AN ABSTRACT OF THE DISSERTATION OF

Aidan B. Estelle for the degree of Doctor of Philosophy in Biochemistry and Biophysics presented on September 13, 2022.

Title: Specificity, Allostery, and Multivalency in Binding to the Hub protein LC8

Abstract approved: _____

Elisar Barbar

Interactions between proteins are essential to life, driving and regulating a majority of processes within all living cells. Study of protein-protein interactions reveals that some proteins act as hubs within networks of interactions, binding to many partner proteins. These hubs therefore are of particular importance to understanding protein function, interwoven as they are with dozens of biological functions. LC8 is one such hub protein, binding to over 100 known clients and playing a role in many unrelated pathways. LC8 binding, mediated by a short linear motif in client proteins, induces a dimeric structure on clients, leading the protein to be referred to as a dimerization engine.

This thesis discusses the function of LC8, examining both the general properties of LC8 that facilitate LC8-client binding, and documenting and characterizing new LC8-binding proteins. Each of the three chapters of original work is a report of primary research. Chapter 2 is a detailed investigation of the thermodynamics of LC8 binding, which necessitated the development of a new method of analysis built on principles of Bayesian statistics. This method allowed us to measure detailed thermodynamics of LC8 binding, and demonstrate that LC8 favors a fully bound state, consistent with its function as an engine for dimerization. Chapter 3 is concerned with characterizing the LC8-binding linear motif, and development of a tool for prediction of LC8 binding. We collate a database of LC8-binding proteins and find that residues flanking the core motif sequence play an important role in regulating binding. The predictive tool uses a library of known LC8-binding and non-binding sequences to generate a scoring matrix for potential clients and has already been adopted by researchers studying LC8 interactions. In chapter 4 we present a characterization of a new LC8-binding protein named Kank1. Kank1, a cytoskeletal regulator found at the cell cortex, binds LC8 multivalently, forming a large

complex consisting of at least five LC8 dimer units. The complex forms with significant cooperativity, and unlike many multivalent LC8-interacting proteins, forms a homogenous stable oligomer, indicating that the complex may play a structural role, rigidifying the scaffold of Kank1. Lastly, chapter 5 discusses the impact of this work, and highlights of the work presented in each chapter. It additionally presents ongoing and future steps in the study of LC8 interactions.

This thesis additionally contains two appendices reporting primary research that is unrelated to LC8. The first is concerned with a protein from the Peroxiredoxin family of redox proteins. Peroxiredoxins unfold during catalysis, and we demonstrated that our model peroxiredoxin unfolds transiently in absence of catalysis, emphasizing that the protein is finely structurally tuned for catalysis. The second appendix discusses the nucleocapsid of the SAR-CoV-2 virus, examining the protein's interaction with RNA, which is essential to viral replication. We find that the protein can interact both specifically and nonspecifically with RNA, and that nonspecific binding is correlated to liquid-liquid phase separation, which is believed to be essential to some viral functions.

Specificity, Allostery, and Multivalency in Binding to the Hub Protein LC8

by
Aidan B. Estelle

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 13, 2022
Commencement June 2023

Doctor of Philosophy dissertation of Aidan B. Estelle presented on September 13, 2022.

APPROVED:

Major Professor, representing Biochemistry and Biophysics

Head of the Department of Biochemistry and Biophysics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Aidan B. Estelle, Author

ACKNOWLEDGEMENTS

First and most importantly I would like to thank my advisor and mentor, Dr. Elisar Barbar. Elisar, I could not have asked for a better mentor, or someone more dedicated to excellent research. Your lab has been such a welcoming place, even before I really earned it, I've always felt as though my input and ideas have been valued, which has always meant a huge amount to me. I've had the chance to learn so much in my time here that I can hardly keep track of it all. Thank you for providing such a wonderful environment, and for providing both the space to go out and learn so many things and help when I needed it. I feel extremely lucky to have worked in your lab these last five years.

Every member of the Barbar lab (including former ones) makes it a wonderful place to work. Nathan, your cheery disposition was, and I assume still is, infectious, and working with you was a treat. Heather, thank you for all your help in the lab, but especially your guidance when I first joined, showing me the ropes. I still laugh about all the trouble I must have caused you when I first joined, emailing you questions about work while you were recovering from a concussion. Thank you for also welcoming me so warmly as a friend and roommate, and for not getting irritated at the sight of me after seeing me for 10 hours a day every day for a year and a half. Kayla, like you said, we've always had such an advantage. I've gained so much from your presence, both professionally as our great organizer and personally as a friend. Thank you for willing to always be a foil to my sometimes admittedly ridiculous antics. We all miss you from the lab already, by the way, even if you're only downstairs. Jesse, thank you for bringing such a positive influence in the group, and for being the other person as interested in LC8 as I am. Zhen, I can divide my time in the lab between before and after you joined, where before everything was chaotic and uncertain, and after the lab runs like a well-oiled machine. Hannah, it's been wonderful to have you in the lab, and to see how far ahead you've made it already. I'd like to take at least a bit of credit for that, as your former TA. Douglas, it has been such a huge relief to have a second member of the lab who knows a bit about computers. Your kindness has made it so easy for me to become reliant on your advice, as well. All the undergraduates and rotation students in the lab have contributed to the wonderful environment as well, but I particularly want to mention Seth Pinckney; you and I joined the lab at the same time, and it was wonderful to have your help on my work.

Beyond the lab, the entire BB department has been a wonderful place to work, and I've thoroughly enjoyed my time here with everyone. I'm particularly grateful to the other

students of my year, Kayla, Amber, Ruben and Phil, it's been such a wild ride. I've been blessed with an unusually large number of collaborators across Oregon and the world, all of whom I am incredibly grateful for, but I particularly want to mention Dan Zuckerman and his lab, who welcomed me into their group as an honorary member while I learned the ropes of computational biophysics. Thank you for providing such a welcoming environment, even for someone who so clearly didn't know a thing.

I also want to thank my committee. Rick, thank you for teaching me the lab skills I still use every day. Andy, thank you for giving me a chance to work with you on a peroxiredoxin, the project was truly some of the most satisfying work I've done in grad school. Dan, thank you for teaching me so much about computational biophysics, and giving me so much of your time when I was a slow learner. Patrick, thank you for teaching me so much about NMR, and also for our countless conversations about my work over the years. Thank you, Claudia, for donating your time to my committee meetings as a GCR.

Lastly but certainly not least, I have to thank my family and friends for all their support. Mom and Dad, I'm sure you know, but I would not have been able to get any of this done without your help. Thank you for believing in me and cheering me on. I have so many more friends to thank than can reasonably fit in one acknowledgement, and I treasure all of you. In particular, I have to thank Heather, Elise and Dan for being wonderful roommates, and Chris and Ian for inspiring and driving me to get to grad school in the first place.

CONTRIBUTION OF AUTHORS

Chapter 2: Dan Zuckerman assisted in conceptualization of the statistical method used in modeling, as well as in the writing and editing process. August George assisted in both conceptualization and with coding of the methods. Chapter 3: Ylva Ivarsson and Cecilia Blikstated performed phage display experiments. Nathan Jespersen performed titration calorimetry experiments, peptide design, and contributed heavily to the analysis, writing and editing of the chapter. Anna Akhmanova and York-Christoph Ammon collected cell images used in the manuscript, Norman Davey contributed initial PSSM development, and David Hendrix assisted in the design of LC8Pred and of the LC8Hub website. Chapter 4: Anna Akhmanova and York-Christoph Ammon contributed in-cell colocalization assays and pulldowns. J. Helena Kinion synthesized peptides of Kank1 and tested peptide binding. Appendix 1: Patrick Reardon assisted with NMR data collection and analysis (in both appendix 1 and 2). Seth Pinckney expressed and purified proteins for the work. P. Andrew Karplus conceptualized the project and was closely involved in the writing process. Appendix 2: Heather M Forsythe performed NMR experiments with g1-1000 RNA, Zhen Yu expressed and purified proteins, and performed EMSA assays. Kaitlyn Hughes performed phase separation assays with g1-1000 RNA, Patrick Allen performed fluorescence anisotropy experiments. Elisar Barbar was closely involved in conceptualization, experimentation, and writing of every chapter.

TABLE OF CONTENTS

Page

1 Introduction..…… 1
LC8 – an essential hub protein……..…….2
Structure and function of LC8 binding……..…….3
The LC8 motif……..……..6
Multivalency in LC8 binding……..……..9
Perspective……..……12
Dissertation contents……..……...13
2: Quantifying Cooperative Multisite Binding through Bayesian Inference.....15
Abstract……..……16
Introduction……..……...16
Results……..……..20
Discussion……..……34
Methods……..……40
Acknowledgements……..……45
Supporting Information……..……46
3: Systematic Identification of Recognition Motifs for the Hub Protein LC8……54
Abstract……..……55
Introduction……..……...55
Results……..……..58
Discussion……..……67
Materials and Methods……..……70
Acknowledgements……..……74
Supporting Information……..……75
4: Multivalency Drives Binding between LC8 and the Cytoskeletal regulator Kank1 ……76
Abstract……77
Introduction……..……...77
Results……..……..79
Discussion……85
Materials and Methods……..……88

TABLE OF CONTENTS (continued)

Page

5: Conclusions……..……..……..91

Impact...…… ……..…….92

Highlights of reported work……..……...92

Ongoing work and future directions……..……..94

References……..……97

Appendices……..……..……114

1: Native State Fluctuations in a Peroxiredoxin Active Site Match Motions Needed for Catalysis……115

Summary……..…116

Introduction……...…116

Results…...……..…120

Discussion.……..…127

Methods....……..…133

Acknowledgements..……..…137

Supporting Information……..…138

2: Specificity and Heterogeneity in RNA-binding Domains of the Sars-Cov-2 Nucleocapsid Protein……144

Introduction……..145

Results…...……..147

Discussion.……..157

Materials and Methods……..……..159

Supporting Information……..162

3:Design and Characterization of a Synthetic Multivalent LC8-binding Protein……....……166

LIST OF FIGURES

Page

Figure 1.1: LC8 is an essential hub protein……2

Figure 1.2: Three binding modes of LC8……4

Figure 1.3: Structure and Thermodynamics of the LC8 motif……8

Figure 1.4: Multivalent LC8-client interactions……..11

Figure 2.1: LC8 binds clients through a two-step mechanism……19

Figure 2.2: Exact degeneracy in binding isotherms……20

Figure 2.3: Analysis of two-step model using synthetic isotherms……24

Figure 2.4: LC8 binding to a peptide from the protein SPAG5……..27

Figure 2.5: Example distributions for thermodynamic parameters from 3 LC8-peptide isotherms……..30

Figure 2.6: Binding between the intermediate chain (IC) and NudE……31

Figure 2.7: Phase diagram of width of posterior distributions as a function of model parameters……32

SI Figure 2.1: Example MCMC traces and marginal distributions for all model parameters……46

SI figure 2.2: Distributions of thermodynamic parameters plotted with total free energies and enthalpies……..47

SI Figure 2.3: Marginal distributions comparing models with uniform and normal-distribution priors……..48

SI Figure 2.4: Effect of concentration priors on marginal posterior distributions for thermodynamic parameters in a 1:1 binding model……48

SI Figure 2.5: Example marginal distributions of replicate models for the LC8-SPAG5 interaction……...49

SI Figure 2.6: Two-dimensional marginal distributions of enthalpy for selected isotherms……49

SI Figure 2.7: Marginal distributions of thermodynamic parameters for individual and global models for three LC8-peptide interactions……..50

SI Figure 2.8: Marginal distributions for thermodynamic parameters of IC-NudE binding isotherms……..51

Figure 3.1: Motif sequence logo and surface analysis of LC8……..57

LIST OF FIGURES (continued)

Page

Figure 3.2: Analysis of LC8-binding and nonbinding motifs reveals distinct positional preferences……..……..60

Figure 3.3: LC8 is structurally variable but conserved in sequence……..62

Figure 3.4: Generation and testing of the LC8Pred algorithm…….64

SI Figure 3.1: Optimization of matrix weights…….75

Figure 4.1: LC8 and Kank1 structure and function…….....79

Figure 4.2: LC8 colocalizes with Kank1 in HeLa cells……..80

Figure 4.3: LC8Pred predictions for the 500-800 region of Kank1……81

Figure 4.4: Isotherms for $Kank1_{595-720}$ and peptides from Kank1 motifs…….82

Figure 4.5: Characterization of the LC8-Kank complex……...84

LIST OF TABLES

Page

Table 2.1: Ranges for thermodynamic parameters for LC8-client binding……..…28

SI Table 2.1: Model priors and sampling lengths for all isotherms…….....52

SI Table 2.2: Credibility regions for ‘sum’ thermodynamic parameters and ratios of concentrations……...52

SI Table 2.3: Ranges of thermodynamic parameters for LC8-client binding when modeled with ±20% LC8 concentration……..…...53

LIST OF APPENDIX FIGURES

Page

Figure A1.1: Overview of *Xc*PrxQ catalysis and structure……119

Figure A1.2: Unusual chemical shift and environment of ser44 amide……121

Figure A1.3: Spin relaxation and modelfree analysis……122

Figure A1.4: Preexisting local unfolding equilibrium helices α2 and α3 revealed by CEST……124

Figure A1.5: Hydrogen exchange behavior of the *Xc*PrxQ C_P-thiolate and disulfide forms……126

Figure A1.6: Suboptimal interactions in the Prx active site and a possible trigger linking the unfolding of helices α2 and α3……130

SI Figure A1.1: Spin relaxation and {1H}-15N NOE values measured at 500 MHz……138

SI Figure A1.2: R_{ex} terms for disulfide and CP-thiolate *Xc*PrxQ and R_2 rates calculated using $R_{1\rho}$ for C_P-thiolate *Xc*PrxQ……139

SI Figure A1.3: Conformational heterogeneity seen in the crystal structure as a possible explanation for R_{ex} terms seen in C_P-thiolate *Xc*PrxQ……140

SI Figure A1.4: CEST profiles of C_P-thiolate *Xc*PrxQ……141

SI Figure A1.5: Comparison of hydrogen exchange in three thioredoxin fold proteins……142

Figure A2.1: Structure of the SARS-CoV-2 Nucleocapsid protein……146

Figure A2.2: Binding between g1-1000 RNA and CoV-N domains……148

Figure A2.3: Phase separation of CoV-N with g1-1000 RNA……149

Figure A2.4: EMSA assays of NTD with ss-14mer and ds-14mer……150

Figure A2.5: Interaction of CoV-N NTD with ss and dsRNA……152

Figure A2.6: Interaction of Y109A NTD with ss and dsRNA……155

Figure A2.7: Binding between the CTD and 14mer RNA……156

Figure A2.8: Fluorescence anisotropy measurements of CoV-N domains with RNA……157

SI Figure A2.1: Phase diagram of CoV-N interaction with g1-1000 RNA……162

SI Figure A2.2: Chemical shift perturbations of binding between the ss-14mer and the CoV-N NTD……163

SI Figure A2.3: Assignments of the Y109A NTD……164

LIST OF APPENDIX FIGURES (continued)

Page

SI Figure A2.4: Phase separation of the NTD with 14mer RNA……165

Figure A3.1: Sedimentation velocity analytical ultracentrifugation (AUC) of LC8 complexes……169

LIST OF APPENDIX TABLES

Page

SI Table A1.3: Model parameters and chi2 values for resides fit to model of exchange ……..……143

Specificity, Allostery, and Multivalency in Binding to the Hub Protein LC8

Chapter 1

Introduction

Interactions between proteins play an essential role across biological systems[1,2]. Nearly all biological processes are accomplished and/or regulated by multiple proteins working in concert[1–3]. As such, careful study of protein-protein interactions is key to understanding cell functions. Examination of networks of protein-protein interaction has revealed that while most proteins function through interaction with a handful of partners, some proteins act as hubs, with dozens or hundreds of interacting partners of diverse function[4–6]. The essential nature of such hubs is demonstrated by the impact of hub knockout mutants, which often exhibit several unconnected phenotypes[4,7]. For these reasons, Hub proteins are of particular interest in the study of protein interactions, sitting at the center of complex networks of function and regulation[5,6].

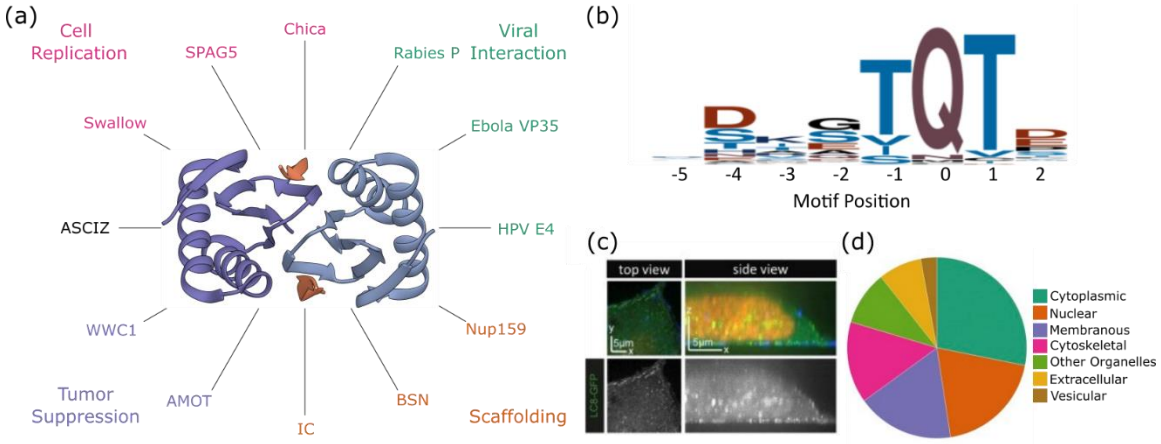


Figure 1.1: LC8 is an essential hub protein. (a) Structure of LC8 colored in shades of purple, with client strands colored orange. Selected LC8-binding proteins are listed in spokes around the diagram, colored corresponding to the functions listed at the edges of the panel. (b) Sequence alignment of known LC8-binding sequences taken from the database LC8Hub. (c) Live cell images containing LC8-GFP and (d) localization of known LC8-interacting proteins demonstrate that LC8 is localized throughout the cell. Panels (c) and (d) are adapted from Jespersen et. al. (2019)[8].

LC8 – an essential hub protein

LC8, a 20 kDa dimer, is one such hub protein[9]. While LC8 takes its name from its initial discovery as a light chain of the axonemal dynein complex, it has since been shown to bind over 100 clients with a wide variety of functions throughout the eukaryotic cell (Fig. 1.1c,d)[8]. All LC8 clients share a short linear motif sequence found in regions of intrinsic disorder that facilitate binding to the hub[8,10]. LC8-interacting proteins play roles in cell replication[10–12], tumor suppression[13,14], structural scaffolding[15–18], and viral infection[8,19,20] (Fig. 1.1a), among other functions[8,21–24]. With binding partners in such a wide variety of

cell functions, LC8 knockout mutations are fatal in several organisms and cell lines[25–27]. Consistent with this importance, LC8 is highly conserved across eukaryotes (with 94% sequence identity between drosophila and human)[9] and is present in all explored eukaryotic organisms, including plants, which lack all other dynein subunits[28].

Structure and function of LC8 binding

LC8 contains a mix of alpha and beta structures, forming two antiparallel beta sheets along its dimer interface, which are flanked by two alpha helices on each side (Fig. 1.1a)[9]. The primary binding grooves of the protein form at the edge of each beta sheet, one on either side of the protein, which accommodate disordered peptides approximately eight amino acids in length[29]. Bound clients take on an induced beta-strand structure, forming backbone hydrogen bonds with the LC8 beta strand at the dimer interface[9,29]. Investigation of sequence conservation has revealed that this binding groove is highly conserved relative to other elements of LC8 structure, pointing to client binding as the primary function of LC8[6,8]. The importance of the groove is additionally supported by the fact that all known LC8-binding interactions occur at this binding groove, making it an example of what has been referred to in the literature as a 'dynamic' or linear motif-binding hub, that accommodates many different clients through the same mechanism[6,9].

As a component of dynein, LC8 was first proposed to act as a cargo adaptor, attaching proteins to dynein for movement around the cell[9,30]. However, there has been little evidence in support of this hypothesis, and in 2008 our lab introduced a new paradigm for LC8 function – that of a dimerization hub[9]. As proposed, LC8 binds to client proteins such as the intermediate chain of dynein to dimerize them (Fig. 1.2a). Since 2008, a substantial body of evidence demonstrating that LC8-interacting proteins take on a dimeric structure has been collected, making clear the association between LC8 and client dimerization. Following is a brief discussion of several LC8-binding interactions studied in-depth over the last decade, where dimerization plays a key role.

The PICTS complex

The Panoramix-induced co-transcriptional silencing (PICTS) complex, consisting of proteins Panoramix, Nxf2 and Nxt1, is an important component of the PIWI-interacting RNA pathway, which silences transposon activity in animal gonads[31,32]. Panoramix contains two LC8-binding sequences within a region of disorder at its C-terminus[31,32].

Mutation of the LC8-binding sequences to abolish binding results in failure of the pathway, evidenced by an increase in transposon activity[31,32]. Crucially, replacement of the LC8-binding sequence with a leucine zipper results in a full rescue of transposon silencing[32]. The leucine zipper acts as a simple dimerization domain, and the zipper's effectiveness as a substitute for LC8 binding indicates that the primary role of LC8 in the PICTS complex is as an engine for dimerization.

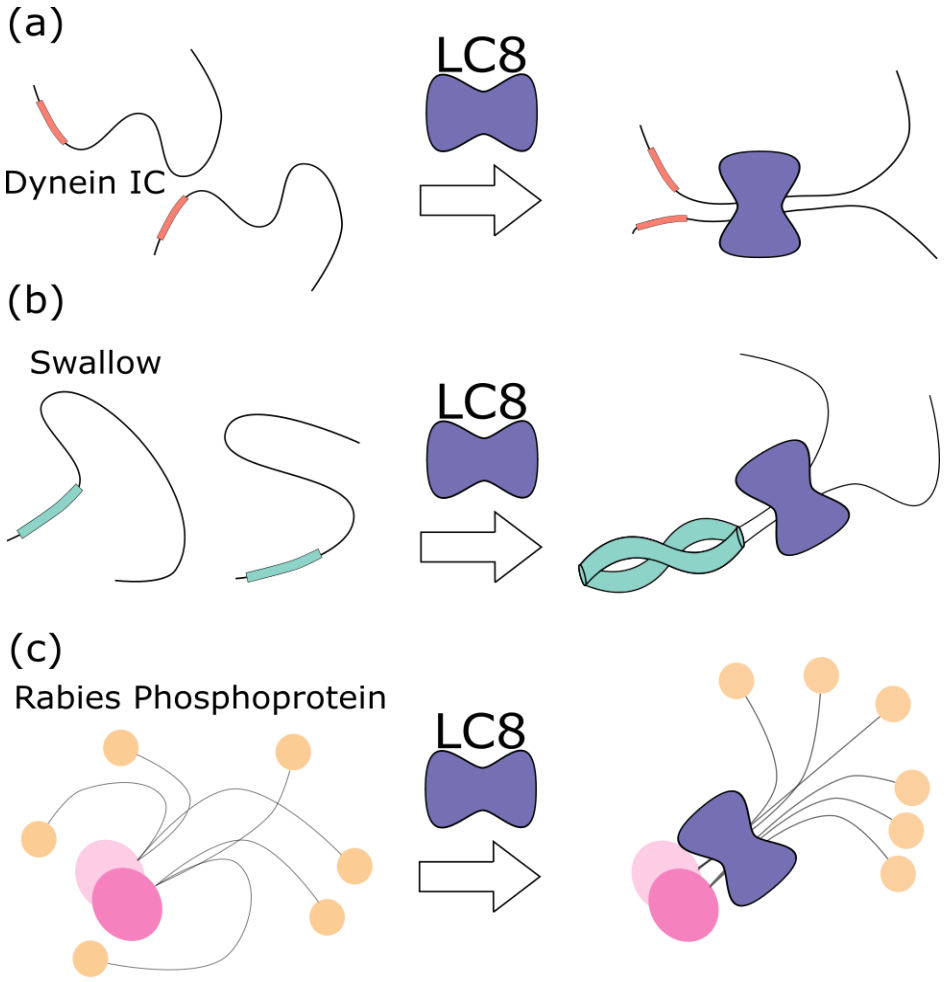


Figure 1.2: Three binding modes of LC8. Cartoon diagrams of the role of LC8 binding in 3 LC8-dependent systems. (a) LC8 dimer inducing a dimeric structure in the intermediate chain (IC) of Dynein. (b) LC8 dimer binding to Swallow, stabilizing a transient coiled-coil domain near the LC8-binding site. (c) LC8 dimer binding to the rabies phosphoprotein, which restricts the orientation of the C-terminal domain (yellow) and compacts the structure of the protein complex.

Swallow

Swallow is a predominantly disordered protein roughly 62 kDa in size that is essential for proper localization of mRNA in drosophila oocytes[11,12]. Essential to this function is swallow's only element of structure, a 71-residue predicted coiled-coil roughly in the middle of the protein sequence[12]. Experiments probing the coiled-coil domain of the protein reveal it to be only marginally stable, and prone to aggregation in vitro[12]. Addition of LC8 dramatically stabilizes the coiled-coil structure, resulting in a high-affinity, stable complex[12]. The domain can also be stabilized through mutation, resulting in a coiled-coil structure, confirming that LC8 is stabilizing an already-present dimer[12]. Corroborating the findings, biochemical investigations have revealed that LC8 is essential to swallow function, and loss of the LC8 motif from swallow results in mis-regulation of mRNA localization, eventually leading to embryonic defects[12,33]. The function of LC8 in this complex is therefore of a stabilizer, binding to an already-dimeric protein and inducing a tightly-bound complex, allowing swallow to perform downstream functions (Fig. 1.2b).

Rabies Virus Phosphoprotein

The rabies virus phosphoprotein (RavP) is one of five constituent proteins of the rabies virus and plays several roles in viral function, including regulation of viral transcription by connecting the nucleoprotein that wraps around the RNA to the RNA-dependent polymerase[19,34]. Roughly 300 residues in length, the protein forms a 70 kDa homodimer driven by a dimerization domain roughly 1/3 of the way through its sequence[19,19]. It additionally contains a folded C-terminal domain, separated from the dimerization domain by a long (~60 residue) linker containing an LC8-binding motif. Abolishing LC8 binding through mutation of the linker results in a dramatic drop in virus lethality, suggesting LC8 is critical for virus function[35]. While RavP is already a strong dimer in absence of LC8, binding to LC8 restricts the motion of the linker between dimerization domain and C-terminal domain, compacting the structural ensemble of the RavP and locking the relative orientation of the C-terminal domains (Fig. 1.2c)[19]. Downstream, this restricted conformation increases the rate of viral transcription, by increasing the rate that the protein 'walks' along RNA[19]. This indicates that LC8 effectively acts as a switch for viral transcription, where the transcription rate, a limiting step of viral replication, is maximized in the presence of LC8.

LC8 is a dimerization hub

While the details of the role that LC8 plays in each of these complexes varies from system to system, the common factor is the importance of a dimer structure in the complex. It appears that while LC8 does induce dimerization in some clients, in others it is merely adopting or stabilizing an already-present dimer structure. Indeed, a significant percentage of LC8-binding proteins contain coiled-coil or dimerization domains, in close sequence proximity to their LC8-binding motif[8,36]. It bears mentioning that several LC8-binding proteins, such as ANA2 or Ebolavirus VP35 are tetrameric, and remain tetrameric in the presence of LC8, indicating that LC8 complexes at other stoichiometries are possible in some cases[20,37]. Individual exceptions aside, it is now clear from available examples of LC8-binding proteins that the core function of LC8 is as a driver, stabilizer, and modifier of dimeric complexes containing regions of disorder.

Structurally, the dimerization hub theory has also been aided by detailed investigation into the structure and thermodynamics of LC8-client binding. Nuclear magnetic resonance (NMR) studies of LC8 indicate the presence of an intermediate LC8 state, bound at only one groove[38]. The intermediate state appears to have an increased affinity for client strands, resulting in LC8 favoring a fully occupied induced-dimeric state. Additionally, structural examination reveals this increased affinity may be driven by shear movement at the dimer interface and suggests that the effect may drive homologous binding – i.e., binding where LC8 binds the same client at both sites, thereby dimerizing the client[38]. The thermodynamics of LC8 binding have not been explored in detail, however, and a deeper investigation into LC8-client thermodynamics is the focus of chapter 2 of this work.

The LC8 Motif

LC8 binds to intrinsically disordered regions (IDRs) of client proteins which lack rigid structure. As is the case for many binding interactions within IDRs, the sequence alone is therefore the determinant of binding, and such binding sequences are referred to as short linear motifs[6,8,9] (SLiMs). The essential component of the LC8 linear motif is a TQT residue triad (Fig. 1.1b, 1.3a). All known LC8-binding proteins contain a TQT or TQT-like sequence[8,10]. Structurally conservative substitutions such as T->S,I,V and Q->N,M,L are tolerated without abolishing binding, although motifs only rarely contain more than one such substitution. Outside the TQT, five additional residues (one towards the C terminus

and four towards the N terminus) contact LC8 in bound structures. In contrast to the essential TQT, sequence alignments of known LC8-interacting proteins reveal no clear trend in the sidechains present at these other five motif positions[8] (referred to as 'flanking' positions hereafter) (Fig. 1.1b). Structural alignment of known LC8 motifs also reveals a great degree of structural heterogeneity at these positions (fig. 1.3b,c), suggesting the LC8 binding groove can accommodate a variety of structures at flanking positions. A 2016 examination of crystal structures of LC8 bound to clients revealed that, when bound, the flanking residues are relatively flexible, and the TQT relatively rigid[10]. This has led to the 'anchored flexibility' hypothesis of LC8 binding – the TQT anchors the client strand within LC8, and the remainder of the binding motif remains flexible and therefore capable of accommodating a variety of sequences[10]. Such sequence-permissivity at the flanking sites raises a question of function – is the binding grove agnostic to these flanking sequences, or do they play a functional role?

While the TQT anchor is essential for LC8 interaction, it is not the sole determinant of binding. A 2019 investigation of new LC8-binding sequences, published alongside work presented in this dissertation (chapter 3), utilized phage display experiments to generate potential new LC8-binding sequences[8]. A set of 53 sequences selected by phage display were tested for binding to LC8 by isothermal titration calorimetry (ITC). Surprisingly, while nearly all contained a TQT anchor, only 16 of the peptides bound to LC8 with a measurable affinity , indicating that the flanking sequences do play a significant role in determining whether LC8 binds a potential motif[8]. Succinctly, the TQT is essential but not sufficient for LC8 binding, a poor flanking sequence can eliminate LC8 binding even in the presence of a TQT. By extension, the binding affinity between a motif and LC8 is determined in part by the flanking sequence, as TQT-containing motifs can vary in affinity for LC8 over several orders of magnitude, from nanomolar affinities up to tens or hundreds of micromolar[8].

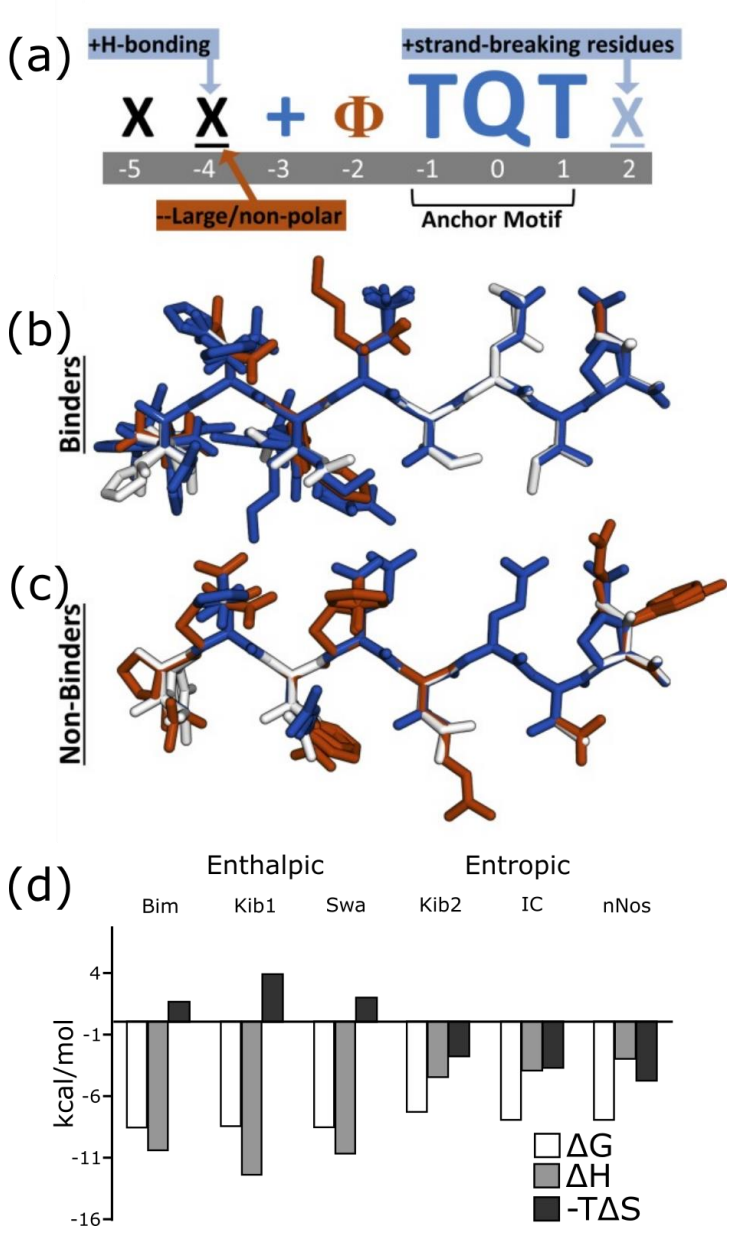


Figure 1.3: Structure and thermodynamics of the LC8 motif. (a) Diagram of preferred residues at each position in the LC8-binding motif. "Φ" denotes hydrophobic residues; "X" signifies any residue (unless certain residues are disfavored); underlined "X" signifies any residue but with strong preferences for specific residues; "+" denotes positively charged amino acids. Physiochemical properties beneficial for binding are colored dark blue or light blue, based on magnitude, and deleterious properties are colored in red. (b,c) overlay of modeled structure (assembled in Chimera) for tight (K_D < 10 μM) LC8-binding (b) and LC8-nonbinding (c) sequences. Residues are colored based on whether they are beneficial (blue) deleterious (red) or neutral (white) for binding. (d) Thermodynamics of several LC8-binding interactions, organized by the extent to which they are driven by enthalpy (left) or entropy (right). Panels a-c are adapted from Jespersen et. al., (2019)[8], and d from figures and data presented in Nyarko et. al., (2012)[39].

Investigation of LC8-client binding by ITC has revealed a substantial degree of variability in the thermodynamics of binding. Variations in the entropy of binding indicate that LC8-client binding is seemingly dictated in part by entropy-enthalpy compensation (Fig. 1.3d). While all LC8-binding interactions are enthalpically favorable, the entropic favorability varies – some interactions are driven entirely by enthalpy with an entropic cost (e.g. Bim, Swallow in Fig. 1.3d), while others are driven by both entropy and enthalpy (e.g. IC, nNos in Fig. 1.3d)[8,39]. In correlation with binding entropy, NMR analyses of LC8 find that the protein is flexible, in motion on the micro-millisecond timescale[39,40]. Binding to clients rigidifies LC8, eliminating these motions, but the degree of rigidification is not uniform across client sequences[39,40]. Particularly, the flexibility of client-bound LC8 correlates with the entropy of binding to a given client: Clients that are entropically favorable do not alter the motions of LC8 when bound, where enthalpically-driven LC8-binding sequences rigidify LC8[39]. As tested sequences all share an identical anchor sequence, it falls to the flanking sequences to be the source of this variation in flexibility in the bound state. The exact relationship between binding sequence and binding entropy remains unclear, however, as no obvious trends are present in the currently available data.

The flanking sequences, as the presumptive source of variation in both the overall binding affinity as well as the degree of LC8 rigidification, are therefore arguably the more important components of a given LC8 motif. Given LC8's role as a hub with many client proteins, it appears advantageous to have some variation in the binding affinity between LC8 and various clients. The flanking sequences offer a method for tuning LC8-client binding, varying individual interactions for the needs of individual systems. A substantial amount of work remains to fully understand the interplay between flanking sequences and LC8 affinity, and chapter 3 of this thesis presents a study of the LC8 motif, what flanking residues are preferred in LC8 binding, and development of a method for predicting whether a given sequence will bind to LC8.

Multivalency in LC8 Binding

The simple and flexible nature of linear motifs makes them ideally suited for multivalent recruitment of large complexes. LC8-binding sites are no exception, and a growing list of LC8-binding proteins with multiple motifs in a row has appeared in recent years[10,41]. Multivalency in binding introduces a substantial challenge to structural characterization of these complexes, as multivalency introduces a great degree of heterogeneity of both

structure and conformation, rendering many standard methods of structural investigation ineffective[41]. Nevertheless, multivalent LC8-client complexes play an important role in cell function. Following is a brief overview of several examples of such interactions.

Nup159

Multivalent binding appears to play a predominantly structural role in many complexes, as demonstrated in the complex between Dyn2 (LC8's ortholog in yeast) and Nup159. Nup159 contains five Dyn2 motifs within a span of ~120 residues towards its C terminus[15]. The five motifs, separated by short linkers, induce a unique ladder-like 'polybivalent' structure in Nup159 on binding[15,41]. Detailed thermodynamic investigation reveals that each motif appears to add to the overall affinity of binding between Dyn2 and Nup159, indicating that the motifs act cooperatively, in concert, to stabilize a high-occupancy state (Fig. 1.4a)[15]. Electron micrographs of the Nup82 subcomplex of the nuclear pore (which includes Nup159 and Dyn2) reveals a beads-on-a-string-like structure, where each Dyn2 appears as an individual bead[16]. While conformational heterogeneity complicates structural analysis of the complex[42], this bound state is rigid, relative to a disordered strand[15,16], indicating that the rigidity of Nup159 is controlled by the amount of Dyn2 present around it – In an apo state, the Dyn2-binding region of Nup159 is disordered and flexible. The presence of a switchable structured domain dependent on Dyn2 binding likely allows the protein to be either flexible or rigid when needed, a useful trait for Nup159's function as a structural scaffold.

Other LC8-interacting proteins that act as structural scaffolds in large protein complexes include RSP3[22], Bassoon[17], P53BP1[30], and the newly discovered Kank1, which is discussed in detail in chapter 4 of this work. These proteins have similar architecture to Nup159 – large regions of disorder, and a sequence containing multiple LC8-binding motifs separated by short linkers[10,36]. Their similarity to Nup159 makes it tempting to suggest that LC8-binding plays a similar structural role in these complexes, although the exact role of multivalent LC8 binding in these proteins remains to be seen.

Chica

The LC8-binding protein Chica is associated with the mitotic spindle, responsible for asymmetric localization of Dynein during cell division. This function is dependent on LC8, and knockdown of LC8 or mutation of Chica aimed at abolishing LC8 binding results in

incorrect Dynein localization[10,43]. The protein contains four motifs, three of which are confirmed to bind LC8, in a ~70 residue IDR towards the C terminus of the protein, between two structured domains[10]. While the exact function of LC8 in this complex is currently unknown, Chica presents an interesting contrast to Nup159 in its thermodynamics of binding. While each LC8 motif in Nup159 is relatively weak binding, they work cooperatively to bind tightly to LC8. Chica, in contrast, contains one tight-binding motif and several weak-binding ones[10]. In fact, the over-all binding affinity between LC8 and Chica is the same as the tight-binding individual motif (0.4 μM)[10]. This difference in affinity likely plays a role in Chica's function. Furthermore, it highlights that multivalent LC8-client interactions are surprisingly thermodynamically varied. This variation is expected to impact the occupancy state of such complexes, impacting downstream function.

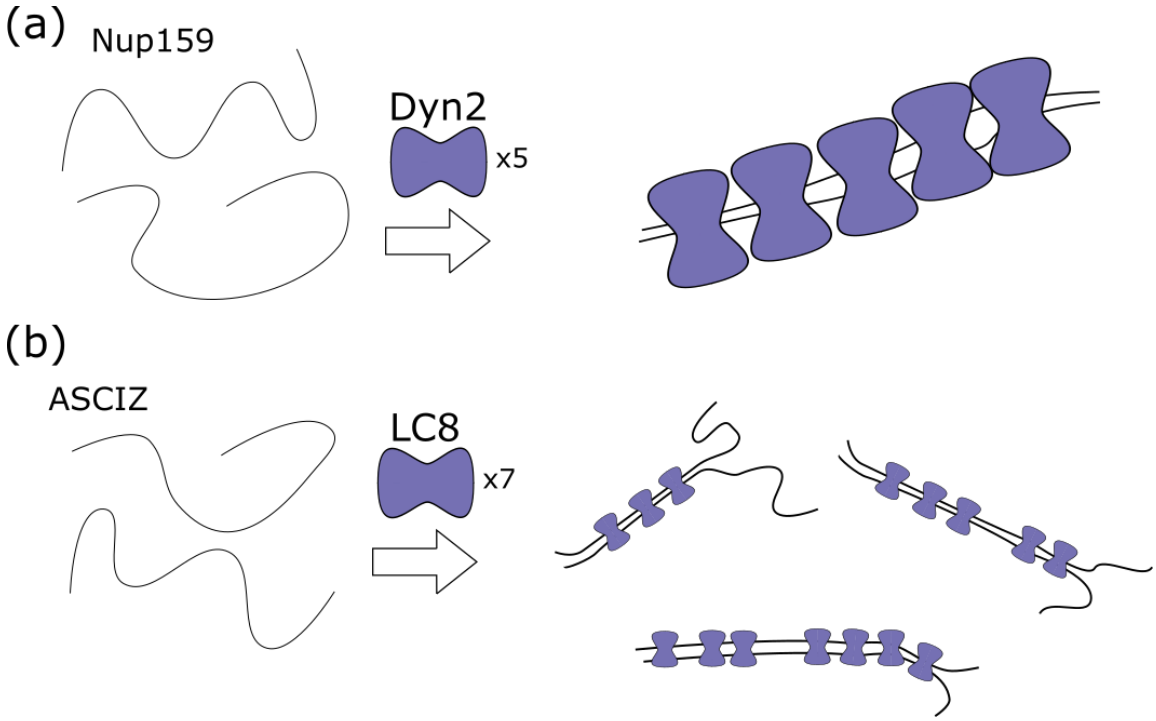


Figure 1.4: Multivalent LC8-client interaction. (a) Cartoon of binding between Nup159 and Dyn2. The interaction forms a rod-like complex, which forms part of a structural scaffold in the yeast nuclear pore. (b) Cartoon of binding between LC8 and the transcription factor ASCIZ. The proteins form a heterogeneous mix of different complexes, favoring a mix of varied occupancies over a rigid bound state.

ASCIZ

ASCIZ is an LC8-binding transcription factor responsible for regulating LC8 expression which also utilizes multivalent LC8 interaction to sense the cellular LC8 concentration[44,45]. ASCIZ consists of an N-terminal zinc finger domain, followed by a region of disorder that stretches to the end of the protein[44]. In drosophila, the IDR contains 7 recognizable LC8 motifs (>11 in human ASCIZ), spaced by linkers of variable length[44,45]. ASCIZ activates transcription of LC8, meaning it both binds LC8 and is responsible regulating the production of LC8[44]. In fact, ASCIZ acts as both a sensor and a regulator of LC8 concentration within the cell, inducing transcription when the concentration is low, and limiting it when the concentration is high. The multivalent nature of ASCIZ allows for a complexly tunable sensor for the level of LC8 present in the cell – the number of bound LC8 sites is believed to directly report on the presence or absence of LC8[44]. Distinct from both Nup159 and Chica, biophysical characterization of ASCIZ-LC8 binding reveals a mix of positive and negative cooperativity resulting in a heterogeneous mixture of several partially bound LC8-ASCIZ states (Fig. 1.4b)[44]. While the exact structural mechanism of this mix of different allosteric effects is not yet clear, it appears that the length of the linkers between LC8 motifs is a key ingredient of binding allostery[44,46]. As a result of this mix of effects, the LC8-ASCIZ complex remains deeply heterogeneous in both conformation and occupancy, even at saturating LC8 concentrations. It is therefore plausible that the negative allostery that strongly disfavors saturation of the ASCIZ-LC8 complex is tuned for the needs of the cell, which always requires at least a low level of LC8.

Perspective

The unique nature of each individual LC8-binding system complicates broad classification of LC8-binding proteins. Without detailed investigation, it is therefore difficult to predict the specific role of LC8 binding in each newly discovered client protein. The common factor of LC8 interaction is the bridging together of two client strands. LC8's general function can then be thought of as a molecular staple – in some cases sticking two disordered strands together into a dimeric structure, in others simply restricting the conformational flexibility of the client. And indeed, the current conception of the thermodynamic details of LC8 binding supports this theory. In many multivalent cases, a series of staples hold together two strands along a region of sequence, inducing the formation of a whole structured domain that can be turned on or off dependent on LC8 binding. A great number of

questions remain, however, and in this work, I present several investigations into details of this fascinating system. My work includes detailed thermodynamic analysis of LC8-client binding cooperativity using Bayesian statistical modeling (chapter 2), Examination of the LC8 motif and development of a method for predicting LC8 binding (chapter 3), and biophysical characterization of a newly discovered interaction between LC8 and Kank1 (chapter 4).

Dissertation Contents

This dissertation includes three chapters and two appendices of original work, each prepared in the style of a research manuscript. These include two published papers and three manuscripts in preparation or the review process. Chapter two, a manuscript currently under review at *Biophysical Journal,* covers the analysis of calorimetric data with a focus on binding between LC8 and clients. The manuscript uses Bayesian statistical modeling methods to determine thermodynamic parameters of binding interactions with a two-step mechanism and applies these methods to characterize the thermodynamics of LC8 binding in microscopic detail. In the process, the work examines the intrinsic determinability of multi-step binding parameters and the impact of uncertainty in analyte concentration on parameter determinability.

Chapter three consists of a portion of an article published in *Life Science Alliance* examining the LC8 motif. The chapter describes a comprehensive analysis of known LC8-binding motifs, with a focus on examining the flanking motif region, outside the TQT anchor. It additionally describes a method for predicting LC8-binding, that utilizes known sequences to predict a propensity for binding in new sequences. It additionally describes an online resource consisting of a database of LC8-interacting proteins, as well as a public form of the tool for LC8 interaction prediction.

Chapter four is a manuscript prepared for journal submission, focused on characterization of binding between LC8 and the tumor suppressor Kank1. The work demonstrates that Kank1 binds LC8 in cells, recruiting it to the cell cortex. Kank1 defies expectations, containing only a single predicted LC8-binding motif, but binding LC8 multivalently at seven sites. We demonstrate that the protein contains no strong-binding LC8 motifs, despite the full sequence binding tightly to LC8, indicating the complex is strongly driven by cooperativity in binding.

Lastly, the fifth chapter of this dissertation summarizes the important findings in chapters two through four, and the impact of the work on the study of LC8 interactions. It highlights the importance of studying the thermodynamics of complex binding interactions in detail, and discusses the unique challenges presented by multivalent LC8-binding interactions, as well as what is being done to begin tackling these challenges.

Two additional manuscripts of original work can be found in appendices 1 and 2. These consist of significant works completed during my PhD that bear no connection to my primary thesis project on LC8. The first appendix consists of a manuscript published in *Structure* on the solution dynamics of a protein from the peroxiredoxin family of redox proteins. Peroxiredoxins unfold locally as an essential component of their catalytic cycle, and we demonstrated that our model peroxiredoxin unfolds transiently even in absence of catalysis, emphasizing that the folding-unfolding equilibrium in these proteins is delicately tuned for the protein's function. The second appendix is a manuscript prepared for journal submission which discusses the Sars-CoV-2 Nucleocapsid protein, its structure, and binding between the protein and RNA. The protein drives the formation of viral particles through multivalent binding to RNA, and our investigation focuses on the structure of nucleocapsid-RNA interactions, examining the protein's preference for different RNA structures, and demonstrating that phase separation, which is connected to nucleocapsid formation in the virus, is driven by weak, nonspecific interactions between protein and RNA.

Chapter 2

Quantifying Cooperative Multisite Binding through Bayesian Inference

Aidan B Estelle, August George, Elisar J Barbar, Daniel M Zuckerman

In review at *Biophysical Journal*, July 2022
Also available as a preprint on BioRxiv:
https://doi.org/10.1101/2022.06.29.498022

Abstract

Multistep protein-protein interactions underlie most biological processes, but their characterization through methods such as isothermal titration calorimetry (ITC) is largely confined to simple models that provide little information on the intermediate, individual steps. We examine the hub protein LC8, which binds to disordered regions of 100+ client proteins in a wide range of stoichiometries. Despite evidence that LC8 binds clients cooperatively, prior ITC thermodynamic analyses have relied on models that do not accommodate allostery, and furthermore do not account for critical uncertainties in analyte concentrations. To characterize allostery in a more rigorous fashion, we build on existing Bayesian approaches to ITC to quantify thermodynamic parameters for multi-step binding interactions impacted by significant uncertainty in protein concentration. Notably, we account for a previously unrecognized intrinsic ambiguity in concentrations in standard binding models and clarify how this ambiguity impacts the extent to which binding parameters can be determined in cases of highly uncertain analyte concentrations. Our approach is applicable to a host of multi-step binding interactions, and we use it to investigate two systems. First, we deeply examine 2:2 LC8 binding and find it to be significantly positively cooperative with high confidence for multiple clients. Building on observations in the LC8 system, we develop a system-agnostic 'phase diagram' calculated from synthetic data demonstrating that certain binding parameters intrinsically inflate parameter uncertainty in ITC analysis, independent of experimental uncertainties. Second, we study 2:2 binding between the dynein intermediate chain and binding protein NudE, where in contrast, we find little evidence of allostery.

Introduction

Intracellular processes frequently depend on complex, multistep interactions between proteins or between protein and small-molecule ligands[3,48,49]. The hub protein LC8 provides an extreme example of binding complexity, accommodating over 100 client proteins via two symmetrical binding grooves[9,36] – often binding in multivalent fashion with a range of stoichiometries[8,10,15,41,44]. LC8 is found throughout the eukaryotic cell, and involved in a host of cell functions, with client proteins including transcription factors[8,44], tumor suppressors and oncogenes[13,50], viral proteins[19,51,52], and cytoskeletal proteins[10,18].

Structurally, LC8 forms a small 20 kDa homodimer (Fig. 2.1a), with two identical binding grooves formed at the dimer interface[9,36]. These binding sites induce a beta-strand

structure in a well-characterized linear motif anchored by a TQT amino acid sequence within disordered regions of client proteins[8,10]. Despite extensive studies[8,18,39], the mechanisms and thermodynamics of LC8 binding are still not fully understood, due to the difficulty of deconvoluting a multiplicity of microscopic states in its complex binding processes.

While usually fit to a simple model, LC8-client binding is likely impacted by allostery. The first evidence indicating allosteric behavior arose from nuclear magnetic resonance (NMR) titrations of peptides with LC8. A partially bound intermediate was detected half-way through titrations, with an estimated 2.5 to 6-fold higher affinity for the second binding step relative to the first[38]. Further evidence of allostery emerged from isothermal titration calorimetry (ITC) studies. Although ITC data are commonly fit to a simple *n*-independent sites binding model[53,54], this model is inadequate for a number of LC8-client systems that exhibit non-sigmoidal behavior, dipping slightly in heat per injection during early titration points instead of forming a flat plateau (Fig. 2.1b)[8]. This non-canonical behavior raises the possibility that these isotherms may fit well to a two-step model of binding, more representative of the expectation of dimeric LC8-client binding[55].

The use of ITC to interrogate complex systems and multi-step binding is challenging, as ITC data is of relatively low information, and individual isotherms often fit well to varied model parameters[55,56]. Despite this, well designed experiments can utilize ITC to measure cooperativity or allostery[57,58], entropy-enthalpy compensation[39,59], changes in protonation state[60,61], and competition between multiple ligands[62,63]. In general, these studies rely on fitting data globally to a model that includes several isotherms collected at varied conditions to reduce ambiguity of fit parameters[55,56], or a ‘divide and conquer’ type approach, where subsections of a complex binding network can be isolated and examined [18,57].

Concentration uncertainty is a critical concern in analysis of ITC data. In principle, accurate determination of protein and ligand concentration is a prerequisite for obtaining reliable thermodynamic quantities by ITC, yet these values are challenging if not impossible to obtain for many systems[54,64–66]. The most common software package for fitting ITC data, built in Origin 7.0 and distributed with calorimeters, attempts to account for this uncertainty in its simplest multi-ligand model through the stoichiometric parameter *n*, which can fit to non-integer values to correct for error in cell concentrations[53,67]. However, this implementation ignores uncertainty in concentration of the titrant in the

syringe, and is only applicable to the simple binding model, as complex binding models in Origin have no comparable correction factor. The popular and highly flexible fitting software SEDPHAT greatly improves on Origin's capabilities, allowing for both explicit or implicit (i.e. an 'inactive fraction' correction) uncertainty corrections[56,68]. As the authors note, however, allowing for variation in both analyte concentrations makes binding constants indeterminable within SEDPHAT due to correlative effects among model parameters.

Bayesian analysis offers a natural framework for incorporating uncertainty in concentration measurements in ITC analysis[54,69]. In a Bayesian framework, thermodynamic parameter determination is guided by a mix of experimental data and 'prior' information, such as uncertainty ranges/models, that weights the overall 'posterior' probability of a given set of thermodynamic parameters. The posterior distribution of estimated binding parameters generated through Bayesian analysis is a complete description of the probability range of each model parameter – and correlations among parameters – based on the input data and priors. With a meaningful prior description of concentration uncertainty, there is reduced risk of underestimating uncertainty in thermodynamic binding parameters.

We build on earlier applications of Bayesian inference to ITC. Nguyen et al. (2018)[54] studied 1:1 binding using a Bayesian statistical framework accounting for concentration uncertainty and performed sensitivity analysis on concentration priors. For a two-site binding model, Duvvuri et al. (2018)[70] demonstrated that a Bayesian method can accurately and precisely determine two separate affinities when applied as a global model to several isotherms, but the work assumes no uncertainty in measured concentrations[70], raising the possibility that parameter uncertainty is underestimated[54,56]. Cardoso et al. (2020)[69] used a simplified 4-step binding model with a single common binding enthalpy for a set of isotherms to determine 3 of 4 distinct affinities between protein and ligand, with the fourth being uncertain across a range of several orders of magnitude. Although Cardoso et al. (2020)[69] include concentrations as model parameters, they greatly narrow concentration priors using a preliminary 'calibration' assuming independent sites. We note that the *n*-independent sites model is not appropriate for complex systems, particularly in cases where the independent-sites model does not fit well to the isotherm shape. A sensitivity analysis regarding concentration uncertainty was not performed in

either multisite study, and neither work probed the information content of single isotherms for multisite systems.

Here, we report a Bayesian analysis of allosteric effects in two-site systems with a careful accounting of concentration effects critical for reliable analysis. We show that LC8-client interactions unambiguously exhibit positive allostery, driving binding towards a fully bound state. In contrast, symmetric two-site binding between the coiled coil domain of the dynein cargo adaptor NudE and the intermediate chain (IC) of dynein[71] shows no significant evidence for allostery.

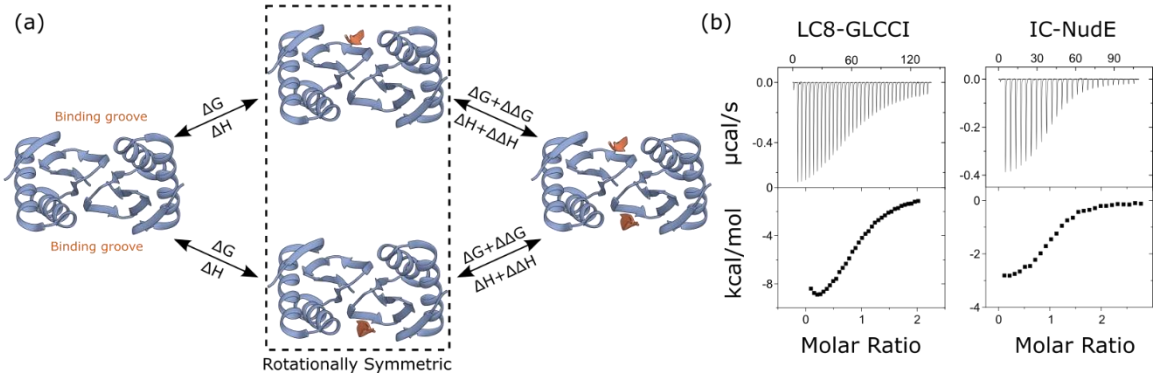


Figure 2.1: LC8 binds clients through a two-step mechanism. (a) Diagram of LC8-client binding, showing a structure of apo LC8 on the left, and a fully bound structure (PDB 3E2B) on the right. Intermediates are boxed to indicate they are symmetric and indistinguishable species. (b) example isotherms for binding between LC8 and client peptide taken from GLCCI (left) and binding between the intermediate chain (IC) and partner NudE (right).

We also provide methodological advances. First, we derive simple mathematical relations that govern the influence of concentration uncertainties on different binding parameters, providing a fundamental basis for the previously noted strong sensitivity of enthalpies – but not free energies – to concentration uncertainty[56]. Second, by using synthetic models, we systematically characterize the causes of binding-parameter uncertainties in two ways: we demonstrate that substantial uncertainty can result from the binding parameters themselves, e.g., strong vs. weak binding; and we also determine the effects of different prior functional forms and uncertainty ranges in a multisite context, extending the work of Nguyen et al. (2018)[54]. Finally, we outline best practices for determining model parameters and uncertainties in a multisite Bayesian framework.

Results

A mathematical "degeneracy" in thermodynamic parameters impacts analysis at any stoichiometry

We first present a simple mathematical analysis that explains previously reported correlation effects among titrant and titrand concentrations[56], and which significantly impacts the overall analysis of ITC data. Importantly, our analysis applies to monovalent or multivalent binding. Specifically, when the concentrations are uncertain, as is common in analysis of ITC data[54,56], we show below that only the *ratio* of titrant:titrand concentrations can be estimated, rather than the individual values, and this ambiguity propagates to all thermodynamic parameters. Hence, there is a "degeneracy" in that multiple solutions (sets of concentration values and thermodynamic parameters) will equally describe even idealized ITC data lacking experimental noise (Fig. 2.2).

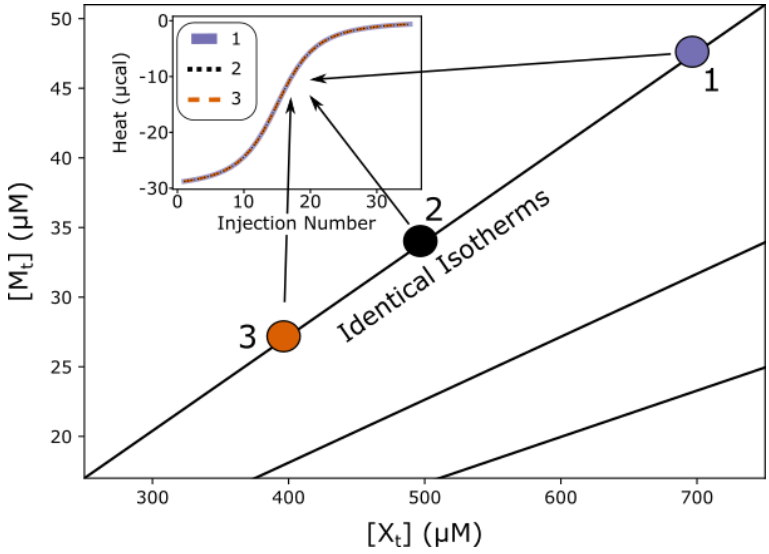


Figure 2.2: Exact degeneracy in binding isotherms. Based on the scaling relations of Eq (2), for any set of ligand and total macromolecule concentrations (X_t, M_t), there are infinitely many alternative concentrations (e.g., filled circles) on a diagonal line in the ([X_t], [M_t]) plane which yield exactly equivalent isotherms (inset) for a fixed set of thermodynamic parameters. For any given point in parameter space, equivalent degenerate lines can be drawn in a radial manner (e.g. the two additional solid black lines), passing through the point and the origin. The plotted synthetic isotherms are for 1:1 binding, but analogous degeneracy also holds for multivalent binding - see text.

We first describe the degeneracy for standard 1:1 binding between a macromolecule M and ligand X, following the scheme

$$\mathrm{M} + \mathrm{X} \rightleftharpoons \mathrm{MX} \tag{1}$$

The heat, Q, of a 1:1 binding system at any titration point can be described using the standard quadratic binding equation used in the independent sites model[47,53]:

$$\frac{Q}{V_0} = \frac{[M_t]\Delta H}{2}\left\{1 + \frac{[X_t]}{[M_t]} + \frac{K_d}{[M_t]} - \sqrt{\left(1 + \frac{[X_t]}{[M_t]} + \frac{K_d}{[M_t]}\right)^2 - \frac{4[X_t]}{[M_t]}}\right\} \tag{2}$$

where $[M_t]$ and $[X_t]$ are the concentrations of macromolecule and ligand (i.e., cell component and syringe component) respectively, while K_d and ΔH are the binding affinity and enthalpy.

The degeneracy is demonstrated by introducing a linear scaling of all parameters by an arbitrary number denoted α. Specifically, we apply the following transformations:

$$\begin{aligned} [M_t] &\longrightarrow \alpha[M_t] \\ [X_t] &\longrightarrow \alpha[X_t] \\ K_d &\longrightarrow \alpha K_d \\ \Delta H &\longrightarrow \frac{\Delta H}{\alpha} \end{aligned} \tag{3}$$

Applying this set of transformations, we can rewrite the binding equation:

$$\frac{Q}{V_0} = \frac{\alpha[M_t]\frac{\Delta H}{\alpha}}{2}\left\{1 + \frac{\alpha[X_t]}{\alpha[M_t]} + \frac{\alpha K_d}{\alpha[M_t]} - \sqrt{\left(1 + \frac{\alpha[X_t]}{\alpha[M_t]} + \frac{\alpha K_d}{\alpha[M_t]}\right)^2 - \frac{4\alpha[X_t]}{\alpha[M_t]}}\right\} \tag{4}$$

Regardless of the value of the factor α, all introduced factors cancel leaving Q unchanged.

Nearly identical considerations apply in the two-step binding model of primary interest here. As detailed in the methods, the value of Q is unchanged when both concentrations and both K_d values are multiplied by α and both ΔH values are divided by α. The underlying model is more complex as it requires solving a system of nonlinear equations (see Methods for details), but the result is that α is propagated through the nonlinear equation solutions, and once again cancels in the calculation of Q, leaving the heat value unchanged.

For reference, the corresponding concentration degeneracy scaling relations for 2:2 binding derived in Methods are as follows:

$$
\begin{aligned}
[M_t] &\longrightarrow \alpha[M_t] \\
[X_t] &\longrightarrow \alpha[X_t] \\
K_{d1} &\longrightarrow \alpha K_{d1} \\
\Delta G_1 &\longrightarrow \Delta G_1 + RT\log\alpha \\
K_{d2} &\longrightarrow \alpha K_{d2} \\
\Delta G_2 &\longrightarrow \Delta G_2 + RT\log\alpha \\
\Delta H_1 &\longrightarrow \frac{\Delta H_1}{\alpha} \\
\Delta H_2 &\longrightarrow \tfrac{\Delta H_2}{\alpha}
\end{aligned} \tag{5}
$$

To facilitate analysis and discussion of allostery below, from this point on we parameterize our model using ΔG, ΔΔG, ΔH and ΔΔH. The ΔΔG and ΔΔH value correspond to the differences between the first and second binding steps. Thus $K_{d1} = e^{\Delta G/RT}$, $K_{d2} = e^{(\Delta G+\Delta\Delta G)/RT}$, and $\Delta H_2 - \Delta H_1 = \Delta\Delta H$. The energy-like formulation allows for easy assessment of allostery, as ΔΔG is the free energy of allostery (which will be zero in the absence of allostery and positive or negative for negative or positive cooperativity, respectively), and ΔΔH is the change in enthalpy between binding steps with analogous characterization.

The degeneracy and associated scaling relationships in Eq (3) provide important insight into assessment of thermodynamic parameters inferred from ITC data. We see directly that binding enthalpy changes proportionately to concentrations of titrant and titrand. That is, a given percent error in an assumed concentration of either ligand (characterized by alpha) translates to the same scale of error in ΔH. On the other hand, the binding free energy ΔG, is less sensitive to concentration errors, due to scaling with ln(α), rather than directly multiplied by α.

The scaling relationships of Eq. (3) also presage a significant issue in Bayesian inference, namely, sensitivity to the choice of priors. Within the set of degenerate solutions (diagonal lines of concentration pairs in Fig. 2.2), the Bayesian ‘likelihood’ probability – which describes how well a parameter set fits the data in the absence of prior information – will be constant, as solutions are mathematically identical. Thus, within any degenerate set, the assumed prior distributions for concentrations, will determine the overall posterior distributions (see Methods). Because the posterior distributions ultimately determine the uncertainty ranges, this is a key point.

Below, we continue to examine the ramifications of the concentration degeneracy, demonstrating concretely that enthalpy is more impacted by uncertainty in concentrations than free energy. We also examine the influence of priors on parameter distributions, and discuss parameter distributions determined from isotherms in cases of high concentration uncertainty.

Validation of Bayesian inference pipeline with synthetic data

To test our Bayesian pipeline (Methods), we generated 'synthetic' simulated isotherms using hand-chosen sets of thermodynamic parameters ΔG, ΔΔG, ΔH, ΔΔH (see Fig. 2.1) inserted in Methods Eq (15) with added Gaussian noise. Following an exploration using synthetic data of how allostery impacts binding isotherms (e.g. Fig. 2.3a), we selected synthetic model parameters to mimic the isotherm shape seen in LC8-peptide binding examples. Specifically, slight positive allostery (ΔΔG = -1, ΔΔH = -1.5 kcal/mol) was best-suited to imitating real LC8-peptide isotherms, along with ΔG = -7 and ΔH = -10 kcal/mol. Synthetic noise is taken from a Gaussian distribution with a zero mean and standard deviation σ = 0.2 μcal. As shown in Fig. 2.3, we used our pipeline to sample posterior distributions for these isotherms. For concentrations, we chose uniform prior distributions of ±10% of the true value (which simply limits sampled concentration values to these ranges). The choice of 10% approximates what we view to be an attainable level of uncertainty for experimental protein concentrations.

Under these representative conditions, inferred posterior distributions fell around the known model parameters, and model parameters equate to isotherms which closely matched the isotherm shape (Fig. 2.3b,c). The finite widths of the distributions are due to synthetic experimental noise. The posterior distribution for ΔG covers a range of ~1 kcal/mol distributed around the true value of -7 kcal/mol. Examination of the distribution lets us define a 'credibility region,' that contains 95% of the distribution probability (i.e., from the 2.5 to 97.5%ile of the distribution), which is directly analogous to a frequentist confidence interval. For ΔG, the 95% credibility region is -7.5 to -6.4 kcal/mol. Similarly, the 95% credibility region for ΔΔG covers a range of ~1.5 kcal/mol, evenly distributed around -1 kcal/mol. ΔH and ΔΔH both have slightly wider credibility regions, with widths of 2.3 and 3.3 kcal/mol respectively, but both are distributed around the true values of -10 and -1.5 kcal/mol respectively.

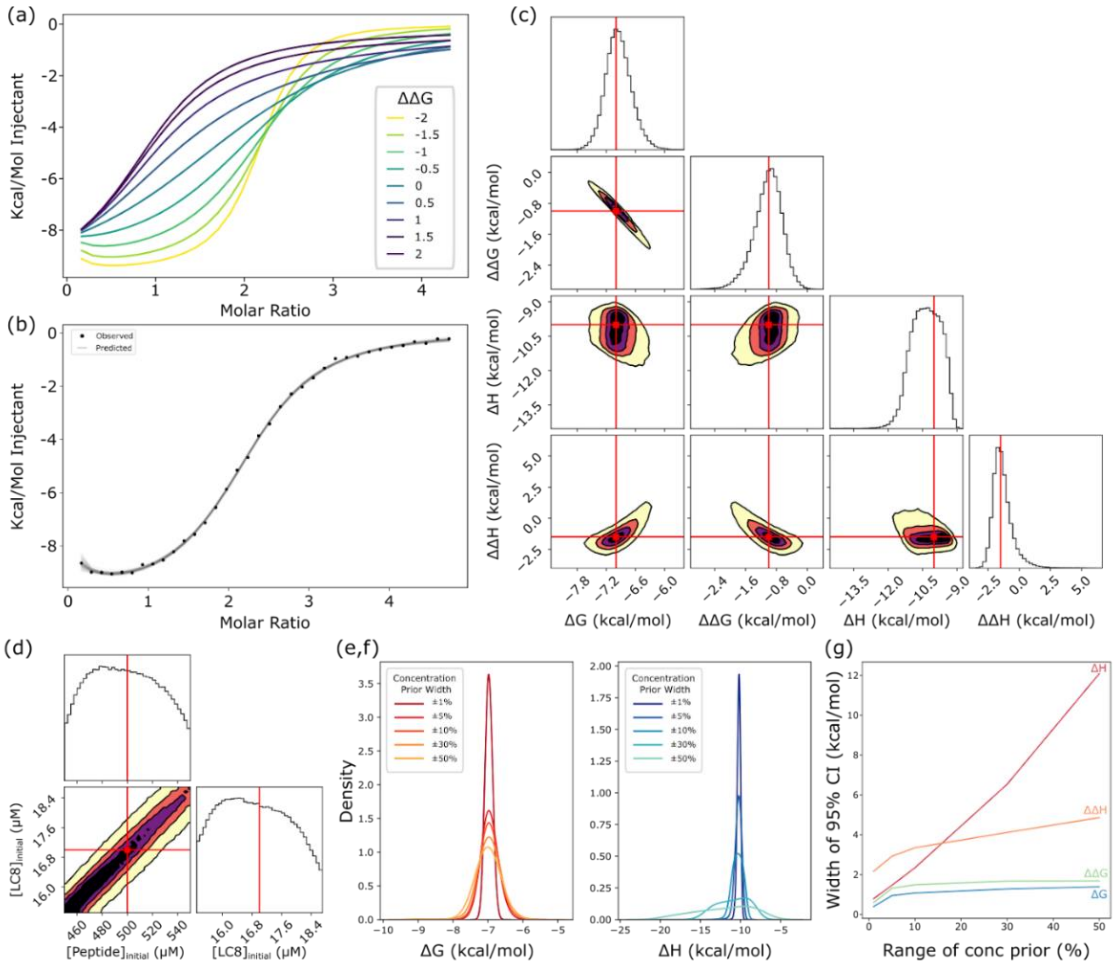


Figure 2.3: Analysis of two-step model using synthetic isotherms. (a) A set of synthetic isotherms for two-step binding with varied ΔΔG parameters demonstrating how allostery changes isotherm shape. Thermodynamic parameters are ΔG = -7, ΔH=-10, and ΔΔH=0. Concentrations are set at 17 and 500 μM for cell and syringe respectively, and injection volumes are 6 μL. (b) A synthetic isotherm with added gaussian noise (points) with 50 fitted isotherms (lines) generated through the Bayesian pipeline, i.e., sampled from the posterior. (c) One and two-dimensional marginal distributions for thermodynamic parameters, with contours in the two-dimensional plots set at 95 (yellow), 75 (orange), 50(purple) and 25%(black) confidence. Red lines and dots indicate true values for the synthetic isotherm. Marginal distributions, along with MCMC chains for all eight model parameters, including nuisance parameters can be found in SI Figure 2.1. (d) Marginal distributions for concentration parameters, exhibiting characteristic diagonal shape (Fig. 2.2) with contours as in (c). (e,f) One-dimensional distributions for ΔG (e) and ΔH (f) plotted for models with prior ranges for concentrations of 1, 5, 10, 30 and 50% of the stated concentration. (g) Width of the 95% Bayesian credibility region, akin to a confidence interval, for thermodynamic parameters as a function of the width of the concentration prior used in modeling, plotted from models with prior ranges for concentrations of +/-1, 5, 10, 30 and 50% of the stated concentration.

One benefit of Bayesian inference is the ability to examine multi-dimensional likelihood distributions to obtain correlations between model parameters. For example, in our two-dimensional distributions for the thermodynamic parameters, the ΔG and ΔΔG values are strongly negatively correlated (Fig. 2.3c), indicating a compensatory effect in the model, where increases in ΔG can be compensated by decreases in ΔΔG to arrive at similar solutions. Resultantly, the distribution for both ΔG and ΔΔG are broader than the 'total' free energy (i.e., $2\Delta G + \Delta\Delta G$), evidence that we can know the overall energy of binding more precisely than we can know the energy of each step (SI Fig. 2.2). Additionally, the mathematical degeneracy for concentrations described above can clearly be seen in these two-dimensional correlations: the two-dimensional marginal distribution for each concentration is a noise-broadened straight line covering the entire prior range (Fig. 2.3d). The scaling relationship of the model parameters outlined previously means that each point along this diagonal corresponds to a degenerate solution, i.e., each point has equivalent likelihood based on the data.

Impact of concentration degeneracy on two-site thermodynamic parameters assessed via synthetic data.

Bayesian inference enables determination of distributions for thermodynamic parameters even in cases of a concentration degeneracy. The net result, as will be seen, is a broadening of (posterior) parameter distributions based on multiple equally likely solutions, constrained by the priors used. Despite intrinsic limitations surrounding concentrations, the *ratio* of concentrations can be quantified with relatively high precision even when individual concentrations are highly uncertain.

To quantify the impacts of the concentration degeneracy within a Bayesian inference pipeline, we examined a series of uniform prior distributions for concentrations, ranging from ±1% to ±50% for both concentrations. These priors were applied to a synthetic isotherm mimicking experimental parameters, as described in the pipeline validation above. The choice of concentration priors – which embody assumed or estimated experimental uncertainties – greatly impacts the predicted uncertainty of thermodynamic parameters. The distributions for ΔG and ΔH, not surprisingly, both widen as the prior range is increased (Fig. 2.3e,f). As anticipated by the degeneracy scaling relations of Eq (3), the width of the distributions for ΔH and ΔΔH increases roughly linearly with the concentration prior range, while the distributions for ΔG and ΔΔG increase initially

at low concentration ranges then level off. This can be explained by the logarithmic relationship between the K_D (which is what scales with the degeneracy) and free energy. Functionally, high uncertainty in concentrations therefore only slightly increases uncertainty in binding free energy, while having a more significant impact on binding enthalpy.

The concentration degeneracy of the model limits the degree to which erroneously determined individual concentrations can be corrected. As discussed above, the fact that the Bayesian likelihood is equal at any point along the degeneracy lines (Fig. 2.2) means that the data have little impact on the posterior distributions for *individual* concentrations, which instead takes the shape of the prior used. This can be seen in the model validation example (Fig. 2.3d), where the posterior distribution is approximately uniform, echoing the uniform prior.

The ratio of concentrations (‘macromolecule’ to ‘ligand’), on the other hand, is a more meaningful parameter, and the quality of the ratio can improve a single uncertain concentration. For example, when we sample the posterior for the same isotherm, but use a normal (i.e. Gaussian) distribution for one concentration prior and a uniform distribution for the other, both posteriors take the shape of a normal distribution (SI Fig. 2.3). This is a direct result of the degeneracy identified above. SI Table 2.2 shows concentration ratio credibility regions for the experimental systems.

Because of the nearly determinative relationship between the prior and posterior concentration distributions, we elected to use uniform priors for concentrations throughout this work to avoid undue influence on our results from model priors. We believe varying the widths of uniform priors is the best way to probe concentration uncertainty effects.

For completeness, we also examined 1:1 binding with synthetic data. Overall, the impact of the concentration degeneracy on model parameters is similar (SI Fig. 2.4): binding enthalpy posterior distributions are wider than free energy distributions. In response to changes in concentration prior ranges, the posterior for ΔG is more impacted than in the two-step model, but the distribution remains much narrower than that of the enthalpy in every case.

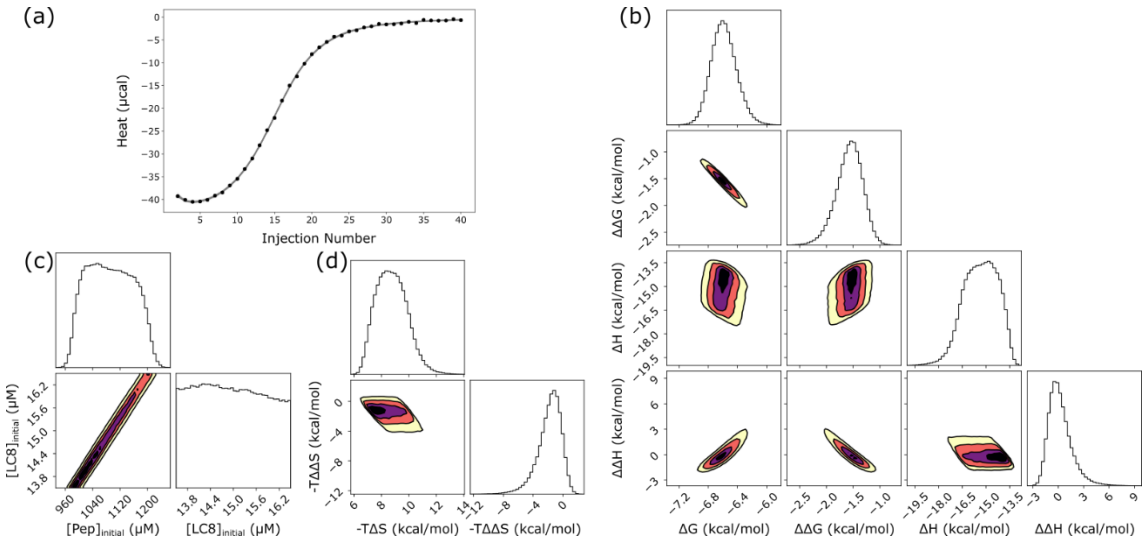


Figure 2.4: LC8 binding to a peptide from the protein SPAG5. (a) Experimental titration isotherm of SPAG5 into LC8 (points) with 50 example traces (lines) drawn from the posterior distribution of thermodynamic parameters and concentrations. (b) One and two-dimensional marginal distributions for thermodynamic parameters, with contours in the two-dimensional plots set at 95 (yellow), 75 (orange), 50(purple) and 25%(black) credibility. (c) Marginal distributions for concentrations of LC8 and peptide, showing a line of degenerate solutions, which may be compared to Fig. 2.2. (d) Marginal distributions for entropy (-TΔS) and entropy of allostery (-TΔΔS).

Application to 2:2 LC8:IDP Systems

We applied the Bayesian analysis pipeline to a set of 7 experimental isotherms of binding between LC8 and client peptides, all of which bind in a 2:2 ratio. Note that the two LC8's form a strong homodimer (Kd ~ 60 nM)[72] and this initial homodimer formation is excluded from our analysis. The systems were selected from a prior study[8] for tight binding and their deviation from the standard sigmoidal isotherm shape. As noted above, the user-supplied uncertainties for concentrations may impact uncertainty in other parameters. Following analysis with priors of ±10% and ±20% of the measured LC8 concentration as determined by absorbance at 280 nm, we have elected to focus on results at ±10% (Table 2.1), as moving to ±20% does not greatly alter the posterior distributions (SI. Table 2.3), and we believe ±10% to be an achievable uncertainty in protein concentration for most cases. The high degree of purity (>95%) and high absorbance at 280 nm, due to the presence of 6 chromophores (1 Trp, 5 Tyr) allow for a high signal-to-noise ratio for the absorbance, reducing uncertainty in the measurement. Comparatively, because of the difficulty in accurately measuring concentration for peptides with few or no chromophores[65,73] (1 Tyr

residue for the peptides discussed here[8]), we used a prior of increased width for the peptide concentration, up to a limit of ±50% of the initially measured value estimated by absorbance at 280 nm. As discussed above, the posterior distributions processed through the Bayesian pipeline are limited by the most restrictive prior used, owing to the concentration ratio being well defined (SI. Table 2.2). As a result, this approach ensures that posterior distributions are limited to the range around the measured concentration of LC8, allowing us to effectively infer the uncertain peptide concentration.

Bayesian analysis of the seven systems reveals significant heterogeneity in the precision with which binding parameters can be determined (Table 2.1). As will be described in detail below, this is only partially reflective of apparent data quality (e.g., noise level). Instead, certain binding parameters, particularly binding enthalpies, are intrinsically more difficult to characterize. Variations in precision do not stem from inadequate sampling in the Bayesian pipeline: triplicate runs are performed to confirm sampling quality (see Methods) (example in SI Fig. 2.5).

Table 2.1: Ranges for thermodynamic parameters for LC8-client binding. Values delineate 95% Bayesian credibility regions from sampled posterior distributions, which are akin to 95% confidence intervals.

Peptide	ΔG		ΔΔG		ΔH		ΔΔH		-TΔS		-TΔΔS	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
SPAG5	-6.9	-6.2	-2.1	-1.1	-18	-14	-1.8	3.6	6.9	11	-5.7	0.6
BSN (I)	-5.6	-4.9	-2.1	-0.9	-37	-14	0.1	45	8.8	32	-51	-2.6
BSN (II)	-7.1	-6.5	-1.2	-0.4	-5.8	-4.8	-9.1	-7.0	-1.9	-1.0	6.1	8.3
SLC9A2	-6.8	-5.5	-2.6	-0.5	-24	-10	-5.0	23	3.7	18.3	-26	4.4
VP35	-7.4	-6.8	-1.7	-0.9	-14	-11	-0.7	1.5	4.0	6.7	-3.0	-0.2
GLCCI	-6.1	-5.1	-2.6	-1.0	-27	-9.7	-0.5	36	3.7	22	-39	-0.5
BIM	-9.5	-7.1	-2.0	0.8	-12	-10	-0.5	2.9	1.4	4.4	-0.7	2.2

In particularly tractable cases, such as for SPAG5 binding in Figure 2.4, the analysis provides marginal distributions of similar precision to those seen with synthetic data. For binding between a peptide from the protein SPAG5 and LC8, Bayesian analysis yields a 95% credibility region of -6.9 to -6.2 kcal/mol for ΔG (Table 2.1), equivalent to a range for K_{d1} of 8.7 μM to 27 μM. The 95% credibility region for ΔΔG, the allosteric difference between the first and second binding event, is -2.1 to -1.1 kcal/mol, roughly

equivalent to a 6 to 30-fold increase in affinity for the second binding step relative to the first. The change in binding enthalpy between first and second events, ΔΔH, is distributed around zero (Fig. 2.4c), with uncertainty >2 kcal/mol for all cases, meaning we are unable to discern conclusively if there is any allosteric change in enthalpy between binding steps. From ΔG and ΔH values for both binding steps, we can additionally calculate -TΔS and -TΔΔS, for the entropy of binding and the change in entropy across binding steps respectively. Although the marginal distributions for these terms are quite broad (Fig. 2.4d), the -TΔΔS mostly sits at negative values, indicating that binding allostery has a greater probability of being entropically driven. See Table 2.1 for the full set of credibility regions.

Some general conclusions about allostery are apparent from the full set of data (Table 2.1). In all cases except one (binding to BIM), the distribution for ΔΔG is negative, indicating that all isotherms exhibit some positive cooperativity. Even for BIM, which has the widest ΔΔG distribution, the range predominantly covers negative values. All isotherms exhibit precisely determined free energies: 95% credibility regions cover a range of 2 kcal/mol or less for all cases except BIM. A common feature among some isotherms, seen clearly in the 'fair' and 'poor' examples in Figure 2.5, is an apparent loss of precision in our ability to determine model enthalpies, as both show wide distributions for ΔH and ΔΔH. For these isotherms (e.g., SLC9A2, GLCCI, and BIM), the two-dimensional marginal distribution for ΔH and ΔΔH shows a clear correlative effect (SI Fig. 2.6), and the one-dimensional distribution for the 'total' enthalpy (i.e. $2\Delta H + \Delta\Delta H$) is narrower than the individual parameter distributions (SI Fig. 2.2). In sum, the wide enthalpy distributions represent an inability to precisely determine 'microscopic' enthalpies for individual binding events, even when the overall enthalpy can be determined with high precision.

Parameter inference from multiple isotherms

The use of additional experimental information is expected to increase the precision of parameter determination, and Bayesian inference is readily adapted to employ multiple isotherms, whether at matching or different experimental conditions[70]. For some systems where two isotherms were available, we therefore used a 'global' model that included both isotherms. Despite the higher dimensionality resulting from additional nuisance parameters (see Methods), we found it relatively easy to sample the parameter space for

a two-isotherm model (SI Fig. 2.7). In some cases, such as for GLCCI, the addition of a second isotherm usefully narrowed posterior distributions, while in others (e.g. BSN motif I) it proved less impactful, largely just taking the same shape as the distribution for individual isotherms. We note that the isotherms examined were designed as technical replicates, not as optimized isotherms at different conditions for a global model. We expect results on multiple isotherms with varied experimental setups, e.g., different concentrations, to be more consistently valuable. Nevertheless, the global models demonstrate our ability to apply the pipeline to multiple isotherms simultaneously, a key step toward improved precision going forward.

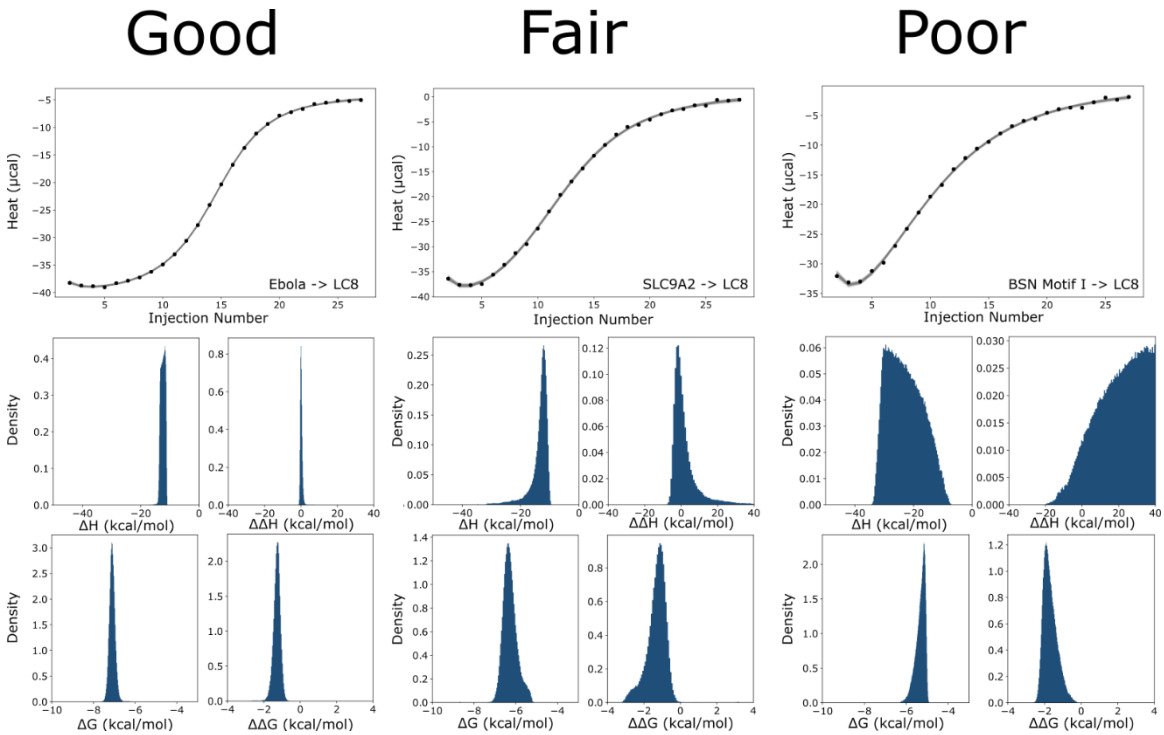


Figure 2.5: Example distributions for thermodynamic parameters from 3 LC8-peptide isotherms. Binding between LC8 and peptides from Ebola VP35 (left), SLC9A2 (middle) and motif 1 from BSN (right). Isotherms are shown at the top, and distributions for thermodynamic parameters are shown below. Horizontal axes represent the full width of the uniform prior range for each parameter to allow for direct comparison between each isotherm.

IC-NudE binding

To confirm the utility of the Bayesian pipeline for a range of systems, we tested it on binding of the intermediate chain of dynein (IC) to the non-dynein protein NudE. Binding

between IC and NudE can be described by the same model as binding between LC8 and clients – NudE forms a dimeric coiled-coil structure which then accommodates two strands of monomeric disordered IC for a 2:2 complex stoichiometry (Fig. 2.6a)[71]. Prior characterization of NudE-IC binding used a simple independent sites binding model without taking in consideration any binding allostery, and thus provides a good system for re-analysis as well as for comparison to LC8-client binding[71,74].

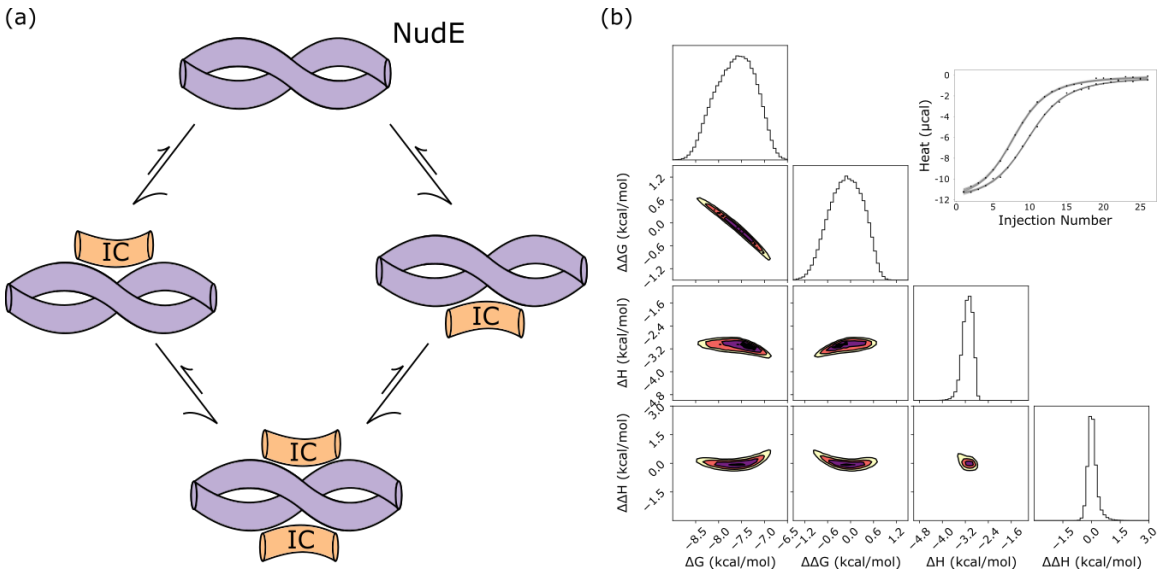


Figure 2.6: binding between the intermediate chain (IC) and NudE. (a) A model of IC-NudE binding, which forms a 2:2 complex, similar to what is seen in LC8. A cartoon diagram of NudE is shown in purple and IC in orange. (b) Sampled distributions modeled from two isotherms for binding between IC and NudE from yeast. Marginal distributions for thermodynamic parameters are shown on the left, and the top right corner contains the experimental isotherms (points) with model values (lines) drawn from the posterior.

For IC-NudE binding, a two-step model recapitulates the parameters determined in fits to independent sites modeling with little evidence of allostery. For high confidence in model parameters, we applied a global model, identical to the one used in LC8-client binding, to two titrations of IC into NudE. Bayesian sampling returns distributions that are narrowly dispersed for all thermodynamic parameters, both for individual-isotherm models (SI. Fig. 2.8), and for the global, 2-isotherm model (Fig. 2.6). Neither ΔΔG nor ΔΔH are significantly shifted from a distribution around zero, suggesting little, if any, allostery in binding. Simple models on these data indicate that binding has an enthalpy of -3.1 kcal/mol, and an affinity of 2.3 µM (i.e. a ΔG of -7.6 kcal/mol) implying a TΔS value of 4.5 kcal/mol, meaning binding is entropically favored[75]. Our two-step model predicts a ΔH

distribution centered near -3 kcal/mol, and a ΔG distribution centered near -7.5 kcal/mol, aligning well with the published values. This binding interaction works well as a counterexample to LC8-client binding: distributions for allosteric terms are centered around zero and determined distributions match closely to reported values modeled from an independent sites model.

'Phase diagram' analysis reveals weak binding affinities underlie loss of precision in binding enthalpies

We exploit synthetic isotherms to systematically survey binding parameters and determine the extent to which the physical parameters themselves intrinsically lead to lower precision in parameter inference. That is, for a fixed level of experimental noise, we quantify the widths of posterior marginal distributions and array the information in an interpretable 'phase diagram.' This effort was motivated by initial anecdotal observations that weaker binding was correlated with increased uncertainty, i.e., broader posterior marginals, in binding parameters, especially ΔH and ΔΔH. We created a series of synthetic isotherms on a grid of ΔG and ΔΔG values and determined posterior distributions for each isotherm. Two-dimensional plots of the width of these distributions (Fig. 2.7) as a function of ΔG and ΔΔG capture trends in our ability to determine model parameters.

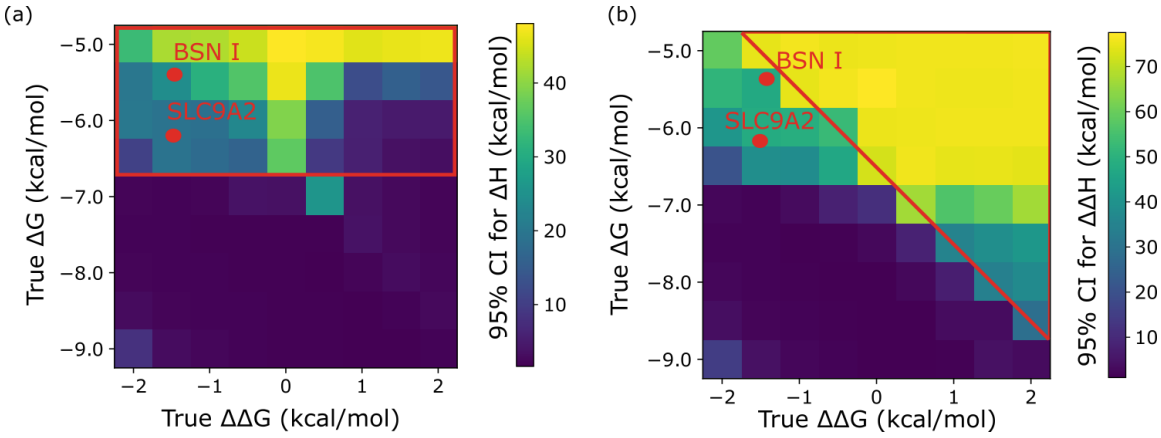


Figure 2.7: Phase diagram of width of posterior distributions as a function of model parameters. Two-dimensional plots along axes of ΔG and ΔΔG, wherein synthetic data with those parameters were generated, then sampled for posterior distributions at each point. Boxes are colored by the width of the 95% credibility region for ΔH (a) and ΔΔH (b), with lighter colors correspond to wider credibility regions (color bars). Red polygons demonstrate where each K_d (K_{d1} for left, K_{d2} for right) is greater than 17 μM, the cell concentration set for these synthetic isotherms. Red dots indicate mean values for experimental isotherms for binding for BSN motif I and SLC9A2

Generally, we lose precision in binding enthalpy in situations of weaker binding. Interestingly, the relationship appears to differ somewhat between ΔH (Fig. 2.7a) and ΔΔH (Fig. 2.7b). For ΔH, the primary dependence appears to be on the value of ΔG, with precision decreasing when ΔΔG is 0 or negative (top left corner of 7a). Conversely, the precision for ΔΔH appears dependent on both ΔG and ΔΔG, with the worst precision found in the top right quarter of the plot, where binding is weak and allostery is positive. In particular, precision reduces meaningfully in the red boxed areas shown in Fig. 2.7, where the affinity of each binding step is above the cell concentration (17 μM). This is consistent with our experimental results, wherein tighter-binding isotherms, higher magnitude ΔG value) such as Ebola VP35 and SPAG5 (Fig. 2.4, Fig. 2.5) perform better in terms of enthalpy determination than slightly weaker-binding isotherms such as SLC9A2 (Fig. 2.5), and much better than weakly-binding isotherms such as for BSN motif I (Fig. 2.5).

This overall loss in precision in cases of weaker binding is consistent with the concept of c in ITC experimental design[47,76], with additional considerations. The parameter $c = n[\mathrm{cell}]/K_d$, where n is binding stoichiometry, is a guide for setting experimental concentrations: ideally, c should be between 10 and 1000. For example, when n=1, the concentration in the cell should be 10-1000 times the K_d. In a multisite system, the exact relationship between each K_d and enthalpy determination is complicated by the existence of two binding constants, both of which may be outside of the relevant range and therefore limit precision. Conversely, model precision appears highest when both K_{d1} and K_{d2} are within the 10 to 1000 range for their respective c values. This neatly explains the steep drop-off in precision we see above -6.5 kcal/mol ΔG values for precision in ΔH, boxed in red in Figure 2.7. Similarly, when ΔG_2 (i.e. ΔG + ΔΔG) is above -6.5, the precision drops off steeply for ΔΔH, seen along a diagonal boxed in red in the figure.

For weak-binding isotherms such as BSN motif I, the lack of precision in enthalpy determination can be explained by a value for K_{d1} that is above the experimental concentration. For BSN motif I (location marked in Fig. 2.7, along with SLC9A2), ΔG is near -5 kcal/mol, compensated for by a negative ΔΔG value such that overall binding still appears relatively strong. Functionally, this means that the first-glance evaluation of binding affinity excluding allosteric considerations can be somewhat misleading, hiding the fact that ΔG (and therefore K_{d1}) is relatively weak. Even hidden in this fashion, the weakness of K_{d1} nonetheless hurts our ability to accurately determine binding enthalpies

and explains the variation we see in precision in enthalpy parameters throughout our tested data.

Discussion

This work examines the application of two-step binding models to isothermal titration calorimetry binding data, focusing on the hub protein LC8 and on accounting for critical uncertainties. LC8 binds over 100 client proteins in eukaryotic cells, and is involved in regulating a host of cell functions, motivating the detailed mechanistic study of LC8 binding. Using a Bayesian framework, we sought to determine precisely how much information can be extracted from a single isothermal titration (ITC) calorimetry isotherm, and examine how uncertainty in analyte concentration impacts model parameters, an investigation greatly aided by the use of simulated 'synthetic' isotherms with known parameters. Building on prior work[54,69,70], we have advanced Bayesian analysis of binding, and applied it to rigorous biophysical characterization of binding between LC8 and client peptides, as well as binding between the intermediate chain of dynein and the coiled-coil domain of NudE. We also used synthetic data to unambiguously separate effects of experimental error from intrinsic limitations, and we systematically surveyed the latter to generate a 'phase diagram' of intrinsic (in)tractability.

Allostery in LC8 binding

Our data show that LC8 can bind client proteins with significant positively cooperative allostery. Of the 7 peptides examined here, Bayesian analysis for all except one (BIM) yields a highly certain negative ΔΔG value, in agreement with early NMR studies that suggested positive allostery in LC8 binding[38]. Further investigation is required to determine whether such behavior is universal for LC8 client peptides. For the present study, we selected test isotherms with preference for two criteria we anticipated would leverage Bayesian modeling: (1) tight-binding to LC8 and (2) an isotherm shape that breaks from a strict sigmoid. Because this was neither a comprehensive nor random selection of systems, more work will be needed to determine conclusively whether LC8 binding is uniformly positively cooperative and whether the degree of allostery is sequence dependent.

Our findings are a step toward understanding the underlying biological function of LC8 allostery. While LC8-client complexes are varied, the putative functional unit of many LC8-client interactions is a 2:2 bound structure, where LC8 promotes dimerization in client proteins[11,21,77]. Further, In proteins where LC8-binding plays a structural role, such as at the nuclear pore in yeast[15,16], fully bound states driven by cooperativity are more likely to be highly rigidified. In both the case of 2:2 binding, and structural complexes, then, the functional state is promoted by cooperativity. We have proposed that positive allostery could drive the formation of homodimeric complexes[38], with the same client bound to each LC8 motif, and that allostery could effectively encourage homologous complexes while discouraging heterologous ones. In the future, ITC or nuclear magnetic resonance titrations of a second peptide into a bound LC8-peptide complex could be used to examine whether heterologous binding is indeed disfavored. The picture of LC8-client binding is additionally complicated by the discovery over the last decade of multivalent LC8-binding proteins such as the nucleoporin NUP159 or the transcription factor ASCIZ. Complexes between LC8 and multivalent clients are often highly heterogeneous in both stoichiometry and conformation[42,44,46], and particularly in the case of ASCIZ, the fully bound state is highly disfavored by some form of negative cooperativity, thought to be mediated through the linker sequences between LC8 motifs. This negative cooperativity ensures that ASCIZ is sensitive to LC8 even at high LC8 concentrations. The role that allostery of LC8 binding plays in these interactions is likely very complicated, and its relationship to effects seen in multivalent binding that rely on the length and structure of linkers between motifs[46,78], remains to be seen.

The mechanism of allostery appears to be entropically driven. While entropy is often the term with the widest distribution (Table 2.1), owing to its dependence on both the free energy and the enthalpy, there is a clear trend in our results towards positive $T\Delta\Delta S$ values, which equates to the second binding step being more entropically favorable than the first. Relatedly, NMR dynamics measurements indicate LC8's flexible core is rigidified on binding to clients[39,40]. Since LC8-binding allostery necessarily requires some change in the structural ensemble of LC8, it is possible that the first binding step can be thought of as 'paying up-front' for the entropic cost of both binding steps–i.e., rigidifying the whole LC8 core. This mechanism would also allow for variation in allostery on a per-peptide basis, as the degree of rigidification in the core seen by NMR is dependent on

client sequence[39]. Future molecular dynamics simulations can examine the differences in rigidity of the LC8 core in different bound states and across binding to different peptides.

Bayesian inference in binding analysis

"How much information is contained in an ITC isotherm?" is a fundamental biophysics question that Bayesian inference is uniquely suited to answer. Building on prior work[54,69,70], we have improved the ability of the Bayesian approach to account for the uncertainty that intrinsically occurs in *both* titrant and titrand concentrations. Our approach was motivated in large part by the apparently novel recognition of a mathematical "degeneracy" in ITC analysis, i.e., the existence of multiple solutions even in the absence of experimental noise, which prevents inference of a fully unique set of thermodynamic parameters. This degeneracy holds for simple 1:1 binding and apparently for arbitrary stoichiometry, as described in the Results.

While fitting ITC data to multi-step binding and other complex models is challenging, Bayesian inference allows for quantified "posterior" probability distributions for model parameters, reducing the risk of overfitting a complex model to insufficient data. These posterior distributions – or more accurately, the joint distribution over all binding parameters – fundamentally answer the question of the information contained in an ITC isotherm[54,70]. Bayesian inference is particularly powerful both for handling the degenerate nature of binding models that account for concentrations, and for experiments like ITC, which are very 'low-information' by nature[55,56].

The Bayesian approach offers several advantages over frequentist fitting methods[54,56] that are particularly useful in the case of ITC-measured binding: (1) Bayesian inference is not hampered by correlative or degenerate model solutions, allowing for inclusion of concentration parameters, (2) inferred distributions offer insight into correlative relationships in model parameters, and (3) the Bayesian model is highly flexible, and allows for incorporation of additional experimental data, whether additional isotherms or through the implementation of a variety of prior distributions.

Our investigation of how concentrations impact model parameters has shown that, as expected from the correlative relationships of the model parameters, that uncertainty in concentration significantly increases uncertainty in binding enthalpy, and has a reduced impact on free energy. This agrees well with Nguyen et al. (2018) who reported results on 1:1 binding, suggesting the concentration-enthalpy relationship is likely to be generic to all

binding models. We have shown that while the individual concentrations may be indeterminable from the model alone, the ratio of concentrations can be readily determined, provided the underlying stoichiometry of binding is known (SI Table 2.3).

Our primary goal has been to quantify uncertainty as completely as possible in determination of thermodynamic binding parameters. From a single isotherm, we sample marginal posterior distributions with widths on the scale of 1-2 kcal/mol for a two-step model of binding with four thermodynamic parameters, consistent with prior Bayesian analysis[69,70]. Although this is much higher uncertainty than the fractions of a kcal/mol usually reported in the analysis of ITC data[57,58], the difference can be explained to a great extent by the complexity of the two-step model, which intrinsically includes other correlative effects, e.g., between ΔG and ΔΔG, which are not accounted for in frequentist fitting methods. Additional uncertainty, beyond what can be attributed to the two-step binding model, arises from our 'skeptical' consideration of analyte concentrations, modeled by realistically wide concentration priors (± 10% for LC8, up to ±50% for peptides) that contribute to uncertainty in determined parameters. While 'microscopic' free energy and enthalpy parameters for individual binding steps cannot always be determined with good precision, the total values accounting for both steps show improved precision (SI Table 2.3, SI Fig. 2.2).

We believe that, in cases where higher precision is required for binding parameters, uncertainty can be decreased through the use of careful concentration determination through multiple methods, and the use of global models derived from multiple isotherms at varied concentrations.

Synthetic datasets guide experimentation

Our investigation has benefited greatly from the use of synthetic isotherms. Built from known thermodynamic parameters, and modeled using our Bayesian pipeline, the value of synthetic isotherms as an aid in experimental design is well-established[54,56]. They are particularly valuable in cases of complex binding, where it is not necessarily clear how determinable model parameters are, such as in our case. Synthetic isotherms have allowed us to test and troubleshoot our pipeline (Fig. 2.3), probe the information content of isotherms under variable conditions of concentration and priors (Fig. 2.3, SI Fig. 2.4), and examine how thermodynamic parameters themselves impact our ability to determine information from isotherms, resulting in the 'phase diagram' of relative tractability (Fig.

2.7). In the context of multi-isotherm modeling, we believe that utilizing synthetic data to design new experiments, such as is done with frequentist fitting in the program SEDPHAT[56,68] will be particularly valuable.

Practical limitations of Bayesian sampling and global modeling

Bayesian statistical analysis is much more computationally expensive than frequentist fitting methods. It usually relies on Markov chain Monte Carlo (MCMC) sampling, which requires simulating a sufficient number of steps to adequately explore the parameter space, potentially including a need to locate and sample multiple probability peaks (akin to energy basins in conformation space). For our ITC model, simple MCMC sampling methods proved unable to adequately sample the model space, even following sampling times of several days and over 4 million samples. While the ensemble sampler[79] used by us and others applying Bayesian models to ITC[69,70] has been robust for our purposes, sampling continues to be an important consideration, especially when considering future study of more complex models. For all work presented here, wall-clock sampling times were on the scale of hours, and hence readily feasible. More complex models could require significantly more sampling, although there is no simple scaling law that applies because of the uncertain nature of the parameter-space 'landscape'. Global modeling of multiple isotherms may also require additional sampling: as additional isotherms are added to a global model, each one brings with it a new set of nuisance parameters (4 per isotherm in our work - see Methods). In our hands, global models of two isotherms could be well-sampled within half a day. While global models of technical replicates may improve signal to noise ratios, ideally, global experiments should be designed with the intent of covering several experimental conditions[69,70], and all experiments must be high quality to ensure they contribute to global fits.

As an aid to investigators employing Bayesian inference in future studies, we have developed a set of guiding best practices (see manuscript).

Concluding remarks and future steps

Bayesian inference has allowed us to characterize the binding and allostery with high confidence for two different protein-protein interactions of 2:2 stoichiometry, despite meaningful uncertainties in analyte concentrations and inherent limitations of isothermal titration calorimetry. Our analysis was enabled by improvements to prior work [54,69,70] in

treating concentration uncertainties, and further demonstrates the value of Bayesian inference to ITC analysis. We used synthetic data to systematically characterize the uncertainty landscape for 2:2 binding based on intrinsic binding properties, an approach that readily can be extended to other models.

We examined two multi-step binding systems, the hub protein LC8 and the dynein intermediate chain (IC), For LC8, every client peptide studied showed evidence of allostery, corroborating hypotheses from a decade ago[38], and thus serving as an important step toward quantitative characterization of more complex LC8-client complexes. In contrast, the dynein IC/NudE complex showed minimal evidence of allostery.

While our focus here has been on two-step symmetric-site binding systems, Bayesian methods can be applied to other complex models investigated by ITC. Measurement of complex multivalent systems, enthalpy-entropy compensation, and ternary complexes or competition binding are all likely to benefit from analysis under a Bayesian framework. Although there is a limit on how much information can be gained from individual isotherms, investigation utilizing synthetic data can guide design, to help determine experimental conditions that maximize gain from additional ITC experiments within a given system.

Methods

Binding Models - 1:1 binding

For a 1:1 binding interactions between some macromolecule M, and ligand X:

$$\mathrm{M} + \mathrm{X} \rightleftharpoons \mathrm{MX} \quad (1)$$

The energy of binding is described by the following quadratic equation[47]

$$\frac{Q}{V_0} = \frac{[M_t]\Delta H}{2}\left\{1 + \frac{[X_t]}{[M_t]} + \frac{K_d}{[M_t]} - \sqrt{\left(1 + \frac{[X_t]}{[M_t]} + \frac{K_d}{[M_t]}\right)^2 - \frac{4[X_t]}{[M_t]}}\right\} \quad (2)$$

$[M_t]$ and $[X_t]$ are the total concentrations of macromolecule and ligand after each injection, ΔH is the binding enthalpy, K_d is the binding affinity, V_0 is the volume of the cell, and Q is the heat of the system. This is directly equivalent to Origin's independent-sites model[53] when n=1. The observed measurement is dQ_i (i.e. the heat of injection i), which is calculated from Q using the following equation:

$$dQ_i = Q_i + \frac{V_i}{V_0}\left(\frac{Q_i + Q_{i-1}}{2}\right) - Q_{i-1} + \Delta H_0 \quad (3)$$

where V_i is the injection volume for injection i, to account for the change in volume associated with the injection. ΔH_0 is a correction term to account for heat of dilution and other effects that can shift, assumed constant over all injections in a given isotherm.

Two-step binding

Two-step binding is modeled in a standard fashion, such as in the binding polynomial model [80] as:

$$\mathrm{M} + \mathrm{X} \rightleftharpoons \mathrm{MX} + \mathrm{X} \rightleftharpoons \mathrm{MX_2} \quad (1)$$

Under this scheme, each binding affinity is as follows:

$$2K_{d1} = \frac{[X][M]}{[MX]}$$

$$\frac{1}{2}K_{d2} = \frac{[X][MX]}{[MX_2]} \quad (5,6)$$

where K_{d1} and K_{d2} are the affinities for the first and second binding step. Factors of 2 and ½ account for the existence of indistinguishable rotationally symmetric intermediates in our model. The total concentrations of X and M can be written as:

$$[M_t] = [M] + [MX] + [MX_2]$$
$$[X_t] = [X] + [MX] + 2[MX_2] \quad (7,8)$$

Through rearrangement and substitution of equations 5 and 6, the total concentration equations can be rewritten only in terms of [M] and [X], the concentrations of free macromolecule and ligand:

$$[M_t] = [M] + 2\frac{[X][M]}{K_{d1}} + \frac{[X]^2[M]}{K_{d1}K_{d2}}$$
$$[X_t] = [X] + 2\frac{[X][M]}{K_{d1}} + 2\frac{[X]^2[M]}{K_{d1}K_{d2}} \quad (9,10)$$

This system of equations is solved numerically for each given injection point to determine the unbound concentrations [M] and [X]. With both free concentrations determined, the system heat can be calculated:

$$\frac{Q}{V_0} = \Delta H_1[MX] + (\Delta H_1 + \Delta H_2)[MX_2] \quad (11)$$

where ΔH_1 and ΔH_2 are the enthalpies of binding step one and two respectively. The concentrations of each bound state can be calculated from [X] and [M] and equations 5 and 6. As in the 1:1 binding model, equation 3 is used to calculate the observed heat of injection, dQ, for each injection.

Degeneracy in two-step binding

When protein concentrations are included as model parameters, degenerate solutions are introduced. As outlined in the manuscript, the degeneracy is exposed from the following transformation:

$$[M_t] \longrightarrow \alpha[M_t]$$
$$[X_t] \longrightarrow \alpha[X_t]$$
$$K_{d1} \longrightarrow \alpha K_{d1}$$
$$\Delta G_1 \longrightarrow \Delta G_1 + RT\log\alpha$$
$$K_{d2} \longrightarrow \alpha K_{d2}$$

$$\Delta G_2 \longrightarrow \Delta G_2 + RT\log\alpha$$
$$\Delta H_1 \longrightarrow \frac{\Delta H_1}{\alpha}$$
$$\Delta H_2 \longrightarrow \frac{\Delta H_2}{\alpha}$$

(12)

Here, α can be any positive number. Following this transformation, the equations used to calculate [X] and [M] (eq. 5 and 6 in the methods) are transformed:

$$\alpha[M_t] = [M] + 2\frac{[X][M]}{\alpha K_{d1}} + \frac{[X]^2[M]}{\alpha^2 K_{d1}K_{d2}}$$
$$\alpha[X_t] = [X] + 2\frac{[X][M]}{\alpha K_{d1}} + 2\frac{[X]^2[M]}{\alpha^2 K_{d1}K_{d2}}$$

(13,14)

In these transformed concentration-sum equations, the new solutions for both [X] and [M] are exactly the previous solutions multiplied by α, as can be verified by substitution. Finally, applying the transformed values into the equation for Q yields

$$\frac{Q}{V_0} = 2\frac{\Delta H_1}{\alpha}\frac{\alpha[X]\alpha[M]}{\alpha K_{d1}} + \frac{(\Delta H_1 + \Delta H_2)}{\alpha}\frac{\alpha[M](\alpha[X])^2}{\alpha^2 K_{d1}K_{d2}}$$

(15)

As in the 1:1 binding model, cancellation of α shows there is no change in the value of Q for any α value. This demonstrates the degeneracy for 2:2 binding, which we can expect to generalize to higher stoichiometries.

Bayesian inference

Bayesian inference is a method to calculate a "posterior" *distribution* of model parameter values based on prior assumptions (encoded as prior distributions for parameters presumed to hold in the absence of data) and the data. In general, as more data is analyzed, the influence of the prior will decrease[81,82] . The posterior distribution of parameters provides rich information such as the parameter means and confidence intervals (technically "credibility regions"), in addition to correlation information regarding whether and how parameters vary together.

Bayesian inference is based on Bayes' rule[81,83] which enables us to infer a distribution of parameters θ (e.g., binding free energy and enthalpy, etc.) consistent with a given set of data D (e.g., ITC isotherms):

$$P(\theta|D) = P(D|\theta)\,P(\theta) / P(D) \quad (16)$$

where $P(\theta|D)$ is the (posterior) probability distribution of the model parameters, θ, given the data, D; $P(D|\theta)$ (the likelihood) is the probability distribution of the data given the model parameters and is given below; $P(\theta)$ (the prior) is the probability of the model parameters, specified below; and $P(D)$ (the evidence) is the probability of the data. For a given set of data, the unknown denominator $P(D)$ is constant, independent of parameters, so it does not affect the inference of posteriors. Typically, it is not possible to analytically solve Bayes' rule, so numerical methods such as Markov chain Monte Carlo are used to determine the target (posterior) distribution[84–86]. Details of our implementation are given below.

Bayesian model

Following prior work[65,70], we assume the data has Gaussian noise with a mean of zero and an unknown standard deviation. The ITC model parameters θ include concentration terms ($X_{initial}$,$M_{initial}$) and thermodynamic terms (ΔG, ΔΔG, ΔH, ΔΔH), as well as the nuisance parameters (ΔH_0 and σ) for heat of dilution and Gaussian noise. We use uniform prior distributions for the model parameters specified below and the unknown noise standard deviation unless otherwise stated. For global models (e.g. SI Fig. 2.7, Fig. 2.6), while it may be possible to assume a global noise or concentration model, we instead elected to apply global models with an additional set of concentration and nuisance parameters for each additional isotherm (bringing the total parameter count up to 12 for two-isotherm models). Uniform prior ranges for thermodynamic parameters were identical for all models, listed in SI Table 2.1. For nuisance parameters ΔH_0 and σ, uniform priors of -10 to 10 μcal and 0.001 to 1 μcal respectively were used in all models.

The likelihood for a set of data $D = \{x_1, x_2, \ldots\}$, denoted ($p(D|\theta)$), is the product of the probabilities at all data points x_i based on a normal distribution of standard deviation σ centered around $\mu_i(\theta)$, the calculated value of point *i* for the binding model and parameters θ. It therefore takes the following form:

$$p(D|\theta) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu_i)^2}{2\sigma^2}\right\} \quad (17)$$

and we note that σ is assumed unknown and sampled as part of the Bayesian inference process. When the priors are uniform, as we most often assume, the posterior is simply proportional to the likelihood given here.

Sampling

We use the affine-invariant Markov chain Monte Carlo sampling method[87] to perform Bayesian inference, as also used by Duvvuri et al. (2018) and Cardoso et al. (2020). The affine-invariant sampler is an ensemble-based method in which multiple walkers move through the sample space in a correlated fashion. We empirically found this method to sample significantly better than the standard Metropolis-Hastings [84,85] sampler for our model. In our hands, the Metropolis-Hastings method was unable to converge on the target distribution after 4,000,000 sampling steps, whereas the affine-invariant sampler was able to converge after 100,000 sampling steps.

Implementation

We used the EMCEE package[79] in Python to perform the affine sampling, using a 20%:80% mix of the “differential evolution” and “stretch” move sets with 25-50 walkers. For each experiment, 3 replicas are run for 50,000-200,000 sampling steps/replica until convergence. Each replica converged, as determined by the autocorrelation time, where sampled steps must be greater than 50x the autocorrelation. Convergence was additionally assessed through examination of posterior distributions from model replicas, which were nearly identical in all cases (SI Fig. 2.5). This implementation runs at ~9 samples for each walker per second on 4 cores of a node on the Oregon State College of Science computing cluster.

The code, data, and an example notebook are available at:
https://github.com/ZuckermanLab/Bayesian_ITC

Experimental ITC

All isothermal titration calorimetry experiments used here have been previously reported in other publications[8,75]. Briefly, LC8, IC and NudE were all expressed in BL21 or Rosetta cell lines, and purified to 95% purity using a combination of 6xHis TALON affinity purification and size exclusion chromatography. LC8-binding peptides were

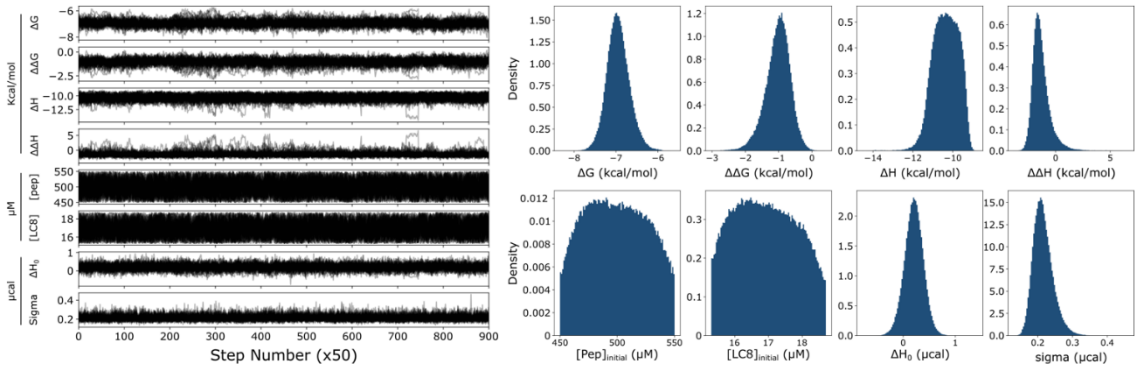
purchased from Genscript. Proteins were dialyzed prior to calorimetry into a buffer of 50 mM NaPO4, 50 mM NaCl, 5 mM β-mercaptoethanol and 1 mM NaN3, at pH 7.5. In the case of LC8-peptide binding, Peptides were dissolved into buffer following dialysis to ensure minimal buffer mismatch between peptide and protein. All ITC experiments were performed at 25 C, with an initial injection of 2 µL, which was discarded to account for the first injection anomaly. Peak integration was performed in Origin 7.0.

Synthetic ITC isotherms

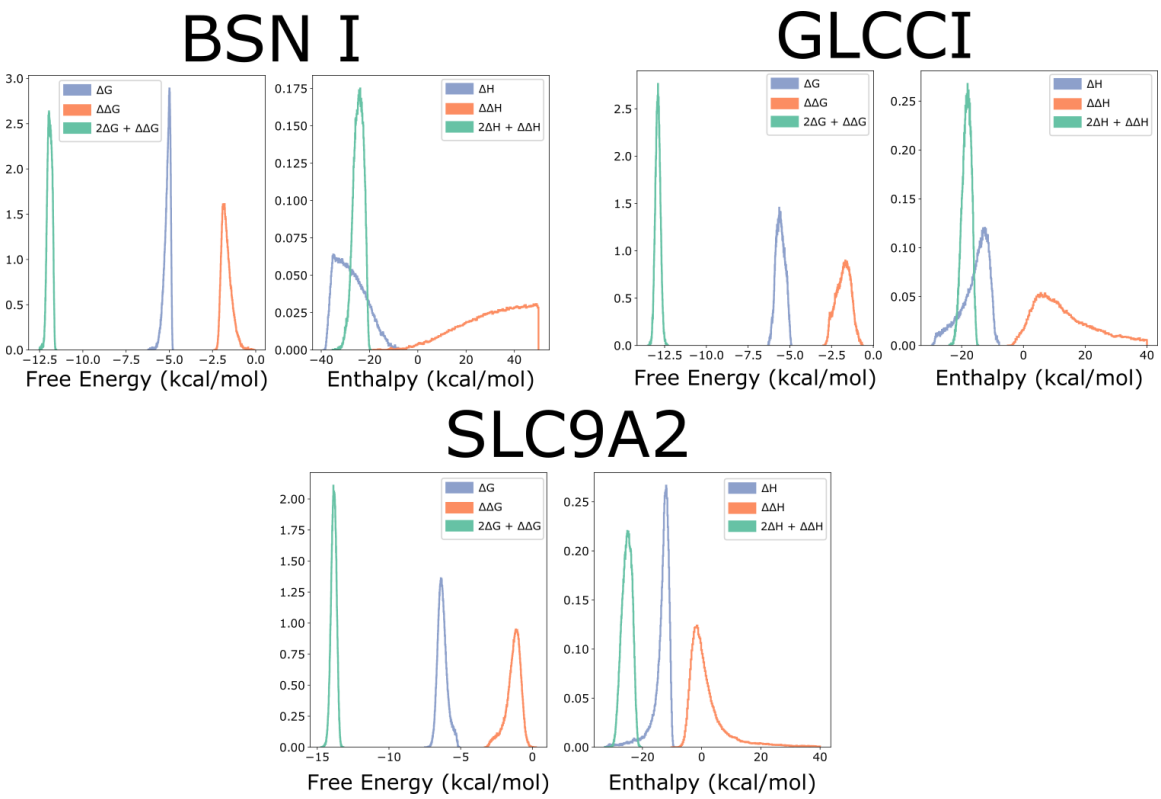
Synthetic isotherms for 1:1 and two-step binding were generated following equation 2 for 1:1 binding and equations 8,9 and 10 for two-step binding. Parameters were chosen to mimic typical experimental conditions employed in our group. For 1:1 binding (SI Fig. 2.4), we used ΔG and ΔH values of -12 and -8 respectively, and concentrations of 34 µM in the cell and 500 µM in the syringe. For two-step binding, varied thermodynamic parameters were used (e.g. Fig. 2.3, Fig. 2.7), but concentrations were fixed at 17 µM in the cell and 500 µM in the syringe. For all synthetic isotherms under both models, we simulated one injection of 2µL followed by 34 injections of 6 µL, with a ΔH_0 of 0 µcal, and added synthetic noise from a Gaussian distribution with standard deviation 0.2 µcal. To accurately replicate experimental conditions, we eliminated the first injection when applying models to this data.

Acknowledgements

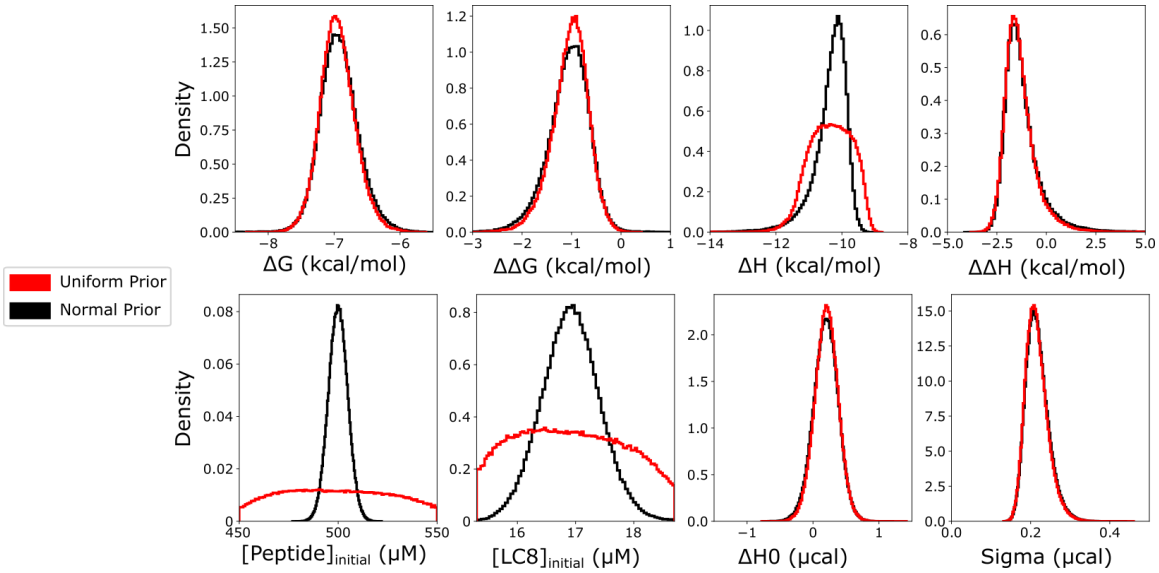
This work was supported by the U.S. National Institutes of Health grant R01-GM141733 and by U.S. National Science Foundation Grant MCB 2119837. We additionally acknowledge the support in the form of computational resources of the Oregon State University NMR Facility funded in part by the National Institutes of Health, HEI Grant 1S10OD018518, and by the M. J. Murdock Charitable Trust grant #2014162. We appreciate helpful discussions with John Chodera, David Minh, and Trung Hai Nguyen.



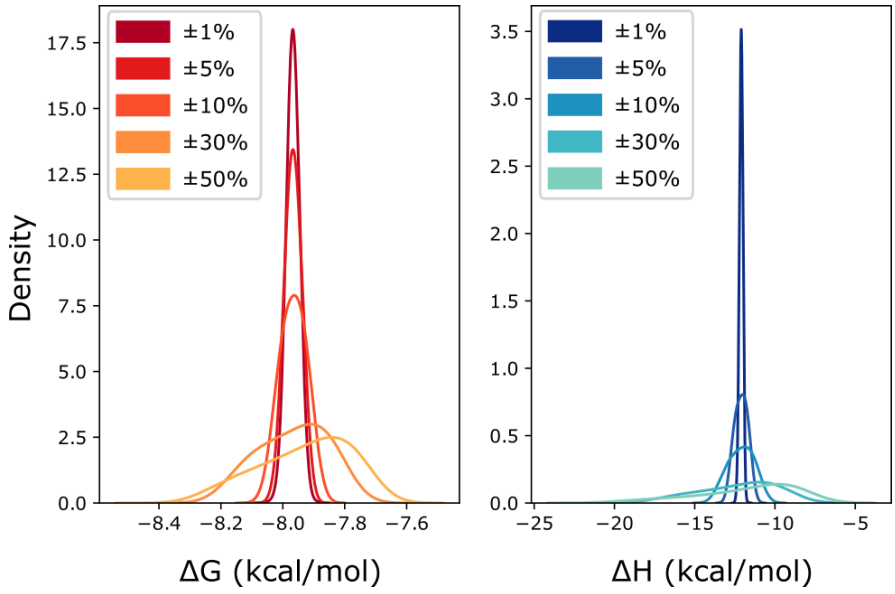
SI Figure 2.1: example MCMC traces and marginal distribution for all model parameters. MCMC traces and distributions are drawn from model on synthetic isotherm from Fig. 2.3, generated from parameters ΔG = -7, $\Delta\Delta G$ = -1, ΔH = -10, $\Delta\Delta H$ = -1.5, $[peptide]_{initial}$ = 500, $[LC8]_{initial}$ = 17, ΔH_0 = 0 and sigma = 0.2. Each trace includes multiple walkers, each of which is drawn as its own chain (see methods for details). Traces are thinned to one step for every fifty.



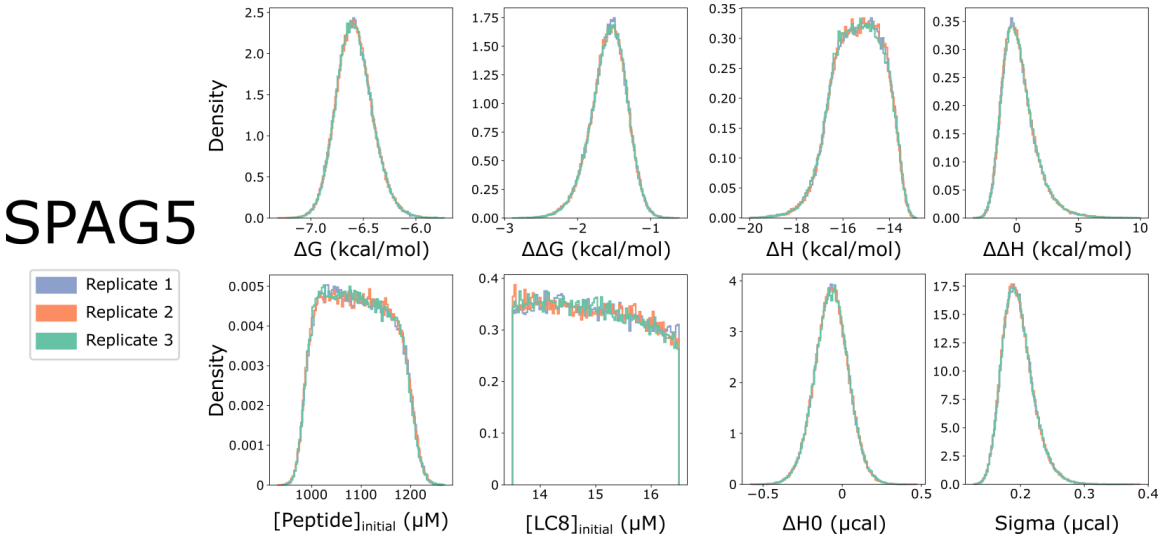
SI Figure 2.2: Distributions of thermodynamic parameters plotted with total free energies and enthalpies. Each plot shows a set of either ΔG and ΔΔG or ΔH and ΔΔH, along with the 'total' value for that parameter, i.e. 2ΔG+ΔΔG or 2ΔH+ΔΔH. The distributions for this sum value are often narrower than the individual parameters, as the total enthalpy and free energy of binding can be determined with higher precision from a given isotherm than the individual values. ΔG,ΔH are the energy and enthalpy of binding step 1, while ΔG+ΔΔG,ΔH+ΔΔH are the energy and enthalpy of binding step 2, making the total values reported here the energy and enthalpy of both binding steps combined.



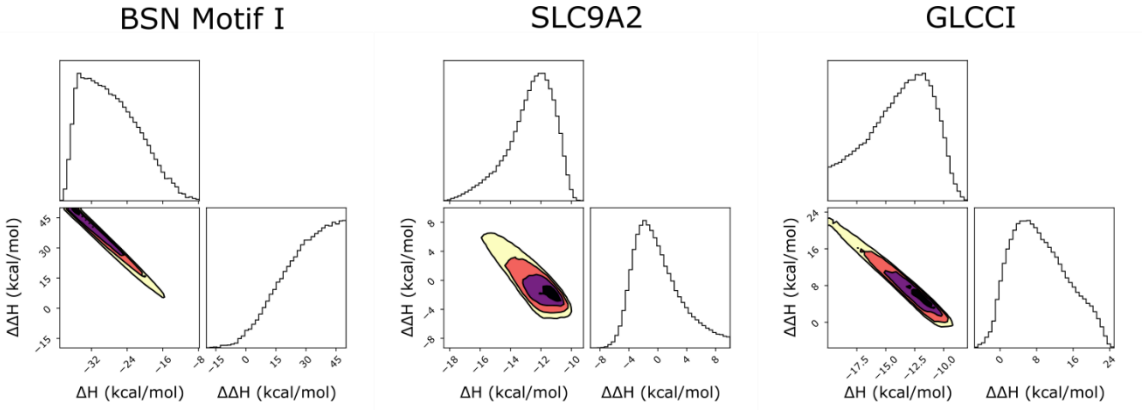
SI Figure 2.3: Marginal distributions comparing models with uniform and normal-distribution priors. Distributions are taken from models on an identical synthetic isotherm generated from parameters ΔG = -7, ΔΔG = -1, ΔH = -10, ΔΔH = -1.5, $[peptide]_{initial}$ = 500, $[LC8]_{initial}$ = 17, ΔH_0 = 0 and sigma = 0.2. All priors are identical except for [peptide]initial, where the uniform prior model (red) was run with a ±10% of stated value uniform prior, and the normal prior model (black) was run with a normal distribution prior with standard deviation = 1% of stated value.



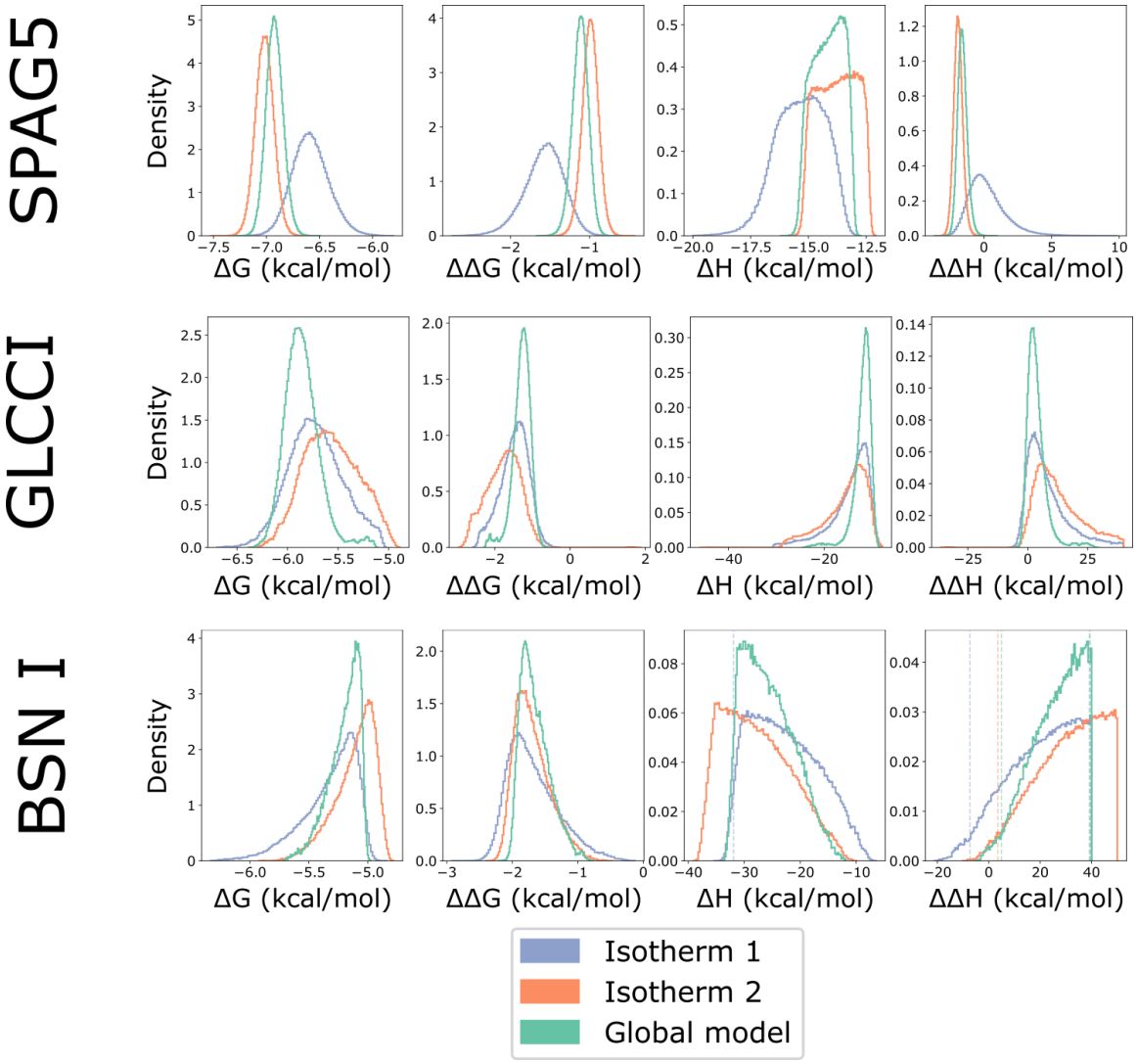
SI Figure 2.4: Effect of concentration priors on marginal posterior distributions for thermodynamic parameters in a 1:1 binding model. Distributions are taken from models on an identical synthetic isotherm generated from parameters ΔG = -8, ΔH = -12, $[X]_{initial}$ = 500, $[M]_{initial}$ = 34, ΔH_0 = 0 and sigma = 0.2. Model priors ar0e uniform distributions of varied width in each plot for $[X]_{initial}$ and $[M]_{initial,}$ varied from ±1% to ±50%.



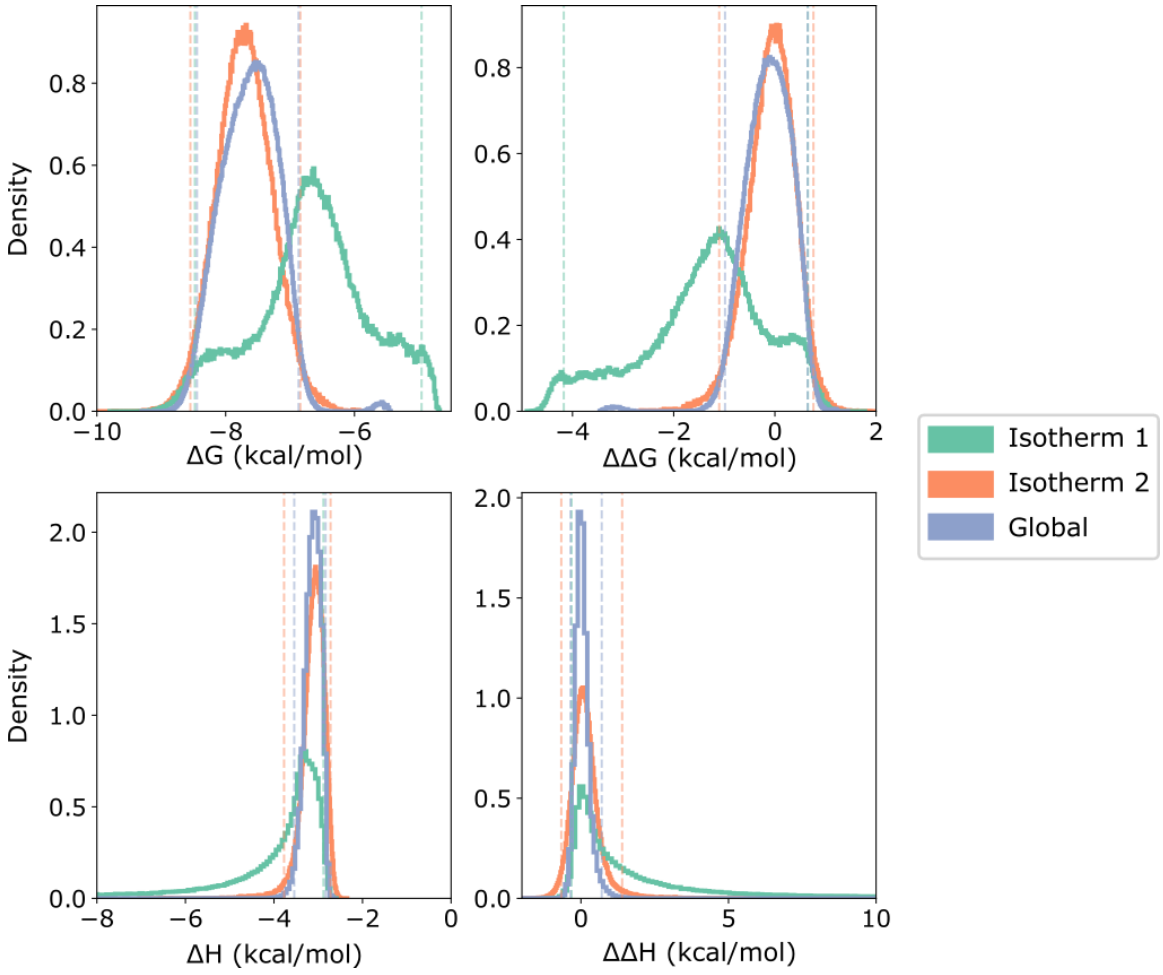
SI Figure 2.5: Example marginal distributions of replicate models for the LC8-SPAG5 interaction. Each model replicate is run on an identical isotherm with a different random seed dictating random starts for MCMC chains and trial move selections. Each model returns near-identical marginal distributions.



SI Figure 2.6: two dimensional marginal distributions of enthalpy for selected isotherms. Marginal distributions for BSN I, SLC9A2, and GLCCI are shown, each of which has wide 1D distributions for both ΔH and ΔΔH. Enthalpy parameters are closely correlated, resulting in a diagonal two-dimensional distribution within the enthalpy space.



SI Figure 2.7: Marginal distributions of thermodynamic parameters for individual and global models for three LC8-peptide interactions. Distributions for each individual isotherm and distributions for the global model are shown in purple, orange and green respectively. While the global model improves precision in determined parameters in some cases (e.g. GLCCI), in others it appears to follow the shape of the distributions for individual isotherms (e.g. BSN I).



SI Figure 2.8: Marginal distributions for thermodynamic parameters for IC-NudE binding isotherms. Distributions for each individual isotherm and distributions for the global model are shown in green, orange, and purple respectively.

SI Table 2.1: Model priors and sampling lengths for all isotherms.

Isotherm	samples	walkers	dG prior (kcal/mol)	dH prior (kcal/mol)	ddG prior (kcal/mol)	ddH Prior (kcal/mol)	X initial Prior (µcal)	M initial Prior (µcal)
Synthetic isotherm models	50,000	25-50	-3 to -10	-50 to 0	-4 to 4	-40 to 40	Varied*	Varied*
Single isotherm experimental models	100,000	50	-3 to -10	-50 to 0	-4 to 4	-40 to 40	± 10% of stated	Varied*
Two isotherm experimental models	200,000	50	-3 to -10	-50 to 0	-4 to 4	-40 to 40	± 10% of stated	Varied*

* ±10% of stated unless otherwise noted

SI Table 2.2: Credibility regions for 'sum' thermodynamic parameters and ratios of concentrations. 95% credibility region from sampled posterior distributions for the ΔG sum(2ΔG+ΔΔG) and dH sum(2ΔH+ΔΔH) as well as the ratio of concentrations ([peptide]/[LC8]). Credibility regions for ΔG and ΔH sums are frequently narrower than the credibility regions for individual parameters (Table 2.1).

Isotherm	ΔG sum min	ΔG sum max	ΔH sum min	ΔH sum max	conc ratio min	conc ratio max
SPAG5	-15.0	-14.5	-34	-28	71	75
BSN (I)	-12.2	-11.7	-29	-21	64	86
BSN (II)	-14.5	-14.1	-20	-17	61	66
SLC9A2	-14.2	-13.5	-29	-22	72	84
VP35	-15.7	-15.3	-27	-22	21	22
GLCCI	-13.2	-12.6	-21	-16	85	99
BIM	-18.2	-16.1	-26	-22	25	28

SI Table 2.3: Ranges of thermodynamic parameters for LC8-client binding when modeled with ±20% LC8 concentration. Table 2.1 in the main text contains equivalent information at ±10% LC8 concentration. Values delineate 95% Bayesian credibility regions from sampled posterior distributions, when modeled with ±20% priors for LC8 concentration. Distributions are largely very similar to those presented in Table 2.1, with a slight decrease in precision. BSN I, for which posterior distributions are significantly broader, is the only notable exception.

Isotherm	ΔG min	ΔG max	ΔΔG min	ΔΔG max	ΔH min	ΔH max	ΔΔH min	ΔΔH max	-TΔS min	-TΔS max	-TΔΔS min	-TΔΔS max
SPAG5	-6.93	-6.2	-2.12	-1.13	-19.11	-12.59	-1.87	3.7	5.97	12.61	-5.78	0.69
BSN (I)	-7	-4.76	-2	0.28	-36.87	-7.57	-15.87	48.08	0.77	31.97	-50	16.02
BSN (II)	-7.09	-6.44	-1.22	-0.39	-6.41	-4.44	-9.75	-6.53	-2.22	-0.46	5.68	8.99
SLC9A2	-6.85	-5.47	-2.64	-0.53	-24.97	-9.75	-5.19	24.76	3.19	19.49	-27.42	4.6
VP35	-7.4	-6.81	-1.67	-0.92	-15.23	-10.29	-0.77	1.45	3.22	8.08	-3.11	-0.18
GLCCI	-6.11	-5.21	-2.27	-0.99	-19.49	-9.35	-0.52	21.35	3.43	14.12	-23.57	-0.48
BIM	-9.5	-6.95	-2.18	0.8	-13.5	-9.57	-3.17	-0.47	1.05	5.54	-0.93	2.32

Chapter 3

Systematic Identification of Recognition motifs for the hub protein LC8

Aidan B Estelle, Nathan Jespersen, Nathan Waugh, Norman E Davey, Cecilia Blikstad, York-Christoph Ammon, Anna Akhmanova, Ylva Ivarsson, David A Hendrix, Elisar Barbar

This chapter is adapted from the following publication:
Life Science Alliance (2019) 2, e201900366

Abstract

Hub proteins participate in cellular regulation by dynamic binding of multiple proteins within interaction networks. The hub protein LC8 reversibly interacts with more than 100 partners through a flexible pocket at its dimer interface. To explore the diversity of the LC8 partner pool, we screened for LC8 binding partners using a proteomic phage display library composed of peptides from the human proteome, which had no bias toward a known LC8 motif. Of the identified hits, we validated binding of 29 peptides using isothermal titration calorimetry. Of the 29 peptides, 19 were entirely novel, and all had the canonical TQT motif anchor. A striking observation is that numerous peptides containing the TQT anchor do not bind LC8, indicating that residues outside of the anchor facilitate LC8 interactions. Using both LC8-binding and nonbinding peptides containing the motif anchor, we developed the "LC8Pred" algorithm that identifies critical residues flanking the anchor and parses random sequences to predict LC8-binding motifs with ~78% accuracy. Our findings significantly expand the scope of the LC8 hub interactome.

Introduction

Most proteins interact with few partners, but a class of proteins referred to as hubs interact with a large number of partners in complex protein–protein interaction networks[88,89]. Hubs can be static or dynamic. Static hubs bind a large number of partners simultaneously at different sites, for example, BRCA2[5]. Dynamic hubs bind multiple partners that compete for the same site[90,91]. Well-known examples of dynamic hubs include calmodulin and 14-3-3 proteins[92–94]. A more recently discovered member of dynamic hub proteins is the dynein light chain LC8[9].

There are more than 280 binary interactions for human LC8 in the Mentha database[95], some of which have been extensively studied, including the dynein intermediate chain (IC)[96–98] and the transcription factor ASCIZ[44,99,100]. In addition, expression patterns show that LC8 is highly expressed across a wide variety of cell types[101] and is broadly distributed within individual cells[102,103].

LC8 is an 89–amino acid homodimeric protein first identified as a subunit of the dynein motor complex. Colocalization and binding studies with dynein led to a common perception that LC8 functions as a dynein "cargo adaptor" to facilitate transport of dynein cargo[104,105]. However, further studies have shown that LC8 interacts with many proteins not associated with dynein at the same symmetrical grooves in the LC8 dimer interface

(Fig 3.1A). Because of the symmetry of the binding sites of the LC8 dimer, and its association with dimeric proteins, it is now generally accepted that LC8 serves not as a cargo adaptor in the dynein machinery but rather as a dimerization hub in a variety of systems[9].

LC8 interacts with an 8–amino acid recognition motif within intrinsically disordered regions of its partners. Sequences bound to LC8 form a single β-strand structure integrated into an LC8 antiparallel β-sheet[106] (Fig 3.1). Although there is some variation in the binding motif, it is most frequently anchored by a TQT sequence[10]. The glutamine in the TQT anchor is typically numbered as position 0 because it is the most highly conserved amino acid[29]. The flanking threonines are therefore defined as positions −1 and +1. The TQT anchor is highly enriched among known LC8 partners and will be referred to in this article as the “motif anchor”[10] (Fig 3.1B).

A dynamic binding interface, determined from nuclear magnetic resonance (NMR) relaxation and hydrogen/deuterium exchange experiments[29,40,107], allows for large sequence variation in LC8 binding partners; however, several steric and enthalpic restrictions are placed on binding sequences. One restriction is inferred from analysis of solvent accessible surface areas of peptides bound to LC8[10] (Fig 3.1C). The side chains of the amino acids at positions –1 and 1 of the peptide (both threonines in Fig 3.1C) are completely buried, leading to a strong preference for amino acids with branched side chains that are either hydrophobic or, as is the case for threonine, participate in hydrogen bonding. In fact, these two positions are the only side chains that are completely buried (Fig 3.1C, orange versus pink side chains), suggesting that these residues are under more stringent selective pressures. Interestingly, even though the amino acids on both sides of the anchor are highly variable, their side chains are easily fit within discrete pockets (Fig 3.1B). In contrast, outside of the 8–amino acid LC8-binding motif, there is higher variability in amino acid sequence and in side chain rotamer conformations (Fig 3.1B). Analysis of these structures explains the preference for the “TQT” anchor within the LC8 recognition motif but falls short of capturing the spectrum of amino acids that can flank the anchor in potential binding sequences.

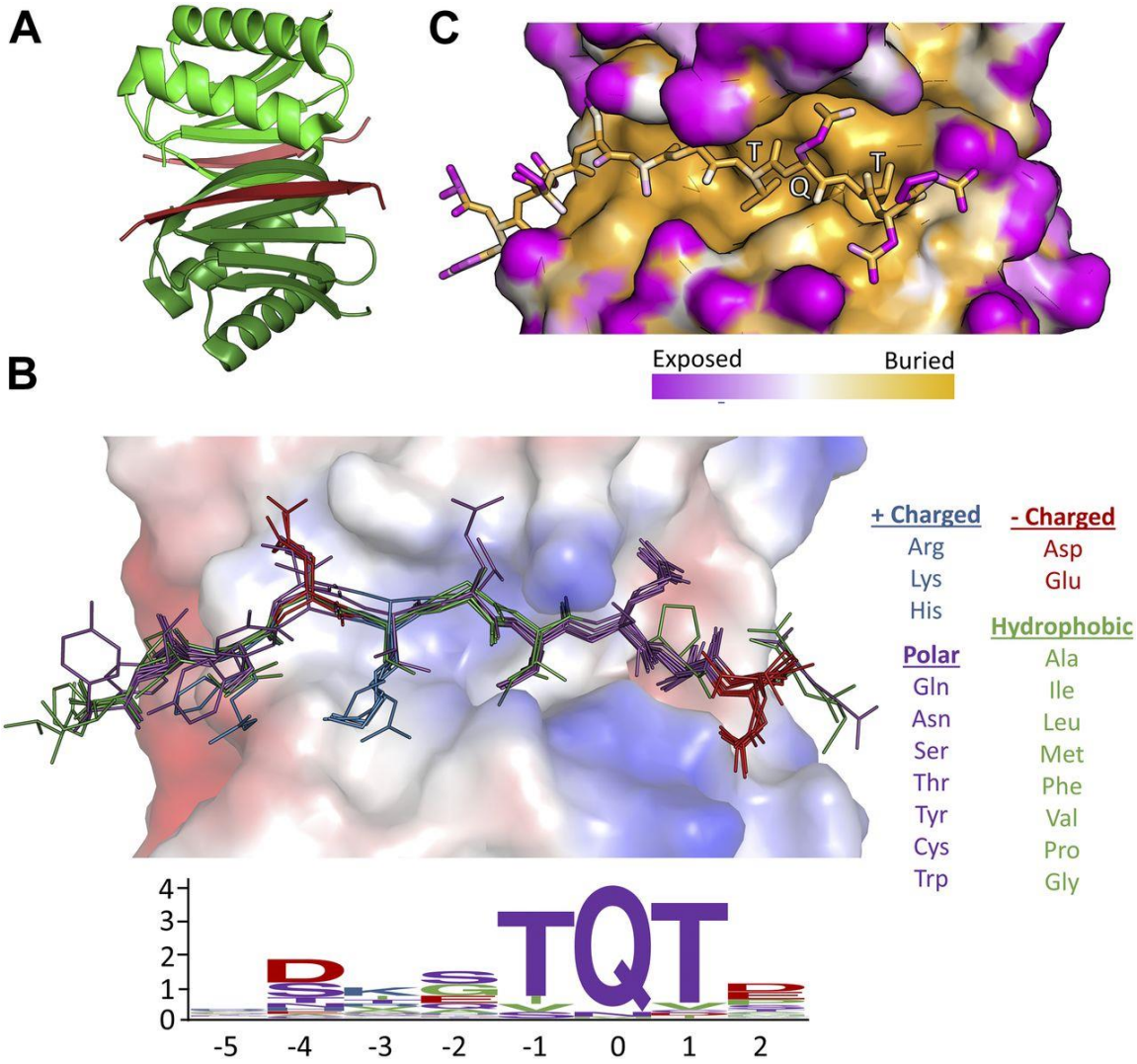


Figure 3.1: Motif sequence logo and surface analysis of LC8. (A) Crystal structure of a representative LC8 dimer (protomers shown in shades of green) bound to a peptide (shades of red). (B) Electrostatic charge potential for the LC8 pocket structure using PyMOL's charge-smoothed potential calculator, with positive potentials shown in blue, negative in red, and neutral in white. Peptides from available crystal structures of bound LC8 are shown, and colored based upon amino acid chemical characteristics (right). Amino acid enrichment is shown below each position within the LC8-binding motif, calculated from 79 known binder motifs listed on the LC8 database (http://lc8hub.cgrb.oregonstate.edu). Amino acid letter heights represent relative enrichment of that amino acid. (C) Solvent accessible surface area depiction of the same LC8/peptide pair shown in (A). Color scheme was defined at the atomic level using the GetArea program[108], with magenta representing more solvent exposed and orange regions more buried atoms.

In an effort to determine a consensus binding motif, Rapali et al (2011)[109] used phage display and randomized all 8 amino acid positions of the motif except for the

conserved glutamine at position 0, and determined VSRGTQTE to be the most thermodynamically favorable binding sequence[109]. Although this experiment led to the discovery of multiple LC8 binding partners, the idea of a specific “consensus sequence” belies the dynamics of the LC8 binding site. In addition, by selecting for the tightest binder, many weaker binders were likely outcompeted and therefore not visible in their study. Our goal in this work is to determine the extent of the variability in LC8 binding sequences flanking the motif anchor.

LC8 motif prediction analyses have increased the number of known binding sequences, and enhanced our understanding of the motif specificity[13,109]; however, algorithms generated in these studies were designed for initial screening and are therefore not sufficiently stringent for general use nor made publicly available. Here we build on initial proteomic peptide phage display (ProP-PD) experiments and position specific scoring matrix (PSSM) methods of identifying LC8-binding interactions to examine determinants of LC8 binding. We examine validated sequences from both these methods to pick out general trends in both LC8-binding and -nonbinding sequences. A database that includes partners identified in this work along with published interactions is now available and contains all 82 LC8 interacting motifs validated through biochemical or biophysical techniques. Finally, we used this database to develop an algorithm that incorporates both binding and nonbinding sequences to effectively predict LC8 partners and define rules for LC8 partner recognition that underscore the plasticity of the LC8 binding pocket.

Results

ProP-PD and PSSM-guided experiments reveal new LC8-binding sequences

To build our initial dataset of LC8-binding proteins, we performed ProP-PD experiments and constructed an initial PSSM to determine new LC8-binding motifs. Initial Pro-PD experiments on a library of disordered proteins from the human revealed 53 potential LC8-binding partners, which were validated by isothermal titration calorimetry (ITC). Surprisingly, of the 53 potential sequences, only 16 were demonstrated to bind LC8 by ITC (K_D < 50 μM). Further potential sequences were identified through scans of the human proteome. Following initial filtering that utilized the sequence’s propensity for disorder and degree of evolutionary conservation (see manuscript for details), sequences were scored for similarity to known LC8-binding proteins with a PSSM generated from weighing the

frequency of each amino acid at each position in known LC8-binding sequences against the expected background frequency (see Methods). ITC was performed on the 19 sequences predicted in this manner, and surprisingly only 7 of the 19 total sequences bound to LC8 with a measurable affinity. For a detailed discussion of these experiments, and a complete list of tested peptides, see Jespersen et. al., (2019)[8].

Common motif features that promote LC8 binding

To assess common features for binding from this growing dataset of interactions, we overlaid all known tight binding partners, (50 sequences with Kds <10 μM, Fig 3.2B) and all nonbinding sequences (determined by ProP-PD and PSSM, Fig 3.2C). This comparison revealed some conspicuous differences between binders and nonbinders, allowing for the determination of the position-based rules that follow (Fig 3.2A and D).

The anchor is extremely well conserved in both amino acid type and volume. There is a strong preference for a mid-sized H-bonding/hydrophobic residue at positions −1 and +1 and a clear preference for a glutamine at position 0. Any deviation from this anchor, such as the RQT seen in EIF4G3, leads to a nonbinding sequence. Both threonines are completely buried in crystal structures (Fig 3.1C), and therefore, deviations to a charged group are highly unfavorable (Fig 3.2D, Poor Anchor).

Position +2, which has no β-strand backbone interactions in any crystal structures, shows a large preference for proline, aspartate, and glutamate residues. Interestingly, these three residues are classically depleted in β-strands[110,111], providing a potential explanation for these residues acting as “strand-breaking” amino acids at the periphery of the LC8 binding pocket. An alternative explanation for their enrichment is that the negative charge for E and D can interact with the positive electrostatic charge on LC8 (Fig 3.1B). Proline, however, might energetically assist in binding by reducing the change in entropy, as both proline and pre-proline residues are conformationally restricted[112]. Hydrophobic amino acids are not well accommodated at this position (Fig 3.2D, Hydrophobic +2).

Position −2 shows little charge preference and allows positive, negative, polar, and hydrophobic residues; however, there are no examples of bulky aromatic side chains at this position among the tightly binding peptides, indicating that there are some steric constraints (Fig 3.2D, Bulky Hydrophobic −2).

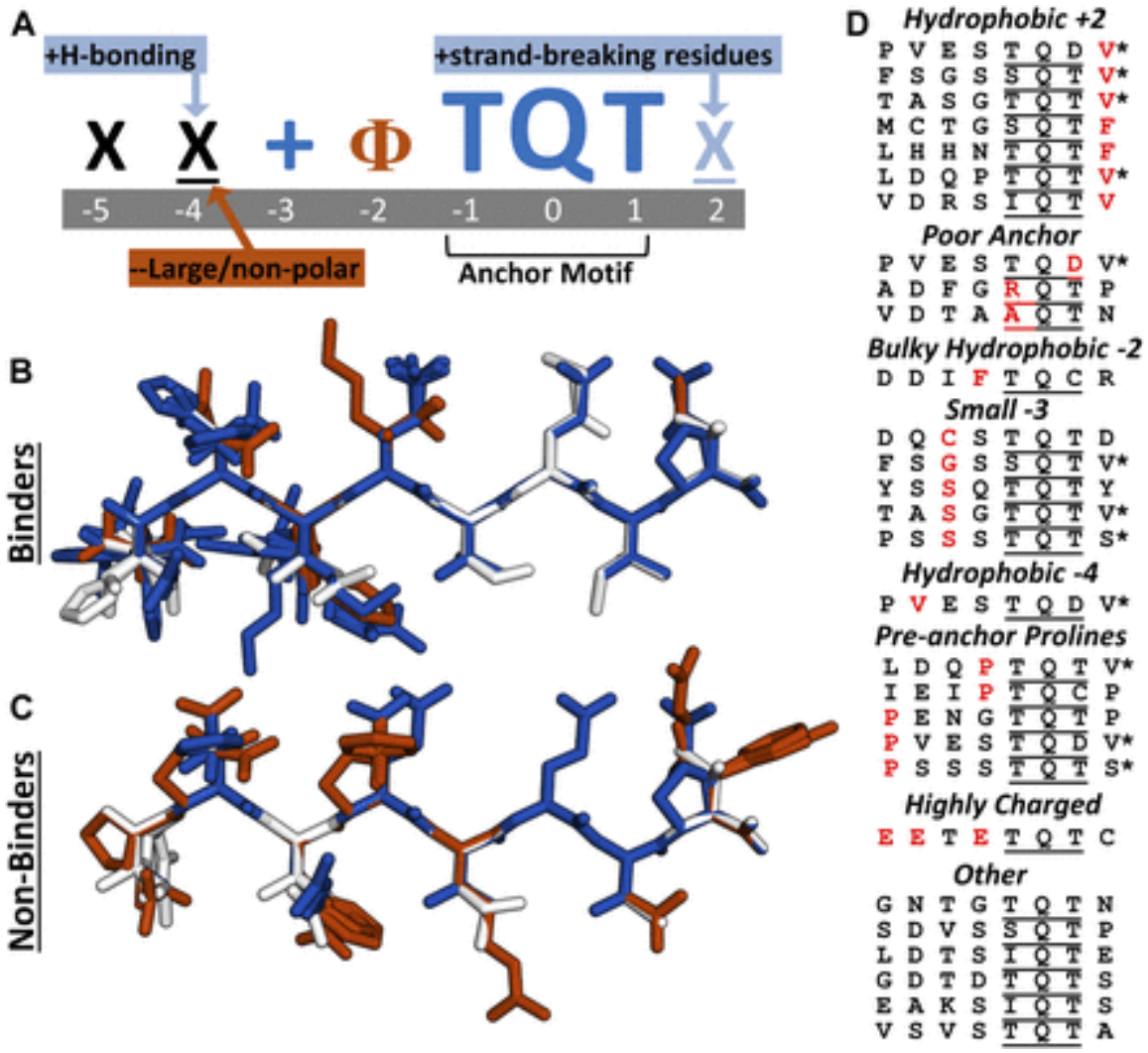


Figure 3.2: Analysis of LC8-binding and nonbinding motifs reveals distinct positional preferences. (A) Motif preferences for LC8 binding partners. "Φ" denotes hydrophobic residues; "X" signifies any residue (unless certain residues are disfavored); underlined "X" signifies any residue but with strong preferences for particular residues; "+" denotes positively charged amino acids. Physiochemical properties beneficial for binding are colored dark blue or light blue, based on magnitude, and deleterious properties are colored in red. (B) Known tightly binding sequences (Kd < 10 μM) are cropped to 8 amino acid motifs and built using the Chimera molecular modeling software. This includes LC8 sequences found on the LC8Hub database, and those determined in this article. (C) Overlay of all nonbinding peptides used in this study. Residues are colored based upon whether they are beneficial (blue), deleterious (red), or neutral (white) for binding, using the amino acid enrichment and depletion in known motifs (Fig 3.4A). (D) Categories of nonbinding sequences. Residues highlighted in red depict the reason the sequence is placed within a given category. *Denotes sequences placed in multiple categories.

Position −3 favors large side chains as nearly all tight binders contain an amino acid at least as large as valine at this position, with only two occurrences of an alanine. Fig 3.1B reveals a binding pocket where large side chains can fit, which is often occupied by lysines or arginines. A small side chain at the −3 position does not immediately exclude a sequence from binding, as in CAPRIN2, but seems to be less favorable based on the depletion of these residues (Fig 3.2D, Small −3).

Position −4 favors amino acids capable of making a polar contact, such as aspartate, and no sequences identified to date have hydrophobic residues larger than alanine at this position (Fig 3.2D, Hydrophobic −4). Finally, the −5 position shows a slight bias toward positively charged residues (Fig 3.2B and C), but it is unclear whether this effect is significant.

In general, partners must bind within a deep hydrophobic pocket and form a β-strand structure; therefore, multiple similar charges within a peptide, or sterically challenging prolines at any internal position, makes binding unfavorable. Even with this systematic comparison, a number of the nonbinding sequences could not be categorized (Fig 3.2D).

The partner-binding pocket is conserved in LC8 sequences but is structurally variable

A comparison of LC8 amino acid sequences from 58 different eukaryotic species using the ConSurf program[113] reveals that the partner binding site is strictly conserved across these diverse organisms (Fig 3.3A). Interestingly, the conservation of residues creates a noticeable gradient pattern that radiates out from the dimeric interface/partner binding site, with the most conserved residues near the core (maroon), and the least conserved residues at the peripheries (blue).

We used the Ensemblator program[114], which aligns independently determined 3D structures and identifies regions of structural conservation or plasticity, to visualize how a sequence that is strictly conserved is capable of binding such a wide variety of sequences. By overlaying the protomers from five published crystal and NMR structures of free LC8, we observed that the β-strand that directly binds to partners is highly variable[29,40,107] (Fig 3.3B) and has the highest root mean squared deviation (RMSD) values between structures. It is of note that the most sequence-conserved region is also the most structurally variable part of the protein. This structural plasticity allows accommodation of

a diverse set of partners with a wide range of properties and sheds light on why definitive identification of LC8-binding motifs is such a difficult task.

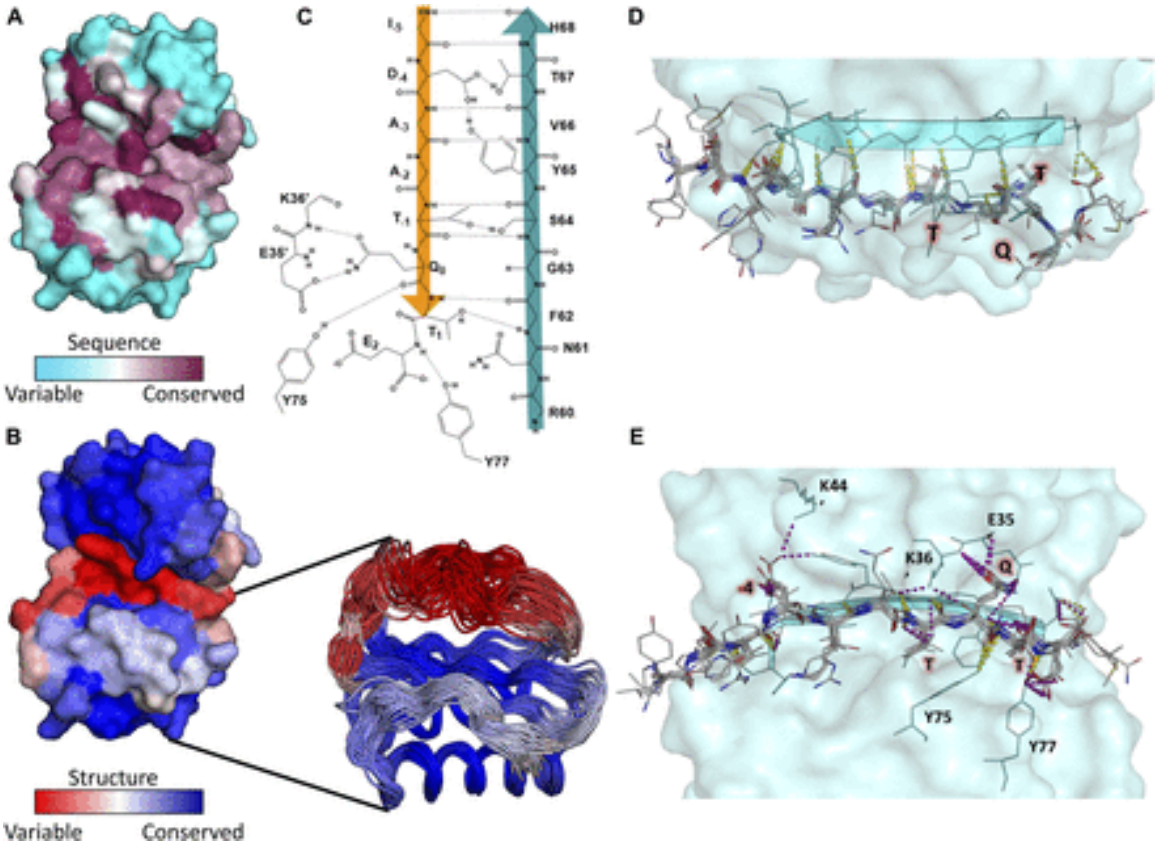


Figure 3.3: LC8 is structurally variable but conserved in sequence. (A) Surface representation of LC8 colored by sequence conservation using ConSurf. More sequence-conserved regions are shown in magenta, less sequence-conserved regions are shown in cyan. Highly conserved residues map to those within the LC8 binding site. (B) Surface representation of LC8 colored by structural conservation in the free protein using the Ensemblator. Regions that are more structurally variable are shown in red, whereas more structurally conserved regions are shown in blue. An overlay of NMR and crystal structure protomers used for the structural analysis is shown as a cut-out in (B). (C) 2D depiction of the binding interface between an example peptide (orange) and the binding β-strand within LC8 (Teal). (D, E) Polar bonds between LC8 and peptides from crystal structures are shown in (D) (top down view, only backbone interactions) and (E) (pocket view). Colors of polar contacts are based on whether the polar contacts stem from backbone (yellow) or side chain (purple) residues on the peptide. Peptide residues with frequent side chain interactions are labeled in red. (C, E) Residues outside of the binding β-strand that are important interaction sites shown in (C) are labeled in (E).

Incorporation of physicochemical features and nonbinder data improves binding predictions

Based on position preferences described above, we developed an LC8Pred algorithm that captures common features observed in binding peptides, including size and charge preferences, and features present in the 32 anchor-containing nonbinding peptide sequences (Fig 3.4A). For each matrix, positive values within the matrix indicate that the given amino acid is enriched in binding sequences and depleted in nonbinding sequences, whereas high negative values signify depletion of that amino acid in binding sequences and enrichment in nonbinding sequences. The addition of nonbinding sequence information significantly improved the algorithm's capacity to differentiate between binding and nonbinding sequences; however, with only 32 nonbinding sequences, our data were notably sparse, and separation between the two groups was incomplete. To improve our differentiation capacity, we binned the 20 amino acids into four categories and developed additional PSSMs using these bins, thereby reducing the overall number of matrix terms. The first PSSM separated amino acids based on polarity, whereas the second PSSM separated according to volume.

These matrices largely confirm groupings as described in the "common feature" section above, but with some exceptions. Notably, although there is a preference for large amino acids at the −3 position, the polarity matrix also shows an enrichment in positively charged residues. In addition, although the −5 position is the most varied in the matrix, it has a high score for positively charges residues (Fig 3.4A). This discrepancy is because of the lack of positively charged residues at −5 in the nonbinding sequences rather than from any strong enrichment of positive charge in the binding sequences. The −5 position also shows a slight enrichment for very large amino acids and is the only position to do so. Crystal structures show that the −5 position is not buried within LC8's binding groove and therefore experiences much less steric restriction (Fig 3.1B).

Using the described matrices, we scored all known binders and nonbinders to determine the discriminatory capabilities of the PSSMs (Fig 3.4B). Although the amino acid, volume, and polarity matrices were each moderately successful at separating binding from nonbinding sequences in isolation, the best separation was achieved when every matrix was combined. We combined the volume and polarity matrices to determine a volume and polarity score, and the amino acid matrix was used to determine an amino acid score (Fig 3.4B).

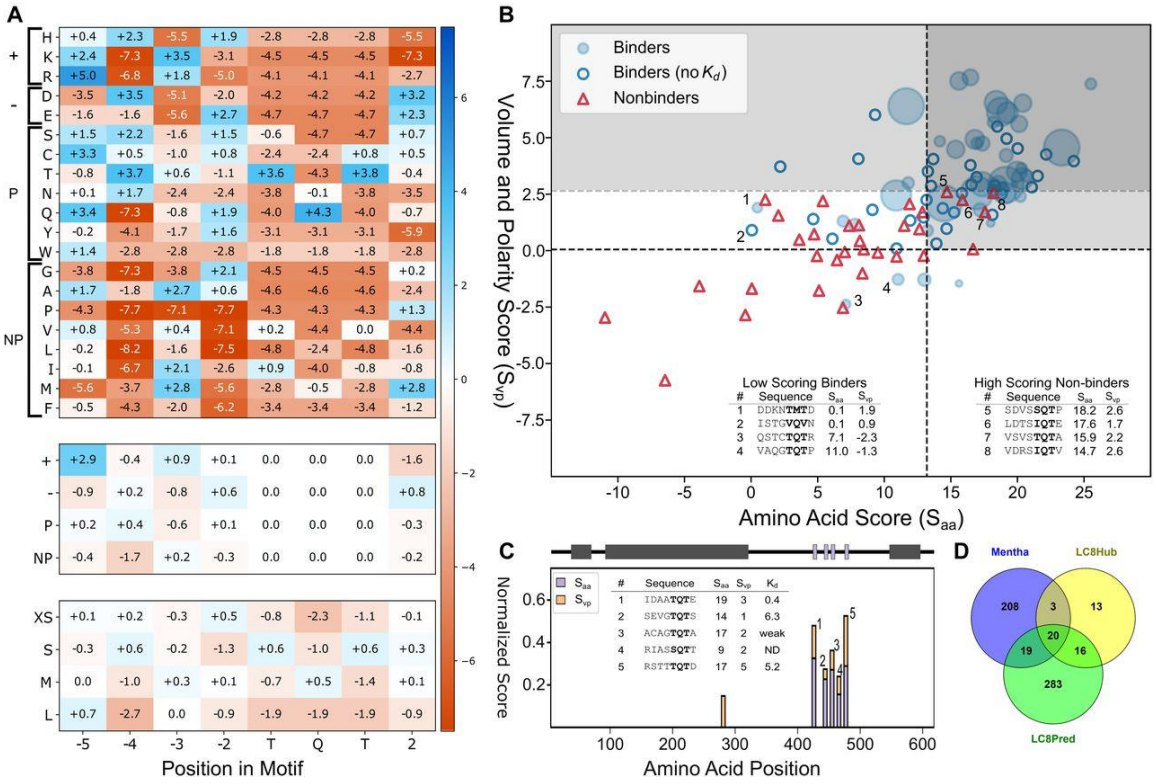


Figure 3.4: Generation and testing of The LC8Pred algorithm. (A) PSSMs for amino acids (A, top), bins by chemical property—positively charged, negatively charged, polar, or nonpolar (middle), and bins by volume—less than 106 A3, 122 to 142 A3, 155 to 171 A3, and greater than 200 A3 (bottom). Values correspond to the combined weight at a given position for the binder-only matrix and the nonbinder-normalized matrix. (B) Scatterplot of available sequences scored using a leave-one-out method of cross validation. For binders with a known Kd, the size of the bubble was varied inversely with the Kd, with binders with a Kd below 0.5 μM represented as the maximum possible dot size. Binder sequences with an unknown binding affinity were plotted as hollow circles and nonbinders as red triangles. The light grey box denotes predicted binding sequences using this scoring system. A second threshold for the volume and polarity axis indicates the very high confidence region, above which the specificity is unity. Outliers are noted in the tables (inset) and numbered in figure. (C) Normalized scores from matrices used to evaluate known LC8-binding protein Chica, where a score of one equates to the ideal amino acids of physicochemical properties at all positions. A sliding window to evaluate Chica for predicted binding sites across the protein was used, with the "0" position within the motif plotted (i.e., at 400, the 0 position is the 400th amino acid within Chica). A diagram of Chica showing secondary structure prediction (grey) and LC8 binding sites (purple) is above, and sequences predicted to bind are on the right, along with their corresponding scores. (D) Venn diagram of human proteins in the LC8Hub database, proteins that contain at least one LC8-binding sequence as determined by LC8Pred, and proteins reported to bind LC8 in the protein–protein interaction database Mentha[95].

Because our goal is to predict partners with high reliability, strict thresholds were used to determine what constitutes a binder and a nonbinder. A minimal score of 12.9 on the amino acid matrix, and 0.1 on the volume and polarity matrix, is used to determine whether a sequence is likely to be considered a binder. These thresholds result in only four false positives and 20 false negatives with our available data set, corresponding to a 75% true-positive rate and an 88% true negative rate (Fig 3.4B). Interesting, although the volume and polarity matrices only provide a small increase in accuracy overall at these thresholds, they are extremely proficient at separating binders from nonbinders when applied stringently. A threshold of 2.7 on the volume and polarity matrix alone results in a 0% false-positive rate, while retaining 57% of the true positives (Fig 3.4B).

Although we achieve an accuracy of 78%, there are a number of outliers: both high-scoring nonbinders and low-scoring binders. Within the binders, the first sequence, DDKNTMTD, is from Myosin Va (Fig 3.4B). It is unsurprising that this sequence scores poorly, as it is the only “TMT” anchor with verified binding data, and therefore has a low score because of the M instead of Q. However, binding is likely salvaged by the presence of the highly favourable amino acids at the other positions and by the presence of adjacent coiled-coil domains in the full-length protein. The remaining three lowest scores belong to proteins with multiple LC8-binding sequences proximal to one another (namely ASCIZ/ATMIN, and BSN), which would facilitate binding of weaker motifs because of their bivalency. Within the nonbinders, three of the four well-scoring nonbinders are listed in Fig 3.2D as “other,” indicating that there is consistency between algorithm predictions and our ability to recognize binders/nonbinders based on sequence. This also suggests that there are some deleterious interactions that we have yet to understand and will require more data to decipher. The fourth sequence contains a hydrophobic valine at the +2 position (Fig 3.4B, sequence 8), which is very rare, as this position is often fully solvent exposed and prefers β-strand breaking residues (Fig 3.1B). Although LC8Pred weights valine at +2 negatively (Fig 3.4A), the remaining residues score well enough to result in the erroneous categorization of this sequence as a binder. Further accumulation of LC8-binding and nonbinding sequences will no doubt help to clarify the importance of one poorly scoring residue and improve LC8Pred accuracy. Our LC8 motif algorithm is available on the database web page for public use (http://lc8hub.cgrb.oregonstate.edu/LC8Pred.php) for any sequence of interest.

Predictive scores for the human protein Chica: a known LC8 binder

To test the ability of LC8Pred to identify binding sequences, we scored a test protein on each matrix using a sliding window. For this test, we selected Chica, a protein that contains a series of LC8-binding sequences between residues 400 and 475[10]. To prevent algorithmic bias, peptides from Chica were not used in the development of our scoring matrix. Upon applying the LC8Pred algorithm, six positive scores were returned within Chica (Fig 3.4C). One of these scores fell far below threshold and was ignored. The remaining five scores were within the LC8-binding region; four of which have previously been determined experimentally to bind LC8[10]. The other is an SQT-containing sequence that scored below the designated threshold in the amino acid matrix, indicating that although this particular sequence may bind LC8, the prediction is of low confidence (Fig 3.4C). These test results provide strong evidence of the discriminatory power of our algorithm, as it can successfully recognize sequences that bind LC8 while excluding those that do not.

Human proteome scan identifies 374 potential binding sequences

After determining LC8Pred's reliability and ability to distinguish potential motifs, we used it to scan the human proteome to identify high-confidence binding partners. In total, 785 sequences scored above our PSSM thresholds. These sequences were then further filtered using IUpred to eliminate motifs within ordered regions. This process yielded 374 high-confidence hits from 338 proteins (Table S2 – see manuscript). Of these, 36 have been previously described in direct interaction studies and are listed on our LC8Hub database (Fig 3.4D). A further 19 partners have been identified in high throughput proteomics studies, such as pull-down mass spectrometry, including the highest scoring hit (FAM117B)[115,116]. Our data validate these interactions and define likely binding regions within these partners. It is of note that several of the identified partners contain multiple putative LC8 sites in close succession. The ability of LC8 to "zip up" partners with multiple recognition motifs has been described for both Nup159[15] and ASCIZ[44], and it is possible that many partners within this list contain weaker LC8 sites proximal to these tight-binding motifs.

Prior studies on LC8 interactions have noted an enrichment in LC8 partners within the Hippo signalling pathway[13]. Our proteome scan has identified these same partners (e.g., AMOT, WWC1, and WWC2) and additional novel binders from the hippo pathway,

such as STK4 and DLG5. Interestingly, this pathway is the only “biological process” significantly enriched in LC8 binding partners, based on gene ontology analysis using the WebGestalt program[117].

To verify that LC8Pred is correctly predicting partners, we synthesized three peptides from Table S2 (see manuscript) and tested their capacity to interact with LC8 via ITC. The three peptides were derived from the human proteins: HIV Tat-specific factor 1 (HTATSF1), a cofactor required for the Tat protein activation of human immunodeficiency virus transcription; otoferlin (OTOF), a calcium ion sensor involved in vesicle-plasma membrane fusion and neurotransmitter release, associated with hearing loss; and ninein (NIN), a component of the core centrosome and a dynein activator protein. These peptides were selected based on their mid-level scores and lack of prior data detailing LC8 interactions (Table S2 – see manuscript). All three peptides bound to LC8, although only HTATSF1 was a “strong” binder with a fittable thermogram (Kd of 10 μM). These data support the effectiveness of our LC8Pred algorithm and demonstrate that it is capable of predicting binding partners of varying affinities despite noncanonical motifs (Table S3 – see manuscript).

Discussion

Hub proteins are essential for cell viability as they are central in protein–protein interaction networks. Dynamic hubs such as LC8 often have a recognizable binding motif, which should allow for the prediction of binding partners without the need for exhaustive testing of each individual interaction[118]; however, no such program is available for LC8. Instead, binding partners are often identified via high-throughput pull-down experiments. For example, the interaction between LC8 and OFD1 was initially identified via pull-down mass spectrometry study in cilia[116]. In most cases, follow-up experiments for validation of direct binding are not performed, as it is prohibitively expensive to verify these interactions in a systematic fashion. Here, we validate purported and previously unreported LC8 binding partners (including OFD1), measure their binding affinities and thermodynamic properties, and establish a database of known LC8–partner interactions to define and describe generalizable requirements for LC8 motif recognition. We use these rules, along with amino acid preferences in nonbinding sequences, to develop an algorithm that effectively distinguishes between binding and nonbinding sequences, with the aim of facilitating a

priori prediction and discovery of LC8–partner interactions with much greater confidence and accuracy than has been possible before now.

Of the 72 synthesized tetradecameric peptides, we verified binding for 29 peptides derived from 27 distinct proteins. Of these 27 proteins, 19 are newly identified LC8 binding partners. It is of note that all validated sequences contain the canonical TQT anchor (or variation thereof) at the C-terminus of the peptide, supporting the idea that a C-terminal anchor is vital for LC8 binding. Although the LC8 binding site is structurally dynamic, there are distinct preferences and exclusions for each position within the binding motif (Fig 3.2). In addition to the presence of an anchor, binders often have −4 positions capable of H-bonding, larger positive side chains at −3 positions, and strand breaking +2 positions. However, the presence of pre-anchor prolines, a high concentration of charges, or bulky hydrophobic groups at the −2 position will each limit the likelihood that a sequence will bind LC8 (Fig 3.2).

Algorithms for motif identifications have been developed for both 14-3-3 and calmodulin to efficiently predict potential binding partners. In the case of calmodulin, its diverse set of binding motifs has led to multiple programs[119–121], which predict potential binding partners via a mixture of sequence similarity to known binders, α-helical propensity, or the number of canonical calmodulin-binding motifs within a given sequence. In the case of 14-3-3, which binds phosphorylated sequences within disordered segments of proteins, the algorithm makes use of support vector machines and artificial neural networks[118] and scores potential binding sequences using a PSSM. Here we succeeded in generating LC8Pred, an algorithm with a 78% accuracy rate, by incorporating nonbinder data and by reducing the PSSM dimensionality from 20 amino acids to four physicochemical categories, based on either polarity or volume. We have tested LC8Pred on the known LC8 binder Chica and by scanning the human proteome. In case of Chica, LC8Pred efficiently recognized known binding sites and excluded all other regions (Fig 3.4C). Our proteome scan identified 338 potential LC8 binding partners, including 19 binding partners that have been identified previously via high-throughput proteomics studies (Fig 3.4D and Table S2 – see manuscript), providing a new set of high-confidence LC8-interacting proteins. Three peptides were selected from these potential partners and shown to indeed bind LC8.

The ability to bind a wide variety of sequences despite an extremely conserved binding interface is a hallmark of dynamic hubs, as exemplified by calmodulin[122] and 14-

3-3 proteins[123]. Crystal and NMR structures for LC8 show that the β3 strand at the partner binding interface has the highest sequence conservation (Fig 3.3A), and surprisingly, it is also the most dynamic region (Fig 3.3B). Consistent with the dynamic nature of the binding grooves, thermodynamic analyses of tight binding sequences demonstrate a wide range of entropy/enthalpy compensation, including some sequences that bind with a favorable change in entropy, such as ICE1 and VP4 (see manuscript). Previous studies on LC8 dynamics of binding to dynein IC and the protein swallow (Swa) show that increases in ordered structure upon binding are peptide dependent[40]. With Swa, the complex is more compact, rigid, and homogeneous than with IC, indicating that the IC peptide retains more freedom of motion in the bound state than does the Swa peptide. Consistent with these observations, IC binds with a favorable entropy, whereas Swa does not. Our work here demonstrates that these different modes of binding are not limited to IC and Swa but rather that entropic factors commonly modulate LC8 binding to accommodate extraordinary variation in binding sequences.

Hub proteins like LC8 are essential for cell homeostasis as they sit at the center of complex interaction networks; therefore, it is imperative to understand the rules that govern hub protein interactions. The dynamic nature of the LC8 pocket, and entropic contributions to binding, make it difficult to predict partners with high confidence, and yet it is this very dynamic characteristic that makes LC8 such a powerfully effective hub protein. Here we have amalgamated our experimentally verified LC8-binding sequences with all previously described binding sequences and developed an algorithm that significantly advances our ability to predict LC8 partners based solely on sequence. Confidence in a potential LC8-binding sequence can be further improved by considering the structure and conservation of the binding site, and we have therefore linked LC8Pred to ProViz, a tool that analyzes protein structure and conservation. In addition, it is important to note that LC8Pred is optimized for stringency and predicting tight binding interactions and does not account for adjacent oligomerization sites, which would increase binding affinities. Future versions of the algorithm will incorporate parameters to account for other factors impacting binding, such as oligomerization state or subcellular localizations. We also anticipate that the predictive power of our algorithm will improve dramatically as more LC8-binding and nonbinding sequences are identified and deposited in the LC8hub database, resulting in a comprehensive view of the LC8 hub interaction network

Materials and Methods

ProP-PD selections

Phage display selections were performed using a proteomic library designed from the disordered regions of the human proteome described in the study by Davey et al (2017)[124]. Selections were performed with minor adjustments. GST-LC8 (0.1 mg/ml in 100 µl TBS, 50 mM Tris–HCl, 150 mM NaCl, pH 7.4) was coated on a Maxisorp 96-well plate (Nunc) via overnight shake-incubation at 4°C. Plates were blocked with 0.5% BSA in TBS for 1 h at 4°C and washed with TBS. The phage library was added to the well (100 µl) and incubated for 2 h at 4°C. Unbound phages were removed by washing plates five times with 300 µl TBS + 0.05% Tween. Bound phages were eluted by infection into 100 µl log-phase Escherichia coli Omnimax cells (Invitrogen; OD: 0.3–0.8) in 2xYT media (10 g bacto-yeast extract, 16 g bacto-tryptone, 5 g NaCl per liter) supplemented with 10 µg/ml tetracyclin. After a 30-min shake-incubation at 37°C, the bacteria were hyperinfected with M13K07 helper phages for 45 min to allow phage production. Cultures were transferred into 5 ml 2xYT, 0.3 mM IPTG, and grown overnight with antibiotics (25 µg/ml kanamycin and 100 µg/ml carbenicillin). The bacteria were pelleted by centrifugation. 1 mL of the phage supernatant was extracted and heat inactivated at 65°C for 20 min. Finally, the solution was pH neutralized using 10× TBS, and the phage pool was used in the next round of selection. Five rounds of phage selections were performed in total. The phage pool from the fourth day of selection was used for clonal phage ELISAs and sequencing. For next-generation sequencing, 5 µl of the phage pool from the fourth day of selection was used as template in a barcoding PCR. The sample was prepared and analyzed as described in detail elsewhere[125].

Peptide synthesis

A total of 72 putative binding partners identified from ProP-PD selections and algorithm predictions were commercially synthesized from either Genscript, or Synpeptide, as 14–16 amino acid sequences. Non-native residues were added to the termini of some peptides to facilitate solubility and peptide concentration determination (Tables 1 and 2 – see manuscript). All peptides were derived from either human or viral proteins.

Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) experiments for the interactions of LC8 with peptides were performed using a Microcal VP-ITC microcalorimeter at 25°C in buffer composed of 50 mM sodium phosphate, 50 mM NaCl, 1 mM sodium azide, and 5 mM β-mercaptoethanol, pH 7.5. Some peptides contained cysteine residues, so 5 mM β-mercaptoethanol was included in all solutions for consistency. In all experiments, an initial 2 µl injection was followed by 26–50 injections of 3–10 µl peptide (500 µM) into 25 µM LC8 in the sample cell. Number and volume of injections were adjusted for each experiment to minimize ambiguity in the shape behaviour of isotherms and thermograms. Peptide concentrations were determined from absorbances at 280 nm using molar extinction coefficient values computed with the Protparam tool on the ExPASy website[126]. Peptides lacking aromatic residues were weighed and resuspended in the proper volumes to ensure 500 µM final concentrations. Protein samples and buffer were degassed before data collection. Data were processed using Origin 7.0 (Microcal) and fit to a single-site binding model. Final values for binding parameters are averages of two to three independent experiments.

LC8Pred algorithm generation

The LC8Pred algorithm was developed using 79 LC8 binding sequences and 32 anchor-containing nonbinding sequences (See Manuscript – Table S4). We selected sequences that bind LC8 with high confidence, on which direct interaction data are available. In addition, all sequences with a Kd above 25 µM were not included. The TQT (or variation thereof) anchor-containing nonbinders were those peptides shown by ITC to have no binding to LC8.

In addition, a new series of matrices were developed which binned amino acids into categories based on physicochemical properties. Specifically, a matrix that separates amino acids into positively charged, negatively charged, hydrophobic, or polar and uncharged, and a matrix that separates amino acids into four groups based on volume, with volume bins being selected to minimize the range of volumes within each bin. We built these matrices to overcome the limitation of our small dataset, as reducing the number of groups from 20 amino acids to four possible properties improves the likelihood that some information is available for a given position and a given property within the motif.

In total, six matrices were developed, two for each set of bins (amino acid, polarity, and volume). For a given bin, one matrix was normalized to the background frequency of a given amino acid or a given property within the disordered eukaryotic proteome taken from the DisProt database of intrinsically disordered regions[127]. For the other matrix, normalization was done for the frequency of a given amino acid or property in the nonbinder dataset. As nonbinding sequences were selected based on the presence of an anchor, there is no enrichment or depletion at the anchor positions of −1 to +1. These positions were therefore ignored in these matrices.

To simplify our scoring system, we combined the matrices into two simple scoring metrics, Saa and Svp, where Saa is a combination of the two matrices that use amino acid–type bins, and Svp is a combination of the four matrices that use volume or polarity bins. To determine how effective each individual matrix was at separating binding and nonbinding sequences, we scored our available sequences using leave-one-out cross validation, where a given sequence was excluded from the matrix and then scored. The leave-one-out approach was used to combat the difficulty of our limited dataset.

We used receiver operating characteristic (ROC) curves (SI Figure 3.1) as a metric of the effectiveness of each score. The area under these curves corresponds to the ability of each matrix to separate binding sequences from nonbinding sequences. We then combined scores into the Saa and Svp scores described above, where each individual matrix score was weighted through a grid search of possible weights, where the largest area under the ROC was taken to be the optimal weight for each score. Surprisingly, the area under the ROC curve was highest when the binder-only polarity matrix was removed from the Svp score. Positions −1, 0, and 1 are therefore not weighted in the polarity matrix (Fig 3.4A) because the nonbinder normalized matrix was also excluded at those positions because of a lack of anchor enrichment, as discussed above.

The LC8 motif repository

We have manually curated a database that compiles information for all known LC8 binding partners. Including the 19 binding partners identified in this work, there are currently 80 experimentally confirmed LC8 interacting partners containing 116 individual anchor motifs. Of these binding motifs, 98 have been confirmed by in vivo or in vitro experiments, with a further 18 identified through biochemical screening methods. The database serves to (1) provide a source of up-to-date information on LC8 and its cellular role, (2) organize and

classify LC8 binding proteins in an easily searchable manner, and (3) list the sequences of all TQT motifs to aid in identification of new binding partners. Access to the motif repository is available at http://LC8hub.cgrb.oregonstate.edu. For each protein, the following information is provided: the species, TQT peptide sequence, number of motifs in the protein, Protein Data Bank (PDB) ID (if a structure exists), reference link, and interaction type. The interaction type has three levels of classification, depending on the method by which the LC8–partner interaction was identified: (1) high-throughput biochemical method, such as yeast-2-hybrid, where the interaction has not been confirmed by in vivo or in vitro experiments; (2) in vivo experiments, such as mutation or knockout experiments, where a function for the LC8-partner complex has been identified; and (3) in vitro experiments that determine the binding affinity, structure, or other information about the LC8–partner interaction. In addition, sequences of interest can be tested at LC8Hub by inputting a .fasta file or a string of letters corresponding to the protein sequence of interest. Output provides both the Saa and the Svp scores, and indicating sequences that are likely to bind LC8 according to available data. Finally, sequences determined to either bind or not bind LC8 despite the presence of an anchor sequence can be submitted for incorporation into the database. It is our hope that the information in this database will facilitate research on LC8 and, by enhancing our understanding of the TQT motif, enable more robust prediction of new binding partners.

Structure and motif analysis

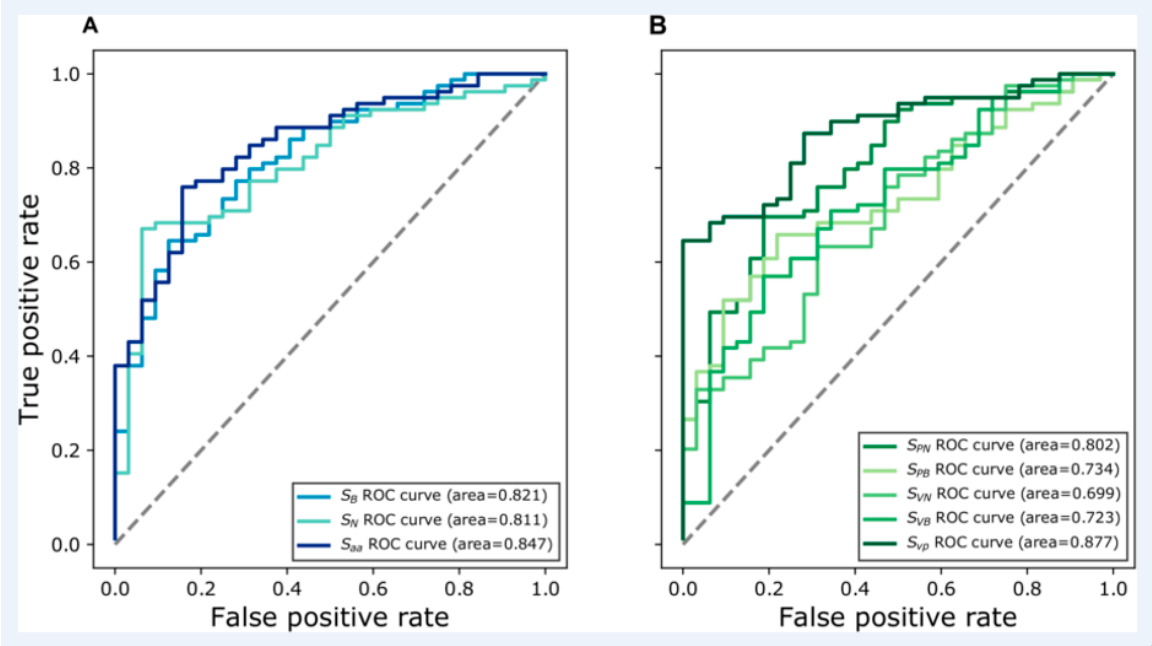
Structures of LC8 were obtained from the PDB (free LC8 PDB codes: 1PWJ, 1PWK, 1RE6, 3BRI, 5WOF; bound to peptides: 2XQQ, 4QH7, 3E2B, 2P2T, 3BRL, 3DVP, 3P8M, 3ZKE, 4D07, 4HT6, 5E0M). All images were generated using PyMol. Peptides without structures available were built in silico using Chimera[128]. Peptides in Fig 3.2 are colored according to enrichment and depletion tables for amino acids, shown in Fig 3.4A (blue for scores >1, white for scores between 1 and −1, and red for scores <1). Solvent accessible surface area analysis was performed using a representative LC8 crystal structure (2XQQ) with the GETAREA program[108]. Protein charge potential was calculated for LC8 using PyMol's built-in charge-smoothed potential calculator. Two-dimensional lig-plots were generated using ChemDraw.

Alignment of LC8 structures was done using the Ensemblator[114] program, and the RMSD for residues in free LC8 structures (listed above) was calculated using the built-in

local alignment tool. This tool works by aligning each dipeptide within the protein and calculating the RMSD for the next amino acid within the protein sequence. A representative structure was then colored based on these values to demonstrate structural conservation. Sequence-based conservation was performed using ConSurf[113], with LC8 sequences from 58 different eukaryotic species.

Acknowledgements

This work was supported by the National Science Foundation grant MCB 1617019 to E Barbar, by the Swedish Research Council (C0509201) to Y Ivarsson, by Lennanders foundation and Ingegerd Bergh's foundation to C Blikstad. The NGS was performed in collaboration with Eduard Resch, Fraunhofer Institute for Molecular Biology and Applied Ecology IME TMP, Frankfurt am Main, Germany. Y-C Ammon is supported by the MARIE SKŁODOWSKA-CURIE ACTIONS Innovative Training Network 675407 PolarNet.



SI Figure 3.1: Optimization of matrix weights. (A) ROC curves of both amino acid matrices and Saa. The larger the area under the curve (AUROC), the more effective the curve is at separating binder sequences from non-binder sequences. (B) ROC curves of each volume and polarity matrix, and Svp. Notably, the Svp curve performed substantially better than each volume and polarity matrix individually, which suggests that volume and polarity are both essential to understanding the preferences within the LC8 motif.

Chapter 4

Multivalency Drives binding between LC8 and the cytoskeletal regulator Kank1

Aidan B Estelle, York-Christoph Ammon, J. Helena Kinion, Anna Akhmanova, Elisar Barbar

Abstract

Kank1 is a cytoskeletal regulator localized to the cortex of the cell, where it binds to both focal adhesions, which regulate the actin cytoskeleton, and cortical microtubule stabilizing complexes. The protein LC8 is a small dimeric protein that acts as a dimerization hub for many clients through binding at a short linear motif. Many LC8 clients bind the protein multivalently, through repetition of the LC8-binding linear motif. While the exact function of each multivalent LC8 client is unique, these interactions are thought to play a structural role, rigidifying a disordered region of the client protein. Here, we present work that demonstrates that despite containing only a single predicted LC8 motif, Kank1 binds multivalently to LC8, driving LC8's localization to the cell cortex. Kank1-LC8 binding is highly cooperative, with an overall binding affinity at least than two orders of magnitude greater than the affinity of each individual LC8-binding motif. This cooperativity results in a complex that favors a homogenous, fully bound, and rigid state. We believe this cooperative complex serves a structural role, with cooperativity ensuring that the protein transitions efficiently between a rigid and flexible state.

Introduction

LC8 is a small (20 kDa) dimeric protein that acts as a binding hub via interaction with intrinsically disordered protein (IDP) clients mediated by a short linear motif[8,9]. Hub proteins like LC8 interact with many binding partners, making them central points of networks of protein-protein interactions, and therefore important points of regulation[5,6]. The LC8-binding linear motif is found within disordered regions of client proteins and is defined by a TQT amino acid sequence which anchors the client within LC8's binding sites (Fig. 4.1a,b)[8,10]. LC8 contains two symmetric binding grooves, at either side of the protein, making it capable of accommodating two client strands[29]. In many complexes, LC8 acts as a dimerization engine, binding two strands of the same protein at each site and inducing or modifying a dimeric structure in the client protein (Fig. 4.1a)[9,12,19,32].

Due to the structural simplicity of linear motifs, such as the LC8-binding motif, they are ideally suited to the facilitation of multivalent binding, with several motifs in the same sequence of a protein[41,129]. Such multivalent complexes frequently play essential roles in large macromolecular complexes, where IDPs act as scaffolds[41]. These LC8 complexes are involved in a host of cellular functions: at the nuclear pore[15,16], regulating transcription[44,45,100], at neuronal synapses[17], and regulating the cytoskeleton during cell

division[10], among other functions[22,37,50]. Multivalent LC8-binding interactions result in the formation of a ladder-like, 'polybivalent' complex, where two strands of the client IDP form the body of the ladder and are held together by LC8 rungs[41,44]. These complexes present a significant technical challenge to study, owing to the thermodynamic complexity of binding, and heterogeneity in structure and occupancy of bound complexes[42,44]. Indeed, characterized multivalent LC8 complexes are often highly structurally heterogeneous when examined by ultracentrifugation or electron microscopy, and this heterogeneity may play a functional role, such as in the LC8-binding protein ASCIZ, which senses the LC8 concentration in the cell and regulates LC8 expression[44].

KN motif and ankyrin repeat protein 1 (Kank1) is a ~150 kDa member of the Kank protein family, which share an N-terminal KN motif and C-terminal ankyrin repeat domain[130,131]. Of the Kank family of proteins (Including Kank2-4) Kank1 is by far the most extensively studied, although each protein in the family is thought to function similarly to Kank1. The protein shares many characteristics with known multivalent LC8-binding proteins, including large regions of disorder, coiled-coil domains, and involvement in large macromolecular complexes. Beyond the KN motif and repeat domain, Kank1 consists of a mix of predicted disorder and coiled-coil structure, including several coiled-coils predicted from residues 258-501, and two long stretches of predicted disorder, named L1 and L2, spanning residues 60-258, and 501-1161 (Fig. 4.1c). Kank1 is a tumor suppressor[130,132], and frequently found downregulated in cancer tissues in the kidney[132], brain[133,134], and lungs[135]. The exact mechanism of Kank1's tumor suppressor activity is unclear, although it may be connected to the role that Kank1 plays in regulation of the cytoskeleton at the cell cortex.

Uniquely, Kank1 can be found in complexes regulating both the actin and microtubule cytoskeleton[131,136,137]. The protein's KN motif binds tightly to Talin1[131], which is involved in the formation of focal adhesions (FAs) (Fig. 4.1d), large protein complexes at the cell cortex which regulate the growth of the actin cytoskeleton. In addition to this, Kank1's coiled-coil domains bind to Liprin-β1, a component of cortical microtubule stabilizing complexes[137] (CMSCs), regulatory complexes that control microtubule growth (Fig. 4.1d). Kank1 recruits the kinesin KIF21A to these stabilizing complexes, slowing the growth of microtubules at the cortex and reducing the incidence of 'catastrophes' where microtubules grow into the cell membrane[137]. As a member of both actin and microtubule-

regulating complexes, Kank1 offers a potential route for crosstalk between the two systems.

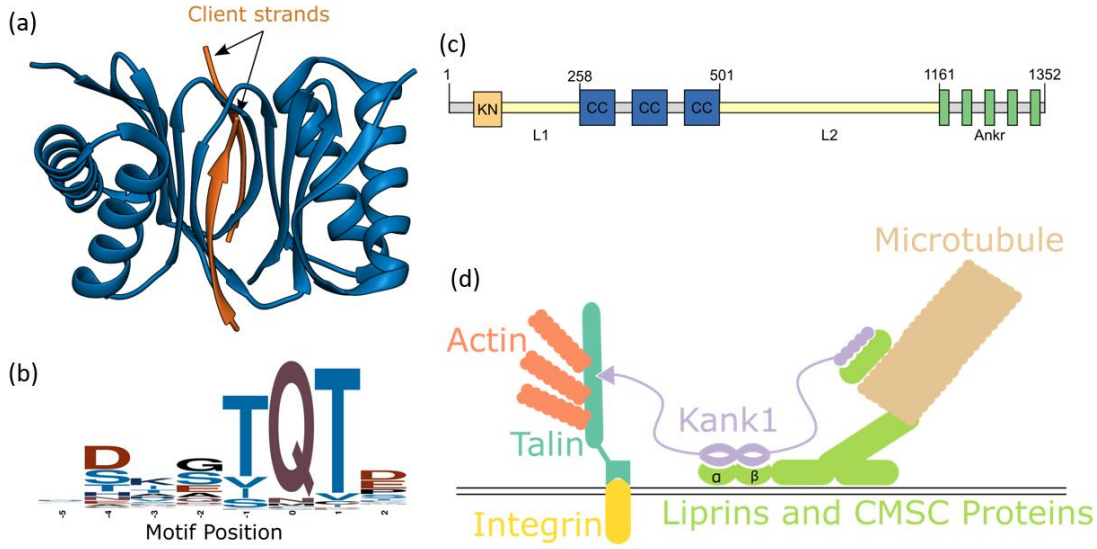


Figure 4.1: LC8 and Kank1 structure and function. Ribbon diagram of LC8, drawn in blue, with two client strands (PDB 2P2T). Clients take a beta-strand structure when bound to LC8. (b) Sequence logo of the LC8 motif, constructed from all known LC8-binding sequences in the LC8Hub database[8]. (c) Diagram outlining Kank1 structure, including KN motif (orange), ankryin repeat domain (green), coiled-coils (blue) and linker sequences (yellow). (d) Schematic of Kank1 (purple) binding to FAs (left) and cortical microtubule stabilizing complexes (CMSCs) (right). Kank1 interacts with Talin1 (teal) to bind FAs and is additionally recruited to CMSCs by liprins (light green).

Here, we present work that demonstrates that Kank1 binds to LC8 multivalently, at a site within Kank1's disordered L2 region. We demonstrate that in human cells, Kank1 recruits LC8 to the edges of focal adhesions at the cell cortex. Despite containing only a single predicted LC8 motif, we show that Kank1 binds LC8 multivalently, forming a cooperativity-driven ladder-like complex. Uniquely among multivalent LC8-binding proteins, the LC8-Kank1 complex is structurally homogenous, suggesting that binding may play an inducible structural role, allowing for flexibility when needed during macromolecular complex formation, and rigidity when needed to protect the full complex.

Results

Kank1 colocalizes with LC8 at focal adhesions

Pulldown mass spectrometry experiments on Kank1 provided the first evidence of interaction between Kank1 and LC8, demonstrating that Kank1 pulls down LC8 along with

known Kank1-interacting proteins such as talin1 and liprin-β1[136]. To examine this interaction in cells, we stained for paxillin (an FA protein), and Kank1 or Kank2 in HeLa cells stably expressing LC8 tagged with C-terminal GFP. Examining localization of each protein, the signal for LC8 strongly overlaps with Kank1 in patches at the edge of FAs. Kank2 does not exhibit the same pattern of localization, suggesting Kank1 may be the only protein in the Kank family to bind LC8. Cells with Kank1 expression suppressed via siRNA knockdown see elimination of LC8 localization, confirming that Kank1 is responsible for localization of LC8 to FAs. Knockdown of Kank2 does not appear to have an impact on LC8 localization, confirming that Kank2 plays no LC8-interacting role.

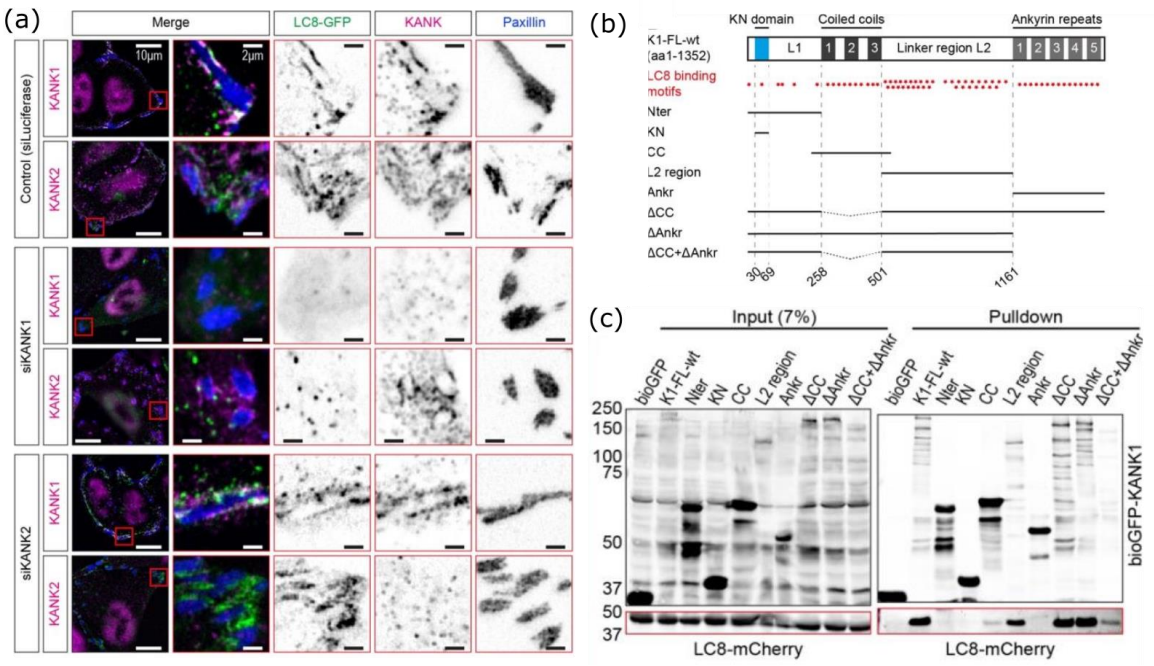


Figure 4.2: LC8 Colocalizes in Kank1 in HeLa cells. (a) Cells expressing LC8-GFP (green), stained for Paxillin (blue) and Kank1 or Kank2 (red). LC8 and Kank1 colocalize at the edges of focal adhesions. Cells are transfected with siRNA for luciferase (as a control), siKank1 and siKank2. Silencing of Kank1 abolishes LC8 colocalization. (b) Diagram of Kank1 structure, with bio-GFP tagged constructs designed for pulldowns shown. Motif-like amino acid sequences are marked with a star at the top of the diagram. (c) Streptavidin pulldowns of LC8-mCherry with bioGFP-Kank1. Left is input of crude cell lysate being pulled down, while right are blotted for GFP (top) and mCherry (bottom). Figures are adapted from chapter 4 of Ammon et. al., (2020)[136].

Kank1's intrinsically disordered L2 region binds LC8

To determine an approximate LC8-binding region for Kank1, we performed a series of streptavidin-based pulldown assays on LC8, using constructs of bioGFP-Kank1 as bait.

These experiments revealed that the L2 region (residues 501 to 1161) is the primary LC8-binding component of Kank1 (Fig. 4.2b,c). The L2 construct on its own strongly pulls down LC8, and other LC8-binding constructs all contain L2. Detailed examination of the L2 sequence reveals a canonical TQT anchor motif at residue 710 of the protein, further suggesting that LC8 binding occurs at L2 in Kank1.

Using these pulldown experiments as a guide, we were able to further narrow down the LC8-interacting region to a predicted segment of disorder between residues 500 and 800 of the L2 region. We scored the L2 region on the LC8-motif predicting tool LC8Pred[8] and found that all predicted LC8-binding motifs are within residues 600-720. Within this region, LC8pred predicts a single LC8-binding motif, a sequence of ASRGVNTE with the N instead of a Q at residue 651 (Fig. 4.3). In addition to this, the analysis yields 5 additional sequences with prediction scores above 0. It should be noted that these scores, per LC8pred criteria, are not predicted to bind LC8 (which requires a total score over 13)[8] but are nonetheless worthy of consideration. Of the remaining 5, the lowest-scoring sequence (anchored with an unlikely TAT sequence at position 689) can be discarded, due both to the alanine at motif position 0, which has never been observed in LC8 binding, and to its position between two TNT-sequence containing motifs which overlap the TAT sequence. This leaves us with a total of 5 potential LC8-binding motifs: an SNT-containing motif at position 605, a VNT-containing motif at 651, two TNT-containing motifs at 685 and 697, and the aforementioned canonical TQT-containing motif at 710 (Fig. 4.3).

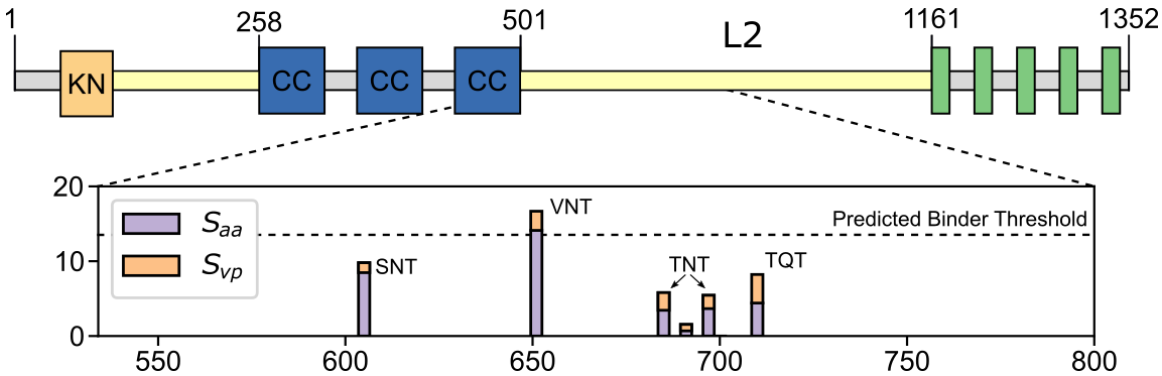


Figure 4.3: LC8pred predictions for the 500-800 region of Kank1. Diagram of kank1 (above) showing folded domains, as well as the L2 region where LC8 binding takes place. LC8pred scores along a sliding window for residues 500-800 of the L2 region. Scores are shown as a stacked bar plot combining the amino acid score (purple, S_{aa}) and volume and polarity score (Orange, S_{vp}). A combined score of 13 is sufficient to predict LC8 binding. The anchor sequence is listed for each motif.

Kank1 binds LC8 multivalently

To investigate binding between Kank1 and LC8, we recombinantly expressed a section of Kank1's intrinsically disordered region stretching from residue 595 to 720, to include all predicted LC8-binding motifs. To measure binding between LC8 and Kank1$_{595\text{-}720}$, we performed isothermal titration calorimetry (ITC) experiments titrating LC8 into a sample of Kank1$_{595\text{-}720}$ (Fig. 4.4, right). We found binding to be on the low micromolar scale (1.3 μM), with an enthalpy of -6.6 kcal/mol, consistent with expectations for a tight-binding LC8-client complex. Notably, the value of *n*, as calculated in the ITC analysis package implemented in Origin[53], is 5.92, suggesting a very high stoichiometry of LC8:kank binding. While *n* is not always a trustworthy parameter in complex cases due to the assumptions in the independent-sites model[138], this high value of n can be taken as an indicator that Kank1 binds LC8 at multiple motifs, indicating that LC8pred (which is not designed with multivalent cooperativity in mind) is underpredicting LC8 binding in this case, and several of Kank1's other LC8 motifs play a role in LC8 binding.

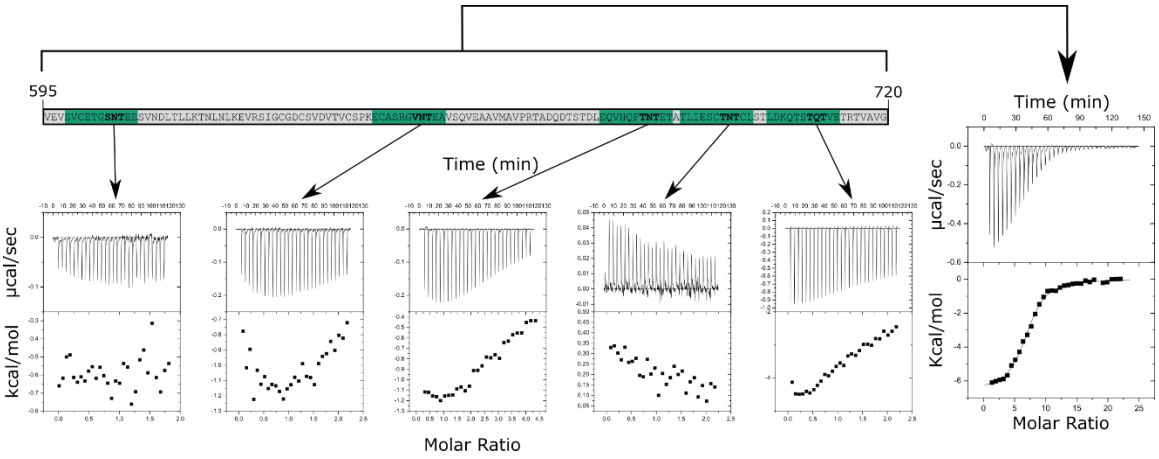


Figure 4.4: Isotherms for Kank1$_{595\text{-}720}$ and peptides from kank1 motifs. Figure displays the sequence of kank1$_{595\text{-}720}$, with each predicted motif from 4.3 highlighted in green. Isotherms for titration of a peptide for each motif into LC8 is shown below. No isotherms bind tightly enough to be fit to a model, although isotherms 2 through 5 do show some evidence of weak binding. An additional isotherm for the entire Kank1$_{595\text{-}720}$ sequence is on the right, which displays tight binding between Kank1 and LC8.

To provide further confirmation of binding stoichiometry we performed size exclusion chromatography - multiangle light scattering (SEC-MALS) experiments on Kank$_{595\text{-}720}$. First, examining the protein in isolation, we see a single dominant peak eluting from the SEC, suggesting a homogenous sample (Fig. 4.5b). The mass of the peak

determined from MALS is 18.1 kDa, close to the expected monomer mass of 16.4 kDa for the construct. To examine the LC8-Kank$_{595-720}$ complex, we mixed LC8 and Kank595-720 at an excess of LC8 and purified the LC8-Kank$_{595-720}$ complex by SEC, to isolate a sample of the bound complex. Unfortunately, we see some dissociation of this complex on the SEC-MALS column, evidenced by a long tail on the dominant peak, and a second peak that elutes at ~28 minutes, consistent with the expected elution time for LC8 on this column. However, a single dominant peak for the LC8-Kank$_{595-720}$ complex can still be seen (Fig. 4.5c). Fitting a mass for only this front peak returns a complex mass of 177 kDa, close to the expected mass for a 14:2 LC8:Kank$_{595-720}$ complex (181 kDa). The mass is not uniform along the peak however, due in part to the fact that the complex falls apart somewhat on the column, so it is difficult to use this metric to definitively determine binding stoichiometry. Nevertheless, this confirms expectations set by ITC experiments that the stoichiometry of binding is above 5:1, most strongly suggesting a 7:1 (14:2) complex.

Kank-LC8 binding is highly cooperative

To investigate the details of LC8-client binding and examine the five potential LC8-binding motifs in isolation, we synthesized peptides of each motif, and measured binding between each peptide and LC8 by ITC. To our surprise, we found that none of the motif peptides bound to LC8 at an affinity measurable by ITC. The isotherm for peptide 1 shows no evidence of binding at all, while isotherms for peptides 2 through 5 show only very weak binding, with isotherms 3 and 5 most closely approaching an acceptable binding curve (Fig. 4.4). All isotherms were collected at an LC8 concentration of 60 μM, indicating that the affinity between LC8 and peptides is likely well above 60 μM in all cases. This contrasts with the measured binding between LC8 and Kank1$_{595-720,}$ suggesting that LC8-Kank1 binding is strongly driven by cooperativity between multiple LC8 motifs within the Kank1 sequence.

To further investigate cooperativity of binding, we performed a series of analytical ultracentrifugation (AUC) experiments titrating LC8 into Kank1$_{595-720}$, which further confirmed the high degree of cooperativity in binding (Fig. 4.5a). At 1.25:1 LC8:Kank1, three AUC peaks appear, one corresponding roughly to the expected mass for free Kank1, one corresponding to some intermediate state at ~3 svedbergs, and one peak corresponding to a larger complex at ~5.5 svedbergs. As the titration progresses, the minor intermediate state and the peak corresponding to free Kank$_{595-720}$ both disappear.

Specifically, a shift away from a peak of excess $Kank_{595-720}$ and a towards a new peak for excess LC8 occurs between the 5:1 and 10:1 titration point, suggesting a binding stoichiometry between 5 and 10:1, and in agreement with the measured *n* value by ITC. The bound state peak appears to shift between the 5:1 and 10:1 titration point from 6 svedbergs up to 7. This movement could be due to a difference in bound structures at these titration points but is more likely due to a shift in equilibrium towards the bound state, as the simple sedimentation velocity model used will be impacted by the dynamic equilibrium between free and bound states, and therefore be shifted towards lower svedberg units for the complex the larger the unbound fraction of protein in the system.

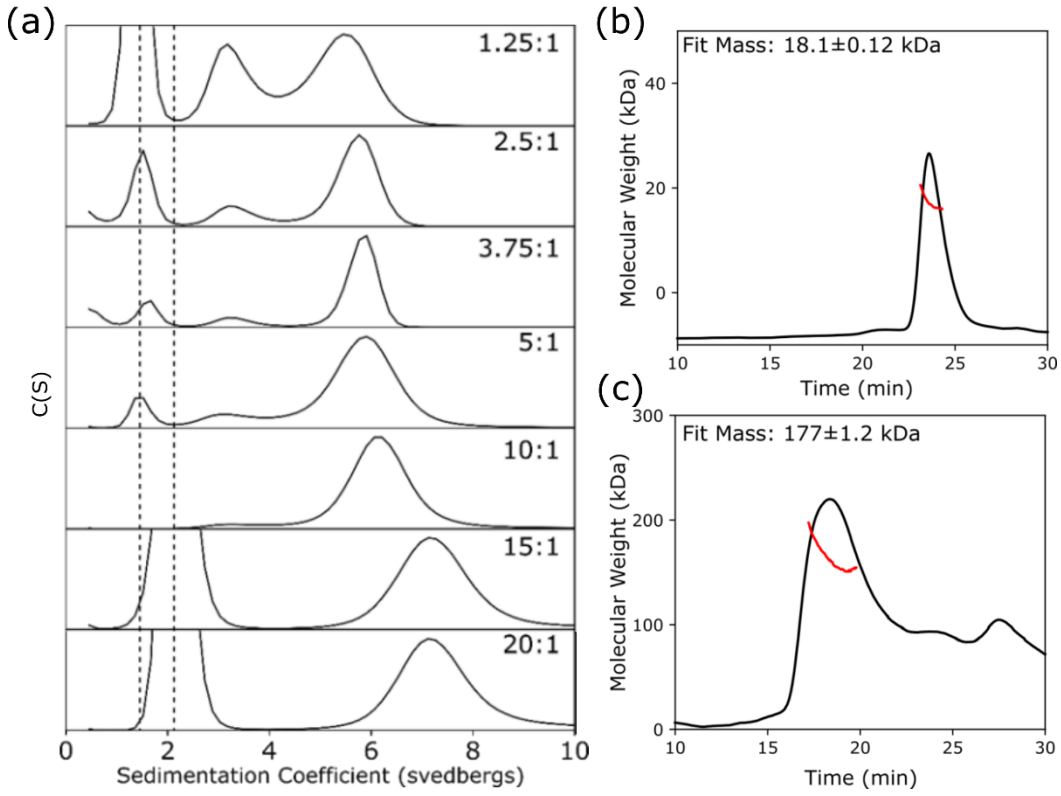


Figure 4.5: Characterization of the LC8-Kank1 complex. (a) A series of sedimentation velocity AUC experiments performed at 7.2 µM Kank1, and increasing ratios of LC8. Dotted lines show centers of AUC peaks for Kank1 and LC8 in absence of their binding partner. To account for variation in the measured wavelength at high LC8 concentrations, Y axes are normalized to the tallest bound state (<2.5 svedbergs) peak. (b,c) MALS of Kank1 (b) and Kank1 bound to LC8 (c). Red line represents measured mass, with the average fit mass listed at the top of each plot.

Discussion

Multivalency in LC8-Kank1 binding

Despite LC8Pred predicting a single LC8 site, Kank1 clearly binds LC8 multivalently. The MALS-determined mass of the LC8-Kank1 complex suggests a complex of 14:2 stoichiometry, with 7 dimers of LC8 bound to dimeric Kank1. An alternate possibility is that LC8 and Kank1 do not form a polybivalent complex, and instead form a trimeric or tetrameric (or higher order) structure. Indeed, several other combinations of $Kank_{595\text{-}720}$ and LC8 would exhibit a similar mass, such as a 12:3 LC8:Kank complex, with a theoretical mass of ~176 kDa. AUC and ITC results seem to confirm a binding stoichiometry between 10:2 and 14:2, however. The value of *n* from the ITC fit is near 6, and the $Kank_{595\text{-}720}$ peak in the AUC titration disappears between 5:1 and 10:1 LC8:Kank. While the stoichiometry determined by ITC does not perfectly match with the stoichiometry of the MALS data, this could be due either to error in the measured concentration of Kank595-720 used for ITC, or due to imprecision of the value of *n* in cases where all binding events are not energetically identical. Beyond these experiments, no evidence has yet appeared of a multivalent LC8-client complex that takes a non-polybivalent structure, suggesting that such binding appears to be the preferred structure for LC8 complexes[15,41,42,44]. Further, Kank1's LC8-binding domain directly follows a coiled-coil, which we believe will increase the favorability of a dimeric structure in the complex, as seen in the structure of the multivalent LC8-binding protein Nup159[15,16]. Based on these facts, we believe that the LC8-Kank1 complex is polybivalent, which raises an additional question. With 6 or 7 LC8 dimers bound, but only 5 sequences containing recognizable motifs, there must be 1-2 additional sites that have evaded notice. $Kank_{595\text{-}720}$ does not include any other sequence that resembles an LC8 motif, meaning that binding may be happening at a previously unrecognized motif sequence. While TQT (or TQT-like sequences as seen in Kank1) is established as the most common LC8-binding motif, new variations on the motif do occur, such as the LC8-binding site in MAG, discovered in 2019 with a TLT motif sequence[139].

Cooperativity in the Kank1 complex

The degree to which Kank1-LC8 binding is driven by cooperativity is remarkable among multivalent LC8-client binding, although Kank1 is not the first LC8-binding protein to exhibit cooperativity. In non-multivalent cases, proteins such as LC8-binding protein Swallow exhibit cooperativity driven by a coiled-coil or other dimerization domain – mutation of the

domain to a tighter dimer increases the affinity for LC8, indicating that an already-present dimer structure enhances LC8 binding[12]. Other multivalent cases such as ASCIZ and Nup159 (with 7 and 5 LC8-binding sites respectively) exhibit cooperativity as well – with an 'overall' binding affinity that is tighter than each motif's LC8 affinity in isolation[15,44]. None of these cases exhibit cooperativity above the scale of ~10fold enhancement however, while Kank1 appears to exhibit cooperativity in the scale of 2 orders of magnitude: $Kank_{595-720}$ binds LC8 with a 1.3 μM affinity, while each individual motif binds LC8 well above the limit of determination by ITC, near 60 μM for our experiments, suggesting a >50-fold enhancement of affinity. The exact source of this cooperativity is unclear, but prior investigation into multivalent LC8 complexes have suggested that the length of linkers between individual motifs plays a role in multivalent cooperativity[44,46]. The density of LC8 motifs within the 595-720 region of Kank1 is high in comparison to other multivalent complexes (e.g. the LC8-binding domain of dASCIZ is 145 residues and contains 7 motifs), suggesting that linkers between each motif must be short, which we believe will contribute to positive allostery.

Cooperativity is also apparent when examining the complex by AUC. Highly cooperative systems will disfavor states of intermediate occupancy, due to the relatively low affinity between LC8 and those low-occupancy states. Effectively, high cooperativity means that the system's equilibrium is balanced to favor either an apo or fully-bound state. For Kank1-LC8 binding, this is exactly what we see. The complex does appear to have a stable intermediate, but even at low concentrations of LC8, the saturated complex is the dominant bound species. This contrasts starkly with similar ultracentrifugation experiments on the LC8-binding protein ASCIZ, which, even at a significant excess of LC8, shows a mix of partially bound and saturated states[44]. While ASCIZ relies on this heterogeneous mix of states to sense the concentration of LC8[44], it seems likely that Kank1-LC8 binding plays a more structural role. Effectively, LC8 binding can act as a switchable element of structure. When unbound, Kank1 is flexible, allowing it to find its many binding partners in focal adhesions and cortical microtubule stabilizing complexes. When bound to LC8, the protein is rigid, holding the two complexes in place. This is the proposed mechanism for the LC8-binding protein Nup159, found in the nuclear pore, which bears several similarities to Kank1 in both structure, containing several coiled-coil domains, and function, as a scaffold in a large macromolecular complex[15,16].

Future directions

While this investigation reveals and characterizes a new LC8-binding interaction, it raises additional questions about the details of the LC8-Kank complex. The exact stoichiometry of binding is not yet clear, as calorimetric data conflicts with the MALS-determined mass. Confirmation of the mass of the complex through sedimentation equilibrium AUC experiments, would assist in determination of its stoichiometry. Additionally, negative stain electron microscopy, which has been a powerful tool for other multivalent LC8-client complexes[42,44], may provide confirmation of complex stoichiometry, as well as confirmation of whether the complex takes a polybivalent structure.

Related to the question of stoichiometry, the location of the one or two additional LC8 motifs in the $Kank_{595\text{-}720}$ sequence also remains a mystery. Investigations using isolated peptides of these sequences are likely to be of limited use, as we have already shown that the predicted LC8-binding motifs in Kank1 do not bind tightly to LC8. As an alternative to examining potential motifs in isolation, prior studies of multivalent LC8-binding proteins have disentangled many details of LC8 binding by examining an LC8-binding domain in fragments containing a subset of the total sequence. We believe this method would assist in elucidating the potential location of Kank1's additional motifs, as well as providing a more comprehensive picture of the thermodynamics of how Kank1-LC8 binding is so strongly driven by co-operativity. It also has the advantage of making Kank1 a more feasible target for nuclear magnetic resonance (NMR) spectroscopy – the $Kank_{595\text{-}720}$ construct contains 23 threonines and 16 serines, resulting in a highly degenerate NMR spectrum that cannot be fully assigned. Smaller constructs of Kank1 would doubtless be better suited to NMR analysis, which may allow for direct structural determination of each LC8-binding site on Kank1.

Finally, many questions remain about the larger context of LC8-Kank1 binding. While we propose that LC8 plays a structural role in the Kank1 complex, the exact function of Kank1-LC8 binding is unconfirmed. The function of Kank1 is not entirely clear either: while Kank1 is known to regulate cytoskeletal growth, it's unclear if this is the source of its tumor suppression activity. Therefore it is difficult to examine whether LC8 modulates Kank1's function as a cytoskeletal regulator, or is connected to some other, uncharacterized function of Kank1.

Conclusions

Within this work we have characterized a new LC8-binding interaction. LC8 binds the tumor suppressor and cytoskeletal regulator Kank1, multivalently and with high affinity. Further, the complex binds with a great degree of cooperativity, more than what is seen in other LC8-client complexes. Driven by this cooperativity, the Kank1-LC8 complex is homogenous in occupancy, rather than forming a mix of partially-bound intermediates. This fact points to the Kank1-LC8 complex playing a structural role, facilitating Kank1's function as a bridge between FAs and CMSCs.

Materials and Methods

Protein Expression and Purification

For all biophysical assays, LC8 and $Kank_{595-720}$ were expressed recombinantly in E. coli. Proteins were cloned into the pET24d expression vector, with an N-terminal 6xHis and TEV-cleavable site and expressed in either Rosetta (LC8) or C41 ($Kank_{595-720}$) E. coli cells, both of which are derived from BL21 DE3 cells. Cells were grown in ZYM5052 auto-induction media at 37 C for 24 hours. LC8 and $Kank_{595-720}$ were both purified by affinity chromatography with TALON cobalt resin, as previously described[44]. $Kank_{595-720}$ expresses into inclusion bodies and was therefore purified using buffers containing 6 M urea as previously described[44]. Following affinity purification, $Kank_{595-720}$ was dialyzed out of urea, and both $Kank_{595-720}$ and LC8 were further purified by size exclusion chromatography (SEC) using a Superdex S75 hi-load column (GE Healthcare). SEC was performed in a buffer of 25 mM Tris, pH 7.5, 150 mM NaCl, 5 mM β-mercaptoethanol, and 1 mM NaN_3. Purified proteins were stored in SEC buffer either used within a week or flash frozen to -80 C for storage.

Isothermal titration calorimetry

We performed isothermal titration calorimetry at 25 C with a VP-ITC microcalorimeter (Microcal) in SEC buffer. For Kank595-720, LC8 was titrated into a cell sample of Kank595-720, with a syringe and cell concentration of 400 and 4 μM respectively. Each injection had an 8 μL volume, and a total of 36 injections were performed. For Kank1 peptides, which were synthesized in-house using solid-phase synthesis and purified by HPLC, peptide was dissolved into SEC buffer at 500 μM and was titrated into a cell sample of LC8 at 60 μM. A total of 28 injections of peptide into LC8 were performed, at injection

volumes of 10 µL. Peaks were integrated and fit to the n-independent sites model in Origin 7.0.

Analytical Ultracentrifugation

We performed sedimentation velocity analytical ultracentrifugation (SV-AUC) on a Beckman Coulter Optima XL-A analytical ultracentrifuge, equipped with optics for absorbance measurements. We mixed purified LC8 and $Kank_{595-720}$ at a series of increasing concentrations, with a fixed $Kank_{595-720}$ concentration of 7.2 µM, and LC8 concentration varied from 9 µM (1.25:1) to 144 µM (20:1). The 1.25, 2.5, and 3.75:1 complexes were measured by absorbance at 280 nm, 5 and 10:1 used absorbance at 292 nm, and the 15 and 20:1 complexes used 298 nm. All experiments were performed in SEC buffer, dialyzed the night before SV-AUC was performed, to ensure minimal β-mercaptoethanol degradation. We loaded the complexes into 12 mm path-length 2-channel cells, and centrifuged the samples at 42,000 rpm and 20 C. We acquired 300 scans at the relevant wavelength, with no delay between scans. Absorbance profiles were fit to a c(S) distribution in SEDFIT, with a calculated buffer density of 1.0009 g/ml, calculated using Sednterp.

Size exclusion chromatography – multiangle light scattering

For size exclusion chromatography paired with multiangle light scattering (SEC-MALS), experiments were performed on a 10/300 Superdex 200 analytical SEC column (GE Healthcare) attached to an AKTA-FPLC (GE Healthcare) and routed through a DAWN multiangle light scattering and Optilab refractive index system (Wyatt Technology). We equilibrated the system to SEC buffer, then injected 100 µL of sample onto the column. For $Kank_{595-720}$ we injected a sample of 30 µM $Kank_{595-720}$, and for the complex, we injected a sample of a theoretical particle concentration (assuming 14:2 binding, i.e. 14 µM LC8 and 2 µM $Kank_{595-720}$) of 1 µM. Molar masses were estimated in the ASTRA software package, using a Zimm scattering model.

Cell culture and streptavidin pulldown assays

All cell culture, siRNA transfections, and streptavidin pulldowns were performed as described in Ammon et. al.,(2020)[136]. Briefly, HeLa stably expressing endogenous LC8-GFP cells were transfected with siRNA using HiPerFect (Quiagen), and analyzed 48-72

hours after transfection. For imaging, cells were fixed in ice-cold methanol, and proteins were visualized with commercially available primary antibodies for Kank1, Kank2 and Paxilin, along with fluorescently labeled secondary antibodies. Cells were imaged with a Nikon Eclipse Ni upright wide field fluorescence microscope and a Nikon DS-Qi2 CMOS camera (Nikon), using Plan Apo Lambda 100x N.A. 1.45 oil objective (Nikon) and Nikon NIS (Br) software (Nikon). For streptavidin pulldowns, briefly, bioGFP-tagged Kank1 constructs were expressed in HKE293T cells and harvested from cell lysates using streptavidin beads. Beads were then incubated with (separately prepared) lysates of cells expressing LC8-mCherry, washed, and the beads were stripped using a SDS-PAGE sample buffer. Gels of pulldown product were visualized using commercially available antibodies against GFP and mCherry.

Chapter 5

Conclusions

Impact

Each study in this thesis is focused on investigation of the function of the hub protein LC8, including investigation of the thermodynamics of LC8 binding, study of the LC8-binding motif, and characterization of a new LC8-binding protein. Chapters 2 and 3 are similar in that they both attempt to answer concrete questions about LC8 binding and provide tools and methods for future research. From these two chapters we have learned that LC8 binding is dependent on more than simply the TQT of the LC8 motif and that many LC8-binding clients bind with positive allostery. Both chapters also present new computational tools – chapter 2 outlines a method for Bayesian statistical modeling of multistep binding interactions, which we hope will be useful both for additional investigation of LC8 binding and broadly in the study of complex protein-protein interactions. Similarly, LC8Pred, the LC8-motif prediction method from chapter 3 is designed as an easy-to-use tool for investigation of new LC8-binding proteins which is already in heavy use by the LC8 binding community. Lastly, my investigation of Kank1-LC8 binding incorporates both LC8Pred as well as several biophysical techniques (SEC-MALS, ITC, AUC) to characterize a previously unknown binding interaction that demonstrates unprecedented binding cooperativity to a large number of sites and opens a new avenue of LC8 function through its interaction with Kank1. Together, this body of work emphasizes the value of combining experimental and computational approaches to maximize the information pulled from experimental data and provides tools for future researchers to do the same in their own work. This chapter briefly highlights important results from the work and discusses plans for ongoing projects and for future work.

Highlights of reported work

In chapter 2, I examine the thermodynamics of LC8-client binding, with a focus on the simple case of a peptide containing a single LC8 site. Although it is simplest LC8-binding system, the dimer-driven interaction necessitated the development of a Bayesian statistical method for modeling isothermal titration calorimetry (ITC) data using a two-step binding model, building on Bayesian approaches to simple ITC models. I dissected the impact that uncertainties in analyte concentration can have on thermodynamic parameters, demonstrating that even when analyte concentrations are highly uncertain, binding free energies can still be determined to within 1-2 kcal/mol uncertainty. I additionally demonstrated that the determinability of thermodynamic parameters,

particularly binding enthalpies, is intrinsically linked to the microscopic binding parameters, and that microscopic parameters that are poorly matched to experimental conditions result in loss of parameter determinability. Returning to LC8, I demonstrated with confidence that LC8 binds selected clients with positive allostery, indicating the half-bound LC8-client complex has a higher affinity for clients than apo LC8. While this may not be universal in LC8 binding, it fits with the conception of LC8 as an engine for dimerization: with positive allostery, the half-bound state is disfavored, and the fully bound induced-dimer state, thought to be the functional state for most LC8 interactions, is favored.

In chapter 3 I describe work investigating the LC8 motif. Newly characterized LC8-binding peptides, as well as a set of motif-containing peptides demonstrated to not bind LC8, provided data we collated into new insights on LC8 binding. Most notably, we found that the TQT anchor, while required for binding, is not a sufficient determinant of LC8 binding, which depends on both a favorable motif and favorable flanking residues. We collated a database of known LC8-interacting proteins and used this dataset, along with the LC8-nonbinding sequences, to build a set of scoring matrices for evaluating LC8 binding, titled LC8Pred. LC8Pred performed with 76% accuracy in validation, demonstrating a fair capacity to separate binding and non-binding sequences. Both the database of LC8-binding proteins (LC8Hub) and the predictive tool (LC8Pred) are maintained online, for anyone interested in studying LC8 binding.

Chapter 4 is focused on characterizing the interaction between LC8 and multivalent client Kank1. We first examine localization of LC8 and Kank1 in cells and find that Kank1 draws LC8 to focal adhesions at the cell cortex, where the proteins colocalize. Further, through pulldowns and predictions with LC8Pred, we were able to target Kank1's LC8-binding site to a region of disorder midway through the protein's sequence. Thermodynamic investigation of LC8-Kank1 binding revealed that Kank1 binds LC8 multivalently, at 6 or 7 sites. The complex is strongly driven by cooperativity, to the extent that the affinity for the complete LC8-binding region is at least two orders of magnitude tighter-binding than each motif in isolation. We believe that the resultant function of the LC8-kank1 complex plays a structural role, stabilizing the macromolecular complex Kank1 forms at the cell cortex.

Ongoing work and future directions

Modeling LC8 binding

The observed allostery in LC8 binding may not be universal, as we selected a tight-binding subset of known LC8-interacting peptides for our study in chapter 2. There is a growing library of data on other LC8-binding peptides available to us, however, and future studies will hopefully widen the net of examined LC8-interacting proteins. Of particular interest is an examination of LC8-binding interactions of seemingly varied entropy. Independent-sites fits of LC8 interactions report wide variations in binding entropy and enthalpy, suggesting there is a degree of entropy-enthalpy compensation occurring in LC8 binding. This correlates with measurements that show that the core of LC8 is rigidified on binding to clients, to a varying degree dependent on client sequence. This variation in entropy also provides a potential explanation for the structural mechanism of allostery, a question left unanswered in the work presented in this thesis. We hope that combining further modeling work with molecular dynamics simulations of LC8-client complexes will help answer these questions and move us towards a complete understanding of two-step LC8 binding.

LC8Hub

We plan to keep LC8Hub and LC8Pred up to date as research reveals new LC8-binding proteins. While scoring criteria will need to be re-evaluated as training data grows, LC8Pred is perfectly adaptable to new data – as new LC8-binding sequences are discovered, they can easily be added to the LC8Pred training sequences. Additionally, as the list of known LC8-binding proteins grows, it will be necessary to update our conception of how LC8 binding functions. As an example, work published in 2018 revealed that LC8 binds to the protein L-MAG through a currently unique TLT binding motif, providing evidence that a position 0 leucine is possible in LC8 binding. LC8Pred, which utilizes an initial filtering step looking for anchor-like sequences, was therefore adjusted accordingly. We expect that new discoveries of LC8-binding proteins will continue to appear, and as they are incorporated into LC8Hub and LC8Pred, the quality of LC8-binding prediction will improve.

Kank1

Several questions remain about the details of Kank1-LC8 binding. In particular, the exact number and location of Kank1's LC8 binding motifs stands out as a particularly important

unanswered question. We hope that additional studies on fragments of Kank1, alongside the application of additional techniques such as nuclear magnetic resonance spectroscopy and negative stain electron microscopy will provide answers to these questions. Beyond these questions, the function of LC8-Kank1 binding remains unknown, and investigating our hypothesis that the interaction plays a structural role will be an important next step for the study of the protein.

Multivalent binding

Multivalent complexes containing intrinsically disordered proteins (IDPs) present a unique biophysical challenge, due largely to their tendency to form heterogeneous ensembles of different states and the underlying thermodynamic complexity of multivalent interactions. Sample heterogeneity confounds many traditional biophysical investigation methods reliant on a homogenous sample. To combat this and aid the study of both LC8 and other multivalent IDP-containing complexes, we are working to develop methods of analysis tailored to multivalent binding.

Building on the work presented in chapter 2, we plan to expand our attempts at Bayesian modeling to multivalent LC8 complexes. An LC8 binding protein with two motifs can form thirteen structurally distinct complexes, with a complicated network of intermediate states. While we do not believe that a single isotherm will be sufficient to confidently model these entire systems, mutation studies examining each motif in isolation, then utilization of global models that incorporate isolated mutation studies will hopefully fill in the gaps. We additionally hope that the same models can be applied to experimental techniques beyond calorimetry, such as nuclear magnetic resonance (NMR) and analytical ultracentrifugation (AUC) experiments, to help fill in the gaps where calorimetric data does not provide a full picture.

Advancements in analysis methods have led native mass spectrometry to become a powerful tool for measuring heterogeneous multivalent complexes. LC8 binding provides an interesting test case for this technology, where protein complexes are preserved in flight and populations of occupancy states can be measured, and our work has provided impetus, samples and data towards its development. A recent manuscript[46] on LC8 binding to a region of the transcription factor ASCIZ demonstrated through native mass spectrometry that LC8 binds ASCIZ in a predominantly ‘in-register’ duplex, where LC8 is bound at symmetric motifs and induces a ladder-like structure on the client. The

manuscript demonstrates the power of native mass spectrometry to answer otherwise-unapproachable questions about multivalent IDP complexes, as they allow individual complexes to be picked out of a heterogeneous mix.

Negative stain electron microscopy also provides an exciting new avenue for investigating these complexes. As with native mass spectrometry, our work has provided impetus, samples and data for the development of negative stain electron microscopy. Each dimer unit of LC8 appears as a small dot in electron micrographs, and multivalent LC8 complexes appear as a series of beads on a string. While images can be clustered into classes with a fixed structure as is traditional in electron microscopy, this oversimplifies the true heterogeneity of multivalent complexes, which retain some structural flexibility. As such, we have worked to assist in the development of a method of EM analysis that utilizes individual images to build a conformational ensemble of the complex structure[42]. I contributed the design and testing of a synthetic LC8-binding peptide to this work, which is the subject of appendix 5 of this thesis. Analysis of complexes by this method allows us to both determine the occupancy of LC8-client complexes and build an ensemble of their structures.

While these methods have all been focused on investigating LC8 binding, it is our hope that they will be applicable to other multivalent systems. Multivalent binding is a common feature to many IDPs, and multivalent IDP binding is a growing area of study, as these complexes complexes play a role in a host of biological functions. Proper tools that account for the heterogeneity of structure and conformation that are characteristic of multivalency will simplify investigation of these systems, both in future studies on LC8-binding proteins, and many other multivalent systems.

References

(1) Jones, S.; Thornton, J. M. Principles of Protein-Protein Interactions. *Proceedings of the National Academy of Sciences* **1996**, *93* (1), 13–20. https://doi.org/10.1073/pnas.93.1.13.

(2) Diversity of Protein–Protein Interactions. *The EMBO Journal* **2003**, *22* (14), 3486–3492. https://doi.org/10.1093/emboj/cdg359.

(3) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*, 4th ed.; Garland Science, 2002.

(4) Yu, H.; Braun, P.; Yıldırım, M. A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, J.-F.; Dricot, A.; Vazquez, A.; Murray, R. R.; Simon, C.; Tardivo, L.; Tam, S.; Svrzikapa, N.; Fan, C.; de Smet, A.-S.; Motyl, A.; Hudson, M. E.; Park, J.; Xin, X.; Cusick, M. E.; Moore, T.; Boone, C.; Snyder, M.; Roth, F. P.; Barabási, A.-L.; Tavernier, J.; Hill, D. E.; Vidal, M. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **2008**, *322* (5898), 104–110. https://doi.org/10.1126/science.1158684.

(5) Komurov, K.; White, M. Revealing Static and Dynamic Modular Architecture of the Eukaryotic Protein Interaction Network. *Molecular Systems Biology* **2007**, *3* (1), 110. https://doi.org/10.1038/msb4100149.

(6) Jespersen, N.; Barbar, E. Emerging Features of Linear Motif-Binding Hub Proteins. *Trends Biochem Sci* **2020**, *45* (5), 375–384. https://doi.org/10.1016/j.tibs.2020.01.004.

(7) Dudley, A. M.; Janse, D. M.; Tanay, A.; Shamir, R.; Church, G. M. A Global View of Pleiotropy and Phenotypically Derived Gene Function in Yeast. *Mol Syst Biol* **2005**, *1*, 2005.0001. https://doi.org/10.1038/msb4100004.

(8) Jespersen, N.; Estelle, A.; Waugh, N.; Davey, N. E.; Blikstad, C.; Ammon, Y.-C.; Akhmanova, A.; Ivarsson, Y.; Hendrix, D. A.; Barbar, E. Systematic Identification of Recognition Motifs for the Hub Protein LC8. *Life Sci Alliance* **2019**, *2* (4), e201900366. https://doi.org/10.26508/lsa.201900366.

(9) Barbar, E. Dynein Light Chain LC8 Is a Dimerization Hub Essential in Diverse Protein Networks. *Biochemistry* **2008**, *47* (2), 503–508. https://doi.org/10.1021/bi701995m.

(10) Clark, S.; Nyarko, A.; Löhr, F.; Karplus, P. A.; Barbar, E. The Anchored Flexibility Model in LC8 Motif Recognition: Insights from the Chica Complex. *Biochemistry* **2016**, *55* (1), 199–209. https://doi.org/10.1021/acs.biochem.5b01099.

(11) Wang, L.; Hare, M.; Hays, T. S.; Barbar, E. Dynein Light Chain LC8 Promotes Assembly of the Coiled-Coil Domain of Swallow Protein. *Biochemistry* **2004**, *43* (15), 4611–4620. https://doi.org/10.1021/bi036328x.

(12) Kidane, A. I.; Song, Y.; Nyarko, A.; Hall, J.; Hare, M.; Löhr, F.; Barbar, E. Structural Features of LC8-Induced Self Association of Swallow†. *Biochemistry* **2013**, *52* (35), 10.1021/bi400642u. https://doi.org/10.1021/bi400642u.

(13) Erdős, G.; Szaniszló, T.; Pajkos, M.; Hajdu-Soltész, B.; Kiss, B.; Pál, G.; Nyitray, L.; Dosztányi, Z. Novel Linear Motif Filtering Protocol Reveals the Role of the LC8 Dynein Light Chain in the Hippo Pathway. *PLOS Computational Biology* **2017**, *13* (12), e1005885. https://doi.org/10.1371/journal.pcbi.1005885.

(14) Rayala, S. K.; den Hollander, P.; Manavathi, B.; Talukder, A. H.; Song, C.; Peng, S.; Barnekow, A.; Kremerskothen, J.; Kumar, R. Essential Role of KIBRA in Co-Activator Function of Dynein Light Chain 1 in Mammalian Cells. *J Biol Chem* **2006**, *281* (28), 19092–19099. https://doi.org/10.1074/jbc.M600021200.

(15) Nyarko, A.; Song, Y.; Nováček, J.; Žídek, L.; Barbar, E. Multiple Recognition Motifs in Nucleoporin Nup159 Provide a Stable and Rigid Nup159-Dyn2 Assembly. *Journal of Biological Chemistry* **2013**, *288* (4), 2614–2622. https://doi.org/10.1074/jbc.M112.432831.
(16) Gaik, M.; Flemming, D.; von Appen, A.; Kastritis, P.; Mücke, N.; Fischer, J.; Stelter, P.; Ori, A.; Bui, K. H.; Baßler, J.; Barbar, E.; Beck, M.; Hurt, E. Structural Basis for Assembly and Function of the Nup82 Complex in the Nuclear Pore Scaffold. *Journal of Cell Biology* **2015**, *208* (3), 283–297. https://doi.org/10.1083/jcb.201411003.
(17) Fejtova, A.; Davydova, D.; Bischof, F.; Lazarevic, V.; Altrock, W. D.; Romorini, S.; Schöne, C.; Zuschratter, W.; Kreutz, M. R.; Garner, C. C.; Ziv, N. E.; Gundelfinger, E. D. Dynein Light Chain Regulates Axonal Trafficking and Synaptic Levels of Bassoon. *J Cell Biol* **2009**, *185* (2), 341–355. https://doi.org/10.1083/jcb.200807155.
(18) Hall, J.; Karplus, P. A.; Barbar, E. Multivalency in the Assembly of Intrinsically Disordered Dynein Intermediate Chain. *Journal of Biological Chemistry* **2009**, *284* (48), 33115–33121. https://doi.org/10.1074/jbc.M109.048587.
(19) Jespersen, N. E.; Leyrat, C.; Gérard, F. C.; Bourhis, J.-M.; Blondel, D.; Jamin, M.; Barbar, E. The LC8-RavP Ensemble Structure Evinces A Role for LC8 in Regulating Lyssavirus Polymerase Functionality. *Journal of Molecular Biology* **2019**, *431* (24), 4959–4977. https://doi.org/10.1016/j.jmb.2019.10.011.
(20) Kubota, T.; Matsuoka, M.; Chang, T.-H.; Bray, M.; Jones, S.; Tashiro, M.; Kato, A.; Ozato, K. Ebolavirus VP35 Interacts with the Cytoplasmic Dynein Light Chain 8. *J Virol* **2009**, *83* (13), 6952–6956. https://doi.org/10.1128/JVI.00480-09.
(21) Becker, J. R.; Cuella-Martin, R.; Barazas, M.; Liu, R.; Oliveira, C.; Oliver, A. W.; Bilham, K.; Holt, A. B.; Blackford, A. N.; Heierhorst, J.; Jonkers, J.; Rottenberg, S.; Chapman, J. R. The ASCIZ-DYNLL1 Axis Promotes 53BP1-Dependent Non-Homologous End Joining and PARP Inhibitor Sensitivity. *Nat Commun* **2018**, *9* (1), 5406. https://doi.org/10.1038/s41467-018-07855-x.
(22) Gupta, A.; Diener, D. R.; Sivadas, P.; Rosenbaum, J. L.; Yang, P. The Versatile Molecular Complex Component LC8 Promotes Several Distinct Steps of Flagellar Assembly. *Journal of Cell Biology* **2012**, *198* (1), 115–126. https://doi.org/10.1083/jcb.201111041.
(23) Lajoix, A.-D.; Gross, R.; Aknin, C.; Dietz, S.; Granier, C.; Laune, D. Cellulose Membrane Supported Peptide Arrays for Deciphering Protein-Protein Interaction Sites: The Case of PIN, a Protein with Multiple Natural Partners. *Mol Divers* **2004**, *8* (3), 281–290. https://doi.org/10.1023/b:modi.0000036242.01129.27.
(24) Lee, K. H.; Lee, S.; Kim, B.; Chang, S.; Kim, S. W.; Paick, J.-S.; Rhee, K. Dazl Can Bind to Dynein Motor Complex and May Play a Role in Transport of Specific MRNAs. *EMBO J* **2006**, *25* (18), 4263–4270. https://doi.org/10.1038/sj.emboj.7601304.
(25) King, A.; Li, L.; Wong, D. M.; Liu, R.; Bamford, R.; Strasser, A.; Tarlinton, D. M.; Heierhorst, J. Dynein Light Chain Regulates Adaptive and Innate B Cell Development by Distinctive Genetic Mechanisms. *PLOS Genetics* **2017**, *13* (9), e1007010. https://doi.org/10.1371/journal.pgen.1007010.
(26) Dick, T.; Ray, K.; Salz, H. K.; Chia, W. Cytoplasmic Dynein (Ddlc1) Mutations Cause Morphogenetic Defects and Apoptotic Cell Death in Drosophila Melanogaster. *Molecular and Cellular Biology* **1996**, *16* (5), 1966–1977. https://doi.org/10.1128/MCB.16.5.1966.
(27) Goggolidou, P.; Hadjirin, N. F.; Bak, A.; Papakrivopoulou, E.; Hilton, H.; Norris, D. P.; Dean, C. H. Atmin Mediates Kidney Morphogenesis by Modulating Wnt Signaling. *Hum Mol Genet* **2014**, *23* (20), 5303–5316. https://doi.org/10.1093/hmg/ddu246.

(28) Pfister, K. K.; Shah, P. R.; Hummerich, H.; Russ, A.; Cotton, J.; Annuar, A. A.; King, S. M.; Fisher, E. M. C. Genetic Analysis of the Cytoplasmic Dynein Subunit Families. *PLoS Genet* **2006**, *2* (1), e1. https://doi.org/10.1371/journal.pgen.0020001.
(29) Benison, G.; Karplus, P. A.; Barbar, E. Structure and Dynamics of LC8 Complexes with KXTQT-Motif Peptides: Swallow and Dynein Intermediate Chain Compete for a Common Site. *J Mol Biol* **2007**, *371* (2), 457–468. https://doi.org/10.1016/j.jmb.2007.05.046.
(30) Lo, K. W.-H.; Kan, H.-M.; Chan, L.-N.; Xu, W.-G.; Wang, K.-P.; Wu, Z.; Sheng, M.; Zhang, M. The 8-KDa Dynein Light Chain Binds to P53-Binding Protein 1 and Mediates DNA Damage-Induced P53 Nuclear Accumulation *. *Journal of Biological Chemistry* **2005**, *280* (9), 8172–8179. https://doi.org/10.1074/jbc.M411408200.
(31) Schnabl, J.; Wang, J.; Hohmann, U.; Gehre, M.; Batki, J.; Andreev, V. I.; Purkhauser, K.; Fasching, N.; Duchek, P.; Novatchkova, M.; Mechtler, K.; Plaschka, C.; Patel, D. J.; Brennecke, J. Molecular Principles of Piwi-Mediated Cotranscriptional Silencing through the Dimeric SFiNX Complex. *Genes Dev.* **2021**, *35* (5–6), 392–409. https://doi.org/10.1101/gad.347989.120.
(32) Eastwood, E. L.; Jara, K. A.; Bornelöv, S.; Munafò, M.; Frantzis, V.; Kneuss, E.; Barbar, E. J.; Czech, B.; Hannon, G. J. Dimerisation of the PICTS Complex via LC8/Cut-up Drives Co-Transcriptional Transposon Silencing in Drosophila. *Elife* **2021**, *10*, e65557. https://doi.org/10.7554/eLife.65557.
(33) Schnorrer, F.; Bohmann, K.; Nüsslein-Volhard, C. The Molecular Motor Dynein Is Involved in Targeting Swallow and Bicoid RNA to the Anterior Pole of Drosophila Oocytes. *Nat Cell Biol* **2000**, *2* (4), 185–190. https://doi.org/10.1038/35008601.
(34) Raux, H.; Flamand, A.; Blondel, D. Interaction of the Rabies Virus P Protein with the LC8 Dynein Light Chain. *Journal of Virology* **2000**, *74* (21), 10212–10216. https://doi.org/10.1128/JVI.74.21.10212-10216.2000.
(35) Tan, G. S.; Preuss, M. A. R.; Williams, J. C.; Schnell, M. J. The Dynein Light Chain 8 Binding Motif of Rabies Virus Phosphoprotein Promotes Efficient Viral Transcription. *Proc Natl Acad Sci U S A* **2007**, *104* (17), 7229–7234. https://doi.org/10.1073/pnas.0701397104.
(36) Rapali, P.; Szenes, Á.; Radnai, L.; Bakos, A.; Pál, G.; Nyitray, L. DYNLL/LC8: A Light Chain Subunit of the Dynein Motor Complex and Beyond. *The FEBS Journal* **2011**, *278* (17), 2980–2996. https://doi.org/10.1111/j.1742-4658.2011.08254.x.
(37) Slevin, L. K.; Romes, E. M.; Dandulakis, M. G.; Slep, K. C. The Mechanism of Dynein Light Chain LC8-Mediated Oligomerization of the Ana2 Centriole Duplication Factor. *J Biol Chem* **2014**, *289* (30), 20727–20739. https://doi.org/10.1074/jbc.M114.576041.
(38) Benison, G.; Karplus, P. A.; Barbar, E. The Interplay of Ligand Binding and Quaternary Structure in the Diverse Interactions of Dynein Light Chain LC8. *Journal of Molecular Biology* **2008**, *384* (4), 954–966. https://doi.org/10.1016/j.jmb.2008.09.083.
(39) Nyarko, A.; Hall, J.; Hall, A.; Hare, M.; Kremerskothen, J.; Barbar, E. Conformational Dynamics Promote Binding Diversity of Dynein Light Chain LC8. *Biophysical Chemistry* **2011**, *159* (1), 41–47. https://doi.org/10.1016/j.bpc.2011.05.001.
(40) Hall, J.; Hall, A.; Pursifull, N.; Barbar, E. Differences in Dynamic Structure of LC8 Monomer, Dimer, and Dimer−Peptide Complexes. *Biochemistry* **2008**, *47* (46), 11940–11952. https://doi.org/10.1021/bi801093k.
(41) Clark, S. A.; Jespersen, N.; Woodward, C.; Barbar, E. Multivalent IDP Assemblies: Unique Properties of LC8-Associated, IDP Duplex Scaffolds. *FEBS Letters* **2015**, *589* (19, Part A), 2543–2551. https://doi.org/10.1016/j.febslet.2015.07.032.

(42) Mostofian, B.; McFarland, R.; Estelle, A.; Howe, J.; Barbar, E.; Reichow, S. L.; Zuckerman, D. M. Continuum Dynamics and Statistical Correction of Compositional Heterogeneity in Multivalent IDP Oligomers Resolved by Single-Particle EM. *Journal of Molecular Biology* **2022**, *434* (9), 167520. https://doi.org/10.1016/j.jmb.2022.167520.
(43) Dunsch, A. K.; Hammond, D.; Lloyd, J.; Schermelleh, L.; Gruneberg, U.; Barr, F. A. Dynein Light Chain 1 and a Spindle-Associated Adaptor Promote Dynein Asymmetry and Spindle Orientation. *Journal of Cell Biology* **2012**, *198* (6), 1039–1054. https://doi.org/10.1083/jcb.201202112.
(44) Clark, S.; Myers, J. B.; King, A.; Fiala, R.; Novacek, J.; Pearce, G.; Heierhorst, J.; Reichow, S. L.; Barbar, E. J. Multivalency Regulates Activity in an Intrinsically Disordered Transcription Factor. *eLife* **2018**, *7*, e36258. https://doi.org/10.7554/eLife.36258.
(45) Rapali, P.; García-Mayoral, M. F.; Martínez-Moreno, M.; Tárnok, K.; Schlett, K.; Albar, J. P.; Bruix, M.; Nyitray, L.; Rodriguez-Crespo, I. LC8 Dynein Light Chain (DYNLL1) Binds to the C-Terminal Domain of ATM-Interacting Protein (ATMIN/ASCIZ) and Regulates Its Subcellular Localization. *Biochem Biophys Res Commun* **2011**, *414* (3), 493–498. https://doi.org/10.1016/j.bbrc.2011.09.093.
(46) Reardon, P. N.; Jara, K. A.; Rolland, A. D.; Smith, D. A.; Hoang, H. T. M.; Prell, J. S.; Barbar, E. J. The Dynein Light Chain 8 (LC8) Binds Predominantly "in-Register" to a Multivalent Intrinsically Disordered Partner. *Journal of Biological Chemistry* **2020**, *295* (15), 4912–4922. https://doi.org/10.1074/jbc.RA119.011653.
(47) Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L.-N. Rapid Measurement of Binding Constants and Heats of Binding Using a New Titration Calorimeter. *Analytical Biochemistry* **1989**, *179* (1), 131–137. https://doi.org/10.1016/0003-2697(89)90213-3.
(48) Schneider, R.; Blackledge, M.; Jensen, M. R. Elucidating Binding Mechanisms and Dynamics of Intrinsically Disordered Protein Complexes Using NMR Spectroscopy. *Current Opinion in Structural Biology* **2019**, *54*, 10–18. https://doi.org/10.1016/j.sbi.2018.09.007.
(49) Weng, J.; Wang, W. Dynamic Multivalent Interactions of Intrinsically Disordered Proteins. *Current Opinion in Structural Biology* **2020**, *62*, 9–13. https://doi.org/10.1016/j.sbi.2019.11.001.
(50) West, K. L.; Kelliher, J. L.; Xu, Z.; An, L.; Reed, M. R.; Eoff, R. L.; Wang, J.; Huen, M. S. Y.; Leung, J. W. C. LC8/DYNLL1 Is a 53BP1 Effector and Regulates Checkpoint Activation. *Nucleic Acids Research* **2019**, *47* (12), 6236–6249. https://doi.org/10.1093/nar/gkz263.
(51) Luthra, P.; Jordan, D. S.; Leung, D. W.; Amarasinghe, G. K.; Basler, C. F. Ebola Virus VP35 Interaction with Dynein LC8 Regulates Viral RNA Synthesis. *Journal of Virology* **2015**, *89* (9), 5148–5153. https://doi.org/10.1128/JVI.03652-14.
(52) Rodriguez Galvan, J.; Donner, B.; Veseley, C. H.; Reardon, P.; Forsythe, H. M.; Howe, J.; Fujimura, G.; Barbar, E. Human Parainfluenza Virus 3 Phosphoprotein Is a Tetramer and Shares Structural and Interaction Features with Ebola Phosphoprotein VP35. *Biomolecules* **2021**, *11* (11), 1603. https://doi.org/10.3390/biom11111603.
(53) MicroCal. Data Analysis in Origin, 1998.
(54) Nguyen, T. H.; Rustenburg, A. S.; Krimmer, S. G.; Zhang, H.; Clark, J. D.; Novick, P. A.; Branson, K.; Pande, V. S.; Chodera, J. D.; Minh, D. D. L. Bayesian Analysis of Isothermal Titration Calorimetry for Binding Thermodynamics. *PLOS ONE* **2018**, *13* (9), e0203224. https://doi.org/10.1371/journal.pone.0203224.
(55) Freiburger, L.; Auclair, K.; Mittermaier, A. Global ITC Fitting Methods in Studies of Protein Allostery. *Methods* **2015**, *76*, 149–161. https://doi.org/10.1016/j.ymeth.2014.12.018.

(56) Zhao, H.; Piszczek, G.; Schuck, P. SEDPHAT – a Platform for Global ITC Analysis and Global Multi-Method Analysis of Molecular Interactions. *Methods* **2015**, *76*, 137–148. https://doi.org/10.1016/j.ymeth.2014.11.012.
(57) Feng, C.; Roy, A.; Post, C. B. Entropic Allostery Dominates the Phosphorylation-Dependent Regulation of Syk Tyrosine Kinase Release from Immunoreceptor Tyrosine-Based Activation Motifs. *Protein Science* **2018**, *27* (10), 1780–1796. https://doi.org/10.1002/pro.3489.
(58) Lee, A. L.; Sapienza, P. J. Thermodynamic and NMR Assessment of Ligand Cooperativity and Intersubunit Communication in Symmetric Dimers: Application to Thymidylate Synthase. *Frontiers in Molecular Biosciences* **2018**, *5*.
(59) Chodera, J. D.; Mobley, D. L. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu Rev Biophys* **2013**, *42*, 121–142. https://doi.org/10.1146/annurev-biophys-083012-130318.
(60) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation Changes upon Ligand Binding to Trypsin and Thrombin: Structural Interpretation Based on PKa Calculations and ITC Experiments. *Journal of Molecular Biology* **2007**, *367* (5), 1347–1356. https://doi.org/10.1016/j.jmb.2007.01.022.
(61) Steuber, H.; Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Tracing Changes in Protonation: A Prerequisite to Factorize Thermodynamic Data of Inhibitor Binding to Aldose Reductase. *Journal of Molecular Biology* **2007**, *373* (5), 1305–1320. https://doi.org/10.1016/j.jmb.2007.08.063.
(62) Leavitt, S.; Freire, E. Direct Measurement of Protein Binding Energetics by Isothermal Titration Calorimetry. *Current Opinion in Structural Biology* **2001**, *11* (5), 560–566. https://doi.org/10.1016/S0959-440X(00)00248-7.
(63) Velazquez-Campoy, A.; Kiso, Y.; Freire, E. The Binding Energetics of First- and Second-Generation HIV-1 Protease Inhibitors: Implications for Drug Design. *Archives of Biochemistry and Biophysics* **2001**, *390* (2), 169–175. https://doi.org/10.1006/abbi.2001.2333.
(64) Boyce, S. E.; Tellinghuisen, J.; Chodera, J. D. *Avoiding Accuracy-Limiting Pitfalls in the Study of Protein-Ligand Interactions with Isothermal Titration Calorimetry*; preprint; Biochemistry, 2015. https://doi.org/10.1101/023796.
(65) Contreras-Martos, S.; Nguyen, H. H.; Nguyen, P. N.; Hristozova, N.; Macossay-Castillo, M.; Kovacs, D.; Bekesi, A.; Oemig, J. S.; Maes, D.; Pauwels, K.; Tompa, P.; Lebrun, P. Quantification of Intrinsically Disordered Proteins: A Problem Not Fully Appreciated. *Frontiers in Molecular Biosciences* **2018**, *5*.
(66) Myszka, D. G.; Abdiche, Y. N.; Arisaka, F.; Byron, O.; Eisenstein, E.; Hensley, P.; Thomson, J. A.; Lombardo, C. R.; Schwarz, F.; Stafford, W.; Doyle, M. L. The ABRF-MIRG'02 Study: Assembly State, Thermodynamic, and Kinetic Analysis of an Enzyme/Inhibitor Interaction. *J Biomol Tech* **2003**, *14* (4), 247–269.
(67) MicroCal. VP-ITC Users Manual, 2001.
(68) Houtman, J. C. D.; Brown, P. H.; Bowden, B.; Yamaguchi, H.; Appella, E.; Samelson, L. E.; Schuck, P. Studying Multisite Binary and Ternary Protein Interactions by Global Analysis of Isothermal Titration Calorimetry Data in SEDPHAT: Application to Adaptor Protein Complexes in Cell Signaling. *Protein Science* **2007**, *16* (1), 30–42. https://doi.org/10.1110/ps.062558507.
(69) Cardoso, M. V. C.; Rivera, J. D.; Vitale, P. A. M.; Degenhardt, M. F. S.; Abiko, L. A.; Oliveira, C. L. P.; Salinas, R. K. CALX-CBD1 Ca2+-Binding Cooperativity Studied by NMR Spectroscopy

and ITC with Bayesian Statistics. *Biophysical Journal* **2020**, *119* (2), 337–348. https://doi.org/10.1016/j.bpj.2020.05.031.

(70) Duvvuri, H.; Wheeler, L. C.; Harms, M. J. Pytc: Open-Source Python Software for Global Analyses of Isothermal Titration Calorimetry Data. *Biochemistry* **2018**, *57* (18), 2578–2583. https://doi.org/10.1021/acs.biochem.7b01264.

(71) Nyarko, A.; Song, Y.; Barbar, E. Intrinsic Disorder in Dynein Intermediate Chain Modulates Its Interactions with NudE and Dynactin. *Journal of Biological Chemistry* **2012**, *287* (30), 24884–24893. https://doi.org/10.1074/jbc.M112.376038.

(72) Benison, G.; Chiodo, M.; Karplus, P. A.; Barbar, E. Structural, Thermodynamic, and Kinetic Effects of a Phosphomimetic Mutation in Dynein Light Chain LC8. *Biochemistry* **2009**, *48* (48), 11381–11389. https://doi.org/10.1021/bi901589w.

(73) Anthis, N. J.; Clore, G. M. Sequence-Specific Determination of Protein and Peptide Concentrations by Absorbance at 205 Nm. *Protein Sci* **2013**, *22* (6), 851–858. https://doi.org/10.1002/pro.2253.

(74) Jie, J.; Löhr, F.; Barbar, E. Interactions of Yeast Dynein with Dynein Light Chain and Dynactin: GENERAL IMPLICATIONS FOR INTRINSICALLY DISORDERED DUPLEX SCAFFOLDS IN MULTIPROTEIN ASSEMBLIES. *Journal of Biological Chemistry* **2015**, *290* (39), 23863–23874. https://doi.org/10.1074/jbc.M115.649715.

(75) Jie, J. *Protien Disorder in the Evolution of Dynein Regulation.*; Oregon State University, 2016.

(76) Turnbull, W. B.; Daranas, A. H. On the Value of c: Can Low Affinity Systems Be Studied by Isothermal Titration Calorimetry? *J. Am. Chem. Soc.* **2003**, *125* (48), 14859–14866. https://doi.org/10.1021/ja036166s.

(77) Manneville, J.-B.; Jehanno, M.; Etienne-Manneville, S. Dlg1 Binds GKAP to Control Dynein Association with Microtubules, Centrosome Positioning, and Cell Polarity. *Journal of Cell Biology* **2010**, *191* (3), 585–598. https://doi.org/10.1083/jcb.201002151.

(78) Sørensen, C. S.; Jendroszek, A.; Kjaergaard, M. Linker Dependence of Avidity in Multivalent Interactions Between Disordered Proteins. *Journal of Molecular Biology* **2019**, *431* (24), 4784–4795. https://doi.org/10.1016/j.jmb.2019.09.001.

(79) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. Emcee : The MCMC Hammer. *Publications of the Astronomical Society of the Pacific* **2013**, *125* (925), 306–312. https://doi.org/10.1086/670067.

(80) Freire, E.; Schön, A.; Velazquez-Campoy, A. Chapter 5 Isothermal Titration Calorimetry: General Formalism Using Binding Polynomials. In *Methods in Enzymology*; Biothermodynamics, Part A; Academic Press, 2009; Vol. 455, pp 127–155. https://doi.org/10.1016/S0076-6879(08)04205-5.

(81) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. Bayesian Data Analysis Third Edition. 677.

(82) Hogg, D. W.; Foreman-Mackey, D. Data Analysis Recipes: Using Markov Chain Monte Carlo. *ApJS* **2018**, *236* (1), 11. https://doi.org/10.3847/1538-4365/aab76e.

(83) Bayes, T. An Essay towards Solving a Problem in the Doctrine of Chances. *Phil. Trans. of the Royal Soc. of London* **1763**, *53*, 370–418.

(84) Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57* (1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

(85) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21* (6), 1087–1092. https://doi.org/10.1063/1.1699114.

(86) Sharma, S. Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy. *Annual Review of Astronomy and Astrophysics* **2017**, *55* (1), 213–259. https://doi.org/10.1146/annurev-astro-082214-122339.
(87) Goodman, J.; Weare, J. Ensemble Samplers with Affine Invariance. *CAMCoS* **2010**, *5* (1), 65–80. https://doi.org/10.2140/camcos.2010.5.65.
(88) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabási, A.-L. The Large-Scale Organization of Metabolic Networks. *Nature* **2000**, *407* (6804), 651–654. https://doi.org/10.1038/35036627.
(89) Jeong, H.; Mason, S. P.; Barabási, A.-L.; Oltvai, Z. N. Lethality and Centrality in Protein Networks. *Nature* **2001**, *411* (6833), 41–42. https://doi.org/10.1038/35075138.
(90) Patil, A.; Kinoshita, K.; Nakamura, H. Hub Promiscuity in Protein-Protein Interaction Networks. *International Journal of Molecular Sciences* **2010**, *11* (4), 1930–1943. https://doi.org/10.3390/ijms11041930.
(91) Wu, X.; Guo, J.; Zhang, D.-Y.; Lin, K. The Properties of Hub Proteins in a Yeast-Aggregated Cell Cycle Network and Its Phase Sub-Networks. *PROTEOMICS* **2009**, *9* (20), 4812–4824. https://doi.org/10.1002/pmic.200900053.
(92) Aitken, A. 14-3-3 Proteins: A Historic Overview. *Seminars in Cancer Biology* **2006**, *16* (3), 162–172. https://doi.org/10.1016/j.semcancer.2006.03.005.
(93) Uchikoga, N.; Matsuzaki, Y.; Ohue, M.; Akiyama, Y. Specificity of Broad Protein Interaction Surfaces for Proteins with Multiple Binding Partners. *Biophysics and Physicobiology* **2016**, *13*, 105–115. https://doi.org/10.2142/biophysico.13.0_105.
(94) Uhart, M.; Flores, G.; Bustos, D. M. Controllability of Protein-Protein Interaction Phosphorylation-Based Networks: Participation of the Hub 14-3-3 Protein Family. *Sci Rep* **2016**, *6* (1), 26234. https://doi.org/10.1038/srep26234.
(95) Calderone, A.; Castagnoli, L.; Cesareni, G. Mentha: A Resource for Browsing Integrated Protein-Interaction Networks. *Nat Methods* **2013**, *10* (8), 690–691. https://doi.org/10.1038/nmeth.2561.
(96) Makokha, M.; Hare, M.; Li, M.; Hays, T.; Barbar, E. Interactions of Cytoplasmic Dynein Light Chains Tctex-1 and LC8 with the Intermediate Chain IC74. *Biochemistry* **2002**, *41* (13), 4302–4311. https://doi.org/10.1021/bi011970h.
(97) Benison, G.; Nyarko, A.; Barbar, E. Heteronuclear NMR Identifies a Nascent Helix in Intrinsically Disordered Dynein Intermediate Chain: Implications for Folding and Dimerization. *Journal of Molecular Biology* **2006**, *362* (5), 1082–1093. https://doi.org/10.1016/j.jmb.2006.08.006.
(98) Nyarko, A.; Barbar, E. Light Chain-Dependent Self-Association of Dynein Intermediate Chain*. *Journal of Biological Chemistry* **2011**, *286* (2), 1556–1566. https://doi.org/10.1074/jbc.M110.171686.
(99) Jurado, S.; Conlan, L. A.; Baker, E. K.; Ng, J.-L.; Tenis, N.; Hoch, N. C.; Gleeson, K.; Smeets, M.; Izon, D.; Heierhorst, J. ATM Substrate Chk2-Interacting Zn2+ Finger (ASCIZ) Is a Bi-Functional Transcriptional Activator and Feedback Sensor in the Regulation of Dynein Light Chain (DYNLL1) Expression*. *Journal of Biological Chemistry* **2012**, *287* (5), 3156–3164. https://doi.org/10.1074/jbc.M111.306019.
(100) Zaytseva, O.; Tenis, N.; Mitchell, N.; Kanno, S.; Yasui, A.; Heierhorst, J.; Quinn, L. M. The Novel Zinc Finger Protein DASCIZ Regulates Mitosis in Drosophila via an Essential Role in Dynein Light-Chain Expression. *Genetics* **2014**, *196* (2), 443–453. https://doi.org/10.1534/genetics.113.159541.

(101) Petryszak, R.; Keays, M.; Tang, Y. A.; Fonseca, N. A.; Barrera, E.; Burdett, T.; Füllgrabe, A.; Fuentes, A. M.-P.; Jupp, S.; Koskinen, S.; Mannion, O.; Huerta, L.; Megy, K.; Snow, C.; Williams, E.; Barzine, M.; Hastings, E.; Weisser, H.; Wright, J.; Jaiswal, P.; Huber, W.; Choudhary, J.; Parkinson, H. E.; Brazma, A. Expression Atlas Update—an Integrated Database of Gene and Protein Expression in Humans, Animals and Plants. *Nucleic Acids Research* **2016**, *44* (D1), D746–D752. https://doi.org/10.1093/nar/gkv1045.

(102) Chen, Y.-M.; Gerwin, C.; Sheng, Z.-H. Dynein Light Chain LC8 Regulates Syntaphilin-Mediated Mitochondrial Docking in Axons. *J. Neurosci.* **2009**, *29* (30), 9429–9438. https://doi.org/10.1523/JNEUROSCI.1472-09.2009.

(103) Wang, X.; Olson, J. R.; Rasoloson, D.; Ellenbecker, M.; Bailey, J.; Voronina, E. Dynein Light Chain DLC-1 Promotes Localization and Function of the PUF Protein FBF-2 in Germline Progenitor Cells. *Development* **2016**, *143* (24), 4643–4653. https://doi.org/10.1242/dev.140921.

(104) Theerawatanasirikul, S.; Phecharat, N.; Prawettongsopon, C.; Chaicumpa, W.; Lekcharoensuk, P. Dynein Light Chain DYNLL1 Subunit Facilitates Porcine Circovirus Type 2 Intracellular Transports along Microtubules. *Arch Virol* **2017**, *162* (3), 677–686. https://doi.org/10.1007/s00705-016-3140-0.

(105) Rodríguez-Crespo, I.; Yélamos, B.; Roncal, F.; Albar, J. P.; Ortiz de Montellano, P. R.; Gavilanes, F. Identification of Novel Cellular Proteins That Bind to the LC8 Dynein Light Chain Using a Pepscan Technique. *FEBS Letters* **2001**, *503* (2–3), 135–141. https://doi.org/10.1016/S0014-5793(01)02718-1.

(106) Liang, J.; Jaffrey, S. R.; Guo, W.; Snyder, S. H.; Clardy, J. Structure of the PIN/LC8 Dimer with a Bound Peptide. *Nat Struct Mol Biol* **1999**, *6* (8), 735–740. https://doi.org/10.1038/11501.

(107) Fan, J.-S.; Zhang, Q.; Tochio, H.; Zhang, M. Backbone Dynamics of the 8 KDa Dynein Light Chain Dimer Reveals Molecular Basis of the Protein's Functional Diversity. *J Biomol NMR* **2002**, *23* (2), 103–114. https://doi.org/10.1023/A:1016332918178.

(108) Fraczkiewicz, R.; Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry* **1998**, *19* (3), 319–333. https://doi.org/10.1002/(SICI)1096-987X(199802)19:3<319::AID-JCC6>3.0.CO;2-W.

(109) Rapali, P.; Radnai, L.; Süveges, D.; Harmat, V.; Tölgyesi, F.; Wahlgren, W. Y.; Katona, G.; Nyitray, L.; Pál, G. Directed Evolution Reveals the Binding Motif Preference of the LC8/DYNLL Hub Protein and Predicts Large Numbers of Novel Binders in the Human Proteome. *PLOS ONE* **2011**, *6* (4), e18818. https://doi.org/10.1371/journal.pone.0018818.

(110) Chou, P. Y.; Fasman, G. D. Prediction of Protein Conformation. *Biochemistry* **1974**, *13* (2), 222–245. https://doi.org/10.1021/bi00699a002.

(111) Ashok Kumar, T. CFSSP: Chou and Fasman Secondary Structure Prediction Server. *Wide Spectrum* **2013**, *1* (9), 15–19. https://doi.org/10.5281/zenodo.50733.

(112) Ho, B. K.; Brasseur, R. The Ramachandran Plots of Glycine and Pre-Proline. *BMC Structural Biology* **2005**, *5* (1), 14. https://doi.org/10.1186/1472-6807-5-14.

(113) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf 2016: An Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules. *Nucleic Acids Res* **2016**, *44* (W1), W344-350. https://doi.org/10.1093/nar/gkw408.

(114) Brereton, A. E.; Karplus, P. A. Ensemblator v3: Robust Atom-Level Comparative Analyses and Classification of Protein Structure Ensembles. *Protein Science* **2018**, *27* (1), 41–50. https://doi.org/10.1002/pro.3249.

(115) Hein, M. Y.; Hubner, N. C.; Poser, I.; Cox, J.; Nagaraj, N.; Toyoda, Y.; Gak, I. A.; Weisswange, I.; Mansfeld, J.; Buchholz, F.; Hyman, A. A.; Mann, M. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **2015**, *163* (3), 712–723. https://doi.org/10.1016/j.cell.2015.09.053.

(116) Boldt, K.; van Reeuwijk, J.; Lu, Q.; Koutroumpas, K.; Nguyen, T.-M. T.; Texier, Y.; van Beersum, S. E. C.; Horn, N.; Willer, J. R.; Mans, D. A.; Dougherty, G.; Lamers, I. J. C.; Coene, K. L. M.; Arts, H. H.; Betts, M. J.; Beyer, T.; Bolat, E.; Gloeckner, C. J.; Haidari, K.; Hetterschijt, L.; Iaconis, D.; Jenkins, D.; Klose, F.; Knapp, B.; Latour, B.; Letteboer, S. J. F.; Marcelis, C. L.; Mitic, D.; Morleo, M.; Oud, M. M.; Riemersma, M.; Rix, S.; Terhal, P. A.; Toedt, G.; van Dam, T. J. P.; de Vrieze, E.; Wissinger, Y.; Wu, K. M.; Apic, G.; Beales, P. L.; Blacque, O. E.; Gibson, T. J.; Huynen, M. A.; Katsanis, N.; Kremer, H.; Omran, H.; van Wijk, E.; Wolfrum, U.; Kepes, F.; Davis, E. E.; Franco, B.; Giles, R. H.; Ueffing, M.; Russell, R. B.; Roepman, R. An Organelle-Specific Protein Landscape Identifies Novel Diseases and Molecular Mechanisms. *Nat Commun* **2016**, *7* (1), 11491. https://doi.org/10.1038/ncomms11491.

(117) Wang, J.; Vasaikar, S.; Shi, Z.; Greer, M.; Zhang, B. WebGestalt 2017: A More Comprehensive, Powerful, Flexible and Interactive Gene Set Enrichment Analysis Toolkit. *Nucleic Acids Research* **2017**, *45* (W1), W130–W137. https://doi.org/10.1093/nar/gkx356.

(118) Madeira, F.; Tinti, M.; Murugesan, G.; Berrett, E.; Stafford, M.; Toth, R.; Cole, C.; MacKintosh, C.; Barton, G. J. 14-3-3-Pred: Improved Methods to Predict 14-3-3-Binding Phosphopeptides. *Bioinformatics* **2015**, *31* (14), 2276–2283. https://doi.org/10.1093/bioinformatics/btv133.

(119) Yap, K. L.; Kim, J.; Truong, K.; Sherman, M.; Yuan, T.; Ikura, M. Calmodulin Target Database. *J Struct Func Genom* **2000**, *1* (1), 8–14. https://doi.org/10.1023/A:1011320027914.

(120) Mruk, K.; Farley, B. M.; Ritacco, A. W.; Kobertz, W. R. Calmodulation Meta-Analysis: Predicting Calmodulin Binding via Canonical Motif Clustering. *Journal of General Physiology* **2014**, *144* (1), 105–114. https://doi.org/10.1085/jgp.201311140.

(121) Abbasi, W. A.; Asif, A.; Andleeb, S.; Minhas, F. ul A. A. CaMELS: In Silico Prediction of Calmodulin Binding Proteins and Their Binding Sites. *Proteins: Structure, Function, and Bioinformatics* **2017**, *85* (9), 1724–1740. https://doi.org/10.1002/prot.25330.

(122) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational Entropy in Molecular Recognition by Proteins. *Nature* **2007**, *448* (7151), 325–329. https://doi.org/10.1038/nature05959.

(123) Johnson, C.; Crowther, S.; Stafford, M. J.; Campbell, D. G.; Toth, R.; MacKintosh, C. Bioinformatic and Experimental Survey of 14-3-3-Binding Sites. *Biochemical Journal* **2010**, *427* (1), 69–78. https://doi.org/10.1042/BJ20091834.

(124) Davey, N. E.; Seo, M.-H.; Yadav, V. K.; Jeon, J.; Nim, S.; Krystkowiak, I.; Blikstad, C.; Dong, D.; Markova, N.; Kim, P. M.; Ivarsson, Y. Discovery of Short Linear Motif-Mediated Interactions through Phage Display of Intrinsically Disordered Regions of the Human Proteome. *The FEBS Journal* **2017**, *284* (3), 485–498. https://doi.org/10.1111/febs.13995.

(125) Wu, C.-G.; Chen, H.; Guo, F.; Yadav, V. K.; Mcilwain, S. J.; Rowse, M.; Choudhary, A.; Lin, Z.; Li, Y.; Gu, T.; Zheng, A.; Xu, Q.; Lee, W.; Resch, E.; Johnson, B.; Day, J.; Ge, Y.; Ong, I. M.; Burkard, M. E.; Ivarsson, Y.; Xing, Y. PP2A-B′ Holoenzyme Substrate Recognition, Regulation and Role in Cytokinesis. *Cell Discov* **2017**, *3* (1), 1–19. https://doi.org/10.1038/celldisc.2017.27.

(126) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols*

Handbook; Walker, J. M., Ed.; Springer Protocols Handbooks; Humana Press: Totowa, NJ, 2005; pp 571–607. https://doi.org/10.1385/1-59259-890-0:571.

(127) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A.; Gasparini, A.; Hatos, A.; Kajava, A. V.; Kalmar, L.; Leonardi, E.; Lazar, T.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Meszaros, A.; Minervini, G.; Murvai, N.; Pujols, J.; Roche, D. B.; Salladini, E.; Schad, E.; Schramm, A.; Szabo, B.; Tantos, A.; Tonello, F.; Tsirigos, K. D.; Veljković, N.; Ventura, S.; Vranken, W.; Warholm, P.; Uversky, V. N.; Dunker, A. K.; Longhi, S.; Tompa, P.; Tosatto, S. C. E. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Research* **2017**, *45* (D1), D219–D227. https://doi.org/10.1093/nar/gkw1056.

(128) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *Journal of Computational Chemistry* **2004**, *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.

(129) Fung, H. Y. J.; Birol, M.; Rhoades, E. IDPs in Macromolecular Complexes: The Roles of Multivalent Interactions in Diverse Assemblies. *Curr Opin Struct Biol* **2018**, *49*, 36–43. https://doi.org/10.1016/j.sbi.2017.12.007.

(130) Kakinuma, N.; Zhu, Y.; Wang, Y.; Roy, B. C.; Kiyama, R. Kank Proteins: Structure, Functions and Diseases. *Cell Mol Life Sci* **2009**, *66* (16), 2651–2659. https://doi.org/10.1007/s00018-009-0038-y.

(131) Bouchet, B. P.; Gough, R. E.; Ammon, Y.-C.; van de Willige, D.; Post, H.; Jacquemet, G.; Altelaar, A. M.; Heck, A. J.; Goult, B. T.; Akhmanova, A. Talin-KANK1 Interaction Controls the Recruitment of Cortical Microtubule Stabilizing Complexes to Focal Adhesions. *eLife* **2016**, *5*, e18124. https://doi.org/10.7554/eLife.18124.

(132) Sarkar, S.; Roy, B. C.; Hatano, N.; Aoyagi, T.; Gohji, K.; Kiyama, R. A Novel Ankyrin Repeat-Containing Gene (Kank) Located at 9p24 Is a Growth Suppressor of Renal Cell Carcinoma. *J Biol Chem* **2002**, *277* (39), 36585–36591. https://doi.org/10.1074/jbc.M204244200.

(133) Cui, Z.; Shen, Y.; Chen, K. H.; Mittal, S. K.; Yang, J.-Y.; Zhang, G. KANK1 Inhibits Cell Growth by Inducing Apoptosis through Regulating CXXC5 in Human Malignant Peripheral Nerve Sheath Tumors. *Sci Rep* **2017**, *7*, 40325. https://doi.org/10.1038/srep40325.

(134) Guo, X.; Fan, W.; Bian, X.; Ma, D. Upregulation of the Kank1 Gene-Induced Brain Glioma Apoptosis and Blockade of the Cell Cycle in G0/G1 Phase. *International Journal of Oncology* **2014**, *44* (3), 797–804. https://doi.org/10.3892/ijo.2014.2247.

(135) Gu, Y.; Zhang, M. Upregulation of the Kank1 Gene Inhibits Human Lung Cancer Progression in Vitro and in Vivo. *Oncol Rep* **2018**, *40* (3), 1243–1250. https://doi.org/10.3892/or.2018.6526.

(136) Ammon, York-Christoph; Akhmanova, A.S.; University Utrecht. Linking Cell-Matrix Adhesions and Microtubules: The Role of KANK Proteins, Utrecht University, 2020. https://doi.org/10.33540/346.

(137) van der Vaart, B.; van Riel, W. E.; Doodhi, H.; Kevenaar, J. T.; Katrukha, E. A.; Gumy, L.; Bouchet, B. P.; Grigoriev, I.; Spangler, S. A.; Yu, K. L.; Wulf, P. S.; Wu, J.; Lansbergen, G.; van Battum, E. Y.; Pasterkamp, R. J.; Mimori-Kiyosue, Y.; Demmers, J.; Olieric, N.; Maly, I. V.; Hoogenraad, C. C.; Akhmanova, A. CFEOM1-Associated Kinesin KIF21A Is a Cortical Microtubule Growth Inhibitor. *Dev Cell* **2013**, *27* (2), 145–160. https://doi.org/10.1016/j.devcel.2013.09.010.

(138) Estelle, A. B.; George, A.; Barbar, E. J.; Zuckerman, D. M. Quantifying Cooperative Multisite Binding through Bayesian Inference. bioRxiv July 1, 2022, p 2022.06.29.498022. https://doi.org/10.1101/2022.06.29.498022.
(139) Myllykoski, M.; Eichel, M. A.; Jung, R. B.; Kelm, S.; Werner, H. B.; Kursula, P. High-Affinity Heterotetramer Formation between the Large Myelin-Associated Glycoprotein and the Dynein Light Chain DYNLL1. *J Neurochem* **2018**, *147* (6), 764–783. https://doi.org/10.1111/jnc.14598.
(140) Hall, A.; Karplus, P. A.; Poole, L. B. Typical 2-Cys Peroxiredoxins – Structures, Mechanisms and Functions. *The FEBS Journal* **2009**, *276* (9), 2469–2477. https://doi.org/10.1111/j.1742-4658.2009.06985.x.
(141) Perkins, A.; Nelson, K. J.; Parsonage, D.; Poole, L. B.; Karplus, P. A. Peroxiredoxins: Guardians against Oxidative Stress and Modulators of Peroxide Signaling. *Trends in Biochemical Sciences* **2015**, *40* (8), 435–445. https://doi.org/10.1016/j.tibs.2015.05.001.
(142) Wood, Z. A.; Poole, L. B.; Karplus, P. A. Peroxiredoxin Evolution and the Regulation of Hydrogen Peroxide Signaling. *Science* **2003**, *300* (5619), 650–653. https://doi.org/10.1126/science.1080405.
(143) Portillo-Ledesma, S.; Randall, L. M.; Parsonage, D.; Dalla Rizza, J.; Karplus, P. A.; Poole, L. B.; Denicola, A.; Ferrer-Sueta, G. Differential Kinetics of Two-Cysteine Peroxiredoxin Disulfide Formation Reveal a Novel Model for Peroxide Sensing. *Biochemistry* **2018**, *57* (24), 3416–3424. https://doi.org/10.1021/acs.biochem.8b00188.
(144) Adimora, N. J.; Jones, D. P.; Kemp, M. L. A Model of Redox Kinetics Implicates the Thiol Proteome in Cellular Hydrogen Peroxide Responses. *Antioxid Redox Signal* **2010**, *13* (6), 731–743. https://doi.org/10.1089/ars.2009.2968.
(145) Winterbourn, C. C. Reconciling the Chemistry and Biology of Reactive Oxygen Species. *Nature Chemical Biology* **2008**, *4* (5), 278–286. https://doi.org/10.1038/nchembio.85.
(146) Hall, A.; Parsonage, D.; Poole, L. B.; Karplus, P. A. Structural Evidence That Peroxiredoxin Catalytic Power Is Based on Transition-State Stabilization. *J Mol Biol* **2010**, *402* (1), 194–209. https://doi.org/10.1016/j.jmb.2010.07.022.
(147) Kriznik, A.; Libiad, M.; Le Cordier, H.; Boukhenouna, S.; Toledano, M. B.; Rahuel-Clermont, S. Dynamics of a Key Conformational Transition in the Mechanism of Peroxiredoxin Sulfinylation. *ACS Catal* **2020**, *10* (5), 3326–3339. https://doi.org/10.1021/acscatal.9b04471.
(148) Randall, L. M.; Dalla Rizza, J.; Parsonage, D.; Santos, J.; Mehl, R. A.; Lowther, W. T.; Poole, L. B.; Denicola, A. Unraveling the Effects of Peroxiredoxin 2 Nitration; Role of C-Terminal Tyrosine 193. *Free Radic Biol Med* **2019**, *141*, 492–501. https://doi.org/10.1016/j.freeradbiomed.2019.07.016.
(149) Horta, B. B.; de Oliveira, M. A.; Discola, K. F.; Cussiol, J. R. R.; Netto, L. E. S. Structural and Biochemical Characterization of Peroxiredoxin Qβ from Xylella Fastidiosa. *Journal of Biological Chemistry* **2010**, *285* (21), 16051–16065. https://doi.org/10.1074/jbc.M109.094839.
(150) Perkins, A.; Nelson, K. J.; Williams, J. R.; Parsonage, D.; Poole, L. B.; Karplus, P. A. The Sensitive Balance between the Fully Folded and Locally Unfolded Conformations of a Model Peroxiredoxin. *Biochemistry* **2013**, *52* (48), 8708–8721. https://doi.org/10.1021/bi4011573.
(151) Perkins, A.; Parsonage, D.; Nelson, K. J.; Ogba, O. M.; Cheong, P. H.-Y.; Poole, L. B.; Karplus, P. A. Peroxiredoxin Catalysis at Atomic Resolution. *Structure* **2016**, *24* (10), 1668–1678. https://doi.org/10.1016/j.str.2016.07.012.

(152) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. Intrinsic Dynamics of an Enzyme Underlies Catalysis. *Nature* **2005**, *438* (7064), 117–121. https://doi.org/10.1038/nature04105.

(153) Venkitakrishnan, R. P.; Zaborowski, E.; McElheny, D.; Benkovic, S. J.; Dyson, H. J.; Wright, P. E. Conformational Changes in the Active Site Loops of Dihydrofolate Reductase during the Catalytic Cycle. *Biochemistry* **2004**, *43* (51), 16046–16055. https://doi.org/10.1021/bi048119y.

(154) Wolf-Watz, M.; Thai, V.; Henzler-Wildman, K.; Hadjipavlou, G.; Eisenmesser, E. Z.; Kern, D. Linkage between Dynamics and Catalysis in a Thermophilic-Mesophilic Enzyme Pair. *Nature Structural & Molecular Biology* **2004**, *11* (10), 945–949. https://doi.org/10.1038/nsmb821.

(155) Berlow, R. B.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E. Role of Backbone Dynamics in Modulating the Interactions of Disordered Ligands with the TAZ1 Domain of the CREB-Binding Protein. *Biochemistry* **2019**, *58* (10), 1354–1362. https://doi.org/10.1021/acs.biochem.8b01290.

(156) Ceccon, A.; Schmidt, T.; Tugarinov, V.; Kotler, S. A.; Schwieters, C. D.; Clore, G. M. Interaction of Huntingtin Exon-1 Peptides with Lipid-Based Micellar Nanoparticles Probed by Solution NMR and Q-Band Pulsed EPR. *J Am Chem Soc* **2018**, *140* (20), 6199–6202. https://doi.org/10.1021/jacs.8b02619.

(157) Delaforge, E.; Kragelj, J.; Tengo, L.; Palencia, A.; Milles, S.; Bouvignies, G.; Salvi, N.; Blackledge, M.; Jensen, M. R. Deciphering the Dynamic Interaction Profile of an Intrinsically Disordered Protein by NMR Exchange Spectroscopy. *J. Am. Chem. Soc.* **2018**, *140* (3), 1148–1158. https://doi.org/10.1021/jacs.7b12407.

(158) Meinhold, D. W.; Wright, P. E. Measurement of Protein Unfolding/Refolding Kinetics and Structural Characterization of Hidden Intermediates by NMR Relaxation Dispersion. *Proc Natl Acad Sci U S A* **2011**, *108* (22), 9078–9083. https://doi.org/10.1073/pnas.1105682108.

(159) Hall, A.; Nelson, K.; Poole, L. B.; Karplus, P. A. Structure-Based Insights into the Catalytic Power and Conformational Dexterity of Peroxiredoxins. *Antioxid Redox Signal* **2011**, *15* (3), 795–815. https://doi.org/10.1089/ars.2010.3624.

(160) Nelson, K. J.; Knutson, S. T.; Soito, L.; Klomsiri, C.; Poole, L. B.; Fetrow, J. S. Analysis of the Peroxiredoxin Family: Using Active-Site Structure and Sequence Information for Global Classification and Residue Analysis. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79* (3), 947–964. https://doi.org/10.1002/prot.22936.

(161) Adén, J.; Wallgren, M.; Storm, P.; Weise, C. F.; Christiansen, A.; Schröder, W. P.; Funk, C.; Wolf-Watz, M. Extraordinary Ms-Ms Backbone Dynamics in Arabidopsis Thaliana Peroxiredoxin Q. *Biochim Biophys Acta* **2011**, *1814* (12), 1880–1890. https://doi.org/10.1016/j.bbapap.2011.07.011.

(162) Liao, S.-J.; Yang, C.-Y.; Chin, K.-H.; Wang, A. H.-J.; Chou, S.-H. Insights into the Alkyl Peroxide Reduction Pathway of Xanthomonas Campestris Bacterioferritin Comigratory Protein from the Trapped Intermediate-Ligand Complex Structures. *J Mol Biol* **2009**, *390* (5), 951–966. https://doi.org/10.1016/j.jmb.2009.05.030.

(163) Buchko, G. W.; Perkins, A.; Parsonage, D.; Poole, L. B.; Karplus, P. A. Backbone Chemical Shift Assignments for Xanthomonas Campestris Peroxiredoxin Q in the Reduced and Oxidized States: A Dramatic Change in Backbone Dynamics. *Biomol NMR Assign* **2016**, *10* (1), 57–61. https://doi.org/10.1007/s12104-015-9637-8.

(164) Tüchsen, E.; Hayes, J. M.; Ramaprasad, S.; Copie, V.; Woodward, C. Solvent Exchange of Buried Water and Hydrogen Exchange of Peptide NH Groups Hydrogen Bonded to Buried

Waters in Bovine Pancreatic Trypsin Inhibitor. *Biochemistry* **1987**, *26* (16), 5163–5172. https://doi.org/10.1021/bi00390a040.

(165) Kjaergaard, M.; Poulsen, F. M. Sequence Correction of Random Coil Chemical Shifts: Correlation between Neighbor Correction Factors and Changes in the Ramachandran Distribution. *J. Biomol. NMR* **2011**, *50* (2), 157–165. https://doi.org/10.1007/s10858-011-9508-2.

(166) Hwang, T.-L.; van Zijl, P. C. M.; Mori, S. Accurate Quantitation of Water-Amide Proton Exchange Rates Using the Phase-Modulated CLEAN Chemical EXchange (CLEANEX-PM) Approach with a Fast-HSQC (FHSQC) Detection Scheme. *J Biomol NMR* **1998**, *11* (2), 221–226. https://doi.org/10.1023/A:1008276004875.

(167) Zhang, Y.-Z. Protein and Peptide Structure and Interactions Studied by Hydrogen Exchanger and NMR. *Dissertations available from ProQuest* **1995**, 1–203.

(168) Fitzkee, N. C.; Torchia, D. A.; Bax, A. Measuring Rapid Hydrogen Exchange in the Homodimeric 36 KDa HIV-1 Integrase Catalytic Core Domain. *Protein Sci* **2011**, *20* (3), 500–512. https://doi.org/10.1002/pro.582.

(169) Skinner, J. J.; Lim, W. K.; Bédard, S.; Black, B. E.; Englander, S. W. Protein Dynamics Viewed by Hydrogen Exchange. *Protein Sci* **2012**, *21* (7), 996–1005. https://doi.org/10.1002/pro.2081.

(170) Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. Protein Stability Parameters Measured by Hydrogen Exchange. *Proteins* **1994**, *20* (1), 4–14. https://doi.org/10.1002/prot.340200103.

(171) Englander, S. W.; Sosnick, T. R.; Englander, J. J.; Mayne, L. Mechanisms and Uses of Hydrogen Exchange. *Current Opinion in Structural Biology* **1996**, *6* (1), 18–23. https://doi.org/10.1016/S0959-440X(96)80090-X.

(172) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313* (5793), 1638–1642. https://doi.org/10.1126/science.1130258.

(173) Karplus, P. A.; Pearson, M. A.; Hausinger, R. P. 70 Years of Crystalline Urease: What Have We Learned? *Acc. Chem. Res.* **1997**, *30* (8), 330–337. https://doi.org/10.1021/ar960022j.

(174) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res* **2004**, *14* (6), 1188–1190. https://doi.org/10.1101/gr.849004.

(175) Trivelli, X.; Krimm, I.; Ebel, C.; Verdoucq, L.; Prouzet-Mauléon, V.; Chartier, Y.; Tsan, P.; Lauquin, G.; Meyer, Y.; Lancelin, J.-M. Characterization of the Yeast Peroxiredoxin Ahp1 in Its Reduced Active and Overoxidized Inactive Forms Using NMR. *Biochemistry* **2003**, *42* (48), 14139–14149. https://doi.org/10.1021/bi035551r.

(176) Lian, F.-M.; Yu, J.; Ma, X.-X.; Yu, X.-J.; Chen, Y.; Zhou, C.-Z. Structural Snapshots of Yeast Alkyl Hydroperoxide Reductase Ahp1 Peroxiredoxin Reveal a Novel Two-Cysteine Mechanism of Electron Transfer to Eliminate Reactive Oxygen Species. *J Biol Chem* **2012**, *287* (21), 17077–17087. https://doi.org/10.1074/jbc.M112.357368.

(177) Perkins, A.; Gretes, M. C.; Nelson, K. J.; Poole, L. B.; Karplus, P. A. Mapping the Active Site Helix-to-Strand Conversion of CxxxxC Peroxiredoxin Q Enzymes. *Biochemistry* **2012**, *51* (38), 7638–7650. https://doi.org/10.1021/bi301017s.

(178) Bolduc, J. A.; Nelson, K. J.; Haynes, A. C.; Lee, J.; Reisz, J. A.; Graff, A. H.; Clodfelter, J. E.; Parsonage, D.; Poole, L. B.; Furdui, C. M.; Lowther, W. T. Novel Hyperoxidation Resistance Motifs in 2-Cys Peroxiredoxins. *J Biol Chem* **2018**, *293* (30), 11901–11912. https://doi.org/10.1074/jbc.RA117.001690.

(179) Nelson, K. J.; Parsonage, D.; Karplus, P. A.; Poole, L. B. Evaluating Peroxiredoxin Sensitivity toward Inactivation by Peroxide Substrates. *Methods Enzymol* **2013**, *527*, 21–40. https://doi.org/10.1016/B978-0-12-405882-8.00002-7.

(180) Poynton, R. A.; Peskin, A. V.; Haynes, A. C.; Lowther, W. T.; Hampton, M. B.; Winterbourn, C. C. Kinetic Analysis of Structural Influences on the Susceptibility of Peroxiredoxins 2 and 3 to Hyperoxidation. *Biochem J* **2016**, *473* (4), 411–421. https://doi.org/10.1042/BJ20150572.

(181) Hall, A.; Parsonage, D.; Horita, D.; Karplus, P. A.; Poole, L. B.; Barbar, E. Redox-Dependent Dynamics of a Dual Thioredoxin Fold Protein: Evolution of Specialized Folds. *Biochemistry* **2009**, *48* (25), 5984–5993. https://doi.org/10.1021/bi900270w.

(182) Bhutani, N.; Udgaonkar, J. B. Folding Subdomains of Thioredoxin Characterized by Native-State Hydrogen Exchange. *Protein Science* **2003**, *12* (8), 1719–1731. https://doi.org/10.1110/ps.0239503.

(183) Jeng, M. F.; Dyson, H. J. Comparison of the Hydrogen-Exchange Behavior of Reduced and Oxidized Escherichia Coli Thioredoxin. *Biochemistry* **1995**, *34* (2), 611–619. https://doi.org/10.1021/bi00002a028.

(184) Nirudodhi, S.; Parsonage, D.; Karplus, P. A.; Poole, L. B.; Maier, C. S. Conformational Studies of the Robust 2-Cys Peroxiredoxin Salmonella Typhimurium AhpC by Solution Phase Hydrogen/Deuterium (H/D) Exchange Monitored by Electrospray Ionization Mass Spectrometry. *Int J Mass Spectrom* **2011**, *302* (1–3), 93–100. https://doi.org/10.1016/j.ijms.2010.08.018.

(185) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J Biomol NMR* **1995**, *6* (3), 277–293. https://doi.org/10.1007/BF00197809.

(186) Coggins, B. E.; Werner-Allen, J. W.; Yan, A.; Zhou, P. Rapid Protein Global Fold Determination Using Ultrasparse Sampling, High-Dynamic Range Artifact Suppression, and Time-Shared NOESY. *J. Am. Chem. Soc.* **2012**, *134* (45), 18619–18630. https://doi.org/10.1021/ja307445y.

(187) Farrow, N. A.; Muhandiram, R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Forman-Kay, J. D.; Kay, L. E. Backbone Dynamics of a Free and Phosphopeptide-Complexed Src Homology 2 Domain Studied by 15N NMR Relaxation. *Biochemistry* **1994**, *33* (19), 5984–6003. https://doi.org/10.1021/bi00185a040.

(188) Johnson, B. A. Using NMRView to Visualize and Analyze the NMR Spectra of Macromolecules. *Methods Mol Biol* **2004**, *278*, 313–352. https://doi.org/10.1385/1-59259-809-9:313.

(189) Korzhnev, D. M.; Skrynnikov, N. R.; Millet, O.; Torchia, D. A.; Kay, L. E. An NMR Experiment for the Accurate Measurement of Heteronuclear Spin-Lock Relaxation Rates. *J. Am. Chem. Soc.* **2002**, *124* (36), 10743–10753. https://doi.org/10.1021/ja0204776.

(190) Palmer, A. G.; Kroenke, C. D.; Loria, J. P. Nuclear Magnetic Resonance Methods for Quantifying Microsecond-to-Millisecond Motions in Biological Macromolecules. *Methods Enzymol* **2001**, *339*, 204–238. https://doi.org/10.1016/s0076-6879(01)39315-1.

(191) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity. *J. Am. Chem. Soc.* **1982**, *104* (17), 4546–4559. https://doi.org/10.1021/ja00381a009.

(192) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. *J. Am. Chem. Soc.* **1982**, *104* (17), 4559–4570. https://doi.org/10.1021/ja00381a010.

(193) d'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models I. Minimisation Algorithms and Their Performance within the Model-Free and Brownian Rotational Diffusion Spaces. *J Biomol NMR* **2008**, *40* (2), 107–119. https://doi.org/10.1007/s10858-007-9214-2.
(194) d'Auvergne, E. J.; Gooley, P. R. Optimisation of NMR Dynamic Models II. A New Methodology for the Dual Optimisation of the Model-Free Parameters and the Brownian Rotational Diffusion Tensor. *J Biomol NMR* **2008**, *40* (2), 121–133. https://doi.org/10.1007/s10858-007-9213-3.
(195) Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M. Deviations from the Simple Two-Parameter Model-Free Approach to the Interpretation of Nitrogen-15 Nuclear Magnetic Relaxation of Proteins. *J. Am. Chem. Soc.* **1990**, *112* (12), 4989–4991. https://doi.org/10.1021/ja00168a070.
(196) Barbar, E.; Hare, M.; Daragan, V.; Barany, G.; Woodward, C. Dynamics of the Conformational Ensemble of Partially Folded Bovine Pancreatic Trypsin Inhibitor. *Biochemistry* **1998**, *37* (21), 7822–7833. https://doi.org/10.1021/bi973102j.
(197) Barbar, E.; Hare, M.; Makokha, M.; Barany, G.; Woodward, C. NMR-Detected Order in Core Residues of Denatured Bovine Pancreatic Trypsin Inhibitor †. *Biochemistry* **2001**, *40* (32), 9734–9742. https://doi.org/10.1021/bi010483z.
(198) Hall, J.; Hall, A.; Pursifull, N.; Barbar, E. Differences in Dynamic Structure of LC8 Monomer, Dimer, and Dimer-Peptide Complexes. *Biochemistry* **2008**, *47* (46), 11940–11952. https://doi.org/10.1021/bi801093k.
(199) d'Auvergne, E. J.; Gooley, P. R. The Use of Model Selection in the Model-Free Analysis of Protein Dynamics. *J Biomol NMR* **2003**, *25* (1), 25–39. https://doi.org/10.1023/A:1021902006114.
(200) d'Auvergne, E. J.; Gooley, P. R. Model-Free Model Elimination: A New Step in the Model-Free Dynamic Analysis of NMR Relaxation Data. *J Biomol NMR* **2006**, *35* (2), 117. https://doi.org/10.1007/s10858-006-9007-z.
(201) Vallurupalli, P.; Bouvignies, G.; Kay, L. E. Studying "Invisible" Excited Protein States in Slow Exchange with a Major State Conformation. *J. Am. Chem. Soc.* **2012**, *134* (19), 8148–8161. https://doi.org/10.1021/ja3001419.
(202) Forsythe, H. M.; Vetter, C. J.; Jara, K. A.; Reardon, P. N.; David, L. L.; Barbar, E. J.; Lampi, K. J. Altered Protein Dynamics and Increased Aggregation of Human ΓS-Crystallin Due to Cataract-Associated Deamidations. *Biochemistry* **2019**, *58* (40), 4112–4124. https://doi.org/10.1021/acs.biochem.9b00593.
(203) Ritchie, H.; Mathieu, E.; Rodés-Guirao, L.; Appel, C.; Giattino, C.; Ortiz-Ospina, E.; Hasell, J.; Macdonald, B.; Beltekian, D.; Roser, M. Coronavirus Pandemic (COVID-19). *Our World in Data* **2020**.
(204) Cong, Y.; Ulasli, M.; Schepers, H.; Mauthe, M.; V'kovski, P.; Kriegenburg, F.; Thiel, V.; de Haan, C. A. M.; Reggiori, F. Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *J Virol* **2020**, *94* (4), e01925-19. https://doi.org/10.1128/JVI.01925-19.
(205) Quayum, S. T.; Hasan, S. Analysing the Impact of the Two Most Common SARS-CoV-2 Nucleocapsid Protein Variants on Interactions with Membrane Protein in Silico. *J Genet Eng Biotechnol* **2021**, *19* (1), 138. https://doi.org/10.1186/s43141-021-00233-z.
(206) Lu, S.; Ye, Q.; Singh, D.; Cao, Y.; Diedrich, J. K.; Yates, J. R.; Villa, E.; Cleveland, D. W.; Corbett, K. D. The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive

Condensates with RNA and the Membrane-Associated M Protein. *Nat Commun* **2021**, *12* (1), 502. https://doi.org/10.1038/s41467-020-20768-y.

(207) Lu, X.; Pan, J.; Tao, J.; Guo, D. SARS-CoV Nucleocapsid Protein Antagonizes IFN-β Response by Targeting Initial Step of IFN-β Induction Pathway, and Its C-Terminal Region Is Critical for the Antagonism. *Virus Genes* **2011**, *42* (1), 37–45. https://doi.org/10.1007/s11262-010-0544-x.

(208) Zhou, B.; Liu, J.; Wang, Q.; Liu, X.; Li, X.; Li, P.; Ma, Q.; Cao, C. The Nucleocapsid Protein of Severe Acute Respiratory Syndrome Coronavirus Inhibits Cell Cytokinesis and Proliferation by Interacting with Translation Elongation Factor 1alpha. *J Virol* **2008**, *82* (14), 6962–6971. https://doi.org/10.1128/JVI.00133-08.

(209) Mason, R. J. Pathogenesis of COVID-19 from a Cell Biology Perspective. *European Respiratory Journal* **2020**, *55* (4). https://doi.org/10.1183/13993003.00607-2020.

(210) Forsythe, H. M.; Barbar, E. The Role of Dancing Duplexes in Biology and Disease. *Prog Mol Biol Transl Sci* **2021**, *183*, 249–270. https://doi.org/10.1016/bs.pmbts.2021.06.004.

(211) Forsythe, H. M.; Rodriguez Galvan, J.; Yu, Z.; Pinckney, S.; Reardon, P.; Cooley, R. B.; Zhu, P.; Rolland, A. D.; Prell, J. S.; Barbar, E. Multivalent Binding of the Partially Disordered SARS-CoV-2 Nucleocapsid Phosphoprotein Dimer to RNA. *Biophys J* **2021**, *120* (14), 2890–2901. https://doi.org/10.1016/j.bpj.2021.03.023.

(212) Redzic, J. S.; Lee, E.; Born, A.; Issaian, A.; Henen, M. A.; Nichols, P. J.; Blue, A.; Hansen, K. C.; D'Alessandro, A.; Vögeli, B.; Eisenmesser, E. Z. The Inherent Dynamics and Interaction Sites of the SARS-CoV-2 Nucleocapsid N-Terminal Region. *J Mol Biol* **2021**, *433* (15), 167108. https://doi.org/10.1016/j.jmb.2021.167108.

(213) Dinesh, D. C.; Chalupska, D.; Silhan, J.; Koutna, E.; Nencka, R.; Veverka, V.; Boura, E. Structural Basis of RNA Recognition by the SARS-CoV-2 Nucleocapsid Phosphoprotein. *PLOS Pathogens* **2020**, *16* (12), e1009100. https://doi.org/10.1371/journal.ppat.1009100.

(214) Khan, A.; Tahir Khan, M.; Saleem, S.; Junaid, M.; Ali, A.; Shujait Ali, S.; Khan, M.; Wei, D.-Q. Structural Insights into the Mechanism of RNA Recognition by the N-Terminal RNA-Binding Domain of the SARS-CoV-2 Nucleocapsid Phosphoprotein. *Comput Struct Biotechnol J* **2020**, *18*, 2174–2184. https://doi.org/10.1016/j.csbj.2020.08.006.

(215) Zhou, R.; Zeng, R.; von Brunn, A.; Lei, J. Structural Characterization of the C-Terminal Domain of SARS-CoV-2 Nucleocapsid Protein. *Mol Biomed* **2020**, *1*, 2. https://doi.org/10.1186/s43556-020-00001-4.

(216) Carlson, C. R.; Asfaha, J. B.; Ghent, C. M.; Howard, C. J.; Hartooni, N.; Safari, M.; Frankel, A. D.; Morgan, D. O. Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for Its Dual Functions. *Mol Cell* **2020**, *80* (6), 1092-1103.e4. https://doi.org/10.1016/j.molcel.2020.11.025.

(217) Roden, C.; Gladfelter, A. S. RNA Contributions to the Form and Function of Biomolecular Condensates. *Nat Rev Mol Cell Biol* **2021**, *22* (3), 183–195. https://doi.org/10.1038/s41580-020-0264-6.

(218) Scherer, K. M.; Mascheroni, L.; Carnell, G. W.; Wunderlich, L. C. S.; Makarchuk, S.; Brockhoff, M.; Mela, I.; Fernandez-Villegas, A.; Barysevich, M.; Stewart, H.; Suau Sans, M.; George, C. L.; Lamb, J. R.; Kaminski-Schierle, G. S.; Heeney, J. L.; Kaminski, C. F. SARS-CoV-2 Nucleocapsid Protein Adheres to Replication Organelles before Viral Assembly at the Golgi/ERGIC and Lysosome-Mediated Egress. *Sci Adv* **2022**, *8* (1), eabl4895. https://doi.org/10.1126/sciadv.abl4895.

(219) Iserman, C.; Roden, C. A.; Boerneke, M. A.; Sealfon, R. S. G.; McLaughlin, G. A.; Jungreis, I.; Fritch, E. J.; Hou, Y. J.; Ekena, J.; Weidmann, C. A.; Theesfeld, C. L.; Kellis, M.; Troyanskaya,

O. G.; Baric, R. S.; Sheahan, T. P.; Weeks, K. M.; Gladfelter, A. S. Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid. *Mol Cell* **2020**, *80* (6), 1078-1091.e6. https://doi.org/10.1016/j.molcel.2020.11.041.

(220) Wu, C.; Qavi, A. J.; Hachim, A.; Kavian, N.; Cole, A. R.; Moyle, A. B.; Wagner, N. D.; Sweeney-Gibbons, J.; Rohrs, H. W.; Gross, M. L.; Peiris, J. S. M.; Basler, C. F.; Farnsworth, C. W.; Valkenburg, S. A.; Amarasinghe, G. K.; Leung, D. W. Characterization of SARS-CoV-2 Nucleocapsid Protein Reveals Multiple Functional Consequences of the C-Terminal Domain. *iScience* **2021**, *24* (6), 102681. https://doi.org/10.1016/j.isci.2021.102681.

(221) Chauhan, A.; Avti, P.; Shekhar, N.; Prajapat, M.; Sarma, P.; Bhattacharyya, A.; Kumar, S.; Kaur, H.; Prakash, A.; Medhi, B. Structural and Conformational Analysis of SARS CoV 2 N-CTD Revealing Monomeric and Dimeric Active Sites during the RNA-Binding and Stabilization: Insights towards Potential Inhibitors for N-CTD. *Comput Biol Med* **2021**, *134*, 104495. https://doi.org/10.1016/j.compbiomed.2021.104495.

(222) Li, P.; Banjade, S.; Cheng, H.-C.; Kim, S.; Chen, B.; Guo, L.; Llaguno, M.; Hollingsworth, J. V.; King, D. S.; Banani, S. F.; Russo, P. S.; Jiang, Q.-X.; Nixon, B. T.; Rosen, M. K. Phase Transitions in the Assembly of Multi-Valent Signaling Proteins. *Nature* **2012**, *483* (7389), 336–340. https://doi.org/10.1038/nature10879.

Appendices

Appendix 1

Native State Fluctuations in a Peroxiredoxin Active Site Match Motions Needed for Catalysis

Aidan B Estelle, Patrick N Reardon, Seth Pinckney, Leslie B Poole, Elisar Barbar, P. Andrew Karplus

Published in *Structure* (2022) 30(2):278-288

Summary

Peroxiredoxins are ubiquitous enzymes that detoxify peroxides and regulate redox signaling. During catalysis, a 'peroxidatic' cysteine (C_P) in the conserved active site reduces peroxide while being oxidized to a C_P-sulfenate, prompting a local unfolding event which enables formation of a disulfide with a second, 'resolving' cysteine. Here, we use nuclear magnetic resonance spectroscopy to probe the dynamics of the C_P-thiolate and disulfide forms of *Xanthomonas campestris* Peroxiredoxin Q. Chemical exchange saturation transfer behavior of the resting enzyme reveals 26 residues in and around the active site exchanging at a rate of 72 s^{-1} with a locally-unfolded, high-energy (2.5% of the population) state. This unequivocally establishes that a catalytically-relevant local-unfolding equilibrium exists in the enzyme's C_P-thiolate form. Also, faster motions imply an active site instability that could promote local unfolding and, based on other work, be exacerbated by C_P-sulfenate formation so as to direct the enzyme along a functional catalytic trajectory.

Introduction

Peroxiredoxins (Prxs) are ubiquitous enzymes which efficiently reduce peroxides using simple cysteine chemistry[140–142]. They are highly efficient peroxide scavengers, with rate constants for peroxide reduction up to 10^8 M^{-1} s^{-1}, nearing the diffusion limit[141,143]. Historically, Prxs have taken a backseat to well characterized peroxidases such as catalase, but evidence implies that Prxs reduce upwards of 90% of cellular peroxides, serving as key protectors against oxidative stress, and also, in eukaryotic cells, key regulators of redox signaling[144,145].

Prxs have a thioredoxin fold and conserve an active-site PxxxS/TxxC sequence[141] that ends with a so-called 'peroxidatic' cysteine (C_P) and forms a loop preceding helix α2 (the C_P-loop) and the first turn of helix α2 (Fig. A1.1b). The Prx catalytic cycle begins with C_P attacking peroxide substrate to yield a C_P-sulfenate and a reduced water or alcohol product (Fig. A1.1a)[141,142], with an active site exquisitely set up to stabilize the transition state of this S_N2 displacement reaction[146]. Then an obligatory local unfolding event allows a so-called 'resolving' cysteine (C_R) to attack the C_P-sulfenate to form a C_P-C_R disulfide bond. The resolving Cys may come from a different protein, but commonly comes from another part of the Prx chain or partner subunit, in which case the required local unfolding involves at least one structural element in addition to the catalytic pocket region involving

the C_P-loop and the first turn of helix α2[141]. Disulfide formation covalently locks the protein in a locally-unfolded conformation until the Prx is reduced back to an active thiolate state, usually by a thioredoxin[141]. Thus, for 2-Cys Prxs, the enzyme cycles from a resting fully-folded (FF) C_P-thiolate form through a C_P-SOH form, to a locally-unfolded (LU) C_P-SS-C_R-disulfide form before being recycled to complete the cycle (Fig. A1.1a). Several cleverly designed kinetics studies have additionally shown that the local-unfolding event itself can, for some Prxs, be the rate-limiting step in disulfide formation[147,148]. While not the focus of this work, many Prxs in eukaryotes are especially sensitive to inactivation via a hyperoxidation reaction that is in competition with the local unfolding (Fig. A1.1a) and these "sensitive" Prxs are thought to be involved in the regulation of redox signaling[141,142,147].

Given the prominent conformational changes during catalysis, the dynamic properties of Prxs remain poorly understood because well-studied Prxs form dimeric or decameric complexes that are too large for facile study by solution nuclear magnetic resonance (NMR) spectroscopy. A major open question in the field is how much catalytically-relevant local unfolding already occurs in the resting C_P-thiolate form of 2-Cys Prxs, versus how much the oxidation to a C_P-sulfenate triggers a conformational change to the LU state. While few details are known about the thermodynamics or kinetics of local unfolding, crystallographic studies have implied that sulfenate formation induces local unfolding[149–151], and a recent study of hyperoxidation kinetics concluded that sulfenate formation promotes local unfolding by over 100-fold[147]. NMR-based investigation into solution dynamics has the potential to answer these questions, and conformational exchange measurements in particular have a track record of providing insight into catalytically relevant motions in absence of catalysis[152–154], as well as folding-unfolding equilibria [155–158].

The PrxQ subfamily of peroxiredoxins[159,160] – found in bacteria and many plants and fungi, and with subgroups based on C_R location (in helix α2, α3 or no C_R) – includes monomeric members potentially amenable to study by NMR[151,161]. We selected PrxQ from the plant pathogen *Xanthomonas campestris* (*Xc*PrxQ) as a model system that is both monomeric and crystallographically characterized[162]. *Xc*PrxQ is a 17 kDa monomer with its peroxidatic Cys (Cys48) at the N-terminal end of helix α2 (as it is for all Prxs), and its resolving Cys (Cys84) in the middle of helix α3 (Fig. A1.1b,c). Also, helix α2 is kinked, due to a proline at position 60, giving it distinct N- and C-terminal parts we call here $\alpha2_N$ and

$\alpha2_C$. Disulfide formation for *Xc*PrxQ involves a large shift in the C_P-loop, very little change in the N-terminal end of helix α2 and the complete unfolding of helix α3 that allows the C_P- and C_R-residues to come together (Fig. A1.1b). Despite the minimal structure change in helix α2, based on studies of a closely related PrxQ, Horta et al., (2010) proposed that its dynamics were nevertheless very important and that C_P- sulfenate formation triggered a local unfolding of helix α2 as a necessary intermediate, and that the helix refolded after disulfide formation.

As groundwork for the dynamics studies reported here, we carried out an ~1 Å resolution crystallographic analysis of the *Xc*PrxQ catalytic cycle[151] that included views of the C_P-sulfenate and C_P-sulfinate forms– formed in the catalytically active crystals – and a highly unusual structure for the C_P-sulfenate led to the conclusion that its formation does indeed destabilize the FF active site. We also assigned the NMR spectra of the C_P-thiolate and disulfide forms of *Xc*PrxQ[163]. Notably, while the C_P-thiolate spectrum of *Xc*PrxQ could be nearly fully assigned (149/152), the disulfide spectrum was missing over half of the expected resonances (68/152 assigned), consistent with these regions experiencing peak broadening due to intermediate exchange and providing evidence of interesting motion in this form. Intriguingly, the one other monomeric PrxQ studied by NMR is from the subgroup with C_R in α2, and while it also showed extensive peak loss due to intermediate exchange, in that case it was for the C_P-thiolate form of the enzyme[161].

Here, we use heteronuclear NMR dynamics experiments to characterize the fast, intermediate, and slow motions in both the C_P-thiolate and disulfide forms of *Xc*PrxQ. We find little difference between fast-timescale spin-relaxation behavior of the C_P-thiolate and disulfide forms in spite of dramatic differences in NMR spectra, identify a slow exchanging core that is common to both forms, and most significantly, report unequivocal evidence that the resting C_P-thiolate form of the enzyme contains, at about a 2.5% level, a population of a catalytically-relevant locally-unfolded excited state.

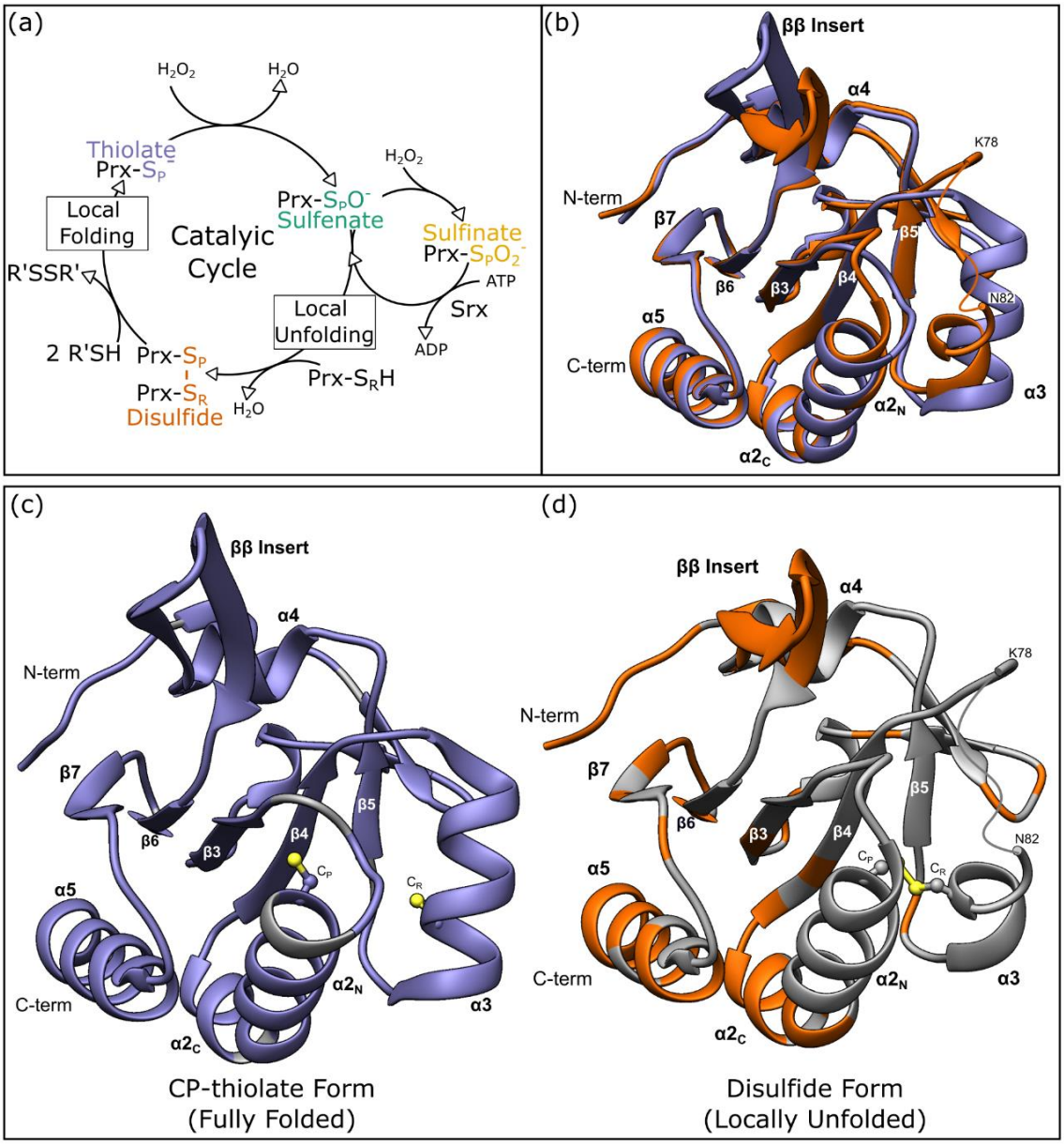


Figure A1.1: Overview of *Xc*PrxQ catalysis and structure. (a) Catalytic cycle of Prxs, identifying the C_P-thiolate, C_P-sulfenate, C_P-sulfinate and disulfide enzyme forms. Srx is sulfiredoxin. (b) Overlay of ribbon diagrams of *Xc*PrxQ C_P-thiolate (purple; PDB code 5IIZ) and disulfide (orange; PDB code 5IOX) forms. Major secondary structure elements are labeled, and a thin trace represents residues 79-81 that are not observed in the disulfide form crystal structure. (c,d) Ribbon diagrams as in panel B, but showing the C_P and C_R residues as sticks and with the ribbon colored only for residues with assigned amide resonances[163]. Excluding prolines, the C_P-thiolate and disulfide forms were 97% and 45% assigned, respectively.

Results

NMR spectral analysis

NMR spectra collected on ^{15}N- and ^{13}C,^{15}N-labelled *Xc*PrxQ (Fig. A1.2a) under reducing (C_P-thiolate form) and non-reducing (disulfide form) conditions closely match available assignments[163]. About half of the backbone amide resonances, including those of the C_P (Cys48) and C_R (Cys84) residues, are not present in spectra of the disulfide form. While the lack of peaks for the catalytic cysteines complicates verification of their oxidation state, our buffer closely matches that used in crystallization of the disulfide form[151,163]. Furthermore, reduction by DTT restores the C_P-thiolate spectrum, confirming that the peak disappearance is not an artefact but is specific to the disulfide form. Both the C_P and C_R amides are assigned in the *Xc*PrxQ spectrum under reducing conditions, and the Cβ chemical shifts of the C_P and C_R residues obtained from HNCACB experiments are consistent with expected shifts for a Cys thiol/thiolate, confirming the protein is the C_P-thiolate state.

The Ser-44 amide makes an NH...pi hydrogen bond

The ^{1}H-^{15}N HSQC spectrum of the *Xc*PrxQ C_P-thiolate form has a peak with an unusual amide proton chemical shift of 4.4 ppm (Fig. A1.2a), that was not present in the previously assigned spectra collected at 500 MHz. Using HNCA and HN(CO)CA datasets collected at 800 MHz, we unambiguously assigned the peak to the Ser44 NH (Fig. A1.2c), one of just three previously unassigned backbone amides. In crystal structures, Ser44 directly precedes the beginning of the active site helix α2 (adjacent to the conserved Thr45), with its amide proton interacting with the electron-rich face of the aromatic ring of Phe83 from helix α3 (Fig. A1.2b). Such an amide-pi hydrogen bond, while uncommon, has been observed by NMR in other proteins, and this environment provides a satisfactory rationale for the unusual chemical shift, as for example, it matches closely with the ^{1}H shift of 4.3 ppm seen for Gly 37 of the bovine pancreatic trypsin inhibitor[164].

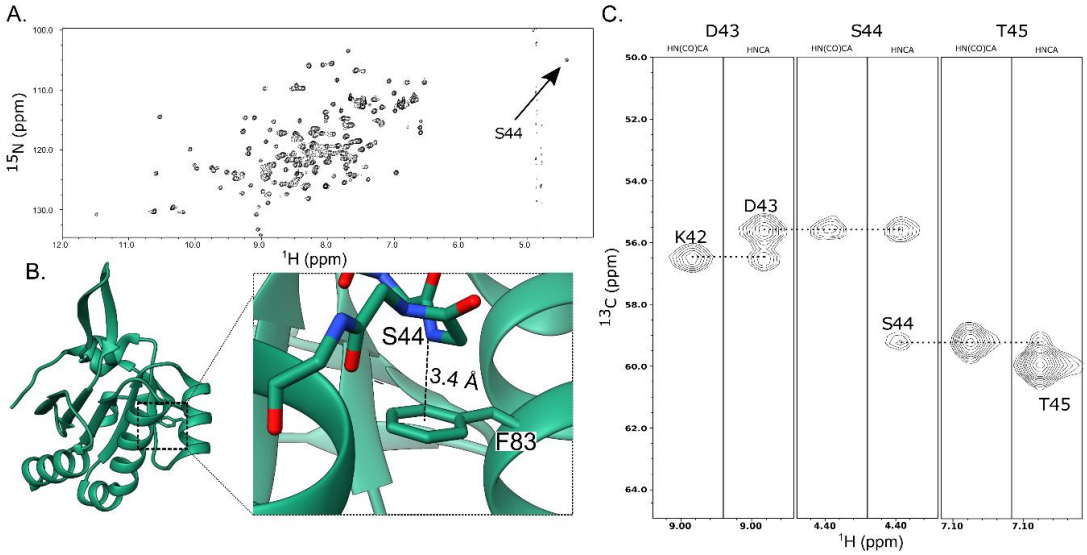


Figure A1.2: Unusual chemical shift and environment of Ser44 amide. (a) 1H-15N HSQC of *Xc*PrxQ under reducing conditions, with the Ser44 amide peak indicated. (b) Context and close-up view of the interaction between the Ser44 amide and the Phe83 side chain in the *Xc*PrxQ C_P-thiolate crystal structure (PDB code 5IIZ). (c) 1H-13C strips of HNCA and HN(CO)CA spectra, allowing the assignment of Ser44 amide.

Modelfree analysis reveals that the thiolate and disulfide have similar ps-ns dynamics

To examine the ps-ns timescale internal motions of *Xc*PrxQ, we measured ^{15}N R_1, ^{15}N R_2 relaxation rates and the steady state ^{1}H-^{15}N heteronucelar NOE ({^{1}H}-^{15}N NOE) of *Xc*PrxQ in C_P-thiolate and disulfide forms (Fig. A1.3a) at 800 and 500 MHz (SI Fig. A1.1). Reliable relaxation rates and {^{1}H}-^{15}N NOE values were obtained for 131 of 149 assigned residues in the C_P-thiolate form, with the remaining being unresolved due to peak overlaps. Modelfree analysis for the C_P-thiolate form revealed that the N- and C-termini and the ββ insert (residues 110-117) are undergoing rapid motion in the ps-ns timescale, with S^2 values dropping below 0.6 in these regions, indicating that they are at least partially disordered in solution (Fig. A1.3a). Consistent with this, these three regions have the highest B-factors in the crystal structure (Fig. A1.3b). We observed substantive R_{ex} terms (> 3 s^{-1}) for 17 residues in total (SI Fig. A1.2). These residues occur throughout the protein, and notably include C_P (Cys48) and its neighbors Thr45 and Thr49, as well as a set of residues for which alternate conformations were seen in the crystal structures (SI Fig. A1.3). To provide an alternate measurement of ^{15}N R_2 rates we collected an additional measurement of ^{15}N $R_{1\rho}$ and used the measured ^{15}N $R_{1\rho}$ and ^{15}N R_1 rates to calculate ^{15}N R_2 relaxation (SI Fig. A1.2). Two residues in particular – C_P (Cys48) and Thr49 – had

abnormally high ^{15}N R_2 rates when determined this way, similar to what was seen in our direct ^{15}N R_2 measurements, and indicative of exchange contributing to relaxation.

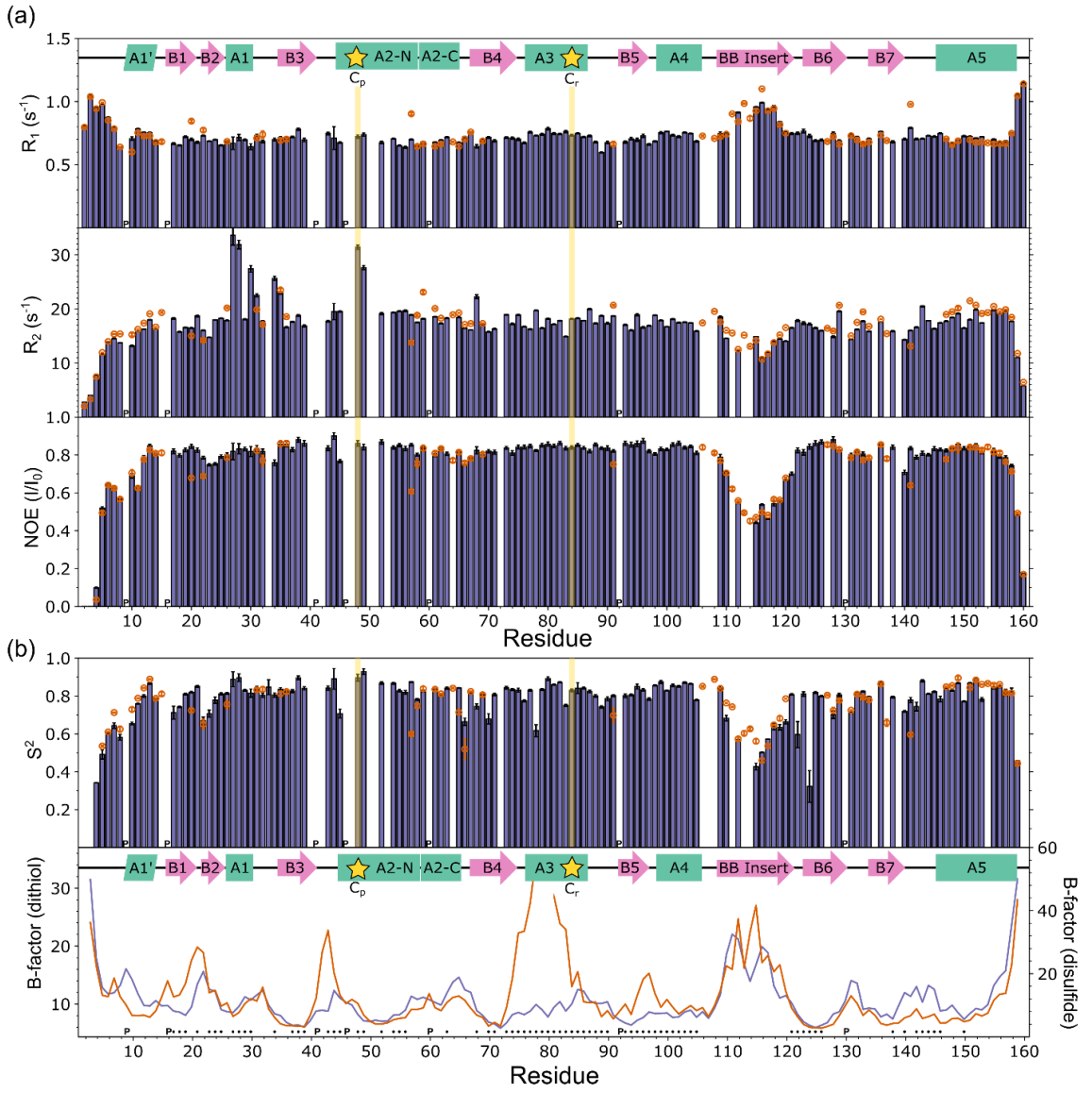


Figure A1.3: Spin relaxation and modelfree analysis. (a)^{15}N R_1 (top), R_2 (middle) and {^1H}-^{15}N NOE (bottom) values for C_P-thiolate (purple bars) and disulfide (orange circles) forms of *Xc*PrxQ. Secondary structure from the C_P-thiolate crystal structure is shown at the top of the figure (α-helices as rectangles, β-strands as arrows, and 3_{10}-helices as parallelograms) with the peroxidatic and resolving Cys locations highlighted in yellow. (b) Order parameter (top) for disulfide and C_P-thiolate forms, and average backbone crystallographic B-factors (bottom) from the C_P-thiolate (Purple; PDB code 5IIZ) and disulfide (Orange; PDB code 5IOX) crystal structure. Residues with missing disulfide-state NMR assignments are marked with a dot. Plot scaled to the median B-factor align.

For the disulfide form of the protein, the S^2 values are generally similar to those of the C_P-thiolate form, suggesting that the fast motions of these parts of the protein are largely unchanged by disulfide formation. The modelfree spectral density functions selected for the disulfide form were also generally the same on a residue-by-residue basis (SI Tables 1 and 2 – see manuscript), further supporting the conclusion that the fast time scale dynamics are similar between the two forms of the protein. With only a handful of exceptions, residues with R_{ex} terms in the C_P-thiolate form also have R_{ex} terms in the disulfide form, and notably all 17 residues with R_{ex}>3 s^{-1} in the C_P-thiolate form are either missing or also have an R_{ex} term in the disulfide form.

CEST studies reveal dynamics of local unfolding involving the catalytic helices α*2 and* α*3*

To examine chemical exchange in C_P-thiolate *Xc*PrxQ, we performed ^{15}N-Chemical Exchange Saturation Transfer (CEST) experiments. Eighteen residues in C_P-thiolate *Xc*PrxQ showed obvious exchange, as evidenced by two distinct dips in their CEST profiles (Fig. A1.4a). Notably, these included C_P, C_R and most of the residues changing conformation in the FF to LU transition: the C_P-loop, the N-terminal part of helix α2 ($\alpha2_N$) and helix α3. Overall, we identified 26 residues that fit a global 2-state model for chemical exchange with a k_{ex} = 72 s^{-1} and a pE = 2.5%. These parameters imply a rate of transition from the ground to the excited state (k_{GE}) of 1.8 s^{-1} and a rate from the excited back to the ground state (k_{EG}), of 71 s^{-1}. The exchanging regions were generally clustered between residues 43 - 55, and 75 - 100 (Fig. A1.4b and SI Table A1.3). Residues Gly21 and Val143 are outside of this cluster, but their exchange is reasonably explained by the C_P-thiolate crystal structure: the amide hydrogen of Gly21 hydrogen bonds to the Asp81 carboxylate (in helix α3), and the sidechain of Val143 packs against the $\alpha2_N$ (Fig. A1.4b inset).

In addition to the 26 residues described above, Asp96 appeared to be in a system of 3-state exchange, with two separate minor states (SI Fig. A1.S4). However, no other residues within the protein showed evidence of such exchange, so we could not confidently fit this profile to a 3-state model of exchange. Given its proximity to residue 99, which fits well to a model with a high k_{ex} (SI Table A1.3), this residue provides some evidence that the loop connecting strand β5 and helix α4 is in a 3-state exchange regime. A handful of residues in these regions that either did not fit to the global model (e.g. 96, 99), or were not measured unambiguously due to peak overlap (e.g. 50, 51), are shown in gray in Figure A1.4b. Additional residues in the region for which the CEST

profile showed no evidence of an excited state (e.g. 77, 79, 88 and 91) may have a $\Delta\omega$~0, such that no excited state peak would be observed.

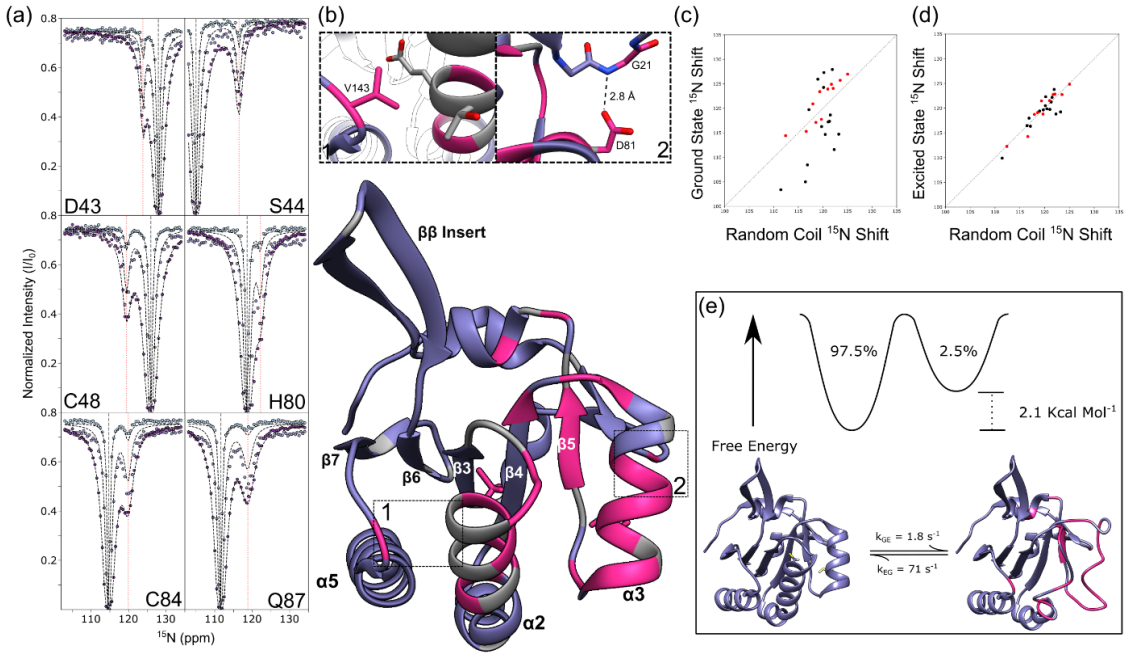


Figure A1.4: Preexisting local unfolding equilibrium of helices α2 and α3 revealed by CEST. (a) CEST profiles at B1 frequencies of 50 Hz (dark purple), 25 Hz (light purple) and 10 Hz (blue) are shown for six clearly exchanging residues along with curves based on the global fit (dashed lines). (b) C_P-tholate ribbon diagram of C_P-thiolate *Xc*PrxQ (PDB code 5IIZ) distinguishing the 26 residues fit to the global exchange model (pink) from residues for which no exchange was detected (purple) or which were not evaluated due to lack of assignment, overlapping or otherwise unfittable CEST profiles (grey). Inset boxes show the interactions of Val143 and Gly21 with residues from helices α3 and α2, respectively. (c-d) Scatter plot of ground state (c) and CEST-measured excited state (d) chemical shifts versus those expected for a random coil (Kjaergaard and Poulsen, 2011). The 11 residues not initially identified as exchanging but included in the global fit are indicated (red). (e) Cartoon summary of the kinetics (bottom) and thermodynamics (top) for the inferred transition between a ground state fully-folded C_P-thiolate conformation and an excited state conformation with 26 residues (pink) experiencing different environments due to the local unfolding of helices α2 and α3.

With these CEST experiments revealing that the $\alpha2_N$/α3 region of the protein is in equilibrium with a distinct higher energy conformation, we next asked what could be learned about the conformation of the excited state. A comparison of the ground state and excited state ^{15}N chemical shifts with reference chemical shifts for a random coil[165] shows that while the ground state shifts are highly dispersed (Fig. A1.4c), the excited

state shifts consistently match the random coil values (Fig. A1.4d). This provides evidence that the excited state of this region is largely unfolded, consistent with this process representing a preexisting FF to LU transition in the C_P-thiolate form of the protein.

Hydrogen exchange reveals a slow-exchanging core distinct from catalytic helices
We assessed solvent exchange at multiple timescales, using CLEANEX[166] to characterize very fast exchanging amides and conventional hydrogen-deuterium exchange experiments[167] to measure the slower exchanging amides in the range of minutes to weeks. Based on their exchange rates, we grouped 140 (of 149) assigned residues in the C_P-thiolate form and 66 (of 68) residues in the disulfide form into four classes (Fig. A1.5): very fast (CLEANEX detected, log(P) ~ 0), fast(CLEANEX invisible, 0 < log(P) < 2.5), intermediate (2.6 < log(P) < 4.5), slow (4.5 < log(P) < 7), and very slow (log(P) > 7) as described in the methods. In the C_P-thiolate structure, the fast exchanging amides are at the N and C termini and at the ββ insert, matching well with the residues having low order parameters in the modelfree analysis (Fig. A1.3a). Interestingly, the active site residues, Arg123 and Cys48 (C_P), also exchange on this timescale. Both these amides are solvent-exposed, with the C_P amide notably involved in coordinating the incoming peroxide, but the very rapid exchange implies not just a static exposure, but a localized flexibility that allows for formation of the chemical intermediates involved in amide exchange chemistry[168,169].

In the disulfide form (Fig. A1.5), most CLEANEX-visible residues are in the ββ insert and at the termini, similar to the C_P-thiolate form. Over-all, the protection factors are lower by one to three orders of magnitude in the disulfide state, indicating some loosening of the structure However, the pattern of exchange –regions of the protein that are slow-exchanging and those that are fast-exchanging – is the same between the two proteins. There remain some “very slow” exchanging amides which cluster to form a smaller stable core that is still centered on the ‘lower’ halves of β3 and β6 and the central portion of α5 (Fig A1.5). As the ‘upper’ (i.e., closer to the peroxide-binding pocket) halves of β3, β4 and β6 do not have assigned resonances, we cannot know the extent to which they are protected. The majority of amides in the C_P-thiolate form that were “fast” exchanging (including those in the $\alpha2_N$ and α3 regions) are missing from the disulfide form spectra.

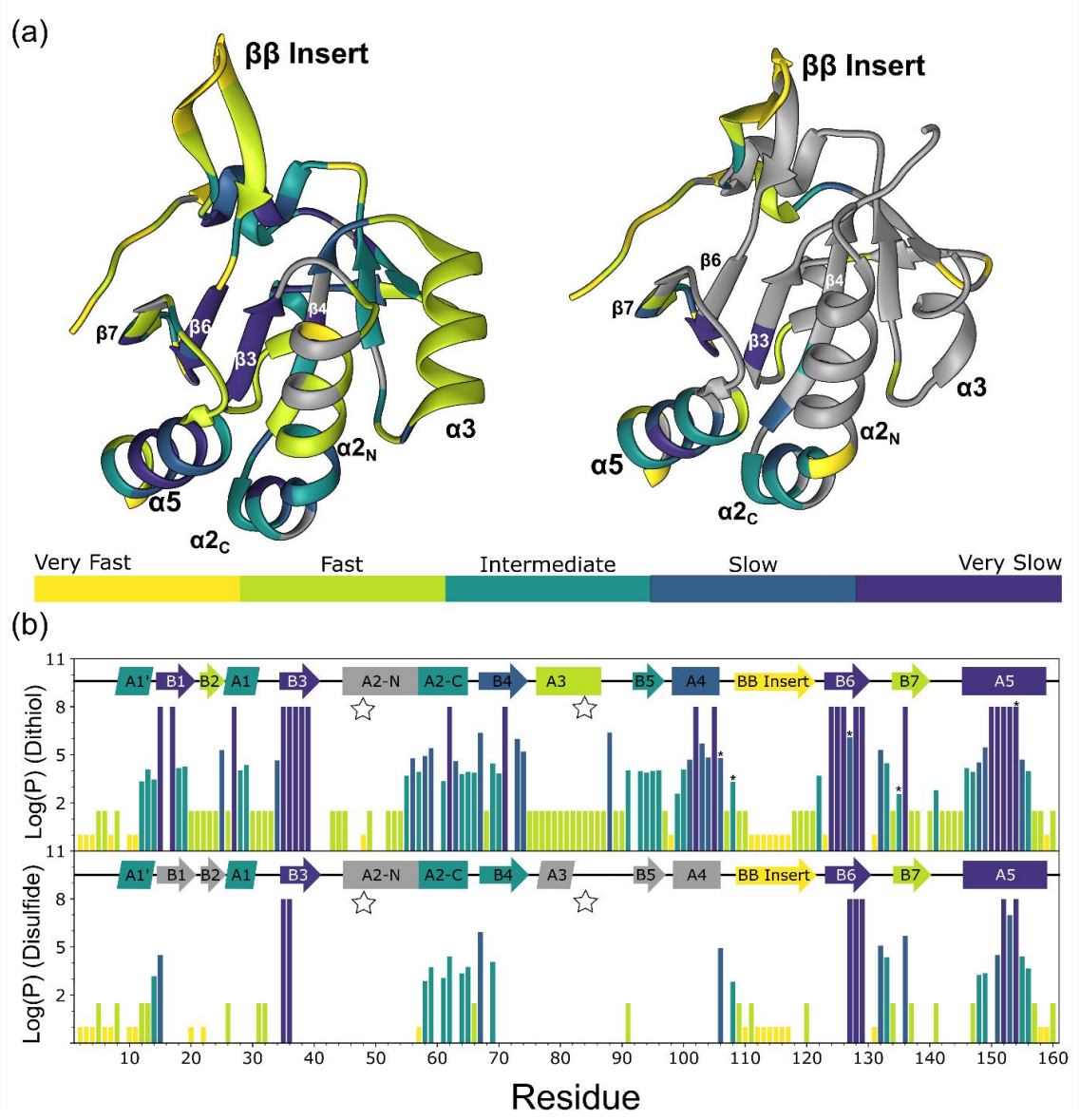


Figure A1.5: Hydrogen exchange behavior of the *Xc*PrxQ C_P-thiolate and disulfide forms. (a) Ribbon diagrams of C_P-thiolate (left) and disulfide (right) *Xc*PrxQ colored based on each residue's rate of hydrogen exchange (see methods): very fast (yellow), fast (green), intermediate (teal), low (blue), and very slow (purple). (b) Plots of logarithm of the protection factors for each residue seen in C_P-thiolate (top) and disulfide (bottom) *Xc*PrxQ, with the coloring as in panel (a). Residues with protection factors inferred despite ambiguities due to peak overlap are indicated (asterisks). A schematic of the C_P-thiolate secondary structure elements (as in Fig. 3) is provided at the top of each plot, with each element colored by its predominant exchange category.

Discussion

The three types of NMR measurements we have carried out probe a broad range of timescales and paint a consistent and extensive picture of the dynamics of *Xc*PrxQ, including the direct observation and detailed characterization of a catalytically-relevant local-unfolding equilibrium that occurs for the C_P-thiolate form on the ms-s timescale.

A comprehensive picture of C_P-thiolate Xc*PrxQ dynamics*

The hydrogen exchange behavior (Fig. A1.5) provides a useful framework on which to build an understanding of *Xc*PrxQ dynamics. The very-slow-exchanging core (log(P)>7) is the most stably-folded part of *Xc*PrxQ and centers on strands β3, β6 (with one residue each from β4 and β7) and includes residues from α5 and $\alpha2_C$ on one face of the sheet and from α1', β1, α1 and α4 on the other. Side chains from these elements form well-packed hydrophobic clusters on the front and back faces of the central part of the β-sheet. Assuming a protection factor of at least ~10^8 for these core amide groups implies an overall protein stability of ~11 kcal/mol or higher[170].

The slow-exchanging positions (4.5≤logP<7) include residues mostly adjacent to the very-slow-exchanging core residues. The intermediate group (2.5<logP<4.5) encompasses all the central elements of the fold except for helices $\alpha2_N$ and α3 along with strand β5 which is exposed by the unfolding of α3 and only has a single sidechain (Val94) contributing to the main hydrophobic core. The fast exchanging group (0≤logP≤2) includes helices $\alpha2_N$ and α3, exposed loops and edge strands of the sheet (β2 and β7), and finally, the fast (logP~0) exchanging residues are largely the N- and C-termini and the ββ-hairpin.

The relaxation experiments and modelfree analysis show that the N- and C-termini and the ββ insert (residues 110-117) are disordered on the ps-ns timescale (Fig. A1.3), fully rationalizing the fast solvent exchange (i.e., log(P)~0) in these regions. In addition, the details of the coordinated unfolding of helices $\alpha2_N$ and α3 revealed by the CEST measurements match well with the level of protection of these residues observed by hydrogen exchange. Specifically, the unfolding and folding rate constants of a 1.8 s^{-1} and 71 s^{-1}, respectively, mean that 1 in 40 molecules (~2.5%) are unfolded at any moment. This predicts a protection factor of P=40, or Log(P)= 1.6, which sits squarely within the 0≤log(P)≤2 range[171].

How the disulfide form dynamics differ

The dramatic loss of over half of the expected ^{1}H-^{15}N HSQC peaks in the disulfide form was taken to mean that even though only the C_P-loop and helix α3 locally unfold in the disulfide bonded form, their lack of tight packing impacted about half of the structure[163]. However, the minimal chemical shift differences among the assigned residues, especially in the very slow exchanging core, implied that the core structure does not significantly change in the disulfide form. Rather remarkably, the spin-relaxation measurements reveal no substantive differences in fast-timescale motions for the observable parts of the disulfide structure, with the only significant differences in motion found at the borders of missing segments (e.g. Ala57, Phe91).

Similarly, the hydrogen exchange measurements show a conservation of both the highly dynamic nature of the N- and C-termini and ββ-hairpin and the location of the slow-exchanging core, which remains centered on strands β3, β6 and helix α5 (Fig. A1.5). Since 7 residues remain in the very-slow exchanging group, it is even possible that the stability of the fold as a whole is unchanged. But, a lowered stability of much of the protein is evidenced by many residues near the core – such as in β1, $\alpha2_C$ and other α5 residues – showing 1 to 3 orders of magnitude less protection from exchange.

Insights into Xc*PrxQ catalysis*

The coordinated conformational transition of helices α2 and α3 revealed by the CEST experiments unequivocally establishes that the C_P-thiolate form of *Xc*PrxQ has a preexisting catalytically relevant local-unfolding equilibrium. The close match of the excited state chemical shifts with random coil reference values (Fig. A1.4d) leaves no doubt that this truly is a folded-unfolded or order-disorder transition rather than a transition between two alternate fully structured conformations. This is reminiscent of the paradigm-setting work on dihydrofolate reductase showing how the energy landscape directs the system along a functional catalytic trajectory with each intermediate in the catalytic cycle sampling a low-lying excited state conformation that resembles the ground-state structure of the following intermediate[172].

For Prxs, the next catalytic intermediate is not the disulfide, but the C_P-sulfenate (Fig. A1.1a), and there is evidence that the locally-unfolded (LU) conformation is indeed the ground-state conformation for this form. In addition to structural evidence that C_P-sulfenate formation creates strain in the FF conformation[151], a recent stopped-flow kinetics

study of yeast Tsa1[147] showed that C_P-sulfenate form favors the LU conformation by over 100-fold (k_{LU} = ~65 s^{-1} and k_{FF} <0.6 s^{-1}). Since the Prx active site is well conserved, we propose all Prxs will be qualitatively similar in this regard. Assuming that for Prxs in general the C_P-thiolate form favors the FF conformation by at least 10-fold (to allow for efficient catalysis), we infer that C_P-sulfenate formation shifts the equilibrium toward the LU form by over 1000-fold.

The CEST results also reveal that the unfolding of helices α2 and α3 are tightly coupled, and validates the proposal of Horta et al., (2010) that, even though helix α2 is virtually unchanged between the C_P-thiolate and disulfide forms, disulfide formation does not just involve movement of the C_P loop, but goes through an intermediate that has helix $\alpha2_N$ largely unfolded. Also notable is that the FF-LU equilibrium constant, which is ~40-fold (or ~2.2 kcal/mol; Fig. A1.4d) in favor of the FF conformation strikes an effective balance that allows for near maximal activity of the enzyme (with >97% of the enzyme with an FF active site), while also ensuring that the unfolding required for disulfide formation can readily occur.

Interestingly, two residues (48 and 49) that exhibit both high modelfree R_{ex} terms and elevated $R_{1\rho}$-derived R_2 rates (SI Fig. A1.2) are associated with the active site and may indicate an additional dynamic process in the first turn of helix α2 that is faster than the local unfolding process identified by our CEST experiments. Furthermore, increased solvent exchange at residue Cys48 and Arg123 (which is involved in the same hydrogen bond network) vs others that are directly involved with the FF-LU transition is consistent with an additional faster timescale motion. We propose these motions are related to suboptimal H-bonding interactions in the resting FF active site structure (Fig. A1.6a), including possible rearrangements in the active site H-bonding network that could be associated with water binding in the place of hydrogen peroxide and/or protonation/deprotonation of the C_P-thiolate. Since this stretch of the backbone has a highly conserved conformation among Prxs[146], we propose that such fluctuations will be a common feature of Prx family members and that they reflect an intrinsic marginal stability of the empty Prx active site pocket that is important for the optimal transition state stabilization needed for the high peroxide specificity and catalytic power. Such a role for

suboptimal interactions in promoting catalytic power is similar to what we have proposed for the enzyme urease, which also binds a small highly polar substrate [173].

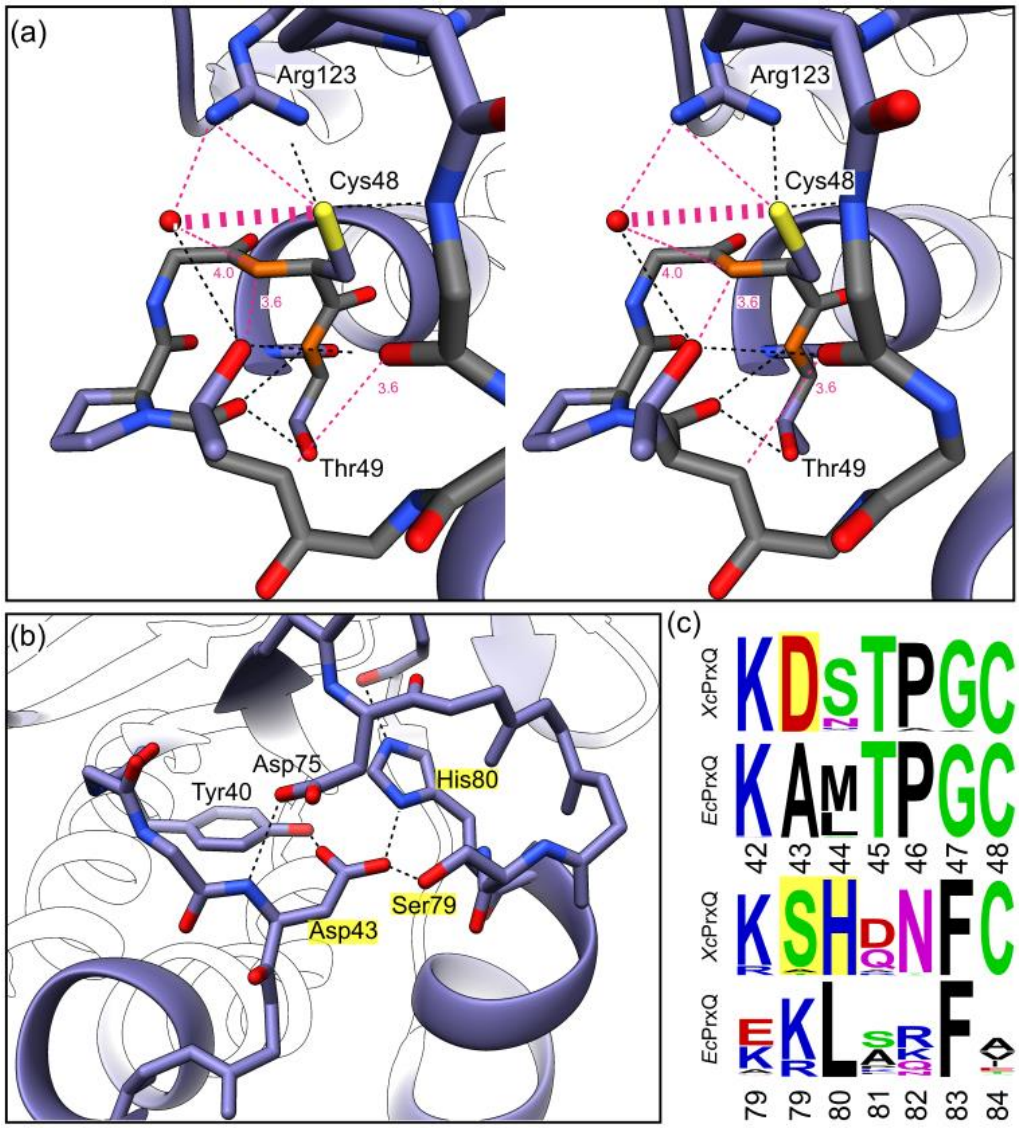


Figure A1.6: Suboptimal interactions in the Prx active site and a possible trigger linking the unfolding of helices α2 and α3. (a) Stereoview of the active site region highlighting three amides (Thr45, Cys 48 and Thr 49). Cys48 and Thr49 (orange) have large R_{ex} terms, and high $R_{1\rho}$-derived R_2 rates. Nearby H-bonds are shown (dashed lines) with suboptimal (long or non-linear) H-bonds highlighted (magenta, with distances if >3.5 Å). A close approach (3.2 Å) between the water and the C_P-sulfur is also highlighted (wide dashes). (b) H-bond network surrounding interactions of Asp43 in the C_P-loop with Ser79 and His80 at the start of helix α3. All H-bonds shown are <3 Å. (c) Conservation pattern seen near residues 43 and 80 in PrxQ sequences ≥60% identical to either *Xc*PrxQ with C_R in helix α3 (top) or EcBCP with C_R in helix α2 (bottom). See methods for details. Logos were generated using the online WebLogo tool[174].

Not all X*cPrxQ dynamic properties are common to Prxs*

Given the universal importance of transient unfolding events to Prx function, it is important to discern which properties observed here might be true for Prxs in general and which are unique to *Xc*PrxQ and its close relatives. Even though Prx dynamics are little studied, two other Prxs – yeast Ahp1[175], and *Arabidopsis thaliana* PrxQ (*At*PrxQ)[161] – have been studied by NMR and provide useful comparisons. As Ahp1 and *At*PrxQ have their C_R residues in different places than both *Xc*PrxQ and each other, their properties can provide insight into possible function-related differences in Prx dynamics. For Ahp1, C_R was later shown to be in an N-terminal loop[176], and by NMR, this C_R-containing loop was highly disordered making it readily available for disulfide formation. The C_P-loop and start of helix 2 included 5 unassigned residues and showed little protection from hydrogen exchange consistent with a localized conformational plasticity. For *At*PrxQ, a striking difference – which we noted earlier[163] – is a global behavior opposite to *Xc*PrxQ with the C_P-thiolate form spectrum missing over half of the peaks (including C_P and C_R) and the disulfide form mostly assigned. The missing assignments in the C_P-thiolate form include most residues of helix α2 (both in $\alpha2_N$ and $\alpha2_C$) and segments surrounding it (α5, β3, the C_P-loop and the loop preceding β6; *Xc*PrxQ numbering) and are plausibly all due to a preexisting catalytically-relevant local unfolding equilibrium centering on the whole of helix α2. Consistent with this, crystal structures of other Prxs with C_R in helix 2 showed that $\alpha2_C$ has high B-factors in the C_P-thiolate form and becomes more ordered in the disulfide form[177].

While the existence of FF-LU related dynamics thus appears to be common among Prxs, a notable difference in the *At*PrxQ dynamics is that the preexisting FF-LU interchange in C_P-thiolate *At*PrxQ appears to be at least 20-times faster [161] than the exchange rate we observed in this work for *Xc*PrxQ. Another observation is that the regions involved in the FF-LU transition differ in a manner consistent with the location of C_R, again illustrating how the energy landscapes of Prxs have evolved to specifically sample a low-lying excited state conformation that resembles the structure of the following intermediate[172]. In addition to parts of the structure surrounding C_R being dynamic, it is also notable that for both Ahp1 and *At*PrxQ helix α3 is a particularly stable part of the molecule. Similarly, for Prx1 subfamily members with C_R in a C-terminal tail segment, it has been shown that the unfolding of the C_P-loop specifically triggers an unfolding of the C_R-containing C-terminal tail region while again leaving helix α3 stably folded[150]. Looking for a “trigger” that could explain the coupled unfolding of the α2 and α3 helices in *Xc*PrxQ,

we find a buried H-bonding network between the side chains of Asp43, Ser79 and His80 that is conserved in this branch of the PrxQs group and not in PrxQs having their C_R in helix α2 (Fig. A1.6b,c), such as the PrxQ homologue from *Escherichia coli* (*Ec*PrxQ). We propose that the movement of Asp43 upon local unfolding of the C_P-loop destabilizes this buried position of His80 and triggers the coordinated unfolding of helix α3.

There also appear to be commonalities regarding Prx dynamics. Two related ones are first the intrinsic marginal stability of the FF active site loop in the C_P-thiolate form (mentioned above) that helps promote catalysis, and second, as we suggested above, an ~1000-fold destabilization of this loop associated with C_P-sulfenate formation. Such a Prx-wide intrinsic instability centered on the first turn of helix α2 implies that the macroscopic FF-LU equilibrium for any Prx, i.e. the "set point" that governs its sensitivity to hyperoxidation (see Fig. 9 of Perkins et al., 2013), will be a function of how well *other parts* of the folded protein stabilize the FF C_P-loop. This has already been well documented for those Prxs sensitive to hyperoxidation, in which interactions with other parts of the protein are behind the greater stability and slower-unfolding of the FF-conformation[142,147,178–180]. An awareness of this concept allows us to rationalize why *At*PrxQ has a much more dynamic C_P-thiolate form than does *Xc*PrxQ. Specifically, for *Xc*PrxQ the very stable $\alpha2_C$ portion of α2 is a well-ordered anchor that helps stabilize the FF C_P-loop (i.e., $\alpha2_N$), whereas for reduced *At*PrxQ this anchor is missing because $\alpha2_C$ is part of the conformational change. In reduced *At*PrxQ, $\alpha2_C$ is a very high B-factor region interacting only loosely with the protein core, which then switches to a lower B-factor region interacting more strongly with the protein core in the disulfide form (Fig. 6 of Perkins et al., 2012).

Functionally relevant shifts in the slow exchanging core in thioredoxin fold proteins

Looking beyond Prxs to the broader thioredoxin (Trx) superfamily, this study extends our earlier work[181] by providing another example of how the position of the slow exchanging core varies among Trx-fold proteins in ways that make functional sense. The *Xc*PrxQ core differs from both that of the *S. typhimurium* AhpF N-terminal domain (NTD) and that of *E. coli* Trx (SI Fig. A1.5). In Trx, the core includes the central beta sheet, especially strands equivalent to β3 and β6 in *Xc*PrxQ[182,183], with little participation of residues from helices, and the much more extensive NTD core, which is built on two tandem Trx domains, resides primarily in the catalytically inactive vestigial Trx domain and encompasses elements

equivalent to β3, β6 α2 and α5, meaning it is shifted toward the interface with the catalytically active Trx domain[181]. While the data are less extensive, the mass spectrometry-based hydrogen exchange profile of *Salmonella typhimurium* AhpC shows it has a similar slow exchange of the β3, β6 and α5 trio, but helix α3 is also part of the core and supports the decamer-forming interface[184].

Outlook

This study provides another example in which NMR dynamics measurements greatly extends insights beyond what could be learned from the static crystal structures themselves. In addition to answering the key question motivating this study – showing that C_P-thiol form of *Xc*PrxQ does indeed readily sample a catalytically-relevant LU conformation – this work sets the stage for follow-up NMR and kinetics studies of both *At*PrxQ and *Xc*PrxQ. These will target understanding the similarities and differences of the two wild type enzymes and correlating those properties with their catalytic properties, as well as answering additional questions of broad importance for the Prx field, such as characterizing how the commonly used mutations of C_P or C_R to Ser or Ala alter Prx dynamics and using C_R to Ala and/or Ser mutants to assess the impacts converting C_P-thiolate to C_P-sulfenate and C_P-sulfinate.

Methods

Protein Expression and Purification

Untagged *Xc*PrxQ encoded in the pTHCm plasmid[151] was transformed into the C41(DE3) *E. coli* cell line, and was expressed at 37° C in modified M9 minimal media, supplemented with 1 g/L $^{15}NH_4Cl$ and 2 g/L U-^{13}C glucose as needed. Cultures, on reaching OD600 ~0.6-0.7, were induced with 0.4 mM IPTG and harvested after 5-7 h. Cells were lysed by sonication, centrifuged for 1 h at 15,000 rpm (~26,000xg) in a sorvall fixed angle rotor, and the cleared lysate dialyzed overnight in 20 mM Tris pH 7.5. The dialysate was loaded onto a Macro-Prep High Q ion exchange resin (Bio-Rad, Hercules, CA) and eluted at NaCl concentration of less than 100 mM, as previously described[151,163]. The eluate was concentrated to 3 mL and further purified on a superdex HiLoad 75 column (GE Healthcare, Chicago, IL) in 20 mM Tris pH 7.3, 100 mM NaCl, and 1 mM sodium azide. The C_P-thiolate form was prepared by adding 5 mM dithiothreitol (DTT) and incubating for 20 min at room temperature. NMR samples were concentrated to 0.8 - 1.0 mM *Xc*PrxQ,

with a protease inhibitor cocktail (Roche Applied Sciences, Madison, WI), 7.5% D_2O, and 1 mM dimethylsilapentane-5-sulfonic acid (DSS) as a reference.

NMR Spectroscopy

Spectra were collected at 20° C on either a Bruker Avance III HD 800 MHz spectrometer equipped with a 5 mm triple resonance (HCN) cryogenic probe, or a Bruker Avance III 500 MHz spectrometer with a conventional triple resonance (HCN) probe. 3D HNCA, HNCACB, HN(CO)CA and HN(CO)CACB spectra collected at 800 MHz with non-uniform sampling on C_P-thiolate *Xc*PrxQ samples with additional uniform-^{13}C labelling confirmed published assignments and allowed us to assign residue Ser44. Spectra were processed in nmrPipe [185], using a sine-bell window function and one zero-filled point for every real point. We used scrub [186] for NUS reconstruction on 3D experiments.

Fast Dynamics

^1H-^{15}N HSQC-type experiments were used for measurement of ^{15}N R_1, ^{15}N R_2 rates and ^1H-^{15}N Heteronuclear NOE ({^1H}-^{15}N NOE) data at both 800 and 500 MHz using the same delays. Temperature compensated versions of published pulse sequences[187], were used for the ^{15}N R_1 and ^{15}N R_2 measurements, with the latter incorporating CPMG pulses for eliminating exchange contributions to R_2 relaxation. ^{15}N R_1 measurements were collected with delays in a range from 20 to 1200 ms, with a 60 ms delay collected in triplicate for error analysis. ^{15}N R_2 relaxation experiments used delays in a range from 17 to 237 ms, with a 34 ms delay collected in triplicate for error analysis. Peak intensities were fit to a single exponential decay in NMRViewJ[188] to determine the relaxation rates. Error was determined in NMRViewJ by the Monte Carlo method, utilizing the replicated measurements. {^1H}-^{15}N NOE experiments were collected using the[187] pulse sequence with an 8 s recycle delay at 800 MHz and a 5 s recycle delay at 500 MHz. I/I_o ratios were calculated in NMRViewJ. Uncertainty was estimated for intensities using the standard deviation of the spectrum noise level and propagated through I/I_o calculations. $R_{1\rho}$ rates were measured using a pulse sequence that incorporated adiabatic pulses and randomized phase-altered continuous wave 1H decoupling during the spinlock[189]. The spinlock field strength of 2902 Hz was measured using the off resonance continuous wave decoupling method[190]. The $R_{1\rho}$-derived R_2 rate was calculated from the measured $R_{1\rho}$ and

R_1 rate, using the equation $R_{1\rho} = R_1\cos^2\theta + R_2\sin^2\theta$, where $\theta = \arctan(\omega_1/\Omega)$. ω_1 is the spin-lock frequency and Ω is the offset from the carrier frequency[189].

Modelfree analysis[191,192] was performed in Relax[193,194]. Relax uses the extended modelfree analysis[191,192,195], and builds on the frequently-used[196–198] 5-model modelfree method to select a model from one of nine possibilities for each spin, guided by AIC[194,199,200]. The global model is then optimized in response to these models, and the process is repeated until convergence. This loop is carried out on each of four global diffusion models of increasing anisotropy, and final global model selection performed by AIC. The analysis incorporated ^{15}N R_1, ^{15}N R_2 and {^{1}H}-^{15}N NOE measurements collected at 800 and 500 MHz. Data were fit to a diffusion tensor which incorporated available crystal structures of the C_P-thiolate (PDB 5IIZ) and disulfide (PDB 5IOX) forms of *Xc*PrxQ. In the case of 5IIZ, we selected conformation A in places where multiple rotamers were present. For both proteins, the global model selected by this analysis was fully anisotropic – the 'ellipsoid' model, in relax terminology.

Chemical Exchange Saturation Transfer

Chemical Exchange Saturation Transfer (CEST) measurements used the ^{1}H-^{15}N HSQC-CEST pulse sequence from the Kay lab[201] adapted for Bruker spectrometers, including temperature compensation and a 90x240y90x 1H decoupling during the exchange time (T1). We collected spectra with a T1 delay of 400 ms and B1 frequencies of 10, 25 and 50 Hz. 63 CEST slices and one reference spectrum were collected, with B1 increments of 0.5 ppm in the ^{15}N dimension, stretching from 103 to 133.5 ppm.

Peak intensities were extracted from all 64 slices in NMRViewJ, and I/I_o ratios were calculated with in-house scripts. Uncertainty for peak intensities was approximated using the standard deviation of the noise floor of the spectrum and propagated through I/I_o calculations. For residues exhibiting an exchange profile in plots of I/I_o, ChemEx[201] was used to fit profiles to a 2-state system of exchange modelled by the Bloch-McConnell equation[201]. Our initial fit did not assume the existence of a global model, so for each residue, we fit data collected with B1 frequencies of 10, 25 and 50 Hz to a model which included six parameters: the relaxation rates $R_{2,a}$ $R_{2,b}$ and R_1, the difference in chemical shift between the ground and excited state $\Delta\omega$, as well as the rate constant for the overall exchange k_{ex} and the population of the excited state pE. After this initial fit, 15 residues with similar rates of exchange and pE values were fit to a global model, with $R_{2,a}$ $R_{2,b}$ R_1,

and $\Delta\omega$ fit for each residue, but k_{ex} and pE fit globally. Using the k_{ex} and pE from our global model, we then fit every residue in the region of the 15 residues (residues 43 to 55 and 75 to 100), and identified additional 9 residues and 2 other residues, Gly21 and Val143, both of which are in contact with regions containing a high number of residues undergoing exchange. These total 11 additional residues had small $\Delta\omega$ values *so* that the Lorentzian for the excited state was less notable in the CEST profile.

Hydrogen Exchange Measurements

Fast-exchanging amide protons were identified using CLEAN chemical EXchange (CLEANEX) pulse sequences with a mixing time of 100 ms[166]. Peaks determined to be CLEANEX-visible were those that could be detected above the noise level of the spectrum. For hydrogen-deuterium exchange (HD-X) experiments, samples prepared for NMR were freeze dried and then resuspended in a volume of D_2O matching the sample volume before lyophilization, following methods published elsewhere[29,181,202]. Resuspended samples were incubated at 20° C, and best-HSQC spectra were collected periodically until 48 h, with final spectra taken after 2 weeks. The midpoint of the first best-HSQC was 7.2 min after resuspension. Non-overlapped peaks that dropped substantially in peak height within 48 h were fit to rate constants for exchange obtained by a least-squares fit to a single exponential decay function, $I = I_0 * e^{-k_{ex}t} + C$ using in-house python scripts, where k_{ex} is the rate of amide exchange, and C is a flat offset to account for variations in the spectral noise floor. Uncertainties in the rate constants were calculated from the variance of the determined exchange rate. Protection factors (P) were calculated relative to the expected random coil exchange rate (k_{rc}) as provided by SPHERES[165] for residues exchanging on an amenable timescale (46 in C_P-thiolate, 21 in disulfide), with log(P) values ranging from ~2.6 to ~7 (Fig. A1.6b,c, SI Table 4 – See manuscript).

From these experiments, we grouped each residue into one of five classes: very fast, fast, intermediate, slow, and very slow exchanging. “Very fast” exchanging residues were those with detectable peaks in the CLEANEX spectrum (i.e. exchanging on the timescale of milliseconds). “Fast” exchanging residues were those not appearing in the CLEANEX spectrum, but exchanging fast enough that they did not have appreciable peaks in our HD-X experiments (i.e. exchanging on the sub-seconds to seconds timescale). For the residues that could be fit to an exchange rate (i.e. exchanging on a

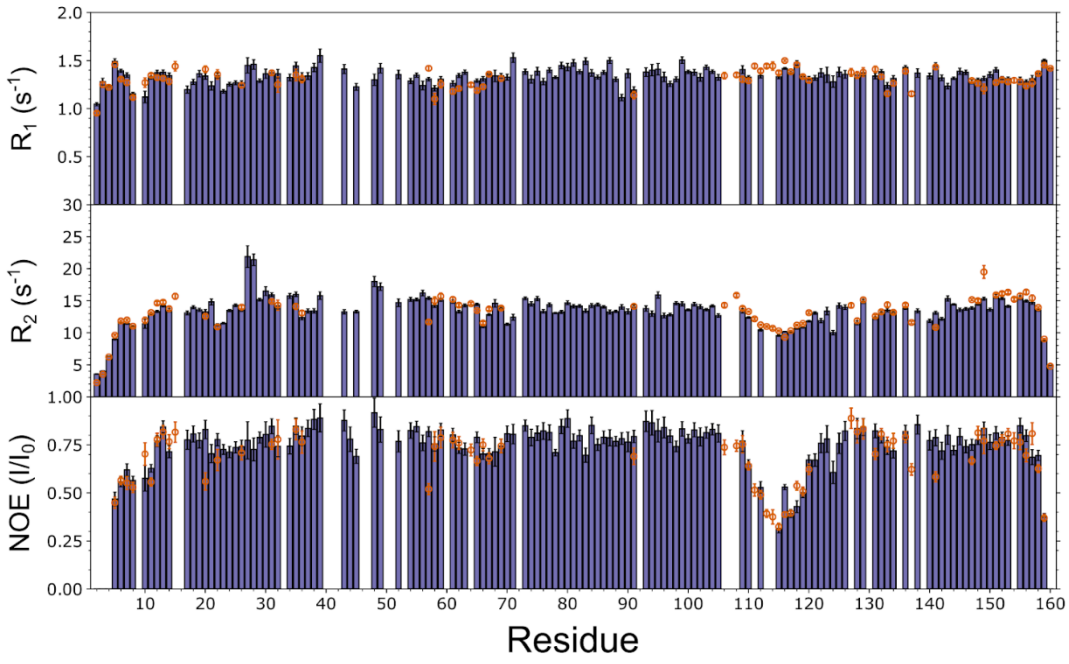
timescale of minutes to hours), those with 2.5<log(P)<4.5 were termed “intermediate” and those with 4.5≤log(P)<7 were termed “slow.” Finally, “very slow” residues were those for which the peak intensity decrease after 48 h was not enough to allow accurate fitting to an exchange rate (i.e. exchanging on the scale of days or longer). We assigned Log(P)=0 to residues in the fast exchanging group, because the timescales of reference intrinsic exchange rates for residues in a random coil is 1 - 50 s^{-1}, which spans the 100 ms mixing time of our CLEANEX spectra. Similarly, we assigned log(P)=1.5 to the fast group, as they are expected to range from 0<log(P)<2.5 since they exchange more rapidly than the fastest-exchanging residue in the intermediate category which had log(P)=2.6. Finally, we plotted the very slow exchanging class at a Log(P)=8, a value just larger than the upper end of the “slow” category.

*Xc*PrxQ and *Ec*PrxQ conservation analysis

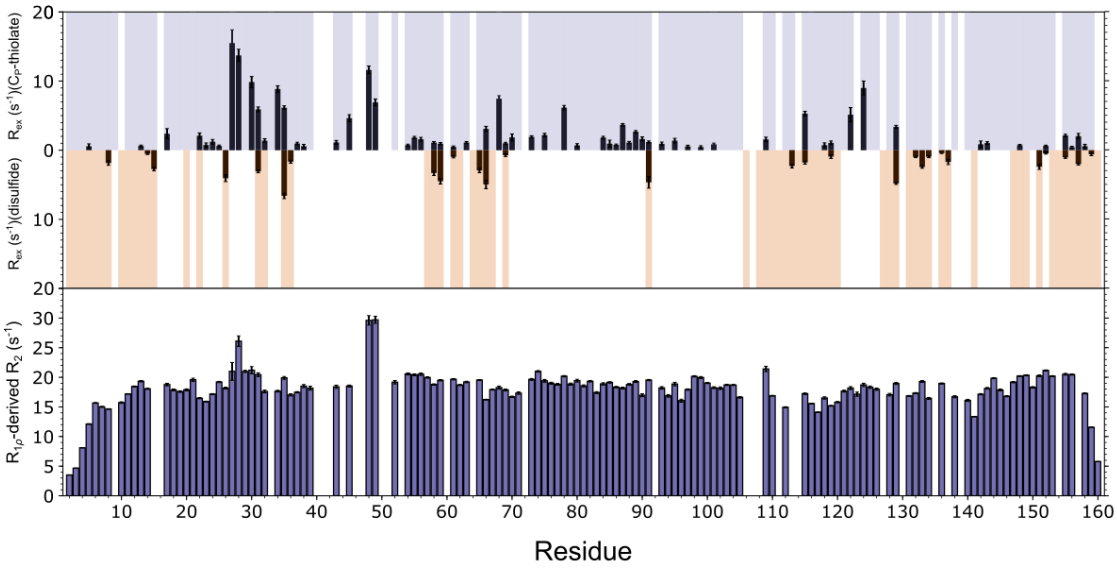
Amino acid sequences for proteins with 60% or greater identity to *Xc*PrxQ and EcPrxQ (originally called bacterioferritin comigratory protein or BCP), were retrieved from the NCBI using BLAST[35]. We clustered sequences using CD-HIT[36,37] at 90% similarity, which yielded 166 sequences related to *Xc*PrxQ, and 214 sequences related to *Ec*PrxQ. For both datasets, clustered sequences were aligned using MUSCLE[38], then filtered to remove all proteins which did not have a resolving Cys at the CxxxxC motif in *Ec*PrxQ or at the aligned position of C_R84 in *Xc*PrxQ.

Acknowledgements

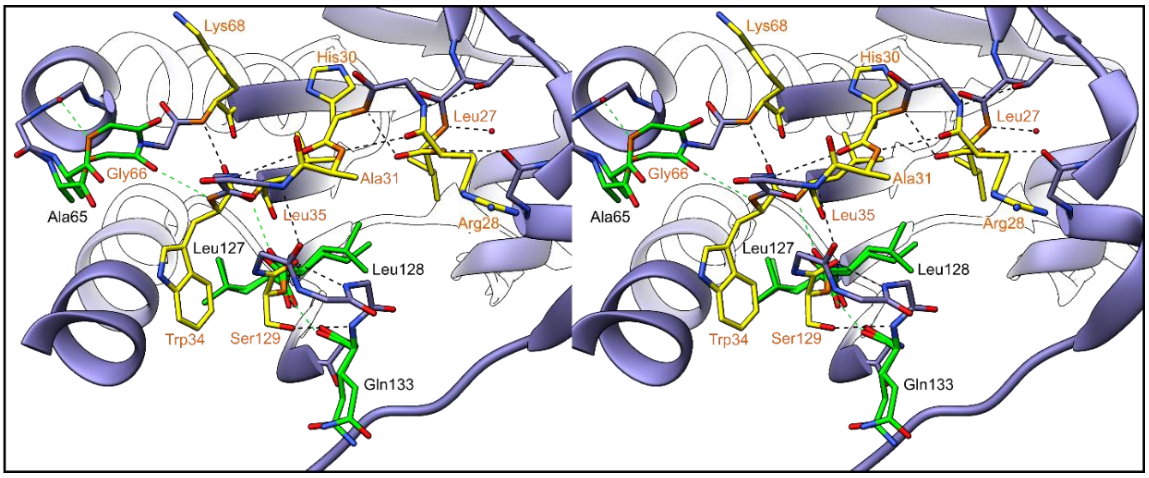
The authors would like to thank Garry Buchko for helpful discussion regarding assignment of NMR spectra. We acknowledge the support of the Oregon State University NMR Facility funded in part by the National Institutes of Health, HEI Grant 1S10OD018518, and by the M. J. Murdock Charitable Trust grant #2014162. This work was supported in part by the U.S. National Institutes of Health grants R01-GM119227 (to PAK and LBP), R35-GM135179 (to LBP), and the National Science Foundation Award 1617019 (to EB).



SI Figure A1.1: Spin-relaxation and {^{1}H}-^{15}N NOE values measured at 500 MHz. Shown are ^{15}N R_1 (top), ^{15}N R_2 (middle) and {^{1}H}-^{15}N NOE measurements for the C_P-thiolate (purple bars) and disulfide (orange circles) forms of *Xc*PrxQ. This is equivalent to Figure A1.3a, but showing data measured at 500 MHz.



SI Figure A1.2: R_{ex} terms for disulfide and C_P-thiolate *Xc*PrxQ and R2 rates calculated using $R_{1\rho}$ for C_P-thiolate *Xc*PrxQ. (a) Transverse relaxation due to exchange (R_{ex}) values determined by modelfree analysis are plotted as a function of residue number for the C_P-thiolate (black bars going up from zero) and the disulfide (black bars going down from zero) forms. Shading identifies residues which could be fit to a modelfree model, while blank spaces mark residues for which information was not available, either due to missing spin-relaxation data or a lack of convergence during modelfree analysis. (b) R_2 relaxation rates for the C_P-thiolate form of *Xc*PrxQ, calculated using measured $R_{1\rho}$ and R_1 relaxation rates collected at 800 MHz.



SI figure A1.3: Conformational heterogeneity seen in the crystal structure as a possible explanation for R_{ex} terms seen in C_P-thiolate *Xc*PrxQ. Stereoview of a central portion β-of C_P-thiolate *Xc*PrxQ (purple carbons) highlighting nine residues that have R_{ex} >3 (yellow carbons and orange amide nitrogen atoms and residue labels) and five nearby residues that were seen to adopt two conformations in crystal structure (green carbons). Relevant H-bonds involving backbone amides are shown (dashed lines), and those involving residues with multiple conformations are colored green.

SI

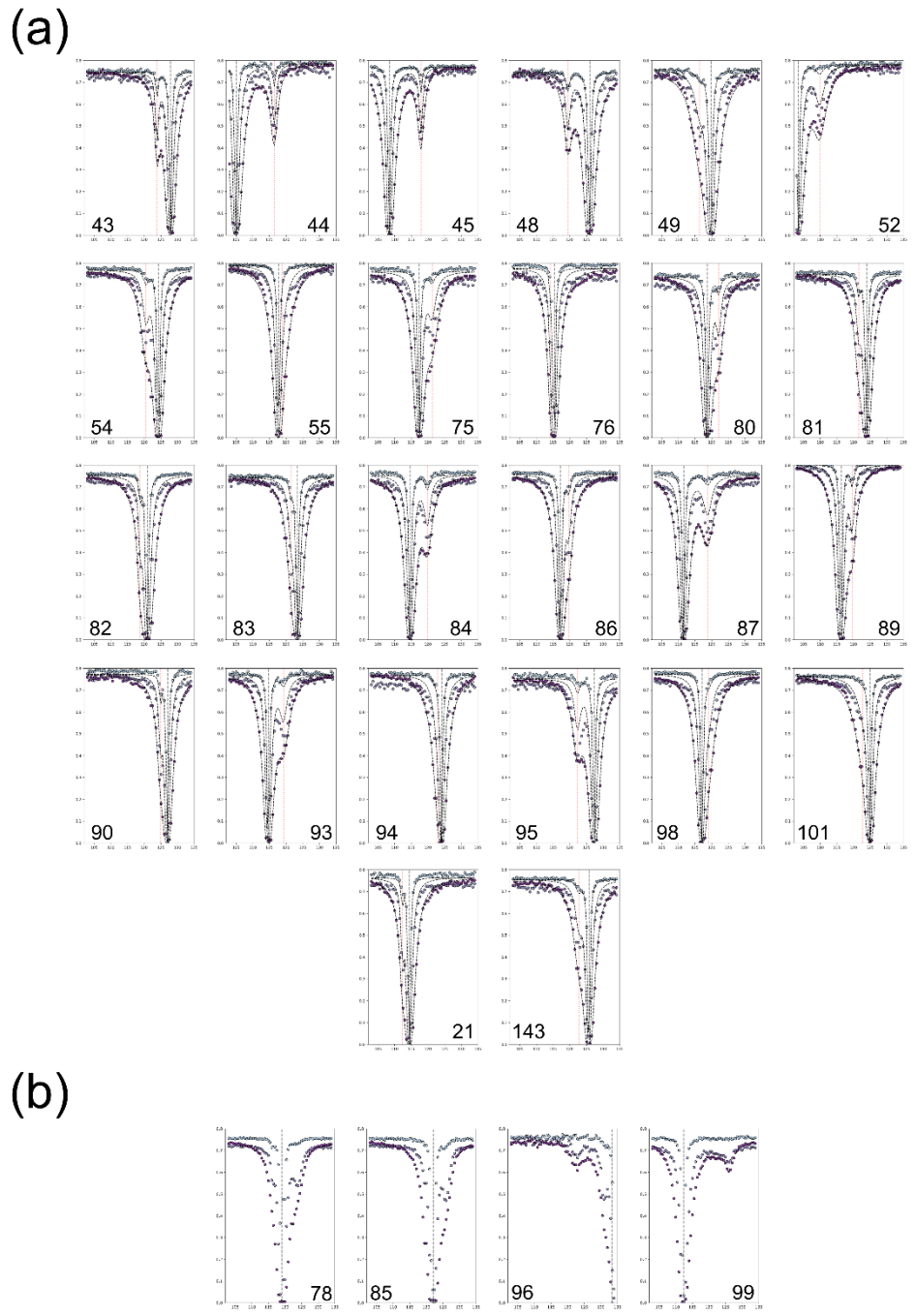
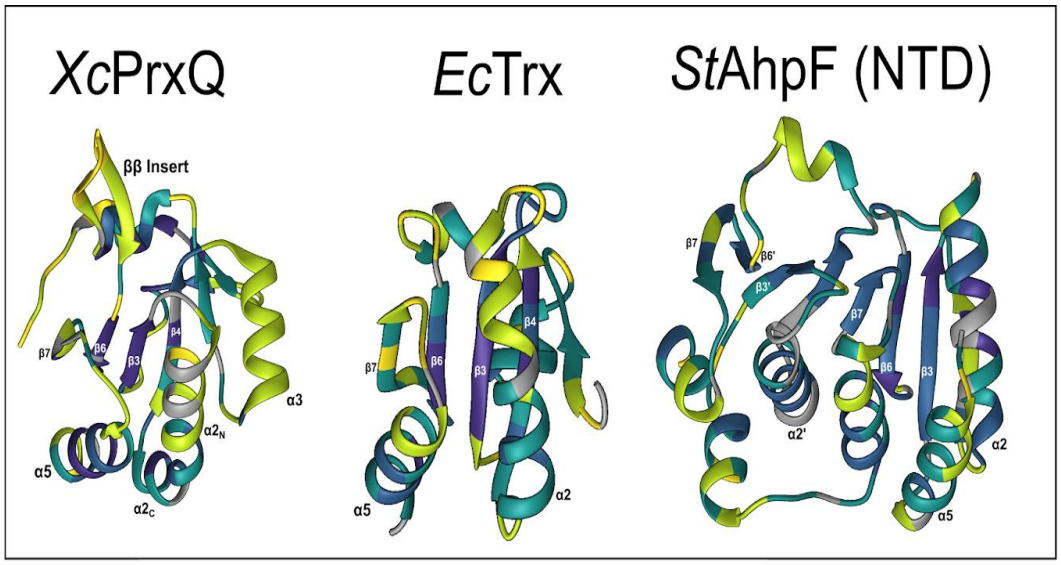


Figure A1.4: CEST profiles of C_P-thiolate *Xc*PrxQ. (a) CEST plots at B1 frequencies of 50 Hz (dark purple), 25 Hz (light purple), and 10 Hz (blue) for all 26 residues included in our global model, along with curves based on the global model (dashed lines). (b) The same as panel A, but for the 4 residues with clearly evident exchange profiles that did not fit well to the global model.



SI Figure A1.5: Comparison of hydrogen exchange in three thioredoxin fold proteins. The C_P-thiolate *Xc*PrxQ ribbon diagram (left) is shown colored as in Figure A1.5 by hydrogen exchange rate group from very fast to very slow (see Fig. A1.5). Ribbon diagrams of published hydrogen exchange rates for *Ec*Trx[182,183](center) and *St*AhpF NTD[181] (right) are colored similarly, but to aid the comparison by accounting for differences in global stability, the Log(P) boundaries on the groups were adjusted. For *Ec*Trx, the boundaries between Very fast (yellow), Fast (green), intermediate (teal) and slow (blue) are 0.5, 2.5, 4.5 and 6, respectively. For the *St*AhpF NTD, the cutoffs are 2.5, 3, 4.5, and 7, respectively. In *Ec*Trx and the *St*AhpF NTD, secondary structural elements are labelled based on names of homologous structural elements in XcPrxQ, with primes used to denote the C-terminal half of the NTD, as that protein contains two Trx domains.

SI Table A1.3: Model parameters and chi^2 values for residues fit to model of exchange.

Residue	Number	*kex*	± *kex*	pE	± pE	Chi^2
D	43	80	8	0.02	0.0013	1.56
S	44	85	17	0.019	0.002	2.95
T	45	90	6	0.0176	0.0008	2.57
C	48	77	10	0.024	0.0018	1.74
T	49	70	20	0.02	0.004	5.47
G	52	78	11	0.019	0.0014	1.47
D	54	80	10	0.03	0.002	5.32
D	75	66	7	0.032	0.002	1.95
K	78*	640	130	0.009	0.0009	56.1
H	80	97	15	0.028	0.003	5.22
C	84	71	6	0.026	0.0013	1.63
A	85*	60	24	0.015	0.004	35.46
K	86	93	8	0.022	0.001	5.07
Q	87	58	5	0.028	0.0017	4.32
F	89	72	6	0.024	0.0014	4.26
L	93	46	5	0.034	0.003	3.67
S	95	42	5	0.043	0.004	4.93
E	99*	1010	60	0.0051	0.0002	4.36

*based on these fits, marked residues were excluded from the global model

Appendix 2

Specificity and Heterogeneity RNA-binding Domains of the Sars-CoV-2 Nucleocapsid Protein

Aidan B Estelle, Heather M Forsythe, Zhen Yu, Kaitlyn Hughes, Brittany Lasher, Patrick Allen, Patrick Reardon, David Hendrix, Elisar J Barbar

Introduction

The ongoing public health crisis known as the COVID-19 pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has, since origination in 2019, killed over 6 million worldwide at this time of writing[203]. While development of vaccines has aided the crisis to a degree, appearance of vaccine-resistant mutants has highlighted the potential value of therapeutics that target highly conserved structures and interactions. One potential target is the nucleocapsid (CoV-N) protein. The most abundant protein in infected cells[204], CoV-N interacts with viral RNA, binds to the membrane protein (CoV-M)[205,206], interferes with the host immune response through binding to INF-β[207], and inhibits host cell proliferation through binding to EF1α[208]. The most essential function of CoV-N is its interaction with viral RNA, however: The protein coats the 32 kb genomic RNA (gRNA), protecting the RNA and promoting assembly of the ribonucleocapsid (RNP) complex[209]. N-RNA interactions are essential for genome packaging, virion assembly, viral transcription, and viral replication, making structural understanding of N-RNA interactions essential to understanding how SARS-CoV-2 functions[204].

Structurally, the protein contains a mix of folded domains and disordered linkers (Fig. A2.1a). The protein forms a dimer structure binds multivalently to CoV-2's genomic RNA, having been shown to bind RNA at both domains as well as at disordered linkers[210,211]. The N-terminal domain (NTD) spans residues 44 to 182, and takes a predominantly beta structure, forming a core beta sheet with surrounding flexible loops. The core beta sheet, and the beta hairpin above it, form a cup-like shape that is positively charged, and believed to be the RNA-binding site of the domain[212–214] (Fig. A2.1c). The C-terminal domain (CTD) spans residues 247 to 366 and takes a stable dimer structure, forming a disc-like shape[215]. The CTD structure can be divided into an alpha and beta face, corresponding to the predominant secondary structure at each (fig. A2.1d). The beta face forms a single sheet linking together the two CTD domains, while the alpha face consists of a series of interlocking helices. The alpha face is thought to be the RNA-binding face of the protein, due in part to its strong positive charge[215] (Fig. A2.1).

The nucleocapsid protein is known to phase separate with RNA, driven by multivalent interaction with domains of CoV-N to form liquid droplets[206,216,217]. While the exact details are not yet clear, phase separation is thought to be critical to viral function, and infected cells have been shown to contain droplets of N-RNA condensates[218]. One hypothesis suggest that phase separation allows for processing of the viral genome,

preventing nucleocapsid formation until required[216]. Evidence suggests that phase separation may be RNA sequence or structure specific, as some RNAs appear to induce separation to a greater degree than others[219]. The mechanism of separation is unclear, however, due in part to a lack of detailed information on the structural specificity of RNA-N interactions.

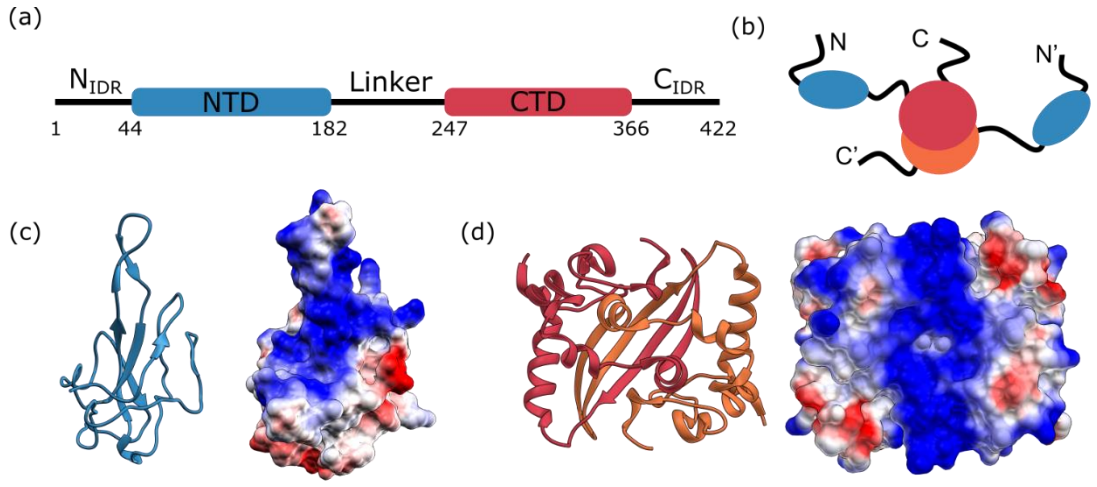


Fig. A2.1: Structure of the SARS-CoV-2 Nucleocapsid protein. (a) Domain architecture of CoV-N. Folded domains are shown in blue and red, while disordered linkers are black lines. (b) Cartoon diagram of CoV-N structure, showing dimerization at CTD. (c) Crystal structure of the CoV-N NTD (PDB 7CDZ), shown as a ribbon diagram (left) and protein surface (right), colored by surface charge calculated in chimera. (d) Crystal structure of the CoV-N CTD (PDB 7CE0), shown as a ribbon diagram (left) and protein surface (right), colored by surface charge calculated in chimera.

The NTD has been shown to bind both single stranded RNA (ssRNA) and double stranded RNA (dsRNA) in studies examining RNAs of length 7-30 nucleotides[212–214]. Mutation studies reveal that binding is strongly driven by charge, as mutations that reverse the NTD's positive charge negatively impact binding[213]. Structural studies confirm that that both RNA structures interact with the positively charged cup of the NTD (Fig. A2.1c)[213]. However, nuclear magnetic resonance (NMR) investigations reveal qualitative differences between the interaction of the NTD with ssRNA and dsRNA, suggesting that the mode of binding is dependent on the strand state of the RNA[213]. Docking models suggest that the more flexible ssRNA wraps around the beta hairpin at the center of the NTD's binding site, forming a U shape, while dsRNA sits on one side of the binding domain[213]. Whether the NTD has a preference for a distinct RNA state (i.e. ssRNA or dsRNA) remains unknown, along with what the downstream function of any RNA preference.

The CTD also binds to RNA, thought to be mediated through its positively charged face (Fig. A2.1d). However, much less structural information on CTD-RNA interaction is available. Gel shift assays have demonstrated that the CTD shifts with ssRNA, ssDNA and dsDNA[215], pointing to a relatively non-specific binding interaction, and fluorescence anisotropy experiments between the CTD and RNA suggest that the two interact weakly, further pointing towards nonspecific binding[220]. Molecular dynamics simulations of the CTD bound to RNA suggest that the N-terminus of the domain is critical to binding to RNA, but do not provide evidence of any significant conformational change in the CTD upon RNA binding[221].

In this work, we examine binding between the CoV-N domains and RNA, looking first at the first 1000 nucleotides of the CoV-2 genomic RNA (g1-1000) and then at small 14-nucleotide RNAs. We find that the NTD binds to ssRNA with a greater degree of specificity and binds nonspecifically to dsRNA. Further, a mutation of the NTD thought to damage NTD-RNA interactions shows an increase in nonspecific binding, suggesting that the NTD is in balance between specific and nonspecific interactions. For the CTD, we confirm that the protein interacts only weakly with both ss and dsRNA, binding slightly tighter to dsRNA. Examining the propensity of each domain to form phase separated droplets, we find that the degree of weak, nonspecific interaction correlates closely with the degree of observed phase separation, suggesting separation is driven by weak multivalent interactions, and that the NTD in particular is capable of both specific and nonspecific binding, providing structural evidence for the dependence of phase separation character on specific RNAs.

Results

The CoV-N domains bind 1-1000 genomic RNA at their positively charged faces

To examine interactions between the CoV-N domains and RNA, we first took NMR measurements of the domains following addition of the first 1000 bases of the CoV-2 genome (g1-1000). The RNA does not encode any viral proteins and is thought to be a primary driver of condensation in the virus[211]. We have previously examined interactions between the full length nucleocapsid protein (FL-N) and g1-1000 and determined that the NTD tightly binds g1-1000[211]. Here, we examine each domain in isolation, and find that both domains bind to g1-1000. Both RNA-bound spectra exhibit no significant chemical shift perturbation, but dips in peak intensities, indicative of slow exchange with a large

bound state. The peak intensity ratios (Fig. A2.2a,c) report on the proximity of each residue to the site of binding, allowing us to draw a crude map of binding by mapping intensity ratios the protein structure.

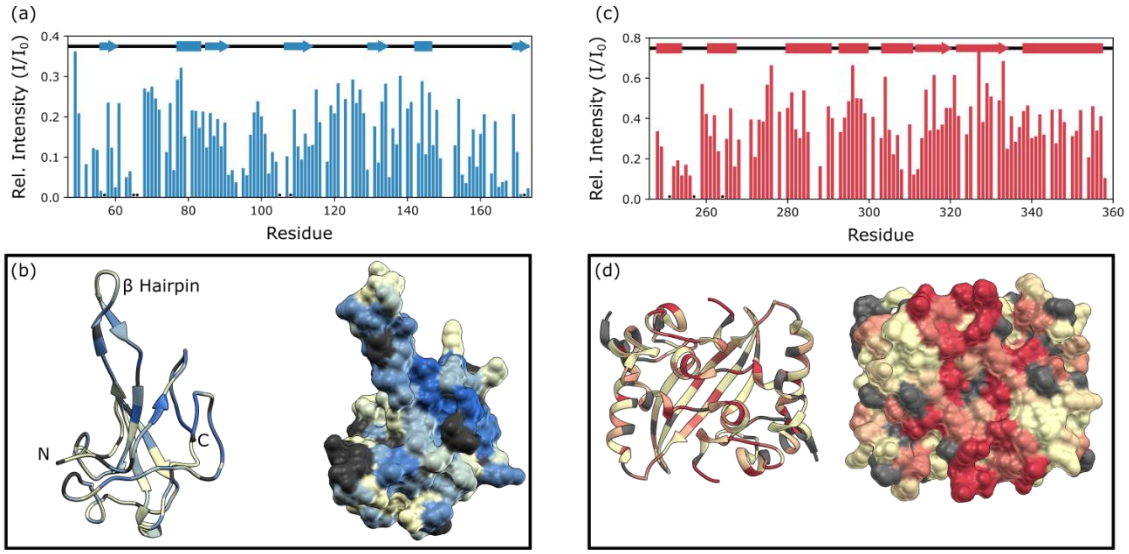


Figure A2.2: Binding between g1-1000 RNA and CoV-N domains. Peak intensity ratios for the NTD (a) and CTD (c) bound to g1-1000 RNA at a ratio of 100:1 protein:RNA. Peaks which disappear completely are marked with stars at the bottom of the plot. Structures of the NTD (b) and CTD (d) with peak intensity ratios mapped to their structure, in the form of both ribbon diagrams (left) and maps of the protein surface. For the NTD, blue represents an I/I_0 of 0, and tan represents an I/I0 above 0.25. For the CTD, red represents an I/I_0 near 0, and tan represents an I/I0 above 0.4. For both structures, unassigned or overlapped residues and prolines are colored gray.

The NTD binds RNA in the positively charged groove (Fig. A2.1b) formed around domain's β Hairpin. The sharpest decrease in peak intensities can be seen at the base of the β hairpin at residues 91-96 and 101-107, as well as at the β strand at the N terminus of the protein, and the C-terminal strand to the right in Fig A2.2b (Fig. A2.2a,b). This agrees well with previously published work, including NMR studies of RNA-protein interactions of the NTD[212,213], and docked models of the NTD-RNA complex which both indicate that the RNA-NTD complex forms at the charged surface of the NTD[213].

The CTD binds RNA along the positively charged alpha face (Fig. A2.1d) of the CTD dimer. Here, peak intensity drops are greatest at the N-terminus of the CTD (residues 247-270), as well as at the helical region 305-314 (Fig. A2.2d). Both these regions are on the alpha face of the domain, indicating binding likely happens along this face. The N-terminal end of the CTD (247-252), which retains flexibility and does not appear in crystal

structures, also dips significantly in intensity, suggesting it is also involved in RNA binding and confirming initial molecular dynamics studies which indicated the importance of the N-terminus of the CTD to RNA binding[221].

The CTD drives phase separation with g1-1000

To examine the role that each CoV-N domain plays in phase separation with RNA, we fluorescently labeled g1-1000 and incubated the RNA with samples of each nucleocapsid domain, as well as the full length CoV-N protein (FL-N). At the initial tested conditions, we found that the both the FL-N and the CTD phase separate with g1-1000 at 37 C. The NTD, in contrast, did not phase separate at the tested conditions with g1-1000. To further explore the phase space, we prepared samples of each protein with g1-1000 at varied pH, temperature, and concentrations (SI Figure A2.1). At all tested conditions, the CTD formed droplets with g1-1000, while the NTD did not. FL-N appeared most sensitive to phase conditions, forming droplets most readily at low concentrations, high temperatures, and neutral pH. The tendency of the CTD to phase separate under all conditions suggests that the CTD may be the dominant driver of liquid droplet formation in the nucleocapsid protein, to the point that it phase separates very readily, and presence of both domains together in FL-N construct allows for separation to be condition dependent.

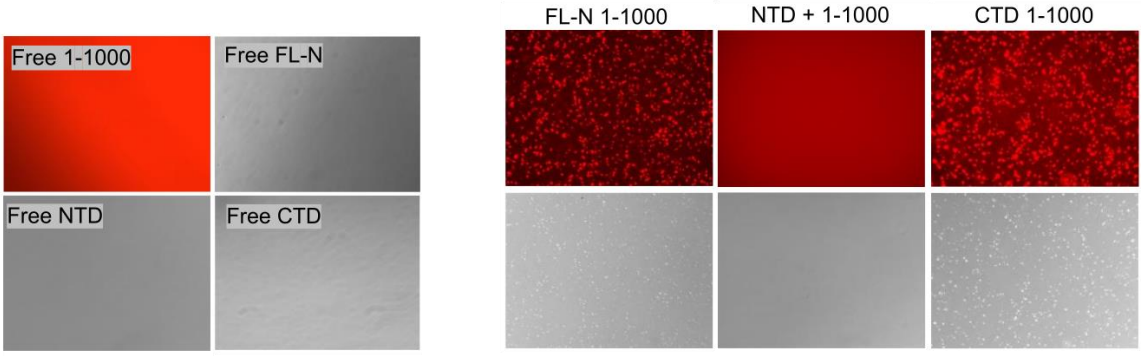


Figure A2.3: Phase separation of CoV-N with g1-1000 RNA. (left) Samples of free g1-1000 RNA, along with each CoV-N construct – FL-N, NTD and CTD. (right) Fluorescence (top) and brightfield images of FL-N, NTD and CTD mixed with g1-1000 RNA at 3.5 μM protein and 50 nM g1-1000.

The N-terminal domain preferentially binds ssRNA

Interested in examining the domain-RNA interactions at a detailed structural level, we synthesized a 14-nucleotide RNA fragment (ss-14mer), along with a reverse complement for the formation of a double stranded fragment (ds-14mer). Our intention in the sequence

design was to minimize the possibility of internal base pairing within the 14-base fragment, and indeed, ^{1}H NMR spectra of the ss-14mer shows no peaks in the paired imino proton region. Upon annealing with the reverse complement sequence, the fingerprint of 1H peaks in the 6-8 ppm region changes, and new peaks characteristic of paired imino protons appear at 11-15 ppm, indicating the sample contains only the ds-14mer.

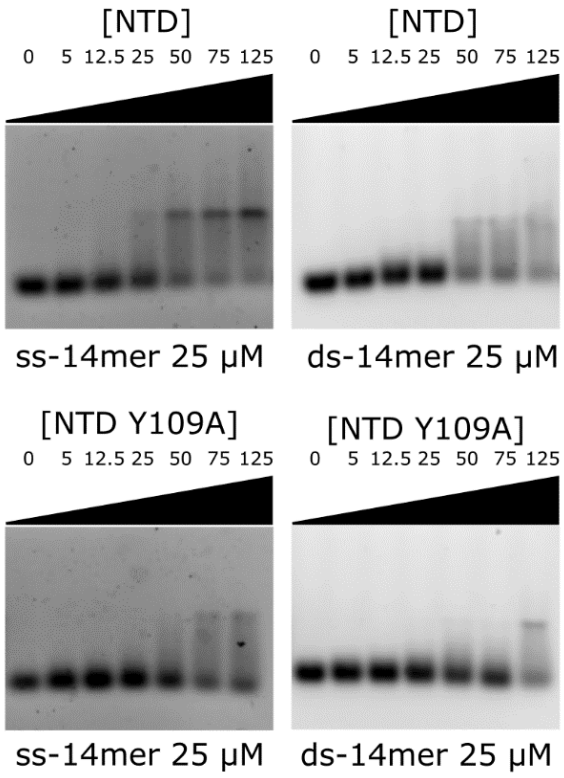


Fig. A2.4: EMSA assays of NTD with ss-14mer and ds-14mer. (a-d) EMSA of the NTD (top) and Y109A mutant NTD (bottom) with the ss-14mer (left) and ds-14mer (right). All lanes contain 25 μM 14-mer, and NTD at ratios of 0,0.2,0.5,1,2,3 and 5:1. The NTD concentration in μM is listed above each lane.

The NTD binds to the ss-14mer more tightly than the ds-14mer. To provide an estimate of the relative binding affinity between the NTD and our 14-mer RNAs, we performed electrophoretic mobility shift (EMSA) assays on both RNAs, titrating the NTD into each (Fig. A2.4). The NTD binds both RNAs, resulting in bands shifting up and smearing in the agarose gel. Examination reveals that the ss-14mer binds more tightly to the NTD, an effect that is particularly apparent at the 1:1 (25 μM) titration point, where the ss-14mer is smeared along the lane, and the ds-14mer remains predominantly on the free-RNA band.

Titrations of the ss-14mer and ds-14mer RNAs into ^{15}N labeled samples of the NTD reveal intermediate exchange at the binding interface. The titrations induce a mix of fast and intermediate exchange in the NTD spectrum, resulting in a heterogeneous mixture of peak behaviors, including ~30 peaks that disappear from the spectrum early in the titration for both the ss-14mer and ds-14mer. Mapping the disappearing peaks to the structure (Fig. A2.5c) reveals they are located at the RNA-binding interface, suggesting the intermediate exchange is induced by the protein-RNA interaction. For fast-exchanging residues, we were unsuccessful in attempting to fit an affinity from shifted peaks, however plotting chemical shift perturbations (CSPs) >σ (where σ is the standard deviation CSP) for the ss-14mer titration, reveals a site distal to the RNA-binding site that exhibits significant CSPs (SI Fig. A2.2). The underlying cause of this perturbation is unclear but is unlikely to be a second binding site due to its dependence on RNA concentration.

Examining the ds-14mer titration reveals evidence of nonspecific binding. Repeating the same chemical shift perturbation analysis for the ds-14mer reveals a much wider pattern of chemical shift perturbations, with no clear structural localization. Analysis of these perturbed peaks reveals several that shift linearly with the RNA concentration, in contrast to the asymptotic behavior of other shifting peaks (Fig. A2.5d). This behavior is characteristic of nonspecific binding, and mapping residues exhibiting this behavior reveals a face of the NTD where nonspecific-binding residues cluster (Fig. A2.5c). This face corresponds to residues 80-84, 119-120 and 140-146, and is situated opposite the canonical binding face. Intriguingly, the ss-14mer titration shows no evidence of nonspecific interaction (Fig. A2.5c), indicating nonspecific interaction occurs only between the ds-14mer and the NTD.

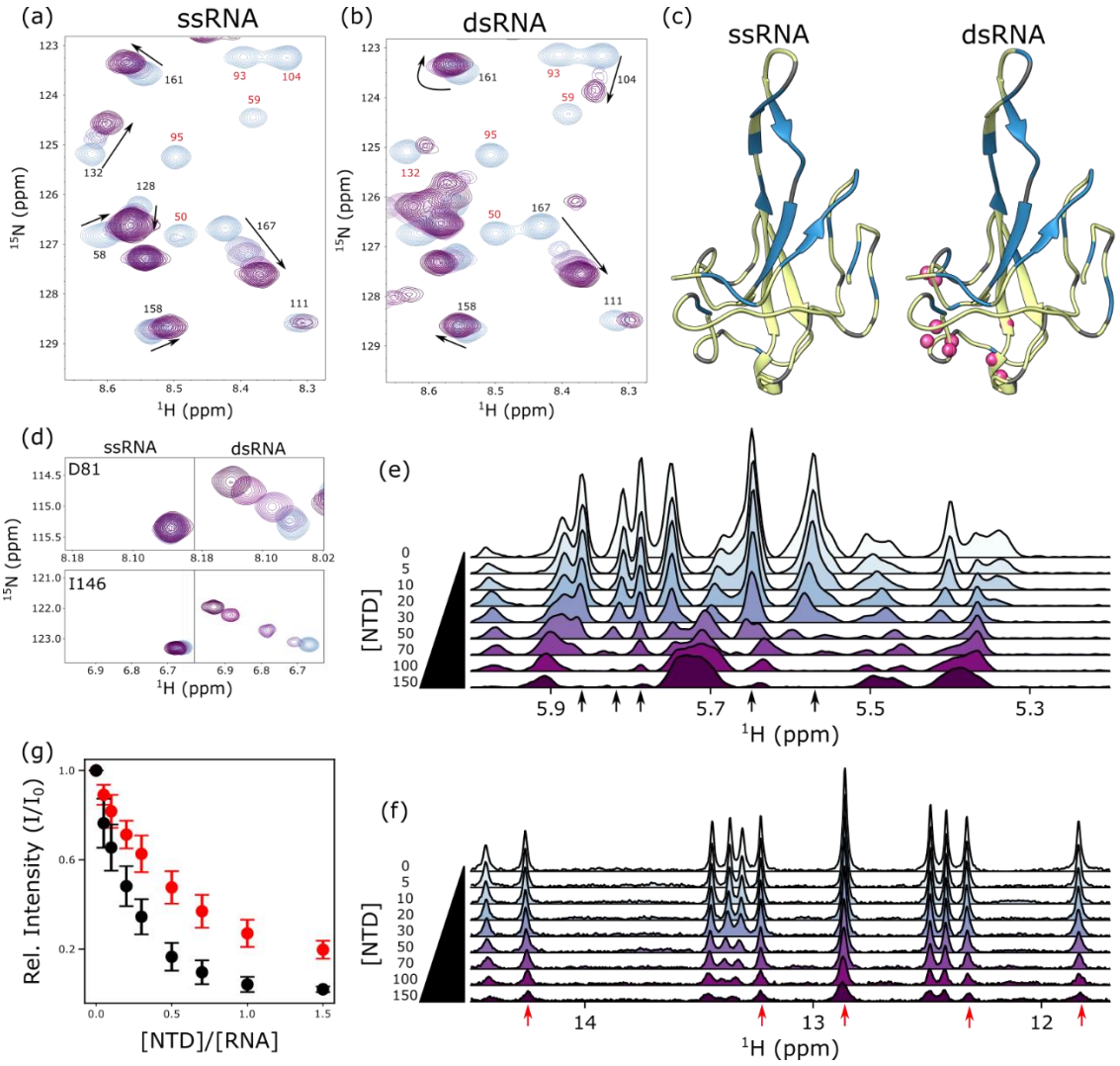


Figure A2.5: Interaction of CoV-N NTD with ss and dsRNA. (a,b) Sample window from ^{1}H-^{15}N HSQC spectra of the NTD (100 μM) at a titration series with the ss- and ds-14mer. 14-mer concentrations are 0,25,50,100,150, and 200 μM respectively, colored blue at 0 μM and purple at 200 μM. Spectra show a mix of intermediate and fast exchange – resulting in some peaks disappearing due to exchange broadening (labeled in red) and others shifting due to fast exchange (labeled in black). (c) Structure of NTD colored to show intermediate-exchanging residues in ss-14mer (left) and ds-14mer (right) titration. Assigned residues are colored yellow, and intermediate-exchanging blue. For the dsRNA titration, a pink orb is drawn to represent all residues exhibiting evidence of nonspecific binding. (d) Slice of 1H-15N HSQC spectra highlighting residues D81 and I146. Colors are the same as in panels a,b. slices show evidence of nonspecific binding to the ds-14mer. (e,f) ^{1}H NMR spectra of titration of NTD into ss-14mer (e) and ds-14mer (f). NTD concentration for each spectrum is listed to the right. Selected peaks labeled with an arrow (red for ds-14mer, black for ss-14mer) are analyzed in panel (g). (g) average peak intensities as a function of ratio of concentration of NTD to 14mer. Values shown are averages taken from selected 5 peaks in e,f. Uncertainty is standard deviation of the set.

To further examine the difference between ss-14mer and ds-14mer binding, we performed NMR measurements of titrations of the NTD into solutions of RNA (Fig. A2.5e,f). ^{1}H NMR spectra of the ds-14mer are ideally suited for this analysis, due to the presence of imino proton peaks in the range of 11-15 ppm (Fig. A2.5f), outside the chemical shift region for protein amides. The ^{1}H spectra of the ss-14mer also contains a series of peaks where no protein resonances appear, at 5-6 ppm (Fig. A2.5e). For both the ss-14mer and ds-14mer, peaks were significantly attenuated upon addition of NTD. We selected 5 peaks with minimum overlap from each spectrum (black and red arrows in Fig. A2.5e,f) for analysis and plotted peak intensities as a function of the ratio of NTD to RNA (Fig. A2.5g), and found significantly more peak attenuation for the titration of NTD into the ss-14mer. This attenuation points to tighter binding between the NTD and the ss-14mer, offering confirmation of our EMSA results above that the NTD binds the ss-14mer with a greater affinity than the ds-14mer.

The Y109A mutation interrupts specific binding between the NTD and RNA

To further explore NTD-RNA interaction, we began investigation of the Y109A mutant of the NTD, where a tyrosine thought to be involved in RNA binding through pi-stacking interactions is mutated to alanine. The Y109A mutation is reported in the literature to weaken binding between the NTD and RNA and thought to reduce the propensity for phase separation. We measured binding between the Y109A NTD and both the ss-14mer and ds-14mer. EMSA assays (Fig. A2.4) confirmed weaker binding between the Y109A NTD and both RNAs, in comparison to the wt NTD. In fact, at the measured concentration of 25 μM RNA, barely any binding between the ds-14mer and Y109A NTD could be seen, even at 5:1 NTD:RNA.

We performed NMR titrations of RNA into the Y109A NTD and found that the Y109A mutation induces a greater degree of nonspecific interactions between the protein and RNA. NMR spectra of the Y109A NTD (which we were able to assign using the wt NTD assignments and an HSQC-NOESY spectrum, see SI Fig. A2.3) reveals chemical shifts proximal to the Y109 position, suggesting that the global structure is not significantly perturbed by the mutation (SI Fig. A2.3). Titrations of RNA into samples of the Y109A NTD reveal a similar pattern of fast and intermediate exchange as seen in the wt domain (Fig. A2.6a,b). For both the ss-14mer and ds-14mer, a significant number of residues disappear at the binding groove (Fig. A2.6c), although notably, fewer residues disappear for the ds-

14mer titration, suggesting the exchange regime differs slightly for this titration. Surprisingly, we see similar evidence of nonspecific binding as was present in the wt NTD-ds-14mer titration, but now for both the ss-14mer and ds-14mer, although the nonspecific binding can be seen most clearly in the ds-14mer titration.

Titrations of Y109A NTD into RNA, as performed with the wt NTD, revealed weak binding between the mutant NTD and the ds-14mer. Titration of Y109A NTD into ds-14mer (Fig. 6x) induced less peak attenuation than what is seen for the wt NTD, confirming EMSA assays revealing this interaction to be weak. The ss-14mer titration is less conclusive – small RNA peaks remain even at saturating conditions, but within noise, this curve cannot be said to distinct from the wt NTD+ss-14mer titration.

To determine whether the tendency to bind nonspecifically has an impact on NTD function, we performed a series of liquid-droplet formation experiments on the NTD with the short RNAs at NMR conditions (SI Fig. A2.4) and found that several RNA-NTD combinations formed liquid droplets. We tested each of the four combinations of NTD and RNA at both 25 and 37 C, at a 2:1 ratio of RNA:NTD, and found that only the ss-14mer and wt NTD did not form liquid droplets. At 25 C, only the combination of ds-14mer and Y109A NTD formed visible droplets – a fact we first observed when performing titrations in the NMR. At 37 C, however, both ds-14mer samples, as well as the ss-14mer+Y109A NTD sample all separated readily into liquid droplets.

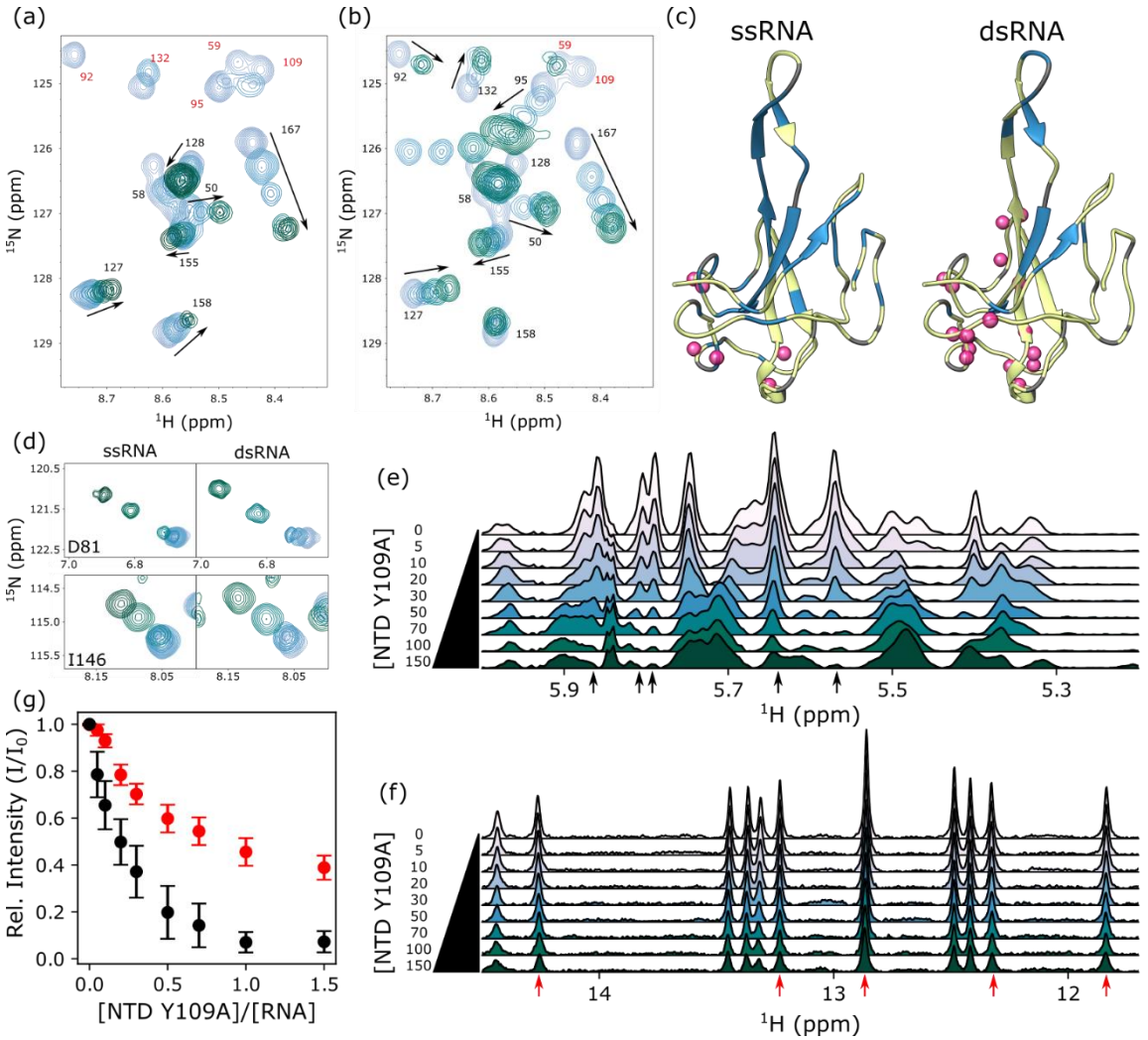


Figure A2.6: Interaction of Y109A NTD with ss and dsRNA. (a,b) Sample window from ^{1}H-^{15}N HSQC spectra of the Y109A NTD (100 μM) at a titration series with the ss- and ds-14mer. 14-mer concentrations are 0,25,50,100,150, and 200 μM respectively, colored blue at 0 μM and green at 200 μM. The final titration point for the ds-14mer could not be obtained, due to liquid droplet formation in the sample. Spectra show a mix of intermediate and fast exchange – resulting in some peaks disappearing due to exchange broadening (labeled in red) and others shifting due to fast exchange (labeled in black). (c) Structure of NTD colored to show intermediate-exchanging residues in ss-14mer (left) and ds-14mer (right) titration. Assigned residues are colored yellow, and intermediate-exchanging blue. A pink orb is drawn to represent all residues exhibiting evidence of nonspecific binding. (d) Slice of 1H-15N HSQC spectra highlighting residues D81 and I146. Colors are the same as in panels a,b. slices show evidence of nonspecific binding to the ds-14mer. (e,f) ^{1}H NMR spectra of titration of Y109A NTD into ss-14mer (e) and ds-14mer (f). Y109A NTD concentration for each spectrum is listed to the right. Selected peaks labeled with an arrow (red for ds-14mer, black for ss-14mer) are analyzed in panel (g). (g) average peak intensities as a function of ratio of concentration of NTD to 14mer. Values shown are averages taken from selected 5 peaks in e,f. Uncertainty is standard deviation of the set.

The CoV-N CTD preferentially binds dsRNA

We last turned our attention to interactions between the CTD and our 14-mer RNAs, and found that interactions between the CTD and RNA are weak. Attempts to assay this interaction by NMR were hampered both by peak broadening due to the size of the bound complex, meaning that peak shifts due to binding such as those collected on the NTD could not be collected. Peak intensity ratios collected at 0.66:1 RNA:CTD showed heterogeneity due to binding, offering the clearest picture of binding we were able to obtain (Fig. A2.7a,c). Comparing this intensity ratio between the ss-14mer and ds-14mer shows a pattern of high structural specificity for the ds-14mer, with much less specificity present for the ss-14mer interaction. Particularly the N-terminal region, and the residue 290-320 region on the alpha face of the CTD shows significantly more peak attenuation in the ds-14mer titration, with most peaks in these regions disappearing from the spectrum. The ss-14mer spectrum also dips in these regions, but to a much less significant degree. This suggests that the ds-14mer titration is highly specific to the alpha face of the CTD, where the ss-14mer shows no clear preference of binding site. no evidence of droplet formation was detected at either of these conditions.

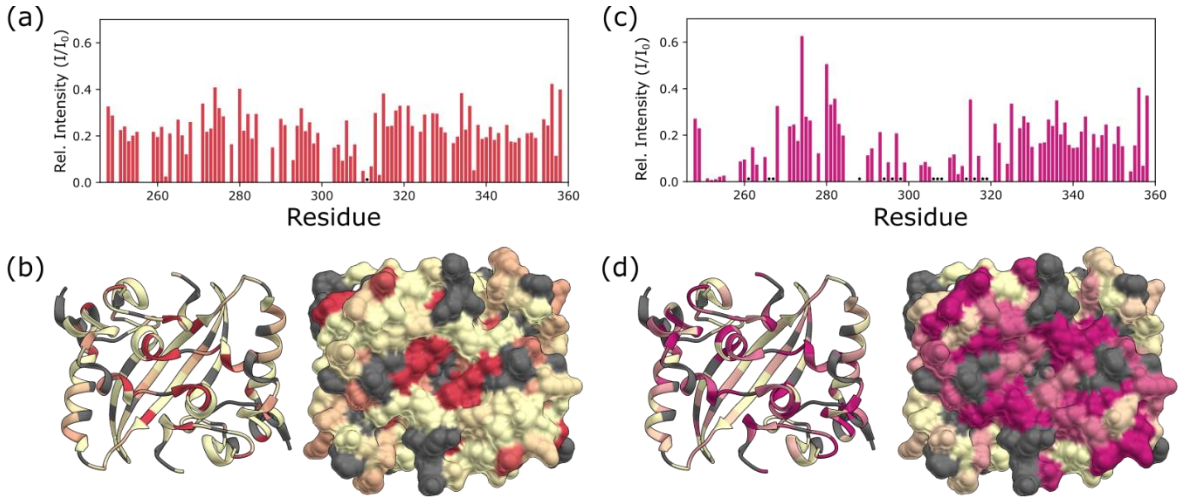


Figure A2.7: binding between the CTD and 14-mer RNA. (a,b) peak intensity ratios (relative to apo CTD) for the CTD bound to the ss-14mer (a) and the ds-14mer (b). Residues that disappear in the bound spectrum are marked with a star. (c,d) Ribbon diagram (left) and surface (right) of the CTD with colors mapped from the intensity ratios when bound to the ss-14mer (c) and ds-14mer (d). Darker red/pink corresponds to intensities closer to 0, and tan represents an intensity ratio >0.25.

Anisotropy reveals CTD-RNA binding is weak

To assess the affinity of binding between the 14mer RNAs and the CTD, we performed fluorescence anisotropy experiments between the 14mers and the CTD, as well as the

NTD. Titrations reveal binding on the scale of 20-30 μM between the NTD and the 14mers (22 μM for NTD+ss-14mer, 30 μM for NTD+ds-14mer), in agreement with published work on short RNAs. For the CTD, however, anisotropy changes were small, and we were unable to fit them to a model of binding, suggesting that binding between the CTD and RNA is weak in comparison to the NTD. A greater degree of change in anisotropy was observed for the ds-14mer titration, which, taken with our NMR results, suggests that the CTD may bind the ds-14mer slightly more tightly than the ss-14mer.

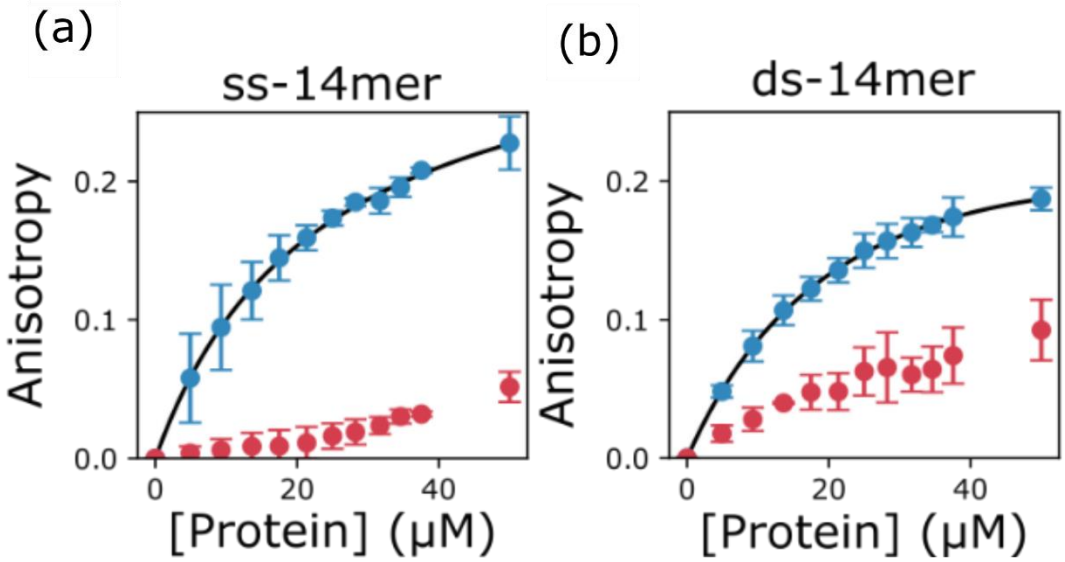


Fig. A2.8: Fluorescence anisotropy measurements of CoV-N domains with RNA. (a) titration of NTD (blue) and CTD (red) into fluorescently labeled ss-14mer RNA at 50 nM. (b) Titration of NTD (blue) and CTD (red) into fluorescently labeled ds-14mer RNA at 50 nM. Both plots are normalized to the anisotropy of the RNA in absence of protein.

Discussion

Binding affinities between CoV-N domains and 14-mer RNA

We have examined binding between both the N and C terminal domains of CoV-N with single and double stranded 14-mers of RNA to determine whether the domains have any preference for one RNA over the other. We found that the NTD binds ssRNA slightly more tightly than dsRNA, a finding confirmed both by EMSA assays, titrations of NTD into RNA, and fluorescence anisotropy measurements. It appears that the reverse is true for the CTD, which, based on NMR peak disappearances and anisotropy measurements favors dsRNA. These differences are relatively minor, and the more striking difference is between the NTD and the CTD. While the NTD binds the 14-mer RNA on the scale of the tens of

micromolar, the CTD-RNA binding was weaker than could be determined by any methods used, suggesting an affinity >100 μM. Mutation of Y109 to alanine appears to weaken binding of the NTD to RNA, based on EMSAs and NMR titration data, and confirming expectations set by prior studies.

Nonspecific binding and phase separation

Examining binding between the NTD and the 14-mer RNAs, we were surprised to find a varying degree of nonspecific binding. There is a surprising degree of nonspecific binding between the NTD and RNA. Indeed, the degree of nonspecific binding we see evidence of seems to be inversely correlated to the affinity of binding to RNA. For the tightest-binding complex, between the wt NTD and the ss-14mer, we see no evidence of nonspecific interaction, while for the ds-14mer, and for both RNAs with the Y109A NTD, we see evidence of nonspecific interaction. In fact, the Y109A-dsRNA titration shows the most evidence of nonspecific interaction, with many residues showing evidence of nonspecific binding. One possible explanation for this trend is that the NTD is only capable of one binding mode at a time – in other words, NTD that is tightly bound to RNA (as in the wt+ss-14mer case) is unable to bind nonspecifically, and vice versa. This may be due to specific NTD-RNA binding effectively closing off the nonspecific binding interface, which may be consistent with the differences in NTD-RNA complex seen by docking simulations. The Y109A mutation may also disrupt this specific structure, increasing the relative favorability of the nonspecific interaction.

Nonspecific binding is also closely correlated with the phase separation we see for the NTD. While dilute NTD does not phase separate with g1-1000, the protein will phase separate with RNA under more concentrated NMR conditions (100 μM NTD). Further, phase separation of the NTD correlates with nonspecific binding – the ssRNA+wtNTD complex does not form droplets, while the other three complexes do. Multivalent, weak interactions are the hallmark of liquid-liquid phase separation[222], so it seems natural to conclude that the separation of these samples is driven by the nonspecific binding visible by NMR. The CTD follows this same general trend, binding weakly to RNA – the domain binds only weakly to RNA, but phase separates with it most readily.

We can conclude that while the CTD promotes phase separation, the NTD appears tuned to promote separation only under certain conditions, dependent on the structure of the RNA being bound. As mentioned, the degree to which the nucleocapsid forms liquid-

liquid droplets is known to be dependent on the RNA being used. Here, we present one possible explanation for the variation seen – different structures of RNA bind CoV-N with different propensities for nonspecific binding, and those propensities in turn drive phase separation to a varying degree. Indeed, studies of phase separation with larger RNAs have also pointed towards dsRNA being a significant driver of phase separation, which meshes well with the structural information provided here.

Conclusions

While many questions remain, we have demonstrated here that the CoV-N domains do differentiate between different RNA structures. In particular, we have shown that the NTD binds RNA more tightly than the CTD, and that it binds to ssRNA more tightly than dsRNA. More critically, the tightness of binding is negatively correlated with nonspecific interactions visible by NMR, and those nonspecific interactions may drive phase separation. This provides a first glimpse into the structural determinants regulating the phase conditions of CoV-N, which is thought to be an essential point of regulation in viral function.

Materials and Methods

Expression and purification

FL-N, NTD, and CTD were expressed as described previously[211]. Briefly, FL-N was grown in 2xYT or MJ9 media to an OD_{600} of 0.6, then induced with 1 mM isopropylthio- beta-galactoside (IPTG). CTD was grown to an OD_{600} of 2 in terrific broth, then induced with 0.5 mM IPTG at 18 C overnight. For ^{15}N labeling, cells of either FL-N or CTD were instead pelleted at an OD of 0.7, and washed with ^{15}N-enriched MJ9 media, grown for an hour then induced with 0.5 mM IPTG at 30C for 4 hours (FL-N) or 18C overnight (CTD). The NTD was expressed in either ZYM-5052 autoinduction media (for unlabeled protein)or MD-5052 autoinduction media (for 15N labeling)

Proteins were purified under native conditions using TALON His-tag purification protocol (Clonetech). Cells were lysed in lysis buffer (50mM sodium phosphate buffer, 1M NaCl, 1mM NaN3, 5mM Imidazole, pH 8.0, 0.6mg/mL lysozyme with proteinase inhibitor) for 1hr at 4°C. Cells were further sonicated and centrifuged at 20000 RPM for 45 min to collect the supernatant. The supernatant was incubated with Cobolt resin in a gravity column. All washing steps in 3M NaCl to ensure removal of bound RNA contaminant. FL-

N and CTD were eluted with 4 CV elution buffer (50mM sodium phosphate buffer, 300mM NaCl, 1mM NaN3, 350mM Imidazole, pH 8.0). Purified NTD was eluted by on-column proteolysis with 30 nM untagged fast-acting protease bdSENP1 for 1 hr at 4 C. Proteins were further purified on a Superdex 75 gel filtration column (GE Health) in 50mM sodium phosphate, 150mM NaCl, pH 6.5. The purity of the recombinant proteins, assessed by SDS-polyacrylamide gels, was >95%. Protein concentrations were determined from absorbance at 280 nm using molar extinction coefficient values. Purified proteins were either stored at 4 C and used within one week or flash frozen to -80 C for long term storage.

NMR spectroscopy

NMR spectra of the NTD and CTD were collected on a Bruker 800 MHZ Avance II HD spectrometer equipped with a triple resonance cryogenic probe. All NMR samples contained 7.5% D_2O and a commercial protease inhibitor cocktail (Roche applied Science) Experiments on the NTD were carried out at 25 C in a buffer of 20 mM PO4 and 150 mM NaCl, at pH 6.5. Experiments on the CTD were carried out at 25 C in a buffer of 20 mM PO4 and 200 mM NaCl at pH 6 in a shaped NMR to minimize the impact of salt on the spectrum quality. Titrations with RNA were performed with concentrated RNA/NTD, to ensure less than 10% sample dilution across the titration.

Microscopy

Fluorescence microscopy images were taken on a Keyence BZ-X700/BZ-X710 microscope and a 384-well plate (Cellvis P384-1.5H-N); images were processed using BZ-x viewer and BZ-x analyzer software. For this experiment, cy3-labeled RNA was diluted into Invitrogen UltraPure DNase/RNase-Free Distilled Water to reach a final concentration of 50 nM, when added to protein sample for the 1-1000 RNA, respectively. Stocks of Atto 488 NHS ester (Sigma 41698) labeled FL-N, unlabeled CTD, and unlabeled NTD were prepared by diluting into 20 mM HEPES, 150 mM NaCl, 1 mM DTT, pH 7.5 droplet buffer to reach 4 µM and 10 µM final protein concentrations. Protein staining was accomplished by mixing Atto 488 NHS ester with protein and shaking at 4 C for 1hr according to the manufacture's protocol. Unbound dye was removed by PD-10 column (Cytiva). Unstained protein samples were prepared by combining 27 µL of protein stock with 3 µL cy3 labeled desired RNA for a total sample volume of 30 µL. For comparison, RNA alone samples were prepared with 27 µL of droplet buffer and 3 µL of the desired

RNA while protein alone samples were prepared with 27 µL of protein stock and 3 µL of Invitrogen UltraPure DNase/RNase-Free Distilled Water. The samples were then incubated at 37 C for at least one hour and subsequent imaging was taken.

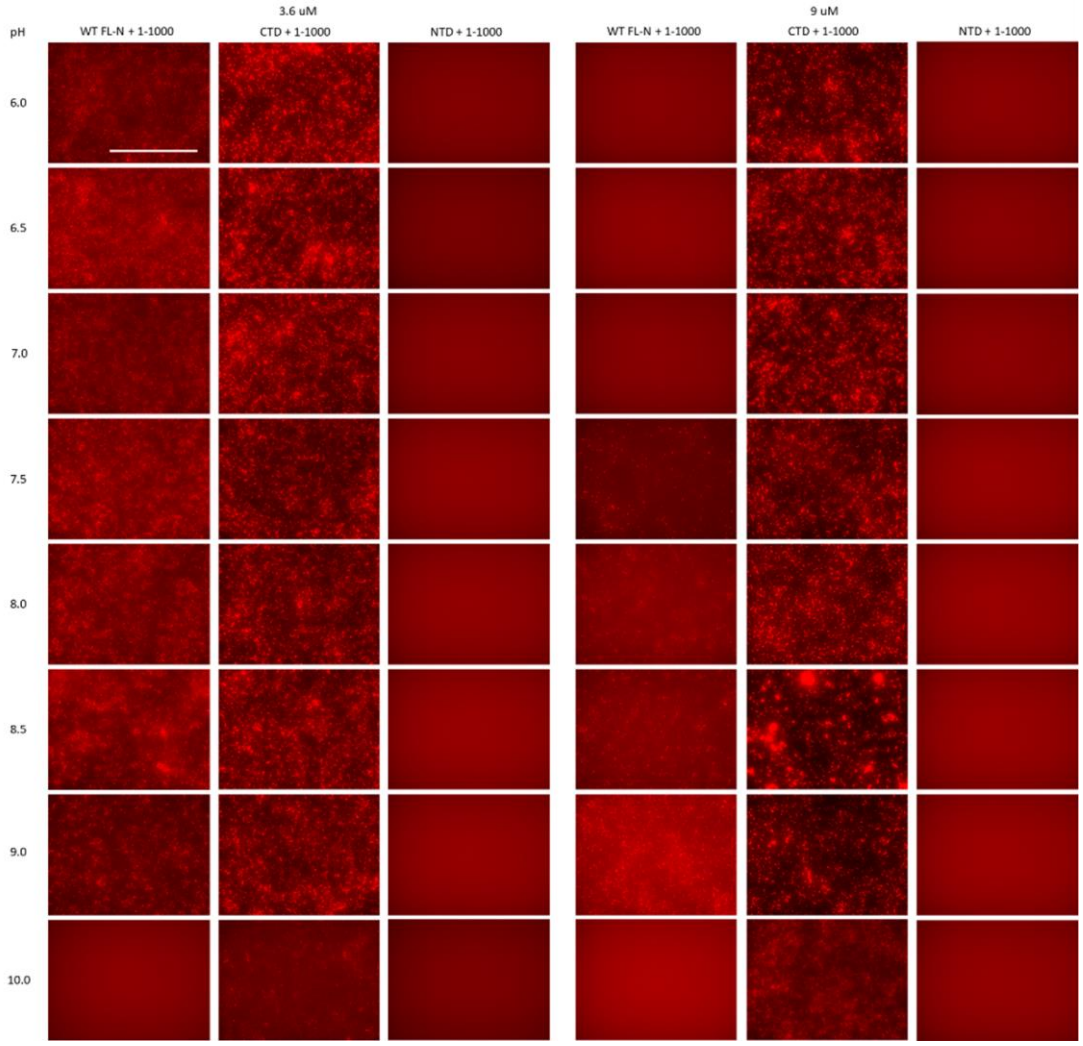
RNA Design Electrophoretic mobility shift assay (EMSA)

14-mer RNAs purified by HPLC were purchased from Genscript or Eurofins genomics. The ss-14mer sequence was GGCACAUGGACGUC, and the reverse complement used to generate the ds-14mer was GACGUCCAUGUGCC. The ds-14mer was generated by mixing equal parts ss-14mer and its reverse complement, heating the sample to 75 C, then annealing by allowing the sample to cool to room temperature. RNA sequences were designed to minimize the possibility of internal base pairing in the ss-14mer, and RNA samples were confirmed to be homogeneous by NMR and gel shift assays. The in vitro transcription and purification of the 1-1000 RNA generation followed protocols described previously [Forsythe et al 2020].

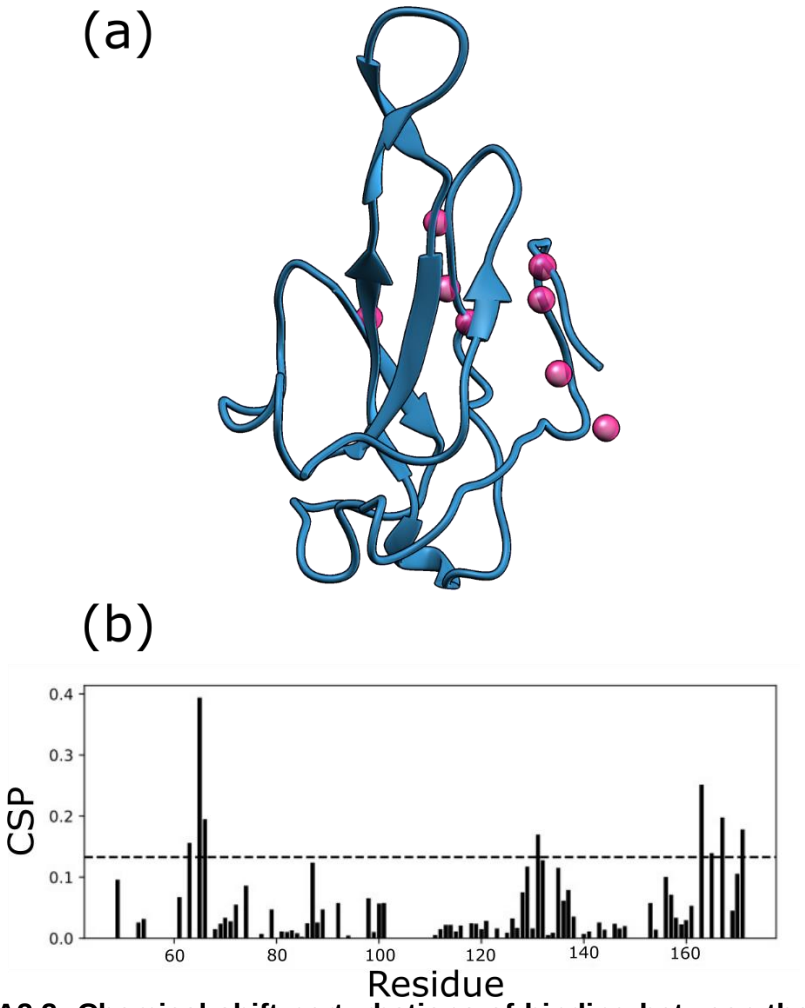
RNA was visualized by electrophoretic mobility shift assay in 1% agarose gel. RNA at 200 ng/ul was added to increasing concentrations of protein in the range of 0-125 µM, and incubated for 30 min at room temperature in a total reaction volume of 10µl. RNA bands were stained by the Midori Green Nucleic Acid staining solution (Bulldog Bio. Inc. Portsmouth, NH) and visualized by Bio-Rad Gel Doc Image system.

Fluorescence anisotropy

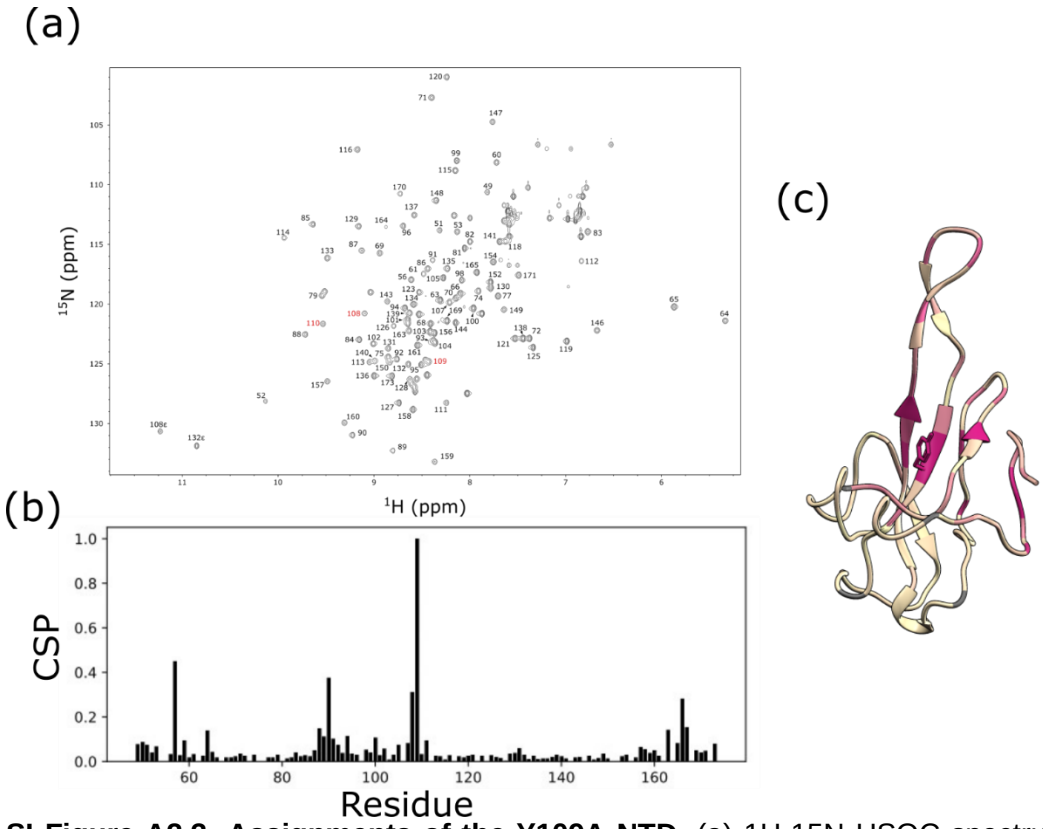
For fluorescence anisotropy, Both the ss-14mer and ds-14mer were labeledas described above in the microscopy. The measurement was performed on a FluoroMax-3 spectrophotometer (Horiba Scientific). The excitation wavelength was set to 492 nm and the emission wavelength to 516 nm. The concentration of RNA was set to 50 nM in 50 mM NaPO4 pH 6.5, 150 mM NaCl, 1 mM NaN3 buffer and the protein was titrated in the concentration range from 0 to 1.5 uM. The data were fitted in GraphPad Prism 9 using the OneSite-Total biding model.



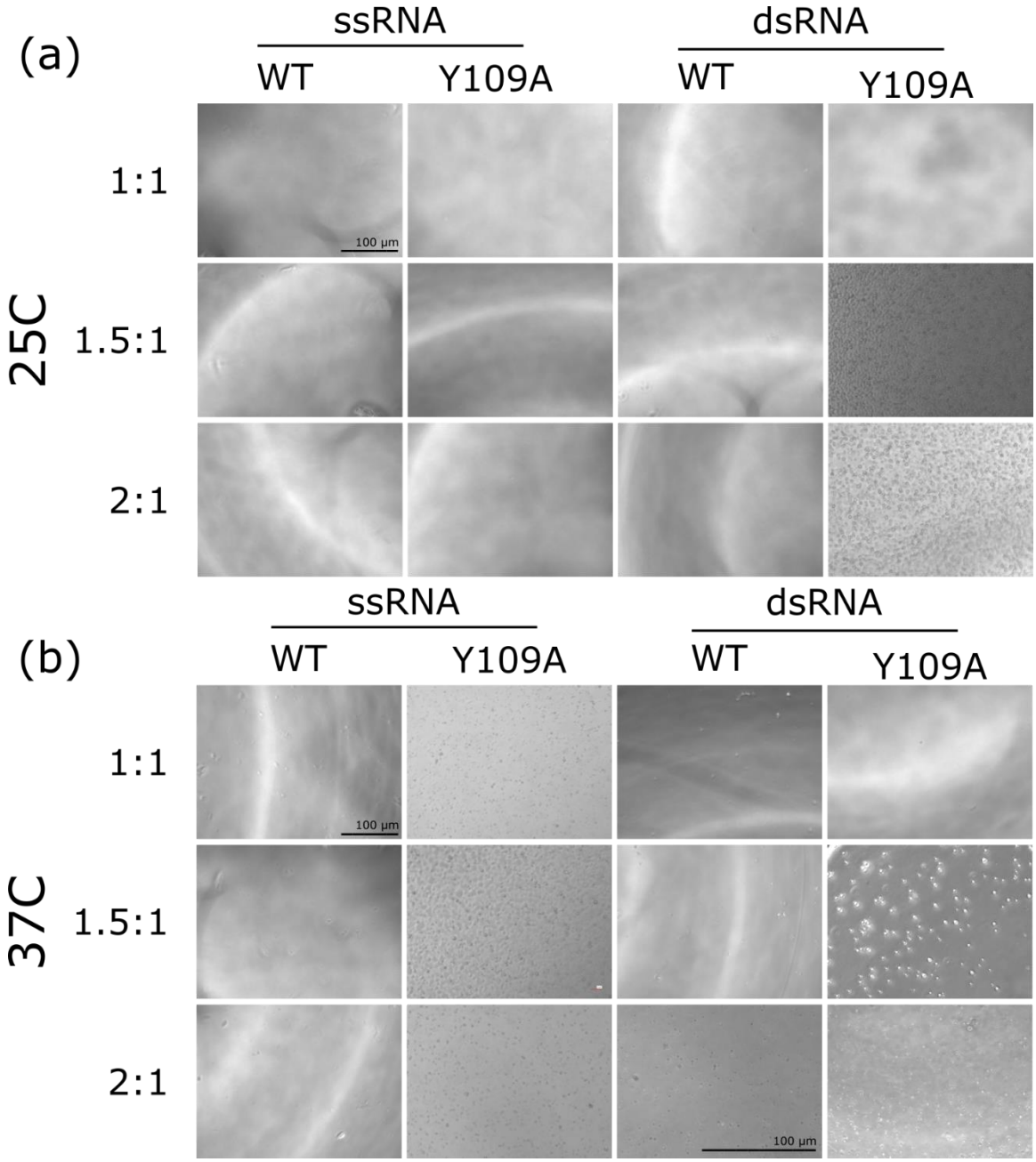
SI Figure A2.1: Phase diagram of CoV-N interaction with g1-1000 RNA . Bright-field and red fluorescent imaging investigating LLPS after 1.5 hours incubation at 37° C in droplet buffer (20 mM Tris, 150 mM NaCl, 1 mM DTT, pH 7.5) of 3.6 and 9 μM of FL-N, CTD, and NTD with 50 nM of 1-1000 gRNA at pH of 6.0 to 10.0 with 0.5 intervals in between. Scale bar is 200 μm. Imaging done using Keyence BZ-X700/BZ-X710 microscope with a 40X objective lens and a 384-well plate (Cellvis P384-1.5H-N).



SI Figure A2.2: Chemical shift perturbations of binding between the ss-14mer and the CoV-N NTD. (a) Structure of the NTD with residues where CSP > σ drawn as balls, concentrated on a region of the protein behind the main cup of the NTD structure. (b) Plot of chemical shift perturbations at each residue. The dotted line represents σ, the standard deviation of the chemical shift perturbations.



SI Figure A2.3: Assignments of the Y109A NTD. (a) 1H-15N HSQC spectrum of the Y109A NTD, with peaks labeled by assignment. Assigned peaks in red are near the site of mutation and significantly shifted from the WT spectrum. (b) Mutation-induced CSPs plotted for each residue. (c) Ribbon diagram of the NTD with atoms for Y109 drawn in. Diagram is colored by CSP, where beige represents low CSP and pink represents high. The CSPs are concentrated to an area around the site of mutation.



SI Figure A2.4: Phase separation of NTD with 14-mer RNA. (a,b) brighfield images of the NTD, both WT and Y109A (100 µM) collected at 25 C (a) and 37 C (b), mixed with RNA at ratios of 1:1,1.5:1, and 2:1 RNA:NTD. 100 µm Scale bar in top left image of (a) and (b) applies to all images except where an alternate bar is show.

Appendix 3

Design and characterization of a synthetic multivalent LC8-binding protein

Aidan B Estelle, Elisar Barbar

Excerpted from ‘Continuum dynamics and statistical correction of compositional heterogeneity in multivalent IDP oligomers resolved by single-particle EM’, published in in *Journal of Molecular Biology,* May 2022.

The following appendix is taken from a manuscript published in May 2022 on the characterization of LC8-IDP complexes through negative stain electron microscopy. I designed and performed initial characterization on a synthetic protein designed to tightly bind to LC8 in a multivalent fashion, described below. The protein binds tightly to LC8, forming a homogeneous 8:2 LC8:IDP complex ideal for testing electron microscopy methods.

Design, Expression and purification of the *syn*-4mer

we designed a novel LC8-binding peptide (termed *syn*-4mer) using a series of 4 repeats of the amino acid sequence RKAIDAATQTE, taken from the tight-binding LC8 motif of the protein CHICA (Uniprot Q9H4H8), which has a 0.4 μM affinity to LC8, making it one of the tightest-known LC8-binding motifs. The motif is spaced by uniform disordered linker sequences, totaling 3 linkers, and flanking GSYGS sequences were added to the N- and C-termini of the constructs to allow for quantification by absorbance at 280 nm. The final sequence is:
GSYGSRKAIDAATQTEPKETRKAIDAATQTEPKETRKAIDAATQTEPKETRKAIDAATQT
EGSYGS.

A gene sequence for the LC8-binding syn-4mer peptide was purchased as a block (integrated DNA technologies, Coralville, Iowa) and cloned into a pET24d expression vector with an N-terminal Hisx6 affinity tag and a tobacco etch virus protease cleavable site. The protein was expressed in ZYM-505256 auto-induction media at 37 °C for 24 hr. Cells were harvested, lysed by sonication and purified in denaturing buffers containing 6 M urea on TALON resin. The 4-mer was dialyzed into non-denaturing buffer (25 mM tris pH 7.5, 150 mM NaCl) and further purified by gel filtration on a Superdex 75 Hi-load column (GE Health), in the same buffer. Full length LC8 of Drosophila melanogaster was all expressed and purified as previously described. All proteins were stored at 4 °C and used within one week of purification.

SEC-MALS, Isothermal titration calorimetry , and Analytical ultracentrifugation

Size-exclusion chromatography (SEC) coupled to a multiangle light scattering (MALS) instrument was performed using an analytical SEC column of Superdex S200 resin (GE Healthcare) on an AKTA-FPLC (GE Healthcare), then routed through a DAWN multiple-angle light scattering and Optilab refractive index system (Wyatt Technology). The column

was equilibrated to a buffer of 25 mM tris (pH 7.5), 150 mM NaCl, and 5 mM BME, then injected with 100 µL of LC8/syn-4mer complex in the same buffer at an estimated 2 µM particle concentration (16 µM LC8 + 4 µM syn-4mer, assuming 2:8 binding stoichiometry). We estimated the molar mass using the ASTRA software package, with a Zimm scattering model.

Isothermal titration calorimetry was carried out at 25 °C using a VP-ITC microcalorimeter (Microcal) in a buffer of 25 mM tris (pH 7.5), 150 mM NaCl and 5 mM BME. A cell containing 9 µM syn-4mer was titrated with a solution of 300 µM LC8, across 32 injections of 8 µL. Peaks were integrated and fit to a single-site binding model in Origin 7.0.

Samples of the syn-4mer peptide in complex with LC8 were prepared for sedimentation velocity analytical ultracentrifugation (SV-AUC) by mixing excess (8:1) LC8 with syn-4mer, then purifying the complex by gel filtration on a Superdex 200 column in a buffer of 25 mM tris (pH 7.5), 150 mM NaCl, and 5 mM β-mercaptoethanol. The estimated concentration of the syn-4mer/LC8 complex applied to SV-AUC was at a 4:1 ratio of syn-4mer (13.8 µM) and LC8 (55 µM). The SV-AUC titration of LC8 into Nup159 was performed by mixing Nup159 (12.5 µM) and LC8 at LC8:Nup159 ratios of 0.5:1 to 8:1 in a buffer of 50 mM sodium phosphate (pH 7.5), 50 mM NaCl, 5 mM TCEP and 1 mM sodium azide. SV-AUC was performed on a Beckman Coulter Optima XL-A ultracentrifuge, equipped with optics for absorbance. Complexes were loaded into two-channel sectored centerpieces with a 12-mm path length and centrifuged at 42,000 rpm and 20 °C. We collected 300 scans at 280 nm with no interscan delay, and fit data to a c(S) distribution using SEDFIT.57 Buffer density was calculated using Sednterp.

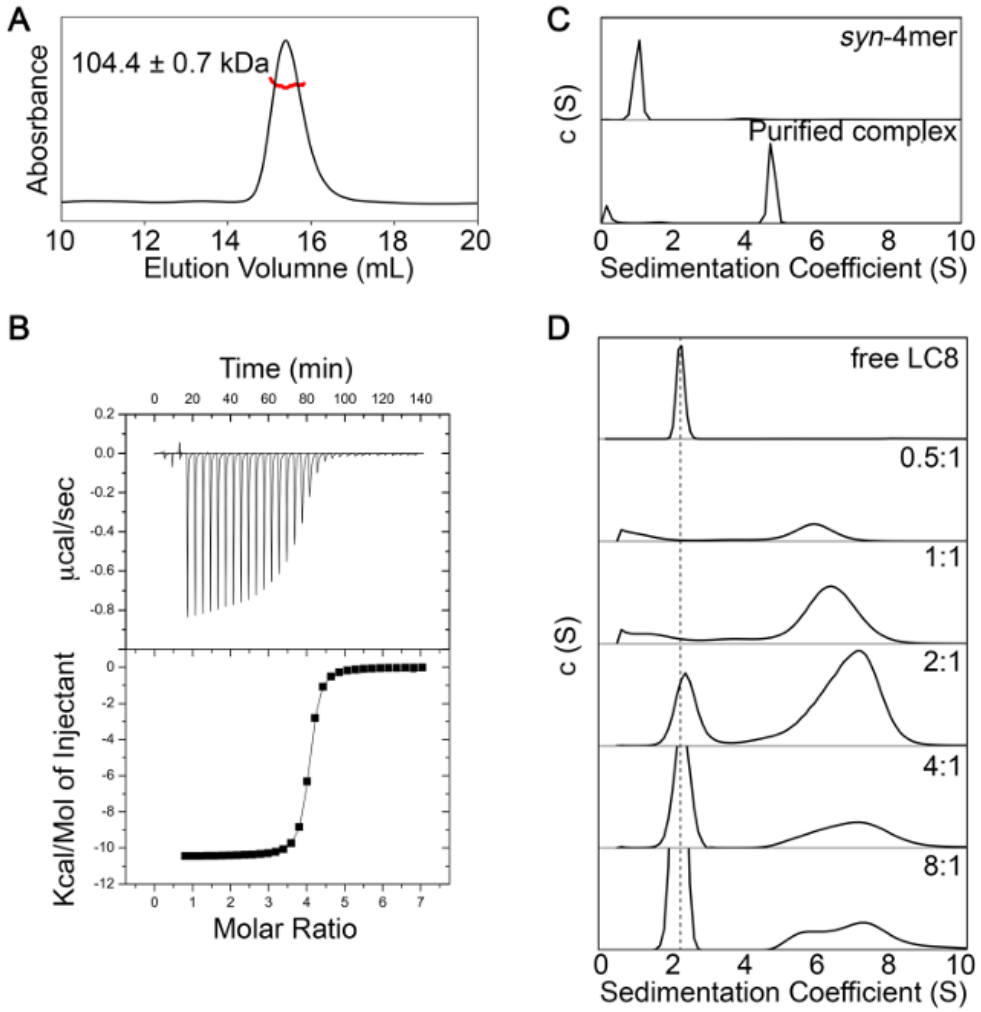


Figure A3.1: Sedimentation velocity analytical ultracentrifugation (AUC) of LC8 complexes. A) SEC-MALS of syn-4mer in complex with LC8. Purified complex eluted as a single peak, with a mass of 104.4±0.7 kDa, within uncertainty of the expected mass for a 2:8 complex, 105.2 kDa. B) Isotherm of binding between LC8 and the syn-4mer. The isotherm fits well to a simple binding model with Kd = 36±3 nM, DH = -10.47±0.04 kcal/mol, and N=3.98±0.01. Model fit is shown as a line. C) AUC data for the syn-4mer and size exclusion purified LC8/syn-4mer complex. A sharp peak at a sedimentation coefficient of 4.7 S indicates a tight and homogeneous complex. D) AUC data for LC8 and LC8/Nup159 complexes formed at increasing ratios of LC8. The dashed line is centered on the LC8 peak. The multiple peaks in the 6-8 S for the complex indicates heterogeneity of the complex and with an S value close to 8, it suggests a higher order assembly than a 5-mer and two Nup159 chains.