Calorimetric Determination of Dystrophin ABD1 Unfolding Energetics

A THESIS SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE BY

Christian Scott Coffman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN CHEMISTRY

Dr. Anne Hinderliter

December, 2018

Acknowledgements

I want to say thanks to all of those around me who supported the laborious purification process of Dystrophin ABD1. Elliot for helping kick off the project with designing the DNA constructs, Robby and Tori for the partnership all throughout experimentation including all derivations as well as for not throwing up on me while appreciating the shear beauty that is physical chemistry, Ellen for putting up with all of the late nights/ nights I wouldn't return home for the project, and the rest of the undergraduates in the Hinderliter lab. I would also like to acknowledge Lisa Ito and Kassidy Rodrigues for their support. A special thanks to Dr. Hinderliter for her unwavering support no matter how painful it was at sometimes it made me a better researcher and person. And a thanks to Dr. Cembran and the Dr. Dave Thomas group for their work in this collaboration. A special thanks to my committee for the helpful advice in finishing this thesis.

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Abstract

Muscular Dystrophy (MD) is a disease that effects the structural integrity of muscle cells. Studies have linked the Dystrophin protein to MD as the most commonly altered gene in patients with MD. Using empirical models established to predict the change in heat capacity associated with protein unfolding, we we correlated the likelihood of a mutation as being disease causing with an associated change in the heat capacity at that amino acid position. These studies focused on the first Actin Binding Domain of Dystrophin (ABD1, 27kDa) using Differential Scanning Calorimetry (DSC), as it is the region with the highest density of disease causing mutations. ABD1 is comprised of two Calmodulin Homology domains (CH1 and CH2) connected by a short linker region and is predicted to be slightly disordered. [1] Analyzing the data acquired from DSC proved to be rather difficult as it was highly dependent on the baseline definition, which can be rather noisy. This thesis describes the evolution of our DSC analysis starting with an analysis published in the Biophysical Journal. This method suggested the a change in heat capacity (ΔC_p of 5 $\pm$ 5 kcal/mol). However, the model showed some systematic deviation from the experimental data so the data was fit to a Gaussian and Hubbert distribution. Then a deconvolution approach revealed the presence of an appreciable occupancy (approximately 50%) of intermediate states that helps account for the deviations from a two-state model. Deconvoluting the transitions revealed at least one intermediate transition with $\Delta G(37°C)$ of 2 ± 1kcal/mol and an unfolding free energy of $2.2 \pm 0.6 kcal/mol$ and a change in heat capacity that is smaller than predicted. This free energy is comparable to that which has been determined for actin binding thus implicating unfolding of ABD1 upon binding actin, possibly through separating the CH domains or some other mechanism.

Contents

List of Tables iv

List of Figures viii

List of Abbreviations xvi

1 Introduction 1

1.1 Muscular Dystrophy 2

1.2 The Dystrophin protein 4

1.3 Materials 11

1.3.1 Reagents 11

1.3.2 Glassware 13

1.3.3 Instrumentation 13

1.4 Methods 14

1.4.1 Plasmid construct design 14

1.4.2 Cell growth and protein purification 15

1.4.3 Lysis and protein purification 16

1.4.4 Differential scanning calorimetry (DSC) 18

1.4.5 Software 19

2 Theory and comparison between methods of analysis 21

2.1 Construct results 22

2.2 Theory 25
2.2.1 Two-state equilibrium derivation 25
2.2.2 Three-state equilibrium derivation 29
2.2.3 Two independent transitions 31
2.3 Current methodology 32
2.4 Rationale for new methodology 36
2.5 Efficiency 36
2.6 Other benefits of automated work flow 39
2.6.1 Change in heat capacity estimate 40
2.6.2 Creating a baseline 40
2.6.3 Excess enthalpy and transition temperature estimates 44
2.7 The search for a global minimum 44
2.7.1 Assessing the minimum 48
2.8 Regression of a boltzmann distribution to ABD1 data 50
2.9 Gaussian distributions 53
2.9.1 Fitting two peaks 53
2.9.2 Fitting baseline subtracted data 55
2.10 Gaussian modeling of ABD1 data 55
2.10.1 The Hubbert function 57
2.11 Model free approach to fitting 64

3 Conclusion **75**
3.0.1 Future directions 76

Bibliography **79**

List of Tables

1.1 A comparison of the variances and means between ? samples and other samples to determine if the ? samples are from the same population as the + or - populations. The first value in each cell represents the ANOVA p-value The second value in each cell (in parentheses) represents the p-value from a t-test of the appropriate nature depending on the output from the ANOVA test. Significant values are underlined. When variances were equivalent we used a two sample t-test, when they were not equal we used Welch's t-test. 8

1.2 p-values from ANOVA (first value) and t-test (value in parentheses) between buried/exposed and disease causing(+)/non-disease causing mutations(-). p-values adjusted by Tukey's honest significant differences method and the Benjamini/Hochberg method.[27] Statistically significant values are underlined. The second value in each cell (in parentheses) represents the p-value from a t-test using Tukey's Honest Significant differences method. Significant values are underlined. . . . 9

2.1 Results from regressions with different seeding values on Synaptotagmin C2B data originally reported by Fealey et al.[40] The data was interpretted from a graph in the original publication.[40] The fit values as well as reported modeling errors from regression are reported for each parameter. The last column reports the mean squared residual from the regression output calculated by dividing the RSS by the number of data points in the scan. 34

2.2 Evaluations of different sets of data encompassing both experimental and theoretical data. One scan was analyzed for each set of reported values above to remain consistent across all rows as some of the experiments only have one set of data. The two-state model was generated with the following parameters ΔH_{T_m}=140 kcal/mol, ΔC_p = 2 kcal/mol, T_m 336. The three- state model was generated with parameters ΔH_{T_m} = 100, 120 (kcal/mol); ΔC_p = 1, 1.5; T_m = 325, 327. Only one set of scans was obtained for MBP and MBP+ABD1 experiments where MBP+ABD1 represents the uncleaved protein construct containing both MBP and ABD1. Models that did not converge due to singularities are represented as DNC. Models utilizing more than one transition list the first set of values as the top value in each cell, and the second transition as the lower set of values in each cell. 51

2.3 Results from baseline subtracted experimental ABD1 data fit with a single Hubbert function. The fit parameters define the following features of the curve: k defines the height, SD the width, and T_m the center. The average is reported along with the 95% confidence interval. 64

2.4 Results from baseline subtracted experimental ABD1 data fit with two Hubbert functions. The fit parameters define the following features of the curve: k defines the height, SD the width, and T_m the center. The number in the first column directly following the parameter name describes which transition the fit parameter corresponds to. Averages are reported along with the 95% confidence interval. 66

2.5 Thermodynamic values reported from the model free approach to analysis. The 95% confidence intervals were generated from triplicate scans of ABD1. 72

2.6 Thermodynamic values documented from triplicate analysis of ABD1 data. All values are reported with 95% confidence intervals. Each column represents a different method of analysis. Values which are not calculated by a given method of analysis are marked by a minus sign (-). Gaussian fitting methods were not included due to the noted deviation it showed from theoretical data in Fig. 2.13. 74

List of Figures

1.1 A representation of a mutated Dystrophin gene (red) which is located on the 'X' chromosome. Left: The image represents a female's sex chromosomes. Right: An image representing a male's sex chromosomes. 3

1.2 A cartoon representation of Dystrophin. The yellow regions denote actin binding regions, the red vertical lines represent the f-actin cytoskeleton matrix, the curved dashed line represents the sarcolemma. Each circle along Dystrophin represents a folded domain with the left two overlapping circles representing CH1 and CH2 of ABD1, all subsequent circles representing the large portion of SPR's, and the rectangular region representing the WW domain, which is rich in cysteine residues and couples Dystrophin to the sarcolemma (dashed, curved line). The second actin binding domain (ABD2) can be seen as the second string of yellow circles in the central region of Dystrophin. All reported domains are based off of speculative interaction predictions from homology mapping and are provided by Uniprot.[26] 5

1.3 Left: The total number of mutations at each amino acid position. Protein regions are designated by color. Right: The mutation density of each protein region. The red circle represents ABD1 and the blue square represents regions that are undefined as per Uniprot. 7

2.1 Sequence of the Dystrophin ABD1 amino acid sequence cloned for purification and thermodynamic studies. Amino acids are colored by their identities. 22

2.2 A representation of the pet28-MBP-TEV plasmid design. All of the features are noted included opening reading frames where the ABD1 gene was inserted right after the TEV site and the MBP gene denoted in light maroon at the BamHI restriction site. 23

2.3 SDS-PAGE gels displaying the purity test for Dys ABD1. The top represents a purification for the first two sets of scans. The contents of each lane are as follows: Unstained Protein Ladder, Broad Range (10-250 kDa) from NEB with molecular weight bands representing 10kDa, 15kDa, 20kDa, 25kDa, 30kDa, 40kDa, 60kDa, 80kDa, 100kDa, 150kDa, and 250kDa; the second column elution with bands at 44kDa and two smaller bands around 10kDa; empty lane; flow through fractions from column two with molecular weights 27kDa. Bottom: Purification for the last scan, with the ladder in the left most lane, an empty lane, then two lanes with the flow through fraction. The molecular weights were difficult to interpret for the second gel but the molecular weight ladder was the same. 24

2.4 Example fits from the reported regression solutions reported in Tab. 2.1. The raw data is represented as discrete blue squares and the model is represented by the continuous orange line. Starting with the top left, the outputs from Tab. 2.1 are graphed from left to right in order. . . . 35

2.5 An image taken from the interactive user interface designed in R to aid in DSC analysis. Pictured here is the baseline assist where the user drags the definition of their baseline which updates the graph in live time with a projection of the baseline as well as the slope. 38

2.6 Top-left: The integrated heat capacity throughout the entirety of the transition. Bottom-left: The integrated function, normalized, before readjusting the max to be equivalent to the estimated change in heat capacity. Top-right: Experimental data with the estimated baseline graphed together. Bottom-right: The normalized integral of the experimental data after subtraction of the sigmoidal baseline. 41

2.7 Raw data of the three ABD1 scans represented after an initial baseline estimate. The final baseline estimate is projected by the red dotted lines with the slopes printed above the given lines. The blue vertical lines display the innermost bounds used in the definition of the baselines. . 43

2.8 Geometric representation of the change in enthalpy (maroon) and the calculated sigmoidal baseline (lower bound) for all three scans. The computed baseline can also be visualized as the lower bound for the maroon region. 45

2.9 An example of experimental ABD1 data (black) fit with a two independent transitions model (red). Note the qualitatively good fit of the raw data, but in evaluating the confidence regions of the fit, it is clear that one transition has both a negative enthalpy and change in heat capacity indicative of overfitting the data. 46

2.10 Experimental data representing a melt of an MBP-ABD1 construct (black) fit with both a sequential model and an independent transitions model. Both have reported confidence regions as well. The independent transitions model appears to fit better. 47

2.11 Left: Pairwise determination of the log error of the model fit as a heat map with the 95% confidence denoted with a red circle. Right: Confidence estimate using a random sampling of the nearby parameter space, where only results satisfying the 95% confidence criteria are printed on the graph as an 'x'. 49

2.12 Nonlinear regression fits of a two-state Boltzmann distribution regression model (yellow) on each individual scan (black) with the reported fit parameters in the table below. The average and 95% confidence interval of the fit parameters are reported in the table along with the Van't Hoff enthalpy (ΔH_{VH}) and free energy at physiological temperatures ($\Delta G(37°C)$). 52

2.13 The first three panels represent DSC data that were simulated using two transitions of 100 and 120 kcal/mol, respectively, with varying differences in transition temperatures and no baseline change in heat capacity. The simulated DSC data were then fit to normal distributions through regressing a Gaussian distribution to the data. Red and blue shading represent the separate Gaussian distributions and purple shading represents the sum of the two. Top-left: Transition temperatures that were sufficiently separate resulted in the proper fitting of two distinct transitions with the fit enthalpies being 96% accurate. Top-right: Two overlapping transitions with the fit enthalpy 96% accurate. Bottom: Two transitions of identical transition temperature, where the fit enthalpy was reportedly equal leading to 80% accuracy. 54

2.14 Top-left: Simulated DSC data with from the following parameters for a single transition ΔH = 140 kcal/mol, ΔC_p = 5 kcal/mol K, T_m =336 K. Top-right: Data from the graph to the left treated with the baseline subtraction procedure. The enthalpy was measured to be 98% accurate and also unaffected by the presence of noise. Bottom-left: Simulated DSC data displaying overlapping transitions and a nonzero change in heat capacity. Bottom-right: Gaussian distributions fit to the baseline treated data the enthalpies were within 5% agreement of the values with which they were generated. The y-axis for both figures are the same and represent the same units (kcal/ mol K). 56

2.15 Fitting DSC data after baseline subtraction to Gaussian distributions. Each set of experimental data is represented by a separate column. Row 1: represents a one Gaussian fit. Reported ΔH's: 121, 137, and 104 kcal/mol, respectively Row 2: Fitting ABD1 data to two Gaussian distributions. Red populations are centered on 329K and blue are centered on 327K. Reported ΔH's red = 52, 82, and 36 kcal/mol, respectively; blue = 80, 74, and 75 kcal/mol respectively. 58

2.16 Theoretically generated two-state transitions with enthalpies ranging from 25kcal/mol to 250kcal/mol. The experiments are colored by the generated enthalpy values. Model fits of the curves are represented as dashed lines. 59

2.17 Experimental data on ABD1 were fit to single and double Hubbert distributions where the black line represents experimental data after baseline subtraction and the shaded red region represents the total modeled enthalpy. Top: Baseline subtracted experimental ABD1 data fit with a single Hubbert distribution. Bottom:Baseline subtracted experimental ABD1 data fit with two Hubbert distributions. 60

2.18 Compiled height and width values determined from fitting theoretically generated two-state transitions of enthalpies varying from 25kcal/mol to 250kcal/mol to a Hubbert distribution. The top two graphs display the relationship of the height (top-left) and width (top-right) fit values to the enthalpy value with which the data was generated. The points are colored by the enthalpy values with which they were generated. The second two graphs represent the linearization and fit (bottom-left) of the height and width fit values where as the bottom-right graph represents the relationship between the fit values on linear axes. The coloring represents sectioning of the space into values that would represent transitions that are too broad (red) or too narrow (blue) to be represented as a two-state transition. The power model $H = a * W^b$ returned fit values of $a = 55.394$ and $b = -1.705$. 62

2.19 Suggested system for classification of thermodynamic transitions. The blue region represents transitions of disproportionately sharp transitions indicative of cooperative effects, the red region represents disproportionately wide transitions (likely multiple transitions), and a purple region where an acceptable definition of a two-state transition could be adopted. The two-state region (purple) was arbitrarily defined by introducing noise into the fit parameters of the Hubbert function. . . . 63

2.20 Theoretical data generated by Zhou et al. with the sigmoidal baseline subtracted (black points) fit with a Hubbert function (red line). The resulting height and width of the Hubbert fit parameters are 43.0 kcal/mol and 1.1 K, respectively. Rights provided to this Protein Science article provided by the Copyright Clearance Center's RightsLink. 65

2.21 Three graphs representing the the fraction population of the reference state $F_0(T)$ (black), final state $F_n(T)$ (red), and intermediate states $F_I(T)$ (blue) as a function of temperature for all three scans. 70

List of Abbreviations

A260: Absorbance at 260nm

A280: Absorbance at 280nm

ABD1: N-terminal actin-binding domain

ABD2: Second actin binding domain

APS: Ammonium persulfate

BMD: Becker muscular dystrophy

BME: beta-mercaptoethanol

CD: Circular Dichroism

CH: Calponin-homology domains

ΔC_p: Change in baseline heat capacity

DEER: Double electron-electron resonance

ΔG: Free energy of unfolding

$\Delta G(37°)$: Free energy of unfolding at 37° Celsius

ΔH_{T_m}: Transition enthalpy

DMD: Duchenne muscular dystrophy

dmd.nl: Leiden muscular dystrophy database

DNA: Deoxyribonucleic acid

DSC: Differential scanning calorimetry

DTT: Dithiothreitol

Dys: Dystrophin

E. coli: Escherichia coli

EDTA: Ethylenediaminetetraacetic acid

EPR: Electron paramagnetic resonance

f-actin: filamentous actin

HCl: Hydrochloric acid

HEPES: 2-[4-(2-hydroxyethyl)piperazin-1-yl] ethane sulfonic acid

His tag: 6x polyhistidine tag

IDPs: Intrinsically disordered proteins

IPTG: Isopropyl-β-D-1-thiogalactopyranoside

KCl: Potassium chloride

LB: Luria broth

MBP: Maltose binding protein

MD: Muscular Dystrophy

MOPS: 3-(N-morpholino) propanesulfonic acid

NaCl: Sodium chloride

NaOH: Sodium hydroxide

PAGE: Polyacrylamide gel electrophoresis

PDB: Protein Databank

SASA: Solvent accessible surface area

SDS: Sodium dodecyl sulfate

TAE: Tris base, acetic acid and EDTA

TEMED: Tetramethylethylenediamine

TEV: Tobacco etch virus

T_m: Transition temperature

Utr: Utrophin

Utr ABD1: Wild type Utr ABD1

Introduction

1

> *In understanding the thermodynamic stability and characterizing unfolding transitions in Dystrophin we are offered insight into the major energetic states.*
>
> ...
>
> *This thesis presents the analysis of the denaturation of ABD1 as an evolution of our understanding of the nuances encompassed within thermodynamic transitions.*
>
> — **Main Ideas**

Understanding the energetics governing cells and their processes is an extremely complex question. Even when narrowing the question to the point of only considering a single class of macro-molecules or even a single protein or structure researchers debate on the nature of the energetics.[2, 3] The problem lies in defining the experimental conditions: what is the ionic strength?, at what pH?, what temperature?, is mechanical stress involved?, are the protein binding partners accounted for?, what experimental signal will be measured?, etc. All of these are questions that are important to the function of a protein and a cell, but are nearly impossible to probe thoroughly.[4] Researchers in the medical sciences are particularly burdened

with problem of generalizing in vitro principles to in vivo applications because of the associated complexity.[5–9] Studies on Muscular Dystrophy (MD) focusing on the Dystrophin protein are troubled by a multitude of difficulties that will be highlighted in Sec. 1.2 and a particular emphasis on our studies using DSC, but first an introduction to MD is needed for physiological and medical context.

1.1 Muscular Dystrophy

Muscular Dystrophy (MD) is the most common genetic disease affecting muscle tissue.[10] MD affects roughly 1:3,500 males.[11] It is almost exclusively documented in males given that it is an X-chromosome sex-linked disease (see Fig. 1.1.[11, 12] Patients expressing MD typically have genomic and sometimes proteomic sequencing completed to document underlying causes.[13] Thorough documentation has highlighted differential effects based on the nature of the mutation at the amino acid level, which are stored in the Leiden database on Muscular Dystrophin ($dmd.nl$).[11, 12, 14, 15] The two major classes of MD are classified as Duchenne's (DMD) and Becker's MD (BMD) which correspond to premature truncation and point mutations in the protein product, respectively.[13, 14] The severity of the phenotype mirrors the extent of amino acids changed in the Dystrophin protein making DMD diagnoses far more problematic than BMD.[13, 15, 16] Disease causing point mutations have been documented in every folding domain of Dystrophin. Given the length (3,685 amino acids) of Dystrophin and its repeating structural features (Spectrin repeat domains, SPR), it seems unusual that point mutations would cause a loss of function at almost any point along its sequence. Disease causing point mutations suggests the importance of every amino acid, or at least important regions of amino acids, in the overall function of Dystrophin.

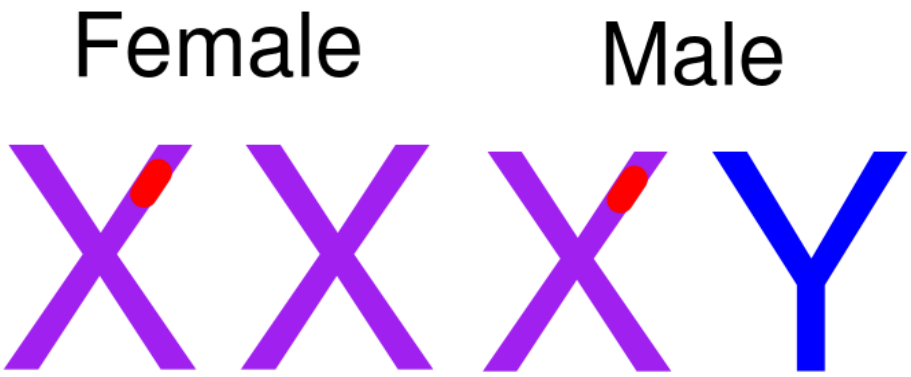


Fig. 1.1: A representation of a mutated Dystrophin gene (red) which is located on the 'X' chromosome. Left: The image represents a female's sex chromosomes. Right: An image representing a male's sex chromosomes.

Beyond the extensive documentation linking Dystrophin to MD, not much is known mechanistically about Dystrophin. Physiologically, the presence of Dystrophin has been shown to increase the stability of membrane surfaces to sustained stresses.[17, 18] Furthermore, patients with MD are marked by a decreased resilience in their muscular tissues from repeated stress.[19] Given the noted loss in structural integrity of muscles and the chemical integrity of the sarcoplasmic membrane, the current understanding is that Dystrophin stabilizes the sarcoplasmic membrane with respect to mechanical strain, but it is not known how.[20, 21] Dystrophin has also been shown to affect the structural dynamics of f-actin, thus likely implicating ABD1.[21]

Not every point mutation leads to a diseased phenotype, and even still some disease causing point mutations are more problematic than others.[13] The differential effects of these mutations based on what domain they occur in are indicative of the overall importance of each respective protein domain and has even inspired the design of gene therapies aiming to deliver shortened protein transcripts of Dystrophin (micro-Dystrophin) and its homolog Utrophin.[20–23] The designed proteins for treatment are shortened Utrophin or Dystrophin segments that fit within the size lim-

itations of current gene therapy techniques. The design of these shortened proteins is based off a design principle proposed to retain the key active features of Dystrophin while reducing or removing other portions that are less necessary for proper function, which is hinted at through relating the severity of a given point mutation to the domain in which it occurs.[21] Since disease causing point mutations have been noted in just about every folding domain in Dystrophin, we defined the protein domain for our studies through analyzing the documented mutation data.[13] Our aim was to study a domain of Dystrophin that was representative of the diseased condition and hopefully propose a mechanism from the wild type protein that would offer insight into its malfunction. The next section (Sec. 1.2) will focus on the Dystrophin protein at the amino acid level, the folding domain level, and its location in muscle cells to provide a molecular context.

1.2 The Dystrophin protein

Dystrophin is one of the largest proteins expressed in the human body with 3,685 amino acids and a molecular weight of approximately 427 kDa. It is incorporated in a larger system within cells called the Dystrophin glycoprotein complex.[20] Dystrophin is a naturally occurring protein implicated in the mechanical integrity of muscle cells as it is likely a mechano-transducing protein.[17, 18] Its biophysical characteristics have proven hard to study as its large size makes it hard to purify. The challenge lies in expressing and purifying a protein of its massive size without aggregation and the limited expression capabilities in E. coli.[24] Beyond this, having reversible processes would be extremely unlikely. Furthermore, the lack of enzymatic activity makes Dystrophin difficult to study in that it does not have a chemical assay to probe its function because it likely serves as a structural protein with no known catalysis.[20, 25] To address the difficulties in size, we chose to study a domain that

would be both representative of the mechanism for the fully functional Dystrophin protein and encompass a large portion of the relevant disease causing mutations; ABD1 covered both of these criterion. Choosing a small representative portion of Dystrophin is consistent with the design philosophy researchers are currently using to develop mini-Dystrophin as discussed in Sec. 1.1.[21]

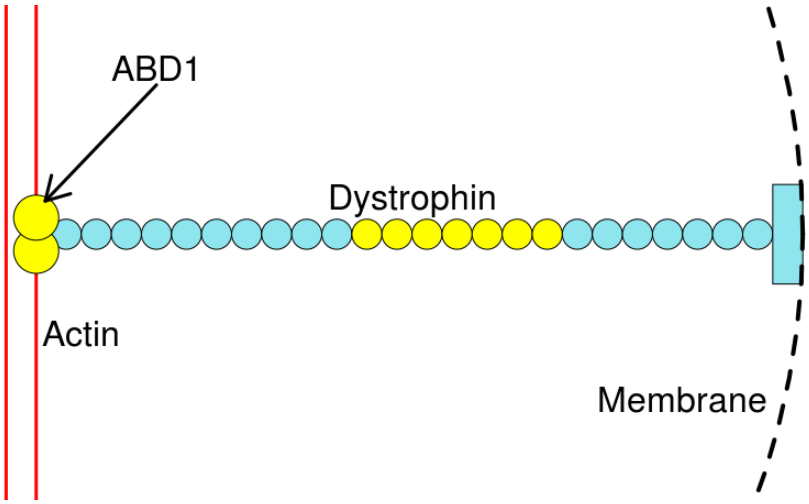


Fig. 1.2: A cartoon representation of Dystrophin. The yellow regions denote actin binding regions, the red vertical lines represent the f-actin cytoskeleton matrix, the curved dashed line represents the sarcolemma. Each circle along Dystrophin represents a folded domain with the left two overlapping circles representing CH1 and CH2 of ABD1, all subsequent circles representing the large portion of SPR's, and the rectangular region representing the WW domain, which is rich in cysteine residues and couples Dystrophin to the sarcolemma (dashed, curved line). The second actin binding domain (ABD2) can be seen as the second string of yellow circles in the central region of Dystrophin. All reported domains are based off of speculative interaction predictions from homology mapping and are provided by Uniprot.[26]

A cartoon image of Dystrophin is provided in Fig. 1.2 displaying the basic arrangement of Dystrophin joining the actin cytoskeleton to the sarcoplasmic membrane. The putative actin binding domains are represented in yellow with ABD1 being represented by the two left most circles joined to the cytoskeleton represented as red lines, see Fig. 1.2.[20] 27 SPR's are represented as small circles all of which are highly homogeneous. As noted in Sec. 1.1, disease causing point mutations have

been noted in just about every folding domain. This suggests that while the SPR's are extremely homologous domains, they are not redundant, with respect to the function of Dystrophin. On the contrary they all must be necessary to some extent for the proper function of Dystrophin. So, in studying the function of Dystrophin we had to be deliberate in choosing a small representative section.

The identity of mutations that were reported were mined from the dmd.nl database.[13] The total number of mutations occurring at each amino acid position are graphed in the left hand chart in Fig. 1.3. Some hot spots for mutations are noted but are difficult to interpret at face value given the complex splicing of Dystrophin and how that might naturally lead to mutation hot spots. However, the practically uniform distribution of mutations that were documented in the ABD1 region were particularly intriguing given that no other region showed such a uniform distribution (left graph 1.3, in red). Statistical analysis was done on the documented sequences and was utilized to hone in on studying the ABD1 domain. One key finding revealed a large density of mutations occurring in the ABD1 region, as seen in Fig. 1.3 as the Red circle. These calculations took the total number of disease causing mutations and divided it by the total amino acid length of that region. The result was called the 'Mutation Density' which is seen along the y-axis on the right hand graph of Fig. 1.3. A correlation to the disease causing likelihood of the mutation to the associated change in heat capacity was noted.

A set of statistical tests were ran comparing corresponding changes in physical parameters to see if the differences in these physical parameters were significantly different. Comparisons were developed between buried and exposed amino acids (determined through a script ran in Pymol), and the consequences of the mutation: being diseased (+), non-diseased (-), or uncertain diagnosis(?). Having a trinary scale doesn't offer any extra insight into characteristics of the mutations, so we

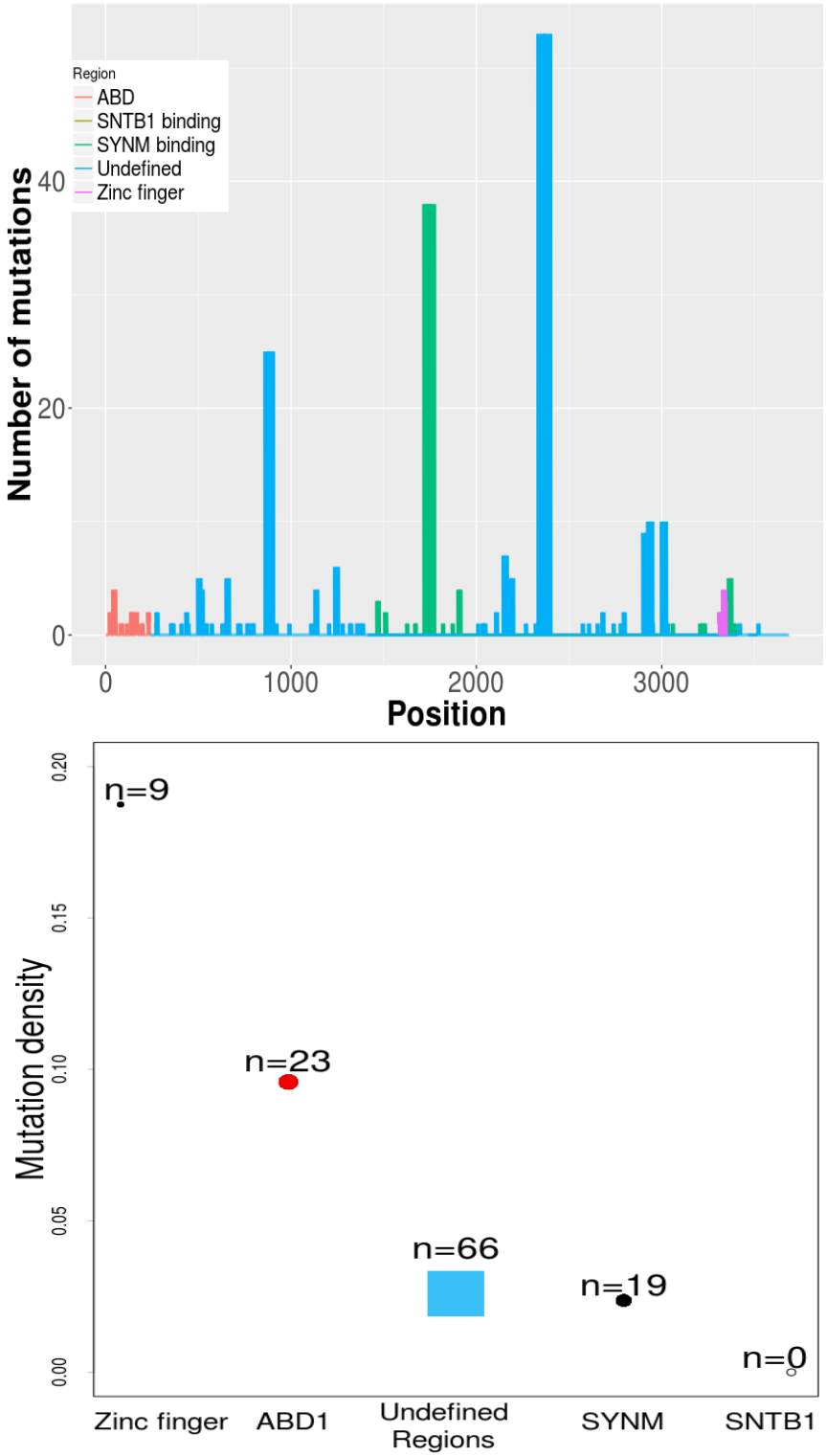


Fig. 1.3: Left: The total number of mutations at each amino acid position. Protein regions are designated by color. Right: The mutation density of each protein region. The red circle represents ABD1 and the blue square represents regions that are undefined as per Uniprot.

p-values for ANOVA and t-tests on unsure diagnoses (?) vs. disease (+) and non-disease causing (-) comparing changes in heat capacity

	Buried (?)	Exposed (?)
Buried (+)	0.74 (0.34)	0.86 (0.02)
Buried (?)	-	0.69 (0.65)
Buried (-)	0.35 (0.66)	0.62 (0.35)
Exposed(+)	0.73 (0.03)	0.73 (0.02)
Exposed(?)	0.69 (0.65)	-
Exposed(-)	0.01 (0.01)	0.02 (0.03)

Tab. 1.1: A comparison of the variances and means between ? samples and other samples to determine if the ? samples are from the same population as the + or - populations. The first value in each cell represents the ANOVA p-value The second value in each cell (in parentheses) represents the p-value from a t-test of the appropriate nature depending on the output from the ANOVA test. Significant values are underlined. When variances were equivalent we used a two sample t-test, when they were not equal we used Welch's t-test.

compared uncertain diagnoses(?) with positive and negative diagnoses to see if they were significantly different. This was a two-tailed test with the null hypothesis being that both populations are the same. The table represents the p-values from ANOVA and t-test evaluations. The left-hand value represents the p-value from an ANOVA, which is the likelihood of the compared values to have the same variation in values. The right hand value reported within parentheses represents the p-value for the two values to have different means through a t-test. The t-test that was ran was dependent on the results from the ANOVA analysis. If the variance was the same, a standard two sample t-test was used. For differing variances, the Welch's t-test was used. Values of significance ($p < 0.05$) are underlined. The results are recorded in Tab. 1.1. The results showed enough difference between the uncertain diagnoses and the certain diagnoses, that it was determined that they didn't fit cleanly into either category. Therefore, uncertain diagnoses were not considered for further evaluations. The positive and negative diagnoses were then compared using the same statistical analysis.

Next, we wanted to test whether the characteristics of disease causing mutations differed depending on the solvent exposure of a given residue. Given what is normally taught as the hydrophobic effect (that hydrophobic residues are driven to the center of a protein fold and serve as the main stabilizing feature), the hypothesis was that disease causing mutations that are buried would likely have a negative change in heat capacity, whereas for exposed residues the change would be positive. The results from the comparison of the predicted changes in heat capacity from the mutations comparing buried to exposed residues are reported in Tab. 1.2. A statistically significant difference was noted between Buried and Exposed disease causing mutation. This is consistent with a disease mechanism that destabilizes the folded structure of ABD1, either decreasing the hydrophobicity for buried residues or increasing the hydrophobicity for exposed residues as these would both serve to destabilize the folded structure of ABD1 according to the canon of the hydrophobic effect.

p-values for ANOVA and t-tests comparing changes in heat capacity

	Buried (+)	Buried (-)	Exposed (+)
Buried (-)	0.59(0.32)	-	-
Exposed(+)	0.003(0.003)	0.50(0.32)	-
Exposed(-)	0.14(0.10)	0.69(0.34)	0.99(0.88)

Tab. 1.2: p-values from ANOVA (first value) and t-test (value in parentheses) between buried/exposed and disease causing(+)/non-disease causing mutations(-). p-values adjusted by Tukey's honest significant differences method and the Benjamini/Hochberg method.[27] Statistically significant values are underlined. The second value in each cell (in parentheses) represents the p-value from a t-test using Tukey's Honest Significant differences method. Significant values are underlined.

Given the predicted changes in physical parameters of the wild type and mutated amino acids as is shown in Tab. 1.2, we decided to utilize DSC to study the construct because it is the only method that can directly measure the change in heat capacity of a protein unfolding transition. DSC also measures an innate signal, not requiring any

dyes or other tags, which might mar the physiological interpretation. The mutations primarily affected the change in heat capacity predicted between the folded and unfolded states thus implicating them in the overall stability of ABD1 among other physical predictions.[28, 29] The correlation of changes in heat capacity with disease causing mutations is also supported by the work of Mallela et al. who noted effects on aggregation and secondary structure formation from a few of the noted disease causing mutations.[25] We then hypothesized that the functionality of Dys ABD1 is dependent on the change in heat capacity between folded and unfolded states. Specifically, in this study, we aimed to test whether Dys ABD1 had a smaller change in heat capacity than would be predicted from its sequence because of the large number of exposed hydrophobic residues noted in its crystal structure. We tested this hypothesis using DSC.[28]

DSC is an extremely useful technique that utilizes the innate thermal signature of a molecule (i.e. heat capacity) in determining the strength of interaction by pushing it through a phase transition and directly measuring the energy differential between a cell containing the molecule of interest and cell containing the buffer solution.[28] Through using the innate thermal signature of the molecule of interest, there is no need for any extra dyes or tags to be used thus introducing fewer parameters to constrict its generalizability and physiological relevance. The difficulty with the technique however, is that it is difficult to analyze. Problems such as the difficulty in determining a baseline, determining the number of transitions represented in a given peak, or the reversibility of the transition .[28, 30–36]

Beyond this, interpreting the confidence of fitted values is oftentimes overlooked. Given its current importance, many papers have been recently published on how to increase the verity of DSC data interpretation.[33, 34] Here I present the evolution of analysis that we have developed that implemented techniques that

have been previously put forth (fitting to Gaussian and Hubbert distributions, and deconvolution analysis)[35, 36] and contribute our methods for establishing bounds for the pre-transition baseline definition in accordance with two-state theory and a method for analyzing the shape of the confidence regions.

1.3 Materials

1.3.1 Reagents

The reagents used for thermodynamic studies were assured, high quality reagents as necessary for quantitative determination of thermodynamic parameters. Potassium chloride (KCl) and sodium chloride (NaCl) salts were Puriss-grade with a purity of >99.5%. They were stored at room temperature. The buffer reagents 3-(N-morpholino) propanesulfonic acid (MOPS) and 2-[4-(2- hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (HEPES) were Biochemika grade powders and received from Fluka Chemical Corp and they were stored at room temperature. Reagents responsible for divalent cation sequestration 2,2',2",2'"-(Ethane-1,2-diyldinitrilo)tetraacetic acid (EDTA) and Ethylene glycol-bis(2-aminoethylether)-N,N,N',N'-tetraacetic acid (EGTA) were Biochemika grade powders and received from Fluka Chemical Corp, as well and were stored at room temperature. Urea (>98% pure) and imidazole (>99% purity) were obtained from Sigma-Aldrich as dry reagents and stored at room temperature with a carefully sealed lid. DNase and RNase free Dithiothreitol (DTT) was obtained from ThermoFisher Scientific as a dry reagent and refrigerated in a desiccator at 4°C. Potassium chloride tablets used for pHing were >85% pure from Sigma Aldrich and stored at room temperature. 2-Mercaptoethanol (BME) was obtained 99.0% pure from Sigma Aldrich and stored at 4°C. Two grades of methanol were used. The higher purity methanol was 99.8% pure, obtained from

Sigma Aldrich and was stored at room temperature. The lower purity methanol was 95% pure, obtained from Thermofisher Scientific and stored at room temperature. Laboratory grade ethanol used for disinfectant and some nucleotide techniques was 98% pure received from Thermofisher Scientific and stored at room temperature. Glacial Acetic Acid that was 99+% pure was received from ACROS organics. Anhydrous Cholorom (purity >99%) was obtained from Sigma-Aldrich for washing dishware that would also be used for lipids experiments. Household bleach supplies were used as a sterilization technique for discarding liquid biohazard waste such as bacterial cultures. Phenylmethylsufonyl Fluoride (PMSF, 98.5% pure) was obtained from Sigma-Aldrich to serve as protease inhibitors for the lysing process. 98% pure Sodium Hydroxide (NaOH) was received from Sigma-Aldrich. Tris-HCl was order from Thermofisher Scientific. Tris Base was obtained from Thermofisher Scientific. TAE stock buffer was obtained from Thermofisher Scientific. Glycerol (99.5% pure) was obtained from Sigma-Aldrich. DNase and RNase free, deionized water was obtained from Thermofisher Scientific. Tris base was obtained from Thermofisher Scientific. Sodium Dodecyl Sulfate was ordered from Sigma-Aldrich and was >98.5% pure. Hydrochloric acid (HCl) was obtained from Thermofisher Scientific at 25% purity. Sodium dodecyl sulfate Polyacrylamide Gel Electrophoresis (SDS PAGE) reagents were all obtained in a kit from Bio-Rad including 40% acrylamide solution, 1% bisacrylamide solution, 1.5 M Tris-HCl pH8.7, 10% ammonium persulfate (APS), and TEMED. The 10% crosslinking gel utilized was made from 7.5 mL 40% acrylamide solution, 3.9 mL 1% bisacrylamide solution, 7.5 mL 1.5 M Tris-HCl, pH 8.7, Add water to 30 mL, 0.3 mL 10% APS, 0.3 mL 10% SDS, 0.03 mL TEMED. Dry premixed Luria Broth (LB) growth media was acquired from Thermo Fisher Scientific. Individual components to make our own were acquired separately. Yeast Extract and Peptone were acquired from Sigma-Aldrich as dry powders. Isopropyl-β-

D-thiogalactoside (IPTG) was acquired from Sigma-Aldrich as a dry powder. Ni-NTA nickel agarose affinity resin was bought from Qiagen. Guanidinium Hydrochloride (98% purity) was acquired from Sigma-Aldrich. Benzonuclease (>250 units/μL, >90% purity) DNase and general nuclease was acquired from Sigma-Aldrich. Anhydrous Magnesium Chloride (98% purity) was bought from Sigma-Aldrich as a dry powder. Snakeskin Dialysis tubing 10k and 15k MWCO 22mm was acquired from Sigma-Aldrich.

1.3.2 Glassware

Pyrex Erlenmeyer flasks ranging from 15mL to 500mL, 2.8L Pyrex growth flasks rated for 1.2L of growth, Pyrex rated beakers ranging from 10mL to 4L, glass spacer plates (0.75 and 1.0 mm spacings) and short mini-PROTEAN plates from Bio-Rad, 1.5mL Eppendorf tubes, 1.5mL cryogenic storage tubes from Thermofisher Scientific with color coded screw tops, 15mL borosilicate test tubes from Corning Scientific, 15 and 50 mL falcon tubes, 50mL Oak Ridge tubes, 500mL Centrifuge tubes from Thermofisher Scientific, single use culture plating sticks from Bio-Rad, Petri dishes from Thermofisher Scientific, and a bunsen burner with a flint starter were all utilized for different aspects of the experimental work from transforming cells to cell growth to protein purification.

1.3.3 Instrumentation

Many different instruments were utilized to complete the work outlined in this thesis. A NanoDSC Differential Scanning Calorimeter (DSC) from TA Instruments (New Castle, DE) was used to directly measure the heat capacity of the system being measured. A Fluoromax was used to correlate with calorimetric data, but ultimately didn't give altogether useful data. A Jasco J-1000 CD was utilized when

data was still being taken to measure other unfolding signals from the protein construct. A couple types of spectrophotometers were used. A Nanodrop was utilized to determine concentrations of nucleic acids and proteins based on absorbance profiles. Similarly, a Varian Cary 50 Bio and a Beckman DU 640 spectrophotometer were utilized to measure the concentrations of protein and nucleotides in a given sample. SDS-PAGE gels were run using a PowerPac Basic distributed by Bio-Rad. A Sorvall RT 6000B and a Fisher Scientific accuSpin Micro 17 were generally used for centrifugation processes. Cell cultures were grown up in a Large growth coffin shaker from New England Biolabs (NEB). Gel images were taken utilizing a Chemidoc gel viewer from Bio-Rad with a Universal Hood II from Bio-Rad. Buffer Samples were dissolved utilizing a Vortex Genie 2 from Scientific Industries. Cell lysis was carried out utilizing a Branson digital sonifier with a metal sonicating tip. Gel staining and destaining was carried out in Tupperware containers and speed up by the use of a Sanyo standard Microwave. Gas was evolved out of samples prepared for measurement in the DSC utilizing an Gast pump capable of creating a -1 bar pressure differential. Dry reagents were massed on a Mettler Toledo XS64 analytical balance. Cell transformations were heated in a Precision Microprocessor Controlled 280 Series Water Bath. Some reactions were incubated on a Fisher Scientific, Nutating Mixer, 1.75-lb Load Capacity.

1.4 Methods

1.4.1 Plasmid construct design

A DNA plasmid construct was designed (Fig. 2.2) following that outlined by David Waugh.[37] The design calls for the desired protein construct to be bound to a Tobacco Etch Virus protease recognition site and ultimately to a Maltose binding

protein with a 6x His tag on the N-terminal. To make the construct design, sequences were found on Uniprot for the Maltose binding protein, the TEV recognition sequence was taken from Waugh's documentation of the protease, and six extra histidines were entered on the N terminus. With the design laid out on the amino acid level, we translated the amino acid sequence into a nucleotide sequence through a reverse translation calculator that is available online called "in-silico". $(http://in-silico.net/tools/biology/sequence_conversion)$ The sequence was checked to assure its compatibility within an E. coli system. Once designed, the nucleotide sequence was sent to Genewiz for synthesis and verification. The sequence can be seen in Fig. 2.1

The designed construct was verified and inserted into a well behaved protein expression plasmid called pet-28-MBP-TEV by Genewiz (picture in Fig. 2.2). The plasmid also contained key features necessary for expression such as antibiotic resistance, and a Lac Operon promoter region. Upon receiving the plasmid with the mouse Dys ABD1 protein fragment (residues 8-246 from Uniprot sequence P11531, Fig. 2.1). The plasmid was transformed into BL21 DE3 E. coli (DNase wounded) and subsequently grown to an OD of 0.7 at which point live stocks were put in a 50% glycerol stock and frozen in liquid Nitrogen and stored in -80$^{\circ}C$. Then, the plasmid was transformed into Rosetta strain E. coli (protease wounded) via heat shock. The cell line was also stored in glycerol stocks at -80$^{\circ}C$. The Rosetta strain was subsequently used for all downstream protein expression and purification purposes.

1.4.2 Cell growth and protein purification

Once competent cells were verified to contain the plasmid of interest, stored in the -80$^{\circ}C$ freezer, and plated on an antibiotic containing plate, a small seeding culture (approximately 12mL) was grown up overnight in a test tube containing LB

media with an selective antibiotic. The Culture was grown up at 37°C overnight and shaken at 250rpm. A negative control of the same LB media containing antibiotics was grown up to control for experimental technique. When turbidity was observed in the experimental test tube only, the second set of small growths were seeded. These test tubes contained fresh LB growth media with antibiotic and were also approximately 12mL. The second set of growths were seeded from the first overnight growth. Then, once the second set of growths reach the end of log phase, with an OD 0.8, the second set of small growths were used to seed large growths varying from 1L to 1.2 L in volume. Large growths were carried out in 2.4L baffled bottom flasks to ensure proper oxygen exchange once the growths become particularly dense. The each large growth was monitored separately using the OD600 reading. Once the OD reached 0.4- 0.5 the cells were induced to start expressing the protein construct by doping the growths with 2mM final concentration of Isopropyl β-D-1-thiogalactopyranoside (IPTG). Growths were continued at 37°C while the cells remained in log phase and it was terminated once the cells reached a phase of slowed growth.

1.4.3 Lysis and protein purification

Cells were then pelleted in a Beckman-Coulter Avanti J-E centrifuge with F10BClx500cy rotor at 16,000rpm, for 15 minutes. Cell pellets were then resuspended in an appropriate amount of lysis buffer containing PMSF, BME, EDTA, at pH 8.0. Lysis was done utilizing a sonicating tip at 70% power alternating between one second bursts and four second rests for a total time of 40 minutes. Cellular debris was separated from soluble material by centrifuging in a Beckman-Coulter Avanti J-E centrifuge with J25.5 rotor at 16,000 rpm for 40 minutes. The unsettled insoluble particles were then removed utilizing an 0.2μm size syringe filter. The

buffer conditions were then controlled via a round of dialysis with 100mM KCl 20mM HEPES. Subsequently, large pieces of DNA were broken down, to prevent binding to the Nickel column, through an incubation with Benzonuclease and its cofactor Magnesium. The extra Magnesium and smaller DNA fragments through dialysis. The product was then bound to a Nickel-Agarose bead column and brought to equilibrium by rocking gently for three hours on a New Brunswick Scientific Co., INC. Series 25 Incubator shaker. Then the non-bound protein was washed away by letting the column release its soluble fraction. Nonspecifically bound material was washed from the column with an imidazole buffer gradient containing 20mM HEPES, 20mM NaCl, and upwards of 250mM Imidazole. Other nonspecific electrostatically bound particles were washed from the column with a 250 mM KCl wash. The protein construct containing Dystrophin ABD1 connected to Maltose binding protein and a 6x Histidine tail through a TEV cleavage site was then eluted off of the column with the same imidazole buffer containing 150mM imidazole. Imidazole was dialyzed out and the purity of the construct was checked using SDS-PAGE gel verification and can be seen in Fig. 2.3. The calculated molecular weight of Maltose binding protein is 44kDa and of ABD1 is 27kDa and were deemed pure based off of the SDS-PAGE analysis based on the fact that the only major band noted in flow through fractions was at 27kDa. Purity of the protein was also considered using a comparison of 260nm and 280nm absorbances on both the Nanodrop and a Beckman DU 640 spectrophotometer. The A260/A280 ratio was considered pure if it was 1.0 or smaller and this sample had a ratio of 0.8. The final protein concentration was then determined using Beer's law, the expected A280 absorbance value for the construct based on the number of aromatic residues in its sequence, and the A280 as measured in replicate by the Beckman spectrophotometer. Cleavage was then carried out by an overnight incubation with in-house purified Tobacco Etch Virus protease. A second

Nickel column was ran, but this time the ABD1 portion flowed through the column without binding. The same purity tests and concentration determination was done on the purified ABD1 region and can be seen in Fig. 2.3. ABD1 was deemed pure from these images based on the fact that the only clear band appears at 27kDa.

1.4.4 Differential scanning calorimetry (DSC)

DSC experiments were performed on a NanoDSC (TA Instruments, New Castle, DE) at a scan rate of 1°C/min and 1.5°C/min and under a pressure of 3 atm, as described previously.35,69 All scans were conducted in buffers containing 20 mM MOPS, 100 mM KCl, pH 7.5. Hysteresis was minimized by measuring the heat capacity of the experimental buffer first and assuring that the difference between two subsequent scans was purely encompassed by the standard resting fluctutations, then the experiment with protein sample would begin. The reference cell would always contain either double deionized water or the buffer, whereas the sample cell would contain water (for experiments to keep the instrument warm and reduce hysteresis before buffer scans), buffer (to prime the cells and remove hysteresis effects before loading the protein sample), or the protein sample (for the final experimental work).

All samples including water and reference samples were degassed prior to loading for 15 minutes while being stirred vigorously to reduce the likelihood of gas bubble expulsion during experiments. The solubility of gases in liquids decreases with temperature and thus makes the evolution of a bubble more likely if samples are not properly degassed. The expulsion of bubbles ruin experimental measurements given that the measured heat capacity of the cell is drastically changed when gaseous Oxygen is taking up much of the surface area where the heating coil can be in contact with the sample. To further ensure no gas bubbles will escape the sample,

experimental verification of our loading technique was verified by letting the DSC come to equilibrium after loading, then pressurizing the cells. Since, there were never any drastic changes as defined by the manufacturer it is unlikely any bubbles have evolved and the experiment can be carried out. After verifying that our technique did not introduce bubbles on multiple separate instances, this step was omitted in an effort to limit the samples exposure to room temperatures.

1.4.5 Software

R- R is a statistical programming language and environment specifically developed for computing and graphics with large data sets. The 'Shiny' package was used to create interactive scripts in R. Not only are they interactive, but they are interactive through graphics making them GUI's and therefore much more user friendly as they don't require any knowledge of R scripting to use. The 'Shinydashboard' package helped speed the development of shiny interactive apps, by handling many of the aesthetic details by default and creating a dashboard like set up that is very intuitive and user friendly. The 'ggplot2' package was used to generate all plots. The package uses a 'grammar of graphics' that is state-of-the-art in the graphic design community. The 'dplyr' package was utilized for fast manipulation of the mutation population data. It utilizes systems of filters and data selection that made the functional programming of general arguments possible. This in turn made it possible to program the mutation analysis into an GUI application thus making the mutation analysis inheritable. The 'nls2' package was used to add additional rigor to the minimization with a new regression option 'brute force'. The 'brute force' approach allows for a vector input of starting values that minimizes each of them separately thus probing a small grid of seeding values for the regression as opposed to a single point. In this thesis an option for utilizing the brute force algorithm was

accounted for by creating a series of points centered about the graphical prediction with user defined radius of values around the estimate. The 'nlstools' package was used to allow for the analysis of confidence from nonlinear regression analysis with confint2, nlsConfRegions, and nlsContourRSS. These algorithms were specifically designed following the text of Bates and Watts originally proposed by Beale. [38, 39] Microsoft Excel was used to develop the initial models interactively in spreadsheets. The spreadsheets served as an intuitive environment to develop the models because all values are visible to the user. However, bugs became much harder to identify and perhaps more importantly, in running nonlinear regressions on large datasets computers were prone to crashing. The 'Solveraid' Microsoft Excel add-in was used for the statistical evaluation of nonlinear regressions via post hoc analysis. Nanoanalyze software that accompanies the DSC running software was used to read the .dsc file extension and was used to convert into .xls and .csv file types for use with the in-house scripts and Excel worksheets discussed.

2 Theory and comparison between methods of analysis

> *[This analysis lead to more] consistent results as well as more reliable confidence estimates to accompany the model fit parameters*
>
> — **Main Result**

This chapter outlines the theory used for DSC analysis and expands into more complex models and introduces a new method for the nonlinear regression and evaluating the number of states. It outlines the methods used for estimating parameter fits through utilizing simpler distribution functions (Gaussian and Hubbert) that lead to a more intuitive interpretation and compares it to the current method. It describes some of the new features such as confidence contour analysis and different seeding approaches to assure the fit is a minimum. We also compiled fit parameters on multiple data sets to test if this approach is generalizable. Then we evaluate ABD1 as displaying transition behavior suggestive of the occupancy of one or more intermediate states through a deconvolution approach.[28, 36] First, however I will discuss the results of the design and transformation of the gene into E. coli and the purification of the expressed protein.

2.1 Construct results

MLWWEEVEDCTERFDVQKKTFTKWINAQFSKFGKQHIDNLFSDLQDGKRLLD
LLEGLTGQKLPKEKGSTRVHALNNVNKALRVLQKNNVDLVNIGSTDIVDGNHK
LTLGLIWNIILHWQVKNVMKTIMAGLQQTNSEKILLSWVRQSTRNTPQVNVINF
TSSWSDGLALNALIHSHRPDLFDWNSVVSQHSATQRLFHAFNIAKCQLGIEKL
LDPEDVATTTPDKKSILMTITSLFQVLP

Fig. 2.1: Sequence of the Dystrophin ABD1 amino acid sequence cloned for purification and thermodynamic studies. Amino acids are colored by their identities.

The details of the gene designed to express Dystrophin ABD1 are discussed within the methods section. The designed gene is represented in Fig. 2.1 where amino acids are colored by their identities. This sequence was taken from Uniprot and then converted to a DNA sequence by the Genewiz company through a reverse translation algorithm. Genewiz synthesized the gene and subsequently ligated it into the pet28-MBP-TEV plasmid. The final ligated product was confirmed through a sequence analysis. The final plasmid construct is visualized in Fig. 2.2. The plasmid construct contains the synthesized ABD1 gene inserted at the BAMHI site. The plasmid backbone already contains antibiotic resistance (Kanamycin) as well as the His-tag, MBP, and TEV cleavage site preceding the BAMHI cut site. After transformation into E. coli cells and expression of the ABD1 protein the protein was purified. However, the originally designed construct was unable to be cleaved via TEV protease. Therefore a plasmid construct that had been utilized by the Thomas group from the University of Minnesota was borrowed. Studies on this construct are reported in a 2012 paper from the Thomas group.[21] The construct acquired from the Thomas lab was then expressed and the protein product purified, the results of which are shown in Fig. 2.3. The protein gels (SDS-PAGE) indicated the presence of a single protein with a molecular weight of approximately 27kDa (the same as

the predicted molecular weight of the protein construct). The protein construct was considered to be pure based off of these gels.

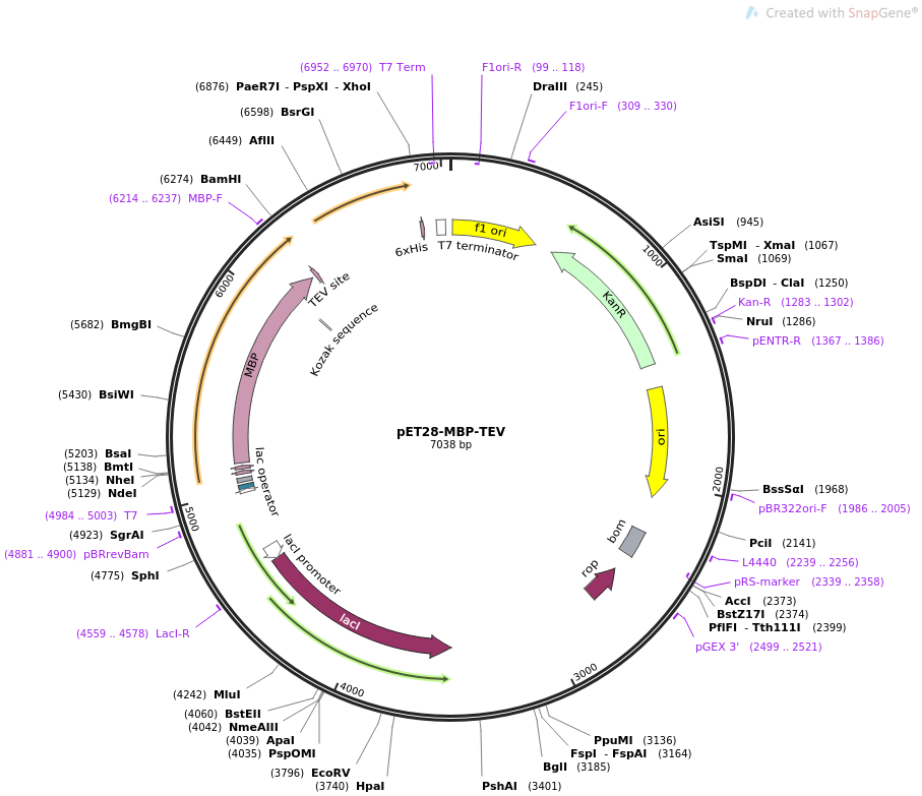


Fig. 2.2: A representation of the pet28-MBP-TEV plasmid design. All of the features are noted included opening reading frames where the ABD1 gene was inserted right after the TEV site and the MBP gene denoted in light maroon at the BamHI restriction site.

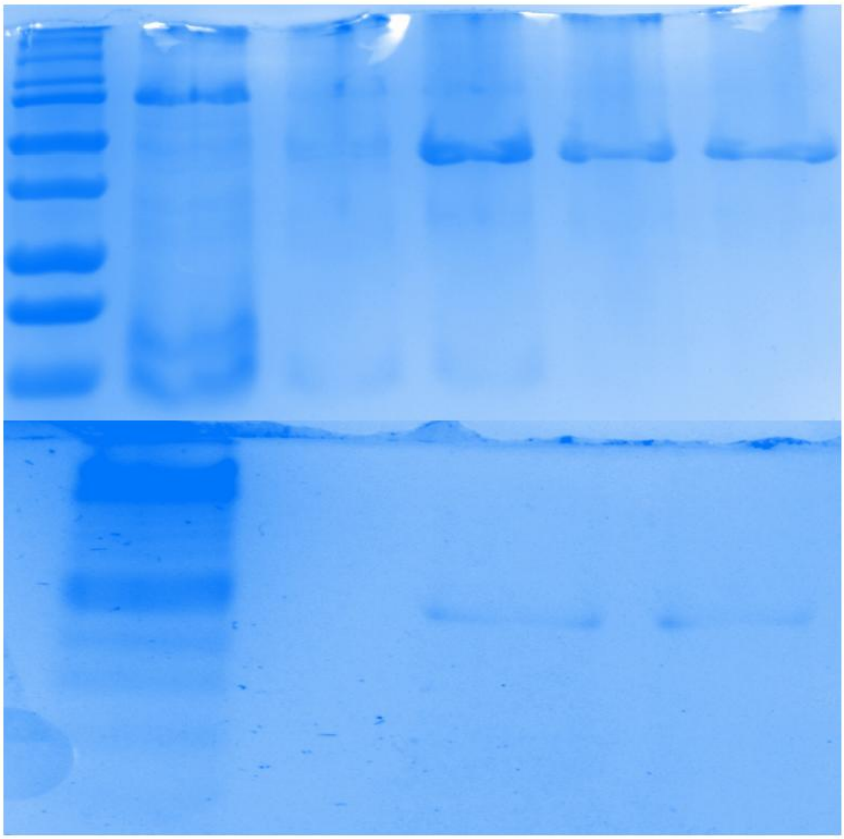


Fig. 2.3: SDS-PAGE gels displaying the purity test for Dys ABD1. The top represents a purification for the first two sets of scans. The contents of each lane are as follows: Unstained Protein Ladder, Broad Range (10-250 kDa) from NEB with molecular weight bands representing 10kDa, 15kDa, 20kDa, 25kDa, 30kDa, 40kDa, 60kDa, 80kDa, 100kDa, 150kDa, and 250kDa; the second column elution with bands at 44kDa and two smaller bands around 10kDa; empty lane; flow through fractions from column two with molecular weights 27kDa. Bottom: Purification for the last scan, with the ladder in the left most lane, an empty lane, then two lanes with the flow through fraction. The molecular weights were difficult to interpret for the second gel but the molecular weight ladder was the same.

2.2 Theory

2.2.1 Two-state equilibrium derivation

Starting with a few well described thermodynamic equations a model for the heat capacity for a two-state protein unfolding process can be derived.[28, 40] Particularly, the chemical process can be represented in the following equilibrium model where N represents the fraction folded (native), U represents the fraction unfolded, and K represents the equilibrium constant between those two-states.

$$\mathrm{N} \xrightleftharpoons{K(T)} \mathrm{U} \tag{2.1}$$

The equilibrium expression is given by

$$K(T) = e^{\frac{-\Delta G(T)}{RT}} \tag{2.2}$$

Where equilibrium (K(T)) is a function of temperature. Eq. 2.3 is the familiar equation relating Free energy to contributions from enthalpy and entropy in Eq. 2.3.

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \tag{2.3}$$

Where $\Delta G(T)$ is a function of temperature (as shown be Eq. 2.3), and $\Delta H(T)$ and $\Delta S(T)$ are also functions of temperature as will be shown in Eqs. 2.4 and 2.5. Kirchoff's law (Eq. 2.4) states that the change in enthalpy between two states is the integral of the heat capacity over the given temperature range.

$$\Delta H(T) = \Delta H_{Tm} + \int_{Tm}^{T} C_p(T)dT \tag{2.4}$$

Here, ΔH_{T_m} is a reference enthalpy which is in this case taken at the transition temperature T_m. The $C_p(T)$ term will simplify to a constant as will be shown in the next couple of steps. Lastly, given that our DSC system is at constant pressure (namely 3atm) We can rearrange the equation $\Delta H = T\Delta S + V\Delta P$ (at constant pressure ΔP =0) to get

$$\Delta S(T) = \frac{\Delta H(T)}{T} \tag{2.5}$$

Putting Kirchoff's equation (Eq. 2.4) into this yields,

$$\Delta S(T) = \frac{\Delta H_{Tm}}{Tm} + \int_{Tm}^{T} \frac{C_p(T)}{T} dT \tag{2.6}$$

Now, putting Eq. 2.4 and 2.6 into Eq. 2.3 we get

$$\Delta G(T) = \Delta H_{Tm} + \int_{Tm}^{T} C_p(T)dT - T(\frac{\Delta H_{Tm}}{Tm} + \int_{Tm}^{T} \frac{C_p(T)}{T} dT) \tag{2.7}$$

Note: It is important to select T_m as the reference temp, because at $T = T_m$, $\Delta G_{T_m} = 0$. So, we set $\Delta G_{T_m} = \Delta G_0 = 0$ by choosing $T = T_m$ as the reference temperature. This is important because the difference in free energy between two states (1 and 2) is given by $\Delta G_{trans} = \Delta G_1 - \Delta G_0$ relative to state 1, where ΔG_0 refers to the absolute free energy of a reference state, ΔG_1 the free energy of another state, and ΔG_{trans} corresponding difference in free energy. However, it is impossible to determine the absolute free energy of a state. By referencing the transition temperature (T_m) where $\Delta G_0 = 0$ we are able to get around this problem because $\Delta G = \Delta G_1 - \Delta G_0 = \Delta G_1$.[28, 40]

The heat capacity ($C_p(T)$) here is due to the difference in heat capacity between the folded and unfolded states of the protein. This difference can be assumed to be

constant with respect to temperature over the scale we will be measuring meaning $C_p(T) = \Delta C_p$.[28]

$$\Delta G(T) = \Delta H_{T_m}(1 - \frac{T}{T_m}) + \Delta C_p(T - Tm - T(ln(\frac{T}{T_m})) \tag{2.8}$$

Eq. 2.8 is called the Gibbs-Helmholtz equation. We bounded our system by the rules of mass conservation. This means that while the protein may change states, it will never be gained or lost, rather distributed differently between these states. For the two-state model this can be expressed as $P_{tot} = P_{unfolded}(T) + P_{folded}(T)$. Where P(T) describes the concentration of the species described by its respective subscript. Here we can express the fractions of the folded an folded states as follows by normalizing the concentration by the total amount of protein in solution to $1 = \Theta_{folded}(T) + \Theta_{unfolded}(T)$, where $\Theta_{folded}(T) = \frac{P_{folded}(T)}{P_{tot}(T)}$ and $\Theta_{unfolded}(T) = \frac{P_{unfolded}(T)}{P_{tot}(T)}$. The equilibrium between these two states is given as $K(T) = \frac{\Theta_{unfolded}(T)}{\Theta_{folded}(T)}$. Plug this into the mass conservation law as follows

$$\Theta_{folded}(T) + \Theta_{unfolded}(T) = 1 \qquad K(T) = \frac{\Theta_{unfolded}(T)}{\Theta_{folded}(T)}$$

$$\Theta_{folded}(T) + \Theta_{unfolded}(T) = 1 \qquad \Theta_{folded}(T)K(T) = \Theta_{unfolded}(T)$$

Plugging the RHS into the LHS

$$\Theta_{folded}(T) + K(T)\Theta_{folded}(T) = 1$$

$$\Theta_{folded}(T)(1 + K(T)) = 1$$

$$\Theta_{folded}(T) = \frac{1}{1 + K(T)}$$

Then given that $\Theta_{folded}(T)K(T) = \Theta_{unfolded}(T)$ we arrive at Eqs.

$$\Theta_{unfolded}(T) = \frac{K(T)}{K(T)+1} = \frac{1}{1+\frac{1}{K(T)}} \tag{2.9}$$

and

$$\Theta_{folded}(T) = \frac{1}{K(T)+1} \tag{2.10}$$

Finally, given $\Delta G(T) = -RTln(K(T))$ and using Eq. 2.9 we get

$$\Theta_{unfolded}(T) = \frac{1}{1+\exp\frac{\Delta G(T)}{RT}}. \tag{2.11}$$

And subsequently, with Eq. 2.8 this becomes

$$\Theta_{unfolded}(T) = \frac{1}{1+e^{\frac{\Delta H_{Tm}(1-\frac{T}{Tm})+\Delta C_p(T-Tm+\Delta C_p(ln(\frac{T}{Tm}))}{RT}}}. \tag{2.12}$$

As the calorimeter increases in temperature, the protein unfolds and changes the heat capacity of our system as follows using the product rule

$$C_p(T) = \frac{d}{dT}(\Theta_u(T)\Delta H(T)) = \frac{d}{dT}(\Theta_u(T))\Delta H(T) + \Theta_u(T)\frac{d}{dT}(\Delta H(T)) \tag{2.13}$$

So, we differentiate $\Theta_u(T)$ and $\Delta H(T)$ with respect to Temperature, remembering that $\Delta H(T)$ is a function of temperature $\Delta H(T)$. Differentiating $\Delta H(T)$ with respect to temperature given our assumption that $dC_p(T)$ between folded and unfolded states is constant yields ΔC_p. Differentiating $\Theta_u(T)$ is a little more difficult because $\Theta(K(\Delta G(T)))$. Keeping in mind that the derivative of a function of the form $y = e^{ax}$

looks like this: $\frac{dy}{dx} = ae^{ax}$ we can take the first derivative of $\frac{\Delta G(T)}{RT}$ given our expression in Eq. 2.8 .

$$\frac{d}{dT}(\frac{\Delta G(T)}{RT}) = \frac{d}{dT}(\frac{\Delta H_{Tm}(\frac{1}{T} - \frac{1}{Tm}) + \Delta C_p(1 - \frac{Tm}{T} + \frac{\Delta C_p}{T}(ln(\frac{T}{Tm})))}{R}) \tag{2.14}$$

$$\frac{d}{dT}(\frac{\Delta G(T)}{RT}) = \frac{1}{\Delta H_{Tm}}T^2 + \Delta C_p(\frac{Tm}{T^2} - \frac{\Delta C_p}{T^2}(ln(\frac{T}{Tm}) + \frac{\Delta C_p}{Tm}\frac{1}{TTm} \tag{2.15}$$

$$\frac{d}{dT}\Theta_u(T) = \frac{d}{dT}(\frac{1}{1 + e^{\frac{\Delta H_{Tm}(1-\frac{T}{Tm}) + \Delta C_p(T - Tm + \Delta C_p(ln(\frac{T}{Tm}))}{RT}}}) \tag{2.16}$$

$$\frac{d}{dT}\Theta_u(T) = \frac{\frac{\frac{\Delta H_{Tm}}{T} - \frac{\Delta C_p Tm}{T} + \Delta C_p}{RT}(K(T)^2 + K(T)) - \frac{K(T)^2}{RT}}{(K(T) + 1)^2} \tag{2.17}$$

Now we can substitute this into Eq. 2.13 to get

$$C_p(T) = \Delta H(T)\frac{(\Delta C_p - \frac{\Delta C_p T_m}{T} + \frac{\Delta H_{Tm}}{T})\frac{K(T)}{RT}}{(K(T) + 1)^2} + \frac{\Delta C_p K(T)}{K(T) + 1} \tag{2.18}$$

which simplifies to

$$C_p(T) = \Delta H(T)\frac{(\Delta C_p T - \Delta C_p T_m + \Delta H_{Tm})K(T)}{RT^2(K(T) + 1)^2} + \frac{\Delta C_p K(T)}{K(T) + 1} \tag{2.19}$$

Where $\Delta H(T)$ is expressed in Eq. 2.4 Eq. 2.18 is the final equation we can use to fit the change in enthalpy of a two-state transition.

2.2.2 Three-state equilibrium derivation

Under conditions of equilibrium the occupancy of three states can be represented as follows.

$$\mathrm{N} \underset{}{\overset{K_1(T)}{\rightleftarrows}} \mathrm{I} \overset{K_2(T)}{\rightleftarrows} \mathrm{U} \tag{2.20}$$

Thus, mass balance dictates $P_{total} = N(T) + I(T) + U(T)$. Where P_{total} represents the total concentration of protein in solution, and N, I, and U represent the Native, Intermediate, and Unfolded populations. Using the two equilibrium constants, we can solve for the concentrations of all of the states in terms of N (the 'reference' state) and the equilibria constants

$$\Theta_N(T) = \frac{N(T)}{N(T) + I(T) + U(T)} = \frac{1}{Q(T)} \tag{2.21}$$

$$\Theta_I(T) = \frac{I(T)}{N(T) + I(T) + U(T)} = \frac{K_1(T)}{Q(T)} \tag{2.22}$$

$$\Theta_U(T) = \frac{U(T)}{N(T) + I(T) + U(T)} = \frac{K_1(T)K_2(T)}{Q(T)} \tag{2.23}$$

Where, $Q(T) = 1 + K_1(T) + K_1(T)K_2(T)$. The heat capacity is then given by the following differential:

$$C_p(T) = \frac{d}{dT}(\Theta_I(T)\Delta H_I(T)) + \frac{d}{dT}(\Theta_U(T)\Delta H_U(T)) \tag{2.24}$$

Plugging in values for $\Theta_I(T)$ and $\Theta_U(T)$ as well as $\Delta H_I(T)$ and $\Delta H_U(T)$ yields

$$
\begin{aligned}
C_p(T) \quad = \quad & (\Delta H_1 + (\Delta C_{p,1}(T - T_{m,1}))) * \\
& \frac{K_2(T)((T\Delta C_{p,1} - \Delta C_{p,1}T_{m,1} + \Delta H_1)K_1(T)K_2(T)}{RT^2} - \\
& \frac{\Delta C_{p,2}T + (\Delta C_{p,2}T_{m,2}) - \Delta H_2)}{RT^2} * \\
& (((K_1(T)+1)K_2(T)+1)^2) + \Delta C_{p,1}\frac{K_1(T)}{1 + K_1^{-1}(T)(1 + (K_2^{-1}(T)))} + \\
& \Delta H_2 + (\Delta C_{p,2}(T - T_{m,2})) * \\
& T(\frac{(((\Delta C_{p,2} + \Delta C_{p,1})T - \Delta C_{p,2}T_{m,2} - (\Delta C_{p,1}T_{m,1}) + \Delta H_2 + \Delta H_1)K_1}{(RT^2)(((K_1(T)+1)(K_2(T))+1))^2} + \\
& \frac{T\Delta C_{p,2} - (\Delta C_{p,2}T_{m,2}) + \Delta H_2)(K_2(T))}{(RT^2)(((K_1(T)+1)(K_2(T))+1))^2}) + \\
& (\Delta C_{p,2}(\frac{K_2(T)}{1 + K_1^{-1}(T) + K_2(T)}))
\end{aligned}
$$

Where from equations 2.2 and 2.8:

$$K_x(T) = e^{-\frac{\Delta G_x(T)}{RT}}$$

and

$$\Delta G_x(T) = \Delta H_{x,T_{x,m}}(1 - \frac{T}{T_{x,m}}) + \Delta C_{x,p}(T - T_{x,m} - T(ln(\frac{T}{T_{x,m}}))$$

Where x represents either of the two transitions. These models were adapted from previous studies.[41, 42]

2.2.3 Two independent transitions

Transitions of more complex phenomena are expressed in a very similar way, such as independent unfolding events:[41]

$$\mathrm{N}_1 \underset{}{\overset{K_1(T)}{\rightleftharpoons}} \mathrm{U}_1 \quad and \quad \mathrm{N}_2 \overset{K_2(T)}{\rightleftharpoons} \mathrm{U}_2 \tag{2.25}$$

Where K_1 and K_2 represent equilibrium transitions for two independent events. For complex folded proteins this may be the unfolding of two separate domains or a dissociation reaction followed by and unfolding. For ABD1 the two independent transitions would most likely represent the separate CH domains unfolding. The fit equation is taken from utilizing Eq. 2.19 twice in the same expression to account for the two independent processes which can be seen in Eq. 2.26.

$$C_p(T) = \Delta H_1(T)\frac{(\Delta C_{p,1}T - \Delta C_{p,1}T_{m,1} + \Delta H_{Tm,1})K_1(T)}{RT^2(K_1(T)+1)^2} + \frac{\Delta C_{p,1}K_1(T)}{K_1(T)+1} + \Delta H_2(T)\frac{(\Delta C_{p,2}T - \Delta C_{p,2}T_{m,2} + \Delta H_{Tm,2})K_2(T)}{RT^2(K_2(T)+1)^2} + \frac{\Delta C_{p,2}K_2(T)}{K_2(T)+1} \tag{2.26}$$

2.3 Current methodology

To discuss the changes introduced through this method of analysis, it is first necessary to describe more wholly the analysis that is currently used to fit nonlinear thermodynamics data. The current method of analysis is run in Excel. The other software Nanoanalyze is too rigid in its estimate of partial specific volume which accounts for the volume occupied by the species in solution. Generally, the analysis uses a model, as described in the previous sections, which is fit with a nonlinear regression algorithm to the raw data. The specifics of the current method are described below.

Initial model guesses for the nonlinear regression are required. The initial modeling estimates are created from a combination of inference from looking at the data graphed and model predictions based off of empirical models for the dependency of thermodynamic parameters on the molecular weight and amino acid sequence of the protein.[28, 29, 43, 44] Estimates vary slightly depending on the

model used.[45] The transition temperature is estimated from the experimental data itself. Because the empirical models for estimating thermodynamic parameters were generally developed from highly stable globular proteins, the initial guesses won't specifically fit experimental data so a nonlinear regression algorithm is necessary to minimize the modeling error.

In Excel the nonlinear regression algorithm is called 'Solver'. Solver allows the use of three different types of minimization algorithms GRG Nonlinear Solving Method, Simplex LP Solving Method, or Evolutionary Solving Method. All of these methods targeted the Residual Sum of Squares (RSS), which is given as the difference between experimental data and the modeled estimate squared, as an object of minimization. The current method utilized the default GRG Nonlinear Solving Method. Since these nonlinear regression problems have multiple dimensions (one for each of the fit parameters, and one for the RSS) all of these methods are prone to error through being trapped in local minima and not finding the global minimum. Each of these methods proceed in a way such that they produce a step (how the step is produced is dependent on the particular algorithm) in the parameter space defined by the variables being fit, calculates the RSS and moves in a direction that will ultimately make the RSS smaller. In order to determine which point along the multitude of steps taken during the minimization process, all parameter values, and all corresponding RSS values must be stored. With extremely large data sets, this can put an extremely high demand on the memory of the computer and oftentimes crashes computers, especially when involving slower scan rates or a wider temperature range.

The current method uses 'solver' in Excel to fit data, but is known to return different fit values depending on the initial estimates for the nonlinear regression. This can be a problem when using initial estimates based on different prediction

models. We generated a table of initial seeding guesses that varied within the deviation of the two prediction methods used to generate initial seeding guess for ΔH_{T_m} and ΔC_p (predictions differ by as much as 50% for ΔC_p and 25% for ΔH_{T_m}).[28, 29] The results of tests ran on a Synaptotagmin 1 C2B data set originally published in PLOS ONE are outlined in Tab. 2.1.[40]

	ΔH_{T_m} (kcal/mol)	ΔC_p (kcal/mol K)	T_m (K)	$\Delta G(37°C)$ (kcal/mol)	Mean Resid.[2]
Input 1	70	1	319.6	-	-
Output 1	60 ± 8	2.0 ± 0.7	320 ± 1	1.5 ± 0.7	0.500
Input 2	60	3	319.6	-	-
Output 2	59 ± 9	3.2 ± 0.6	320 ± 1	1.3 ± 0.6	0.756
Input 3	40	1	319.6	-	-
Output 3	60 ± 112	1.99 ± 1	288 ± 1500	-6.26 ± 9	4.925

Tab. 2.1: Results from regressions with different seeding values on Synaptotagmin C2B data originally reported by Fealey et al.[40] The data was interpretted from a graph in the original publication.[40] The fit values as well as reported modeling errors from regression are reported for each parameter. The last column reports the mean squared residual from the regression output calculated by dividing the RSS by the number of data points in the scan.

The first set of inputs listed in Tab. 2.1 agree most closely with the published best fit results ($\Delta H_{T_m} = 69.6 \pm 0.6$ kcal/mol, $T_m = 319.6 \pm 0.1$K).[40] The second set of inputs utilized the fit enthalpy from the first regression (Output 1) as the seeding guess, and varied the change in heat capacity by 50% from the fit value of Output 1. The third set of guesses seeded a lower estimate for the transition enthalpy. All three regressions returned similar mean values for the transition enthalpy, though the associated standard deviation varied. The change in heat capacity varied by roughly 50% and the transition temperature for the first two regressions were consistent, but not for the third regression. The model fits are graphed in Fig. 2.4. The differences in the reported fit parameters are likely then due to the algorithm finding a local minimum. Since the mean squared residuals is smallest for the first reported

regression in Tab. 2.1, it is more likely that they correspond to the global minimum. This highlights the fact that analyzing DSC data is highly dependent on seeded values.

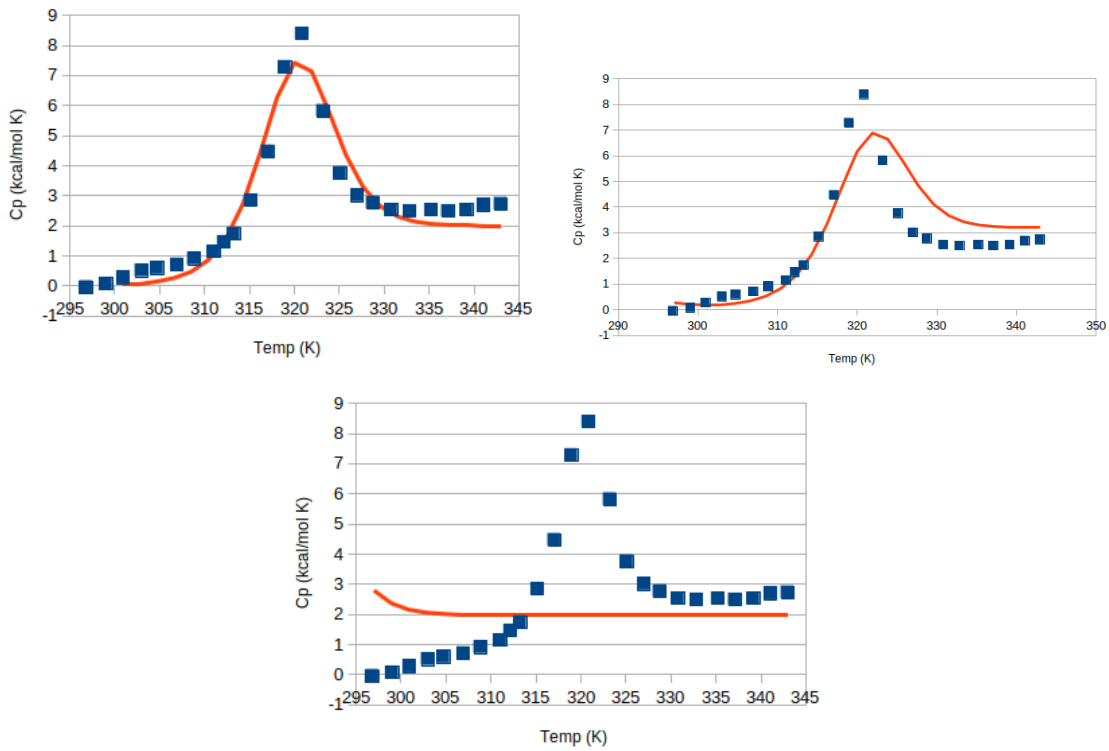


Fig. 2.4: Example fits from the reported regression solutions reported in Tab. 2.1. The raw data is represented as discrete blue squares and the model is represented by the continuous orange line. Starting with the top left, the outputs from Tab. 2.1 are graphed from left to right in order.

The amount of memory needed for a nonlinear regression with upwards of 10,000 data points causes many computers to crash, freeze, or otherwise abort the analysis. Coupled with the fact that the current methodology can take roughly a day to get a model fit, shows the great demand this method places on people hours. Fitting more complex models places even higher demands by doubling the number of fit parameters (at a minimum) increasing the rate of computers crashing. The new methodology addresses each of these concerns through doing all calculations in

a new language, R. The specifics of how each of these are addressed will be covered in the following sections.

2.4 Rationale for new methodology

The new methodology was designed to address the difficulties noted in the current methods of fitting experimental data. Namely, the challenges that are currently faced concern the reproducibility of analysis, the difficulty in discerning between a two-state denaturation vs. a more complex model, and the demand on both computer and people resources. All of these concerns and how they are addressed within the new methodology are elaborated upon below.

2.5 Efficiency

The first major benefit of utilizing the analysis developed in R was the increased efficiency in multiple ways. The amount of time spent defining a baseline, running a solver fit algorithm and getting publishable figures takes on the order of 10's of hours for the analysis of a single data set. Given that a single study requires three to ten scans to be analyzed the amount of people hours required to analyze data can take a week or longer to finish. The new methodology added new efficiency to the analysis in two ways: it decreased the computer time needed for nonlinear regression and linear regression (for establishing the baselines) as well as decreased the instances of computers crashing and it improved efficiency by automating more of the work-flow.

A decrease in the computer time was noted when running the analysis in R rather than Excel. Excel took approximately three seconds for the nonlinear regression algorithm to run on small data sets, approximately six seconds on larger data sets with greater than 10,000 rows and five varying columns. On computers

with less processing power the run time increased to approximately six seconds and for larger data sets Excel crashed and froze a majority of cases leading to a difficulty in estimating how much time the regression took. Running a nonlinear regression on the same data sets in R returned almost instantaneous results with no noticeable difference in computer time between the smaller and larger data sets. This means that R is also scalable to fit larger data sets which is necessary when fitting multiple data sets simultaneously.

Additional contributions to efficiency come from automating more portions of the work flow. Finding baselines through alternating the baseline model definition and a solver regression in Excel can take roughly six hours to find a working baseline for a single data set. Baseline definitions can not be automated in Excel because it introduces too many fit parameters and crashes the spreadsheet and even still often crashes leaving the user to restart the analysis. It has also been noted to lead to unreliable results.[28] In the R program developed, the user has two possibilities in varying the baseline, the first is entering the lower temperature for the baseline domain definition into the Rscript, the second is directly varying the temperature bounds for the lower baseline domain directly on a graph through the model fitting GUI developed in Shiny within the R environment. An image of the GUI interface for fitting a baseline to the data can be seen in Fig. 2.5.

After every new baseline definition in R, the nonlinear regression is recomputed and gives the user live feedback of the model fit both with the graph and fit values returned. R solves the regression and returns the graph and fit values instantly. This can speed up analysis roughly by four hours per data set. The graphs returned are of publication quality which also greatly increases efficiency. Through automating the regression and assisting with the baseline definition, the process can be ran on an arbitrary number of data sets simultaneously which reduces the amount of people

app.png

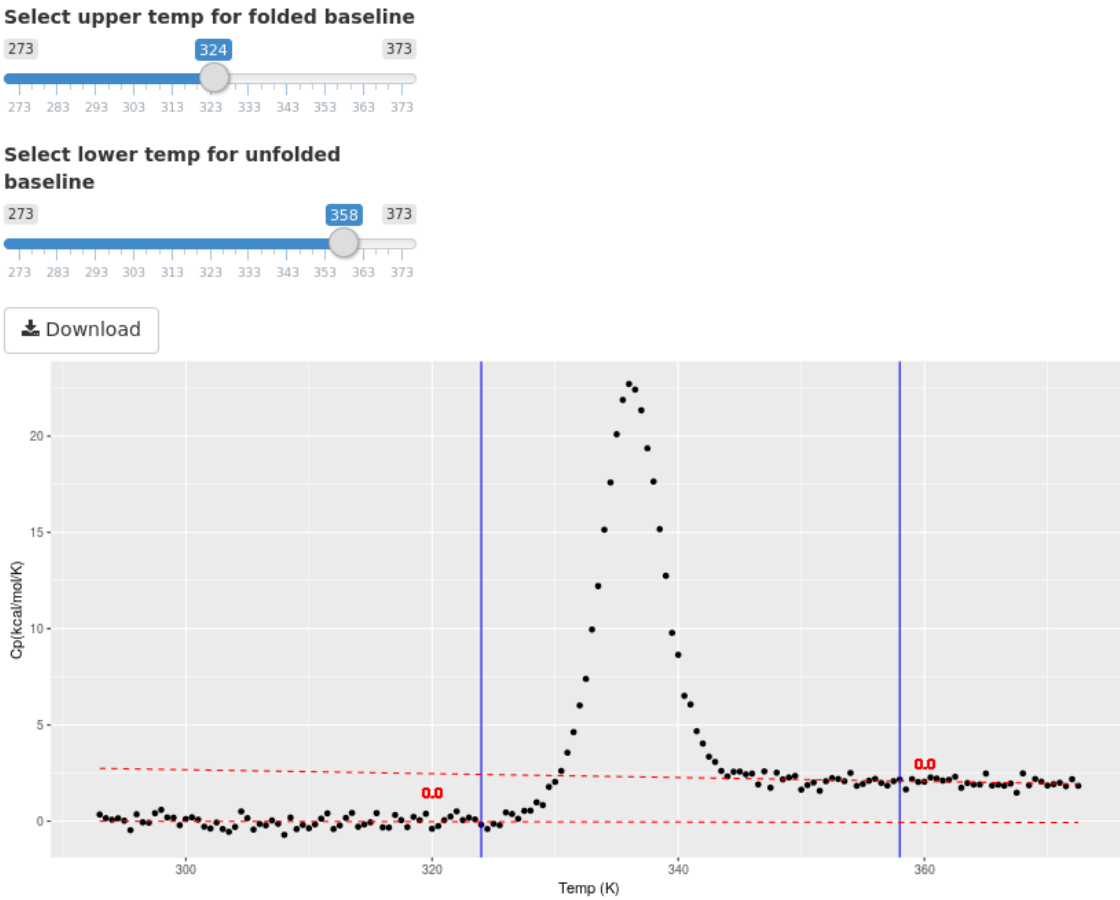


Fig. 2.5: An image taken from the interactive user interface designed in R to aid in DSC analysis. Pictured here is the baseline assist where the user drags the definition of their baseline which updates the graph in live time with a projection of the baseline as well as the slope.

hours necessary for analysis by approximately $4 * n$ hours where n is the number of data sets being fit without accounting for the number of crashes that might be faced in Excel. Up to four data sets have been fit simultaneously without any noticeable loss in run time.

2.6 Other benefits of automated work flow

The automated work flow introduced in the R program also has the benefit of making the analysis more consistent. The analysis developed in R creates intelligent estimates of fit parameters utilizing the raw data in a novel way utilizing methodologies proposed by Freire and Karplus which take into account the geometry of the data while also regressing a vector list of starting values.[28, 30] Here I will discuss putting the theory into action in the analysis which includes developing estimates for the fit parameters (ΔH_{T_m}, ΔC_p, and T_m), creating a sigmoidal baseline estimate for the basal shift in heat capacity, and generating a model estimate off of the estimated values. The general outline for generating estimates is as follows: 1) Determine the absolute shift in basal heat capacity for an estimate for ΔC_p, 2) determine the complete 'change in enthalpy' between the pre- and post-transition states in order to normalize the extent of reaction (denaturation), 3) create an estimate for the sigmoid that describes the continuous shift in basal heat capacity, 4) Utilize the defined sigmoid to create the lower threshold for the integration for estimating the transition enthalpy(ΔH_{T_m}), 5) estimate the change in transition temperature (T_m) by finding the temperature at which exactly half of the transition enthalpy is reached. The order of the approximations are done precisely in this order as they build off of one another. The method of developing estimations are outlined in the next couple of sections.

2.6.1 Change in heat capacity estimate

After the experimental data has had the folded baseline set to zero by creating a line model off of the first 25 values and subtracting the line model from the raw data, the data appears to be flat before and after the transition. The change in heat capacity (ΔC_p) is described by the baseline shift between the folded and unfolded states. The pre- and post-transitional baselines are given by line models, the differences between them is constant.[28] Therefore, we generated an estimate for the change in heat capacity by finding the difference between the first and last portions of the experimental curve. We averaged the estimated heat capacity over 25 points (to reduce the susceptibility to noise) at the highest and lowest temperatures present in the data and took the difference to be the change in heat capacity.

2.6.2 Creating a baseline

The most difficult part in estimating the thermodynamic parameters from the experimental data comes from defining the baseline during the transition portion of the curve.[30] One method of estimating the curve utilizes the definition of the model as being two-state, to estimate the baseline as being a step function. The step function would have the lower step be defined by the folded baseline and the upper step defined by the unfolded baseline. They are joined together by a non-continuous step function at the halfway point of the transition.[30] The problem with this approach is that at the joining of the two baselines, it assumes the transition is all or nothing. Statistical thermodynamics suggests that the two states have a difference in free energy where the difference in free energy relates to the distribution of molecules existing in the two states. As energy is added to the system it shifts the balance of the equilibrium (distribution of states, see Eq. 2.2) which is a continuous

process and is described by a continuous function the Gibbs-Helmholtz equation and (see Eq. 2.8).

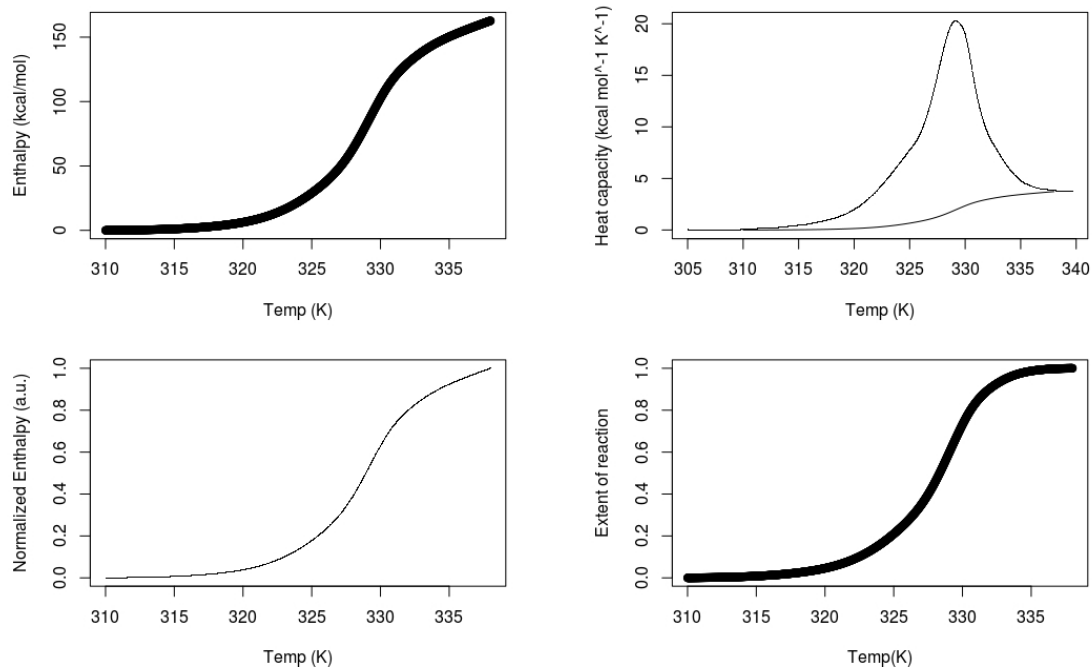


Fig. 2.6: Top-left: The integrated heat capacity throughout the entirety of the transition. Bottom-left: The integrated function, normalized, before readjusting the max to be equivalent to the estimated change in heat capacity. Top-right: Experimental data with the estimated baseline graphed together. Bottom-right: The normalized integral of the experimental data after subtraction of the sigmoidal baseline.

Assuming the equilibrium shifts as a function of temperature, naturally produces a sigmoidal baseline with the steepest portion of the sigmoid occurring at the transition temperature. Furthermore, the steepness of the sigmoid is directly proportional to the corresponding heat capacity at a given temperature meaning that the sigmoid is the integrated expression of the heat capacity. The challenge at this point however, can be seen in the top left panel of Fig. 2.6. The top and bottom portions of the sigmoid are expected to be flat given the two-state assumption, however there is a noticeable slope at the end of the sigmoid at high temperatures. Given, that the sigmoid is the integrated expression of the heat capacity measurements, and

conversely that the heat capacity is the differential of the total energy of the system, it follows that the constant slope seen at high temperatures is due to the baseline shift in heat capacity (ΔC_p) which leads to its positive, but constant slope. Since energy is conserved we can then account for this with the following expression:

$$\Delta H_{tot}(T) = \Delta H_{base}(T) + \Delta H_{excessenthalpy}(T). \tag{2.27}$$

We directly solved ΔH_{tot} as described above by integrating the entire area under the curve between two user defined points where the transition is supposed to begin and end. We constructed $\Delta C_{p,basal(T)}$ supposing it takes a value from zero to ΔC_p and varies as a function of the total area under the transition curve ($\Delta H_{tot}(T)$). To do this, we first normalized the $\Delta H_{tot}(T)$ curve between zero and one (lower left panel Fig. 2.6) and then multiplied by ΔC_p. The final baseline definition is shown in the upper right panel of Fig. 2.6. This allows for the direct subtraction of the baseline from the experimental data. The excess enthalpy can then be taken as the area between the curves and the extent of reaction represented as the enthalpy normalized between zero and one as pictured in the lower right panel of Fig. 2.6. The resulting sigmoid has a flat post-transitional slope at high temperatures. To aid in properly establishing a post-transitional baseline it is recommended to push the denaturation to higher temperatures for all scans once a decent reversibility scan is established in future experiments. Baseline projections are included along with the associated slope value of the baseline to aid in baseline definition. This process is shown in Fig. 2.7. The ability of this method of baseline subtraction to retain the overall shape of transition is discussed in later sections.

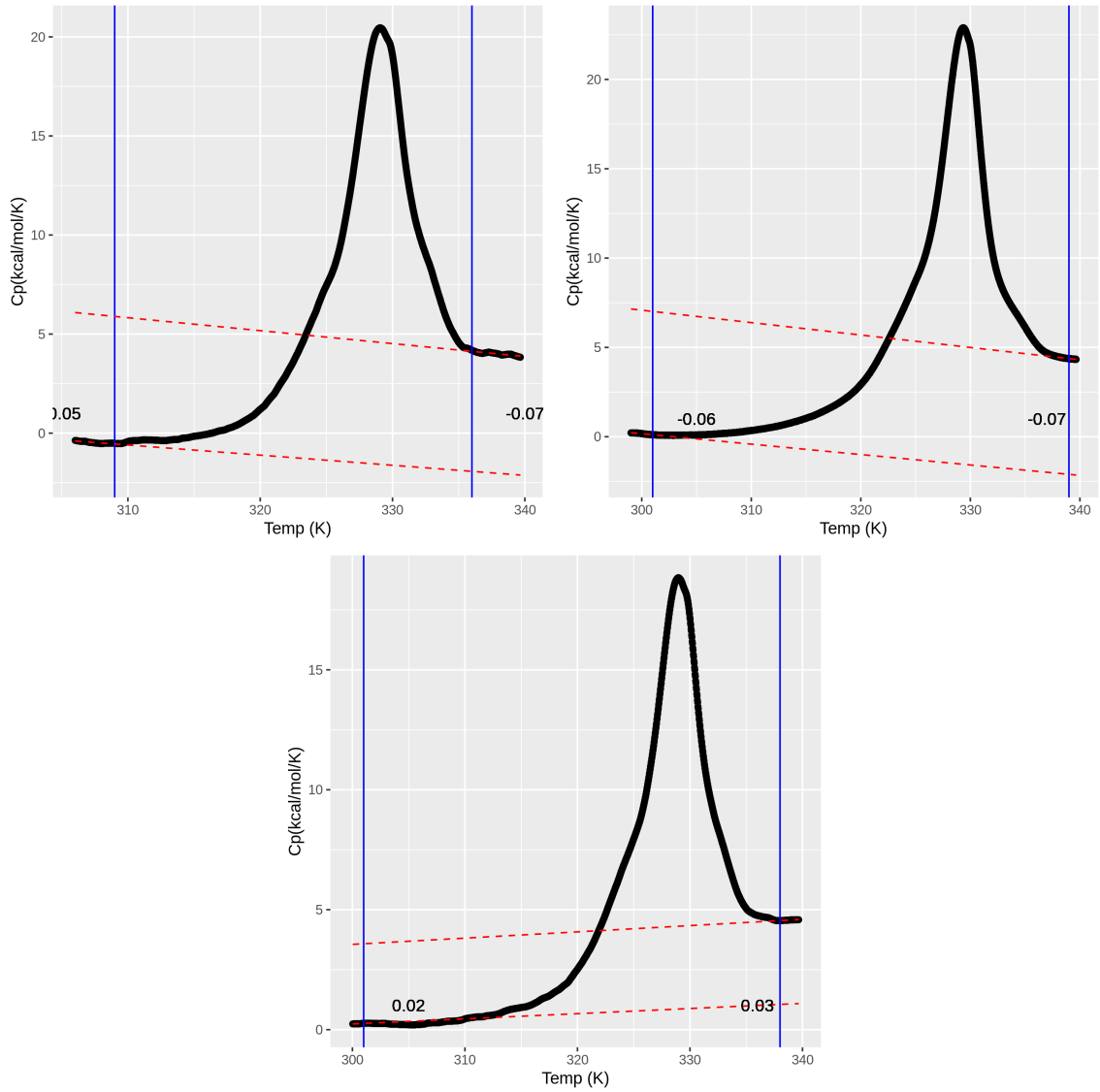


Fig. 2.7: Raw data of the three ABD1 scans represented after an initial baseline estimate. The final baseline estimate is projected by the red dotted lines with the slopes printed above the given lines. The blue vertical lines display the innermost bounds used in the definition of the baselines.

2.6.3 Excess enthalpy and transition temperature estimates

With the baseline established, estimating the excess enthalpy of transition is relatively straight forward to consider as it is represented as the area between the experimental curve and the sigmoid baseline.[28] The integral can be evaluated by a rectangular Riemann summation starting at the user defined pre-transition point and ending at the user defined post-transition point, for the upper and lower bounds of integration, respectively (Maroon area in Fig. 2.8).

The transition temperature of a two-state process is represented by the temperature at which there is 50:50 population split between states. The transition temperature was estimated by halving the estimated excess enthalpy (Maroon area in Fig. 2.8) and finding the corresponding temperature.[28] Thus, we've found a way to establish estimates for the three thermodynamic parameters needed to describe a two-state transition ΔH_{T_m}, ΔC_p, and T_m and provide the first initial model fit guesses to seed the nonlinear regression.

2.7 The search for a global minimum

With the increased efficiency in run time it became possible to run many nonlinear regressions on the same set of data to have a higher probability of finding a global minimum. This is because any given minimum is not necessarily a global minimum, but might instead be a local minima (a local minimum is defined as the smallest value within a bounded region, a global minimum is defined as the smallest value within an unbounded region). There is no method that can determine whether it is indeed a global minimum. We utilized a "brute-force" methodology in which an entire vector list of values are minimized. This method is essentially running

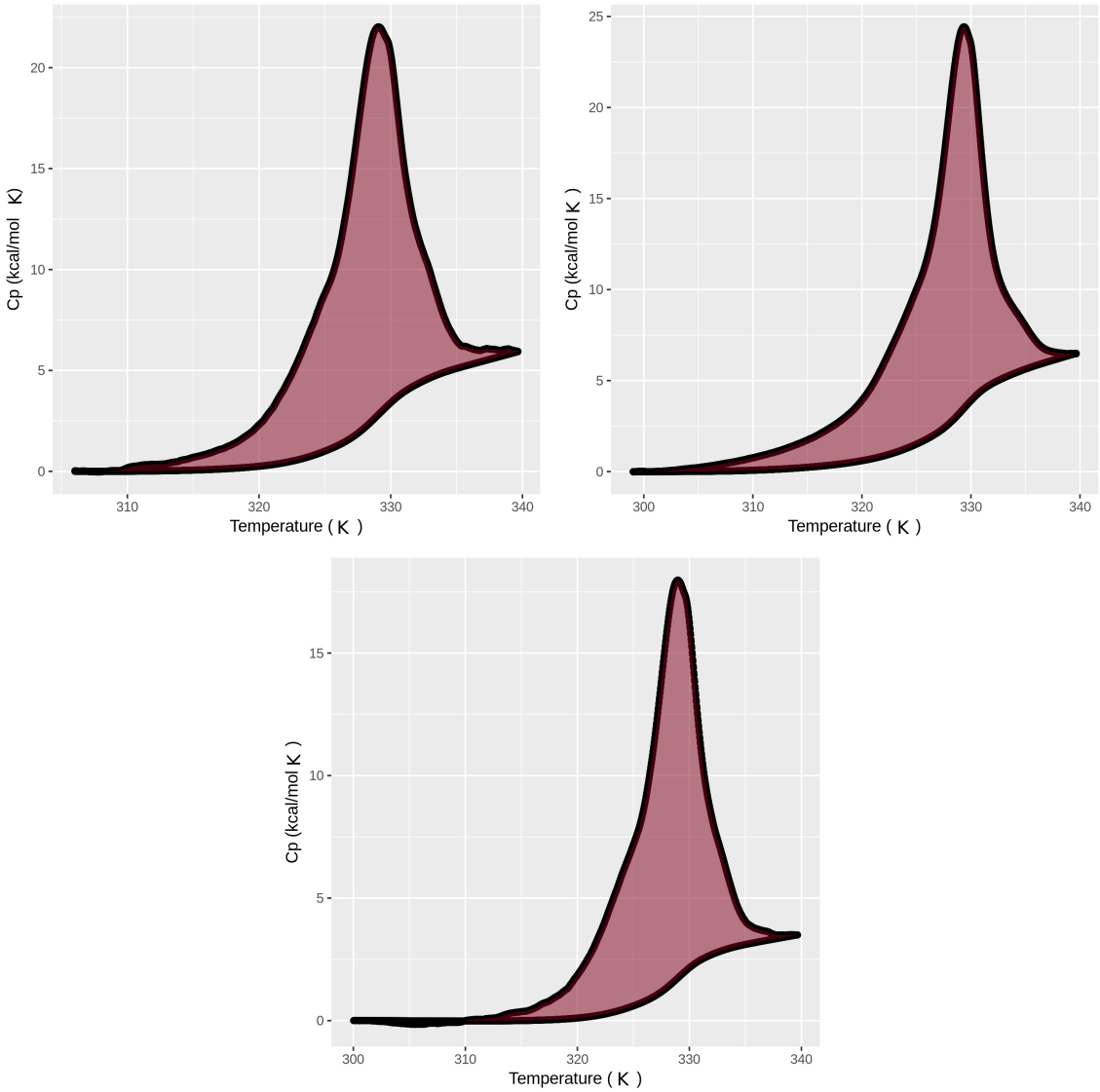


Fig. 2.8: Geometric representation of the change in enthalpy (maroon) and the calculated sigmoidal baseline (lower bound) for all three scans. The computed baseline can also be visualized as the lower bound for the maroon region.

the nonlinear regression from multiple starting points given by the vector list of start values. This approach becomes increasingly necessary as the model to be fit increases in complexity. In fact, without it it can be hard to find a minimum at all (a model does not converge) for instance in fitting a three-state model.

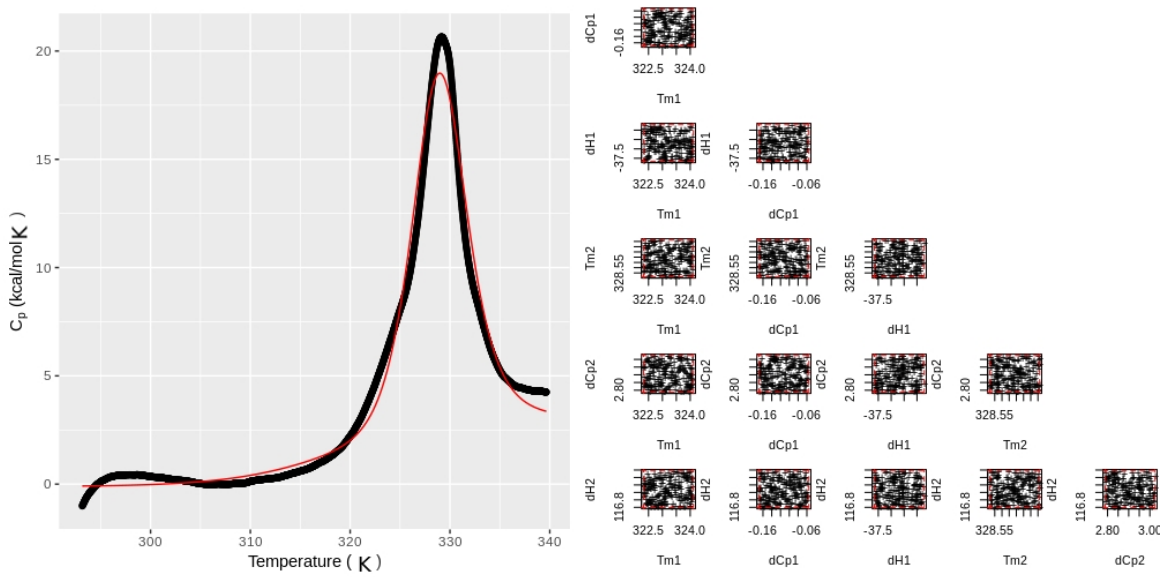


Fig. 2.9: An example of experimental ABD1 data (black) fit with a two independent transitions model (red). Note the qualitatively good fit of the raw data, but in evaluating the confidence regions of the fit, it is clear that one transition has both a negative enthalpy and change in heat capacity indicative of overfitting the data.

To probe the parameters space thoroughly we created an evenly spaced hypercube of n-dimensions (n being the number of parameters to be fit) that was centered on values given from the geometric estimation of fit parameters.[46] Using the estimates, the user must only define the breadth of the search along each dimension. Since the search will grow exponentially with the number of parameters, it is important for the user to have the option to define how many points will be probed along the established dimensions between the maximum values. The regression used in the brute-force algorithm is the Levenberg-Marquardt. Calling this value p and the number of dimensions n, the number of regressions to be run is p^n. With

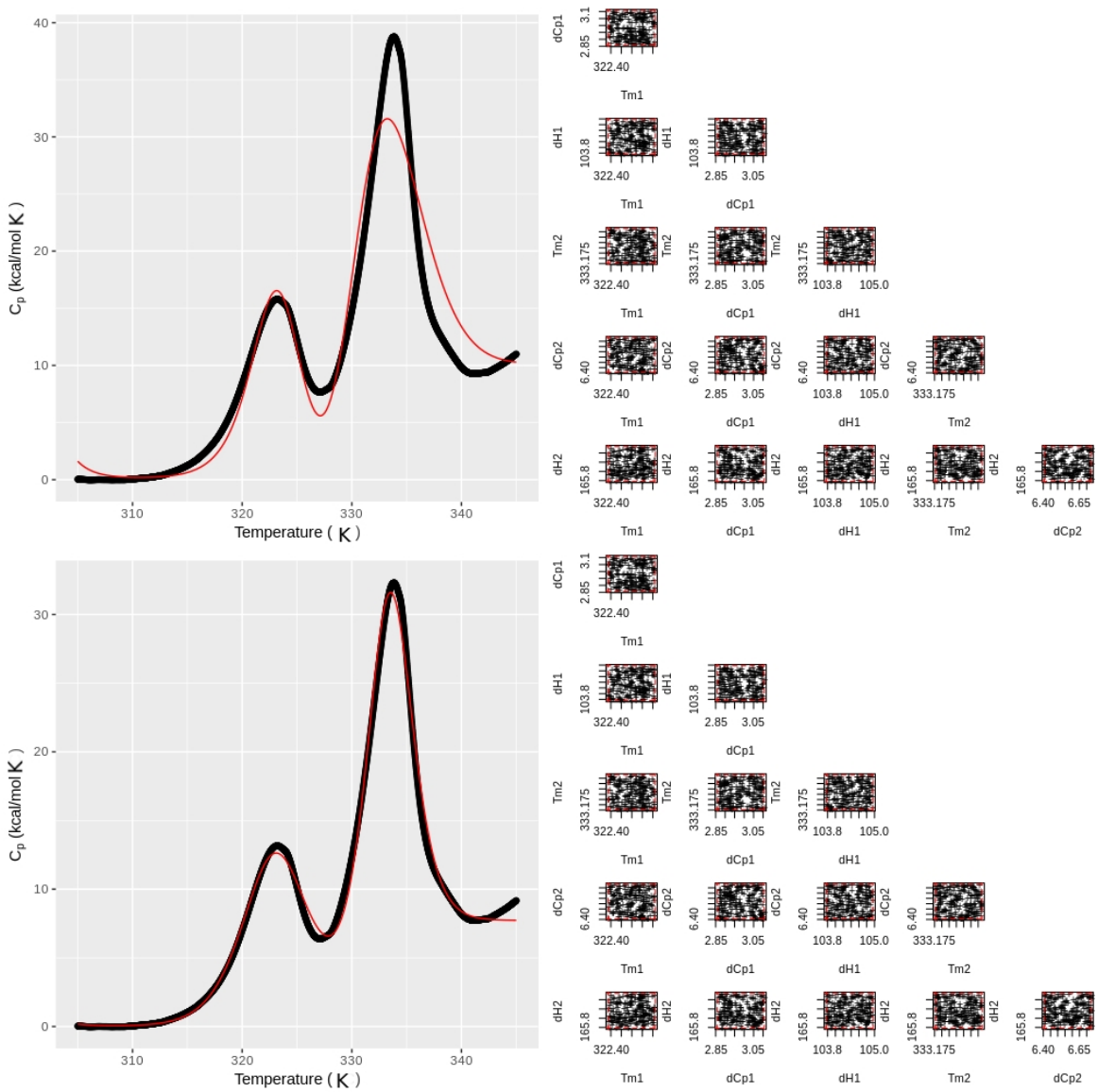


Fig. 2.10: Experimental data representing a melt of an MBP-ABD1 construct (black) fit with both a sequential model and an independent transitions model. Both have reported confidence regions as well. The independent transitions model appears to fit better.

this approach it is recommended to calculate the number of regressions p^n before running the regression to assure a reasonable number of regressions is to be ran, as the number of regressions increases exponentially. (i.e. for a three dimensional search, probing ten points along each dimension results in 1,000 searches, four points in 10,000, five points in 100,000, etc.). Generally for 700 regressions it takes about five minutes to run. While the hypercube approach assure a global minimum is found, it is a step in the right direction especially when coupled with the knowledge about minima shape that is acquired from confidence analysis that will be discussed in the next section. This methodology was successfully applied to complex models and can be seen in Fig. 2.9 and Fig. 2.10.

2.7.1 Assessing the minimum

Once a minima was identified by the regression, we analyzed its characteristics. To assess the fit confidence for the nonlinear regression we utilized a tool originally developed by Baty et al.[47] The package, developed in R, generates two sets of results to estimate the confidence regions centered at the identified minimum for the regression. The first method considers how the modeling error changes as a function of the parameter estimates. The parameters are varied pairwise leading to 2D cross sections of the modeling error as a function of the model definition. Visually this is represented as a heat map in Fig. 2.11. Regions of relatively low modeling error are represented in green and regions of high modeling error in red. The coloring is based on a logarithmic scale. This method also allows for an estimate of the 95% confidence interval utilizing a criterion initially proposed by Beale that is indicated as a red-dotted circle on the graph in Fig. 2.11.[38, 39] The second method fills in the gaps left by the pairwise likelihood estimates by taking a random sampling of points, centered at the identified minimum, determining the modeling error and only

graphing the ones that satisfy the 95% confidence criteria established by Beale.[38, 39] The graphs produced from these methods can be seen in Fig. 2.11.

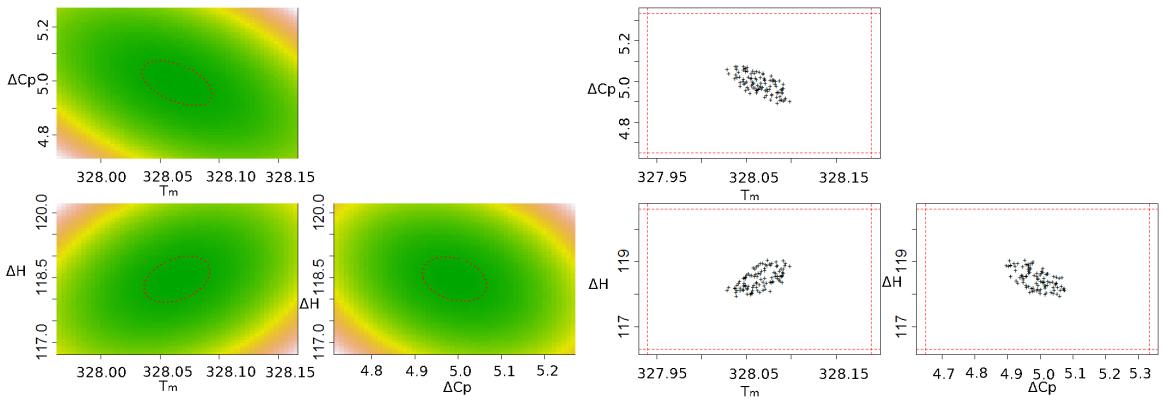


Fig. 2.11: Left: Pairwise determination of the log error of the model fit as a heat map with the 95% confidence denoted with a red circle. Right: Confidence estimate using a random sampling of the nearby parameter space, where only results satisfying the 95% confidence criteria are printed on the graph as an 'x'.

The log-likelihood cross sections are plotted on the left hand side of Fig. 2.11. The confidence regions appear to be elliptical in shape and have well bounded regions of 95% confidence (circled in red on the graphs). The set of graphs of the right of Fig. 2.11 represents the random sampling of points that satisfied the 95% confidence criterion which also suggest a well defined confidence region that happens to overlap with the log-likelihood estimates. Together, this set of graphs depicts a well bounded confidence region. This mode of confidence analysis was coupled with the brute-force nonlinear regression algorithm to provide a rigorous way to get consistent fitting results. The combination of the brute-force approach with the assessment of the confidence region allows for a detailed determination of the best fit. The brute-force algorithm addresses the issue of local minima, at least within the user-defined grid spacing (with a semi-exhaustive search), and the log-likelihood analysis coupled with the confidence region analysis describe the shape and dimensions of the confidence region leading to more consistent results

between different users as well as more reliable confidence estimates to accompany the model fit parameters.[28] The combination of these tools for analysis was applied to multiple sets of data to determine if it was generally applicable (see Tab. 2.2).

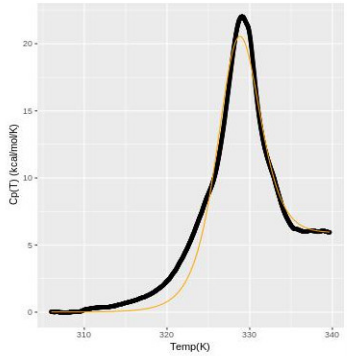
2.8 Regression of a boltzmann distribution to ABD1 data

Regression analysis was conducted on ABD1 data using a Boltzmann distribution on the separate DSC scans and yielded the results reported in Fig. 2.12 using Eq. 2.18. The individual fits can be seen in the top portion of Fig. 2.12 along with their reported values.

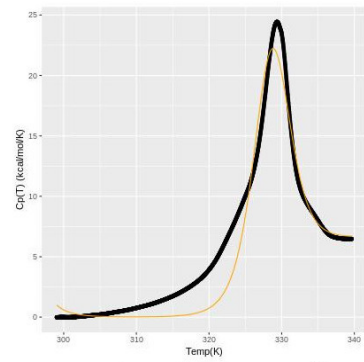
Fitting with the Boltzmann distribution led to the results outlined in Fig. 2.12. The transition has well defined thermodynamic parameters: $\Delta H_{T_m} = 118 \pm 5$ kcal/mol, $\Delta C_p = 5 \pm 2$ kcal/mol K, and $T_m = 328.0 \pm 0.2$ Kelvin. The standard deviations of these fit values are comparable to those reported in other studies.[40, 42] However, the models can be seen to have a systematic deviation from the experimental data where the low end of the experimental transition does not agree with the proposed two-state model. To assess this deviation we set out to determine if the deviation could be explained by the introduction of more complex models which account for more states. Applying more complex model quickly makes the fit equation far more complicated as can be seen in Eq. 2.25 making it difficult to grasp intuitively as well as by introducing new thermodynamic fit parameters making the problem of finding a global minimum. The next couple of sections outline how we applied distribution functions of increasing complexity to keep model fits intuitively simple. We evaluate an approach using Gaussian curves, an approach using Hubbert functions, and finally an approach that doesn't use any predetermined models.

Data	Value	Geometric	NLS (Two-State)	NLS (Three-State)	NLS (Two-ind.)
Two-State	ΔH_{T_m} (kcal/mol)	140.5	139.6 ± 0.4	129 ± 1 24.4 ± 0.2	DNC
	ΔC_p (kcal/mol)	1.81	2.03 ± 0.03	4.51 ± 0.01 0.12 ± 0.03	DNC
	T_m (K)	336.5	336.00 ± 0.02	338.92 ± 0.04 320.61 ± 0.22	DNC
	$\Delta G(37°C)$ (kcal/mol)	9.13	8.67 ± 0.03	5.28 ± 0.01 0.77 ± 0.03	DNC
Three-State	ΔH_{T_m} (kcal/mol)	136	DNC	100 ± 1 120.1 ± 0.1	DNC
	ΔC_p (kcal/mol)	1.9	DNC	1.3 ± 0.4 1.493 ± 0.008	DNC
	T_m (K)	336.5	DNC	325.02 ± 0.07 327.1 ± 0.1	DNC
	$\Delta G(37°C)$ (kcal/mol)	8.66	DNC	4.12 ± 0.4 5.56 ± 0.01	DNC
MBP	ΔH_{T_m} (kcal/mol)	139	128.6 ± 0.1	DNC	DNC
	ΔC_p (kcal/mol)	0.964	-2.0 ± 0.4	DNC	DNC
	T_m (K)	334.93	339.59 ± 0.02	DNC	DNC
	$\Delta G(37°C)$ (kcal/mol)	7.33	13.78 ± 0.4	DNC	DNC
Karplus	ΔH_{T_m} (kcal/mol)	193.8	188.9 ± 0.8	DNC	189 ± 1 60 ± 10
	ΔC_p (kcal/mol)	2.7	2.0 ± 0.9	DNC	-3 ± 2 2 ± 1
	T_m (K)	323.3	323.34 ± 0.04	DNC	323.52 326.5 ± 0.9
	$\Delta G(37°C)$ (kcal/mol)	7.44	7.2 ± 0.9	DNC	8.65 ± 2 2.17 ± 1
ABD1	ΔH_{T_m} (kcal/mol)	135 ± 4	118 ± 5	DNC	-36.67 ± 0.5 117.3 ± 0.2
	ΔC_p (kcal/mol)	5.1 ± 0.5	5 ± 2	DNC	-0.11 ± 0.03 2.91 ± 0.06
	T_m (K)	328.1 ± 0.1	328.0 ± 0	DNC	323.2 ± 0.5 328.58 ± 0.02
	$\Delta G(37°C)$ (kcal/mol)	4.83	4 ± 2	DNC	-1.45 ± 0.03 5.05 ± 0.06
MBP-ABD1	ΔH_{T_m} (kcal/mol)	259	NA	127 ± 1 66.2 ± 0.6	104.4 ± 0.3 166.2 ± 0.2
	ΔC_p (kcal/mol)	NA	NA	5.8 ± 0.4 10.1 ± 0.2	2.98 ± 0.06 6.57 ± 0.08
	T_m (K)	NA	NA	323.31 ± 0.05 326.79 ± 0.03	322.4 ± 0.02 333.192 ± 0.008
	$\Delta G(37°C)$ (kcal/mol)	NA	NA	3.6 ± 0.4 -1.0 ± 0.2	3.26 ± 0.06 6.13 ± 0.08

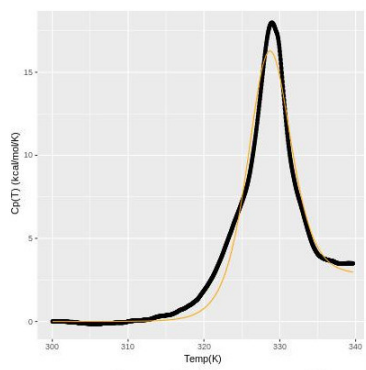
Tab. 2.2: Evaluations of different sets of data encompassing both experimental and theoretical data. One scan was analyzed for each set of reported values above to remain consistent across all rows as some of the experiments only have one set of data. The two-state model was generated with the following parameters ΔH_{T_m}=140 kcal/mol, ΔC_p = 2 kcal/mol, T_m 336. The three- state model was generated with parameters ΔH_{T_m} = 100, 120 (kcal/mol); ΔC_p = 1, 1.5; T_m = 325, 327. Only one set of scans was obtained for MBP and MBP+ABD1 experiments where MBP+ABD1 represents the uncleaved protein construct containing both MBP and ABD1. Models that did not converge due to singularities are represented as DNC. Models utilizing more than one transition list the first set of values as the top value in each cell, and the second transition as the lower set of values in each cell.
.



	Geom	Fit	Model.Error	Vant.Hoff
ΔH_{T_m}	140.62	118.93	0.70	111.1
ΔC_p	5.92	5.92	0.10	NA
T_m	328.28	327.97	0.04	NA



	Geom	Fit	Model.Error	Vant.Hoff
ΔH_{T_m}	164.92	122.83	0.94	128.8
ΔC_p	6.49	6.67	0.11	NA
T_m	327.99	327.95	0.05	NA



	Geom	Fit	Model.Error	Vant.Hoff
ΔH_{T_m}	113.44	111.74	0.62	90.2
ΔC_p	3.49	2.86	0.10	NA
T_m	328.22	328.20	0.04	NA

	ΔH_{T_m} (kcal/mol)	ΔC_p (kcal/mol K)	T_m (K)	ΔH_{VH} $(kcal/mol)$	$\Delta G(37°C)$ $(kcal/mol)$
Ave	118	5	328.0	110	4
95% conf	12	5	0.5	119	5

Fig. 2.12: Nonlinear regression fits of a two-state Boltzmann distribution regression model (yellow) on each individual scan (black) with the reported fit parameters in the table below. The average and 95% confidence interval of the fit parameters are reported in the table along with the Van't Hoff enthalpy (ΔH_{VH}) and free energy at physiological temperatures ($\Delta G(37°C)$).

2.9 Gaussian distributions

2.9.1 Fitting two peaks

We fit a model of two Gaussian distributions to generated data with no baseline shift (i.e. a ΔC_p of 0) and multiple transitions to determine if DSC transitions can be modeled by Gaussian distributions (See Fig. 2.13, shaded areas represent Gaussian distribution fit lines). The data was produced using two transitions of 100 and 120 kcal/mol, but with varied differences in transition temperature to probe the accuracy as the transitions overlap more. When peaks were completely separated it was noted that the Gaussian distribution consistently underestimated the theoretically generated enthalpy.[**Giancolaa1997DSCConcentration**] The difference is clearest in the top-left graph Fig. 2.13 where separation between theoretical data (black line) has a small deviation from the Gaussian model fit (shaded region) specifically at the extremities of the distributions. The equation for a Gaussian distribution that the transitions were fit to is given by the following equation

$$C_p(T) = A * e^{-\frac{(T-T_m)^2}{2B^2}} \tag{2.28}$$

Where A represents the height of the transition and B represents the width of the transition.

For transitions differing by 6K or more, a thermodynamic transition can be modeled to within 96% accuracy. The accuracy drops off markedly as the transition temperatures become closer as is shown in the top right of Fig. 2.13. The enthalpies predicted by the Gaussian modeling analysis were calculated using rectangular Riemann sums over the entire temperature scale and would introduce only a small

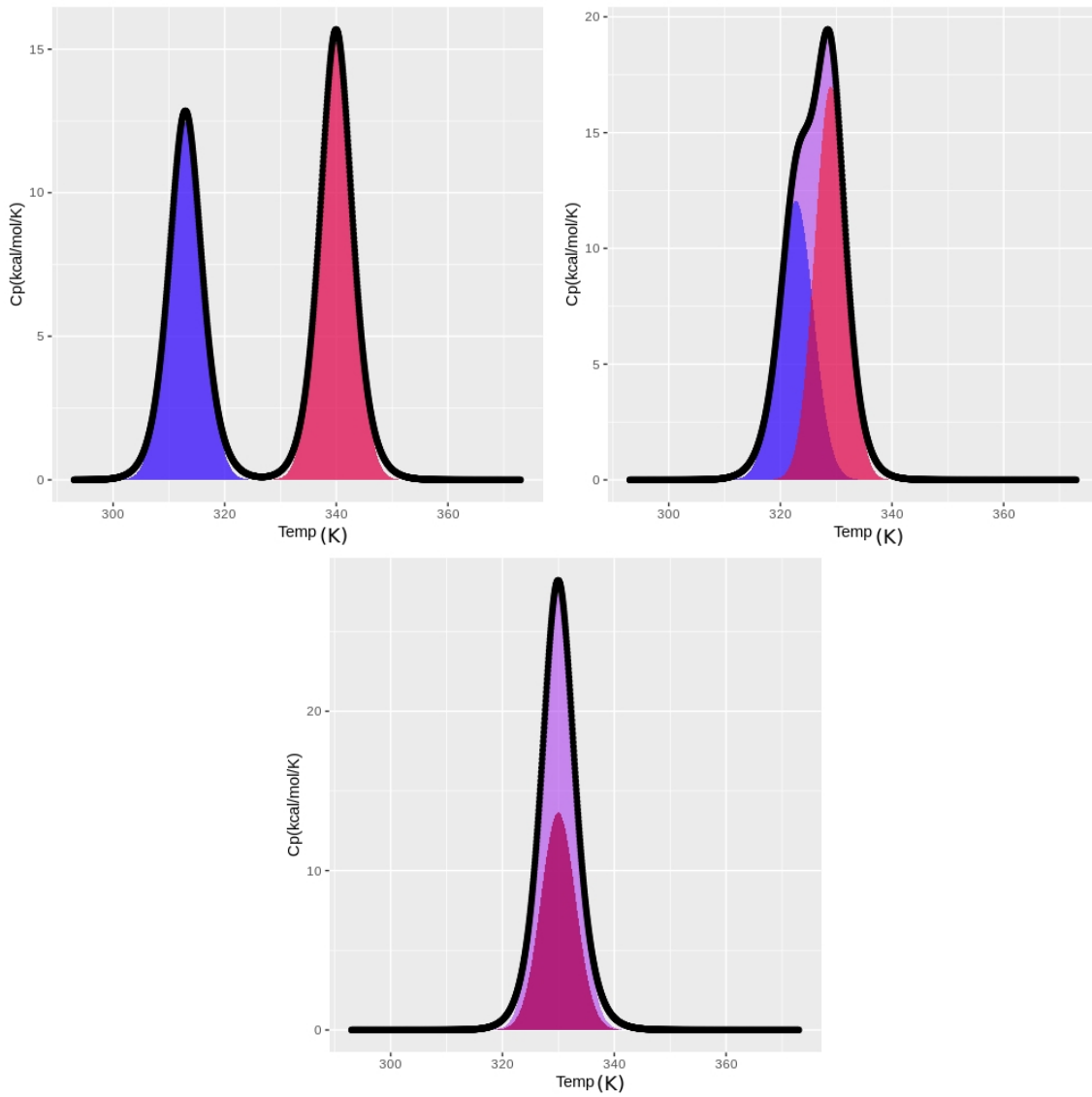


Fig. 2.13: The first three panels represent DSC data that were simulated using two transitions of 100 and 120 kcal/mol, respectively, with varying differences in transition temperatures and no baseline change in heat capacity. The simulated DSC data were then fit to normal distributions through regressing a Gaussian distribution to the data. Red and blue shading represent the separate Gaussian distributions and purple shading represents the sum of the two. Top-left: Transition temperatures that were sufficiently separate resulted in the proper fitting of two distinct transitions with the fit enthalpies being 96% accurate. Top-right: Two overlapping transitions with the fit enthalpy 96% accurate. Bottom: Two transitions of identical transition temperature, where the fit enthalpy was reportedly equal leading to 80% accuracy.

amount of error given that each curve represents 300 points making the regions effectively infinitesimally narrow.

2.9.2 Fitting baseline subtracted data

Next, we analyzed theoretically generated data with a nonzero change in baseline heat capacity to make sure that the sigmoidal baseline that will be subtracted from transitions that have a nonzero change in heat capacity retains the shape of the transition. The first generated experiments had only one transition (see top of Fig. 2.14), the others had two(see bottom of Fig. 2.14). Testing varying numbers of transitions probes the ability of the baseline construction method we created to be generalized to more than one equilibrium process. The generated data can be seen on the left of Fig. 2.14 and the right displays data with the sigmoidal baseline subtracted off as well as fit to Gaussian distributions. Overall, the baseline subtraction procedure retained the same shape as they returned an enthalpy that only differed by 5% from the value used to generate it, which is accounted for by the systematic deviation between a Gaussian distribution and a transition process that was noted in the previous section. Passing this set of positive controls, the method of fitting multiple Gaussian distributions was applied to the experimental ABD1 data.

2.10 Gaussian modeling of ABD1 data

The graphics from this analysis can be seen in Fig. 2.8. Specifically, the sigmoidal baseline depicted as the lower bound to the enthalpy integral is the baseline used for subtraction to result in Gaussian distribution. The data after baseline subtraction were fit to one and two Gaussian models, (see. Fig. 2.15). The raw data appears to be more closely modeled by a two Gaussian model than by a single Gaussian by interpreting the visual fit of the graphs. There appears to be two transitions one

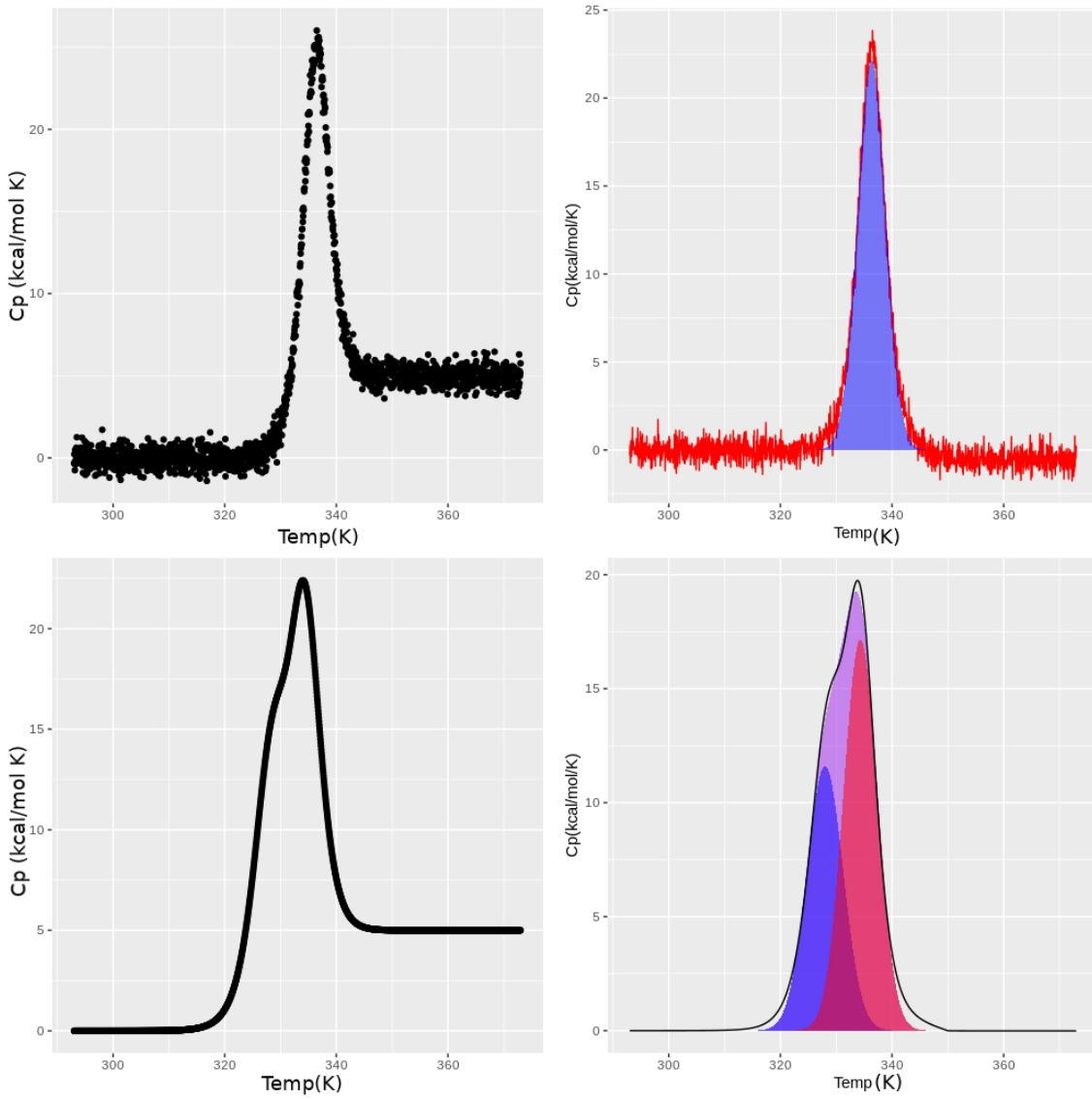


Fig. 2.14: Top-left: Simulated DSC data with from the following parameters for a single transition ΔH = 140 kcal/mol, ΔC_p = 5 kcal/mol K, T_m =336 K. Top-right: Data from the graph to the left treated with the baseline subtraction procedure. The enthalpy was measured to be 98% accurate and also unaffected by the presence of noise. Bottom-left: Simulated DSC data displaying overlapping transitions and a nonzero change in heat capacity. Bottom-right: Gaussian distributions fit to the baseline treated data the enthalpies were within 5% agreement of the values with which they were generated. The y-axis for both figures are the same and represent the same units (kcal/ mol K).

centered at 327K and one centered on 329K represented in the raw data. There are a couple of possible interpretations of this data. 1) There is a contaminant in each of the samples to a varying degree. However, with the resolution offered by the SDS-PAGE gels, the presence of a contaminant can not be substantiated due to the presence of one band seen in both sets of gels shown in Fig. 2.3. 2) There are subtle transitions in ABD1 that might correspond with the different globular CH domains unfolding at different temperatures, which suggests the significant population of intermediate states throughout ABD1's transition. However, the Gaussian distributions still showed some areas of systematic deviation from the experimental data. So we wanted to determine whether this systematic deviation could be addressed by utilizing a different distribution function and if using a different distribution function might alter the observation of existing intermediate states.

2.10.1 The Hubbert function

The Hubbert function describes the geometry of a theoretically generated two-state transition much more accurately than a Gaussian distribution.[35] It is a derived logistic distribution curve originally developed for geologists in the charting and depletion of resources.[35] The Hubbert function doesn't have any physical significance in fitting the data other than it models experimental data well making it a purely phenomenological explanation.[35] The equation is as follows

$$C_p(T) = C_{p,0}(T) + \frac{4A\exp(\frac{T-T_m}{w})}{(1+\exp(\frac{T-T_m}{w}))^2} \tag{2.29}$$

where T_m is the center of the peak (the transition temperature), w is the width of the peak, and A is the amplitude or height of the peak.

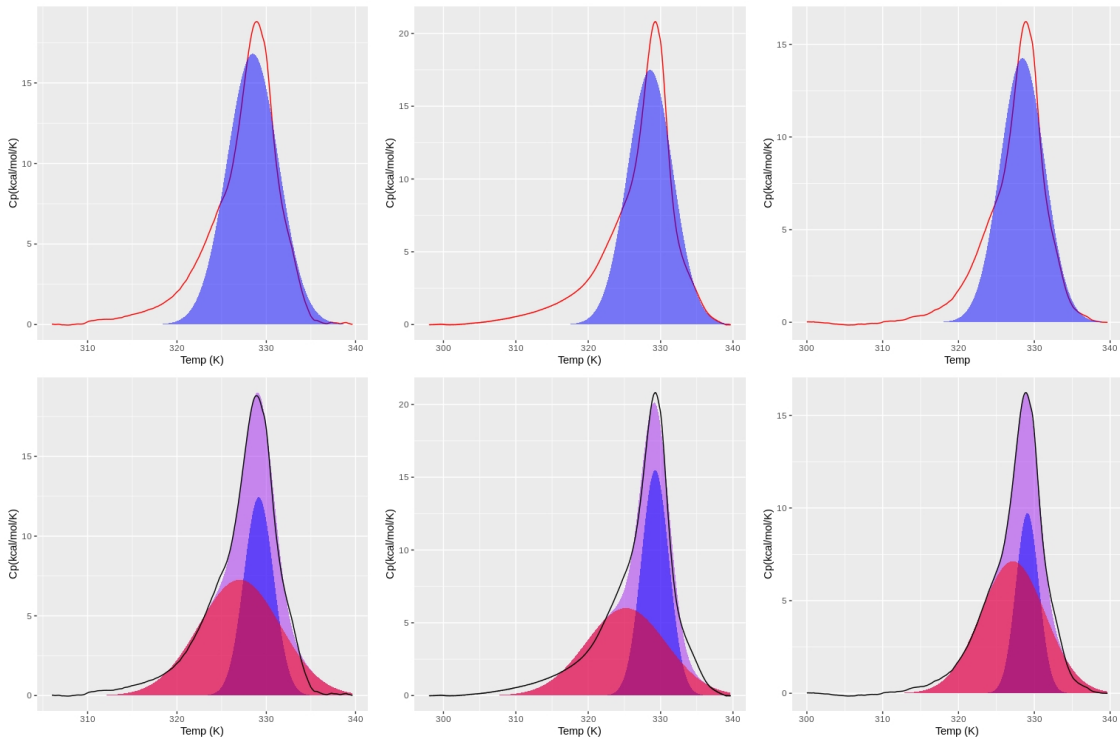


Fig. 2.15: Fitting DSC data after baseline subtraction to Gaussian distributions. Each set of experimental data is represented by a separate column. Row 1: represents a one Gaussian fit. Reported ΔH's: 121, 137, and 104 kcal/mol, respectively Row 2: Fitting ABD1 data to two Gaussian distributions. Red populations are centered on 329K and blue are centered on 327K. Reported ΔH's red = 52, 82, and 36 kcal/mol, respectively; blue = 80, 74, and 75 kcal/mol respectively.

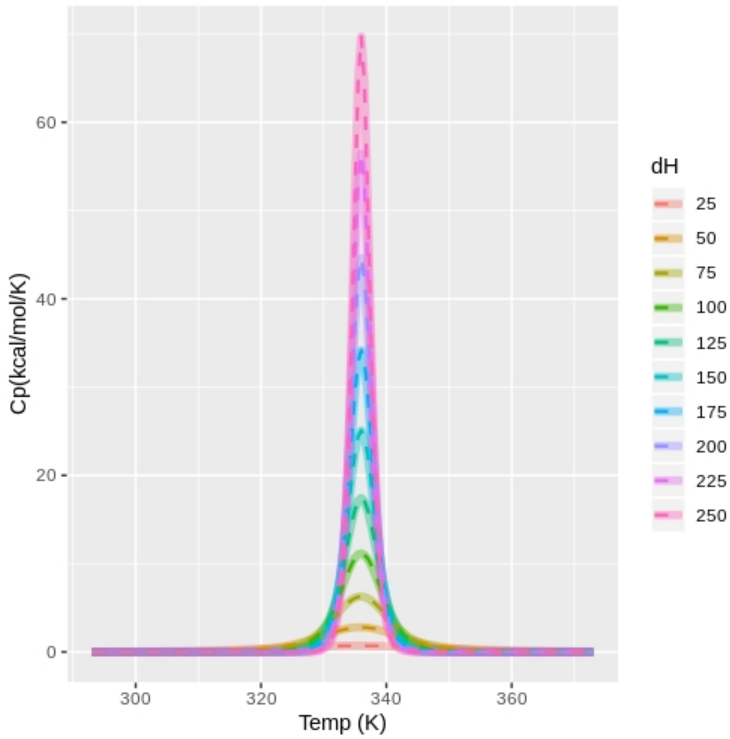


Fig. 2.16: Theoretically generated two-state transitions with enthalpies ranging from 25kcal/mol to 250kcal/mol. The experiments are colored by the generated enthalpy values. Model fits of the curves are represented as dashed lines.

Theoretically produced heat capacity scans assuming a two-state transition were fit with the Hubbert function. The function appears to account for the shape of the theoretically generated heat capacities over a range of enthalpies from 25 to 250 kcal/mol (see Fig. 2.16). Being that the Hubbert function has better agreement with two-state transition theory than a Gaussian distribution, the estimated thermodynamic values were taken from fits to a Hubbert functions rather than Gaussian distributions (see Tabs. 2.3 and 2.4). The fits can be visualized in Fig. 2.17.

Regression with the Hubbert distributions don't show convergent results for the presence of a second state as interpretted by the one degree standard deviation in the definition of the second transition temperature. Moreover, the two Hubbert model appears to fit on the higher temperature portion of the transition that is shown in

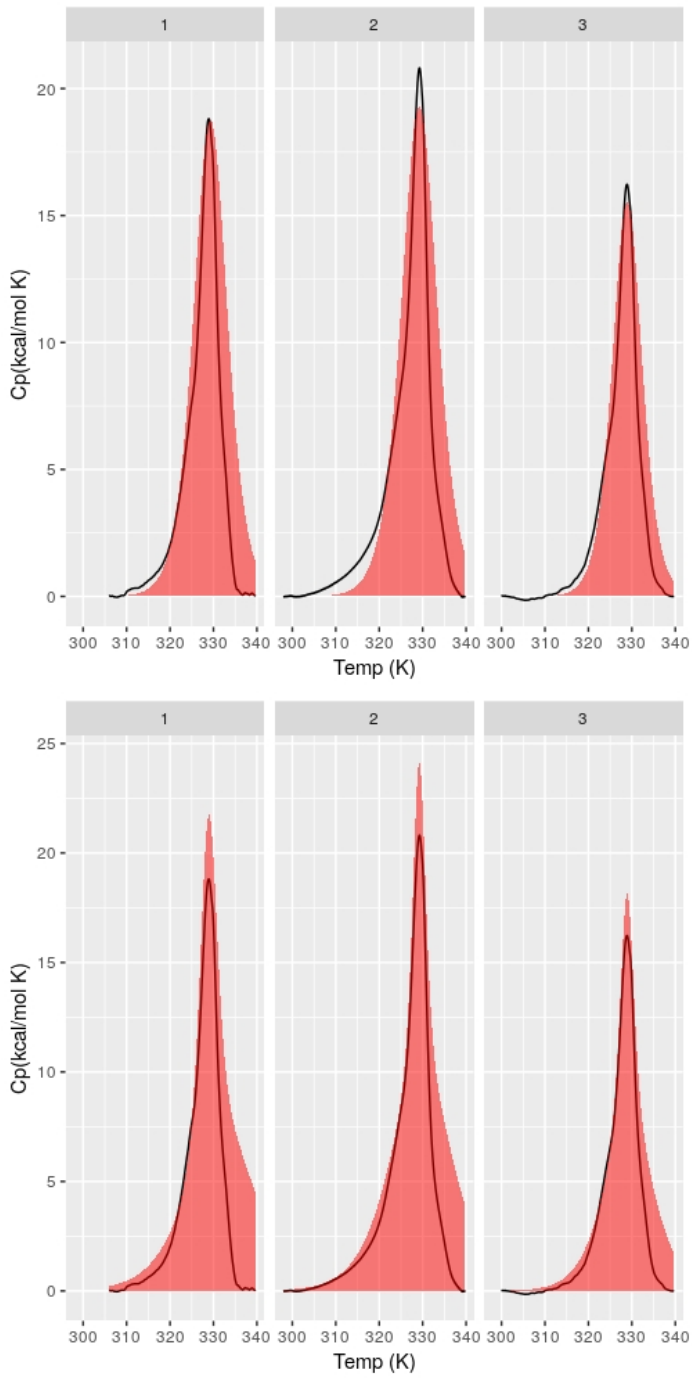


Fig. 2.17: Experimental data on ABD1 were fit to single and double Hubbert distributions where the black line represents experimental data after baseline subtraction and the shaded red region represents the total modeled enthalpy. Top: Baseline subtracted experimental ABD1 data fit with a single Hubbert distribution. Bottom:Baseline subtracted experimental ABD1 data fit with two Hubbert distributions.

the bottom of Fig. 2.16 as compared to the single Hubbert model fits on the top of Fig. 2.16.

The Hubbert function might serve as a metric for future experiments to determine if a single peak is in fact described by a two-state model or not. The geometry of a two-state transition without a baseline change in heat capacity governed by one parameter (ΔH_{T_m}). The Hubbert function has another degree of freedom in that the curve is described by its height and its width. This means that if the Hubbert function can fit transitions with a wide range in enthalpies (as was previously shown, see Fig. 2.16) the height and width of the Hubbert function must be correlated such that one can be described as a function of the other. We explored this possibility which can be visualized in Fig. 2.18.

In Fig. 2.18 the height of the Hubbert function seems to increase exponentially with the defined enthalpy of transition, whereas the width of the Hubbert function appears to decrease exponentially with respect to the enthalpy. Taken together, this suggests a power law relationship between the height and width of a two-state transition as defined by the Hubbert function. The lower two graphs in Fig. 2.18 show the power law relationship ($H = a * W^b$, where H and W represent height and width, respectively and a and b are fit parameters) between the height and width of the Hubbert parameters where each colored point on the plot represents a two-state transition generated with a different enthalpy value (the coloring is retained from Fig. 2.16. The lower left graph of Fig. 2.18 represents a log-log plot of parameters where the lower right plot represents the parameters with linear axes. This relationship was fit first by linearization in log space as a proof of concept that it is described by a power law as well as for the seeding values for a nonlinear regression fit. The final fit parameters are listed in the caption of Fig. 2.18. It is possible that the definition of these power law relations might be used to differentiate a two-state process from

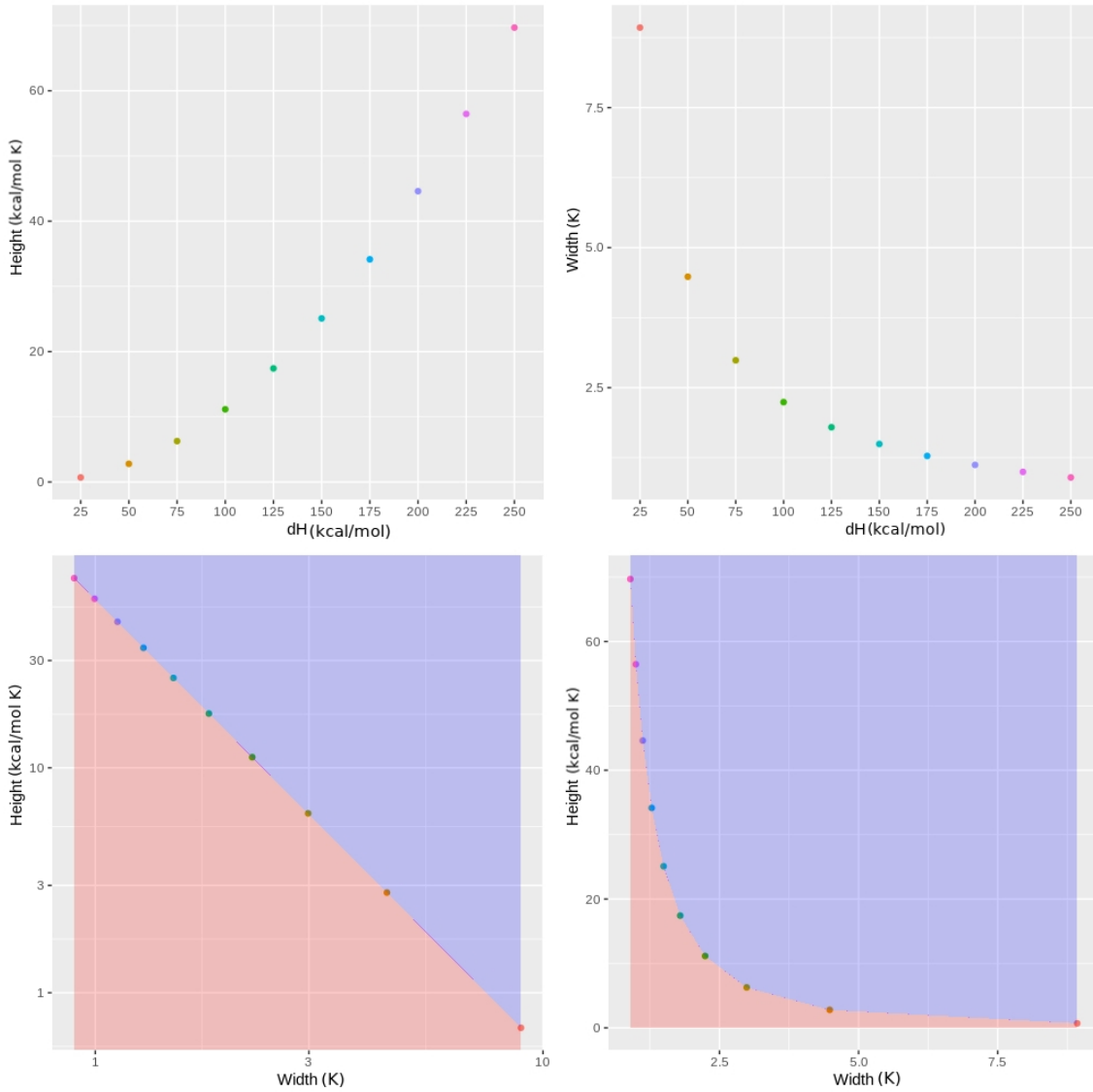


Fig. 2.18: Compiled height and width values determined from fitting theoretically generated two-state transitions of enthalpies varying from 25kcal/mol to 250kcal/mol to a Hubbert distribution. The top two graphs display the relationship of the height (top-left) and width (top-right) fit values to the enthalpy value with which the data was generated. The points are colored by the enthalpy values with which they were generated. The second two graphs represent the linearization and fit (bottom-left) of the height and width fit values where as the bottom-right graph represents the relationship between the fit values on linear axes. The coloring represents sectioning of the space into values that would represent transitions that are too broad (red) or too narrow (blue) to be represented as a two-state transition. The power model $H = a * W^b$ returned fit values of $a = 55.394$ and $b = -1.705$.

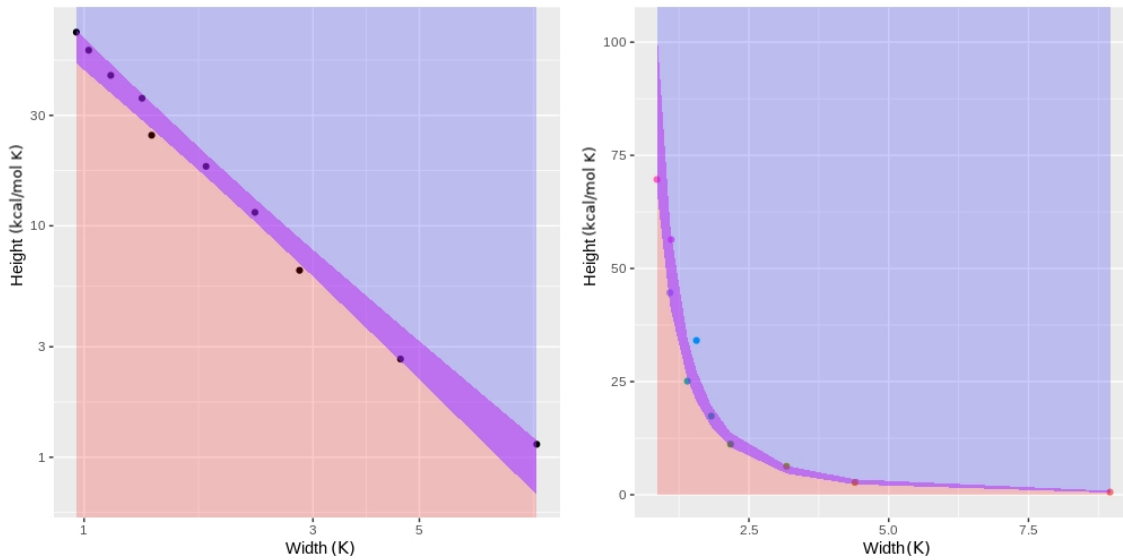


Fig. 2.19: Suggested system for classification of thermodynamic transitions. The blue region represents transitions of disproportionately sharp transitions indicative of cooperative effects, the red region represents disproportionately wide transitions (likely multiple transitions), and a purple region where an acceptable definition of a two-state transition could be adopted. The two-state region (purple) was arbitrarily defined by introducing noise into the fit parameters of the Hubbert function.

other types of transitions. A hypothetical examples of this type of definition is shown in Fig. 2.19.

In Fig. 2.19 the graph is segmented into three sections: the blue section representing transitions with a disproportionately sharp transition (suggestive of cooperative effects), a red portion representing transitions that are disproportionately wide (suggesting multiple transitions or out of equilibrium processes), and a purple portion representing a two-state process. The regions defined in the graph are meant to display the qualitative development of this mode for classifying thermodynamic transitions. The two-state region was arbitrarily defined by introducing a small amount of normally distributed noise into the height and width fit parameters to exemplify the geometry of the different regions. The Hubbert function was fit to the example data generated by Karplus et al. with which they exemplified the idea that

	1	2	3	Ave
k (kcal/mol)	18.8 ± 0.1	19.3± 0.1	15.5± 0.1	17± 5
SD (K)	2.61 ± 0.02	2.78± 0.02	2.30± 0.02	2.6± 0.5
T_m (K)	329.31± 0.02	329.24± 0.03	328.93± 0.02	329.2± 0.5

Tab. 2.3: Results from baseline subtracted experimental ABD1 data fit with a single Hubbert function. The fit parameters define the following features of the curve: k defines the height, SD the width, and T_m the center. The average is reported along with the 95% confidence interval.

the Van't Hoff criterion is not sufficient to define a two-state transition. The results from the fit are shown in Fig. 2.20. The resulting height and width parameters for the fit are 43.0 kcal/mol and 1.1 K, respectively. Comparing this to Fig. 2.18, this would fall within the purple region that accounts for transitions that are disproportionately wide which suggests the presence of more than one transition. Therefore, data fit to the Hubbert distribution can be used to determine if a transition is two-state through comparing the height and width of the fit. The transitions for ABD1 when fit to a Hubbert function display similar fit parameters in that they are disproportionately wide as are listed in Tab. 2.3, whereas the Van't Hoff analysis of ABD1 lead to the conclusion that the transition was two-state in nature because the fitted enthalpy falls within the standard deviation of the determined Van't Hoff enthalpy.

2.11 Model free approach to fitting

This beginning of this section summarizes the paper in which this method of analysis was originally described by Biltonen and Freire.[36]

Utilizing the innate enthalpic signature it is possible for one to directly assess the number of transitions encompassed within the experimental data without pre-supposing any sort of model. The basis for the model defines the partition function

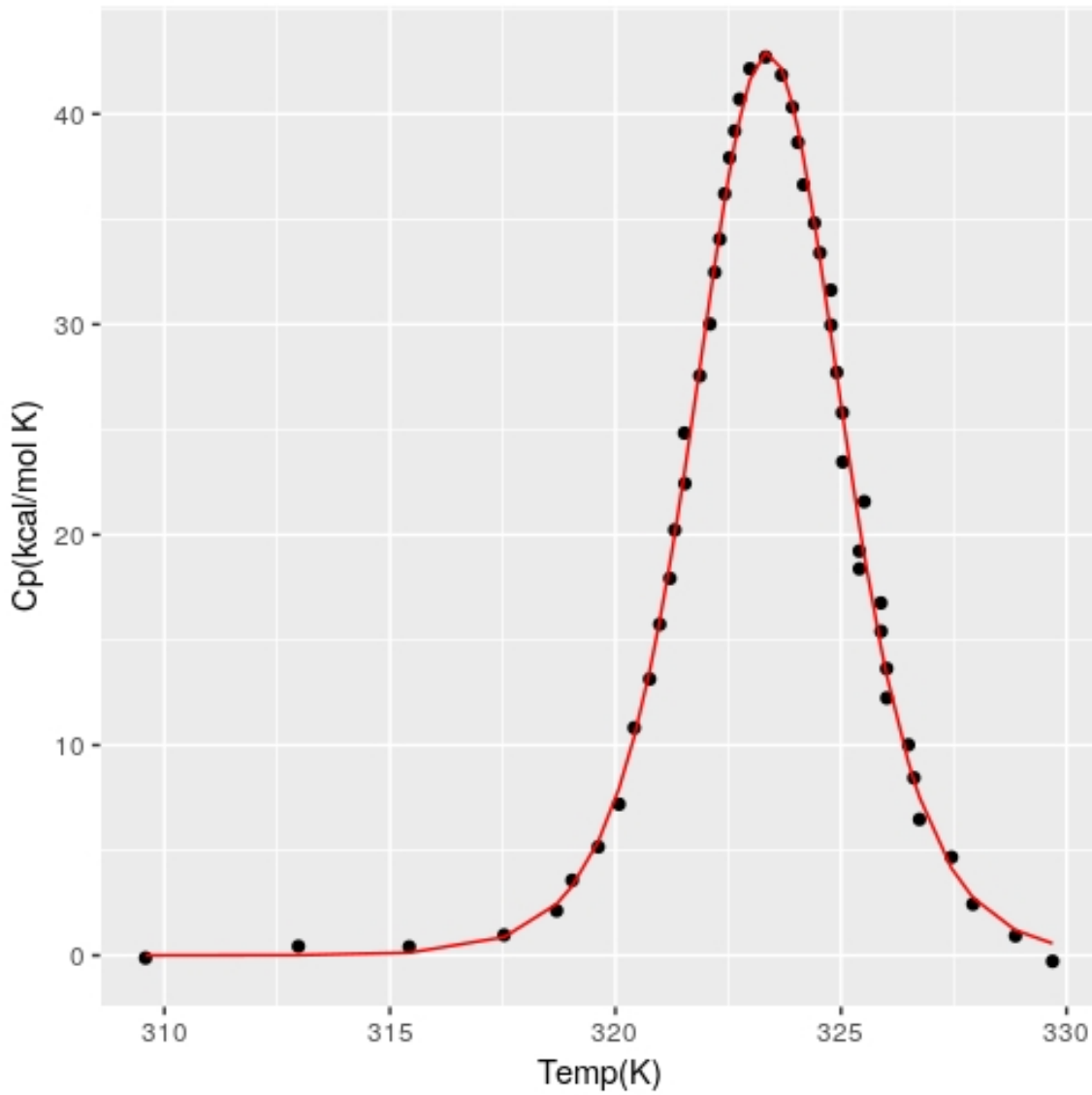


Fig. 2.20: Theoretical data generated by Zhou et al. with the sigmoidal baseline subtracted (black points) fit with a Hubbert function (red line). The resulting height and width of the Hubbert fit parameters are 43.0 kcal/mol and 1.1 K, respectively. Rights provided to this Protein Science article provided by the Copyright Clearance Center's RightsLink.

	1	2	3	Ave
k_1 (kcal/mol K)	14.8 ± 0.1	14.02± 0.07	10.4± 0.1	13 ± 5
SD_1 (K)	1.31 ± 0.01	0.979± 0.007	1.01± 0.01	1.1 ± 0.5
$T_m 1$ (K)	328.910 ± 0.009	329.212± 0.007	328.91± 0.01	329.0 ± 0.5
k_2 (kcal/mol K)	7.5 ± 0.1	10.17± 0.06	7.8± 0.2	8.5 ± 2
SD_2 (K)	5.28± 0.07	4.76± 0.03	3.78± 0.04	4.6± 2
$T_m 2$ (K)	331.78 ± 0.09	329.73± 0.02	329.3 ± 0.03	330± 2

Tab. 2.4: Results from baseline subtracted experimental ABD1 data fit with two Hubbert functions. The fit parameters define the following features of the curve: k defines the height, SD the width, and T_m the center. The number in the first column directly following the parameter name describes which transition the fit parameter corresponds to. Averages are reported along with the 95% confidence interval.

as C where $C = \sum_c W_c$, where W represents the probability of existing in the conformational state c. The probability of existing in a given state can be represented using the Boltzmann distribution $W_c(T) = \omega_c e^{\frac{-G_c(T)}{RT}}$. Therefore,

$$C(T) = \sum_i^n \omega_i(T) e^{\frac{-G_i(T)}{RT}} \tag{2.30}$$

$\omega_i(T)$ represents the degeneracy of states that are at a given energy level $G_i(T)$ and n is all states in solution. Absolute free energy is impossible to determine, but this partition function has two alternative modes of expression $Z(T)$ and $Q(T)$ that are described as follows. $Q(T)$ is the ratio of $C(T)$ and the statistical weight of the lowest energy state and $Z(T)$ is the ratio of $C(T)$ and the statistical weight of the highest energy state. It is convenient to use these two states as references as they can be measured easily experimentally at either very low or very high temperatures, respectively. For a system with multiple chemical species, let the mole fraction

$y(T)$ of each species k be represented by $y_k(T) = \frac{\#_k(T)}{N_T}$ where $\#$ represents the total number of molecules of k and N_T the total number of molecules in solution. Therefore, this will always satsify the identity $\sum y_k(T) = 1$ for all positive values of T. Then any property of the system $X(T)$ can be described by the average over all populations in solution namely $\bar{X}(T) = \sum_k y_k < X_k(T) >$ where $\bar{X}(T)$ is the average value over all species in solution and $< X_k(T) >$ is the per mole value over all configurations of species k. By definition this average is given by[36]

$$< X_k(T) >= \sum F_{k,i}(T) X_{k,i}(T) \tag{2.31}$$

For the present investigation there is assumed to be only one molecule of interest (the purified ABD1 protein) therefore the k subscript is unecessary and $\bar{X}(T) =< X(T) >$. Utilizing a Boltzmann distribution to calculate the fraction occupancy of each energetic state of energy i we get

$$F_i(T) = \frac{\omega_i(T) e^{-G_i(T)/RT}}{\sum_i \omega_i e^{-G_i(T)/RT}} \tag{2.32}$$

Writing this in terms of the relative partition coefficient $Q(T)$ described above, this becomes[36]

$$F_i(T) = \frac{\omega_i(T) e^{-\Delta G_i(T)/RT}}{Q(T)} \tag{2.33}$$

where $\Delta G_i(T) = G_i(T) - G_0(T)$, or the difference in free energy between the ith state and the reference state. Utilizing equations Eq. 2.31 and 2.33 the observed enthalpy is

$$< H(T) >= RT^2 \frac{\delta ln(Q(T))}{\delta T} \tag{2.34}$$

Separating the differentials and integrating leads to the equation

$$\psi = ln(Q(T)) = \int_{T_0}^{T} \frac{\Delta \bar{H}}{RT^2} dT \tag{2.35}$$

Freire has shown Eq. 2.35 satisfies the Lipschitz condition for ψ so that the solution is uniquely defined and is stable with respect to the lower bound of integration.[36] For the case of one chemical species in solution $e^{\psi} = Q$, thus giving the partition function for that species. For multistate transitions, $Q(T)$ can be expressed as follows for simplicity $Q(T) = 1 + \sum_{i-1}^{n} e^{\frac{-\Delta G_i(T)}{RT}}$ and the fraction at energy state i is given by[36]

$$F_i(T) = \frac{e^{\frac{-\Delta G_i(T)}{RT}}}{Q(T)} \tag{2.36}$$

Remembering that $< \Delta H(T) >= \sum_{i=0}^{n} \Delta H_i(T) F_i(T)$. We can plug in Eq. 2.36 which results in the following expression for observed enthalpy values

$$< \Delta H(T) >= \frac{\sum_{i=0}^{n} \Delta H_i(T) e^{-\frac{\Delta G_i(T)}{RT}}}{Q(T)} \tag{2.37}$$

The observed enthalpy of transition can be evaluated directly after implementing the algorithm proposed above for the elimination of the baseline shift in ΔC_p. It is directly integrated as follows $< \Delta H(T) >= \int_{T_0}^{T} (C_p(T) - C_{p,0}(T)) dT$ where $C_{p,0}(T)$ is the baseline heat capacity of the folded state. Integration of this equation yields

$$ln(Q(T)) = \int_{T_0}^{T} \frac{< \Delta H(T) >}{RT^2} dT \tag{2.38}$$

which can in turn be used to express the fraction of the reference state $F_0(T)$.

$$F_0(T) = \frac{1}{Q(T)} = exp(-\int_{T_0}^{T} \frac{< \Delta H(T) >}{RT^2} dT) \tag{2.39}$$

and similarly for the final state F_n

$$F_n(T) = \frac{1}{Z(T)} = \exp(-\int_T^{T_n} (\Delta H_n(T) - < \Delta H(T) >) \frac{1}{RT^2} dT) \tag{2.40}$$

Given the identity expressed above for the conservation of mass, the fraction occupying any number of intermediate states $I(T)$ is given by

$$F_I(T) = 1 - F_0(T) - F_n(T) \tag{2.41}$$

Using this development the observed change in enthalpy can be normalized to the fraction not in the ground state by evaluating[36]

$$\frac{< \Delta H(T) >}{1 - F_0(T)} = \Delta h_1(T) + \frac{\sum \Delta H_i(T) e^{\frac{-\Delta G_i(T)}{RT}}}{Q_1(T)} \tag{2.42}$$

In this way, this method can be simply used to evaluate the fractions of the native, unfolded, and intermediate states indiscriminate of the number of intermediate states. This ends the summary of the work of Biltonen and Freire in developing this method of model free fitting.[36]

In the study of ABD1 this method of analysis revealed a population of intermediate states as high as 65%. The population of the intermediate states occurs towards the center of the transition, thus likely implying their role in the small bumps that are noted on either side of the transition (see Fig. 2.21).

The population of the ground state (black) in Fig. 2.21 can be seen trending from complete occupancy to extremely small population as a function of temperature which fits with general knowledge that the population of this state will be depleted and not repopulated as temperature increases. The population of the denatured state (red) in Fig. 2.21 starts very close to zero and trends towards complete occupancy as

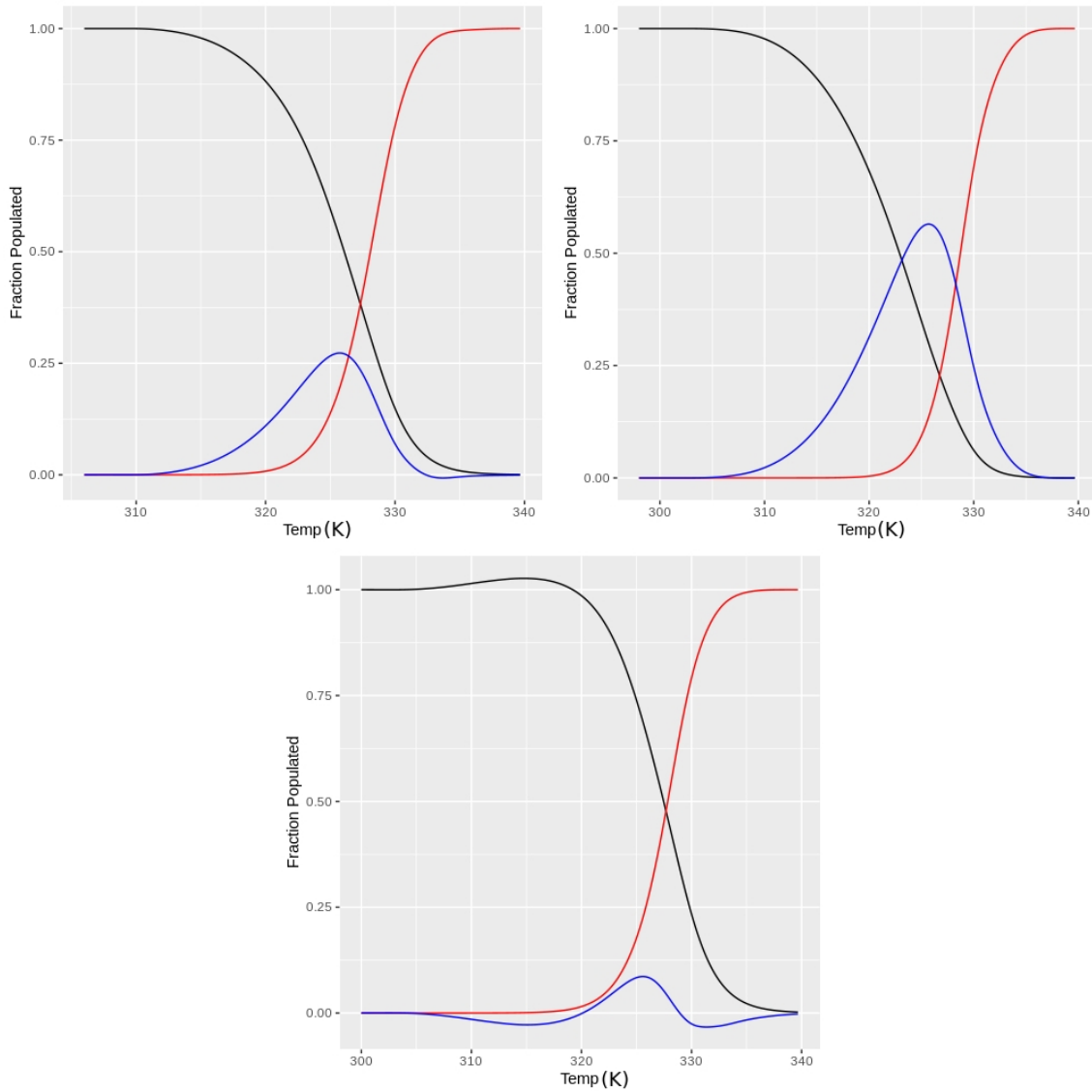


Fig. 2.21: Three graphs representing the the fraction population of the reference state $F_0(T)$ (black), final state $F_n(T)$ (red), and intermediate states $F_I(T)$ (blue) as a function of temperature for all three scans.

a function of temperature thus implying that the denature state is more energetically favorable at high temperatures. The intermediate states are seen to largely be populated with a max of about 65% as seen in Fig. 2.21, however the population of the intermediate state is highly varied as can be seen comparing the blue curves. The curves for the last scan also seem to show irregularities in that they fluctuate above and below one and zero fractions of occupancy. The average maximum percent occupancy also has a standard deviation that encompasses 0. These two observations in the data made it inadvisable to further fit the intermediate transitions of ABD1 as they would likely be prone to erroneous interpretation due to overfitting. Future studies using DSC analysis will be able to utilize this approach to tease out the population of intermediate states and the associated enthalpies of transition. The third scan, which has a higher scan rate ($1.5\frac{^\circ C}{min}$) and the smallest population of the intermediate state, is consistent with the population of the intermediate states being kinetically controlled. The slight changes that the kinetics of the intermediate states might play a significant role in the mechanism of ABD1. The transition enthalpies between the reference state and the intermediate states accounted for 63.9, 51.38, and 88.3 kcal/mol of the overall change in enthalpy for scans 1, 2, and 3, respectively thus leaving the rest of the enthalpy to the one or more intermediate transitions to the final unfolded states. Assuming that ΔC_p is proportional to the enthalpy $\Delta G(37^\circ C)$ was calculated to be 1.99, 1.39, and 3.26 kcal/mol respectively. For all three scans this amounted to a free energy of unfolding of 2 ± 1 kcal/mol. The complete listing of thermodynamic values from both detected transitions are represented in Tab. 2.5. The results listed within Tab. 2.5 suggest the two transitions have indistinguishable enthalpies and free energies of transition and only differ significantly in their transition temperatures. The observation of indistinguishable enthalpies and free energies are likely due to a sequential unfolding of the CH

	Transition 1	Transition 2
ΔH (kcal/mol)	68 ± 45	72 ± 20
ΔC_p (kcal/mol K)	2.21 ± 2.23	3.1 ± 2.23
T_m (K)	325.4 ± 5.2	328.0 ± 0.7
$\Delta G(37°C)$ (kcal/mol)	2 ± 1	2.24 ± 0.6

Tab. 2.5: Thermodynamic values reported from the model free approach to analysis. The 95% confidence intervals were generated from triplicate scans of ABD1.

domains as their similar size would likely lend themselves to similar unfolding enthalpies and free energies.

A comparison between the different methods of analysis can be seen in Tab. 2.6. The reported confidence intervals for the thermodynamic fit parameters from each of the methods are large. The size of the confidence intervals is likely attributable to the noted differences that existed between scans in that they displayed varying peak heat capacities as well as the change in heat capacity (ΔC_p) values. None of the regression methods (regressing Gaussian, Hubbert, and Boltzmann distributions) were able to detect the presence of more than one transition based off of the overlapping confidence intervals measured. Some of the regression methods were unable to converge to a single minimum as is reported in Tab. 2.2. Two methods of analysis suggested the presence of more than one transition: the method of comparing the height and width proportions of a Hubbert function and the model free approach. Comparing the Hubbert fit proportions serves only as a metric for determining if set of data can be considered a two-state transition, whereas the model free approach can also be used to measure the thermodynamic parameters.

The model free approach is also unique in that it does not presuppose any given model, but is rather a direct readout of the experimental data. The confidence intervals for the transition temperatures (T_m) determined from the model free

approach also overlapped. However, it is still valid to conclude that the transition temperatures (T_m) reported from the model free approach are indeed separate. The model free approach allows for the evaluation of multiple transitions even if the transition temperatures (T_m) are not significantly different. This is because the model free approach extrapolates the fraction population of states directly from experimental data. After extrapolating the fraction population of the folded and unfolded states (F_n and F_u, respectively) the sum of the total fraction occupancy of these two states would have to be 100% if they were the only states populated. Since the sum of the percent occupancy of the folded and unfolded states deviated from 100% (most notably in the temperature range 325K to 333K), the population of an intermediate state or set of states is implied, thus necessitating a second transition. The presence of intermediate unfolding states as determined by the model free approach is not dependent upon measured transition temperatures (T_m) that are significantly different. The presence of intermediate states is dependent upon the direct evaluation of a significant population of intermediate states from experimental data. The experimental evidence presented in this thesis therefore suggests the presence of intermediate unfolded states (and similarly two notable transitions) in the denaturation pathway of ABD1, while maintaining that the transition temperatures do not differ significantly. This conclusion follows from the fact that the calculation of populated states (folded, intermediate, unfolded) precedes the calculation of the transition temperatures (T_m) in the model free approach.

Further analysis on the thermodynamics of the intermediate states could result in the determination of a higher number of intermediate states, but will require new higher quality scans of ABD1 to be done. Alternatively, The presence of intermediate states could be explored utilizing methods with region specific information on protein

	Van't Hoff	Boltzmann two-state	Hubbert 1	Hubbert 2	Model free
$\Delta H_{T_m,1}$ (kcal/mol)	110 ± 119	118 ± 12	147 ± 42	57.2 ± 22	68 ± 45
$\Delta C_{p,1}$ (kcal/mol K)	-	5 ± 5	3.5 ± 1.6	0.9 ± 1	2.21 ± 2.23
$T_{m,1}$ (K)	-	328.0 ± 0.5	329.2 ± 0.5	329.0 ± 0.5	325.4 ± 5.2
$\Delta G_1(37°C)$ $(kcal/mol)$	-	4 ± 5	6.539 ± 1.6	2.782 ± 1	2 ± 1
$\Delta H_{T_m,2}$ (kcal/mol)	-	-	-	176 ± 41	72 ± 20
$\Delta C_{p,2}$ (kcal/mol K)	-	-	-	2.62 ± 1.5	3.1 ± 2.23
$T_{m,2}$ (K)	-	-	-	330 ± 2	328.0 ± 0.7
$\Delta G_2(37°C)$ $(kcal/mol)$	-	-	-	9 ± 2	2.2 ± 0.6

Tab. 2.6: Thermodynamic values documented from triplicate analysis of ABD1 data. All values are reported with 95% confidence intervals. Each column represents a different method of analysis. Values which are not calculated by a given method of analysis are marked by a minus sign (-). Gaussian fitting methods were not included due to the noted deviation it showed from theoretical data in Fig. 2.13.

folding such as NMR, EPR, or even circular dichroism or fluorescence lifetime (should the intermediate state account for a measurable change in secondary structure or solvent exposure of Tryptophan residues).

Conclusion

3

Through this thesis I've described how we utilized mutation data on Muscular Dystrophy inform our studies. Analysis showed that ABD1 was the most representative domain of Muscular Dystrophy based on the total number (ABD1 accounted for 33% of disease causing mutations in Dystrophin) and density of disease causing mutations (ABD1 had the second highest density of disease causing mutations). We evaluated the physical characteristics of the disease causing mutations which revealed a strong correlation to the predicted change in heat capacity of the given mutation with the likelihood of it being disease causing. We then chose an experimental methodology that could test a hypothesis consistent with this observation: DSC. We designed the protein into a DNA plasmid, transformed it into E. Coli, and purified it by methods outlined in the Materials and Methods section. The purity of ABD1 from the methods described above were affirmed based on SDS-PAGE analysis Fig. 2.3.

The experimental data was with a two-state model, but showed a systematic deviation. Through increasing the efficiency of analysis by automating portions of the work flow within R, we were able to test how different distribution functions agreed with theory and ultimately modeled the experimental data. Our tests suggested that a Gaussian distribution matches theory well (to within 5%), but the Hubbert function fits theory much more reliably (to within 1% over a large range of predicted transition enthalpies). However, neither of these functions provided unequivocal

evidence for the existence of intermediate states in the experimental data. We then analyzed the data with a model free approach outlined by Biltonen and Freire and found the population of intermediate states, that were not detected by a Van't Hoff analysis, though the resolution was insufficient to determine the exact number of intermediate states. The model free approach did suggest that the intermediate states were kinetically controlled due to the noted scan rate dependence.

Nonetheless, we determined that the change in heat capacity did not differ significantly from predictions based on molecular weight, but was smaller than predicted from a sequence based prediction.[28, 29] Ultimately, these observations are consistent with the crystal structure of Dystrophin ABD1 (1DXX) and provide the upper energetic barrier for which ABD1 can withstand, without causing a depletion of the native state in solution.[48]

The folded stability of Dystrophin's ABD1 domain likely plays a role in the development of DMD. However, more detail is needed to determine the importance of the intermediate unfolded states of ABD1 as well as the possibility that the population of intermediate states is kinetically controlled. Details on the role folded stability plays in the function of ABD1 is needed in order to determine the importance of individual mutations on the function of Dystrophin as well as develop intelligent treatment options for MD. An alternative hypothesis that the mechanism for the proper function of ABD1 with a kinetically controlled intermediate state is proposed here with the observation of an intermediate state that shows a slight kinetic dependence and is likely due to the sequential unfolding of the separate CH domains.

3.0.1 Future directions

In order to determine if the mutations noted affect the folded stability of ABD1, it is necessary to purify the ABD1 domain with point mutations corresponding to

mutations noted in the disease. Then, a thermodynamic comparison, similar to the one just presented, using DSC methods would allow for a direct comparison of the folded stability through a comparison of free energies at $37°$C. However, determining the role of folded stability requires in vivo experiments to be done in order to determine the reasonable range of forces and/or energies Dystrophin and ABD1 are expected to encounter naturally. For the DSC experiments it is suggested that once the researcher determines the reversibility of the transition, all subsequent scans be run to a much higher final temperature so as to mitigate any error in establishing the post-transitional baseline and also run scan rate dependent scans ranging from $1\frac{°C}{min}$ to $2\frac{°C}{min}$ in order to assess the possible impact of mutations on the kinetically controlled intermediate state. The researcher is advised not to go any slower as it may lead to less reversibility and aggregation which was noted with ABD1 when scan rates lower than $1\frac{°C}{min}$ were used.

The second key set of experiments to gain a mechanistic insight into the operation of Dystrophin and the actin cytoskeleton will include determining the binding affinity and mechanism of ABD1. The self polymerizing nature of actin will provide a bit of a challenge for quantifying the affinity. The polymerization of actin is not well understood in quantitative detail and would need to be characterized before any quantitative analysis on binding can be done. A possible method to evaluate the binding affinity would be using fluorescence studies given the number of aromatic residues noted in both ABD1 and monomeric actin sequences. Thus, the studies could trace either the fluorescence intensity or fluorescent lifetime of the native tryptophans. Both signals would increase with the extent of binding because upon binding, exposed aromatic residues would be shielded from quenching due to water exposure. However, by monitoring the lifetime we would be able to obtain a better signal to noise ratio given the more sensitive response that is commonly noted

with lifetime as opposed to fluorescence intensity. Fluorescence approaches will be complicated by the scattering nature of polymerized actin. Thus, the binding model should only include four states in reference to actin 1) unbound g-actin, 2) bound g-actin, 3) unbound f-actin, 4) bound f-actin. The researcher will be able to assess transitions from states 1) to 3) and 2) to 4) by monitoring the binding at another far removed wavelength that would account for the extent of f-actin polymerization as monitored through scattering. If the intensity at that wavelength is relatively constant then the binding model is greatly reduced in complexity to ABD1 binding g-actin and f-actin, respectively.

Once the experimentalist has quantified how to control for polymerization of actin with respect to ABD1 titrations binding could then be measured using Isothermal titration calorimetry (ITC). ITC would give additional information such as whether the binding event is enthalpically or entropically driven, which ultimately serves to inform the understanding of interactions and binding events between Dystrophin and actin and possibly for other binding events involving actin. Analyzing these titration experiments would also necessitate the development of a similar R application being developed for ITC as was developed for DSC.

Overall, upon completing these studies, our understanding of Dystrophin's association to the cytoskeleton, as well as Dystrophin's stability would be better understood. This information will be indispensable in understanding the complex story that is the mechanism of Dystrophin in the human body and the amount of forces they can withstand.

Bibliography

(1) Uversky, V. N.; Oldfield, C.; Dunker, A. K. *J Mol Recognit* **2005**, *18*, 343–84 (cit. on p. ii).

(2) Dimova, R. *Advances in Planar Lipid Bilayers and Liposomes* **2012**, *16*, 1–50 (cit. on p. 1).

(3) Kuzmin, P. I.; Akimov, S. A.; Chizmadzhev, Y. A.; Zimmerberg, J.; Cohen, F. S. *Biophysical Journal* **2005**, *88*, 1120–1133 (cit. on p. 1).

(4) Hlavacek, W. S.; Faeder, J. R. *Science Signaling* **2009**, *2*, 46 (cit. on p. 1).

(5) Kumar, P.; Libchaber, A. *Biophysical journal* **Aug. 2013**, *105*, 783–93 (cit. on p. 2).

(6) Kirmizis, D.; Logothetidis, S. *International journal of nanomedicine* **Apr. 2010**, *5*, 137–45 (cit. on p. 2).

(7) Suchyna, T. M.; Sachs, F. *The Journal of physiology* **May 2007**, *581*, 369–87 (cit. on p. 2).

(8) Suchyna, T. M.; Markin, V. S.; Sachs, F. *Biophysical journal* **Aug. 2009**, *97*, 738–47 (cit. on p. 2).

(9) Decker, M. L.; Behnke-Barclay, M.; Cook, M. G.; Lesch, M.; Decker, R. S. *Circ Res.* **1991**, *69*, 86–94 (cit. on p. 2).

(10) Zhou, L.; Haiyan, L. *Journal of Neuropathology & Experimental Neurology* **2010**, *69*, 771–776 (cit. on p. 2).

(11) Centers for Disease Control and Prevention (CDC). *Prevalence of Duchenne/Becker muscular dystrophy among males aged 5-24 years – four states, 2007.* Tech. rep. 40; Atlanta: Centers for Disease Control and Prevention (CDC)., 2009, pp 1119–22 (cit. on p. 2).

(12) Tuffery-Giraud, S. et al. *Human Mutation* **June 2009**, *30*, 934–945 (cit. on p. 2).

(13) Johan T. den Dunnen Leiden Muscular Dystrophy pages., 2014 (cit. on pp. 2–4, 6).

(14) Krieger, C. C.; Bhasin, N.; Tewari, M.; Brown, A. E. X.; Safer, D.; Sweeney, H. L.; Discher, D. E. *Cytoskeleton (Hoboken).* **2010**, *67*, 796–807 (cit. on p. 2).

(15) Levine, B. A.; Moir+, A. J. G.; Patchell, V. B.; Perry, S. V. *FEBS Lett.* **1990**, *263*, 159–162 (cit. on p. 2).

(16) Fairclough, R. J.; Bareja, A.; Davies, K. E. *Experimental Physiology Exp Physiol* **2013**, 1101–1113 (cit. on p. 2).

(17) Fabbrizio, E.; Bonetkerrache, A.; Limas, F.; Hugon, G.; Mornet, D. *Biochemical and Biophysical Research Communications* **Aug. 1995**, *213*, 295–301 (cit. on pp. 3, 4).

(18) Fabbrizio, E.; Harricane, M.-C.; Pons, F.; Leger, J.; Mornet, D. *Biol Cell* **1992**, *76*, 167–174 (cit. on pp. 3, 4).

(19) Siciliano, G.; Simoncini, C.; Giannotti, S.; Zampa, V.; Angelini, C.; Ricci, G. *Acta Myol.* **2015**, *34*, 3–8 (cit. on p. 3).

(20) Ervasti, J. M. *Madame Curie Bioscience Database* **2013** (cit. on pp. 3–5).

(21) Lin, A. Y.; Prochniewicz, E.; Henderson, D. M.; Li, B.; Ervasti, J. M.; Thomas, D. D. *Journal of molecular biology* **June 2012**, *420*, 87–98 (cit. on pp. 3–5, 22).

(22) Gao, Q.; McNally, E. M. *Comprehensive Physiology* **2015** (cit. on p. 3).

(23) Crone, M.; Mah, J. K. *Current Treatment Options in Neurology* **2018** (cit. on p. 3).

(24) Singh, S.; Kongari, N.; et al.; Mallela, K. *Proc. Natl. Acad. Sci.* **2010**, 15069–15074 (cit. on p. 4).

(25) Singh, S. M.; Kongari, N.; Cabello-Villegas, J.; Mallela, K. M. G. *Proceedings of the National Academy of Sciences of the United States of America* **Aug. 2010**, *107*, 15069–74 (cit. on pp. 4, 10).

(26) Chen C Huang H, W. C. *Methods Mol. Biol.* **2017**, *1558*, 3–39 (cit. on p. 5).

(27) Benjamini, Y.; Hochberg, Y. *Journal of the Royal Statistical Society. Series B (Methodological),* **1995**, *57*, 289–300 (cit. on p. 9).

(28) Freire, E. In *Protein Stability and Folding*; Humana Press: New Jersey, 1995, pp 191–218 (cit. on pp. 10, 21, 25–27, 32, 34, 37, 39, 40, 44, 50, 76).

(29) Spolar, R. S.; Livingstone, J. R.; Record, M. T. *Biochemistry* **1992**, *31*, 3947–3955 (cit. on pp. 10, 32, 34, 76).

(30) Zhou, Y.; Hall, C. K.; Karplus, M. *Protein science : a publication of the Protein Society* **May 1999**, *8*, 1064–74 (cit. on pp. 10, 39, 40).

(31) Privalov, P.; Khechinashvili, N. *Journal of Molecular Biology* **July 1974**, *86*, 665–684 (cit. on p. 10).

(32) Brandts, J. F. *J. Am. Chem. Soc.* **1964**, *86*, 4302–4314 (cit. on p. 10).

(33) Mazurenko, S.; Kunka, S.; Beerens, K.; Johnson, C. M.; Damborsky, C. M.; Prokop, C. M. *Nature Scientific Reports* **2017**, *7* (cit. on p. 10).

(34) Mazurenko, S.; Stourac, J.; Kunka, A.; Nedeljković, S.; Bednar, D.; Prokop, Z.; Damborsky, J. *Nucleic Acids Research* **2018**, *46*, W344–W349 (cit. on p. 10).

(35) Vega, S.; Garcia-Gonzalez, M. A.; Lanas, A.; Velazquez-Campoy, A.; Abian, O. *Scientific reports* **Jan. 2015**, *5*, 7988 (cit. on pp. 10, 11, 57).

(36) Biltonen, R.; Freire, E. *Critical Reviews in Biochemistry* **1978**, *5*, 85–124 (cit. on pp. 10, 11, 21, 64, 67–69).

(37) Nallamsetty, S.; Waugh, D. S. *Nature Protocols* **Mar. 2007**, *2*, 383–391 (cit. on p. 14).

(38) Bates, D. M.; Watts, D. G., *Nonlinear Regression Analysis and Its Applications*, 2nd ed.; Wiley-Interscience: 1990 (cit. on pp. 20, 48, 49).

(39) Beale, E. M. L. Confidence Regions in Non-Linear Estimation., 1960 (cit. on pp. 20, 48, 49).

(40) Fealey, M.; Gauer, J.; Kempka, S.; Miller, K.; Nayak, K.; Sutton, B.; Hinderliter, A. *PLOS one* **2012** (cit. on pp. 25, 26, 34, 50).

(41) Montgomery, D.; Jordan, R.; McMacken, R.; Freire, E. *Journal of Molecular Biology* **July 1993**, *232*, 680–692 (cit. on p. 31).

(42) Fealey, M. E.; Mahling, R.; Rice, A. M.; Dunleavy, K.; Kobany, S. E. G.; Lohese, K. J.; Horn, B.; Hinderliter, A. *Biochemistry* **May 2016**, *55*, 2914–2926 (cit. on pp. 31, 50).

(43) Makhatadze, G.; Privalov, P. *Journal of Molecular Biology* **1990**, *213*, 375–384 (cit. on p. 32).

(44) Makhatadze, G. I.; Privalov, P. *Protein Sci.* **1996**, *5*, 507–510 (cit. on p. 32).

(45) Livingstone, J. R.; Spolar, R. S.; Record, M. T. *Biochemistry* **Apr. 1991**, *30*, 4237–44 (cit. on p. 33).

(46) Grothendieck, G. *CRAN* **2013** (cit. on p. 46).

(47) Baty, F.; Delignette-Muller, M.-L.; Charles, S.; Flandrois, J.-P.; Ritz, C. *CRAN* **2015** (cit. on p. 48).

(48) Fealey, M. E.; Horn, B.; Coffman, C.; Miller, R.; Lin, A. Y.; Thompson, A. R.; Schramel, J.; Groth, E.; Hinderliter, A.; Cembran, A.; Thomas, D. D. *Biophysical Journal* **2018**, *115*, 445–454 (cit. on p. 76).