{max}^{unf}$ [pN]
R15	6.8 ± 0.3	-3.3 ± 0.1	9.5 ± 0.4	2.1 ± 0.2	4.9	21.1
ROSS (N)	6.3 ± 0.8	-3.5 ± 0.1	7.2 ± 0.5	2.9 ± 0.3	6.0	16.5
ROSS (NL)	-	-1.7 ± 0.1	-	2.5 ± 0.4	-	10.3

Table 8: Kinetic ZB-model fit parameters of R15 and ROSS based on force distribution histograms collected from 500 nm/s constant velocity measurements, see [Figure 51](#). Based on [Equation 67](#), the fraction  $\varphi_{NL}$  of NL unfolding among all unfoldings of N and NL was determined to be 0.25.  $F_{max}^{fold}$  and  $F_{max}^{unf}$  refer to the mode of the respective distribution.

Besides force spectroscopic fingerprints and force distributions, the evaluation of constant velocity measurements also provides experimentally determined contour length gains upon forced protein unfolding. Here, the observed contour length gains<sup>5</sup> from all constant velocity measurements of R15 and ROSS yield an average of  $L_p^{CV,R15} = 33.4 \pm 1.1$  nm and  $L_p^{CV,ROSS} = 34.7 \pm 0.9$  nm. These values are in excellent agreement with the contour length gains expected for full unfolding of the respective structure, i. e.,  $L_p^{calc,R15} = 34.47$  nm and  $L_p^{calc,ROSS} = 34.51$  nm which were calculated according to [Equation 5](#). By this means, both R15 and ROSS have to be fully folded since otherwise significantly shorter gains would have been measured.

In summary, both R15 and ROSS are fully folded. While R15 has only one typical fingerprint, ROSS shows at least four different typical unfolding pat-

<sup>5</sup> Average contour gains of R15 | ROSS are deduced from WLC-Fits to the folded and maximally unfolded states of N = 204 | 809 force-extension traces from 4 | 11 molecules. Persistence lengths of  $p_p = 0.5$  | 0.7 nm were chosen to model the unfolded polypeptide chain. Average DNA-handle parameters are  $p_D = 23$  | 31 nm,  $L_D = 360$  | 364 nm, and  $K = 180$  | 200 pN.

terns followed by refolding. This observation is supported by the respective force distributions from which we can deduce that R15 is a highly cooperative and simple two-state folder, whereas the mechanical behaviour of ROSS involves multi-pathway conformational transitions within a multidimensional energy landscape. Further, all transitions of ROSS between its folded and unfolded states typically involve short-lived on-pathway intermediates which are direct reporters of a rough energy landscape.

### 6.2.2 Split and kinked rate plots

As a next step, constant distance measurement results are introduced to complement the available data set of R15 and ROSS. In this subsection, the major focus lies on force-dependent transition rates which, in case of ROSS, give clear indications for a rough energy landscape. In contrast, R15 continues being a perfect 'role model' for simple two-state behaviour.

In [Figure 52](#), rates extracted from constant distance measurements based on [HMM](#) analysis are represented by full circles while empty squares refer to rates derived from constant velocity measurements using the OBS method. In case of R15, unfolding (grey) and refolding (brick red) rates are each nicely fitted over the entire accessible force range by a single SB-model fit (thick solid lines) as shown in [Figure 52A](#). This confirms that R15 folding can be adequately modelled by simple two-state mechanics. The kinetic parameters provided by the SB-model fits are given in [Table 9](#). Thin dotted lines represent the ZB-model which, instead of directly fitting the data in [Figure 52A](#), uses the parameters obtained from fitting the force distributions in [Section 6.2.1](#), see [Table 8](#). Especially the good agreement with the OBS rates speaks for a consistent data set. However, with respect to a correct treatment of all involved energetic contributions, see [Section 3.5](#), the SB-model outperforms the ZB-model. In [Table 10](#), a comparison of the respective folding free energies  $\Delta G_0^{k_0}$  and  $\Delta G_0^{k_0(\text{ZB})}$  to the energy  $\Delta G_0^{\text{bulk}}$  from bulk confirms a much better zero-force extrapolation by the SB-model. For a detailed discussion of the predictive capabilities of different kinetic models see [Section 5.3.2](#).

For ROSS, the rate plots in [Figure 52B](#) look way more complicated as compared to R15. Due to multiple kinks, here, the SB-model only locally fits the data, see [Table 9](#) for all obtained fit parameters. The kinks hint at an unfolding and refolding process of ROSS which involves the crossing of multiple sequential barriers. This in turn can be interpreted as a fingerprint for roughness, see also [Section 4.2.4](#). In case of refolding (dark blue data points), three different force ranges (I - III) can be discerned for fitting, see all lines in [Figure 52B](#) with a negative slope. The respective force ranges are I: 5 - 6 pN, II: 6 - 7 pN, and III: 7 - 9.5 pN. Force range III comprises refolding rates from both constant velocity and constant distance mode. The only direct fit (thick grey line) to the unfolding rates of ROSS (all grey data points presumably belong to 'native' (N) unfolding, see below) is made for the highest accessible force range (IV) covering 16.5 up to 18 pN. The red dashed line is not directly fit to the unfolding rates but modelled based on its 'counterpart' which fits

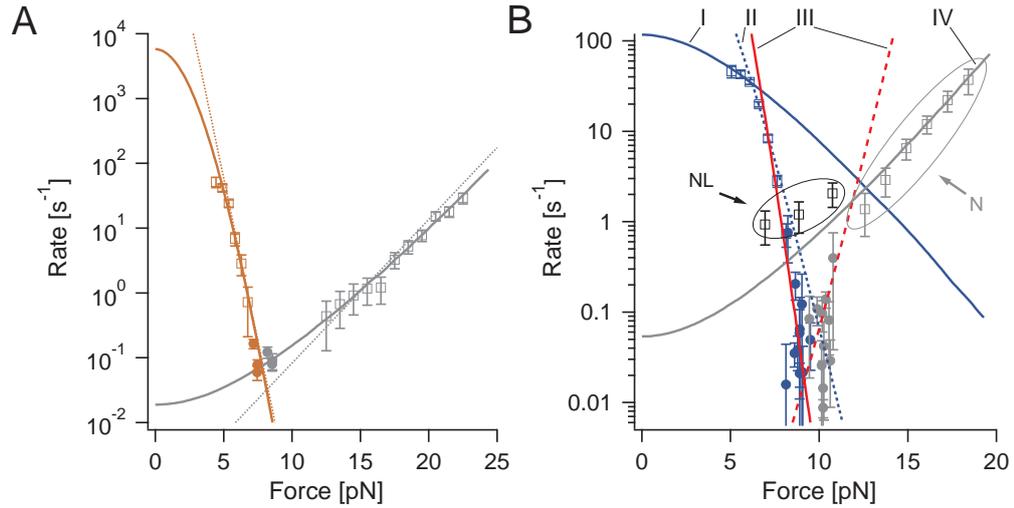


Figure 52: ROSS has split and kinked rate plots. Empty squares represent OBS rates, full circles are rates from constant distance measurements extracted by HMM analysis. (A) Thick solid lines are SB-model fits to the data. Dotted lines represent the ZB-model according to previously extracted fit parameters from corresponding force distributions, see Table 8. (B) All lines are local SB-model fits to the data. Fitting ranges are I: 5 - 6 pN, II: 6 - 7 pN, and III: 7 - 9.5 pN for refolding; IV: 16.5 - 18 pN for unfolding. The red dashed line with a positive slope is derived from its 'counterpart' (red line) based on detailed balance and assuming  $\Delta x_u + \Delta x_f = L_P$ .

refolding within force range III (red line). Therefore, the unknown zero-force rate constant  $k_0^u$  is derived according to the principle of detailed balance, see Equation 32, upon using the fitted  $k_0^f$  in connection with the experimentally determined folding free energy of  $\Delta G_0^{CD,indiv} = 27.7 k_B T$  for ROSS, see Table 10. The unknown slope  $\Delta x_u$  is inferred by the assumption that transition state positions obtained with the SB-model add up to about the full contour length change which is involved in the transition<sup>6</sup>, i. e.,  $\Delta x_u + \Delta x_f = L_P$  in our case. While measured rates from constant distance measurements are not precise enough to confirm the correctness of the 'inferred' fit, the shape of this fit adequately catches the overall trend within the experimental data.

Besides multiple kinks, the unfolding rates of ROSS also feature splitting. The two encircled unfolding branches in Figure 52B mark the two regimes where either NL (black empty squares) or N (grey empty squares) predominate. Here, the split rates are the rate plot analogy to the bi-modal unfolding force histogram in Figure 51B which, again, proclaims multi-pathway transitions in a multidimensional energy landscape [170]. It is important to note that the unfolding rates extracted from constant distance measurements do not overlap with the unfolding rates of NL. This means that a constant force bias allows the separation of N and NL kinetics and that all folding transitions into the 'native' configuration of ROSS can be correctly identified.

<sup>6</sup> Note that this assumption is confirmed for FLred in Section 5.2.2.

PROTEIN	$\log_{10}k_0^f$ [s <sup>-1</sup> ]	$\log_{10}k_0^u$ [s <sup>-1</sup> ]	$\Delta x_f$ [nm]	$\Delta x_u$ [nm]	$\lambda_{\text{mid}}$ [s <sup>-1</sup> ]	$F_{\text{mid}}$ [pN]
R15	$3.8 \pm 0.1$	$-1.7 \pm 0.1$	$23.7 \pm 0.2$	$2.9 \pm 0.1$	0.1	7.8
ROSS (I)	$2.1 \pm 0.1$	-	$3.1 \pm 0.7$	-	-	-
ROSS (II)	$3.8 \pm 0.1$	-	$12.6 \pm 0.2$	-	-	-
ROSS (III)	$6.4 \pm 0.2$	-	$22.4 \pm 0.4$	-	< 0.1	9.5
ROSS (IV)	-	$-1.3 \pm 0.4$	-	$3.0 \pm 1.1$	-	-

Table 9: Kinetic SB-model fit parameters of R15 and ROSS based on OBS and HMM rates extracted from all constant velocity and constant distance measurements, see [Figure 52](#). Parameters were obtained by local fits covering the following force ranges I: 5 - 6 pN, II: 6 - 7 pN, and III: 7 - 9.5 pN for refolding; IV: 16.5 - 18 pN for unfolding.

The fit results collected in [Table 9](#) suggest that the folding energy landscape of ROSS has at least four distinct transition state positions. Further, local DHS-model fits (not shown) confirm that the observed kinks within the rate plots are stronger than the bending which is expected due to mere Hammond behaviour<sup>7</sup>. By this means, the amount of four transition states is not a result from overfitting the data. On the contrary, all four identified transition states stand in agreement with the up to three transiently populated high-energy on-pathway intermediates identified upon closer inspection of all transition events in [Section 6.4.3](#). Note that the force ranges over which the SB-model fits capture the tendencies within experimental rates become broader with rising force. This additionally supports the view of force-induced transition state switching: transition states which are closer to the folded state become predominant with increasing force while, at the same time, the force lever becomes shorter.

[Table 9](#) also includes the combined rate  $\lambda_{\text{mid}} = k_{\text{mid}}^u + k_{\text{mid}}^f$  measured at the mid-force  $F_{\text{mid}}$  where unfolding and refolding rates equal each other. To a large extent, the substantially higher  $F_{\text{mid}}$  value of ROSS is a result of its more than two times higher folding free energy. On top of that, the smaller  $\lambda_{\text{mid}}$  attests ROSS a higher durability than R15 under a constant force bias.

As the kinked rate plots suggest a rough energy landscape which implies a multitude of states one might obviously ask: Why not using [HMM](#) for a more detailed analysis and characterization of the equilibrium fluctuations of ROSS? One reason for not doing so is the high numeric cost which is involved when the applied [HMM](#) algorithm faces hundreds of minutes of high-resolution 200 kHz data in combination with at least 6 necessary states<sup>8</sup>. Due to the very short-lived intermediates and the very slow overall folding and un-

<sup>7</sup> For a discussion of transition state movement inferred by Hammond behaviour see [Section 5.2.4](#) together with [Section 4.2.4](#) and [Section 5.3.1](#).

<sup>8</sup> The necessary states are two folded states representing N and NL, three on-pathway intermediates and one unfolded state.

folding rates under a constant force bias, unfortunately both high resolution and very long trajectories are necessary.

PROTEIN	$\Delta G_0^{\text{bulk}}$	$\Delta G_0^{\text{k}_0}$	$\Delta G_0^{\text{k}_0(\text{ZB})}$	$\Delta G_0^{\text{CD,indiv}}$	$L_p^{\text{calc}}$	$L_p^{\text{CV}}$
			[k <sub>B</sub> T]			[nm]
R15	11.5	$12.7 \pm 1.3$	$23.3 \pm 2.3$	$16.1 \pm 1.6$	34.47	$33.4 \pm 1.1$
ROSS	25.2	-	$22.6 \pm 2.3$	$27.7 \pm 2.8$	34.51	$34.7 \pm 0.9$

Table 10: Average contour length gains and folding free energies of R15 and ROSS.  $L_p^{\text{CV}}$  values represent the average from all constant velocity data.

Table 10 lists differently obtained folding free energy values for R15 and ROSS. In case of R15, the  $\Delta G_0^{\text{k}_0}$  value of  $12.7 \pm 1.3$  k<sub>B</sub>T is further considered as the best experimental result of this study as it is based on the largest amount of available data. For ROSS,  $\Delta G_0^{\text{CD,indiv}}$  provides the most reliable result of  $27.7 \pm 2.8$  k<sub>B</sub>T. For both R15 and ROSS, the single molecule folding free energies are in very good agreement with the results reported from bulk measurements.

The investigation of force-dependent transition rates combined from both constant distance and constant velocity measurements confirmed that R15 mechanics can be well modelled by assuming a simple two-state mechanism over the entire force range. Equally, signs for both roughness and multidimensionality of ROSS are found in its rate plots represented by at least four distinct slopes and two distinct unfolding branches. Further, under a constant force bias, its higher folding free energy makes ROSS more durable than R15.

### 6.2.3 Transition averages and prolonged transition path times

In this part we start to move on from the mere detection of protein conformational transitions towards a direct evaluation of transition events themselves. Here, both transition averages and individual transition path times will reveal that typical time scales of R15 and ROSS folding are separated by more than two orders of magnitude.

As elaborated in Section 4.3.3, averages of detected protein folding transitions are dynamic fingerprints which provide information on typical protein folding time scales. To get down to this information, detected averages need to be compared to averages from unperturbed (= 'protein-free') bead diffusion. Figure 53 presents such a comparison for R15 and ROSS which includes averaged transitions as well as bead relaxation curves. In Figure 53A, the averaged transitions of R15 unfolding and refolding (empty circles) are nearly as fast as what is expected for unperturbed bead diffusion (black dashed line). Here, the expected shape is based on over a thousand transitions extracted from a one-dimensional bead diffusion simulation in the experimentally determined bead potential according to Equation 60 using the known bead friction coefficient  $\gamma_B = 9.4 \cdot 10^{-6}$  pN · s · nm<sup>-1</sup>. Note that the higher noise level

within the detected averages is due to their limited statistics. In contrast to R15, the averaged transitions of ROSS are much slower as one would expect for unperturbed bead diffusion, see [Figure 53B](#). Now, following argumentation in [Section 4.3.3.2](#), merely by comparing detected to expected averaged transitions we can already tell that the upper limit for R15 folding is a few tens of microseconds while for ROSS it takes about a millisecond to fold.

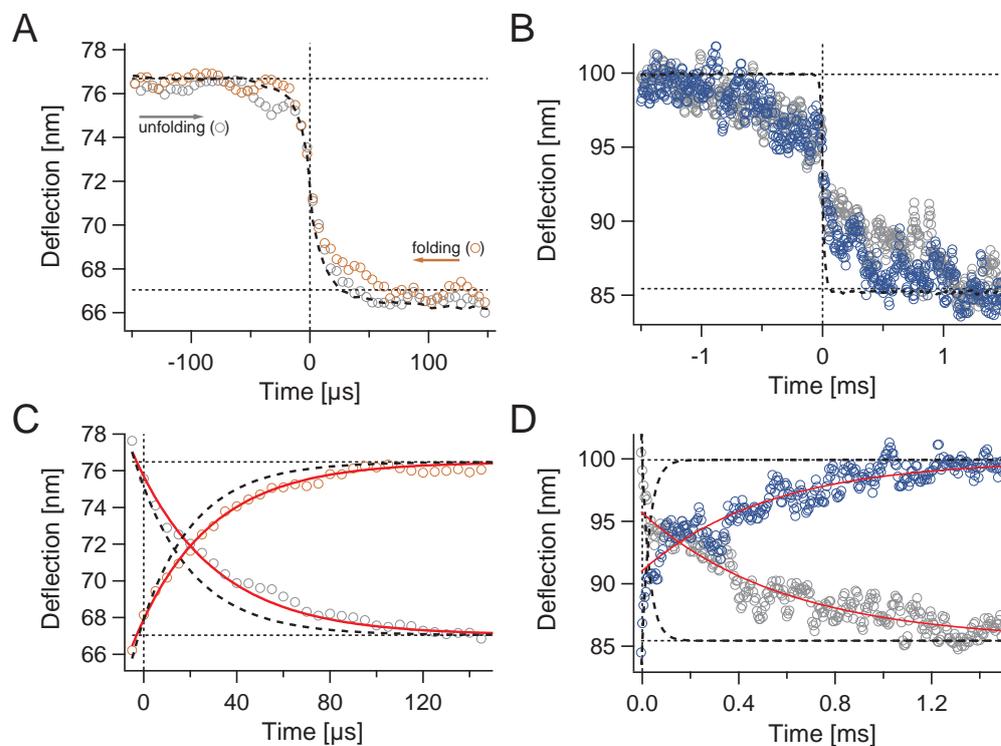


Figure 53: Transition averages of R15 and ROSS. (A) Empty circles represent averaged transitions of 55 unfolding (grey) and 55 time-reversed refolding (brick red) events from one individual constant distance trajectory of R15. The black dashed line is the average of 1817 transitions obtained from a one-dimensional bead diffusion simulation which mimics ‘protein-free’ diffusion. (B) Averaged ROSS transitions from 6 unfolding and 6 refolding events. The thick black dashed line is based on 4563 simulated transitions. (C) Bead relaxation curves of R15 based on the same data set as in (A). Red lines are single-exponential fits. Thick black dashed lines are respective fits to the ‘protein-free’ signal. (D) Bead relaxation curves of ROSS.

The bead relaxation curves shown in [Figure 53C](#) and [Figure 53D](#) clearly illustrate the retardation effect inferred by additional protein diffusion for both R15 and ROSS. A simple way to quantify this effect is by comparing the measured relaxation times  $\tau_{\text{relax}}$  to the expected ones  $\tau_{\text{relax}}^{\text{1D}}$ . The corresponding single-exponential fit results are listed in [Table 11](#). While for R15 the fastest possible system response represented by  $\tau_{\text{relax}}^{\text{1D}}$  is delayed by only +10 μs, for ROSS it is already about +0.5 ms.

After studying transition averages, now we go into further detail by evaluating the transition path time  $\tau_{\text{TP}}$  provided by each transition. Examples of

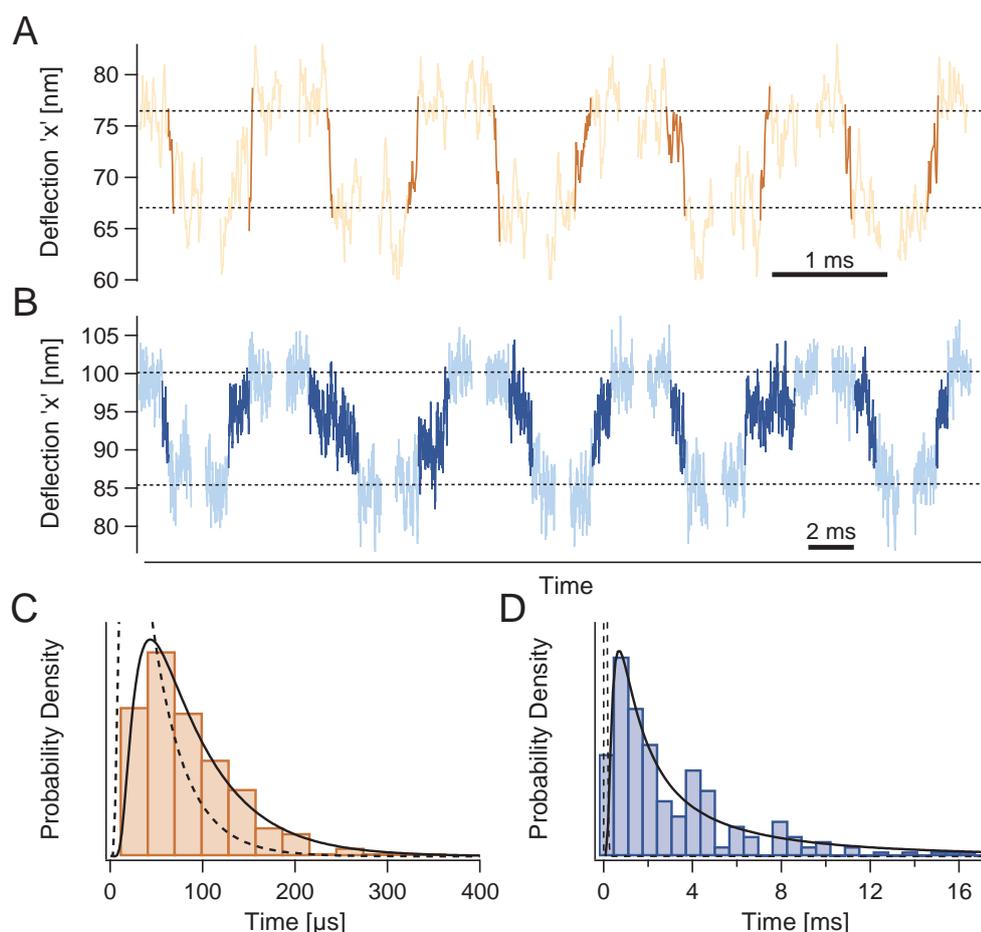


Figure 54: Individual equilibrium transitions and transition path time distributions of R15 and ROSS. (A) and (B) Zoom into 10 successive unfolding/refolding transitions of R15 and ROSS (light colors). Transition paths are highlighted (dark colors). Horizontal lines mark the folded/unfolded state levels. (B) and (C) Transition path time distributions of R15 and ROSS based on 259 and 201 equilibrium transitions. Black lines are fits according to Equation 57. Dashed lines are fits referring to unaltered bead diffusion.

individual R15 and ROSS transitions are illustrated in Figure 54A and Figure 54B. Here, dashed horizontal lines mark the levels of the folded (top) and unfolded (bottom) state. In case of R15, transition paths are defined as the shortest interconnection between these two levels within direct proximity of a two-state transition identified by HMM analysis. The identification by HMM is necessary to discern actual transitions from mere thermal noise. In case of ROSS, already the beginning and end of transition paths are thermally blurred due to their relatively long duration. This is why transition paths of ROSS are treated as if they were composed of a series of two adjoining, short-lived, and rapidly exchanging intermediate states<sup>9</sup>. The corresponding HMM

<sup>9</sup> The choice of two on-pathway intermediates is not by chance. Figure 66 in Section 6.4.3 reveals that there are at least two obligatory on-pathway intermediates involved in (un-)folding transitions of ROSS.

analysis then effectively does the differentiation between thermal noise along the transition path and actual crossing of the folded and unfolded state level, see also [Figure 20B](#) in [Section 3.8.5](#). As both folded and unfolded states have much longer lifetimes than the assumed intermediates along the transition paths, the HMM-assisted transition path identification is readily implemented.

Graphically the information provided by the analysis of extracted transition path times  $\tau_{\text{TP}}$  is presented in [Figure 54C](#) and [Figure 54D](#). Here, the transition path time distributions of R15 and ROSS are fitted according to [Equation 57](#) (black solid lines) and compared to respective fits of unperturbed bead diffusion (black dashed lines). Besides a fixed temperature of  $T = 298$  K, the ‘protein-free’ reference fits as well as the fit to the R15 distribution assume a fixed diffusion coefficient of  $D = k_{\text{B}}T/\gamma_{\text{B}}$ , where  $\gamma_{\text{B}}$  is the bead friction coefficient. In case of ROSS,  $D$  needs to be a free parameter to enable convergence of the fit. However, the parameters provided by all fits which further include the barrier height  $\Delta G^{\text{TS}}$  and curvature  $\omega_{\text{TS}}^2$  at the barrier top, are not of primary interest. Mainly this is due to the numerical study performed in [Section 4.2.1](#) which revealed that parameters obtained from fitting transition path time distributions are highly error-prone. The two characteristic parameters which are deduced from the transition path time distributions in [Figure 54](#) are the most probable transition path times  $\tau_{\text{TP}}^{\text{max}}$  as well as transition path time averages  $\langle \tau_{\text{TP}} \rangle$ , see [Table 11](#).

PROTEIN	$\tau_{\text{relax}}^{\text{1D}}$	$\tau_{\text{relax}}$	$\tau_{\text{TP}}^{\text{max,1D}}$	$\tau_{\text{TP}}^{\text{max}}$	$\langle \tau_{\text{TP}}^{\text{1D}} \rangle$	$\langle \tau_{\text{TP}} \rangle$
	[ $\mu\text{s}$ ]					
R15	23	33	19	43	60	83
ROSS	31	555	32	690	57	3923

Table 11: Transition path time averages and bead relaxation times of R15 and ROSS. The additional ‘1D’ superscript is given to reference values from one-dimensional bead diffusion simulations unaltered by protein diffusion.

All experimentally determined time scales listed in [Table 11](#) are characteristic for the protein folding transitions of R15 and ROSS. However, all these time scales are basically a convolution of two time scales comprising bead and protein diffusion. The results of the extensive numerical study in [Section 4.3.3.2](#) give insight into the relative and absolute effects of different protein friction coefficients on expected and measured characteristic time scales, see [Table 2](#). By mere comparison, here, we can make a first rough guess on the absolute protein folding transition path times of R15 and ROSS under force. In case of R15, which leads to an increase of  $+10 \mu\text{s}$  for  $\tau_{\text{relax}}$  and  $+23 \mu\text{s}$  for  $\langle \tau_{\text{TP}} \rangle$  with respect to the ‘protein-free’ values, a rough estimate for the actual R15 transition path time is  $\langle \tau_{\text{TP}}^{\text{R15}} \rangle \sim 10 - 30 \mu\text{s}$ . For ROSS, the additional contribution by bead diffusion is negligible and, hence,  $\langle \tau_{\text{TP}}^{\text{ROSS}} \rangle \sim 3.9$  ms. Further, it may be of interest to note that in case of ROSS there is a relatively huge difference between the most probable transition path time and its average due to a quite

broad distribution. With respect to the results from the previous sections it seems legitimate to claim that this broad transition path time distribution provides another piece of evidence for a roughened energy landscape.

Of course, results in [Table 2](#) and [Table 11](#) are based on different energy landscape parameters and, hence, are not directly interconvertible. But, after all, the dimensions of the deconvolved protein folding energy landscapes of R15 and ROSS, see [Section 6.2.4](#), at most differ only by a factor of 2 from the landscape used in the numerical study, see [Section 4.3.1](#).

Taken together, this section's results confirm that the direct evaluation of individual transitions provides useful information to get rough time scale estimates of the involved protein folding. Here, in particular, the estimated average transition path times are  $\langle \tau_{TP}^{R15} \rangle \sim 10 - 30 \mu\text{s}$  for the natural protein R15 and  $\langle \tau_{TP}^{ROSS} \rangle \sim 3.9 \text{ms}$  for the artificial Rossmann fold. Based on these values, the artificial protein is lagging behind the natural one by over two orders of magnitude.

#### 6.2.4 Deconvolution and commitment probabilities

In the following, deconvolution will be applied to get hold of the folding energy landscapes of R15 and ROSS in their entirety. Alongside, the committor-based barrier shape reconstruction, see [Section 3.8.2](#), is put into practice. While the latter yields meaningful results with respect to deconvolution, the deconvolved energy landscapes themselves partially stand in conflict with previous kinetic results.

The deconvolved energy landscapes shown as dashed light blue lines in [Figure 55B](#) and [Figure 56B](#) are Boltzmann-inverted probability distributions deconvolved from measured distributions of R15 and ROSS. Here, measured distributions are obtained from single equilibrium trajectories recorded for several minutes in constant distance mode around the respective mid-force, see [Table 9](#) for  $F_{\text{mid}}$ . Red dashed lines represent energy landscapes directly inferred by Boltzmann inversion of the measured probability distributions prior deconvolution. The relevant probability distributions themselves are shown in [Section 6.2.5](#) in [Figure 58A](#) and [Figure 58D](#) respectively. The barrier shapes of the directly measured and deconvolved energy landscapes can be used to calculate the corresponding committors by solving [Equation 51](#) for  $p_i(x)$  followed by rescaling of the resulting  $p_i(x)$  to its correct range between 0 and 1. Final results are indicated as dashed lines in [Figure 55A](#) and [Figure 56A](#) where full lines represent committors directly calculated from the respective equilibrium trajectory according to [Equation 49](#). Finally, [Figure 55C](#) and [Figure 56C](#) presents the transition path probabilities directly derived from the committors in [Figure 55A](#) and [Figure 56A](#) according to [Equation 53](#).

For both R15 and ROSS, the overall shape of their committors which are directly calculated from the respective equilibrium trajectory (solid lines in [Figure 55A](#) and [Figure 56A](#)) are more resemblant to the commitment probabilities derived from the deconvolved energy landscapes (dashed light blue lines) than to those derived from the directly measured potential (dashed

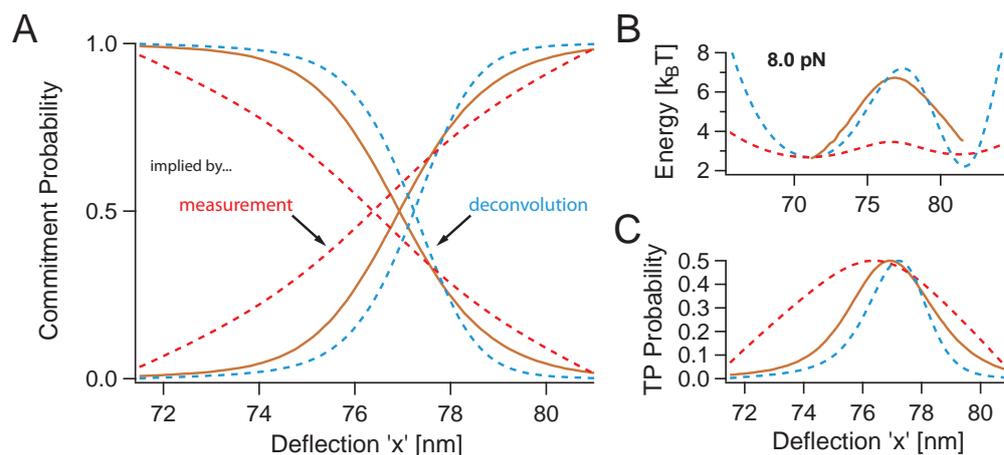


Figure 55: Committor-based barrier reconstruction for R15. Dashed lines are theoretically expected curves based on the measured (red) and deconvolved (light blue) probability distribution of the analysed trajectory. (A) Commitment probability of R15 (full line) derived from an equilibrium trajectory according to Equation 49. (B) Committor-based barrier shape reconstruction according to Equation 51. Dashed lines represent the convolved (red) and deconvolved (light blue) bead potential. The indicated force is the mean between forces acting on the folded and unfolded state levels. (C) Transition path probability derived from committors in (A) based on Equation 53.

red lines). By this means, also the barrier heights in Figure 55B and Figure 56B are reconstructed reasonably well with respect to the deconvolved potential. However, one can also see that the barriers' curvatures are more resemblant to those of the energy landscapes directly inferred by the recorded motion of the beads. Like their respective committors, transition path probabilities in Figure 55C and Figure 56C are clearly more similar to expectations risen by deconvolution than by direct measurement. Taken together, these results attest the committor calculation based on Equation 49 as well as the committor-based barrier reconstruction according to Equation 51 reasonable performance with respect to deconvolved energy landscapes.

Next, the results of this section are viewed in relation to simulations in the same matter presented in Section 4.3.4. Since the deconvolved energy landscapes of R15 and ROSS are both slightly asymmetric, simulation results presented in Figure 37 are of highest interest followed by results in Figure 34. What experiments and simulations definitely have in common is the 'conservation' of the barrier shape of the directly measured bead potential when trying to reconstruct the deconvolved barrier. A simple but yet important conclusion directly follows from the commitment probabilities of R15 and ROSS by being more resemblant to the ones expected from deconvolution than to those expected for mere bead diffusion: both R15 and ROSS are not folding at the speed limit (see black lines in Figure 34A and Figure 37A) which is in perfect agreement with our previous results. Further there are hints which potentially reveal that R15 effectively experiences less while ROSS most likely experiences more friction than the beads. All these hints basically originate

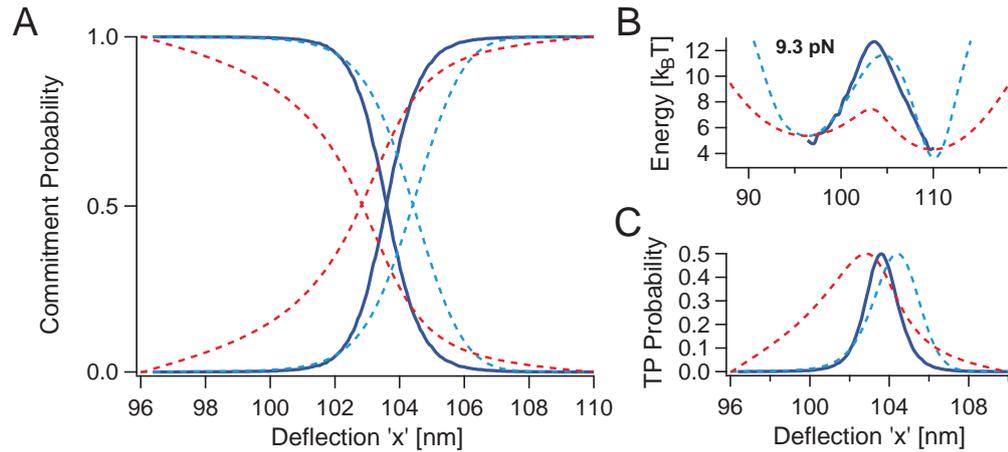


Figure 56: Committor-based barrier reconstruction for ROSS. Dashed lines are theoretically expected curves based on the measured (red) and deconvolved (light blue) probability distribution of the analysed trajectory. (A) Commitment probability of ROSS (full line) derived from an equilibrium trajectory according to Equation 49. (B) Committor-based barrier shape reconstruction according to Equation 51. Dashed lines represent the convolved (red) and deconvolved (light blue) bead potential. The indicated force is the mean between forces acting on the folded and unfolded state levels. (C) Transition path probability derived from committors in (A) based on Equation 53.

from a shallower slope of the commitment probability of R15 and a steeper slope of ROSS when being compared to the respective shape expected from deconvolution. Hence, other resulting hints might be the lower versus higher than expected reconstructed barriers as well as the broader versus more narrow than expected transition path probabilities. With respect to Section 4.3.2, effectively experiencing different friction than the bead corresponds to R15 and ROSS having a friction coefficient different from the 'sweet spot' where transition path times in contour as well as deflection space are about equal. Along these lines, the results from this section potentially imply that  $\gamma_p^{R15} < 1.5 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  while  $\gamma_p^{ROSS} > 3 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . Additionally, in case of R15, the protein folding speed limit implies a lower border for the friction coefficient. Based on the simple formula  $N/100 \mu\text{s}$ , where  $N$  represents the number of amino acids involved in unfolding, full-length speed limit unfolding of all 114 amino acids of R15 would happen as fast as in  $1.14 \mu\text{s}$  [119]. From this very short transition path time it follows that  $\gamma_p^{R15} > 4.8 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , where specifically '>' instead of '≥' is used since R15 folding is slower than expected for speed limit protein folding.

Altogether, the direct commitment probability calculations from equilibrium trajectories, their comparison to expectations risen by the respective deconvolved energy landscapes, and the review of this section's results in the context of similar simulations revealed a couple of things. Concerning the committor-based barrier shape reconstruction, barrier heights are reasonable with respect to deconvolution while their shapes tend to be more resemblant

to the measured bead potential which has been similarly observed in simulations. Independent of deconvolution and, hence, more importantly, both R15 and ROSS are not folding at the speed limit which directly follows from the strong deviation of their commitment probabilities from the shape expected for mere bead diffusion. Further, R15 effectively experiences less while ROSS effectively experiences higher friction than the beads. In combination, this yields  $4.8 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1} < \gamma_{\text{p}}^{\text{R15}} < 1.5 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  for the friction coefficient of R15 and  $\gamma_{\text{p}}^{\text{ROSS}} > 3 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  for ROSS.

#### 6.2.4.1 Kinetic parameter mismatch: deconvolution versus kinetics

This subsection briefly compares kinetic key parameters provided by deconvolution to those obtained from fitting force-dependent rates. Despite a generally good agreement, two discrepancies are noteworthy as they corroborate an insensitivity of deconvolution towards strong asymmetries and short-lived intermediates.

Figure 57 presents deconvolved energy landscapes of R15 and ROSS which were transformed from bead deflection into (unfolded) contour length space. Note that the darker coloured lines in Figure 57 correspond to the landscapes shown as dashed light blue lines in Figure 55B and Figure 56B. After a zero-force transformation (lighter colours), the folding free energy  $\Delta G_{0^*}$  between folded and unfolded states is readily extracted, see dashed horizontal lines. The obtained values are in very good agreement with  $\Delta G_0^{\text{CD,indiv}}$  calculated from individual constant distance trajectories, see Table 10. By this means, the relative energy between the two states separated by a barrier is correctly determined by deconvolution. Another important characteristic of a barrier is the position of the barrier top, i. e., the transition state position typically denoted as  $\Delta x_{\text{u}}$  and  $\Delta x_{\text{f}}$  for unfolding and folding transitions. In Figure 57, the transition state position provided by deconvolution is marked by black horizontal arrows. In contrast, red horizontal arrows reflect transition state positions derived from SB-model fits to the force-dependent rates of R15 and ROSS, see Table 9. The discrepancies between transition states marked by black and red arrows obviously raise a few questions which are discussed in the following.

In case of R15, the transition state position identified as  $\Delta x_{\text{f}}$  for folding transitions in Figure 57A is about the same for both deconvolution and measured folding kinetics. However, for unfolding, the values for  $\Delta x_{\text{u}}$  differ by more than a factor of four. By this means it seems as if deconvolution is not capable of reproducing the high asymmetry which is measured for R15 kinetics, see Figure 52A. Instead, the transition state position ratio  $\Delta x_{\text{u}}/\Delta x_{\text{f}}$  formerly introduced as a measure of symmetry in Section 5.3.1 changes from 0.12 implied by kinetics to 0.63 implied by deconvolution which already approaches perfect symmetry which would be given in case of 1.0. Two aspects need to be addressed in view of the observed discrepancy between the (unfolding) transition state position provided by either kinetics or deconvolution: Hammond behaviour and the ‘missing’ contour length when trying to reproduce the full contour length  $L_{\text{p}}$  of R15 by adding up  $\Delta x_{\text{u}}$  and  $\Delta x_{\text{f}}$  provided by

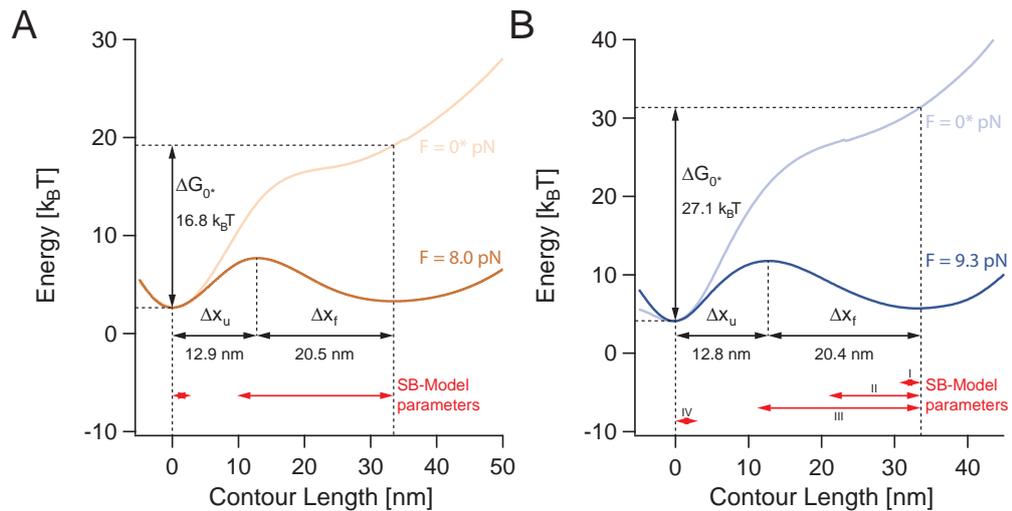


Figure 57: R15 and ROSS point towards two fundamental weaknesses of deconvolution. Red arrows indicate transition state positions according to the SB-model fits in Figure 52, for parameters see Table 9. (A) Asymmetry: despite the strongly asymmetric rate plots of R15, the deconvolved potentials shows only weak asymmetry. (B) Short-lived high-energy intermediates: deconvolution is insensitive to the energy landscape roughness of ROSS which is suggested by its kinked rate plots.

the SB-model. As the SB-model fits the unfolding rates of R15 over the entire experimentally accessible force range, a mere force-induced transition state movement known as Hammond behaviour can be excluded as an explanation for the observed discrepancy<sup>10</sup>. Concerning the ‘missing’ contour length, one possible interpretation could be a very broad transition state ensemble of R15 which, to a certain extent, would prove both  $\Delta x_u$  values to be right. Another possible explanation would be a folding mechanism of R15 which is partially orthogonal to the acting force and, hence, invisible to the measurement signal. As deconvolution innately assumes a purely one-dimensional folding mechanism, the transition state position provided by the deconvolved energy landscape may be interpreted as some sort of effective position. So does the hypothesis of a limited capability of reconstructing strong asymmetries by deconvolution hold? The strongest argument in support of this hypothesis is given by the two-dimensional diffusion simulations performed in an asymmetric protein folding potential in Section 4.3.4. In particular, in Figure 36B, deconvolution of the simulated deflection signal followed by a contour length transformation yields a more central transition state position than the used protein folding energy landscape actually had. Upon closer inspection of all deconvolved energy landscapes, see Figure 34B, Figure 37B, Figure 55B, and Figure 56B, it is striking that all deconvolved transition state positions are within close proximity of the transition state of the convolved potential. This leads to a more refined hypothesis: deconvolution is capable to resolve asym-

<sup>10</sup> See Section 4.2.4, Section 5.3.1, and Section 5.3.2 for detailed discussions about rate plots, transition state positions, Hammond behaviour, deconvolution and different kinetic models.

metries in protein folding energy landscape transition state positions to an extent which goes only slightly beyond what is already provided by the convolved potential.

When it comes to ROSS, we lack a pair of  $\Delta x_u$  and  $\Delta x_f$  values from SB-model fits with an overlapping force range which could be directly compared to deconvolution. Nevertheless, the indicated force bias of 9.3 pN which acts on the deconvolved energy landscape of ROSS in [Figure 57B](#) falls into force range III of refolding rates fitted by the SB-model in [Figure 52B](#). The appropriately marked red arrow in [Figure 57B](#) indicates a similar  $\Delta x_f$  for both deconvolution and the SB-model. What is remarkable about the deconvolved energy landscape of ROSS is that any sign of roughness is missing. By contrast, the kinked rate plots of ROSS strongly suggest an underlying rough energy landscape with multiple distinct transition states the existence of which will be further proven in [Section 6.4.3](#) and [Section 6.4.4](#). Then why does the deconvolved energy landscape of ROSS only have one transition state with a smooth barrier? A simple explanation is the small weight of the short-lived on-pathway intermediates within the overall probability distribution. In other words, the roughness of ROSS lies beyond the resolution of deconvolution.

In short, transition state positions cannot be correctly resolved by deconvolution unless they lie within proximity of the transition state positions already provided by the convolved potential. Short-lived and rarely populated intermediates which are 'deeply buried' within the thermodynamic noise of predominantly populated states are likely not to be resolved by deconvolution at all. Differently speaking, deconvolution does not provide the most accurate kinetic parameters and details. However, by representing a smoothed average, deconvolved energy landscapes can be expected to provide a robust overall picture which roughly characterizes the underlying kinetics.

What do the above findings mean for our results and their interpretation? Since the deconvolved energy landscape shapes are put under question, so have to be the thereupon based transition path time and friction coefficient calculations. In general, transition state positions derived by deconvolution may not provide the best reference for reaction coordinate quality tests. An exception are systems which do have a relatively central transition state and a folding mechanism which fully projects onto the acting force axis as, for instance, is given for [DNA-hairpin](#) or [coiled-coil protein unzipping](#).

### 6.2.5 *Transition path ensembles and probabilities*

In addition to the temporal information provided by transition path times, transition paths also provide spatial information which is now introduced by the experimentally determined transition path ensembles of R15 and ROSS. While the measured ensembles confirm the presence of protein folding, the thereupon based transition path probabilities fail to directly localise deconvolved transition state positions.

[Figure 58](#) presents all ingredients which are necessary to calculate the conditional probability of being on a transition path from experimental trajec-

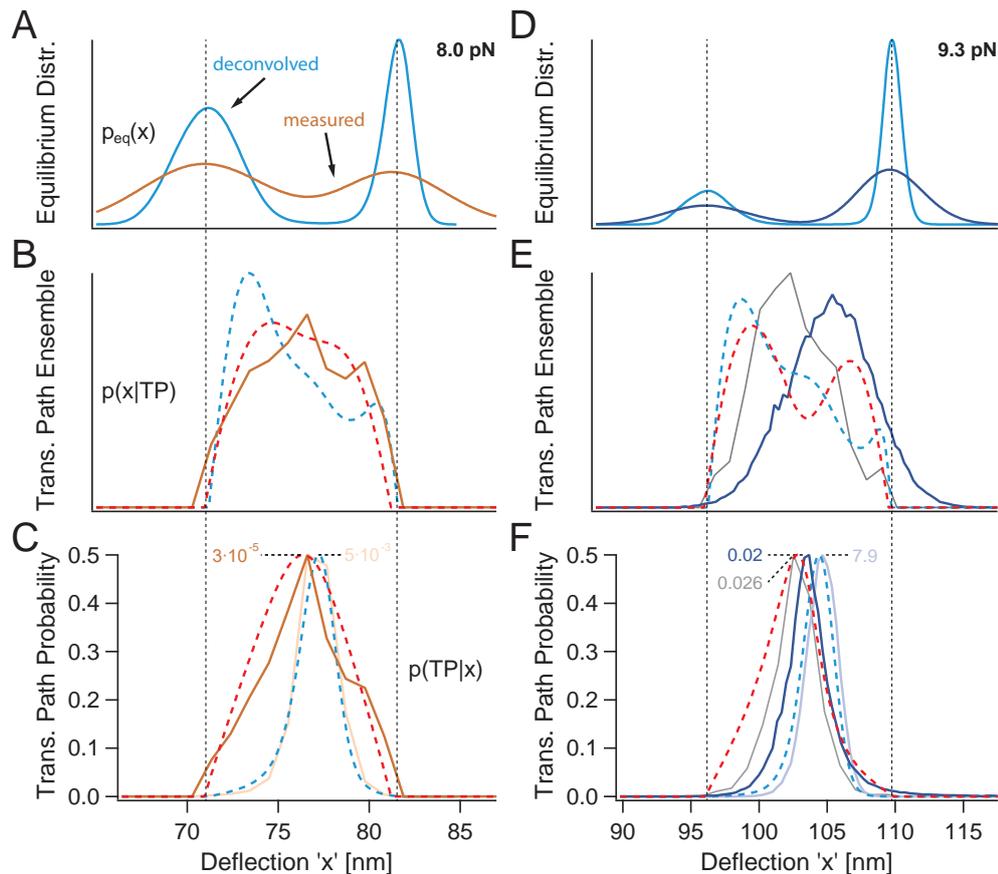


Figure 58: Transition path probability calculations for ROSS and R15. Dashed red and light blue lines are theoretically expected curves based on measured and deconvolved probability distributions. (A) Measured and deconvolved probability distributions of R15. (B) Transition path ensemble of R15 based on 110 consecutive unfolding/refolding transitions. (C) Transition path probabilities of R15 according to Equation 52 using measured (brick red) or deconvolved (orange) data (not to scale). (D) - (F) Equilibrium distributions, transition path ensembles and probabilities of ROSS based on 12 consecutive unfolding/refolding transitions. See text for more details.

ries based on Equation 52. The measured (brick red) and deconvolved (light blue) probability distributions of an equilibrium trajectory of R15 are shown in Figure 58A. Below, in Figure 58B, the measured transition path ensemble of R15, which is the probability distribution of all points recorded along extracted transition paths, is illustrated. Here, dashed red and light blue lines represent theoretically expected ensembles based on the measured and deconvolved probability distributions from above. The derivation of expected transition path ensembles is described in detail in Section 4.3.2.2. And finally, (conditional) transition path probabilities of R15 implied by Equation 52 are depicted in Figure 58C. In more detail, the brick red transition path probability is derived upon using the measured equilibrium distribution as  $p_{\text{eq}}(x)$  in Equation 52, while the pale red probability is based on the deconvolved equilibrium distribution and an additional normalization factor according to [155].

Note that in [Figure 58C](#), expected shapes are drawn to scale while the peak values of experimentally determined probabilities are indicated separately. All corresponding graphs of ROSS are presented in [Figure 58D](#) - [Figure 58F](#) right next to the results of R15. The additional thin grey line in [Figure 58E](#) depicts the transition path ensemble of ROSS obtained upon extracting transitions paths delimited by mere folded and unfolded state level crossing as opposed to passing through two HMM-identified on-pathway intermediates, see also [Section 6.2.3](#). The associated transition path probability is represented by the thin grey line in [Figure 58F](#) and only calculated using the measured equilibrium distribution.

The transition path ensembles of R15 and ROSS neither match shapes expected for unperturbed bead diffusion (red dashed lines) nor shapes derived from deconvolved probability distributions (light blue dashed lines), see [Figure 58B](#) and [Figure 58E](#). By contrast, both measured ensembles are more or less centrally peaked distributions which, according to simulations in [Section 4.3.2.3](#), is expected to arise from and to increase with protein friction. By this means, the measured ensembles indirectly reveal the presence of protein folding. Compared to ROSS, deviations in case of R15 are less pronounced which is in accord with previous results suggesting that ROSS has a much higher protein friction coefficient. However, as it is R15 which behaves like a simple two-state folder, the deviation of its transition path ensemble is further analysed in [Section 6.4.2](#) to make an estimate of the involved protein friction based on simple diffusion.

There are two important aspects when it comes to transition path probabilities: the position and the absolute probability value of their main peak. While the peak position is supposed to identify the relevant transition state position, the absolute value is considered as a measure of reaction coordinate quality which is best if the theoretical maximum of 0.5 is reached, see [Section 3.8.3](#). In the following, we will first focus on the overall transition path probability shapes and peak positions.

The brick red line in [Figure 58C](#) illustrates the transition path probability of R15 where the directly measured probability distribution enters [Equation 52](#) as  $p_{\text{eq}}(x)$ . The resulting shape and peak position pretty much fulfil expectations for unperturbed bead diffusion (dashed red line). By contrast, if the deconvolved equilibrium probability distribution is used as denominator for the transition path probability calculation, the resulting orange line is almost identical to the dashed light blue one which represents the shape and peak position implied by the deconvolved potential. In congruence with simulation results obtained in [Section 4.3.4](#), the strong dependence on  $p_{\text{eq}}(x)$  discloses a general weakness of the transition path probability calculation: the overall shape of the resulting probability is basically predefined by the probability distribution entering [Equation 52](#). By this means, an intended transition state localization essentially loses its foundation as the final result already enters the calculation. When moving on to the shape of the transition path probability of ROSS, a very similar scenario can be seen. If the deconvolved light blue probability distribution from [Figure 58D](#) enters [Equation 52](#), the shape of the

resulting pale blue transition path probability in [Figure 58F](#) is almost identical to the expected dashed light blue one. At least, the dark blue probability, which is based on the measured probability distribution of ROSS, reveals a clear tendency away from the shape expected for unperturbed bead diffusion towards the deconvolved shape. Here, this is mainly due to high enough protein friction which accordingly alters bead diffusion. However, the thin grey probability reveals that the information of a slight transition state shift away from the one of the bead potential is lost when the transition path ensemble of ROSS is derived from transition paths that are merely delimited by folded and unfolded state level crossings.

Now, the absolute transition path probability values get into focus. Judging from the extremely small transition path probability maxima of R15 in [Figure 58C](#), the reaction coordinate quality seems to be very poor. To a large extent, these extremely low probabilities are due to a very small  $p(\text{TP})$ , i. e., the relative amount of time spent on transition paths. As transition paths of R15 are preselected by HMM, all 'pseudo' transitions within the noise are not considered which technically reduces  $p(\text{TP})$ . On the other hand, in case of the measured equilibrium distribution, all noise remains still included which yields relatively large values for  $p_{\text{eq}}(x)$  around the transition state position. In combination, these two effects strongly reduce the transition path probability of R15. As deconvolution basically removes noise from the measured equilibrium distribution, application of the deconvolved equilibrium distribution as  $p_{\text{eq}}(x)$  in [Equation 52](#) automatically yields a much higher transition path probability. This probability is then further increased by the additional normalization factor which, in principle, can be interpreted as being supposed to come up for a reduced  $p(\text{TP})$  value. However, a maximal probability of  $5 \cdot 10^{-3}$  still indicates a poor reaction coordinate. A poor reaction coordinate, in turn, would match the previously mentioned possibility that folding transitions of R15 partially proceed orthogonal with respect to the acting force, see [Section 6.2.4.1](#). Similar to R15, the transition path probability of ROSS which is based on the deconvolved equilibrium distribution is about two orders of magnitude higher than the one which is derived directly, see [Figure 58F](#). However, reaching an impossible probability of 7.9 puts the whole transition path probability calculation under question. In case of ROSS one can argue that protein folding occurs on longer than expected time scales which obviously yields a too high  $p(\text{TP})$  value.

Two central results were elaborated in this section. First, transition path ensembles can directly reveal the presence of protein folding. Second, transition path probabilities do not localise correct transition state positions unless they are provided with this information upon their calculation. Further, a meaningful interpretation of transition path probabilities only seems possible when additional information is available. Other than that, new insights which go beyond the information provided by transition path ensembles are not offered. All these findings are backed by simulations performed in [Section 4.3.4](#).

## 6.3 MUTANTS

This section is dedicated to a further investigation of the folding mechanism of ROSS. At first, the proline-switch hypothesis, which is proposed due to PrL fingerprints in [Section 6.2.1](#), is tested by measuring the mechanics of proline-free mutants. Thereafter, different pulling geometries are realized to facilitate the structural interpretation of on-pathway intermediates which are indicated in [Figure 50B](#) of [Section 6.2.1](#) and postulated in [Section 6.2.2](#).

### 6.3.1 P86A and P86S: proline-free mutants reveal that proline serves as a stabilizing design element

By providing arguments for as well as against the proline-switch hypothesis, the following mechanical study of proline-free mutants of ROSS neither fully confirms nor disproves it. Alongside, the collected data disclose that proline plays a crucial role in the folding mechanism of ROSS by noticeably influencing the unfolding barrier and yielding the highest folding free energy.

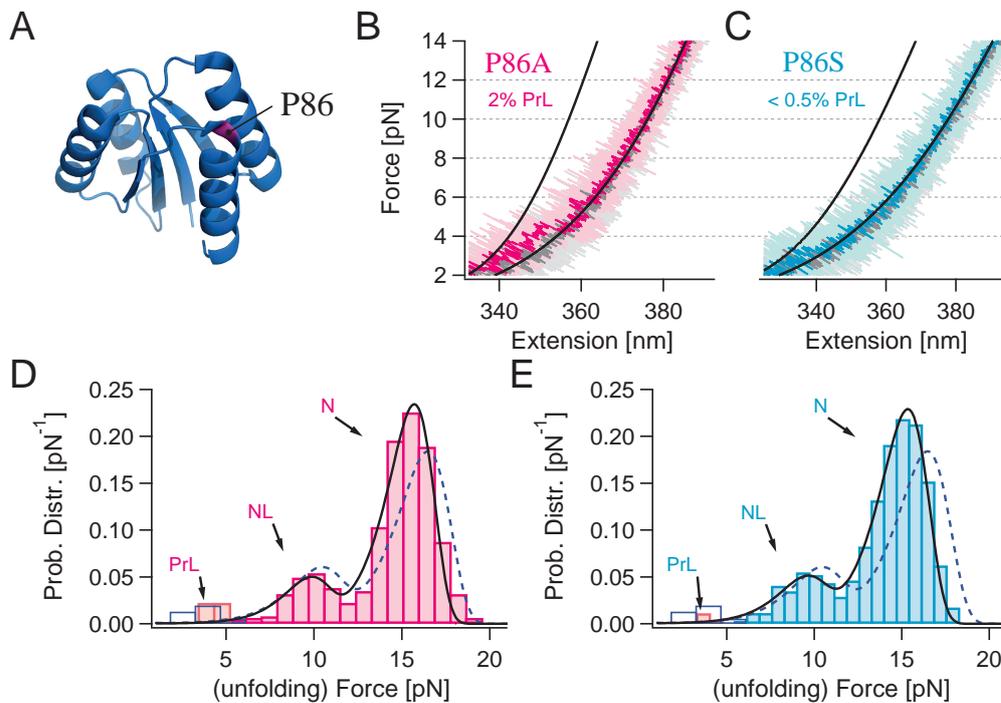


Figure 59: Proline-free mutants of ROSS by substitution: P86A and P86S. (A) Position of P86 within the solution NMR structure of ROSS. (B) and (C) PrL fingerprints of P86A and P86S recorded at 500 nm/s. (D) and (E) Unfolding force histograms of P86A and P86S based on all stretch-and-relax cycles performed at 500 nm/s. Dashed lines and unfilled histograms refer to N, NL, and PrL of ROSS with proline P86 being present, see [Figure 51B](#).

[Figure 59A](#) highlights the position of the only proline within the sequence of the designed Rossmann fold. Within the solution NMR structure of ROSS, P86 is located at the N-terminal end of the  $\alpha$ -helix which comprises the last

C-terminal secondary structural element. To realize proline-free mutants of ROSS, alanine (A) and serine (S) were chosen as two alternative substitutes for P86. The reasons for choosing these two replacements are, in brief, alanine’s high helix propensity and serine being favourable at N-capping positions, for more details see [Section A.1.2](#) [63, 165]. In [Figure 59B](#) and [Figure 59C](#), stretch-and-relax cycles of P86A and P86S are shown which look like typical PrL fingerprints of ROSS, compare also to [Figure 50D](#) in [Section 6.2.1](#). Though occurring less often than in case of ROSS, the mere observation of PrL fingerprints within proline-free mutants proclaims that a proline-switch cannot be their only explanation. Here, another detail becomes relevant: other than observed for ROSS, in case of P86A and P86S the number of successive PrL fingerprints is strongly reduced down to very few repetitions within 500 nm/s stretch-and-relax cycles. While for ROSS at least 4 and typically even more PrL fingerprints were observed in a row, P86A and P86S often only showed single isolated PrL fingerprints. The unfolding force distributions of P86A and P86S depicted in [Figure 59D](#) and [Figure 59E](#) feature the same multi-modal shape already seen for ROSS. This indicates that the point mutations within P86A or P86S do not strongly affect the overall folding mechanism of ROSS which involves multiple fingerprints and multi-modal unfolding. However, there are two differences. First, the relative fractions of N, NL and PrL are altered. And, second, the peak positions of N and NL are shifted towards lower forces. The latter can be directly seen when comparing the fits to the double-peaked distributions representing N and NL (full black lines) to the respective fit from ROSS (dashed dark blue lines) taken from [Figure 51B](#). This shift towards lower forces is an indication for a lower mechanical stability of P86A and P86S with respect to ROSS.

PROTEIN	$L_p^{CV}$ [nm]	N	NL	PrL [%]	Rc	$\Delta G_0^{CD,indiv}$ [k <sub>B</sub> T]
ROSS/P86	$34.7 \pm 0.9$	$71 \pm 3$	$23 \pm 2$	<b><math>5 \pm 2</math></b>	$1.3 \pm 0.9$	<b><math>27.7 \pm 2.8</math></b>
P86A	$34.4 \pm 0.9$	$81 \pm 3$	$16 \pm 3$	<b><math>2.5 \pm 1.2</math></b>	$0.7 \pm 0.7$	<b><math>25.9 \pm 2.6</math></b>
P86S	$34.5 \pm 0.9$	$84 \pm 2$	$15 \pm 2$	<b><math>0.4 \pm 0.4</math></b>	$1.0 \pm 0.8$	<b><math>26.1 \pm 2.6</math></b>

Table 12: Contour length gains, relative fingerprint occurrences during 500 nm/s constant velocity measurements, and folding free energies from constant distance measurements of ROSS, P86A, and P86S.

The numbers listed in [Table 12](#) support the above discussion of [Figure 59](#). With  $L_p^{CV,P86A} = 34.4 \pm 0.9$  nm and  $L_p^{CV,P86S} = 34.5 \pm 0.9$  nm, the average contour length gains<sup>11</sup> of P86A and P86S are identical to the gain observed for ROSS. This confirms that both proline-free mutants are fully folded. Further,

<sup>11</sup> Average contour gains of P86A | P86S are deduced from WLC-Fits to the folded and maximally unfolded states of N = 747 | 913 force-extension traces from 5 | 10 molecules. The persistence length  $p_p = 0.7$  nm was chosen as for ROSS. Average DNA-handle parameters are  $p_D = 32$  | 32 nm,  $L_D = 364$  | 362 nm, and  $K = 200$  | 210 pN.

Table 12 quantifies the above mentioned changes in relative occurrence of N, NL, and PrL fingerprints<sup>12</sup>. While the drop in relative occurrence of NL is nearly the same for both proline-free mutants, the drop of PrL occurrence is way more drastic for P86S than for P86A. More precisely, for P86A a drop from 5 to only 2.5% corresponds to a change by a factor of 2, whereas in case of P86S the occurrence of PrL fingerprints drops by about a factor of 10. However, for both proline-free mutants, the occurrence of PrL does not drop down to zero as would have been expected if only a proline-switch was its reason. Interestingly, the relative occurrence of Rc fingerprints as defined by Figure 50E remains about the same for ROSS, P86A, and P86S. This indicates that the refolding competence of partially unfolded protein against force remains conserved for both proline-free mutants. And, finally, we reach to the folding free energies  $\Delta G_0^{\text{CD,indiv}}$  of P86A and P86S which were extracted from constant distance measurements. As one can see, both proline-free mutants have a very similar folding free energy of about 26  $k_B T$  which is almost 2  $k_B T$  less than for ROSS. Together with the above mentioned higher mechanical stability, the higher folding free energy directly approves the intended overall stable design for ROSS.

PROTEIN	Type	$\varphi_{\text{NL}}$	$\log_{10} k_0^{\text{u,ZB}}$ [s <sup>-1</sup> ]	$\Delta x_{\text{u}}^{\text{ZB}}$ [nm]	$F_{\text{max}}^{\text{unf}}$ [pN]	$\lambda_{\text{mid}}$ [s <sup>-1</sup> ]	$F_{\text{mid}}$ [pN]
ROSS/P86	N	-	$-3.5 \pm 0.1$	$2.9 \pm 0.3$	16.5	$< 0.1$	9.5
P86A	N	-	$-3.8 \pm 1.1$	$3.2 \pm 0.2$	15.7	$\lesssim 0.4$	8.9
P86S	N	-	$-3.5 \pm 0.1$	$3.1 \pm 0.3$	15.4	$\lesssim 0.5$	9.2
ROSS/P86	NL	$0.25 \pm 0.02$	$-1.7 \pm 0.1$	$2.5 \pm 0.4$	10.3	-	-
P86A	NL	$0.17 \pm 0.03$	$-1.8 \pm 0.5$	$3.1 \pm 1.0$	9.7	-	-
P86S	NL	$0.15 \pm 0.02$	$-1.5 \pm 0.1$	$3.4 \pm 0.5$	9.4	-	-

Table 13: Comparing the kinetics of ROSS and the proline-free mutants P86A and P86S. Kinetic ZB-model fit parameters are based on unfolding force distribution histograms collected from 500 nm/s constant velocity measurements, see Figure 51B, Figure 59D, and Figure 59E. Combined rates  $\lambda_{\text{mid}}$  at mid-forces  $F_{\text{mid}}$  represent averages from multiple constant distance trajectories which yield similarly scattered data points for P86A and P86S as shown for ROSS in Figure 52B.

As already stated above, the occurrence of the same fingerprints as for ROSS suggests that P86A and P86S still share the same folding mechanism. The kinetic fit parameters listed in Table 13 confirm this view, see  $k_0^{\text{u,ZB}}$  and  $\Delta x_{\text{u}}^{\text{ZB}}$  values. Here, the tendency of slightly broader  $\Delta x_{\text{u}}^{\text{ZB}}$  values for the NL unfolding force distribution simply results from the smaller relative fraction  $\varphi_{\text{NL}}$  which causes an effectively broader distribution. Note that  $\varphi_{\text{NL}}$  values

<sup>12</sup> Errors in relative occurrence were derived by bootstrapping and indicate confidence intervals of 95%.

identify a reduction of NL fingerprints by about 10% with respect to the entire double-peaked distributions. Relatively, this means a 40% reduction of NL occurrence. The specified  $F_{\max}^{\text{unf}}$  values indicate an overall shift of the unfolding force distributions of P86A and P86S by up to 1 pN towards lower forces while the relative distance of about 6 pN between the two peaks of N and NL stays about constant, see [Figure 59D](#) and [Figure 59E](#). Mid-forces  $F_{\text{mid}}$  only slightly drop by about 0.5 pN which is in agreement with the relatively small loss in folding free energy. However, combined rates  $\lambda_{\text{mid}}$  increase by about one order of magnitude despite slightly smaller mid-forces. This is strong evidence for an effectively reduced unfolding barrier which confirms the lower mechanical resistance of P86A and P86S already suspected above.

What can we learn from the above results about the folding mechanism of ROSS? First of all, proline P86 plays a central role in clamping together the designed Rossmann fold by some sort of 'key lock' mechanism which forces the C-terminal helix to close. This follows from the position of P86 in the solution NMR structure, see [Figure 59A](#), in connection with the higher unfolding barrier<sup>13</sup> and higher folding free energy of ROSS as opposed to the proline-free mutants. However, the higher stability obviously comes at the cost of undesired side reactions represented by NL and PrL fingerprints. Next, we can discuss possible reasons for the observed reduction of undesired side reactions for the proline-free mutants. The 'population' of NL fingerprints is equally reduced by about one half for both P86A and P86S. Since both mutants are missing the proline, the equal reduction may point towards the lost forced key lock mechanism of P86 which, in turn, seems to block rearrangements necessary to avoid NL fingerprints. One way to test this hypothesis may be by comparing the refolding forces which, in case the key lock snaps in too fast, would be lower for P86A and P86S. Now what about the much higher reduction of PrL fingerprints in case of P86S as opposed to P86A? When trying to judge about this effect, the main problem is that there is no clue about the mechanism which causes successive PrL fingerprints in the absence of a proline-switch other than 'dead' states, see also [212]. Because of that, at this point, only a vague guess can be made: serine might be a slightly better promoter of  $\alpha$ -helix formation at a N-capping position and, at the same time, provide higher backbone flexibility of the unfolded polypeptide chain. In particular the backbone flexibility may play a crucial role in finding the correct native fold.

To close this section, we finish with a summary of facts and insights concerning the folding mechanism of ROSS which were gained by comparison to the proline-free mutants P86A and P86S. Contrary to expectations, PrL fingerprints, which were previously thought to be caused by a proline-switch, still occur even without any proline being present. However, in view of an overall reduced occurrence and a strongly reduced accumulation of successive PrL fingerprints, the proline-switch hypothesis cannot be completely disproved.

<sup>13</sup> Basically, P86 increases the energy of the transition state for unfolding which is most likely represented by a loosened C-terminal  $\alpha$ -helix. Note that this is in accord with the first on-pathway unfolding intermediate identified in [Section 6.4.3](#) and characterized in [Section 6.3.2](#).

Importantly, for both proline-free mutants a reduction of unfolding barriers, folding free energies (by about  $2 k_B T$ ), and the occurrence (by about one half and more) of undesirable side reactions represented by NL and PrL fingerprints is observed. Together with the position of P86 within the solution NMR structure of ROSS, all these results support the hypothesis of a key lock mechanism of the C-terminal  $\alpha$ -helix which is enforced by proline P86 to clamp ROSS together. This hypothesis is further manifested in [Section 6.3.2](#).

### 6.3.2 *S49Cc: directed unfolding of the C-terminal half of ROSS uncovers an energetic imbalance*

Thanks to the S49Cc construct, a detailed mechanical characterization of an obligatory high-energy on-pathway intermediate of ROSS based on constant distance measurements becomes possible. Corresponding results provide evidence for an energetic imbalance within ROSS. Moreover, indications for a structural interpretation of NL fingerprints are found and further elaborated.

By changing DNA-handle attachment positions, it is possible to target mechanical forces at selected substructures of a protein. In this section, the Rossmann fold is subjected to force after using the mutation S49C as attachment position instead of the N-terminal cysteine, for details see [Section A.1.2](#). Since the resulting S49Cc construct automatically restricts unfolding to the C-terminal half of the Rossmann fold, here, this is termed 'directed unfolding'.

The typical force-extension trace of S49Cc shown in [Figure 60A](#) features the very same pre-flipping already seen for the predominating native fingerprint of ROSS in [Figure 50B](#). Further, the unfolding force distribution of S49Cc presented in [Figure 60B](#) reveals the same multi-modal unfolding which originates from N, NL, and PrL fingerprints. In addition, the peak positions of the N and NL 'populations' of S49Cc are exactly the same as for ROSS, see also [Table 15](#). Taken together, these results are very strong evidence for S49Cc to have the same unfolding pathway as ROSS. This is a very important finding as it implies that many insights gained upon studying S49Cc are valid for ROSS, too. One of these insights is that unfolding of ROSS starts from the C-terminus and passes through an obligatory on-pathway intermediate. For S49Cc, the contour length gain histogram of this intermediate and its only possible structural interpretation are depicted in [Figure 60C](#). As highlighted in red within the inset NMR structure of ROSS, the intermediate of S49Cc comprises unfolding of the C-terminal  $\alpha$ -helix and  $\beta$ -sheet. Resulting average contour length gains<sup>14</sup> of the intermediate as well as of full unfolding of S49Cc are listed in [Table 14](#). Since  $L_p^{CV,S49Cc} = 16.5 \pm 0.6$  nm very well matches the expectation for full unfolding of S49Cc, here, a fully folded protein is on hand.

Despite the similarities between S49Cc and ROSS, there are also interesting differences when it comes to their relative fingerprint occurrences listed in

<sup>14</sup> For S49Cc, average contour gains are deduced from WLC-Fits to  $N = 338$  force-extension traces from eight molecules. In accord with ROSS, the protein's persistence length was set to  $p_p = 0.7$  nm. Average DNA-handle parameters are  $p_D = 33$  nm,  $L_D = 359$  nm, and  $K = 220$  pN.

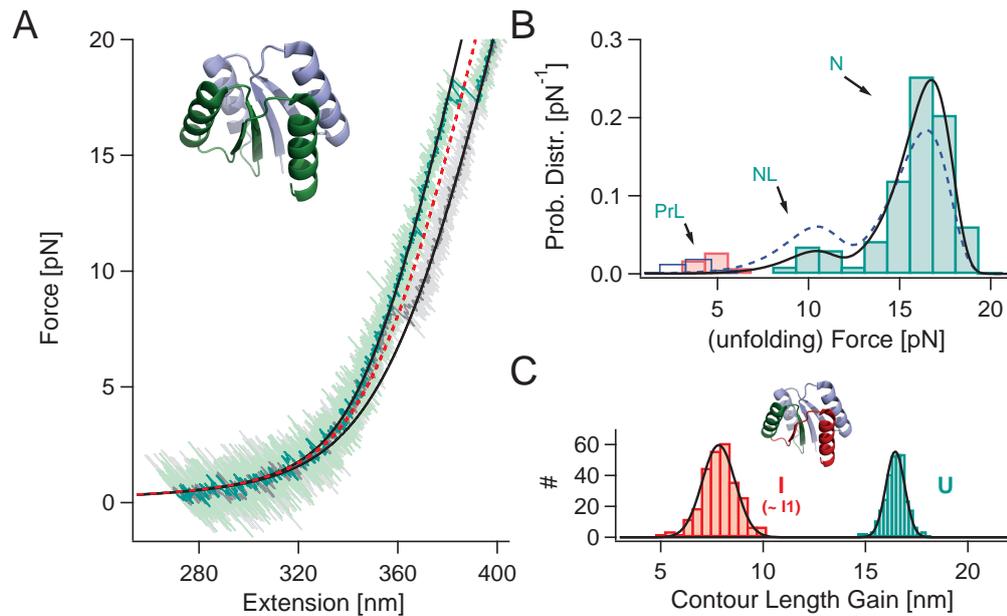


Figure 60: Directed unfolding of the C-terminal half of ROSS: S49Cc. (A) Typical force-extension trace of S49Cc recorded at 500 nm/s featuring an obligatory on-pathway intermediate upon unfolding during stretching (green) followed by refolding during relaxation (grey). Black lines are WLC-fits to the folded and fully unfolded state. The red dashed line is a fit to the intermediate. *Inset*: Solution NMR structure of ROSS: the C-terminal half which is subjected to force when pulling on the S49Cc construct is highlighted in green. (B) The unfolding force histogram of S49Cc, which is based on 500 nm/s cycles, has three distinct peaks (green and pink). The black line fits the green double-peaked histogram according to Equation 67. The dashed line and the unfilled histogram refer to N, NL, and PrL of ROSS. (C) Contour length gain histograms of S49Cc based on WLC-fits to the intermediate (red) and unfolded (green) state of S49Cc. *Inset*: The C-terminal  $\alpha$ -helix and  $\beta$ -sheet, which unfold upon formation of the intermediate, are highlighted in red.

Table 14. While PrL fingerprints occur about as frequently and cumulated for both S49Cc and ROSS, in case of S49Cc, the relative amount of observed NL fingerprints drops drastically and Rc fingerprints seem to have completely disappeared. Since P86 is located within the C-terminal half of ROSS, it is also involved in the unfolding and refolding process of S49Cc. Therefore, the similar PrL occurrence makes sense, see also Section 6.3.1. The drastic drop of NL fingerprints, by contrast, potentially hints towards their actual origin: a non-native interaction between the C- and N-terminal halves of ROSS which is, at least partially, inhibited by the pulling geometry of S49Cc. With the still upcoming folding free energies of S49Cc being in support of this idea, the potential origin of NL fingerprints is further discussed in Section 6.3.2.1. The disappearance of Rc fingerprints may simply be owed to difficulties in resolving them within S49Cc measurements. Alternatively, statistics may not be high enough to detect a strong reduction of Rc fingerprints. In the latter

case, Rc fingerprints would directly relate to NL fingerprints. This, in turn, would explain Rc fingerprints as ‘rescue’ transitions out of a NL into a N configuration via partial unfolding followed by rearrangement and refolding.

PROTEIN	Transition	$L_p^{\text{calc}}$	$L_p^{\text{CV}}$	N	NL	PrL	Rc
		[nm]				[%]	
ROSS	$F \rightleftharpoons U$	34.51	$34.7 \pm 0.9$	$71 \pm 3$	$23 \pm 2$	$5 \pm 2$	$1.3 \pm 0.9$
S49Cc	$F \rightleftharpoons U$	16.74	$16.5 \pm 0.6$	$88 \pm 6$	$7 \pm 5$	$5 \pm 2$	0
S49Cc	$F \rightleftharpoons I$	8.68	$8.0 \pm 0.9$	-	-	-	-

Table 14: Contour length gains and relative fingerprint occurrences of S49Cc compared to ROSS.

Especially for N fingerprints, the kinetic ZB-model fit parameters listed in Table 15 manifest the similarities between S49Cc and ROSS unfolding. For NL fingerprints, the zero-force extrapolated rate constant  $k_0^{u,ZB}$  and the transition state position  $\Delta x_u^{ZB}$  of S49Cc have to be handled with care as the corresponding part of the distribution provides too few data points for good convergence of the fit, see Figure 60B. However,  $F_{\text{max}}^{\text{unf}}$  values are in perfect agreement with those of ROSS for both N and NL fingerprints. The remarkable difference between the combined rates and mid-forces of S49C and ROSS already foretells two advantages of the S49Cc construct with respect to mechanical studies: the about two orders of magnitude faster rates at significantly higher forces provide much better statistics within shorter time at increased spatial resolution. These advantages are exploited next for a thorough mechanical characterization of S49Cc based on constant distance measurements.

PROTEIN	Type	$\varphi_{\text{NL}}$	$\log_{10} k_0^{u,ZB}$	$\Delta x_u^{ZB}$	$F_{\text{max}}^{\text{unf}}$	$\lambda_{\text{mid}}$	$F_{\text{mid}}$
			[s <sup>-1</sup> ]	[nm]	[pN]	[s <sup>-1</sup> ]	[pN]
ROSS	N	-	$-3.5 \pm 0.1$	$2.9 \pm 0.3$	16.5	< 0.1	9.5
S49Cc	N	-	$-4.1 \pm 0.3$	$3.2 \pm 0.3$	16.8	5.6	12.9
ROSS	NL	$0.25 \pm 0.02$	$-1.7 \pm 0.1$	$2.5 \pm 0.4$	10.3	-	-
S49Cc	NL	$0.08 \pm 0.05$	$(-1.8 \pm 1.4)$	$(4.0 \pm 3.9)$	10.2	-	-

Table 15: Comparing the kinetics of S49Cc and ROSS. Kinetic ZB-model fit parameters are based on unfolding force distribution histograms collected from 500 nm/s constant velocity measurements, see Figure 51B and Figure 60B. For S49Cc, combined rates  $\lambda_{\text{mid}}$  at mid-forces  $F_{\text{mid}}$  refer to the intersection of SB-model fits to  $F \rightarrow I$  and  $U \rightarrow I$  transitions in Figure 61B.

To assign states within constant distance trajectories of S49Cc like the one depicted in Figure 61A, the HMM algorithm introduced in Section 3.4.3 is configured to assume a three-state model which only allows transitions between

next neighbours. The resulting lifetime distributions are single exponential and thereby support the finding of an obligatory on-pathway intermediate for unfolding as well as refolding of S49Cc. By scanning through the force range which is accessible to equilibrium fluctuations, force-dependent rates of all three states of S49Cc can be extracted and fitted as shown in Figure 61B. Corresponding state occupancy probabilities are presented in Figure 61C. The kinetic results obtained from analysing all constant distance trajectories of S49Cc are summarized in Table 16.

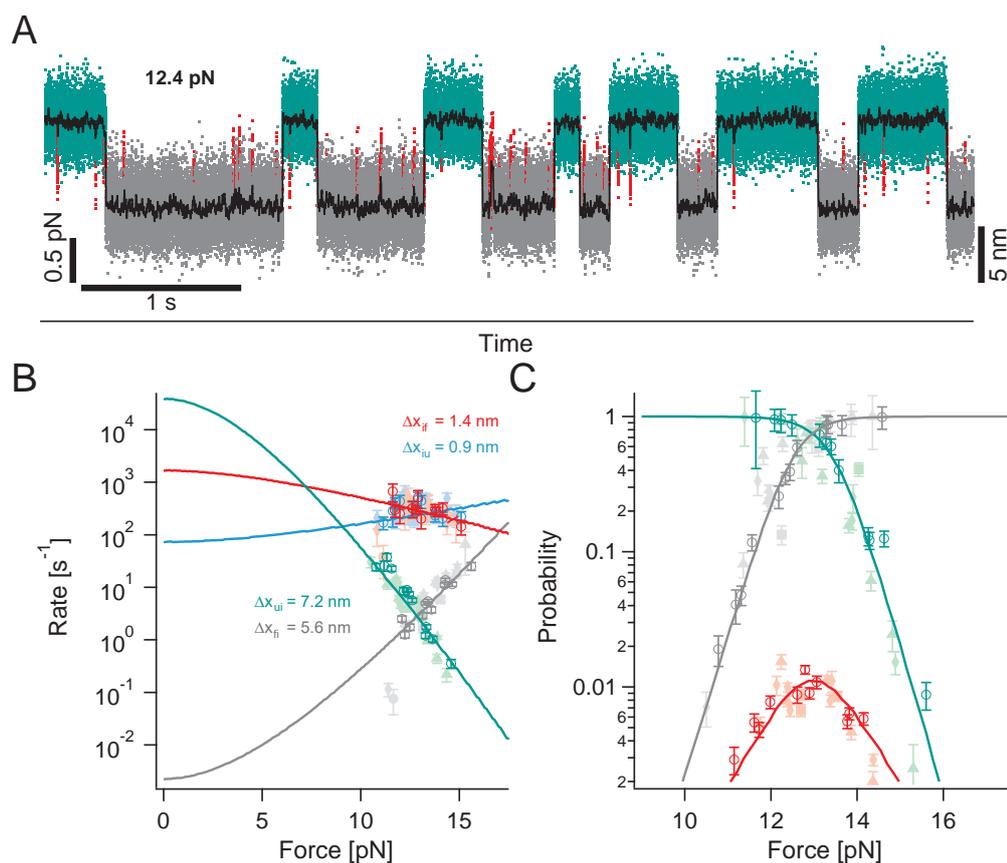


Figure 61: Detailed characterization of the on-pathway intermediate of S49Cc by constant distance measurements. (A) Constant distance trajectory of S49Cc. Folded (green dots), unfolded (grey) and intermediate (red) states are classified by HMM. The indicated force of 12.4 pN acts on the intermediate state level. The black trace is a moving average filtered version of the trajectory. (B) Force-dependent rates (empty circles) of one single molecule. Lines are SB-Model fits. (C) State occupancy probabilities (empty circles) and fits (lines). For comparison, (B) and (C) are complemented with additional data points from other molecules (different shapes in light colours).

As confirmed by the transition state positions listed in Table 16, the rate plot of S49Cc is relatively symmetric with respect to the position of the intermediate state. Together, all transition state distances add up to 15.1 nm which is almost as much as the contour length gain measured upon full unfolding of S49Cc, see Table 14. The confirmation of this expectable but not always given

relation will prove beneficial for the energy landscape reconstruction of ROSS in Section 6.4.4. And, with respect to the rate plot of S49Cc, another aspect is also worth mentioning: while having very shallow slopes and, hence, being relatively insensitive against force, rates for leaving the on-pathway intermediate stem from dwell times typically being of the order of milliseconds. By this means, the dwell-time of the on-pathway intermediate which is characterized in this section can be expected to substantially contribute to the overall transition path time of ROSS which is 3.9 ms on average at forces around 9.5 pN, see also Section 6.2.3.

Transition	$\log_{10}k_0^f$ [s <sup>-1</sup> ]	$\log_{10}k_0^u$ [s <sup>-1</sup> ]	$\Delta x_f$ [nm]	$\Delta x_u$ [nm]	$\Delta G_0^{\text{CD,indiv glob}}$ [k <sub>B</sub> T]
F ⇌ I	3.2 ± 0.3	-2.7 ± 0.1	1.4 ± 0.4	5.6 ± 0.2	14.9   15.3 ± 1.5
I ⇌ U	4.6 ± 0.1	1.9 ± 0.1	7.2 ± 0.2	0.9 ± 0.3	6.0   6.1 ± 0.6
F ⇌ U	-	-	-	-	20.9   21.4 ± 2.1

Table 16: SB-model fit parameters and folding free energies of S49Cc and its obligatory on-pathway intermediate.

An important key to understand the folding mechanics of ROSS is provided by the folding free energies derived from the constant distance measurements performed with the S49Cc construct. With respect to the folded state, the intermediate has a folding free energy of about 15 k<sub>B</sub>T while full unfolding of S49Cc releases about 21 k<sub>B</sub>T, see Table 16. Together with the fact that full unfolding of the entire Rossmann fold releases up to 28 k<sub>B</sub>T, see Table 10 in Section 6.2.2, these numbers allow two conclusions. First, the difference in released folding free energy upon full unfolding of S49Cc and ROSS indicates that a residual energy of about 7 k<sub>B</sub>T is still stored within the N-terminal half of the otherwise unfolded S49Cc construct. And second, upon unfolding into the intermediate more than half of the entire folding free energy of the designed Rossmann fold is set free while three quarters of the entire protein still remain folded.

Since the N-terminal half of S49Cc needs to hold up a certain structure to store the residual energy after unfolding of the C-terminal half, non-native interactions between these two halves are likely to be suppressed. As non-native interactions are suspected to cause NL fingerprints, their indirectly proven partial inhibition can explain the drastic decrease of NL fingerprints for S49Cc. Furthermore, when a significant amount of energy is still being stored within a substructure of an otherwise unfolded protein, also another aspect comes to mind: cooperativity. Here, the mere fact of a residual energy suggests that the design of ROSS is geared towards sequential folding.

The release of more than half of the folding free energy of ROSS upon unfolding of only one quarter of its structure points towards an energetic imbalance. However, the imbalance between the N- and C-terminal part of ROSS is not as drastic as it seems at first sight. One reason is that upon

unfolding of the C-terminal  $\alpha$ -helix and  $\beta$ -sheet of S49Cc, a huge amount of energy is lost due to the exposure of a large interface area of the protein to the surrounding solvent. To get an idea by how much the folding free energy stored within the C-terminal half of ROSS does exceed the energy stored within the N-terminal half, we do a gedankenexperiment. Let's assume an equally distributed folding free energy throughout the Rossmann fold. By this means, the N- and C-terminal halves would each store  $7 k_B T$  by themselves as this is the remaining energy after the C-terminal half is unfolded, see above. Accordingly, another  $14 k_B T$  are stored within the interface when the N- and C-terminal halves are brought together to form the Rossmann fold. Now, if the C-terminal  $\alpha$ -helix and  $\beta$ -sheet are being unfolded, a maximal amount of half the interface energy, i. e.,  $7 k_B T$ , is being released due to symmetry and the assumption of equally distributed energies. Further, upon breaking up 4 out of 5 interfaces within the N-terminal half of ROSS, i. e., one interface between the two  $\beta$ -sheets and three out of four interfaces between  $\alpha$ -helices and  $\beta$ -sheets, another  $4/5 \cdot 7 k_B T$  are maximally released. Together, this results in  $12.6 k_B T$  for the unfolding of the C-terminal  $\alpha$ -helix and  $\beta$ -sheet which is  $2.3 k_B T$  less than observed. This gedankenexperiment oversimplifies the actual situation, but it provides a lower estimate of the energetic imbalance between the N- and C-terminal halves of ROSS being on the order of 10% more energy within the C-terminal half.

Based on an energetic imbalance of at least 10% between the two halves of the Rossmann fold, another idea suggests itself. The imbalance might explain the relatively high amount of undesirable side reactions of ROSS into mechanically weaker or even completely labile folds represented by NL and PrL fingerprints. Figuratively speaking, the C-terminal half of ROSS is so eager to fold onto any interface provided by the N-terminal half that it does not wait until the correct interface has been formed. This idea is supported by the fact that the amount of NL fingerprints is significantly reduced for proline-free mutants of ROSS which have about  $2 k_B T$  less folding free energy stored within their C-terminal half, see [Table 12](#) in [Section 6.3.1](#). Note that these  $2 k_B T$  less are also in the same ballpark as the lower estimate for the energetic imbalance from above.

So far, directed unfolding based on the S49Cc construct confirmed that unfolding of the Rossmann fold starts from the C-terminus and passes through an obligatory on-pathway intermediate with millisecond lifetimes. In addition, a residual energy of  $7 k_B T$  within the N-terminal half of the otherwise unfolded S49Cc points towards a sequential folding mechanism of ROSS. The uncovered energetic imbalance between the two halves of the Rossmann fold may actually be an important reason for the relatively high amount of undesirable non-native side reactions represented by NL, PrL, and Rc fingerprints. Further, the results obtained in this section indicate that NL fingerprints occur upon non-native interactions between the N- and C-terminal halves of ROSS. Moreover, Rc fingerprints are likely to be a spin-off of NL fingerprints by representing rescue attempts out of NL and into the native configuration. And, for completeness, no further insight into an alternative origin has been

found for PrL fingerprints which, hence, remain vaguely related to problems of the protein backbone.

### 6.3.2.1 Swapped $\beta$ -sheets might explain NL fingerprints

This subsection addresses the question about the origin of the most prominent non-native NL fingerprint of the Rossmann fold. After finding evidence for a non-native structural arrangement, potential explanations including swapped  $\beta$ -sheets and flipped  $\alpha$ -helices are discussed.

In [Section 6.3.2](#), it has been shown that after full unfolding of the C-terminal half, the N-terminal half of S49Cc basically remains folded. This automatically reduces the number of potential non-native interactions between the N- and C-terminal halves of ROSS (N/C-interactions). At the same time, the occurrence of NL fingerprints is drastically decreased for S49Cc. Though being non-sufficient as proof, the concurrence of these two results suggests that non-native N/C-interactions might lead to NL fingerprints.

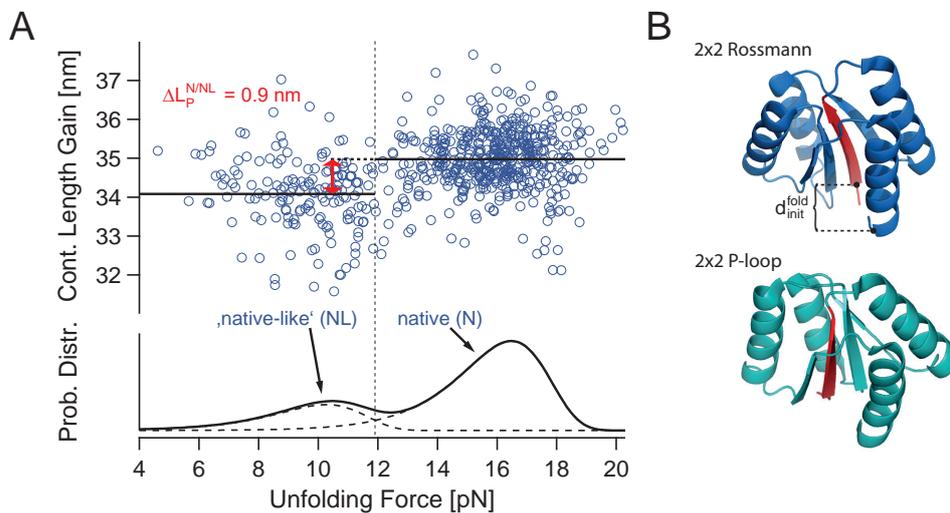


Figure 62: Why  $\beta$ -sheet swapping might cause 'native-like' fingerprints. (A) Upper panel: scatter plot of full contour length gains versus unfolding forces of ROSS (500 nm/s cycles). Black horizontal lines mark average  $L_p$  values for forces below and above 11.9 pN. Lower panel: the corresponding fit to the unfolding force distribution of ROSS is composed of the native and 'native-like' populations (dashed curves) which intersect at 11.9 pN. (B) Solution NMR structures of the 2x2 Rossmann fold (pdbID: 2LV8) and P-loop (pdbID: 2LVB) illustrate that the N-terminal  $\beta$ -sheet (coloured in red) can take different positions within a four  $\beta\alpha$ -repeat protein. A non-native N-terminal  $\beta$ -sheet position automatically changes  $d_{init}^{fold}$  and, by this means, measured contour length gains  $L_p$ , see [Equation 5](#).

As illustrated in the lower panel of [Figure 62A](#), so far, the only way to discriminate between N and NL fingerprints of ROSS is by force. If that was the only difference between N and NL, a force-induced pathway switch involving a lower unfolding barrier for NL would be the only explanation on hand for this kind of mechanical behaviour while the actual origin of this

switch remained unclear. However, motivated by indications of non-native N/C-interactions to cause NL fingerprints, now, the most prominent and directly measured structural information is examined in more detail: the total contour length gain upon unfolding of ROSS. Therefore, a scatter plot of contour length gains versus unfolding forces is presented in the upper panel of [Figure 62A](#). And in fact, when separated by force, one can already see by eye that the average contour length gain of N and NL fingerprints differs by  $\Delta L_p^{N/NL} = 0.9 \pm 0.6$  nm which is direct evidence for a structural difference prior unfolding. In particular, as NL fingerprints show shorter contour length gains upon unfolding, the initial distance  $d_{\text{init}}^{\text{fold}}$  between the N- and C-termini of the NL-related structure of ROSS has to be larger, see also [Equation 5](#). Note that when being in the unfolded state, force-extension traces of N and NL fingerprints perfectly overlap.

What kind of wrong arrangement could possibly occur upon attempting to fold into the Rossmann fold structure? An interesting option are swapped  $\beta$ -sheets right in the core of the Rossmann fold which would result in a P-loop as illustrated in [Figure 62B](#). The folding motifs of these two artificial proteins cannot be discriminated by the rules according to which they were designed and the preference for one of these motifs was only influenced by adapting secondary structure lengths [115]. Further, unintentional  $\beta$ -sheet swapping occurred for designed 2x3 Rossmann folds which folded into their P-loop variant when being expressed<sup>15</sup>. In view of the fact that the designed P-loop has a 3 times lower folding free energy than Rossmann, also a reduced unfolding barrier would be plausible and, by this means, the difference in unfolding forces of N and NL fingerprints could be explained. The remaining all-dominant question is whether a  $\beta$ -sheet swap in the core of ROSS can result in a difference in  $d_{\text{init}}^{\text{fold}}$  of the order of 0.9 nm? Distances between the N-terminal ends of the four  $\beta$ -sheets and the C-terminal end of the C-terminal  $\alpha$ -helix within the solution NMR structure of the Rossmann fold range from 1.26 nm to 1.58 nm. Importantly, the shortest distance is the one between the N- and C-termini of the native Rossmann fold. The distance from the C-terminus towards the  $\beta$ -sheet residing at the P-loop position is only slightly larger with 1.33 nm. The longest distance is measured with respect to the surface exposed  $\beta$ -sheet within the N-terminal half of ROSS. So, if the remaining structure of ROSS would still decently fold upon non-native P-loop formation, another 0.8 nm would be missing to reach the difference in length observed for NL fingerprints. By contrast, in case of a  $\beta$ -sheet swap within the N-terminal half of ROSS, only 0.6 nm were missing. Considering that a hydrophobic  $\beta$ -sheet is removed from the core and instead being solvent exposed, here, an additional 0.6 nm due to a less densely packed fold seem possible. So, in principle,  $\beta$ -sheet swapping is an option for non-native interactions which result in NL fingerprints.

Other important aspects which may influence the occurrence of non-native N/C-interactions or, differently speaking, misfolds, are the energetic imbalance as well as the sequential folding mechanism of ROSS, see [Section 6.3.2](#).

<sup>15</sup> By way of personal communication from David Baker, University of Washington, 2014.

Further, the restricted and uneven spatial separation of secondary structure elements due to the pulling geometry needs to be taken into account, see also [Figure 45B](#) in [Section 5.3.3](#).

In the future, NL fingerprints can be studied in more detail by a number of further investigative steps. One step would be to separate refolding events which precede either N or NL unfolding to look for any potential differences concerning their distribution, duration, etc. Based on unfolding and refolding cycles, folding free energies could be determined for NL fingerprints using, e. g., Crooks theorem. Further, Molecular dynamics (MD) simulations could be helpful to check different 'swapping' geometries for sanity. An experimental approach towards deciphering the true nature behind NL fingerprints may be given by photo-induced crosslinking of the N-terminal  $\beta$ -sheet to one of its neighbouring secondary structure elements within the N-terminal half of the Rossmann fold. This way, potential non-native interactions could be selectively inhibited. For more suggestions see also [Section 6.5](#).

In summary, a direct comparison between the measured contour length gains of N and NL fingerprints of ROSS revealed that NL fingerprints provide by  $\Delta L_p^{N/NL} = 0.9 \pm 0.6$  nm shorter lengths. This is direct proof for a differently folded structure with a larger initial distance  $d_{init}^{fold}$  between N- and C-termini. Inspired by the structurally very similar P-loop fold, the idea of swapped  $\beta$ -sheets and/or eventually even flipped  $\alpha$ -helices with respect to the original Rossmann fold may well explain the occurrence of NL fingerprints due to non-native interactions between the N- and C-terminal halves of ROSS.

### 6.3.3 *S49Cn and S85Cn: tracing another intermediate*

Directed N-terminal unfolding of ROSS reveals a much higher mechanical barrier as opposed to C-terminal unfolding. In addition, unfolding of S85Cn discloses another on-pathway intermediate of ROSS with sub-millisecond average lifetimes.

As for the S49Cc construct, the S49Cn construct is based on using the S49C point mutation within a surface exposed loop region of the Rossmann fold as an alternative DNA-handle attachment position. In case of S49Cn, this attachment position is used instead of the C-terminal cysteine of the ROSS construct, for details see [Section A.1.2](#). This way, unfolding of the Rossmann fold is restricted to its N-terminal half coloured in yellow in the inset of [Figure 63A](#). The same principle is valid for the S85Cn construct which uses the S85C point mutation instead of S49C.

Already on first sight, typical force-extension traces of S49Cn and S85Cn in [Figure 63](#) disclose relatively high unfolding forces. For both constructs, unfolding in 500 nm/s stretch-and-relax cycles occurs at about 37 pN on average. This is far above all unfolding forces observed for ROSS and speaks for a very high mechanical unfolding barrier in case of blocked C-terminal 'un-locking' due to the new pulling geometry inferred by construct design. Especially in case of S49Cn, unfolding directly involves pulling the N-terminal  $\beta$ -sheet out of the hydrophobic core of the Rossmann fold. In view of this scenario, a

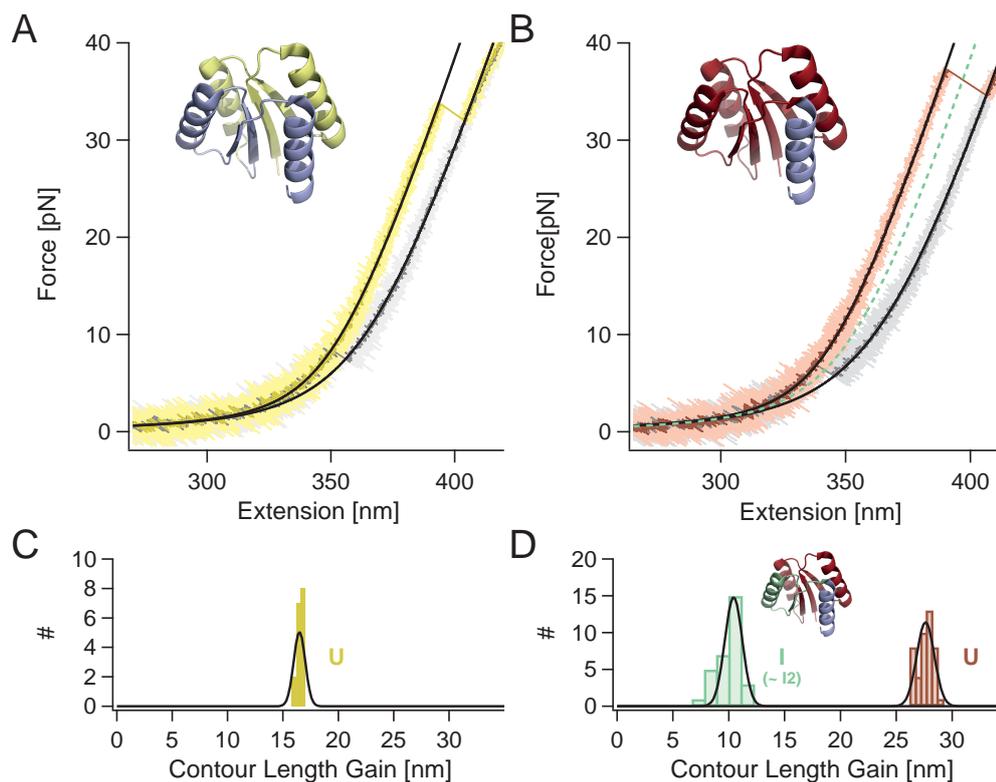


Figure 63: Directed N-terminal unfolding of ROSS: S49Cn and S85Cn. (A) Typical force-extension trace of S49Cn recorded at 500 nm/s. Black lines are WLC-fits to the folded and fully unfolded state. *Inset*: NMR structure of ROSS: the N-terminal half which is subjected to force when pulling on the S49Cn construct is highlighted in yellow. (B) S85Cn features an on-pathway intermediate upon unfolding indicated by the pale green dashed WLC-fit. *Inset*: The part which is subjected to force when pulling on the S85Cn construct is highlighted in dark red. (C) Contour length gain histogram of S49Cn. (D) Contour length gain histograms of S85Cn based on WLC-fits to the intermediate (pale green) and unfolded (dark red) state of S85Cn. *Inset*: The C-terminal structural elements, which are suspected to be unfolded upon formation of the intermediate of S85Cn, are highlighted in pale green.

higher unfolding barrier with respect to (sequential) C-terminal unfolding is expected. Further, S49Cn and S85Cn do not unfold during every constant velocity cycle, see 'p(U)'-labeled column in Table 17, and sometimes only during relaxation after stretching. In view of the limited force range of the experimental setup, this indicates that corresponding force distributions are not fully sampled with respect to high forces. Also, the above average unfolding force of 37 pN does only represent a lower estimate. Compared to S49Cn, S85Cn unfolds over 10 times less often. This is most likely owed to the difference in pulling geometry which involves shearing of a larger internal surface area in combination with a potentially less effective force lever in case of S85Cn. Furthermore, unfolding of S85Cn might start along a different path as

compared to S49Cn. Note that the average contour length gains<sup>16</sup> of  $L_p^{CV,S49Cn} = 16.5 \pm 0.3$  nm and  $L_p^{CV,S85Cn} = 27.8 \pm 0.7$  nm listed in Table 17 for full unfolding of S49Cn and S85Cn very well match expectations indicated by  $L_p^{calc}$  values. This confirms correctly folded as well as fully (re-)folding competent constructs.

PROTEIN	Transition	$L_p^{calc}$ [nm]	$L_p^{CV}$ [nm]	p (U) [%]	p (I U) [%]	min( $F_{unf}$ ) [pN]
S49Cn	F $\rightleftharpoons$ U	16.65	16.5 $\pm$ 0.3	44 $\pm$ 10	0	13
S85Cn	F $\rightleftharpoons$ U	28.06	27.8 $\pm$ 0.7	3 $\pm$ 1	72 $\pm$ 13	7
S85Cn	F $\rightleftharpoons$ I	9.68	10.4 $\pm$ 1.2	-	-	-

Table 17: Contour length gains, the probability p (U) to observe unfolding and the conditional probability p (I|U) of seeing intermediates upon unfolding, and minimal unfolding forces of S49Cn and S85Cn. Here, unfolding means within a single 500 nm/s stretch-and-relax cycle covering forces from zero up to at least 30 pN or, typically, more. This also includes unfolding only during relaxation after stretching. By contrast, minimal unfolding forces min( $F_{unf}$ ) only refer to unfolding during stretching.

Even though the force distributions of S49Cn and S85Cn were not fully sampled during 500 nm/s constant velocity cycles, it is still possible to make a few remarks on additional NL, PrL and Rc fingerprint occurrences which characterize the multi-modal unfolding of ROSS. The easiest candidate are PrL fingerprints as unfolding out of their mechanically unstable configuration is easy to discern from other rupture events: PrL fingerprints were neither observed for S49Cn nor for S85Cn. This can be expected, because the entire C-terminal  $\alpha$ -helix which includes proline P86 at the critical N-capping position is never directly involved in unfolding and refolding. With 13 and 7 pN, minimal unfolding forces min( $F_{unf}$ ) which were observed during stretching of S49Cn and S85Cn, see Table 17, are relatively small with respect to the lower estimate of an average unfolding force of 37 pN. Especially the 7 pN for S85Cn could originate from an additional unfolding force distribution similar to the additional population observed for NL fingerprints in case of ROSS. Moreover, if such an additional population was missing for S49Cn, this would be another indication for non-native interactions between the N- and C-terminal halves of ROSS (N/C-interactions) to cause NL fingerprints: with respect to S85Cn, N/C-interactions are automatically more suppressed for S49Cn where only the N-terminal half is forced to unfold. However, current statistics do not clearly speak for or against an additional population within the unfolding forces of either S49Cn or S85Cn. Alternatively, another straightforward expla-

<sup>16</sup> Average contour gains of S49Cn | S85Cn are deduced from WLC-Fits to N = 108 | 820 force-extension traces from 8 | 12 molecules. The persistence length  $p_P = 0.7$  nm was chosen as for ROSS. Average DNA-handle parameters are  $p_D = 28$  | 20 nm,  $L_D = 364$  | 363 nm, and  $K = 320$  | 300 pN.

nation for the relatively small  $\min(F_{\text{unf}})$  values could be given by a very small transition state distance  $\Delta x_{\text{u}}^{\text{ZB}}$  which leads to a broad distribution of unfolding forces. Finally, no Rc fingerprints were observed during directed N-terminal unfolding experiments. As previous results suggest that Rc fingerprints represent rescue transitions out of the NL configuration, a lack of Rc fingerprints is expected upon the very few and uncertain unfolding events which might be attributed to NL unfolding. Along with the just made considerations one also has to be aware of the fact that S49Cn and S85Cn are forced to have a different initiation of unfolding than ROSS. By this means, both constructs may choose completely different pathways for unfolding as opposed to ROSS, P86A, P86S, and S49Cc.

The contour length gain histogram in [Figure 63D](#) reveals that unfolding of S85Cn typically involves an on-pathway intermediate. As listed in [Table 17](#), the corresponding average contour length gain is  $L_{\text{p}}^{\text{CV,S85Cn-I}} = 10.4 \pm 1.2$  nm. Among all possible configurations, unfolding of the C-terminal secondary structural elements  $\beta 4\alpha 3$ , see inset of [Figure 63D](#) for their position within the solution NMR structure of ROSS, yields the most likely structural interpretation of the S85Cn intermediate with an expected contour length gain  $L_{\text{p}}^{\text{calc}}$  of 9.68 nm which is closest to the measured value. In contrast to S85Cn, not even slight hints of any intermediates were observed upon unfolding of S49Cn. This indirectly supports the structural interpretation of the S85Cn intermediate which comprises unfolding of the outermost C-terminal  $\beta$ -sheet and  $\alpha$ -helix which are subjected to force within the S85Cn construct but not within S49Cn.

Besides a structural interpretation, what else can we say about the S85Cn intermediate? First of all, the S85Cn intermediate (S85Cn-I) does not show any sign of pre-flipping as opposed to the unfolding intermediate of S49Cc which basically corresponds to the first encountered unfolding intermediate of ROSS. This indicates that S85Cn-I is only populated after passing the predominant transition state of unfolding. With the additional information that S85Cn-I is populated at unfolding forces ranging from 6.5 up to 50 pN, one can conclude that the main transition state of S85Cn unfolding is shorter than S85Cn-I at almost all accessible forces. Another important quantity which characterizes S85Cn-I is the (uncorrected) observed average dwell time<sup>17</sup> being  $\tau_{\text{DW}}^{\text{S85Cn-I}} = 438$   $\mu\text{s}$ . Already in the sub-millisecond regime,  $\tau_{\text{DW}}^{\text{S85Cn-I}}$  is one order of magnitude shorter than the total transition path time of ROSS. Due to the fact that most S85Cn data were only sampled at 30 kHz, dwell times of less than  $\tau_{\text{c}} = 100$   $\mu\text{s}$  are likely to be missed<sup>18</sup>. To account for missed events, here, it suffices to simply subtract  $\tau_{\text{c}}$  from the measured average dwell time to get a good estimate for the 'true' dwell time  $\tau^*$  which is hence given as

<sup>17</sup> The average dwell time  $\tau_{\text{DW}}^{\text{S85Cn-I}}$  is based on 26 unfolding events during constant velocity and 7 unfolding or refolding events during constant distance measurements without differentiating between forces.

<sup>18</sup> To clearly detect intermediates, at least two successive points with similar deflection values and out of reach of both folded and unfolded states are necessary within a transition. With the additional time needed to diffuse into and out of the intermediate, this results in roughly 3 necessary point 'durations' when sampling at 30 kHz which adds up to  $\sim 100$   $\mu\text{s}$ .

$\tau_{\text{DW}}^{*,\text{S85Cn-I}} = 338 \mu\text{s}$ , see also [212]. According to these numbers, the expected fraction of missed events is then given by  $f_{\text{miss}} = 1 - \exp(-\frac{\tau_{\text{e}}}{\tau^*})$  which yields about 26% [214]. This value, in turn, fully explains why only 72% of all unfolding events clearly show an intermediate, see the conditional probability  $p(\text{I|U})$  in Table 17. By this means, S85Cn-I is not only a short-lived but also an obligatory on-pathway intermediate of S85Cn.

To be able to fully sample the unfolding force distributions of S85Cn or similarly mechanically resistant constructs within reasonable time, one might consider to complement data from optical tweezers with data from AFM experiments. Here, the recently reported microsecond resolution could further be useful to resolve and characterize intermediates in more detail [250].

Taken together, the much higher mechanical resistance which acts against directed N-terminal unfolding of ROSS confirms that forced mechanical unfolding of the native Rossmann fold starts from the C-terminus. Further, the on-pathway intermediate of the Rossmann fold which was detected within folding transitions of S85Cn is obligatory. Structurally, this intermediate most likely represents a fully folded N-terminal half of the Rossmann fold and still aligned core  $\beta$ -strands. After a correction for missed events, the average dwell time of  $\tau_{\text{DW}}^{*,\text{S85Cn-I}} \sim 300 \mu\text{s}$  places the intermediate's lifetime into the sub-millisecond regime.

## 6.4 EXTENDED DISCUSSION

The following discussion aims at studying two main aspects. One aspect is the further investigation of observed protein diffusion of R15 and ROSS from a simple two-state perspective. Here, analysis results from Section 6.2 are directly used to derive an effective roughness  $\epsilon_{\text{rms}}$  as well as friction coefficients  $\gamma_{\text{P}}$  for both proteins. Further, a merely diffusion based method for protein friction determination from the deformation of transition path ensembles is introduced and applied to R15 data. The second aspect comprises a more detailed characterization of the folding energy landscape of the Rossmann fold. Inspired by results from Section 6.3, on-pathway intermediate state positions are localised. Thereafter, a detailed energy landscape reconstruction based on transition path time distributions and folding rates of ROSS is conducted in the style of the technique elaborated in Section 4.2.3.

### 6.4.1 Effective roughness and friction under force

Based on experimentally determined transition path time averages and deconvolved energy landscapes from Section 6.2.3 and Section 6.2.4, here, the protein friction coefficient  $\gamma_{\text{P}}$  and the effective roughness  $\epsilon_{\text{rms}}$  of R15 and ROSS are derived according to Equation 54 which relates transition path times to energy landscape parameters and Equation 63 which characterizes the roughness concept of Zwanzig, see Section 3.8.4 and Section 4.2.2.

Deriving an effective roughness  $\epsilon_{\text{rms}}$  based on Equation 63 basically means to compare a measured quantity to a reference value which both relate to

the diffusion constant  $D$ . So, strictly speaking, the thereupon obtained  $\epsilon_{\text{rms}}$  only represents an *additional* effective roughness with respect to the reference. Now, since transition path times are inversely proportional to the diffusion coefficient  $D$ , see Equation 54, we commence with the straightforward derivation of the effective roughness of R15 and ROSS directly inferred by their measured transition path time averages with respect to the theoretical protein folding speed limit given by  $N/_{100} \mu\text{s}$ , where  $N$  is the number of amino acids which need to fold [119]. While relevant transition path times are listed in Table 18, the inferred additional effective roughness of  $\epsilon_{\text{rms}}^{\text{R15}} = 1.7 k_{\text{B}}\text{T}$  and  $\epsilon_{\text{rms}}^{\text{ROSS}} = 2.9 k_{\text{B}}\text{T}$  is illustrated by the gray potentials in Figure 64A and Figure 64B. Note that the superimposed effective roughness shown in Figure 64 is only an illustration which is not drawn to scale with respect to the x-axis. Here, the applied root-mean-squared roughness has a typical length scale of  $\Delta x_{\text{appl}}^r = 5 \text{ nm}$  which is fifty times longer than necessary to meet Zwanzig's specification of  $\Delta x^r = 0.1 \text{ nm}$  in case of nanometre scaling.

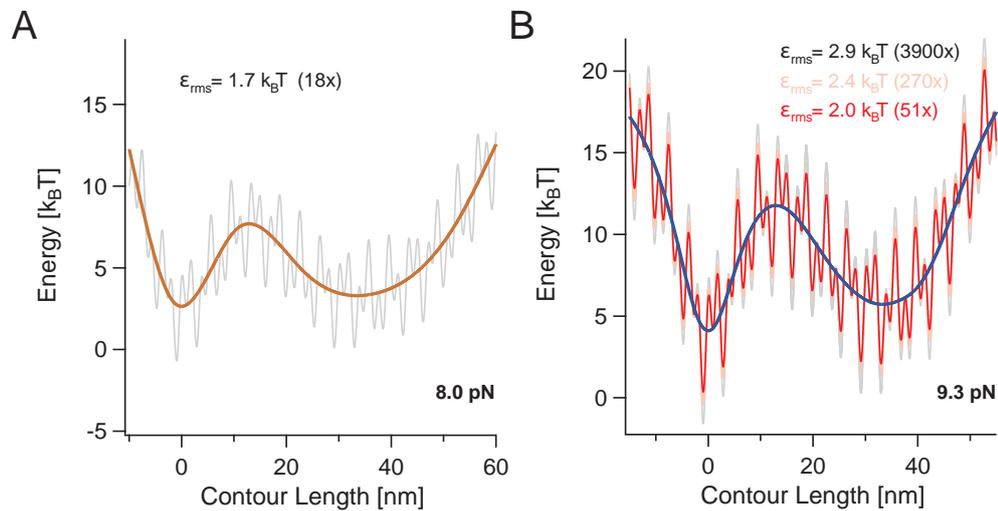


Figure 64: Effective energy landscape roughness of R15 and ROSS derived from transition path times measured under force. (A) With  $\langle \tau_{\text{TP}}^{\text{R15}} \rangle \approx 20 \mu\text{s}$ , R15 is 18 times slower than  $N/_{100} \mu\text{s} = 1.14 \mu\text{s}$  which would be expected for speed limit folding of the  $N = 114$  amino acids of R15. This difference can be explained by a smaller diffusion coefficient inferred by an additional effective roughness  $\epsilon_{\text{rms}}$  superimposed onto the folding energy landscape of R15, see Section 4.2.2. A representative illustration of the corresponding rough energy landscape is shown in light grey. (B). The additional roughness of ROSS with respect to speed limit folding is illustrated in light grey. Light red is the relative gain in roughness of ROSS with respect to R15 after ruling out differences in the respective energy landscape shapes. Red is the gain in roughness with respect to the diffusion coefficient of the designed protein FLred.

An additional roughness of  $\epsilon_{\text{rms}}^{\text{R15}} \sim 2 k_{\text{B}}\text{T}$  as derived for R15 is of the same order as the additional roughness which has been reported for RNA with respect to DNA hairpin unzipping [154]. Concerning ROSS, with respect to literature, a value of  $\epsilon_{\text{rms}}^{\text{ROSS}} \sim 3 k_{\text{B}}\text{T}$  already corresponds to the additional

roughness observed for misfolding of the prion protein PrP which represents a massive energy landscape distortion [249]. But how do R15 and ROSS directly compare in terms of roughness? This question is addressed next.

Given the deconvolved energy landscape shapes of R15 and ROSS as well as their average transition path times, it is possible to derive their corresponding friction coefficient  $\gamma_{(P)} = k_B T/D$  according to Equation 54. To account for the asymmetric barrier shapes within calculations, Equation 54 was individually applied to the four transition path segments delimited by the minima (which provide the relevant curvatures within the respective path segment), the maximum, and the deflection points in between. This way, as shown by diffusion simulations in Section 4.2.1, transition path time predictions based on Equation 54 are more accurate. Resulting friction coefficients  $\gamma_P(\tau_{TP}, G(L_P))$  are listed in Table 18 and can be used to directly compare R15 and ROSS with one another since now differences in energy landscape shapes are ruled out. The 270 times higher friction coefficient of ROSS yields an additional effective roughness of  $2.4 k_B T$  with respect to R15 which is illustrated by the light red potential in Figure 64B. Note that without accounting for the different deconvolved energy landscape shapes of R15 and ROSS, the mere difference between their transition path time averages only yields a 195 times higher friction coefficient.

PROTEIN	Speed limit		Experiment	
	$\langle \tau_{TP} \rangle$ [ $\mu\text{s}$ ]	$\gamma_P(\tau_{TP}, G(L_P))$ [ $\text{pN} \cdot \text{s} \cdot \text{nm}^{-1}$ ]	$\langle \tau_{TP} \rangle$ [ $\mu\text{s}$ ]	$\gamma_P(\tau_{TP}, G(L_P))$ [ $\text{pN} \cdot \text{s} \cdot \text{nm}^{-1}$ ]
R15	1.14	$4.8 \cdot 10^{-8}$	20	$9.7 \cdot 10^{-7}$
ROSS	1	$6.6 \cdot 10^{-8}$	3900	$2.6 \cdot 10^{-4}$
FLred	0.77	-	-	$(5.1 \cdot 10^{-6})$

Table 18: Theoretical speed limit versus experiment: protein friction coefficients implied by transition path times. Speed limit folding times are estimated according to  $N/100 \mu\text{s}$ . Protein friction values  $\gamma_P(\tau_{TP}, G(L_P))$  are inferred by an optimized sequential application of Equation 54 to respective deconvolved energy landscape shapes  $G(L_P)$  in connection with the indicated overall transition path times  $\tau_{TP}$  (see text). The friction coefficient of FLred is listed for reference, see Section 5.3.1 for its (different) derivation.

With respect to protein friction coefficients explicitly reported from single-molecule force spectroscopy experiments (corresponding values are marked at the bottom of Figure 29B), the value of  $\gamma_P^{R15}(\tau_{TP}, G(L_P)) = 9.7 \cdot 10^{-7} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , which is obtained for R15, appears to be the lowest [143, 177, 247]. Compared to the friction coefficient of the designed protein FLred which was derived in Section 5.3.1, the Rossmann fold has an only about 50 times higher value which corresponds to an additional roughness of  $2.0 k_B T$ .

Note that the friction coefficients inferred by the theoretical speed limit folding times of R15 and ROSS very well match the value of  $\gamma_P = 4 \cdot 10^{-8} \text{ pN} \cdot$

$\text{s} \cdot \text{nm}^{-1}$  used to simulate speed limit protein folding in [Section 4.3](#). In principle, this confirms that the choice of unfolded contour length space as reaction coordinate is suitable to actually model speed limit protein folding.

In summary, R15 experiences an effective roughness of  $\epsilon_{\text{rms}}^{\text{R15}} \sim 2 \text{ k}_B\text{T}$  with respect to speed limit protein folding while ROSS has a significantly higher roughness of  $\epsilon_{\text{rms}}^{\text{ROSS}} \sim 3 \text{ k}_B\text{T}$ . Upon direct comparison, ROSS has a remarkable additional roughness of  $2.4 \text{ k}_B\text{T}$  with respect to R15. Further, together with the deconvolved energy landscape shapes, measured transition path times imply  $\gamma_{\text{p}}^{\text{R15}}(\tau_{\text{TP}}, G(\text{L}_P)) = 9.7 \cdot 10^{-7} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  and  $\gamma_{\text{p}}^{\text{ROSS}}(\tau_{\text{TP}}, G(\text{L}_P)) = 2.6 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  which lie both within the ranges found in [Section 6.2.4](#).

#### 6.4.2 Transition-path-assisted protein friction determination

As suggested by two-dimensional diffusion simulations in [Section 4.3.2](#) and [Section 4.3.4](#), the shapes of detected transition path ensembles are sensitive towards changes in protein friction. Here, after introducing a measure  $\Delta p(x|\text{TP})$  for relative transition path ensemble deformation, the detected friction-induced deformation of the transition path ensemble of R15 is used to derive its corresponding friction coefficient  $\gamma_{\text{p}}^{\text{R15}}(\Delta p)$ .

A straightforward way to characterize the relative deformation of an ensemble  $p(x|\text{TP})$  with respect to a reference  $p_0(x|\text{TP})$  is by deriving their non-overlapping area. This can simply be done by integrating over the absolute values of the ensemble difference:

$$\Delta p(x|\text{TP}) = \int_{x_{\text{min}}}^{x_{\text{max}}} dx |p(x|\text{TP}) - p_0(x|\text{TP})|. \quad (68)$$

Here, the borders  $x_{\text{min}}$  and  $x_{\text{max}}$  should be chosen such that both distributions are completely included. In the following, [Equation 68](#) will be applied to already normalized distributions. Then, together with a common reference  $p_0(x|\text{TP})$ ,  $\Delta p(x|\text{TP})$  values which result from different ensembles  $p(x|\text{TP})$  can be directly compared to each other. If, as in our case, the distributions are normalized and the scaling in 'x' is given in nanometres, a maximal deformation of 2 nm can be reached for non-overlapping distributions<sup>19</sup>.

The top row of [Figure 65](#) introduces the principle of the merely diffusion-based protein friction determination. While the full line in [Figure 65A](#) shows the experimentally determined transition path ensemble of R15, full lines in [Figure 65B](#) display transition path ensembles obtained from two-dimensional diffusion simulations based on the deconvolved energy landscape of R15. Dashed lines in [Figure 65A](#) and [Figure 65B](#) refer to ensembles expected for unperturbed one-dimensional bead diffusion according to theory (red) and simulations sampled at 200 kHz (black). In [Figure 65B](#), the difference in shape among the simulated ensembles results from varying the respective protein friction coefficient  $\gamma_{\text{p}}$  from low (black) towards high friction (light blue). Next, the deformation values  $\Delta p(x|\text{TP})$  of the simulated ensembles are

<sup>19</sup> Non-overlapping distributions could, in principle, arise due to endless friction which, in turn, would result in a delta function.

plotted against their corresponding friction coefficients which yields the master 'curve' represented by full squares in Figure 65C. Finally, the friction coefficient of R15 is given by the intersection of the friction-induced deformation observed for R15 with the master curve. With  $\Delta p$  (R15) being represented by the dashed horizontal line in Figure 65C, the intersection yields  $\gamma_P^{R15}(\Delta p) \sim 1 \cdot 10^{-6 \pm 1.5} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  as highlighted by the empty square. Here, the error in friction indicates the maximal uncertainty implied by the error in  $\Delta p$  which, in turn, was found by bootstrapping, see below.

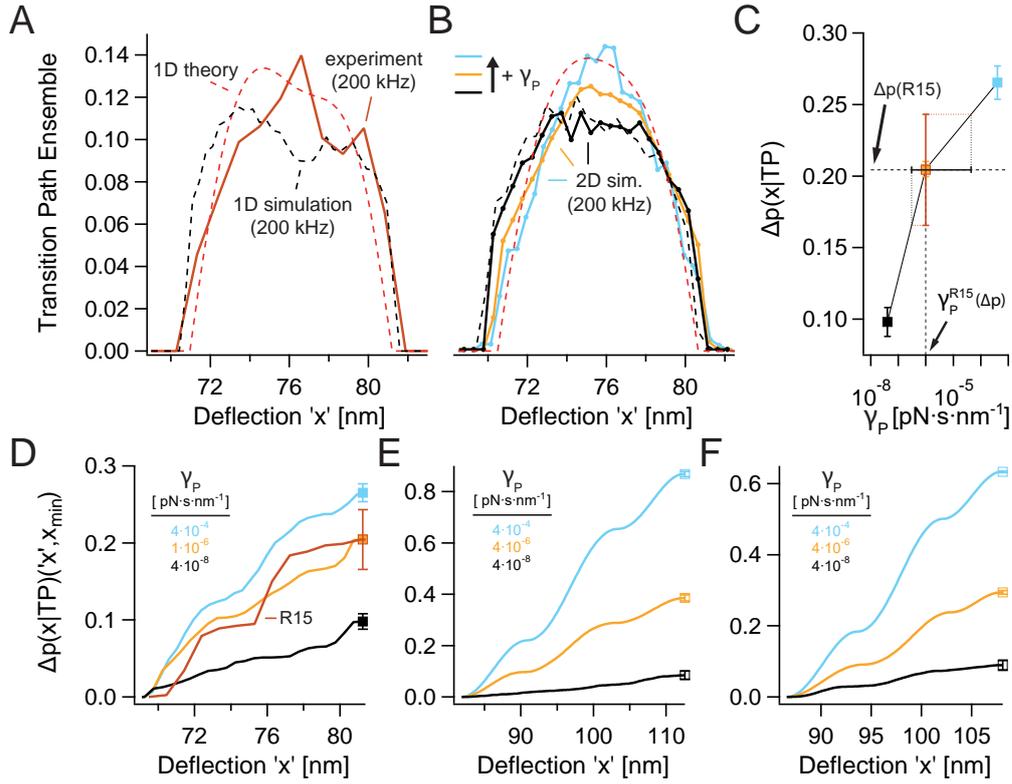


Figure 65: Friction coefficient determination of R15 based on the friction-induced deformation of the measured transition path ensemble. (A) Measured transition path ensemble of R15 (brick red). Dashed lines in (A) and (B) refer to ensembles expected for unperturbed one-dimensional bead diffusion. (B) Black, orange, and light blue lines represent ensembles derived from two-dimensional R15 diffusion simulations where  $\gamma_P = 4 \cdot 10^{-8}$ ,  $1 \cdot 10^{-6}$ , and  $4 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . (C) Deformation values  $\Delta p(x|TP)$  provided by bootstrapped ensembles in (A) and (B). (D) Stepwise evaluation of Equation 68 reaches at final values shown in (C). (E) and (F) Stepwise evaluation of Equation 68 along transition path ensembles from highly resolved two-dimensional diffusion simulations in symmetric and asymmetric potentials, see Figure 35 and Figure 38. See text for more details.

The bottom row of Figure 65 illustrates the stepwise evaluation of Equation 68 which reaches final  $\Delta p(x|TP)$  values (squares) when 'x' reaches  $x_{\max}$ . In Figure 65D, the measured and simulated ensembles of R15 from Figure 65A and Figure 65B are evaluated. To account for effects arising from limited reso-

lution, see [Section 4.3.4](#), here, ensembles from unperturbed one-dimensional bead diffusion simulations sampled at 200 kHz are used as common reference  $p_0(x|TP)$  in [Equation 68](#) to better comply with experiments. [Figure 65E](#) and [Figure 65F](#) depict the stepwise evaluation of ensembles from other friction-dependent simulations performed in [Section 4.3](#) which, in this context, serve as a proof of principle. Owing to much higher sampling at 1 GHz and better spatial resolution, in this case, theoretical ensembles for unperturbed bead diffusion were used as  $p_0(x|TP)$ . The respective results clearly show that an increase in friction results in an increase of deformation values calculated according to [Equation 68](#).

In [Figure 65](#), all final  $\Delta p(x|TP)$  values and their errors were determined by bootstrapping<sup>20</sup>, i. e., random sampling with replacement of transition paths to repeatedly create transition path ensembles based on a given transition path pool<sup>21</sup>. This, in turn, allows a repeated evaluation of [Equation 68](#) to be made. Note that lines in [Figure 65D](#) - [Figure 65F](#) do not represent an average from repeated evaluation of [Equation 68](#), but a single evaluation of an overall transition path ensemble average which was adapted to match final  $\Delta p(x|TP)$  values if required. Opposed to simulations in [Figure 65E](#) and [Figure 65F](#), in [Figure 65B](#) - [Figure 65D](#) a friction coefficient of  $\gamma_P^{R15, sim} = 1 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  was chosen instead of  $4 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The different choice is intended to account for potential non-linearities of the friction-dependence of  $\Delta p(x|TP)$  around the friction expected for R15 based on the previously derived value listed in [Table 18](#).

Despite their similar protein friction coefficients, simulations in the bottom row of [Figure 65](#) yield considerably different final deformation values  $\Delta p(x|TP)$ . This divergence is mainly caused by differences in spatial separation of the folded and unfolded states which are affected by the acting force<sup>22</sup> in combination with the unfolded contour length<sup>23</sup>, see deflection values ' $\chi$ '. Further, the relatively large error of the final deformation value of the experimental transition path ensemble of R15 is owed to relatively small statistics provided by only 49 transitions compared to at least 325 transitions from simulations sampled at the same frequency. However, in combination with bootstrapping, still a very reasonable result is obtained which speaks for the robustness of the presented method.

As transition path ensembles relate to their underlying energy landscapes, so do the friction-induced transition path ensemble deformations. Note that stretches of steep ascent of the stepwise evaluation of [Equation 68](#) in [Figure 65D](#) - [Figure 65F](#) correspond to regions with locally increased or decreased 'path-dwelling' times which are expected around the maxima or min-

<sup>20</sup> Here, errors in  $\Delta p(x|TP)$  indicate confidence intervals of 95 %.

<sup>21</sup> Results in [Figure 65](#) are based on the following amount of transition paths. R15, [Figure 65A](#), [Figure 65C](#), [Figure 65D](#) - experiment: 49; simulations from low to high  $\gamma_P$ : 970, 2924, 325. Simulations, [Figure 65E](#) - symmetric potential from low to high  $\gamma_P$ : 136, 557, 105. Simulations, [Figure 65F](#) - asymmetric potential from low to high  $\gamma_P$ : 392, 3488, 3133.

<sup>22</sup> Here,  $k_{eff} \sim 0.1 \text{ pN/nm}$  for all simulations as well as the experiment. Acting forces are given by multiplication of  $k_{eff}$  with corresponding deflection values ' $\chi$ '.

<sup>23</sup> Unfolded contour lengths in [Figure 65D](#) - [Figure 65F](#) are:  $\sim 33 \text{ nm}$ ,  $\sim 76 \text{ nm}$ ,  $\sim 52 \text{ nm}$ .

ima of the underlying potential. Corresponding tendencies can already be seen in [Figure 65A](#) and [Figure 65B](#). Here, upon close inspection, the friction-induced deformation has a slight asymmetry with a higher path weight on the folded side. This correctly relates to the transition state position of R15 which was already found to be much closer to the folded than the unfolded state based on kinetics, see [Table 8](#) or [Table 9](#).

Another interesting point of speculation is the pronounced deviation between the experimental and simulated trajectory of R15 in [Figure 65D](#) which are both based on ensemble averages and which both end up having about the same final  $\Delta p(x|TP)$  value. Partially, this may be explained by differences between the respective unperturbed bead diffusion references  $p_0(x|TP)$ , see black dashed lines in [Figure 65A](#) and [Figure 65B](#). However, an even more intriguing explanation would be a position-dependent, non-uniform friction coefficient  $\gamma_p^{R15}(\Delta p(x|TP))$  to which measured transition path ensembles should be sensitive in accord with similar thoughts published elsewhere [[137](#), [158](#)]. Given the method introduced in this section, direct proof for position-dependent protein friction is, hence, only a bit more statistics and preferably slightly better resolution away.

In addition to potentially paving the way towards detecting position-dependent protein friction, what are the advantages of the herein introduced alternative method for deriving protein friction coefficients? By being purely diffusion-based, this method probably represents the most direct way of deriving a diffusion constant  $D$  where the complex interdependence of protein and bead diffusion is automatically considered by a relatively simple simulation. Given that, the method is independent of any additional model which is typically needed to relate diffusion-altered kinetic rates or transition path times to the shape of an underlying potential. Further, provided that statistics are high enough, deriving protein friction coefficients based on the deformation of detected transition path ensembles should even be possible for older force-spectroscopic measurement data including [AFM](#) experiments.

Taken together, the friction coefficient of  $\gamma_p^{R15}(\Delta p) \sim 1 \cdot 10^{-6 \pm 1.5} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  which results from the detected friction-induced transition path ensemble deformation of R15 very well matches the transition path time based value  $\gamma_p^{R15}(\tau_{TP}, G(L_P))$  from [Section 6.4.1](#). In return, the value from this section implies an average transition path time of  $24 \mu\text{s}$  which is in perfect agreement with previous results from [Section 6.2.3](#). By this means, the herein introduced method to derive protein friction coefficients based on detectable friction-induced transition path ensemble deformations  $\Delta p(x|TP)$  yields very reasonable results. Importantly, already a relatively small number of transition events, here 49, is enough to get meaningful protein friction coefficient estimates.

### 6.4.3 *Localization and structural interpretation of high-energy intermediates of the Rossmann fold*

This section is dedicated to a more detailed inspection of the folding transitions of ROSS. As already pointed out during analysis of different mutants in [Section 6.3.2](#) and [Section 6.3.3](#), transitions of the Rossmann fold typically involve passing through intermediates. In the following, these intermediates are localised based on *WLC*-fits as well as transition path histograms and ensembles from constant velocity and constant distance data.

The top of [Figure 66](#) displays a selection of ROSS unfolding transitions recorded during 500 nm/s constant velocity cycles. Thanks to the contour length transformation outlined in [Section 3.1.1](#), events which occur at different forces can be directly aligned for better comparison. [Figure 66B](#) presents a contour length gain histogram based on *WLC*-fits to intermediates within unfolding transitions as well as to the fully unfolded state. Here, data are based on 309 transitions of one single molecule without making a difference between N, NL, and Rc fingerprints. The transition path histogram shown in [Figure 66C](#) stems from another molecule's 85 unfolding transition paths observed during 500 nm/s constant velocity cycles. In [Figure 66A](#), 8 of these paths are highlighted by darker colors<sup>24</sup>. Again, all except PrL fingerprints are included. Transition path ensembles depicted in [Figure 66D](#) represent five different constant distance trajectories from four different molecules before (dashed lines) and after deconvolution (full lines). To better fit into one graph, ensembles from before and after deconvolution do not have the same y-axis scaling. On average, ensembles in [Figure 66D](#) include 13 transition paths extracted from successive equilibrium unfolding and refolding transitions, see [Figure 54B](#) for examples. [Figure 66E](#) details the deconvolved ensembles from [Figure 66D](#) after their transformation into contour space. Here, the contour length transformation makes possible the comparison of the spatial information provided by transition path ensembles from different experiments.

An overview of the results obtained from analysing the data presented in [Figure 66](#) is given in [Table 19](#). The top row indicates contour length gains which are expected upon unfolding of the secondary structural elements specified in brackets below. Here, the particular choice of unfolded secondary structure assumes that unfolding proceeds from C- to N-terminus and that expected contour length gains should be as close as possible to the experimental ones. The second row holds the results from fitting a Gaussian four-peak distribution (thick black line) to the histogram shown in [Figure 66B](#). While originating from only one single molecule, the value for the completely unfolded state (U) compares very well to  $L_p^{CV} = 34.7 \pm 0.9$  from [Table 10](#) which is based on all measured molecules. Hence, values obtained for the three predominant on-pathway intermediate positions can be expected to very well represent all molecules. Results from the Gaussian three-peak distribution

<sup>24</sup> As for constant distance trajectories, transition paths from constant velocity measurements of ROSS were identified by *HMM* analysis to overcome issues with thermal noise, see also [Section 6.2.3](#). Here, the analysis was directly done in contour space.

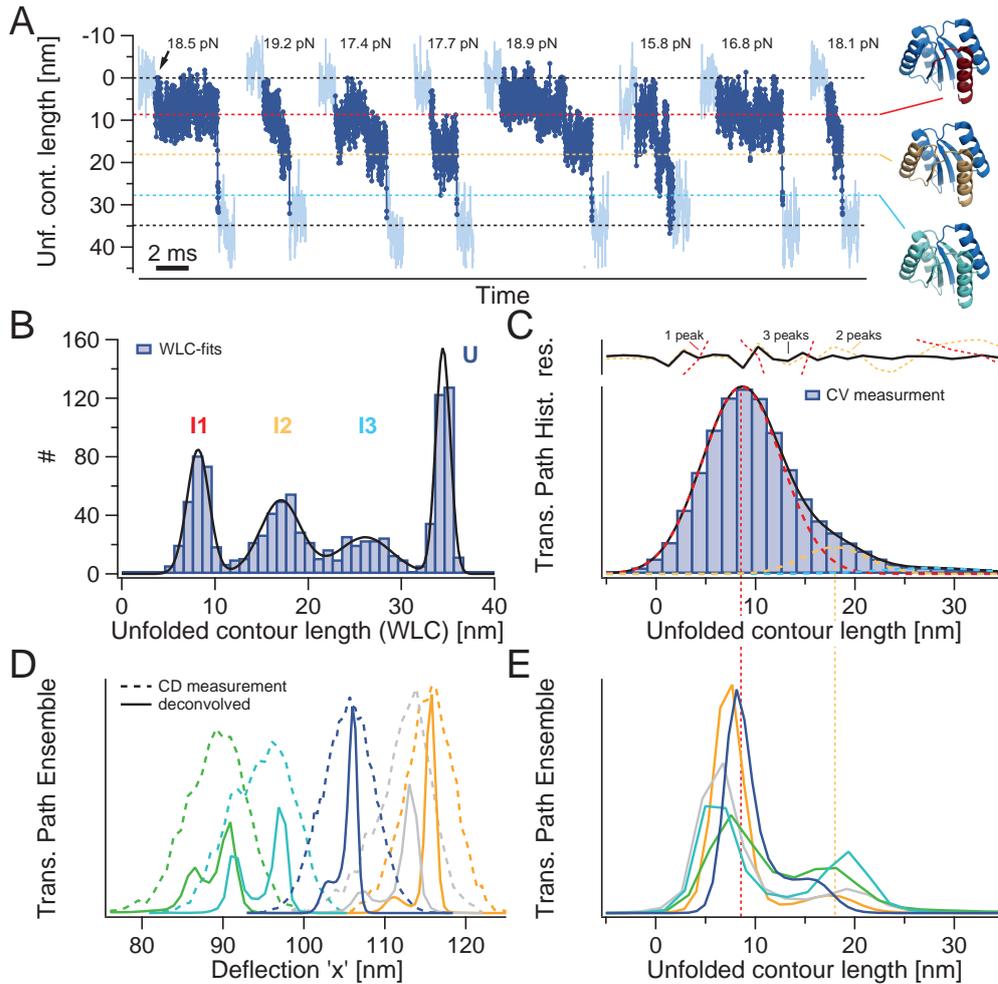


Figure 66: Localization and structural interpretation of short-lived on-pathway intermediates of ROSS. (A) Unfolding transitions from 500 nm/s pulls transformed into contour space. (B) Contour length gain histogram based on *WLC*-fits. (C) and (D) Transition path histogram and ensembles extracted from constant velocity and constant distance data. (E) Deconvolved ensembles from (D) transformed into contour space. See text for more details.

(thick black line) to the transition path histogram depicted in Figure 66C are listed in the next to last row of Table 19. As one can see, the contribution of the third peak (light blue dashed line) to the entire histogram in Figure 66C is very small by only having a relative area of 4% with respect to 86% and 10% for the other two peaks. However, residuals in the top of Figure 66C confirm that the best fit result is obtained with a three-peak distribution. The good agreement with the *WLC*-fit results from the above row is another proof for the existence of a third on-pathway intermediate. All errors for contour lengths extracted from constant velocity (CV) data are standard deviations derived from the Full Width at Half Maximum (FWHM) values of the fitted Gaussian peaks according to  $\sigma = \text{FWHM}/2\sqrt{2 \cdot \ln(2)}$ . An important difference between Figure 66B and Figure 66C is that in addition to spatial information,

the transition path histogram in [Figure 66C](#) already includes temporal information in the sense of relative state occupancies weighted according to the probed force distribution. By this means, the difference in relative peak area in [Figure 66C](#) already suggests that the first intermediate I<sub>1</sub> has the longest and the third intermediate I<sub>3</sub> the shortest dwell time average which can also be estimated by eye in [Figure 66A](#). The last row in [Table 19](#) contains the two main peak averages of the deconvolved transition path ensembles shown in [Figure 66E](#). While the third intermediate I<sub>3</sub> could obviously not be resolved by deconvolution due to its very short dwell time, the very good agreement of the I<sub>1</sub> and I<sub>2</sub> positions with respect to alternative results confirms that it is possible to extract on-pathway intermediate positions by deconvolving transition path ensembles. Note that these intermediates were invisible upon deconvolution of the entire equilibrium distributions.

METHOD	I <sub>1</sub>	I <sub>2</sub>	I <sub>3</sub>	U
	[nm]			
$L_p^{\text{calc}}$ ( <a href="#">Equation 5</a> )	7.49 / 9.9 ( $\alpha 4 / \alpha 4 \beta 4$ )	16.13 / 19.31 ( $\alpha 4 - \alpha 3 / \alpha 4 - \beta 3$ )	26.82 ( $\alpha 4 - \beta 2$ )	34.51 ( $\alpha 4 - \beta 1$ )
WLC-fits (CV)	$8.3 \pm 1.1$	$17.6 \pm 1.9$	$26.5 \pm 2.1$	$34.6 \pm 0.8$
TP histogram (CV)	$8.6 \pm 4.0$	$18.0 \pm 3.3$	$27.6 \pm 5.3$	-
TP ensemble (CD)	$7.3 \pm 0.8$	$17.9 \pm 1.7$	-	-

Table 19: Predominant on-pathway intermediate positions of ROSS given in measures of unfolded contour length. Results are based on data presented in [Figure 66](#). The fully unfolded state (U) is included for completeness. Structural elements which are unfolded upon reaching the corresponding intermediate are indicated in brackets.

According to results from [Table 19](#), up to three on-pathway intermediates were resolved upon a detailed inspection of the folding transitions of the Rossmann fold. In [Figure 66A](#), coloured dashed horizontal lines indicate the three positions which were derived from the transition path histogram in [Figure 66C](#). Here, upon closer inspection, it even seems as if identified intermediates were comprised of more fast interchanging intermediates. This is why found intermediate positions are referred to as predominant. For this reason, let us have another look at intermediate I<sub>1</sub> of ROSS which ought to be the same as intermediate 'I' observed for S<sub>49</sub>Cc (S<sub>49</sub>Cc-I) in [Section 6.3.2](#). For ROSS, in [Table 19](#), the WLC-fit average for I<sub>1</sub> is considerably longer than the calculated value expected for the mere unfolding of the C-terminal  $\alpha$ -helix  $\alpha 4$  but still too short to be identified as the full unfolding of the C-terminal  $\alpha$ -helix and  $\beta$ -strand  $\alpha 4 \beta 4$ . In [Table 14](#), the average for S<sub>49</sub>Cc-I is already further away from  $\alpha 4$  and closer to  $\alpha 4 \beta 4$  but still remarkably shorter than expected for full unfolding of  $\alpha 4 \beta 4$ . Hence, I<sub>1</sub> could be an effective superposition of two fast interchanging intermediates. A plausible reason for such fluctuations may, e. g., be the recently reported differences in internal friction

involved in the formation of ( $\alpha$ -)helices and ( $\beta$ -)hairpins where, according to all-atom simulations, hairpins experience less friction [252]. By this means, the intermediate I<sub>1</sub> of ROSS might comprise  $\beta_4\beta_3$  hairpin fluctuations. Along the same lines, I<sub>2</sub> of ROSS might include  $\beta_3\beta_2$  hairpin fluctuations which are slightly more biased towards an open hairpin configuration as compared to S85Cn.

Another point of debate with respect to the structural interpretation of the predominant intermediates of the Rossmann fold is the following question. Is the assumption that unfolding of ROSS proceeds from C- to N-terminus justified? In short: yes. The strongest evidence for this assumption to be correct is given by the independent structural interpretation of the intermediates S49Cc-I and S85Cn-I from Section 6.3.2 and Section 6.3.3 which basically correspond to the intermediates I<sub>1</sub> and I<sub>2</sub> of ROSS. And, in anticipation of their derivation in Section 6.4.4, by lasting  $\sim 2.5$  ms and  $\sim 135$   $\mu$ s, the typical dwell times of I<sub>1</sub> and I<sub>2</sub> also match those observed for S49Cc-I and S85Cn-I which were independently found to lie within the millisecond and sub-millisecond regime. In view of these numbers, the difference in relative area of the three peaks in Figure 66C implies that I<sub>3</sub> should have typical dwell times of less than 100  $\mu$ s. With respect to its relatively broad distribution in Figure 66B, an exact structural interpretation of I<sub>3</sub> as given in Table 19 is not really possible. Similar to previous thoughts on fast interchanging intermediates, I<sub>3</sub> might be any combination of two or three secondary structural elements of the N-terminal half of ROSS.

In addition to a direct investigation of the productive transitions of the Rossmann fold as conducted in this section, there exist other ways of tracking down folding intermediates which are not resolvable by mere deconvolution of entire equilibrium fluctuations. One of these ways has been applied in reference [248], where histograms of the folded and unfolded state are fitted separately with multiple skewed Gaussian distributions which represent the system's characteristic PSF. Considering the multitude of unsuccessful folding attempts of the Rossmann fold which can be seen in the lower (unfolded) state level of the constant distance trajectory in Figure 12, this kind of analysis may be used to look for additional off-pathway intermediates which do not match positions of the on-pathway intermediates characterized here. Another way to reveal very fast transitions which go beyond the detection limit of all other techniques is by autocorrelation analysis, see Section 3.4.4. This could be of particular interest for a further analysis of the folded state of ROSS.

In conclusion, the principle result of this section is that the folding transitions of the Rossmann fold pass through three predominant on-pathway intermediates I<sub>1</sub>, I<sub>2</sub>, and I<sub>3</sub> which corroborate the picture of a sequential folding and unfolding process. Data are consistent with unfolding to start from the C-terminus and to proceed towards the N-terminal end which is strongly supported by close to identical structural interpretations and dwell times of I<sub>1</sub> and I<sub>2</sub> with respect to the independently measured obligatory on-pathway intermediates of S49Cc and S85Cn. Further, intermediates I<sub>1</sub> and I<sub>2</sub> seem to involve fast  $\beta_4\beta_3$  and  $\beta_3\beta_2$  hairpin fluctuations. Within the experimentally

accessible range, I<sub>3</sub> was found to have the shortest dwell times and the most undefined structure.

#### 6.4.4 *Transition-path-assisted energy landscape reconstruction*

If a rough energy landscape is composed of a number of localised high-energy on-pathway intermediates, overall transition path times basically reflect a sum of dwell times. To decipher the energy landscape roughness of the Rossmann fold, here, the dwell-time-based energy landscape reconstruction technique, which was developed and tested using simulations in [Section 4.2.3](#), is now applied to experimental data.

At this point, all important landmarks which characterize the (native) folding energy landscape of the Rossmann fold are known. With respect to the folded state (N), unfolded contour lengths of the four transition states (TS<sub>1</sub> - TS<sub>4</sub>), three intermediates (I<sub>1</sub> - I<sub>3</sub>), and the fully unfolded state (U) are listed in [Table 9](#) and [Table 19](#). All these positions are marked by dark blue squares in the reconstructed landscape in [Figure 67A](#). Now, starting only from the known positions, the already plotted barrier heights of the energy landscape can be iteratively reconstructed by using the information provided by the transition path time distribution and the rate plots shown in [Figure 67B](#) and [Figure 67C](#). In a nutshell, the fits included in both graphs provide dwell times which characterize different barrier crossings at different forces. Given that relevant transition state positions  $\Delta x^{\text{TS}}$  are already known, expected dwell times for crossing an assumed harmonic barrier can be directly calculated as a function of the barrier height  $\Delta G^{\text{TS}}$  based on [Equation 55](#) in combination with Kramers rate equation [Equation 29](#), see [Section 4.2.3](#) for details. From such a dwell time versus barrier height plot, an unknown barrier height can be directly read off as the value corresponding to the measured dwell time, see also [Figure 24B](#). Naturally, the friction coefficient which was previously derived for ROSS based on its average transition path time is not applicable in this part. Instead, here,  $\gamma_{\text{P}} = \gamma_{\text{FLred}} = 5.1 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  from the more 'well-behaved' artificial protein FLred is used for calculations. Further, owing to the relatively long dwell times, transition path times for actual barrier crossing can, for the moment, be neglected.

According to requirements in [Section 4.2.3.2](#), we are now fully set to begin with the reconstruction. Fitting [Equation 64](#) to the transition path time distribution in [Figure 67B](#) yields  $\tau_1 = 2.5 \pm 0.4 \text{ ms}$  and  $\tau_2 = 135 \pm 46 \mu\text{s}$ . This leaves us with two dwell times for transition path times arising from passing through three intermediate states. Since we already know that I<sub>3</sub> is barely resolved within constant distance measurements,  $\tau_1$  and  $\tau_2$  can essentially be attributed to I<sub>1</sub> and I<sub>2</sub>. Importantly, these values completely agree with the millisecond and sub-millisecond lifetimes which were independently derived for I<sub>1</sub> and I<sub>2</sub> by the S49Cc and S85Cn constructs. To be still able to determine a barrier height for I<sub>3</sub>, we make use of the information provided by the transition path histogram in [Figure 66C](#) to derive an upper dwell time limit  $\tau_3$  for I<sub>3</sub> with respect to  $\tau_2$ : compared to the 10% of total area for I<sub>2</sub>, the

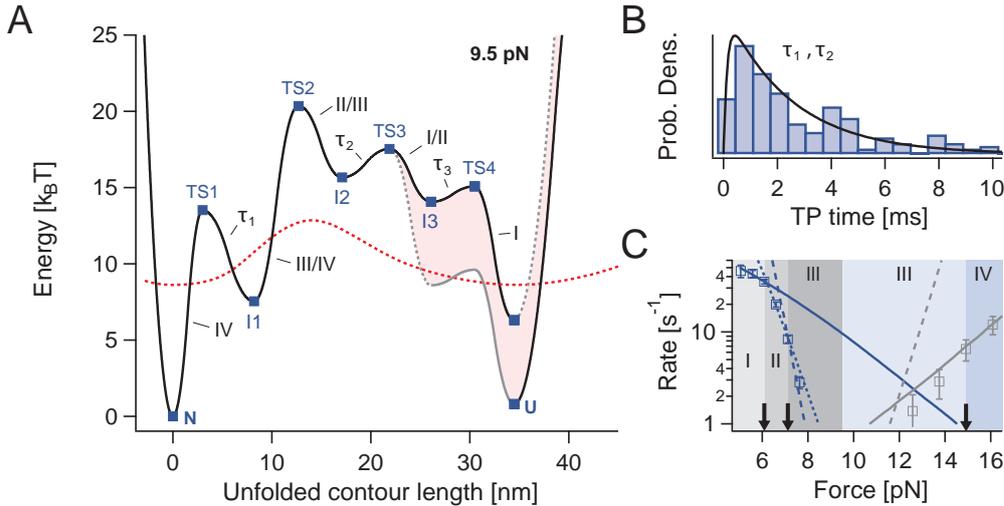


Figure 67: Transition-path-assisted energy landscape reconstruction for ROSS at 9.5 pN. (A) Assembled energy landscape of ROSS reconstructed according to the method elaborated in Section 4.2.3. Blue squares are landmarks based on average contour length gains of WLC-fits (minima) and slopes of SB-model fits (maxima), see Table 19 and Table 9. Barrier heights are marked according to their derivation being either based on dwell times  $\tau$ , predominant rates, or transition state switches within effective rates. See text for details. For reference, the dashed red line shows a deconvolved energy landscape of ROSS tilted towards 9.5 pN. (B) Transition path time distribution from Figure 54C fitted by Equation 64. (C) Zoom into the force-dependent rate plots from Figure 52B. Background colours mark force ranges I to IV of the only locally valid SB-model fits. Arrows point towards forces with a transition state switch.

4 % of  $I_3$  correspond to 40 %. As these areas correlate with the relative time spent within proximity of the respective states, this leads us to an estimate of  $\tau_3 \sim 50 \mu\text{s}$ . Next, we need to know which barrier is determinant for each dwell time. This is where the rate plots shown in Figure 67C come into play as they indicate which transition state is predominant at which force. Since the distribution in Figure 67B only comprises transition path times from constant distance trajectories recorded around the mid-force  $F_{\text{mid}}$  of 9.5 pN, see full circles in Figure 52B, the previously determined dwell times belong to force regime number III where transition state TS2 is predominant. From an energetic perspective, this puts TS1 and TS3 below TS2, and TS4 below TS3. To prevent confusion: regimes I - IV, i. e., going from low to high force in Figure 67C, provide transition state positions TS4 - TS1 in Figure 67A. Finally, after finding the determinant barrier positions, we can derive the following barrier heights<sup>25</sup> for intermediates  $I_1 - I_3$ :  $\Delta G_{9.5}^{I_1 \cdot \text{TS}_1} \sim 6 \text{ k}_B\text{T}$ ,  $\Delta G_{9.5}^{I_2 \cdot \text{TS}_3} \sim 2 \text{ k}_B\text{T}$ , and  $\Delta G_{9.5}^{I_3 \cdot \text{TS}_4} \sim 1.5 \text{ k}_B\text{T}$ .

After having determined one barrier for each intermediate, how can we derive the second one from which we only know that, at 9.5 pN, it must

<sup>25</sup> Note that in this section, all barrier heights are unsigned for convenience.

be significantly higher? The key to answer to this question is provided by force-induced predominant transition state switches. When a transition state switch is on hand, rates for crossing either of the two neighbouring transition states are equal when starting from the minimum located in between them, see also [Figure 25A](#). Moreover, according to findings in [Section 4.2.4](#), force-induced transition state switches cause kinks within force-dependent rate plots which allows to localise them easily. In [Figure 67C](#), forces at which a transition state switch occurs within the folding transitions of the Rossmann fold are marked by black arrows. Now, after tilting<sup>26</sup> a known barrier of an intermediate towards the force of the relevant transition state switch, the tilted barrier provides a transition rate from which the unknown barrier height can be directly derived since it has to provide the exact same rate. Thereafter, the newly found barrier height simply needs to be tilted back to 9.5 pN. As an example, let us consider the second barrier for intermediate I<sub>1</sub>. Here, the relevant transition state switch III/IV between regime III and IV occurs at 15 pN. Tilting  $\Delta G_{9.5}^{I_1 \cdot TS_1}$  towards this force yields  $\Delta G_{15}^{I_1 \cdot TS_1} \sim 10 \text{ k}_B T$ . By setting  $k_{15}^{I_1 \cdot N} = k_{15}^{I_1 \cdot I_2}$ , we reach at  $\Delta G_{15}^{I_1 \cdot TS_2} \sim 10 \text{ k}_B T$ . After tilting barriers back to 9.5 pN, we arrive at  $\Delta G_{9.5}^{I_1 \cdot TS_2} \sim 13 \text{ k}_B T$ . The same procedure was done for I<sub>2</sub> and I<sub>3</sub> as well. Note that the example of I<sub>1</sub> further means that at 15 pN a maximal local roughness of about 10 k<sub>B</sub>T is attained.

The last two missing barrier heights are  $\Delta G_{9.5}^{N \cdot TS_1}$  and  $\Delta G_{9.5}^{U \cdot TS_4}$ . Both barriers are the only ones which can, in principle, be directly derived from measured transition rates which lie far enough within force ranges IV and I, respectively. However, to be on the safe side, it is best to use the rates provided at transition state switch positions III/IV and I/II as there the respective measured overall transition rate is exactly half the rate for crossing the barrier we want to estimate, see also [Figure 25C](#). Now, before stitching all energy barriers together, we add one more essential piece of information to the overall picture. In addition to transition path times and effective rates, another known quantity is the zero-force folding free energy  $\Delta G_0$  of the Rossmann fold which indicates the relative energy between the folded and unfolded state. By tilting this energy difference, we arrive at  $\Delta G_8^{N \cdot U} \sim 0.8 \text{ k}_B T$ . Note that the folding free energy was only tilted towards 8.0 instead of 9.5 pN. This is done to account for a difference in force of about 1.5 pN between the folded and unfolded state which arises due to the passive mode of our constant distance measurements.

What can we say about the final result of the fully reconstructed energy landscape of the Rossmann fold shown in [Figure 67A](#)? In general, compared to the red dashed line which represents a deconvolved energy landscape tilted towards 9.5 pN, the reconstructed landscape has a much higher level of detail. Notably, the main transition state position as well as the relative energy levels between the native and unfolded states are almost the same for the deconvolved and reconstructed landscapes. However, within the reconstructed landscape, there is a shaded area which illustrates an energy mismatch of about 5 k<sub>B</sub>T which occurs around transition state position TS<sub>4</sub>. In

<sup>26</sup> Note that the folding energy landscape needs to be tilted in deflection space to correctly apply changes in acting force which, basically, corresponds to  $\Delta G_F = \Delta G_0 - F \cdot \Delta x_{\text{tether}}$ .

principle, this mismatch arises from matching the energy difference between folded and unfolded state to the value obtained from tilting the overall folding free energy. Among potential reasons for the origin of this mismatch are artefacts arising from inaccuracies of the reconstruction, the already known multi-dimensionality of the Rossmann fold which involves multiple fingerprints, position-dependent friction, additional unresolved intermediates, or anything else which can originate from projecting the folding of a globular protein on our single measurement coordinate along the vector of the acting force. Then, how do we know whether the reconstructed energy landscape provides any meaningful information? Here, we can test this by comparing the zero-force extrapolated folding free energy of intermediate I<sub>1</sub> provided by the reconstructed landscape to the energy derived for the intermediate I from S49Cc measurements in Section 6.3.2. Since ROSS and S49Cc share the same unfolding pathway and (first) intermediate, the respective energies should match. And indeed, the value of  $\Delta G_0^{N \cdot I_1} \sim 17 \text{ k}_B\text{T}$  from the reconstructed landscape corresponds very well to the  $\Delta G_0^{F \cdot I} \sim 15 \text{ k}_B\text{T}$  from S49Cc listed in Table 16. Such a good match is very strong evidence for a meaningful energy landscape reconstruction.

Note that the superposition of the reconstructed and deconvolved landscapes in Figure 67A is somewhat reminiscent of the superposition of a random roughness onto the deconvolved energy landscape of the Rossmann fold depicted in Figure 64B. However, in agreement to findings in Section 6.4.3, here, the previously assumed relatively high effective roughness of the Rossmann fold is unmasked to essentially be constituted of a limited number of discrete on-pathway intermediates.

Altogether, here, we showed that in combination with a more detailed analysis of transition paths, a five state energy landscape can be reconstructed from data provided by an otherwise simple two-state analysis. By reaching a much higher level of detail as compared to deconvolution, the herein applied dwell-time-based energy landscape reconstruction technique, which was developed in this work, has proven to be very useful. In addition to the result that the energy landscape roughness of the Rossmann fold is essentially constituted of 3 predominant intermediates, we also found that at 15 pN, a maximal local roughness of  $10 \text{ k}_B\text{T}$  is reached in terms of the highest on-pathway intermediate energy well depth. Notably, the reconstructed energy landscape provides almost the exact same folding free energy of  $\Delta G_0^{N \cdot I_1} \sim 17 \text{ k}_B\text{T}$  for the intermediate I<sub>1</sub> as has been previously found for the corresponding intermediate I of S49Cc from constant distance measurements.

## 6.5 SUMMARY AND OUTLOOK

This final section is divided into two parts. In the first part, previous experimental results are reviewed in the context of comparing the artificial proteins ROSS and FLred to the natural protein R15. This includes two additional features one of which focuses on the folding mechanism of the Rossmann fold and the other one summarizes gained insights into energy landscape rough-

ness in general. The second part is dedicated to answer the question of how and what we can learn from studying transition paths. After a brief summary, prospects of the newly developed method of protein friction determination based on transition path ensemble deformations  $\Delta p(x|TP)$  are discussed in view of a number of other protein systems which have been measured during this thesis. Further, an expansion of the analysis of transition paths from equilibrium towards off-equilibrium data is proposed.

### 6.5.1 *Folding reliability: evolution outperforms design by means of reduced complexity*

While R15 is a highly cooperative and simple two-state folder, ROSS is a sequential multi-state folder with high stability. This statement is based on a multitude of fingerprints which involve off-pathway transitions, a multimodal distribution of unfolding forces as well as on kinked rate plots which were all observed to characterize the mechanics of the artificial Rossmann fold in [Section 6.2](#). By contrast, folding of the natural R15 protein is well described with one distinct fingerprint, an unimodal unfolding force distribution, and rate plots having constant slopes over the entire accessible force range.

[Figure 68A](#) and [Figure 68B](#) present a few key parameters which distinguish the mechanics of R15 and ROSS. With respect to general protein-folding, there are two crucial differences between R15 and ROSS, namely the success rate and the duration of folding into the native (N) configuration. With a 100% success rate and transition path times in the low microsecond regime, R15 is an extremely reliably and fast folding protein. On the other hand, the Rossmann fold is prone to getting trapped in non-native configurations which, in case of 500 nm/s constant velocity cycles, almost make up a total of 30% of all folding transitions, see [Figure 68C](#). In addition to unreliable folding under force, in case of ROSS, the process of folding itself lasts a few milliseconds. Hence, while R15 appears to have a relatively smooth energy landscape, ROSS seems to experience an additional energy landscape roughness which can be quantified by an additional effective roughness of  $\epsilon_{rms} = 2.4 k_B T$  with respect to R15, see [Section 6.4.1](#). Another important consequence of the long transition path times of the Rossmann fold is its slow folding rate at zero force  $k_0^{f,ROSS}$  which only reaches about  $100 s^{-1}$ , see also force range I in [Table 9](#). With  $k_0^{f,R15}$  being  $10.000 s^{-1}$ , R15 folds way more efficiently than ROSS. Note that transition path times of less than 100  $\mu s$  are necessary to achieve folding rates of  $10.000 s^{-1}$  while transition path times of 4 ms only allow folding rates of up to  $250 s^{-1}$  to be achieved.

With respect to reliability, speed, and efficiency, the natural protein R15 clearly outperforms the artificial Rossmann fold. However, ROSS has one advantage over R15 which is given by storing twice the folding free energy per amino acid. In combination with its sequential folding mechanism, the high folding free energy enables ROSS to refold against significantly higher forces and to longer resist against unfolding around its higher mid-force compared to R15, see  $F_{max}^{fold}$  in [Figure 68A](#) and [Figure 68B](#), and see also [Table 9](#). In conclu-

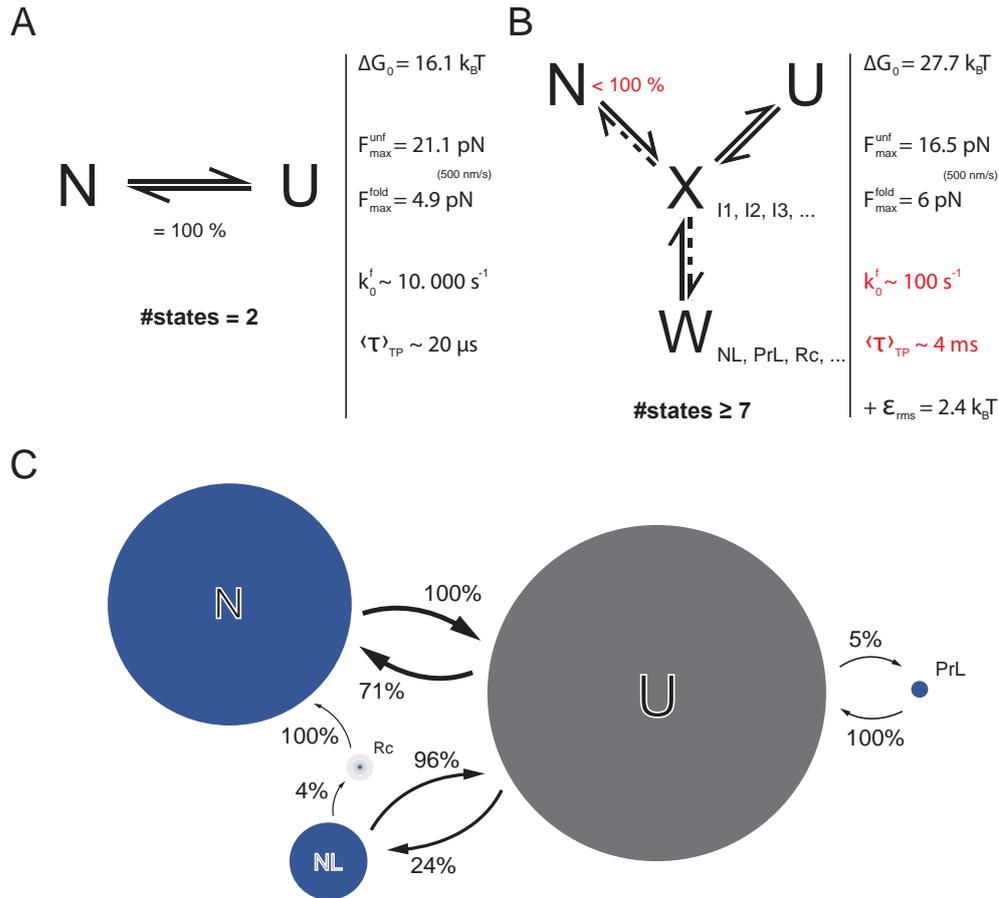


Figure 68: Reaction schemes for R15 and ROSS. (A) The measured mechanical behaviour of R15 corresponds to behaviour expected for an ideal two-state folder. (B) The mechanics of ROSS are much more complex than simple two-state. In addition to obligatory on-pathway intermediates, the rough energy landscape of the artificial Rossmann fold also comprises misfolded states in the style of kinetic traps. (C) Interconnections between observed fingerprints of ROSS. Percentages symbolize the reaction flux based on fingerprint occurrences during 500 nm/s constant velocity cycles.

sion, the comparison between natural and artificial proteins seems to result in having to choose between higher cooperativity which apparently comes along with reliable and efficient folding or higher folding free energies which tend to promote undesirable non-native interactions in parallel to providing more resistant structures. In cells, proteins need to function properly and reliably at all times, they do not typically need to store a lot of energy. Or to put it differently: like happiness does not lie in money alone, the protein-folding problem is not solved by stability alone.

Hence, to guarantee optimal protein functionality in terms of reliable, repeated and fast folding, the folding free energy is not the only parameter that needs attention when designing proteins. This finding is indirectly supported by the fact that a designed destabilization helped Top7 to become a cooperative folder, see [245]. Importantly, this work even provides 'designed' proof

of how a labile design can result in more efficient and reliable folding. This proof is given by the labile re-design of the Ferredoxin-like fold called FLred in [Chapter 5](#). Compared to ROSS, FLred shows much faster folding rates at zero force, i. e.,  $1.000 - 10.000 \text{ s}^{-1}$  (see [Table 4](#)), while not providing any evidence for multiple pathways or misfolding. On the other hand, compared to ROSS, FLred stores less than one third of folding free energy per amino acid.

### *Folding mechanism of ROSS*

According to a detailed investigation of individual folding transitions in [Section 6.4.3](#), native folding of the Rossmann fold involves passing through three predominant on-pathway intermediates I<sub>1</sub>, I<sub>2</sub>, and I<sub>3</sub>. This corroborates the picture of a sequential folding and unfolding process which is further quantified in [Section 6.4.4](#). The mechanical data suggest that unfolding starts from the C-terminus and proceeds towards the N-terminal end which is strongly supported by close to identical structural interpretations and dwell times of I<sub>1</sub> and I<sub>2</sub> with respect to the independently measured obligatory on-pathway intermediates of S49Cc and S85Cn, see [Section 6.3.2](#) and [Section 6.3.3](#). Moreover, intermediates I<sub>1</sub> and I<sub>2</sub> seem to involve fast  $\beta_4\beta_3$  and  $\beta_3\beta_2$  hairpin fluctuations. An essential feature of the overall stability of the Rossmann fold is provided by the enforced key lock mechanism of the C-terminal  $\alpha$ -helix. The enforcement seems to arise from a proline-induced stiffening of the backbone at the N-terminal capping position of this helix, namely at position P86, see [Section 6.3.1](#).

In addition to native (N) folding transition pathways (N/U), the Rossmann fold sometimes 'loses its way' and ends up in non-native configurations as revealed by at least four different typical fingerprints in [Section 6.2.1](#). These additional fingerprints involve at least 3 more possible pathways which can be identified as transitions between NL/U, PrL/U, and NL-Rc/Rc-N in [Figure 68C](#). Concerning the nature of NL (native-like) transitions, a comparison of its contour length gains to those provided by native (N) transitions in [Section 6.3.2.1](#) revealed by  $\Delta L_p^{N/NL} = 0.9 \pm 0.6 \text{ nm}$  shorter lengths for NL. This is direct proof for a different structural arrangement which seems to explicitly involve non-native interactions between the N- and C-terminal halves of ROSS. Since the relative occurrence of Rc (rescue-transition) fingerprints scales with the occurrence of NL, Rc is likely to be a spin-off of NL fingerprints by representing rescue attempts out of NL and into the native configuration, see also [Section 6.3.2](#). The mechanical study of proline-free mutants performed in [Section 6.3.1](#) could neither confirm nor disprove that a proline-switch is the only cause for the occurrence of PrL (proline-like) fingerprints. Since no further insight into an alternative origin has been found, the PrL configuration remains vaguely related to problems of the protein backbone.

An important reason for the relatively high amount of undesirable non-native interactions may be given by the energetic imbalance between the N- and C-terminal halves of the Rossmann fold which was uncovered in [Section 6.3.2](#) and which attests the C-terminal half to store more energy than the

N-terminal half. A tangible result which confirms this hypothesis is given by findings for the proline-free mutants P86A and P86S in [Section 6.3.1](#). Here, a reduction in folding free energy by about  $2 k_B T$  within the C-terminal half of ROSS led to a significant reduction of the occurrence of undesirable non-native interactions represented by NL and PrL, see [Table 12](#). Furthermore, this observation confirmed that the use of proline P86 is a design element which increases the overall mechanical stability of the Rossmann fold by stiffening its backbone. However, the reduction of undesirable non-native interactions upon replacement of the proline, which comes along with a reduction in folding free energy, reconfirms the above statement that the folding free energy is not the only thing which needs to be optimized by design.

If all intermediates as well as their hypothesized  $\beta$ -hairpin fluctuations are considered, solely native folding of ROSS already involves at least 7 states. Furthermore, the four identified different pathways contribute at least another two additional states, namely NL and PrL. With four different pathways and 9 different states, to date, the Rossmann fold probably provides the most complicated folding network which has been characterized with optical tweezers for a protein of this size. In this context, it is of high interest to mention that all  $\alpha\beta$ -proteins including the Rossmann fold which originate from the same design study have one issue in common: while all  $\alpha\beta$ -proteins have well-resolved [NMR](#) structures, none of them does crystallize to provide a crystal structure<sup>27</sup> [[115](#), [199](#)]. This observation could imply that all these proteins do not exist in exactly one configuration<sup>28</sup> which would be absolutely crucial for crystallization. From this perspective, the complicated folding network found for ROSS is the first experimental explanation for the crystallization issues of the designed  $\alpha\beta$ -proteins.

Note that the folding mechanism of the designed 2x2 Rossmann fold has a lot in common with the naturally occurring thermophilic variant of Adenylate kinase (thADK) which has a very similar folding motif [[166](#)]. The two on-pathway intermediates which were identified upon unfolding of thADK are structurally identical to I1 and I2 of the Rossmann fold. Additionally, unfolding of I1 also involves a typical pre-flipping. Unfolding out of the second intermediate appears to be initiated by a traceable hairpin unzipping. Furthermore, there seems to be a hint towards a third, very short-lived on-pathway intermediate which is located very close to the fully unfolded state of thADK. And finally, thADK, which incorporates a number of prolines, shows the same proline-like behaviour as the designed Rossmann fold. Now, in perspective of the folding mechanism of a naturally occurring protein with a similar fold, the folding issues of ROSS can be boiled down to one single problem: NL fingerprints, i. e., the non-native interactions between the N- and C-terminal halves of ROSS.

With respect to its complicated folding network, ROSS is reminiscent of being a miniaturized version of the much bigger Hsp90 protein, a large molec-

<sup>27</sup> By way of personal communication from Po-Ssu Huang after a talk he gave at the Technical University of Munich in September 2017.

<sup>28</sup> The particular reason why additional populations expected for NL and PrL are not 'resolved' by [NMR](#) remains unsettled.

ular machine which has been reported to also show multi-pathway folding and misfolding in single-molecule experiments [109]. For Hsp90, a strongly reduced reaction scheme has been used to separate successful (N) and unsuccessful (W = 'wrong') folding attempts out of the unfolded (U) state according to:  $N \rightleftharpoons U \rightleftharpoons W$ . The similarity to the reduced scheme shown in Figure 68B, where the additional X merely represents the entire interconnecting space within the folding funnel except for states N, W, and U, is obvious. Now, if we consider all non-native interactions of the Rossmann fold to be included in a single population W, then the logarithm of the ratio of the probabilities of being either native or non-native should yield the energy difference between these two hypothetical states which is  $\ln(p^{(N)}/p^{(W)}) \sim 1 k_B T$ . In view of an overall folding free energy of almost  $30 k_B T$  this seems quite small, however, this energy difference of about  $1 k_B T$  has the same order of magnitude which caused a significant reduction of non-native interactions due to the replacement of proline P86, see above.

*Energy landscape roughness and internal protein friction - two sides of the same coin*

What is energy landscape roughness? And what does internal protein friction mean? In literature, roughness is described in different ways. Sometimes it is simply expressed as elevated internal friction due to 'frustration' [40, 238], alternatively, changes in friction are paraphrased as additional effective roughness  $\epsilon_{rms}$  which can typically range from 2 up to  $5 k_B T$  [159, 209], or, distinct local energy minima of up to  $15 k_B T$  are brought into play to characterize a rough energy landscape [148].

All the above ways of describing roughness are also used in this work. While R15 was found to have an internal friction coefficient of  $\gamma_p^{R15} = 9.7 \cdot 10^{-7} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ , ROSS had a coefficient of  $\gamma_p^{ROSS} = 2.6 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . Hence, ROSS has an elevated internal friction that can be expressed as an additional energy landscape roughness of about  $2.4 k_B T$  with respect to R15, see Section 6.4.1. A detailed analysis of the transition paths of the Rossmann fold revealed that the actual reason for its apparently elevated friction are three distinct predominant on-pathway intermediates which reach a distinct local roughness of up to  $10 k_B T$ , see Section 6.4.3 and Section 6.4.4. This observation is a key finding which directly relates elevated internal friction to distinct minima in an energy landscape. Such an understanding of friction can also be found in recently published simulation studies [253].

Friction  $\gamma$  is linked to diffusion D by temperature according to  $\gamma = k_B T / D$ . Similar to the speed of light which sets a limit to the propagation speed of light in space, diffusion sets a limit to the speed of protein folding. The friction which corresponds to the experimentally determined diffusion limit of cytochrome c (CyC) equals  $\gamma_p^{CyC} = 8.8 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  at a temperature of 298 K [92]. Compared to this value, the friction measured for R15 is about one order of magnitude higher, see above paragraph. From this perspective, one can deduce that there is still a remaining ruggedness present which is 'invisible' to our experiments. In this context, when assuming that the internal

friction coefficient of ROSS was that of cytochrome c instead of FLred as being assumed in [Section 6.4.4](#), then the maximal local roughness which was found for ROSS would even reach up to  $15 k_B T$ .

In summary, friction or effective roughness are qualitative measures of quantitative effects inferred by distinct energy landscape roughness. Whenever experiments cannot resolve its origin, speaking of elevated internal friction serves as a circumvention to describe roughness. In this context, effective roughness can be considered as a theoretical construct which translates apparent differences in friction into the energy landscape perspective. In other words, effective roughness is an intuitively clear illustration of something which has not yet been understood in detail. A consequence of our better understanding of energy landscape roughness is that experimentally determined friction coefficients should be considered as upper limit estimates as there always remains the possibility of experimentally unresolved ruggedness.

#### *An outlook towards more data, theories, design, and mutants*

Most constant velocity measurements of the Rossmann fold, its mutants, and different pulling geometries were performed at a speed of 500 nm/s. By screening through different speeds and looking for changes within the relative fingerprint occurrences one might learn more about the hierarchy of folding and unfolding events or force-induced pathway switches. In particular, a loading rate dependent study of the essentially bi-modal unfolding force distribution of ROSS seems to be an ideal occasion to test a theory which was recently developed to distinguish signatures of multipathway conformational transitions [170]. Following experiments which were performed on Hsp90, one could test the effect of additional waiting times at zero or very low force on the success rate of folding into the native state of ROSS [109].

Since the designed Rossmann fold showed many similarities to the naturally occurring thADK with a very similar fold, see above, it seems intriguing to compare the ideal two-state mechanics of R15 to a designed three-helix bundle like, e. g.,  $\alpha_3D$  [233]. Furthermore, the mechanical characterization of any re-design of a naturally occurring protein which has already been mechanically characterized like, e. g., calmodulin [216], can lead to a lot of new insights on folding mechanisms, design rules, the impact on a protein's functionality, etc.

Proline-free versions of S49Cc, the construct for directed C-terminal unfolding in [Section 6.3.2](#), are likely to provide more precise results on the effect of the P86A and P86S mutation on the folding mechanism of ROSS particularly with regard to the folding free energy of its first predominant on-pathway intermediate I1. Besides, the 'simple' three-state mechanics which were observed for S49Cc may serve as an experimental reference for the effect of transition state switches on overall transition rates which is discussed on the basis of one-dimensional diffusion simulations in [Section 4.2.4](#).

A new S9Cc construct may help to figure out whether the N-terminal  $\beta$ -sheet is truly not involved in the formation of unfolding intermediate I<sub>2</sub>.

Sometimes, truncation mutants can help to find out whether a substructure of a protein is stable enough to keep up a folded structure on its own. Potential candidates for truncation mutants of ROSS could be its N- and C-terminal halves for which there are two reasons. First, measurements on S49Cc revealed that after unfolding of the C-terminal half, a residual energy of  $7 k_B T$  still resided in the N-terminal half, see [Section 6.3.2](#). And second, upon full unfolding, intermediate I<sub>2</sub> represents a fully folded C-terminal half, see [Section 6.4.3](#). However, it should be noted that already the attempt of measuring a less radical truncation mutant, which had only the C-terminal  $\alpha$ -helix deleted, unfortunately failed<sup>29</sup>.

Circular permutations are a very promising approach to address a couple of questions related to the folding mechanism of the Rossmann fold. For such permutations, the proximity of the N- and C-terminal ends of ROSS proves beneficial. To illuminate interactions within the N-terminal half of ROSS which comprises the very short-lived intermediate I<sub>3</sub>, an interesting circular permutation would be given by closing the N- and C-termini with a loop and opening another loop right after position S49. This would yield a new variant of the S49Cn construct where unfolding is not restricted to start by pulling the N-terminal  $\beta$ -sheet out of the hydrophobic core. Another interesting circular permutation could address the enforced key lock mechanism and the bending issue of the C-terminal  $\alpha$ -helix around position P86. Therefore, again, N- and C-termini are closed by inserting an additional loop and instead, the loop next to P86 is opened.

### 6.5.2 Transition paths - high potential with imposed limits

In force spectroscopic single-molecule trajectories, transition paths enclose temporal and spatial information. These are provided by their transition path times and the shape of their transition path ensembles. The direct evaluation of individual transitions provides useful information to derive rough time scale estimates of the involved protein folding. Here, estimated average transition path times were  $\langle \tau_{TP}^{R15} \rangle \sim 20 \mu s$  for the natural protein R15 and  $\langle \tau_{TP}^{ROSS} \rangle \sim 3.9 ms$  for the artificial Rossmann fold, see [Section 6.2.3](#). In case of time scales being similar to the response time of the measurement system, special care has to be taken to account for mixing between actual protein folding and its detection, see [Section 4.3.3.2](#). In this work, both transition path times and ensembles were used to derive the measured protein's internal friction  $\gamma_P$ , see [Section 6.4.1](#) and [Section 6.4.2](#). Furthermore, based on the information provided by transition paths, a newly developed approach enabled the reconstruction of a protein-folding energy landscape with a much higher level of detail as compared to deconvolution which itself is one of the most advanced reconstruction techniques at hand, see [Figure 67A](#) in [Section 6.4.4](#).

<sup>29</sup> Besides smeared out multiple peaks during SEC-runs, no interpretable fingerprints different from randomly arranged polypeptide were seen during trap measurements. Data not shown.

In agreement with results obtained from simulations in [Chapter 4](#), transition path probability calculations, as described in [Section 3.8.3](#), were found to be incapable of providing correct information on transition state positions or reaction coordinate quality without additional information provided by, e. g., deconvolution, see [Section 6.2.5](#). Concerning committor-based barrier shape reconstructions described in [Section 3.8.2](#) and performed in [Section 6.2.4](#), barrier heights are reasonable with respect to deconvolution while their shapes tend to be more resemblant to the measured bead potential which also has been similarly observed in simulations, see in particular [Section 4.3.4](#). Essentially, the observed limitations of the predictive capabilities of different transition path analysis tools are imposed by our measurement procedure. As elaborated in [Section 4.3.2](#), the projection of a multidimensional folding process into one dimension is additionally distorted by differences in friction with respect to protein and bead diffusion.

Similar to transition path times and ensembles which are altered by the additional presence of protein diffusion, the shape of committors can be used to detect the presence of protein folding by comparison to the shape expected for unperturbed bead diffusion. Furthermore, the slopes of committors indicate whether the involved protein friction effectively lies below or above the friction inherent to the detection system. This way it was found that both R15 and ROSS are not folding at the speed limit which directly follows from the strong deviation of their commitment probabilities from the shape expected for mere bead diffusion, see [Figure 55A](#) and [Figure 56A](#) in [Section 6.2.4](#). Further, R15 effectively experiences less while ROSS effectively experiences higher friction than the beads. In combination, this yielded  $4.8 \cdot 10^{-8} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1} < \gamma_{\text{P}}^{\text{R15}} < 1.5 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  for the friction coefficient of R15 and  $\gamma_{\text{P}}^{\text{ROSS}} > 3 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  for ROSS. In case of ROSS, the indicated values only hold upon assuming two-state folding.

#### *From landscape reconstruction towards friction maps*

In total, three different ways of protein friction coefficient determination were used in this thesis. A typical way to derive  $\gamma_{\text{P}}$  is based on solving Kramers rate equation for  $\gamma$ , see [Equation 29](#). One way to implement this approach is given by using zero-force extrapolated folding rate constants and the shape of their corresponding deconvolved energy landscape which provides necessary curvatures at critical barrier positions, i. e.,  $\gamma_{\text{P}}$  is given as  $\gamma_{\text{P}}(k_0^f, G_0(L_{\text{P}}))$ . This derivation was applied for the designed protein FLred in [Section 5.3.1](#) of [Chapter 5](#) and yielded  $\gamma_{\text{P}}^{\text{FLred}} = 5.1 \cdot 10^{-6} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The two other approaches which were used in [Chapter 6](#) to determine  $\gamma_{\text{P}}$  are based on information provided from transition paths. In [Section 6.4.1](#), average transition path times  $\langle \tau_{\text{TP}} \rangle$  were used in connection with their corresponding deconvolved energy landscapes and yielded values  $\gamma_{\text{P}}(\tau_{\text{TP}}, G(L_{\text{P}}))$  after solving [Equation 54](#) for  $\gamma = k_{\text{B}}T/D$ . Together with the corresponding deconvolved energy landscape shapes, measured transition path times of R15 and ROSS imply  $\gamma_{\text{P}}^{\text{R15}} = 9.7 \cdot 10^{-7} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  and  $\gamma_{\text{P}}^{\text{ROSS}} = 2.6 \cdot 10^{-4} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$ . The third

approach which was newly developed in this thesis to derive protein friction based on a minimal (implicit) mechanical model is introduced and applied in [Section 6.4.2](#). Here, the measurable friction-induced deformation of transition path ensembles is used in combination with corresponding two-dimensional Langevin Dynamics simulations to derive the inherent protein friction according to  $\gamma_P(\Delta p(x|TP))$ , see also [Equation 68](#). Application to data provided by the measurements on R15 yielded  $\gamma_P^{R15}(\Delta p) \sim 1 \cdot 10^{-6 \pm 1.5} \text{ pN} \cdot \text{s} \cdot \text{nm}^{-1}$  which is in excellent agreement with the result obtained from transition path times.

Both transition-path-based protein friction determinations have certain advantages over the determination based on extrapolated rates. While transition path times are experimentally directly accessible, zero-force folding rate constants strongly depend on the model which is used for their extrapolation. Furthermore, while transition path times only weakly depend on experimentally determined barrier heights, see [Figure 19A](#), Kramers rate equation involves an exponential barrier height dependence which can lead to much larger errors. A particular advantage of the newly developed protein friction determination technique which is based on friction-induced transition path ensemble deformations  $\Delta p(x|TP)$  is that already a relatively small number of transition events is enough to get meaningful protein friction coefficient estimates. Here, a single constant distance trajectory with a few transitions is already enough to find the right order of magnitude in protein friction. By contrast, multiple trajectories with many transitions over a large enough force range are necessary to be able to reasonably model rates. Owing to their asymmetric distribution, see [Figure 19B](#), transition path times also need enough sampling to yield a good average value.

The theoretical possibility of extracting protein friction coefficients from shapes of dominant transition paths, which, in our case, are represented by transition path ensembles, can also be found in literature [[137](#)]. After seeing remarkable differences among diffusion constants which were derived from transition path times of different DNA-hairpins, experimentalists foresee a possibility to measure position-dependent diffusion constants in the near future [[158](#)]. Here, a way to directly extract position-dependent information on the diffusion constant, i. e., the inverse friction coefficient, has already been proposed to be given by the stepwise evaluation of [Equation 68](#) along measured transition path ensembles. Strong deviations from corresponding simulations which assume constant friction and, by this means, a fixed diffusion constant, directly indicate corresponding changes in  $\gamma$  and  $D$ , respectively; for an example see [Figure 65D](#). Further, stretches of steep and shallow ascent of the incremental evaluation of [Equation 68](#) along measured and simulated transition path ensembles shown in all graphs at the bottom row of [Figure 65](#) can be related to energy landscape features such as transition state positions or inflection points.

In the future, a possible proof of principle measurement, which could also help to fully establish the herein introduced transition-path-ensemble-based friction determination, is the measurement of a 'speed limit' or low-friction reference. Such a reference would be given by a protein which folds so fast

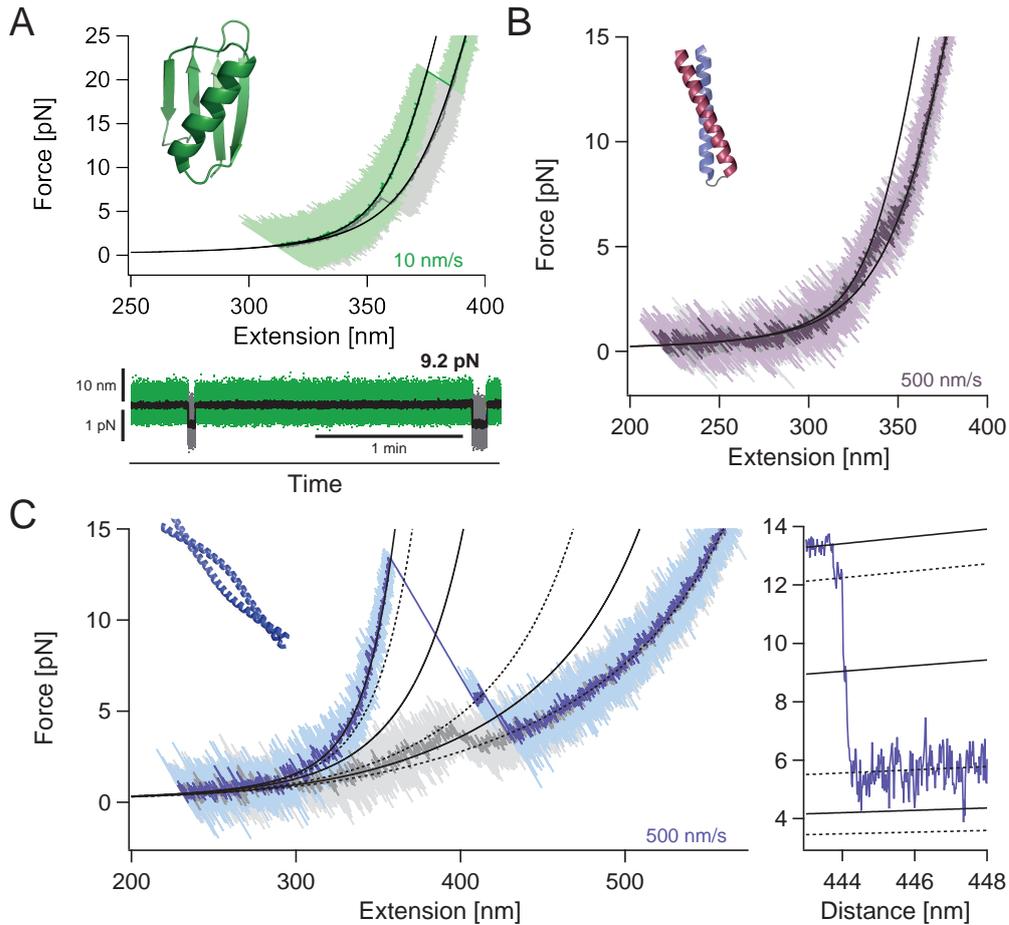


Figure 69: A ‘speed limit’ reference and candidates for sequence-resolved friction maps. (A) GB1: a potential ‘speed limit’ reference. (B) Anti-parallel heterodimeric coiled-coil: U2LZ shows fast folding/unfolding transitions around 6 pN. (C) Parallel homodimeric coiled-coil from *Drosophila melanogaster*: NS2n shows neck, stalk 1, and stalk 2 unfolding transitions. The force-distance relation on the right is a zoom into the high-resolution force-extension relation from the left (light blue).

that it yields transition path ensembles which are almost identical to what would be expected for unperturbed bead diffusion. A candidate for such a reference may already be given by GB1 which, during this thesis, was found to undergo folding transitions at constant distance measurements around 10 pN, see Figure 69A. Further, statistics on transitions can be easily increased by force-jump experiments to higher forces which simultaneously increases resolution due to correspondingly larger changes in the deflection signal. Importantly, from FRET measurements it is already known that the transition path time of GB1 can be expected to be shorter than 10  $\mu$ s [45].

While measuring absolute values for internal protein friction might always be a bit tricky due to experimental limitations, relative changes in friction among different proteins of similar size and/or structure can be readily assessed by calculating their  $\gamma_P$  ( $\Delta p(x|TP)$ ) values. In this context, another class

of proteins which has also been studied during this thesis might be of particular interest: coiled coils. The main graph in [Figure 69C](#) depicts a constant velocity cycle of a 834 amino acid<sup>30</sup>, homodimeric coiled coil from the *Drosophila melanogaster* Kinesin(-1) Heavy Chain. This NS2n construct comprises the neck (N), stalk 1 (S1), and stalk 2 (S2) domains and is being unzipped from the N-terminal (n) end, see [Section A.2](#) for more details. The six WLC-fits mark five distinct transitions starting from the neck and proceeding towards the second stalk domain. Relative to each other, all these transitions can provide a certain amount of friction which is involved in their unfolding and refolding. Due to the simple coiled coil structure, it is also easy to isolate all these transitions within correspondingly shorter unzipping constructs for their detailed study. The results from such studies may then be used to build sequence resolved friction maps as a complement for sequence resolved stability maps. Does friction go hand in hand with stability? Is friction unaltered by stability? Does friction depend on the direction of unzipping? These and many more questions may be addressed in the future. Along these lines, another question could be: is there a fundamental difference in friction between parallel homodimeric and antiparallel heterodimeric coiled coils? This could be answered by studying the antiparallel heterodimeric U2LZ construct which shows fast folding/unfolding transitions around 6 pN and, hence, should yield good statistics on short time scales, see [Figure 69B](#). Note that the reaction coordinate for coiled coil unzipping is typically way more 'direct' than for forced globular protein unfolding. This way, coiled coils provide an ideal model system to further fathom the potential which is concealed within transition paths.

#### *Transition path analysis beyond equilibrium*

In this work, transition path analysis mainly focused on evaluating equilibrium trajectories from constant distance (CD) measurements. Due to the relatively slow equilibrium rates of both R15 and ROSS, the accumulation of measured equilibrium transitions for good enough statistics was very time consuming. However, these measurements were necessary to prove that measured unfolding and refolding pathways provide the same information. A very efficient way to increase statistics is by using off-equilibrium techniques which are given by constant velocity (CV) measurements or force jump experiments [[189](#), [202](#)].

The high potential which lies within gathering information from off-equilibrium measurements is illustrated in [Figure 70A](#). The upper panel shows a scatter plot of force versus transition path times which were extracted from repeated stretch-and-relax cycles of one single molecule and the lower panel shows corresponding data which were pooled from all available CD trajectories of all molecules. While the CV measurement yields 532 data points within 10 minutes, data which were collected from many CD trajectories include 201

<sup>30</sup> To our knowledge, this is the longest coiled coil construct which, up to date, has ever been unzipped at once.

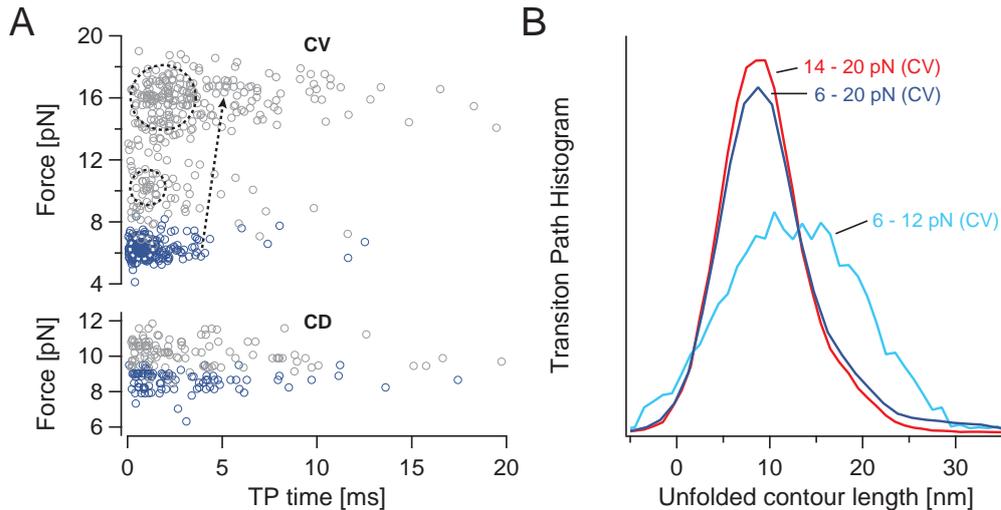


Figure 70: Force-dependent transition path times, histograms, and ensembles. (A) Scatter plots of force versus transition path times. The upper panel displays transition path times extracted from unfolding and refolding transitions (grey and dark blue empty circles) of an individual molecule studied in 500 nm/s constant velocity (CV) mode. Below, transition path times from constant distance (CD) trajectories are pooled together from all molecules. (B) Transition path histograms from CV measurements clearly provide spatial information that changes with force.

data points from measurements lasting for hours. Note that indicated forces refer to the force bias which acts on the transition's initial state. When extracting transition paths from CV measurements (as opposed to direct force jump experiments), a point of concern might be the constantly changing inter-trap distance. For transition paths which last only a few milliseconds, this additional change in force bias during the actual transition event is almost negligible, see also Figure 12. In fact, during stretch-and-relax cycles, the direction of movement of the traps typically compensates for the non-constant force bias which is inferred by changes in unfolded contour length.

As highlighted by dashed circles in the upper panel of Figure 70A, the unfolding and refolding force distributions of the Rossmann fold can be found again in the scatter plot. Furthermore, the dashed arrow suggests a slight trend towards longer transition path times with increasing force. While this trend might be surprising on first sight, it can be directly explained with the fact that intermediate  $I_1$ , which has the highest impact on transition path time durations of ROSS, reaches its highest energy well depth not until 15 pN, see also Figure 67A. This correlation can also be seen in Figure 70B which shows normalized histograms that are composed of transition paths from CV measurements separated by force. Clearly, in Figure 70B, force-dependent differences in resolution and individual intermediate state contributions are well noticeable and reveal a wealth of information.

When restricting transition path histograms from CV measurements to within a very narrow force range, these histograms automatically become

ensembles. In view of the possibility of measuring force-dependent transition path times and ensembles, also force dependent transition path analysis might be put into practice. This could theoretically start with estimating force-dependent transition state movement and end at a fully transition-path-assisted energy landscape reconstruction without the need for additional information like rates. Abrupt force-dependent changes of transition path ensemble shapes could indicate a change in predominant reaction pathways. Furthermore, a potential force-dependence of internal protein friction could be investigated to find out whether force might be one of the things which cause elevated friction.

In view of the results presented in this work, an appropriate evaluation of single-molecule protein-folding transition paths provides very valuable information which can, for example, be used for energy landscape reconstruction or internal protein friction determination. Hence, the analysis of transition paths measured in as well as out of equilibrium represents a promising way of learning more about the 'secrets' of protein-folding.

## Part V

### APPENDIX

Protein sequences, protocols, and troubleshooting.



## MATERIALS AND METHODS

---

### A.1 PROTEIN SEQUENCES

If not stated differently, all listed proteins were expressed in *Escherichia coli* (*E.coli*) using a pET-28a(+) vector. Details of applied molecular cloning techniques and protocols for protein expression, protein purification and attachment of oligos are given in [Section A.2](#) and [Section A.3](#). All necessary protein parameters, such as their molecular weight or extinction coefficients, were calculated using the Protparam tool which can be found under <http://web.expasy.org/protparam/>.

For an easier reading of the sequences the following rules have been applied. The sequence of the protein which is meant to be unfolded has bold, black **letters**. Mutations are indicated by bold, blue **letters**. Parts serving as unstructured spacers and/or originating from restriction sites are **gray**. The cysteines (**C**) for oligo attachments are red. The rest in normal, black 'letters' has a certain function, i. e., the first methionine (M) to start expression, lysins (K) next to red cysteines to create 'hot-cysteines' for better oligo attachment, the His-Tag comprising 6 histidines for purification, as well as additional entire proteins. Among the latter are ubiquitin, GB1 and fluorophores. The first two serve as additional spacers to spatially separate oligo attachment points, to especially prevent crosslinking of reactive cysteines within a construct.

#### A.1.1 *Ferredoxin-like fold*

The mechanically extremely labile Ferredoxin-like fold (FLred) used to engineer a highly elastic and tough protein-based hydrogel [76] is a re-design of the *de novo* designed ideal Ferredoxin-like fold (FL) [115]. The sequence of the latter is taken from the published solution NMR structure (pdbID: 2KL8). Including the first methionine, the core sequence of FL comprises 77 amino acids. A sequence alignment with FLred reveals 16 mutations which mechanically destabilize the fold, namely: M<sub>3</sub>F, G<sub>9</sub>T, L<sub>12</sub>D, A<sub>14</sub>Q, A<sub>18</sub>V, I<sub>23</sub>N, Q<sub>25</sub>R, F<sub>29</sub>D, L<sub>37</sub>R, L<sub>42</sub>F, V<sub>49</sub>L, P<sub>50</sub>S<sup>1</sup>, V<sub>53</sub>N, A<sub>61</sub>V, F<sub>68</sub>Q and I<sub>76</sub>E. Underlined amino acids of FL and blue coloured amino acids of FLred indicate the mutations within the given sequences. In brief, the mutations reflect a weakening of the hydrophobic effect and a potential increase of solvent contact area, both leading to destabilization. This can be deduced from the fact that almost all mutations replace hydrophobic side-chains with bulkier, polar or even charged side-chains. The mutations are located especially at transitions between secondary structural motifs and surface exposed loops, see [Figure 39](#).

---

<sup>1</sup> Structurally this mutation very much resembles the P86S mutation in the proline-free version of the Rossmann fold where it was shown to indeed mechanically destabilize the protein.

The sequence of FL is only given for comparison. FL was designed and expressed at the University of Washington in Prof. Baker's group with the sequence having been cloned into plasmid pET-29b by GenScript. It has not been designed and expressed as a construct suitable for mechanical studies with optical tweezers yet.

FL - STABLE DESIGN. MEMDIRFRGDDLEAFEKALKEMIRQARKFAG  
TVTYTLDGNDLEIRITGVPEQVRKELAKEAERLAKEFNITVYTIIRLEHH  
HHHH

The FLred sequence features a design for N- to C-terminal mechanical unfolding/folding measurements with optical tweezers. For an easier thiol oligo attachment, incorporated cysteines are spatially separated via an additional GB1 at the N-terminus. The construct was designed and cloned into the expression vector pQE-8oL, expressed, purified, and lyophilized at the University of British Columbia in Prof. Li's group (UBC).

FLRED - LABILE DESIGN. MRGSHHHHHHGS<sup>C</sup>MTYKLILNGKTLKGETT  
TEAVDAATAEKVFKQYANDNGVDGEWYDDATKTFTVTERSMG<sup>E</sup>FDIR  
FR<sup>T</sup>DD<sup>D</sup>E<sup>Q</sup>FEK<sup>V</sup>LKEM<sup>N</sup>RRARK<sup>D</sup>AGTVTY<sup>R</sup>RDGND<sup>F</sup>EIRITG<sup>I</sup>SEQ<sup>N</sup>  
RKELAKE<sup>V</sup>ERLAKE<sup>Q</sup>NITVY<sup>T</sup>ER<sup>G</sup>SLERSK<sup>C</sup>

#### A.1.2 Rossmann fold

The investigated Rossmann fold is also a *de novo* designed ideal protein [115]; (pdbID: 2LV8). As sequence similarity suggests, for the first trap construct of the Rossmann fold the same design approach as for FLred was used.

ROSS-GB1. MRGSHHHHHHGS<sup>C</sup>MTYKLILNGKTLKGETTTEAVDAATA  
EKVFKQYANDNGVDGEWYDDATKTFTVTERSM<sup>L</sup>LYVLIISNDK<sup>K</sup>LIEE  
ARKMAEKANLELRTVKTEDELKKYLEEFRKESQ<sup>N</sup>IKVLILVSNDEELDK  
AKELAQKMEIDVRTRKVTSPDEAKRWIKEFSEEG<sup>G</sup>SLERSK<sup>C</sup>

A previous mechanical characterization of GB1 with AFM yielded average unfolding forces above 150 pN [35]. At first glance this suggests an even more stable behaviour than for ubiquitin, which also had been characterized using an AFM showing a little bit smaller unfolding forces around 150 pN [195]. Nevertheless, in our optical tweezers experiments, GB1 already unfolds within tens of seconds at a constant force-load of merely 10pN (data not shown). As this is not the case for ubiquitin [83, 108, 188, 216, 261], new constructs were designed using ubiquitins instead of GB1 (constructs not shown). Finally, the use of maleimide oligos made it possible to use TCEP during the attachment reaction. Hence, a pure Rossmann fold with cysteines directly at its contiguous N- and C-termini was provided by UBC and named ROSSorg.

ROSSORG. **MAC**KMLLYVLIISNDK**KLIEE**ARKMAEKANLELRTVK**TED**  
**EL**KKYLEEFRKESQ**NIK**VLILVSNDEELDKAKELA**Q**KMEIDVRTRK**VTS**  
**P**DEAKRWIK**E**FSEEG**G**SK**CL**EH**HHHHH**

The relatively low yield of the oligo-attachments and the very short spacer sequences between the reactive cysteines and the protein motivated an optimized Rossmann fold design called ROSSopt. Besides slightly longer spacers, the C-terminal negatively charged glutamic acid (E) of the *XhoI* restriction site of the expression vector, was removed. In fact, an improved attachment efficiency with respect to ROSSorg was attained: the ratio of proteins with two attached oligos versus one oligo increased significantly from 3:4 to almost 2:1, see Figure 77 in Section A.3.3. This indicates that enough space between the reactive cysteines indeed is an issue not only for undesirable inter-crosslinking. ROSSopt and all following Rossmann fold constructs were designed and expressed by the author. Genes, which were cloned into plasmid pET-28a(+), were obtained from GenScript.

ROSSOPT. **MAC**KGSM**LLY**VLIISNDK**KLIEE**ARKMAEKANLELRTVK**T**  
**E**DELKKYLEEFRKESQ**NIK**VLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**  
**T**SPDEAKRWIK**E**FSEEG**G**SSG**KCL**HH**HHHH**

#### *Proline-free Mutants*

The Rossmann fold showed an additional population of unfolding forces around 5 pN very similar to the additional population found for filamin A caused by a so called proline-switch [189]. Hence, we suspected Proline P86 to be the reason for our observation as it is the only proline within the entire 100 amino acid long sequence of the Rossmann fold. While in filamin A the responsible proline P2225 is located within a loop region between 2  $\beta$ -sheets, P86 sits right at the beginning of an  $\alpha$ -helix. To test the proline-switch hypothesis, two proline-free Rossmann fold mutants were designed, namely P86A and P86S.

The alanine in P86A was chosen by following the sequence design protocol of the Rossmann fold itself [115]. This protocol states that proline is only allowed in loops and at the beginning of helices and strands, i. e., the beginning of a helix in our case. Even though glycine (G) seems to be a good candidate to replace proline due to the similar small size, it is only allowed for loops and therefore glycine was omitted. Next in size comes alanine (A), which is allowed for helices and loops in the core and at the boundary. Due to its high helix propensity, this made alanine the first choice [165].

P86A. **MAC**KGSM**LLY**VLIISNDK**KLIEE**ARKMAEKANLELRTVK**T**ED**E**  
**L**KKYLEEFRKESQ**NIK**VLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**TS**A**  
**D**EAKRWIK**E**FSEEG**G**SSG**KCL**EH**HHHHH**

Nevertheless, while being slightly solvent-exposed, P86 is not located at a middle position, but right at the N-terminal end of the helix (N-capping posi-

tion), such that alanin may not be an optimal replacement after all. Therefore, serine (S), the next in size, was considered in P86S, as it has polar uncharged side chains, is allowed at the surface according to the sequence design protocol, and is favourable at a N-capping position [63]. Furthermore, P86S resembles the P50S mutation made in the re-design of the Ferredoxin-like fold, where P50 is located at the beginning of an  $\alpha$ -helix as well.

P86S. **MAC**KGSMMLLYVLIISNDK**KL**IEEARKMAEKANLELRTVKTEDE  
LKKYLEEFRKESQ**NI**KVLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**TSS  
DEAKRWIK**E**FSEEGGSSG**K**CL**E**HHHHHHH

#### *Directed unfolding mutants*

To gain deeper insight into the unfolding pathway of the Rossmann fold, further mutants with altered pulling geometries were designed. As the 2x2 Rossmann fold has a very symmetric  $\alpha$ - $\beta$ -repeat structure, it is obvious to pull on either the N- or the C-terminal half. This was realized with the S49Cn and S49Cc constructs, where 'n' indicates unfolding of the N-terminal half and 'c' indicates unfolding of the C-terminal half. As the serine S49 is located at a surface exposed loop position right in the middle of the Rossmann fold it is the ideal candidate to be mutated into a structurally almost identical cysteine for oligo attachment. Besides the S49C mutation, the entire sequence is kept exactly the same as for the ROSSopt construct with two exceptions: the unnecessary cysteine is mutated into a serine and the C-terminal glutamic acid is not left out (as it has not been for any Rossmann fold mutant).

S49CN. **MAC**KGSMMLLYVLIISNDK**KL**IEEARKMAEKANLELRTVKTEDE  
ELKKYLEEFRKE**C**Q**NI**KVLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**TSS  
PDEAKRWIK**E**FSEEGGSSG**K**SLEHHHHHHH

S49CC. **MA**SKGSMMLLYVLIISNDK**KL**IEEARKMAEKANLELRTVKTEDE  
ELKKYLEEFRKE**C**Q**NI**KVLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**TSS  
PDEAKRWIK**E**FSEEGGSSG**K**CL**E**HHHHHHH

As the C-terminal  $\alpha$ -helix plays a crucial role in unfolding, a third directed unfolding mutant has been designed, where the entire protein without this helix is being unfolded. This is realized by pulling at the N-terminus and the surface exposed loop position S85 in the S85Cn mutant.

S85CN. **MAC**KGSMMLLYVLIISNDK**KL**IEEARKMAEKANLELRTVKTEDE  
ELKKYLEEFRKESQ**NI**KVLILVSNDEELDKAKELA**Q**KMEIDVRTRK**V**T**C**  
PDEAKRWIK**E**FSEEGGSSG**K**SLEHHHHHHH

#### A.1.3 R15

The three-helix bundle protein R15 is part of the natural cytoskeletal protein spectrin (pdbID: 1U4Q). The trap construct comprises the amino acids 1658-

1771 of spectrin, which include 4 N- and C-terminal amino acids from the neighbouring spectrin domains to provide a good folding environment. The construct was provided by Thomas Suren with oligos already attached, for details see [221].

R15. MA**C**KELGGKLKEANKQQNFNTGIKDFDFWLSEVEALLASEDY  
GKDLASVNNLLKKHQLEADISAHEDRLKDLNSQADSLMTSSAFDTS  
QVKDKRETINGRFQRIKSMMAARRAKLNESHRLSHRLGGTK**C**LEHHH  
HHH

#### A.1.4 GB1

To prove that GB1 causes additional unfolding events in the GB1 containing constructs FLred and ROSS-GB1, first a GB1-tetramer (GB1-4) was designed (pdbID: 3GB1). All GB1 constructs were provided by UBC.

GB1-4. MRGSHHHHHHGS**C**MTYKLILNGKTLKGETTTEAVDAATAE  
KVFKQYANDNGVDGEWYDDATKTFTVTERSMTYKLILNGKTLKGET  
TTEAVDAATAEKVFKQYANDNGVDGEWYDDATKTFTVTERSMTYK  
LILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWYDDATK  
TFTVTERSMTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGV  
DGEWYDDATKTFTVTERS**C**

Next, a single GB1 construct flanked by two ubiquitins (pdbID: 1UBQ) was designed. Note that an additional aspartic acid (D) was inserted right after the first methionine of GB1.

GB1-2UBI. MA**C**KMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPP  
DQQLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGELMDTYKLILNG  
KTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWYDDATKTFTVT  
EGTMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQLIFAGKQ  
LEDGRTLSDYNIQKESTLHLVLRRLRGGK**C**LEHHHHHHH

Finally, a pure GB1 construct was designed to avoid additional ubiquitin unfoldings.

GB1. MA**C**KMDTYKLILNGKTLKGETTTEAVDAATAEKVFKQYAND  
NGVDGEWYDDATKTFTVTEK**C**LEHHHHHHH

#### A.1.5 Tension sensors: Ypet-HP35(st)-mCherry

The *in vivo* tension sensors (TS) consist of three proteins which comprise a FRET-pair connected via the C-terminal subdomain of the actin-binding protein villin as force-sensing element. The subdomain, also called villin headpiece (HP), includes the amino acids 42 to 76 of villin (pdbID: 1YU5), i. e., 35 amino acids, and is being called HP35. The N-terminal fluorophore

called Ypet is an evolutionary optimization of a previously used already optimized variant of Yellow Fluorescent Protein (YFP) called Venus (pdbID: 1MYW) [88, 152]. The mutations of Ypet with respect to Venus are: I47L<sup>2</sup>, L68V, S208F, and V224L [160]. The last 11 amino acids of Ypet comprise an unstructured region and were omitted to increase fluorophore proximity within the fusion protein. Two native cysteines of Ypet (marked in green) were not removed as they are not expected to interfere with the oligo attachment – C71<sup>3</sup> is buried inside the core, while the potentially reactive group of C49 points inside the protein. The C-terminal fluorophore is mCherry (pdbID: 2H5Q). Two single GS-linkers connect the fluorophores with HP35. The entire construct is called HP35-TS.

The genes encoding the sensors were cloned into pLPCX plasmid and expressed using HEK293 cells. Design and expression were performed in Dr. Grashoff's group at the Max Planck Institute of Biochemistry (MPIB).

HP35-TS. **MA**C**K**MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGD  
 ATYGKLT**LK****L****C**TTGKLPVPWPTLVTTLG**Y****G****V****Q****C**FARYPDHMKQHDFFK  
 SAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGN  
 ILGHKLEYNYNshNVYITADKQKNGIKANFKIRHNIEDGGVQLADHYQQ  
 NTPIGDGPVLLPDNHYSYQSAL**F**KDPNEKRDHMLLE**F****L****T**AAG**S****L****S****D****E**  
**F****K****A****V****F****G****M****T****R****S****A****F****A****N****L****P****L****W****K****Q****Q****N****L****K****K****E****K****G****L****F****G****S****M****V****S****K****G****E****E****D****N****M****A****I****I****K****E****F**  
 MRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDI  
 LSPQFMYGSKAYVKHPADIPDYLLKLSFPEGFKWERVMNFEDGGVVTVTQ  
 DSSLQDGEFIYKVKLRGTNFPDGPVMQKKTMGWEASSERMYPEDGALK  
 GEIKQRLKLDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNED  
 YTIVEQYERAEGRHSTGGMDELY**K****C**LEHHHHHH

By two point mutations deduced from rational protein design, namely N27A and K29M<sup>4</sup>, HP becomes mechanically more stable (st) [22]. Hence, a different force range can be probed with HP35st-TS.

HP35st-TS. **MA**C**K**MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGE  
 GDATYGKLT**LK****L****C**TTGKLPVPWPTLVTTLG**Y****G****V****Q****C**FARYPDHMKQHDF  
 FKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDG  
 NILGHKLEYNYNshNVYITADKQKNGIKANFKIRHNIEDGGVQLADHYQ  
 QNTPIGDGPVLLPDNHYSYQSAL**F**KDPNEKRDHMLLE**F****L****T**AAG**S****L****S****D****E**  
**D****F****K****A****V****F****G****M****T****R****S****A****F****A****N****L****P****L****W****K****Q****Q****A****L****M****K****E****K****G****L****F****G****S****M****V****S****K****G****E****E****D****N****M****A****I****I****K****E****F**  
 EFMRFKMHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAW  
 DILSPQFMYGSKAYVKHPADIPDYLLKLSFPEGFKWERVMNFEDGGVVTVT  
 QDSSLQDGEFIYKVKLRGTNFPDGPVMQKKTMGWEASSERMYPEDGAL  
 KGEIKQRLKLDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNE  
 DYTIVEQYERAEGRHSTGGMDELY**K****C**LEHHHHHH

<sup>2</sup> In [160] the count starts after the first methionine of Ypet, i. e., valin in this case.

<sup>3</sup> Here the count starts from the first methionine of Ypet (M<sub>5</sub> of YVCwt).

<sup>4</sup> Here, the amino acid count starts from the first out of the 35 amino acid long villin subdomain HP35. The corresponding mutations in villin are N68A and K70M.

### A.1.6 Anti-parallel heterodimeric coiled coil: U2LZ

The heterodimeric coiled coil construct is an anti-parallel leucine zipper (LZ) flanked by two ubiquitins (U2) and named U2LZ. The two zipper sequences that are separated by a triple GS-repeat, have been designed and shown to promote anti-parallel dimerization [85]. When folded, the N- to C-terminal first 29 amino acid sequence aligns with the C- to N-terminal second 30 amino acid sequence. Note, that the C-terminal glycine (G) has originally been a glutamin (Q). The gene of U2LZ was cloned into a pET-28a(+) vector, expressed, and purified at MPIB.

U2LZ. **M**A**C**KMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQ  
RLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGEL**A**L**K**KELQ**A**N**K**KEL  
**A**QLKWELQ**A**L**K**KEL**A**QGS**G**SG**S**E**Q**LE**K**KL**Q**ALE**K**KL**A**QLEWKN**Q**ALE  
**K**KL**A**GGTMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIF  
AGKQLEDGRTLSDYNIQKESTLHLVLRRLRGGGGMDELY**K**LEHHHHHH

### A.1.7 Homodimeric coiled coils from kinesin-1

Sequences of coiled coil unzipping constructs derived from *Drosophila melanogaster* Kinesin(-1) Heavy Chain are given in [Section A.2](#).

## A.2 MOLECULAR CLONING

During this work, a lot of effort has been spent on designing, cloning and measuring the unzipping of homodimeric coiled coil constructs isolated from *Drosophila melanogaster* Kinesin(-1) Heavy Chain (DmKHC, hereafter also referred to as 'kinesin'). Hence, some background information, used techniques, new findings and successful designs as well as further possible constructs are presented to encourage and to support future projects.

### A.2.1 Functional regions and domains of kinesin-1

The pGEX-6P vector<sup>5</sup> containing the gene of DmKHC between the *Bam*HI and *Not*I restriction sites, was kindly provided by Stephan Roche. Thorough gene sequencing by GATC revealed 20 silent mutations<sup>6</sup> and a stop codon after amino acid Q950. Therefore the rest of kinesin's overall 975 amino acids is shown in gray. Functional regions have cyan **letters**, predicted coiled coil domains interesting for unzipping experiments have bold **letters**.

DMKHC. MSAEREIPAEDSIKVV**C**RFRLNDSEEKAGSKFVVKFPNNVEE  
N**C**ISIAGKVYLFDKVFKPNASQEKVYNEAAKSIVTDVLAGYNGTIFAYGQT

<sup>5</sup> Be aware that in contrast to the kanamycin resistance of the predominantly used pET-28a vector, the pGEX-6P vector carries an ampicillin resistance.

<sup>6</sup> Most likely attributed to a sequence optimization, the silent mutations are important to know for optimal amplification and mutation primer design.

SSGKTHMTEGVIGDSVKQGGIIPRIVNDIFNHIYAMEVNLEFHIKVSYYEIM  
 DKIRDLLDVSKVNLSVHEDKNRVPYVKGATERFVSSPEDVFEVIEEGKSNR  
 HIAVTNMNEHSSRSHSVFLINVKQENLENQKKLSGKLYLVDLAGSEKVS  
 TGAEGTVLDEAKNINKSLSALGNVISALADGNKTHIPYRDSKLTRILQESL  
 GGNARTTIVICCPASFNESKSTLDFGRRAKTVKNVVCVNEELTAEEWK  
 RRYEKEKEKNARLKGKVEKLEIELARWRAGETVKAEEQINMEDLMEAS  
 TPNLEVEAAQTAEEAALAAQRTALANMSASVAVNEQARLATECERLY  
 QQLDDKDEEINQQSQYAEQLKEQVMEQEELIANARREYETLQSEMARI  
 QQENESAKEEVKEVLQALEELAVNYDQKSQEIDNKNKDIDALNEELQ  
 QKQSVFNAASTELQQLKDMSSHQKKRITEMLTNLLRDLGEVGOAIAFGE  
 SSIDLKMSALAGTDASKVEEDFTMARLFISKMKTEAKNIAQRCSNMETQQ  
 ADSNKKISEYEKDLGEYRLLISQHEARMKSLQESMREAENKKRTLLEEQI  
 DSLREECAKLKAAEHVSAVNAEEKQRAEELRSMFDSQMDELREAHTR  
 QVSELRDEIAAKQHEMDEMKDVBHQKILLAHQQMTADYEKVRQEDAE  
 KSELQNIILTNERREQARKDLKGLDTPVAKELQTLHNLKLFVQDLQQR  
 IRKNVVNEESEEDGGSLAQKQKISFLENNLDQLTKVHKQLVRDNADLR  
 CELPKLEKRLRCTMTERVKALETALKEAKEGAMRDRKRYQYEVDRIKEAV  
 RQKHLGRRGPQAQIAKPIRSQQGAIAIRGGGAVGGPSPLAQVNPVNS

Kinesin consists of a N-terminal globular head domain (roughly up to K331) and a rather extended rod domain (starting at the green A345 until end). While being the catalytic motor domain, the head is connected via a short, flexible neck-linker (T332 - T344, derived from T324 - T336 in human kinesin via sequence alignment, [149]) to the rod domain [180]. The mainly  $\alpha$ -helical rod is subdivided into neck (N), stalk (S), and tail (T) domains [239]. The predicted coiled coil regions of these domains are connected via rather unstructured regions, i. e., the hinge (H<sub>1</sub>) between neck and stalk and the kink, also called hinge 2 (H<sub>2</sub>), between the first (S<sub>1</sub>) and second coiled coil (S<sub>2</sub>) of the stalk [54, 87, 90]. The middle of H<sub>2</sub> is marked by the helix breaker proline P587 directly followed by glycine G588 [210]. After S<sub>2</sub>, follows a stalk-tail linker region to which the Kinesin Light Chains (KLC) bind (V810 - K834, derived from V789 - K813 in human kinesin via sequence alignment, [57]), which in turn are able to bind cargo. Finally, after the tail forms a coiled coil [217], it terminates in a globular tail region. This terminal region carries a highly conserved motif for regulatory function of the kinesin-1 superfamily, here QIAKPIRS (Q941 - S948, derived from Q919 - P926 in human kinesin via sequence alignment), which directly interacts with the Switch I region in the head domain (R197 - H212, derived from R190 - H205 in human kinesin via sequence alignment). The Switch I region is associated with nucleotide binding and, hence, is essential for motor activity. It ends with the highly conserved SSRSH motif [58].

#### *DmKHC structure and coiled coil prediction*

While the globular head domain of DmKHC has a well resolved crystal structure (pdbID: 2Y5W), the extended rod domain is only known to consist of

multiple extended parts due to coiled coil formation. The exact coiled coil regions and their stability remain unclear. A combined approach based on coiled coil prediction and direct mechanical single-molecule characterization is expected to clarify the true nature of the rod domain. This is especially interesting due to the fact that besides dimerization the rod domain is involved in many processes necessary for correct biological function of the kinesin motor, e. g., KLC binding or regulatory functions (see above).

Already in 1953 Francis Crick described the 'knobs-into-holes' packing of two  $\alpha$ -helices forming a coiled coil structure [50]. A typical 'heptad' repeat (positions indicated with a, b, c, d, e, f, and g) of generally apolar amino acids at positions a and d would give rise to a hydrophobic inner surface between the two helices. Furthermore, charged residues would be preferentially solvent exposed and thus found at the other positions [46]. Based on these rules, coiled coil prediction algorithms evolved which calculate the probability of a given amino acid sequence to form a coiled coil. Generally, these algorithms also consider amino acid sequences already known to form coiled coils. The two algorithms used in this theses were COILS [134] and Paircoil2 [144]. Analysing the entire DmKHC sequence with Paircoil2 yielded an 85 - 100 % coiled coil probability for the bold typed regions of the sequence. On average a probability of 92 % at a window size of 26 residues is reached. These regions, hence, mark the dimerization of kinesin driven through coiled coil formation of the neck, stalk (S1 & S2), and tail. The rest of DmKHC has a probability of less than 15 % to form a coiled coil. Interestingly, COILS predicts a further extension of coiled coil formation especially for the C-terminal tail region even at reduced window size. Else, all local features of Paircoil2, e. g., the helix breaker P587, are reproduced by COILS as well.

#### A.2.2 *Design and cloning of coiled coil unzipping constructs*

The unzipping geometry of a coiled coil is the same as for DNA unzipping [25]. The major difference are two identical  $\alpha$ -helices in case of a homodimeric coiled coil as compared to the different complementary DNA strands. As repetitive pulling cycles of entire coiled coil unzipping and re-zipping of one single molecule are desired, strand separation of the dimeric construct needs to be prevented. This has been achieved by introducing a covalent cysteine crosslink at the opposite end with respect to force application [30]. To be able to use our cysteine chemistry for oligo attachment, an additional ubiquitin molecule is inserted between reactive cysteine and individual coil on the side of force application [83]. So in contrast to the cysteines for crosslinking, the cysteines for oligo attachment are spatially separated. As the mechanical stability of ubiquitin is known to be much higher than the stability of the investigated coiled coils, the ubiquitins can be expected not to interfere with the mechanical characterization measurements [195].

*Cloning procedure*

After a certain coiled coil region of DmKHC has been chosen as investigative target, it was transferred from the pGEX-6P vector into a specially designed pET-28a-2Ubi vector already containing two ubiquitins. The specific order of relevant restriction sites, cysteines (C) for oligo attachment, ubiquitins (Ubi) and His-Tag (H<sub>6</sub>) for purification is: *NcoI*<sup>7</sup>—C—Ubi—*SacI*<sup>8</sup>—*HindIII*—Ubi—C—*XhoI*—H<sub>6</sub>. To create a coiled coil construct, where unzipping starts from the N-terminus, the pET-28a-2Ubi vector is cut at the restriction sites *SacI/XhoI* and treated with Antarctic Phosphatase<sup>9</sup> to prevent re-ligation. The same restriction sites are accordingly added to the gene of the chosen coiled coil region upon PCR amplification from the pGEX-6P vector. For this purpose specially designed amplification primers<sup>10</sup> are used, which also insert the cysteine necessary for crosslinking. After cutting the amplification PCR products as well, they are ligated into the cut pET-28a-2Ubi vector. Finally, the resulting expression vector encoding a N-terminal ubiquitin followed by a coiled coil is transformed into competent cells. Using the restriction sites *NcoI/HindIII* yields the opposite construct, i. e., a coiled coil that can be unzipped from the C-terminus. An example for resulting amino acid sequences of constructs being unzipped starting from either N- or C-terminus can be found in [Section A.2.2](#). To create monomer constructs, where ubiquitins are usually added to the N- and C-termini of the respective protein [188, 216, 261], the restriction sites *SacI/HindIII* would be used<sup>11</sup>.

**KITS, CULTURES AND DETAILS.** After treatment with restriction enzymes or polymerases, plasmids and PCR products were always purified with the QIAGEN Plasmid Mini Kit and QIAquick PCR Purification Kit (Qiagen). All restriction enzymes<sup>12</sup> as well as ligases were from New England Biolabs (NEB). Besides the Quick Ligation Kit (M2200L), the standard T<sub>4</sub> DNA Ligase (Mo202S) was used, especially for very long inserts. Due to its 50-fold lower error rate as compared to usual Taq Polymerase, the Phusion High-Fidelity Taq DNA Polymerase (F-530L, Finnzymes as part of Thermo Fischer Scientific) was used for amplification PCR. Successful treatment with Antarctic Phosphatase (Mo289S, NEB) was tested via a parallel attempt to ligate the

- 
- 7 For amplification primer design an eye should be kept on the frame-shifted encoding of a methionine within the *NcoI* restriction site, especially, as expression can start from there.
- 8 The used vector actually has two *SacI* restriction sites. Hence, the amount of used restriction enzyme has to be adapted. For simplicity the second *SacI* was omitted in the main text.
- 9 Antarctic Phosphatase and the necessary amount of buffer were added during the last 15 minutes of cutting.
- 10 In principle, the restriction sites are introduced as 5'-overhangs to the amplification primers. To ensure proper cutting, a 'non-sense' ACCTCTG-sequence is added to the 5'-end resulting in: 5'-ACCTCTG—restriction-site—coiled coil.
- 11 In the pET-28a-2Ubi vector there is another restriction site right between *HindIII* and the following ubiquitin, namely *KpnI*, which can be used as well.
- 12 Note that the units per volume [U/mL] can vary for different enzyme supplies. Hence, the used amounts need to be adapted, especially, when applying multiple restriction enzymes in the same reaction. Enzyme activity should be tested separately.

cut pET-28a-2Ubi vector without addition of untreated insert. Ideally, after transformation no cells should be able to grow.

After transformation, cells were plated on agarose gels and grown overnight. The next day, Crimson Taq DNA Polymerase (NEB) was used for colony PCR to screen multiple colonies for the correct insert size. Therefore, colonies were picked, shortly dipped into a prepared PCR reaction and put into LB medium. Positive colonies were grown overnight and their plasmid was purified using the QIAprep Spin MiniPrep Kit and sent to GATC for sequencing. The glycerol cultures of competent cells with confirmed exact sequences were stored at  $-80^{\circ}\text{C}$ . Glycerol cultures consisted of 250  $\mu\text{L}$  cell culture being well mixed with 750  $\mu\text{L}$  glycerol.

For transformation of ligation products either XL10-Gold Ultracompetent or XL1-Blue Supercompetent Cells were used. These cells served as indirect plasmid 'storage' since fresh plasmid could be purified from overnight cultures inoculated from the glycerol stocks whenever needed. For expression BL21-CodonPlus (DE3)-RIPL Competent Cells<sup>13</sup> were used. All cells originated from Agilent Technologies (AT).

Cell competence was tested by transforming uncut plasmid. Cell growth was tested with untreated cells in LB medium without antibiotics. Untreated cells were also tried to be grown in LB medium with antibiotics to assure selectivity for plasmids containing antibiotics resistance. If cells without intrinsic resistance were still able to grow, they must have gotten contaminated. A 50 mL falcon containing only LB medium without cells but other than that being treated all the same was incubated overnight as well and served as a zero growth reference and a control of uncontaminated LB medium. In this thesis carbenicillin was used instead of ampicillin.

All primers used for amplification or mutation reactions were ordered from metabion. Mutations include removal of native restriction sites, cysteine replacement and tryptophan insertions as described below. Mutations were performed using the QuikChange Lightning Multi Site-Directed Mutagenesis Kit (AT), which allowed up to five successful mutations in one step.

Essentially, protocols were adopted as supplied by the companies.

### *Choosing crosslinking positions*

The design of coiled coil unzipping constructs requires a cysteine at one end to establish a covalent crosslink between the two coils. To promote the formation of the disulfide bond between the two reactive thiol groups of the cysteines, they are positioned on either a or d positions within the heptad repeat. Furthermore, the amino acid, which is being replaced by the cysteine, should originally be hydrophobic as required for good coiled coil formation. This increases chances for choosing a working crosslinking position as the hydrophobic residues can be expected to 'hide away' from the solvent by pointing inside the coiled coil and thus towards one another, which should

<sup>13</sup> For BL21-Codon Plus cells 20 seconds are recommended for heat shock transformation. This is only half of the 45 seconds typically used for usual BL21, XL1 or XL10 cells.

promote crosslinking. All four coiled coil regions of DmKHC highlighted in bold letters were chosen such that they start and end at either a or d position. Confirmed and potential crosslinking positions are coloured in green and orange, respectively.

*Cysteine-free mutant design based on multiple sequence alignment and coiled coil prediction*

The cysteine-based chemistry for attaching oligos as well as the engineered cysteine crosslink of the dimer can interfere with naturally occurring cysteines within the rest of the protein sequence. In the sequence of DmKHC all natural cysteines are coloured in magenta, i. e., five cysteines within the head domain (C<sub>17</sub>, C<sub>45</sub>, C<sub>309</sub>, and C<sub>310</sub>) and neck-linker region (C<sub>338</sub>) and five cysteines within the rod domain (C<sub>441</sub>, C<sub>632</sub>, C<sub>695</sub>, C<sub>880</sub>, and C<sub>891</sub>).

To find a good substitute for each cysteine, sequences of 18 different kinesins were aligned. This helped to determine the either rather hydrophobic (suggested substitutes from alignment were valine, leucine or isoleucine) or rather hydrophilic (suggested substitutes from alignment were serine, threonine, lysine or arginine) surrounding of each cysteine. Interestingly, for the rod domain these results match perfectly well with the assigned positions within the heptad repeat, i. e., a and d positions for hydrophobic and the rest for hydrophilic environments. The best steric mimic of a cysteine was then chosen for substitution, i. e., serine for a rather hydrophobic and valine for a rather hydrophobic environment.

As oligo attachment and cysteine crosslinking occur after the monomeric protein has folded, in the case of the head domain only surface exposed cysteines need to be mutated. Hence, only mutations C<sub>17</sub>S and C<sub>338</sub>S were made to create 'cysteine-light' constructs to investigate proposed neck-linker docking of an active motor with the optical tweezers setup [52, 93, 105, 180]. Importantly, motor activity, though reduced in speed with respect to literature values, was confirmed for these constructs. When mutations within an active domain are made, verification of activity is crucial as completely inhibited motor activity has been reported for entirely 'cysteine-free' mutants [93].

The following mutations were made<sup>14</sup> for the rod domain: C<sub>441</sub>V, C<sub>632</sub>V, C<sub>695</sub>V, C<sub>880</sub>S, and C<sub>891</sub>S. While the first three mutations concern cysteins on hydrophobic a positions within the heptad repeat of the predicted coiled coil formation, the last two mutations are cysteins on positions f and c.

*Silent mutations to delete native restriction sites*

As explained above, stretches of DmKHC were amplified via PCR from the pGEX-6P vector and ligated into a specially designed pET-28a(+) vector either using the restriction sites *NcoI*/*HindIII* or *SacI*/*XhoI*. As the DmKHC gene itself also contained *NcoI*, *SacI* and *XhoI* restriction sites, these needed

<sup>14</sup> Mutation primer design can be performed with the online tool QuikChange Primer Design: <http://www.genomics.agilent.com/primerDesignProgram.jsp>

to be deleted via silent mutations. These mutations are called silent as they comprise changes on DNA level, which result in an unchanged amino acid sequence. Or, in other words, a different amino acid codon is used. All silent mutations are underlined within the DmKHC sequence: *dNcoI*-T612T, *dXhoI*-L684L, *dSacI*-L740L, and *dNcoI*-T892T. The 'd' stands for deletion of the respective restriction site. Upon designing mutation primers, special care was taken to use the most frequent and possible codon of the expression system, i. e., *E.coli*, based on a modified overview from [139]. In case of *dNcoI*-T612T, the DNA sequence ACC.ATG.GCG encoding the amino acids T.M.A was mutated into ACC.ATG.GCG. The nucleic acids in bold type mark the restriction site specific sequence. As methionine only has one unique codon and the encoding for alanine needs to start with a guanine, only the threonine could be considered for a silent mutation.

#### *Tryptophan-enhanced protein absorption*

If a selected coiled coil region has neither a tryptophan nor a few tyrosines, it is very difficult to detect the expressed protein during purification. Since the additional ubiquitin carries only a single tyrosin, the absorption signal of the entire construct will be very low up to almost invisible when using standard ultraviolet spectrophotometry. To overcome this problem, additional tryptophans were inserted into the expression vector to increase the detectable absorption signal of the expressed protein, see sequences in Section A.2.2. As a result, up to 8-fold higher extinction coefficients were reached, which extremely facilitated protein purification and subsequent attachment of oligos.

#### *Explicit example: stalk 2 domain unzipping*

Construct names refer to the kinesin regions where the construct originates from. The neck coiled coil is called N, the coiled coils of stalk 1 and 2 are S1 and S2 and the shortest tail coiled coil is simply called T. As the length of the tail coiled coil is unclear and three potential C-terminal crosslinks are hypothesized, the longer versions are called T2 and T3. The ultimate goal is to map the entire energy landscape of the kinesin rod domain. Hence, combined constructs are needed to characterize the transition regions between the coiled coils. Therefore almost all possible subparts, e. g., NS1<sup>15</sup>, S1S2, NS2 or S2T2, were cloned including the longest possible construct<sup>16</sup>: NT3.

The S2 constructs S2n and S2c serve as representative examples for all cloned homodimeric DmKHC coiled coil constructs for mechanical unzipping studies. The ending in the name indicates N- or C-terminal unzipping from the side where additional ubiquitins were engineered into the construct to serve as oligo attachment points. While cysteines for oligo attachment are

<sup>15</sup> NS1 means that the construct starts with the neck coiled coil N and ends with the C-terminal end of S1 including the hinge between the two.

<sup>16</sup> The construct NT3 ranges from A345 to V928 and comprises two times 584 amino acids resulting in an expected contour gain of about 426 nm.

still coloured in red, blue cysteines at the opposite end are mutations meant to form a crosslink upon dimerization. While the C695V mutation is coloured in blue as well, the silent mutations dXhoI-L684L and dSacI-L740L are additionally underlined. Additional tryptophans (**W**), which increase the protein's absorption for enhanced detection during purification, are purple.

S<sub>2N</sub>. MA**W**G**C**KMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPD  
 QQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRG**G**ELISQHEARMKSL  
 QESMREAENKKRT**L**EEQIDSLREE**V**AKLKAAEHVSAVN**A**E**E**KQRAEE  
 LRSMFDSQMDELREA**H**TRQVSEL**R**DEIAAKQH**E**MD**E**MKDVH**Q**K**C**LE  
**W**HHHHHH

S<sub>2C</sub>. MG**C**SQHEARMKSLQESMREAENKKRT**L**EEQIDSLREE**V**AKLK  
 AA**E**HVSAVN**A**E**E**KQRA**E**ELRSMFDSQMDELREA**H**TRQVSEL**R**DEIAA  
 KQH**E**MD**E**MKDVH**Q**KLK**L**GT**M**QIFVKTLTGKTITLEVEPSDTIENVKAKI  
 QDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRRLRG**G**K**C**LE  
**W**HHHHHH

#### *Successful and potential crosslinking positions*

Successful unzipping experiments were performed with the constructs S<sub>1n</sub>, S<sub>2n</sub>, S<sub>2c</sub>, NS<sub>1c</sub> and NS<sub>2n</sub>. These experiments confirm the following positions to work as engineered cysteine crosslinks: A345, L556, I660 and L761. The position W376 has already shown successful crosslinking in [31].

Crosslinking tests with SDS-gels of expressed constructs comprising the tail yielded that especially V928 seems to be a good crosslinking candidate. The positions Q850 and L882 also showed crosslinking potential, while position A910 remained unclear.

### A.3 PROTOCOLS

This section specifies all essential steps in advance to the mechanical characterization of a single protein at the optical tweezers setup. The steps comprise protein expression and purification, oligo attachment, DNA-handle PCR and measurement chamber preparation. Many footnotes and troubleshooting sections include crucial details and helpful recommendations.

#### A.3.1 *Protein expression and purification*

The prerequisite for protein expression is a cell culture containing an expression vector encoding the desired protein's sequence. Here, this is a glycerol culture of BL21-CodonPlus (DE3)-RIPL Competent Cells stored at -80°C.

### Expression

- **Inoculation 1.** Inoculation of 20 mL LB medium containing suitable antibiotics (usually kanamycin) with cells picked from a glycerol culture using a 200  $\mu$ L-pipette. Grow overnight (ON) in an incubator set to 37°C and shake at ~169 rpm.
- **Inoculation 2.** Inoculation of 400 mL LB medium containing suitable antibiotics in a 1 L-flask with the ON-culture from Inoculation 1. Grow at 37°C and shake at ~169 rpm for 2 to 5 h until  $OD_{600} = 0.6$  is reached.
- **Induction with IPTG.** Add IPTG<sup>17</sup> and induce for 3 h.
- **Harvesting.** Centrifuge for 30 min at 4.600 rpm in two 200 mL-flasks in a centrifuge<sup>18</sup> pre-cooled to 4°C. Discard supernatant and resolve pellet: add 10 mL of NaP<sup>19</sup> buffer to the first flask and vortex thoroughly. Transfer everything into the second flask and vortex again. Collect in 50 mL-falcon. Repeat procedure with another 10 mL of Soni buffer to collect remains.
- **Storage.** Store at -20°C until purification.

To test, whether the expression worked fine, compare expression test references from before and after IPTG-induction with a SDS-PAGE<sup>20</sup> run (SDS-gel). As a reference, simply withdraw 2 mL cell culture, spin down, remove the supernatant, and store the pellet at -20°C until the actual expression test. Note the corresponding  $OD_{600}$ -value to better adapt the amount of cell extract being loaded on the SDS-gel. The reference from after induction should show an over-expressed band corresponding to the size of the desired protein.

### Purification

Always work on ice and as clean as possible. Ensure that all used buffers are fresh, clean, filtered and degassed where necessary. When working with chromatography systems avoid air bubbles by all means. Never allow a direct mixture of salt containing buffers with ethanol containing cleaning and/or storage buffers.

- **Thaw.** Thaw harvested cells at RT (to be faster) or gently on ice. When thawed, keep on ice.
- **Disrupt cells.** Cells are disrupted mechanically<sup>21</sup> with an automated French Press<sup>22</sup> at a pressure of 1.6 bar. Right after the 20 mL of thawn

<sup>17</sup> Isopropyl  $\beta$ -D-1-thiogalactopyranoside.

<sup>18</sup> Rotanta 460R, Hettich.

<sup>19</sup> 50 mM  $NaH_2PO_4 \times 2 H_2O$ , 50 mM  $Na_2HPO_4 \times 12 H_2O$ , 300 mM NaCl, adjust pH to 8.0. Do not yet add imidazole to prevent increased foam formation during vortexing.

<sup>20</sup> Sodium Dodecyl Sulfate PolyAcrylamide GEL.

<sup>21</sup> Additional cell lysis may be considered prior or instead of mechanical cell disruption. This was not done for the purification of expressed coiled coil or Rossmann fold protein constructs. Nevertheless, for the activity of purified kinesin motor domains, addition of protease inhibitors into the lysis buffer turned out to be crucial.

<sup>22</sup> TS Series Benchtop, Constant Systems.

cell culture have passed, another 20 mL of NaP buffer are applied to the French Press to collect remains. From now on, work fast to allow as little unwanted reactions, e. g., the formation of cystein crosslinks, to happen as possible.

- **Centrifuge.** Centrifuge<sup>23</sup> disrupted cells for 30 min at 17.000 rpm and 4°C using the rotor JA-17. The supernatant contains the desired protein.
- **Sterile filtration and imidazole addition.** Filter the supernatant two times using membrane filters<sup>24</sup> with 0.2 µm pore size. Add 20 mM imidazole while maintaining pH 8.0.
- **Purification 1: metal ion affinity.** In the first purification step, a Ni-NTA column<sup>25</sup> is mounted onto an automated affinity chromatography protein purification system<sup>26</sup>. The buffers used for washing and for elution during this standard purification procedure are based on the NaP buffer supplemented with 20 mM and 500 mM imidazole respectively and pH adjusted to 8.0. The 2.5 mL eluate is collected in a 15 mL-falcon set on ice. Right along with elution, the system evaluates the protein concentration X based on the provided absorption coefficient<sup>27</sup>.
- **DTDP activation/passivation.** Immediately after elution, a 5- to 10-fold excess of DTDP with respect to reactive cysteines is added to the eluate. Therefore, a fresh solution of 100 mM DTDP solved in dimethyl sulfoxide<sup>28</sup> (DMSO) is prepared in advance. For a protein with two reactive cysteines, the necessary volume V of DMSO-solution for adding a 5-fold excess of DTDP to the 2.5 mL eluate can be calculated according to:

$$\frac{250 \cdot X [\text{mg/mL}]}{m [\text{kDa}]} = V [\mu\text{L}] , \quad (69)$$

where m is the molecular weight of the purified protein. As too much DTDP can promote precipitation, right after addition of DTDP the sample needs careful and at the same time quick mixing; avoid a final concentration of more than 5 mM DTDP. Let stand for 45 to 60 min at RT.

*Background:* The disulfide bond of a DTDP molecule gets broken open by a reactive thiol group of the protein's cysteine resulting in a thiol-pyridine activated cysteine and a pyridine-2-thione [39]. DTDP-excess is necessary to prevent not yet activated cysteines from reacting with already activated ones by breaking up the disulfide bond with the pyridine ring. When all cysteines are thiol-pyridine activated, the protein solution is passivated against cystein crosslinking.

<sup>23</sup> Avanti J-E, Beckman Coulter.

<sup>24</sup> Filtropur S 0.2, Sarstedt.

<sup>25</sup> HisTrap, GE Healthcare; or Ni-NTA Superflow Cartridge, Qiagen; both 1 mL.

<sup>26</sup> Profinia, Bio-Rad.

<sup>27</sup> Extinction coefficient divided by molecular weight.

<sup>28</sup> DMSO has a melting temperature above 0°C, hence, do not put on ice.

- **Concentrate.** Pre-spin the passivated protein for 10 min at 4,600 rpm and 4°C to get rid of eventual aggregates which appear as a pellet. Concentrate or dilute the supernatant up or down to 14 mg/mL<sup>29</sup>. Use a centrifugal filter<sup>30</sup> for concentrating at 4,600 rpm and 4°C for 10 to 60 min<sup>31</sup>.
- **Purification 2: size-exclusion.** For SEC an appropriate column<sup>32</sup> is connected to a HPLC<sup>33</sup> system<sup>34</sup> and equilibrated in PBS buffer (pH 7.4<sup>35</sup>) in advance. Mount a 500 µL-loop and load it with 7 mg<sup>36</sup> of protein. Start a 60 min run at a flow-rate of 0.5 mL/min and collect<sup>37</sup> 30 s long fractions. Here, the relevant collection window of the mounted column ranges from ~15 to 45 min. Besides getting rid of larger unspecifically bound proteins and protein multimers, all excess DTDP (peak at ~57 min) is removed, too. If the collector was not set on ice, set collected fractions on ice immediately.
- **SDS-gel control.** To identify the fractions containing the desired protein, the most direct way is to screen all potential fractions on an SDS-gel. Therefore at least 10 µL per fraction are needed. In case of coiled coil constructs a preliminary test of the engineered cystein crosslink can be performed by running the same fraction with and without the reducing agent β-mercaptoethanol in the laemmli buffer, see Figure 71.
- **Storage.** Pool and concentrate<sup>38</sup> desired fractions to reach 2 nmol<sup>39</sup> of protein within less than 100 µL. Flash freeze<sup>40</sup> the aliquots and store at -80°C. These aliquots will be directly used for the oligo attachment discussed in Section A.3.2.

### A.3.2 Oligonucleotide attachment details

To covalently attach oligos to the thiol groups of two cysteines specifically engineered into the protein of interest, either thiol or maleimide modified

29 Consider the concentration provided by the first purification step.

30 Amicon Ultra 4 mL with matching membrane size, Merck Millipore.

31 High amounts of imidazole can cause clogging of the membrane resulting in very long centrifugation times. Switching to a second centrifugal filter can be helpful.

32 Superdex 200 10/300 GL, GE Healthcare.

33 High-Performance Liquid Chromatography.

34 HPLC system from Jasco.

35 If a thiol oligo attachment is intended, one should consider pH 8.0 to reach higher reactivity. In case of a maleimide oligo attachment a further buffer exchange will follow to reestablish the cysteines' thiol groups and to set the pH below 7.4.

36 Unspecific binding might falsify the estimated protein amount in the eluate of purification step 1. Therefore the specified maximal load of 10 mg is avoided.

37 Fraction collector: CHF 122SC, Advantec MFS.

38 Use centrifugal concentrators (Vivaspin 500 with appropriate pore size, Vivaproducts) for a benchtop centrifuge ideally pre-cooled to 4°C (Perfect Spin 24R, Peqlab).

39 This amount is adapted to match the 4 nmol of reactive oligos added to the oligo attachment reaction, where, in principle, a 1:1 ratio between reactive cysteines and oligos is intended. For coiled coils try to reach 4 nmol as there is only one reactive cysteine per protein monomer.

40 Before freezing, addition of 10% glycerol may be considered to impose less stress on the protein construct. This was not necessary for any of the investigated constructs though.

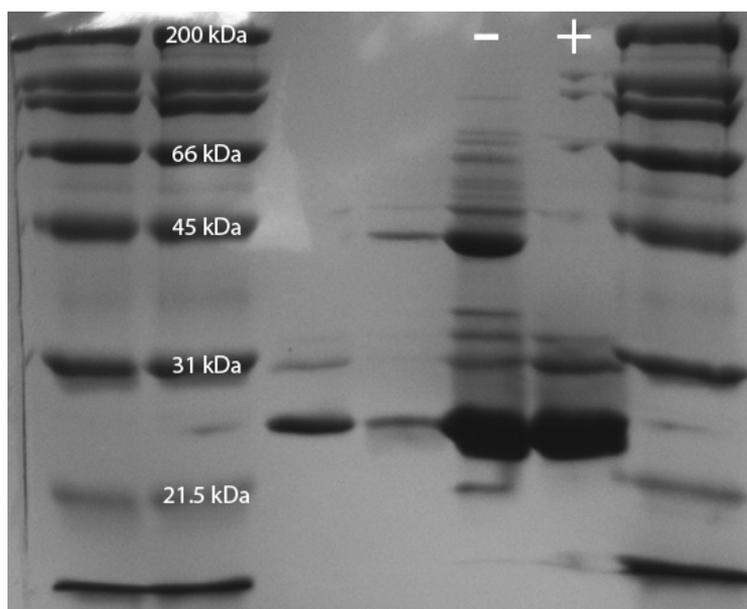


Figure 71: Test of an engineered cystein crosslink on an SDS-gel. SDS-gel after metal ion affinity purification of the coiled coil construct S1n with 21 kDa per coil. The (-)-lane containing laemmli buffer w/o reducing agent  $\beta$ -mercaptoethanol clearly shows a thick double-sized band whereas the (+)-lane with reducing agent does not. This indicates that the cystein mutation L556C within DmKHC works as a crosslink. Thinner bands are degradation products to be removed by subsequent SEC.

oligos<sup>41</sup> were used. The position of the modification is located at the 3'-end of the 34 nucleotides long sequence and marked with bold letters:

OLIGO. 5'-GGCAGGGCTGACGTTCAACCAGACCAGCGAGTC**G**-3'

#### *Thiol oligos*

After their activation and ethanol precipitation, thiol oligos are directly added to thiol-pyridine activated protein; see [Section A.3.1](#) for preparative protein treatment. Here, the activation and ethanol precipitation of thiol oligos are described in a step by step protocol.

1. Thaw necessary amount of 100  $\mu$ M thiol oligo solution.
2. **Activation:** add 10 mM TCEP<sup>42</sup> and let stand at RT for 30 min.
3. Add 1/10 volume sodium acetate (3 M, pH 5.2) and mix gently. By now the sample should be in a 1.5 mL LoBind reaction tube.
4. **Precipitation:** add 2 volumes 100 % EtOH (-20°C pre-cooled).

<sup>41</sup> Thiol oligo synthesis was performed by IBA GmbH and delivered as 100  $\mu$ M solution, maleimide oligos were synthesized and lyophilized by biomers.net GmbH.

<sup>42</sup> For example by adding 1/10 volume of 100 mM TCEP solved in either H<sub>2</sub>O or 10x PBS. Both solutions are acidic with pH 2 to 4, both of them worked fine.

5. Keep at  $-80^{\circ}\text{C}$  for 1 h.
6. Centrifuge for 30 min in a benchtop centrifuge pre-cooled to  $0^{\circ}\text{C}$  at maximum speed.
7. Carefully remove supernatant, pellet may or may not be visible.
8. Wash with 200  $\mu\text{L}$  of 70% EtOH.
9. Centrifuge for 10 min at 0 to  $4^{\circ}\text{C}$  at maximum speed.
10. Carefully remove supernatant.
11. Let air dry<sup>43</sup> for  $\sim 15$  min at RT or  $37^{\circ}\text{C}$  until white.
12. Resuspend pellet in appropriate amount<sup>44</sup> of buffer; pH 8.0 is recommended.

Next, the attachment reaction for thiol oligos with thiol-pyridine activated protein and the successive purification by SEC are outlined.

1. Measure the thiol oligo concentration<sup>45</sup> of resuspended oligo pellet(s).
2. **Attachment reaction mix:** add purified thiol-pyridine activated protein to the oligo solution in a 1:1 up to 1:2 ratio between cysteins and thiol oligos. Excess of thiol oligos showed higher protein-oligo construct yields<sup>46</sup>. Let stand at RT for 24 h.  
*Optional:* As oxygen promotes thiol reactivity, insert a small pipette tip into the reaction tube as a stirring element and place on a rotor<sup>47</sup>.
3. **Purification:** make a SEC run similar to the second protein purification step in Section A.3.1 - 0.5 mL/min, PBS, 100  $\mu\text{L}$ -loop, 30 s fractions, Superdex 200 column. An example is shown in Figure 72. Make 5  $\mu\text{L}$  aliquots<sup>48</sup> from the correct fraction(s), flash freeze, and store at  $-80^{\circ}\text{C}$ .

Figure 72 shows a typical SEC purification run of a thiol oligo attachment similar to the run shown for maleimide oligos in Figure 7. Besides the two peaks of the final protein-oligo construct with two attached oligos and protein

43 Eventually cover open reaction tube with Nescofilm and let only a small slit open to prevent (dust-) contamination.

44 To reach higher concentrations, an initial volume of, e. g., 40  $\mu\text{L}$  thiol oligos is being resuspended in half that volume at the end.

45 The final concentrations have been found to fluctuate a lot no matter what may be expected from the initial 100  $\mu\text{M}$  concentration. Therefore an individual determination at the Nanodrop is highly recommended for each reaction tube.

46 Not all thiol oligos are reactive after precipitation and resuspension. This is supported by the reactivity test of thiol oligos performed with the crosslinker BM(PEG)<sub>2</sub>. Importantly, cross-reaction between thiol oligos themselves was observed to be very poor, see Figure 73.

47 Additional rotation and stirring resulted in 40% more protein-oligo construct as compared to a parallel reaction from the same attachment reaction mix.

48 As a rule of thumb, peaks lower than 50 mV should be concentrated, below 25 mV neighbouring fractions should be pooled and concentrated.

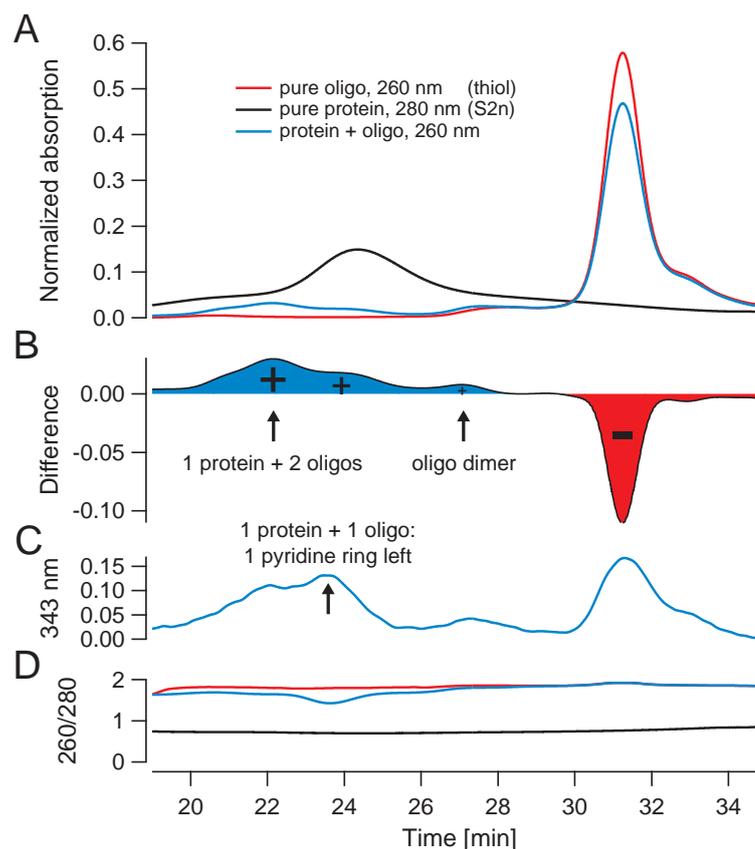


Figure 72: SEC purification of a thiol oligo attachment. (A) Normalized SEC runs of pure protein (280 nm), pure thiol oligos (260 nm), and oligo attachment reaction products (260 nm). (B) The difference of attachment product and pure oligos visualizes the elution of individual (+)-marked products. The first peak corresponds to protein with two bound oligos, the second to protein with one oligo, and the third to dimerized oligos. (C) 343 nm absorption signal (in mV) tracking thiol-pyridine activated cysteines. (D) 260/280 ratio of the SEC runs. The pure protein peak shows a value below 1. The peaks of the other runs show a ratio larger than 1, indicating that absorption is dominated by the contribution of DNA.

with only one attached oligo, the difference signal shows an additional peak due to the self-dimerization of thiol oligos. Due to a large enough protein size of S2c (about 45 kDa), the 260/280 ratio shows a dip around the elution volume of the protein with only one oligo being attached. This directly monitors the relatively higher protein absorption contribution per oligo as opposed to the protein-construct with two oligos. Furthermore, the absorption signal measured at a wavelength of 343 nm helps to track the remaining pyridine ring still bound to the thiol-pyridine activated cysteine where no oligo yet got attached to. The switched relative peak height at 343 nm as compared to 260 nm is due to all pyridine rings being replaced by thiol oligos in the elution volume of the desired protein-oligo construct.

### Reactivity and cross-reaction tests for thiol oligos

Inspired by the reactivity test for maleimide oligos presented in [Section 2.2](#), here, 1,8-Bismaleimido-diethyleneglycol (BM(PEG)<sub>2</sub>) instead of DTT was used as crosslinking agent to test the reactivity of thiol oligos. [Figure 73](#) shows that at least 40% of the thiol oligos, pretreated as described above, are reactive. Another test was performed to investigate the cross-reactivity of thiol oligos, i. e., their tendency to form dimers, by letting them stand at RT for 24 hours after resuspension in buffer without any further additives. As a result, [Figure 73](#) shows that only up to 7% of the thiol oligos dimerized within a day. This is an important result, as it allows for adding excess oligos into the attachment reaction mix without risking that all reactive oligos form dimers and, hence, cannot attach to the protein any more.

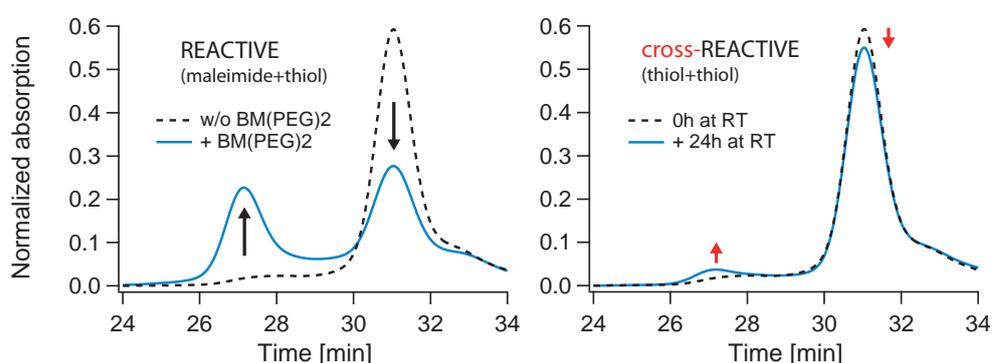


Figure 73: Thiol oligo reactivity test with BM(PEG)<sub>2</sub>. **Left:** The crosslinking reaction of activated thiol oligos with BM(PEG)<sub>2</sub> for 30 min at RT shows that at least 40% of the oligos are reactive. **Right:** A cross-reaction of thiol oligos among themselves for 24 h at RT reveals that only 7% of the oligos dimerized.

Screening the pH for crosslinking maleimide oligos with DTT confirmed the highest thiol reactivity to be around pH 8.0, see [Figure 9](#). To further improve the thiol oligo attachment, the influence of temperature and additional oxygen were investigated. Therefore one attachment reaction mix was split into three reactions with different reaction conditions, where the first was kept standing at RT, the second stood at 4°C, and the third was placed on a rotor at RT with a small pipette tip inserted into the reaction tube as a stirring element, which helps increasing the air contact area of the reaction mix. All three reactions lasted 24 ± 1 hours. With respect to the first reference reaction, the reaction at 4°C resulted in -20% and the reaction with an increased air contact area in +40% final protein-oligo construct (data not shown).

### Maleimide oligos

If a maleimide oligo attachment is intended, the DTDP activation/passivation step of the protein purification can be omitted. To prevent protein oligomerization via the cysteines, enough reducing agent should be added along with the purification. Right before the attachment reaction, 10 mM TCEP are added

to the purified protein to ensure reduction of all cysteines. However, as TCEP showed to reduce maleimide oligo reactivity, see Figure 9, it needs to be removed before final mixing of the attachment reaction. Similarly, thiol-pyridine activated protein can be reduced for attaching maleimide oligos.

Here, the preparation and characterization of maleimide oligos are described in a step by step protocol. The characterization comprises the determination of concentration, monomer fraction, and reactivity of the maleimide oligos and should be performed for every new batch (see Figure 76A).

1. Carefully dissolve lyophilized maleimide oligos in PBS, pH 6.7<sup>49</sup>.
2. Measure the concentration<sup>50</sup> of maleimide oligos.
3. Aliquot<sup>51</sup>, flash freeze, and store at -80°C.
4. Determine fraction of potentially reactive monomers via SEC<sup>52</sup>.
5. Test reactivity with DTT as described in Section 2.2.

Next, the protein activation for an attachment reaction with maleimide oligos, the attachment reaction and successive purification by SEC are outlined.

1. **Activation:** reduce protein with 10 mM TCEP<sup>53</sup> for 30 min at RT.
2. Remove TCEP by exchanging buffer with a desalting column<sup>54</sup>. Exchange towards PBS, pH 6.7.
3. Concentrate<sup>55</sup> eluate to above 100 μM<sup>56</sup> and verify final concentration.
4. **Attachment reaction mix:** add activated protein to an aliquot of freshly thawed maleimide oligo solution in a 1:1 ratio between activated cysteines and maleimide oligos - a formula is given in the following text. Let stand on ice or at RT for about 2 h<sup>57</sup>.
5. **Purification:** make a SEC run - 0.5 mL/min, PBS, 100 μL-loop, 30 s fractions, Superdex 200 column. An example is shown in Figure 7. Make 5 μL aliquots from the correct fraction(s), flash freeze, and store at -80°C.

49 Make sure the buffer is filtered and thoroughly degassed. The low pH ensures specificity of the maleimide-thiol reaction and should not surpass a maximum of 7.4.

50 Typically 80 μM are reached when the indicated amount of buffer to reach 100 μM are added.

51 Typically 2 nmol oligos are aliquoted in 10 to 20 μL depending on concentration.

52 Use one aliquot for the SEC run and determine the relative area under the monomer peak; examples of obtained results are given in Figure 76.

53 Note that TCEP is highly acidic. Prepare a 100 mM TCEP-solution in PBS and titrate to a suitable pH using 10 M NaOH. Add 1/10 volume.

54 PD MiniTrap G-25, GE Healthcare. Using the gravity protocol yielded best results with less than 10 % protein loss.

55 Vivaspin 500 with appropriate pore size, Vivaproducts.

56 Higher concentrations help minimize the final reaction volume leading to higher attachment yields, see Figure 9.

57 In special cases ON reactions in the cold room may be considered to increase very low yields. In general, reaction times exceeding 1 hour did not show any increase in yield neither for the test with DTT nor for proteins.

To calculate the appropriate amount of protein to be added to a maleimide solution for an efficient attachment reaction, the following formula was used:

$$\frac{c_{\text{oligo}} \cdot V_{\text{oligo}}}{2 \cdot c_{\text{prot}}} \cdot \text{frac}(\text{mono}) \cdot \text{frac}(\text{reactive}) = V_{\text{prot}}. \quad (70)$$

Here,  $V$  stands for volume,  $c$  for concentration,  $\text{frac}(\text{mono})$  for the fraction of oligo monomers, and  $\text{frac}(\text{reactive})$  for the fraction of reactive monomers. The denominator 2 reflects the amount of activated cysteines per protein. The highest observed value for  $\text{frac}(\text{reactive})$  was 0.77 for the U2LZ construct.

#### *Protein-oligo construct test*

Besides a sophisticated interpretation of the detected absorption signals of oligo attachment SEC purification runs, a simple and direct test can be performed to identify the fractions which contain protein with two covalently bound oligos [82]. For this test, purified and eventually pooled and concentrated protein-oligo constructs are mixed with DNA-handles carrying a ss-overhang complementary to the oligos bound to the protein and incubated at RT for at least 30 minutes. Thereafter, the reaction mix is subjected to an agarose gel electrophoresis run. With the DNA-handles having a characteristic size slightly bigger than 500 base pairs (see Section A.3.4), a protein-oligo construct with two bound oligos and two handles hybridized to them will typically appear as a band slightly larger than 1 kilobase, see Figure 74.

This test is also helpful for tuning the mixing ratio between protein-oligo construct and DNA-handles as well as the incubation time for the final measurement mix, see Section A.3.5. Nevertheless, a detectable band on an agarose gel is not obligatory for successful single molecule experiments.

#### *Troubleshooting - purification issues and attachment reaction conditions*

A well conducted purification of the oligo attachment reaction is crucial for successful experiments. Initially, the combination of bad column performance, product contamination with dimerized oligonucleotides and uncorrected collection time shifts, made it almost impossible to purify small protein-oligo constructs. Figure 75A compares the performance of an old and a new column on the same sample. Characteristic collection time shifts are shown in Figure 75B and Figure 75C. The issue with dimerized oligonucleotide contamination is dealt with in Section A.3.3.

To correctly interpret SEC purification runs of oligo attachment reactions, it is highly recommended to make reference runs of the two individual components, namely pure oligos and pure protein, as depicted in Figure 8 and Figure 72. If, like in case of the tension sensors, larger fluorescent proteins are involved, consider that the protein peak with no oligos bound can contribute a substantial additional peak close to the other peaks of interest.

Minimize losses by filling up the sample to full loop volume before loading.

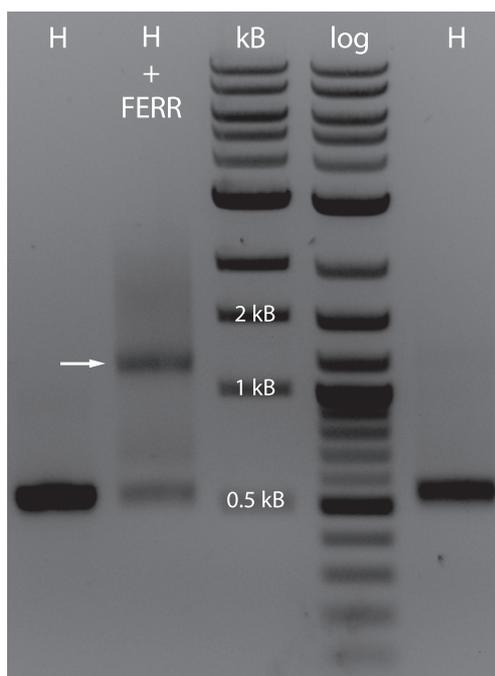


Figure 74: Protein-oligo construct test by agarose gel electrophoresis. Lanes labelled with H were loaded with purified DNA-handles; as expected, their characteristic band size is slightly bigger than 500 base pairs. The lane labelled with H + FERR is the corresponding protein-oligo construct incubated with DNA-handles - a white arrow points towards the double-sized band of two handles hybridized to the two oligos of the construct.

To minimize the effective loop size for an increased SEC resolution, an interesting option is to use an additional small Ni-NTA column in front of the SEC column as described in Section A.3.3.

Thorough vortexing of the buffer used to resuspend thiol oligo pellets is expected to increase the amount of oxygen bound in the buffer and therewith to increase thiol reactivity.

Even if it may seem contradictory, increasing the protein amount with respect to oligos sometimes helped a lot to improve the yield of a thiol oligo attachment and, hence, may be considered when dealing with low yields. A possible explanation can be that not all cysteines got (thiol-pyridine) activated or are somehow hard to access.

### A.3.3 SEC in series with Ni-NTA

After an oligo attachment reaction, products are purified by SEC. Here, the major goal is to get rid of any oligos which did not attach to the protein. This is crucial as all oligos are complementary to the overhangs of the DNA-handles and, hence, compete with the protein-oligo construct during incubation with handles to form the final protein-DNA chimeras.

When thiol oligos are used, they can also react with each other during an attachment reaction and form oligo dimers, see Figure 73. Such dimers can

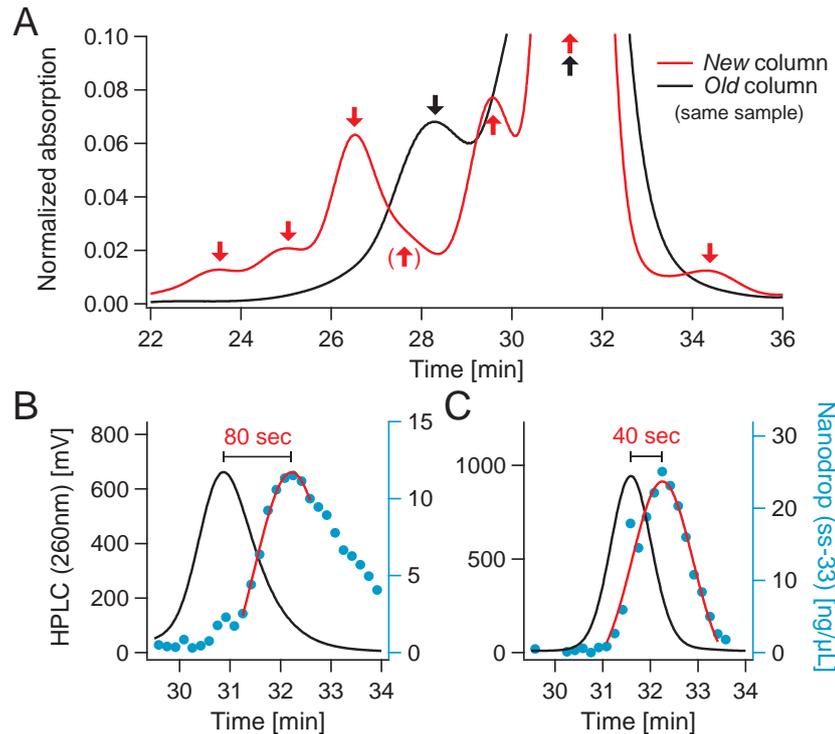


Figure 75: Superdex 200 SEC column characterization: resolution and collection time shift. SEC runs on PBS, at 0.5 mL/min flow-rate, and with 100  $\mu$ L-loop. (A) Maleimide oligonucleotide attachment reaction with ROSSorg purified with a new (red line) and an old (black line) column. Seven peaks and shoulders distinguished by the new column collapse into two peaks resolved by the old column marking an extreme loss in resolution. The two main peaks were aligned in time. (B), (C) Black curves are oligo monomer peaks detected with the HPLC systems in lab 3167 (chemistry, (B)) and lab 3118 (formerly origami, (C)). Blue dots are concentrations of collected 10 s fractions measured at the Nanodrop. Red lines are local Gaussian fits.

bind two DNA-handles forming constructs which result in stretching mere DNA during mechanical measurements at the trap (D events, see troubleshooting part in Section A.3.5). These events can heavily undermine experimental efficiency. Importantly, even though maleimide oligos are not able to form dimers during an attachment reaction, their lyophilized supply was found to contain contaminations showing dimer-like properties including the occurrence of D events which strongly corrupted measurement efficiency, too; see Figure 76A for the contaminations<sup>58</sup>. Hence, complete removal of oligo dimers or dimer-like constructs is key to successful experiments. An explanation for the dimer-like constructs observed with maleimide oligos could be a contamination of the basic oligonucleotides with DTT as reported in [127].

If a Superdex 200 column is used to purify the protein-oligo construct after an attachment reaction, it is not possible to completely remove oligonu-

<sup>58</sup> The relative amount of contaminants was derived from the area under their peaks relative to the overall area including all peaks.

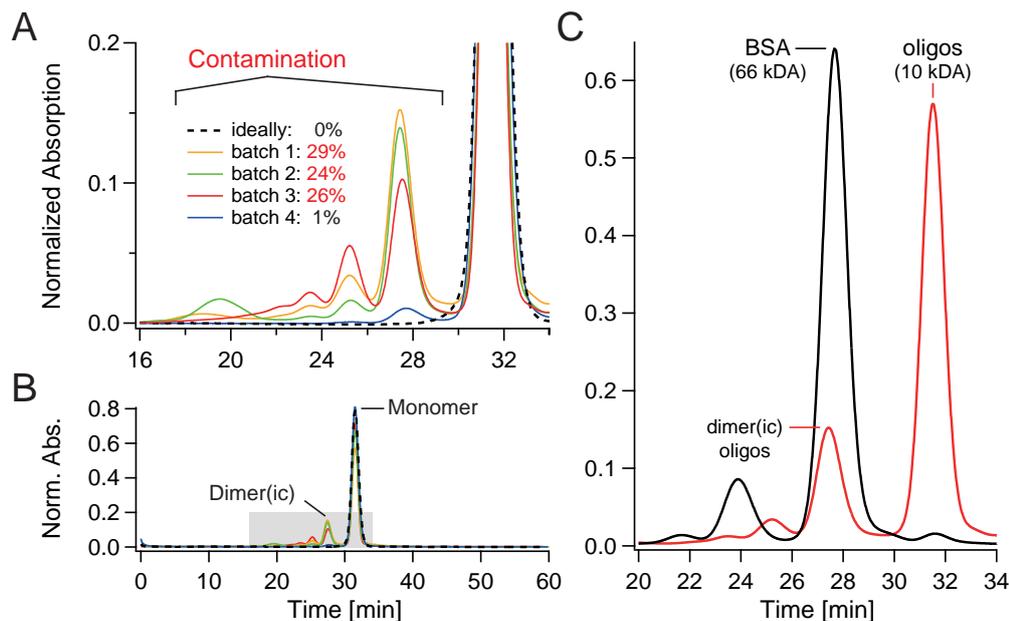


Figure 76: Contamination and elution volume of maleimide oligonucleotides revealed by SEC. SEC runs on PBS, at 0.5 mL/min flow-rate, and with 100  $\mu$ L-loop. (A) Zoom into the gray shaded area of (B) Different batches of ordered maleimide oligonucleotide monomers show an up to 30% contamination with bigger side products. The dashed curve is a re-run of a purified monomer fraction showing no further multimer formation. (C) A comparison with BSA (from Sigma Aldrich) shows that the elution volume of the dimer-like fraction of maleimide oligos equals the one of BSA.

cleotide dimers in case of small proteins like, e. g., the Rossmann fold construct with about 14 kDa, see Figure 7. In SEC runs, the rather stretched conformation of the short oligonucleotides makes them appear bigger than compact globular proteins of similar weight<sup>59</sup>. This leads to an overlap of the products of the attachment reaction. Even though a single oligonucleotide has a molecular weight of about 10 kDa, it already elutes around a volume of 15.5 mL, i. e., 31 minutes at a flow-rate of 0.5 mL/min, along with globular proteins of about 35 kDa in size according to the column's specifications (Instructions 71-5017-96 AK). The oligonucleotide dimers yet elute after 14 mL along with BSA (66 kDa), see Figure 76C.

To circumvent this problem, an additional metal ion affinity chromatography purification step using a Ni-NTA column was taken before the SEC run. This step makes use of the His-Tag engineered into the protein constructs. After loading the entire oligo attachment reaction on the Ni-NTA column, unreacted oligos and oligo dimers can be washed away, while all protein remains bound, i. e., protein with no, one or two attached oligos, see Figure 77.

Next, the three major steps of Ni-NTA column cleaning, loading and washing a sample on a Ni-NTA column, and elution of a sample from the Ni-NTA

<sup>59</sup> The same was observed for coiled coil constructs as well. In SEC runs they would elute much earlier than expected due to their rather stretched conformation.

column in series with a SEC column, are logged. Individual steps are separated by semicolon.

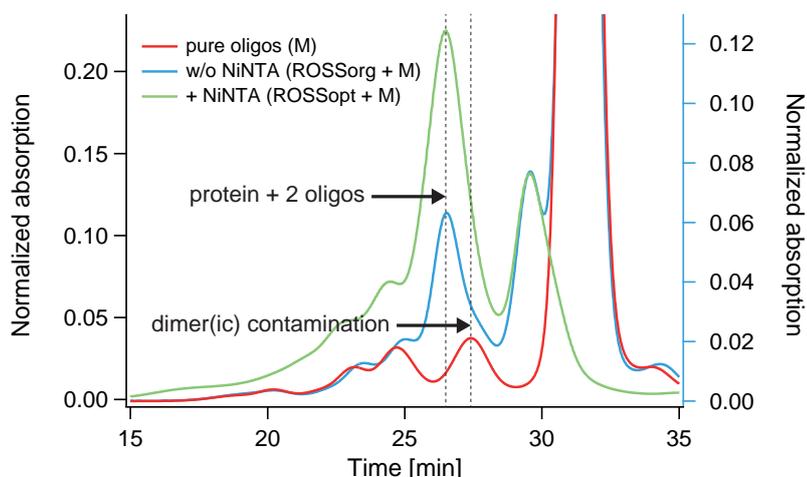


Figure 77: Improved purification of an oligo-attachment reaction by an additional Ni-NTA column. The green run shows that all unreacted maleimide oligos still present in the blue and red trace were successfully removed. Most importantly, the same holds for the removal of the dimer-like contamination. Note that ROSSopt binds two oligos way more efficiently than ROSSorg due to a more flexible design of attachment points (see sequences in Section A.1): the ratio of the peak height of the elution volume between two oligos bound to the Rossmann fold and only one bound oligo increased from 3:4 for ROSSorg to almost 2:1 for ROSSopt.

#### *Clean and restore column*

Clean and restore a Ni-NTA column. Wash with 10 column volumes (CV) of distilled water (ddH<sub>2</sub>O); 10 CV 0.5 M EDTA<sup>60</sup>; 10 CV ddH<sub>2</sub>O; 10 CV 0.2 M NaOH; 5 CV ddH<sub>2</sub>O; 10 CV cleaning buffer: 6 M Guanidinium chloride + 0.5 M Imidazole in 50mM Tris pH 7.5; 10 CV ddH<sub>2</sub>O; 5 CV 0.1 M Ni-II-Chloride Hexahydrate<sup>61</sup>; 10 CV ddH<sub>2</sub>O; for long-term storage: 5 CV 20% EtOH.

#### *Load and wash sample*

Load the sample<sup>62</sup> on a Ni-NTA column and wash away unbound oligonucleotides. If the column is not yet on water, first equilibrate with: 10 CV ddH<sub>2</sub>O; then: 10 CV sample buffer<sup>63</sup>; load sample: use a volume appropriate for easy sample loading while minimizing losses; optional: 10 CV sample

<sup>60</sup> Solve in ddH<sub>2</sub>O; for solubility: stepwise adjust to pH 7-8 with 10 M NaOH.

<sup>61</sup> Solve in Tris, pH can be anything as long as no protein is involved; avoid buffers containing phosphates to not reduce solubility. Even as little as 5 mM may be used: less Ni will be bound and, hence, the column will be very specific.

<sup>62</sup> Here, the sample is the entire oligo-attachment reaction.

<sup>63</sup> If the protein's pI supports it, here, up to pH 8 may be considered for high His-Tag affinity.

buffer; wash away unbound oligonucleotides: 10 CV 2 M potassium chloride<sup>64</sup> (KCl); optional: 10 CV sample buffer; 10 CV SEC buffer<sup>65</sup>.

#### *Elute and separate by size*

Elution of a sample bound to a Ni-NTA column coupled to a SEC column. Mount Ni-NTA column before SEC column<sup>66</sup>; inject 0.5 mL of 0.5 M Imidazole buffer<sup>67</sup> as 'pseudo'-sample into the 0.5 mL-loop of the equilibrated<sup>68</sup> HPLC system; start SEC run: 60 min run length, 0.5 mL/min flow-rate, and 30 s time intervals for collecting fractions are typical settings.

#### A.3.4 DNA-handle PCR

The two types of DNA-handles that are able to either bind to streptavidin or to anti-digoxigenin coated beads, are produced in one single PCR on a  $\lambda$ -phage DNA template. The single reaction is accomplished by using a 1:1 ratio of two different sense primers<sup>69</sup> carrying either three biotin or three digoxigenin sites, other than that their sequences are exactly the same. The anti-sense primer is the same for both types of handles and has an abasic site in the middle, where the polymerase falls off. The remaining ss-overhang is complementary to the oligos reacted to the protein, see Section A.3.2. This way the DNA-handles can hybridize to the protein-oligo construct after mixing<sup>70</sup>.

Bold letters in the sense primer's sequence mark the position of the biotin or digoxigenin modifications. The position of the abasic site within the anti-sense primer is marked with a bold x. The gray sequence is complementary to the protein-oligo.

SENSE PRIMER. 5'-GGCGATCTGGTCGTTGATTG-3'

ANTI-SENSE PRIMER. 5'-CGACTCGCTGGTCTGGTTGAACGTCAGCCCTGCCxCTGCCCGGCTCTGGACAGG-3'

The primers bind to a region with especially high GC-content, i. e., 61 % for the 511 nucleotide bases ranging from number 10557 to 11067 of  $\lambda$ -phage DNA. Together with the additional 34 bases from the overhang, two handles yield an overall expected contour length of approximately 371 nm<sup>71</sup>.

64 Solved in sample buffer; adapt pH to be protein- + HisTag-friendly.

65 Usually PBS pH 7.4.

66 The Ni-NTA column's additional volume yields a corresponding time shift in SEC runs.

67 Solve imidazole in appropriate buffer and adjust pH.

68 Pass 2 CV of SEC column for equilibration. Superdex 200: 100 min run at 0.5 mL/min flow-rate.

69 All primers for the DNA-handle PCR are purchased from IBA GmbH.

70 Note that this statistically leads to 50% of all trapping constructs having two DNA-handles with the same functionalization. These constructs will most likely bind to a single bead with the respective coating.

71 Here, 0.34 nm contour length per base pair are assumed. This exactly yields 371.28 or 370.6 nm when including or excluding the abasic site upon counting.

Reagent	Volume [ $\mu\text{L}$ ]
ddH <sub>2</sub> O	232
ThermoPol buffer	30
dNTP	6
$\lambda$ -phage DNA	6
sense primer <i>biotin</i>	5
sense primer <i>digoxigenin</i>	5
anti-sense primer	10
<i>Taq</i> Polymerase	6

Table 20: Reaction premix for DNA-handle PCR.

Step	Temperature [C°]	Time [min : sec]
1	95	2:00
2	95	0:25
3	55	0:45
4	68	0:37
5	go to step 2	repeat 44x
6	68	5:00
7	4	hold

Table 21: Temperature step protocol for DNA-handle PCR.

### PCR protocol details

All primers for the DNA-handle PCR were ordered from IBA. Lambda DNA (N<sup>6</sup>-methyladenine-free) was used as template (N3013S, NEB). The PCR is conducted using the *Taq* DNA Polymerase with ThermoPol Buffer (Mo267S, NEB) in a Routine PCR, i. e., an extra annealing step is included in contrast to a Routine 2-Step PCR. The total 300  $\mu\text{L}$  reaction premix itemized in Table 20 is equally split into 6 standard PCR tubes. The temperature<sup>72</sup> step protocol for the PCR is given in Table 21. The PCR-products are purified with the QIAquick PCR Purification Kit (Qiagen).

### DNA-handle test

The DNA-handles' ability to bind to the protein-bound oligos can be easily tested. After reaction with dimerized oligonucleotides, an agarose gel electrophoresis is run including a pure DNA-handle reference band. A band at about double the size of a single handle should show up in case of working

<sup>72</sup> Melting temperatures for primers can be calculated using the Oligo Calc tool: <http://www.basic.northwestern.edu/biotools/oligocalc.html>

handles; also see [Figure 74](#) for the resulting gel of a protein-oligo construct test with identical principle. For a better interpretation of the test, it is recommended to include an additional reference consisting of DNA-handles that are known to work.

Dimerized oligonucleotides mimicking a protein-oligo construct can be obtained from DTT- or BM(PEG)<sub>2</sub>-tests for maleimide or thiol reactivity, thiol oligo reactions with protein, and maleimide-oligo dimer-like contaminations.

### *Troubleshooting*

According to its specification, the QIAquick PCR Purification Kit removes all unreacted primers or fragments smaller than 40mers from a PCR-product. As the anti-sense primer is a 54mer, unreacted primers can remain in the final elution. This is critical as the anti-sense primer includes the sequence complementary to the protein-bound oligo and, hence, can compete against the much slower diffusing DNA-handles when mixed with the protein-oligo construct. Subsequent removal of unreacted primers can be achieved using the QIAquick Gel Extraction Kit (Qiagen) after running the purified DNA-handles through an agarose gel.

To minimize chances of having unreacted primers, an increased number of 44 repeated PCR-cycles has been chosen. Further the amount of anti-sense primers was kept slightly below the sum of both sense primers.

As primers are delivered on different days and aliquoted with different pipettes by different people, aliquot volumes should be reviewed to guarantee the same amount of the two different sense primers.

#### A.3.5 *Sample preparation and measurement at the trap*

All measurements were performed in filtered and degassed PBS buffer with pH 7.4 (P4417, Sigma Aldrich).

#### *Sample chamber assembly*

To assemble a sample chamber, two stripes of Parafilm M (P7793, Sigma Aldrich) are placed on a bigger cover slip (LH25.1, Carl Roth) such that a channel along the longer axis is formed. A smaller cover slip (LH22.1, Carl Roth) is placed on top of the channel. Next, the sample chamber arrangement is placed on a heat plate preheated to 85°C and the smaller cover slip is slightly pressed down using the cleaned<sup>73</sup> edge of a microscope slide. The ready-to-use sample chamber is removed from the heat plate. Prior filling with the final measurement mix (see below for its preparation), the channel is being incubated with about 40 µL BSA buffer (10 mg/mL BSA (A0281, Sigma Aldrich) solved in PBS and filtered) for at least 10 minutes to passivate the glass surfaces of the channel. After incubation, the channel is flushed with 100 µL of PBS buffer. Now the final mix can be filled into the channel which

<sup>73</sup> Note that prior use, all slips and slides are manually cleaned with Kimwipes (7551, Kimtech Science by Kimberly-Clark).

is thereafter being sealed with vacuum grease (Baysilone-Paste of medium viscosity, GE Bayer Silicones).

#### *Final measurement mix*

The necessary components for the final measurement mix are DNA-handles (for preparation see Section A.3.4), the protein-oligo construct (for preparation see Section A.3.2), PBS measurement buffer, commercially available streptavidin-coated silica beads (CS<sub>01N</sub> 1 μm, Bangs Laboratories), in-house functionalized anti-digoxigenin-coated and fluorescently labelled silica beads (anti-digoxigenin: Anti-digoxigenin Fab fragments, 11214667001, Roche; fluorescent label: albumin from bovine serum (BSA), tetramethylrhodamine (TMR) conjugate, A23016, Molecular Probes as part of Thermo Fischer Scientific; beads: SC<sub>03N</sub> 1 μm, Bangs Laboratories) and the oxygen scavenging system GODCAT (26 U/mL glucose oxidase (G2133, Sigma Aldrich), 1.700 U/mL catalase (C3155, Sigma Aldrich) and 0.6 % (w/v) glucose (G8270, Sigma Aldrich)).

The amounts given for the composition of GODCAT refer to the final measurement mix and are deduced from [259]. The amount of glucose matches well with the v/v values given in [123, 258]. Instead of PIPES buffer, MOPS has been used without addition of magnesium and EDTA. All three components of GODCAT were prepared as 50 x solutions, aliquoted, and stored<sup>74</sup> at -20°C. Therefore, glucose oxidase and catalase solutions contained 50 % (v/v) glycerol.

An example for pipetting a final measurement mix is given in Table 22 and explained in detail in this paragraph. First, the final protein-DNA chimeras have to be formed by letting the DNA-handles (H) hybridize to the protein-oligo construct (P) in an optimized ratio (H:P usually 1:2 up to 1:10) during an optimized time (ranging from 10 minutes to 2 hours as well as overnight reactions) on either ice or at RT. During and/or after the hybridization reaction, additional measurement buffer (B) is added for further dilution; this results in the PHB mix. Next, 0.3 μL of PHB are added to 40 μL of B and well mixed. Then 0.5 μL of a 1:10 dilution<sup>75</sup> of commercially available streptavidin beads (cS<sub>10</sub>) in measurement buffer B are added to react for 7 minutes at RT; this results in the cS<sub>PHB</sub><sup>\*</sup> mix (the reaction time highly depends on the binding affinity of the beads; for low binding affinities also overnight reactions on a rotor in a cold room may be considered). In parallel 0.5 μL of a 1:10 dilution of anti-digoxigenin beads (A<sub>10</sub>) in measurement buffer B are added to an extra 40 μL of B; this results in the A<sup>\*</sup> mix (similar treatment of both bead types showed better reproducibility of amounts of beads in the final mixes, especially in the case of multiple sample chambers per day). The final mix is prepared such that a final volume of 100 μL is reached. First, 3 times 2 μL of

74 As claimed by the supplier Calbiochem, stock solutions of catalase are stable for 1 month at -20°C. Hence, stocks need to be refreshed accordingly.

75 Usually, both types of beads are kept in private 1:10 dilutions that can be used over months. Hence, dilutions help saving stock solutions and keeping the stocks clean through less frequent usage. While dilutions in PBS were observed to be stable for months, dilutions in Tris seem to be stable for a few weeks only.

Step	Mixed amounts [ $\mu\text{L}$ ]	Comment
1.1	5 P + 1 H + 0 B	2 hours at RT
1.2	+ 4 B = PHB	optional dilution
2.1	40 B + 0.3 PHB	mix well
2.2	+ 0.5 cS <sub>10</sub> = cS <sub>PHB</sub> *	mix well; 7 minutes at RT
3	40 B + 0.5 A <sub>10</sub> = A*	in parallel to step 2.2
4	84 B + 3 x 2 GODCAT = F	mix well; start before 2.2 ends
5	F + 8 cS <sub>PHB</sub> * + 2 A* = F*	mix well

Table 22: Final measurement mix example as applied for HP35st-TS.

the 50 x stock solutions of the GODCAT are added to 84  $\mu\text{L}$  of B and well mixed; this results in the F mix. Second, 8  $\mu\text{L}$  of cS<sub>PHB</sub>\* and 2  $\mu\text{L}$  of A\* are added to F and mixed; this results in the F\* mix. The apparent uneven ratio of cS<sub>PHB</sub>\* and A\* is due to uneven bead concentrations and needs to be adapted whenever a new bead stock is being used. Now, the final measurement mix is ready to be filled into the sample chamber.

#### Measurement procedure

After mounting the sample chamber between the two water immersion objectives (use immersion medium Immersol W 2010, Carl Zeiss), force spectroscopic measurements can begin at the warmed up<sup>76</sup> setup shown in Figure 5. At the beginning, the two trap centres are kept apart at an initial distance of about 8  $\mu\text{m}$ . Now, two differently functionalized beads that can be distinguished by fluorescence are caught within the two traps. After recording their power spectra for calibration according to [225], the beads are brought in close proximity to each other by moving the mobile trap at constant velocity towards the fixed trap until both beads slightly touch each other. During this initial approach<sup>77</sup> no tether is yet established and, hence, the measured signals serve as a trap distance dependant baseline for zero (tether-induced) force<sup>78</sup>, for details see [212]. After the beads have touched, the mobile trap is again moved away from the fixed trap and repetitive stretch-and-relax cycles are performed. If a tether forms, first, its integrity is tested and then data at desired constant (mobile trap) velocities are being collected. Typical stretch-and-relax cycles resulting from constant velocity measurements are shown in

<sup>76</sup> Warming up takes about an hour with the trapping laser switched on. During this time a filled and sealed (dummy) sample chamber is used to let the detection side warm up as well.

<sup>77</sup> Local correlations between the two traps' signals ideally allow to exactly assign the point of zero tether extension, i. e., the inter-trap distance at which the surfaces of the two beads start touching [212]. This point coincides with the inter-trap distance at which the apparent force quite suddenly drops in direction of smaller distances while the signals' noise gets reduced at the same time.

<sup>78</sup> The small apparent forces that are detected in the baseline are most likely originating from crosstalk due to the proximity of the two laser foci and eventually they may even reflect hydrodynamic interactions between the two beads [147].

**Figure 10.** Next, various constant trap distances are stepped through to perform equilibrium measurements at a constant force bias. Typical equilibrium data measured in constant distance mode are shown in **Figure 11**. During extremely long equilibrium measurements, a few short constant velocity cycles are run between each constant trap distance step to not inadvertently miss eventual drift during the measurement. Even data containing only small drift have to be handled with care. If all data were collected and the measurement was neither terminated by a breaking tether nor by a third bead accidentally falling into one of the traps, the tether was ripped apart. After a couple of empty cycles marking the end of a successful measurement, both beads were calibrated (ideally for a second time) after the traps were moved back to their initial distance of about 8  $\mu\text{m}$ . If the calibration values lie within the expectable range, the measurement was not corrupted by small nearly invisible particles eventually falling into the trap during data recording.

#### *Troubleshooting - influences on successful tether formation*

The important factors influencing successful tether formation can be split into two categories. One category concerns the individual quality and performance of the used components, namely the protein-oligo construct, the DNA-handles, both types of beads and the oxygen scavenging system<sup>79</sup>. The second category applies to the interplay of all components, i. e., the use of the right amounts, concentrations and reaction times for mixing the final measurement mix.

For creation of the final protein-DNA chimeras, optimal amounts of protein-oligo construct and DNA-handles as well as the reaction time used in step 1.1 of **Table 22** can be screened with the protein-oligo construct test through agarose gel electrophoresis as described at the end of **Section A.3.2**. This test presumes correct function of the DNA-handles, which can be independently tested with dimerized oligonucleotides mimicking the protein-oligo construct according to **Section A.3.4**. Nevertheless, single molecule measurements may still be possible even if the protein-oligo construct test did not show a detectable positive result on an agarose gel.

In step 2, the order of first diluting PHB in B and then adding the beads is very important. If the wrong order is used, concentrated protein-DNA chimeras will be locally added by pipetting into well diluted beads. This results in a lot of constructs being bound to very few beads while most of the beads will not have any construct bound at all.

The most important way to find the right concentrations for incubating protein-DNA chimeras with functionalized beads in step 2.2, is to 'read' and interpret the experiments. Therefore, statistics<sup>80</sup> of observed events for all tested bead pairs are collected. These events include: nothing (N; no tether formation at all), sticking (S; beads would completely stick together and even-

<sup>79</sup> Uncontaminated, filtered and degassed measurement buffer as well as the purity of any other additive, e. g., nucleotides, are assumed.

<sup>80</sup> When a component was changed, at least 25 entirely different pairs of beads per sample chamber were tested to establish relevant statistics.

tually rip apart at very high forces), DNA only (D; tether formation with the correct contour length of DNA-handles but without any protein unfolding), too much DNA (2mD; tethers with too short contour length without any protein unfolding), protein (P; the desired protein fingerprint is there), too much protein (2mP; too short tether with protein-like unfolding events usually of a different way from the expected fingerprint), unzipping (U; a plateau region around 15 pN as known from DNA(-hairpin) unzipping [70, 183]), weird and long tethers (wLT; tethers with too long contour length and indefinable unfolding events with little or no hysteresis), half tethers (hT; D or P events with half of the expected contour length for the DNA-handles), DNA-handle crap (aka crappy linker: cL; ~2 to 12 nm contour length flipping events localized or distributed over the entire force range and protein-like unfolding/refolding events repeatedly showing considerable hysteresis), and ripping tethers (rT; P events, where tethers rip off at forces below 30 pN after only a very few stretch-and-relax cycles).

A good rule of thumb is to tune step 2 towards a probability of tether formation equal to about one third. In this case, the probability of having a single tether, i. e., the desired single molecule event P, equals approximately 80% [146]. This can be derived upon assumption of a Poisson distribution of protein-DNA chimeras randomly attached to the bead similar to the attachment of molecular motors to beads in *in vitro* assays [24, 186]. The given rule is not that strict though, as there are more indications for a true single molecule event, e. g., the contour lengths of the DNA-handles and of the protein unfolding as well as the measured kinetics.

If the number of events N is above two thirds of all events, too few protein-DNA chimeras have bound to the beads during incubation in step 2.2. Hence, the amount of PHB needs to be increased in step 2.1 and/or the incubation time prolonged in step 2.2. For rather short incubation times where diffusion will play an important role, one might also consider decreasing the overall reaction volume by decreasing the amount of buffer B in step 2.1. In case that the events S, 2mD and 2mP dominate, obviously too many tethers are forming at the same time and the amount of PHB and/or its reaction time with the beads need/s to be decreased.

A high amount of D events usually originates from a contamination with dimerized oligonucleotides. How to get rid of this contamination is described in Section A.3.3.

Observation of unzipping (U) events are an indication of too many DNA-handles being used with respect to the protein-oligo construct. In contrast, too few DNA-handles may cause the occurrence of hT events with only half of the expected contour length. The latter can also show up because of protein-oligo constructs that have only one oligo attached. DNA-handle crap (cL) points towards deficient handles and a new DNA-handle PCR is highly recommended.

Weird long tethers (wLT) can be observed when beads are overloaded with protein and/or DNA-handles. An explanation could be that the formation of protein-DNA chimeras is highly inefficient such that a lot of PHB is applied in step 2. This further leads to unspecific unfolding events with too long contour

lengths instead of sticking which is the usual observation for application of too much PHB.

The frequent rupture of tethers (rT) at forces below 30 pN can have multiple reasons. In brief, either the oxygen scavenging system has gone bad, or the functionalized DNA-handle primers or the handles themselves are not in best shape, or the binding to the beads is somehow corrupted. To exclude the latter, a fresh preparation of new bead stocks surely is one option<sup>81</sup>. Besides, one might also consider to simply wash the beads to get rid of gradually detached streptavidin or anti-digoxigenin from the storage buffer to not interfere during incubation in step 2.2.

In principle there is no clear preference for the type of beads being used to incubate with the protein-DNA chimeras in step 2.2. Here, the commercially available beads with a more reproducible reactivity were preferred. If the other bead type is less reactive, more repeats per bead pair will be needed upon trying to establish a tether. This can be inverted by letting a higher concentration of protein-DNA chimeras react with the less reactive beads over a longer time (preferably on a rotor). Picking up a tether will be possible with fewer attempts provided that in step 2.2 a few protein-DNA chimeras have reacted to the beads at all.

To end up with about the same amount of each bead type in the final mix F\* after step 5, it is helpful to estimate the relative concentrations of their initial 1:10 stock dilutions. This is achieved best by making a 1:1000 dilution and directly filling it into an unpassivated sample chamber. In this chamber, the beads will stick to the glass surface and an average number of beads per bright field screen can be counted. From this number the necessary ratio for the final mix can be derived. Another way would be to make statistics of the beads encountered during a measurement and to adjust their amounts accordingly for the next sample. This presumes perfect distinction through the fluorescent label as well as perfect mixing<sup>82</sup> during all the steps of creating the final measurement mix. So far, the best distinction between different types of beads was possible through slightly differently sized beads.

Good mixing between each step is crucial and achieved best by tapping the reaction tubes; mostly, vortexing (more foam) or pipetting up and down (additional shear stress; consider cutting off a little bit from the pipette tip for a bigger opening) is fine for mixing as well but it was passed down to be harmful to the activity of sensitive proteins such as molecular motors. One can slightly shake the pipette to initially better spread the beads upon addition to solution. Directly prior to pipetting beads, their solution always needs to be (re-)mixed.

81 A test to quantify the binding ability of beads is presented in the supplement of [123].

82 Perfect mixing of beads has not been observed during this thesis; the two types of beads would rather show up in larger areas of one type. Concerning the fluorescent labelling, its homogeneity and intensity have shown quite some variability within one stock.

To assure good reproducibility of pipetted amounts, always use the same pipettes and avoid volumes of less than  $0.5 \mu\text{L}$ <sup>83</sup> unless very precise pipettes are available.

---

<sup>83</sup> The correct amount of  $0.5 \mu\text{L}$  can still be estimated quite well by checking the volume in a standard  $10 \mu\text{L}$  pipette tip by eye.

## BIBLIOGRAPHY

---

- [1] C. E. Aitken, R. A. Marshall, and J. D. Puglisi. An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. *Biophysical Journal*, 94:1826–1835, 2008. (Cited on page 19.)
- [2] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–1314, 1961. (Cited on page 3.)
- [3] J. M. Antos, G. L. Chew, C. P. Guimaraes, N. C. Yoder, G. M. Grotenbreg, M. W. L. Popp, and H. L. Ploegh. Site-specific N- and C-terminal labeling of a single polypeptide using sortases of different specificity. *Journal of the American Chemical Society*, 131:10800–10801, 2009. (Cited on page 24.)
- [4] S. A. Arrhenius. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für Physikalische Chemie*, 4:226–248, 1889. (Cited on pages 26 and 41.)
- [5] A. Ashkin. Acceleration and trapping of particles by radiation pressure. *Physical Review Letters*, 24:156–159, 1970. (Cited on page 15.)
- [6] A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters*, 11:288–290, 1986. (Cited on page 15.)
- [7] A. Ashkin, J. M. Dziedzic, and T. Yamane. Optical trapping and manipulation of single cells using infrared laser beams. *Nature*, 330:769–771, 1987. (Cited on page 15.)
- [8] M. Atakhorrami, K. M. Addas, and C. F. Schmidt. Twin optical traps for two-particle cross-correlation measurements: eliminating cross-talk. *Review of Scientific Instruments*, 79:43103—1–5, 2008. (Cited on page 17.)
- [9] K. Austen, P. Ringer, A. Mehlich, A. Chrostek-Grashoff, C. Kluger, C. Klingner, B. Sabass, R. Zent, M. Rief, and C. Grashoff. Extracellular rigidity sensing by talin isoform-specific mechanical linkages. *Nature Cell Biology*, 17:1597–1606, 2015. (Cited on pages 35, 117, 139, and 140.)
- [10] M. Balsera, S. Stepaniants, S. Izrailev, Y. Oono, and K. Schulten. Reconstructing potential energy functions from simulated force-induced unbinding processes. *Biophysical Journal*, 73:1281–1287, 1997. (Cited on page 47.)

- [11] J. M. Baskin, J. A. Prescher, S. T. Laughlin, N. J. Agard, P. V. Chang, I. A. Miller, A. Lo, J. A. Codelli, and C. R. Bertozzi. Copper-free click chemistry for dynamic in vivo imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 104:16793–16797, 2007. (Cited on page 24.)
- [12] D. Bauer, D. R. Merz, B. Pelz, K. E. Theisen, G. Yacyshyn, D. Mokranjac, R. I. Dima, M. Rief, and G. Žoldák. Nucleotides regulate the mechanical hierarchy between subdomains of the nucleotide binding domain of the Hsp70 chaperone DnaK. *Proceedings of the National Academy of Sciences of the United States of America*, 112:10389–10394, 2015. (Cited on pages 26, 28, and 33.)
- [13] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970. (Cited on page 39.)
- [14] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 94:6185–6190, 1997. (Cited on page 15.)
- [15] G. I. Bell. Models for the specific adhesion of cells to cells. *Science*, 200:618–627, 1978. (Cited on pages 28 and 42.)
- [16] A. M. Berezhkovskii, A. Szabo, N. Greives, and H.-X. Zhou. Multidimensional reaction rate theory with anisotropic diffusion. *The Journal of Chemical Physics*, 141:204106—1–6, 2014. (Cited on page 89.)
- [17] K. Berg-Sørensen and H. Flyvbjerg. The colour of thermal noise in classical Brownian motion: a feasibility study of direct experimental observation. *New Journal of Physics*, 7:010038—1–10, 2005. (Cited on page 65.)
- [18] F. Berkemeier, M. Bertz, S. Xiao, N. Pinotsis, M. Wilmanns, F. Gräter, and M. Rief. Fast-folding alpha-helices as reversible strain absorbers in the muscle protein myomesin. *Proceedings of the National Academy of Sciences of the United States of America*, 108:14139–14144, 2011. (Cited on pages 38 and 137.)
- [19] F. G. Berkemeier. *Kraftspektroskopische Untersuchungen der Myomesin-Elastizität*. PhD thesis, Technische Universität München, 2011. (Cited on pages 31, 33, 34, 37, 38, 40, and 137.)
- [20] R. B. Best and G. Hummer. Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences of the United States of America*, 102:6732–6737, 2005. (Cited on page 29.)

- [21] R. B. Best and G. Hummer. Coordinate-dependent diffusion in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 107:1088–1093, 2010. (Cited on page 46.)
- [22] Y. Bi, J. H. Cho, E. Y. Kim, B. Shan, H. Schindelin, and D. Raleigh. Rational design, structural and thermodynamic characterization of a hyperstable variant of the villin headpiece helical subdomain. *Biochemistry*, 46:7497–7505, 2007. (Cited on page 216.)
- [23] O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello, and T. Kiefhaber. The speed limit for protein folding measured by triplet-triplet energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 96:9597–9601, 1999. (Cited on page 41.)
- [24] S. M. Block, L. S. Goldstein, and B. J. Schnapp. Bead movement by single kinesin molecules studied with optical tweezers. *Nature*, 348:348–352, 1990. (Cited on page 244.)
- [25] U. Bockelmann, B. Essevaz-Roulet, and F. Heslot. Molecular stick-slip motion revealed by opening DNA with piconewton forces. *Physical Review Letters*, 79:4489–4492, 1997. (Cited on page 219.)
- [26] P. G. Bolhuis. Transition-path sampling of beta-hairpin folding. *Proceedings of the National Academy of Sciences of the United States of America*, 100:12129–34, 2003. (Cited on page 54.)
- [27] P. G. Bolhuis and W. Lechner. On the relation between projections of the reweighted path ensemble. *Journal of Statistical Physics*, 145:841–859, 2011. (Cited on page 57.)
- [28] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry*, 53:291–318, 2002. (Cited on pages 29, 52, 53, and 54.)
- [29] T. Bornschlög. *Gleichgewichtsmechanik und Fluktuationen von Superhelices*. PhD thesis, Technische Universität München, 2008. (Cited on page 31.)
- [30] T. Bornschlög and M. Rief. Single molecule unzipping of coiled coils: sequence resolved stability profiles. *Physical Review Letters*, 96:118102–1–4, 2006. (Cited on pages 136 and 219.)
- [31] T. Bornschlög, G. Woehlke, and M. Rief. Single molecule mechanics of the kinesin neck. *Proceedings of the National Academy of Sciences of the United States of America*, 106:6992–6997, 2009. (Cited on page 224.)
- [32] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84:7524–7528, 1987. (Cited on page 6.)

- [33] C. Bustamante, J. F. Marko, E. D. Siggia, and S. B. Smith. Entropic Elasticity of  $\lambda$ -Phage DNA. *Science*, 265:1599–1600, 1994. (Cited on page 31.)
- [34] C. Bustamante, Y. Chemla, and J. Moffitt. High-Resolution dual-trap optical tweezers with differential detection. In P. R. Selvin and T. Ha, editors, *Single-Molecule techniques: a laboratory manual*, chapter 14, pages 297–324. Cold Spring Harbor Laboratory Press, New York, 2008. (Cited on page 17.)
- [35] Y. Cao and H. Li. Polyprotein of GB1 is an ideal artificial elastomeric protein. *Nature Materials*, 6:109–114, 2007. (Cited on pages 21 and 212.)
- [36] Y. Cao, C. Lam, M. Wang, and H. Li. Nonmechanical protein can have significant mechanical stability. *Angewandte Chemie International Edition*, 45:642–645, 2006. (Cited on page 134.)
- [37] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez. Mechanical and chemical unfolding of a single protein: a comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 96:3694–3699, 1999. (Cited on page 134.)
- [38] A. R. Carter, Y. Seol, and T. T. Perkins. Precision surface-coupled optical-trapping assay with one-basepair resolution. *Biophysical Journal*, 96:2926–2934, 2009. (Cited on pages 15 and 16.)
- [39] C. Cecconi, E. A. Shank, F. W. Dahlquist, S. Marqusee, and C. Bustamante. Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers. *European Biophysics Journal*, 37:729–738, 2008. (Cited on pages 19 and 226.)
- [40] T. Cellmer, E. R. Henry, J. Hofrichter, and W. A. Eaton. Measuring internal friction of an ultrafast-folding protein. *Proceedings of the National Academy of Sciences of the United States of America*, 105:18320–18325, 2008. (Cited on pages 41, 45, 46, and 200.)
- [41] S. Chaudhury and D. E. Makarov. A harmonic transition state approximation for the duration of reactive events in complex molecular rearrangements. *The Journal of Chemical Physics*, 133:034118—1–11, 2010. (Cited on page 59.)
- [42] J. D. Chodera and V. S. Pande. Splitting probabilities as a test of reaction coordinate choice in single-molecule experiments. *Physical Review Letters*, 107:08102—1–5, 2011. (Cited on pages 28, 53, 54, 107, and 108.)
- [43] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, and N. S. Hinrichs. Bayesian hidden Markov model analysis of single-molecule force spectroscopy: characterizing kinetics under measurement uncertainty. *arXiv.org*, arXiv:1108.1430 [cond-mat.stat-mech], 2011. (Cited on page 40.)

- [44] H. S. Chung, J. M. Louis, and W. A. Eaton. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proceedings of the National Academy of Sciences of the United States of America*, 106:11837–11844, 2009. (Cited on pages 3, 45, 46, 57, and 78.)
- [45] H. S. Chung, K. McHale, J. M. Louis, and W. a. Eaton. Single-molecule fluorescence experiments determine protein folding transition path times. *Science*, 335:981–984, 2012. (Cited on pages 4, 7, and 205.)
- [46] C. Cohen and D. A. D. Parry.  $\alpha$ -helical coiled coils - a widespread motif in proteins. *Trends in Biochemical Sciences*, 11:245–248, 1986. (Cited on page 219.)
- [47] I. Coluzza. Computational protein design. *Journal of Physics: Condensed Matter*, 29:143001—1–16, 2017. (Cited on page 8.)
- [48] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players. Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760, 2010. (Cited on page 8.)
- [49] P. Cossio, G. Hummer, and A. Szabo. On artifacts in single-molecule force spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 112:14248–14253, 2015. (Cited on page 89.)
- [50] F. H. C. Crick. The packing of  $\alpha$ -helices: simple coiled-coils. *Acta Crystallographica*, 6:689–697, 1953. (Cited on page 219.)
- [51] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60:2721–2726, 1999. (Cited on page 26.)
- [52] A. Czövek, G. J. Szöllösi, and I. Derényi. Neck-Linker docking coordinates the kinetics of kinesin’s heads. *Biophysical Journal*, 100:1729–1736, 2011. (Cited on page 222.)
- [53] R. Dawkins. *The blind watchmaker: why the evidence of evolution reveals a universe without design*. W. W. Norton & Company, Inc., New York, 1986. (Cited on page 4.)
- [54] M. de Cuevas, T. Tao, and L. S. B. Goldstein. Evidence that the stalk of *Drosophila* kinesin heavy chain is an alpha-helical coiled coil. *Journal of Cell Biology*, 116:957–965, 1992. (Cited on page 218.)
- [55] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *The Journal of Chemical Physics*, 108:1964–1977, 1998. (Cited on pages 29 and 52.)
- [56] J. E. J. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Society for Industrial and Applied Mathematics, 1996. (Cited on page 50.)

- [57] R. J. Diefenbach, E. Diefenbach, M. W. Douglas, and A. L. Cunningham. The heavy chain of conventional kinesin interacts with the SNARE proteins SNAP25 and SNAP23. *Biochemistry*, 41:14906–14915, 2002. (Cited on page 218.)
- [58] K. A. Dietrich, C. V. Sindelar, P. D. Brewer, K. H. Downing, C. R. Cremona, and S. E. Rice. The kinesin-1 motor protein is regulated by a direct interaction of its head and tail. *Proceedings of the National Academy of Sciences of the United States of America*, 105:8938–8943, 2008. (Cited on page 218.)
- [59] H. Dietz and M. Rief. Exploring the energy landscape of GFP by single-molecule mechanical experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 101:16192–16197, 2004. (Cited on pages 26, 32, 41, and 134.)
- [60] H. Dietz and M. Rief. Protein structure by mechanical triangulation. *Proceedings of the National Academy of Sciences of the United States of America*, 103:1244–1247, 2006. (Cited on pages 26 and 32.)
- [61] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4:10–19, 1997. (Cited on page 5.)
- [62] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042–1046, 2012. (Cited on pages 3, 5, and 6.)
- [63] A. J. Doig and R. L. Baldwin. N- and C-capping preferences for all 20 amino acids in  $\alpha$ -helical peptides. *Protein Science*, 4:1325–1336, 1995. (Cited on pages 166 and 214.)
- [64] O. K. Dudko, A. E. Filippov, J. Klafter, and M. Urbakh. Beyond the conventional description of dynamic force spectroscopy of adhesion bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 100:11378–11381, 2003. (Cited on page 147.)
- [65] O. K. Dudko, G. Hummer, and A. Szabo. Intrinsic rates and activation free energies from single-molecule pulling experiments. *Physical Review Letters*, 96:108101—1–4, 2006. (Cited on pages 37, 45, and 47.)
- [66] O. K. Dudko, J. Mathé, A. Szabo, A. Meller, and G. Hummer. Extracting kinetics from single-molecule force spectroscopy: nanopore unzipping of DNA hairpins. *Biophysical Journal*, 92:4188–4195, 2007. (Cited on page 131.)
- [67] O. K. Dudko, G. Hummer, and A. Szabo. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 105:15755–15760, 2008. (Cited on pages 28, 37, and 48.)
- [68] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322:549–560, 1905. (Cited on page 46.)

- [69] G. L. Ellman. Tissue sulfhydryl groups. *Archives of Biochemistry and Biophysics*, 82:70–77, 1959. (Cited on page 22.)
- [70] B. Essevaz-Roulet, U. Bockelmann, and F. Heslot. Mechanical separation of the complementary strands of DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 94:11935–11940, 1997. (Cited on page 244.)
- [71] E. Evans and K. Ritchie. Dynamic strength of molecular adhesion bonds. *Biophysical Journal*, 72:1541–1555, 1997. (Cited on page 47.)
- [72] E. Evans and K. Ritchie. Strength of a weak bond connecting flexible polymer chains. *Biophysical Journal*, 76:2439–2447, 1999. (Cited on page 47.)
- [73] M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Transactions of the Faraday Society*, 31:875–894, 1935. (Cited on page 7.)
- [74] H. Eyring. The Activated Complex in Chemical Reactions. *The Journal of Chemical Physics*, 3:107–115, 1935. (Cited on page 7.)
- [75] E. Fällman and O. Axner. Design for fully steerable dual-trap optical tweezers. *Applied Optics*, 36:2107–2113, 1997. (Cited on page 16.)
- [76] J. Fang, A. Mehlich, N. Koga, J. Huang, R. Koga, X. Gao, C. Hu, C. Jin, M. Rief, J. Kast, D. Baker, and H. Li. Forced protein unfolding leads to highly elastic and tough protein hydrogels. *Nature Communications*, 4:2974–2983, 2013. (Cited on pages 115, 116, 138, and 211.)
- [77] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Physical Review Letters*, 63:1195–1198, 1989. (Cited on page 50.)
- [78] A. Fersht. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. W. H. Freeman and Company, New York, 2nd edition, 1999. (Cited on page 81.)
- [79] J. T. Finer, R. M. Simmons, and J. A. Spudich. Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, 368:113–119, 1994. (Cited on page 15.)
- [80] I. Forné, J. Ludwigsen, A. Imhof, P. B. Becker, and F. Mueller-Planitz. Probing the conformation of the ISWI ATPase domain with genetically encoded photoreactive crosslinkers and mass spectrometry. *Molecular & Cellular Proteomics*, 11:012088—1–11, 2012. (Cited on page 24.)
- [81] A. Freikamp, A. Mehlich, K. Christoph, and C. Grashoff. Investigating piconewton forces in cells by FRET-based molecular force microscopy. *Journal of Structural Biology*, in press, 2016. (Cited on pages 36 and 139.)

- [82] J. C. M. Gebhardt. *Energielandschaften einzelner Proteine gemessen mit der optischen Pinzette*. PhD thesis, Technische Universität München, 2009. (Cited on pages 17, 19, 33, and 233.)
- [83] J. C. M. Gebhardt, T. Bornschlöggl, and M. Rief. Full distance-resolved folding energy landscape of one single protein molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 107:2013–2018, 2010. (Cited on pages 3, 16, 19, 21, 28, 41, 42, 49, 51, 81, 131, 135, 212, and 219.)
- [84] E. B. Getz, M. Xiao, T. Chakrabarty, R. Cooke, and P. R. Selvin. A comparison between the sulfhydryl reductants tris(2-carboxyethyl)phosphine and dithiothreitol for use in protein biochemistry. *Analytical Biochemistry*, 273:73–80, 1999. (Cited on pages 22 and 23.)
- [85] I. Ghosh, A. D. Hamilton, and L. Regan. Antiparallel leucine zipper-directed protein reassembly: application to the green fluorescent protein. *Journal of the American Chemical Society*, 122:5658–5659, 2000. (Cited on page 217.)
- [86] R. Godoy-Ruiz, E. R. Henry, J. Kubelka, J. Hofrichter, V. Muñoz, J. M. Sanchez-Ruiz, and W. A. Eaton. Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data. *Journal of Physical Chemistry*, 112:5938–5949, 2008. (Cited on page 46.)
- [87] L. S. B. Goldstein. Structural features involved in force generation in the kinesin superfamily. *Biophysical Journal*, 68:260S–265S, 1995. (Cited on page 218.)
- [88] C. Grashoff, B. D. Hoffman, M. D. Brenner, R. Zhou, M. Parsons, M. T. Yang, M. A. McLean, S. G. Sligar, C. S. Chen, T. Ha, and M. A. Schwartz. Measuring mechanical tension across vinculin reveals regulation of focal adhesion dynamics. *Nature*, 466:263–266, 2010. (Cited on pages 139, 140, and 216.)
- [89] W. J. Greenleaf, M. T. Woodside, E. a. Abbondanzieri, and S. M. Block. Passive all-optical force clamp for high-resolution laser trapping. *Physical Review Letters*, 95:208102—1–4, 2005. (Cited on page 27.)
- [90] M. Grummt, G. Woehlke, U. Henningsen, S. Fuchs, M. Schleicher, and M. Schliwa. Importance of a flexible hinge near the motor domain in kinesin-driven motility. *EMBO Journal*, 17:5536–5542, 1998. (Cited on page 218.)
- [91] A. N. Gupta, A. Vincent, K. Neupane, H. Yu, F. Wang, and M. T. Woodside. Experimental validation of free-energy-landscape reconstruction from non-equilibrium single-molecule force spectroscopy measurements. *Nature Physics*, 7:631–634, 2011. (Cited on pages 26 and 50.)

- [92] S. J. Hagen, J. Hofrichter, A. Szabo, and W. a. Eaton. Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 93:11615–11617, 1996. (Cited on pages 6, 66, and 200.)
- [93] K. Hahlen, B. Ebbing, J. Reinders, J. Mergler, A. Sickmann, and G. Woehlke. Feedback of the kinesin-1 neck-linker position on the catalytic site. *Journal of Biological Chemistry*, 281:18868–18877, 2006. (Cited on page 222.)
- [94] S. Hammond. A correlation of reaction rates. *Journal of the American Chemical Society*, 77:334–338, 1955. (Cited on pages 46, 80, and 125.)
- [95] P. Hänggi, P. Talkner, and M. Borkovec. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62:251–341, 1990. (Cited on pages 45 and 89.)
- [96] B. Hellenkamp, P. Wortmann, F. Kandzia, M. Zacharias, and T. Hugel. Multidomain structure and correlated dynamics determined by self-consistent FRET networks. *Nature Methods*, 14:174–180, 2016. (Cited on page 3.)
- [97] G. T. Hermanson. *Bioconjugate Techniques*. Elsevier Inc., Burlington, 2nd edition, 2008. (Cited on page 23.)
- [98] M. Hinczewski, J. C. M. Gebhardt, M. Rief, and D. Thirumalai. From mechanical folding trajectories to intrinsic energy landscapes of biopolymers. *Proceedings of the National Academy of Sciences of the United States of America*, 110:4500–4505, 2013. (Cited on pages 28 and 50.)
- [99] J. Howard. *Mechanics of motor proteins and the cytoskeleton*. Sinauer Associates, Inc., Sunderland, 2001. (Cited on pages 26 and 45.)
- [100] P.-S. Huang, S. E. Boyken, and D. Baker. The coming of age of de novo protein design. *Nature*, 537:320–327, 2016. (Cited on page 8.)
- [101] G. Hummer. From transition paths to transition states and rate coefficients. *The Journal of Chemical Physics*, 120:516–523, 2004. (Cited on pages 7, 29, 56, 57, 59, and 68.)
- [102] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98:3658–3661, 2001. (Cited on pages 26 and 50.)
- [103] G. Hummer and A. Szabo. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophysical Journal*, 85:5–15, 2003. (Cited on page 26.)

- [104] G. Hummer and A. Szabo. Free energy profiles from single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107:21441–21446, 2010. (Cited on page 50.)
- [105] C. Hyeon and J. N. Onuchic. A structural perspective on the dynamics of kinesin motors. *Biophysical Journal*, 101:2749–2759, 2011. (Cited on page 222.)
- [106] C. Hyeon and D. Thirumalai. Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? *Proceedings of the National Academy of Sciences of the United States of America*, 100:10249–10253, 2003. (Cited on pages 41, 46, and 71.)
- [107] C. Hyeon and D. Thirumalai. Forced-unfolding and force-quench refolding of RNA hairpins. *Biophysical Journal*, 90:3410–3427, 2006. (Cited on pages 125 and 135.)
- [108] M. Jahn, A. Rehn, B. Pelz, B. Hellenkamp, K. Richter, M. Rief, J. Buchner, and T. Hugel. The charged linker of the molecular chaperone Hsp90 modulates domain contacts and biological function. *Proceedings of the National Academy of Sciences of the United States of America*, 111:17881–17886, 2014. (Cited on pages 21 and 212.)
- [109] M. Jahn, J. Buchner, T. Hugel, and M. Rief. Folding and assembly of the large molecular machine Hsp90 studied in single-molecule experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:1232–1237, 2016. (Cited on pages 200 and 201.)
- [110] P. A. Jansson. *Deconvolution with applications in spectroscopy*. Academic Press, New York, 1st edition, 1984. (Cited on page 49.)
- [111] E. A. Jares-Erijman and T. M. Jovin. FRET imaging. *Nature Biotechnology*, 21:1387–1395, 2003. (Cited on page 139.)
- [112] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78:2690–2693, 1997. (Cited on page 26.)
- [113] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science*, 276:1112–1116, 1997. (Cited on page 15.)
- [114] J. C. Kendrew, G. Bodo, H. M. Dintzis, G. Parrish, and H. Wyckoff. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958. (Cited on page 3.)
- [115] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montellione, and D. Baker. Principles for designing ideal protein structures. *Nature*, 491:222–227, 2012. (Cited on pages 9, 116, 144, 176, 199, 211, 212, and 213.)

- [116] H. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1940. (Cited on pages 28 and 45.)
- [117] O. Kratky and G. Porod. Röntgenuntersuchung gelöster Fadenmoleküle. *Recueil des Travaux Chimiques des Pays-Bas*, 68:1106–1122, 1949. (Cited on page 31.)
- [118] S. S. Krishna, R. I. Sadreyev, and N. V. Grishin. A tale of two ferredoxins: sequence similarity and structural differences. *BMC Structural Biology*, 6:8, 2006. (Cited on page 9.)
- [119] J. Kubelka, J. Hofrichter, and W. a. Eaton. The protein folding ‘speed limit’. *Current Opinion in Structural Biology*, 14:76–88, feb 2004. (Cited on pages 6, 57, 84, 158, and 182.)
- [120] J. Kubelka, E. R. Henry, T. Cellmer, J. Hofrichter, and W. A. Eaton. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proceedings of the National Academy of Sciences of the United States of America*, 105:18655–18662, 2008. (Cited on page 46.)
- [121] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003. (Cited on pages 8 and 9.)
- [122] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The Weighted Histogram Analysis Method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13:1011–1021, 1992. (Cited on page 50.)
- [123] M. P. Landry, P. M. McCall, Z. Qi, and Y. R. Chemla. Characterization of photoactivated singlet oxygen damage in single-molecule optical trap experiments. *Biophysical Journal*, 97:2128–2136, 2009. (Cited on pages 19, 241, and 245.)
- [124] M. J. Lang, C. L. Asbury, J. W. Shaevitz, and S. M. Block. An automated two-dimensional optical force clamp for single molecule studies. *Biophysical Journal*, 83:491–501, 2002. (Cited on page 27.)
- [125] J. Langer. Statistical theory of the decay of metastable states. *Annals of Physics*, 54:258–275, 1969. (Cited on page 89.)
- [126] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487:545–574, 2011. (Cited on page 8.)

- [127] C. Y. Lee, H. E. Canavan, L. J. Gamble, and D. G. Castner. Evidence of impurities in thiolated single-stranded DNA oligomers and their effect on DNA self-assembly on gold. *Langmuir*, 21:5134–5141, 2005. ISSN 07437463. doi: 10.1021/la0472302. (Cited on page 235.)
- [128] J. E. Leffler. Parameters for the description of transition states. *Science*, 117:340–341, 1953. (Cited on page 135.)
- [129] S. Lepthien, L. Merkel, and N. Budisa. In vivo double and triple labeling of proteins using synthetic amino acids. *Angewandte Chemie International Edition*, 49:5446–5450, 2010. (Cited on page 24.)
- [130] C. Levinthal. How to fold graciously. *Mössbauer Spectroscopy in Biological Systems. Proceedings of a meeting held at Allerton House, Monticello, Illinois*, 67:22–24, 1969. (Cited on page 4.)
- [131] M. O. Lindberg and M. Oliveberg. Malleability of protein folding pathways: a simple reason for complex behaviour. *Current Opinion in Structural Biology*, 17:21–29, 2007. (Cited on pages 79 and 131.)
- [132] F. Liu, M. Nakaema, and M. Gruebele. The transition state transit time of WW domain folding is controlled by energy landscape roughness. *The Journal of Chemical Physics*, 131:195101—1–9, 2009. (Cited on pages 46 and 71.)
- [133] G. V. Los, L. P. Encell, M. G. McDougall, D. D. Hartzell, N. Karassina, C. Zimprich, M. G. Wood, R. Learish, R. F. Ohana, M. Urh, D. Simpson, J. Mendez, K. Zimmerman, P. Otto, G. Vidugiris, J. Zhu, A. Darzins, D. H. Klaubert, R. F. Bulleit, and K. V. Wood. HaloTag: a novel protein labelling technology for cell imaging and protein analysis. *ACS Chemical Biology*, 3:373–382, 2008. (Cited on page 24.)
- [134] A. Lupas, M. van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252:1162–1164, 1991. (Cited on page 219.)
- [135] M. P. Lutolf and J. a. Hubbell. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. *Nature Biotechnology*, 23:47–55, 2005. (Cited on page 115.)
- [136] A. Maitra and G. Arya. Model accounting for the effects of pulling-device stiffness in the analyses of single-molecule force measurements. *Physical Review Letters*, 104:108301, 2010. (Cited on page 49.)
- [137] D. E. Makarov. Shapes of dominant transition paths from single-molecule force spectroscopy. *The Journal of Chemical Physics*, 143:194103—1–5, 2015. (Cited on pages 187 and 204.)
- [138] S. V. Malinin and V. Y. Chernyak. Transition times in the low-noise limit of stochastic dynamics. *The Journal of Chemical Physics*, 132:014504—1–11, 2010. (Cited on pages 57 and 59.)

- [139] S. Maloy, V. Stewart, and R. Taylor. *Genetic analysis of pathogenic bacteria: a laboratory manual*. Cold Spring Harbor Laboratory Press, New York, 1996. (Cited on page 223.)
- [140] P. Mangeol and U. Bockelmann. Interference and crosstalk in double optical tweezers using a single laser source. *Review of Scientific Instruments*, 79:083103—1–8, 2008. (Cited on page 16.)
- [141] M. Manosas and F. Ritort. Thermodynamic and kinetic aspects of RNA pulling experiments. *Biophysical Journal*, 88:3224–3242, 2005. (Cited on page 17.)
- [142] M. Manosas, J. D. Wen, P. T. X. Li, S. B. Smith, C. Bustamante, I. Tinoco, and F. Ritort. Force unfolding kinetics of RNA using optical tweezers. II. Modeling experiments. *Biophysical Journal*, 92:3010–3021, 2007. (Cited on pages 65 and 67.)
- [143] A. P. Manuel, J. Lambert, and M. T. Woodside. Reconstructing folding energy landscapes from splitting probability analysis of single-molecule trajectories. *Proceedings of the National Academy of Sciences of the United States of America*, 112:7183–7188, 2015. (Cited on pages 7, 28, 53, 55, 78, 86, 107, 108, and 183.)
- [144] A. V. McDonnell, T. Jiang, A. E. Keating, and B. Berger. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, 22:356–358, 2006. (Cited on page 219.)
- [145] S. A. McKinney, C. Joo, and T. Ha. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophysical Journal*, 91:1941–1951, 2006. (Cited on page 38.)
- [146] A. Mehlich. *Kurze DNA-Stücke als Rauschunterdrücker in der Kraftspektroskopie mit optischen Fallen*. Diploma thesis, Technische Universität München, 2009. (Cited on pages 17, 19, and 244.)
- [147] J.-C. Meiners and S. R. Quake. Direct measurement of hydrodynamic cross correlations between two particles in an external potential. *Physical Review Letters*, 82:2211–2214, 1999. (Cited on page 242.)
- [148] L. Milanesi, J. P. Waltho, C. a. Hunter, D. J. Shaw, G. S. Beddard, G. D. Reid, S. Dev, and M. Volk. Measurement of energy landscape roughness of folded and unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 109:19563–19568, 2012. (Cited on pages 41, 46, and 200.)
- [149] Y. Miyazono, M. Hayashi, P. Karagiannis, Y. Harada, and H. Tadakuma. Strain through the neck linker ensures processive runs: a DNA-kinesin hybrid nanomachine study. *The EMBO Journal*, 29:93–106, 2010. (Cited on page 218.)

- [150] J. R. Moffitt, Y. R. Chemla, D. Izhaky, and C. Bustamante. Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proceedings of the National Academy of Sciences of the United States of America*, 103:9006–9011, 2006. (Cited on pages 16 and 17.)
- [151] J. Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15:285–289, 2005. (Cited on page 5.)
- [152] T. Nagai, K. Ibata, E. S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnology*, 20:87–90, 2002. (Cited on page 216.)
- [153] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proceedings of the National Academy of Sciences of the United States of America*, 104:2655–2660, 2007. (Cited on page 66.)
- [154] K. Neupane, D. B. Ritchie, H. Yu, D. a. N. Foster, F. Wang, and M. T. Woodside. Transition path times for nucleic acid folding determined from energy-landscape analysis of single-molecule trajectories. *Physical Review Letters*, 109:068102—1–5, 2012. (Cited on pages 7, 46, 58, 78, and 182.)
- [155] K. Neupane, A. P. Manuel, J. Lambert, and M. T. Woodside. Transition-path probability as a test of reaction-coordinate quality reveals DNA hairpin folding is a one-dimensional diffusive process. *The Journal of Physical Chemistry Letters*, 6:1005–1010, 2015. (Cited on pages 7, 29, 103, 104, 107, 108, and 162.)
- [156] K. Neupane, D. A. N. Foster, D. R. Dee, H. Yu, F. Wang, and M. T. Woodside. Direct observation of transition paths during the folding of proteins and nucleic acids. *Science*, 352:239–242, 2016. (Cited on pages 7, 70, and 107.)
- [157] K. Neupane, A. P. Manuel, and M. T. Woodside. Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nature Physics*, 2016. ISSN 1745-2473. (Cited on pages 7 and 107.)
- [158] K. Neupane, F. Wang, and M. T. Woodside. Direct measurement of sequence-dependent transition path times and conformational diffusion in DNA duplex formation. *Proceedings of the National Academy of Sciences of the United States of America*, 114:1329–1334, 2017. (Cited on pages 187 and 204.)
- [159] R. Nevo, V. Brumfeld, R. Kapon, P. Hinterdorfer, and Z. Reich. Direct measurement of protein energy landscape roughness. *EMBO Reports*, 6:482–486, 2005. (Cited on pages 46, 71, and 200.)

- [160] A. W. Nguyen and P. S. Daugherty. Evolutionary optimization of fluorescent proteins for intracellular FRET. *Nature Biotechnology*, 23:355–360, 2005. (Cited on page 216.)
- [161] L. Oberbarnscheidt, R. Janissen, and F. Oesterhelt. Direct and model free calculation of force-dependent dissociation rates from force spectroscopic data. *Biophysical Journal*, 97:L19–L21, 2009. (Cited on pages 26, 37, and 38.)
- [162] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997. (Cited on page 5.)
- [163] D. E. Otzen and M. Oliveberg. Conformational plasticity in folding of the split beta-alpha-beta protein S6: evidence for burst-phase disruption of the native state. *Journal of Molecular Biology*, 317:613–627, 2002. (Cited on page 131.)
- [164] C. Pabo. Molecular technology. Designing proteins and peptides. *Nature*, 301:200, 1983. (Cited on page 8.)
- [165] C. N. Pace and J. M. Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical journal*, 75:422–427, 1998. (Cited on pages 166 and 213.)
- [166] B. Pelz. *Enzyme mechanics studied by single molecule force spectroscopy*. PhD thesis, Technische Universität München, 2014. (Cited on pages 17, 32, 137, and 199.)
- [167] B. Pelz, G. Žoldák, F. Zeller, M. Zacharias, and M. Rief. Subnanometre enzyme mechanics probed by single-molecule force spectroscopy. *Nature Communications*, 7:10848—1–9, 2016. (Cited on page 28.)
- [168] E. J. G. Peterman, F. Gittes, and C. F. Schmidt. Laser-induced heating in optical traps. *Biophysical Journal*, 84:1308–1316, 2003. (Cited on pages 17 and 19.)
- [169] C. A. Pierse and O. K. Dudko. Kinetics and energetics of biomolecular folding and binding. *Biophysical Journal*, 105:L19–L22, 2013. (Cited on page 47.)
- [170] C. A. Pierse and O. K. Dudko. Distinguishing signatures of multipathway conformational transitions. *Physical Review Letters*, 118:088101—1–5, 2017. (Cited on pages 147, 150, and 201.)
- [171] I. Popa, R. Berkovich, J. Alegre-Cebollada, C. L. Badilla, J. A. Rivas-Pardo, Y. Taniguchi, M. Kawakami, and J. M. Fernandez. Nanomechanics of HaloTag tethers. *Journal of the American Chemical Society*, 135:12762–12771, 2013. (Cited on page 24.)

- [172] L. R. Pratt. A statistical method for identifying transition states in high dimensional problems. *The Journal of Chemical Physics*, 85:5045, 1986. (Cited on page 52.)
- [173] T. Proft. Sortase-mediated protein ligation: an emerging biotechnology tool for protein modification and immobilisation. *Biotechnology Letters*, 32:1–10, 2010. (Cited on page 24.)
- [174] S. B. Prusiner. Prions. *Proceedings of the National Academy of Sciences of the United States of America*, 95:13363–13383, 1998. (Cited on page 3.)
- [175] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989. (Cited on pages 27 and 38.)
- [176] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3:4–16, 1986. (Cited on page 38.)
- [177] B. Ramm, J. Stigler, M. Hinczewski, D. Thirumalai, H. Herrmann, G. Woehlke, and M. Rief. Sequence-resolved free energy profiles of stress-bearing vimentin intermediate filaments. *Proceedings of the National Academy of Sciences of the United States of America*, 111:11359–11364, 2014. (Cited on pages 50, 51, 66, 78, 86, and 183.)
- [178] S. T. Rao and M. G. Rossmann. Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, 76:241–256, 1973. (Cited on page 9.)
- [179] Y. M. Rhee and V. S. Pande. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *Journal of Physical Chemistry B*, 109:6780–6786, 2005. (Cited on page 54.)
- [180] S. Rice, A. W. Lin, D. Safer, C. L. Hart, N. Naber, B. O. Carragher, S. M. Cain, E. Pechatnikova, E. M. Wilson-Kubalek, M. Whittaker, E. Pate, R. Cooke, E. W. Taylor, R. A. Milligan, and R. D. Vale. A structural change in the kinesin motor protein that drives motility. *Nature*, 402:778–784, 1999. (Cited on pages 218 and 222.)
- [181] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, 276:1109–1112, 1997. (Cited on pages 3, 15, 115, and 134.)
- [182] M. Rief, J. Fernandez, and H. Gaub. Elastically coupled two-level systems as a model for biopolymer extensibility. *Physical Review Letters*, 81:4764–4767, 1998. (Cited on page 35.)
- [183] M. Rief, H. Clausen-Schaumann, and H. E. Gaub. Sequence-dependent mechanics of single DNA molecules. *Nature Structural Biology*, 6:346–349, 1999. (Cited on page 244.)

- [184] M. Rief, R. S. Rock, A. D. Mehta, M. S. Mooseker, R. E. Cheney, and J. A. Spudich. Myosin-V stepping kinetics: a molecular model for processivity. *Proceedings of the National Academy of Sciences of the United States of America*, 97:9482–9486, 2000. (Cited on page 76.)
- [185] P. Ringer, A. Weißl, A.-L. Cost, A. Freikamp, B. Sabass, A. Mehlich, M. Tramier, M. Rief, and C. Grashoff. Multiplexing molecular tension sensors reveals piconewton force gradient across talin-1. *Nature Methods*, (accepted), 2017. (Cited on page 141.)
- [186] R. S. Rock, M. Rief, A. D. Mehta, and J. a. Spudich. In vitro assays of processive myosin motors. *Methods*, 22:373–381, 2000. (Cited on page 244.)
- [187] L. Rognoni. *Molecular Mechanism of the Mechanosensor Filamin*. PhD thesis, Technische Universität München, 2014. (Cited on page 135.)
- [188] L. Rognoni, J. Stigler, B. Pelz, J. Ylänne, and M. Rief. Dynamic force sensing of filamin revealed in single-molecule experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 109:19679–19684, 2012. (Cited on pages 21, 28, 135, 212, and 220.)
- [189] L. Rognoni, T. Möst, G. Žoldák, and M. Rief. Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin. *Proceedings of the National Academy of Sciences of the United States of America*, 111:5568–5573, 2014. (Cited on pages 5, 37, 145, 206, and 213.)
- [190] M. G. Rossmann, D. Moras, and K. W. Olsen. Chemical and biological evolution of nucleotide-binding protein. *Nature*, 250:194–199, 1974. (Cited on page 9.)
- [191] I. E. Sánchez and T. Kiefhaber. Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *Journal of Molecular Biology*, 325:367–376, 2003. (Cited on pages 79 and 131.)
- [192] I. E. Sánchez and T. Kiefhaber. Hammond behavior versus ground state effects in protein folding: evidence for narrow free energy barriers and residual structure in unfolded states. *Journal of Molecular Biology*, 327:867–884, 2003. (Cited on page 134.)
- [193] M. Schlierf. *Einzelmolekülmechanik der Faltung und Entfaltung von Proteinen*. PhD thesis, Technische Universität München, 2008. (Cited on pages 31 and 34.)
- [194] M. Schlierf and M. Rief. Single-molecule unfolding force distributions reveal a funnel-shaped energy landscape. *Biophysical Journal*, 90:L33–L35, feb 2006. (Cited on pages 26 and 134.)
- [195] M. Schlierf, H. Li, and J. M. Fernandez. The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. *Proceedings*

- of the National Academy of Sciences of the United States of America, 101: 7299–7304, 2004. (Cited on pages 21, 134, 212, and 219.)
- [196] M. Schlierf, F. Berkemeier, and M. Rief. Direct observation of active protein folding using lock-in force spectroscopy. *Biophysical Journal*, 93: 3989–3998, 2007. (Cited on pages 26, 28, 42, 47, and 134.)
- [197] M. Schlierf, Z. T. Yew, M. Rief, and E. Paci. Complex unfolding kinetics of single-domain proteins in the presence of force. *Biophysical Journal*, 99:1620–1627, 2010. (Cited on pages 5, 26, 81, 131, and 134.)
- [198] S. Schmid, M. Götz, and T. Hugel. Single-molecule analysis beyond dwell times: demonstration and assessment in and out of equilibrium. *Biophysical Journal*, 111:1375–1384, 2016. (Cited on page 3.)
- [199] M. Schneider, X. Fu, and A. E. Keating. X-ray vs. NMR structures as templates for computational protein design. *Proteins: Structure, Function and Bioinformatics*, 77:97–110, 2009. (Cited on page 199.)
- [200] B. Schuler and H. Hofmann. Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Current Opinion in Structural Biology*, 23:36–47, 2013. (Cited on page 3.)
- [201] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419:743–748, 2002. (Cited on pages 41 and 46.)
- [202] I. Schwaiger, M. Schleicher, A. A. Noegel, and M. Rief. The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiments. *EMBO reports*, 6:46–51, 2005. (Cited on page 206.)
- [203] K. A. Scott, S. Batey, K. A. Hooton, and J. Clarke. The folding of spectrin domains I: wild-type domains have the same stability but very different kinetic properties. *Journal of Molecular Biology*, 344:195–205, 2004. (Cited on page 144.)
- [204] G. Senavirathne, J. Liu, M. A. Lopez, J. Hanne, J. Martin-Lopez, J.-B. Lee, K. E. Yoder, and R. Fishel. Widespread nuclease contamination in commonly used oxygen-scavenging systems. *Nature Methods*, 12:901–902, 2015. (Cited on page 19.)
- [205] R. F. Service. Problem Solved\* (\*sort of). *Science*, 321:784–786, 2008. (Cited on pages 3 and 8.)
- [206] D. E. Shafer, J. K. Inman, and A. Lees. Reaction of Tris(2-carboxyethyl)phosphine (TCEP) with maleimide and  $\alpha$ -Haloacyl groups: anomalous elution of TCEP by gel filtration. *Analytical Biochemistry*, 282:161–164, 2000. (Cited on page 23.)

- [207] E. M. Sletten and C. R. Bertozzi. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angewandte Chemie International Edition*, 48:6974–6998, 2009. (Cited on page 24.)
- [208] B. L. Smith, T. E. Schaffer, M. Viani, J. B. Thompson, N. A. Frederick, J. Kindt, A. Belcher, G. D. Stucky, D. E. Morse, and P. K. Hansma. Molecular mechanistic origin of the toughness of natural adhesives, fibres and composites. *Nature*, 399:761–763, 1999. (Cited on page 115.)
- [209] A. Solanki, K. Neupane, and M. T. Woodside. Single-molecule force spectroscopy of rapidly fluctuating, marginally stable structures in the intrinsically disordered protein  $\alpha$ -Synuclein. *Physical Review Letters*, 112:158103—1–6, 2014. (Cited on pages 36 and 200.)
- [210] Y.-H. Song and E. Mandelkow. Paracrystalline structure of the stalk domain of the microtubule motor protein kinesin. *Journal of Structural Biology*, 112:93–102, 1994. (Cited on page 218.)
- [211] R. Stevens, L. Stevens, and N. C. Price. The stabilities of various thiol compounds used in protein purifications. *Biochemical Education*, 11:70–70, 1983. (Cited on page 22.)
- [212] J. Stigler. *Complex ligand-dependent folding of single proteins observed with optical tweezers*. PhD thesis, Technische Universität München, 2012. (Cited on pages 31, 38, 65, 122, 135, 168, 181, and 242.)
- [213] J. Stigler and M. Rief. Calcium-dependent folding of single calmodulin molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 109:17814–17819, 2012. (Cited on pages 3, 28, and 135.)
- [214] J. Stigler and M. Rief. Hidden Markov analysis of trajectories in single-molecule experiments and the effects of missed events. *ChemPhysChem*, 13:1079–1086, 2012. (Cited on pages 38, 39, 40, and 181.)
- [215] J. Stigler and M. Rief. Ligand-induced changes of the apparent transition-state position in mechanical protein unfolding. *Biophysical Journal*, 109:365–372, 2015. (Cited on page 28.)
- [216] J. Stigler, F. Ziegler, A. Gieseke, J. C. M. Gebhardt, and M. Rief. The complex folding network of single calmodulin molecules. *Science*, 334:512–516, 2011. (Cited on pages 5, 19, 21, 27, 38, 40, 51, 201, 212, and 220.)
- [217] M. F. Stock, J. Guerrero, B. Cobb, C. T. Eggers, T. G. Huang, X. Li, and D. D. Hackney. Formation of the compact conformer of kinesin requires a COOH-terminal heavy chain domain and inhibits microtubule-stimulated ATPase activity. *Journal of Biological Chemistry*, 274:14617–14623, 1999. (Cited on page 218.)
- [218] T. R. Strick, J. F. Allemand, D. Bensimon, A. Bensimon, and V. Croquette. Elasticity of a single supercoiled DNA molecule.pdf. *Science*, 271:1835–1837, 1996. (Cited on page 15.)

- [219] M. Sugita and T. Kikuchi. Analyses of the folding properties of ferredoxin-like fold proteins by means of a coarse-grained Gō model: relationship between the free energy profiles and folding cores. *Proteins: Structure, Function and Bioinformatics*, 82:954–965, 2014. (Cited on pages 131 and 132.)
- [220] J.-Y. Sun, X. Zhao, W. R. K. Illeperuma, O. Chaudhuri, K. H. Oh, D. J. Mooney, J. J. Vlassak, and Z. Suo. Highly stretchable and tough hydrogels. *Nature*, 489:133–136, 2012. (Cited on page 115.)
- [221] T. Suren. *Kraftspektroskopie an Spektrindomänen*. Diploma thesis, Technische Universität München, 2012. (Cited on page 215.)
- [222] K. Svoboda and S. M. Block. Biological Applications of Optical Forces. *Annual Review of Biophysics and Biomolecular Structure*, 23:247–285, 1994. (Cited on page 16.)
- [223] K. Svoboda, C. F. Schmidt, B. J. Schnapp, and S. M. Block. Direct observation of kinesin stepping by optical trapping interferometry. *Nature*, 365:721–727, 1993. (Cited on page 15.)
- [224] I. Tinoco and C. Bustamante. The effect of force on thermodynamics and kinetics of single molecule reactions. *Biophysical Chemistry*, 101-102: 513–533, 2002. (Cited on page 31.)
- [225] S. F. Tolić-Nørrelykke, E. Schäffer, J. Howard, F. S. Pavone, F. Jülicher, and H. Flyvbjerg. Calibration of optical tweezers with positional detection in the back focal plane. *Review of Scientific Instruments*, 77:103101—1–11, 2006. (Cited on pages 16, 17, and 242.)
- [226] Z. Tshiprut, J. Klafter, and M. Urbakh. Single-molecule pulling experiments: when the stiffness of the pulling device matters. *Biophysical Journal*, 95:L42–L44, 2008. (Cited on page 49.)
- [227] M. S. Turner. Origin of the universe. *Scientific American*, 301:36–43, 2009. (Cited on page 4.)
- [228] J. H. van't Hoff. *Etudes de dynamique chimique*. Frederik Muller & Co., Amsterdam, 1884. (Cited on page 41.)
- [229] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967. (Cited on page 39.)
- [230] Y. von Hansen, A. Mehlich, B. Pelz, M. Rief, and R. R. Netz. Auto- and cross-power spectral analysis of dual trap optical tweezer experiments using Bayesian inference. *Review of Scientific Instruments*, 83:095116—1–19, Sept. 2012. (Cited on pages 16 and 17.)
- [231] S. A. Waldauer, O. Bakajin, and L. J. Lapidus. Extremely slow intramolecular diffusion in unfolded protein L. *Proceedings of the National*

*Academy of Sciences of the United States of America*, 107:13713–13717, 2010. (Cited on page 66.)

- [232] J. E. Walker, M. Saraste, M. Runswick, and N. J. Gay. Distantly related sequences in the  $\alpha$ - and  $\beta$ -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *The EMBO Journal*, 1:945–951, 1982. (Cited on page 144.)
- [233] S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado. Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proceedings of the National Academy of Sciences of the United States of America*, 96:5486–5491, 1999. (Cited on page 201.)
- [234] L. Wang, A. Brock, B. Herberich, and P. G. Schultz. Expanding the genetic code of *Escherichia coli*. *Science*, 292:498–500, 2001. (Cited on page 24.)
- [235] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. Stretching DNA with Optical Tweezers. *Biophysical Journal*, 72:1335–1346, 1997. (Cited on pages 15 and 31.)
- [236] E. Weinan and E. Vanden-Eijnden. Towards a theory of transition paths. *Journal of Statistical Physics*, 123:503–523, 2006. (Cited on pages 7 and 52.)
- [237] E. Weinan and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annual Review of Physical Chemistry*, 61:391–420, 2010. (Cited on pages 7, 52, and 54.)
- [238] B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*, 463:685–688, 2010. (Cited on pages 6, 71, and 200.)
- [239] G. Woehlke and M. Schliwa. Walking on two heads: the many talents of kinesin. *Nature Reviews Molecular Cell Biology*, 1:50–58, 2000. (Cited on page 218.)
- [240] P. G. Wolynes. Energy landscapes and solved protein-folding problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363:453–467, 2005. (Cited on page 5.)
- [241] P. G. Wolynes, W. a. Eaton, and A. R. Fersht. Chemical physics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 109:17770–17771, 2012. (Cited on page 5.)
- [242] M. T. Woodside and S. M. Block. Reconstructing folding energy landscapes by single-molecule force spectroscopy. *Annual Review of Biophysics*, 43:19–39, 2014. (Cited on page 3.)
- [243] M. T. Woodside, P. C. Anthony, W. M. Behnke-Parks, K. Larizadeh, D. Herschlag, and S. M. Block. Direct measurement of the full,

- sequence-dependent folding landscape of a nucleic acid. *Science*, 314: 1001–1005, 2006. (Cited on pages 28 and 49.)
- [244] D. N. Woolfson, G. J. Bartlett, A. J. Burton, J. W. Heal, A. Niitsu, A. R. Thomson, and C. W. Wood. De novo protein design: how do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology*, 33:16–26, 2015. (Cited on page 8.)
- [245] S. Yadahalli and S. Gosavi. Designing cooperativity into the designed protein Top 7. *Proteins: Structure, Function and Bioinformatics*, 82:364–374, 2014. (Cited on page 197.)
- [246] Z. T. Yew, M. Schlierf, M. Rief, and E. Paci. Direct evidence of the multidimensionality of the free-energy landscapes of proteins revealed by mechanical probes. *Physical Review E*, 81:1–4, 2010. (Cited on page 134.)
- [247] H. Yu, A. N. Gupta, X. Liu, K. Neupane, A. M. Brigley, I. Sosova, and M. T. Woodside. Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time, and rates. *Proceedings of the National Academy of Sciences of the United States of America*, 109:14452–14457, 2012. (Cited on pages 28, 46, 49, 78, 86, 100, and 183.)
- [248] H. Yu, X. Liu, K. Neupane, A. N. Gupta, A. M. Brigley, A. Solanki, I. Sosova, and M. T. Woodside. Direct observation of multiple misfolding pathways in a single prion protein molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 109:5283–5288, 2012. (Cited on pages 93, 97, and 191.)
- [249] H. Yu, D. R. Dee, X. Liu, A. M. Brigley, I. Sosova, and M. T. Woodside. Protein misfolding occurs by slow diffusion across multiple barriers in a rough energy landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 112:8308–8313, 2015. (Cited on pages 29, 36, 46, 66, 71, 131, and 183.)
- [250] H. Yu, M. G. W. Siewny, D. T. Edwards, A. W. Sanders, and T. T. Perkins. Hidden dynamics in the unfolding of individual bacteriorhodopsin proteins. *Science*, 355:945–950, 2017. (Cited on pages 3 and 181.)
- [251] B. W. Zhang, D. Jasnow, and D. M. Zuckerman. Transition-event durations in one-dimensional activated processes. *The Journal of Chemical Physics*, 126:074507—1–18, 2007. (Cited on page 59.)
- [252] W. Zheng, D. De Sancho, T. Hoppe, and R. B. Best. Dependence of internal friction on folding mechanism. *Journal of the American Chemical Society*, 137:3283–3290, 2015. (Cited on page 191.)
- [253] W. Zheng, D. De Sancho, and R. B. Best. Modulation of folding internal friction by local and global barrier heights. *Journal of Physical Chemistry Letters*, 7:1028–1034, 2016. (Cited on page 200.)

- [254] P. I. Zhuravlev, M. Hinczewski, S. Chakrabarti, S. Marqusee, and D. Thirumalai. Force-dependent switch in protein unfolding pathways and transition state movements. *Proceedings of the National Academy of Sciences of the United States of America*, 113:E715–E724, 2016. (Cited on page 81.)
- [255] S. N. Zhurkov. Kinetic concept of the strength of solids. *International Journal of Fracture Mechanics*, 1:311–323, 1965. (Cited on page 41.)
- [256] S. N. Zhurkov and V. E. Korsukov. Atomic mechanism of fracture of solid polymers. *Journal of Polymer Science: Polymer Physics Edition*, 12:385–398, 1974. (Cited on page 41.)
- [257] F. B. Ziegler. *Einzelmolekülexperimente zum Einfluss von Knoten und makromolekularem Crowding auf die Faltung von Proteinen*. PhD thesis, Technische Universität München, 2015. (Cited on pages 17, 24, 31, 32, 33, and 123.)
- [258] D. Zimmermann. *Involvement of myosin V in organelle transport and its unconventional interaction with microtubules*. PhD thesis, Ludwig-Maximilians-Universität München, July 2012. (Cited on page 241.)
- [259] D. Zimmermann, B. Abdel Motaal, L. Voith von Voithenberg, M. Schliwa, and Z. Ökten. Diffusion of myosin V on microtubules: a fine-tuned interaction for which e-hooks are dispensable. *PLoS ONE*, 6:e25473—1–10, 2011. (Cited on page 241.)
- [260] G. Žoldák and M. Rief. Force as a single molecule probe of multidimensional protein energy landscapes. *Current Opinion in Structural Biology*, 23:48–57, 2013. (Cited on page 3.)
- [261] G. Žoldák, J. Stigler, B. Pelz, H. Li, and M. Rief. Ultrafast folding kinetics and cooperativity of villin headpiece in single-molecule force spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 110:18156–18161, 2013. (Cited on pages 7, 15, 21, 29, 35, 36, 40, 51, 68, 117, 135, 140, 212, and 220.)
- [262] R. Zwanzig. Diffusion in a rough potential. *Proceedings of the National Academy of Sciences of the United States of America*, 85:2029–2030, 1988. (Cited on pages 6 and 71.)
- [263] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal’s paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89:20–22, 1992. (Cited on page 4.)



## CURRICULUM VITAE

---

Alexander Mehlich

22 February 1984      Born in Berlin

Nationality            German

1990 - 1993            Enrolment at 1<sup>st</sup> Primary School Berlin-Mitte

1993 - 2002            Enrolment at Nicolaus-Cusanus-Gymnasium, Bonn

Jan. - May 2000        Enrolment at Secondary School Nr. 1527, Moscow;  
guest student at the Faculty of Mechanics  
and Mathematics, Moscow State University

2000 - 2001            Guest math student at the University of Cologne

July 2002              Abitur (general qualification for university entrance)

2002 - 2003            Civil service at the Goethe-Institut Inter Nationes, Bonn

2003 - 2006            Basic studies in physics at the TU Munich

2006 - 2007            Studying abroad in the Master's course

"Physique et Applications" at the  
Université Pierre et Marie Curie in Paris;  
scientific internship at Saint-Gobain Recherche, Paris

2007 - 2009            Main studies in physics at the TU Munich

November 2009        Final degree as Diplom-Physiker (univ.)

since January 2010    Postgraduate studies guided by Prof. Dr. Matthias Rief,  
Chair of Biophysics E22, TU Munich

2010 - 2011            Speaker of the NIM student board,  
cluster of excellence Nanosystems Initiative Munich

July - Sept. 2014      Research stay at the University of British Columbia,  
Vancouver, Canada, incl. a visit to the  
University of Washington, Seattle, United States