UNIVERSITY OF CALIFORNIA,
IRVINE

Development and application of molecular modeling methods for characterization of drug targets in *Mycobacterium tuberculosis*

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Pharmaceutical Sciences

by

Kalistyn H. Burley

Dissertation Committee:
Professor David L. Mobley, Chair
Professor Celia W. Goulding
Professor Thomas L. Poulos

2020

DEDICATION

To my children, Aubrey and Baron – being your mother is, and will always be, my greatest accomplishment.

To Scott, an amazing partner in life and in parenthood, who has never doubted me, and has also never let me settle.

To my mom, dad, and brothers Kristopher and Jacob, who gave me the confidence to "jump off cliffs and build my wings on the way down." - Ray Bradbury, 1986

And to my mentors and role models, particularly the women before me who have paved the road for my successes, I am forever indebted and committed to paying it forward.

TABLE OF CONTENTS

LIST OF FIGURES iv

LIST OF TABLES vi

ACKNOWLEDGMENTS vii

VITA ix

ABSTRACT OF THE DISSERTATION xii

CHAPTER 1: **An introduction to tuberculosis and computational methods for structure-based drug design** 1
References 16

CHAPTER 2: **Enhancing side chain sampling using non-equilibrium candidate Monte Carlo** 19
References 59

CHAPTER 3: **Structure of a** *Mycobacterium tuberculosis* **heme-degrading protein, MhuD, variant in complex with its product** 63
References 92

CHAPTER 4: **Insights from the structure and conformational dynamics of** *Mycobacterium tuberculosis* **malic enzyme, MEZ** 97
References 134

CHAPTER 5**: Summary and Conclusions - Complementing structural information using computational methods for drug-design** 139
References 146

APPENDIX A: Chapter 2 Supporting Information 148

APPENDIX B: Chapter 3 Supporting Information 154

APPENDIX C: Chapter 4 Supporting Information 157

LIST OF FIGURES

		Page
Figure 1.1	Percentage of new TB cases that are MDR or rifampicin resistant (RR-TB)	2
Figure 1.2	*Mtb* cell wall architecture	5
Figure 1.3	*Mtb* iron uptake	7
Figure 1.4	Computational methods for drug development	12
Figure 1.5	Approximate timescales of molecular motions	14
Figure 2.1	Cycling of atom interactions during execution of NCMC side chain move	26
Figure 2.2	Move biasing based on known side chain rotamer states	30
Figure 2.3	Workflow illustration of BLUES side chain proposals	32
Figure 2.4	Rotameric states of valine side chain	35
Figure 2.5	Umbrella sampling of valine-alanine peptide in explicit solvent	38
Figure 2.6	Valine-Alanine rotamer transition data where the torsional force constant, k = 3.8 kcal/mol	40
Figure 2.7	Valine-Alanine rotamer transition data where k = 10 kcal/mol.	42
Figure 2.8	Val111 χ_1 rotamer data for BLUES simulations of p-xylene bound T4 lysozyme L99A in explicit solvent	47
Figure 2.9	Val111 χ_1 transition rates in T4 lysozyme L99A bound to p-xylene for BLUES and MD	47
Figure 2.10	Backbone RMSDs for T4 lysozyme L99A simulations	50
Figure 2.11	Overlay of backbone RMSDs of T4 lysozyme L99A for BLUES and MD (first 100ns)	51
Figure 2.12	Val111 χ_1 state transitions per 10^6 FEs from simulations from 3 distinct starting conformations of p-xylene bound T4 lysozyme L99A	53
Figure 3.1	Structures of tetrapyrroles and WT-MhuD-monoheme	65
Figure 3.2	Heme and αBV affinity of MhuD and the R26S variant	73
Figure 3.3	Structure of the MhuD-R26S-αBV complex	76
Figure 3.4	Structural comparison of MhuD-heme-CN and MhuD-R26S-αBV	81
Figure 3.5	Helical stability in MD simulations of MhuD	82
Figure 3.6	Orientation of His75 in MD simulations of MhuD	83
Figure 3.7	Position of Arg79 side chain during MD simulations of MhuD	85
Figure 4.1	Malic enzyme structural classes and oligomeric states	100
Figure 4.2	Structural domains of MEZ	103
Figure 4.3	Biophysical and biochemical analyses of MEZ	109
Figure 4.4	Different modes of $NAD(P)^+$ binding after docking with Mn^{2+} and malate	112

Figure 4.5 Divalent cation stabilizes malate in binding site 114
Figure 4.6 Average RMSF per MEZ residue 115
Figure 4.7 Conformational dynamics of MEZ active site 116
Figure 4.8 Putative $NAD(P)^+$ binding modes in MEZ after MD simulations 121

LIST OF TABLES

		Page
Table 2.1	Data summary for simulations of T4 Lysozyme L99A with p-xylene from alternate starting conformations.	53
Table 3.1	Heme and αBV binding affinities (K_d) for MhuD	72
Table 3.2	Statistics for X-ray diffraction data collection and atomic refinement for the MhuD-R26S-αBV complex.	75
Table 4.1.	Data collection and refinement statistics for MEZ crystal structure.	106

ACKNOWLEDGMENTS

In my time as a graduate student, I have been fortunate to have had the guidance and mentorship of two outstanding research advisors: Dr. Celia Goulding and Dr. David Mobley. They are each highly revered in their respective fields and are deeply committed to their responsibility to nurture and train future scientists. Thank you, Celia, for applying steady pressure to move me forward in my research, persistently advocating on my behalf, and for helping me navigate the highs, lows, and idiosyncrasies of life as a graduate student. Thank you to David, for always making yourself available, encouraging and providing opportunities to engage in the wider computational chemistry community, looking out for my general welfare and for trusting in my ability to work independently. Thank you both for unequivocally supporting and respecting my decision to start and grow my family while in graduate school. To that end, I would also like to acknowledge and thank my husband Scott - I could not have done this without you. Thank you for being an amazing father, partner, and coach. Finally, a special thank you to Mary and Henry Bugielski - for providing outstanding care and a warm and nurturing environment for our children, each and every weekday.

Chapter 2 is a minimally modified reprint of the published work: Burley KH, Gill SC, Lim NM, Mobley DL. Enhancing Side chain sampling using non-equilibrium Monte Carlo. *J Chem Theory Comput*. 2019 Jan 24;15(3):1848-1862. I would like to acknowledge financial support from the National Science Foundation (CHE 1352608) and the National Institutes of Health (1R01GM108889-01 and T32GM108561), and the Vertex Fellowship as well as computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513 and the infrastructure support from the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing Center (SDSC). I particularly thank my co-authors, Sam Gill for developing and pioneering the BLUES method and Nathan Lim for helping transform BLUES into a developer-friendly and user-friendly software package. I would also like to recognize Sukanya Sasmal (UCI) for help with docking, Gaetano Calabró (UCI) for basic MD training when joining the lab, and Victoria Lim (UCI) for help with umbrella sampling.

Chapter 3 is a minimally modified reprint of the published work: Chao A, Burley KH, Sieminski PJ, de Miranda R, Chen X, Mobley DL, Goulding CW. Structure of Mycobacterium tuberculosis heme-degrading protein, MhuD, variant in complex with its product. *Biochemistry*. 2019 Oct 22;58(46):4610-20. This work and its authors were financially supported by the National Institutes of Health (NIH) (P01-AI095208, T32GM108561), the National Science Foundation (DGE-1321846). The authors are grateful to Tom Poulos for discussions and critical reading of the manuscript, as well as Dr. Fengyun Ni, Baylor College of Medicine for advice on the more difficult aspects of structure refinement. The authors also acknowledge the Advanced Light Source at Berkeley National Laboratories (ALS) and the Stanford Synchrotron Radiation Lightsource (SSRL) for their invaluable help in data collection, and The Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing Center (SSDC) for their computing support. Regarding my co-authors, I especially thank Alex Chao for solving the structure, Xiaorui Chen for her

contributions in data processing and refinement, as well as Paul Sieminski and Rodger de Miranda for their biochemical characterization of the MhuD variant with its product.

Chapter 4 is part of a manuscript in preparation. I would like to acknowledge Bonnie Cuthbert for her tenacity and expertise in refining the crystal structure of MEZ, her analysis of the structure, as well as her moral support and friendship, Ervin Irimpan for his assistance in preparing protein and setting up crystal trays, Ilona Foik for the well-kept records of her prior crystallization attempts and purification methods, and Robert Quechol for his work in performing critical biochemical assays. I also appreciate financial support from the National Institutes of Health (NIH) (P01-AI095208, T32GM108561) and the National Science Foundation (DGE-1321846), as well as computing support from The Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputing Center (SSDC) and UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513.

VITA

Kalistyn Hope Burley

EDUCATION

Ph.D. in Pharmaceutical Sciences **June 2020**
University of California, Irvine Irvine, CA
Co-Advisors: Dr. David Mobley, Dr. Celia Goulding

A.B. in Biomedical Engineering **June 2009**
Dartmouth College Hanover, NH

PROFESSIONAL EXPERIENCE

Graduate Student Researcher September 2015 – April 2020
Mobley Lab, UC Irvine, Department of Pharmaceutical Sciences Irvine, CA

Graduate Student Researcher January 2016 – April 2020
Goulding Lab, UC Irvine, Department of Pharmaceutical Sciences Irvine, CA

Modeling and Informatics Intern June 2019 – September 2019
Vertex Pharmaceuticals San Diego, CA

Research Associate January 2014 – July 2015
Formulations Department, BioMarin Pharmaceuticals Novato, CA

Research Associate November 2011 - December 2013
Oda Lab, Children's Hospital Oakland Research Institute Oakland , CA

Research Assistant May 2011 – October 2011
Weier Lab, Biosciences Division, Lawrence Berkeley National Lab Berkeley, CA

Machine Shop Assistant April 2010 – July 2010
Thayer School of Engineering, Dartmouth College Hanover, NH

Undergraduate Research Assistant March 2008 – March 2010
Gerngross Lab, Thayer School of Engineering, Dartmouth College Hanover, NH

PUBLICATIONS

Burley KH, Cuthbert BJ, Basu P, Bhatt A, Irimpan EM, Quechol R, Foiks I, Beste DJV, Goulding CW. Structure and conformational dynamics of *Mycobacterium tuberculosis* malic enzyme. (*Manuscript in preparation)*

Chao A‡, **Burley KH‡**, Sieminski PJ, de Miranda R, Chen X, Mobley DL, Goulding CW. Structure of Mycobacterium tuberculosis heme-degrading protein, MhuD, variant in complex with its product. *Biochemistry*. 2019 Oct 22;58(46):4610-20. ‡*Contributed equally*

Burley KH, Gill SC, Lim NM, Mobley DL. Enhancing Side chain sampling using non-equilibrium Monte Carlo. *J Chem Theory Comput*. 2019 Jan 24;15(3):1848-1862.

Bannan CC, **Burley KH**, Chiu M, Shirts MR, Gilson MK, Mobley DL. Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge. *J Comput Aided Mol Des*. 2016 Nov;30(11):927-944.

Cuthbert BJ, **Burley KH**, Goulding CW. Introducing the new bacterial branch of the RNase A superfamily. *RNA Biol*. 2018 Jan 2;15(1):9-12.

Burley K, Goulding CW. Protein engineering: Redirecting membrane machinery. *Nat Chem Biol*. 2017 Aug 18;13(9):927-928.

Lemke KH, Weier JF, Weier HG, Lawin-O'Brien AR. High performance DNA probes for perinatal detection of numerical chromosome aberrations. *Adv Tech Biol Med.* 2015 Nov 3;3(3):155-164.

Stukas S, Robert J, Lee M, Kulic I, Carr M, Tourigny K, Fan J, Namjoshi D, **Lemke K**, DeValle N, Chan J, Wilson T, Wilkinson A, Chapanian R, Kizhakkedathu JN, Cirrito JR, Oda MN, Wellington CL. Intravenously injected human apolipoprotein A-I rapidly enters the central nervous system via the choroid plexus. *J Am Heart Assoc.* 2014 Nov 12;3(6):e001156.

Weier JF, Hartshorne C, Nguyen HN, Baumgartner A, Polyzos AA, **Lemke KH**, Zeng H, Weier HU. Analysis of human invasive cytotrophoblasts using multicolor fluorescence in situ hybridization. *Methods.* 2013 Dec 1;64(2):160-8.

O'Brien B, Zeng H, Polyzos AA, **Lemke KH**, Weier JF, Wang M, Zitzelsberger HF, Weier HU. Bioinformatics tools allow targeted selection of chromosome enumeration probes and aneuploidy detection. *J Histochem Cytochem*. 2013 Feb;61(2):134-47.

Hsu JH, Zeng H, **Lemke KH**, Polyzos AA, Weier JF, Wang M, Lawin-O'Brien AR, Weier HU, O'Brien B. Chromosome-specific DNA repeats: rapid identification in silico and validation using fluorescence in situ hybridization. *Int J Mol Sci.* 2012 Dec 20;14(1):57-71.

SELECTED ORAL PRESENTATIONS

Using publicly available structural data to identify and characterize integral membrane proteins. September 2019*,* Vertex Pharmaceuticals, San Diego, CA.

Enhancing side chain sampling with non-equilibrium candidate Monte Carlo. August 2019, ACS Women in COMP Symposium, San Diego, CA.

Applying NCMC sampling methods to aid in drug fragment screening of Mtb malic enzyme. December 2018, Advancement to Candidacy Exam, Irvine, CA.

Addressing sampling challenges in molecular simulations for drug discovery. March 2018, Vertex Day at UCI, Irvine, CA.

HONORS AND FELLOWSHIPS

NIH T32 Chemical and Structural Biology Predoctoral Trainee	2018 - 2019
Vertex Scholar Fellowship	2017 - 2018
NSF Graduate Research Fellowship Program Honorable Mention	2017
NSF Graduate Research Fellowship Program Honorable Mention	2016

ABSTRACT OF THE DISSERTATION

Development and application of molecular modeling methods for characterization of drug targets in *Mycobacterium tuberculosis*

By

Kalistyn H. Burley

Doctor of Philosophy in Pharmaceutical Sciences

University of California, Irvine, 2020

Professor David Mobley and Professor Celia Goulding

For decades, tuberculosis (TB) has persisted as a global health burden with over a million people succumbing to the disease each year. As cases of multi-drug resistant and extensively drug-resistant increase, there is an urgent need for the development of novel therapeutics to combat the disease. Fortunately, advances in protein structure determination and development of computational modeling tools provide valuable insight into the specific interactions the mediate binding between proteins and ligands. Nonetheless, the results of these methods are sometimes ambiguous or inconclusive. Taken together, structural determination and computational modeling can complement each other to provide more robust and interpretable data. In Chapter 2 of this work, I focus on the development of a computational method to enhance sampling of side chains, which can rearrange in the presence of different ligands. Specifically, I describe the development and validation of a non-equilibrium candidate Monte Carlo method for the enhanced sampling of side chain rotamers during molecular dynamics (MD) simulations. I validate

this method on a simple valine-alanine dipeptide and demonstrate that it significantly improves rotamer sampling of a buried valine sidechain in the binding site of model system T4 lysozyme L99A. In Chapters 3 and 4, we use X-ray crystallography and MD to elucidate and study the structures of two potential TB drug targets from Mycobacterium tuberculosis (Mtb): Mtb heme oxygenase (MhuD) and Mtb malic enzyme (MEZ). In the structure of MhuD in complex with product, we observe the formation of a novel α-helix; however, the unusual 5:2 ratio of product to protein subunit in the asymmetric unit, as well as the proximity of the helix to the crystallographic interface, provoke questions regarding its biological relevance. Using MD, I confirm that formation of the α-helix is favored in the presence of product and likely associated with product-turnover. In Chapter 3, I describe the X-ray structure of apo-MEZ along with results from differential scanning fluorimetry and gel filtration. MD simulations provide insight into interactions of MEZ with $NAD(P)^+$, Mn^{2+}, and malate, while also corroborating our empirical observations that MEZ is unusually disordered.

CHAPTER 1: An introduction to tuberculosis and computational methods for structure-based drug design

Forty years ago, in the early 1980s, tuberculosis (TB) was thought to be nearly eradicated, with global health experts targeting 2010 for total elimination of the disease.[1] Around this time, the October 1981 cover of Fortune magazine proclaimed that computer-aided drug design (CADD), as pioneered by Merck, was positioned to be the next industrial revolution.[2] So why then, decades later, does TB persist as the most lethal infectious disease year after year, ranking in the top 10 of causes of death globally? Why, in the age of increasing computing power and artificial intelligence, do pharmaceutical companies still require expansive teams of chemists, biologists and engineers to develop new therapeutics? Herein, I discuss the persistent and evolving challenges of treating tuberculosis as well as the contributions and limitations of computational methods for drug design.

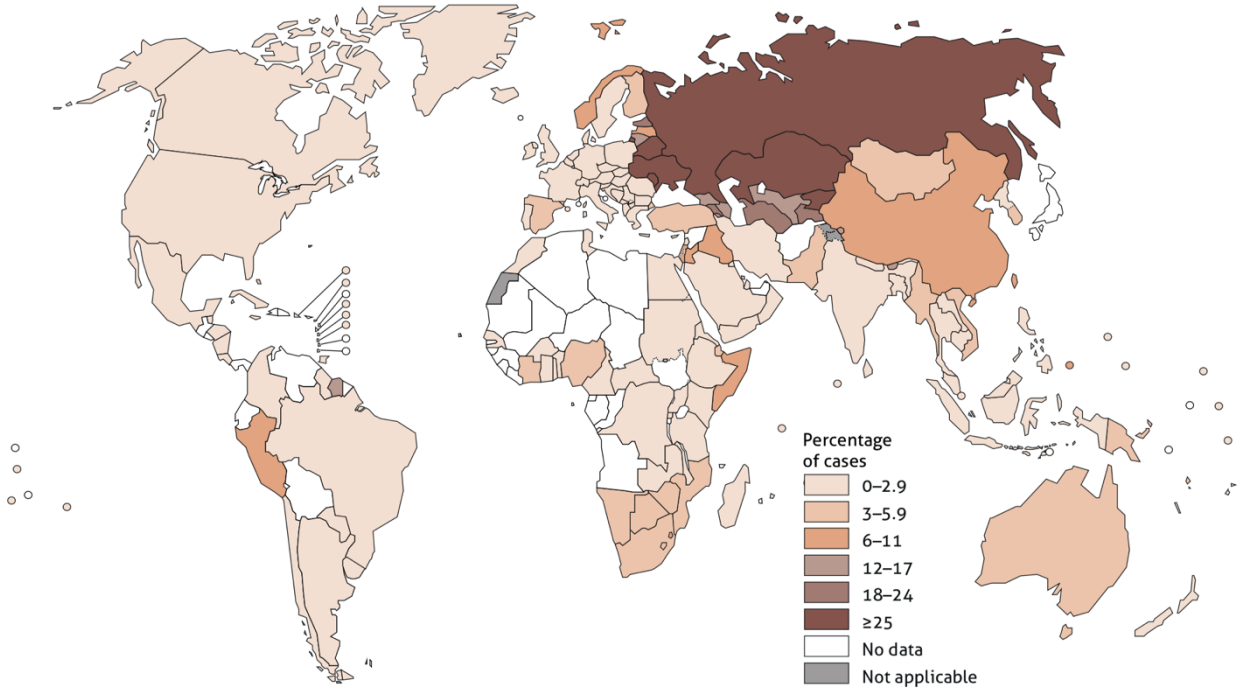
What is tuberculosis?

Tuberculosis is an infectious disease caused by a biological pathogen known as *Mycobacteria tuberculosis* (Mtb). It typically infects the pulmonary tissue and as such is most commonly transmitted via airborne exposure. Once it colonizes the lungs, it forms granulomas which help shield the mycobacteria from attack by the host immune system. Curiously, Mtb has a unique capacity to persist in the host cell macrophages in a quiescent state, wherein the infected individual is asymptomatic, and thus no longer contagious. This form of TB infection is known as latent TB and is generally harmless to the host at this stage. What makes latent TB so pernicious, is that, decades later if the affected individual

becomes immunocompromised via unrelated health issues or ageing, latent mycobacteria can re-emerge from its non-replicating state and transition into an active infection.

Resurgence of Tuberculosis

In fact, this is exactly what happened in the 1980s and 1990s with the rampant spread of HIV/AIDS. Accompanying the dramatic spike in the number of immunocompromised men and women among the general population, cases of active TB surged and thus, a growing number of otherwise healthy people were newly infected by Mtb. Tragically, individuals with HIV are particularly vulnerable to TB as reflected in the 251,000 deaths in 2018 among people co-infected with HIV/TB.[3] For HIV-



[a] Percentages are based on the most recent data point for countries with representative data from 2004 to 2019. Model-based estimates for countries with data before 2004 are not shown. MDR-TB is a subset of RR-TB.

Figure 1.1 Percent of new TB cases that are MDR or rifampicin resistant (RR-TB).[3] This figure from the WHO Global Tuberculosis Report 2019[3] highlights the prevalence of drug resistance among new TB cases. Drug resistance rates are highest in the Russian Federation, Ukraine, Belarus, Kazahkstan and their neighboring countries.

negative populations, TB took the lives of approximately 1.2 million people in 2018.[3] In all, the World Health Organization (WHO) estimates that 10 million people became ill with TB in 2018 and approximately one quarter of the world's population is infected with latent TB. While the overall rate of TB infection has plateaued in recent years, the number of drug-resistant (DR) or multi-DR (MDR) cases is increasing, accounting for 5% or 500,000 incidences of infection (**Figure 1.1**).[3] Alarmingly, extensively-DR (XDR) infections are also increasing and becoming more widespread with over 13,000 new cases reported from 81 different countries in 2018.[3]

How is tuberculosis treated?

Guidelines for the treatment of TB are set forth by both the WHO[4] and the Centers for Disease Control and Prevention (CDC)[5] and vary depending on the degree of drug resistance (DR) of the infecting Mtb strain. For drug-susceptible TB, the first line of treatment recommended by the CDC consists of a cocktail of drugs: isoniazid (INH), rifampin, ethambutol, and pyrazinamide, which are orally administered over 6-9 months.[5] For rifampin-resistant (RR), MDR, or XDR TB, a secondary line of drugs is available, including: levofloxacin, moxifloxacin, bedaquiline, delamanid, linezolid, and pretomanid, some of which require intramuscular administration. Until recently, the treatment guidelines favored use of injectable therapeutics for MDR and XDR for a shorter treatment of 9-12 months.[5,6] However, the WHO has shifted its recommendation to prioritize orally administered drugs for a longer course of 18-20 months, as the shorter treatment required a daily injection for the first 160 days.[6] Unfortunately, orally administered drugs also lead to unfavorable side effects, including but not limited to, gastrointestinal discomfort, rashes, joint pain, tingling in the extremities, and/or liver damage. While these side effects may be

tolerable for more typical 7-10 day prescriptions of antibiotics, the protracted 4-20 month administration period leads to patient non-compliance, wherein infected individuals discontinue treatment once primary symptoms subside. Sadly, partial or incomplete treatment facilitates further development of MDR and XDR Mtb strains, which require harsher, more prolonged regimens and thus perpetuate the cycle. As MDR strains continue to emerge, the development of new, multifaceted TB therapeutics is paramount for preventing this global epidemic from becoming an intractable global health crisis.

Challenges in TB Drug Development

Unfortunately, drug development is a tedious and risky endeavor, fueled by economic incentives. Because MDR TB is most prominent in poor countries, the economic incentives – that ultimately persuade pharmaceutical companies to risk investing hundreds of millions of dollars in the development of new drugs – are fewer. The biological and economic challenges associated with TB drug development are reflected in the creeping pace by which new TB drugs have entered the market. In fact, the first line treatment for drug-susceptible TB – consisting of a combination of isoniazid, rifampin, ethambutol and pyrazinamide – is nearly 50 years old.[7] Prior to approval of Bedaquiline in 2012 and Pretomanid in 2019, both for the treatment of MDR TB, the last TB-specific drug was developed in the late 1960s.[7]

Apart from economic hurdles, the infectious agent that causes tuberculosis, Mtb, also presents some unique biological challenges for drug development. As noted previously, Mtb can survive in its host in either an active or latent state; in its latent state, Mtb can persist for decades despite the acidic and nutrient-poor conditions of host

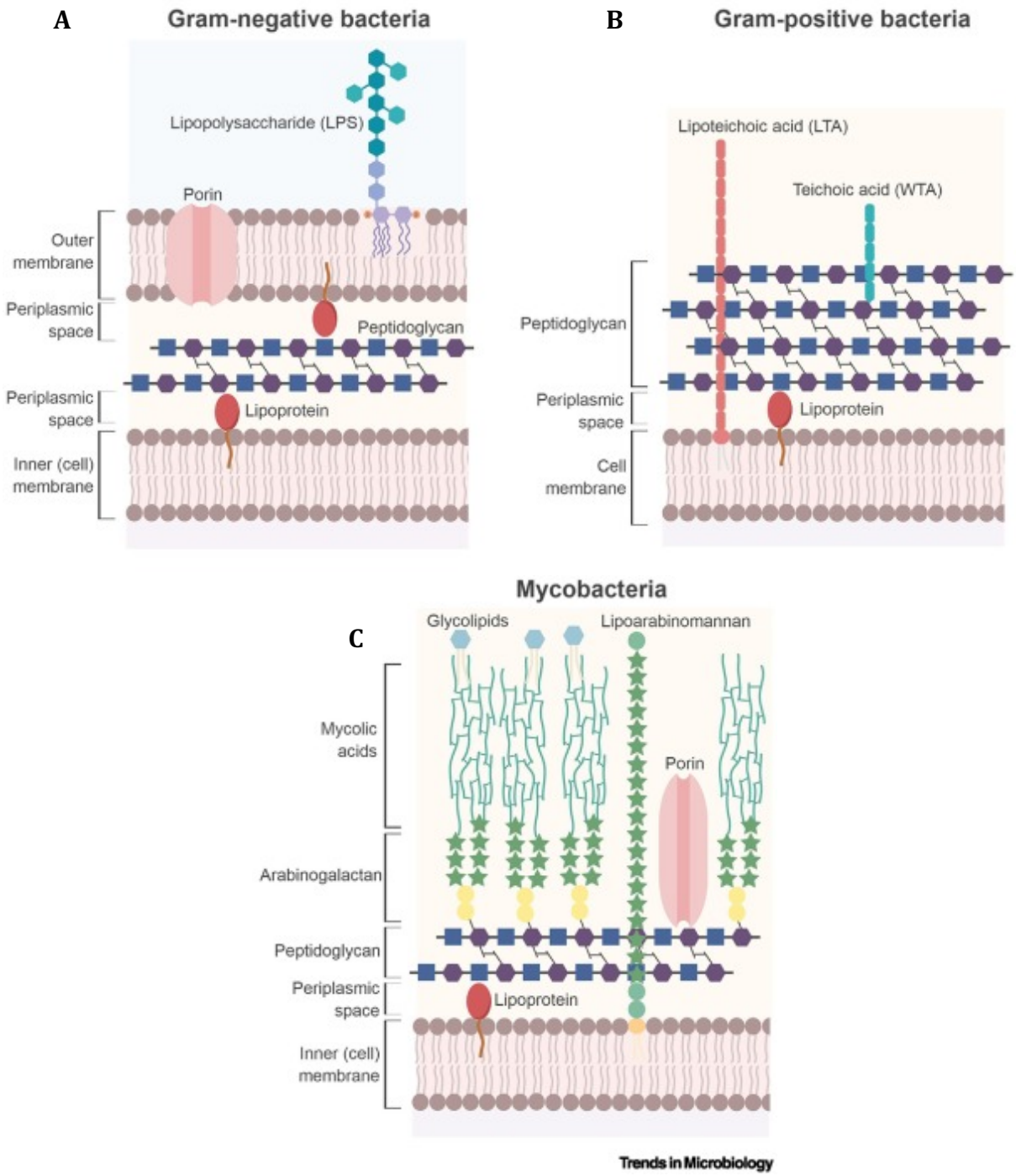


Figure 1.2 Comparing *Mycobacterium* cell wall composition.[8] **A**. Gram negative bacteria possess two phospholipid bilayer membranes mediated by the periplasmic space, which also contains peptidoglycans. **B**. Gram-positive bacteria have a single outer phospholipid bilayer cell membrane which is coated with peptidoglycans. **C**. Mycobacteria also possess a single phospholipid bilayer as an inner membrane. However, the outer cell wall is densely packed with peptidoglycans, arabinogalactans, mycolic acids, and glycolipids, which form a unique, protective waxy coating.

macrophages. Although the primary mode of transmission is airborne, infection by Mtb can also be spread indirectly; the bacterium has been shown to survive outside of the body in dust for several months.[9] The resilience of Mtb is due in part to the unique architecture of its cell wall (**Figure 1.2**). In addition to protecting against dehydration and withstanding attacks from a host's immune system, the thick, waxy cell envelope also creates a formidable barrier against the uptake of antibiotics.

Composition of the mycobacterial cell wall

Mycobacteria are neither classified as gram-negative (**Figure 1.2A)** nor gram-positive (**Figure 1.2B**) but rather represent a distinct class of bacteria. While the cell wall of gram-negative bacteria includes a set of inner and outer phospholipid bilayers mediated by the periplasmic space, the mycobacterial cell wall contains a single conventional phospholipid bilayer with a unique and complex outer membrane.[8] Also known as the mycomembrane, the outer membrane is rich in carbohydrates and lipids, notably arabinogalactans and mycolic acids which form a waxy coating on the its cell surface.[8] Apart from mycobacteria, arabinogalactans are also found in plants and are a major constituent of natural gum;[10] mycolic acids are unique to mycobacteria and form a mesh of long fatty acid chains which coat the cells with a thick hydrophobic layer. Mycolic acid synthesis is critical to survival and pathogenesis; the waxy mycomembrane of Mtb serves as armor, protecting the bacterium from assaults by the host immune system.[11–13] Because of the impenetrability of its cell wall, Mtb relies on a number of active transport mechanisms to regulate the uptake of nutrients and export of waste to maintain cell

homeostasis. Targeting enzymes critical for mycolic acid synthesis or nutrient uptake and processing are both promising means of treating tuberculosis.

Iron uptake and homeostasis in Mtb

One nutrient indispensable for Mtb survival is iron, which plays a critical role in facilitating bio-catalysis, electron transport, oxygen transport and storage.[14–17] Although iron acts favorably as a potent catalyst for enzymes, "free" iron has toxic effects; specifically, in its aqueous state, it activates free-radical chemistry through the Fenton reaction and catalyzes the formation of highly reactive superoxide radicals and peroxide. Due to its critical role in biological processes and the accompanying liabilities of its free form, iron homeostasis is tightly regulated through high-affinity import and export and storage mechanisms.[16,18–23]

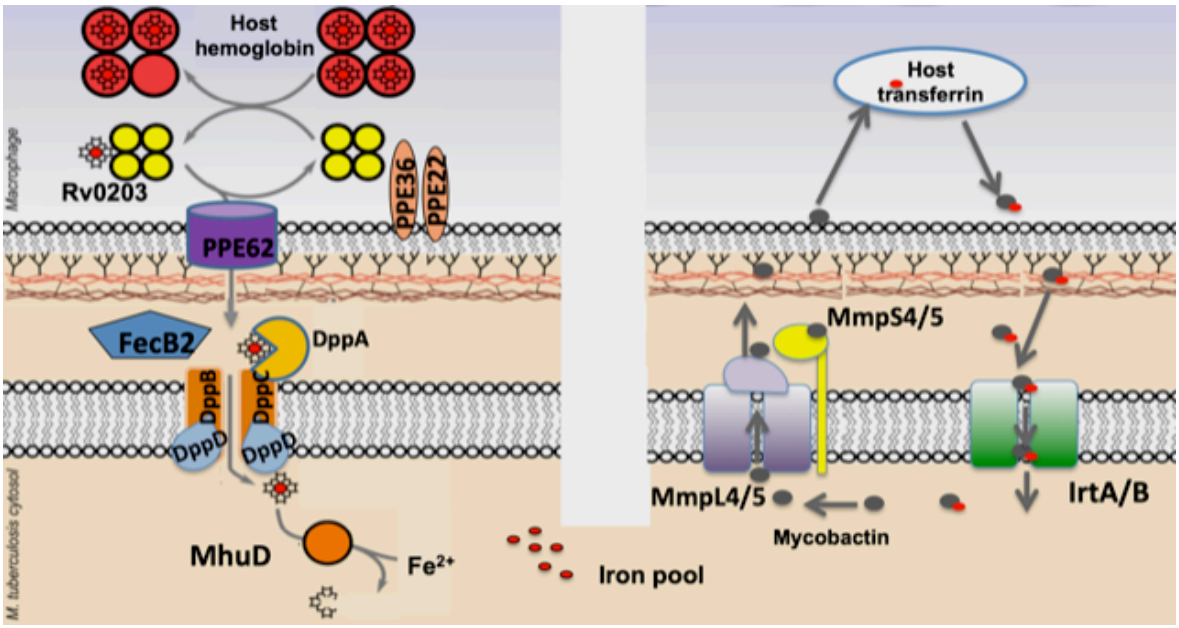


Figure 1.3 Heme and iron uptake pathways in Mtb.[18] The two primary means by which Mtb acquires iron are via the heme uptake pathway (left) and via the iron uptake system (right). Heme is acquired from host heme-binding proteins and actively transported into the cytosol, where it is oxidized by MhuD to release free iron. Alternatively (right), iron is also sequestered from host transferrin by way of siderophores which Mtb synthesizes and exports. Siderophores are recycled and re-exported upon release of the free iron in the cytosol.

Mtb has two well-characterized means of iron sequestration and uptake (**Figure 1.3**)[14,19,20]. One such method involves the synthesis and export of small (MW < 1000 Da) iron chelating compounds known as siderophores, which scavenge iron from transferrin, lactoferrin, ferritin and other iron-containing proteins in the host. In Mtb, these siderophores, known as mycobactin and carboxymycobactin, are hypothesized to diffuse freely across the macrophage phagosome to retrieve iron from more nutrient rich environments and return to the macrophage, wherein iron is otherwise limited.[21] Notably, free iron makes up just 1-2% of the host iron while 80% of host iron is found in in heme bound to hemoglobin and myoglobin.[22] As such, Mtb also has a highly evolved heme-uptake pathway to import and process heme from host hemoglobin to replenish and maintain its iron supply (**Figure 1.3**). In this process, heme is scavenged from host hemoproteins and imported through porins and active transporters into the Mtb cytosol, where it is degraded by the Mtb heme-degrading protein, MhuD, to release free iron.[19,23–25] The structure of MhuD in complex with its product, for which it has nanomolar affinity, is further discussed and described in Chapter 3.

Targeting cell wall synthesis is also promising

Another avenue for overcoming the formidable barrier of the Mtb cellular envelope is to target enzymes that play a role in directly synthesizing or providing substrates for the synthesis of cell wall components. Ethionamide (ETH) and INH are two such drugs that both eventually inhibit InhA, an enoyl-acyl carrier protein reductase, where the end result is the disruption of mycolic acid biosynthesis.[26,27] MDR TB co-resistance to first line drug INH and second line drug ETH is mediated by increased intracellular $NADH/NAD^+$

ratios.[28] INH and ETH must form adducts with NAD^+ in the cell to successfully inhibit their targets.[28] Recent work has revealed that Mtb malic enzyme (MEZ) knockout strains (Mtb∆*mez*) accumulate mycolic acids in the cytosol instead of the cell wall, which results in an abnormal colony morphology; furthermore, Mtb∆*mez* exhibits reduced uptake by macrophages.[11] It is speculated that MEZ serves as a source for the reductants NADH and NADPH, which are required by enzymes in the Mtb fatty acid synthase systems I and II.[11] Because of its demonstrated role in facilitating formation of the mycomembrane, MEZ is a promising drug target that may increase Mtb sensitivity to other antibiotics. In Chapter 4, I describe the apo X-ray structure of MEZ, propose $NAD(P)^+$ co-factor binding modes based on observations from molecular dynamics simulations, and discuss opportunities for structure-based design of MEZ inhibitors.

Using computational methods to expedite drug development

While striking, the dearth of novel TB drugs is not surprising given that drug development is a slow, costly, and failure-laden process. On average it takes 12-15 years to bring a drug from the initial discovery phase to market and only 1 in 5000 drug candidates is ultimately successful.[29] The estimated cost per drug is more than $2.5 billion,[29,30] in part due to the poor success rates as well as the redundant efforts among pharmaceutical companies that pursue overlapping drug targets.

The longest and most costly phase of drug development is the discovery phase, wherein a compound is identified and structurally modified until its binding, metabolic, and pharmacokinetic properties are optimized.[30] In-silico simulations and screening methods serve as tools for expediting this process and help make drug discovery more

accessible to academic and other non-profit institutions. One well known method for virtual screening is docking, which in its basic form, involves querying a library of three-dimensionally enumerated ligands against the structure of your target protein. Each ligand is placed or "docked" into the binding pocket of the protein and a scoring function approximates binding affinities of each ligand based on the resulting geometry, as well as the quality and quantity of intermolecular contacts formed in the docked protein-ligand complex. The ligands are then ranked according to their likelihood of binding the target, as predicted by the scoring function. The top scoring ligand, however, does not necessarily translate, in practice, to the ligand with the greatest binding affinity.

The objective of a high-throughput docking screen is not to identify *the* pharmaceutically active ligand that will enter into clinical trials, but rather to narrow down a much larger pool of ligands (say from 1,000,000 to 100,000) that will then be used for biologically guided high-throughput screening. Filtering out ligands that are less likely to bind the target, increases the likelihood of identifying positive hits while using a smaller, and thus cheaper, subset of the full ligand library for subsequent in-vitro screening efforts. If this process is successful, you will see an enrichment in the number of positive hits among your top-ranking compounds (**Figure 1.4A**).[31] It is generally understood and accepted that during this filtering step, ligands with potentially favorable properties will likely be discarded. To maximize the likelihood of success in docking, it is critical to begin with a structure of your target protein in a representative conformation that reflects the active (or inactive depending on the desired outcome) form. Ensemble docking, flexible docking, and algorithms which mix docking with molecular dynamics simulations are just a

few of the more elaborate adaptations that attempt to account for some of the dynamics of the protein and ligand that are otherwise ignored in a purely rigid approach.[32–35]

Another useful computational tool, which has more applicability in later stages of drug discovery, is the calculation of binding free energies (BFE). BFE calculations can accelerate efforts to optimize the binding properties of a lead compound. BFEs, once validated on a target, can help identify which chemical modifications to the lead molecule are most likely to enhance binding, thereby guiding decisions about how to focus synthesis efforts. Briefly, a BFE calculation computes the binding free energy ($\Delta G_{binding}$) between a ligand and a protein. Direct computation of this value, by monitoring association and dissociation events in simulations, is intractable and impractical given the multitude of interactions amongst the hundreds of thousands of atoms in a protein-ligand-water system and the values for typical on- and off-rates. Thus for drug discovery purposes, BFE calculations are performed by way of indirect calculations that are facilitated by taking advantage of the thermodynamic cycle (**Figure 1.4B**). The $\Delta G_{binding}$ can be computed indirectly by first computing the energy of de-solvation of the ligand from the surrounding water (ΔG_{water}) and secondly calculating the energy of decoupling the ligand from the protein ($\Delta G_{protein}$) as described in **Figure 1.4B**. For ease of computation, the calculations are performed by alchemically segmenting the transition from one state (eg. ligand in water to ligand in vacuum) into a series of small steps, also known as lambda (λ) windows. The number and length of these λ windows directly correlate to the computational cost of the calculation and have some impact on the accuracy of the calculations.

One challenge of BFE calculations is that the accuracy can depend greatly on the degree of conformational sampling accessed over the course of the λ windows.[36,37] As for

docking, if the most relevant conformational state(s) for that particular ligand-protein complex are not explored or included, the results can be negatively affected.[38,39] This particular obstacle is known as the "sampling problem."

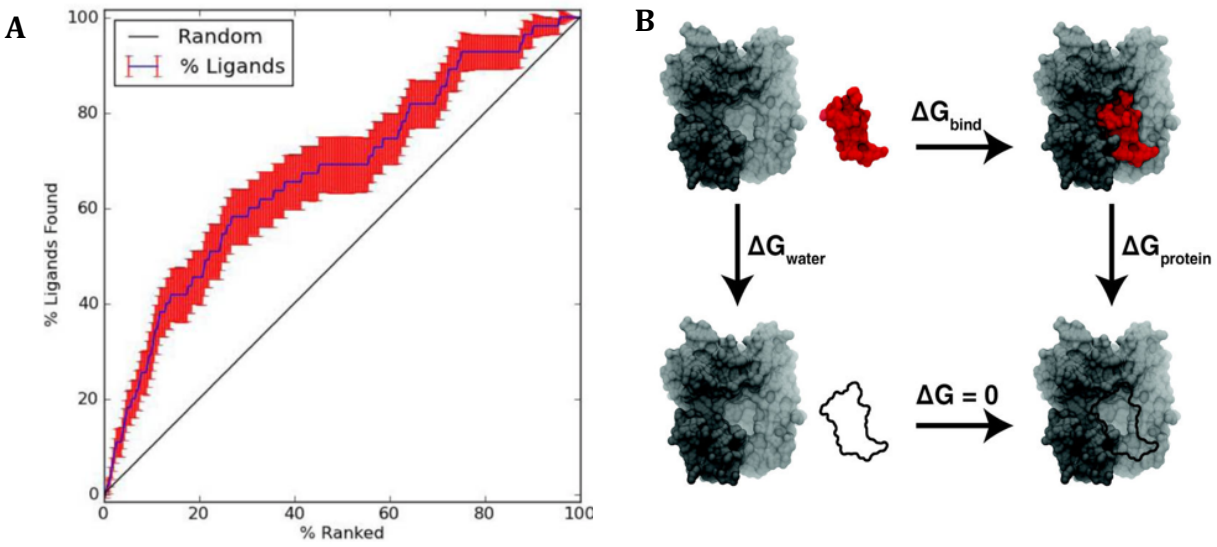


Figure 1.4 Computational drug development methods.[31,40] **A**.[31] Enrichment plots demonstrate the value of docking to screen ligands. In a test set of ligands containing known hits, docking can be used to rank the ligands according to their likelihood of binding the target receptor. If 10% of ligands were selected at random, one would expect to find 10% of the known hits (line, y=x). However, within the top 10% of compounds ranked by docking, you would expect to find >10% of the ligand hits (red/blue line). **B**.[40] Binding free energy calculations are used to estimate the binding affinity between a ligand and receptor. The thermodynamics cycle shown here demonstrates how the $\Delta G_{binding}$ between a ligand and receptor can be computed indirectly by computing the de-solvation energy of the ligand from the surrounding water (ΔG_{water}) and secondly calculating the energy of decoupling the ligand from the protein ($\Delta G_{protein}$).

The Sampling Problem

The degree of sampling in a classical molecular dynamics simulation is directly associated with the length of the simulation. If you could simulate a protein-ligand system for an infinite amount of time, assuming your system is set up correctly and your force field is accurate, you could explicitly identify the most dominant and therefore most relevant

motions and conformations of your system. As described in **Figure 1.5**, different types of protein motions have different timescales which are correlated to their energy barriers. Molecular simulations are typically performed on the order of nanoseconds and microseconds, although millisecond simulations are becoming more accessible with advancements in computing power. The Anton computers at Pittsburgh's Supercoputing Center, for example, can simulate a 100,000 atom system at a rate of more than 20μs/day.[41] One relatively accessible motion that is critical for accurate BFE calculations, are side chain rearrangements.

Sidechain rotamer states play a pivotal role in ligand binding; an extensive analysis of paired apo/holo structures from the Protein Data Bank (PDB) revealed that sidechain rotamer flips occur in nearly 90% of the binding sites with a bound ligand.[42] Depending on local energy barriers, the timescales of rotamer rearrangements can range from ps to microseconds, with higher energy barriers expected in more crowded environments like the interior of proteins. Although sidechain flips may occur on accessible simulation timescales, acquiring a representative ensemble of residue conformations in a binding site remains challenging, particularly if only a handful of transitions are observed over the course of a simulation. Poor sampling of these transitions can bias BFE calculations, leading to inaccurate results.[38,39]

Beyond improving accuracy of BFE calculations, sufficiently sampling sidechain rotamers during molecular simulations is also important for correctly predicting ligand binding modes that guide structure-based drug design, as well as in silico high throughput screening efforts. In an effort to remedy some of the sampling limitations in classical

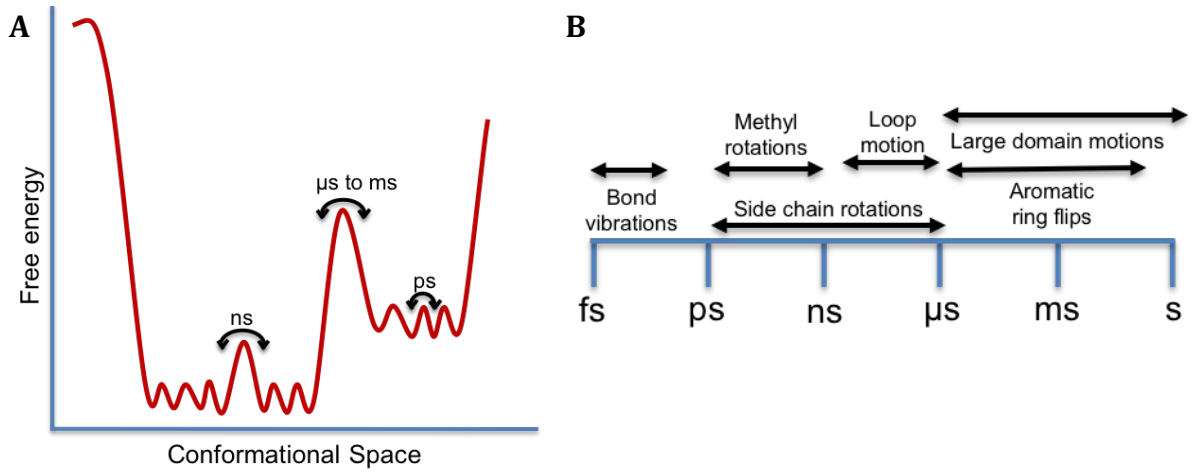


Figure 1.5 Approximate timescales of molecular motions.[43] **A.** The time scale of a particular molecular motion is correlated with the free energy difference between the starting and ending states. Fast motions, on the order of ns or ps, have relatively smaller energy barriers as compared with slower μs or ms motions. **B.** Several types of molecular motions are displayed here with approximate time scales (eg – bond vibrations are on the order of fs while larger domain motions range from μs to s). For some motions, the time scale can vary dramatically depending on the local environment, which may restrict motions by way of increased energy barriers.

molecular dynamics, I have developed and validated a novel method for enhancing side chain sampling using non-equilibrium candidate Monte Carlo; the details of this work are described at length in Chapter 2.

Conclusion

As a student of two distinct labs and disciplines, I have had the unique opportunity to work at the intersection of computational method development and X-ray crystal structure determination. While both are rooted deeply in physical principles and seek to produce "correct" results, method development also champions consistency and reproducibility. In that pursuit, method developers focus on model systems, for good reason, where variables and ambiguities are kept to a minimum. By comparison, X-ray crystallographers deal in uncertainty and must regularly confront the reality that reproducibility, while valiant, is not always possible. When it comes to applying

computational methods to real systems, the answers are not always clear nor even directly verifiable. For drug development, computational chemists and structural biologists do not have the luxury of choosing protein targets that are most amenable to structure determination or give the most consistent results. Rather, the most important consideration is the biology: does manipulating the selected protein target elicit the desired pharmacological outcome? Thus, in practice, computational modeling and structure determination become tools that can inspire alternative ways of thinking as opposed to correctly answering pre-defined questions.

Herein, I begin this dissertation in an area of certainty and precision with a method development project (Chapter 2), then show how computational methods and structure can complement each other (Chapter 3), and finally arrive in the land of speculation where computational and structural methods inspire questions rather than produce certainties (Chapter 4). Chapter 2 describes development of a general tool to improve side chain sampling in molecular simulations. Chapter 3 is a compelling example of how computational methods can be used to help validate structural observations. And, finally, Chapter 4 veers further into the land of uncertainty, as I describe a lower resolution structure of an unusually disordered enzyme and use molecular simulations to speculate with regards to how it interacts with its substrate, metal, and cofactors.

References

1. Dowdle, W. R. & Centers for Disease Control (CDC). A strategic plan for the elimination of tuberculosis in the United States. *MMWR supplements* **38**, 1–25 (1989).
2. Van Drie, J. H. Computer-aided drug design: the next 20 years. *J Comput Aided Mol Des* **21**, 591–601 (2007).
3. World Health Organization. *Global Tuberculosis Report 2019.* (World Health Organization, 2019).
4. World Health Organization. *Guidelines for treatment of drug-susceptible tuberculosis and patient care: 2017 update.* (2017).
5. Nahid, P. *et al.* Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. *Clinical Infectious Diseases* **63**, e147–e195 (2016).
6. WHO | Public Notice: Guideline Development Group meeting to update the WHO guidelines on drug-resistant tuberculosis. *WHO* http://www.who.int/tb/areas-of-work/drug-resistant-tb/treatment/drug-resistant-tb-gdg/en/ (2020).
7. Chakraborty, S. & Rhee, K. Y. Tuberculosis Drug Development: History and Evolution of the Mechanism-Based Paradigm. *Cold Spring Harb Perspect Med* **5**, (2015).
8. Porfírio, S., Carlson, R. W. & Azadi, P. Elucidating Peptidoglycan Structure: An Analytical Toolset. *Trends in Microbiology* **27**, 607–622 (2019).
9. Kramer, A., Schwebke, I. & Kampf, G. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC Infectious Diseases* **6**, 130 (2006).
10. Cell and Developmental Biology of Arabinogalactan-Proteins | Eugene A. Nothnagel | Springer. https://www.springer.com/gp/book/9780306464690.
11. Basu, P. *et al.* The anaplerotic node is essential for the intracellular survival of Mycobacterium tuberculosis. *J. Biol. Chem.* **293**, 5695–5704 (2018).
12. Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs, W. R. & Sacchettini, J. C. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. *Science* **279**, 98–102 (1998).
13. Marrakchi, H. *et al.* MabA (FabG1), a Mycobacterium tuberculosis protein involved in the long-chain fatty acid elongation system FAS-II. *Microbiology (Reading, Engl.)* **148**, 951–960 (2002).
14. Chao, A., Sieminski, P. J., Owens, C. P. & Goulding, C. W. Iron Acquisition in Mycobacterium tuberculosis. *Chem. Rev.* **119**, 1193–1220 (2019).
15. Pandey, M., Talwar, S., Bose, S. & Pandey, A. K. Iron homeostasis in Mycobacterium tuberculosis is essential for persistence. *Sci Rep* **8**, 1–9 (2018).
16. Rodriguez, G. M. Control of iron metabolism in Mycobacterium tuberculosis. *Trends in Microbiology* **14**, 320–327 (2006).

17. Ratledge, C. Iron, mycobacteria and tuberculosis. *Tuberculosis (Edinb)* **84**, 110–130 (2004).

18. Owens, C. P., Chim, N. & Goulding, C. W. Insights on how the Mycobacterium tuberculosis heme uptake pathway can be used as a drug target. *Future Medicinal Chemistry* **5**, 1391–1403 (2013).

19. Tullius, M. V. *et al.* Discovery and characterization of a unique mycobacterial heme acquisition system. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5051–5056 (2011).

20. Wells, R. M. *et al.* Discovery of a Siderophore Export System Essential for Virulence of Mycobacterium tuberculosis. *PLOS Pathogens* **9**, e1003120 (2013).

21. Luo, M., Fadeev, E. A. & Groves, J. T. Mycobactin-mediated iron acquisition within macrophages. *Nat Chem Biol* **1**, 149–153 (2005).

22. Zhang, A.-S. & Enns, C. A. Iron Homeostasis: Recently Identified Proteins Provide Insight into Novel Control Mechanisms. *J. Biol. Chem.* **284**, 711–715 (2009).

23. Owens, C. P. *et al.* The Mycobacterium tuberculosis secreted protein Rv0203 transfers heme to membrane proteins MmpL3 and MmpL11. *J. Biol. Chem.* **288**, 21714–21728 (2013).

24. Graves, A. B. *et al.* Crystallographic and spectroscopic insights into heme degradation by Mycobacterium tuberculosis MhuD. *Inorg Chem* **53**, 5931–5940 (2014).

25. Chim, N., Iniguez, A., Nguyen, T. Q. & Goulding, C. W. Unusual diheme conformation of the heme-degrading protein from Mycobacterium tuberculosis. *J. Mol. Biol.* **395**, 595–608 (2010).

26. Nguyen, L. Antibiotic resistance mechanisms in M. tuberculosis: an update. *Arch Toxicol* **90**, 1585–1604 (2016).

27. Vilchèze, C. & Jacobs, W. R. Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis: Genes, Mutations, and Causalities. *Microbiol Spectr* **2**, MGM2-0014–2013 (2014).

28. Vilchèze, C. *et al.* Altered NADH/NAD+ Ratio Mediates Coresistance to Isoniazid and Ethionamide in Mycobacteria. *Antimicrobial Agents and Chemotherapy* **49**, 708–720 (2005).

29. Petrova, E. Innovation in the Pharmaceutical Industry: The Process of Drug Discovery and Development. in *Innovation and Marketing in the Pharmaceutical Industry: Emerging Practices, Research, and Policies* (eds. Ding, M., Eliashberg, J. & Stremersch, S.) 19–81 (Springer New York, 2014). doi:10.1007/978-1-4614-7801-0_2.

30. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33 (2016).

31. Mobley, D. L. *et al.* Blind prediction of HIV integrase binding from the SAMPL4 challenge. *J Comput Aided Mol Des* **28**, 327–345 (2014).

32. Amaro, R. E. *et al.* Ensemble Docking in Drug Discovery. *Biophys. J.* **114**, 2271–2278 (2018).

33. Kelley, B. P., Brown, S. P., Warren, G. L. & Muchmore, S. W. POSIT: Flexible Shape-Guided Docking For Pose Prediction. *Journal of Chemical Information and Modeling* **55**, 1771–1780 (2015).

34. Totrov, M. & Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **18**, 178–184 (2008).

35. Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W. & Shoichet, B. K. Testing a Flexible-receptor Docking Algorithm in a Model Binding Site. *Journal of Molecular Biology* **337**, 1161–1182 (2004).

36. Mobley, D. L. *et al.* Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.* **371**, 1118–1134 (2007).

37. Deng, Y. & Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *Journal of Chemical Theory and Computation* **2**, 1255–1273 (2006).

38. Jiang, W. & Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *Journal of Chemical Theory and Computation* **6**, 2559–2565 (2010).

39. Mobley, D. L., Chodera, J. D. & Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J. Chem. Theory Comput.* **3**, 1231–1235 (2007).

40. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **9**, 71 (2011).

41. Request for proposals for biomolecular simulation time on Anton. *Pittsburg Supercomputing Center* https://www.psc.edu/anton-rfp (2020).

42. Gaudreault, F., Chartier, M. & Najmanovich, R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* **28**, i423–i430 (2012).

43. Dr. Peter Bolhuis. Bridging length and time scales in biomolecular systems. (2015).

CHAPTER 2: Enhancing side chain sampling using non-equilibrium Monte Carlo

Abstract

Molecular simulations are a valuable tool for studying biomolecular motions and thermodynamics. However, such motions can be slow compared to simulation timescales, yet critical. Specifically, adequate sampling of side chain motions in protein binding pockets is crucial for obtaining accurate estimates of ligand binding free energies from molecular simulations. The timescale of side chain rotamer flips can range from a few ps to several hundred ns or longer, particularly in crowded environments like the interior of proteins. Here, we apply a mixed non-equilibrium candidate Monte Carlo (NCMC)/molecular dynamics (MD) method to enhance sampling of side chain rotamers. The NCMC portion of our method applies a switching protocol wherein the steric and electrostatic interactions between target side chain atoms and the surrounding environment are cycled off and then back on during the course of a move proposal. Between NCMC move proposals, simulation of the system continues via traditional molecular dynamics. Here, we first validate this approach on a simple, solvated valine-alanine dipeptide system and then apply it to a well-studied model ligand binding site in T4 lysozyme L99A. We compute the rate of rotamer transitions for a valine side chain using our approach and compare it to that of traditional molecular dynamics simulations. Here, we show that our NCMC/MD method substantially enhances side chain sampling, especially in systems where the torsional barrier to rotation is high (≥ 10 kcal/mol). These barriers can be intrinsic torsional barriers or steric barriers imposed by the environment. Overall, this may provide a promising strategy to selectively improve side chain sampling in molecular simulations.

Introduction

Proteins are highly dynamic as they interact with hormones, ions, and other types of signaling molecules as well as other proteins. The motions that mediate these interactions comprise both large and small remodeling events, such as the shifting of domains to facilitate access to a catalytic site or the flipping of side chain rotamers in an active site to accommodate a binding event. Classical molecular simulations have proven a valuable tool for understanding these motions, and in some cases, quantitatively predicting properties like binding affinity or the kinetics for structural transitions.[1–4] Unfortunately, some of these motions are difficult to model and predict even as advances in computing power make longer timescale motions more accessible. In some cases, simulation timescales are still not long enough to capture the relevant motions, and such difficulties are often called "sampling problems".

Some motions are expected to be slow. For example, allosteric remodeling, protein unfolding, and ligand binding can often take microseconds to seconds. In contrast, simple side chain rotations are often faster in comparison but can still take anywhere from ps to several hundred ns, or even longer in the interior of proteins[5] — thus, they too can present sampling problems. Such motions are often simpler to detect and more tractable for sampling improvements than larger-scale rearrangements in biomolecules. As such, our particular focus here is on accelerating side chain motions with our enhanced sampling methodology.

Adequately sampling the populations of local conformational states and understanding the impact of ligand binding on local configurational entropy is critical for accurate free energy predictions. Although rotamer flips may occur on accessible

timescales, predicting the correct population distribution of a set of side chain rotamer states still presents challenges, particularly if only a handful of transitions are observed over the course of a simulation. Thus, poor sampling of these transitions can bias free energy calculations used for predicting ligand binding, leading to inaccurate results.[6,7] Previous studies on T4 lysozyme L99A have demonstrated that insufficient sampling of side chain motions during simulations for free energy calculations can lead to errors of several kcal/mol.[8–10]

Side chain rotamer states play a pivotal role in ligand binding. An extensive analysis of a carefully curated library of protein structures in the Protein Data Bank revealed that side chain rotamer flips in binding sites are both common and critical to binding. Among the >1000 distinct apo/holo pairs evaluated for conformational changes, nearly 90% of those structures undergo at least one rotamer flip in the binding site upon ligand binding; reconfiguration of five or fewer side chains account for 90% of those cases.[11]

Here our focus is on accelerating sampling of side chain rearrangements in proteins. Although conventional Monte Carlo (MC) can facilitate transitions of side chain rotamers across energy barriers, acceptance rates of such MC move proposals are particularly low for crowded systems. Thus, here, we use non-equilibrium candidate Monte Carlo (NCMC)[12] to improve acceptance of these MC moves. By mixing NCMC side chain sampling moves with classical molecular dynamics (MD) simulations, we improve side chain rotamer sampling in solvated systems relative to standard MD.

Theory

Ultimately, one of our long-term goals is to accurately predict binding thermodynamics, even when accompanied by protein conformational change. However, such predictions require that the representative states of the system be sampled with their correct populations. Thus, slow side chain rearrangements can pose substantial problems for accurate predictions. By focusing on enhancing side chain sampling in biomolecular simulations, our goal is to ultimately improve the accuracy of downstream applications used for prediction of binding modes and binding thermodynamics.

Other methods can be used to observe side chain flips

Side chain rotamer flips are observed both deliberately and incidentally by many simulation methods. Depending on the system and the magnitude of torsional energy barriers, classical MD may sample side chain rotamer flips and rearrangements. However, obtaining accurate populations requires simulating long enough to obtain many transitions of all relevant side chain rotamers, which could in some cases be prohibitively long.

Monte Carlo (MC) schemes can accelerate sampling across energy barriers by allowing hops *over barriers* (given suitable move proposals) rather than having barrier-crossing times be limited by normal kinetics. However, MC is not well suited for solvated or other crowded systems (eg. interior of proteins) as the overwhelming majority of proposed moves result in clashes and hence will be rejected. Thus for side chains, while very small rotational moves may be accepted, larger moves are very likely to result in collisions with the surrounding atoms and be rejected.

Since direct simulation using MD or MC is relatively impractical for improving sampling of side chain rotamer transitions, given the exponential dependence of barrier-crossing rates on the barrier height, biased or enhanced sampling techniques such as umbrella sampling may be more suited to this problem. Umbrella sampling applies biasing techniques to augment the sampling of selected motions and can be used to enhance side chain sampling[8,13] Although it is an efficient tool for generating the free energy landscape (or potential of mean force, PMF) of individual side chain rotamers, its utility declines as simulations and systems become more complex.

Specifically, umbrella sampling is challenging to apply to larger systems because of practical considerations arising from the difficulty of biasing more than one or two degrees of freedom at a time; even applications to sample rearrangement of a few side chain rotamers quickly becomes intractable. Imagine umbrella sampling an arginine side chain where we need to sample the rotation of multiple torsions simultaneously; this would require construction of a multidimensional PMF which, while tractable, involves substantial computational cost. The situation would be further complicated if the arginine neighbored another residue that we also needed to umbrella sample. Furthermore, integrating with other types of simulations would be laborious – for example, a simulation which sampled across ligand binding modes might need a separate umbrella sampling study and associated PMF for each binding mode.

Beyond logistical challenges related to implementation, a final challenge for umbrella sampling is that it requires advance knowledge about which side chains in a binding pocket are likely to rotate (in order to avoid wasting time umbrella sampling unimportant motions). Ideally one would like a tool which automatically determines which

motions are most likely to occur. Thus, while umbrella sampling is a valuable tool, its application can quickly become complex and require considerable human intervention and planning.

Tempering and annealing MD methods, wherein the simulation temperature is allowed to vary between high and low values to help overcome energy barriers, can also be applied to enhance sampling of side chain rearrangements. However, for slow motions, very high temperatures can be required in order to drive transitions over high energy barriers; this can lead to instability or unfolding of other regions of the system. Furthermore, raising the temperature shifts the equilibrium distribution, meaning that samples from higher temperature simulations require re-weighting for accurate free energy calculations. Methods such as replica exchange with solute tempering (REST/REST2)[8,14] provide an alternate approach for enhancing sampling that preserves ensemble distributions; in REST2, the intramolecular potential energy for regions of the protein can be scaled so as to alchemically decrease energy barriers and facilitate exploration of alternate states. Typically, REST2 is applied in binding free-energy calculations whereas here, we seek to develop a more general tool for enhanced side chain sampling in molecular simulations.

Non-equilibrium candidate Monte Carlo provides potentially a more general tool

Here, we apply non-equilibrium candidate Monte Carlo (NCMC) to improve sampling of side chain rotamer moves.

While related, MC and NCMC moves are executed and evaluated differently. In MC, a move is instantaneously attempted and accepted or rejected according to its impact on the

potential energy of the system. For NCMC, a move proposal proceeds via a non-equilibrium switching protocol[12] consisting of a series of perturbation and propagation steps wherein the structural and thermodynamic degrees of freedom are progressively sampled. During this stepwise process, the surrounding environment of the system can react to the perturbation before the move is accepted or rejected. This is especially useful for condensed phase systems as the surrounding solvent can make way for the repositioning of other atoms. Unlike MC, move acceptance depends not on the resultant potential energy, but rather the (nonequilibrium) work performed over the course of the NCMC move. The total work is tallied and the move is either accepted or rejected according to an acceptance criterion; this ensures the resulting states are sampled from the Boltzmann distribution and are representative of the equilibrium populations. If the simulation were done at constant volume and energy (NVE), the work done in this case would simply be the total change in potential energy as in conventional Metropolis Monte Carlo; however, because we use Langevin dynamics here, dissipative work is also done in the process and contributes to the total, so the total work is not equal to the change in potential energy. Rather, we accumulate the protocol work for the proposed perturbation (the work done in the context of making the perturbation) and accept or reject based on this work.[12,15]

In our case, a side chain rotational move proposal proceeds via a series of smaller steps in which the steric and electrostatic interactions between the side chain and the rest of the system are alchemically turned off, the side chain is rotated, and then the interactions are progressively cycled back on (**Figure 2.1**). The move is either accepted or rejected based on the nonequilibrium work done during this process.

Breaking the move into smaller, discrete segments allows the system to gradually relax and respond to the perturbation, thereby facilitating more frequent acceptance of larger positional perturbations. Particularly, nonequilibrium relaxation in NCMC mitigates steric clashes and other difficulties which would result in rejection of proposals in traditional MC, allowing for more ambitious move proposals.

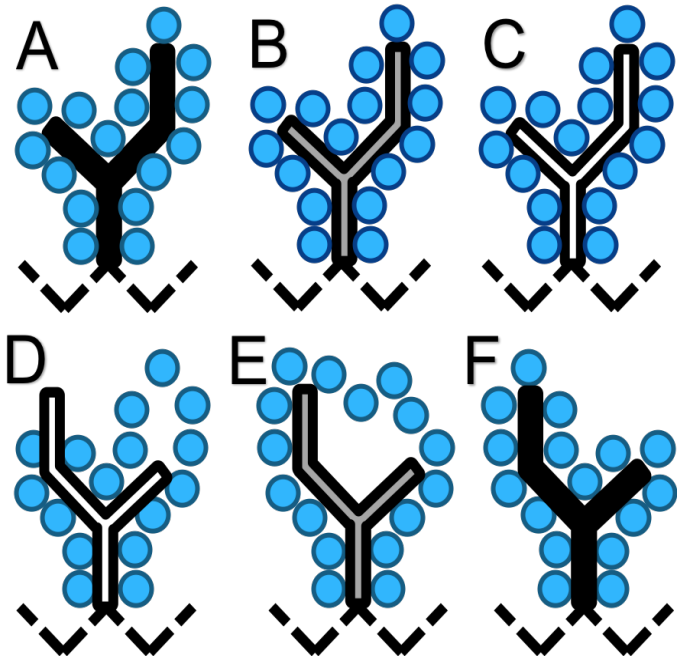


Figure 2.1: Cycling of atom interactions during execution of NCMC side chain move. The branched isoleucine side chain is depicted with a solid black outline while the backbone atoms are represented by a black dashed line. The blue circles represent all atoms proximal to the side chain, including the surrounding solvent. The fill color of the branched side chain reflects the degree of interaction with the surrounding atoms: black – full interaction, white – no interaction, and gray – partial interaction. A) The side chain is fully interacting with the environment. B) The side chain's interactions are partially off, allowing gradual relaxation of the surrounding atoms. C) The side chain's interactions are fully turned off. D) The side chain is randomly rotated around a significant rotatable bond; its interactions remain off. E) The side chain's interactions are partially turned on and the NCMC propagation steps facilitate relaxation of the rotated side chain to resolve clashes. F) At the end of the NCMC protocol the side chain fully interacts with the surrounding environment in a new orientation. The NCMC move is then accepted or rejected based on the work performed.

This is not the first study to apply NCMC to side chain sampling; another recent study applied it with some success but limited overall benefit, using a different protocol

than that employed here. Particularly, in the prior study, the nonequilibrium switching protocol only addressed the sampling of *structural* degrees of freedom without perturbing interactions within the system,[16] meaning that rotation of side chains across energy barriers involved applying a force to rotate the side chain across any torsional barrier and past any steric obstacles. In contrast, we apply NCMC to side chain rotations where we perturb thermodynamic properties — particularly, the strength of the interactions between the side chain and the rest of the system — removing steric barriers to rotation (though still retaining any inherent torsional barriers).

We implement NCMC side chain rotations as an extension of BLUES

BLUES[15] is a simulation package designed to enhance sampling of ligand binding modes by combining NCMC move proposals with intervals of standard MD. When applied to a model T4 lysozyme L99A system, the BLUES approach of mixing random NCMC ligand rotation move proposals with MD enhances binding mode sampling efficiency by more than two orders of magnitude, as compared with classical MD.

Implementation of NCMC and MD in BLUES

NCMC moves are implemented in BLUES by first cycling off the interactions between the target and surrounding atoms prior to the move. The electrostatic and then steric interactions are turned off and on by scaling λ (a variable controlling the strength of interactions between the side chain and the rest of the system) over a given number of n NCMC steps. First, the interactions are scaled from 1 to 0 over a series of $\frac{n}{2}$ steps, where λ=1 corresponds to full interactions and λ=0 corresponds to no interactions. When all interactions are turned off (at $\frac{n}{2}$), the target atoms are re-positioned — in this

case, a side chain is rotated — and the interactions are scaled back on in reverse order (steric and then electrostatic) from λ=0 to λ=1 over $\frac{n}{2}$ steps. The total work of decoupling, repositioning and recoupling of the atoms is summed and used to accept or reject the move. NCMC moves are alternated with intervals of traditional molecular dynamics to allow the system to undergo normal dynamics and relax further before attempting additional moves. The number of intermediate MD steps is specified by the user. BLUES also provides options for adding extra NCMC steps immediately preceding and following the rotational move (via an option called nprop) as well as freezing atoms distant from the region being perturbed during the NCMC simulation to help limit unintended motion and reduce variance in the work distribution, thereby improving move acceptance.[12,15,16] More robust documentation and details as well as the full BLUES package are freely available on GitHub at https://github.com/MobleyLab/blues, in the BLUES documentation (https://mobleylab-blues.readthedocs.io), and detailed in the work of Gill et al.[15]

Addition of side chain moves to BLUES

Here, we build upon the existing BLUES framework by introducing a new move type, wherein selected side chains are rotated by way of NCMC move proposals. Given a list of residue indices, BLUES identifies all significant rotatable bonds within the selected side chains, where a significant rotatable bond is here defined to mean a rotatable bond for a torsion for which neither terminal atom is a hydrogen. We use OpenEye's OEChem Toolkit to traverse and interrogate bonds in the target residues.[17] For each NCMC move proposal, one significant rotatable bond is randomly selected and rotated after scaling off

interactions between the downstream atoms and the surrounding environment. During this process, the work done during the NCMC move proposal is tracked and the move is either accepted or rejected accordingly using the Metropolis acceptance criterion.

We bias our moves according to known side chain rotamer states

Side chain moves can either be proposed by way of random rotations or by biasing move proposals according to known side chain rotamer distributions. In the case of side chain rotations, random moves would likely result in many proposed moves being rejected because they would place the torsion in a state with a particularly high energy. Because each NCMC move proposal has an associated cost, regardless of whether the move is ultimately accepted or not, it is advantageous to propose rational moves (i.e. moves to favorable rotameric states) rather than random ones. In addition to avoiding move proposals to high energy states, we also want to avoid using NCMC to attempt small, favorable moves within the same rotamer state; while small moves are more likely to be accepted, they are readily (and more economically) accessed using traditional MD.

Fortunately, in the case of side chain flips, the rotamer states are well-known; the structural conformations of the twenty, natural amino-acids have been extensively characterized;[18–23] here we use this *a priori knowledge* of low energy rotamer states to conditionally bias move proposals to known, favorable rotamer states. By avoiding move proposals to high energy states, we increase the likelihood that a move will be accepted. Move proposal biasing however, if applied haphazardly, can disrupt detailed balance, distorting the resulting distribution of states sampled. In order to ensure the reversibility of our move proposals, a biased move is only proposed when the rotamer is in

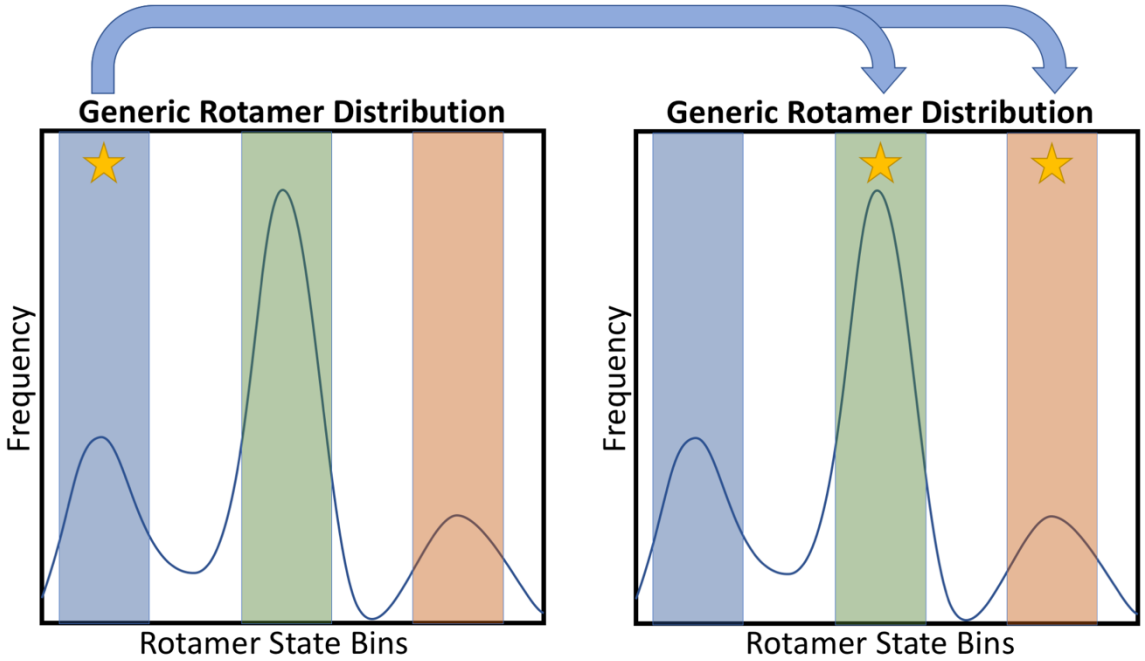


Figure 2.2: Move biasing is based on known side chain rotamer states. Shown here is a generic rotamer distribution possessing three dominant states. From these dominant states, one can define three equivalently sized bins highlighted in blue, green and yellow. In order for a biased move to be executed, the rotamer must start from one of these three bins. If the rotamer is in the blue state (starred on left plot), a move is randomly proposed to one of the two alternate states (starred on right plot). To ensure reversibility of biased moves, the final state of the rotamer following NCMC propagation must still fall within the defined region of one of these two alternate states. Although overall move acceptance decreases (as smaller, inconsequential moves are no longer attempted), side chain move biasing ultimately helps enhance efficiency by ensuring that only substantial moves that sample alternate rotamer states are attempted.

one of the pre-defined favored states, so that move proposals and their corresponding reverse rotations have equivalent probabilities (**Figure 2.2**). For the random rotations of ligands in BLUES, the intermixing of NCMC and MD proceeds in an alternating fashion — every *m* MD steps, an NCMC move is proposed (e.g., NCMC→MD→NCMC→MD). When using BLUES with biased side chain rotations, however, move proposals may be more irregular (**Figure 2.3**). Specifically, after *m* MD steps, an NCMC move is only proposed if the current rotamer state falls within one of the favored rotamer states, otherwise more another round

of *m* MD steps is executed executed and the process is repeated until the condition is satisfied (eg. MD→NCMC→MD→MD→MD→NCMC→MD).

To preserve move reversibility, the collection of starting states must match the states considered in rotational move proposals (**Figure 2.2**). Thus if, at the point of evaluation, the side chain rotamer is within the pre-defined range of favorable states, an NCMC move is proposed to another favorable state; if not, no move is proposed, and we follow up with another round of *m* MD steps (**Figure 2.3**). Relatedly, if an NCMC move is executed and the resulting state (after relaxation) falls outside of the range of favorable states, the move is immediately rejected. This scenario arises because the physical rotation of the atoms occurs at the midpoint of the NCMC switching protocol; as the intermolecular interactions are turned back on, the side chain atoms can move and sometimes fall outside of the acceptable range.

As noted previously, we bias our move proposals towards "large" rotations to ensure that NCMC is efficiently enhancing sampling alternate rotamer states; NCMC rotational moves are only proposed to a rotameric state that differs from the starting state, as shown in **Figure 2.2**. For example, in the case of valine, a rotamer in the trans ($\chi_1 \approx 180^\circ$) state will be limited to move proposals that would result in it occupying either the gauche(-) ($\chi_1 \approx -60^\circ$) or gauche(+) ($\chi_1 \approx 60^\circ$) rotameric state.

While rotamer biasing minimizes proposals to high energy states and increases the overall acceptance of larger moves (to alternate rotamer states), the overall acceptance rate suffers. Small rotational moves within the same starting state, which have a higher likelihood of acceptance, are never proposed, resulting in an overall lower acceptance rate

relative to unbiased moves. It is worth noting that, in contrast to conventional MC (and as in our previous work) the acceptance ratio is NOT the key criteria for efficiency here.

Specifically, we are much more interested in the acceptance rate of substantial moves than in the overall acceptance rate; one protocol might result in high overall acceptance rate but very poor acceptance of substantial torsional moves, whereas a protocol with a lower overall acceptance rate might result in much better acceptance of substantial moves. So here we examine not just overall acceptance rate but acceptance rate of substantial moves.

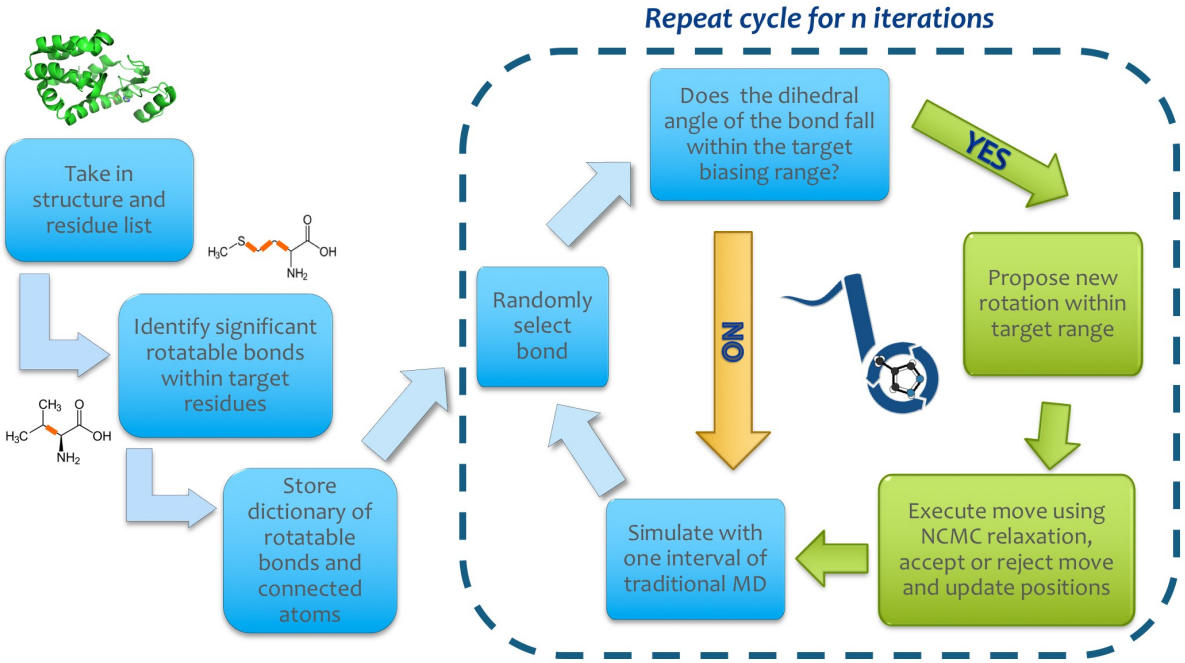


Figure 2.3: Workflow illustration of BLUES side chain proposals. Prior to executing any side chain type moves, a dictionary of target bonds and associated atoms is generated according to user inputs (outside of dashed box). Given the structure and a list of residue ID numbers, BLUES identifies significant rotatable bonds in those residues; these are bonds for which neither terminus is a hydrogen. It also identifies all upstream atoms that would move as a result of each significant bond being rotated. This information is compiled and stored in a dictionary. Thereon, BLUES cycles through a series of steps as it executes *n* NCMC side chain rotation move proposals (shown in the dashed box). First a bond is randomly selected from the dictionary. If the angle falls within one of the rotamer bins (as described in **Figure 2.2**), a move to a new rotamer bin is proposed and executed via the NCMC protocol (represented by green path). Otherwise (yellow path), the side chain move proposal is skipped and an additional round of molecular dynamics is executed before restarting the cycle by randomly selecting a bond.

Validation

To validate NCMC with MD for side chain rotations, we tested our method on a simple model system and compared it with brute force MD as well as umbrella sampling. Ultimately, we were interested in whether our BLUES with side chain rotations produces correct rotamer populations consistent with those obtained from umbrella sampling.

We use a valine-alanine dipeptide as our model system

Our chosen test system consists of a valine-alanine dipeptide, explicitly solvated in water. We chose to focus our tests on a simple dipeptide because this minimizes backbone motions and reduces the possibility of the side chains interacting with one another, which could make sampling more challenging and complicate validation or conflate sampling challenges with implementation errors.

Alanine dipeptide is a commonly used model system for molecular simulations;[24–31] however, the absence of rotatable bonds involving central heavy atoms in the alanine side chain makes it unsuited for our purposes. The next simplest dipeptide option is valine-alanine, which possesses a single significant rotatable bond in the valine side chain that has 3 distinct rotamers as shown in **Figure 2.4**. Initially a valine-valine system was tested, but there were challenges associated with controlling for the interdependence and influence of one residue's rotamer state on the other when generating reference results for comparison with BLUES (data not shown). A valine-glycine system was also explored, however, the absence of a glycine side chain resulted in the dipeptide backbone collapsing on itself.

We opted to include explicit water in our model system so as to better represent the typical solvated state of biological systems. We also sought to evaluate whether this method is viable in a solvated system; while previous BLUES ligand rotations were carried

out in an explicitly solvated system,[15] the ligand itself is buried in a binding pocket mostly devoid of local solvent molecules. Further, in recent work, others have encountered challenges when forming rotational moves on valine and methionine side chains in explicit solvent using a hybrid MD and NCMC approach,[14] providing an opportunity to see whether the approach employed here might fare better.

Here, in our valine-alanine dipeptide system, we use BLUES with side chain rotations to sample rotamer states for the valine side chain and compute the populations. We compare these results with populations obtained from a separate set of umbrella sampling simulations performed with more conventional techniques in order to validate our BLUES-based approach.

Preparation of valine-alanine input files

The input files for valine-alanine were prepared using tleap from AmberTools 15.[31] A linear valine-alanine peptide was generated with an N-terminal valine and a C-terminal alanine and parameterized using ff99SBildn. [32] The dipeptide was explicitly solvated in tleap with a 14Å rectangular box of TIP3P water, extending from the surface of the peptide to the box edge, and Cl atoms were added to neutralize the charge of the system. The system was then minimized using sander from AMBER14 [31] with steepest descents running for 20,000 steps and heated from 100K to 300K with constant volume for 25,000 steps using 2fs timesteps. Equilibration proceeded with sander for 500,000 steps and 2fs timesteps under constant pressure with positional restraints initially applied on all non-water atoms with a force constant of 50 kcal/mol/Å^2. and progressively lifted in increments of 5 kcal/mol/Å^2. over ten 50,000 step segments. The resulting topology and

coordinate files of valine-alanine were used as inputs for umbrella sampling, MD, and BLUES runs.

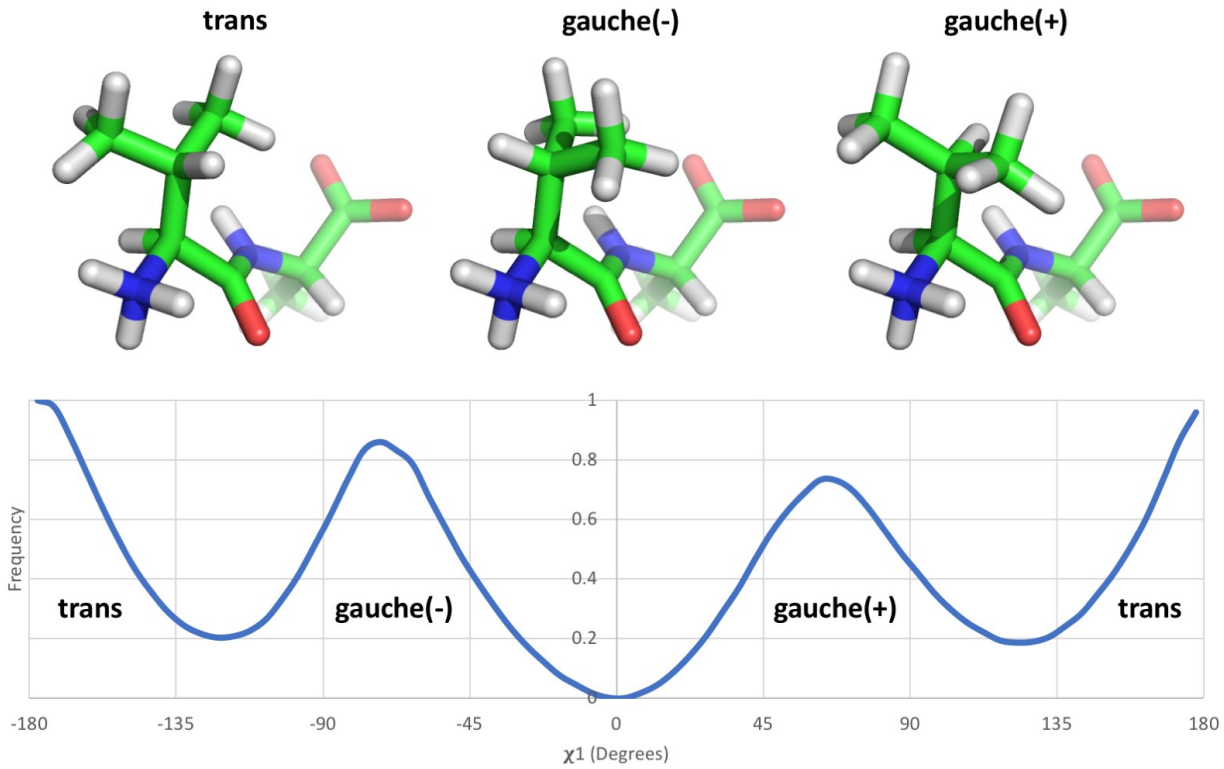


Figure 2.4: Rotameric states of valine side chain. Valine has three distinct rotamer states. The valine-alanine dipeptide structure is represented here with the N-terminus and valine side chain positioned at the front and the C-terminus and alanine side chain at the posterior. The three structures show the valine side chain χ_1 rotamer in each of its three states [trans, gauche(-), and gauche(+)]. The bottom panel shows the frequency of occurrence of the different states indicated above from simulations in solution; the data shown here is based on umbrella sampling of the valine-alanine peptide which is described and discussed further in **Figure 2.5**. While the qualitative rotameric preferences of valine are expected to remain consistent whether the side chain is in solution or in a binding site, the actual frequency of each state is likely to vary substantially depending on its surroundings.

Umbrella sampling methods

Umbrella sampling was executed in OpenMM 7.1.0[34] using a Langevin integrator with a 1fs timestep and a friction coefficient of 10/picosecond. The dipeptide backbone atoms were restrained with a force constant of 5 kcal/mol/Å^2. The atoms forming the dihedral angle of the valine χ_1 were harmonically restrained with a force constant of 200

kcal/mol/Å^2 for 36 10° χ_1 simulation windows (0°-350°). The simulation length for each umbrella window was 3ns (3,000,000 steps).

For analysis, the reduced potential energy — a dimensionless generalized form of the potential as a function of inverse temperature, pressure, volume, chemical potential and number of particles — was computed from the trajectory of each umbrella sampling window and the multistate Bennett acceptance ratio (MBAR) was used to estimate free energies.[35] The reduced potential is used by MBAR to help ensure the framework can easily extend to different ensembles without reformulation. The resulting free energy profile is shown in **Figure 2.5**. Full details are available in scripts deposited in Appendix A.

MD simulation details

MD simulations were executed with OpenMM 7.1.0[34] using a Langevin integrator with a 2fs timestep and a friction coefficient of 10/picosecond. Backbone atoms were restrained with a force constant of 5 kcal/mol/Å^2. For simulations wherein the torsional barrier was inflated (further details below), the periodic torsion force constant on the valine χ_1 dihedral was increased by 2.6X (from 3.8 to 10 kcal/mol) to mimic the highest barrier to rotation found for valine in the crowded environment of a binding pocket where the surroundings provide a steric barrier to rotation.[36]

BLUES simulation details

BLUES with side chain rotations was executed for 50,000 iterations, where an iteration is defined as one attempted NCMC move followed by at least one round of MD. Backbone atoms were restrained with a force constant of 5 kcal/mol/Å^2. After each NCMC move, at least 2ps of MD was carried out before the valine χ_1 dihedral angle was evaluated.

If the angle fell within the favorable range, an NCMC move was proposed and attempted; otherwise additional rounds consisting of 2ps of MD were run, with the angle re-evaluated after each, until the angle fell within specified favorable range, in which case an NCMC move was proposed. The three favorable dihedral angles for the χ_1 rotamer of valine are approximately centered near -60°, +60°, and ±180°; for these simulations, three equivalently sized rotamer biasing bins for the valine side chain were defined as: [-74°to -52°], [52°to 74°], and [169°to -169°] as conceptually illustrated in **Figure 2.2**.

BLUES generates accurate rotamer populations

Umbrella sampling was used to obtain a potential mean of force (PMF) for the valine side chain in the dipeptide and to obtain reference rotamer populations to validate our BLUES-based approach.

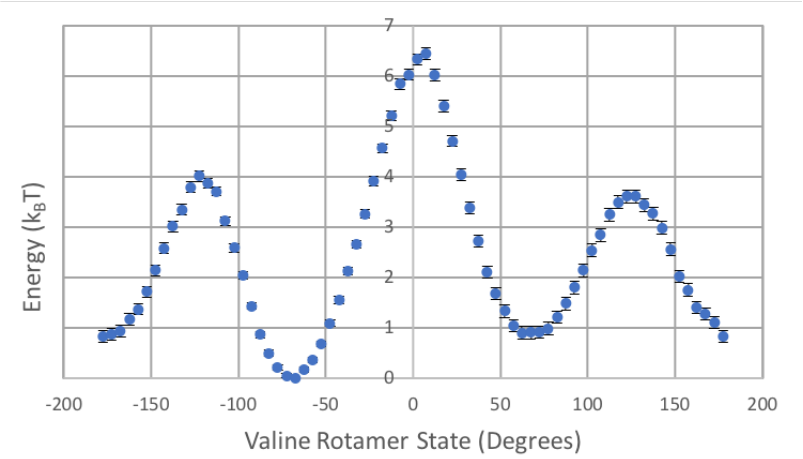
The output from umbrella sampling is the free energy landscape as a function of dihedral angle, which we can invert to estimate the population of particular rotameric states (**Figure 2.5**). The population is proportional to the negative exponential of the free energy divided by the Boltzmann factor:[37]

$$F_{(\chi 1)} \ \alpha \ e^{-G/k_B T} \qquad (1)$$

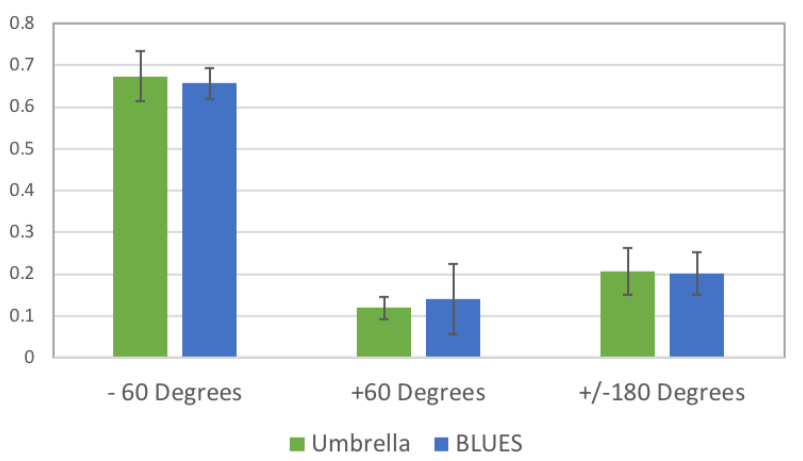
where $F_{(\chi 1)}$ is the frequency of a given rotamer state χ_1, G is the energy of that state, k_B is Boltzmann's constant and T thermodynamic temperature. To minimize population discrepancies arising from differences in backbone sampling using the different methods, the dipeptide backbone was restrained during the simulations.

Here, populations for each of the three dominant rotamer states for valine were estimated directly from the PMF and normalized to 1. The ranges for each dihedral rotamer

bin were defined as gauche(-): [-115°, 0°], gauche(+): [0°, 115°], and trans: [-115°, -180°] and [115°, 180°]. When compared to the normalized rotamer states sampled by BLUES (**Figure 2.5b**), the populations agreed within statistical uncertainty. We expect these results to be independent of the selected bin ranges as the state populations should be comparable



(a) PMF for valine in solvated val-ala peptide



(b) Histogram of valine rotamer populations

Figure 2.5: Umbrella sampling of valine-alanine peptide in explicit solvent. (*a*) As expected, the PMF analysis of the solvated valine-alanine system reflects energy valleys at the three dominant valine rotamer states (**Figure 2.4**). Error bars were generated using estimates of the standard error from MBAR analysis as provided by the pymbar package. (*b*) Here the populations of each valine rotamer state from umbrella sampling are plotted in green and populations from BLUES are plotted in blue. Umbrella sampling populations were estimated from the PMF using the Boltzmann relationship (Equation 1) Uncertainties were estimated by splitting the umbrella data into 10 chunks and splitting the BLUES data into 5 chunks prior to analysis, and computing the standard deviation in the population estimate across chunks.

between the two simulation methods (after convergence). We have confirmed this by recomputing populations using alternate bin ranges (data not shown). This serves to validate that BLUES with side chain rotations is indeed sampling the correct rotamer distribution and is implemented correctly.

We use the number of force evaluations to compare sampling efficiency

To compare the efficiency of classical MD and BLUES, we must account for the non-trivial cost of performing the NCMC switching protocol. Because BLUES includes intervals of both MD and NCMC, we consider the total number of force evaluations resulting from each simulation type when comparing with MD rather than comparing the total simulation time run. In terms of wallclock time, this may not be strictly comparable – e.g. alchemical perturbations in NCMC can incur some additional costs relative to force evaluations in standard MD, but also can be made faster by holding some parts of the system fixed. However, accounting for force evaluations still provides a good starting point for examination of efficiency.

For each perturbation or propagation step, NCMC executes one force evaluation. Hence a BLUES simulation consisting of MD and NCMC will have a total cost, in force evaluations (*FEs*) of:

$$FEs = (N + M)\times n + K \quad (2)$$

where N is the number of MD steps per move, M is the number of NCMC steps per move, n is the number of proposed NCMC moves, and K is the number of additional MD steps added for iterations where a rotamer is in a state outside of the target states (see yellow path in **Figure 2.3**).

For classical MD, the calculation is much simpler:

$$FEs = K \qquad (3)$$

where K is the total number of MD steps.

MD outperforms BLUES for simple valine-alanine system

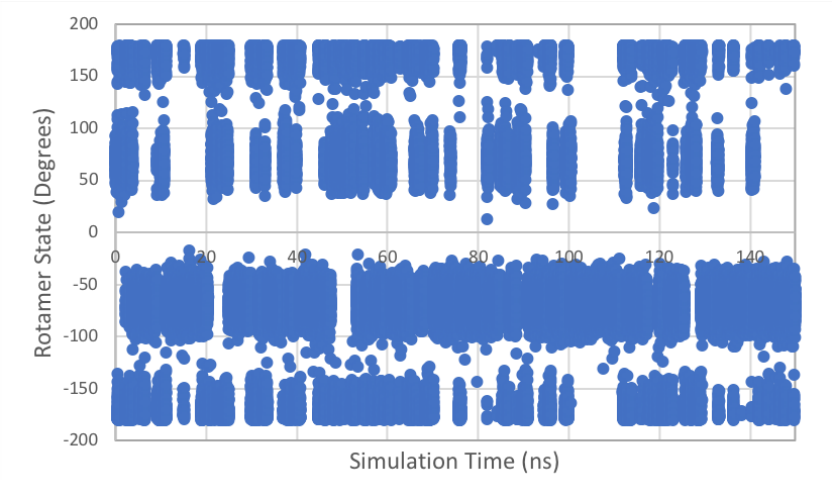
To compare the sampling efficiency of BLUES versus MD, we compute and compare the number of transitions per nanosecond of simulation time as well as the transitions per million force evaluations, accounted as described above.

Here, we define a transition as being a move of the valine χ_1 rotamer from one stable conformational state to another (eg. when χ_1 transitions from *gauche(-)* to *trans* or -60°to ±180°).

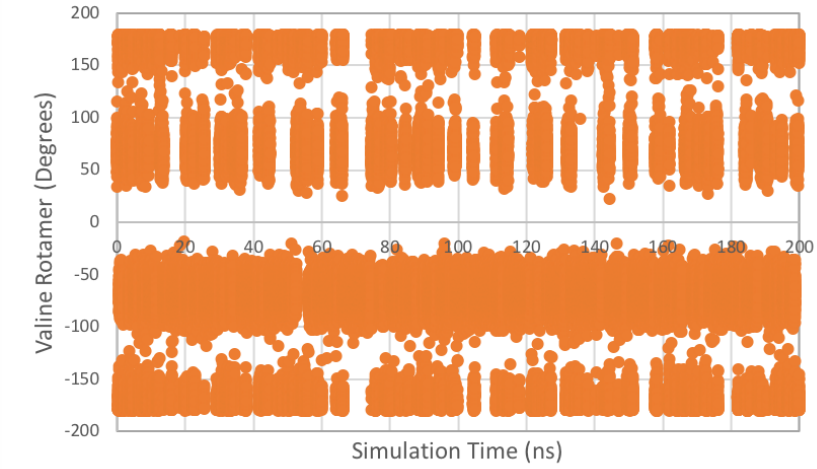
For this particular system, BLUES confers no advantage over classical MD (**Figure 2.6c**). There is no significant difference in the number of transitions per nanosecond of simulation time (1.9±0.3 vs 2.1±0.3 for BLUES and MD, respectively), but classical MD is significantly more efficient at sampling rotamer states by our metric of transitions per million force evaluations (2.3±0.4 vs 4.2±0.6 for BLUES and MD, respectively). As evidenced by the relatively high frequency of rotamer transitions for both BLUES and MD (**Figure 2.6**), the valine side chain in our dipeptide is one that is quite mobile and readily sampled

Figure 2.6: Valine-Alanine rotamer transition data where the torsional force constant, *k* = 3.8 kcal/mol. Rotamer data for valine χ_1 in solvated val-ala dipeptide system shown for BLUES and MD simulations. The *x* axes of (*a*) and (*b*), while different, represent roughly equivalent numbers of force evaluations (FE) when accounting for the costs of NCMC side chain moves. Each measurement represents 2ps of simulation time. (*a*) The dihedral angles of the valine χ_1 rotamer from the BLUES simulation are plotted in blue. (*b*) The dihedral angles of the valine χ_1 rotamer from the MD simulation are plotted in orange. (*c*) The transition frequencies from one rotamer state to another (eg. from *trans* to *gauche(-)*) are plotted for both BLUES (in blue) and MD (in orange). On the left are the transitions per ns and on the right are the transitions per 1e6 FEs. Error bars were generated by splitting the data into five chunks for analysis.

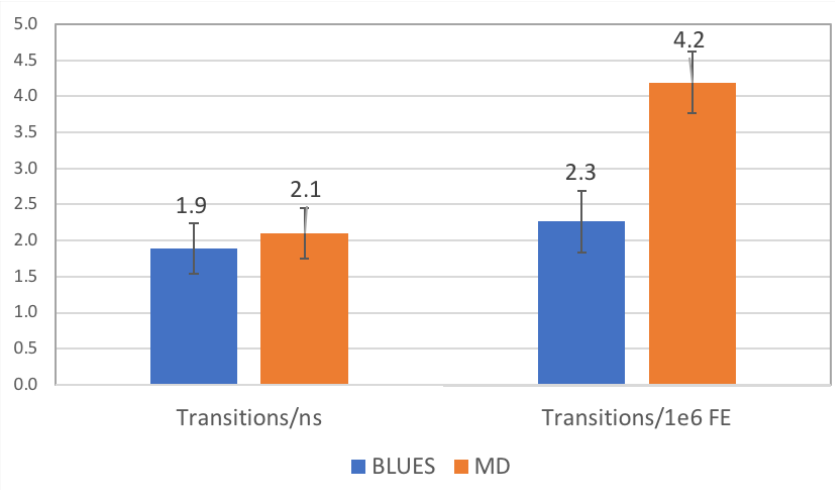
Torsional Force Constant k = 3.6 kcal/mol



(a) Valine rotamer transitions from BLUES



(b) Valine rotamer transitions from MD



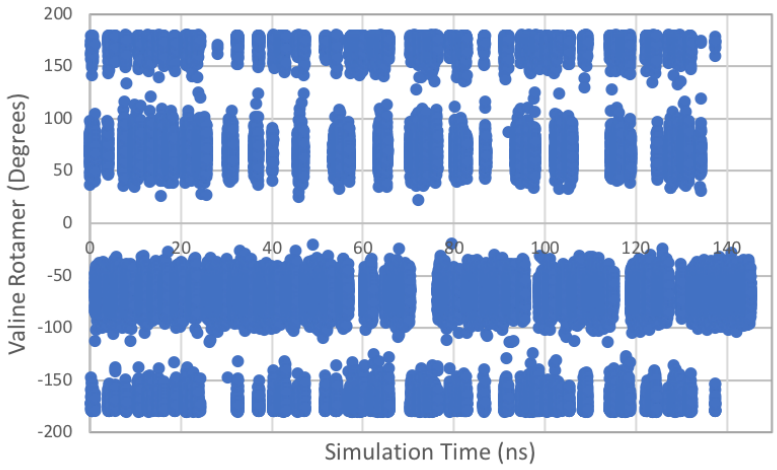
(c) Rotamer transition rates

by classical MD and thus unlikely to benefit from methods that specifically enhance side chain rotamer sampling at an added cost. As such, it is not a good test case for evaluating the potential benefits of BLUES in systems when transition barriers are large.

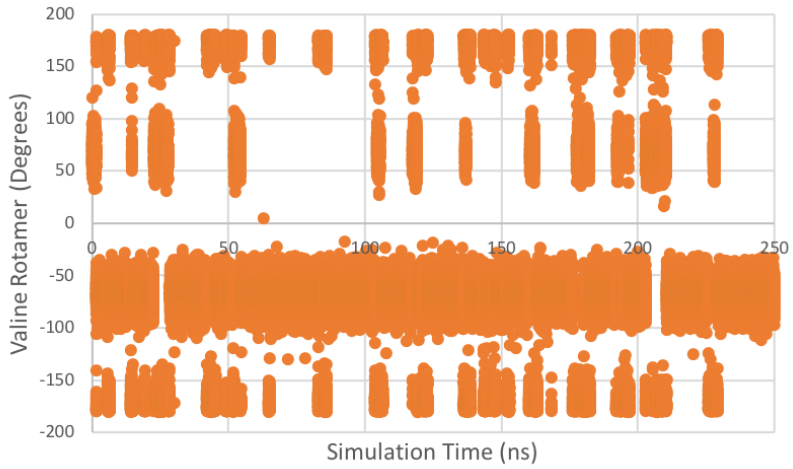
As noted above however, there are systems or regions of systems where side chain transitions are slow and the benefit of applying enhanced side chain sampling methods can perhaps in some cases outweigh the costs. In the interior of proteins (e.g. as in the case of the valine 111 side chain in the widely-studied L99A mutant of T4 lysozyme L99A[8,38,39]) the maximal barrier to rotation can be >10kcal/mol due to tight packing of the surroundings such as neighboring side chains.[38] To assess the relative performance of BLUES versus MD when barrier heights are higher (e.g. when torsional rotation is hindered by steric interactions with the surroundings) we updated our model system to have a higher barrier to side chain rotation. Specifically, by artificially elevating the torsional barriers for valine χ_1 (by increasing the torsional force constants), we can use the valine-alanine system as a test case for BLUES side chain sampling.

Figure 2.7: Valine-Alanine rotamer transition data where k = 10 kcal/mol. Rotamer data for valine χ_1 in solvated val-ala dipeptide system is plotted for BLUES and MD simulations where the torsional force constant (k) of valine χ_1 has been increased to 10 kcal/mol. The x axes of (a) and (b), while different, represent roughly equivalent numbers of force evaluations (FE) when accounting for the costs of NCMC side chain moves. Each measurement represents 2ps of simulation time. (a) The dihedral angles of the valine χ_1 rotamer from the BLUES simulation are plotted in blue. (b) The dihedral angles of the valine χ_1 rotamer from the MD simulation are plotted in orange. (c) The frequency of transitions from one rotamer state to another (eg. from *gauche(+)* to *gauche(-)*) are plotted for both BLUES (in blue) and MD (in orange). Bars on the left reflect the number of transitions per ns while the transitions per 1e6 FEs are shown on the right. Error bars were generated by analyzing the data in five chunks and computing the standard deviation across chunks.

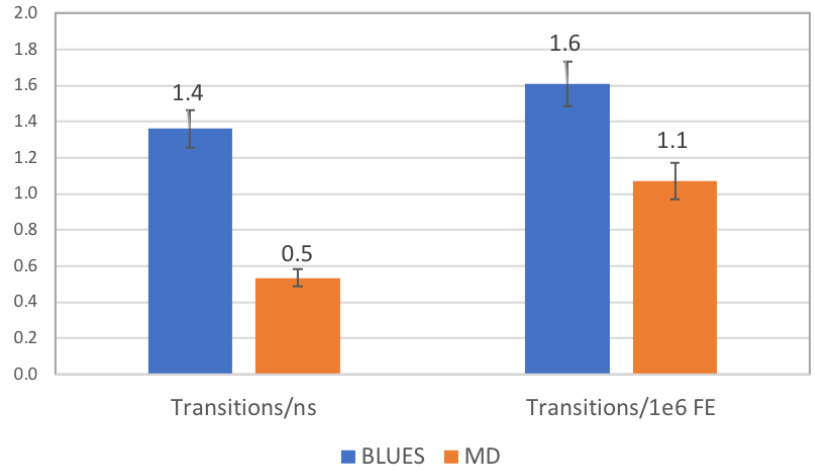
Torsional Force Constant k = 10 kcal/mol



(a) Valine rotamer transitions from BLUES



(b) Valine rotamer transitions from MD



(c) Rotamer transition rates

BLUES confers an advantage in systems where the torsion barriers are high

In the binding pocket of T4 lysozyme L99A, torsional rotation is blocked by the surrounding environment—leading to very high barriers to rotamer swaps.[38] Here, to mimic this phenomenon, we increase the periodic torsion force constant for the valine χ_1 by a factor of 2.6 (from 3.8 kcal/mol) such that the highest barrier to rotation is roughly equivalent to 10kcal/mol. Given the periodicity of the torsion being rotated, the maximum barrier to rotation may not represent the rate limiting barrier; nonetheless, by increasing the torsional force constant and thus the height of all barriers, we decrease the rotational mobility of the valine χ_1 and increase the typical timescale for rotation.

Compared to our unmodified valine-alanine system, there is a reduction in the total number of observed transitions between rotamer states as well as in the rates of transitions (**Figure 2.7**). When comparing transitions per nanosecond and transitions per force evaluation, BLUES exhibits a more favorable transition rate by both metrics when compared with MD (1.4±0.1 transitions/ns and 1.6±0.1 transitions/1e6 FEs for BLUES versus 0.50±0.04 transitions/ns and 1.1±0.1 transitions/1e6 FEs for MD). Thus, roughly as expected, BLUES becomes a better option for enhanced side chain sampling as barriers to rotation grow larger.

While we have demonstrated that side chain moves in BLUES can be used to enhance rotamer sampling, they should not be blindly applied to every system. NCMC can be costly, especially for side chain rotations where move acceptance is low, and one must first evaluate whether there is a likely barrier to rotation that is significant enough to be worth the additional cost. When barriers are low, standard MD may be more efficient than NCMC.

Testing on a model protein binding site

In line with our ultimate goal of improving accuracy of binding free energy calculations in pharmaceutically relevant systems, we are interested in applying NCMC side chain rotations to the binding sites of receptors, enzymes, and other biomolecules. Thus, we sought to assess the impact of our method on sampling of a side chain rotamer in a ligand binding site.

We use T4 lysozyme L99A as a model system

T4 lysozyme's L99A mutant (which forms a simple buried binding site that binds a series of nonpolar ligands) was chosen because it has studied extensively both crystallographically and as a computational model.[6,8,9,39–42] A cursory search of the RCSB PDB[43] for T4 lysozyme L99A returns over 120 structures containing nearly more than 90 distinct ligands and recently, absolute binding free energy calculations were calculated between T4 lysozyme L99A and 141 distinct ligands, using results from a more careful search.[44]

Previously it has been observed that a valine side chain in the binding pocket of T4 lysozyme L99A remodels its rotamer state in the presence of p-xylene, as compared with other ligands like toluene.[6,8,9] With benzene and toluene bound, valine 111 (Val111) adopts a trans conformation (χ_1=180°); however, in the presence of modestly larger ligands like p-xylene and o-xylene, the Val111 rotamer flips to the gauche(-) conformation (χ_1=-60°). This particular rotamer rearrangement has been used previously to test methods for sampling side chain rearrangement on ligand binding,[6,9,14] and has proven challenging to adequately

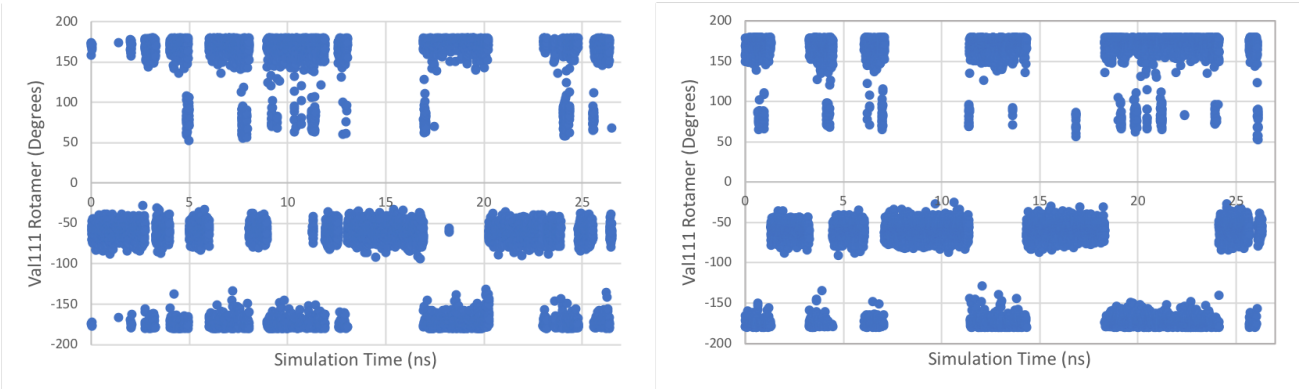
sample with standard MD simulations, driving applications of umbrella sampling[6] and Hamiltonian replica exchange.[8,14]

Preparation of input files for T4 lysozyme L99A with p-xylene system

The input files for T4 lysozyme L99A structure with p-xylene bound were prepared from the crystal structure (PDBID: 187L) retrieved the RCSB Protein Data Bank. Waters were removed and hydrogens were added to the system using pdb4amber from AmberTools15.[31] P-xylene was parameterized using antechamber from AmberTools15[31] with GAFF version 1.7 and AM1-BCC charges. Missing atoms of the lysozyme structure were added using tleap in AmberTools15,[31] and parameterized using ff99SBildn.[32] The system was explicitly solvated in tleap with a 10Å rectangular box of TIP3P water, extending from the surface of the protein to the box edge, and chloride atoms were added to neutralize the charge of the system.

BLUES results and comparison with MD

Compared to the 500 million FEs performed in the microsecond MD simulation, there were approximately 35 million FEs for each of the two BLUES runs comprising just under 30ns of simulation time as shown in **Figure 2.8** and described in **Table 2.1**. The total number of Val111 χ_1 rotamer transitions in each of these two simulations was 71 and 61. The rate of transitions (**Figure 2.9**) for these BLUES simulations were 2.0±0.8 and 1.7±1.0 transitions/1e6 FEs and 2.7±1.0 and 2.3±1.4 transitions/ns. The rotamer transition frequency per million FEs of BLUES shows nearly 20 fold improvement over traditional MD. However, because of the peculiarities in the microsecond trajectory noted previously,



(a) BLUES Replicate 1 (b) BLUES Replicate 2

Figure 2.8: Val111 χ_1 rotamer data for BLUES simulations of p-xylene bound T4 lysozyme L99A in explicit solvent. Here, Val111 χ_1 angle data is plotted for two BLUES simulations with identical user inputs for T4 lysozyme L99A in explicit solvent. Initial velocities were assigned using different random seeds in these two trials, so simulation results are distinct given the rapid divergence in trajectories which results. (a) The dihedral angles of the Val111 χ_1 rotamer for the first replicate are plotted in dark blue. A total of 71 transitions of Val11 χ_1 are recorded over 26.5 ns of simulation time with 35×10^6 FEs. (b) The dihedral angles of the Val111 χ_1 rotamer for the second replicate are plotted in light blue. A total of 61 transitions of Val11 χ_1 are recorded over 26.3 ns of simulation time with 35×10^6 FEs.

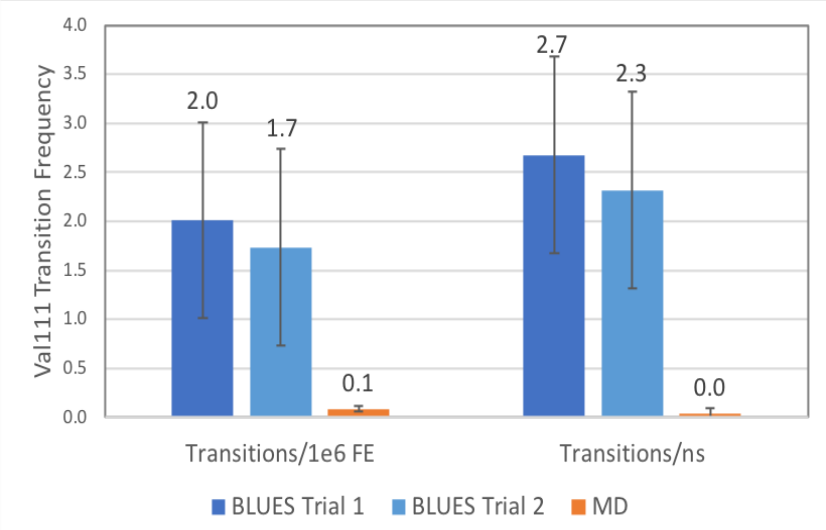


Figure 2.9: Val111 χ_1 transition rates in T4 lysozyme L99A bound to p-xylene for BLUES and MD. Rotamer state transition rates for two replicates of BLUES simulations are shown in dark blue and light blue while those for MD are shown in orange. On the left are transitions per 1×10^6 FEs and on the right are transitions per ns of simulation time. Error bars for transition rates were generated by splitting the simulation data into 3 chunks and computing the standard deviation across chunks.

(eg - decrease in transitions over time, low residence time at alternate rotamer states), we decided it was necessary to further analyze these simulations.

Slow relaxation of T4 lysozyme L99A backbone impacts sampling of the Val111 χ_1 rotamer

To further analyze the trajectory of T4 lysozyme bound to p-xylene, we used MDTraj to compute the backbone RMSDs of the whole system, of the local residues, and of Val111 for both the microsecond simulation and one of the BLUES simulations.

Details of MDTraj analysis

For all RMSD calculations, the initial input coordinates for T4 lysozyme L99A with p-xylene were used as the reference state; all simulations started from this state. The global backbone RMSD was computed using all backbone atoms in the system. The local backbone RMSD was computed using backbone atoms from all residues within 5Å of the Val111 side chain and the Val111 backbone RMSD was computed using the Val111 backbone atoms.

Analysis of backbone trajectories for T4 lysozyme L99A with p-xylene in MD and BLUES

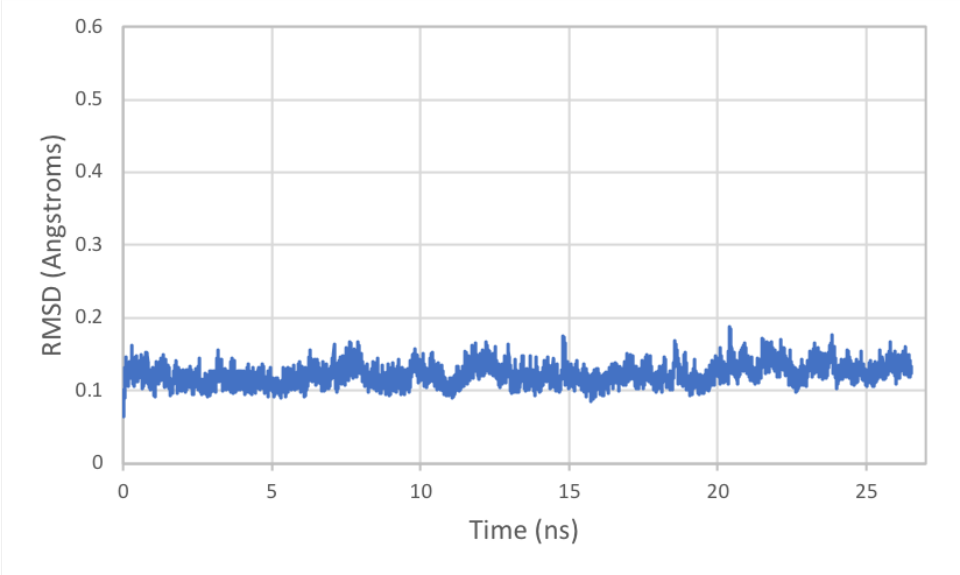
While there were no obvious differences for the backbone of the Val111 residue or backbones of residues local to the binding site (**Figure A.1**), there appears to be some larger systemic relaxation that occurs over the course of the longer MD simulation (**Figure 2.10**). Previous long-timescale simulations (5μs) of the T4 lysozyme L99A mutant in its apo form capture larger conformational changes as the protein transitions from a ground state to an excited state. During this conformational shift, the binding pocket expands and the χ_1 of phenylalanine 114[45] relaxes to an alternate rotamer state. Here, in our much shorter microsecond simulation, we may be observing some relaxation towards either the excited

or ground states; further analysis is needed to determine if and how the motions we see relate to those previously reported.

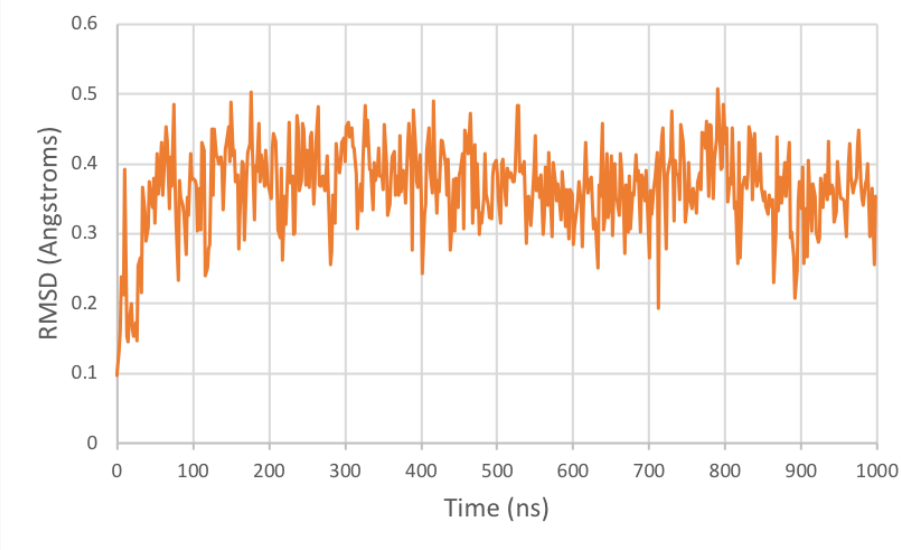
Overlaying the whole backbone RMSDs of BLUES and MD simulations (**Figure 2.11**) suggests that the BLUES simulations have not yet transitioned to this alternate state in the relatively short simulation time. As such, it could imply that the improvement in side chain sampling of Val111 by BLUES may be somewhat attributable to the fact that it is limited to sampling a particular conformational state when compared with MD, a topic we address further below.

BLUES enhances Val111 rotamer sampling in T4 lysozyme L99A at alternate conformational starting points

In order to more clearly explore and evaluate the performance of BLUES as compared to MD given the observed backbone remodeling, simulations of T4 lysozyme L99A were run using different snapshots from the microsecond trajectory as starting points. The first set was run using the initial input coordinates for the system (time 0). The next set commenced from the 250ns point, just after the observed backbone transition. And the last set of simulations was run using a snapshot from the end point of the microsecond simulation. Hereafter, these starting points will be referred to as T0, T250, and T1000, respectively. For each starting point, 60ns of classical MD data was generated such that the trajectory frames were written with the same frequency as BLUES; this was important for ensuring that no fast transitions were missed as a result of differences in reporting frequency (as could have also affected the prior comparison for the microsecond simulation).



(a) MD - all residues



(b) BLUES - all residues

Figure 2.10: Backbone RMSDs for T4 lysozyme L99A simulations. RMSDs of T4 lysozyme L99A backbone atoms in BLUES and MD simulations RMSDs of backbone atoms from the microsecond MD simulation of T4 lysozyme L99A with p-xylene bound are plotted in orange while those from the shorter BLUES simulations are plotted in blue. (*a*) and (*b*) RMSDs of all backbone atoms in T4 lysozyme L99A bound to p-xylene.

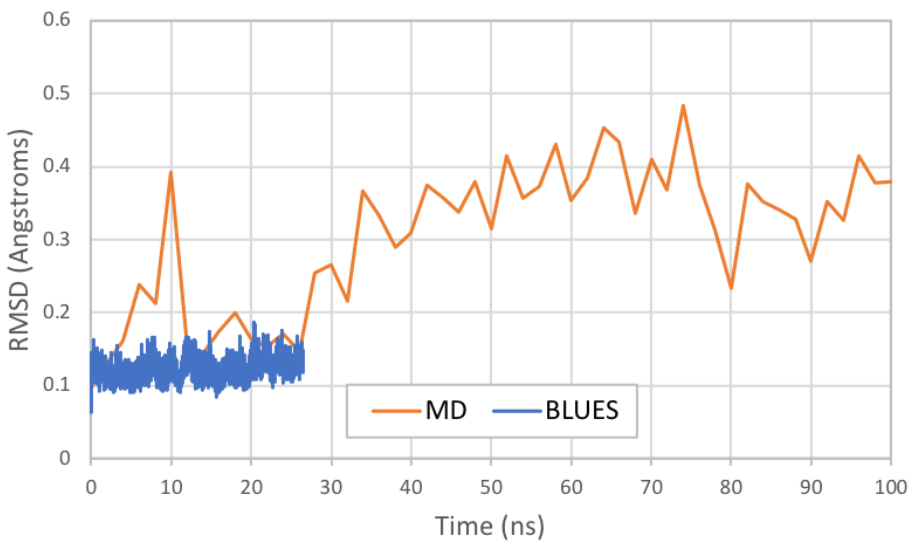


Figure 2.11: Overlay of backbone RMSDs of T4 lysozyme L99A for BLUES and MD (first 100ns). The backbone RMSDs for all residues in T4 lysozyme L99A for the BLUES simulation (**Figure 2.10b**) are plotted in blue while those from the first 100ns of the MD simulation (**Figure 2.10a**) are plotted in orange.

MD and BLUES simulation details for protein relaxation tests

The three sets of starting coordinates were generated from the microsecond trajectory of T4 lysozyme L99A bound to p-xylene, with the first being the input coordinates (T0), the second being a snapshot after 250ns (T250), and the third from the final state after 1 microsecond of simulation (T1000). MD was executed in OpenMM 7.1.0[34] using a Langevin integrator with a 2fs timestep and a friction coefficient of 10/picosecond.

BLUES simulations were run for 10,000 iterations with 2ps increments of MD between each Val111 dihedral state evaluation, repeated as needed until the rotamer fell within the biased dihedral range described previously. Each NCMC move was executed for 1004 steps with *nprop = 3*, such that NCMC steps were weighted toward the middle (lambda 0.2 to 0.8) of the NCMC switching protocol as described previously. Each BLUES simulation consisted of 22-24 million FEs comprising 24 - 28ns of simulation time. Comparatively, each MD simulation ran for 30 million FEs, equivalent to 60 ns of simulation

time. See **Table 2.1** for specific FE and timescale data. Full details are available in scripts deposited in **Appendix A**.

Comparison of MD and BLUES from three starting conformations

As shown in **Figure 2.12** and **Table 2.1**, the Val11 χ_1 rotamer transition frequency decreases for the simulations that start after the backbone relaxation (T250, T1000).

For the MD simulations, transitions are only observed in the T0 simulation using the original input files. For all three starting points, BLUES exhibits a marked improvement in sampling the rotamer states compared to MD, with MD showing one or zero transitions to the alternate rotamer state. However, as observed in our long MD simulation, there is a decline in rotamer transition rate (and associated move acceptances) for the T250 and T1000 simulations, compared to T0. (**Table 2.1**). Thus, the rotamer transitions per million force evaluations in BLUES drops from 3.8 to 2.1 after the backbone relaxation and down to 1.3 in simulations starting from the final structure (**Figure 2.12**). This further suggests that some sort of system relaxation is causing an increase in the torsional barrier for Val111. Nonetheless, BLUES still performs favorably compared to standard MD, dramatically accelerating rotamer transitions.

For both BLUES and MD, the gauche(-) rotamer is the dominant state regardless of the starting point; however, the increase in the Val111 torsional barrier post-backbone relaxation is reflected in the relative increase in occupancy of this state. For BLUES, the relative occupancies of the gauche(-) conformation for T0, T250, and T1000 are 0.58, 0.78 and 0.88 respectively (with not enough data yet collected for convergence). By comparison,

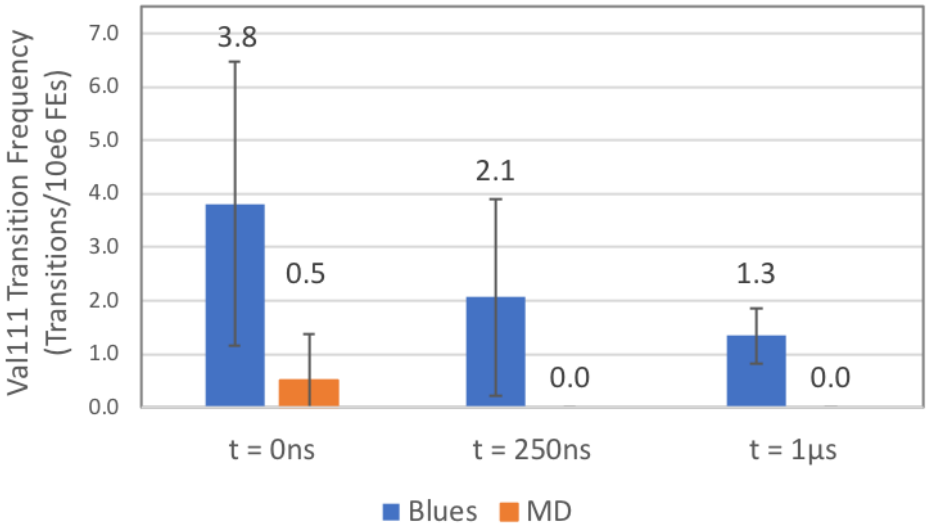


Figure 2.12: Val111 χ_1 state transitions per 10^6 FEs from simulations from 3 distinct starting conformations of p-xylene bound T4 lysozyme L99A. The rotamer transition rate for the 3 BLUES and 3 MD simulations are plotted in blue and orange, respectively and are in units of transitions per million FEs. As shown in **Figure A.2d** and **Figure A.2f**, there are 0 recorded transitions for MD simulations starting from the 250ns frame as well as the 1µs frame. Error bars for transition rates were generated by splitting the simulation data into 3 chunks and computing the standard deviation across chunks.

Table 2.1: Data summary for simulations of T4 Lysozyme L99A with p-xylene from alternate starting conformations.

Method [Starting Conformation]	*Acceptance Rate*	*# of Trans*	*Time (ns)*	*FEs (10^6)*	*Trans/ 10^6 FEs*
MD [T0]	n/a	15	60.0	30.0	0.5
Blues [T0]	0.11%	91	27.7	23.9	3.8
MD [T250]	n/a	0	60.0	30.0	0.0
Blues [T250]	0.04%	47	25.3	22.7	2.1
MD [T1000]	n/a	0	60.0	30.0	0.0
Blues [T1000]	0.08%	30	24.8	22.5	1.3

the relative occupancy of the gauche(-) rotamer state for the MD simulations is 0.97, 1.0, and 1.0, though the occupancies computed from MD are even further from convergence than those from BLUES since the number of transitions is far lower. Overall, it seems that some slow protein relaxation is impacting at least the likelihood of rotamer transitions, but potentially also the relative preference of different rotameric states. From this data we can

see that BLUES is effectively enhancing the sampling of the Val111 χ_1 rotamer. However, in the absence of efficient backbone sampling, the improvements to binding free energy calculations may be limited. While we think the slow backbone relaxation observed here may be relatively uncommon amongst receptor-ligand systems, further study of this observed relaxation and its impact on ligand binding and binding free energy calculations now seem warranted.

Despite this challenge, our results clearly show that BLUES provides accelerated sampling of side chain motions in systems with substantial barriers to rotation. This was modestly true in our model dipeptide system, but is especially borne out for side chain sampling in the case of p-xylene bound the T4 lysozyme L99A mutant.

Discussion and Future Work

Side chain BLUES enhances sampling of side chain rotamer states, as compared with classical MD, for systems where there is an existing (or artificially inflated) high torsional barrier. Here we have demonstrated an increase in sampling efficiency for both our valine-alanine dipeptide as well as our larger T4 lysozyme L99A/p-xylene system which has known side chain sampling problems,[6,8,14] and this improved sampling has the potential to improve binding free energy calculations. However, in the case of the T4 lysozyme L99A/pxylene system studied here, without long exploratory classical simulations, alternate backbone configurations could be missed; in classical MD, these appear to essentially lock the relevant side chain in a single rotameric state by increasing the transition barrier, whereas with BLUES, transitions are still possible. Nonetheless, the BLUES framework readily allows for integration of various move types and in the future, side chain sampling could be combined with enhanced backbone sampling.

In our work here, the application of side chain BLUES has been limited to sampling of a single valine side chain with one significant rotatable χ bond; most amino acid side chains, however, have multiple rotatable bonds. Furthermore, rotamer flips in a binding pocket are not always isolated but often coupled with remodeling of neighboring amino acid rotamers.[11] Moving forward we are interested in evaluating how this method performs in sampling more complex residues as well multiple side chains in a single simulation.

As written, the side chain BLUES method can be used to sample multiple χ angles in a given side chain, as well as multiple such angles from several side chains; this is implemented in our existing BLUES framework by randomly and sequentially applying the NCMC procedures to these target bonds, as described below. Given a user-input list of residues, all significant rotatable bonds are identified (**Figure 2.3**). Each side chain bond torsion or χ_i has an associated distribution of states (where *i* is the selected chi index in a given side chain): valine has one χ with three predominant states; lysine has four χ each with three prevailing states; phenylalanine has two χ, one with three dominant states and the other with two states (the aromatic bonds are not rotatable and thus are ignored). The preferred rotamer states for each of the 20 natural amino acids has been described previously and is readily obtained from the literature.[18–20,23] The biasing ranges are set by identifying the favored angles for each χ and then then creating a range (eg.±15°) around those angles. Currently, only one χ is rotated per move proposal.

So let's imagine we are interested in specifically sampling a lysine side chain in our BLUES simulation. For each NCMC move proposal, one of the four χ within lysine would be randomly chosen and a move would be proposed to an angle within 15° of the three favored angles for that selected χ_i. In simulations where multiple residues are sampled, all χ

torsions are identified prior to simulation and one is randomly selected for each NCMC move proposal. For example, a series of move proposals where Val111 and Lys104 are sampled might proceed as follows: Move Proposal 1 - Val111 χ_1 rotated to 63°; Move Proposal 2 - Lys104 χ_4 rotated to -170°; Move Proposal 3 - Lys104 χ_2 rotated to 91°. In the future, we may bias moves according to a continuous rotamer library and also rotate more than one bond at a time.

Given the flexibility of its implementation, BLUES can be used to explore and more readily identify residues in a protein binding site that are most likely to undergo some sort of conformational reformation, and sample such rearrangements much more efficiently than MD. It can also be mixed with different BLUES move types (some established and some in development) such as ligand flips and internal bond rotations, ligand translocations, and water translocations.

As noted, BLUES is not practical or beneficial for all applications (ie. sampling of solvent exposed residues as in our simple dipeptide), and one should think carefully about whether and how to apply the method. In practice, one may not know which residues may have high barriers to rotation and would thus benefit from enhanced sampling. In these cases, we suggest running short exploratory MD simulations for which the binding pocket residue dihedrals are plotted and evaluated to identify those with few or no transitions. This analysis could be further aided by examination of existing crystal structures (if available) of the target protein; any side chains with slow transition rates in-silico that have alternate orientations in crystals would most likely benefit by enhanced sampling by NCMC/MD.

Furthermore, the rotamer state of a given side chain may depend on the positions of surrounding residues. Increasing the number of MD steps between NCMC move proposals can help facilitate transient exploration of alternate conformations of the proteins and may allow for some degree of coupling among side chain conformations. However, in its current form, side chain moves in BLUES may not be suited for systems with side chains that must rotate simultaneously. Future work will include testing on a system wherein a series of side chains in the binding pocket are known to rotate in concert to assess how well BLUES works for such problem.

While NCMC is an exciting way to accelerate sampling of specific degrees of freedom known to be problematic within MD simulations, one remaining challenge is that, given the high cost of NCMC and low acceptance of side chain moves, more work may need to be done to improve acceptance of these move types and/or reduce the additional computational cost. One potential side benefit that remains to be explored, despite the challenge of low move acceptances, can be derived from the Jarzynski relationship, which relates differences in free energies between states to the irreversible work done in moving between them. Currently BLUES tracks the work done in turning on and off interactions between the side chain and the surrounding system, regardless of whether an NCMC move is accepted. The accumulation of this data could be used to readily identify favored or unfavored rotamer states in a system regardless of the rate of acceptance, via the Jarzynski nonequilibrium free energy relationship.[46]

As noted, further studies of lysozyme L99A binding may be needed to assess the impact of the long-timescale protein relaxation observed here and by others,[45] and it may

have implications for the diverse modeling studies of binding which have previously been conducted on this lysozyme binding site.

Supporting Information: Supplementary figures, run and analysis scripts for umbrella sampling, BLUES and MD simulations, as well as input files are provided in **Appendix A**.

References

1. Michel, J.; Essex, J. W. Hit Identification and Binding Mode Predictions by Rigorous Free Energy Simulations. *J. Med. Chem.* 2008, *51*, 6654–6664.

2. Zeevaart, J. G.; Wang, L.; Thakur, V. V.; Leung, C. S.; Tirado-Rives, J.; Bailey, C. M.; Domaoal, R. A.; Anderson, K. S.; Jorgensen, W. L. Optimization of Azoles as Anti-Human Immunodeficiency Virus Agents Guided by Free-Energy Calculations. *J. Am. Chem. Soc.* 2008, *130*, 9492–9499.

3. Chipot, C.; Rozanska, X.; Dixit, S. B. Can Free Energy Calculations Be Fast and Accurate at the Same Time? Binding of Low-Affinity, Non-Peptide Inhibitors to the SH2 Domain of the Src Protein. *JComput Aided Mol Des* 2005, *19*, 765–770.

4. Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* 2011, *133*, 9181–9183.

5. Mobley, D. L. Let's Get Honest about Sampling. *J Comput Aided Mol Des* 2012, *26*, 93–95.

6. Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *Journal of Molecular Biology* 2007, *371*, 1118–1134.

7. Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* 2006, *2*, 1255–1273.

8. Jiang, W.; Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* 2010, *6*, 2559–2565.

9. Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J. Chem. Theory Comput.* 2007, *3*, 1231–1235.

10. Lim, N. M.; Wang, L.; Abel, R.; Mobley, D. L. Sensitivity in Binding Free Energies Due to Protein Reorganization. *Journal of Chemical Theory and Computation* 2016, *12*, 4620–4631.

11. Gaudreault, F.; Chartier, M.; Najmanovich, R. Side-Chain Rotamer Changes upon Ligand Binding: Common, Crucial, Correlate with Entropy and Rearrange Hydrogen Bonding. *Bioinformatics* 2012, *28*, i423–i430.

12. Nilmeier, J. P.; Crooks, G. E.; Minh, D. D. L.; Chodera, J. D. Nonequilibrium Candidate Monte Carlo Is an Efficient Tool for Equilibrium Simulation. *PNAS* 2011, *108*, E1009–E1018.

13. Beutler, T. C.; Bremi, T.; Ernst, R. R.; van Gunsteren, W. F. Motion and Conformation of Side Chains in Peptides. A Comparison of 2D Umbrella-Sampling Molecular Dynamics and NMR Results. *J.Phys. Chem.* 1996, *100*, 2637–2645.

14. Wang, L.; Berne, B. J.; Friesner, R. A. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein–Ligand Binding Affinities. *PNAS* 2012, *109*, 1937–1942.

15. Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* 2018, *122*, 5579–5598.

16. Kurut, A.; Fonseca, R.; Boomsma, W. Driving Structural Transitions in Molecular Simulations Using the Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* 2018, *122*, 1195-1204.

17. OpeneEye Scientific Software, I. OEChem Toolkit. 2010. (accessed June 16, 2015).

18. Scouras, A. D.; Daggett, V. The Dynameomics Rotamer Library: Amino Acid Side Chain Conformations and Dynamics from Comprehensive Molecular Dynamics Simulations in Water. *Protein Sci.* 2011, *20*, 341–352.

19. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins: Structure, Function, and Bioinformatics* 2000, *40*, 389–408.

20. Hintze, B. J.; Lewis, S. M.; Richardson, J. S.; Richardson, D. C. Molprobity's Ultimate Rotamer-Library Distributions for Model Validation. *Proteins: Structure, Function, and Bioinformatics* 2016, *84*, 1177–1189.

21. Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* 2011, *19*, 844–858.

22. Jr, R. L. D.; Karplus, M. Conformational Analysis of the Backbone-Dependent Rotamer Preferences of Protein Side chains. *Nature Structural & Molecular Biology* 1994, *1*, 334–340.

23. Dunbrack, R. L.; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci* 1997, *6*, 1661–1681.

24. Ishizuka, R.; Huber, G. A.; McCammon, J. A. Solvation Effect on the Conformations of Alanine Dipeptide: Integral Equation Approach. *The Journal of Physical Chemistry Letters* 2010, *1*, 2279– 2283.

25. Drozdov, A. N.; Grossfield, A.; Pappu, R. V. Role of Solvent in Determining Conformational Preferences of Alanine Dipeptide in Water. *Journal of the American Chemical Society* 2004, *126*, 2574–2581.

26. Kalko, S. G.; Guàrdia, E.; Padró, J. A. Molecular Dynamics Simulation of the Hydration of the Alanine Dipeptide. *The Journal of Physical Chemistry B* 1999, *103*, 3935–3941.

27. Marrone, T. J.; Gilson, M. K.; McCammon, J. A. Comparison of Continuum and Explicit Models of Solvation: Potentials of Mean Force for Alanine Dipeptide. *The Journal of Physical Chemistry* 1996, *100*, 1439–1441.

28. Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-Time Conformational Transitions of Alanine Dipeptide in Aqueous Solution: Continuous and Discrete-State Kinetic Models. *The Journal of Physical Chemistry B* 2004, *108*, 19487–19495.

29. Weise, C. F.; Weisshaar, J. C. Conformational Analysis of Alanine Dipeptide from Dipolar Couplings in a Water-Based Liquid Crystal. *The Journal of Physical Chemistry B* 2003, *107*, 3265–3277.

30. Gaigeot, M.-P. Unravelling the Conformational Dynamics of the Aqueous Alanine Dipeptide with First-Principle Molecular Dynamics. *The Journal of Physical Chemistry B* 2009, *113*, 10059–10062.

31. Cruz, V.; Ramos, J.; Martínez-Salazar, J. Water-Mediated Conformations of the Alanine Dipeptide as Revealed by Distributed Umbrella Sampling Simulations, Quantum Mechanics Based Calculations, and Experimental Data. *The Journal of Physical Chemistry B* 2011, *115*, 4880–4886.

32. Case, D. A. et al. AmberTools15. 2015.

33. 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*, 1950– 1958.

34. Eastman, P.; Swails, J.; Chodera, J. D., McGibbon,R.T.;Zhao, Y.;Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Computational Biology* 2017, *13*, e1005659.

35. Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* 2008, *129*, 124105.

36. Mobley, D. L.; Chodera, J. D.; Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of Chemical Theory and Computation* 2007, *3*, 1231–1235.

37. Bach, A. Boltzmann's Probability Distribution of 1877. *Archive for History of Exact Sciences* 1990, *41*, 1–40.

38. Mobley, D. L.; Chodera, J. D.; Dill, K. A. The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J Chem Theory Comput* 2007,*3*, 1231–1235.

39. Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* 1995, *34*, 8576–8588.

40. Morton,A.; Baase,W.A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* 1995, *34*, 8564–8575.

41. Merski, M.; Fischer, M.; Balius, T. E.; Eidam, O.; Shoichet, B. K. Homologous Ligands Accommodated by Discrete Conformations of a Buried Cavity. *Proceedings of the National Academy of Sciences* 2015, *112*, 5039–5044.

42. Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A Cavity Containing Mutant of T4 Lysozyme Is Stabilized by Buried Benzene. *Nature* 1992, *355*, 371–373.

43. 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*, 235–242.

44. Xie, B.; Nguyen, T. H.; Minh, D. D. L. Absolute Binding Free Energies between T4 Lysozyme and 141 Small Molecules: Calculations Based on Multiple Rigid Receptor Configurations. *J. Chem. Theory Comput.* 2018, *13*, 2930-2944.

45. Schiffer, J. M.; Feher, V. A.; Malmstrom, R. D.; Sida, R.; Amaro, R. E. Capturing Invisible Motions in the Transition from Ground to Rare Excited States of T4 Lysozyme L99A. *Biophys. J.* 2016, *111*, 1631–1640.

46. Jarzynski, C. Equilibrium Free-Energy Differences from Nonequilibrium Measurements: A Master-Equation Approach. *Phys. Rev. E.* 1997, *56*, 5018–5035.

CHAPTER 3: The structure of a *Mycobacterium tuberculosis* heme-degrading protein, MhuD, variant in complex with its product

Abstract

Mycobacterium tuberculosis (Mtb), the causative agent of tuberculosis, requires iron for survival. In Mtb, MhuD is the cytosolic protein that degrades imported heme. MhuD is distinct, both in sequence and structure, from canonical heme oxygenases (HOs) but homologous with IsdG-type proteins. Canonical HO is found mainly in eukaryotes, while IsdG-type proteins are predominantly found in prokaryotes including pathogens. While there are several published structures of MhuD and other IsdG-type proteins in complex with heme substrate, no structures have been reported of IsdG-type proteins in complex with product, unlike HOs. We recently showed that the Mtb variant MhuD-R26S produces biliverdin IXα (αBV) rather than the wild-type mycobilin isomers as product. Given that mycobilin and other IsdG-type protein products like staphylobilin are difficult to isolate in quantities sufficient for structure determination, here we use the MhuD-R26S variant and its product αBV as a proxy to study the IsdG-type protein/product complex. First we show that αBV has nanomolar affinity for MhuD and the R26S variant. Second we determined the MhuD-R26S-αBV complex structure to 2.5 Å, which reveals two notable features (1) two αBV molecules bound per active site and (2) a novel α-helix (α3) unobserved in previous MhuD-heme structures. Finally, through molecular dynamics simulations we show that α3 is stable with the proximal αBV alone. With MhuD's high affinity for product along with observed structural and electrostatic changes that accompany substrate turnover suggest that there may be an unidentified class of proteins responsible for product extraction from MhuD and other IsdG-type proteins.

Introduction

Heme degradation is critical for a variety of biological functions, including iron re-utilization, cell signaling, and antioxidant defense[1-3]. The well-studied canonical heme oxygenase (HO), human HO (hHO-1), catalyzes the oxidative cleavage of heme to release biliverdin IXα (αBV, **Figure 3.1A**), ferrous iron, and carbon monoxide (CO)[4-6]. HO homologs are also present in eukaryotes and have been found in some prokaryotes including the pathogens: *Corynebacterium diphtheriae*, *Pseudomonas aeruginosa,* and *Neisseria meningitidis*[7-11]. In eukaryotes, the HO reaction is coupled with the conversion of biliverdin (BV) to bilirubin by biliverdin reductase (BVR)[12]. After conjugation of bilirubin with glucuronic acid, bilirubin is excreted[13]. The fate of HO-produced BV in prokaryotes has not been well studied thus far; although in *P. aeruginosa*, heme is degraded by a HO homolog, HemO, and the BV by-product is excreted, without further reduction, through an unknown[14].

Iron surface determinan*t* G (IsdG)-type proteins, found mainly in bacteria, comprise another heme-degrading protein family distinct from canonical HOs, in both sequence and structure[15-17]. *Staphylococcus aureus* IsdG and IsdI were the first members characterized[16,17], and instead of degrading heme to BV, iron, and CO, these enzymes cleave and oxidize heme at the β- and δ-meso carbon sites to produce staphylobilin isomers (**Figure 3.1A**), free iron, and formaldehyde[18,19]. Other IsdG-type enzymes include MhuD from *Mycobacterium tuberculosis* (Mtb) and LFO1 from eukaryotic *Chlamydomonas reinhardtii*, which also degrade heme into unique products[15,20]. While the product(s) of LFO1 heme degradation are yet-to-be-determined[20], MhuD degrades heme into iron and mycobilin isomers (**Figure 3.1A**)[21]. Like staphylobilins, the mycobilin isomers are also

oxidized at the β- or δ-meso carbons; however, cleavage occurs at the α-meso carbon with no observed loss of a C1-product[21]. The lack of C1-product may be physiologically important, as the CO by-product of hHO-1 triggers the transition of Mtb from its active to latent state[21]. Unlike HO-produced BV, the fate of the IsdG-type protein heme degradation tetrapyrrole products, like staphylobilin and mycobilin, is unknown; however, they may have antioxidant properties similar to BV[22].

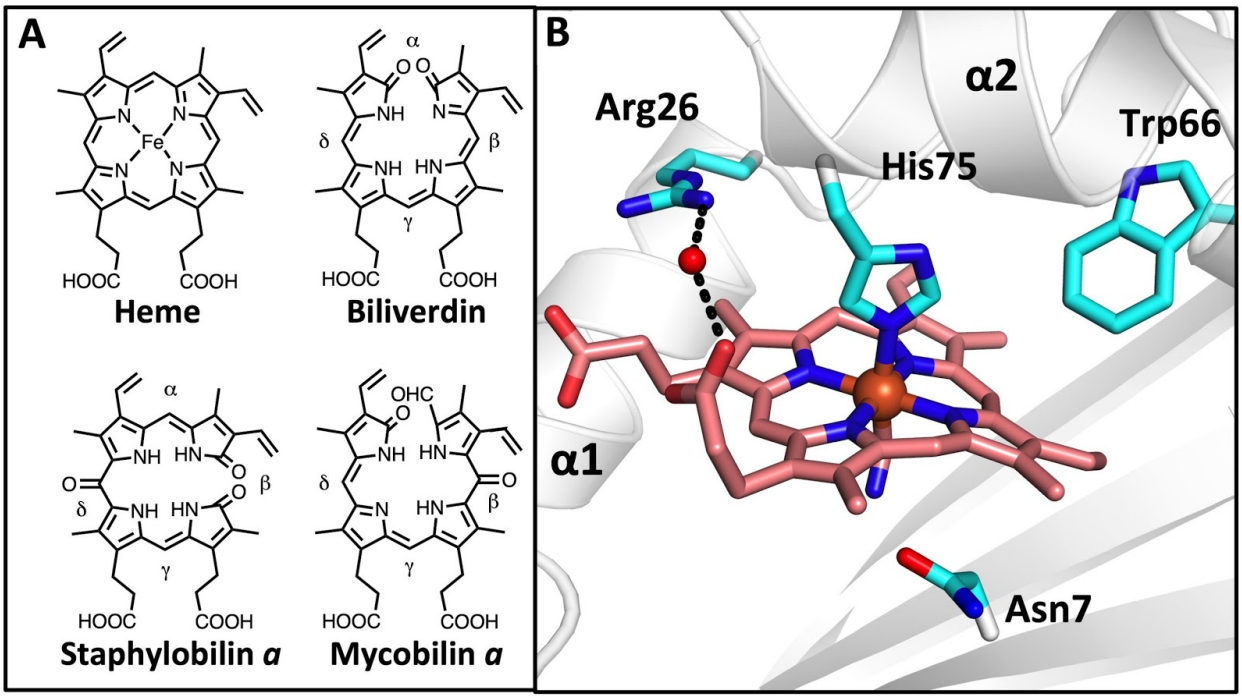


Figure 3.1 Structures of tetrapyrroles and WT-MhuD-monoheme. A. The structure of heme and heme tetrapyrrole degradation products. **B**. The structure of the active site of the Mtb heme-degrading protein, MhuD (cartoon, white), in its cyano-derivatized monoheme form (stick, pink). Depicted in stick representation (cyan) are essential Asn7, Trp66 and His75 (which coordinates heme-iron) residues, and Arg26 that forms a water-mediated H-bond with one of the heme propionates.

The structures of HO and IsdG-type proteins are distinct; HOs are predominately monomeric, comprised of a α-helical domain[23], while IsdG-type proteins consist of a homodimeric β-barrel decorated with two α-helices from each subunit[17]. Unsurprisingly,

the two distinct classes of heme-degrading enzymes utilize different mechanisms to degrade heme[24]. Although heme is coordinated by a proximal His residue in both HO and IsdG-type enzymes, the heme molecule in HOs is near-planar, while the heme is ruffled for IsdG-type proteins[17,23]. Furthermore, HOs have a distal pocket with a network of ordered water molecules, which facilitates the three consecutive monooxygenase steps required for heme degradation[6]. In contrast, the distal heme pocket is quite hydrophobic for IsdG-type proteins, with only one or two ordered waters observed in the proximal pocket[17,25]. In MhuD, it has been proposed that the hydrophobic pocket and the ruffled heme both contribute to the sequential monooxygenase and dioxygenase steps required for MhuD to degrade heme[24].

The structures of both human and *C. diphtheriae* HO complexed with αBV illustrate that the HO heme degradation reaction is coupled with a conformational change from a 'closed 'to 'open' state[26,27]. In both the eukaryotic and prokaryotic HO structures, the open-product bound state results from relaxation of the distal and proximal α-helices, and a rotation of the catalytic His side chain out of the active site pocket, as it is no longer coordinated to heme-iron[26,27]. This structural shift suggests that a degree of protein flexibility is necessary amongst the HO homologs to facilitate catalysis. Structures of apo and monoheme bound forms of IsdG-type proteins reveal an analogous conformational change in the catalytic His residue, as it is absent or disordered in the apo structures; however, the ordering of the elongated L2 loop region upon heme binding results in a much more drastic conformational shift compared with those of the HOs (apo-MhuD; Protein Data Bank (PDB) ID: 5UQ4)[17,25,28]. Furthermore unlike other IsdG-type proteins studied to date, MhuD is exceptionally flexible with its active site capable of accommodating two

molecules of heme, resulting in protein inactivation[15,25]. The biological significance of this conformational plasticity and its role in product turnover is not well understood as there is no structure of an IsdG-type enzyme in its product-bound form.

The structure of an IsdG-type protein in complex with its heme degradation product would further our understanding of the mechanism of action of this protein family. Unfortunately, staphylobilin and mycobilin products of IsdG-type proteins are difficult to purify[19,21], which has presumably hindered the structural determination of the product-bound form. Recently, we demonstrated that a MhuD variant, MhuD-Arg26Ser (**Figure 3.1B**), upon heme degradation produces αBV, formaldehyde, and iron[29]. In this current study, we determine the affinity of wild-type (WT) MhuD and the MhuD-R26S variant to both heme and αBV, and show they both bind heme and αBV in the nanomolar range. This high affinity to αBV allows the utilization of the MhuD-R26S variant as a proxy to study IsdG-type proteins in complex with product. Upon solving the crystal structure of the MhuD-R26S-αBV complex, we observed the formation of a novel secondary structural element unseen in the other heme-bound MhuD structures, and its implications will be discussed further.

Methods

Fluorescence-detection of ligand binding

Fluorescence-detected titrations of heme and αBV were carried out using a previously described protocol[30]. Stock solutions of MhuD (80 nM), heme (8 μM), and αBV (8 μM) were prepared in 50 mM Tris/HCl pH 7.4, 150 mM NaCl. Heme or αBV was titrated and gently mixed in 16 nM or 32 nM increments into MhuD-WT or MhuD-R26S. Following a 1-min incubation, fluorescence emission spectra were acquired between 320 to 500 nm on

a Hitachi F-4500 Fluorescence Spectrophotometer through excitation at 285 nm with the following parameters: 1/3 nm step size, scan speed of 240 nm/min, PMT voltage of 700 V, and slit widths at 10 nm (MhuD WT) and 20 nm (MhuD-R26S).

Fluorescence emission spectral analysis

Results from the fluorescence-based assay were fit to Eqn. 1 derived from Conger et el [30, 31], to determine the equilibrium dissociation-constant (K_d) of heme or αBV with MhuD and its mutants.

Eqn. 1:

$$F = \frac{[\text{MhuD}] + [\text{ligand}] + K_\text{d} - \sqrt{([\text{MhuD}] + [\text{ligand}] + K_\text{d})^2 - 4[\text{MhuD}][\text{heme}]}}{2} \times \left(\frac{F_\text{min} - F_\text{max}}{[\text{MhuD}]}\right) + F_\text{max}$$

where [MhuD] is the total concentration of MhuD or mutant MhuD, [ligand] is the total concentration of heme or αBV, F_{max} is the emission intensity without ligand, and F_{min} is the emission intensity for fully ligand-bound MhuD. Fitting of the fluorescence emission intensity at 340 nm for K_d determination was performed using Origin 2018.

Expression and purification of Mtb MhuD and the R26S variant

WT MhuD and the R26S variant were purified as previously reported[15,29]. In brief, *E. coli* B21-Gold (DE3) cells transformed with pET22b-MhuD plasmid were grown in LB medium containing 50 μg/mL ampicillin at 37°C. Overexpression was induced at OD_{600} of ~ 0.6 using 1 mM IPTG. The cells were harvested 4 hours post induction and resuspended in lysis buffer (50 mM Tris/HCl pH 7.4, 350 mM NaCl and 10 mM imidazole). Cells were lysed via sonication and the resulting lysate was centrifuged at 14,000 rpm. The cell supernatant was loaded onto a Ni^{2+}-charged HiTrap chelating column (5 mL) and washed with lysis buffer. Bound protein was eluted from the column with increasing concentrations of

imidazole. Next, MhuD, which elutes at 50 and 100 mM imidazole, was concentrated (Amicon, 5 kDa molecular mass cutoff) and was further purified on a S75 gel filtration column in 20 mM Tris/HCl pH 8, and 10 mM NaCl. A final purification step was performed on an ion-exchange column (HiTrap Q HP, 5 mL) with MhuD eluting at 150 mM NaCl.

Crystallization of MhuD-R26S-αBV complex

To prepare an αBV solution, approximately 2 mg of αBV hydrochloride (SigmaAldrich) was dissolved in 500 µL 0.1 M NaOH followed by 500 µL 1 M Tris/HCl pH 7.4 before dilution into 50 mM Tris/HCl pH 7.4, 150 mM NaCl. A ferric chloride solution was prepared by dissolving 27.3 mg of ferric chloride hexahydrate (SigmaAldrich) in 10 mL water. A 1.3 fold excess of a 1:1 molar ratio solution of αBV and ferric chloride was gradually added to 100 µM apo-MhuD R26S and incubated overnight at 4°C before being concentrated to 10 mg/mL (Lowry assay)[32]. Light blue crystals appeared in 0.1 M HEPES pH 6.5, 4.6 M NaCl, 30 mM glycyl-glycyl-glycine after 2 days. The crystals were flash frozen in 100% NVH oil and a data set to 1.9 Å was collected at 70K. The collected data was indexed, integrated, and reduced using iMOSFLM[33]. Due to the extremely high degree of anisotropy at higher resolution in the a^* direction (http://services.mbi.ucla.edu/anisoscale), the data was cut off at 2.5 Å. Initial phase determination was carried out using Phaser in the PHENIX suite[34] using WT MhuD-heme-CN structure as a search model (PDB ID 4NL5)[25]. αBV molecules were positioned into appropriate positive electron density in the vicinity of the active site, and the structure was refined using phenix.refine [34]. For each monomer, the electron density of the loop region between Ala24 to Asn32 is poorly defined and residues His25-Val30 were modeled as

alanines. The only Ramanchandran outliers are Val30 modeled as Ala in both Chain A and B, which are in this poorly defined region of electron density.

Molecular dynamics (MD) simulations

Input files for the MhuD^{4NL5}-heme and MhuD^{4NL5}-αBV simulations were prepared using the structure of dimeric MhuD-heme-CN with the cyano groups removed (PDB ID: 4NL5). For the MhuD-αBV complex, input files were prepared from the MhuD-R26S-αBV structure, wherein the R26S mutation was reversed and just one αBV (proximal) was retained per active site. All crystallographic waters were removed with the exception of the ordered waters in the active site of MhuD^{4NL5} (HOH numbers: 313, 325, 349). Hydrogen atoms were added to the system using pdb4amber from AmberTools15 with default protonation states[35]. For simulations of MhuD^{4NL5}-αBV, each αBV was manually docked to mimic the approximate orientation and position of the proximal αBV from the MhuD-R26S-αBV structure. αBV was parameterized using antechamber from AmberTools15 [35] with GAFF version 1.7 and AM1-BCC charges. The ligated His75-Fe-heme ligand was parameterized according to previously published Density Functional Theory calculations[36]. Missing atoms for each structure were added using tleap in AmberTools15[35], and parameterized using ff99SBildn[37]. Each system was explicitly solvated in tleap with a 10 Å rectangular box of TIP3P water, extending from the surface of the protein to the box edge, and sodium ions were added to neutralize the charge of the system.

Minimization proceeded using *sander* from Amber14[35] with steepest descents running for 20,000 steps, followed by heating from 100 K to 300 K with constant volume for 25,000 timesteps of 2 fs. Equilibration continued using *sander* for 500,000 timesteps of

2 fs under constant pressure with positional restraints initially applied on all non-water atoms at 50 kcal/mol/Å^2 and progressively lifted in increments of 5 kcal/mol/Å^2 over ten 50,000 step segments. The resulting topology and coordinate files for each system were used as inputs for MD simulations. Production simulations were executed in OpenMM 7.1.0[38] using a Langevin integrator with a 2 fs timestep and a friction coefficient of 10/ps. For each of the three systems, five independent simulations of 100 ns each were initiated with randomized velocities. Among the five simulations, one simulation was restarted for each system for an additional 500 ns, bringing the total simulation time to 1 μs/system.

MD analysis

Distances and secondary structure assignments were computed using MDTraj [39]. For analysis, each monomer of MhuD was treated independently with the assumption that long-range interactions between each subunit are negligible for the time scales simulated here. To analyze the orientation of His75 during simulations, two orientations were defined: 1) *Active Site* – side chain pointing into binding pocket, towards the center of the tetrapyrrole, and 2) *Solvent Exposed* – side chain flipped away from the binding pocket, as in the MhuD-αBV structure (see **Figure 3.4Bi**). To assign these positions, we computed the distance between the ε nitrogen atom (furthest from the backbone) on His75 and the nitrogen atom located between the α and β carbons on either αBV or heme. If the distance was less than 6 Å, the His75 residue was classified as being oriented in the *Active Site* position; otherwise the position was classified as *Solvent Exposed*.

For the Arg79 positional analysis, we identified four positions; 1) *Active site* – Arg79 directed into the binding pocket interacting with the ligand, 2) *Helix 1* – Arg79 side chain interacting with residues on α-helix-1 (i.e residues 16-25) of MhuD, 3) *Helix 2* – Arg79 side chain interacting with residues on α-helix-2 (residues 60-75), and 4) *Solvent Exposed* – Arg79 side chain oriented into the surrounding solvent. To analyze the Arg79 in our simulations, position assignments were defined as follows: 1) *Active Site* – guanidinium carbon of Arg79 side chain (CZ) within 4.5 Å of either terminal carbon on the propionate groups of the heme or αBV ligands. 2) *Helix 1* - Arg79 CZ atom within 6 Å of Glu16 α carbon 3) *Helix 2* – Arg79 CZ atom within 6 Å of His75 or Ile72 backbone oxygen and more than 4.5 Å from terminal carbons on the propionate groups of the heme or αBV ligands. 4) *Solvent Exposed* – Any position of Arg79 falling outside the description of positions 1-3.

Results

Affinity of substrate and product to MhuD

Previously, the affinity (K_d) of WT-MhuD for heme was measured to be ~6 nM[30]; empirically it has been observed that MhuD also has a high affinity for its product, as isolation of MhuD mycobilin products requires iron chelation followed by protein denaturation[21,29]. Recently, we demonstrated that the MhuD-R26S variant degrades heme

Table 3.1 Heme and αBV binding affinities (K_d) were determined for WT MhuD and the MhuD-R26S variant with 1:1 heme or αBV to protein ratio. Each experiment was performed in triplicate.

	heme K_d (nM)	αBV K_d (nM)
MhuD-WT	6.0 ± 2.9	36.4 ± 2.1
MhuD-R26S	9.5 ± 3.0	92.4 ± 6.8

to produce αBV as its predominate tetrapyrrole product rather than the WT mycobilin isomers[21,29]. As it is difficult to isolate large quantities of mycobilin, in this study the MhuD-R26S variant and its αBV product is utilized as a model system to study IsdG-type proteins in complex with product. First, we determined the affinities of heme and αBV to both WT-MhuD and the MhuD-R26S variant using a previously described fluorescence-based assay

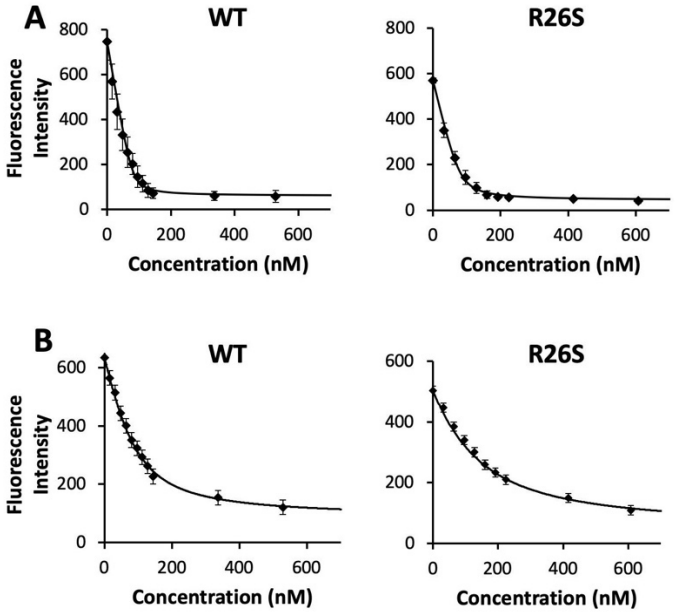


Figure 3.2 Heme and αBV affinity of MhuD and the R26S variant. Representative fluorescent emission intensities at 340 nm after excitation at 280 nm of WT-MhuD (left panels) and the MhuD-R26S variant (right panels) with increasing concentrations of **A.** heme and **B.** αBV.

(**Figure 3.2**)[30]. The affinities of WT-MhuD and MhuD-R26S for heme are ~6nM and ~9nM, respectively (**Table 3.1)**. The reduced affinity of heme to the MhuD-R26S variant compared to WT-MhuD may be due to the loss of the water-mediated interaction of Arg26 with the heme propionate (**Figure 3.1B**) when Arg26 is mutated to Ser. By comparison, both WT-MhuD and MhuD-R26S have weaker affinity for αBV of ~36 nM and ~92 nM, respectively (**Table 3.1**). However, we observe the same trend as for heme, where WT-MhuD binds αBV with a higher affinity than the MhuD-R26S variant (even though the WT product of MhuD is

mycobilin rather than αBV), which suggests that the Arg26 residue may form contacts with the tetrapyrrole product.

The structure of the MhuD-R26S-αBV complex

To investigate the structural impacts on MhuD arising from heme degradation into product, we turned our attention to structure determination of MhuD-R26S in complex with αBV. The MhuD-R26S-αBV complex crystallized in the presence of ferric chloride, and light blue crystals appeared in 0.1 HEPES pH 6.5, 4.6 M NaCl and 30 mM glycyl-glycyl-glycine after 2 days. The structure of MhuD-R26S-αBV was solved to 2.5 Å resolution with a final R/R_{free} of 23.2/28.5 (**Table 3.2**). The asymmetric unit contains five stacked αBV molecules that link individual subunits from two adjacent, biologically relevant homodimers (**Figure 3.3A).** The monomers in the asymmetric unit, each in complex with two αBV molecules, superimpose with a root-mean-square deviation (RMSD) of 0.2 Å, and the five stacked αBV molecules model well into the electron density (**Figure B.1**). The biologically relevant MhuD-R26S homodimeric structure is observed in the crystallographic 2-fold symmetry (**Figure 3.3B**), similar to apo- and heme-bound MhuD homodimer structures (apo-MhuD; PDB code: 5UQ4)[15,25]. By size exclusion chromatography, the MhuD-R26S-αBV complex was confirmed to be a dimer in solution, **Figure B.2**.

The MhuD-R26S-αBV homodimeric structure is similar to that of the MhuD-heme-CN and MhuD-diheme structures[15,25], where each subunit forms a ferredoxin-like fold. To form the homodimer, each polypeptide chain donates four β-strands to form the eight-stranded antiparallel β-barrel. Additionally, each monomer has three α-helices together

with two flexible loop regions; the first loop connects α-helix-1 to β-strand-2 and the second connects α-helix-3 to β-strand-4 (**Figure 3.3B**). It should be noted that the overall

Table 3.2 Statistics for X-ray diffraction data collection and atomic refinement for the MhuD-R26S-αBV complex.

Space Group	I222
Unit cell dimensions (Å)	37.22 x 113.61 x 113.65
pH of crystallization condition	6.5
Protein concentration (mg/mL)	20
Data Collection	
Wavelength, Å	1.0
Resolution range, Å	35.94 – 2.5[#]
Unique reflections (total)	8717 (120788)
Completeness, %*	100.00 (100.00)
Redundancy*	13.9 (14.4)
R_{merge}*, †	0.092 (0.592)
I/σ*	22.5 (8.9)
NCS copies	2
Model refinement	
Resolution range, Å	35.93 - 2.5
No. of reflections (working/free)	8707 (867)
No. of protein + ligand atoms	1430
No. of water molecules	16
No. of αBVs in NCS	5
R_{work}/R_{free}, %¶	23.16/28.53
Rms deviations	
Bond lengths, Å	0.009
Bond angles, °	1.52
Ramachandran plot	
Most favorable region, %	90.96
Additional allowed region, %	7.98
	1.06
Outliers (Val30), %	5J43
PDB ID code	6PLE

[#]The data was cut off at 2.5 Å due to extreme anisotropy at high resolution.
*Statistics for the highest-resolution shell are given in parentheses.
† $R_{merge} = \Sigma_{hkl}\Sigma_i |I_i(hkl) - (I(hkl))| / \Sigma_{hkl}\Sigma_i I_i(hkl)$
¶ $R_{work} = \Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$. R_{free} was computed identically except where all reflections belong to a test set of 10% randomly selected data.

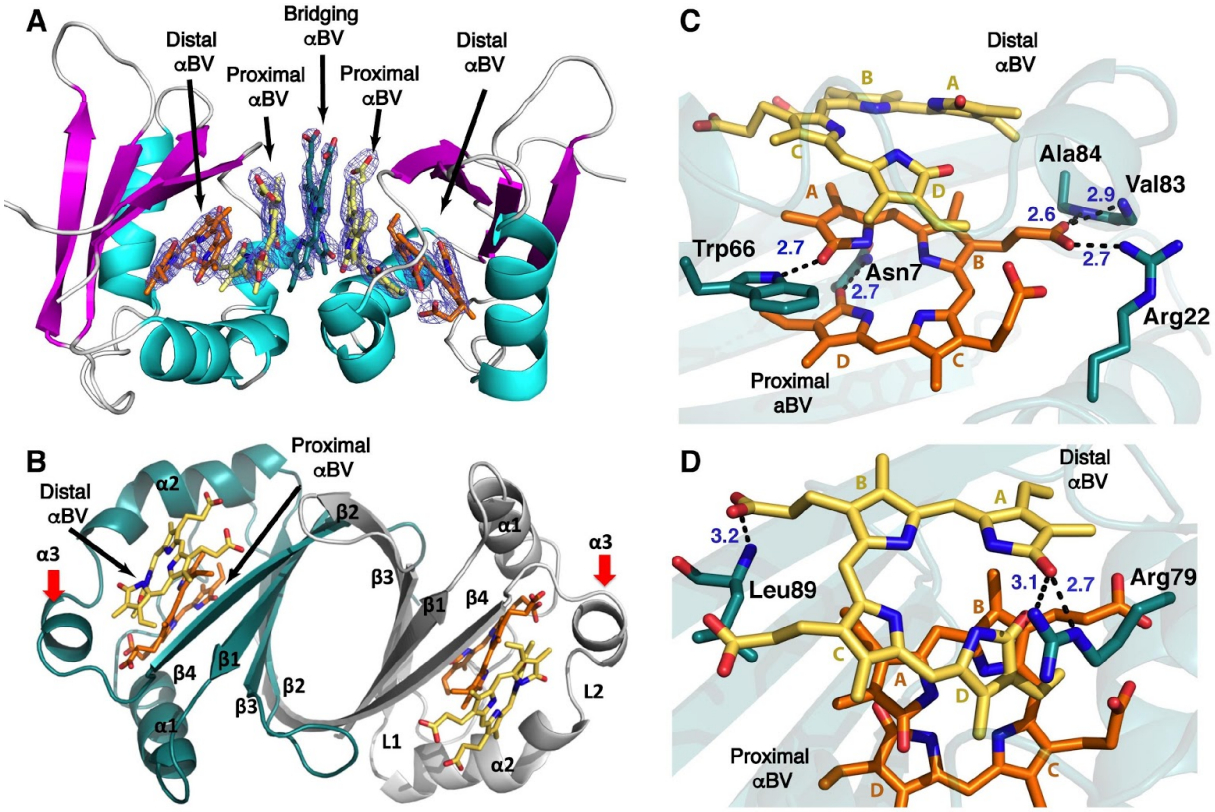


Figure 3.3 Structure of the MhuD-R26S-αBV complex. A. The asymmetric unit that contains five molecules of αBV stacked connecting the active sites of two MhuD monomers. **B.** The dimeric biological assembly with two molecules of αBV per active site. The new structural helix, α3, is denoted with a red arrow. **C.** Interactions of the proximal αBV (orange) with MhuD (green). Blue dashed lines represent H-bonds with their length in Å. **D.** Interactions of the distal αBV (yellow) with MhuD (green).

topology of the MhuD-R26S-αBV monomer differs from those of the apo-MhuD and MhuD-heme-CN structures[25]. Within the MhuD-R26S-αBV structure, there is an additional short α-helix-3 formed by residues Ala76-Asn81, which was not observed in the monoheme, diheme and apo-MhuD structures[15,25], and thus the latter structures have a slightly longer extended L2 loop region. There are also other subtle conformational changes that will be discussed later.

αBV Binding to MhuD

Each MhuD-R26S active site binds two molecules of αBV (**Figures 3.3B**). The solvent-protected or proximal αBV interacts with both MhuD and the distal αBV, and the distal αBV also interacts with MhuD. The proximal αBV forms five electrostatic and eleven hydrophobic interactions with MhuD residues. The αBV propionate-6 carboxylate group hydrogen-bonds (H-bonds) with the backbone amides of Val83 and Ala84 (2.9 and 2.6 Å, respectively), and forms a salt-bridge with the NH1 group of Arg22 (2.7 Å). Additionally, Trp66 NE1 H-bonds (2.7 Å) with the A-ring lactam oxygen (=O), and Asn7 ND2 H-bonds (2.7 Å) to the D-ring lactam oxygen (**Figure 3.3C**). The tetrapyrrole plane of the distal αBV is nearly parallel to that of the proximal αBV.

The proximal and distal αBV molecules interact through hydrophobic and van der Waals interactions only. The distal αBV is rotated ~205° and flipped ~180° with respect to the αBV tetrapyrrole plane, and the distal αBV tetrapyrrole plane is positioned ~3.4 Å above and translated ~5.0 Å relative to the proximal αBV plane. Consequently, the C-ring of the distal αBV is partially placed over the A-ring of the proximal αBV resulting in the distal αBV being more solvent exposed than the proximal one (**Figure 3.3D**). The proximal αBV forms five H-bonds with MhuD whereas the distal αBV forms only three and has fewer hydrophobic interactions with MhuD than the proximal αBV (**Figure 3.3D**). The A-ring lactam oxygen of the distal αBV H-bonds with Arg79 NE and NH1 (3.1 and 2.7 Å, respectively), and the distal αBV propionate-6 H-bonds to the Leu89 amide group (3.2 Å).

Along with the proximal and distal αBV subunits, there is a third 'bridging' αBV, which connects the two symmetrical subunits each bound to two αBV molecules (**Figure 3.3A**). The third 'bridging' αBV is parallel to the tetrapyrrole plane of the flanking distal

αBV molecule from each subunit and is separated by ~3.8 Å from each distal αBV. Additionally, the bridging αBV is rotated ~90° and translated ~2.5 Å with respect to the distal αBV molecules. The bridging αBV interacts with the distal αBVs by hydrophobic and van der Waals interactions, and also forms one H-bond to each, where the bridging αBV ring-D/A lactam oxygen H-bonds to the ring-A/D pyrrole nitrogen of the distal αBVs (2.8 Å and 3.3 Å), respectively. Finally, the bridging αBV forms a H-bond to each MhuD subunit, ring-A/D lactam oxygen H-bonds with Arg79 NH2 group on respective subunits (3.0 and 2.6 Å, respectively).

Comparison of MhuD substrate and product bound structures

There are several distinct structural differences between the inactive substrate-bound (MhuD-heme-CN) [25] and product-bound (MhuD-R26S-αBV) MhuD structures, despite an RMSD of 1.1 Å (**Figure 3.4A**). The most notable difference between the two structures is within α-helix-2 and the sequential loop region L2. In MhuD–heme–CN, the α-helix-2 is kinked after residue Asn68 and terminates at His75, where this region in the MhuD-R26S-αBV structure has a looser helical geometry. This kinked helical region in the MhuD-heme-CN structure positions the catalytic His75 within the active site so it coordinates with heme-iron. However, in MhuD-R26S-αBV, His75 flips out of the active site (90° rotation and translation of 2.8 Å of the Cα, **Figure 3.4Bi**) and is stabilized by a H-bond between the His75 imidazole nitrogen to the backbone carbonyl of Ala27 (3.2 Å). In combination with the slight unraveling of the C-terminus of α-helix-2 in MhuD-R26S-αBV, the extended loop L2 region has an additional α-helix (α3, Ala76-Asn81) as compared to MhuD-heme-CN [25] (**Figure 3.4A**). Within the L2 loop region of MhuD-heme-CN, both His78 and Arg79 are solvent-exposed with no observable electron density for the Arg79 side

chain. In contrast, in MhuD-R26S-αBV, Arg79 is located in α-helix-3, and its side chain is positioned towards the active site and is stabilized by a H-bond between its NH1 group with the backbone carbonyl of Ile72 (3.1 Å) (**Figure 3.4Bii**); notably His78 is still solvent-exposed and weakly stabilized by a cation-π interaction between its imidazole side chain and Arg22 (3.7 Å).

Another minor structural difference between substrate and product bound MhuD complexes is the unraveling of the C-terminal α-helix-1 in MhuD-R26S-αBV compared to MhuD-heme-CN. Arg26 is situated in this location and participates in a water-mediated H-bond with one of the heme propionates in the MhuD-heme-CN structure [25] (**Figures 3.1B & 3.4A**). Within the MhuD-R26S-αBV structure, this region is no longer helical. Thus, this observed unraveling of α-helix-1 may be the result of the Arg26Ser mutation, however, this conformational change ensures that the backbone carbonyl group of the subsequent residue, Ala27, is in H-bonding distance of the imidazole nitrogen of His75, to stabilize the flipped out His75 in the MhuD-R26S-αBV structure [21].

The conformational change in going from substrate to product bound also alters the active site pocket volume and the electrostatic potential of the molecular surface (**Figures 3.4C & 3.4D**). The active site volume increases dramatically from heme bound to that of αBV, from 188 to 354 $Å^3$, calculated utilizing CASTp [40]. In conjunction with the increased active site volume, the molecular surface region surrounding the active site and the L2 loop region undergoes an electrostatic potential change. The MhuD molecular surface surrounding one side of the exposed active site is positively and negatively charged in the presence of heme (**Figure 3.4C**), whereas in the presence of αBV it becomes more hydrophobic and negatively charged with a positively charged bridge capping the active

site (**Figure 3.4D**). Furthermore, when rotated 90° about the y-axis, there is altered molecular surface electrostatics from positively charged and hydrophobic to predominately negatively charged in the presence of heme and αBV, respectively (**Figures 3.4C & 3.4D**, right panels**)**.

MD Simulations

MD simulations were performed to gain further insight into the protein conformational changes associated with heme and αBV binding. In particular, we wanted to evaluate the *in silico* stability of the MhuD α-helix-3 (Ala76-Asn81) in the presence of proximal αBV alone. MD simulations were set up using the biological dimer of the MhuD-R26S-αBV structure with the R26S mutation modeled as WT Arg26 and only the proximal αBVs retained. Heretofore, we refer to this system in our MD analysis as MhuD-αBV. Over the 1 μs of combined simulation time, including a continuous 600 ns simulation, the α-helix-3 persisted in MhuD-αBV for over 70% of the run, suggesting that the α-helix-3 is stable when there is just one αBV molecule present per active site. Interestingly, we also observe some helix formation within the L1 loop region, an otherwise highly flexible region of the protein (**Figure 3.5)**.

To test if the α-helix-3 forms during turnover from heme to product, two more sets of simulations were carried out. The first contained the MhuD-heme-CN structure without the cyano group (MhuD^{4NL5}-heme) while the second was comprised of the MhuD^{4NL5} protein structure with αBV docked in place of heme to give MhuD^{4NL5}-αBV. For this system, the αBVs were manually docked in the orientation and position of the proximal αBV from the MhuD-R26S-αBV structure. As with the MhuD-αBV simulations, the MhuD^{4NL5}-αBV and MhuD^{4NL5}-heme simulations each ran for a total of 1 μs. Although the two αBV MD systems

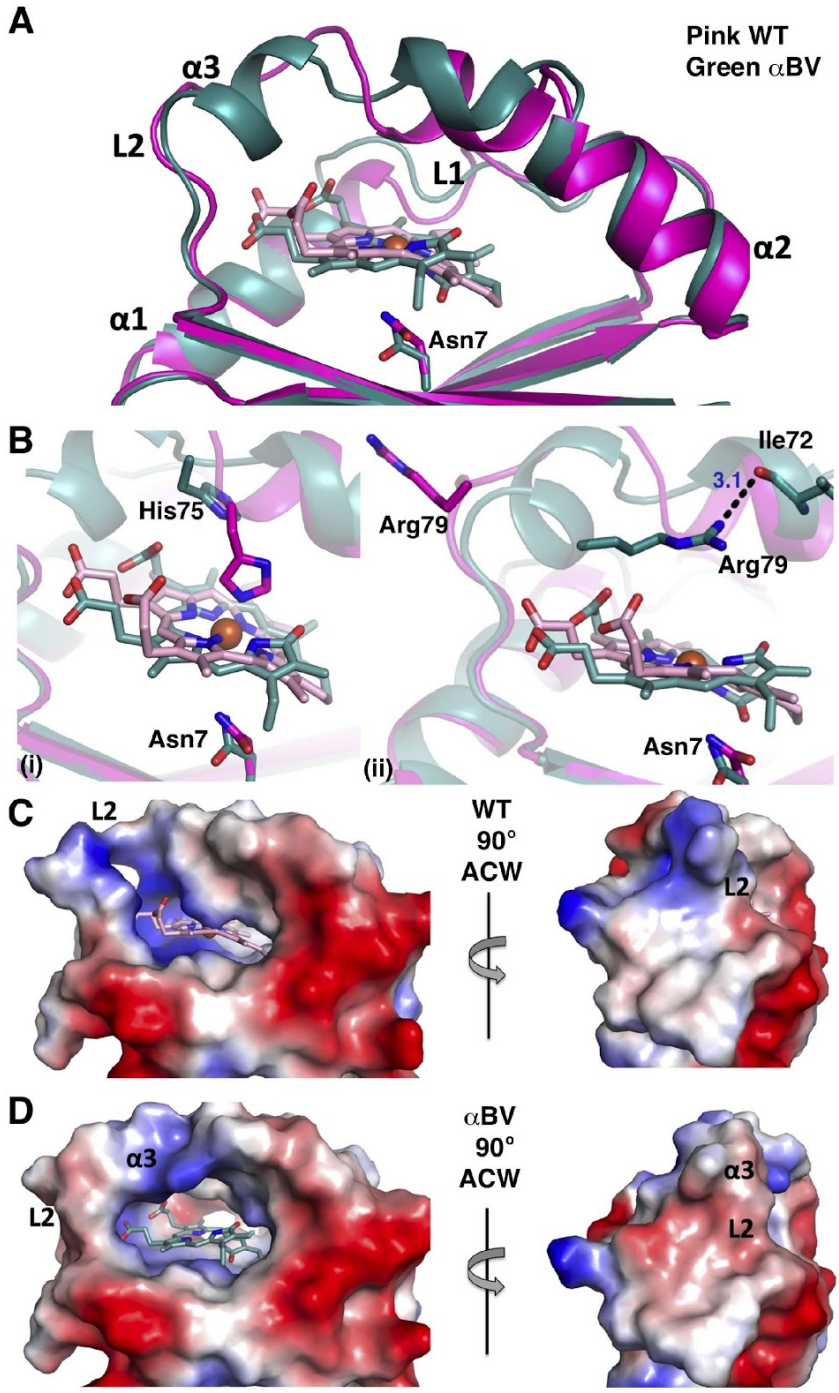


Figure 3.4 Structural comparison of MhuD-heme-CN and MhuD-R26S-αBV. A. Superposition of the MhuD-heme-CN (pink, PDB ID 4NL5, the cyano group is omitted for clarity) and MhuD-R26S-αBV (green). **B**. Active site comparison with **(i)** catalytic residues Asn7 and His75, and **(ii)** residues Asn7 and Arg79, are shown in stick representation. Black dashed lines represent H-bonds with their length in Å. **C-D** Electrostatic molecular surface representations, where blue and red are positively and negatively charged surfaces, respectively. Right panel is the left panel rotated 90° anticlockwise (ACW). **C.** MhuD-heme-CN (WT) and **D.** MhuD-R26S-αBV (αBV).

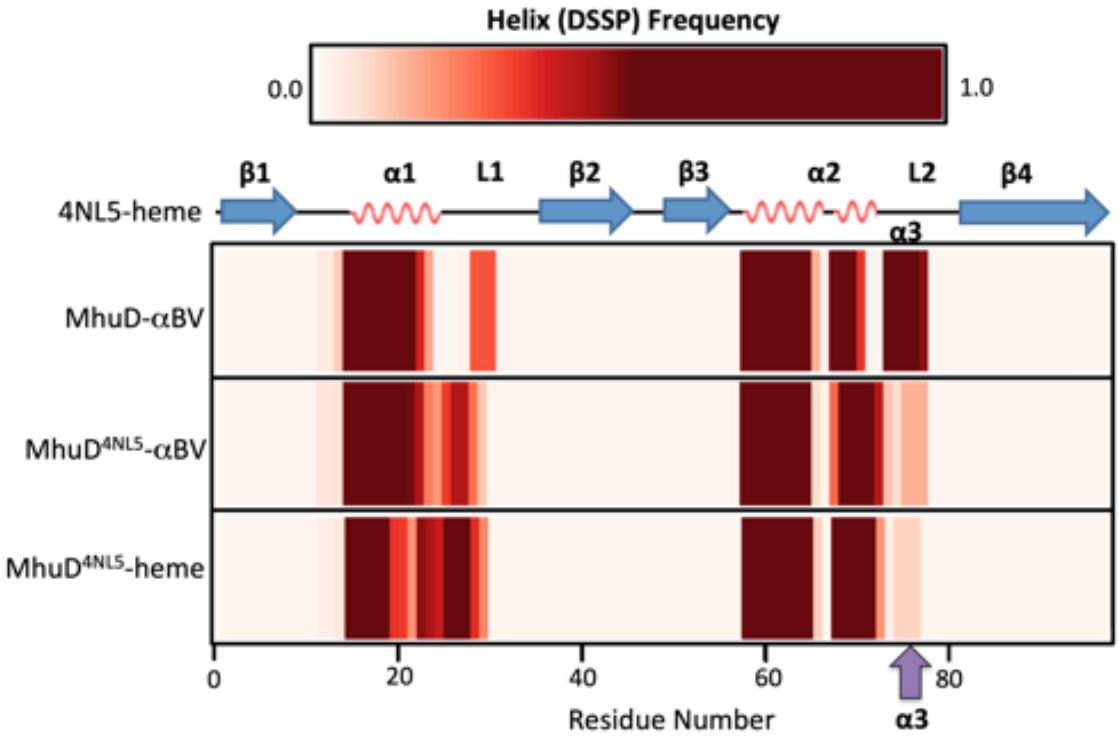


Figure 3.5 Helical stability in MD simulations of MhuD. The frequency of helix formation (DSSP) is plotted for each residue of MhuD in simulations with αBV and heme. For simulations of MhuD-αBV initiated from the coordinates of the MhuD-αBV structure, the novel α-helix (α3) that forms in L2 persists. In simulations of MhuD^{4NL5}-αBV initiated from the coordinates of the MhuD-monoheme structure (PDB: 4NL5), α3 transiently forms in the L2 region. In the presence of heme (MhuD^{4NL5}-heme), formation of α3 also is observed but less frequently than in MhuD^{4NL5}-αBV.

(MhuD-αBV and MhuD^{4NL5}-αBV) are identical in composition, they have distinct initial positions and velocities; MhuD-αBV simulations start from the coordinates of the αBV-bound crystal structure while MhuD^{4NL5}-αBV simulations start from those of the heme-bound structure. Given sufficient simulation time, the dynamics of these two systems should eventually converge. While we did not reach convergence, it is compelling that we see some *de novo* formation of the α-helix-3 in our simulation of MhuD^{4NL5}-αBV (**Figure 3.5**), suggesting that it is a relevant structural motif that forms in the presence of αBV. We speculate that the α-helix-3 may be further stabilized by contact of MhuD with accessory proteins. This α-helix-3 is also transiently observed in the MhuD^{4NL5}-heme simulation (**Figure 3.5)**; however, its occurrence is less frequent compared to MhuD^{4NL5}-αBV. In

addition, the binding mode of the modeled αBV, in both the MhuD-αBV and MhuD^{4NL5}-αBV simulations, is stable and consistent with the hypothesis that the proximal αBV in the MhuD-R26S-αBV structure mimics the orientation of the WT MhuD product.

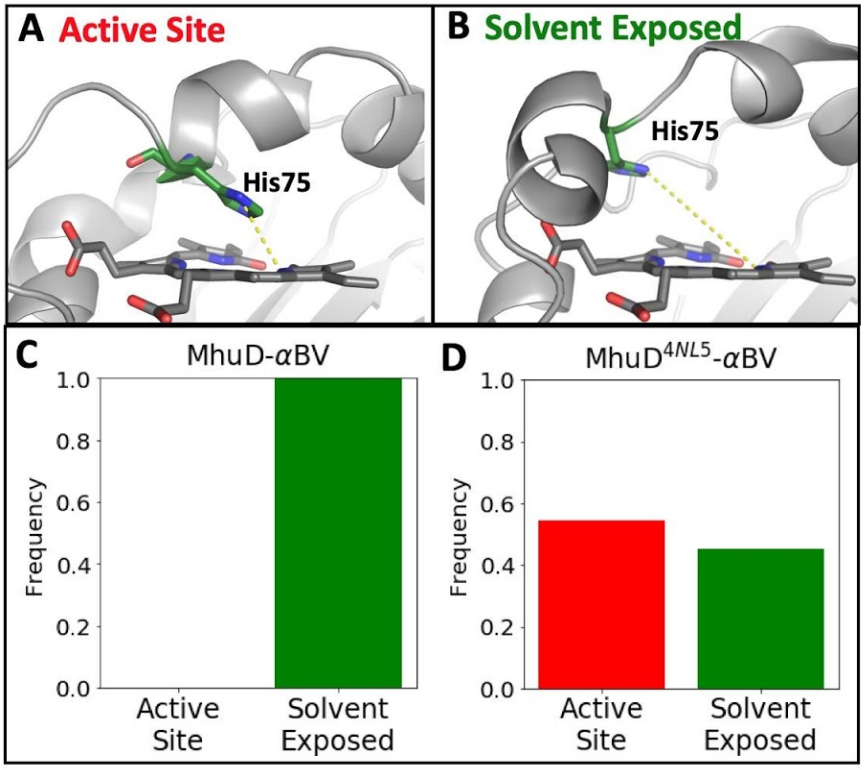


Figure 3.6 Orientation of His75 in MD simulations of MhuD. During MD simulations of MhuD with αBV, the position of His75 can be classified as either a) directed into the active site or b) rotated out of the active site. This designation was assigned based on the distance between the His75 ε nitrogen atom and one of the αBV nitrogens, as described in the methods section. The structure in **A.** is a snapshot from the MhuD^{4NL5}-αBV simulation while **B.** shows a snapshot of the MhuD-αBV simulation. The frequency of each position is plotted for the MhuD-αBV simulations initiated from **C.** the MhuD-R26S-αBV crystal structure and **D.** the MhuD^{4NL5}-αBV simulations, where αBV has docked in place of heme in the MhuD-monoheme structure. For the MhuD-αBV simulations, His75 remains oriented out of the active site while in the MD of MhuD^{4NL5}-αBV, His75 flips in and out of the active site. Notably, data for MhuD^{4NL5}-heme simulations of the MhuD-monoheme structure is not shown as the His75 residue is ligated to the heme-iron and thus cannot explore alternate positions.

We also sought to evaluate whether the 90° rotation of the catalytic His75 side chain away from the active site, as observed in the MhuD-αBV structure (**Figures 3.4Bi & 3.6**), is facilitated by substrate turnover to αBV. In the MhuD-αBV simulations, His75 residue

remains in the "flipped out" or solvent exposed orientation (**Figure 3.6B**) while in the MhuD^{4NL5}-heme simulations, it is ligated to the heme-iron atom and thus remains tethered to the active site. For MhuD^{4NL5}-αBV, the His75 alternates between the two orientations (**Figure 3.6C**), as shown in **Figures 3.6A & 3.6B**. The solvent exposed position may be further stabilized in the MhuD^{4NL5}-αBV simulations after extended simulation time and upon full formation of the α-helix-3, as observed in the MhuD-αBV structure, where the Arg79 side chain forms a hydrogen bond with the backbone carbonyl of Ile72 (**Figure 3.4Bii**).

Because the Arg79 side chain is unresolved in the MhuD^{4NL5}-heme structure but stabilized in the MhuD-αBV structure, we were also interested in exploring its dynamics in the presence of heme versus αBV. Upon inspection of the MhuD^{4NL5}-heme simulations, four classifications of Arg79 positions were identified: helix 1 and helix 2 (interacting with residues on α-helix-1 or on α-helix-2), active site (interacting with heme or αBV) and solvent exposed (**Figures 3.7A-D**). In the MhuD^{4NL5}-heme simulations, the varied distribution of the Arg79 positions mirrors its disorder in the crystal structure (**Figure 3.7E**). Unexpectedly, we found at times that the Arg79 forms H-bonds with the propionate groups of the heme; in fact, in the extended 600ns simulation of MhuD^{4NL5}-heme, the Arg79 residue for one MhuD subunit flipped into the active site and remained coordinated to the heme ligand for over 90% of the simulation (data not shown). In simulations with αBV, Arg79 interacts with residues on helix 2 or becomes solvent exposed. Consistent with the MhuD-αBV structure, the predominant state in the MhuD-αBV simulations shows it interacting with helix 2 (**Figure 3.7F)**. In our MhuD^{4NL5}-αBV simulation, the favored state is solvent exposed (**Figure 3.7G)**, but we suspect the interaction of Arg79 with helix 2 may be

further stabilized upon full formation of α-helix-3. Given that we only observe Arg79 interacting with the ligand propionate groups in simulations where His75 is oriented into the active site, we hypothesize that Arg79 plays a critical role in promoting catalysis (when heme is present) and facilitating product egress (via formation of α-helix-3).

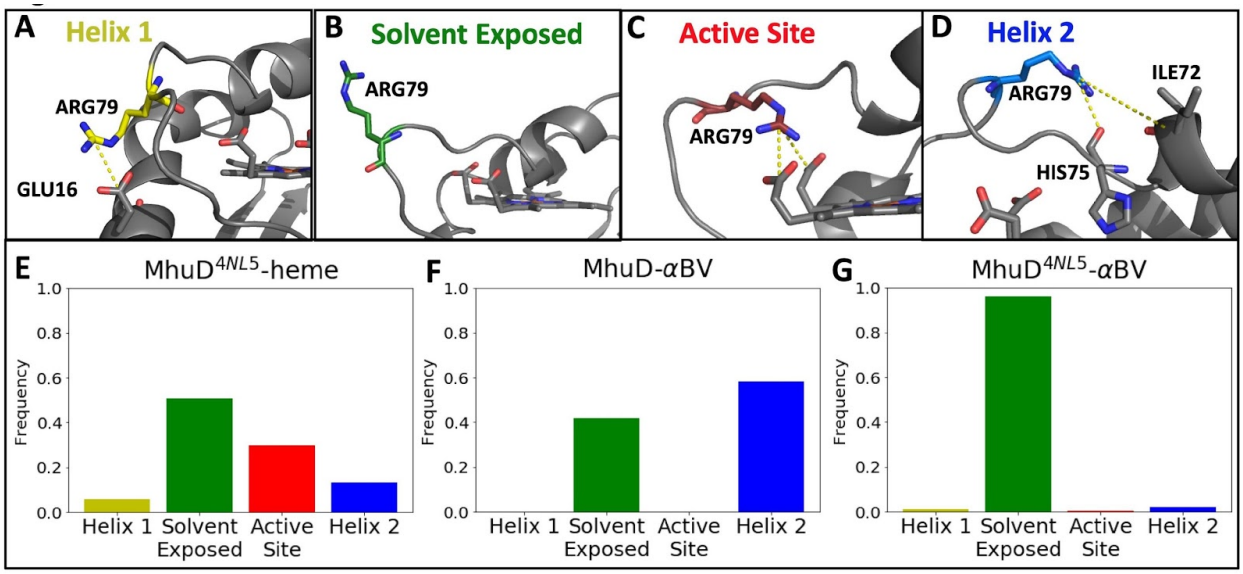


Figure 3.7 Position of Arg79 side chain during MD simulations of MhuD. In the MhuD-monoheme structure (PDB ID 4NL5), the Arg79 side chain is unresolved. MD simulations of the MhuD- monoheme structure (MhuD^{4NL5}-heme) highlight the flexibility of this residue. We have classified its positions during simulations into four categories: **A.** *Helix 1* shows Arg79 interacting with α-helix-1 (residues 16-25), **B.** *Solvent Exposed* refers to Arg79 freely oriented in the surrounding water, **C.** *Active Site* is where Arg79 interacts with the ligand (heme or αBV), and **D.** *Helix 2* refers to the Arg79 side-chain interacting with residues in α-helix-2 (residues 60-75). All representations in **A-D** are snapshots from the simulation of MhuD^{4NL5}-heme. The frequency of each Arg79 position is plotted in panels **E-G** for all three simulation types. For **E.** MhuD^{4NL5}-heme, Arg79 is highly motile and at times, coordinates with the propionate groups of the heme ligand. For **F.** MhuD-αBV, where simulations are initiated from the coordinates of the MhuD-αBV structure, Arg79 predominantly interacts with helix 2 residues and may be stabilized by its participation in α-helix-3. For **G.** MhuD^{4NL5}-αBV, where heme has been removed from the MhuD-monoheme structure and αBV is docked in its place, the Arg79 remains mostly solvent exposed.

Together these results suggest that in the MhuD-R26S-αBV structure we are (1) observing the proximal αBV in an orientation and location equivalent to the turnover product, (2) that the additional α-helix-3 present in the MhuD-product complex is not a crystallographic artifact and, (3) the 90° rotation of His75 out of the active site is consistent

with substrate turnover. In addition, the simulations implicate Arg79 as a putative catalytic residue that also facilitates formation of α-helix-3 upon MhuD turnover—an observation that warrants further biochemical investigation

Proximal αBV is representative of the MhuD physiological product

Within the MhuD-R26S-αBV structure, the orientation of the proximal αBV is representative of the MhuD physiological product, mycobilin. The proximal αBV adopts a similar orientation as heme in the active site of the MhuD-heme-CN structure (**Figures 3.4A-B**) [25]; however, αBV is rotated approximately 20° about the plane compared to heme-CN and the tetrapyrrole ring structure of αBV is considerably more twisted compared to heme. The reduced affinity of αBV to the MhuD-R26S variant, compared to WT-MhuD, supports that the proximal αBV in the MhuD-R26S-αBV structure adopts a similar orientation as the WT MhuD product. As observed for heme-bound MhuD, where Arg26 interacts with the heme propionate, Arg26 likely forms a non-covalent interaction with the tetrapyrrole product as well; thus, after catalysis, one would expect that location of the product would mirror that of substrate heme, as we observe for the proximal αBV. Additionally, the overall conformation of MhuD proximal αBV is similar to the αBV product of *C. diphtheriae* HmuO (PDB 4GPC)[27], representative of the all-Z-all-syn type BV conformation. However, in another structure of hHO-1 in complex with αBV (PDB 1S8C) [26], the αBV occupies an internal cavity adjacent to the active site and exhibits a more linear extended conformation. This was proposed to be the route of αBV dissociation from hHO-1 in the absence of BVR[26]. As αBV is bound tightly to MhuD, we did not anticipate that we would observe a partially dissociated product conformation or location, as seen in the hHO-1 structure. Furthermore, in the MD simulations of the proximal αBV alone in the MhuD-

R26S-αBV structure, αBV remains in the same relative location and orientation also supporting that this is the correct product orientation. Together these observations suggest that the proximal αBV within MhuD-R26S-αBV structure is the physiological orientation of the MhuD tetrapyrrole product, mycobilin, within the active site.

Discussion

Within the PDB there is only one other protein structure with a strikingly similar αBV stacking conformation in its active site, a BVR from cyanobacteria [41]. As with the MhuD-R26S-αBV structure, the BVR-αBV structure has offset nearly parallel tetrapyrrole planes stacked at van der Waals distance and although the proximal and distal αBV molecules in MhuD have no inter-molecular H-bonds, the BVR αBV molecules form an intermolecular H-bond between the lactam oxygen and pyrrole nitrogen, as observed between the distal and 'bridging' αBV molecules in the MhuD-R26S-αBV asymmetric unit (**Figure 3.3A**).

Tetrapyrrole stacking is sometimes important for enzyme activity; however, it has also been observed that heme-heme stacking can be the product of crystallization. Heme stacking has previously been shown to be involved in protein electron transfer reactions; for example, NapB, a cytochrome subunit of nitrate reductase, requires two stacked heme molecules for electron transfer[42]. More recently it was demonstrated that the cyanobacterial BVR required two stacked BVs to reduce BV to bilirubin[41] by an unprecedented mechanism. In contrast, MhuD can also accommodate two stacked heme molecules per active site, although this renders the enzyme inactive[15]. Tetrapyrrole stacking at the crystallographic interface is not unprecedented, as observed in the structures of MhuD-diheme[15], and the ChaN-heme[43], an iron-regulated lipoprotein

implicated in heme acquisition in *Campylobacter jejuni*. As MhuD is known to be active only in its monoheme form[15], and thus only produces one molecule of product per active site, we hypothesize that the proximal αBV is in the correct orientation of the MhuD tetrapyrrole product; consequently, the three additional stacked αBV molecules adjoining the two active sites of adjacent MhuD monomers are likely a product of crystallization.

The conformational changes between the MhuD substrate and product bound structures, while relatively subtle based on RMSD, are significant compared to those of canonical HOs. These differences are borne out in the fluctuation of the active site volume (**Figure S3A**). The structures of *C. diphtheriae* HO (HmuO) in its substrate- and product-bound form (PDB 1IW0, 4GPC) show some shifting and unwinding of the active site proximal and distal helices[27], while the change in the active site pocket volume is trivial (from ~250 Å^3 to ~260 Å^3)[40]. By comparison, MhuD demonstrates much greater conformational versatility. When MhuD binds one heme molecule in its active form, the C-terminal of α-helix-1 unravels to accommodate the heme molecule and the C-terminal region of α-helix-2 extends to all the catalytically essential His75 to coordinate heme-iron, resulting in a kinked α-helix-2. The MhuD-monoheme complex can also bind another molecule of heme resulting in its diheme inactive form, whereby the kinked α-helix-2 is now extended and His75 binds to heme-iron of the solvent exposed heme molecule, nearly tripling the volume of the active site from ~190 Å^3 to ~530 Å^3 compared to the monoheme structure[40]. Alternatively, when MhuD-monoheme turns over in the presence of an electron source, it forms the MhuD-product structure, which leads to the further unraveling of the C-terminal of α-helix-1 and the kinked α-helix-2 along with the formation of α-helix-3. With the transformation of substrate to product, the active site of MhuD doubles in

volume, from ~190 Å^3 to ~360 Å^3[3, 40]. The four different conformational states of MhuD highlight this protein's inherent flexibility (**Figure B.3B**), which may be harnessed to produce MhuD inhibitors.

Little is known about the fate of IsdG-like protein products; but removal of their tetrapyrrole products requires protein denaturation[19,21]. Indeed, our previous and current work suggest that both MhuD substrate and product bind in the low nanomolar range and therefore the displacement of product by substrate would only occur at high heme concentrations *in vivo*[30]. We propose three possible mechanisms of MhuD product release; (1) a dramatic conformational change would reduce product affinity and result in dissociation, (2) IsdG-type proteins are 'suicide' proteins that after one turnover require degradation or (3) an accessory protein is required for the removal of product from the MhuD active site.

Although MhuD is an inherently flexible protein, as described above, its high affinity for tetrapyrroles decreases the likelihood that a conformational change alone would promote product release. Because it has been shown that *S. aureus* IsdG is degraded *in vivo* in its apo form, yet stabilized in the presence of heme[44], it seems unlikely that IsdG-type enzymes are also degraded when bound to product. As the MhuD-R26S-αBV complex structure has a novel structural element and an associated shift in molecular surface electrostatics in comparison to MhuD-monoheme complex, these conformational changes may promote protein-protein interactions to aid in product removal. Notably, the ^{75}HisXXXArg^{79} motif encompassing the newly formed α-helix-3 is also observed in IsdG-type proteins *S. aureus* IsdG and IsdI (**Figure B.4**). Formation of this helix upon substrate

turnover may be a common feature among all IsdG-type proteins that facilitates removal of their tetrapyrrole products, although further work is required to validate this hypothesis.

Protein-protein interaction-induced product removal is reminiscent of human HO-1 BV removal. Human BVR interacts with hHO-1, albeit via a weak interaction, to remove the product BV and further reduce it to bilirubin[45]. In contrast, the well-studied bacterial HO from *P. aeruginosa* excretes BV without further reduction[14]. We hypothesize that a yet-to-be-identified Mtb protein removes product from MhuD and potentially aids in its eventual excretion, as observed in the human HO system. Surprisingly, Mtb has four close homologs of BVR even though Mtb does not have a conventional BV-producing HO enzyme[15]. One of these homologs, Rv2074, has BV reduction activity although its electron donating cofactor is the flavin cofactor F420 [46], a deazaflavin cofactor that is a low potential hydride transfer agent[47], rather than flavin mononucleotide (FMN) as observed for eukaryotic BVRs[48]. In contrast, Rv1155 does not readily reduce BV[49]. Consequently, it was proposed that one of the BV inactive Mtb BVRs catabolizes mycobilins. Rv2607 and Rv2991 have not been tested for BVR activity and could also act in MhuD product breakdown [49]. Mtb has both heme and siderophore-mediated iron acquisition systems [50]; however, *Mycobacterium leprea* only has a heme uptake and catabolism pathway. Furthermore, the *M. leprea* proteome only has the Mtb BVR homologs, Rv1155 and Rv2607, suggesting one of these proteins is perhaps involved in MhuD product removal.

The MhuD-product complex structure has an additional α-helix-3 and an accompanying change in electrostatic surface potential compared to the MhuD-substrate complex[15]. The structure and MD simulations suggest that α-helix-3 forms when His75 is no longer coordinated to the heme-iron and can freely flip out of the binding

pocket. Furthermore this new α-helix-3 is stabilized by Arg79 moving into the active site vicinity, whereby Arg79 is a residue previously unresolved in both the heme-bound and apo-MhuD structures[15, 25]. Given our results, the conservation of ^{75}HisXXXArg^{79} in other IsdG-type proteins, and the presence of BVR homologs in MhuD, we hypothesize that α-helix-3 formation may be a common structural feature among all product-bound IsdG-type proteins that facilitates protein-protein interactions in order to promote product egress.

Supporting information, including accession IDs and supplemental figures are available in **Appendix B**.

References

1. Dore, S., Takahashi, M., Ferris, C. D., Zakhary, R., Hester, L. D., Guastella, D., and Snyder, S. H. (1999) Bilirubin, formed by activation of heme oxygenase-2, protects neurons against oxidative stress injury. Proc. Natl. Acad. Sci. U S A 96, 2445-2450.

2. Ferris, C. D., Jaffrey, S. R., Sawa, A., Takahashi, M., Brady, S. D., Barrow, R. K., Tysoe, S. A., Wolosker, H., Baranano, D. E., Dore, S., Poss, K. D., and Snyder, S. H. (1999) Haem oxygenase-1 prevents cell death by regulating cellular iron. Nat. Cell. Biol. 1, 152-157.

3. Brouard, S., Otterbein, L. E., Anrather, J., Tobiasch, E., Bach, F. H., Choi, A. M., and Soares, M. P. (2000) Carbon monoxide generated by heme oxygenase 1 suppresses endothelial cell apoptosis. J. Exp. Med. 192, 1015-1026.

4. Tenhunen, R., Marver, H. S., and Schmid, R. (1969) Microsomal heme oxygenase. Characterization of the enzyme. J. Biol. Chem. 244, 6388-6394.

5. Yoshida, T., Noguchi, M., and Kikuchi, G. (1980) Oxygenated form of heme - heme oxygenase complex and requirement for second electron to initiate heme degradation from the oxygenated complex. J. Biol. Chem. 255, 4418-4420.

6. Matsui, T., Unno, M., and Ikeda-Saito, M. (2010) Heme oxygenase reveals its strategy for catalyzing three successive oxygenation reactions. Acc. Chem. Res. 43, 240-247.

7. Schmitt, M. P. (1997) Utilization of host iron sources by *Corynebacterium diphtheriae*: identification of a gene whose product is homologous to eukaryotic heme oxygenases and is required for acquisition of iron from heme and hemoglobin. J. Bacteriol. 179, 838-845.

8. Wilks, A., and Schmitt, M. P. (1998) Expression and characterization of a heme oxygenase (Hmu O) from *Corynebacterium diphtheriae.* Iron acquisition requires oxidative cleavage of the heme macrocycle. J. Biol. Chem. 273, 837-841.

9. Zhu, W., Wilks, A., and Stojiljkovic, I. (2000) Degradation of heme in gram-negative bacteria: the product of the hemO gene of Neisseriae is a heme oxygenase. J. Bacteriol. 182, 6783-6790.

10. Ratliff, M., Zhu, W., Deshmukh, R., Wilks, A., and Stojiljkovic, I. (2001) Homologues of neisserial heme oxygenase in gram-negative bacteria: degradation of heme by the product of the pigA gene of *Pseudomonas aeruginosa.* J. Bacteriol. 183, 6394-6403.

11. Wilks, A. (2002) Heme oxygenase: evolution, structure, and mechanism. Antioxid. Redox Signal 4, 603-614.

12. Noguchi, M., Yoshida, T., and Kikuchi, G. (1979) Specific requirement of NADPH-cytochrome c reductase for the microsomal heme oxygenase reaction yielding biliverdin IX alpha. FEBS Lett. 98, 281-284.

13. Mantle, T. J. (2002) Haem degradation in animals and plants, *Biochem Soc Trans 30*, 630-633.

14. Barker, K. D., Barkovits, K., and Wilks, A. (2012) Metabolic flux of extracellular heme uptake in Pseudomonas aeruginosa is driven by the iron-regulated heme oxygenase (HemO). J. Biol. Chem. 287, 18342-18350.

15. Chim, N., Iniguez, A., Nguyen, T. Q., and Goulding, C. W. (2010) Unusual diheme conformation of the heme-degrading protein from *Mycobacteriuim tuberculosis*. J. Mol. Biol. 395, 595-608.

16. Skaar, E. P., Gaspar, A. H., and Schneewind, O. (2004) IsdG and IsdI, heme-degrading enzymes in the cytoplasm of *Staphylococcus aureus*. J. Biol. Chem. 279, 436-443.

17. Wu, R., Skaar, E. P., Zhang, R., Joachimiak, G., Gornicki, P., Schneewind, O., and Joachimiak, A. (2005) *Staphylococcus aureus* IsdG and IsdI, heme-degrading enzymes with structural similarity to monooxygenases. J. Biol. Chem. 280, 2840-2846.

18. Matsui, T., Nambu, S., Ono, Y., Goulding, C. W., Tsumoto, K., and Ikeda-Saito, M. (2013) Heme degradation by *Staphylococcus aureus* IsdG and IsdI liberates formaldehyde rather than carbon monoxide. Biochemistry 52, 3025-3027.

19. Reniere, M. L., Ukpabi, G. N., Harry, S. R., Stec, D. F., Krull, R., Wright, D. W., Bachmann, B. O., Murphy, M. E., and Skaar, E. P. (2010) The IsdG-family of haem oxygenases degrades haem to a novel chromophore. Mol. Microbiol. 75, 1529-1538.

20. Lojek, L. J., Farrand, A. J., Wisecaver, J. H., Blaby-Haas, C. E., Michel, B. W., Merchant, S. S., Rokas, A., and Skaar, E. P. (2017) *Chlamydomonas reinhardtii* LFO1 Is an IsdG Family Heme Oxygenase. mSphere *2*, e00176-17.

21. Nambu, S., Matsui, T., Goulding, C. W., Takahashi, S., and Ikeda-Saito, M. (2013) A new way to degrade heme: the *Mycobacterium tuberculosis* enzyme MhuD catalyzes heme degradation without generating CO. J. Biol. Chem. 288, 10101-10109.

22. Vanella, L., Barbagallo, I., Tibullo, D., Forte, S., Zappala, A., and Li Volti, G. (2016) The non-canonical functions of the heme oxygenases. Oncotarget 7, 69075-69086.

23. Schuller, D. J., Wilks, a., Ortiz de Montellano, P. R., and Poulos, T. L. (1999) Crystal structure of human heme oxygenase-1. Nat. Struct. Biol. 6, 860-867.

24. Matsui, T., Nambu, S., Goulding, C. W., Takahashi, S., Fujii, H., and Ikeda-Saito, M. (2016) Unique coupling of mono- and dioxygenase chemistries in a single active site promotes heme degradation. Proc. Natl. Acad. Sci. U S A 113, 3779-3784.

25. Graves, A. B., Morse, R. P., Chao, A., Iniguez, A., Goulding, C. W., and Liptak, M. D. (2014) Crystallographic and spectroscopic insights into heme degradation by *Mycobacterium tuberculosis* MhuD. Inorg. Chem. 53, 5931-5940.

26. Lad, L., Friedman, J., Li, H., Bhaskar, B., Ortiz de Montellano, P. R., and Poulos, T. L. (2004) Crystal structure of human heme oxygenase-1 in a complex with biliverdin. Biochemistry 43, 3793-3801.

27. Unno, M., Ardevol, A., Rovira, C., and Ikeda-Saito, M. (2013) Structures of the substrate-free and product-bound forms of HmuO, a heme oxygenase from *Corynebacterium diphtheriae*: X-ray crystallography and molecular dynamics investigation. J. Biol. Chem. 288, 34443-34458.

28. Lee, W. C., Reniere, M. L., Skaar, E. P., and Murphy, M. E. (2008) Ruffling of metalloporphyrins bound to IsdG and IsdI, two heme-degrading enzymes in *Staphylococcus aureus*. J. Biol. Chem. 283, 30957-30963.

29. Chao, A., and Goulding, C. W. (2019) A Single Mutation in the *Mycobacterium tuberculosis* Heme-Degrading Protein, MhuD, Results in Different Products. Biochemistry 58, 489-492.

30. Thakuri, B., Graves, A. B., Chao, A., Johansen, S. L., Goulding, C. W., and Liptak, M. D. (2018) The affinity of MhuD for heme is consistent with a heme degrading function in vivo. Metallomics 10, 1560-1563.

31. Conger, M. A., Pokhrel, D., and Liptak, M. D. (2017) Tight binding of heme to *Staphylococcus aureus* IsdG and IsdI precludes design of a competitive inhibitor. Metallomics 9, 556-563.

32. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) Protein measurement with the Folin phenol reagent. J. Biol. Chem. 193, 265-275.

33. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R., and Leslie, A. G. (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr. D Biol. Crystallogr. 67, 271-281.

34. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol .Crystallogr. 66, 213-221.

35. Case, D. A., Cerutti, D. S., Cheatham III, T. E., Darden, T. A., Duke, R. E., Giese, T. J., Gohlke, H., Goetz, A. W., Greene, D., Homeyer, N., S. Izadi, P. Janowski, J. Kaus, A., and Kovalenko, T. S. L., LeGrand S., Li P., Luchko T., Luo R., Madej B., Merz K.M., Monard G., Needham P., Nguyen H., Nguyen H.T., Omelyan I., Onufriev A., Roe D.R., Roitberg A., Salomon-Ferrer R., Simmerling C.L., Smith W., Swails J., Walker R.C., Wang J., Wolf R.M., Wu X., York D.M. and Kollman P.A. (2015) AmberTools15.

36. Harris, D., Loew, G., and Waskell, L. (2001) Calculation of the electronic structure and spectra of model cytochrome P450 compound I. J. Inorg. Biochem. 83, 309-318.

37. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78, 1950-1958.

38. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. 1*3*, e1005659.

39. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernandez, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J., and Pande, V. S. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories, Biophys. J. 109, 1528-1532.

40. Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018) CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 4*6*, W363-W367.

41. Takao, H., Hirabayashi, K., Nishigaya, Y., Kouriki, H., Nakaniwa, T., Hagiwara, Y., Harada, J., Sato, H., Yamazaki, T., Sakakibara, Y., Suiko, M., Asada, Y., Takahashi, Y., Yamamoto, K., Fukuyama, K., Sugishima, M., and Wada, K. (2017) A substrate-bound structure of cyanobacterial biliverdin reductase identifies stacked substrates as critical for activity. Nat. Commun. 8, 14397.

42. Brige, A., Leys, D., Meyer, T. E., Cusanovich, M. A., and Van Beeumen, J. J. (2002) The 1.25 A resolution structure of the diheme NapB subunit of soluble nitrate reductase reveals a novel cytochrome c fold with a stacked heme arrangement. Biochemistry 41, 4827-4836.

43. Chan, A. C., Lelj-Garolla, B., F, I. R., Pedersen, K. A., Mauk, A. G., and Murphy, M. E. (2006) Cofacial heme binding is linked to dimerization by a bacterial heme transport protein. J. Mol. Biol. 362, 1108-1119.

44. Reniere, M. L., Haley, K. P., and Skaar, E. P. (2011) The flexible loop of *Staphylococcus aureus* IsdG is required for its degradation in the absence of heme. Biochemistry 50, 6730-6737.

45. Maines, M. D., and Trakshel, G. M. (1993) Purification and characterization of human biliverdin reductase. Arch. Biochem. Biophys. 300, 320-326.

46. Ahmed, F. H., Mohamed, A. E., Carr, P. D., Lee, B. M., Condic-Jurkic, K., O'Mara, M. L., and Jackson, C. J. (2016) Rv2074 is a novel F420 H2 -dependent biliverdin reductase in *Mycobacterium tuberculosis*. Protein Sci. 25, 1692-1709.

47. Bashiri, G., Antoney, J., Jirgis, E. N. M., Shah, M. V., Ney, B., Copp, J., Stuteley, S. M., Sreebhavan, S., Palmer, B., Middleditch, M., Tokuriki, N., Greening, C., Scott, C., Baker, E. N., and Jackson, C. J. (2019) A revised biosynthetic pathway for the cofactor F420 in prokaryotes. Nat. Commun. 10, 1558.

48. Sugishima, M., Wada, K., and Fukuyama, K. (2019) Recent Advances in the Understanding of the Reaction Chemistries of the Heme Catabolizing Enzymes HO and BVR Based on High Resolution Protein Structures. Curr. Med. Chem. 26, 1-12.

49. Ahmed, F. H., Carr, P. D., Lee, B. M., Afriat-Jurnou, L., Mohamed, A. E., Hong, N. S., Flanagan, J., Taylor, M. C., Greening, C., and Jackson, C. J. (2015) Sequence-Structure-Function Classification of a Catalytically Diverse Oxidoreductase Superfamily in Mycobacteria. J. Mol. Biol. 427, 3554-3571.

50. Chao, A., Sieminski, P. J., Owens, C. P., and Goulding, C. W. (2019) Iron Acquisition in *Mycobacterium tuberculosis*. Chem. Rev. 119, 1193-1220.

CHAPTER 4: Insights from the structure and conformational dynamics of *Mycobacterium tuberculosis* malic enzyme, MEZ

Abstract

Tuberculosis (TB) is the most lethal infectious disease worldwide. It is notoriously difficult to treat, requiring a cocktail of antibiotics administered over several months. The dense, waxy outer membrane of the TB-causing agent, *Mycobacterium tuberculosis* (Mtb), acts as a formidable barrier against uptake of antibiotics. Enzymes involved in maintaining the integrity of the Mtb wall are promising drug targets. Recently, we demonstrated that knocking out the *mez* gene, which encodes for Mtb malic enzyme (MEZ), alters the morphology of the Mtb cell wall and results in deficient uptake by macrophages. Here, we present the structure of MEZ and compare it with known structures of prokaryotic and eukaryotic malic enzymes. We use biochemical assays to determine its oligomeric state and to evaluate the effects of pH and allosteric regulators on its thermal stability. To explore and compare the interactions between MEZ and its substrate malate and cofactors $NAD(P)^+$, and Mn^{2+}, we ran a series of molecular dynamics (MD) simulations. Our MD analysis corroborates our empirical observations that MEZ is unusually disordered, and that this disorder persists despite the addition of cofactors $NAD(P)^+$ and substrate. MD simulations also reveal that MEZ subunits alternate between open and closed states, and that MEZ can stably bind its $NAD(P)^+$ cofactor in multiple conformations, including an inactive, compact conformation. Together the structure of MEZ and insights from its dynamics could be harnessed to inform design of a MEZ inhibitor.

Introduction

The pathogen *Mycobacterium tuberculosis* (Mtb) is the causative agent of tuberculosis (TB). In the late 20th century, TB was believed to be nearly eradicated world-wide.[1] However, in the 1980s, with the rapid and widespread dissemination of the HIV virus, immunocompromised AIDS patients started to die from TB. The resurgence of TB was further compounded by its complicated treatment, which comprises a 3- to 4-antibiotic regimen over at least a six-month period.[2] Non-compliance with the prescribed TB therapy has resulted in the emergence of multi- and extensively-drug resistant Mtb strains. Nearly 40 years later, TB is the most lethal infectious disease in the world, with 1.5 million TB-related deaths and 1.1 million new Mtb infections in 2018 alone.[3] New anti-TB therapeutics, with streamlined treatment regimens and fewer side effects, are urgently needed.

In a recent study, our collaborators demonstrated that enzymes at the anaplerotic node of central carbon metabolism, which replenish the intermediates of the citric acid cycle, are essential for Mtb pathogenesis and may serve as novel therapeutic targets.[4] One such enzyme, malic enzyme (ME), reversibly converts malate and NAD(P)+ to pyruvate, CO_2 and NAD(P)H. Although our previous study suggests that Mtb ME (MEZ) plays just a minor back-up role in CO_2-dependent anaplerosis (acting in the reverse direction as a pyruvate carboxylase), we demonstrate that MEZ has a significant accessory role in biosynthesis of lipids for the mycomembrane. While the MtbΔ*mez* strain grows comparably to wild type Mtb on gluconeogenic or glycolytic substrates, the *mez* deletion strain had impaired macrophage invasion. Furthermore, the MtbΔ*mez* colonies had a glossy and viscous morphology, similar to Mtb mutants lacking cell-wall lipid transporters

and a trehalose dimycolate esterase[5–8], suggesting that the lipid composition of the cell wall in in Mtb *mez* knockouts is compromised. In fact, the *mez* deletion strain accumulates apolar free fatty and mycolic acids in the cytosol, resulting in an accompanying decrease in the levels of cell-wall bound mycolates. These results suggest that Mtb MEZ predominantly functions as a L-malate decarboxylase producing NAD(P)H for utilization in lipid biosynthesis.

The role of MEZ in lipid metabolism is further reflected in the fact that *mez* is part of the *phoPR* regulon, a two-component system that regulates genes essential for the biosynthesis of complex lipids required for Mtb virulence.[9] Like Mtb∆*mez,*, the Mtb∆*phoP* mutant also results in a structurally distinct cell envelope; thus its likely MEZ functions as a source of NAD(P)H for the synthesis of complex lipids. The connection between lipid synthesis and NAD(P)H generating MEs is not unique to Mtb. Studies of eukaryotic oleaginous microorganisms, particularly of algae and yeast, reveal that MEs produce NAD(P)H for fatty-acid rich storage compounds such as triacylglycerol (TAG)[10]. Within prokaryotes, mostly only actinobacteria have the ability to accumulate TAG; although in *Streptomyces coelicolor*, the deletion of the two genes that express NAD- and NADP-dependent MEs results in a mutant that has decreased production of TAG.[11] Additionally, in *Rhodococcus jostii*, the overexpression of ME while growing *R. jostii* on glucose promotes an increased NADP-dependent ME activity resulting in a near 2-fold increase in fatty acid synthesis.[12] Altogether these studies, demonstrate clear precedence for the relationship between MEs and lipid biosynthesis. Given our previous study, we believe MEZ plays a major role in supplying a NAD(P)H reducing agents required for lipid biosynthesis, and

therefore virulence, in Mtb. Thus, MEZ presents as a compelling candidate for structure-based anti-TB drug design.

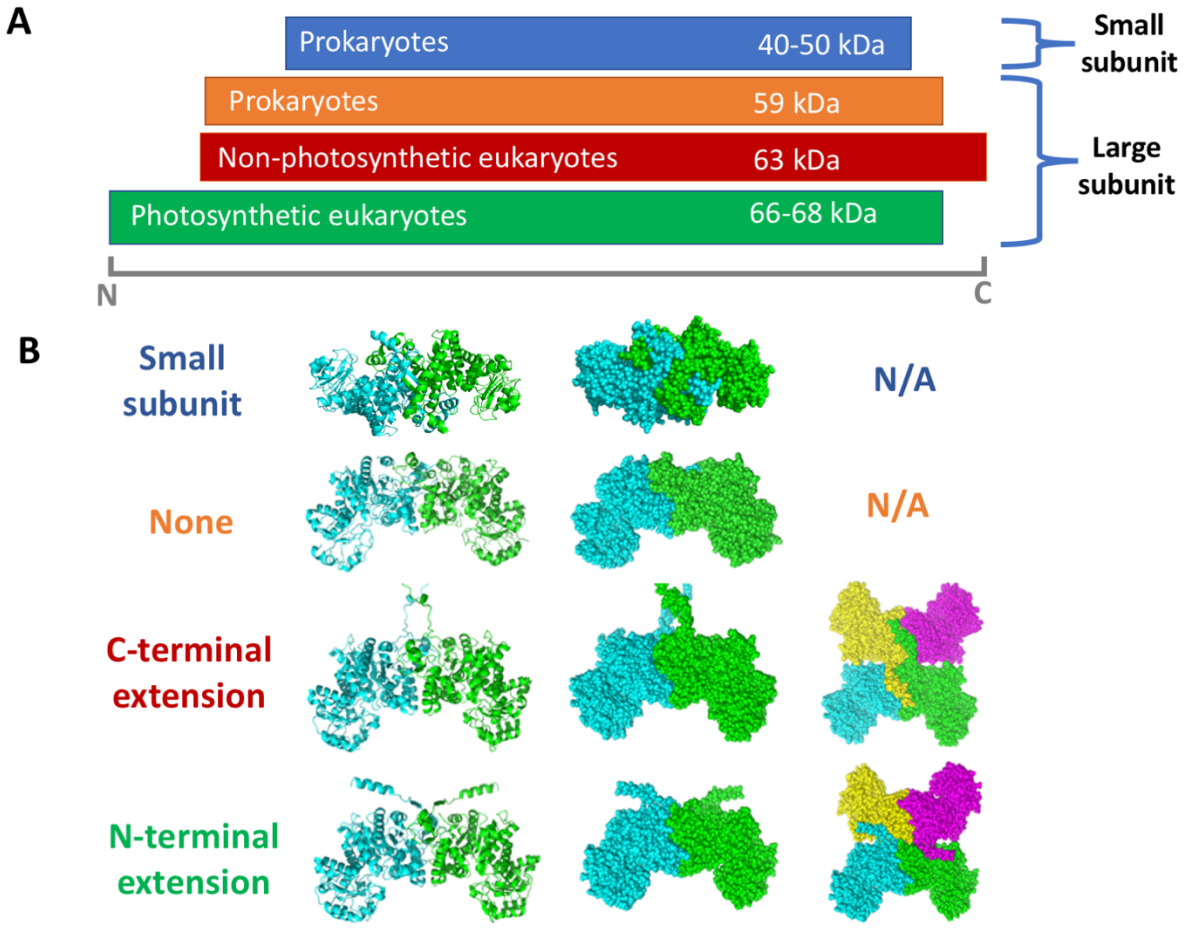


Figure 4.1 Malic Enzyme Structural Classes and Oligomeric States. (A) Malic enzymes (ME) can be classified as small subunit (blue) and large subunit (orange, red, green). Small subunit MEs are typically prokaryotic and range from 40-50kDa, while large subunit MEs have been found in both prokaryotes and eukaryotes and range from 59-68kDa in size. Among large subunit Mes, the presence and/or absence of N- and C- terminal tails broadly corresponds with the biological domain of its source species. (B) Small subunit MEs form dimers such as *Pyrococcus horikoshii* ME (PDB ID:1WW8). Among large subunit MEs, slight variations are observed in the lengths of the N- and C-terminal tails. Tetrameric models of the most complete chains from representative structures of *Mtb* MEZ, human mitochondrial ME (C-terminal extension PDB ID:3WJ6), and maize ME (N-terminal extension, PDB ID:5OU5) reveal how the extended termini interact with subunits of neighboring dimer species to stabilize formation of the tetrameric species. *E. coli* ME (PDB ID:6AGS), represented in the second row, has neither an N- nor C-terminal extension and exists as a dimer, similar to MEZ.

The critical role of MEs in metabolism and NAD(P)H generation is reflected in its ubiquity amongst many species, including eukaryotes, prokaryotes and archaea. It follows

that there is a sizeable group of diverse X-ray crystal structures of NADP- and NAD-MEs in the Protein Data Bank (PDB, **Table C.1**) ranging from high eukaryotes, to plants, to prokaryotes, all of which contain a NAD(P)-binding Rossman-fold domain. Among those studied, MEs typically adopt dimeric or tetrameric assemblies with subunits ranging from 40-50 kDa for smaller prokaryotic MEs and around 60 kDa for many larger subunit MEs found in eukaryotes and prokaryotes (**Figure 4.1A**).

To-date, there have been three distinct structural classes of MEs reported: eukaryotic large subunit MEs, prokaryotic small subunit MEs, and hybrid chimeric MEs; the subset of hybrid chimeric MEs will not be discussed further. The structures of eukaryotic large subunit MEs, which are approximately 60 kDa, consist of a tetrameric assembly comprised of a dimer of dimers (Figure **4.1B**). Each monomer has four distinct domains (A-D). Within higher eukaryotes, the presence of a C-terminal tail in Domain D facilitates the tetramerization of two dimers. In contrast, in plant ME proteins, dimer tetramerization is facilitated by an N-terminal tail in Domain A. There is one large subunit ME in the PDB from *E. coli* with a dimeric, non-tetrameric assembly. The other distinct class consists of prokaryotic small subunit MEs, approximately ~40kDa and dimeric, which have a minimal Domain A and no Domain D. In fact, assembly of the dimer is required for full formation of the active site, wherein each monomer subunit contributes a Lys and a Tyr residue that are both required for catalysis.[13] By comparison, larger subunit ME proteins have fully formed active sites, independent of dimerization. Alignment of representative prokaryotic and eukaryotic large subunit ME sequences highlights the presence of an extended N and/or C-terminal tail among the larger subunit MEs, which is unique to eukaryotes and may be required to form tetrameric oligomeric states (**Figure 4.1B)**. In all structural classes of MEs

reported, Domains B and C have conserved motifs and a high degree of structural similarity.

Complementing the structural analyses of MEs, there is a large body of biochemical data on these enzymes, particularly for higher eukaryotic and plant MEs, where their mechanism of action, catalytic residues and activators and inhibitors have been well studied.[14–18] This has been particularly well-reviewed for higher eukaryotes by Tong *et al*[19]. In contrast, there is less known about the biochemistry of larger subunit prokaryotic MEs (apart from *E. coli*[20,21]).

Understanding the X-ray crystal structure of Mtb ME (MEZ) is paramount to understanding its biochemistry and assessing its viability as an anti-TB drug. Herein, we discuss the structure of MEZ in comparison with human ME structures and in the context of the Mtb MEZ biochemical data. Additionally, we use docking and molecular simulations to explore the conformational flexibility of MEZ and to characterize its interaction with its malate substrate and $NAD(P)^+$ cofactors.

Results

Structure of Mtb Malic Enzyme (MEZ)

The asymmetric unit contains four subunits or two dimers of MEZ. Dimer 1 is comprised of subunits A and B, while dimer 2 is formed by subunits C and D. Dimer 1 and dimer 2 are very similar with root mean square deviation (RMSD) for αC atoms of 0.88 Å. Similarly, the subunits align closely, where subunit D with RMSD of 0.72 Å and subunit B is most like subunit C with RMSD of 1.05 Å. In the dimers, subunit A aligns to subunit B with RMSD of 1.37 Å, and subunit C aligns to subunit D with an RMSD of 1.12 Å.

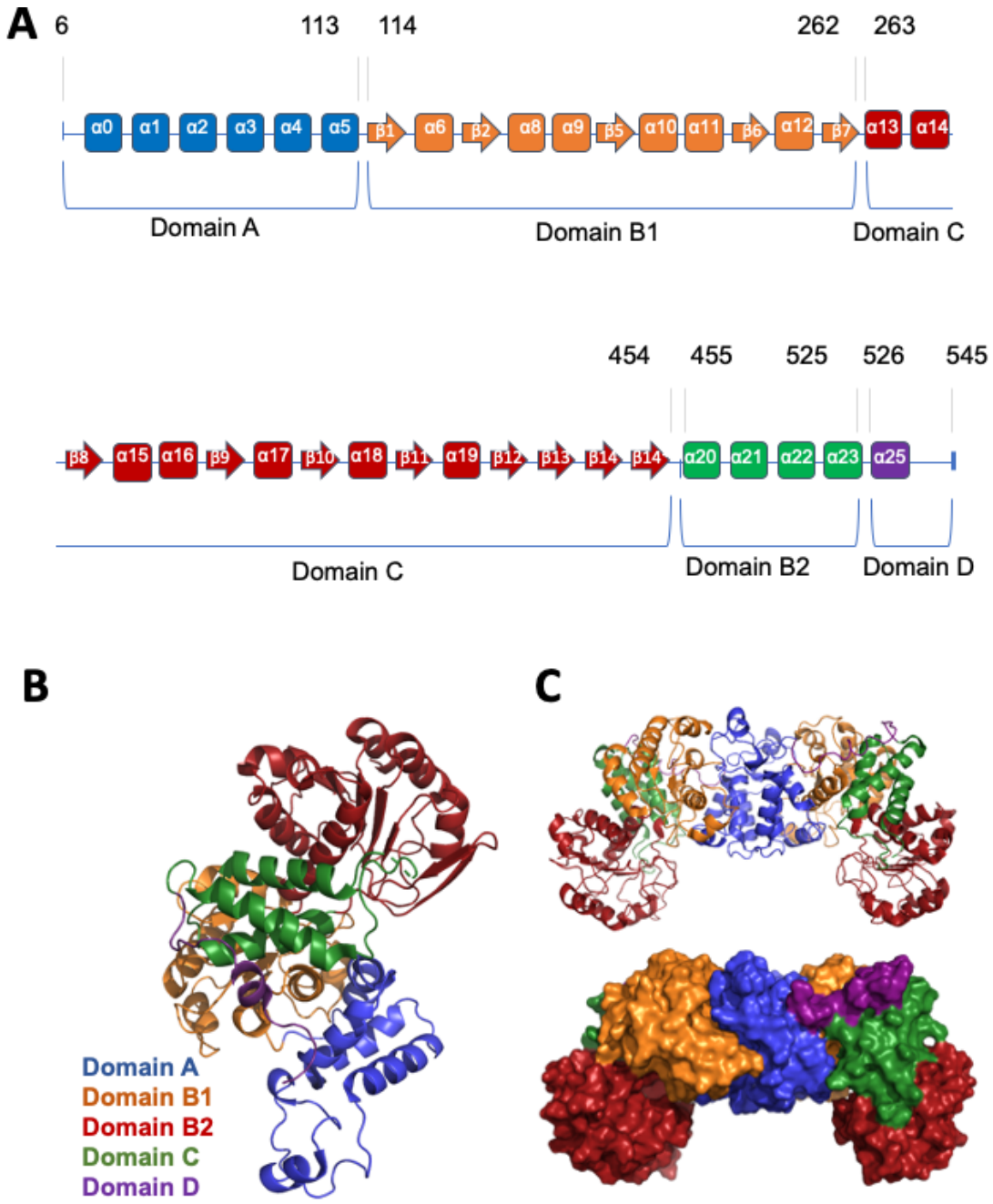


Figure 4.2 Structural Domains of MEZ. A The structural domains of MEZ corresponding to the previously described domains of homologous human mitochondrial NAD(P)$^+$ malic enzyme are shown here in this diagram with the corresponding range of residue numbers at the start and end of each domain. B-strands are represented with arrows while helices are depicted with rectangles. **B.** Structural cartoon of the MEZ monomer with the domains colored as (A). **C.** The cartoon and surface representations of dimeric MEZ are colored by domain.

Like other MEs, MEZ is composed of four domains (**Figure 4.2 A&B**). Domain A (residues 1-113) is composed of five α-helices (α1-α5). Domain B (residues 114-262, and 455-525) has an internal β-sheet composed of 5 parallel β-strands (β1/β5/β2/β6/β7). The rest of domain B is helical: α6-α12 and α20-α24. Domain C is formed by residues 263-454, and bears the canonical Rossmann fold of ME proteins. The Rossmann-fold has a 6-stranded parallel β-sheet (β9/β8/β10/β11/β12/β14). Preceding the final β-strand of the β-sheet is a conserved β-hairpin composed of strands β13 and β14. The β-sheet is enclosed by seven α-helices: α13-α19. Lastly, the tail of the C-terminus (residues 526-545) composes domain D, which is largely unstructured but for a final α-helix: α24.

The four subunits of MEZ are very similar; however, there are some noteworthy differences. A rather unexpected difference is that within subunit B, the proposed Rossmann-fold NAD(P) binding site is unstructured and, thus, not modeled due to an absence of observable electron density, potentially indicating increased mobility in this region of the structure. Additionally, there is some variability in modeled α-helices between subunits. Subunit C has an additional helix (residues 11-15) before α1, which aligns to a α-helix in human ME (PDB ID: 1QR6). All subunits show a substantial kink in α5, where in subunit D the α-helix is interrupted between residues 96-102; similar to subunit D, a break is also observed in human ME (residues 111-113). Finally, α17, which is structured in subunits A, C, D, is not structured in subunit B.

More striking is the difference in the strands of the β-sheets between subunits. ME proteins have a conserved β-hairpin that is not present in subunit A, but is observed in subunit C (β2a/ β2b), and is only partially formed in subunits B and D. Additionally, in the Rossmann fold β-sheet, strand β6 is interrupted in subunit C due to a complete loss of

density. Lastly, while the Rossmann fold of subunit A shows the canonical 6-stranded β-sheet with residues 451 and 452 forming an abridged β-strand (β14*) following β-hairpin strand β14, in the other subunits these residues are in one continuous β-strand (β14) with varying likelihoods of participating in the Rossmann fold. Notably, in chain B and C, β14 is shifted away from the Rossmann fold β-sheet, however, in subunit B, β13 appears equally likely or unlikely to participate in the β-sheet as β14. While in subunit D, β14 aligns well to β14* positioning and is likely to participate in the Rossmann fold β-sheet. This suggests that in some MEZ conformations the Rossmann-fold β-sheet stabilizes the β-hairpin and reduces its mobility.

Interestingly, in β9 there is a defined kink that is not observed in other MEs due to MEZ Pro328. When MEZ is aligned to human ME (PDB: 1QR6) or the minimal bacterial ME (PDB: 5CEE), measuring the angle between the α-carbon of P328 to the α-carbon of human ME I341 and to α-carbon of human ME M343 carbon or from α-carbon of P329 to the α-carbon of 5CEE ME V341 to the α-carbon of 5CEE ME L213 gives the resulting angle of 6° or 17°, respectively.

The MEZ structure shows an 'open' conformation when compared to other structures, which is unsurprising considering the NAD(P) cofactor and substrate-binding sites are unoccupied but for a glycerol molecule from the cryoprotectant in subunit A. Notably, relative to other ME structures in the PDB, MEZ appears to be the least well ordered as (1) many secondary structure elements are random coils or have no observable electron density in the MEZ structure compared to others and (2) MEZ appears to be relatively unstable as considerable efforts were made to improve the diffraction quality of MEZ crystals to no avail.

Table 4.1. Data collection and refinement statistics for MEZ crystal structure.

	MEZ
Data collection	
Space group	P 2_1
Cell dimensions	
a, *b*, *c* (Å)	94.6, 144.1, 118.1
α ,β, γ(°)	90, 109.5, 90
Resolution (Å)	89.2-3.6 (3.7-3.6)[a]
R_{merge}[b]	0.178 (0.917)
I / σI	2.9 (0.87)
Completeness (%)	97.62 (94.86)
Redundancy	2.0 (1.9)
Refinement	
Resolution (Å)	89.2-3.6 (3.7-3.6)
Total reflections	66726 (6289)
Unique reflections	33948 (3301)
R_{work} / R_{free}[c]	32.1 / 35.8
Ramachandran favored (%)	90.0
Ramachandran outliers (%)	0.25
No. atoms	
Protein	13575
Ligands	72
Water	145
B-factors (Å^2)	
Protein	86.31
Ligands	86.78
Water	54.01
R.m.s. deviations	
Bond lengths (Å)	0.01
Bond angles (°)	1.06
PDB-ID	**6URF**

[a]. Values within parentheses refer to the highest resolution shell.
[b]. $R_{merge} = \sum\sum|I_{hkl}—I_{hkl}(j)|/\sum I_{hkl}$, where $I_{hkl}(j)$ is observed intensity and I_{hkl} is the final average value of intensity. $R_{pim} = \sum\{1/[N_{hkl}-1]\}^{1/2} \times \sum|I_{hkl}(j) —I_{hkl}|/\sum\sum I_{hkl}(j)$. $R_{meas} = \sum\{N_{hkl}/[N_{hkl}-1]\}^{1/2} \times \sum|I_{hkl}(j) — I_{hkl}|/\sum\sum I_{hkl}(j)$. $CC = \sum (x-(x))(y-(y))/ [\sum(x-(x))^2 \sum(y-(y))^2]^{1/2}$.
[c]. $R_{work} = \sum||F_{obs}| - |F_{calc}||/\sum|F_{obs}|$ and $R_{free} = \sum||F_{obs}| - |F_{calc}||/\sum|F_{obs}|$, where all reflections belong to a test set of 5% data randomly selected in Phenix.

The oligomeric state of MEZ

Most of the reported large subunit ME proteins are tetrameric, and MEZ falls into the large subunit category. The MEZ X-ray crystal structure has four subunits in the asymmetric unit, where the two dimers have a similar assembly to the large subunit ME

dimers (**Figure 4.1B & Figure 4.2C**). The two dimers do weakly interact to form a tetramer; however, the MEZ tetrameric state does not resemble any previously observed ME tetramer assemblies. Furthermore, the MEZ sequence does not have N- or C-terminal tails that facilitate tetramer formation in plant and higher eukaryote MEs, respectively (**Figure 4.2**).

The dimer interface, similar to those seen in other large subunit ME proteins, is facilitated predominantly by hydrophobic interactions (greater than 80%) and some H-bonds between Domain A from each monomer. Additionally, Domain A from one monomer also interacts with a couple of loop regions in Domain B (residues 201-207) from the other monomer. The surface areas buried in Dimer 1 (subunits A&B) and Dimer 2 (subunits C&D) are 2431 and 2428 $Å^2$, respectively, where the monomers have a surface area of 24381-27347 $Å^2$. At the dimer/dimer interface of the MEZ structure there is very little surface area buried, 396 $Å^2$ and 379 $Å^2$ for Subunits A&D and B&C, respectively, suggesting that the tetramer in the asymmetric unit is an artifact of crystal packing rather than MEZ forming a stable tetramer in solution. The only other prokaryotic large subunit ME structure (PDB ID: 6AGS, *E. coli* ME, SfcA) is also dimeric; however this structure lacks an accompanying publication and prior biochemical studies of SfcA suggest that it forms a tetramer by size exclusion chromatography (SEC). Nonetheless, the *E. coli* ME dimer superimposes well with the MEZ dimer (RMSD 2.201 Å over all atoms) and both have similar dimeric interfaces. Notably, in the *E. coli* ME dimer, the first and last β-strand from alternate monomers form two 2-stranded β-sheets, in contrast to the MEZ structure, where the N- and C-termini of all monomers are random coils (**Figure 4.2**). To confirm the

oligomeric state of MEZ in solution, we carried out SEC and the results suggest that MEZ is predominately dimeric in solution (**Figure 4.3A**).

For large subunit MEs, it has been observed that the preferred multimeric state shifts at lower pH[17]; additionally allosteric regulators like fumarate have been shown to stabilize tetramer formation for human mitochondrial ME.[15] In our previous study, we demonstrated that at lower pH, MEZ was more active in both the forward and reverse directions. Here, we found that the activity of MEZ in the presence of alanine, succinate and fumarate is elevated 100%, 200% and 300%, respectively (**Figure 4.3B**). To evaluate the effects pH and allosteric regulators on the oligomeric assembly of MEZ, we also performed SEC analysis of MEZ at pH 6.0 and following incubation with 100-fold molar excess of fumarate, succinate and alanine (and including 2mM of the small molecule in the running buffer). In all conditions studied, MEZ remained predominately dimeric (data not shown).

Allosteric Regulation and Thermal Stability of MEZ

To test whether low pH and allosteric regulators elevate the activity of MEZ by stabilizing the structure, we examined the impact of ligand binding and pH on the thermal stability of MEZ. Using differential scanning fluorimetry, we determined the melting temperature (T_m) of MEZ in the presence of 200-fold molar excess of known cofactors NAD and NADPH, malic acid, and malic acid with KCl and at pHs 6.0, 6.5, 7.0 and 7.5. All conditions tested also contained 10mM $MnCl_2$. When controlling for pH, the addition of cofactors, substrate and KCl, as shown in **Figure 4.3C**, had minimal impact on the T_m of MEZ, with the greatest shift (~1.7°C) occurring at pH 7.5 between apo MEZ and MEZ + NAD.

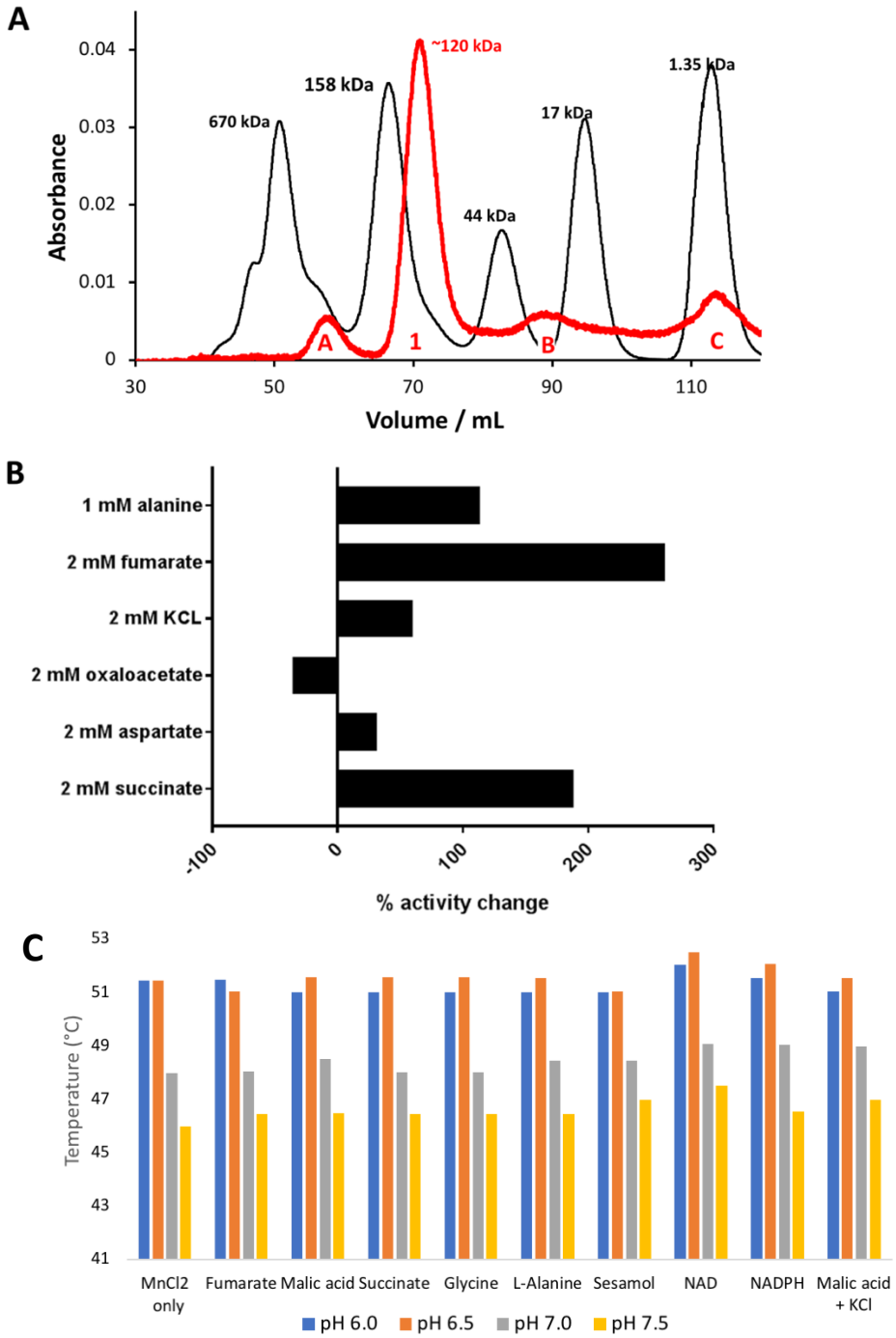


Figure 4.3 Biophysical and Biochemical Analyses of MEZ. A. Gel Filtration of MEZ. To identify its oligomeric state, His-tagged MEZ was purified via Ni-chromatography. The predominant species is the dimeric species in Peak C, which runs at an approximate size of 120kDa. SDS-PAGE analysis of peak B reveals a contaminant with an estimated molecular weight of 75kDa. Species in Peak A are likely higher order aggregates of MEZ and/or the contaminant species in Peak B and Peak C is a small molecule such as imidazole. **B.** Potential allosteric regulators of MEZ. Kinetic studies of MEZ in the presence of various small molecules suggest that MEZ is upregulated 2-4 fold by alanine, succinate, and fumarate. **C.** Thermal stability of MEZ at different pHs with various small molecules (2mM), measured using differential scanning fluorimeter (DSF). MEZ exhibits greater thermal

stability by DSF at lower pHs but is not significantly affected by the presence of small molecules. All samples contain 10mM $MnCl_2$.

While incubation with ligands had little effect, there is a clear trend between MEZ T_m and pH. At low pHs (6.0-6.5), the T_ms are comparable, ranging from 51-52.5°C while at relatively higher pHs (7.0-7.5), the measured T_ms are lower, ranging from 45.5-49°C. The dependence of MEZ thermal stability on pH suggests that the enzyme may be partially regulated by pH, potentially through stabilizing conformations that favor a particular reaction direction.

The divalent metal and active site

Among MEs studied thus far, all require binding of either Mn^{2+} or Mg^{2+} in the active site to aid in substrate, intermediates and product binding during catalysis as well as to reinforce the structural integrity of the enzyme.[22,23] The metal binding site is highly conserved among both prokaryotes and eukaryotes and comprises two Asp residues and one Glu residue: this conservation is also observed in MEZ that consists of residues Glu240, Asp241, and Asp264 residues (Asp263 may also coordinate with the divalent metal). Additional critical catalytic residues in other MEs studied are Lys169 and Tyr96 (numbering for MEZ). In eukaryotes, an Arg at position 165 in human mitochondrial ME (position 151 in MEZ), is also critical for catalysis and its modification disrupts malate binding without disrupting the cofactor binding. While this Arg is highly conserved among eukaryotic species, it is less conserved among prokaryotes. The substitution of Arg for Ala in in this position for MEZ (Ala151) distinguishes it from the other structurally resolved large subunit prokaryotic *E. coli* ME, where the Arg is present (Arg157).

We solved the X-ray crystal structure of MEZ in its 'open' apo conformation, even though we added high concentrations of malate, $MgCl_2$ and NAD(P)H (2-10mM range) to

the crystallization condition and rescreened in their presence. To fully understand the structure of MEZ in the presence of its substrate, metal and cofactor, we utilized PyMol to manually dock these components into the apo structure based on known modes of small molecule binding in the human mitochondrial ME structure (PDB ID: 1PJ2). After docking, we energy minimized the structural complexes before performing molecular dynamics simulations.

For MEZ, after minimizing the docked Mn^{2+}, malate and $NAD(P)^+$, we observe that within subunit A, the metal ion, malate and $NAD(P)^+$ molecules are not significantly displaced, and malate is stabilized by both the divalent metal and the nicotinamide group of $NAD(P)^+$. However, within subunit B, we observed a greater degree of flexibility and drift of the divalent metal and small molecules (**Figure 4.4A**). In the presence of NAD, in subunit A, MEZ retains contacts along the entire length of the dinucleotide, whereas the contacts of the adenine base with MEZ are lost within subunit B (**Figure 4.4A & B**). The same is observed for NADP where the contacts of MEZ Chain A with NADPH form along the length of molecule, and in contrast, there is no contact between chain B and the adenine base, **Figure 4.4C**. Even though the ribose 2' phosphate forms no contacts with MEZ in Chain B, the subunit B NADP is only slightly displaced from NAPD in subunit A (**Figure 4.4A**).

Conformational Flexibility of MEZ

While some MEs prefer either NAD^+ or $NADP^+$ as an oxidizing agent to catalyze the conversion of malate to pyruvate, MEZ is among those MEs which can use either dinucleotide cofactor. As noted previously, in one MEZ subunit in the ASU, a glycerol molecule is observed within the putative cofactor binding site. To explore the $NAD(P)^+$ binding modes, examine the potential cooperative dynamics between subunits within the

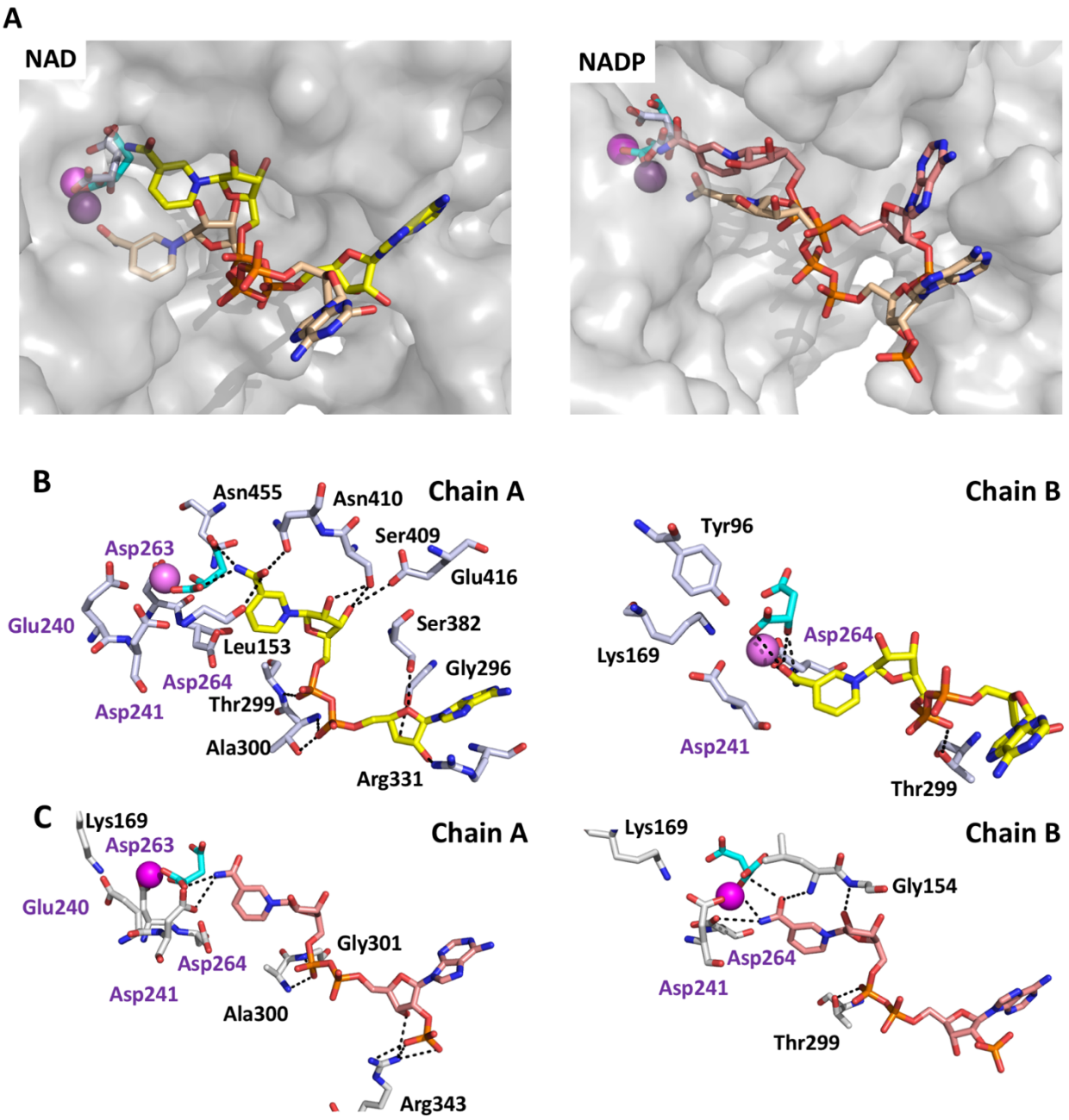


Figure 4.4 Different modes of NAD(P)$^+$ binding after docking with Mn^{2+} and malate A. MEZ is in surface representation (grey) with both subunits superimposed and the divalent metal and small molecules in stick representation. After minimization, you can observe that in the left panel with NAD, the chain 'A' Mn^{2+}, malate and NAD (colored magenta, cyan and yellow, respectively) are in the correct orientation for the reaction to occur, whereas in chain B, the divalent metal (purple) and malate (grey) have shifted and NAD (wheat) is in such a conformation where it is no longer setup to carry out the chemistry as the NAD(P)$^+$ nicotinamide group is no longer in coordinated correctly to malate and in turn malate to Mn^{2+}. Additionally, the chain 'B' NAD adenine base is oriented 180 away from the NAD adenine base group of chain A. In the right panel with NADP a similar observation is seen between chain 'A' and chain ''B'. In chain 'A', NADP (salmon) is in position to carry out the chemistry but in chain 'B' the NADP (salmon) nicotinamide group is oriented

differently with respect to malate and Mn^{2+}; however, its adenine base group is not shifted to the same extent with respect to chain 'A' as that when observing the NAD chain A&B comparison. **B & C.** The polar contacts of NAD (yellow) with protein residues (white sticks) and malate (cyan) are represented by black-dashed lines. **B.** In the right panel is Chain 'A', where one can observe that MEZ makes polar contacts throughout the entire NAD molecule and NAD also H-bonds with malate. However, in chain 'B', Mn^{2+}, malate and NAD are no longer oriented correctly to support the chemistry, and the ribose group of the adenine base no longer has H-bonded to MEZ. **C.** In the right panel is Chain 'A', where one can observe that MEZ makes polar contacts throughout the entire NADP molecule including to the ribose 2' phosphate group (not in NAD), and additionally NADP also H-bonds with malate. However, in chain 'B', Mn^{2+}, malate and NAD are no longer oriented correctly to support the chemistry, and the NADP ribose 2' phosphate group is no longer H-bonded to MEZ.

dimer, and provide insight into the specific interactions that facilitate use of either NAD^+ or $NADP^+$ for catalysis, we ran molecular dynamics simulations of MEZ with malate, NAD^+ or $NADP^+$, and Mn^{2+}. Because metals often present challenges in parameterization of classical molecular dynamics simulations, we initially simulated MEZ in the presence of malate and either NAD^+ or $NADP^+$ only. However, in the absence of Mn^{2+}, we observed that malate is unstable in the binding site (**Figure 4.5**) and dissociates within the first 10 ns of simulation.

To evaluate and compare the overall flexibility of MEZ in its apo (open) and holo (closed) forms, we computed the root mean squared fluctuation (RMSF) of the Cα atom for each residue and averaged the values over the five simulation repeats (**Figure 4.6A**). For domains A, B1 and B2 the RMSF values generally range between 1 and 3 Å and these regions are relatively stable compared to the rest of the protein chain. Domains D, and especially domain C, exhibit the greatest instability; Domain D residue Cα atoms have RMSF values ranging from 1.5-3.5 Å while domain C has the broadest range from 1-7 Å for the apo MEZ simulation. When compared to other systems, domains C and D are particularly flexible; RMSF values computed from simulations of lysozyme are on the order

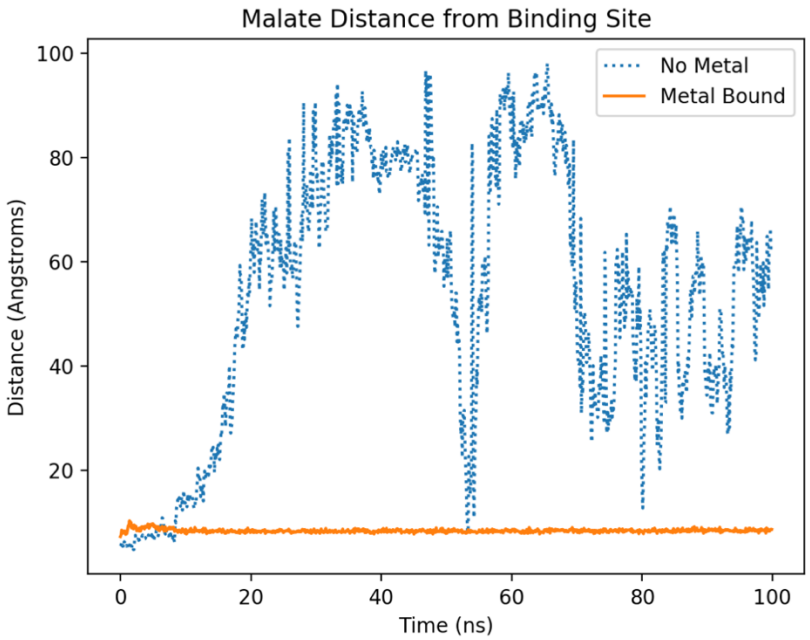


Figure 4.5 Divalent Cation Stabilizes Malate in Binding Site. In molecular dynamics simulations of MEZ with malate and NAD, the malate ligand exits the binding site within the first 10 ns of simulation and briefly returns and exits again around 55 ns. Comparatively, in simulations of MEZ with malate, NAD, and Mn^{2+}, the malate remains bound within the active site of MEZ for the duration of simulations.

of <1 Å,[24] are <2.5 Å for HIV-protease I[25] and range from 1-6 Å for Amyloid β(1-40) dimers, a primary constitutent of amyloid plaques in Alzheimer's disease.[26] Interestingly, the addition of malate, Mn^{2+} and $NAD(P)^+$ to the simulations had no impact on the overall stability of the MEZ protein backbone, apart from some reduction in computed Cα RMSF values for residues between Gln355 and Ala430 (**Figure 4.6B**). This is consistent with our empirical observations that soaking or cocrystallizing substrates, cofactors and metals with MEZ failed to improve the diffraction quality of our crystals.

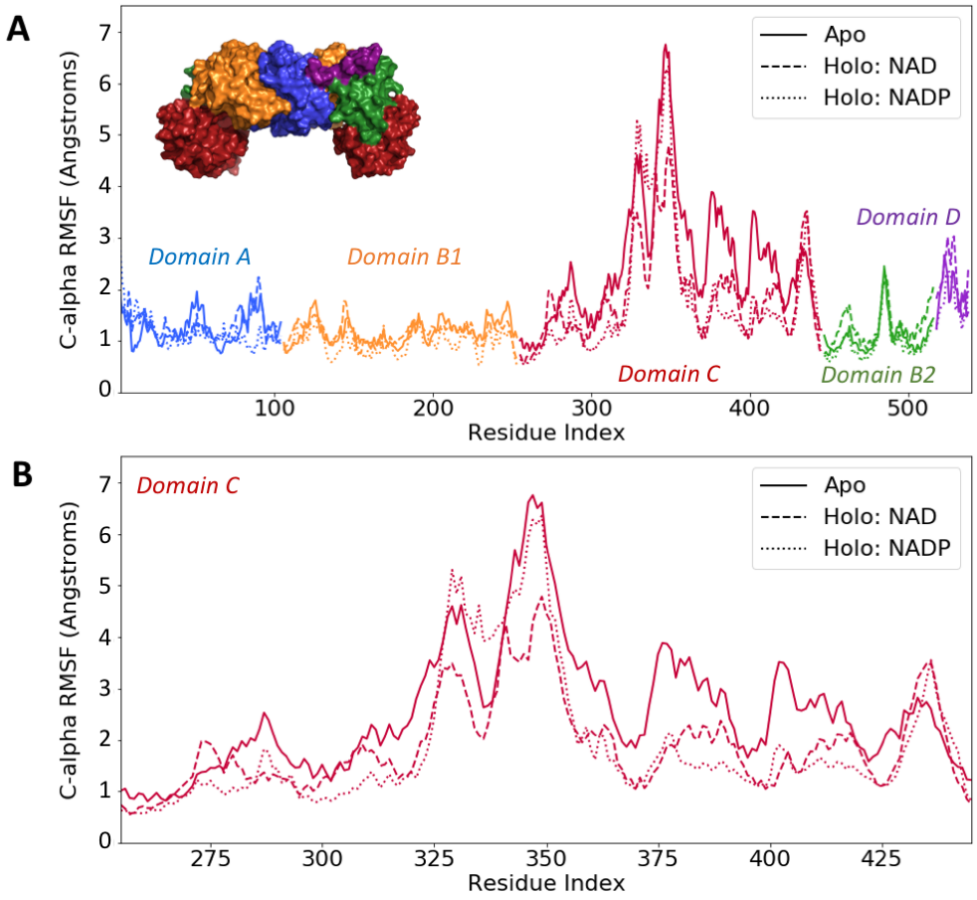


Figure 4.6. Average RMSF per MEZ Residue. Average root mean square fluctuations of the Cα atoms of each residue were calculated from simulations of MEZ (Apo: Mn^{2+} only), MEZ with Mn^{2+}, malate, and NAD (Holo: NAD), and MEZ with Mn^{2+}, malate, and NADP (Holo: NADP). RMSFs are colored according to the previously described structural domains: A (blue), B1 (orange), C (red), B2 (green), and D (purple). Overall, A) RMSFs are comparable among the simulation types, with B) a decreased RMSF of some residues between Gln355 and Ala430 in domain C for the holo simulations.

Fluctuations Between Open and Closed States

Previous studies of MEZ homologs suggest that malic enzymes can adopt open and closed states. As shown in **Figure 4.7**, in our simulations we also observe expansion and contraction of the MEZ active site which can be tracked by recording the distance between Cα atoms on two neighboring helices. Given the observation in the MEZ structure that a glycerol molecule associates in the cofactor binding region in just one subunit of one dimer, we were interested in evaluating the relative stability of cofactor binding in neighboring

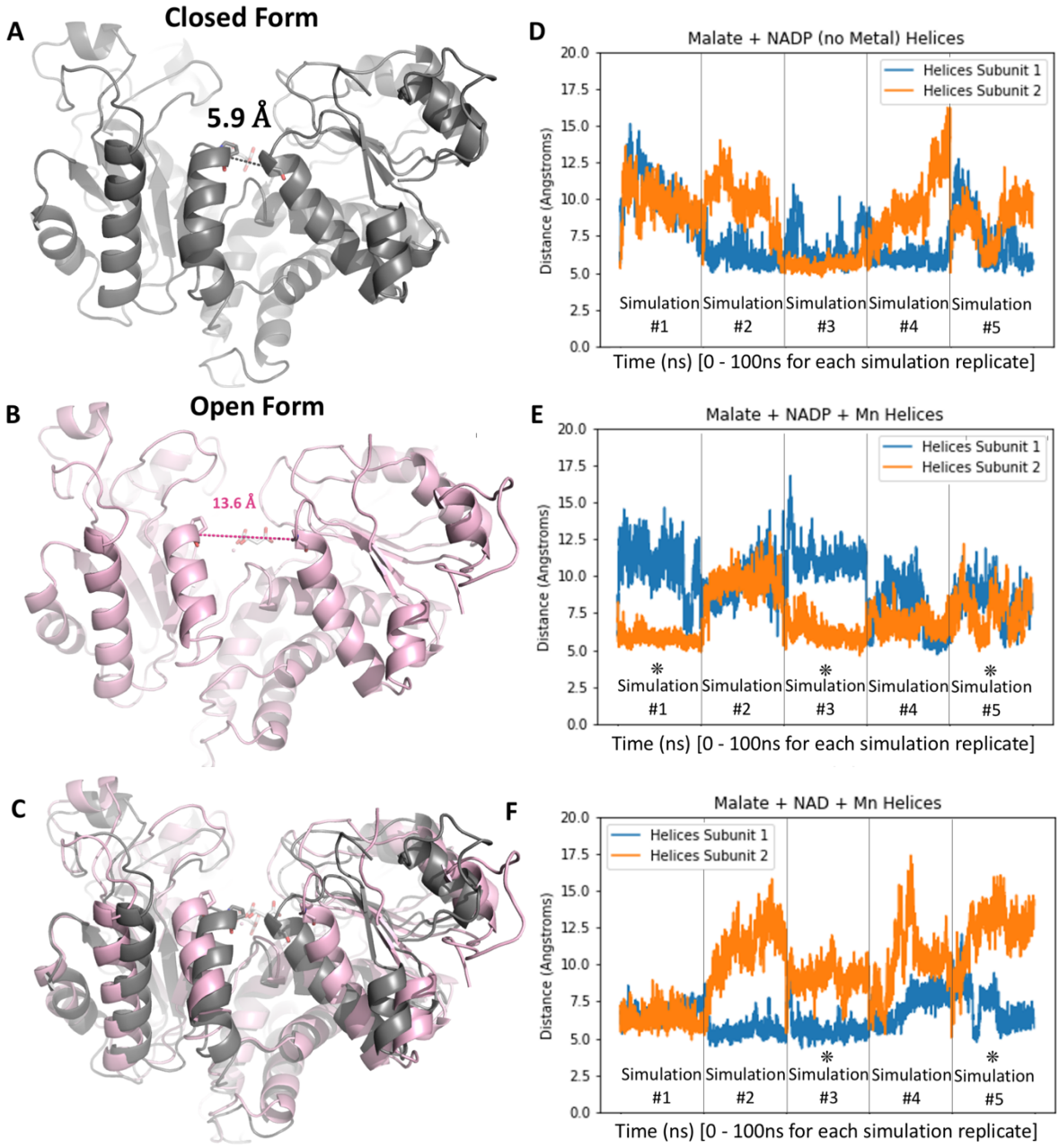


Figure 4.7 Conformational Dynamics of MEZ Active Site. Shown in **A** and **B** are two snapshots from molecular dynamics simulations of MEZ with Mn^{2+}, malate, and NAD^+. For each image, the distance between Pro244 and Ala300 is depicted with a dashed line and labeled in angstroms with the closed form shown in **A** and the open form shown in **B**. In **C** the two MEZ subunits from **A** and **B** have been aligned to provide a more global view of the overall opening of the active site. Plotted in **D-F** are the distances between the Pro244 and Ala300 for the each 100 ns simulation replicate where **D** shows data for MEZ with $NADP^+$ and malate, **E** shows data for MEZ with $NADP^+$, malate, and Mn^{2+}, and **F** gives data for MEZ with NAD^+, malate, and Mn^{2+}. Each simulation replicate represents an independent simulation for which the starting velocities have been randomized and are unique to each replicate. All subunits in all simulations are initiated from the closed form. The ⁕ mark denotes simulations where the $NAD(P)^+$ cofactors and malate substrate remain stably

bound in the active site of both subunits. For these ⁕ simulations, the neighboring MEZ subunits are either 1) stabilized in the opposite form (open vs. closed) as for Simulations 1, or 2) fluctuate together between the open in closed states, suggesting some conformationally cooperative dynamics between the dimer subunits. This cooperative motion suggests that each MEZ subunit catalyzes the reaction in a sequential rather than parallel manner.

subunits as well as whether there exists an interdependence of open and closed states.

For simulations of MEZ, malate and $NADP^+$ without the divalent cation, each subunit in the MEZ dimer fluctuates between the open and closed states independently of the neighboring subunit (**Figure 4.7D**). However, there are no examples among the five 100 ns simulations where the cofactor stably binds to both subunits. Among the five simulations of MEZ in the presence of Mn^{2+}, malate and $NADP^+$, the subunits adopt both closed and open states. In simulations where the cofactor remains stably bound in an active orientation with the nicotinamide group near the malate and Mn^{2+} (Simulation 1, 3, 5, **Figure 4.7**), the subunits stably adopt opposite forms or in the case of Simulation 5 (**Figure 4.7**) appear to rapidly fluctuate between the open and closed states. In the case of MEZ with Mn^{2+}, malate and NAD^+, the subunits also transition between open and closed states. For simulations where the NAD^+ cofactor is stably bound to both subunits in a catalytically relevant orientation (Simulations 3,5, **Figure 4.7**), the subunits adopt opposite open-closed forms.

Because molecular mechanics simulations cannot simulate catalytic mechanisms, specifically the making and breaking of bonds and electron transfer, these simulations do not provide direct insight into the mechanism of malate turnover. It is also worth noting that the 100ns duration of each simulation may be insufficient to precisely capture and characterize any coordinated motion between the neighboring subunits. Nonetheless, we can use this data to develop hypotheses regarding the global dynamics of the protein in

association with its substrate and cofactors, which can be further evaluated with kinetic studies. Generally, we observe a fluctuation between the open and closed forms amongst all simulation types. Furthermore, in simulations with NAD(P)$^+$, malate and Mn^{2+}, where the nicotinamide group of the cofactor remains proximal to the substrate and divalent cation in both subunits, the data suggests the closed form is only stabilized when the neighboring unit stably adopts the open form.

NAD(P)$^+$ Co-factor Binding Modes

To identify the residues that participate in cofactor binding and that specifically facilitate the binding of both NAD$^+$ and NADP$^+$, we selected the simulations for which the cofactor binding was most stable for both MEZ subunits and used the final simulation frame as the reference for analysis. To compare relative stability amongst simulations, we computed and plotted the average RMSF for each cofactor atom in both subunits (**Figure C.1**). For MEZ with Mn^{2+}, malate and NAD$^+$, we selected Simulation 3; while for MEZ with Mn^{2+}, malate and NADP$^+$, we selected Simulation 1.

For MEZ with Mn^{2+}, malate and NAD$^+$, we observe two stable conformations: one elongated and the other collapsed (**Figure 4.8B**). Indeed, multiple conformations of NAD(H) have been observed and described previously in crystallographic, solution NMR, and molecular dynamics studies.[27–30] Specifically, the collapsed form is commonly observed for unbound NAD(H) in solution[31], but has also been observed in holo structures[30], including in the exo-site of human mitochondrial ME (PDB ID: 1PJ3, 1PJL). To identify additional examples of enzymes in complex with the collapsed form of NAD$^+$, we retrieved all instances of NAD$^+$ molecules deposited in the PDB (a total of 4,163 sets of coordinates) and computed Tanimoto scores with OpenEye Scientific's Shape Toolkit using

the collapsed form we observe in our MEZ simulations as the reference structure. The three highest scoring NAD^+ molecules (with the most similar conformation to the collapsed NAD^+ in MEZ) were observed in ADP ribosyl cyclase which cyclizes NAD^+ into cyclic-ADP ribose (PDB ID: 3ZWM), as well as prokaryotic CBS domain proteins (PDB IDs: 2RC3, 4FRY), whereby the domain regulates enzyme and transport activities in response to adenosyl group-binding. Thus this collapsed form is not unprecedented.

Here we suggest that the collapsed form of NAD^+ in complex with MEZ represents an inactive, albeit stable conformation; the nicotinamide group is oriented away from the malate substrate and Mn^{2+} cation, and instead makes polar contacts with the backbone of Glu150 (**Figure 4.8A (i)**). This form is also stabilized by polar contacts with malate, Asp264, Arg414, and the backbones of Leu153 and Asp241. As noted previously, human mitochondrial ME (PDB: 1PJ3) as well as all other structurally resolved large subunit MEs, have a conserved 'GER (Glu163-Arg165)' motif in the active site of human mitochondrial ME, corresponding to residues 148-151 in MEZ. In MEZ (and other small subunit prokaryotic MEs), this motif is substituted for residues 'AEA,' with an alanine residue at position 151. Given that the collapsed form of NAD^+ is proximal to Ala151, it is possible that presence of an Arg residue at this position could interrupt or at minimum, affect the stability of the collapsed conformation in the MEZ active site. Binding of this collapsed form of NAD^+ may be unique among non-'GER' containing large subunit MEs. The elongated conformation of NAD^+ appears to adopt a catalytically relevant orientation (**Figure 4.8A (ii)**), with the nicotinamide group oriented toward the malate and Mn^{2+} cation. This active form is stabilized by several polar interactions with the malate substrate, the sidechains of Arg343, Asp330, Thr383, Ser409, Arg414 as well as the backbone of Gly301 and Asn410.

By comparison, in simulations of MEZ with Mn^{2+}, malate and $NADP^+$, the cofactor adopts a semi-collapsed form (**Figure 4.8B**) which appears to be well oriented for catalysis, with the nicotinamide group proximally located near the malate substrate and Mn^{2+} cation (**Figure 4.8C**). $NADP^+$ makes several polar contacts with malate as well as sidechain and backbone atoms of the surrounding residues. The ligand is stabilized by backbone and/or sidechain interactions with Arg248, Thr268, Thr299, Ala300, Arg331, and Ser382. Notably the kinking of β9 by Pro328, helps orient Arg331 such that it can stabilize the additional 2' phosphate group on ribose (absent in NAD^+). Tanimoto analysis was similarly performed (as described above for NAD^+) using the semi-collapsed structure of $NADP^+$ and all deposited structures of $NADP^+$ in the PDB (3,164 sets of coordinates). The three top-scoring $NADP^+$ molecules are found in unpublished structures of Mtb nicotinic acid mononucleotide adenylyltransferase (NadD) (PDB ID: 4S1O) and *Mycobacterium abseccus* NadD (PDB ID: 4YMI,5VIR).

Methods

MEZ purification

MEZ was purified as described in prior work.[4] Briefly a pNIC28-BSA4 plasmid containing the MEZ gene with an N-terminal His-tag was transformed into BL21(DE3) cells and grown in LB media containing 50 μg ml-1 kanamycin at 37°C. Once the cells reached 0.6-0.8 OD, expression was induced with 1mM IPTG and the temperature was reduced to 18 °C for approximately 18 hours. Cells were harvested by centrifugation for 30 minutes at 5000 rpm. Pellets were washed with STE buffer (10 mM Tris-HCl, pH8, 100 mM NaCl, 1 mM EDTA) and stored at -20°C. Cell pellets were resuspended in Buffer A (50 mM Tris pH 7.6, 500 mM NaCl, 50 mM Imidazole, 0.14 mM BME, 5mM MgCl2, 4% glycerol) containing

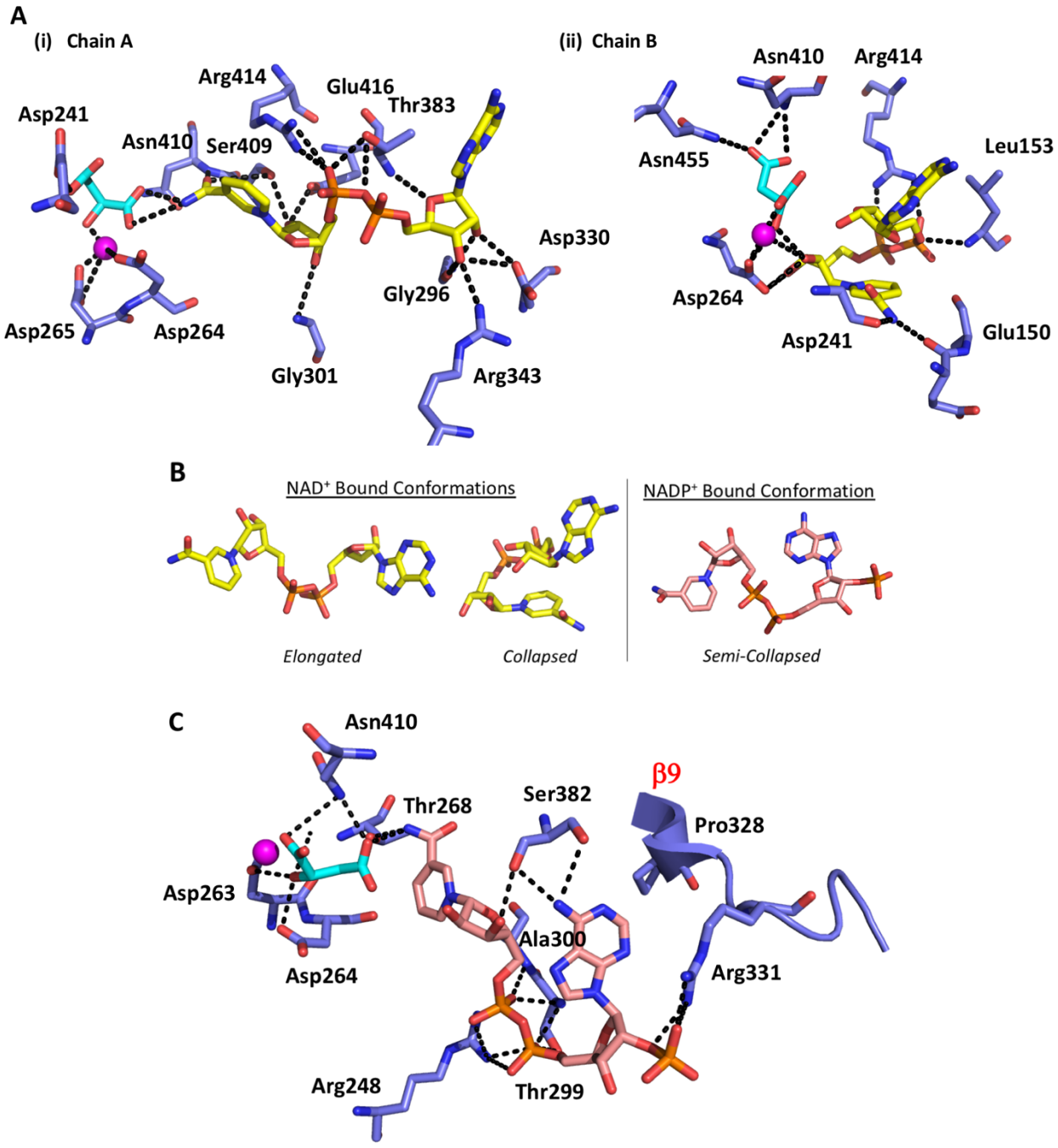


Figure 4.8 Putative NAD(P)$^+$ Binding Modes in MEZ after MD simulations. A. Putative NAD$^+$ Binding Modes in MEZ. In simulations of MEZ with NAD, Mn^{2+}, and malate, the NAD cofactor adopts two distinct binding modes. In **i**) the NAD$^+$ extends along the pocket in an elongated conformation. The nicotinamide interacts with the malate (cyan) while making polar contacts with the backbone of Asn410 and the sidechain of Ser409. The neighboring ribose also forms contacts with Ser409 as well as Glu416 and Gly301. The phosphate groups are coordinated by Arg414, and Thr383 while the ribose near the adenine group can form polar contacts with Gly296. Asp330 and Arg343. In **ii**) the NAD$^+$ cofactor is folded over in a collapsed conformation and occupies a much more limited space of the MEZ binding pocket. The nicotinamide interacts with the backbone of Glu151 and

Asp241 and is oriented away from the Mn^{2+} (magenta) and malate substrate, and thus this likely represents an inactive bound conformation. The phosphate groups are oriented by Arg414 and the backbone of Leu153 while the ribose makes polar contacts with Asp264. **B.** Bound Cofactor Conformations. Representative conformations of the NAD^+ and $NADP^+$ binding modes in simulations of MEZ with Mn^{2+} and malate are shown. The conformations were selected by taking the final frame of the simulation trajectories where the cofactors were most stably bound. The elongated form of NAD^+ represents an active conformation while the collapsed form of NAD^+ occupies the binding site in an inactive formation, as the nicotinamide group is oriented away from the malate substrate. For $NADP^+$ simulations, the cofactor occupies a single semi-collapsed, active form with the nicotinamide oriented near the malate substrate. **C.** Putative $NADP^+$ Binding Mode in MEZ. In simulations of MEZ with $NADP^+$, Mn^{2+}, and malate the NADP cofactor (pink) makes several polar interactions with residues in the MEZ binding pocket. The nicotinamide group is stabilized by interactions with the malate (cyan) substrate and Thr268, while its neighboring ribose forms contacts with Ser382, which also makes a polar contact with the adenine moiety. The phosphate groups form interactions with Arg248, Ala299 and Thr300, respectively and the 2' phosphate group (absent in NAD^+) makes interactions with Arg331. The distinct kinking of the β9 strand in MEZ by Pro328 may favorably position Arg331 for interaction with the $NADP^+$ 2' phosphate group.

lysozyme and 1 mM PMSF (phenylmethylsulfonyl fluoride). Following sonication and centrifugation to remove debris, the cell lysate was loaded onto a nickel-bound His-Trap column and washed with several column volumes of Buffer A, followed by a high salt wash (three column volumes of Buffer A with 2M NaCl). The MEZ was eluted with a gradient of Buffer A and Buffer A containing 500mM Imidazole. MEZ containing fractions were identified via SDS-PAGE analysis and collected and pooled. For DSF and size exclusion chromatography, MEZ was dialyzed into Buffer A overnight at 4°C.

MEZ crystallization

Prior to crystallization studies, MEZ was concentrated and dialyzed via diafiltration (Amicon, 30 kDa MWCO) into 50mM trisodium Citrate-citric acid, 100mM NaCl, pH 6.0, per the results of a pre-crystallization DSF buffer screen performed using the Meltdown program[32], wherein MEZ exhibited a trend of greater thermal stability at lower pH (pHs 5.5-10 were tested). Notably, prior attempts at crystallization (before identification of an optimal protein buffer and pH), yielded few diffracting crystals and none were

reproducible. Crystallization conditions were identified via a collection of Hampton Research crystal screens. Crystal hits were optimized using sparse matrix screens of the pH, salt, and precipitate concentrations. Additionally, protein concentration, drop sizes, and cryo-conditions were varied. The use of crystal seeding, in-situ proteolysis, dehydration, additive screens and the addition of small organic compounds including alanine, glycine, tartronic acid, malic acid and cofactors NADH, NAD^+ and NADPH (2-10mM) via soaking and co-crystallization were all explored. Of the more than 2000 crystals that were harvested, most diffracted no better than 8.0 Å and just two crystals diffracted to 4.0 Å or better.

The crystal condition that gave the highest resolution diffracting crystal was 0.2M NaCl, 0.1M HEPES, pH 7.5, 2mM Alanine, 12% PEG 8000 and the crystal was cryo-frozen in 20% glycerol. The diffraction data set was collected at 70K to a resolution of 3.6Å at ALS and was indexed and integrated in MOSFLM, and scaled in pointless. A solution was generated via molecular replacement using Phaser[33] with an I-TASSER MEZ dimer model of human malic enzyme (PDB ID: 2AW5, chains B+C). The model was refined in Coot and phenix.refine.[33,34]

Molecular Dynamics Simulations

To prepare MEZ for simulations, a dimer of the most complete chains (A & D) was generated by aligning chain D with chain B (the dimeric partner of chain A) in PyMol and submitting the A/D dimer to pdbfixer[35] to model in missing atoms, residues and loops using default protonation states at pH 7.0. Malate, NAD^+ and $NADP^+$ were manually docked in PyMol so as to mimic binding modes observed in MEZ's closest structural homolog by sequence, human mitochondrial ME (PDB ID: 1PJ2).

Cofactors NAD^+ and $NADP^+$ were prepared using previously optimized Amber parameters.[36,37] The Mn^{2+} metal site parameters were built using Amber's MCPB.py program.[38] Malate and Mn^{2+} were prepared using antechamber from AmberTools18[39] with GAFF version 1.7 and AM1-BCC charges. In the MCPB.py input file, the terminal oxygens of residues Asp241 and Asp264 were listed as having bonded pairs with the Mn^{2+} ion and the malate residue was identified as a non-amino acid included in the metal complex. Geometry optimization and force constant calculations were performed with input files generated by MCPB.py using Gaussian09. Output parameters evaluated with parmed[40] and satisfied the criteria outlined in the MCPB.py tutorial.

Each system was solvated using tleap from AmberTools18[39] with a 10 Å rectangular box of TIP3P water, extending from the surface of the protein to the box edge with sodium ions added to neutralize the charge of the system and the protein was parameterized using ff99SB. Minimization proceeded using sander from Amber14 with steepest descents running for 20000 steps, followed by heating from 100 to 300 K with a Berendsen thermostat and a constant volume for 25000 time steps of 2 fs. Equilibration continued using sander for 500000 time steps of 2 fs under constant pressure via a Monte Carlo barostat and positional restraints initially applied on all non-water atoms at 50 kcal/mol/$Å^2$ and progressively increased in increments of 5 kcal/mol/$Å^2$ over ten 50000-step segments. The resulting topology and coordinate files for each system were used as inputs for production MD simulations. Production simulations were executed in OpenMM 7.1.0[41] using a Langevin integrator with a 2 fs time step and a friction coefficient of 10 ps–1. For each of four systems (MEZ only, MEZ/malate/NAD^+, MEZ/Mn^{2+}/malate/NAD^+, and

MEZ/Mn^{2+}/malate/$NADP^{+}$), five 100 ns simulations, with shared equilibration, were initiated with random starting velocities.

Molecular Dynamics Analysis

Analysis of trajectory data was performed using MDTraj 1.9.3. In order to systematically track the stability of malate in the binding site, we computed the distance between C1 atom of malate and the Cα atom of Asn455 for the duration of each simulation. Similarly, the relative positioning of $NAD(P)^{+}$ in relation to the active site was tracked by recording the distance between the C4 atom of the nicotinamide group of $NAD(P)^{+}$, which is reduced to form NAD(P)H, and the Cα atom of Asn455. To quantify differences in the conformational stability of MEZ alone or in complex with malate and $NAD(P)^{+}$, we computed and averaged the RMSF values of each Cα atom in the protein backbone for all simulations and plotted them as shown in (**Figure 4.6**). Visual inspection of the trajectories revealed open and closed states for the MEZ subunits; in order to assign the relative state, we tracked the relative proximity of two neighboring helices by computing the distance between Cα atoms of Pro244 and Ala300 (**Figure 4.7**). For analysis of $NAD(P)^{+}$ binding modes, we identified one simulation among the five replicates of MEZ/Mn^{2+}/malate/NAD and MEZ/Mn^{2+}/malate/NADP for which the cofactor was most stable. Sustained binding of the cofactor in the active site of MEZ was assessed by plotting the distance between the $NAD(P)^{+}$ carbon atom at position 4 of the nicotinamide group and the Cα atom of a stable residue in the binding pocket, Asn455 (**Figure C.2**). Relative stability of the cofactor amongst the replicates was imputed by comparing the RMSF for each atom in the NAD^{+} or $NADP^{+}$ ligand (**Figure C.1**); the final frame was used to examine the contacts between the protein and each cofactor.

Protein Thermal Stability Measurements

Protein thermal stability was determined using differential scanning fluorimetry (DSF). For DSF measurements, proteins were incubated with 25 or 40 μM SYPRO orange dye in 20 mM sodium phosphate (pH 7.4), 150 mM NaCl. Samples were heated from 25 °C to 96 °C at 1 °C/min using an Mx3005P QPCR machine (Agilent Technologies). The dye was excited at 492 nm, and fluorescence emission was monitored at 610 nm. Melting curves were obtained in duplicate, and each experiment was conducted independently three times. Melting temperatures (T_ms) were determined using nonlinear regression to determine melting-curve inflection points.

Size exclusion chromatography

Size exclusion chromatography for MEZ was run on a S75 HiLoad 16/600 column at 0.5mL/min with protein standards from Bio-Rad to determine the oligomeric state of MEZ in solution.

Discussion

Many species have multiple copies of genes encoding for ME[16,20,42], which underscores its physiological importance in carbon metabolism as well as in providing NAD(P)H for reducing energy. The redundancy of ME genes can also be attributed to the need for differential regulation of its expression and activity in response to environmental and metabolic stimuli. For example, *Bacillus subtilis* has four distinct ME isoforms ranging from 45-60kDa: YtsJ, MleA, MalS and MaeA. While all are dispensable for growth in glucose minimal media, YtsJ - a chimeric ME - is upregulated in malate medium, plays a major role in the utilization of malate for growth, and has a 70-fold greater activity with $NADP^+$ over

NAD^+ as a cofactor.[42] Conversely, MleA (~48kDa) is differentially expressed in complex media depending on the growth phase of the cells, and the enzyme has no detectable activity with $NADP^+$.[42] The two closest *B. subtilis* ME homologs to MEZ, by sequence identity, are MalS and MaeA; these enzymes show no specific preference for NAD^+ versus $NADP^+$ but have slightly enhanced catalytic efficiency in the presence of NAD^+. All four isoforms can decarboxylate oxaloacetate (OAA), an intermediate metabolite in the conversion of malate to pyruvate.

Many eukaryotes also possess multiple ME isoforms, which are often distinct from each other in their cofactor preference for NAD^+ or $NADP^+$. Humans possess three MEs; one NADP-dependent ME is found in the cytosol while the other two, one with $NADP^+$ specificity and the other $NAD(P)^+$ activity, are localized to the mitochondria. Analogously, photosynthetic eukaryotes possess both plastidic C4- associated tetrameric ME that plays a role in photosynthesis, as well as a non-C4 dimeric isoform of ME.[43–45] For Mtb, however, MEZ is the only ME type protein, and as such, it is likely subject to more distinct mechanisms of regulation.

Unlike other characterized large subunit MEs, MEZ forms a stable dimer

Generally, eukaryotic MEs are known to conform to a morpheein (adopting multiple homo-oligomeric forms) model of allosteric regulation, wherein the conformational state of the monomeric or dimeric subunits reforms in the presence of particular ligands, and thereby stabilizes the higher-order tetrameric state.[44–46]

Stabilization of the higher order tetrameric state for larger subunit type MEs is facilitated by the presence of a C- or N- terminal extension.[19,44] Mutations in the C-terminal extensions of pigeon ME and human cytosolic ME disrupt formation of the tetramer

assembly.[15,18] In photosynthetic eukaryotes, the N terminal tail extension is long (50-60 residues) when compared to the moderately extended C-termini of non-photosynthetic eukaryotes (20-30 residues). Structurally, both the N and C termini localize at the dimer and tetramer interfaces (**Figure 4.1B**). In non-photosynthetic eukaryotes, the C terminus extends up and away from the dimer along the middle of the solvent facing region of the adjacent dimer subunit, while in photosynthetic eukaryotes, the N-terminus forms a β strand that interacts with the neighboring dimers' N-termini to form a pair of antiparallel β sheets at the vertex of the tetramer interface (**Figure 4.1B**). *E. coli* ME (SfcA), MEZ's closest structural homolog, has a modest 7 residue extension at both the N- and C- termini (PDB ID: 6AGS). Structurally these two extensions, which each form a β-strand, come together in a parallel β-sheet near the canonical tetrameric interface; although the crystal structure suggests formation of a dimer, by SEC, the *E. coli* SfcA ME assembles as a tetramer.[20] Conversely, MEZ has neither an N- nor C-terminal extension, and the canonical tetramer is not formed within the ASU nor can it be generated in the crystallographic symmetry; the assembly of MEZ as a stable dimer is also supported by SEC data presented in this work.

MEZ contains a conserved metal binding site

All MEs studied thus far require binding of either Mn^{2+} or Mg^{2+} in the active site to stabilize the substrate, intermediates and product during catalysis, and as well as to reinforce the structural integrity of the enzyme. Previous studies of human cytosolic ME demonstrate that binding of a divalent cation confers a conformational change that protects the monomer ME against aggregation in acidic conditions or in the presence of chemical denaturants.[48] The metal binding site is highly conserved among both

prokaryotic and eukaryotic MEs and comprises two aspartate residues and one glutamate residue: Glu240, Asp241, and Asp264, as numbered in Mtb. Studies of other MEs reveal that binding of trivalent Lu^{3+} in this site results in enzymatic inhibition[49]; similarly binding of other metal ions Zn^{2+}, Cu^{2+}, or Fe^{2+} drive conformational remodeling that is unfavorable for activity. In MD simulations of MEZ, we observe that Mn^{2+} is required to stabilize the malate substrate in the active site (**Figure 4.5**).

MEZ active site lacks Arg conserved in large subunit MEs

Additional critical catalytic residues for ME activity include Lys169 and Tyr96 (Mtb numbering).[19] These residues are highly conserved among all species; however, structural elucidation of the small subunit prokaryotic MEs has revealed that, while these residues are conserved in the primary sequence, the structural organization is such that the Lys residue is absent from the active site. Consequently, in small subunit prokaryotic MEs, formation of the dimer is critical for catalysis, with one monomer forming the metal binding site with the Lys and the neighboring monomer contributing the catalytic Tyr.[13] In human mitochondrial ME, pigeon cytosolic ME and maize ME, Arg165 is also critical for catalysis and its modification disrupts malate binding without disrupting the cofactor binding. While Arg165 (Ala151 in MEZ) is highly conserved among eukaryotic species, it is less conserved among prokaryotes. The absence of this residue in MEZ distinguishes it from the only other structurally resolved large subunit prokaryotic ME, *E. coli* ME, where Arg165 is present (PDB ID: 6AGS). MEs lacking Arg165 possess a Gly or Ala substitution at the Arg165 site and often require a monovalent cation such as K^+ or NH_4^+, in addition to the Mn^{2+} or Mg^{2+}, for catalysis.[50–54] We suspect that the monovalent cation acts as a positively

charged proxy for the arginine residue to help stabilize the malate substrate. While the catalytic efficiency of ME likely suffers as a result of this Arg substitution, it is worth nothing that replacement with a smaller, aliphatic residue may permit promiscuous binding of other substrates, inhibitors, or cofactors that would otherwise be sterically and/or electrostatically excluded. In organisms like Mtb, where MEZ is the only ME, the lack of this conserved arginine residue may facilitate greater regulatory control of the enzyme. Furthermore greater substrate promiscuity could also help drive NAD(P)H generation under various metabolic and environmental stimuli.

MEZ may be partially regulated by pH

The regulation of MEs by pH has been studied in other organisms. In eukaryotic species, it has been observed that acidic pH drives the dissociation of the tetramer/dimer equilibrium toward dimeric and monomeric species; relatedly, disruption of tetramer formation has been linked to substrate inhibition of ME by malate at pHs less than 7.0.[55–59] Recently, the molecular basis of pH dependent inhibition of the photosynthetic C4-ME by malate, particularly during nocturnal periods where the stromal pH drops to 7.0-7.5 from pH 8.0, was determined to be related to destabilization of the higher order oligomeric state.[44] In our DSF studies, we also observe that MEZ is stabilized at lower pHs; however, by SEC the oligomeric state persists as a dimer at both pH 7.5 (**Figure 4.4B**) and pH 6.0 (data not shown).

The biological relevance of the increased stability of MEZ at lower pHs may be tied to the fact that *mez* is part of the *phoPR* regulon, which is upregulated at low pH.[9] The PhoPR system regulon includes genes essential for the biosynthesis of complex lipids required for Mtb virulence.[9] In future work we will evaluate how pH affects the kinetics of

MEZ and whether pH drives the gluconeogenic (forward) or anaplerotic (reverse) reactions, or shifts the NAD(P)$^+$ cofactor preference.

MEZ is allosterically regulated by fumarate and succinate

Many eukaryotic MEs have been shown to partake in cooperative or anti-cooperative kinetics with respect to malate binding in neighboring subunits. In human mitochondrial ME and *Ascaris suum* ME, fumarate acts as an allosteric activator and supplants the cooperative kinetic effects of substrate binding. [56,60] In human mitochondrial ME, the fumarate binds at two symmetric allosteric sites at the dimer interface; these two sites are coordinated by Arg67 and Arg91 residues from each subunit. While our MEZ structure has no ligands bound at these sites, the structurally corresponding residues, Arg51 and Arg75 in MEZ are organized in a similar configuration at the dimer interface, suggesting that it may be also be modulated by binding of ligands at the two symmetric allosteric sites. Kinetic studies of the homologous *E. coli* ME, suggest that its activity is upregulated by addition of aspartate and inhibited by OAA, CoA, palmotyl-CoA, and acetyl-phosphate.[20] Fumarate, glutamate, succinate, glycine and alanine have no effect on its activity.

In gluconeogenic kinetic studies of MEZ, the small molecules fumarate, succinate and alanine boost activity by 2-4 fold (**Figure 4.4B**); MEZ shows a slight preference for fumarate, which could be related to its relative rigidity compared to succinate, where fumarate binding might be more energetically favored due to reduced entropic costs. Modulation by alanine may be tied to activity of alanine dehydrogenase (ALD), which converts pyruvate (the product of MEZ) to alanine along with the concomitant oxidation of NADH to NAD$^+$. Like MEZ, the gene for ALD is upregulated under hypoxic

conditions[61]; it is possible that MEZ play a role in preserving redox balance and supplying NADH when ALD activity increases.

The plasticity of MEZ: its biological relevance and druggability

Generally speaking, it has been observed that structural protein disorder is associated with either 1) increased promiscuity and/or 2) a disorder-to-order transition upon protein-ligand interactions, or 3) a disorder-to-order transition upon protein-protein interactions.[62] Given that we do not observe a decrease in disorder nor an increase in thermal stability (**Figure 4.4A**) in our simulations and biochemical analyses of MEZ in the presence of cofactors and substrate, there is little evidence to support a disorder-to-order transition upon ligand binding. At this point, it remains to be seen whether MEZ forms protein-protein interactions that facilitate a transition to a more ordered state. However, a recent study demonstrated that cytosolic human ME1 activates and forms a stable hetero-complex with 6-Phosphogluconate dehydrogenase (6GPD), which notably also generates NADPH and may help maintain redox homeostasis.[63] Mtb has an analogous coenzyme F420-dependent 6-Phosphogluconate dehydrogenase (FGD1), which generates reduced coenzyme F420. While FGD1 is structurally distinct from human 6GPD,[64] MEZ may interact with FGD1 or have other protein-interaction partners. Ultimately, we suspect that the biological significance of the structural plasticity of MEZ may be more directly related to facilitating cofactor or substrate promiscuity such that the enzyme can help generate reductants for lipid synthesis and the maintenance of redox homeostasis in response to various metabolic and environmental stressors.

Regarding the druggability of MEZ, we note some distinct structural differences between it and its human ME homologs. In our simulations of MEZ with NAD^+ and $NADP^+$,

we observe that the NAD^+ cofactor can adopt multiple stable binding modes (**Figure 4.8**). In its compact form, the ribose group near the NAD^+ adenine motif forms polar contacts with the Mn^{2+} ion and malate substrate, but its nicotinamide group is oriented away from the active site and thus, is not suitably positioned for catalysis (**Figure 4.8A (ii)**). In order to accommodate this orientation, MEZ adopts a semi-open conformation wherein the β-hairpin turn (β3/β4) of residues 150-154 is pushed further away from the binding pocket as its backbone makes interactions with the folded NAD^+ molecule. Notably, in other ME structures like that of human mitochondrial ME, the residues at the apex of this β-hairpin form a helix that is flanked by the Arg165 residue, which anchors this turn to the ME active site by forming polar contacts with malate (PDB ID: 1PJ2). In MEZ, where the arginine is substituted for an alanine (position 151), this β-hairpin has greater flexibility as it no longer possesses a positively charged sidechain to coordinate the malate substrate. We suspect that the stable binding of NAD^+ in a compact form is unique to MEZ and other prokaryotic MEs that lack the conserved Arg165 residue. As such, this structural distinction may present an opportunity to design a bacteria-specific inhibitor for MEs that does not bind human MEs.

References

1. Dowdle, W. R. & Centers for Disease Control (CDC). A strategic plan for the elimination of tuberculosis in the United States. *MMWR supplements* **38**, 1–25 (1989).

2. Nahid, P. *et al.* Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. *Clin Infect Dis* **63**, 853–867 (2016).

3. World Health Organization. *Global Tuberculosis Report 2019.* (World Health Organization, 2019).

4. Basu, P. *et al.* The anaplerotic node is essential for the intracellular survival of Mycobacterium tuberculosis. *J. Biol. Chem.* **293**, 5695–5704 (2018).

5. Ojha, A. K., Trivelli, X., Guerardel, Y., Kremer, L. & Hatfull, G. F. Enzymatic hydrolysis of trehalose dimycolate releases free mycolic acids during mycobacterial growth in biofilms. *J. Biol. Chem.* **285**, 17380–17389 (2010).

6. Owens, C. P. *et al.* The Mycobacterium tuberculosis secreted protein Rv0203 transfers heme to membrane proteins MmpL3 and MmpL11. *J. Biol. Chem.* **288**, 21714–21728 (2013).

7. Pacheco, S. A., Hsu, F.-F., Powers, K. M. & Purdy, G. E. MmpL11 Protein Transports Mycolic Acid-containing Lipids to the Mycobacterial Cell Wall and Contributes to Biofilm Formation in Mycobacterium smegmatis. *J Biol Chem* **288**, 24213–24222 (2013).

8. Fay, A. *et al.* Two Accessory Proteins Govern MmpL3 Mycolic Acid Transport in Mycobacteria. *mBio* **10**, (2019).

9. Baker, J. J., Johnson, B. K. & Abramovitch, R. B. Slow growth of Mycobacterium tuberculosis at acidic pH is regulated by phoPR and host-associated carbon sources. *Mol Microbiol* **94**, 56–69 (2014).

10. Ratledge, C. The role of malic enzyme as the provider of NADPH in oleaginous microorganisms: a reappraisal and unsolved problems. *Biotechnol Lett* **36**, 1557–1568 (2014).

11. Rodriguez, E., Navone, L., Casati, P. & Gramajo, H. Impact of Malic Enzymes on Antibiotic and Triacylglycerol Production in Streptomyces coelicolor. *Appl Environ Microbiol* **78**, 4571–4579 (2012).

12. Hernández, M. A. & Alvarez, H. M. Increasing lipid production using an NADP+-dependent malic enzyme from Rhodococcus jostii. *Microbiology* **165**, 4–14 (2019).

13. Alvarez, C. E. *et al.* The crystal structure of the malic enzyme from Candidatus Phytoplasma reveals the minimal structural determinants for a malic enzyme. *Acta Cryst D* **74**, 332–340 (2018).

14. Xu, Y., Bhargava, G., Wu, H., Loeber, G. & Tong, L. Crystal structure of human mitochondrial NAD(P)(+)-dependent malic enzyme: a new class of oxidative decarboxylases. *Structure* **7**, 877–889 (1999).

15. Hsieh, J. Y. *et al.* Structural characteristics of the nonallosteric human cytosolic malic enzyme. *Biochim.Biophys.Acta* (2014) doi:10.2210/pdb3wja/pdb.

16. Bukato, G., Kochan, Z. & Swierczyński, J. Different regulatory properties of the cytosolic and mitochondrial forms of malic enzyme isolated from human brain. *Int. J. Biochem. Cell Biol.* **27**, 1003–1008 (1995).

17. Alvarez, C. E. *et al.* Molecular adaptations of NADP-malic enzyme for its function in C 4 photosynthesis in grasses. *Nat. Plants* **5**, 755–765 (2019).

18. Chang, H.-C. & Chang, G.-G. Involvement of Single Residue Tryptophan 548 in the Quaternary Structural Stability of Pigeon Cytosolic Malic Enzyme. *J. Biol. Chem.* **278**, 23996–24002 (2003).

19. Chang, G.-G. & Tong, L. Structure and Function of Malic Enzymes, A New Class of Oxidative Decarboxylases. *Biochemistry* **42**, 12721–12733 (2003).

20. Bologna, F. P., Andreo, C. S. & Drincovich, M. F. Escherichia coli Malic Enzymes: Two Isoforms with Substantial Differences in Kinetic Properties, Metabolic Regulation, and Structure. *J Bacteriol* **189**, 5937–5946 (2007).

21. Wang, B. *et al.* Biochemical properties and physiological roles of NADP-dependent malic enzyme in Escherichia coli. *J Microbiol.* **49**, 797–802 (2011).

22. Chang, H.-C., Chou, W.-Y. & Chang, G.-G. Effect of metal binding on the structural stability of pigeon liver malic enzyme. *J. Biol. Chem.* **277**, 4663–4671 (2002).

23. Jernejc, K. & Legiša, M. The influence of metal ions on malic enzyme activity and lipid synthesis in Aspergillus niger. *FEMS Microbiol Lett* **217**, 185–190 (2002).

24. Borkotoky, S. & Murali, A. A computational assessment of pH-dependent differential interaction of T7 lysozyme with T7 RNA polymerase. *BMC Structural Biology* **17**, 7 (2017).

25. Costa, M. G. *et al.* Impact of M36I polymorphism on the interaction of HIV-1 protease with its substrates: insights from molecular dynamics. *BMC Genomics* **15**, S5 (2014).

26. Watts, C. R., Gregory, A. J., Frisbie, C. P. & Lovas, S. Structural properties of amyloid β(1-40) dimer explored by replica exchange molecular dynamics simulations. *Proteins* **85**, 1024–1045 (2017).

27. Parthasarathy, R. & Fridey, S. M. Conformational variability of NAD+ in the free and bound states: a nicotinamide sandwich in NAD+ crystals. *Science* **226**, 969–971 (1984).

28. Guillot, B., Lecomte, C., Cousson, A., Scherf, C. & Jelsch, C. High-resolution neutron structure of nicotinamide adenine dinucleotide. *Acta Cryst D* **57**, 981–989 (2001).

29. Cui, Q. & Karplus, M. Molecular Properties from Combined QM/MM Methods. 2. Chemical Shifts in Large Molecules. *J. Phys. Chem. B* **104**, 3721–3743 (2000).

30. Tanner, J. J., Tu, S. C., Barbour, L. J., Barnes, C. L. & Krause, K. L. Unusual folded conformation of nicotinamide adenine dinucleotide bound to flavin reductase P. *Protein Sci.* **8**, 1725–1732 (1999).

31. Smith, P. E. & Tanner, J. J. Conformations of nicotinamide adenine dinucleotide (NAD(+)) in various environments. *J. Mol. Recognit.* **13**, 27–34 (2000).

32. Rosa, N. *et al.* Meltdown: A Tool to Help in the Interpretation of Thermal Melt Curves Acquired by Differential Scanning Fluorimetry. *Journal of Biomolecular Screening* **20**, 898–905 (2015).

33. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

34. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

35. Eastman, P. *et al.* OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **9**, 461–469 (2013).

36. Ryde, U. Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion. *Proteins* **21**, 40–56 (1995).

37. Ryde, U. On the role of Glu-68 in alcohol dehydrogenase. *Protein Sci* **4**, 1124–1132 (1995).

38. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).

39. Case, D. A. *et al. AmberTools17*. (2017).

40. Swails, J. *et al. Parmed*. (2015).

41. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **13**, e1005659 (2017).

42. Lerondel, G., Doan, T., Zamboni, N., Sauer, U. & Aymerich, S. YtsJ Has the Major Physiological Role of the Four Paralogous Malic Enzyme Isoforms in Bacillus subtilis. *Journal of Bacteriology* **188**, 4727–4736 (2006).

43. Rao, S. R., Kamath, B. G. & Bhagwat, A. S. Identification of Metal-Binding Site of Maize NADP-Malic Enzyme by Affinity Cleavage by Fe2+-Ascorbate. in *Photosynthesis: Mechanisms and Effects: Volume I–V: Proceedings of the XIth International Congress on Photosynthesis, Budapest, Hungary, August 17–22, 1998* (ed. Garab, G.) 3591–3594 (Springer Netherlands, 1998). doi:10.1007/978-94-011-3953-3_838.

44. Alvarez, C. E. *et al.* Molecular adaptations of NADP-malic enzyme for its function in C4 photosynthesis in grasses. *Nat. Plants* **5**, 755–765 (2019).

45. Grover, S. D. & Wedding, R. T. Modulation of the activity of NAD malic enzyme from Solanum tuberosum by changes in oligomeric state. *Archives of Biochemistry and Biophysics* **234**, 418–425 (1984).

46. Gerald E, E. & Carlos S, A. NADP-malic enzyme from plants. *Phytochemistry* **31**, 1845–1857 (1992).

47. Hsieh, J.-Y., Chen, S.-H. & Hung, H.-C. Functional roles of the tetramer organization of malic enzyme. *J. Biol. Chem.* **284**, 18096–18105 (2009).

48. Chang, H.-C. *et al.* Metal Ions Stabilize a Dimeric Molten Globule State between the Open and Closed Forms of Malic Enzyme. *Biophys J* **93**, 3977–3988 (2007).

49. Kuo, C.-W., Hung, H.-C., Tong, L. & Chang, G.-G. Metal-Induced reversible structural interconversion of human mitochondrial NAD(P)+-dependent malic enzyme. *Proteins* **54**, 404–411 (2004).

50. Lamed, R. & Zeikus, J. G. Thermostable, ammonium-activated malic enzyme of Clostridium thermocellum. *Biochimica et Biophysica Acta (BBA) - Enzymology* **660**, 251–255 (1981).

51. Garrido-Pertierra, A., Marcos, C. M., Fernandez, M. M. & Ruiz-Amil, M. Properties and function of malate enzyme from Pseudomonas putida. *Biochimie* **65**, 629–635 (1984).

52. Driscoll, B. T. & Finan, T. M. Properties of NAD(+)- and NADP(+)-dependent malic enzymes of Rhizobium (Sinorhizobium) meliloti and differential expression of their genes in nitrogen-fixing bacteroids. *Microbiology (Reading, Engl.)* **143 (Pt 2)**, 489–498 (1997).

53. Gourdon, P., Baucher, M.-F., Lindley, N. D. & Guyonvarch, A. Cloning of the Malic Enzyme Gene from Corynebacterium glutamicum and Role of the Enzyme in Lactate Metabolism. *Appl Environ Microbiol* **66**, 2981–2987 (2000).

54. Spaans, S. K., Weusthuis, R. A., van der Oost, J. & Kengen, S. W. M. NADPH-generating systems in bacteria and archaea. *Front. Microbiol.* **6**, (2015).

55. Tronconi, M. A. *et al.* Allosteric substrate inhibition of Arabidopsis NAD-dependent malic enzyme 1 is released by fumarate. *Phytochemistry* **111**, 37–47 (2015).

56. Hsieh, J.-Y. *et al.* Fumarate Analogs Act as Allosteric Inhibitors of the Human Mitochondrial NAD(P)+-Dependent Malic Enzyme. *PLoS ONE* **9**, e98385 (2014).

57. Chang, G. G., Huang, T. M. & Chang, T. C. Reversible dissociation of the catalytically active subunits of pigeon liver malic enzyme. *Biochem. J.* **254**, 123–130 (1988).

58. Hsieh, J.-Y., Shih, W.-T., Kuo, Y.-H., Liu, G.-Y. & Hung, H.-C. Functional Roles of Metabolic Intermediates in Regulating the Human Mitochondrial NAD(P) + -Dependent Malic Enzyme. *Sci Rep* **9**, 1–14 (2019).

59. Murugan, S. & Hung, H.-C. Biophysical Characterization of the Dimer and Tetramer Interface Interactions of the Human Cytosolic Malic Enzyme. *PLOS ONE* **7**, e50143 (2012).

60. Karsten, W. E., Pais, J. E., Rao, G. S. J., Harris, B. G. & Cook, P. F. Ascaris suum NAD-malic enzyme is activated by L-malate and fumarate binding to separate allosteric sites. *Biochemistry* **42**, 9712–9721 (2003).

61. Giffin, M. M., Modesti, L., Raab, R. W., Wayne, L. G. & Sohaskey, C. D. ald of Mycobacterium tuberculosis Encodes both the Alanine Dehydrogenase and the Putative Glycine Dehydrogenase. *Journal of Bacteriology* **194**, 1045–1054 (2012).

62. Lieutaud, P. *et al.* How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord Proteins* **4**, (2016).

63. Yao, P. *et al.* Evidence for a direct cross-talk between malic enzyme and the pentose phosphate pathway via structural interactions. *J. Biol. Chem.* **292**, 17113–17120 (2017).

64. Bashiri, G., Squire, C. J., Moreland, N. J. & Baker, E. N. Crystal Structures of F420-dependent Glucose-6-phosphate Dehydrogenase FGD1 Involved in the Activation of the Anti-tuberculosis Drug Candidate PA-824 Reveal the Basis of Coenzyme and Substrate Binding. *J. Biol. Chem.* **283**, 17531–17541 (2008).

CHAPTER 5: Summary and Conclusions - Complementing structural information using computational methods for drug-design

Static information provided by structural methods like X-ray crystallography and cryo-EM establishes the critical foundation required for understanding the structure-function relationship of proteins. Frequently, additional information regarding protein dynamics can also be gleaned directly from structural data: B-factors and unresolved electron density highlight the relative flexibility of certain regions; differences amongst subunits within a single ASU can reveal multiple low energy conformations; and relatedly, it is not uncommon to obtain both apo and holo (ligand-bound) conformations from a single crystal. For X-ray crystallography, there are concerted efforts to harness molecular simulations in order to interpret diffuse scattering in X-ray diffraction data.[1,2] Typically, this weak electron density, which is dispersed around and distinct from high intensity Bragg's peaks, is ignored; however, in theory, it can provide insight into protein dynamics within the crystal. In the field of cryo-EM, advances in imaging acquisition and processing are facilitating the collection and resolution of distinct conformational states that provide insight into the global motions of the protein.[3] This is particularly compelling in driving our mechanistic understanding of multi-subunit protein complexes[4] and membrane protein channels and pores.[5] Furthermore, additional advances in nuclear magnetic resonance and molecular dynamics (MD) simulations also help complement structural data to give a more comprehensive picture of a protein's behavior. Determining the structure and relevant motions of a protein provides valuable mechanistic and geometric insight that can be used to further inform drug design efforts. In the chapters preceding, I have presented work that both seeks to improve upon methods that shed light on protein

dynamics as well as to practically apply them in further efforts towards developing novel therapeutics to treat TB.

In Chapter 2, I describe a novel method for enhancing side chain sampling in molecular simulations. Side chain rearrangements in binding pockets are actually critical and common.[6] Effectively sampling these motions during molecular simulations is crucial for accurately computing thermodynamic properties, particularly binding free energies between ligands and receptors.[7–10] Prior knowledge of side chain conformations is also relevant when performing virtual screening via docking. When available, X-ray crystal structures of proteins complexes with ligands can provide information about preferred side chain orientations. However, depending on the particular system and the quality of the structural data, there may still be ambiguities in the rotameric states. Among the 144,000+ structures deposited in the Protein Data Bank (PDB) with resolutions up to 5Å, more than 20% fall in the range of 2.5Å - 5Å[11] wherein the accurate modeling of side chain rotamers (based on the available electron density map) deteriorates. Even in high resolution structures, dose-dependent radiation damage during data collection can lead to ambiguous or worse, misleading density maps resulting in misinterpretations.[12–14] Specifically, radiation damage during cryogenic X-ray data collection has been shown to elongate and disrupt disulfide bonds, reduce metallocenters, decarboxylate Asp and Glu sidechains, cleave Tyr hydroxyl groups, and interrupt metal coordination by protein side chains.[12] Even if data collection is free of radiation damage, the conditions of crystal preparation and collection can also affect the preferred side rotamer states. An X-ray crystallography study of cyclophilin A (CypA) at both cryo- and ambient temperatures reveals that side chain rearrangements can be critically affected by the X-ray collection

temperature, which can ultimately obscure identification of catalytically relevant orientations.[15]

In its current form, my work in Chapter 2 serves primarily as a proof of concept for a computational method using non-equilibrium candidate Monte Carlo (NCMC) mixed with MD to enhance side chain sampling. Further development and validation of this method is required to realize its application on more complex side chains and systems. In the culmination of my Chapter 2 study, I apply the method to a simple valine side chain in T4 lysozyme L99A and show that it significantly enhances sampling of Val111 compared to classical MD; however, it remains to be seen whether this method is effective on more flexible residues, such as Arg, Lys, Gln, Met, and Asn which are shown to rearrange most frequently upon ligand binding.[6] Furthermore, given that aromatic motifs can have high steric barriers to rotation, it would be interesting to validate this approach in the binding pocket of protein with a Tyr or Phe. Finally, in an ideal setting, this approach could be used to enhance sampling of side chains that rotate in concert, as is observed in the previously noted example of CypA, where several side chains shift their rotameric states in room temperature versus cryogenic X-ray structures. This system would be the penultimate test case for the method – can our mixed NCMC/MD approach help resolve the catalytically relevant side chain orientations when initiated from the starting coordinates of the cryogenic X-ray structure?

In Chapter 3, I shift to a more applied project and demonstrate how in one of its most basic forms, classical MD methods can be used to effectively complement and validate structural observations. Elucidating the structure of Mtb heme-degrading protein, MhuD, in complex with product is critical to understanding how the enzyme rearranges to release

its tetrapyrrole product. In kinetic studies of MhuD with heme and an electron donor, enzyme catalysis is limited to single turnover events. My co-authors in Chapter 3 demonstrate that the binding affinity of MhuD for its product is in the nanomolar range and thus passive diffusion of the product out of the binding site is inefficient and unlikely. In the structure of the MhuD variant bound to its biliverdin (BV) product, we observed the formation of a novel α-helix not observed in the substrate bound structure. This conformational shift may facilitate recognition by an accessory protein to promote removal of product. However, in our MhuD structure, we also observed an unusual ratio of ligand to protein (5:2) which arises from the stacking of BV molecules. Additionally, the relative proximity of the α-helix to the BV and the protein subunit interface gave rise to uncertainty regarding the legitimacy of the new helical motif. Using classical MD simulations of MhuD with heme and MhuD with BV, initiated from the structural coordinates of both the substrate and product bound crystal structures, I demonstrated that the formation of the novel α-helix is mediated by MhuD turnover of heme to its tetrapyrrole product. Analysis of the MD reveals that: 1) the α-helix is stable; 2) there is de-novo formation of the α-helix when substituting BV in place of heme in the heme-bound MhuD structure; 3) formation of the α-helix is driven by the rotation of the otherwise iron-coordinated His75 residue out of the binding pocket. MD analysis also identified a previously unexamined residue, Arg79, which may be involved in coordinating heme in the active site through interactions with its propionate groups.

MhuD belongs to a family of IsdG-type heme-degrading proteins that have also been shown to have high affinity for their substrates. Future work will involve determining whether structurally homologous proteins in this family, like IsdI and IsdG from

Staphylococcus aureus also form a novel α-helix or undergo a related conformational change upon product turnover. Additionally, we would like to identify and characterize the accessory protein in Mtb that is responsible for removing MhuD's product. As noted in Chapter 3, we have identified a few candidates in Mtb based on their homology to biliverdin reductases, which excise biliverdin from human heme oxygenase (HO-1). Finally, the observation in MD simulations that Arg79 may play a role in stabilizing heme in the active site, warrants further investigation through site-directed mutagenesis studies. Identification of MhuD's accessory protein and further characterization of MhuD's binding site could ultimately guide the design of future TB therapeutics that target iron homeostasis in Mtb.

In Chapters 2 and 3, I have had the luxury of working with high quality structures with resolutions of 2.5Å or better. In Chapter 4, I describe the structure of MEZ, Mtb malic enzyme, that plays a role in maintaining the supply of NAD(P)H for use as a reductant in the biosynthesis of lipids that are integral components of the Mtb cell wall. Of the projects I have worked on in my graduate career and elsewhere, the crystallization of MEZ has proved to be the most challenging and requiring the greatest perseverance. While I have chosen not to pursue a career in structural biology, I have developed a profound respect and appreciation for those scientists that dedicate themselves to elucidating novel protein structures. Not only is this pursuit of upmost importance (and particularly integral to the work of computational chemists), it is incredibly difficult and unpredictable endeavor that requires immense dedication and unending optimism.

After harvesting and freezing over 2000 crystals in a multitude of conditions, my best data set for MEZ diffracted to 3.6Å, with no ligands bound. Although there are

ambiguities in the rotameric states for many side chains and poor electron density for the backbone of some residues, the MEZ structure nonetheless provides insight into its molecular organization and relative disorder. Using MD simulations and docking, I was able to identify residues in MEZ that form contacts with its cofactors and substrate, identify putative binding modes and conformations for $NAD(P)^+$ and better understand its relative indifference with regard to its preferences for NAD^+ versus $NADP^+$. MD simulations also highlighted a unique structural distinction between MEZ and other eukaryotic MEs, specifically the absence of a conserved Arg in the binding site which is substituted by Ala151 in MEZ.

In MD simulations of MEZ, I observe two stable conformations of the NAD^+ cofactor – an elongated form that is well positioned for catalysis and a collapsed conformation where the reactive nicotinamide group is oriented away from the substrate. Interestingly, the collapsed form makes contacts with the backbone of residues near Ala151 and traps MEZ in a semi-open state. I hypothesize that this collapsed, stable NAD^+ binding mode is unique to Mtb MEZ and unlikely to form stably in human MEs where the Arg residue stabilizes the closure of the binding pocket by interacting with the malate substrate. In designing an inhibitor specific for MEZ, one could imagine taking the approach of stabilizing this semi-open form of MEZ by emulating and building on the binding mode of the collapsed NAD^+ form with a ligand that also forms additional interactions in core of the active site typically occupied by the absent MEZ Arg residue.

The results and future directions of the three projects described in Chapters 2, 3, and 4 are distinct in their certitude and clarity. The path forward for my computational methods work (Chapter 2) is straightforward: improve the functionality of the method and

validate it on increasingly complex side chains and systems. For my research on MhuD (Chapter 3), the results are both conclusive and prospective. My MD results substantiate the structural observations and in association with the biochemical data, provide greater assurance that there is likely another protein responsible for MhuD product removal. The computational modeling also prompted new research questions about a residue that may be involved in the mechanism of action of heme degradation. Finally, in Chapter 4, my results and future directions, while interesting, are more speculative and difficult to validate. It is unlikely, given the challenges I confronted in crystallizing MEZ, that there will be a follow up crystal structure that directly corroborates the observations from the computational modeling. That said, there are a number of future biochemical studies that could be done to better characterize the influence of pH on the rate and directionality of MEZ. Finally, the semi-open conformation of MEZ in complex with collapsed NAD^+ could provide a starting point for virtual screening using either docking or a ligand-based approach based on the folded form of the NAD^+ molecule.

References

1. Wych, D., Wall, M. E., Mobley, D. & IUCr. Molecular dynamics simulations of protein X-ray crystallographic diffuse scattering. *Acta Crystallographica Section A: Foundations and Advances* http://scripts.iucr.org/cgi-bin/paper?S0108767319097575 (2019).

2. Wych, D. C., Fraser, J. S., Mobley, D. L. & Wall, M. E. Liquid-like and rigid-body motions in molecular-dynamics simulations of a crystalline protein. *Structural Dynamics* **6**, 064704 (2019).

3. Bonomi, M., Pellarin, R. & Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophysical Journal* **114**, 1604–1613 (2018).

4. Parey, K. *et al.* High-resolution cryo-EM structures of respiratory complex I: Mechanism, assembly, and disease. *Science Advances* **5**, eaax9484 (2019).

5. Fan, C. *et al.* Ball-and-chain inactivation in a calcium-gated potassium channel. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2116-0.

6. Gaudreault, F., Chartier, M. & Najmanovich, R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* **28**, i423–i430 (2012).

7. Mobley, D. L. *et al.* Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.* **371**, 1118–1134 (2007).

8. Deng, Y. & Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *Journal of Chemical Theory and Computation* **2**, 1255–1273 (2006).

9. Jiang, W. & Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *Journal of Chemical Theory and Computation* **6**, 2559–2565 (2010).

10. Mobley, D. L., Chodera, J. D. & Dill, K. A. Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of Chemical Theory and Computation* **3**, 1231–1235 (2007).

11. RCSB PDB - Histograms. http://www.rcsb.org/pdb/statistics/histogram.do?mdcat=refine&mditem=ls_d_res_high&minLabel=0&maxLabel=5&numOfbars=10&name=Resolution.

12. Gerstel, M., Deane, C. M. & Garman, E. F. Identifying and quantifying radiation damage at the atomic level. *J Synchrotron Rad* **22**, 201–212 (2015).

13. Matsui, Y. *et al.* Specific damage induced by X-ray radiation and structural changes in the primary photoreaction of bacteriorhodopsin. *J. Mol. Biol.* **324**, 469–481 (2002).

14. Kort, R., Hellingwerf, K. J. & Ravelli, R. B. G. Initial events in the photocycle of photoactive yellow protein. *J. Biol. Chem.* **279**, 26417–26424 (2004).

15. Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).

Appendix A: Supporting Information for Chapter 2

Input files, simulations scripts, and analysis scripts are available on the ACS Publications website at DOI: 10.1021/acs.jctc.8b01018. Descriptions and details for the analysis and run scripts are organized and described following the supplemental figures; file paths and file names are italicized.

Supplemental Figures

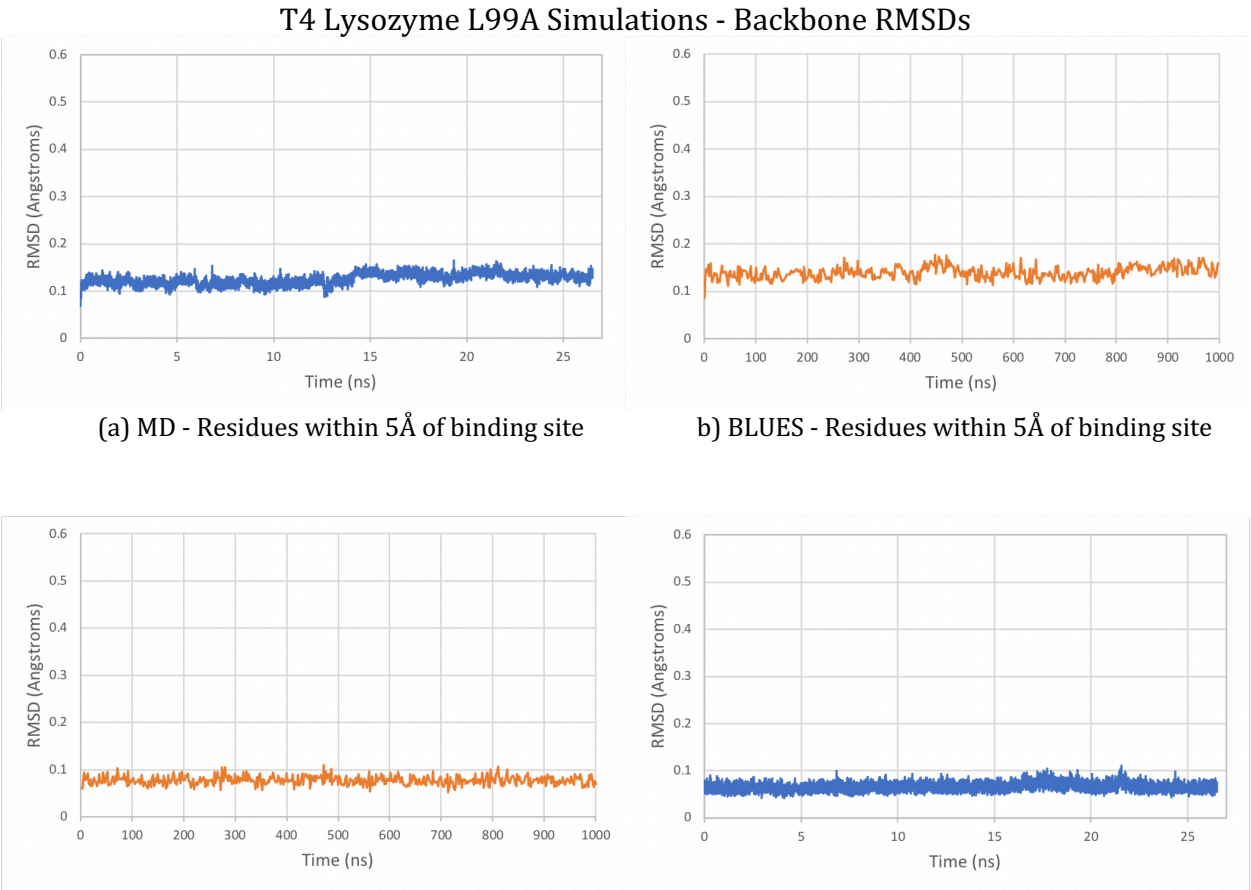


Figure A.1: **RMSDs of T4 lysozyme L99A backbone atoms in BLUES and MD simulations.** RMSDs of backbone atoms from the microsecond MD simulation of T4 lysozyme L99A with p-xylene bound are plotted in orange while those from the shorter BLUES simulations are plotted in blue. (a) and (b) RMSDs of all backbone atoms residues within 5Å of the p-xylene binding site. (c) and (d) RMSDs of backbone atoms from Val111.

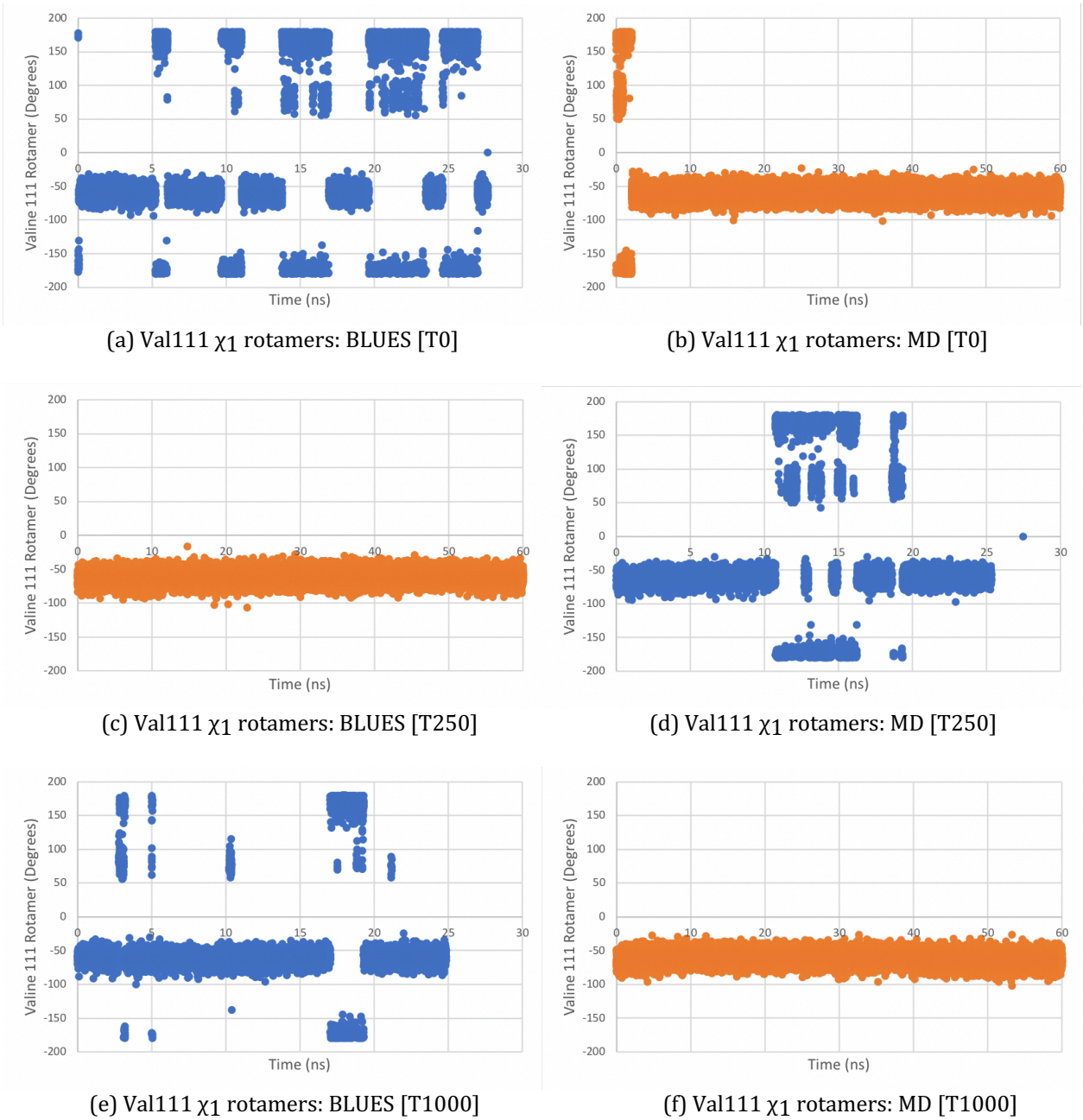


Figure A.2: **Val111 χ1 rotamer data for BLUES and MD simulations of p-xylene bound T4 lysozyme L99A in explicit solvent from three starting states.** The BLUES simulations are plotted in blue (a,c,e) and the MD simulations are plotted in orange (b,d,f). (a) and (b) These simulations started from the initial input files generated from the p-xylene bound T4 lysozyme L99A crystal structure. The BLUES simulation was 27.7ns, with 23.9 × 10^6 FEs, and a total of 91 rotamer transitions. The MD simulation was 60ns, with 30 × 10^6 FEs, and a total of 15 rotamer transitions. (c) and (d) These simulations started from a snapshot of p-xylene bound to T4 lysozyme L99A after

250ns of MD simulation. The BLUES simulation was 25.3ns, with 22.7 × 10^6 FEs, and a total of 47 rotamer transitions. The MD simulation was 60ns, with 30 × 10^6 FEs, and 0 observed rotamer transitions. (e) and (f) These simulations started from a snapshot of p-xylene bound to T4 lysozyme L99A after 1μs of MD simulation. The BLUES simulation was 24.8ns, with 22.5 × 10^6 FEs, and a total of 30 rotamer transitions. The MD simulation was 60ns, with 30 × 10^6 FEs, and 0 observed rotamer transitions.

Input Files

inputfiles/

The input files folder contains the Amber coordinate and parameter files for the T4 lysozyme L99A with p-xylene system as well as those for the valine-alanine dipeptide. For lysozyme, there are three coordinate files - *lastpxyMD.inpcrd* was used for the T1000 simulations, *pxyLys_MD_250ns,inpcrd* was used for the T250 simulations, and *lysozyme_pxy.inpcrd* was used for all other simulations, including T0.

Run Scripts

run_scripts/

The *run_scripts/* folder contains the run scripts for MD simulations using Amber and openMM as well as those for running BLUES.

run_scripts/openMM_MD/openmm_md.py

Backbone restraints can be set by uncommenting lines 30-38.

The periodic torsion force constant can be adjusted by uncommenting lines 44-57; the factor by which the force constant is adjusted is set in line 55.

run_scripts/BLUES/

moves.py and *simulation.py* are modified BLUES package scripts used to specifically implement biased sidechain rotations for the chi1 rotamer of valine 111 in T4 lysozyme.

Biasing is manually implemented in *moves.py* and *simulation.py* as follows:

In *moves.py*-

Line 513-516 the dihedral atoms of the chi1 torsion are defined and the current angle is computed.

Line 522-555, the bins are defined, and a theta is randomly generated until one that results in a move to an alternate bin state is identified.

Lines 625-650, following execution of NCMC, the resultant dihedral is checked to make sure it still falls into one of the biasing bins.

In *simulation.py*-

Lines 564 - 585, a boolean function (evalDihedral) that defines the acceptable rotamer bins for valine and determines if the current angle falls within one of those bins

Lines 598-608, evalDihedral is executed in an if statement that determines whether or not an NCMC move proposal will be executed. If not, another round of MD is performed.

sc_test_pxy.py was used in BLUES simulations for T4 lysozyme with p-xylene and *sc_test_watVA.py* was used in BLUES simulations for the solvated valine-alanine peptide. For each, the input structure files are loaded on lines 23-35 and the residue to be rotated is specified on line 41.

All other parameters are set in lines 29-38.

Lines 67-75 are used to compute the dihedral angle of the target bond with the atom indices specified on line 71.

run_scripts/amber_MD

md.in contains the input parameters while *equil_din.sh* is the shell script used for execution.

umbrella_sampling/

Contains the run and analysis scripts for umbrella samping the valine rotamer in valine-alanine.

For *umbrella_sampling/run_umbrella_sampling.py*

This script uses openMM to perform umbrella sampling. The inputs are specified on lines 25-26. The torsion is harmonically restrained according to lines 66-71. Backbone restraints are added on lines 73-82. The atoms that make up sampled torsion are specified in line 132.A separate file (*centers.dat*) specifies the force constant at each umbrella sampling window. This value can be adjusted as needed to improve overlap of sampling windows.

For *umbrella_sampling/umbrella_sampling_analysis.py*

This script processes the output trajectory files generated from *run_umbrella_sampling.py* and uses MBAR to generate a 1D PMF. The parameter file is specified on line 27 and the sampled atoms are specified on 28. This script also pulls the force constants used for each window of the sampling from the *centers.dat* file.

Analysis Scripts

dihedral_analysis/

dihedraldata.py is a simple script for computing and writing out the dihedral angle of atoms specified on line 8 for a system trajectory described on lines 4 and 5.

trans_big_moves.py is a script that, given a list of dihedral angle data on line 95, computes the number of transitions amongst the 3 valine rotamer states (gauche(-) or m60, gauche(+) or p60, and trans or mp180). It also computes the relative probability of each

state for the input data and the average counts of each state. Additionally, the average amount of time spent in each rotamer state without transitioning is accounted in lines 104-111. Lines 136-152 were used with BLUES log files to affirm that only "large" moves to alternate rotamer states were being proposed and accepted.

chunk_analysis.py was adapted from *trans_big_moves.py* and was used to split a list of dihedral angles (specified on line 61) into x number of chunks (specified on line 25) such that the frequency of each rotamer state as well as the total number of rotamer state transitions for each chunk would be computed separately and could be used to generate error bars.

Appendix B: Supporting Information for Chapter 3

Accession IDs
hHO-1 P09601

HmuO Q54AI1

MhuD P9WKH3

IsdG Q7A649

IsdI Q7A827

Coordinates and structure factors have been deposited in the Protein Data Bank with a PDB ID: 6PLE.

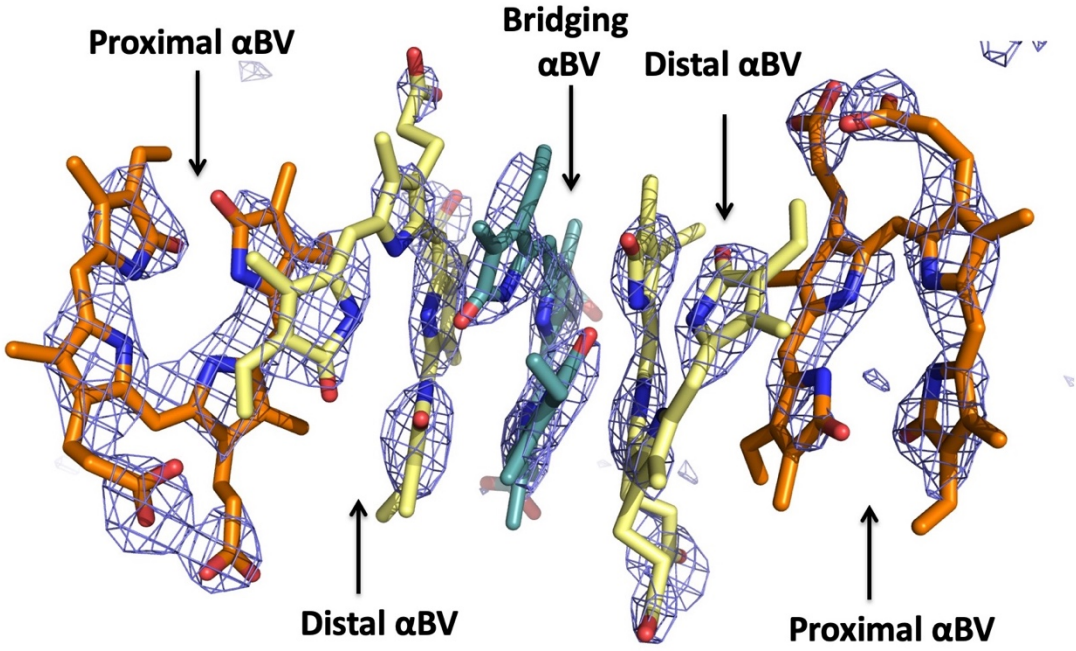


Figure B.1 A simulated annealing m*F*o-Δ*F*c omit map (generated by phenix.refine) of the αBV molecules, contoured at 3.0 σ, with the mesh colored in blue. The five αBV molecules bridging the two MhuD-R26S-αBV monomers in the asymmetric unit are colored as in Figure 3A.

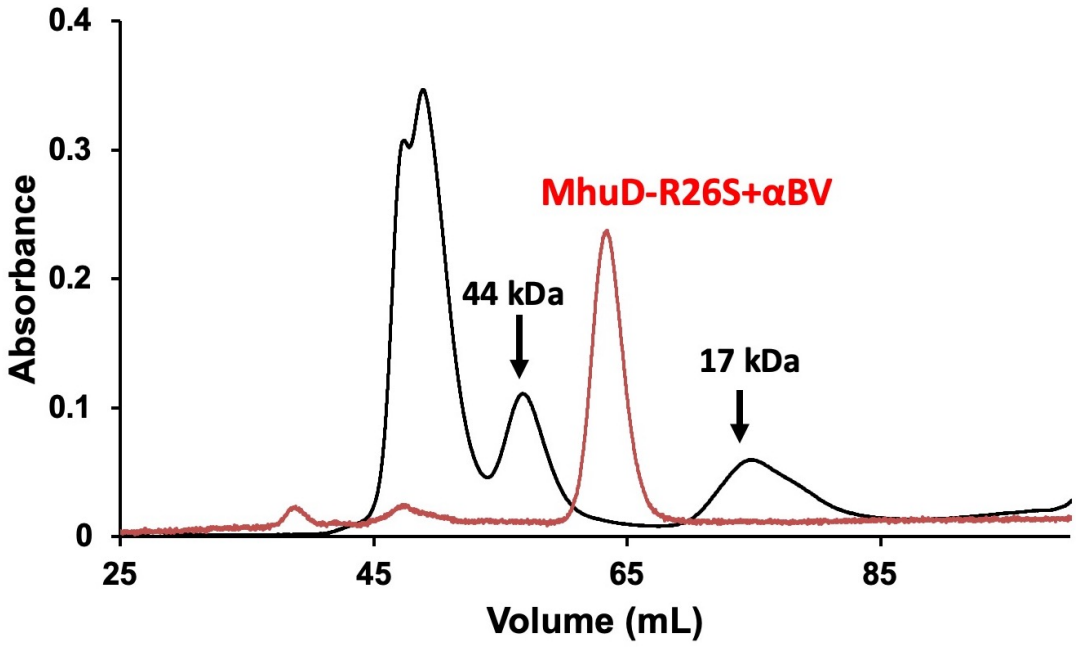


Figure B.2 Size exclusion chromatography (S75, HiLoad 16/600) of the MhuD+αBV complex (red trace) and protein standards (black trace, Bio-Rad), which clearly demonstrates that the MhuD-R26S+αBV complex elutes as dimer.

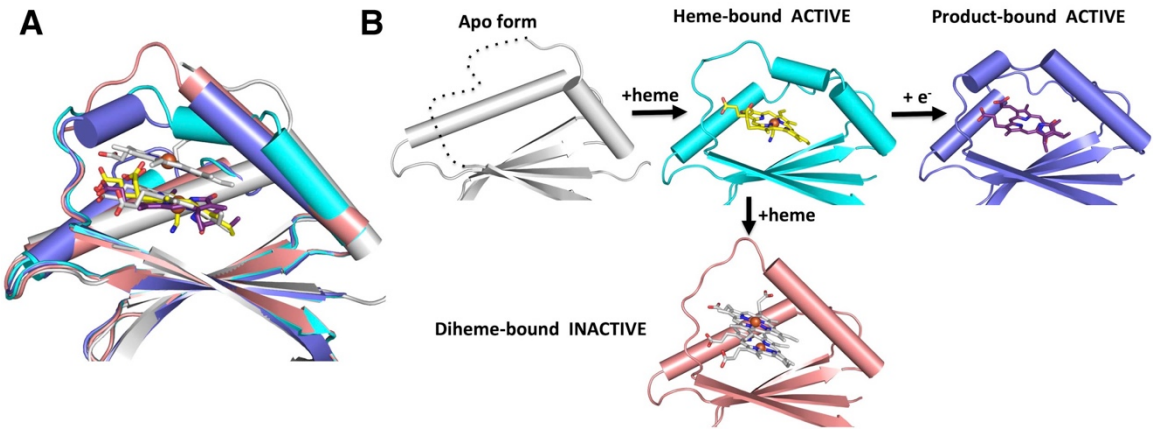


Figure B.3 Depiction of MhuD's conformational flexibility along its reaction pathway. A. Superposition of apo-MhuD (white, PDB ID: 5UQ4), MhuD-monoheme (cyan, PDB ID: 4NL5), MhuD-diheme (pink, PDB ID: 3HX9) and MhuD-R26S-αBV (blue). **B.** Apo-MhuD binds heme to form the closed MhuD-monoheme active form and in the presence of an electron donor it can degrade heme to the product bound form or alternatively, if there are high concentrations of heme, then it forms an open diheme inactive form.

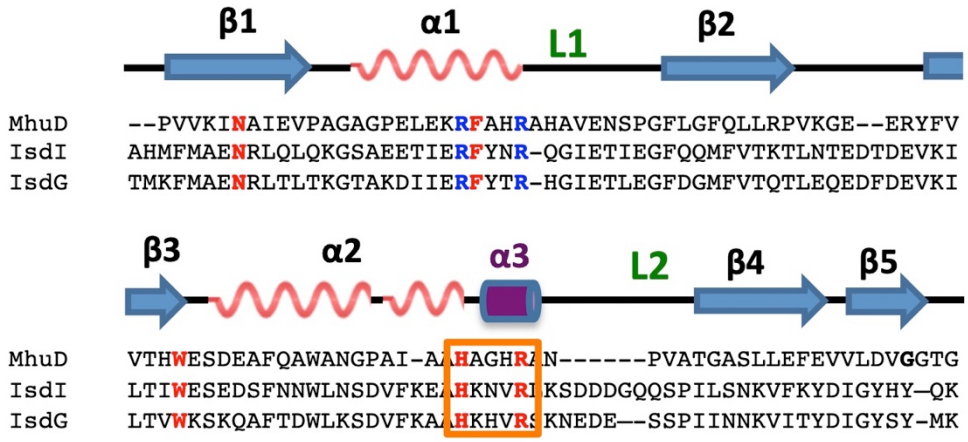


Figure B.4 Structure- and sequence-based alignment of MhuD and *S. aureus* IsdG and IsdI. Structure-based alignment of MhuD, IsdG and IsdI. The secondary structural elements (arrows are β-strands and waves are α-helices) are based on the MhuD-monoheme structure (PDB ID: 4NL5), apart from the purple cylinder α-helix-3, that is the new helix observed in the MhuD-R26S-αBV structure. Important conserved residues are in red and blue. The MhuD ^{75}HXXXR^{79} motif is highlighted by an orange box.

Appendix C: Supporting Information for Chapter 4

Coordinates and structure factors have been deposited in the Protein Data Bank with a PDB ID: 6URF.

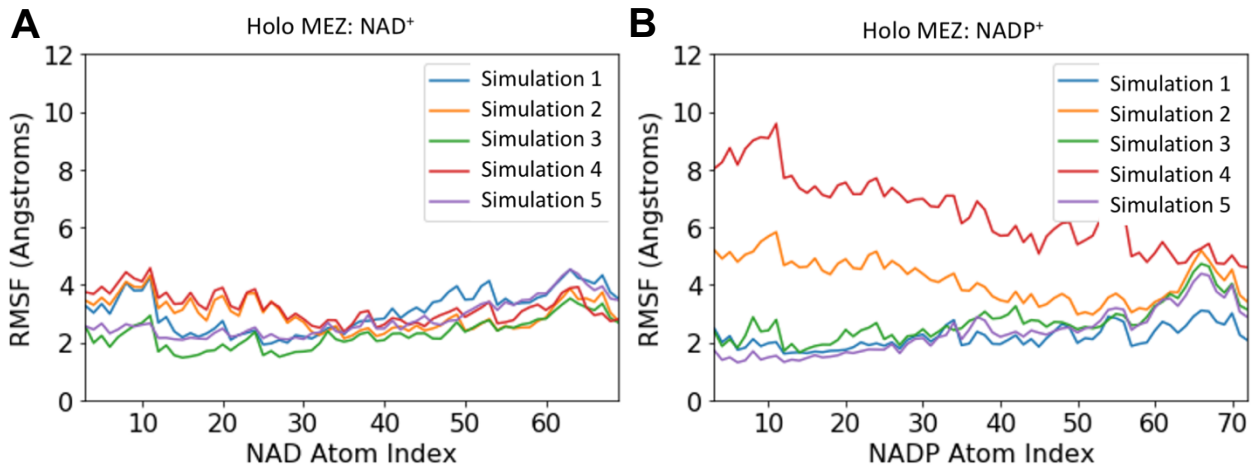


Figure C.1 RMSF of NAD(P)+ Cofactors in Molecular Dynamics Simulations. To evaluate the overall stability of the cofactors for each simulation and identify stable conformations for analysis, the RMSF of each non-hydrogen cofactor atom was computed and averaged for the two chains for each of the five 100 ns simulations. In panel a) Holo MEZ NAD+, each simulation includes MEZ with Mn2+/malate/NAD+ while in panel b) Holo MEZ: NADP+, each simulation includes MEZ with Mn2+/malate/NADP. These plots were used to help identify simulations where the binding modes for the cofactors were more stable: Job ID: 6024 for NAD + and Job ID: 2283 for NADP +. Conformations from the final frames of these two simulations were used for binding mode analysis.

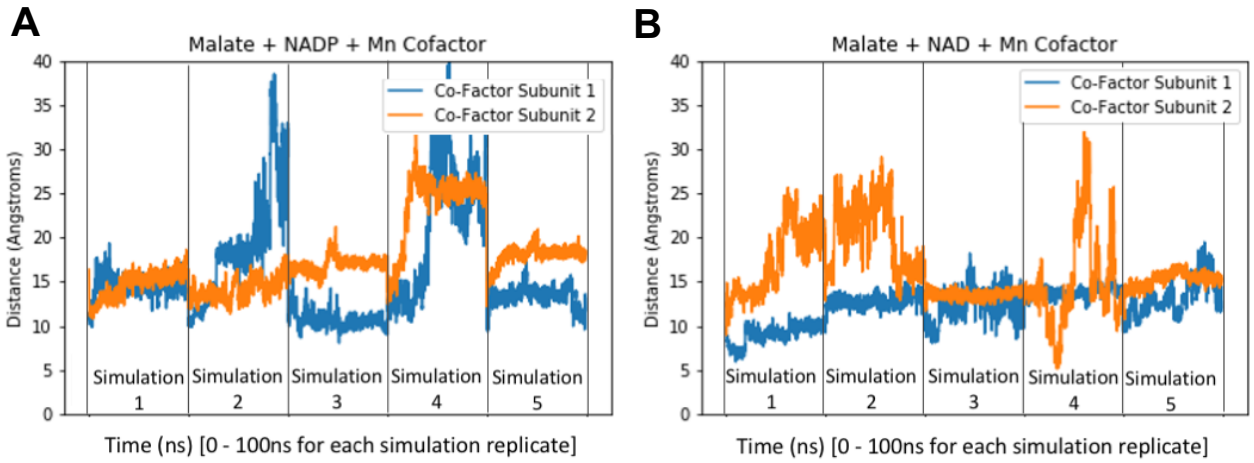


Figure C.2 NAD(P)+ Cofactor Stability in Binding Site in Molecular Dynamics Simulations. To identify simulations where the cofactor persistently occupied the binding site, the distance between the NAD(P)$^+$ carbon atom at position 4 of the nicotinamide group and the Cα atom of a stable residue in the binding pocket, Asn455. A) In simulations of MEZ with NADP$^+$, malate and Mn^{2+}, the

cofactor remained stably bound for 3 out of 5 of the 100 ns replicates (Simulations 1, 3, 5). B) In simulations of MEZ with NAD^+, malate and Mn^2, the cofactor consistently occupied the binding site for 2 of the 5 replicates (Simulations 3, 5).

Table C.1 Malic Enzyme Structures in Protein Data Bank. Representative structures of malic enzymes (MEs) in the PDB, including oligomeric state indicated as D = dimer, T = tetramer with cofactor, substrate/inhibitor, divalent metal and activators.

Organism - Oligomeric state	Cofactor	Substrate /inhibitor	Cation	Acti-vator	Res. (Å)	PDB code	RMSD(Å) /over residue #s	Ref
ME Large Subunit								
Human - D	N/A	N/A	N/A	N/A	2.5	2AW5	2.0 (512/531)	-
Ascaris suum - T	NAD				2.3	1LLQ	1.6 (515/592)	[1]
Zea mayes (Maize) - T					2.2	5OU5	1.8 (514/557)	[2]
Human mitochondrial - T	NAD				2.1	1QR6	1.8 (515/551)	[3]
Sorghum bicolor - T	NADP	pyruvate			2.0	6C7N	1.9 (514/557)	[2]
Human cytosolic - T					2.5	3WJA	1.8 (515/563)	[4]
E. coli - D					2.3	6AGS	2.0 (514/557)	-
Human mitochondrial - T	NAD		Lu^{3+}		2.9	1PJL	1.8 (515/551)	[5]
Columba livia (Pigeon) - T	NADP	oxalate	Mn^{2+}		2.5	1GQ2	2.5 (515/555)	[6]
Ascaris suum -T	NAD	tartronate			2.0	1O0S	1.6 (515/592)	[7]
Human mitochondrial - T	tartronate	fumarate	Mn^{2+}	ATP	2.2	1GZ4	2.9 (514/551)	[8]
Human mitochondrial - T	ATP	D-malate	Mn^{2+}	FUM	2.3	1PJ4	2.6 (514/552)	[9]
Human mitochondrial - T	NAD	pyruvate	Mn^{2+}	FUM	2.1	1PJ3	2.6 (515/553)	[9]
Human mitochondrial - T	NADH	L-malate	Mn^{2+}	FUM	2.3	1PJ2	2.5 (514/553)	[9]
Human mitochondrial - T	NAD	oxalate	Mn^{2+}		2.2	1DO8	2.6 (514/553)	[10]
Human mitochondrial - T	NAD	tartronate	Mg^{2+}		2.6	1EFL	2.6 (514/553)	[10]
Human mitochondrial - T	NAD	ketomalate	Mg^{2+}		2.6	1EFK	2.6 (514/553)	[10]
Human mitochondrial - T	ATP	oxalate	Mn^{2+}	FUM	2.3	1GZ3	2.6 (514/554)	[8]
Human mitochondrial - D								
ME Small Subunit	NAD	MES			1.6			
Pyrococcus horikoshii - D					2.5	2DVM	2.7 (353/434)	-
Pyrococcus horikoshii - D	NAD		Zn^{2+}		3.1	1WW8	2.8 (354/433)	-
Thermotoga maritima - D			Mg^{2+}		2.5	2HAE	2.5 (320/373)	-
Streptococcus pyogenes - D					2.6	2A9F	2.6 (323/380)	-
Thermotoga maritima - D		acetate			2.3	1VL6	2.7 (321/377)	-
Entamoeba histolytica - D	NAD		Mg^{2+}		2.5	3NV9	3.3 (358/486)	-
Candidatus phytoplasma - D						5CEE	2.6 (325/387)	[11]

References

1. Coleman, D.E., et al., *Crystal structure of the malic enzyme from Ascaris suum complexed with nicotinamide adenine dinucleotide at 2.3 A resolution.* Biochemistry, 2002. **41**(22): p. 6928-38.
2. Alvarez, C.E., et al., *Molecular adaptations of NADP-malic enzyme for its function in C4 photosynthesis in grasses.* Nat Plants, 2019. **5**(7): p. 755-765.
3. Xu, Y., et al., *Crystal structure of human mitochondrial NAD(P)+-dependent malic enzyme: a new class of oxidative decarboxylases.* Structure, 1999. **7**(8): p. R877-89.
4. Hsieh, J.Y., et al., *Structural characteristics of the nonallosteric human cytosolic malic enzyme.* Biochim Biophys Acta, 2014. **1844**(10): p. 1773-83.
5. Yang, Z., et al., *Potent and competitive inhibition of malic enzymes by lanthanide ions.* Biochem Biophys Res Commun, 2000. **274**(2): p. 440-4.

6. Yang, Z., et al., *Structural studies of the pigeon cytosolic NADP(+)-dependent malic enzyme.* Protein Sci, 2002. **11**(2): p. 332-41.
7. Rao, G.S., et al., *Crystallographic studies on Ascaris suum NAD-malic enzyme bound to reduced cofactor and identification of an effector site.* J Biol Chem, 2003. **278**(39): p. 38051-8.
8. Yang, Z., C.W. Lanks, and L. Tong, *Molecular mechanism for the regulation of human mitochondrial NAD(P)+-dependent malic enzyme by ATP and fumarate.* Structure, 2002. **10**(7): p. 951-60.
9. Tao, X., Z. Yang, and L. Tong, *Crystal structures of substrate complexes of malic enzyme and insights into the catalytic mechanism.* Structure, 2003. **11**(9): p. 1141-50.
10. Yang, Z., et al., *Structure of a closed form of human malic enzyme and implications for catalytic mechanism.* Nat Struct Biol, 2000. **7**(3): p. 251-7.
11. Alvarez, C.E., et al., *The crystal structure of the malic enzyme from Candidatus Phytoplasma reveals the minimal structural determinants for a malic enzyme.* Acta Crystallogr D Struct Biol, 2018. **74**(Pt 4): p. 332-340.