Uncovering Hierarchical Cellular Mechanisms: Linking Molecular Regulation and Biological Topology

Chengxuan Li

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Margaret S Cheung, Chair

Jens Gundlach

Marcel Den Nijs

Program Authorized to Offer Degree:

Physics

University of Washington

Abstract

Uncovering Hierarchical Cellular Mechanisms: Linking Molecular Regulation and Biological Topology

Chengxuan Li

Chair of the Supervisory Committee:
Margaret S Cheung
Department of Physics

This dissertation investigates biological phenotypes across spatial scales, emphasizing hierarchical and topological features through computational physics and machine learning. At the molecular scale, coarse-grained simulations revealed diverse conformations of cofilin oligomers stabilized by disulfide bonds, which regulate actomyosin dynamics via redox-sensitive modifications. At the mesoscale, mechanochemical simulations and network theory uncovered topological transitions, or "avalanches," in branched actomyosin networks controlled by Arp2/3, with machine learning models predicting these events. At the cellular scale, the GRIP-Tomo 2.0 framework integrated synthetic cryo-electron tomography with graph-based learning, enabling robust protein classification through conserved topological fingerprints under limited data. Together, these studies demonstrate how topological insights across scales illuminate the

physical principles underlying biological function, highlighting the power of physics-based computation in complex living systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank my advisor, Professor Margaret S Cheung, for her steady guidance and unwavering support throughout my Ph.D journey. I've learned a great deal under your mentorship—not only in research, but in how to think more clearly, communicate more effectively, and work more independently. I'm grateful for the trust and freedom you gave me to explore my ideas, and for always being available when it counted. To the members of our research group at UW — Jiayi Wang and Jules Nde — thank you for the helpful discussions, technical advice, and for occasionally keeping things light when the deadlines piled up. It's been a pleasure to learn alongside you. To my committee members, Professor Jens Gundlach, Marcel Den Nijs, Armita Nourmohammad, Jiun-Haw Chu and Eli Shlizerman, thank you for your time, thoughtful feedback, and for engaging with my work at each stage. Special thanks to Ms Catherine Provost as well — for being warm and supportive during my Ph.D. study here.

I would like to thank my collaborators, particularly August George, Doo Nam Kim, Trevor Moser and James Evans at Pacific Northwest National Laboratory, for the opportunity to work on joint projects that expanded the scope of my research. I also want to thank my early graduate mentors James Liman, Yossi Eliaz, Carlos Bueno and Pengzhi Zhang as well as faculty supervisors Drs Peter G. Wolynes and Neal Waxham for their advice and encouragement during the formative stages of my Ph.D. Your guidance helped me navigate the steepest parts of the learning curve.

I am grateful to the Department of Physics at the University of Washington, the Department of Physics at the University of Houston, the Environmental Molecular Sciences Laboratory (EMSL) at Pacific Northwest National Lab and the Center for Theoretical Biological Physics (CTBP) at

Rice University for providing an intellectually stimulating environment and the resources necessary to carry out research in this dissertation.

I owe special thanks to my parents, whose constant support has kept me grounded through this long process. I'm especially grateful to your company during the early months of the COVID pandemic, when I was quarantined in the apartment. Even if my research topic is still a mystery to you, your encouragement and belief in me never wavered—and that made all the difference. To my friends and cohorts Ruoyu Zhang, Wenqin Chen, Congcong Xu and Zhongxin Liang—thank you for all the good times we shared along the way.

Finally, I'd like to thank everyone I've had the chance to work with or learn from during my Ph.D. life—you've all contributed in ways both big and small, and I'm grateful for it.

CHAPTER 1: INTRODUCTION

1.1 MOTIVATION AND SCOPE

Living cells are composed of dynamic, spatially organized macromolecular assemblies that interact across multiple scales to carry out essential biological functions. Understanding how molecular structures and interactions translate into larger-scale cellular behaviors remains a central challenge in modern biophysics and cell biology. [1-3] This dissertation seeks to address this challenge by adopting a hierarchical and topological framework, grounded in computational physics and enriched by machine learning, to explore biological phenotypes ranging from molecular to cellular levels.

Phenotypes are often studied at isolated biological scales, such as single-protein structure or global cell morphology. However, many emergent properties of life arise from the coupling across scales: the molecular geometry of a protein can influence mesoscale cytoskeletal architecture, which in turn can affect whole-cell dynamics and morphology. [4-7] To capture these relationships, this thesis integrates multi-scale simulations and data-driven methods that track how structural information propagates through levels of biological organization.

Topology offers a powerful and generalizable language for describing biological form and function. It provides a means of encoding connectivity, symmetry, and robustness in systems ranging from molecular complexes to cytoskeletal networks and spatial proteomes. [8-10] Importantly, topological features often persist under noise or deformation, making them ideal descriptors in both modeling and experimental data.

Recent advances in computational modeling have enabled increasingly detailed representations of biological systems, including coarse-grained molecular dynamics, force-based

mechanochemical models, and agent-based simulations. [11-13] In parallel, machine learning—especially graph-based and interpretable models—has emerged as a complementary tool for identifying relevant patterns and predictive features in high-dimensional, often noisy biological data. [14-16]

This dissertation builds upon these developments to investigate three core problems: the structural diversity of cofilin oligomers (Chapter 2), dynamic reorganization in Arp2/3-mediated actomyosin networks (Chapter 3), and robust protein identification from noisy cryo-electron tomography data using GRIP-Tomo 2.0 (Chapter 4). These projects are unified by a central hypothesis: that topology, when examined across hierarchical scales, can bridge structure and function in living systems.

Ultimately, this work aims to contribute not only specific biological insights but also broadly applicable computational frameworks. By leveraging physical modeling, topological abstraction, and machine learning, it seeks to advance a predictive understanding of cellular organization grounded in the principles of physics.

1.2 BIOLOGICAL PHENOTYPES ACROSS SCALES

Biological phenotypes emerge from the interplay of molecular interactions, supramolecular structures, and dynamic cellular environments. These phenotypes span a broad range of spatial and temporal scales, from the folding and binding specificity of individual proteins to the collective behavior of cytoskeletal networks and the morphological responses of entire cells. [4-6] To understand these multiscale phenomena, it is necessary to analyze how information is organized and transmitted across levels of biological hierarchy.

At the molecular scale, proteins often undergo conformational changes or oligomerize in response to biochemical signals or post-translational modifications. These events can encode functional states and regulatory mechanisms that propagate upward in scale. [7, 8] For example, the redox-sensitive oligomerization of cofilin—an actin-binding protein—influences filament severing and bundling dynamics, ultimately affecting cytoskeletal structure and contractility. [17, 18]

At the mesoscale, cytoskeletal assemblies such as actomyosin networks exhibit emergent behaviors, including force generation, mechanical feedback, and topological reorganization. [19-21]. The Arp2/3 complex mediates the branching of actin filaments, and its concentration or activity modulates the global connectivity and stiffness of the actin network. These networks can undergo sudden, avalanche-like transitions in structure, redistributing stresses and reorganizing intracellular architecture. [22]

At the cellular scale, phenotypes include whole-cell morphology, protein localization patterns, and adaptive responses to environmental stimuli. Cryo-electron tomography (cryo-ET) offers a powerful tool to visualize macromolecular structures in situ, but the noisy and

incomplete nature of cryo-ET data poses significant challenges for identifying protein identities and distributions. [23, 24] Recent developments in synthetic data generation and graph-based machine learning, such as the GRIP-Tomo 2.0 pipeline, enable more reliable phenotype recognition by extracting conserved topological features that remain robust under imaging noise. [25, 26]

By studying phenotypes at these three representative levels—molecular, mesoscale, and cellular—this dissertation aims to reveal organizing principles that bridge physical structure and biological function. Each level not only presents distinct challenges but also offers unique insights into how biological systems maintain robustness, adaptability, and precision despite complexity and stochasticity.

1.3 PHYSICS-BASED MODELING IN CELL BIOLOGY

The complexity of biological systems often obscures the underlying physical principles that govern their structure and dynamics. Physics-based modeling provides a powerful framework to bridge this gap by offering simplified yet mechanistically grounded representations of biological processes. These approaches allow researchers to dissect key parameters, simulate emergent behaviors, and formulate predictive theories that can guide experimental design. [12, 27]

In molecular systems, coarse-grained simulations reduce the degrees of freedom by representing groups of atoms or residues as single units. This abstraction makes it feasible to model large biomolecular complexes or long timescale processes, such as protein folding, oligomerization, and allosteric transitions. [28, 29] In Chapter 2 of this dissertation, coarse-

grained molecular dynamics was employed to investigate cofilin oligomerization and its topological consequences on actin binding.

At larger scales, mechanochemical models and simulations offer frameworks for studying active cellular materials such as cytoskeletal networks. These models incorporate physical laws—force generation, energy dissipation, and elastic deformation—to simulate how cytoskeletal elements self-organize, adapt to stress, and reorganize under perturbations. [19, 30] In Chapter 3, such modeling was used to reveal avalanche-like reorganization events in branched actomyosin networks.

Physics-based modeling also facilitates the integration of experimental data. Parameters can be inferred from imaging, biochemical measurements, or force spectroscopy, allowing for simulations that are both hypothesis-driven and experimentally anchored. Furthermore, these models provide a controlled environment to test biological hypotheses under systematically varied conditions, something often unachievable in vivo. [25, 31]

Importantly, physical models are not merely tools for simulation—they help identify conserved organizational principles, such as symmetry breaking, topological transitions, or feedback regulation, that are relevant across systems and scales. [32, 33] This aligns with the hierarchical approach of this thesis, in which physical constraints at the molecular level propagate through mesoscale networks to influence whole-cell behavior.

By integrating physics-based modeling with computational and machine learning techniques, this dissertation aims to uncover how cellular structures encode and process information. The synergy between modeling and data analysis provides a foundation for uncovering general rules of biological organization.

1.4 MACHINE LEARNING IN BIOLOGICAL PHYSICS

Machine learning (ML) has emerged as a transformative tool in biological physics, enabling the analysis of complex, high-dimensional data that often elude traditional modeling approaches. When applied thoughtfully, ML complements physics-based methods by uncovering patterns, classifying behaviors, and making predictions in systems characterized by stochasticity, noise, and incomplete information. [34-36]

In structural biology, deep learning models have demonstrated remarkable success in protein structure prediction, molecular docking, and cryo-EM map interpretation, most notably with tools like AlphaFold and Cryo-DRGN. [14, 37, 38] These applications exemplify how data-driven models can extract meaningful structure-function relationships from large-scale biological datasets.

Beyond structure, ML has been applied to dynamic biological processes. In cytoskeletal research, supervised and unsupervised learning methods have been used to classify filament morphologies, predict mechanical responses, and track network evolution. [22, 39] In Chapter 3 of this thesis, machine learning was used to identify critical reorganization events—"avalanches"—in simulated actomyosin networks, revealing predictive features of network collapse.

A particularly promising area involves the integration of graph-based ML methods with biophysical modeling. Biological systems are inherently networked, whether in the form of molecular interactions, cytoskeletal connectivity, or spatial protein distributions. Neural networks (NNs) and related algorithms can encode topological information, making them especially suited for datasets with structural complexity. [15, 40, 41]

Chapter 4 of this dissertation presents GRIP-Tomo 2.0, a novel pipeline that integrates synthetic cryo-ET data with graph-based learning. By extracting topological fingerprints robust to imaging noise, the model enables accurate protein classification in data-scarce experimental regimes. Importantly, GRIP-Tomo 2.0 emphasizes model interpretability, allowing biologically meaningful features to be linked back to structural organization—a key requirement in biomedical applications. [42, 43]

While ML methods are often criticized for their "black-box" nature, this work advocates for interpretable, hypothesis-driven applications of machine learning within biological physics. When grounded in physical understanding and combined with synthetic data or simulations, ML becomes a powerful framework for inferring latent structure and bridging biological scales.

1.5 TOPOLOGY AS A UNIFYING LENS

Topology offers a powerful conceptual framework for understanding biological structure and organization across scales. Unlike geometric measures that depend on precise distances or angles, topological features describe properties preserved under continuous deformations, such as connectivity, loops, and symmetries. These properties are particularly valuable in biological systems, where noise, variability, and dynamic rearrangement are inherent. [10, 44, 45]

At the molecular level, topology plays a role in determining protein folding pathways, domain organization, and interaction interfaces. In Chapter 2 of this dissertation, diverse topological conformations of cofilin dimers—such as symmetric versus asymmetric arrangements—were shown to influence actin binding and filament remodeling. These

topological states arise from specific intermolecular interactions, yet they manifest functional consequences that persist across cellular contexts. [46, 47]

In mesoscale cytoskeletal networks, topology governs how forces propagate, how connectivity changes during reorganization, and how structures resist or adapt to stress. The actomyosin networks explored in Chapter 3 exhibit avalanche-like transitions, which are topological in nature: abrupt shifts in global connectivity and stress distribution triggered by localized perturbations. Persistent topological motifs in these networks serve as indicators of stability, resilience, or impending collapse. [47, 48]

At the cellular scale, topological abstractions are crucial for interpreting complex spatial data from cryo-electron tomography. In Chapter 4, GRIP-Tomo 2.0 captures these abstractions through graph-based representations, where each node and edge encode structural and spatial information. Topological graph features—such as clustering coefficients, centralities, or motif counts—enable robust classification of macromolecular components even under imaging noise and data sparsity. [26]

Across all three projects, topology emerges as a consistent and interpretable descriptor that bridges physical modeling and biological insight. Its scale-invariance makes it well-suited to unify molecular, mesoscale, and cellular perspectives, offering a generalizable language for structure-function relationships in living systems.

By foregrounding topological analysis, this dissertation demonstrates how biological function can be traced through persistent structural patterns. This approach offers new tools for quantifying organization, predicting dynamics, and comparing systems across experimental modalities or model conditions.

1.6 OVERVIEW OF DISSERTATION

This dissertation is organized into five chapters. Chapter 1 introduces the central theme of exploring biological phenotypes through a hierarchical and topological lens, integrating physics-based modeling and machine learning. The motivation, conceptual framework, and methodological tools are presented in the context of studying structure-function relationships across molecular, mesoscale, and cellular scales.

Chapter 2 focuses on the molecular-level regulation of cytoskeletal dynamics through cofilin oligomerization. Coarse-grained molecular dynamics simulations were employed to uncover stable topological arrangements of cofilin dimers formed via disulfide bonding. These conformational states influence actin filament remodeling and highlight how redox-sensitive protein structure can regulate cytoskeletal function.

Chapter 3 examines the mesoscale behavior of branched actomyosin networks using mechanochemical modeling and network analysis. The study identifies avalanche-like transitions in network organization, driven by Arp2/3 complex-mediated branching. Machine learning models were developed to predict these events, linking topological network properties with emergent mechanical dynamics.

Chapter 4 introduces GRIP-Tomo 2.0, a graph-based machine learning framework for protein classification in cryo-electron tomography. By combining synthetic training data with interpretable topological fingerprints, the method achieves robust identification of macromolecular components in noisy and incomplete experimental datasets.

Chapter 5 synthesizes the findings of the three projects and discusses their broader implications for biological physics. It also outlines current limitations and future directions,

emphasizing how topological and hierarchical reasoning can be extended to more complex cellular systems and integrated with experimental workflows.

Together, these chapters demonstrate how topological features, when examined across multiple biological scales, can serve as consistent descriptors of structure and function. This work underscores the value of interdisciplinary methods in unraveling the physical principles that govern life at the cellular level.

CHAPTER 2: DECIPHERING THE COFILIN OLIGOMERS VIA INTERMOLECULAR DISULFIDE BOND FORMATION: A COARSE-GRAINED MOLECULAR DYNAMICS APPROACH TO UNDERSTANDING COFILIN'S REGULATION ON ACTIN FILAMENTS

This chapter is based on Chengxuan Li's first author publication: Chengxuan Li, Tingyi Wei, Margaret S. Cheung and Min-Yeh Tsai. "Deciphering the Cofilin Oligomers via Intermolecular Disulfide Bond Formation: A Coarse-grained Molecular Dynamics Approach to Understanding Cofilin's Regulation on Actin Filaments." The Journal of Physical Chemistry B 128 (19) (2024): 4590–4601. https://doi.org/10.1021/acs.jpcb.3c07938

Codes related to this work can be found in the GitHub repository: https://github.com/Cheung-group/Cofilin_Oligomer_AWSEM *or* https://github.com/pnnl/PTMPSI/tree/master/ptmpsi-awsem

Cofilin, a key actin-binding protein, orchestrates the dynamics of the actomyosin network through its actin-severing activity and by promoting the recycling of actin monomers. Recent experiments suggest that cofilin forms functionally distinct oligomers via thiol post-translational modifications (PTMs) that promote actin nucleation and assembly. Despite these advances, the structural conformations of cofilin oligomers that modulate actin activity remain elusive because there are combinatorial ways to oxidize thiols in cysteines to form disulfide bonds rapidly. This study employs molecular dynamics simulations to investigate human cofilin 1 as a case study for exploring cofilin dimers via disulfide bond formation. Utilizing a biasing scheme in simulations, I focus on analyzing dimer conformations conducive to disulfide bond formation. Additionally, I

explore potential PTMs arising from the examined conformational ensemble. Using the free energy profiling, the simulations unveil a range of probable cofilin dimer structures not represented in current Protein Data Bank entries. These candidate dimers are characterized by their distinct population distributions and relative free energies. Of particular note is a dimer featuring an interface between cysteines 139 and 147 residues, which demonstrates stable free energy characteristics and intriguingly symmetrical geometry. In contrast, the experimentally proposed dimer structure exhibits a less stable free energy profile. I also evaluate frustration quantification based on the energy landscape theory in the protein–protein interactions at the dimer interfaces. Notably, the 39–39 dimer configuration emerges as a promising candidate for forming cofilin tetramers, as substantiated by frustration analysis. Additionally, docking simulations with actin filaments further evaluate the stability of these cofilin dimer-actin complexes. My findings thus offer a computational framework for understanding the role of thiol PTM of cofilin proteins in regulating oligomerization, and the subsequent cofilin-mediated actin dynamics in the actomyosin network.

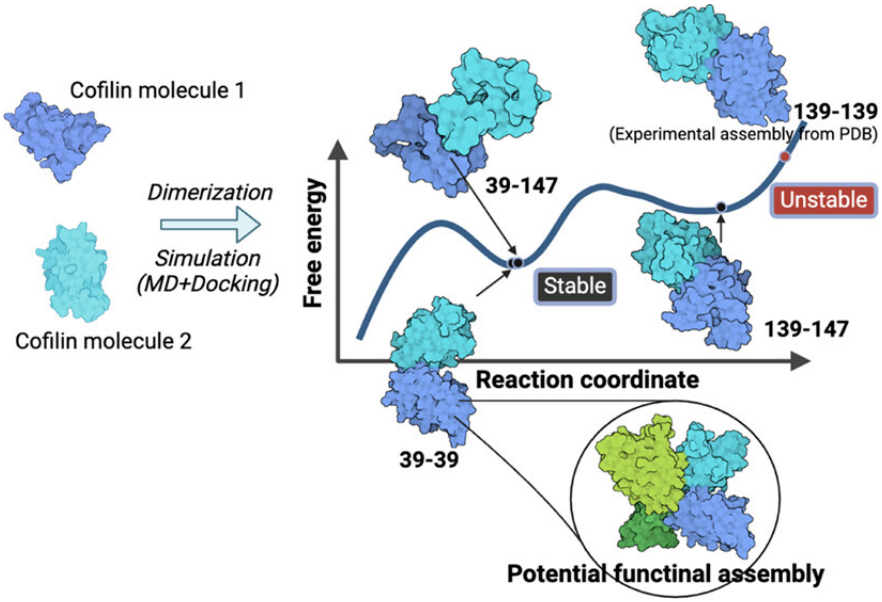


Figure 2.1. Illustration of the cofilin oligomerization study.

2.1 INTRODUCTION

The actomyosin network serves as a critical cellular scaffold that regulates a variety of cell behaviors, including cell division, shape changes, and movement, [49, 50] regulated by actin-binding proteins. Among these numerous actin-binding proteins, cofilin holds particular significance. Cofilin monomers act as molecular scissors, severing actin filaments and facilitating the recycling process through the collaborative efforts of capping proteins, profilins, and other actin-binding proteins. [18, 51, 52] When interacting with actin molecules, cofilin monomers employ distinct binding sites for globular actin monomers (G-actin) compared to actin filaments (F-actin), as highlighted in a previous study. [53] A recent cryo-electron microscopy (cryo-EM) investigation has elucidated the overall structure of F-actin with cofilin molecules decorating the filament's surface, leading to the formation of what is termed cofilactin. [54] These studies collectively suggest that filament surface heterogeneity plays a crucial role in modulating filament severing. The variation in filament severing boundaries arises from different combinations of cofilin-actin interfaces, contributing to the stochastic nature of filament severing.

Interestingly, experimental evidence substantiates that cofilin is not merely functional in its monomeric form; it also exists as oligomers with divergent roles in actin regulation. [17] Cofilin has four cysteine (Cys) residues that can undergo redox-dependent post-translational modifications (PTM) to form specific intramolecular or intermolecular disulfide bonds. [55] Specifically, rather than severing actin filaments, when cofilin forms oligomers through intermolecular disulfide bonds, they are implicated in promoting actin nucleation, assembly, and bundling. [56, 57] The formation and functioning of the cofilin oligomers are regulated by the

local cofilin concentration and the phosphorylation pathway, which also governs the monomeric form. [57, 58] However, there are combinatorial ways to form intermolecular disulfide bonds. Establishing the causal relation between the molecular arrangement of cofilin oligomers and their distinct role in regulating actomyosin networks in response to redox is unclear.

Various experimental works have posited that these cofilin dimers likely emerge through intermolecular disulfide bonds between specific cysteine residues ─ namely, cysteine 39 and cysteine 147─on adjacent cofilin monomers, [17] instead of a cofilin symmetric dimer through the disulfide bonds between their cysteine 139 residues. [59] Moreover, prevailing theories suggest that the biologically active form of cofilin is not a dimer but rather a tetramer, leaving the dimeric state as a potentially transient state in oligomer formation. [60] However, the structural and functional intricacies of these cofilin oligomers in response to thiol PTMs remain unknown, thereby offering ample opportunities for the application of coarse-grained simulation techniques.

While the atomistic simulations have shed light on crucial aspects of the mechanical stress from the composition of filament structures, they are often constrained by limited time scales and length scales, which pose challenges for achieving robust statistical analyses. Coarse-grained protein models provide an avenue to surmount these limitations. The application of coarse-grained protein modeling to explore the network dynamics of CaMKII/F-actin bundles, in conjunction with protein array experiments and electron microscopy imaging, has effectively elucidated the multivalent binding interactions between CaMKII and F-actin, aligning with experimental findings. [42] Collectively, these studies underscore the capacity of coarse-grained protein models to capture sufficient molecular details, enabling them to replicate numerous features of filamentous behavior and, in turn, to address the regulatory roles of actin-binding proteins within the actin network.

The advantages of coarse-grained simulations in addressing complex protein behaviors over long time scales make them particularly well-suited for investigating the formation, stability, and function of cofilin oligomers and their interaction with actin filaments. In this study, I focus on human cofilin 1, with specific attention to the structural representation denoted by the PDB ID 4BEX. [59] While it provides a symmetric cofilin dimeric structure comprised of two monomers, it is noteworthy that the authors themselves suggest that this assembly, formed during crystallization, may not faithfully represent a biologically relevant dimeric state. To fill the existing knowledge gaps, I provide a computational framework to evaluate the cofilin oligomers formed with distinct intermolecular disulfide bonds and their interaction with a short fragment of the actin filament.

2.2 SIMULATION AND ANALYSIS

Associative-Memory, Water-Mediated, Structure, and Energy Model (AWSEM)

The present study leveraged the AWSEM (Associative-Memory, Water-Mediated, Structure, and Energy Model) coarse-grained protein force field for an in-depth analysis of cofilin protein dimerization. [61] This force field, implemented on the LAMMPS simulation platform, employs a three-bead model for each amino acid residue, representing C_α, C_β, and O atoms. This coarse graining preserves ideal peptide bond geometry, thus enabling efficient simulations of protein folding dynamics, [28] structure prediction, [62] and protein aggregation. [63-65]

The energy function of AWSEM comprises transferable and physically motivated potentials, encapsulating complex residue-specific physicochemical properties such as hydrophobicity and electrostatic interactions within the framework of protein secondary and

tertiary structures. Implicitly, the enthalpic contributions of protein–protein contacts are included in the AWSEM contact energies.

In AWSEM's coarse-grained scheme, the energy function is sectioned into three main components: $V_{backbone}$, $V_{nonbackbone}$, and V_{FM} (Fragment Memory), responsible for the backbone geometries, protein's physicochemical attributes, and local structural tendencies, respectively. $V_{backbone}$ encompasses five terms that regulate chain connectivity, bond angles around the C_α atom, orientations of the C_β atoms, backbone dihedral angles, and excluded volume interactions. Harmonic potentials control both the chain connectivity and the bond angles around the C_α atom. $V_{nonbackbone}$ is further split into three terms: $V_{contact}$, V_{burial}, and $V_{helical}$. These terms individually consider aspects like tertiary fold contact interactions, residue exposure/burial preferences, and helical structure propensity. V_{FM} is tailored to bias the local structure toward those found in a "fragment memory" library of protein fragments with similar local sequences. This term also accounts for the local steric effects influenced by the protein's local sequence. For this study, a single fragment memory scheme within AWSEM was employed to accentuate the effects of physical forces on folding and binding landscapes.

For further details on the force field, readers are referred to the work by Davtyan et al. [12] The AWSEM code is publicly accessible via Github: AWSEM Repository (https://github.com/adavtyan/awsemmd).

Structural Preparation and Simulation Protocol for Cofilin Dimer Formation through Intermolecular Disulfide Bond

The wild-type sequence of human cofilin-1 (hCof1) is depicted in Figure 2.2A. Due to the limited availability of monomeric cofilin structures, a crystal structure with high resolution was selected for the monomeric form (PDB ID: 4BEX). [59] This structure was found to possess a mutation at residue 147, where cystine was replaced by alanine. To revert the structure to its wild-type form, I mutated residue 147 back to cystine using the "Mutate Residue" module available in VMD software, as shown in Figure 2.2B. [66]

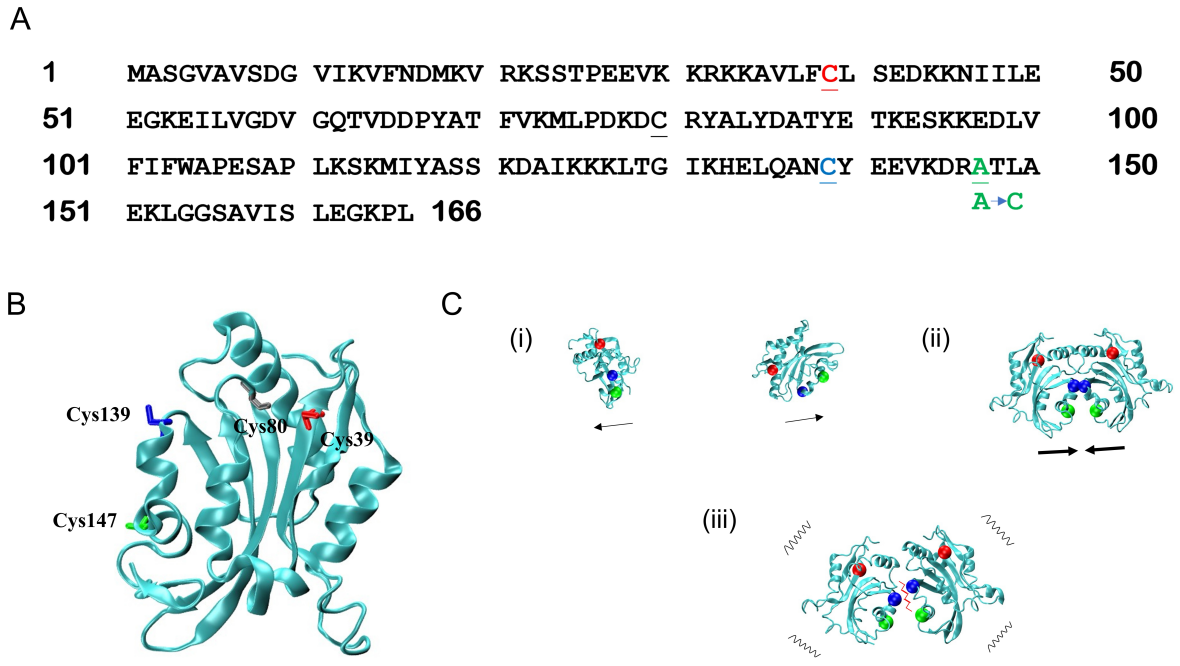


Figure 2.2. Mutated cofilin sequences, their monomeric structure, and simulated dimer formation workflow are described. (A) Sequence representation of human cofilin 1 (PDB ID: 4BEX). Cysteine residues 39, 139, and 147 are highlighted in red, blue, and green, respectively. The latter was mutated to cysteine in my simulations to mirror in vivo conditions. While cysteine residue 80 is also underlined, it is not color-coded. The color patterns introduced here are consistently used in subsequent figures. (B) Monomeric cofilin structure after mutation at residue 147. The protein chain is rendered using the 'NewCartoon' method in cyan, while cysteine residues employ the 'Licorice' visualization with colors as designated in (A). Due to spatial proximity in relation to residue 39, cysteine residue 80 is labeled. (C) The stages of simulated dimer formation. (i) Monomers are drawn apart using a gentle spring force centered on the total

protein structures. (ii) The force from stage (i) is discontinued and replaced by a potent spring force ─ set at an equilibrium distance of 2 Å, the approximate length of a disulfide bond ─ encouraging the monomers to assemble (as demonstrated using residues cys139 and cys139). (iii) Post removal of the bias from (ii), the monomers are at liberty to oscillate or dissociate. All molecular visualizations were made using VMD.

The simulation process for forming cofilin dimers was organized into three distinct steps, illustrated in Figure 2.2C: (1) Initiation (using pushing forces), (2) Association (using pulling forces), and (3) Relaxation (absence of biasing forces). In the Initiation step, a spring force characterized by an equilibrium distance greater than the intermolecular disulfide bond was employed to two cofilin monomers. During the Association phase, a spring force equivalent to the equilibrium distance of a disulfide bond was applied between the Cys39-Cys147 and Cys139-Cys139 residues of adjacent cofilin monomers, to simulate oxidized thiol cysteine residues from post-translational modifications. For simplicity, Figure 2.2C shows only the Cys139-Cys139 bond as an example. The Relaxation step was designed to allow the mutual configurations of cofilin monomers in the dimer assembly to stabilize naturally, simulating the reduced state of cysteine. Accordingly, the biasing spring force was removed during this phase.

The simulated system was held at a constant temperature of 300 K throughout the study. Each simulation trajectory consisted of three phases: (1) Initiation, (2) Association, and (3) Relaxation, as previously outlined. These phases encompassed 0.5, 4, and 4 million simulation time steps, respectively. To ensure the robustness of the statistical analysis, I conducted a total of 30 individual trajectories.

Importance Sampling and Free Energy Analysis

For the purpose of free energy analysis, (21) I employed two distinct configurations of cofilin dimers as reference structures for umbrella sampling. The first configuration featured

Cys139-Cys147 at the interface (139–147 cofilin dimer), which was identified during the previous unbiased relaxation simulation within the context of the 3-stage dimer formation simulation. The presence of this configuration remained consistent throughout the dimer formation simulation, regardless of the residue pairs subjected to spring force application (i.e., Cys39-Cys147 and Cys139-Cys139 residues of adjacent cofilin monomers). The second configuration corresponded to the crystallized 4BEX assembly (139–139 cofilin dimer) obtained through experimental means. [59] Notably, the reference structure was derived by averaging over the last 50 simulation frames of a single trajectory, specifically chosen from among the 30 trajectories to ensure the stable dimer formation of interest.

In the umbrella sampling, a harmonic biasing force was applied to the α-carbon atoms of the target molecule, utilizing the global Q value of the entire dimer assembly as the sampling coordinate.

The global Q value was calculated based on the following equation [12]

$$Q = \frac{2}{(N-2)(N-3)} \sum_{i<j-2} \exp\left[-\frac{\left(r_{ij}-r_{ij}^{N}\right)^{2}}{2\sigma_{ij}^{2}}\right]$$

where N is the total number of residues, r_{ij} is the instantaneous distance between C_α atoms of residues i and j, r_{ij}^{N} is the same distance in the reference structure for calculating the formation of a native contact, and σ_{ij} is given as $\sigma_{ij} = (1 + |i - j|)^{0.15}$. Q quantifies the similarity between two protein structures by assessing the fraction of all pairwise residues' native contacts. "Q global" specifically denotes the computation of native contacts for all residues within a protein structure, as applied in the context of a dimeric structure in this study.

I performed umbrella sampling analyses on five replicate sets, specifically selecting binding simulations where the two cofilin monomers consistently remained associated with each other (i.e., $Q \geq 0.60$). The global Q values obtained from these binding simulations were then

concatenated into an array and visually inspected for consistent overlapping. Subsequently, I categorized the data into 300 discretized centers, which are generated using the *numpy.linspace* function. [67]

I then generated a list of biased centers for the sampled Q values, ranging from 0.60 to 0.975 with an interval of 0.025. Other predefined parameters, such as force constants (1000 in the unit of kcal/mol $Å^2$), maximum iterations (100 K), and energy unit kT (0.593 kcal/mol), were set accordingly. All other parameters were kept consistent with default values utilized in previous work. [62] Next, I utilized the Pyemma thermo module function '*pyemma.thermo.estimate_umbrella_sampling*' with the estimator 'WHAM' to calculate the free energy profile. Finally, I visualized the calculated free energy profile alongside the discretized centers. [68] The jupyter notebooks for umbrella sampling analysis can be found on the Github (https://github.com/pnnl/PTMPSI/tree/master/ptmpsi-awsem).

Exclusion of Long-Range Electrostatics in Binding Simulations

Recognizing the significance of electrostatic interactions in governing protein structure and dynamics is fundamental to the understanding of molecular behavior. Notably, while the standard AWSEM code incorporates local electrostatic interactions with the solvent, it does not consider long-range electrostatic interactions. Previous research has emphasized the pivotal role played by long-range electrostatic interactions (namely, Debye–Hückel potentials) in predicting the structural characteristics of specific proteins and shaping the binding energy landscapes of various protein binders. [63] However, when considering the stability of cofilin, these interactions appear to have a limited impact. A comparison of the simulated annealing results for

cofilin, with and without the inclusion of long-range electrostatics (e.g., screened electrostatics at high salt concentration), clearly demonstrates their negligible effect on the native structure's stability, as visualized in Figure 2.3 in the Supporting Information.

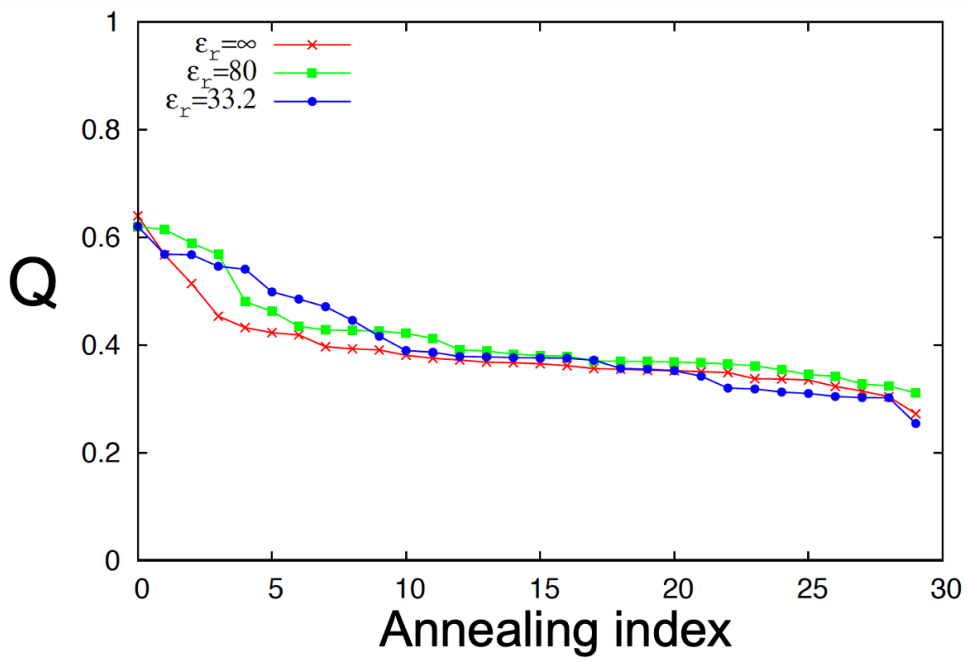


Figure 2.3. Effects of long-range electrostatic interactions on the structural stability of human cofilin 1 is evaluated using AWSEM. (a) Structural stability test. The Q value of cofilin is calculated as a function of simulation time. The Q value remains 0.8~0.9 (high Q value), independent of the strength of long-range electrostatics. (b) Simulated annealing profiles. The annealing result shows similar profile, irrespective of the electrostatic strengths. The Debye-Hückel potentials were used to mimic the long-range electrostatic effects. ε represents the dielectric constant with ε=∞, 80 (water), 33.2, and 16.6 meaning "no", "mild", "strong", and "extreme strong" electrostatic strength, respectively.

Furthermore, even in scenarios where multiple cofilin molecules engage in dimerization, the frustration analysis (refer to Figure 2.4 in the Supporting Information) confirms that these long-range electrostatic interactions do not exert a significant influence. Figure 2.4 vividly illustrates the minimal role of electrostatic-induced frustration interactions at the cofilin binding interfaces. Consequently, despite the capability of AWSEM to incorporate Debye–Hückel potentials for simulating long-range electrostatics in simulations, I have chosen to exclude these

interactions from the binding simulations due to their limited impact on the proteins under investigation, essentially resembling the assumption of a strong screening effect due to the high concentration of salt.

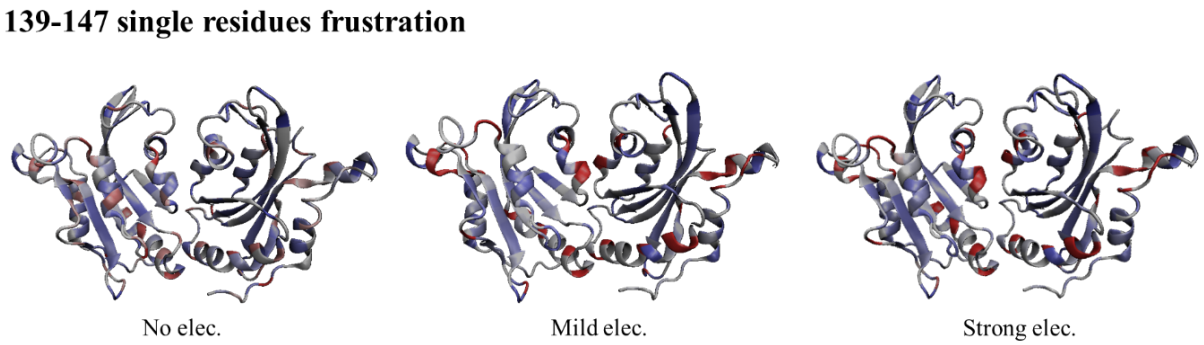


Figure 2.4 - Single residues frustration analysis. Electrostatically induced frustration is evaluated using single residue frustration mode for four cofilin dimer configurations (From top to bottom: 4BEX experimental assembly, 39-39 dimer, 39-147 dimer, 139-147 dimer). The frustration is calculated (from left to right) in the context of "No electrostatics (No elec.)", "Mild electrostatics (Mild elec.)", and "strong electrostatics (Strong elec.)", which represents dielectric constant in the Debye Hückel potentials $\varepsilon = \infty$, 80, 33, respectively. The cofilin structure is shown using cartoon diagram with highly frustrated residues colored in red, neural residues colored in gray, and minimally frustrated residues colored in purple.

Visualization, Clustering, and Frustration Analysis

The analysis of AWSEM simulations was conducted using various tools and techniques. To visualize the simulations, I employed VMD. [66] The total protein structure was visualized using the 'New Cartoon' drawing method, while the specific cysteine residues (39, 139, and 147) were highlighted using the 'Beads' drawing method, each assigned a distinct color (Cys39 in red, Cys139 in blue, and Cys147 in green) for easy identification within the dimer.

For further trajectory analysis, I initially processed the PDB files containing the cofilin dimer configurations using the Python package MDTraj. [69] Subsequently, I generated contact

maps illustrating the residue positions within the cofilin dimers using the Contact Map Explorer (https://github.com/dwhswenson/contact_map), which is an open-source tool built on MDTraj. The scripts for generating these contact maps are available on Github repository (https://github.com/pnnl/PTMPSI/tree/master/ptmpsi-awsem).

To perform structural clustering analysis, I curated a total of 1360 structures obtained from biased simulations conducted for umbrella sampling. In pursuit of a comprehensive exploration of the cofilin dimer ensemble, I systematically selected these structures across five separate replicate simulations, each representing 16 distinct trajectories (resulting in a total of 5 × 16 = 80 trajectories). Within each trajectory, I extracted a structure every 50 simulation frames, yielding a grand total of 5 × 16 × 17 = 1360 representative structures. The clustering analysis began by transforming the Q-global data, into a square form. The Q value serves as a structural similarity measure, ranging from 0 to 1, where ″1″ signifies an identical reference structure and ″0″ denotes a completely dissimilar counterpart. Subsequently, I employed hierarchical clustering analysis, utilizing the Python package Seaborn and the function “seaborn.clustermap.″ In this analysis, I specified critical parameters, including metric = “correlation” for the pairwise distance metric and method = “average” for the linkage method. These choices facilitated the generation of informative visualizations. Please consult the GitHub repository (https://github.com/pnnl/PTMPSI/tree/master/ptmpsi-awsem/clustering_analysis) for further reference and access to detailed scripts related to the clustering analysis of the dimer configurations.

To investigate frustration within the cofilin structures, I utilized the Frustratometer Server, accessible at http://frustratometer.qb.fcen.uba.ar. [70] Based on energy landscape theory, the concept of frustration in biomolecules and its contemporary perspective on folding, function, and assembly have been discussed in detail in various sources. [71, 72] This analysis was

performed by uploading the PDB structures of cofilin oligomer candidates to the Frustratometer Server with a sequence separation parameter of 3 and without considering long-range electrostatics (see Exclusion of electrostatics in binding simulations section above).

Protein frustration analysis employs computational techniques to identify energetically frustrated residues or regions within a protein structure. These methods evaluate factors like conflicting interactions, steric clashes, and geometric constraints to pinpoint areas where local interactions favor different conformations, leading to overall structural instability. By quantifying and visualizing frustration, researchers gain insights into protein folding, stability, and function, which can inform studies on protein engineering and drug design.

In this study, I explored mutational frustration in cofilin oligomers to assess the favorability of current residues compared to hypothetical ones at the oligomer interface. [73] When the interface within a cofilin oligomer experiences minimal frustration (mostly green lines), it is energetically stabilized. Conversely, a highly frustrated interface (with many red lines) suggests instability and may serve as potential sites for allosteric regulation or protein–protein interactions.

While the current frustration analysis algorithm does not directly incorporate the formation of disulfide bonds between residues, I acknowledge the importance of considering how such chemical modifications can influence molecular interactions and configurations. Indeed, the formation of disulfide bonds could potentially mitigate the destabilizing effects of frustrated interactions by stabilizing the interface, as suggested by experts in this field.

2.3 AWSEM SIMULATIONS REVEAL MULTIPLE PROBABLE BINDING INTERFACES IN A COFILIN DIMER

In the AWSEM simulations, I identified multiple cofilin dimer candidates established by intermolecular disulfide bonds with distinct binding interfaces, of which six were particularly noteworthy. Cofilin's mutual configurations in the dimer are represented using three key cysteine residues (39, 139, 147) for visual inspection. Previous experimental findings had proposed a dimer interface involving cysteine 39 and 147 residues, referred to as the 39–147 cofilin dimer. [74] Additionally, the PDB structure of human cofilin 1 (PDB ID: 4BEX) reported a disulfide bond formed between adjacent, symmetric monomers at cysteine 139 residues, known as the 139–139 cofilin dimer. [59] While the simulations confirmed the presence of the 39–147 cofilin dimer, the experimentally reported 4BEX assembly involving disulfide bonds between cysteine 139 residues was notably absent. The simulations also unveiled several previously unreported binding interfaces, with the 39–39 and 139–147 interfaces being the most prevalent. In Figure 2A, the population distribution of these cofilin dimer candidates is illustrated using structural clustering analysis, while Figure 2.5B visualizes the centroid structures based on their constituent cysteine residues. It is important to note that the population size shown in Figure 2.5A serves as a qualitative indicator for structural categorization purposes and does not represent the realistic population of representative clusters.

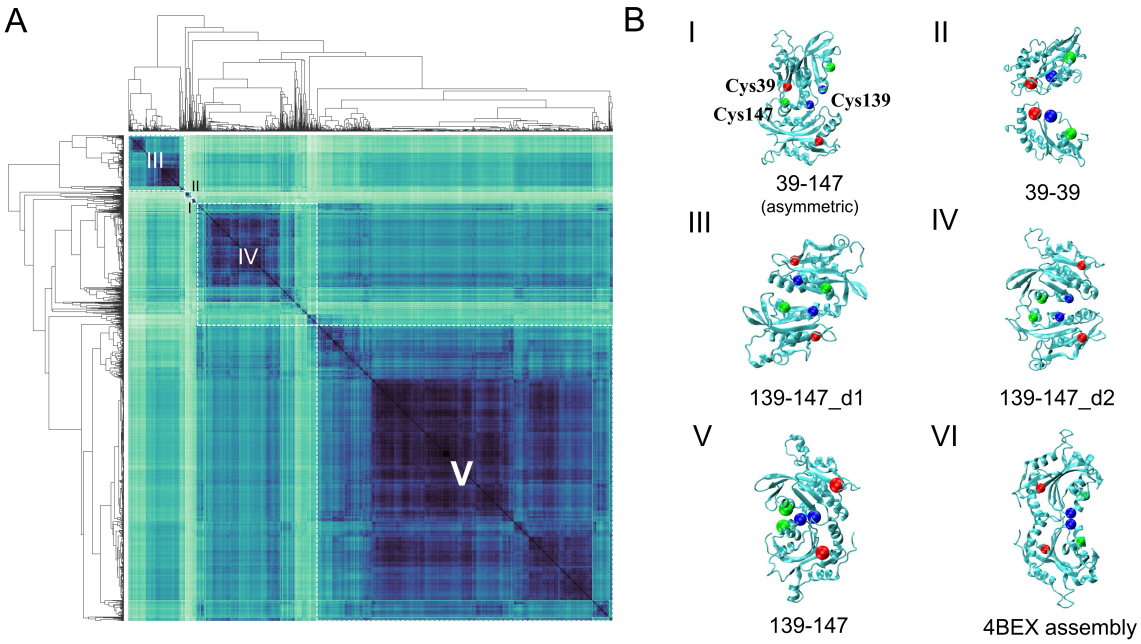


Figure 2.5. Analysis of structural clustering for various cofilin dimer configurations. (A) Clustered heatmap with dendrograms of cofilin dimeric configurations derived from AWSEM MD simulation trajectories using the Q value for structure. Dark clusters correspond to the cofilin dimer candidates depicted in (B). (B) Cofilin dimer candidates indexed I–V, correlating with the cluster map in (A). Candidate VI (139–139) represents the experimental assembly from 4BEX, which was not identified in the simulations. Candidates I–V are denoted by the cysteine residues detected at the interface. The cysteine residues 39, 139, and 147 are represented as red, blue, and green spheres, respectively. Candidates III and IV (both belonging to 139–147) are further distinguished as d1 and d2, representing degenerative states 1 and 2, respectively. Notably, the candidate I (39–147) dimer exhibits asymmetry, while the other candidates are symmetric.

The simulation results predict that the 39–147 cofilin dimer is the only asymmetric assembly among the other six configurations. The asymmetry of this assembly can be visualized in three-dimensional space by representing the three key cysteine residues (39, 139, and 147) with red, blue, and green balls, respectively (refer to Figure 2.5B–I). One can conceptualize a virtual triangle formed by connecting these three cysteine residues (colored balls) within the cofilin molecule. The asymmetric binding configuration of a cofilin dimer is akin to envisioning two virtual triangles stacked together with their vertices misaligned. This asymmetric assembly

suggests a generic intermolecular interaction between cofilin molecules in vivo and is known to exhibit bundling activity under oxidative stress. [17, 18, 74, 75] The interface residue pairs in this asymmetric assembly consist of a combination of mild hydrophobic contacts (involving Ser-Ile, Ser-Cys, Tyr-Cys, Cys-Gly, Leu-Gly) and scattered charge pairs (including Glu-Lys, Glu-Glu).

The experimentally determined cofilin assembly, the 139–139 cofilin dimer, exhibits C2 symmetry, with two virtual triangles mirroring each other in the crossing plane. In this symmetric assembly, the interface residue pairs engage in numerous strong hydrophobic contacts (e.g., Phe-Leu, Ala-Leu, Met-Met, Ala-Met, Ala-Ala, Cys-Cys, Cys-Tyr, Asn-Pro) and a significant number of repulsive charged pairs (e.g., Glu-Glu, Lys-Lys). Despite its high degree of symmetry, the distribution of the resulting interaction pairs does not optimize at the interface, thus remaining unobserved in the simulations.

On the other hand, the predicted 139–147 cofilin dimer displays considerable conformational dynamics among its subspecies. These subspecies share a significant number of strong hydrophobic contacts (e.g., Met-Leu, Ala-Leu, Leu-Ile, Met-Met, Met-Ala, Cys-Cys) along with mild stabilizing charge pairs (e.g., Asp-Glu). This combination of chemical interactions underscores the conformational flexibility and dynamic exchange of contacts at the interface.

Interestingly, the model of 39–39 cofilin dimer exhibits a highly symmetric molecular structure (C2 symmetry) with very limited hydrophobic contacts, nearly none at all. The majority of the binding interface is composed of charged residues such as Lys(+), Glu(−), and Asp(−), resulting in charge-complementary clusters, including major attractive charge pairs, Lys(+)-Glu(−) and Lys(+)-Asp(−), and several repulsive charge pairs like Glu(−)-Glu(−) and Glu(−)-

Asp(−). This phenomenon highlights a unique charge-pair distribution that compensates for repulsive charge forces through effective distribution. Note that these charge pairs are formed due to short-range electrostatic contributions from the default AWSEM potential. This version does not take into account the long-range electrostatic effects (See Methods for the details). One thing worth mentioning is that Arg(+) is not observed as part of the electrostatics-driven binding interface, which is notable considering its common role in electrostatic interactions. One possible explanation for the exclusion of Arg(+) at the binding interface is the interplay between steric considerations, the structural specificity required for protein–protein interactions, and the relatively large and complex side chain of Arginine (Arg) compared to some other charged residues like Lysine (Lys). Protein–protein interactions necessitate a precise fit between interacting surfaces, and the choice of residues in the binding interface often aims to achieve optimal interactions, prevent steric clashes, and maintain a specific structural arrangement.

To assess the stability of the predicted cofilin dimer complexes with greater precision, I built all-atom structures for these dimers and performed individual 40 ns MD simulations using the NAMD simulation package with the CHARMM27 force field. [76] The findings from these simulations consistently affirm the stability of the predicted dimer complexes (Refer to Figure 2.6).

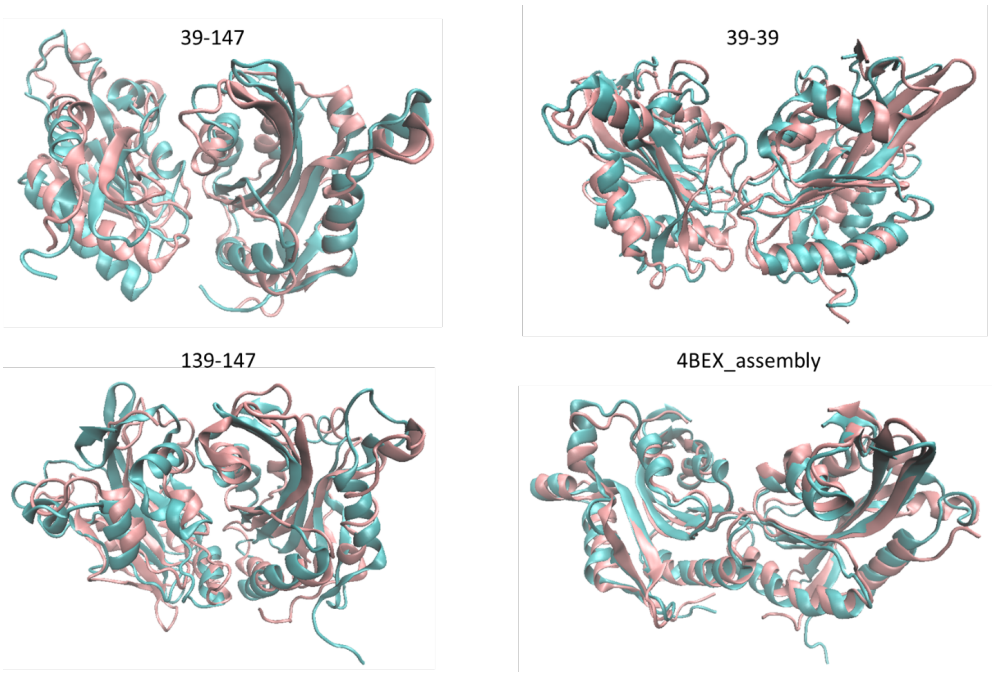


Figure 2.6. Structural stability of predicted cofilin dimer complexes is analyzed via RMSD analysis. The RMSD values were calculated for each predicted cofilin dimer complex: (a) 39-147 cofilin dimer, (b) 39-39 cofilin dimer, (c) 139-147 cofilin dimer, and (d) 139-139 cofilin dimer (4BEX assembly). The corresponding cofilin dimer complexes are depicted below, with the pink cartoon diagram representing the initial structure (reference structure) and the cyan cartoon diagram representing the final frame (at 40ns) from the atomistic simulation. The atomistic simulations were performed using NAMD with the CHARMM27 force field, including explicit water molecules within a simulation box, counter ions for charge neutralization, and periodic boundary conditions.

2.4 FREE ENERGY LANDSCAPES OF MAJOR COFILIN DIMER CANDIDATES: INSIGHTS FROM IMPORTANCE SAMPLING

Despite the plethora of cofilin dimer configurations obtained through clustering analysis, their thermodynamic stability remains unknown. To assess their relative stability among the six

cofilin dimers, I employed the umbrella sampling technique to explore the binding free energy landscape of cofilin dimers. Figure 2.3 presents the free energy profile of cofilin dimer candidates using the 139–147 dimer as the reference structure. Notably, several basins are evident along the free energy profile, indicating stable populations of the dimer. These free energy basins are labeled A&B, C, and D, with their respective representative structures (dimer candidates I to VI) displayed accordingly.

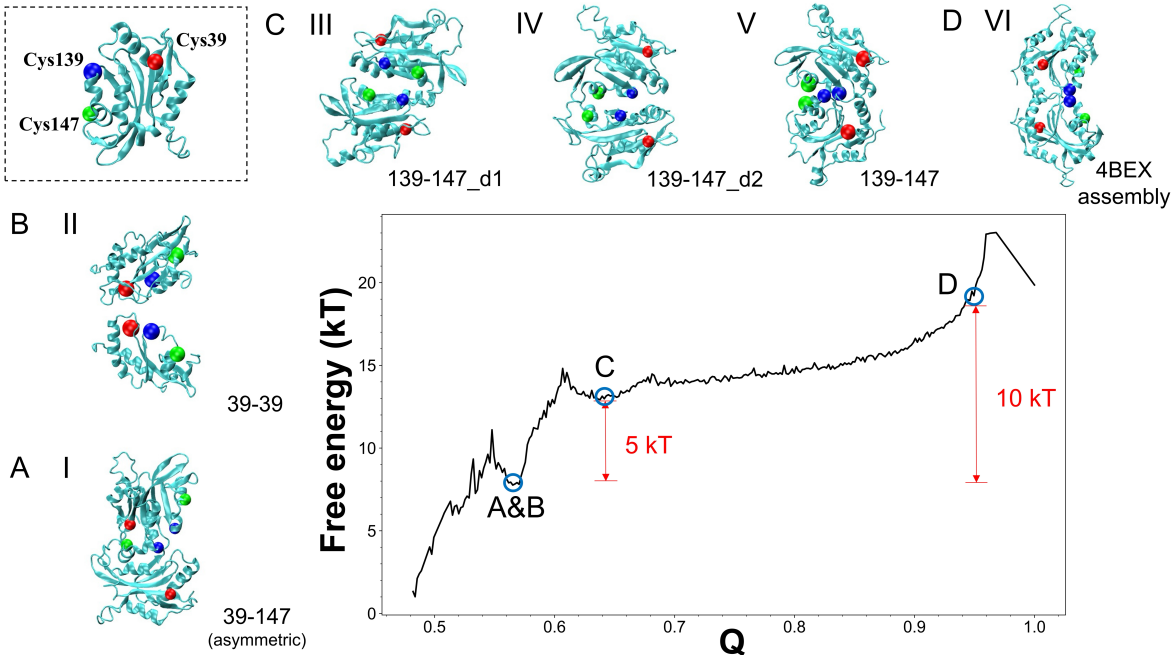


Figure 2.7. Free energy profiles of cofilin dimer candidates are shown. The monomeric cofilin structure is illustrated at the top left: the main protein chain is in cyan, while cysteine residues 39, 139, and 147 are colored red, blue, and green, respectively. This color scheme extends to all dimer structures within this figure. A, B, C, and D denote various free energy profiles, either basin-like or non-basin-like, corresponding to specific dimer configurations. (A) Candidate I: 39–147 dimer. (B) Candidate II: 39–39 dimer. (C) Candidates III–V: 139–147 dimers along with their degenerative states. (D) Candidate VI: 139–139 dimers (4BEX experimental configuration). Notably, Candidates I and II exhibit a local free energy minimum that is 5 kT less than Candidates III–V and 10 kT lower than Candidate VI. Candidate VI does not present a local free energy minimum.

Dimer candidates I (39–147) and II (39–39) are situated within the A&B basin, representing the lowest free energy state among the others, despite their initial low prevalence as observed in the clustered dendrogram. This result suggests a greater thermodynamic stability of approximately 5kT (∼3 kcal/mol at 300 K) compared to other dimer candidates, namely III, IV, and V (e.g., the 139–147 dimer). The 139–147 dimer ensemble is characterized by the free energy basin C, which exhibits a broader range, indicating higher conformational dynamics. Consequently, the 139–147 dimer encompasses several degenerate configurations, including 139–147 itself and a doublet pair, d1/d2. One distinctive feature that distinguishes d1 from d2 is the relative placement of Cys139 and Cys147 within the interface. In the case of d1, the positioning of one cofilin's Cys139 and Cys147 in the interface is on opposite sides compared to those of the other cofilin. Conversely, d2 demonstrates that Cys139 and Cys147 are situated on the same side. This side-by-side variation is primarily attributed to a rotational change at the interface. Alongside this rotational difference, it is worth noting that they also exhibit similar global Q values, as illustrated in the two-dimensional free energy landscape depicted below. The 139–139 dimer (VI) is not stable in the simulation as it does not exhibit any free energy basin, suggesting it is not a biologically relevant dimer (in line with the statement in the paper [59]).

2.5 EXPERIMENTALLY CHARACTERIZED COFILIN ASSEMBLY EXHIBITS A HIGHLY FRUSTRATED INTERFACE WHILE PREDICTED DIMER MODELS EXHIBIT A LESS FRUSTRATED INTERFACE

I conduct frustration analysis [70] for the aforementioned set of six cofilin dimer candidates (see Figure 2.8). Frustration in the context of protein analysis refers to energetically unfavorable interactions within specific regions of a given protein. Frustration analysis typically involves comparing the energy of a given interaction or region with the statistical energy distribution of different decoy states. [71, 77, 78] If the energy of the interaction is higher than the average energy of the decoys, it is considered frustrated (indicated by red lines). This signifies that the local interaction is not stable within the current protein structure. Conversely, if the energy of the interaction is lower than that of the decoys, it is termed minimally frustrated (indicated by green lines), implying that the local interaction energy cannot be further stabilized through mutation or state changes. Please refer to the Methods section for a detailed definition of frustration and a quantitative description. The dimer interfaces of candidates I, II, and V exhibit predominantly green lines with few instances of red lines, indicative of minor frustrated interactions as compared to candidates III and IV. This observation suggests that the interfaces of candidates I, II, and V are favorable binding counterparts. However, candidate VI, corresponding to the 4BEX experimentally determined assembly, displays many red lines in its frustration analysis, implying a high degree of frustrated interactions. Consequently, the candidate VI interface is unlikely to be a favorable binding interface in the biological context.

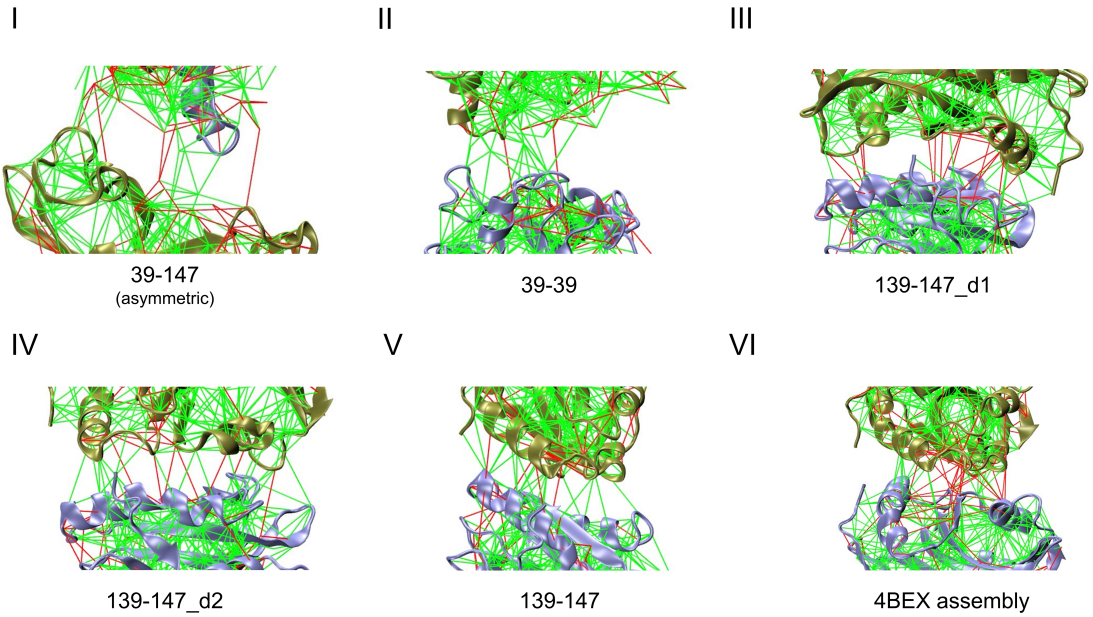


Figure 2.8. Frustration analysis of cofilin dimer interfaces. Interactions at the binding interface across various cofilin dimer candidates are analyzed for frustration. Green lines highlight minimally frustrated residue pairs, while red lines highlight highly frustrated residue pairs. Notably, Candidates I, II, and V demonstrate fewer frustrating interfaces compared to Candidates III and IV. In contrast, Candidate VI displays pronounced frustration at the binding interface.

The predicted frustrated pairs are illustrated through contact maps, presented in Figures S4–S7 in the Supporting Information. Additionally, Figure 2.9 provides a summary of the frustration analyses in terms of the percentage of frustration. Furthermore, detailed frustrated pairs have been tabulated and are available for download in an Excel file.

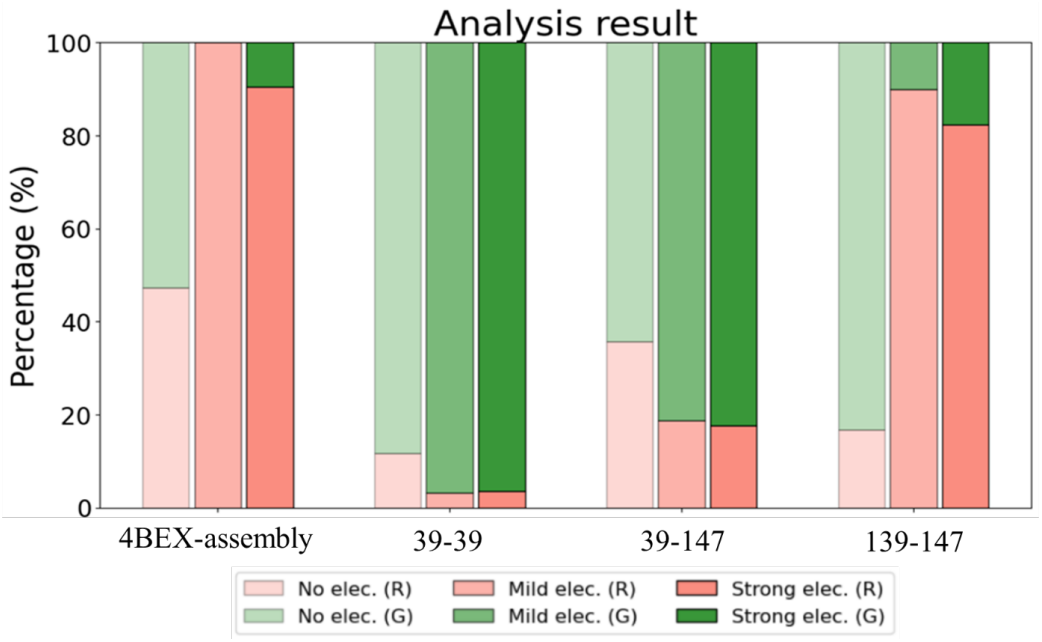


Figure 2.9. Comparison of the four cofilin dimers' mutational frustration analysis. This histogram illustrates the percentage of frustrated interactions versus minimally frustrated interactions influenced by electrostatic forces for four cofilin dimers. Each column, from left to right, is labeled as "No electrostatic (No elec.)", "Mild electrostatic (Mild elec.)", and "Strong electrostatic (Strong elec.)", corresponding to the Debye-Hückel potential with dielectric constants $\varepsilon = \infty$, 80, and 33, respectively. Frustrated residues are depicted in red, while minimally frustrated residues are shown in green.

2.6 POPULATION SHIFTS OF DIMERIC STRUCTURE FROM I TO V ARE DYNAMIC DUE TO LOW FREE ENERGY BARRIERS IN BETWEEN BASINS

In the context of relatively low free energy barriers separating the basins populated with predicted cofilin dimer configurations, I employed importance (umbrella) sampling simulations to construct two-dimensional free energy landscapes. These landscapes profile free energy is shown as a function of critical coordinates. Figure 2.10 illustrates these landscapes, depicting the free energy in terms of the global Q value and the radius of gyration (Figure 2.10A), as well as the system's potential energy as a function of the radius of gyration (Figure 2.10B).

Thermodynamically speaking, candidate V readily interconverts with candidates III and IV, owing to negligible energy barriers. It is noteworthy that both III and IV exhibit comparable Q values, approximately around Q = 0.6, within the same free energy basin. This similarity in Q values indicates that they share similar contacts formed at the interface. This similarity in the interface suggests a degree of dynamic flexibility inherent to this binding interface. It is characterized by variations in the relative orientations of individual monomers, which results in a degeneracy in energy, designated as d1 and d2. In contrast, dimer configuration V presents a significantly distinct Q value, approximately 0.9, indicating a structurally different configuration. Remarkably, the radius of gyration (Rg) for III–V falls within the range of 20 to 22 Å, consistent with their similarity in shape. Furthermore, these local minima are more thermodynamically favorable for conformational transitions into candidates I and II, as indicated by the directional arrows in Figure 2.10.

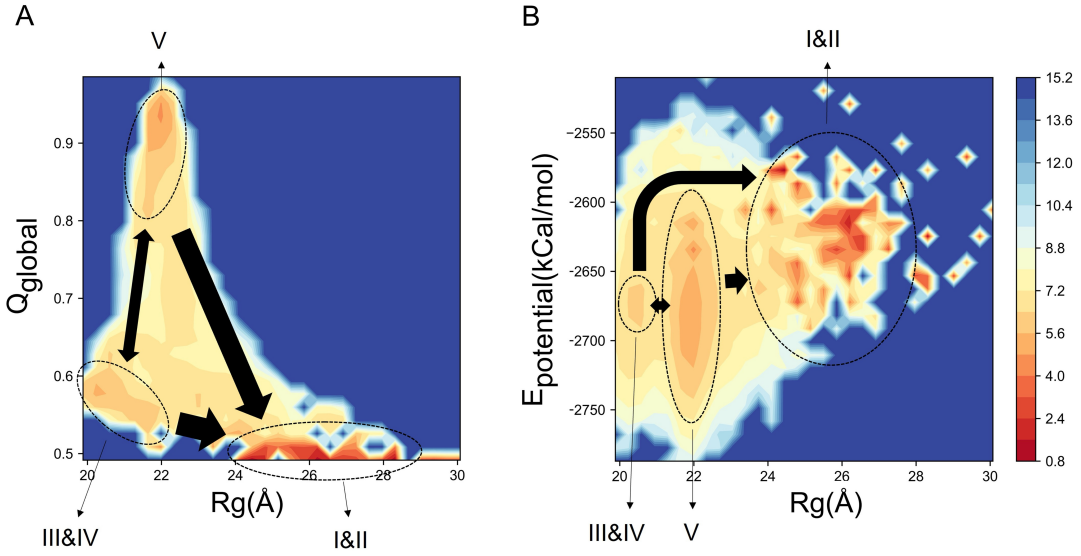


Figure 2.10. Free energy landscape of cofilin dimer transitions is explored using two-dimensional free energy surfaces. (A) Depiction of the two-dimensional free energy surface based on the radius of gyration (Rg) and the global Q value (with reference to candidate V). (B)

Free energy landscape mapped as a function of Rg and potential energy. Dimers are pinpointed at the free energy local minima, represented by dashed circles, and labeled I to V. Arrows of varying thickness indicate the probability of transitions, influenced by relative free energy differences.

2.7 COFILIN MONOMERS AND DIMERS INTERACT DIFFERENTLY WITH ACTIN FILAMENT FRAGMENTS

To explore the binding affinities of various cofilin forms with actin, I utilized the ClusPro2.0 server for docking simulations between a cofilin dimer model and a fragment of F-actin. As a control, I first demonstrated that monomeric cofilin readily docks onto an F-actin fragment where the cofilin monomer (orange) places at the junction of actin units, as depicted in Figure 2.11A─an observation consistent with its role in actin severing. Conversely, cofilin dimers (for example Model I) can only dock at the terminal ends of F-actin, not between adjacent actin units, as illustrated in Figure 2.11C. Figure 2.11B,D further illustrate that both monomeric and dimeric forms of cofilin exhibit strong binding affinity for G-actin. Based on these findings, I propose that the smaller size of the cofilin monomer facilitates its docking between actin units on F-actin, thus enhancing its role in severing. Conversely, the larger size of dimeric cofilin limits its binding to either G-actin or the terminal ends of F-actin, aligning with their respective roles in actin nucleation and assembly.

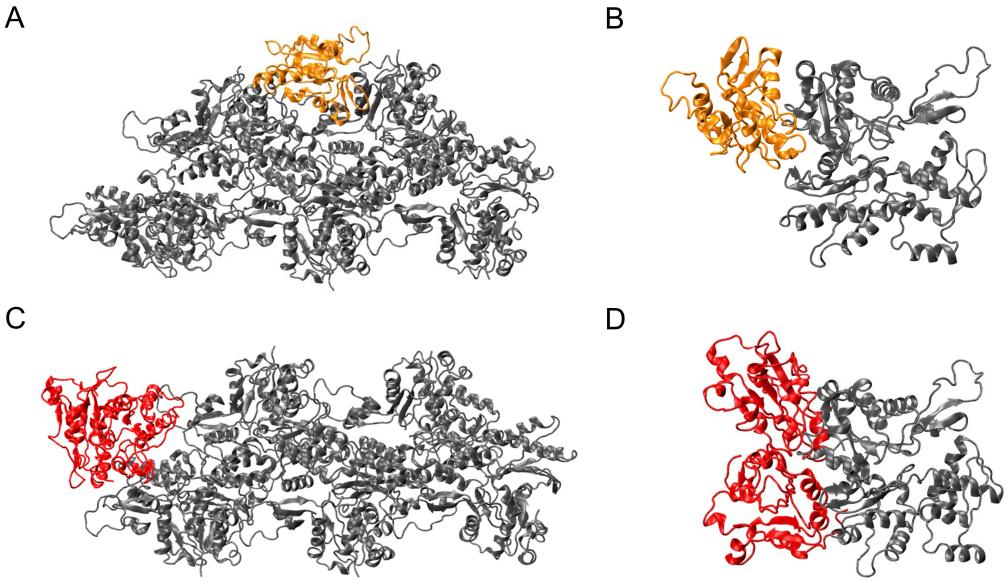


Figure 2.11. Interaction dynamics of cofilin with actin structures are predicted and shown. (A) Interaction of a monomeric cofilin (displayed in orange) with a filament-actin (F-actin) structure (shown in gray), which consists of 5 actin subunits. (B) Monomeric cofilin (orange) interacts with a G-actin structure (gray). (C) Dimeric cofilin structure (colored in red) docked onto an F-actin structure (gray), composed of 5 actin units. (D) Interaction of a dimeric cofilin (red) with a G-actin (gray). Visualization of the docking process of cofilin monomers and dimers onto filament-actin is facilitated by the ClusPro server.

2.8 DISCUSSION

Symmetry and Stability of Cofilin Dimers: Insights to Cofilin Tetramers from AWSEM Simulations

In this simulation study, I have predicted various cofilin dimer configurations that could potentially be guided by intermolecular disulfide bond formation. I should note that each cofilin dimer is characterized by distinct binding interfaces. Employing free energy analysis [79] through the AWSEM coarse-grained force field, [12] these proposed dimeric configurations display diverse population distributions and relative free energies. Among these, one particular candidate, designated as “dimer 139–147″, stands out due to its prevalence, characterized by an interface involving cysteine residues at positions 139 and 147. Additionally, I identify two other notable dimer candidates named “dimer 39–39″ and “dimer 39–147″. The configuration of “dimer 39–147″ bears a resemblance to the experimentally proposed dimer structure, while “dimer 39–39″ holds promise as a potential precursor to cofilin tetramer formation.

Some of these configurations find support in experimental evidence, either directly or indirectly, while others are theoretical predictions lacking experimental validation. Nevertheless, the free energy calculations have demonstrated the capacity to predict the stability of these configurations. Notably, among all the predicted cofilin dimer configurations, the 39–39 assembly stands out due to its remarkable stability, high degree of symmetry, and optimized charge-pair distributions at the interface, facilitated by short-range electrostatic interactions.

It is important to note that, in the current version of AWSEM, long-range effects such as Debye–Hückel potentials have not been implemented, under the assumption of a strong screening effect due to the high concentration of salt. This prediction raises the intriguing

possibility of a prevalent population when cofilin oligomerizes into larger entities ─ from cofilin dimers to tetramers, where two cofilin dimers randomly diffuse and collide.

The 39–39 dimer's C2 symmetry positions it as a strong candidate for participating in the formation of a cofilin tetramer. To test this hypothesis, I conducted AWSEM simulations. Initially, I applied two biasing forces to maintain disulfide bond equilibrium distances between the Cys39 residues of adjacent 39–39 dimers. Subsequently, these forces were removed to assess the tetramer's structural stability. As illustrated in Figure 2.12A, these simulations revealed a stable cofilin tetramer configuration, with symmetry properties of the tetrahedral-like point group (Td-like). Further analysis of mutational frustration within the tetramer, as shown in Figure 2.12B, indicated minimal frustration at the interfaces between the four chains. I speculate that this tetrameric configuration is energetically favorable, thereby providing support for its potential as a biologically relevant assembly. While higher-order reaction schemes are possible, I consider the most probable reaction channel to involve a second-order reaction, wherein two activated cofilin dimers interact to form a tetramer. Such simulations will require more sophisticated simulations; therefore, it is beyond the scope of this study.

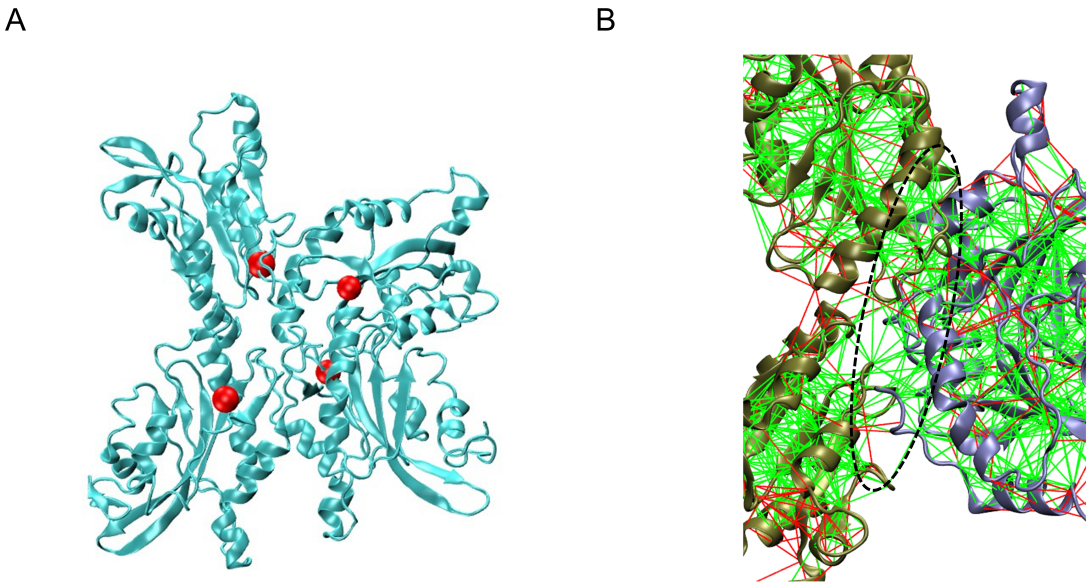


Figure 2.12. Proposed tetrameric configuration of cofilin using 39–39 dimers. (A) Display of the simulated tetrameric cofilin structure, with protein chains rendered in cyan and cysteine 39 residues emphasized using red balls. Notably, this configuration remains stable in the AWSEM simulation without any external force biases. (B) Mutational frustration analysis was conducted at the interface between two contrasting dimers, colored tan and gray. A dominant minimally frustrated interface is inferred, as the majority of interactions within the delineated ellipsoid are green, indicative of an optimized binding interface.

Given the scarcity of experimentally determined conformations, I relied on the monomeric structure derived from crystallographic data as the baseline (zero-order approximation). Although a thorough exploration of cofilin's monomeric structural ensemble is feasible, it lies beyond the scope of current study. However, such an investigation could offer valuable insights complementary to this approach. I remain receptive to future updates that may yield further discoveries in this regard.

Cofilin Oligomerization and Its Functional Implications in Actin Regulation through Post-Translational Modifications

The regulatory influence of cofilin monomers or oligomers on filamentous actin, for instance, involves a complex interplay of cooperative binding and mechanical coupling with the filament, resulting in diverse severing activities at the interfaces between bare and decorated segments. [80] Numerous computational investigations, conducted at various levels, have been dedicated to exploring the mechanical stress modulation induced by cofilin binding. [81-83] Among these investigations, mesoscopic models have proven to be particularly valuable for elucidating the mechano-elastic properties of actin filaments in the context of fragmentation. De La Cruz et al. pioneered a mesoscopic model capable of predicting bending and torsional rigidities based on polymer interaction energies, geometric constraints, and twisting-bending coupling dynamics. [81] This model, as subsequently refined to include considerations of filament helicity and occupancy, has demonstrated its efficacy in mapping filament strain energy across specific lateral and longitudinal interfaces. [83] Since this mechanical stress requires input from subtle changes in protein structures, all-atom molecular dynamics simulations play a pivotal role in furnishing these intricate structural insights at the molecular level that remain elusive in mesoscopic models. For instance, extensive molecular dynamics simulations have the capability to faithfully reproduce the subtle helical twists observed in actin filaments upon cofilin binding, [84] a phenomenon consistent with cryo-electron microscopy findings. [85] The atomistic-scale simulations have also uncovered the asymmetric dynamics of barbed and pointed ends, along with intricate conformational changes in neighboring subunits arising from differences in crossover lengths at slow-severing and fast-severing boundaries. [86, 87]

Even with AlphaFold2 [14] or RoseTTAFold [34] offers people a tool to predict protein structures from sequences accurately, these structure-only models with canonical amino acids lack the interpretation of protein functions that rely on post-translational modification (PTM). PTM is the chemical modification of amino acids in response to changes in the cellular environment. While some PTMs, such as phosphorylation, require specific enzymes to characterize cellular behaviors, oxidation–reduction (redox)-based PTMs (i.e., redox PTMs) of thiol-containing cysteines are the most common in rapid responses to shifting redox conditions. [88] One of the authors recently developed a Python-based workflow, PTM-Psi (A Python Package to Facilitate the Computational Investigation of Post-Translational Modification on Protein Structures and Their Impacts on Dynamics and Functions), that allows users to interpret how chemical perturbations caused by PTMs, particularly thiol PTMs, affect a protein's properties, dynamics, and interactions with its binding partners, leveraging either inferred or experimentally determined protein structures. They demonstrated the utility of PTM-Psi for interpreting sequence-structure–function relationships of a cysteine-rich protein derived from thiol redox proteomics data. [32]

The PTM-Psi workflow commences with a FASTA sequence, utilizing AlphaFold/OpenFold or a provided PDB file to infer its structure, followed by the incorporation of a targeted PTM into the structure. Specifically, PTM-Psi first parametrizes nonstandard residues introduced using NWChem as the quantum-mechanics backend, subsequently preparing GROMACS molecular dynamics simulations to assess the impact of PTMs on structural and dynamical changes within the protein. In the final stage, PTM-Psi facilitates the preparation of input files for small-molecule docking simulations using the Autodock Suite, enabling the

exploration of functional alterations, such as assessing the structural compatibility of a substrate within a catalytic site.

Here, cofilin oligomerization involves thiol PTMs of cysteines to form intermolecular disulfide bonds under oxidation stress. The cofilin dimer models enable to explore these possibilities based on cofilin's distinct binding interfaces. I employed a straightforward docking approach to address these questions, aiming to provide practical insights into predicting cofilin's function in terms of its oligomeric state. I showed that due to the volume exclusion, only cofilin monomers could fit at the cleft of actin filaments, while cofilin dimers have fewer options to bind to the actin filament. Under oxidation stress, cofilins may regulate their binding with actin filament through its thiol PTMs of disulfide formation. I fit this coarse-grained modeling module based on AWSEM simulations into the schema of PTM-Psi (https://github.com/pnnl/PTMPSI/tree/master/ptmpsi-awsem).

2.9 CONCLUSION

This study focuses on cofilin, a vital actin-binding protein known for its role in actin-severing and monomer recycling. Recent experimental discoveries have highlighted cofilin's capacity to form functionally diverse oligomers through intermolecular disulfide bond formation, contributing to actin nucleation and assembly. The formation of cofilin oligomers signifies its importance in regulating actin dynamics under oxidative stress. I provided a computer model to evaluate the structural conformations of these cofilin oligomers and evaluated their impact on actin binding with coarse-grained molecular simulations. These findings enhance comprehension of the intricate interplay between cofilin and actin, which has a far-reaching impact on understanding the cellular cytoskeletal dynamics.

CHAPTER 3: FORECASTING AVALANCHES IN BRANCHED ACTOMYOSIN NETWORKS WITH NETWORK SCIENCE AND MACHINE LEARNING

This chapter is based on Chengxuan Li's first author publication: Chengxuan Li, James Liman, Yossi Eliaz and Margaret S. Cheung. "Forecasting Avalanches in Branched Actomyosin Networks with Network Science and Machine Learning." The Journal of Physical Chemistry B 125 (42) (2021): 11591–11605. https://doi.org/10.1021/acs.jpcb.1c04792

Codes related to this work can be found in the GitHub repository: https://github.com/Cheung-group/Actomyosin-Avalanche_Arp2-3_MEDYAN

In this work, I explored the dynamic and structural effects of actin-related proteins 2/3 (Arp2/3) on actomyosin networks using mechanochemical simulations of active matter networks. On the nanoscale, the Arp2/3 complex alters the topology of actomyosin by nucleating a daughter filament at an angle with respect to a mother filament. At a subcellular scale, they orchestrate the formation of a branched actomyosin network. Using a coarse-grained approach, I sought to understand how an actomyosin network temporally and spatially reorganizes itself by varying the concentration of the Arp2/3 complexes. Driven by motor dynamics, the network stalls at a high concentration of Arp2/3 and contracts at a low Arp2/3 concentration. At an intermediate Arp2/3 concentration, however, the actomyosin network is formed by loosely connected clusters that may collapse suddenly when driven by motors. This physical phenomenon is called an "avalanche" largely due to the marginal instability inherent to the morphology of a branched actomyosin network when the Arp2/3 complex is present. While embracing the data science

approaches, I unveiled the higher-order patterns in the branched actomyosin networks and discovered a sudden change in the "social" network topology of actomyosin, which is a new type of avalanche in addition to the two types of avalanches associated with a sudden change in the size or shape of the whole actomyosin network, as shown in a previous investigation. This new finding promotes the importance of using network theory and machine learning models to forecast avalanches in actomyosin networks. The mechanisms of the Arp2/3 complexes in shaping the architecture of branched actomyosin networks obtained in this paper will help better understand the emergent reorganization of the topology in dense actomyosin networks that are difficult to detect in experiments.

3.1 INTRODUCTION

In muscle cells, actin filaments and myosins are organized into a striped sarcomere, [20] and in non-muscle cells, actomyosin networks tend to be isotropic, especially at the edge of cells such as the actin cortex. [20, 89] I hypothesized that the nanostructure of the actomyosin network dictates the structure and dynamics of the entire system. [20] Actin-binding proteins (ABPs) are the key drivers of changes in the local structure of actomyosin networks. One of such ABP is the Arp2/3 complex [90] which is responsible for geometrical arrangement in a global architecture by being a nucleator for branched actomyosin networks. [91, 92] The Arp2/3 complex, also known as a brancher in the system, creates a junction of a daughter filament nucleated from its mother filament and subsequently orchestrates the formation of branched actin networks. [90] Together with myosin motors, the branched actomyosin networks are responsive to mechanical

perturbations from the environment of a cell, [93] as the network organization controls contractile tension generation in a cell.

Computational models have been used to explore actomyosin contractility, [94, 95] but few of these computational models explore the effect of the Arp2/3 complex on the dynamics of the system. Despite extensive experimental and computational studies, [96-99] only recently has the computational work from Cheung group shown that the presence of the Arp2/3 complex causes sudden collapse dynamics of marginally stable actomyosin networks, called “avalanches”. [22] The avalanche possibly underscores the phenomenon of a “cyto-quake”, [100] a drastic structural change in the actomyosin network within a short period of time. The biophysical importance of this phenomenon is appreciated since it is deeply related to the structural rearrangement of the cytoskeleton. However, how this phenomenon connects to the nanoarchitecture of the branched actomyosin network orchestrated by the Arp2/3 complexes remains unclear.

In this work, I explored the impact of Arp2/3 concentration in modulating avalanches in branched actomyosin networks using coarse-grained simulations. I used the software package MEDYAN, [11] which simulates the organization of actin filaments with mechanochemical feedback from actin-binding proteins, such as those that form catch bonds (non-muscle myosin IIA motors, NMIIA) [101-103]), slip bonds (α-actinin linkers [104, 105]), and filament nucleators (Arp2/3 complexes [22, 106]). In the simulations, even though I only changed the concentration of Arp2/3 complexes while the turnover rates [20, 107] remained the same in the chemical reactions, the patterns of the temporally evolving networks were incredibly complex. Driven by a high concentration of motors, several new global features, or orders, emerge from a locally well-connected network. To quantify and even to forecast such new orders from an

inhomogeneous system that is far from equilibrium, I converted the physical networks into mathematical graphs that reveal the pattern of a higher-order scaffold within the complex network. Using network sciences tools, [39] I discovered a new type of avalanche in actomyosin networks related to a sudden change in the topology of a well-connected actomyosin network.

I then introduced these new features to train machine learning (ML) [108-111] models for forecasting these interesting far-from-equilibrium events. As an exploration of the predictability of avalanches, I trained two supervised machine learning models (support vector machine [112] and XGBoost [113]) with only the mechanical description of the actin filaments. The latter follows a gradient tree-boosting algorithm that is much more sophisticated than the former which follows a simple linear regression algorithm. In ML and supervised ML in particular, data curation and feature extraction are crucial for building reliable prediction models. I used features with physical interoperation from both polymer and network theory.

I utilized the two representative supervised machine learning models to explore order parameters for feature learning. I considered three polymer physics order parameters (the mean filament displacement, the radius of gyration, and the shape) [22] and six network theory order parameters (the density, the average clustering, the clique number, the mean closeness, the mean betweenness, and the assortativity) [39] for feature learning. Not only have I forecasted avalanches with great high probability, but I have also shown that the avalanches are mechanically dominated rather than chemically in the actomyosin network. The consideration of the features from the network theory order parameters into the training greatly improved the performance of both machine learning models by minimizing false negatives, benefiting the support vector machine model more than the XGBoost model. This work has greatly expanded the toolset available for analyzing or interpreting protein-mediated actomyosin networks.

3.2 SIMULATION

Mechanochemical Dynamics of Active Networks (MEDYAN)

I simulated the dynamics of actomyosin networks by using a coarse-grained mechanochemical model of active systems called Mechanochemical Dynamics of Active Networks (MEDYAN), developed by the Papoian group. [27, 31, 107, 114-116] The highlight of MEDYAN is its inclusion of mechanochemical feedback of the active networks, which makes the software more appropriate for research interests compared to the models used in previous studies. MEDYAN consists of four main steps in its mechanochemical loop, as described in Figure 3.1.

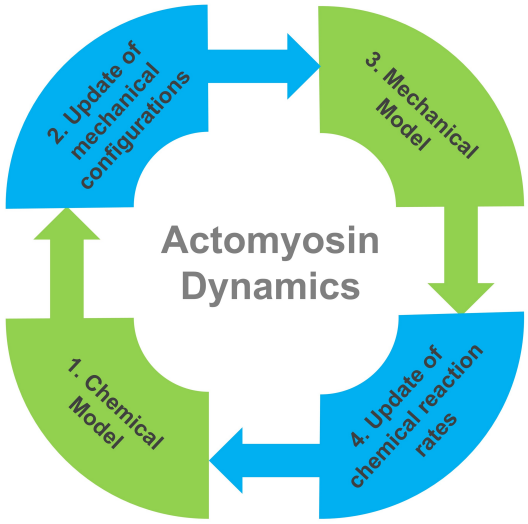


Figure 3.1. The workflow of MEDYAN consists of four steps: (1) Initiate the chemical model: the system evolves the actomyosin network with stochastic chemical reaction diffusion. (2) Update the mechanochemical configurations: the chemical reactions deform the network locally, followed by the formation of a new mechanical configuration. (3) Initiate the mechanical model: the total energy of the new mechanical configuration is minimized with the conjugate gradient

method by reaching a new equilibrium. (4) Update the chemical reaction rates: the chemical reaction rates are mechanochemically updated at the new equilibrium state. These four steps are cycled through for the entirety of the simulations. For a detailed description of MEDYAN, see ref [11].

Simulation Parameters

All simulations were confined to a three-dimensional, 1 μm^3 rigid cubical box to match the size of a typical dendritic spine. The maximum time of the simulation is set to 2000 s, and snapshots are captured every 10 s. Initially, the number of actin filaments was 50, and the filament length was 10 monomers. An example of typical snapshots of the simulations is shown in Figure 3.2, visualized with Mayavi 4.7.0. [117]

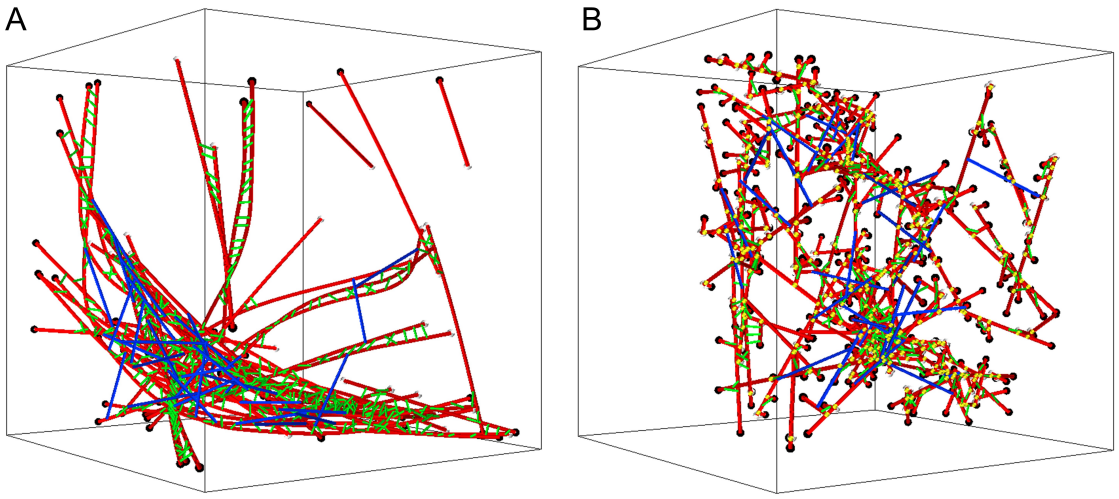


Figure 3.2. Typical snapshots of MEDYAN simulations for the unbranched actomyosin networks without Arp2/3 complexes (A) and for the branched actomyosin networks with Arp2/3 complexes (B). In both snapshots, a red cylinder represents an F-actin filament, a black bead represents a plus end of an F-actin filament, a white bead represents a minus end of an F-actin filament, a blue cylinder represents an ensemble of NMIIA motors that consists of 15–30 motor heads, a green cylinder represents an α-actinin cross-linker, and a yellow bead represents an Arp2/3 complex.

Reaction Rates and Mechanical Constants

The reaction rates used in the simulations are listed in Table 3.1. k_{p+} or k_{p-}: polymerization reactions of F-actin on the plus ends or minus ends; k_{dp+} or k_{dp-}: depolymerization reactions of F-actin on the plus ends or the minus ends; k_{bl} or k_{ubl}: binding or unbinding reactions of α-actinin linkers; k_{bm} or k_{ubm}: binding or unbinding reactions of NMIIA motors; k_{wm}: the walking reactions of NMIIA motors; k_{bf}: the branching reaction of F-actin; k_{df}: the destruction reaction of a short F-actin no longer than one segment. The branching or destruction reactions are included only in branched simulations. For the details of chemical reactions in MEDYAN, please refer to refs [11, 22]

Table 3.1. Reaction Rates in the Chemical Model of MEDYAN

	reaction rates	value
actin filaments	k_{p+}	0.151 s^{-1} [11]
	k_{p-}	0.017 s^{-1} [11]
	k_{dp+}	1.4 s^{-1} [118]
	k_{dp-}	0.8 s^{-1} [118]
linkers	k_{bl}	0.009 s^{-1} [119]
	k_{ubl}	0.3 s^{-1} [119]
motors	k_{bm}	0.2 s^{-1} [120]
	k_{ubm}	1.7 s^{-1} [11]
	k_{wm}	0.2 s^{-1} [11]
branching	k_{bf}	0.0001 s^{-1}
destruction	k_{df}	1.0 s^{-1} (only applied to actin filament with one segment)

The mechanical constants used in the simulations are listed in Table 3.2. k_{bend}: the filament bending constant; $k_{stretch}$: the filament stretching constant; k_{volume}: the volume force constant; k_{motor}: the motor stretching constant; k_{linker}: the cross-linker stretching constant; $k_{boundary}$: the boundary constant; λ: the boundary screening length constant; $k_{stretch}^{branch}$: the branching point stretching constant; k_{bend}^{branch}: the branching point bending constant; θ_0: the branching point bending angle; $k_{diheral}^{branch}$: the branching point dihedral constant; and $k_{position}^{branch}$: the branching point position constant. For detailed force field definitions in MEDYAN, please refer to refs [11, 22].

Table 3.2. Mechanical Constants in the Mechanical Model of MEDYAN

	mechanical constants	value
actin filaments	k_{bend}	2690 pN·nm
	$k_{stretch}$	100 pN/nm
	k_{volume}	100000 pN·nm^4
motors	k_{motor}	2.5 pN/nm
linkers	k_{linker}	8.0 pN/nm
boundary repulsion	$k_{boundary}$	41 pN·nm (10 k_BT)
	λ	2.7 nm
branched filament	$k_{stretch}^{branch}$	100 pN/nm
	k_{bend}^{branch}	100 pN·nm
	θ_0	~70° [90, 91]
	$k_{position}^{branch}$	100 pN·nm

Setting of Actin-Binding Protein Concentration

To explore the extent of actin-binding proteins on actomyosin dynamics, I chose several concentration ratios of actin-binding proteins to total actin. Five sets of motor and linker concentrations were selected to replicate the *in vitro* experiments from the Weitz group, [104] while the three brancher concentrations were selected to investigate the impact of

Arp2/3 concentration on the dynamics of the actomyosin network, as shown in Table 3.3. For referring to the Arp2/3 concentrations in the study, I refer to $x_{b:a} = 0.002$ as the low brancher concentration, $x_{b:a} = 0.02$ as the medium brancher concentration, and $x_{b:a} = 0.2$ as the high brancher concentration.

Table 3.3. Five Sets of Concentration Ratios of Motors or Linkers to Actin and the Three Concentration Ratios of Branchers to Actin

$x_{m:a}$	$x_{l:a}$	motors	linkers
0.01	0.01	low	low
0.01	0.5	low	high
0.5	0.01	high	low
0.05	0.1	medium	medium
0.5	0.5	high	high

$x_{b:a}$	branchers (i.e., Arp2/3)
0.002	low
0.02	medium
0.2	high

$x_{m:a}$ represents the ratio of motor concentration to actin concentration, $x_{l:a}$ represents the ratio of the α-actinin cross linker concentration to the actin concentration, and $x_{b:a}$ represents the ratio of brancher (Arp2/3) concentration to actin concentration.

3.3 DATA ANALYSIS

Polymer Physics Order Parameters

Three polymer physics order parameters, the radius of gyration (R_g), the mean displacement of filaments (δx_F), and the shape parameter (S), are used to describe the macroscopic properties of the system. Their definitions are in eqs 3.1–3.4 as below.

The radius of gyration R_g of actin filaments is used in this study to quantify the size of actomyosin networks. The size changes show the contraction or the expansion of the network. The radius of gyration of the actin filaments R_g is described in Equation 3.1.

$$R_g = \sqrt{\frac{1}{N}\sum_i^N (C_i - C_M)^2} \quad \text{(Eqn.3.1)}$$

where N is the number of actin filaments in the system, C_i is the middle point of an actin filament i, and C_M is the center of mass of the network.

R_g/R_g^i is used to describe the change of network size compared to the initial condition, where R_g is the current radius of gyration, R_g^i is the initial radius gyration at 10 seconds to allow for the initialization of the simulation.

The mean displacement of actin filaments is used to identify an avalanche by tracking the center of mass of every individual filament [22]. A sudden change of the mean filament displacement shows the abrupt transformation of the network structure.

The mean filament displacement $\delta x_F(t)$ is described in Equation 3.2,

$$\delta x_{\mathrm{F}}(t) = \frac{1}{N}\sum_{i}^{N} |C_{Mi}(t) - C_{Mi}(t-1)| \quad \text{(Eqn. 3.2)}$$

where C_{Mi} is the center of mass of filament i and N is the number of filaments in the network.

To quantify the shape of actomyosin networks, I have tracked the successive changes of shape parameter, which also describe and characterize an avalanche.

The shape parameter (*S*) is described in Equation 3.3 [121],

$$S = 27\frac{|\prod_{i=1}^{3}(\lambda_{\mathrm{i}} - \bar{\lambda})|}{(trT)^3} \quad \text{(Eqn. 3.3)}$$

where T is the geometrical inertia tensor as described in Equation 3.4, λ_{i} are the eigenvalues of T and $\bar{\lambda}$ is the average eigenvalue of T,

$$T_{\alpha\beta} = \frac{1}{2N^2}\sum_{i,j=1}^{N}(r_{i\alpha} - r_{j\alpha})(r_{i\beta} - r_{j\beta}) \quad \text{(Eqn. 3.4)}$$

where N is the number of beads in the network. Beads are the ends of actin subunits in the filaments, $r_{i\alpha}$ is the projection of a bead to an axis where *i* is the index of beads and α is the index for *x, y, z* axis.

Network Theory Order Parameters

Network theory is utilized in this study to capture the hidden properties of the actomyosin network. Yossi Eliaz has previously implemented network theory to characterize the complex topology in actomyosin dynamics. Yossi's work reveals hidden properties involving uneven changes in the shape or the size of a network that are not captured by conventional order parameters. [39]

I followed the steps to build a mathematical graph, *G (V, E*), from physical actomyosin networks based on the proximity map of actin filaments. A proximity map is a matrix determined from the positions of actin filaments, where the coordinates of the plus end of each actin subunit are recorded as the position of a node (*V*). I chose 20 nm as the cutoff distance for constructing the proximity map; if the distance between a pair of nodes is less than 20 nm, it is assigned an edge (*E*) on the graph. To profile the topological arrangement that evolves into hierarchical, higher-order complexes in a network of actomyosins, I opted for tracking the filaments that are generally in close contact but not in direct contact by ignoring the chemical connectivity of filaments formed by actin-binding proteins such as a motor or a linker. The lengths of motors and linkers are 200 ± 25 and 35 ± 5 nm, respectively. [122, 123] I also do not include the two adjacent actin monomeric units in a filament measured at 27 nm. Therefore, the choice of short cutoff at 20 nm in constructing a proximity map will satisfy the purpose of tracking the emergent features of a complex network while not capturing the chemical connectivity of actin.

In this way, I converted the actomyosin networks into mathematical graphs. The conformation of the mathematical graph from the physical actomyosin network and the calculation of network theory order parameters were performed with the Python package NetworkX. [124]

Graph (*V, E*): The components of a graph *G (V, E*) include *V*, a set of vertices (also called nodes or points, filled circles in Figure 3.3), and *E*, a set of edges that connect the vertices (black solid lines in Figure 3.3). Figure 3.3A shows an example of a graph composed of six nodes and eight edges. The degree of a node is the number of edges that are connected with it. For example, the degree of node a is 1, while the degree of node b is 4.

Assortativity: Figure 3.3B shows an example graph that has lower assortativity than the graph in Figure 3.3A. The assortativity, ρ, measures the tendency of nodes with similar degrees to be directly connected. The graph in Figure 3.3A has most of the nodes with similar degrees (b, c, d, and e) connected to each other, while the nodes with similar degrees in the graph in Figure 3.3B are not connected. Therefore, the graph in Figure 3.3A has higher assortativity than the graph in Figure 3.3B. I employed the *Z*-score of the time derivative of ρ, $Z_{\Delta\rho}$, as one of the order parameters to probe undetected changes in the patterns. $Z = \frac{\rho - \mu}{\sigma}$, μ is the mean, and σ is the standard deviation of ρ.

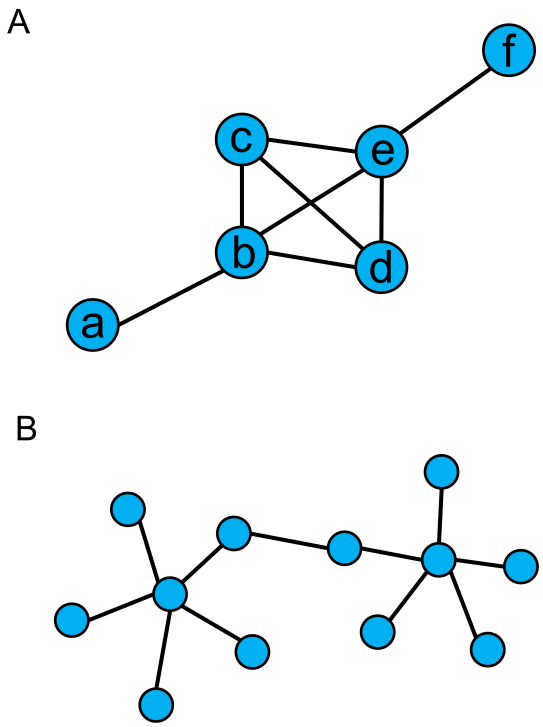


Figure 3.3. Illustrating graph properties with examples: (A) an example graph with high assortativity; (B) an example graph that has lower assortativity than the graph in (A).

Detailed descriptions and definitions of the parameters are shown below in eqs 3.5–3.11.

Density measures the number of edges observed in a graph compared to the maximum number that could be observed. The density of a network shows the number of connections inside the network; higher density means more connections.

The density of a graph *G* is:

$$d = \frac{2m}{n(n-1)} \qquad \text{(Eqn. 3.5)}$$

where *n* is the number of nodes and *m* is the number of edges in *G*.

A clique is a subset of nodes that are all directly connected. The largest clique in the graph of Figure 3.3A is formed by nodes b, c, d and e. By definition, the clique number is the number of nodes in the largest clique of a graph; thus, the clique number of this graph is 4. The clique number of a network shows the size of the biggest 'family' in the network.

Average clustering measures the tendency of nodes in a graph to cluster together. It shows the clustering degree of a network, network with more clusters has higher average clustering value.

The average clustering coefficient of a graph *G* is:

$$C = \frac{1}{n}\sum_{v \in G} c_v \qquad \text{(Eqn. 3.6)}$$

where *n* is the number of nodes in *G*, c_v is the clustering coefficient of the node (Figure 3.4):

$$c_v = \frac{2T(v)}{deg\,(v)(deg(v)-1)} \qquad \text{(Eqn. 3.7)}$$

where *T(v)* is the number of triangles through node *v* and *deg(v)* is the degree of *v*.

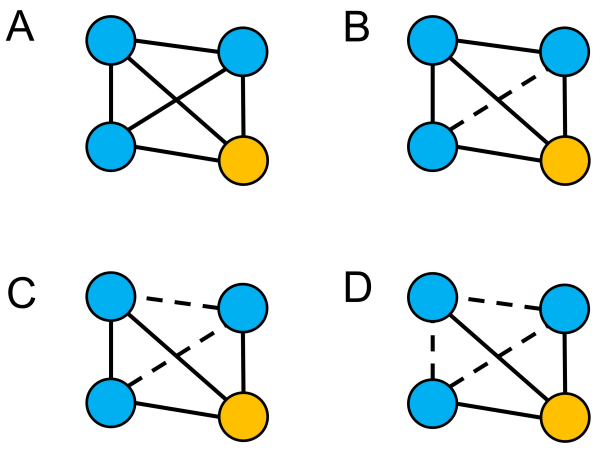


Figure 3.4. The definition of clustering coefficient. The clustering coefficient is the ratio of the observed number of triangles that include a node to the maximum possible number of triangles that include this node. In this figure, the clustering coefficient of the yellow node is measured. The solid lines represent formed edges, and dashed lines represent unformed edges. When all edges are formed (A), there are three possible triangles that include the yellow node. Based on the definition, the yellow node has clustering coefficient of 1 in A, 2/3 in B, 1/3 in C and 0 in D.

The betweenness centrality (or betweenness) of a node calculates the number of shortest paths between other nodes that pass through this node. It measures the centrality of a node by how often it acts as a 'bridge' between other pairs of nodes.

Betweenness centrality of a node v is:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad \text{(Eqn. 3.8)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s,t. If $s=t$, $\sigma(s,t)=1$, and if $v \in s,t$, $\sigma(s,t|v)=0$.

The closeness centrality (or closeness) of a node measures its average farness (inverse distance) to all other nodes. Nodes with a higher closeness score have shorter distances to all other nodes.

The closeness centrality of a node u is:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \quad \text{(Eqn. 3.9)}$$

where *d(v, u)* is the shortest-path distance between *v* and *u*, and *n-1* is the number of nodes that can reach *u*.

Degree assortativity coefficient (or assortativity) measures the similarity of connections in the graph with respect to the node degree. In other words, it measures the tendency of nodes with similar degree to be directly connected.

The degree assortativity coefficient of a graph *G* is [125]:

$$r = \frac{\sum_{x,y} xy(e_{xy} - q_x q_y)}{\sigma_q{}^2} \qquad \text{(Eqn. 3.10)}$$

where e_{xy} is the joint probability distribution of the degrees of two nodes, which is a symmetric matrix and its element equals the number of edges in the graph that connect a node of degree x with another node of degree y:

$$e_{xy} = |\{(u, v): u, v \in E \ \wedge k_u = x\,, k_v = y\}| \qquad \text{(Eqn. 3.11)}$$

q_x and q_y are the fraction of edges that start and end at nodes with degrees x and y, σ_q is the standard deviation of the distribution of q_i values.

Assortativity reveals the morphology of a graph. Graph with high assortativity has ‘centers’ within which the nodes with similar degree are connected to each other (Fig 3.3A), while graph with low assortativity has aster-like structures (Fig 3.3C).

Machine Learning Models to Forecast Avalanches

To predict avalanches in the mesoscopic simulations of actomyosin dynamics, I adopted two types of supervised learning models for training the data set simulated with MEDYAN software: the SVM (support vector machine) model [112] and the XGBoost model. [113] Although both are widely used in machine learning, the XGBoost model has been proven to be more sufficient (better performance) and less expensive (consuming fewer computing resources) than linear regression models in most cases, while the performance of the SVM model depends highly on the choice of the kernels. [126-128] I trained the data with a polynomial kernel in the SVM model provided by the python scikit-learn 0.23.2 package. [112] For the XGBoost model, I trained the data using the python XGBoost 1.3.0 package. [113]

The data set used for training consists of 335 snapshots that precede avalanches as a positive data set ("avalanche"), while for a negative data set ("no avalanche"), I used 493 other snapshots that were not succeeded by an avalanche. All these avalanches were selected from the MEDYAN simulations with the parameters of $x_{b:a} = 0.02$ and $x_{m:a} = 0.5$, representing the actomyosin dynamics at a medium brancher concentration and a high motor concentration, a condition favorable for the avalanches because of its rich connections nucleated by the branchers and abundant forces generated by the motors.

I have justified the upper threshold of 50 nm in the mean filament displacement (δx_F) for the selection of avalanche events from the snapshots of trajectories. I selected the individual snapshots with δx_F over the upper threshold and assigned them to the positive data set for training machine learning models. I visually inspected these snapshots that indeed there is a

structurally large change over a short period of time. I also selected snapshots with a mean filament displacement less than the lower threshold of 20 nm and assigned them into a negative data set for machine learning models. I confirmed them (without an avalanche) by visual inspection. I have included more details in Figure 3.5.

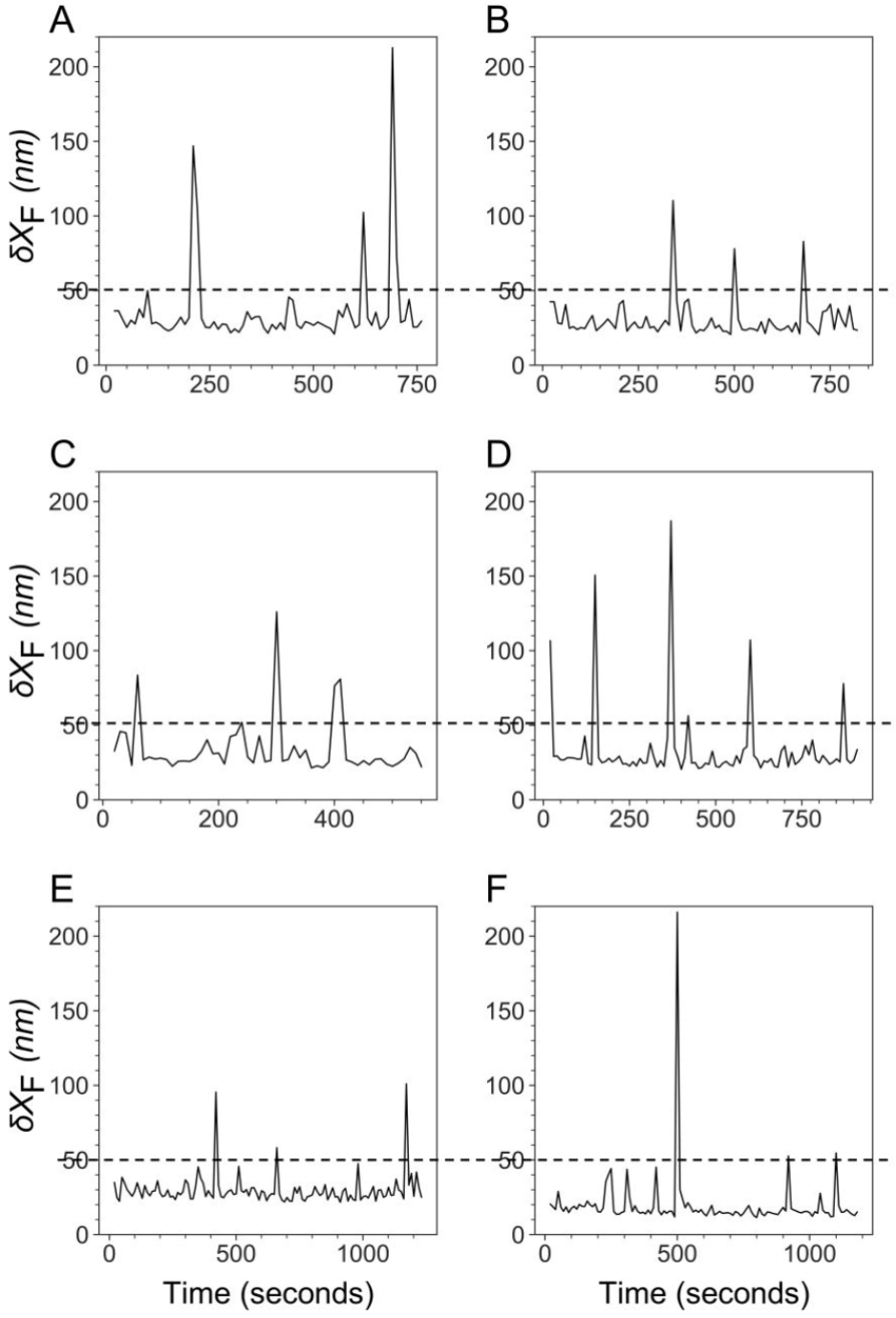


Figure 3.5. Mean filament displacement vs. time from representative simulations with avalanches. δxF is the mean displacement of filaments in the network. A-F show six simulations with the same high motor, low linker and medium brancher concentrations. Dashed horizontal lines represent the upper threshold of 50 nm.

For each snapshot in both the positive and negative the data set, I computed nine order parameters, or features, to characterize the morphologically complex structures in an actomyosin network. There are three from polymer physics: the radius of gyration (R_g), the mean displacement (δx_F), and the shape parameter (S). There are an additional six from network theory: the density, the clique number, the average clustering, the mean betweenness, the mean closeness, and the assortativity (ρ). The total data set of 335 positives and 493 negatives is selected into two parts by the Python scikit-learn package: [112] 60% for training the model and 40% for testing the performance of the model. The training and testing sets, training_data.csv and testing_data.csv, are provided in https://doi.org/10.1021/acs.jpcb.1c04792.

Data Analytics for the Machine Learning Models. Quality Indicators

A true positive (TP)/true negative (TN) is an outcome where the model correctly predicts the positive/negative class. A false positive (FP)/false negative (FN) is an outcome where the model incorrectly predicts the positive/negative class. The true positive rate (TPR), false positive rate (FPR), precision, and recall are defined as follows:

$$TPR = TP/(TP+FN) \quad (1)$$

$$FPR=FP/(FP+TN) \quad (2)$$

$$\text{precision} = TP/(TP+FP) \quad (3)$$

$$\text{recall} = TP/(TP+FN) \quad (4)$$

With such quality indicators, I measured the performance of the machine learning models by using the receiver operating characteristic curve, the precision–recall curve, the area under the curve, and the confusion matrix as defined below:

Receiver operating characteristic (ROC) curve: the ROC curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings randomly selected by the Python scikit-learn package.

Precision–recall (PR) curve: the PR curve is created by plotting the precision against the recall at various threshold settings randomly selected by the Python scikit-learn package.

The area under the curve (AUC): the ratio of the area under the curve to the total area in a figure. The AUC ranges from 0 to 1. A larger AUC value indicates a better performance of a model.

Confusion matrix: a confusion matrix is used to present the details in the performance of the models. The confusion matrix is composed of true class in rows and predicted class in columns. The numbers in the table are the numbers of the four quality indicators (TP, TN, FP, and FN) described above. A confusion matrix with higher true positives (TPs) and true negatives (TNs) indicates a more accurate machine learning model.

3.4 CONTENT OF THE ARP2/3 COMPLEX MODULATES THE DYNAMICS OF ACTOMYOSIN NETWORKS

To explore the influence of the Arp2/3 concentration that impacts the content of branched networks in the contractility of an actomyosin network, I compared the R_g/R_g^i time courses of actomyosin systems with low, medium, and high brancher concentrations, as shown in parts A, B, and C of Figure 3.6, respectively. R_g is the radius of gyration and R_g^i is the initial condition of R_g. The increase and decrease in R_g/R_g^i over time indicate the expansion and contraction of the actomyosin network.

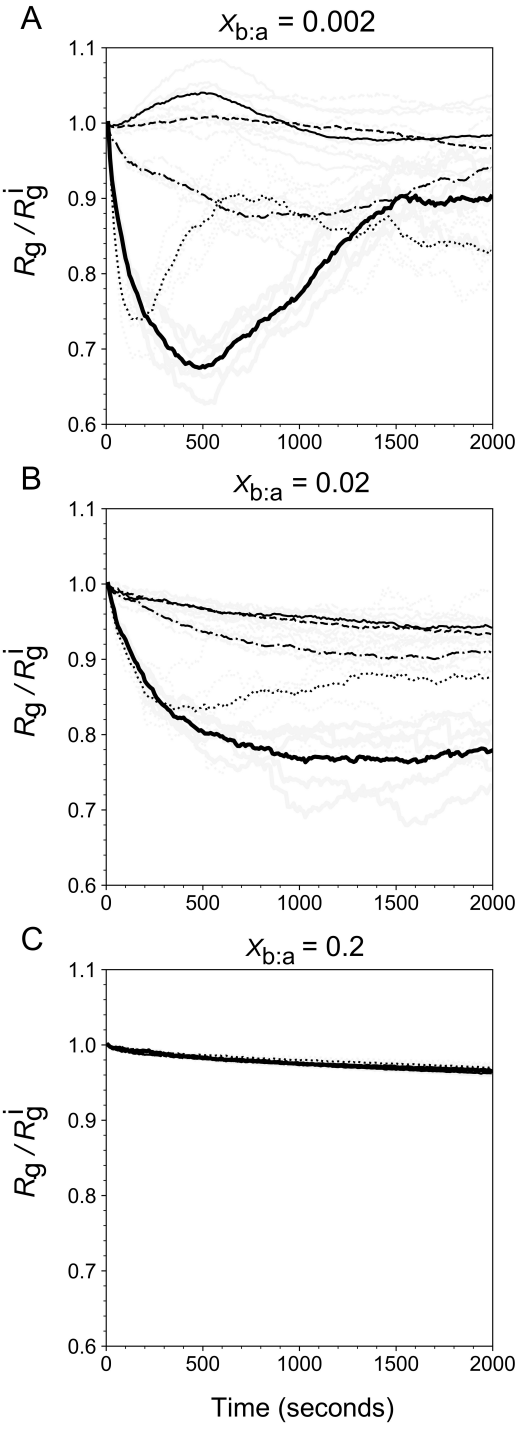


Figure 3.6. Time courses of Rg/Rg^i in branched networks with low, medium, and high brancher concentrations. The black lines in (A), (B), and (C) show the time courses of R_g/R_g^i of simulations with different motor and linker concentration pairs under the conditions of low ($x_{b:a}$ = 0.002), medium ($x_{b:a}$ = 0.02), and high brancher concentrations ($x_{b:a}$ = 0.2), each averaged

from five simulations with the same initial conditions (averaged results are shown in black, and the five original trajectory replicates are shown in light gray). Thin solid lines represent networks with low motor and linker concentrations ($x_{m:a} = 0.01$, $x_{l:a} = 0.01$); dashed lines represent networks with low motor and high linker concentrations ($x_{m:a} = 0.01$, $x_{l:a} = 0.5$); dotted lines represent networks with high motor and low linker concentrations ($x_{m:a} = 0.5$, $x_{l:a} = 0.01$); dashed dotted lines represent networks with medium motor and linker concentrations ($x_{m:a} = 0.05$, $x_{l:a} = 0.1$); and thick solid lines represent networks with high motor and linker concentrations ($x_{m:a} = 0.5$, $x_{l:a} = 0.5$). The definitions of $x_{m:a}$, $x_{l:a}$, and $x_{b:a}$ are described in Table 3.3.

At low Arp2/3 concentrations ($x_{b:a} = 0.002$) in Figure 3.6 A, R_g/R_g^i varies broadly with motor and linker concentrations. At low motor and linker concentrations ($x_{m:a} = 0.01$, $x_{l:a} = 0.01$, thin solid line in Figure 3.6A), R_g/R_g^i initially increases over time, reflecting the expansion of the networks. When motor or linker concentrations both increased ($x_{m:a} > 0.01$ and $x_{l:a} > 0.01$), the contractility of actomyosin networks reacted differently by either the active (i.e., motors) or passive actin-binding proteins (i.e., linkers). R_g/R_g^i shows that an increase in motor concentration always promotes contraction of an actomyosin network (dotted line compared to thin solid line, thick solid line compared to dashed lines in Figure 3.6A, $x_{m:a} = 0.5$ compared to 0.01). Meanwhile, either a high or low linker concentration leads to a similar R_g/R_g^i over time at a steady state (thick solid line and dotted line in Figure 3.6 A, $x_{l:a} = 0.5$ and 0.01). These observations are consistent with the findings from prior experimental [104] and theoretical investigations. [22]

R_g/R_g^i varies less over time when I increase the ratio of the Arp2/3 concentration over actin to 0.02 (Figure 3.6B), but the profiles are still modulated by the motor and linker concentrations. When the ratio is increased to the highest level at 0.2 (Figure 3.6C), R_g/R_g^i does not significantly change over time regardless of the motor and linker concentrations. The dynamics in these networks slow down significantly. The actomyosin network remains highly

branched and static after the motors are unable to walk along the filaments saturated with Arp2/3 complexes.

3.5 NETWORK THEORY FACILITATES DATA VISUALIZATION BY CONVERTING A PHYSICALLY COMPLEX NETWORK TO A MATHEMATICAL GRAPH

I visualized snapshots of actomyosin networks at low, medium, or high Arp2/3 concentrations while the motor and linker activities remained the same in Figures 3.7A, 3.7C, or 3.7E, respectively. Their structural morphologies are distinctive, and the tensions are distributed unevenly throughout a complex network. The length of the filaments at a higher Arp2/3 concentration (Figure 3.7E) are shorter, and the tension within them is lower compared to those at a lower Arp2/3 concentration (Figure 3.7A). When the Arp2/3 concentration is medium (Figure 3.7C), the filament lengths and tension degree are also at a medium level. I speculated that there is a causal relationship between the complexity of the physical network and the distribution of tensions, which dictates the emergent dynamics of actomyosin networks.

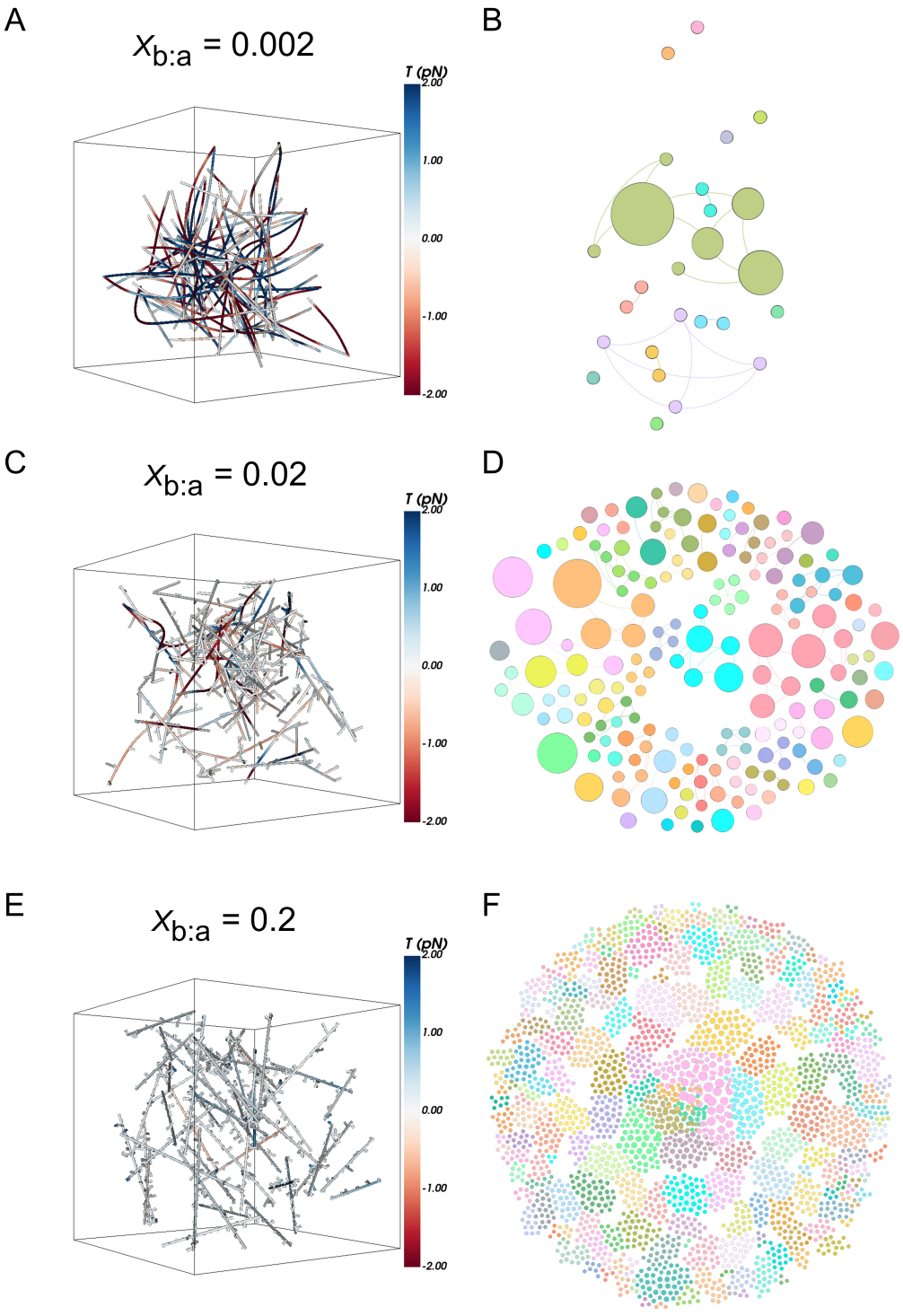


Figure 3.7. Graph networks of physical actomyosin networks at several Arp2/3 concentrations at 500 s. (A), (C), and (E) show snapshots of the simulations at low, medium, and high Arp2/3 concentrations, while they all have the same high motor and linker concentrations. The actomyosin filaments are colored by tension in pN, while motors, linkers, and branchers are not shown on the snapshot. A positive value of tension represents stretched filaments, while a

negative value of tension represents compressed filaments. (B), (D), and (F) are graph representations of the actomyosin network from (A), (C), and (E), respectively. I filtered out nodes with a node degree less than 3 and the nodes with a self-loop for clear visualization. The layout algorithm by Yifan Hu with default parameters in Gephi 0.92 [129] is used. The size of a node depends on its betweenness centrality, while the nodes are colored according to the identification number of a component (a component is a group of nodes that are connected inside the group but not connected to any nodes outside the group).

However, the first step is to quantify the complexity emerging from an actomyosin network by converting a physical network into a mathematically representative graph. This transformation reveals the hidden features of actomyosin networks in the MEDYAN simulations by filtering out unimportant information as noise. I visualized typical snapshots from the simulations with low ($x_{b:a} = 0.002$), medium ($x_{b:a} = 0.02$), or high ($x_{b:a} = 0.2$) Arp2/3 concentrations in Figures 3.7A, 3.7C, or 3.7E, respectively, in mathematical graphs in Figures 3.7B, 3.7D, or 3.7F. Indeed, these mathematical graphs are distinctive in both sizes and structures and reveal hidden features through the sizes and connectedness of a node (which represents an actin filament). At a local level, the number of nodes with high degrees increases with the concentration of Arp2/3. In addition, the mathematical graphs show that there are more connections between nodes (i.e., filaments) at higher Arp2/3 concentrations than those at lower Arp2/3 concentrations.

The network theory order parameters excel at revealing hidden features at a nonlocal level, which is a challenging task in a nonhomogeneous system. I revealed the importance of these hidden features by measuring the "betweenness centrality" of a node. When the Arp2/3 concentration is low, there are only a few nodes with high betweenness (shown by a large node) on the graph in Figure 3.7B, corresponding to a highly connected hub of actomyosin networks in Figure 3.7A. When the Arp2/3 concentration is medium (Figure 3.7D), several high-betweenness components emerge, corresponding to several delocalized and sparsely connected

clusters within the physical network (Figure 3.7C). When the Arp2/3 concentration is high in Figure 3.7F, compared to the networks with low Arp2/3 concentration, there are more nodes with low betweenness emerge (shown by decentralized, small nodes). I interpreted that these nodes are spatially far apart from one another. They are categorized into several communities with different colors. The layout approach created by Yifan Hu is useful to visualize the community in a complex network by arranging the nodes hierarchically on a mathematical graph in Figures 3.7B and 3.7D. Overall, there was no clearly defined central hub in the network (Figure 3.7E). Physically, these nodes are the individual mother filaments with their daughter filaments (Figure 3.7E), while the former are not connected to one another.

3.6 NETWORK THEORY ORDER PARAMETERS REVEAL ASTER-LIKE FEATURES FROM PHYSICALLY COMPLEX ACTOMYOSIN NETWORKS AT VARYING ARP2/3 CONCENTRATIONS

At the community level, network theory order parameters such as assortativity, ρ, provide the outlook of the hierarchical structure within a network. The decrease in assortativity indicates the change of a network morphology from a “centered” to an “aster” topology, as shown in Figure 3.3. While I varied the Arp2/3 complex concentration in the simulations of actomyosin dynamics in Figure 3.8, I showed that an increase in Arp2/3 complex concentration leads to a decrease in assortativity, revealing an altered topology within a network to become aster-like.

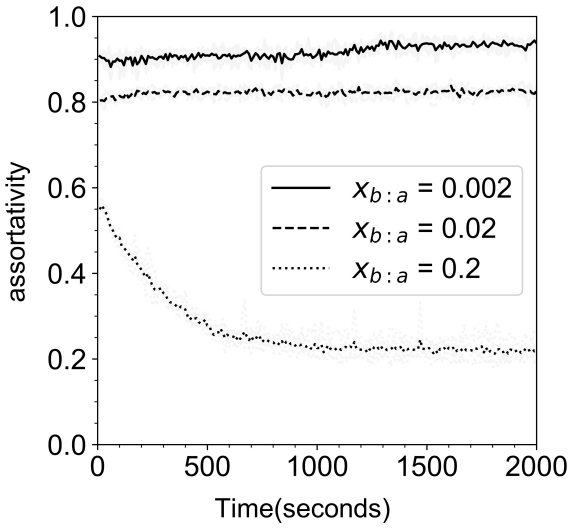


Figure 3.8. Assortativity of networks with different Arp2/3 concentrations. The figure shows the assortativity for actomyosin networks with low, medium, and high brancher concentrations but the same high motor and linker concentrations. Solid lines represent networks with low Arp2/3 concentrations ($x_{b:a}$ = 0.002); dashed lines represent networks with medium Arp2/3 concentrations ($x_{b:a}$ = 0.02); and dotted lines represent networks with high Arp2/3 concentrations ($x_{b:a}$ = 0.2). Each black line is the averaged result of five simulations with the same initial conditions (the five replicates are shown in light gray).

Similarly, in Figure 3.7E, where the Arp2/3 complex concentration is high, I observed numerous short branches on the actin filaments, which is equivalent to the composition of having many “aster-like” branches in the network (Figure 3.7F). The morphology of the actomyosin networks with high Arp2/3 concentrations is totally different from the actomyosin networks with low Arp2/3 concentrations, resulting from the formation of more short branches on the actin filaments mediated by Arp2/3 nucleation behavior.

3.7 CHANGE IN ASSORTATIVITY CAPTURES A NEW TYPE OF AVALANCHE RESULTING FROM DISRUPTION IN THE HIERARCHICAL ORGANIZATION OF AN ACTOMYOSIN NETWORK

An avalanche, or a cyto-quake, is a sudden structural rearrangement of the networks captured by the positional changes of the filaments, δx_F, in simulations or experiments. [22, 130] In the computational investigation from prior work, [22] I have characterized two types of avalanches by employing the polymer physics order parameters such as the radius of gyration (R_g) and the shape parameter (S). R_g and S measure distinctive properties of a network in terms of its size and shape, respectively. However, they do not capture the rearrangement of topology within a network that may distinctively impact neither the overall shape nor the size of a network. In this work, by varying the concentration of Arp2/3, I discovered another classification of avalanche caused by the hierarchical reorganization of a network (Figure 3.9). Such changes are structurally complex and mostly hidden by layers of information. I captured them with the aid of unique network theory order parameters.

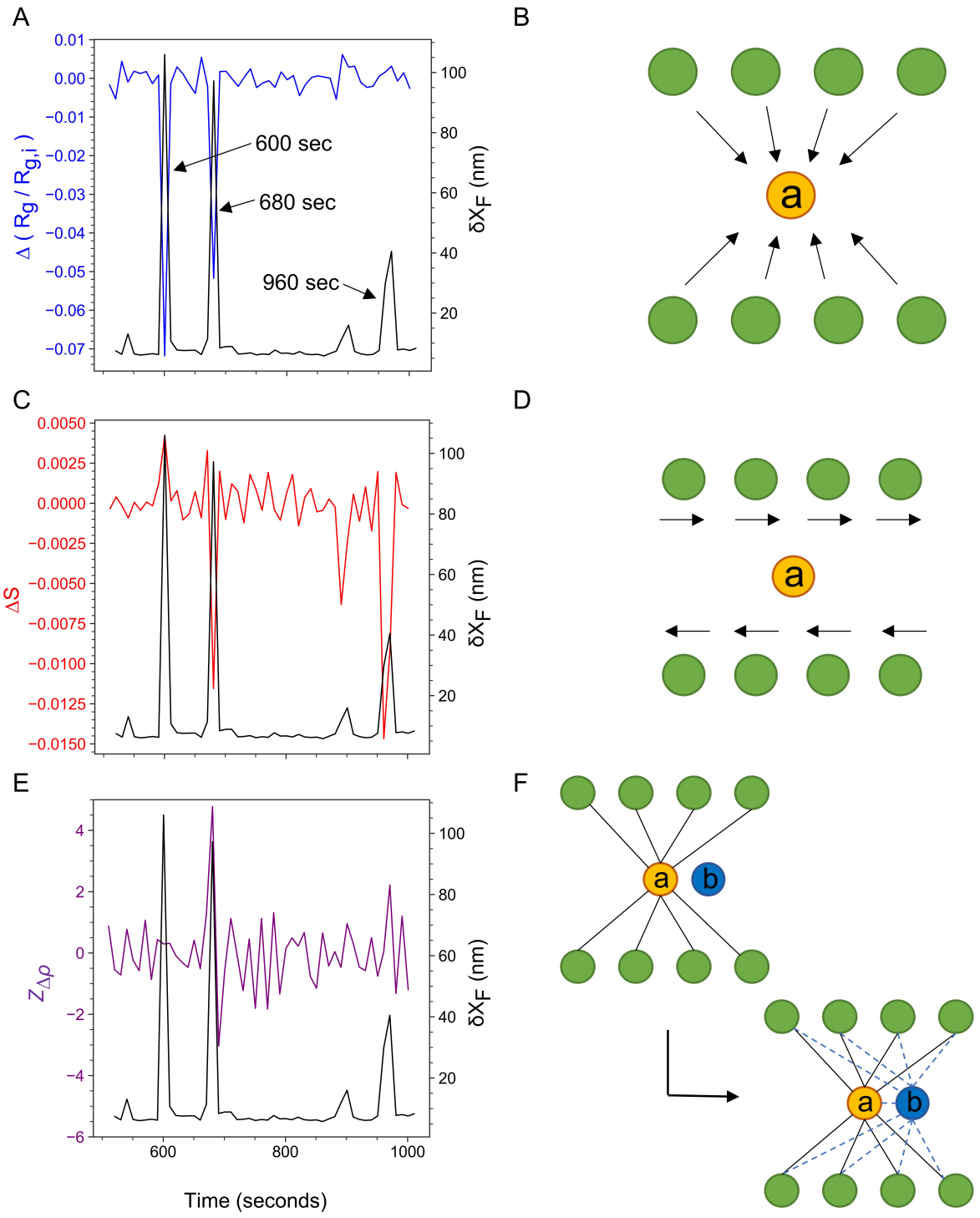


Figure 3.9. Assortativity captures the third classification of avalanche. Δx_F represents the mean displacement of filaments in the network. (A) $\Delta(R_g/R_g^i)$ represents the time derivative of the ratio of the current R_g and the "initial" R_g at 10 s, (C) ΔS represents the time derivative of the shape parameter, (E) $Z_{\Delta\rho}$ represents the Z-score of the time derivatives of assortativity, ρ, an order parameter measuring the topology of a network. (B), (D), and (F) are the cartoons that illustrate

the changes in the size, shape, and topology of a network, corresponding to (A), (C), and (E), respectively. The black arrows in (B) and (D) represent the moving directions of the green nodes relative to node a. The black lines in (F) represent the edges from the green nodes to node a before the avalanche. The blue dashed lines represent newly formed edges connecting other nodes to a nearby node b during the avalanche. An increase in $Z_{\Delta\rho}$ indicates an increase in assortativity of the network. Illustratively, node a and node b now form a "center-like" cluster during the avalanche. The example simulation has high motor, low linker, and medium brancher concentrations.

I ranked the nine order parameters (three from the polymer physics and six from the network theory order parameters) by comparing their Pearson correlation against one another in Figure 3.10. Interestingly, ρ is strongly anticorrelated to the other order parameters, signifying its importance in capturing emergent properties. Indeed, the change in ρ, Δρ, is useful to probe the emergence of a higher-order organization from uneven distribution of local nodes.

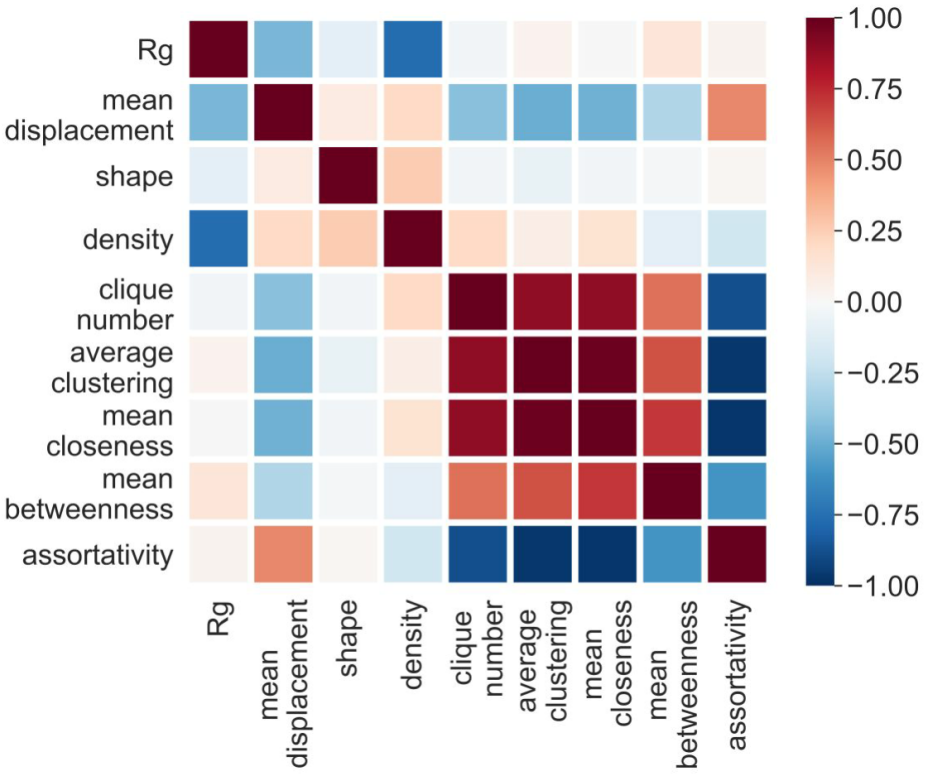


Figure 3.10. The Pearson correlation matrix of the three polymer-physics and six network theory order parameters. *R*g represents the radius of gyration of the network. Positive correlation value represents correlated while negative correlation value represents anti-correlated. Larger absolute correlation value indicates stronger correlation. The matrix is generated from 1000 simulation snapshots with various initial simulation settings.

I demonstrated the new characterization of the avalanches by showing a segment of the trajectory as an example in Figure 3.9 that it is entirely different from the previous two kinds. The plot of δx_F over time in Figure 3.9 indicates three avalanches at 600, 680, and 960 s. I also showed the time derivatives of R_g/R_g^i, the shape parameter S, and the Z-score of time derivative of assortativity in Figures 3.9A, 3.9C, and 3.9E and provided the illustrations corresponding to the movements in Figures 3.9B, 3.9D, and 3.9F. The avalanche at 600 s coincides with a sharp decrease in $\Delta(R_g/R_g^i)$ in Figure 3.9A, indicating that this is a contraction (Figure 3.9B). The avalanche at 960 s coincides with a large drop of ΔS (Figure 3.9C), signifying that the network deforms under shear and the shape becomes more oblate (Figure 3.9D).

The avalanche at 680 s not only coincides with contraction and shape changes but also uniquely coincides with a peak in $Z_{\Delta\rho}$, the Z-score of $\Delta\rho$ in Figure 3.9E, while the events at 600 and 960 s do not. The sudden increase in assortativity at 680 s indicates that the topology of the network changes by altering the hierarchy of connected nodes. I illustrated this movement in Figure 3.9F and focused on the connectivity from the surrounding green nodes to node (a) and then to nearby node (b) during an avalanche. Once other nodes that connect to node (a) also connect to node (b), the node degree and the betweenness of node (b) grow. Consequently, the assortativity of the network increases by 0.11 because node (a) and node (b) that now share a

similarly large node degree and high betweenness connect to each other (e.g., a tongue-in-cheek remark would be “guilt by association”). The connectivity of nodes changes from a “aster-like” to a “centered” topology. Although this new type of avalanche at 680 s may still carry changes in size and shape, it is an entirely new type of avalanche distinctive to the other two avalanches at 600 and 960 s.

Once I discovered the third type of avalanche at 680 s (Figure 3.9E), I visualized the physical networks (Figures 3.11A and 3.11B) on a mathematical graph in Figures 3.11C and 3.11D before and after the avalanche, respectively. Then, I computed non-local order parameters, particularly the betweenness centrality, to detect hidden patterns in a complex network. After the avalanche at 680 s, I observed the emergence of clustered green nodes with high betweenness in Figure 3.11D, which did not exist before the avalanche (Figure 3.11C). The formation of this center with high betweenness nodes increases the assortativity of the network, reflected as a peak in Figure 3.9E.

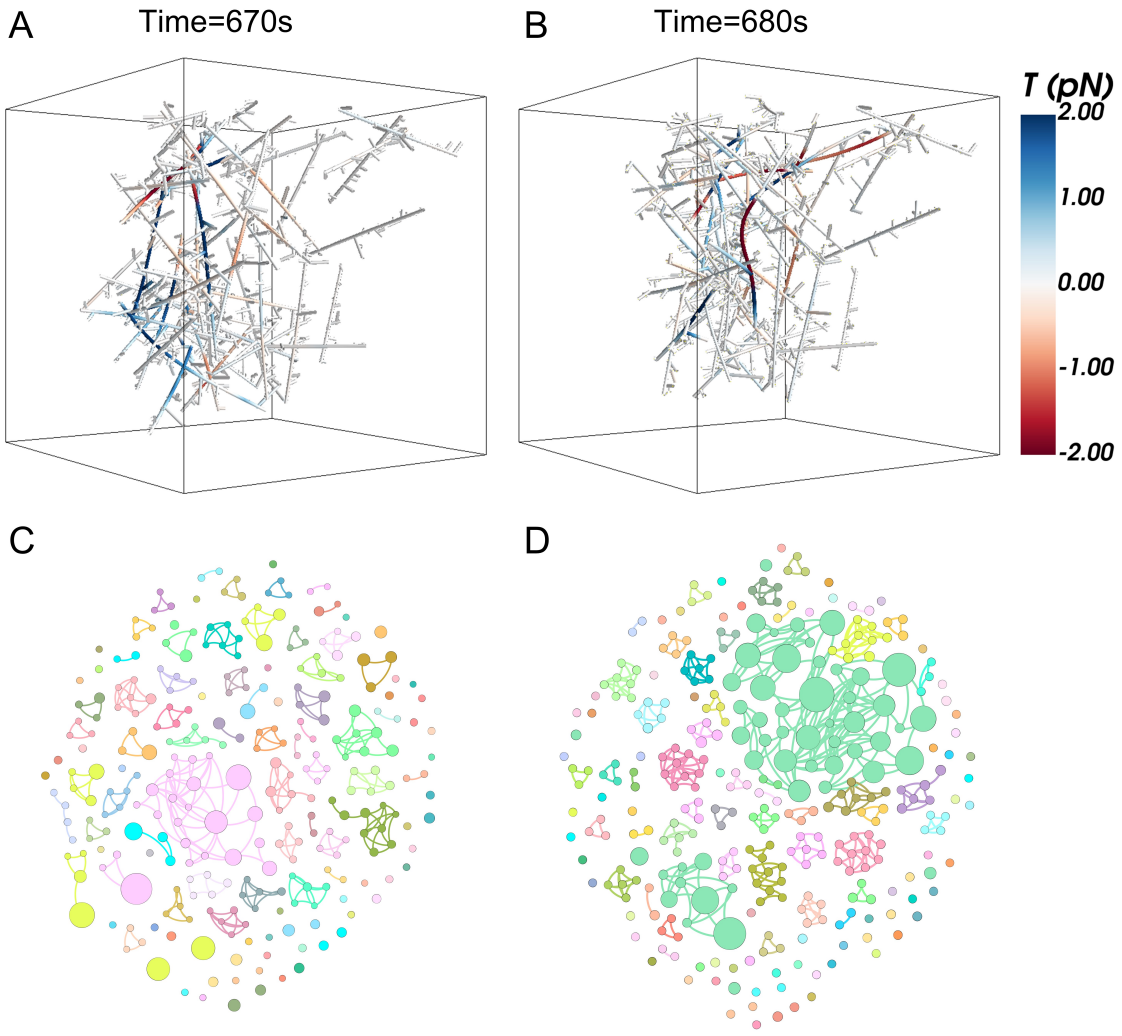


Figure 3.11. Tension snapshots and corresponding visualized graphs for avalanches at 680 s. (A) and (B) are the tension snapshots before (670 s) and during (680 s) the avalanche, and (C) and (D) are the corresponding visualized graphs of the snapshots. These two visualized graphs filtered out nodes with degrees lower than 6 and nodes with self-loops for clear output, and the layout algorithm Yifan Hu with default parameters in Gephi 0.92 [129] was used during the visualization process. The size of a node depends on its betweenness centrality, while the nodes are colored according to component ID.

3.8 MACHINE LEARNING TOOLS WERE APPLIED TO FORECAST AVALANCHES IN ACTOMYOSIN DYNAMICS

I compared the performance of machine learning (ML) models using the receiver operating characteristic (ROC) and precision–recall (PR) curves in Figure 3.12. Both ROC and PR curves provide a diagnostic tool for binary classification models for measuring the ability of a machine learning model to make correct predictions. The area under the curve (AUC) of the two curves provides quantitative scores that summarize the curves and can be used to compare classifiers. An AUC closer to 1 indicates a more skillful model. [131-133]

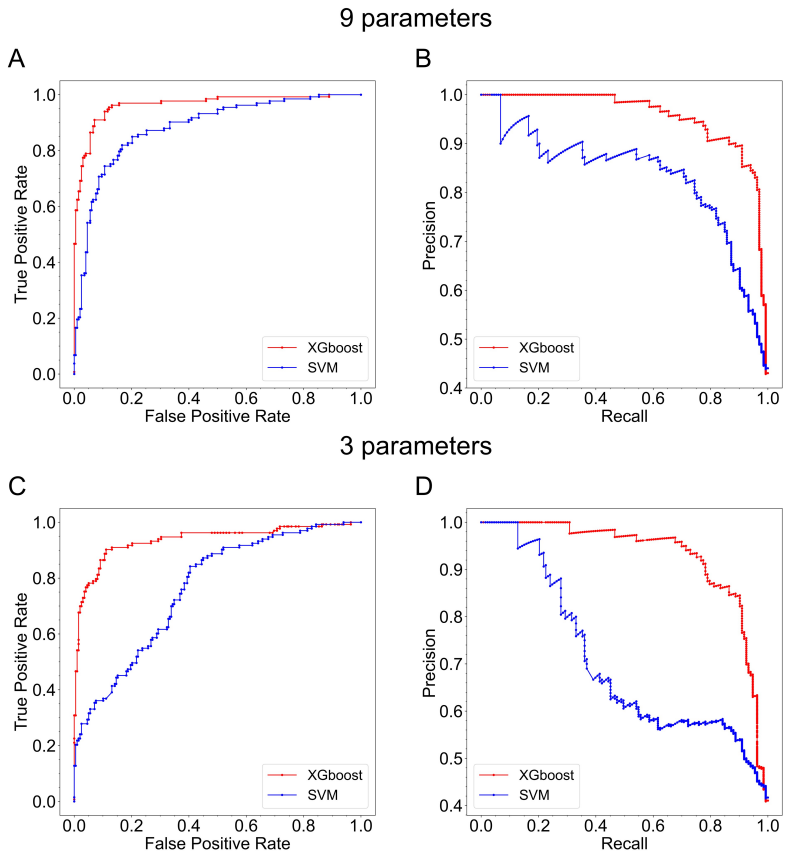


Figure 3.12. ROC and PR curves for XGBoost and SVM models. (A) and (B) show the performance of machine learning models trained with all nine parameters: three polymer physics order parameters and six network theory order parameters. (C) and (D) show the performance of the models trained with only three polymer physics order parameters. (A) shows ROC curves for XGBoost and SVM models trained with all nine parameters, and (B) shows the PR curves for these two models; (C) and (D) show the ROC and PR curves of the same types of models trained with only three parameters.

Because I achieved a comprehensive understanding of the structural characteristics of avalanches, I will employ the order parameters as features to apply machine learning tools to forecast avalanches. While I compared two prominent machine learning (ML) models in predicting emergent avalanches in actomyosin networks, I also explored the importance of the features employed for training data sets.

For each machine learning model (XGBoost or SVM), I employed two sets of order parameters for training the data set. The first set includes all nine order parameters: three polymer physics order parameters and six network theory order parameters. The second set includes only three polymer physics order parameters. As shown in Figures 3.12A and 3.12B, all nine parameters were used in training the model. For the SVM model (blue), the AUCs for the ROC curve and PR curve are 0.88 and 0.83, respectively. For the XGBoost model (red), the AUCs for the ROC curve and PR curve are 0.96 and 0.96, respectively. The XGBoost model performs better than the SVM model, which is expected since the XGBoost model is shown to be sufficient in most applied cases, while the performance of the SVM model relies strongly on the selection of a good kernel. [126, 127] Notably, the AUC values for the XGBoost model from the investigation are high enough (close to 1, the perfect value) to prove the excellent performance of this model in this case.

Next, I diagnosed the outcome with only three polymer physics order parameters used in training the models (Figures 3.12C and 3.12D). For the SVM model (blue), the AUC for the ROC curve and that for the PR curve decrease significantly to 0.76 and 0.70, respectively, in comparison with those trained with nine order parameters. For the XGBoost model (red), the AUCs for the ROC curve and the AUCs for the PR curve were 0.94 and 0.93, respectively. They

remain at the same level as those trained with nine order parameters, indicating that adding more features from network theory order parameters improves the performance of the SVM model more than the performance of the XGBoost model.

3.9 NETWORK THEORY ORDER PARAMETERS STRENGTHEN MACHINE LEARNING MODELS TO FORECAST AVALANCHES BETTER IN ACTOMYOSIN DYNAMICS

I further diagnosed the performance of the two ML models by employing the confusion matrices that label the true and predicted cases in Figure 3.13. The definition of a confusion matrix is explained in the Data Analytics section. For the SVM model, the addition of the six network theory order parameters into the ML training data set (Figure 3.13A and Figure 3.13C) moved 38 counts from the category of false negatives to the category of true positives and one count of true negatives to one count of false positives. Meanwhile, for the XGBoost model, the addition of the six network theory order parameters into the ML training data set (Figure 3.13B and Figure 3.13D) brings only eight counts from the category of false negatives to that of true positives and three counts from the category of false positives to the category of true negatives. The addition of network theory order parameters into the ML model data sets enhances the forecast of avalanches, especially by reducing the number of false negative predictions, indicating more predicted hidden avalanches.

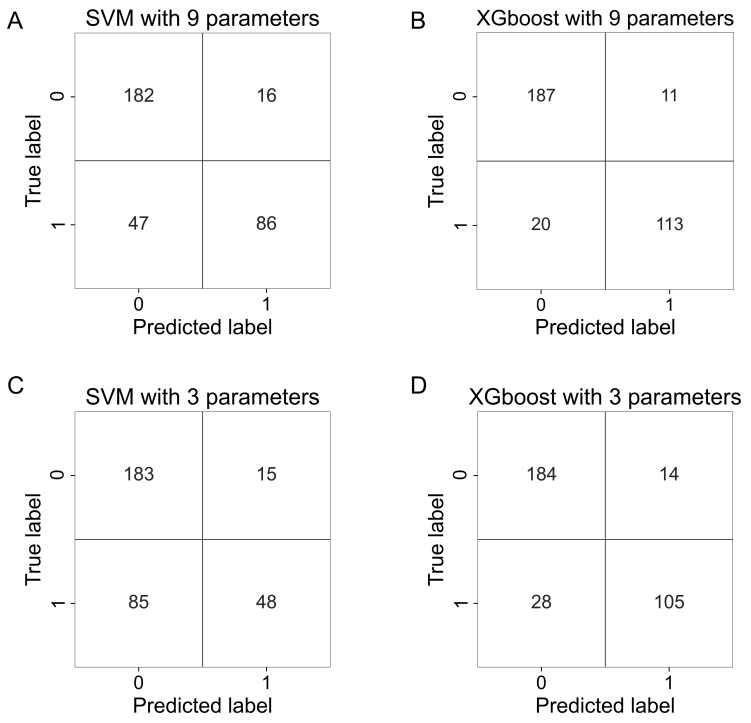


Figure 3.13. Confusion matrices for XGBoost and SVM models. (A) and (B) show the confusion matrices of the XGBoost and SVM models trained with all nine parameters: three polymer physics order parameters and six network theory order parameters. (C) and (D) show the confusion matrices of the XGBoost and SVM models trained with only three polymer physics order parameters. In the confusion matrices, label 1 represents a snapshot that is followed by avalanche while label 0 represents a snapshot that is not followed by avalanche.

The XGBoost model exhibits better performance than the SVM model with both three and nine parameter data sets, indicating that the XGBoost model is potentially a better classifier in this case for the prediction of avalanches. Additionally, the XGBoost model shows less sensitivity to the network theory order parameters than the SVM model. These facts motivate me to further investigate how these nine features work during the training of this model. Therefore, I further evaluated the importance of these features in the XGBoost models in Figure 3.14.

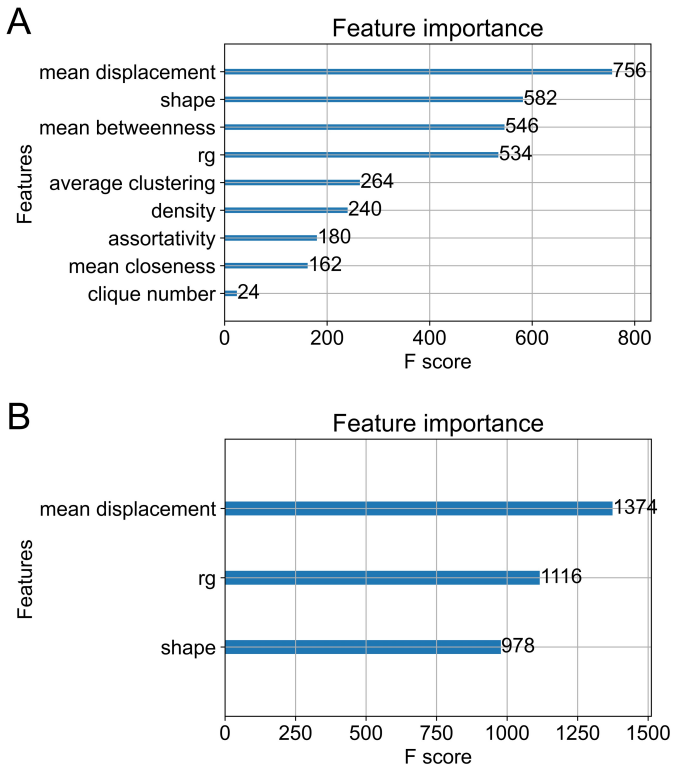


Figure 3.14. Feature importance of parameters in the XGBoost models. (A) and (B) show the feature importance of the parameters used in the two XGBoost models trained with nine and three parameters. (A) and (B) count the number of times each feature is split across all boosting rounds (trees) in the model as the F-score and visualize the result as a bar graph, with the features ordered according to how many times they appear.

As shown in Figure 3.14A, when the XGBoost model was trained with nine features, the three polymer physics order parameters mean displacement, shape and radius of gyration have a prior importance in avalanche prediction. Most of the network theory order parameters have secondary importance with the exception of the mean betweenness. As discussed in the mathematical graphs of Figures 3.7 and 3.11, the betweenness of nodes tracks the centrality distributions in the network, measuring the “shortest pathways” on actin filaments in a physical network. As a mechanical emergent phenomenon in actomyosin networks, avalanches are closely

related to the formation and subsequent delocalization of clustered centers in the physical network. Therefore, the mean betweenness plays a more important role than other network theory order parameters in forecasting avalanches.

Figure 3.14B shows that in the XGBoost model trained with only three polymer physics order parameters the importance of these three parameters is relatively similar, which is consistent with Figure 3.14A. In contrast, when I used only the six network theory order parameters to train the data set, the forecast of an avalanche was poor (see Figure 3.15). The ensemble of decision trees [113] in XGBoost is better at detecting hidden patterns from a complex network than linear regression in the SVM model. Although the introduction of network order parameters in training the data set improves the performance of both models, the extent of performance is more significant for the SVM model than for the XGBoost model.

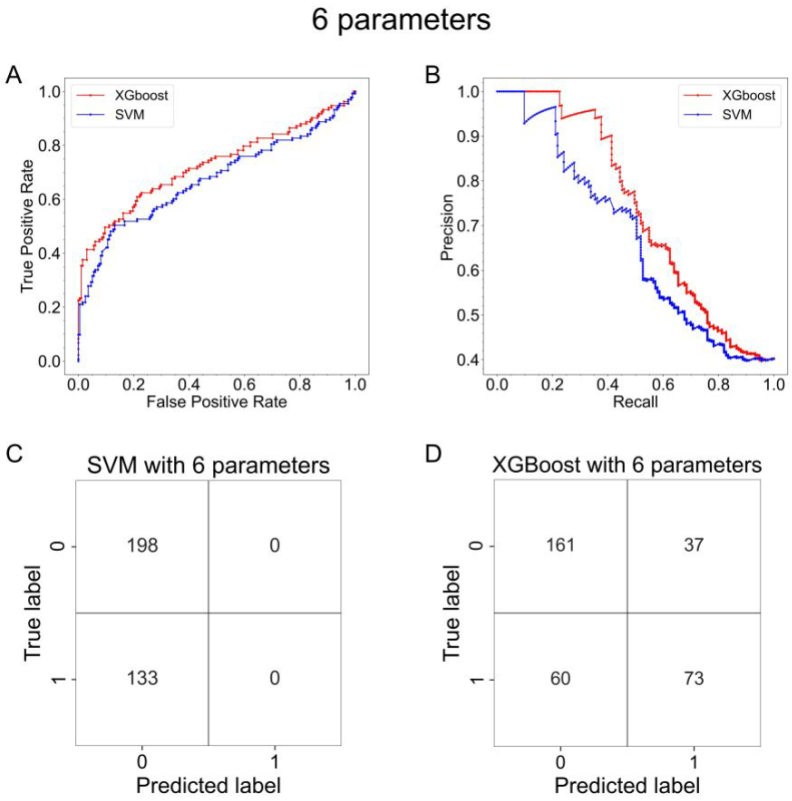


Figure 3.15. ROC, PR curves and confusion matrices for SVM and XGBoost models trained with six network theory order parameters. A shows ROC curves for the SVM and the XGBoost models trained with six network theory order parameters, B shows the PR curves for these two models. C and D show the confusion matrices for these two models, respectively. The six network theory order parameters are density, average clustering, mean closeness, mean betweenness, clique number and assortativity.

3.10 DISCUSSION

Arp2/3 Complex Concentration Tunes the Emergence of Avalanches in Actomyosin Networks

The actin-related proteins 2/3 (Arp2/3) complex, also known as the brancher in the system, initiates a filament branch (daughter filaments) at an angle of 70° on the sides of the preexisting mother, subsequently altering the topology of the network. [91, 92] The binding of the Arp2/3 complex to a filament is an ATP-dependent process [134] to prevent passive unbinding. This piece of experimental evidence shows that it is exceedingly rare for the Arp2/3 complex to unbind itself from actin filaments without an ATP-consuming reaction involving another enzyme. It motivates the reasoning of the parametrization of the Arp2/3 complex in which the event of Arp2/3 unbinding from actin filaments is quite rare in the MEYDAN simulations. [22] In prior investigation, James Liman justified the use of MEDYAN over other codes such as Cytosim [94] or AFiNES [95] because MEDYAN [11] has physically realistic models and mechanochemical feedback, which is critical to describe active processes.

Here, the study shows that a high concentration of the Arp2/3 complex limits linker binding and motor walking, which in turn reduces connectivity and inhibits contraction of the network (Figure 3.6C). However, by decreasing the Arp2/3 concentration in the system, the network not only contracts faster and more robustly (Figure 3.6A) but also has the ability to form larger clusters at the center (Figure 3.6A). At an intermediate concentration of the Arp2/3

complex and a high concentration of motor proteins, the structure of actomyosin is marginally stabilized; thus, the “avalanche” phenomenon is most likely to occur (Figure 3.6B and Figure 3.6C).

As the main part of the cytoskeleton, actomyosin networks play important roles in various cell behaviors. As an essential actin-binding protein, the distribution of Arp2/3 complexes has been experimentally proven to be related to cell motility in non-muscle cells. [135, 136] In addition, the branched actin network is especially rich at the edge of cells, such as the actin cortex, indicating the importance of this protein to the modulation of cell shape changes. [89, 137] This modulation is achieved by the treadmilling of branched networks nucleated by the Arp2/3 complex and other actin-binding proteins that sever filaments such as cofilin. [43, 138, 139] The impact of the Arp2/3 complex concentration on the simulated network structure and dynamics described in this study will advance the understanding of the role of the Arp2/3 complex in these cell behaviors.

Network Theory Reveals a New Type of Avalanche Associated with Topological Changes in a Physical Network

With the concentration of branchers (Arp2/3 complexes) dictating the nanostructure of the actomyosin network, which in turn alters the entire network topology, a proper tool is needed to describe the hierarchical properties of the system. Therefore, I chose network theory to analyze the simulated networks in this study. Mathematical graphs and the tools of data science prove to be superior in detecting hidden patterns within a complex network. [39] Order parameters such as clustering coefficient, betweenness, and closeness measure the microscopic network properties down to a single actin subunit and characterize its role in nonlocal features, while order parameters such as assortativity and density measure the macroscopic properties of

the whole network. Therefore, the network theory order parameters provide the needed mesoscopic descriptors that connect the microscopic properties to macroscopic phenomena in an active system far from equilibrium.

In particular, I discovered that betweenness is most useful for visualizing the connection between the microscopic and the macroscopic network properties when the Arp2/3 concentrations vary (note: Arp2/3 initiates branching). For example, the graphs in Figures 3.7B, D,F have distinguished betweenness distributions. In the mathematical graph of Figure 3.7B, only one community of nodes with the highest betweenness in large circles are connected to each other, while other communities of nodes with a much lower betweenness in small circles are gathered. It underpins a centralized cluster of actomyosin filaments at low Arp2/3 concentration in the snapshot shown in Figure 3.7A. In the mathematical graph of Figure 3.7D, several communities have higher betweenness (in larger circles) than other communities (in smaller circles). It indicates the presence of multiple centers in the actomyosin network in Figure 3.7C. In the mathematical graph of Figure 3.7F, the betweenness of nodes is small among most communities. It indicates delocalized communities in the actomyosin network (Figure 3.7E). The causal relationship between the size of communities and the value of betweenness supports the hypothesis that a higher brancher (i.e., Arp2/3) concentration leads to a global network with delocalized communities, involving significant changes in the rearrangement of network topology.

This new type of avalanche is related to the change of the network topology in a branched actomyosin network. I discovered these subtle changes in the network topology by observing the betweenness from a mathematical graph at the onset of an avalanche at 680 s (Figure 3.11). There is an increase in the number of nodes with high betweenness during the avalanche (Figure

3.11D) compared to that before the avalanche (Figure 3.11C). The changes in the betweenness indicate the altered connectivity from an aster-like to a centered-like node (Figure 3.9E). The sudden creation of distinctive communities in a network initiates an avalanche. I further revealed the hidden hierarchy of the network with assortativity and captured a new type of "avalanche" involving the reorganization of a network from a "delocalized" community to a "centralized" one (Figure 3.11). In cells, actomyosin networks may have similar size and shape but distinguished intra-network topology. The emergence of new higher order risen above layers of actomyosin filaments probably leads to distinct functions. Therefore, it is important to utilize assortativity or other order parameters in the network theory to reveal the hidden topological features among these networks.

Forecast Avalanches in Actomyosin Dynamics with Network Science and Machine Learning

In ML and supervised ML in particular, data curation and feature extraction are crucial steps for building reliable prediction models. For forecasting avalanches, Figure 3.12 shows how adding network theory parameters to SVM models increases their specificity and sensitivity, whereas XGBoost models do not have this strong impact. As a rule of thumb, XGBoost is a suitable option, especially for small data sets such as ours as compared to other machines learning models. [126-128] However, the SVM model is a naiver approach since it is merely a linear regression model that relies heavily on selecting the right kernels and lacks the ability to boost the model multiple times with the same data set.

A key feature is the betweenness centrality that captures the formation and disappearance of a cluster from a complex network. I showed that adding the betweenness centrality into the training set with the polymer physics order parameters greatly increases the performance of SVM, while other network theory order parameters are of secondary importance in Figure 3.14.

I believe that this approach can also be used in predicting an avalanche in experiments where the positions of actin filaments are easier to track than actin-binding proteins. The discovery of hidden patterns can be achieved by converting a physical network into a mathematical graph, and the forecasting of avalanches can be predicted by ML.

3.11 CONCLUSION AND FUTURE OUTLOOK

To my knowledge, I was the first to systematically detect the impact of Arp2/3 complex concentration on the structures and dynamics of actomyosin networks by using mathematical graphs and data science. These tools are shown to be useful for revealing hidden patterns in complex networks, allowing me to leverage this knowledge as crucial features to train machine learning models to forecast avalanches within actomyosin networks.

To forecast the avalanches, two types of machine learning models, the SVM and the XGBoost models, were trained under various conditions. I showed that the XGBoost model performs better at forecasting avalanches than the SVM model. However, the performance of the SVM model significantly increases when the network theory order parameters are trained in the data set. Although the XGBoost model was sufficient compared to the SVM model in predicting avalanches in this work, in some other cases where an outstanding kernel for the SVM model was utilized, the performance of the SVM model supersedes the performance of the XGBoost model. [126-128] Therefore, ML models are not entirely a black box; when trained with physically meaningful features, they provide meaningful predictions with high probability.

Despite the difference in ML models, I have used only the features from the mechanical or topological properties of a network in forecasting avalanches with high sensitivity and

specificity without any knowledge of their chemical dynamics. It is indicative that the avalanche is a mechanically dominant, common phenomenon in the simulated actomyosin systems. Although this finding is consistent with that of another work about avalanches risen from unbranched actomyosin networks, [140] this independent work embraces the emergence of structural hierarchy in a network from sudden topological changes in the nanoarchitectures of branched actomyosin filaments.

CHAPTER 4: GRAPH IDENTIFICATION OF PROTEINS IN TOMOGRAMS (GRIP-TOMO) 2.0: ACCELERATING PROTEIN CLASSIFICATION FOR CRYO-ELECTRON TOMOGRAPHY WITH INTELLIGENT SEARCH

This chapter is based on Chengxuan Li's first author publication which is in preparation: Chengxuan Li, August George, Trevor Moser, Doo Nam Kim, Reece Neff, Malio Nelson, Arsam Firoozfar, Kate Baldwin, James E Evans, and Margaret S Cheung. "Graph Identification of Proteins in Tomograms (GRIP-Tomo) 2.0: Accelerating Protein Classification for Cryo-Electron Tomography with Intelligent Search"

Codes related to this work are not public yet when this dissertation is written.

Cryo-electron tomography (cryo-ET) enables structural characterization of biomolecules under near-native conditions, but low signal-to-noise ratios and structural heterogeneity pose challenges for automated particle classification in sub-tomogram averaging (STA). Existing methods rely heavily on manual annotation and structural templates, limiting efficiency and robustness. I, collaborating with PNNL team members, developed GRIP-Tomo 2.0, a machine-learning pipeline that extracts interpretable topological features of protein structures within noisy experimental backgrounds. The pipeline includes three innovations: synthetic tomogram generation simulating realistic noise, graph-based persistent feature extraction as protein fingerprints, and high-performance computing for acceleration. GRIP-Tomo 2.0 achieves over 90% accuracy in classifying large and small proteins (<160K Dalton) and 80% accuracy in cross-domain classification of proteins and noise, even in challenging real-to-synthetic datasets. This method surpasses current STA practices by accurately classifying smaller protein particles

without requiring refinement, paving the way for automation with limited training data in genome space. GRIP-Tomo 2.0 advances cryo-ET workflows by enhancing automated visual proteomics, improving efficiency and scalability in structural analysis.

4.1 INTRODUCTION

A living cell contains dynamic, spatially complex protein assemblies that drive cellular processes and are highly sensitive to its functional state. [141, 142] Characterization of protein assemblies while they are in action is essential for relating their native structural properties to the phenotype of a cellular state. [24, 143, 144] However, achieving sufficiently high resolution to characterize the three-dimensional (3D) structures of these protein assemblies in situ is challenging. Most current proteomics have little or no spatial resolution, while the methods to determine atomistic structures such as X-ray crystallography [145] and nuclear magnetic resonance (NMR) spectroscopy [146] require biochemically isolated samples. Since “resolution-revolution” [147] in the recent advances of cryogenic electron microscope (cryo-EM), cryogenic electron tomography (cryo-ET) promises 3D visualization of complex cellular architecture of biological samples in situ. [23, 143]

Nevertheless, cryo-ET has its own limitation in poor data quality. To generate high-resolution structures, it requires a complex workflow that prohibits scaling. [148, 149] In cryo-ET, besides proteins, both electron dosage and sample thickness contribute to the signals. It is challenging to balance the trade-off between resolution and sample preservation. [23] The bulk of the signals distributes over a series of tilted 2D images, in which individual images contain only a fraction of the total structural information. [148] An algorithm of back projection is used to construct a full 3D volume called a tomogram. In this process, data loss, distortion, and

defects attributes to low signal to noise ratio in reconstructed tomograms. Cryo-ET thus requires an extensive workflow of refinement involving particle classification to resolve individual molecular structures. [23]

Sub-tomogram averaging (STA) [150-153] is a common technique of refinement in the cryo-ET workflow that uses frequently occurring identifiable particles with rigid shapes to reconstruct atomic-level resolution maps. Refining structures with STA is a resource intensive process, as most workflows require sufficient number of copies in each structural classification for a protein target that exists in multiple conformations. [154-159] Numerous computational programs such as IMOD, [160] EMAN2, [161] RELION, [148] emClarity, [162] PyTom, [152, 163] and Warp [149] have been developed for cryo-ET data processing by offering a user streamlined workflows of automated particle picking from tomograms, classification of various conformations, and fast STA, as shown in Figure 4.1. However, these tools still require expert knowledge and manual annotation by visually inspecting the particles in noisy tomograms. Even with deep learning models to improve individual steps in cryo-ET workflow, they often require large training datasets for verification and lack generalizability and interpretability. [34, 35, 38, 159, 164-170] The challenge in scaling up the cryo-ET for automation still exists when the biological sample is complex, particularly when it involves small proteins without distinguishable features from background noises for classification.

August George, along with team members at Pacific Northwest National Lab (PNNL), developed Graph Identification of Proteins in Tomograms (GRIP-Tomo) [26] to resolve the bottleneck of protein classification in the STA. This algorithm was built based on the notion that protein structures are evolutionary conserved despite of millions of homologues in the database. GRIP-Tomo transformed evolutionarily conserved protein structures into mathematical graphs

and distinguished them from each other according to topologically invariant features on a graph. The previous version of GRIP-Tomo demonstrated the success in a proof of principle for protein identification among synthesized single particle sub-tomograms without realistic background noise.

In this follow up work, I worked with collaborators at PNNL and verified GRIP-Tomo 2.0 by including realistic background noise in the synthesized sub-tomograms for learning signals from noises. It includes the module of synthesizing ‘mock’ sub-tomograms with realistic noises and tunable imaging artifacts such as dosage and thickness. GRIP-Tomo 2.0 deploys topological data analysis (TDA), [16, 46, 171-173] leveraging persistent topological features in spatial scales as structural ‘fingerprints’ to comprehensively describe the shapes of macromolecules. I worked with Reece Neff to deploy the pipeline on High Performance Computing (HPC) platforms to efficiently compute the persistent features. I included a new module for interpretable machine-learning to classify protein structures in sub-tomograms. Through a systematic examination of imaging parameters, including sample thickness and electron dosage, I demonstrate that graph-theoretic features can uncover classification boundaries, inform data calibration, and enable structure-aware interpretation. These fingerprint-like features, derived from machine learning data, provide unique insights that guide users in optimizing noise characterization for accurately distinguishing proteins within mixed samples.

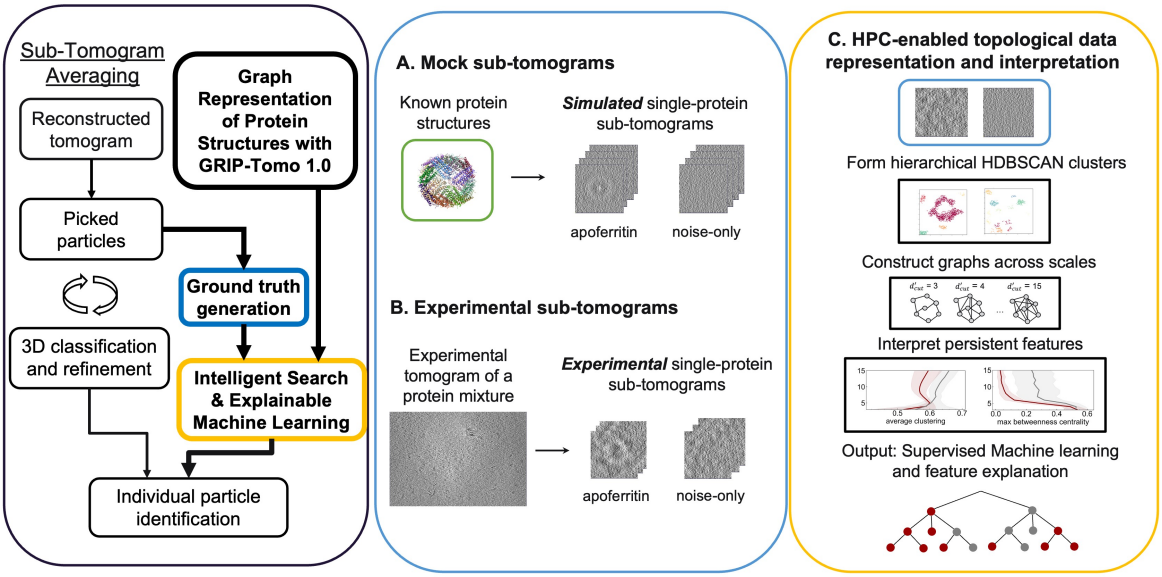


Figure 4.1. GRIP-Tomo 2.0 framework for interpretable macromolecular identification in cryo-ET. Overview of the GRIP-Tomo 2.0 pipeline, designed to improve particle identification in sub-tomogram averaging (STA) through intelligent search and explainable machine learning. The pipeline involves generating reliable ground truth using mock (A) and experimental (B) data, enabling interpretable machine learning workflows (C).

4.2 OVERVIEW OF THE GRIP-TOMO 2.0 PIPELINE FOR PROTEIN IDENTIFICATION IN CRYO-ET

Motivation (Figure 4.1, left Panel)

Traditional STA processes involve particle picking, 3D classification, and refinement steps, which GRIP-Tomo 2.0 aims to enhance by incorporating known protein structures from a

database and graph representation techniques established in GRIP-Tomo 1.0 [26]. To address the challenge of identifying macromolecular structures in noisy cryo-electron tomograms, I worked with my collabrators at PNNL to develop GRIP-Tomo 2.0, a comprehensive pipeline that integrates dual-source ground truth data generation, topological representation, and interpretable machine learning. (Figure 4.1). The pipeline is designed to extract biologically meaningful patterns from sub-tomograms and leverage them for robust protein classification under realistic imaging conditions.

Ground truth data generation (Figure 4.1, middle Panel)

As illustrated in Figure 4.1, GRIP-Tomo 2.0 begins with the ground truth generation of two complementary datasets: synthetic (mock) sub-tomograms simulated from known PDB atomic structures under tunable imaging parameters, and experimental sub-tomograms extracted from real cryo-ET data of mixed macromolecular samples (Figure 4.8). Together, the dual source framework enables to synthesize realistic mock sub-tomograms with tunable imaging parameters while anchoring the evaluation in real experimental data. The mock dataset provides an essential and tunable source for model training, while the experimental set allows to evaluate the generalization capacity and biological relevance of learned representations. The ability to calibrate mock volumes to match real tomograms lays a strong foundation for the hybrid validation strategy employed throughout GRIP-Tomo 2.0. This dual-source framework enables both controlled model training and realistic validation, establishing a foundation for systematic exploration of imaging effects and model generalizability.

Topological representation of protein structures and interpretable machine learning (Figure 4.1, right Panel)

Following data generation, sub-tomograms are converted into graph-based representations that capture evolutionarily conserved topological features of protein structures. This transformation allows GRIP-Tomo to encode structural information in a format amenable to scalable computation and statistical learning. The graph construction and feature extraction steps are implemented in a high-throughput, HPC-compatible pipeline capable of analyzing thousands of sub-tomograms in parallel. Finally, GRIP-Tomo employs explainable machine learning to classify protein structures and uncover which topological features are most predictive. This integration of interpretable modeling supports not only accurate classification but also mechanistic insights into what distinguishes proteins from background or from one another in tomograms.

4.3 MOCK SUB-TOMOGRAMS GENERATED BY TUNABLE PIPELINE ARE WELL-MATCHED TO EXPERIMENTAL SUB-TOMOGRAMS

Mock sub-tomogram

The mock data pipeline (Figure 4.2A and 4.8A) begins with atomic structures from the Protein Data Bank (PDB), which are input into cisTEM [25] to simulate 2D tilt series projections. The 2D tilt-images are then preprocessed by densmatch, Topaz 2D denoising, [169]

Contrast Transfer Function (CTF) correction and finally reconstructed into 3D sub-tomograms which goes through low-pass filter and density inversion. Detailed description of the entire pipeline is in Methods 2.1. The simulation allows precise control over critical imaging parameters, including electron dosage per tilt and sample thickness, which were the two most impactful factors on the output. The terms ‘dosage’ and ‘thickness’ will be used to describe the electron dosage per tilt and sample thickness in the following texts to avoid redundant descriptions. By adjusting dosage and thickness, I generated mock sub-tomograms that span a wide signal-to-noise ratio (SNR) spectrum—ranging from idealized conditions (Figure 4.2C, first row in golden box) to those mimicking real experimental limitations (Figure 4.2C, grey box). Notably, this mock data generation pipeline is exhibited on HPC platforms and encoded with parallel computing, allowing large scale simulations across various combinations of imaging parameters.

Experimental sub-tomogram

The experimental data pipeline (Figure 4.2B and 4.8B) starts from the tilt series images of a Protein Mixture Solution collected by the Thermo-Fischer Krios Microscope, as described in Methods 2.2, followed by processing such as Contrast Transfer Function (CTF) correction, Topaz 2D denoising and the following 3D reconstruction into tomogram. I worked with Kate Baldwin to manually pick the center coordinates of particles and extracted the sub-tomograms based on the coordinates. The experimental data pipeline serves to generate ground truth data from the real world, mainly used as reference and testing set in later steps. Both mock and experimental data generation pipeline undergo the deep learning-based Topaz denoising [169] on the 2D tilt series images.

Imaging parameter sweep

I first used the same dosage and thickness as experimental settings (2.9 $e^{-}/Å^{2}$, 250 Å) to simulate the mock sub-tomograms. However, the output sub-tomogram (Figure 4.2C, first row in golden box) retained higher contrast and structural clarity, indicating that the noise and distortion levels were shifted versus the real experimental imaging conditions (Figure 4.2C, grey box). I calibrate the imaging parameters that best reproduce the appearance of the experimental sub-tomograms (Figure 4.5) by adjusting the simulated dosage and thickness conditions. I found that mock sub-tomograms simulated at 0.3 $e^{-}/Å^{2}$ dosage and 500 Å thickness were well-matched to the experimental data in real space. This calibrated simulation condition was selected for downstream training of the machine learning models, ensuring representational alignment between synthetic and experimental domains.

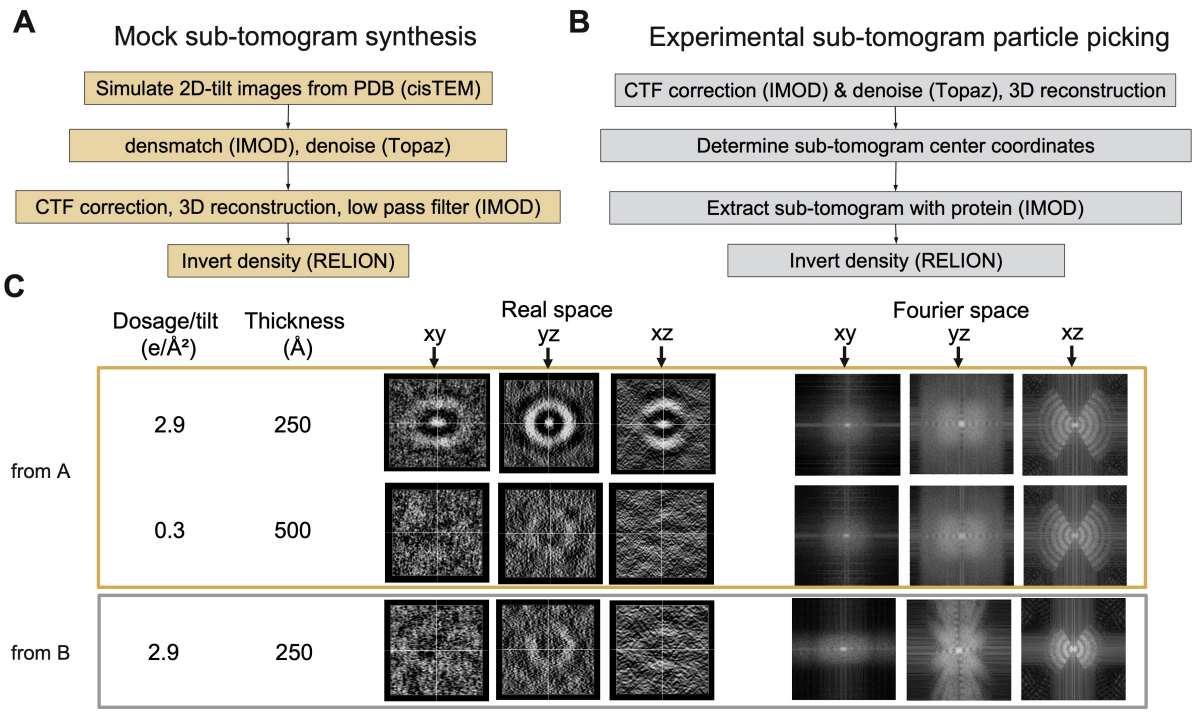


Figure 4.2. Synthetic mock and experimental sub-tomogram preparation pipelines and imaging parameter calibration. Illustration of the workflows used for generating mock and experimental sub-tomograms. (A) The mock data synthesis pipeline begins with cisTEM simulations to generate tilt-series data from PDB structures, followed by densmatch, denoising, CTF correction, and 3D reconstruction. (B) Experimental sub-tomogram generation involves CTF correction, denoising with Topaz, manual particle picking, sub-tomogram extraction, and density inversion. (C) Examples of mock and experimental sub-tomograms. The final synthesized mock sub-tomograms exhibit tunable quality by adjusting imaging parameters, such as dosage per tilt and sample thickness, in golden box. The experimental sub-tomogram is shown in grey box.

4.4 GRIP-TOMO 2.0 EXTRACTS GEOMETRY-PRESERVING GRAPH REPRESENTATIONS FROM 3D SUB-TOMOGRAMS USING SCALABLE HPC WORKFLOWS

Preprocessing and clustering of density volumes

The output of ground truth data generation from the last section are sub-tomograms as 3D density volumes stored in MRC format, a common format for storing image and volume data in fields of cryo-electron microscopy and tomography. Each input MRC volume is a 219 × 219 × 219 voxel cube, resulting in approximately 10 million voxels per sub-tomogram. Maintaining sufficient throughput and resolution at this voxel count is essential for preserving meaningful structural information but also poses significant computational challenges for downstream processing at scale. To address this, I worked with Reece Neff to develop a streamlined pipeline that reduces the data dimensionality while preserving key topological and structural signals. The input sub-tomogram with N_{vox} ~ 10 million is standardized (Figure 4.3a), thresholded (Figure 4.3b–c), density-aware coarsened (Figure 4.3d) and HDBSCAN clustered (Figure 4.3e) to get a final number of voxels in the largest cluster. For apoferritin I find that the largest cluster has a size of N_{vox}^{c}= 42029 voxels — closely matching its known number of atoms of approximately 39000 — and visually recapitulates its spherical geometry even in the presence of tomographic artifacts like the missing wedge effect (Figure 4.3e). This validates that the voxel-to-cluster transformation retains biologically relevant shape information and atom count.

Graph representation and persistent feature calculation

From these clusters, I construct undirected graphs from the remaining voxels based on proximity, applying a range of spatial cutoff distances d'_{cut} to capture both local and global

connectivity (Figure 4.3f). The increase of d'_{cut} leads to an increased number of edges while the number of nodes is conserved (Figure 4.3f). Graphs generated at lower d'_{cut} reveal local coordination, while those at higher d'_{cut} capture broader organizational features such as domains or global shape. The resulting graphs preserve macromolecular structure, including spherical organization affected by missing wedge anisotropy, confirming their geometric fidelity. To quantify these topologies, I extract a set of graph features across scales, such as maximum eigenvector centrality, degree assortativity, and number of communities (Figure 4.3g and Section 4.7.3) into a vectorized "fingerprint". These graph descriptors encode key structural properties and enable machine learning models to distinguish between molecular identities or structural states based on interpretable topological signatures.

Parallelized and scaled pipeline on High Performance Computing (HPC) platforms

To further enhance the throughput and scalability of GRIP-Tomo 2.0, I worked with Reece Neff to deploy the full graph feature extraction pipeline on the NERSC Perlmutter supercomputer using Parsl for workflow parallelization (Figure 4.3h). This deployment enabled concurrent processing across 1,024 compute nodes and 4 workers per node, achieving an estimated 90× speedup compared to serial execution. As a benchmark, the full pipeline—including voxel normalization, density thresholding, HDBSCAN clustering, graph construction, and feature extraction—generated 45,612 graph features from mock datasets in just over two days, consuming approximately 4,323 node hours. This efficient distributed framework allows GRIP-Tomo 2.0 to perform parameter sweeps and process large tomogram sets, enabling comprehensive simulation-to-experiment comparisons across diverse imaging artifact conditions.

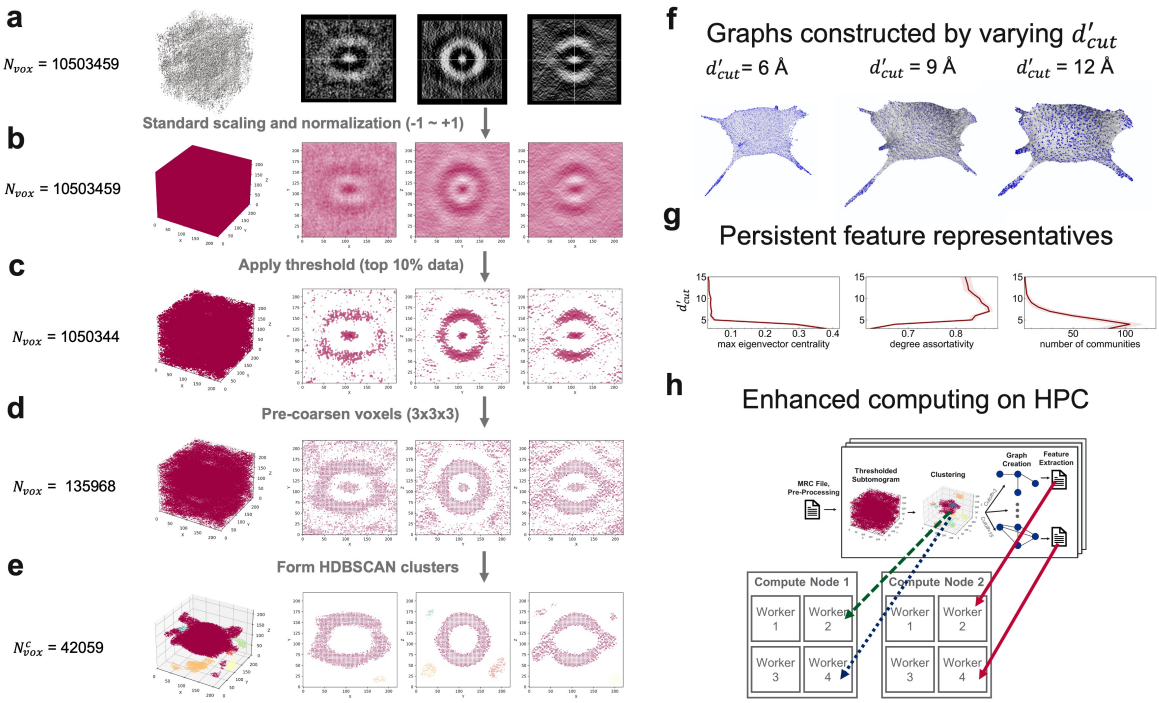


Figure 4.3. Framework for scalable transformation of cryo-ET sub-tomograms into geometry-preserving graph features using High Performance Computing. The GRIP-Tomo 2.0 pipeline transforms high-dimensional sub-tomograms into biologically meaningful, geometry-preserving graph features. N_{vox} stands for the number of voxels preserved this step and N^c_{vox} stands for the final clustered number of voxels, or centroids. (a) Raw input volumes are first standardized. (b–c) The highest-density voxels are retained through thresholding to emphasize structural signal. (d) A density-aware voxel coarsening step reduces node count while maintaining spatial integrity. (e) HDBSCAN clustering identifies stable point clusters. (f) Graphs are constructed from clustered voxels across cutoff distances d'_{cut} revealing multi-scale topological organization. (g) Persistent graph features capture both structural hierarchy and spatial layout. (h) The pipeline is parallelized on HPC infrastructure using Parsl, enabling distributed execution of all steps and generation of over 100,000 graph features to date. This framework captures 3D protein shape in a scalable and interpretable form.

4.5 TOPOLOGICAL FINGERPRINTS PREDICT IMAGING CONDITIONS FAVORABLE FOR PROTEIN CLASSIFICATION AND REVEAL RESOLUTION BOUNDARIES OF GRIP-TOMO

Diagnostic use of topological fingerprints to guide imaging parameter selection in mock datasets:

To evaluate how imaging conditions affect GRIP-Tomo's ability to extract discriminative structural, I systematically simulated mock sub-tomograms across a matrix of electron dosages and sample thicknesses. From each dataset, I extracted the combined graph-based fingerprints—a multiscale trajectory of 11 topological features across 13 graph cutoff values d'_{cut} — for four categories: apoferritin, beta-gal, aldolase, and noise. These fingerprints, visualized in Figure 4.4A, reveal how imaging conditions influence the structural distinctiveness of proteins. Importantly, some imaging conditions clearly produce fingerprints that are highly separable across all four categories, making them favorable for downstream classification. A particularly illustrative example is the condition with dosage = 2.0 $e^{-}/Å^2$ and thickness = 400 Å, where the fingerprint trajectories are visibly distinct even to the human eye. Under this condition, each category exhibits a unique feature profile across d'_{cut}, suggesting that the graph topology of these simulated volumes preserves sufficient biological and geometrical signal to support accurate classification—not only between proteins and noise but also among protein classes. In contrast, other conditions result in overlapping or ambiguous fingerprints. For instance, under dosage = 0.3 $e^{-}/Å^2$ and thickness = 500 Å—the condition later selected for simulation-to-experiment domain transfer—the feature trajectories for apoferritin, beta-galactosidase, and aldolase show considerable overlap. While this condition was chosen for its close match to experimental

fingerprints, its intrinsic limitations—namely the poor class separability among proteins—make it difficult to distinguish structurally similar macromolecules. As a result, while GRIP-Tomo performs well in binary classification (e.g., protein vs. noise in Figure 4.7D) under this condition, it struggles to resolve finer structural differences between proteins like apoferritin and beta-gal (Figure 4.9). This observation helps explain the drop in protein-class classification performance seen in the next section: even though the synthetic data were calibrated to match experimental imaging conditions, the condition itself does not support strong topological contrast between proteins. If experimental data were acquired under a more favorable condition—such as higher dosage and lower thickness, yielding fingerprints like those at dosage = 2.0 $e^{-}/Å^2$ and thickness = 400 Å—then finer distinctions among proteins would likely be more achievable. This highlights an important dual role of GRIP-Tomo fingerprints: not only do they serve as model input, but they also act as diagnostic visual tools that can help researchers pre-screen synthetic imaging conditions and anticipate the success or limitations of downstream classification. In this way, GRIP-Tomo offers a novel lens for experiment design and simulation tuning, providing actionable guidance on whether a given imaging regime is sufficient to separate relevant biological classes.

The classification boundary on mock data revealed by varied sample thickness

To further quantify how sample thickness affects classification ability, Figure 4.4B shows the per-class classification accuracy of GRIP-Tomo models trained and tested on mock data generated with fixed dosage = 2.9 $e^{-}/Å^2$ and varying sample thicknesses, as a study case. I trained using 5 samples per class and tested using 45 samples per class. In general, accuracy decreases with increasing thickness, as expected due to more severe scattering and background

noise. However, the trend is not uniform across categories, and an intriguing exception is observed in the classification of aldolase, the smallest protein in the dataset. Additionally, the accuracy was high at low thickness despite the small training set size of 5 samples, which suggests the GRIP-Tomo is a data efficient method for predictions under desirable imaging conditions. At thickness = 250 Å, aldolase classification performance is lower than that of the larger proteins. However, as thickness increases to 750 Å, aldolase accuracy improves and surpasses its performance at thinner conditions. This observation is counterintuitive under traditional cryo-ET expectations, where smaller proteins typically become more difficult to classify as sample thickness increases. One possible explanation lies in the nature of GRIP-Tomo's feature representation. Unlike conventional methods that rely purely on raw voxel intensity or template matching, GRIP-Tomo constructs graph-based representations that may capture topological structure more robustly—even in noisier or more diffuse density environments. The density of edge connectivity, modularity, and clustering behavior within the graph may still provide discriminative signal for small proteins, even when the voxel-based contrast is weak. Moreover, the relatively high electron dosage of 2.9 $e^{-}/Å^{2}$ used across these simulations may contribute to improved signal for small structures like aldolase. In typical cryo-ET experiments, lower dosages are often used to minimize radiation damage, which can suppress small-protein signal early. In contrast, the higher dose used here might enhance visibility in a way that complements the graph abstraction layer, allowing topological features to remain stable across a wider range of thicknesses than traditional voxel-based approaches.

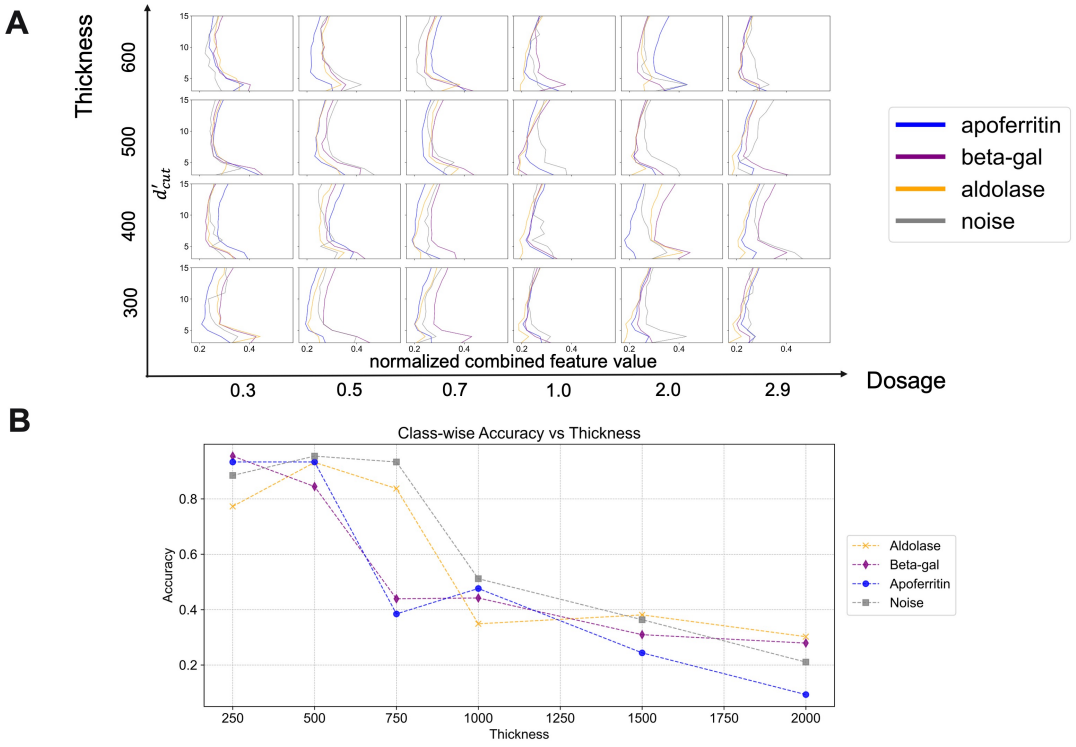


Figure 4.4. GRIP-Tomo fingerprints identify favorable imaging conditions and track classification performance across sample thicknesses. (A) GRIP-Tomo feature fingerprints derived from mock sub-tomograms simulated across a grid of electron dosage (x-axis, 0.3 to 2.9 $e^{-}/Å^{2}$) and sample thickness (y-axis, 300 to 600 Å). Each subplot shows the normalized 13-point feature vector across different d'_{cut}, combining 11 topological features described in Methods 3.1.6, and color-coded by structural category: apoferritin (blue), beta-galactosidase (purple), aldolase (orange), and noise (gray). These vectors serve as 'persistent topological fingerprints' that encapsulate the graph-based signature of each macromolecular structure. (B) Per-class classification accuracy of GRIP-Tomo as a function of increasing sample thickness. Each thickness has its own model trained and tested on the data specifically simulated in this thickness. 10% of mock data are used for training while 90% of mock data are used for testing. Mock sub-tomograms were synthesized with dosage = 2.9 $e^{-}/Å^{2}$.

4.6 CALIBRATION OF IMAGING CONDITIONS REDUCES THE SIMULATION-TO-REAL GAP AND ENABLES INTERPRETABLE CROSS-DOMAIN CLASSIFICATION

Calibration of mock sub-tomograms to experimental sub-tomograms

To bridge the gap between synthetic and experimental sub-tomograms, I investigated whether tuning mock imaging parameters could align topological fingerprints and improve classification performance. Initial mock sub-tomograms were synthesized with imaging parameters in cisTEM (dosage = 2.9 $e^{-}/Å^{2}$, thickness = 250 Å) matching estimated experimental settings. However, when visualizing their topological fingerprints (Figure 4.7A, dashed box), I observed consistent divergence from experimental fingerprints (Figure 4.7B) across several key features, including average clustering, eigenvector centrality, and degree assortativity shown in the figure. These discrepancies reflect a simulation-to-real gap: although imaging parameters appear nominally matched, the effective signal structures differ. To reduce this domain gap, I performed a calibration step, systematically varying mock imaging parameters (see Figure 4.4, Figure 4.5 and Figure 4.6) and identifying the setting—dosage = 0.3 $e^{-}/Å^{2}$, thickness = 500 Å—that minimized the fingerprint discrepancy. The topological fingerprints of the experimental sub-tomograms and the composite similarity scores between the mock and sub-tomograms are shown in Figure 4.6 to support the calibration, along with visual confirmation of the real space images. Using this calibrated condition, I regenerated mock sub-tomograms for three categories (apoferritin, beta-galactosidase, and noise-only), each with 60 rotated instances, yielding a balanced training set of 180 samples. The experimental test set comprised 180 sub-tomograms from the Protein Mixture Dataset (Methods 2.2).

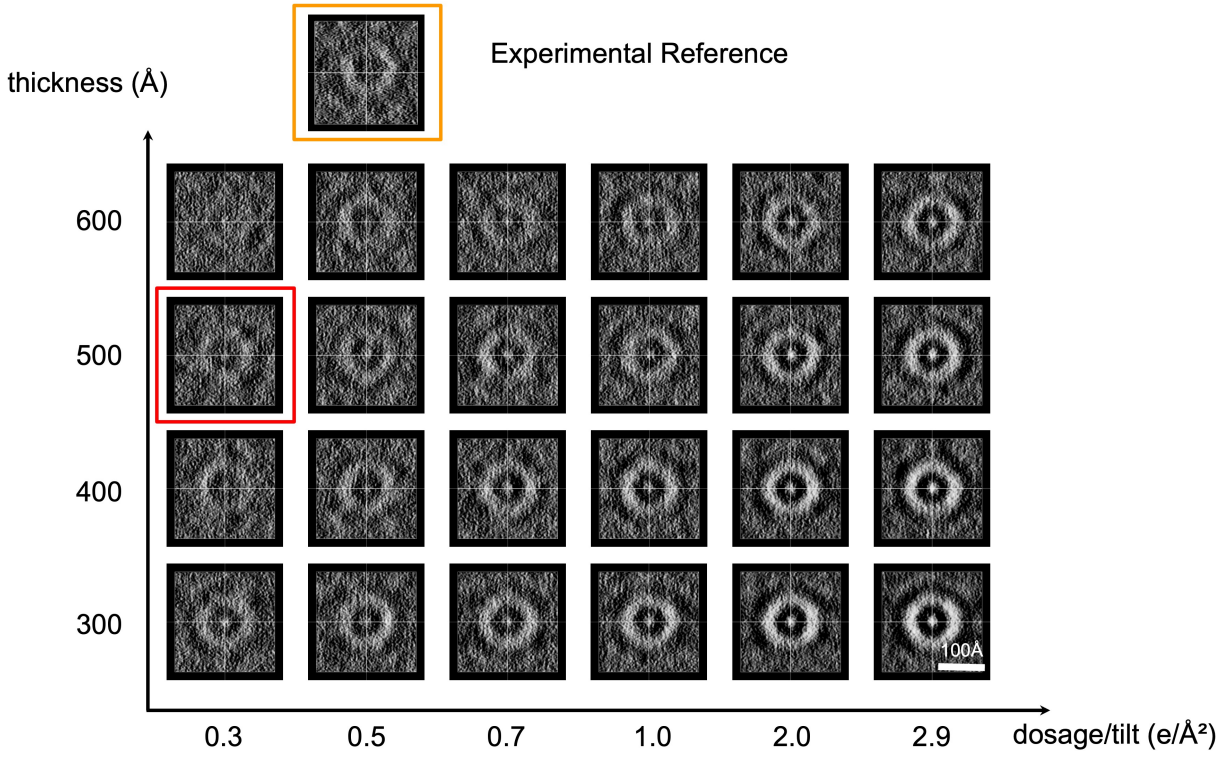


Figure 4.5. Visualization of mock sub-tomograms across simulated imaging conditions using orthoplanes representation. Each panel displays the central z-axis slice of a mock sub-tomogram of apoferritin, rendered using the map → orthoplanes function in UCSF ChimeraX. Sub-tomograms were simulated using varied imaging conditions defined by electron dosage per tilt (horizontal axis, 0.3 to 2.9 e^{-}/Å²) and sample thickness (vertical axis, 300 to 600 Å). The orange-outlined panel (top row) shows the experimental reference slice from real Protein Mixture tomogram, while the red-outlined panel highlights the synthetic condition (dosage = 0.3 e^{-}/Å², thickness = 500 Å) selected for optimal matching to the experimental data. A 100 Å scale bar is shown in the bottom right panel for reference.

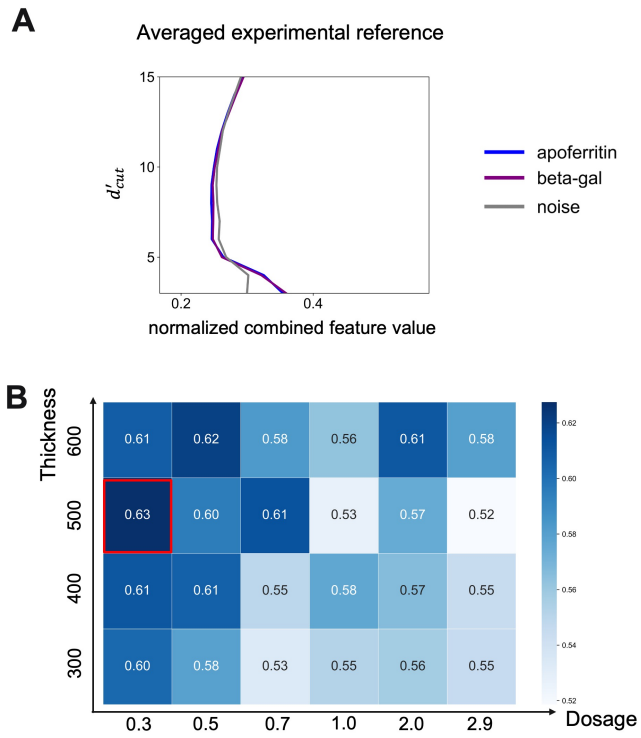


Figure 4.6. Composite similarity scoring between mock and experimental data based on GRIP-Tomo fingerprints. (A) Normalized topological fingerprint averaged from 10 experimental sub-tomograms, plotted across 13 graph cutoff values d'_{cut}. The fingerprint combines 11 graph-based features described in Methods 3.1.6, aggregated across protein categories. (B) Composite similarity scores between each mock condition in Figure 4.4A and the experimental reference fingerprint shown in (A). Each cell represents the similarity between a given mock dataset (simulated under a specific dosage and thickness) and the experimental fingerprint. Scores are computed by comparing the concatenated GRIP-Tomo feature vectors across categories and d'_{cut}. The highest similarity (score = 0.63) occurs at dosage = 0.3 $e^{-}/Å^2$ and thickness = 500 Å, outlined in red, indicating the optimal topological match between mock and experimental samples.

Improved cross-domain classification after calibration

I then extracted graph-based persistent features as topological fingerprints from both the training and testing datasets using the GRIP-Tomo pipeline and trained a Random Forest classifier. Comparing classification results before (Figure 4.7C) and after (Figure 4.7D) calibration highlights a clear performance gain: uncalibrated training yields poor generalization

to experimental data (especially for protein class) with an accuracy of 0.39, whereas calibration recovers strong classification capability, achieving 97% recall for proteins and an overall accuracy of 0.81 (Figure 4.7C and 4.7D).

Interpretability of persistent features and relation to biological scales: To interpret this performance, I examined the feature importances from the trained classifier (Figure 4.7E). Interestingly, the top-ranked features cluster into three scale bands:

- Local range (cutoff 3 Å): capturing local density and node connectivity, potentially related to tightly packed atomic centers or helix cores.
- Intermediate (cutoff 6–8 Å): these cutoffs consistently appeared in earlier GRIP-Tomo work and may correspond to typical intra-domain distances, such as those spanning α-helices, loops, or β-sheet spacing.
- Global range (cutoff 13–14 Å): features like degree assortativity and max centrality at these scales likely reflect global shape, modularity, or domain interfaces. These may correspond to the diameters of the four-fold symmetry channel in apoferritin and central cavity in beta-gal.

This multi-scale distribution suggests that GRIP-Tomo captures biologically meaningful structure-function relationships through its graph-based representation. This scale-based distribution supports the hypothesis that different graph cutoffs capture different levels of structural hierarchy. For instance, the importance of average clustering at cutoff 6 Å may reflect the compactness of secondary structure motifs, while degree assortativity at cutoff 13 Å may relate to global organization patterns such as ion channel diameter or central symmetric domains. Moreover, the calibration strategy demonstrates a practical path to aligning synthetic and real data—a crucial step for enabling domain transfer in machine learning models trained on

simulated biological systems. Together, these results validate both the interpretable nature of GRIP-Tomo features and the effectiveness of imaging condition calibration in closing the simulation-to-experiment gap for downstream classification.

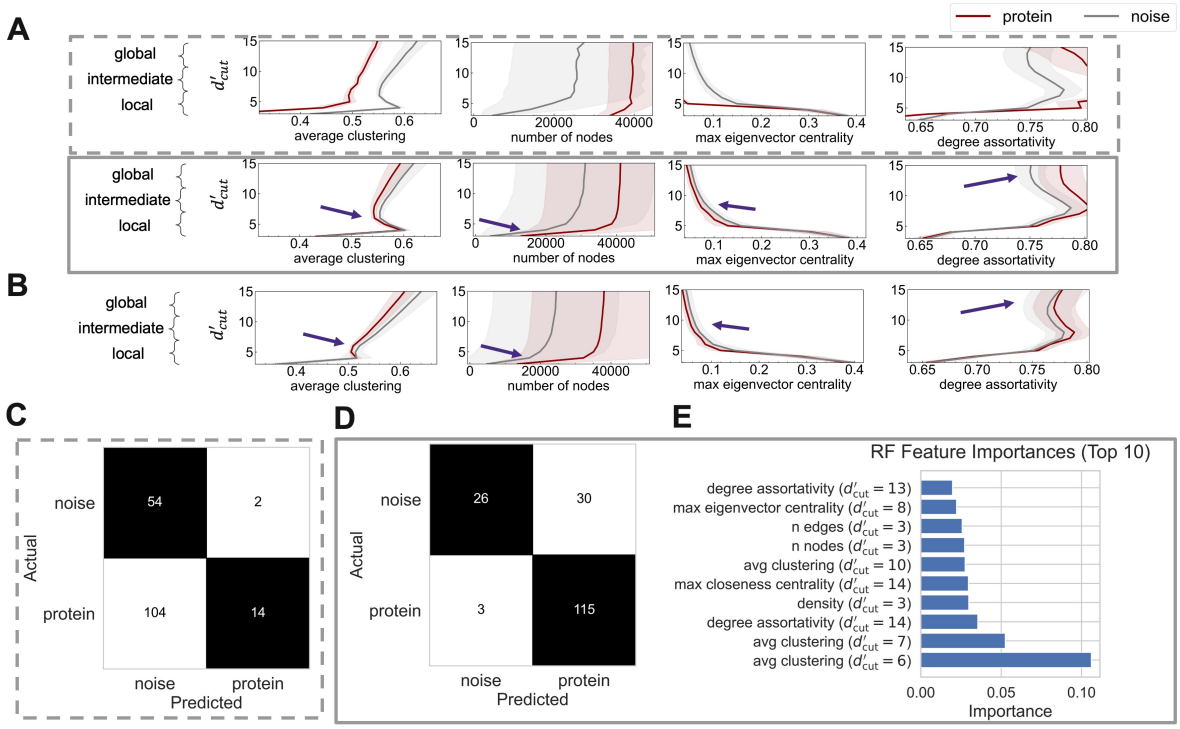


Figure 4.7. Calibrating mock imaging conditions aligns persistent topological fingerprints with experimental data and improves cross-domain classification performance. (A) Persistent topological fingerprints from mock sub-tomograms synthesized under uncalibrated (dashed box, dosage = 2.9 e^{-}/Å², thickness = 250 Å) and calibrated imaging conditions (solid box, dosage = 0.3 e^{-}/Å², thickness = 500 Å), showing persistent feature trends across cutoffs d'_{cut} for representative features: average clustering, number of nodes, max eigenvector centrality, and degree assortativity. Gray and red lines represent noise and protein categories, respectively; shaded regions denote sample variance. (B) Persistent topological fingerprints of experimental data. (C) Confusion matrix of classification before calibration (trained on mock data in dashed box in A and tested on experimental data in B). (D) Confusion matrix of classification after calibration (trained on mock data in solid box in A and tested on experimental data in B) (E) Top 10 feature importances from the Random Forest model trained on calibrated mock data and tested on experimental data.

4.7 DETAILS OF GROUND TRUTH DATA GENERATION, GRAPH FEATURE EXTRACTION AND MACHINE LEARNING CLASSIFICATION

4.7.1 Graph based analysis of density volumes

In the previous work, August George developed a method called GRaph Identification of Proteins in Tomograms (GRIP-Tomo) to identify proteins in pristine synthetic volume densities [26] in one pass, illustrated in Figure 6 of [26]. Briefly, in this method, an input sub-volume is normalized and thresholded. The remaining high-density voxels above the threshold are clustered using DBSCAN. [174] The cluster centroids as assigned as nodes, and edges are added between the nodes if their Euclidean distance is below a cutoff value, d_{cut} and d'_{cut} for PDB structure and density volumes. From this graph there is a vector of 12 topological graph features calculated: number of nodes, number of edges, number of communities, density, average clustering of each node, degree assortativity, diameter, average paths length, clique number, max betweenness centrality, max closeness centrality, max eigenvector centrality. The average relative similarity can be computed from two graph feature vectors. August George used a single short cutoff of 8 and 9 angstrom to distinguish single-domain and multiple-domain proteins.

In GRIP-Tomo 2.0, I build on the previous work of GRIP-Tomo 1.0, creating a mathematical graph representation G = (V, E) using the HDBSCAN algorithm [175, 176] instead of DBSCAN. The set of nodes V correspond to the cluster centroids from HDBSCAN, and the set of edges E accounts for the relations between nodes. An edge exists between any two nodes if they are within a cutoff distance d'_{cut}. The distance between two nodes is the minimum Euclidean

distance. Thus, the mathematical graph representation is defined as an adjacency matrix, where d'_{cut} is an important hyperparameter regulating the connectivity of the graph.

4.7.2 Ground Truth Data Generation

Synthesis of single-particle mock sub-tomograms from known protein structures:

For the synthesis of a mock sub-tomogram with a protein and artifacts (Figure 4.2A and 4.8A), it starts by simulating a tilted series of 2D images from a protein data bank (PDB) [177] structure, using cisTEM [25] to generate noise and experimental artifacts. I simulated combinations of artifacts by varying the electron dosage and sample thickness. The 2D tilt series was simulated from -60 to 60 degrees with 3-degree increments, which reproduces a missing wedge effect.

I first match the densities of the simulated 2D tilt images to an experimental reference dataset (see Methods 2.2) using IMOD. [160] I then applied Topaz 2D denoise [169] to the density-matched 2D tilt images, using the affine model with a patch size of 64. The denoised tilt series was merged together using the EMAN2 [161] and then was contract transfer function (CTF) corrected, reconstructed, and low pass filtered using IMOD. The mock sub-tomograms density was inverted using RELION 4 [148] to harmonize the simulated and experimental datasets.

In addition to the mock sub-tomograms containing protein structures, I also synthesized noise-only sub-tomograms with no proteins as a control data set for the synthesized datasets. These were simulated using the same process as those described above, except that the starting protein structure was replaced by a single carbon atom at the center of the space.

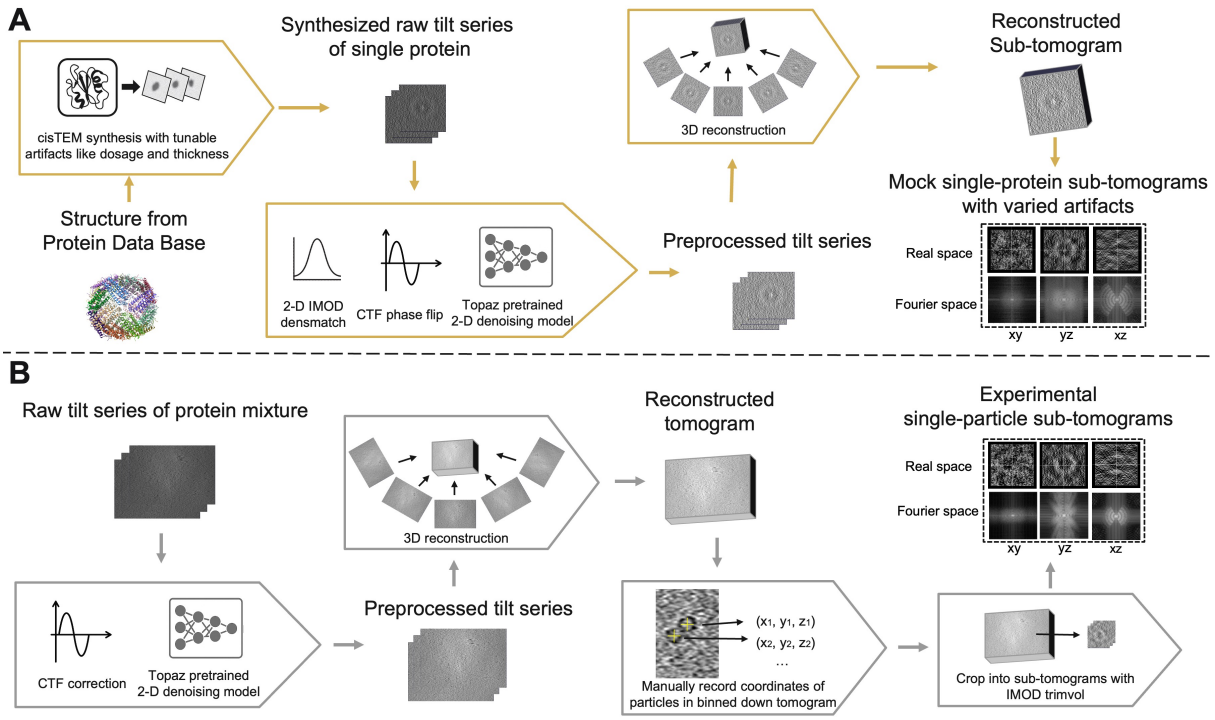


Figure 4.8. Parallel workflows for generating mock and experimental sub-tomograms with tunable control over imaging artifacts. (A) Synthetic sub-tomogram generation starts from known protein structures (e.g., PDB files), which are used to simulate 2D tilt series using cisTEM with controllable imaging parameters (e.g., electron dosage per tilt, sample thickness). The simulated tilt series is further processed to incorporate key cryo-ET artifacts: (1) density distribution alignment using IMOD's densmatch, (2) contrast transfer function (CTF) correction via phase flipping, and (3) denoising using a pretrained 2D Topaz model. The resulting preprocessed tilt series is reconstructed into a 3D tomogram, yielding mock sub-tomograms. These mock volumes exhibit controllable SNR characteristics, making them ideal for model training and benchmarking. (B) Experimental sub-tomograms are generated by reconstructing 3D tomograms from raw tilt series of mixed-protein samples. After preprocessing (CTF correction and denoising), the 3D tomogram is manually inspected to record particle coordinates. Sub-volumes are extracted using IMOD's trimvol, producing experimental sub-tomograms for testing and comparison. Example real-space and Fourier-space slices are shown for mock and experimental sub-tomograms under varying conditions, illustrating structural degradation under low-dose and high-thickness regimes.

Collecting single-particle experimental sub-tomograms:

To validate this approach, I extracted sub-tomograms from an experimental tomogram (Figure 4.2B and 4.8B) containing a mixture of three proteins that are structurally similar but discernable in size and shape: horse spleen light chain apoferritin (PDB ID: 2W0O, stated as

‘apoferritin’ in following texts), -galactosidase (PDB ID: 6DRV, stated as ‘beta-gal’ in following texts) and aldolase (PDB ID:8EW2) (Table 4.1). This solution was made of a sparse mixture of each protein at an individual concentration of 1.5mg/mL in a buffer of 25mM Tris, 2mM MgCl2, 50mM NaCl, and 2mM DTT. 3uL of the protein mixture was loaded onto a 1.2/1.3 holey carbon TEM grid, blotted under 90% RH, and vitrified by plunge freezing into liquid ethane.

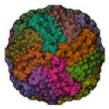
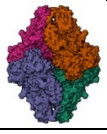
Table 4.1	**apoferritin**	**β-galactosidase (beta-gal)**	**aldolase**
Structural Snapshots			
PDB ID	2W0O	6DRV	8EW2
# of residues	4080	4096	1452
Weight (kDa)	498.24	466.41	157.05
stoichiometry	homo-24-mer	homo-4-mer	homo-4-mer
Symmetry	octahedral	dihedral	dihedral

Table 4.1. The macromolecules in the solution of protein mixture. The protein mixture solution contains three spheroidal macromolecules: apoferritin (horse spleen, light chain), -galactosidase (beta-gal) and aldolase. The size of aldolase was too small for manual particle picking. Therefore, only the particles with the former two proteins were manually picked as a dataset for testing.

Screening and tilt series collection was performed with serialEM [178] in a dose symmetric strategy with images every 3 degrees to +/- 60 degrees. Total cumulative electron dose for the tilt series was 120e/Å^2 at 42,000 times magnification for a pixel size of 1.1 Å/pixel and a 20eV energy filter slit. Raw frames were motion corrected with Motioncor2, [179] denoised with Topaz using the affine model with a patch size of 1024, [169] and the tilt series were aligned and reconstructed in IMOD using back weighted projection. CTF fitting and correction was performed with IMOD’s CTF plotter prior to reconstruction. Tomograms were

visualized with the 3dmod module in IMOD for particle picking, and the sub-volumes were extracted using the ‘trimvol’ command in IMOD. [160]

From a reconstructed tomogram, I worked with Kate Baldwin to manually pick 180 sub-volumes split across each of the three categories: apoferritin, beta-gal and noise-only. Due to its small size, aldolase was not picked. Sub-tomograms centered on the picked particles were extracted using IMOD [160] and inverted with RELION 4. [148] Besides the dosage and thickness, I use consistent cisTEM settings across all the synthesized samples: the tilt angle range is -60 to 60 degrees with an interval of 3.0 degrees, frame is 1, phase is 0.5 and defocus is 30000 angstroms to mimic the experimental settings when collecting the Protein Mixture Solution dataset. The remaining cisTEM settings were set to the default values. All mock sub-tomograms have a volume of 219*219*219 cubic pixels at 1 angstroms/pixel resolution. For computational efficiency I used a uniform box size for all the mock sub-tomograms. The experimental sub-tomograms used a 1.1 angstroms/pixel (‘apix’) resolution, which is close to the mock sub-tomograms mentioned above. I extracted sub-volumes of 219*219*219 cubic pixels to be consistent with the mock sub-tomograms.

4.7.3 HPC enhanced workflow of extracting topological features from mock and experimental density volumes

Density volumes to features

To extract interpretable graph-based features from cryo-electron tomographic data, I

worked with Reece Neff to develop a high-throughput workflow that transforms 3D density volumes (in MRC format) into topological descriptors. This pipeline was optimized for HPC deployment using Parsl [180] and executed on ALCC Perlmutter.

Standard scaling and normalization: a per-volume standard scaling is applied to ensure consistency in density distribution across MRC files. Using scikit-learn's StandardScaler, [112] voxel intensities were rescaled via z-score normalization: $z=(x-\mu)/\sigma$, where μ and σ denote the mean and standard deviation of the MRC volume. This step enables fair comparison between mock and experimental datasets, which often exhibit different dynamic ranges and baseline densities.

Thresholding: A threshold ratioing step retained the top 10% highest-density voxels in each standardized volume. This top-k percentile filtering caps the number of candidate voxels passed downstream, bounding the computational complexity and memory usage of graph construction. The ratio was chosen to maximize feature extraction completeness while ensuring the pipeline could run within time and memory constraints on HPC.

Density coarsening: a density-aware voxel coarsening strategy is introduced with NumPy [67] to reduce spatial and computational load. Voxels were grouped in 3×3×3 windows, and their average position and density were computed to generate a representative centroid. This strategy preserves local density peaks and reduces worst-case complexity up to 1/Csize3 for cube size C, supporting scalability even for volumes with >10 million voxels.

HDBSCAN clustering: To isolate protein signals from background noise, the coarsened point cloud was clustered using HDBSCAN. Unlike DBSCAN, which is sensitive to parameter tuning at low thresholds, HDBSCAN identifies stable clusters over a range of densities, enabling robust segmentation. The largest persistent cluster from each volume was selected as the

candidate region for downstream graph representation. Clustering parameters were tuned to perform Ill across synthetic and experimental datasets (e.g., min_samples=512, min_cluster_size=512).

Graph construction: From the retained cluster points, graphs were constructed across multiple spatial resolutions. Each graph was built by assigning edges between node pairs within a cutoff distance. I varied d'_{cut} from 3 Å to 15 Å to capture persistent structural features across scales. The resulting multi-scale graphs reflect both local density geometry and global topological connectivity.

Persistent graph feature calculation across d'_{cut}: For each graph at every cutoff level, a comprehensive set of 11 graph features was extracted, including degree assortativity, clustering coefficient, eigenvector centrality, closeness centrality, and others. This set of features across every d'_{cut} were aggregated into a single vector representation with a size of 11 graph features x 13 d'_{cut} values = 143 machine learning features per sub-tomogram and saved as CSVs for subsequent machine learning (ML) analysis. To improve scalability, features with poor runtime efficiency are excluded, such as max clique number, which is identified as a key bottleneck during profiling.

By combining these features across biological relevant multiple scales, I created a persistent graph feature vector of the data as a foundation for training explainable machine learning models. These persistent feature vectors are then passed into a Random Forest (RF) [181] model to perform model training and testing, as described in the Methods section 4.

HPC Parallelization Using Parsl

To support scalable feature extraction from thousands of sub-tomograms under various parameter settings, I worked with Reece Neff to deploy the pipeline on the ALCC Perlmutter

supercomputer using Parsl, a Python-based parallel scripting library. [180] Parsl's dataflow-aware scheduler orchestrates concurrent task execution while preserving data dependencies between stages. Each processing step—standardization, thresholding, clustering, graph construction, and feature extraction—was parallelized across independent sub-tomograms or parameter combinations.

On Perlmutter, 1024 nodes can be used to process 4096 sub-tomogram volumes in parallel. Parsl allowed different pipeline stages to run asynchronously across workers, dramatically reducing total wall-clock time. Over 100,000 graph features across the mock and experimental datasets were collected, enabling thorough downstream analysis. Execution time profiling across 45,000 features further guided optimization, identifying steps like clique number and graph construction as major bottlenecks for large, dense graphs. These findings motivated improvements such as feature pruning and the use of coarsening to reduce graph size.

4.7.4 Training, Evaluation and Interpretability of the learning model

Machine learning model:

To evaluate the discriminative power of GRIP-Tomo 2.0 graph-based features and assess their biological relevance, I trained and interpreted a supervised Random Forest (RF) classifier [181] using Python's scikit-learn [112] implementation and 1000 estimators. RF models are Ill suited for classification tasks with low data amounts and provide feature importance values for interpretability.

I constructed multi-class classification including aldolase, apoferritin, beta-gal and noise in Figure 4.4, the Random Forest model was trained and tested on the same mock dataset. For each imaging condition with a certain combination of dosage and thickness, 200 mock sub-

tomograms (50 each of aldolase, apoferritin, beta-gal, and synthetic noise) were simulated following the pipeline in Figure 4.2A and 4.8A. These mock sub-tomograms were then randomly split as 10% for training and 90% for testing.

I also constructed a binary classification task to distinguish between protein-containing and noise-only sub-tomograms in Figure 4.7. The training dataset consisted of 180 mock sub-tomograms (60 each of apoferritin, beta-gal, and synthetic noise), generated under the calibrated imaging condition (electron dosage = 0.3 $e^{-}/Å^{2}$, thickness = 500 Å). The test set comprised 180 experimental sub-tomograms annotated as protein or noise via manual inspection.

Performance evaluation

To evaluate the classification performance of the trained Random Forest model on the testing dataset, I calculated the confusion matrix, accuracy and F1 score using scikit-learn. [112] The confusion matrix provides an overview of the classification performance and contains the counts of true positive, true negative, false positive, and false negative predictions. From these values I can determine the accuracy (ratio of correctly predicted instances to the total instances in the testing dataset) and F1-score (harmonic mean of precision and recall).

Interpretability of feature importance in a learnt model

In GRIP-Tomo 2.0, I calculate the importance of each feature at each d'_{cut} by scikit learn feature importance. [112] The feature importance serves as a metric for measuring the sensitivity of the predicted result to the change of inputs. Thus, this approach provides insights into which features contribute most significantly to the model's prediction.

4.8 DISCUSSION AND CONCLUSION

The evolutionary conserved protein structures are useful features to identify proteins in tomograms with realistic noises

Cryo-electron tomography enables visualization of macromolecular structures in their native context, but its inherently low signal-to-noise ratio (SNR) poses serious challenges for reliable protein classification. GRIP-Tomo 2.0 addresses this limitation by extracting evolutionarily conserved topological patterns—persistent across replicates and imaging conditions—using a graph-based representation of 3D macromolecular architecture. These graph features, by focusing on structural connectivity rather than intensity, remain robust even under noise and missing wedge artifacts, enabling protein detection without the need for segmentation or template matching. Unlike traditional voxel-based workflows, GRIP-Tomo 2.0 emphasizes the geometric backbone of biological structures, allowing classifiers to operate in a topologically informed and biologically meaningful space.

Topological fingerprints enable imaging-aware synthesis design and cross-domain calibration

GRIP-Tomo 2.0 highlights its unique strengths in both predicting classification performance and interpreting data quality through topological fingerprints. By visualizing how persistent features vary across imaging parameters, GRIP-Tomo 2.0 enables researchers to anticipate favorable conditions for both training set design and experimental setup. Moreover, the ability to reveal counterintuitive trends—such as improved aldolase classification at moderate thickness—opens new avenues for investigating the interplay between graph topology,

molecular structure, and imaging physics. These insights point toward future opportunities for mechanistic understanding and methodological refinement in cryo-ET analysis.

By simulating diverse imaging conditions and projecting volumetric information into topological fingerprints, GRIP-Tomo 2.0 establishes a simulation-to-experiment learning framework that informs both data generation and interpretation. Surprisingly, I found that optimal alignment between synthetic and experimental data did not occur under the highest imaging quality, but rather at intermediate conditions—specifically, electron dosage of 0.3 $e^{-}/Å^{2}$ and sample thickness of 500 Å. Under these parameters, synthetic fingerprints most closely resembled those of experimental sub-tomograms, despite increased within-class feature overlap that made protein classification more challenging. These results underscore the importance of realism and task-alignment over maximal contrast in bridging the simulation-to-experiment gap and suggest new principles for generating synthetic data for downstream cryo-ET machine learning tasks.

Graph representation of proteins raises efficient, interpretable and trustworthy machine learning

GRIP-Tomo 2.0 identifies proteins from cryo-ET data by leveraging machine learning on synthetic training data derived from a single atomic structure in the Protein Data Bank (PDB). This approach operates within a limited genomic and structural space, where each protein class originates from a single known conformation. Starting from this minimal input, GRIP-Tomo 2.0 simulates mock sub-tomograms under realistic imaging conditions to generate training examples, enabling automated classification of macromolecules in noisy tomograms. By removing the need for large quantities of annotated, high-quality training data—which are often difficult or impractical to obtain in cryo-ET—GRIP-Tomo 2.0 offers an efficient and accessible alternative for structural discovery. Despite the narrow diversity of structural input, the framework extracts

robust topological features that generalize to experimental data, demonstrating the power of structurally grounded simulation for data-scarce biological inference.

The interpretability of GRIP-Tomo 2.0 stems from its graph-theoretic foundation. Persistent topological features—such as clustering, assortativity, and eigenvector centrality—not only drive classification performance but also correlate with known structural motifs like α-helices, β-sheets, and symmetric domains. Feature importance rankings from trained models consistently identified these biologically relevant descriptors, offering insight into how the cross-domain protein vs. noise classification is achieved. GRIP-Tomo 2.0 fingerprints further allow for systematic benchmarking of imaging conditions: researchers can visually inspect feature separability across a dosage-thickness matrix before choosing optimal simulation parameters for training. This transparency distinguishes GRIP-Tomo 2.0 from black-box deep learning models and enables trustworthy, reproducible workflows.

A key advantage of GRIP-Tomo 2.0 is its ability to generate large volumes of annotated synthetic data tailored to experimental conditions. This capability opens the door for training other AI models, including deep networks, particularly for underrepresented classes such as small proteins or rare conformational states. Researchers in the cryo-ET community can use GRIP-Tomo 2.0 not only as a classification tool, but also as a data generation engine—creating curated mock datasets to bootstrap or augment their own learning pipelines. In addition, the low data requirements needed for classification reduces the need for manual data labeling. By providing both interpretability and extensibility, GRIP-Tomo contributes a versatile framework for structure-aware machine learning in cellular imaging.

Limitations of GRIP-Tomo 2.0

Despite these strengths, several limitations remain. First, the mock data generation requires manual tuning to match experimental conditions. Future work will explore how to automate this process using a data-driven approach. Second, cross-domain classification struggles to distinguish between structurally similar proteins (e.g., apoferritin vs. beta-gal, see Figure 4.9), likely due to fingerprint overlap under current imaging conditions; improved separation may be possible with higher dosage and thinner samples. Third, while I explored the effect of thickness in detail—including the surprising robustness of small proteins like aldolase—future work should systematically vary dosage and build a complete condition-performance matrix. Fourth, feature extraction remains computationally intensive even on high-performance computing clusters; accelerated implementations, potentially on GPUs, are needed for broader adoption. Lastly, thismethod has so far been validated only on a single experimental dataset. Demonstrating its generalizability across other tomograms will be an important next step.

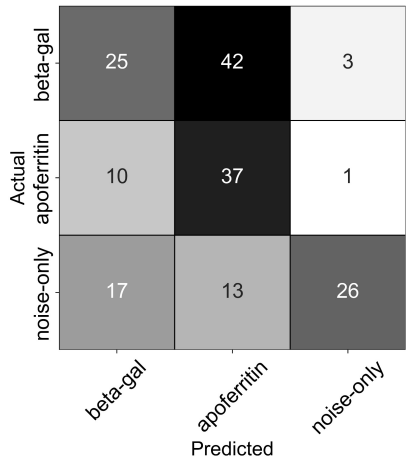


Figure 4.9. Multi-class confusion matrix when RF model is trained on calibrated mock data and tested on experimental data. Although the cross-domain protein vs noise binary classification could achieve 81% accuracy, when assigned the task to classify apoferritin, beta-gal and noise, the RF model shows poor performance in distinguishing the three categories. The overall

accuracy drops to 51%, indicating the fact that apoferritin and beta-gal are not classifiable in this this specific experimental case.

4.9 CONCLUSION

To conclude, GRIP-Tomo 2.0 represents an advancement in computational tools for cryo-electron tomography by providing a novel solution to current limitations in particle classification workflows. Building upon the previous framework, this enhanced platform integrates tunable mock data simulation with realistic noise, and a novel interpretable feature extraction pipeline with high-performance computing acceleration. The key innovation lies in extracting biologically relevant molecular "fingerprints" that enable effective cross-domain learning from simulated to experimental data while requiring minimal training datasets. By improving the interpretability and data-efficiency of macromolecular classification, GRIP-Tomo 2.0 opens new avenues for visual proteomics using large-scale data simulations.

CHAPTER 5: CONCLUSION AND DISCUSSION

This dissertation presents an integrated investigation of cellular phenotypes across biological scales through a physics-informed lens. By combining molecular simulations, mechanochemical modeling, and interpretable machine learning, the work advances a unifying framework centered on topology and hierarchy—concepts that have proven powerful for describing structural organization in living systems. [10, 182]

A central finding across the three projects is that topological descriptors—such as contact interfaces, branching configurations, and graph motifs—remain robust even when system geometry varies or noise is introduced. In Chapter 2, the oligomeric topologies of cofilin dimers were shown to influence actin filament remodeling in a redox-sensitive manner. These conformational states are governed by intermolecular interactions but persist as distinct modes of filament decoration. [54] In Chapter 3, simulations of actomyosin networks revealed that topological transitions—characterized by avalanche-like reorganization—signal instability and precede structural collapse. Predictive features extracted through machine learning, such as network density and centrality variance, enabled early identification of such events. [7, 8]

Chapter 4 extends this topological framework to the cellular scale via GRIP-Tomo 2.0. This pipeline utilizes synthetic cryo-electron tomography (cryo-ET) data to train graph-based classifiers that recognize proteins based on topological fingerprints. Notably, optimal domain alignment between synthetic and experimental datasets did not occur at the highest imaging quality, but rather under intermediate conditions—highlighting a nontrivial structure-noise tradeoff that aligns with recent observations in simulation-to-experiment learning. [37] The

model's interpretability also allowed biological features, such as rotational symmetry or cluster connectivity, to be linked directly to classification outcomes.

Together, these findings suggest that hierarchical topology provides a scale-agnostic vocabulary for bridging molecular detail and whole-cell architecture. By anchoring physical modeling in topological abstraction, one can construct generalizable descriptors of biological function—even under conditions where data are limited, noisy, or heterogeneous. This offers significant promise for emerging fields such as spatial proteomics, integrative structural biology, and in situ cell modeling.

Nonetheless, several limitations remain. First, the simulated and synthetic data used in Chapters 2 and 4, while controlled, cannot fully recapitulate the biochemical heterogeneity or imaging distortions present in live-cell contexts. Second, the classification tasks considered in Chapter 4 focus on a narrow range of protein targets, and it remains unclear how the method scales to more diverse cellular environments. Third, the parameter space for simulation and imaging (e.g., SNR, dosage, defocus) is only partially explored; a more comprehensive sampling could reveal new insights into model generalization. Finally, the computational cost of generating synthetic data and performing graph-based analysis remains a barrier to broad adoption.

Future work should aim to bridge the simulation-to-experiment gap more effectively by calibrating synthetic data against empirical benchmarks and expanding the diversity of protein architectures used for training. Integrating diffusion models or generative adversarial networks (GANs) into synthetic cryo-ET pipelines could improve realism. Additionally, extending GRIP-Tomo 2.0 to handle multi-protein complexes and dynamic interactions—especially with time-resolved cryo-ET—would increase its biological utility. Automation of the entire workflow, from

simulation to classification, is another important direction to enhance scalability and accessibility.

Another promising avenue is to test the robustness of GRIP-Tomo 2.0 in more challenging experimental conditions, such as imaging of crowded cellular environments containing overlapping or interacting macromolecules. Developing strategies to disentangle overlapping topological signatures or adapt classifiers to high-density contexts could substantially improve its applicability to in situ proteomics.

For the cofilin project, future work could leverage advanced structure prediction tools such as AlphaFold-Multimer to systematically explore the oligomerization landscape. This would help identify energetically favorable conformations at larger oligomer sizes and reveal potential structural motifs that govern filament severing efficiency.

Regarding actomyosin networks, recent updates to simulation platforms like MEDYAN now include membrane coupling and curvature feedback. Incorporating these features into the current modeling framework could enable new studies of actin–membrane interaction, cortical tension regulation, and force generation at cellular boundaries.

At the whole-cell level, simulation environments such as Lattice Microbes offer the ability to simulate even more complex intracellular dynamics, including signaling cascades and macromolecular crowding. Coupling GRIP-Tomo's graph abstraction with spatially resolved, stochastic whole-cell simulations may yield unprecedented insight into emergent cellular behavior.

On the modeling front, the development of hybrid frameworks that couple physical simulations with data-driven inference could yield both predictive power and mechanistic

insight. For example, variational coarse-graining techniques or physics-informed neural networks may allow biologically grounded parameter inference from limited experimental input.

Ultimately, this dissertation demonstrates the value of topological reasoning in biological physics. It shows that across scales—from molecular dimers to protein networks to the crowded cellular environment—structural patterns can be distilled into interpretable, machine-readable signatures. These signatures, in turn, provide a foundation for predictive modeling and functional inference. As experimental techniques continue to evolve, and as data integration across modalities improves, topological frameworks like those introduced here may serve as key tools for connecting physics, computation, and biology.

APPENDIX

The research data in this dissertation are stored in University of Washington Hyak Klone High Performance Computing cluster directory: /gscratch/cheung/chengxuan/ChengxuanLi_2025/. The structure of the data is attached below:

(1) The research project studying avalanche and Arp2/3 complex in actomyosin network simulated by MEDYAN, published in https://pubs.acs.org/doi/10.1021/acs.jpcb.1c04792

├── Arp23_avalanche_MEDYAN

│ ├── data_for_paper: the MEDYAN simulations data used in the paper

│ │ ├── 10_14_2020_paperdata_v4_1: simulation data batch 1 for brancher concentration

│ │ ├── 10_14_2020_paperdata_v4_2: simulation data batch 2 for brancher concentration

│ │ ├── 9_23_2019: simulation data batch 3 for brancher concentration

│ │ └── Machine-Learning_data: simulation data for machine learning results

│ ├── manuscript_figures: the figures and codes/slides for creating these figures, in different versions of the manuscript

│ ├── manuscripts: the manuscripts in different versions during the preparation and revision process

│ ├── MEDYAN_analysis&animation_tools: Python codes for animating and analyzing MEDYAN simulation results (Mayavi for animation)

│ └── MEDYAN_softwares: Several versions of MEDYAN software that are used in related research works

(2) The research project studying cofilin oligomerization, published in https://pubs.acs.org/doi/10.1021/acs.jpcb.3c07938

├── cofilin_oligomer

│ ├── manuscripts&figures: the different versions of manuscripts as well as codes/slides for figures during preparation and revision process of the paper

│ ├── simulations: The AWSEM simulations for cofilin oligomerization as well as docking simulations using ClusPro

│ │ ├── 39-39_tetramer_simulations

│ │ ├── 4BEX-147C_mutant_simulations_analysis

│ │ └── ClusPro_docking

│ └── software: The AWSEM software used for the simulations above

│ └── awsemmd-master

(3) The research project about developing the GRIP-Tomo 2.0 framework, manuscript in preparation when this dissertation is written

├── GRIP-Tomo_2.0

│ ├── code_for_manuscript: the codes for plotting figures in the manuscript, can also be found in Gitlab

```
│   │   ├── Figure_3_mrc2features
│   │   ├── Figure_4_and_S3_mock_features_varying_dosage_and_thickness
│   │   └── Figure5_cross-domain_classification
│   ├── data
│   │   ├── csv_for_features: graph features stored in csv files, for fingerprint plotting and machine learning
│   │   │   ├── experimental
│   │   │   ├── mock_before_calibration_dosage=2.9_thickness=250
│   │   │   ├── mock_calibrated_to_experimental_dosage=0.3_thickness=5000
│   │   │   ├── mock_varied_dosage-and-thickness_1-sample-per-categopry
│   │   │   └── mock_varied_thickness_50-samples-per-category
│   │   ├── PDB: PDB files used for cisTEM mock sub-tomogram simulations
│   │   │   ├── apoferritin-betagal-aldolase-noise_50-copies_each_category
│   │   │   └── apoferritin-betagal-noise_60-copies_each_category
│   │   └── subtomograms: sub-tomograms from cisTEM simulations and experimental protein mixture dataset
│   │       ├── experimental_data
│   │       ├── mock_data_before_calibration_dosage=2.9_thickness=250
│   │       ├── mock_data_calibrated_dosage=0.3_thickness=500
```

```
│   │       ├── mock_data_varying_dosage_and_thickness_one_sample_per_condition
│   │       └── mock_data_vary_thickness_explore_boundary
```

LIST OF PUBLICATIONS

- Chengxuan Li, August George, Trevor Moser, Doo Nam Kim, Reece Neff, Arsam Firoozfar, Kate Baldwin, James E Evans, and Margaret S Cheung. “Graph Identification of Proteins in Tomograms (GRIP-Tomo) 2.0: Accelerating Protein Classification for Cryo-Electron Tomography with Intelligent Search” (in preparation).
- Chengxuan Li, Tingyi Wei, Margaret S. Cheung and Min-Yeh Tsai. “Deciphering the Cofilin Oligomers via Intermolecular Disulfide Bond Formation: A Coarse-grained Molecular Dynamics Approach to Understanding Cofilin’s Regulation on Actin Filaments.” The Journal of Physical Chemistry B 128 (19) (2024): 4590–4601. https://doi.org/10.1021/acs.jpcb.3c07938
- Chengxuan Li, James Liman, Yossi Eliaz and Margaret S. Cheung. “Forecasting Avalanches in Branched Actomyosin Networks with Network Science and Machine Learning.” The Journal of Physical Chemistry B 125 (42) (2021): 11591–11605. https://doi.org/10.1021/acs.jpcb.1c04792

BIBLIOGRAPHY

1. Alberts, B., et al., *Molecular Biology of the Cell.* Garland Science, 2015. **6th ed.**
2. Misteli, T., *The concept of self-organization in cellular architecture.* J Cell Biol., 2001. **155(2)**: p. 181-185.
3. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402(6761 Suppl)**: p. C47-C52.
4. Papoian, G.A., et al., *Water in protein structure prediction.* Proc Natl Acad Sci U S A., 2004. **101(10)**: p. 3352-3357.
5. Tojkander, S., G. Gateva, and P. Lappalainen, *Actin stress fibers–assembly, dynamics and biological roles.* J Cell Sci., 2012. **125(Pt 8)**: p. 1855-1864.
6. Dominguez, R. and K.C. Holmes, *Actin structure and function.* Annu Rev Biophys., 2011. **40**: p. 169-186.
7. Kim, T., et al., *Computational analysis of viscoelastic properties of crosslinked actin networks.* PLoS Comput Biol., 2009. **5(7)**: p. e1000439-e1000439.
8. Banerjee, S. and M.C. Marchetti, *Instabilities and oscillations in isotropic active gels.* Soft Matter, 2011. **7**: p. 463-473.
9. Wang, Y. and et al., *A persistent topological feature approach to cryo-electron tomography analysis.* Nature Commun., 2023. **14**: p. 3482-3482.
10. Edelsbrunner, H. and J. Harer, *Computational Topology: An Introduction.* American Mathematical Society, 2010.
11. Popov, K., J. Komianos, and G.A. Papoian, *MEDYAN: Mechanochemical Simulations of Contraction and Polarity Alignment in Actomyosin Networks.* PLoS Comput. Biol., 2016. **12 (4)**: p. e1004877-e1004877.
12. Davtyan, A., et al., *AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing.* J Phys Chem B, 2012. **116**(29): p. 8494-503.
13. Stam, S., et al., *Filament rigidity and connectivity tune the deformation modes of active biopolymer networks.* Proc Natl Acad Sci U S A, 2017. **114**(47): p. E10037-E10045.
14. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
15. Kinman, L.F., et al., *Automated model-free analysis of cryo-EM volume ensembles with SIREn.* bioRxiv, 2024: p. 2024.10.08.617123.
16. Clough, J.R., et al., *A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. **44**(12): p. 8766-8778.
17. Bernstein, B.W. and J.R. Bamburg, *ADF/cofilin: A Functional Node in Cell Biology.* Trends Cell Biol., 2010. **20 (4)**: p. 187-195.
18. Bamburg, J.R., et al., *Cofilin and Actin Dynamics: Multiple Modes of Regulation and Their Impacts in Neuronal Development and Degeneration.* Cells, 2021. **10**: p. 2726-2726.
19. Ennomani, H., et al., *Architecture and Connectivity Govern Actin Network Contractility.* Curr Biol, 2016. **26**(5): p. 616-26.
20. Koenderink, G.H. and E.K. Paluch, *Architecture shapes contractility in actomyosin networks.* Curr. Opin. Cell Biol., 2018. **50**: p. 79-85.

21. Murrell, M., et al., *Forcing cells into shape: the mechanics of actomyosin contractility.* Nat Rev Mol Cell Biol, 2015. **16**(8): p. 486-98.
22. Liman, J., 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**(20): p. 10825-10831.
23. Turk, M. and W. Baumeister, *The promise and the challenges of cryo-electron tomography.* FEBS Letters, 2020. **594**(20): p. 3243-3261.
24. Young, L.N. and E. Villa, *Bringing Structure to Cell Biology with Cryo-Electron Tomography.* Annual Review of Biophysics, 2023. **52**(Volume 52, 2023): p. 573-595.
25. Grant, T., A. Rohou, and N. Grigorieff, *cisTEM, user-friendly software for single-particle image processing.* eLife, 2018. **7**: p. e35383.
26. George, A., et al., *Graph identification of proteins in tomograms (GRIP-Tomo).* Protein Science, 2023. **32**(1): p. e4538.
27. Popov, K., J. Komianos, and G.A. Papoian, *MEDYAN: Mechanochemical Simulations of Contraction and Polarity Alignment in Actomyosin Networks.* PLoS Comput Biol, 2016. **12**(4): p. e1004877.
28. Tsai, M.Y., et al., *Electrostatics, structure prediction, and the energy landscapes for protein folding and binding.* Protein Sci, 2016. **25**(1): p. 255-69.
29. Schafer, N.P., et al., *Learning To Fold Proteins Using Energy Landscape Theory.* Isr J Chem, 2014. **54**(8-9): p. 1311-1337.
30. Lenz, M., et al., *Contractile units in disordered actomyosin bundles arise from F-actin buckling.* Phys Rev Lett, 2012. **108**(23): p. 238107.
31. Chandrasekaran, A., A. Upadhyaya, and G.A. Papoian, *Remarkable structural transformations of actin bundles are driven by their initial polarity, motor activity, crosslinking, and filament treadmilling.* PLoS Comput Biol, 2019. **15**(7): p. e1007156.
32. Mejia-Rodriguez, D., et al., *PTM-Psi: A Python Package to Facilitate the Computational Investigation of Post-Translational Modification on Protein Structures and Their Impacts on Dynamics and Functions.* Protein Sci., 2023. **32**: p. e4822-e4822.
33. Wu, H., P.G. Wolynes, and G.A. Papoian, *AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins.* J Phys Chem B, 2018. **122**(49): p. 11115-11125.
34. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
35. Du, X., et al., *Active learning to classify macromolecular structures in situ for less supervision in cryo-electron tomography.* Bioinformatics, 2021. **37**(16): p. 2340-2346.
36. Goh, G.B., N.O. Hodas, and A. Vishnu, *Deep learning for computational chemistry.* J. Comput. Chem., 2017. **38**: p. 1291-1307.
37. Zhong, E.D., et al., *CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks.* Nat Methods., 2021. **18**: p. 176-185.
38. Powell, B.M. and J.H. Davis, *Learning structural heterogeneity from cryo-electron sub-tomograms with tomoDRGN.* Nature Methods, 2024. **21**(8): p. 1525-1536.
39. Eliaz, Y., et al., *Insights from graph theory on the morphologies of actomyosin networks with multilinkers.* Physical Review E, 2020. **102**(6).
40. Liu, M.-D., L. Ding, and Y.-L. Bai, *Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the ARIMA to wind speed prediction.* Energy Conversion and Management, 2021. **233**: p. 113917.

41. Baek, M., et al., *Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.* Science, 2021. **373**: p. 871-871.
42. Wang, Q., 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**(38): p. 18937-18942.
43. Chan, C., C.C. Beltzner, and T.D. Pollard, *Cofilin dissociates Arp2/3 complex and branches from actin filaments.* Curr Biol, 2009. **19**(7): p. 537-45.
44. Xia, K. and G.-W. Wei, *Persistent homology analysis of protein structure, flexibility, and folding.* International Journal for Numerical Methods in Biomedical Engineering, 2014. **30**(8): p. 814-844.
45. Zhang, B. and P.G. Wolynes, *Topology, structures, and energy landscapes of human chromosomes.* Proc Natl Acad Sci U S A, 2015. **112**(19): p. 6062-7.
46. Porter, M.A., M. Feng, and E. Katifori, *The topology of data.* Physics Today, 2023. **76**(1): p. 36-42.
47. Xia, K. and G.-W. Wei, *Persistent topology for cryo-EM data analysis.* International Journal for Numerical Methods in Biomedical Engineering, 2015. **31**(8).
48. Wang, Y., et al., *A persistent topological feature approach to cryo-electron tomography analysis.* Nature Commun., 2023. **14**: p. 3482-3482.
49. Salbreux, G., G. Charras, and E. Paluch, *Actin Cortex Mechanics and Cellular Morphogenesis.* Trends Cell Biol., 2012. **22 (10)**: p. 536-545.
50. Levayer, R. and T. Lecuit, *Biomechanical Regulation of Contractility: Spatial Control and Dynamics.* Trends Cell Biol., 2012. **22 (2)**: p. 61-81.
51. De La Cruz, E.M. and D. Sept, *The Kinetics of Cooperative Cofilin Binding Reveals Two States of the Cofilin-Actin Filament.* Biophys. J., 2010. **98 (9)**: p. 1893-1901.
52. Bravo-Cordero, J.J., et al., *Functions of Cofilin in Cell Locomotion and Invasion.* Nat. Rev. Mol. Cell Biol., 2013. **14 (7)**: p. 405-415.
53. Paavilainen, V.O., et al., *Structure of the Actin-Depolymerizing Factor Homology Domain in Complex with Actin.* J. Cell Biol., 2008. **182 (1)**: p. 51-59.
54. Galkin, V.E., et al., *Remodeling of Actin Filaments by ADF/cofilin Proteins.* Proc. Natl. Acad. Sci. U.S.A., 2011. **108 (51)**: p. 20568-20572.
55. Kiley, P.J. and G. Storz, *Exploiting Thiol Modifications.* PLoS Biol., 2004. **2 (11)**: p. e400-e400.
56. Pfannstiel, J., et al., *Human Cofilin Forms Oligomers Exhibiting Actin Bundling Activity.* J. Biol. Chem., 2001. **276 (52)**: p. 49476-49484.
57. Goyal, P., et al., *Cofilin Oligomer Formation Occurs in Vivo and Is Regulated by Cofilin Phosphorylation.* PLoS One, 2013. **8 (8)**: p. e71769-e71769.
58. Andrianantoandro, E. and T.D. Pollard, *Mechanism of Actin Filament Turnover by Severing and Nucleation at Different Concentrations of ADF/cofilin.* Mol. Cell, 2006. **24 (1)**: p. 13-23.
59. Klejnot, M., et al., *Analysis of the Human Cofilin 1 Structure Reveals Conformational Changes Required for Actin Binding.* Acta Crystallogr. D Biol. Crystallogr., 2013. **69 (Pt 9)**: p. 1780-1788.
60. Goyal, P., et al., *Cofilin oligomer formation occurs in vivo and is regulated by cofilin phosphorylation.* PLoS One, 2013. **8**(8): p. e71769.

61. Davtyan, A., et al., *AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing.* J. Phys. Chem. B, 2012. **116 (29)**: p. 8494-8503.
62. Tsai, M.Y., et al., *Electrostatics, Structure Prediction, and the Energy Landscapes for Protein Folding and Binding.* Protein Sci., 2016. **25 (1)**: p. 255-269.
63. Zheng, W., et al., *Exploring the Aggregation Free Energy Landscape of the Amyloid-β Protein (1–40).* Proc. Natl. Acad. Sci. U.S.A., 2016. **113 (42)**: p. 11835-11840.
64. Zheng, W., M.Y. Tsai, and P.G. Wolynes, *Comparing the Aggregation Free Energy Landscapes of Amyloid Beta(1–42) and Amyloid Beta(1–40).* J. Am. Chem. Soc., 2017. **139 (46)**: p. 16666-16676.
65. Ma, Y.W., T.Y. Lin, and M.Y. Tsai, *Fibril Surface-Dependent Amyloid Precursors Revealed by Coarse-Grained Molecular Dynamics Simulation.* Front Mol. Biosci, 2021. **8**: p. 719320-719320.
66. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual Molecular Dynamics.* J. Mol. Graph., 1996. **14 (1)**: p. 33-38.
67. Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. **585**(7825): p. 357-362.
68. Scherer, M.K., et al., *PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models.* J. Chem. Theory Comput., 2015. **11 (11)**: p. 5525-5542.
69. McGibbon, R.T., et al., *MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories.* Biophys. J., 2015. **109 (8)**: p. 1528-1532.
70. Parra, R.G., et al., *Protein Frustratometer 2: A Tool to Localize Energetic Frustration in Protein Molecules, Now with Electrostatics.* Nucleic Acids Res., 2016. **44 (W1)**: p. W356-W360.
71. Ferreiro, D.U., E.A. Komives, and P.G. Wolynes, *Frustration in Biomolecules.* Q. Rev. Biophys., 2014. **47 (4)**: p. 285-363.
72. Gianni, S., et al., *Fuzziness and Frustration in the Energy Landscape of Protein Folding, Function, and Assembly.* Acc. Chem. Res., 2021. **54 (5)**: p. 1251-1259.
73. Jenik, M., et al., *Protein Frustratometer: A Tool to Localize Energetic Frustration in Protein Molecules.* Nucleic Acids Res., 2012. **40**: p. W348-W351.
74. Pfannstiel, J., et al., *Human cofilin forms oligomers exhibiting actin bundling activity.* J Biol Chem, 2001. **276**(52): p. 49476-84.
75. Klamt, F., et al., *Oxidant-Induced Apoptosis Is Mediated by Oxidation of the Actin-Regulatory Protein Cofilin.* Nat. Cell Biol., 2009. **11 (10)**: p. 1241-1246.
76. Phillips, J.C., et al., *Scalable molecular dynamics with {NAMD}.* J. Comput. Chem., 2005. **26 (16)**: p. 1781-1802.
77. Ferreiro, D.U., et al., *Localizing Frustration in Native Proteins and Protein Assemblies.* Proc. Natl. Acad. Sci. U.S.A., 2007. **104 (50)**: p. 19819-19824.
78. Ferreiro, D.U., E.A. Komives, and P.G. Wolynes, *Frustration, Function and Folding.* Curr. Opin. Struct. Biol., 2018. **48**: p. 68-73.
79. Mouro, P.R., et al., *Quantifying Nonnative Interactions in the Protein-Folding Free-Energy Landscape.* Biophys. J., 2016. **111 (2)**: p. 287-293.
80. De La Cruz, E.M., *How Cofilin Severs an Actin Filament.* Biophys. Rev., 2009. **1 (2)**: p. 51-59.
81. De La Cruz, E.M., et al., *Origin of Twist-Bend Coupling in Actin Filaments.* Biophys. J., 2010. **99 (6)**: p. 1852-1860.

82. De La Cruz, E.M., J.L. Martiel, and L. Blanchoin, *Mechanical Heterogeneity Favors Fragmentation of Strained Actin Filaments.* Biophys. J., 2015. **108 (9)**: p. 2270-2281.
83. Schramm, A.C., et al., *Actin Filament Strain Promotes Severing and Cofilin Dissociation.* Biophys J, 2017. **112**(12): p. 2624-2633.
84. Fan, J., et al., *Molecular Origins of Cofilin-Linked Changes in Actin Filament Mechanics.* J. Mol. Biol., 2013. **425 (7)**: p. 1225-1240.
85. Galkin, V.E., et al., *Actin Depolymerizing Factor Stabilizes an Existing State of F-Actin and Can Change the Tilt of F-Actin Subunits.* J. Cell Biol., 2001. **153 (1)**: p. 75-86.
86. Hocky, G.M., et al., *Structural Basis of Fast- and Slow-Severing Actin–cofilactin Boundaries.* J. Biol. Chem., 2021. **296**: p. 100337-100337.
87. Huehn, A., et al., *The Actin Filament Twist Changes Abruptly at Boundaries between Bare and Cofilin-Decorated Segments.* J. Biol. Chem., 2018. **293 (15)**: p. 5377-5383.
88. Paulsen, C.E. and K.S. Carroll, *Cysteine-Mediated Redox Signaling: Chemistry, Biology, and Tools for Discovery.* Chem. Rev., 2013. **113 (7)**: p. 4633-4679.
89. Svitkina, T.M., *Actin Cell Cortex: Structure and Molecular Organization.* Trends Cell Biol., 2020. **30**: p. 556-565.
90. Svitkina, T.M. and G.G. Borisy, *Arp2/3 Complex and Actin Depolymerizing Factor/Cofilin in Dendritic Organization and Treadmilling of Actin Filament Array in Lamellipodia.* J. Cell Biol., 1999. **145**: p. 1009-1026.
91. Mullins, R.D., J.A. Heuser, and T.D. Pollard, *The Interaction of Arp2/3 Complex with Actin: Nucleation, High Affinity Pointed EndCapping, and Formation of Branching Networks of Filaments.* Proc. Natl. Acad. Sci. U. S. A., 1998. **95**: p. 6181-6186.
92. Malik-Garbi, M., et al., *Scaling behaviour in steady-state contracting actomyosin networks.* Nat. Phys., 2019. **15**: p. 509-516.
93. Papalazarou, V. and L.M. Machesky, *The cell pushes back: The Arp2/3 complex is a key orchestrator of cellular responses to environmental forces.* Curr. Opin. Cell Biol., 2021. **68**: p. 37-44.
94. Nedelec, F. and D. Foethke, *Collective Langevin dynamics of flexible cytoskeletal fibers.* New J. Phys., 2007. **9**: p. 427-427.
95. Freedman, S.L., et al., *A Versatile Framework for Simulating the Dynamic Mechanical Structure of Cytoskeletal Networks.* Biophys. J., 2017. **113**: p. 448-460.
96. Hu, L. and G.A. Papoian, *Molecular transport modulates the adaptive response of branched actin networks to an external force.* J. Phys. Chem. B, 2013. **117**: p. 13388-96.
97. Hu, L. and G.A. Papoian, *Mechano-chemical feedbacks regulate actin mesh growth in lamellipodial protrusions.* Biophys J, 2010. **98**(8): p. 1375-84.
98. Lan, Y. and G.A. Papoian, *The stochastic dynamics of filopodial growth.* Biophys. J., 2008. **94**: p. 3839-52.
99. Carlsson, A.E., *The effect of branching on the critical concentration and average filament length of actin.* Biophys. J., 2005. **89**: p. 130-40.
100. Alencar, A.M., et al., *Non-equilibrium cytoquake dynamics in cytoskeletal remodeling and stabilization.* Soft Matter, 2016. **12**: p. 8506-8511.
101. Kane, R.E., *Interconversion of Structural and Contractile Actin Gels by Insertion of Myosin during Assembly.* J. Cell Biol., 1983. **97 (6)**: p. 1745-1752.
102. Janson, L.W., J. Kolega, and D.L. Taylor, *Modulation of Contraction by Gelation/Solation in a Reconstituted Motile Model.* J. Cell Biol., 1991. **114**: p. 1005-1015.

103. Bendix, P.M., et al., *A quantitative analysis of contractility in active cytoskeletal protein networks.* Biophys J, 2008. **94**(8): p. 3126-36.
104. Bendix, P.M., et al., *A quantitative analysis of contractility in active cytoskeletal protein networks.* Biophys. J., 2008. **94**: p. 3126-36.
105. Tan, T.H., et al., *Self-organized stress patterns drive state transitions in actin cortices.* Sci. Adv., 2018. **4**: p. eaar2847-eaar2847.
106. Hu, L. and G.A. Papoian, *Molecular transport modulates the adaptive response of branched actin networks to an external force.* J Phys Chem B, 2013. **117**(42): p. 13388-96.
107. Ni, Q. and G.A. Papoian, *Turnover versus treadmilling in actin network assembly and remodeling.* Cytoskeleton (Hoboken), 2019. **76**(11-12): p. 562-570.
108. Baldi, P., P. Sadowski, and D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning.* Nat. Commun., 2014. **5**: p. 4308-4308.
109. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**: p. 115-118.
110. Goh, G.B., N.O. Hodas, and A. Vishnu, *Deep learning for computational chemistry.* J Comput Chem, 2017. **38**(16): p. 1291-1307.
111. Park, Y. and M. Kellis, *Deep learning for regulatory genomics.* Nat Biotechnol, 2015. **33**(8): p. 825-6.
112. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* JMLR, 2011. **12**: p. 2825-2830.
113. Chen, T. and C. Guestrin, *A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: p. 785-794.
114. Floyd, C., G.A. Papoian, and C. Jarzynski, *Quantifying dissipation in actomyosin networks.* Interface Focus, 2019. **9**: p. 20180078-20180078.
115. Hu, L. and G.A. Papoian, *How does the antagonism between capping and anti-capping proteins affect actin network dynamics?* J Phys Condens Matter, 2011. **23**(37): p. 374101.
116. Komianos, J.E. and G.A. Papoian, *Stochastic Ratcheting on a Funneled Energy Landscape Is Necessary for Highly Efficient Contractility of Actomyosin Force Dipoles.* Physical Review X, 2018. **8**(2).
117. Ramachandran, P. and G. Varoquaux, *Mayavi: 3D Visualization of Scientific Data.* Comput. Sci. Eng., 2011. **13**: p. 40-51.
118. Pollard, T.D., *Rate Constants for the Reactions of ATP- and ADP-Actin with the Ends of Actin Filaments.* J. Cell Biol., 1986. **103**: p. 2747-2754.
119. Wachsstock, D.H., W.H. Schwartz, and T.D. Pollard, *Affinity of a-Actinin for Actin Determines the Structure and Mechanical Properties of Actin Filament Gels.* Biophys. J., 1993. **65**: p. 205-214.
120. Kovacs, M., et al., *Functional divergence of human cytoplasmic myosin II: kinetic characterization of the non-muscle IIA isoform.* J. Biol. Chem., 2003. **278**: p. 38132-40.
121. Dima, R.I. and D. Thirumalai, *Asymmetry in the Shapes of Folded and Denatured States of Proteins.* J. Phys. Chem. B, 2004. **108**: p. 6564-6570.
122. Pollard, T.D., *Structure and Polymerization of Acanthamoeba Myosin-II Filaments.* J. Cell Biol., 1982. **95**: p. 816-825.
123. Meyer, R.K. and U. Aebi, *Bundling of Actin Filaments by α-Actinin Depends on Its Molecular Length.* J. Cell Biol., 1990. **110**: p. 2013-2024.

124. Hagberg, A.A., D.A. Schult, and P.J. Swart, *Exploring Network Structure, Dynamics, and Function using NetworkX.* Proceedings of the 7th Python in Science Conference (SciPy 2008), 2008: p. 11-15.
125. Newman, M.E., *Mixing patterns in networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2003. **67**(2 Pt 2): p. 026126.
126. Fan, J., et al., *Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China.* Energy Convers. Manage., 2018. **164**: p. 102-111.
127. Qin, C., et al., *XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring.* Math. Probl. Eng., 2021. **2021**: p. 1-18.
128. Fan, J., et al., *Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China.* Agricultural and Forest Meteorology, 2018. **263**: p. 225-241.
129. Bastian, M., S. Heymann, and M. Jacomy, *Gephi: An Open Source Software for Exploring and Manipulating Networks.* International AAAI Conference on Weblogs and Social Media, 2009.
130. Alencar, A.M., et al., *Non-equilibrium cytoquake dynamics in cytoskeletal remodeling and stabilization.* Soft Matter, 2016. **12**(41): p. 8506-8511.
131. Bradley, A.E., *THE USE OF THE AREA UNDER THE ROC CURVE IN THE EVALUATION OF MACHINE LEARNING ALGORITHMS.* Pattern Recognition, 1997. **30**: p. 1145-1159.
132. Davis, J. and M. Goadrich, *The Relationship Between Precision-Recall and ROC Curves.* Proceedings of the 23 rd International Conference on Machine Learning, 2006.
133. Flach, P.A. and M. Kull, *Precision-Recall-Gain Curves: PR Analysis Done Right.* Advances in Neural Information Processing Systems 28 (NIPS 2015), 2015. **28**.
134. Fäßler, F., et al., *Novel cryo-electron tomography structure of Arp2/3 complex in cells reveals mechanisms of branch formation.* bioRxiv, 2020.
135. Pollard, T.D. and G.G. Borisy, *Cellular Motility Driven by Assembly and Disassembly of Actin Filaments.* Cell, 2003. **112**: p. 453-465.
136. Delatour, V., et al., *Arp2/3 controls the motile behavior of N-WASP-functionalized GUVs and modulates N-WASP surface distribution by mediating transient links with actin filaments.* Biophys. J., 2008. **94**: p. 4890-905.
137. Welch, M.D., et al., *The Human Arp2/3 Complex Is Composed of Evolutionarily Conserved Subunits and Is Localized to Cellular Regions of Dynamic Actin Filament Assembly.* J. Cell Biol., 1997. **138**: p. 375-384.
138. Ichetovkin, I., W. Grant, and J. Condeelis, *Cofilin Produces Newly Polymerized Actin Filaments that Are Preferred for Dendritic Nucleation by the Arp2/3 Complex.* Curr. Biol., 2002. **12**: p. 79-84.
139. Bravo-Cordero, J.J., et al., *Functions of cofilin in cell locomotion and invasion.* Nat Rev Mol Cell Biol, 2013. **14**(7): p. 405-15.
140. Floyd, C., et al., *Understanding cytoskeletal avalanches using mechanical stability analysis.* arXiv, 2021.
141. Rukhlenko, O., B.N. Kholodenko, and W. Kolch, *Systems biology approaches to macromolecules: the role of dynamic protein assemblies in information processing.* Current Opinion in Structural Biology, 2021. **67**: p. 61-68.

142. Russel, D., et al., *The structural dynamics of macromolecular processes.* Current Opinion in Cell Biology, 2009. **21**(1): p. 97-108.
143. Schur, F.K.M., *Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging.* Current Opinion in Structural Biology, 2019. **58**: p. 1-9.
144. Ni, T., et al., *High-resolution in situ structure determination by cryo-electron tomography and subtomogram averaging using emClarity.* Nature Protocols, 2022. **17**(2): p. 421-444.
145. Wang, H.-W. and J.-W. Wang, *How cryo-electron microscopy and X-ray crystallography complement each other.* Protein Science, 2017. **26**(1): p. 32-39.
146. Letertre, M.P.M., P. Giraudeau, and P. de Tullio, *Nuclear Magnetic Resonance Spectroscopy in Clinical Metabolomics and Personalized Medicine: Current Challenges and Perspectives.* Frontiers in Molecular Biosciences, 2021. **Volume 8 - 2021**.
147. Kühlbrandt, W., *The Resolution Revolution.* Science, 2014. **343**(6178): p. 1443-1444.
148. Scheres, S.H.W., *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, 2012. **180**(3): p. 519-530.
149. Tegunov, D. and P. Cramer, *Real-time cryo-electron microscopy data preprocessing with Warp.* Nature Methods, 2019. **16**(11): p. 1146-1152.
150. Chen, M., et al., *A complete data processing workflow for cryo-ET and subtomogram averaging.* Nature Methods, 2019. **16**(11): p. 1161-1168.
151. Galaz-Montoya, J.G., et al., *Alignment algorithms and per-particle CTF correction for single particle cryo-electron tomography.* Journal of Structural Biology, 2016. **194**(3): p. 383-394.
152. Hrabe, T., et al., *PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis.* Journal of Structural Biology, 2012. **178**(2): p. 177-188.
153. Zivanov, J., et al., *A Bayesian approach to single-particle electron cryo-tomography in RELION-4.0.* eLife, 2022. **11**: p. e83724.
154. Scaramuzza, S. and D. Castaño-Díez, *Step-by-step guide to efficient subtomogram averaging of virus-like particles with Dynamo.* PLOS Biology, 2021. **19**(8): p. e3001318.
155. Balyschew, N., et al., *Streamlined structure determination by cryo-electron tomography and subtomogram averaging using TomoBEAR.* Nature Communications, 2023. **14**(1): p. 6543.
156. Zhao, C., et al., *Computational methods for in situ structural studies with cryogenic electron tomography.* Frontiers in Cellular and Infection Microbiology, 2023. **13**.
157. Zheng, T. and S. Cai, *Recent technical advances in cellular cryo-electron tomography.* The International Journal of Biochemistry & Cell Biology, 2024. **175**: p. 106648.
158. Chen, Z., et al., *De novo protein identification in mammalian sperm using in situ cryoelectron tomography and AlphaFold2 docking.* Cell, 2023. **186**(23): p. 5041-5053.e19.
159. Rice, G., et al., *TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining.* Nature Methods, 2023. **20**(6): p. 871-880.
160. Kremer, J.R., D.N. Mastronarde, and J.R. McIntosh, *Computer Visualization of Three-Dimensional Image Data Using IMOD.* Journal of Structural Biology, 1996. **116**(1): p. 71-76.

161. Tang, G., et al., *EMAN2: An extensible image processing suite for electron microscopy.* Journal of Structural Biology, 2007. **157**(1): p. 38-46.
162. Himes, B.A. and P. Zhang, *emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging.* Nature Methods, 2018. **15**(11): p. 955-961.
163. Chaillet, M.L., et al. *Extensive Angular Sampling Enables the Sensitive Localization of Macromolecules in Electron Tomograms*. International Journal of Molecular Sciences, 2023. **24**, DOI: 10.3390/ijms241713375.
164. Moebel, E., et al., *Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms.* Nature Methods, 2021. **18**(11): p. 1386-1394.
165. Zeng, X., et al., *High-throughput cryo-ET structural pattern mining by unsupervised deep iterative subtomogram clustering.* Proceedings of the National Academy of Sciences, 2023. **120**(15): p. e2213149120.
166. Lamm, L., et al., *MemBrain: A deep learning-aided pipeline for detection of membrane proteins in Cryo-electron tomograms.* Computer Methods and Programs in Biomedicine, 2022. **224**: p. 106990.
167. Zhang, H., et al., *A method for restoring signals and revealing individual macromolecule states in cryo-ET, REST.* Nature Communications, 2023. **14**(1): p. 2937.
168. Genthe, E., et al., *PickYOLO: Fast deep learning particle detector for annotation of cryo electron tomograms.* Journal of Structural Biology, 2023. **215**(3): p. 107990.
169. Bepler, T., et al., *Topaz-Denoise: general deep denoising models for cryoEM and cryoET.* Nature Communications, 2020. **11**(1): p. 5208.
170. Moebel, E. and C. Kervrann, *Towards unsupervised classification of macromolecular complexes in cryo electron tomography: Challenges and opportunities.* Computer Methods and Programs in Biomedicine, 2022. **225**: p. 107017.
171. Carlsson, G., *Topological methods for data modelling.* Nature Reviews Physics, 2020. **2**(12): p. 697-708.
172. Chazal, F. and B. Michel, *An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists.* Frontiers in Artificial Intelligence, 2021. **4**.
173. Hartsock, I., et al., *Topological data analysis of pattern formation of human induced pluripotent stem cell colonies.* bioRxiv, 2024: p. 2024.05.07.592985.
174. Schubert, E., et al., *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.* ACM Trans. Database Syst., 2017. **42**(3): p. Article 19.
175. McInnes, L. and J. Healy. *Accelerated hierarchical density based clustering*. in *2017 IEEE international conference on data mining workshops (ICDMW)*. 2017. IEEE.
176. Campello, R., D. Moulavi, and J. Sander, *Density-Based Clustering Based on Hierarchical Density Estimates*. Vol. 7819. 2013. 160-172.
177. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
178. Mastronarde, D.N., *SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position.* Microscopy and Microanalysis, 2003. **9**(S02): p. 1182-1183.
179. Zheng, S.Q., et al., *MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy.* Nature Methods, 2017. **14**(4): p. 331-332.
180. Babuji, Y., et al., *Parsl: Pervasive Parallel Programming in Python*, in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*. 2019, Association for Computing Machinery: Phoenix, AZ, USA. p. 25–36.

181. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
182. Bronstein, M.M., et al., *Geometric deep learning: going beyond Euclidean data.* IEEE Signal Process Mag., 2017. **34(4)**: p. 18-42.