

Copyright

Carlos A. Bueno Basurco

2022

RICE UNIVERSITY

Mesoscale Models for the Study of Emergent Behaviors Arising from Protein Interactions

By

Carlos Andres Bueno Basurco

A THESIS SUBMITTED  
IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE

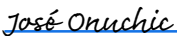
Doctor of Philosophy

APPROVED, THESIS COMMITTEE



Peter Wolynes

Bullard-Welch Foundation Professor of Chemistry  
Professor of BioSciences, Materials Science &  
NanoEngineering, and Physics & Astronomy



[José Onuchic \(Sep 30, 2022 08:13 CDT\)](#)

Jose Onuchic

Harry C. and Olga K. Wiess Chair of Physics  
Professor of Chemistry and BioSciences



[Kaden Hazzard \(Oct 10, 2022 17:59 EDT\)](#)

Kaden Hazzard

Associate Professor, Physics and Astronomy

HOUSTON, TEXAS

September 2022

## ABSTRACT

### **Mesoscale Models for the Study of Emergent Behaviors Arising from Protein Interactions**

by

**Carlos A. Bueno Basurco**

Proteins are versatile biopolymers in living systems; they exhibit a great diversity of functions depending on the order in which their amino acids are arranged. Most protein functions, like mechanical or regulatory functions, only emerge from the interactions with other proteins and macromolecules. This dissertation describes how we have developed and adapted new computational models to investigate emergent structural and dynamic properties of protein interactions.

Chapter 1 presents a review of the two systems of interest to be explored in successive chapters: the regulation of the actin cytoskeleton and the control of DNA transcription by the nuclear factor kappa B (NF- $\kappa$ B). It also introduces some models developed to study the interactions of protein with actin filaments and with DNA.

Chapters 2 and 3 focus on protein interactions in the actin cytoskeleton network. Chapter 2 describes how we have estimated the mechanical and dynamical properties of actin networks using polymer theory. We developed a simplified mathematical mean-field model of F-actin polymerization, cross-linking, and branching based on mass action kinetics. Then we obtained an analytical solution to

the connectivity, rigidity, and force percolation transitions using a generalized version of the Flory-Stockmayer theory. Chapter 3 describes how we used a computational mechano-chemical model to simulate the conditions where the actin networks exhibit rare sudden movements. We show that actin networks containing Arp2/3 undergo sudden releases of strain known as “cytoquakes”.

Chapters 4 and 5 focus on DNA-protein interactions. Chapter 4 describes a new implementation to simulate protein and DNA dynamics for large systems that we developed. This new procedure retains the accuracy of previous methods our group developed with a 30-fold speedup and eases the introduction of new potential energy terms. Chapter 5 describes how we used this protein and DNA model to explain why the NF- $\kappa$ B heterodimer, a transcription factor, can bind the DNA with more affinity than the corresponding NF- $\kappa$ B homodimer. We determined that a higher affinity in the heteromer protein-protein interface can help us explain this feature. We also measured the energy landscape, frustration, and low-frequency vibrations of the complexes.

In conclusion, these novel methods and models reproduce protein interactions in cells. These models have elucidated the origin of several emergent behaviors, such as force percolation and marginal stability in actin networks, as well as signal transduction by protein-DNA complexes.

# Acknowledgments

It is my belief that a Ph.D. degree cannot be achieved alone. Science is a community endeavor, and we need to rely on the support of each other to reach places that have not been reached before.

I want to thank my advisor during my Ph.D., Peter Wolynes, who always suggested novel ideas. I enjoyed working on the few ones I could undertake. I learned from Peter that science has a significant component of communication and human interaction, and even simple plots describe an underlying hypothesis of the main factors contributing to a phenomenon. He taught me to be thoughtful and organized in the way I explore new theories, make simulations, and analyze results to make sure no new findings are overlooked. He also trained me not to define any problem as impossible but to think seriously about what steps would be needed to solve them and overcome the limitations our current tools have. I would also like to thank Margaret Cheung, who acted as a co-advisor for the projects related to actin networks and treated me as one of her students. I felt welcomed in her group during our collaborations.

I learned a lot from the conversations with Nicholas Schafer, who acted as a mentor, encouraging me to push myself as a person and not let anyone, even myself, put arbitrary limits on my goals. I also learned much from Davit Potoyan and Ryan Cheng about the details of executing and communicating science. I also appreciate my time with James Liman, who worked with me to understand and discuss

passionately about the details and assumptions of every model we used and was always there as a friend.

I also enjoyed the discussions I had with all the CTBP members I had the opportunity to meet: Wei Lu, Qian Wang, Victor Tsai, Aram Davtyan, Xingcheng Lin, Shikai Jin, Hana Jaafari, Xinyu Gu, Mingchen Chen, Antonio Oliveira, Vinicius Contessoto, Andrei Gassic, Esteban Doderro-Rojas. I am also thankful to Neal Waxham, Garegin Papoian, Juan de Pablo, and Peter Eastman, who helped us develop the models used in this paper. I would also like to thank the support of my SSPB cohort, especially Ilenne Del Valle, who helped me navigate the difficulties of doing a Ph.D. in a new country.

Finally, I want to thank my parents, Luis and Frida, for their unconditional support, even if they could not understand why I was investing so much time and effort in studying the shape and movement of exotic molecules. Finally, I would also like to thank my wife, Susana, whom I married during this Ph.D. and who was always by my side, sharing the good and the bad times.

# Contents

<b>Acknowledgments</b> .....	<b>iii</b>
<b>Contents</b> .....	<b>v</b>
<b>List of Figures</b> .....	<b>viii</b>
<b>List of Tables</b> .....	<b>xi</b>
<b>List of Equations</b> .....	<b>xii</b>
<b>Nomenclature</b> .....	<b>xiv</b>
<b>Introduction</b> .....	<b>1</b>
1.1. Motivation .....	1
1.2. Protein interactions with actin filaments.....	3
1.2.1. Actin filaments .....	3
1.2.2. Actin-binding proteins .....	5
1.2.3. Mesoscopic models for actin networks.....	7
1.3. Protein interactions with DNA .....	8
1.3.1. 3SPN.2.....	9
1.3.2. AWSEM .....	10
<b>Connectivity and rigidity percolation of cytoskeletal networks</b> .....	<b>11</b>
2.1. Introduction.....	11
2.2. Results .....	16
2.2.1. Macroscopic chemical kinetics laws recapitulate MEDYAN simulations of binding stoichiometries .....	16
2.2.2. Actomyosin networks undergo two sol-gel transitions when modeled using the two-step model of linker binding.....	19
2.2.3. Arp2/3 complex changes the percolation threshold.....	23
2.2.4. Contraction occurs when the network is gelled by motors and linkers and not gelled by only linkers .....	26
2.3. Discussion .....	29
2.3.1. Arp2/3 changes the requirements for the network to percolate .....	29
2.3.2. The actin network is in the sol state when the linker concentration is much greater than the actin concentration .....	32

2.3.3. The difference between the chemical kinetic model and MEDYAN can be attributed to the heterogeneous distribution when the system is percolated .....	33
2.4. Conclusion .....	35
2.5. Methods .....	35
2.5.1. Mass action chemical kinetics model .....	36
2.5.2. Flory-Stockmayer Theory.....	39
2.5.3. Rigidity Percolation.....	41
2.5.4. Coarse-grained mechanochemical model of actomyosin systems (MEDYAN)42	
<b>The role of the Arp2/3 complex in shaping the dynamics and structures of branched actomyosin networks .....</b>	<b>44</b>
3.1. Introduction.....	44
3.2. Results .....	47
3.2.1. Linkers modulate contraction in unbranched actomyosin networks .....	47
3.2.2. Branched actomyosin networks display convulsive movements.....	51
3.2.3. Mechanistic insights into the convulsive movements during an avalanche ...	55
3.3. Discussion .....	61
3.3.1. Contractility of actomyosin networks depends on their topologies.....	61
3.3.2. The non-equilibrium dynamics of the branched actomyosin networks displays the behavior of jammed assemblies.....	62
3.4. Methods .....	64
3.4.1. Coarse-grained mechanochemical model of actomyosin systems (MEDYAN)64	
3.4.2. Cluster size distribution .....	67
<b>OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations .....</b>	<b>69</b>
4.1. Introduction.....	69
4.2. Design and implementation .....	72
4.3. Results .....	77
4.3.1. Protein-only simulations benchmark .....	77
4.3.2. DNA-only simulations benchmark .....	78
4.3.3. Protein-DNA simulations benchmark .....	79
4.3.4. Protein-DNA interface prediction example .....	80



4.3.5. Potentials that depend on locations of residues relative to a membrane.....	83
4.3.6. Describing many-body saturating disulfide bonds .....	91
4.4. Discussion .....	97
4.5. Methods .....	99
4.5.1. Simulation setup .....	99
4.5.2. Q-value definition .....	100
<b>Resolving the NF-<math>\kappa</math>B Heterodimer Binding Paradox: Strain and Frustration Guide the Binding of Dimeric Transcription Factors.....</b>	<b>102</b>
5.1. Introduction.....	102
5.2. Methods .....	109
5.3. Results and Discussion .....	112
5.4. Conclusion .....	125
<b>References .....</b>	<b>127</b>
<b>Appendix A .....</b>	<b>166</b>
<b>Appendix B.....</b>	<b>182</b>
<b>Appendix C.....</b>	<b>197</b>
<b>Appendix D .....</b>	<b>227</b>

## List of Figures

<b>Fig 2.1. Time course of concentrations of bound species provided from the macroscopic chemical kinetic model (solid lines) and the MEDYAN simulations (dots) of actomyosin networks. ....</b>	<b>19</b>
<b>Fig 2.2. Percentage of binding sites to the total number of binding sites in different states for a two-step linker binding model is shown as a function of the total number of linkers in a system. ....</b>	<b>20</b>
<b>Fig 2.3. Heatmap of the proportion of the concentration of crosslinks to the concentration of total binding sites as a function of the linker binding equilibrium constant and the concentration of linkers. ....</b>	<b>22</b>
<b>Fig 2.4. Plot showing the location of different experiments on actin crosslinking plotted in the two-step model phase space. ....</b>	<b>23</b>
<b>Fig 2.5. Fraction of actin monomers in finite clusters (<math>P_s</math>) without Arp2/3 (left) and with Arp2/3 (right). ....</b>	<b>24</b>
<b>Fig 2.6. Fraction of actin monomers in finite clusters (<math>P_s</math>) as a function of Arp2/3 concentration and crosslinker concentration with the exception of motors. ....</b>	<b>25</b>
<b>Fig 2.7. Fraction of actin monomers in finite clusters (<math>P_s</math>) including motor and linker crosslinks (A) or only when considering linker crosslinks (B). ....</b>	<b>26</b>
<b>Fig 2.8. Schematic phase diagrams of actomyosin systems as a function of linker and motor concentrations of actomyosin networks without Arp2/3 (A) and with Arp2/3 (B). ....</b>	<b>28</b>
<b>Fig 2.9.- Rigidity percolation limits including both motor and linker crosslinks (left lines) or those found when only considering linker crosslinks (right lines). ....</b>	<b>29</b>
<b>Fig 2.10. Diagram of connections of F-actin monomers to other F-actin monomers. ....</b>	<b>38</b>
<b>Figure 3.1. Time courses of the ratio <math>Rg/Rgi</math> indicate contractile or extensile motion in unbranched actomyosin networks at several conditions of motors and linkers. ....</b>	<b>50</b>

<b>Figure 3.2. The time courses of radius gyration ratios for contractile and extensile motions in branched actomyosin networks at several conditions of motors and linkers. ....</b>	<b>52</b>
<b>Figure 3.3. Changes in the tension of a branched network occur during an avalanche. ....</b>	<b>54</b>
<b>Figure 3.4. Examples of avalanches in a branched network containing 16 motors and 3000 linkers. ....</b>	<b>57</b>
<b>Figure 3.5. Probability densities of the weighted mean cluster size (<math>N_w</math>) of the unbranched networks (A) and branched networks (B). ....</b>	<b>60</b>
<b>Figure 3.6. Typical snapshots of MEDYAN simulations without Arp2/3 protein complexes (A) or with Arp2/3 protein complexes (B). ....</b>	<b>65</b>
<b>Figure 3.7. Schematic diagram of a cluster. ....</b>	<b>67</b>
<b>Figure 4.1 Benchmark timing results for AWSEM simulations with the LAMMPS and the OpenMM implementations on a linear scale (left) and on a log scale (right). ....</b>	<b>78</b>
<b>Figure 4.2. Benchmark timing results for 3SPN2 simulations with the LAMMPS implementation of 3SPN2 and the OpenMM implementation of 3SPN2 on a linear scale (left) and on a log scale (right). ....</b>	<b>79</b>
<b>Figure 4.3. Benchmark results for AWSEM-3SPN2 simulations of protein-DNA complexes using the LAMMPS and the OpenMM implementations of both forcefields on a linear scale (left) and on a log scale (right). ....</b>	<b>80</b>
<b>Figure 4.4 Scatter plot of the interaction energy between the DNA and the protein versus the fraction of the symmetrized native contacts formed at each time frame during the last 7.5 million steps of simulations from 10 runs. ....</b>	<b>82</b>
<b>Figure 4.5. A schematic figure for the Z-dependent contact potential. ....</b>	<b>84</b>
<b>Figure 4.6. Structure prediction results using the three contact potential schemes evaluated using Qwater (left) and Qmem (right). ....</b>	<b>89</b>
<b>Figure 4.7. Overlay of the native structures and the best Qwater and Qmem structures using the membrane burial depth dependent contact potential. ...</b>	<b>90</b>

<b>Figure 4.8. The fraction of correct location assignments of the residues relative to the membrane using a purely sequence-based method (PureseqTM) and that yielded by running OpenAWSEM simulations (AWSEM).</b>	<b>91</b>
<b>Figure 4.9. Structure prediction results for six disulfide rich proteins using various strengths of the saturable disulfide bond interaction.</b>	<b>94</b>
<b>Figure 4.10. The fractions of correct disulfide bonds in the predictions of several disulfide rich proteins.</b>	<b>95</b>
<b>Figure 4.11. The formation of disulfide bonds in a single annealing trajectory with <math>k = 5</math>.</b>	<b>96</b>
<b>Figure 4.12. The average formation of disulfide bonds as a function of time over the 20 annealing runs, with the patterns from the standard AWSEM shown on the left and patterns from the nonadditive disulfide potential runs with <math>k = 5</math> shown on the right.</b>	<b>97</b>
<b>Figure 5.1. DNA-NF<math>\kappa</math>B interactions.</b>	<b>105</b>
<b>Figure 5.2. Differences in elastic and electrostatic energies between DNA-bound and free forms of NF-<math>\kappa</math>B.</b>	<b>115</b>
<b>Figure 5.3. Distributions of Cartesian and strain PCs for all three dimers of NF-<math>\kappa</math>B in their free and DNA-bound structural ensembles.</b>	<b>119</b>
<b>Figure 5.4. Frustration in NF-<math>\kappa</math>B dimers.</b>	<b>122</b>
<b>Figure 5.5. Change in the frustration along the twist modes.</b>	<b>123</b>
<b>Figure 5.6. Profiles of the elastic energy along the DNA dissociation coordinate for DNA dissociation from I<math>\kappa</math>B-bound (red) and I<math>\kappa</math>B-free (blue) heterodimer p50p65 forms.</b>	<b>125</b>

## List of Tables

<b>Table 2.1. Reactions included in the chemical kinetic model. ....</b>	<b>37</b>
--	-----------

## List of Equations

Equation 2.1 .....	39
Equation 2.2 .....	40
Equation 2.3 .....	41
Equation 3.1 .....	55
Equation 3.2 .....	55
Equation 3.3 .....	55
Equation 3.4 .....	58
Equation 4.1 .....	85
Equation 4.2 .....	85
Equation 4.3 .....	85
Equation 4.4 .....	86
Equation 4.5 .....	87
Equation 4.6 .....	87
Equation 4.7 .....	87
Equation 4.8 .....	87
Equation 4.9 .....	87
Equation 4.10.....	87
Equation 4.11.....	88
Equation 4.12.....	88
Equation 4.13.....	92
Equation 4.14.....	92

<b>Equation 4.15</b> .....	<b>92</b>
<b>Equation 4.16</b> .....	<b>92</b>
<b>Equation 4.17</b> .....	<b>92</b>
<b>Equation 4.18</b> .....	<b>93</b>
<b>Equation 4.19</b> .....	<b>100</b>
<b>Equation 5.1</b> .....	<b>110</b>
<b>Equation 5.2</b> .....	<b>112</b>
<b>Equation 5.3</b> .....	<b>112</b>
<b>Equation 5.4</b> .....	<b>113</b>

# Nomenclature

3SPN	3-site-per-nucleotide model of DNA
Arp2/3	Actin-related protein complex 2/3
AFINES	Active Filament Network Simulation
AWSEM	Associative Memory, Water Mediated, Structure and Energy Model
CaMKII	Calcium/calmodulin-dependent protein kinase II
DNA	Deoxyribonucleic acid
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
LTP	Long-term potentiation
MEDYAN	Mechanochemical Dynamics of Active Networks model
NF- $\kappa$ B	Nuclear factor kappa B
NMIIA	Non-muscle myosin IIA
PCA	Principal component analysis
$R_g$	Radius of gyration



## Introduction

*"All models are wrong, but some models are useful."- George E. P. Box*

### 1.1. Motivation

Living systems are characterized by complex, chaotic interactions that paradoxically give rise to ordered and predictable behaviors at the cellular level [1]. Individual isolated components of the cell cannot reproduce the behaviors that arise from the interactions of many molecules and their environment [2]. Although computational advances have accelerated the modeling of individual molecules [3], biological time scales remain inaccessible to detailed models due to their complexity [4]. Averaging properties of large groups of molecules using simple rules allows us to bridge the gap between theory and simulation and explain complex emergent behaviors on time scales and size scales relevant to living cells [5].

The complexity of living organisms is a result of evolution. While DNA is the molecule responsible for transmitting genetic information across generations, proteins generate complexity in living systems due to the various physicochemical properties of their monomers and the vast available sequence space [6]. The protein sequence mainly determines its 3D structure and functions [7]. Proteins seldom work in isolation; their functions are highly dependent on the environment in which they are found. Most structural, regulatory, or signaling functions, require the presence of multiple other components of the cell system and are challenging to study experimentally and theoretically.

A classic example of a complex system in the cell is the actin cytoskeleton. Actin networks enable cells to grow, migrate and support themselves. Actin is one of the most abundant proteins in eukaryotic cells and is involved in more protein-protein interactions than any other known protein [8]. Actin binds to itself, forming filaments in the cell known as F-actin. Actin also binds capping proteins, severing proteins, motor proteins, crosslinker proteins, nucleation factors, elongation factors, and many other actin-binding proteins [8,9]. The interaction of actin with these actin-binding proteins determines not only the structural properties of individual actin filaments but also the structure and dynamics of the actin network [10,11].

Another example of emergent properties arising from protein interactions can be found in the nucleus of the cell. Several proteins, for example histones, transcription factors, and helicases, interact with the DNA to control the translation of genes and the replication of DNA. A fascinating example is NF- $\kappa$ B, which forms

dimers that bind the DNA, regulate DNA transcription, and determine cell survival. Paradoxically experimental observations have found that the NF- $\kappa$ B heterodimer binds with more affinity than expected as a sum of the monomer interactions, suggesting an emergent property arising from the dimer interactions.

This dissertation focusses on the emergent behaviors arising from the interactions of actin with actin-binding proteins, the interactions of proteins with DNA, and computational and analytical methods developed to study these interactions. The emergent behaviors are observed at different scales, so the theories and models used to understand these systems correspond to different fields of physics. The first problem was studied using insights from polymer physics, soft active matter, and condensed matter. In contrast, the second problem utilizes theories related to protein folding and assembly as well as molecular biophysics.

## **1.2. Protein interactions with actin filaments**

### **1.2.1. Actin filaments**

Actin was first discovered in muscle extracts by Straub in 1942 [12]. Since then, actin has also been found in all eukaryotic cells [13], and actin homologs have been identified in bacterial cells [14]. Actin networks are the primary drivers of cell movement [15], including muscle contraction, cell protrusion [16], and neuronal plasticity [17]. Actin is present both in the cytoplasm and in the nucleus as globular actin (G-actin), the monomeric form of actin, or as filamentous actin (F-actin). F-actin is composed of hundreds to thousands of F-actin protomers [18].

The monomeric form of actin, G-actin, is a globular protein of around 42 kDa and 375 amino acids and is made up of four subdomains: A1, A2, A3, and A4. The subdomain A1 is attached to the structurally related subdomain A3 [8]. Subdomains A1 and A3 face toward the minus (pointed) end of the filament, while subdomains A2 and A4 face towards the plus (barbed) end of the filament. The A3 and A4 subdomains can be grouped into a large domain, close to the actin filament center, while the A1 and A2 subdomains form a small domain closer to the actin filament surface [19].

In the cell, G-actin polymerizes to F-actin. The F-actin structure was elucidated first in 1963 by Hanson and Lowy [20] and can be described as a left-handed helix or a right-handed double-helix. Actin has a width of around 9-9.5 nm. When it is described as a double-helix, it has a half-helical length of around 36 nm, which corresponds to 13 F-actin protomers approximately [21]. Actin filaments are semiflexible polymers with a persistence length of around 17  $\mu\text{m}$  [22]. The structural properties of the helix not only depend on the bound nucleotide (ATP vs. ADP) and cations (calcium or magnesium) that are bound to the protomers, but also depend on the actin isoform [23], additional proteins bound to the fiber such as cofilin [24] or Arp2/3 [25,26], and the mechanical stresses supported by the fiber [27].

Actin filaments are in constant dynamic turnover because they have higher affinity for new actin monomers at the plus end compared to the minus end [28]. Actin polymerization is an active process; ATP-actin binds with more affinity to the

plus end of the filament, where it then hydrolyzes to Actin-ADP and ultimately dissociates more frequently from the minus end [29]. At steady-state, the continuous growth of the filament at the plus end accompanied by the shrinking at the minus end is known as actin treadmilling [30]. Actin treadmilling is a non-equilibrium process that consumes free energy in the form of ATP, generates force, and participates in cell motility.

### **1.2.2. Actin-binding proteins**

Actin treadmilling is highly regulated in cells by capping proteins, nucleation factors, severing proteins, actin-monomer-binding proteins, and actin polymerases [9], which control the kinetics of filament polymerization and depolymerization. Actin crosslinkers, such as  $\alpha$ -actinin or fascin, alter the higher-order structures and dynamics of the network by creating larger assemblies such as bundles, asters, or isotropic networks, as well as steering the forces in the network [31]. Most actin-binding proteins bind to the same conserved binding site in F-actin [32].

Two actin-binding proteins of interest are myosin motors and Arp2/3. Myosin motors can also act as crosslinkers and generate forces in the actin networks by walking along the actin filaments, acting as a second non-equilibrium force-exerting process in actin networks [33,34]. Non-muscle myosin IIA (NMIIA) can form minifilaments, bipolar ensembles of around 10-30 individual myosin motors, that can generate contractility in actin networks [35]. Arp2/3 acts as a nucleator of new actin filaments and as a brancher. Arp2/3 binds on the side of an actin filament

and creates a nucleation point that allows a new actin filament to grow at an angle of 70 degrees from the parent filament [36].

Actin-binding proteins interact in complex ways with the actin network. For example, it has been reported that actin networks can only contract under specific concentrations of crosslinkers and motors [37] and that crosslinkers modulate the distinct steady-state dynamics of actin networks [38]. It has also been shown that fascin and  $\alpha$ -actinin can phase-separate in filament bundles [39].

Actin networks also play a central role in forming new memories in the brain, known as long-term potentiation (LTP) [40,41]. During long-term potentiation, the presynaptic neuron axon sends high-frequency electrical signals to the post-synaptic neuron [40,42]. High-frequency electric signals trigger a signaling cascade in the dendritic spine, allowing it to grow and become more susceptible to further signals from the presynaptic neuron [42].

Numerous proteins mediate the growth of the post-synaptic dendrite, and their interaction with actin is necessary to form the characteristic mushroom shape of the mature dendritic spine. It has been shown that filopodia are formed instead of the mushroom shape when Arp2/3 is absent [43], when  $\alpha$ -actinin is knocked-down or overexpressed [44], or when myosin is inhibited [45,46]. Other proteins, such as CPEB or CAMKII, also play a vital role in the maturation of the dendritic spine and have been described elsewhere [47–50].

Because of these interactions, it is helpful to consider actin networks as a system with multiple overlapping transitions happening as a function of the

concentration of actin-binding proteins. Some aspects of interest in this system are, for example, whether the system is connected or disconnected (connectivity percolation) [51], whether it deforms as a liquid or as a solid (rigidity percolation) [52], or whether it is capable of transmitting forces across the system (force percolation) [53,54]. The system can also separate on regions of different composition (phase-separate) [37], undergo alignment of filaments [55–57], and aggregate as a glass [58,59] or in a periodic structure as a crystal [60].

### **1.2.3. Mesoscopic models for actin networks**

Some coarse-grained models have been developed to study interactions in dynamic actin networks that can simulate the effects of treadmilling, motor activity, and crosslinker connectivity in actin networks, such as AFINES [61], Cytosim [62], and MEDYAN. MEDYAN (Mechanochemical Dynamics of Active Networks) is a coarse-grained mechanochemical model of actin networks that has been developed by Papoian and coworkers. [63–67].

MEDYAN models stochastic chemical reactions and deterministic mechanical representations of active systems such as the actin cytoskeleton or microtubules. This model treats unbound molecules as diffusive entities and bound molecules as structural mechanical entities. Diffusive entities, which are not bound to filaments, do not have specific positions but can react chemically with other entities at a rate determined by the local concentration. They can also diffuse between discrete compartments in the system. On the other hand, mechanical entities have specific positions and experience forces depending on their relative position to other

molecules. In this dissertation, we have used MEDYAN to model actin networks containing NMIIA, Arp2/3, actin, and  $\alpha$ -actinin [59]. A complete description of MEDYAN can be found in Chapter 3.

To obtain a more analytical insight into the connectivity percolation, rigidity percolation, and force percolation of the modeled cytoskeletal networks we also employ the theory of Flory and Stockmayer [68,69], which describes the conversion of monomers into heavy insoluble gels by branching and cross-linking. Tavares et al. recently generalized the Flory-Stockmayer theory [70,71] to handle complex polyfunctional monomers, like the F-actin protomer [53]. A complete description of the Flory-Stockmayer theory can be found in Chapter 2.

### **1.3. Protein interactions with DNA**

More than 1000 proteins in the human proteome interact with DNA, of which around half are transcription factors [72]. The other half includes proteins with functions such as DNA repair, DNA replication, and association with chromatin [73]. Some proteins, such as the histones, bind the DNA non-specifically mainly through electrostatic interactions with the DNA phosphate groups. On the other hand, transcription factors are thought to bind to a specific sequence motif in the DNA, perhaps by interacting directly with the nucleobases.

To study protein-DNA interactions, atomistic models have been created for proteins and DNA, which can be used to study fast conformational transitions[74,75]. These atomistic models struggle to simulate the time and size



scales relevant to understand more complex phenomena. DNA is the longest molecule found in a cell, and models of different scales can be used to understand the emerging phenomena arising from its interactions with proteins, from the atomistic scale to the mesoscale and chromosome-scale models [76].

Several coarse-grained DNA models have been developed to represent the DNA in the mesoscale. Among them, the martini DNA model, which models each nucleic acid using 6-7 sites per nucleotide [77], and the SIRAH DNA model [78,79], which models the DNA with six sites per nucleotide, have been used in combination with protein models to study protein-DNA interactions[80,81]. Another DNA Coarse-grained model is oxDNA [82,83] which has been used in combination with an anisotropic network model of proteins to simulate histone-DNA complexes [84]. In this dissertation we focus on the 3SPN.2 coarse-grained DNA model in combination with the AWSEM coarse-grained protein model.

### **1.3.1. 3SPN.2**

3SPN.2 is a Coarse-Grained DNA model developed by the de Pablo group that models the DNA molecule using three sites per nucleotide: the phosphate group site, the sugar site and the nucleobase site [85]. 3SPN.2 provides a flexible representation of the DNA backbone, nucleobases interactions, and electrostatics, while 3SPN.2C, a modification of the 3SPN.2 potential, incorporates the DNA sequence-dependent curvature [86]. The 3SPN.2 model has been used in conjunction with the AICG2+ protein model for protein-DNA simulations [87] of histones[88,89] and transcription factors [90] and to study the opening of cohesin's

SMC ring [91]. AICG2+ also implements protein-RNA interactions and protein-lipid interactions [92,93].

### **1.3.2. AWSEM**

The Associative memory, Water-mediated, Structure and Energy Model (AWSEM) is a coarse-grained protein model. Our group has used 3SPN.2C in combination with the AWSEM to study protein-nucleic acid complexes, such as the nucleosome [94], SMC complexes [95], NF- $\kappa$ B DNA complexes [96], T7 helicase [97], and CPEB3 [49]. AWSEM parameters were optimized by a machine learning strategy based on the principles of the energy landscape theory of protein folding and specifically the Principle of Minimal Frustration [98]. AWSEM represents each amino acid with three explicit particles, which correspond to the carbon- $\alpha$  (CA), oxygen (O), and carbon- $\beta$  (CB) atoms [98], while the positions of other backbone atoms can be calculated directly from their coordinates using the assumption of a planar peptide bond

The AWSEM Hamiltonian includes a backbone potential, a contact potential, a burial potential, a hydrogen bonding potential, and an associative memory potential based on bioinformatic input [98]. Also, depending upon the problem of interest, AWSEM may include a desolvation barrier potential [99], an electrostatic potential [100], and a membrane potential [101]. AWSEM has performed among the top 10 forcefields in the recent CASP13 protein folding competition [102,103]. A more detailed description of the AWSEM forcefield and the 3SPN.2 forcefield can be found in Chapter 4 and in Appendix C.

# Connectivity and rigidity percolation of cytoskeletal networks.

This chapter is a slightly modified version of an article published in the PLOS Computational Biology journal [53].

### 2.1. Introduction

Actomyosin networks are essential for crucial aspects of muscular contraction, cellular locomotion, endocytosis, the sensing of obstacles in the environment, and even for the synaptic plasticity of dendritic spines [17]. In muscle cells actin filaments and the accompanying myosin minifilaments are parallel and organized, and the mechanism of contraction is relatively well understood [104]. On the other hand, in non-muscle cells, actomyosin networks are non-equilibrium dynamic systems of actin filaments connected by actin-binding proteins [9,105]. The non-muscle actin filaments can be randomly oriented, or can form branched bundles [106,107], asters [66,108], or meshes [109,110]. The filaments can be in a homogeneous solution [106], form a distinct phase [37], or may display more

complex architectures [111]. For example in neurons, the complex structure of actomyosin networks in the dendritic spines are regulated by actin-binding proteins such as non-muscle myosin IIA heavy chain (NMIIA) motors,  $\alpha$ -actinin, actin-related protein complex 2/3 (Arp2/3), and calcium/calmodulin-dependent protein kinase II (CaMKII) [47,112].

Crosslinkers, like  $\alpha$ -actinin, bind actin filaments at binding sites located on the sides of the filaments [113]. When the system has become sufficiently connected by  $\alpha$ -actinin, the network rheology changes. The network behaves as an elastic solid when the concentration of  $\alpha$ -actinin is less than the bundling threshold, but the network behaves as a viscous fluid when the  $\alpha$ -actinin concentration is higher than the bundling threshold [106]. The stiffness of the network changes by several orders of magnitude even for small alterations of cross-linker concentration [114]. Other crosslinkers lead to a similar pattern of mechanical response depending on their structures and sizes [115]. Arp2/3 is also an actin crosslinker, but it binds and caps the minus end of a daughter filament. Thus Arp2/3 acts both as a nucleator and as a brancher [25]. Experimentally, Arp2/3 has been shown to nucleate actin filaments and form branched actomyosin networks [116]. Branched actin networks display different dynamical and rheological behavior than do randomly crosslinked networks [10,59,105]. This variety of behaviors allows actin networks with Arp2/3 to carry out distinct cellular functions. Branched actin networks can also display rare convulsive large scale remodeling events called avalanches or “cytoquakes” [59,117].

Advances in reconstituting actin myosin systems have given insight into how actin interacts with specific actin-binding proteins [118]. A reconstituted network is able to contract when the system has more than a threshold concentration of motors but only over a limited window of concentration of linkers [37]. Changing the concentration of crosslinkers in a reconstituted system with myosin affects the steady state dynamics of actin networks [38]. Experiments have also shown that at high myosin density, crosslinkers are not needed to promote contractility on cellular length scales [119].

The variety of orientations, architectures and biochemical compositions of actin networks makes it difficult to develop a grand unifying theory that can explain all aspects of cytoskeletal contraction. Nevertheless, one of the most important factors determining the behavior of an actomyosin network is the connectivity of the network. This connectivity modulates non-monotonically the network's ability to contract [37,38,106]. Motor activity is also needed to allow actin networks to contract and modulates contractility in a non-monotonic way. Motors encourage contractility at medium levels of activity, but decrease contractility when their activity is high but their processivity is low [120–122]. The buckling of the filaments is also necessary for the contraction of highly connected actomyosin networks [26,34,123–126]. Other structural features such as filament bundling [127], the alignment of the filaments [56], branching by Arp2/3 [25], shrinking of the actin filaments [128] and global changes in the actin network architecture [129] also modulate the ability of the network to contract. All of these features depend on the biochemical composition of the actomyosin system [26].

Several models have been developed to simulate interactions between actin-binding proteins and actin filaments such as MEDYAN [67], Cytosim [62], and AFINES [61]. We have previously studied the reorganization of actin networks caused by Arp2/3 using the MEDYAN model, which includes mechanochemical feedback on the binding and unbinding of actin-binding proteins to actin filaments and represents actin filaments as mechanical objects [59]. The completeness of these simulation models is a virtue, but in this paper, we develop a simple analytical model that allows us to highlight and appreciate how the connectivity of the cytoskeletal network develops in time and influences the dynamics and rheology of actomyosin systems.

We previously have explored an actin contractility model focused on the load response of individual actin filaments and active motor-like events [34,125,130–134] and explicitly connected the concept of rigidity percolation with glass transitions in network materials [135]. In work related to the present effort, Zilman and Safran have predicted the structural behavior of non-motorized actin networks with a single crosslinker type using a theoretical mean-field model based on the Flory polymer theory [136]. These models however do not deal explicitly with the branching nature of Arp2/3, which is a key biochemical component in cortical actin. The approach taken here is based on the seminal work of Flory and Stockmayer on condensation polymer networks in a Bethe lattice [68,137]. A recent generalization of Flory's work allows us to model branched networks where multivalent monomers have multiple different binding sites, an important feature of biological actin networks [71]. We also highlight the distinction between connectivity

percolation, also called conductivity percolation, which simply monitors the existence of an infinite cluster that is connected and rigidity percolation which determines when the infinite cluster becomes elastically stable [138–140]. Alvarado et al. have proposed a schematic phase diagram for active systems, with 4 regimes, where the network can be described as being an active solution, a prestressed gel, able to undergo global contractions or only local contractions [54].

In the present analysis we have developed a macroscopic chemical kinetic model based on binding and unbinding kinetics of actin-binding proteins. We found that the transient concentrations obtained from the chemical kinetic model are comparable with the results of a coarse-grained mechanochemical model (MEDYAN) before the percolation transition. We also showed how the mechanism of binding between actin-binding proteins and actin filaments and the binding cooperativity can alter the concentrations needed to observe connectivity percolation. The calculations show that low concentrations of motors are not able to produce contractile motions in the actin networks without additional linkers, but that at high concentrations motors are sufficient to produce contractile motions. We also locate the connectivity percolation transition as a function of linker composition and explore how the rigidity percolation transition differs from the connectivity percolation transition when the connections made by the linkers are not themselves individually rigid. We find that, unlike other actin-binding proteins, Arp2/3, an actin brancher that generates complex architectures, modulates the actomyosin percolation in the network in a non-monotonic way. In conclusion, the present model based on the Flory-Stockmayer theory allows us to determine how

the biochemical composition, branching, and the linker binding mechanism are linked to the connectivity in the system and the observation of contraction.

## **2.2. Results**

### **2.2.1. Macroscopic chemical kinetics laws recapitulate MEDYAN simulations of binding stoichiometries**

We used a macroscopic kinetic description to predict the number of both the connections made during the growth, and the number of plus and minus ends of the actin filaments, which determines the length and their treadmilling rate. We compared the transient concentrations of the different kinds of F-actin binding species obtained from a chemical kinetic model to the predicted transient concentrations of F-actin binding species obtained using MEDYAN. MEDYAN is a state-of-the-art coarse-grained mechanochemical model of the actomyosin networks. MEDYAN, unlike the chemical kinetic model, includes stochastic chemical reactions, mechanical representations and mechanochemical feedback of far-from-equilibrium systems. The chemical kinetic model allows us to find an analytical solution to the percolation of actomyosin networks over time.

In the main, the transient concentrations from the chemical kinetic model and MEDYAN simulations agree with each other as shown in Fig 2.1. For both models we started the system with small filaments of F-actin that act as nucleators. During the first part of the trajectories, G-actin polymerizes into F-actin filaments and actin binding sites become available for actin-binding proteins, such as  $\alpha$ -



actinin, myosin, and Arp2/3 for them to bind. As the simulation progresses the binding and unbinding rates even up and the concentrations of bound actin-binding proteins reach a steady state.

Late in the growth of the network there are some differences between the transient concentrations of bound sites predicted by the chemical kinetic model and those from the MEDYAN simulation. First, the concentration of F-actin monomer direct connections ( $[F_m \cdot F_p]$ ) obtained from MEDYAN was slightly lower than the chemical kinetic model result (purple lines in Fig 2.1). We attribute this small difference to the fact that, in MEDYAN, the polymerization rate of those filaments that are near the wall is decreased by mechanochemical feedback when they collide with the wall. The chemical kinetic model does not take such mechanochemical feedback or wall interactions into account.

The concentration of bound motors ( $[F_c \cdot M \cdot F_c]$ ) obtained from the MEDYAN simulation does not differ from the concentration of bound motors in the macroscopic chemical kinetic model (yellow lines in Fig 2.1). On the other hand, the concentrations of bound linkers ( $[F_c \cdot L \cdot F_c]$ ) obtained from MEDYAN differs from the concentrations of bound linkers calculated using the chemical kinetic model (orange lines in Fig 2.1). This discrepancy arises from the heterogeneous distribution of the binding sites in the system. In the chemical kinetic model, a homogeneous distribution of binding sites and an isotropic network conformation is assumed, while in MEDYAN the distribution is spatially heterogeneous and the filaments can form bundles. The heterogeneous distribution of binding sites in

MEDYAN implies that fewer binding sites are available at a given time to be bound by linkers due to the small distance of search that is possible for linkers that have already been bound ( $d_C^{\min} = 30 \text{ nm}$ ,  $d_C^{\max} = 40 \text{ nm}$ ). The concentration of bound motors ( $[F_C \cdot M \cdot F_C]$ ) acquired from MEDYAN and from macroscopic kinetics are similar because the search distance for a motor is greater than the search distance for a linker in the MEDYAN model ( $d_M^{\min} = 175 \text{ nm}$ ,  $d_C^{\max} = 225 \text{ nm}$ ).

Finally, the concentration of bound branchers ( $[F_C \cdot B \cdot F_m]$ ) found in MEDYAN is greater than the concentration of bound branchers obtained from the chemical kinetic model (red lines in Fig 2.1). This difference is a consequence of the results for other species that we have just discussed. The large concentration of non-polymerized G-actin molecules in the system with Arp2/3 (Fig 2.1B) predicted by the MEDYAN model comes from there being a slower effective polymerization rate caused by collisions between actin filaments and the boundary. The resulting larger concentration of available binding sites allows the branching reaction to occur faster in MEDYAN compared with what happens in the macroscopic chemical kinetic model. The concentrations at steady state of the bound species in MEDYAN tend to converge to those from the chemical kinetic model.

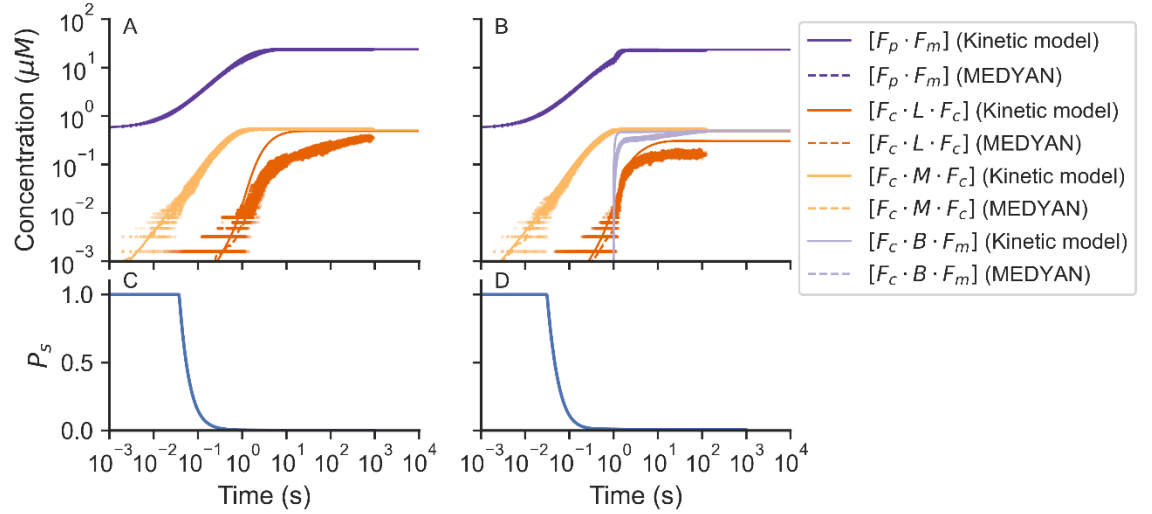


Fig 2.1. Time course of concentrations of bound species provided from the macroscopic chemical kinetic model (solid lines) and the MEDYAN simulations (dots) of actomyosin networks.

Results for networks without Arp2/3 are shown in (A) and results with Arp2/3 are shown in (B). The average MEDYAN concentrations are plotted as dotted lines. Fraction of F-actin monomers in finite clusters obtained from the chemical kinetic model are shown in (C) without Arp2/3 and results with Arp2/3 are shown in (D).  $[F_m \cdot F_p]$  is the concentration of plus sites of F-actin monomers bound to a minus site of another actin monomer.  $[F_c \cdot L \cdot F_c]$  is the concentration of F-actin monomer binding sites bound to another F-actin monomer binding site through a linker ( $\alpha$ -actinin).  $[F_c \cdot M \cdot F_c]$  is the concentration of actin monomer binding sites bound to another actin monomer binding site through a motor (NMIIA).  $[F_c \cdot B \cdot F_m]$  is the concentration of actin monomer binding sites bound to a minus site of another actin monomer through a brancher (Arp2/3).

### 2.2.2. Actomyosin networks undergo two sol-gel transitions when modeled using the two-step model of linker binding

Some actomyosin models simplify the binding of linkers and motors to F-actin filaments as a one-step reaction, binding two actin filaments at the same time [62,67]. Other models consider a two-step model for actin binding, where each reaction happens at different moments [106,141,142]. Here we examine the effect of

a two-step non-cooperative model on the percolation of actin networks where both actin-binding domains have the same affinity to actin-binding sites.

As they grow, actomyosin networks undergo a sol-gel transition. In the two-step model of linker binding, both heads of the linker bind independently to the actin-binding sites (non-cooperative binding). At an intermediate linker concentration, a maximum concentration of crosslinker connections ( $[F_c \cdot L \cdot F_c]$ ) is found. Above this linker concentration the binding sites have become saturated with linkers, increasing the single bound linker concentration ( $[F_c \cdot L]$ ). The number of crosslinker connections is maximum when the concentration of linkers is equal to the concentration of binding sites, as shown in Fig 2.2.

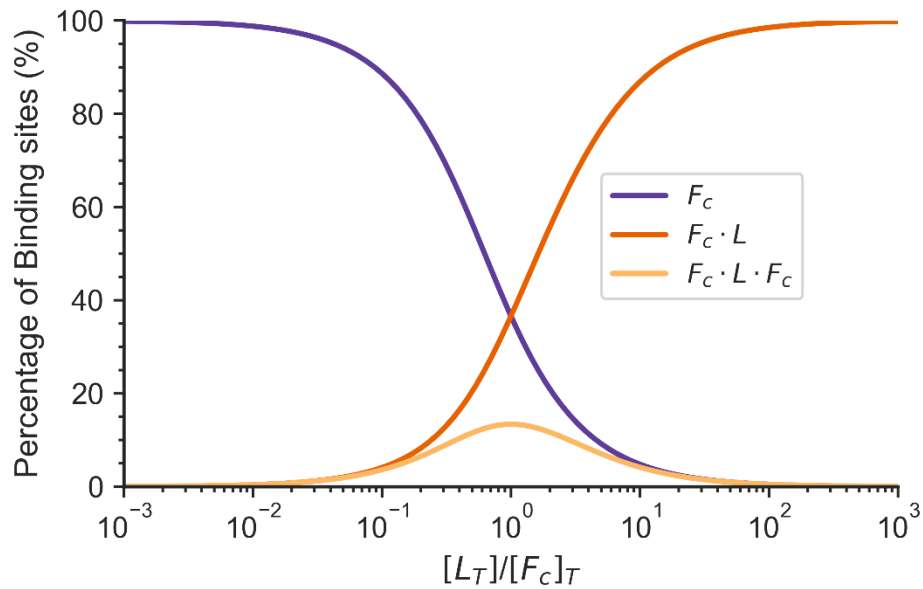


Fig 2.2. Percentage of binding sites to the total number of binding sites in different states for a two-step linker binding model is shown as a function of the total number of linkers in a system.

$F_c$  is the percentage of the concentration of free binding sites to the total concentration of binding sites.  $F_c \cdot L$  is the percentage of the concentration of binding sites bound to a linker

to the total concentration of binding sites.  $F_c \cdot L \cdot F_c$  is the percentage of the concentration of crosslinks to the total concentration of binding sites. The total concentration of binding sites  $[F_c]_T$  in the system is  $25 \mu\text{M}$ ,  $k_c^+ = 1 \mu\text{M}^{-1}\text{s}^{-1}$ ,  $k_c^- = 1\text{s}^{-1}$ .

For this model both heads of the linker have the same probability to bind to an actin-binding site, independent of the state of the opposite head. The number of connections in this system depends on multiple factors, including the total concentration of binding sites ( $[F_c]_T$ ), the total concentration of linkers ( $[L]_T$ ), and the linker binding equilibrium constant ( $K_c$ ). The number of connections formed in the system is shown in a 2D plot by normalizing the number of connections with the total concentration of binding sites.

The maximum concentration of crosslinker connections ( $[F_c \cdot L \cdot F_c]$ ) for this system occurs when the binding constant ( $K_c$ ) is larger than the total concentration of binding sites ( $[F_c]_T$ ) and the total linker concentration ( $[L]_T$ ) is equal to the total concentration of binding sites ( $[F_c]_T$ ) (Fig 2.3). A small survey of experiments in the literature shows that the first connectivity percolation transition has been observed when the system is not saturated by linkers. (Fig 2.4).

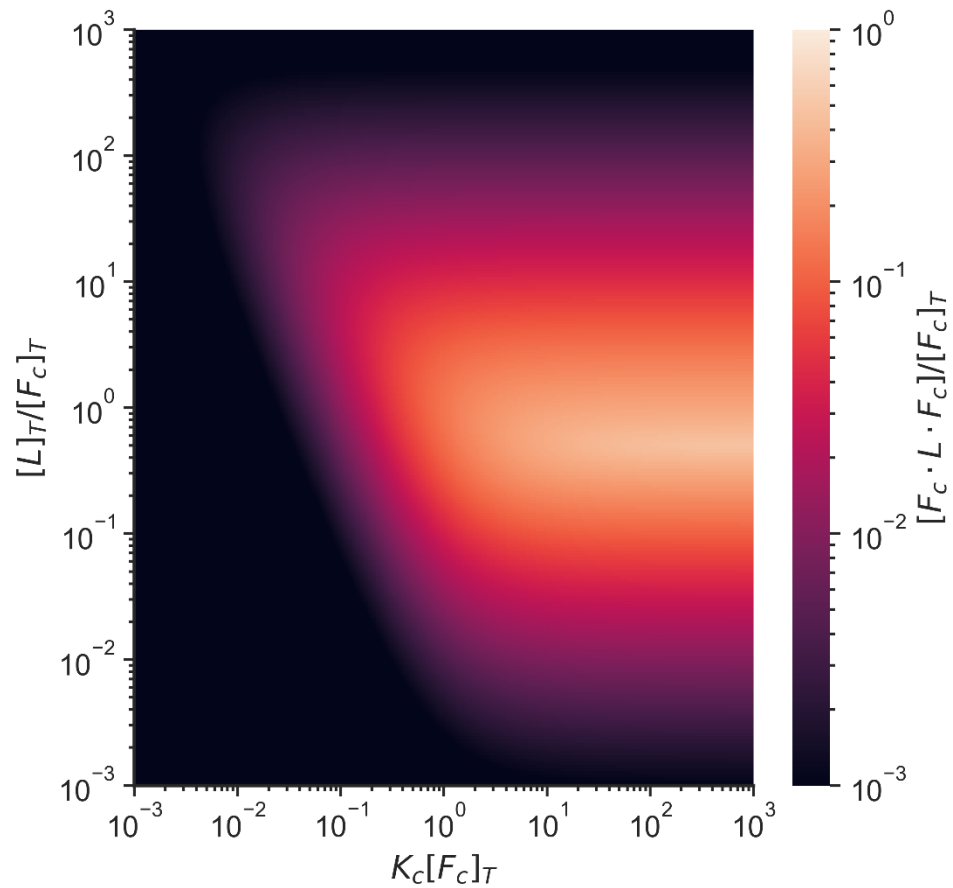


Fig 2.3. Heatmap of the proportion of the concentration of crosslinks to the concentration of total binding sites as a function of the linker binding equilibrium constant and the concentration of linkers.

All axes have been normalized by the concentration of total binding sites in the system.  $[L]_T$  is the total linker concentration,  $[F_c]_T$  is the total concentration of binding sites,  $[F_c \cdot L \cdot F_c]$  is the concentration of crosslinks, and  $K_c$  is the linker binding equilibrium constant.

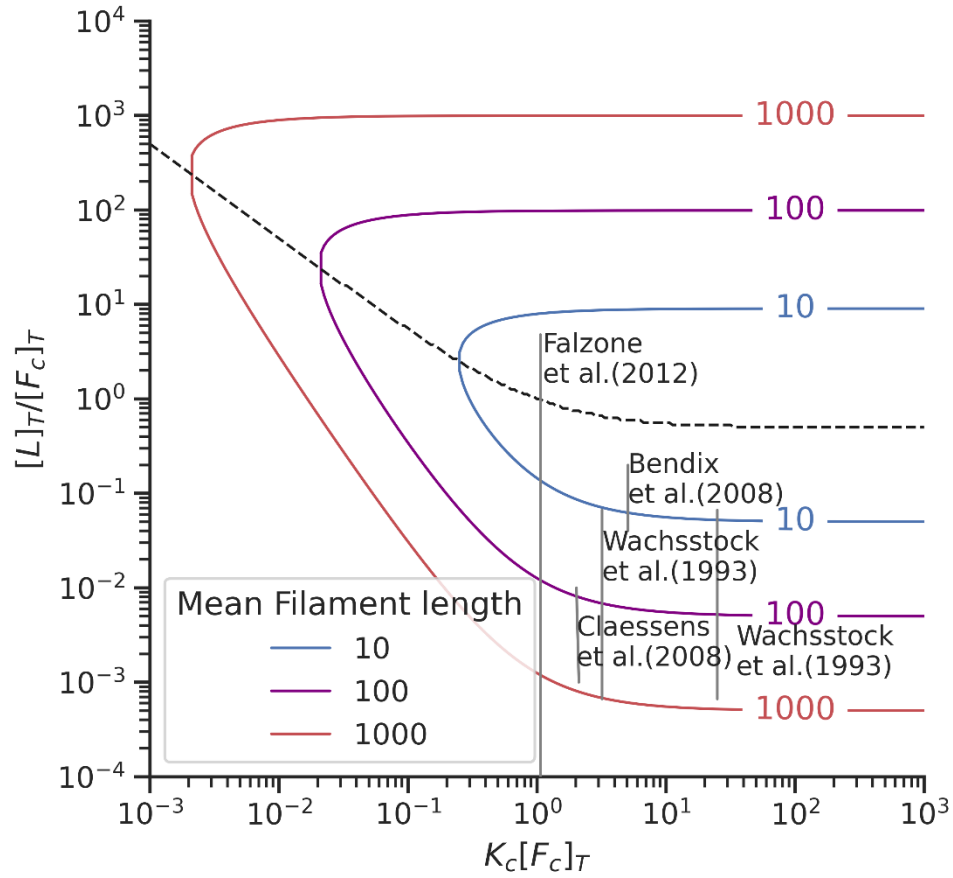


Fig 2.4. Plot showing the location of different experiments on actin crosslinking plotted in the two-step model phase space.

The curved lines indicate percolation transitions for filaments of different sizes. The dotted black line indicates the region where the maximum number of crosslinks can be observed.  $[L]_T$  is the total linker concentration,  $[F_c]_T$  is the total concentration of binding sites, and  $K_c$  is the linker binding equilibrium constant. The experimental concentrations of linkers and equilibrium constants were obtained from literature. [37,106,143,144]

### 2.2.3. Arp2/3 complex changes the percolation threshold

To understand the effects of motors and Arp2/3 complexes on the connectivity of the actomyosin network we included them in the analysis of the macroscopic chemical kinetic model and calculated the probability that an actin

monomer is in a finite cluster ( $P_s$ ). When  $P_s < 1$ , there is at least one infinite cluster in the system, and the system has formed a gel.

Motors connect the system in much the same way as linkers do since they also connect two binding sites. Therefore, motor binding increases the crosslinking probability ( $p_c$ ) in an additive manner with linker binding (Fig 2.5A). Including  $0.5 \mu\text{M}$  Arp2/3 to the system reduces the total number of linkers or motors required to gelate the network (Fig 2.5B).

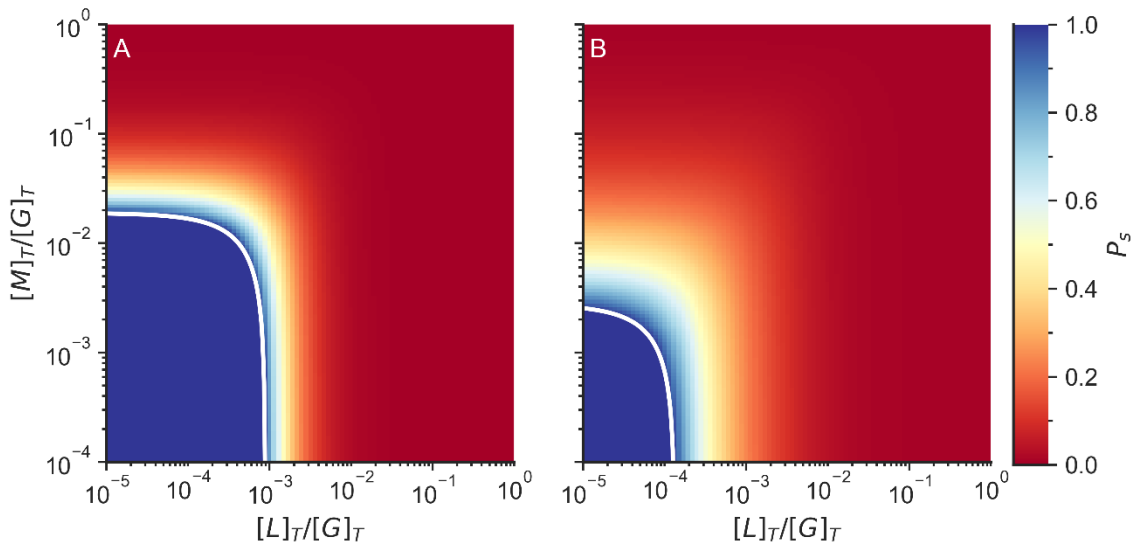


Fig 2.5. Fraction of actin monomers in finite clusters ( $P_s$ ) without Arp2/3 (left) and with Arp2/3 (right).

The color indicates the probability that an F-actin monomer is in a finite cluster. The white line indicates the connectivity percolation transition. The system is not gelled when  $P_s = 1$ , while the system is gelled when  $P_s < 1$ .  $[L]_T$  is the total linker concentration,  $[M]_T$  is the total motor concentration, and  $[G]_T$  is the total G-actin concentration. The total concentration of G actin in the system was  $25 \mu\text{M}$  and the total concentration of Arp2/3 on the simulations with Arp2/3 was  $0.5 \mu\text{M}$ .



Increasing the Arp2/3 concentration decreases the concentrations of linkers or motors needed to gelate the network. Only at high Arp2/3 concentrations do we find that the system is unable to form a gel even in the presence of high linker concentration (Fig 2.6). This is due to the saturation of binding sites by Arp2/3 which competes with linker binding, and the saturation of minus sites ( $F_m$ ) which compete with polymerization.

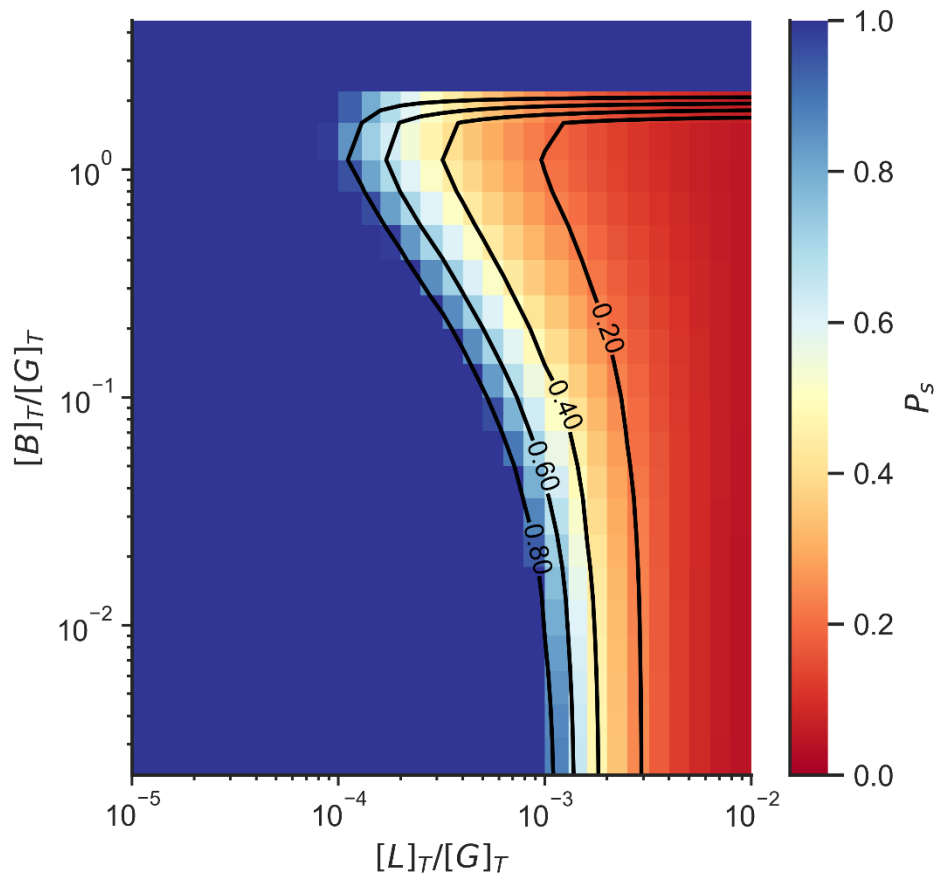


Fig 2.6. Fraction of actin monomers in finite clusters ( $P_s$ ) as a function of Arp2/3 concentration and crosslinker concentration with the exception of motors.

The color indicates the probability that an F-actin monomer is in a finite cluster. The system is not gelled when  $P_s = 1$ , while the system is gelled when  $P_s < 1$ .  $[L]_T$  is the total linker concentration,  $[M]_T$  is the total motor concentration, and  $[G]_T$  is the total actin concentration. The total concentration of G actin in the system was 25  $\mu\text{M}$ .

## 2.2.4. Contraction occurs when the network is gelled by motors and linkers and not gelled by only linkers

Linkers and motors behave differently in actomyosin networks. Motors tend to walk over filaments, exerting forces in the network, while linkers act as structural beams making the network more rigid. In the previous section we defined percolation as occurring when the network is simply fully connected by a combination of linkers and motors, but it is also possible to define a connectivity percolation limit monitoring when the system is connected by linkers alone (Fig 2.7).

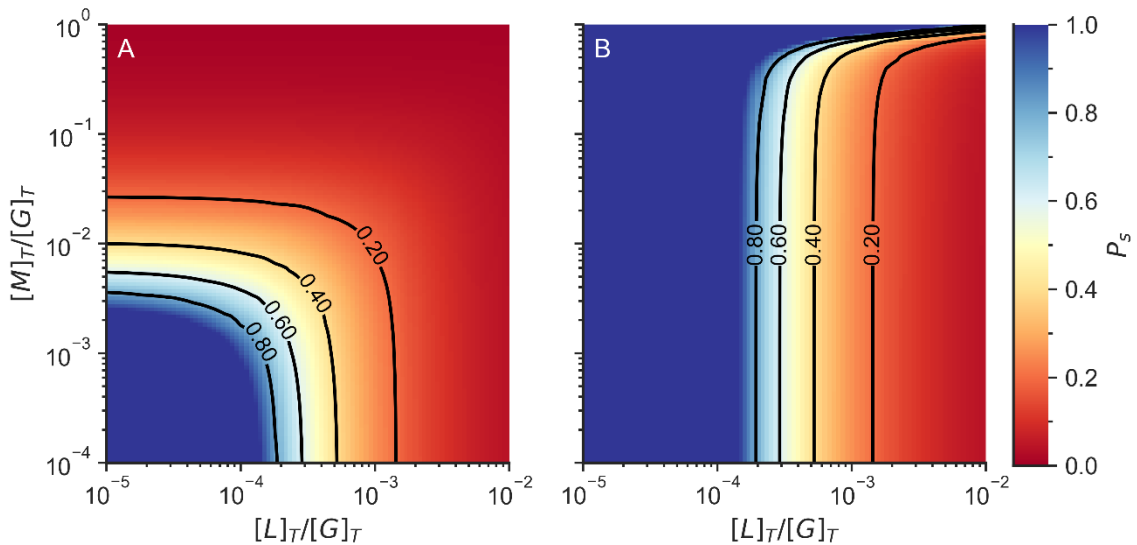


Fig 2.7. Fraction of actin monomers in finite clusters ( $P_s$ ) including motor and linker crosslinks (A) or only when considering linker crosslinks (B).

The color indicates the probability that an F-actin monomer is in a finite cluster. The system is not gelled when  $P_s = 1$ , while the system is gelled when  $P_s < 1$ .  $[L]_T$  is the total linker concentration,  $[M]_T$  is the total motor concentration, and  $[G]_T$  is the total actin concentration

Taking into consideration the different behavior of motors and linkers, we see there are three regimes for our system (Fig 2.8). In a first regime at low concentrations of motors and passive linkers (purple region in Fig 2.8) the system cannot form a gel. In a second regime at high motor concentration and low linker concentration (white region in Fig 2.8) the system is gelled by motors but does not form a gel by linkers considered by themselves. Finally in a third regime at high linker concentration the system is gelled by linkers and motors acting together (green region in Fig 2.8).

When both the linker and motor connections with the actin are individually rigid, the number of degrees of freedom lost by binding equals the total number of degrees of freedom of the actin monomer, therefore the threshold for connectivity percolation will be the same as for rigidity percolation. Noting this, we suggest that these three regimes can explain the different mechanical behaviors manifested by the actomyosin network. In regime 1, the system is floppy and cannot transmit or exert forces. In regime 2, the motors can exert forces to the system and the system is able to contract. In regime 3, the linkers provide structure to the network so the network can transmit forces, but the network has become so rigid that it is unable to contract significantly through motor action.

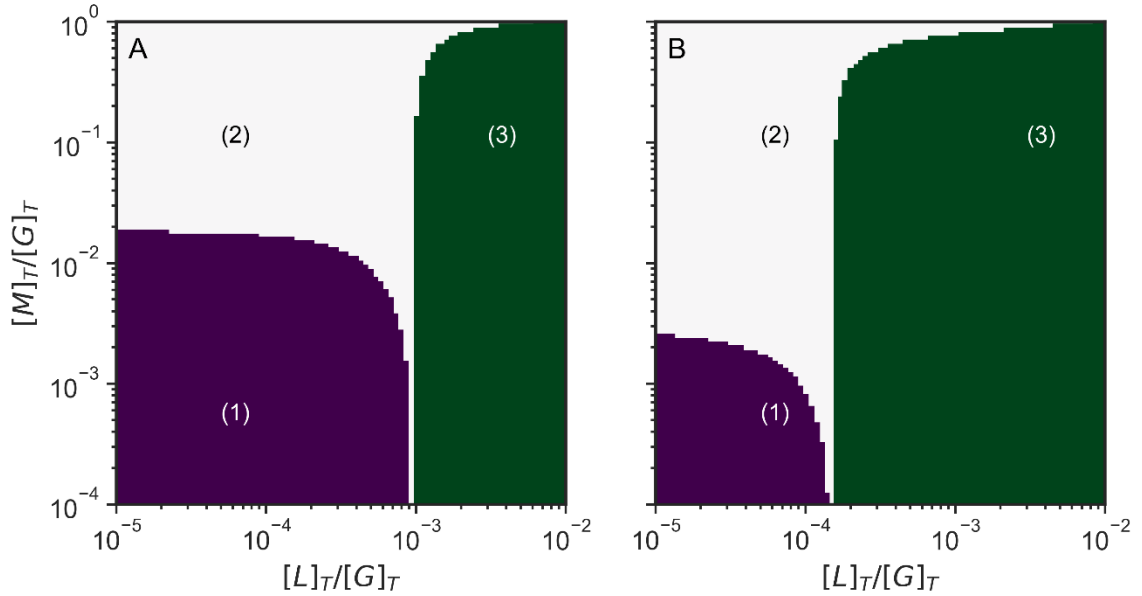


Fig 2.8. Schematic phase diagrams of actomyosin systems as a function of linker and motor concentrations of actomyosin networks without Arp2/3 (A) and with Arp2/3 (B). In region (1) the system is not gelled. In region (2) the system is not gelled only by linker connections, but the system is connected fully when we also consider the motor connections. In region (3) the system is gelled just by linkers alone.  $[L]_T$  is the total linker concentration,  $[M]_T$  is the total motor concentration, and  $[G]_T$  is the total actin concentration.

When the individual linker connections are flexible, the number of degrees of freedom of the system depends on the number and the rigidity of individual linker connections. A totally rigid connection takes away 6 degrees of freedom from the system, while a connection that only preserves the distance between two monomers would only take away one degree of freedom from the system. Fascin, a small globular crosslinker, creates rigid connections [145] that take away 6 degrees of freedom once formed, while forming a crosslink with  $\alpha$ -actinin, takes away only 1 degree of freedom when the system is at rest, since the actin-binding domain of

alpha actinin can rotate and bend with respect to the rod domain [146]. In general, as the linker connections become more flexible, more linker connections are required to reach the rigidity percolation transition (Fig 2.9).

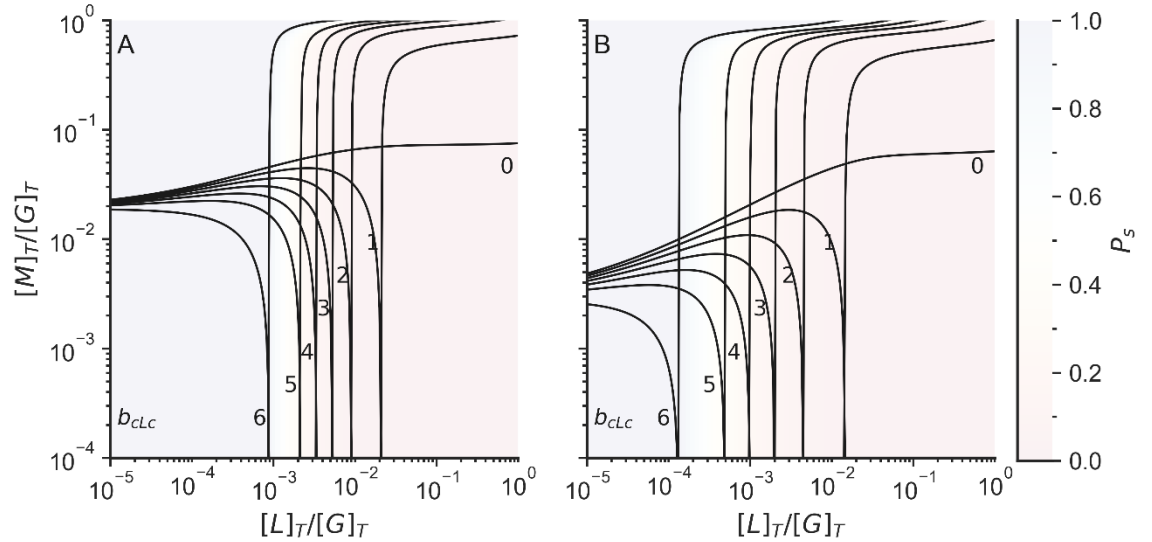


Fig 2.9.- Rigidity percolation limits including both motor and linker crosslinks (left lines) or those found when only considering linker crosslinks (right lines). The number indicates the rigidity of the crosslinks ( $b_{cLc}$ ). We assume that the connections between plus and minus sites, the connections between binding sites and minus sites, and the connections between motors and binding sites are rigid ( $b_{p \rightarrow m} = b_{c \rightarrow m} = b_{cMc} = 6$ ).  $[L]_T$  is the total linker concentration,  $[M]_T$  is the total motor concentration, and  $[G]_T$  is the total actin concentration. The color of the background indicates the probability that an actin is in a finite cluster ( $P_s$ ).

## 2.3. Discussion

### 2.3.1. Arp2/3 changes the requirements for the network to percolate

There are 3 possible regimes of mechanical behavior that depend on connectivity. In the regime where neither linkers nor motors form a percolated

cluster, any forces exerted by the motors cannot be transmitted through the system over large distances; the system therefore does not contract and will exhibit only local fluctuations. In the regime where the linkers alone do not form a percolation cluster, but the motors and linkers together do, the system is not yet rigid, but the forces can be transmitted throughout the system, allowing global contraction. In the regime where the linkers by themselves percolate, the system becomes highly rigid so that the motors are unable to contract the system. At high motor concentrations, motors also act as crosslinkers and can form a percolation cluster and allow by themselves contraction. This picture obtained from the Flory-Stockmayer kinetic analysis agrees with experimental results that indicate that contraction can occur at high motor concentrations even without the presence of other crosslinkers [119].

Arp2/3 increases the connectivity and the rigidity of the network, allowing the system to exhibit global contraction at smaller concentrations of crosslinkers and motors. Arp2/3 also makes the network rigid at smaller concentrations of linkers. At high concentrations of Arp2/3, however Arp2/3 reduces the average size of the filaments, and when the concentration of Arp2/3 becomes larger than the concentration of F-actin monomers, the network becomes disconnected. The limit for rigidity percolation coincides with the limit for connectivity percolation when the individual motor and linker connections are rigid by themselves.

Each newly formed connection between F-actin monomers and an actin cluster adds a new monomer and six degrees of freedom to the cluster in the Bethe lattice percolation model. When these connections are rigid, each connection also

removes six degrees of freedom from the cluster, keeping the cluster as a rigid object. In contrast when the connections are flexible so that monomers can bend or slide while remaining together, each connection removes only up to five degrees of freedom, allowing the cluster to remain flexible.

If the crosslinks formed by the linkers and the motors are rigid, the three connectivity percolation regimes shown in Fig 2.10 coincide with the rigidity percolation regimes. The rigidity assumption is valid if the linkers and the motors do not bend or slide along filaments. The assumption of complete rigidity may be strictly valid for only some crosslinkers such as fascin and espin. Arp2/3 branchers and the connections formed between actin monomers act as rigid connections. For more flexible crosslinkers such as  $\alpha$ -actinin and filamin, the rigidity approximation does not hold since these types of crosslinkers show high flexibility between their actin-binding domains and the rod domains [146]. Rigidity and force propagation through other mechanisms in the actin network, such as hydrodynamics, could also play a significant factor in contraction. These mechanisms are outside of the scope of this paper.

An actin system having only highly flexible linkers will not reach rigidity percolation regimes unless closed loops are formed in the system. The chemical kinetics model combined with the Flory-Stockmayer theory, is implemented on a Bethe lattice, which does not contain closed loops. Bethe lattice models can be made to account for rigid percolation regimes by anchoring multiple monomers to a single boundary, as shown in previous literature [138,147]. A complete theoretical

treatment of a rigidity percolation model of actomyosin networks must deal with the formation of closed loops in such networks.

### **2.3.2. The actin network is in the sol state when the linker concentration is much greater than the actin concentration**

In our previous works [59], we modeled linker binding as a termolecular reaction in which a linker must simultaneously bind two actin filaments, forming a crosslink. Termolecular reactions in biology can however be decomposed into two separate bimolecular steps. Here we explored the behavior of a non-cooperative linker binding where the binding sites are distributed homogeneously. The binding rate constants for these actin-binding domains with actin filaments are independent of each other. Under this condition, actin-binding domains of different linkers compete for filament binding, inhibiting the formation of crosslinks between two distinct filaments in the system (see Fig 2.2). This mechanism causes the network to remain a sol when the linker concentration is much greater (10-100 times) than the actin concentration (see Fig 2.4). However, previous experiments have shown that the actin-binding domains of some linkers such as  $\alpha$ -actinin and filamin bind independently to actin filaments in a cooperative way, when actin bundles form [106]. A full theoretical treatment of cytoskeletal percolation must deal both with the bundling of actin filaments and the cooperativity of linker binding.



### **2.3.3. The difference between the chemical kinetic model and MEDYAN can be attributed to the heterogeneous distribution when the system is percolated**

Despite the results from chemical kinetic differential equations and from MEDYAN showing similar trends, there are noticeable differences of the transient concentrations of plus sites bound to minus sites ( $[F_m \cdot F_p]$ ), bound linkers ( $[F_c \cdot L \cdot F_c]$ ), bound branchers ( $[F_c \cdot B \cdot F_m]$ ) in the systems between the two models. These differences occur once the system has formed a percolation cluster. The differences for the transient concentration of plus sites bound to minus sites ( $[F_m \cdot F_p]$ ) are caused by the diminishing polymerization rate of those filaments that are near the boundary due to the mechanochemical feedback in MEDYAN. An infinite system without boundary has been assumed in the chemical kinetic model, therefore this mechanical feedback does not occur in that model. We attribute the discrepancy in the transient concentrations of the bound linkers ( $[F_c \cdot L \cdot F_c]$ ) to the heterogeneous distribution of the binding sites in the system in MEDYAN. In the chemical kinetic model, in contrast a homogeneous distribution of binding sites is assumed.

There are two types of structural connections formed between actin monomers: lateral connections and longitudinal connections. In this work, we have only used only the lateral connection in our model to calculate the connectivity percolation. This assumption excludes cyclic connections which are not defined in this version of the Flory-Stockmayer theory [70]. There is also a slight difference between the number of binding sites and the number of actin monomers on a

filament since binding sites may exist on the interface of two or more actin monomers instead of on a single F-actin monomer. Nevertheless, as the length of an actin filament increases, this difference tends to be negligible. In the chemical kinetic description, we have not included the connections of a single linker or a motor to a single binding site ( $F_c \cdot L$  and  $F_c \cdot M$ ). These connections do not alter the connectivity percolation since they do not connect actin monomers to larger networks. We have also simplified the connections made by a brancher to a single binding site ( $F_c \cdot B \cdot F_m$ ). We recognize that it is possible that a brancher may simultaneously connect to multiple F-actin monomers. Nevertheless, the connections between branchers, mother filaments, and daughter filaments do not alter the connectivity percolation since actin monomers attached to branchers are already in the same cluster.

The heterogeneous distribution of binding sites in MEDYAN causes fewer binding sites to be available to be bound by linkers due to the small search distance of the  $\alpha$ -actinin linkers ( $d_c^{\min} = 30$  nm,  $d_c^{\max} = 40$  nm). The difference in the concentration of bound branchers ( $[F_c \cdot B \cdot F_m]$ ) can be explained because of the lower concentration of non-polymerized G-actin in the chemical kinetic model. This lesser concentration of G-actin is due to a faster effective polymerization rate in the chemical kinetic model than that of the MEDYAN model since collisions between actin filaments and the boundary are not represented in the chemical kinetic model. The faster effective polymerization rate leads to an increase of available binding sites ( $[F_c]$ ) in the chemical kinetic model, which in turn allows the branching reaction to occur earlier. Finite-size effects exist in MEDYAN. While such effects can

be reproduced in a chemical kinetic model, we consider the chemical kinetic model, which assumes an infinite system, to be sufficient to explain the processes happening during the early stages of the connectivity percolation in the network.

## **2.4. Conclusion**

In this work, we have used a generalization of the Flory-Stockmayer theory of percolation to show three possible phase regimes for a cytoskeletal network depending on the connectivity achieved by motors, linkers and branchers. These connections give rise to local or global contraction depending on the percolation regime. Actin-binding proteins modulate the structure and dynamics of the network, allowing the cell to exhibit different behavior and functions. We also show that Arp2/3 increases the connectivity of the network when the concentration of Arp2/3 is lower than the concentration of F-actin monomers.

## **2.5. Methods**

We have modeled reactions between actin filaments (F-actin), monomeric globular actins (G-actin), and actin-binding proteins using either a mass action chemical kinetics model described by a system of ordinary differential equations, which we call the chemical kinetic model, or a stochastic mechanochemical model (MEDYAN). We have quantified the number of connections between F-actin monomers in the system and then used a mean field model based on the Flory-

Stockmayer theory [68–70] to calculate whether the system should behave like a liquid (sol state) or like a semi-solid (gel state).

### **2.5.1. Mass action chemical kinetics model**

We have modeled five actin binding and unbinding processes in the actomyosin networks using a chemical kinetic model based on mass action kinetics. The equations from the chemical kinetic model parallel the chemical reactions described in MEDYAN [63–67], a mechanochemical model of actomyosin networks detailed in a section below. The reactions, shown in Table 2.1, include the polymerization and depolymerization at both the plus ends ( $F_p$ ) and the minus ends ( $F_m$ ) of actin filaments, along with the binding and the unbinding of  $\alpha$ -actinin linkers (L), NMIIA motors (M), and the Arp2/3 complex (B) to F-actin. To account for the NMIIA minifilaments, a motor (M) consists of 22.5 myosin molecules, which corresponds to the average number of motor heads in the MEDYAN model [59,66]. The rates for these reactions are shown in Table A in S1 Text. The chemical kinetic model assumes a homogeneous system with perfect mixing, and we do not model the spatial effects of the diffusion of chemical species. The chemical kinetic model also assumes an infinite volume, so the stochasticity of the processes is averaged out.

Reaction	Description
$F_p + G \xrightleftharpoons[k_p^-]{k_p^+} F_p \cdot F_m + F_p + F_c$	Actin polymerization at the plus end
$F_m + G \xrightleftharpoons[k_m^-]{k_m^+} F_p \cdot F_m + F_m + F_c$	Actin polymerization at the minus end
$F_c + B + G \xrightleftharpoons[k_B^-]{k_B^+} F_c \cdot B \cdot F_m + F_p + F_c$	Brancher binding
$F_c + L + F_c \xrightleftharpoons[k_C^-]{k_C^+} F_c \cdot L \cdot F_c$	Linker binding
$F_c + M + F_c \xrightleftharpoons[k_M^-]{k_M^+} F_c \cdot M \cdot F_c$	Motor binding

Table 2.1. Reactions included in the chemical kinetic model.

We have used mass-action kinetics to model the binding of G-actin (G), motors (M), linkers (L), and branchers (B) to actin filaments (F) as well as their unbinding. The kinetic equations replicate the MEDYAN reactions, where a three-body reaction takes place (see section I.A of the Appendix A for more details). We have also included a two-step binding reaction of linkers (L) to actin filaments (F) based on experimental observations (see section I.B of the Appendix A for more details). We have defined three different interfaces for the F-actin monomers: the plus site ( $F_p$ ), the minus site ( $F_m$ ), and the ABP binding site ( $F_c$ ) since actin, motors,

linkers, and branchers can be connected to the actin filament monomers through these interfaces as shown in Fig 2.10.

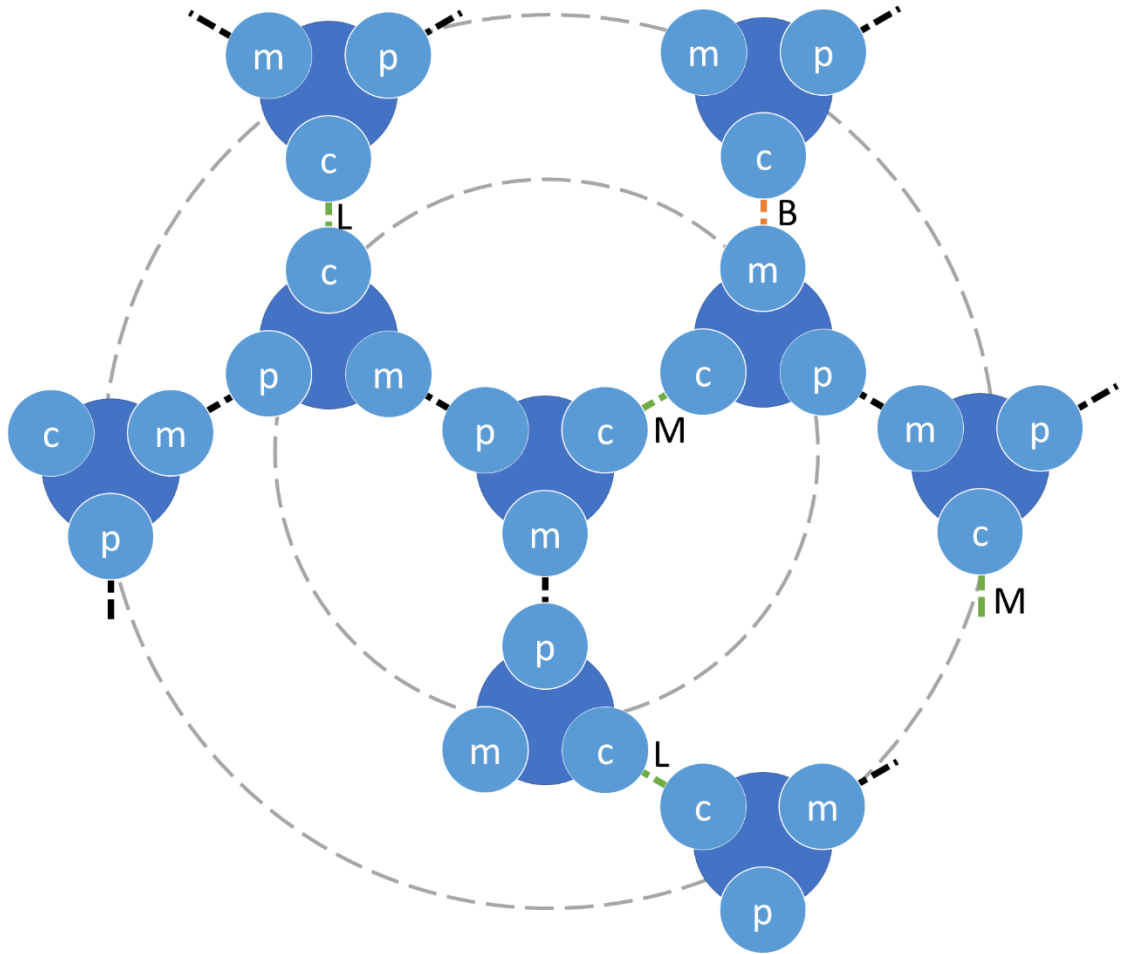


Fig 2.10. Diagram of connections of F-actin monomers to other F-actin monomers. The F-actin monomers are shown in blue and have 3 sites: the plus site (p), the minus site (m), and the binding site (c). The dotted lines indicate connections from the site of an F-actin monomer to another F-actin monomer. The connections are formed by polymerization (black dotted lines), linkers (L), motors (M) or branchers (B). The actin cluster can be represented as a treelike cluster, where the particle in the center is the root,

and can be connected to up to 3 particles in the first layer, 6 particles in the second layer, and so on.

### 2.5.2. Flory-Stockmayer Theory

The theory of Flory and Stockmayer [68,69] describes the conversion of monomers first into soluble low molecular weight polymers and then into heavy insoluble gels by branching and crosslinking. In these theories the polymers are described using a mean field model, where the probability of finding a bound functional group depends only on the nature of the functional group. When the fraction of reacted polymers reaches a threshold, then the weighted average monomer size becomes infinite.

Tavares et al. recently developed a generalization of the Flory-Stockmayer theory to describe patchy colloids [70,71], which can be understood as polyfunctional branched monomers. We have used this generalization of the Flory-Stockmayer theory to calculate when the percolation transitions occur. The crosslinking probabilities ( $\theta_{\alpha \rightarrow \beta}$ ) were calculated using the ratio of the concentration of species in the bound state ( $[\alpha \cdot \beta]$ ) to the total concentration of the species ( $[\alpha]_T$ ) as quantified from the chemical kinetic model or the MEDYAN simulations (Equation 2.1).

$$\theta_{\alpha \rightarrow \beta} = \frac{[\alpha \cdot \beta]}{[\alpha]_T} = \frac{[\alpha \cdot \beta]}{[\alpha] + [\alpha \cdot \beta]} \quad \text{Equation 2.1}$$

Where  $\theta_{\alpha \rightarrow \beta}$  is the probability of having an F-actin monomer connected through the site  $\alpha$  to the site  $\beta$  of another F-actin monomer, and  $\alpha$  and  $\beta$  can be the plus site (p), the minus site (m), or the actin-binding site (c) (Fig 2.10).

The probabilities of having an F-actin monomer connected to another F-actin monomer from one site to another are shown in Equation 2.2.

$$\begin{aligned}
 \theta_{p \rightarrow m} &= \frac{[F_p \cdot F_m]}{[F_p \cdot F_m] + [F_p]} \\
 \theta_{m \rightarrow p} &= \frac{[F_p \cdot F_m]}{[F_p \cdot F_m] + [F_c \cdot B \cdot F_m] + [F_m]} \\
 \theta_{c \rightarrow c} &= \frac{2[F_c \cdot L \cdot F_c] + 2[F_c \cdot M \cdot F_c]}{2[F_c \cdot L \cdot F_c] + 2[F_c \cdot M \cdot F_c] + [F_c \cdot B \cdot F_m] + [F_c]} \\
 \theta_{c \rightarrow m} &= \frac{[F_c \cdot B \cdot F_m]}{2[F_c \cdot L \cdot F_c] + 2[F_c \cdot M \cdot F_c] + [F_c]} \\
 \theta_{m \rightarrow c} &= \frac{[F_c \cdot B \cdot F_m]}{[F_p \cdot F_m] + [F_c \cdot B \cdot F_m] + [F_m]}
 \end{aligned}
 \tag{Equation 2.2}$$

Where  $\theta_{p \rightarrow m}$  and  $\theta_{m \rightarrow p}$  are the probability that an F-actin monomer plus site ( $F_p$ ) is connected to the minus site ( $F_m$ ) of another F-actin monomer and vice versa through actin filament polymerization. The probability that F-actin monomer binding site ( $F_c$ ) is connected to the binding site ( $F_c$ ) of another F-actin monomer is denoted as  $\theta_{c \rightarrow c}$ . Connections through the binding sites are formed by motor and linker binding. Finally,  $\theta_{c \rightarrow m}$  and  $\theta_{m \rightarrow c}$  are the probability that F-actin monomer binding site ( $F_c$ ) is connected to the minus site ( $F_m$ ) of another F-actin monomer, and vice versa through brancher binding.

$[F_p \cdot F_m]$  is the concentration of plus sites bound to minus sites, as in polymerized F-actin,  $[F_c \cdot L \cdot F_c]$  is the concentration of pairs of F-actin monomers bound through the binding sites with linkers,  $[F_c \cdot M \cdot F_c]$  is the concentration of pairs of F-actin monomers bound through the binding sites with motors, and  $[F_c \cdot B \cdot$



$F_m]$  is the concentration of F-actin monomers bound with a brancher.  $[F_p]$  is the concentration of unbound plus sites of F-actin monomers,  $[F_m]$  is the concentration of unbound minus sites of F-actin monomer and  $[F_c]$  is the concentration of unbound binding sites. The probability that an F-actin is connected to an infinite cluster is detailed in the section I.C of the Appendix A, the solution for a simple case of linkers and actin is detailed in the section I.D of Appendix A.

### 2.5.3. Rigidity Percolation

To understand how a rigid lattice is formed in the network we use Maxwell counting, which has also been used as a first step to understand the rigidity of glasses [138,139,147–153]. In short this counting procedure is based on the fact that the number of floppy modes per connection ( $f$ ) is related to the number of degrees of freedom per F-actin monomer ( $g$ ) minus the number of constraints given by other connections. (Equation 2.3)

$$f = g - \sum_{\alpha}^z \theta_{\alpha} b = 6 - \frac{1}{2} \sum_{\alpha} \sum_{\beta} b_{\alpha \rightarrow \beta} \frac{\theta_{\alpha \rightarrow \beta} \left(1 - \frac{P_s}{Q_{\alpha}} \frac{P_s}{Q_{\beta}}\right)}{(1 - P_s)} \quad \text{Equation 2.3}$$

Where  $p$  is the probability of forming a contact,  $z$  is the coordination number and  $b$  is the number of constraints given by the connection. Every F-actin monomer has 6 degrees of freedom in 3D space ( $g = 6$ ): 3 translational degrees of freedom and 3 rotational degrees of freedom. In addition, the F-actin can connect through 3 possible sites ( $z = 3$ ). When an F-actin monomer connects to another F-actin monomer, the system loses degrees of freedom depending on the rigidity of the connection,  $b$ . For this model we considered that each direct connection between

two F-actin monomers accounts for a loss of 6 degrees of freedom ( $b = 6$ ) since we assume each actin filament is a rigid object. For example, when there is no connection between two F-actin monomers there would be two separate filaments with 6 degrees of freedom each (a total of 12 degrees of freedom). When the connection forms, the system contains only one rigid filament with 6 degrees of freedom.

Similarly, rigid connections of two F-actin monomers with linkers, motors, and branchers also account for a loss of 6 degrees of freedom ( $b = 6$ ). The connectivity percolation is the same as the rigidity percolation ( $b = g$ ) when the connections are rigid.

On the other hand, when the connections of two F-actin monomers with linkers and motors are floppy ( $b < 6$ ) the connectivity percolation is not the same as the rigidity percolation, since more than one connection is needed to make the system rigid.

#### **2.5.4. Coarse-grained mechanochemical model of actomyosin systems (MEDYAN)**

We have used an elegant coarse-grained mechanochemical model of actomyosin systems called MEDYAN (Mechanochemical Dynamics of Active Networks) developed by Papoian and his group [63–67]. MEDYAN models both stochastic chemical reactions and deterministic mechanical representations of far-from-equilibrium systems. In this study, we have included some important actin-

binding proteins in actomyosin networks: non-muscle myosin IIA heavy chain (NMIIA) motors,  $\alpha$ -actinin linkers, and actin-related protein complex 2/3 (Arp2/3) branchers, all in a fixed geometry (See section I.E of the Appendix A for more details).

### **2.5.5. Acknowledgements**

We thank both Neal Waxham and Garegin Papoian for their helpful discussions and acknowledge with gratitude the use of the MEDYAN code provided by Dr. Papoian. CB, JL, NPS, MSC, and PGW were supported by the NSF Division of Chemistry RAISE grant 1743392 and by the Center for Theoretical Biological Physics, sponsored by the NSF Division of Physics grant 2019745. CB, PGW were supported by the PoLS Student Research Network sponsored by the NSF Division of Physics grant 1522550. PGW was supported by the D. R. Bullard-Welch Chair at Rice University, Grant C-0016 and by the Welch Foundation (grant C-1792). JL, MSC had additional support to use the uHPC and Sabine clusters managed by the Core facility for Advanced Computing and Data Science at the University of Houston and sponsored by the NSF Office of Advanced Cyberinfrastructure grant 1531814. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

# **The role of the Arp2/3 complex in shaping the dynamics and structures of branched actomyosin networks**

This chapter is a slightly modified version of an article published in the Proceedings of the National Academy of Sciences [59].

### **3.1. Introduction**

Actomyosin networks are systems of actin filaments that are organized both by their interaction with myosin and passive cross-linkers. The non-equilibrium activity of actomyosin networks gives cells the ability to grow, move and divide [105,154]. The contraction of actomyosin networks has been recapitulated *in vitro* using reconstituted solutions of actin filaments containing several types of actin-binding proteins that merely act as cross-linkers along with active myosins [26,37,38,155]. Contractility of the network appears to require a threshold concentration of myosin and the presence of ATP but is only seen at an intermediate concentration of bivalent cross-linkers. Early studies of reconstituted solutions of

actin filaments *in vitro* generally lacked many of the key constituents of the cytoskeleton, notably the actin-related protein 2/3 (Arp2/3) complex. The Arp2/3 complex specifically alters the topology of actin filament networks by forming branches. One of many examples of actomyosin networks with the Arp2/3 complex is the dendritic spine. Dendritic spines are small membranous actin-filled protrusions attached to neuronal dendrites whose morphological plasticity is commonly hypothesized to underlie learning and retrieving memory [40,154,156]. While the structure of the Arp2/3 complex has been explored extensively [33,106,115,116,118,157–161], how the Arp2/3 complex changes the architectural dynamics of the actomyosin network raises many questions. Indeed, recent experimental studies have shown that the Arp2/3 complex qualitatively changes the dynamics of the network [26,38,162].

Here, we employ a powerful computational software for flexibly modeling the complexity of cytoskeletons, MEDYAN, that was developed by Papoian and his group [63–67]. The MEDYAN computational framework incorporates a stochastic description of individual chemical reaction events appropriate to the nanoscale and simultaneously employs a deterministic treatment of the mechanics of the cellular assembly at micron scales through cycles of mechanochemical feedback. Using a predecessor of the MEDYAN model, Papoian and his group have modeled unmotorized branched actin networks and have studied the effects of branchers and capping proteins on the speed of protrusion of a flexible membrane [163–166]. The mechano-chemical aspects of MEDYAN make it possible to investigate how the nucleation and branching initiated by the Arp2/3 complex change the dynamics and

structures of actomyosin systems. Here, we report on a computational investigation of how branchers influence network contractility.

Several other computational frameworks have been developed for modeling the dynamics of actomyosin networks [34,61,62,125,167]. One code, Cytosim, models the actomyosin network at the mesoscopic scale [62] but lacks the mechanochemical feedback which is critical to describe active processes. Another code, AFiNES [61], has been used to study a two-dimensional model of an actomyosin network and describes network rheology under mechanical stress. AFiNES lacks, however, the volume exclusion for actin filaments which we believe is important for correctly determining the mechanics of jammed states. MEDYAN captures all these features which we believe are crucial to understanding the actual biological system.

Contractile motions of actomyosin networks can be caused by the forces generated by motors (across actin filaments) but may also be driven by compression and expansion (along actin filaments) due to polymerization. These forces deform the network and build up strains through the expenditure of free energy [168,169]. These strains then feed back into various biochemical processes. Our focus is on exploring the contrast between the dynamics of branched and unbranched networks, via conducting simulations both with and without Arp2/3. We show that the branched networks (with Arp2/3) can become marginally unstable. Fluctuations continually cause local collapse events which are directly connected to this marginal stability. In systems that are far-from equilibrium and

active, marginal stability does not occur at a specific concentration but depends on both the initial configuration of the network and the history of the system. The marginal stability of branched actomyosin assemblies resembles what is found in jammed granular assemblies and colloidal glasses [134,170,171].

## 3.2. Results

### 3.2.1. Linkers modulate contraction in unbranched actomyosin networks

Our basic approach is the simulation of actin molecular assemblies in solution with various mixtures of actin binding proteins, all in a fixed geometry. To study unbranched networks, we considered nine different concentration ratios of non-muscle myosin IIA heavy chain (NMIIA) motor proteins to actin monomers ( $x_{m:A}$ ) and eight concentration ratios of  $\alpha$ -actinin linker molecules to actins ( $x_{\alpha:A}$ ) in our simulations. In total then we studied 72 distinct actomyosin network assembly scenarios. All simulations were confined to a  $1 \mu\text{m} \times 1 \mu\text{m} \times 1 \mu\text{m}$  box with  $25 \mu\text{M}$  of actin in total. The concentration of actin,  $25 \mu\text{M}$ , was specifically chosen to replicate the *in vitro* experiments from the Weitz group [37]. The monomers, the filaments, the motors, the linkers, and the branchers were initially distributed randomly within a cubic container. Other various physical parameters characterizing simulated systems are described in detail in Table S3 in Appendix B. In our discussion we will highlight four representative conditions of  $x_{m:A}$  and  $x_{\alpha:A}$  from these scenarios to elucidate the role of linkers in unbranched networks: (1) low motor and low linker concentrations ( $x_{m:A} = 0.01$  and  $x_{\alpha:A} = 0.01$ ), (2) high motor

and low linker concentrations ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.01$ ), (3) high motor and high linker concentration ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.5$ ), and (4) medium motor and medium linker concentrations ( $x_{m:A} = 0.05$  and  $x_{\alpha:A} = 0.1$ ). Note that for every combination of motor and linker to actin ratios, sixteen simulation replicates were performed, differing only in their random initializations. The average length of a filament in unbranched networks is approximately  $0.85 \mu\text{m}$ . Throughout the simulations, we recorded the structure of the networks once every 10 seconds. In addition to plotting the time courses of these parameters, in some cases we have also plotted the changes of the parameters that take place between successive pairs of snapshots to highlight when large sudden changes in the parameters occur.

In order to follow the assembly process, we show in Figure 3.1 how the radius of gyration of the whole system ( $R_g$ ) normalized by the “initial”  $R_g$  value ( $R_g^i$ ),  $R_g/R_g^i$ , changes with time.  $R_g^i$  is actually the radius of gyration measured after an initialization run without motors of 10 seconds, so as to allow the filament length distribution to reach a steady-state before the motors are turned on. When an actomyosin network contracts due to the motor action, the radius ratio  $R_g/R_g^i$  will become less than 1, while, in contrast, when  $R_g/R_g^i$  becomes greater than 1 it indicates that the system has expanded during assembly. Unbranched actomyosin systems with low motor concentration ( $x_{m:A} = 0.01$ ) and low linker concentration ( $x_{\alpha:A} = 0.01$ ) expand rather slowly (Figure 3.1A) in comparison with what occurs for the other three conditions. The expansions are caused by actin polymerization



without any significant countervailing forces being exerted by motors that would pull the actin filaments together.

Adding linkers leads to a competition between actin polymerization and motor pulling. This competition accounts for the differing extent of contraction. We calculated  $\langle R_g/R_g^i \rangle$ , which is the average of the normalized radius of gyration ratio taken over all replicates for a given condition, varying motor and linker concentrations for both unbranched and branched networks. The radius ratios  $\langle R_g/R_g^i \rangle$  achieved at steady state for the high motor concentration and high linker concentration systems ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.5$ , Figure 3.1C) are approximately 16% lower than the steady state ratios  $\langle R_g/R_g^i \rangle$  for high motor concentration and low linker concentration systems ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.01$ , Figure 3.1B). Contractility is only manifested for unbranched systems when the motor concentration exceeds a threshold of  $x_{m:A} = 0.01$  (Figure 3.1B, C, and D).

Assembly dynamics depends on the fraction of the linkers that are unmotorized. For the unbranched systems with  $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.01$  (high motor and low linker concentrations), whose time-courses of the radius ratio  $R_g/R_g^i$  are depicted in Figure 3.1B, the pulling forces from motors lead to the rapid reduction in the radius ratio  $R_g/R_g^i$  during the first 130 seconds. When we increased the concentration of linkers to  $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.5$ , the actin network contracts much more slowly as shown in Figure 3.1C. These results semi-quantitatively agree with the experimental data from the Weitz lab [37] who studied these unbranched system.

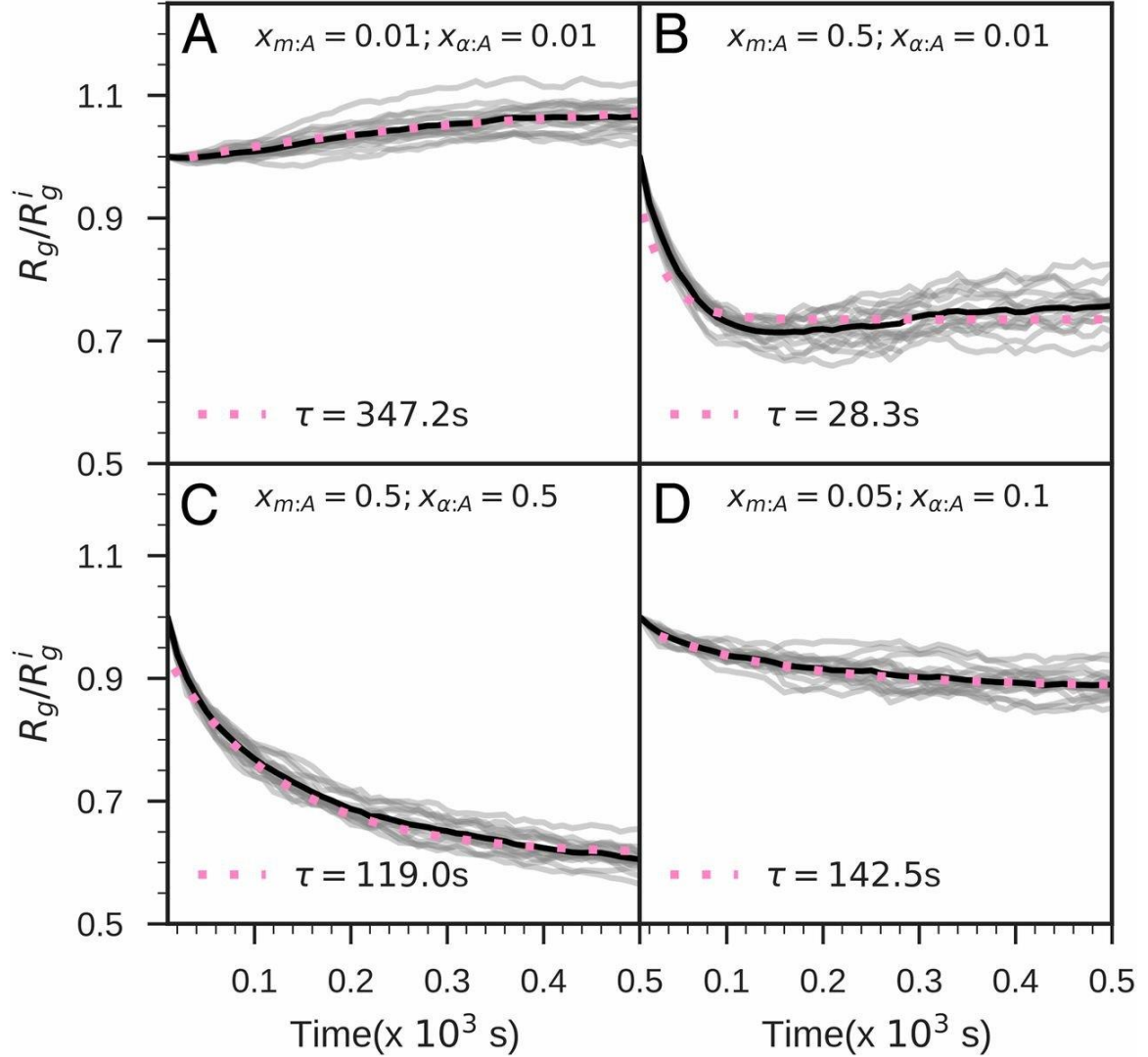


Figure 3.1. Time courses of the ratio  $R_g/R_g^i$  indicate contractile or extensile motion in unbranched actomyosin networks at several conditions of motors and linkers.

(A) Systems with low motor and low linker concentrations ( $x_{m:A} = 0.01$  and  $x_{\alpha:A} = 0.01$ ).  
 (B) Systems with high motor and low linker concentrations ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.01$ ).  
 (C) Systems with high motor and high linker concentrations ( $x_{m:A} = 0.5$  and  $x_{\alpha:A} = 0.5$ ).  
 (D) Systems with medium motor and medium linker concentrations ( $x_{m:A} = 0.05$  and  $x_{\alpha:A} = 0.1$ ). The pink-dotted lines show the single exponential fits to the sets of the time courses of  $R_g/R_g^i$  for each simulation condition. The traces for simulation replicates are indicated with light gray lines and their averages are shown in black lines.  $\tau$  is the single exponential fitting time constant (see Appendix B for more details).

### 3.2.2. Branched actomyosin networks display convulsive movements

Next, we included Arp2/3 protein complexes to simulate branched networks. The radius of gyration ratios  $R_g/R_g^i$  for the branched networks (Figure 3.2) behave differently during assembly from what was seen for the unbranched networks (Figure 3.1). The branched networks with low concentrations of motors and low concentrations of linkers ( $x_{m:A} = 0.01$  and  $x_{\alpha:A} = 0.01$ , Figure 3.2A) contract rather than expand as did the unbranched assemblies (Figure 3.1A). The overall contraction of the branched assemblies results from the creation of branches that inhibit actin polymerization and depolymerization. Adding a brancher prevents the actin depolymerization reaction at the minus end of a daughter filament because the brancher positions itself at the branch junction once a daughter filament has been created from its mother filament. When there is depolymerization at the minus end of a filament that is faster than the dissociation of a brancher from the filament, the brancher will inhibit the depolymerization reactions at the minus end of a mother filament. Actin polymerization and depolymerization which require actin turnover are thus significantly inhibited by the presence of branchers. The concentration of filamentous actin (F-actin) in the branched simulations is larger than the concentration of F-actin in the unbranched simulations. The increased F-actin concentration is due to the capping of the minus end of the mother and the daughter filaments, as well as the new F-actin plus ends created by Arp2/3 nucleation. Nevertheless, due to the increased number of total filaments, the average length of a filament is in fact reduced when Arp2/3 is added. The average length of a filament in branched networks is approximately  $0.16 \mu\text{m}$ . Unbranched networks turn out to

consist of long and parallel filaments displaying liquid crystalline order, while branched networks are comprised of short and bifurcated filaments and appear more nearly isotropic.

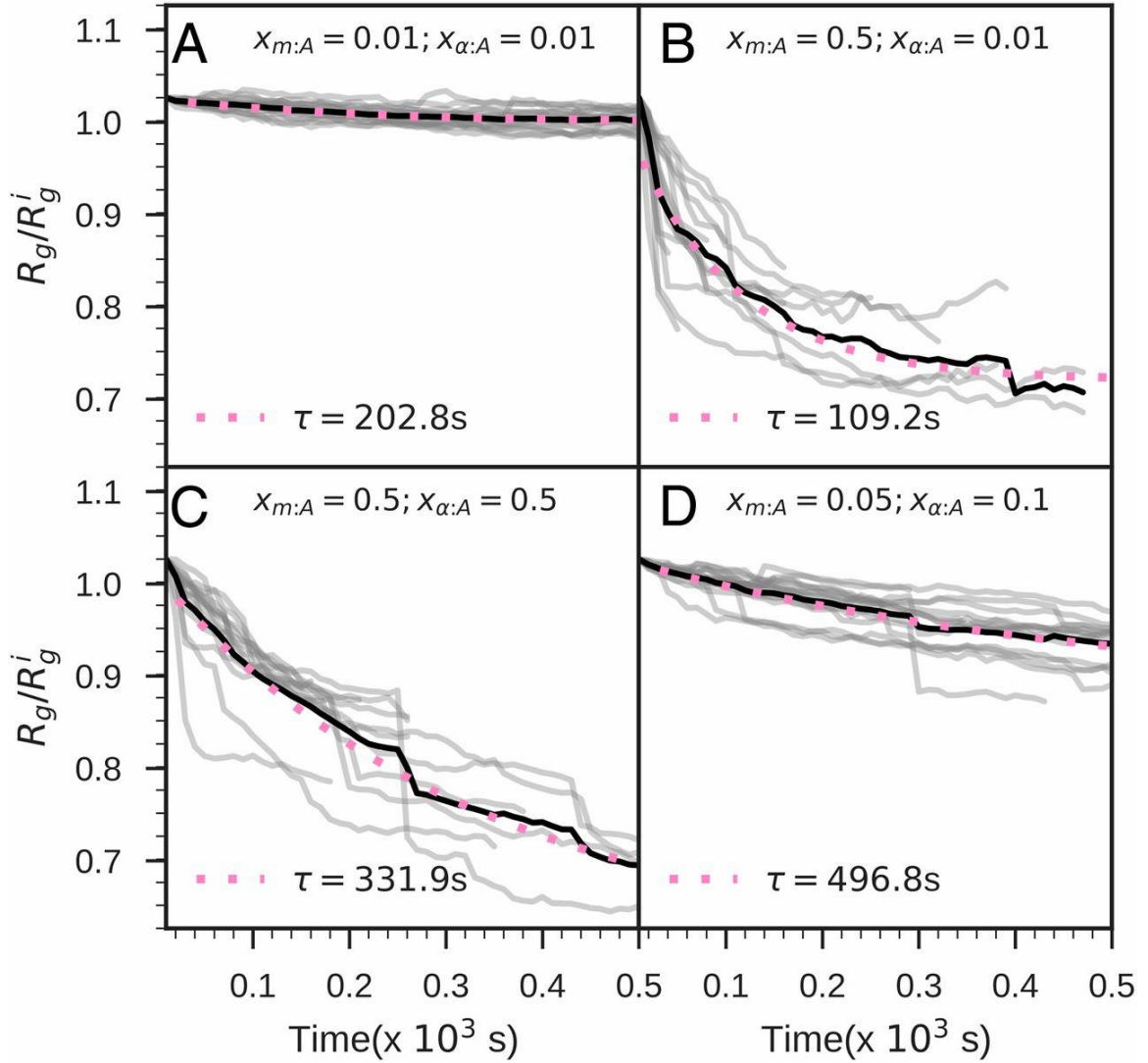


Figure 3.2. The time courses of radius gyration ratios for contractile and extensile motions in branched actomyosin networks at several conditions of motors and linkers. (A-D) The indications for concentrations of motors and linkers are the same as those used in Figure 3.1.  $\tau$  is the single exponential fitting time constant (see SI for more details).

The contraction of branched networks occurs in a more irregular fashion than does the contraction of unbranched assemblies. The contraction of the assemblies is rather intermittent, occurring by unpredictable abrupt drops of the radius of gyration ratio  $R_g/R_g^i$ . These convulsive events are inherently stochastic and hence do not occur at the same time in different simulation replicate runs. These abrupt drops of  $R_g/R_g^i$  occur very quickly. Individually, these events have a duration of less than 10 seconds. The sharpest drops can be as large as 20%. These drops, which we call avalanches, are observed more frequently at higher concentrations of motors (Figure 3.2B and C) than at lower concentrations (Figure 3.2A). For the range of parameters studied in our simulations, we have never observed such discrete and large-scale avalanches in the unbranched actomyosin networks created whenever Arp2/3 was absent.

To get a better picture of the mechanism underlying the avalanches, we examined changes in the ratio of the radius of gyration after an event and we have also visualized snapshots of the system to uncover the corresponding structural changes. Snapshots for medium motor concentration and medium linker concentration ( $x_{m:A} = 0.05$  and  $x_{\alpha:A} = 0.1$ ) are shown in Figure 3.3.  $\Delta(R_g/R_g^i)$  corresponds to the successive temporal changes of the radius of gyration ratios  $R_g/R_g^i$ . An abrupt drop in the radius of gyration ratio occurs after 440 seconds (Figure 3.3A). At this point, we note that the network contains several regions of high tension. These regions are shown in a white color inside a blue square in Figure 3.3B. The tension is rather heterogeneously distributed immediately prior to the

avalanche. The high-tension regions are created by the stalling of motors. Once the motors have successfully reorganized the actin filaments, the regions of high tension disappear as shown in Figure 3.3C. No motor or linker unbinding was seen immediately prior to an avalanche. Instead a motor, which was previously walking on a short filament and on a filament bound by Arp2/3 branchers, became jammed before an avalanche was triggered.

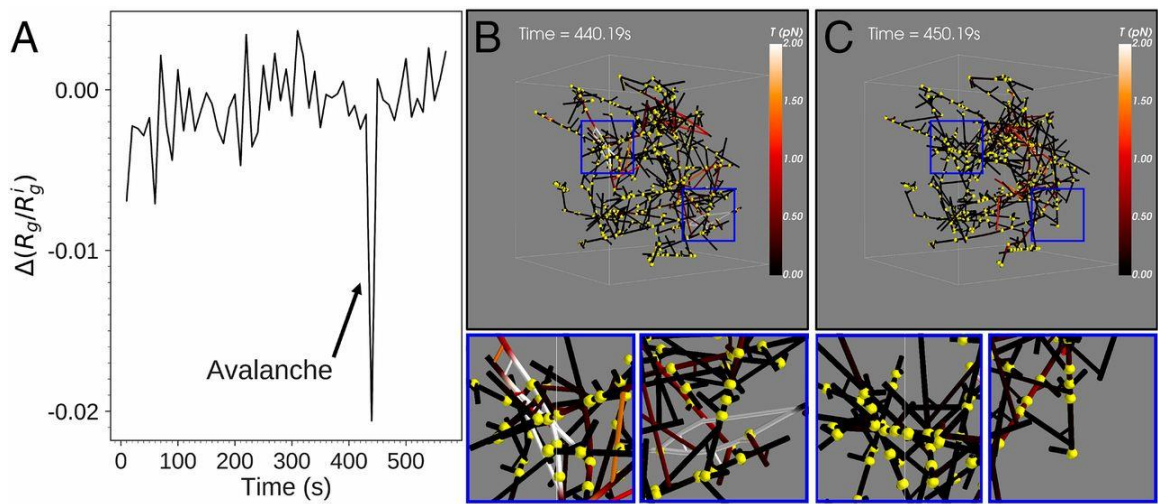


Figure 3.3. Changes in the tension of a branched network occur during an avalanche. (A) The time courses of the successive values of temporal changes over a 10 second time interval in the ratio of the radius of gyration normalized by the  $R_g$  value at 10 second. The quantity  $\Delta(R_g/R_g^i)$  is shown for a branched network simulation with medium motor and medium linker concentrations ( $x_{m:A} = 0.05$  and  $x_{\alpha:A} = 0.1$ ) over time. (B-C) Two snapshots of actin filaments, motors, and linkers where the tension is indicated by color; these show the morphology of the network before (B) and after (C) the abrupt drop of  $\Delta(R_g/R_g^i)$  when an avalanche occurs. The dimensions of the cubic simulation box are  $1 \mu\text{m} \times 1 \mu\text{m} \times 1 \mu\text{m}$ . A blue square highlights a concentrated high-tension region that can be seen prior to the  $\Delta(R_g/R_g^i)$  drop which then becomes a dispersed low-tension area after the  $\Delta(R_g/R_g^i)$  drop.

### 3.2.3. Mechanistic insights into the convulsive movements during an avalanche

The mean of the displacements of the center of mass of each actin filament at time  $t$  ( $\delta x_F(t)$ ) is defined by Equation 3.1

$$\delta x_F(t) = \frac{1}{N} \sum_i^N |CoM_i(t) - CoM_i(t-1)| \quad \text{Equation 3.1}$$

where  $CoM_i$  is the center of mass of filament  $i$  and  $N$  is the number of filaments in a system. The  $\delta x_F$  between each successive pair of snapshots (black curves) turns out to be a useful parameter to identify and characterize avalanches, as shown in Figure 3.4.

In addition, we studied the changes in the shapes of the actin assemblies using the shape parameter ( $S$ ). This is defined for an actomyosin network by Equation 3.2 [172].

$$S = 27 \frac{|\prod_{i=1}^3 (\lambda_i - \bar{\lambda})|}{(trT)^3} \quad \text{Equation 3.2}$$

$T$  is the geometrical inertia tensor of an assembly as described in Equation 3.3, while the  $\lambda_i$  are the eigenvalues of the inertia tensor  $T$ , and  $\bar{\lambda}$  is the average eigenvalue of the inertia tensor  $T$ .

$$T_{\alpha\beta} = \frac{1}{2N^2} \sum_{i,j=1}^N (r_{i\alpha} - r_{j\alpha})(r_{i\beta} - r_{j\beta}) \quad \text{Equation 3.3}$$

Here  $N$  is the number of beads in the network. Each bead represents an end of a cylindrical actin segment.  $r_{i\alpha}$  is the projection of bead  $i$  on the  $\alpha$  axis where  $\alpha$  can be  $x, y, z$ .

There appear to be two types of avalanches that display differences in their shape and in the size of the reconfigured regions. In one group, corresponding to shear events, the temporal changes are manifested by large changes of the shape parameter ( $\Delta S$ ) between successive pairs of snapshots, changes that occur through filament sliding (e.g. labeled as Avalanche 1 in Figure 3.4). A second group of avalanches involve collapses of local regions and are characterized by significant changes in the normalized radius of gyration ratio between successive pairs of snapshots (e.g. labeled as Avalanche 2 in Figure 3.4). We believe that the type of avalanches described here are examples of two extremes of a broad distribution of avalanches.



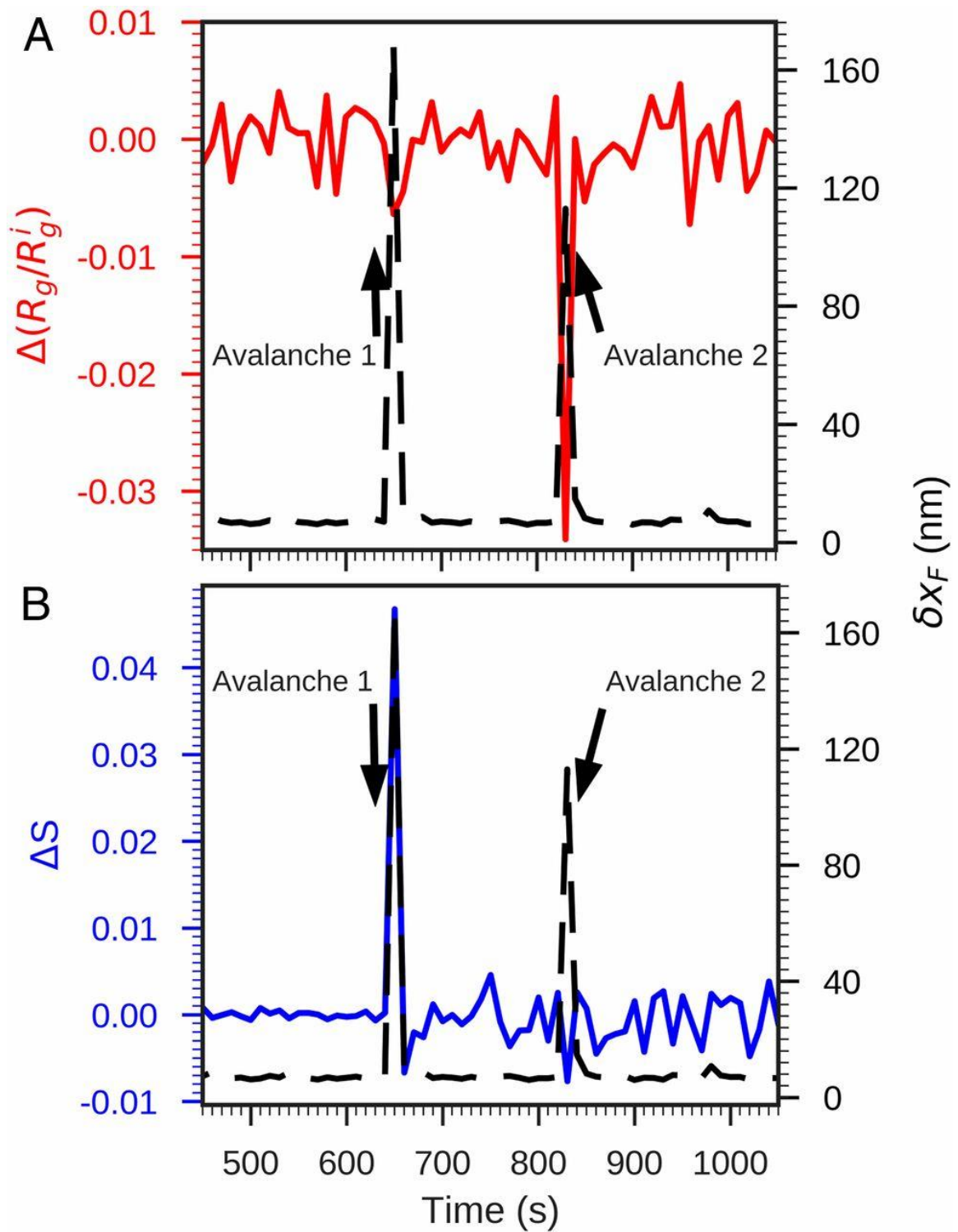


Figure 3.4. Examples of avalanches in a branched network containing 16 motors and 3000 linkers.

In Panel A, the time course of the changes in the ratio of the radius of gyration normalized by the  $R_g$  value at the initial 10 second point between successive pairs of snapshots that are separated by 10 seconds,  $\Delta(R_g/R_g^i)$ , is plotted in red. In Panel B, the time course of the changes in the shape parameter between successive pairs of snapshots that are separated by 10 seconds,  $\Delta S$ , is plotted in blue. In both Panels A and B, the time course of the mean filament displacement  $\delta x_F$  is plotted in black.

The underlying mechanism for these abrupt behaviors seems to reflect the rather wide distribution of sizes of filamentous actin (F-actin) connectivity clusters in the network. We calculated the distributions of connectivity cluster sizes and used those distributions to compute the weighted averages of the cluster size ( $N_w$ ) found in both the unbranched and the branched networks. This was done for all 72 concentration scenarios while also analyzing whether avalanches occur or not. The weighted mean cluster size ( $N_w$ ) was calculated in units of actin monomers.  $N_w$  represents the average cluster size of randomly selected monomers and is defined as the ratio between the first moment and the second moment of the cluster size distribution  $p(n)$  as shown in Equation 3.4.

$$N_w = \frac{\sum_n n^2 p(n)}{\sum_n n p(n)} \quad \text{Equation 3.4}$$

Here  $n$  is the number of actin monomers in a cluster and  $p(n)$  is the probability of finding a cluster of size  $n$  in the system.

The weighted mean cluster size of F-actins,  $N_w$ , provides a parameter describing heterogeneity of the cluster sizes of a network. Assemblies with a small value of  $N_w$  contain multiple small and isolated clusters that are not connected to the rest of the system through motors, linkers or branchers. A network with a large

value of  $N_w$  contains at least one large cluster of actins all of which are mechanically linked.

In Figure 3.5, we plot histograms of the weighted mean cluster size ( $N_w$ ) for the unbranched and branched networks. Panel A shows that for the unbranched networks, there are two predominant actin cluster sizes: one peak for small clusters having at most 2,000 actin monomers and another peak for large clusters with more than 12,000 actin monomers. The smaller connectivity clusters are typically isolated, while there is usually only a single cluster of more than 12,000 actin monomers. The unbranched networks with many small and isolated F-actin connectivity clusters generally are only weakly connected through motors or linkers, while the networks with large F-actin connectivity clusters have become fully connected thereby allowing force to be transmitted globally throughout the networks.

In Panel B, we see that branched networks have a rather wide distribution of  $N_w$  (Figure 3.5B) with a broad continuous range of intermediate size actin clusters. To characterize an event as an avalanche during an individual trajectory, we calculated the Z-score of the mean filament displacement  $\delta x_F$  between each successive pair of snapshots separated by 10 seconds. We classified an event as being an avalanche when the mean filament displacement  $\delta x_F$  between each successive pair of snapshots is sufficiently large: that is, we require the Z-score relative to the displacements for other intervals to exceed 5 in order for the interval to be classified as having an avalanche event. As shown in Figure 3.5B, there is a

significant difference in cluster size distributions for snapshots when an avalanche occurs and when there has been no avalanche. This comparison of the distributions of the actin cluster sizes suggests that forces build up in the branched systems with large connectivity clusters eventually resulting in an avalanche. Avalanches rarely occur in networks containing only small and isolated actin clusters.

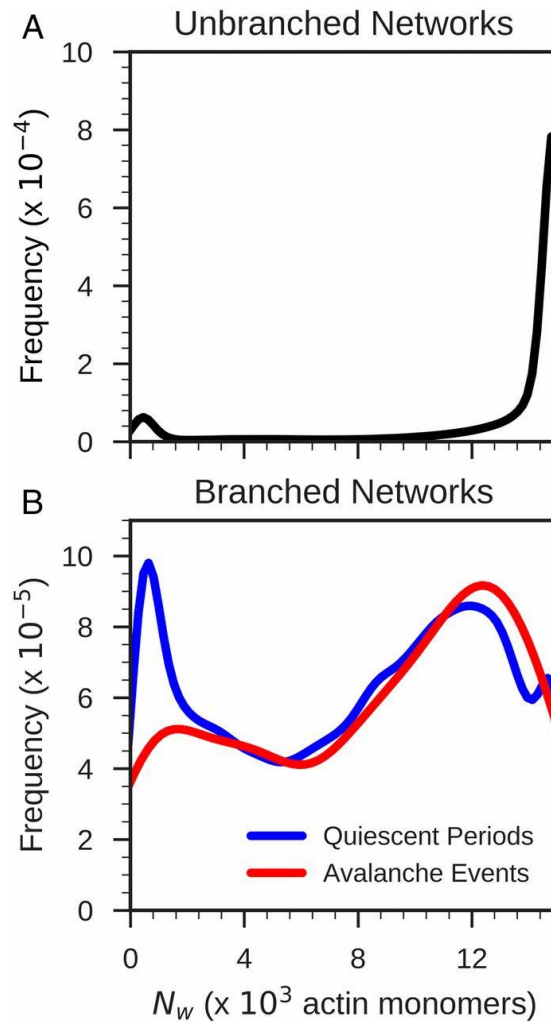


Figure 3.5. Probability densities of the weighted mean cluster size ( $N_w$ ) of the unbranched networks (A) and branched networks (B).

The branched networks were further analyzed in the event of no avalanche (blue curve) or avalanches (red curve). 1,152 simulations were carried out and analyzed both for the unbranched networks and for the branched networks.

### **3.3. Discussion**

#### **3.3.1. Contractility of actomyosin networks depends on their topologies**

Consistent with earlier studies, we find that the contraction of unbranched actomyosin networks depends on the concentrations of motor protein and linkers [26,37,38,155]. The “phase diagram” of macroscopic contractility is re-entrant: systems having an intermediate value of linker concentration contract more robustly than do systems with either a high or a low concentration of linkers. Re-entrance of the transition with changing linker concentration was one of the key findings of several experimental studies [26,37]. These studies show that the macroscopic contractility in actin reconstituted systems requires a threshold concentration of motors but is strongest with an intermediate concentration of linkers. In our simulations, re-entrance arises because when the concentration of linkers is low the network is not mechanically connected while at high linker concentrations the network of an unbranched actomyosin system becomes too rigid for the motors to displace fibers significantly.

Structural differences between branched actomyosin networks and the unbranched systems lead to their having a different phase diagram. The branched networks with low concentrations of motor globally contract more than do unbranched networks with the same low motor concentrations. These contractions apparently are caused by the Arp2/3 branchers inhibiting the polymerization and

depolymerization of the actin fibers, which would otherwise favor the expansion of the network. In our model, an Arp2/3 brancher initially deposits at the branched junction and, once there, nucleates a daughter filament. Given that the depolymerization rate at the minus end of an actin filament is faster than the dissociation rate for branchers, the Arp2/3 branchers effectively inhibit actin filament polymerization and depolymerization.

The relaxation times of the branched actomyosin networks are longer than those of the unbranched systems, by as much as a factor of four. The slowing due to the addition of branchers can be explained by two effects: 1) motors are not able to walk past branchers and 2) on average, the actin filaments of the branched networks are considerably shorter than the filaments of the unbranched networks. The persistence of motor walking in the unbranched networks, which contain longer filaments, drives fast relaxation. In the branched networks, on the other hand, the motors constantly unbind and rebind to short actin filaments that are studded with branches, thereby slowing the relaxation.

### **3.3.2. The non-equilibrium dynamics of the branched actomyosin networks displays the behavior of jammed assemblies**

The trajectories for the branched actomyosin networks show a large number of sudden variations of individual filament displacements. These events are not seen in the trajectories for the unbranched systems at the same concentrations of motors and linkers. Branched networks also have a larger variation in actin cluster size compared with what is seen for their unbranched counterparts. In branched

networks, sudden releases of strain energy occur that lead to local contraction and filament sliding. We call these rare events avalanches. Avalanches in general are triggered by motor driven movements which then in turn cause bound motors to disengage, bound linkers to rupture, and bent filaments to recoil.

It has been suggested that such avalanches occur during the remodeling of the cytoskeleton *in vivo* [117,173]. In these experiments, beads attached to the actomyosin cortex in cells undergo large, step-like displacements over several micrometers. Taken together with the present simulations, these findings suggest that branched actomyosin networks form marginally stable states reminiscent of jammed granular systems or colloidal glasses [134,170]. The mechanical properties of jammed systems depend on their past history, a signature that the system remains far from equilibrium [170]. Our findings suggest that the branched networks are more likely to become marginally stable and jammed than the unbranched networks. This feature seems to arise because branched networks have a wider distribution of F-actin cluster sizes than their unbranched counterparts have. In the branched assemblies, motors can change the conformation of the networks in a convulsive avalanche-like manner once there has been a sufficient buildup of high tension locally in the network [174].

Actin-binding proteins such as the Arp2/3 complex alter the non-equilibrium and history-dependent properties of actomyosin by changing the morphology of the networks. A full theoretical treatment of cytoskeletal dynamics must deal both with their assembly and their glassy preparation dependence.

## **3.4. Methods**

### **3.4.1. Coarse-grained mechanochemical model of actomyosin systems (MEDYAN)**

We have used a computational implementation of a coarse-grained mechanochemical model of actomyosin systems (Mechanochemical Dynamics of Active Networks or MEDYAN) [67] developed by Papoian and his group. MEDYAN models the stochastic nature of reactions and the diffusion of chemical entities (which we will refer to as “the chemical model”) and provides simultaneously a coarse-grained physical representation of filaments, branchers, motors, and linkers (which we will refer to as “the mechanical model”). The coarse-grained actomyosin system is initialized with F-actin filaments that are made up of segments consisting of 10 G-actin monomers each, non-muscle myosin IIA heavy chain (NMIIA) motor proteins (i.e. motors),  $\alpha$ -actinin cross-linking proteins (i.e. linkers), and actin-related protein 2/3 (Arp2/3) branching protein complexes (i.e. branchers). The simulations of branched assemblies of filaments include the Arp2/3 protein complexes while the unbranched simulations of filaments do not. Figure 3.6 shows a typical snapshot of a simulation of unbranched filaments (Figure 3.6A) and branched filaments (Figure 3.6B). The monomers (G-actin), the filaments (F-actin), the motors (NMIIA), the linkers ( $\alpha$ -actinin) and the branchers (Arp2/3) were initially distributed randomly within a cubic container.



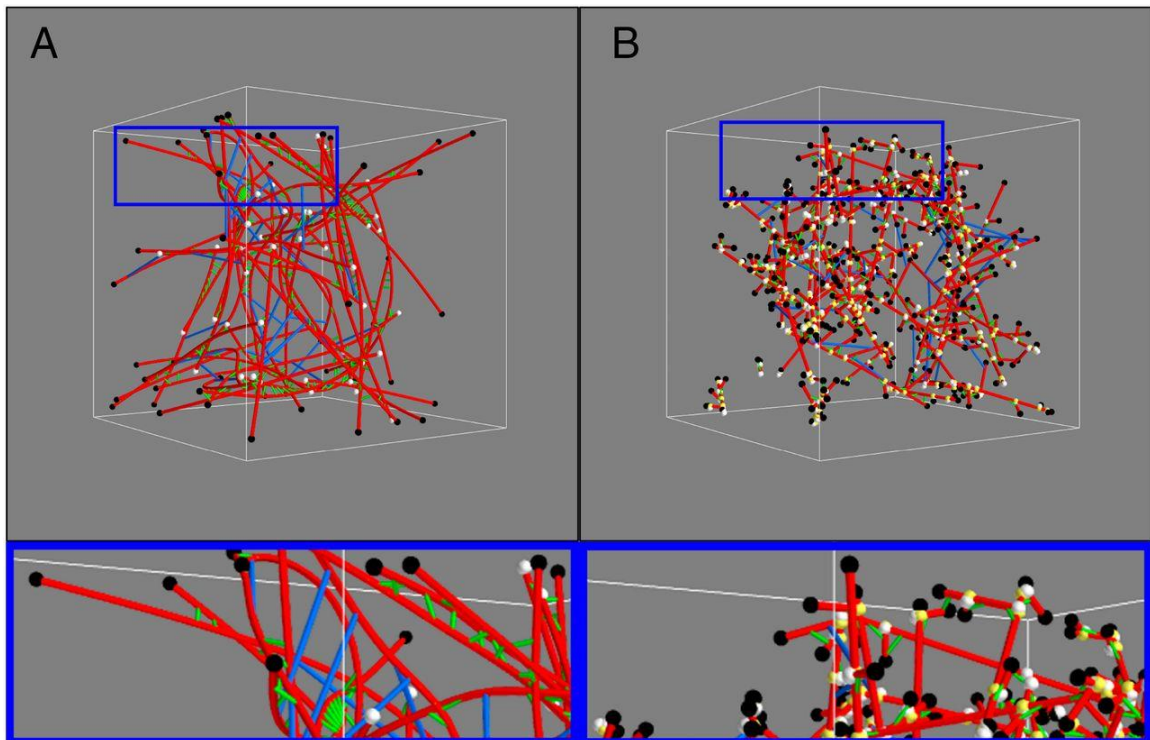


Figure 3.6. Typical snapshots of MEDYAN simulations without Arp2/3 protein complexes (A) or with Arp2/3 protein complexes (B).

A red cylinder represents an F-actin filament. A black bead represents a positive (or a barbed) end of an F-actin filament. A white bead represents a negative (or a pointed end) of an F-actin filament. A green cylinder represents an  $\alpha$ -actinin linker protein. A blue cylinder represents an ensemble of NMIIA motor proteins. A yellow bead represents Arp2/3 protein complexes. The angle between the mother and its daughter filaments through a brancher is  $70^\circ$  as reported experimentally [116,159]s.

We have included the polymerization reactions of F-actin at both the plus and the minus ends of the fibers, the depolymerization reactions of F-actin from both the plus and the minus ends, the binding and the unbinding reactions of  $\alpha$ -actinin linkers, the binding and the unbinding reactions of NMIIA motors, the walking reactions of NMIIA, the branching reaction of F-actin, and the destruction

reaction of short F-actin no longer than one segment in our model. We did not include the unbinding reactions of branchers because of the known high stability of Arp2/3-actin protein complexes [116,159]. We also did not include linker cross-linking between the first binding sites of mother and its daughter filaments due to geometric constraints. The motors, the linkers, and the branchers occupied same binding sites on the filaments.

All simulations were confined to a  $1 \mu\text{m} \times 1 \mu\text{m} \times 1 \mu\text{m}$  box with  $25 \mu\text{M}$  of actin in total. The concentration of actin,  $25 \mu\text{M}$ , was specifically chosen to replicate the *in vitro* experiments from the Weitz group [37]. We considered nine different concentration ratios of motor proteins to actin monomers ( $x_{m:a}$ ) and eight concentration ratios of linker molecules to actins ( $x_{\alpha:a}$ ) in our simulations. We set  $x_{m:a}$  to be 0.0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, and 1.0 whereas  $x_{\alpha:a}$  was set to be 0.0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, and 1.0. All branched simulations contained  $0.5 \mu\text{M}$  of Arp2/3 branchers in total to create sufficient branching structures in the simulation. Moreover, we used the filament branching rate of  $0.0001 \text{ s}^{-1}$  and the short filament destruction rate of  $1.0 \text{ s}^{-1}$ . Furthermore, we have also included actin filament bending, stretching, branching, and exclusion volume potentials, motor and linker stretching harmonic potentials, and filament-boundary interaction potential (see SI for more details). The reaction rates and the parameters of our model are listed in Tables S1, S2, and S3.

### 3.4.2. Cluster size distribution

To analyze how each monomer in a filament connects to another, we defined a cluster as a group of monomers in an F-actin that are connected through the plus ends (p), the minus ends (m), or through binding sites (c) as shown in Figure 3.7A. A motor (M) or a linker (L) can connect two actin monomers through their binding sites (c). A brancher (B) connects two F-actin monomers through the binding site (c) of one to the minus end (M) of the other one.

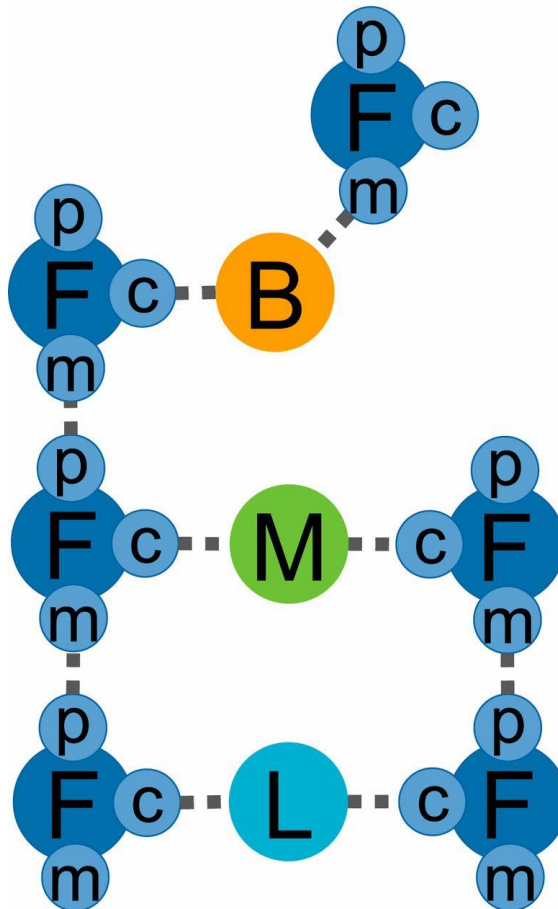


Figure 3.7. Schematic diagram of a cluster.

A cluster is a group of F-actin monomers (F) that are connected through their binding sites (c), the plus ends (p) or the minus ends (m). The motor (M) and the linkers (L) connect the

F-actin monomers through the binding sites. The brancher (B) connects two F-actin monomers by binding on the minus end (m) of one and on the binding site (c) of the other one.

### **3.5. Acknowledgments**

We thank Garyk Papoian and his group for sharing the MEDYAN code with us. The assistance of James Komianos and Aravind Chandrasekaran in implementing Arp2/3 branchers into the code is greatly appreciated. This work was supported by National Science Foundation Grants CHE 1743392, PHY 1522550, and Center for Theoretical Biological Physics Grant PHY 1427654. We acknowledge the use of the uHPC OAC 1531814 and Sabine clusters managed by the core facility for Advanced Computing and Data Science at the University of Houston. Additional support was also provided by the D. R. Bullard-Welch Chair at Rice University, Grant C-0016, and an endowment from the William Wheless III Professorship at the University of Texas Health Science Center at Houston.

# **OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations**

This chapter is a slightly modified version of an article published in the PLoS computational biology journal [175]. OpenAWSEM is available at <https://github.com/npschafer/openawsem> website, and Open3SPN2 is available at <https://github.com/cabb99/open3spn2> website.

### **4.1. Introduction**

In recent decades, experimental methods for studying biological systems have made great strides providing dynamic and structural information across a range of scales. Nevertheless, most experimental probes are still very indirect, with a wide gap between what can be measured directly and what scientists want to understand and visualize. Modern theoretical frameworks for organizing our thinking along with computational simulation codes begin to allow the detailed

mechanisms of biomolecular assemblies to be laid bare. The development of physical simulation models allows mechanistic ideas that are often only inferred indirectly from structural biology to be tested rigorously in a quantitative way rather than remaining attractive but qualitative hypotheses. Biomolecular simulations, in fact, are now beginning to uncover previously unforeseen mechanisms on the molecular level.

When writing down a mathematical description of the forces acting on biomolecules, an important first decision to make is what degree of detail is needed to represent the relevant motions of the biomolecules within their environment. In particular, one must decide which of the atomic degrees of freedom should be kept and which can be averaged over. Retaining all of the atomic degrees of freedom gives rise to the popular all-atom models of biomolecules immersed in a solvent which is also described in atomic detail. While these models are computationally costly to simulate, they can be quite accurate and have recently been used successfully to fold small proteins and even now begin to allow study of the dynamics of larger systems [3,176]. The great amount of detail in the all-atom representation often leads us to forget that all-atom models today still make physical assumptions like the additivity of the intermolecular forces, which may not be fully accurate in all situations. Averaging over the solvent degrees of freedom yields tremendous computational cost savings. The gain in efficiency arises from two factors: first, when we simulate a solvated biomolecule in full atomic detail, most of the atoms belong to the solvent. Eliminating them from detailed consideration then greatly reduces the number of computational operations needed

to follow the dynamics. Second, as parts of the biomolecule move through the solvent, they are constantly buffeted by collisions with the nearby solvent molecules. These collisions dramatically slow down the large-scale motions that usually are of the most interest, yet in the main these frictional effects do not change the structural character of the motions.

Averaging over all the solvent degrees of freedom while retaining a fully atomically detailed representation of the biomolecule thus already yields significant computational advantages. While solvent averaging alone increases computational efficiency, additional computational savings can be had by simplifying the representation of the biomolecule itself. Here again, there are two ways computational time is saved. First, there is a direct savings related to the need to compute a still smaller number of forces. Second, one can choose to intentionally speed up certain internal motions that are otherwise slow in a typical all-atom model by lowering torsional barriers, such as the rotation of backbone Ramachandran dihedral angles. Opting for a coarse-grained representation of a biomolecule, by facilitating sampling, greatly expands the number of biological questions that can be effectively studied.

While it is convenient to average over the solvent and detailed side chain degrees of freedom, the thermodynamic effects of the solvent and the side chains are subtle—considerably more subtle than the buried surface area model. In proteins, it is well known that bulk aqueous solvent gives rise to an effective hydrophobic attraction between non-polar residues [177]. This effect motivated the

buried surface area approximation. It is less widely known that specifically bound water molecules also mediate interactions between pairs of polar residues; these give rise to an effective hydrophilic interaction [178,179]. These water-mediated interactions are quite important in protein complexes. One efficient way of handling such phenomena is to alias such interactions back onto the protein degrees of freedom. Doing this leads to strongly nonadditive forces. It is commonly believed that averaging over any of the degrees of freedom lowers the reliability of a model. For biomolecules, however, the all-atom force fields have themselves generally been parameterized by experimental data just as the coarse-grained models are. The greater freedom of formulating coarse grained models however has long encouraged the use of machine learning strategies to determine these parameters. Such machine learning increases the accuracy of the description [180]. The resulting sophisticated coarse-grained models have proved surprisingly effective in describing biomolecular dynamics both in folding and function, even in a quantitative sense [181].

## **4.2. Design and implementation**

The coarse-grained protein folding force field known as the Associative memory, water-mediated, structure and energy model (AWSEM) is the latest iteration of a series of coarse-grained models that have been primarily developed in the Wolynes and Papoian groups over the last several decades [98]. AWSEM employs a detailed backbone representation along with a single interaction site for each side chain. The AWSEM force field includes an implicit solvent model with a



hydrophobic burial term along with explicit water-mediated nonadditive interactions between the residues. AWSEM-MD is an implementation of the AWSEM model in the LAMMPS molecular dynamics package [182]. AWSEM-MD has been successful in predicting the structures of globular  $\alpha$ -helical proteins [98], both designed and natural  $\alpha/\beta$  proteins [101], and polytopic  $\alpha$ -helical membrane proteins [183]. AWSEM-MD has also been used to study protein association [184] and aggregation [185]. Recently, AWSEM-MD has been used to predict the folds of large proteins by incorporating co-evolutionary information [186] and 3D template information [187]. It has also performed quite well in the CASP13 competition [102].

Nucleic acids are important partners with proteins in biology and it is desirable to study their dynamics with compatible computational tools. 3SPN.2 is a Coarse-Grained DNA model developed by the de Pablo group that models the DNA molecule using 3-sites-per-nucleotide: a particle for the phosphate group, a particle for the sugar and a particle for the nucleobase [85]. 3SPN.2 provides a flexible representation for the DNA backbone and employs a detailed representation of the base pairing interaction and DNA electrostatics. 3SPN.2C also describes the DNA sequence dependent curvature [86]. 3SPN.2C has already been used in combination with AWSEM to study protein-DNA complexes, such as the nucleosome [94] and NF- $\kappa$ B DNA complexes [96].

As the force fields that are used to model protein and protein-DNA systems become more complicated, and as the systems being studied become larger, the

software used to model these systems must also evolve. The challenges are clear: for example, in a recent study of chromosome organization proteins [95], AWSEM combined with co-evolutionary information was used to study a protein complex having a total of 3964 residues. For these large systems, even relatively short simulation runs of 100 ns laboratory time took up to 24 hours to obtain using LAMMPS code. In the present paper, we will show how the OpenMM framework can be used to speed up such simulations using GPUs and how OpenMM framework allows one to introduce novel interactions in the simulation force field models with relative ease.

The LAMMPS simulation package employs a parallelization scheme that is based on spatial decomposition, with each CPU handling a separate contiguous region of space. Information about the forces that act across the boundaries of these domains is passed between the processors at each timestep. This parallelization scheme is relatively simple to implement due to its nearly universal structure with respect to different forcefields. This approach to parallelization scales very well for simulations of bulk liquids and solids, where the system has a nearly uniform density. For simulations of biomolecules with an implicit solvent forcefield, like AWSEM and 3SPN2, however, spatial decomposition can be inefficient because the systems have highly heterogeneous local densities. Processors that compute the interactions inside of the mostly empty boxes will ordinarily then be idle while waiting for the processors that compute the interactions inside of those boxes that are full of atoms. A spatial-decomposition scheme that dynamically adjusts the sizes of the CPU-domains can only partially compensate for this effect. For implicit

solvent models, the force-based parallelization scheme employed by OpenMM turns out to be much more efficient, especially when implemented on GPUs [188].

OpenMM was developed with the intention of being compatible with multiple hardware platforms including GPUs. It provides a high level application programming interface (API) that removes the burden of writing platform specific codes. Traditionally, computational scientists have designed forcefields for single CPUs and then only later would spend time modifying their codes to support simulations on multiple CPUs and even more time on adding GPU support. With OpenMM, one only needs to write down the equations describing the forcefields once, and the software automatically compiles optimized code that can be run on all platforms including a single CPU, multiple CPUs, and GPUs (with both CUDA and OpenCL support).

OpenMM provides various flexible custom force templates to ease the implementation of forcefields with new functional forms. To implement OpenAWSEM and Open3SPN2, we used the custom force template that best fits each term in the Hamiltonians. For example, the “CustomNonbondedForce” is the best choice for the excluded volume term, which acts between every pair of atoms, while the “CustomBondForce” supports a very wide range of functional forms and is appropriate for terms that involve only a small subset of the system’s atoms. Another interesting situation that OpenMM flexibly encodes is AWSEM’s water-mediated interaction. Since the water-mediated interactions depend on the local density around each interacting atom, the local density around each residue has to be computed first before computing the mediated interactions. This two-stage

feature can be implemented using the “CustomGBForce” template, which was originally intended to support another two stage energy term: the Generalized Born-type potentials.

The custom force templates allow for rapid prototyping of new potential terms. For each new potential, only the energy formula needs to be specified, while its derivatives are automatically computed for the purposes of computing the forces. By automating the derivative calculation, even non-experts can design and implement new force fields readily. In this paper, we will illustrate this capability of the OpenMM framework by introducing two new features into AWSEM. The first new feature is a contact term that depends on the degree of burial of a residue in a biological membrane. This energy can be used to describe proteins that have both cytoplasmic parts that are surrounded by water, and other parts that are buried in a membrane, which are thus surrounded by lipid primarily. The second new nonadditive potential we introduce and explore is a many-body disulfide bond term that prevents the unphysical clustering of Cysteines that can occur when disulfide bonds are modeled using a naïve pair potential that must per force be very strong. This potential allows us to recapitulate the early experiments of Anfinsen on ribonuclease that started the experimental study of protein folding mechanism [7,189].

## 4.3. Results

### 4.3.1. Protein-only simulations benchmark

When AWSEM was first implemented using LAMMPS 8 years ago, dynamic studies of proteins mostly focused on proteins having less than a thousand residues. This limited focus was due both to the computational cost of studying larger system, and partly, to the scarcity of experimentally solved structures of large biological machines. The structures of larger proteins and their complexes are now being obtained at an unprecedented pace, thanks especially to the development of Cryo-EM structure determination methods. One recently solved large protein; gamma secretase has drawn lots of attention due to its role in Alzheimer' disease. Gamma secretase contains 1542 residues [190]. Figure 4.1 shows comparative benchmark results for simulations using OpenAWSEM and using LAMMPS for proteins with various lengths. For a protein with 3724 residues (PDBid: 4qqw), a simulation of 4 million steps corresponding roughly to 20 $\mu$ s in laboratory time took more than 200 hours (8 days) using LAMMPS. The same simulation takes only about 8 hours using OpenAWSEM, thus making millisecond simulations feasible within a few days.

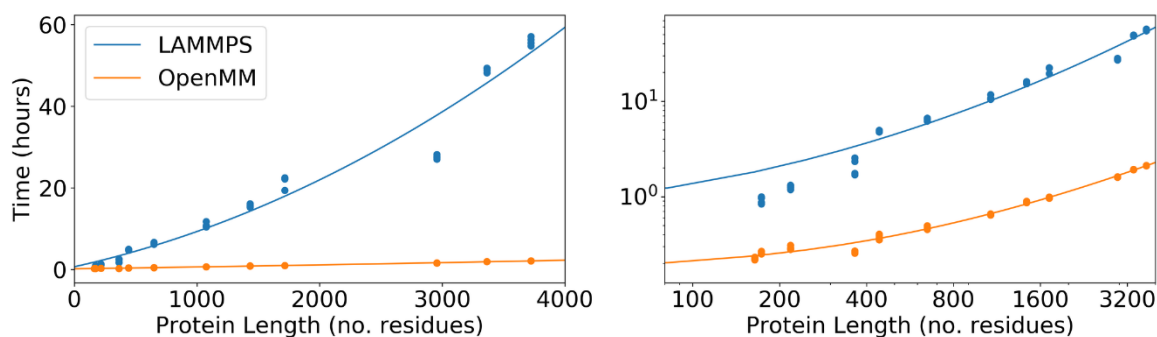


Figure 4.1 Benchmark timing results for AWSEM simulations with the LAMMPS and the OpenMM implementations on a linear scale (left) and on a log scale (right).

The x-axis is the number of residues in the proteins that are being simulated. The y-axis shows the number of computer hours needed to run a 1 million-step simulation. Each protein was simulated 5 times using each implementation. The lines are quadratic fits. The simulation protein set was chosen to have a wide range of protein sequence lengths ranging from 164 residues to 3724 residues.

#### **4.3.2. DNA-only simulations benchmark**

To test the scaling of the runtime of Open3SPN2 for nucleic acids, we ran several random DNA sequences of different lengths using the 3SPN2.C forcefield. The DNA strands were simulated using LAMMPS and using OpenMM for 1 hour and, from these test runs, we estimated the time needed to run 1 million steps. As shown in Figure 4.2, the OpenMM implementation of 3SPN2.C reduces the simulation time of long DNA strands ranging in size from 250 bp up to 1.5kb DNA strands. For short sequences the GPU is underutilized, and the greater overhead associated with using the GPU results in longer overall simulation times. For the 1.5 kb case, we found a fourfold improvement in simulation speed. For longer DNA strands, the speedup will be greater due to better scaling. This improvement in the simulation speed allows the study of DNA dynamics on much longer timescales even for more complex systems such as DNA origamis or small sections of chromosomes.

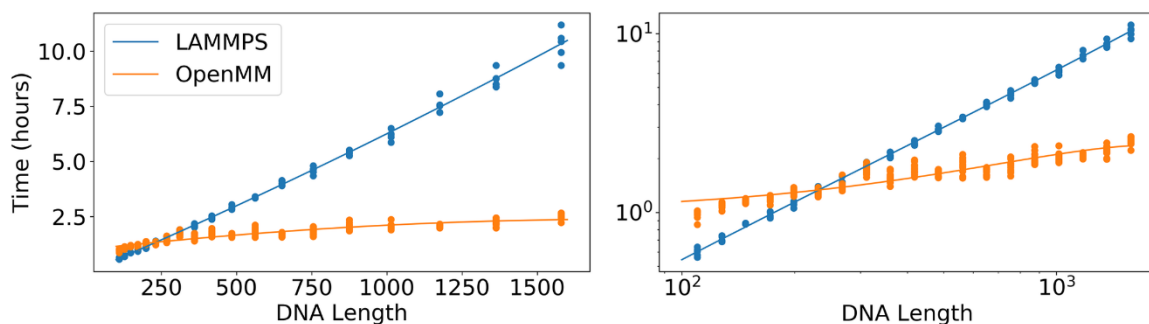


Figure 4.2. Benchmark timing results for 3SPN2 simulations with the LAMMPS implementation of 3SPN2 and the OpenMM implementation of 3SPN2 on a linear scale (left) and on a log scale (right).

The x-axis is the number of nucleotides in the DNA that is being simulated. The y-axis shows the number of computer hours that are needed to run a 1 million-timestep simulation. Each DNA length was simulated 5 times using each implementation. The lines are quadratic fits. The DNA lengths range from 110 nucleotides to 1580 nucleotides.

### 4.3.3. Protein-DNA simulations benchmark

To assess the speedup of DNA-protein simulations we selected several protein-DNA complexes that have a diverse range of lengths for both the protein and the DNA sequences. We included in this test set only structures from the PDB that contained a single protein chain and a single DNA chain. We simulated each complex 5 times for 1 hour using each implementation and estimated how much time would be required to run 1 million steps. Figure 4.3 shows an improvement of the simulation speed of protein-DNA complexes by 1 to 2 orders of magnitude. The largest structure that we simulated was RecA, a protein with 2050 amino acids, in complex with a 18 nucleotides ssDNA (PDBid: 3cmu). In this case, we obtained a 300-fold speedup.

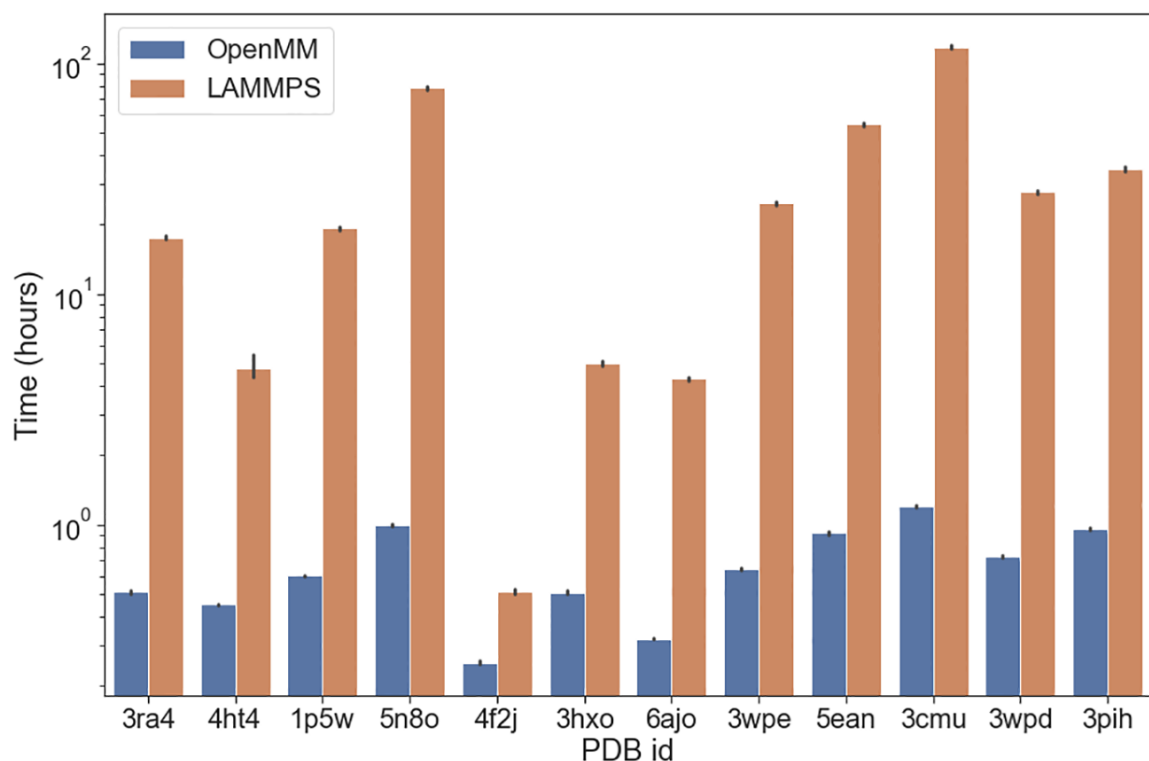


Figure 4.3. Benchmark results for AWSEM-3SPN2 simulations of protein-DNA complexes using the LAMMPS and the OpenMM implementations of both forcefields on a linear scale (left) and on a log scale (right).

The x-axis shows the PDB ID. The y-axis shows the computer hours needed to simulate for 1 million steps. Each complex was simulated 5 times using each implementation. The protein length ranges from 52 nucleotides to 2050 amino acids, while the DNA length ranges from 2 to 40 nucleotides.

#### 4.3.4. Protein-DNA interface prediction example

As an example of simulating protein-DNA interactions, we characterized the capability of the AWSEM-3SPN2 Hamiltonian to predict the correct protein-DNA binding interface of the sporulation specific transcription factor Ndt80 (PDBid: 1mn). At a constant temperature of 300K, the protein and DNA in the crystal structure were first pulled apart and run for 2.5 million steps; following this, a weak,



non-specific force was used to pull them back together while being run for another 2.5 million steps. Following this, the pulling force was released, and the complex was simulated for another 5 million steps to let it relax. To reduce the effects of binding to only a short length of DNA, we extended the crystallized DNA by adding DNA made with 100 A/T base pairs to both ends of the double stranded DNA using the 3DNA package [191].

The OpenAWSEM-Open3SPN2 cross-interaction is given by electrostatic interactions between the DNA phosphates and charged residues of the protein, as well as excluded volume terms. The current implementation lacks specific interactions that depend on the nucleotide type and amino acid type. Therefore, would it not through indirect DNA conformation-mediated effects, the protein would not be expected to prefer binding to any particular stretch of nucleotides on the DNA. The part of the protein surface that binds to the DNA and the orientation of the bound protein with respect to the DNA, however, is somewhat specific. To evaluate the quality of the DNA-protein interface, while focusing on finding the native binding pocket of the protein, we quantified the quality of the docking in terms of the number of contacts that the protein makes with any location along the DNA. A residue in the protein is said to make such a “symmetrized” contact with DNA when the C $\beta$  atom in the residue is closer than 1.8 nm to a Phosphate of DNA in the crystal structure and where also, in the predicted structure, this C $\beta$  atom is found within 1.8nm of a Phosphate of the DNA. For PDB ID 1mnn, there are 135 such native contacts. The interface energy is defined as the sum of protein-DNA excluded volume energy and the electrostatic interaction energy between the

protein and the DNA. As can be seen in Figure 4.4, there is a strong correlation between the protein-DNA interface energy and the quality of the protein-DNA interface, and the orientation of bound protein relative to the DNA matches that found by experiment.

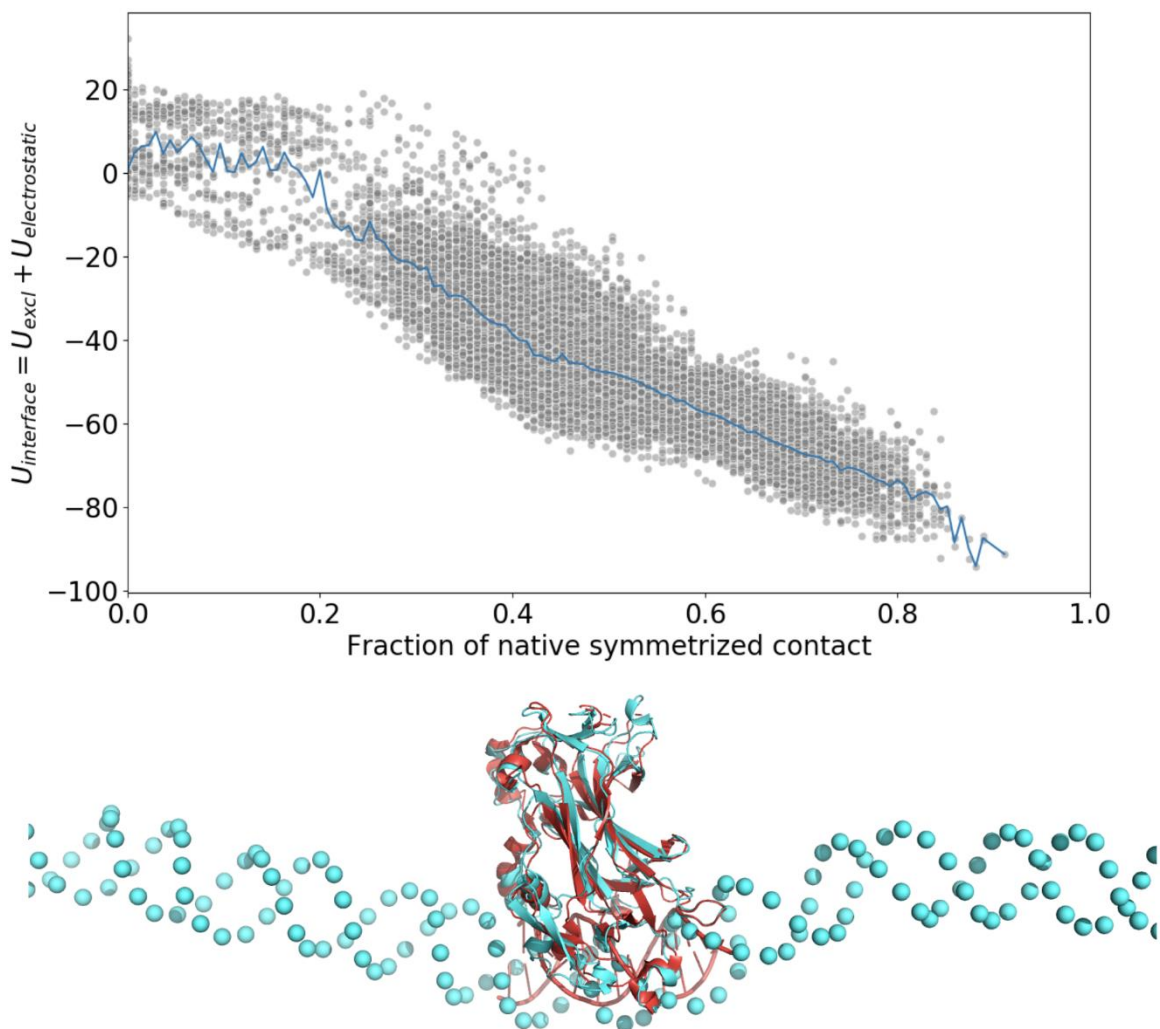


Figure 4.4 Scatter plot of the interaction energy between the DNA and the protein versus the fraction of the symmetrized native contacts formed at each time frame during the last 7.5 million steps of simulations from 10 runs.

The average energy as a function of the number of symmetrized native contacts is indicated with blue line. A simulation snapshot showing the overlap of the crystal structure (colored in red) and the predicted structure (colored in cyan) that has the lowest interface energy.

There is a high correlation between the protein-DNA interface energy and the number of symmetrized contacts, indicating that the binding process is funneled to the correct interface. The overlap figure was created by aligning only the protein parts of the crystal structure and the predicted structure. We see that the DNA in both structures turns out to be aligned quite well, showing good structural agreement between the lowest energy simulated structure and the experimental structure.

#### **4.3.5. Potentials that depend on locations of residues relative to a membrane**

The water-mediated potential introduced by Papoian et al. [98] acknowledged that residues interact not only when they are directly in contact but also when they perturb the surrounding water, which in turn changes the energetics of more distant residues. The parameters for this potential were optimized using an energy landscape theory inspired machine learning algorithm [181,192–195]. Energy landscape theory provides a recipe whereby a transferable energy function can be learned by searching for the most funnel like landscape in a class of energy models. The funnel-like character of the landscape is measured by a Z score,  $Z = (E_{native} - E_{mg})/\sigma(E_{mg})$ . This quantity is then maximized while maintaining  $E_{mg}$  constant.  $E_{mg}$  is the average energy of a set of misfolded decoy structures. Using this strategy leads to an optimal set of parameters to discriminate between properly folded and misfolded structures. In the simplest model these parameters are the strengths of the interactions for different types of residue pairs at various distances and how these interactions vary with the local density of protein and by contrast with the local density of solvent water. The AWSEM potential has proved to be very successful in structure prediction and has allowed exploration of many aspects of

protein functional motions [184,185]. The water-mediated potential was originally designed for globular proteins, but the same optimization scheme was used also to find a transferable energy function that would fold membrane proteins, [183] in their membrane environment; the residue pair interactions then are mediated by lipids instead of by water. Following the same procedures as used for the globular proteins, the parameters for proteins that are found entirely inside the membrane were optimized to discriminate proper folds. Many proteins, however, have some of their parts inside the membrane while other parts of the protein remain outside in the cytoplasm. To study such systems we need a potential that can dynamically switch from being water-mediated to lipid mediated based on the position of the residues relative to the bilayer. Figure 4.5 shows the schematic of this potential.

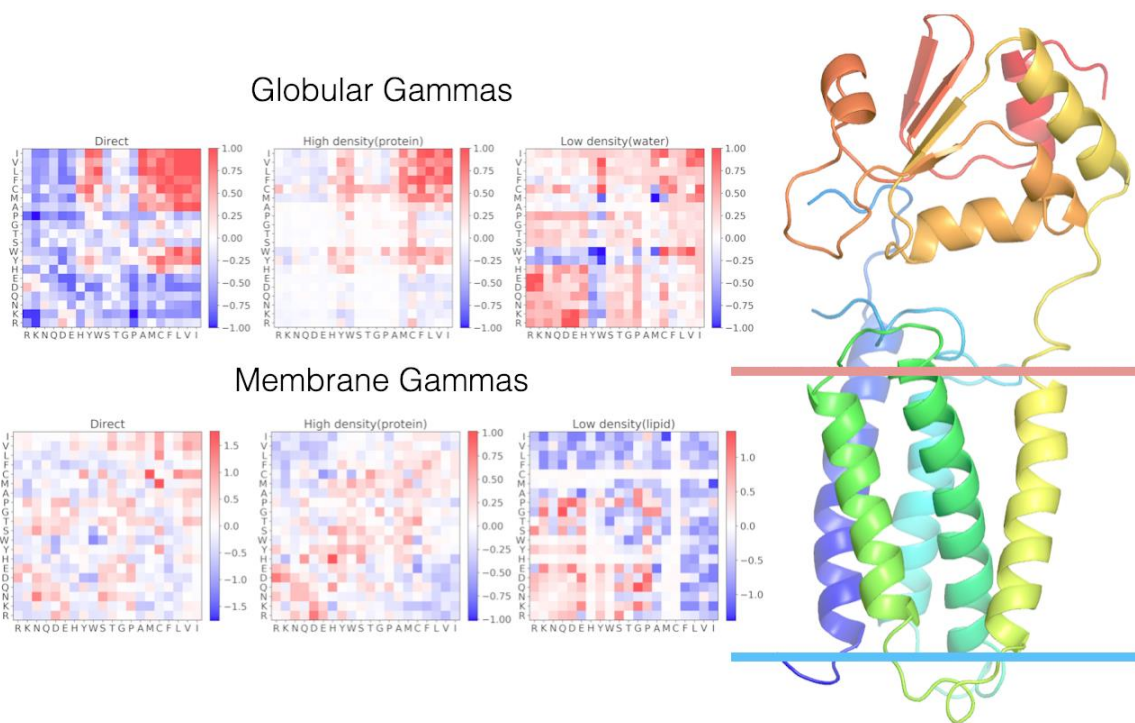


Figure 4.5. A schematic figure for the Z-dependent contact potential.

The residues outside of the membrane, where the membrane boundary is indicated by the two colored lines, interact using the globular parameters. The residues inside the membrane interact using the membrane-optimized parameters. If one residue is inside, while another one is outside, the pair interacts as if they both were in water. In the heat maps on the left side of the figure, red color indicates a favorable interaction between the pair of residues indicated on the horizontal and vertical axes, whereas blue color indicates an unfavorable interaction. Separate heat maps are shown for the direct, low-density, and high-density interaction matrices in the water (globular) and membrane environments.

Here, we introduce a z-dependent contact term that allows such dynamic switching. The interactions smoothly transition between the membrane mediated interactions and water-mediated interactions depending on the location of the interacting residues with respect to the membrane as measured by a height z. We define the new contact potential term  $V_{contact}$  through the following equations:

$$V_{contact} = \sum_{j-i>9} V_{contact}(i,j) \quad \text{Equation 4.1}$$

$$V_{contact}(i,j) = (1 - \alpha_i\alpha_j)V_{water}(i,j) + k_{relative}\alpha_i\alpha_jV_{memb}(i,j) \quad \text{Equation 4.2}$$

$$\alpha_i = \frac{1}{2}(\tanh(\eta(b + z_i)) + \tanh(\eta(b - z_i))) \quad \text{Equation 4.3}$$

where  $b = 1.5nm$ ,  $\eta = 10nm^{-1}$ .  $V_{water}(i,j)$  and  $V_{memb}(i,j)$  are the contact terms for water mediated and membrane mediated interactions as defined in previous paper [98,183].

Since both sets of parameters in the Hamiltonian were previously optimized without acknowledging the presence of the other terms, we also need to introduce a new parameter  $k_{relative}$  that controls the relative strength of the membrane mediated and the water-mediated interactions. A high  $k_{relative}$  favors forming

contacts inside the membrane, while a low  $k_{relative}$  favors forming contacts in water. To determine the optimal value of  $k_{relative}$ , we again employ the energy landscape optimization learning scheme. The decoys for implementing this scheme were generated by shifting the proteins vertically and rotating them. One then optimizes the  $k_{relative}$  while keeping the previously determined parameters fixed. This machine learning scheme was employed using a test set obtained by downloading the complete Alpha-helical polytopic database, a total of 1561 proteins, from the Orientations of Proteins in Membranes (OPM) database. [196]. The advantage of the OPM database over the traditional RCSB protein data bank is that it also spatially aligns membrane proteins relative to the membrane. The training proteins must have significant parts both inside and outside the membrane. Therefore, for each protein, we computed the fraction of the residues that are found inside the membrane

$$\chi = \frac{1}{L} \sum_{i=1}^L (abs(z_i) < 15\text{\AA}) \quad \text{Equation 4.4}$$

where  $z_i$  is the z coordinate of CA of residue  $i$ ,  $L$  is the protein length. For training we only kept those proteins with  $\chi$  between 0.2 and 0.8. We also removed those proteins that have more than 2000 residues in order to speed up the optimization. This yielded a set of 1116 training proteins. For each protein, we then generated 240 decoys. These were generated first by rotating them along the x axis with 12 different orientation at: 0, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165 degrees, and then shifting the structure vertically by 20 different displacements: -40, -36, -32, -28, -24, -20, -16, -12, -8, -4, 0, 4, 8, 12, 16, 20, 24, 28, 32, 36 angstroms

along the z-axis. To carry out this optimization, the total energies are evaluated using the following equations:

$$E = k_{wat}\phi_{wat} + k_{memb}\phi_{memb} + k_{memb_{burial}}\phi_{memb_{burial}} \quad \text{Equation 4.5}$$

$$\phi_{wat} = \sum_{j-i>9} (1 - \alpha_i\alpha_j)V_{water}(i, j) \quad \text{Equation 4.6}$$

$$\phi_{memb} = \sum_{j-i>9} \alpha_i\alpha_j V_{memb}(i, j) \quad \text{Equation 4.7}$$

$$\phi_{memb_{burial}} = \sum_i A(\sigma_i)\Theta(z_i, z_m = 15\text{\AA}) \quad \text{Equation 4.8}$$

$$\Theta(z_i, z_m) = \frac{1}{2}(\tanh(k_m(z_m + z_i)) + \tanh(k_m(b - z_i))) \quad \text{Equation 4.9}$$

In these expression the values of  $A(\sigma_i)$  are the amino acid hydrophobicities on the octanol scale of Wimley and White [197–200]. We include  $\phi_{memb_{burial}}$  here because the membrane burial term also depends on the position of protein with respect to the membrane [201]. In the machine learning algorithm thus we want to find the values of  $k_{wat}$ ,  $k_{memb}$ , and  $k_{memb_{burial}}$  that maximize the Z score for the correct positioning and orientations of the proteins with the membrane. Since some decoys are more similar to the native positioning than are others, we reweighted the decoys when computing the decoy averages in  $\langle\phi\rangle_{mg}$ .

$$\langle\phi\rangle_{mg} = \frac{1}{\sum_{d=1}^N (1 - \theta_d)} \sum_{d=1}^N (1 - \theta_d)\phi_d \quad \text{Equation 4.10}$$

where N is the number of decoys. For each decoy, the fraction of residues that have the same pattern of burial as the native structure is defined to be  $\theta_d$ . Two

residues are said to have the same burial assignment when either they are both inside the membrane or they are both in the cytoplasm.

$$\theta_d = \frac{1}{L} \sum_{i=1}^L \delta_i \quad \text{Equation 4.11}$$

$$\delta_i = \begin{cases} 1 & \text{if } abs(z_i^0) < 15\text{\AA} = abs(z_i) < 15\text{\AA} \\ 0 & \text{if } abs(z_i^0) < 15\text{\AA} \neq abs(z_i) < 15\text{\AA} \end{cases} \quad \text{Equation 4.12}$$

, where  $(z_i)$  is the z coordinate of CA of residue i in the native(decoy) structure. The optimal values of the coefficients that maximize the Z score turn out to be 1, 3.3, 3.3 for  $\phi_{wat}$ ,  $\phi_{memb}$ , and  $\phi_{memb_{burial}}$  respectively.

To demonstrate the effectiveness of the force field obtained in this way, we selected from the database 15 proteins that have both membrane and globular parts. The folding of membrane proteins is sometimes thought to have two stages. [202]The first stage is imagined to be the insertion of the transmembrane helices into the membrane. In vivo this process is sometimes helped by the translocon [203]. The second stage of membrane folding is then the rearrangement of the now buried helices inside the membrane. To imitate the first stage, we used Pureseq™ [204]first to provide an initial idea of the topology with respect to the membrane. Based on the Pureseq™ prediction result, we wrote a script to assign each residue to three different regions: cytoplasmic, membrane or extracellular. Each residue is then pulled into its preliminarily predicted region according to the resulting initial assignment using a force field that only contains the backbone terms. Then, a force



is applied to the two ends of the protein while applying a strong membrane term, so that the helices become well separated but still live within the membrane. Finally, the residue type dependent membrane potential is introduced along with the contact terms and an annealing protocol of 8 million steps is followed with the temperature decreasing from 800 to 200. The results for the structure prediction runs using the z-dependent contact term are compared with the results using the original contact potential in the Figure 4.6.

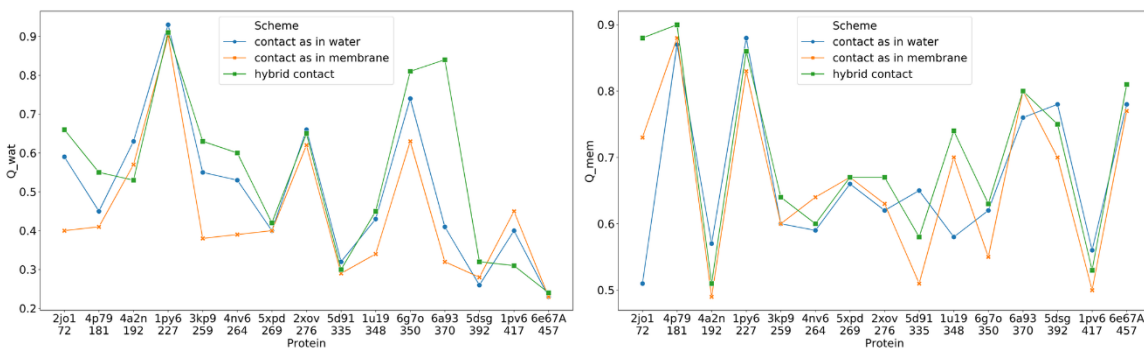


Figure 4.6. Structure prediction results using the three contact potential schemes evaluated using  $Q_{water}$  (left) and  $Q_{mem}$  (right).

$Q_{water}$  measures the structural similarity to the native structure using only the residues that are outside of the membrane, whereas  $Q_{mem}$  measures the structural similarity of the structures for those residues found inside the membrane. The closer the similarity score is to 1.0, the more native like is the prediction. The hybrid potential in general performs better than either the pure globular protein model or the pure membrane model.

Figure 4.7 shows the aligned structures of the native structures and the predicted structures using the new membrane burial depth dependent contact potential.

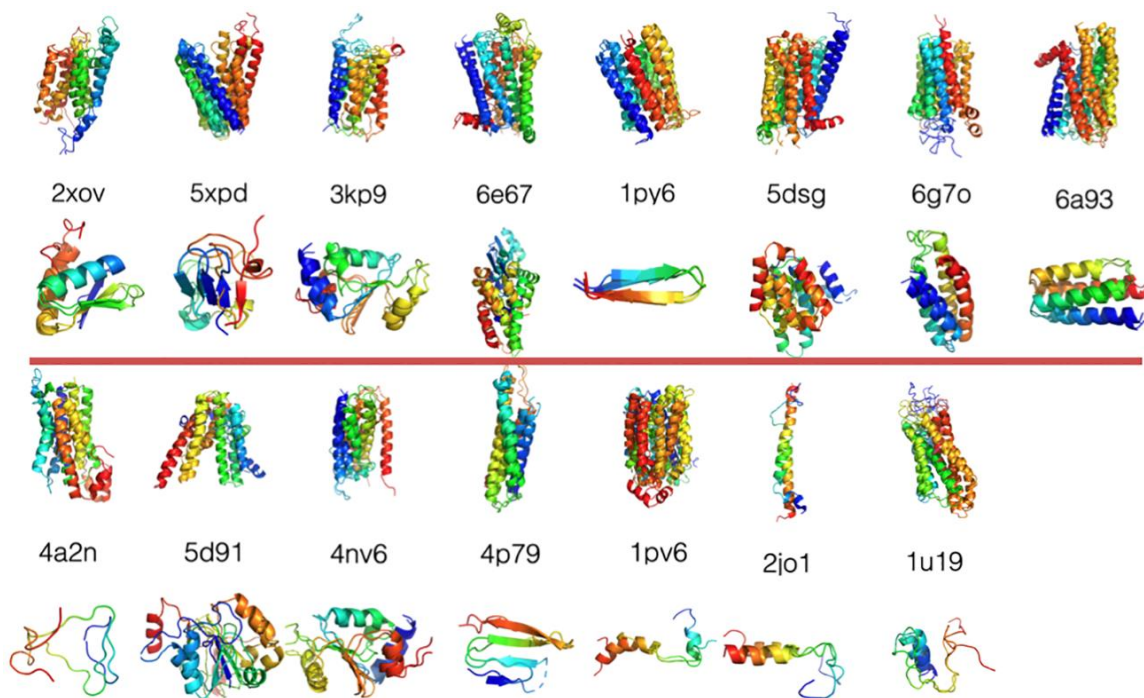


Figure 4.7. Overlay of the native structures and the best Qwater and Qmem structures using the membrane burial depth dependent contact potential.

For each protein, the upper figure shows the part of the protein that is found buried in the membrane and the lower part of the figure shows the globular domain.

The AWSEM annealing yields an improved assignment of the location of the helices relative to the purely sequence based method Pureseq<sup>TM</sup> that was used for initial structures. In Figure 4.8, we see that for 10 out of 15 proteins tested, the fraction of correctly assigned location is increased after the folding. In this test set, 3kp9, 5xpd, 1u19 now have more than 10 additional residues that take on their correct native location assignments compared to what is used initially based on the Pureseq<sup>TM</sup> results.

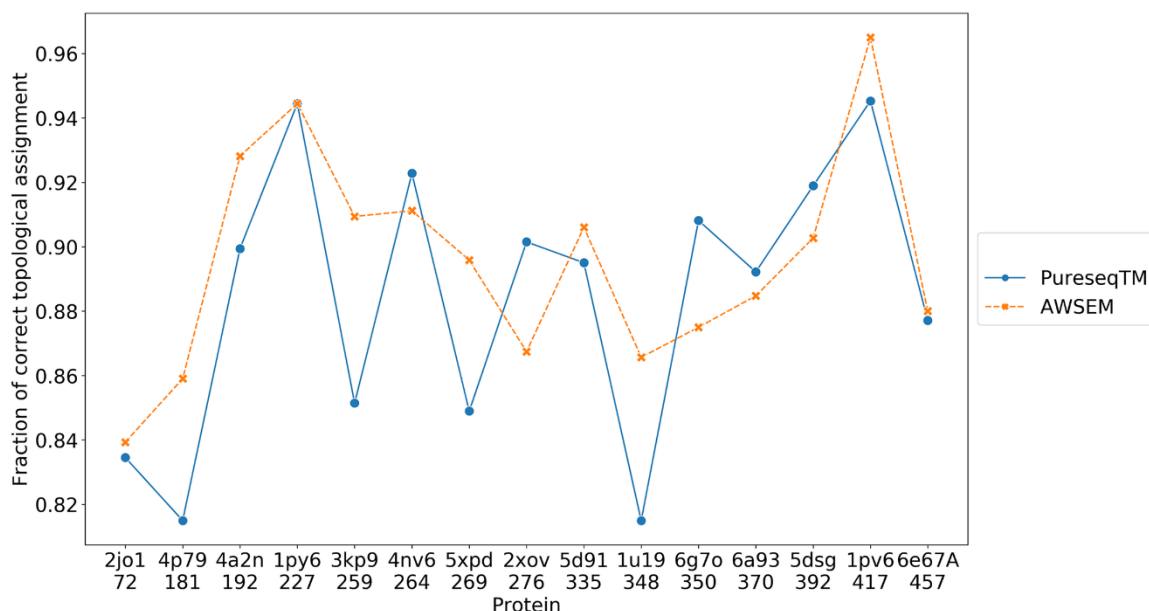


Figure 4.8. The fraction of correct location assignments of the residues relative to the membrane using a purely sequence-based method (PureseqTM) and that yielded by running OpenAWSEM simulations (AWSEM).

#### 4.3.6. Describing many-body saturating disulfide bonds

The disulfide bond forms a very strong interaction between two Cysteines. These bonds restrain the dynamics of the protein and often control protein stability and function. Very often the smaller extracellular proteins are dominantly stabilized by a large number of disulfide linkages. If treated as a pair interaction, the strong disulfide bonds tend to condense and cluster. The covalent chemical bond, in contrast to the pair interacting potential, saturates: only one bond can be formed by each Cysteine, not more. The famous protein ribonuclease A was originally studied by Christian Anfinsen. It has four disulfide bonds. Monitoring the formation of these bonds was a key part of Anfinsen's exploration [7]. Two of the four bonds have been

shown to be important for conformational stability and the other two are needed for catalytic activity [205]. Because covalent chemical bonds saturate, a simple pairwise potential cannot model accurately Anfinsen's experiment. The saturation effect is critical: when there are only two cystines, they form a single strong disulfide bond, but when a third Cystine comes near to the two Cystines that have already formed a bond, the third Cystine shouldn't feel any strong attracting force. This is a many-body effect. In this study, we tackled this saturation problem by developing a saturable many-body disulfide bond interaction using the openAWSEM framework. In this potential, displayed in Equation 4.13, the saturation is accounted for using a density variable  $\rho_i^{cys}$  that reflects the number of Cystines around residue  $i$  smoothed by a tanh function. The disulfide interaction term is then a pair interaction that is modulated by two  $\rho_i^{cys}$  dependent switching functions,  $\theta_{ij}^{near}$  and  $\theta_{ij}^{small}$ . These two switching functions are defined in Equation 4.16 and Equation 4.17.

$$V_{disulfide} = \sum V_{ij} \quad \text{Equation 4.13}$$

$$V_{ij} = \theta_{ij}^{near} \theta_{ij}^{small} \alpha(r_{ij}) \quad \text{Equation 4.14}$$

$$\alpha(r_{ij}) = \frac{1}{2} \left( \tanh \left( \kappa (r_{ij} - r_c) \right) - 1 \right) \quad \text{Equation 4.15}$$

$$\theta_{ij}^{near} = \frac{1}{2} \left( \tanh \left( \kappa_s (0.2 - |\rho_i^{cys} - \rho_j^{cys}|) \right) + 1 \right) \quad \text{Equation 4.16}$$

$$\theta_{ij}^{small} = \frac{1}{2} \left( \tanh \left( \kappa_s \left( 2.2 - (\rho_i^{cys} + \rho_j^{cys}) \right) \right) + 1 \right) \quad \text{Equation 4.17}$$

$$\rho_i^{cys} = \sum_{|j-i|>1} \frac{1}{2} (1 - \tanh(\kappa(r - r_c))) \quad \text{Equation 4.18}$$

where  $i, j$  label all the Cystine residues, and  $r_{ij}$  is the Cb distance between residue  $i$  and  $j$ .  $\kappa$  is set to 20, so that  $\theta_{ij}^{near}$  is 0 when the difference between  $\rho_i^{cys}$  (the Cystine density around residue  $i$ ) and  $\rho_j^{cys}$  (the Cystine density around residue  $j$ ) is larger than 0.4, and  $\theta_{ij}^{near}$  is 0 when the sum of those two densities is larger than 2.4. The parameters introduced to quantify the rapidity of saturation were calibrated using a database search for disulfide bonds in known crystallographic structures. To determine a reasonable potential well size  $\kappa$  for determining the Cystine density, our survey showed that the Cb-Cb distances between residues that form disulfide bonds fall in the range of 3.6 Å to 4.1 Å. We therefore chose a 0.5 Å interval over which to turn on the interaction by setting  $\kappa = 10\text{Å}^{-1}$  and  $r_c = 4.2\text{Å}$  in Equation 4.18.

To illustrate the efficiency of using the new nonadditive Cystine density dependent disulfide bond term, we simulated the folding of ribonuclease A (1fs3), bovine pancreatic trypsin inhibitor (1bpi), alpha thrombin (1ppb) and several other cystine rich proteins selected from [206]. We tested 3 different strengths for the new potential, ( $k = 0, 2, 5$ ), as well as the model that has the pairwise additive potential, which we call “original”. We can see from Figure 4.9 that as the strength of the saturable disulfide bond term increases, the predictions become closer to the correct structure (as evaluated by the Q value). The saturable disulfide bond term

significantly improves the structure prediction quality for ribonuclease A. This improvement is mainly due to the correct formation of the Cys26-Cys84 bond, which was also shown by experiment to be essential for protein stability [205].

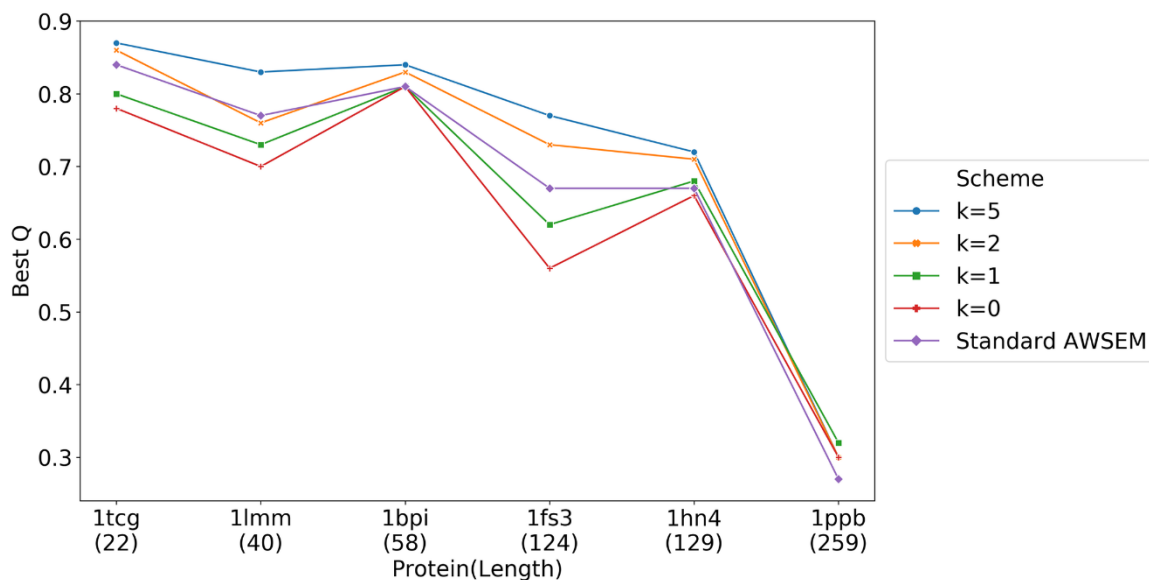


Figure 4.9. Structure prediction results for six disulfide rich proteins using various strengths of the saturable disulfide bond interaction.

We plot the best Q from 20 simulated annealing runs that started from different random velocity seeds for each different value of the disulfide interaction strength. As the strength of the disulfide interactions increases, the best Q increases. 1tcg, 1lmm, 1bpi and 1ppb all have 3 disulfide bond. 1fs3 has 4 disulfide bonds, and 1hn4 has 7 disulfide bonds. The relatively modest best Q for thrombin (1ppb) probably comes from the fact that we have only modeled the main chain of the molecule, but thrombin also has a short chain that has been experimentally shown to be important for function [207].

The new disulfide bond term helps specifically to form the native disulfide bonds, rather than allow the formation of mispaired Cysteines as shown in Figure 4.10. Even though in some cases (1tcg, 1lmm and 1ppb), the prediction quality measured by Q was not significantly affected by using the saturable disulfide interaction, the fraction of correct disulfide bonds was improved in all six proteins we tested.

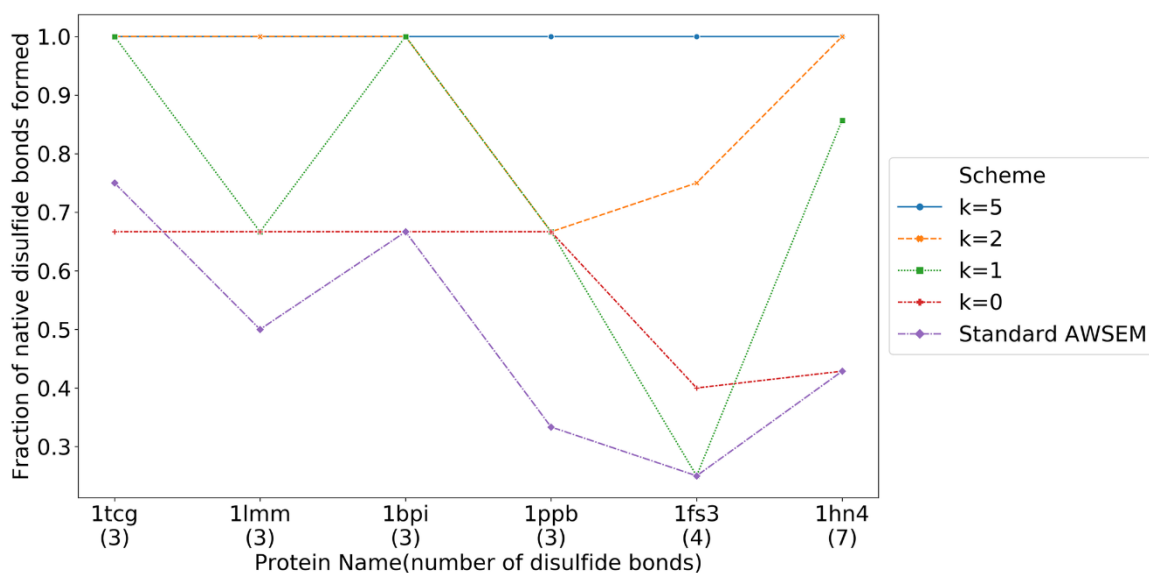


Figure 4.10. The fractions of correct disulfide bonds in the predictions of several disulfide rich proteins.

These fractions are shown for several different strengths of the saturable interaction. At full strength, nearly all the pairs form correctly.

When we follow the annealing trajectories for these disulfide rich proteins, we find that, consistent with the funneled nature of the energy landscape, disulfide bonds do not always form in a specific unique order, and indeed non-native disulfide bonds occasionally form and revert back to being unpaired, finally

achieving a native like structure. Of course, we must bear in mind that in the laboratory this process must involve chemically tuning the oxidation of these bonds. Figure 4.11 shows the sequence of formation of disulfide bonds from each frame in a simulated annealing trajectory of ribonuclease A. As the extended protein starts to fold from high temperature, some non-native disulfide bonds do form, but, in the end, the protein is funneled to form the correct native disulfide bonds.

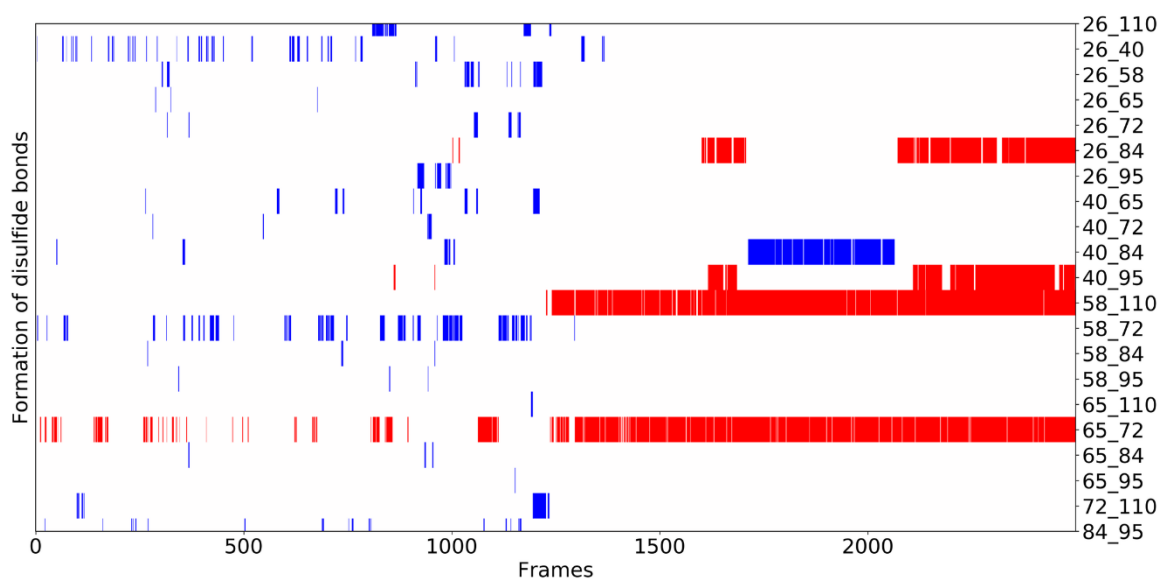


Figure 4.11. The formation of disulfide bonds in a single annealing trajectory with  $k = 5$ . Following the trajectory in time, disulfide pairs are darkened in when they are formed. Red indicates that a native disulfide bond has been formed. Blue indicates that a non-native disulfide bond has formed. The alignment of the best Q structure from this trajectory with the crystal structure is shown in SI. Its Q value is 0.77.

As shown in Figure 4.12, using the standard AWSEM, only one native disulfide bond (residue 58 and residue 110) ends up being formed in most of the 20 trajectories, while the other native pairs(26-84, 40-95, 65-72) are rarely formed. In



comparison, using the new Cystine density dependent disulfide bond potential, all the native pairs are finally formed.

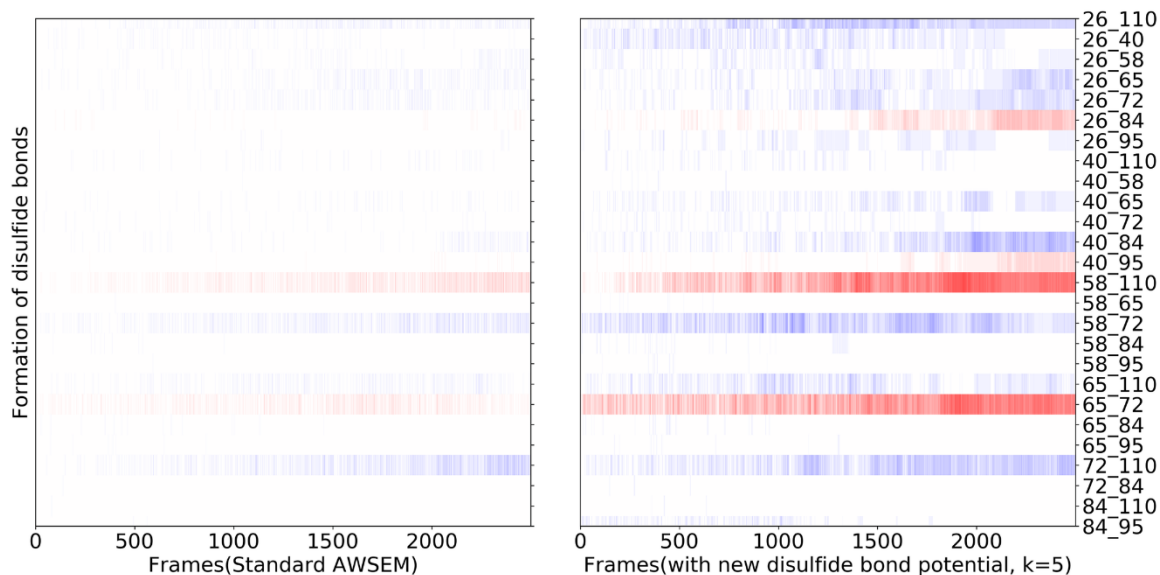


Figure 4.12. The average formation of disulfide bonds as a function of time over the 20 annealing runs, with the patterns from the standard AWSEM shown on the left and patterns from the nonadditive disulfide potential runs with  $k = 5$  shown on the right.

Red indicates that native disulfide bond has formed. Blue indicates the formation of a non-native disulfide bond. The darker the color, the larger fraction of the trajectories that form this disulfide bond during this time frame. We see that, occasionally, even with the full strength saturable interactions, sometimes non-native disulfides persist after the rapid annealings.

#### 4.4. Discussion

We have described a new computational simulation framework for carrying out coarse grained protein-DNA simulations: OpenAWSEM and Open3SPN2. In this new framework, simulations using GPUs can achieve speedups of a factor of thirty for the simulation of proteins that have more than two thousand residues. Large lengths of DNA also can be studied more efficiently than existing CPU-based

implementations. The minimal time scale for protein folding is at least microseconds [208], which indicates the size of the computational burden required to study such systems via all-atom simulations. With OpenAWSEM, folding and functional mechanisms of even very large proteins can be simulated within a reasonable amount of clock time (hours or days), thereby opening the door for a wide range of functional biomolecular dynamics studies. The codes are written entirely with Python 3, including the user interfaces. The computationally costly part of the simulations is handled by the OpenMM library, which was coded with efficiency in mind. Python 3 provides great code readability and modification efficiency, and since the codes are automatically compiled while running, the time spent in compilation of the code is eliminated. Also, using the automatic computation of the derivatives of the Hamiltonian instead of explicitly coding the forces greatly simplifies the introduction and implementation of new energy terms to accommodate new physics. To illustrate this feature of OpenAWSEM, we have designed and implemented two sophisticated potentials for some specialized folding situations. One of these involves the introduction of a membrane burial dependent contact potential to describe proteins that are only partially buried in membranes. We have demonstrated that using this potential for structure prediction leads to more accurate structures than when the proteins are treated as uniformly living in one environment or the other. Another energy term that was easy to code was a density dependent disulfide bonding potential that mimicks the saturation of chemical bonds. Introducing this term generally improved structure predictions and also allowed us to computationally recapitulate Anfinsen's Nobel prize winning

experiments on ribonuclease. These two new potentials serve to illustrate the flexibility and extendability of the OpenAWSEM framework and will encourage the design of future coarse grained force fields for large biomolecular simulations using this computational software framework.

## **4.5. Methods**

### **4.5.1. Simulation setup**

The default values of the parameters in the annealing protocol for all the simulations performed in this study are given below. We maintained those values as being consistent with those typically used in the LAMMPS implementation of AWSEM-MD. (listed in S1 Document) We point out that for many problems involving very large systems, these run parameters should be revised for optimal efficiency. As a default in the structure prediction runs, we used the langevin integrator with friction of  $1\text{ps}^{-1}$ , time steps of nominal 5fs, and temperature going from 800K to 200K during simulated annealing. The simulations were carried out for 8 million steps. This corresponds roughly to  $40\ \mu\text{s}$  of laboratory time. Default forces included in our study are the connectivity, chain, chi, exclude volume, rama, rama modulated by proline, rama modulated by secondary structure input file "ssweight", contact, beta, pap and fragment memory terms. Each term can be turned on and off and vary in strength and setting in the force\_setups.py file. All OpenAWSEM and Open3SPN2 simulations were carried out with Nvidia V100 and all LAMMPS version simulations were carried out with Intel Xeon CPU E5-2650 v2 on the Rice NOTS server.

### 4.5.2. Q-value definition

The Q-value is a measure of how similar a predicted structure is to the correct native structure. To evaluate the quality of the protein predictions we used the Q value which is defined in Equation 4.19

$$Q = \frac{2}{(N-2)(N-3)} \sum_{i < j-2} e^{-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}} \quad \text{Equation 4.19}$$

where  $N$  is the total number of residues,  $i$  and  $j$  are sequence positions,  $r_{ij}$  is the distance between the CA of residue  $i$  and the CA of residue  $j$ .  $r_{ij}^N$  is the distance between CA of residue  $i$  and CA of residue  $j$  in native structure,  $\sigma_{ij} = (1 + |i - j|^{0.15})\text{\AA}$ . For  $Q_{water}$ ,  $N$  is the number of residues outside of the membrane, and the sum is taken over all those residues. For  $Q_{membrane}$ ,  $N$  is the number of residues outside the membrane, and  $\sigma_{ij} = 2(1 + |i - j|^{0.15})\text{\AA}$ .

## 4.6. Acknowledgments

The authors would like to thank Dr. Peter Eastman the developer of OpenMM for helpful discussions during the development of the OpenAWSEM and Open3SPN2 package. WL, CB, NPS, SJ, XC, MC, XG, AD and PGW were supported by the Center for Theoretical Biological Physics and sponsored by an NSF grant (PHY- 2019745) and by the D. R. Bullard-Welch Chair at Rice University, Grant C-0016. CB, PGW were supported by the PoLS Student Research Network sponsored by the NSF Division of

Physics grant 1522550. JM and JJP were supported by NSF grant BIO/MCB 1818328.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

# **Resolving the NF- $\kappa$ B Heterodimer Binding Paradox: Strain and Frustration Guide the Binding of Dimeric Transcription Factors**

This chapter is a slightly modified version of an article published in the Journal of the American Chemical Society and has been adapted here with permission from Potoyan et al. [209]. Copyright 2022 American Chemical Society.

### **5.1. Introduction**

Transcription factor oligomers regulate the flow of genetic information [210]. Even at the dimer level, different combinations of partners bind differently to specific genomic regions and can thereby trigger different sequences of downstream events, eventually leading to different cell fates [210,211]. There are many reasons why forming and acting as oligomers may be biologically adaptive for transcriptional regulation. The combinatorial diversity of pairing can allow a finer tuning of gene expression by turning a limited number of transcription factor coding

genes into a much larger pool of disparate transcription factor complexes with widely varying binding specificity toward different regulatory genomic sequences. The association of proteins by itself can then serve as a switch or a control point for exerting further levels of post-translational regulation [211]. Finally, oligomerization can be seen as adding nonlinearity to the system-wide kinetic network, which allows the response to changes in protein copy numbers to be more sensitive [212]. Such nonlinearity also offers noise buffering benefits [213,214] and can turn steady states of regulatory networks into oscillatory attractor patterns [215]. We thus see that uncovering the fundamental principles behind transcription factor association and DNA recognition may lead to a deeper understanding of basic cell biology and perhaps may allow one to devise better therapeutic strategies to intervene when regulatory networks fail.

In this chapter, we look at these basic questions by focusing on one specific important gene regulatory system that involves the transcription factor NF- $\kappa$ B. This transcription factor is the central regulator of an extensive and complex genetic broadcasting system [216]. We have shown that the behavior of this broadcasting network is controlled by the kinetic features of transcription factor binding and release from DNA [96,217]. The NF- $\kappa$ B family of proteins has many members, and these function in various combinations of homodimers and heterodimers. The heterodimeric form, made of two different subunits, p50p65, happens to be the most abundant in cells. This heterodimer binds to DNA significantly more strongly than do either of the p50p50 and p65p65 homodimers [218]. The steady-state ratio of different dimers in cells is clearly controlled via intricate gene regulatory

pathways, which cannot be explained solely by DNA binding affinities or dimerization equilibrium constants [219]. A priori, if binding to DNA were made up of individual additive contributions from each monomer, we would expect the heterodimer to have an affinity somewhat between the affinities of the two homodimers, not to be dramatically stronger than either of them. The fact that the heterodimer binds more strongly than either homodimer is, therefore, something of a paradox.

Using a combination of gel mobility shift and fluorescence anisotropy assays, Phelps et al. [218,220] have established that p65p50 binds to the kb target site of immunoglobulin enhancer with an affinity of  $\sim 10$  nM; p50p50 has a weaker affinity of  $\sim 50$  nM, and p65p65 has the weakest affinity of all three dimers at  $\sim 150$  nM. These measurements were all done at near-physiological conditions. The genome-wide patterns of binding are expected to be more complex, with different dimers displaying a distribution of affinities that show mutually exclusive preferences for some sequences. Nevertheless, protein binding microarray experiments [221] have shown that the distributions for 10-bp kb sites with different flanking bases still place p65p50 consistently as a stronger binder than both the p50p50 and p65p65 dimers. The tight binding of the heterodimeric NF- $\kappa$ B complex is apparently not an accident. In fact, cells have elaborated a mechanism we have termed “molecular stripping” to rapidly and completely remove the heterodimeric NF- $\kappa$ B from its many genomic target sites [96,217,222,223] (Figure 5.1A). The relative concentration of free and bound heterodimeric forms is thus under kinetic control. The homodimers, on the other hand, because of their weak binding, do not need molecular stripping,



since their dissociation is sufficiently rapid [224] (Figure 5.1A). Their relative concentrations of DNA-bound and free forms can equilibrate thermodynamically. The steady-state composition of the different dimers will thus be seen to be under kinetic, not thermodynamic, control owing to the strong heterodimeric association and molecular stripping.

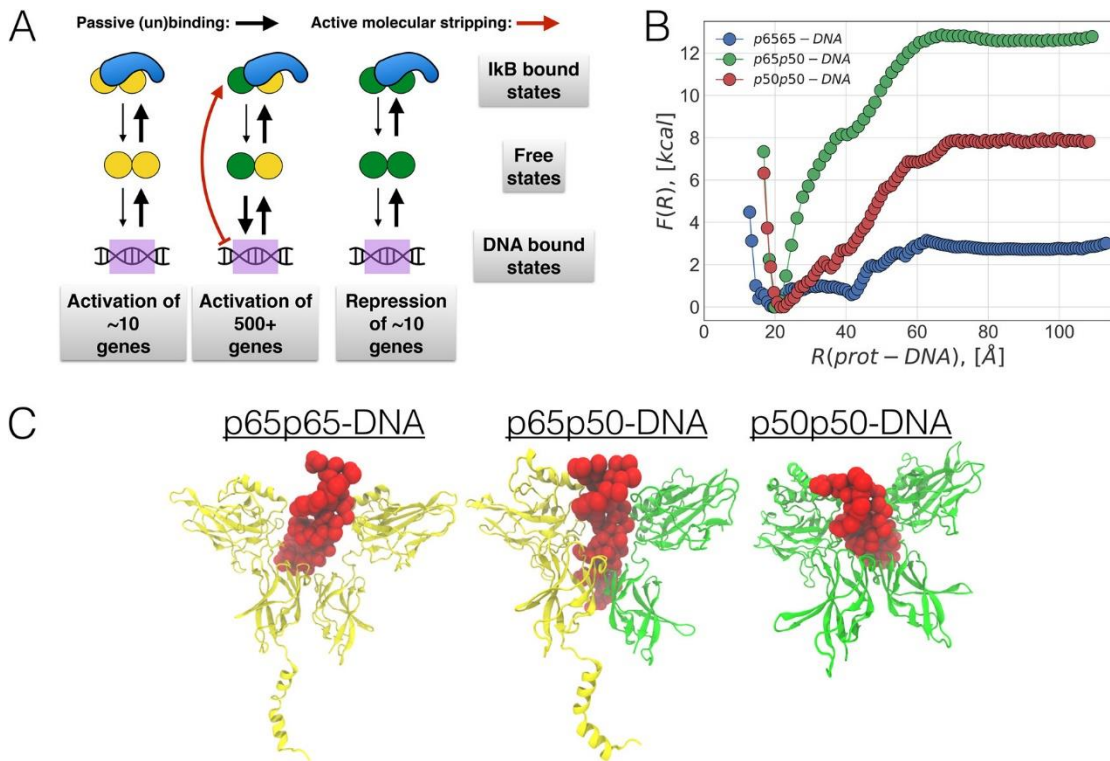


Figure 5.1. DNA-NFκB interactions.

(A) Simplified kinetic scheme that depicts the kinetics of interconversion between DNA-bound, free, and IκB-bound NFκB dimers. Due to a combination of factors including large number of binding sites and, on average, higher affinities, the p65p50 heterodimer of NF-κB relies on molecular stripping for generating a timely response to appropriate external signals. For the p65p65 and p50p50 homodimers, either there is no molecular stripping (p50p50 lacks NLS domain) or any effect of stripping is negligible because of the low binding affinity (p65p65). The nearly irreversible stripping process indicated by the red arrow renders the heterodimer population under kinetic control. (B) Free energy profiles of DNA dissociation from different NF-κB-DNA complexes. The p50p65 heterodimer is shown in blue, p65p65 homodimer is shown in green, and p50p50 homodimer is shown in

red. (C) Representative structures from the simulations of NF- $\kappa$ B dimers bound to the same 20-bp-long DNA segment.

In the present work, we are concerned with understanding the basic molecular features that resolve the “binding paradox” by using predictive coarse-grained models for energy landscapes of the proteins and the DNA in order to simulate the structural and mechanical changes that accompany the dissociation of these transcription factors from the DNA-bound complexes.

To set the stage for describing the nature of these changes and the computational tools used for analyzing them, it is instructive to first briefly survey some basic tenets of the theory of protein dynamics. The modern view of protein dynamics is provided by energy landscape theory [225]. A key notion from this theory is that the landscape overall resembles a funnel, allowing rapid sampling of the global conformational space of foldable proteins. The funnel-like shape of the global energy landscape is a result of the evolutionary minimization of the conflict between different residual interactions in proteins, which is also known as “frustration”. This principle of minimal frustration [226,227] explains that evolution has sculpted funnel-like landscapes for proteins in order to eliminate deep kinetic traps, which would otherwise lead to unreasonably long folding times. The funnel topography ensures rapid and reliable folding into functional states. Because of the funnel-like shape of the landscape, most conformational changes can be described either as arising from distortions of the native structure or as arising from the making and breaking of native interactions; completely non-native structures are

rare as mechanistic intermediates (but do arise on occasion). The funnel-like shape of the folding landscape can also be exploited for deriving potentials for protein structure prediction using machine learning strategies [228]. The resulting optimized landscapes provide predictive force fields that can be used to simulate functional dynamics, as is done in the present work. The bottom of the funnel has low but nonzero entropy and contains highly structured conformations which can occupy multiple distinct structural basins. The multiplicity of these basins arises partly from the residual frustration [229]. The frustration can be readily quantified by computational tools such as the Frustratometer [230]. Residual frustration is not necessarily an accident of nature coming from some failure of evolutionary optimization but can signal an adaptation to facilitate ligand-induced functional allosteric changes of protein conformational states [227]. Because of the multibasin character of most functional proteins, the dynamics of native proteins is anisotropic [229,231]. While individual minima can be well described by quasi-harmonic local fluctuations with a good fit to crystallographic B factors, jumps among the different basins are accomplished via larger amplitude changes also known as “proteinquakes” [232,233]. Such large-amplitude motions involve high-order nonlinear deformations, “cracking” [234–237], and local unfolding [229], all of which require going beyond the harmonic approximations in the energy intrinsic to the idea of elasticity [229,231]. These large-amplitude motions occur predominantly by moving along the directions of some of the few low-frequency soft modes which can be extracted by techniques such as principal component analysis (PCA) of fluctuations in long molecular dynamics simulation runs. Thus, one can use concepts

from the elasticity theory of materials [232], multibasin approximations [238,239], and linear response theory [240,241] to predict and explain the basis of ligand-induced structural changes. Below we use an array of these tools to characterize the protein dynamics in the NF- $\kappa$ B complexes to uncover the structural and dynamic resolution of the “binding paradox” of NF- $\kappa$ B transcription factors. In particular, our analysis in some sense should also be considered an extension of the approach first introduced by Miyashita et al. [232], where elasticity theory was employed for scrutinizing the allosteric movements of adenylate kinase in terms of mechanical strain and strain-induced cracking. A novelty of the present approach is the use of a nonstandard set of collective coordinates: the principal components of pair distances involved in the space of physical contacts. These new collective variables enable us to carry out very detailed quantification and visualization of how the mechanical strain results from stretching numerous individual physical contacts distributed within and between protein domains. Similarly, we introduce a binary-valued representation of the dynamical contacts (0 = broken, 1 = unbroken) in order to visualize the specific cracking transitions where numerous contacts become completely broken. We call the principal components associated with the contact pair distances “strain PCs” and the principal components of the fluctuations of the binary contact observables “cracking PCs”. These two kinds of collective coordinates allow us to visualize the rich cooperative interplay between straining and cracking of specific contacts, shedding light on how the allosteric coupling between the dimer interaction interface and the DNA binding interface is made possible through soft modes of motion of different NF- $\kappa$ B dimers, which then contribute to the differences

in their binding affinities. We also pinpoint how energetic frustration in the heterodimeric interface allows the heterodimer to have greater flexibility at this key hinge, which enables the heterodimer to bind more intimately to its DNA targets than do the more rigid and unfrustrated homodimers, which cannot clamp down as effectively onto the DNA. We believe the tools introduced in the present analysis of the NF- $\kappa$ B system shed light on the relationship between binding affinity and molecular architecture of transcription factors in general and also can be used to analyze and understand other systems. Our calculations can be further tested through protein engineering studies that probe how transcription factor–DNA interactions can be changed through seemingly remote mutations.

## 5.2. Methods

We use physics-based, coarse-grained predictive protein and DNA models in order to simulate the process of DNA dissociation from the NF- $\kappa$ B–DNA bound complexes as well as to generate equilibrium conformational ensembles for the endpoint states of the free and DNA-bound NF- $\kappa$ B dimers. We use the predictive protein energy landscape model provided by the Associative memory, Water-mediated, Structure and Energy Model (AWSEM) developed in the Wolynes’s group [98] together with the 3 Sites Per Nucleotide (3SPN.2C) model for DNA developed by de Pablo’s group [86,242]. The total Hamiltonian,  $H_{tot} = H_{AWSEM} + H_{3SPN.2C} + H_{pp-DNA}$ , consists of terms describing both the interactions within the protein components and within the DNA and the interactions between the two molecular species. Besides steric terms preventing molecular overlap, the interactions

between protein and DNA are largely electrostatic. We have chosen to approximate these electrostatic interactions with a simple mean-field Debye–Hückel potential,

$$H_{pp-DNA} = \sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij}} e^{-\frac{r_{ij}}{\kappa_D}} \quad \text{Equation 5.1}$$

where  $\epsilon$ , the dielectric constant, is set to 80, and  $\kappa_D$ , the Debye length, is set to 10 nm in order to approximate the physical environment of the cell.

To carry out simulations, we use as initial input the crystal structures provided in the Protein Data Bank (PDB) for p65p65 (2RAM), p65p50 (1LE5), and p50p50 (1SVC) NF- $\kappa$ B dimers. The raw PDB structures of the dimers, however, are missing important structural moieties which we had to computationally reconstruct before running the equilibrium sampling of dimers and their complexes. The reconstruction of structures with complete sequences has been made possible by using the predictive force field of AWSEM [98], which allows us to rapidly locate and sample equilibrium configurations consistent with crystal structures.

For the constant-temperature simulations, we perform routine minimization and equilibration protocols for all structures before running long-time-scale production runs, which are then subjected to analysis. A 20-bp DNA fragment with the same sequence that was used in the in vitro experiments [224] on molecular stripping is used for simulating DNA-bound NF- $\kappa$ B complexes. The orientation of the DNA has been constrained, however, in all of the windows of the umbrella sampling simulations to remain parallel to the initial orientation in DNA-bound complex. By this constraining of the orientation, we exclude contributions from rotational

entropy as the DNA dissociates, as well as finite chain effects of DNA, so as to make sure that one measures the free energy of DNA detachment without any contributions from sliding to the end of a short fragment.

We have carried out several kinds of PCA on the conformational ensembles generated via constant-temperature sampling (equivalent to 500  $\mu\text{s}$  of real time [96]) as well as on ensembles sampled along the DNA dissociation coordinate (equivalent to 10  $\mu\text{s}$  of real time [96]). Cartesian principal components are computed using the ordinary Cartesian coordinates of the  $\text{C}\alpha$  atoms of the long protein chains [ $r^{3N}(t)$ ]. We then use the principal eigenvectors found for the free NF- $\kappa$ B dimers as a basis for projecting the trajectories of complexes of NF- $\kappa$ B bound with DNA. This allows us to compare directly how the same twist/breathing modes evolve upon binding to the DNA.

To compute the contact-based principal components, we have used the same structural ensembles as were used for the Cartesian PCA. First we use a trajectory-averaged structure to define the  $\text{C}\alpha$  Cartesian coordinates  $\langle r^{3N} \rangle$ . The coordinates of this averaged structure  $\langle r^{3N} \rangle$  are then used to compute the set of all pairwise distances,  $d'_{ij}$ , which we use for fiducial purposes. Next these pairwise distances are filtered to exclude contacts with sequence spacing closer than four amino acid residues apart as well as those that are farther apart than 10–12  $\text{\AA}$  in the 3D averaged structure. Once these pairs of interactions are defined, the set of pairwise distances, filtered in the same way, is then computed for all the snapshots of the trajectory  $d_{ij}(t)$ . By this means, we have selected only those nonbonded contact

pairs which contribute significantly to the energy of elastic strain. For the strain PCA, we monitor the absolute magnitudes of the contact distances  $d_{ij}(t)$ . To compute the cracking principal components, we use binary quantities  $q_{ij} = 0,1$  to indicate the extent of contact formation, where  $q_{ij} = 0$  corresponds to a contact being broken when the pairs become too far separated,  $d_{ij}(t) > 8\text{\AA}$ , while  $q_{ij} = 1$  corresponds to a contact having been formed with a close distance,  $d_{ij}(t) < 8\text{\AA}$ . Using these contact definitions, we can quantify the elastic energy of proteins by a uniform quasi-harmonic approximation [232],

$$\Delta E_{elast}(t) = \frac{1}{2} \left\langle \sum_{i,j} (d_{ij}(t) - d_{ij}^{ref})^2 \right\rangle \quad \text{Equation 5.2}$$

where  $\langle \rangle$  denotes the ensemble average, and  $d_{ij}^{ref}$  denotes the pairwise distances of the reference structure. One can also estimate the fractional energetic contribution from any individual strain or cracking principal component,  $PC_n(t)$ , by using a similar quasi-harmonic approximation:

$$\Delta E_{PC_n}(t) = \frac{1}{2\lambda_n^2} \left\langle (PC_n(t) - PC_n^{ref})^2 \right\rangle \quad \text{Equation 5.3}$$

where  $PC^{ref}$  is the value for the average reference structure, and  $\lambda_n$  is the eigenvalue associated with the  $n$ th principal component.

### 5.3. Results and Discussion

When binding to DNA or to other ligands, many proteins are forced to undergo large-scale conformational rearrangements, where some intramolecular



contacts inevitably must break in order to form new contacts that are more suited to the ligand. This is the basis of allosteric communication.

The dissociation of stably bound protein–DNA complexes is a slow process. A rare, activated event is needed for unbinding to occur. We therefore used umbrella sampling simulations with a center-of-mass distance restraint between the DNA and NF- $\kappa$ B dimers in order to extract relevant free energy profiles along this zeroth-order reaction coordinate (see Methods and Appendix D). Through such simulations, we are able to determine binding affinities at least qualitatively. The most important result of this analysis is that simulations using the AWSEM force field actually do recapitulate the “paradoxical” pattern of dissociation observed in the experiments in which the heterodimer binds more strongly to the DNA than do either of homodimers (Figure 5.1B). The predicted differences in binding affinities mirror those found by experiment. The ratios of the predicted to the measured binding free energy,

$$\eta = \frac{\Delta F^{AWSEM-3SPN.2C}}{\Delta F^{exp}} \quad \text{Equation 5.4}$$

for the various dimers are as follows: The p65p50 heterodimer binds with a predicted affinity that is about right,  $\eta \approx 1$ . The predictions of the binding free energies for p50p50 and p65p65 are each larger than are the experimental values: for the p50p50 homodimer  $\eta \approx 2$ , while  $\eta \approx 3$  for the p65p65 homodimer. We can see that, in quantitative terms, the magnitude of the paradoxical discrepancy is somewhat underestimated by the simulation in comparison with experiment, but nevertheless the predicted disparities in the binding energies show very clearly that

the heterodimer binding is much stronger than binding of either homodimer, as is seen in the laboratory.

The model we use has no DNA sequence-specific effects other than those that would come from the sequence-dependent shapes of the DNA grooves that are encoded in the 3SPN.2C Hamiltonian. We note that NF- $\kappa$ B binding is known to be relatively promiscuous, [243] suggesting this is a reasonable approximation. While any additional specific interactions in the Hamiltonian could change the picture quantitatively, we see that the large free energy difference between the binding of the heterodimer and the homodimers is already captured by the present purely electrostatic sequence-independent model of the direct protein–DNA interactions (Figure 5.1B). These results then suggest that the model is sufficiently realistic to justify extracting more of the relevant structural features that give rise to the observed dissociation trends. In Figure 5.1C, we show sampled configurations of the final DNA-bound structures of the various dimers. We immediately see that the thermodynamic binding differences are also accompanied by large-scale structural differences in the DNA-bound species.

By calculating the average elastic and electrostatic energy changes (see Materials and Methods) between the structural ensembles of the DNA-bound and the DNA-free NF- $\kappa$ B dimers, we see that the structural changes in the dimers upon DNA binding are accompanied by a changing balance between elastic energy and electrostatic energy. These two energies play off each other in a seesaw fashion. There is significant lowering in the internal elastic energy for the p65p50

heterodimer, which exceeds the energetic stabilization of the p65p65 and p50p50 homodimers (Figure 5.2). On the other hand, the electrostatic stabilization upon DNA binding appears comparable for p65p50 and p50p50 dimers but is less favorable for the p65p65 homodimer. Thus, both our calculations and the experiments point out that resolving the binding paradox requires one to look beyond the protein–DNA interaction surface itself and to inquire further into the way transcription factors are built to see how their inner elastic strains and the wiring of their contacts enables these proteins to sterically accommodate and release the DNA.

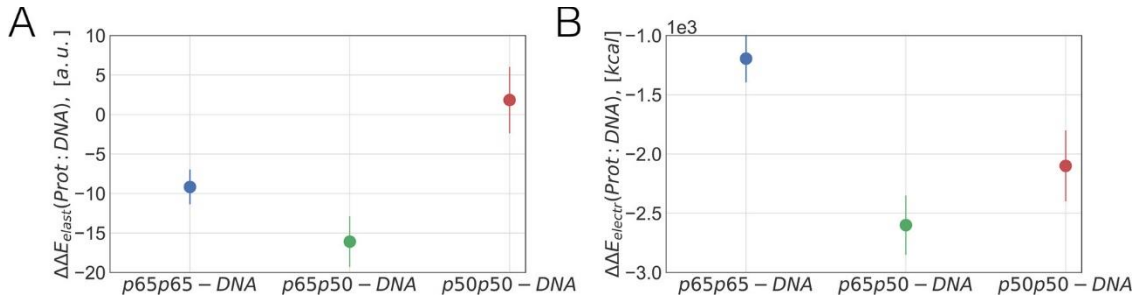


Figure 5.2. Differences in elastic and electrostatic energies between DNA-bound and free forms of NF- $\kappa$ B.

(A) Differences in average protein elastic energies between DNA-bound and free forms of NF- $\kappa$ B:  $\Delta\Delta E_{elast}(prot:DNA) = \Delta E_{elast}(prot:DNA) - \Delta E_{elast}(prot)$ .

(B) Differences in average electrostatic energies (estimated at a mean-field Debye–Hückel level) between the DNA-bound and free forms of NF- $\kappa$ B:  $\Delta\Delta E_{elec}(prot:DNA) = \Delta E_{elec}(prot:DNA) - \Delta E_{elec}(prot)$ .

We therefore carried out long-timescale (equivalent to  $\sim 10^2 \mu s$  in laboratory time) constant-temperature simulations of the homo- and heterodimeric forms of NF- $\kappa$ B in their free and DNA-bound forms. We then analyzed the ensembles of

conformations generated in this way using several types of PCA. PCA provides a way of decoupling correlated fluctuations to find dominant collective coordinates that may function as reaction coordinates in exploring conformational transitions [229,244,245]. This analysis allows us to uncover how the accommodation of the protein to the DNA is made possible not only by rigid-body domain motions and localized strain rearrangements but also ultimately by cracking at specific sites in the dimer interfaces. These dimer interfaces apparently have evolved to be frustrated in the ground-state conformation so as to facilitate their rupture.

To disentangle these different collective motions, we employ several different forms of PCA. One of these analyses is the conventional PCA based on describing protein conformations using the Cartesian positions of residues in three-dimensional space [246]. Many papers [96,229,244,245] have documented that these Cartesian principal components generally track the rigid-body motions of protein subdomains relative to each other.

Two other different kinds of PCA give further insight into the way these motions occur and allow us to identify the specific contacts whose local stretching and contraction contribute to the changes in elastic energy. We use one form employing pair distances in the space of physical contacts that shows where the molecule stretches and contracts (strain PCA) and a distinct form of analysis involving another set of observables that measures whether any contacts are completely made or broken (cracking PCA), which thereby monitors cracking (see Materials and Methods). Years ago, cracking PCA in contact space was used to

predict and follow the subtle conformational changes induced by phosphorylation, a common form of post-translational modifications in proteins [241,247]. Finally, by following both the dominant conventional Cartesian principal components and the strain principal components, we find the specific locally strained contacts which are most correlated with the global structural changes in binding.

As visualized by Cartesian PCA, the dominant thermal motions of the free NF- $\kappa$ B dimers are domain twisting and breathing motions of the two DNA-binding domains relative to each other [96]. These motions account for over 90% of the thermal fluctuations in the Cartesian coordinates of the protein (Appendix D). The twist motion persists in the DNA-bound form, where it now appears as a shearing motion accompanied by one-dimensional sliding along the DNA. The breathing motion in the free protein dimers, on the other hand, is suppressed by binding DNA, which fills the cavity between the NF- $\kappa$ B binding domains. The suppression of the twist mode and the activation of the breathing mode, therefore, are seen as important contributors to the dissociation of NF- $\kappa$ B–DNA complexes. In a previous paper [96], we have shown how the naturally evolved inhibitor protein I $\kappa$ B facilitates dissociation by freezing the domain twist in an open state. In this way, I $\kappa$ B holds one of the domains of NF- $\kappa$ B in a twisted configuration so that one of the domains can no longer make strong electrostatic contact with the DNA, thereby allowing the DNA to escape the embrace of NF- $\kappa$ B more easily since it now contacts only a single NF- $\kappa$ B domain. In the absence of I $\kappa$ B, the dissociation can only take place by means of energetically costly thermal fluctuations, which accomplish the domain twist and allow the breathing motion and thus DNA escape.

While the Cartesian principal components give a global picture of the motions needed for binding in each of the individual dimers, the energetic differences between these dimers come from changes that are localized in the protein structure (Figure 5.1C). These structural changes are captured by the local forms of PCA (Figure 5.3), which allow us to pin down those specific contacts that facilitate the global motions that are highlighted in the Cartesian principal components. The principal components of the fluctuations of the relative distances involved in local individual contacts localize the pivots for the large-amplitude motions of the molecule. We used these strain principal components to detect the regions most prone to deformation and cracking. Thus, when the strain becomes large enough, we see that some specific contacts break while others may form (Figure 5.4 and Figure 5.5). The principal components of the correlated fluctuations of occupation of contacts thus detect and localize the sites of cracking; i.e., they locate the fault lines that initiate the “proteinquake”.

While globally the main principal Cartesian motions appear similar for all of the dimeric forms of NF- $\kappa$ B (Figure 5.3 and Appendix D), the localized strain motions that drive these global motions are quite different for the heterodimer and for the homodimers. The strain principal components thus allow quantification of the type and magnitude of various contact contributions to the overall elastic energy changes (Figure 2). Additionally, using the correlations between Cartesian and strain principal components, we find that the twist results in a rather cooperative contact straining/breaking at the dimerization interface, while the breathing motion and lower amplitude thermal motions take place via a coordinated exchange of

contacts, some contacts being made and others being broken more or less at the same time. This pattern is seen by computing the principal components of the binary contact probability fluctuations that monitor cracking (Appendix D).

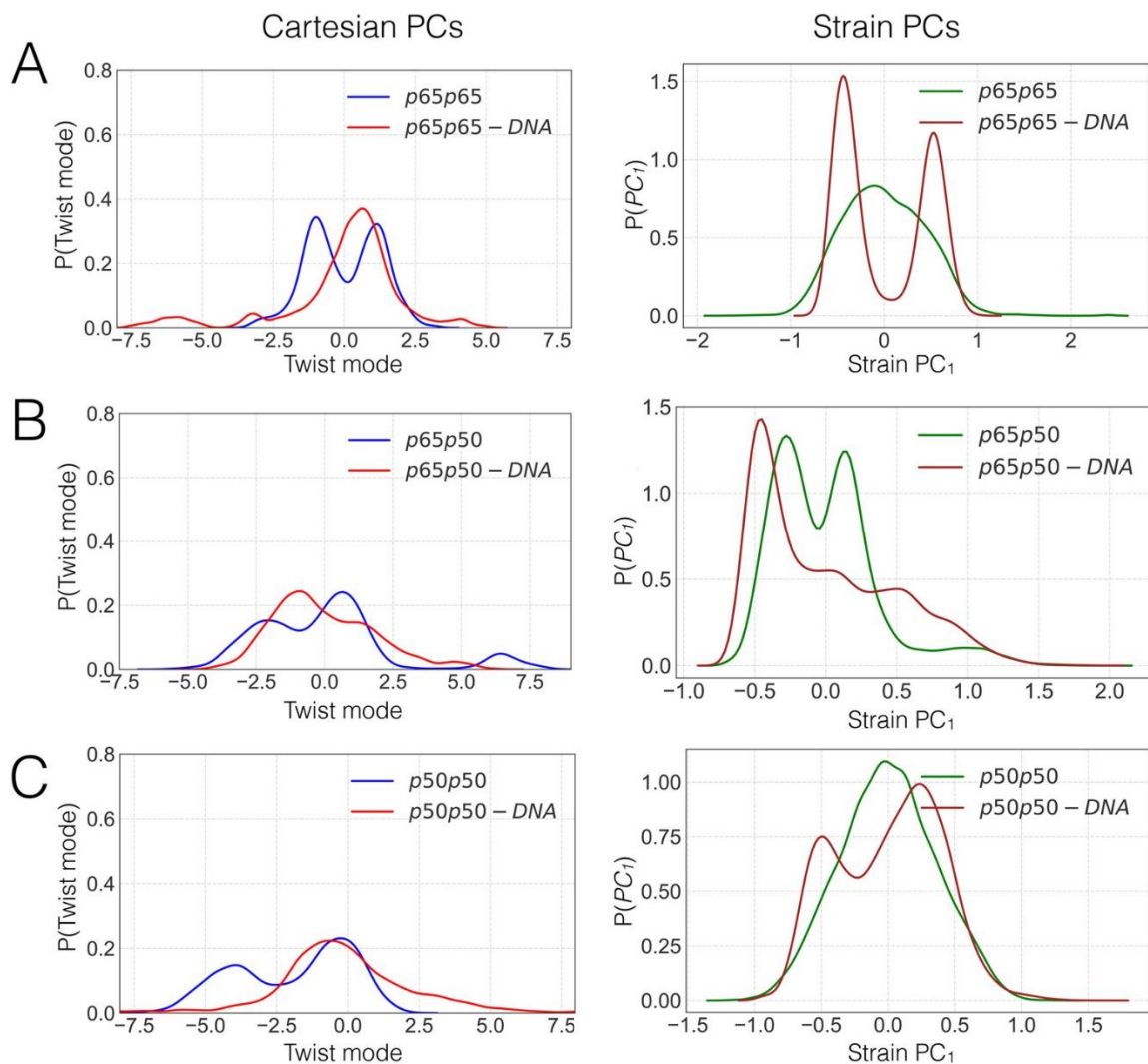


Figure 5.3. Distributions of Cartesian and strain PCs for all three dimers of NF- $\kappa$ B in their free and DNA-bound structural ensembles.

(A) Distribution of twist and first strain PC for p65p65 and p65p65–DNA. (B) Distribution of twist and first strain PC for p65p50 and p65p50–DNA. (C) Distribution of twist and first strain PC for p50p50 and p50p50–DNA.

Projecting the conformational ensemble of the DNA-bound NF- $\kappa$ B complex onto the first strain principal component, we see that, upon binding to DNA, the multiple conformational states of the free p65p50 heterodimer NF- $\kappa$ B coalesce into one single macrostate (Figure 5.3). For the homodimers we see an opposite pattern, where the relatively more rigid dimer interface in the free dimer becomes more dynamic and the strain principal component distribution bifurcates into multiple states upon binding to DNA. This change in the strain PCs indicates that there is global frustration (i.e., frustration at a distance) between, on the one hand, the tendency of NF- $\kappa$ B dimers to form more stable rigid dimers through contacts in the dimerization interface and, on the other hand, their need to allow some conformational flexibility at the dimerization interface so as to be stronger DNA binders.

These local structural changes contribute toward the total elastic energy, which in the case of both homodimers is less favorable in the DNA-bound form than in the free form (Figure 5.2). The fluctuations of the first strain principal component in heterodimer p65p50 happen to be the most correlated with the global twist mode (Appendix D). On the other hand, owing to the greater rigidity of the homodimers, the global twist mode is most correlated with a minor principal component, the 11th strain principal component, PC<sub>11</sub> (Appendix D). This high-order principal component makes a factor of  $\sim 10$  smaller contribution to the total fluctuations than does the first strain principal component. The first 10 strain principal components in both homodimers are uncorrelated with global motions of domains relative to each other but rather correspond to a variety of intradomain local deformations.



This is why the total internal elastic energy of homodimers quantified by the quasi-harmonic approximation is much lower than it is for the p65p50 heterodimer (Figure 5.2). If one compares the average electrostatic free energy stabilization upon DNA binding, then electrostatically the p65p50 heterodimer appears to be comparable to p50p50. We see that what truly sets these dimers apart from each other is the nature of the elastic energy changes upon the DNA binding.

We therefore now turn to an analysis of the local frustration patterns in NF- $\kappa$ B dimers to understand why these local deformations are so different in these different dimeric molecules while their global motions may appear to be not too different from each other. In Figure 5.4, we show the frustration pattern of the native contact map along with the strain principal component and the corresponding contact occupation principal component that shows where there is cracking in the region of high strain (see also Appendix D). We see that frustration, strain, and cracking are all quite well colocalized at a specific fault in the protein, much like the fault lines in seismology.

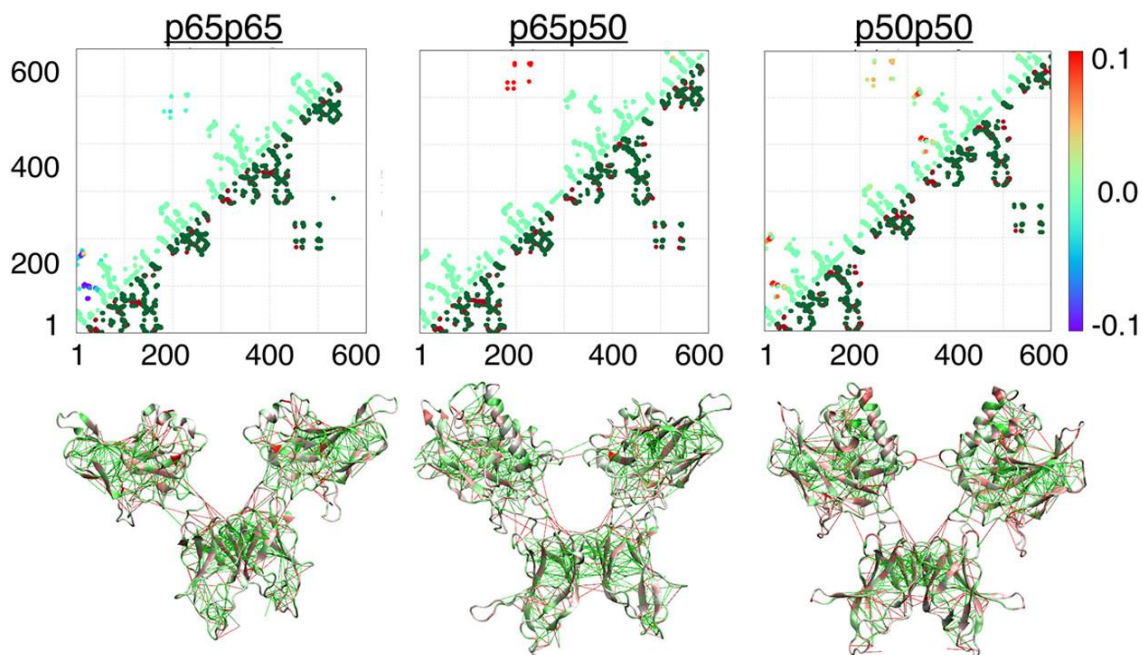


Figure 5.4. Frustration in NF- $\kappa$ B dimers.

Contact maps color-coded to show the frustration patterns (lower triangle) and the main contacts that contribute to strain motion (upper triangle) (PC1) for p65p650, p65p50, and p50p50 dimers. The color bar at the right indicates the magnitude of fluctuation in strain PC motion. The color of the contacts for the frustration patterns shown in both the contact map and the three-dimensional structures is indicated as red for highly frustrated contacts and green for minimally frustrated contacts.

Using the PCA in the space of the pairwise distances between residues in physical contacts from constant-temperature equilibrium sampling allows us to easily identify the most strained contacts present in the free NF- $\kappa$ B dimers (Figure 5.4). These contacts form along the interface between the dimerization domains of NF- $\kappa$ B. These contacts, therefore, contribute significantly to the total changes in elastic energy during the accommodation of the DNA, as quantified by the quasi-harmonic approximations used by Miyashita et al. [232] (Figure 5.2). The dimer interface region also turns out to be more frustrated in the heterodimer than it is in

either of the homodimers (Figure 5.4). The fault in the p65p50 heterodimer has apparently evolved to be easy to break. The more frustrated interfacial contacts of the heterodimer are precisely localized (Figure 5.4) so as to favor the shear motions that enable the twist mode in both the free and the DNA-bound forms.

The role of frustration can be quantified by following the changes in the number of frustrated contacts along the twist-like motion for each dimer (Figure 5.5). We find that the heterodimer stands out as the only dimer that forms more contacts and becomes less frustrated at the end of the twist motion, while both the p65p65 and p50p50 homodimers break more contacts and become more frustrated as the dimer deviates from its equilibrium-averaged configuration to accommodate better the DNA. This shows that the high mobility of the heterodimer correlates with its unusually high binding affinity to the DNA. The accommodating motions are facilitated by the positioning of frustrated contacts near the dimerization interface region (Figure 5.4).

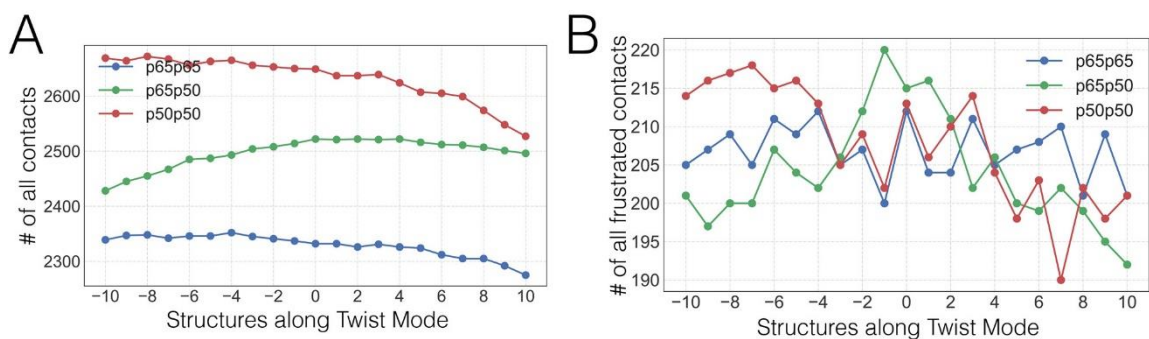


Figure 5.5. Change in the frustration along the twist modes.

(A) Change in the total number of all contacts. (B) Change in the total number of frustrated contacts as the dimeric molecules traverse along the twist coordinate.

In both the free p65p65 and p50p50 homodimers, the protein dimerization interface is less frustrated (Figure 5.4 and Appendix D) than in the heterodimer. Thus, the homodimers are indeed correspondingly more rigid, so that the twist motion is suppressed in these homodimers to a great extent compared to the heterodimer. The low frustration of the dimerization domain interfaces in homodimers likely follows from symmetry principles [248,249]. Indeed, both directed evolution experiments and simulations have provided ample evidence that symmetric associations are energetically most stable and are therefore more likely to be discovered by evolution [250]. There are symmetry arguments for the apparent over-representation of homodimers in the functional protein universe [251]. Thus, we suspect that heterodimerization in NF- $\kappa$ B has been selected by evolution to be the dominant form, thanks to its manifesting a more beneficial trade-off between thermodynamic stability and allosteric functional DNA binding that is needed to allow kinetic control of a large genetic broadcasting network [216].

Finally, by computing the protein elastic energy for each conformational ensemble along the DNA dissociation coordinate and by monitoring the cracking contacts along with the strain contacts in the quasi-harmonic approximation, we see that the elastic energy is strongly correlated with the rupture of interdomain contacts in the NF- $\kappa$ B as it pulls away from DNA (Figure 5.6). By comparing these events during passive DNA dissociation to what happens when the I $\kappa$ B induces the dissociation of the DNA and strips the heterodimer p65p50 from DNA sites, we see that I $\kappa$ B's catalytic action can be traced to reducing the large elastic barrier (Figure

5.6C), which allows DNA to dissociate with less breakage of the dimerization interdomain contacts.

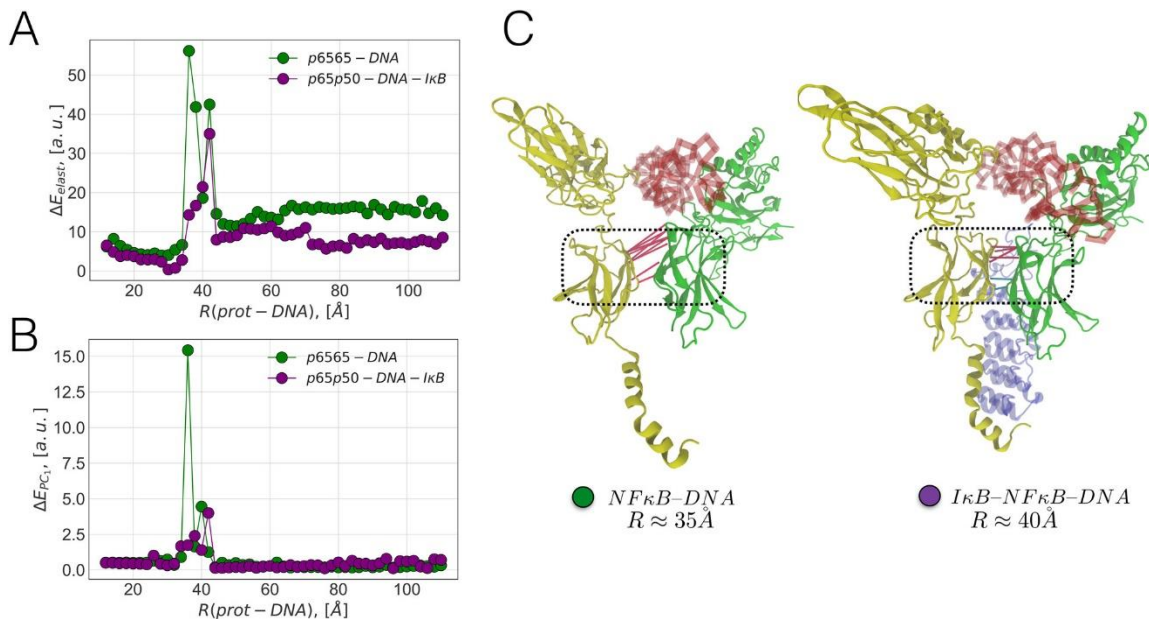


Figure 5.6. Profiles of the elastic energy along the DNA dissociation coordinate for DNA dissociation from IκB-bound (red) and IκB-free (blue) heterodimer p50p65 forms. (A) Change in the total elastic energy of NF-κB. (B) Elastic energy associated with the cracking PC1, which is localized at the dimerization interface of the p65p50 heterodimer. (C) Structures that correspond to transition states in the cracking PC1 profile, which monitors making/breaking of contacts at the dimerization interface. Contacts between the dimerization domains are highlighted, with red lines indicating cracked contacts ( $q_{ij} = 0$ ) and blue ones indicating uncracked ones ( $q_{ij} = 1$ ).

## 5.4. Conclusion

In the present work, we have tried to illustrate the way the interactions between the dimerization interfaces of protein domains in transcription factors determine the nature of the allosteric motions when the complexes bind to DNA. The differences in the free energy landscapes of the binding and dissociation of

heterodimers and homodimers from DNA reflect the conflicts between the elastic energies determined by the flexibility of dimerization interfaces of the transcription factors and the need for forming a strong electrostatic contact at the interface between the DNA and the DNA-binding domains of the protein complexes. We have carried out long-timescale simulations with a coarse-grained predictive protein-DNA force field showing the way DNA binding modulates the elastic deformations of the free transcription factors. Employing principal component analysis in the space of physical contacts along with the more conventional principal component analysis in the space of Cartesian coordinates of all backbone atoms allows us to understand how frustration at the dimeric interfaces controls motional flexibility by lubricating the fault that leads to the “proteinquake” necessary for DNA binding. The more frustrated interface of the heterodimer p50p65 leads to increased conformational disorder in the free form of the heterodimer, allowing p50p65 to bind more intimately to DNA than do either of the two homodimers, p65p65 and p50p50, which by having less frustrated and thus more rigid dimerization interfaces are prevented from optimally accommodating the DNA.

## **5.5. Acknowledgments**

We acknowledge financial support by the D.R. Bullard-Welch Chair at Rice University, Grant C-0016, and PPG Grant P01 GM071862 from the National Institute of General Medical Sciences.

# References

1. Schrodinger E. What is life? 1944.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*. 1999;402: C47-52. doi:10.1038/35011540. PubMed PMID: 10591225
3. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. *Science* (80- ). 2011;334: 517–520. doi:10.1126/science.1208351. PubMed PMID: 22034434
4. Zwier MC, Chong LT. Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol*. 2010;10: 745–752. doi:10.1016/j.coph.2010.09.008. PubMed PMID: 20934381
5. Ingólfsson HI, Lopez C a., Uusitalo JJ, de Jong DH, Gopal SM, Periole X, et al. The power of coarse graining in biomolecular simulations. *Wiley Interdiscip Rev Comput Mol Sci*. 2014;4: 225–248. doi:10.1002/wcms.1169. PubMed PMID: 25309628
6. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 2005;102: 14338–14343. doi:10.1073/pnas.0504070102. PubMed PMID: 16176987
7. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181: 223–30. doi:10.1126/science.181.4096.223. PubMed PMID:

4124164

8. Dominguez R, Holmes KC. Actin structure and function. *Annu Rev Biophys.* 2011;40: 169–186. doi:10.1146/annurev-biophys-042910-155359. PubMed PMID: 21314430
9. Pollard TD. Actin and Actin-Binding Proteins. *Cold Spring Harb Perspect Biol.* 2016;8: 1–18. doi:10.1101/cshperspect.a018226. PubMed PMID: 26988969
10. Stricker J, Falzone T, Gardel ML. Mechanics of the F-actin cytoskeleton. *J Biomech.* 2010;43: 9–14. doi:10.1016/j.jbiomech.2009.09.003. PubMed PMID: 19913792
11. Mofrad MRK. Rheology of the Cytoskeleton. *Annu Rev Fluid Mech.* 2009;41: 433–453. doi:10.1146/annurev.fluid.010908.165236
12. Straub FB. Actin. *Stud Inst Med Chem Univ Szeged.* 1942;2: 3–15.
13. Poglazov BF. Actin and coordination of metabolic processes. *Biochem Int.* 1983;6: 757–765. . PubMed PMID: 6385979
14. Shaevitz JW, Gitai Z. The structure and function of bacterial actin homologs. *Cold Spring Harb Perspect Biol.* 2010;2: 1–18. doi:10.1101/cshperspect.a000364. PubMed PMID: 20630996
15. Ananthkrishnan R, Ehrlicher A. The forces behind cell movement. *Int J Biol Sci.* 2007;3: 303–317. doi:10.7150/ijbs.3.303. PubMed PMID: 17589565
16. Mueller J, Szep G, Nemethova M, de Vries I, Lieber AD, Winkler C, et al. Load



- Adaptation of Lamellipodial Actin Networks. *Cell*. 2017;171: 188-200.e16.  
doi:10.1016/j.cell.2017.07.051. PubMed PMID: 28867286
17. Svitkina TM. The actin cytoskeleton and actin-based motility. *Cold Spring Harb Perspect Biol*. 2018;10: 1–22. doi:10.1101/cshperspect.a018267.  
PubMed PMID: 29295889
  18. Sept D, Xu J, Pollard TD, McCammon JA. Annealing accounts for the length of actin filaments formed by spontaneous polymerization. *Biophys J*. 1999;77: 2911–2919. doi:10.1016/S0006-3495(99)77124-9. PubMed PMID: 10585915
  19. Page R, Lindberg U, Schutt CE. Domain motions in actin. *J Mol Biol*. 1998;280: 463–474. doi:10.1006/jmbi.1998.1879. PubMed PMID: 9665849
  20. Hanson J, Lowy J. The structure of F-actin and of actin filaments isolated from muscle. *J Mol Biol*. 1963;6: 46–60. doi:10.1016/S0022-2836(63)80081-9
  21. dos Remedios C, Chhabra D. Actin-binding proteins and disease. Springer Science & Business Media; 2008.
  22. Gittes F, Mickey B, Nettleton J, Howard J. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J Cell Biol*. 1993;120: 923–934. doi:10.1083/jcb.120.4.923. PubMed PMID: 8432732
  23. Orlova A, Chen X, Rubenstein PA, Egelman EH. Modulation of yeast F-actin structure by a mutation in the nucleotide-binding cleft. *J Mol Biol*. 1997;271: 235–243. doi:10.1006/jmbi.1997.1163. PubMed PMID: 9268655

24. McGough A, Pope B, Chiu W, Weeds A. Cofilin changes the twist of F-actin: Implications for actin filament dynamics and cellular function. *J Cell Biol.* 1997;138: 771–781. doi:10.1083/jcb.138.4.771. PubMed PMID: 9265645
25. Papalazarou V, Machesky LM. The cell pushes back: The Arp2/3 complex is a key orchestrator of cellular responses to environmental forces. *Curr Opin Cell Biol.* 2021;68: 37–44. doi:10.1016/j.ceb.2020.08.012. PubMed PMID: 32977244
26. Ennomani H, Letort G, Guérin C, Martiel J-L, Cao W, Nédélec F, et al. Architecture and Connectivity Govern Actin Network Contractility. *Curr Biol.* 2016;26: 616–26. doi:10.1016/j.cub.2015.12.069. PubMed PMID: 26898468
27. De La Cruz EM, Roland J, McCullough BR, Blanchoin L, Martiel JL. Origin of twist-bend coupling in actin filaments. *Biophys J.* 2010;99: 1852–1860. doi:10.1016/j.bpj.2010.07.009. PubMed PMID: 20858430
28. Mohapatra L, Goode BL, Jelenkovic P, Phillips R, Kondev J. Design Principles of Length Control of Cytoskeletal Structures. *Annu Rev Biophys.* 2016;45: 85–116. doi:10.1146/annurev-biophys-070915-094206. PubMed PMID: 27145876
29. Fujiwara I, Vavylonis D, Pollard TD. Polymerization kinetics of ADP- and ADP-Pi-actin determined by fluorescence microscopy. *Proc Natl Acad Sci U S A.* 2007;104: 8827–8832. doi:10.1073/pnas.0702510104. PubMed PMID: 17517656

30. Wegner A. Head to tail polymerization of actin. *J Mol Biol.* 1976;108: 139–50. doi:10.1016/s0022-2836(76)80100-3. PubMed PMID: 1003481
31. Castaneda N, Park J, Kang EH. Regulation of Actin Bundle Mechanics and Structure by Intracellular Environmental Factors. *Front Phys.* 2021;9: 1–7. doi:10.3389/fphy.2021.675885
32. Dominguez R. A Common Binding Site for Actin-Binding Proteins on the Actin Surface. *Actin-Monomer-Binding Proteins.* New York, NY: Springer New York; 2007. pp. 107–115. doi:10.1007/978-0-387-46407-7\_10
33. Stachowiak MR, McCall PM, Thoresen T, Balcioglu HE, Kasiewicz L, Gardel ML, et al. Self-organization of myosin II in reconstituted actomyosin bundles. *Biophys J.* 2012;103: 1265–1274. doi:10.1016/j.bpj.2012.08.028. PubMed PMID: 22995499
34. Wang S, Wolynes PG. Active patterning and asymmetric transport in a model actomyosin network. *J Chem Phys.* 2013;139. doi:10.1063/1.4848657. PubMed PMID: 24359394
35. Verkhovskiy AB, Svitkina TM, Borisy GG. Myosin II filament assemblies in the active lamella of fibroblasts: Their morphogenesis and role in the formation of actin filament bundles. *J Cell Biol.* 1995;131: 989–1002. doi:10.1083/jcb.131.4.989. PubMed PMID: 7490299
36. Pizarro-Cerdá J, Chorev DS, Geiger B, Cossart P. The Diverse Family of Arp2/3 Complexes. *Trends Cell Biol.* 2017;27: 93–100. doi:10.1016/j.tcb.2016.08.001.

PubMed PMID: 27595492

37. Bendix PM, Koenderink GH, Cuvelier D, Dogic Z, Koeleman BN, Briehar WM, et al. A quantitative analysis of contractility in active cytoskeletal protein networks. *Biophys J*. 2008;94: 3126–3136. doi:10.1529/biophysj.107.117960. PubMed PMID: 18192374
38. Tan TH, Malik-Garbi M, Abu-Shah E, Li J, Sharma A, MacKintosh FC, et al. Self-organized stress patterns drive state transitions in actin cortices. *Sci Adv*. 2018;4: eaar2847. doi:10.1126/sciadv.aar2847. PubMed PMID: 29881775
39. Winkelman JD, Suarez C, Hocky GM, Harker AJ, Morgenthaler AN, Christensen JR, et al. Fascin- and  $\alpha$ -Actinin-Bundled Networks Contain Intrinsic Structural Features that Drive Protein Sorting. *Curr Biol*. 2016;26: 2697–2706. doi:10.1016/j.cub.2016.07.080. PubMed PMID: 27666967
40. Baudry M, Bi X, Gall C, Lynch G. The biochemistry of memory: The 26year journey of a “new and specific hypothesis”. *Neurobiol Learn Mem*. 2011;95: 125–33. doi:10.1016/j.nlm.2010.11.015. PubMed PMID: 21134478
41. Okamoto K, Bosch M, Hayashi Y. The Roles of CaMKII and F-Actin in the Structural Plasticity of Dendritic Spines: A Potential Molecular Identity of a Synaptic Tag? *Physiology*. 2009;24: 357–366. doi:10.1152/physiol.00029.2009. PubMed PMID: 19996366
42. Lømo T. Discovering long-term potentiation (LTP) - recollections and reflections on what came after. *Acta Physiol*. 2017. doi:10.1111/apha.12921.

PubMed PMID: 28719040

43. Spence EF, Kanak DJ, Carlson BR, Soderling SH. The Arp2/3 Complex Is Essential for Distinct Stages of Spine Synapse Maturation, Including Synapse Unsilencing. *J Neurosci*. 2016;36: 9696–709. doi:10.1523/JNEUROSCI.0876-16.2016. PubMed PMID: 27629719
44. Hodges JL, Vilchez SM, Asmussen H, Whitmore LA, Horwitz AR.  $\alpha$ -Actinin-2 mediates spine morphology and assembly of the post-synaptic density in hippocampal neurons. *PLoS One*. 2014;9. doi:10.1371/journal.pone.0101770
45. Ryu J, Liu L, Wong TP, Wu DC, Burette A, Weinberg R, et al. A critical role for myosin IIB in dendritic spine morphology and synaptic function. *Neuron*. 2006;49: 175–182. doi:10.1016/j.neuron.2005.12.017. PubMed PMID: 16423692
46. Hodges JL, Newell-Litwa K, Asmussen H, Vicente-Manzanares M, Horwitz AR. Myosin IIB activity and phosphorylation status determines dendritic spine and post-synaptic density morphology. *PLoS One*. 2011;6. doi:10.1371/journal.pone.0024149
47. Wang Q, Chen M, Schafer NP, Bueno C, Song SS, Hudmon A, et al. Assemblies of calcium/calmodulin-dependent kinase II with actin and their dynamic regulation by calmodulin in dendritic spines. *Proc Natl Acad Sci U S A*. 2019;116: 18937–18942. doi:10.1073/pnas.1911452116. PubMed PMID: 31455737

48. Gu X, Schafer NP, Wang Q, Song SS, Chen M, Waxham MN, et al. Exploring the F-actin/CPEB3 interaction and its possible role in the molecular mechanism of long-term memory. *Proc Natl Acad Sci U S A*. 2020;117: 22128–22134. doi:10.1073/pnas.2012964117. PubMed PMID: 32848053
49. Gu X, Schafer NP, Bueno C, Lu W, Wolynes PG. A structural dynamics model for how CPEB3 binding to SUMO2 can regulate translational control in dendritic spines. *Manuscr Submitt Publ*. 2022.
50. Sanabria H, Swulius MT, Kolodziej SJ, Liu J, Waxham MN.  $\beta$ CaMKII regulates actin assembly and structure. *J Biol Chem*. 2009;284: 9770–9780. doi:10.1074/jbc.M809518200. PubMed PMID: 19208632
51. Wang S, Wolynes PG. Active patterning and asymmetric transport in a model actomyosin network. *J Chem Phys*. 2013;139: 235103. doi:10.1063/1.4848657. PubMed PMID: 24359394
52. Gurmessa BJ, Bitten N, Nguyen DT, Saleh OA, Ross JL, Das M, et al. Triggered disassembly and reassembly of actin networks induces rigidity phase transitions. *Soft Matter*. 2019;15: 1335–1344. doi:10.1039/c8sm01912f
53. Bueno C, Liman J, Schafer NP, Cheung MS, Wolynes PG. A generalized Flory-Stockmayer kinetic theory of connectivity percolation and rigidity percolation of cytoskeletal networks. *PLoS Comput Biol*. 2022;18: e1010105. doi:10.1371/journal.pcbi.1010105. PubMed PMID: 35533192
54. Alvarado J, Sheinman M, Sharma A, MacKintosh FC, Koenderink GH. Force

- percolation of contractile active gels. *Soft Matter*. 2017;13: 5624–5644.  
doi:10.1039/c7sm00834a. PubMed PMID: 28812094
55. Popp D, Yamamoto A, Iwasa M, Maéda Y. Direct visualization of actin nematic network formation and dynamics. *Biochem Biophys Res Commun*. 2006;351: 348–353. doi:10.1016/j.bbrc.2006.10.041. PubMed PMID: 17067551
  56. Dalhaimer P, Discher DE, Lubensky TC. Crosslinked actin networks show liquid crystal elastomer behaviour, including soft-mode elasticity. *Nat Phys*. 2007;3: 354–360. doi:10.1038/nphys567
  57. Coppin CM, Leavis PC. Quantitation of liquid-crystalline ordering in F-actin solutions. *Biophys J*. 1992;63: 794–807. doi:10.1016/S0006-3495(92)81647-8. PubMed PMID: 1330036
  58. Wang S, Wolynes PG. Communication: Effective temperature and glassy dynamics of active matter. *J Chem Phys*. 2011;135. doi:10.1063/1.3624753
  59. Liman J, Bueno C, Eliaz Y, Schafer NP, Waxham MN, Wolynes PG, et al. The role of the Arp2/3 complex in shaping the dynamics and structures of branched actomyosin networks. *Proc Natl Acad Sci U S A*. 2020;117: 10825–10831. doi:10.1073/pnas.1922494117. PubMed PMID: 32354995
  60. Soares E Silva M, Alvarado J, Nguyen J, Georgoulia N, Mulder BM, Koenderink GH. Self-organized patterns of actin filaments in cell-sized confinement. *Soft Matter*. 2011;7: 10631–10641. doi:10.1039/c1sm06060k
  61. Freedman SL, Banerjee S, Hocky GM, Dinner AR. A Versatile Framework for

Simulating the Dynamic Mechanical Structure of Cytoskeletal Networks.

Biophys J. 2017;113: 448–460. doi:10.1016/j.bpj.2017.06.003. PubMed PMID: 28746855

62. Nedelec F, Foethke D. Collective Langevin dynamics of flexible cytoskeletal fibers. *New J Phys*. 2007;9. doi:10.1088/1367-2630/9/11/427
63. Ni Q, Papoian GA. Turnover versus treadmilling in actin network assembly and remodeling. *Cytoskeleton (Hoboken)*. 2019;76: 562–570. doi:10.1002/cm.21564. PubMed PMID: 31525282
64. Floyd C, Papoian GA, Jarzynski C. Quantifying dissipation in actomyosin networks. *Interface Focus*. 2019;9: 20180078. doi:10.1098/rsfs.2018.0078. PubMed PMID: 31065344
65. Komianos JE, Papoian GA. Stochastic Ratcheting on a Funneled Energy Landscape Is Necessary for Highly Efficient Contractility of Actomyosin Force Dipoles. *Phys Rev X*. 2018;8: 21006. doi:10.1103/PhysRevX.8.021006
66. Chandrasekaran A, Upadhyaya A, Papoian GA. Remarkable structural transformations of actin bundles are driven by their initial polarity, motor activity, crosslinking, and filament treadmilling. *PLoS Comput Biol*. 2019;15: e1007156. doi:10.1371/journal.pcbi.1007156. PubMed PMID: 31287817
67. Popov K, Komianos J, Papoian GA. MEDYAN: Mechanochemical Simulations of Contraction and Polarity Alignment in Actomyosin Networks. *PLoS Comput Biol*. 2016;12: e1004877. doi:10.1371/journal.pcbi.1004877. PubMed PMID:



27120189

68. Flory PJ. Molecular Size Distribution in Three Dimensional Polymers. I. Gelation. *J Am Chem Soc.* 1941;63: 3083–3090. doi:10.1021/ja01856a061
69. Stockmayer WH, Jacobson H. Gel formation in vinyl-divinyl copolymers. *J Chem Phys.* 1943;11: 393. doi:10.1063/1.1723863
70. Tavares JM, Teixeira PIC, Telo Da Gama MM. Percolation of colloids with distinct interaction sites. *Phys Rev E - Stat Nonlinear, Soft Matter Phys.* 2010;81: 1–4. doi:10.1103/PhysRevE.81.010501. PubMed PMID: 20365311
71. Tavares JM, Teixeira PIC, Telo Da Gama MM, Sciortino F. Equilibrium self-assembly of colloids with distinct interaction sites: Thermodynamics, percolation, and cluster distribution functions. *J Chem Phys.* 2010;132. doi:10.1063/1.3435346. PubMed PMID: 20572716
72. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. hPDI: A database of experimental human protein-DNA interactions. *Bioinformatics.* 2010;26: 287–289. doi:10.1093/bioinformatics/btp631. PubMed PMID: 19900953
73. Hu S, Xie Z, Onishi A, Yu X, Jiang L, Lin J, et al. Profiling the Human Protein-DNA Interactome Reveals ERK2 as a Transcriptional Repressor of Interferon Signaling. *Cell.* 2009;139: 610–622. doi:10.1016/j.cell.2009.08.037. PubMed PMID: 19879846
74. Pérez A, Luque FJ, Orozco M. Frontiers in molecular dynamics simulations of DNA. *Acc Chem Res.* 2012;45: 196–205. doi:10.1021/ar2001217. PubMed

PMID: 21830782

75. Winogradoff D, Aksimentiev A. Molecular Mechanism of Spontaneous Nucleosome Unraveling. *J Mol Biol.* 2019;431: 323–335.  
doi:10.1016/j.jmb.2018.11.013. PubMed PMID: 30468737
76. Moller J, de Pablo JJ. Bottom-Up Meets Top-Down: The Crossroads of Multiscale Chromatin Modeling. *Biophys J.* 2020;118: 2057–2065.  
doi:10.1016/j.bpj.2020.03.014. PubMed PMID: 32320675
77. Uusitalo JJ, Ingólfsson HI, Akhshi P, Tieleman DP, Marrink SJ. Martini Coarse-Grained Force Field: Extension to DNA. *J Chem Theory Comput.* 2015;11: 3932–3945. doi:10.1021/acs.jctc.5b00286. PubMed PMID: 26574472
78. Dans PD, Darré L, Machado MR, Zeida A, Brandner AF, Pantano S. Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level. *Adv Bioinforma Comput Biol.* 2013; 71–81. doi:10.1007/978-3-319-02624-4\_7
79. Dans PD, Zeida A, Machado MR, Pantano S. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J Chem Theory Comput.* 2010;6: 1711–1725. doi:10.1021/ct900653p. PubMed PMID: 277408500024
80. Machado MR, Pantano S. Exploring LacI-DNA Dynamics by Multiscale Simulations Using the SIRAH Force Field. *J Chem Theory Comput.* 2015;11: 5012–5023. doi:10.1021/acs.jctc.5b00575. PubMed PMID: 26574286

81. Paissoni C, Jussupow A, Camilloni C. Martini bead form factors for nucleic acids and their application in the refinement of protein–nucleic acid complexes against SAXS data. *J Appl Crystallogr.* 2019;52: 394–402. doi:10.1107/S1600576719002450
82. Snodin BEK, Randisi F, Mosayebi M, Šulc P, Schreck JS, Romano F, et al. Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. *J Chem Phys.* 2015;142. doi:10.1063/1.4921957. PubMed PMID: 26093573
83. Ouldridge TE, Louis AA, Doye JPK. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J Chem Phys.* 2011;134. doi:10.1063/1.3552946. PubMed PMID: 21361556
84. Procyk J, Poppleton E, Šulc P. Coarse-grained nucleic acid-protein model for hybrid nanotechnology. *Soft Matter.* 2021;17: 3586–3593. doi:10.1039/d0sm01639j. PubMed PMID: 33398312
85. Hinckley DM, Freeman GS, Whitmer JK, de Pablo JJ. An experimentally-informed coarse-grained 3-Site-Per-Nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization. *J Chem Phys.* 2013;139: 144903. doi:10.1063/1.4822042. PubMed PMID: 24116642
86. Freeman GS, Hinckley DM, Lequeieu JP, Whitmer JK, de Pablo JJ. Coarse-grained modeling of DNA curvature. *J Chem Phys.* 2014;141: 165103. doi:10.1063/1.4897649. PubMed PMID: 25362344

87. Tan C, Jung J, Kobayashi C, Torre DU La, Takada S, Sugita Y. Implementation of residue-level coarsegrained models in GENESIS for large-scale molecular dynamics simulations. *PLoS Comput Biol.* 2022;18: 1–30. doi:10.1371/journal.pcbi.1009578. PubMed PMID: 35381009
88. Brandani GB, Niina T, Tan C, Takada S. DNA sliding in nucleosomes via twist defect propagation revealed by molecular simulations. *Nucleic Acids Res.* 2018;46: 2788–2801. doi:10.1093/nar/gky158. PubMed PMID: 29506273
89. Lequieu J, Schwartz DC, De Pablo JJ. In silico evidence for sequence-dependent nucleosome sliding. *Proc Natl Acad Sci U S A.* 2017;114: E9197–E9205. doi:10.1073/pnas.1705685114. PubMed PMID: 29078285
90. Tan C, Terakawa T, Takada S. Dynamic Coupling among Protein Binding, Sliding, and DNA Bending Revealed by Molecular Dynamics. *J Am Chem Soc.* 2016;138: 8512–8522. doi:10.1021/jacs.6b03729. PubMed PMID: 27309278
91. Sakata R, Niwa K, Ugarte La Torre D, Gu C, Tahara E, Takada S, et al. Opening of cohesin's SMC ring is essential for timely DNA replication and DNA loop formation. *Cell Rep.* 2021;35: 108999. doi:10.1016/j.celrep.2021.108999. PubMed PMID: 33909997
92. Ugarte La Torre D, Takada S. Modeling lipid-protein interactions for coarse-grained lipid and C $\alpha$  protein models. *J Chem Phys.* 2021;155. doi:10.1063/5.0057278. PubMed PMID: 34686048
93. Ugarte La Torre D, Takada S. Coarse-grained implicit solvent lipid force field

- with a compatible resolution to the C $\alpha$  protein representation. *J Chem Phys.* 2020;153: 205101. doi:10.1063/5.0026342. PubMed PMID: 33261497
94. Zhang B, Zheng W, Papoian GA, Wolynes PG. Exploring the Free Energy Landscape of Nucleosomes. *J Am Chem Soc.* 2016;138: 8126–33. doi:10.1021/jacs.6b02893. PubMed PMID: 27300314
95. Krepel D, Davtyan A, Schafer NP, Wolynes PG, Onuchic JN. Braiding topology and the energy landscape of chromosome organization proteins. *Proc Natl Acad Sci U S A.* 2020;117: 1468–1477. doi:10.1073/pnas.1917750117. PubMed PMID: 31888987
96. Potoyan DA, Zheng W, Komives EA, Wolynes PG. Molecular stripping in the NF- $\kappa$ B/I $\kappa$ B/DNA genetic regulatory network. *Proc Natl Acad Sci U S A.* 2016;113: 110–115. doi:10.1073/pnas.1520483112. PubMed PMID: 26699500
97. Jin S, Bueno C, Lu W, Wang Q, Chen M, Chen X, et al. Computationally Exploring the Mechanism of Bacteriophage T7 Helicase gp4 Translocating along ssDNA. *Manuscr Submitt Publ.* 2022.
98. Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B.* 2012;116: 8494–503. doi:10.1021/jp212541y. PubMed PMID: 22545654
99. Sirovetz BJ, Schafer NP, Wolynes PG. Water Mediated Interactions and the

- Protein Folding Phase Diagram in the Temperature-Pressure Plane. *J Phys Chem B*. 2015;119: 11416–11427. doi:10.1021/acs.jpcc.5b03828. PubMed PMID: 26102155
100. Tsai M-Y, Zheng W, Balamurugan D, Schafer NP, Kim BL, Cheung MS, et al. Electrostatics, structure prediction, and the energy landscapes for protein folding and binding. *Protein Sci*. 2016;25: 255–69. doi:10.1002/pro.2751. PubMed PMID: 26183799
101. Truong HH, Kim BL, Schafer NP, Wolynes PG. Predictive energy landscapes for folding membrane protein assemblies. *J Chem Phys*. 2015;143: 243101. doi:10.1063/1.4929598. PubMed PMID: 26723586
102. Jin S, Chen M, Chen X, Bueno C, Lu W, Schafer NP, et al. Protein Structure Prediction in CASP13 Using AWSEM-Suite. *J Chem Theory Comput*. 2020;16: 3977–3988. doi:10.1021/acs.jctc.0c00188. PubMed PMID: 32396727
103. Jin S, Contessoto VG, Chen M, Schafer NP, Lu W, Chen X, et al. AWSEM-Suite: A protein structure prediction server based on template-guided, coevolutionary-enhanced optimized folding landscapes. *Nucleic Acids Res*. 2020;48: W25–W30. doi:10.1093/NAR/GKAA356. PubMed PMID: 32383764
104. Caruel M, Truskinovsky L. Physics of muscle contraction. *Reports Prog Phys*. 2018;81. doi:10.1088/1361-6633/aa7b9e. PubMed PMID: 28649969
105. Murrell MP, Oakes PW, Lenz M, Gardel ML. Forcing cells into shape: the mechanics of actomyosin contractility. *Nat Rev Mol Cell Biol*. 2015;16: 486–

98. doi:10.1038/nrm4012. PubMed PMID: 26130009
106. Wachsstock DH, Schwartz WH, Pollard TD. Affinity of alpha-actinin for actin determines the structure and mechanical properties of actin filament gels. *Biophys J.* 1993;65: 205–14. doi:10.1016/S0006-3495(93)81059-2. PubMed PMID: 8369430
107. Tempel M, Isenberg G, Sackmann E. Temperature-induced sol-gel transition and microgel formation in  $\alpha$ -actinin cross-linked actin networks: A rheological study. *Phys Rev E - Stat Physics, Plasmas, Fluids, Relat Interdiscip Top.* 1996;54: 1802–1810. doi:10.1103/PhysRevE.54.1802. PubMed PMID: 9965260
108. Köster DV, Husain K, Iljazi E, Bhat A, Bieling P, Mullins RD, et al. Actomyosin dynamics drive local membrane component organization in an in vitro active composite layer. *Proc Natl Acad Sci U S A.* 2016;113: E1645–E1654. doi:10.1073/pnas.1514030113. PubMed PMID: 26929326
109. Heinemann F, Vogel SK, Schwille P. Lateral membrane diffusion modulated by a minimal actin cortex. *Biophys J.* 2013;104: 1465–1475. doi:10.1016/j.bpj.2013.02.042. PubMed PMID: 23561523
110. Vogel SK, Heinemann F, Chwastek G, Schwille P. The design of MACs (minimal actin cortices). *Cytoskeleton.* 2013;70: 706–717. doi:10.1002/cm.21136. PubMed PMID: 24039068
111. Freedman SL, Suarez C, Winkelman JD, Kovar DR, Voth GA, Dinner AR, et al.

- Mechanical and kinetic factors drive sorting of F-actin cross-linkers on bundles. *Proc Natl Acad Sci U S A*. 2019;116: 16192–16197.  
doi:10.1073/pnas.1820814116. PubMed PMID: 31346091
112. Konietzny A, Bär J, Mikhaylova M. Dendritic Actin Cytoskeleton: Structure, Functions, and Regulations. *Front Cell Neurosci*. 2017;11: 1–10.  
doi:10.3389/fncel.2017.00147. PubMed PMID: 28572759
113. Sjöblom B, Salmazo A, Djinović-Carugo K.  $\alpha$ -Actinin structure and regulation. *Cell Mol Life Sci*. 2008;65: 2688–2701. doi:10.1007/s00018-008-8080-8.  
PubMed PMID: 18488141
114. Gardel ML, Shin JH, MacKintosh FC, Mahadevan L, Matsudaira P, Weitz DA. Elastic behavior of cross-linked and bundled actin networks. *Science (80- )*. 2004;304: 1301–1305. doi:10.1126/science.1095087. PubMed PMID: 15166374
115. Wagner B, Tharmann R, Haase I, Fischer M, Bausch AR. Cytoskeletal polymer networks: The molecular structure of cross-linkers determines macroscopic properties. *Proc Natl Acad Sci U S A*. 2006;103: 13974–13978.  
doi:10.1073/pnas.0510190103. PubMed PMID: 16963567
116. Mullins RD, Heuser JA, Pollard TD. The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proc Natl Acad Sci U S A*. 1998;95: 6181–6186. doi:10.1073/pnas.95.11.6181. PubMed PMID: 9600938



117. Alencar AM, Ferraz MSA, Park CY, Millet E, Trepas X, Fredberg JJ, et al. Non-equilibrium cytoquake dynamics in cytoskeletal remodeling and stabilization. *Soft Matter*. 2016;12: 8506–8511. doi:10.1039/c6sm01041e. PubMed PMID: 27722665
118. Kane RE. Interconversion of structural and contractile actin gels by insertion of myosin during assembly. *J Cell Biol*. 1983;97: 1745–1752. doi:10.1083/jcb.97.6.1745. PubMed PMID: 6227627
119. Linsmeier I, Banerjee S, Oakes PW, Jung W, Kim T, Murrell MP. Disordered actomyosin networks are sufficient to produce cooperative and telescopic contractility. *Nat Commun*. 2016;7: 12615. doi:10.1038/ncomms12615. PubMed PMID: 27558758
120. Mizuno D, Tardin C, Schmidt CF, Mackintosh FC. Nonequilibrium mechanics of active cytoskeletal networks. *Science*. 2007;315: 370–3. doi:10.1126/science.1134404. PubMed PMID: 17234946
121. Banerjee S, Liverpool TB, Marchetti MC. Generic phases of cross-linked active gels: Relaxation, oscillation and contractility. *EPL (Europhysics Lett)*. 2011;96: 58004. doi:10.1209/0295-5075/96/58004. PubMed PMID: 23003998
122. Alvarado J, Koenderink GH. Reconstituting cytoskeletal contraction events with biomimetic actin-myosin active gels. *Methods Cell Biol*. 2015;128: 83–103. doi:10.1016/bs.mcb.2015.02.001. PubMed PMID: 25997344
123. Murrell MP, Gardel ML. F-actin buckling coordinates contractility and

- severing in a biomimetic actomyosin cortex. *Proc Natl Acad Sci U S A*. 2012;109: 20820–20825. doi:10.1073/pnas.1214753109. PubMed PMID: 23213249
124. Soares E Silva M, Depken M, Stuhmann B, Korsten M, MacKintosh FC, Koenderink GH. Active multistage coarsening of actin networks driven by myosin motors. *Proc Natl Acad Sci U S A*. 2011;108: 9408–9413. doi:10.1073/pnas.1016616108. PubMed PMID: 21593409
125. Wang S, Wolynes PG. Active contractility in actomyosin networks. *Proc Natl Acad Sci*. 2012;109: 6446–6451. doi:10.1073/pnas.1204205109. PubMed PMID: 22493220
126. Lenz M, Thoresen T, Gardel ML, Dinner AR. Contractile units in disordered actomyosin bundles arise from f-actin buckling. *Phys Rev Lett*. 2012;108: 1–5. doi:10.1103/PhysRevLett.108.238107
127. Thoresen T, Lenz M, Gardel ML. Reconstitution of contractile actomyosin bundles. *Biophys J*. 2011;100: 2698–2705. doi:10.1016/j.bpj.2011.04.031. PubMed PMID: 21641315
128. Oelz DB, Rubinstein BY, Mogilner A. A Combination of Actin Treadmilling and Cross-Linking Drives Contraction of Random Actomyosin Arrays. *Biophys J*. 2015;109: 1818–1829. doi:10.1016/j.bpj.2015.09.013. PubMed PMID: 26536259
129. Reymann A-C, Boujemaa-Paterski R, Martiel J-L, Guérin C, Cao W, Chin HF, et

- al. Actin network architecture can determine myosin motor activity. *Science*. 2012;336: 1310–4. doi:10.1126/science.1221708. PubMed PMID: 22679097
130. Shen T, Wolynes PG. Statistical mechanics of a cat's cradle. *New J Phys*. 2006;8. doi:10.1088/1367-2630/8/11/273
131. Wang S, Shen T, Wolynes PG. The interplay of nonlinearity and architecture in equilibrium cytoskeletal mechanics. *J Chem Phys*. 2011;134: 014510. doi:10.1063/1.3518450. PubMed PMID: 21219010
132. Wang S, Wolynes PG. On the spontaneous collective motion of active matter. *Proc Natl Acad Sci U S A*. 2011;108: 15184–9. doi:10.1073/pnas.1112034108. PubMed PMID: 21876141
133. Wang S, Wolynes PG. Tensegrity and motor-driven effective interactions in a model cytoskeleton. *J Chem Phys*. 2012;136. doi:10.1063/1.3702583. PubMed PMID: 22502548
134. Wang S, Wolynes PG. Microscopic theory of the glassy dynamics of passive and active network materials. *J Chem Phys*. 2013;138. doi:10.1063/1.4773349. PubMed PMID: 23556772
135. Hall RW, Wolynes PG. Microscopic theory of network glasses. *Phys Rev Lett*. 2003;90: 085505. doi:10.1103/PhysRevLett.90.085505. PubMed PMID: 12633439
136. Zilman AG, Safran SA. Role of cross-links in bundle formation, phase separation and gelation of long filaments. *Europhys Lett*. 2003;63: 139–145.

doi:10.1209/epl/i2003-00489-5

137. Stockmayer WH. Theory of molecular size distribution and gel formation in branched-chain polymers. *J Chem Phys.* 1943;11: 45–55.  
doi:10.1063/1.1723803
138. Thorpe MF, Stinchcombe RB. Two exactly soluble models of rigidity percolation. *Philos Trans R Soc A Math Phys Eng Sci.* 2014;372.  
doi:10.1098/rsta.2012.0038. PubMed PMID: 24379428
139. Bresser W, Boolchand P, Suranyi P. Rigidity Percolation and Molecular Clustering in Network Glasses. *Phys Rev Lett.* 1986;56: 2493–2496.  
doi:10.1103/PhysRevLett.56.2493. PubMed PMID: 10033007
140. Moukarzel C, Duxbury PM. Comparison of rigidity and connectivity percolation in two dimensions. *Phys Rev E - Stat Physics, Plasmas, Fluids, Relat Interdiscip Top.* 1999;59: 2614–2622. doi:10.1103/PhysRevE.59.2614
141. Kim T, Hwang W, Kamm RD. Computational analysis of a cross-linked actin-like network. *Exp Mech.* 2009;49: 91–104. doi:10.1007/s11340-007-9091-3
142. Müller KW, Meier C, Wall WA. Resolution of sub-element length scales in Brownian dynamics simulations of biopolymer networks with geometrically exact beam finite elements. *J Comput Phys.* 2015;303: 185–202.  
doi:10.1016/j.jcp.2015.09.038
143. Falzone TT, Lenz M, Kovar DR, Gardel ML. Assembly kinetics determine the architecture of  $\alpha$ -actinin crosslinked F-actin networks. *Nat Commun.* 2012;3:

861. doi:10.1038/ncomms1862. PubMed PMID: 22643888
144. Claessens MMAE, Semmrich C, Ramos L, Bausch AR. Helical twist controls the thickness of F-actin bundles. *Proc Natl Acad Sci.* 2008;105: 8819–8822. doi:10.1073/pnas.0711149105. PubMed PMID: 18579789
145. Yang S, Huang FK, Huang J, Chen S, Jakoncic J, Leo-Macias A, et al. Molecular mechanism of fascin function in filopodial formation. *J Biol Chem.* 2013;288: 274–284. doi:10.1074/jbc.M112.427971. PubMed PMID: 23184945
146. Hampton CM, Taylor DW, Taylor KA. Novel Structures for  $\alpha$ -Actinin:F-Actin Interactions and their Implications for Actin-Membrane Attachment and Tension Sensing in the Cytoskeleton. *J Mol Biol.* 2007;368: 92–104. doi:10.1016/j.jmb.2007.01.071. PubMed PMID: 17331538
147. Duxbury PM, Jacobs DJ, Thorpe MF, Moukarzel C. Floppy modes and the free energy: Rigidity and connectivity percolation on Bethe lattices. *Phys Rev E - Stat Physics, Plasmas, Fluids, Relat Interdiscip Top.* 1999;59: 2084–2092. doi:10.1103/PhysRevE.59.2084. PubMed PMID: 28363908
148. Garboczi, Thorpe. Effective-medium theory of percolation on central-force elastic networks. II. Further results. *Phys Rev B Condens Matter.* 1985;31: 7276–7281. doi:10.1103/physrevb.31.7276. PubMed PMID: 9935648
149. Jacobs DJ, Thorpe MF. Generic rigidity percolation: The pebble game. *Phys Rev Lett.* 1995;75: 4051–4054. doi:10.1103/PhysRevLett.75.4051. PubMed PMID: 9964678

150. Garboczi, Thorpe. Effective-medium theory of percolation on central-force elastic networks. III. The superelastic problem. *Phys Rev B Condens Matter*. 1986;33: 3289–3294. doi:10.1103/physrevb.33.3289. PubMed PMID: 9938708
151. Thorpe MF. Rigidity percolation in glassy structures. *J Non Cryst Solids*. 1985;76: 109–116. doi:10.1016/0022-3093(85)90056-0
152. Chubynsky M V., Thorpe MF. Algorithms for three-dimensional rigidity analysis and a first-order percolation transition. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2007;76: 041135. doi:10.1103/PhysRevE.76.041135. PubMed PMID: 17994964
153. Moukarzel C, Duxbury PM, Leath PL. Infinite-Cluster Geometry in Central-Force Networks. *Phys Rev Lett*. 1997;78: 1480–1483. doi:10.1103/PhysRevLett.78.1480
154. Cingolani LA, Goda Y. Actin in action: The interplay between the actin cytoskeleton and synaptic efficacy. *Nat Rev Neurosci*. 2008;9: 344–356. doi:10.1038/nrn2373. PubMed PMID: 18425089
155. Janson LW, Kolega J, Taylor DL. Modulation of contraction by gelation/solution in a reconstituted motile model. *J Cell Biol*. 1991;114: 1005–1015. doi:10.1083/jcb.114.5.1005. PubMed PMID: 1651941
156. Honkura N, Matsuzaki M, Noguchi J, Ellis-Davies GCR, Kasai H. The subspine organization of actin fibers regulates the structure and plasticity of dendritic

- spines. *Neuron*. 2008;57: 719–729. doi:10.1016/j.neuron.2008.01.013.  
PubMed PMID: 18341992
157. Smith BA, Daugherty-Clarke K, Goode BL, Gelles J. Pathway of actin filament branch formation by Arp2/3 complex revealed by single-molecule imaging. *Proc Natl Acad Sci U S A*. 2013;110: 1285–1290.  
doi:10.1073/pnas.1211164110. PubMed PMID: 23292935
158. Luo W, Yu C han, Lieu ZZ, Allard J, Mogilner A, Sheetz MP, et al. Analysis of the local organization and dynamics of cellular actin networks. *J Cell Biol*. 2013;202: 1057–1073. doi:10.1083/jcb.201210123. PubMed PMID: 24081490
159. Svitkina TM, Borisy GG. Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia. *J Cell Biol*. 1999;145: 1009–26.  
doi:10.1083/jcb.145.5.1009. PubMed PMID: 10352018
160. Kovar DR, Harris ES, Mahaffy R, Higgs HN, Pollard TD. Control of the assembly of ATP- and ADP-actin by formins and profilin. *Cell*. United States; 2006. pp. 423–435. doi:10.1016/j.cell.2005.11.038. PubMed PMID: 16439214
161. Zalevsky J, Lempert L, Kranitz H, Mullins RD. Different WASP family proteins stimulate different Arp2/3 complex-dependent actin-nucleating activities. *Curr Biol*. 2001;11: 1903–1913. doi:10.1016/s0960-9822(01)00603-0.  
PubMed PMID: 11747816

162. Lombardo AT, Nelson SR, Kennedy GG, Trybus KM, Walcott S, Warshaw DM. Myosin Va transport of liposomes in three-dimensional actin networks is modulated by actin filament density, position, and polarity. *Proc Natl Acad Sci U S A*. 2019;116: 8326–8335. doi:10.1073/pnas.1901176116. PubMed PMID: 30967504
163. Hu L, Papoian GA. Mechano-chemical feedbacks regulate actin mesh growth in lamellipodial protrusions. *Biophys J*. 2010;98: 1375–1384. doi:10.1016/j.bpj.2009.11.054. PubMed PMID: 20409456
164. Hu L, Papoian GA. Molecular transport modulates the adaptive response of branched actin networks to an external force. *J Phys Chem B*. 2013;117: 13388–13396. doi:10.1021/jp405179e. PubMed PMID: 23962335
165. Lan Y, Papoian GA. The stochastic dynamics of filopodial growth. *Biophys J*. 2008;94: 3839–3852. doi:10.1529/biophysj.107.123778. PubMed PMID: 18234810
166. Hu L, Papoian GA. How does the antagonism between capping and anti-capping proteins affect actin network dynamics? *J Phys Condens Matter*. 2011;23: 374101. doi:10.1088/0953-8984/23/37/374101. PubMed PMID: 21862844
167. Åström JA, Kumar PBS, Vattulainen I, Karttunen M. Strain hardening, avalanches, and strain softening in dense cross-linked actin networks. *Phys Rev E*. 2008;77: 051913. doi:10.1103/PhysRevE.77.051913



168. Feng J, Levine H, Mao X, Sander LM. Nonlinear elasticity of disordered fiber networks. *Soft Matter*. 2016;12: 1419–1424. doi:10.1039/c5sm01856k. PubMed PMID: 26616428
169. Kim J, Feng J, Jones CAR, Mao X, Sander LM, Levine H, et al. Stress-induced plasticity of dynamic collagen networks. *Nat Commun*. 2017;8: 842. doi:10.1038/s41467-017-01011-7. PubMed PMID: 29018207
170. Lubchenko V, Wolynes PG. Aging, Jamming, and the Limits of Stability of Amorphous Solids. *J Phys Chem B*. 2018;122: 3280–3295. doi:10.1021/acs.jpcc.7b09553. PubMed PMID: 29216433
171. Lubchenko V, Wolynes PG. Theory of Structural Glasses and Supercooled Liquids. 2006; 235–268. doi:10.1146/annurev.physchem.58.032806.104653. PubMed PMID: 17067282
172. Dima RI, Thirumalai D. Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B*. 2004;108: 6564–6570. doi:10.1021/jp037128y
173. Shi Y, Porter CL, Crocker JC, Reich DH. Dissecting fat-tailed fluctuations in the cytoskeleton with active micropost arrays. *Proc Natl Acad Sci U S A*. 2019;116: 13839–13846. doi:10.1073/pnas.1900963116. PubMed PMID: 31239336
174. Lootens D, Van Damme H, Hébraud P. Giant stress fluctuations at the jamming transition. *Phys Rev Lett*. 2003;90: 178301.

- doi:10.1103/PhysRevLett.90.178301. PubMed PMID: 12786109
175. Lu W, Bueno C, Schafer NP, Moller J, Jin S, Chen X, et al. OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *PLoS Comput Biol.* 2021;17: 1–21.  
doi:10.1371/JOURNAL.PCBI.1008308. PubMed PMID: 33577557
176. Suomivuori C-M, Latorraca NR, Wingler LM, Eismann S, King MC, Kleinhenz ALW, et al. Molecular mechanism of biased signaling in a prototypical G protein-coupled receptor. *Science.* 2020;367: 881–887.  
doi:10.1126/science.aaz0326. PubMed PMID: 32079767
177. Kauzmann W. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry.* Elsevier; 1959. pp. 1–63.
178. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Water in protein structure prediction. *Proc Natl Acad Sci U S A.* 2004;101: 3352–3357. doi:10.1073/pnas.0307851100. PubMed PMID: 14988499
179. Papoian GA, Ulander J, Wolynes PG. Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc.* 2003;125: 9170–9178. doi:10.1021/ja034729u. PubMed PMID: 15369374
180. Schafer NP, Kim BL, Zheng W, Wolynes PG. Learning To Fold Proteins Using Energy Landscape Theory. *Isr J Chem.* 2014;54: 1311–1337.  
doi:10.1002/ijch.201300145. PubMed PMID: 25308991
181. Papoian GA. Coarse-grained modeling of biomolecules. CRC Press; 2017.

182. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J Comput Phys.* 1995;117: 1–19. doi:10.1006/jcph.1995.1039
183. Kim BL, Schafer NP, Wolynes PG. Predictive energy landscapes for folding  $\alpha$ -helical transmembrane proteins. *Proc Natl Acad Sci U S A.* 2014;111: 11031–11036. doi:10.1073/pnas.1410529111. PubMed PMID: 25030446
184. Zheng W, Schafer NP, Davtyan A, Papoian GA, Wolynes PG. Predictive energy landscapes for protein-protein association. *Proc Natl Acad Sci U S A.* 2012;109: 19244–9. doi:10.1073/pnas.1216215109. PubMed PMID: 23129648
185. Chen M, Tsai M, Zheng W, Wolynes PG. The Aggregation Free Energy Landscapes of Polyglutamine Repeats. *J Am Chem Soc.* 2016;138: 15197–15203. doi:10.1021/jacs.6b08665. PubMed PMID: 27786478
186. Sirovetz BJ, Schafer NP, Wolynes PG. Protein Structure Prediction: Making AWSEM AWSEM-ER by Adding Evolutionary Restraints. *Proteins.* 2017; 1–33. doi:10.1002/prot.25367. PubMed PMID: 28799172
187. Chen M, Lin X, Lu W, Schafer NP, Onuchic JN, Wolynes PG. Template-Guided Protein Structure Prediction and Refinement Using Optimized Folding Landscape Force Fields. *J Chem Theory Comput.* 2018;14: 6102–6116. doi:10.1021/acs.jctc.8b00683. PubMed PMID: 30240202
188. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular

dynamics. PLoS Comput Biol. 2017;13: 1–17.

doi:10.1371/journal.pcbi.1005659. PubMed PMID: 28746339

189. HABER E, ANFINSEN CB. Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. J Biol Chem. 1962;237: 1839–1844. .  
PubMed PMID: 13903380
190. Bai X-C, Yan C, Yang G, Lu P, Ma D, Sun L, et al. An atomic structure of human  $\gamma$ -secretase. Nature. 2015;525: 212–217. doi:10.1038/nature14892. PubMed PMID: 26280335
191. Lu XJ, Olson WK. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003;31: 5108–5121. doi:10.1093/nar/gkg680. PubMed PMID: 12930962
192. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc Natl Acad Sci U S A. 1998;95: 2932–7. doi:10.1073/pnas.95.6.2932. PubMed PMID: 9501193
193. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. Protein Sci. 1996;5: 1043–1059. doi:10.1002/pro.5560050607. PubMed PMID: 8762136
194. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl

- Acad Sci U S A. 1992;89: 9029–9033. doi:10.1073/pnas.89.19.9029. PubMed PMID: 1409599
195. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci U S A*. 1992;89: 4918–4922. doi:10.1073/pnas.89.11.4918. PubMed PMID: 1594594
196. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*. 2012;40: D370-6. doi:10.1093/nar/gkr703. PubMed PMID: 21890895
197. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol*. 1996;3: 842–848. doi:10.1038/nsb1096-842. PubMed PMID: 8836100
198. Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*. 1996;35: 5109–5124. doi:10.1021/bi9600153. PubMed PMID: 8611495
199. White SH, Wimley WC. Hydrophobic interactions of peptides with membrane interfaces. *Biochim Biophys Acta*. 1998;1376: 339–352. doi:10.1016/s0304-4157(98)00021-5. PubMed PMID: 9804985
200. White SH, Wimley WC. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct*. 1999;28: 319–365. doi:10.1146/annurev.biophys.28.1.319. PubMed PMID: 10410805

201. Lu W, Schafer NP, Wolynes PG. Energy landscape underlying spontaneous insertion and folding of an alpha-helical transmembrane protein into a bilayer. *Nat Commun.* 2018;9. doi:10.1038/s41467-018-07320-9. PubMed PMID: 30470737
202. Popot JL, Engelman DM. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry.* 1990;29: 4031–4037. doi:10.1021/bi00469a001. PubMed PMID: 1694455
203. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, et al. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature.* 2005;433: 377–381. doi:10.1038/nature03216. PubMed PMID: 15674282
204. Wang Q, Ni C, Li Z, Li X, Han R, Zhao F, et al. Efficient and accurate prediction of transmembrane topology from amino acid sequence only. 2019. doi:10.1101/627307
205. Klink TA, Woycechowsky KJ, Taylor KM, Raines RT. Contribution of disulfide bonds to the conformational stability and catalytic activity of ribonuclease A. *Eur J Biochem.* 2000;267: 566–572. doi:10.1046/j.1432-1327.2000.01037.x. PubMed PMID: 10632727
206. Cheek S, Krishna SS, Grishin N V. Structural classification of small, disulfide-rich protein domains. *J Mol Biol.* 2006;359: 215–237. doi:10.1016/j.jmb.2006.03.017. PubMed PMID: 16618491

207. Papaconstantinou ME, Bah A, Di Cera E. Role of the A chain in thrombin function. *Cell Mol Life Sci.* 2008;65: 1943–1947. doi:10.1007/s00018-008-8179-y. PubMed PMID: 18470478
208. Kubelka J, Hofrichter J, Eaton WA. The protein folding “speed limit”. *Curr Opin Struct Biol.* 2004;14: 76–88. doi:10.1016/j.sbi.2004.01.013. PubMed PMID: 15102453
209. Potoyan DA, Bueno C, Zheng W, Komives EA, Wolynes PG. Resolving the NFκB Heterodimer Binding Paradox: Strain and Frustration Guide the Binding of Dimeric Transcription Factors. *J Am Chem Soc.* 2017;139: 18558–18566. doi:10.1021/jacs.7b08741. PubMed PMID: 29183131
210. Ptashne M. *A genetic switch: Gene control and phage lambda.* 1986.
211. Ptashne M, Gann A. *Genes & signals.* Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.; 2002.
212. Ghim C-M, Almaas E. Genetic noise control via protein oligomerization. *BMC Syst Biol.* 2008;2: 94. doi:10.1186/1752-0509-2-94. PubMed PMID: 18980697
213. Burger A, Walczak AM, Wolynes PG. Abduction and asylum in the lives of transcription factors. *Proc Natl Acad Sci U S A.* 2010;107: 4016–4021. doi:10.1073/pnas.0915138107. PubMed PMID: 20160109
214. Bundschuh R, Hayot F, Jayaprakash C. The role of dimerization in noise reduction of simple genetic networks. *J Theor Biol.* 2003;220: 261–269.

doi:10.1006/jtbi.2003.3164. PubMed PMID: 12468297

215. Potoyan DA, Wolynes PG. On the dephasing of genetic oscillators. *Proc Natl Acad Sci U S A*. 2014;111: 2391–2396. doi:10.1073/pnas.1323433111. PubMed PMID: 24469814
216. Potoyan DA, Wolynes PG. Stochastic dynamics of genetic broadcasting networks. *Phys Rev E*. 2017;96: 52305. doi:10.1103/PhysRevE.96.052305. PubMed PMID: 29347696
217. Potoyan DA, Zheng W, Ferreira DU, Wolynes PG, Komives EA. PEST Control of Molecular Stripping of NF $\kappa$ B from DNA Transcription Sites. *J Phys Chem B*. 2016;120: 8532–8538. doi:10.1021/acs.jpcc.6b02359. PubMed PMID: 27098223
218. Phelps CB, Sengchanthalangsy LL, Malek S, Ghosh G. Mechanism of kappa B DNA binding by Rel/NF-kappa B dimers. *J Biol Chem*. 2000;275: 24392–24399. doi:10.1074/jbc.M003784200. PubMed PMID: 10825175
219. Rackov G, Hernández-Jiménez E, Shokri R, Carmona-Rodríguez L, Mañes S, Álvarez-Mon M, et al. p21 mediates macrophage reprogramming through regulation of p50-p50 NF- $\kappa$ B and IFN- $\beta$ . *J Clin Invest*. 2016;126: 3089–3103. doi:10.1172/JCI83404. PubMed PMID: 27427981
220. Hoffmann A, Natoli G, Ghosh G. Transcriptional regulation via the NF-kappaB signaling module. *Oncogene*. 2006;25: 6706–6716. doi:10.1038/sj.onc.1209933. PubMed PMID: 17072323



221. Siggers T, Chang AB, Teixeira A, Wong D, Williams KJ, Ahmed B, et al. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF- $\kappa$ B family DNA binding. *Nat Immunol.* 2011;13: 95–102. doi:10.1038/ni.2151. PubMed PMID: 22101729
222. Dembinski HE, Wismer K, Vargas JD, Suryawanshi GW, Kern N, Kroon G, et al. Functional importance of stripping in NF $\kappa$ B signaling revealed by a stripping-impaired I $\kappa$ B $\alpha$  mutant. *Proc Natl Acad Sci U S A.* 2017;114: 1916–1921. doi:10.1073/pnas.1610192114. PubMed PMID: 28167786
223. Fagerlund R, Behar M, Fortmann KT, Lin YE, Vargas JD, Hoffmann A. Anatomy of a negative feedback loop: the case of I $\kappa$ B $\alpha$ . *J R Soc Interface.* 2015;12: 262. doi:10.1098/rsif.2015.0262. PubMed PMID: 26311312
224. Bergqvist S, Alverdi V, Mengel B, Hoffmann A, Ghosh G, Komives EA. Kinetic enhancement of NF-kappaBxDNA dissociation by IkappaBalpha. *Proc Natl Acad Sci U S A.* 2009;106: 19328–19333. doi:10.1073/pnas.0908797106. PubMed PMID: 19887633
225. Onuchic JN, Wolynes PG. Theory of protein folding. *Curr Opin Struct Biol.* 2004;14: 70–75. doi:10.1016/j.sbi.2004.01.009. PubMed PMID: 15102452
226. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci.* 1987;84: 7524–7528. doi:10.1073/pnas.84.21.7524. PubMed PMID: 3478708
227. Ferreira DU, Hegler J a, Komives E a, Wolynes PG. On the role of frustration in

- the energy landscapes of allosteric proteins. *Proc Natl Acad Sci U S A*. 2011;108: 3499–3503. doi:10.1073/pnas.1018980108. PubMed PMID: 21273505
228. Stein DL. *Spin glasses and biology*. World Scientific; 1992.
229. Zhuravlev PI, Papoian GA. Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Q Rev Biophys*. 2010;43: 295–332. doi:10.1017/S0033583510000119. PubMed PMID: 20819242
230. Parra RG, Schafer NP, Radusky LG, Tsai M-Y, Guzovsky AB, Wolynes PG, et al. Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res*. 2016;44: W356-60. doi:10.1093/nar/gkw304. PubMed PMID: 27131359
231. Kitao A, Takemura K. High anisotropy and frustration: the keys to regulating protein function efficiently in crowded environments. *Curr Opin Struct Biol*. 2017;42: 50–58. doi:10.1016/j.sbi.2016.10.014. PubMed PMID: 27815984
232. Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci U S A*. 2003;100: 12570–5. doi:10.1073/pnas.2135471100. PubMed PMID: 14566052
233. Ansari A, Berendzen J, Bowne SF, Frauenfelder H, Iben IE, Sauke TB, et al. Protein states and proteinquakes. *Proc Natl Acad Sci U S A*. 1985;82: 5000–

5004. doi:10.1073/pnas.82.15.5000. PubMed PMID: 3860839
234. Whitford PC, Miyashita O, Levy Y, Onuchic JN. Conformational transitions of adenylate kinase: switching by cracking. *J Mol Biol.* 2007;366: 1661–1671. doi:10.1016/j.jmb.2006.11.085. PubMed PMID: 17217965
235. Whitford PC. Disorder guides protein function. *Proc Natl Acad Sci U S A.* 2013;110: 7114–7115. doi:10.1073/pnas.1305236110. PubMed PMID: 23610426
236. Shan Y, Arkhipov A, Kim ET, Pan AC, Shaw DE. Transitions to catalytically inactive conformations in EGFR kinase. *Proc Natl Acad Sci U S A.* 2013;110: 7270–7275. doi:10.1073/pnas.1220843110. PubMed PMID: 23576739
237. Li W, Wolynes PG, Takada S. Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. *Proc Natl Acad Sci U S A.* 2011;108: 3504–3509. doi:10.1073/pnas.1018983108. PubMed PMID: 21307307
238. Okazaki K, Koga N, Takada S, Onuchic JN, Wolynes PG. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc Natl Acad Sci U S A.* 2006;103: 11844–11849. doi:10.1073/pnas.0604375103. PubMed PMID: 16877541
239. Okazaki K-I, Takada S. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift

- mechanisms. *Proc Natl Acad Sci U S A*. 2008;105: 11182–11187.  
doi:10.1073/pnas.0802524105. PubMed PMID: 18678900
240. Ikeguchi M, Ueno J, Sato M, Kidera A. Protein structural change upon ligand binding: linear response theory. *Phys Rev Lett*. 2005;94: 78102.  
doi:10.1103/PhysRevLett.94.078102. PubMed PMID: 15783858
241. Lätzer J, Shen T, Wolynes PG. Conformational switching upon phosphorylation: A predictive framework based on energy landscape principles. *Biochemistry*. 2008;47: 2110–2122. doi:10.1021/bi701350v.  
PubMed PMID: 18198897
242. Lequieu J, Córdoba A, Schwartz DC, de Pablo JJ. Tension-Dependent Free Energies of Nucleosome Unwrapping. *ACS Cent Sci*. 2016;2: 660–666.  
doi:10.1021/acscentsci.6b00201. PubMed PMID: 27725965
243. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, et al. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A*. 2003;100: 12247–12252.  
doi:10.1073/pnas.2135255100. PubMed PMID: 14527995
244. Ernst M, Sittel F, Stock G. Contact- and distance-based principal component analysis of protein dynamics. *J Chem Phys*. 2015;143: 244114.  
doi:10.1063/1.4938249. PubMed PMID: 26723658
245. Ernst M, Wolf S, Stock G. Identification and Validation of Reaction Coordinates Describing Protein Functional Motion: Hierarchical Dynamics of T4 Lysozyme.

- J Chem Theory Comput. 2017;13: 5076–5088. doi:10.1021/acs.jctc.7b00571
246. Amadei A, Linssen AB, Berendsen HJ. Essential dynamics of proteins. *Proteins*. 1993;17: 412–425. doi:10.1002/prot.340170408. PubMed PMID: 8108382
247. Shen T, Zong C, Hamelberg D, McCammon JA, Wolynes PG. The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *FASEB J Off Publ Fed Am Soc Exp Biol*. 2005;19: 1389–1395. doi:10.1096/fj.04-3590hyp. PubMed PMID: 16126906
248. Wolynes PG. Symmetry and the energy landscapes of biomolecules. *Proc Natl Acad Sci U S A*. 1996;93: 14249–14255. doi:10.1073/pnas.93.25.14249. PubMed PMID: 8962034
249. Blundell TL, Srinivasan N. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc Natl Acad Sci U S A*. 1996;93: 14243–14248. doi:10.1073/pnas.93.25.14243. PubMed PMID: 8962033
250. Plaxco KW, Gross M. Protein complexes: the evolution of symmetry. *Curr Biol*. 2009;19: R25-6. doi:10.1016/j.cub.2008.11.004. PubMed PMID: 19138586
251. Bergendahl LT, Marsh JA. Functional determinants of protein assembly into homomeric complexes. *Sci Rep*. 2017;7: 4932. doi:10.1038/s41598-017-05084-8. PubMed PMID: 28694495

# Appendix A

Supplementary information for

**A generalized Flory-Stockmayer kinetic theory of connectivity percolation and rigidity percolation of branched cytoskeletal networks**

Carlos Bueno, James Liman, Nicholas P. Schafer, Margaret S. Cheung, and Peter G. Wolynes

## I. Methods

### A. Mass action chemical kinetics model

We quantified the concentration of the F-actin interfaces and their states (bound or unbound) using a mass action chemical kinetics model. Each F-actin monomer has 3 interfaces: the plus site ( $F_p$ ), the minus site ( $F_m$ ), and the binding site ( $F_c$ ). The change of the concentration of plus sites bound to minus sites over time ( $\frac{d[F_p \cdot F_m]}{dt}$ ) is given by the rate of polymerization at both ends, from which is subtracted the rate of depolymerization at both ends (Equation A). The polymerization rate on the plus end is proportional to the G-actin concentration ( $[G]$ ), the concentration of unbound plus sites ( $[F_p]$ ), and the rate constant for the polymerization on the plus end ( $k_p^+$ ). The polymerization rate on the minus end is also proportional to the G-actin concentration ( $[G]$ ), the concentration of unbound minus sites ( $[F_m]$ ), and the rate constant for the polymerization on the minus end ( $k_m^+$ ). The depolymerization rates do not depend on the G-actin concentration ( $[G]$ ), only on the concentration of unbound plus or minus sites and on the respective rate constant ( $k_p^-$ ,  $k_m^-$ ). Free G-actin monomers ( $[G]$ ) are consumed during the polymerization, so the change of G-actin concentration over time is the negative of the change of the concentration of plus sites bound to minus sites over time.

$$\frac{d[F_p \cdot F_m]}{dt} = -\frac{d[G]}{dt} = k_p^+[F_p][G] + k_m^+[F_m][G] - k_p^-[F_p] - k_m^-[F_m] \quad \text{Equation A}$$

The change of the concentration of unbound plus sites over time ( $\frac{d[F_p]}{dt}$ ) is given by the rate of creation of new filaments by branching, from which is subtracted the rate of destruction by unbranching (Equation B). The rate of creation of new filaments by branching is proportional to the concentration of unbound binding sites ( $[F_c]$ ), the concentration of unbound minus sites ( $[F_m]$ ), the concentration of unbound branchers ( $[B]$ ), and the brancher binding rate constant ( $k_B^+$ ). The rate of destruction of filaments by unbranching is proportional to the concentration of bound branchers ( $[F_c \cdot B \cdot F_m]$ ), and the brancher unbinding rate constant ( $k_B^-$ ).



$$\frac{d[F_p]}{dt} = k_B^+[F_c][G][B] - k_B^-[F_c \cdot B \cdot F_m] \quad \text{Equation B}$$

On the other hand, since the branching reaction does not create new unbound minus sites ( $F_m$ ), the concentration of unbound minus sites ( $[F_m]$ ) is constant over time (Equation C). We did not include the actin filament nucleation or destruction processes in the chemical kinetics model.

$$\frac{d[F_m]}{dt} = 0 \quad \text{Equation C}$$

The change of the concentration of bound branchers over time ( $\frac{d[F_c \cdot B \cdot F_m]}{dt}$ ) is proportional to the creation rate of filaments by branching and is also the negative of the change in brancher concentration over time ( $\frac{d[B]}{dt}$ ) (Equation D).

$$\frac{d[F_c \cdot B \cdot F_m]}{dt} = -\frac{d[B]}{dt} = k_B^+[F_c][G][B] - k_B^-[F_c \cdot B \cdot F_m] \quad \text{Equation D}$$

Unless stated otherwise, the linker binding reaction occurs in a single step and the change in unbound linker concentrations over time ( $\frac{d[L]}{dt}$ ) is given by the linker unbinding rate, from which is subtracted the linker binding rate. The linker unbinding rate is proportional to the concentration of bound linkers ( $[c \cdot L \cdot c]$ ) and the linker unbinding rate constant ( $k_c^-$ ). The linker binding rate is proportional to the concentration of unbound linkers ( $[L]$ ), the square of the concentration of free binding sites ( $[F_c]$ ), the linker binding rate constant ( $k_c^+$ ) and the spatial factor for the linker binding reaction ( $\alpha_L$ ). The change of the concentration of bound linker over time ( $\frac{d[c \cdot L \cdot c]}{dt}$ ) is the negative of the concentration of unbound linkers (Equation E).

$$\frac{d[L]}{dt} = -\frac{d[c \cdot L \cdot c]}{dt} = -\alpha_L k_c^+[F_c]^2[L] + k_c^-[c \cdot L \cdot c] \quad \text{Equation E}$$

The change of the unbound motor concentration over time ( $\frac{d[M]}{dt}$ ) is given by the motor unbinding rate, subtracted by the motor binding rate. The motor unbinding rate is proportional to

the concentration of bound motors ( $[c \cdot M \cdot c]$ ) and the motor unbinding rate constant ( $k_M^-$ ). The motor binding rate is proportional to the concentration of unbound motors ( $[M]$ ), the square of the concentration of free binding sites ( $[F_c]$ ), the motor binding rate constant ( $k_M^+$ ), and the spatial factor for the motor binding reaction ( $\alpha_M$ ). The change of the concentration of bound motor over time ( $\frac{d[c \cdot M \cdot c]}{dt}$ ) is the negative of the concentration of unbound motors (Equation F).

$$\frac{d[M]}{dt} = -\frac{d[c \cdot M \cdot c]}{dt} = -\alpha_M k_M^+ [F_c]^2 [M] + k_M^- [c \cdot M \cdot c] \quad \text{Equation F}$$

Finally, the change of the concentration of free binding sites over time ( $\frac{d[F_c]}{dt}$ ) increases due to actin polymerization and decreases when the motor, linker, or brancher binds to a binding site (Equation G). A single free binding site becomes bound during the branching reaction, while two free actin binding sites become bound during the linker and motor binding reactions.

$$\begin{aligned} \frac{d[F_c]}{dt} = & \chi k_p^+ [F_p][G] - \chi k_p^- [F_p] + \chi k_m^+ [F_m][G] - \chi k_m^- [F_m] \\ & - k_B^+ [F_c][G][B] + k_B^- [F_c \cdot B \cdot F_m] \\ & - 2\alpha_C k_C^+ [F_c]^2 [L] + 2k_C^- [F_c \cdot L \cdot F_c] \\ & - 2\alpha_M k_M^+ [F_c]^2 [M] + 2k_M^- [F_c \cdot M \cdot F_c] \end{aligned} \quad \text{Equation G}$$

This set of ordinary differential equations is based on the reactions of the Mechanochemical Dynamics of Active Networks (MEDYAN) model [1–6], developed by Papoian and his group (see section E of the supplementary information for more details). To make our set of equations comparable with the kinetic scheme of the MEDYAN model, we made the estimate that there is only one binding site for every 10 monomers, so unless stated otherwise  $\chi = 0.1$ . The spatial factor is defined as the probability that two actin binding sites are within the search distance in a homogeneous mixture.

Table A. Parameters used in the chemical kinetic model

Description	Constant	Value	Units
Polymerization rate coefficient on the plus end	$k_p^+$	11.6	$\mu\text{M}^{-1}\text{s}^{-1}$
Depolymerization rate coefficient on the plus end	$k_p^-$	1.4	$\text{s}^{-1}$
Polymerization rate coefficient on the minus end	$k_m^+$	1.3	$\mu\text{M}^{-1}\text{s}^{-1}$
Depolymerization rate coefficient on the plus end	$k_m^-$	0.8	$\text{s}^{-1}$
Linker binding rate coefficient	$k_C^+$	0.7	$\mu\text{M}^{-1}\text{s}^{-1}$
Linker unbinding rate coefficient	$k_C^-$	0.3	$\text{s}^{-1}$
Motor binding rate coefficient	$k_M^+$	0.7	$\mu\text{M}^{-1}\text{s}^{-1}$
Motor unbinding rate coefficient	$k_M^-$	1.7	$\text{s}^{-1}$
Brancher binding rate coefficient	$k_B^+$	0.0001	$\mu\text{M}^{-2}\text{s}^{-1}$
Brancher unbinding rate coefficient	$k_B^-$	$1 \times 10^{-10}$	$\text{s}^{-1}$
Linker minimum search distance	$d_C^{min}$	30	<i>nm</i>
Linker maximum search distance	$d_C^{max}$	40	<i>nm</i>
Motor minimum search distance	$d_M^{min}$	175	<i>nm</i>
Motor maximum search distance	$d_M^{max}$	225	<i>nm</i>

## B. The two-step model for the linker binding reaction

The affinity of the actin-binding proteins to single filaments has been shown to be important for domain separation of crosslinkers in experiments [7,8]. To reflect these experimental findings, we studied an alternative chemical kinetic model for linker binding. In this alternative model, each head of the linker binds an actin filament independently with the same binding affinity without cooperativity. We only included the polymerization, the depolymerization reactions, the linker binding, and the unbinding reactions (Table B).

Table B. Reactions included in the two-step chemical kinetic model for linker binding

REACTION	DESCRIPTION
$F_p + G \xrightleftharpoons[k_p^-]{k_p^+} F_p \cdot F_m + F_p + F_c$	Actin polymerization and depolymerization at plus end
$F_m + G \xrightleftharpoons[k_m^-]{k_m^+} F_p \cdot F_m + F_m + F_c$	Actin polymerization and depolymerization at minus end
$F_c + L \xrightleftharpoons[k_c^-]{2k_c^+} F_c \cdot L$	Linker binding and unbinding
$F_c \cdot L + F_c \xrightleftharpoons[k_c^-]{k_c^+} F_c \cdot L \cdot F_c$	Crosslinker formation and dissociation

In this model, the concentrations of the unbound plus sites ( $F_p$ ), the unbound minus sites ( $F_m$ ), the plus sites bound to minus sites ( $F_p \cdot F_m$ ), and unbound G-actin ( $G$ ) follow the Equation C, Equation A, and Equation B respectively assuming there is no brancher present. The concentration of free binding sites ( $F_c$ ) is shown in Equation H.

$$\frac{d[F_c]}{dt} = -2 k_c^+ [F_c][L] - k_c^+ [F_c \cdot L][F_c] + k_c^- [F_c \cdot L] + k_c^- [F_c \cdot L \cdot F_c] + k_p^+ [F_p][G] + k_m^+ [F_m][G] - k_p^- [F_p] - k_m^- [F_m] \quad \text{Equation H}$$

The concentration of linkers in the different states: unbound (L), single bound ( $F_c \cdot L$ ), and double bound ( $F_c \cdot L \cdot F_c$ ), are shown in Equation I, Equation J, and Equation K respectively.

$$\frac{d[L]}{dt} = -2 k_c^+ [F_c][L] + k_c^- [F_c \cdot L] \quad \text{Equation I}$$

$$\frac{d[F_c \cdot L]}{dt} = 2 k_c^+ [F_c][L] - k_c^+ [F_c \cdot L][F_c] - k_c^- [F_c \cdot L] + k_c^- [F_c \cdot L \cdot F_c] \quad \text{Equation J}$$

$$\frac{d[F_c \cdot L \cdot F_c]}{dt} = k_c^+ [F_c \cdot L][F_c] - k_c^- [F_c \cdot L \cdot F_c] \quad \text{Equation K}$$

### C. Probability that an F-actin monomer is not connected to an infinite cluster

(P<sub>s</sub>)

The probability that an F-actin monomer is not connected to an infinite cluster through the site  $\alpha$  ( $Q_\alpha$ ), is the sum of two terms. The first term is the probability of the site not being connected to another site ( $1 - \theta_\alpha$ ). The second term is the probability that the site is connected to another F-actin monomer, with the condition that the neighboring bound F-actin monomer is not connected to an infinite cluster (Equation L).

$$Q_\alpha = 1 - \theta_\alpha + \sum_\beta \frac{\theta_{\alpha \rightarrow \beta}}{Q_\beta} \prod_\gamma Q_\gamma \quad \text{Equation L}$$

In this equation  $\theta_\alpha = \sum_\beta \theta_{\alpha \rightarrow \beta}$ , and  $\alpha$ ,  $\beta$ , and  $\gamma$  can be the plus site (p), the minus site (m), or the actin binding site (c). The probability that an F-actin monomer is not connected to an infinite cluster through any site (P<sub>s</sub>) is the product of the probabilities of the F-actin monomer not being bound to an infinite cluster from each site ( $Q_\alpha$ ). (Equation M).

$$P_s = \prod_\alpha Q_\alpha \quad \text{Equation M}$$

In this equation  $\alpha$  can be the plus site (p), the minus site (m), or the binding site (c). The system of equations given by Equation M has a trivial solution when  $Q_p = Q_m = Q_c = 1$ , which is the only solution when  $P_s=1$  and the system is not percolated. This system can be solved exactly to calculate the fraction of actin monomers in finite clusters given the connectivity probabilities ( $\theta_\alpha$ ) (Figure A).

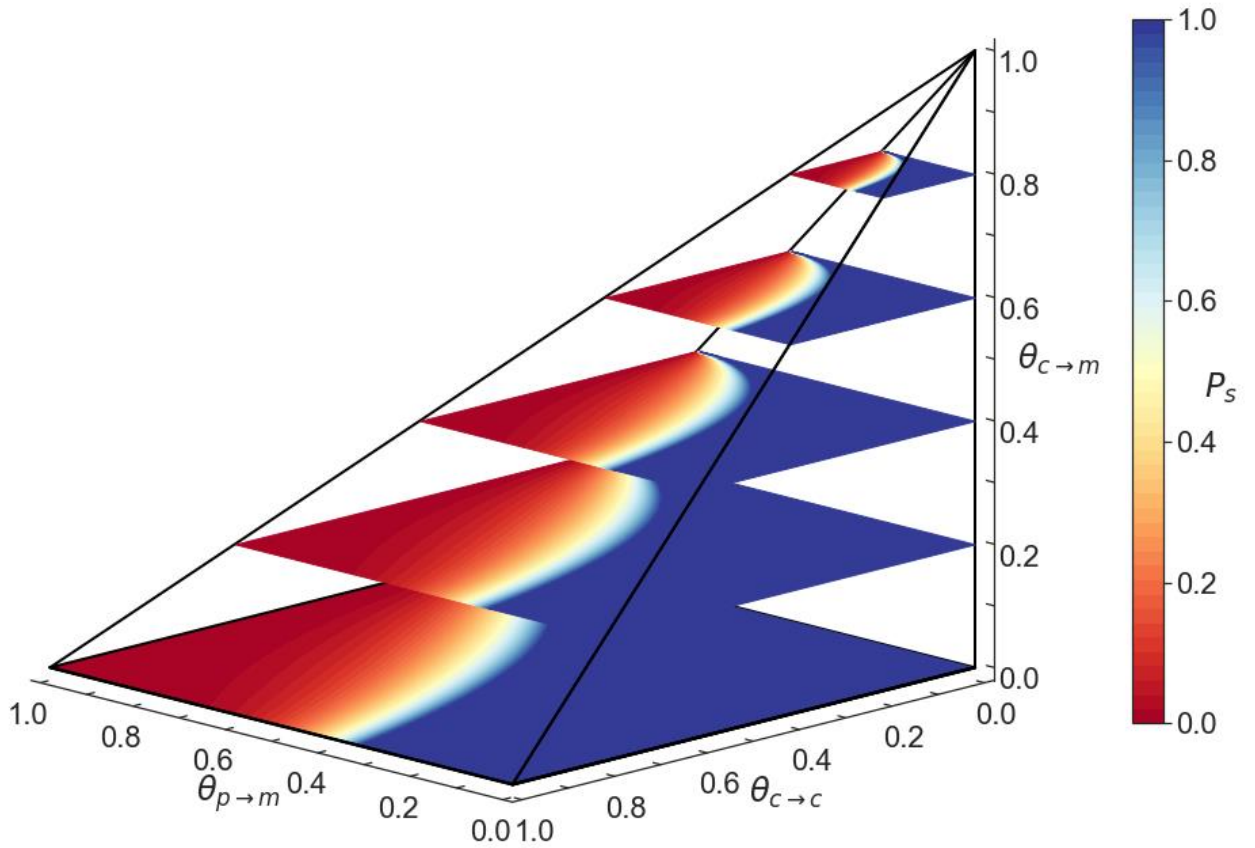


Figure A. Fraction of actin monomers in finite clusters ( $P_s$ ) as a function of the crosslinking probabilities between plus and minus sites ( $\theta_{p \rightarrow m}$ ), binding sites ( $\theta_{c \rightarrow c}$ ), and binding sites to minus sites ( $\theta_{c \rightarrow m}$ ). The color indicates the probability that an F-actin monomer is in a finite cluster.

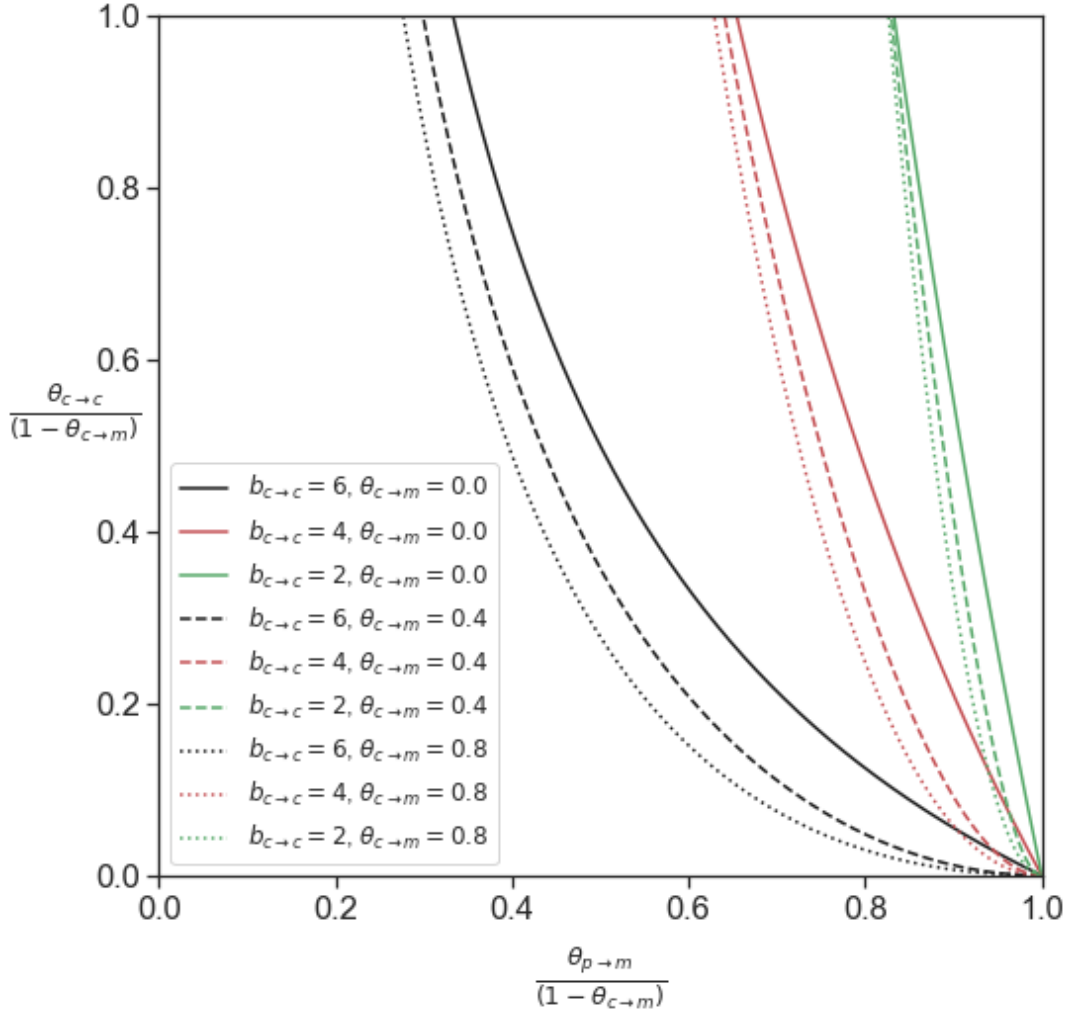


Figure B. Rigidity percolation boundaries as a function of the crosslinking probabilities between plus and minus sites ( $\theta_{p \rightarrow m}$ ), binding sites ( $\theta_{c \rightarrow c}$ ), binding sites and minus sites ( $\theta_{c \rightarrow m}$ ), and the rigidity of the crosslinkers ( $b_{c \rightarrow c}$ ). We assume that the connections between plus and minus sites, as well as the connections between binding sites and minus sites are rigid ( $b_{p \rightarrow m} = b_{c \rightarrow m} = 6$ )

#### D. Analytical solution of the Flory-Stockmayer equations for a system without branchers

We solved analytically the Flory-Stockmayer equations for a system where there is only actin and bivalent crosslinkers, like  $\alpha$ -actinin or filamin. For this system, we calculated the

minimum concentration of linkers needed to percolate the system as a function of the connectivity probabilities ( $\theta_{\alpha \rightarrow \beta}$ ). The probability of an actin monomer to be in a finite cluster is calculated by  $P_s = \prod_{\alpha} Q_{\alpha}$ , as shown in Equation M, where  $\alpha$  can be the plus site (p), the minus site (m) or a binding site (c).  $Q$  is the probability that a site is not connected to an infinite cluster, as shown in Equation L. The analytical solution for  $Q$  for this system can be expressed by the following set of equations:

$$Q_p = 1 - \theta_{p \rightarrow m} + \theta_{p \rightarrow m} Q_p Q_c$$

$$Q_m = 1 - \theta_{m \rightarrow p} + \theta_{m \rightarrow p} Q_m Q_c$$

$$Q_c = 1 - \theta_{c \rightarrow c} + \theta_{c \rightarrow c} Q_p Q_m$$

In this system all connected minus sites are bound to plus sites. Then  $\theta_p = \theta_{p \rightarrow m} = \theta_{m \rightarrow p} = \theta_m = \theta_a$ , where we define  $\theta_a$  as the probability of an F-actin minus site or plus site to be bound. We also define  $\theta_c = \theta_{c \rightarrow c}$  as the probability that an F-actin binding site is bound. The previous system of equations can be reduced to the following set of equations.

$$Q_p = 1 - \theta_a + \theta_a Q_p Q_c$$

$$Q_m = 1 - \theta_a + \theta_a Q_m Q_c$$

$$Q_c = 1 - \theta_c + \theta_c Q_p Q_m$$

This set of equations has two solutions: a trivial solution  $Q_c = Q_p = Q_m = 1$ , and a non-trivial solution as shown below.

$$Q_c = \frac{1}{\theta_a} - \frac{\theta_c}{2} - \sqrt{\frac{\theta_c (4 - 4\theta_a + \theta_a \theta_c)}{4\theta_a}}$$

$$Q_p = Q_m = \sqrt{\frac{(4 - 4\theta_a + \theta_a \theta_c)}{4\theta_a \theta_c}} - \frac{1}{2}$$

The system percolates when  $P_s = \prod_{\alpha} Q_{\alpha} < 1$ , which occurs when:



$$\theta_c = \frac{1 - \theta_a}{2\theta_a}$$

The average length of the filament ( $\langle L \rangle$ ) is related to  $\theta_a$  by the following equation:

$$\langle L \rangle = \frac{1}{1 - \theta_a}$$

This equation shows that the percolation of the actomyosin network requires crosslinking. If the actin does not polymerize ( $\theta_a = 0$ ), all F-actin monomers need to be crosslinked for the system to be percolated ( $\theta_c = 1$ ). As the degree of polymerization increases, the length of the filament decreases and less crosslinkers are required to percolate the system. The proportion of crosslinkers required to percolate the network decreases hyperbolically as the length of the filament increases (Figure C).

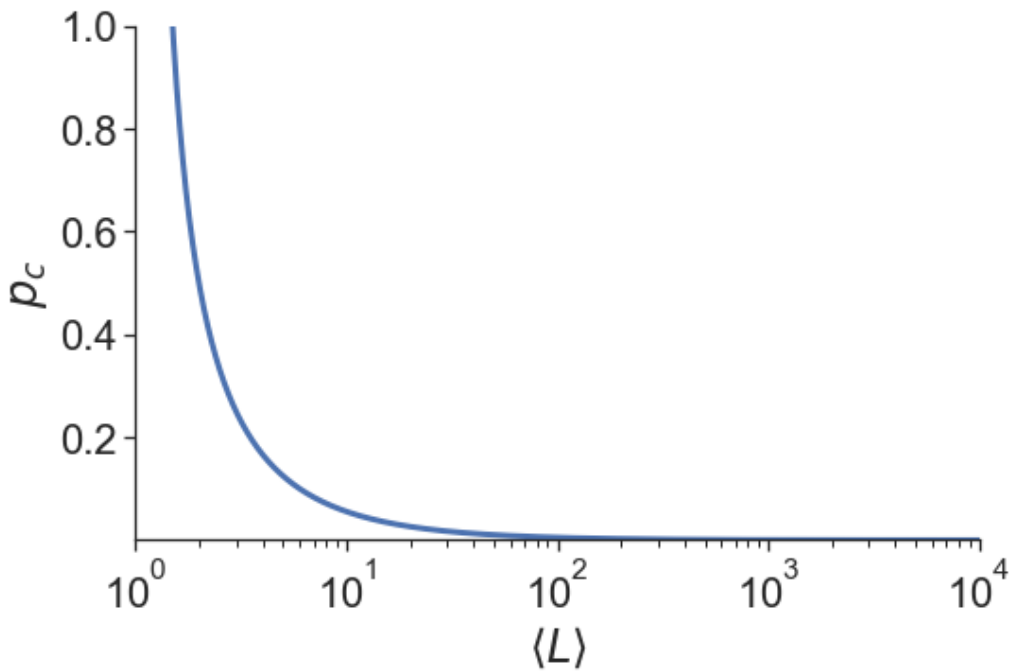


Figure C. Minimum crosslinking probability ( $\theta_c$ ) needed to percolate a system containing only linkers as a function of the average filament length ( $\langle L \rangle$ )

### **E. Coarse-grained mechanochemical model of actomyosin systems (MEDYAN)**

We have used a coarse-grained mechanochemical model of actomyosin systems called MEDYAN (Mechanochemical Dynamics of Active Networks) developed by Papoian and his group [1–4,6]. MEDYAN models both stochastic chemical reactions and deterministic mechanical representations of far-from-equilibrium systems. In this study, we have included some important actin-binding proteins in actomyosin networks: non-muscle myosin IIA heavy chain (NMIIA) motors,  $\alpha$ -actinin linkers, and actin-related protein complex 2/3 (Arp2/3) branchers, all in a fixed geometry.

MEDYAN represents actin filaments, linkers, motors, and branchers mechanically. For example, the actin filaments in MEDYAN are modeled as stretchable and bendable rods that have repulsive interactions with other filaments and the boundary. In our simulations a motor mini filament is an ensemble of, on average, 22.5 NMIIA subunits. MEDYAN simulations model chemical reactions stochastically using reaction-diffusion equations. Our MEDYAN simulations employ eight different compartments therefore allowing us to consider heterogeneous distributions of chemical species. Lastly, MEDYAN simulations consider mechanochemical feedback between the mechanical representations and the chemical reactions in the system. For example, the unbinding reaction of motors is modeled using a catch bond; hence motors are more likely to stay bound to a filament when pulling forces are applied to the motor.

We simulated a system containing 15000 actin monomers, 300 branchers, 333 motor mini-filaments, and 1500 linkers in a  $1\mu\text{m}^3$  cubic box, which corresponds to an actin concentration of  $25\mu\text{M}$ , a linker concentration of  $2.5\mu\text{M}$ , a motor concentration of  $12.5\mu\text{M}$ , and a brancher concentration of  $0.5\mu\text{M}$ . The branchers were introduced after 1 second. The parameters for the simulations are the same as the ones shown in Table A. The simulations were run with and without

branchers. Sample snapshots of the simulation are shown in Figure D. We followed the same procedure as described by in our previous work [5]. We compared the concentration of bound species to the concentration of bound species found in the chemical kinetic model. To compare the discrete snapshots of MEDYAN to the continuum chemical kinetic model, we ran 5 different simulations with different snapshot intervals, each 0.001s, 0.01s, 0.1s, 1s, and 10s, and interpolated the obtained data using a linear interpolation with the closest snapshots to obtain a mean MEDYAN concentration. G-actin monomers, F-actin filaments,  $\alpha$ -actinin linkers, NMIIA motors, and Arp2/3 branchers were initially distributed randomly inside a cubic container. All MEDYAN simulations were enclosed in a  $1 \mu\text{m} \times 1 \mu\text{m} \times 1 \mu\text{m}$  container with  $25 \mu\text{M}$  of actin in total as described in our previous work [5].

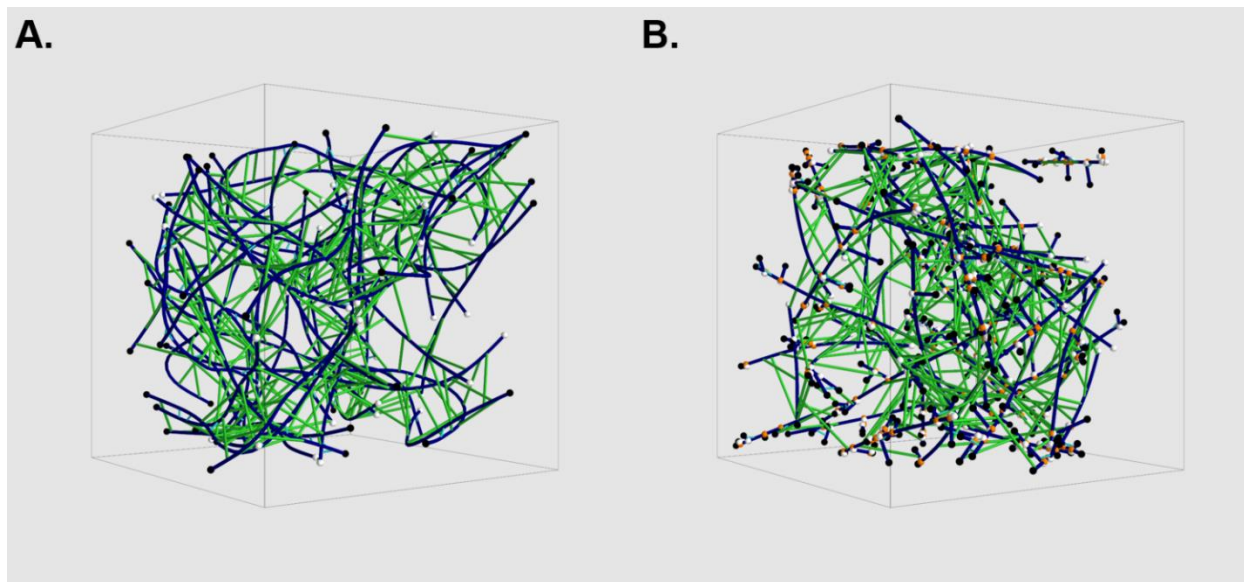


Figure D. Typical snapshot of MEDYAN simulation of an unbranched network simulation (A) and a branched network simulation (B). Blue cylinders represent actin filaments, green cylinders represent motor mini-filaments, cyan cylinders represent crosslinkers, white beads represent the minus end, black beads represent the plus end, and orange beads represent branchers.

## References

1. Popov K, Komianos J, Papoian GA. MEDYAN: Mechanochemical Simulations of Contraction and Polarity Alignment in Actomyosin Networks. *PLoS Comput Biol.* 2016;12:

- 1–35. doi:10.1371/journal.pcbi.1004877. PubMed PMID: 27120189
2. Ni Q, Papoian GA. Turnover versus treadmilling in actin network assembly and remodeling. *Cytoskeleton* (Hoboken). 2019;76: 562–570. doi:10.1002/cm.21564. PubMed PMID: 31525282
  3. Floyd C, Papoian GA, Jarzynski C. Quantifying dissipation in actomyosin networks. *Interface Focus*. 2019;9: 20180078. doi:10.1098/rsfs.2018.0078. PubMed PMID: 31065344
  4. Komianos JE, Papoian GA. Stochastic Ratcheting on a Funneled Energy Landscape Is Necessary for Highly Efficient Contractility of Actomyosin Force Dipoles. *Phys Rev X*. 2018;8: 21006. doi:10.1103/PhysRevX.8.021006
  5. Liman J, Bueno C, Eliaz Y, Schafer NP, Waxham MN, Wolynes PG, et al. The role of the Arp2/3 complex in shaping the dynamics and structures of branched actomyosin networks. *Proc Natl Acad Sci U S A*. 2020;117: 10825–10831. doi:10.1073/pnas.1922494117. PubMed PMID: 32354995
  6. Chandrasekaran A, Upadhyaya A, Papoian GA. Remarkable structural transformations of actin bundles are driven by their initial polarity, motor activity, crosslinking, and filament treadmilling. *PLoS Comput Biol*. 2019;15: e1007156. doi:10.1371/journal.pcbi.1007156. PubMed PMID: 31287817
  7. Freedman SL, Suarez C, Winkelman JD, Kovar DR, Voth GA, Dinner AR, et al. Mechanical and kinetic factors drive sorting of F-actin cross-linkers on bundles. *Proc Natl Acad Sci U S A*. 2019;116: 16192–16197. doi:10.1073/pnas.1820814116. PubMed PMID: 31346091
  8. Winkelman JD, Suarez C, Hocky GM, Harker AJ, Morganthaler AN, Christensen JR, et al.

Fascin- and  $\alpha$ -Actinin-Bundled Networks Contain Intrinsic Structural Features that Drive Protein Sorting. *Curr Biol.* 2016;26: 2697–2706. doi:10.1016/j.cub.2016.07.080. PubMed PMID: 27666967

# Appendix B

Supplementary Information for

**The role of the Arp2/3 complex in shaping the dynamics and structures of branched actomyosin networks**

James Liman, Carlos Bueno, Yossi Eliaz, Nicholas P. Schafer, M. Neal Waxham, Peter G. Wolynes, Herbert Levine, and Margaret S. Cheung

\* Corresponding authors: Peter G. Wolynes ([pgwolynes@rice.edu](mailto:pgwolynes@rice.edu)), Herbert Levine ([h.levine@northeastern.edu](mailto:h.levine@northeastern.edu)), and Margaret S. Cheung ([mscheung@uh.edu](mailto:mscheung@uh.edu))

**This PDF file includes:**

Supplementary text  
Figures S1 to S4  
Tables S1 to S3  
SI References

## MEDYAN software package

MEDYAN models the dynamics of actomyosin networks using rounds of four mechanochemical steps. First, the chemical reactions that will take place are chosen using a stochastic reaction-diffusion model. Second, the topology of the actomyosin network is updated according to the chemical reactions that occurred during the previous step. Third, we minimize the energy of the network with this updated topology. Fourth, the forces exerted on the network components are used to update the chemical rates. Rounds of these four steps are performed repeatedly. A diagram of this process is shown in Figure S1.

### Step 1: The chemical model

For the chemical reaction step, we considered the polymerization reactions of F-actin on both the plus ends ( $k_{p+}$ ) and the minus ends ( $k_{p-}$ ), the depolymerization reactions of F-actin on both the plus ends ( $k_{dp+}$ ) and the minus ends ( $k_{dp-}$ ), the binding ( $k_{bl}$ ) and the unbinding ( $k_{ubl}$ ) reactions of  $\alpha$ -actinin linkers, the binding ( $k_{bm}$ ) and the unbinding reactions ( $k_{ubm}$ ) of NMIIA motors, the walking reactions ( $k_{wm}$ ) of NMIIA motors, the branching reaction of F-actin ( $k_{bf}$ ), and the destruction reaction of short F-actin no longer than two monomers ( $k_{df}$ ). New to this current study, we modeled the Arp2/3 binding to actin filaments as an irreversible reaction because the stability of the Arp2/3-actin protein complexes is known from experiments to be high (1-3). All simulations were confined to a  $1 \mu\text{m} \times 1 \mu\text{m} \times 1 \mu\text{m}$  box with  $25 \mu\text{M}$  of actin in total. The concentration of actin,  $25 \mu\text{M}$ , was specifically chosen to replicate the concentrations of the *in vitro* experiment by the Weitz group (4). Table S1 compiles the reaction rates used in our simulations.

### Step 2: Update of mechanical representations

After the changes of the chemical composition had been computed, we updated the mechanical representation in the simulated system. For example, if a polymerization reaction occurs during the previous step, the mechanical model then will change the representation of the cylinder to reflect the addition of an actin monomer. This change in polymer size may move the system to a mechanically unstable configuration. The system will be returned to a stable configuration in the next step.

### Step 3: The mechanical model

In our simulations, MEDYAN mechanically represents filaments, branchers, motors, and linkers inside a rigid box. An F-actin filament is modeled as a semi-flexible rod that is bendable, stretchable, but however, non-twistable (5). The bending potential  $U_i^{bend}$  of the semi-flexible rod is described in Equation S1.

$$U_i^{bend} = K_{bend}[1 - \cos(\theta_{i,i+1})] \quad (\text{S1})$$

where  $K_{bend}$  is the bending energy constant and  $\theta_{i,i+1}$  is the angle between the cylindrical segment  $i$  and the neighboring cylindrical segment  $i + 1$ . The stretching potential of the semi-flexible rod  $U_i^{stretch}$  is described in Equation S2.



$$U_i^{stretch} = \frac{1}{2} K_{stretch} (|\vec{l}_i| - l_0)^2 \quad (S2)$$

where  $K_{stretch}$  is the stretching energy constant,  $\vec{l}_i$  is the vector between endpoints of the  $i^{th}$  cylindrical monomer segment, and  $l_0$  is the equilibrium length of a cylinder. The stretch elasticity constant for actin was chosen for computational efficiency. The exact value of this constant was shown to have a negligible effect on the actin dynamics.

The polymer branching potential  $U_{i,j}^{branch}$  is described in Equation S3.

$$U_{i,j}^{branch} = U_{i,j}^{branch,stretch} + U_{i,j}^{branch,angular} \quad (S3)$$

where the cylindrical segment  $i$  is the mother filament and the cylindrical segment  $j$  is the daughter filament.

The branched polymer stretching potential  $U_{i,j}^{stretch}$  is described in Equation S4

$$U_{i,j}^{stretch} = K_{stretch}^{branch} (|\vec{d}_{i,j}| - d_0)^2 \quad (S4)$$

where  $K_{stretch}^{branch}$  is the stretching constant of branched polymers,  $\vec{d}_{i,j}$  is the distance between the binding site in cylinder  $i$  and the minus end of cylinder  $j$ , and  $d_0$  is the equilibrium length of  $\vec{d}_{i,j}$ .

The branched filament angular potential  $U_{i,j}^{angular}$  is described in Equation S5.

$$U_{i,j}^{angular} = K_{angular}^{branch} [1 - \cos(\theta_{i,j} - \theta_0)] \quad (S5)$$

where  $K_{angular}^{branch}$  is the angular energy constant for the branched polymer,  $\theta_{i,j}$  is the angle between mother  $i$  and daughter  $j$  polymers as defined by the angle between the vector defined from the minus end to the plus end of the mother filament and the vector defined from the minus end to the plus end of the daughter filament.  $\theta_0$  is the equilibrium angle between branched polymers.

The motor stretching harmonic potential  $U_{i,j}^{motor}$  is described in Equation S6.

$$U_{i,j}^{motor} = \frac{1}{2} K_{motor} (|\vec{l}_{ij}^m| - l_0^m)^2 \quad (S6)$$

where  $K_{motor}$  is the stretching energy constant of motors,  $l_{ij}^m$  is the instantaneous length of the motor, as measured by the distance between the binding sites it is bound to, and  $l_0^m$  is the equilibrium length of the motor. Linkers are also modeled with a stretching harmonic potential analogous to the potential described in S7 but with a different stretching energy constant  $K_{linker}$ . For the implementation of the motor stretching harmonic potential in MEDYAN, please refer to (5). Motors can bind simultaneously to pairs of actin cylinders whose binding sites are separated by a distance between 175 to 225 nm. Linkers can bind simultaneously to pairs of actin cylinders whose binding sites are separated by a distance between 30 to 40 nm.

Arp2/3 nucleates daughter filaments by first binding to a mother filament. The angle between the mother and its daughter filament is approximately  $70^\circ$  (1.22 radians). Linkers are not allowed to bind pairs of binding sites immediately downstream towards the plus ends from a brancher.

The exclusion volume potential  $U_{i,j}^{vol}$  between two cylindrical segments on two adjacent polymers is given in Equation S7.

$$U_{i,j}^{vol} = \int_{l_i} \int_{l_j} \delta U (|\vec{r}_i - \vec{r}_j|) dl_i dl_j \quad (S7)$$

where the pair potential  $\delta U (|\vec{r}_i - \vec{r}_j|)$  for pure excluded volume repulsion is the inverse fourth-power of distance  $\frac{1}{(\vec{r}_i - \vec{r}_j)^4}$  where  $i$  and  $j$  are the indices for the two interacting cylindrical segments. For the implementation of the exclusion volume potential in MEDYAN, please refer to (5).

The interaction potential between a filament and the boundary of the cubic container  $U_i^{boundary}$  is described in Equation S8.

$$U_i^{boundary} = \begin{cases} K_{boundary} e^{-\frac{d_i}{\lambda}} & d_i \leq d_{cutoff} \\ 0 & d_i > d_{cutoff} \end{cases} \quad (S8)$$

where  $K_{boundary}$  is the repulsive energy constant,  $d_i$  is the closest distance between each boundary and the  $i^{th}$  cylindrical segment, and  $\lambda$  is the screening length. The calculation of the interaction potential between a filament and the boundary was done after energy minimization. Table S2 lists all parameters of the mechanical model used in our simulations.

In this mechanical model phase, MEDYAN minimizes the energy of the system using a Polak–Ribière conjugate gradient method. After the energy minimization is finished, a new mechanical configuration of the system—including new locations and stress profiles of filaments, bound motors and linkers, and branchers—is formed.

#### Step 4: Update of chemical rates

Having reached a new configuration after the minimization step, the different components in the system such as motors, linkers or filament ends are subject to different stresses. During this step the forces acting on the components change the reaction rates by modifying the rate constants. For example: a polymerization reaction is less likely to occur on a plus end of a filament that is located closer to the boundary since the plus end of the filament experiences greater repulsion boundary forces. The unbinding reaction of linkers is modeled using a slip bond and the unbinding reaction of motors is modeled using a catch bond. Hence, linker is more likely to unbind when pulling forces are applied to the linker and motor is less likely to unbind when pulling forces are applied to the motor. Table S3 lists all mechanochemical parameters used in our simulations.

## The radius of gyration ( $R_g$ ) of actomyosin networks with varying concentrations of motor and linker proteins can be used to distinguish contracted and non-contracted systems

We calculated the radius of gyration ( $R_g$ ) of the actin filaments to quantify the contraction of actomyosin networks. The radius of gyration of the actin filaments  $R_g$  is described in Equation S9.

$$R_g = \sqrt{\frac{1}{N} \sum_i^N (C_i - CoM)^2} \quad (S9)$$

where  $N$  is the total number of cylinders,  $C_i$  is the center of mass of cylinder  $i$ , and CoM is the center of mass of the network. We calculated  $R_g$  for networks with different concentration ratios of motor and linker to actin for both unbranched and branched networks. We plotted the ratio of the instantaneous radius of gyration,  $R_g$ , to the radius of gyration at 10 s,  $R_g^i$ , ( $R_g/R_g^i$ ) for both systems.  $R_g^i$  is measured at 10 seconds because the filament length distribution is allowed to reach a steady-state before the motors are added at 10 s. Note that for every combination of motor and linker to actin ratio, sixteen simulation replicates were performed, differing only in their random initializations. We calculated  $\langle R_g/R_g^i \rangle$ , which is the average of the normalized radius of gyration ratio taken all replicates for a given condition of each actomyosin network with varying motor and linker concentrations for both unbranched and branched networks.

Next, we fitted these  $\langle R_g/R_g^i \rangle(t)$  using Equation S10.

$$\langle R_g/R_g^i \rangle(t) = A - B \cdot e^{-\frac{t}{\tau}} \quad (S10)$$

where  $A$ ,  $B$ , and  $\tau$  are parameters of the fitting equation,  $t$  is time in seconds, and  $\tau$  is the exponential time constant. The difference of  $B$  and  $A$ , ( $A - B$ ) is the  $\langle R_g/R_g^i \rangle$  at  $t = 0$  while  $A$  is the  $\langle R_g/R_g^i \rangle$  at the steady-state.

## A state diagram of the boundary interaction energy shows that the contractility is re-entrant for unbranched actomyosin networks

We chose the total boundary interaction energy as a complementary measure of contractility to the radius of gyration  $R_g$ . Smaller values of the total boundary interaction energy correspond to more contracted systems. While the radius of gyration ( $R_g$ ) of the actin filaments can be used to distinguish between contracted and non-contracted actomyosin networks,  $R_g$  is not the only parameter that can be used to quantify the degree of contractedness. A comparison of the radii of gyration and boundary interaction energies for contracted and non-contracted system is shown in Figure S2. The  $R_g$  and the boundary interaction energy of the less contracted network (Figure S2A) are 0.34  $\mu\text{m}$  and 0.0071 pN nm respectively, while the  $R_g$  and the boundary interaction energy of the more contracted network (Figure S2B) are 0.22  $\mu\text{m}$  and 0.0014 pN nm respectively. We observe that in this case the boundary interaction energy better distinguishes between contracted and non-contracted systems.

State diagram of the boundary interaction energy of unbranched actomyosin networks is shown in Figures S3. We observe a re-entrant feature—where, above a motor threshold

concentration, the systems with intermediate concentrations of linker are more contracted than the systems with either high or low concentration of linker—for the unbranched networks. The re-entrant feature is observed for unbranched systems with  $x_{m:A}$  approximately between 0.05 and 0.2 as highlighted by the dotted green lines in Figure S3.

A plot comparing the distributions of the mean filament displacements  $\delta x_f$  for both unbranched network and branched network simulations is shown in Figure S4. The displacements of the branched networks have a rather wide distribution compared with that for their unbranched counterparts. Defining a specific threshold of  $\delta x_f$  to classify an avalanche is somewhat arbitrary since filament displacements of a network depend on their past topology and history. Therefore, we use the Z-score relative to the displacements for other intervals to classify an event as being an avalanche rather than using a simple displacement cutoff.

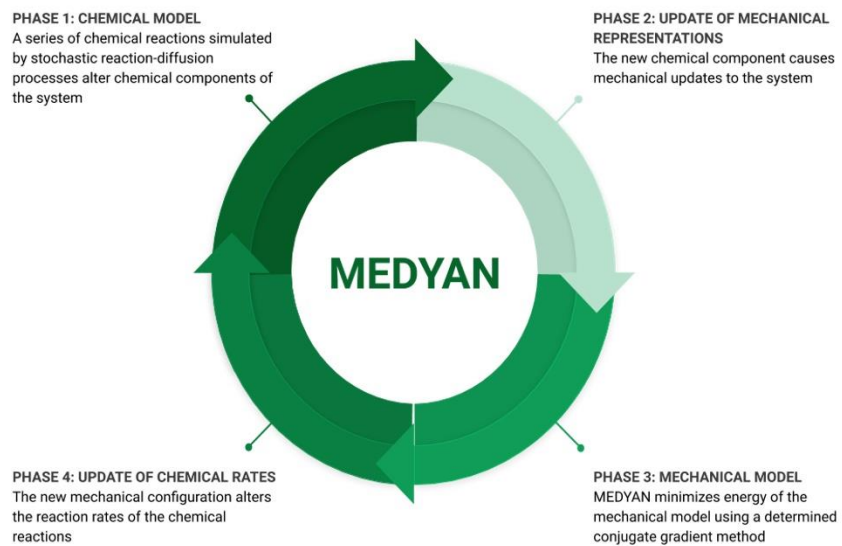


Figure S1. MEDYAN software package flow diagram (adapted from Fig 4 in Ref. (5)).

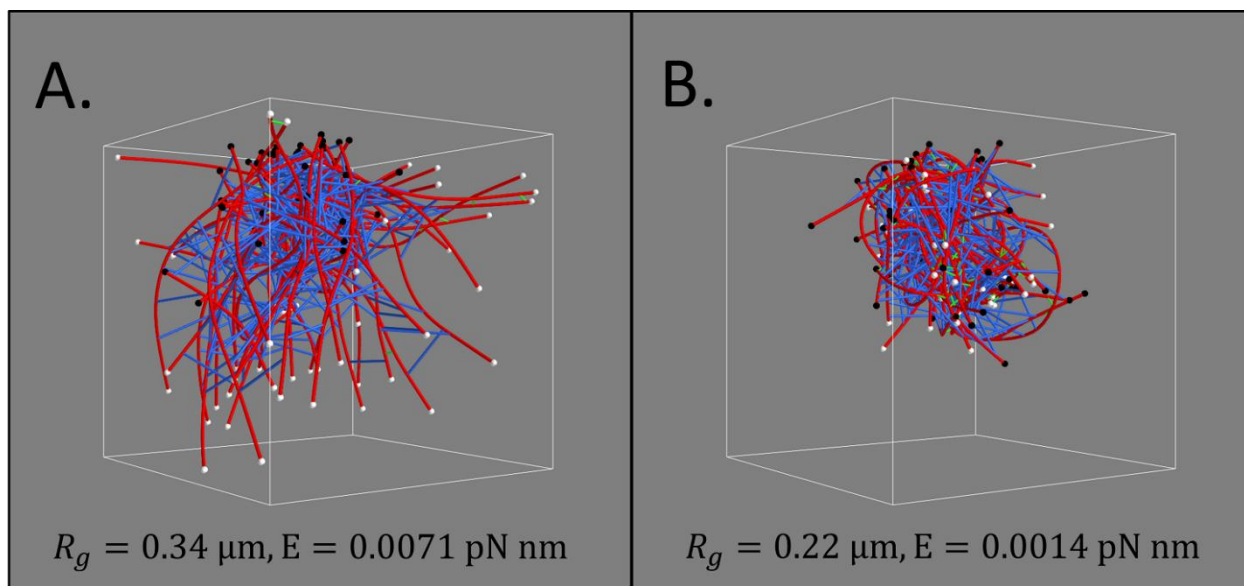


Figure S2. Comparison of two unbranched actomyosin networks with different degrees of contractedness. The  $R_g$  and the boundary interaction energy of the less contracted network (A) are  $0.34 \mu\text{m}$  and  $0.0071 \text{ pN nm}$  respectively, while the  $R_g$  and the boundary interaction energy of the more contracted network (B) are  $0.22 \mu\text{m}$  and  $0.0014 \text{ pN nm}$  respectively.

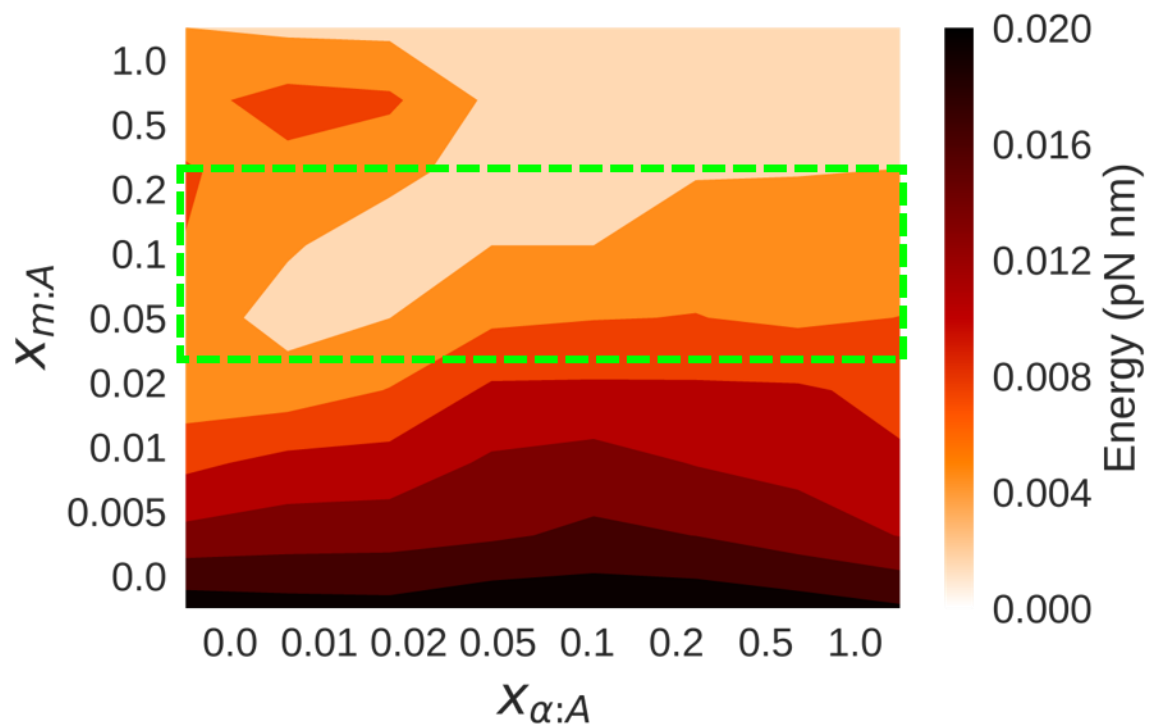


Figure S3. State diagram of the boundary interaction energy of the unbranched actomyosin networks at 2000 s. The area surrounded by dotted green lines highlights the re-entrant feature of contractility for the unbranched networks.

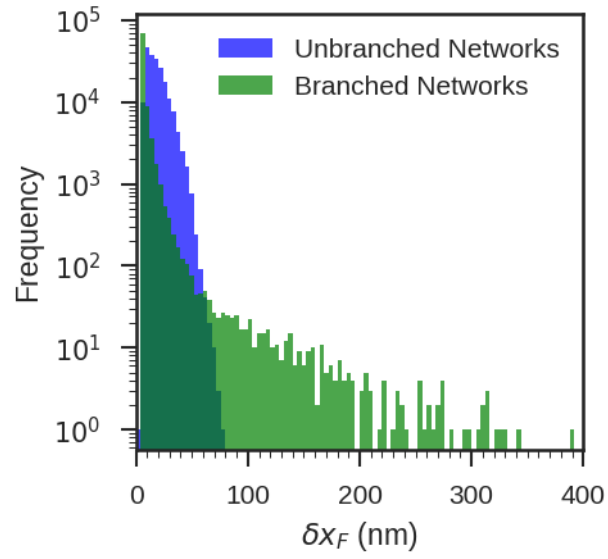


Figure S4. A plot of the distributions of the mean filament displacements  $\delta x_f$  of both unbranched network and branched network simulations.



Table S1: Reaction rates used in the chemical models for MEDYAN simulations in this study.

<b>Symbols</b>	<b>Reaction rates</b>
$k_{p+}$	$11.6 \mu\text{M}^{-1} \text{s}^{-1}$ (6)
$k_{p-}$	$1.3 \mu\text{M}^{-1} \text{s}^{-1}$ (6)
$k_{dp+}$	$1.4 \text{s}^{-1}$ (6)
$k_{dp-}$	$0.8 \text{s}^{-1}$ (6)
$k_{bl}$	$0.7 \mu\text{M}^{-1} \text{s}^{-1}$ (7)
$k_{ubl}$	$0.3 \text{s}^{-1}$ (7)
$k_{bm}$	$0.7 \mu\text{M}^{-1} \text{s}^{-1}$ (8)
$k_{ubm}$	$1.7 \text{s}^{-1}$ (5)
$k_{wm}$	$0.2 \text{s}^{-1}$ (5)
$k_{bf}$	$0.0001 \text{s}^{-1}$ (limited to number of branchers)
$k_{df}$	$1.0 \text{s}^{-1}$ (only applied to 1 actin segment filament)

Table S2: Parameters of the mechanical models used for the MEDYAN simulations in this study.

<b>Symbol</b>	<b>Constants</b>
$K_{bend}$	2690 pN.nm
$K_{stretch}$	100 pN/nm
$K_{stretch}^{branch}$	100 pN/nm
$d_0$	6 nm
$K_{angular}^{branch}$	100 pN.nm
$\theta_0$	$\sim 70^\circ$ (2, 9)
$K_{motor}$	2.5 pN/nm
$K_{linker}$	8.0 pN/nm
$K_{boundary}$	41 pN.nm (equivalent to 10 $k_bT$ )
$\lambda$	2.7 nm
$d_{cutoff}$	200 nm

Table S3: Parameters of the mechanochemical models used for the MEDYAN simulations in this study.

<b>Description</b>	<b>Constants</b>
NMIIA motor head stall force	15 pN
NMIIA motor head unbinding force	12.62 pN
$\alpha$ -actinin characteristic unbinding length	0.24 nm
Actin polymerization characteristic length	2.7 nm
Motor walking speed	0.2 steps/s

## References

1. R. D. Mullins, J. A. Heuser, T. D. Pollard, The interaction of Arp2/3 complex with actin: Nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proceedings of the National Academy of Sciences* **95**, 6181-6186 (1998).
2. T. M. Svitkina, G. G. Borisy, Arp2/3 Complex and Actin Depolymerizing Factor/Cofilin in Dendritic Organization and Treadmilling of Actin Filament Array in Lamellipodia. *The Journal of Cell Biology* **145**, 1009-1026 (1999).
3. B. A. Smith, K. Daugherty-Clarke, B. L. Goode, J. Gelles, Pathway of actin filament branch formation by Arp2/3 complex revealed by single-molecule imaging. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 1285-1290 (2013).
4. P. M. Bendix *et al.*, A quantitative analysis of contractility in active cytoskeletal protein networks. *Biophys J* **94**, 3126-3136 (2008).
5. K. Popov, J. Komianos, G. A. Papoian, MEDYAN: Mechanochemical Simulations of Contraction and Polarity Alignment in Actomyosin Networks. *PLoS Comput Biol* **12**, e1004877 (2016).
6. T. D. Pollard, Rate constants for the reactions of ATP- and ADP-actin with the ends of actin filaments. *The Journal of Cell Biology* **103**, 2747-2754 (1986).
7. D. H. Wachsstock, W. H. Schwartz, T. D. Pollard, Affinity of alpha-actinin for actin determines the structure and mechanical properties of actin filament gels. *Biophysical Journal* **65**, 205-214 (1993).
8. M. Kovács, F. Wang, A. Hu, Y. Zhang, J. R. Sellers, Functional Divergence of Human Cytoplasmic Myosin II: KINETIC CHARACTERIZATION OF THE NON-MUSCLE IIA ISOFORM. *Journal of Biological Chemistry* **278**, 38132-38140 (2003).
9. R. D. Mullins, J. A. Heuser, T. D. Pollard, The interaction of Arp2/3 complex with actin: Nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6181-6186 (1998).

# Appendix C

# Supporting Information

OpenAWSEM with Open3SPN2: a fast, flexible, and accessible framework for large scale coarse-grained biomolecular simulations

## Contents

<b>1</b>	<b>OpenAWSEM</b>	<b>1</b>
1.1	Connectivity term . . . . .	2
1.2	Chain term . . . . .	2
1.3	Chirality term . . . . .	2
1.4	Rama term . . . . .	2
1.5	Excluded Volume term . . . . .	3
1.6	Contact term . . . . .	4
1.7	$\beta$ -hydrogen bonding and P-AP terms . . . . .	4
<b>2</b>	<b>Open3SPN2</b>	<b>5</b>
2.1	Bonded terms . . . . .	7
2.2	Stacking, BasePairing and CrossStacking terms . . . . .	13
2.3	Non-bonded terms . . . . .	18
2.4	Protein-DNA Excluded Volume term . . . . .	19
2.5	Protein-DNA Electrostatics term . . . . .	19
<b>3</b>	<b>Energy validation of the OpenMM implementation of AWSEM, 3SPN.2, and 3SPN.2C</b>	<b>20</b>
3.1	Energy evaluation comparison with LAMMPS AWSEM . . . . .	20
3.2	Energy evaluation comparison with LAMMPS 3SPN.2 and 3SPN.2C	20
<b>4</b>	<b>Tutorial</b>	<b>21</b>
4.1	open3SPN2 . . . . .	21
4.1.1	Example DNA system . . . . .	21
4.1.2	Example Protein-DNA system . . . . .	22
<b>5</b>	<b>Supplementary figures</b>	<b>26</b>
5.1	Structure prediction results using three contact potential schemes evaluated using the overall Q . . . . .	26
5.2	Example of over saturation of disulfide bonds observed in original AWSEM simulation. . . . .	26
5.3	Bets Q for each run. . . . .	26
5.4	The predicted structure of alpha-thrombin(PDB: 1ppb) aligned with the crystal structure. . . . .	28
5.5	The predicted structure of ribonuclease A(PDB: 1fs3) aligned with the crystal structure. . . . .	28

## 1 OpenAWSEM

In AWSEM coarse grained simulations, the amino acids are represented by six particles, (CA, CB, O, C, N and H) except for Proline and Glycine both of which are represented by five particles. For Proline, no hydrogen is connected to the

nitrogen inside its amide group. Glycine has no CB. Among those 6 particles in the standard representation, C, N and H are designated as "virtual sites" which means their coordinates are not dynamical variables but instead are computed based on the positions of the other particles, which are dynamical variables.

The standard AWSEM potential is made up of several term:

$$V_{AWSEM} = V_{con} + V_{chain} + V_{chi} + V_{rama} + V_{excl} + V_{contact} + V_{beta} + V_{pap} + V_{frag} \quad (1)$$

### 1.1 Connectivity term

The connectivity term is designed to maintain the bonded distances between  $C\alpha_i$  and  $O_i$ ,  $C\beta_i$  and  $C\alpha_{i+1}$ . and between  $O_i$  to  $C\alpha_{i+1}$ .

$$V_{con} = k_{con} \left( \sum_i^N (r_{C\alpha_i O_i} - r_{C\alpha O}^0)^2 + \sum_{res_i != GLY} (r_{C\alpha_i C\beta_i} - r_{C\alpha\beta}^0)^2 \right) \quad (2)$$

$$+ \sum_i^{N-1} ((r_{C\alpha_i C\alpha_{i+1}} - r_{C\alpha C\alpha_{i+1}}^0)^2 + (r_{O_i C\alpha_{i+1}} - r_{OC\alpha_{i+1}}^0)^2) \quad (3)$$

### 1.2 Chain term

The chain term models the positions of C' and N atoms.

$$V_{chain} = \lambda_{chain} \left[ \sum_{i=2}^N (r_{N_i C\beta_i} - r_{N_i C\beta_i}^0)^2 + \sum_{i=1}^{N-1} (r_{C'_i C\beta_i} - r_{C'_i C\beta_i}^0)^2 + \sum_{i=2}^{N-1} (r_{N_i C'} - r_{N C'}^0)^2 \right] \quad (4)$$

We implemented the connectivity term and the chain term using "HarmonicBondForce".

### 1.3 Chirality term

The chirality term is used to fix the direction of the  $C\beta_i$  relative to the plane formed by  $C'_i$ ,  $C\alpha_i$  and  $N_i$ .

$$V_\chi = \lambda_\chi \sum (\chi_i - \chi_0)^2 \quad (5)$$

$$\chi_i = \frac{(\mathbf{r}_{C'_i C\alpha_i} \times \mathbf{r}_{C\alpha_i N_i}) \cdot \mathbf{r}_{C\alpha_i C\beta_i}}{|\mathbf{r}_{C'_i C\alpha_i} \times \mathbf{r}_{C\alpha_i N_i}| \cdot |\mathbf{r}_{C\alpha_i C\beta_i}|} \quad (6)$$

### 1.4 Rama term

The rama term is used to fix the  $\phi$ ,  $\psi$  angles within a reasonable range.

$$V_{rama} = -\lambda_{rama} \sum_{i=2}^{N-1} \sum_j W_j e^{-\sigma_j (\omega_{\phi_j} (\cos(\phi_i - \phi_j^0) - 1)^2 + \omega_{\psi_j} (\cos(\psi_i - \psi_j^0) - 1)^2)} \quad (7)$$

The chirality term  $V_\chi$  and Rama term was implemented using "CustomCompoundBondForce".

Table 1: parameters

Parameter	Value	Units
$\lambda_{con}$	120	kcal/Å <sup>2</sup> mol
$\lambda_{chain}$	120	kcal/Å <sup>2</sup> mol
$\lambda_{\chi}$	60	kcal/ mol
$\lambda_{rama}$	2	kcal/ mol
$\lambda_{excl}$	20	kcal/Å <sup>2</sup> mol
$r_{C\alpha_i C\alpha_{i+1}}^0$	3.816	Å
$r_{C\alpha_i C O_i}^0$	2.40	Å
$r_{C O_i C\alpha_i}^0$	2.76	Å
$r_{C\alpha_i C\beta_i}^0$	1.53	Å
$r_{N_i C\beta_i}^0$	2.46	Å
$r_{C'_i C\beta_i}^0$	2.52	Å
$r_{N_i C'_i}^0$	2.46	Å
$\chi_0$	-0.71	Å <sup>3</sup>

	General Case			Alpha Helix	Beta Sheet	Proline	
W	1.3149	1.32016	1.0264	2.0	2.0	2.17	2.15
$\sigma$	15.398	49.0521	49.0954	419.0	15.398	105.52	109.09
$\omega_\phi$	0.15	0.25	0.65	1.0	1.0	1.0	1.0
$\phi_0$	-1.74	-1.265	1.041	-0.895	-2.25	-1.153	-0.95
$\omega_\psi$	0.65	0.45	0.25	1.0	1.0	0.15	0.15
$\psi_0$	2.138	-0.318	0.78	-0.82	2.16	2.4	-0.218

## 1.5 Excluded Volume term

The excluded volume term prevents the overlapping of backbone atoms.

$$V_{excl} = \lambda_{excl} \sum_{ij} [H(r_{C_i C_j} - r_{ex}^C)(r_{C_i C_j} - r_{ex}^C)^2 + H(r_{O_i O_j} - r_{ex}^O)(r_{O_i O_j} - r_{ex}^O)^2] \quad (8)$$

$$H(r) = \begin{cases} 1 & x \geq 0 \\ 0 & x \leq 0 \end{cases} \quad (9)$$

The excluded volume term used "CustomNonbondedForce". All the parameters are the same as those defined in the original AWSEM paper. The parameters are defined in Table 1



## 1.6 Contact term

The transferable interactions have the form:

$$V_{contact} = V_{direct} + V_{water} \quad (10)$$

$$V_{direct} = \sum_{j-i>9} \gamma_{ij}(a_i, a_j) \Theta_{i,j}^I \quad (11)$$

$$V_{water}(i, j) = \sum_{j-i>9} \Theta_{i,j}^{II} (\sigma_{ij}^{wat} \gamma_{ij}^{wat}(a_i, a_j) + \sigma_{ij}^{prot} \gamma_{ij}^{protwat}(a_i, a_j)) \quad (12)$$

$$\Theta_{i,j}^\mu = \frac{1}{4} (1 + \tanh(\eta(r_{ij} - r_{min}^\mu))) (1 + \tanh(\eta(r_{max}^\mu - r_{ij}))) \quad (13)$$

$$\sigma_{ij}^{water} = \frac{1}{4} (1 - \tanh(\eta_\sigma(\rho_i - \rho_0))) (1 - \tanh(\eta_\sigma(\rho_j - \rho_0))) \quad (14)$$

$$\sigma_{ij}^{prot} = 1 - \sigma_{ij}^{water} \quad (15)$$

## 1.7 $\beta$ -hydrogen bonding and P-AP terms

We made some modification of these terms in order to make more efficient implementation of the force fields.

$$\theta_{i,j} = \exp\left(-\frac{(r_{O_i N_j} - r_{ON})^2}{2\sigma_{ON}^2} - \frac{(r_{O_i H_j} - r_{OH})^2}{2\sigma_{OH}^2}\right) \quad (16)$$

$$\theta_{j,i} = \exp\left(-\frac{(r_{O_j N_i} - r_{ON})^2}{2\sigma_{ON}^2} - \frac{(r_{O_j H_i} - r_{OH})^2}{2\sigma_{OH}^2}\right) \quad (17)$$

$$\theta_{j,i+2} = \exp\left(-\frac{(r_{O_j N_{i+2}} - r_{ON})^2}{2\sigma_{ON}^2} - \frac{(r_{O_j H_{i+2}} - r_{OH})^2}{2\sigma_{OH}^2}\right) \quad (18)$$

$$V1_{ij} = \lambda_1(i, j) \theta_{i,j} \quad (19)$$

$$V2_{ij} = \lambda_2(i, j) \theta_{i,j} \theta_{j,i} \quad (20)$$

$$V3_{ij} = \lambda_3(i, j) \theta_{i,j} \theta_{j,i+2} \quad (21)$$

$$V_{ij} = V1_{ij} + V2_{ij} + V3_{ij} \quad (22)$$

$$V_{beta} = -k_{beta} \sum_{ij} V_{ij} \quad (23)$$

In previous the LAMMPS implementation,  $V_{beta} = -k_{beta} \sum_{ij} V_{ij} v_i v_j$ , the additional term  $v_i v_j$  was used to ensure that the hydrogen bonds do not occur within a span of 5 residues that is shorter than  $12\text{\AA}$ . Now this constraint is incorporated onto the pap term. The  $V_{beta}$  defined here can be fit into the "CustomHbondForce" template. Since for  $V2_{ij}$ , we can define  $O_i, N_i, H_i$ , the oxygen, hydrogen and nitrogen of residue i as the donor, and  $N_j, H_j, O_j$  as the acceptor. We could have implemented the exact same version as the LAMMPS version using "CustomCompoundBondForce", but computing bonded forces is much slower than computing non-bonded forces like "CustomHbondForce". When two residues are far apart, computing their interaction is unnecessary.

$$v_i = \frac{1}{2}(1 + \tanh(\mu_1 * (r_{ca_i ca_{i+4}} - r_{CHB}))) \quad (24)$$

$$\theta_{i,j}^1 = \frac{1}{2}(1 + \tanh(\eta_{pap} * (r_0 - r_{ca_i n_j}))) \quad (25)$$

$$\theta_{i,j}^2 = \frac{1}{2}(1 + \tanh(\eta_{pap} * (r_0 - r_{ca_{i+4} n_{j+4}}))) \quad (26)$$

$$\theta_{i,j}^3 = \frac{1}{2}(1 + \tanh(\eta_{pap} * (r_0 - r_{ca_{i+4} n_{j-4}}))) \quad (27)$$

$$V_{i,j} = (\gamma_1(i, j) + \gamma_2(i, j)\theta_{i,j}^1\theta_{i,j}^2 + \gamma_3 i, j\theta_{i,j}^3)v_i \quad (28)$$

$$V_{pap} = \sum_{i,j} k_{pap} V_{i,j} \quad (29)$$

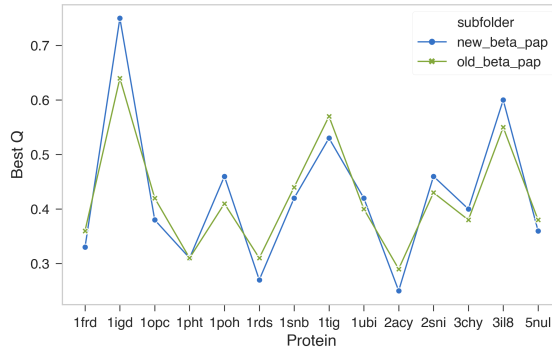


Figure 1: No significant different between structure prediction results using new and old beta hydrogen bonding term and pap term implementation.

## 2 Open3SPN2

The open3SPN2 software framework implements the 3SPN.2 [1] forcefields for A-DNA and B-DNA, and the 3SPN.2C [2] forcefield. The 3SPN.2 forcefield has been previously parametrized taking into account the free energy of nucleic acid hybridization, the intra strand base stacking energy, the DNA persistence length and the width of minor and major groves [1]. The 3SPN.2C forcefield is an extension of the 3SPN.2 forcefield that is able to reproduce sequence dependent curvature in the DNA [2].

In 3SPN.2 and 3SPN.2C each residue is represented by a three sites: a phosphate site (P), a sugar site (S) and a nucleobase site (B), where the nucleobase can be adenine (A), guanine (G), cytosine(C), or thymine (T). The 3SPN.2 potentials are the sum of eight terms (see equation 30). Three of this terms are bonded terms, which include a two-site bond term ( $V_{Bond}$ ), a three-site angle term ( $V_{Angle}$ ), and a four-site dihedral term ( $V_{Dihedral}$ ). Another three terms depend on the angles between nucleobases. Among them is a three-site stacking term ( $V_{Stacking}$ ) between consecutive nucleotides, a four-site basepairing term ( $V_{BasePair}$ ) between complementary nucleobases, and a five-site cross-stacking

term ( $V_{CrossStacking}$ ). The last two non-bonded terms depend only on the pairwise distances between sites and include an exclusion term ( $V_{Exclusion}$ ) and an electrostatics term ( $V_{Electrostatics}$ ).

$$V_{3SPN2} = V_{Bond} + V_{Angle} + V_{Dihedral} + V_{Stacking} + V_{BasePair} + V_{CrossStacking} + V_{Exclusion} + V_{Electrostatics} \quad (30)$$

For 3SPN.2C a reference atomistic structure needs to be created using the 3DNA software [3]. The reference structure is given by a set of base-step and base-pair geometric parameters (Tables 2 and 3) suited for protein-DNA binding [2]. The base step-parameters depend on the type of the base ( $B_o$ ) and the type of the neighboring sequence-adjacent base ( $B_n$ ). The base-pair parameters depend only on the type of the base ( $B_o$ ), since we expect a Watson-Crick basepair ( $B_p$ ). After the atomistic reference structure is created, the structure is Coarse Grained and the distances, angles and dihedrals from the structure will become the equilibrium distances, angles and dihedrals for the 3SPN.2C forcefield.

Table 2: open3SPN2 base-step reference geometric parameters

$B_o$	$B_n$	twist ( $^\circ$ )	roll ( $^\circ$ )	tilt ( $^\circ$ )	shift ( $\text{\AA}$ )	slide ( $\text{\AA}$ )	rise ( $\text{\AA}$ )
A	A	35.31	0.76	-1.84	-0.05	-0.21	3.27
A	T	31.21	-1.39	0	0	-0.56	3.39
A	C	31.52	0.91	-0.64	0.21	-0.54	3.39
A	G	33.05	3.15	-1.48	0.12	-0.27	3.38
T	A	36.2	5.25	0	0	0.03	3.34
T	T	35.31	0.76	1.84	0.05	-0.21	3.27
T	C	34.8	3.87	1.52	0.27	-0.03	3.35
T	G	35.02	5.95	0.05	0.16	0.18	3.38
C	A	35.02	5.95	-0.05	-0.16	0.18	3.38
C	T	33.05	3.15	1.48	-0.12	-0.27	3.38
C	C	33.17	3.86	0.4	0.02	-0.47	3.28
C	G	35.30	4.29	0	0	0.57	3.49
G	A	34.8	3.87	-1.52	-0.27	-0.03	3.35
G	T	31.52	0.91	0.64	-0.21	-0.54	3.39
G	C	34.38	0.67	0	0	-0.07	3.38
G	G	33.17	3.86	-0.4	-0.02	-0.47	3.28

Table 3: open3SPN2 base-pair reference geometric parameters

$B_o$	$B_p$	buckle ( $^\circ$ )	propeller ( $^\circ$ )	opening ( $^\circ$ )	shear ( $\text{\AA}$ )	stretch ( $\text{\AA}$ )	stagger ( $\text{\AA}$ )
A	T	1.8	-15	1.5	0.07	-0.19	0.07
T	A	-1.8	-15	1.5	-0.07	-0.19	0.07
C	G	-4.9	-8.7	-0.6	0.16	-0.17	0.15
G	C	4.9	-8.7	-0.6	-0.16	-0.17	0.15

## 2.1 Bonded terms

The bond term is a quartic function of the pairwise distance between two sites that doesn't include the cubic term. The coefficient for the quartic term is 100 times greater per  $\text{\AA}^2$  than the coefficient for the harmonic term (Eq. 31). The quartic function allows the bond to have a wider well than an harmonic potential with a comparable coefficient (Figure 2).

$$V_{Bond} = \sum_i^{Bonds} k_{b_i} (r_{b_i} - r_{b_i}^o)^2 + 100k_{b_i} (r_{b_i} - r_{b_i}^o)^4 \quad (31)$$

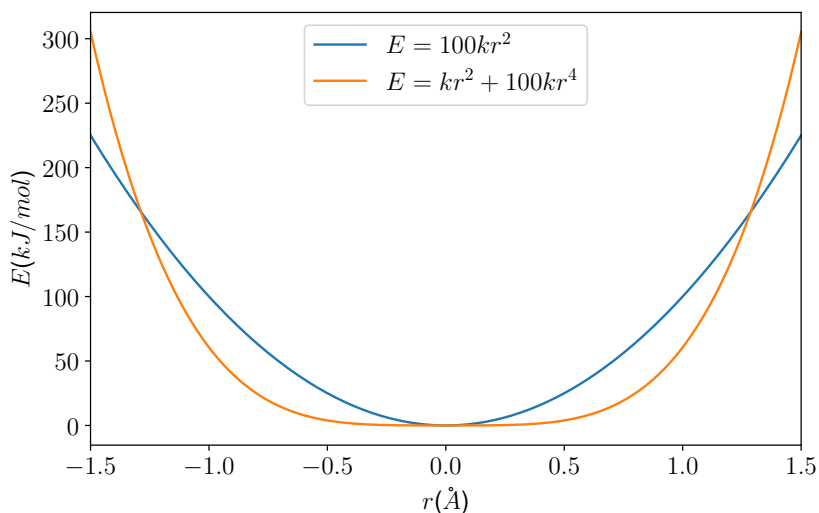


Figure 2: The harmonic potential function is shown in blue, compared with the open3SPN2 quartic potential function with  $k_b = 0.143403 \text{kcal/mol/\AA}$ . The open3SPN2 bond potential function shows a wider well than a comparable harmonic potential function.

There are 6 types of bonds defined for each forcefield: a bond from a phosphate ( $P$ ) to a sugar ( $S$ ), a bond between a sugar ( $S$ ) and the phosphate of the next residue ( $P_1$ ), and four bonds from the sugar to the nucleobase that depend on the nucleobase type (Figure 3).

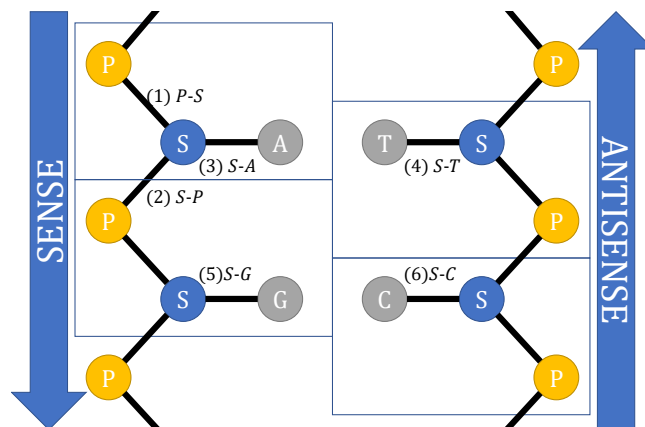


Figure 3: List of bonds in the 3SPN.2 and 3SPN.2C forcefields. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site (A, C, T, or G). The 6 types of bonds are listed from 1 to 6.

The parameters of the bonds are listed on the table 4, where  $i$  is the first site type,  $j$  is the second site type,  $r_b^o$  is the equilibrium distance of the bond, and  $k_b$  is the coefficient for the harmonic term. In 3SPN.2C the equilibrium distances ( $r_b^o$ ) are computed from a reference structure generated using the equilibrium base-pair and base-step parameters, so they are not shown in the table.

Table 4: open3SPN2 bond parameters

<b>Forcefield</b>	<i>i</i>	<i>j</i>	$r_b^o(\text{\AA})$	$k_b(\text{kcal/mol/\AA}^2)$
3SPN.2 (A-DNA)	P	S	4.157	0.143 403
3SPN.2 (A-DNA)	S	$P_1$	3.78	0.143 403
3SPN.2 (A-DNA)	S	A	4.697	0.143 403
3SPN.2 (A-DNA)	S	T	4.22	0.143 403
3SPN.2 (A-DNA)	S	G	4.852	0.143 403
3SPN.2 (A-DNA)	S	C	4.066	0.143 403
3SPN.2 (B-DNA)	P	S	3.899	0.143 403
3SPN.2 (B-DNA)	S	$P_1$	3.559	0.143 403
3SPN.2 (B-DNA)	S	A	4.67	0.143 403
3SPN.2 (B-DNA)	S	T	4.189	0.143 403
3SPN.2 (B-DNA)	S	G	4.829	0.143 403
3SPN.2 (B-DNA)	S	C	4.112	0.143 403
3SPN.2C	P	S	—	0.143 403
3SPN.2C	S	$P_1$	—	0.143 403
3SPN.2C	S	A	—	0.143 403
3SPN.2C	S	T	—	0.143 403
3SPN.2C	S	G	—	0.143 403
3SPN.2C	S	C	—	0.143 403

<sup>1</sup> The suffix 1 in the names of the sites indicates that the site is part of the next residue.

— The equilibrium distances ( $r_b^o$ ) for the 3SPN.2C forcefield is sequence dependent and computed from a template created based on the geometric parameters.

The angle term is an harmonic function of the angle  $\theta_a$  between 3 sites  $i$ ,  $j$  and  $k$ , where  $j$  is the center site. The term coefficient ( $k_a$ ) is  $200 \text{ kJ/mol/rad}^2$  in the 3SPN.2 forcefield. There are 10 possible angles: P-S-P, S-P-S, P-S-B and B-P-S, where B can be any nucleobase (A,C,T,G) (Figure 4 and their parameters are listed in the table 5).

$$V_{Angle} = \sum_i^{Angles} k_{ai}(\theta_{ai} - \theta_{ai}^o)^2 \quad (32)$$

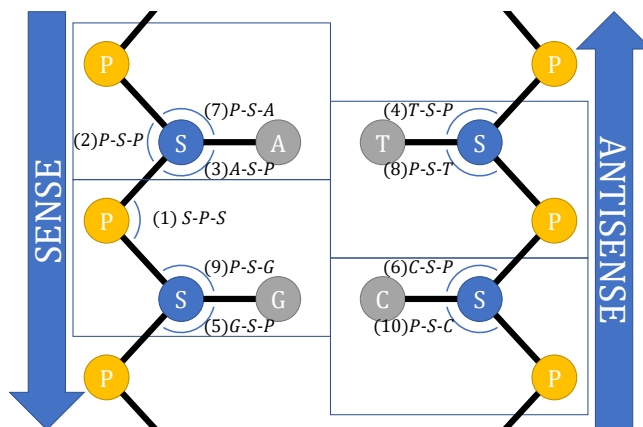


Figure 4: List of angles in the 3SPN.2 and 3SPN.2C forcefields. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site (A, C, T, or G). The 10 types of angles are listed from 1 to 10.

Table 5: Open3SPN.2 angle parameters

Forcefield	i	j	k	$\theta_a^\circ$ ( $^\circ$ )
3SPN.2 (A-DNA)	S	$P_1$	$S_1$	92.77
3SPN.2 (A-DNA)	P	S	$P_1$	91.24
3SPN.2 (A-DNA)	A	S	$P_1$	104.86
3SPN.2 (A-DNA)	T	S	$P_1$	110.58
3SPN.2 (A-DNA)	G	S	$P_1$	103.86
3SPN.2 (A-DNA)	C	S	$P_1$	106.94
3SPN.2 (A-DNA)	P	S	A	103.71
3SPN.2 (A-DNA)	P	S	T	93.27
3SPN.2 (A-DNA)	P	S	G	107.49
3SPN.2 (A-DNA)	P	S	C	97.58
3SPN.2 (B-DNA)	S	$P_1$	$S_1$	94.49
3SPN.2 (B-DNA)	P	S	$P_1$	120.15
3SPN.2 (B-DNA)	A	S	$P_1$	112.07
3SPN.2 (B-DNA)	T	S	$P_1$	116.68
3SPN.2 (B-DNA)	G	S	$P_1$	110.12
3SPN.2 (B-DNA)	C	S	$P_1$	114.34
3SPN.2 (B-DNA)	P	S	A	103.53
3SPN.2 (B-DNA)	P	S	T	92.06
3SPN.2 (B-DNA)	P	S	G	107.4
3SPN.2 (B-DNA)	P	S	C	96.96

In 3SPN.2C the equilibrium angles ( $\theta_a^\circ$ ) are computed from a reference structure generated using the equilibrium base-pair and base-step parameters. The equilibrium constant also depends on the nucleobase type ( $B_o$ ) and the neighboring bases ( $B_n$ ) as shown in the table 7.

Table 7: Open3SPN.2C angle parameters

Forcefield	i	j	k	$k_a$ kcal/mol/rad <sup>2</sup>	$\theta_a$ (°)	$B_o$	$B_n$
3SPN.2C	S	$P_1$	$S_1$	355	—	A	A
3SPN.2C	S	$P_1$	$S_1$	464	—	A	C
3SPN.2C	S	$P_1$	$S_1$	368	—	A	G
3SPN.2C	S	$P_1$	$S_1$	147	—	A	T
3SPN.2C	S	$P_1$	$S_1$	273	—	C	A
3SPN.2C	S	$P_1$	$S_1$	165	—	C	C
3SPN.2C	S	$P_1$	$S_1$	478	—	C	G
3SPN.2C	S	$P_1$	$S_1$	368	—	C	T
3SPN.2C	S	$P_1$	$S_1$	442	—	G	A
3SPN.2C	S	$P_1$	$S_1$	228	—	G	C
3SPN.2C	S	$P_1$	$S_1$	165	—	G	G
3SPN.2C	S	$P_1$	$S_1$	464	—	G	T
3SPN.2C	S	$P_1$	$S_1$	230	—	T	A
3SPN.2C	S	$P_1$	$S_1$	442	—	T	C
3SPN.2C	S	$P_1$	$S_1$	273	—	T	G
3SPN.2C	S	$P_1$	$S_1$	355	—	T	T
3SPN.2C	P	S	$P_1$	300	—	any	any
3SPN.2C	A	S	$P_1$	460	—	A	A
3SPN.2C	A	S	$P_1$	442	—	A	C
3SPN.2C	A	S	$P_1$	358	—	A	G
3SPN.2C	A	S	$P_1$	370	—	A	T
3SPN.2C	T	S	$P_1$	120	—	T	A
3SPN.2C	T	S	$P_1$	383	—	T	C
3SPN.2C	T	S	$P_1$	206	—	T	G
3SPN.2C	T	S	$P_1$	460	—	T	T
3SPN.2C	G	S	$P_1$	383	—	G	A
3SPN.2C	G	S	$P_1$	336	—	G	C
3SPN.2C	G	S	$P_1$	278	—	G	G
3SPN.2C	G	S	$P_1$	442	—	G	T
3SPN.2C	C	S	$P_1$	206	—	C	A
3SPN.2C	C	S	$P_1$	278	—	C	C
3SPN.2C	C	S	$P_1$	278	—	C	G
3SPN.2C	C	S	$P_1$	358	—	C	T
3SPN.2C	P	S	A	460	—	A	$A_{-1}$
3SPN.2C	P	S	A	206	—	A	$C_{-1}$
3SPN.2C	P	S	A	383	—	A	$G_{-1}$
3SPN.2C	P	S	A	120	—	A	$T_{-1}$
3SPN.2C	P	S	T	370	—	T	$A_{-1}$
3SPN.2C	P	S	T	358	—	T	$C_{-1}$
3SPN.2C	P	S	T	442	—	T	$G_{-1}$
3SPN.2C	P	S	T	460	—	T	$T_{-1}$
3SPN.2C	P	S	G	358	—	G	$A_{-1}$
3SPN.2C	P	S	G	278	—	G	$C_{-1}$
3SPN.2C	P	S	G	278	—	G	$G_{-1}$
3SPN.2C	P	S	G	206	—	G	$T_{-1}$
3SPN.2C	P	S	C	442	—	C	$A_{-1}$



Table 7: Open3SPN.2C angle parameters (continued)

Forcefield	i	j	k	$k_a(kJ/mol)$	$\theta_a^\circ$ ( $^\circ$ )	$B_o$	$B_n$
3SPN.2C	P	S	C	278	—	C	$C_{-1}$
3SPN.2C	P	S	C	336	—	C	$G_{-1}$
3SPN.2C	P	S	C	383	—	C	$T_{-1}$

1 The suffix 1 in the names of the sites indicates that the site is part of the next residue.

-1 The neighboring base on the 5' direction or behind in the sequence.

— The equilibrium angles ( $\theta_a^\circ$ ) for the 3SPN.2C forcefield is sequence dependent and computed from a template based on the geometric parameters.

The open3SPN2 forcefield includes two dihedral potentials, a gaussian potential and a cosine potential (Eq. 33).

$$V_{Dihedral} = \sum_i^{Dihedrals} -k_G i e^{-\frac{-(\phi_i - \phi_i^o)^2}{2\sigma_i^2}} + k_C i (1 - \cos(\phi_i - \phi_i^o)) \quad (33)$$

Where  $k_G$  is the coefficient for the gaussian potential,  $k_C$  is the coefficient for the cosine potential,  $\phi$  is the dihedral angle between the sites i, j, k, and l. The parameters are listed in the table 8.

In 3SPN.2 only the gaussian potential is used, while in 3SPN.2C a mixture of the gaussian potential and the cosine potential is used for the dihedrals S-P-S-P and P-S-P-S. 3SPN.2C also adds a dihedral potential for the dihedrals B-S-P-S and S-P-S-B, where B can be any nucleobase (Figure 5). In 3SPN.2C the equilibrium angles ( $\phi_D^o$ ) are computed from a template structure generated using the equilibrium basepair and base stacking parameters.

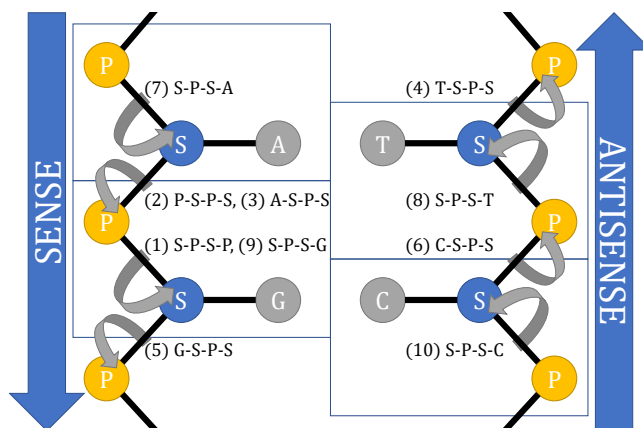


Figure 5: List of dihedrals in the 3SPN.2 and 3SPN.2C forcefields. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site (A, C, T, or G). The 10 types of dihedrals are listed from 1 to 10. 3SPN.2 only includes the dihedrals (1) and (2).

Table 8: Open3SPN.2C dihedral parameters

Forcefield	i	j	k	l	$K_C(kcal/mol)$	$K_G(kcal/mol)$	$\sigma(rad)$	$\phi^o (^\circ)$
3SPN.2 (A-DNA)	S	$P_1$	$S_1$	$P_2$	0	1.434034	0.3	-9.58
3SPN.2 (A-DNA)	P	S	$P_1$	$S_1$	0	1.434034	0.3	-328.4
3SPN.2 (B-DNA)	S	$P_1$	$S_1$	$P_2$	0	1.434034	0.3	-359.17
3SPN.2 (B-DNA)	P	S	$P_1$	$S_1$	0	1.434034	0.3	-334.79
3SPN.2C	S	$P_1$	$S_1$	$P_2$	0.478011	1.67304	0.3	—
3SPN.2C	P	S	$P_1$	$S_1$	0.478011	1.67304	0.3	—
3SPN.2C	A	S	$P_1$	$S_1$	0.478011	0	0.3	—
3SPN.2C	T	S	$P_1$	$S_1$	0.478011	0	0.3	—
3SPN.2C	G	S	$P_1$	$S_1$	0.478011	0	0.3	—
3SPN.2C	C	S	$P_1$	$S_1$	0.478011	0	0.3	—
3SPN.2C	S	$P_1$	$S_1$	$A_1$	0.478011	0	0.3	—
3SPN.2C	S	$P_1$	$S_1$	$T_1$	0.478011	0	0.3	—
3SPN.2C	S	$P_1$	$S_1$	$G_1$	0.478011	0	0.3	—
3SPN.2C	S	$P_1$	$S_1$	$C_1$	0.478011	0	0.3	—

1 The suffix 1 in the names of the sites indicates that the site is part of the next residue.

— The equilibrium dihedral angles ( $\phi^o$ ) for the 3SPN.2C forcefield is sequence dependent and computed from a template based on the geometric parameters.

## 2.2 Stacking, BasePairing and CrossStacking terms

The Stacking, BasePairing and CrossStacking terms are non-bonded terms that depend on the distance between the nucleobases ( $r_{BS}$ ,  $r_{BP}$ , and  $r_{CS}$  respectively), as well as angles defined between the residues. All the terms include a modulating function ( $f$ ) of an angle ( $\theta$ ). The modulating function can be understood as depending in the position of the second nucleobase relative to two cones in 3D space. If the second nucleobase is inside the interior cone, the modulating function is equal to 1, and if it is outside the cone, the modulating function is equal to 0. Between this two cones the modulating function has a value between 1 and 0 that depends on the angle (Eq 34). The coefficient  $K$  defines the width of the conical section.

$$f(\theta|K, \theta^o) = \begin{cases} 1 & \frac{\pi}{2} \geq |K(\theta - \theta^o)| \\ 1 - \cos^2(K(\theta - \theta^o)) & \frac{\pi}{2} < |K(\theta - \theta^o)| \leq \pi \\ 0 & \pi < |K(\theta - \theta^o)| \end{cases} \quad (34)$$

For the stacking term  $K_{BS} = 6$ , which defines a inner cone of 30 degrees and an outer cone of 60 degrees. The stacking potential is a mixture of a repulsive potential and an attractive potential. The depth of the attractive well is  $\epsilon$  and fluctuates with the modulating function. The steepness of the repulsive potential is  $\alpha_{BS} = 3\text{\AA}^{-2}$ . The parameters are listed in the table 9.

$$F_{BS}(\theta_{BS}) = f(\theta_{BS}|K_{BS}, \theta_{BS}^o) \quad (35)$$

$$V_{BS} = \sum_i \begin{cases} \epsilon_i(1 - e^{-\alpha_{BS}(r_{BSi} - r_{BSi}^o)^2}) - \epsilon_i F_{BS}(\theta_{BSi}) & , r_{BSi} < r_{BSi}^o \\ \epsilon_i(1 - e^{-\alpha_{BS}(r_{BSi} - r_{BSi}^o)^2}) F_{BS}(\theta_{BSi}) - \epsilon_i F_{BS}(\theta_{BSi}) & , r_{BSi} \geq r_{BSi}^o \end{cases} \quad (36)$$

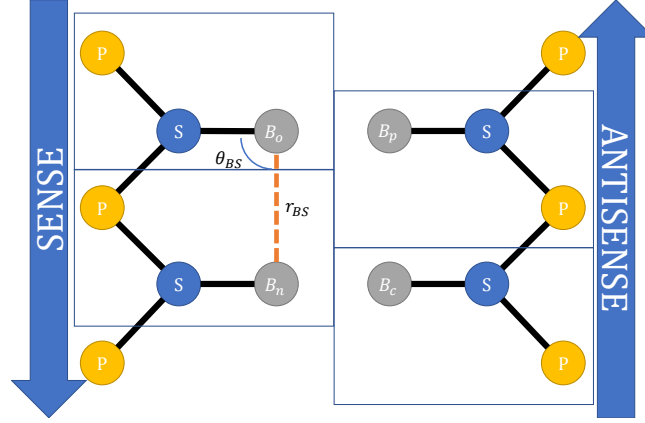


Figure 6: Important variables used for the stacking term. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site. The nucleobases shown are the reference nucleobase ( $B_o$ ), the neighboring nucleobase ( $B_n$ ), the base-pairing nucleobase ( $B_p$ ) and the cross-stacking nucleobase ( $B_c$ ). The variables shown are the distance between  $B_o$  and  $B_n$  ( $r_{BS}$ ), and the angle between S,  $B_o$ , and  $B_n$  ( $\theta_{BS}$ ).

Table 9: open3SPN2 base stacking parameters

DNA	$B_o$	$B_n$	$\epsilon(kcal/mol)$	$r_{BS}^o(\text{\AA})$	$\theta_{BS}^o(^{\circ})$
3SPN.2 (A-DNA)	A	A	3.439293	4.022	108.32
3SPN.2 (A-DNA)	A	T	3.427342	3.344	96.74
3SPN.2 (A-DNA)	A	G	3.166826	4.261	111.32
3SPN.2 (A-DNA)	A	C	3.467973	3.737	97.36
3SPN.2 (A-DNA)	T	A	2.478489	4.794	103.33
3SPN.2 (A-DNA)	T	T	3.193117	4.031	94.85
3SPN.2 (A-DNA)	T	G	2.471319	5.064	105.36
3SPN.2 (A-DNA)	T	C	3.080784	4.445	94.51
3SPN.2 (A-DNA)	G	A	3.539675	3.855	108.25
3SPN.2 (A-DNA)	G	T	3.721319	3.217	95.59
3SPN.2 (A-DNA)	G	G	3.568356	4.077	111.66
3SPN.2 (A-DNA)	G	C	3.678298	3.592	96.71
3SPN.2 (A-DNA)	C	A	2.729446	4.499	111.39
3SPN.2 (A-DNA)	C	T	3.056883	3.708	102.73
3SPN.2 (A-DNA)	C	G	2.51434	4.772	113.47
3SPN.2 (A-DNA)	C	C	3.164436	4.116	102.14
3SPN.2 (B-DNA)	A	A	3.439293	3.716	101.15
3SPN.2 (B-DNA)	A	T	3.427342	3.675	85.94

3SPN.2 (B-DNA)	A	G	3.166826	3.827	105.26
3SPN.2 (B-DNA)	A	C	3.467973	3.744	89
3SPN.2 (B-DNA)	T	A	2.478489	4.238	101.59
3SPN.2 (B-DNA)	T	T	3.193117	3.984	89.5
3SPN.2 (B-DNA)	T	G	2.471319	4.416	104.31
3SPN.2 (B-DNA)	T	C	3.080784	4.141	91.28
3SPN.2 (B-DNA)	G	A	3.539675	3.576	100.89
3SPN.2 (B-DNA)	G	T	3.721319	3.598	84.83
3SPN.2 (B-DNA)	G	G	3.568356	3.664	105.48
3SPN.2 (B-DNA)	G	C	3.678298	3.635	88.28
3SPN.2 (B-DNA)	C	A	2.729446	4.038	106.49
3SPN.2 (B-DNA)	C	T	3.056883	3.798	93.31
3SPN.2 (B-DNA)	C	G	2.51434	4.208	109.54
3SPN.2 (B-DNA)	C	C	3.164436	3.935	95.46
3SPN.2C	A	A	3.303059	3.58	100.13
3SPN.2C	A	T	3.597036	3.56	90.48
3SPN.2C	A	G	3.183556	3.85	104.39
3SPN.2C	A	C	3.781071	3.45	93.23
3SPN.2C	T	A	2.186902	4.15	102.59
3SPN.2C	T	T	2.973231	3.93	93.32
3SPN.2C	T	G	2.289675	4.32	103.7
3SPN.2C	T	C	3.133365	3.87	94.55
3SPN.2C	G	A	3.288719	3.51	95.45
3SPN.2C	G	T	3.487094	3.47	87.63
3SPN.2C	G	G	3.530115	3.67	106.36
3SPN.2C	G	C	3.625717	3.42	83.12
3SPN.2C	C	A	2.210803	4.15	102.69
3SPN.2C	C	T	2.968451	3.99	96.05
3SPN.2C	C	G	2.110421	4.34	100.46
3SPN.2C	C	C	3.34847	3.84	100.68

The modulating function for the base pair term depends on the angles between the sugar (S), the base( $B_o$ ) and the complementary base( $B_n$ ). There are two angles that can be defined on this way( $\theta_{BP1}, \theta_{BP2}$ ). It also depends on the cosine of the dihedral between both sugars and bases (S-B-B-S) ( $\phi_{BP}$ ). The cone for basepairing is much narrower ( $K_{BP} = 12$ ) where the inner cone is 15 degrees and the outer cone is 30 degrees.

$$F_{BP} = \frac{1 + \cos(\phi_{BP})}{2} f(\theta_{BP1} | K_{BP}, \theta_{BP1}^o) f(\theta_{BP2} | K_{BP}, \theta_{BP2}^o) \quad (37)$$

The steepness parameter,  $\alpha_{BP}$ , is  $2nm^{-2}$ . The parameters are listed on the table 10.

$$V_{BP} = \sum_i \begin{cases} \epsilon_i (1 - e^{-\alpha_{BP}(r_{BPi} - r_{BPi}^o)^2}) - \epsilon_i F_{BP} & , r_{BPi} < r_{BPi}^o \\ \epsilon_i (1 - e^{-\alpha_{BP}(r_{BPi} - r_{BPi}^o)^2}) F_{BP} - \epsilon_i F_{BP} & , r_{BPi} \geq r_{BPi}^o \end{cases} \quad (38)$$

Table 10: open3SPN2 basepairing parameters

Forcefield	$B_o$	$B_p$	$r_{BP}^o$ (Å)	$\epsilon_{BP}$ (kcal/mol)	$\phi_{BP}$ (°)	$\theta_{BP1}$ (°)	$\theta_{BP2}$ (°)
3SPN.2 (A-DNA)	A	T	5.861	3.99874	50.17	160.91	140.49
3SPN.2 (A-DNA)	G	C	5.528	5.06241	38.33	165.25	147.11
3SPN.2 (B-DNA)	A	T	5.941	3.99874	-38.35	156.54	135.78
3SPN.2 (B-DNA)	G	C	5.530	5.06241	-42.98	159.81	141.16
3SPN.2C	A	T	5.82	3.44292	-38.18	153.17	133.51
3SPN.2C	G	C	5.52	4.35873	-35.75	159.5	138.08

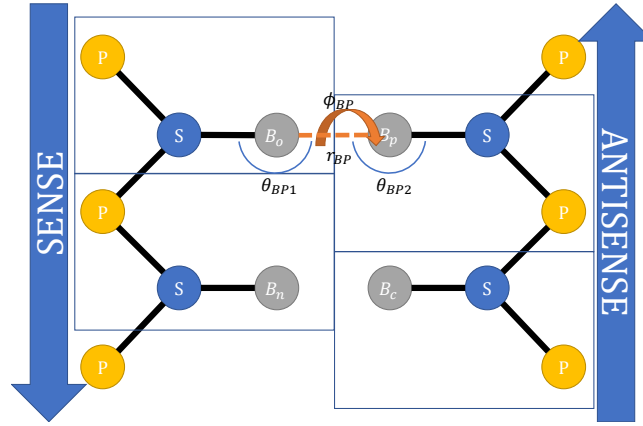


Figure 7: Important variables used for the base-pairing term. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site. The nucleobases shown are the reference nucleobase ( $B_o$ ), the neighboring nucleobase ( $B_n$ ), the base-pairing nucleobase ( $B_p$ ) and the cross-stacking nucleobase ( $B_c$ ). The variables shown are the distance between  $B_o$  and  $B_p$  ( $r_{BP}$ ), the angle between the sugar from the the reference nucleotide,  $B_o$ , and  $B_p$  ( $\theta_{BP1}$ ), the angle between the sugar from the the base-pairing nucleotide,  $B_p$ , and  $B_o$  ( $\theta_{BP2}$ ), and the dihedral between the sugar from the the reference nucleotide,  $B_o$ , and  $B_p$  and the sugar from the the base-pairing nucleotide ( $\phi_{BP}$ ).

The CrossStacking potential has only an attractive potential. the modulating function depends on the angle between the sugar, the base, and the cross-stacking base. It also depends on the vector angle between defined between the sugar-base vectors.  $K$  for the first angle is 8, while for the second angle is 12.  $\alpha$  is  $2nm^{-2}$ .

$$V_{CS} = \sum_i \begin{cases} -\epsilon_i F_{CS} & , r_{CSi} < r_{CSi}^o \\ \epsilon_i (1 - e^{-\alpha_i (r_{CSi} - r_{CSi}^o)^2}) F_{CS} - \epsilon_i F_{CS} & , r_{CSi} \geq r_{CSi}^o \end{cases} \quad (39)$$

$$F_{CS} = f(\phi_{CS} | K_{BP}, \phi_{CS}^o) f(\theta_{CS} | K_{CS}, \theta_{CS}^o) \quad (40)$$

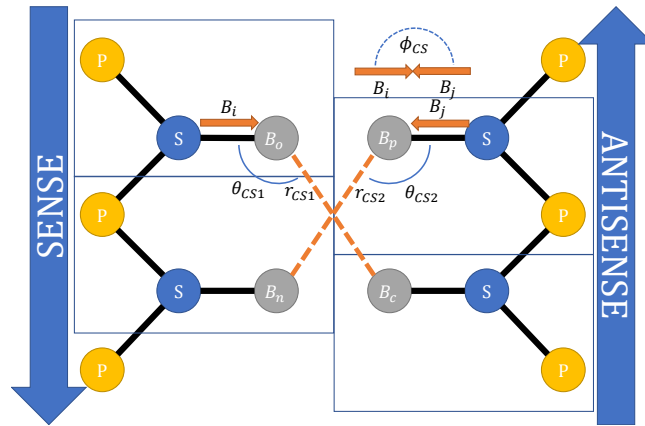


Figure 8: Important variables used for the cross-stacking term. Each square encloses a residue, which contains a phosphate site (P), a sugar site (S) and a nucleobase site. The nucleobases shown are the reference nucleobase ( $B_o$ ), the neighboring nucleobase ( $B_n$ ), the base-pairing nucleobase ( $B_p$ ) and the cross-stacking nucleobase ( $B_c$ ). Some variables shown are the distance between  $B_o$  and  $B_c$  ( $r_{CS1}$ ), the angle between S,  $B_o$ , and  $B_c$  ( $\theta_{CS1}$ ). This variables are also mirrored for the base-pairing nucleotide. Also shown are the vectors  $B_i$  and  $B_j$ , which are defined as the vectors originating from the sugar to the nucleobase of the reference nucleotide and the base-pairing nucleotide respectively.  $\phi_{CS}$  is the vector angle between  $B_i$  and  $B_j$ .

Table 11: open3SPN2 CrossStacking parameters

Forcefield	$B_o$	$B_p$	$B_c$	$\phi_{CS}^\circ$ (°)	$\theta_{CS1}$ (°)	$\theta_{CS2}$ (°)	$r_{CS1}$ (Å)	$r_{CS2}$ (Å)	$\epsilon_{CS1}$ (kcal/mol)	$\epsilon_{CS2}$ (kcal/mol)
3SPN.2 (A-DNA)	A	T	A	126.57	147.44	130.5	7.344	4.624	0.522452	0.522452
3SPN.2 (A-DNA)	A	T	T	126.57	148.97	138.73	8.081	5.095	0.662942	0.662942
3SPN.2 (A-DNA)	A	T	G	126.57	146.21	126.68	7.187	4.464	0.677075	0.712197
3SPN.2 (A-DNA)	A	T	C	126.57	150.17	134.18	7.99	5.162	0.46634	0.60683
3SPN.2 (A-DNA)	T	A	A	126.57	138.42	130.41	8.081	5.095	0.662942	0.662942
3SPN.2 (A-DNA)	T	A	T	126.57	141.67	134.68	8.755	5.693	0.522452	0.522452
3SPN.2 (A-DNA)	T	A	G	126.57	136.64	127.69	7.952	4.896	0.60683	0.46634
3SPN.2 (A-DNA)	T	A	C	126.57	141.64	131.38	8.697	5.724	0.712197	0.677075
3SPN.2 (A-DNA)	G	C	A	134.71	147.67	130.57	7.187	4.464	0.677075	0.712197
3SPN.2 (A-DNA)	G	C	T	134.71	148.28	140.17	7.952	4.896	0.60683	0.46634
3SPN.2 (A-DNA)	G	C	G	134.71	146.84	126.44	7.019	4.315	0.901943	1.1478
3SPN.2 (A-DNA)	G	C	C	134.71	150.02	135.31	7.844	4.968	0.269738	0.269738
3SPN.2 (A-DNA)	C	G	A	134.71	145.83	132.69	7.99	5.162	0.46634	0.60683
3SPN.2 (A-DNA)	C	G	T	134.71	148.39	138.21	8.697	5.724	0.712197	0.677075
3SPN.2 (A-DNA)	C	G	G	134.71	144.24	129.73	7.844	4.968	0.269738	0.269738
3SPN.2 (A-DNA)	C	G	C	134.71	148.74	134.45	8.63	5.759	1.1478	0.901943
3SPN.2 (B-DNA)	A	T	A	116.09	154.38	116.88	6.208	5.435	0.522452	0.522452
3SPN.2 (B-DNA)	A	T	T	116.09	159.1	121.74	6.876	6.295	0.662942	0.662942
3SPN.2 (B-DNA)	A	T	G	116.09	152.46	114.23	6.072	5.183	0.677075	0.712197
3SPN.2 (B-DNA)	A	T	C	116.09	158.38	119.06	6.811	6.082	0.46634	0.60683
3SPN.2 (B-DNA)	T	A	A	116.09	147.1	109.42	6.876	6.295	0.662942	0.662942
3SPN.2 (B-DNA)	T	A	T	116.09	153.79	112.95	7.48	7.195	0.522452	0.522452
3SPN.2 (B-DNA)	T	A	G	116.09	144.44	107.32	6.771	6.028	0.60683	0.46634
3SPN.2 (B-DNA)	T	A	C	116.09	151.48	110.56	7.453	6.981	0.712197	0.677075
3SPN.2 (B-DNA)	G	C	A	124.93	154.69	119.34	6.072	5.183	0.677075	0.712197
3SPN.2 (B-DNA)	G	C	T	124.93	157.83	124.72	6.771	6.028	0.60683	0.46634
3SPN.2 (B-DNA)	G	C	G	124.93	153.43	116.51	5.921	4.934	0.901943	1.1478
3SPN.2 (B-DNA)	G	C	C	124.93	158.04	121.98	6.688	5.811	0.269738	0.269738
3SPN.2 (B-DNA)	C	G	A	124.93	152.99	114.6	6.811	6.082	0.46634	0.60683
3SPN.2 (B-DNA)	C	G	T	124.93	159.08	118.26	7.453	6.981	0.712197	0.677075
3SPN.2 (B-DNA)	C	G	G	124.93	150.53	112.45	6.688	5.811	0.269738	0.269738
3SPN.2 (B-DNA)	C	G	C	124.93	157.17	115.88	7.409	6.757	1.1478	0.901943
3SPN.2C	A	T	A	110.92	154.04	116.34	6.42	5.58	0.449831	0.449831
3SPN.2C	A	T	T	110.92	158.77	119.61	6.77	6.14	0.570793	0.570793
3SPN.2C	A	T	G	110.92	153.88	115.19	6.27	5.63	0.582961	0.613202
3SPN.2C	A	T	C	110.92	157.69	120.92	6.84	6.18	0.401518	0.52248
3SPN.2C	T	A	A	110.92	148.62	107.4	6.77	6.14	0.570793	0.570793
3SPN.2C	T	A	T	110.92	155.05	110.76	7.21	6.8	0.449831	0.449831
3SPN.2C	T	A	G	110.92	147.54	106.33	6.53	6.07	0.52248	0.401518
3SPN.2C	T	A	C	110.92	153.61	111.57	7.08	6.64	0.613202	0.582961
3SPN.2C	G	C	A	120.45	153.91	121.61	6.27	5.63	0.582961	0.613202
3SPN.2C	G	C	T	120.45	155.72	124.92	6.53	6.07	0.52248	0.401518
3SPN.2C	G	C	G	120.45	151.84	120.52	5.74	5.87	0.776573	0.988256
3SPN.2C	G	C	C	120.45	157.8	124.88	6.86	5.66	0.232244	0.232244
3SPN.2C	C	G	A	120.45	152.04	112.45	6.84	6.18	0.401518	0.52248
3SPN.2C	C	G	T	120.45	157.72	115.43	7.08	6.64	0.613202	0.582961
3SPN.2C	C	G	G	120.45	151.65	110.51	6.86	5.66	0.232244	0.232244
3SPN.2C	C	G	C	120.45	154.49	115.8	6.79	6.8	0.988256	0.776573

### 2.3 Non-bonded terms

The exclusion potential contains the repulsive section of a Lennard-Jones potential.

$$V_{Exclusion} = \sum_{ij} \begin{cases} \epsilon_r \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \epsilon_r & , r < \sigma_{ij} \\ 0 & , r \geq \sigma_{ij} \end{cases} \quad (41)$$

The electrostatics potential is based on the Debye-Huckel potential.

$$V_{Electrostatics} = \sum_{ij} \frac{q_i q_j e^{-\frac{r_{ij}}{\lambda_D}}}{4\pi\epsilon_0\epsilon(T, C) r_{ij}} \quad (42)$$

The dielectric coefficient depends on the temperature and the concentration of ions.

$$\lambda_D = \frac{\sqrt{\epsilon_0\epsilon(T, C) r_{i,j}}}{2\beta N_A e_c^2 I} \quad (43)$$

Table 12: open3SPN2 Exclusion and electrostatics parameters

Forcefield	Particle	$\epsilon$ (kcal/mol)	$r$ ( $\text{\AA}$ )	mass (Da)	charge ( $e$ )
3SPN.2 (A-DNA)	P	0.239006	4.5	94.9696	-0.6
3SPN.2 (A-DNA)	S	0.239006	6.2	83.1104	0
3SPN.2 (A-DNA)	A	0.239006	4.46	134.122	0
3SPN.2 (A-DNA)	T	0.239006	5.5	125.1078	0
3SPN.2 (A-DNA)	G	0.239006	4.2	150.1214	0
3SPN.2 (A-DNA)	C	0.239006	5.7	110.0964	0
3SPN.2 (B-DNA)	P	0.239006	4.5	94.9696	-0.6
3SPN.2 (B-DNA)	S	0.239006	6.2	83.1104	0
3SPN.2 (B-DNA)	A	0.239006	5.4	134.122	0
3SPN.2 (B-DNA)	T	0.239006	7.1	125.1078	0
3SPN.2 (B-DNA)	G	0.239006	4.9	150.1214	0
3SPN.2 (B-DNA)	C	0.239006	6.4	110.0964	0
3SPN.2C	P	0.239006	4.5	94.9696	-0.6
3SPN.2C	S	0.239006	6.2	83.1104	0
3SPN.2C	A	0.239006	5.4	134.122	0
3SPN.2C	T	0.239006	7.1	125.1078	0
3SPN.2C	G	0.239006	4.9	150.1214	0
3SPN.2C	C	0.239006	6.4	110.0964	0

## 2.4 Protein-DNA Excluded Volume term

To prevent protein and DNA overlap each other, we added a Lennard-Jones interaction between protein atoms and protein atoms.

$$V_{LJ}(r) = \begin{cases} 4\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6] - E_{cut} & r < r_c \\ 0 & r \geq r_c \end{cases} \quad (44)$$

with  $\epsilon = 0.03$  kcal/mol,  $\sigma = 5.7\text{\AA}$ ,  $r_c = 2.5\sigma$  and  $E_{cut} = 4\epsilon[(\sigma/r_c)^{12} - (\sigma/r_c)^6]$ . The detail of calibration of parameters for this term and the next term can be found in the SI of ref [4].

## 2.5 Protein-DNA Electrostatics term

The protein and DNA electrostatic interaction is modeled as a Debye-Huckel term.

$$V_{DH} = k_{elec} \sum_{i < j} \frac{q_i q_j}{\epsilon_r r_{ij}} e^{-r_{ij}/l_D} \quad (45)$$

where  $k_{elec} = (4\pi\epsilon_0)^{-1} = 332.24$  kcal  $\text{\AA}$ /mol,  $\epsilon_r = 78$ ,  $l_D = 9.6\text{\AA}$ ,  $q_i$  and  $q_j$  are charges of residue  $i$  and  $j$ . The distance  $r_{ij}$  is the distance between the P atom of DNA residue  $i$  and the CB atom of protein residue  $j$ . Among protein residues,  $q = 1$  for arginine and lysine and  $q = -1$  for aspartate and glutamate. the charge of protein residue is assigned to CB atom. The charge of  $q = -0.6$  is assigned to the P atom of DNA.



### 3 Energy validation of the OpenMM implementation of AWSEM, 3SPN.2, and 3SPN.2C

#### 3.1 Energy evaluation comparison with LAMMPS AWSEM

We ran a short simulation of protein phage 434 repressor (PDBID: 1r69) using LAMMPS AWSEM. The structures are saved every 4000 steps. The energies of each energy terms for the first 6 frames evaluated using both OpenMM and LAMMPS implementation are shown here.

	Scheme	Frame	Chain	Chi	Con	Excluded	Rama	Burial	Water	Frag_Mem
0	OpenAWSEM	0	11.22	24.90	18.88	59.50	-236.05	-53.08	-21.00	-331.41
0	LAMMPS	0	11.13	24.99	224.57	59.53	-232.28	-53.08	-21.00	-331.41
1	OpenAWSEM	1	33.69	9.11	63.93	3.52	-281.10	-54.55	-36.98	-323.85
1	LAMMPS	1	33.56	9.14	77.72	3.52	-277.75	-54.55	-36.98	-323.85
2	OpenAWSEM	2	31.17	6.26	46.97	6.91	-288.55	-55.54	-36.49	-327.89
2	LAMMPS	2	31.10	6.28	48.74	6.89	-284.84	-55.54	-36.49	-327.89
3	OpenAWSEM	3	22.94	7.50	46.28	7.82	-293.50	-57.06	-32.85	-327.66
3	LAMMPS	3	22.87	7.53	46.99	7.83	-290.39	-57.06	-32.85	-327.66
4	OpenAWSEM	4	24.39	6.77	51.75	6.68	-284.85	-56.07	-38.37	-324.24
4	LAMMPS	4	24.29	6.79	54.58	6.68	-282.27	-56.07	-38.38	-324.25
5	OpenAWSEM	5	27.66	8.29	47.92	4.34	-286.99	-57.27	-30.29	-325.43
5	LAMMPS	5	27.56	8.29	48.75	4.34	-283.10	-57.27	-30.30	-325.43
6	OpenAWSEM	6	5.41	1.77	0.90	0.85	-300.72	-57.81	-37.07	-331.64
6	LAMMPS	6	5.35	1.78	0.90	0.85	-297.06	-57.81	-37.07	-331.64

Note, the small difference (less than 1 percent) between these two implementations in the multiple terms like the chain term are due the coordination conversion from LAMMPS output format "lammprj" to OpenAWSEM format "pdb". In "lammprj" format, the positions of atoms are save as the relative position to the simulation box. The difference in the Con term is due a small design change: OpenAWSEM doesn't have the bond between CB and CA for Glycine, but LAMMPS include this bond by using virtual HB as CB.

#### 3.2 Energy evaluation comparison with LAMMPS 3SPN.2 and 3SPN.2C

We ran three short simulations in lammprj using the USER-3SPN2 package of a double stranded DNA with sequence ATACAAAGGTGCGAGGTTTCTATGCTCCCACG. The simulations were run for 50000 steps with a timestep of 0.02 ps using the forcefields 3SPN.2 for A-form DNA, B-form DNA and 3SPN.2C respectively. The simulations were ran at a temperature of 300K and a salt concentration of 100mM. A snapshot was taken every 2000 steps.

To make a fair comparison of the implementations in openMM and LAMMPS we recomputed the energies of the resulting 25 frames. We also implemented this comparisons as unit tests in the open3SPN2 software package. The results for the last 6 frames are shown in the tables below.

Forcefield	Frame	Scheme	Angle	Basepair	Bond	CrossStacking	Dihedral	Electrostatics	Exclusion	Stacking	$E_{total}$ (kcal/mol)
ADNA	20	LAMMPS	59.32	-120.75	32.94	-23.62	-157.27	14.14	0.60	-170.80	-365.44
ADNA	20	OpenMM	59.32	-120.75	32.94	-23.62	-157.27	14.14	0.60	-170.80	-365.44
ADNA	21	LAMMPS	55.73	-123.53	28.93	-25.83	-162.05	15.32	0.31	-177.98	-389.10
ADNA	21	OpenMM	55.73	-123.54	28.93	-25.83	-162.05	15.32	0.31	-177.98	-389.10
ADNA	22	LAMMPS	62.99	-125.08	25.42	-23.48	-155.80	14.37	0.93	-170.82	-371.46
ADNA	22	OpenMM	62.99	-125.08	25.42	-23.48	-155.80	14.37	0.93	-170.82	-371.46
ADNA	23	LAMMPS	53.84	-119.08	25.86	-22.89	-158.21	14.69	0.12	-178.91	-384.58
ADNA	23	OpenMM	53.84	-119.08	25.86	-22.89	-158.21	14.69	0.12	-178.91	-384.58
ADNA	24	LAMMPS	52.63	-128.30	21.00	-25.61	-159.19	13.59	0.20	-177.21	-402.90
ADNA	24	OpenMM	52.63	-128.29	21.00	-25.61	-159.19	13.59	0.20	-177.21	-402.90
ADNA	25	LAMMPS	64.75	-123.48	28.13	-26.96	-160.75	14.24	0.18	-169.62	-373.51
ADNA	25	OpenMM	64.75	-123.48	28.13	-26.96	-160.75	14.24	0.18	-169.62	-373.51
BDNA	20	LAMMPS	65.46	-127.77	27.30	-30.59	-155.67	11.75	0.29	-173.75	-382.97
BDNA	20	OpenMM	65.46	-127.77	27.30	-30.59	-155.67	11.75	0.29	-173.75	-382.97
BDNA	21	LAMMPS	49.08	-125.87	26.10	-30.35	-157.98	11.39	0.55	-178.76	-405.84
BDNA	21	OpenMM	49.08	-125.87	26.10	-30.35	-157.98	11.39	0.55	-178.76	-405.84
BDNA	22	LAMMPS	57.63	-131.75	27.05	-29.72	-159.12	11.16	0.05	-176.13	-400.83
BDNA	22	OpenMM	57.63	-131.75	27.05	-29.72	-159.12	11.16	0.05	-176.13	-400.83
BDNA	23	LAMMPS	48.37	-132.69	24.26	-31.00	-156.43	11.40	0.24	-177.33	-413.19
BDNA	23	OpenMM	48.37	-132.69	24.26	-31.00	-156.43	11.40	0.24	-177.33	-413.19
BDNA	24	LAMMPS	54.43	-135.04	25.73	-28.52	-155.46	11.01	0.62	-169.71	-396.96
BDNA	24	OpenMM	54.43	-135.04	25.73	-28.52	-155.46	11.01	0.62	-169.71	-396.96
BDNA	25	LAMMPS	65.59	-128.54	23.00	-29.56	-155.69	11.37	0.73	-167.10	-380.20
BDNA	25	OpenMM	65.59	-128.54	23.00	-29.56	-155.69	11.37	0.73	-167.10	-380.20
B_curved	20	LAMMPS	50.13	-104.57	19.94	-23.45	-181.98	10.24	0.21	-172.50	-401.98
B_curved	20	OpenMM	50.13	-104.57	19.94	-23.45	-181.98	10.24	0.21	-172.50	-401.98
B_curved	21	LAMMPS	64.00	-92.13	27.78	-22.03	-179.08	10.52	0.32	-173.11	-363.73
B_curved	21	OpenMM	64.00	-92.13	27.78	-22.03	-179.08	10.52	0.32	-173.11	-363.73
B_curved	22	LAMMPS	62.70	-108.68	20.72	-25.38	-184.63	10.37	0.27	-171.34	-395.96
B_curved	22	OpenMM	62.70	-108.68	20.72	-25.38	-184.63	10.37	0.27	-171.34	-395.96
B_curved	23	LAMMPS	52.67	-112.61	17.80	-25.01	-188.50	10.12	0.24	-167.98	-413.28
B_curved	23	OpenMM	52.67	-112.61	17.80	-25.01	-188.50	10.12	0.24	-167.98	-413.28
B_curved	24	LAMMPS	60.59	-102.59	20.37	-19.99	-182.96	10.09	1.00	-159.94	-373.44
B_curved	24	OpenMM	60.59	-102.59	20.37	-19.99	-182.96	10.09	1.00	-159.94	-373.44
B_curved	25	LAMMPS	58.18	-99.51	33.32	-19.96	-181.88	10.06	0.25	-163.45	-362.98
B_curved	25	OpenMM	58.18	-99.51	33.32	-19.96	-181.88	10.06	0.25	-163.45	-362.98

## 4 Tutorial

### 4.1 open3SPN2

#### 4.1.1 Example DNA system

The following code is also available at [https://github.com/cabb99/open3spn2/tree/master/examples/from\\_sequence](https://github.com/cabb99/open3spn2/tree/master/examples/from_sequence)

```

1 # Initialize the DNA from a sequence.
2 # DNA type can be changed to 'A' or 'B'
3
4 seq='ATACAAAGGTGCGAGTTTCTATGCTCCCACG'
5 dna=open3SPN2.DNA.fromSequence(seq,dna_type='B_curved')
6
7 # Compute the topology for the DNA structure.
8 # Since the dna was generated from the sequence using X3DNA,
9 # it is not necessary to recompute the geometry.
10
11 dna.computeTopology(template_from_X3DNA=False)
12
13 # Create the system.
14 # To set periodic boundary conditions (periodicBox=[50,50,50]).
15 # The periodic box size is in nanometers.
16 dna.periodic=False
17 s=open3SPN2.System(dna, periodicBox=None)
18
19 #Add 3SPN2 forces
20 s.add3SPN2forces(verbose=True)
21
22 import simtk.openmm
23 import simtk.openmm.app
24 import simtk.unit
25 import sys
26 import numpy as np
27

```

```

28 #Initialize Molecular Dynamics simulations
29 s.initializeMD(temperature=300 * simtk.unit.kelvin,platform_name='
    OpenCL')
30 simulation=s.simulation
31
32 #Set initial positions
33 simulation.context.setPositions(s.coord.getPositions())
34
35 energy_unit=simtk.openmm.unit.kilojoule_per_mole
36 #Total energy
37 state = simulation.context.getState(getEnergy=True)
38 energy = state.getPotentialEnergy().value_in_unit(energy_unit)
39 print('TotalEnergy',round(energy,6),energy_unit.get_symbol())
40
41 #Detailed energy
42 energies = {}
43 for force_name, force in s.forces.items():
44     group=force.getForceGroup()
45     state = simulation.context.getState(getEnergy=True, groups=2**
        group)
46     energies[force_name] =state.getPotentialEnergy().value_in_unit(
        energy_unit)
47
48 for force_name in s.forces.keys():
49     print(force_name, round(energies[force_name],6),energy_unit.
        get_symbol())
50
51 #Add simulation reporters
52 dcd_reporter=simtk.openmm.app.DCDReporter(f'output.dcd', 1000)
53 energy_reporter=simtk.openmm.app.StateDataReporter(sys.stdout,
    1000, step=True,time=True,
54     potentialEnergy=True, temperature=True)
55 simulation.reporters.append(dcd_reporter)
56 simulation.reporters.append(energy_reporter)
57
58 #Run simulation
59 simulation.step(10000)
60

```

#### 4.1.2 Example Protein-DNA system

The following code is also available at [https://github.com/cabb99/open3spn2/tree/master/examples/Protein\\_DNA](https://github.com/cabb99/open3spn2/tree/master/examples/Protein_DNA)

```

1 # If you want to specify the package address
2 # you can add them to the PYTHONPATH environment variable.
3 # Also you can add them on the run time uncommenting the lines
    below
4 # import sys
5 # open3SPN2_HOME = '/Users/weilu/open3spn2/'
6 # openAWSEM_HOME = '/Users/weilu/openmmawsem/'
7 # sys.path.insert(0,open3SPN2_HOME)
8 # sys.path.insert(0,openAWSEM_HOME)
9
10 #Import openAWSEM, open3SPN2 and other libraries
11 import open3SPN2
12 import ffAWSEM
13 import pandas
14 import numpy as np
15 import simtk.openmm
16 from functools import partial

```

```

17 import sys
18
19 #Fix the system (adds missing atoms)
20 fix=open3SPN2.fixPDB("1lmb.pdb")
21
22 #Create a table containing both the proteins and the DNA
23 complex_table=open3SPN2.pdb2table(fix)
24
25 # Create a single memory file
26 ffAWSEM.create_single_memory(fix)
27
28 #Generate a coarse-grained model of the DNA molecules
29 dna_atoms=open3SPN2.DNA.CoarseGrain(complex_table)
30
31 #Generate a coarse-grained model of the Protein molecules
32 protein_atoms=ffAWSEM.Protein.CoarseGrain(complex_table)
33
34 #Merge the models
35 Coarse=pandas.concat([protein_atoms,dna_atoms],sort=False)
36 Coarse.index=range(len(Coarse))
37 Coarse['serial']=list(Coarse.index)
38
39 #Save the protein_sequence
40 ffAWSEM.save_protein_sequence(Coarse,sequence_file='protein.seq')
41
42 # Create a merged PDB
43 ffAWSEM.writePDB(Coarse,'clean.pdb')
44
45 #Create the merged system
46 pdb=simtk.openmm.app.PDBFile('clean.pdb')
47 top=pdb.topology
48 coord=pdb.positions
49 forcefield=simtk.openmm.app.ForceField(ffAWSEM.xml,open3SPN2.xml)
50 s=forcefield.createSystem(top)
51
52 #Create the DNA and Protein Objects
53 dna=open3SPN2.DNA.fromCoarsePDB('clean.pdb')
54 with open('protein.seq') as ps:
55     protein_seq=ps.readlines()[0]
56 protein=ffAWSEM.Protein.fromCoarsePDB('clean.pdb',
57                                     sequence=protein_seq)
58 dna.periodic=False
59 protein.periodic=False
60
61 #Copy the AWSEM parameter files
62 ffAWSEM.copy_parameter_files()
63
64 #Clear Forces from the system (optional)
65 keepCMMotionRemover=True
66 j=0
67 for i, f in enumerate(s.getForces()):
68     if keepCMMotionRemover and i == 0 and f.__class__ == simtk.
        openmm.CMMotionRemover:
69         # print('Kept ', f.__class__)
70         j += 1
71         continue
72     else:
73         # print('Removed ', f.__class__)
74         s.removeForce(j)
75 if keepCMMotionRemover == False:
76     assert len(s.getForces()) == 0, 'Not all the forces were
        removed'

```

```

77 else:
78     assert len(s.getForces()) <= 1, 'Not all the forces were
       removed'
79
80 #Initialize the force dictionary
81 forces={}
82 for i in range(s.getNumForces()):
83     force = s.getForce(i)
84     force_name="CMMotionRemover"
85
86 #Add 3SPN.2 forces
87 for force_name in open3SPN2.forces:
88     print(force_name)
89     force = open3SPN2.forces[force_name](dna)
90     if force_name in ['BasePair', 'CrossStacking']:
91         force.addForce(s)
92     else:
93         s.addForce(force)
94     forces.update({force_name:force})
95
96 #Add AWSEM forces
97 ft=ffAWSEM.functionTerms
98 openAWSEMforces = dict(Connectivity=ft.basicTerms.con_term,
99                        Chain=ft.basicTerms.chain_term,
100                       Chi=ft.basicTerms.chi_term,
101                       Excl=ft.basicTerms.excl_term,
102                       rama=ft.basicTerms.rama_term,
103                       rama_pro=ft.basicTerms.rama_proline_term,
104                       contact=ft.contactTerms.contact_term,
105                       frag = partial(ft.templateTerms.
106                                     fragment_memory_term,
107                                     frag_file_list_file = "./
       single_frags.mem",
108                                     npy_frag_table = "./
       single_frags.npy",
109                                     UseSavedFragTable = False,
110                                     k_fm = 0.04184/3),
111                       beta1 = ft.hydrogenBondTerms.beta_term_1,
112                       beta2 = ft.hydrogenBondTerms.beta_term_2,
113                       beta3 = ft.hydrogenBondTerms.beta_term_3,
114                       pap1 = ft.hydrogenBondTerms.pap_term_1,
115                       pap2 = ft.hydrogenBondTerms.pap_term_2,
116                       )
117 protein.setup_virtual_sites(s)
118 #Add DNA-protein interaction forces
119 for force_name in open3SPN2.protein_dna_forces:
120     print(force_name)
121     force = open3SPN2.protein_dna_forces[force_name](dna, protein)
122     s.addForce(force)
123     forces.update({force_name: force})
124
125 #Fix excludions
126 for force_name in openAWSEMforces:
127     print(force_name)
128     if force_name in ['contact']:
129         force = openAWSEMforces[force_name](protein,
130                                             withExclusion=False,
131                                             periodic=False)
132         print(force.getNumExclusions())
133         open3SPN2.addNonBondedExclusions(dna, force)
134         print(force.getNumExclusions())

```

```

135     elif force_name in ['Excl']:
136         force = openAWSEMforces[force_name](protein)
137         print(force.getNumExclusions())
138         open3SPN2.addNonBondedExclusions(dna, force)
139         print(force.getNumExclusions())
140     else:
141         force = openAWSEMforces[force_name](protein)
142         s.addForce(force)
143         forces.update({force_name: force})
144
145 #Initialize the simulation
146 temperature=300 * simtk.openmm.unit.kelvin
147 platform_name='OpenCL' #'Reference', 'CPU', 'CUDA', 'OpenCL'
148 integrator = simtk.openmm.LangevinIntegrator(temperature,
149         1 / simtk.openmm.unit.picosecond,
150         2 * simtk.openmm.unit.femtoseconds)
151 platform = simtk.openmm.Platform.getPlatformByName(platform_name)
152 simulation = simtk.openmm.app.Simulation(top,s, integrator,
153         platform)
154 simulation.context.setPositions(coord)
155 energy_unit=simtk.openmm.unit.kilojoule_per_mole
156 state = simulation.context.getState(getEnergy=True)
157 energy = state.getPotentialEnergy().value_in_unit(energy_unit)
158 print(energy)
159
160 #Obtain total energy
161 energy_unit=simtk.openmm.unit.kilojoule_per_mole
162 state = simulation.context.getState(getEnergy=True)
163 energy = state.getPotentialEnergy().value_in_unit(energy_unit)
164 print('TotalEnergy', round(energy,6), energy_unit.get_symbol())
165
166 #Obtain detailed energy
167 energies = {}
168 for force_name, force in forces.items():
169     group=force.getForceGroup()
170     state = simulation.context.getState(getEnergy=True,
171         groups=2**group)
172     energies[force_name] =state.getPotentialEnergy().value_in_unit(
173         energy_unit)
174
175 for force_name in forces.keys():
176     print(force_name, round(energies[force_name],6),
177         energy_unit.get_symbol())
178
179 #Add simulation reporters
180 dcd_reporter=simtk.openmm.app.DCDReporter(f'output.dcd', 10000)
181 energy_reporter=simtk.openmm.app.StateDataReporter(sys.stdout,
182     10000, step=True,time=True, potentialEnergy=True, temperature=
183     True)
184 simulation.reporters.append(dcd_reporter)
185 simulation.reporters.append(energy_reporter)
186
187 #Run simulation
188 simulation.minimizeEnergy()
189 simulation.context.setVelocitiesToTemperature(temperature)
190 simulation.step(100000)

```

## 5 Supplementary figures

### 5.1 Structure prediction results using three contact potential schemes evaluated using the overall Q

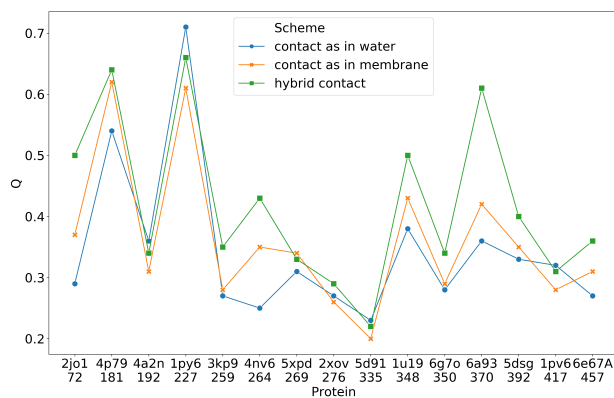


Figure 9: Structure prediction results using three contact potential schemes evaluated using the overall Q.

### 5.2 Example of over saturation of disulfide bonds observed in original AWSEM simulation.

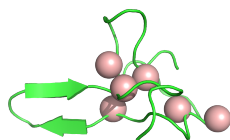


Figure 10: Example of over saturation of disulfide bonds observed in original AWSEM simulation. One cysteine is in contact with three other cystines.

### 5.3 Bets Q for each run.

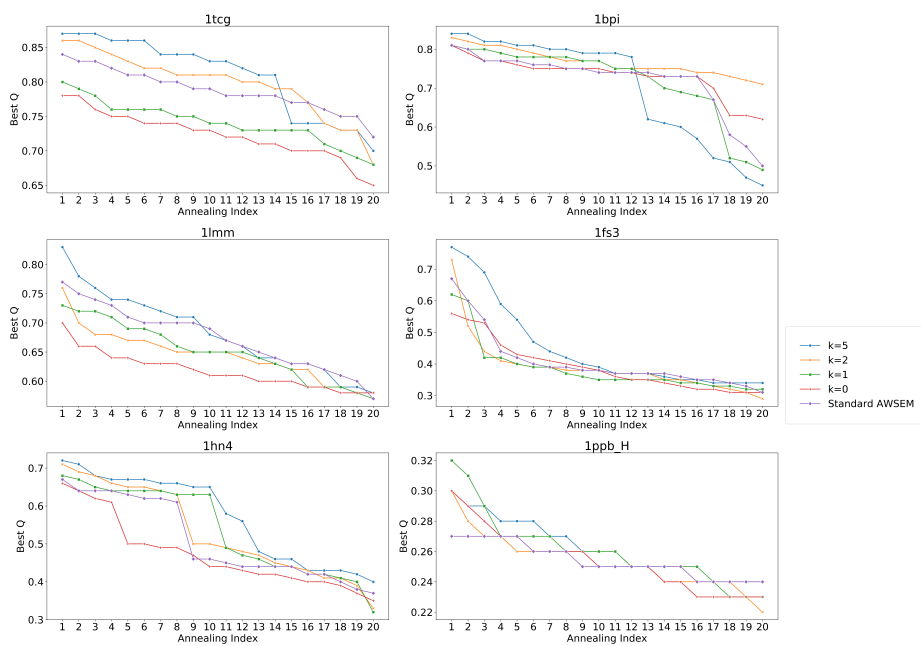


Figure 11: **Best Q value for each run.** a large strength of the disulfide bond potential leads to higher Q value. The annealing indexes are given by sorting their Q value from high to low.



#### 5.4 The predicted structure of alpha-thrombin(PDB: 1ppb) aligned with the crystal structure.

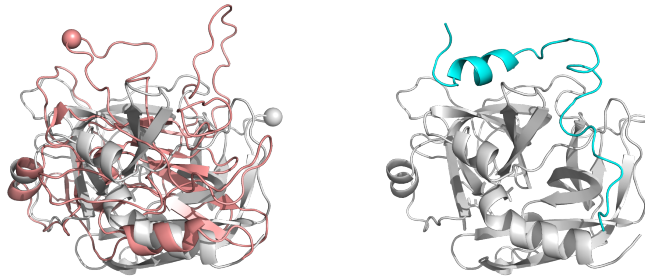


Figure 12: **Left: The structure alignment of predicted structure(red) and crystal structure(white). Right: The complete thrombin crystal structure.** Overall, the lower left region (the C terminal region; residue 168-259) is well aligned. But there is a partial mirror image shown in upper right, residue 150 is shown as sphere as an indication of the mirror image. This partial native folding might be due to that we didn't model the short chain that is experimentally proven to be important for thrombin function [5].

#### 5.5 The predicted structure of ribonuclease A(PDB: 1fs3) aligned with the crystal structure.

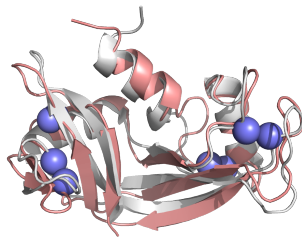


Figure 13: **The structure alignment of predicted structure(red) and crystal structure(white).** The sphere is the CB of Cystines. All Cystine pairs in the predicted structure is matched with the Cystine pairs in crystal structure.

## References

- [1] Hinckley DM, Freeman GS, Whitmer JK, De Pablo JJ. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *Journal of Chemical Physics*. 2013;139(14). doi:10.1063/1.4822042.
- [2] Freeman GS, Hinckley DM, Lequieu JP, Whitmer JK, de Pablo JJ. Coarse-grained modeling of DNA curvature. *The Journal of chemical physics*. 2014;141(16):165103. doi:10.1063/1.4897649.
- [3] Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic acids research*. 2003;31(17):5108–5121.
- [4] Zhang B, Zheng W, Papoian GA, Wolynes PG. Exploring the free energy landscape of nucleosomes. *Journal of the American Chemical Society*. 2016;138(26):8126–8133.
- [5] Papaconstantinou M, Bah A, Di Cera E. Role of the A chain in thrombin function. *Cellular and Molecular Life Sciences*. 2008;65(12):1943–1947.

# Appendix D

## Supporting Information to

Resolving the NF $\kappa$ B hetero-dimer binding paradox: Strain and frustration guide the binding of dimeric transcription factors.

Davit A. Potoyan, Carlos Bueno, Weihua Zheng, Elizabeth A. Komives, Peter G. Wolynes.

### A. Simulation protocols.

All the simulations reported this work have been carried out using a custom built force field which combines the electrostatically enhanced version of AWSEM for proteins (1) with the 3SPN.2C force field for the DNA (2). This force field has been implemented in the LAMMPS suite (version Oct. 9th 2012) for molecular dynamics simulations (3,4). All the simulations reported in this work follow the standard protocols of energy minimization and subsequent equilibration under constant molecular number, volume and temperature conditions (NVT ensemble). The following PDB structures of NF $\kappa$ B were used for setting up NVT simulations: *2RAM* for p65p65, *IILE* for p65p50 and *ISVC* for p50p50. Some of these structures were missing N-terminal DNA binding domains which were reconstructed by running simulated annealing runs with AWSEM starting from modified PDB structures which were supplemented with the missing residues. We used T=300K as the temperature for all simulations including the ones using enhanced sampling via harmonic umbrella potentials. The equations of motion were integrated using the Langevin thermostat with a damping constant of 2000 femtoseconds. The time step for simulations was 5 femtoseconds which ensured proper energy conservation and smooth dynamics. Following the minimization with conjugate gradient and short equilibration steps we ran a set of 50-60 independent simulations initiated from different randomly assigned velocity distributions. The convergence of the simulations was assessed by testing two or more subsets of simulations for consistency in thermodynamic measures such as the domain-domain distances, principal components and potential of mean force profiles.

All free and DNA bound NF $\kappa$ B complexes were subjected to two kinds of simulations: long constant temperature runs used for computing principal components and umbrella sampling runs used for extracting potentials of mean force. To convert simulation time using the coarse grained force field to real time we used FRET estimates of domain motions which take place on the order of 1-10 micro-seconds (5). Close to ~100 domain motions were registered in the simulations which allowed us to estimate the real simulation time to be over 500 microseconds.

Principal component analysis of the free NF $\kappa$ B dimers was performed on structural ensembles generated by constant temperature simulation runs. Routine backbone RMSD alignment of all structures was made prior to all cartesian PCA.

We used the positions of all the backbone C $_{\alpha}$  atoms of protein chains for the conventional Cartesian Principal Component analysis. This was done using the python libraries of MDAnalysis (6). For analyzing the DNA bound NF $\kappa$ B dimers we have used the eigenvectors of the respective free forms. The principal components obtained by projecting Cartesian coordinates of NF $\kappa$ B-DNA structural ensemble on the eigenvectors of NF $\kappa$ B quantify relative changes in the low frequency modes upon DNA binding.

To obtain dissociation free energy profiles we have used umbrella sampling simulations with center of mass distance defined between the DNA and dimerization two domains of NF $\kappa$ B. All of the umbrella sampling simulations were initiated from pre-equilibrated NF $\kappa$ B-DNA constant temperature runs. In order to remove entropic contributions due to overall rotational and translational motions of DNA we subjected all of the umbrellas to the same orientational constraint. This orientational constraint keeps DNA molecules parallel to the orientation in the initial structure of the DNA bound complexes. The orientational constraint is realized by defining a collective variable by using the quaternions of the double helical DNA. The orientation angle has been constrained to remain close to zero indicating a parallel orientation of the DNA to the starting configuration by adding a harmonic bias term with a spring constant of 100 *kcal/mol* Å<sup>2</sup>. This biasing term essentially suppressed any rotational motion of the DNA around the zero angle with respect to the orientation of the DNA in the bound complex while leaving intact the internal motions of the DNA molecule. Free energy profiles were obtained by using the Weighted Histogram Analysis Method (7) on the center of mass distance extracted from simulations with umbrella sampling.

## B. Description of AWSEM+3SPN.2C force fields

AWSEM is a predictive coarse-grained protein force field. Each residue is represented via 3 atoms: C $_{\alpha}$ , C $_{\beta}$  and O. AWSEM is an implicit solvent model which captures the solvent environment through parameters which were trained using energy landscape based algorithms along with a protein structural databases (3,4).

A sequence-independent backbone term,  $V_{\text{backbone}}$ , maintains polypeptide like conformations for the chain. This term enforces planarity of the peptide bond and the chirality of the C $_{\beta}$  atom. A tertiary contact term,  $V_{\text{contact}}$ , models both physical contacts induced by direct interaction between pairs of closely packed residues and water or protein mediated interactions between pairs of residues that are far apart in space. The nature of these interactions which could be either water-mediated or protein-mediated, is determined by the local protein density of the two residues. Therefore this effective potential is non-additive. A burial term,  $V_{\text{burial}}$ , models the burial environment of residues which can be either exposed, partially buried or completely buried depending on hydrophobicity. A hydrogen bonding potential,  $V_{\text{HB}}$ , has two parts. The first part is sequence independent, long range and favors cooperative  $\beta$  sheets formation; the second part is sequence dependent and is more sensitive with respect to the distance and relative orientation of the interacting residues. The relative contribution of each of these energy

functions is determined by scaling factors which ranged from 0 to 1. The scale of each energy term is adjusted so as to better capture the optimal balance of forces that shape secondary and tertiary levels of protein structure in the environment of DNA molecule. The values of scaling factors which we have used in all our simulations are:  $V_{\text{backbone}} = 0.6$ ,  $V_{\text{contact}}=0.75$   $V_{\text{burial}}=1.0$  and  $V_{\text{HB}}=0.5$ .

The associative memory term,  $V_{\text{AM}}$ , models local-in-sequence interactions ( $3 \leq$  sequence separation  $\leq 9$ ). Combining known protein structures along with an algorithm for aligning a target sequence to those structures, AWSEM can be used to limit or guide the local secondary structure conformational search. We used PDB structures 1LE5 for (p65p50), 1SVC (p50p50) and 2RAM (p65p50) as the only memories determining the local in sequence interactions. Employing the single memory setting, the local secondary structure of each protein is weakly biased towards the respective crystal structures. Unlike the local secondary structures, however, all of the intra- and inter-residue tertiary interactions are fully predicted by the physics based transferable interactions in AWSEM potential. The details of AWSEM force-field are described in depth in the publication by Davtyan et al. (3). The Debye Huckel term  $V_{\text{DH}}$ , models the electrostatic interactions between charges in protein-protein, DNA-DNA and protein-DNA atom contacts in a mean field approximation. We chose the Debye screening parameter to be  $\kappa = (10nm)^{-1}$  in order to roughly mimic the ionic screening at  $\sim 100$  mM which corresponds to a physiological concentration of monovalent ions at temperature  $T=300$  K when using a dielectric constant for water  $\epsilon \sim 80$ . The cutoff for this potential is set to be four times the Debye length at 40nm. The role of electrostatic interactions and their modeling in the context of AWSEM is discussed in greater depth in the publication of Tsai et al (1). The AWSEM code is available for download free of charge including numerous examples and tutorials which can all be accessed at <https://github.com/adavtyan/awsemmd/wiki>

The 20bp DNA molecule which is used in NF $\kappa$ B-DNA simulations is modeled by using the 3SPN.2C coarse grained force field (2). Both AWSEM and 3SPN.2C codes have been ported onto LAMMPS package <http://lammps.sandia.gov>. In the 3SPN.2C force field nucleic acids are mapped onto three coarse-grained sites, each corresponding to sugar, phosphate and base groups. The 3SPN.2C is a physics based and empirically parametrized force field consisting of bonded interactions represented by  $V_{\text{bond}}$ ,  $V_{\text{angle}}$  and  $V_{\text{dihed}}$  terms which model harmonic fluctuations in bonds, angles and dihedral angles respectively. The non-bonded interactions are modeled by a stacking term  $V_{\text{stack}}$ , a base-pairing term  $V_{\text{bp}}$ , an excluded volume term  $V_{\text{excl}}$ , a solvent mediated interaction  $V_{\text{solve}}$  and an electrostatic term which is modeled by Debye Huckel potential  $V_{\text{DH}}$  Consistent with interactions in AWSEM. An in depth description of the 3SPN.2C force field can be found in the paper by Freeman et al (2). Besides the electrostatic interactions between protein and DNA charged atoms which are modeled by Debye Huckel terms, we also have a Lennard Jones potential between all protein-DNA pairs which accounts for the excluded volume effects.

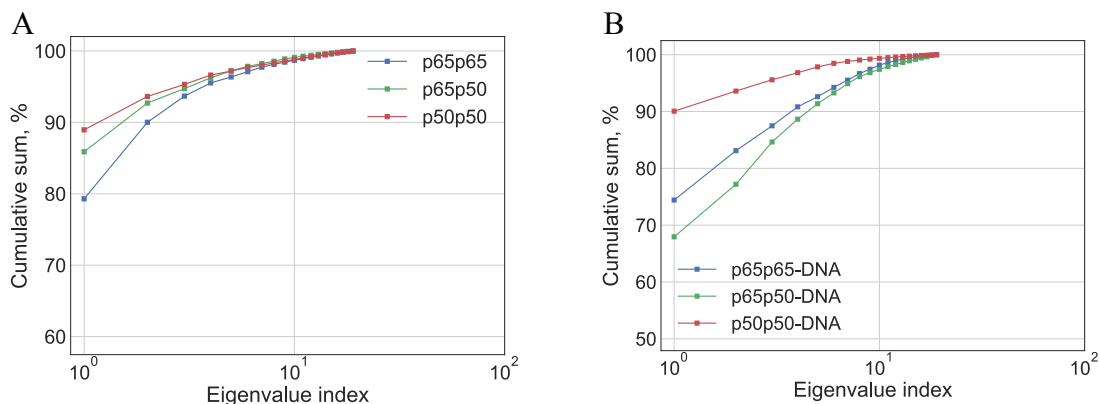


Fig. S1. Shown are the cumulative percentages of eigenvalues for the Cartesian principal components of the (A) free and (B) DNA bound NF $\kappa$ B dimers.

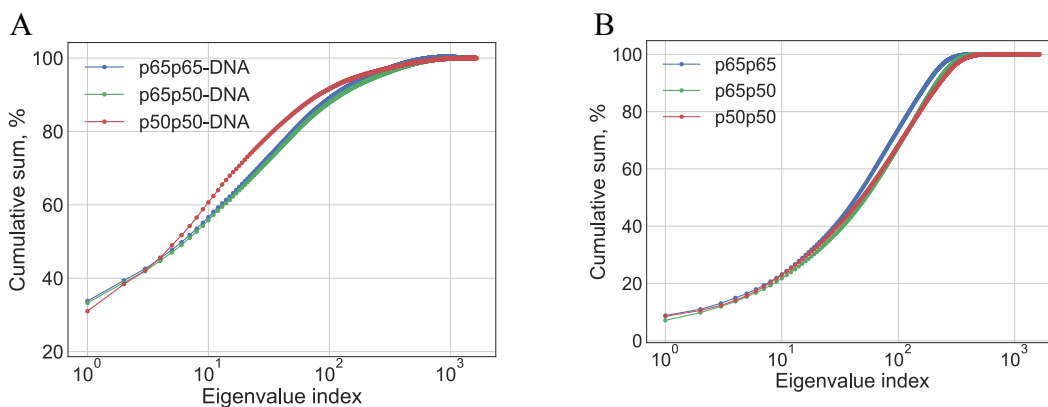


Fig. S2. Shown are the cumulative percentages of eigenvalues for the strain principal components of the (A) free and (B) DNA bound NF $\kappa$ B dimers.

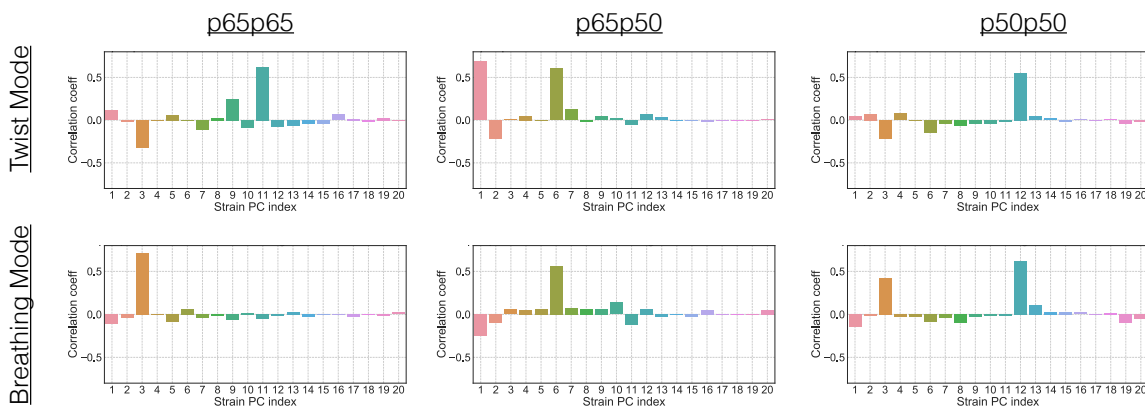


Fig S3. Shown are the correlation coefficients (Y-axis) quantifying correlation between different local strain models (X-axis) and global twist and breathing modes of different free and DNA bound NF $\kappa$ B molecules.

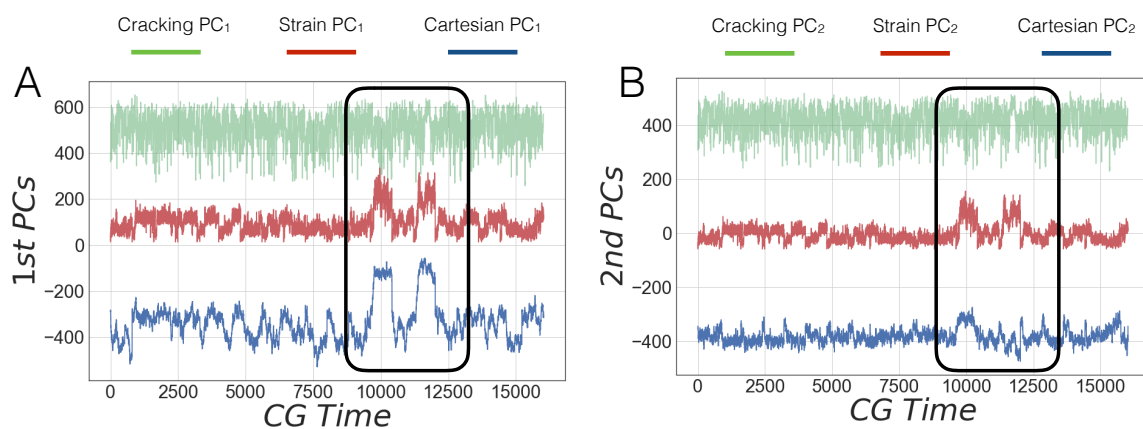


Fig S4. Shown are raw trajectories of principal components corresponding to cracking, straining and global cartesian motion. (A) 1st principal opponents of p65p50 hetero-dimer (B) 2nd principal components of p65p50 hetero-dimer.

### Supplementary References

1. Tsai, M.-Y., Zheng W., Balamurugan D., Schafer NP., Kim BL., Cheung MS., Wolynes PG. *Prot. Sci.* 2015, 1, 1-15
2. Freeman, G. S., Hinckley, D. M., Lequieu, J. P., Whitmer, J. K., de Pablo, J. J. *J. Chem. Phys.* 2014, 141, 165103
3. Davtyan, A., Schafer, N.P., Zheng, W., Clementi, C., Wolynes, P.G., Papoian, G.A. *J. Phys. Chem. B.* 2012, 116, 8494-8503
4. Papoian, G.A., Ulander, J., Eastwood, M.P., Luthey-Schulten, Z., Wolynes, P.G. *Proc. Natl. Acad. Sci.* 2004, 101, 3352-3357
5. Henzler-Wildman K, Kern D. *Nature* 2007, 145, 964-972
6. Michaud-Agrawal, N., Denning, E.J., Woolf, T.B., Beckstein, O. *J. Comp. Chem.* 2011, 32, 2319-2327
7. Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H. and Kollman, P.A. *J. Comp. Chem.* 1992, **13**, 1011-1021