SPECTRAL APPROACHES FOR IDENTIFYING KINETIC FEATURES IN MOLECULAR DYNAMICS SIMULATIONS OF GLOBULAR PROTEINS

by

Andrej J. Savol

B.S., Applied Mathematics, B.A., Music, University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Andrej J. Savol

It was defended on
April 16, 2015
and approved by

Dr. Carl Kingsford
Associate Professor, Computational Biology Department, School of Computer Science, Carnegie Mellon University

Dr. David Koes
Assistant Professor, Dept. of Computational and Systems Biology, School of Medicine, University of Pittsburgh

Dr. Chris Stanley
Instrument Scientist, Oak Ridge National Laboratory

Dr. Dan Zuckerman
Associate Professor, Dept. of Computational and Systems Biology, School of Medicine, University of Pittsburgh

Dissertation Advisor: **Dr. Chakra Chennubhotla**
Assistant Professor, Dept. of Computational and Systems Biology, School of Medicine, University of Pittsburgh

SPECTRAL APPROACHES FOR IDENTIFYING KINETIC FEATURES IN MOLECULAR DYNAMICS SIMULATIONS OF GLOBULAR PROTEINS

Andrej J. Savol, B.S.

University of Pittsburgh, 2015

Proteins live in an environment of random thermal vibrations yet they convert this constant disorder into selective biological function. As data acquisition methods for resolving protein motions improve more of the randomness is also captured; there is thus a parallel need for analysis methods that filter out the disorder and clarify functionally-relevant protein behavior. Few behaviors are more relevant than folding in the first place, and this thesis opens by addressing which conformational states are kinetically relevant for promoting or inhibiting attainment of the folded native state. Our modeling approach discretizes simulation data into a network of nodes and edges representing, respectively, different protein conformations and observed conformational transitions. A perturbative strategy is then invoked to quantify the importance of each node, i.e. conformational substate, with regard to theoretical folding rates. On a test of 10 proteins this framework identifies unique 'kinetic traps' and 'facilitator substates' that sometimes evade detection with traditional RMSD-based analysis. We then apply spectral approaches and auto-regressive models to (1) address efficiency concerns for more general networks and (2) mimic protein flexibility with compact linear models.

Contents

List of Figures

List of Tables

Acknowledgements

It is a great luxury to dig deep into a scientific problem. But if a thinker's dream upon reflection, grad school is still a challenge on the ground. Sometimes it feels like constantly raking leaves and having no protection against a ruinous gust. At all times, however, there has been the support, encouragement, and good humor of some spectacularly kind individuals. They have made the entire experience rewarding and importantly also offered gentle reminders of my good luck to be in higher learning in the first place.

Classmates and collaborators that I'd like to thank include Virginia Burger, Keith Callenberg, Justin Hogg, Grace Huang, Jacob Joseph, Yevgeniya Monisova, David Mowrey, Rumi Naik, Arvind Ramanathan, John Sekar, Tim Travers, and Guy Zinman. At various times these wonderful scientists have been my teachers, housemates, officemates, co-authors, or commiserators; their intellect has often left me humbled but their friendship has always made me grateful.

A good share of my extant sanity is due to the Pittsburgh Compline Choir and the efforts of its fine musicians and instrumentalists. This ensemble remains my personal paragon of group cooperation and a reminder of the fruits of joint efforts. I especially appreciate the compositional guidance of Alastair Stout and the friendship of Elena Swann and Betty Rieley from the ranks. Other musical mentors from my undergraduate studies continued to enrich my experience in Pittsburgh, Roger Zahab, Don Franklin, and Natalie Phillips foremost among them.

The following friends were cheerleaders even when the buzzer was far in the future: Lee and Betsy Shaw, Gene and Marcia Gruver, Ron and Ping Chan, Mark Jordan, Eddie Skolnik, and Jan Wyse. This last person has been something of a guardian angel over my studies; her kindness has impacted three generations of the Savol family. Other friends were likewise so important their names will always come to mind when hearing the word 'Pittsburgh', especially Charles Chapman, John Vassallo, Brent Jackson, Tom Willoughby, Evan Mallory, Sachem Clark, Tom Menditto, Zach Koopmans, Harmandeep Singh, Josh Buchholz, Alex Ryan, Brian Bennett, Robert Tate, and James Keller.

It has been a great privilege to work with Chakra on many fascinating topics. His boundless enthusiasm, insight, and patience are etched in my mind now as the key attributes of an exemplary scientist and mentor. Thank you Chakra! Other unforgettable personalities at BST3 are Sandy Yates, Kelly Gentille, Jason Boles, and Gengkon Lum. The assistance provided by these coworkers made the difference between jumping ship and setting sail. Thom Gulish at CMU has been another constant whose support I greatly appreciate.

My thesis committee provided me with fantastic feedback and suggestions. Thank you David, Carl, Dan, and Chris.

My gratitude now turns west. Paul, Lana, and Linnea Norton offered sunny southern Californian encouragement from the very beginning of my studies. Thanks also go warmly to Colin McArthur for his uplifting letters and deep understanding. Lastly, I'm grateful to a man whose life provides a reservoir of guidance and inspiration, George Azadian.

I conclude with a brief retelling of a kitchen experiment. It is prior to 1990 and my father has just placed a hard-boiled egg on the rim of a glass flask after lowering scraps of burning newspaper into the container. The egg seals the air in the interior, but at first nothing happens.

Then suddenly and on its own, the egg jiggles and squeezes through the flask's neck, landing next to the charred Seattle P–I masthead. Jaro, MaryLee, and I could only laugh and wonder what properties of nature could explain what we just observed. If we kids didn't comprehend the equations and concepts quite right at first, what my parents Toni and Martin conveyed perfectly then and always was a fascination with the natural world and a willingness to 'poke it and see what happens'. Their encouragement and support has been fundamental. I wondered then about an egg sucking in its own gut, and I wonder much more now about my good fortune to have them as parents. Jaro and MaryLee have in the same way never wavered in their encouragement and special understanding.

Thank you all for being a part of my graduate studies.

Chapter 1

Introduction

If biologists permit statisticians to become arbiters of biologic questions, scientific disaster is inevitable.
Joseph Berkson (1899–1982)

Molecular dynamics (MD) simulations have dramatically improved our atomistic understanding of protein motions, energetics, and function. That these experiments are performed substantially from first principles illustrates that biology is yet another beneficiary of the 'aggressive territorial expansion' of mathematics[1]. Fundamental phenomena that power the cell, for example, can be replicated by the emergent behavior of atoms that are modeled entirely digitally in a computer. However, when performing protein simulations, and especially equilibrium simulations, often the behavior of interest does not emerge. Or perhaps it sort of emerges, but the collaborator isn't so sure. In silico experiments of protein dynamics are increasingly producing datasets too large to unambiguously assess visually and often too varied or subtle to interpret with basic tools. Biological functions inhabit a broad spectrum of timescales (Fig. 1.1), and atomistic modeling has only somewhat recently entered the temporal territory of protein folding. It hardly made sense to discuss the statistics of protein folding within a simulation when folding only occurred once. But, at least for polypeptides of modest length and with well-structured native states, this discussion is open and computational biology is tasked with developing analysis tools that add to it.

A first high-level question might be why we need computational tools for data that were themselves generated computationally. The modeling community could respond that relating MD data to bulk experimental data from NMR, SANS, FRET, etc. requires the filtering out of a lot of unneeded detail in the simulations. Naturally, this problem goes way beyond filtering: very few experimental observables have a direct one-to-one correspondence with MD's primary output, the instantaneous position of all atoms in a molecule/solvent system. In short, the classical Newtonian equations propelling MD are still hiding what an experimenter, leaning over the same fluctuating protein, would be able to observe with current techniques. This line of inquiry argues that when the experiment and observables agree, non-observable details in the

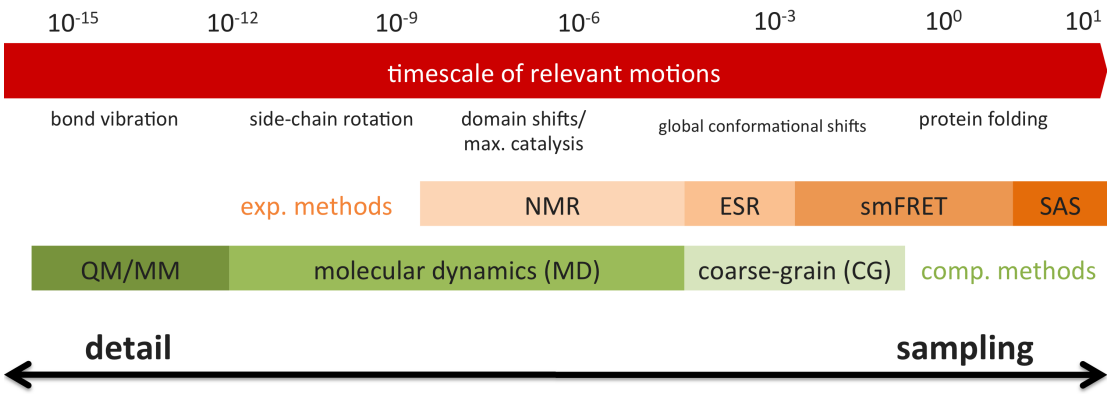


Figure 1.1: A comparison of timescales and methods for protein flexibility. Approximate timescales, in seconds, are shown above some representative biophysical behaviors. The temporal resolution of some common experimental methods (nuclear magnetic resonance (NMR), electron spin resonance (ESR), single molecule Forster resonance energy transfer (smFRET), and small angle scattering (SAS)) is diagrammed in orange. Computational methods (quantum mechanics/molecular mechanics (QM/MM), etc.) are compared in green.

simulation can be considered. A second and optimistic justification for dissecting synthetic data might be that the collaboration can work in the other direction, that MD simulations properly analyzed can suggest experiments difficult to justify otherwise. Here we take this standpoint–that many relevant and testable features of protein motions are likewise hiding in MD data. We argue that kinetic models of conformational change can tease them out in Chapter 2, Kinetic frustration in protein folding, and build compact models of protein motions using statistical invariants in Chapter 4, Auto-regressive models of protein motions. The intervening Chapter 3, Faster f-score approximations, will discuss the efficiency of the introduced kinetic calculations, so its results are motivated by the intricacies of molecular motions but its applications are hopefully not limited to them.

If it's not yet apparent, time in various senses is a common theme here. The thermodynamics of folding pathways dictate which routes are faster or slower and also which structural features (e.g. helicity, native contacts, etc.) confer acceleration or inhibition of the folding process. We use the term *kinetic features* to denote changes in folding rates that result from the presence or absence of a specific conformational substate. We identify kinetic features in a normative and quantitative way in Chapter 2. Our method is based on a simple idea: what happens to theoretical folding times when individual conformational substates are energetically forbidden? This question is somewhat difficult to investigate experimentally because it involves altering a protein's

energetics in a certain state-space region while keeping the rest of the territory unchanged (mutational and solvent perturbation studies sometimes have this aim). Imagine adding a localized, infinite mesa to the free-energy landscape (FEL) in Fig. 1.2. This is exactly the question graph theory can address, however, since each FEL subregion can be examined and perturbed independently. Importantly, this perturbation approach is only valid if individual FEL subregions, or *conformational substates*, are identified with reasonable clustering protocols. Provided our clustering does indeed group together those individual conformers that interconvert freely, the kinetic influence of individual conformational substates on overall folding times can be quantified.

Computing these kinetic features, called *kinetic traps* or *facilitators* in the context of protein folding, turns out to be expensive, so we try to save some of that time in Chapter 3. Once a system is schematized as a network of interconverting states, small alterations can often be made without recomputing the global topology and its behavior. To do so, we map the original intact network to a spectral representation such that update equations from matrix perturbation theory are feasible. A drawback is that our approach still requires the intact network's full spectrum, so efficiency tests, where we attain a one- to five-fold speedup, are performed on medium-sized social, biological, and infrastructure networks where initial, dense eigendecomposition is tractable.

Returning to protein flexibility, Chapter 4 investigates the time spans over which protein motions are predictable. Consider again the folding trajectory depicted in white in Fig. 1.2. Are some protein motions duplicated each time the peptide conformation approaches the low-energy native state? We show that there are indeed compact representations of protein motions that exploit dynamic regularities in observed simulations. We are then able to synthesize novel protein motions that reproduce the statistics of all-atom MD simulations but are encoded compactly in linear expressions, that is, without the higher-order energy terms of electrostatic or van der Waals interactions.

■ 1.1 Publications list

Most of the work in this thesis has been published. Results from non-first author publications are not included with the exception of Fig. 4.2, generated by the author and included in publication [1] below.

9. **Savol A** and Chennubhotla CS. **Approximating frustration scores in complex networks via perturbed Laplacian spectra**. *Scientific Reports*, 2015. [In Review]

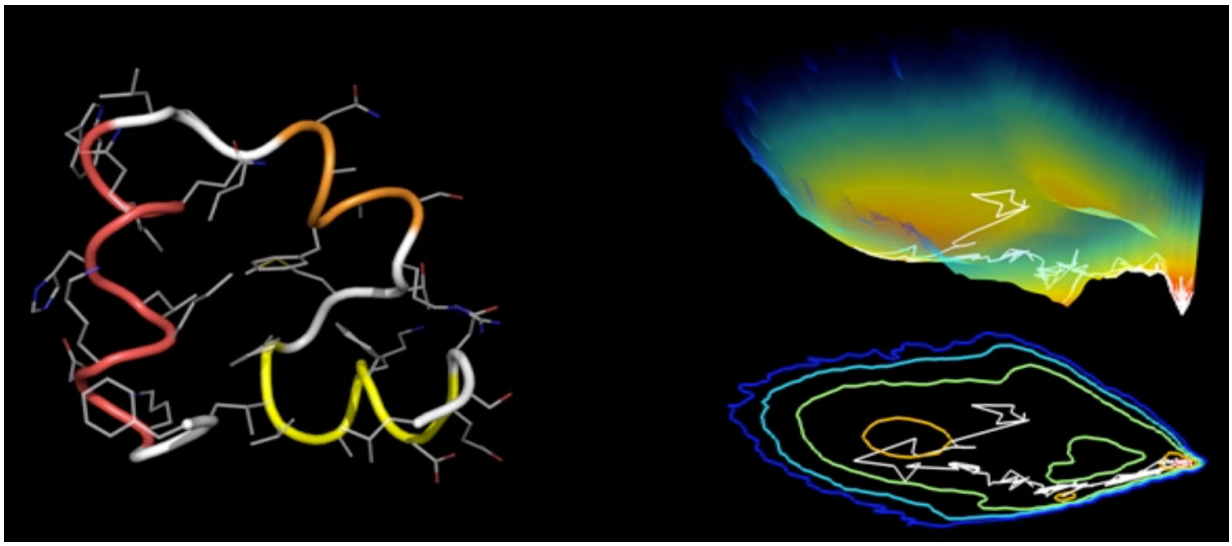


Figure 1.2: A free-energy landscape and an individual folding event. A model fast-folding protein, villin headpiece (VHP), is shown in its folded conformation (left). Backbone colors indicate important structural components, helices 1 (yellow), 2 (orange), and 3 (red). To arrive in that conformation the polypeptide accesses a sequence of conformations shown in white on the free-energy landscape visualization (right). Landscape is composed of the dihedral principal components of the VHP simulation discussed in Chap. 2. Energetically favorable regions of the state-space are depicted in red.

8. **Savol A** and Chennubhotla CS. **Quantifying the Sources of Kinetic Frustration in Folding Simulations of Small Proteins**. *Journal of Chemical Theory and Computation*, 10(8):2964–2974, August 2014.

7. Ramanathan A, **Savol A**, Burger V, Chennubhotla CS, and Agarwal PK. **Protein Conformational Populations and Functionally Relevant Substates**. *Accounts of chemical research*, 47(1):149–156, August 2013.

6. Taylor DP, Wells JZ, **Savol A**, Chennubhotla CS, and Wells A. **Modeling boundary conditions for balanced proliferation in metastatic latency.** *Clinical cancer research*, 19(5):1063–1070, March 2013.

5. Ramanathan A, **Savol A**, Agarwal PK, and Chennubhotla CS. **Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: application to enzyme adenylate kinase.** *Proteins*, 80(11):2536–2551, November 2012.

4. Burger VM, Ramanathan A, **Savol A**, Stanley CB, Agarwal PK, and Chennubhotla CS. **Quasi-anharmonic analysis reveals intermediate States in the nuclear co-activator receptor binding domain ensemble.** *Pacific Symposium on Biocomputing*, pages 70–81, 2012.

3. Ramanathan A, **Savol A**, Burger V, Quinn S, Agarwal PK, and Chennubhotla CS. **Statistical Inference for Big Data Problems in Molecular Biophysics**. *Proceedings of Big Data*, 2012.

2. **Savol A**, Burger VM, Agarwal PK, Ramanathan A, and Chennubhotla CS. **QAARM: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin.** *Bioinformatics*, 27(13):i52–60, July 2011.

1. Ramanathan A, **Savol A**, Langmead CJ, Agarwal PK, and Chennubhotla CS. **Discovering conformational sub-states relevant to protein function.** *PLoS ONE*, 6(1):e15827, 2011.

Chapter 2

Kinetic frustration in protein folding

We are bound withal to Time, and the amounts of it spent getting from one end of a journey to another.

Thomas Pynchon
Mason & Dixon

Experiments and atomistic simulations of polypeptides have revealed structural intermediates that promote or inhibit conformational transitions to the native state during folding. We invoke a concept of *kinetic frustration* to quantify the prevalence and impact of these behaviors on folding rates within a large set of atomistic simulation data for ten fast-folding proteins, where each protein's conformational space is represented as a Markov state model of conformational transitions. Our graph theoretic approach addresses what conformational features correlate with folding inhibition and therefore permits comparison among features within a single protein network and also more generally between proteins. Nonnative contacts and nonnative secondary structure formation can be quantitatively implicated in inhibiting folding for several of the tested peptides.

2.1 Introduction

Theoretical and computational modeling has provided many insights into the remarkable ability of proteins to rapidly fold from unstructured coils into their native, functional conformations.[2,3] Especially for small structured proteins, entire folding processes can be investigated via atomistic, equilibrium molecular dynamics (MD) simulations, where the ability to sample multiple folding events (with μs simulations) with a transferable force field is an important milestone in algorithm development and hardware parallelization[4–8]. When multiple folding events are observed, the underlying kinetics and conformational features that promote structural transitions and the eventual attainment of the native state can be statistically compared. Such studies reveal important characteristics of the underlying free energy landscape (FEL), the high-dimensional surface of hills and valleys that govern the likelihood of structural transitions and the occupancy probabilities of energetically coherent states, called *conformational substates*[9,10]. For structured

proteins the FEL has been conceptualized as a funnel with a low-energy *native ensemble* at its global minimum (where near-native intermediates are kinetic neighbors) and a *nonnative ensemble* comprised of freely interconverting conformers at some further reaction distance[11,12]. While the majority of protein functions are accomplished via the native ensemble, quantifying the structural and kinetic characteristics of the nonnative ensemble can aid calibration of coarse-grained polypeptide models[13–16] and improve our understanding of folding initiation pathways[17,18], protein misfolding[19], protein aggregation[20,21], and synergistic folding (i.e., folding in tandem with a binding partner)[22].

Although nonnative ensembles recapitulate several properties of idealized random-coil models,[23,24] they have also been shown to deviate from polymeric predictions in important ways. Substantial secondary structure can accrue in the nonnative ensemble[7,25,26], and these nucleation locations have been implicated as consistent waypoints in folding pathways[27,28]. Lindorff-Larson et al.[6] likewise showed that for transition pathways specifically, secondary structure accumulates before native contacts are formed, a temporal preference that is inconsistent with an idealized nonnative ensemble. From the kinetic perspective, another surprise is that the nonnative ensemble can be modeled as a hub-like transition map, where interchange between unfolded peptide geometries is mediated preferentially via the native (hub) ensemble instead of by direct routes[29] (but see Ref. 30). The minimally frustrated model of protein folding harmonizes some of these observations by recognizing that folding is energetically downhill and will thus avoid the enthalpic frustration of nonnative structure formation[31–34]. Analogously, we can ask whether the nonnative ensemble is minimally frustrated in a kinetic sense. Does folding proceed sequentially[35] from unfolded to folded substates or are there off-pathway kinetic inhibitors populating the FEL? For some but not all of our studied peptides, we will see that substantial kinetic traps that retard folding can be identified.

Computational studies have invoked cartesian, angular, topological, subspace-projection, or other structural descriptors to identify nonnative conformational states[27,36–38], but their potential impact on folding rates requires additional analysis. To quantify the kinetic contribution to folding of specific conformational substates, we present here a methodology that (1) permits comparisons of kinetic inhibition across multiple folding events and between multiple proteins and (2) can query any proposed structural parameters that may impact folding kinetics. Instead of only specifying

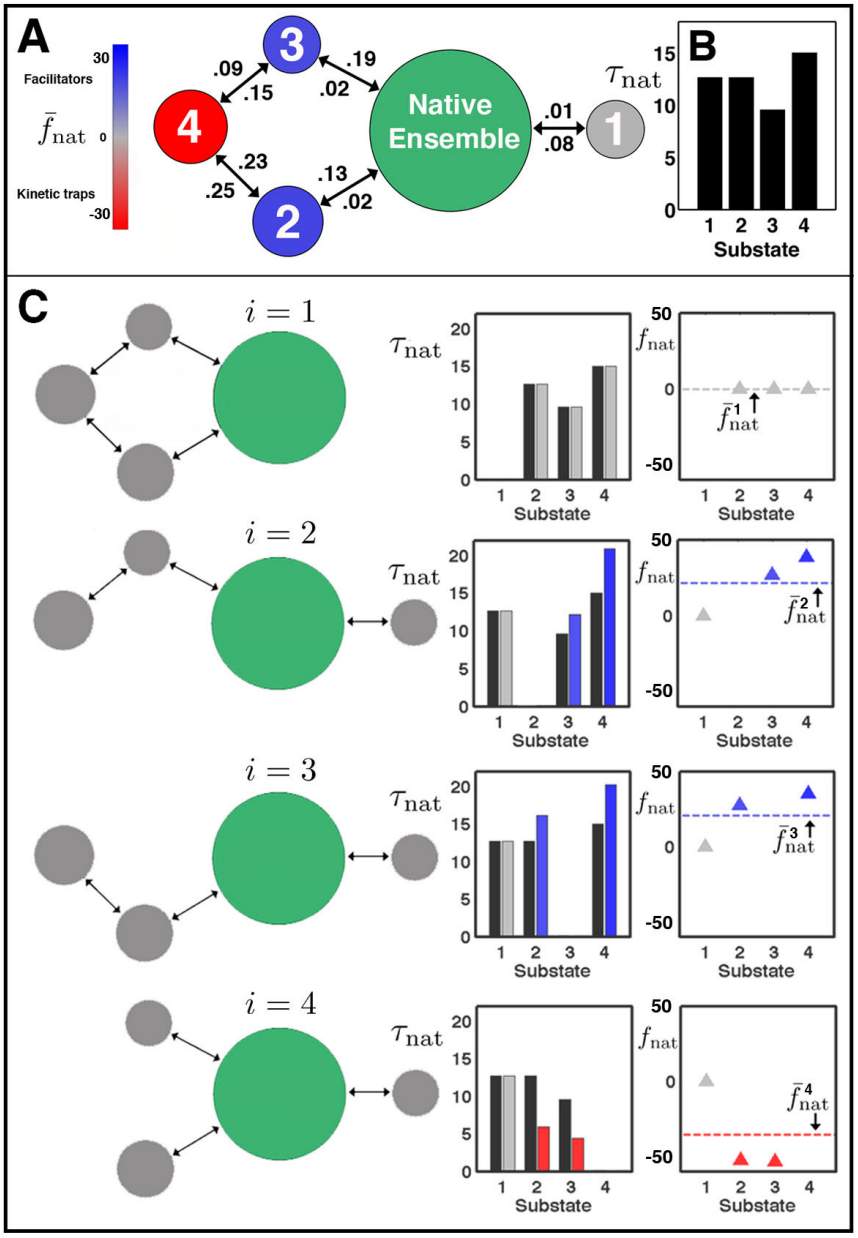


Figure 2.1: Computing frustration scores, $\bar{f}_{\rm nat}$, for a model transition network. (A) Each conformational substate in the nonnative ensemble ($1 \ldots k_{\rm nn}$) is colored according to frustration scores, $\bar{f}_{\rm nat}$; substate diameters indicate the stationary probability. The native ensemble is represented as a single green substate. Transition probabilities are shown along observed transitions where values above the transition path always denote left→right transitions and values beneath the arrow refer to moving left←right. (B) Computed MFPTs, $\tau_{\rm nat}$ values, for each nonnative substate to reach the native ensemble. (C) Procedure for computing $\bar{f}_{\rm nat}$. Each nonnative substate is removed from the transition matrix (states $i = 1 \ldots 4$, top to bottom) and transit times for all remaining $k_{\rm nn} - 1$ substates are compared with unperturbed values (left panels: black bars, unperturbed; red, gray, or blue bars, perturbed). Relative changes in transit times (wedges, right panels) are averaged over remaining substates to yield $\bar{f}_{\rm nat}$ (dashed lines). These frustration scores are then depicted by the color scale on the original intact network (A). Substates 1 and 2 have identical MFPTs, whereas $\bar{f}_{\rm nat}$ values indicate substate 2 is a facilitator and increases folding rates from all other substates by an average of 25%, while substate 1 is kinetically neutral. Substate 4 is a kinetic trap, slowing all transit times by 30% on average in the unperturbed network.

the existence of kinetic traps, hubs, or preferential pathways in MD trajectories, we quantify the overall kinetic burden, or *kinetic frustration*, that structural deformations (secondary structure, tertiary structure, standard RMSD-to-native, or others) effect in protein simulations. In the rest of this introduction we overview our model's assumptions, justification for a topological definition of kinetic frustration, and primary results.

We invoke a kinetic modeling of MD simulation data where a simulation trajectory and its conformers are represented as (1) sets of clusters, or conformational substates, whose kinetically-indistinguishable members share conformational features, and (2) transitions, which capture the observed jumps between substates. Such a network of substates (nodes) and edges, when constructed with an appropriate lag-time between sampled trajectory snapshots and clustering criteria, satisfies the properties of a Markov State Model (MSM)[39,40]. These models are guided by the motivation to equate conformational transitions with probability flow, enabling multi-step transition pathways to be associated with a probability and expected duration even if the path itself was never observed within contiguous trajectory frames. By representing a protein's FEL as an evolving finite markov chain, MSMs permit computation of the stationary distribution, the unique set of substate probabilities that is stable over time. It is then possible to calculate the expected time for any substate to transition to the assigned native ensemble, i.e. the mean first passage time (MFPT) or transit time[41]. MFPT values express temporal expectations for random walks along the weighted edges of the conformational network[42,43]. They are robust[44,45] and can be compared to diffusional models of folding[46,47] and nanosecond laser T-jump experiments[48]. Whereas MFPTs necessarily are a function of two specified endpoints, our concern is only with those transition paths that terminate at the native ensemble, a convention implicit throughout this study and indicated by the subscript of MFPT values, τ_{nat}. Can these values tell us which substates are responsible for accelerating or hindering folding? Not directly, but that exact information is revealed when substates are theoretically removed from the transition network and the change in τ_{nat} values among the rest of the nonnative ensemble compared. Kinetic frustration, quantified in frustration scores, $\bar{f}_{\text{nat}}$, captures these changes and quantifies the degree to which a particular conformational substate state inhibits or facilitates transitions to the native state. The terms *inhibit* and *facilitate* summarize a substate's topological neighborhood with respect to the native ensemble: a substate that facilitates folding is highly connected to native or native-like

substates, whereas a folding inhibitor promotes transitions to non-native regions of the FEL.

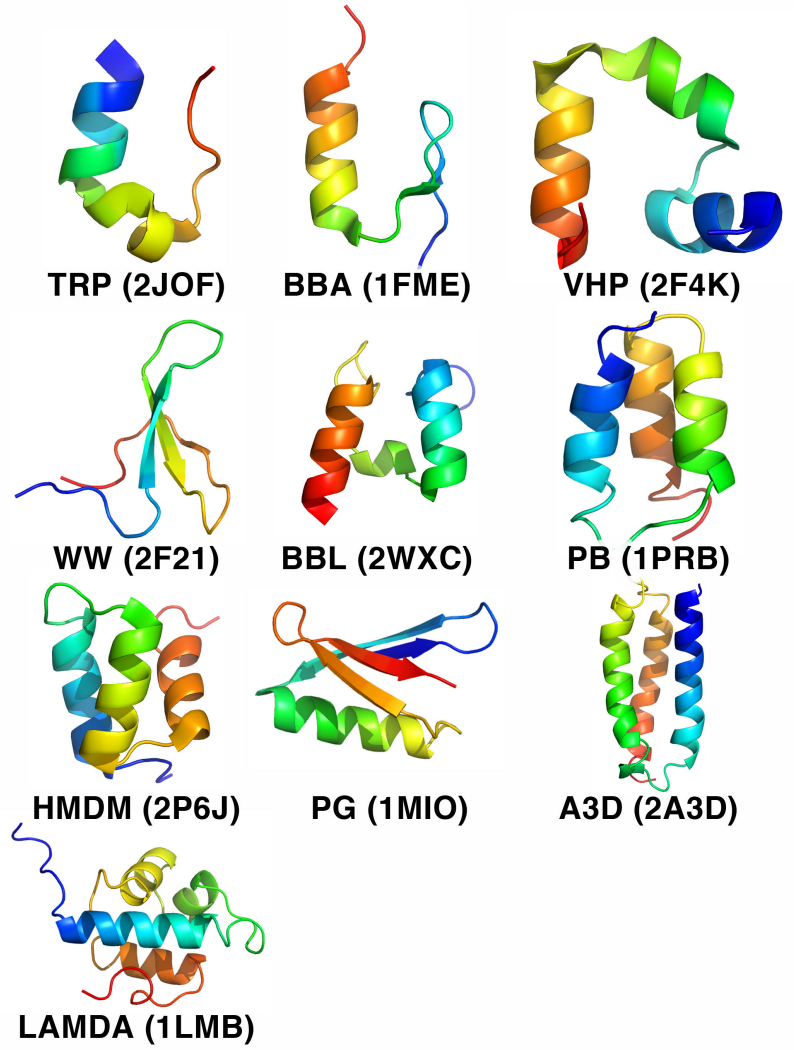


Figure 2.2: Native conformations for 10 tested proteins. Peptide and simulation details are provided in Table 2.1. Proteins are labeled by nickname; PDB accession numbers are shown in parentheses.

MFPTs and frustration scores, $\bar{f}_{\mathrm{nat}}$, are therefore related but, importantly, distill different information. τ_{nat} values reflect expected transit times given the network structure, whereas frustration scores quantify the impact on the network given the substate of interest. Nodes which have equal transit times need not share frustration scores, for example (see Figs. 2.1 and 2.6B). Additionally, MFPT values have been shown to be less informative in large graphs because transition paths quickly 'forget' from where they started and as a result only the in-degree of the target node, and not topological structure, is important[49]. Frustration scores, our

preferred centrality measure, are individually computed for all nodes in the nonnative ensemble by observing changes in τ_{nat} values when each is removed from the network model (Fig. 2.1)[50], a process akin to the eigenvalue estimation problem in matrix perturbation theory[51]. Observations obtained by altering the transition network in this way provide a quantitative framework for understanding its unperturbed behavior. Specifically, each frustration score $\bar{f}^i_{\text{nat}}$ is interpretable as the mean percentage change in all transit times from all possible paths as a result of node i. Substates with $\bar{f}_{\text{nat}} > 0$ are labeled facilitators since folding rates would decrease (i.e., τ_{nat} increase) in their absence; states with $\bar{f}_{\text{nat}} < 0$ are inhibitors, or kinetic traps, in that folding rates would increase (τ_{nat} decrease) if they were to be removed from the conformational landscape. We thus invoke the concept of kinetic frustration because MFPT values alone cannot elucidate these causal relationships.

Within our simulation dataset of ten fast-folding proteins, substantial kinetic traps were observed for four proteins (TRP, BBL, PB, and HMDM) whereas kinetic inhibition was chiefly absent in the nonnative ensembles of WW, PG, and A3D. The largest frustration scores (most extreme facilitators) were observed in the WW and PG simulations, and about half of the ten globular proteins (Table 2.1) presented a noticeable collection of facilitators that appeared to be topologically distinct from the native ensemble itself. As shown in Fig. 2.6, kinetic traps were unequally distributed throughout the nonnative ensembles. The transition networks, or transition maps, did display unique topological features, and we were able to ask to what relative degree secondary structure, tertiary structure, and nonnativeness (standard RMSD-to-native) were associated with positive or negative kinetic frustration. We chose these structural parameters because of their broad interpretability and popularity for monitoring folding progress[52,53], but emphasize that the approach is compatible with any geometric feature that can be computed for all trajectory frames.

Folding is a conformationally heterogenous process[54], but the recognized prevalence of preferred folding routes[55] and transition pathways[56] highlights the need for tools linking specific nonnative substates to folding kinetics. Quantifying these relationships is a legitimate aim in its own right, but our findings relate to the larger problem of predicting the kinetic impact of direct perturbations to protein systems. Mutations, small molecule ligands, or solvent conditions that modulate the populations of conformational substates can influence folding rates or folding

routes[57–59], and quantifying any such changes therefore has applications to pathway inhibition, aggregation-based diseases, and protein engineering[60–63].

Table 2.1: Simulation data. A summary of the proteins and simulations studied, adapted from Lindorff-Larson et al.[6] Data columns indicate sequence length (N_{res}), total aggregated simulation duration (t_{total}), number of conformational substates (and any substates excised during transition matrix MLE) (k), Protein Data Bank accession code (residue indices), simulation temperature, number of folding events (N_f), number of unfolding events (N_f), and the native-ensemble RMSD cutoff (r_{nc}). All figures and tables order proteins according to increasing sequence length. Native conformations are shown in Fig. 2.2.

Protein name	N_{res}	t_{total} (μs)	k	PDB	Temp (K)	N_f	N_u	r_{nc} (Å)
trp-cage (TRP)	20	208	417	2JOF	290	12	12	1.5
BBA	28	325	$999-1$	1FME	325	14	14	2.6
villin headpiece (VHP)	35	125	251	2F4K	360	34	34	1.3
WW-domain (WW)	35	1137	2274	2F21 (4–39)	360	12	11	1.4
BBL	47	429	860	2WXC	298	12	11	4.8
protein B (PB)	47	104	208	1PRB (7–53)	340	19	19	3.4
homeodomain (HMDM)	52	327	654	2P6J	360	27	28	3.7
protein G (PG)	56	1155	$2310-2$	1MI0 (10–65)	350	12	13	1.2
alpha 3D (A3D)	73	707	1414	2A3D	370	12	12	2.9
lambda repressor (LAMDA)	80	643	$1293-1$	1LMB (6–85)	350	10	12	1.9

2.2 Methods

2.2.1 MD Simulations

We applied our analysis to ten proteins within a large simulation dataset generated by D. E. Shaw Research as reported in Ref. 6 and analyzed further elsewhere[30,34,64,65]. Aggregate simulations of the ten proteins comprise 5.1ms of total sampling where each protein undergoes at least ten folding and unfolding events (Table 2.1). The proteins selected for simulation by the original authors were chosen such that a variety of local and global protein structure would be represented across the dataset. Additional considerations for the original authors were that the folding rates observed in the simulation could be compared to those from experiments; this consideration also required that the simulations use different temperatures for different sequences, from a minimum of 290 K (TRP) to a maximum of 370 K (A3D). We chose this dataset because the simulations (1) were performed with a consistent forcefield[66] and (2) allowed for sampling of multiple folding and unfolding events. While other large simulation datasets were available, this collection of trajectories has already been shown to sample reasonably close to the experimentally observed native states and has also been subjected to kinetic analyses that we can compare with our

results[30]. Our analysis here does not require that folding and unfolding events be precisely defined, but our results on the FEL more generally are unlikely to be statistically useful if they are derived from simulations that only capture a low number of folding events. Figure 2.4 illustrates how these folding events are commonly defined. Each trajectory conformer (trajectory frame) is assigned a native contact value Q:

$$Q(t) = \frac{\sum_{i=1}^{N_{res}} \sum_{j=1}^{N_i} \frac{1}{1+e^{10(d_{ij}(t)-(d_{ij}^0+1))}}}{\sum_{i=1}^{N_{res}} N_i} \tag{2.1}$$

where i and j are residue indices, N_{res} is the total number of residues, N_i is the number of residues in contact with residue i in the native substate (based on a cutoff threshold), d_{ij} is the instantaneous distance between residues i and j, and d_{ij}^0 is the average interresidue distance between i and j in the native substate[6]. Contacts for VHP are shown in Fig. 2.3. $Q(t)$ values range from 0 (completely denatured) to 1 (completely folded). The resultant $Q(t)$ trace is analyzed for full traversals of folded/unfolded thresholds $Q = 0.9$ and $Q = 0.1$, respectively. Such a trace for VHP is shown in Fig. 2.4, where folding and unfolding events are indicated by superimposed blue and red panels. Duration variation of these events for the entire 125 μs simulation is shown in Fig. 2.5. Although specific folding events are not invoked in subsequent analysis, we emphasize that our transition networks result from multiple folding occurrences of various duration.

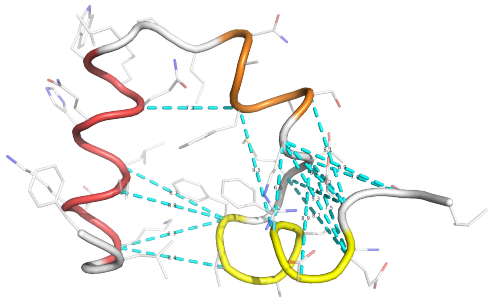


Figure 2.3: **N**ative contacts for VHP. Dashed cyan lines indicate distances d_{ij} that are compared with native distances d_{ij}^0 in Eqn. 2.1. Yellow, orange, and red backbone regions denote helices 1, 2, and 3, respectively (see also Fig. 2.10, upper right panel).

Clustering and all subsequent analysis was performed on the C_α coordinates. Snapshots were recorded every 200 ps. Multiple simulations for the same protein, if present, were concatenated.

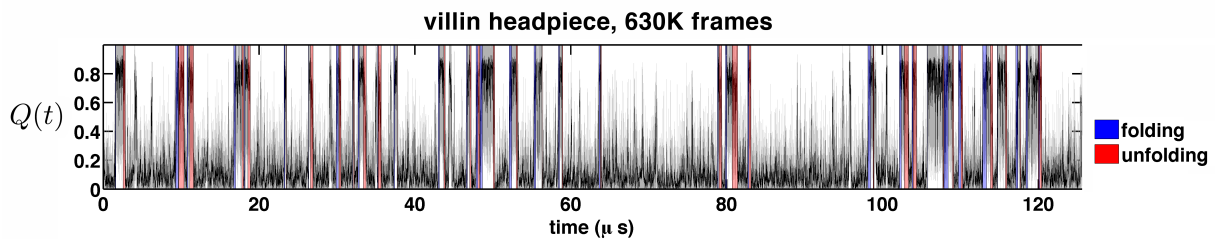


Figure 2.4: $Q(t)$, a reaction coordinate for monitoring folding progress Defined in Eqn. 2.1, $Q(t)$ is shown for a 125μs molecular dynamics simulation of villin headpiece. Folding and unfolding events are highlighted by blue and red panels, respectively. Histograms of these durations are given in Fig. 2.5. Contact maps for all 10 proteins are shown in Fig. 2.9.

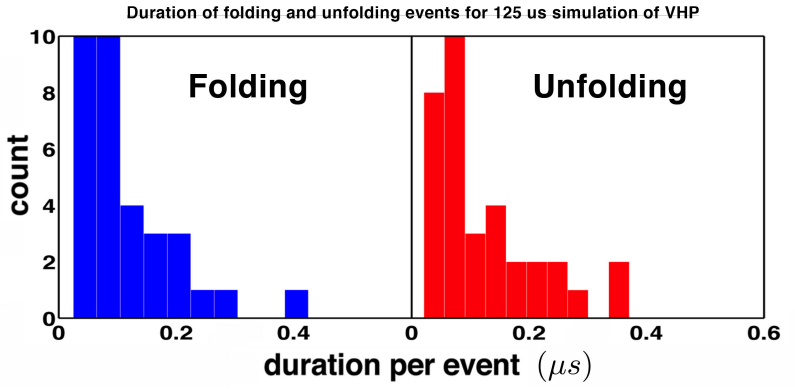


Figure 2.5: Duration of folding and unfolding events for VHP.

■ 2.2.2 Determining conformational substates

To identify conformational substates for each protein, we performed hierarchical clustering with MSMBuilder2[67]. Trajectories were first subsampled to obtain snapshots every 50ns based on implied timescales (Fig. 2.7), then clustered into substates using root mean squared distance (RMSD) and Ward's algorithm[69]. Remaining trajectory frames were discarded. The number of substates, k (see Table 2.1), is a heuristic user parameter that was selected to be approximately equal to (simulation frames)/10[70]. This parameter has been shown to have little dependence on peptide length, N_{res}[71]. The k values chosen here correspond closely to those in Ref. 30. The transition probability matrix P was then approximated using the MSMBuilder2 maximum likelihood estimation (MLE) routine, and substates not included in the estimated matrix (i.e., those separate from the primary connected component due to being at the termini of trajectories and isolated during the MLE routine) were excised from subsequent analysis. Connected singletons

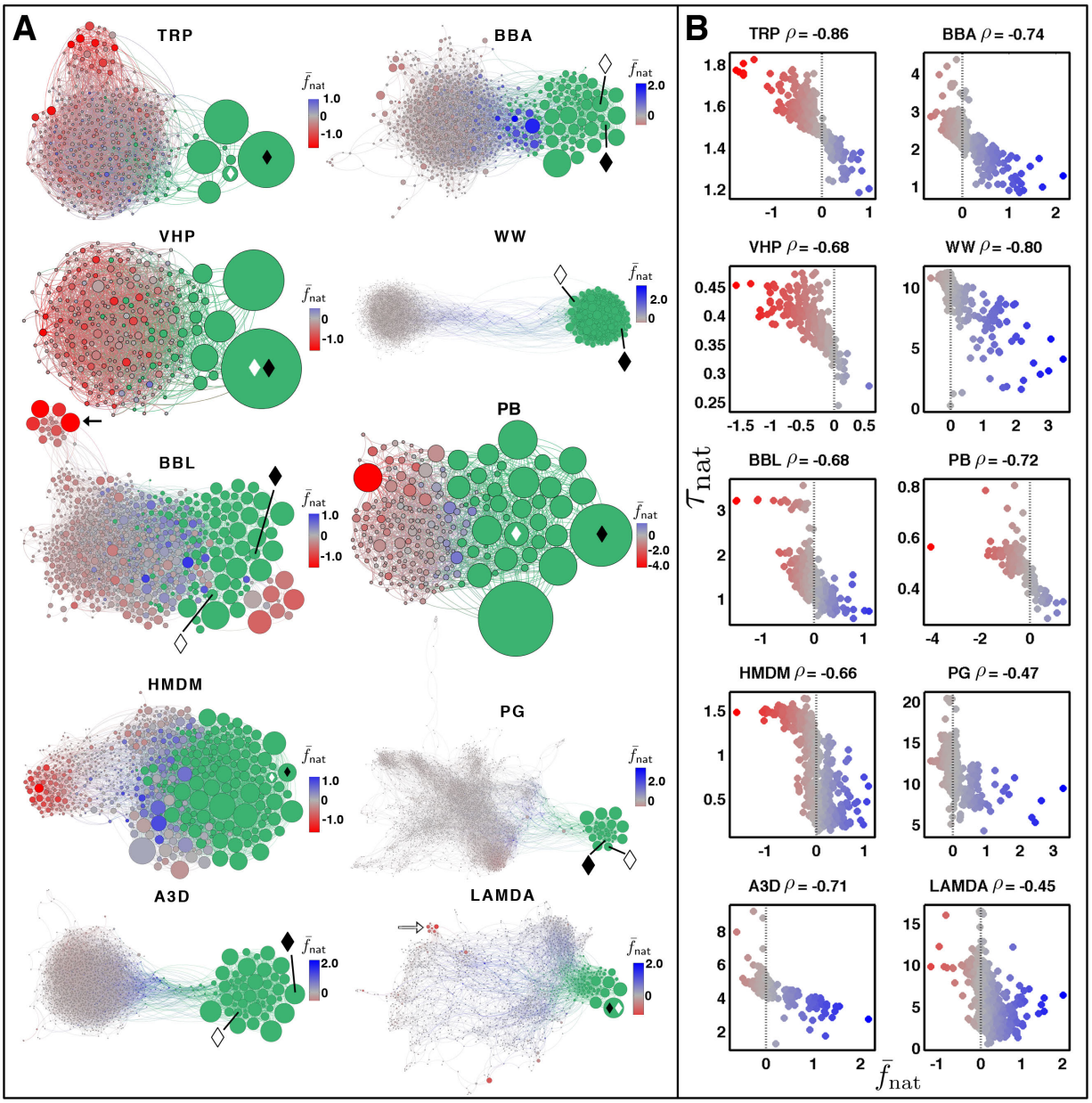


Figure 2.6: Network representations of substates and transitions. (A) Nodes indicate conformational substates determined with RMSD clustering[67]; node diameters are proportional to substate probability. The native ensemble, green, was determined by modularity optimization[68]. White diamonds indicate the substate containing the conformation closest to the experimental structure. Black diamonds indicate the substate containing the native conformation (see Methods). Frustration scores, $\bar{f}_{\mathsf{nat}}$, are denoted by the color spectrum, centered at $\bar{f}_{\mathsf{nat}} = 0$. Positive scores, blue, indicate substates that facilitate transition to the native ensemble; negative scores, red, indicate kinetic traps. (B) A comparison of frustration scores, $\bar{f}_{\mathrm{nat}}$, and transit times, τ_{nat} (μs), for nonnative substates in (A). Color values correspond to frustration scores as in (A). See Table 2.2 and Section 2.3.1 for additional details.

(substates with a single member) were retained, however, and constituted 0% of total conformers for BBA, BBL, PB, and TRP and 0.8% – 10% for A3D, HMDM, LAMDA, PG, VHP, and WW. Distributions of cluster sizes (number of member conformers) and widths (defined as mean pairwise RMSD of any two substate members), are given in the Appendix (Figs. A.2 and A.8). As summary, the conformational substates are determined through the geometric property of structural homogeneity (i.e. RMSD), whereas matrix P is determined by the kinetic property of transitions observed within the (downsampled) trajectory.

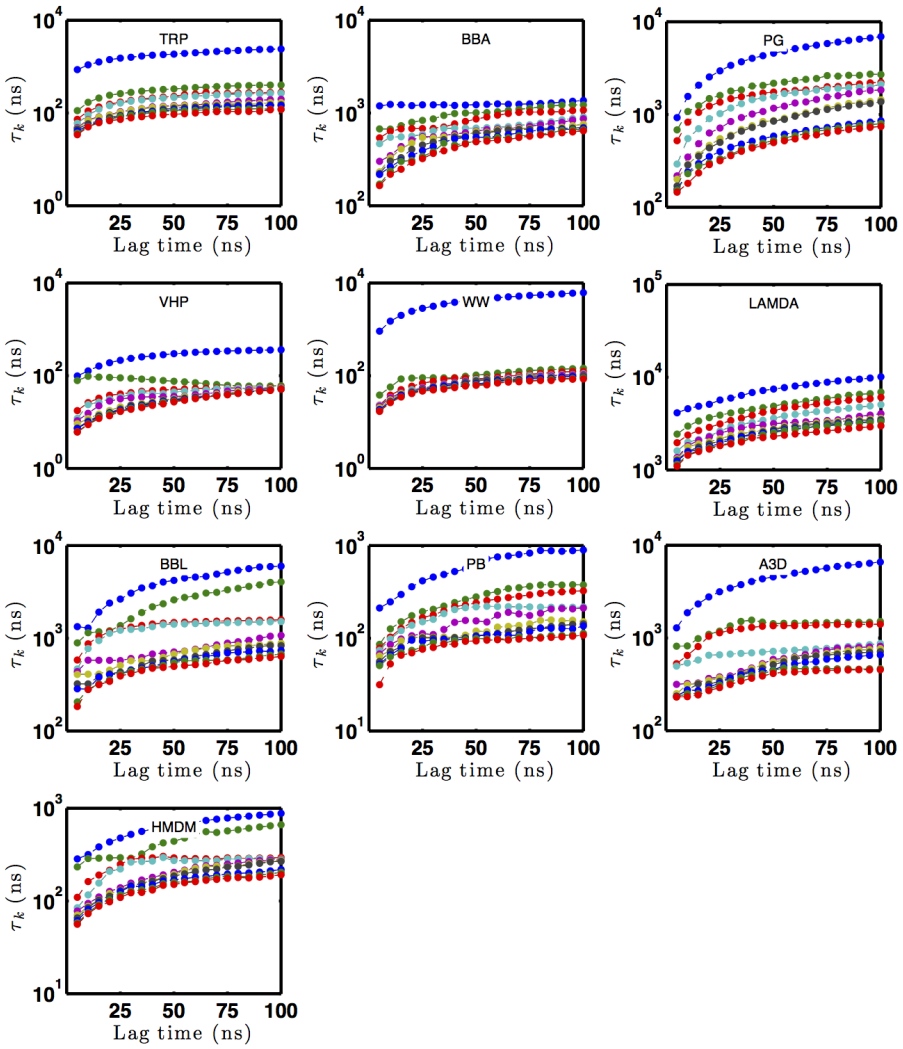


Figure 2.7: Implied Timescales. Ten slowest implied timescales (τ_k) as a function of lag time computed from MSMs constructed for each aggregate simulation.

■ 2.2.3 Defining the native ensemble

Our network folding model requires a demarcated native state to function as a kinetic endpoint, i.e., a theoretical absorbing state[41] where folding is defined as complete. Selecting the largest conformational substate[6], the substate closest to the PDB-deposited coordinates, or a hard RMSD threshold is too restrictive, excluding many substates with 'native-like' properties and artificially increasing theoretical τ_{nat} values[72]. Instead we chose to designate a *native ensemble*, or a set of conformational substates that interconvert more frequently with each other than with outside substates. Such a graph property is captured by an algorithm called modularity optimization[68], and is particularly suited for this classification task in that it reflects and adapts to the actual network topology, unlike an RMSD threshold. Modularity optimization proceeds by initially designating each substate as its own ensemble and then iteratively combining them until only highly intra-connected ensembles remain, at which point modularity is maximized. For a transition network, modularity is defined as

$$W = \frac{1}{2m} \sum_{i,j} \left[c_{ij} - \frac{k_i k_j}{2m} \right] \delta\left(s_i, s_j\right) \tag{2.2}$$

where c_{ij} is the number of transitions between substates i and j, k_i is the total number of transitions to substate i, k_j is the number of transitions to substate j, $2m$ is the total transition count in the network, and $\delta\left(s_i, s_j\right) = 1$ when substates i and j reside in the same ensemble s and 0 otherwise (elsewhere l_{nn} or l_{n} denote total edges among nodes in the nonnative or native ensembles, respectively). Other formulations for optimizing modularity are possible, including normalized cut and conductance criteria[73]. Maximizing W yields multiple ensembles for each of ten analyzed transition networks, but only one per network will be defined as the native ensemble. Within each candidate ensemble, five random conformers were sampled from every constituent substate, and the aggregate number of conformers within a cutoff r_{nc} to the PDB-deposited native structure were counted and compared with similar counts from the remaining candidate ensembles. The ensemble with the most conformers under the cutoff was designated to be the native ensemble, and all its substates, not only those under the cutoff, were included. All substates not in the native ensemble were defined to be in the nonnative ensemble.

The cutoff itself, r_{nc}, was determined by identifying the RMSD value such that 5% of total trajectory frames were within r_{nc} Å from the PDB-deposited crystal structure (Table 2.1). For 8

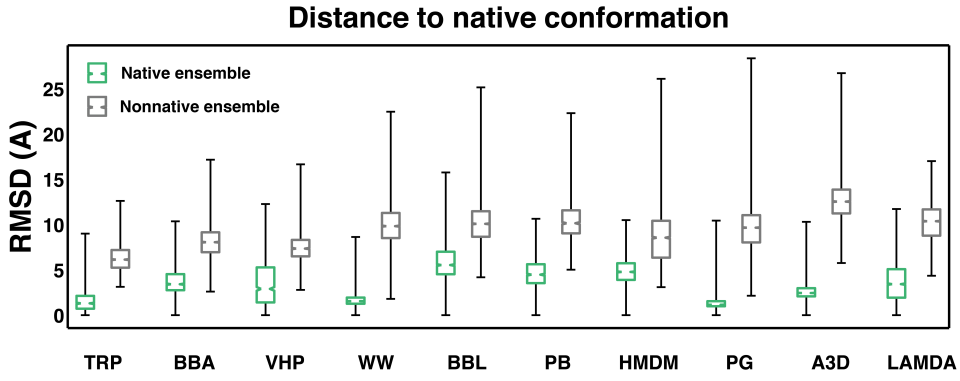


Figure 2.8: RMSD of native and nonnative ensembles to native conformation. All trajectory frames from the native (green) and nonnative (gray) ensembles were aligned to the defined native conformation. The resulting distribution of deformations (one value per conformer) is depicted as a boxplot with box limits demarcating the 25th and 75th quartiles and the whiskers showing the total data range.

of 10 proteins, 100% of the frames with RMSD $< r_{nc}$ were found in the native ensemble; the same values were 82% for BBA and 93% for BBL. These last two percentages confirm our qualitative suspicion of ensemble misclassification for these two systems, visible in Fig. 2.6. For BBL there is a collection of nodes very close to the native ensemble (lower right) that has been classified as nonnative and represents kinetic traps according to the computed frustration scores. For both BBL and BBA the landscape topography suggests the native ensemble needs to be slightly more inclusive. Naturally, other definitions of the native or native ensemble are possible[74,75], but the protocol followed here permits comparison with Ref. 30 and allows us to apply a classification method without invoking knowledge-based thresholds.

Substates containing the snapshot nearest the experimental native structure were always contained in the assigned native ensemble (Fig. 2.6, white diamonds). The number of substates assigned to the native and nonnative ensembles was k_{n} and k_{nn}, respectively (Table 2.1). Although the nonnative ensembles were partitioned variously during iterations of modularity optimization, the constituent substates of the (eventually-defined) native ensemble were in fact identically preserved through all iterations for all proteins. The algorithm converged in seconds for all networks. Modularity optimization operates exclusively on transitions counts, however we observed that secondary structure also separated cleanly between the native and nonnative ensembles as a result of this classification (Fig. 2.10). The identified native ensembles are shown in green in the network

representations (Fig. 2.6). For purposes of computing RMSD, a *native conformation* was defined to be the trajectory snapshot nearest the theoretical mean structure of the entire native ensemble (Fig. 2.6, black diamonds and Fig. 2.15, tube representations).

■ 2.2.4 Defining $Q(t)$ and secondary structure

The proportion of native contacts present in any trajectory conformer, Q, is a useful reaction coordinate for monitoring folding progress[76,77] or modeling energy barriers[78]. It is a degenerate quantity in that many distinct conformers could map to an identical Q value[79,80]. We defined native contacts as those residue pairs whose separation ($C^\alpha - C^\alpha$) was less than 10 Å for at least 65% of the conformers within the native ensemble. Native contacts separated by fewer than 7 amino acids in the primary structure were excluded. We denote the percentage of native contacts as $Q_{\mathrm{n}}(t)$ and the percentage of nonnative contacts as $Q_{\mathrm{nn}}(t)$ for some time t.

We quantified the presence of secondary structure in each trajectory frame using Protein Secondary Element Assignment (P-SEA), which labels every residue as in either an unstructured coil, alpha helix, beta sheet, or 'other' configuration[81] (Fig. 2.10). An 'ideal' sequence of native secondary structure assignment was defined as the residue-wise assignment most common within the native ensemble and termed the structure sequence. The presence of native secondary structure throughout the simulation was quantified by dividing the number of native-like P-SEA assignments by the total number of beta sheet and alpha helix assignments within the structure sequence. This value is denoted $H_{\mathrm{n}}(t)$. Nonnative secondary structure, which captures the percentage of alpha and beta secondary structure assignment that is unlike that found in the structure sequence, is denoted $H_{\mathrm{nn}}(t)$.

■ 2.2.5 Mean first passage times

Having determined the set of substates defining the native state, we next derived the expected mean first passage time of each nonnative substate to the native ensemble as put forward in Ref. 50 (alternative algorithms for computing transit times are given in Ref. 82). First, we estimated the symmetric transition probability matrix P from the clustering results using the MSMBuilder2 MLE method to guarantee detailed balance[83]. This matrix carries jump probabilities for the embedded discrete Markov chain[41], but can also be expressed as a rate matrix $\mathbb{K}$ that approximates the continuous time transition rates[40].

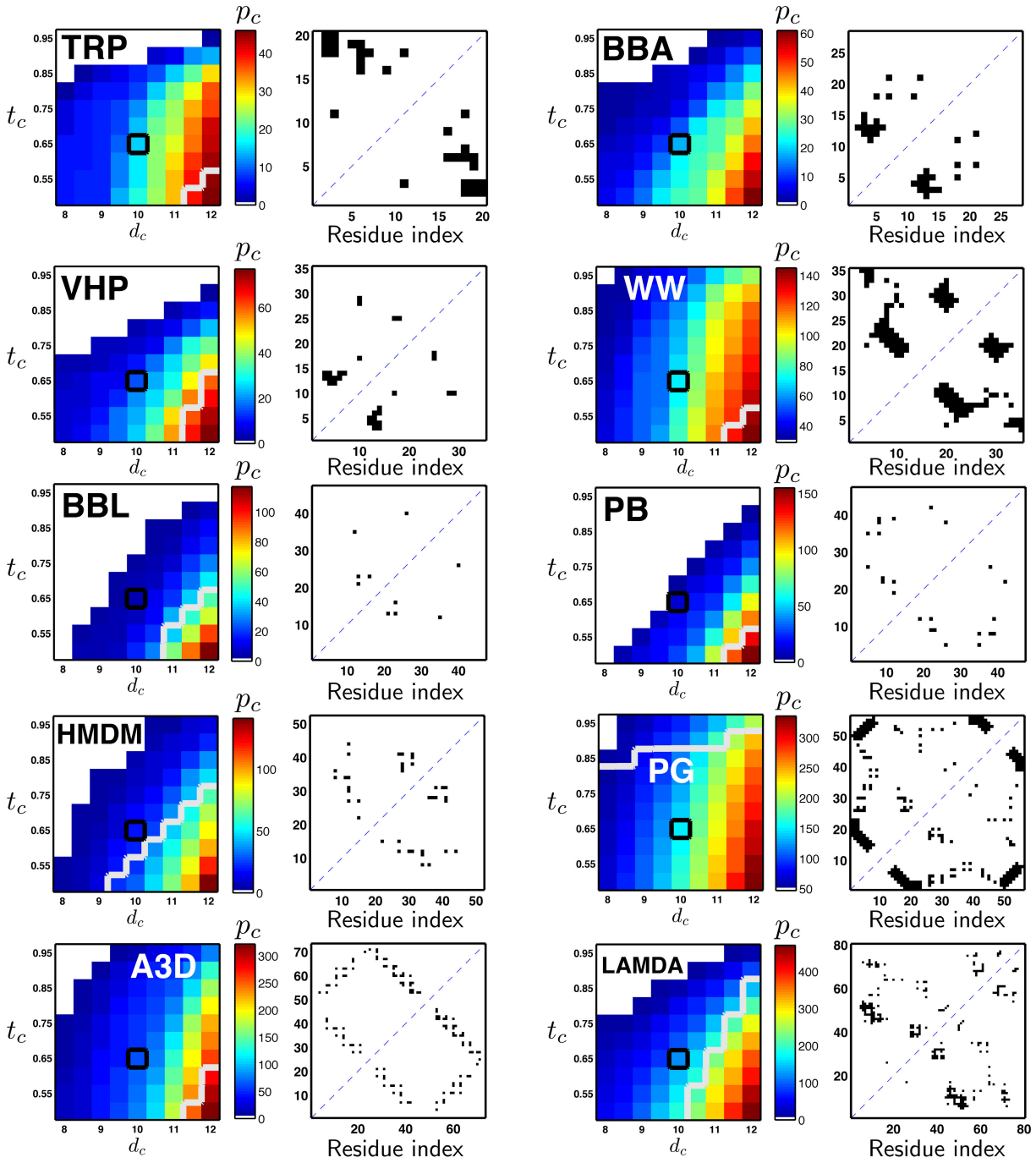


Figure 2.9: Temporal and spacial thresholds for defining native contacts. Heat maps (left columns) show the number of native contacts, p_c, defined for a range of interresidue distances, d_c (Å), and temporal thresholds, t_c (proportion of native frames). The black box indicates the selected threshold pair selected. Gray contours delineate threshold pairs that produce native contacts in the *nonnative* ensemble. Resulting contacts per protein are depicted in the contact maps (right columns).

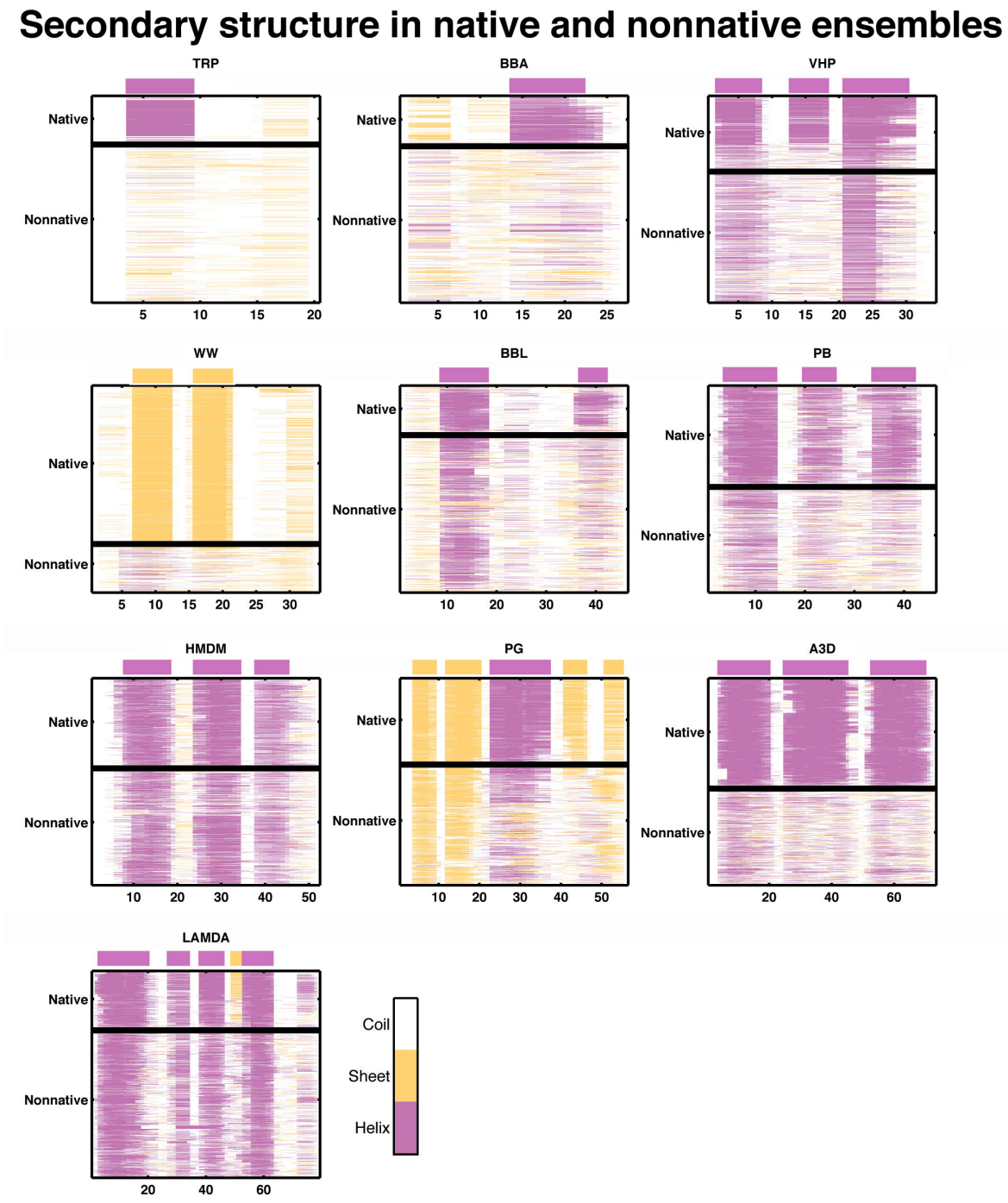


Figure 2.10: Secondary structure is shown for every trajectory frame used in the clustering and subsequent analysis. Frames classified as belonging to the nonnative ensemble are shown in panels' lower portions; native conformers are in the upper divisions. The *structure sequence* (see Section 2.2.4) is depicted above the upper x-axis. Lower abscissa labels denote residue indices for each peptide, starting from 0. Residue-wise secondary structure assignment was performed on C_α coordinates with P-SEA[81].

For each nonnative substate i we modify $\mathbb{K}$ to have zero transition rates to all substates previously connected to i. We then compute the formal matrix exponentiation $e^{(\mathbb{K}_{0i}t)}$ for geometrically-spaced t values ($t = 50(1.2^r)$, $r \in [0, 1, \ldots 40]$. That is, $t \sim 50ns \ldots 74\mu s$). The fraction of trajectories, starting at (nonnative) state i, that will arrive at the native ensemble N before time t is then given by:

$$P_{iN} = \sum_{j \in N} \left[e^{\mathbb{K}_{0i}t} \right]_{ij} \tag{2.3}$$

where j indexes substates in the native ensemble. This fraction consistently converged for all substates (i.e. $\min_i P_{iN} = 0.9964$ at t_{max} out of all proteins). The mean first passage time (see Fig. 2.16) is then given by

$$\tau_{iN} = \int_0^\infty \frac{\mathrm{d}P_{iN}(t)}{\mathrm{d}t} t \mathrm{d}t. \tag{2.4}$$

The starting state can be revisited before arriving at the target substate in this formulation of MFPT's.

■ 2.2.6 Frustration scores

We then ask how these mean first passage times to the native ensemble, or transit times, change in response to network perturbation, the removal of a substate in the nonnative ensemble. To that end, we remove a substate i in the nonnative ensemble from the network and then observe the percentage change between unperturbed (τ_{jN}) and perturbed (τ^*_{jN}) transit times, in both cases for all nonnative substates $m \in [1 \ldots k_{\mathrm{nn}} \neq i]$ (see Fig. 2.1). That is,

$$\bar{f}^i_{\mathsf{nat}} = \frac{100}{k_{\mathrm{nn}} - 1} \sum_{m \neq i} \frac{\tau^*_{mN} - \tau_{mN}}{\tau_{mN}}, \tag{2.5}$$

where the bar thus indicates the average percentage change in transit time over all $k_{\mathrm{nn}}-1$ substates in the nonnative ensemble. Substates in the native ensemble are never removed throughout the procedure. Any isolates resulting from removing node i were discarded when computing $\bar{f}^i_{\mathsf{nat}}$, but this was rare ($\frac{\mathsf{isolates}}{k_{\mathrm{nn}}} < 0.01$ for all proteins except LAMDA, $\frac{\mathsf{isolates}}{k_{\mathrm{nn}}} = .021$). Frustration scores $\bar{f}^i_{\mathsf{nat}}$ quantify the kinetic impact, positive or negative, for each nonnative substate i, expressed as percentages in Fig. 2.11 and Table 2.2. States with positive frustration scores are termed *facilitators*; those with negative frustration scores are termed *inhibitors* or *kinetic traps*. All analysis subsequent to clustering was performed in MATLAB[84]. Due to the matrix exponentiation,

complexity of $\bar{f}_{\mathsf{nat}}$ computation is $\sim \mathcal{O}(N^3)$, and runtimes were between 4min (PB, $k_{\mathrm{nn}} = 167$) and 20hr (PG, $k_{\mathrm{nn}} = 2248$) on a 12-core cluster.

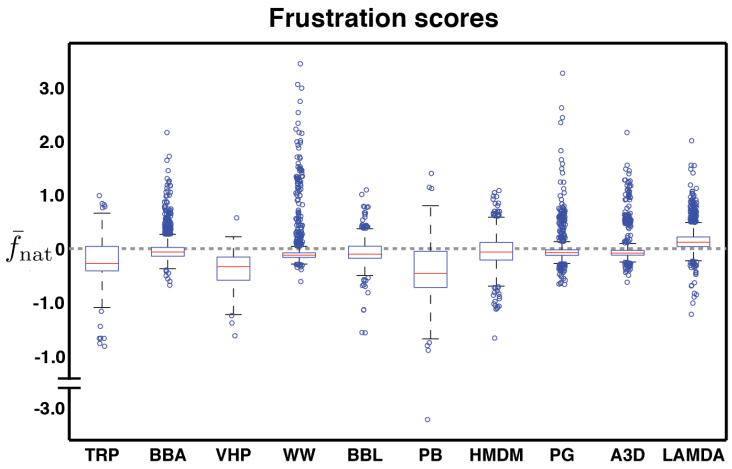


Figure 2.11: Frustration scores. Nonnative substate $\bar{f}_{\mathsf{nat}}$ values are shown for all ten proteins. The number of data points corresponds to the number of substates in the nonnative ensemble, k_{nn} (Table 2.1). Values less than zero indicate a kinetic trap, those above zero indicate a substate that facilitates transition to the native ensemble. Central red marks indicate the median; box edges are the 25th and 75th percentiles. See Fig. 2.16 for a comparison with frustration scores of phantom networks.

■ 2.2.7 Network representations

Substate transition matrices are usually very sparse, especially in the nonnative ensemble (Table 2.2). Most transitions are forbidden due to the involved steric clashes, backbone geometry restrictions, and repulsive electrostatics. Graph-based visualizations of the FEL thus have interpretive value in conveying only the transitions that do take place as well as the relative sizes of conformational states[85,86]. We used Gephi[87] to represent each protein's transition network (Fig. 2.6). Network layouts were optimized using the Force Atlas algorithm, first allowing and then penalizing node overlap, in both cases with an inter-node repulsion strength of 200. Edge weights were scaled according to the transition matrix, specifically $1000 \cdot \mathbb{K}$, but are not differentiated visually in the figure. The repulsion force acts between all nodes, whereas node attraction is relative to connecting edge weight, so unconnected nodes feel zero direct attraction. Node diameters reflect their relative populations, but the smallest node is shown no smaller than $1/30^{\mathrm{th}}$ the size of the largest node for clarity. Network radial orientation was rooted with the native ensemble facing east.

2.3 Results

Conformational landscapes of the ten proteins can be conveniently depicted as networks of nodes and edges that illustrate folding properties of each peptide. These representations are shown in Fig. 2.6A. The native ensembles as defined in Methods are colored green, though the maps' layouts themselves were created without pre-knowledge of native or nonnative substate assignments. All nonnative substates are colored according to their computed $\bar{f}_{\mathrm{nat}}$ values.

Table 2.2: Transition network summary statistics. Details of transition networks in Fig. 2.6. Columns 1-7: parenthetical values denote properties of the native ensemble, all others to the nonnative ensemble. The range of frustration scores is given in the first two columns. Median substate size refers to the number of trajectory snapshots clustered into each conformational substate. Substate width refers to average intra-substate pairwise RMSD. The number of substates classified by modularity optimization as being in the nonnative (native) ensemble is denoted by $k_{\mathrm{nn}}(k_{\mathrm{n}})$. Maximum modularity value, W, for the modularity optimization algorithm utilized (see Methods) is also given. The last column shows the ratio between (1) total transition edges connecting nonnative and native ensembles ($l_{\mathrm{nn}\to\mathrm{n}}$) and (2) the total number of edges in the nonnative ensemble (l_{nn}), expressed as a percentage.

	min $\bar{f}_{\mathrm{nat}}$	max $\bar{f}_{\mathrm{nat}}$	median substate size	mean substate width, Å	median neighbors per substate	$k_{\mathrm{nn}}(k_{\mathrm{n}})$	transition matrix density, %	W	$\frac{l_{\mathrm{nn}\to\mathrm{n}}}{l_{\mathrm{nn}}}$
TRP	-1.8	0.1	7 (11.5)	3.4 (2.6)	14 (12)	387 (30)	3.6 (22.3)	0.30	5.74
BBA	-0.7	2.2	5 (12)	4.3 (2.9)	8 (16)	905 (93)	1.0 (16.2)	0.42	3.89
VHP	-1.6	0.6	7 (9.5)	5.3 (4.6)	13 (16)	207(44)	6.1 (19.3)	0.23	10.23
WW	-0.6	3.5	2 (68)	4.7 (1.7)	4 (86)	2067 (207)	0.2 (39.2)	0.46	0.68
BBL	-1.6	1.1	7 (15)	6.2 (4.8)	11 (18)	758 (102)	1.5 (12.7)	0.42	8.36
PB	-4.2	1.4	5 (21)	7.0 (4.7)	10 (20)	167 (41)	5.8 (39.7)	0.34	6.61
HMDM	-1.7	1.1	5 (18)	5.6 (4.0)	8 (29)	517 (137)	1.8 (17.8)	0.44	20.86
PG	-0.7	3.3	4 (78)	5.5 (2.5)	6 (27)	2248 (62)	0.4 (37.5)	0.66	0.50
A3D	-0.6	2.2	4 (88.5)	8.2 (3.6)	8 (40)	1346 (68)	0.6 (48.5)	0.45	0.51
LAMDA	-1.2	2.0	6 (18.5)	6.0 (4.6)	5 (18)	1181 (112)	0.6 (14.7)	0.68	2.25

2.3.1 Properties of transition networks

Many general phenomenological aspects of protein folding are visible in these abstractions in addition to features that distinguish the folding behavior of specific polypeptides. The prevalence of large substates within the highly-interconnected native ensemble (see Table 2.2), for example, reflects the loss of entropy a folding peptide experiences upon attaining the energetically favorable folded conformation (cluster size and widths given in Figs. A.2 and A.8). Secondly, folding facilitators (blue substates) are as expected mostly proximal to the native ensemble due to being conformationally very native-like[88]. It is important not to over-interpret properties of the facilitators because the classification between native and nonnative is subjective. Here we invoked

modularity optimization for the classification task, but by using splitting probabilities[89] or other reaction coordinates we could perhaps reassign some of our facilitator substates into the native ensemble, or vice versa. However, the topological isolation of the native ensemble, especially for TRP, WW, PG, A3D, and LAMDA, suggests modularity optimization effectively classifies the folded and unfolded ensembles without invoking any protein-specific parameters. A particularly evident separation between the two ensembles characterized the transition maps of WW, PG, and A3D, all of which had less than one percent of nonnative edges connecting the nonnative and native ensembles (range for all proteins: $0.5 - 20.9\%$, see $l_{\mathrm{nn}\to\mathrm{n}}/l_{\mathrm{nn}}$ in Table 2.2). Higher values of this measure indicate less homogeneous folding pathways[90], most evident in HMDM and PB transition maps. A fair general statement is therefore that low-entropy bottlenecks are suggested by the networks of TRP, WW, PG, A3D, and LAMBDA but are harder to identify for BBL, PB, and HMDM.

Large kinetic traps, transition bottlenecks, and facilitators, among other topological motifs, are unequally prominent among the networks. Several of the maps depict a nonnative ensemble of freely interconverting structures that form no apparent energetically-coherent substates (min $\bar{f}_{\mathrm{nat}}$ values for WW, PG, and A3D are $-0.6, -0.7$, and -0.6, respectively), whereas substantial kinetic traps characterize the nonnative ensembles of TRP, BBL, PB, and HMDM (min. $\bar{f}_{\mathrm{nat}} = -1.8, -1.6, -4.2$, and -1.7, Table 2.2). The distribution of frustration scores for each peptide is shown in Fig. 2.11. We also compared τ_{nat} and $\bar{f}_{\mathrm{nat}}$ values to corresponding quantities computed for phantom (i.e. synthesized) networks in Fig. 2.16. The comparison reveals that the degree distribution inherent to the transition network of each peptide is sufficient input for approximating τ_{nat} and $\bar{f}_{\mathrm{nat}}$ in generated networks.

However, a more quantitative analysis is necessary to reveal the specific conformational features, or structural parameters, that are responsible for the frustration scores unique to each protein's unfolded ensemble. We focus primarily on properties of kinetic traps because facilitators inherently border a native/nonnative delineation that is convenient but imposed; any conformational differences between facilitators and native substates are likely to be subtle with regard to the structural parameters used here[88]. We first address clustering properties that could be thought to cause negative outlier $\bar{f}_{\mathrm{nat}}$ values and then discuss the structural features that indeed correlate with kinetic frustration.

■ 2.3.2 Is kinetic frustration a clustering artifact?

As shown in Fig. 2.11, most substates within the nonnative ensemble are kinetically neutral; their individual presence in the FEL has little impact on the expected transit times of other substates. Importantly, these substates need not be small or have few constituent conformers. While substates with substantial positive or negative frustration scores tend to be above average in size, the converse is not true (Fig. A.2). That kinetic traps, as identified through $\bar{f}_{\mathrm{nat}}$, must contain a substantial number of conformers reflects our intuition that kinetic traps represent local energetic minima with stabilizing intramolecular interactions in the nonnative FEL[91]. Substate size (i.e. number of members) as a descriptive trait can be contrasted with substate width (i.e., the mean pairwise RMSD of any two of its members), which provides an approximation of local entropy. If kinetic traps presented increasing substate width as $\bar{f}_{\mathrm{nat}}$ values became more extreme, we could conclude that $\bar{f}_{\mathrm{nat}}$ values are actually artifacts of the clustering step. In this scenario larger and larger peripheral regions of the configurational state space are unfairly grouped together during clustering, resulting in artificially exaggerated $\bar{f}_{\mathrm{nat}}$ values. We observed, in contrast, that kinetic traps display decreasing width values, suggesting they represent genuine local energy minima (Figs. 2.15B and A.8).

■ 2.3.3 Properties of kinetic traps

Frustration scores directly reflect the transition topology. Having discussed that clusters in our networks are well-formed, we next investigate conformational causes of this observed topology. Specifically, are there general structural features that cause kinetic traps[92]? We selected five structural parameters that share the desirable properties of normalizability and interpretability and computed them for all substates (all nonnative and native trajectory frames) in the transition networks. Definitions for native contacts, Q_{n}, nonnative contacts, Q_{nn}, native secondary structure H_{n}, and nonnative secondary structure H_{nn} are given in Section 2.2.4, and our fifth structural parameter was standard RMSD (against the native conformation). Figure 2.12 illustrates the relationship between $\bar{f}_{\mathrm{nat}}$ and fractional contacts (Q_{nn} or Q_{n}) for all nonnative substates within the conformational landscape of LAMDA. The relationship observed confirms our expectation that kinetic traps display more nonnative contacts than the nonnative ensemble in general. The presence of interresidue contacts, both native and nonnative, is normalized against the corresponding

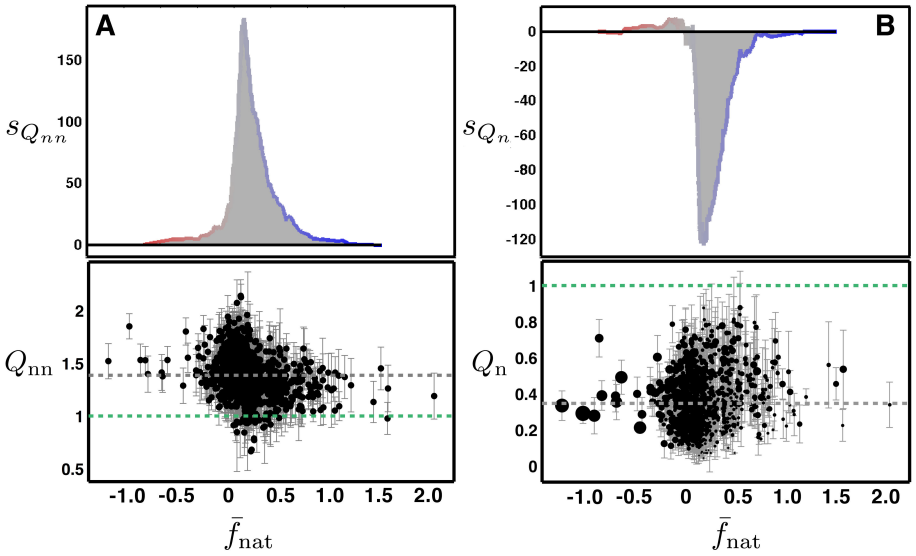


Figure 2.12: Structural features in the nonnative ensemble are related to kinetic frustration. Average intra-substate Q_{nn} (C) and Q_n (D) values for LAMDA are plotted against frustration scores, $\bar{f}_{\mathrm{nat}}$, showing that nonnative contacts are associated with kinetic frustration. Structural parameter values, Q_{n} and Q_{nn} included, are normalized against the average corresponding values within the native ensemble. Marker widths are scaled according to substate populations, and error bars indicate one standard deviation. Dashed lines show normalized average values for the nonnative (gray) and native (green) ensembles. Cumulative sums, (A) $s_{Q_{\mathrm{nn}}} = \sum_{\bar{f}^{\mathrm{min}}}^{\bar{f}^i}(Q_{\mathrm{nn}} - \bar{Q}_{\mathrm{nn}})$ and (B) $s_{Q_n} = \sum_{\bar{f}^{\mathrm{min}}}^{\bar{f}^i}(Q_{\mathrm{n}} - \bar{Q}_{\mathrm{n}})$, (see main text) show the propensity of structural features to be more associated with negative or positive $\bar{f}_{\mathrm{nat}}$ values. When integrated these curves yield the bias values, β, that allow quantitative comparison between proteins (Fig. 2.14 and Table 2.3). Color values along the curve correspond to substate color values in Fig. 2.6.

quantity observed in the native ensemble. Mean values for these features are shown with dashed horizontal lines, gray for the entire nonnative ensemble, green for the entire native ensemble. For LAMDA, we conclude that the enrichment of nonnative contacts among kinetically frustrated substates is one hypothesis for the appearance of outlying 'red' substates in LAMDA's transition network (Fig. 2.6, white arrow).

In normalizing Q_n and Q_{nn} against their respective values in the native ensemble, we can evaluate their correlative relationship to frustration scores and then compare among protein systems. We thus quantify whether a feature is more enriched among kinetic traps or facilitators with a *bias value*

$$\beta_F = -\int_{\bar{f}^{\mathrm{min}}_{\mathrm{nat}}}^{\bar{f}^{\mathrm{max}}_{\mathrm{nat}}} \sum_{\bar{f}^{\mathrm{min}}_{\mathrm{nat}}}^{\bar{f}^i_{\mathrm{nat}}} (F - \bar{F})\mathrm{d}f \tag{2.6}$$

for any feature F with mean value $\bar{F}$ (within the nonnative ensemble), where f indexes the

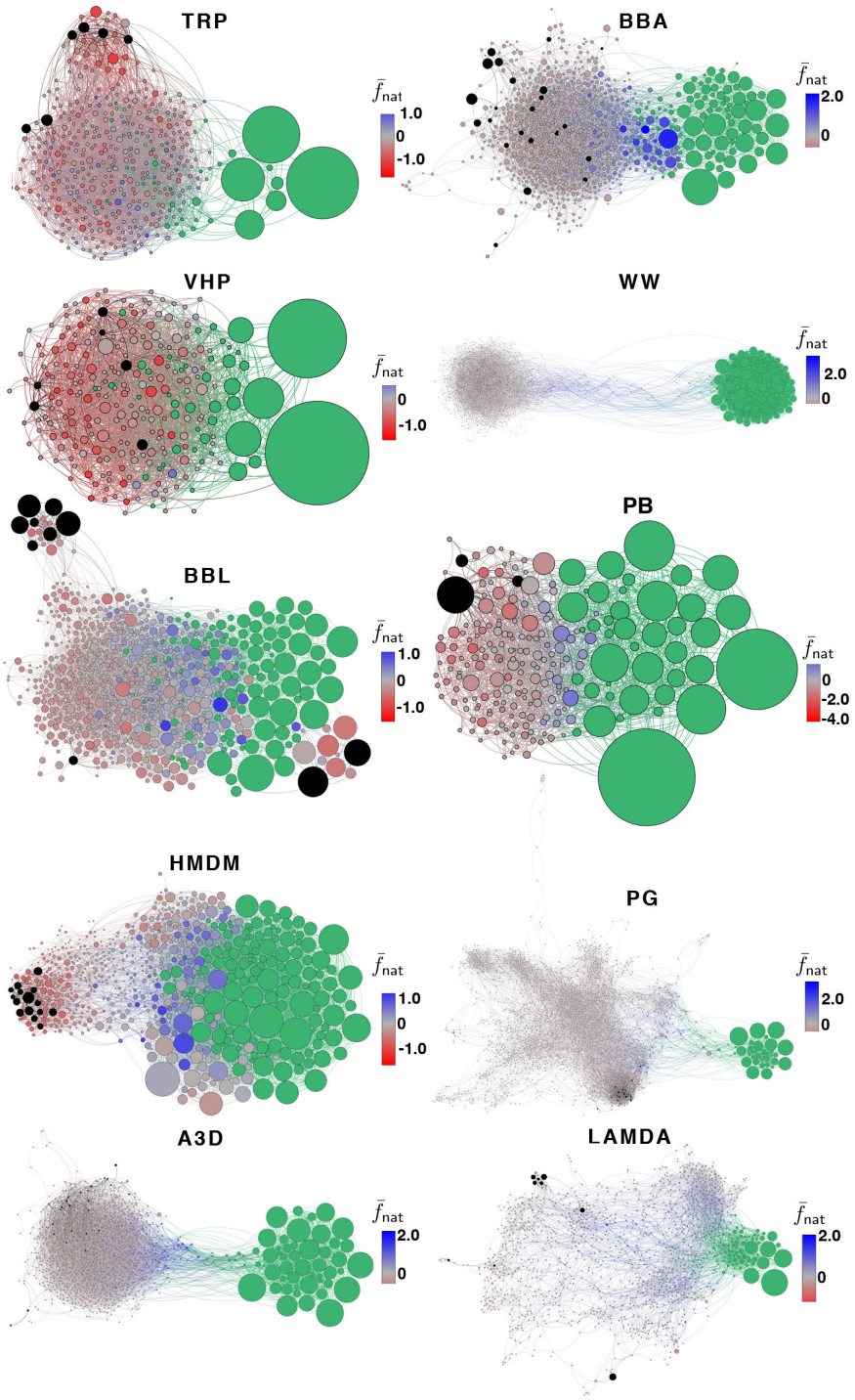


Figure 2.13: Location of major kinetic traps. Conformational substates depicted in structural ensembles in Fig. 2.15 are colored black.

ascendingly-sorted $\bar{f}_{\text{nat}}$ values. Bias values convey whether feature F is more enriched for negative or positive $\bar{f}_{\text{nat}}$ values (Fig. 2.12 panels A and B), where the negative sign allows us to compare with standard linear correlation, which we performed with the addition of weighted substate size (Table 2.3).

Table 2.3: Correlation, ρ, between $\bar{f}_{\text{nat}}$ and five structural parameters compared with bias values, β. Correlation values are weighted according to substate populations[93], whereas frustration biases are non-weighted. Statistically significant β-values are indicated in bold ($p < 0.005$ according to permutation test). See also Fig. 2.14.

	TRP	BBA	VHP	WW	BBL	PB	HMDM	PG	A3D	LAMDA
$\rho_{\bar{f}_{\text{nat}},\text{RMSD}}$	-0.00	-0.36	-0.02	-0.29	-0.14	0.02	-0.46	-0.08	-0.30	-0.35
β_{RMSD}	0.01	**-0.29**	0.04	**-0.32**	**-0.21**	0.01	**-0.30**	**-0.18**	**-0.28**	**-0.28**
$\rho_{\bar{f}_{\text{nat}},H_{\text{nn}}}$	0.02	-0.11	0.05	-0.08	0.01	0.04	-0.04	-0.08	-0.09	-0.21
$\beta_{H_{\text{nn}}}$	-0.02	**-0.13**	0.14	-0.09	-0.09	-0.03	-0.03	**-0.19**	**-0.12**	**-0.10**
$\rho_{\bar{f}_{\text{nat}},H_{\text{n}}}$	-0.00	0.46	0.03	0.40	0.01	-0.21	0.06	-0.01	0.14	-0.08
$\beta_{H_{\text{n}}}$	0.05	**0.39**	-0.05	**0.39**	**0.11**	-0.00	0.01	**0.13**	**0.14**	-0.07
$\rho_{\bar{f}_{\text{nat}},Q_{\text{nn}}}$	-0.45	-0.16	-0.07	-0.01	-0.31	-0.24	-0.09	-0.14	-0.14	-0.34
$\beta_{Q_{\text{nn}}}$	**-0.33**	**-0.13**	-0.03	0.00	**-0.18**	-0.17	-0.05	**-0.15**	**-0.13**	**-0.28**
$\rho_{\bar{f}_{\text{nat}},Q_{\text{n}}}$	-0.21	0.17	0.13	0.56	-0.01	-0.22	0.30	0.06	0.23	0.16
$\beta_{Q_{\text{n}}}$	**-0.16**	**0.15**	0.08	**0.58**	**0.16**	-0.04	**0.21**	**0.17**	**0.20**	**0.16**

Bias values for all five structural parameters are presented in Fig. 2.14. Values near zero indicate the structural parameter is not strongly associated with kinetic frustration. Large positive or negative values indicate a strong relationship. We performed permutation tests to check the statistical significance of these bias values (Table 2.3). As we would expect, native secondary structure and native contacts frequently have positive bias values (TRP $\beta_{Q_{\text{n}}}$ is the only statistically significant exception), indicating that facilitator substates contain many native-like structural features, whereas kinetic traps do not. Significant nonnative secondary structure bias values were observed for BBA, PG, A3D, and LAMDA ($\beta_{H_{\text{nn}}} = -0.13, -0.19, -0.12$, and -0.10 respectively), and significant nonnative contacts bias values were observed for TRP, BBA, BBL, PG, A3D, LAMDA ($\beta_{Q_{\text{nn}}} = -0.33, -0.13, -0.18, -0.15, -0.13$, and -0.28 respectively). Because RMSD-to-native is so commonly invoked as a distance for how far a simulation has progressed, we also computed bias values for RMSD, which were negative and statistically significant for all proteins except TRP, VHP, and PB. Especially for BBA, WW, BBL, HMDM, and A3D, non-specific structural deformity, the characteristic summarized by RMSD, appears more associated

with kinetic frustration than the specific structural features tested. VHP and PB simulations did not present statistically significant bias values for any structural parameters, perhaps due to modularity optimization defining the native ensemble too inclusively for these networks, see Fig. 2.10.

■ 2.3.4 Visualizing kinetic traps

Conformational ensembles consisting of snapshots from the most kinetically frustrated substates are shown in Fig. 2.15, rendered with PyMOL[94]. The five percent of frames that were members of substates with the most negative frustration scores were aligned to their collective mean structure and represented as ensembles (topological context shown in Fig. 2.13). Then the native conformation (see Methods) was added, aligned, and shown in a thicker tube representation. These ensembles illustrate properties of the nonnative ensemble that were suggested by the transition networks in Fig. 2.6. The diffuse nonnative ensembles of WW, PG, and A3D, for example, have few stabilizing nonnative interactions, so even their most kinetically frustrated states appear almost completely unstructured (Fig. 2.15A). In contrast, the nonnative transition maps with more topological isolation among inhibitor substates, especially BBL and PG, show much more homogeneity in their respective structural ensembles. BBL's bias for Q_{nn} was -0.18, suggesting that the nonnative tertiary structure apparent in the ensemble is responsible for the cluster of kinetic traps evident in the network (Fig. 2.6 black arrow). The kinetic trap ensemble for TRP shows that the peptide can get conformationally 'stuck' in a nonnative but stabilized geometry. Although the nonnative configuration in the ensemble and the superimposed native state have very different backbone geometries, the relative compactness of the former may explain why TRP presented a statistically significant negative β_{Q_n}, a property not observed for any other peptide. Peptides with simple contact topologies have been shown in lattice models to allow more interplay between native and nonnative contacts[95], consistent with our findings on TRP. That stabilizing interactions generally may be responsible for kinetic traps is suggested by Fig. 2.15B, which shows that kinetic traps commonly have smaller widths (lower average pairwise RMSD of constituent members) than the nonnative ensemble.

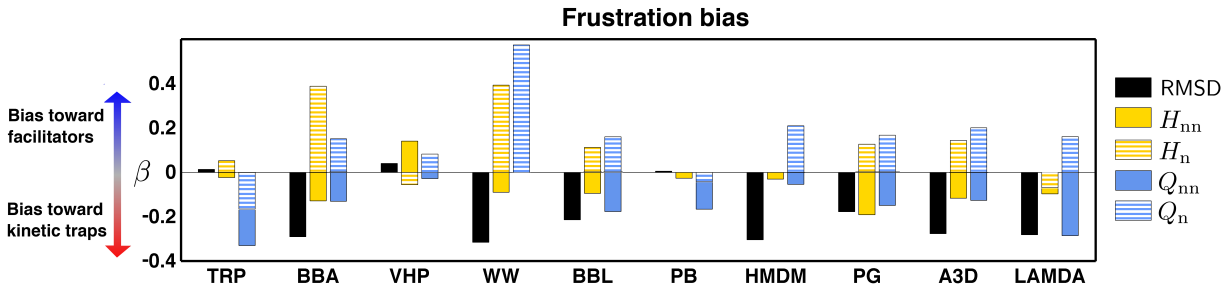


Figure 2.14: A comparison of β values for five structural parameters. Frustration bias values relate structural features to kinetic frustration. Negative values indicate the structural feature is strongly associated with kinetic frustration, i.e., slowing transition to the native state for that protein. Positive values indicate the feature is associated with states that facilitate attainment of the native state. The RMSD distance from the native conformation has the largest negative bias value for BBA, WW, BBL, HMDM, and A3D. Nonnative contacts have the largest negative biases for TRP, PB, and LAMDA. Nonnative secondary structure, $H_{\rm nn}$, is the most biased structural parameter only for PG. Some bias values close to zero are not statistically significant (Table 2.3), indicating the structural parameter has little kinetic impact on folding for that protein. Bars H_{nn} and H_n and also Q_{nn} and Q_n are shown as overlapped pairs. See also Figs. A.3, A.4, A.5, A.6, and A.7.

■ 2.4 Discussion

To compare the folding properties of ten protein sequences we have exploited both quantitative and interpretive aspects of network models of protein folding. Our definition of kinetic frustration is grounded in graph theoretic principles while being consistent with a qualitative understanding of kinetic features, such as kinetic traps. Importantly, kinetic frustration looks at *changes* in MFPT values rather than raw values which are known to collapse to simple recapitulations of node in-degree (Fig. A.1)[49].

The method thus allows direct comparison between temporal folding behaviors and conformational features, the latter summarized by five standard structural metrics that were normalized against their prevalence in the native ensemble. While nonnative intramolecular interactions or nonnative secondary structure formation have been recognized as contributing factors to misfolding[58,97,98] or folding rate reductions[99,100], we quantified the influence of these structural malformations through a normative process that requires no prior domain knowledge of the protein of interest. Specifically, folding for TRP, VHP, PB, PG and LAMDA was most kinetically frustrated by deformations other than that characterized by RMSD, suggestive of stabilizing forces that trap a folding protein in a semi-structured but nonnative conformation. These details were resolvable because we chose to perturb individual substates rather than substate collections within

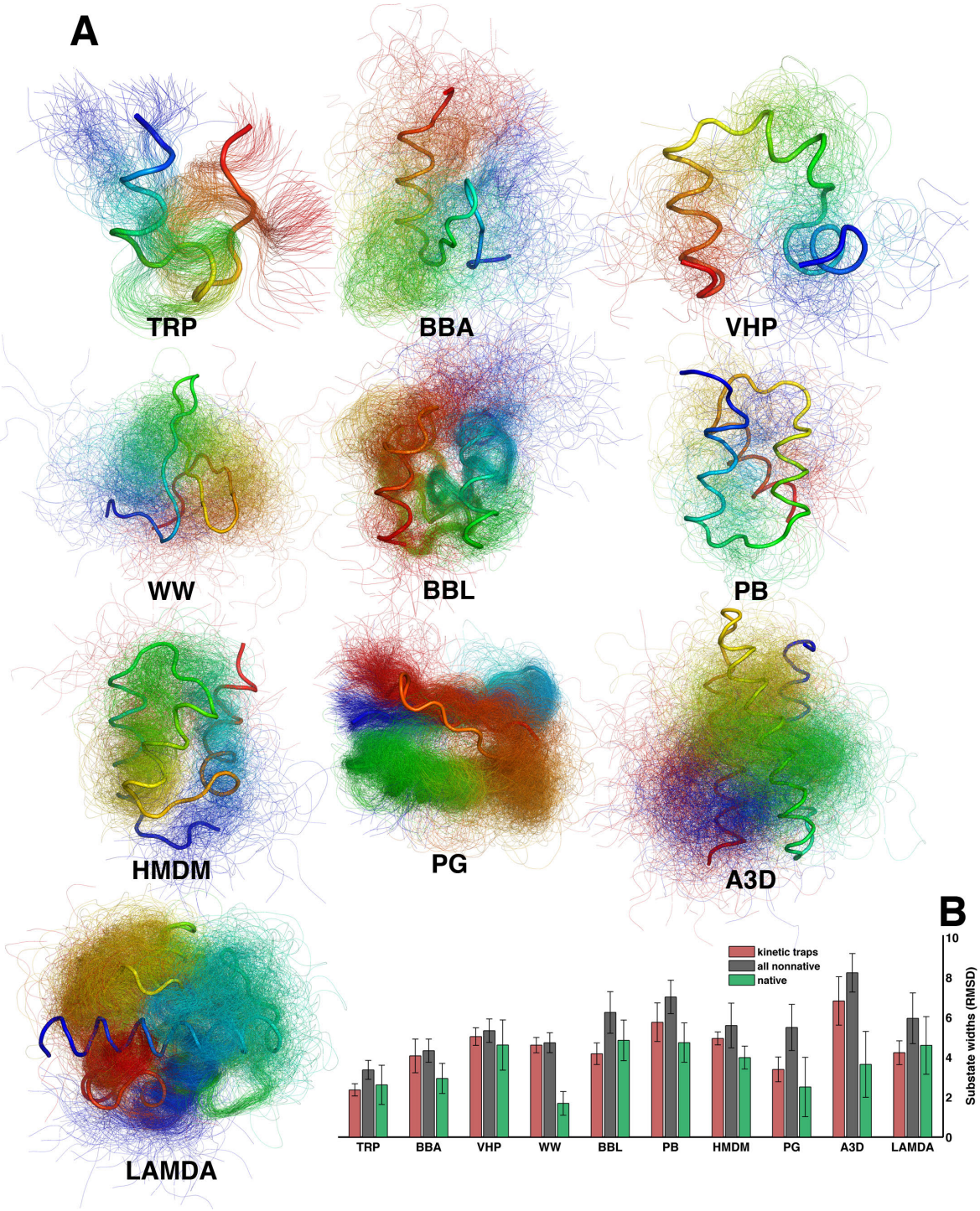


Figure 2.15: Ensemble representation of kinetic traps compared with native structure. (A) Representative structures of kinetic traps, also shown in topological context in Fig. 2.13. (B) A comparison of substate widths (intra-substate pairwise RMSD). Red, substates classified as kinetic traps; gray, all nonnative substates; green, all native substates.

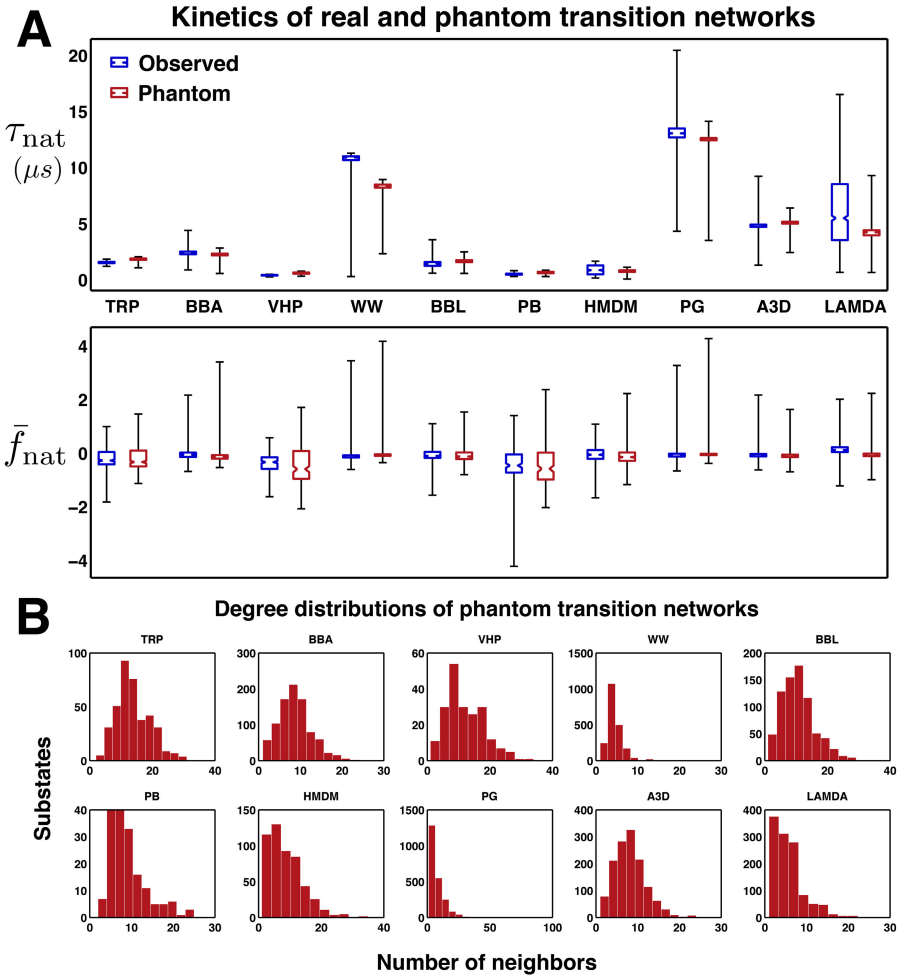


Figure 2.16: (A) A comparison of kinetic properties (transit times, top; frustration scores, bottom) for observed, blue, and phantom (i.e., synthesized), red, kinetic transition networks. Box notches indicate the median, box edges indicate the 25th and 75th percentiles, and whiskers denote data limits. (B) Degree distributions of phantom networks. Phantom nonnative ensembles with substate counts and degree distributions matching those of the observed networks were synthesized with Complex Networks Package for Matlab[96]. The resulting transition count matrix was symmetrized and edge weights were assigned based on corresponding distributions within observed networks. As in Fig. 2.1, a single substate was then added to represent the entire native ensemble, and edges connecting the native and nonnative ensembles were introduced in accordance with their prevalence in the observed networks, c.f. $\frac{l_{\mathrm{nn}\to\mathrm{n}}}{l_{\mathrm{nn}}}$ in Table 2.2. Native ensemble self transitions were assigned to equate with total intra-ensemble transitions from the native ensembles in the observed networks. The resulting transition count matrix then underwent the perturbation process in Methods to yield $\bar{f}_{\mathrm{nat}}$ values.

the transition networks[30]. Regarding the properties and prevalence of facilitator substates in the networks, we make few specific claims. While the WW, PG, A3D, and LAMDA networks suggested bottleneck transition states, different native/nonnative ensemble classifications could lead potentially to different conclusions. For example, it would be interesting to apply transition path analyses to the facilitators to determine if such substates 'commit' to the native and nonnative ensembles equally[101]. In contrast, our conclusions regarding the kinetic traps are unlikely to be overturned by small reclassifications of the native ensemble.

We additionally observed that phantom networks constructed by mimicking bulk topological attributes of the observed networks mostly reproduced emergent kinetic properties τ_{nat} and $\bar{f}_{\text{nat}}$ (Fig. 2.16). Specifically, the weighted degree distributions and edge distributions from the observed networks (Table 2.2) were sampled to generate synthetic or *phantom* networks with matching characteristics, and a single native substate of equal weight to the entire native ensemble was then added to the network. Finally, the appropriate number and weight of edges to connect the nonnative ensemble and the introduced native substate were added such that the ratio $\frac{l_{\text{nn}\to\text{n}}}{l_{\text{nn}}}$ from column 9 of Table 2.2 was preserved. That is, the synthetic native ensemble (a single native substate now) was connected to the synthetic nonnative ensemble in the same fashion as in the observed networks. Paired box plots in Fig. 2.16A show the distribution of MFPT times to arrive at the native conformation τ_{nat} (top) and frustration scores $\bar{f}_{\text{nat}}$ for each nonnative conformation (bottom) for both real (blue) and phantom (red) networks. Clearly, the correspondence between extreme values suggests there is a mapping between a given degree distribution and the range of possible kinetic traps or facilitators given that the neighborhood topology of the native substate is also preserved. Put another way, the two properties of (1) total connectivity between native and nonnative ensembles and (2) overall degree distribution apparently are sufficient for fully describing the expected kinetic frustration in the network. These two properties should not fairly be invoked in a null model since it would be difficult to foresee these topological characteristics from protein first principles. However, the reproducibility does argue that kinetic features, even fairly substantial ones, are probably not exclusive to protein transition networks (see Chapter 3). An additional caveat against over-interpretation is that these boxplots are not weighted by substate size.

If transition networks directly reflect the kinetic barriers, traps, and pathways caused by con-

formational fluctuations, as argued, then further topological properties can hopefully be linked to more nuanced categories of structural deformation. Certainly, subjective concepts such as misfolded intermediates, unstructured intermediates, and kinetic traps, often invoked in the literature of misfolding pathologies[21,102–104], can especially benefit from this type of quantification since simulations are increasingly sampling distant or rare FEL regions where these events occur.

■ 2.5 Conclusion

Finally, we discuss some additional implications of the present work. Whereas certain structural elements have been implicated in promoting or inhibiting folding for particular protein sequences[105], the method here permits comparison *between* protein sequences because all of the descriptive quantities (β, ρ, and $\bar{f}_{\mathsf{nat}}$) are normalized and the method requires no domain knowledge of the studied protein; even the native ensemble classification is performed automatically with modularity optimization (Section 2.2.3). For purposes of protein engineering, candidate mutant sequences can be simulated and the structural feature most responsible for retarding folding determined. For example, if a single point mutation induces a substantial change in folding rates observed in simulation, our method can determine what structural property (helicity, intramolecular contacts, or general disorder (RMSD)) is most responsible. To invoke a comparison to urban infrastructure, our method can deduce for each city (i.e. protein transition network derived from a specific peptide/environment simuation) whether the problematic intersections share some common feature like a median, turn lane, or visual distraction (nonnative helicity, nonnative contacts, etc.).

Turning to our impact on protein inhibition or drug binding, we remind the reader that receptor flexibility is now playing a larger role in discovery efforts[106]. Because docking studies are computationally demanding, a parsimonious set of receptor poses is desirable, but experimenters are faced with the important question of how to select this set. The native state is not the only reasonable choice[107]. Not only might other poses have higher binding affinity, but kinetic traps could also be good target poses for non-competitive inhibitors. Our reasoning is that a conformational state that *hinders* folding to the functional, pathological state is a natural candidate for docking efforts to stabilize that state further. Kinetic traps and preferred folding routes have indeed been identified with landscape visualization techniques and Markov state

models[108], but our framework here permits their kinetic properties to be compared quantifiably, meaning that candidate binding poses according to this line of reasoning could be suggested optimally. In the contrasting scenario that a kinetic trap is associated with aggregation of a therapeutic agent (antibody, etc.), one could evaluate the ability of point mutations to reduce the kinetic frustration of the trap and promote attainment of the functional, non-agglomerative native conformation[109].

Chapter 3

Faster f-score approximations

Complex network models can provide unique insight into the collective phenomena of many-component systems in physics, biology, and the social sciences. A network science subfield exists for predicting and understanding kinetic changes in these networks as a result of node or edge alteration, but the appropriate centrality metric to employ for ranking nodes in a given network and perturbation instance remains an open question. In pursuit of a kinetically-interpretable centrality score, we discuss the f-score, or frustration score, as a quantification of node importance. Each f-score value quantifies the role of the selected node in accelerating or retarding average global transit times to a destination target node. After discussing merits of the f-score centrality metric, we employ spectral and matrix perturbation theory in order to calculate fast approximations to the exact values, illustrated by tests on both synthetic and real medium-sized networks. We report a modest computational improvement (0-400%, network depending) for networks $N > 500$ with low average error ($< 3\%$).

3.1 Introduction

Complicated dynamics that mimic a range of systems from the physical, social, and biological sciences can emerge from simple node and edge models[110]. The basic components encode probabilistic relationships between pairs of model units and collectively constitute the network as a closed-system and induce its behavior. Direct probability flow is permitted between connected nodes; the absence of an edge between nodes means probability can travel between them only indirectly. Despite these minimalistic constituents, such networks can display diverse emergent characteristics and have provided flexible models for disease propagation[111,112], neuronal dynamics[113], router communication[114], protein folding pathways[86], utility grids[115], collaboration histories[116], protein interaction networks[117], and other phenomena at wide-ranging spatial and temporal scales[118–120]. Importantly, these systems frequently confront outside intervention or internal damage whose impact must be predicted or minimized[121–125]. Quantifying vulnerability in the face of targeted or random attacks motivates the more general network science

question: Which network nodes or edges are 'important' or 'central' with respect to the entire graph[126–130]? This question is open because, as worded by one practitioner, 'precise translation' of these terms to a computable metric is required[131]. In response, spectral techniques and graph theoretic principles offer comparisons in global network characteristics before and after a node or edge is altered[132,133]. One drawback is that such characteristics do not always correspond to concrete dynamic or physical quantities. Some interpretable network properties commonly compared between intact and perturbed networks are synchronization[134], diffusion[135], and relaxation rates[136]. Recently, a raft of other centrality metrics have quantified node importance in the context of global graph behavior[137–139]. Some metrics can be combined to improve centrality predictions[140], but many are strongly correlated[141,142].

Here we suggest a different kind of centrality metric for scenarios where a network contains a specific node that is of more interest *a priori* than others[143] (termed a 'target node'). This concern may arise given a known resource sink in utility networks, a functional low-energy conformation in protein folding simulations[50,144], or a master server in computer networks[145]. In these situations we are not concerned with global network behavior *per se* but rather how behavior changes kinetically at the pre-specified target node in response to perturbations (e.g., damage) made elsewhere. The appropriate centrality metric, which associates a scalar quantity with each non-target node, should thus be sensitive to this choice of target node. Our proposed centrality score is now a function of three entities: a user-selected target-node n_t, a queried node n_p whose centrality value is desired, and a global graph structure $\mathcal{H}$ (Fig. 3.1).

Additionally, although centrality is often discussed in the context of traps or hubs[30,54], a target node could very reasonably in fact be less densely connected to the graph (have non-maximal degree) than other nodes of less interest. As an example, we might ask how the infection risk faced by a particular individual (target node, n_t) with few social contacts changes in response to vaccination of a second individual (perturbed node, n_p) with more social contacts. Surprisingly, such low-degree nodes can sometimes have greater network influence than higher degree nodes[146]. In such a case, a useful centrality metric should therefore be sensitive to both local and global topologies, close social contacts and global disease reservoirs respectively[118]. What interpretable metric can quantify the importance of each perturbed node n_p, vis-a-vis the target node n_t, in light of these requirements?

Our choice is called an f-score, f_{n_p}, and is based on the concept of *trapping time*, the average time required to arrive at the target node from any other node (start node) in the network [43,143]. Trapping time is the weighted average of mean first passage times (MFPT's, equivalent to *hitting times* [147] or *transit times*) to n_t over every other node. An individual MFPT value itself expresses the temporal expectation for a random walk starting at one node to arrive at another [41]. As opposed to the shortest path distance, a MFPT value $\tau_{n\to m}(\mathcal{H})$ reflects the influence of all possible paths between nodes n and m in graph $\mathcal{H}$. Whereas MFPT's necessarily are a function of two specified endpoints, in this work concern is restricted to those transition paths that terminate at n_t, and trapping time is then the average over all start nodes:

$$\bar{\tau}_{n_t} = \frac{1}{N-1}\sum_{n\neq n_t}^{N} \tau_{n\to n_t},$$

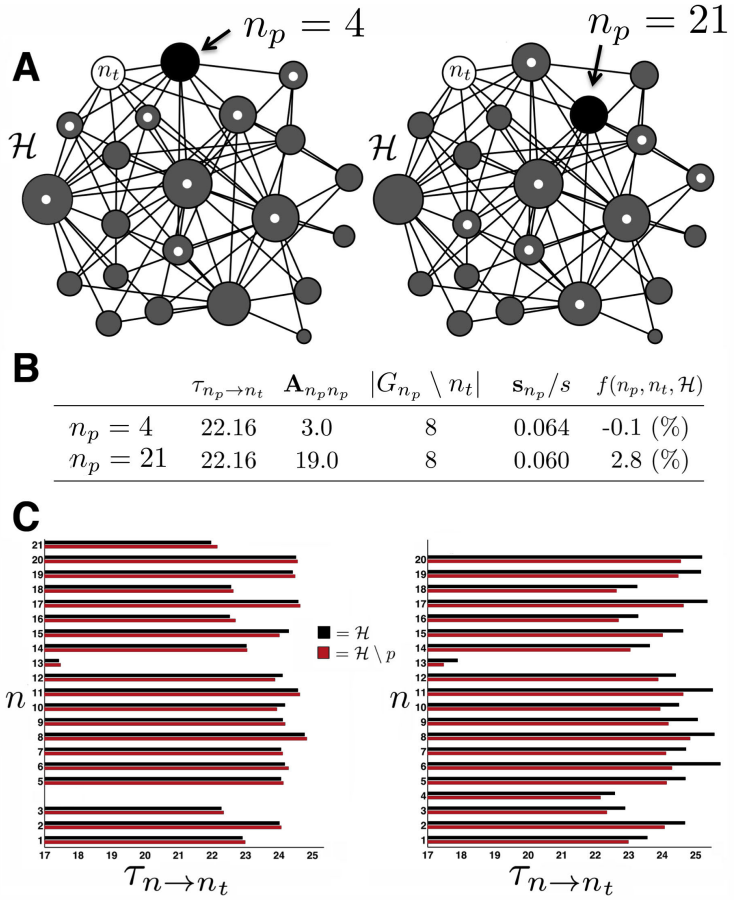
where there are N nodes in the intact network $\mathcal{H}$. We then ask how much this trapping time changes in response to individual removal of non-target nodes $\{n_p\}$ from the network, a kind of derivative of the trapping time. The resulting quantity for excised node n_p, denoted $f(n_p, n_t, \mathcal{H})$, therefore tells us the mean relative change, or frustration, in all MFPT's to n_t as a result of node n_p (Fig. 3.1).

Whereas *frustration* has been invoked in various synchronization contexts [148,149], here the word captures the propensity of a single node to help or hinder transition paths to n_t due to graph topology and not due to any outside intervention. Formally,

$$f(n_p, n_t, \mathcal{H}) = f_{n_p} = 100 * \frac{\left(\frac{1}{N-2}\sum_{n\neq n_t}^{N} \tau_{n\to n_t}(\mathcal{H}_p)\right) - \left(\frac{1}{N-1}\sum_{n\neq n_t}^{N} \tau_{n\to n_t}(\mathcal{H})\right)}{\frac{1}{N-1}\sum_{n\neq n_t}^{N} \tau_{n\to n_t}(\mathcal{H})}, \tag{3.1}$$

where $\mathcal{H}_p$ is identical to $\mathcal{H}$ except node n_p has been excised, so $\mathcal{H}_p = \mathcal{H} \setminus n_p$. Eqn. 3.1 includes a scaling coefficient to emphasize that f-scores convey percentages. In summary, this metric tells use precisely how much *all* paths to n_t are inhibited ($f_{n_p} < 0$) or accelerated ($f_{n_p} > 0$) as a result of node n_p *in the intact graph* $\mathcal{H}$ (Fig. 3.1C). Interpretability of this sort is key to a successful centrality metric [150]. In the following we first connect spectral theory with MFPT's and trapping times and then secondly propose a protocol for approximating f-scores using basic

matrix perturbation theory that is more efficient than direct matrix inversion methods we know of. Examples and tests are conducted with synthetic and real datasets, in all cases using sparse, nonregular, and undirected graphs.



	$\tau_{n_p \to n_t}$	$\mathbf{A}_{n_p n_p}$	$\lvert G_{n_p} \setminus n_t \rvert$	$\mathbf{s}_{n_p}/s$	$f(n_p, n_t, \mathcal{H})$
$n_p = 4$	22.16	3.0	8	0.064	-0.1 (%)
$n_p = 21$	22.16	19.0	8	0.060	2.8 (%)

Figure 3.1: Scheme for computing f-scores, $f(n_p, n_t, \mathcal{H})$. (A) Example network $\mathcal{H}$ with 22 nodes; node widths indicate total degree $\mathbf{s}_n$. Target node n_t is shown in white, removed node n_p in black; $n_p = 4$ left panel, $n_p = 21$ right panel. Neighbors of n_p, denoted G_{n_p}, are indicated with white dots. (B) A comparison of network properties for $n_p = 4$ and $n_p = 21$. In intact graph $\mathcal{H}$ both nodes have the same MFPT to n_t ($\tau_{n_p \to n_t}$); neighbor counts $|G_{n_p} \setminus n_t|$, and total normalized degree $\mathbf{s}_{n_p}/s$ are also similar. Because of differing topology vis-a-vis n_t and contrasting self-loop weights $\mathbf{A}_{n_p n_p}$, nodes $n_p = 4$ and $n_p = 21$ impact graph dynamics differently when removed ($f(n_p, n_t, \mathcal{H})$). (C) Impact in MFPT's to n_t when node n_p is removed. Black bars show MFPTs for intact graph, red bars show travel times when $n_p = 4$ (left) or $n_p = 21$ (right) is removed. Horizontal axes cropped for clarity.

■ 3.2 Methods

For some chosen target node n_t in graph $\mathcal{H}$, denominator and subtrahend in Eq. 3.1 need only be computed once for any desired set of perturbed nodes $\{n_p\} = N_p$. Because the topology in $\mathcal{H}$

is mostly preserved for any single node perturbation, we can therefore exploit spectral properties of $\mathcal{H}$ in order to approximate quickly the first numerator term given that we already know the second, which has no n_p dependence. We begin in this direction by introducing nomenclature relevant to mean first passage times and perturbation theory in the context of complex networks.

Let $\mathcal{H} = (V, E)$ be a weighted, undirected graph where V is the set of vertices and E is the set of edges weights. The vertices or nodes are indexed by $n, m \in \{1 \dots N\}$, with special symbols n_t for the target node and $\{n_g\} = n \in G_{n_p}$ for the perturbed node and neighbors of the perturbed node, respectively. Nodes that are not neighbors of some node n are denoted $\bar{G}_n$. The graph laplacian $\mathbf{L}$, an N x N matrix, is defined as $\mathbf{L} = \mathbf{S} - \mathbf{A}$, where $\mathbf{A}$, the symmetric adjacency matrix is defined such that $\mathbf{A}_{nm} = \mathbf{A}_{mn} = a_{nm} \in E$ is the nonnegative weight of the edge connecting nodes n and m, and $\mathbf{A}_{mm}$ is the weight of self-loops for node m. Because $\mathbf{L}$ contains no information of node self-loops, which are essential for modeling many complex phenomena, our expressions often require matrix $\mathbf{S}$, whose diagonal carries weighted node degrees, i.e., $\mathbf{S}_{mm} = \mathbf{s}_m = \sum_{n=1}^{N} \mathbf{A}_{mn}$. A column vector of these degrees is denoted as $\mathbf{s}$, and $s = \mathbf{s}^T\mathbf{1}$ is the total edge weight in the network, sometimes called $vol(\mathcal{H})$. Estimating the degree distribution (Table 3.2) involves the nonweighted degree vector $\mathbf{x}$ where $\mathbf{x}_m = \sum_{n=1}^{N} (\mathbf{A}_{mn} \neq 0)$. Perturbation of a single node amounts to decreasing all the node's edges, including self-transitions by some relative amount $\epsilon \in [0\ 1]$, i.e., $\mathbf{L}_{p_{n_p,n_p}} = (1 - \epsilon) \times \mathbf{L}_{n_p n_p}$ with corresponding values decreased at nodes G_{n_p} so that $\sum_{m=1}^{N} \mathbf{L}_{p_{nm}} = 0\ \forall n$. Node removal occurs when $\epsilon = 1$. A perturbation impacts the adjacency matrix analogously,

$$\mathbf{A}_p = \mathbf{A} - (\epsilon \mathbf{A}_{[n_p,1:N]} + \epsilon \mathbf{A}_{[1:N,n_p]}).$$

Here and elsewhere brackets denote index ranges.

■ 3.2.1 Mean first passage times, trapping times, and f-scores

With these and a few additional definitions we can compute the pairwise MFPT matrix for all nodes in a weighted, symmetric network $\mathcal{H}$. First, the 'fundamental matrix' $\mathbf{Z}$ from Markov chain literature is defined as

$$\mathbf{Z} = (\mathbf{I} - (\mathbf{P} - \mathbf{A}))^{-1}, \tag{3.2}$$

where $\mathbf{P} = \mathbf{S}^{-1}\mathbf{A}$ is the row-stochastic transition probability matrix and $\mathbf{I}$ is the identity matrix. The traditional expression for computing all pairwise MFPT values then is

$$\mathbf{M}(\mathcal{H}) = \{\tau_{n\to m}(\mathcal{H})\} = (\mathbf{I} - \mathbf{Z} + \mathbf{E}\mathbf{Z}_{diag})\,\mathbf{D}, \tag{3.3}$$

where $\mathbf{Z}_{diag}$ is equivalent to $\mathbf{Z}$ but with vanished off-diagonals, $\mathbf{E}$ is a constant matrix of all 1's, and $\mathbf{D}$ is also diagonal and carries in its diagonal the inverse of the stationary distribution (or limiting probability): $\mathbf{D}_{nn} = \frac{1}{\alpha_n}$, where $\vec{\alpha}$ is the dominant eigenvector of $\mathbf{P}$[41]. Trapping times $\bar{\tau}_{n_t}$ for some target node n_t are then computed by averaging over the appropriate column of $\mathbf{M}$:

$$\bar{\tau}_{n_t} = \frac{1}{N-1}\sum_{m=1\neq n_t}^{N} \mathbf{M}_{m,n_t}, \tag{3.4}$$

such that our exact f-score definition (3.1) becomes

$$f(n_p, n_t, \mathcal{H}) = 100 * \frac{\bar{\tau}_{n_t}(\mathcal{H}_p) - \bar{\tau}_{n_t}(\mathcal{H})}{\bar{\tau}_{n_t}(\mathcal{H})}. \tag{3.5}$$

Even though the adjacency matrix is generally sparse and $\mathbf{S}$ being diagonal is cheaply invertible, the inversion in (3.2) produces a dense $\mathbf{Z}$. That is, each desired f_{n_p} value requires an expensive matrix inversion and no dynamic or topological information about $\mathcal{H}$ is recycled when iterating over user-selected $\{n_p\}$. Moreover, non-n_t columns of $\mathbf{M}$ are left unused despite the expense of their calculation.

One alternative formulation for $\bar{\tau}_{n_t}$ that flexibly allows n_t to be comprised of an arbitrary set of target nodes is presented in Ref. 50, but efficiency is still an issue because matrix exponents must be evaluated multiple times for each n_p of interest. This approach was employed in Chapter 2. Thankfully, trapping times $\bar{\tau}_{n_t}$ can be computed without explicitly calculating individual transit times $\tau_{n\to n_t}$ and averaging over n as in (3.4). Specifically, a spectral formulation presented in Ref. 151 permits $\bar{\tau}_{n_t}$ to be expressed via Laplacian eigenvectors $\mathbf{u}_{1\ldots N}$ and eigenvalues $\lambda_{1\ldots N}$:

$$\bar{\tau}_{n_t} = \frac{N}{N-1}\sum_{k=2}^{N}\frac{1}{\lambda_k}\left(su_{n_t k}^2 - u_{n_t k}\mathbf{s}^T\mathbf{u}_k\right), \tag{3.6}$$

where the first eigenpair is excluded because $\lambda_1 = 0$. A related treatment with adjacency matrix spectra is also possible[43]. Equation 3.6 invokes all non-dominant eigenpairs, defined as the associated quantities $\{\mathbf{u}_k, \lambda_k\}$ such that $\mathbf{L}\mathbf{u}_k = \lambda_k\mathbf{u}_k$, where as before $\mathbf{L} = \mathbf{S} - \mathbf{A}$. Eigenpairs are indexed by eigenindices $j, k \in \{1\ldots N\}$ and sorted: $\lambda_1 = 0 \leq \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$. Eigenvectors

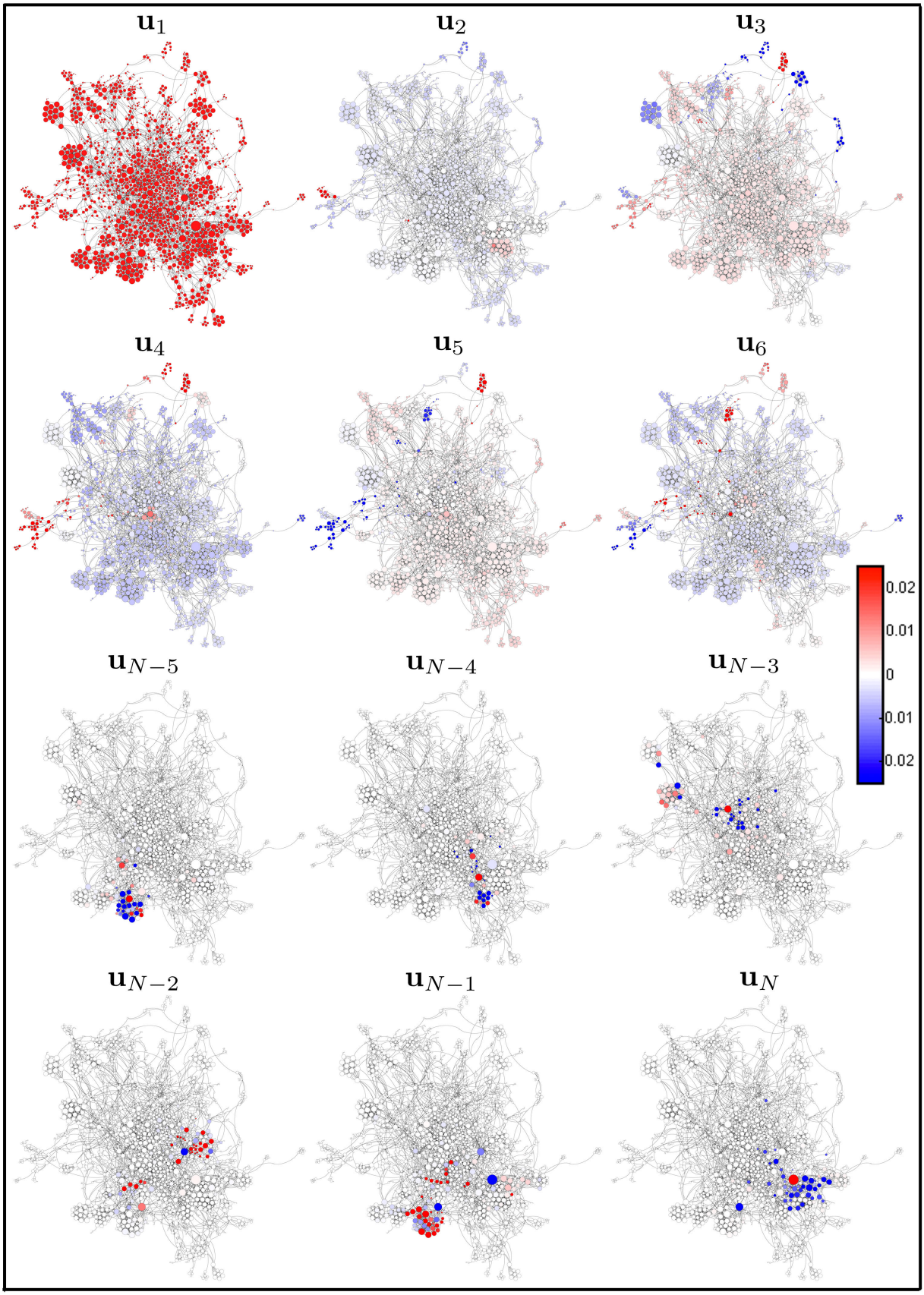


Figure 3.2: (Continued on following page)

Figure 3.2: Visual representation of extreme Laplacian eigenvectors of test network $\mathcal{H}_{\mathsf{YST}}$. High-index eigenvectors ($k = N-5:N$) have very localized 'density' at major topological hubs. Low-index eigenpairs have more dispersed patterns. All eigenvectors are normalized and sum to zero except $\mathbf{u}_1$ which is normalized but has a positive sum: $\mathbf{1}^T\mathbf{u}_1 > 0$. See also Fig. 3.3.

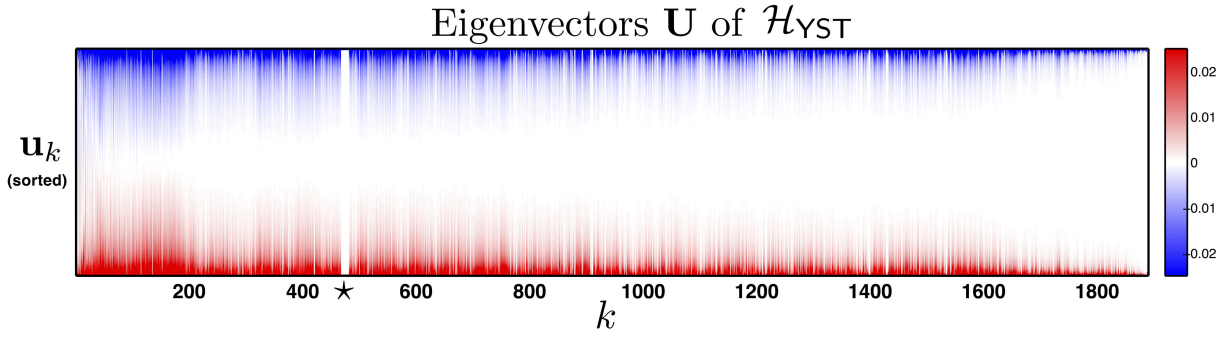


Figure 3.3: Laplacian eigenvectors of test network $\mathcal{H}_{\mathsf{YST}}$. Individual elements of each eigenvector $\mathbf{u}_k$ are sorted to illustrate that low index modes are more disparate than high index modes whose 'densities' are more localized. The range of eigenvector elements was [-0.9754,0.9831]; the range of the colorbar is more limited to emphasize the presence of intermediate values. Eigenindices 468 to 480 are degenerate and have less dispersed 'density' patterns ($\star$).

together form the columns of a matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$, where $\mathbf{U}_k$ or $\mathbf{u}_k$ indicates the kth column and $\mathbf{U}_{ij}$ or u_{ij} denotes the ith element of the jth column of $\mathbf{U}$.

Across many disciplines, these Laplacian eigenpairs are used to map the topology encoded in $\mathbf{L}$ to an alternate or lower-dimensionality basis (Fig. 3.2), often to facilitate coarse-graining[152,153], clustering[154–156], or link prediction tasks[157], and many centrality measures and robustness measures have naturally been formulated from them. For example, one may ask which link or node removals maximally or minimally impact the *algebraic connectivity* λ_2, or the eigenratio λ_2/λ_N[158], both of which are summary measures of dynamic synchronization[159,160]. Spectrum-based centralities have also been formulated without invoking the concept of node removal at all: the *eigenvector centrality* provides direct node rankings by simply comparing elements of the first non-trivial eigenvector(s)[161]. One may also examine an individual row of the eigenvector matrix, i.e. $\mathbf{U}_{[n_p,1\ldots N]}$, whose elements convey the dynamical importance of node n_p within each eigenfrequency[131]. Most such interpretations of $\mathbf{U}$ and λ relate to global behavior over the entire graph. Part of the appeal of synchronization and eigenvector centrality measures is that only dominant and/or extreme eigenpairs are required, meaning these centrality values even for very large graphs are feasible with sparse eigensolvers[162,163]. Formally, Eqn. 3.6 requires the entire

spectrum and cannot take advantage of these numerical methods. However, Eqn. 3.6 favorably permits us to consider each eigenpair separately, and so we associate a symbol $\bar{\tau}_{n_t}^k$ with the trapping time contribution of each distinct eigenpair k: $\bar{\tau}_{n_t}^k = \frac{N}{N-1}\left(su_{n_tk}^2 - u_{n_tk}\mathbf{s}^T\mathbf{u}_k\right)$ such that total trapping time is their sum: $\bar{\tau}_{n_t} = \sum_{k=2}^{N} \bar{\tau}_{n_t}^k$. The central concept is that the spectra of $\mathbf{L}$ and $\mathbf{L}_p$ are closely related and therefore many $\bar{\tau}_{n_t}^k$ values will be unchanged upon network perturbation. That is, given trapping time contributions $\bar{\tau}_{n_p}^k \ \forall k \neq 1$ for the intact graph $\mathcal{H}$, we can selectively estimate only those eigenpairs in $\mathcal{H}_p$ (and thus only those $\bar{\tau}_{n_p}^k$ values) that non-negligibly impact a node's associated f-score. The other variables in Eqn. 3.6, s and $\mathbf{s}$, are known observables of $\mathcal{H}_p$. In summary, instead of an exact f_{n_p} we compute an estimate $\tilde{f}_{n_p}$ by (1) identifying 'free' eigenindices k_F that substantially alter total trapping time $\sum_{k=2}^{N} \bar{\tau}_{n_t}^k$, and then (2) efficiently estimating quantities $\tilde{\mathbf{u}}_k$ and $\tilde{\lambda}_k$ necessary for Eqn. 3.6.

■ 3.2.2 Estimating λ_p and $\mathbf{U}_p$

For λ specifically, convenient analytic expressions that determine the exact eigenvalues are possible in the case that the network presents very controlled structure[143,151]. With complex networks, however, alternatives other than dense eigensolvers include approximations[164], perturbation theory applied to the intact $\mathbf{L}$, or eigenvalue bounds. In the latter, one can bound the maximum shift of the perturbed eigenvalues $|\lambda - \lambda_p|$ given the local topology of n_p[165–167], but in our experience these bounds are not adequately tight and, besides, eigenvalue perturbation (Eqn. 3.14) is more accurate and almost as fast. Regardless, it is the estimation of the eigenvectors $\tilde{\mathbf{U}}$ that represents the largest computational expense.

For notational clarity, tildes are assigned to approximate/estimated quantities of the perturbed spectrum and subscript or superscript p's indicate exact quantities or indices. A matrix of estimated Laplacian eigenvectors is therefore denoted $\tilde{\mathbf{U}}$, while dense eigendecomposition would yield $\mathbf{U}_p$ given $\mathbf{L}_p$.

Using classical first order perturbation theory, for some eigenpair k:

$$\tilde{\lambda}_k - \lambda_k = \frac{\mathbf{u}_k^T \epsilon \mathbf{B} \mathbf{u}_k}{\mathbf{u}_k^T \mathbf{u}_k}, \tag{3.7}$$

where $\mathbf{L}_p = \mathbf{L} + \epsilon\mathbf{B}$ is the Laplacian of $\mathcal{H}_p$[168]. However, in the case that the perturbation impacts a single node n_p, meaning all connected edges (and self-loops) are proportionally decreased by ϵ, the expression can be simplified:

$$\frac{\Delta\lambda_k}{\epsilon} = \frac{\tilde{\lambda}_k - \lambda_k}{\epsilon} = \mathbf{u}_k^T \mathbf{B} \mathbf{u}_k \text{ (subscript in } \mathbf{u}_k \text{ implied hereafter)} \tag{3.8}$$

$$= \left(\sum_{n \in G_{n_p}} u_n \left(\mathbf{u}^T \mathbf{B}_n \right) \right) + u_{n_p} \mathbf{u}^T \mathbf{B}_{n_p} \tag{3.9}$$

$$= \left(\sum_{n \in G_{n_p}} \mathbf{B}_{nn} u_n^2 \right) + u_{n_p} \left(\mathbf{u}^T \mathbf{B}_{n_p} - u_{n_p} \mathbf{B}_{n_p n_p} \right) + u_{n_p} \left(\mathbf{u}^T \mathbf{B}_{n_p} \right) \tag{3.10}$$

$$= \left(-\mathbf{u}^T \mathsf{diag}(\mathbf{B}_{n_p}) \mathbf{u} + u_{n_p}^2 \mathbf{B}_{n_p n_p} \right) + u_{n_p n_p}^2 \left(-\lambda - \mathbf{B}_{n_p n_p} \right) + u_{n_p n_p}^2 \left(-\lambda \right) \tag{3.11}$$

$$= \mathbf{u}^T \mathsf{diag}(\mathbf{L}_{n_p}) \mathbf{u} + u_{n_p}^2 \left(-\mathbf{L}_{n_p n_p} - \lambda + \mathbf{L}_{n_p n_p} - \lambda \right) \tag{3.12}$$

$$= \left(\mathbf{u}.^2 \right)^T \mathbf{L}_{n_p} - 2\lambda u_{n_p}^2 \tag{3.13}$$

$$\Rightarrow \tilde{\lambda}_k - \lambda_k = \epsilon * \left(\left(\mathbf{u}_k.^2 \right)^T \mathbf{L}_{n_p} - 2\lambda_k u_{n_p k}^2 \right). \tag{3.14}$$

where the notation $\left(.^2\right)$ signifies the element-wise exponent, $\mathsf{diag}(\mathbf{x})$ is a zero matrix with $\mathbf{x}$ along its diagonal, and $\mathbf{L}_{n_p}$ denotes the n_pth column of the *intact* Laplacian. Equation 3.8 lacks the denominator present in Eqn. 3.7 because proper Laplacian matrices already have normalized eigenvectors. First-order matrix perturbation is not limited to symmetric matrices or normalized eigenvectors[169], but those properties allow us to make the simplifications that result in Eqn. 3.14 (c.f. Ref. 170 for perturbation of adjacency matrix $\mathbf{A}$).

Likewise, we can also update the eigenvectors using standard perturbation approaches[175,176]:

$$\tilde{\mathbf{u}}_k = \mathbf{u}_k + \Delta\mathbf{u}_k = \mathbf{u}_k + \sum_{j=1 \neq k}^{N} \frac{\mathbf{u}_j^T \left(\mathbf{L}_p - \mathbf{I}\tilde{\lambda}_k \right) \mathbf{u}_k}{\tilde{\lambda}_k - \tilde{\lambda}_j} \mathbf{u}_j. \tag{3.15}$$

In matrix notation the update term for all $k \in k_F$ becomes

$$\Delta\mathbf{U}_{[1:N,k_F]} \leftarrow \mathbf{U}_{[1:N,k_F]} \left\{ \mathbf{U}_{[n_F,k_F]}^T \left(\mathbf{L}_{p[n_F,1:N]} \mathbf{U}_{[1:N,k_F]} - \mathbf{U}_{[n_F,k_F]} * \mathbf{I}\tilde{\lambda}_{k_F} \right) .* \Lambda_{[k_F,k_F]} \right\}, \tag{3.16}$$

where $\Lambda_{ij} = \left(\tilde{\lambda}_i - \tilde{\lambda}_j \right)^{-1}$ and $\Lambda_{ii} = 0,\ i,j \in \{2 \ldots N\}$. This update step has complexity $\mathcal{O}(n^2)$, and updating N eigenvectors of the spectrum costs $\mathcal{O}(n^3)$. Naively implemented, this would constitute a profligate linear estimate to the eigenbasis when exact, direct eigensolvers have the same approximate cost, sparse solvers being cheaper still. In practice, however, the perturbations here require only the subset k_F of the spectrum to be updated for accurate estimates, and the corrections themselves are small and vanish rapidly (see Fig. 3.9). Higher-order approximations for λ and $\mathbf{U}$ are possible[177] but iterative application of first-order estimates trumps these more

Table 3.1: Dataset summary. Six networks are compared based on node count N, edge count nnz, permissibility of self-transitions, degree distribution exponent α, algebraic connectivity λ_2, and spectral radius λ_N. In $\mathcal{H}_{\mathsf{A}}$ edge weights denote average total daily capacity between airport pairs. Edge weights in $\mathcal{H}_{\mathsf{YST}}$ reflect confidence in functional interactions based on aggregated screening studies. In social network $\mathcal{H}_{\mathsf{UC}}$ edges denote the symmetrized number of communicated messages. Networks $\mathcal{H}_{1000}$ and $\mathcal{H}_{\mathsf{A}}$ are shown in Fig. 3.4; the remaining four networks are shown in Fig. 3.8. Power law exponent and standard error was estimated as $\alpha_{\mathcal{H}} = N\left[\sum_{m=1}^{N} \ln \frac{\mathbf{x}_m}{\mathbf{x}_{\mathsf{min}}}\right]^{-1}$ and $\sigma_\alpha = \frac{\alpha-1}{\sqrt{N}}$, where $\mathbf{x}$ is the vector of *unweighted* node degrees[171]. Later in this chapter we also perform tests on the protein transition networks introduced in Table 2.2.

Name	Description	N	nnz	self-loops?	α [171]	λ_2	λ_N
	Synthetic networks:						
$\mathcal{H}_{500}$		500	1896	yes	2.46 ± 0.07	5.02	1.41e+4
$\mathcal{H}_{1000}$		1000	4199	yes	2.26 ± 0.04	17.31	2.37e+4
$\mathcal{H}_{2000}$		2002	9725	yes	2.13 ± 0.03	34.46	8.20e+4
	Real networks:						
$\mathcal{H}_{\mathsf{A}}$	Busiest commercial US airports [172]	500	5960	no	1.64 ± 0.03	0.2	1.4e+05
$\mathcal{H}_{\mathsf{YST}}$	Yeast protein interactions [173]	1890	9464	yes	1.80 ± 0.02	0.39	1.20e+03
$\mathcal{H}_{\mathsf{UC}}$	UC Irvine social network [174]	1893	27670	no	1.56 ± 0.01	0.17	809.1

expensive updates in the literature we have encountered[175]. Additionally, we will show that the selected eigenpairs are often non-extreme and non-adjacent, and most efficient eigensolvers are not traditionally amenable to updating simultaneously non-contiguous eigenpairs[178]. It is for this reason that we choose to update iteratively $\tilde{\mathbf{U}}$ using the method least efficient in traditional implementation but well-suited to the specific perturbation structure $\mathbf{B}$ and stopping criterion $|\Delta\tilde{f}_{n_p}| < f^*$.

■ 3.2.3 A heuristic for k_F

As mentioned, we accelerate Eqn. 3.15 by limiting the summation to selected eigenindices k_F. We identify this set of indices by observing that when a local perturbation is made in a network, some Laplacian eigenpairs are impacted more than others. Efficient computation of the perturbed spectrum should ignore unimpacted eigenpairs, and we can discriminate between eigenpairs further by considering only those whose contributions to trapping time at n_t change substantially upon the perturbation, that is $|\Delta\tau_{n_t}^k| > \tau_0^*$. In order to effectively classify eigenpairs into a 'free' class, k_F and a 'locked' class, k_L, we need a heuristic for $|\tilde{\Delta}\tau_{n_t}^k|$ that avoids direct eigendecomposition. Our choice is

$$|\tilde{\Delta}\bar{\tau}_{n_t}^k| = \tilde{\bar{\tau}}_{n_t}^k(\mathcal{H}_p) - \bar{\tau}_{n_t}^k(\mathcal{H}) \tag{3.17}$$

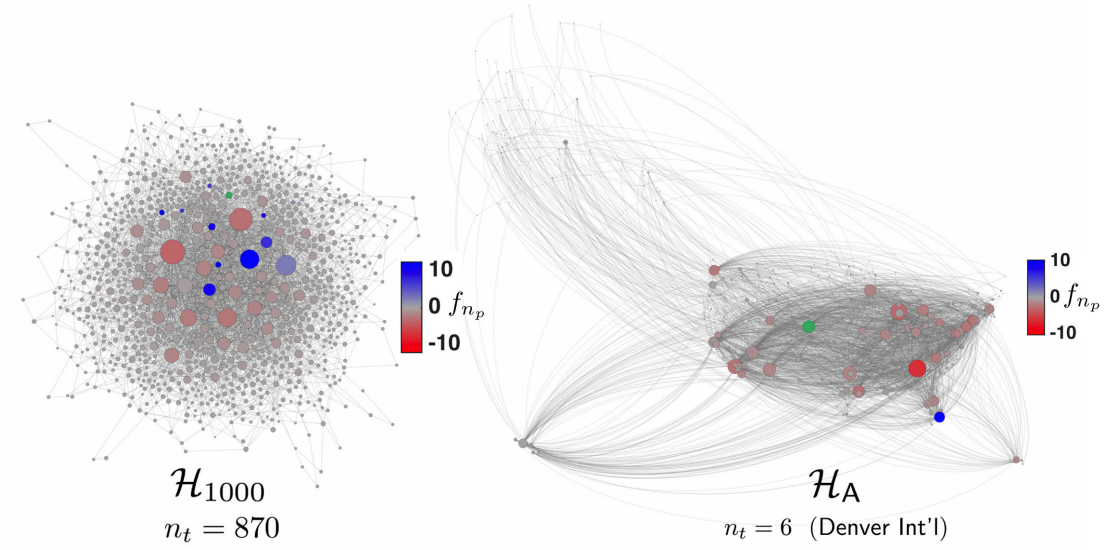


Figure 3.4: F-scores for $\mathcal{H}_{1000}$ and $\mathcal{H}_{\mathsf{A}}$. A representative target node (green) for each network was selected and f-scores for all other nodes were computed and shown by colorscale. Node widths reflect total edge weight including self-loops for each node, and the spatial arrangement results from the Gephi Force Atlas algorithm[87] (left), or geographical location (right). Edge weights are not depicted. (Right) Most major airports are densely connected throughout the network and by their presence retard average transit times of a random walk to n_t, Denver International Airport. One major airport, Miami's, however, has a substantial positive f-score, meaning average MFPT's to Denver would in fact drop by 10.3% if MIA were removed from the network[179]. F-score ranges were -3.8 to 12.3 ($\mathcal{H}_{1000}$) and -8.0 to 10.3 ($\mathcal{H}_{\mathsf{A}}$). Visualizations for remaining test networks are in Fig. 3.8.

where

$$\tilde{\bar{\tau}}^k_{n_t} = \frac{1}{\tilde{\lambda}_k}\left(\frac{N}{N-1}\right)\left(s_p \tilde{u}^2_{n_t k} - \mathbf{s}_p^T \tilde{\mathbf{u}}_k \tilde{u}_{n_t,k}\right). \tag{3.18}$$

Vector $\tilde{\mathbf{u}}_k$ is a column of $\tilde{\mathbf{U}}$, itself equal to $\mathbf{U}$ with the exception of rows corresponding to the perturbed node n_p and its neighbors G_{n_p}. Specifically,

$$\tilde{\mathbf{U}}_{[\{n_p \cup n_g\},1:N]} = \mathbf{U}_{[\{n_p \cup n_g\},1:N]} - 2\left(\mathbf{L}_{p[\{n_p \cup n_g\},1:N]} * \mathbf{U} - \mathbf{U}_{[\{n_p \cup n_g\},1:N]} * \mathbf{I}\vec{\tilde{\lambda}}\right). \tag{3.19}$$

Changes in the elements of the approximation vectors $\tilde{\mathbf{U}}$ correspond to the gradient of the Rayleigh quotient[51] evaluated only at n_p and G_{n_p} since the gradient at all other nodes will be negligible. Tildes over returned values emphasize that Eqns. 3.18 and 3.19 are not exact and simply provide a convenient heuristic for selecting the initial free eigenindices:

$$k_F = \underset{k}{\mathsf{find}}\left(|\tilde{\Delta}\bar{\tau}^k_{n_t}| > \bar{\tau}^*\right). \tag{3.20}$$

Intuitively, Eqn. 3.19 tells us about the impact of the perturbation given (i) the graph $\mathcal{H}$ and (ii) the perturbed node n_p, whereas Eqn. 3.18 tells us about the impact of the perturbation given

all three involved entities: graph $\mathcal{H}$, node n_p, *and* target node n_t. Together, the expressions reveal which k eigenindices give rise to large predicted $|\Delta\bar{\tau}^k|$ values. We only employ this routine at $\mathsf{iter} = 0$, before vectors $\mathbf{U}_{k_F}$ have been updated with linear estimate Eqn. 3.15. Subsequently, provided with $\tilde{\mathbf{U}}_{\mathsf{iter}>0}$, we can utilize the observed changes in trapping time contributions $|\bar{\tau}^k_{n_t}|$ to select k_F for the next iteration.

■ 3.2.4 Degenerate eigenvalues

The denominator in Eqn. 3.15 suggests that problems with the update expressions will occur if our network contains degenerate, or duplicate, eigenvalues. Even for fully connected, non-regular networks, this issue can arise if there are some regularities in the graph structure. To illustrate, Fig. 3.5 shows the gap between eigenvalue pairs for the six networks we consider. It is apparent that networks $\mathcal{H}_{\mathsf{YST}}$ and $\mathcal{H}_{\mathsf{UC}}$ have degenerate regions of their spectra that will cause 'blowups' if we use standard perturbation theory. We can manage this issue with only a small efficiency penalty.

Following the notation in Ref. 175 let β_d and μ_d be the respective eigenvectors and eigenvalues of the M by M matrix $\mathbf{U}_d^T\mathbf{L}_p\mathbf{U}_d$ where $\mathbf{U}_d$ is the $N \times M < N$ column matrix of the degenerate modes in $\mathbf{U}$. Then $\tilde{\mathbf{u}}_{k\in d} = \sum_{j=d}\beta_{jk}\mathbf{u}_j$ and $\tilde{\lambda}_{k\in d} = \mu_d$. The degenerate eigenvector $\tilde{\mathbf{u}}_{k\in d}$ is further updated with Eqn. 3.15 where eigenindices d are excluded from the sum, thereby avoiding a vanishing denominator. In summary, these degenerate eigenpairs are solved using brute force eigensolvers while the rest of the spectrum can take advantage of the perturbation strategy.

■ 3.2.5 Methods Summary

Our protocol works by perturbing node n_p by a small amount $\epsilon \sim 10e{-}4$ and iteratively correcting eigenvectors $\mathbf{U}$ from the intact graph $\mathcal{H}$ to approximate the basis of the altered graph, $\mathcal{H}_p$. However, we choose to update only vectors that make significant ($> \tau^*$) contribution to the trapping time, $\bar{\tau}_{n_t}$, given the user-chosen target node n_t. That is, we choose to permit small non-orthogonalities in the updated spectrum as long as the estimated f-score $\tilde{f}_{n_p}$ stabilizes. Specifically, at each iteration the set of vectors that gets updated is denoted $k_F \subset \{2 \dots N\}$ (subscript F for 'free'), and this set is guaranteed to be non-increasing with each iteration (see Section 3.3.1). Those eigenvectors that are already converged are called 'locked' and denoted k_L such that $k_L \cap k_F = \emptyset$. Moreover, when $\mathsf{iter} = 0$, most eigenvector elements do not change, so we

Fast f-score estimation

INPUT: Laplacians $\mathbf{L}$ and $\mathbf{L}_p$, target node index n_t, and perturbed node indices N_p

OUTPUT: $\tilde{f}\,(n_p, n_t)\ \ \forall n_p \in \{N_p\}$.

$(\mathbf{U}_0, \lambda) \leftarrow \mathsf{eig}\,(\mathbf{L})$ ▷ Direct eigendecomposition

$\mathbf{U} \leftarrow \mathbf{U}_0$

$\bar{\tau}^k_{n_t} \leftarrow \left(\frac{N}{N-1}\right)\left(\frac{su^2_{kn_t} - \left(\mathbf{s}^T\mathbf{u}_k\right)u_{kn_t}}{\lambda_k}\right)\ \ \forall k \neq 1$

Predict free/locked modes, k_F, k_L, by estimating $\Delta\bar{\tau}^k_{n_t}$

for $n_p \in N_p$ **do**

  $\mathbf{U}_{[n_p \cup n_g, 2:N]} \leftarrow \mathbf{U}_{[n_p \cup n_g, 2:N]} - \nabla r\left(\mathbf{U}_{[n_p \cup n_g, 2:N]}\right)$ ▷ see Eqn. 3.19

  $\mathbf{U}_k = \mathbf{U}_k / \|\mathbf{U}_k\|$ ▷ Normalize all columns of $\mathbf{U}$

  $\tilde{\Delta}\bar{\tau}^k_{n_t} \leftarrow \left(\frac{N}{N-1}\right)\left(\frac{s_p u^2_{kn_t} - \left(\mathbf{s}_p^T\mathbf{u}_k\right)u_{kn_t}}{\lambda_k}\right) - \bar{\tau}^k_{n_t},\ \ \forall k \neq 1$

  $k_F \leftarrow \underset{k}{\mathsf{find}}\left(|\tilde{\Delta}\bar{\tau}^k_{n_t}| > \bar{\tau}^*\right),\ k_L \leftarrow \{2 \ldots N\} \setminus k_F$ ▷ Select free/locked eigenpairs

Estimate perturbed eigenvalues

  Select $\epsilon \sim 10^{-4}$

  $\mathbf{U} \leftarrow \mathbf{U}_0$

  $\tilde{\lambda}_k \leftarrow k + \epsilon * \left(\left(\mathbf{U}_k.^2\right)^T \mathbf{L}_{n_p} - 2\lambda_k u^2_{kk}\right)\ \ \forall k \neq 1$

  Generate matrix of update weights: $\Lambda_{ij} = \left(\tilde{\lambda}_i - \tilde{\lambda}_j\right)^{-1},\ \Lambda_{ii} = 0,\ i, j \in \{2 \ldots N\}$

Update U iteratively until $\tilde{f}(n_p, n_t)$ converges

  $\mathsf{iter} \leftarrow 0$

  Store free node indices: $n_F = \{n_p \cup n_g\}$ ▷ only n_p and neighborhood eligible for update

  while $\mathsf{converged} == 0$ **do** ▷ Begin iteration for $\tilde{f}_{n_p}$

    $\Delta\mathbf{U}_{[1:N,k_F]} \leftarrow \mathbf{U}_{[1:N,k_F]}\left\{\mathbf{U}^T_{[n_F,k_F]}\left(\mathbf{L}_{p[n_F,1:N]}\mathbf{U}_{[1:N,k_F]} - \mathbf{U}_{[n_F,k_F]} * \mathbf{I}\tilde{\lambda}_{k_F}\right) .* \Lambda_{[k_F,k_F]}\right\}$ ▷ see Eqn. 3.15

    $\tilde{\mathbf{U}} \leftarrow \mathbf{U} + \Delta\mathbf{U}$

    $\tilde{\bar{\tau}}^k_{n_t} \leftarrow \left(\frac{N}{N-1}\right)\left(\frac{s_p\tilde{u}^2_{kn_t} - \left(\mathbf{s}_p^T\tilde{\mathbf{u}}_k\right)\tilde{u}_{kn_t}}{\tilde{\lambda}_k}\right),\ \forall k \in k_F$ ▷ Compute updated $\tilde{\bar{\tau}}^k_{n_t}$

    $\tilde{f}_{\mathsf{iter}}(n_p, n_t) \leftarrow (1/\epsilon) * \frac{\sum_{k=2}^N \tilde{\bar{\tau}}^k - \sum_{k=2}^N \bar{\tau}^k}{\sum_{k=2}^N \bar{\tau}^k}$ ▷ Estimate new f_{n_p}

    $\mathsf{converged} \leftarrow \left|\tilde{f}_{\mathsf{iter}} - \tilde{f}_{\mathsf{iter}-1}\right| / \left|\tilde{f}_{\mathsf{iter}-1}\right| < f^*$

    if !converged **then**

      $k_F \leftarrow \mathsf{find}\left(|\tilde{\Delta}\bar{\tau}^{k_F}_{n_t}| > \bar{\tau}^*\right)$

      $n_F \leftarrow \{1 \ldots N\}$ ▷ All nodes now eligible for update

      $\mathbf{U} \leftarrow \tilde{\mathbf{U}}$

      $\mathsf{iter} \leftarrow \mathsf{iter} + 1$

    end if

  end while

end for

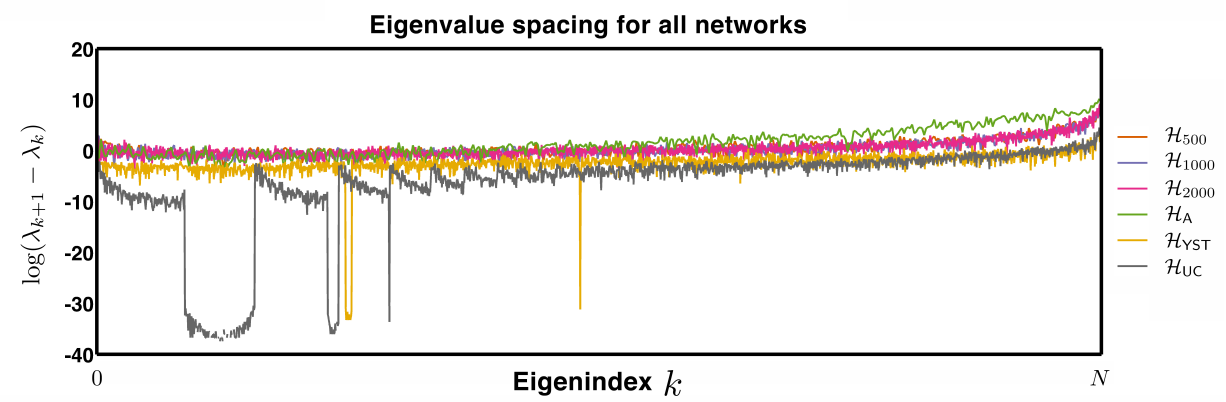


Figure 3.5: Eigenvalue spacing for all test networks. Eigenvalues $\lambda_k, k > 0$ for intact networks are sorted ascendingly and their differences plotted as $\log(\lambda_{k+1} - \lambda_k)$. Degenerate eigenvalues are present in $\mathcal{H}_{\mathsf{YST}}$ (yellow) and $\mathcal{H}_{\mathsf{UC}}$ (gray). In tests we selected a 'degenerate threshold' $\Delta\lambda^* = 0.00001$. For comparison $\log(0.00001) \approx -11.5$.

can restrict the update to elements corresponding to n_F, that is, 'free' elements row-wise of the current eigenvectors $\mathbf{U}$. In subsequent iterations, when $\mathsf{iter} > 0$, $n_F = \{1 \dots N\}$. Boxed pseudocode is given in **Fast f-score estimation**. All computations were performed with Matlab[84] on an eight core desktop computer with Intel Core i7-4770 CPUs @ 3.40GHz, 256K L2 cache; the operating system was Ubuntu 14.04.2. Network visualizations were produced with Gephi[87] as in Sec. 2.2.7.

■ 3.3 Numerical Results

We tested our algorithm on six small to medium networks, both synthesized and naturally occurring (Table 3.2). Symmetric synthesized networks $\mathcal{H}_{500}$, $\mathcal{H}_{1000}$, and $\mathcal{H}_{2000}$ were first generated with Complex Networks[96] and then self and non-self weights were assigned randomly to existing edges. Visualizations for $\mathcal{H}_{1000}$ and $\mathcal{H}_{\mathsf{A}}$ are provided in Fig. 3.4. To illustrate the relationship between (i) the free eigenspectrum k_F and (ii) f-score predictions as the algorithm progresses for the synthetic networks, we randomly chose a n_t in each synthetic network and charted algorithm execution for multiple representative nodes $\{n_p\}$ (Fig. 3.6). Specifically, convergence properties for one example node n_p are shown in red while selected n_p generally are denoted with black ($n_p \in G_{n_t}$) or gray ($n_p \in \bar{G}_{n_t}$) curves (panels B and C). Iterations for a single representative perturbed node in $\mathcal{H}_{500}$ is shown in Fig. 3.9. Qualitatively, convergence behavior was shared among the three networks. We observed that the size of the free eigenspectrum $|k_F|$ decreases

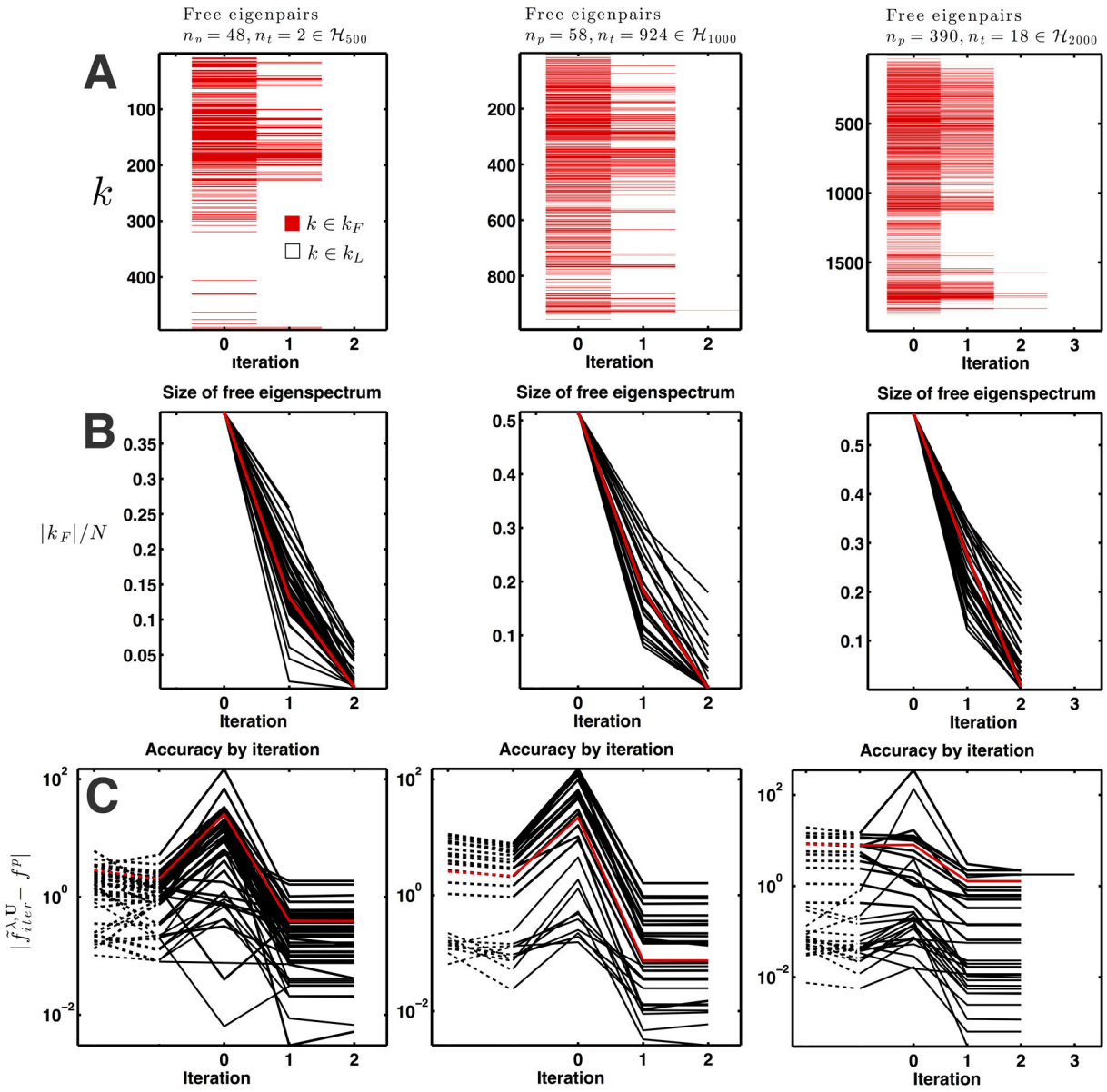


Figure 3.6: The number of free eigenindices $|k_F|$ decreases each iteration. (A) Free eigenindices per iteration are shown; $\mathcal{H}_{500}$ (left col.), $\mathcal{H}_{1000}$ (center col.), $\mathcal{H}_{2000}$ (right col.). (B) Convergence of k_F shown for all neighbors of selected $n_t \in \mathcal{H}$, values for representative n_p from row (A) shown in red. Vertical axis gives proportion of total spectrum. (C) Absolute accuracy of $\tilde{f}$ at each iteration. Dashed lines show accuracy change with only eigenvalue update $\tilde{\lambda}$. Red curves as in (B). Algorithm terminates when $\tilde{f}$ changes by less than f^*.

quasi-linearly each iteration (Fig. 3.6B) given a selection threshold of $\tau^* = 0.995$ and also that $\tilde{f}$ convergence is attained within three iterations for $\mathcal{H}_{500}$ and $\mathcal{H}_{1000}$ and four iterations for $\mathcal{H}_{2000}$ (Fig. 3.6C).

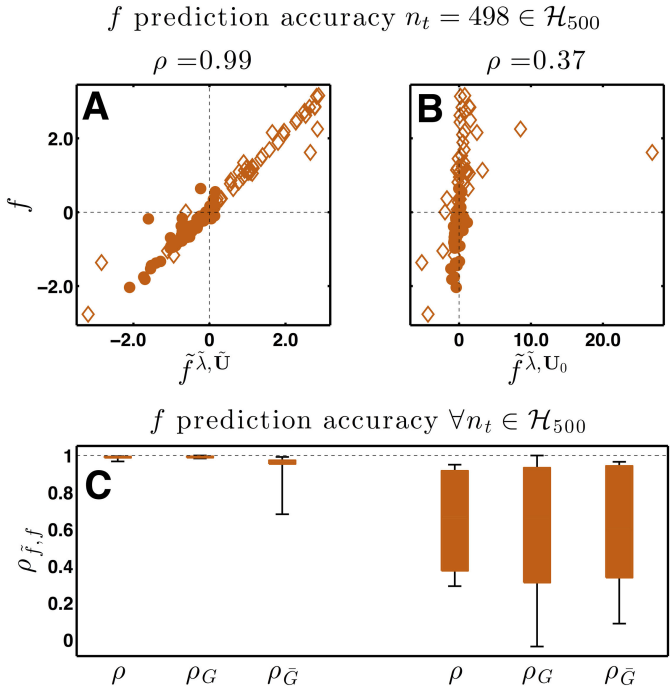


Figure 3.7: Both perturbed eigenvalues and eigenvectors must be estimated for accurate f-score prediction. (A) F-score scatter plot for representative target node $n_t = 492$ in graph $\mathcal{H}_{500}$: vertical axis is the exact f-score (f), horizontal axis is the predicted f-score ($\tilde{f}$) for all nodes $n_p \neq n_t \in \mathcal{H}_{500}$. Diamonds denote neighbors of n_t (G_{n_t}), dots foreigners ($n_{\bar{G}_{n_t}}$). (B) Estimated f-scores $\tilde{f}$ computed from intact, unperturbed eigenvectors $\mathbf{U}_0$ and estimated eigenvalues ($\tilde{\lambda}$); axes as in (A). (C) The distribution of prediction accuracy for all target nodes in $\mathcal{H}_{500}$; accuracy over only neighbors of each n_t is labeled ρ_G, accuracy for foreigners of each n_t is labeled $\rho_{\bar{G}}$, and correlation over all perturbed nodes is labeled as ρ. Box limits indicate upper and lower quartiles; whiskers show complete data range.

The free eigenpairs were distributed throughout the spectra, consistent with our claim that changes in trapping time cannot be fully recovered by extreme eigenpairs alone. Some pairs remain free through several iterations, but only free eigenpairs can remain free. Once locked an eigenpair will not be updated further. Even though $|k_F|$ apparently decreases, it is not the case that estimated f-scores likewise converge monotonically toward the true f_{n_p}, and in fact they often get worse during the first iteration, iter $= 0$ (Fig. 3.6C). That is, a single iteration of eigenvector update Eqn. 3.15 often produces worse f predictions than scores estimated with only approximated eigenvalues (D, dashed lines). This illustrates that transit/trapping times are many-to-one indirect functions of the spectrum; the objective formally being minimized in (3.15) and pseudocode line 16 is not $\tilde{f}$ but the gradient of the Rayleigh quotient (at nodes n_F). Consequently, as free eigenpairs adjust to the graph structure in $\mathcal{H}_p$ our estimates $\tilde{f}$ can temporarily

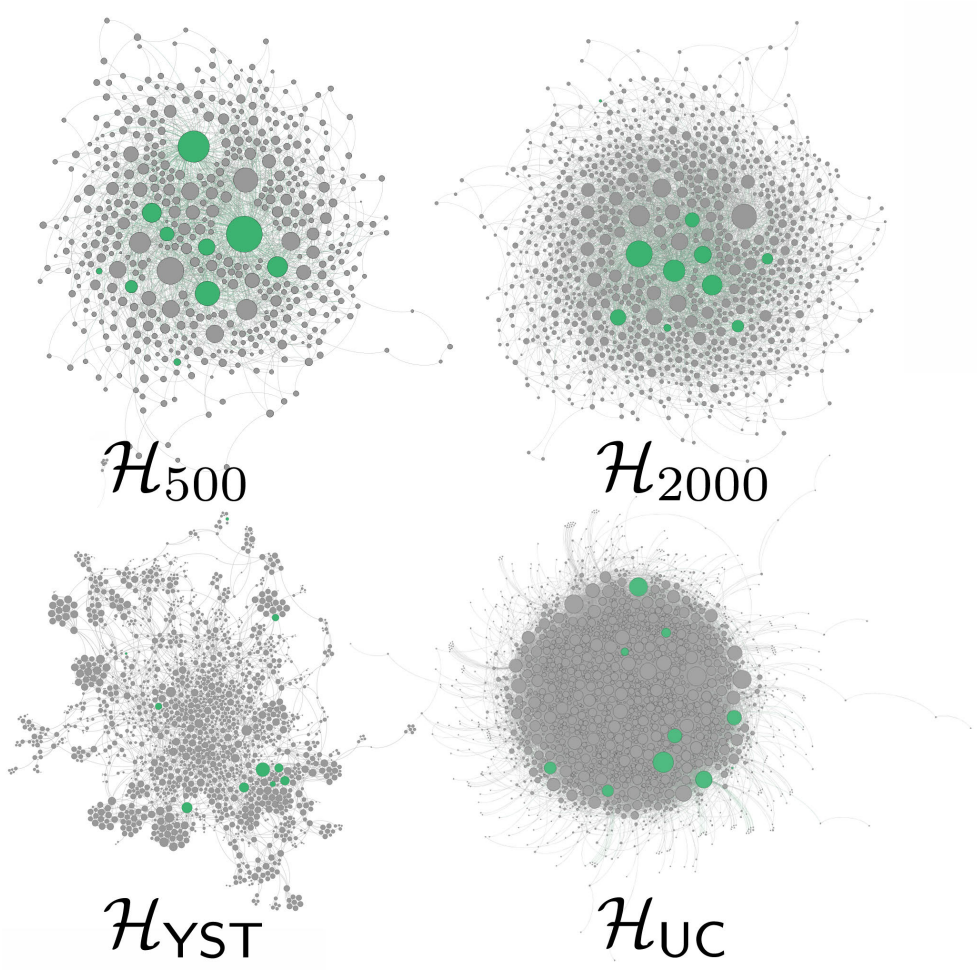


Figure 3.8: Network visualizations for $\mathcal{H}_{500}$, $\mathcal{H}_{2000}$, $\mathcal{H}_{\mathsf{YST}}$, and $\mathcal{H}_{\mathsf{UC}}$. All nodes are colored gray except those target nodes n_t which were tested for accuracy and efficiency in Fig. 3.10, here colored green. The visualization procedure followed the protocol in Sec. 2.2.7.

suffer. However, as k_F diminishes and trapping time contributions $\bar{\tau}^k$ stabilize the predicted f-score $\tilde{f}$ generally approaches the true value. A final prediction error $|f - \tilde{f}_{\mathsf{iter}>0}|$ worse than starting prediction error $|f - \tilde{f}_{\mathsf{iter}=0}|$ suggests either (i) a failed k_F selection heuristic (pseudocode lines 4-8) or (ii) overly permissive convergence thresholds f^* and τ^*.

When altering a physical network such that n_t trapping times are impacted, f-score accuracy rather than eigenvector convergence is the more relevant statistic. While f-scores are often close to zero for low-degree nodes distant from n_t, nodes that are first and second degree neighbors of n_t often have appreciable f_{n_p} values. Figure 3.7 shows predicted and exact f-scores for neighbor and selected non-neighbor nodes of $n_t = 2 \in \mathcal{H}_{500}$. In the upper panels direct neighbors of n_t

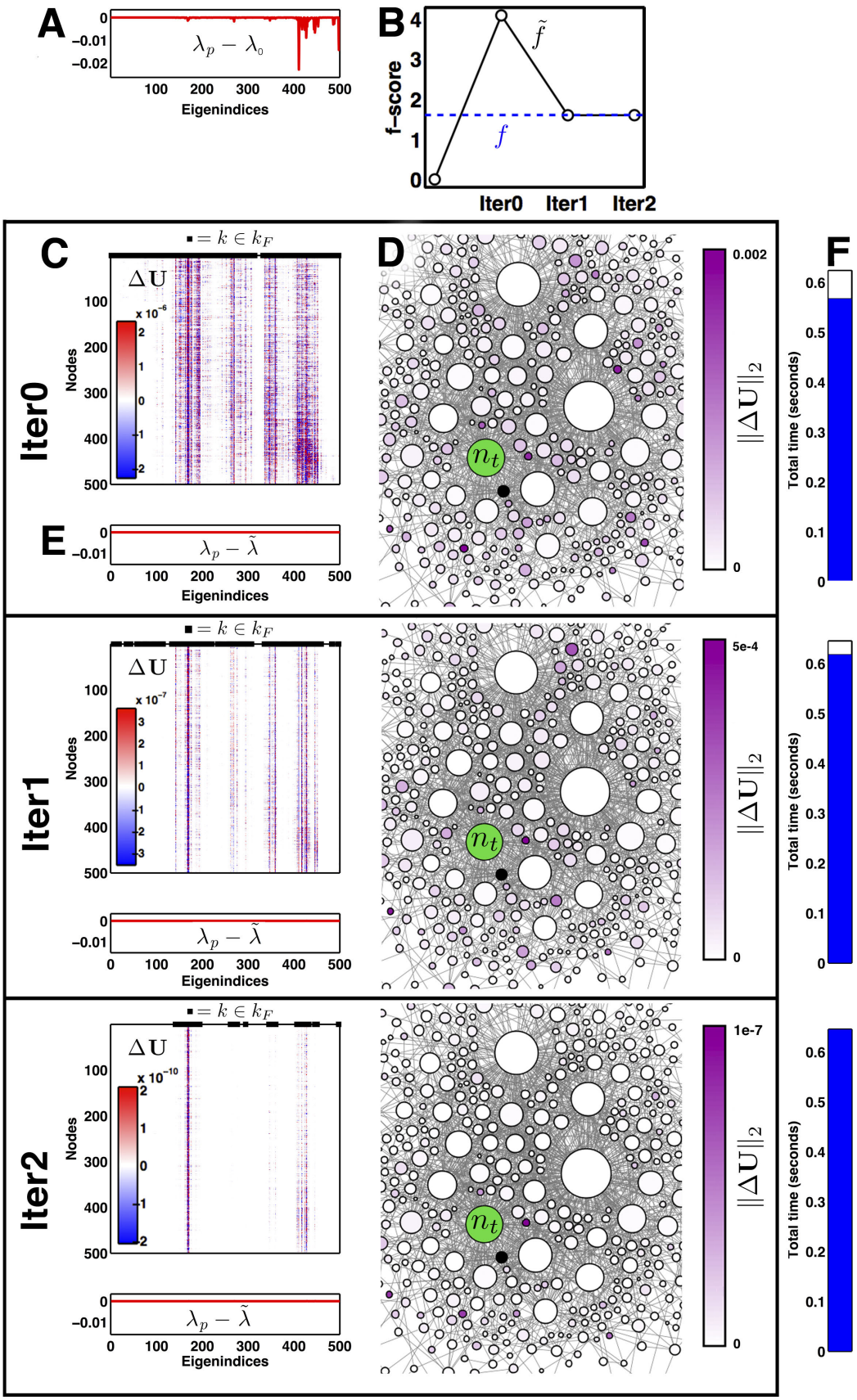


Figure 3.9: (Continued on following page)

Figure 3.9: Visualization for $\tilde{f}(n_t = 498, n_p = 438, \mathcal{H}_{500})$ estimation over three iterations. (B) F-score estimate $\tilde{f}$, black. True value, f, shown as dashed blue line. (C) Eigenvector update $\Delta\mathbf{U}$ (Eqn. 3.15, Algorithm line 16); rows are nodes (n), columns are eigenindices (k). Black squares indicate free eigenindices k_F (Eqn. 3.20). (D) Magnitudes of eigenvector update displayed at each node n, $\|\Delta\mathbf{U}_{[n,1:N]}\|_2$. Only a subset of $\mathcal{H}_{500}$ is shown to illustrate changes in relative update magnitude. Target node $n_t = 498$, green; perturbed node $n_p = 438$, black. The magnitude of the updates decreases approximately two orders of magnitude each iteration. (E) Error of predicted eigenvalues, $\tilde{\lambda} - \lambda_p$. (F) Aggregate runtime. (A) Pre-procedure eigenvalue error, $\lambda_p - \lambda_0$.

are designated with diamonds while non-neighbors are filled circles. The upper panels illustrate that accurate λ_p values cannot produce good $\tilde{f}$ values if $\mathbf{U}_0$ is left uncorrected. Throughout we will use this linear correlation as an accuracy measure: $\rho_{G_{n_t}}$ for neighbors of n_t, $\rho_{\bar{G}_{n_t}}$ for non-neighbors of n_t, and ρ for the combined set of both neighbors and non-neighbors of n_t. Correlation accuracy was tested by individually treating every node $n \in \mathcal{H}_{500}$ as a target node n_t, and for the other networks test n_t nodes were selected uniformly across the degree distribution (Figs. 3.7C and 3.10A).

3.3.1 Algorithm thresholds

There are two user selected parameters that control the trade-off between speed and accuracy within the procedure. The first, τ^*_{iter} controls whether a given eigenvector $\mathbf{U}_{k \in k_F}$ (1) remains in k_F after an iterative update or (2) gets moved into the set of locked eigenvectors k_L. Presently, $\bar{\tau}^*_{\mathsf{iter}>0}$ is set so that k_F after each iteration includes those eigenvectors that contribute 99.5% percent of the total change in τ. Iteration histories of $|k_F|$ are shown in Figs. 3.6 and 3.9.

The second user parameter determines when the algorithm terminates . Once $\tilde{f}_{n_p}$ stabilizes, changing less than f^* per iteration, the algorithm terminates. A threshold of $f^* = 0.01$ generally produces good accuracy correlations (Fig. 3.7).

3.4 Discussion

Several theoretical challenges remain for elucidating the behavior of complex networks. A first might be to understand the exact physical or dynamical significance of their graph spectra[180] in different modeling contexts. One general interpretation in this direction is from Van Mieghem: squared elements of each row vector $\mathbf{U}_{[n_p,1:N]}$ quantify the impact of removing node n_p from the network at eigenfrequencies $1 \ldots N$[131]. At least for Markov-type networks that evolve temporally,

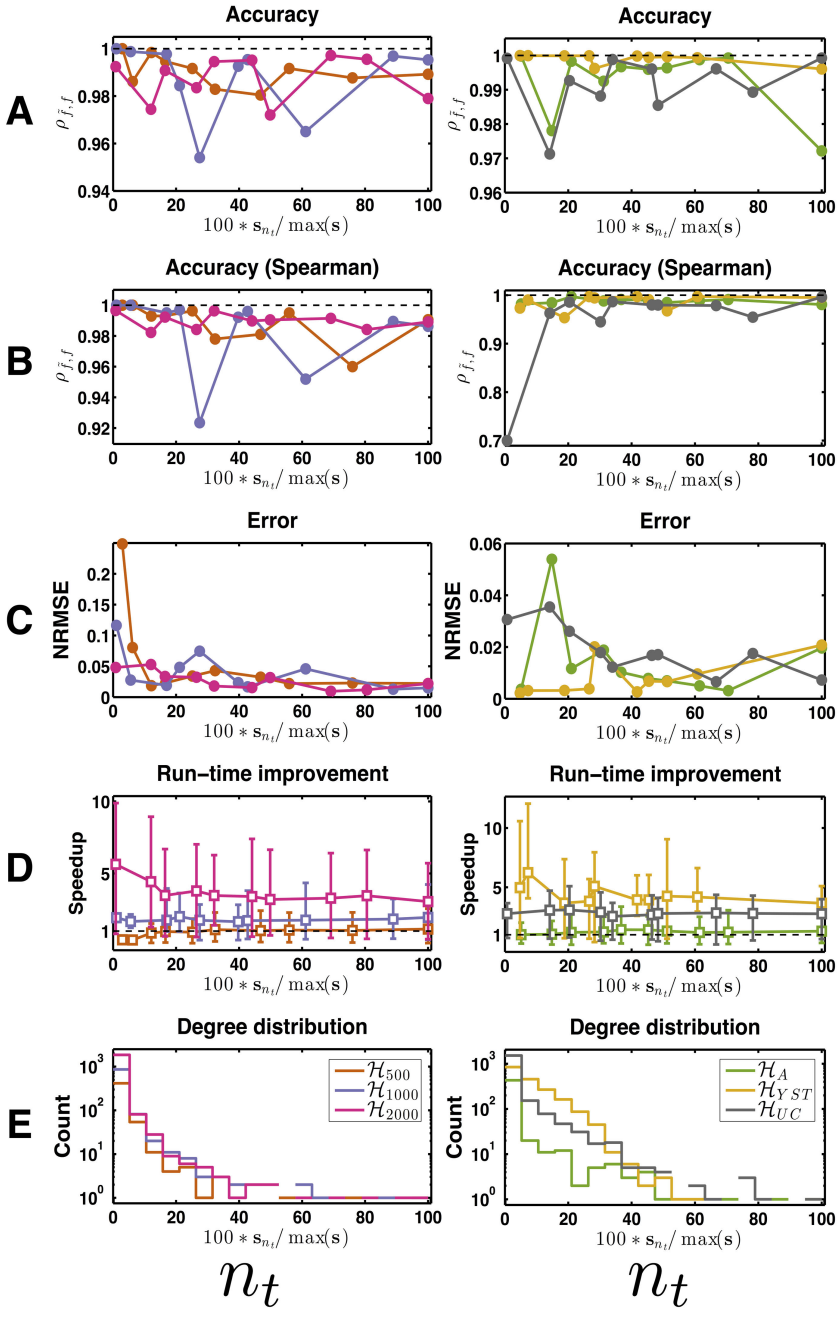


Figure 3.10: F-score accuracy and efficiency. Results for synthetic (left column) and real (right column) networks. Horizontal axes in all panels indicate the weighted degree of n_t as a percentage of the maximally-weighted node, $\max_{n_t} \mathbf{s}_{n_t \in \mathcal{H}}$. Target nodes n_t were selected by binning all nodes into 20 equal bins according to degree and then randomly selecting 10 n_t nodes equally spaced across nonempty bins. (A) Accuracy as determined by correlation of predicted f-scores, $\tilde{f}$, with ground truth f-scores, f. (B) Normalized root mean squared error. (C) Run-time improvement against direct method, where whiskers show maximum and minimum values. (D) Weighted degree distributions. Colors indicate network selection; see Table 3.2 and Fig. 3.11 for results summary by network. Average runtimes for a single n_p, in seconds, were: 0.08 ($\mathcal{H}_{500}$), 0.64 ($\mathcal{H}_{1000}$), 3.48 ($\mathcal{H}_{2000}$), 0.47 ($\mathcal{H}_{\mathsf{A}}$), 1.06 ($\mathcal{H}_{\mathsf{YST}}$), and 1.43 ($\mathcal{H}_{\mathsf{UC}}$).

Table 3.2: Accuracy and efficiency of predicted f-scores. Algorithm accuracy evaluated with correlation ρ, Spearman rank correlation ρ_s, and root mean squared error normalized by the range of exact scores, $\overline{\mathsf{NRMSE}}$. As controls we also show accuracies for f-score estimates derived without eigenvector updates, $\overline{\mathsf{NRMSE}}_{\tilde{\lambda},\mathbf{U}_0}$ and those derived from the intact spectrum, $\overline{\mathsf{NRMSE}}_{\lambda_0,\mathbf{U}_0}$. When present the overline indicates weighted average over all tested n_t's, i.e., all values in Fig. 3.10B. The NRMSE values are also compared in Fig. 3.11. Standard deviation of the accuracy values are provided in Table A.1. Some n_p nodes are tested more than once with different target nodes n_t, so total n_p count can exceed the network size. The lower ten rows show results for the protein folding networks discussed in Chapter 2. Protein folding networks are ordered according to increasing N = Total n_p+1 in order to show the general trend in efficiency.

	Total n_t	Total n_p	ρ	ρ_s	$\overline{\mathsf{NRMSE}}_{\tilde{\lambda},\tilde{\mathbf{U}}}$	$\overline{\mathsf{NRMSE}}_{\tilde{\lambda},\mathbf{U}_0}$	$\overline{\mathsf{NRMSE}}_{\lambda_0,\mathbf{U}_0}$	Avg. speedup
$\mathcal{H}_{500}$	10	607	0.99	0.98	0.027	0.181	0.192	1.05
$\mathcal{H}_{1000}$	10	837	0.99	0.98	0.026	0.173	0.200	1.82
$\mathcal{H}_{2000}$	10	1880	0.99	0.99	0.021	0.108	0.144	3.38
$\mathcal{H}_{\mathsf{A}}$	10	880	0.99	0.99	0.012	0.102	0.109	1.28
$\mathcal{H}_{\mathsf{YST}}$	10	550	1.00	0.99	0.009	0.174	0.234	4.27
$\mathcal{H}_{\mathsf{UC}}$	10	1117	0.99	0.97	0.016	0.096	0.127	2.83
					$\mathsf{NRMSE}_{\tilde{\lambda},\tilde{\mathbf{U}}}$	$\mathsf{NRMSE}_{\tilde{\lambda},\mathbf{U}_0}$	$\mathsf{NRMSE}_{\lambda_0,\mathbf{U}_0}$	
$\mathcal{H}_{\mathsf{PB}}$	1	167	1.00	1.00	0.007	0.036	0.140	0.48
$\mathcal{H}_{\mathsf{VHP}}$	1	207	1.00	1.00	0.012	0.045	0.238	0.60
$\mathcal{H}_{\mathsf{TRP}}$	1	387	1.00	1.00	0.004	0.014	0.157	0.82
$\mathcal{H}_{\mathsf{HMDM}}$	1	517	1.00	0.99	0.009	0.130	0.120	0.97
$\mathcal{H}_{\mathsf{BBL}}$	1	758	1.00	1.00	0.005	0.034	0.084	1.23
$\mathcal{H}_{\mathsf{BBA}}$	1	905	0.99	1.00	0.017	0.057	0.108	1.04
$\mathcal{H}_{\mathsf{LAMDA}}$	1	1181	1.00	1.00	0.002	0.051	0.074	1.72
$\mathcal{H}_{\mathsf{A3D}}$	1	1346	1.00	1.00	0.001	0.039	0.099	1.46
$\mathcal{H}_{\mathsf{WW}}$	1	2067	1.00	0.99	0.001	0.059	0.076	1.02
$\mathcal{H}_{\mathsf{PG}}$	1	2248	1.00	1.00	0.001	0.021	0.049	1.13

we think this interpretation can be better reified by invoking mean first passage times. Indeed, Eqn. 3.6 formulates this same element-squared row vector into a convenient quantity $\bar{\tau}_{n_t}$ where we do not need to inspect individual eigenfrequencies in order to assess the topological importance of n_p. That is, individual elements of $\mathbf{U}_{[n_p,1:N]}$ may increase or decrease upon network perturbation, but we can always interpret an f-score to signify that node n_p helps ($f_{n_p} > 0$) or hinders ($f_{n_p} < 0$) graph transitions to n_t. Interestingly, these small transit time changes manifest themselves in various and discontiguous regions of the Laplacian spectrum (Fig. 3.6), precluding use of many traditional eigensolvers or methods limited to extreme or connected eigenpairs [181–183].

Our primary focus has been to show that, algorithmically, careful selection of eigenpairs k_F can produce a less expensive approximation $\tilde{f}(n_p, n_t)$ that avoids the fundamental matrix $\mathbf{Z}$. This selection cannot be made by comparing the intact and perturbed spectra (since it would require directly computing the latter), but we can guess that nodes with large Rayleigh

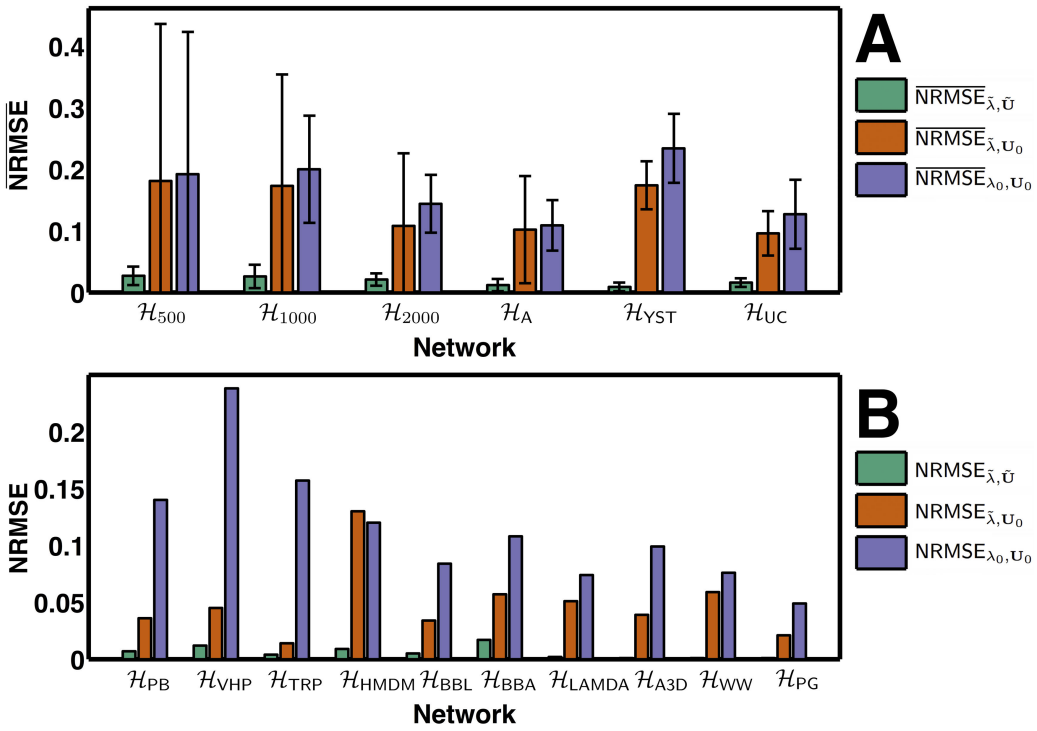


Figure 3.11: Comparison of f-score accuracy. Normalized root mean squared error for f-scores generated with (1) the intact spectrum, $\overline{\text{NRMSE}}_{\lambda_0,\mathbf{U}_0}$ (violet), (2) updated eigenvalues but intact eigenvectors, $\overline{\text{NRMSE}}_{\tilde{\lambda},\mathbf{U}_0}$ (orange), and (3) both updated eigenvalues and eigenvectors, $\overline{\text{NRMSE}}_{\tilde{\lambda},\tilde{\mathbf{U}}}$ (green). (A) Six test networks. Error bars indicate $\pm\sigma$. The overline in $\overline{\text{NRMSD}}$ indicates weighted average over all tested n_t's, i.e., all values in Fig. 3.10B. (B) Protein folding networks from Chap. 2. Target node n_t for each network is the entire native ensemble from Fig. 2.6.

quotient gradients (Alg. line 5) reveal those eigenpairs which will move substantially upon node perturbation (k_F) and which will remain approximately stationary (k_L). Iterative application of first-order perturbation theory to both $\tilde{\lambda}$ and $\tilde{\mathbf{U}}$ for only this selected subspace then reveals an approximate updated spectrum (and updated $\bar{\tau}_{n_t}$) faster than dense eigendecomposition. At least with our tested system specifications, networks larger than $N = 500$ generally experience a speedup; this threshold could relate to L2 cache availability[184]. It is important to emphasize that MFPT's can be variously formulated and other approximations for τ and by extension f_{n_p} that do not invoke spectral theory could also be compared[185–187].

Because f-scores are usually linear functions of the perturbation magnitude $\epsilon \in [0, 1]$, it is not necessary to completely remove node n_p from the graph and problematically decrement the rank of $\mathbf{U}$ in order to estimate f_{n_p}. Instead, we chose a very small ϵ so that the eigenvector shifts are small and linear estimates are accurate. This approach has the additional advantage that nodes

are never disconnected from the primary graph component when a bottleneck node is perturbed. In these situations the f-score cannot fairly be viewed as the change in transit times were n_p to be removed since some paths to n_t would become impossible. The interpretation in these cases should be that f_{n_p} represents changes in transit times were n_p to be almost completely removed from the network.

If a second network science challenge is to understand dynamically what happens to networks when they are altered[158,188,189], then f-scores perhaps contribute here as well; it is often behavior at some target node that is more important than global properties across a whole graph, and f-score values reveal exactly that. Though many networks in the biological and social sciences surpass in size those considered here, simple coarse-graining methods[152] can be applied so that the resultant network is amenable to our method.

Chapter 4

Auto-regressive models of protein motions

Molecular dynamics (MD) simulations have dramatically improved the atomistic understanding of protein motions, energetics, and function. These growing data sets have necessitated a corresponding emphasis on trajectory analysis methods for characterizing simulation data, particularly since functional protein motions and transitions are often rare and/or intricate events. Observing that such events give rise to long-tailed spatial distributions, we recently developed a higher-order statistics based dimensionality reduction method, called quasi-anharmonic analysis (QAA), for identifying biophysically-relevant reaction coordinates and substates within MD simulations. Further characterization of conformation space should consider the temporal dynamics specific to each identified substate. Our model uses hierarchical clustering to learn energetically coherent substates and dynamic modes of motion from a 0.5μs ubiqutin simulation. Autoregressive (AR) modeling within and between states enables a compact and generative description of the conformational landscape as it relates to functional transitions between binding poses. Lacking a predictive component, QAA is extended here within a general autoregressive (AR) model appreciative of the trajectory's temporal dependencies and the specific, local dynamics accessible to a protein within identified energy wells. These metastable states and their transition rates are extracted within a QAA-derived subspace using hierarchical Markov clustering to provide parameter sets for the second-order AR model. We show the learned model can be extrapolated to synthesize trajectories of arbitrary length.

■ 4.1 Introduction

Conformational changes in proteins constitute the underlying behavior of cellular regulation. As part of regulating cellular homeostasis, proteins perform a number of functions through native fluctuations at multiple length- and timescales. A variety of experimental techniques have illuminated the linkage between protein dynamics and function; however, resolving the precise

spatio-temporal relationships in protein motions which confer biological function remains a long-standing challenge in protein biochemistry[190].

Governing the protein's rich conformational space is a high-dimensional energy landscape with multiple *hills* and *valleys*[9,10]. To characterize this energy surface, theoretical and computational modeling of protein dynamics have been widely used, as have Molecular dynamics (MD) and Monte Carlo techniques to provide atomistic insights into protein fluctuation. These techniques are now being extensively used to investigate various biophysical and biochemical processes including protein-ligand binding[191], protein folding [29,192] and enzyme catalysis[193].

As the timescales accessible to all-atom MD (and other coarse-grained approaches) continue to reach the microsecond and millisecond timescales, the data generated from such simulations can potentially reach $\mathcal{O}$(petabytes). The availability of large data-sets that cover the native-state dynamics and folding and unfolding pathways of the entire foldome, called Dynameomics[194], has allowed scientists to simulate over 2,000 proteins with a combined timescale of 340 μs. Projects such as Folding@home[195] have also accelerated the availability of large data-sets of protein folding trajectories as have specialized hardware, such as Anton[196], field-programmable gate-arrays (FPGA)[197], and GPUs[198].

The availability of such data-sets, while useful, has created new challenges in (a) extracting low-dimensional, biophysically relevant coordinates that elucidate how the protein functions (for example, how a protein recognizes its binding partner), (b) separating the landscape spanned by the simulations (or even groups of simulations) into a coherent set of conformational substates, (c) quantifying the intrinsic structural and dynamical properties within a substate and finally, (d) determining transition rates between these conformational substates. Indeed, important dynamical phenomena within simulated trajectories must be extracted from an enormous quantity of non-specific, ambient fluctuations. Clustering techniques for mining this noisy conformational space often use structural similarity measures, such as root-mean square deviation (RMSD) which quantifies an average value of structural deviation. However, *functional* motions need not elicit large global RMSD values; indeed, localized protein regions commonly exhibit small but important flexibility.

These observations motivated us to examine the statistical nature of atomic fluctuations from long timescale simulations[199,200]. Our studies across multiple simulations (and multiple force-

fields) reveal that functionally relevant motions generally occur rarely. These events are reflected in higher-order correlations, manifested in long-tailed spatial (fluctuation) distributions[201]. Techniques reliant on second-order statistics (variance) are poorly suited to resolve such higher-order correlations in the data, and we have observed that linear orthogonal bases (as in principal component analysis[202]) poorly describe some energy landscapes. Thus, the current frameworks to analyze long timescale trajectories *do not guarantee that identified substates are correlated with biophysically relevant events.*

We put forward a low-dimensional representation of protein motions at long timescales using a novel technique, *quasi-anharmonic analysis* (QAA)[203]. QAA partitions the conformational landscape using fourth-order *spatial*-fluctuation statistics and detects substates with *energetic* coherence. Each region contains conformers that show similarity with respect to biophysically relevant order parameters. The insights gained from QAA were effectively used to resolve higher-order dependencies in *spatial* fluctuations in the context of molecular recognition and enzyme catalysis.

While QAA effectively captures spatial correlations, it lacks a stochastic model of the underlying dynamics and substate transitions. To address this shortcoming, we build auto-regressive (AR) models to both encode local protein dynamics accessible within energetically coherent substates and permit transitions between connected regions in the landscape. We call this method the *quasi-anharmonic auto-regressive model* (QAARM). Within a QAA-derived subspace, metastable states and their transition rates are extracted using hierarchical Markov clustering which provides parameter sets for the second-order AR model. We show that the learned AR model can be extrapolated to synthesize trajectories of arbitrary length. We exploit the time-invariant statistical regularities within protein motions to investigate equilibrium fluctuations of ubiquitin, a widely studied protein involved in the proteosomal degradation pathway. We show that QAARM can extract and synthesize pathways by which ubiquitin adapts its binding surface to recognize a variety of substrates.

■ 4.2 Related Work

Previous studies have focused on the use of auto-regressive models in the frequency domain to understand memory functions in MD simulations[204]. The approach has been used to interpret

quasi-elastic neutron scattering experiments[205] and to accelerate MD simulations[206]. Using principal component analysis (PCA),[207] pursued time-series analysis of MD simulations[208–210] to hierarchically describe the energy landscape and analyze explicit solvent effects on protein dynamics. However, PCA-based representations and their extensions are limited in their description of the conformational landscape [211,212] due to assumed Gaussian fluctuations, and hence such approaches may not sufficiently describe conformational diversity[213,214].

More recent kinetic modeling, based on Markov state models (MSM), can describe the kinetics associated with protein folding[29,40,215]

MSMs commonly use RMSD values to first cluster simulation conformations into kinetically accessible micro-states and then iteratively merge these micro-states into several macro-states. MSMs can provide insights into macro-state dwell times (residence time) and can characterize mean first passage times. However, structure-based clustering need not result in energetically coherent substates. Our complementary but generative approach here explicitly pursues energetically coherent substates clusters which correspond intuitively to separated energy wells. Chiang et al., in Ref. 39, developed a related approach based on Markov dynamic models (MDM) which includes a set of hidden states to capture conformational dependencies. The generative models resulting from MDM were applied on small systems such as alanine dipeptide. In comparison, we illustrate our results on real proteins such as ubiquitin and also demonstrate the utility of our approach to reveal molecular recognition pathways.

■ 4.3 Approach

An overview of quasi-anharmonic auto-regressive model (QAARM) is shown in Fig. 4.1. MD simulation data is first processed to remove rotational and translational degrees of freedom. Quasi-anharmonic analysis (QAA) is then applied (Section 4.5) which outputs a reduced dimensional representation of the original MD data. Motivated to detect biophysically relevant energy wells, or highly populated regions, in the low-dimensional QAA space, we next use a simple Markov diffusion model to cluster the conformations into meta-stable substates (Section 4.6). Local dynamics within each substate are then captured by a linear, second-order auto-regressive model (Section 4.7) which explicitly models spatial fluctuations. The AR model thus extends the time-insensitive QAA model by considering temporal relationships between successive MD frames.

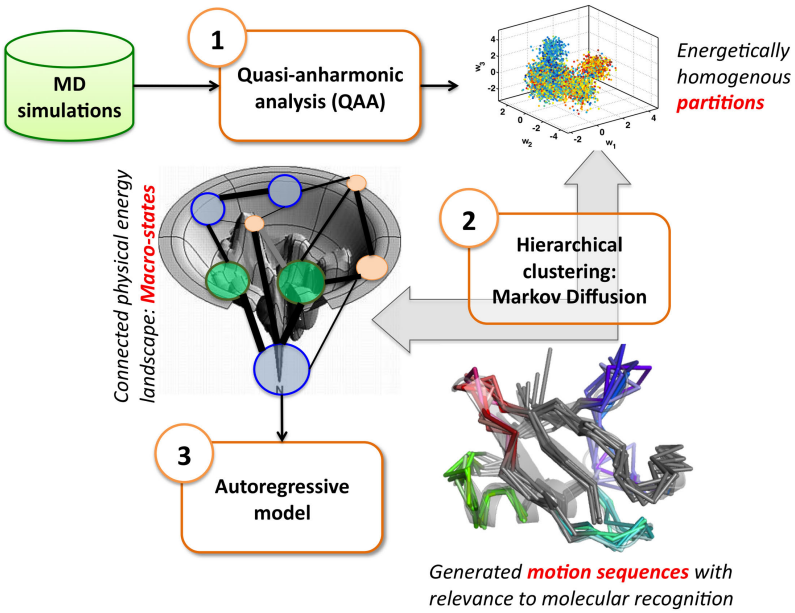


Figure 4.1: Overview of QAARM: We use MD simulations as input to quasi-anharmonic analysis (QAA). The output of QAA is a reduced dimensional space, in which conformers clustered together represent *micro-states*. This reduced-dimensional space is then input into a Markov diffusion framework to identify clusters of conformations that are kinetically accessible. These clusters represent *metastable macro-states*. We then build second order auto-regressive models for each substate to identify pathways between metastable states.

■ 4.4 Molecular dynamics simulation of human ubiquitin

Ubiquitin, a small globular protein, is involved in the proteosomal degradation pathway. It consists of 76 residues and folds into a well defined β-grasp fold. Ubiquitin's structure is evolutionarily conserved across all eukaryotes, consisting of five anti-parallel β-strands ($\beta_1 - \beta_5$) as well as two α-helices. The primary binding surface (R1 in Fig. 4.4) of ubiquitin is composed of a small number of residues proximal to the flexible $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ loops. A secondary binding interface consists of the $\beta_4 - \alpha_2$ region. Ubiquitin binds to over 300 or more targets in the human cell and naturally has been the focus of many experimental and computational efforts to characterize molecular recognition[216]. With a large number of crystal structures and NMR conformers available (both substrate-free and substrate-bound), ubiquitin provides an ideal platform for studying protein dynamics in the context of biomolecular recognition.

The protocol for simulating ubiquitin in solution is described elsewhere[199]. Briefly, eight crystal structures of ubiquitin (PDB codes: 1UBQ, 1P3Q, 1S1Q, 1TBE, 1YIW, 2D3G, 2G45 and 2FCQ) were used for our simulation. Each simulation was carried out using the AMBER suite of

tools, and each production run lasted a total of 62.5 ns. Hydrogen atoms were simulated using SHAKE algorithm, while electrostatics were evaluated using the particle mesh-ewald (PME) technique. A cut-off of 10 Å was used for long-range interactions (electrostatic and van der Waals). Conformations were stored every picosecond resulting in a total of 62,500 conformations per simulation. The simulations cumulatively constitute 0.5 μs of sampling in the ubiquitin landscape. For analyses only C^{α} atoms were used. All trajectory processing was performed with MATLAB.

■ 4.5 Quasi-anharmonic representation of protein dynamics

QAA is a general, statistically rigorous approach to identify non-Gaussian and rare behavior within extensive atomistic molecular dynamics (MD) trajectories. It utilizes higher-order statistics of protein motions and is not restricted to orthogonal basis directions, a major compromise of existing techniques. QAA identifies energetically coherent substates in the conformational hierarchy and also possible transitions between these substates, consistent with the understanding that proteins sample from a hierarchical, multi-level energy landscape, with minima/ maxima separated by energy barriers[9,10]. Internal protein motions, driven by thermal energy in the solvent, enable proteins to explore this rugged landscape.

Here, we summarize quasi-anharmonic representation of protein motions in long timescale simulation trajectories based on diagonalization of a tensor of fourth-order statistics. This tensor describes positional fluctuations and their couplings. We use an efficient algebraic technique called joint-diagonalization of cumulant matrices (JADE), a well known algorithm in the machine learning literature for analyzing multi-variate data[217].

First, we assume that overall rotation/translation degrees of freedom have been removed and hence that positional fluctuations $\vec{x}$ are centered around the origin. Second, second-order correlations are removed from the fluctuation data. In particular, a covariance matrix G is estimated: $G = E\{\vec{x}\vec{x}^T\}$, which is then diagonalized by orthogonal eigenvectors B and eigenvalues Σ using $G = B\Sigma B^T$, followed by elimination of second-order correlations in $\vec{x}$ with $\vec{\alpha} = \Sigma^{-1/2}B^T\vec{x}$, leaving $E\{\vec{\alpha}\vec{\alpha}^T\} = I$, an identity matrix of size $3N \times 3N$ for N atoms under consideration.

Third, a fourth order cumulant tensor $\mathcal{K}$ is estimated comprising both auto- and cross-

cumulants given by

$$\kappa(\alpha_i) = E\{\alpha_i^4\} - 3E^2\{\alpha_i^2\}, \tag{4.1}$$

and

$$\kappa(\alpha_i, \alpha_j, \alpha_k, \alpha_l) = E\{\alpha_i, \alpha_j, \alpha_k, \alpha_l\} - E\{\alpha_i, \alpha_j\}E\{\alpha_k, \alpha_l\} - E\{\alpha_i, \alpha_k\}E\{\alpha_j, \alpha_l\} - E\{\alpha_i, \alpha_l\}E\{\alpha_k, \alpha_j\}, \tag{4.2}$$

respectively. This expression is further simplified because $E\{\vec{\alpha}\vec{\alpha}^T\} = I$, and hence $E\{\alpha_i\alpha_j\} = 1$ when $i = j$ and 0 when $i \neq j$. The cumulant tensor will have a total $3N \times (3N + 1)/2$ matrices each of size $3N \times 3N$ accounting for auto- and cross-cumulant terms.

Fourth, the fourth order dependencies denoted by the sum of the cross-cumulant terms are minimized, a procedure equivalent to diagonalizing the tensor $\mathcal{K}$. No closed form solution exists for diagonalizing a tensor, however an approximate solution can be found using efficient algebraic techniques, such as Jacobi rotations[218]. Just as the eigenbasis B diagonalizes the covariance matrix G, a rotation matrix J can be found which approximately diagonalizes the cumulant tensor $\mathcal{K}$, leading to:

$$\vec{w} = J\vec{\alpha}. \tag{4.3}$$

Substituting for $\vec{\alpha}$ from above:

$$\vec{w} = J\Sigma^{-1/2}B^T\vec{x}, \tag{4.4}$$

and thus $\vec{w} = U^{-1}\vec{x}$ implying

$$U = B\Sigma^{1/2}J^T. \tag{4.5}$$

Thus, U represents anharmonic modes of motion derived by minimizing the fourth-order dependencies in positional fluctuations, in addition to eliminating the second-order correlations (as is the case with quasi-harmonic analysis) (see Fig. 4.2). Unlike in approaches that use principal component analysis, U can be non-orthogonal and hence intrinsically coupled. The anharmonic modes of motion U_i, each a column vector of matrix U, are sorted decreasingly by amplitude ($\|U_i\|$).

Finally, we paint each conformer in the QAA subspace by internal energy, the sum of electrostatic and Van der Waals interactions (computed with NAMDenergy[219]) over each conformation. We emphasize that resultant energy coherence within observed substates is an emergent property

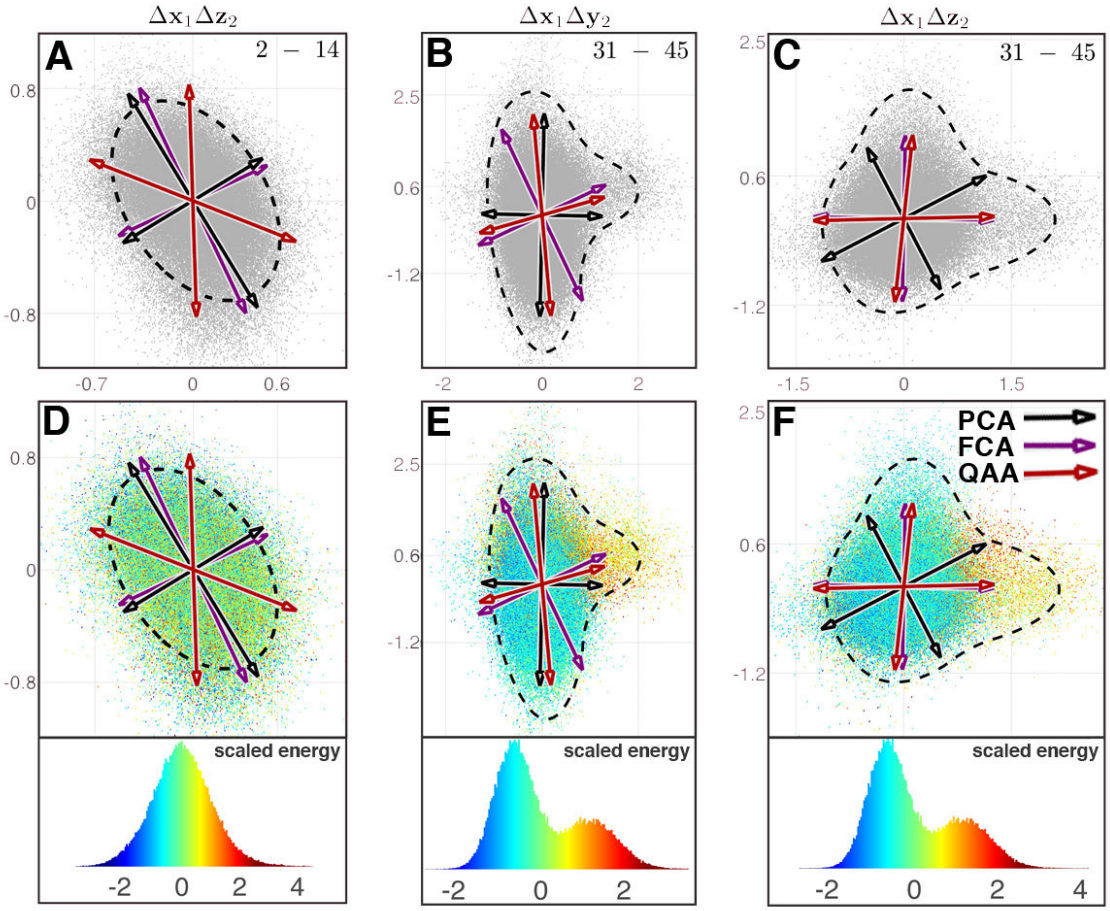


Figure 4.2: QAA vectors on 2D atomic displacement distributions, compared with FCA and PCA. Spatial fluctuation distributions for three atom pairs from simulations of ubiquitin are shown, each datapoint representing two spatial coordinates for a single conformer, one from the first atom and one from the second. (A) Fluctuations in the x direction for C^{α} 2 are shown on the vertical axis. Fluctuations in the z direction for C^{α} 14 are shown on the horizontal axis. Pairwise fluctuations are mostly Gaussian or harmonic, manifested as an oval probability distribution. (B) Residue pair 31–45 experiences anharmonic fluctuations in the x and y directions, respectively, indicated by a bulge in the probability distribution. (A–F) Principal component analysis (PCA) models the data according to the black arrows, the directions that maximize variance. Purple arrows indicate the modes derived from full correlation analysis[214], which minimizes mutual information but is restricted to orthogonal modes. QAA basis vectors (red) point in the direction of greatest anharmonicity and need not be orthogonal. Notice that QAA vectors point toward the directions of increasing energy, especially panels (B) and (C). (D–F) Energy[219] distributions are shown for each joint distribution. The color range in each (lower panels) is thresholded above and below $\pm 2.5\sigma$ for visual clarity. All spatial units are in Å. For each residue pair a total of 100,000 conformers were used from $0.5\mu s$ of simulation[203].

of the method, that is, conformer internal energies are not considered during the projection onto the QAA subspace.

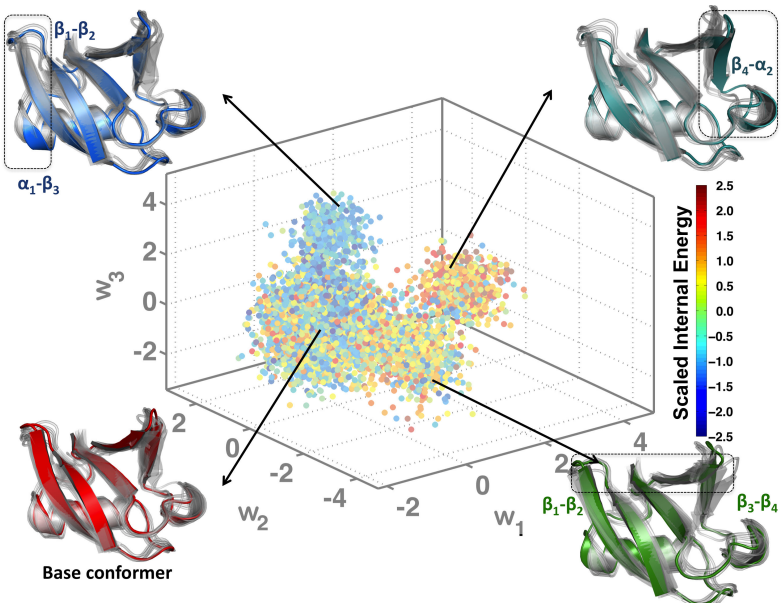


Figure 4.3: Organizing the conformational landscape of ubiquitin into energetically homogenous regions: Projections ($w_1 - w_3$) of the conformations from the ubiquitin simulation onto the top three anharmonic basis vectors are colored according to the scaled internal energy values of the conformers. Also shown are the results from hierarchical clustering (Section 4.6), where a total of 78 clusters were identified. The arrows indicate portions of the landscape from where some of the clusters originated. Illustrative examples of cluster centers are shown as colored cartoons; overlaps with the other cluster centers are shown as transparent gray cartoons. In each of the example cluster centers, the region that undergoes the maximal conformational change is highlighted using a dotted rectangle. The corresponding secondary structures are marked for ease of identification.

4.5.1 Organizing ubiquitin conformational landscape into energetically homogenous regions

From the original simulation consisting of nearly 500,000 conformations (0.5 μs), 10,000 equally spaced conformations were collected for training the QAA basis. We performed QAA within a 30-dimensional PCA subspace which covers 95% of the input variance. The anharmonic modes of motion reveal the ability of ubiquitin to modulate both the primary and secondary binding surfaces ($\beta_1 - \beta_2$ and $\beta_3 - \beta_4$; $\beta_4 - \alpha_2$), as shown in Fig. 4.3. In addition, the distances between $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ can also serve as order parameters to describe the anharmonic landscape spanned by our simulations. Motions along each of the anharmonic modes permits ubiquitin to adopt a conformation that resembles the substrate-bound conformation. Our AR models will be deemed successful if they can recover this property.

■ 4.6 Hierarchical clustering for metastable substates

Energy wells in the 30 dimensional QAA-space determine biophysically relevant substates; the structure and dynamics of each can be characterized through clustering. Neighboring conformers in QAA-space have similar internal energies (Fig. 4.3) and thus are dynamically and kinetically related. To facilitate clustering, we model the MD trajectory as an undirected network where edges connect energetically adjacent conformers in QAA-space (using Euclidean distance within the QAA space). The energetic coherence of a neighbors is simply an observation given the proximity of energetically coherent conformers shown in Fig. 4.3. We can then cluster this network using a hierarchical Markov diffusion framework. This approach is an extension of earlier applications of spectral graph partitioning algorithms for segmenting natural images[162,220], understanding protein dynamics and allosteric propagation[221], and relating signal propagation on a protein structure to its equilibrium dynamics[222].

We begin hierarchical clustering by constructing a Markov transition matrix using edge weights between conformer pairs. Weights are chosen according to distance within QAA-space between connected conformers.

We then initiate a Markov chain (or random walk) on the weighted undirected network. As Markov transition probabilities homogenize through diffusion, an implicit clustering emerges from the network. First, a set of nodes representing the putative clusters are identified. Then, a Markov transition matrix is newly constructed using this reduced representation. The principle behind this construction is that upon reaching a stationary distribution at the coarsest hierarchy level, the Markov chain has also converged at finer (more local) network levels. This consistency regulates the overall topology of the network and helps build a multi-resolution representation of metastable states.

We expect that fine-grained hierarchy levels will produce many small clusters containing close neighbors in QAA space; that is, most cluster members will be from the same time-window (and single trajectory). As Markov diffusion progresses (fine-grained to coarse), conformers that are more distant neighbors will be connected by edges in the diffused network, and will therefore be assigned to the same cluster. Thus, the hierarchical clustering can highlight dynamical connections between conformers at different timescales.

■ 4.6.1 Markov diffusion framework

Initiation: The MD simulation is modeled as an undirected graph by placing an edge with weight 1 from each data point to its six nearest Euclidean neighbors in the QAA space. At hierarchy level $t = 0$, each data point is considered a node. Let n_o be the number of trajectory frames. The $n_0 \times n_0$ adjacency matrix C_0 gives the edge weights between each data point pair and the $n_0 \times n_0$ diagonal degree matrix D_0 (Step 1) gives the connectivity at each node in that $D_t(i,i)$ contains the total number of connections to node i at hierarchy level t. Nodes with high degrees can be seen as hubs, and nodes with very low degrees can be seen as isolates. The stationary distribution of the Markov chain is given by the normalized degree vector $\vec{\pi}_0(i) = \frac{D_0(i,i)}{\sum_j D_0(j,j)}$, and represents the probability of a Markov Chain residing in a particular node after infinite iterations.

Iteration: For $t = 1$ until done:

1. Compute the diagonal degree matrix D_{t-1}, with entries

$$D_{t-1}(i,j) = \begin{cases} \sum_{j=1}^{n_{t-1}} C_{t-1}(i,j) & i = j \\ 0 & i \neq j \end{cases}$$

   and the Markov transition matrix $M_{t-1} = C_{t-1} D_{t-1}^{-1}$.

2. Diffuse the Markov transition matrix by repeated multiplication $M_{t-1}^d = M_{t-1} \times M_{t-1}$ to reveal distant connectivity.

3. Determine the $(n_{t-1} \times n_t)$ kernel matrix K_t to carry network information from hierarchy level $(t-1)$ to level (t). The kernel matrix is made up of a subset of $n_t \ll n_{t-1}$ columns of M_{t-1}^d, which is selected such that the size of n_t is minimized while maintaining nonzero transition probability among all n_{t-1} points.

4. Solve $\vec{\pi}_{t-1} = K_t \vec{\pi}_t$ for $\vec{\pi}_t$ with an expectation-maximization algorithm to find a low-dimensional representation $\vec{\pi}_t$ of the stationary distribution $\vec{\pi}_{t-1}$.

5. Compute C_t using $\vec{\pi}_t$:
   $C_t = \text{diag}(\vec{\pi}_t) K_t^T \text{diag}(K_t \vec{\pi}_t)^{-1} K_t \text{diag}(\vec{\pi}_t)$, where K_t^T denotes the transpose of K_t and $\text{diag}(\vec{\pi}_t)$ indicates the diagonal matrix formed from the vector $\vec{\pi}_t$.

6. $t \to t + 1$

Termination: End if $n_t \leq 2$. At this point, the network has been divided into one or two clusters.

Backwards iteration along the hierarchy allows computation of an $(n_{t-1} \times n_t)$ ownership matrix O_t for each hierarchy level t, in which $O_t(i,j)$ gives the probability that data point i belongs to cluster j at level t of the hierarchy:

$$O_t(i,j) = \frac{K_t(i,j)\vec{\pi}_t(j)}{\sum_{k=1}^{n_t} K_t(i,k)\vec{\pi}_t(k)},$$

where $\sum_{j=1}^{n_t} O_t(i,j) = 1$. The ownership matrix gives the probability distribution for the likelihood that a data point belongs to any metastable state of the trajectory, providing a soft partitioning of the data. A hard partitioning is determined by assigning each data point to the cluster to which it has maximal ownership probability.

■ 4.6.2 Characterizing metastable substates in the ubiquitin landscape

The connectivity matrix C_0 at clustering initialization is shown in Fig. 4.4. The connectivity matrix shows several regions of high cross-talk. Iterative diffusion of the Markov chain derived from this connectivity matrix, followed by kernel selection, results in six hierarchy levels with 10000, 4486, 978, 78, 11, and 2 clusters at each respective level.

To provide parameters for the AR model (Section 4.7), a membership threshold must be chosen that is fine enough to capture local dynamics, but still coarse enough to allow flexibility. We chose a membership threshold such that all cluster members were reachable from the cluster center within 50 ps, where a substate's center is defined as the closest conformer to the mean of that substate. The mean QAA-space distance between conformers in successive frames is $\hat{\vec{d}}$, and the standard deviation is σ_d. The hierarchy level at which 99.7% of the conformers are within $\hat{\vec{d}} + 3\sigma_d$ of their substate center is selected for further processing. Following this criterion, auto-regressive analysis was pursued using statistics from level 4 of the hierarchy.

Four cluster centers from the connectivity matrix are shown for example clusters in Fig. 4.4. Clusters can be mapped from QAA-space onto the connectivity matrix to visualize accessibility between substates. As an example, *dynamically distant* clusters are illustrated in Fig. 4.4 by the red and green enclosed regions. The metastable substates identified share significant similarities in both conformational and energetic space, however, they do not interact directly in QAA-space. This produces a partitioning of the landscape that is quite unique from the perspective of understanding ubiquitin's equilibrium fluctuations: the landscapes's extrema represent distinct

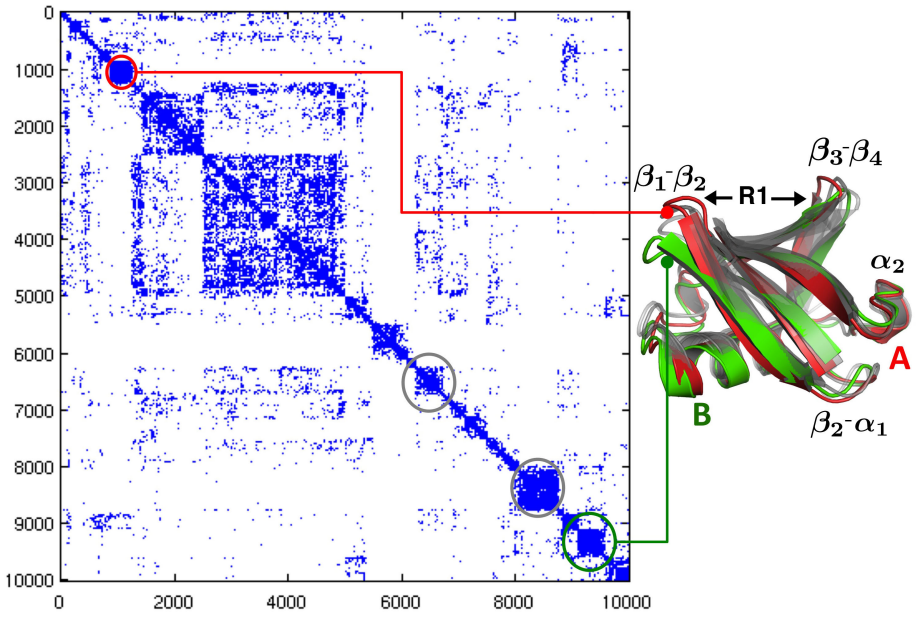


Figure 4.4: Markov diffusion clustering of QAA shows ubiquitin motions involved in binding substrates: The 30 dimensional space determined form QAA is used to construct a set of meta-stable states that are energetically accessible. From a group of 10,000 conformers, we show how the network is modeled with the adjacency matrix C_0 shown here. The Markov diffusion produces a total of 78 macro-states at level 4 of the hierarchy. To illustrate the extremum points in the network, we depict two representative clusters (A and B shown in red and green respectively) representing changes within the binding regions of ubiquitin. The primary binding region is indicated by **R1**.

conformations of ubiquitin's binding regions. Note that while $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ adapt an "open" conformation in the structure shown in green (average separation of over 18 Å), the red structure shows the binding regions "close" to each other (average separation of 13.5 Å). Thus, the inherent motions of ubiquitin involve sampling the two metastable states with almost exclusively no cross-talk. However, note that both the red and green structures can interconvert between the metastable states highlighted in gray. These metastable states represent the so-called *intermediates* which are necessarily visited before sampling either open or closed conformations. Intermediate metastable states also highlight the importance of ancillary structural changes that ubiquitin might have to undergo in order to sample either the open or closed conformations. These changes are predominantly located along $\beta_2 - \alpha_1$ and the C-terminal tip of α_1 helix. Thus, for the opening and closing of the binding region in ubiquitin, our pathways deduced from the Markov transition matrices reveal that it is energetically more favorable to undergo conformational changes along the two regions highlighted in gray.

■ 4.7 Building Auto-regressive models

Motivated by the need for a compact, linear, and generative model of protein dynamics, we extend our findings from QAA and hierarchical clustering with a stochastic auto-regressive (AR) model inspired by problems in control theory and signal processing[223]. Understood as a second-order stochastic differential equation that has been sampled at regular intervals, the model relates each successive protein conformation to the previous two. It consists of an appearance model and a dynamic model, which can be conceptualized, respectively, as encoding the protein in a meaningful low-dimensional (embedded) space and modeling characteristic motions within that subspace. Because we learn summarizing parameters for the protein's dynamics in the model, we can synthesize extrapolated trajectories of arbitrary length. As a contrast to molecular dynamics methods, where the *system* and *environment* are simulated and dynamics result, or fragment models of protein structure[224], the approach here models *dynamics* explicitly and exploits the statistical regularities intrinsic to a natively fluctuating protein. Within this approach, time-evolution of the protein's conformation $\vec{x}_t$ results from coupling the appearance and dynamics models:

$$\vec{x}_t = U\vec{w}_t + \vec{\epsilon}_t, \quad \vec{\epsilon}_t \backsim N(0, R), \tag{4.6}$$

$$\vec{w}_t = A_1\vec{w}_{t-1} + A_2\vec{w}_{t-2} + \vec{\eta}_t, \quad \vec{\eta}_t \backsim N(0, Q), \tag{4.7}$$

where weights $\vec{w}_t$ constitute the projection of $\vec{x}_t$ onto the subspace spanned by U, or its *state*. Determining a physically meaningful subset of basis vectors, and the vectors themselves, is frustrated by the enormous conformational space accessible to a fluctuating protein. A poor choice of basis vectors (or selecting too few) would increase the reconstruction error $\vec{\epsilon}_t$ upon mapping each embedded state back to full conformational space. The basis chosen here, U, is the first 30 anharmonic modes which are extracted using QAA and which span the conformational subspace available to the dynamic model. Both deterministic and stochastic elements are contained in the model's dynamic component (Eqn. 4.7). The deterministic element is a second order Markov model in which the state at time t, $\vec{w}_t$, is a linear combination of states $\vec{w}_{t-1}$ and $\vec{w}_{t-2}$. Stochasticity is introduced by the Gaussian driving distribution $\vec{\eta}_t$, which quantifies motions that are not fully captured by the linear model. More temporal information is available to this model than to a first order model; we show that this permits characterization of complex motion patterns

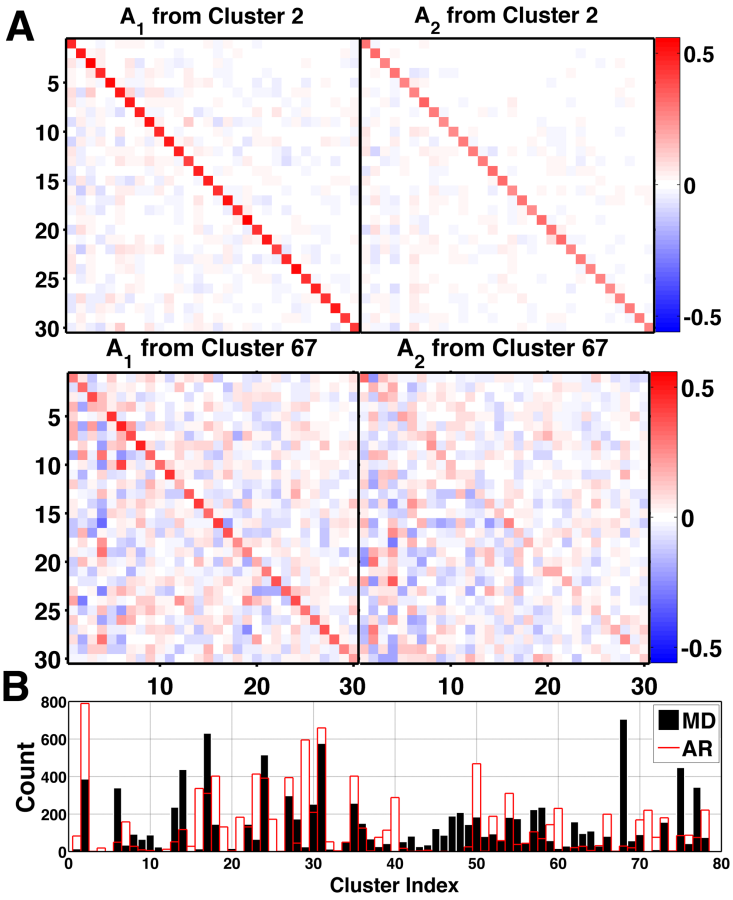


Figure 4.5: Representative transition matrices are highly diagonal: (**A**) A_1 and A_2 for the most populated cluster, cluster 2, which contained 29108 structures or 5.84% of the entire .5μs simulation. Cross correlations between QAA modes are highly reduced, yielding low off-diagonal elements. Distinctions between A_1 and A_2 indicate the constituent structures (from cluster 2) carried dynamic information across multiple frames. The lower two panels show less strongly diagonal transition matrices for a less populated cluster, 67, which contained 627 structures. Elements of A_1 and A_2 range from $-.84$ to .72 over all clusters ($-.33$ to .5597 over clusters 2 and 67). (**B**) Cluster memberships for MD training data (black) and AR-synthesized (red) ubiquitin conformations, 10,000 frames each.

that extend over several timeframes (or MD conformations during training). Distinct from the connectivity matrix in Section 4.4, the *transition matrices* A_1 and A_2 here constitute the second-order transition matrices of the stochastic process and must be learned from training simulation data. In the following subsections we address learning these dynamical model parameters and generating synthetic trajectories.

■ 4.7.1 Learning the dynamical model

The dynamical model, Eqn. 4.7, exploits our knowledge of past states (conformations) to propose a future state. Before we can compute transition matrices A_1 and A_2 from training data, we first

project the ubiquitin simulation into the embedded 30-dimensional QAA-space to yield training states $\vec{w}_t$:

$$\vec{w}_t \equiv U^T X, \tag{4.8}$$

where columns of X, $\vec{x}_1 \dots \vec{x}_T$, are $3N$ vectors carrying the protein's coordinates (for N residues). Following the derivation put forward in Ref. 225, the auto-regressive model is defined sequentially over the weights:

$$\begin{aligned} \vec{w}_3 &\approx A_1 \vec{w}_2 + A_2 \vec{w}_1 \\ \vec{w}_4 &\approx A_1 \vec{w}_3 + A_2 \vec{w}_2 \\ &\vdots \\ \vec{w}_T &\approx A_1 \vec{w}_{T-1} + A_2 \vec{w}_{T-2} \end{aligned} \tag{4.9}$$

with unknowns A_1 and A_2. We concatenate state vectors and transition matrices with the notation $W_{i,j} \equiv [\vec{w}_i \ \vec{w}_{i+1} \dots \vec{w}_j]$ and $\mathbf{A} \equiv [A_1 \ A_2]$ to express the system in matrix form:

$$W_{3,T} = \mathbf{A}\mathbf{W}_1^2 \quad \text{where} \quad \mathbf{W}_1^2 \equiv \begin{bmatrix} W_{2,T-1} \\ W_{1,T-2} \end{bmatrix}. \tag{4.10}$$

The total squared error between the true states and the predicted states is minimized with the Frobenius norm $\| \cdot \|_F$:

$$\mathbf{A} = \underset{\hat{\mathbf{A}}}{\operatorname{argmin}} \left\| \hat{\mathbf{A}}\mathbf{W}_1^2 - W_{3,T} \right\|_F . \tag{4.11}$$

Generally the state subspace is much smaller than the number of observations (training simulation frames), so $W_{i,j}$ is rarely square. The solution to Eqn. 4.11 then follows:

$$\mathbf{A} = W_{3,T}\mathbf{W}_1^{2*}, \tag{4.12}$$

where $F^* \equiv F^T \left(FF^T\right)^{-1}$ denotes the pseudo-inverse of a matrix F. Representative A_1 and A_2 matrices are shown in Fig. 4.5A. The stochastic term, $\vec{\eta}_t$, represents those dynamics that are inadequately captured by the second-order linear model, and is drawn from a Gaussian distribution with covariance equal to that of the prediction error averaged over the training sequence. That is,

$$\begin{aligned} &R = \mathbb{E}\left[PP^T\right], \text{where the prediction error is} \\ &P = W_{3,T} - \left(A_1 W_{2,T-1} + A_2 W_{1,T-2}\right). \end{aligned} \tag{4.13}$$

Interpreted physically, each A_1 and A_2 pair encodes the local, *time invariant* dynamics. The eigendecomposition of $\left[\begin{smallmatrix} 0 & I \\ A_2 & A_1 \end{smallmatrix}\right]$ yields the exponential decay constants $\beta_m = \frac{1}{\tau}\log\frac{1}{\lambda_m}$ for these local dynamics, where $\lambda_m < 1$ denotes any positive eigenvalue[223].

■ 4.7.2 Synthesizing new motion sequences

The learned transition matrices A_1 and A_2, unique to each cluster, can be used to generate novel structure sequences of arbitrary length. That is, $\vec{w}_t$ and $\vec{x}_t$ now constitute the unknowns in Eqns. 4.6 and 4.7. Starting from a randomly selected frame from our training trajectory, we propagate the model using only the learned transition matrices. Within the QAA space defined by the column vectors of U, we compare each generated conformer to the mean structure of every cluster. At every step, the nascent trajectory is assigned to the cluster center with nearest Euclidean distance, and permitted to evolve according to that cluster's A_1, A_2, and $\vec{\eta}_t$ until it moves closer to a different cluster center. We generated a synthetic trajectory of 25,000 frames, during which the protein visited 76 of the 78 clusters present in the training data (cluster membership for 10,000 training and testing frames is shown in Fig. 4.5). Other than error/transition parameters and the determined local mean of each cluster, no temporal information from the training data was necessary for the time-evolution of ubiquitin's dynamics. Additionally, it should be noted that the entirety of the generative process is carried out in the embedded QAA subspace, and the appearance model, Eqn. 4.6, is only used in post processing to return to full, $3N$ conformation space. We can conceptualize each substate's transition matrices as linearly encoding local dynamics, whereas reconstruction information from embedded to full conformational space is carried in the QAA basis vectors U. Far less storage and computing resources are required to propagate the AR-derived dynamics than with conventional sampling. Temporally, the synthetic trajectory employs the same time step found in the training data; the 25,000 synthetic trajectory frames compares with approximately 25 ns of MD simulation.

■ 4.7.3 Predicting pathways of molecular recognition in ubiquitin

The underlying stochastic dynamical model allows us to synthesize new conformations of arbitrary length. This is particularly useful when one has to predict the binding mode of ubiquitin with another protein. Note that in our simulations, ubiquitin was simulated in its substrate-free form. Hence, no explicit knowledge of the substrate-bound form was available. However, when we

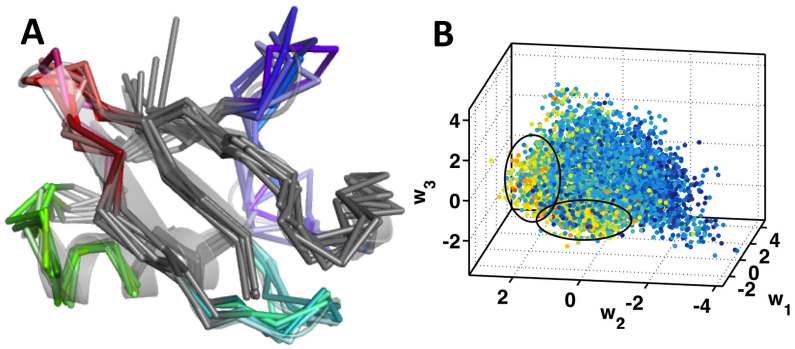


Figure 4.6: Synthesized conformations from QAARM reveal novel binding modes of ubiquitin substrate. (A) A movie-like representation of the fluctuations in ubiquitin synthesized from the QAARM model. Note that regions undergoing large conformational changes include the binding regions and the ancillary regions of ubiquitin. These motions have a direct implication on binding a variety of substrates. (B) Synthetically generated 25,000 conformers are projected back onto the QAA bases to reveal the number of potential contacts that each of the synthesized conformer can make with a known substrate, Rabex 5. Note that the substate highlighted by the ellipse consists of a small number of conformers in the synthesized data that can form a large number of contacts with Rabex 5.

synthesize 25,000 conformers from QAARM, its utility becomes quite evident. The conformers show fluctuations along the flexible regions of ubiquitin (highlighted in Fig. 4.6A). Further, these motions are largely similar to the fluctuations in the ubiquitin simulations, as evidenced by projecting the synthesized conformations back onto the QAA basis vectors.

It is also interesting to note that the projection of the synthesized conformers onto the QAA basis space reveals novel pathways of ubiquitin binding. To illustrate this, we chose the PDB id: 2FIF where ubiquitin interacts with Rabex-5 along the secondary binding site of ubiquitin ($\beta_2 - \alpha_1$ and $\beta_3 - \beta_4$ interface). We computed the estimated number of contacts each synthesized conformer would form if it were to be superimposed onto the crystal structure. Since we use only C^{α} atoms to generate the conformers, we consider two atoms to be in contact whenever they are separated by less than 7.3 Å[226]. The results of this computation are illustrated in Fig. 4.6B. By projecting the synthesized conformers onto the QAA basis space, we discover that a small number of conformers (highlighted by ellipses) form a large number of contacts with the substrate. Furthermore, the other parts of the landscape (in QAA) show meager contacts with the substrate. This allows us to pinpoint a specific mechanism by which only a small number of the generated conformers can bind to ubiquitin in a specific manner. Although it remains to be seen if these conformations are also energetically favorable, we believe that QAARM has allowed one to predict complex formation by exploiting the statistical regularities in the substrate-free simulations of

ubiquitin. Thus, in line with previous studies that proposed conformational selection to be a predominant pathway for recognizing binding partners in ubiquitin[227], our studies also predict a similar mechanism (at least at the C^{α} resolution).

■ 4.8 Discussion

Well-sampled conformational space is more useful with organizational principles which can describe and characterize it. Methods for extracting meaningful features and events must cope with longer simulations of increasingly larger and more complicated systems, and should eventually be used for validating and error-checking MD simulations themselves. The trajectory studied here samples many of the unique binding poses ubiquitin must adopt for specific recognition of diverse ligands, providing a rich platform for studying functionally relevant structural transitions. However, ubiquitin's structural shifts are subtle when compared to those of hinge or multidomain proteins; that these motions and connecting pathways are distinctly resolved with our method speaks to the utility and suitability of higher-order statistics for decomposing conformational space.

We exploited long-tail spatial distributions in former work (quasi-anharmonic analysis), and we extended it in this paper with linear stochastic models which account for temporal dependencies. This allows us to explore specific, local dynamics accessible to a protein within energetically homogeneous wells. Clustering to determine substates was performed here within the 30-dimensional QAA subspace, and we plan to compare our identified substates to those from other clustering methods in the future. Additionally, increasingly subtle or energetically local behaviors can be encoded by learning AR models at successive clustering levels; how we couple AR-models at potentially disparate hierarchical levels to give a coherent picture of protein fluctuations is a topic of future work. In addition, we plan to apply maximum entropy methods to enable the dynamics of even poorly-sampled energy wells to be incorporated into the AR-model.

Chapter 5

Conclusions

. . . your errors grow in proportion
to the distance you cover.
Vladimir Nabokov
Ultima Thule

Being dynamic and somewhat unpredictable systems, proteins are fun to study. Their behavior gets forced into 'the conceptual boxes supplied by professional education'[228] when we want to make useful claims about how they work. We've taken two broad approaches to this task where, we hope, our methods contributions are valid and illuminating. Both are dependent on the conformational landscape paradigm, a powerful conceptualization of protein behavior at a molecular level. Projection-based methods as in Chapter 4 identify a smaller subspace of the landscape that captures important correlations in the data and allows comparison with experiment and also visualization. While it is recognized that choosing useful reaction coordinates is often challenging, the model as a whole adapts well to new phenomenological concepts. Intrinsically disordered proteins live in 'less-funneled' landscapes[229] and proteins that fold in tandem are said to 'synergistically' alter the conformational landscape[22], for example. We tried to show that some motions in this subspace are statistically regular and can be modeled with simple first-order regression. When the high-dimensional free energy surface is discretized and conformational changes are mapped to state transitions we thereby arrive at the second organizing structure considered in this dissertation: network models (Chapters 2 and 3). These too have many interpretive conveniences. Native or misfolded states can be identified as 'hubs' or 'traps' for example, but, at least at the current maturation of network science, many topological or kinetic features remain poorly-defined. Additionally, the discretization step does not avoid the reaction coordinate question since any number of features could be invoked for delineating the network nodes. So, given these challenges and compromises, where might network modeling and subspace projections of protein simulations be applied?

One area of study is protein aggregation, caused by small solvent or sequence perturbations leading to pathologic protein misfolding[230]. That phenotype can be so drastically altered by such small environmental or amino acid changes argues for perturbation studies of the conformational

landscape like those undertaken in the opening chapters. That is, even a relatively small conformational substate made inaccessible via inhibitor or solvent adjustments could potentially impact overall folding behavior, and these impacts, even when phenotypically silent, are increasingly discernable by experiment[231]. 'Velvet gloving' a protein just a little bit away from its native state in this fashion could dramatically change which binding pose is presented and which oligomerizations are possible[20,107]. F-scores are discussed here as a way of quantifying these subtle kinetic changes. Our thinking is that as human modulation of protein dynamics increases in subtlety and specificity, we will need more accurate ways of quantifying–and predicting–how the target's kinetics are impacted.

In sequence, Chapters 2 and 3 highlight our departure from the assumption that network dynamics are always dominated by a native, densely connected hub. At first, more general approaches might appear unneeded. Looking at WW in Fig. 2.6, we see a FEL characterized by well-defined folded and unfolded ensembles and a topology that is dominated by a well-like native ensemble. But we should remember that the polypeptides studied in that chapter were chosen for simulation by the original authors specifically to identify folding characteristics that promote attainment of the native state. Many proteins with key disease-associated function or dysfunction were necessarily excluded. Yet, proteins with less defined native states are in fact over-represented among key regulatory and signaling pathways in higher organisms[232,233], so functionally-relevant regions of the FEL should not reflexively be labelled 'native' any longer. Although unstructured proteins present less static pharmacophores, initial studies suggest they could in fact be *more* druggable than conventional drug targets[234], so we must properly understand complex networks that lack topologically-dominating hubs[235]. In Chaper 3 we discussed in more detail the problems arising from network centrality measures that overlook this reality. Importantly, others have recognized that topologically important nodes in a network need not be the most highly connected[146], so we feel our work complements the current toolbox of network analysis methods.

Additional Figures

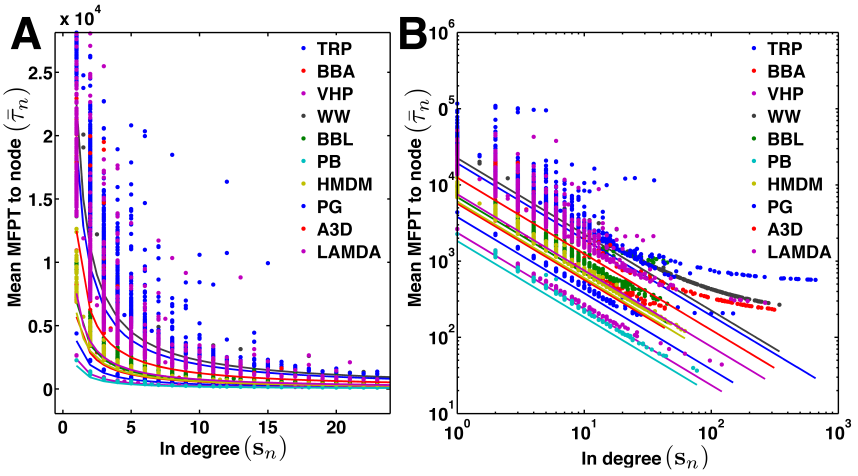


Figure A.1: Trapping times are largely a function of in-degree $\mathbf{s}_n$. For the ten protein networks tested in Chapters 2 and 3, $\mathbf{s}_n$ is plotted against average trapping time. (A) Linear axes, (B) logarithmic axes.

Table A.1: Accuracy of predicted f-scores. A duplicate of Table 3.2 with accuracy standard deviations, as depicted with error bars in Fig. 3.11.

	Total n_t	Total n_p	ρ	ρ_s	$\overline{\text{NRMSE}}_{\tilde{\lambda},\tilde{\mathbf{U}}}$	$\overline{\text{NRMSE}}_{\tilde{\lambda},\mathbf{U}_0}$	$\overline{\text{NRMSE}}_{\lambda_0,\mathbf{U}_0}$	Avg. speedup
$\mathcal{H}_{500}$	10	607	0.99	0.98	0.027 ± 0.015	0.181 ± 0.255	0.192 ± 0.231	1.05
$\mathcal{H}_{1000}$	10	837	0.99	0.98	0.026 ± 0.019	0.173 ± 0.181	0.200 ± 0.087	1.82
$\mathcal{H}_{2000}$	10	1880	0.99	0.99	0.021 ± 0.010	0.108 ± 0.118	0.144 ± 0.047	3.38
$\mathcal{H}_{A}$	10	880	0.99	0.99	0.012 ± 0.010	0.102 ± 0.087	0.109 ± 0.041	1.28
$\mathcal{H}_{YST}$	10	550	1.00	0.99	0.009 ± 0.007	0.174 ± 0.039	0.234 ± 0.056	4.27
$\mathcal{H}_{UC}$	10	1117	0.99	0.97	0.016 ± 0.007	0.096 ± 0.036	0.127 ± 0.056	2.83

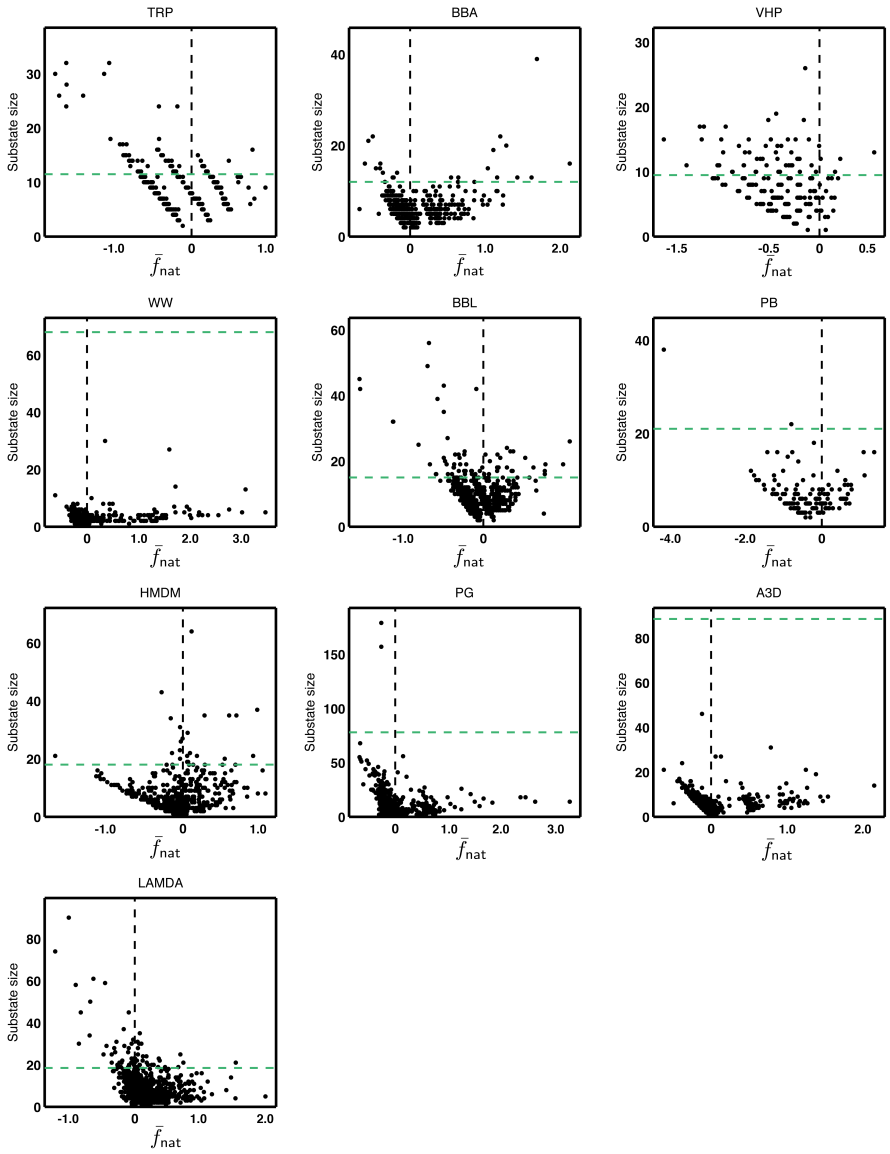


Figure A.2: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus substate populations. The green dashed line indicates the median substate size within the native ensemble.

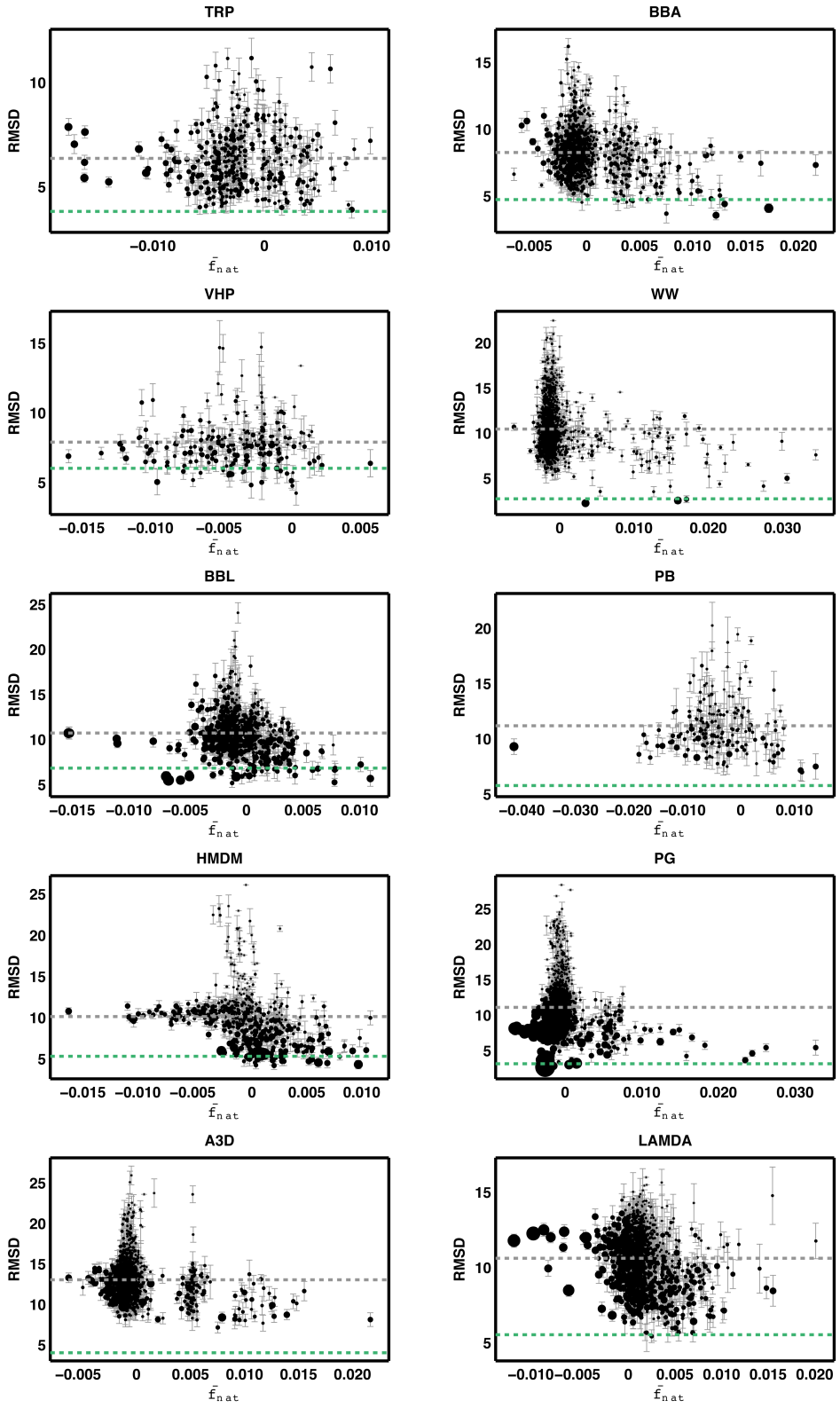


Figure A.3: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus RMSD-to-native. The green (gray) dashed line indicates the median value within the native (nonnative) ensemble.

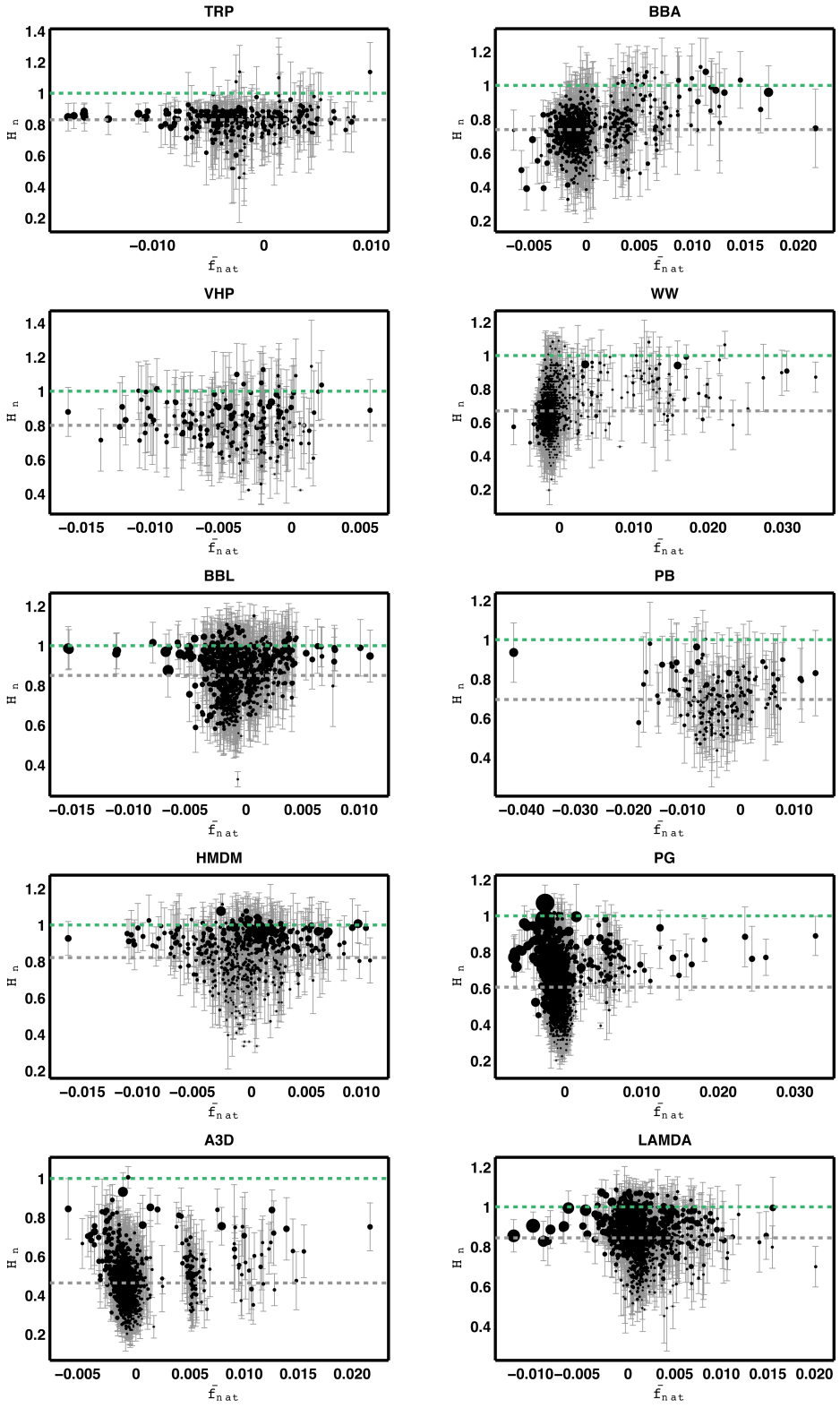


Figure A.4: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus native helicity, H_{n}. The green (gray) dashed line indicates the median value within the native (nonnative) ensemble.

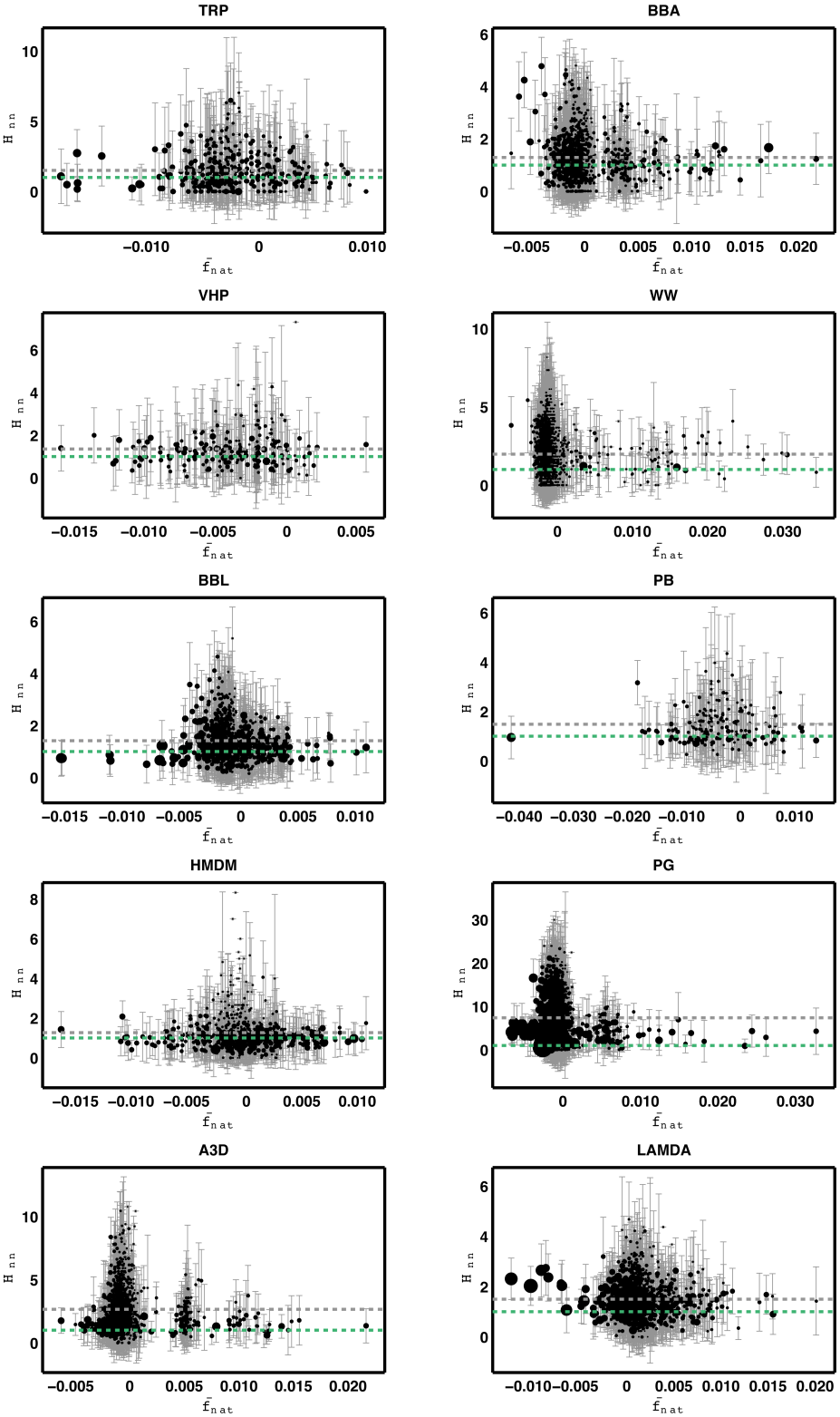


Figure A.5: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus nonnative helicity, H_{nn}. The green (gray) dashed line indicates the median value within the native (nonnative) ensemble.

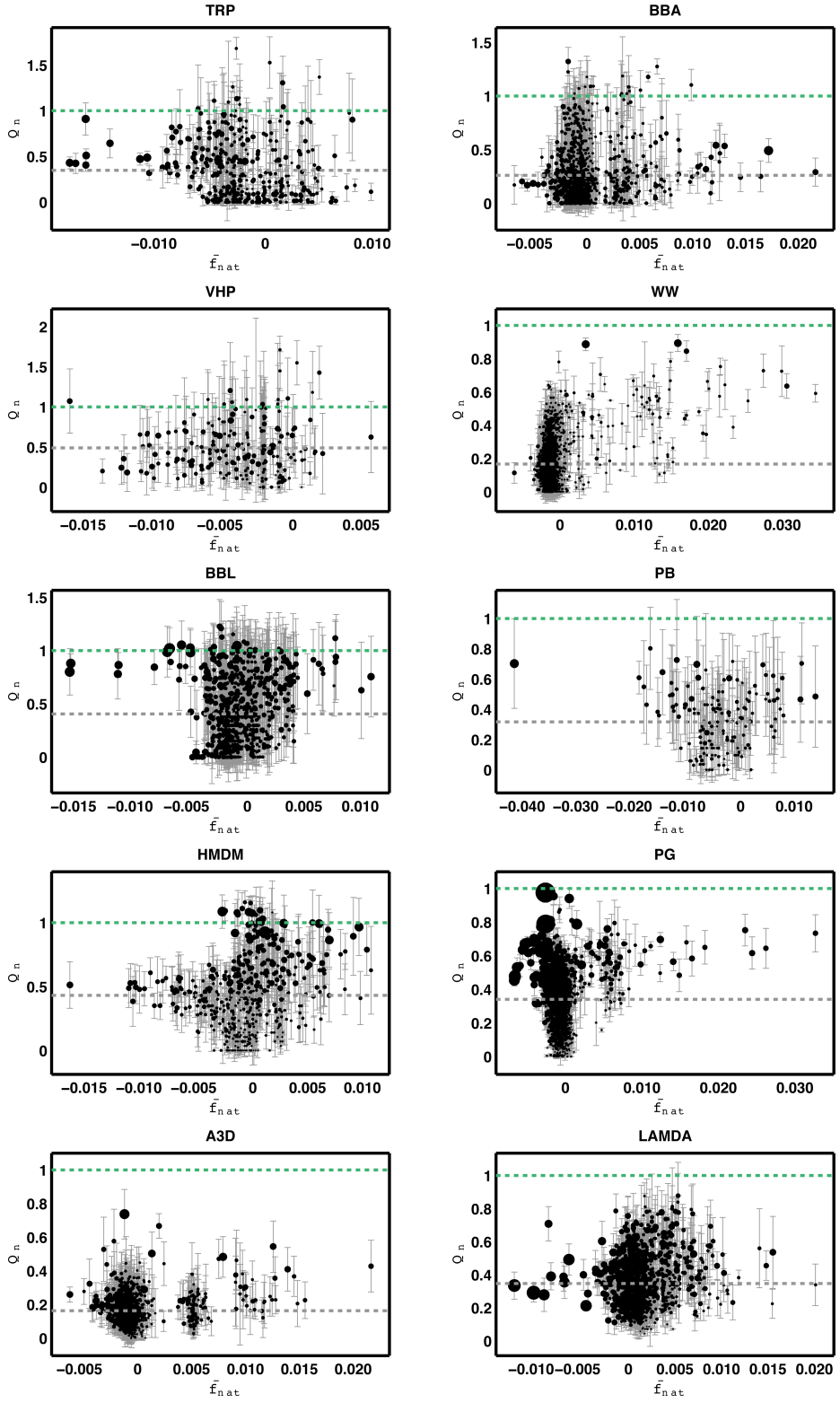


Figure A.6: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus native contacts, Q_{n}. The green (gray) dashed line indicates the median value within the native (nonnative) ensemble.

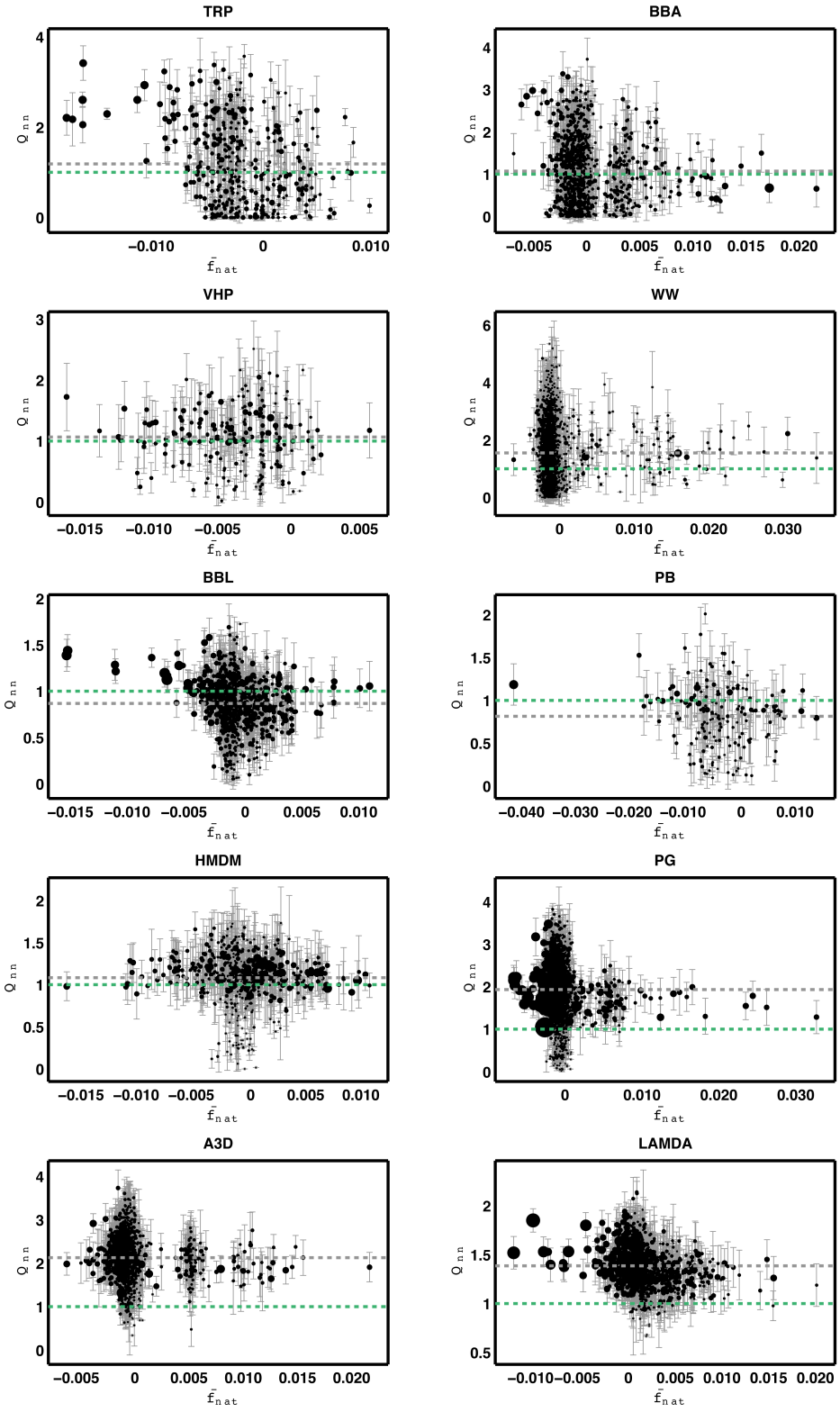


Figure A.7: Frustration scores, $\bar{f}_{\mathrm{nat}}$, versus nonnative contacts, Q_{nn}. The green (gray) dashed line indicates the median value within the native (nonnative) ensemble.

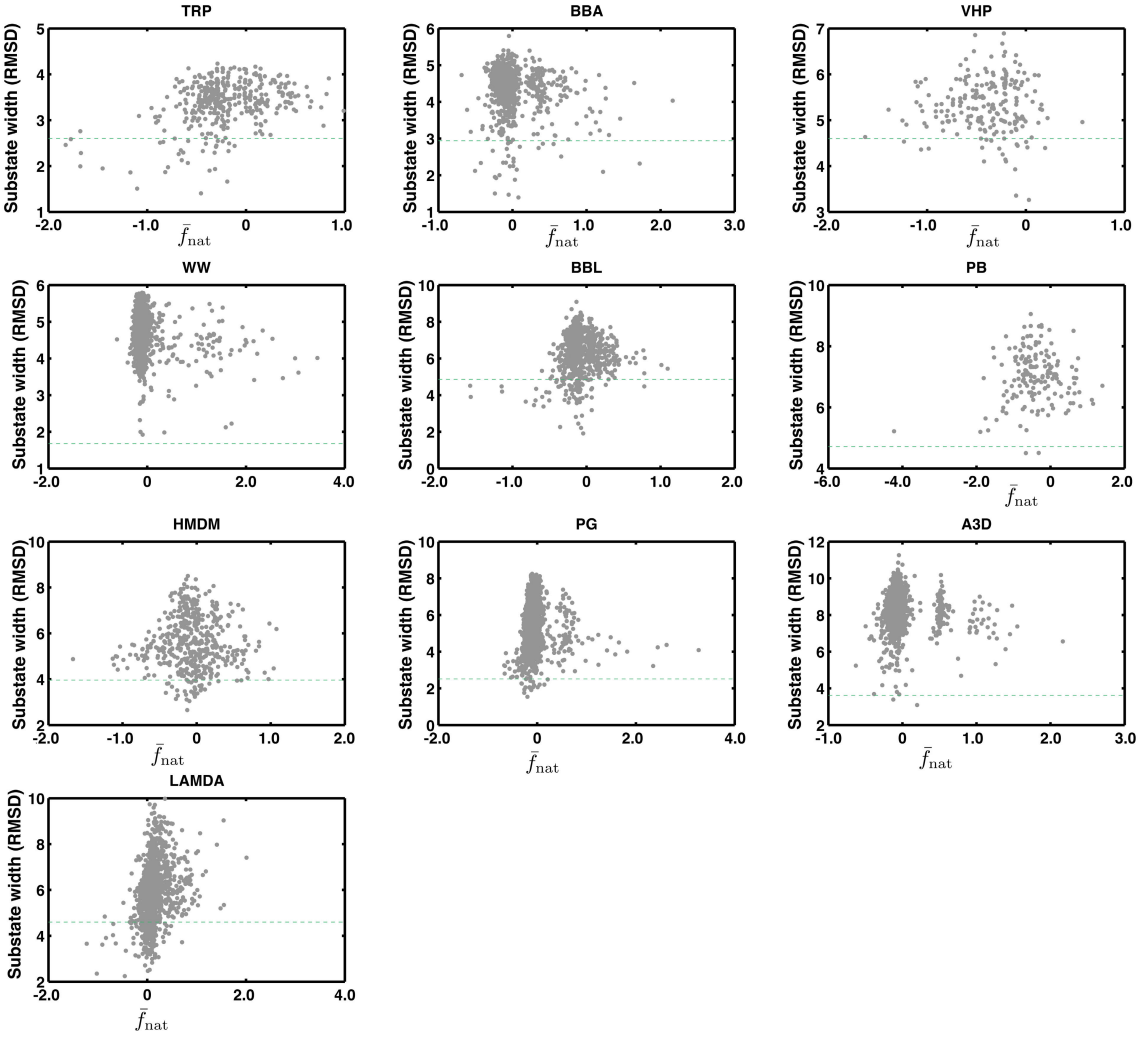


Figure A.8: Substate widths. Frustration scores, $\bar{f}_{\mathrm{nat}}$, plotted against substate widths, defined as average intra-substate pairwise RMSD. The green dashed line indicates the median cluster width of all conformational substates within the native ensemble. Singletons are excluded.

References

[1] Jordan Ellenberg. *How not to be wrong : the power of mathematical thinking*. The Penguin Press, New York, 2014. (page 1).

[2] C Levinthal. **Are there pathways for protein folding**. *J. Chim. Phys*, **65**(1):44–45, 1968. (page 5).

[3] C B Anfinsen. **Principles that govern the folding of protein chains**. *Science*, **181**(96):223–230, July 1973. (page 5).

[4] Alan R Fersht. **On the simulation of protein folding by short time scale molecular dynamics and distributed computing.** *Proceedings of the National Academy of Sciences of the United States of America*, **99**(22):14122–14125, October 2002. (page 5).

[5] Vijay S Pande, Ian Baker, Jarrod Chapman, Sidney P Elmer, Siraj Khaliq, Stefan M Larson, Young Min Rhee, Michael R Shirts, Christopher D Snow, Eric J Sorin, and Bojan Zagrovic. **Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing.** *Biopolymers*, **68**(1):91–109, January 2003. (page 5).

[6] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. **How Fast-Folding Proteins Fold**. *Science*, **334**(6055):517–520, October 2011. (pages 5, 6, 11, 12, 16).

[7] Peter L Freddolino and Klaus Schulten. **Common Structural Transitions in Explicit-Solvent Simulations of Villin Headpiece Folding**. *Biophysical Journal*, **97**(8):2338–2347, October 2009. (pages 5, 6).

[8] Robert B Best. **Atomistic molecular simulations of protein folding.** *Current opinion in structural biology*, **22**(1):52–61, February 2012. (page 5).

[9] H Frauenfelder, S G Sligar, and PG G Wolynes. **The energy landscapes and motions of proteins.** *Science*, **254**(5038):1598–1603, December 1991. (pages 5, 61, 65).

[10] H Frauenfelder, F Parak, and R D Young. **Conformational substates in proteins.** *Annual review of biophysics and biophysical chemistry*, **17**:451–479, 1988. (pages 5, 61, 65).

[11] H S Chan and Ken A Dill. **Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics.** *Proteins*, **30**(1):2–33, January 1998. (page 6).

[12] Martin Karplus. **Behind the folding funnel diagram.** *Nature Chemical Biology*, **7**(7):401–404, July 2011. (page 6).

[13] Justin Spiriti and Daniel M Zuckerman. **Tunable Coarse Graining for Monte Carlo Simulations of Proteins via Smoothed Energy Tables: Direct and Exchange Simulations.** *Journal of Chemical Theory and Computation*, **10**(11):5161–5177, November 2014. (page 6).

[14] Yuzo Ueda, Hiroshi Taketomi, and N Gō. **Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A Three dimensional lattice model of lysozyme**. *Biopolymers*, **17**(6):1531–1548, 1978. (page 6).

[15] Steven S Plotkin. **Speeding protein folding beyond the Gō model: How a little frustration sometimes helps**. *Proteins*, **45**(4):337–345, 2001. (page 6).

[16] Mark T Oakley, David J Wales, and Roy L Johnston. **The Effect of Nonnative Interactions on the Energy Landscapes of Frustrated Model Proteins**. *Journal of Atomic, Molecular, and Optical Physics*, **2012**(4096):1–9, 2012. (page 6).

[17] Hongxing Lei, Chun Wu, Haiguang Liu, and Yong Duan. **Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations.** *Proceedings of the National Academy of Sciences of the United States of America*, **104**(12):4925–4930, March 2007. (page 6).

[18] Cecilia Clementi and Steven S Plotkin. **The effects of nonnative interactions on protein folding rates: Theory and simulation**. *Protein science*, **13**(7):1750–1766, July 2004. (page 6).

[19] Steven Hayward and E James Milner-White. **The geometry of α-sheet: Implications for its possible function as amyloid precursor in proteins**. *Proteins*, **71**(1):415–425, 2008. (page 6).

[20] R V Pappu, R Srinivasan, and G D Rose. **The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding.** *Proceedings of the National Academy of Sciences of the United States of America*, **97**(23):12565–12570, November 2000. (pages 6, 80).

[21] P Neudecker, P Robustelli, A Cavalli, P Walsh, P Lundstrom, A Zarrine-Afsar, S Sharpe, M Vendruscolo, and L E Kay. **Structure of an Intermediate State in Protein Folding and Aggregation**. *Science*, **336**(6079):362–366, April 2012. (pages 6, 34).

[22] Weihong Zhang, Debabani Ganguly, and Jianhan Chen. **Residual Structures, Conformational Fluctuations, and Electrostatic Interactions in the Synergistic Folding of Two Intrinsically Disordered Proteins**. *PLoS computational biology*, **8**(1):e1002353, January 2012. (pages 6, 79).

[23] Jonathan E Kohn, Ian S Millett, Jaby Jacob, Bojan Zagrovic, Thomas M Dillon, Nikolina Cingel, Robin S Dothager, Soenke Seifert, P Thiyagarajan, Tobin R Sosnick, M Zahid Hasan, Vijay S Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W Plaxco. **Random-coil behavior and the dimensions of chemically unfolded proteins.** *Proceedings of the National Academy of Sciences of the United States of America*, **101**(34):12491–12496, August 2004. (page 6).

[24] Ian S Millett, Sebastian Doniach, and Kevin W Plaxco. **Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins.** *Advances in protein chemistry*, **62**:241–262, 2002. (page 6).

[25] Michael Knott and Robert B Best. **A Preformed Binding Interface in the Unbound Ensemble of an Intrinsically Disordered Protein: Evidence from Molecular Simulations**. *PLoS computational biology*, **8**(7):e1002605, July 2012. (page 6).

[26] Per Rogne, Przemysław Ozdowy, Christian Richter, Krishna Saxena, Harald Schwalbe, and Lars T Kuhn. **Atomic-Level Structure Characterization of an Ultrafast Folding Mini-Protein Denatured State**. *PLoS ONE*, **7**(7):e41301, July 2012. (page 6).

[27] Hongxing Lei, Yao Su, Lian Jin, and Yong Duan. **Folding network of villin headpiece subdomain.** *Biophysical Journal*, **99**(10):3374–3384, November 2010. (page 6).

[28] Oxana V Galzitskaya and Anna V Glyakina. **Nucleation-based prediction of the protein folding rate and its correlation with the folding nucleus size.** *Proteins*, **80**(12):2711–2727, December 2012. (page 6).

[29] Gregory R Bowman and Vijay S Pande. **Protein folded states are kinetic hubs.** *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24):10890–10895, June 2010. (pages 6, 61, 63).

[30] Alex Dickson and Charles Brooks, III. **Native States of Fast-Folding Proteins Are Kinetic Traps**. *Journal of the American Chemical Society*, **135**(12):4729–4734, March 2013. (pages 6, 11, 12, 13, 17, 33, 37).

[31] N Gō. **Theoretical studies of protein folding.** *Annual review of biophysics and bioengineering*, **12**:183–210, 1983. (page 6).

[32] J D Bryngelson, José N Onuchic, N D Socci, and PG G Wolynes. **Funnels, pathways, and the energy landscape of protein folding: a synthesis.** *Proteins*, **21**(3):167–195, March 1995. (page 6).

[33] Joan-Emma Shea, José N Onuchic, and Charles Brooks, III. **Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A.** *Proceedings of the National Academy of Sciences of the United States of America*, **96**(22):12512–12517, October 1999. (page 6).

[34] Robert B Best, Gerhard Hummer, and William A Eaton. **Native contacts determine protein folding mechanisms in atomistic simulations.** *Proceedings of the National Academy of Sciences of the United States of America*, **110**(44):17874–17879, October 2013. (pages 6, 11).

[35] P S Kim and R L Baldwin. **Intermediates in the folding reactions of small proteins.** *Annual Review of Biochemistry*, **59**:631–660, 1990. (page 6).

[36] Cecilia Clementi, H Nymeyer, and José N Onuchic. **Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins.** *Journal of Molecular Biology*, **298**(5):937–953, May 2000. (page 6).

[37] A. Jain, R Hegger, and G Stock. **Hidden Complexity of Protein Free-Energy Landscapes Revealed by Principal Component Analysis by Parts**. *The Journal of Physical Chemistry Letters*, **1**:2769–2773, 2010. (page 6).

[38] Debabani Ganguly, Weihong Zhang, and Jianhan Chen. **Synergistic folding of two intrinsically disordered proteins: searching for conformational selection.** *Molecular bioSystems*, **8**(1):198–209, January 2012. (page 6).

[39] Tsung-Han Chiang, David Hsu, and Jean-Claude Latombe. **Markov dynamic models for long-timescale protein motion.** *Bioinformatics*, **26**(12):i269–77, June 2010. (pages 8, 63).

[40] John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope. **Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics.** *The Journal of Chemical Physics*, **126**(15):155101, April 2007. (pages 8, 18, 63).

[41] J G Kemeny and James Laurie Snell. *Finite Markov Chains*. Springer Verlag, New York, 1976. (pages 8, 16, 18, 38, 41).

[42] Zhongzhi Zhang, Tong Shan, and Guanrong Chen. **Random walks on weighted networks**. *Physical Review E*, **87**(1):012112, January 2013. (page 8).

[43] Zhongzhi Zhang, Alafate Julaiti, Baoyu Hou, Hongjuan Zhang, and Guanrong Chen. **Mean first-passage time for random walks on undirected networks**. *The European Physical Journal B*, **84**(4):691–697, 2011. (pages 8, 38, 41).

[44] Nina Singhal Hinrichs. *Algorithms for Building Models of Molecular Motion from Simulations*. PhD thesis, Stanford, September 2007. (page 8).

[45] Jeffrey K Weber and Vijay S Pande. **Protein Folding Is Mechanistically Robust**. *Biophysical Journal*, **102**(4):859–867, February 2012. (page 8).

[46] Amandeep K Sangha and T Keyes. **Proteins fold by subdiffusion of the order parameter.** *Journal of Physical Chemistry B*, **113**(48):15886–15894, December 2009. (page 8).

[47] Ronaldo J Oliveira, Paul C Whitford, Jorge Chahine, Jin Wang, José N Onuchic, and Vitor B P Leite. **The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding.** *Biophysical Journal*, **99**(2):600–608, July 2010. (page 8).

[48] Troy Cellmer, Eric R Henry, James Hofrichter, and William A Eaton. **Measuring internal friction of an ultrafast-folding protein.** *Proceedings of the National Academy of Sciences of the United States of America*, **105**(47):18320–18325, November 2008. (page 8).

[49] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. **Hitting and commute times in large graphs are often misleading**. *arXiv preprint arXiv:1003.1266*, 2010. (pages 9, 30).

[50] Alex Dickson and Charles Brooks, III. **Quantifying hub-like behavior in protein folding networks.** *Journal of Chemical Theory and Computation*, **8**(9):3044–3052, 2012. (pages 10, 18, 37, 41).

[51] Loyd Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997. (pages 10, 47).

[52] P Ferrara, J Apostolakis, and Amedeo Caflisch. **Computer simulations of protein folding by targeted molecular dynamics.** *Proteins*, **39**(3):252–260, May 2000. (page 10).

[53] Hyunbum Jang, Carol K Hall, and Yaoqi Zhou. **Protein folding pathways and kinetics: molecular dynamics simulations of beta-strand motifs.** *Biophysical Journal*, **83**(2):819–835, August 2002. (page 10).

[54] Amedeo Caflisch and Peter Hamm. **Complexity in Protein Folding: Simulation Meets Experiment**. *Current Physical Chemistry*, **2**(1):4–11, 2012. (pages 10, 37).

[55] Daniel-Adriano Silva, Gregory R Bowman, Alejandro Sosa-Peinado, and Xuhui Huang. **A role for both conformational selection and induced fit in ligand binding by the LAO protein.** *PLoS computational biology*, **7**(5):e1002054, May 2011. (page 10).

[56] Nan-jie Deng, Weihua Zheng, Emillio Gallicchio, and Ronald M Levy. **Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations**. *Journal of the American Chemical Society*, **133**(24):9387–9394, June 2011. (page 10).

[57] Beth G Wensley, Sarah Batey, Fleur A C Bone, Zheng Ming Chan, Nuala R Tumelty, Annette Steward, Lee Gyan Kwa, Alessandro Borgia, and Jane Clarke. **Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family**. *Nature*, **463**(7281):685–688, April 2010. (page 11).

[58] Dalit Shental-Bechor, Martin T J Smith, Duncan Mackenzie, Aron Broom, Amir Marcovitz, Fadila Ghashut, Chris Go, Fernando Bralha, Elizabeth M Meiering, and Yaakov Levy. **Non-native interactions regulate folding and switching of myristoylated protein.** *Proceedings of the National Academy of Sciences of the United States of America*, **109**(44):17839–17844, October 2012. (pages 11, 30).

[59] Yunxiang Sun and Dengming Ming. **Energetic Frustrations in Protein Folding at Residue Resolution: A Homologous Simulation Study of Im9 Proteins**. *PLoS ONE*, **9**(1):e87719, January 2014. (page 11).

[60] Aitziber L Cortajarena, Fang Yi, and Lynne Regan. **Designed TPR modules as novel anticancer agents.** *ACS chemical biology*, **3**(3):161–166, March 2008. (page 11).

[61] Gerard Said, Seden Grippon, and Peter Kirkpatrick. **Tafamidis.** *Nature Reviews Drug Discovery*, **11**(3):185–186, March 2012. (page 11).

[62] Glenn L Butterfoss and Brian Kuhlman. **Computer-based design of novel protein structures.** *Annual review of biophysics and biomolecular structure*, **35**:49–65, 2006. (page 11).

[63] Lukasz A Joachimiak, Tanja Kortemme, Barry L Stoddard, and David Baker. **Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface**. *Journal of Molecular Biology*, **361**(1):195–208, August 2006. (page 11).

[64] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. **Protein folding kinetics and thermodynamics from atomistic simulation.** *Proceedings of the National Academy of Sciences of the United States of America*, **109**(44):17845–17850, October 2012. (page 11).

[65] Eric R Henry, Robert B Best, and William A Eaton. **Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations.** *Proceedings of the National Academy of Sciences of the United States of America*, **110**(44):17880–17885, October 2013. (page 11).

[66] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. **How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?** *Biophysical Journal*, **100**(9):L47–L49, May 2011. (page 11).

[67] Kyle A Beauchamp, Gregory R Bowman, Thomas J Lane, Lutz Maibaum, Imran S Haque, and Vijay S Pande. **MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale.** *Journal of Chemical Theory and Computation*, **7**(10):3412–3419, October 2011. (pages 13, 14).

[68] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. **Fast unfolding of communities in large networks**. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10):P10008, October 2008. (pages 14, 16).

[69] Joe H Ward, Jr. **Hierarchical grouping to optimize an objective function**. *Journal of the American statistical association*, **58**(301):236–244, 1963. (page 13).

[70] Kyle A Beauchamp, Robert McGibbon, Yu-Shan Lin, and Vijay S Pande. **Simple few-state models reveal hidden complexity in protein folding.** *Proceedings of the National Academy of Sciences of the United States of America*, **109**(44):17807–17813, October 2012. (page 13).

[71] Vijay S Pande, Kyle A Beauchamp, and Gregory R Bowman. **Everything you wanted to know about Markov State Models but were afraid to ask.** *Methods*, **52**(1):99–105, September 2010. (page 13).

[72] Guha Jayachandran, V Vishal, and Vijay S Pande. **Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece**. *The Journal of Chemical Physics*, **124**(16):164902, 2006. (page 16).

[73] Jure Leskovec, Kevin J Lang, and Michael Mahoney. **Empirical comparison of algorithms for network community detection**. *Proceedings of the 19th international conference on World Wide Web Conference Committee*, pages 631–640, 2010. (page 16).

[74] Pavel I Zhuravlev, Christopher Kroboth Materese, and Garegin A Papoian. **Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins.** *Journal of Physical Chemistry B*, **113**(26):8800–8812, July 2009. (page 17).

[75] Charles K Fisher and Collin M Stultz. **Protein structure along the order-disorder continuum.** *Journal of the American Chemical Society*, **133**(26):10022–10025, July 2011. (page 17).

[76] H Nymeyer, A E García, and José N Onuchic. **Folding funnels and frustration in off-lattice minimalist protein landscapes.** *Proceedings of the National Academy of Sciences of the United States of America*, **95**(11):5921–5928, May 1998. (page 18).

[77] Yaakov Levy, PG G Wolynes, and José N Onuchic. **Protein topology determines binding mechanism**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(2):511–516, January 2004. (page 18).

[78] Joachim Lätzer, Tongye Shen, and PG G Wolynes. **Conformational Switching upon Phosphorylation: A Predictive Framework Based on Energy Landscape Principles**. *Biochemistry*, **47**(7):2110–2122, February 2008. (page 18).

[79] Rose Du, Vijay S Pande, Alexander Yu Grosberg, Toyoichi Tanaka, and Eugene S Shakhnovich. **On the transition coordinate for protein folding**. *The Journal of Chemical Physics*, **108**(1):334, 1998. (page 18).

[80] Rudesh D Toofanny, Amanda L Jonsson, and Valerie Daggett. **A comprehensive multidimensional-embedded, one-dimensional reaction coordinate for protein unfolding/folding.** *Biophysical Journal*, **98**(11):2671–2681, June 2010. (page 18).

[81] Gilles Labesse, N Colloc'h, Joël Pothier, and J-P Mornon. **P-SEA: A new efficient assignment of secondary structure from Cα trace of proteins**. *Computer applications in the biosciences*, **13**(3):291–295, 1997. (pages 18, 20).

[82] Mieczyslaw Torchala, Przemyslaw Chelminiak, Michal Kurzynski, and Paul A Bates. **RaTrav: a tool for calculating mean first-passage times on biochemical networks.** *BMC systems biology*, **7**:130, 2013. (page 18).

[83] Gregory R Bowman, Kyle A Beauchamp, G Boxer, and Vijay S Pande. **Progress and challenges in the automated construction of Markov state models for full protein systems**. *The Journal of Chemical Physics*, **131**:124101, 2009. (page 18).

[84] MATLAB. *version 7.14.0.739 (R2012a)*. The MathWorks Inc., Natick, Massachusetts. (pages 21, 50).

[85] Jeffrey W Peng. **Modeling conformational ensembles of slow functional motions in Pin1-WW.** *PLoS computational biology*, **6**(12):e1001015, 2010. (page 22).

[86] John D Chodera and Vijay S Pande. **The social network (of protein conformations).** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(32):12969–12970, August 2011. (pages 22, 36).

[87] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. **Gephi: an open source software for exploring and manipulating networks.** *ICWSM*, pages 361–362, 2009. (pages 22, 47, 50).

[88] Rebeca García-Fandiño, Pau Bernadó, Sara Ayuso-Tejedor, Javier Sancho, and Modesto Orozco. **Defining the Nature of Thermal Intermediate in 3 State Folding Proteins: Apoflavodoxin, a Study Case**. *PLoS computational biology*, **8**(8):e1002647, August 2012. (pages 23, 24).

[89] S S Cho. **P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes**. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(3):586–591, January 2006. (page 24).

[90] Oxana V Galzitskaya and A V Finkelstein. **A theoretical search for folding/unfolding nuclei in three-dimensional protein structures.** *Proceedings of the National Academy of Sciences of the United States of America*, **96**(20):11299–11304, September 1999. (page 24).

[91] Maksym Tsytlonok and Laura S Itzhaki. **The how's and why's of protein folding intermediates.** *Archives of Biochemistry and Biophysics*, October 2012. (page 25).

[92] Joanna I Sulkowska, Jeffrey K Noel, and José N Onuchic. **Energy landscape of knotted protein folding.** *Proceedings of the National Academy of Sciences of the United States of America*, **109**(44):17783–17788, October 2012. (page 25).

[93] F Pozzi, T Di Matteo, and T Aste. **Exponential smoothing weighted correlations - Springer**. *The European Physical Journal B*, 2012. (page 28).

[94] PyMOL. *The PyMOL Molecular Graphics System*. Schrodinger LLC, April 2010. (page 29).

[95] Patrícia F N Faísca, Ana Nunes, Rui D M Travasso, and Eugene S Shakhnovich. **Non-native interactions play an effective role in protein folding dynamics.** *Protein science*, **19**(11):2196–2209, November 2010. (page 29).

[96] Lev Muchnik. *Complex Networks Package for MatLab (Version 1.6)*. www.levmuchnik.net. (pages 32, 50).

[97] Carlo Camilloni, Daniel Schaal, Kristian Schweimer, Stephan Schwarzinger, and Alfonso De Simone. **Energy landscape of the prion protein helix 1 probed by metadynamics and NMR.** *Biophysical Journal*, **102**(1):158–167, January 2012. (page 30).

[98] Yantao Chen and Jiandong Ding. **Roles of non-native hydrogen-bonding interaction in helix-coil transition of a single polypeptide as revealed by comparison between Gō-like and non-Gō models.** *Proteins*, **78**(9):2090–2100, July 2010. (page 30).

[99] Allan Chris M Ferreon, Crystal R Moran, Josephine C Ferreon, and Ashok A Deniz. **Alteration of the α-Synuclein Folding Landscape by a Mutation Related to Parkinson's Disease**. *Angewandte Chemie International Edition*, **49**(20):3469–3472, April 2010. (page 30).

[100] M M Gromiha and S Selvaraj. **Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction.** *Journal of Molecular Biology*, **310**(1):27–32, June 2001. (page 30).

[101] Robert B Best and Gerhard Hummer. **Reaction coordinates and rates from transition paths.** *Proceedings of the National Academy of Sciences of the United States of America*, **102**(19):6732–6737, May 2005. (page 33).

[102] Vikram Khipple Mulligan and Avijit Chakrabartty. **Protein misfolding in the late-onset neurodegenerative diseases: Common themes and the unique case of amyotrophic lateral sclerosis**. *Proteins*, **81**(8):1285–1303, 2013. (page 34).

[103] Christopher B Stanley, Tatiana Perevozchikova, and Valerie Berthelier. **Structural Formation of Huntingtin Exon 1 Aggregates Probed by Small-Angle Neutron Scattering**. *Biophysical Journal*, **100**(10):2504–2512, May 2011. (page 34).

[104] Christopher M Dobson and Martin Karplus. **The fundamentals of protein folding: bringing together theory and experiment.** *Current opinion in structural biology*, **9**(1):92–101, February 1999. (page 34).

[105] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P. Eastwood, Joseph A Bank, John M Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. **Atomic-level characterization of the structural dynamics of proteins.** *Science*, **330**(6002):341–346, October 2010. (page 34).

[106] Gloria Fuentes, Shubhra Ghosh Dastidar, Arumugam Madhumalar, and Chandra S Verma. **Role of protein flexibility in the discovery of new drugs**. *Drug Development Research*, **72**(1):26–35, 2011. (page 34).

[107] William Sinko, César de Oliveira, Sarah Williams, Adam Van Wynsberghe, Jacob D Durrant, Rong Cao, Eric Oldfield, and J Andrew McCammon. **Applying Molecular Dynamics Simulations to Identify Rarely Sampled Ligand-bound Conformational States of Undecaprenyl Pyrophosphate Synthase, an Antibacterial Target**. *Chemical Biology & Drug Design*, **77**(6):412–420, March 2011. (pages 34, 80).

[108] Yanping Yin, Gia G Maisuradze, Adam Liwo, and Harold A Scheraga. **Hidden Protein Folding Pathways in Free-Energy Landscapes Uncovered by Network Analysis**. *Journal of Chemical Theory and Computation*, **8**(4):1176–1189, April 2012. (page 35).

[109] Naresh Chennamsetty, Vladimir Voynov, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. **Design of therapeutic proteins with enhanced stability.** *Proceedings of the National Academy of Sciences*, **106**(29):11937–11942, July 2009. (page 35).

[110] Réka Albert and Albert-Laszlo Barabasi. **Statistical mechanics of complex networks**. *Reviews of modern physics*, **74**(1):47, 2002. (page 36).

[111] Naoki Masuda. **Immunization of networks with community structure**. *New Journal of Physics*, **11**(12):123018, 2009. (page 36).

[112] Piet Van Mieghem. **Epidemic phase transition of the SIS type in networks**. *EPL (Europhysics Letters)*, **97**(4):48004, 2012. (page 36).

[113] Ed Bullmore and Olaf Sporns. **Complex brain networks: graph theoretical analysis of structural and functional systems**. *Nature Reviews Neuroscience*, **10**(3):186–198, February 2009. (page 36).

[114] A T Lawniczak, A Gerisch, and K Maxie. **Effects of randomly added links on a phase transition in data network traffic models**. *Proc of the 3rd International DCDIS Conference*, 2003. (page 36).

[115] Giuliano Andrea Pagani and Marco Aiello. **The Power Grid as a complex network: A survey**. *Physica A: Statistical Mechanics and its Applications*, **392**(11):2688–2700, June 2013. (page 36).

[116] D J Watts and S H Strogatz. **Collective dynamics of 'small-world' networks.** *Nature*, **393**(6684):440–442, June 1998. (page 36).

[117] H Jeong, S P Mason, Albert-Laszlo Barabasi, and Z N Oltvai. **Lethality and centrality in protein networks.** *Nature*, **411**(6833):41–42, May 2001. (page 36).

[118] B A Prakash, J Vreeken, and C Faloutsos. **Efficiently spotting the starting points of an epidemic in a large graph**. *Knowledge and information systems*, 2014. (pages 36, 37).

[119] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008. (page 36).

[120] Deepak Mangal, Niladri Sett, Sanasam Ranbir Singh, and Sukumar Nandi. **Link prediction on evolving social network using spectral analysis**. *2013 IEEE International Conference on Advanced Networks and Telecommuncations Systems*, pages 1–6, 2013. (page 36).

[121] Hui Wang, Jinyuan Huang, Xiaomin Xu, and Yanghua Xiao. **Damage attack on complex networks**. *Physica A: Statistical Mechanics and its Applications*, pages 1–15, April 2014. (page 36).

[122] Ricardo Gutiérrez, Irene Sendiña-Nadal, Massimiliano Zanin, David Papo, and Stefano Boccaletti. **Targeting the dynamics of complex networks.** *Scientific Reports*, **2**:396–396, January 2012. (page 36).

[123] Christian M Schneider, André A Moreira, José S Andrade, Shlomo Havlin, and Hans J Herrmann. **Mitigation of malicious attacks on networks**. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(10):3838–3841, 2011. (page 36).

[124] B Wang, H W Tang, C H Guo, Z L Xiu, and T Zhou. **Optimization of network structure to random failures**. *Physica A: Statistical Mechanics and its Applications*, **368**(2):607–614, 2006. (page 36).

[125] Benjamin Shargel, Hiroki Sayama, Irving R Epstein, and Yaneer Bar-Yam. **Optimization of robustness and connectivity in complex networks.** *Physical Review Letters*, **90**(6):068701, February 2003. (page 36).

[126] Venky Soundararajan and Murali Aravamudan. **Global connectivity of hub residues in Oncoprotein structures encodes genetic factors dictating personalized drug response to targeted Cancer therapy**. *Scientific Reports*, **4**:7294, December 2014. (page 37).

[127] Michele Benzi and Christine Klymko. **A matrix analysis of different centrality measures**. *arXiv preprint arXiv:1312.6722*, 2013. (page 37).

[128] Gergana Bounova and Olivier de Weck. **Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles**. *Physical Review E*, **85**(1):016117, January 2012. (page 37).

[129] Konstantin Klemm, M Ángeles Serrano, Víctor M Eguíluz, and Maxi San Miguel. **A measure of individual role in collective dynamics**. *Scientific Reports*, **2**, February 2012. (page 37).

[130] Ernesto Estrada and Naomichi Hatano. **A vibrational approach to node centrality and vulnerability in complex networks**. *Physica A: Statistical Mechanics and its Applications*, **389**(17):3648–3660, September 2010. (page 37).

[131] Piet Van Mieghem. **Graph eigenvectors, fundamental weights and centrality metrics for nodes in networks**. *arXiv preprint arXiv:1401.4580*, 2014. (pages 37, 43, 55).

[132] Gitanjali Yadav and Suresh Babu. **NEXCADE: Perturbation Analysis for Complex Networks**. *PLoS ONE*, **7**(8):e41827, August 2012. (page 37).

[133] Andrew Y Ng, Alice X Zheng, and Michael I Jordan. **Link analysis, eigenvectors and stability**. *International Joint Conference on Artificial Intelligence*, **17**(1):903–910, 2001. (page 37).

[134] Juan Chen, Jun-an Lu, Choujun Zhan, and Guanrong Chen. **Laplacian spectra and synchronization processes on complex networks**. In *Handbook of Optimization in Complex Networks*, pages 81–113. Springer, 2012. (page 37).

[135] Remi Monasson. **Diffusion, localization and dispersion relations on "small-world" lattices**. *The European Physical Journal B-Condensed Matter and Complex Systems*, **12**(4):555–567, 1999. (page 37).

[136] Patrick N McGraw and Michael Menzinger. **Laplacian spectra as a diagnostic tool for network structure and dynamics**. *Physical Review E*, **77**(3):031102, March 2008. (page 37).

[137] J Martin Hernandez. *Measuring Robustness of Complex Networks*. PhD thesis, Delft University of Technology, September 2013. (page 37).

[138] Gourab Ghoshal and Albert-Laszlo Barabasi. **Ranking stability and super-stable nodes in complex networks**. *Nature Communications*, **2**:392–7, July 2011. (page 37).

[139] Scott D Pauls and Daniel Remondini. **Measures of centrality based on the spectrum of the Laplacian**. *Physical Review E*, **85**(6):066127, June 2012. (page 37).

[140] S Navlakha and C Kingsford. **The power of protein interaction networks for associating genes with diseases**. *Bioinformatics*, **26**(8):1057–1063, April 2010. (page 37).

[141] C Li, H Wang, W de Haan, C J Stam, and Piet Van Mieghem. **The correlation of metrics in complex networks with applications in functional brain networks**. *Journal of Statistical Mechanics: Theory and Experiment*, **2011**(11):P11018, 2011. (page 37).

[142] L da F Costa, F A Rodrigues, G Travieso, and P R Villas Boas. **Characterization of complex networks: A survey of measurements**. *Advances in Physics*, **56**(1):167–242, January 2007. (page 37).

[143] Hongxiao Liu and Zhongzhi Zhang. **Laplacian spectra of recursive treelike small-world polymer networks: Analytical solutions and applications**. *The Journal of Chemical Physics*, **138**(11):114904, 2013. (pages 37, 38, 44).

[144] Andrej Savol and Chakra S Chennubhotla. **Quantifying the Sources of Kinetic Frustration in Folding Simulations of Small Proteins**. *Journal of Chemical Theory and Computation*, **10**(8):2964–2974, August 2014. (page 37).

[145] Toru Ohira and Ryusuke Sawatari. **Phase transition in a computer network traffic model**. *Physical Review E*, **58**(1):193, 1998. (page 37).

[146] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. **Identification of influential spreaders in complex networks**. *Nature Physics*, **6**(11):888–893, August 2010. (pages 37, 80).

[147] Peter G Doyle and James Laurie Snell. *Random Walks and Electric Networks*. Carus Monographs. Mathematical Association of America, Washington, 1984. (page 38).

[148] Murray Shanahan. **Metastable chimera states in community-structured oscillator networks.** *Chaos*, **20**(1):013108–013108, March 2010. (page 38).

[149] Pablo Villegas, Paolo Moretti, and Miguel A Muñoz. **Frustrated hierarchical synchronization and emergent complexity in the human connectome network.** *Scientific Reports*, **4**:5990–5990, January 2014. (page 38).

[150] Eleanor R Brush, David C Krakauer, and Jessica C Flack. **A Family of Algorithms for Computing Consensus about Node State from Network Data**. *PLoS computational biology*, **9**(7):e1003109, July 2013. (page 38).

[151] Yuan Lin and Zhongzhi Zhang. **Random walks in weighted networks with a perfect trap: An application of Laplacian spectra.** *Physical Review E*, **87**(6-1):062140, June 2013. (pages 41, 44).

[152] David Gfeller and Paolo De Los Rios. **Spectral coarse graining and synchronization in oscillator networks.** *Physical Review Letters*, **100**(17):174104–174104, May 2008. (pages 43, 59).

[153] Stéphane S Lafon and Ann B AB Lee. **Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(9):1393–1403, September 2006. (page 43).

[154] Ulrike von Luxburg. **A tutorial on spectral clustering**. *Statistics and Computing*, **17**(4):395–416, August 2007. (page 43).

[155] Dilip Krishnan, Raanan Fattal, and Richard Szeliski. **Efficient preconditioning of laplacian matrices for computer graphics**. *ACM Transactions on Graphics*, **32**(4):1, July 2013. (page 43).

[156] Nina Singhal Hinrichs and Vijay S Pande. **Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics**. *The Journal of Chemical Physics*, **126**(24):244101, 2007. (page 43).

[157] Jérôme Kunegis and Andreas Lommatzsch. **Learning spectral graph transformations for link prediction**. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1–8, Montreal, 2009. ACM Press. (page 43).

[158] Piet Van Mieghem, Dragan Stevanović, Fernando Kuipers, Cong Li, Ruud van de Bovenkamp, Daijie Liu, and Huijuan Wang. **Decreasing the spectral radius of a graph by link removals.** *Physical Review E*, **84**(1 Pt 2):016101, July 2011. (pages 43, 59).

[159] Mauricio Barahona and Louis M Pecora. **Synchronization in small-world systems.** *Physical Review Letters*, **89**(5):054101, July 2002. (page 43).

[160] Alexander C Kalloniatis. **From incoherence to synchronicity in the network Kuramoto model.** *Physical Review E*, **82**(6 Pt 2):066202–066202, December 2010. (page 43).

[161] Phillip Bonacich. **Factoring and weighting approaches to status scores and clique identification**. *Journal of Mathematical Sociology*, **2**(1):113–120, 1972. (page 43).

[162] Chakra S Chennubhotla and Allan D Jepson. **Hierarchical eigensolver for transition matrices in spectral methods**. *Advances in Neural Information Processing Systems*, **17**:273–280, 2005. (pages 43, 69).

[163] John C Urschel, Xiaozhe Hu, Jinchao Xu, and Ludmil T Zikatanov. **A Cascadic Multigrid Algorithm for Computing the Fiedler Vector of Graph Laplacians**. *arXiv:1412.0565*, December 2014. (page 43).

[164] Yong Liu and Shizhong Liao. **An Error Bound for Eigenvalues of Graph Laplacian with Bounded Kernel Function**. *2011 Seventh International Conference on Computational Intelligence and Security*, pages 436–440, December 2011. (page 44).

[165] Steve Butler. **Interlacing for weighted graphs using the normalized Laplacian**. *Electronic Journal of Linear Algebra*, **16**(90-98):87, 2007. (page 44).

[166] Aida Abiad, Miquel A Fiol, Willem H Haemers, and Guillem Perarnau. **An interlacing approach for bounding the sum of Laplacian eigenvalues of graphs**. *Linear Algebra and its Applications*, **448**:11–21, 2014. (page 44).

[167] Baofeng Wu, Jiayu Shao, and Xiying Yuan. **Deleting vertices and interlacing Laplacian eigenvalues**. *Chinese Annals of Mathematics, Series B*, **31**(2):231–236, February 2010. (page 44).

[168] J H Wilkinson. *The algebraic eigenvalue problem*. Oxford University Press, 1965. (page 44).

[169] R A Marcus. **Brief comments on perturbation theory of a nonsymmetric matrix: The GF matrix**. *Journal of Physical Chemistry A*, **105**(12):2612–2616, 2001. (page 45).

[170] Attilio Milanese, Jie Sun, and Takashi Nishikawa. **Approximating spectral impact of structural perturbations in large networks**. *Physical Review E*, **81**(4):046112, April 2010. (page 45).

[171] M E J Newman. **Power laws, Pareto distributions and Zipf's law**. *Audio and Electroacoustics Newsletter, IEEE*, December 2004. (page 46).

[172] Vittoria Colizza, Romualdo Pastor-Satorras, and Alessandro Vespignani. **Reaction–diffusion processes and metapopulation models in heterogeneous networks**. *Nature Physics*, **3**(4):276–282, March 2007. (page 46).

[173] Lars Kiemer, Stefano Costa, Marius Ueffing, and Gianni Cesareni. **WI-PHI: A weighted yeast interactome enriched for direct physical interactions**. *Proteomics*, **7**(6):932–943, March 2007. (page 46).

[174] Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. **Patterns and Dynamics of Users' Behavior and Interaction: Network Analysis of an Online Community**. *Journal of the American Society for Information Science and Technology*, **60**(5):911–932, May 2009. (page 46).

[175] X L Liu and C S Oliveira. **Iterative modal perturbation and reanalysis of eigenvalue problem**. *Communications in Numerical Methods in Engineering*, **19**(4):263–274, January 2003. (pages 45, 46, 48).

[176] David MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003. (page 45).

[177] Xin Yan, Yang Wu, Xiaohui Li, Chunlin Li, and Yaogai Hu. **Eigenvector perturbations of complex networks**. *Physica A: Statistical Mechanics and its Applications*, pages 1–13, April 2014. (page 45).

[178] V Hernandez, J E Roman, A Tomas, and V Vidal. **Arnoldi methods in SLEPc**. *SLEPc Technical Report STR-4*, 2007. (page 46).

[179] T Verma, N A M Araújo, and Hans J Herrmann. **Revealing the structure of the world airline network**. *Scientific Reports*, **4**, July 2014. (page 47).

[180] Stefano Boccaletti, V Latora, Y Moreno, and M Chavez. **Complex networks: Structure and dynamics**. *Physics reports*, 2006. (page 55).

[181] Juan G Restrepo, Edward Ott, and Brian R Hunt. **Characterizing the dynamical importance of network nodes and links.** *Physical Review Letters*, **97**(9):094102, September 2006. (page 57).

[182] Romain Allez and Jean-Philippe Bouchaud. **Eigenvector dynamics: General theory and some applications.** *Physical Review E*, **86**(4 Pt 2):046202, October 2012. (page 57).

[183] Yousef Saad. *Numerical methods for large eigenvalue problems*, **158**. SIAM, 1992. (page 57).

[184] Cleve Moler. **MATLAB Incorporates LAPACK** [online]. 2000. (page 58).

[185] Jacob D Stevenson and David J Wales. **Communication: Analysing kinetic transition networks for rare events.** *The Journal of Chemical Physics*, **141**(4):041104, July 2014. (page 58).

[186] Zhuo Qi Lee, Wen-Jing Hsu, and Miao Lin. **Estimating mean first passage time of biased random walks with short relaxation time on complex networks.** *PLoS ONE*, **9**(4):e93348, 2014. (page 58).

[187] Vincent Tejedor. *Random walks and first-passage properties*. PhD thesis, Technische Universität München, July 2012. (page 58).

[188] Yuan Lin and Zhongzhi Zhang. **Controlling the efficiency of trapping in a scale-free small-world network**. *Scientific Reports*, **4**:6274, September 2014. (page 59).

[189] Takamitsu Watanabe and Naoki Masuda. **Enhancing the spectral gap of networks by node removal.** *Physical Review E*, **82**(4 Pt 2):046102, October 2010. (page 59).

[190] Katherine Henzler-Wildman and Dorothee Kern. **Dynamic personalities of proteins**. *Nature*, **450**(7172):964–972, 2007. (page 61).

[191] Thomas Simonson, Georgios Archontis, and Martin Karplus. **Free energy simulations come of age: protein-ligand recognition.** *Accounts of chemical research*, **35**(6):430–437, June 2002. (page 61).

[192] Jochen Balbach, Vincent Forge, Nico A J van Nuland, Steve L Winder, Peter J Hore, and Christopher M Dobson. **Following protein folding in real time using NMR spectroscopy**. *Nature Structural Biology*, **2**(10):865–870, October 1995. (page 61).

[193] Pratul K Agarwal. **Enzymes: An integrated view of structure, dynamics and function.** *Microbial cell factories*, **5**:2, 2006. (page 61).

[194] Marc W van der Kamp, R Dustin Schaeffer, Amanda L Jonsson, Alexander D Scouras, Andrew M Simms, Rudesh D Toofanny, Noah C Benson, Peter C Anderson, Eric D Merkley, Steven Rysavy, Dennis Bromley, David A C Beck, and Valerie Daggett. **Dynameomics: a comprehensive database of protein dynamics.** *Structure*, **18**(4):423–435, March 2010. (page 61).

[195] Adam L Beberg, Daniel L Ensign, Guha Jayachandran, Siraj Khaliq, and Vijay S Pande. **Folding@home: Lessons from eight years of volunteer distributed computing**. In *Distributed Processing (IPDPS)*, pages 1–8. IEEE, 2009. (page 61).

[196] David E Shaw, Martin M. Deneroff, Ron O Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, Stanley C. Wang, David E Shaw, Martin M. Deneroff, Ron O Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. **Anton, a special-purpose machine for molecular dynamics simulation**. *ACM SIGARCH Computer Architecture News*, **35**(2):1–12, June 2007. (page 61).

[197] S R Alam, Pratul K Agarwal, M C Smith, J S Vetter, and D Caliga. **Using FPGA Devices to Accelerate Biomolecular Simulations**. *Computer*, **40**(3):66–73, March 2007. (page 61).

[198] M J Harvey, G Giupponi, and G De Fabritiis. **ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale**. *Journal of Chemical Theory and Computation*, **5**(6):1632–1639, June 2009. (page 61).

[199] Arvind Ramanathan and Pratul K Agarwal. **Computational Identification of Slow Conformational Fluctuations in Proteins**. *Journal of Physical Chemistry B*, **113**(52):16669–16680, 2009. (pages 61, 64).

[200] Arvind Ramanathan, Ji Oh Yoo, and Christopher J Langmead. **On-the-fly identification of conformational substates from molecular dynamics simulations**. *Journal of Chemical Theory and Computation*, **7**(3):778–789, 2011. (page 61).

[201] J Andrew McCammon. **Molecular dynamics of ferrocytochrome c: anharmonicity of atomic displacements.** *Biopolymers*, **21**(10):1979–1989, October 1982. (page 62).

[202] A Amadei, A B Linssen, and H J Berendsen. **Essential dynamics of proteins.** *Proteins*, **17**(4):412–425, December 1993. (page 62).

[203] Arvind Ramanathan, Andrej Savol, Christopher J Langmead, Pratul K Agarwal, and Chakra S Chennubhotla. **Discovering conformational sub-states relevant to protein function.** *PLoS ONE*, **6**(1):e15827, 2011. (pages 62, 67).

[204] Gerald R Kneller and K Hinsen. **Computing memory functions from molecular dynamics simulations**. *The Journal of Chemical Physics*, **115**(24):11097–11105, 2001. (page 62).

[205] Gerald R Kneller. **Quasielastic neutron scattering and relaxation processes in proteins: analytical and simulation-based models.** *Physical chemistry chemical physics : PCCP*, **7**(13):2641–2655, July 2005. (page 63).

[206] B Brutovsky, T Mülders, and Gerald R Kneller. **Accelerating molecular dynamics simulations by linear prediction of time series**. *The Journal of Chemical Physics*, **118**(14):6179, 2003. (page 63).

[207] Burak Alakent, M C Camurdan, and Pemra Doruker. **Hierarchical structure of the energy landscape of proteins revisited by time series analysis. I. Mimicking protein dynamics in different time scales**. *The Journal of Chemical Physics*, **123**(14), 2005. (page 63).

[208] Burak Alakent and Pemra Doruker. **Application of time series analysis on molecular dynamics simulations of proteins: A study of different conformational spaces by principal component analysis**. *The Journal of Chemical Physics*, **121**(10):4759, January 2004. (page 63).

[209] Burak Alakent and Pemra Doruker. **Hierarchical structure of the energy landscape of proteins revisited by time series analysis. II. Investigation of explicit solvent effects**. *The Journal of Chemical Physics*, **123**(14):144911, January 2005. (page 63).

[210] M C Camurdan and Pemra Doruker. **Mimicking Protein Dynamics by the Integration of Elastic Network Model with Time Series Analysis**. *International Journal of High Performance Computing Applications*, **21**(1):59–65, February 2007. (page 63).

[211] Kaihsu Tai, Tongye Shen, Ulf Börjesson, Marios Philippopoulos, and J Andrew McCammon. **Analysis of a 10-ns Molecular Dynamics Simulation of Mouse Acetylcholinesterase**. *Biophysical Journal*, **81**(2):715–724, November 2008. (page 63).

[212] M A Balsera, W Wriggers, Y Oono, and Klaus Schulten. **Principal component analysis and long time protein dynamics**. *Journal of Physical Chemistry*, **100**(7):2567–2572, 1996. (page 63).

[213] Oliver F Lange and Helmut Grubmüller. **Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales?** *Journal of Physical Chemistry B*, **110**(45):22842–22852, November 2006. (page 63).

[214] Oliver F Lange and Helmut Grubmüller. **Full correlation analysis of conformational protein dynamics.** *Proteins*, **70**(4):1294–1312, March 2008. (pages 63, 67).

[215] Anthony M A West, Ron Elber, and David Shalloway. **Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide**. *The Journal of Chemical Physics*, **126**(14):145104, 2007. (page 63).

[216] G. Meisenberg and W.H. Simmons. *Principles of medical biochemistry*. Mosby Elsevier, 2006. (page 64).

[217] J F Cardoso. **Blind signal separation: statistical principles**. In *Proceedings of the IEEE*, pages 2009–2025, 1998. (page 65).

[218] G H Golub and C F Van Loan. *Matrix computations*. Johns Hopkins University Press, 1989. (page 66).

[219] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. **Scalable molecular dynamics with NAMD**. *Journal of Computational Chemistry*, **26**(16):1781–1802, 2005. (pages 66, 67).

[220] Chakra S Chennubhotla and Allan D Jepson. **Half-lives of eigenflows for spectral clustering**. *Advances in Neural Information Processing Systems 15*, 2002. (page 69).

[221] Chakra S Chennubhotla and Ivet Bahar. **Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES.** *Molecular systems biology*, **2**:36, 2006. (page 69).

[222] Chakra S Chennubhotla and Ivet Bahar. **Signal propagation in proteins and relation to equilibrium fluctuations.** *PLoS computational biology*, **3**(9):1716–1726, September 2007. (page 69).

[223] Andrew Blake and Michael Isard. *Active Contours : The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer, February 1998. (pages 73, 76).

[224] W Boomsma, K V Mardia, C C Taylor, J Ferkinghoff-Borg, A Krogh, and T Hamelryck. **A generative, probabilistic model of local protein structure**. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(26):8932–8937, July 2008. (page 73).

[225] Midori Hyndman. *Dynamic Texture Modelling*. Master's thesis, University of Toronto, September 2006. (page 75).

[226] Q Cui and Ivet Bahar. *Normal mode analysis: theory and applications to biological and chemical systems*. Mathematical and Computational Biology Series. Chapman & Hall/CRC, Boca Raton, 2005. (page 77).

[227] Oliver F Lange, N A Lakomek, C Fares, G F Schroder, K F A Walter, S Becker, J Meiler, C Griesinger, and B L de Groot. **Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution**. *Science*, **320**(5882):1471–1475, June 2008. (page 78).

[228] Thomas S Kuhn. *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, 1962. (page 79).

[229] Charles K Fisher and Collin M Stultz. **Constructing ensembles for intrinsically disordered proteins.** *Current opinion in structural biology*, **21**(3):426–431, June 2011. (page 79).

[230] Massimo Stefani. **Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world.** *Biochimica et biophysica acta*, **1739**(1):5–25, December 2004. (page 79).

[231] Mahdi Muhammad Moosa, Allan Chris M Ferreon, and Ashok A Deniz. **Forced folding of a disordered protein accesses an alternative folding landscape.** *ChemPhysChem*, **16**(1):90–94, January 2015. (page 80).

[232] Monika Fuxreiter, Peter Tompa, István Simon, Vladimir N Uversky, and Francisco J Asturias. **Malleable machines take shape in eukaryotic transcriptional regulation.** *Nature Chemical Biology*, **4**(12):728–737, December 2008. (page 80).

[233] A K Dunker, Z Obradovic, P Romero, E C Garner, and C J Brown. **Intrinsic protein disorder in complete genomes.** *Genome informatics. Workshop on Genome Informatics*, **11**:161–171, 2000. (page 80).

[234] Yugang Zhang, Huaiqing Cao, and Zhirong Liu. **Binding cavities and druggability of intrinsically disordered proteins.** *Protein science*, January 2015. (page 80).

[235] Diwakar Shukla, Carlos X Hernández, Jeffrey K Weber, and Vijay S Pande. **Markov state models provide insights into dynamic modulation of protein function.** *Accounts of chemical research*, **48**(2):414–422, February 2015. (page 80).

Author Index