

FROM METABOLITES TO MACROMOLECULES: COMPUTATIONAL MODELS OF THE
E. COLI AND *H. SAPIENS* CYTOPLASMS

BY

MEREDITH RICKARD

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Professor Martin Gruebele, Chair and Co-Director of Research
Professor Taras Pogorelov, Co-Director of Research
Professor Zaida Luthey-Schulten
Professor Wilfred van der Donk

ABSTRACT

In recent years, there has been an increasing emphasis on the impact of the cellular environment on biomolecular dynamics. In particular, there is growing evidence that the cell's environment is finely tuned to modulate the stability, kinetics, binding, and activity of proteins within it. This thesis uses computational techniques to better understand the atomistic mechanisms underlying these impacts on protein dynamics.

I begin by describing the motivation and challenges of simulating biomolecules in-cell. In Chapter 1, I discuss the in-cell environment and introduce crowding, sticking, and quinary interactions, the basic interactions that affect protein-dynamics in-cell. I briefly highlight several experimental studies that have demonstrated the impact of the cellular environment on protein dynamics over the last fifteen years. Then, I discuss the history and current landscape of simulating biomolecules in cellular conditions, including notable studies and ongoing challenges. In Chapter 2, I describe the design, construction, and simulation of my contribution to this landscape—several ~200,000 atom models that depict small sections of the *E. coli* and *H. sapiens* cytoplasm in atomistic detail.

Then, I move into the results of these cytosolic simulations, including the effect of the cellular environment on ATP (Chapter 3), the partial folding of a small, fast-folding WW domain in the *E. coli* models (Chapter 4), and the nonspecific protein-protein sticking observed throughout my simulations (Chapter 5). Additionally, I present preliminary results from the *H. sapiens* cytoplasm simulations, including partial folding of another fast-folding construct and quinary interactions between two enzymes (Chapter 6).

Finally, I describe the results of a collaboration between the Gruebele group, the Pogorelov group, Professor Stephen Taylor (UIUC Department of Music), Kurt Hebel, and Carla Scaletti (both of the sound design software company Symbolic Sound). Together, we have developed data sonification techniques to use sound as a teaching tool and data analysis aid in protein folding research (Chapter 7).

To Mom and Dad

ACKNOWLEDGMENTS

As I finish my PhD, I realize how lucky I have been to have not one, but two amazing advisors. Professor Martin Gruebele and Professor Taras Pogorelov have taught me so much about science and leadership and have equipped me with the skills to move forward to my next adventure. Particularly, I want to thank Martin’s willingness to let me drive my projects—I was always allowed to pursue my own ideas, and I was always supported and guided when things didn’t go to plan. This gave me the freedom to grow intellectually and form my own approach to research problems. Taras’s guidance and support has been crucial to my success here. I needed to learn computational skills I needed for this project very quickly, and he invested the time to make sure I understood what I was doing. Throughout my time in the Pogorelov group, he has celebrated my successes and encouraged me through the challenges every step of the way. I am proud to call both of them advisors and friends.

I was incredibly lucky to have a “dream team” of mentors as I was starting out. Yi Zhang started the cytoplasm modeling project and taught me everything I know about molecular dynamics and simulations. Caitlin Davis has been an amazing mentor and friend throughout this process. From reading over an abstract to taking a long lunch to look at my data and discuss my results, she has offered so much help and has taught me so much about being a scientist. I also want to thank Shahar Sukenik, Drishti Guin, Kapil Dave, and Lydia Kisley for their support and mentorship. In turn, I want to thank the students who patiently taught me how to be a mentor: my three longtime officemates, Mayank Boob, Gopika Gopan, and Yuhan Wang.

All members of both the Pogorelov and Gruebele groups have been amazing sources of friendship and support. From the Pogorelov group, I would like to thank Haolin Luo (my stellar undergraduate researcher), Ashley De Lio, Kevin Cheng, Zachary Wang, Mike Hallock, Alex Bieniek, Artem Mustaev, Juan David Campolargo, Andrew Lee, and Vadim Mustaev. From the Gruebele group, I would like to thank Drishti Guin, Kapil Dave, Caitlin Davis, Shahar Sukenik, Lydia Kisley, Duc Nguyen, Shu-Han Chao, Tanya Perlova, Ruopei Feng, Huy Nguyen, Yuhan Wang, Gopika Gopan, Alison Wallum, Aniket Ravan, Zane Thornburg, Chenghao Zhang, Arash Firouzbakht, Mahima Unnikrishnan, Zhaleh Ghaemi, and Nikolai Kocherginski. A special thanks must be extended to Mayank Boob, my “double labmate” in both the Pogorelov and Gruebele groups.

I must also thank my collaborators on the data sonification project, Dr. Carla Scaletti, Professor Stephen Taylor, Dr. Kurt Hebel, and Franz Danksagmüller. I am grateful for their eagerness to learn about protein folding and their willingness to teach me about music. All of them lent their considerable talent to the sonifications, for which I am very grateful.

Further, I owe thanks to the incredible support staff in the School of Chemical Sciences. The IMP office staff, including Karen Watson, Beth Myler, Theresa Struss, Stacy Walker, Katie Trabaris, Lisa Johnson, and Wendy Wimmer, have been fantastic friends and incredible resources to help me navigate graduate school. In the departmental office, I would like to thank Connie Knight and Lloyd Munjanja.

I would also like to thank my fellow trainees both at UIUC and beyond. My physical chemistry cohort, which consists of lifelong friends, have helped me in so many ways—from working together on stat mech homework to giving me pep talks on the hard days. In the larger protein-folding field, I have been so lucky to meet a fantastic network of graduate students and postdocs at various conferences. They have all been incredibly generous with their time, know-how, and encouragement. I look forward to the day that these trainees are, themselves, leaders in our field.

Martin and Taras are merely the latest in a long list of inspiring educators I have been taught by. Mrs. Romera (kindergarten), Mrs. Ross (first grade), Mr. McConnell (third grade), Mrs. Cybulski (fourth grade), Mrs. Sherman (eighth grade), Mr. Massie (high school), Dr. Dixon (undergraduate research advisor), and so many more have challenged me and encouraged me to continue pursuing my education.

Last, I would like to thank my family for their support and love. I am grateful to my boyfriend, Jesse, who has been the very best part of the last five years here in Champaign-Urbana. My brother and grandparents have cheered me on from afar every step of the way. Finally, I would like to thank my parents, whose unwavering belief in me has carried me through this experience.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
CHAPTER 2: METHODS TO CONSTRUCT MODELS OF THE <i>E. COLI</i> AND HUMAN CYTOPLASMS IN <i>SILICO</i>	15
CHAPTER 3: THE IMPACT OF THE CELLULAR ENVIRONMENT ON SIMULATED METABOLITE CONFORMATIONS.....	34
CHAPTER 4: CROWDING, STICKING, AND PARTIAL FOLDING OF GTT WW DOMAIN IN A SMALL <i>E. COLI</i> CYTOPLASM MODEL	45
CHAPTER 5: IN-CELL PROTEIN-PROTEIN CONTACTS: TRANSIENT INTERACTIONS IN THE CROWD	60
CHAPTER 6: PROTEIN DYNAMICS IN AN ALL-ATOM MODEL OF THE <i>H. SAPIENS</i> CYTOPLASM	74
CHAPTER 7: DATA SONIFICATION TO ENHANCE TEACHING AND RESEARCH	85
APPENDIX A: SUPPLEMENTARY INFORMATION FOR CH. 2: METHODS	109
APPENDIX B: SUPPLEMENTARY INFORMATION FOR CH. 4: FOLDING OF GTT IN <i>E. COLI</i> MODELS.....	120
APPENDIX C: SUPPLEMENTARY INFORMATION FOR CH. 5: IN-CELL PROTEIN-PROTEIN CONTACTS	125
APPENDIX D: SUPPLEMENTARY INFORMATION FOR CH. 7: DATA SONIFICATION	128

CHAPTER 1: INTRODUCTION

PROTEIN FOLDING AND DYNAMICS

Protein folding has been studied quantitatively *in vitro* for over half a century starting with experiments that revealed the structural, kinetic and thermodynamic principles behind folding of small proteins,^{2,3} soon followed by computational efforts to model folding and unfolding events.^{4,5} The energy landscape theory of protein folding that crystallized in the 1990s quantified the paradigm of entropy-enthalpy compensation that allows protein folding to be such a fast (microseconds to minutes) chemical reaction with small barriers and shallow free energy minima.^{6,7} Starting about 20 years ago, *in vitro* computational and experimental efforts began to converge,^{8,9} culminating in computational reversible folding of a dozen small proteins¹⁰ and more since then.¹¹

Simultaneously, methods to predict the structure of a folded protein from its amino acid sequence have been developed, including associative memory Hamiltonians^{12,13} and fragment based approaches.^{14–16} Most recently, AlphaFold, an artificial intelligence program that can predict a protein's backbone conformations with sub-angstrom resolution, was released.¹⁷ While these efforts have advanced the field's ability to predict folded structures and identify pathways between the unfolded and folded state *in vitro*, questions about protein folding in the cellular environment remain.¹⁸

A protein that is folding *in vivo* must navigate a heterogeneous environment filled with metabolites, inorganic ions, water, and many types of macromolecules (Figure 1.1). Neighboring macromolecules are considered to be particularly important because macromolecules are estimated

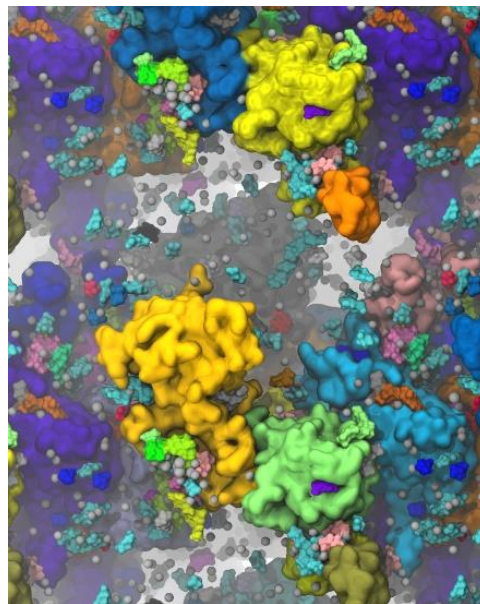


Figure 1.1: Proteins fold and perform their biological functions in a crowded, complex milieu of inorganic ions, metabolites, and other macromolecules.

* Some writing and figures in this chapter are reproduced with permission from:

Gopan, G.; Gruebele, M.; Rickard, M. M. In-Cell Protein Landscapes: Making the Match between Theory, Simulation and Experiment. *Curr. Opin. Struct. Biol.* **2021**, 66, 163–169.

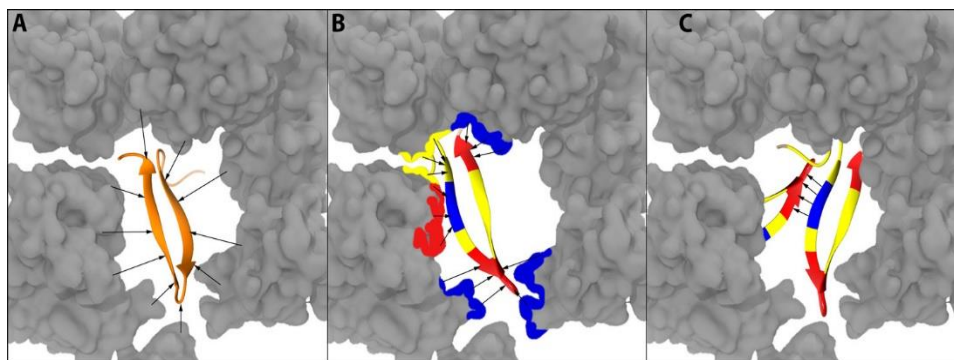


Figure 1.2: Crowding, sticking, and quinary interactions drive many effects of the cellular environment on macromolecules. (a) Crowding: a protein of interest (orange) experiences steric effects from neighboring macromolecules (gray). (b) Sticking: hydrophobic (yellow) and charged (red and blue) patches on the protein of interest (ribbon) and neighboring macromolecules nonspecifically interact. (c) Quinary Interactions: negatively charged residues on the protein of interest (blue, ribbon) functionally form transient interactions with specific positive (red) residues on a neighboring protein.*

to occupy 10-40% of a cell's volume,¹⁹ but ions and small metabolites also play a big role by screening and binding.²⁰ A folding protein's shallow free energy minima means that cell cycle, cell compartments, cytoskeleton and other features of the cell will modulate protein folding and protein interactions in space and time.^{20,21}

EXPERIMENTAL STUDIES OF IN-CELL PROTEIN FOLDING, BINDING, AND DYNAMICS

A number of recent reviews have covered the advances in theoretical, computational and experimental analysis of in-cell protein folding stability, diffusion and protein-protein interactions prior to the last three years.^{21–24} Here, I discuss the basic concepts of crowding, sticking, and quinary interactions, and review a number of studies over the last decade that have illustrated the effect of the cellular environment on proteins.

Crowding, Sticking and Quinary Interactions

The root causes of major in-cell effects on proteins observed experimentally are largely attributed to three major areas: crowding, sticking, and quinary structure (Figure 1.2). Crowding, or the steric repulsion of neighboring macromolecule in the cellular environment, is predicted to

*Figure prepared by Drishti Guin, Reprinted with permission from:

Guin, D.; Grubele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chemical Reviews* 2019, 119 (18), 10691–10717. Copyright 2019 American Chemical Society.

stabilize more compact (folded) structures.^{22,25} However, the interior of the cell also offers electrostatic and hydrophobic interactions mediated by water, ions, and other macromolecules which can destabilize proteins²⁶ and interfere with their diffusion around the cell.^{27–29} These interactions are referred to collectively as ‘sticking’ and are cited as the cause of observed anomalous diffusion inside cells. A number of *in vitro* studies of proteins using inert crowders and cell lysate have shown that the interplay of both repulsive (crowding) and attractive (sticking) interactions affects proteins in-cell.^{30–34} When the weak, transient interactions caused by sticking have evolved to benefit the organism, they are referred to as quinary structure.^{21,35} For a more thorough review of crowding, sticking, and quinary interactions and their effects on protein folding, see ref. 23.

Effect of Cellular Environment on Protein Folding, Binding and Dynamics

Experimental techniques to study protein folding in cells started to develop in the early 2000s. The first quantitative method to compare protein stability *in vitro* and inside cytoplasm of *E. coli* cells using mass spectroscopy was reported by the Oas group in 2001.³⁶ Such early techniques were destructive to the cell by extracting its contents prior to measurement. However, with improved *in vivo* fluorescent labeling, protein stability and aggregation could be observed directly in live bacterial cells.³⁷ Fluorescence imaging further emerged as a powerful technique to study protein folding stability and dynamics directly in live mammalian cells,³⁸ enabling the observation of protein stability and folding kinetics even of single molecules.³⁹

From the earliest experimental studies of in-cell protein folding, differences between *in vitro* and *in vivo* protein stability and folding kinetics have been observed.^{29,36–38,40–43} Proteins are both stabilized and destabilized *in vivo*,^{20,31} and studies have shown that both crowding and sticking interactions affect stability changes.^{31,33} Further, different features of proteins, including charged surface residues and protein shape, can make proteins more sensitive to crowding and sticking.^{44,45} Importantly, these changes vary across a single organism: in-cell studies of protein folding have shown that protein folding stability and kinetics can be affected by cell type, cell cycle, and cell stress.^{46–48}

Interactions with neighboring macromolecules also greatly alter a protein’s energy landscape. Experiments have shown that crowded conditions can impact proteins’ folded and unfolded conformational ensembles *in vivo* compared to crystal or solution structures.^{32,40,49–52} These effects on conformational ensembles are particularly important in light of the recent

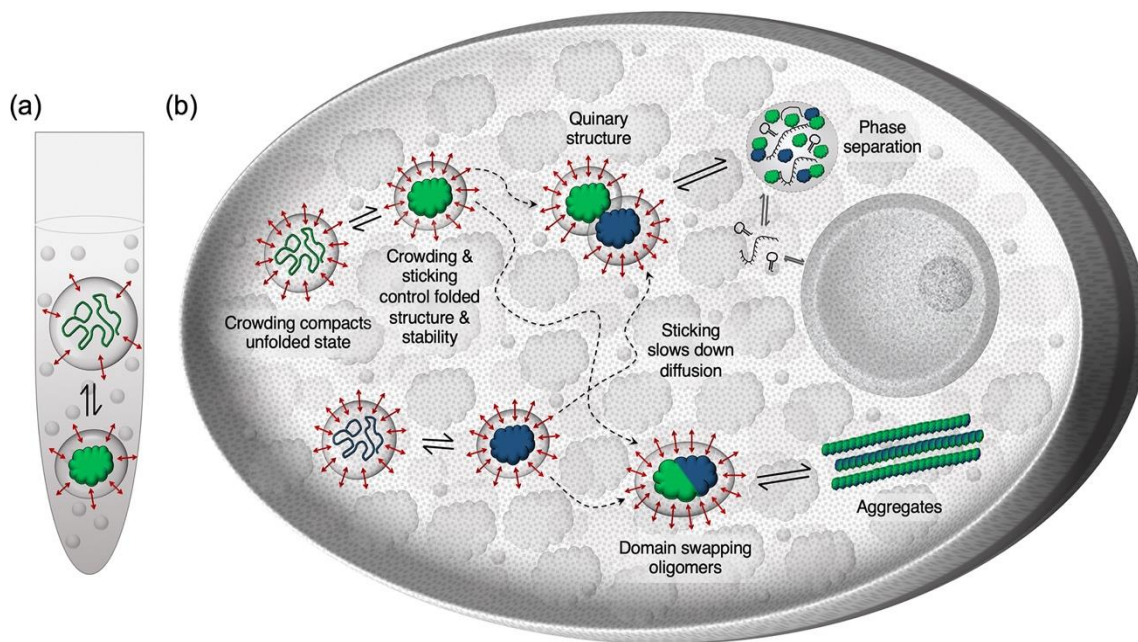


Figure 1.3: The complexity and impact of cellular environment on protein structure, stability and function are difficult to reproduce in a test tube. Figure summarizes the factors in play based on a few experimental results of interest. (a) *In vitro* experiments of proteins in dilute aqueous buffer solutions or with inert crowders often neglect the non-steric interactions between macromolecules and variability in composition inside cells.^{22,37} (b) Effects of macromolecular crowding and sticking forces on compactness of unfolded state,⁵⁸ structure, and stability of folded state^{36,59,60} and diffusion of a protein inside the cell to reach the required locations or in search of binding partners.²⁰ Protein-protein interactions can result in functional (quinary) or non-functional complexes/oligomers, whose shape determines the stabilization by crowding.⁶² Quinary structures allow for multivalent interactions, extreme cases leading to phase separation in the crowded cellular environment.⁶⁶ Figure prepared by Gopika Gopan, published in ref. 1.

emphasis on cryptic pockets—previously not observed drug targets that only transiently become accessible.⁵³

In addition to showing the cell's influence on protein stability and folding kinetics, in-cell experiments have shown that environment impacts protein dynamics and function. A recent study has shown that the chaperone Hsp70 relies heavily on the cellular environment to function, and only non-specifically sticks to clients *in vitro*.⁵⁴ Like folding reactions, the cellular environment can both stabilize and destabilize binding reactions between macromolecules.^{55,56}

Experimental studies also indicate a high degree of evolved modulation of protein dynamics by the cellular environment. In 2017, Oliveberg and coworkers, using in-cell NMR spectrometry, observed a marked decrease of rotational diffusion of human proteins expressed in the *E. coli* cytoplasm compared to their bacterial homologs, indicating the human proteins were sticking badly to their bacterial neighbors.⁵⁷ However, mutating a single charged residue on the

surface of the protein allowed the rotational diffusion to recover enough to produce an NMR signal, validating earlier studies that showed that electrostatic mutations affect the tumbling of a protein in-cell.^{57,58} Intriguingly, a later study by the same group found that the opposite did not necessarily apply—when the same proteins were studied in human cells, the bacterial proteins still showed higher mobility in the human cell than the human proteins in their native environments.²⁷ It is hypothesized that this effect occurs as a result of an evolved electrostatic repulsion throughout the *E. coli* proteome to minimize protein-protein sticking in the more crowded bacterial intracellular environment.^{27,57} Additionally, phenomena such as fold-switching proteins,^{59,60} disorder-to-order conformational changes in intrinsically disordered proteins,^{19,61} and the formation of membraneless organelles^{62,63} are modulated by local conditions inside the cell (Figure 1.3).

Taken together, these experimental studies indicate that evolution has finely tuned the interactions between a protein and its local environment to optimize function and prevent aggregation. However, measurement limits in spatial resolution and timescale make it difficult to determine the exact atomistic mechanisms of these phenomena experimentally (Figure 1.4). Therefore, computational studies of proteins *in vivo* are needed to capture the impact of the cellular environment on macromolecules.

THEORETICAL AND COMPUTATIONAL STUDIES OF IN-CELL PROTEIN DYNAMICS

History of Simulating Biomolecules in vivo

Theoretical and computational efforts to determine the effect of the cellular environment on proteins began in the 1980s by exploring the effects of crowding²⁵ and macromolecular sticking⁶⁴ on protein folding and protein-protein interactions. Early computational efforts to simulate proteins *in vivo* modeled macromolecules as hard spheres and focused on the diffusion and neighboring interactions of these particles.^{65,66} In 2010, McGuffee and Elcock simulated a sizeable segment of the *E. coli* cytoplasm using Brownian dynamics.⁶⁷ Following their work, coarse-grained cellular models and all-atom models of crowded protein solutions were developed.^{68,69} In 2016, Sugita and Feig used all-atom molecular dynamics to evaluate protein stability and protein-protein interactions in a large (>100 million atom) model of a bacterial cytoplasm.⁷⁰ They found that ensembles are subject to a high degree of copy-to-copy variation, underscoring the importance of local environment on in-cell protein structure.⁷⁰

Since this study, the field of in-cell protein dynamics has continued to grow.^{28,71–75} Recently, coarse-grained simulations of proteins in the presence of inert crowders demonstrate the impact of crowding on structural ensembles of both folded and unfolded proteins.^{51,76} Additionally, ‘whole cell’ simulation techniques have been developed that merge coarse-grained real-space dynamics on a lattice with reaction networks for small solutes and chemical reactions that cannot be handled on the lattice.⁷⁷ This work is being extended from bacterial to eukaryotic cells.⁷⁸ As coarse-graining approaches the all-atom scale, it remains important to include reaction networks. Quantum mechanical treatments of reactions in the cell are the gold standard, but more efficient methods to combine weak interactions with chemistry (e.g. bond breaking and making in metabolism or replication) will be needed to bring a cell to life on the computer.

Current Challenges

A major challenge of computational studies of crowded environments is reaching the microsecond or longer timescales necessary to fold a protein or to bring two proteins together. This is particularly challenging because early results point to slower folding in cells due to more restricted conformational dynamics.⁷² Prior to the project discussed in this dissertation, which was

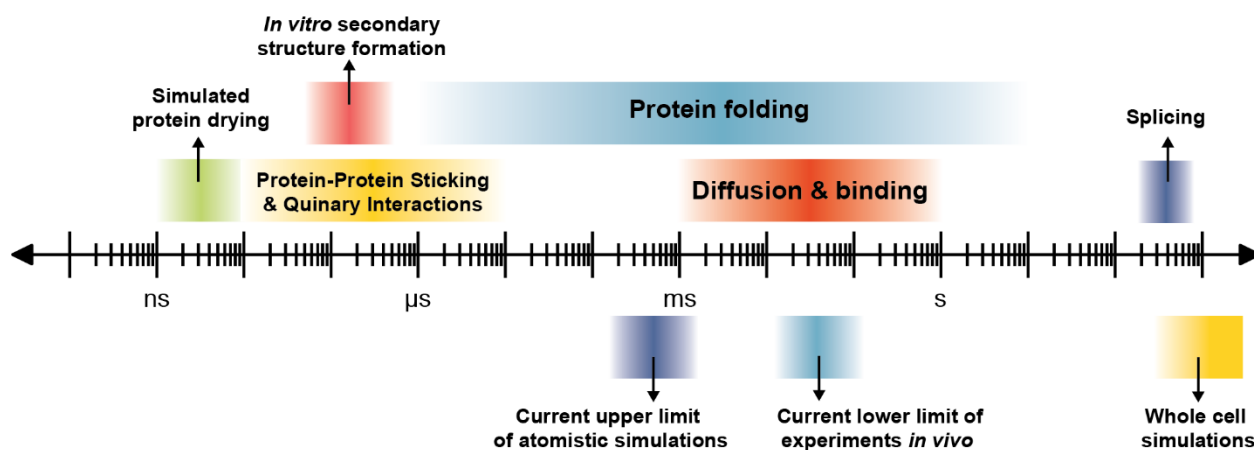


Figure 1.4: Timescales of notable biological processes (top)^{13,43,44,67} and experimental and computational methods (bottom)^{25,28,29,65}. To converge, the timescales of these techniques must overlap with each other and the process of interest. Figure published in ref. 1.

started by Yi Zhang,⁷⁹ no atomistic simulation of proteins in crowded conditions had reached the microsecond timescale. To date, the models discussed here have been simulated for over 250 μs. However, despite simulating for much longer than necessary in water-box calculations,^{73,80} no complete folding events have been observed in our cytoplasm models. Recently, we simulated multiple unfolded copies of a fast-folding WW domain for over 220 μs and observed only partial

folding (see Chapter 4); instead, persistent non-native structure or even intermolecular structures were formed.⁷³

Multiple simulations of proteins in cell-like environments have observed excessive nonspecific protein-protein interactions and clustering (see Chapter 5).^{28,71,72,74} Due to the chemical nature of protein-protein sticking, a number of forces can contribute to these interactions. Current experimental evidence emphasizes the importance of electrostatic interactions in protein-protein sticking.^{33,57} Computational results also point to the importance of hydrophobic patches. While all-atom studies of concentrated protein solutions do observe protein-protein contact formation between charged residues,^{74,75} other studies revealed both electrostatic and hydrophobic surface interactions in a model of the cytoplasm,^{72,73} or in dense homogeneous protein solutions.⁷¹

The lack of protein folding and excessive protein-protein sticking observed in models of proteins in various crowded environments are strong indicators that the parameters used in the

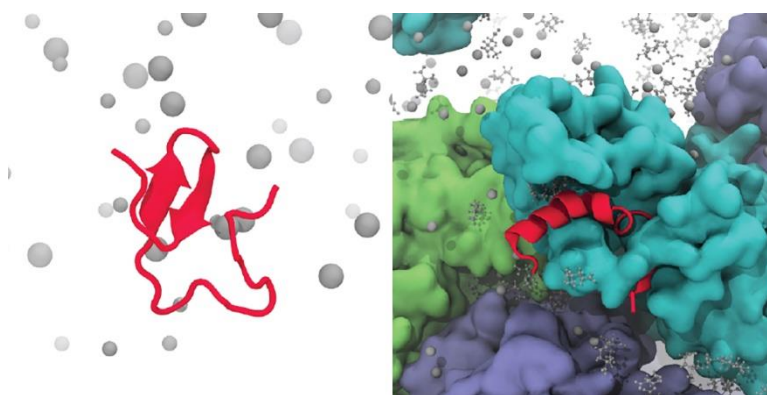


Figure 1.5: Classical MD force fields' tendency to stabilize protein-protein interactions makes computationally modeling crowded environments challenging. (a) While a WW domain (red) folds easily in a simulation of a waterbox with ions (gray spheres represent Na^+ , Cl^-), (b) excessive sticking interactions between surrounding macromolecules and metabolites cause long-lived misfolded states to appear in crowded simulations. The WW domain (in red) has formed non-native helical structure.⁵¹ Figure published in ref. 1.

models, collectively referred to as force fields, introduce bias to the simulations. Current force fields tend to over-stabilize protein-protein sticking.⁸¹ To better simulate these environments, a number of modifications have been made to reduce sticking to the right level and better replicate experimental diffusion coefficients: using the TIP4P-D water model (in conjunction with the Amber force field),^{71,82} modifying the strength of protein-water interactions,⁷⁴ and using NBFIX corrections.⁷² Despite these improvements, the comparison of computed and experimentally observed protein-protein sticking demonstrates that force fields developed to fold single proteins *in vitro* still need refinement to work for the cytoplasm (Figure 1.5).

TOWARDS A MATCH

As both experiments and simulations of proteins *in vivo* continue to develop, a number of questions remain. To answer these questions, the Gruebele and Pogorelov groups have developed a number of atomistic models of proteins in cell-like environments to further elucidate the effect of the cellular environment on biomolecules.

In Chapter 2, I describe the design, construction, and simulation of these models, which realistically represent small sections of the *E. coli* and *H. sapiens* cytoplasm with atomistic resolution. The *E. coli* model simulated with the CHARMM 22* force field⁸³ was the first reported atomistic simulation of proteins in crowded environments to reach the microsecond timescale.⁷² In Chapters 3-5, I analyze the results from the *E. coli* simulations and address how the cellular environment affects metabolite dynamics, protein folding,⁷³ and nonspecific protein-protein sticking.⁷² In Chapter 6, I compare results from the *E. coli* simulations with a similar model designed to mimic the conditions of the human cytoplasm to determine how different species evolved to modulate their cellular environments. Finally, in Chapter 7, I use data sonification to describe protein folding reactions and analyze protein transition states in a project with musician collaborators from Urbana-Champaign and abroad.

REFERENCES

- (1) Gopan, G.; Gruebele, M.; Rickard, M. In-Cell Protein Landscapes: Making the Match between Theory, Simulation and Experiment. *Curr. Opin. Struct. Biol.* **2021**, *66*, 163–169. <https://doi.org/10.1016/j.sbi.2020.10.013>.
- (2) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by x-Ray Analysis. *Nature* **1958**, *181* (4610), 662–666. <https://doi.org/10.1038/181662a0>.
- (3) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230.
- (4) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* **1975**, *253* (5494), 694–698. <https://doi.org/10.1038/253694a0>.
- (5) Taketomi, H.; Ueda, Y.; Gō, N. Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. *Int. J. Pept. Protein Res.* **1975**, *7* (6), 445–459. <https://doi.org/10.1111/j.1399-3011.1975.tb02465.x>.
- (6) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels. *Nat. Struct. Biol.* **1997**, *4* (1), 10–19. <https://doi.org/10.1038/nsb0197-10>.
- (7) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins Struct. Funct. Bioinforma.* **1995**, *21* (3), 167–195. <https://doi.org/10.1002/prot.340210302>.
- (8) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998**, *282* (5389), 740–744. <https://doi.org/10.1126/science.282.5389.740>.
- (9) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. Absolute Comparison of Simulated and Experimental Protein-Folding Dynamics. *Nature* **2002**, *420* (6911), 102–106. <https://doi.org/10.1038/nature01160>.
- (10) Kresten Lindorff-Larsen; Stefano Piana; Ron O. Dror; David E. Shaw. How Fast-Folding Proteins Fold | Science. *Science* **2011**, *334* (6055), 517–520.
- (11) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. <https://doi.org/10.1016/j.neuron.2018.08.011>.
- (12) Friedrichs, M. S.; Wolynes, P. G. Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians. *Science* **1989**, *246* (4928), 371–373. <https://doi.org/10.1126/science.246.4928.371>.
- (13) Jin, S.; Contessoto, V. G.; Chen, M.; Schafer, N. P.; Lu, W.; Chen, X.; Bueno, C.; Hajitaheri, A.; Sirovetz, B. J.; Davtyan, A.; Papoian, G. A.; Tsai, M.-Y.; Wolynes, P. G. AWSEM-Suite: A Protein Structure Prediction Server Based on Template-Guided, Coevolutionary-Enhanced Optimized Folding Landscapes. *Nucleic Acids Res.* **2020**, *48* (W1), W25–W30. <https://doi.org/10.1093/nar/gkaa356>.
- (14) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. Edited by F. E. Cohen. *J. Mol. Biol.* **1997**, *268* (1), 209–225. <https://doi.org/10.1006/jmbi.1997.0959>.

- (15) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>.
- (16) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V.; Baker, D. Structure Prediction for CASP8 with All-Atom Refinement Using Rosetta. *Proteins* **2009**, *77* (0 9), 89–99. <https://doi.org/10.1002/prot.22540>.
- (17) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (18) Powers, E. T.; Gierasch, L. M. The Proteome Folding Problem and Cellular Proteostasis. *J. Mol. Biol.* **2021**, *433* (20), 167197. <https://doi.org/10.1016/j.jmb.2021.167197>.
- (19) Theillet, F.-X.; Binolfi, A.; Frembgen-Kesner, T.; Hingorani, K.; Sarkar, M.; Kyne, C.; Li, C.; Crowley, P. B.; Gierasch, L.; Pielak, G. J.; Elcock, A. H.; Gershenson, A.; Selenko, P. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114* (13), 6661–6714. <https://doi.org/10.1021/cr400695p>.
- (20) Gruebele, M.; Dave, K.; Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *Annu. Rev. Biophys.* **2016**, *45* (1), 233–251. <https://doi.org/10.1146/annurev-biophys-062215-011236>.
- (21) Wirth, A. J.; Gruebele, M. Quinary Protein Structure and the Consequences of Crowding in Living Cells: Leaving the Test-Tube Behind. *BioEssays* **2013**, *35* (11), 984–993. <https://doi.org/10.1002/bies.201300080>.
- (22) Rivas, G.; Minton, A. P. Toward an Understanding of Biochemical Equilibria within Living Cells. *Biophys. Rev.* **2017**, *10* (2), 241–253. <https://doi.org/10.1007/s12551-017-0347-6>.
- (23) Guin, D.; Gruebele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chem. Rev.* **2019**, *119* (18), 10691–10717. <https://doi.org/10.1021/acs.chemrev.8b00753>.
- (24) Cohen, R. D.; Pielak, G. J. A Cell Is More than the Sum of Its (Dilute) Parts: A Brief History of Quinary Structure. *Protein Sci. Publ. Protein Soc.* **2017**, *26* (3), 403–413. <https://doi.org/10.1002/pro.3092>.
- (25) Minton, A. P. Excluded Volume as a Determinant of Macromolecular Structure and Reactivity. *Biopolymers* **1981**, *20* (10), 2093–2120. <https://doi.org/10.1002/bip.1981.360201006>.
- (26) Wennerström, H.; Estrada, E. V.; Danielsson, J.; Oliveberg, M. Colloidal Stability of the Living Cell. *Proc. Natl. Acad. Sci.* **2020**. <https://doi.org/10.1073/pnas.1914599117>.
- (27) Leeb, S.; Sörensen, T.; Yang, F.; Mu, X.; Oliveberg, M.; Danielsson, J. Diffusive Protein Interactions in Human versus Bacterial Cells. *Curr. Res. Struct. Biol.* **2020**. <https://doi.org/10.1016/j.crstbi.2020.04.002>.

- (28) Nawrocki, G.; Karaboga, A.; Sugita, Y.; Feig, M. Effect of Protein–Protein Interactions and Solvent Viscosity on the Rotational Diffusion of Proteins in Crowded Environments. *Phys. Chem. Chem. Phys.* **2019**, *21* (2), 876–883. <https://doi.org/10.1039/C8CP06142D>.
- (29) Guo, M.; Gelman, H.; Gruebele, M. Coupled Protein Diffusion and Folding in the Cell. *PLoS One* **2014**, *9* (12), e113040. <https://doi.org/10.1371/journal.pone.0113040>.
- (30) Davis, C. M.; Deutsch, J.; Gruebele, M. An in Vitro Mimic of In-Cell Solvation for Protein Folding Studies. *Protein Sci.* **2020**, *29* (4), 1060–1068. <https://doi.org/10.1002/pro.3833>.
- (31) Caitlin M. Davis; Martin Gruebele. Non-Steric Interactions Predict the Trend and Steric Interactions the Offset of Protein Stability in Cells. *ChemPhysChem* **19** (18), 2290–2294. <https://doi.org/10.1002/cphc.201800534>.
- (32) Dhar, A.; Samiotakis, A.; Ebbinghaus, S.; Nienhaus, L.; Homouz, D.; Gruebele, M.; Cheung, M. S. Structure, Function, and Folding of Phosphoglycerate Kinase Are Strongly Perturbed by Macromolecular Crowding. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (41), 17586–17591. <https://doi.org/10.1073/pnas.1006760107>.
- (33) Cohen, R. D.; Pielak, G. J. Electrostatic Contributions to Protein Quinary Structure. *J. Am. Chem. Soc.* **2016**, *138* (40), 13139–13142. <https://doi.org/10.1021/jacs.6b07323>.
- (34) Cheng, K.; Wu, Q.; Zhang, Z.; Pielak, G. J.; Liu, M.; Li, C. Crowding and Confinement Can Oppositely Affect Protein Stability. *ChemPhysChem* **2018**, *19* (24), 3350–3355. <https://doi.org/10.1002/cphc.201800857>.
- (35) McConkey, E. H. Molecular Evolution, Intracellular Organization, and the Quinary Structure of Proteins. *Proc. Natl. Acad. Sci.* **1982**, *79* (10), 3236–3240. <https://doi.org/10.1073/pnas.79.10.3236>.
- (36) Ghaemmamghami, S.; Oas, T. G. Quantitative Protein Stability Measurement in Vivo. *Nat. Struct. Biol.* **2001**, *8* (10), 879–882. <https://doi.org/10.1038/nsb1001-879>.
- (37) Ignatova, Z.; Gierasch, L. M. Monitoring Protein Stability and Aggregation in Vivo by Real-Time Fluorescent Labeling. *Proc. Natl. Acad. Sci.* **2004**, *101* (2), 523–528. <https://doi.org/10.1073/pnas.0304533101>.
- (38) Ebbinghaus, S.; Dhar, A.; McDonald, J. D.; Gruebele, M. Protein Folding Stability and Dynamics Imaged in a Living Cell. *Nat. Methods* **2010**, *7* (4), 319–323. <https://doi.org/10.1038/nmeth.1435>.
- (39) König, I.; Zarrine-Afsar, A.; Aznauryan, M.; Soranno, A.; Wunderlich, B.; Dingfelder, F.; Stüber, J. C.; Plückthun, A.; Nettels, D.; Schuler, B. Single-Molecule Spectroscopy of Protein Conformational Dynamics in Live Eukaryotic Cells. *Nat. Methods* **2015**, *12* (8), 773–779. <https://doi.org/10.1038/nmeth.3475>.
- (40) Wang, Y.; Sukenik, S.; Davis, C. M.; Gruebele, M. Cell Volume Controls Protein Stability and Compactness of the Unfolded State. *J. Phys. Chem. B* **2018**, *122* (49), 11762–11770. <https://doi.org/10.1021/acs.jpcb.8b08216>.
- (41) Miklos, A. C.; Sarkar, M.; Wang, Y.; Pielak, G. J. Protein Crowding Tunes Protein Stability. *J. Am. Chem. Soc.* **2011**, *133* (18), 7116–7120. <https://doi.org/10.1021/ja200067p>.
- (42) Guzman, I.; Gelman, H.; Tai, J.; Gruebele, M. The Extracellular Protein VlsE Is Destabilized inside Cells. *J. Mol. Biol.* **2014**, *426* (1), 11–20. <https://doi.org/10.1016/j.jmb.2013.08.024>.
- (43) Danielsson, J.; Mu, X.; Lang, L.; Wang, H.; Binolfi, A.; Theillet, F.-X.; Bekei, B.; Logan, D. T.; Selenko, P.; Wennerström, H.; Oliveberg, M. Thermodynamics of Protein Destabilization in Live Cells. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (40), 12402–12407. <https://doi.org/10.1073/pnas.1511308112>.

- (44) Guseman, A. J.; Goncalves, G. M. P.; Speer, S. L.; Young, G. B.; Pielak, G. J. Protein Shape Modulates Crowding Effects. *Proc. Natl. Acad. Sci.* **2018**, *115* (43), 10965–10970. <https://doi.org/10.1073/pnas.1810054115>.
- (45) Smith, A. E.; Zhou, L. Z.; Gorenssek, A. H.; Senske, M.; Pielak, G. J. In-Cell Thermodynamics and a New Role for Protein Surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (7), 1725–1730. <https://doi.org/10.1073/PNAS.1518620113>.
- (46) Feng, R.; Gruebele, M.; Davis, C. M. Quantifying Protein Dynamics and Stability in a Living Organism. *Nat. Commun.* **2019**, *10* (1), 1–7. <https://doi.org/10.1038/s41467-019-09088-y>.
- (47) Gnutt, D.; Sistemich, L.; Ebbinghaus, S. Protein Folding Modulation in Cells Subject to Differentiation and Stress. *Front. Mol. Biosci.* **2019**, *6*. <https://doi.org/10.3389/fmolb.2019.00038>.
- (48) Wirth, A. J.; Platkov, M.; Gruebele, M. Temporal Variation of a Protein Folding Energy Landscape in the Cell. *J. Am. Chem. Soc.* **2013**, *135* (51), 19215–19221. <https://doi.org/10.1021/ja4087165>.
- (49) Yang, Y.; Chen, S.-N.; Yang, F.; Li, X.-Y.; Feintuch, A.; Su, X.-C.; Goldfarb, D. In-Cell Destabilization of a Homo-Dimeric Protein Complex Detected by DEER Spectroscopy. *bioRxiv* **2020**, 2020.03.27.011510. <https://doi.org/10.1101/2020.03.27.011510>.
- (50) Tanaka, T.; Ikeya, T.; Kamoshida, H.; Suemoto, Y.; Mishima, M.; Shirakawa, M.; Güntert, P.; Ito, Y. High-Resolution Protein 3D Structure Determination in Living Eukaryotic Cells. *Angew. Chem. Int. Ed.* **2019**, *58* (22), 7284–7288. <https://doi.org/10.1002/anie.201900840>.
- (51) Gasic, A. G.; Boob, M. M.; Prigozhin, M. B.; Homouz, D.; Daugherty, C. M.; Gruebele, M.; Cheung, M. S. Critical Phenomena in the Temperature-Pressure-Crowding Phase Diagram of a Protein. *Phys. Rev. X* **2019**, *9* (4), 041035. <https://doi.org/10.1103/PhysRevX.9.041035>.
- (52) Dalaloyan, A.; Martorana, A.; Barak, Y.; Gataulin, D.; Reuveny, E.; Howe, A.; Elbaum, M.; Albeck, S.; Unger, T.; Frydman, V.; Abdelkader, E. H.; Otting, G.; Goldfarb, D. Tracking Conformational Changes in Calmodulin in Vitro, in Cell Extract, and in Cells by Electron Paramagnetic Resonance Distance Measurements. *ChemPhysChem* **2019**, *20* (14), 1860–1868. <https://doi.org/10.1002/cphc.201900341>.
- (53) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc. Chem. Res.* **2020**, *53* (3), 654–661. <https://doi.org/10.1021/acs.accounts.9b00613>.
- (54) Guin, D.; Gelman, H.; Wang, Y.; Gruebele, M. Heat Shock-Induced Chaperoning by Hsp70 Is Enabled in-Cell. *PLoS ONE* **2019**, *14* (9). <https://doi.org/10.1371/journal.pone.0222990>.
- (55) Davis, C. M.; Gruebele, M. Cellular Sticking Can Strongly Reduce Complex Binding by Speeding Dissociation. *J. Phys. Chem. B* **2021**, *125* (15), 3815–3823. <https://doi.org/10.1021/acs.jpcc.1c00950>.
- (56) Guseman, A. J.; Speer, S. L.; Perez Goncalves, G. M.; Pielak, G. J. Surface Charge Modulates Protein–Protein Interactions in Physiologically Relevant Environments. *Biochemistry* **2018**, *57* (11), 1681–1684. <https://doi.org/10.1021/acs.biochem.8b00061>.
- (57) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in Escherichia Coli. *Proc. Natl. Acad. Sci.* **2017**, *114* (23), E4556–E4563. <https://doi.org/10.1073/pnas.1621227114>.
- (58) Barbieri, L.; Luchinat, E.; Banci, L. Protein Interaction Patterns in Different Cellular Environments Are Revealed by In-Cell NMR. *Sci. Rep.* **2015**, *5*, 14456. <https://doi.org/10.1038/srep14456>.

- (59) Porter, L. L.; Looger, L. L. Extant Fold-Switching Proteins Are Widespread. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (23), 5968–5973. <https://doi.org/10.1073/pnas.1800168115>.
- (60) Zuber, P. K.; Schweimer, K.; Rösch, P.; Artsimovitch, I.; Knauer, S. H. Reversible Fold-Switching Controls the Functional Cycle of the Antitermination Factor RfaH. *Nat. Commun.* **2019**, *10* (1), 702–702. <https://doi.org/10.1038/s41467-019-08567-6>.
- (61) Fonin, A. V.; Darling, A. L.; Kuznetsova, I. M.; Turoverov, K. K.; Uversky, V. N. Intrinsically Disordered Proteins in Crowded Milieu: When Chaos Prevails within the Cellular Gumbo. *Cell. Mol. Life Sci.* **2018**, *75* (21), 3907–3929. <https://doi.org/10.1007/s00018-018-2894-9>.
- (62) Schuster, B. S.; Reed, E. H.; Parthasarathy, R.; Jahnke, C. N.; Caldwell, R. M.; Bermudez, J. G.; Ramage, H.; Good, M. C.; Hammer, D. A. Controllable Protein Phase Separation and Modular Recruitment to Form Responsive Membraneless Organelles. *Nat. Commun.* **2018**, *9* (1), 1–12. <https://doi.org/10.1038/s41467-018-05403-1>.
- (63) Lin, Y.; Protter, D. S. W.; Rosen, M. K.; Parker, R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol. Cell* **2015**, *60* (2), 208–219. <https://doi.org/10.1016/j.molcel.2015.08.018>.
- (64) Muramatsu, N.; Minton, A. P. Tracer Diffusion of Globular Proteins in Concentrated Protein Solutions. *Proc. Natl. Acad. Sci.* **1988**, *85* (9), 2984–2988. <https://doi.org/10.1073/pnas.85.9.2984>.
- (65) Bicout, D. J.; Field, M. J. Stochastic Dynamics Simulations of Macromolecular Diffusion in a Model of the Cytoplasm of Escherichia Coli. *J. Phys. Chem.* **1996**, *100* (7), 2489–2497. <https://doi.org/10.1021/jp9525191>.
- (66) Ridgway, D.; Broderick, G.; Lopez-Campistrous, A.; Ru'aini, M.; Winter, P.; Hamilton, M.; Boulanger, P.; Kovalenko, A.; Ellison, M. J. Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm. *Biophys. J.* **2008**, *94* (10), 3748–3759. <https://doi.org/10.1529/biophysj.107.116053>.
- (67) McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput. Biol.* **2010**, *6* (3).
- (68) Wang, Q.; Cheung, M. S. A Physics-Based Approach of Coarse-Graining the Cytoplasm of Escherichia Coli (CGCYTO). *Biophys. J.* **2012**, *102* (10), 2353–2361. <https://doi.org/10.1016/j.bpj.2012.04.010>.
- (69) Cossins, B. P.; Jacobson, M. P.; Guallar, V. A New View of the Bacterial Cytosol Environment. *PLoS Comput. Biol.* **2011**, *7* (6). <https://doi.org/10.1371/journal.pcbi.1002066>.
- (70) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm. *eLife* **2016**, *5*, e19274. <https://doi.org/10.7554/eLife.19274>.
- (71) Bülow, S. von; Siggel, M.; Linke, M.; Hummer, G. Dynamic Cluster Formation Determines Viscosity and Diffusion in Dense Protein Solutions. *Proc. Natl. Acad. Sci.* **2019**, *116* (20), 9843–9852. <https://doi.org/10.1073/pnas.1817564116>.
- (72) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein–Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (73) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740.

- (74) Nawrocki, G.; Wang, P.; Yu, I.; Sugita, Y.; Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. *J. Phys. Chem. B* **2017**, *121* (49), 11072–11084. <https://doi.org/10.1021/acs.jpcb.7b08785>.
- (75) Nawrocki, G.; Im, W.; Sugita, Y.; Feig, M. Clustering and Dynamics of Crowded Proteins near Membranes and Their Influence on Membrane Bending. *Proc. Natl. Acad. Sci.* **2019**, *116* (49), 24562–24567. <https://doi.org/10.1073/pnas.1910771116>.
- (76) Zagarra, F. C.; Homouz, D.; Gasic, A. G.; Babel, L.; Kovermann, M.; Wittung-Stafshede, P.; Cheung, M. S. Crowding-Induced Elongated Conformation of Urea-Unfolded Apoazurin: Investigating the Role of Crowder Shape in Silico. *J. Phys. Chem. B* **2019**, *123* (17), 3607–3617. <https://doi.org/10.1021/acs.jpcb.9b00782>.
- (77) Chew, W.-X.; Kaizu, K.; Watabe, M.; Muniandy, S. V.; Takahashi, K.; Arjunan, S. N. V. Reaction-Diffusion Kinetics on Lattice at the Microscopic Scale. *Phys. Rev. E* **2018**, *98* (3), 032418. <https://doi.org/10.1103/PhysRevE.98.032418>.
- (78) Ghaemi, Z.; Peterson, J. R.; Gruebele, M.; Luthey-Schulten, Z. An In-Silico Human Cell Model Reveals the Influence of Spatial Organization on RNA Splicing. *PLOS Comput. Biol.* **2020**, *16* (3), e1007717. <https://doi.org/10.1371/journal.pcbi.1007717>.
- (79) Zhang, Y. Frustration of Protein Folding from in Vitro to in Vivo, University of Illinois at Urbana-Champaign, Urbana, IL, 2018.
- (80) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β -Sheet Protein. *J. Mol. Biol.* **2011**, *405* (1), 43–48. <https://doi.org/10.1016/j.jmb.2010.10.023>.
- (81) Petrov, D.; Zagrovic, B. Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Comput. Biol.* **2014**, *10* (5), e1003638–e1003638. <https://doi.org/10.1371/journal.pcbi.1003638>.
- (82) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950–1958. <https://doi.org/10.1002/prot.22711>.
- (83) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–49. <https://doi.org/10.1016/j.bpj.2011.03.051>.

CHAPTER 2: METHODS TO CONSTRUCT MODELS OF THE *E. COLI* AND HUMAN CYTOPLASMS *IN SILICO*

INTRODUCTION

Molecular Dynamics

Molecular dynamics (MD) is a classical mechanical computational method that has been used to simulate the motions of bio-macromolecules since 1977.³ The fundamental principle of atomistic MD is the classical treatment of molecules. Forces on individual atoms are calculated by summing the various energetic interactions the atoms can have (bonded interactions, such as a bond stretch or an atom rotating around a dihedral angle, and nonbonded interactions, the electrostatic and Lennard-Jones interactions), shown in Equation 2.1. Once forces are calculated, Newton's second law $F = ma$ relates the force to acceleration, and thus, to that atom's position at the next timestep. The movements of electrons are done away with entirely and are treated as partial point charges located at the nuclei of their respective atoms. This assumption allows molecular movements to be simulated classically, without quantum effects, at a great reduction of computational cost. Because electron dynamics is not included explicitly and harmonic potentials

$$V = \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1 + \cos(n\phi - \delta)] + \sum_{\text{impropers}} k_\omega(\omega - \omega_0)^2 + \sum_{\text{nonbonded}} \epsilon \left[\left(\frac{R_{\text{minij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{\text{minij}}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon r_{ij}} \quad (2.1 \text{Equation 2.1})$$

Equation 2.1: Potential energy function for CHARMM force field. Bonds, angles, and impropers are described as springs. Non-bonded terms are divided into Coulombic and Lennard-Jones (van der Waals) approximations. Equation modified from ref. 4.

* Some writing and figures in this chapter are reproduced in part with permission from:

Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, 10 (18), 5667–5673.¹ Copyright American Chemical Society 2019.

Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, 124 (23), 4732–4740.² Copyright American Chemical Society 2020.

are used, bonds cannot be made or broken, meaning no chemical reactions occur during the simulations.

Force Fields

In order to perform classical MD simulations, one must develop a force field, an approximation for the quantum mechanical Born-Oppenheimer potential energy surface.⁴ Force fields consist of a variety of parameters that must be provided to the software, or MD “engine,” of choice. For example, to determine the forces on a given bond in Equation 2.1, both k_b (the strength of the bond) and b_0 (the ideal length of the bond) must be known, among similar parameters for all bonded interactions. Additionally, the non-bonded calculations require that the van der Waals radii, Lennard-Jones well depth, and atomic partial charges are provided.

A variety of atomistic MD force fields have been developed and are currently available for use, including CHARMM, AMBER, OPLS, and GROMOS.^{5–9} Additionally, these force fields are frequently updated to more accurately represent these parameters and correct any identified biases. For example, while previous force field versions overstabilized the α -helices of folding proteins,^{10,11} current force fields fold globular proteins in test-tube like conditions well.¹² Recently, force field development has turned to mediating the energetics of protein-protein contacts and charge-charge interactions to better replicate the collapse of IDP's.^{13–16}

A critical aspect for accurately describing these models here and throughout is the “stickiness” of MD force fields. Even with continuing improvements, current MD force fields tend to be too sticky—that is, the force fields favor protein-protein contacts over protein-solvent contacts.^{15,17} This is particularly important during simulations of proteins in crowded environments like the cell—the force fields may overly stabilize interactions between neighboring macromolecules, rather than other aspects of the local environment.^{1,17,18}

A frequently used technique to mediate the stickiness of atomistic force fields is the use of non-bonded fix (NBFIX) corrections.¹⁹ NBFIX corrections modify the Lennard-Jones parameters²⁰ of specific atom pairs to “surgically” alter the attraction between those atoms. NBFIX corrections can be implemented in a number of ways. For example, in 2019, the Sugita and Feig groups increased the Lennard-Jones well depth (ϵ , Equation 2.1) of all protein-water atom pairs by a factor of 1.09.¹⁸ In contrast, in 2018, Aksimentiev and coworkers developed a set of NBFIX corrections that lessened the attraction between positively- and negatively-charged functional groups and ions (this set is known as Champaign-Urbana fix, or CUFIX).²¹

All simulations discussed in this thesis were performed using different versions of the CHARMM force field: CHARMM 22* (C22*),²² which was originally developed to fold small proteins in test-tube like conditions, CHARMM 36m (C36m),²³ which was developed to accurately replicate the collapse of both globular and disordered proteins, and CHARMM 36m with CUFIX corrections (C36mCU).^{21,23}

The CHARMM force field was initially selected due to its compatibility with NAMD,²⁴ the software used to minimize and pre-equilibrate the systems, and to compare more closely to existing simulations of the *M. genitalium* cytoplasm.²⁵ Since the start of this project, the field has continued to develop. Simulations of crowded protein environments have indicated that a combination of the TIP4P-D water model and the Amber99SB*-ILDN-Q force field may better mediate protein-protein sticking than CHARMM force fields and TIP3P water.^{26–31} Additionally, over the last five years, multiple versions of Amber designed to mediate protein-protein sticking and IDP collapse have been developed.^{14,32} Given these results, in the future, using Amber force fields in the *E. coli* and *H. sapiens* models described here may be more appropriate.

Brownian Dynamics

Like MD, Brownian Dynamics (BD) is a classical simulation technique that has been used to simulate the motions of biomolecules for decades.³³ Unlike MD, where molecules are described as a series of interconnected springs and can be simulated with explicit or implicit solvent, BD focuses on the interactions a solute has with an implicit solvent. Therefore, a solute (in this case, a protein) is free to diffuse in the simulation and is affected by long range forces. In essence, BD takes into account two major parameters: the general diffusion of the molecule through the (implicit) solvent, and any attractive or repulsive forces that occur between nearby

$$r(t + \Delta t) = r(t) + \frac{DF(t)}{k_B T} + R(t) \quad (2.2)$$

Equation 2.2: Equation for translational motion in Brownian Dynamics simulations, where r is the position at time t , D is the diffusion coefficient, F is the solute-solute interaction forces at time t , and R is a random number.

macromolecules. Finally, a random number generator is used to mimic random “kicks” from the solvent to the molecule of interest (Equation 2.2). Due to its “overdamped” molecular motion, BD

simulations use a much longer timestep than MD simulations do. This allows longer timescales to be more computationally accessible with BD.

While BD is useful for a number of applications, including ligand binding,³⁴ efficiently modeling large or slow systems,³⁵ or even simulating proteins in-cell,³⁶ here we use BD to allow the macromolecules in our models to equilibrate by translating and rigidly rotating in our desired box for long timescales.

E. coli and *H. sapiens* Models

The ultimate goal of these models was to capture an atomistic, realistic, cellular environment that could be simulated at multi-microsecond timescales or longer to determine the

Organism	Force Field	Abbreviation
<i>E. coli</i>	CHARMM 22*	EC22*
<i>E. coli</i>	CHARMM 36m	EC36m
<i>E. coli</i>	CHARMM 36m with CUFIX	EC36mCU
<i>H. sapiens</i>	CHARMM 36m	HC36m

Table 2.1: The various models constructed and their respective abbreviations.

effects of local environment on biomolecular dynamics. Due to the high computational costs of atomistic simulation at microsecond timescales, our strategy was to build a model that we estimated to be the smallest “unit cell” that could accurately replicate the unique electrostatic and steric environment of the cytoplasm—a cube of ~13 nm on each size, or roughly 200,000 atoms. Although these models represent a very small fraction (~“5 ppm” of the total volume of an *E. coli* cell) of the cellular environments they are meant to model, they are realistic and based on our current best knowledge of these environments. The models discussed here have a volume of ~2 x 10⁻⁶ μm³, which is smaller than *E. coli*, human U-2 OS cells, and minimal bacterial cells by a

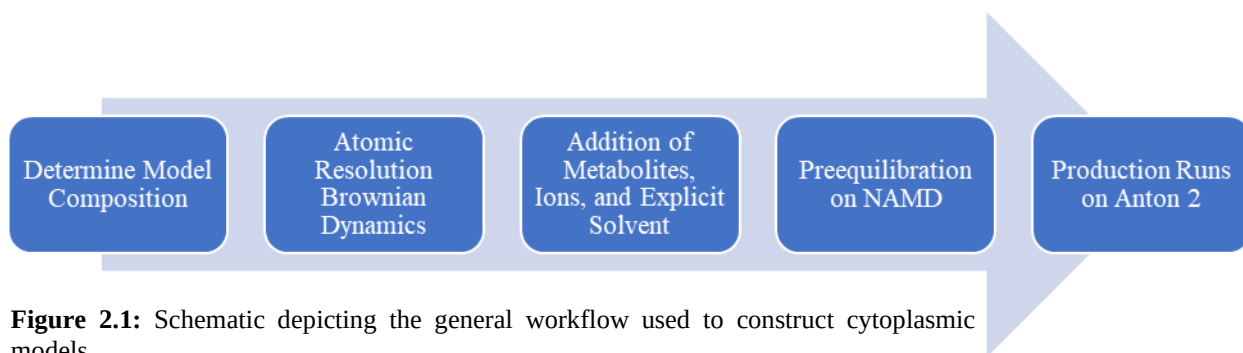
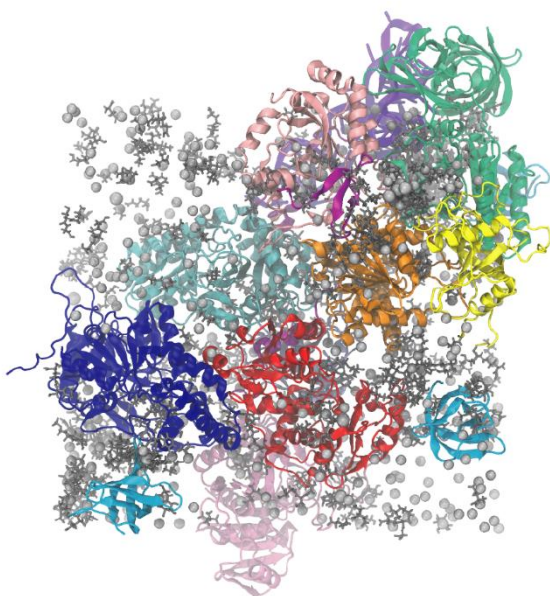


Figure 2.1: Schematic depicting the general workflow used to construct cytoplasmic models.

factors of approximately 235,000, 2 x 10⁹, and 15,000 respectively.^{37–39} Because the model is relatively small, we do not attempt to span a large section of the proteome or metabolome. Rather, we aimed to create a model with a “reasonable” composition: that is, an actual cell might have a segment with these components coexisting for some period of time. This allowed us to evaluate



	Adenylosuccinate Synthetase (ADSS) PDB ID: 1ADE		Mn Superoxide Dismutase (Mn SOD) PDB ID: 1D5N
	Elongation Factor Tu (EFTu) PDB ID: 1OB2		Inorganic Pyrophosphatase (PPA) PDB ID: 2EIP
	Phosphoserine Aminotransferase (PSAT) PDB ID: 1BJO		Cyclophilin A (CYC A) PDB ID: 1LOP
	Aspartate Transcarbamoylase (ATCase) PDB ID: 1ZA1		Cold Shock Protein (3 copies) (CSP) PDB ID: 1MJC
	Malate Dehydrogenase (MD) PDB ID: 2CMD		GTT WW domain (3 copies) (GTT) PDB ID: 2F21 (modified)
	Glyceraldehyde 3-phosphate Dehydrogenase (GAPDH) PDB ID: 1S7C		Phe-tRNA (RNA) PDB ID: 1OB2

Figure 2.2: A snapshot of the EC36m cytoplasm model system. Color-coded macromolecules comprising the system are listed below along with copy number and PDB ID. Ions and small metabolite molecules are shown in gray. Water molecules are not shown for clarity. Figure published in ref. 1.

the general effects the cellular environment might have on various biomolecules over longer time frames. Other simulations have spanned the breadth of the proteome and metabolome more broadly than the models described here.^{25,36}

The design, construction, and simulation of the models follow the general workflow shown in Figure 2.1. First, the composition of the model—the target organism, which macromolecules and metabolites to include, and the approximate volume (and thus protein density) must be decided. These initial decisions are made from a combination of factors, including computational cost,⁴⁰ available data on cell composition,^{41–43} previous computational and experimental studies on the cells in question (for example, the *H. sapiens* model was constructed with U2-OS cells in mind, to better pair with experimental data),^{36,44} and high-resolution structure availability.⁴⁵ Then, the macromolecules only are modelled as rigid bodies using Atomic Resolution Brownian Dynamics (ARBD)³⁵ to “pack” them into the chosen box size and prepare the

initial model. After simulation with ARBD, the rest of the cytoplasm components are added, including metabolites, inorganic ions, and solvent. The completed model is preequilibrated on local computer clusters under a variety of conditions using NAMD.²⁴ Finally, the models are simulated without restraints on the special-purpose supercomputer Anton 2.⁴⁰

In this chapter and throughout, we describe multiple atomistic cellular models run with several MD force fields. We created models designed to mimic cellular conditions in two different organisms: *E. coli* and *H. sapiens* with several different force fields. These models and their abbreviations are described in Table 2.1.

MACROMOLECULAR COMPONENTS OF MODELS

E. coli Models

Because the resulting model is relatively small, we did not aim to span the breadth of the *E. coli* proteome. All proteins selected matched two initial criteria. First, we limited the size of macromolecules to <50 kDa in order to limit the size of the final model. We only included macromolecules that had published experimental structures with 100% identity with the *E. coli* canonical sequence. GTT WW domain, an engineered protein designed to be a “speed limit” folder,⁴⁶ was included to study the interactions between folded and unfolded proteins. The other selected macromolecules were either found to be highly abundant in the *E. coli* proteome⁴⁷ (as is the case with EFTu) or used in previous Brownian Dynamics simulations.³⁶ Further, to avoid biologically functional or specific protein-protein interactions, we attempted to sample proteins from different pathways that had no known biological interactions with each other and checked each protein’s known interactions on UniProt.⁴⁸ We acknowledge that because of the sheer size of a complete proteome, there are dozens of other macromolecules that fit these criteria and could have been selected instead.

Both the EC22* and the EC36m/EC36mCU models contain all proteins listed in Figure 2.2. Initial simulations of the EC22* model indicated the DNA binding domain was unstable in the cellular environment, so it was removed in the EC36m and EC36mCU models. The macromolecular components of all three *E. coli* models are listed in Table A.1.

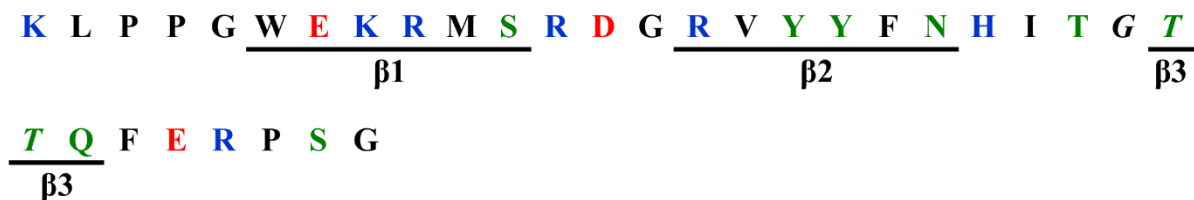


Figure 2.3: Sequence of GTT, the fast-folding WW domain used in all *E. coli* models. Underlined residues are beta strands. Residues are colored by their residue type: hydrophobic (black), polar (green), basic (blue), and acidic (red).

GTT Model Construction

No crystal structure of the fast-folding WW domain GTT exists in the PDB. However, a similar construct (Fip) has been crystallized experimentally.⁴⁹ The first 37 residues of this crystal structure (PDB ID: 2F21) were used to construct the initial GTT structures. The GTT structures were constructed by mutating the existing crystal structure at residues N26G, A27T and S28T. The full sequence of the construct is shown in Figure 2.3. Initial unfolded structures were obtained with temperature-jump molecular dynamics simulations using NAMD 2.9.²⁴ Unfolded GTT was subjected to equilibrium MD with C22* force field²² in a 5x5x5 nm³ box with TIP3P water⁵⁰ and 0.15 M NaCl, in the NPT ensemble. The Langevin piston and Nosé–Hoover procedure were used to maintain temperature and pressure at constant values (500 K and 1 atm).^{51,52} The final structures were equilibrated at 318 K before being placed in the cytoplasm models. The simulation of GTT alone was run in the identical cube of water with a size of 5 nm initially used to obtain the

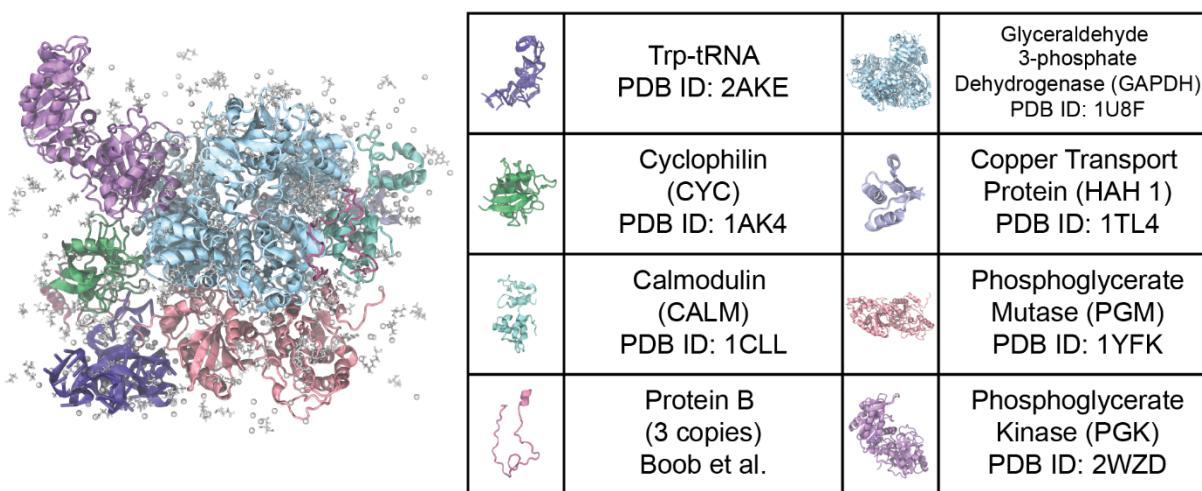


Figure 2.4: A snapshot of the HC36m cytoplasm model system. Color-coded macromolecules comprising the system are listed below along with copy number and PDB ID. Ions and small metabolite molecules are shown in gray. Water molecules are not shown for clarity.

unfolded GTT structures. The EC22* cytoplasm model contained three unfolded copies of GTT, while the EC36m and EC36mCU models contained 1 folded and 2 unfolded copies of GTT^{21,23} each.

H. sapiens Models

Proteins in the *H. sapiens* models were selected in a similar manner to the *E. coli* models. We started with human proteomic data of the U-2 OS osteosarcoma cell line.³⁸ This proteomic data was used to confine ourselves to the hundred or so most abundant proteins in the proteome.

By cross-referencing with the protein database UniProt,⁴⁸ we eliminated any proteins that were not shown to be cytoplasmic. Further, with one notable exception (GAPDH, which was selected to observe the quinary interactions between it and a known partner, PGK),⁵³ we opted to included proteins that were relatively small (≤ 50 kDa) to ensure free translation of the proteins throughout the model. The UniProt database of protein sequences was used to check for quaternary structures and to identify appropriate experimental protein structures in the Protein Data Bank.⁴⁵ Incomplete structures were predicted using the homology modelling package MODELLER.⁵⁴ A snapshot of the completed system and a chart of all macromolecular components of the HC36m model is shown in Figure 2.4.

Protein B Model Construction

All three copies of Protein B in the HC36m and HDA models were in the unfolded state. Protein B structures were randomly selected from the unfolded ensemble of simulations described in ref.⁵⁵. The sequence of the construct is shown in Figure 2.5.

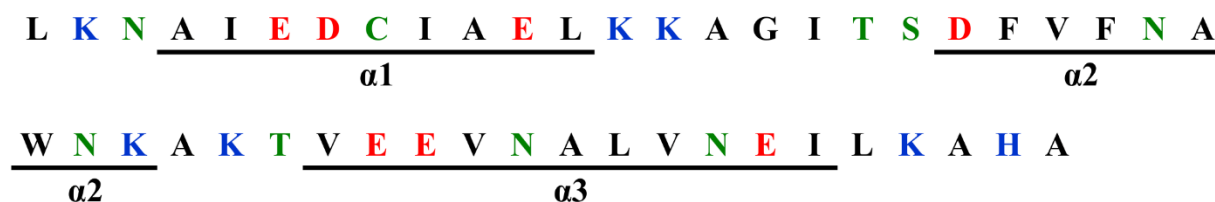


Figure 2.5: Sequence of Protein B, the fast-folding domain used in all *H. sapiens* models. Underlined residues are alpha helices. Residues are colored by their residue type: hydrophobic (black), polar (green), basic (blue), and acidic (red).

ATOMIC RESOLUTION BROWNIAN DYNAMICS

E. coli Models

All macromolecules were initially equilibrated together using the Atomic Resolution Brownian Dynamics (ARBD) software,³⁵ which allowed the macromolecules to fit into and sample the computational periodic unit cell. Macromolecules were packed into a 120 Å cube and modelled as rigid bodies with an electrostatic and Lennard-Jones potential. Potentials were distributed by a 3-D 2 Å grid using VMD's VolMap plugin. The rigid bodies were modelled according to a Langevin equation with a 200 fs timestep. Langevin damping coefficients were estimated using the Hydropro software package (Calculated diffusional and rotational Langevin damping coefficients were on the order of 500-2000 ns⁻¹ and 4000-8000 ns⁻¹ respectively).⁵⁶ After ARBD

ADDITION OF METABOLITES, INORGANIC IONS, AND WATER

E. coli Models

Due to the finite size of the box, only the most abundant metabolite species were included. Metabolite numbers for each model were either picked from the experimental concentration distribution directly, with a cutoff at <1 molecule/box (EC22*), or an in-house Monte Carlo code (see Code A.1) was used to sample the number of copies of each metabolite based on the experimental metabolite concentrations (Figure 2.6).

In the EC22* model, metabolites were selected based solely on known experimental concentrations and assumed even distribution of these metabolites throughout the cell.⁴¹ In the EC36m/EC36mCU models, copy numbers of metabolites were determined using an in-house Monte Carlo sampling scheme based on experimental data; in brief we have randomly drawn metabolite names from the experimental distribution (Figure 2.6). In both cases, unparameterized metabolites were parameterized by analogy using CGenFF^{57,58} webserver. Metabolites were packed into the model randomly using PACKMOL.⁵⁹ The metabolite counts in each model are shown in Table A.2. For the metabolite data used to make the randomized distribution, please see supplementary file A.1 “Bacterial”. Across all systems, about 45 species of metabolites were sampled, from the most common (glutamate, 96 mM) to 59th most common (methionine, 0.15 mM). Figure 2.6 illustrates that the Monte Carlo sampling scheme selected near-biological concentrations of each metabolite, while allowing for selecting a wider range of low concentration metabolites.

VMD⁶⁰’s Solvate and Autoionize plugins were used to add TIP3P water⁵⁰ and inorganic ions. To stabilize the tRNA, 20 Mg²⁺ ions were selectively placed around the tRNA using VMD’s CIonize plugin. After the addition of inorganic ions and explicit solvent, the final system contains approximately 200,000 atoms. The macromolecular concentrations of the systems are ~300 mg/mL with ~220 mM metabolites and ~350 mM inorganic ions. Final details of the constructed *E. coli* models are shown in Table 2.2.

	EC22*	EC36m/EC36mCU	HC36m
Total Atoms	206,821	190,817	203,903
Box size (Approximate)*	12.5 nm cube	12.5 nm cube	12.5 nm cube
Macromolecules			
Atom Count	49,082	48,343	45,270
Mass (kDa)	356	351	328
Concentration (mg/mL)	303	299	279
Metabolites			
Atom Count	6,688	7,884	4,097
Ions (total)	439	427	216
K ⁺	353	353	188
Na ⁺	30	24	5
Mg ²⁺ (free/bound [†])	40/0	20/23	0/10
Cl ⁻	16	7	13
Water (atom count)	150,612	134,163	154,320

Table 2.2: Components and details of the models. *Because the system was run in an NPT ensemble, slight volume fluctuations occur throughout the simulation. Therefore, changes in concentration are also expected. [†]Bound Mg²⁺ refers only to the ions that were selectively placed near the tRNA. They are not restrained in any way. Modified table published in ref 1.

H. sapiens Models

Determining Consensus Metabolite Concentrations

For the *H. sapiens* models, we used two different experimental studies to determine a consensus metabolite concentration for human cells.^{42,43} In each study, we opted to use “whole-cell” measurements whenever possible. In cases where multiple whole-cell measurements were available, we averaged all replicates to get an average concentration of each metabolite from each study. We then averaged the concentrations given from each of these studies to form a “consensus” metabolite concentration in mammalian cells. The full consensus concentrations in the *H. sapiens* models are shown in Supplementary File A.1 “Mammalian”.

As in the EC36m and EC36mCU models, copy numbers of metabolites were determined using an in-house Monte Carlo sampling scheme based on experimental data. In the mammalian models, 44 different species of metabolites were sampled, ranging from the most abundant (glutamate, 32 mM) to 77th most abundant (GMP, 0.1 mM). A comparison of the metabolite concentrations represented in the *H. sapiens* models and the consensus experimental concentrations is shown in Figure 2.7. For details of metabolite copy numbers, please see Table A.3.

describe the “standard conditions” I use for these simulations. Unless otherwise noted, these simulations were run on NAMD 2.9.²⁴ Additionally, the following parameters were used for all NAMD simulations, unless explicitly otherwise noted. A 12 Å cutoff was used for short-range non-bonded interactions, with a switching distance starting at 10 Å. Long range electrostatics were calculated using the Particle Mesh Ewald (PME) method⁶² with a grid density of 1 Å⁻³. The SHAKE algorithm⁶³ was used to hold bonds including hydrogen rigid, and the timestep was set at 2 femtoseconds.

E. coli Models

To equilibrate the structure before production runs on Anton 2, all three *E. coli* models were equilibrated for ~150 ns using NAMD 2.9.²⁴ To maximize stability of the macromolecules, artificial restraints were applied throughout preequilibration—first, only water molecules were free to move, with all other atoms fixed in their original positions, then metabolites and inorganic ions were released. Side chains of proteins were released 200 ps before the molecular backbone. However, even after the molecular backbone was released, we added extra harmonic constraints between C α atoms within the same proteins in order to force the protein to remain folded. Details of the preequilibration are shown in Table 2.3. For more information about the C α constraints, see Code A.2.

Restrained species	Time
Metabolites, ions, macromolecules	200 ps
Macromolecules	200 ps
Macromolecular backbone	200 ps
C α constrained*	
Heat from 318 K to 498 K (10 K increments)	100 ps/temperature
Hold at 498 K	66 ns
Cool from 498 K to 318 K (10 K increments)	100 ps/temperature
Hold at 318 K	53 ns

Table 2.3: Details of preequilibration runs with NAMD 2.9. *C α (and P, in the case of the tRNA) atoms were constrained during PR equilibration to prevent macromolecules from unfolding. All macromolecules were free to translate throughout the periodic cell. For more information about C α constraints, please see Code A.2. Table published in ref 1.

H. sapiens Models

In the interest of time, *H. sapiens* models were minimized for 6000 steps and equilibrated

briefly (~7.3 ns) using NAMD 2.9²⁴ without constraints before being transferred for extensive production runs on Anton 2.

PRODUCTION RUNS

All 3 models were simulated using an NPT ensemble with a multigrator⁶⁴ scheme at 1 atm pressure and 318 K. This temperature was selected to maximize protein sampling while staying realistic for the cellular environment. A sample Anton 2 input file is included in appendix A (Code A.3).

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (2) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcb.0c02536>.
- (3) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585–590. <https://doi.org/10.1038/267585a0>.
- (4) Guvench, O.; MacKerell, A. D. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Methods Molecular Biology™; Humana Press: Totowa, NJ, 2008; pp 63–88. https://doi.org/10.1007/978-1-59745-177-2_4.
- (5) Fröhlich, T.; Bernetti, M.; Calonaci, N.; Bussi, G. Toward Empirical Force Fields That Match Experimental Observables. *J. Chem. Phys.* **2020**, *152* (23), 230902. <https://doi.org/10.1063/5.0011346>.
- (6) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. <https://doi.org/10.1021/ja00214a001>.
- (7) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197. <https://doi.org/10.1021/ja00124a002>.
- (8) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616. <https://doi.org/10.1021/jp973084f>.
- (9) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676. <https://doi.org/10.1002/jcc.20090>.
- (10) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. Force Field Bias in Protein Folding Simulations. *Biophys. J.* **2009**, *96* (9), 3772–3780. <https://doi.org/10.1016/j.bpj.2009.02.033>.
- (11) Lopes, P. E. M.; Guvench, O.; MacKerell, A. D. Current Status of Protein Force Fields for Molecular Dynamics. *Methods Mol. Biol. Clifton NJ* **2015**, *1215*, 47–71. https://doi.org/10.1007/978-1-4939-1465-4_3.
- (12) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517–520. <https://doi.org/10.1126/science.1208351>.
- (13) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10* (11), 5113–5124. <https://doi.org/10.1021/ct500569b>.

- (14) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (21), E4758–E4766. <https://doi.org/10.1073/pnas.1800690115>.
- (15) Petrov, D.; Zagrovic, B. Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Comput. Biol.* **2014**, *10* (5), e1003638–e1003638. <https://doi.org/10.1371/journal.pcbi.1003638>.
- (16) Yu, L.; Li, D.-W.; Brüschweiler, R. Systematic Differences between Current Molecular Dynamics Force Fields To Represent Local Properties of Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2021**, *125* (3), 798–804. <https://doi.org/10.1021/acs.jpcb.0c10078>.
- (17) Nawrocki, G.; Wang, P.-H.; Yu, I.; Sugita, Y.; Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. **2017**. <https://doi.org/10.1021/acs.jpcb.7b08785>.
- (18) Nawrocki, G.; Karaboga, A.; Sugita, Y.; Feig, M. Effect of Protein–Protein Interactions and Solvent Viscosity on the Rotational Diffusion of Proteins in Crowded Environments. *Phys. Chem. Chem. Phys.* **2019**, *21* (2), 876–883. <https://doi.org/10.1039/C8CP06142D>.
- (19) Bernèche, S.; Roux, B. Energetics of Ion Conduction through the K⁺ Channel. *Nature* **2001**, *414* (6859), 73–77. <https://doi.org/10.1038/35102067>.
- (20) Lennard-Jones, J. E. Cohesion. *Proc. Phys. Soc.* **1931**, *43* (5), 461–482. <https://doi.org/10.1088/0959-5309/43/5/301>.
- (21) Yoo, J.; Aksimentiev, A. New Tricks for Old Dogs: Improving the Accuracy of Biomolecular Force Fields by Pair-Specific Corrections to Non-Bonded Interactions. *Phys Chem Chem Phys* **2018**, *20* (13), 8432–8449. <https://doi.org/10.1039/C7CP08185E>.
- (22) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–49. <https://doi.org/10.1016/j.bpj.2011.03.051>.
- (23) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71–73. <https://doi.org/10.1038/nmeth.4067>.
- (24) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802. <https://doi.org/10.1002/jcc.20289>.
- (25) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm. *eLife* **2016**, *5*. <https://doi.org/10.7554/eLife.19274>.
- (26) Bülow, S. von; Siggel, M.; Linke, M.; Hummer, G. Dynamic Cluster Formation Determines Viscosity and Diffusion in Dense Protein Solutions. *Proc. Natl. Acad. Sci.* **2019**, 201817564–201817564. <https://doi.org/10.1073/PNAS.1817564116>.
- (27) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119* (16), 5113–5123. <https://doi.org/10.1021/jp508971m>.
- (28) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950–1958. <https://doi.org/10.1002/prot.22711>.

- (29) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix–Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113* (26), 9004–9015. <https://doi.org/10.1021/jp901540t>.
- (30) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725. <https://doi.org/10.1002/prot.21123>.
- (31) Best, R. B.; de Sancho, D.; Mittal, J. Residue-Specific α -Helix Propensities from Molecular Simulation. *Biophys. J.* **2012**, *102* (6), 1462–1467. <https://doi.org/10.1016/j.bpj.2012.02.024>.
- (32) Piana, S.; Robustelli, P.; Tan, D.; Chen, S.; Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes. *J. Chem. Theory Comput.* **2020**, *16* (4), 2494–2507. <https://doi.org/10.1021/acs.jctc.9b00251>.
- (33) Ermak, D. L.; McCammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.* **1978**, *69* (4), 1352–1360. <https://doi.org/10.1063/1.436761>.
- (34) Seitz, C.; Casalino, L.; Konecny, R.; Huber, G.; Amaro, R. E.; McCammon, J. A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys. J.* **2020**, *119* (11), 2275–2289. <https://doi.org/10.1016/j.bpj.2020.10.024>.
- (35) Comer, J.; Aksimentiev, A. Predicting the DNA Sequence Dependence of Nanopore Ion Current Using Atomic-Resolution Brownian Dynamics. *J. Phys. Chem. C* **2012**, *116* (5), 3376–3393. <https://doi.org/10.1021/jp210641j>.
- (36) McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput. Biol.* **2010**, *6*, e1000694. <https://doi.org/10.1371/journal.pcbi.1000694>.
- (37) Wang, Y.; Hammes, F.; Düggelein, M.; Egli, T. Influence of Size, Shape, and Flexibility on Bacterial Passage through Micropore Membrane Filters. *Environ. Sci. Technol.* **2008**, *42* (17), 6749–6754. <https://doi.org/10.1021/es800720n>.
- (38) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szyborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. The Quantitative Proteome of a Human Cell Line. *Mol. Syst. Biol.* **2011**, *7* (549), 1–8. <https://doi.org/10.1038/msb.2011.82>.
- (39) Breuer, M.; Earnest, T. M.; Merryman, C.; Wise, K. S.; Sun, L.; Lynott, M. R.; Hutchison, C. A.; Smith, H. O.; Lapek, J. D.; Gonzalez, D. J.; de Crécy-Lagard, V.; Haas, D.; Hanson, A. D.; Labhsetwar, P.; Glass, J. I.; Luthey-Schulten, Z. Essential Metabolism for a Minimal Cell. *eLife* **8**, e36842. <https://doi.org/10.7554/eLife.36842>.
- (40) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. B.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE: New Orleans, LA, USA, 2014; pp 41–53. <https://doi.org/10.1109/SC.2014.9>.

- (41) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in *Escherichia Coli*. *Nat. Chem. Biol.* **2009**, *5*, 593–599. <https://doi.org/10.1038/nchembio.186.Absolute>.
- (42) Diener, C.; Muñoz-Gonzalez, F.; Encarnación, S.; Resendis-Antonio, O. The Space of Enzyme Regulation in HeLa Cells Can Be Inferred from Its Intracellular Metabolome. *Sci. Rep.* **2016**, *6* (1), 28415. <https://doi.org/10.1038/srep28415>.
- (43) Chen, W. W.; Freinkman, E.; Wang, T.; Birsoy, K.; Sabatini, D. M. Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism. *Cell* **2016**, *166* (5), 1324–1337.e11. <https://doi.org/10.1016/j.cell.2016.07.040>.
- (44) Ebbinghaus, S.; Dhar, A.; Douglas McDonald, J.; Gruebele, M. Protein Folding Stability and Dynamics Imaged in a Living Cell. *Nat. Publ. Group* **2010**, *7* (4), 319–319. <https://doi.org/10.1038/nmeth.1435>.
- (45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. <https://doi.org/10.1093/nar/28.1.235>.
- (46) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β -Sheet Protein. *J. Mol. Biol.* **2011**, *405* (1), 43–48. <https://doi.org/10.1016/j.jmb.2010.10.023>.
- (47) Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrné, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M. The Quantitative and Condition-Dependent *Escherichia Coli* Proteome. *Nat. Biotechnol.* **2016**, *34*, 104–110. <https://doi.org/10.1038/nbt.3418>.
- (48) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
- (49) Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. Structure-Function-Folding Relationship in a WW Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (28), 10648–10653. <https://doi.org/10.1073/pnas.0600511103>.
- (50) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. <https://doi.org/10.1063/1.445869>.
- (51) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621. <https://doi.org/10.1063/1.470648>.
- (52) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189. <https://doi.org/10.1063/1.467468>.
- (53) Sukenik, S.; Ren, P.; Gruebele, M. Weak Protein–Protein Interactions in Live Cells Are Quantified by Cell-Volume Modulation. *Proc. Natl. Acad. Sci.* **2017**, *114* (26), 201700818–201700818. <https://doi.org/10.1073/pnas.1700818114>.
- (54) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **2016**, *54*, 5.6.1–5.6.37. <https://doi.org/10.1002/cpbi.3>.
- (55) Boob, M.; Sukenik, S.; Gruebele, M.; Pogorelov, T. TMAO: Protecting Proteins from Feeling the Heat. **2020**. <https://doi.org/10.26434/chemrxiv.12665648.v1>.

- (56) Ortega, A.; Amorós, D.; García de la Torre, J. Prediction of Hydrodynamic and Other Solution Properties of Rigid Proteins from Atomic- and Residue-Level Models. *Biophys. J.* **2011**, *101*, 892–898. <https://doi.org/10.1016/j.bpj.2011.06.046>.
- (57) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155–3168. <https://doi.org/10.1021/ci3003649>.
- (58) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154. <https://doi.org/10.1021/ci300363c>.
- (59) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. Packmol: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164. <https://doi.org/10.1002/jcc.21224>.
- (60) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (61) Theillet, F.-X.; Binolfi, A.; Frembgen-Kesner, T.; Hingorani, K.; Sarkar, M.; Kyne, C.; Li, C.; Crowley, P. B.; Gierasch, L.; Pielak, G. J.; Elcock, A. H.; Gershenson, A.; Selenko, P. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114* (13), 6661–6714. <https://doi.org/10.1021/cr400695p>.
- (62) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. <https://doi.org/10.1063/1.464397>.
- (63) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. [https://doi.org/10.1016/0021-9991\(77\)90098-5](https://doi.org/10.1016/0021-9991(77)90098-5).
- (64) Lippert, R. A.; Predescu, C.; Ierardi, D. J.; Mackenzie, K. M.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Accurate and Efficient Integration for Molecular Dynamics Simulations at Constant Temperature and Pressure. *J. Chem. Phys.* **2013**, *139*, 164106. <https://doi.org/10.1063/1.4825247>.

CHAPTER 3: THE IMPACT OF THE CELLULAR ENVIRONMENT ON SIMULATED METABOLITE CONFORMATIONS

INTRODUCTION

Proteins *in vivo* perform their biological functions in a complex environment of water, inorganic ions, metabolites, and other macromolecules. The cell's interior contains ~10-40% macromolecules by volume,¹ creating a crowded environment where nonspecific, nonfunctional interactions between various components are inevitable. Experimental and computational studies have demonstrated that the cellular environment greatly impacts proteins' stability, folding kinetics, structural ensembles, binding, and quinary structure.²⁻¹¹ However, the impact of the cellular environment on small molecules is largely unknown.

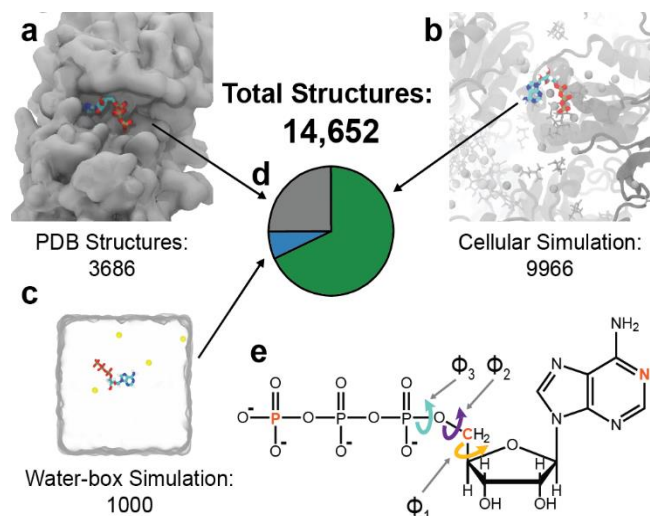


Figure 3.1: Environments overview (a) Representative snapshot of ATP (color) bound to a protein (gray). PDB ID: 1ZYD. (b) Snapshot of the EC36mCU cellular simulation. ATP copies are highlighted in color. (c) Snapshot of Water-box Simulation. (d) Pie chart of structures used in this study: cellular simulation (green), PDB structures (gray), and water-box simulation (blue). (e) Drawing of the structure of ATP, with the geometric coordinates of interest highlighted: Φ_1 (yellow), Φ_2 (purple), and Φ_3 (cyan). The atoms that make up the PCN angle are highlighted in orange.

Adenosine triphosphate, or ATP, is one of the most abundant metabolites in the cell, with a concentration in the ~1-10 mM range across a variety of cell types and organisms.¹²⁻¹⁴ ATP is canonically known as the energy source of the cell.¹⁵ However, recent studies have found that the presence of ATP has a variety of effects on nearby macromolecules, including increasing solubility,^{16,17} stabilizing nearby proteins,¹⁸ and modulating the liquid-liquid phase separation of intrinsically disordered proteins.¹⁹ These findings point to additional functions that likely evolved

* This chapter is reproduced in part with permission from:

Rickard, M.M.; De Lio, A.M.; Luo, H.; Gruebele, M.; Pogorelov, T.V. "The Impact of Local Environment on Simulated Metabolite Conformations." *In preparation, J. Phys. Chem. Lett.* **2021**.

from the close proximity of ATP molecules and macromolecules in the dense environment of the cell.

Previous studies have also shown that ATP's geometry in water-box MD simulations differs from the structures found in the PDB.²⁰ However, it is unknown whether the unique steric and electrostatic environment of the cell affects ATP's conformational distribution.

Here, we study the effect of local environment on the conformation of ATP. By comparing over 14,000 ATP structures in the Protein Data Bank²¹ (PDB), in atomistic simulations of ATP in the cytoplasm, and in a waterbox environment, we find that ATP's geometry is minimally impacted by the cellular environment, but occupies a far wider conformational space in the PDB.

METHODS

For details on EC36mCU model construction and simulation, please see Chapter 2.

Dataset Assembly

To determine the role of local environment on ATP dynamics, we opted to compare our in-cell simulations to two datasets—a water-box simulation of ATP to compare ATP's dynamics in a test-tube like environment, and the PDB, which represents the structures adopted by ATP when bound to the active site of a protein (Figure 3.1).

To study the effects of the cellular environment on ATP conformational distribution, we used data from the EC36mCU model. This force field was selected due to observations that both the EC36m and EC36mCU models had less macromolecular sticking than the EC22* model (see Ch. 5 for more details). Therefore, we hypothesized that the EC36m and EC36mCU models more accurately modeled the in-cell environment.

With 11 ATP copies in the EC36mCU model, we analyzed the ATP conformations sampled in 906 frames (snapshots every 24 ns for ~21.7 μ s), resulting in 9,966 unique ATP structures from the cellular simulations. To further separate the microenvironments within the cellular model, we applied filters to determine the identity of ATP's neighbors. These filters sort ATP structures by whether or not they are within a 5 Å cutoff of macromolecules using VMD's "pbwithin" function.²²

Water-box simulations of ATP were performed using identical parameters to the cellular simulations.²³ These simulations consisted of 1 ATP molecule, four sodium ions to balance charge,

and TIP3P water molecules.²⁴ This system was simulated for 100 ns in the NPT ensemble (T = 303 K, P = 1 atm), using NAMD 3.0 Alpha.²⁵ Because experimental studies have shown that similarly-sized free amino acids and short peptides form intermolecular contacts on a similar timescale, we hypothesized that 100 ns would be sufficient to sample ATP’s conformational space.²⁶ Snapshots from the trajectory were taken every 100 ps, resulting in 1000 unique ATP structures.

Dataset		Number of Unique ATP Structures
In-Cell Simulation		9966
	Macromolecular	7776
	Non-macromolecular	2190
Water-box Simulation		1000
Protein Data Bank		3686
Total		14,652

Table 3.1: Number of unique structures in each dataset.

Structures in the PDB dataset were obtained by searching “Chemical ID in ATP” in the PDB’s Advanced Search Query Builder.²¹ This resulted in 1,526 unique PDB structures. Because multiple copies of ATP were often found in a single unit cell, this analysis contained 3,686 unique ATP structures. Altogether, between the cellular simulations, water-box simulations, and the PDB dataset, 14,652 unique ATP structures were analyzed (Table 3.1).

RESULTS AND DISCUSSION

Geometric Analysis

Given recent results that both the nucleoside base and the phosphate tail are necessary for ATP’s function as a protein hydrotrope,¹⁶ we hypothesize that the interactions between and geometries of the base and tail play an important role in ATP dynamics and function. To better understand the relative positioning of the nucleoside base and phosphate tail, we calculated four geometric coordinates of interest: the $P\gamma$ -C5’-N1 angle (PCN angle); the C3’-C4’-C5’-O5’ dihedral (Φ_1), the C4’-C5’-O5’-P α dihedral (Φ_2), and the C5’-O5’-P α -O3 α dihedral (Φ_3) (Figure 3.1). The results of these calculations are shown in Figure 3.2.

By comparing the geometric coordinates of interest in the four local environments, we observed that ATP’s structure is affected by the local environment, particularly the presence of macromolecules (in this case, the PDB and macromolecular in-cell simulation structures). Qualitatively, we observe similarities in the Φ_3 and PCN angle distribution in the macromolecular

in-cell simulation and the PDB datasets. This could indicate that the presence of nearby macromolecules may alter ATP's conformational landscape whether or not ATP is specifically bound (as it is in the PDB).

Additionally, we see similarities between the non-macromolecular in-cell simulation and water-box structures. Notably, the water-box and non-macromolecular cell simulation datasets appear to have a higher population of low-PCN structures (PCN angle $< 90^\circ$) than the PDB or in-cell macromolecular structures (Figure 3.2, fourth column). We hypothesize that these low-PCN structures are the result of ATP molecules coordinating neighboring inorganic ions, and this effect may be mitigated when ATP is interacting with nearby macromolecules.

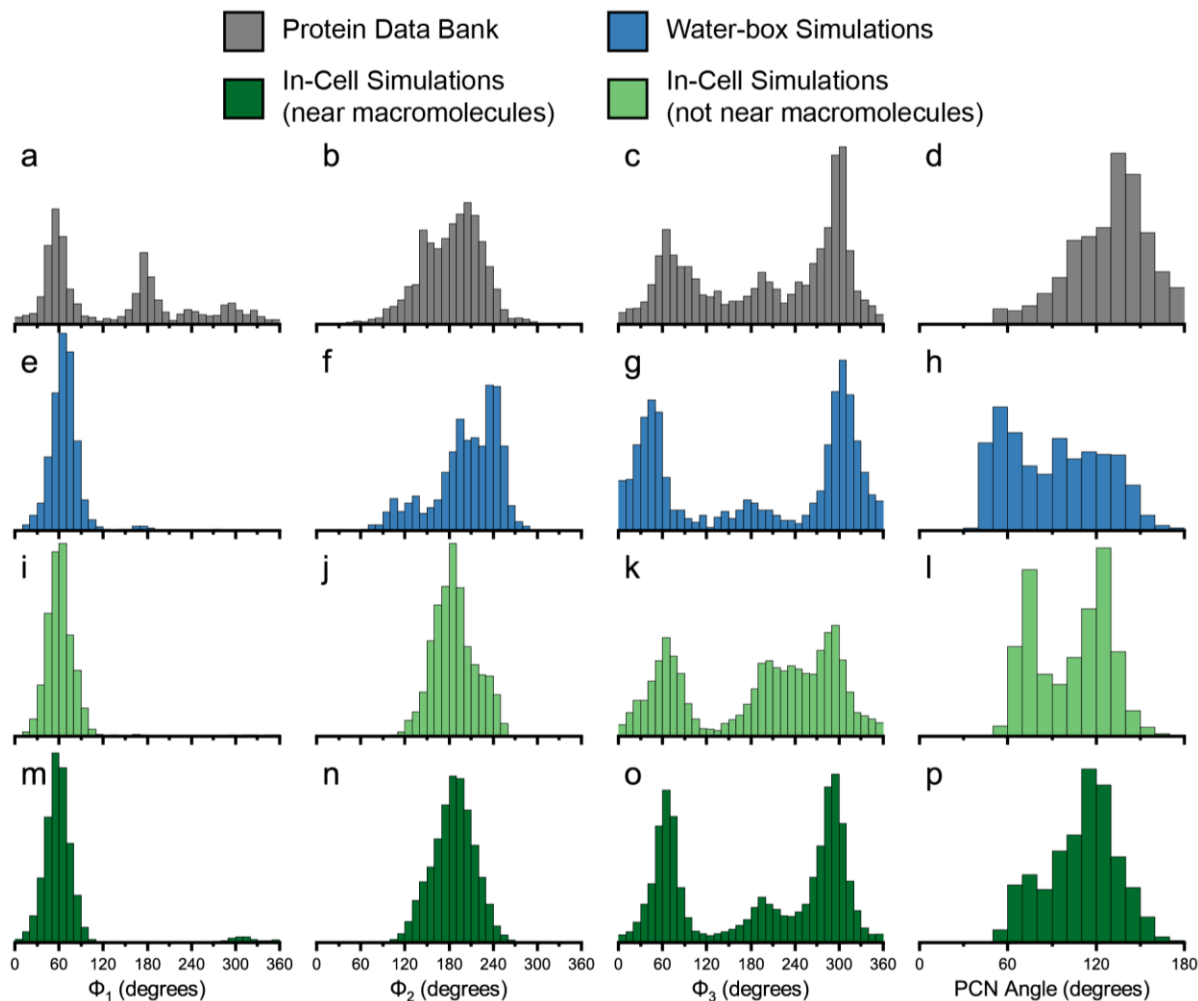


Figure 3.2: Comparison of Φ_1 , Φ_2 , Φ_3 , and PCN angles in the PDB samples (top row, gray), water-box simulation (second row, blue), and in different conditions in the cellular model: not within 5 Å of a macromolecule (third row, light green) and within 5 Å of macromolecules (bottom row, dark green).

However, despite the similarities we observe between the geometric distributions of various datasets, the PDB dataset has a dramatically different Φ_1 distribution (Figure 3.2a) than we observe in any of the MD simulations (Figure 3.2e, i, and m). Although in-cell macromolecular data (Figure 3.2m) had a small population of structures with Φ_1 values greater than 110° (315/7776 structures, 4.0%), particularly between 250° and 360° (290/7776 structures, 3.7%), nearly half of

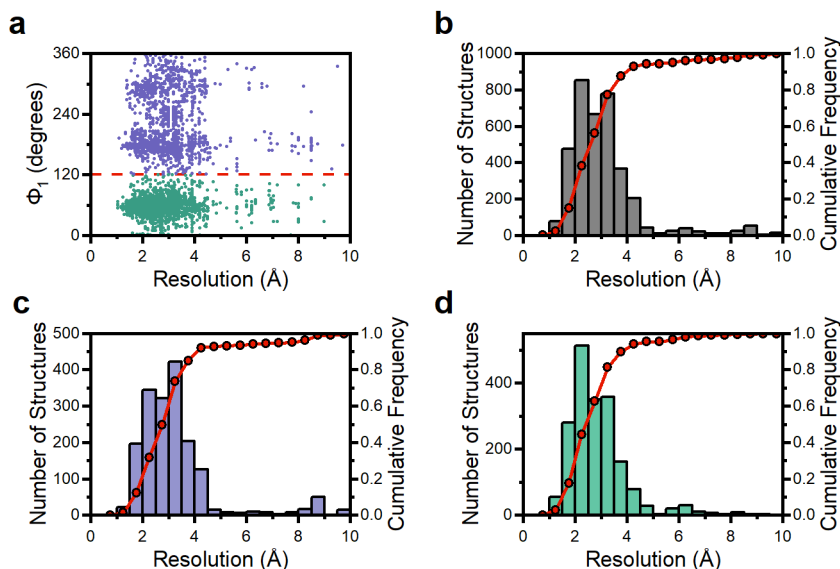


Figure 3.3: Comparison of resolution of ATP PDB structures. (a) Scatter plot of Φ_1 versus PDB structure resolution. Structures with $\Phi_1 > 120^\circ$ are shown in blue, and structures with $\Phi_1 \leq 120^\circ$ are shown in green. (b) Distribution of the resolution of all structures in the PDB dataset (gray).* The cumulative fraction of this distribution is overlaid in red. (c) Distribution of any structure in the PDB dataset with $\Phi_1 > 120^\circ$ (blue).* The cumulative fraction of this distribution is overlaid in red. (d) Distribution of any structure in the PDB dataset with $\Phi_1 \leq 120^\circ$ (green).* The cumulative fraction of this distribution is overlaid in red. *10/3686 structures had no resolutions listed. 9 structures (5 structures with $\Phi_1 > 120^\circ$) had a resolution larger than $10 \text{ \AA}$ and are not shown in the plot.

the PDB dataset had ATP structures greater than 110° (1791/3686 structures, 48.5%) and over four times more structures in the 250° to 360° range (567/3686 structures, 15.4%) than the macromolecular in-cell data.

Intriguingly, the wide distribution of Φ_1 values found in the PDB dataset were not observed in earlier studies, which sampled far fewer PDB structures.²⁰ To ensure the high- Φ_1 conformations were not an artifact of poorly resolved PDB structures, we compared the resolution of the PDB structures with $\Phi_1 > 120^\circ$ to the entire PDB dataset (Figure 3.3). Although there was a slightly higher proportion of resolution $> 5 \text{ \AA}$ structures that had high Φ_1 values, overall, the two distributions were very similar. Additionally, we observed no correlation between structure

resolution and Φ_1 value (Figure 3.3a). Therefore, we concluded that the Φ_1 distribution in the PDB dataset accurately reflected ATP poses bound to macromolecules.

Potential Energy Surface Scans

To better understand the energy landscape of ATP's base/tail interactions, we conducted rigid potential energy scans of Φ_1 , Φ_2 , and Φ_3 in representative PDB structures. Initial structures were randomly selected from the PDB dataset (PDB ID: 1A01 and PDB ID: 2AQX). In both cases, the ATP's charge was set to -4. The presence of solvent was estimated by the polarizable continuum model.²⁷ Calculations were performed using the B3LYP method²⁸ and the 6-31G (d) basis set.²⁹ In each calculation, the dihedral of interest was rotated fully over 36 steps (10 degree increments) and the potential energy at each dihedral value calculated. Results are shown in Figure 3.4.

Overall, the rigid dihedral scan calculations agree well with the geometric distributions observed in the dataset (Figure 3.4). Strikingly, the potential energy surface calculations somewhat validate the unusual Φ_1 distribution observed in the PDB dataset, matching both the relative position and relative well depths ($60^\circ > 180^\circ > 300^\circ$) of the PDB distribution (Figure 3.4a). However, the large potential differences between these minima (~ 5 kcal/mol) do not match the PDB distributions, whose peak areas suggest a potential energy difference on the order of 1-3 kcal/mol. In the future, we plan to further analyze these results to determine whether proteins in the PDB dataset are stabilizing these unusual Φ_1 values.

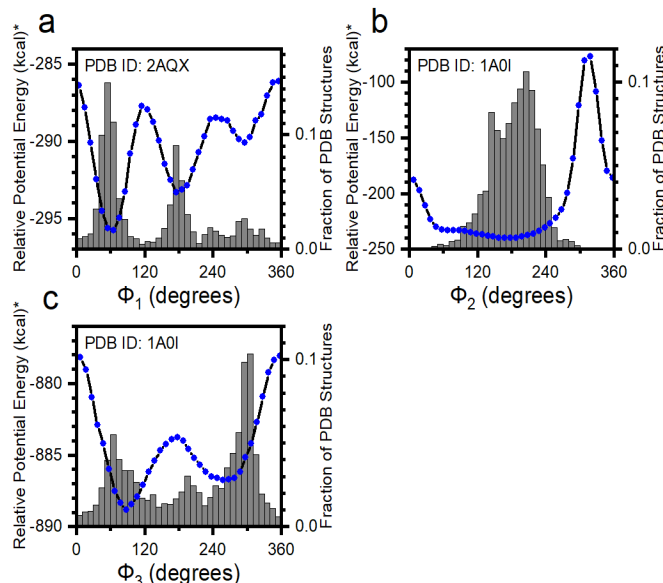


Figure 3.4: (a-c) Results from rigid potential energy surface calculations of Φ_1 , Φ_2 , and Φ_3 (left axis, line) overlaid on the distribution of the respective dihedral angles in the PDB dataset (right axis, histogram). Blue markers on the potential energy surface indicate where the incremented stops on the scan were. PDB ID's noted on each panel indicate the starting structure used for each potential energy scan. *Relative energy was calculated by adding 1,671,800 kcal to absolute values output by Gaussian.

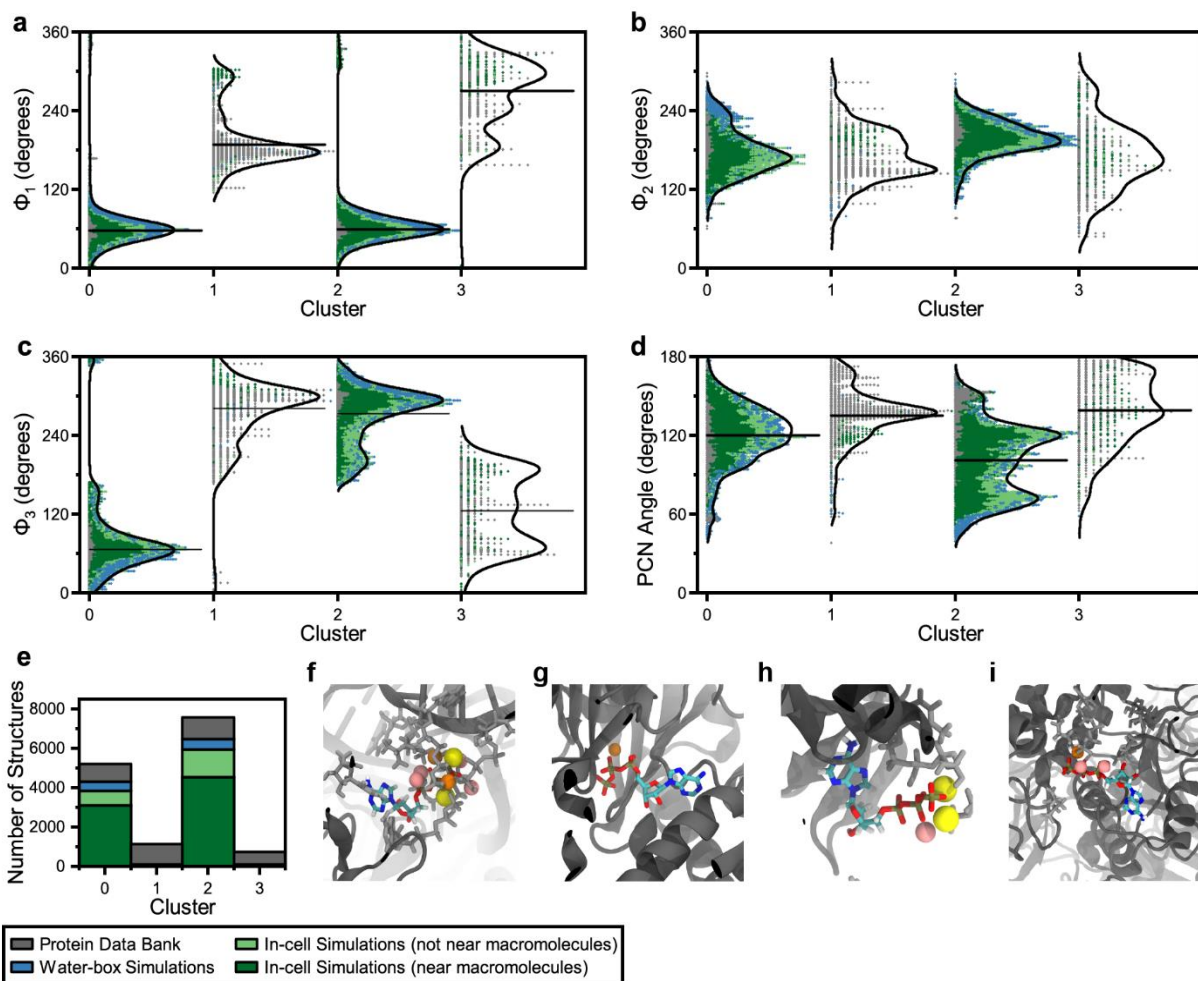


Figure 3.5: Results of k-means clustering analysis. (a-d) Observed Φ_1 , Φ_2 , Φ_3 , and PCN observed in each cluster. Individual structures are displayed as points, colored by the dataset the structure originally came from. Black curves represent the overall distribution of each geometric coordinate in each cluster. Black horizontal lines in (a), (c), and (d) indicate the location of calculated kmeans centroids. (e) Number of structures sorted into each cluster. Bars are colored by the dataset the structure originally came from. In panels (a-e), results are colored by the origin of the ATP structure in question: PDB structures (gray), water-box simulations (blue), in-cell simulations (not near macromolecules) (light green) and in-cell simulations (near macromolecules) (dark green). (f-h) Representative structures of the centroid of clusters 0, 1, 2, and 3 respectively. The representative ATP structures are shown in colored licorice representation. Nearby metabolites are shown in gray licorice. Surrounding macromolecules are shown as gray ribbon structures. K^+ ions are shown in pink, Na^+ ions are shown in yellow, and Mg^{2+} ions are shown in orange.

Clustering Analysis

To determine whether ATP had any characteristic structures in this dataset, we performed k-means clustering on the ATP structures using the Python package `pyclustering`.³⁰ Because the dihedral values were represented as angles between 0 and 360, we modified the Euclidean distance function in the package to reflect that the angles 0 and 360 are equivalent to avoid artifacts in the

clustering analysis. Due to its narrow and consistent distribution throughout all four local environments, the Φ_2 data (second column, Figure 3.2) was not used in the clustering analysis.

We found that four clusters best described the data (Figure 3.5). The clustering appears to primarily separate structures based on their Φ_1 and Φ_3 values and can be roughly described as combinations of low and high Φ_1 and Φ_3 values (clusters 0, 1, 2, and 3 are low-low, high-high, low-high, and high-low respectively). Unsurprisingly, the two clusters with centroids at higher Φ_1 values are almost entirely comprised of structures from the PDB dataset (PDB structures make up 90 and 86% of clusters 1 and 3 respectively).

CONCLUSIONS AND FUTURE WORK

Overall, the geometric and clustering results point to a distinct set of conformations that are only occupied by ATP when it is bound to certain proteins. However, the biological reason for these rare conformations is still unknown. We hypothesize that these proteins may have evolved to “lock” ATP into these unusual geometries for a variety of possible biological reasons. Studies have shown that ATP binds to proteins in a wider range of conformations when it is acting as an allosteric modulator than it does when it binds as an enzyme substrate.³¹ Additionally, it is well-established that enzymes place strain on their substrates in order to catalyze reactions.³²

In the future, we plan to investigate the proteins that bind ATP in the unusual, high- Φ_1 conformations we see in the PDB dataset. Any common features in these proteins (such as a common function or subcellular location) could indicate the biological roots of the conformations we observe in the PDB dataset.

Additionally, we plan to compare the geometries of ATP molecules in the *E. coli* models to those in the *H. sapiens* model. Preliminary results indicate that the lower ion concentrations in the HC36m (see Ch. 6) may help mediate the protein-protein sticking observed in the *E. coli* models (see Ch. 5). Because ATP is a highly charged molecule, we predict that it will be particularly sensitive to the electrostatics in its local environment and may adopt different conformations in the HC36m model.

REFERENCES

- (1) Theillet, F.-X.; Binolfi, A.; Frembgen-Kesner, T.; Hingorani, K.; Sarkar, M.; Kyne, C.; Li, C.; Crowley, P. B.; Gierasch, L.; Pielak, G. J.; Elcock, A. H.; Gershenson, A.; Selenko, P. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* 2014, 114 (13), 6661–6714. <https://doi.org/10.1021/cr400695p>.
- (2) Davis, C. M.; Gruebele, M. Non-Steric Interactions Predict the Trend and Steric Interactions the Offset of Protein Stability in Cells. *ChemPhysChem* 2018, 19 (18), 2290–2294. <https://doi.org/10.1002/cphc.201800534>.
- (3) Dhar, A.; Samiotakis, A.; Ebbinghaus, S.; Nienhaus, L.; Homouz, D.; Gruebele, M.; Cheung, M. S. Structure, Function, and Folding of Phosphoglycerate Kinase Are Strongly Perturbed by Macromolecular Crowding. *Proc. Natl. Acad. Sci. U. S. A.* 2010, 107 (41), 17586–17591. <https://doi.org/10.1073/pnas.1006760107>.
- (4) Feng, R.; Gruebele, M.; Davis, C. M. Quantifying Protein Dynamics and Stability in a Living Organism. *Nat. Commun.* 2019, 10 (1), 1179–1179. <https://doi.org/10.1038/s41467-019-09088-y>.
- (5) Wang, Y.; Sukenik, S.; Davis, C. M.; Gruebele, M. Cell Volume Controls Protein Stability and Compactness of the Unfolded State. *J. Phys. Chem. B* 2018, 122 (49), 11762–11770. <https://doi.org/10.1021/acs.jpcb.8b08216>.
- (6) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* 2020, 124 (23), 4732–4740. <https://doi.org/10.1021/acs.jpcb.0c02536>.
- (7) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* 2019, 10 (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (8) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in Escherichia Coli. *Proc. Natl. Acad. Sci.* 2017, 114 (23), E4556–E4563. <https://doi.org/10.1073/pnas.1621227114>.
- (9) Cohen, R. D.; Pielak, G. J. Electrostatic Contributions to Protein Quinary Structure. *J. Am. Chem. Soc.* 2016, 138 (40), 13139–13142. <https://doi.org/10.1021/jacs.6b07323>.
- (10) Miklos, A. C.; Sarkar, M.; Wang, Y.; Pielak, G. J. Protein Crowding Tunes Protein Stability. *J. Am. Chem. Soc.* 2011, 133 (18), 7116–7120. <https://doi.org/10.1021/ja200067p>.
- (11) Kasahara, K.; Re, S.; Nawrocki, G.; Oshima, H.; Mishima-Tsumagari, C.; Miyata-Yabuki, Y.; Kukimoto-Niino, M.; Yu, I.; Shirouzu, M.; Feig, M.; Sugita, Y. Reduced Efficacy of a Src Kinase Inhibitor in Crowded Protein Solution. *Nat. Commun.* 2021, 12 (1), 4099. <https://doi.org/10.1038/s41467-021-24349-5>.
- (12) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia Coli. *Nat. Chem. Biol.* 2009, 5 (8), 593–599. <https://doi.org/10.1038/nchembio.186>.
- (13) Park, J. O.; Rubin, S. A.; Xu, Y.-F.; Amador-Noguez, D.; Fan, J.; Shlomi, T.; Rabinowitz, J. D. Metabolite Concentrations, Fluxes and Free Energies Imply Efficient Enzyme Usage. *Nat. Chem. Biol.* 2016, 12 (7), 482–489. <https://doi.org/10.1038/nchembio.2077>.
- (14) Chen, W. W.; Freinkman, E.; Wang, T.; Birsoy, K.; Sabatini, D. M. Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism. *Cell* 2016, 166 (5), 1324–1337.e11. <https://doi.org/10.1016/j.cell.2016.07.040>.

- (15) Kamerlin, S. C. L.; Sharma, P. K.; Prasad, R. B.; Warshel, A. Why Nature Really Chose Phosphate. *Q. Rev. Biophys.* 2013, 46 (1), 1–132. <https://doi.org/10.1017/S0033583512000157>.
- (16) Patel, A.; Malinovska, L.; Saha, S.; Wang, J.; Alberti, S.; Krishnan, Y.; Hyman, A. A. ATP as a Biological Hydrotrope. *Science* 2017, 356 (6339), 753–756. <https://doi.org/10.1126/science.aaf6846>.
- (17) Sridharan, S.; Kurzawa, N.; Werner, T.; Günthner, I.; Helm, D.; Huber, W.; Bantscheff, M.; Savitski, M. M. Proteome-Wide Solubility and Thermal Stability Profiling Reveals Distinct Regulatory Roles for ATP. *Nat. Commun.* 2019, 10, 1155. <https://doi.org/10.1038/s41467-019-09107-y>.
- (18) Ou, X.; Lao, Y.; Xu, J.; Wutthinitikornkit, Y.; Shi, R.; Chen, X.; Li, J. ATP Can Efficiently Stabilize Protein through a Unique Mechanism. *JACS Au* 2021. <https://doi.org/10.1021/jacsau.1c00316>.
- (19) Song, J. Adenosine Triphosphate Energy-Independently Controls Protein Homeostasis with Unique Structure and Diverse Mechanisms. *Protein Sci.* 2021, 30 (7), 1277–1293. <https://doi.org/10.1002/pro.4079>.
- (20) Kobayashi, E.; Yura, K.; Nagai, Y. Distinct Conformation of ATP Molecule in Solution and on Protein. *Biophys. Nagoya-Shi Jpn.* 2013, 9, 1–12. <https://doi.org/10.2142/biophysics.9.1>.
- (21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* 2000, 28 (1), 235–242. <https://doi.org/10.1093/nar/28.1.235>.
- (22) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* 1996, 14 (1), 33–38, 27–28. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (23) Pavelites, J. J.; Gao, J.; Bash, P. A.; Mackerell Jr., A. D. A Molecular Mechanics Force Field for NAD⁺ NADH, and the Pyrophosphate Groups of Nucleotides. *J. Comput. Chem.* 1997, 18 (2), 221–239. [https://doi.org/10.1002/\(SICI\)1096-987X\(19970130\)18:2<221::AID-JCC7>3.0.CO;2-X](https://doi.org/10.1002/(SICI)1096-987X(19970130)18:2<221::AID-JCC7>3.0.CO;2-X).
- (24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 1983, 79 (2), 926–935. <https://doi.org/10.1063/1.445869>.
- (25) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* 2020, 153 (4), 044130. <https://doi.org/10.1063/5.0014475>.
- (26) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. Measuring the Rate of Intramolecular Contact Formation in Polypeptides. *Proc. Natl. Acad. Sci.* 2000, 97 (13), 7220–7225. <https://doi.org/10.1073/pnas.97.13.7220>.
- (27) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-PCM Solvation Model. *J. Comput. Chem.* 2003, 24 (6), 669–681. <https://doi.org/10.1002/jcc.10189>.
- (28) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* 1994, 98 (45), 11623–11627. <https://doi.org/10.1021/j100096a001>.
- (29) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. 6-31G* Basis Set for Third-Row Atoms. *J. Comput. Chem.* 2001, 22 (9), 976–984. <https://doi.org/10.1002/jcc.1058>.

- (30) Novikov, A. V. PyClustering: Data Mining Library. J. Open Source Softw. 2019, 4 (36), 1230. <https://doi.org/10.21105/joss.01230>.
- (31) Lu, S.; Huang, W.; Wang, Q.; Shen, Q.; Li, S.; Nussinov, R.; Zhang, J. The Structural Basis of ATP as an Allosteric Modulator. PLoS Comput. Biol. 2014, 10 (9), e1003831. <https://doi.org/10.1371/journal.pcbi.1003831>.
- (32) Pauling, L. Nature of Forces between Large Molecules of Biological Interest*. Nature 1948, 161 (4097), 707–709. <https://doi.org/10.1038/161707a0>.

CHAPTER 4: CROWDING, STICKING, AND PARTIAL FOLDING OF GTT WW DOMAIN IN A SMALL *E. COLI* CYTOPLASM MODEL

INTRODUCTION

In vitro experiments,² theory,³ and computational models⁴ have shown that protein folding free energy landscapes are shallow. That is, the free energy differences between folded and denatured states and activation free energies are relatively small (0-40 kJ/mole) over the physiological temperature range. Consequently most proteins fold relatively quickly (microseconds to minutes),⁵ and their stability is susceptible to changes in solvation conditions including tuning of the temperature,⁶ pressure,⁷ crowding,⁸ ionic strength, acidity,⁹ chao-/kosmotopes^{10,11} and other solvation parameters.

The native solvation environment for many proteins, the cytoplasm, is no exception to this rule, and experiments have reported cell type,¹² cell cycle,¹³ and site-to-site dependence¹⁴ of protein stability and folding kinetics in the cytoplasm, although the general folding mechanism remains unchanged. In cells, proteins are subject to crowding (excluded volume)¹⁵ and sticking interactions¹⁶ in addition to quinary structure (weak functional interactions).¹⁷ For example, a human homolog of a bacterial protein sticks in bacterial cells, while the bacterial protein diffuses

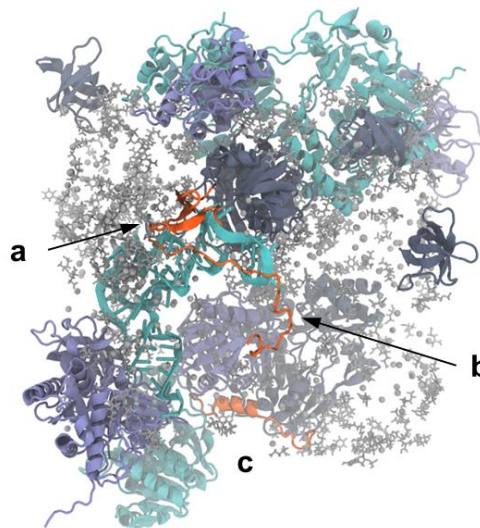


Figure 4.1: Overall view of the EC36mCU simulation after ~4.99 μ s of all-atom simulation, showing three states of GTT in shades of orange. (a) native structure with $Q=0.93$. (b) highly extended unfolded structure with $Q=0.01$. (c) unfolded with non-native helical structure, $Q=0.03$. Other macromolecules in the model are illustrated in blue/green and metabolites and inorganic ions are shown in gray. Figure published in ref. 1.

* This chapter is reproduced in part with permission from:

Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, 124 (23), 4732–4740.¹

more freely, due to differences in electrostatics and surface hydrophobicity.¹⁸

Here, we examine how a small (~16 proteins, ~262 small solutes, ~430 ions and ~50,000 water molecules) all-atom model of the *E. coli* cytoplasm¹⁹ affects the folding of the 33-residue GTT WW domain compared to similar simulations of GTT in aqueous solution.²⁰ The now > 200 μ s simulation occupies a ‘protein-folding niche’ between long-time Brownian dynamics simulations²¹ and short-time all-atom simulations of larger cytoplasmic models.²²

GTT is a triple-stranded β -sheet protein derived from the FiP mutant of the signaling module WW domain,²³ with hairpin 1 formed by strands 1+2, and hairpin 2 formed by strands 2+3 (Figure 4.1). In *in vitro* experiments, GTT folds in ~4.3 μ s (at 353 K) and slightly faster using the CHARMM22* force field, via a mechanism that forms hairpin 1 first *ca.* two-thirds of the time, and hairpin 2 *ca.* one-third of the time.²⁴ Thus GTT is an ideal testbed for simulating the effect of cytoplasmic crowding and sticking on folding dynamics.

We simulated two initially folded and seven initially unfolded copies of GTT in our cytoplasm model using three versions of the CHARMM force field: CHARMM22* (C22*), CHARMM36m (C36m), and CHARMM36m with CUFIX corrections (C36mCU).^{25–27} These force fields were selected due to their propensity to optimize folding (C22*) or reduce ‘sticking’ interactions between the protein and its surrounding environment (C36m and C36mCU). In all simulations, the folded copies remained folded. None of the unfolded copies completely folded despite an overall simulation time at least twenty times longer than the expected folding time at the temperature, although native-like structure formed in hairpins 1 and 2 several times during the simulation. Although the optimization of current force fields for folding of isolated small proteins leads them to overestimate macromolecular sticking, a number of general questions can be asked. Here we focus on three: How do crowding and sticking slow down folding dynamics of unfolded GTT and affect fluctuations of the GTT native state? How disruptive can sticking interactions in the cytoplasm become? And what is the role of in-cell water in mediating native-like contact formation?

ANALYSIS METHODS

For details on model construction, please see Chapter 2

The secondary structure of GTT was calculated every 24 ns using the `compute_dssp` function in MDTraj.^{28,29} The RMSD of GTT 2 in the EC36mCU model was calculated using

VMD's RMSD trajectory tool.³⁰ To determine the buried surface area of GTT, the solvent accessible surface area (SASA) of the protein alone was compared to the SASA of the protein in the cytoplasmic environment (see Code C.1). SASA calculations were performed in VMD with a radius of the sphere set to the radius of a water molecule set, 1.4 Å.

RESULTS AND DISCUSSION

Overall simulation

We ran three simulations of *ca.* 200,000 atoms at 318 K, with each with a macromolecular concentration of 300 mg/mL. The EC22* simulation of ≈ 190 μ s started with three unfolded states, and the EC36m and EC36mCU simulations of ≈ 25 μ s with one folded and two unfolded GTT WW domains each. In none of the simulations did GTT completely fold, although formation of paired strands 1-2 and 2-3 was observed at least 5 times (Figure B.1-Figure B.3). The two folded WW domains also did not unfold for 25 μ s duration of their simulations. The stability of other proteins in the simulation varied, with C α -RMSD ranging from about 2 to 10 Å, and coldshock (Csp) proteins particularly rigid with small RMSD (Figure B.4).

Three typical structures of GTT are shown in a snapshot of the EC36mCU simulation in Figure 4.1 (WW domains in orange): native (top), highly extended (middle) and non-native helical residual structure (bottom). They were also observed in aqueous solution simulations where GTT folded (see below). Although this propensity may be exaggerated by C22*, it is plausible based on the very high helical propensity as evidence by Agadir³¹ scores of residues W11 through M15 (numbering as Figure 4.6g, scores in Table B.1).

GTT in aqueous solution

A long simulation of GTT in aqueous solution on Anton 2 using C22* has previously shown that GTT folds in 10 ± 2 μ s at 395 K, near the melting temperature of GTT with 22*-TIP3P.²⁰ This is similar to the experimentally measured mean refolding time of 5.7 μ s at 363 K.²⁰ To check that a similar time scale can be obtained at our lower simulation temperature of 318 K, well below T_m , we simulated a 25 μ s trajectory using C22* (Figure 4.2). GTT showed large (range 5-20 Å) RMSD fluctuations in the unfolded state, and after populating a short-lived intermediate with a non-native hairpin 1 at 9 μ s, it formed hairpin 2 and within 2 μ s the native state at 19 μ s. Thus GTT can fold on a time scale almost as fast as in the higher temperature simulations, consistent

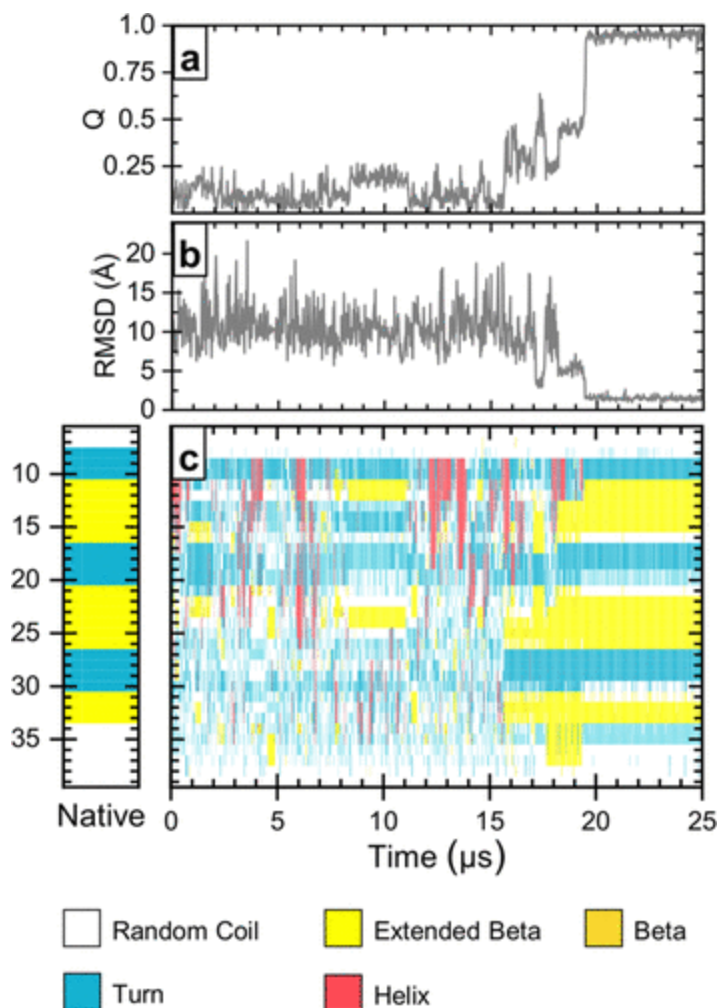


Figure 4.2: CHARMM22* (C22*) trajectory of GTT in aqueous solution at 318 K, showing a complete refolding event in less than 20 μ s, consistent with ref (20). (a) $Q > 0.85$ in the folded state. (b) RMSD undergoes fluctuations between 5 and 20 Å before settling to <3 Å upon folding. (c) Secondary structure involves fluctuating helix especially between residues 3–10 (strand 1). Turn 1 forms temporarily at 8 μ s, until folding is initiated by formation of turn 2 at >15 μ s and subsequent formation of a native-like turn 1 at 20 μ s. Figure published in ref. 1.

with the prediction of faster pre-factors for WW domain offsetting higher barriers at low temperature.³²

The formation of transient helical structure itself is not inconsistent with complete refolding. Figure 4.2 reveals the formation and melting of non-native helix multiple times, generally for < 1 μ s. The N-terminal β -strand is most susceptible to forming non-native helix, as seen in the overall simulation of the cytoplasm model.

Secondary structure of GTT in the cytoplasm model

We begin by examining the dynamics of GTT in the cytoplasm model via secondary structure. Figure 4.3 highlights one 5 μ s region in the trajectories for each GTT WW domain where significant (if any) changes in secondary structure were observed. The full secondary structure trajectories for the EC22*, EC36m and EC36mCU simulations are shown in Figure B.1-

Figure B.3. In four of the seven non-

native WW domains non-helical structure persisted throughout the entire simulation, in contrast to Figure 4.2 and previous simulations of WW domain subject to pressure or temperature perturbations, where transient helical structure lasts for less than a few μ s.³³ As expected from the secondary structure propensity using Agadir,³¹ all but one of the four GTTs with helical structure throughout show long-lived helix near the N-terminus. Therefore, the cytoplasmic environment

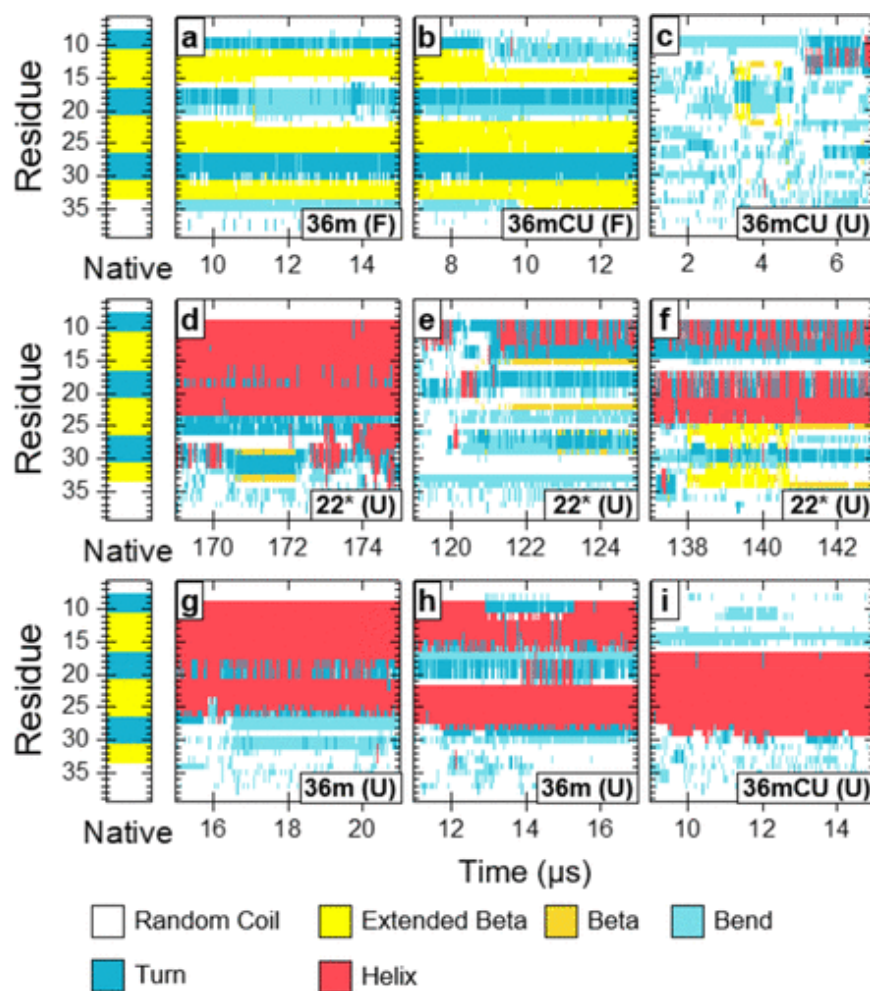


Figure 4.3: Secondary structure of GTT in the cytoplasm models throughout trajectories as calculated by DSSP.^{28,29} The secondary structure at each residue of the native structure is given at the left. Each panel (a–i) represents a 6 μ s slice of each GTT copy’s trajectory. The lower right of each panel indicates the force field used for the simulation and whether the protein started in the folded (F) or unfolded (U) state. Partial β -hairpin formation was observed in several of the unfolded trajectories (c–f) but was frequently register shifted. Secondary structure calculations of the full trajectories can be found in Figure B.1–Figure B.3. Figure published in ref. 1.

enhances helix propensity of WW domain relative to aqueous solution, and this contributes to the reduced folding rate.

The region of turns 1 and 2 is slightly less likely to form non-native helical structure than the β -sheet regions. This is particularly evident in Figure B.2 of GTT2 and GTT3 in the EC36m simulation, where turn 1 (residues 16–20) frequently switches from helical to random coil or turn structure, whereas the adjacent sheet regions remain helical. This is consistent with nucleation of the native state occurring at turns, and not in the β -strands, as will be discussed later in conjunction with desolvation during turn formation.

Compactness of GTT in aqueous solution vs. cytoplasm

In the aqueous solution simulation, GTT rapidly explores radii of gyration greater than 20 Å and then swiftly collapses to near-native radius of gyration (Figure 4.4). This propensity for fast collapse to compact states has been seen in single molecule experiments on unfolded proteins under native conditions.³⁴ In contrast, one might expect that GTT rarely explores states of high

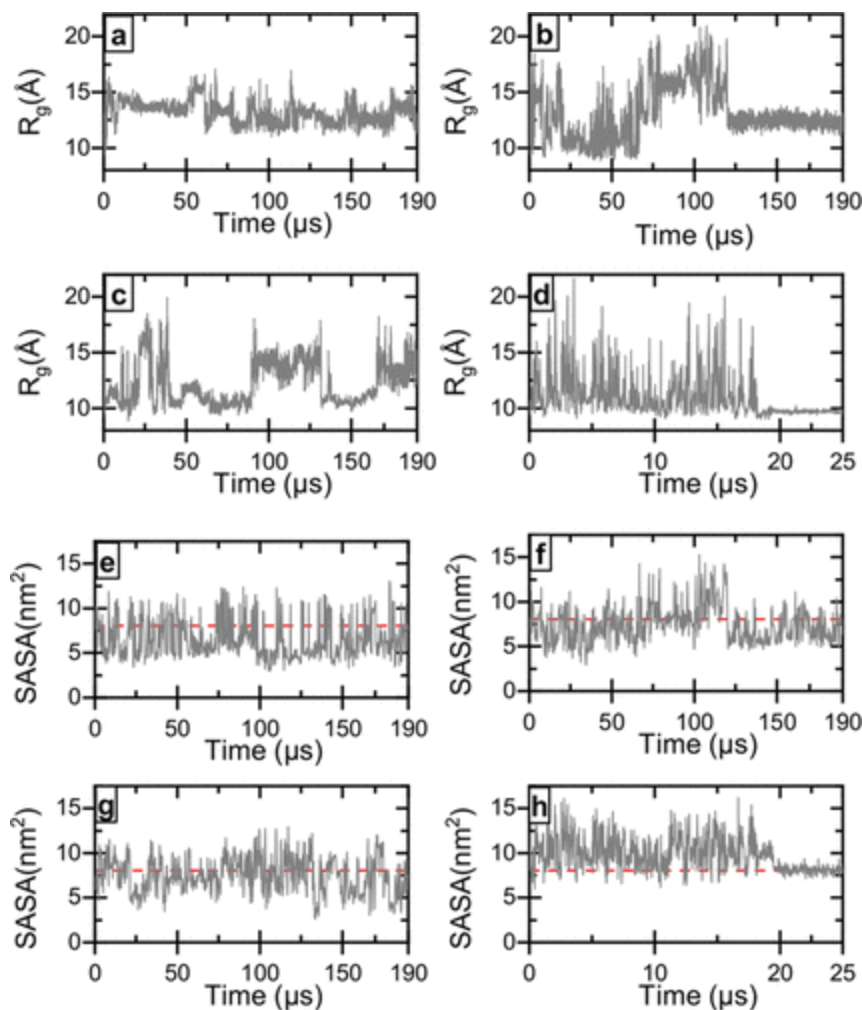


Figure 4.4: Top 4 panels: Radius of gyration R_g of GTT in the CHARM22* (C22*) cytoplasm model (a–c) and in water (d). In water, the protein makes excursions between native-like R_g and >20 Å while in the unfolded state. In the cytoplasm model, R_g fluctuates much less in the unfolded state (GTT1 in panel a), switches between long-lived states of similar compactness (GTT2 in panel b), and can end up in very long-lived metastable states due to sticking to the surface of other macromolecules (see Figure 4.5). Bottom 4 panels: In the aqueous solution C22* simulation, the exposed hydrophobic surface area of GTT (h) is almost always larger than in the native state (red dotted line). In the C22* cytoplasm model, unfolded GTT1-3 (e–g) most often exposes even less hydrophobic area than aqueous folded GTT due to burial against other macromolecules (see Figure 4.5). Figure published in ref. 1.

radius of gyration in the crowded cytoplasmic simulations. This is indeed the case, with GTT frequently getting stuck in states with $R_g < 15$ Å. (Figure 4.4).

Two main differences in the dynamics of unfolded GTT in the cytoplasm and buffer are evident. First of all, whereas large fluctuations occur in aqueous solution on a timescale < 100 ns, in the three EC22* cytoplasmic copies the protein frequently spends many microseconds in a state with a narrow range of radius of gyration, before switching to another such state. The residence time of GTT3 in Figure 4.4 for example is greater than 25 μ s in three cases, and GTT2 gets stuck in a state with nearly constant R_g for over 70 μ s. GTT1 experiences a highly reduced dynamic range of R_g and several long residence times also. The slow time evolution of R_g goes hand in hand with reduced translational diffusion through the cytoplasm, averaging 3.0 nm²/ μ s in aqueous buffer vs. 0.03 nm²/ μ s for the cytoplasmic copies (Figure B.5), as determined by fitting the slope of the mean squared displacements vs. time of the MD simulation trajectories. Previous experimental measurements in *E. coli* show a slowdown closer to a factor of 10 than a factor of 100,^{35,36} indicating that our cytoplasm models overestimate the stickiness of the cytoplasm for GTT.

Secondly, the cytoplasmic GTTs have far less solvent-exposed surface area (Figure 4.4, bottom) in the EC22* simulation, routinely falling below the native state average of 7.5 nm² in aqueous solution. This observation is consistent with reduced diffusion due to macromolecular sticking and crowding. Both super-resolution displacement mapping of mammalian cell³⁷ and bleaching measurements in mammalian cells³⁸ have revealed heterogeneous diffusion consistent with the intermittent sticking and surface area burial seen in the bacterial cytoplasm model in Figure 4.4.

Sticking of GTT to nearby macromolecules

The retarded dynamics and lack of exposed surface area of GTT in the cytoplasmic simulations are strong evidence that GTT is not only crowded (excluded volume) but sticks (forms weak surface bonds) to other proteins. We analyzed such sticking in a shorter version of the simulation previously for a series of proteins.¹⁹ Here we consider two examples of how GTT interacts with the surrounding cytoplasmic matrix.

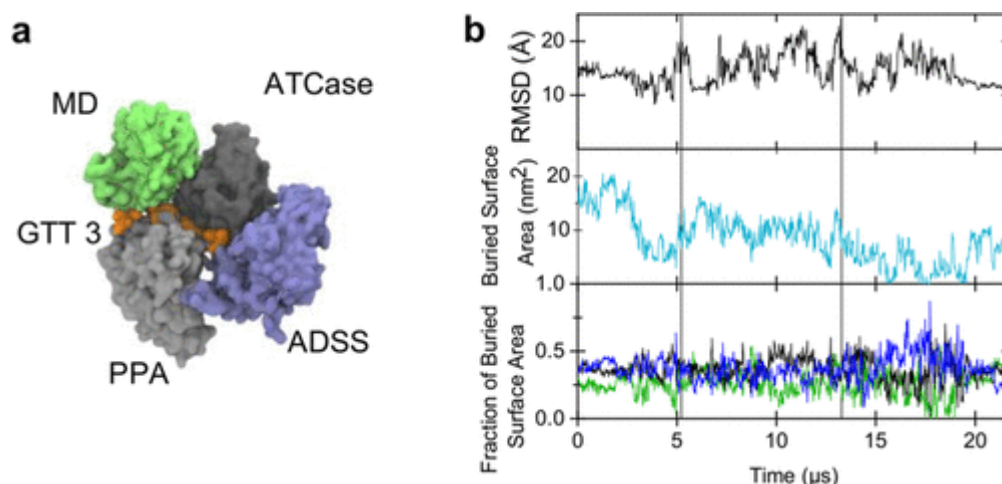


Figure 4.5: (a) Snapshot of GTT3 buried by surrounding macromolecules in the CHARMM36mCU (C36mCU) trajectory. (b) RMSD of GTT2 in the C36mCU trajectory, surface area of GTT that is buried by protein-protein contacts throughout the C36mCU trajectory. (c) Contribution of different residue types: polar (green), hydrophobic (black), and charged (blue) to the buried surface area. Figure published in ref. 1.

Figure 4.5a shows GTT3 from the EC36mCU simulation buried in a matrix of four other cytoplasmic proteins. Another copy's (GTT2) surface area also is buried by other proteins (Figure 4.5b) and varies widely from 20 nm² to near-zero, showing that the simulation is long enough to sample release and reburial on a time scale of several microseconds. A somewhat shorter time scale for lifetime of protein-protein contacts, following a power law distribution with a broad range from 1 ns to 1 μs has been observed for an average over many fully folded proteins.¹⁹ As discussed above, RMSD is able to range to its highest values only when the protein is less buried, so the buried states tend to be more compact. Figure 4.5b also decomposes the buried area into hydrophobic, charged and polar interactions. For the CHARMM force fields used here, hydrophobic, polar and charged patches make a similar contribution to GTT-protein contacts. For a larger protein analyzed previously hydrophobic contacts were more dominant.¹⁹

Chimeric interaction between GTT and other proteins

Cooperativity between interacting proteins in our model cytoplasm can be even more extreme: the intermingling of unfolded GTT with another proteins to form intramolecular tertiary structure, which we will refer to as a 'chimeric protein,' in analogy to genetically chimeric proteins where proteins originating from different genes are linked.⁴³ Such chimeras have been observed experimentally for the WW domain, when monomers are linked together so they can interact frequently, as would be the case with proteins in the cytoplasm.⁴⁴

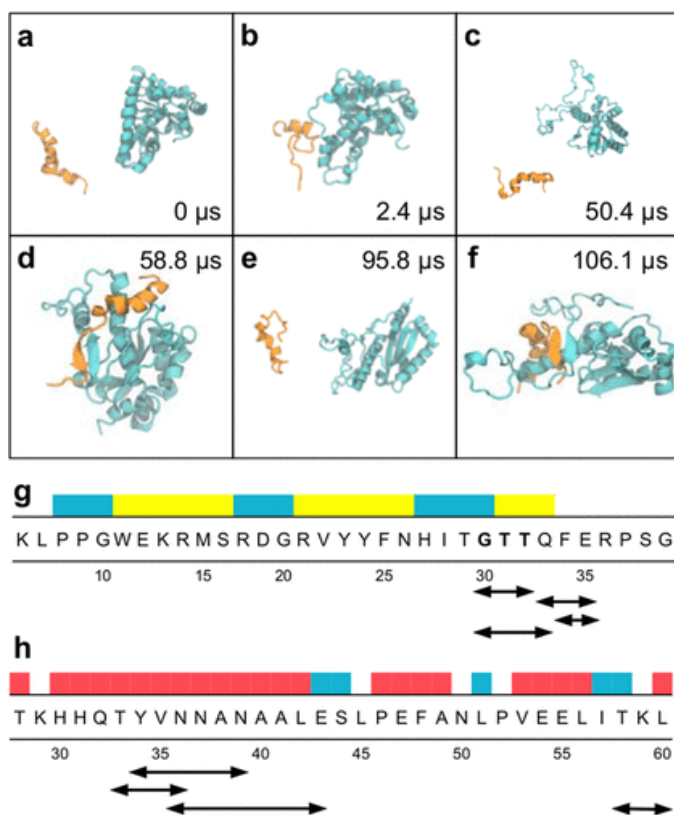


Figure 4.6: Chimeric interactions between GTT1 (orange) and Mn SOD (blue) in the EC22* trajectory. (a) The proteins were not in contact at the beginning of the simulation. (b) GTT forms a contact with Mn SOD and unfolds Mn SOD's α -helix (residues 30–42). (c–f) GTT and Mn SOD continue to form and break protein–protein contacts including 1 antiparallel chimeric interaction (d) and multiple parallel chimeric interactions (f). These chimeric interactions can be remarkably long-lived. The interaction shown in part d lasted for $\sim 4.44 \mu\text{s}$. The complete sequence of GTT (g) and a partial sequence of Mn SOD (h) are shown with their secondary structures indicated. The residues that form chimeric interactions during each event are indicated with black arrows below the respective sequences. Figure published in ref. 1.

GTT mutation that distinguishes GTT from wild-type human Pin1 WW domain.²³

Another such interaction occurred in the EC22* simulation between Mdh and GTT2. Thus, chimeric interactions are not rare in this force field. We did not observe them in the albeit shorter EC36m and EC36mCU simulation (50 μs total), indicating that these force fields may be less sticky. However, such events may present an extreme form of protein-protein interaction in the cytoplasm, and such non-native intramolecular β -sheet structure could eventually resemble pre-amyloid oligomers.

Figure 4.6 shows a specific example where GTT1 interacts with superoxide dismutase (SodA). In (b), the proteins make contact and a long helix in SodA partly unfolds. In (c), GTT has diffused away again. In (d), GTT and the unfolded region interact to form an intermolecular β -sheet composed of two strands (shown in orange and cyan). In (d), they have diffused away again and the β -sheet structure is lost. Finally, in (f), they associate again and the same intermolecular β -sheet strand forms again. Overall, an initially transiently populated helical region in one protein (GTT) interacts with part of a helix in another protein (SodA) to form β -sheet structure. The β -sheet structure is very dynamical and lasted anywhere from several ns to several μs . It was mainly governed by hydrogen bonds between polar residues 30 to 35 in GTT1 (black arrows in Figure 4.6g), with an Arginine-rich region in SodA (Figure 4.6h). This happens to be the region of the

Drying transition of GTT during partial folding in cytoplasm model

Hairpin formation between either strand 1 and 2, or 2 and 3, was observed several times throughout the three simulations. Figure 4.7 illustrates an example of the forward process for loop 2 formation (desolvation and hairpin formation) from the EC36mCU simulation in Figure 4.3c, and Figure B.6 illustrates another example of the reverse process (solvation to open the hairpin).

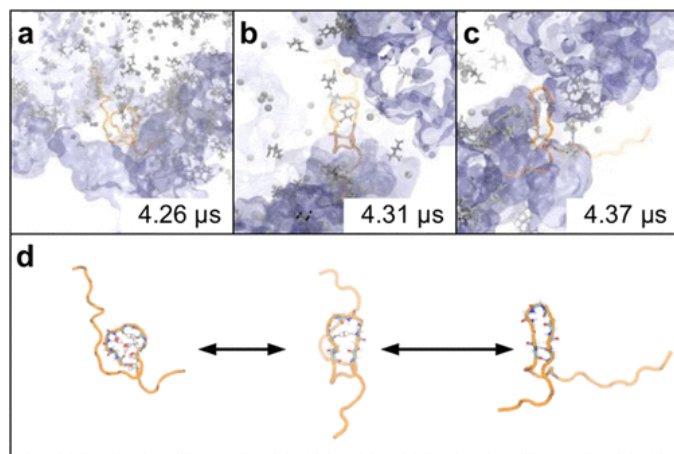


Figure 4.7: Representative snapshots of turn hydration during the β -hairpin formation of GTT2 (U) in the EC36mCU simulation. We observe a potential transition state that involves a single water bridging the backbone of the nascent turn (b). This state in a few observed cases either immediately precedes or immediately follows hairpin formation (c), or further solvation of the turn (a) (also see Figure B.6). Panel d shows the same three states for clarity without the surrounding proteins, ions, and metabolites and highlights the bridging water molecules. Figure published in ref. 1.

Among the three cases of GTT2's partial hairpin formation and break-up in the EC36mCU model, some are more native-like, while others are one residue out of registration. In all cases, we observe a state with a single water molecule intercalating the loop, either just before the 'dry' hairpin forms, or just before additional water molecules solvate the loop while it breaks up. Five samples are not enough for a full transition path analysis, but we propose that the loop intercalated by a single water molecule lies close to the transition state ensemble for loop formation. In all cases the less solvated

loop forms before the β -strands interact, and a drying transition that removes the final intercalating water molecule allows the two-stranded β -sheet to form. This 'loop first, then zip up the sheet' is consistent with *in vitro* measurements and models that residues near the turn have the largest effect on folding/unfolding kinetics.^{45,46}

CONCLUSION

We find that the C22*, C36m and C36mCU force fields all reduce the in-cell folding rate of GTT WW domain, at least by a factor of *ca.* 10 for C22* relative to aqueous buffer. The main reason is sticking of the unfolded protein against other macromolecules in the simulations. In these force fields, the contribution of hydrophobic interactions at protein-protein interfaces is significant.

Slow refolding is corroborated by slow diffusion, slow intermittent changes in radius of gyration and surface burial, formation of salt bridges, H-bonds and hydrophobic contacts between proteins, and even formation of chimeric structures. Experimental observations of a few larger proteins have revealed in-cell folding rates comparable to *in vitro*,⁴⁷ so the force fields are likely to exacerbate sticking. This should not come as a surprise because they were modified from C22* to simulate folding of a single protein in aqueous buffer, where enhanced sticking of a protein (to itself) may enhance collapse and folding. Although sticking in the cell is likely not as severe as in the simulations, sticking will slow down in-cell folding, whereas crowding reduces available volume only by a factor of ca. 2, and has a similarly small effect on folding *in vitro* vs. in-cell or *in vivo*.^{12,48}

Another explanation for the sticking of GTT is that the region of the three mutations made to speed up folding (residues 30-32 in Figure 4.6h) is strongly involved in chimeric structure between GTT and SodA. Thus, mutating surface-exposed residues away from the wild-type sequence could enhance sticking. Two types of experimental results support the importance of surface residues to avoid in-cell sticking, and hence co-evolution of protein surfaces with their cellular environment.⁴⁹ Firstly, in-cell NMR studies have shown that protein surface composition of human vs. bacterial proteins strongly influences free vs. hindered rotational diffusion in bacterial cells.¹⁸ Secondly, comparison of an intracellular eukaryotic protein (PGK) vs. an extracellular prokaryotic protein (VISE) shows that the latter folds more slowly in mammalian cells.

Finally, it is interesting to note that formation of chimeric proteins with an intermolecular β -sheet (Figure 4.6) is a mechanism by which an unfolded protein, here GTT, can initiate formation of β -rich protein oligomers. Such strong sticking could eventually become a nucleus for in-cell formation of amyloid structure, although we did not observe that on the time scale of our simulations.

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcc.0c02536>.
- (2) Yang, W. Y.; Gruebele, M. Folding at the Speed Limit. *Nature* **2003**, *423* (6936), 193–197. <https://doi.org/10.1038/nature01609>.
- (3) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. Toward an Outline of the Topography of a Realistic Protein-Folding Funnel. *PNAS* **1995**, *92* (8), 3626–3630. <https://doi.org/10.1073/pnas.92.8.3626>.
- (4) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. Force Field Bias in Protein Folding Simulations. *Biophysical Journal* **2009**, *96* (9), 3772–3780. <https://doi.org/10.1016/j.bpj.2009.02.033>.
- (5) Nölting, B.; Golbik, R.; Fersht, A. R. Submillisecond Events in Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America* **1995**, *92* (23), 10668–10672. <https://doi.org/10.1073/pnas.92.23.10668>.
- (6) Griko, Y. V.; Privalov, P. L. Calorimetric Study of the Heat and Cold Denaturation of Beta-Lactoglobulin. *Biochemistry* **1992**, *31* (37), 8810–8815. <https://doi.org/10.1021/bi00152a017>.
- (7) Hawley, S. A. Reversible Pressure–Temperature Denaturation of Chymotrypsinogen. *Biochemistry* **1971**, *10* (13), 2436–2442. <https://doi.org/10.1021/bi00789a002>.
- (8) Zimmerman, S. B.; Minton, A. P. Macromolecular Crowding: Biochemical, Biophysical, and Physiological Consequences. *Annual Review of Biophysics and Biomolecular Structure* **1993**, *22*, 27–65. <https://doi.org/10.1146/annurev.bb.22.060193.000331>.
- (9) Brazhnikov, E. V.; Chirgadze YuN, null; Dolgikh, D. A.; Ptitsyn, O. B. Noncooperative Temperature Melting of a Globular Protein without Specific Tertiary Structure: Acid Form of Bovine Carbonic Anhydrase B. *Biopolymers* **1985**, *24* (10), 1899–1907. <https://doi.org/10.1002/bip.360241005>.
- (10) Putnam, F. W.; Erickson, J. O.; Volkin, E.; Neurath, H. NATIVE AND REGENERATED BOVINE ALBUMIN: I. PREPARATION AND PHYSICOCHEMICAL PROPERTIES. *The Journal of General Physiology* **1943**, *26* (6), 513–531. <https://doi.org/10.1085/jgp.26.6.513>.
- (11) Washabaugh, M. W.; Collins, K. D. The Systematic Characterization by Aqueous Column Chromatography of Solutes Which Affect Protein Stability. *The Journal of Biological Chemistry* **1986**, *261* (27), 12477–12485.
- (12) Feng, R.; Gruebele, M.; Davis, C. M. Quantifying Protein Dynamics and Stability in a Living Organism. *Nature Communications* **2019**, *10* (1), 1179–1179. <https://doi.org/10.1038/s41467-019-09088-y>.
- (13) Wirth, A. J.; Platkov, M.; Gruebele, M. Temporal Variation of a Protein Folding Energy Landscape in the Cell. *Journal of the American Chemical Society* **2013**, *135* (51), 19215–19221. <https://doi.org/10.1021/ja4087165>.
- (14) Ebbinghaus, S.; Gruebele, M. Protein Folding Landscapes in the Living Cell. *The Journal of Physical Chemistry Letters* **2011**, *2* (4), 314–319. <https://doi.org/10.1021/jz101729z>.
- (15) van den Berg, B.; Ellis, R. J.; Dobson, C. M. Effects of Macromolecular Crowding on Protein Folding and Aggregation. *The EMBO journal* **1999**, *18* (24), 6927–6933. <https://doi.org/10.1093/emboj/18.24.6927>.

- (16) Chen, T.; Dave, K.; Gruebele, M. Pressure- and Heat-Induced Protein Unfolding in Bacterial Cells: Crowding vs. Sticking. *FEBS letters* **2018**, 592 (8), 1357–1365. <https://doi.org/10.1002/1873-3468.13025>.
- (17) McConkey, E. H. Molecular Evolution, Intracellular Organization, and the Quinary Structure of Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1982**, 79 (10), 3236–3240. <https://doi.org/10.1073/pnas.79.10.3236>.
- (18) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in *Escherichia Coli*. *Proceedings of the National Academy of Sciences* **2017**, 114 (23), E4556–E4563. <https://doi.org/10.1073/pnas.1621227114>.
- (19) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *The Journal of Physical Chemistry Letters* **2019**, 10 (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (20) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β -Sheet Protein. *Journal of Molecular Biology* **2011**, 405 (1), 43–48. <https://doi.org/10.1016/j.jmb.2010.10.023>.
- (21) McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS computational biology* **2010**, 6 (3), e1000694. <https://doi.org/10.1371/journal.pcbi.1000694>.
- (22) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm. *eLife* **2016**, 5. <https://doi.org/10.7554/eLife.19274>.
- (23) Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. Structure-Function-Folding Relationship in a WW Domain. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, 103 (28), 10648–10653. <https://doi.org/10.1073/pnas.0600511103>.
- (24) Davis, C. M.; Zanetti-Polzi, L.; Gruebele, M.; Amadei, A.; Dyer, R. B.; Daidone, I. A Quantitative Connection of Experimental and Simulated Folding Landscapes by Vibrational Spectroscopy. *Chemical Science* **2018**, 9 (48), 9002–9011. <https://doi.org/10.1039/C8SC03786H>.
- (25) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal* **2011**, 100 (9), L47-49. <https://doi.org/10.1016/j.bpj.2011.03.051>.
- (26) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nature Methods* **2017**, 14 (1), 71–73. <https://doi.org/10.1038/nmeth.4067>.
- (27) Yoo, J.; Aksimentiev, A. New Tricks for Old Dogs: Improving the Accuracy of Biomolecular Force Fields by Pair-Specific Corrections to Non-Bonded Interactions. *Phys. Chem. Chem. Phys.* **2018**, 20 (13), 8432–8449. <https://doi.org/10.1039/C7CP08185E>.
- (28) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, 109 (8), 1528–1532. <https://doi.org/10.1016/j.bpj.2015.08.015>.

- (29) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, 22 (12), 2577–2637. <https://doi.org/10.1002/bip.360221211>.
- (30) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, 14 (1), 33–38, 27–28. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (31) Lacroix, E.; Viguera, A. R.; Serrano, L. Elucidating the Folding Problem of Alpha-Helices: Local Motifs, Long-Range Electrostatics, Ionic-Strength Dependence and Prediction of NMR Parameters. *Journal of Molecular Biology* **1998**, 284 (1), 173–191. <https://doi.org/10.1006/jmbi.1998.2145>.
- (32) Liu, F.; Nakaema, M.; Gruebele, M. The Transition State Transit Time of WW Domain Folding Is Controlled by Energy Landscape Roughness. *The Journal of Chemical Physics* **2009**, 131 (19), 195101. <https://doi.org/10.1063/1.3262489>.
- (33) Wirth, A. J.; Liu, Y.; Prigozhin, M. B.; Schulten, K.; Gruebele, M. Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations. *Journal of the American Chemical Society* **2015**, 137 (22), 7152–7159. <https://doi.org/10.1021/jacs.5b02474>.
- (34) Lipman, E. A. Single-Molecule Measurement of Protein Folding Kinetics. *Science* **2003**, 301 (5637), 1233–1235. <https://doi.org/10.1126/science.1085399>.
- (35) Elowitz, M. B.; Surette, M. G.; Wolf, P. E.; Stock, J. B.; Leibler, S. Protein Mobility in the Cytoplasm of Escherichia Coli. *Journal of Bacteriology* **1999**, 181 (1), 197–203.
- (36) Mullineaux, C. W.; Nenninger, A.; Ray, N.; Robinson, C. Diffusion of Green Fluorescent Protein in Three Cell Environments in Escherichia Coli. *Journal of Bacteriology* **2006**, 188 (10), 3442–3448. <https://doi.org/10.1128/JB.188.10.3442-3448.2006>.
- (37) Xiang, L.; Chen, K.; Yan, R.; Li, W.; Xu, K. *Super-Resolution Displacement Mapping of Unbound Single Molecules Reveals Nanoscale Heterogeneities in Intracellular Diffusivity*; preprint; Biophysics, 2019. <https://doi.org/10.1101/559484>.
- (38) Guo, M.; Gelman, H.; Gruebele, M. Coupled Protein Diffusion and Folding in the Cell. *PloS One* **2014**, 9 (12), e113040. <https://doi.org/10.1371/journal.pone.0113040>.
- (39) Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. Evaluating the Strength of Salt Bridges: A Comparison of Current Biomolecular Force Fields. *Journal of Physical Chemistry B* **2014**, 118 (24), 6561–6569. <https://doi.org/10.1021/jp500958r>.
- (40) Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T. Further along the Road Less Traveled: AMBER Ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **2016**, 12 (8), 3926–3947. <https://doi.org/10.1021/acs.jctc.6b00567>.
- (41) Nawrocki, G.; Wang, P.-H.; Yu, I.; Sugita, Y.; Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. **2017**. <https://doi.org/10.1021/acs.jpcb.7b08785>.
- (42) Bülow, S. von; Siggel, M.; Linke, M.; Hummer, G. Dynamic Cluster Formation Determines Viscosity and Diffusion in Dense Protein Solutions. *Proceedings of the National Academy of Sciences* **2019**, 201817564–201817564. <https://doi.org/10.1073/PNAS.1817564116>.
- (43) Campbell, R. K.; Bergert, E. R.; Wang, Y.; Morris, J. C.; Moyle, W. R. Chimeric Proteins Can Exceed the Sum of Their Parts: Implications for Evolution and Protein Design. *Nature Biotechnology* **1997**, 15 (5), 439–443. <https://doi.org/10.1038/nbt0597-439>.

- (44) Dave, K.; Gasic, A. G.; Cheung, M. S.; Gruebele, M. Competition of Individual Domain Folding with Inter-Domain Interaction in WW Domain Engineered Repeat Proteins. *Physical Chemistry Chemical Physics* **2019**, 21 (44), 24393–24405. <https://doi.org/10.1039/C8CP07775D>.
- (45) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. Folding Dynamics and Mechanism of Beta-Hairpin Formation. *Nature* **1997**, 390 (6656), 196–199. <https://doi.org/10.1038/36626>.
- (46) Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. A Statistical Mechanical Model for Beta-Hairpin Kinetics. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, 95 (11), 5872–5879. <https://doi.org/10.1073/pnas.95.11.5872>.
- (47) Gelman, H.; Wirth, A. J.; Gruebele, M. ReAsH as a Quantitative Probe of In-Cell Protein Dynamics. *Biochemistry* **2016**, 55 (13), 1968–1976. <https://doi.org/10.1021/acs.biochem.5b01336>.
- (48) Ebbinghaus, S.; Dhar, A.; Douglas McDonald, J.; Gruebele, M. Protein Folding Stability and Dynamics Imaged in a Living Cell. *Nature Publishing Group* **2010**, 7 (4), 319–319. <https://doi.org/10.1038/nmeth.1435>.
- (49) Guin, D.; Gruebele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chem. Rev.* **2019**, 119 (18), 10691–10717. <https://doi.org/10.1021/acs.chemrev.8b00753>.

CHAPTER 5: IN-CELL PROTEIN-PROTEIN CONTACTS: TRANSIENT INTERACTIONS IN THE CROWD

INTRODUCTION

Proteins *in vivo* are surrounded by a complex cellular milieu composed of macromolecules and small metabolite molecules, in addition to the water and inorganic ions that typically surround proteins in experiments *in vitro*. As discussed earlier, of particular importance is the high concentration of macromolecules, which can range from 10-40% of the cell by volume.²⁻⁴ In such a crowded environment, interactions between various proteins are inevitable. In response to this highly crowded environment, proteins have evolved to form transient functional contacts (quinary structure) and to minimize non-functional competitive binding (sticking).^{5,6} These adaptations are hypothesized to allow proteins to diffuse more freely throughout the cytoplasm, help prevent aggregation, and increase biological function.^{5,7}

High-stability functional protein-protein interactions have been studied for decades, including as potential drug targets.^{8,9} More transient interactions are recently receiving more attention.¹⁰ A number of databases have been created to characterize the total space of the interactome.¹¹ Both computational and experimental studies have been conducted to investigate the physical mechanisms underlying quinary structure (electrostatics, hydrophobic interactions,

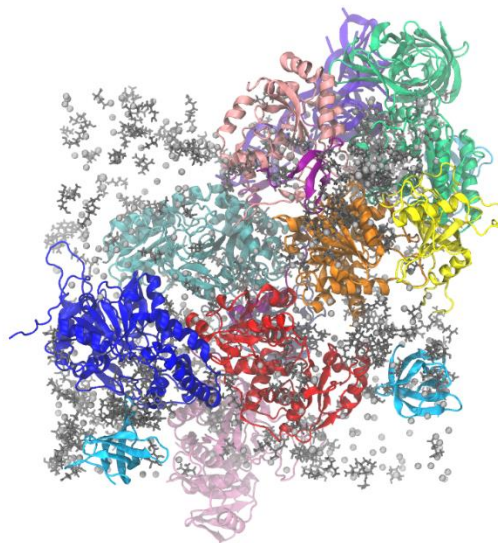


Figure 5.1: Snapshot of the EC36mCU model cytoplasm. Multiple protein-protein contacts, including macromolecular clusters (top right), are observed. Figure published in ref. 1.

* This chapter is reproduced in part with permission from:

Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *The Journal of Physical Chemistry Letters* **2019**, 10 (18), 5667–5673.¹

and protein shape).^{5,6,12–14} Experiments have indicated that crowding and non-specific interactions between neighboring proteins also contribute to the differences in protein dynamics observed *in vitro* and *in vivo*.^{5,15}

Recent experiments have highlighted the dependence of a protein's stability¹⁶ and unfolded ensemble¹⁷ on its local environment. Further, several results emphasize the importance of the local electrostatic environment that cannot be replicated with inert crowders.^{6,15,18} These results point to a bigger picture of a “free for all” of nonspecific protein-protein interactions caused by local charged or hydrophobic patches sticking to complementary patches of neighboring proteins.

However, within this “free for all,” there is evidence of evolved mitigation of protein-protein sticking. Attempts to characterize sticking have largely focused on the experimental observation of anomalous translational or rotational diffusion *in vivo*.^{5,19,20} Studies by Oliveberg and coworkers have indicated that negative surface charges on proteins are crucial to ensure free rotational diffusion of proteins in both mammalian and bacterial cells.^{5,21,22} Additionally, they found that bacterial proteins are less sticky than their human homologs whether they are in bacterial or human cells, indicating that bacterial proteins in general may be more resistant to protein-protein sticking.^{5,21}

Atomistic computational studies in crowded environments have been limited by the computational cost of simulating such large systems. Early attempts to simulate the crowded cellular environment have involved modelling proteins as hard spheres.^{23,24} More recently, protein diffusion *in vivo* has been simulated using Brownian Dynamics or coarse-graining techniques.^{25–29} Others have simulated crowded environments by including multiple copies of small proteins.^{12,30–32} A $>10^8$ atom model of the *Mycoplasma genitalium* cytosol simulated by all-atom molecular dynamics for 100 ns has heralded a new series of simulations studying proteins in crowded environments.³³ Before the simulations discussed in this thesis, no in-cell simulations using all-atom MD had reached the μ s timescale relevant for sampling sticking processes. Since the publication of reference 1, computational studies of protein-protein sticking have continued.^{31,34} The results continue to highlight the role of electrostatics in protein-protein sticking, but do not emphasize the role of hydrophobicity in protein-protein contacts discussed here.

In this chapter, we analyze results from the EC22*, EC36m, and EC36mCU models to determine the extent, mechanism, and timescale of non-specific protein-protein interactions between various components of the bacterial cytoplasm model. Because we deliberately selected

macromolecular components that are not known to interact with each other biologically, our study focuses on protein-protein sticking, rather than quinary interactions or more stable protein-protein interactions.

ANALYSIS METHODS

For details on model construction, please see Chapter 2.

To analyze protein-protein contacts, snapshots were taken once every 24 ns throughout the trajectories. Although the Phe-tRNA forms multiple contacts with neighboring macromolecules, these contacts are not considered here because we are focusing on protein-protein interactions in this work. To calculate the area of the protein-protein contacts, we measured the difference between the solvent accessible surface area (SASA) of each protein alone and in a protein-protein complex for every potential protein-protein interaction in the system. To account for periodic

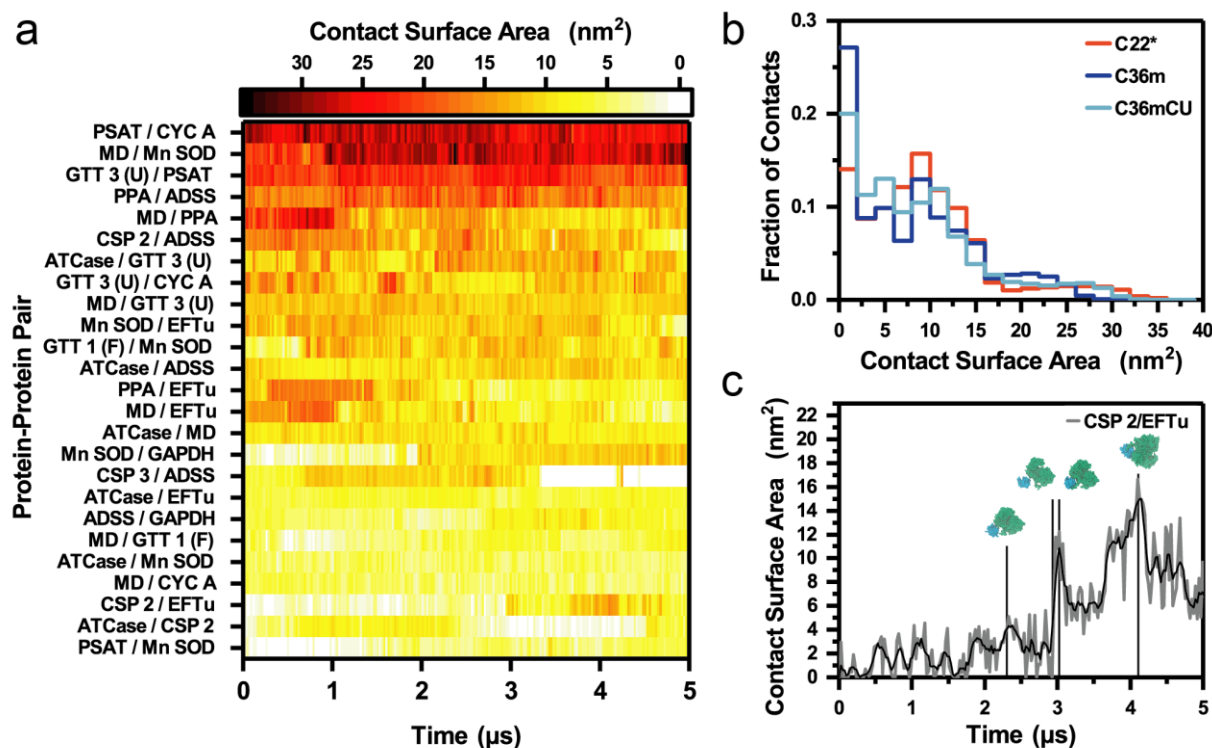


Figure 5.2: (a) Contact size for largest protein-protein contacts during the EC36mCU simulation. (b) Distribution of protein-protein contact sizes for CHARMM 22*, 36m, and 36m with CUFIX corrections. (c) Surface area of the CSP2/EFTu contact in the C36mCU model across the trajectory with snapshots at 2.3, 2.9, 3.0, and 4.1 μs. Figure published in ref. 1.

boundary conditions, the contacts included any extension of a contact area into wrapped regions (see Code C.1).

RESULTS AND DISCUSSION

Protein Clustering

As has been observed previously even in simulations with less than half the protein concentration,^{12,25,30,32} the cytoplasmic proteins tend to cluster, forming protein-rich and water rich regions in the simulation box (Figure 5.1). As discussed in ref. 35, protein-protein interactions are probably overestimated relative to protein-water interactions in the current generation of force fields, the same effect that is also known to overestimate protein stability in simulations.³⁵ To quantify such sticking, we calculated the number of unique neighbors each protein encountered throughout the trajectory. All proteins in all three models formed at least one protein-protein contact; all but one protein in all three models contacts at least two proteins simultaneously at some point. We find that at any given time, proteins on average are in contact with 4.6 unique protein neighbors. These results agree well with previous Brownian Dynamics simulations of

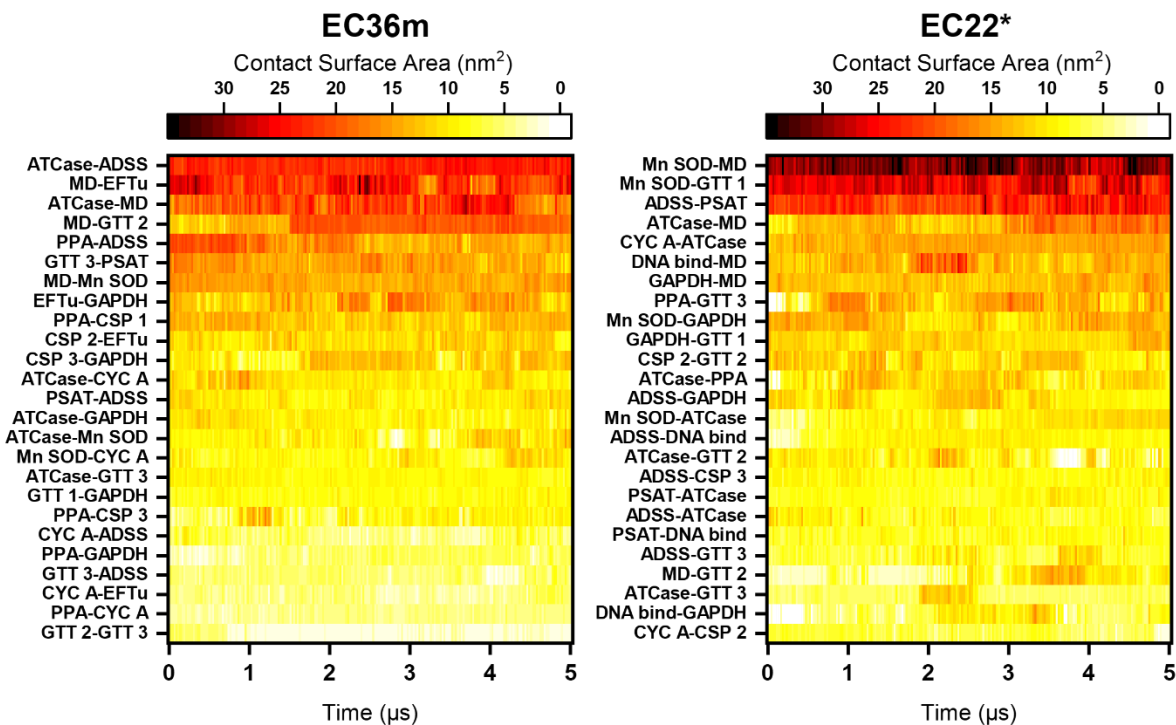


Figure 5.3: Heatmaps of protein-protein contact surface area of the 25 largest protein-protein contacts (by average) in the and EC36m and EC22* models. Figure published in ref. 1.

crowded cellular environments,²⁵ which predict 2-7 unique neighbors for similar-sized proteins. Our observation of $\sim 35 \pm 3$ simultaneous pair contacts out of 105 possible at any given time is also consistent with about 5 neighbors.

The difference of protein-protein contact size distributions between force fields (Figure 5.2b) indicates that the EC36m system is markedly less sticky than EC22* and EC36mCU. EC22* also has more very large ($>30 \text{ nm}^2$) protein-protein contacts. CHARMM 22* was originally designed to fold single proteins in water, and thus may over-stabilize contacts between amino acids that are not close together in protein sequence, such as tertiary structure within a single protein, or surface amino acid contacts between different proteins.

Previous experimental studies of nonspecific protein-protein interactions in crystal structures have indicated that on average, these have a contact area of $\sim 5.7 \text{ nm}^2$.³⁶ In our three systems, the mean contact area was 9.5 nm^2 (EC22*), 8.1 nm^2 (EC36m) and 8.5 nm^2 (EC36mCU). In addition, a plateau in the distribution was observed at $\sim 10 \text{ nm}^2$ for all three models. The plateaus

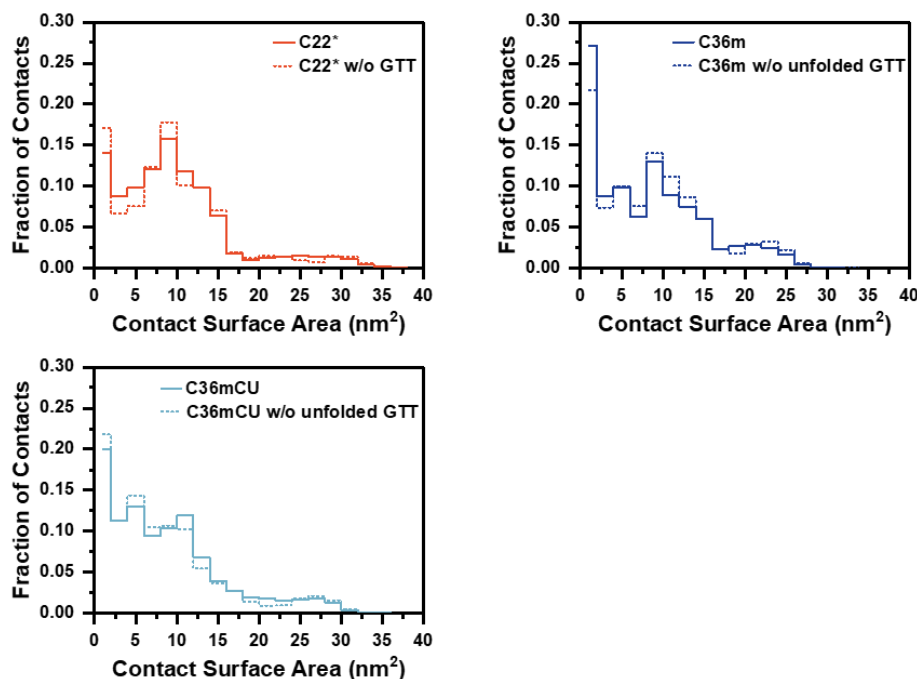


Figure 5.4: Distribution of protein-protein contact surface areas with and without unfolded GTT. Figure published in ref. 1.

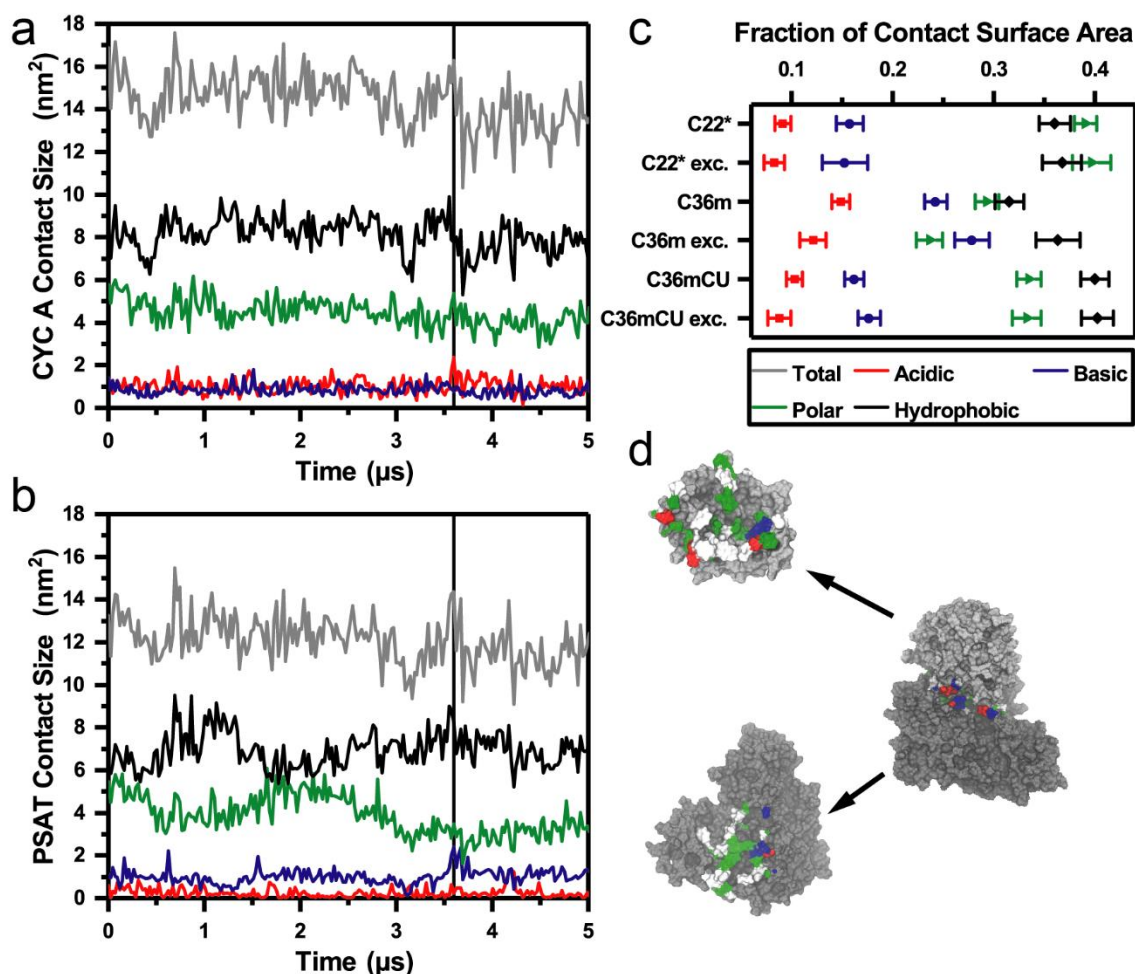


Figure 5.5: (a), (b) Contact surface areas of the PSAT/CYC A contact in the C36mCU simulation. For each contributing protein, hydrophobic, polar, acidic, and basic contributions are shown. (c) Relative contributions of hydrophobic, polar, acidic, and basic residues in protein-protein contacts across all systems. Error bars are based on standard deviations, which in turn are calculated over time. “Exc.” refers to the exclusion of oligomer-oligomer contacts from the dataset. (d) Snapshot of the protein-protein contact at 3.6 μ s. Residues at the protein-protein interface are colored according to residue type: hydrophobic (white), polar (green), basic (blue), and acidic (red). Figure published in ref. 1.

cannot be wholly attributed to unfolded WW domain model proteins included in the simulation (see below), as they persist when these proteins are removed from the average (Figure 5.4).

Unfolded Protein Contacts

The unfolded copies of the GTT model protein (a 33 residue WW domain) are more prone to forming large, long-lived contacts than folded protein surfaces in the simulation. Of the 10 largest contacts on average, a single unfolded copy of GTT participates in 4 (Figure 5.2a) even though it represents only 1 of up to 17 macromolecules in the simulation. We attribute this to

exposure of hydrophobic core residues, in agreement with previous experimental reports^{20,37,38} and computational modeling of in-cell systems²⁵ that point to hydrophobic interactions as an important source of nonspecific protein-protein sticking.

Additionally, other computational¹² and experimental^{5,6} studies have pointed to electrostatics driving protein-protein contacts. Figure 5.5c shows, for contacts in the >10 nm² range, what fraction of the contact surface is covered by hydrophobic, hydrophilic, and charged amino acid side chains, relative to their abundances on the complete set of protein surfaces (averaged for the 3 models) simulated here. Hydrophobic and polar residues tend to dominate over the formation of salt bridges between charged residues in all three models.

Diffusion

Diffusion is the transport mechanism preceding macromolecular sticking. The effective diffusion could be slowed down and made anomalous by sticking, particularly if measured over long times. As has been discussed,³⁹ the bacterial cytoplasm has glassy properties at long scales, modulated by metabolic activity. Protein clustering that contributes to slow dynamics can be seen in Figure 5.1, and may be exacerbated by the small volume and periodic boundary condition of our simulation. Figure 5.6 shows the squared displacement of individual proteins versus elapsed time in all three models. All models show evidence of anomalous diffusion over long time periods: There is a ‘fast’ linear regime from ~30 to 200 ns, a ‘slow’ linear regime from ~200 to 800 ns, and an overall roll-over at long times. EC36m in particular is least well separated into ‘fast’ and ‘slow’ linear regimes, with just a gradual roll-over.

Previous all-atom simulations of proteins in crowded environments^{12,31,33} have indicated larger diffusion coefficients than we find here. Even our ‘fast’ linear regime produces diffusion coefficients ($\sim 1.4\text{--}2$ nm²/μs) slower than experiment⁴⁰ by about a factor of 3-5 when left uncorrected for finite size effects.⁴¹ The subsequent ‘slow’ linear regime goes hand-in-hand with

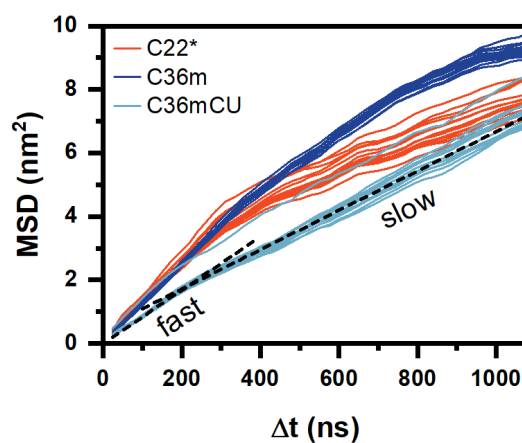


Figure 5.6: Mean squared displacement of proteins in C22* (orange), C36m (dark blue), and C36mCU (light blue) models versus elapsed time. The straight dotted lines superimposed on the C36m CU simulations illustrate the ‘fast’ and ‘slow’ linear regimes. Cold Shock Protein 1 in the C22* model and Cold Shock Proteins 1 & 3 in the C36mCU model were outliers that diffused more quickly than other proteins and were omitted from this figure. Figure published in ref. 1.

the protein sticking noted above—as proteins crowd or stick together and form clusters, they become less free to diffuse about the box. It is worth noting that previous simulations measure diffusion coefficients using much shorter timescales than the work presented here, highlighting the potential importance protein sticking for long-time transport dynamics.

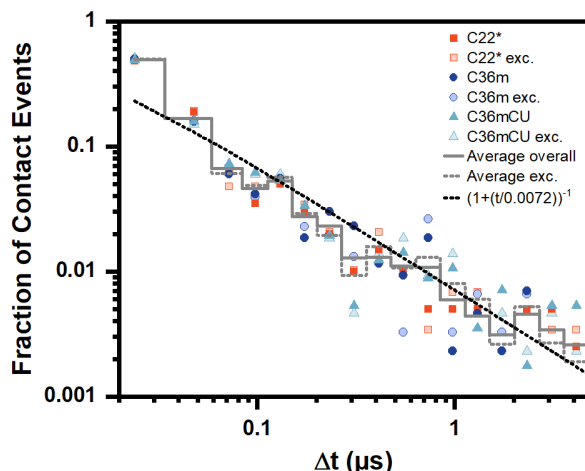


Figure 5.7: Distribution of protein-protein contact lifetimes. Only $\approx 7\%$ of transient contacts last longer than $1 \mu\text{s}$ in all three models; about half of these lasted throughout the simulation window of $5 \mu\text{s}$, so their ultimate potential duration is unknown and not shown. “Exc.” refers to the exclusion of oligomer-oligomer contacts from the dataset. A minimum 0.15 nm^2 contact area threshold was used. Figure published in ref. 1.

the timescales of protein-protein contacts do not differ appreciably between the three force fields simulated: with a $15 \text{ \AA}^2$ threshold defining a contact, 90% the probability $P(\Delta t)$ decays on the order of $\sim 1 \mu\text{s}$.

Whether interactions between oligomeric proteins are excluded (“exc.” in Figure 5.7) or not has little effect on the probability distribution in Figure 5.7. This supports our hypothesis that while oligomeric proteins are of course prone to interacting with their partners, as monomers they are not any more prone to interact with other proteins in the cytoplasmic matrix than are monomeric proteins. The average distribution (gray stepped curves in Figure 5.7) is essentially the same either way. Thus, it appears that binding interfaces of oligomeric proteins, just like surfaces of monomeric proteins, have co-evolved with the cytosol to avoid non-specific interactions, as indicated in previous studies of protein-protein interaction networks.^{42–45}

$P(\Delta t)$ was fitted between 50 ns and $4 \mu\text{s}$ to a stretched exponential and to power laws. The best fit was obtained with a power law probability distribution finite as t approaches 0,⁴⁶ which

Protein-protein Contact Lifetimes

Our results for all protein-protein pairs indicate that contacts are highly dynamic (Figure 5.2a, Figure 5.3). For example, the contact between the cold shock protein number 2 (CSP 2) and the elongation factor Tu (EFTu) is broken and reformed repeatedly; in one case, a contact forms and reaches $>10 \text{ nm}^2$ in under 50 ns (Figure 5.2c). To determine the timescales of the contact events, we compare the distribution of protein-protein contact lifespans Δt in all three simulations (Figure 5.7). Despite differences observed in contact area distribution between the three force fields,

yielded the equation $P(\Delta t) \sim (1 + (\Delta t / 0.0072))^{-1}$. As discussed in ref. ⁴⁷, for ‘birth-death’ processes with constant rates (here: sticking-unsticking processes with constant rate coefficients for making or breaking protein-protein contact pairs), both stretched exponentials and power laws with exponents close to -1 are possible.⁴⁷ The latter happens when the number of contacts at any given time is proportional to the lifespan Δt of contacts and the distribution of contacts is sampled rapidly. In terms of the power spectrum $P(\Delta \omega)$, where $\Delta \omega$ is the contact frequency Fourier-conjugate to contact lifespan Δt , a $1/\Delta t$ probability distribution simply implies a constant power spectrum $P(\Delta \omega)$.⁴⁸

CONCLUSION

The decreased diffusion coefficients, observed protein clustering, and larger than expected protein-protein contacts indicate that all three force fields are “too sticky”— they over-stabilize the frequency of formation and persistence of protein-protein contacts. The large contribution of hydrophobic residues to the largest contacts we observe here indicate that the source of this stickiness may lie in hydrophobic interactions. The 36m model, which has the smallest overall protein-protein contact areas, shows a larger contribution of charged residues relative to hydrophobic residues. We hypothesize that the CUFIX corrections, which have focused largely on charged and polar interactions,⁴⁹ destabilize these charge-charge interactions but not hydrophobic interactions that drive the protein-protein sticking we observe.

Hydrophobicity results from a water-water interactions at the expense of water-protein interactions.⁵⁰ Previous simulations have shown that the TIP3P water model we use leads to overly compact disordered states of unfolded proteins.⁵¹ If water-water interactions are stabilized in current CHARMM force fields, excessive protein-protein sticking and lowered macromolecular diffusion would result. This would also explain current force fields’ tendencies to overestimate protein melting temperatures,³⁵ as proteins would be more likely to favor protein-protein over protein-water interactions and remain folded at higher temperatures. Recent results indicate that simulations using the TIP4P-D water model⁵¹ can reproduce experimental viscosity measurements of crowded protein solutions.³¹ Other groups have attempted to correct this sticking problem by increasing Lennard-Jones protein-water interactions in crowded systems.^{12,52}

We note that the timescale of protein-protein contacts decays on the order of $\sim 1 \mu\text{s}$, with $\sim 7\%$ of protein-protein contacts surviving longer than $1 \mu\text{s}$ without being broken (Figure 5.7). This $\sim 1 \mu\text{s}$ timescale is on the order of the speed limit for downhill protein folding.^{53–55} In the crowded cytoplasm, the speed limits of intra- and inter-molecular protein contacts are not appreciably different. Just as tertiary structure is formed by hydrophobically-driven docking of two secondary structural elements within a single protein, protein-protein contacts in Figure 5.5 are driven by hydrophobic and polar interactions in close vicinity.

In ref. 1, we reported the first μs -timescale all-atom MD simulation of protein dynamics in a bacterial cytoplasm model. Here, we find that three versions of the CHARMM forcefield over-stabilize non-specific sticking between proteins and their surrounding environment, with C36m the least sticky. In the screened high-ion-concentration milieu of our model cytoplasm, protein-protein interaction is driven mainly by hydrophobic and polar contacts and leads to translational diffusion 4–8 times slower than experimental values. Sticking occurs on the μs timescale or faster, similar to the time scale of the intra-molecular tertiary structure formation. Competition with sticking may have driven evolution of fast tertiary contact (re)formation in already-folded proteins that become partly unfolded. Our study demonstrates the potential of crowded all-atom models to offer new insight into protein dynamics *in vivo*. As molecular dynamics continues to grow as a technique to study biomolecules *in vivo*, care must be taken to develop force fields that accurately replicate sticking and non-competitive binding in crowded environments.

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (2) Theillet, F.-X.; Binolfi, A.; Frembgen-Kesner, T.; Hingorani, K.; Sarkar, M.; Kyne, C.; Li, C.; Crowley, P. B.; Gierasch, L.; Pielak, G. J.; Elcock, A. H.; Gershenson, A.; Selenko, P. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114*, 6661–6714. <https://doi.org/10.1021/cr400695p>.
- (3) Minton, A. P. The Influence of Macromolecular Crowding and Macromolecular Confinement on Biochemical Reactions in Physiological Media. *J. Biol. Chem.* **2001**, *276*, 10577–105780. <https://doi.org/10.1074/jbc.R100005200>.
- (4) Mittal, S.; Chowhan, R. K.; Singh, L. R. Macromolecular Crowding: Macromolecules Friend or Foe. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2015**, *1850*, 1822–1831. <https://doi.org/10.1016/J.BBAGEN.2015.05.002>.
- (5) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E4556–E4563. <https://doi.org/10.1073/pnas.1621227114>.
- (6) Cohen, R. D.; Pielak, G. J. Electrostatic Contributions to Protein Quinary Structure. *J. Am. Chem. Soc.* **2016**, *138*, 13139–13142. <https://doi.org/10.1021/jacs.6b07323>.
- (7) Guin, D.; Gruebele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chem. Rev.* **2019**, *119* (18), 10691–10717. <https://doi.org/10.1021/acs.chemrev.8b00753>.
- (8) De Las Rivas, J.; Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput. Biol.* **2010**, *6*, e1000807. <https://doi.org/10.1371/journal.pcbi.1000807>.
- (9) Modell, A. E.; Blosser, S. L.; Arora, P. S. Systematic Targeting of Protein–Protein Interactions. *Trends Pharmacol. Sci.* **2016**, *37*, 702–713. <https://doi.org/10.1016/j.tips.2016.05.008>.
- (10) Sukenik, S.; Ren, P.; Gruebele, M. Weak Protein–Protein Interactions in Live Cells Are Quantified by Cell-Volume Modulation. *Proc. Natl. Acad. Sci.* **2017**, *114*, 6776–6781. <https://doi.org/10.1073/pnas.1700818114>.
- (11) De Las Rivas, J.; Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput. Biol.* **2010**, *6*, e1000807. <https://doi.org/10.1371/journal.pcbi.1000807>.
- (12) Nawrocki, G.; Wang, P.-H.; Yu, I.; Sugita, Y.; Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. *J. Phys. Chem. B* **2017**, *121*, 11072–11084. <https://doi.org/10.1021/acs.jpcc.7b08785>.
- (13) Guseman, A. J.; Goncalves, G. M. P.; Speer, S. L.; Young, G. B.; Pielak, G. J. Protein Shape Modulates Crowding Effects. *Proc. Natl. Acad. Sci.* **2018**, *115* (43), 10965–10970. <https://doi.org/10.1073/pnas.1810054115>.
- (14) Guseman, A. J.; Speer, S. L.; Perez Goncalves, G. M.; Pielak, G. J. Surface Charge Modulates Protein–Protein Interactions in Physiologically Relevant Environments. *Biochemistry* **2018**, *57* (11), 1681–1684. <https://doi.org/10.1021/acs.biochem.8b00061>.

- (15) Davis, C. M.; Gruebele, M. Non-Steric Interactions Predict the Trend and Steric Interactions the Offset of Protein Stability in Cells. *ChemPhysChem* **2018**, *19*, 2290–2294. <https://doi.org/10.1002/cphc.201800534>.
- (16) Feng, R.; Gruebele, M.; Davis, C. M. Quantifying Protein Dynamics and Stability in a Living Organism. *Nat. Commun.* **2019**, *10*, 1179. <https://doi.org/10.1038/s41467-019-09088-y>.
- (17) Wang, Y.; Sukenik, S.; Davis, C. M.; Gruebele, M. Cell Volume Controls Protein Stability and Compactness of the Unfolded State. *J. Phys. Chem. B* **2018**, *122*, 11762–11770. <https://doi.org/10.1021/acs.jpcc.8b08216>.
- (18) Monteith, W. B.; Cohen, R. D.; Smith, A. E.; Guzman-Cisneros, E.; Pielak, G. J. Quinary Structure Modulates Protein Stability in Cells. *Proc. Natl. Acad. Sci.* **2015**, *112*, 1739–1742. <https://doi.org/10.1073/pnas.1417415112>.
- (19) Wang, Q.; Zhuravleva, A.; Gierasch, L. M. Exploring Weak, Transient Protein–Protein Interactions in Crowded in Vivo Environments by in-Cell Nuclear Magnetic Resonance Spectroscopy. *Biochemistry* **2011**, *50*, 9225–9236. <https://doi.org/10.1021/bi201287e>.
- (20) Guo, M.; Gelman, H.; Gruebele, M. Coupled Protein Diffusion and Folding in the Cell. *PLoS ONE* **2014**, *9*, e113040. <https://doi.org/10.1371/journal.pone.0113040>.
- (21) Leeb, S.; Sörensen, T.; Yang, F.; Mu, X.; Oliveberg, M.; Danielsson, J. Diffusive Protein Interactions in Human versus Bacterial Cells. *Curr. Res. Struct. Biol.* **2020**, *2*, 68–78. <https://doi.org/10.1016/j.crstbi.2020.04.002>.
- (22) Wennerström, H.; Vallina Estrada, E.; Danielsson, J.; Oliveberg, M. Colloidal Stability of the Living Cell. *Proc. Natl. Acad. Sci.* **2020**, *117* (19), 10113–10121. <https://doi.org/10.1073/pnas.1914599117>.
- (23) Bicout, D. J.; Field, M. J. Stochastic Dynamics Simulations of Macromolecular Diffusion in a Model of the Cytoplasm of Escherichia Coli. *J. Phys. Chem.* **1996**, *100*, 2489–2497.
- (24) Ridgway, D.; Broderick, G.; Lopez-Campistrous, A.; Ru’aini, M.; Winter, P.; Hamilton, M.; Boulanger, P.; Kovalenko, A.; Ellison, M. J. Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm. *Biophys. J.* **2008**, *94*, 3748–3759. <https://doi.org/10.1529/biophysj.107.116053>.
- (25) McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput. Biol.* **2010**, *6*, e1000694. <https://doi.org/10.1371/journal.pcbi.1000694>.
- (26) Ando, T.; Skolnick, J. Crowding and Hydrodynamic Interactions Likely Dominate in Vivo Macromolecular Motion. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18457–184562. <https://doi.org/10.1073/pnas.1011354107>.
- (27) Mereghetti, P.; Gabdouline, R. R.; Wade, R. C. Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. *Biophys. J.* **2010**, *99*, 3782–3791. <https://doi.org/10.1016/j.bpj.2010.10.035>.
- (28) Mereghetti, P.; Wade, R. C. Atomic Detail Brownian Dynamics Simulations of Concentrated Protein Solutions with a Mean Field Treatment of Hydrodynamic Interactions. *J. Phys. Chem. B* **2012**, *116*, 8523–8533. <https://doi.org/10.1021/jp212532h>.
- (29) Wang, Q.; Cheung, M. S. A Physics-Based Approach of Coarse-Graining the Cytoplasm of Escherichia Coli (CGCYTO). *Biophys. J.* **2012**, *102*, 2353–2361. <https://doi.org/10.1016/J.BPJ.2012.04.010>.

- (30) Cossins, B. P.; Jacobson, M. P.; Guallar, V. A New View of the Bacterial Cytosol Environment. *PLoS Comput. Biol.* **2011**, *7*, e1002066.
- (31) von Bülow, S.; Siggel, M.; Linke, M.; Hummer, G. Dynamic Cluster Formation Determines Viscosity and Diffusion in Dense Protein Solutions. *Proc. Natl. Acad. Sci.* **2019**, *116*, 9843–9852. <https://doi.org/10.1073/PNAS.1817564116>.
- (32) Petrov, D.; Zagrovic, B. Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Comput. Biol.* **2014**, *10*, e1003638. <https://doi.org/10.1371/journal.pcbi.1003638>.
- (33) Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular Interactions Modulate Macromolecular Structure and Dynamics in Atomistic Model of a Bacterial Cytoplasm. *eLife* **2016**, *5*, e19274. <https://doi.org/10.7554/eLife.19274>.
- (34) Nawrocki, G.; Im, W.; Sugita, Y.; Feig, M. Clustering and Dynamics of Crowded Proteins near Membranes and Their Influence on Membrane Bending. *Proc. Natl. Acad. Sci.* **2019**, *116* (49), 24562–24567. <https://doi.org/10.1073/pnas.1910771116>.
- (35) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E4758–E4766. <https://doi.org/10.1073/pnas.1800690115>.
- (36) Janin, J.; Rodier, F. Protein-Protein Interaction at Crystal Contacts. *Proteins Struct. Funct. Genet.* **1995**, *23*, 580–587. <https://doi.org/10.1002/prot.340230413>.
- (37) Smith, A. E.; Zhou, L. Z.; Gorenssek, A. H.; Senske, M.; Pielak, G. J. In-Cell Thermodynamics and a New Role for Protein Surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 1725–1730. <https://doi.org/10.1073/PNAS.1518620113>.
- (38) Dhar, A.; Girdhar, K.; Singh, D.; Gelman, H.; Ebbinghaus, S.; Gruebele, M. Protein Stability and Folding Kinetics in the Nucleus and Endoplasmic Reticulum of Eucaryotic Cells. *Biophys. J.* **2011**, *101*, 421–430. <https://doi.org/10.1016/j.bpj.2011.05.071>.
- (39) Parry, B. R.; Surovtsev, I. V.; Cabeen, M. T.; O'Hern, C. S.; Dufresne, E. R.; Jacobs-Wagner, C. The Bacterial Cytoplasm Has Glass-like Properties and Is Fluidized by Metabolic Activity. *Cell* **2014**, *156*, 183–194. <https://doi.org/10.1016/J.CELL.2013.11.028>.
- (40) Nenninger, A.; Mastroianni, G.; Mullineaux, C. W. Size Dependence of Protein Diffusion in the Cytoplasm of Escherichia Coli. *J. Bacteriol.* **2010**, *192*, 4535–4540. <https://doi.org/10.1128/JB.00284-10>.
- (41) Yeh, I.-C.; Hummer, G. System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions. *J. Phys. Chem. B* **2004**, *108*, 15873–15879. <https://doi.org/10.1021/JP0477147>.
- (42) Heo, M.; Maslov, S.; Shakhnovich, E. Topology of Protein Interaction Network Shapes Protein Abundances and Strengths of Their Functional and Nonspecific Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (10), 4258–4263. <https://doi.org/10.1073/pnas.1009392108>.
- (43) Johnson, M. E.; Hummer, G. Nonspecific Binding Limits the Number of Proteins in a Cell and Shapes Their Interaction Networks. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (2), 603–608. <https://doi.org/10.1073/pnas.1010954108>.
- (44) Levy, E. D.; De, S.; Teichmann, S. A. Cellular Crowding Imposes Global Constraints on the Chemistry and Evolution of Proteomes. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (50), 20461–20466. <https://doi.org/10.1073/pnas.1209312109>.

- (45) Peleg, O.; Choi, J.-M.; Shakhnovich, E. I. Evolution of Specificity in Protein-Protein Interactions. *Biophys. J.* **2014**, *107* (7), 1686–1696. <https://doi.org/10.1016/J.BPJ.2014.08.004>.
- (46) Ervin, J.; Sabelko, J.; Gruebele, M. Submicrosecond Real-Time Fluorescence Sampling: Application to Protein Folding. *J. Photochem. Photobiol. B* **2000**, *54*, 1–15. [https://doi.org/10.1016/S1011-1344\(00\)00002-6](https://doi.org/10.1016/S1011-1344(00)00002-6).
- (47) Reed, W. J.; Hughes, B. D. From Gene Families and Genera to Incomes and Internet File Sizes: Why Power Laws Are so Common in Nature. *Phys. Rev. E - Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **2002**, *66*, 067103. <https://doi.org/10.1103/PhysRevE.66.067103>.
- (48) Bracewell, R. N. Fourier and Other Mathematical Transforms. In *Digital Encyclopedia of Applied Physics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2009; pp 1149–1175. <https://doi.org/10.1002/3527600434.eap144.pub2>.
- (49) Yoo, J.; Aksimentiev, A. New Tricks for Old Dogs: Improving the Accuracy of Biomolecular Force Fields by Pair-Specific Corrections to Non-Bonded Interactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 8432–8449. <https://doi.org/10.1039/C7CP08185E>.
- (50) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2001**, *106*, 521–533. <https://doi.org/10.1021/JP015514E>.
- (51) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123. <https://doi.org/10.1021/jp508971m>.
- (52) Nawrocki, G.; Karaboga, A.; Sugita, Y.; Feig, M. Effect of Protein–Protein Interactions and Solvent Viscosity on the Rotational Diffusion of Proteins in Crowded Environments. *Phys. Chem. Chem. Phys.* **2019**, *21*, 876–883. <https://doi.org/10.1039/C8CP06142D>.
- (53) Yang, W. Y.; Gruebele, M. Folding λ -Repressor at Its Speed Limit. *Biophys. J.* **2004**, *87*, 596–608. <https://doi.org/10.1529/biophysj.103.039040>.
- (54) Hagen, S. J.; Hofrichter, J.; Szabo, A.; Eaton, W. A.; Wipf, P.; Kelly, J. W.; Gruebele, M. Diffusion-Limited Contact Formation in Unfolded Cytochrome c: Estimating the Maximum Rate of Protein Folding. *Proc. Natl. Acad. Sci.* **1996**, *93* (21), 11615–11617. <https://doi.org/10.1073/pnas.93.21.11615>.
- (55) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. An Experimental Survey of the Transition between Two-State and Downhill Protein Folding Scenarios. *Proc. Natl. Acad. Sci.* **2008**, *105*, 2369–2374. <https://doi.org/10.1073/pnas.0711908105>.

CHAPTER 6: PROTEIN DYNAMICS IN AN ALL-ATOM MODEL OF THE *H. SAPIENS* CYTOPLASM

INTRODUCTION

As discussed in Chapter 1, the cellular environment can greatly perturb a protein's dynamics, particularly when compared to *in vitro* studies. However, cell-to-cell environmental differences—including cell species, cell type, and cell stress—can also affect these dynamics.¹⁻³

Our initial *E. coli* models have revealed a great idea about the impact of local environment on metabolite and protein dynamics. However, it was unclear how much of these results were due to the cellular environment in general or the *E. coli* environment in particular. Bacterial and mammalian cellular conditions can vary greatly; for example, metabolite concentrations of glutamate are three times higher in *E. coli* than in *H. sapiens* and

macromolecular crowding varies greatly across both species and cell type.⁴⁻⁷ Further, experimental studies have shown that proteins have evolved to thrive in their unique physicochemical environments.^{1,8,9} We hypothesize that these differences lead to a fundamentally different electrostatic and steric environment in the *E. coli* and *H. sapiens* cytoplasm, which in turn affects the dynamics of the proteins within them.

To further determine the how the differences in the *E. coli* and *H. sapiens* cellular environments affect the dynamics of their respective proteins, we constructed a ~200,000 atom model of a small section of a human cytoplasm and simulated it for 24 μ s (Figure 6.1). Due to the excessive protein-protein sticking and misfolding observed with GTT in the *E. coli* models (Chapter 4), we opted to substitute the copies of GTT with 3 copies of protein B (PRB), a small, α -helical, fast-folding protein that has also been extensively simulated in water-box conditions.^{10,11}

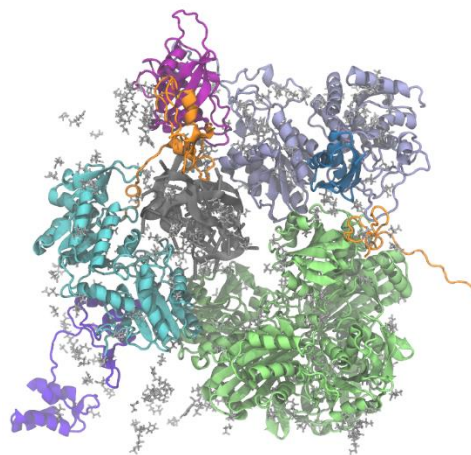


Figure 6.1: Snapshot of the HC36m model after 13.2 μ s of simulation. Proteins are shown in colored ribbon structures, and metabolites as gray licorice structures. For simplicity, inorganic ions and water molecules are not shown.

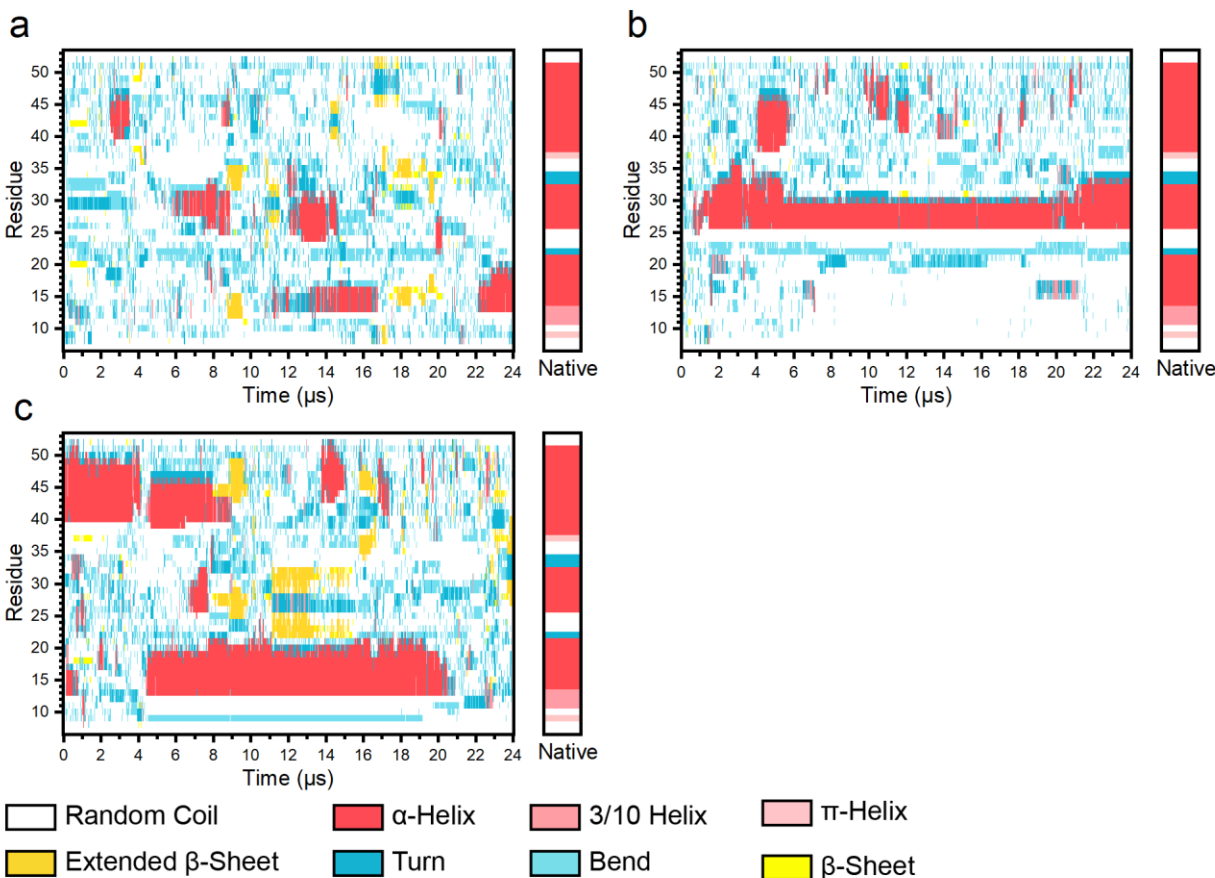


Figure 6.2: Secondary structure of PRB in the HC36m model throughout trajectories as calculated by DSSP.^{2,3} The secondary structure at each residue of the native structure is given at the right. Each panel (a–c) represents one copy of PRB in the first 24 μ s the HC36m simulation. Unlike GTT in the *E. coli* simulations (see Chapter 4), PRB forms long lived, correctly registered, secondary structure in the simulations (b, 2–24 μ s; c, 4–20 μ s).

Here, we discuss preliminary results from this simulation, including PRB’s dynamics and partial folding, and quinary interactions between a pair of proteins that have also been extensively experimentally studied in human cells.^{12–14}

ANALYSIS METHODS

For a detailed description of model construction and simulation, please see Chapter 2.

Backbone root mean squared deviation (RMSD), radius of gyration (R_g), solvent accessible surface area (SASA), hydrophobic SASA, and secondary structure of the three PRB copies in the HC36m model were calculated every 24 ns for the first 24 μ s of C36mCU trajectory. RMSD, R_g , SASA, and hydrophobic SASA were calculated using VMD¹⁵ and secondary structure was calculated using the compute_DSSP function of the python package MDTraj.^{16,17} To calculate the

backbone RMSD and compare to a native secondary structure, a structure taken from the folded ensemble of reference 11 was used.

RESULTS

Protein B partially folds in HC36m model

Like the WW domains in the *E. coli* models,¹⁸ we observed no complete folding events despite simulating longer than the times needed to fold PRB in a water-box calculation.¹⁰ As with GTT, I began by calculating the secondary structure of the three PRB copies throughout the first 24 μ s of simulation. As in we observed in the *E. coli* models, PRB forms long-lived misfolded

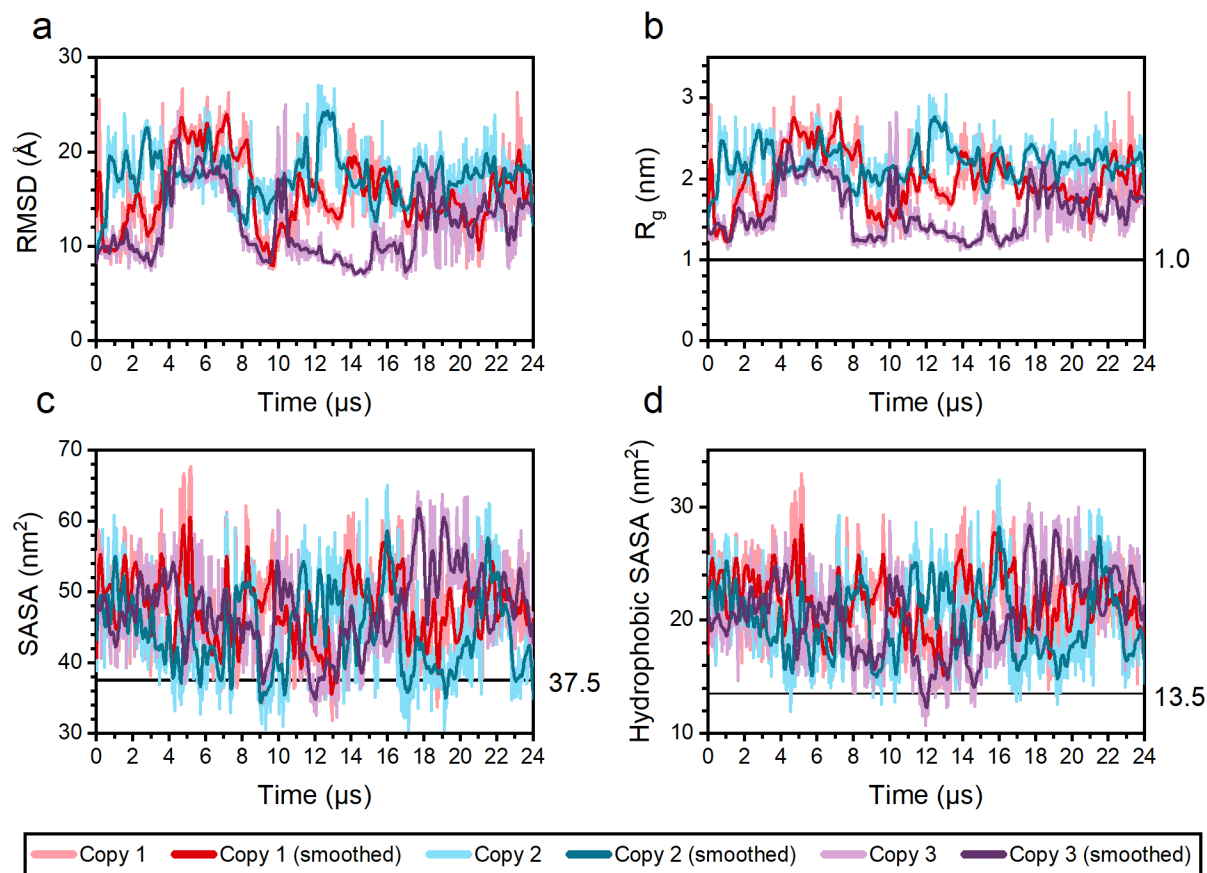


Figure 6.3: Reaction coordinates monitoring the folding of PRB in the HC36m model: (a) Backbone root mean squared deviation (RMSD), (b) Radius of gyration (R_g), (c) Solvent accessible surface area (SASA), (d) Hydrophobic SASA. Copies one, two, and three of PRB in the HC36m model are represented by red, blue, and purple lines respectively. In all panels, the raw data is represented by a light colored line, with a dark colored trace representing the smoothed (adjacent average over 240 ns) data. In panels (b), (c), and (d), a black line represents the value of the reaction coordinate in the native structure and is labeled on the right hand side of the plot.

structures. Here, PRB 3 forms a β -hairpin that persists for several microseconds, despite the native structure of PRB containing no β -strands (Figure 6.2c, 12-14 μ s). Additionally, we observe the formation of register shifted helices, similar to the register shifted β -hairpins we observe in the *E. coli* models (Figure 6.2a, 13 μ s). Finally, when correct secondary structure does form, these partially folded structures persist for similar timescales in both models (GTT 3 forms a partial hairpin that persists for ~ 20 μ s in the EC22* model (Figure B.1); PRB 2's second helix has persisted for over 20 μ s in the HC36m model (Figure 6.2b).

Despite these similarities to our previous simulations, preliminary results suggest that PRB partially folds better in HC36m than GTT did in the *E. coli* models. PRB is more likely than any GTT in our previous simulations to form correct (and correctly registered) secondary structure. Further, PRB appears to stick less to neighboring macromolecules in the HC36m model than GTT did in the *E. coli* models. We observe that in our HC36m model, PRB occasionally has SASA and hydrophobic SASA values below that of its native state. In the *E. coli* models, this was an indication that GTT copies were sticking excessively to their neighbors (see Chapter 4).¹⁸ However, PRB seems to visit these low-SASA states less often than GTT did in the EC22* model, which may indicate PRB sticks less to the *H. sapiens* cytoplasm than GTT does to the *E. coli* cytoplasm.

Most notably, PRB appears to get closer to its native state than GTT did at any point in the >200 μ s of *E. coli* simulations. At ~ 7 μ s, PRB 3 partially forms all three helices, a structure that persists for nearly a microsecond. While this three-helical structure does not fully hydrophobically collapse and has a high backbone RMSD, partial hydrophobic core formation is visible before the protein unfolds again (Figure 6.4). Additionally, the relative positioning of the three helices suggest that some level of tertiary structure formation occurred. While the protein never fully enters the folded ensemble, the near-complete formation of secondary structure and possible tertiary structure make this the most complete folding event of any protein ever atomistically simulated in crowded environments.¹⁸

We hypothesize that two factors account for PRB's relative folding success in the HC36m model. First, MD force fields have historically over stabilized α -helical structures, even folding WW domains into α -helical structures.¹⁹ While these force field issues have largely been resolved in water-box simulations,^{10,20} it is possible that the crowded nature of these in-cell simulations exaggerates any remaining bias towards helical structures (and thus stabilizes PRB more than

GTT). However, the transient β -hairpin formed in by PRB 3 in the HC36m model seems to contradict this hypothesis.

We find it is more likely that a combination of two factors explains PRB's and GTT's relative ability to fold in the cytoplasm: the electrostatics of the folding constructs and the less crowded conditions of the mammalian model. Experimentally, it has been shown that surface charges, modulate protein-protein interactions in-cell.^{1,9,21–23} Further, bacterial proteins have been experimentally shown to stick less than their human homologs in both bacterial and human cells.^{1,23} It is hypothesized that this effect occurs because bacterial proteins evolved more negatively charged surface residues to electrostatically repel neighboring macromolecules in order to compensate for the more-crowded environment of the bacterial cytoplasm.^{7,23} GTT has a

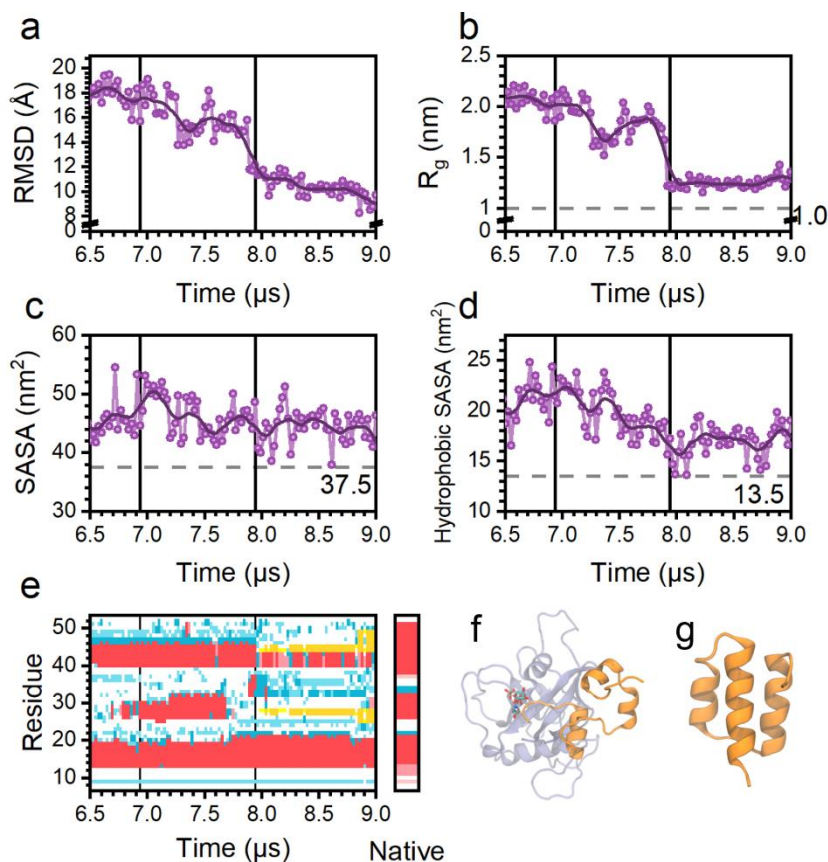


Figure 6.4: Details of PRB 3's partial folding event between 6.5 and 9 μ s. (a-d) RMSD, R_g , SASA, and hydrophobic SASA of PRB 3 between 6.5 and 9 μ s. Raw data is displayed by light purple points, while the smoothed (averaged over 240 ns) data is shown as a dark purple line. Vertical lines delineate the times where secondary structure is formed and lost, and horizontal dashed lines indicate a reaction coordinate's value in the native state. (e) Secondary structure of protein B copy 3 between 6.5 and 9 μ s. Helical residues, beta strands, and turns are shown in red, blue, and yellow respectively. (f) Snapshot of protein B copy 3 (orange) and a neighboring macromolecule (blue) at 7.92 μ s. (g) Protein B's native structure.

calculated isoelectric point (pI) of 5.67, which means it carries a positive charge at physiological pH.²⁴ Like the human proteins observed in *E. coli* experimentally, this charge may have contributed to GTT's excessive protein-protein sticking in the *E. coli* models. In contrast, PRB is negatively charged at physiological pH (calculated pI = 9.52).²⁴ Further, GTT is a fast-folding construct whose original template came from multiple mammalian proteins, while PRB is a lightly mutated bacterial protein.^{25,26} Additionally, the *E. coli* and *H. sapiens* models were not created equal: in order to more closely mimic the *H. sapiens* cellular environment, the HC36m model has a lower concentration of macromolecules, metabolites, and inorganic ions (see Table 2.2). We hypothesize that the electrostatic differences between the two proteins and the more crowded environment in the *E. coli* models likely exaggerated GTT's protein-protein sticking and mitigated PRB's, leaving PRB free to explore its conformational space and fold.

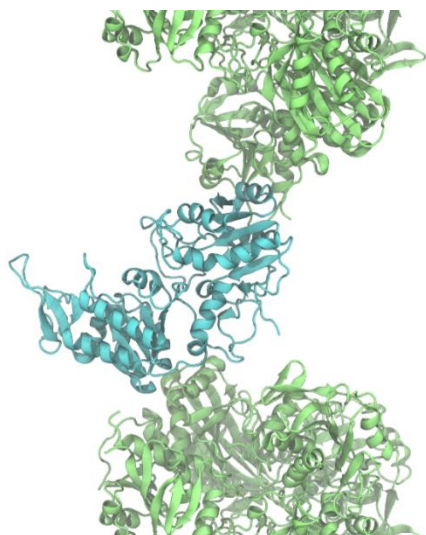


Figure 6.5: Snapshot of PGK (cyan) and GAPDH (lime green) interacting across the periodic boundary conditions in the HC36m model at 6.96 μ s. At multiple points in the simulation, PGK interacts with two different subunits of GAPDH across the periodic boundary conditions.

Interactions between PGK and GAPDH

. Experiments have shown that phosphoglycerate kinase (PGK) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH, which forms nearly-ubiquitous tetramers in the cytoplasm,²⁷ GAPDH₄) form quinary interactions with each other, with an experimentally measured dissociation constant of 14 ± 6 μ M (which is on the order of PGK and GAPDH₄ concentrations in U2-OS cells).^{12,28,29} It is hypothesized that such a quinary interaction evolved in order to “pass” substrates (in this case, 1,3-bisphosphoglycerate) between enzymes to reduce unnecessary metabolite diffusion between steps in a reaction pathway.³⁰

While we observe both the N- and C- domains of PGK interacting with GAPDH throughout the simulation, the C-domain forms contacts with GAPDH much more throughout the simulation (C-domain and N-domain heavy atoms were within a 5 Å cutoff of GAPDH for ~ 21 μ s and ~ 2.5 μ s of the trajectory respectively). While there is insufficient sampling to state definitively that the C-domain interactions are favored, especially given the known protein-protein sticking and decreased diffusion in cellular simulations,^{31,32} this is an intriguing possibility. This C-domain/GAPDH

interaction is particularly interesting because PGK's N-domain holds the binding site for the passed substrate in the interaction, 1,3-bisphosphoglycerate.³³ A C-domain/GAPDH interaction would increase necessary metabolite diffusion more than an N-domain/GAPDH interaction. However, machine-learning based predictions of the PGK/GAPDH interface showed both of PGK's domains interacting with GAPDH, which may support a bi-domain mechanism of interaction.³⁴

Additionally, in the HC36m simulations, both of PGK's domains were interacting with GAPDH (across periodic boundary conditions) in $\sim 1.5 \mu\text{s}$ of the trajectory (Figure 6.5). While proteomic data shows that GAPDH₄ and PGK occur in roughly equal concentrations in human cells, in-cell FRET experiments have pointed to a 2:1 GAPDH₄:PGK ratio as a likely stoichiometry for this quinary interaction.^{12,29} It is likely that the 2:1 GAPDH₄:PGK interactions here are an artifact of the HC36m model's relatively small size and the periodic boundary conditions, but we plan to investigate these bridging interactions further to determine if PGK may, in fact, use both domains to bind GAPDH₄.

CONCLUSIONS AND FUTURE WORK

While preliminary results of the HC36m model look promising to elucidate atomistic mechanisms of protein folding in-cell and quinary interactions of GAPDH and PGK, a number of questions must be addressed in future studies. First, the relative stickiness of the HC36m and EC36m models must be evaluated. Due to the excessive charge-charge interactions often observed in current MD force fields,^{35,36} the HC36m model may, indeed show less protein-protein sticking than its EC36m counterpart. To determine if lowered salt and metabolite concentrations and the different protein surface charges effectively mediate the protein-protein sticking in the HC36m model, we will repeat the analysis used in Chapter 5 to evaluate protein sticking.

Further, the HC36m model offers an important opportunity to directly compare to experimental studies of proteins, particularly PGK, in-cell. Further avenues of inquiry include comparing the PGK structural ensemble in the HC36m model to experimental studies in U2-OS cells. Experimentally, it has been shown that PGK adopts a more compact structure in crowded *in vitro* systems and in-cell.^{13,37,38} Comparing PGK's structure in the HC36m model to known crystal structures and other simulations may offer information about the atomistic mechanism behind this

compaction, or be used to further tune the crowding vs. sticking aspect of current MD force fields.³⁵

REFERENCES

- (1) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. V.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in *Escherichia Coli*. *Proc. Natl. Acad. Sci.* **2017**, *114* (23), E4556–E4563. <https://doi.org/10.1073/pnas.1621227114>.
- (2) Feng, R.; Gruebele, M.; Davis, C. M. Quantifying Protein Dynamics and Stability in a Living Organism. *Nat. Commun.* **2019**, *10* (1), 1179–1179. <https://doi.org/10.1038/s41467-019-09088-y>.
- (3) Wirth, A. J.; Platkov, M.; Gruebele, M. Temporal Variation of a Protein Folding Energy Landscape in the Cell. *J. Am. Chem. Soc.* **2013**, *135* (51), 19215–19221. <https://doi.org/10.1021/ja4087165>.
- (4) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in *Escherichia Coli*. *Nat. Chem. Biol.* **2009**, *5* (8), 593–599. <https://doi.org/10.1038/nchembio.186>.
- (5) Chen, W. W.; Freinkman, E.; Wang, T.; Birsoy, K.; Sabatini, D. M. Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism. *Cell* **2016**, *166* (5), 1324–1337.e11. <https://doi.org/10.1016/j.cell.2016.07.040>.
- (6) Diener, C.; Muñoz-Gonzalez, F.; Encarnación, S.; Resendis-Antonio, O. The Space of Enzyme Regulation in HeLa Cells Can Be Inferred from Its Intracellular Metabolome. *Sci. Rep.* **2016**, *6* (1), 28415. <https://doi.org/10.1038/srep28415>.
- (7) Theillet, F.-X.; Binolfi, A.; Frembgen-Kesner, T.; Hingorani, K.; Sarkar, M.; Kyne, C.; Li, C.; Crowley, P. B.; Gierasch, L.; Pielak, G. J.; Elcock, A. H.; Gershenson, A.; Selenko, P. Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114* (13), 6661–6714. <https://doi.org/10.1021/cr400695p>.
- (8) Ando, N.; Barquera, B.; Bartlett, D. H.; Boyd, E.; Burnim, A. A.; Byer, A. S.; Colman, D.; Gillilan, R. E.; Gruebele, M.; Makhataдзе, G.; Royer, C. A.; Shock, E.; Wand, A. J.; Watkins, M. B. The Molecular Basis for Life in Extreme Environments. *Annu. Rev. Biophys.* **2021**, *50* (1), 343–372. <https://doi.org/10.1146/annurev-biophys-100120-072804>.
- (9) Wennerström, H.; Vallina Estrada, E.; Danielsson, J.; Oliveberg, M. Colloidal Stability of the Living Cell. *Proc. Natl. Acad. Sci.* **2020**, *117* (19), 10113–10121. <https://doi.org/10.1073/pnas.1914599117>.
- (10) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517–520. <https://doi.org/10.1126/science.1208351>.
- (11) Boob, M.; Sukenik, S.; Gruebele, M.; Pogorelov, T. TMAO: Protecting Proteins from Feeling the Heat. **2020**. <https://doi.org/10.26434/chemrxiv.12665648.v1>.
- (12) Sukenik, S.; Ren, P.; Gruebele, M. Weak Protein–Protein Interactions in Live Cells Are Quantified by Cell-Volume Modulation. *Proc. Natl. Acad. Sci.* **2017**, *114* (26), 201700818–201700818. <https://doi.org/10.1073/pnas.1700818114>.

- (13) Ebbinghaus, S.; Dhar, A.; Douglas McDonald, J.; Gruebele, M. Protein Folding Stability and Dynamics Imaged in a Living Cell. *Nat. Publ. Group* **2010**, *7* (4), 319–319. <https://doi.org/10.1038/nmeth.1435>.
- (14) Davis, C. M.; Gruebele, M. Non-Steric Interactions Predict the Trend and Steric Interactions the Offset of Protein Stability in Cells. *ChemPhysChem* **2018**, *19* (18), 2290–2294. <https://doi.org/10.1002/cphc.201800534>.
- (15) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38, 27–28. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- (16) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637. <https://doi.org/10.1002/bip.360221211>.
- (17) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532. <https://doi.org/10.1016/j.bpj.2015.08.015>.
- (18) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcc.0c02536>.
- (19) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. Force Field Bias in Protein Folding Simulations. *Biophys. J.* **2009**, *96* (9), 3772–3780. <https://doi.org/10.1016/j.bpj.2009.02.033>.
- (20) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–49. <https://doi.org/10.1016/j.bpj.2011.03.051>.
- (21) Cohen, R. D.; Pielak, G. J. A Cell Is More than the Sum of Its (Dilute) Parts: A Brief History of Quinary Structure. *Protein Sci. Publ. Protein Soc.* **2017**, *26* (3), 403–413. <https://doi.org/10.1002/pro.3092>.
- (22) Cohen, R. D.; Pielak, G. J. Electrostatic Contributions to Protein Quinary Structure. *J. Am. Chem. Soc.* **2016**, *138* (40), 13139–13142. <https://doi.org/10.1021/jacs.6b07323>.
- (23) Leeb, S.; Sörensen, T.; Yang, F.; Mu, X.; Oliveberg, M.; Danielsson, J. Diffusive Protein Interactions in Human versus Bacterial Cells. *Curr. Res. Struct. Biol.* **2020**, *2*, 68–78. <https://doi.org/10.1016/j.crstbi.2020.04.002>.
- (24) Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C. Expasy, the Swiss Bioinformatics Resource Portal, as Designed by Its Users. *Nucleic Acids Res.* **2021**, *49* (W1), W216–W227. <https://doi.org/10.1093/nar/gkab225>.
- (25) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β -Sheet Protein. *J. Mol. Biol.* **2011**, *405* (1), 43–48. <https://doi.org/10.1016/j.jmb.2010.10.023>.
- (26) Wang, T.; Zhu, Y.; Gai, F. Folding of A Three-Helix Bundle at the Folding Speed Limit. *J. Phys. Chem. B* **2004**, *108* (12), 3694–3697. <https://doi.org/10.1021/jp049652q>.

- (27) Minton, A. P.; Wilf, J. Effect of Macromolecular Crowding upon the Structure and Function of an Enzyme: Glyceraldehyde-3-Phosphate Dehydrogenase. *Biochemistry* **1981**, *20* (17), 4821–4826. <https://doi.org/10.1021/bi00520a003>.
- (28) Tomokuni, Y.; Goryo, K.; Katsura, A.; Torii, S.; Yasumoto, K.; Kemnitz, K.; Takada, M.; Fukumura, H.; Sogawa, K. Loose Interaction between Glyceraldehyde-3-Phosphate Dehydrogenase and Phosphoglycerate Kinase Revealed by Fluorescence Resonance Energy Transfer-Fluorescence Lifetime Imaging Microscopy in Living Cells. *FEBS J.* **2010**, *277* (5), 1310–1318. <https://doi.org/10.1111/j.1742-4658.2010.07561.x>.
- (29) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. The Quantitative Proteome of a Human Cell Line. *Mol. Syst. Biol.* **2011**, *7* (549), 1–8. <https://doi.org/10.1038/msb.2011.82>.
- (30) Srivastava, D. K.; Bernhard, S. A. Metabolite Transfer via Enzyme-Enzyme Complexes. *Science* **1986**, *234* (4780), 1081–1086. <https://doi.org/10.1126/science.3775377>.
- (31) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (32) Nawrocki, G.; Wang, P.-H.; Yu, I.; Sugita, Y.; Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. **2017**. <https://doi.org/10.1021/acs.jpcb.7b08785>.
- (33) Rojas-Pirela, M.; Andrade-Alviárez, D.; Rojas, V.; Kemmerling, U.; Cáceres, A. J.; Michels, P. A.; Concepción, J. L.; Quiñones, W. Phosphoglycerate Kinase: Structural Aspects and Functions, with Special Emphasis on the Enzyme from Kinetoplastea. *Open Biol.* *10* (11), 200302. <https://doi.org/10.1098/rsob.200302>.
- (34) Das, S.; Chakrabarti, S. Classification and Prediction of Protein–Protein Interaction Interface Using Machine Learning Algorithm. *Sci. Rep.* **2021**, *11* (1), 1761. <https://doi.org/10.1038/s41598-020-80900-2>.
- (35) Petrov, D.; Zagrovic, B. Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Comput. Biol.* **2014**, *10* (5), e1003638–e1003638. <https://doi.org/10.1371/journal.pcbi.1003638>.
- (36) Yoo, J.; Aksimentiev, A. New Tricks for Old Dogs: Improving the Accuracy of Biomolecular Force Fields by Pair-Specific Corrections to Non-Bonded Interactions. *Phys Chem Chem Phys* **2018**, *20* (13), 8432–8449. <https://doi.org/10.1039/C7CP08185E>.
- (37) Dhar, A.; Samiotakis, A.; Ebbinghaus, S.; Nienhaus, L.; Homouz, D.; Gruebele, M.; Cheung, M. S. Structure, Function, and Folding of Phosphoglycerate Kinase Are Strongly Perturbed by Macromolecular Crowding. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (41), 17586–17591. <https://doi.org/10.1073/pnas.1006760107>.
- (38) Gasic, A. G.; Boob, M. M.; Prigozhin, M. B.; Homouz, D.; Daugherty, C. M.; Gruebele, M.; Cheung, M. S. Critical Phenomena in the Temperature-Pressure-Crowding Phase Diagram of a Protein. *Phys. Rev. X* **2019**, *9* (4), 041035. <https://doi.org/10.1103/PhysRevX.9.041035>.

CHAPTER 7: DATA SONIFICATION TO ENHANCE TEACHING AND RESEARCH

INTRODUCTION TO DATA SONIFICATION

A recurring challenge in communicating and understanding information about protein dynamics, in both educational and research settings, is the often-high dimensionality of the data. A protein's structure does not rely on just one or two strong bonds, as is the case for organic reactions or inorganic catalysis,² but rather on many weak interactions to hold the folded protein together. Therefore, to effectively study and communicate about protein dynamics, a few critical reaction coordinates must be selected to report on many of these weak interactions and represented carefully.^{3,4} Visualizations, including movies of simulated proteins folding,⁵ fluorescence spectra,⁶ and graphs reflecting various reaction coordinates,⁷ have thus far formed the backbone of protein dynamics representation.³

One has only to recall the crackle of a Geiger counter, alerting its users to the presence of alpha particles for over 100 years, to recognize that sound, too, is an effective tool for discovering, understanding, and communicating the time-dependent behavior of physical phenomena. We are, at the most fundamental level, multi-modal creatures, having evolved to navigate our 3D world using redundant, synergistic combinations of sensory inputs — vision, audition, proprioception, olfaction, and more — to form an accurate and reliable composite model of our surroundings.⁸ Data sonification is a field that uses this multi-modal human nature to enhance communication and understanding of data, in this case, protein dynamics.

Data sonification is a mapping from data (whether generated by a model, captured in an experiment, or gathered through observation) to one or more parameters of an audio signal or sound synthesis model for the purpose of better understanding, communicating or reasoning about the original model, experiment or underlying phenomenon.⁹ For example, you could map the value

* This chapter is reproduced in part with permission from:

Scaletti, C.; Rickard, M.M.; Hebel, K.; Pogorelov, T.V.; Taylor, S.A.; Gruebele, M. “Enhancing Visualization with Sonification to Explain Protein Folding Concepts.” *Under revision, J. Chem. Educ.* **2021**.¹
All data sonifications and videos referenced in this chapter are the work of C. Scaletti, K. Hebel, and S. Taylor. M. Rickard assisted in the construction of the lattice models referenced and helped (with M. Gruebele and T. Pogorelov) write the homework questions discussed in the lattice model section. In the transition state section, all analysis of raw MD trajectories was performed by M. Rickard, and all PCA was performed by C. Scaletti.

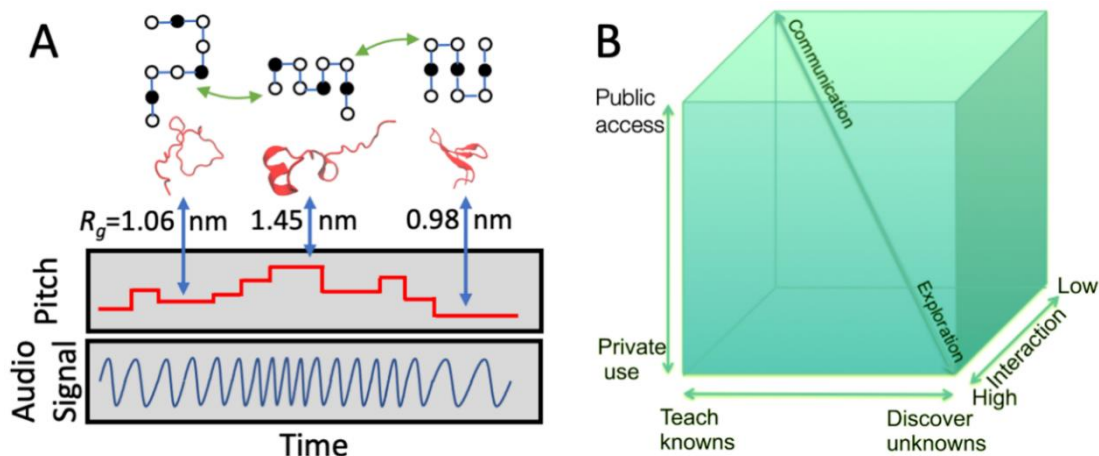


Figure 7.1: (A) A sonification mapping the radius of gyration (size) of a WW domain protein model to pitch of an audio signal. The lattice model as well as three ribbon structures of an atomistic model of WW domain are shown: unfolded (left), misfolded (middle, formed a non-native helix) and folded (right, all beta sheet). (B) The space of data sonification has three dimensions²⁷: intended use (vertical axis), instruct vs. discover (front axis) and interactivity. Our applications occupy two areas of the cube: one is a public presentation for teaching known concepts with a moderate degree of interactivity, and the other is a small, highly interactive group of researchers that use data sonification to guide protein folding research. Figure published in ref. 1.

of a measurement, such as the radius of gyration, to the pitch of an audio signal (Figure 7.1).

A well-designed data sonification (or visualization) mapping should be inference-preserving; in other words, any observations you make or conclusions you can draw in the target (sound) domain should also hold true in the source (simulation/measurement/observation) domain. This is the primary design goal for any data sonification — to preserve the critical features of and interrelationships within the original phenomenon.

Any data visualization, even one as simple as a 2D graph, is the outcome of multiple design decisions; the choice of which variables to represent, whether to use linear or log scales, how to normalize the data, whether to use dotted or solid lines, which colors to use, and so on, can all impact the clarity and communicative power of a visual graph. In data sonification, too, multiple design decisions — for example, the choice of synthesis model, deciding which data variables to map to which sound parameters, whether to use discrete or continuous values, and the scaling and timing of parameter changes — can all contribute to the intelligibility and clarity of the result.²⁷

Data sonification can lie anywhere along a continuum from exploration and discovery of previously unknown patterns to presentation of known information; it can range from highly interactive within a tight feedback loop to passive listening, and it can be carried out privately by a single individual, by a small research group, or in the public sphere. A particular instance of data

sonification can exist at any point in this continuum of passive-interactive, public-private, known-unknown space shown in Figure 7.1B.

Here, we seek to harness the power of data sonification to enhance our understanding of an abstract chemical concept—the energy landscape of protein folding—in two different areas of the cube shown in Figure 7.1B: teaching protein folding concepts to undergraduate students and informing scientific research about protein folding pathways.

SONIFICATION-ENHANCED LATTICE MODEL ANIMATIONS FOR TEACHING THE PROTEIN FOLDING REACTION

Introduction

Protein folding is one of the most important chemical reactions of life: it produces the enzymes that run our metabolism, the transcriptases that copy DNA, and the signaling proteins that tell our cells whether everything is functioning normally.¹⁰ Yet folding and its cousin, protein structure prediction,¹¹ are difficult subjects at the undergraduate level in a physical chemistry, biochemistry, or computational chemistry class, and often omitted despite their importance for chemistry and biology majors.¹²

Conceptually, it has been understood since Anfinsen's work in the 1960s that the folding mechanism is encoded in the amino acid sequence.¹³ Lattice models,¹⁴ invented in the 1970s and applied more widely starting in the 1980s, supported the idea that proteins have evolved to maximize consistency of many distributed interactions¹⁵ and minimize frustration among these interactions.¹ Even so, large proteins require unfolding/folding by chaperones because some frustrated interactions remain.¹⁶

A picture is worth 1000 words, and so visualization plays an important role in educating students about protein dynamics and communicating research in the field. This includes methods for teaching about energy landscapes using the funnel picture,¹⁷ combining visualization^{18–20} with a hands-on mechanical model of folding,²¹ watching color changes when proteins fold and unfold reversibly because "seeing is believing,"²² interacting with computer games that teach protein folding concepts,¹⁰ and creating physical models to help students visualize proteins in three dimensional space²³ or, if visually impaired, to explore those structures using the hands or mouth.²⁴

Whereas sonification has been applied to spectroscopy education,^{25,26} and data sonification of proteins has focused on structure,^{27–30} the goal of our present project was, instead, to convey to

our students some of the *dynamics* of protein folding. We apply a real-time lattice model of folding as a new educational tool and combine this new visualization with sonification to allow students to see and hear changes in the energy and conformation of a protein as it transitions from state to state. Feedback from a diverse set of science and music students indicates that presenting folding as an animated, sonified lattice model increases students' interest in the material and is perceived by them as useful in solving homework problems related to folding.

After introducing key concepts of protein folding, we describe how we constructed the sonified lattice models used to create the videos. We discuss three folding examples using the models, using lecture and homework (sample lecture and sample homework problems with solutions are included in the SI of ref. 1). Finally, we discuss students' responses to these materials and speculate on the outlook for future applications of sonification to dynamics in chemistry.

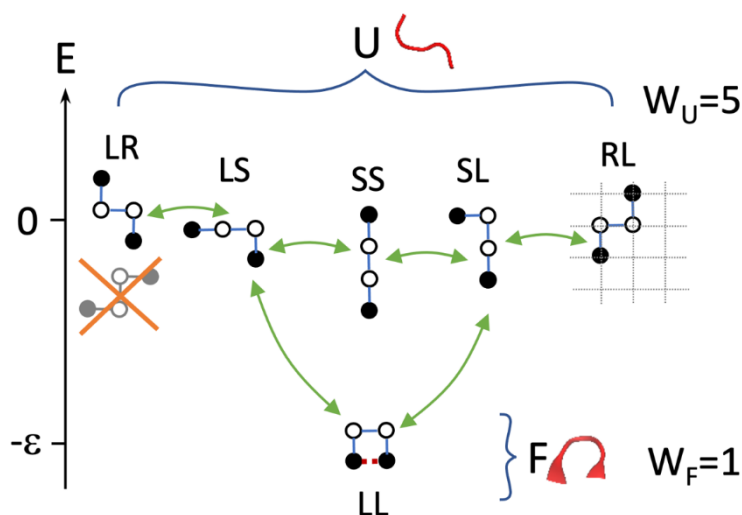


Figure 7.2: A lattice model with hydrophilic (light) and hydrophobic (dark) beads representing a 4 amino acid tetrapeptide that can form a ‘hairpin.’ The gray, dashed lines represent the square lattice on which beads are allowed to move. The axis on the left shows the energy of each state. The folded ‘state’ F at energy $-\epsilon$ contains only $W=1$ conformation, ‘LL’. In contrast, the unfolded ‘state’ U contains $W=5$ conformations and has conformational entropy $S_c=k_B \ln 5$. Green arrows show the allowed interconversions, by flipping one bond 90° , between conformations. The conformation crossed out in red is not counted because it can be obtained from the one above it by 2D rotation. The ‘LSR’ nomenclature for each conformation is explained in the Methods section, e.g., ‘LL’ refers to the chain making two left turns after the first segment, which is always oriented straight up. The red chain next to U and ribbon next to F show a conventional chain/ribbon representation of a disordered chain or hairpin turn.

Protein Folding Concepts

In order to help our students learn to speak the language used by protein scientists, we introduce four key concepts of the folding reaction. These are: solvation and hydrophobicity; energy and conformational entropy; funneled energy landscape; and finally, frustration and traps.

Solvation and hydrophobicity

While hydrogen bonds between amino acids organize secondary structure (e.g., α -helices and β -sheets),³¹ the solvent is critical in bringing amino acids together to form tertiary structure. When water solvates hydrophobic amino acid side chains such as phenylalanine, it forms cages around them with reduced water mobility and entropy. Folding a protein so these hydrophobic side chains release water into the bulk increases the mobility of water and hence overall entropy.³² This hydrophobic driving force is not entirely entropic, especially at low temperature, but increasing solvent entropy does play a major role in folding.

Energy and conformational entropy

When a protein folds, native contacts are made among side chains, ranging from salt bridges (a positively and a negatively charged amino acid coming together) to the hydrophobic contacts mentioned above. This reduces the energy E (or at constant pressure, the enthalpy H) of the protein as it folds. At the same time, the protein comes to occupy a much smaller number W of conformations, reducing its conformational entropy $S_c = \ln W$ (Figure 7.2).³³

Funneled energy landscape

Large molecules have an energy landscape, a generalization of the ‘potential surface’ for small molecules.³⁴ The full energy landscape depends on many coordinates and can include relevant solvent degrees of freedom that are part of the folding process, such as water molecules excluded from amino acid interfaces due to hydrophobicity. A minimal plot of the energy landscape features E as the vertical axis, and S_c as the horizontal axis, indicating how many conformations are possible at each energy. The energy landscapes of unevolved macromolecules have many local minima with many conformations at each energy (Figure 7.3, top).

A remarkable feature of many macromolecular biological reactions is energy-entropy compensation. For example, during folding of an evolved sequence, the reduction of conformational entropy of the peptide chain (Figure 7.3) is accompanied by simultaneous reduction in energy (or enthalpy at constant pressure) through contacts between amino acids: The evolved energy landscape has a funnel shape, with states of lower S_c also having lower E .³⁵ The sequence of amino acids is organized in a way that allows each residue to lower its energy on the order of RT while losing conformational entropy on the order of R ($R = 0.00831 \text{ kJ/[K}\cdot\text{mol]}$ is the universal gas constant). Thus, the free energy $FE = E - TS$ upon folding changes much less than its constituents E (energy) and TS (product of temperature and entropy) (Figure 7.3, bottom right),

resulting in low protein stability and small free energy barriers for folding; low stability is

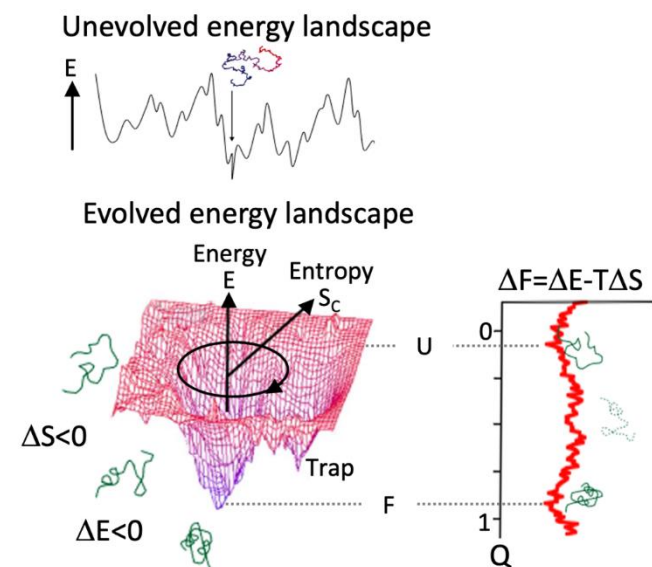


Figure 7.3: Top left: an unevolved energy landscape. Random contacts can lower the energy somewhat, but because the amino acid sequence does not facilitate all such contacts being made simultaneously, the energy landscape is ‘frustrated.’ Bottom left: an evolved energy landscape has a sequence that facilitates multiple simultaneous contacts, offsetting the loss of conformational entropy. It ‘minimizes frustration.’⁶ Bottom right: energy-entropy compensation means that the folding free energy $\Delta F = \Delta E - T\Delta S$ ($G = H - TS$ at constant pressure) is small and free energy barriers are small, making folding a very fast (milliseconds to hours at room temperature) reaction between states (here U and F) that have similar free energy under physiological conditions. Representative structures of the polypeptide chain are shown.

important for function by allowing fluctuations, and small barriers are important for getting to the folded state quickly.

Frustration and traps

In large molecules such as proteins, not all contacts that lower the energy can be made simultaneously, a concept known as frustration. Evolved sequences position amino acids in such a way that a specific 3D protein structure allows a maximum number of native contacts.¹⁵ This structure is thus minimally frustrated.³⁶ However, protein function, and the finite number (~ 20) of amino acids dictate that frustration cannot be eliminated completely, creating frustrated structures that are local minima on the energy landscape. Thus, even evolved landscapes retain some roughness, and local

minima can act as traps that delay the folding process (Figure 7.3, bottom left): in order to fold, the protein must first unfold in order to try a different path down the funnel. However, when a protein can escape a local minimum directly towards the folded state, we call it an ‘intermediate,’ and it could even accelerate the folding process.

Methods

A sonified lattice model for protein folding

To help explain the four key protein folding concepts to students having a minimum of thermodynamics and kinetics knowledge, we implemented a protein “lattice model”,^{37–39} not widely used previously in chemical education, and used this model to generate data-visualization animations with data-sonification sound tracks. We briefly describe our simple model and its implementation in software as a state machine, with more details provided in ref. 1.

The protein is represented on a 2D square lattice as in Figure 7.2: dark beads correspond to hydrophobic amino acids that interact with an energy $-\epsilon$ when they are adjacent; light beads correspond to hydrophilic amino acids that interact with energy $-\epsilon'$ when they are adjacent (see Figure 7.5 for such hydrophilic pairs). The total energy E of the system is the sum of all the pair energies $-\epsilon$ and $-\epsilon'$. We specifically used $\epsilon = 1$ kJ/mole and $\epsilon' = 0.25$ kJ/mole in our implementation of the problem for lecture demonstrations and homework assignments.

The configuration with the lowest energy is considered the ‘folded state’ (F in Figure 7.2); all others are lumped together as the ‘unfolded state’ (U in Figure 7.2), although one may further differentiate long-lived traps (defined below) within U. F and U are the macroscopically distinguished states in the model, which may contain one or more conformations. Thus, the entropy of state F is $S_F = k_B \ln(W_F) = 0$ as it corresponds to a single folded conformation, and the entropy of the state U is $S_U = k_B \ln(W_U) = k_B \ln 5$ in Figure 7.2 because U consists of 5 macroscopically indistinguishable conformations. Of course, the folded state of a real protein has several slightly different conformations, and better experiments could distinguish some of the conformations in the unfolded state from one another.

The protein folds by moving from one configuration to the next. In real-life, this motion is driven by thermal excitation (e.g. water molecules bumping into the protein) and forces between amino acids (e.g. attraction of two hydrophobic amino acids). Folding kinetics of the polypeptide chain in this model follows three simple rules: (1) A single randomly selected bond per time step can be flipped by 90° to make a transition between two conformations, as indicated by the green arrows in Figure 7.2. (2) Beads cannot be superimposed on the same lattice point, they avoid each other due to steric hindrance. (3) To satisfy thermodynamic equilibrium, flips are chosen by Metropolis sampling: ⁴⁰ if a randomly chosen flip lowers the total energy E , it is automatically accepted. If the flip would raise the energy, it is only accepted if the Boltzmann factor $\exp[-\epsilon E/k_B T]$ is greater than a number randomly chosen between 0 and 1. This procedure produces the correct thermodynamic equilibrium. Thus, the model will tend to fold into the lowest energy state at low temperature, and it will tend to unfold to the state U at high temperature because that state contains more conformations that can be visited at high temperature. A ‘trap’ is a conformation or a group of conformations, of energy higher than the folded state F, from which all allowed transitions lead upward in energy.

In this model, each move corresponds to a time step Δt from time t to $t + \Delta t$; the real-world characteristic duration for such steps is the time required for the diffusion of an amino acid residue over a distance comparable to its size, which is about 10 ns.⁴¹ The ‘flipping’ dynamics of the model thus mimic the real dynamics of a polypeptide chain moving continuously in space.

Implementing the lattice model of protein dynamics as a state machine

We used the sound design platform Kyma⁴² to implement the lattice model as a state machine and to map observables (variables) from each conformation to sound parameters and images. Here we give a brief overview of the implementation, with a more detailed description provided in the supplementary information of ref. 1. As described in detail in the SI of ref. 1, a lattice protein object is defined with attributes such as the length of the protein (number of beads). Then Kyma generates a sequence in time by changing the coordinates of the lattice protein object. Finally, interesting observables, such as the size of the protein, are mapped out in time as well.

A conformation is represented in Kyma as a string of directions, starting with the first bond pointing ‘up’, and successively labeling each joint (angle between two bonds) as 90° to the left (L), 90° to the right (R), or straight ahead (S). For the 4-bead hairpin in Figure 7.2, this representation yields $3^3 = 27$ shapes, identified by their joint directions (starting from the first bead): {LS, SL, SR, RS, RL, SS, LR, RR, LL}. If the first and last beads are considered equivalent and conformations that interconvert by rotation in the 2D plane are considered to be identical, then the simple model {LS=SR, SL=RS, RL, SS, LR, LL} in Figure 7.2 is obtained. Transitions between conformations can be represented by a transition probability matrix. For example, if a flip from ‘SL’ to ‘LL’ is randomly chosen in Figure 7.2, it is accepted with probability 1 because it is downhill in energy.

Calculating chemically interesting observables

Observables analogous to those computed in all-atom molecular dynamics (MD) simulations (e.g., Q) or measured in folding experiments (e.g., R_g) can introduce a conceptual link between these 2D lattice models and the kinds of behaviors students might see reflected in data found in the literature.

The observables, $\mathbf{O} = (E, R_g, d_{ee}, Q)$, computed for each conformation, are defined as:

- The total energy E of the conformation, which depends on the number and type of bonds present

- The radius of gyration R_g , which measures the diameter of the protein (see Appendix D for full definition)
- The end-to-end distance d_{ee} , also labeled ‘E2E’ in some videos, from the first bead to the last bead⁴³
- The fraction of native contacts Q , ranging from 0 (unfolded) to 1 (folded) (see Appendix D for full definition)
- To generate a characteristic sound for each conformation (with observables $\mathbf{v}(t)$) or transitions (using $\mathbf{v}(t+\Delta t)-\mathbf{v}(t)$), Kyma maps each reaction coordinate associated with that time step to a sound synthesis or processing parameter and (optionally) displays an image associated with that conformation.

Implementing a lattice model as a nondeterministic finite state machine — the basis of formal languages and the theory of computation — invites analogies between chemical reactions and information processing. For example, in the lectures (included in the SI of ref. 1), biochemical processes are described in terms of information theory, noting that chiral synthesis of a protein chain that encodes its own folded structure through the amino acid sequence (the information) relies on information that was pre-stored at great ATP expense over billions of years of evolution.

Results

In the classroom and homework, we focused on the folding dynamics of four simple lattice proteins: an unevolved polypeptide, a β -hairpin and α -helix,³¹ and the WW domain, a small protein that helps prevent cell apoptosis (death) among its other functions.⁴⁴

Unevolved sequence

An arbitrary chain of amino acids generally does not fold into a unique shape: it has many low-energy states where the peptide can get trapped (Figure 7.3, top). Similarly, a lattice model with an arbitrary string of hydrophobic and hydrophilic beads does not necessarily result in a well-defined folding funnel with a single folded state.

For example, the 6-bead chain ①②③④⑤⑥ (where beads 2 and 6 are hydrophobic) does not yield a graph with a single minimum energy state; instead, it has 6 lower energy conformations connected to 30 conformations at higher energy (Figure 7.4 and [Video D.2](#)). A chain with this kind of energy landscape cannot fold to a unique native state.

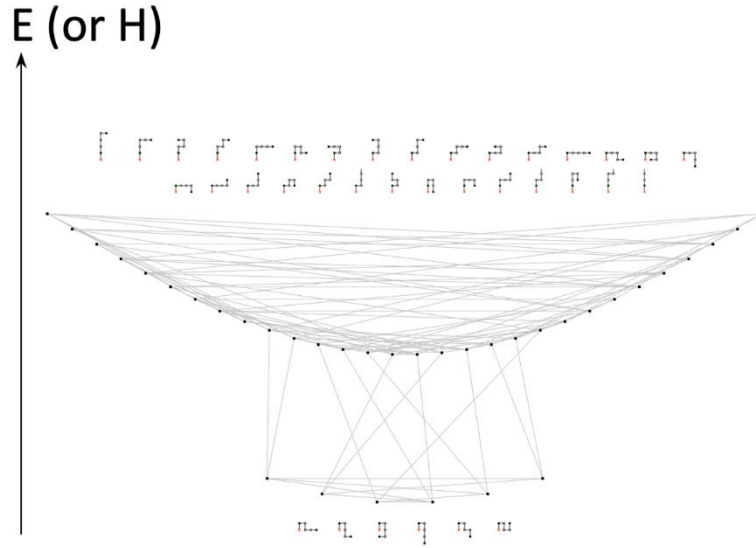


Figure 7.4: Energy landscape for a random (unevolved) sequence of 6 beads. There is no well-defined native state, with six very different structures at the same energy, analogous to Figure 7.3. The low and high energy structures are shown, not in any particular order. Sonification at [Video D.2](#). Figure prepared by Carla Scaletti and published in ref. 1.

Beta hairpin

To see an example of how the process of evolution can store information to generate spatial structure, try changing the position of one of the hydrophobic beads from position 6 to 5 to make the sequence ①②③④⑤⑥ (so beads 2 and 5 are hydrophobic). The state space for this 6-bead β -hairpin model, like the 4-bead example, is generated by taking the space of all $3^4 = 81$ possible conformations, removing any conformations that overlay beads, that are rotations of another conformation in the 2D plane, or that are equivalent under order-reversal. One can also enforce a preference for chirality (e.g., every chain begins as either a straight or as a right turn), resulting in a state space of 22 shapes and a 4-level energy landscape (left example in Figure 7.5A and [Video D.3](#)).

Alpha helix

The α -helix lattice model is based on a chain of 8 beads ①②③④⑤⑥⑦⑧ with hydrophobic elements in positions 2 and 7. As with the 4-bead and β -hairpin models, we start by taking the space of all $3^6 = 729$ possible combinations of left, right, or straight angles at each joint, remove any conformations that cross themselves or that are rotations of another conformation in the 2D plane. To mimic chirality in actual proteins, we allow only the right-handed variant of each shape. To mimic the N- and C- termini of actual proteins, in this model, we do not consider the

first and last beads to be equivalent, so we do not remove shapes that would be equivalent under order-reversal of the beads. After this pruning, there are 272 shapes left. Since α -helical peptides form near-optimal linear hydrogen bonds, we give only the helical hydrogen bonds $\square=1$ kJ/mole (Figure 7.5B and [Video D.4](#)).

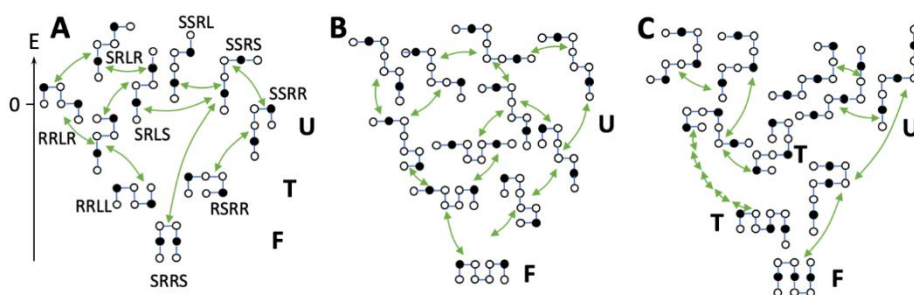


Figure 7.5: Summary of (A) the beta hairpin ([Video D.3](#)), (B) alpha helix ([Video D.4](#)) and (C) WW domain ([Video D.5](#)) lattice models, showing a small selection of the conformations for each, with allowed transitions as green arrows. Each protein has a unique lowest energy conformation, the folded state 'F', unfolded conformations that make up the unfolded state U, and in some cases traps T which have no path downhill in energy towards the folded state. The energy landscapes are funnel-shaped. The hairpin landscape in (A) also highlights the 'RLS' nomenclature discussed in the text. Figure prepared by Carla Scaletti and published in ref. 1.

WW domain

The WW domain is our most complex lattice model, suitable for introduction after students have first worked with the simpler examples in homework. This model is based on a 9-bead string where the first and last beads are not equivalent, and with hydrophobic residues in positions 2, 5, and 8 (Figure 7.5C, Figure 7.6, and [Video D.5](#)). There are 740 shapes in the state space and 10 discrete energy levels in the energy landscape. In this model, the traps correspond to incorrect secondary structure, such as a helix or mis-registered hairpins.

Visualizing and Sonifying Lattice Models

For each state, we map one or more features of the current conformation to one or more parameters of a sound synthesis algorithm; simultaneously we display an image of the conformation (see Figure 7.6) alongside an image of the energy funnel highlighting the current state (as a yellow square) and the potential transitions from that state to the next state (as purple edges). As discussed earlier, a single time-step (one Monte Carlo move of the chain) corresponds to about 10 nanoseconds for a natural amino acid chain.

In the lecture ([Playlist D.7](#)) and homework ([Playlist D.8](#)) videos, we used a variety of sound mappings, each one designed to bring out a particular aspect of the model in order to help the students make a comparison or answer a question.

Teaching the Four Key Concepts of Protein Folding

A course instructor could utilize a combination of the slides (provided in the SI of ref. 1), figures, and videos to introduce the four key concepts of protein, for example:

Solvation and hydrophobicity: The WW model forms the folded state when all three

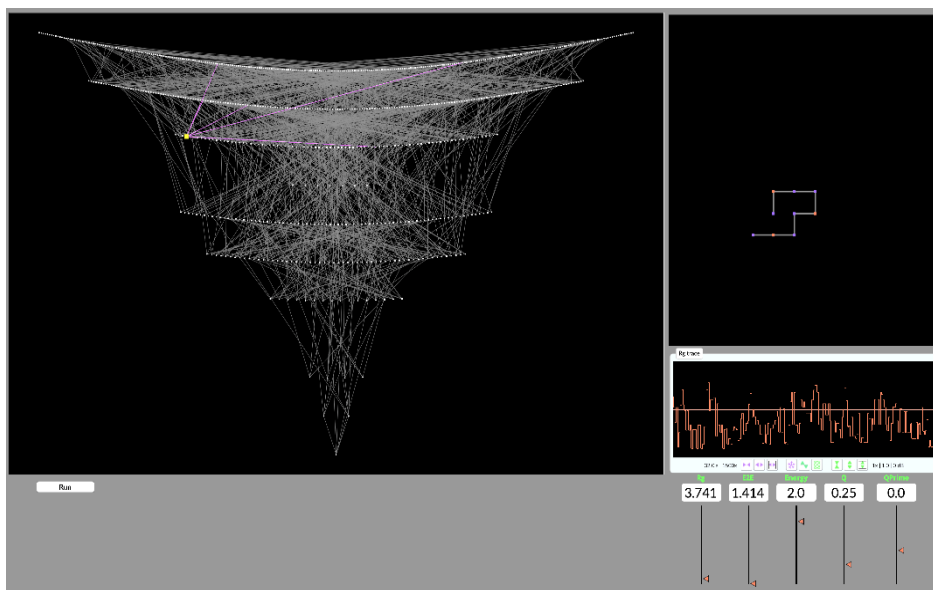


Figure 7.6: Screenshot from an animation of a state machine that models WW domain folding. The funnel is shown on the left, with the current state highlighted in yellow and all possible transitions to the next conformation highlighted in purple; other allowed transitions are shown as a white web connecting conformations. The eight traps, which can be escaped only by going to higher energy, are in the fourth row from the top. The currently occupied conformation (whose shape is shown on the right) is a “degenerate trap” which can be escaped only by going to a conformation at the same energy (degeneracy = equal energy). The “oscilloscope display” on the right can highlight variables such as fraction of native contacts Q or (in this case) radius of gyration R_g , and the observables are displayed on meters below the oscilloscope. In [Video D.6](#), you hear the model at slightly above its folding temperature T_m , where it explores both the folded state and unfolded conformations. Figure prepared by Carla Scaletti and published in ref. 1.

hydrophobic beads are lined up, excluding the solvent as much as possible from their surfaces (Figure 7.5C and [Video D.9](#)). At high temperature, solvation “wins” over hydrophobicity, even though hydrophobicity increases with temperature,³² because the solvated chain has more conformational entropy. ([Video D.10](#))

Energy and conformational entropy: In the unevolved peptide (Figure 7.4 and [Video D.2](#)), the entropy of the lowest energy state is $S_c = k_B \ln(6) \neq 0$. For the evolved proteins in Figure 7.5, $S_c = k_B \ln(1) = 0$ ([Video D.3](#)). The energy of the native state of the WW domain lattice model is -2.5

kJ/mole, due to two hydrophobic interactions ($\square=1$ kJ/mole each) and two hydrophilic interactions (pairs of light beads, $\square'=0.25$ kJ/mole each) ([Video D.5](#)).

Funneled energy landscape: Figure 7.6 shows the funneled landscape for WW domain, and all the videos except for [Video D.2](#) show the protein hopping between conformations, occasionally folding into the unique folded conformation.

Frustration and Traps: In the WW domain ([Video D.11](#)), and hairpin ([Video D.12](#)) models the model protein gets stuck in traps that have a small radius of gyration but are not the folded state. In **Sample homework problem e** (below), students are encouraged to find these traps by listening to the value of R_g mapped to the frequency of the audio signal (Figure 7.1A and [Video D.6](#)).

Sample lectures We developed two sample lectures without previously copyrighted material that are suitable at various levels from sophomore (very little background in thermodynamics or partition functions) to seniors (some background in thermodynamics and partition functions). (Included with SI in ref. 1)

Sample homework problems

The problem set and solutions are presented in the SI of ref. 1 and the solution to each problem relies on the information presented in the associated animation / sonification video.

Problems a and b: To introduce students to the experience of analytical listening and give them some practice extracting information from sound, we started with a simple coin-toss model ([Video D.13](#)). Each virtual coin toss triggers a sound event: a high frequency tone corresponds to a result of Heads and a lower frequency tone corresponds to Tails. Students are asked to identify whether a coin is “fair” or biased towards heads or tails ([Video D.14](#)), and then to quantitatively establish their reasoning.

Problems c and d: The homework progresses to more difficult questions involving protein folding models of a 6-bead hairpin such as the one in Figure 7.4B. After listening to examples of the hairpin model at temperatures above ([Video D.15](#)), below ([Video D.16](#)), and at the ‘folding temperature’ ([Video D.17](#)), students are asked to determine whether the video of the protein at an unknown temperature is at a temperature low enough to be mostly in the folded state ($T < T_m$) ([Video D.18](#)). The sound-mapping is similar to the one used in the coin toss so as to facilitate comparisons between the simple coin-toss model and the 6-bead lattice model.

Problem e: These problems focus on the WW domain model (Figure 7.5C and Figure 7.6) and ask more subtle questions about identifying traps by ear (this can be solved by noticing that a smaller radius of gyration is mapped to a lower pitch) ([Video D.6](#)).

Problem f: In a final, open-ended homework question, the students are invited to view a visualization / sonification of an all-atom molecular dynamics simulation the CHARMM22* force field⁴⁵ and then to speculate about how one could enhance the lattice model simulations to make them closer-to-life. ([Video D.19](#))

We presented the lecture and homework assignment to interested participants in the remote learning course PHYS 498 at the University of Illinois at Urbana-Champaign (38% undergraduates, 62% graduates). The qualitative feedback was positive, with listeners generally noting that watching the lattice movies and hearing various observables **O** of the protein sonified gave them a much more dynamic picture of protein folding, as opposed to a reaction that simply jumps directly from reactant to product in a single step. Comments included *‘I don’t know how I can incorporate sonification in my research but I’ll definitely learn more about it,’* and *‘lectures are very intuitive and easy to understand even if the subjects are complicated.’* For a more detailed description of student responses and feedback, see ref. 1.

Discussion

In the real world, human beings acquire information via combinations of multiple sensory input channels. Even in machine learning, performance on information retrieval and classification tasks is more accurate and robust when two sensory modalities are combined.⁴⁶ In this project, we tried to engage our students’ multi-modal perceptual abilities by using both visualization and sonification of a lattice model to help them understand the dynamics of protein folding.

Sonification can reinforce visualization. For example, high/low-pitched sound conveys high/low energy levels of a conformation (e.g. Figure 7.5 and Figure 7.6) as intuitively as its visual location along the energy axis, while increasing accessibility. While it is known that adding speech audio to written text to increase redundancy can hinder learning,⁴⁷ when non-speech audio and imagery channels are redundant (and not conflicting), memory for information increases with increasing redundancy.⁴⁸

Perhaps more importantly, sonification can complement visualization, that is, it can present aspects of the data that are difficult to visualize. In agreement with earlier experiments,⁴⁹ we also saw evidence that data-driven sound tracks can increase the bandwidth of scientific visualization,

and that sound and animation lend themselves particularly well to the presentation of time-dependent phenomena. When the image and sound channels are complementary, the sound is not just reinforcing the image; it is crucial to conveying the desired information. The authors have found that with minimal practice, they could watch protein visualizations while listening to a sonification to monitor changes in a reaction coordinate that was not explicitly present in the visualization, such as R_g (size) of the conformation. Indeed, we found it easier to identify compact states based on sound than it was to identify them in the visual oscilloscope trace in Figure 7.6. Also, use of sonification allowed us to identify compact states while running the simulation at a faster pace, making it more engaging for the students (who were under time pressure to finish their homework quickly). This phenomenon can enable students to process other variables that are difficult to visualize quickly in real-time, such as protein size (often distorted by the orientation of the protein in a simulation on a 2D computer screen) or solvation (where the huge number of water molecules becomes impossible to count or visually assess).

Outlook

With the ongoing shift to electronic presentation of educational materials, there are now fewer barriers to including information-bearing sound to accompany other modalities (graphics, animation, and interactive problem-solving discussions) in science classes. By increasing the sensory channels of educational materials to include data sonification (non-speech audio), educators can increase the bandwidth of these materials and engage the multiple learning styles of students.

Several future improvements could make these materials even more useful. For instance, adding interactive controls to the lattice models could allow students to formulate hypotheses (“the average radius of gyration of a protein increases with temperature”) and perform computer-experiments to test these hypotheses in an interactive way. Thus, the use of sonification could expand to virtual labs. Sonification could also be applied to more realistic models, such as solvation of a protein visualized by an all-atom molecular dynamics simulation, where plotting the water molecules can obscure the protein, and solvation can be difficult to judge visually.

Data-driven sound has been applied as an educational tool in the past to understand structure²⁷ and spectra,²⁶ and here we see that dynamics, or other data that are easily represented by time series, are also excellent candidates for sonification.

Our initial foray into using data-driven sound and animation as an adjunct to lectures, discussion sections and homework for undergraduate and graduate students has been encouraging. The listeners found that sonification was not only engaging but was a useful tool in problem-solving; most importantly, even specialist co-authors in the protein folding field of research found that it helped increase their intuition for how proteins fold and misfold over time. This encourages us to continue to explore other opportunities for enhancing our lectures and homework with data-driven sound and animations, as well as future assessment with control groups.

USING DATA SONIFICATION TO ANALYZE THE TRANSITION STATES OF A WW DOMAIN

Introduction

In addition to using data sonification as an educational tool to teach protein folding at an intermediate level, we also used data sonification extensively to guide our own protein folding research. As discussed above, sonification is a powerful tool to add dimensionality to a data representation by taking advantage of humans' natural ability to process auditory and visual input in parallel.⁸ To put this to the test in a more interactive, exploratory part of the cube shown in Figure 7.1B, we used data sonification to characterize the folding pathways of GTT, a fast-folding WW domain.^{50,51}

Energy landscape theory states proteins may navigate any number of parallel paths to reach the native state, but experimental and computational limitations make it difficult to observe and characterize these paths.^{4,52} Further, effective reaction coordinates of interest^{53,54} must be chosen to identify the different intermediates that occur as the protein folds along these pathways. Indeed, it has been suggested that researchers' struggle to represent and visualize the high-dimensional nature of a protein folding funnel may hinder the discovery of rare protein folding pathways.⁴

Multiple computational and experimental studies have addressed the folding paths of WW domains.^{55–58} According to current experimental and computational evidence, GTT folds via two main pathways—one in which hairpin 1 forms first, and one in which hairpin 2 folds first.^{55,57} However, we hypothesized that data sonification could aid in identifying rarer parts of the energy landscape and enable us to more effectively characterize protein transition states.

Dataset Selection and Reaction Coordinates

In our early sonifications to track protein folding, we used our simulation of GTT in water ([Video D.19](#)).^{45,51} After these initial tests, we chose to use data from a > 650 μ s atomistic simulation of GTT WW domain in water conducted by Shaw and co-workers.⁵⁰ This trajectory, with snapshots taken every 200 ps, contained more than 20 complete folding and unfolding transitions of GTT, allowing us to more broadly sample GTT's conformational space ([Video D.20](#)).

To monitor the folding of the GTT in this simulation, we relied on three reaction coordinates: radius of gyration (R_g), solvent accessible surface area (SASA), and hydrophobic SASA. These three coordinates were selected because they all (indirectly) report on the solvation of the protein of interest and do not rely on prior knowledge of the protein's native state (unlike RMSD or Q).

Principal Component Analysis

Gaussian weighted averages and standard deviations of each reaction coordinate were calculated in 500 ns, 250 ns, and 125 ns intervals around each time point. To quantify GTT's folding transition, we performed Principal Component Analysis (PCA) to these sets of means and standard deviations. In all three sets (500 ns window, 250 ns window, and 125 ns window), we found that two principal components capture > 97% of the variance in the data. Plotting these principal components against each other clearly separates the folded and unfolded ensembles of GTT (Figure 7.7a). To identify transitions between these two ensembles and incomplete transitions (in which a protein starts to fold or unfold, but does not complete the transition and rejoins its original ensemble), we applied a rarity function to PC 1 and imposed a cutoff between the folded and unfolded ensembles and "transitional space." Using this rarity function, we identified 26 folding and unfolding events and 92 incomplete transitions.

A Folding "Superhighway"

When we viewed the transition events in PC 1/PC 2 space, we observed a large fraction of transitions (approximately half of completed transitions) passed through a relatively small section of the transition space (shaded gray in Figure 7.7b).

We hypothesize that this narrow section is a “superhighway”—that is, a more accessible place in the energy landscape for a folding or unfolding GTT to pass through. GTT appears to fold and unfold more quickly when it passes through the superhighway (superhighway transitions only

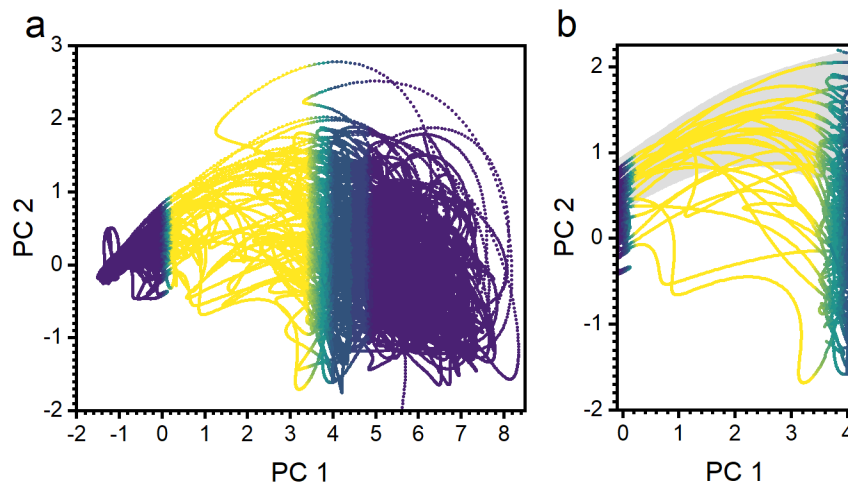


Figure 7.7: (a) GTT's dynamics mapped in PC 1/PC 2 space (averages over 500 ns). Points are colored by rarity (yellow—rare, transition state, purple, common). The folded and unfolded ensemble are shown in purple on the left and right sides respectively. (b) Folding/unfolding transitions of GTT mapped in PC 1/PC 2 space. Gray shading denotes the hypothesized “superhighway.”

meet the rarity function cutoff described above for ~100 ns, while off-superhighway transitions can take up to five times longer). However, the presence of complete transitions off this superhighway suggests that while the superhighway is the most favorable folding path, other folding pathways exist.

When we evaluated the structures of GTT that pass through the superhighway and those that do not, we find striking differences in their conformational ensembles. Structures that pass through the superhighway region agree well with experimental and computational studies of WW domain folding pathways— folding begins by turns forming, followed by the β -hairpin “zipping” into place, and finally, the other hairpin folds (Figure 7.8).^{55,56,59,60} In these superhighway transitions, both hairpin 1 and hairpin 2 form first, which agrees with experimental findings that indicate that GTT’s folding intermediate is comprised of a mixture of hairpin 1 folded and hairpin 2 folded states (although hairpin 1 appears to be favored in our simulations).⁵⁵

In contrast, transitions and incomplete transitions that occur off the superhighway do not agree with currently known pathways.^{55,57} For example, in the unfolding event that occurs at 552 μ s, strand 1 of GTT unfolds and forms an α -helix before the protein unfolds completely (Figure 7.8). In fact, strand 1 forms an α -helix in many off-superhighway transitions and incomplete

transitions, which is consistent with the α -helical character we observed in GTT in the *E. coli* models.⁴⁵ These differences in the conformational ensembles of the superhighway and off-superhighway transitions further support the hypothesis that the off-superhighway transitions represent rare routes through the energy landscape.

Outlook

Here, we used a combination of data sonification and visualization to find previously-unknown rare transition pathways of GTT WW domain. Preliminary results indicate that so-called “superhighway” transition pathways occupy a distinct place in the conformational space of the protein. As we continue this line of inquiry, we plan to further

study the differences in the superhighway and off-superhighway transitions’ respective conformational ensembles. For example, we hypothesize that local solvation, particularly of GTT’s turns, may drive the conformational differences observed in different pathways. In future sonifications, we plan to explore the role of solvent in these observed pathways by probing the intra-protein and protein-solvent hydrogen bonds that form during both superhighway and off-superhighway transitions. We will use data sonification to simultaneously monitor the dozens of potential hydrogen bonding sites on the protein during folding to determine how the hydrogen bonding network differs along different folding pathways.

Throughout the project, data sonification was integral to identifying the various transition pathways and interpreting the data. Frequently, unusual transition paths were identified not just by their principal components, but by the accompanying sonification sounding different than a superhighway transition. These different sounds then prompted us to go back and evaluate the data more closely, often discovering structural information that would have easily been missed in a visualization. Additionally, data sonification was very effective at adding dimensionality to

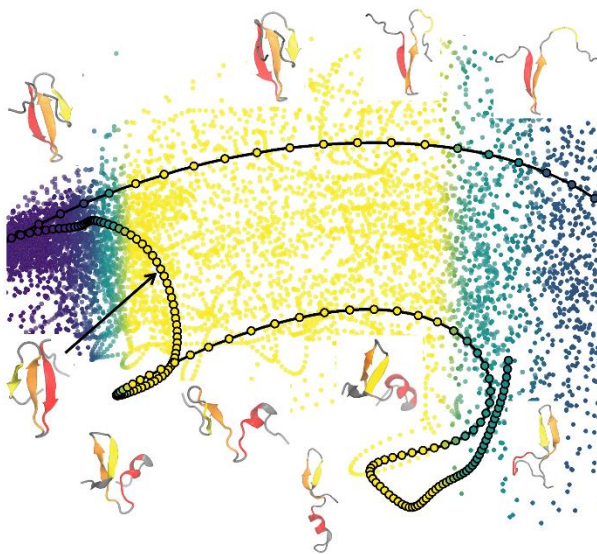


Figure 7.8: Selected transitions that are and are not on the hypothesized superhighway (averages over 250 ns). (Black line, top) “Superhighway” transition from 409.75 to 409.85 μ s, with snapshots of GTT at 409.8, 409.82, 409.84, and 409.86 μ s. (Black line, bottom) “Non-superhighway” transition from 552.3 to 552.8 μ s, with snapshots of GTT at 552.4, 552.46, 552.54, 552.64, 552.78, and 552.89 μ s. Points in each transition are plotted 6 ns apart. Residues that make up β -strands 1, 2, and 3 in GTT’s native structure are colored red, orange, and yellow respectively in the snapshots.

existing visualizations—a sonification could report on not just a single reaction coordinate unrelated to the visualization, but on multiple reaction coordinates mapped to different sound qualities, adding multiple dimensions of data to the visualization. Strikingly, despite the added dimensionality of the dataset, we found watching and listening to a visualization/sonification pair easier than visually checking multiple graphs or reaction coordinates simultaneously. These projects illustrate the advantages of data sonification as a tool to both reinforce information in an educational setting and to analyze higher dimensional datasets than can be described with visualization alone.

REFERENCES

- (1) Scaletti, C.; Rickard, M. M.; Hebel, K.; Pogorelov, T. V.; Taylor, S.; Gruebele, Martin. Enhancing Visualization with Sonification to Explain Protein Folding Concepts. *J Chem Educ* **2021**, *Under revision*.
- (2) Guin, D.; Gruebele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chem. Rev.* **2019**, *119* (18), 10691–10717. <https://doi.org/10.1021/acs.chemrev.8b00753>.
- (3) Ferina, J.; Daggett, V. Visualizing Protein Folding and Unfolding. *J. Mol. Biol.* **2019**, *431* (8), 1540–1564. <https://doi.org/10.1016/j.jmb.2019.02.026>.
- (4) Lapidus, L. J. The Road Less Traveled in Protein Folding: Evidence for Multiple Pathways. *Curr. Opin. Struct. Biol.* **2021**, *66*, 83–88. <https://doi.org/10.1016/j.sbi.2020.10.012>.
- (5) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *J. Phys. Chem. Lett.* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpcllett.9b01556>.
- (6) Gasic, A. G.; Boob, M. M.; Prigozhin, M. B.; Homouz, D.; Daugherty, C. M.; Gruebele, M.; Cheung, M. S. Critical Phenomena in the Temperature-Pressure-Crowding Phase Diagram of a Protein. *Phys. Rev. X* **2019**, *9* (4), 041035. <https://doi.org/10.1103/PhysRevX.9.041035>.
- (7) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcb.0c02536>.
- (8) Ghazanfar, A.; Schroeder, C. Is Neocortex Essentially Multisensory? *Trends Cogn. Sci.* **2006**, *10* (6), 278–285. <https://doi.org/10.1016/j.tics.2006.04.008>.
- (9) Scaletti, C. *Sonification ≠ Music*; Dean, R. T., McLean, A., Eds.; Oxford University Press, 2018; Vol. 1. <https://doi.org/10.1093/oxfordhb/9780190226992.013.9>.
- (10) Miller, J. A.; Khatib, F.; Hammond, H.; Cooper, S.; Horowitz, S. Introducing Foldit Education Mode. *Nat. Struct. Mol. Biol.* **2020**, *27* (9), 769–770. <https://doi.org/10.1038/s41594-020-0485-6>.
- (11) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (12) Barton, J. S. Protein Denaturation and Tertiary Structure. *J. Chem. Educ.* **1986**, *63* (4), 367. <https://doi.org/10.1021/ed063p367>.
- (13) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>.
- (14) Hinds, D. A.; Levitt, M. A Lattice Model for Protein Structure Prediction at Low Resolution. *Proc. Natl. Acad. Sci.* **1992**, *89* (7), 2536–2540. <https://doi.org/10.1073/pnas.89.7.2536>.
- (15) Go, N. The Consistency Principle in Protein Structure and Pathways of Folding. *Adv. Biophys.* **1984**, *18*, 149–164. [https://doi.org/10.1016/0065-227X\(84\)90010-8](https://doi.org/10.1016/0065-227X(84)90010-8).

- (16) Guin, D.; Gruebele, M. Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function. *Chem. Rev.* **2019**, *119* (18), 10691–10717. <https://doi.org/10.1021/acs.chemrev.8b00753>.
- (17) Martínez, L. Introducing the Levinthal's Protein Folding Paradox and Its Solution. *J. Chem. Educ.* **2014**, *91* (11), 1918–1923. <https://doi.org/10.1021/ed300302h>.
- (18) Franco, J. Online Gaming for Understanding Folding, Interactions, and Structure. *J. Chem. Educ.* **2012**, *89* (12), 1543–1546. <https://doi.org/10.1021/ed200803e>.
- (19) Ship, N. J.; Zamble, D. B. Analyzing the 3D Structure of Human Carbonic Anhydrase II and Its Mutants Using Deep View and the Protein Data Bank. *J. Chem. Educ.* **2005**, *82* (12), 1805. <https://doi.org/10.1021/ed082p1805>.
- (20) Abualia, M.; Schroeder, L.; Garcia, M.; Daubenmire, P. L.; Wink, D. J.; Clark, G. A. Connecting Protein Structure to Intermolecular Interactions: A Computer Modeling Laboratory. *J. Chem. Educ.* **2016**, *93* (8), 1353–1363. <https://doi.org/10.1021/acs.jchemed.5b00910>.
- (21) Prigozhin, M. B.; Scott, G. E.; Denos, S. Mechanical Modeling and Computer Simulation of Protein Folding. *J. Chem. Educ.* **2014**, *91* (11), 1939–1942. <https://doi.org/10.1021/ed400719c>.
- (22) Carlson, T. M.; Lam, K. W.; Lam, C. W.; He, J. Z.; Maynard, J. H.; Cavagnero, S. Naked-Eye Detection of Reversible Protein Folding and Unfolding in Aqueous Solution. *J. Chem. Educ.* **2017**, *94* (3), 350–355. <https://doi.org/10.1021/acs.jchemed.6b00507>.
- (23) Meyer, S. C. 3D Printing of Protein Models in an Undergraduate Laboratory: Leucine Zippers. *J. Chem. Educ.* **2015**, *92* (12), 2120–2125. <https://doi.org/10.1021/acs.jchemed.5b00207>.
- (24) Baumer, K. M.; Lopez, J. J.; Naidu, S. V.; Rajendran, S.; Iglesias, M. A.; Carleton, K. M.; Eisenmann, C. J.; Carter, L. R.; Shaw, B. F. Visualizing 3D Imagery by Mouth Using Candy-like Models. *Sci. Adv.* **2021**, *7* (22), eabh0691. <https://doi.org/10.1126/sciadv.abh0691>.
- (25) Garrido, N.; Pitto-Barry, A.; Soldevila-Barreda, J. J.; Lupan, A.; Boyes, L. C.; Martin, W. H. C.; Barry, N. P. E. The Sound of Chemistry: Translating Infrared Wavenumbers into Musical Notes. *J. Chem. Educ.* **2020**, *97* (3), 703–709. <https://doi.org/10.1021/acs.jchemed.9b00775>.
- (26) Pereira, F.; Ponte-e-Sousa, J. C.; Fartaria, R. P. S.; Bonifácio, V. D. B.; Mata, P.; Aires-de-Sousa, J.; Lobo, A. M. Sonified Infrared Spectra and Their Interpretation by Blind and Visually Impaired Students. *J. Chem. Educ.* **2013**, *90* (8), 1028–1031. <https://doi.org/10.1021/ed4000124>.
- (27) Yu, C.-H.; Qin, Z.; Martin-Martinez, F. J.; Buehler, M. J. A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. *ACS Nano* **2019**, *13* (7), 7471–7482. <https://doi.org/10.1021/acsnano.9b02180>.
- (28) Temple, M. D. An Auditory Display Tool for DNA Sequence Analysis. *BMC Bioinformatics* **2017**, *18* (1), 221. <https://doi.org/10.1186/s12859-017-1632-x>.
- (29) Taylor, S. From Program Music to Sonification: Representation and the Evolution of Music and Language. In *Proceedings of the 23rd International Conference on Auditory Display - ICAD 2017*; The International Community for Auditory Display: University Park Campus, 2017; pp 57–63. <https://doi.org/10.21785/icad2017.060>.
- (30) Doak, D. G.; Denyer, G. S.; Gerrard, J. A.; Mackay, J. P.; Allison, J. R. Peppy: A Virtual Reality Environment for Exploring the Principles of Polypeptide Structure. *Protein Sci.* **2020**, *29* (1), 157–168. <https://doi.org/10.1002/pro.3752>.

- (31) Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2. ed., 5. printing.; Freeman: New York, 1997.
- (32) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. A View of the Hydrophobic Effect. *J. Phys. Chem. B* **2002**, *106* (3), 521–533. <https://doi.org/10.1021/jp015514e>.
- (33) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels. *Nat. Struct. Biol.* **1997**, *4* (1), 10–19. <https://doi.org/10.1038/nsb0197-10>.
- (34) Frauenfelder, H.; Sligar, S.; Wolynes, P. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254* (5038), 1598–1603. <https://doi.org/10.1126/science.1749933>.
- (35) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48* (1), 545–600. <https://doi.org/10.1146/annurev.physchem.48.1.545>.
- (36) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. Toward an Outline of the Topography of a Realistic Protein-Folding Funnel. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (8), 3626–3630. <https://doi.org/10.1073/pnas.92.8.3626>.
- (37) Lau, K. F.; Dill, K. A. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules* **1989**, *22* (10), 3986–3997. <https://doi.org/10.1021/ma00200a030>.
- (38) Šali, A.; Shakhnovich, E.; Karplus, M. Kinetics of Protein Folding. *J. Mol. Biol.* **1994**, *235* (5), 1614–1636. <https://doi.org/10.1006/jmbi.1994.1110>.
- (39) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein Folding Funnels: A Kinetic Approach to the Sequence-Structure Relationship. *Proc. Natl. Acad. Sci.* **1992**, *89* (18), 8721–8725. <https://doi.org/10.1073/pnas.89.18.8721>.
- (40) Beddard, G. S. Using the Metropolis Algorithm To Calculate Thermodynamic Quantities: An Undergraduate Computational Experiment. *J. Chem. Educ.* **2011**, *88* (5), 574–580. <https://doi.org/10.1021/ed100414p>.
- (41) Bieri, O.; Wirz, J.; Hellrung, B.; Schutkowski, M.; Drewello, M.; Kiefhaber, T. The Speed Limit for Protein Folding Measured by Triplet-Triplet Energy Transfer. *Proc. Natl. Acad. Sci.* **1999**, *96* (17), 9597–9601. <https://doi.org/10.1073/pnas.96.17.9597>.
- (42) Hebel, K.; Scaletti, C. The Software Architecture of the Kyma System. International Computer Music Association 1993.
- (43) van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; Daggett, V. Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure* **2010**, *18* (4), 423–435. <https://doi.org/10.1016/j.str.2010.01.012>.
- (44) Chen, H. I.; Sudol, M. The WW Domain of Yes-Associated Protein Binds a Proline-Rich Ligand That Differs from the Consensus Established for Src Homology 3-Binding Modules. *Proc. Natl. Acad. Sci.* **1995**, *92* (17), 7819–7823. <https://doi.org/10.1073/pnas.92.17.7819>.
- (45) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcc.0c02536>.
- (46) Baltrusaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41* (2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>.

- (47) Knoop-van Campen, C. A. N.; Segers, E.; Verhoeven, L. Effects of Audio Support on Multimedia Learning Processes and Outcomes in Students with Dyslexia. *Comput. Educ.* **2020**, *150*, 103858. <https://doi.org/10.1016/j.compedu.2020.103858>.
- (48) Lang, A. Audio-Video-Redundancy in Learning. In *Encyclopedia of the Sciences of Learning*; Seel, N. M., Ed.; Springer US: Boston, MA, 2012; pp 382–384. https://doi.org/10.1007/978-1-4419-1428-6_243.
- (49) Scaletti, C.; Craig, A. B. Using Sound to Extract Meaning from Complex Data; Farrell, E. J., Ed.; San Jose, CA, 1991; pp 207–219. <https://doi.org/10.1117/12.44397>.
- (50) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517–520. <https://doi.org/10.1126/science.1208351>.
- (51) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. Computational Design and Experimental Testing of the Fastest-Folding β -Sheet Protein. *J. Mol. Biol.* **2011**, *405* (1), 43–48. <https://doi.org/10.1016/j.jmb.2010.10.023>.
- (52) Bryngelson, J. D.; Onuchic, J. N. J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins Struct. Funct. Bioinforma.* **1995**, *21* (3), 167–195. <https://doi.org/10.1002/prot.340210302>.
- (53) Best, R. B.; Hummer, G. Reaction Coordinates and Rates from Transition Paths. *Proc. Natl. Acad. Sci.* **2005**, *102* (19), 6732–6737. <https://doi.org/10.1073/pnas.0408098102>.
- (54) Peters, B.; Beckham, G. T.; Trout, B. L. Extensions to the Likelihood Maximization Approach for Finding Reaction Coordinates. *J. Chem. Phys.* **2007**, *127* (3), 034109. <https://doi.org/10.1063/1.2748396>.
- (55) Davis, C. M.; Zanetti-Polzi, L.; Gruebele, M.; Amadei, A.; Dyer, R. B.; Daidone, I. A Quantitative Connection of Experimental and Simulated Folding Landscapes by Vibrational Spectroscopy. *Chem. Sci.* **2018**, *9* (48), 9002–9011. <https://doi.org/10.1039/C8SC03786H>.
- (56) Zanetti-Polzi, L.; Davis, C. M.; Gruebele, M.; Dyer, R. B.; Amadei, A.; Daidone, I. Parallel Folding Pathways of Fip35 WW Domain Explained by Infrared Spectra and Their Computer Simulation. *FEBS Lett.* **2017**, *591* (20), 3265–3275. <https://doi.org/10.1002/1873-3468.12836>.
- (57) Matsunaga, Y.; Sugita, Y. Linking Time-Series of Single-Molecule Experiments with Molecular Dynamics Simulations by Machine Learning. *eLife* **2018**, *7*, e32668. <https://doi.org/10.7554/eLife.32668>.
- (58) Gopi, S.; Singh, A.; Suresh, S.; Paul, S.; Ranu, S.; N. Naganathan, A. Toward a Quantitative Description of Microscopic Pathway Heterogeneity in Protein Folding. *Phys. Chem. Chem. Phys.* **2017**, *19* (31), 20891–20903. <https://doi.org/10.1039/C7CP03011H>.
- (59) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133* (45), 18413–18419. <https://doi.org/10.1021/ja207470h>.
- (60) Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. A Statistical Mechanical Model for Beta-Hairpin Kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (11), 5872–5879. <https://doi.org/10.1073/pnas.95.11.5872>.

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CH. 2: METHODS

Reproduced in part with permission from:

- Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *The Journal of Physical Chemistry Letters* **2019**, *10* (18), 5667–5673.¹

Macromolecules in *E. coli* Models

For more information, please see Section 2.2 (Macromolecular Components of Models)

Table A.1: Macromolecular components of all three *E. coli* models. Check marks indicate that a macromolecule was included in the model; multiple check marks indicate multiple copies were included.

Protein Name	C22*	C36m	C36mCU	Protein Name	C22*	C36m	C36mCU
ATCase	✓	✓	✓	EFTu	✓	✓	✓
MD	✓	✓	✓	ADSS	✓	✓	✓
PPA	✓	✓	✓	GAPDH	✓	✓	✓
PSAT	✓	✓	✓	Phe-tRNA	✓	✓	✓
Mn SOD	✓	✓	✓	DNA binding domain	✓	✗	✗
CYC A	✓	✓	✓	GTT (unfolded)	✓✓✓	✓✓	✓✓
CSP	✓✓✓	✓✓✓	✓✓✓	GTT (folded)	✗	✓	✓

Metabolite Concentrations in *E. coli* Models

For more information, please see Section 2.4 (Addition of Metabolites, Inorganic Ions, and Water)

Table A.2: Metabolites in EC22*, EC36m, and EC36mCU models compared to experimental concentrations.² Bolded lines contain metabolites whose parameters are already available in the CHARMM force field. All metabolites that are not bolded were parameterized by analogy using CGenFF^{3,4}. This table only shows metabolites selected to be used in the *H. sapiens* models. For the full list of metabolite concentrations used for randomization, please see file A.1 “Bacterial”.

Name	Exp. Conc. (mM)	Number of copies in model		Conc. in model (mM)		Residue Name
		EC22*	EC36m/EC36mCU	EC22*	EC36m/EC36mCU	
Glutamate	96	113	112	96.11	95.26	GLU
Glutathione	17	20	20	17.01	17.01	GSH
Fructose-1,6-bisphosphate	15	18	18	15.31	15.31	F6P

Table A.2 (cont.) Metabolite Concentrations in *E. coli* Models

ATP	9.6	11	11	9.36	9.36	ATP
UDP-N-acetylglucosamine	9.2	11	9	9.36	7.65	UD1
Hexose-Pa	8.8	10	8	8.5	6.8	BG6
UTP	8.3	10	12	8.5	10.21	UTP
GTP	4.9	6	4	5.1	3.4	GTP
dTTP	4.6	5	6	4.25	5.1	TTP
Aspartate	4.2	5	4	4.25	3.4	ASP
Valine	4	5	6	4.25	5.1	VAL
6-Phosphogluconate	3.8	4	3	3.4	2.55	6PG
Glutamine	3.8	4	4	3.4	3.4	GLN
CTP	2.7	3	3	2.55	2.55	CTP1
NAD ⁺	2.6	3	3	2.55	2.55	NAD
Alanine	2.6	3	3	2.55	2.55	ALA
UDP-glucose	2.5	3	3	2.55	2.55	UPG
Glutathione disulfide	2.4	3	6	2.55	5.1	GDS
Uridine	2.1	2	4	1.7	3.4	URD
Citrate	2	2	2	1.7	1.7	FLC
UDP	1.8	2	0	1.7	0	UDP
Malate	1.7	2	1	1.7	0.85	MLT
3-Phosphoglycerateb	1.5	2	2	1.7	1.7	PG3
Glycerate	1.4	2	0	1.7	0	3PG
Coenzyme A	1.4	2	1	1.7	0.85	COA
Citrulline	1.4	2	1	1.7	0.85	CIR
Pentose-Pc	1.3	2	0	1.7	0	A5P
Glucosamine-6-phosphate	1.2	1	1	0.85	0.85	GLP
Acetylphosphate	1.1	1	0	0.85	0	UVW
Gluconolactone	1	1	3	0.85	2.55	LGC
GDP	0.68	1	0	0.85	0	GDP
Acetyl-CoA	0.61	1	0	0.85	0	ACA
Arginine	0.57	1	0	0.85	0	ARG
Succinate	0.57	1	1	0.85	0.85	SCC
UDP-glucuronate	0.57	0	3	0	2.55	UGL
Lysine	0.41	0	1	0	0.85	LYS
Dihydroxyacetone phosphate	0.37	0	1	0	0.85	ACP
Homocysteine	0.37	0	1	0	0.85	HCY
Leucine	0.3	0	1	0	0.85	LEU

Table A.2 (cont.) Metabolite Concentrations in *E. coli* Models

AMP	0.28	0	1	0	0.85	AMP
Inosine monophosphate	0.27	0	1	0	0.85	I1P
FAD	0.17	0	1	0	0.85	FAD
Methionine	0.15	0	1	0	0.85	MET

Metabolite Concentrations in *H. sapiens* Models

For more information, please see Section 2.4 (Addition of Metabolites, Inorganic Ions, and Water)

Table A.3: Metabolites in *H. sapiens* models compared to experimental concentrations.^{5,6} Bolded lines contain metabolites whose parameters are already available in the CHARMM force field. All metabolites that are not bolded were parameterized by analogy using CGenFF^{3,4}. This table only shows metabolites selected to be used in the *H. sapiens* models. For the full list of metabolite concentrations used for randomization, please see Supplementary file A.1 “Mammalian”.

Metabolite Name	Experimental Concentration (mM)		Consensus Concentration (mM)	No. of Copies Expected in Model	No. of Copies in Model	Concentration in Model (mM)
	<i>Chen et al.</i>	<i>Diener et al.</i>				
Glutamate	24.628	39.138	31.883	42.2	41	31.000
Taurine	18.403	0.000	16.901	22.4	25	18.902
Aspartate	6.219	25.248	15.733	20.8	21	15.878
Glutamine	1.369	30.933	11.006	14.6	13	9.829
Glycine	6.725	13.500	10.112	13.4	13	9.829
GSH	5.551	10.865	8.208	10.9	6	4.537
Spermine	-	7.240	7.240	9.6	5	3.780
Lactate	1.874	11.310	5.461	7.2	5	3.780
Asparagine	0.145	7.290	3.717	4.9	2	1.512
Malate	0.679	5.273	3.651	4.8	10	7.561
Alanine	3.330	3.704	3.517	4.7	6	4.537
Spermidine	-	3.425	3.425	4.5	2	1.512
Beta-Alanine	6.481	0.329	3.405	4.5	4	3.024
Citrate/Isocitrate	1.329	4.904	3.117	4.1	7	5.293
Hydroxyproline	-	2.707	2.707	3.6	5	3.780
ATP	4.641	0.311	2.476	3.3	2	1.512
Putrescine	-	2.084	2.084	2.8	1	0.756
Phosphocholine	3.816	0.000	1.908	2.5	3	2.268
UTP	1.840	0.000	1.840	2.4	4	3.024
Isoleucine	1.401	3.255	1.709	2.3	1	0.756
NAD	0.489	2.712	1.601	2.1	2	1.512
Serine	0.471	4.031	1.588	2.1	2	1.512
Leucine	1.011	3.257	1.572	2.1	1	0.756
Threonine	1.005	2.189	1.278	1.7	2	1.512

Table A.3 (cont.) Metabolite Concentrations in *H. sapiens* Models

Arginine	0.186	3.270	1.234	1.6	4	3.024
Histidine	1.122	1.073	1.098	1.5	1	0.756
Valine	1.348	1.397	1.096	1.5	1	0.756
GSSG	0.064	2.048	1.056	1.4	1	0.756
UDP	0.444	1.264	0.854	1.1	1	0.756
Tyrosine	0.678	1.497	0.838	1.1	1	0.756
3-Hydroxybutyric acid	-	0.059	0.680	0.9	2	1.512
G3P	0.000	1.324	0.662	0.9	1	0.756
Fumarate	0.070	1.175	0.622	0.8	1	0.756
GTP	0.600	0.000	0.600	0.8	1	0.756
ADP	0.140	0.834	0.487	0.6	2	1.512
AMP	0.023	0.898	0.460	0.6	1	0.756
FBP	0.807	0.064	0.436	0.6	1	0.756
Succinate	0.042	0.774	0.408	0.5	1	0.756
Pyruvate	0.281	0.000	0.321	0.4	1	0.756
R5P/RL5P	0.114	0.374	0.244	0.3	1	0.756
Pantothenic Acid	0.235	0.000	0.235	0.3	1	0.756
G6P	0.146	0.322	0.234	0.3	1	0.756
Citrulline	0.009	0.634	0.217	0.3	1	0.756
GMP	0.111	0.079	0.095	0.1	1	0.756

Code to randomly assign metabolites based on known concentrations

Written in Python | For more information, please see Section 2.4 (Addition of Metabolites, Inorganic Ions, and Water)
Reproduced with permission from Evan Curtin

Code reads csv file with metabolite names and concentrations (please see Supp file A.1). Data is summed and normalized such that each metabolite is assigned a probability based on its concentration. For example, if glutamate's given concentration was 100 mM, ATP's was 10 mM and the total of all metabolite concentrations in the file was 500 mM, glutamate would be assigned a probability of $100/500 = .2$, while ATP would be assigned a probability of 0.02. The np.random.multinomial function assigned N_samples of metabolites according to the probability distribution. Then, these results are recombined with the original names of the metabolites and written to an output file.

I wrote the original pseudocode but received help from Evan Curtin implementing the code in Python. I thank him for his time and am printing this code with his permission.

Code A.1: Code to randomly assign number of metabolites based on a text file with concentration data. Results are written to an output text file (“outfile”).

```
# Written by Meredith Rickard on 2018.1.29
# written with a lot of help from Evan Curtin

import numpy as np
import csv

#hardcoded file path because it's consistent and this will rarely be used
with open('C:/Users/Meredith/Documents/temp/human_metab_data.txt', 'r') as f:
    lines = csv.reader(f)
#create dictionary of the data
    d = {l[0]: l[1] for l in lines}

vals = np.fromiter(d.values(), dtype=np.float64)
normalized = vals / np.sum(vals)
#normalized is the probability of each one

#N_samples is calculated based on the expected volume of the model & total
metabolite concentration
N_samples = 300
samples = np.random.multinomial(N_samples, normalized)

dump = {k: v for k, v in zip(d.keys(), samples)}
with open('C:/Users/Meredith/Documents/temp/hela_met_outfile.csv', 'w') as f:
    lines = ['{}{}'.format(k, v) for k, v in dump.items()]
    s = '\n'.join(lines)
    f.write(s)
```

Supplementary File A.1

The supplementary file “Rickard_disseration_supplementary_table.xlsx” contains excel spreadsheets of longer tables not suited for this appendix. The contents are listed below:

E. coli Metabolite Data | Sheet Name: Bacterial

Due to length concerns, the concentrations of all *E. coli* metabolites used for randomization in Section 2.4 are not listed in Table A.2. Supplementary file A.1 contains a sheet that contains all concentrations of *E. coli* metabolites, including those that were not selected to be placed in either *E. coli* model.

Human Metabolite Data | Sheet Name: Human

Due to length concerns, the concentrations of all *H. sapiens* metabolites used for randomization in section 2.4 are not listed in Table A.3. Supplementary file A.1 contains a sheet

that contains all concentrations of *H. sapiens* metabolites, including those that were not selected to be placed in the *H. sapiens* cytoplasm model.

Code to Restrain C α atoms together for equilibration procedures

Written in TCL for use in VMD | For more information, please see Section 2.5 (Preequilibration)

Two top level functions that create extra bonds with $k = 2$ between CA atoms (if protein) or P atoms (if nucleic) that are less than a certain distance apart. It can do this to only its own backbone (extraBonds_self), locking it into its fold, or to another segment (extraBonds_cross), tying two macromolecules together. To do this, it takes each CA and checks every other CA in the target segment. To avoid double counting, in extraBonds_self, only the residue number higher than the one we're checking against is checked. For example, for a 100 residue protein, it finds the distance between CA1 and every CA between 2 and 100. In the next iteration, we're trying to find things that are within 10 of CA2, but we've already checked the CA1-CA2 distance, so we check all distances between 3 and 100. And so on.

getDistance is a function that takes two single-atom atomselections and uses the distance formula to get the distance between the two. getInfo is a function that takes a segment name, determines the number of residues (min and max), and figures out whether to use P or CA as the atom name (by checking if it's nucleic).

Code A.2: Procedure to restrain C α atoms to each other during equilibration. This allows proteins to translate throughout the model, but does not allow them to unfold (or, in the case of extraBonds_cross, can constrain non-covalently bound multimers to each other).

```
proc extraBonds_self {seg} {
    set thresh 10
    set outfile [open "./extrabonds_$seg.xvg" w]
    puts "opened file"
    puts $seg
    set xtra [lassign [getInfo $seg] min_res max_res name]
    puts $min_res
    puts $max_res
    puts $name

    for {set c $min_res} {$c <= $max_res } {incr c} {
        set di [expr {$c+1}]
        set csel [atomselect top "segname $seg and resid $c and name $name"]
        set checkc [string length [$csel get segname]] ;
        # Here we just need to check that our selection exists. So if the
        segname returns an empty (length 0) string, we know that that's not a valid
        selection
    }
}
```

Code A.2 (cont.) Code to restrain C α for equilibration procedures

```

    if {$checkc > 0} {
        for {set d $di} {$d <= $max_res} {incr d} {
            set dsel [atomselect top "segname $seg and resid $d and name
$name"]
            set checkd [string length [$dsel get segname]]
            if {$checkd > 0} {
                set dist [getDistance $csel $dsel]
                if {$dist < $thresh} then {
                    set i1 [$csel get index]
                    set i2 [$dsel get index]
                    puts $outfile "bond $i1 $i2 2 $dist"
                }
            }
            $dsel delete
        }
        $csel delete
    }

    puts "closing"
    close $outfile
}

proc extraBonds_cross {segname1 segname2} {

    set thresh 10
    set outfile [open "./extrabonds_$segname1$segname2.xvg" w]

    set xtra [lassign [getInfo $segname1] min_res1 max_res1 name1]
    set xtra [lassign [getInfo $segname2] min_res2 max_res2 name2]
    puts "$segname1 info"; puts $min_res1; puts $max_res1; puts $name1
    puts "$segname2 info"; puts $min_res2; puts $max_res2; puts $name2

    for {set c $min_res1} {$c <= $max_res1} {incr c} {
        set csel [atomselect top "segname $segname1 and resid $c and name
$name1"]
        set checkc [string length [$csel get segname]]

        if {$checkc > 0} {
            for {set d $min_res2} {$d <= $max_res2} {incr d} {
                set dsel [atomselect top "segname $segname2 and resid $d and name
$name2"]
                set checkd [string length [$dsel get segname]]
                if {$checkd > 0} {
                    puts "d is $d"
                    set dist [getDistance $csel $dsel]
                    if {$dist < $thresh} then {
                        set i1 [$csel get index]
                        set i2 [$dsel get index]
                        puts $outfile "bond $i1 $i2 2 $dist"
                    }
                }
            }
            $dsel delete
        }
    }
}

```

Code A.2 (cont.) Code to restrain C α for equilibration procedures

```
    }
  }
  $csel delete
}
puts "closing"
close $outfile
}

proc getDistance {sel1 sel2} {
  set coordx1 [$sel1 get {x}]
  set coordy1 [$sel1 get {y}]
  set coordz1 [$sel1 get {z}]
  set coordx2 [$sel2 get {x}]
  set coordy2 [$sel2 get {y}]
  set coordz2 [$sel2 get {z}]

  set x1 [lindex $coordx1 0]
  set y1 [lindex $coordy1 0]
  set z1 [lindex $coordz1 0]

  set x2 [lindex $coordx2 0]
  set y2 [lindex $coordy2 0]
  set z2 [lindex $coordz2 0]
  set distance [expr {sqrt(($x2-$x1)*($x2-$x1) + ($y2-$y1)*($y2-$y1) +
($z2-$z1)*($z2-$z1))}]
  return $distance
}

proc getInfo {seg} {
  set sel [atomselect top "segname $seg"]
  set residue [lsort -unique -integer [$sel get resid]]
  set min_res [lindex $residue 0]
  set max_res [lindex $residue end]

  if {[lindex [$sel get nucleic] 0]} {
    set name "P"
  } else {
    set name "CA"
  }
  $sel delete
  return [ list $min_res $max_res $name]
}
```

Sample Anton 2 input file

Modified from Anton 2 “base.ark” basic ark file.

Below is a sample .ark file used to simulate the *H. sapiens* model on Anton 2. Overall, the input file is largely derived from the basic ark file “base.ark” provided by the Pittsburgh Supercomputing Center and D.E. Shaw Research group. However, I have altered the temperature to 318 K. Additionally, I have manually set the `anton.chem.chooser.chem.n_k` value (line 9) to [“32” “64” “64”]. This determines the number of Fourier mesh points along each axis, used to efficiently calculate electrostatic interactions. Normally, the software will automatically select an appropriate `n_k` value. This manual modification was done at the advice of Pittsburgh Supercomputing Center in response to repeated 10114 (PPIP internal underflow) errors on cytoplasm simulations. This should not be altered as a rule in cytoplasm simulations, as the cytoplasm simulations do not always need this modification, and manually altering the `n_k` value decreases Anton 2 performance.⁷

Code A.3: Sample Anton 2 input file for cytoplasm simulations.

```
anton {
  chem {
    angle {
      permitFloppyAngles = "true"
    }
    average_dispersion_type = "auto"
    chooser {
      chem {
        n_k = ["32" "64" "64"]
      }
    }
  }
  tune {
    last_time = "4800000"
    machine_size = ["4" "4" "8"]
    optional {
      ConstraintThreshold = "0.1"
      MaxAllowedNetCharge = "1.2"
    }
    trajectory {
      interval = "240"
    }
  }
}
boot {
  file = "/anton2fs/raw/mmricka2/2020_11_11/mammalian_params.dms"
}
integrator {
```

Code A.3 (cont.) Sample Anton input file

```
Multigrator {
  barostat {
    MTK {
      T_ref = "318.00"
      tau = "0.0416667"
      thermostat {
        NoseHoover {
          chain {
            mts = "4"
            tau = ["0.0416667" "0.0416667" "0.0416667"]
          }
        }
        type = "NoseHoover"
      }
    }
    interval = "480"
    type = "MTK"
  }
  thermostat {
    NoseHoover {
      chains = [{
        mts = "1"
        tau = [".0416667" ".0416667" ".0416667"]
      }]
      use_molecular_ke = "true"
    }
    interval = "24"
    type = "NoseHoover"
  }
}
dt = "0.002"
pressure {
  isotropy = "isotropic"
  p_ref = "1.0"
  tension_ref = ["0." "0." "0." "0." "0." "0." "0." "0." "0."]
}
remove_com_motion = "true"
respa {
  bonded_interval = "1"
  nonbonded_far_interval = "3"
  nonbonded_near_interval = "1"
}
temperature = [{
  T_ref = "318.00"
}]
type = "Multigrator"
}
```

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. V. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *The Journal of Physical Chemistry Letters* **2019**, *10* (18), 5667–5673. <https://doi.org/10.1021/acs.jpclett.9b01556>.
- (2) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia Coli. *Nature Chemical Biology* **2009**, *5*, 593–599. <https://doi.org/10.1038/nchembio.186.Absolute>.
- (3) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *Journal of Chemical Information and Modeling* **2012**, *52*, 3144–3154. <https://doi.org/10.1021/ci300363c>.
- (4) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *Journal of Chemical Information and Modeling* **2012**, *52*, 3155–3168. <https://doi.org/10.1021/ci3003649>.
- (5) Chen, W. W.; Freinkman, E.; Wang, T.; Birsoy, K.; Sabatini, D. M. Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism. *Cell* **2016**, *166* (5), 1324–1337.e11. <https://doi.org/10.1016/j.cell.2016.07.040>.
- (6) Diener, C.; Muñoz-Gonzalez, F.; Encarnación, S.; Resendis-Antonio, O. The Space of Enzyme Regulation in HeLa Cells Can Be Inferred from Its Intracellular Metabolome. *Sci Rep* **2016**, *6* (1), 28415. <https://doi.org/10.1038/srep28415>.
- (7) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. B.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE: New Orleans, LA, USA, 2014; pp 41–53. <https://doi.org/10.1109/SC.2014.9>.

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CH. 4: FOLDING OF GTT IN *E. COLI* MODELS

Reproduced in part with permission from:

- Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, 124 (23), 4732–4740.¹

Secondary Structure of GTT in EC22 Model*

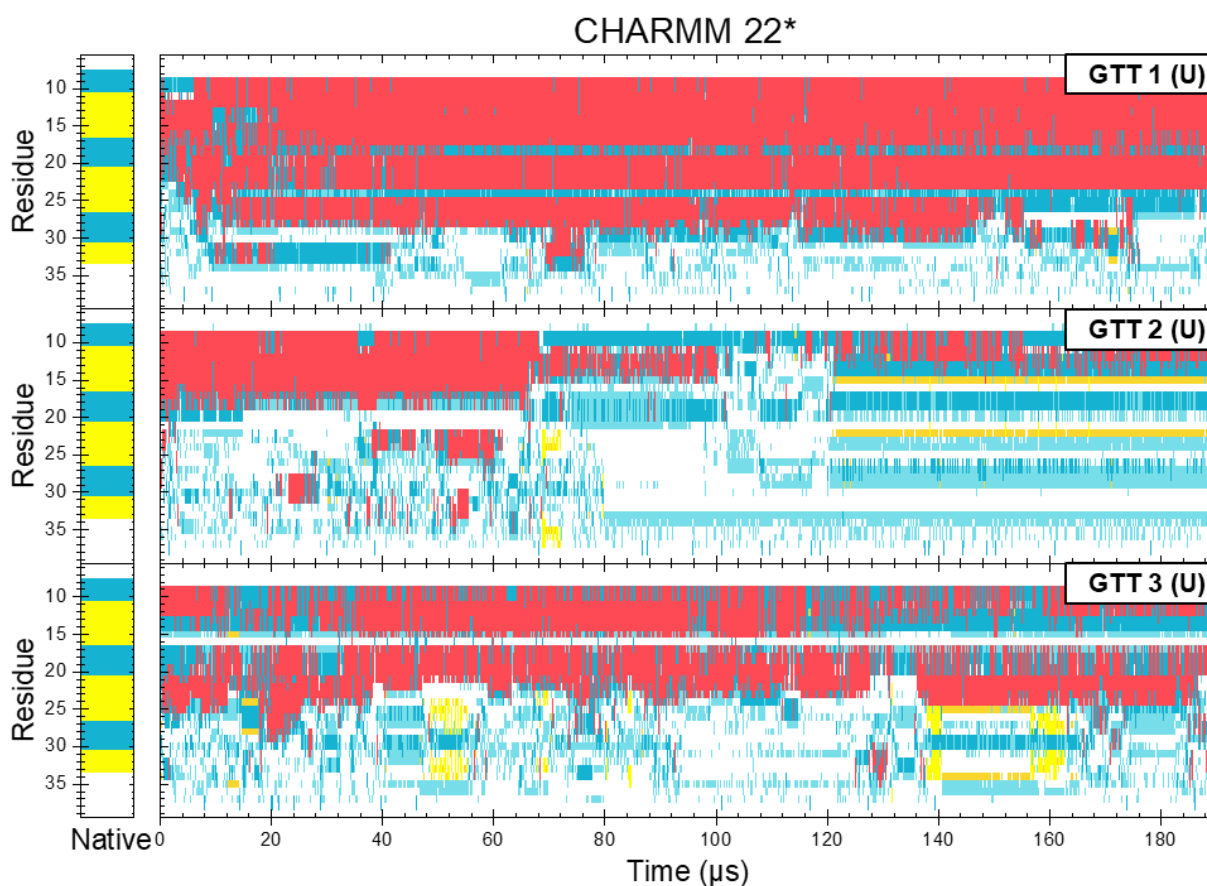


Figure B.1: Simulation of three initially unfolded GTT WW domains in the EC22*-TIP3P model which was run for the longest time (>190 μ s). Secondary structures are coded as yellow (sheet), red (helix) dark blue (turn) and light blue (bend). Figure published in ref. 1

Secondary Structure of GTT in EC36m Model

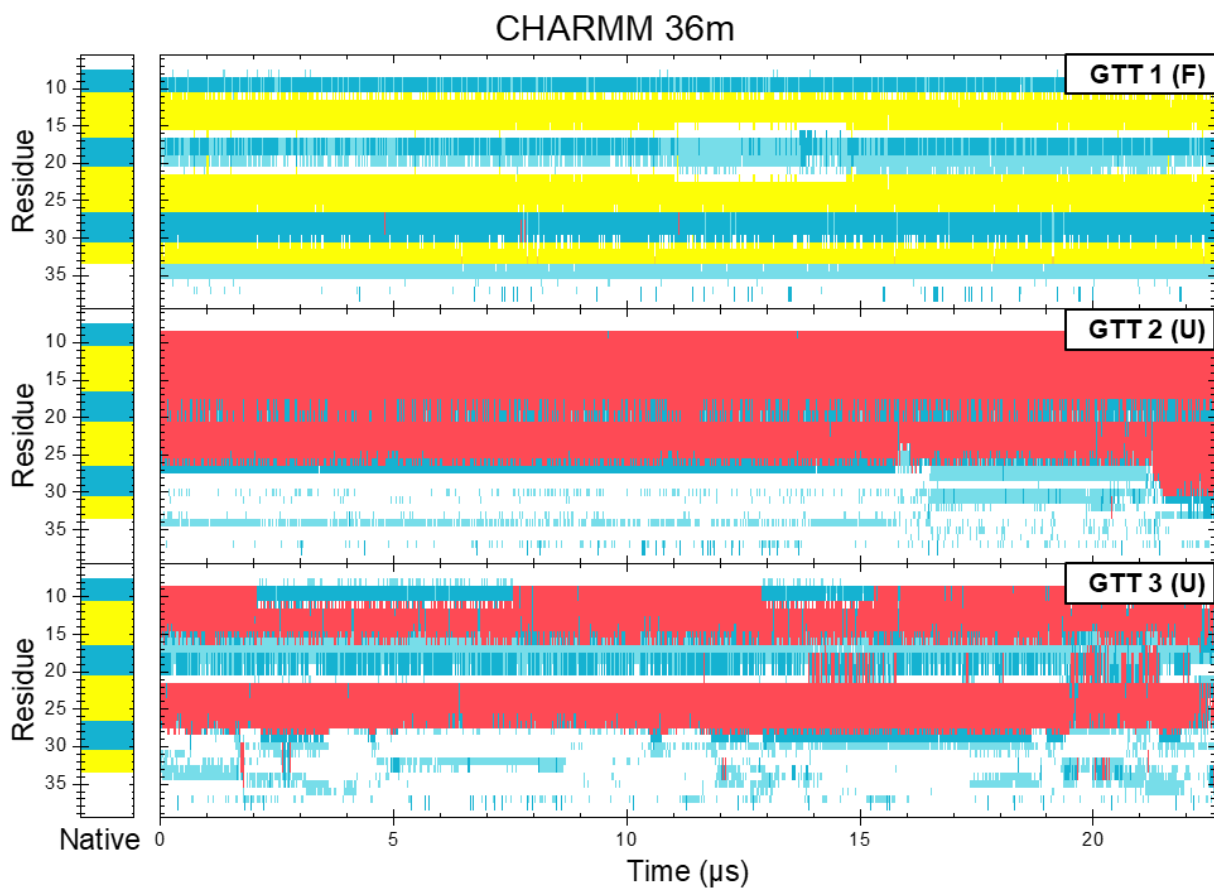


Figure B.2: Simulation of one initially folded (top) and two initially unfolded GTT WW domains in the EC36m-TIP3P model. Secondary structures are coded as yellow (sheet), red (helix) dark blue (turn) and light blue (bend). Figure published in ref. 1

Secondary Structure of GTT in EC36mCU Model

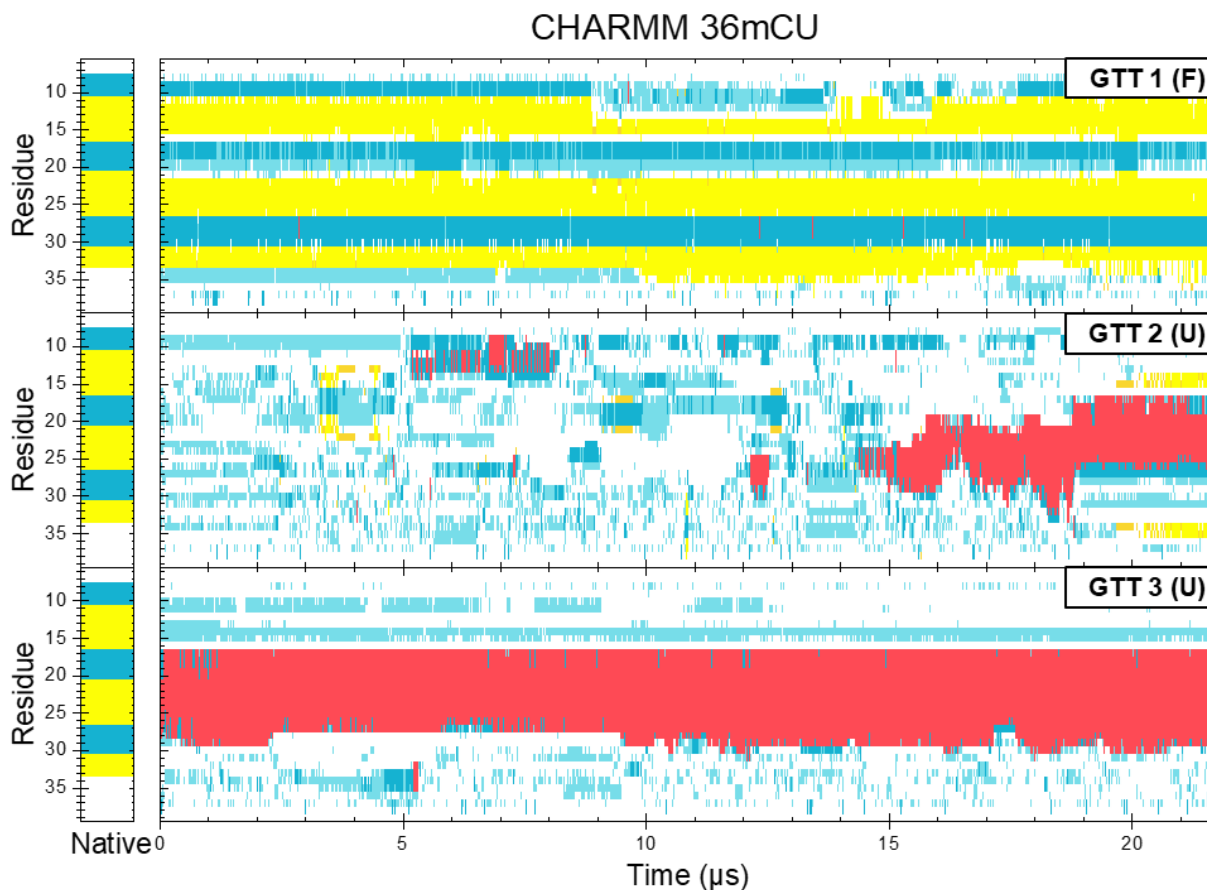


Figure B.3: Simulation of one initially folded (top) and two initially unfolded GTT WW domains in the EC36mCU-TIP3P model. Secondary structures are coded as yellow (sheet), red (helix) dark blue (turn) and light blue (bend). Figure published in ref. 1

RMSD of Proteins in EC22* Simulation

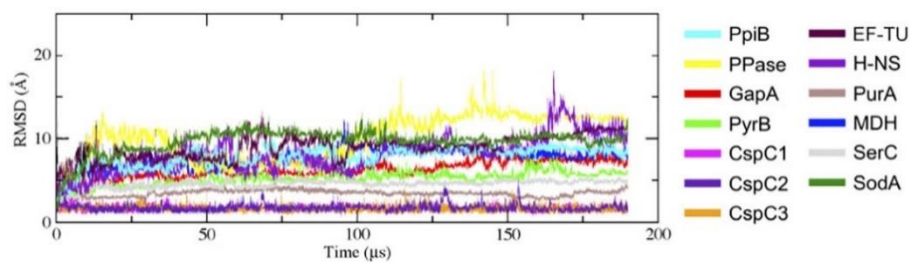


Figure B.4: RMSD values of 13 distinct proteins in the EC22* simulation. Csp1-3 have particularly small RMSD, SerC underwent only minor structural fluctuations, GapA and PpiB remained stable with structural fluctuations in the flexible loop regions, PurA, PyrB and EF-TU remained stable with 1-2 helices partly unfolding, SodA and Mdh partly unfolded by interacting with a copy of GTT, and Hns and a tRNA (not shown, see ref. 1) completely unfolded. In the 36m and 36U simulations, the Mg^{2+} counterion distribution around the tRNA was adjusted and no unfolding was observed over 25 μs , so the RNA unfolding is most likely due to local electrostatic imbalance.

Diffusion of GTT in water-box and EC22* models

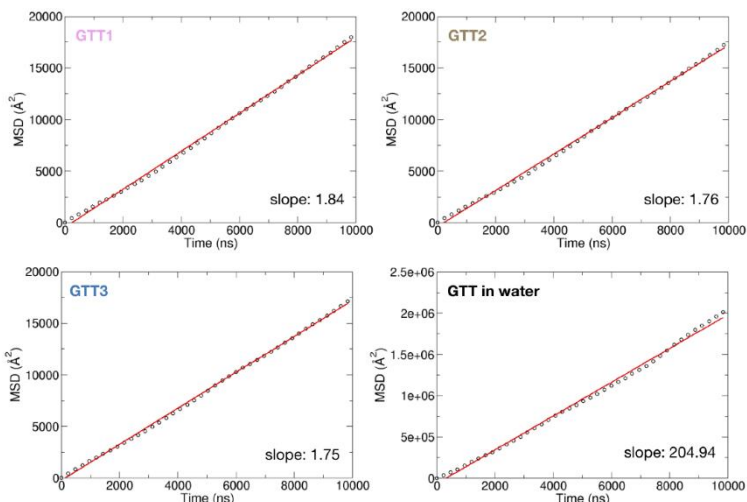


Figure B.5: Diffusion of the three GTT WW domains in the EC22* simulation of bacterial cytoplasm compared to the same force field in buffer.

Solvation of GTT in EC36m model

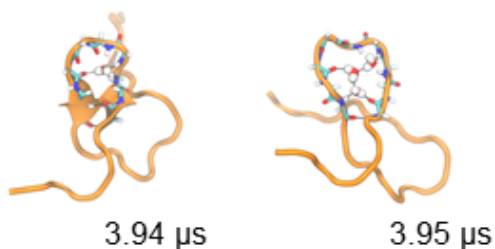


Figure B.6: Example of water bridging transition state preceding solvation of the turn and loss of a β hairpin. Snapshots taken from GTT 2 in the EC36mCU trajectory. This corresponds to the reverse of the case shown in Figure 4.7.

Helical Propensity of GTT

Residue Num	0	1	2	3	4	5	6	7	8	9	10	11	12
Amino Acid	K	L	P	P	G	W	E	K	R	M	S	R	D
Prob(helix)	0	0	0	0	0	0.4	0.4	0.5	0.5	0.4	0.2	0.2	0.1
Residue Num	13	14	15	16	17	18	19	20	21	22	23	24	25
Amino Acid	G	R	V	Y	Y	F	N	H	I	T	G	T	T
Prob(helix)	0.1	0.2	0.2	0.2	0.1	0.1	0	0	0	0	0	0	0
Residue Num	26	27	28	29	30	31	32						
Amino Acid	Q	F	E	R	P	S	G						
Prob(helix)	0	0	0	0	0	0	0						

Table B.1: Helical propensities to the nearest tenth of all residues in GTT WW domain, as predicted by Agadir²

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcc.0c02536>.
- (2) Lacroix, E.; Viguera, A. R.; Serrano, L. Elucidating the Folding Problem of Alpha-Helices: Local Motifs, Long-Range Electrostatics, Ionic-Strength Dependence and Prediction of NMR Parameters. *Journal of Molecular Biology* **1998**, *284* (1), 173–191. <https://doi.org/10.1006/jmbi.1998.2145>.

APPENDIX C: SUPPLEMENTARY INFORMATION FOR CH. 5: IN-CELL PROTEIN- PROTEIN CONTACTS

Code to Calculate Buried SASA of protein-protein contacts in EC36mCU model

Written in TCL for use in VMD | For more information, please see Chapter 5

This code loads a dcd trajectory into memory, selects each possible pair of macromolecules in the trajectory, performs necessary wrapping and unwrapping calculations to ensure accurate SASA, and calculates the SASA buried between each pair at each timestep.

Code C.1: Tcl code to calculate the buried-SASA contacts between all proteins in the EC36mCU model

```
# -----  
# Author: Meredith Rickard  
# Pogorelov Group  
# University of Illinois  
# Written 2019.01.31  
# -----  
#  
  
set root "cufix_nowater_stridel00_contactsize_both_halves_0_900"  
#set nf [molinfo top get numframes]  
set first_f 0  
set last_f 900  
set outfile [open $root.dat w]  
set chain "P"  
set probe 1.4  
  
#In the EC36mCU model, individual macromolecules are all in chain P, with  
different segnames P1-P16 (except for P9)  
#In the EC22* model, individual macromolecules are all in chain P, with  
segnames P1-P16, except the tRNA is called "VN1"  
  
for {set i 1} {$i <= 16} {incr i} {  
    if {$i == 9} {  
        continue  
    } else {  
        set segname "$chain$i"  
    }  
    # if {$i == 17} {  
    #     set segname "VN1"  
    # }
```

Code C.1 (cont.) Code to calculate buried SASA of protein-protein contacts

```
#   } else {
#       set segname "VP$i"
#   }
  for {set k [expr $i+1]} {$k <= 16} {incr k} {
    if {$k == 9} {
      continue
    }
    # set temp_segname "VN1"
  } else {
    set temp_segname "$chain$k"
  }
  set contact_list "$segname - $temp_segname"
  puts "$segname - $temp_segname"
  set molid [mol new 09_13_02_str100_nowater.dcd waitfor all]
  mol top $molid
  mol addfile C36M_nowater.psf
  set_radii_from_charmm par top
  set sel1 [atomselect top "segname $segname"]
  set sel2 [atomselect top "segname $temp_segname"]
  set sel3 [atomselect top "segname $segname or segname
$temp_segname"]
  puts "done"
  for {set j $first_f} {$j <= $last_f} {incr j} {
    molinfo top set frame $j
    set sasa [measure sasa $probe $sel1]
    set sasa2 [measure sasa $probe $sel2]
    set sasa3 [measure sasa $probe $sel3]
    set sasa4 [measure sasa $probe $sel3 -restrict $sel1]
    set sasa5 [measure sasa $probe $sel3 -restrict $sel2]
    set contacts [list [expr [expr $sasa + $sasa2] - $sasa3] [expr
$sasa - $sasa4] [expr $sasa2 - $sasa5]]
    pbc wrap -molid top -now -sel "segname $segname or segname
$temp_segname"
    set sasa [measure sasa $probe $sel1]
    set sasa2 [measure sasa $probe $sel2]
    set sasa3 [measure sasa $probe $sel3]
    set sasa4 [measure sasa $probe $sel3 -restrict $sel1]
    set sasa5 [measure sasa $probe $sel3 -restrict $sel2]
    lappend contacts [expr [expr $sasa + $sasa2] - $sasa3] [expr
$sasa - $sasa4] [expr $sasa2 - $sasa5]
    lappend contact_list $contacts
  }
  set data [join $contact_list ", "]
  puts $outfile "$data"
  mol delete top
}
}
close $outfile
```

REFERENCES

- (1) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcc.0c02536>.
- (2) Lacroix, E.; Viguera, A. R.; Serrano, L. Elucidating the Folding Problem of Alpha-Helices: Local Motifs, Long-Range Electrostatics, Ionic-Strength Dependence and Prediction of NMR Parameters. *Journal of Molecular Biology* **1998**, *284* (1), 173–191. <https://doi.org/10.1006/jmbi.1998.2145>.

APPENDIX D: SUPPLEMENTARY INFORMATION FOR CH. 7: DATA SONIFICATION

Reproduced in part with permission from:

- Scaletti, C.; Rickard, M.M.; Hebel, K.; Pogorelov, T.V.; Taylor, S.A.; Gruebele, M. “Enhancing Visualization with Sonification to Explain Protein Folding Concepts.” *Under revision, J. Chem. Educ.* **2021**.¹

VIDEOS: LATTICE MODEL ANIMATIONS WITH DATA-DRIVEN SOUND

- The videos are presented here as links to Youtube for ease of access. Full videos will also be posted as SI with ref. 1.
- For best results, select video quality HD1080 in the YouTube settings, and listen with stereo headphones or speakers. Additionally, you may want to watch the sonified videos ‘full screen’ using the  icon. Control “playback speed” using the  icon to get a sharp video at the speed you want.
- For brevity, the full homework, lecture slides, and detailed descriptions of the sonification process will not be included here, but will be published as SI with ref. 1 (and are available upon request).

List of all Videos

- [Video D.1](#) is a sonification/visualization of the 4-bead lattice model shown in Figure 7.2.
- [Video D.2](#) Unevolved 6-bead chain where 2 and 6 are hydrophobic shows a ‘glassy’ or unevolved energy landscape that cannot fold into a single native state (See Figure 7.4).
- [Video D.3](#) Beta hairpin, a 6-bead chain where 2 and 5 are hydrophobic, shows a 4-level energy landscape with a single, lowest energy state (See Figure 7.5A).
- [Video D.4](#) Alpha helix, an 8-bead chain with hydrophobic elements in positions 2 and 7 shows a smoother energy landscape funnel. (See Figure 7.5B).
- [Video D.5](#) WW domain, a 9-bead chain with hydrophobic residues in positions 2, 5, and 8, shows a 10-level energy landscape with multiple ‘traps’ or local minima.
- [Video D.6](#) WW domain animation and sonification where the radius of gyration (R_g) is mapped to frequency: the more compact the conformation, the lower the frequency. (See Figure 7.6).
- [Playlist D.7](#) is a collection of videos that can be used in conjunction with the sample lecture slides.
- [Playlist D.8](#) contains the videos that our students used in order to answer the homework questions.
- [Video D.9](#) Using sound to track the formation of native vs non-native bonds in the 9-bead WW lattice model

- [Video D.10](#) WW domain model showing a temperature-dependent bias as to which states are more likely to be visited.
- [Video D.11](#) WW domain model's traps and intermediate states.
- [Video D.12](#) Beta hairpin model's traps and intermediate states. (Starts at a low temperature, then jumps up to a high temperature about halfway through the video).
- [Video D.13](#) Sonification of a simple coin-toss that is either fair, biased toward Heads, or biased toward Tails.
- [Video D.14](#) Coin-toss with an unknown bias
- [Video D.15](#) Beta hairpin model running at a high temperature (and using a sound-mapping similar to the one used in coin toss to facilitate comparisons)
- [Video D.16](#) Beta hairpin model running at a low temperature
- [Video D.17](#) Beta hairpin model at the 'folding temperature', T_m — the temperature at which the likelihood of being in the folded state at the bottom of the funnel equals the likelihood of being in an unfolded state (the ensemble of unfolded structures).
- [Video D.18](#) Beta hairpin model at an unknown temperature. Is the temperature above or below T_m ?
- [Video D.19](#) A visualization/sonification of R_g vs. time (vertical axis units are Å, horizontal axis units are μ s.) of the water-box simulation of GTT referred to in Chapter 4.² WW structures are shown on the right. The pitch you hear corresponds to R_g (higher pitch = higher R_g). R_g is also mapped to pan position. The standard deviation, σ_{R_g} , is mapped to the timbre of the sound and to the color of the trace on the left. VMD visualization of the trajectory is shown on the right.
- [Video D.20](#) A sonification of the mean of R_g vs the standard deviation of R_g (plotted in polar coordinates). Pan position is mapped to radius in polar coordinates. The speed of the sonification slows down according to the rarity function, allowing the listener to pay closer attention to transitions.
- [Playlist D.21](#) The same sound mapping applied to several different lattice models

ADDITIONAL SI METHODS

The construction and sonification of the protein lattice were done by C. Scaletti and K. Hebel. Reproduced with permission. A more detailed description of the sonification process is included in the SI of ref. 1.

Representation of proteins on a lattice

Conformations that can interconvert via 2D rotation are not double-counted, as shown by the crossed-out configuration in Figure 7.2. In a more stringent version of the model, configurations accessible via 3D rotations can also be excluded, (this would eliminate one of the L-shaped conformations in Figure 7.2), and the bead at one end of the chain (e.g., the ‘N-terminus’ in a real protein) can be distinguished from the other end of the chain (see discussion of Figure 7.5).

This simple model, with only two amino acid types, is suitable only for small proteins: for large proteins, there is a high probability of contacts not being consistent with a single minimum energy state. At least 6 different amino acids (beads) are needed to define real protein structures uniquely,³ but we stick with 2 here for simplicity.

Definitions of observables

Radius of gyration R_g is defined as the square root of the sum of the distances squared between each bead and the center of the shape; the ‘mass’ of each bead is assumed to be the same.

The fraction of native contacts Q is defined as the ratio of the number of native contacts in a conformation divided by the number of native contacts in the native state. In the lattice model, a ‘native contact’ is when two dark beads (hydrophobic amino acids) or two light beads (hydrophilic amino acids) are adjacent to one another as they should be in the native (lowest energy) structure.

REFERENCES

- (1) Scaletti, C.; Rickard, M. M.; Hebel, K.; Pogorelov, T. V.; Taylor, S.; Gruebele, Martin. Enhancing Visualization with Sonification to Explain Protein Folding Concepts. *J. Chem. Educ* **2021**, *Under revision*.
- (2) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740. <https://doi.org/10.1021/acs.jpcb.0c02536>.
- (3) Nerattini, F.; Tubiana, L.; Cardelli, C.; Bianco, V.; Dellago, C.; Coluzza, I. Protein Design under Competing Conditions for the Availability of Amino Acids. *Sci Rep* **2020**, *10* (1), 2684. <https://doi.org/10.1038/s41598-020-59401-9>.