UNIVERSITY OF CALIFORNIA,
IRVINE

Development and Application of a Computational Force Field for the Study of Structure, Function and Motion of Enzymes in the Acetate and Non-ribosomal Peptide Pathways

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Andrew Joseph Schaub

Dissertation Committee:
Professor Shiou-Chuan Tsai, Chair
Professor Ray Luo
Professor Rachel Martin

2018

DEDICATION

This dissertation is dedicated to the memory of my father and father-in-law, Jeff Schaub and Hideshi Koyama. I am also grateful to my wife and son, Nanayo and Sora Schaub, for their unconditional love and being the most important individuals in my life.

この論文は、父のシャーブ・ジェフと義父の小山英史に捧げます。

そして、私の人生の中で最も大切であり、無償の愛を与え続けてくれる
妻の七奈代と息子の空良に感謝しています。

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my Ph.D. advisor and mentor, Professor Sheryl Tsai, for being a very supportive and caring mentor. She provided optimism and enthusiasm when I needed it the most to carry this research forward. Her passion for mentoring graduate students, educating undergraduate students and advancing science is truly inspirational. Without her guidance and persistent help this dissertation would not have been possible.

I am also extremely grateful for the mentorship and guidance provided by Professor Ray Luo, in the development of my computational skillset. He was like a father figure, and it was humbling to work for someone who lived and breathed science. Over five years I never saw his love for biophysics and computation fade even the slightest.

I thank my other dissertation committee member, Professor Rachel Martin, as well as my advancement committee members, Professors David Mobley, Ioan Andricioaei, Doug Tobias and Ceila Goulding for providing constructive advice in the development of my projects. I must also thank the Luo, Tsai and Goulding labs for the sharing of ideas, materials and equipment. They have made great colleagues and neighbors, and I am truly indebted to their help and patience. Dr. Felix Grün and Sir Benjamin Katz over at the Mass Spectrometry facility taught me everything I know about liquid chromatography and mass spectrometry.

Current and previous graduate students and postdocs from the Tsai lab have been valuable in their mentorship and encouragement. My deepest gratitude goes to my original graduate student rotation mentor, Dr. Jesus Barajas. I am also grateful to Dr. Jake Milligan, Dr. Dave Jackson, Dr. Tim Valentic, Dr. Chi Nguyen, Dr. Gaurav Shakya, Dr. Joel Bruegger, Dr. Alex White, Gabe Moreno, Bryan Ellis, Shiji Zhao, and Jacob Wolff. I am fortunate for the continued mentorship and advice these people have provided me over the past few years, even after they had departed UCI.

Over the past five years I have had several talented undergraduate students who have helped push this research forward. I am grateful to Andy Tsai, Andrew Ma, Tho (Ethan) Thai, Tho Ho (Tiffany) Nguyen, and Rheem Fawaz. I am especially grateful to Hau Truong, a former undergraduate student who stayed in the lab as a research associate prior to starting graduate school at UCSC.

I thank the Elsevier for permission to include parts of Chapter three of my dissertation, which was originally published in Cell Chemistry & Biology.

CURRICULUM VITAE

Andrew Joseph Schaub

Email: aschaub930@gmail.com

EDUCATION

2013 – 2018	University of California, Irvine Ph.D. in Chemistry Mathematical, Computational and Systems Biology Gateway Program
2009 – 2013	University of Maryland, University College B.S. in Computer Science, minor in Biology

AWARDS AND HONORS

2018 – 2019	NIH Postdoctoral Intramural Research Training Award Fellowship (NIH)
2016 – 2017	JSPS Strategic / NSF GROW Research Fellowship (University of Tokyo)
2015 – 2018	National Science Foundation Graduate Fellowship (UC Irvine)
2013	Upsilon Pi Epsilon National Honor Society for Computing Disciplines
2011	Humanitarian Medal for Operation Tomodachi, US Air Force
2009 – 2013	Dean's List, University of Maryland, University College

PUBLICATIONS

Shakya, G.; Rivera, H., Jr.; Lee, D.J.; Jaremko, M.J.; La Clair, J.J..; Fox, D.T.; Haushalter, R.W.; Schaub, A.J.; Bruegger, J.; Barajas, J.F.; White, A.R.; Kaur, P.; Gwozdziowski, E.R.; Wong, F.; Tsai, S.-C.; Burkart, M.D. Modeling Linear and Cyclic PKS Intermediates Through Atom Replacement. *J. Am. Chem. Soc.* **2014**, *136* (48), 16792-16799.

Barajas, J.F.; Phelan, R.M.; Schaub, A.J.; Kliewer, J.T.; Kelly, P.J.; Jackson, D.R.; Luo, R.; Keasling, J.D.; Tsai, S.-C. Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem. Biol.* **2015**, *22* (8), 1018-1029.

Jackson, D.R.; Tu, S.S.; Nguyen, M.; Barajas, J.F.; Schaub, A.J.; Krug, D.; Pistorius, D.; Luo, R.; Müller, R.; Tsai, S.-C. Structural Insights into Anthranilate Priming During Type II Polyketide Biosynthesis. *ACS Chem. Biol.* **2015**, *11* (1), 95-103.

INVITED TALKS

Schaub, A.J. (January 2018). Development of a Pantetheinyl Force Field for "Unnatural" Natural Product Engineering. Presented as part of the Biophysics and Systems Biology Seminar Series, University of California, Irvine, CA, USA.

Schaub, A.J. (November 2015). Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. Presented as part of the Molecular Dynamics Seminar Series, University of California, Irvine, CA, USA.

RESEARCH EXPERIENCE

January 2014 – July 2018
Graduate Student Researcher: University of California, Irvine, California
Research Advisors: Dr. Shiou-Chuan (Sheryl) Tsai, Dr. Ray Luo
Investigations in structure, function, motion and diversity in the acetate and non-ribosomal peptide pathways through molecular dynamics and crystallographic studies

September 2016 – March 2017
JSPS Strategic / NSF GROW Research Fellow: The University of Tokyo, Japan
Research Advisor: Dr. Ikuro Abe
Probed the effects of pH on product selection in type III polyketide synthases through application of molecular dynamics simulations, biochemical analysis and synthesis of substrate mimics

September 2013 – December 2013
Graduate Student Researcher: University of California, Irvine, California
Research Advisors: Dr. Gregory A. Weiss
Encoded a phagemid with a protein kinase A (PKA) gene through subcloning techniques for use in phage display

September 2006 – March 2007
Undergraduate Student Researcher: The Ohio State University, Columbus, Ohio
Research Advisors: Dr. Charles L. Brooks
Performed structure and function studies of human prolactin and bovine growth hormone mutants

January 2006 – August 2006
Undergraduate Student Researcher: The Ohio State University, Columbus, Ohio
Research Advisors: Dr. Werner Tjarks
Synthesized boronated nucleoside analogs for boron neutron capture therapy

POSTER PRESENTATIONS

Schaub, A.J. (July 2016). Carrier Protein Force Field: An AMBER Force Field for Phosphopantetheine Mediated Biosynthesis. Poster presented at the 2016 Gordon Research Conference: Enzymes, Coenzymes & Metabolic Pathways in Waterville Valley, NH, USA.

Schaub, A.J. (July 2014). Molecular Dynamics Studies of the Myxalamid Biosynthesis PKS-NRPS Reductase Domain. Poster presented at the 28th Annual Symposium of the Protein Society in San Diego, CA, USA.

AUTHORED GRANTS

"Investigation of pH effects on product generation in a type III polyketide synthase Benzalacetone Synthase (BAS)." PI: Abe, I. Agency: XSEDE. Type: Stampede Computing Resources. MCB160168. Effective Dates: January 2017 – December 2017

TEACHING EXPERIENCE

Lecturer – Computational Structural Biology Seminar, University of Tokyo	January 2016
Teaching Assistant – Experimental Microbiology Lab (Bio 118L)	January – March 2015
Teaching Assistant – Molecular Biology (Bio 99)	April – June 2015
Teaching Assistant – Biochemistry (Bio 98)	January – March 2015
Teaching Assistant – Mathematical and Computational Biology Bootcamp	2014 – 2017
Teaching Assistant – National Short Course in Systems Biology	2014 – 2017
Teaching Assistant – Mathematical and Computational Biology Undergraduate Workshop	2016

PROFESSIONAL AFFILIATIONS

2011 – *Present*	AAAS, American Association for the Advancement of Science
2011 – *Present*	ACM, Association for Computing Machinery
2013 – *Present*	ACS, American Chemical Society
2015 – *Present*	BPS, Biophysical Society
2011 – *Present*	IEEE, The Institute of Electrical Engineers
2011 – *Present*	ISCB, International Society for Computational Biology

LANGUAGES

English (Native)
Japanese (Intermediate). Completion of General Intermediate Course, University of Tokyo.

REFERENCES

Dr. Shiou-Chuan (Sheryl) Tsai – Ph.D. Advisor
Professor. Departments of Molecular Biology and Biochemistry, Chemistry, and Pharmaceutical Sciences.
University of California, Irvine. (949) 824-4486. sctsai@uci.edu

Dr. Ray Luo – Ph.D. Advisor
Professor. Departments of Molecular Biology and Biochemistry, Biomedical Engineering, and Chemical Engineering and Materials Science.
University of California, Irvine. (949) 824-9528. ray.luo@uci.edu

Dr. Abe Ikuro – Visiting Research Advisor
Professor. Graduate School of Pharmaceutical Sciences, Laboratory of Natural Products Chemistry.
The University of Tokyo. +81-3-5841-4740. abei@mol.f.u-tokyo.ac.jp

Dr. Rachel Martin – Thesis Committee Member
Professor. Department of Chemistry.
University of California, Irvine. (949) 824-7959. rwmartin@uci.edu

ABSTRACT OF THE DISSERTATION

Development and Application of a Computational Force Field for the Study of Structure, Function and Motion of Enzymes in the Acetate and Non-ribosomal Peptide Pathways

By

Andrew Joseph Schaub

Doctor of Philosophy in Chemistry

University of California, Irvine, 2018

Professor Shiou-Chuan Tsai, Chair

Enzymes in the acetate and non-ribosomal peptide pathways generate chemically diverse and complex bioactive molecules, with the intermediates being chauffeured between catalytic partners via a carrier protein. Recent efforts have been made to engineer these systems to expand their product diversity. A major stumbling block is our poor understanding of the transient protein-protein and protein-substrate interactions between the carrier protein and its many catalytic partner domains. The innate reactivity of pathway intermediates has obfuscated our mechanistic understanding of these interactions during the biosynthesis of these natural products, ultimately impeding the engineering of these systems for the generation of "unnatural" natural products.

Molecular dynamics can be used to provide models of these key interactions that are difficult to capture experimentally, providing the potential to expand the diversity in these systems. Current force fields support basic biochemical building blocks, and specialized force fields support post-translational modifications and non-canonical amino acids, yet none currently exist that are capable of modeling bound intermediates from these systems.

The objective of this dissertation is to address this gap in knowledge and available technology and to present the description of the development of a force field that can be used with MD and other computational techniques to provide models of these experimentally intractable transient interactions. This objective will be divided into three aims.

Aim 1: The development and validation of a force field for use in investigations and engineering efforts involving these pathways. A fragmentation approach was used, providing a degree of modularity to the force field while at the same time reducing the size, complexity and degrees of freedom during the charge fitting and parameterization steps. Tutorials and a web interface were also developed to provide end users access to this tool.

Aim 2: The application of the force field towards understanding dynamics and interactions in these pathways. Regulation of the transient protein-protein interactions and discrete steps in fatty acid biosynthesis remain poorly understood. MD was able to show that specific interactions with its partner are either strengthened or weakened depending on the loaded state of the carrier protein. The force field was also used to model a plant-based polyketide synthase at different pHs, resulting in the identification of an allosterically modulated pH sensor on the surface of the polyketide synthase.

Aim 3: The application of the force field in engineering efforts in these systems to produce biofuels. MD simulations were employed to provide a deeper understanding of protein-substrate interactions in a reductase domain from a non-ribosomal peptide megasynthase. This led to a deeper understanding of this domain, and further identified residues critical for structure integrity and substrate binding, leading to a rationally altered variant with improved activity toward highly reduced substrates.

CHAPTER 1

Introduction

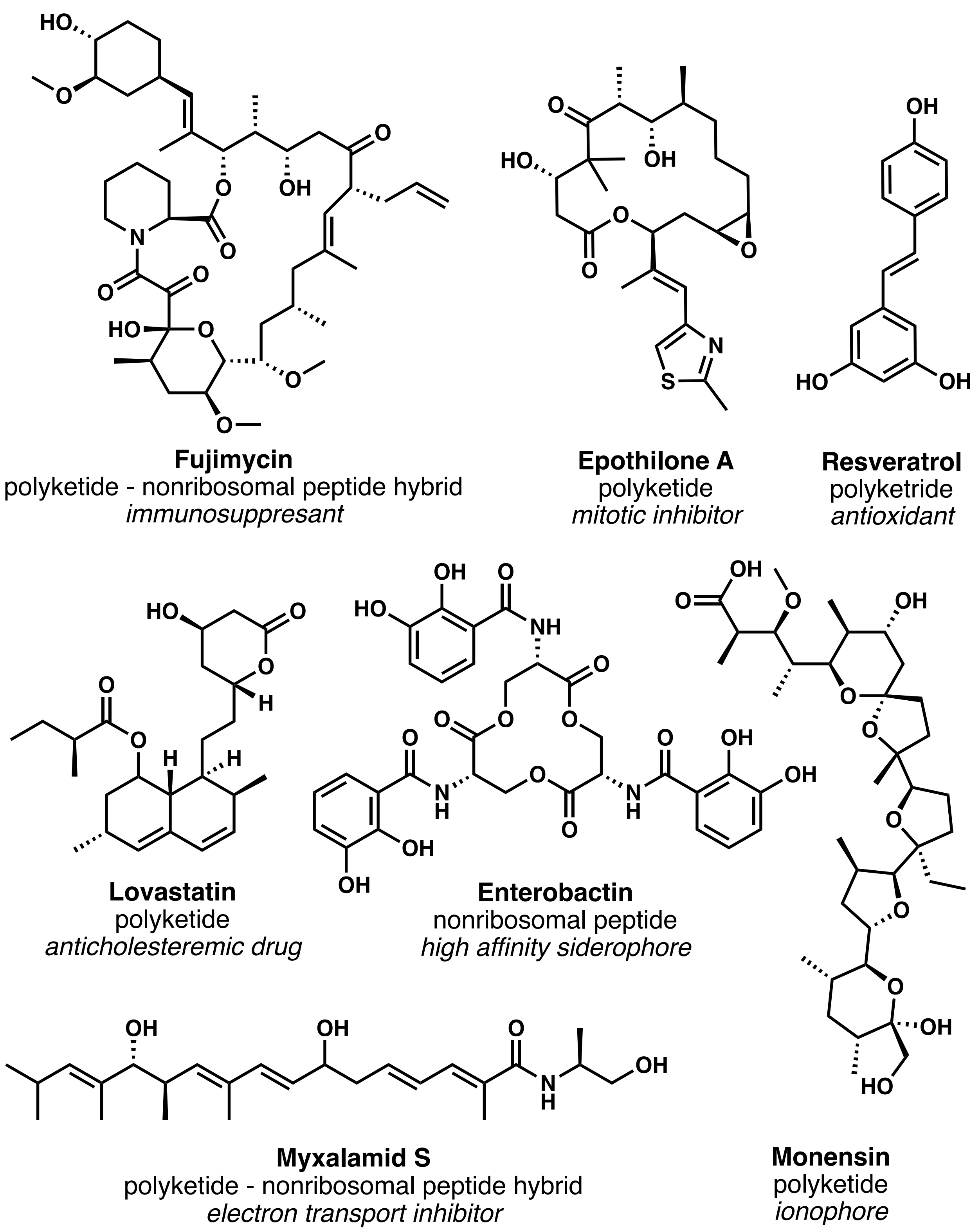
1.1 Natural products

For thousands of years, natural products have captivated the curiosity of humans, and have been intertwined into our culture, with early applications to religious ceremonies, mysticism, and witchcraft.[1] In addition, some have been revered for their healing properties and the source of oldest known medications, with substances found naturally occurring in plants, bacteria and fungi. An important distinction needs to be made between biochemistry and natural product chemistry. Traditionally, biochemistry encompasses primary metabolism, which would include the biosynthesis of nucleic acids, carbohydrates, proteins, fats, and sugars. The primary metabolism is ubiquitous across all species. In comparison, natural product chemistry encompasses secondary metabolism, in which the natural products, or secondary metabolites, may only be synthesized during certain cellular processes or by certain organisms. The field received international recognition in 2015, when Drs. Youyou Tu, William Campbell and Satoshi Omura were awarded the Nobel Prize in Physiology and Medicine for their discoveries of two natural products, artemisinin and avermectin.[2] On a broader scale, the field of pharmacognosy evaluates these naturally occurring medicinal drugs using evidence-based science.[3] This field contains many areas of research, including host cultivation, assay development, analytical chemistry, clinical studies, cell biology, genetics, marine chemistry, ethnobotany, heterologous gene expression, enzymatic structural studies, and organic synthesis, to name a few.[3]

The primary focus of this thesis is molecular dynamic (MD) simulations of natural product-biosynthesizing enzymes. In 2013, the Nobel Prize in Chemistry was awarded to Drs.

Martin Karplus, Michael Levitt and Arieh Warshel, for their contributions in theoretical chemistry that opens up a brand-new field for the simulations of macromolecules.[4] While there are other computational biology techniques such as bioinformatics and genomics, they are not the focus of this thesis, and these two topics are covered in depth in other outstanding reviews.[5-8] Computational structural biology in natural product research includes the development and application of computational techniques to provide much-needed insight to engineer these enzymes in order to generate new "unnatural" natural products, which would be invaluable for the identification of new pharmaceuticals and industrially-relevant compounds such as biofuels. Improved computational methods obtained from this thesis will make significant impacts by bridging a major knowledge gap in our understanding of the protein dynamics involved in the biosynthesis of these natural products.

While natural product compounds may initially look very complex, the majority of them are actually made up of basic building blocks, and a cursory logical inspection can often be used to determine the logical assembly of these compounds. The focus of this thesis will primarily be the natural products constructed from amino acids, or two-carbon units supplied by Coenzyme A derivatives. These classes of natural products include fatty acids, polyketides and non-ribosomal peptides, and have been utilized as antibiotics, immunosuppresants, cholesterol-lowering drugs, antitumoral agents, and biofuels.[9] Other classes of natural products include terpenoids, alkaloids, and phenylpropanoids, and are summarized in other excellent reviews.[10-14] Fatty acids, polyketides and non-ribosomal peptides are medically and industrially useful compounds that are assembled incrementally through the addition of extender units to an initial starter unit by fatty acid



Fujimycin
polyketide - nonribosomal peptide hybrid
immunosuppresant

Epothilone A
polyketide
mitotic inhibitor

Resveratrol
polyketride
antioxidant

Lovastatin
polyketide
anticholesteremic drug

Enterobactin
nonribosomal peptide
high affinity siderophore

Myxalamid S
polyketide - nonribosomal peptide hybrid
electron transport inhibitor

Monensin
polyketide
ionophore

Figure 1-1. Examples of natural products from the acetate and non-ribosomal peptide pathways. These compounds can be classified as fatty acids, polyketides, non-ribosomal peptides, or a hybrid of any of these categories.

synthases (FAS), polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS).[15-17] These natural products have great diversity in their structures and bioactivities (Fig. 1).[18-19] A previous review by Zhang and Rock on the application of computational methods towards the exploration of diversity and structure in FASs reviews this subfield up to 2003.[20] This chapter attempts to summarize the progress up to 2017, and also includes polyketides and non-ribosomal peptides. To the best of our knowledge, this chapter is the first attempt to comprehensively review the development and application of biophysical computational techniques towards understanding the enzymes in the acetate and non-ribosomal peptide pathways.

1.2 Enzymatic machinery

The biosynthesis of fatty acids, polyketides (PKs), and non-ribosomal peptides (NRPs) is typically accomplished by large, multi-domain enzyme complexes in eukaryotes, and by discrete mono-domain enzymes in prokaryotes. The enzymes are called fatty acid synthase (FAS), polyketide synthase (PKS), and non-ribosomal peptide synthetase (NRPS), respectively. Their intermediate products, often highly reactive, are shuttled between the catalytic domains via carrier proteins (CPs) in a well-choreographed order that results in the generation of final products with high fidelity. These "megasynthases" can generate chemically diverse and complex bioactive natural products. The biosynthesis is most often mediated by an acyl carrier protein (ACP) in fatty acid and polyketide biosynthesis, or peptidyl carrier protein (PCP) in non-ribosomal peptide biosynthesis. The growing chain is tethered to a post-translationally modified phosphopantetheine (PPant) prosthetic group attached to a serine residue on the carrier protein (CP), and ultimately released by the hydrolytic cleavage of a thioester bond to generate a macrocycle or linear chain. The valuable compounds generated by these megasynthases are difficult to access through organic synthesis because of their chemical complexity, innate reactivity, and multiple stereocenters.[21] Rather, biosynthesis of fatty acids, polyketides, and non-ribosomal peptides using FAS, PKS and NRPS offer a more economic and efficient way to obtain these natural products.

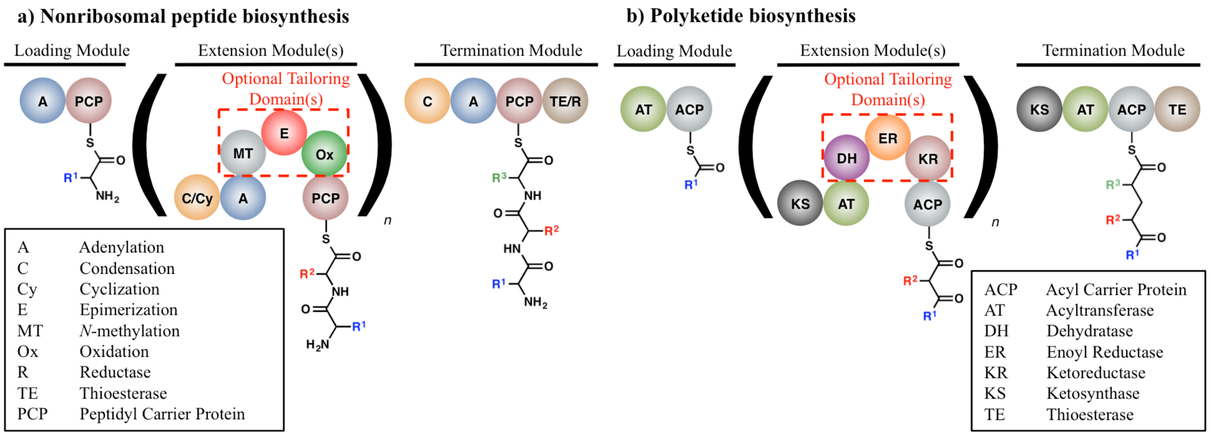


Figure 1-2. Assembly line biosynthesis of a) non-ribosomal peptides in Type A NRPS systems and b) polyketides in Type I modular PKS systems.

In the multi-modular FASs, PKSs, and NRPSs, one extender unit is added for every chain extension, and it is sometimes possible to predict the chemical structure of the final product based on the sequence of gene clusters. There exists an iterative type that is capable of using the same module more than once for multiple extensions, which adds an extra layer of difficulty in chemical structure prediction.[22]

The growing substrate is guided to each binding partner by a peptidyl carrier protein (PCP) in NRPSs and by an acyl carrier protein (ACP) in FASs and PKSs (Fig. 1-2). ACP and PCP are highly similar in structure. Both consist of a four-helix bundle, and a prosthetic phosphopantetheine group is post-translationally attached to a conserved serine residue of CP.

1.2.1 Fatty acid synthase (FAS)

The biosynthesis of fatty acids occurs through the condensation of acetyl units. This was shown experimentally by Rittenberg and Bloch, who in 1945 labeled acetate with ^{13}C at the carbonyl group and deuterium in the methyl group, followed by the detection of isotope locations in the products.[23] They also proved that it is acetyl-CoA that serves as the extender unit of FAS. Ten years later, the same isotope labeling technique was applied to the study of polyketides, with the generation of 6-methylsalicylic acid, which showed a head-to-tail addition of three acetate units. In 1958, Brady did a simple buffer exchange experiment, in which he replaced a bicarbonate buffer with a phosphate buffer, which led to failure in the original radiolabeling assay and revealed the necessity of bicarbonate in the biosynthesis of fatty acids.[23] A few years later, Wakil and Brady independently showed malonyl-CoA as a needed intermediate, which is formed from the ATP dependent carboxylation of acetyl-CoA (Fig. 1-3). The exergonic decarboxylation of malonyl-CoA generates a resonance-stabilized carbanion on acetyl-ACP, which is able to function as a nucleophile.[24] These experiments paved the foundation for current biosynthesis studies of megasynthases.

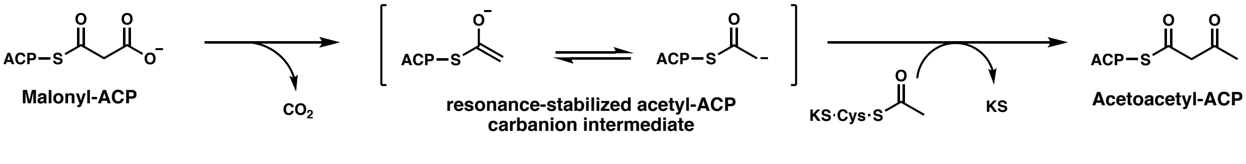


Figure 1-3. Carbon-carbon bond formation in fatty acid biosynthesis as elucidated by Wakil and Brady.

All organisms need to synthesize fatty acids for cellular membranes, and animals store fatty acids as triglycerides for future energy demands. Nature has adopted two solutions for this high demand of fatty acids. One solution employed by eukaryotes and some bacteria involves linking the necessary domains covalently, forming a "megasynthase"; these large polypeptides are known as type I FASs.[23, 25] Another solution found in plants and bacteria is to synthesize fatty acids by use of functional discrete mono-domain enzymes, with ACP shuttling the growing intermediates, and these standalone enzymes are known as type II FASs.

1.2.1.1 Type I FAS

Two comprehensive reviews on type I FAS systems have been written by Smith and Maier.[23, 26] Here, we briefly described the structural biology relevant to this thesis.

1.2.1.1.1 Mammalian Type I FAS

In 2006, Maier *et al.* obtained an initial X-ray structure of mammalian fatty acid synthase from *Sus scrofa* at a resolution of 4.5 Å (PDB ID: 2CF2), and the homodimeric structure revealed that mammalian FAS adopts an overall X-shape.[27] The porcine FAS (UniProtKB: I3LC73) shares a 79% sequence identity with human FAS (UniProtKB: P49327), Furthermore, the structure revealed the relatively long distances between catalytic domains, highlighting the flexibility required by the ACP to interact with its many partners. In 2008, Maier *et al.* again obtained another structure of the porcine mammalian FAS (Fig. 1-4), at an improved resolution of 3.2 Å (PDB IDs: 2VZ8, 2VZ9)[25] These higher quality structures of the wild boar FAS revealed a central axle, by which the ACP is able to rotate around as it swivels to its many partners to incrementally increase the length of the growing fatty acyl chains.

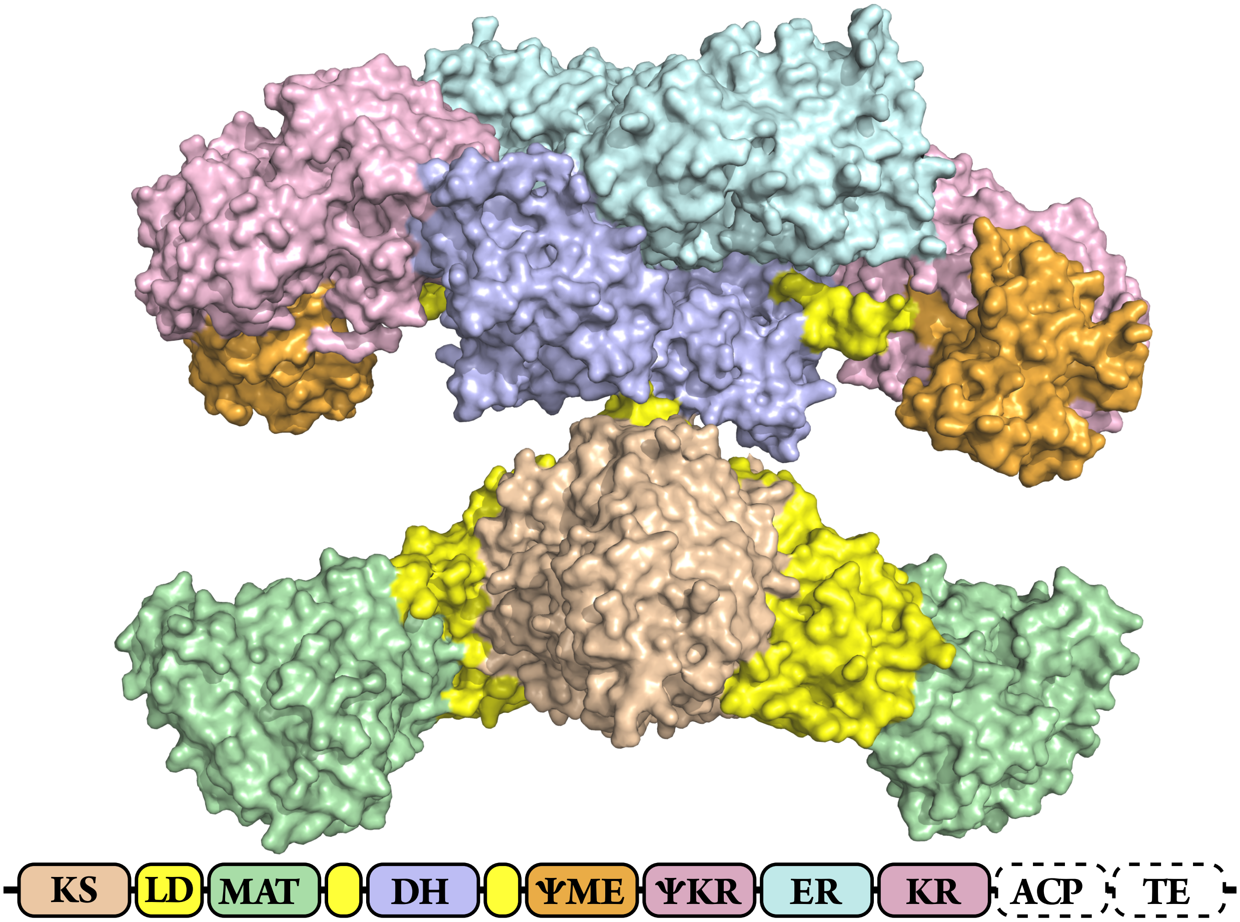


Figure 1-4. The homodimer porcine mammalian FAS (mFAS) structure determined by Maier *et al.* is shown above in "Connolly" surface representation.[25, 28] This figure is an adaptation of an original present in the original work by Maier. Domains are colored for clarity and include the ketosynthase (KS), linker (LD), malonyl-acyl transferase (MAT), dehydratase (DH), inactive methyltransferase (ΨME), inactive ketoreductase (ΨKR), enoylreductase (ER), ketoreductase (KR) domains. The acyl carrier protein (ACP) and thioesterase (TE) domains are missing from the crystal structure. Shown below the crystal structure is the linear organization of the domains on each of the monomer polypeptide chains, with colors matching the cartoon representation above, and the ACP and TE domains are shown in white with dashed lines to represent that they are missing.

Using the above structural information, a murine type I FAS was shown to accept non-native substrates by Rittner *et al.*, who also assigned kinetic constants for native substrates acetyl-CoA and malonyl-CoA, and used these for comparison purposes.[29] The authors noted the malonyl/acyltransferase (MAT) of murine FAS might serve a useful tool for the loading of un-natural substrates due to the domain separation present in its structure. Furthermore, the same MAT, or a similar MAT from other FAS/PKS systems might provide access to incorporation of different starter units.

The full-length structures of human fatty acid synthase (hFAS) have so far eluded structural biologists. It is known that the thioesterase (TE) domain in hFAS is responsible for chain length selection, and two crystal structures of the hFAS thioesterase (hFASTE) domain were reported at 2.6 Å and 1.48 Å, respectively.[30-31] More recently, in 2015, scientists at GlaxoSmithKline (GSK) obtained a structure of a hFAS tri-domain consisting of the ΨME, ΨKR and KR domains with a length of 660 residues and mass of 71.8 kDa.[32] The tri-domain showed high conservation between hFAS and porcine FAS, with an RMSD of 0.44 Å between the two KR domains, and an RMSD of 0.98 Å between the Ψ domains. Hardwicke *et al.* then evaluated the effectiveness of the tri-domain as catalysis for reduction.[32] Unfortunately, they found a markedly reduced activity. Therefore, the authors caution that that these truncated constructs may be suitable for structural studies, but probably not for mechanistic studies.

1.2.1.1.2 Fungal Type I FAS

Interestingly, in fungal systems all of the catalytic domains are separated between two multidomain polypeptide chains, known as the α and β chains. The first glimpse of this type of structure was obtained in 2006 from the Ban lab, and was from the organism *Thermomyces lanuginosus* at a resolution of 5 Å (PDB ID: 2CDH).[33] The following year, the labs of Steitz and Ban independently solved structures of the yeast type I FAS at relatively high resolutions of 4.0 Å and 3.1 Å respectively (PDB IDs: 2PFF, 4V58, 4V59).[34-35] The fungal

FAS structure from T. *lanuginosus* revealed a 2.6 mDa $\alpha_6\beta_6$ heterododecameric FAS, which is contrast to the 540 kDa homodimer mFAS described previously. The α chain contains the acyltransferase (AT), ER, DH and malonyl/palmitoyl transferase (MPT) domains. The β chain contains the ACP, KR, KS and phosphopantetheine transferase (PT) domains. Jenni *et al.* described the fungal FAS structure as containing three parts, a central wheel, and two domes which enclose the reaction chambers.[33, 36] The fungal FAS is less relevant to this thesis and will not be discussed in detail here.

1.2.1.1.3 Bacterial Type I FAS

While Type I FASs are primarily present in eukaryotes, a few are present in some bacterial species. Boehringer *et al.* and Ciccarelli *et al.* obtained cryo-em structures of bacterial type I FASs at 7.5 Å (PDB ID: 4V8L) for *Mycobacterium smegmatis* and 17.5 Å (PDB IDs: 4V8W, 4V8V) for *Mycobacterium tuberculosis*.[37-38] These structures revealed conservation of overall domain architecture similar architecture to the fungal FAS, with a central wheel, and two domes enclosing the reaction chambers. These structures led Boehringer and Ciccarelli to propose two possible origins for these bacterial type I FASs. The first possibility is simply that bacteria adopted these type I FASs through an early horizontal gene transfer, with the second possibility being that these type I FASs might actually represent an ancient minimal type I FAS, which fungi were later able to refine through evolution. Another interesting anomaly in these structures is the complete lack of linker regions, which are present in eukaryotic FASs. Currently it is not known if this is due to a functional purpose or is simply the result of bacterial genome reduction.

1.2.1.2 Type II FAS

Although Type I FAS is proposed to be more efficient due to the localized nature of the domains, Type II FASs display a greater product diversity.[39] The simplest model system is the Type II FAS in *E. coli*. In this system, the acyl carrier protein AcpP has to interact with more than 10 different catalytic partners, and it does so with high efficiency and fidelity.[39] The final products generated by Type II FASs are long fatty-acyl chains.

Several NMR studies have been carried out on the *E. coli* acyl carrier protein (AcpP); however, because of the transient interactions between a CP and its enzyme partner, it has been difficult to capture AcpP-partner domain interactions. Most recently, Nguyen *et al.* from our group reported the crystallographic, NMR and MD studies of a type II FAS di-domain complex of AcpP and a dehydratase (FabA) from *E. coli*, stabilized via a mechanism-based synthetic crosslinker (PDB ID: 4KEH).[40] Further success was recently obtained when a complex crystal structure was solved of the ketosynthase (FabB) from the same pathway in *E. coli* that is covalently bound to AcpP using a different mechanism-based crosslinker (PDB ID: 5KOF), as detailed in Chapter 4. In addition, Zhang *et al.* reported the complex structure of another dehydratase (FabZ) from *E. coli* with *holo*-AcpP (PDB ID: 4ZJB), providing additional details on the differences between loaded and unloaded ACP variants.[41] This class is covered more in-depth in Chapter 4, with focus on the role allostery plays in the regulation and biosynthesis of fatty acids in bacteria.

1.2.2 Polyketide synthase (PKS)

Nature has co-opted the fatty acid biosynthetic assembly line strategy to produce linear and macrocyclic polyketide natural products by utilizing additional tailoring domains for increased chemical diversity and biological function.[42] PKSs can be categorized into three types depending on their domain organizations. The Type I PKSs assemble into "modules," a collection of enzymatic domains that are covalently linked together on the same polypeptide chain. These Type I PKSs can either catalyze polyketide biosynthesis as multi-module complexes, or as a single iterative module.[43] The Type II PKSs, like their FAS counter-parts, are stand-alone enzymes. Type III PKSs are also stand-alone enzymes, but one domain is capable of performing chain elongation, cyclization and chain release independent of other enzymes. More about PKS will be discussed in Chapter 5.

1.2.3 Non-ribosomal peptide synthetase (NRPS)

NRPSs utilize the CP machinery as well, but with amino acids instead of acyl groups as the building blocks. Nonribosomal peptides (NRPs) are assembled in an assembly line fashion similar to polyketides and fatty acids as described previously. Similarly, the NRPS is categorized into three types. The Type A NRPS is also modular, but it uses amino acids as building blocks that consist of over 400 varieties, including the 20 proteinogenic amino acids (Fig. 1-2a).[44] The large pool of starter and extender units from which to build NRPs has generated a wide array of compounds, and includes many antibiotics such as penicillins, vancomycin, and cyclosporines. Efforts in synthetic biology have shown that further expansion of the NRP products generated by the NRPSs can be achieved through the use of non-standard starter units and extender units.[45] More about NRPS will be covered in Chapter 3.

1.2.4 Hybrids

A strategy employed by some organisms for the generation of complex natural products is through the combination of different types of megasynthases, resulting in the generation of hybrids. In the epothilone biosynthetic assembly, an initial polyketide scaffold (generated by a PKS) is passed to a downstream NRPS, which then passes further to downstream PKS.[46-48] A key requirement for these hybrids is the presence of docking domains that are able to recognize the correct downstream and upstream modules, and to successfully pass the intermediates. Another example is fujimycin, an immunosuppressant that is assembled by a PKS and an NRPS using an assortment of substrates including two methoxymalonyl-acyl carrier proteins (-ACPs), an allylmalonyl-CoA, five methylmalonyl-CoAs, two malonyl-CoAs, and one pipecolic acid molecule.[49]

1.3 Computational techniques

There are several computational techniques available to researchers in pharmacognosy. Common computational tools include bioinformatics, homology modeling, molecular docking, and molecular dynamics. Below, we discuss current progress in megasynthase studies using these computational techniques

1.3.1 Bioinformatics

Traditionally, secondary metabolites were discovered through activity-guided screens. Advances in sequencing and genomics have provided investigators an *in silico* way of identifying secondary metabolites through gene cluster analysis.[50] The genomic data also contains pathway information, vital for engineering efforts of directed biosynthesis.

There are several prediction tools available to analyze PKS and NRPS gene clusters, which have been summarized in reviews by Piel and Boddy, and include SMURF, AntiSMASH, NaPDoS, NP.searcher, ClustScan, CLUSEAN, PKMiner, and NRPS-PKS.[51-56] Another excellent resource is prediction informatics for secondary metabolites (PRISM), a software platform with a web component which can aid in the prediction of final products generated from PKS and NRPS gene clusters.[57-58] Often the same secondary metabolite is produced by more than one species; therefore, these prediction tools provide investigators a method to identify novel compounds.

From the plethora of genomic and metagenomic data available, one can construct phylogenetic trees of FAS, PKS and NRPS gene clusters. These trees can provide rich information about the evolutionary history of natural products. In a recent review by Jenke-Kodama and Dittman, many examples were discussed, including one study that concludes that all KS domains from fungi originated from one common source.[59] In 2018, Wood *et al.* performed an extensive phylogenetic analysis on the ACP of trans-acyltransferase (*trans*-AT) and cis-acyltransferase (*cis*-AT) PKS gene clusters, and discovered that the evolution of ACPs is coupled with downstream ketosynthases (KSs), but is not coupled with upstream KSs.[60] This decoupling from upstream KSs is probably influenced by downstream modules that ensure the correct intermediates are passed down, so that the final product can be biosynthesized with high fidelity. In addition to the phylogenetic analysis, Wood *et al.* generated ACP fingerprints from complex structures in the PDB and identified helices and residues important for docking its partner enzymes. The application of ACP fingerprint data generated by Wood might be useful, when combined with molecular dynamics, to predict interface residues between ACPs and partner enzymes. Similar contact predictions utilizing deep learning for other families of enzymes have been used with relative success.[61-62]

1.3.2 Homology modeling

In the absence of a structure from traditional structural biology methods, the structures can be predicted using homology modeling, which gets its name from using a known homolog that shares the same evolutionary origin as a threading template to make a predictive model of the target protein.[63] Popular template-based models include Rosetta and I-TASSER.[64-67] In addition to the template-based models, in the absence of a suitable template, *ab initio* folding methods do exist. QUARK, developed by the Zhang lab, uses an *ab initio* folding approach and applies a force field-based method.[68] Critical Assessment of Protein Structure Prediction (CASP) experiments are performed to evaluate current homology modeling techniques via blind prediction every two years. This bi-annual competition shows methods developers key areas to focus on for future developments. In addition, there is also a dedicated bi-annual competition focused on protein-protein docking modeling known as the Critical Assessment of Predicted Interactions (CAPRI).[69-71] HADDOCK and CLUSPRO are two web servers which have consistently done well in CAPRI competitions in the past.[70, 72-77] An entire review itself could be written on homology modeling applications in the acetate and non-ribosomal peptide pathways; therefore, in the interest of brevity, key examples are provided in the next few sections.

1.3.2.1 FAS

As mentioned in a previous section, no crystal structure is available of human FAS (hFAS), and because hFAS is associated with clinical conditions such as obesity, diabetes and cancer, a structural model of hFAS is much needed. An *apo* homology model was generated by John *et al.* using the porcine FAS structure as the threading template.[78] This model was used to perform docking, and the quality of the *apo* and ligand bound models were validated using short 20 ns MD simulations. The *apo* structure was screened with a ligand database from the National Cancer Institute (NCI), and identified NSC71039 as a potential strong binder. The authors evaluated the efficacy of NSC71039 in a FASN-expressing cancer line, thus highlighting the effectiveness of molecular modeling approaches in drug design, as well as the versatility of homology modeling for hFAS structural prediction.

Using the mammalian FAS (mFAS) structure solved by Maier *et al.* as a template, a high-quality homology model was developed by Viegas *et al.*, which included the missing TE and ACP domains not present in the reported mFAS structure.[79] 20.0 ns MD simulations were performed with and without the TE and ACP domains to evaluate homology model quality, and the model including the TE and ACP domains showed a decrease in RMSD over time, as compared to the model missing the TE and ACP domains. The validation of homology models by analyzing RMSD fluctuations over time obtained from MD simulations is a common technique in modeling, with a detailed protocol provided by Walker *et al.*.[80] Viegas proposed several hydrophobic residues on the KS surface, including Val261, Phe263 and Leu203 which, in combination with Leu2157, Met2158, Val2160 and Val2176 on ACP, were used to restrain the ACP-KS interactions during the generation of full-length homology modeling.

Thoughtfully guided homology modeling can serve as the basis for biochemical investigations in the absence of an experimentally solved structure.

1.3.2.2 PKS

A key focus in the study of these systems is the elucidation of the protein-protein interactions between CPs and their partner enzymes. In Type II PKS systems, similar to the Type II FAS systems, the CP must share interfaces with its various partner enzymes. This is different from the Type I systems, where the CP is included in the same polypeptide, and a single "hot spot" of CP interacts with its partners may not be as necessary. To investigate the differences between Type I and Type II PKSs, Weissman *et al.* in 2006 set out to compare and contrast recognition regions in both types of systems.[81] A homology model of ACP_4 from 6-deoxyerythronolide B synthase (DEBS) was generated and used to identify helix II as the recognition motif, which was validated through mutagenesis studies, with the focus on the interactions of ACP with its partners, the phosphopantetheinyl transferase (PPTase) and TE. In a later study, Alekseyev *et al.* obtained a solution structure of one ACP from DEBS and performed homology modeling to make predictive structures of the other five ACPs present in this system.[82]

We recently obtained a crystal structure of a anthranilate:CoA ligase (AuaEII), which is responsible for the generation of an uncommon starter unit anthraniloyl-CoA used during the biosynthesis of aurachins, a class of quinoline alkaloids.[83] AuaE, which is present in the same pathway, is responsible for transferring the anthraniloyl-CoA onto ACP, and shares high sequence similarity to AuaEII. In the absence of an experimentally derived structure, molecular dynamics guided homology modeling was used to develop high-quality representations of AuaE in two distinct conformations. Short 100 ns simulations were used to measure stability in the two original homology models, as well as refine sidechain geometry. The MD-refined models provided a molecular basis for hinge movement in AuaE, via stabilization of a salt bridge between Arg408 and Asp410, while the same arginine residue in AuaEII is unable to form this salt bridge due to an interaction with a hydroxyl functional group on the anthraniloyl-AMP ribose ring. This study showed the effectiveness of obtaining additional structural and functional details on structures through higher quality MD-guided homology modeling approaches.

1.3.3 Docking

Traditionally, small-molecule docking approaches focus on non-covalent receptor and ligand interactions. Enzymes in these pathways often covalently load their intermediates on a post-translationally modified phosphopantetheine moiety, which is attached to the CP domain. Therefore, covalent docking approaches (as opposed to non-covalent docking) may be needed. A recent review by Scarpino *et al.* evaluated available covalent docking tools, including AutoDock4, CovDock, FITTED, GOLD, ICM-Pro and MOE.[84] The six tools could model covalently bound ligands with 40-70 % success rate, with a lower-success rate for long and flexible ligands that are covalently bound.[84-89] The phosphopantetheine bound intermediates in these pathways are relatively long and flexible; therefore, we expect that its covalent docking will be a major issue that needs to be addressed with careful consideration. Covalent docking was used by our lab to model substrate mimics, though the substrate mimics were compared and contrasted to known substrate models for which NMR data is present in the protein data bank (PDB) (detailed in Chapters 2-4).[90]

1.3.4 Molecular Dynamics (MD)

Experimental techniques, including X-ray crystallography, NMR, FRET, and EM, are powerful tools that allow us to visualize biomolecules, though we're often limited to specific time scales and the ensemble averages of the target proteins. Starting from the first example, sixty years ago, Kendrew *et al.* reported the first crystal structure of a protein, myoglobin (PDB ID: 1MBN).[91] These early crystallography experiments, while providing atomic-resolution models, instilled in us a static-view of proteins, RNA and DNA. Today, researchers show the importance of protein dynamics to its function, and a static picture is no longer satisfactory. Although protein structures solved by Nuclear magnetic resonance (NMR) can capture multiple protein conformations, NMR structures cannot resolve dynamic structures in the femto- and picosecond timescales.

To circumvent the above issues, computational structural biologists model biomolecules *in silico*. Molecular behavior at the atomic level is described by the time-dependent Schrödinger equation (1.1), for which no analytical solution exists for non-trivial systems, and computationally impracticable even on today's fastest computers.[92]

$$i\hbar\frac{\partial}{\partial t}|\Psi(\mathbf{r},t)\,\rangle = \widehat{H}|\Psi(\mathbf{r},t)\rangle \tag{1.1}$$

Using the Born-Oppenheimer approximation, the wave function of a molecule can be described in terms of its nuclear and electronic components. This allows us to treat atoms as particles, and bonds can be represented as springs, thus providing a framework to represent our target systems classically. In MD simulations, we can obtain particle motion as a function of time, and thus resolve time regimes far beyond traditional structural biology techniques. Because MD simulations can sample different time-scales, we can gain valuable insight into protein-protein and protein-substrate interactions, as well as conformational changes. Karplus *et al.* classifies simulations into three main types.[93] The purpose of the first type is to sample configuration space, and this is used in X-ray crystallography refinement through annealing protocols. The second type is used to describe systems at equilibrium, with the goal of obtaining thermodynamic data. These first two types can also be accomplished through Monte Carlo simulations. The final type is to sample protein dynamics, and for this molecular dynamics (MD) is the only option. Investigators can perform a variety of simulations and analyses using traditional MD programs such as AMBER, CHARMM, GROMACS, and NAMD.[94-98] The frontier area is MD method development, which is a result of demands by the end-users whose needs push the method development forward.

Modeling enzymes from the acetate and non-ribosomal peptide pathways presents a challenge, because traditional MD packages are unable to model the long-flexible phosphopantetheine bound intermediates observed in these systems. Therefore, a major hurdle facing MD in these systems is the parameterization of electrostatic and bond properties for the covalently-bound starter units, extender units, and intermediates. To surmount this hurdle, I developed a specialized forcefield for use in AMBER to probe the dynamics of these systems (detailed in Chapter 2). The development of the force field is based on previous force field development methods and is available for free to the molecular dynamics and natural product communities. It includes the forcefields of PPant-tethered starter units, extender units, and biosynthetic intermediates.

1.3.4.1 FAS

Molecular dynamics (MD) simulations performed by Chen *et al.* on the *apo*, *holo*, and acyl forms on the Type II ACP (AcpP) from *E. coli* revealed two substrate binding modes, and AcpP has an adjustable cavity to accommodate substrates of various lengths.[17, 99-100] Simulations were performed in which the loaded phosphopantetheine was positioned in the sequestered ACP cavity, or exposed to the solvent.[99] Different chain lengths were also explored, with the octanoyl-bound intermediate being the most suited for substrate binding in the AcpP cavity. A total of 19 simulations were performed and sampled up to 50 ns. Chen suggests from his computational models that the loop between the second and third helices is important for recognition between different enzyme partners of AcpP, and when it's in the *apo*- or *holo*- forms, this loop region, with highly acidic residues, fluctuates significantly. Upon loading with the fatty acid "cargo", this loop region has an increased flexibility, thus allowing this electronegative patch of ACP to interact with its partner via the "arginine-rich groove".

Medina *et al.* performed a QM/MM study to inspect the proposed catalytic mechanism underlying reduction of β-ketoacyl to β-hydroxyacyl group via oxidation of a NADPH cofactor by the ketoreductase (KR) of hFAS.[101] Several models were tested, and the result suggests that the KR reaction proceeds in two steps. First, the β-carbon of the acyl substrate is subjected to nucleophilic attack by the NADPH hydride, and the proton is replenished by an asynchronous deprotonation of nearby Tyr2034. The model then predicts that the reprotonation of Tyr2034 proceeds by the deprotonation of $NADP^+$ 3'-OH group, which is coordinated with a neighboring Lys1995. This study showed that QM/MM studies can provide high quality models to probe highly reactive intermediates.

1.3.4.2 PKS

An early computational study performed by Yeates, Tang and Houk in 2014 explored the role of allostery in the active site architecture and dynamics of a PKS.[102] Specifically, they investigated a lovastatin PKS (LovD) variant that was 1,000-fold more efficient than the wildtype. Microsecond MD simulations supported that long-range mutations can alter the active site dynamics. Furthermore, MD simulations can reproduce a conformational change of the active site channel, as was observed in crystallographic studies. In addition, QM studies were performed, and the MD simulations matched the QM geometrical arrangements, with a second-shell tyrosine constrained in an active conformation for the majority of the simulation.

Interestingly, several conformations were reported for the crystal structures of LovD variants. The MD studies performed on these variants also showed a sampling of various conformational states due to mutations. Noteworthy from this study was the observation that when the ACP was bound to LovD, there was a stabilization in the active site geometry and reduction in protein dynamics. This suggests that the PKS has an inherent flexibility, and that crystal packing might induce modified conformations in crystallographic studies.[102] The authors summarized that PKS engineering should focus not only on the active site, but should also should take into account the importance of protein-protein interactions and protein dynamics.

Bravo-Rodriguez *et al.* reported the potential application of MD towards engineering PKS systems. [103] Specifically, they explored the potential application of MD to incorporate non-native starter and extender units. An MD model was developed for an acyltransferase ($AT5_{mon}$) of the monensin PKS. The model was used to engineer the active site in $AT5_{mon}$ for incorporation of a non-native starter unit by computationally guided predictions.

In our group, Ellis *et al.* developed oxetane substrate mimics for Type II PKSs. In order to evaluate the capability of the mimics replicating native substrate characteristics, MD simulations and free energy calculations were performed on the native substrates and oxetane mimics. The binding affinities of the native substrate and the mimic as within one standard deviation, suggesting that the mimics accurately depict the natural substrates in regards to binding. Principal component analysis (PCA) was also used to evaluate sampling of the receptor with either substrate present, showing sufficient overlap.[104]

1.3.4.3 NRPS

Blouodoff *et al.* obtained the first crystal structure of the condensation domain of calcium-dependent antibiotic (CDA) synthetase (CDA-C1) from *Streptomyces coelicolor*.[105] During the crystallographic studies, the team noted that the structures obtained were different from previously solved condensation domains. Using small X-ray angle scattering (SAXS), they noted that all of previously reported conformations were present in the ensemble sampled from the SAXS experiment. To further investigate, they performed targeted MD simulations, normal mode analyses (NMAs), and energy-minimized linear interpolation. These three computational techniques reproduced the conformations observed from different X-ray studies and suggested that there were no unfavorable conformations for the condensation domains. Both open and closed conformations are observed, which the authors speculate could be controlled allosterically via a communication network of nearby NRPS domains.

An important area in the study of CPs is the mechanism of chain release of the final product. The myxalamid biosynthetic pathway in *Stigmatella aurantiaca* is composed of six PKS modules and one terminal NRPS module. The terminal reductase domain catalyzes a four-electron, non-processive reduction to produce myxalamids, a family of secondary metabolites. Although common in nature, the lack of structural and dynamic information for termination domains has prevented engineering attempts to improve or alter their function. This was the first MD simulation that deciphers protein-substrate interactions in the active site of a NRPS reductase.[106] In addition, MD and docking studies were used to understand protein-protein interactions between the MxaA reductase (MxaAR) and its corresponding PCP (MxaAPCP). This study is further detailed in Chapter 3.

Dowling *et al.* recently determined the crystal structure of an NRPS cyclization domain, EpoB (EpoBCy), which is responsible for assembling the thioazole moiety of epothilones, a class of PKS-NRPS hybrids with anti-cancer activity.[46] The growing substrates are passed between large multi-domain NRPS modules by way of docking domains. A major engineering goal in these pathways is to change the linear logic of modular processivity. The ability to delete, insert, or swap modules would give the protein designer greater control over chain length and tailoring. It was proposed that the EpoA docking domain (EpoAdd) and the EpoB docking domain (EpoBdd) serve as the recognition domains, which prevent cross communication between wrong modules. More than one conformation of EpoBdd was observed by crystallographic studies. MD simulations revealed high flexibility in the docking domain, consistent with the crystallographic studies. Furthermore, MD simulations were used to produce snapshots of open and closed states of EpoBCy, resulting in a proposed model that correlates protein conformational changes to active site availability. In order to gain a better understanding of intra- and inter-modular interactions in these systems, more native structures need to be solved, including multidomain structures of CP and its partners so that different conformational states of CPs can be evaluated.[107]

1.4 Objectives and overview of dissertation research

Crystal structures of enzymes bound with their natural substrates have proven to be difficult, if not impossible, to obtain for these megasynthases because the intermediates are highly reactive. The innate reactivity of PPant-tethered intermediates has hampered our mechanistic understanding of protein-substrate and protein-protein interactions during the biosynthesis of these natural products. Therefore, there is a need to address the knowledge gap of the protein-protein and protein-substrate interactions in these megasynthases through the development of alternative methods to elucidate the structure, function and dynamics. The goal of my research has been to better understand how PKSs, NRPSs and FASs biosynthesize these complex molecules, and how protein-substrate and protein-protein interactions affect product generation, using MD simulations.

Chapter 2 details the development of a computational force field that provides a tool to model these enzymes *in silico* using molecular dynamics. In chapter 3, MD simulations served as the basis for the biochemical investigations of an NRPS reductase domain, in which MD simulation helps engineer mutants that changed the product outcomes for biofuel engineering. Chapter 4 describes MD simulations and a crystal structures of a protein complex that contains the ketosynthase (FabB) and its partner carrier protein (AcpP) from type II FAS, in which MD simulation guides the mutations of FabB that led to a change of product outcomes that can be utilized for future biofuel production. The dynamics of a type III PKS at different pHs is investigated in Chapter 5, where MD simulations help explain how this PKS biosynthesizes different products at different pHs. The results of the dissertation research demonstrate that the development and application of MD methods in parallel with structural and biochemical studies can aid engineering efforts in the biosynthesis of "unnatural" natural products and engineered biofuels.

References

1. Dewick, P. M., *Medicinal natural products : a biosynthetic approach*. 3rd edition. ed.; Wiley, A John Wiley and Sons, Ltd., Publication: Chichester, West Sussex, United Kingdom, 2009; p x, 539 pages.
2. Van Voorhis, W. C.; Hooft van Huijsduijnen, R.; Wells, T. N., Profile of William C. Campbell, Satoshi Omura, and Youyou Tu, 2015 Nobel Laureates in Physiology or Medicine. *Proc Natl Acad Sci U S A* **2015,** *112* (52), 15773-6.
3. Kinghorn, A. D., The role of pharmacognosy in modern medicine. *Expert Opin Pharmacother* **2002,** *3* (2), 77-9.
4. Fersht, A. R., Profile of Martin Karplus, Michael Levitt, and Arieh Warshel, 2013 nobel laureates in chemistry. *Proc Natl Acad Sci U S A* **2013,** *110* (49), 19656-7.
5. Moore, B. S.; Hertweck, C.; Hopke, J. N.; Izumikawa, M.; Kalaitzis, J. A.; Nilsen, G.; O'Hare, T.; Piel, J.; Shipley, P. R.; Xiang, L.; Austin, M. B.; Noel, J. P., Plant-like biosynthetic pathways in bacteria: from benzoic acid to chalcone. *J Nat Prod* **2002,** *65* (12), 1956-62.
6. Keller, N. P.; Turner, G.; Bennett, J. W., Fungal secondary metabolism - from biochemistry to genomics. *Nat Rev Microbiol* **2005,** *3* (12), 937-47.
7. Gulder, T. A.; Moore, B. S., Chasing the treasures of the sea - bacterial marine natural products. *Curr Opin Microbiol* **2009,** *12* (3), 252-60.
8. Fischbach, M.; Voigt, C. A., Prokaryotic gene clusters: a rich toolbox for synthetic biology. *Biotechnol J* **2010,** *5* (12), 1277-96.
9. Weissman, K. J., The structural biology of biosynthetic megaenzymes. *Nat Chem Biol* **2015,** *11* (9), 660-70.
10. O'Connor, S. E.; Maresh, J. J., Chemistry and biology of monoterpene indole alkaloid biosynthesis. *Nat Prod Rep* **2006,** *23* (4), 532-47.
11. Gershenzon, J.; Dudareva, N., The function of terpene natural products in the natural world. *Nat Chem Biol* **2007,** *3* (7), 408-14.

12. Ferrer, J. L.; Austin, M. B.; Stewart, C., Jr.; Noel, J. P., Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiol Biochem* **2008,** *46* (3), 356-70.

13. Kochanowska-Karamyan, A. J.; Hamann, M. T., Marine indole alkaloids: potential new drug leads for the control of depression and anxiety. *Chem Rev* **2010,** *110* (8), 4489-97.

14. Matsuda, Y.; Abe, I., Biosynthesis of fungal meroterpenoids. *Nat Prod Rep* **2016,** *33* (1), 26-53.

15. Khosla, C.; Herschlag, D.; Cane, D. E.; Walsh, C. T., Assembly line polyketide synthases: mechanistic insights and unsolved problems. *Biochemistry* **2014,** *53* (18), 2875-83.

16. Staunton, J.; Weissman, K. J., Polyketide biosynthesis: a millennium review. *Natural Product Reports* **2001,** *18* (4), 380-416.

17. Chan, D. I.; Vogel, H. J., Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem J* **2010,** *430* (1), 1-19.

18. Korman, T. P.; Crawford, J. M.; Labonte, J. W.; Newman, A. G.; Wong, J.; Townsend, C. A.; Tsai, S. C., Structure and function of an iterative polyketide synthase thioesterase domain catalyzing Claisen cyclization in aflatoxin biosynthesis. *Proc Natl Acad Sci U S A* **2010,** *107* (14), 6246-51.

19. Townsend, C. A., Aflatoxin and deconstruction of type I, iterative polyketide synthase function. *Nat Prod Rep* **2014,** *31* (10), 1260-5.

20. Zhang, Y. M.; Marrakchi, H.; White, S. W.; Rock, C. O., The application of computational methods to explore the diversity and structure of bacterial fatty acid synthase. *J Lipid Res* **2003,** *44* (1), 1-10.

21. Dechert-Schmitt, A. M.; Schmitt, D. C.; Gao, X.; Itoh, T.; Krische, M. J., Polyketide construction via hydrohydroxyalkylation and related alcohol C-H functionalizations: reinventing the chemistry of carbonyl addition. *Nat Prod Rep* **2014,** *31* (4), 504-13.

22. Gaitatzis, N.; Silakowski, B.; Kunze, B.; Nordsiek, G.; Blocker, H.; Hofle, G.; Muller, R., The biosynthesis of the aromatic myxobacterial electron transport inhibitor stigmatellin is directed by a novel type of modular polyketide synthase. *J Biol Chem* **2002,** *277* (15), 13082-90.

23. Smith, S.; Tsai, S. C., The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat Prod Rep* **2007,** *24* (5), 1041-72.

24. Voet, D.; Voet, J. G., *Biochemistry*. 4th ed.; John Wiley & Sons: Hoboken, NJ, 2011; p xxv, 1428, 53 p.

25. Maier, T.; Leibundgut, M.; Ban, N., The crystal structure of a mammalian fatty acid synthase. *Science* **2008,** *321* (5894), 1315-22.

26. Maier, T.; Leibundgut, M.; Boehringer, D.; Ban, N., Structure and function of eukaryotic fatty acid synthases. *Q Rev Biophys* **2010,** *43* (3), 373-422.

27. Maier, T.; Jenni, S.; Ban, N., Architecture of mammalian fatty acid synthase at 4.5 A resolution. *Science* **2006,** *311* (5765), 1258-62.

28. Connolly, M. L., Analytical Molecular-Surface Calculation. *J Appl Crystallogr* **1983,** *16* (Oct), 548-558.

29. Rittner, A.; Paithankar, K. S.; Huu, K. V.; Grininger, M., Characterization of the Polyspecific Transferase of Murine Type I Fatty Acid Synthase (FAS) and Implications for Polyketide Synthase (PKS) Engineering. *ACS Chem Biol* **2018,** *13* (3), 723-732.

30. Chakravarty, B.; Gu, Z.; Chirala, S. S.; Wakil, S. J.; Quiocho, F. A., Human fatty acid synthase: structure and substrate selectivity of the thioesterase domain. *Proc Natl Acad Sci U S A* **2004,** *101* (44), 15567-72.

31. Zhang, W.; Chakravarty, B.; Zheng, F.; Gu, Z.; Wu, H.; Mao, J.; Wakil, S. J.; Quiocho, F. A., Crystal structure of FAS thioesterase domain with polyunsaturated fatty acyl adduct and inhibition by dihomo-gamma-linolenic acid. *Proc Natl Acad Sci U S A* **2011,** *108* (38), 15757-62.

32. Hardwicke, M. A.; Rendina, A. R.; Williams, S. P.; Moore, M. L.; Wang, L.; Krueger, J. A.; Plant, R. N.; Totoritis, R. D.; Zhang, G.; Briand, J.; Burkhart, W. A.; Brown, K. K.; Parrish, C. A., A human fatty acid synthase inhibitor binds beta-ketoacyl reductase in the keto-substrate site. *Nat Chem Biol* **2014,** *10* (9), 774-9.

33. Jenni, S.; Leibundgut, M.; Boehringer, D.; Frick, C.; Mikolasek, B.; Ban, N., Structure of fungal fatty acid synthase and implications for iterative substrate shuttling. *Science* **2007,** *316* (5822), 254-61.

34. Lomakin, I. B.; Xiong, Y.; Steitz, T. A., The crystal structure of yeast fatty acid synthase, a cellular machine with eight active sites working together. *Cell* **2007,** *129* (2), 319-32.

35. Leibundgut, M.; Jenni, S.; Frick, C.; Ban, N., Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **2007,** *316* (5822), 288-90.

36. Jenni, S.; Leibundgut, M.; Maier, T.; Ban, N., Architecture of a fungal fatty acid synthase at 5 A resolution. *Science* **2006,** *311* (5765), 1263-7.

37. Ciccarelli, L.; Connell, S. R.; Enderle, M.; Mills, D. J.; Vonck, J.; Grininger, M., Structure and conformational variability of the mycobacterium tuberculosis fatty acid synthase multienzyme complex. *Structure* **2013,** *21* (7), 1251-7.

38. Boehringer, D.; Ban, N.; Leibundgut, M., 7.5-A cryo-em structure of the mycobacterial fatty acid synthase. *J Mol Biol* **2013,** *425* (5), 841-9.

39. White, S. W.; Zheng, J.; Zhang, Y. M.; Rock, The structural biology of type II fatty acid biosynthesis. *Annu Rev Biochem* **2005,** *74*, 791-831.

40. Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R.; Bruegger, J.; Caldara-Festin, G.; Finzel, K.; Jackson, D. R.; Ishikawa, F.; O'Dowd, B.; McCammon, J. A.; Opella, S. J.; Tsai, S. C.; Burkart, M. D., Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **2014,** *505* (7483), 427-31.

41. Zhang, L.; Xiao, J.; Xu, J.; Fu, T.; Cao, Z.; Zhu, L.; Chen, H. Z.; Shen, X.; Jiang, H.; Zhang, L., Crystal structure of FabZ-ACP complex reveals a dynamic seesaw-like catalytic mechanism of dehydratase in fatty acid biosynthesis. *Cell Res* **2016,** *26* (12), 1330-1344.

42. Sattely, E. S.; Fischbach, M. A.; Walsh, C. T., Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. *Nat Prod Rep* **2008,** *25* (4), 757-93.

43. Staunton, J.; Weissman, K. J., Polyketide biosynthesis: a millennium review. *Nat Prod Rep* **2001,** *18* (4), 380-416.

44. Challis, G. L.; Naismith, J. H., Structural aspects of non-ribosomal peptide biosynthesis. *Curr Opin Struct Biol* **2004,** *14* (6), 748-56.

45. Hur, G. H.; Vickery, C. R.; Burkart, M. D., Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat Prod Rep* **2012,** *29* (10), 1074-98.

46. Dowling, D. P.; Kung, Y.; Croft, A. K.; Taghizadeh, K.; Kelly, W. L.; Walsh, C. T.; Drennan, C. L., Structural elements of an NRPS cyclization domain and its intermodule docking domain. *Proc Natl Acad Sci U S A* **2016,** *113* (44), 12432-12437.

47. Osswald, C.; Zipf, G.; Schmidt, G.; Maier, J.; Bernauer, H. S.; Muller, R.; Wenzel, S. C., Modular Construction of a Functional Artificial Epothilone Polyketide Pathway. *Acs Synth Biol* **2014,** *3* (10), 759-772.

48. Walsh, C. T., Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Natural Product Reports* **2016,** *33* (2), 127-135.

49. Lechner, A.; Wilson, M. C.; Ban, Y. H.; Hwang, J. Y.; Yoon, Y. J.; Moore, B. S., Designed Biosynthesis of 36-Methyl-FK506 by Polyketide Precursor Pathway Engineering. *Acs Synth Biol* **2013,** *2* (7), 379-383.

50. Helfrich, E. J. N.; Reiter, S.; Piel, J., Recent advances in genome-based polyketide discovery (vol 29, pg 107, 2014). *Curr Opin Biotech* **2014,** *29*, 184-184.

51. Adamek, M.; Spohn, M.; Stegmann, E.; Ziemert, N., Mining Bacterial Genomes for Secondary Metabolite Gene Clusters. *Antibiotics: Methods and Protocols* **2017,** *1520*, 23-47.

52. Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R., The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *Plos One* **2012,** *7* (3).

53. Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D., SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* **2010,** *47* (9), 736-741.

54. Blin, K.; Wolf, T.; Chevrette, M. G.; Lu, X. W.; Schwalen, C. J.; Kautsar, S. A.; Duran, H. G. S.; Santos, E. L. C. D. L.; Kim, H. U.; Nave, M.; Dickschat, J. S.; Mitchell, D. A.; Shelest, E.; Breitling, R.; Takano, E.; Lee, S. Y.; Weber, T.; Medema, M. H., antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research* **2017,** *45* (W1), W36-W41.

55. Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R., antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **2011,** *39*, W339-W346.

56. Boddy, C. N., Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. *J Ind Microbiol Biot* **2014,** *41* (2), 443-450.

57. Skinnider, M. A.; Johnston, C. W.; Edgar, R. E.; Dejong, C. A.; Merwin, N. J.; Rees, P. N.; Magarvey, N. A., Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *P Natl Acad Sci USA* **2016,** *113* (42), E6343-E6351.

58. Skinnider, M. A.; Dejong, C. A.; Rees, P. N.; Johnston, C. W.; Li, H. X.; Webster, A. L. H.; Wyatt, M. A.; Magarvey, N. A., Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Research* **2015,** *43* (20), 9645-9662.

59. Jenke-Kodama, H.; Dittmann, E., Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges. *Natural Product Reports* **2009,** *26* (7), 874-883.

60. Wood, D. A. V.; Keatinge-Clay, A. T., The modules of trans-acyltransferase assembly lines redefined with a central acyl carrier protein. *Proteins* **2018,** *86* (6), 664-675.

61. Adhikari, B.; Hou, J.; Cheng, J. L., Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins* **2018,** *86*, 84-96.

62. Wang, S.; Sun, S. Q.; Xu, J. B., Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins* **2018,** *86*, 67-77.

63. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational Methods in Drug Discovery. *Pharmacol Rev* **2014,** *66* (1), 334-395.

64. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **2010,** *5* (4), 725-738.

65. Yang, J. Y.; Yan, R. X.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **2015,** *12* (1), 7-8.

66. Zhang, Y., I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics* **2008,** *9*.

67. Song, Y. F.; DiMaio, F.; Wang, R. Y. R.; Kim, D.; Miles, C.; Brunette, T. J.; Thompson, J.; Baker, D., High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013,** *21* (10), 1735-1742.

68. Xu, D.; Zhang, Y., Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **2012,** *80* (7), 1715-1735.

69. Janin, J., Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins-Structure Function and Genetics* **2002,** *47* (3), 257-257.

70. Janin, J.; Henrick, K.; Moult, J.; Ten Eyck, L.; Sternberg, M. J. E.; Vajda, S.; Vasker, I.; Wodak, S. J., CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **2003,** *52* (1), 2-9.

71. Wodak, S. J.; Janin, J., Modeling protein assemblies: Critical Assessment of Predicted Interactions (CAPRI) 15 years hence. 6TH CAPRI evaluation meeting April 17-19 Tel-Aviv, Israel. *Proteins* **2017,** *85* (3), 357-358.

72. Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J., ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **2004,** *20* (1), 45-50.

73. Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S., PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins* **2006,** *65* (2), 392-406.

74. Kozakov, D.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Vajda, S., How good is automated protein docking? *Proteins* **2013,** *81* (12), 2159-2166.

75. Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S., The ClusPro web server for protein-protein docking. *Nat Protoc* **2017,** *12* (2), 255-278.

76. Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J., HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **2003,** *125* (7), 1731-1737.

77. van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J., The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **2016,** *428* (4), 720-725.

78. John, A.; Umashankar, V.; Samdani, A.; Sangeetha, M.; Krishnakumar, S.; Deepa, P. R., In Silico Structure Prediction of Human Fatty Acid Synthase-Dehydratase: A Plausible Model for Understanding Active Site Interactions. *Bioinform Biol Insights* **2016,** *10*, 143-54.

79. Viegas, M. F.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A., Modeling of Human Fatty Acid Synthase and in Silico Docking of Acyl Carrier Protein Domain and Its Partner Catalytic Domains. *J Phys Chem B* **2018,** *122* (1), 77-85.

80. Nurisso, A.; Daina, A.; Walker, R. C., A practical introduction to molecular dynamics simulations: applications to homology modeling. *Methods Mol Biol* **2012,** *857*, 137-73.

81. Weissman, K. J.; Hong, H.; Popovic, B.; Meersman, F., Evidence for a protein-protein interaction motif on an acyl carrier protein domain from a modular polyketide synthase. *Chemistry & Biology* **2006,** *13* (6), 625-636.

82. Alekseyev, V. Y.; Liu, C. W.; Cane, D. E.; Puglisi, J. D.; Khosla, C., Solution structure and proposed domain-domain recognition interface of an acyl carrier protein domain from a modular polyketide synthase. *Protein Sci* **2007,** *16* (10), 2093-2107.

83. Jackson, D. R.; Tu, S. S.; Nguyen, M.; Barajas, J. F.; Schaub, A. J.; Krug, D.; Pistorius, D.; Luo, R.; Muller, R.; Tsai, S. C., Structural Insights into Anthranilate Priming during Type II Polyketide Biosynthesis. *ACS Chem Biol* **2016,** *11* (1), 95-103.

84. Scarpino, A.; Ferenczy, G. G.; Keseru, G. M., Comparative Evaluation of Covalent Docking Tools. *J Chem Inf Model* **2018**.

85. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **2009,** *30* (16), 2785-91.

86. Zhu, K.; Borrelli, K. W.; Greenwood, J. R.; Day, T.; Abel, R.; Farid, R. S.; Harder, E., Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. *J Chem Inf Model* **2014,** *54* (7), 1932-40.

87. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible ligand docking. *Abstr Pap Am Chem S* **1997,** *214*, 154-Comp.

88. Abagyan, R.; Totrov, M.; Kuznetsov, D., Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry* **1994,** *15* (5), 488-506.

89. Corbeil, C. R.; Williams, C. I.; Labute, P., Variability in docking success rates due to dataset preparation. *J Comput Aid Mol Des* **2012,** *26* (6), 775-786.

90. Shakya, G.; Rivera, H., Jr.; Lee, D. J.; Jaremko, M. J.; La Clair, J. J.; Fox, D. T.; Haushalter, R. W.; Schaub, A. J.; Bruegger, J.; Barajas, J. F.; White, A. R.; Kaur, P.; Gwozdziowski, E. R.; Wong, F.; Tsai, S. C.; Burkart, M. D., Modeling linear and cyclic PKS intermediates through atom replacement. *J Am Chem Soc* **2014,** *136* (48), 16792-9.

91. Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C., 3-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958,** *181* (4610), 662-666.

92. Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H. F.; Shaw, D. E., Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu Rev Biophys* **2012,** *41*, 429-452.

93. Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **2002,** *9* (9), 646-652.

94. Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Derutti, D. S.; Cheatham, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, i.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *AMBER 2018*, University of California, San Francisco: 2018.

95. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry* **2009,** *30* (10), 1545-1614.

96. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.

97. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005,** *26* (16), 1781-1802.

98. Kumar, S.; Huang, C.; Zheng, G.; Bohm, E.; Bhatele, A.; Phillips, J. C.; Yu, H.; Kale, L. V., Scalable molecular dynamics with NAMD on the IBM Blue Gene/L system. *Ibm J Res Dev* **2008,** *52* (1-2), 177-188.

99. Chan, D. I.; Stockner, T.; Tieleman, D. P.; Vogel, H. J., Molecular dynamics simulations of the Apo-, Holo-, and acyl-forms of Escherichia coli acyl carrier protein. *J Biol Chem* **2008,** *283* (48), 33620-9.

100. Chan, D. I.; Tieleman, D. P.; Vogel, H. J., Molecular dynamics simulations of beta-ketoacyl-, beta-hydroxyacyl-, and trans-2-enoyl-acyl carrier proteins of Escherichia coli. *Biochemistry* **2010,** *49* (13), 2860-8.

101. Medina, F. E.; Neves, R. P.; Ramos, M. J.; Fernandes, P. A., A QM/MM study of the reaction mechanism of human beta-ketoacyl reductase. *Phys Chem Chem Phys* **2016,** *19* (1), 347-355.

102. Jimenez-Oses, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N., The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat Chem Biol* **2014,** *10* (6), 431-6.

103. Bravo-Rodriguez, K.; Klopries, S.; Koopmans, K. R.; Sundermann, U.; Yahiaoui, S.; Arens, J.; Kushnir, S.; Schulz, F.; Sanchez-Garcia, E., Substrate Flexibility of a Mutated Acyltransferase Domain and Implications for Polyketide Biosynthesis. *Chem Biol* **2015,** *22* (11), 1425-30.

104. Ellis, B. D.; Milligan, J. C.; White, A. R.; Duong, V.; Altman, P. X.; Mohammed, L. Y.; Crump, M. P.; Crosby, J.; Luo, R.; Vanderwal, C. D.; Tsai, S. C., An Oxetane-Based Polyketide Surrogate To Probe Substrate Binding in a Polyketide Synthase. *Journal of the American Chemical Society* **2018,** *140* (15), 4961-4964.

105. Bloudoff, K.; Rodionov, D.; Schmeing, T. M., Crystal Structures of the First Condensation Domain of CDA Synthetase Suggest Conformational Changes during the Synthetic Cycle of Nonribosomal Peptide Synthetases. *Journal of Molecular Biology* **2013,** *425* (17), 3137-3150.

106. Barajas, J. F.; Phelan, R. M.; Schaub, A. J.; Kliewer, J. T.; Kelly, P. J.; Jackson, D. R.; Luo, R.; Keasling, J. D.; Tsai, S. C., Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem Biol* **2015,** *22* (8), 1018-29.

107. Strieker, M.; Tanovic, A.; Marahiel, M. A., Nonribosomal peptide synthetases: structures and dynamics. *Curr Opin Struct Biol* **2010,** *20* (2), 234-40.

CHAPTER 2

Fatty Acid and Natural Product Biosynthesis Force Field for Investigation of Conformational Dynamics

2.1 Summary

Fatty acid synthases (FAS), polyketide synthases (PKS), and nonribosomal peptide synthetases (NRPS) can generate chemically diverse and complex bioactive natural products in an assembly line fashion. A common thread between these systems is the tethering of the growing intermediates and extender units to a carrier protein that has been post-translationally modified with a phosphopantetheine (PPant) prosthetic group. Lack of knowledge about how PPant-tethered intermediates are sequestered and transported during natural product biosynthesis has hampered our understanding of protein-substrate and protein-protein interactions during the biosynthesis of these natural products, ultimately impeding the engineering of these systems for the generation of "unnatural" natural products. This chapter outlines the development of a pantetheinyl-protein force field with comparison to experimental structural characterization, electronic structure calculations, and normal-mode frequencies. This initial force field represents the first force field specifically designed to probe carrier protein-mediated natural product biosynthesis through *in silico* methods and provides a framework for an all-inclusive force field capable of handling natural or unnatural building blocks. Ultimately, this new tool will provide experimentalists with the ability to model these systems in silico and aid in the exploitation of these systems for the production of novel compounds.

2.2 Introduction

Carrier protein-based biosynthesis of fatty acids, polyketides and nonribosomal peptides provides a plethora of complex, bioactive natural products including valuable pharmaceuticals and precious commodity chemicals (Fig 1-1).[1-4] These industrially useful compounds are assembled incrementally through the addition of extender units to the starter unit by multi-domain fatty acid synthases (FAS), polyketide synthases (PKS), or nonribosomal peptide synthetases (NRPS).[5-7] The building blocks of these systems are primarily acyl malonates or amino acids. Their intermediate products, often highly reactive, are shuttled between the catalytic domains via the carrier proteins (CPs) in a well-choreographed order that results in the generation of the final product with high fidelity.[8] Recent efforts have been made to engineer these systems to expand their product diversity as well as optimize systems for expression in heterologous hosts.[9-10] A major hurdle that remains is our poor understanding of the transient substrate-protein interactions between the CP with its pantetheine bound intermediates, as well as protein-protein interactions between CPs and their catalytic partner domains.[11] Molecular dynamics (MD) and other computational techniques can be used to provide models of these transient interactions that are difficult to capture experimentally, thus providing an additional tool to increase yields and expand diversity when we utilize the mega-synthases for the biosynthesis of "unnatural" natural products.[11]

The simplest model system available for study is the Type II FAS in *E. coli*. In this system, the acyl carrier protein AcpP has to interact with more than 10 different catalytic partners, and it does so with high efficiency and fidelity.[1] The final products generated are long fatty-acyl chains. Next comes PKS. Nature has co-opted the assembly line strategy to produce linear and macrocyclic polyketide natural products by utilizing additional tailoring domains for increased chemical diversity and biological function.[2] Third, NRPSs utilize the carrier protein machinery as well, with elongation by amino acids instead of acyl groups. Further expansion of the products generated by these systems can be achieved through the use of non-standard starter units and extender units. One example is fujimycin, an immunosuppressant that is assembled by a PKS and an NRPS using an assortment of substrates including two methoxymalonyl-acyl carrier proteins (-ACPs), one allylmalonyl-CoA, five methylmalonyl-CoAs, two malonyl-CoAs, and one pipecolic acid molecule (Fig 1-1).[12]

One example that combines *in silico*, *in vitro* and *in vivo* studies was reported by Dowling *et al.*, who recently determined the crystal structure of an NRPS cyclization domain EpoB (EpoBCy) that is responsible for assembling the thioazole moiety of epothilones, a class of natural products with anti-cancer activity (Fig 1-1).[13] Growing substrates are passed between large multi-domain NRPS modules by way of docking domains. It was proposed that the EpoA docking domain (EpoAdd) and the EpoB docking domain (EpoBdd) serve a recognition function, thus preventing cross communication between the wrong modules. More than one conformation of EpoBdd was observed using crystallographic studies. MD simulations of EpoBdd revealed high flexibility in the docking domain consistent with the crystallographic studies. Furthermore, MD was used to produce snapshots of open and closed states of EpoBCy, resulting in a comprehensive model that correlate active site access to product outcome. This in-depth study demonstrates the power of incorporating structural computational biology techniques towards understanding the sequence-structure-function relationships of mega-synthases.

In addition to the potential for interface engineering via docking domain manipulation, computational biology can also help expand starter and extender unit diversity in the mega-synthases. For example, starter unit selection was investigated using MD to develop a model of an acyltransferase ($AT5_{mon}$) in the monensin PKS by Bravo-Rodriguez *et al*.[14] The MD model was successfully used to engineer the active site in $AT5_{mon}$ for incorporation of non-native substrates through computationally guided predictions. A similar study was performed by Barajas *et al*, where MD served as the basis for their biochemical investigations and ultimately the design of an enzyme for use in the biosynthesis of potential biofuels.[15] A surge in creativity has also led to the design of novel cross-linking compounds that have enabled the determination of complex crystal structures involving many classes of the illusive carrier protein with varying partner enzymes (Table 2-S1). These complexes provide an excellent starting structure that, when coupled with MD simulations, can provide a basis for biochemical investigation and eventual engineering and design.

The MD simulations can provide insight into unresolved questions that are difficult to answer experimentally. For example, are the dynamics of the multiple carrier proteins in a megasynthase independent of each other, or are they coupled? Is there a functional advantage to covalently tethering the carrier protein to the megasynthase, or should the CP be tethered close to the location of the catalytic sites in each of the enzyme domains?[5] Is the motion of the carrier protein stochastic when it tries to enter several enzymes until it finds a “perfect fit” for the reaction to occur, or are the pathways more predetermined by the adaptor region between the carrier protein and its partner enzymes? These questions represent a small sampling of what may be explored through molecular modeling techniques.

A major obstacle when studying these multi-protein complex systems using computational structural biology techniques is the covalently-bound phosphopantetheine in the CPs. Current force fields support modeling standard amino acids, nucleic acids, sugars, and lipids.[16-17] A host of non-standard cases can also be investigated, thanks in part to the development of intrinsically disordered protein (IDP), non-canonical amino acid (NCAA), phosphorylated amino acids, and post-translational modified (PTM) force fields.[18-22] At present, no force field exists that is capable of modeling a PPant group bound to a protein. Performing MD simulations on this class of systems requires parameterization of the pantetheine moiety and its relatively large bound intermediates each time, thus reducing the computational accessibility to potentially critical information on protein-protein and protein-substrate interactions.[11, 15, 23] Parameterization of ligands is often straightforward and common practice in molecular modeling applications; however, non-standard residues, such as a phosphopantetheinyl-serine embedded in a protein, require extra parameterization and care. Furthermore, pantetheine is relatively flexible and moderately sized at 40 atoms, and Coenzyme A or phosphopantetheinyl-serine compounds are at minimum 80 or 52 atoms, respectively. A parameterization scheme utilizing modular splitting was employed, resulting in a fragmentation strategy that allowed for construction of a library of many compounds including biosynthetic starter units, extender units, and intermediate units (Fig 2-1). Here we report a pantetheine force field (PFF) built specifically to model and simulate pantetheine-bound biosynthetic intermediates either as a phosphopantetheinyl-serine side chain or as a standalone Coenzyme A.

MD and many other computational approaches rely on the availability and quality of force fields to build and design models, and the current lack of a publicly-available pantetheine force field restricts current and future research in the area of natural product biosynthesis as evidenced by the limited number of published MD simulations on these systems.[11, 15, 23-29] Here, we present a PFF library with 41 compounds, as well as a nomenclature scheme compatible with the Protein Data Bank (Table 2-1). This library will have a significant impact on modelers, engineers and experimentalists who wish to conduct MD simulations of any enzyme that requires PPant as a cofactor.

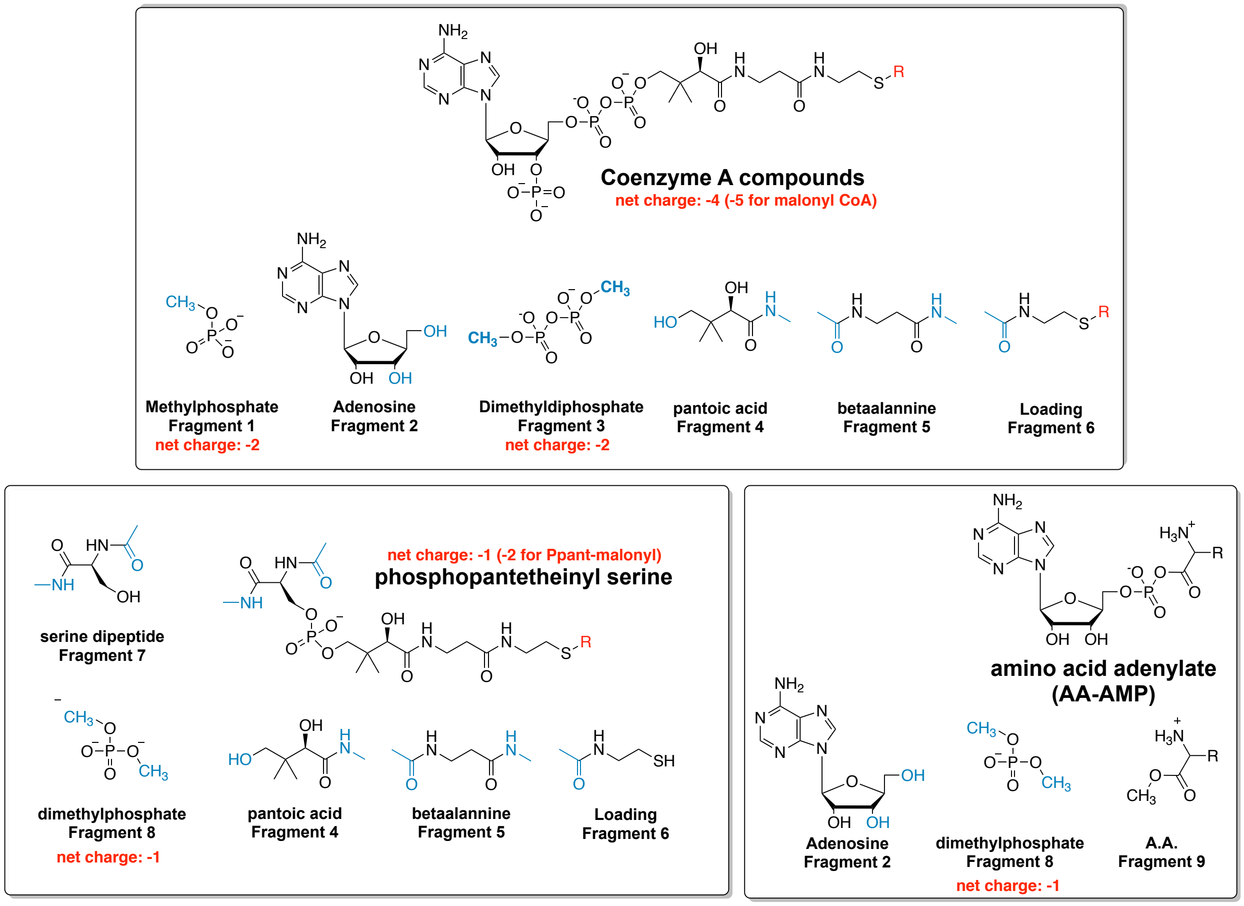


Figure 2-1. The parameterization scheme employed in the development of the PFF involved modular splitting. A) CoA compounds were constructed using six smaller fragments, with fragment **6** providing the structural diversity in the CoA library. CoA bound starter units, extender units and intermediates can be modeled with the above library. B) The phosphopantetheinyl-serine library was constructed using five smaller fragments, with the structural diversity generated by the fragment **5.** This library serves to model starter units, extender units, and intermediates bound to a phosphopantetheinyl-serine as a thioester incorporated in an ACP or PCP. C) The fragmentation strategy allows the possibility of a future expansion including amino acid adenylates, allowing the incorporation of proteinogenic or non-canonical amino acids as shown above. The library can be simply constructed from three fragments, with fragment **9** providing structural diversity.

Table 2-1. Residues present in the pantetheine force field (PFF)

Coenzyme A (CoA) Library			
PFF ID	Library Entry ID	PDB Ligand ID	Structures in PDB
C01	Coenzyme A	COA	459
C02	acetyl-CoA	ACO	177
C03	malonyl-CoA	MLC	11
C04	acetoacetyl-CoA	CAA	30
C05	propionyl-CoA	1VU	9
C06	butyryl-CoA	BCO	8
C07	hexanoyl-CoA	HXC	10
C08	octanoyl-CoA	CO8	9
C09	decanoyl-CoA	MFK	4
C10	dodecyl-CoA	DCC	6
C11	tetradecanoyl-CoA	MYA	54
C12	2-oxopentadecyl-CoA	NHW	19
C13	4-hydroxyphenacyl-CoA	4CO	10
C14	carboxymethyl-CoA	CMC	10
C15	3-hydroxy-3-methylglutaryl-CoA (HMG-CoA)	HMG	9
C16	crotonyl-CoA	COO	7
C17	oxidized-CoA	CAO	7
C18	methylmalonyl-CoA	MCA	5
C19	3-hydroxybuatonyl-CoA	3HC	5
C20	benzoyl-CoA	BYC	4
C21	stearoyl-CoA	ST9	4
C22	acetyltryptamine-CoA	COT	4
C23	persulfide-CoA	COS	4
C24	phenAcyl-CoA	0FQ	3
C25	4-hydroxybenzoyl-CoA	BCA	3
C26	isovaleryl-CoA	IVC	4
C27	*p* -coumaroyl-CoA	WCA	3
C28	*N* "-(2-coenzyme A)-popanoyl-lysine	LYX	3
C29	2-oxopropyl-CoA	SOP	3
C30	S-(2-oxo)pentadecyl-CoA	NHM	3

Phosphopantetheinyl (PPant) Serine Library			
PFF ID	Library Entry ID	PDB Ligand ID	Structures in PDB
S01	4'-phosphopantetheinyl serine	PNS	44
S02	acetyl-	6VG	1
S03	propionyl-	-	-
S04	butyryl-	PSR	2
S05	hexanoyl-	PM4, SXH	3
S06	octanoyl-	SXO	2
S07	decanoyl-	PM8	1

Phosphopantetheine Library			
PFF ID	Library Entry ID	PDB Ligand ID	Structures in PDB
X01	4'-phosphopantetheine	PNS	44
X02	S-acetyl-phosphopantetheine	6VG	1

Pantetheine Library			
PFF ID	Library Entry ID	PDB Ligand ID	Structures in PDB
P01	pantetheine	PNY	2
P02	S-acetyl-pantetheine	-	-

2.3 Results and Discussion

The biosynthesis of fatty acids, polyketides and nonribosomal peptides utilizes an assortment of building blocks, but at their core is the shared pantetheine moiety. A force field was generated of these pantetheine derivatives, including information on partial charges for each atom as well as their bond parameters. All of the members of the force field are compatible with the Amber ff14SB, GAFF and LIPID17 forcefields.[16-17, 30-31] The purpose of the pantetheine force field is to provide anyone who wishes to conduct MD simulations of these megasynthases with appropriate parameters, which in turns improves the understanding of how PKSs, NRPSs, and FASs biosynthesize complex bioactive molecules and how protein-substrate and protein-protein interactions can affect product generation. It is the authors' hope that this tool will help to provide much-needed insight to successfully engineer these systems for product exploitation. At the time of this writing, a search on PubMed revealed ~26,800 publications with the words "molecular dynamics" in conjunction with "enzyme" or "protein." A search for "molecular dynamics" with "Coenzyme A," "polyketide," or "pantetheine" revealed only 62, 18 and 3 publications respectively. This is surprising considering that 4% of enzymes utilize the cofactor Coenzyme A.[32] The lack of work for MD studies of PKS is directly linked to the lack of PPant force field library; therefore, outcomes from this work can have a significant impact in facilitating PKS researchers to conduct MD simulations of PKSs.

2.3.1 A new force field for the study of pantetheine-containing compounds

The electrostatic potential which serves as the basis in RESP charge fitting is dependent on the geometry of the compound. Large, flexible molecules easily produce unwanted intramolecular interactions, resulting in a bias in the charge fitting step. To reduce this bias, a fragmentation approach was employed. This library of compounds provides a consistent charging scheme built on a modular approach. This approach was deemed necessary due primarily to the flexibility and relatively large size of the pantetheine moiety itself. Indeed, it is common for primed Coenzyme A and phosphopantetheinyl-serine compounds to achieve sizes greater than 200 atoms.[33]

2.3.1.1 Compound names, residue names, and atom names nomenclature

When possible, atom names were assigned using the same atom and residue names as found in the protein data bank (PDB). The unloaded variants of Coenzyme A (Ligand ID: COA) and phosphopantetheinyl-serine (Ligand ID: PNS) were the most common with 459 and 44 entries present, respectively. There are 100 Coenzyme A compounds present in the PDB (Table S2-2), and the thirty most common compounds were included in our library (present in 887 of 979 or 90.6% of structures).

2.3.1.2 Restrained electrostatic potential charges derived for 41 pantetheine-containing compounds

Partial charges were calculated for every atom present in the force field using the restrained electrostatic potential (RESP) method utilized for the majority of force fields developed for AMBER.[34] RESP charges were calculated for each fragment with intra- and intermolecular charge constraints applied to ensure integer charges for each member in the library using the R.E.D. Development Server.[35] Partial charges for phosphopantetheinyl-serine were calculated with Ψ specified as the mean value present across structures in the

PDB at -40 degrees (Fig 2-2). During the charge fitting step, charges were compared between individual fragments and the final conjoined compounds. Partial charges, atom names, and atom types for Coenzyme A, Acetyl CoA and phosphopantetheinyl-serine are provided in tabular form (Table 2-2). Differences in charges were investigated for both primary compounds, and only four out of 80 atoms on Coenzyme A had a difference greater than 0.07. O3B on fragment 1, C3B on fragment 2, C9P on fragment 4, and N4P on fragment 6 had differences of 0.0878, 0.0973, 0.0952, and 0.0867, respectively. Four atoms having charges greater than 0.07 but less than 0.15 were considered acceptable based on other modular splitting approaches.[17] The splitting approach was primarily performed at peptide bonds with acetyl and N-methyl groups, or alternatively at methylene groups with methyl caps.

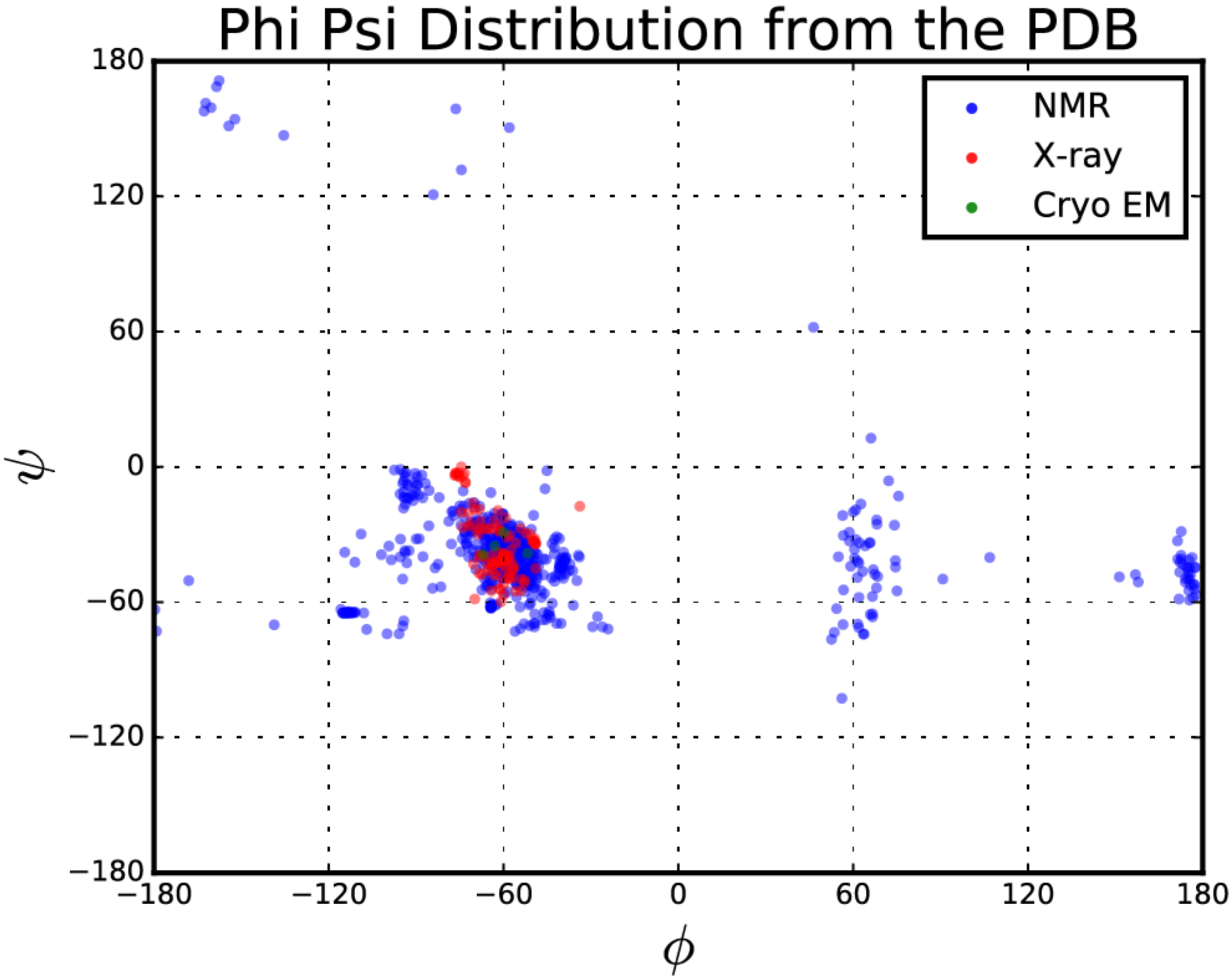


Figure 2-2. Φ/Ψ values of covalently bound pantetheine from the protein data bank (PDB) for a total of 644 data points (17 NMR solution structures, 50 X-ray crystal structures, and 2 cryo-EM structures). All pantetheine moieties were covalently bound to a serine residue. All cryo-EM and X-ray crystallographic structures present in the protein data bank (PDB) have Φ and Ψ angles characteristic of an alpha-helix. All covalently bound phosphopantetheine moieties present in the PDB are bound to a serine residue nested in an alpha-helix.

Table 2-2. Atom names, atom types, and partial charges of phosphopantetheine, Coenzyme A and acetyl CoA from the Protein Data Bank.

PDB Ligand ID: PNS			PDB Ligand ID: COA						PDB Ligand ID: ACO					
Atom Name	Atom Type	RESP Charge	Atom Name	Atom Type	RESP Charge	Atom Name	Atom Type	RESP Charge	Atom Name	Atom Type	RESP Charge	Atom Name	Atom Type	RESP Charge
C28	CI	-0.0978	C1B	CT	0.0391	O8A	O2	-0.913	C1	C	0.4691	O8A	O2	-0.9135
C29	CT	0.301	C2A	CQ	0.5694	O9A	O2	-0.913	C1B	CT	0.0385	O9A	O2	-0.9135
C30	CT	-0.2278	C2B	CT	0.1678	O9P	O	-0.5405	C2	CT	-0.116	O9P	O	-0.5411
C31	CT	-0.2278	C2P	CT	0.0282	OAP	OH	-0.6319	C2A	CQ	0.5688	OAP	OH	-0.6324
C32	CT	0.1051	C3B	CT	-0.016	P1A	P	1.1583	C2B	CT	0.1673	P1A	P	1.1578
C34	C	0.5484	C3P	CT	-0.0979	P2A	P	1.1548	C2P	CT	0.2306	P2A	P	1.1542
C37	CT	0.0864	C4A	CB	0.234	P3B	P	1.2248	C3B	CT	-0.0165	P3B	P	1.2243
C38	CT	-0.1466	C4B	CT	0.1681	S1P	SH	-0.3678	C3P	CT	-0.0361	S1P	SS	-0.3142
C39	C	0.5494	C5A	CB	0.1214	H1B	H2	0.1962	C4A	CB	0.2335	H1B	H2	0.1957
C42	CT	0.0954	C5B	CT	0.0588	H21	H1	0.082	C4B	CT	0.1675	H21	HC	0.0647
C43	CT	-0.0506	C5P	C	0.5358	H22	H1	0.082	C5A	CB	0.1209	H22	HC	0.0647
N36	N	-0.3856	C6A	CA	0.6739	H2A	H5	0.0497	C5B	CT	0.0583	H2A	H5	0.0491
N41	N	-0.513	C6P	CT	-0.063	H2B	H1	0.0962	C5P	C	0.5292	H2B	H1	0.0956
O23	OS	-0.309	C7P	CT	0.0103	H31	H1	0.1088	C6A	CA	0.6733	H2P2	H1	0.0114
O25	O2	-0.7796	C8A	CK	0.1404	H32	H1	0.1088	C6P	CT	-0.0603	H2P3	H1	0.0114
O26	O2	-0.7796	C9P	C	0.5056	H3B	H1	0.0955	C7P	CT	0.0103	H3B	H1	0.0949
O27	OS	-0.4904	CAP	CT	-0.02	H4B	H1	0.1748	C8A	CK	0.1398	H3P1	H1	0.0845
O33	OH	-0.6121	CBP	CT	0.2119	H4P	H	0.2728	C9P	C	0.5051	H3P2	H1	0.0845
O35	O	-0.557	CCP	CT	-0.0476	H5B2	H1	0.0613	CAP	CT	-0.0206	H4B	H1	0.1743
O40	O	-0.5125	CDP	CT	-0.0679	H5B3	H1	0.0613	CBP	CT	0.2114	H4P	H	0.2945
P24	P	1.1792	CEP	CT	-0.1908	H6A1	H	0.4135	CCP	CT	-0.0482	H5B2	H1	0.0608
S44	SH	-0.3519	N1A	NC	-0.7406	H6A2	H	0.4135	CDP	CT	-0.0684	H5B3	H1	0.0608
H281	H1	0.0932	N3A	NC	-0.6578	H6P2	HC	0.0508	CEP	CT	-0.1914	H6A1	H	0.4129
H282	H1	0.0932	N4P	N	-0.4577	H6P3	HC	0.0508	N1A	NC	-0.7412	H6A2	H	0.4129
H301	HC	0.0501	N6A	N2	-0.924	H7P2	H1	0.0648	N3A	NC	-0.6584	H6P2	HC	0.0496
H302	HC	0.0501	N7A	NB	-0.6035	H7P3	H1	0.0648	N4P	N	-0.4748	H6P3	HC	0.0496
H303	HC	0.0501	N8P	N	-0.4511	H8A	H5	0.1885	N6A	N2	-0.9245	H7P2	H1	0.064
H311	HC	0.0501	N9A	N*	0.0091	H8P	H	0.264	N7A	NB	-0.604	H7P3	H1	0.064
H312	HC	0.0501	O1A	O2	-0.8236	HAP	H1	0.1301	N8P	N	-0.4513	H8A	H5	0.1879
H313	HC	0.0501	O2A	O2	-0.7724	HCP2	H1	0.076	N9A	N*	0.0086	H8P	H	0.2635
H32	H1	0.0524	O2B	OH	-0.6376	HCP3	H1	0.076	O1	O	-0.4397	HAP	H1	0.1296
H33	HO	0.3885	O3A	OS	-0.4671	HDP1	HC	0.019	O1A	O2	-0.8242	HC21	HC	0.0647
H36	H	0.229	O3B	OS	-0.6292	HDP2	HC	0.019	O2A	O2	-0.7729	HCP2	H1	0.0754
H371	H1	0.0422	O4A	O2	-0.7729	HDP3	HC	0.019	O2B	OH	-0.6381	HCP3	H1	0.0754
H372	H1	0.0422	O4B	OS	-0.393	HEP1	HC	0.0521	O3A	OS	-0.4676	HDP1	HC	0.0185
H381	HC	0.0701	O5A	O2	-0.8224	HEP2	HC	0.0521	O3B	OS	-0.6297	HDP2	HC	0.0185
H382	HC	0.0701	O5B	OS	-0.5054	HEP3	HC	0.0521	O4A	O2	-0.7734	HDP3	HC	0.0185
H41	H	0.3043	O5P	O	-0.522	HO2B	HO	0.4333	O4B	OS	-0.3935	HEP1	HC	0.0515
H421	H1	0.0692	O6A	OS	-0.5025	HOAP	HO	0.4252	O5A	O2	-0.8229	HEP2	HC	0.0515
H422	H1	0.0692	O7A	O2	-0.913	HS1	HS	0.2007	O5B	OS	-0.506	HEP3	HC	0.0515
H431	H1	0.0826							O5P	O	-0.5214	HO2B	HO	0.4328
H432	H1	0.0826							O6A	OS	-0.503	HOAP	HO	0.4246
H44	HS	0.187							O7A	O2	-0.9135			

2.3.1.3 Force field parameters

Parameters were obtained from the AMBER ff14SB force field when available. Missing parameters were adopted from the General Amber Force Field (GAFF) to serve as a preliminary parameter set to newly defined atom types as was done in the initial LIPID 11 force field.[17] Forcefield_NCAA, another AMBER force field, utilized a similar approach with no additional parameterization, highlighting the promise of such an initial approach for force field development.[19] New atom types were assigned for these existing force constants and equilibrium values (Table 2-3) such that they can be easily revised in future investigations while maintaining compatibility with existing force fields.

Table 2-3. Bond, angle and dihedral parameters, and atom types.

Atom Type	mass	pol	Notes
SS	32.06	2.9	gaff2.dat (ss) thio-ester/thio-ether sp3 sulfur

Bond	K (kcal mol-1 ang -2)	Dist0 (ang)	Notes
C -SS	204.39	1.783	gaff2.dat (c -ss: 1.800) [MP2/aug-cc-PVTZ: 1.78293]
CT-SS	182.96	1.811	gaff2.dat (c3-ss: 1.839) [MP2/aug-cc-PVTZ: 1.81073]
HS-S	294.59	1.337	gaff2.dat (hs-sh: 1.347) [MP2/aug-cc-PVTZ: 1.33612]

Angle	K (kcal mol-1 rad -2)	Theta0 (deg)	Notes
C -CT-C	65.424	111.63	gaff2.dat (c -c3-c : 111.630)
C -CT-OH	85.627	108.79	gaff2.dat (c -c3-oh: 108.790)
C -SS-CT	71.828	104.271	gaff2.dat (c -ss-c3: 99.160) [MP2/aug-cc-PVTZ: 104.271]
CT-CT-SS	63.222	113.075	gaff2.dat (c3-c3-ss: 110.270) [MP2/aug-cc-PVTZ: 113.075]
CT-C -SS	63.438	118.576	gaff2.dat (c3-c -ss: 113.510) [MP2/aug-cc-PVTZ: 118.576]
CT-S -SH	51.361	97.417	gaff2.dat (c3-sh-hs: 96.400) [MP2/aug-cc-PVTZ: 97.417]
H1-CT-SS	42.463	105.425	gaff2.dat (h1-c3-ss: 108.760) [MP2/aug-cc-PVTZ: 105.425]
O -C -SS	79.009	117.999	gaff2.dat (o -c -ss: 123.320) [MP2/aug-cc-PVTZ: 117.999]

Dihedral	Path	V(kcal mol2-1 rad-1)	Phase	Period	Notes
C -SS-CT-CT	3	1	0	3	gaff2.dat (X -c3-ss-X: 3 1.0 0.0 3.0)
C -SS-CT-H1	3	1	0	3	gaff2.dat (X -c3-ss-X: 3 1.0 0.0 3.0)
CT-SS-C -CT	1	2.1	180	2	gaff2.dat (c3-ss-c -c3: 1 2.1 180.0 2)
X -SS-C -X	2	6.2	180	2	gaff2.dat (X -ss-c -X: 2 6.2 180.0 2)
CT-SS-C -O	2	6.2	180	2	gaff2.dat (X -ss-c -X: 2 6.2 180.0 2)

2.3.2 Validation

Existing parameters and new calculated partial charges were validated by performing normal mode analysis. Root mean square fluctuations (RMSF) of a simulation using Coenzyme A were compared to B-factors present in a crystal structure, and root mean square deviation (RMSD) values of the individual fragments of the molecular mechanics optimized structures were compared to optimized electronic structure calculations.

2.3.2.1 Comparison of optimized structures

The compound structures were optimized by Gaussian and compared to experimental values for missing parameters.[36] The use of Mg^{2+} and Na^{+} counterions were employed to find consistent agreement with experimental values as was shown by Schneider *et al* (Table S2-4).[37] The appropriate basis set to match MP2/aug-cc-pVTZ and experimental values using the B3LYP functional was found to be 6-311+G(2d,p). Additional diffuse and polarizable basis functions were needed to find good agreement with the O-P bond lengths and angles of the phosphate groups. The RMSD between QM and MM optimized fragments ranged from 0.072 and 0.606. The methylphosphate and dimethylphosphate fragments had RMSD values of 0.0726 and 0.0982, respectively, between the AMBER minimized fragments and the MP2/aug-cc-pVTZ optimized fragments. The relatively high RMSD of 0.606 Å was observed in adenosine due to the C-N-C-O torsion. This parameter was previously fit for the AMBER ff14SB force field; therefore, it was considered acceptable.

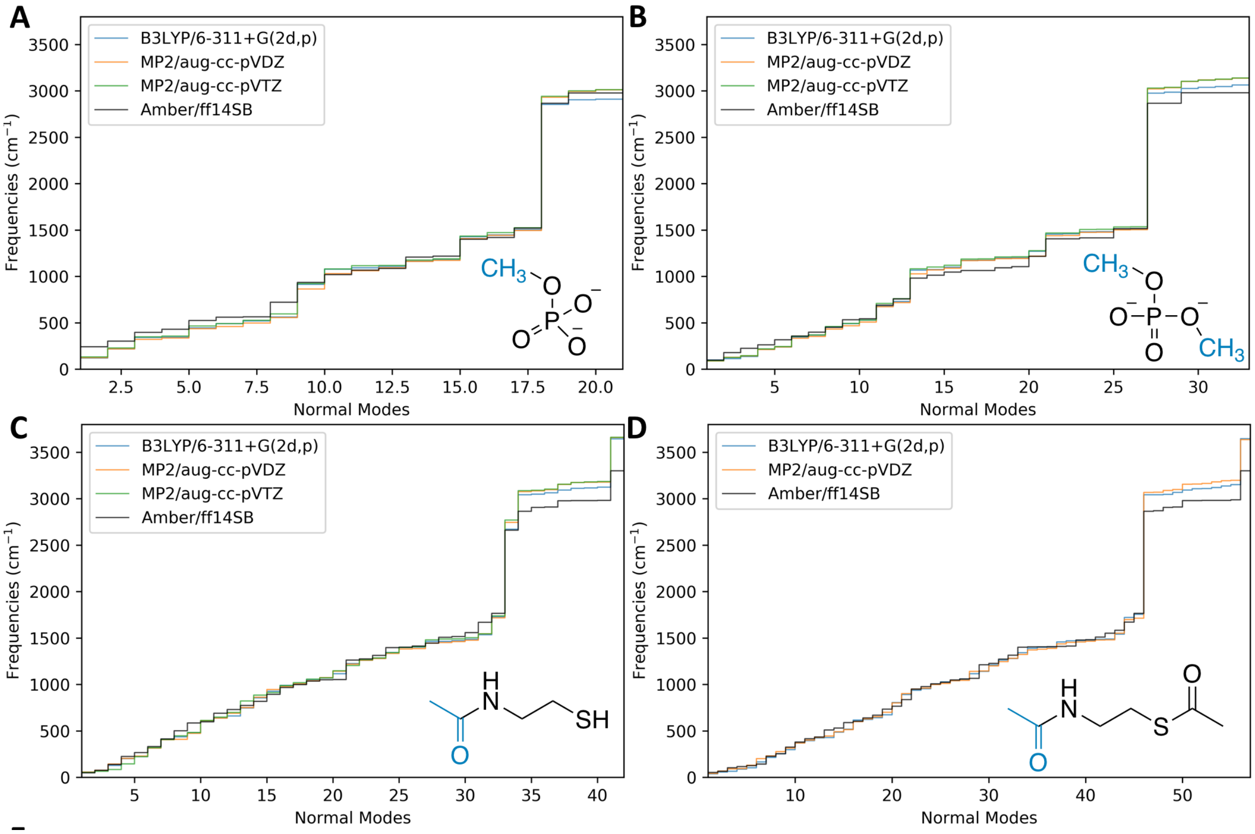


Figure 2-3. Normal mode frequency calculations were performed using B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theory. A low mode search was performed using the Amber ff14SB forcefield with additional phosphopantetheine force field (PFF) parameters.

2.3.2.2 Normal mode analysis

Normal mode frequencies were obtained in AMBER using a low mode search as well as from electronic structure calculations (Fig. 2-3). Calculations were performed at the B3LYP/6-311+G(2d,p) and MP2/aug-cc-pVDZ levels of theory for all the fragments. Fragments containing parameters which were adopted from GAFF, utilized additional electronic structure calculations performed at the MP2/aug-cc-PVTZ level of theory to match the level of theory used in the Forcefield_NCAA, ff03, and Amber FB15 force fields.[19, 38-40] Projections of normal modes agreed well between Amber and Gaussian. The frequencies observed in the 450-1100 cm^{-1} range include C-O and O-P bond-stretching, O-P-O twisting, O-P-O wagging, and O-P-O scissoring. The GAFF parameters were in good agreement and able to reproduce the normal modes for this range. To perform normal mode analysis on the thioester portion, an S-methyl thioacetate fragment was generated. S-C bond stretching was observed at 645 and 749 cm^{-1}, O-C-S scissoring observed at 439 cm^{-1}, and the characteristic intense carbonyl stretch for thioesters at 1720 cm^{-1} at the MP2/aug-cc-pVTZ level of theory.

2.3.2.3 Robustness of parameters over relatively long simulations

An aminoglycoside N3-acetyltransferase, BA2930, was selected as a system to study the robustness of adopted phosphate parameters from GAFF. A crystal structure containing Coenzyme A (PDB ID: 3IJW) was used as a starting structure. A 200 ns simulation was performed with the final 20 ns used to generate an average structure (Fig 2-4). The RMSD of the active site residues and Coenzyme A of the average structure was 0.821 Å, and all hydrogen bonds and electrostatic interactions observed in the crystal structure were observed in the average structure generated from MD. RMSFs for alpha carbons were measured over the final 20 ns of the simulation (Fig 2-5). Normalized RMSF and B-factor values were in relatively good agreement, with the similar fluctuations observed in MD compared to the X-ray crystallographic experiment.

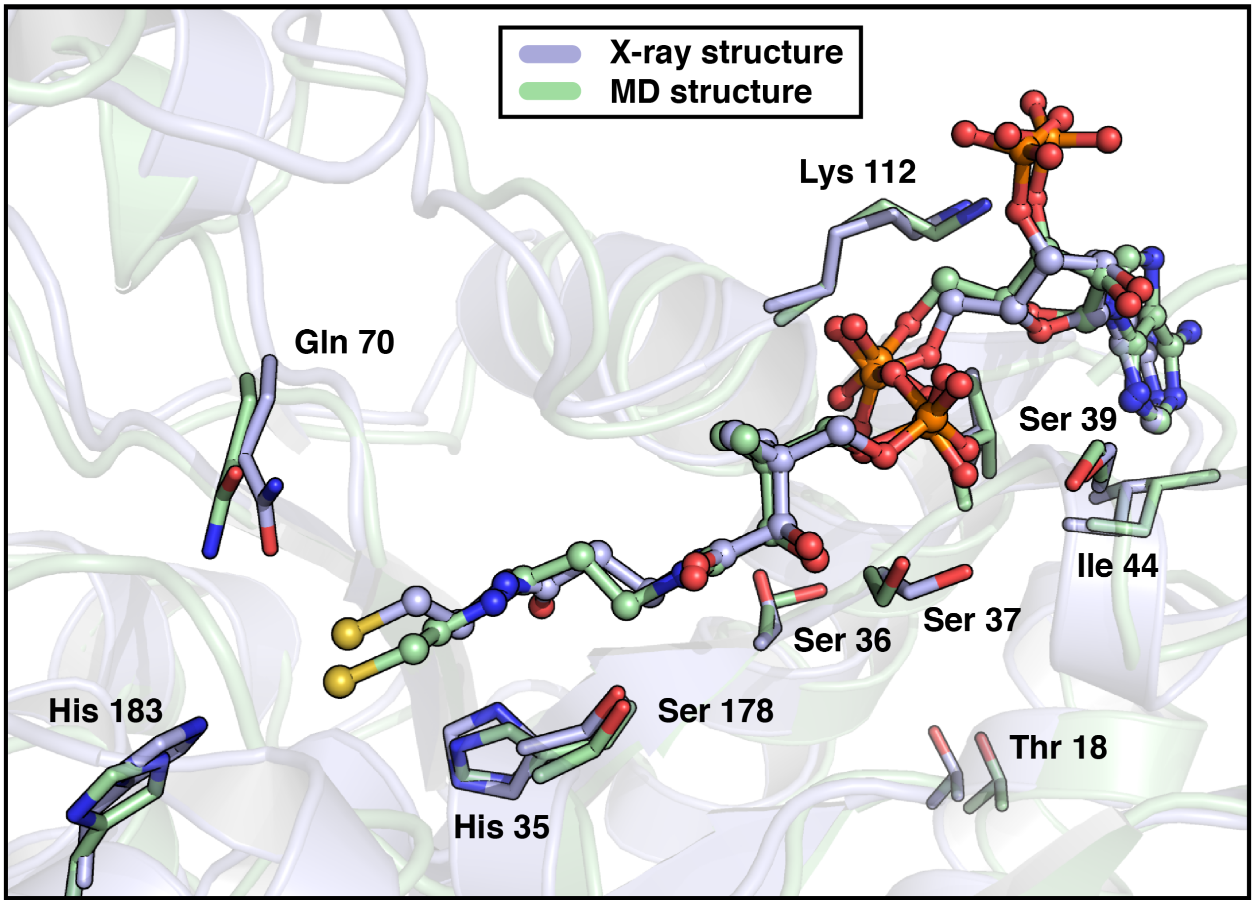


Figure 2-4. Comparison between X-ray structure (PDB ID: 3IJW) and a representative frame of the largest cluster of the final 20 ns of the 200 ns MD simulation performed on BA2930 using the pantetheine force field. The average RMSD for active site residues and Coenzyme A was 0.821, and the RMSD between the frame above compared to the X-ray structure was 0.778. Active site contacts of the MD structure are in consistent agreement with the X-ray structure.

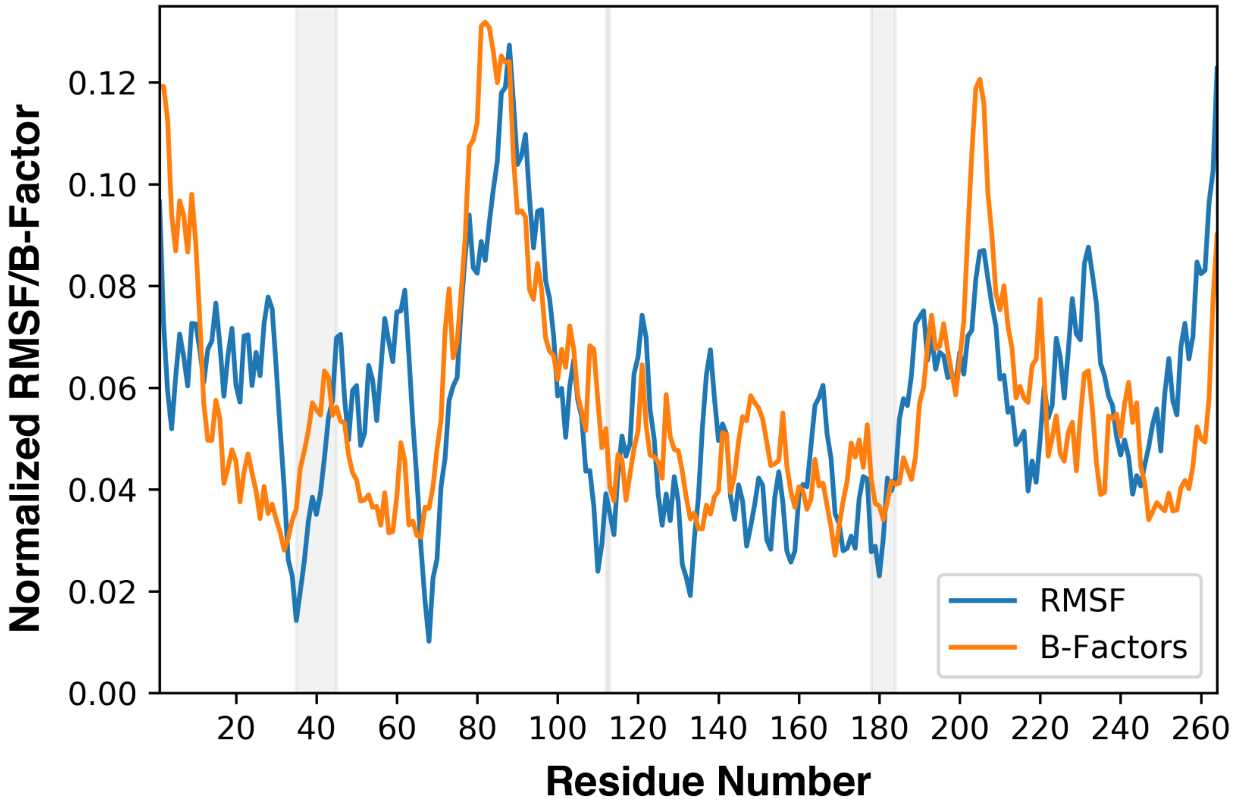
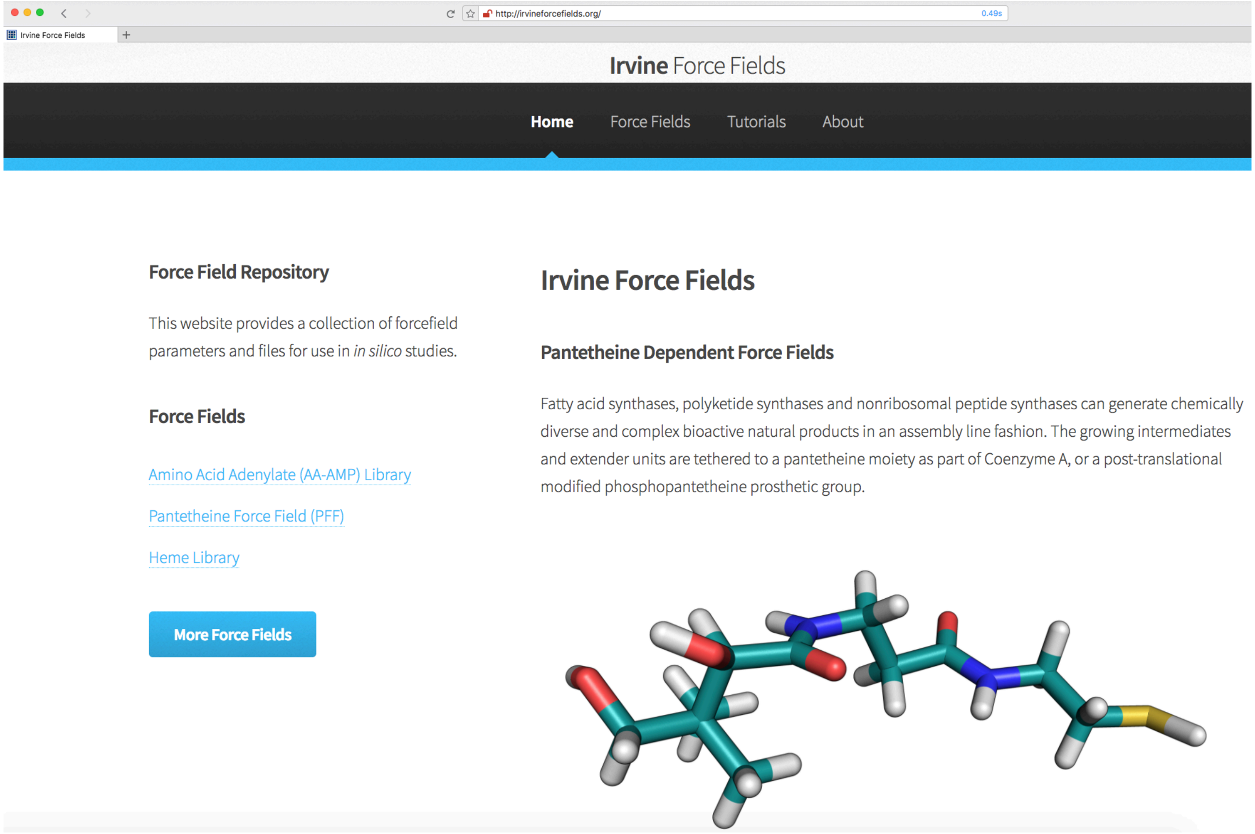


Figure 2-5. RMSFs for alpha carbons were measured over the final 20 ns of the simulation. Alpha carbon RMSFs and B-factors were normalized. Active site residues are shown in grey.

2.3.3 Phosphopantetheine Force Field (PFF) website interface

A website to host the pantetheine force field (http://www.irvineforcefields.org/) has been developed (Fig. 2-6). Contained on the website are four libraries of force fields for Coenzyme A, phosphopantetheinyl-serine, phosphopantetheine and pantetheine compounds. A parameter modification file with all necessary bond parameters, a library file with all structures and charges, and a LEaP configuration file that defines atom types for all compound are also present. Every individual compound is present on the force field page along with their corresponding Cartesian coordinates and RESP charges stored in the AMBER prep and Tripos Mol2 file formats.

An all-encompassing force field can be downloaded from the website free of charge and includes a parameter modification file with all necessary bond parameters, a library file containing all structures with their respective charges, and a LEaP configuration file that defines atom types for all the compounds. In conclusion, this work paved the foundation for future researchers wishing to conduct MD simulations of mega-synthases such as FAS, PKS and NRPS.



Force Field Repository

This website provides a collection of forcefield parameters and files for use in *in silico* studies.

Force Fields

Irvine Force Fields

Pantetheine Dependent Force Fields

Fatty acid synthases, polyketide synthases and nonribosomal peptide synthases can generate chemically diverse and complex bioactive natural products in an assembly line fashion. The growing intermediates and extender units are tethered to a pantetheine moiety as part of Coenzyme A, or a post-translational modified phosphopantetheine prosthetic group.

Figure 2-6. The pantetheine force field (PFF) can be found at the Irvine Force Fields website (http://www.irvineforcefields.org/). Contained on the website are Tripos mol2 files containing Cartesian coordinates and restrained electrostatic potential (RESP) charges for every compound. These structures include the same atom names as present in the Protein Data Bank for increased compatibility. These structures can be used for MD simulations, or as part of docking studies. A LEaP configuration file, parameter modification file and library file contained on the website provides all the necessary charges and parameters to model and simulate structures.

2.3.4 Methods

2.3.5 Parameterization strategy

A fragmentation approach was utilized which provided a degree of modularity to the force field while at the same time reducing the size, complexity and degrees of freedom during the charge fitting and parameterization steps. When possible, bond parameters were assigned using the parameter databases present in the Amber ff14SB, the primary protein and nucleic acid force fields in AMBER.[16-17] As these primary parameter databases are not all inclusive, missing parameters were preliminarily adopted from GAFF.[30] Gaussian 09 was used to optimize the geometries of the fragments using density functional theory with the B3LYP functional at the 6-311+G(2d,p) level of theory.[36] Parameters obtained from geometry optimizations were compared to values present in the existing AMBER force fields, as well as those present in existing experimentally derived structures.

2.3.5.1 Procedure for derivation of partial charges

Fragments were capped with acetyl, *N*-methyl, methyl, and/or hydroxyl caps which mimicked the natural neighboring chemical environment of the fragments. The R.E.D. Development Server was used for charge fitting of the fragments.[35] A two-step restricted electrostatic potential (RESP) fitting strategy used in other AMBER force fields was employed to derive partial charges. The electrostatic potential was calculated using Gaussian 09 with the HF/6-31G(d,p) level of theory. Intramolecular charge constraints were applied on the non-standard residues, forcing an integer charge on the central intramolecular fragment. Fragments were stitched together through the use of intermolecular charge constraints, which were applied between corresponding caps on connecting fragments.

2.3.5.2 Procedure for derivation of missing bond, angle, and dihedral angle parameters

Existing bond parameters in the Amber ff14SB force fields were used where possible to ensure compatibility with existing force fields. Missing parameters were adopted from GAFF and compared to experimental measurements and electronic structure calculations. Electronic structure calculations were performed using MP2/aug-cc-pVTZ level of theory, with Na+ counterions present to achieve measurements in agreement with experimental values.[37]

2.3.5.3 Pantetheine parameters

Atom names of the pantetheine moiety were matched to PDB ligand ID PNY.[41] The pantetheine unit was fragmented into three modules during parameterization: cysteamine linker (**6**), providing the thiol responsible for loading, ß-alanine (**5**), and pantoic acid (**4**).[42] Fragment **4** is capable of existing by itself, or connected to a mono- or di-phosphate species. The cysteamine linker was modeled as the free thiol, or the S-acetyl cysteamine species to mimic the acetyl unit. Di-fragments capable of representing several chemical environments were capped and optimized. Intramolecular and intermolecular charge restraints were applied, and RESP charges were fit using the method described previously.

2.3.5.4 Phosphopantetheine parameters

Atom names of the phosphopantetheine moiety were matched to PDB ligand ID PNS. Fragment **4** was modeled with a connecting mono-phosphate group (**8**). Fragment **8** was capped with a methyl cap to mimic a covalent linkage to a serine moiety. Intra- and inter-molecular charge restraints were applied, and RESP charges fit using the method described previously.

2.3.5.5 Coenzyme A parameters

A chemical component search was performed using PDB ligand ID COA as the initial substructure. A table containing 100 coenzyme A compounds was generated from the PDB, covering a total of 979 structures (Table S2-2). A coenzyme library was generated for the 30 CoA compounds with three or more structures present in the PDB, representing 91% of all PDB structures containing a CoA compound. CoA entries present in the pantetheine force field (PFF) have atom names and residue names matched to those present in the PDB. Three fragments representing a methyl phosphate (**1**), adenosine (**2**), and dimethyl diphosphate (**3**) were joined to the ß-alanine (**5**), and pantoic acid (**4**) fragments of pantetheine. An acetyl cap was added to fragment (**5**). Thirty cysteamine linker fragments were generated and capped with N-methyl caps. Intramolecular and intermolecular charge restraints were applied, and RESP charges fit using the method described previously.

2.3.5.6 Phosphopantetheinyl-serine parameters

Care was taken during the optimization of the phosphopantetheinyl serine residue, with the Ψ dihedral being fixed to -40.0 degrees, while imposing no restriction on the Φ dihedral to match experimental data collected from the Protein Data Bank.

2.3.6 Validation of parameters

The pantetheine force field was validated by comparison of normal-mode frequencies, comparison of QM and MM structures, and comparison of root mean square fluctuations in MD samples to B-factor values from crystallographic experiments.

2.3.6.1 Normal mode analysis

The quality of utilizing existing force field parameters from ff14SB and GAFF was measured through the distribution of normal modes. Experimentally derived parameters for phosphate geometry in the presence of various metal cations was used to benchmark and compare electronic structure calculations.[37] Second order Møller–Plesset perturbation theory (MP2) with an augmented version of the correlation-consistent triple-zeta (aug-cc-pVTZ) basis set was used. It was found that the B3LYP functional with the 6-311+G(2d,p) basis set was in close agreement to MP2/aug-cc-pVTZ. Normal modes were obtained through electronic structure calculations at B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ level of theory. Optimized fragments from electronic structure calculations were used as starting structures in Amber, and were minimized using the pantetheine force field. A low mode search was performed using Kolossváry's algorithm implemented in AMBER in order to obtain the normal mode frequencies.[43] The RMSD between QM and MM optimized fragments ranged from 0.072 and 0.606.

2.3.6.2 Molecular dynamics simulations on BA2930

MD was carried out using AMBER 16.[44] The structure of BA2930, an aminoglycoside N-acetyltransferase, bound to Coenzyme A was acquired from the PDB (PDB ID: 3IJW), and prepared for MD using the program UCSF Chimera.[45-46] Charges and parameters for the amino acids were used from the AMBER ff14SB force field, and charges and parameters for Coenzyme A were obtained from the force field described in this research report. LEaP was used to add hydrogens, neutralize the system through the addition of 10 Na^+ ions, and solvation of the system in a 10-Å water buffer TIP3P truncated octahedron box. Using the program SANDER, the system was subjected to a two-step minimization process to remove any steric clashes present in the initial crystal structure. The first step of minimization was carried out over 5,000 steps for the solvent and ions, with the protein and Coenzyme A restrained by a force constant of 500 kcal/mol/Å^2, followed by a second stage of minimization with no restraints. Using a Langevin temperature equilibration scheme, the system was heated to 298 K using the NVT ensemble over 1 ns with weak 10 kcal/mol/Å^2 restraints on the protein and Coenzyme A. Then, the system was equilibrated using the NPT ensemble for 3 ns. Periodic boundary conditions were used, and hydrogens were restrained using the SHAKE algorithm. The simulation was run over 200 ns. The final 20 ns was used to generate an average structure, which was subsequently minimized. The average structure had an RMSD of 0.821 compared to the crystal structure.

2.3.6.3 B-factor comparisons

Using the program CPPTRAJ, root mean square fluctuations (RMSFs) were measured on the final 20 ns of the simulation.[47] RMSF values from the MD simulation, and B-factors from the crystal structure were normalized and compared. Normalized RMSF and B-factor values for active site residues in contact with Coenzyme A were in close agreement.

2.4 Conclusions and future directions

This chapter presented a new force field to describe 41 compounds, either as CoA analogs, or covalently attached to a serine residue on a carrier protein. This force field is compatible with the PDB and adopted a similar naming scheme. This force field provides an initial framework for performing MD simulations, it has left open-ended the ability to optimize existing parameters, as well as add additional compounds dependent on user demands.

2.5 Supplementary figures and tables

Table S2-1. List of all PDBs with a phosphopantetheine covalently linked to a carrier protein

PDB ID	Ligand IDs	Modification	Native Organism	Complex	Notes
1F80	PN2	holo-ACP	*Bacillus subtilis*	Yes	*in complex with holo-ACP synthase (AcpS)*
1L0I	PSR	butyryl-ACP	*Escherichia coli*	No	*I62M mutant*
2FAC	PM4	hexanoyl-ACP	*Escherichia coli*	No	
2FAD	PM5	heptanoyl-ACP	*Escherichia coli*	No	
2FAE	PM8	decanoyl-ACP	*Escherichia coli*	No	
2X2B	SXM	malonyl-ACP	*Bacillus subtilis*	No	
3EJB	ZMP	tetradecanoyl-ACP	*Escherichia coli*	Yes	*in complex with Cytochrome P450BioI (CYP107H1) from B. subtilis*
3EJD	ZMQ	hexadec-9Z-enoyl-ACP	*Escherichia coli*	Yes	*in complex with Cytochrome P450BioI (CYP107H1) from B. subtilis*
3EJE	ZMO	octadec-9Z-enoyl-ACP	*Escherichia coli*	Yes	*in complex with Cytochrome P450BioI (CYP107H1) from B. subtilis*
3GZL	PNS	disulfide linked-ACP	*Plasmodium falciparum*	No	*disulfide linked-pfACP dimer*
3GZM	PNS	holo-ACP	*Plasmodium falciparum*	No	
3NY7	SXM	malonyl-ACP	*Escherichia coli*	Yes	*in complex with SLC26 anion transporter STAS domain from YchM*
3RG2	PNS	holo-PCP	*Escherichia coli*	Yes	*NRPS EntE adenylation domain and EntB PCP didomain complex*
4BPH	PNS	holo-DCP	*Bacillus subtilis*	No	*D-alanyl carrier protein (Dcp) DltC*
4DG9	DG9	valine-AVS inhibitor-PCP	*Pseudomonas aeruginosa*	Yes	*valine-adenosine vinylsulfonamide (Val-AVS) inhibitor*
4ETW	ZMK	methyl-pimeloyl-ACP	*Shigella flexneri*	Yes	*in complex with BioH S82A*
4H2S	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase*
4H2T	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase*
4H2U	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase*
4H2V	H2V PNS	glycl-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase*
4H2W	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase from Agrobacterium fabrum*
4H2X	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase from Agrobacterium fabrum*
4H2Y	PNS	holo-aa:CP	*Bradyrhizobium diazoefficiens*	Yes	*in complex with an amino acid carrier protein (aa:CP) ligase from Agrobacterium fabrum*
4IHF	1F7	(R)-3-hydroxytetradecanoyl-ACP	*Escherichia coli*	Yes	*in complex with LpxD*
4IHG	PNS	holo-ACP	*Escherichia coli*	Yes	*in complex with LpxD*
4IHH	PNS	holo-ACP	*Escherichia coli*	Yes	*in complex with LpxD*
4IZ6	PNS	DHB-AVS inhibitor-PCP	*Escherichia coli*	Yes	*NRPS EntE adenylation domain and EntB PCP didomain complex via 2,3-dihydroxybenzoate-AVS inhibitor*
4KEH	1R3	sulphonyl-3-alkyne-based probe-ACP	*Escherichia coli*	Yes	*in complex with FabA (Type II FAS Dehydratase)*
4PWV	KH4	4-imidazole carboxyl-PCP	*Streptomyces sp. Acta 2897*	Yes	*in complex with Cytochrome P450sky (CYP163B3) from Streptomyces sp. Acta 2897*
4PXH	KH4	4-imidazole carboxyl-PCP	*Streptomyces sp. Acta 2897*	Yes	*in complex with Cytochrome P450sky (CYP163B3) from Streptomyces sp. Acta 2897*
4ZJB	PNS	holo-ACP	*Helicobacter pylori*	Yes	*in complex with FabZ (Type II FAS Dehydratase)*
4ZXH	PNS	holo-PCP	*Acinetobacter baumannii*	Yes	*complete holo-AB3403 NRPS module (adenylation, condensation, PCP and thioesterase)*
4ZXI	PNS	holo-PCP	*Acinetobacter baumannii*	Yes	*complete holo-AB3403 NRPS module (adenylation, condensation, PCP and thioesterase)*
5CZD	PNS	malemide-based inhibitor-ACP	*Streptomyces halstedii*	Yes	*in complex with acyltransferase (AT) VinK*
5EJD	5PD	holo-PCP	*Penicillium aethiopicum*	Yes	*NRPS TqaA condensation domain and PCP didomain complex*
5ES8	5S4	valine-NH-PCP	*Brevibacillus parabrevis*	Yes	*NRPS LgrA initiation NRPS module (formylation, adenylation, and PCP)*
5ES9	PNS	holo-PCP	*Brevibacillus parabrevis*	Yes	*NRPS LgrA initiation NRPS module (formylation, adenylation, and PCP)*
5H9H	4HH	holo-ACP	*Helicobacter pylori*	No	
5ISX	PNS	holo-PCP	*Brevibacillus brevis*	Yes	*NRPS GrsA epimirization domain and PCP didomain complex*
5JA1	75C	serine-AVS inhibitor-PCP	*Escherichia coli*	Yes	*NRPS EntF NRPS module bound to the MbtH-like protein (MLP) from E. coli*
5JA2	75C	serine-AVS inhibitor-PCP	*Escherichia coli*	Yes	*NRPS EntF NRPS module bound to the MbtH-like protein (MLP) from P. aeruginosa*
5KP7	PNS	holo-ACP	*Lyngbya majuscula*	Yes	*in complex with 3-hydroxy-3-methylglutaryl synthas (HMGS) CurD from Moorea producens*
5KP8	PNS 6VG	acetyl-ACP	*Lyngbya majuscula*	Yes	*in complex with 3-hydroxy-3-methylglutaryl synthas (HMGS) CurD from Moorea producens*
5T3D	75C	serine-AVS inhibitor-PCP	*Acinetobacter baumannii*	Yes	*complete holo-AB3403 NRPS module (adenylation, condensation, PCP and thioesterase)*
5U89	MJ8	glycine-AVS inhibitor-PCP	*Geobacillus sp. Y4.1MC1*	Yes	*NRPS DhbF cross-modular tri-domain (MLP, adenylation, PCP and condenstation)*
			Total Complex Papers Published*	**20**	

* This does not include apo-carrier protein complex structures (PDB IDs: 2FHS, 2XZ0, 4DXE).

Table S2-2. List of all Coenzyme A ligands present in the PDB

Ligand Name	PDB Ligand ID	Structures in PDB	M.W.
4-Chlorophenacyl-coenzyme A	01A	1	920.11
[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3R)-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-4-[[3-oxidanylidene-3-[2-[(2R)-2-oxidanylundecyl]sulfanylethylamino]propyl]amino]butyl] hydrogen phosphate	0ET	1	937.83
phenacyl coenzyme A	0FQ	3	885.67
[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3R)-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-4-[[3-oxidanylidene-3-(4-oxidanylidenepentylamino)propyl]amino]butyl] hydrogen phosphate	0RQ	1	791.53
[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3R)-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-4-[[3-oxidanylidene-3-(propylamino)propyl]amino]butyl] hydrogen phosphate	0T1	3	749.50
(2S)-2-({(3S,5R,9R)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9-trihydroxy-8,8-dimethyl-3,5-dioxido-10,14,20-trioxo-2,4,6-trioxa-18-thia-11,15-diaza-3lambda~5~,5lambda~5~-diphosphaicosan-20-yl}amino)pentanedioic acid (non-preferred name)	1C4	1	954.68
(3S,5S,9R,21S)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9,21-tetrahydroxy-8,8,21-trimethyl-10,14-dioxo-19-thioxo-2,4,6-trioxa-18-thia-11,15-diaza-3,5-diphosphatricosan-23-oic acid 3,5-dioxide	1CZ	1	927.72
3-methylmercaptopropionate-CoA (MMPA-CoA)	1HE	1	869.69
propionyl Coenzyme A	1VU	9	823.60
2-CARBOXYPROPYL-COENZYME A	2CP	2	853.62
METHACRYLYL-COENZYME A	2MC	1	835.61
Salicylyl CoA	2NE	2	887.64
(3R,5S,9R)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9-trihydroxy-8,8-dimethyl-10,14-dioxo-2,4,6-trioxa-11,15-diaza-3,5-diphosphaheptadecane-17-sulfinic acid 3,5-dioxide (non-preferred name)	30N	1	799.53
3-CARBOXYPROPYL-COENZYME A	3CP	1	853.62
(S)-3-Hydroxyhexanoyl-CoA	3H9	2	877.65
3-HYDROXYBUTANOYL-COENZYME A	3HC	5	853.62
3-oxo-4-pregnene-20-carboxyl-Coenzyme A	4BN	1	1094.01
4-HYDROXYBENZYL COENZYME A	4CA	2	873.66
4-HYDROXYPHENACYL COENZYME A	4CO	10	901.67
pivalyl-coenzyme A	52O	1	851.65
(25S)-3-oxocholest-4-en-26-oyl-CoA	5JB	1	1164.14
[[(2~{S},3~{S},4~{R},5~{R})-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3~{R})-4-[[3-[2-[2-[3-[[(2~{R})-4-[[[(2~{R},3~{S},4~{R},5~{R})-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethyldisulfanyl]ethylamino]-3-oxidanylidene-propyl]amino]-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-butyl] hydrogen phosphate	5NG	1	1533.05
~{S}-[2-[3-[[(2~{R})-4-[[[(2~{R},3~{S},4~{R},5~{R})-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethyl] 5-[(2~{R},3~{R},5~{R},6~{S})-6-methyl-3,5-bis(oxidanyl)oxan-2-yl]oxypentanethioate	6QA	1	997.79
(2~{S})-2-[2-[3-[[(2~{R})-4-[[[(2~{R},3~{S},4~{R},5~{R})-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethylsulfanyl]propanoic acid	8HB	2	839.60
(3R,5S,9R,26S)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9-trihydroxy-8,8-dimethyl-10,14,20-trioxo-26-({[(phenylacetyl)amino]acetyl}amino)-2,4,6-trioxa-18-thia-11,15,21-triaza-3,5-diphosphaheptacosan-27-oic acid 3,5-dioxide (non-preferred name)	93M	1	1128.93
(3R,5S,9R,23S)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9-trihydroxy-8,8-dimethyl-10,14-dioxo-23-({[(phenylacetyl)amino]acetyl}amino)-2,4,6-trioxa-18-thia-11,15-diaza-3,5-diphosphatetracosan-24-oic acid 3,5-dioxide (non-preferred name)	93P	1	1071.88
isopentyl-Coenzyme A	A1S	1	837.67
[(2R,3R,4R,5R)-5-(6-amino-9H-purin-9-yl)-3,4-bis(phosphonooxy)tetrahydrofuran-2-yl]methyl (3R)-3-hydroxy-2,2-dimethyl-4-oxo-4-({3-oxo-3-[(2-sulfanylethyl)amino]propyl}amino)butyl dihydrogen diphosphate	AC8	1	847.51
ACETYL COENZYME *A	ACO	177	809.57
AMIDOCARBOXYMETHYLDETHIA COENZYME *A	AMX	1	792.52
4-HYDROXYBENZOYL COENZYME A	BCA	3	887.64
Butyryl Coenzyme A	BCO	8	837.62
benzoyl coenzyme A	BYC	4	871.64
COA-S-TRIMETHYLENE-ACETYL-TRYPTAMINE	CA3	1	1009.85
COA-S-ACETYL 5-BROMOTRYPTAMINE	CA5	1	1046.67
[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3R)-2,2-dimethyl-4-[[3-(4-methylsulfonylbutylamino)-3-oxidanylidene-propyl]amino]-3-oxidanyl-4-oxidanylidene-butyl] hydrogen phosphate	CA6	1	841.61
[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3R)-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-4-[[3-oxidanylidene-3-[4-(phenylsulfonyl)butylamino]propyl]amino]butyl] hydrogen phosphate	CA8	1	903.68
ACETOACETYL-COENZYME A	CAA	30	851.61
ethyl 5-[3-[[(2R)-4-[[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]pentanoate	CAJ	1	835.59
OXIDIZED COENZYME A	CAO	7	783.53
CITRYL-THIOETHER-COENZYME *A	CIC	1	927.66
CARBOXYMETHYL COENZYME *A	CMC	10	825.57
CARBOXYMETHYLDETHIA COENZYME *A	CMX	2	793.51
ISOBUTYRYL-COENZYME A	CO6	2	837.62
S-{(9R,13R,15S)-17-[(2R,3R,4R,5R)-5-(6-amino-9H-purin-9-yl)-3-hydroxy-4-(phosphonooxy)tetrahydrofuran-2-yl]-9,13,15-trihydroxy-10,10-dimethyl-13,15-dioxido-4,8-dioxo-12,14,16-trioxa-3,7-diaza-13,15-diphosphaheptadec-1-yl}(2E)-but-2-enethioate	CO7	1	835.61
OCTANOYL-COENZYME A	CO8	9	893.73
COENZYME A	COA	459	767.53
DEPHOSPHO COENZYME A	COD	8	687.55
TRIFLUOROACETONYL COENZYME A	COF	1	877.57
[(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]METHYL (3R)-3-HYDROXY-4-{[3-({2-[(2-HYDROXYETHYL)DITHIO]ETHYL}AMINO)-3-OXOPROPYL]AMINO}-2,2-DIMETHYL-4-OXOBUTYL DIHYDROGEN DIPHOSPHATE	COK	4	843.65
COENZYME A PERSULFIDE	COS	4	799.60
COA-S-ACETYL TRYPTAMINE	COT	4	967.77
Anthraniloyl-coenzyme A	COW	1	886.66
3-THIAOCTANOYL-COENZYME A	CS8	1	911.77
4-(N,N-DIMETHYLAMINO)CINNAMOYL-COA	DAK	1	940.74
DESULFO-COENZYME A	DCA	4	735.47
DODECYL-COA	DCC	6	949.84
ALPHA-FLUORO-AMIDOCARBOXYMETHYLDETHIA COENZYME A COMPLEX	FAM	1	810.51
Phenylacetyl coenzyme A	FAQ	2	885.67
ALPHA-FLUORO-CARBOXYMETHYLDETHIA COENZYME A COMPLEX	FCX	1	811.50
S-{(9R,13S,15R)-17-[(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]-9,13,15-TRIHYDROXY-10,10-DIMETHYL-13,15-DIOXIDO-4,8-DIOXO-12,14,16-TRIOXA-3,7-DIAZA-13,15-DIPHOSPHAHEPTADEC-1-YL} THIOFORMATE	FYN	1	795.54
glutaryl-coenzyme A	GRA	2	881.63
N-HYDROXYAMIDOCARBOXYMETHYLDETHIA COENZYME *A	HAX	1	808.52
3R-HYDROXYDECANOYL-COENZYME A	HDC	1	937.78
2,4-dihydroxyphenacyl coenzyme A	HFQ	2	917.67
(3R,5S,9R,21S)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9,21-tetrahydroxy-8,8-dimethyl-10,14,19-trioxo-2,4,6-trioxa-18-thia-11,15-diaza-3,5-diphosphatricosan-23-oic acid 3,5-dioxide	HGG	1	897.63
3-HYDROXY-3-METHYLGLUTARYL-COENZYME A	HMG	9	906.62
(S)-3-HYDROXYDECANOYL-COA	HSC	3	933.75
HEXANOYL-COENZYME A	HXC	10	865.68
S-[2-[3-[[(2R)-4-[[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-hydroxy-3-phosphonooxy-oxolan-2-yl]methoxy-hydroxy-phosphoryl]oxy-hydroxy-phosphoryl]oxy-2-hydroxy-3,3-dimethyl-butanoyl]amino]propanoylamino]ethyl] (2R)-2-hydroxy-4-methyl-pentanethioate	IRC	1	881.68
Isovaleryl-coenzyme A	IVC	4	851.65
GDP-N-acetylperosamine-coenzyme A	JBT	1	1381.93
N"-(2-COENZYME A)-PROPANOYL-LYSINE	LYX	3	967.77
METHYLMALONYL-COENZYME A	MCA	5	867.61
METHYLMALONYL(CARBADETHIA)-COENZYME A	MCD	1	849.57
decanoyl-CoA	MFK	4	921.78
MALONYL-COENZYME A	MLC	11	853.58
(R)-2-METHYLMYRISTOYL-COENZYME A	MRR	1	991.92
(S)-2-METHYLMYRISTOYL-COENZYME A	MRS	1	991.92
TETRADECANOYL-COA	MYA	54	977.89
S-4-NITROBUTYRYL-COA	NBC	1	882.62
(3R)-27-AMINO-3-HYDROXY-2,2-DIMETHYL-4,8,14-TRIOXO-12-THIA-5,9,15,19,24-PENTAAZAHEPTACOS-1-YL [(2S,3R,4S,5S)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]METHYL DIHYDROGEN DIPHOSPHATE	NHQ	1	1009.90
2-oxopentadecyl-CoA	NHW	19	991.92
NITROMETHYLDETHIA COENZYME A	NMX	1	794.49
OXALYL-COENZYME A	OXK	1	839.55
3-[(4-AMINO-2-METHYLPYRIMIDIN-5-YL)METHYL]-2-{(1R,11R,15S,17R)-19-[(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]-1,11,15,17-TETRAHYDROXY-12,12-DIMETHYL-15,17-DIOXIDO-6,10-DIOXO-14,16,18-TRIOXA-2-THIA-5,9-DIAZA-15,17-DIPHOSPHANONADEC-1-YL}-5-(2-{[(R)-HYDROXY(PHOSPHONOOXY)PHOSPHORYL]OXY}ETHYL)-4-METHYL-1,3-THIAZOL-3-IUM	OXT	1	1220.86
Palmitoyl-CoA	PKZ	2	1005.94
(R)-IBUPROFENOYL-COENZYME A	RFC	1	956.81
SUCCINYL(CARBADETHIA)-COENZYME A	SCD	1	849.57
(S)-IBUPROFENOYL-COENZYME A	SFC	1	956.81
[(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]METHYL (3R)-3-HYDROXY-2,2-DIMETHYL-4-OXO-4-{[3-OXO-3-({2-[(2-OXOPROPYL)THIO]ETHYL}AMINO)PROPYL]AMINO}BUTYL DIHYDROGEN DIPHOSPHATE	SOP	3	823.60
STEAROYL-COENZYME A	ST9	4	1034.00
(2E)-Hexenoyl-CoA	TC6	1	859.63
3-thiaglutaryl-CoA	TGC	1	899.67
S-{(3S,5R,9R)-1-[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-4-hydroxy-3-(phosphonooxy)tetrahydrofuran-2-yl]-3,5,9-trihydroxy-8,8-dimethyl-3,5-dioxido-10,14-dioxo-2,4,6-trioxa-11,15-diaza-3lambda~5~,5lambda~5~-diphosphaheptadecan-17-yl} undecanethioate	UCC	1	935.81
p-coumaroyl-CoA	WCA	3	913.68
[(2R,3S,4R,5R)-5-(6-AMINO-9H-PURIN-9-YL)-4-HYDROXY-3-(PHOSPHONOOXY)TETRAHYDROFURAN-2-YL]METHYL (3R)-4-({3-[(2-{[(3,5-DIHYDROXYPHENYL)ACETYL]AMINO}ETHYL)AMINO]-3-OXOPROPYL}AMINO)-3-HYDROXY-2,2-DIMETHYL-4-OXOBUTYL DIHYDROGEN DIPHOSPHATE	YE1	5	900.61
[[(2~{R},3~{S},4~{R},5~{R})-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl] [(3~{R})-4-[[3-[2-[(~{E})-2-[3,5-bis(oxidanyl)phenyl]-1-oxidanyl-ethenyl]sulfanylethylamino]-3-oxidanylidene-propyl]amino]-2,2-dimethyl-3-oxidanyl-4-oxidanylidene-butyl] hydrogen phosphate	YE2	1	917.67
TETRADEC-13-YNOIC ACID - COA THIOESTER	YNC	1	973.86

Table S2-3. Charge fitting between di-fragments to generate final partial charges present on Coenzyme A.

	Atom No.	Atom Name	No Restraints	Fragment 1-2	Fragment 2-3	Fragment 3-4	Fragment 4-5	Fragment 5-6	Absolute value of difference between two joined fragments, and one individual fragment.					New Charges	Diff all vs frag
Fragment 1	1	P3B	1.2362	1.2248					0.0114					1.2248	0.0114
	2	O7A	-0.9153	-0.9149					0.0004					-0.9130	0.0023
	3	O8A	-0.9153	-0.9149					0.0004					-0.9130	0.0023
	4	O9A	-0.9233	-0.9092					0.0141					-0.9130	0.0103
	5	O3B	-0.5414	-0.6292					0.0878					-0.6292	0.0878
Fragment 2	6	C1B	0.0381	0.0374	0.0407				0.0007	0.0026				0.0391	0.00095
	7	H1B	0.1986	0.1959	0.1965				0.0027	0.0021				0.1962	0.0024
	8	N1A	-0.7375	-0.7445	-0.7368				0.007	0.0007				-0.7407	0.00315
	9	C2A	0.5625	0.577	0.5617				0.0145	0.0008				0.5694	0.00685
	10	H2A	0.0508	0.0488	0.0505				0.002	0.0003				0.0497	0.00115
	11	C2B	0.1633	0.1716	0.164				0.0083	0.0007				0.1678	0.0045
	12	H2B	0.0863	0.1088	0.0835				0.0225	0.0028				0.0962	0.00985
	13	O2B	-0.6519	-0.626	-0.6492				0.0259	0.0027				-0.6376	0.0143
	14	HO2B	0.4434	0.4263	0.4403				0.0171	0.0031				0.4333	0.0101
	15	C3B	0.0813	-0.1296	0.0976				0.2109	0.0163				-0.0160	0.0973
	16	H3B	0.0569	0.1313	0.0596				0.0744	0.0027				0.0955	0.03855
	17	N3A	-0.6483	-0.6696	-0.6461				0.0213	0.0022				-0.6579	0.00955
	18	C4A	0.2263	0.2454	0.2226				0.0191	0.0037				0.2340	0.0077
	19	C4B	0.2375	0.1456	0.1905				0.0919	0.047				0.1681	0.06945
	20	H4B	0.1439	0.1876	0.162				0.0437	0.0181				0.1748	0.0309
	21	O4B	-0.4209	-0.3791	-0.4069				0.0418	0.014				-0.3930	0.0279
	22	C5A	0.126	0.1138	0.129				0.0122	0.003				0.1214	0.0046
	23	C5B	0.1087	0.1074	0.0102				0.0013	0.0985				0.0588	0.0499
	24	H5B2	0.0402	0.0519	0.0707				0.0117	0.0305				0.0613	0.0211
	25	H5B3	0.0402	0.0519	0.0707				0.0117	0.0305				0.0613	0.0211
	26	C6A	0.6712	0.6777	0.67				0.0065	0.0012				0.6739	0.00265
	27	N6A	-0.9238	-0.9244	-0.9236				0.0006	0.0002				-0.9240	0.0002
	28	H6A1	0.4136	0.4135	0.4134				0.0001	0.0002				0.4135	0.00015
	29	H6A2	0.4136	0.4135	0.4134				0.0001	0.0002				0.4135	0.00015
	30	N7A	-0.6045	-0.6021	-0.6049				0.0024	0.0004				-0.6035	0.001
	31	C8A	0.1403	0.1434	0.1373				0.0031	0.003				0.1404	5E-05
	32	H8A	0.1894	0.1862	0.1907				0.0032	0.0013				0.1885	0.00095
	33	N9A	0.008	0.0074	0.0108				0.0006	0.0028				0.0091	0.0011
Fragment 3	34	P1A	1.1573		1.16	1.1566				0.0027	0.0007			1.1583	0.001
	35	P2A	1.1531		1.1541	1.1554				0.001	0.0023			1.1548	0.00165
	36	O1A	-0.8261		-0.8221	-0.8252				0.004	0.0009			-0.8237	0.00245
	37	O2A	-0.7708		-0.7735	-0.7713				0.0027	0.0005			-0.7724	0.0016
	38	O3A	-0.466		-0.4674	-0.4668				0.0014	0.0008			-0.4671	0.0011
	39	O4A	-0.7706		-0.7728	-0.773				0.0022	0.0024			-0.7729	0.0023
	40	O5A	-0.8245		-0.8238	-0.821				0.0007	0.0035			-0.8224	0.0021
	41	O5B	-0.4849		-0.5263	-0.4846				0.0414	0.0003			-0.5055	0.02055
	42	O6A	-0.4852		-0.4851	-0.5199				0.0001	0.0347			-0.5025	0.0173
Fragment 4	43	CAP	-0.0031			-0.0186	-0.0215				0.0155	0.0184		-0.0201	0.01695
	44	HAP	0.1377			0.1425	0.1177				0.0048	0.02		0.1301	0.0076
	45	OAP	-0.6404			-0.6135	-0.6503				0.0269	0.0099		-0.6319	0.0085
	46	HOAP	0.4318			0.4145	0.4358				0.0173	0.004		0.4252	0.00665
	47	CBP	0.2412			0.1794	0.2444				0.0618	0.0032		0.2119	0.0293
	48	CDP	-0.0468			-0.0483	-0.0875				0.0015	0.0407		-0.0679	0.0211
	49	HDP1	0.0115			0.0177	0.0203				0.0062	0.0088		0.0190	0.0075
	50	HDP2	0.0115			0.0177	0.0203				0.0062	0.0088		0.0190	0.0075
	51	HDP3	0.0115			0.0177	0.0203				0.0062	0.0088		0.0190	0.0075
	52	CEP	-0.1901			-0.1784	-0.2033				0.0117	0.0132		-0.1909	0.00075
	53	HEP1	0.0492			0.0528	0.0513				0.0036	0.0021		0.0521	0.00285
	54	HEP2	0.0492			0.0528	0.0513				0.0036	0.0021		0.0521	0.00285
	55	HEP3	0.0492			0.0528	0.0513				0.0036	0.0021		0.0521	0.00285
	56	C9P	0.4104			0.428	0.5832				0.0176	0.1728		0.5056	0.0952
	57	O9P	-0.5198			-0.52	-0.5611				0.0002	0.0413		-0.5406	0.02075
	58	CCP	-0.0084			-0.0952	-0.0001				0.0868	0.0083		-0.0477	0.03925
	59	HCP2	0.0615			0.0912	0.0607				0.0297	0.0008		0.0760	0.01445
	60	HCP3	0.0615			0.0912	0.0607				0.0297	0.0008		0.0760	0.01445
Fragment 5	61	N8P	-0.5166				-0.3816	-0.5207				0.135	0.0041	-0.4512	0.06545
	62	H8P	0.2827				0.2458	0.2822				0.0369	0.0005	0.2640	0.0187
	63	C7P	-0.015				0.0575	-0.037				0.0725	0.022	0.0103	0.02525
	64	H7P2	0.0747				0.0439	0.0857				0.0308	0.011	0.0648	0.0099
	65	H7P3	0.0747				0.0439	0.0857				0.0308	0.011	0.0648	0.0099
	66	C6P	-0.0163				-0.0341	-0.0919				0.0178	0.0756	-0.0630	0.0467
	67	H6P2	0.0396				0.0411	0.0604				0.0015	0.0208	0.0508	0.01115
	68	H6P3	0.0396				0.0411	0.0604				0.0015	0.0208	0.0508	0.01115
	69	C5P	0.5012				0.4887	0.5828				0.0125	0.0816	0.5358	0.03455
	70	O5P	-0.5158				-0.5137	-0.5303				0.0021	0.0145	-0.5220	0.0062
Fragment 6	71	S1P	-0.3637					-0.3678					0.0041	-0.3678	0.0041
	72	C2P	-0.0013					0.0282					0.0295	0.0282	0.0295
	73	H21	0.091					0.082					0.009	0.082	0.009
	74	H22	0.091					0.082					0.009	0.082	0.009
	75	C3P	-0.0576					-0.0979					0.0403	-0.0979	0.0403
	76	H31	0.1054					0.1088					0.0034	0.1088	0.0034
	77	H32	0.1054					0.1088					0.0034	0.1088	0.0034
	78	N4P	-0.5459					-0.4577					0.0882	-0.4577	0.0882
	79	H4P	0.2963					0.2728					0.0235	0.2728	0.0235
	80	HS1	0.2004					0.2007					0.0003	0.2007	0.0003

Table S2-4. Bond length and angle measurements of phosphate fragments*

Methylphosphate (MP)		MP2/aug-cc-pVTZ	MP2/aug-cc-pVDZ	b3lyp/6-311+g(2d,p)			expt.	
Bond Definition		chg = -2	chg = -2	chg = -2	chg = 0 (2 Na+)	chg = 0 (Mg2+)	chg = -2	std dev
R(1,2)	C-H	1.0996	1.1114	1.1044	1.0974	1.0900	-	0.0080
R(1,3)	C-H	1.0996	1.1114	1.1044	1.0934	1.0900	-	0.0086
R(1,4)	C-H	1.1017	1.1140	1.1083	1.0911	1.0872	-	0.0113
R(1,9)	C-O	1.3957	1.4067	1.3895	1.4180	1.4458	1.433	0.0222
R(5,6)	O-P	1.5430	1.5761	1.5329	1.5836	1.5670	1.514	0.0217
R(5,7)	O-P	1.5430	1.5761	1.5329	1.5233	1.5670	1.514	0.0224
R(5,8)	O-P	1.5329	1.5661	1.5244	1.5164	1.5507	1.514	0.0202
R(5,9)	O-P	1.7553	1.8119	1.7802	1.6441	1.5781	1.621	0.0988
ngle Definition								std dev
A(2,1,3)	H-C-H	107.4476	107.4657	107.1928	109.4472	110.3618	-	1.4301
A(2,1,4)	H-C-H	108.7147	108.5007	108.1537	108.5601	110.2783	-	0.8290
A(2,1,9)	H-C-O	111.5211	111.7074	112.0174	111.2105	110.0271	-	0.7677
A(3,1,4)	H-C-H	108.7147	108.5007	108.1537	109.2831	110.2783	-	0.8307
A(3,1,9)	H-C-O	111.5211	111.7074	112.0173	110.9862	110.0271	-	0.7805
A(4,1,9)	H-C-O	108.8444	108.8597	109.1575	107.2835	105.7745	-	1.4367
A(6,5,7)	O-P-O	114.2659	114.4773	114.5799	108.1003	101.9873	113.0	5.5836
A(6,5,8)	O-P-O	116.3116	116.7081	116.2931	108.8042	103.5278	113.0	5.9295
A(6,5,9)	O-P-O	102.8882	102.4865	102.8622	106.2535	116.0711	107.5/105.2	5.7728
A(7,5,8)	O-P-O	116.3116	116.7081	116.2931	120.5946	103.5278	113.0	6.4948
A(7,5,9)	O-P-O	102.8882	102.4865	102.8622	108.4002	116.0711	107.5/105.2	5.8650
A(8,5,9)	O-P-O	100.9696	100.2950	100.6170	103.8056	113.8957	107.5/105.2	5.7507
A(1,9,5)	C-O-P	112.7463	111.7131	114.5334	118.7649	118.5850	119.1	3.2698

Dimethylphosphate (DMP		MP2/aug-cc-pVTZ	MP2/aug-cc-pVDZ	b3lyp/6-311+g(2d,p)		expt.	
Bond Definition		chg = -1	chg = -1	chg = -1		chg = -1	std dev
R(1,2)	C-H	1.0926	1.1042	1.0955		-	0.0060
R(1,3)	C-H	1.0939	1.1057	1.0974		-	0.0061
R(1,4)	C-H	1.0905	1.1023	1.0945		-	0.0060
R(1,13)	C-O	1.4151	1.4270	1.4131		1.439	0.0075
R(5,10)	O-P	1.5069	1.5370	1.4967		1.485	0.0210
R(5,11)	O-P	1.4968	1.5271	1.4871		1.485	0.0209
R(5,12)	O-P	1.6448	1.6871	1.6483		1.595	0.0235
R(5,13)	O-P	1.6693	1.7135	1.6737		1.595	0.0243
R(6,7)	C-H	1.0903	1.1020	1.0941		-	0.0060
R(6,8)	C-H	1.0930	1.1042	1.0965		-	0.0057
R(6,9)	C-H	1.0908	1.1023	1.0934		-	0.0060
R(6,12)	C-O	1.4177	1.4303	1.4169		1.439	0.0075
ngle Definition							std dev
A(2,1,3)	H-C-H	108.7519	108.8077	108.6778		-	0.0652
A(2,1,4)	H-C-H	109.6238	109.5003	109.3680		-	0.1279
A(2,1,13)	H-C-O	110.8670	111.0550	111.1312		-	0.1360
A(3,1,4)	H-C-H	108.9838	108.8860	108.6534		-	0.1697
A(3,1,13)	H-C-O	111.0381	111.1675	111.3624		-	0.1632
A(4,1,13)	H-C-O	107.5461	107.3832	107.5979		-	0.1120
A(10,5,11)	O-P-O	123.1208	124.0762	123.2799		119.3000	0.5119
A(10,5,12)	O-P-O	109.8847	110.0665	109.9485		111.1/105.1	0.0922
A(10,5,13)	O-P-P	107.3383	107.1974	107.4930		111.1/105.1	0.1479
A(11,5,12)	O-P-O	107.5172	107.2294	107.0773		111.1/105.1	0.2234
A(11,5,13)	O-P-O	109.5324	109.4349	109.5144		111.1/105.1	0.0519
A(12,5,13)	O-P-O	96.0927	94.9448	96.1527		105.0	0.6807
A(7,6,8)	H-C-H	109.2447	109.2887	108.9415		-	0.1890
A(7,6,9)	H-C-H	110.0229	109.9395	109.7894		-	0.1183
A(7,6,12)	H-C-O	107.2196	106.9275	107.3268		-	0.2067
A(8,6,9)	H-C-H	108.7618	108.8593	108.7719		-	0.0536
A(8,6,12)	H-C-O	110.7782	110.8474	110.9800		-	0.1025
A(9,6,12)	H-C-O	110.7955	110.9533	110.9999		-	0.1071
A(5,12,6)	C-O-P	115.8832	114.6512	117.7789		120.1	1.5755
A(1,13,5)	C-O-P	113.9638	112.7531	116.0331		120.1	1.6586

Dimethyldiphosphate (DMDP		MP2/aug-cc-pVDZ	b3lyp/6-311+g(2d,p)		
Bond Definition		chg = -2	chg = -2	hg = 0 (2 Na+)	std dev
R(1,2)	C-H	1.1054	1.0979	1.0896	0.0079
R(1,3)	C-H	1.1048	1.0963	1.0907	0.0071
R(1,4)	C-H	1.1020	1.0935	1.0897	0.0063
R(1,12)	C-O	1.4359	1.4223	1.4388	0.0088
R(5,7)	O-P	1.5327	1.4927	1.5027	0.0208
R(5,8)	O-P	1.5419	1.5015	1.5152	0.0205
R(5,9)	O-P	1.6854	1.6459	1.6239	0.0312
R(5,12)	O-P	1.7004	1.6605	1.6030	0.0490
R(6,9)	O-P	1.6854	1.6459	1.6239	0.0312
R(6,10)	O-P	1.5419	1.5015	1.5152	0.0205
R(6,11)	O-P	1.5327	1.4927	1.5027	0.0208
R(6,13)	O-P	1.7004	1.6605	1.6030	0.0490
R(13,14)	C-O	1.4359	1.4223	1.4388	0.0088
R(14,15)	C-H	1.1048	1.0963	1.0907	0.0071
R(14,16)	C-H	1.1020	1.0935	1.0897	0.0063
R(14,17)	C-H	1.1054	1.0979	1.0896	0.0079
Angle Definition					std dev
A(2,1,3)	H-C-H	109.3877	109.28	109.8949	0.3284
A(2,1,4)	H-C-H	110.0018	109.9917	110.5203	0.3023
A(2,1,12)	H-C-O	106.2729	106.6523	105.9272	0.3627
A(3,1,4)	H-C-H	109.3065	108.9524	109.9569	0.5095
A(3,1,12)	H-C-O	110.5302	110.6254	110.2724	0.1826
A(4,1,12)	H-C-O	111.2923	111.3051	110.2013	0.6336
A(7,5,8)	O-P-O	122.9940	122.0049	113.7134	5.0967
A(7,5,9)	O-P-O	105.5877	106.4240	106.5364	0.5183
A(7,5,12)	O-P-O	105.0411	105.1959	108.9508	2.2139
A(8,5,9)	O-P-O	111.4962	110.9183	110.8481	0.3557
A(8,5,12)	O-P-O	107.7740	108.1309	111.4888	2.0495
A(9,5,12)	O-P-O	101.7973	102.3109	104.8120	1.6128
A(9,6,10)	O-P-O	111.4962	110.9183	110.8481	0.3557
A(9,6,11)	O-P-O	105.5877	106.4241	106.5364	0.5184
A(9,6,13)	O-P-O	101.7973	102.3111	104.8120	1.6128
A(10,6,11)	O-P-O	122.9940	122.0050	113.7134	5.0967
A(10,6,13)	O-P-O	107.7740	108.1310	111.4888	2.0495
A(11,6,13)	O-P-O	105.0411	105.1952	108.9508	2.2141
A(5,9,6)	P-O-P	135.9497	139.8095	135.7177	2.2984
A(1,12,5)	C-O-P	115.2582	118.3887	121.1092	2.9279
A(6,13,14)	C-O-P	115.2582	118.3888	121.1092	2.9279
A(13,14,15)	O-C-H	110.5302	110.6258	110.2724	0.1828
A(13,14,16)	O-C-H	111.2923	111.3051	110.2014	0.6336
A(13,14,17)	O-C-H	106.2729	106.6528	105.9272	0.3629
A(15,14,16)	H-C-H	109.3065	108.9525	109.9569	0.5094
A(15,14,17)	H-C-H	109.3877	109.2794	109.8949	0.3286
A(16,14,17)	H-C-H	110.0018	109.9911	110.5203	0.3025

* Bond standard deviations are colored from blue to red on a scale of 0.0057 – 0.0988.
Angle standard deviations are colored from blue to red on a scale of 0.0519 – 6.4948.

References

1. White, S. W.; Zheng, J.; Zhang, Y. M.; Rock, The structural biology of type II fatty acid biosynthesis. *Annu Rev Biochem* **2005,** *74*, 791-831.
2. Sattely, E. S.; Fischbach, M. A.; Walsh, C. T., Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. *Nat Prod Rep* **2008,** *25* (4), 757-93.
3. Walsh, C. T.; O'Brien, R. V.; Khosla, C., Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew Chem Int Ed Engl* **2013,** *52* (28), 7098-124.
4. Weissman, K. J., The structural biology of biosynthetic megaenzymes. *Nat Chem Biol* **2015,** *11* (9), 660-70.
5. Khosla, C.; Herschlag, D.; Cane, D. E.; Walsh, C. T., Assembly line polyketide synthases: mechanistic insights and unsolved problems. *Biochemistry* **2014,** *53* (18), 2875-83.
6. Staunton, J.; Weissman, K. J., Polyketide biosynthesis: a millennium review. *Natural Product Reports* **2001,** *18* (4), 380-416.
7. Chan, D. I.; Vogel, H. J., Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem J* **2010,** *430* (1), 1-19.
8. Hertweck, C., The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl* **2009,** *48* (26), 4688-716.
9. Nielsen, J.; Keasling, J. D., Engineering Cellular Metabolism. *Cell* **2016,** *164* (6), 1185-97.
10. Yuzawa, S.; Kim, W.; Katz, L.; Keasling, J. D., Heterologous production of polyketides by modular type I polyketide synthases in Escherichia coli. *Curr Opin Biotechnol* **2012,** *23* (5), 727-35.
11. Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R.; Bruegger, J.; Caldara-Festin, G.; Finzel, K.; Jackson, D. R.; Ishikawa, F.; O'Dowd, B.; McCammon, J. A.; Opella, S. J.; Tsai, S. C.; Burkart, M. D., Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **2014,** *505* (7483), 427-31.
12. Lechner, A.; Wilson, M. C.; Ban, Y. H.; Hwang, J. Y.; Yoon, Y. J.; Moore, B. S., Designed Biosynthesis of 36-Methyl-FK506 by Polyketide Precursor Pathway Engineering. *Acs Synth Biol* **2013,** *2* (7), 379-383.
13. Dowling, D. P.; Kung, Y.; Croft, A. K.; Taghizadeh, K.; Kelly, W. L.; Walsh, C. T.; Drennan, C. L., Structural elements of an NRPS cyclization domain and its intermodule docking domain. *Proc Natl Acad Sci U S A* **2016,** *113* (44), 12432-12437.
14. Bravo-Rodriguez, K.; Klopries, S.; Koopmans, K. R.; Sundermann, U.; Yahiaoui, S.; Arens, J.; Kushnir, S.; Schulz, F.; Sanchez-Garcia, E., Substrate Flexibility of a Mutated Acyltransferase Domain and Implications for Polyketide Biosynthesis. *Chem Biol* **2015,** *22* (11), 1425-30.
15. Barajas, J. F.; Phelan, R. M.; Schaub, A. J.; Kliewer, J. T.; Kelly, P. J.; Jackson, D. R.; Luo, R.; Keasling, J. D.; Tsai, S. C., Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem Biol* **2015,** *22* (8), 1018-29.
16. Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: The Amber Lipid Force Field. *J Chem Theory Comput* **2014,** *10* (2), 865-879.
17. Skjevik, A. A.; Madej, B. D.; Walker, R. C.; Teigen, K., LIPID11: a modular framework for lipid simulations using amber. *J Phys Chem B* **2012,** *116* (36), 11124-36.

18. Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A., Forcefield_PTM: Charge and AMBER Forcefield Parameters for Frequently Occurring Post-Translational Modifications. *J Chem Theory Comput* **2013,** *9* (12), 5653-5674.

19. Khoury, G. A.; Smadbeck, J.; Tamamis, P.; Vandris, A. C.; Kieslich, C. A.; Floudas, C. A., Forcefield_NCAA: ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family. *ACS Synth Biol* **2014,** *3* (12), 855-69.

20. Homeyer, N.; Horn, A. H.; Lanig, H.; Sticht, H., AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J Mol Model* **2006,** *12* (3), 281-9.

21. Song, D.; Luo, R.; Chen, H. F., The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **2017,** *57* (5), 1166-1178.

22. Song, D.; Wang, W.; Ye, W.; Ji, D.; Luo, R.; Chen, H. F., ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chem Biol Drug Des* **2017,** *89* (1), 5-15.

23. Jackson, D. R.; Tu, S. S.; Nguyen, M.; Barajas, J. F.; Schaub, A. J.; Krug, D.; Pistorius, D.; Luo, R.; Muller, R.; Tsai, S. C., Structural Insights into Anthranilate Priming during Type II Polyketide Biosynthesis. *ACS Chem Biol* **2016,** *11* (1), 95-103.

24. Chan, D. I.; Stockner, T.; Tieleman, D. P.; Vogel, H. J., Molecular dynamics simulations of the Apo-, Holo-, and acyl-forms of Escherichia coli acyl carrier protein. *J Biol Chem* **2008,** *283* (48), 33620-9.

25. Chan, D. I.; Tieleman, D. P.; Vogel, H. J., Molecular dynamics simulations of beta-ketoacyl-, beta-hydroxyacyl-, and trans-2-enoyl-acyl carrier proteins of Escherichia coli. *Biochemistry* **2010,** *49* (13), 2860-8.

26. Anand, S.; Mohanty, D., Modeling holo-ACP:DH and holo-ACP:KR complexes of modular polyketide synthases: a docking and molecular dynamics study. *BMC Struct Biol* **2012,** *12*, 10.

27. Anand, S.; Mohanty, D., Inter-domain movements in polyketide synthases: a molecular dynamics study. *Mol Biosyst* **2012,** *8* (4), 1157-71.

28. Mugnai, M. L.; Shi, Y.; Keatinge-Clay, A. T.; Elber, R., Molecular dynamics studies of modular polyketide synthase ketoreductase stereospecificity. *Biochemistry* **2015,** *54* (14), 2346-59.

29. Li, X.; Chung, L. W.; Paneth, P.; Morokuma, K., DFT and ONIOM(DFT:MM) studies on Co-C bond cleavage and hydrogen transfer in B12-dependent methylmalonyl-CoA mutase. Stepwise or concerted mechanism? *J Am Chem Soc* **2009,** *131* (14), 5115-25.

30. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-74.

31. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **2006,** *25* (2), 247-60.

32. Mishra, P. K.; Drueckhammer, D. G., Coenzyme A Analogues and Derivatives: Synthesis and Applications as Mechanistic Probes of Coenzyme A Ester-Utilizing Enzymes. *Chem Rev* **2000,** *100* (9), 3283-3310.

33. Ray, L.; Valentic, T. R.; Miyazawa, T.; Withall, D. M.; Song, L.; Milligan, J. C.; Osada, H.; Takahashi, S.; Tsai, S. C.; Challis, G. L., A crotonyl-CoA reductase-carboxylase independent

pathway for assembly of unusual alkylmalonyl-CoA polyketide synthase extender units. *Nat Commun* **2016,** *7*, 13609.

34. Woods, R. J.; Chappelle, R., Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *Theochem* **2000,** *527* (1-3), 149-156.

35. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y., R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* **2011,** *39* (Web Server issue), W511-7.

36. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Gaussian, Inc.: Wallingford, CT, USA, 2009.

37. Schneider, B.; Kabelac, M.; Hobza, P., Geometry of the phosphate group and its interactions with metal cations in crystals and ab initio calculations. *J Am Chem Soc* **1996,** *118* (48), 12207-12217.

38. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P., A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **2003,** *24* (16), 1999-2012.

39. Lee, M. C.; Duan, Y., Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins* **2004,** *55* (3), 620-34.

40. Wang, L. P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martinez, T. J.; Pande, V. S., Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J Phys Chem B* **2017,** *121* (16), 4023-4039.

41. Guy, J. E.; Whittle, E.; Moche, M.; Lengqvist, J.; Lindqvist, Y.; Shanklin, J., Remote control of regioselectivity in acyl-acyl carrier protein-desaturases. *P Natl Acad Sci USA* **2011,** *108* (40), 16594-16599.

42. Mandel, A. L.; La Clair, J. J.; Burkart, M. D., Modular synthesis of pantetheine and phosphopantetheine. *Org Lett* **2004,** *6* (26), 4801-3.

43. Kolossvary, I.; Guida, W. C., Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *J Am Chem Soc* **1996,** *118* (21), 5011-5019.

44. Case, D. A.; Cerutti, D. S.; Cheatham, I., T.E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.;

Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *AMBER 2017*, University of California, San Francisco, 2017.

45. Klimecka, M. M.; Chruszcz, M.; Font, J.; Skarina, T.; Shumilin, I.; Onopryienko, O.; Porebski, P. J.; Cymborowski, M.; Zimmerman, M. D.; Hasseman, J.; Glomski, I. J.; Lebioda, L.; Savchenko, A.; Edwards, A.; Minor, W., Structural analysis of a putative aminoglycoside N-acetyltransferase from Bacillus anthracis. *J Mol Biol* **2011,** *410* (3), 411-23.

46. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-12.

47. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-95.

Reproduced with permission from Barajas, J. F.; Phelan, R. M.; Schaub, A. J.; Kliewer, J. T.; Kelly, P. J.; Jackson, D. R.; Luo, R.; Keasling, J. D.; Tsai, S. C., Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem Biol* **2015,** *22* (8), 1018-29. Copyright 2015 Elsevier Inc. except certain content provided by third parties

CHAPTER 3

Comprehensive Structural Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols

3.1 Summary

Termination domains found in modular mega-synthases such as polyketides synthases (PKS) and non-ribosomal peptide synthetases (NRPS) are responsible for the release of covalently attached intermediates and, in the process, generate functional group diversity contingent on the mechanism employed. The terminal reductase (R) domain from the non-ribosomal peptide synthetase (NRPS) module MxaA in *Stigmatella aurantiaca* Sga15 catalyzes a non-processive four-electron reduction to produce the myxalamide family of secondary metabolites. Despite widespread use in nature, a lack of structural and dynamic information concerning reductive release from polyketide synthase (PKS) and NRPS assembly lines principally limits our ability to redesign R domains with altered or improved activity. The reductase domain of MxaA (MxaAR) was recently solved to 1.90 Å and 1.84 Å in the presence and absence of NADPH, respectively. This structure represents the first cofactor bound, and highest resolution reductase domain. Molecular dynamic simulations delivered an improved picture – beyond traditional structural studies – of key protein-protein and protein-substrate interactions that combined with structural data and provided basis for biochemical investigations. This was also the first time MD was used to decipher the protein-substrate interactions in the active site of a reductase. In addition, MD and docking studies

were used to understand protein-protein interactions between the MxaA reductase and its corresponding peptidyl carrier protein (MxaAPCP). Mutational analysis focused both on the putative catalytic residues and substrate-binding pocket to define the necessity of the catalytic triad and reveal select residues that are highly influential in catalysis. The combined data provides an unparalleled view of this unique termination mechanism that spans from macromolecular movements essential for catalysis to the identification of key substrate-residue interactions. In summary, studies presented here will aid efforts to improve these domains for the production of diverse primary alcohols. This possibility was highlighted by the enhancement of activity towards fully saturated compounds, specifically C_{10} derivatives, through mutation guided by our structural, biochemical and computational results.

3.2 Introduction

The myxobacterium *Stigmatella aurantiaca* Sga15 contains a myxalamid biosynthetic pathway composed of six PKS modules, and one terminal NRPS module (Fig. 3-1). Myxalamids have been identified as potent inhibitors of the respiratory electron transport chain.[1-2] What makes this particular system interesting is its use of a rare termination mechanism to release **6** as a primary alcohol, in contrast to more commonly observed chain release mechanisms using thioesterases (TEs) that produce, for instance, macro–lactones or –lactams.[1,3-5] The biosynthesis of myxalamid is a multi-step process initiated by a type I modular PKS that consists of six modules to biosynthesize **3**, a polyene intermediate that is translocated to the NRPS module for final processing. In the terminal NRPS module, MxaA, the adenylation (A) domain activates alanine as a building block while the condensation (C) domain catalyzes peptide bond formation between alanine and the PKS-generated intermediate to yield the final product.[6]

The last step in biosynthesis requires the reductive release of myxalamid (mxa) from the phosphopantetheine (PPant) prosthetic group covalently attached to the peptidyl carrier protein (PCP). This action is catalyzed by a recently described class of NADPH-dependent terminal reductase (R) domains that execute chain termination by a 4e$^-$ non-processive reduction to generate primary alcohols.[1, 4, 7] To accomplish this, the PCP-bound thioester is first reduced to the aldehyde **5**, which, following reduction by a second NADPH equivalent, affords the final 2-aminopropanol-containing **6**.

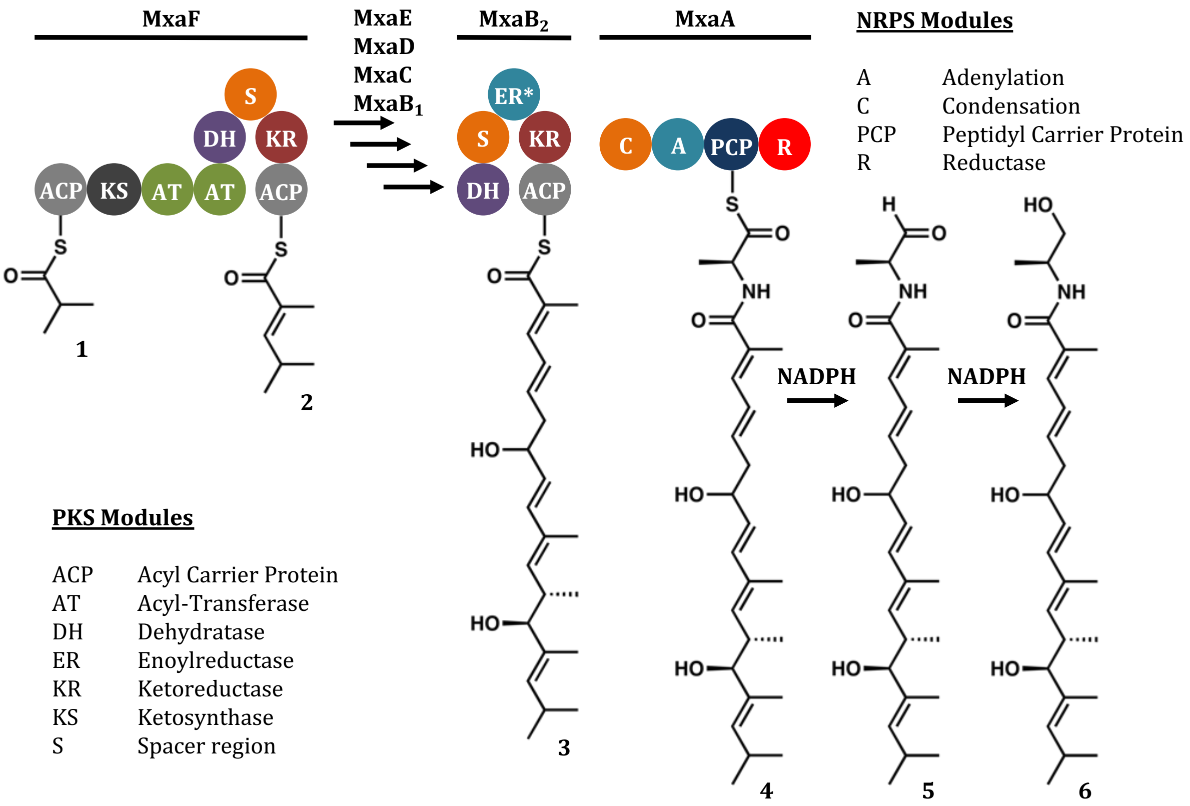


Figure 3-1. Myxalamid biosynthetic pathway: The pathway is composed of six PKS modules (MxaB1 – MxaF) and one terminal NRPS module (MxaA) containing a reductase domain (in red).

Here, we report the 1.84-Å and 1.90-Å structures of the MxaA R domain from *S. aurantiaca* Sga15 in the presence and absence of NADPH, respectively. This, in combination with molecular dynamics (MD) and structure-based mutagenesis, provided an unprecedented view of local and global interactions between the PCP and R domain, and those between the R domain and cofactor/substrate that are essential for catalysis. Furthermore, mutational analysis of the R domain enabled us to rationally mutate a key active site arginine that resulted in an MxaA variant with improved activity towards highly reduced substrates (e.g., dodecanoyl-PCP). Combined structural, computational and biochemical results presented here provide a comprehensive understanding of these unique termination domains and, in the process, set a strong foundation for future efforts to generate new PKS- or NRPS-based routes to diverse terminal alcohol containing compounds.

3.3 Results and discussions

3.3.1 The crystal structure of the MxaR R domain

The MxaA reductase is composed of an N-terminal subdomain that contains NADPH bound in a Rossman fold, and the C-terminal subdomain that contains a helix-turn-helix motif (Fig. 3-2).[8,9] A comparison of the apo and NAPDH-bound MxaA reductase shows a slight conformational change with an overall RMSD of 0.63 Å.

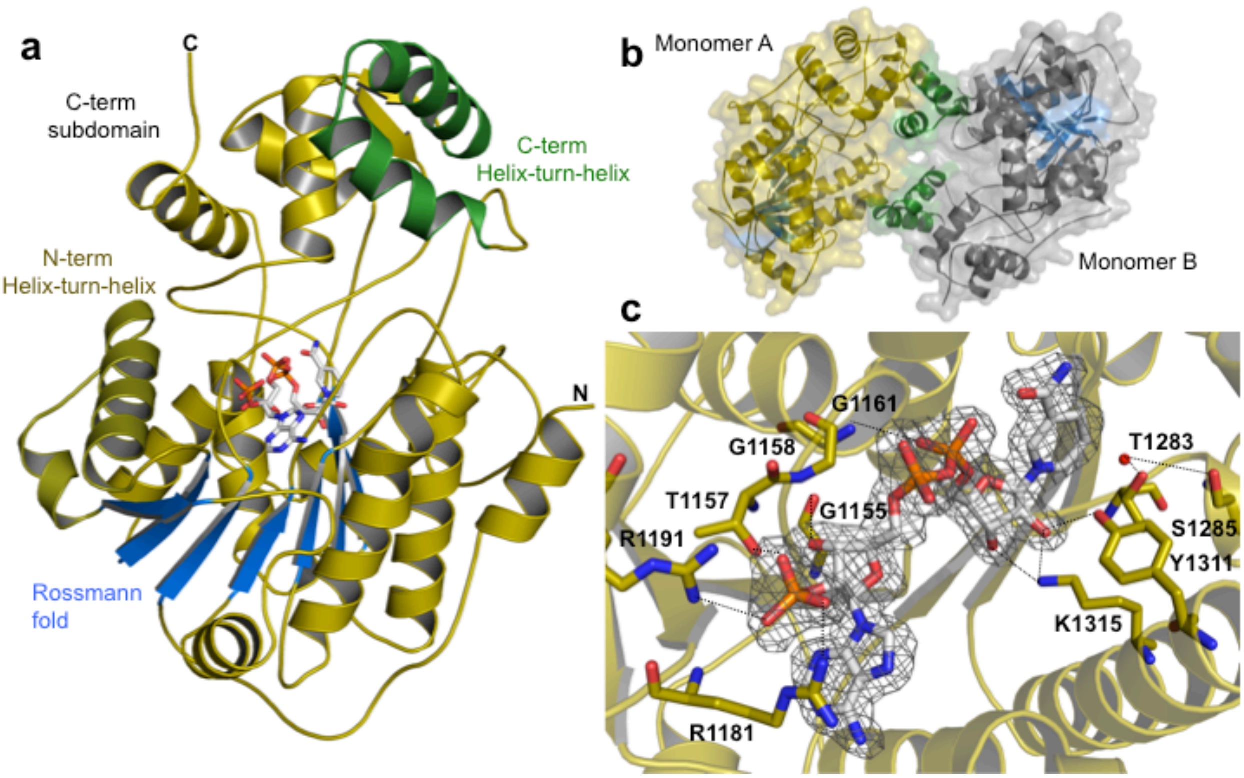


Figure 3-2. Structure of the MxaA R domain. (a) The MxaA R domain monomer is composed of an N-terminal subdomain that contains an NADPH Rossmann fold (in blue) and a C-terminal subdomain which contains a helix-turn-helix motif (shown in green). The NADPH cofactor is displayed in gray sticks. (b) The MxaA R domain crystallizes as a dimer, monomer A shown in yellow and monomer B shown in gray. (c) The cofactor NADPH binds to the TGxxGxxG motif close to the T, Y, and K catalytic site. An SA-omit map of the NADPH co-factor contoured at 1.0σ is shown in the gray isomesh map.

3.3.2 Molecular dynamic studies

To elucidate the structural dynamics of NADPH and substrate binding in MxaA R we conducted molecular dynamic (MD) simulations by analyzing conformational changes in 100 ns MD runs. Atomic coordinates of the MxaA R domain were obtained from the NADPH-bound MxaA R domain (chain B) crystal structure. The ff14SB forcefield in Amber14 was used for the protein and the general AMBER force field (GAFF) was used for the NADPH cofactor.[10-15] NADPH was parameterized using Gaussian 09 to obtain the initial electrostatic potential using the HF/6-31G(d,p) basis set, followed by the use of antechamber to obtain the HF/6-31G(d,p) restricted electrostatic potential (RESP) fit with final overall net charge of -4. The system was explicitly solvated with a buffer of 10 Å TIP3P waters in a truncated octahedron box after neutralizing with counter ions. A two-system minimization was performed using SANDER and PMEMD was used for production runs.[16]

The NADPH-bound MxaA R domain was allowed to equilibrate after heating the system to 300K and subsequently allowed to run over 100 ns. 2D RMSD maps were generated using Chimera and an in-house MATLAB script by comparing RMSD fluctuations of the protein backbone. The maps revealed an RMSD range of 0.61-2.41 Å for the MxaA R domain (Fig. 3-3a). Further dissection of the N- vs. C-term subdomains revealed RMSD ranges of 0.54-1.49 Å and 0.52-2.25 Å, respectively (Fig. 3-3b,c). These results, combined with RMSD values found in our crystal structures, indicate higher flexibility and movement of the C-terminal subdomain.

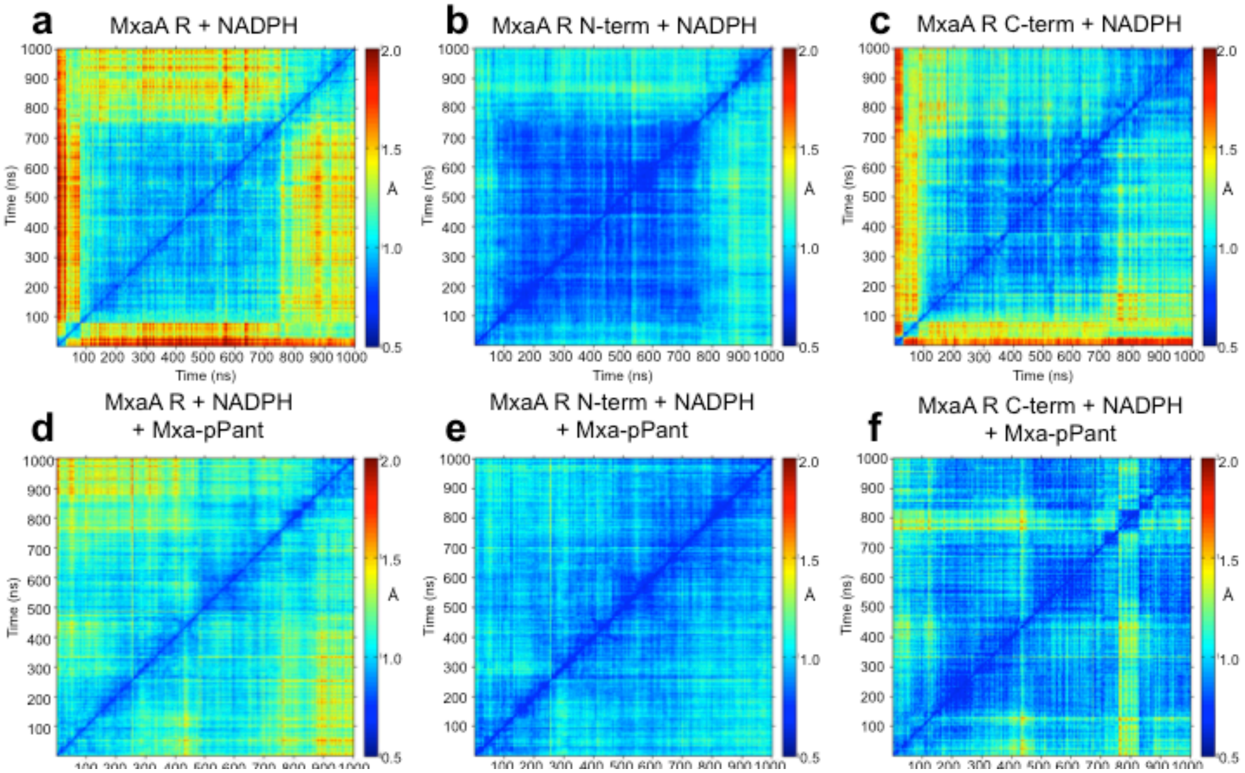


Figure 3-3. Molecular Dynamic Analysis (a) 2D RMSD map analysis of the MxaA R domain backbone with NADPH over the entire 100 ns molecular dynamics simulation. Low RMSD is observed in blue and high RMSD is observed in red. (b,c) Dissecting the N- vs. C-terminal subdomain of MxaA R bound to NAPDH reveals higher RMSD deviations in the C-terminal subdomain. (d) 2D RMSD map was generated with the MxaA R domain bound to NADPH and docked with mxa-PPant. (e,f) Dissection of the N- vs. C-terminal of the bound NADPH, mxa-PPant R domain demonstrate a decrease in movement of the C-terminal subdomain.

The most noticeable region of flexibility was observed in the C-terminal helix-turn-helix (HTH) motif, specifically the conserved hydrophobic residues between Y1430 and Q1455 of α16-α17, which display an average RMSD of 0.82 Å in the NADPH bound model (Fig. 3-2a, Fig. 3-4g-i). Numerous salt bridges are critical in stabilizing the α16-α17 HTH motif, such as R1426 and E1436 (Fig. 3-4a-c). During the 100 ns NADPH bound run, the ζC of R1426 maintains a distance of ≤ 6.0 Å with either εO of E1436. D1444, the turn residue between helix-16 and helix-17, also maintains a tight salt bridge interaction with a distance ≤ 3.5 Å between the ζC of R1364 through stochastic sampling of either helix-13 D1444 δO during 73.2% of the simulation (Fig. 3-4a).

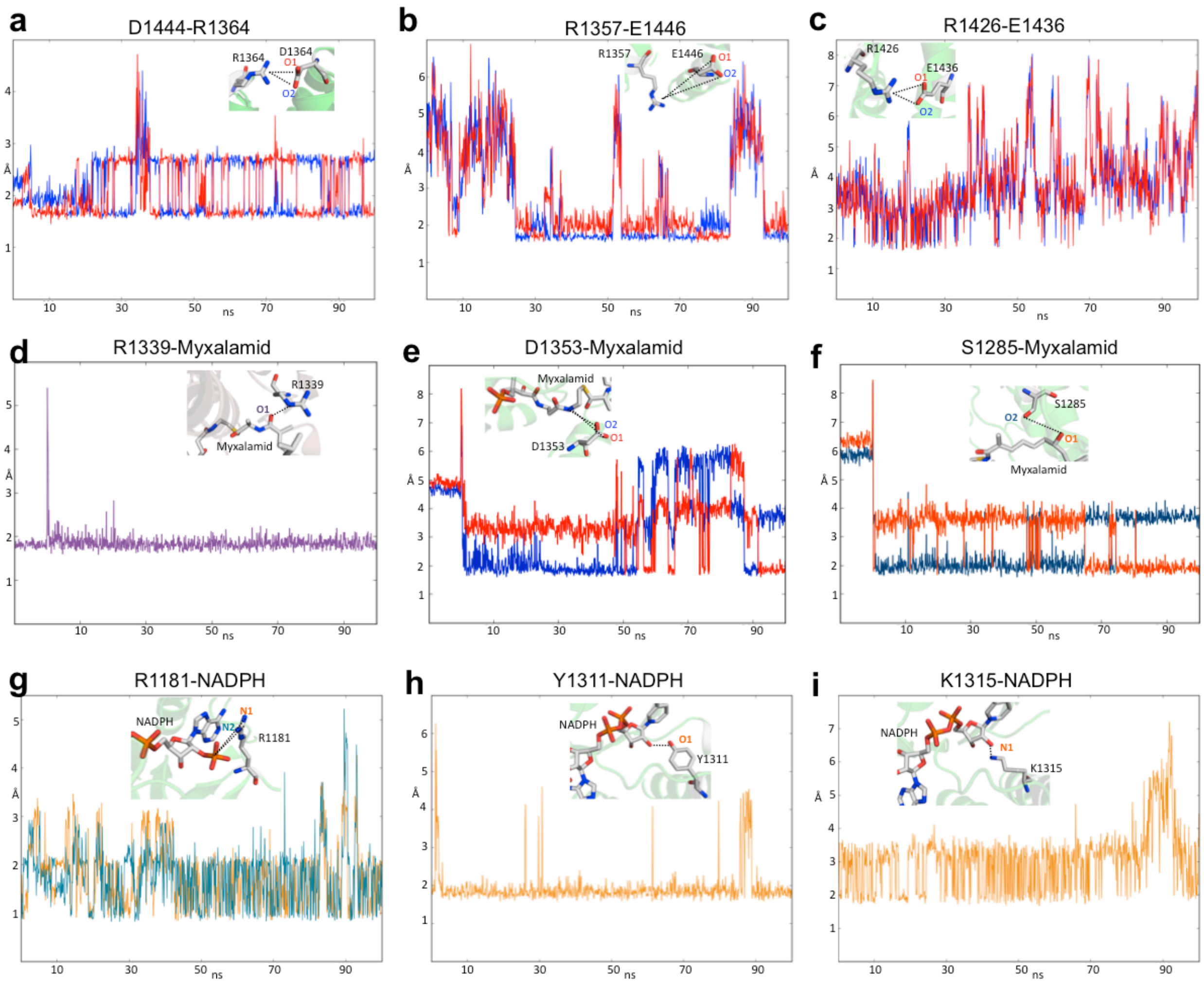
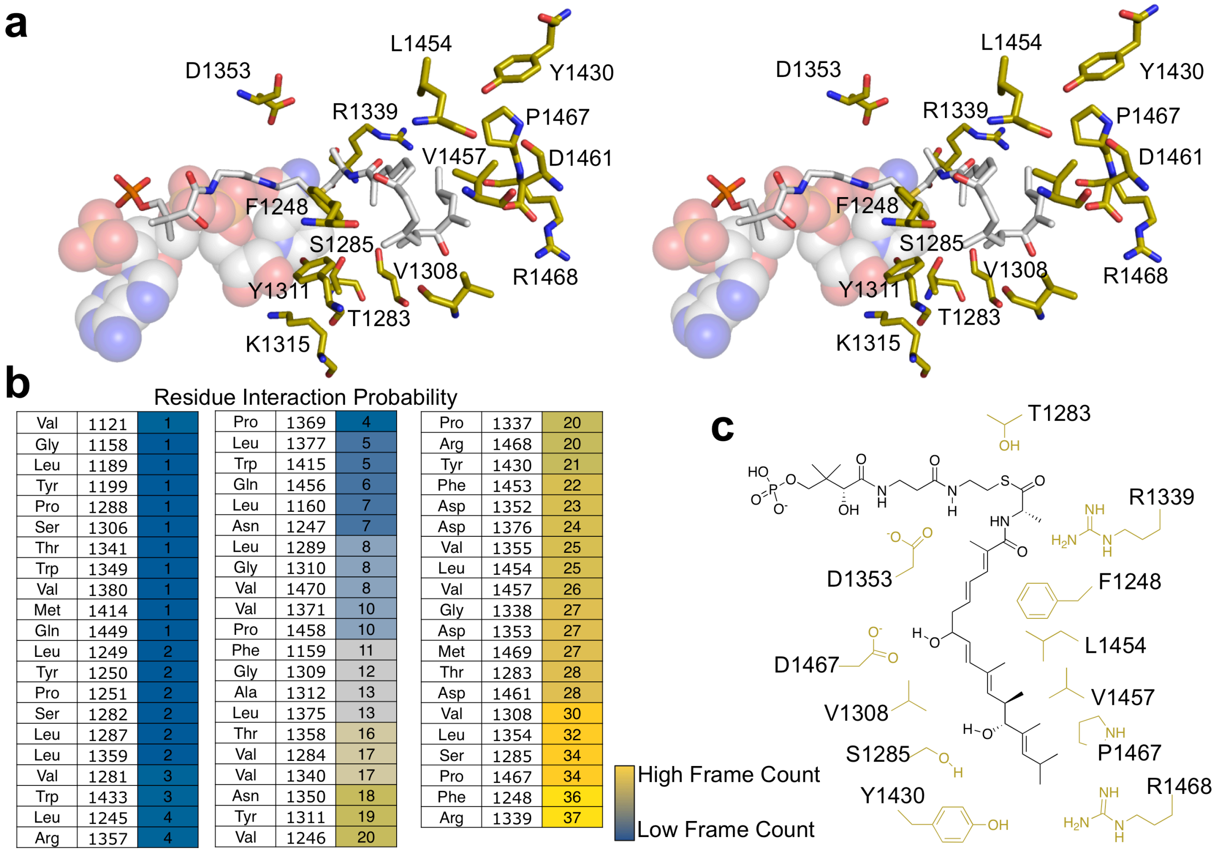


Figure 3-4. Further MD analysis of residues within the HTH of the C-terminal subdomain (a-c). (d-f) Residue analysis of mxa-PPant interactions and NADPH (g-i).

Moderate electrostatic interactions were observed between helix-12 R1357 and helix-17 E1446, with the ζC of R1357 maintaining a distance ≤ 6.0 Å for 68.7% of the simulation. The catalytic triad (T1283, Y1311 and K1315) exhibits little movement with an average of 0.07 Å per residue throughout the entire 100 ns run. The phosphate attached to the nicotinamide ribose 5' carbon remains stable, within 1.99 and 2.09 Å of the Rossmann TGxxGxxG motif.



Residue Interaction Probability

Val	1121	1	Pro	1369	4	Pro	1337	20
Gly	1158	1	Leu	1377	5	Arg	1468	20
Leu	1189	1	Trp	1415	5	Tyr	1430	21
Tyr	1199	1	Gln	1456	6	Phe	1453	22
Pro	1288	1	Leu	1160	7	Asp	1352	23
Ser	1306	1	Asn	1247	7	Asp	1376	24
Thr	1341	1	Leu	1289	8	Val	1355	25
Trp	1349	1	Gly	1310	8	Leu	1454	25
Val	1380	1	Val	1470	8	Val	1457	26
Met	1414	1	Val	1371	10	Gly	1338	27
Gln	1449	1	Pro	1458	10	Asp	1353	27
Leu	1249	2	Phe	1159	11	Met	1469	27
Tyr	1250	2	Gly	1309	12	Thr	1283	28
Pro	1251	2	Ala	1312	13	Asp	1461	28
Ser	1282	2	Leu	1375	13	Val	1308	30
Leu	1287	2	Thr	1358	16	Leu	1354	32
Leu	1359	2	Val	1284	17	Ser	1285	34
Val	1281	3	Val	1340	17	Pro	1467	34
Trp	1433	3	Asn	1350	18	Phe	1248	36
Leu	1245	4	Tyr	1311	19	Arg	1339	37
Arg	1357	4	Val	1246	20			

Figure 3-5. (a) Cartoon and ribbon docking model of MxaA PCP-R domain interactions. The green dot represents the active site cavity of the MxaA R domain. (b) Model of the MxaA PCP-R interface reveals possible electrostatic and pi-stacking interactions between these two domains.

A representative cluster ensemble was generated from MD using RMSD scoring as implemented in Chimera.[17] RMSD scoring reduced the initial set of 1000 frames generated to the 46 most unique frames. *In silico* docking of the PPant-bound substrate using all of the 46 unique frames from the previous MD run by the program GOLD revealed a large binding cavity under the α16-α17 helix-turn-helix motif (Fig. 3-2a).[16] In order to identify substrate-binding residues for MD analysis, we docked the myxalamid substrate using the 46 unique clusters from the NADPH-bound MxaAR MD analysis. We ranked the docking solutions using the ChemPLP scoring function and identified the most consistent binding orientation of the myxalamid substrate by tallying residues involved in substrate binding (Fig. 3-5). The frame from the top ChemPLP solution that was pre-screened by binding orientation was utilized for MD analysis of the MxaA R with the myxalamid substrate. Using the same MD system parameters as before we allowed the PPant-bound substrate and NADPH-bound R domain chain B simulation to run for 100 ns. RMSD 2D map analysis of the MxaA R domain in complex with the PPant substrate revealed a RMSD range of 0.60- 1.99 Å (Fig. 3-3d). This value is lower than the RMSD range for the R domain with NADPH bound, but lacking substrate (Fig. 3-3a), and indicates a decrease in protein motion upon substrate binding. The NADPH binding N-terminal subdomain demonstrates a similar RMSD range of 0.55-1.49 Å, whether substrate is bound or not, while the C-terminus reduces its flexibility upon substrate binding (Fig. 3-3c,f).

In light of results that indicated substrate binding stabilized the C-terminus we opted to focus additional attention on the C-terminal HTH motif, specifically those residues that are steadied through interactions with the substrate (Fig. 3-5a-c). The terminal HTH motif displays a slightly lower average RMSD of 0.76 Å, while the PPant moiety exhibits larger movements than the sequestered myxalamid segment. D1353 shows strong hydrogen bonding with the amide moiety of the PPant group, averaging a 3.0-Å distance for more than 70 ns of the MD run (Fig. 3-4e). The amide carbonyl group of the terminal alanine in the mxa intermediate generates a tight 2.0-Å interaction with R1339, highlighting the likely importance of electrostatic interactions between myxalamid and the R domain (Fig. 3-4d). The methyl-branched diene moiety of the mxa substrate forms intramolecular hydrophobic interactions, kinking the aliphatic substrate back towards the PPant thioester to minimize its hydrophobic surface area. F1248 in MxaA R aids in stabilizing these hydrophobic interactions. The C_9 and C_{15} hydroxyl groups in myxalamid intermediate **4** hydrogen bond with S1285 and D1461, respectively (Fig. 3-5a,c). S1285 forms an average 3.0-Å hydrogen bond with the terminal hydroxyl group in the myxalamid intermediate for more than 95% of the MD simulation. Similarly, D1461 forms an average 3.2-Å hydrogen bond with the first hydroxyl group for more than 80% of the MD run. Both S1285 and D1461 along with R1339 and F1248 appear to play a key role in myxalamid substrate recognition and orienting the substrate for reduction by NADPH. In summary, MD simulation results suggest that the C-terminal domain is highly mobile until myxalamid substrate binding quenches movement, more precisely at the HTH motif, for the first round of reduction.

3.3.3 Docking analysis of PPant-mxa substrate and PCP domain

Initial structural analysis in parallel with the NADPH-bound MxaA R MD analysis revealed a large substrate cavity with various potential substrate binding residues. In depth docking studies were needed because the deep active site cavity contains many potential substrate-binding motifs, and the long chain substrate is inherently flexible. Using *in silico* docking we further probed for residues important in substrate binding by docking the PPant-tethered mxa intermediate in the R domain active site. The 100-ns MD simulation of NADPH bound MxaA R identified 46 unique clusters, indicative of 46 distinct MxaA R domain conformations. One frame from each cluster was obtained and was used as the receptor to dock against the PPant-mxa ligand using the program GOLD.[18] Each frame generated 100 solutions that were scored and ranked using the ChemPLP scoring function.[19] The program LIGPLOT, in parallel with visual inspection, was used to analyze and identify ligand-protein residue interactions (Fig. 3-6).[20]

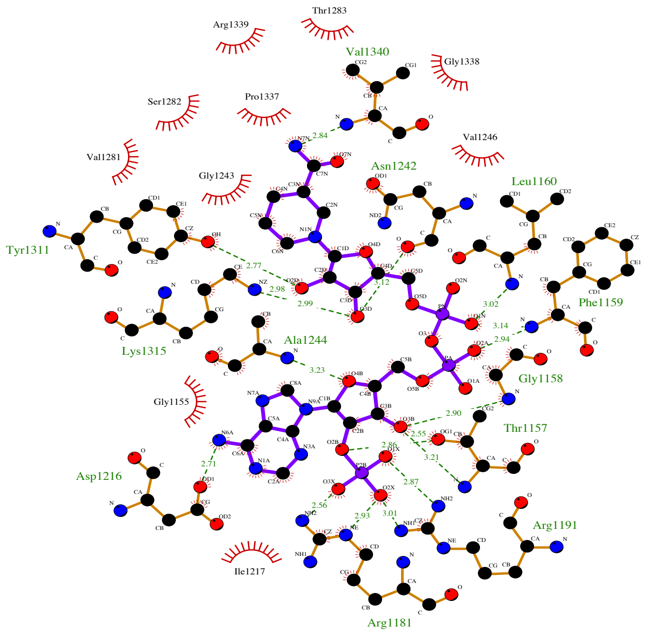


Figure 3-6. LIGPLOT analysis revealing key protein-substrate interactions from cluster-analysis.

The residue–ligand interactions between 2.5-4.0 Å for each frame were tallied, and a heat map was generated, indicative of ligand-residue proximity in different R domain conformations (Fig. 3-5b). Not surprisingly, the catalytic triad was revealed to frequently associate with the PPant-bound substrate. T1283 showed interactions close to the thioester linkage in 28 out of the 46 frames. Additionally, Y1311 interacted with the thioester in 18 out of the 46 frames. The water coordinating S1285 associated with the thioester in 35 out of the 46 frames. The majority of residues that interacted with the MxaA portion of the ligand were localized on the C–terminal subdomain. Out of the 46 frames, 37 showed that R1339 engages in electrostatic interactions with one of the two carbonyl groups in the substrate near the thioester linkage. Residues that outline the back of the substrate-binding pocket revealed several hydrogen bonding and hydrophobic interactions with the substrate, specifically, V1308, Y1430 and R1468. Taken together, residues with high substrate contact probabilities—derived from resultant heat maps—indicate the likely importance of specific residues in substrate recognition and orientation.

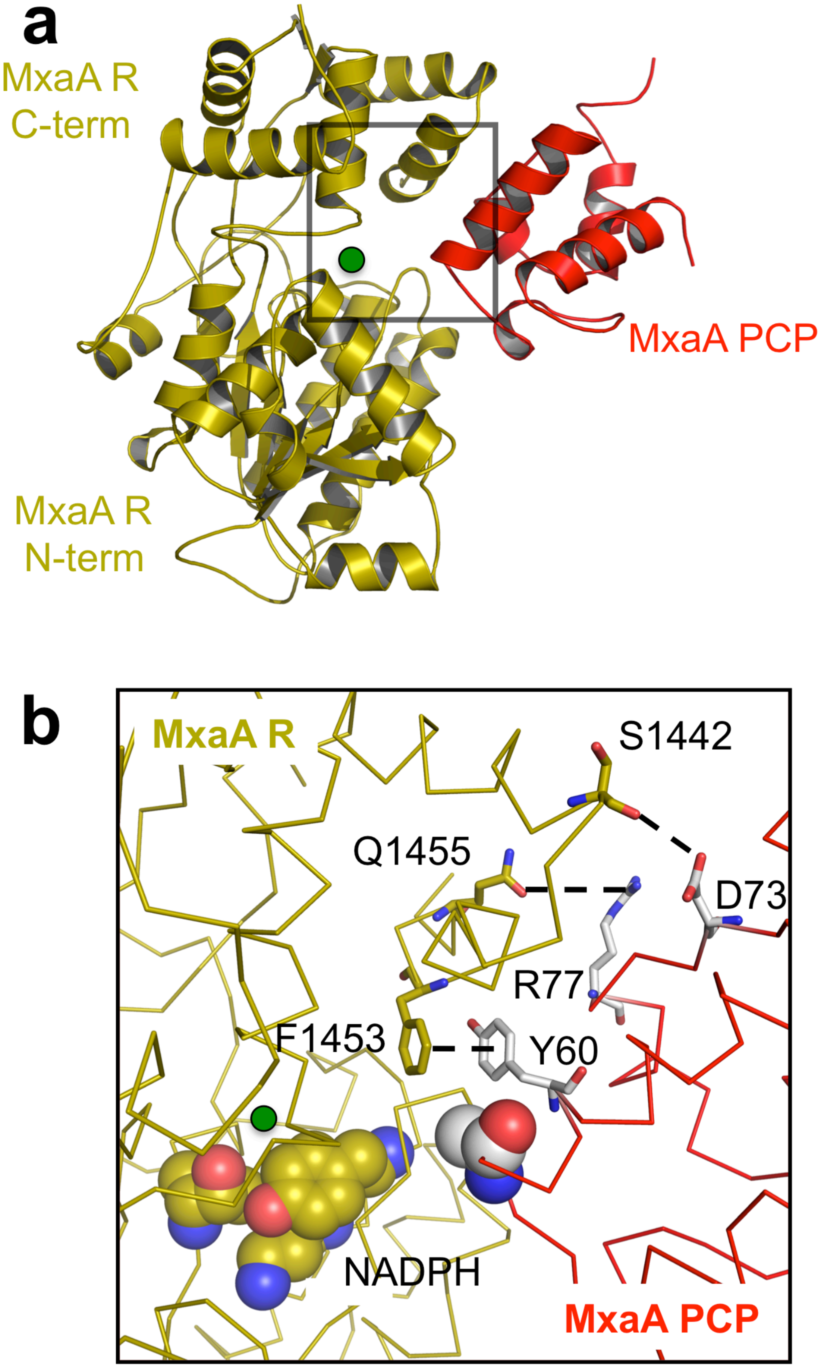


Figure 3-7. (a) Cartoon and ribbon docking model of MxaA PCP-R domain interactions. The green dot represents the active site cavity of the MxaA R domain. (b) Model of the MxaA PCP-R interface reveals possible electrostatic and pi-stacking interactions between these two domains.

To situate the R domain in the perspective of the termination module MxaA, more specifically the protein interactions between PCP and R domain for the reductive release of the final product, we computationally docked the R domain with the PCP domain. The previously solved SrfA-C terminal module structure from *Bacillus subtilis* revealed that the PCP domain is positioned in close proximity to the catalytic C, A and TE domains, thus providing evidence for the spatial relationships of the PCP in the termination module.[21] Accordingly, we used the SrfA-C terminal domain structure (PBD ID: 2VSQ) as a template on which to base our protein-protein docking studies. Using the HHPRED server we generated a tertiary homology model of the MxaA PCP domain and proceeded to dock the R domain using the ZDOCK server.[19,22-23] Most of helix III in the PCP forms contacts with surface residues in both N- and C-terminal subdomains of the MxaA R domain (Fig. 3-7a-b). Dissecting the PCP surface reveals electrostatic interactions between the surfaces of the R and PCP domain. The conserved serine (part of the signature D/HSL motif) contained in the PCP domain that covalently binds the PPant prosthetic group (S56) is located 12.0 Å away from the catalytic triad: T1283, Y1311 and K1315. The HTH motif of the MxaA R C-terminal subdomain has both electrostatic and pi-pi stacking interactions with helix III of the PCP domain. These include the Q1445 of the R domain with R77 from the PCP domain and the carbonyl backbone of S1442 of the R domain with the D73 side chain of the PCP. Additionally, a face-to-face pi stacking interaction occurs between F1453 in the R domain and Y60 of the loop connecting helix II and III in the PCP domain (Fig. 3-7b). These results highlight the importance of electrostatic and aromatic residues on the surfaces of both the R and PCP domains for protein-protein interactions and provide a basis for the engineering of chimeric ACP/PCP-R domain fusions.

3.3.4 Structure of the MxaA R domain

To visualize MxaA R we crystallized the R domain with and without the cofactor, NADPH. Using multi-wavelength anomalous diffraction (MAD) with a selenomethionine-substituted protein the structure of MxaA R was determined to a resolution of 1.95 Å. The apo MxaA R structure was further refined to 1.84 Å (**Table 3-1**). MxaA R was solved as a dimer with an RMSD of 0.45 Å between monomers A and B.

The overall structure contains strong architectural similarities to type E short-chain dehydrogenases/reductases (SDRs) that contain an N-terminal NADPH binding region and a C-terminal substrate-binding subdomain (Fig. 3-2a-c).[24] Structural alignment with the type E SDR from *agrobacterium tumefaciens* (PDB: 4ID9) displays an R.M.S.D. of 3.88 Å through 119 residues of the alpha carbon backbone. Hidden Markov models (HMM) show that MxaA R has structural and sequence similarities with the extended type-E SDRs based on Kallberg *et al.* classification.[25] The N-terminal subdomain contains an extended NADPH-binding α/β Rossmann fold with seven parallel beta sheets (β3- β2- β1- β4- β5- β6- β10) flanked by five alpha helices (α2- α3- α4- α6- α8- α11) (Fig. 3-2a, Fig. 3-8, Fig. 3-9). These structural features correlate well with the previously solved Nrp R domain structure (PDB: 4DQV) with an RMSD of 2.19 Å through 279 residues of the alpha carbon backbone.[7] Similar features include a canonical tyrosine-dependent catalytic triad (T1283, K1315 and Y1311) and a distinctive helix-turn-helix (HTH) motif (α16- α17) found in all structurally known terminating reductase domains.

Table 3-1. Crystallographic statistics.

	MxaA R Native	MxaA R/ NADPH		MxaA R SeMet	
Data collection					
Space group	P21	P21		P21	
Cell dimensions					
a, *b*, *c* (Å)	50.84, 159.30, 54.90	51.11, 159.05, 51.18		50.88, 159.23, 54.92	
a, b, g (°)	90, 108.88, 90	90, 106.14, 90		90, 108.85, 90	
			Peak	*Inflection*	*Remote*
Wavelength	0.9792	0.9775	0.9792	0.9794	0.9611
Resolution (Å)	50.0-1.70 (1.96-1.70)	50.0-1.84 (2.01-1.84)	50.0-1.70 (2.01-1.70)	50.0-1.70 (2.01-1.70)	50.0-1.70 (2.01-1.70)
R_{sym} or R_{merge}	0.075 (0.200)	0.211 (0.545)	0.140 (0.488)	0.142 (0.499)	0.136 (0.460)
CC*	0.995	0.980	0.992	0.991	0.989
I /*sI*	23.0 (7.1)	9.3 (3.9)	66.6 (7.4)	66.7 (7.5)	58.5 (6.7)
Completeness (%)	98.1 (98.1)	100.0 (100.0)	96.2(93.3)	97.3 (92.5)	97.4 (92.4)
Redundancy	2.7 (2.6)	3.6 (3.1)	6.3 (5.0)	6.3 (5.0)	6.3 (5.0)
Refinement					
Resolution (Å)	29.18-1.84 (1.91-1.84)	41.82-1.90 (1.96-1.90)		33.37-1.95 (2.02-1.95)	
No. reflections	367015	440613		508072	
R_{work} / R_{free}	0.1750/ 0.2150 (0.2233/0.2898)	0.1630/0.2080 (0.2818/0.3260)		0.2068/0.2393 (0.2380/0.2965)	
No. atoms					
Protein	5991	6052		5959	
Ligand/ion	8	104		8	
Water	934	786		822	
B-factors					
Protein	24.0	29.6		20.1	
Ligand/ion	17.7	18.2		20.0	
Water	32.3	35.8		28.7	
R.m.s deviations					
Bond lengths (Å)	0.008	0.015		0.020	
Bond angles (°)	1.05	1.4		1.53	

Substrate recognition in the SDR family occurs in the C-terminal subdomain, consequently, while the N-terminal subdomains in SDRs are highly conserved, C-terminal domains often differ in sequence.[26] The C-terminal subdomain of MxaA R consists of 5 helices (α12- α15- α16- α17- α20) and two parallel beta sheets (β9- β11), which are substantially larger (~130 residues) than those found in typical SDRs (Fig 3-9a-c).[24-26] A notable inserted helix-turn-helix motif (α16- α17) between residues Y1431 and Q1456 contains several conserved hydrophobic residues (W1433, L1437, L1450, L1451) frequently present in R domains that conduct PKS or NRPS chain termination with 2- or 4-electron reductions (Fig. 3-1c).[1, 3, 27-30]

To further distinguish true biological interfaces from lattice contacts in the crystal structure, we further analyzed the MxaA R domain utilizing the Evolutionary Protein-Protein Interface Classifier (EPPIC) server, which relies on evolutionary data to detect biological interfaces, and PDBePISA.[31,32] The EPPIC server was unable to reliably determine biologically relevant surface interfaces due to the lack of homolog sequences for comparison.

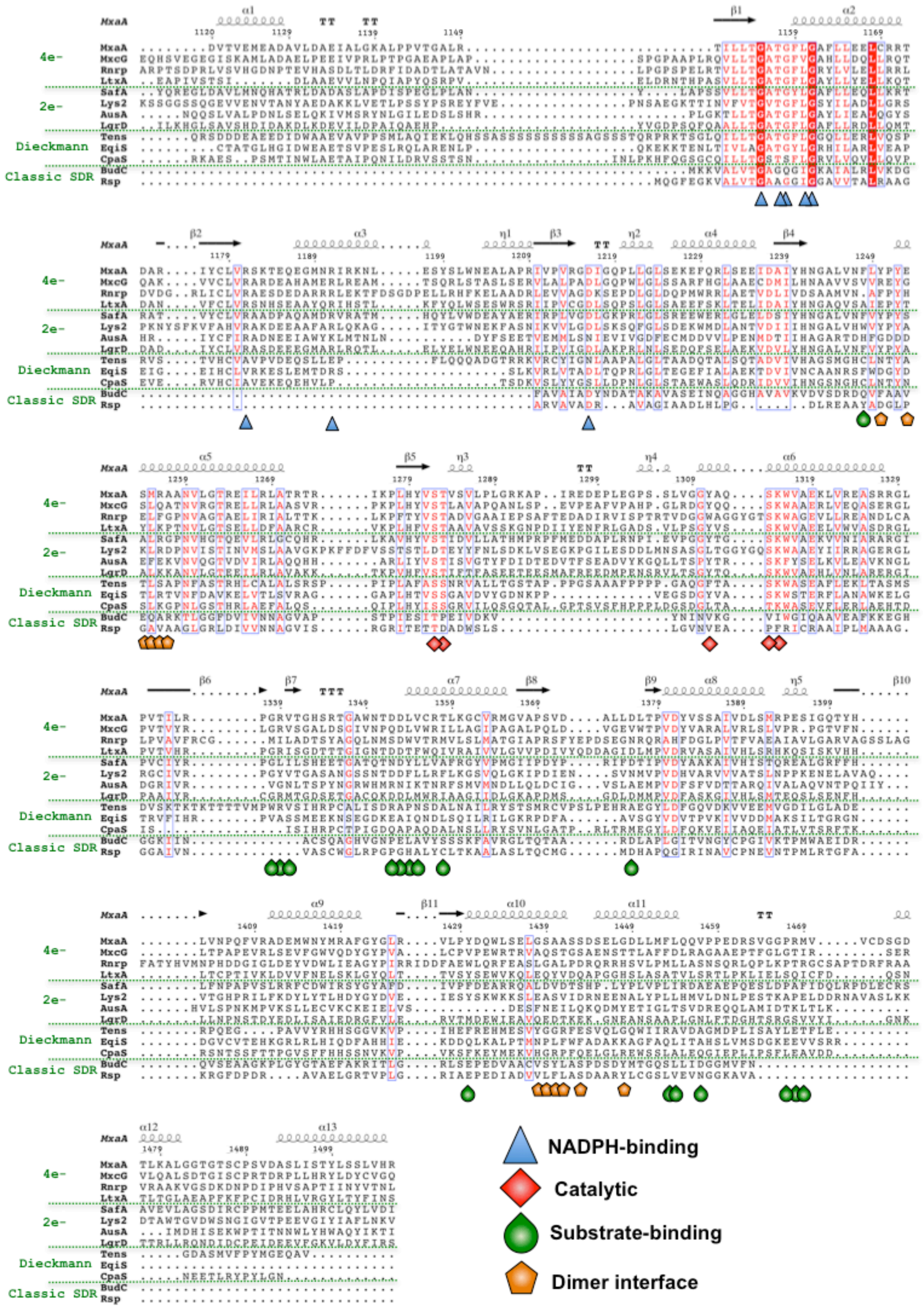


Figure 3-8. Multiple Sequence Alignment Analysis. Representation was generated using ESPript.

PDBePISA generated a Complex Formation Significance Score (CSS) of 0.00, suggesting that the surface interface displayed by the MxaA homodimer is a result of crystal packing. The average interface area between both monomers was calculated to 656.9 Å^2, which is 3.85 % of the total solvent accessible area. This constituted a total of 22 and 20 buried surface residues for monomer A and B respectively. It is also well known that biological interfaces tend to exhibit large areas, with the majority of cases exceeding 1000 Å^2.[33]

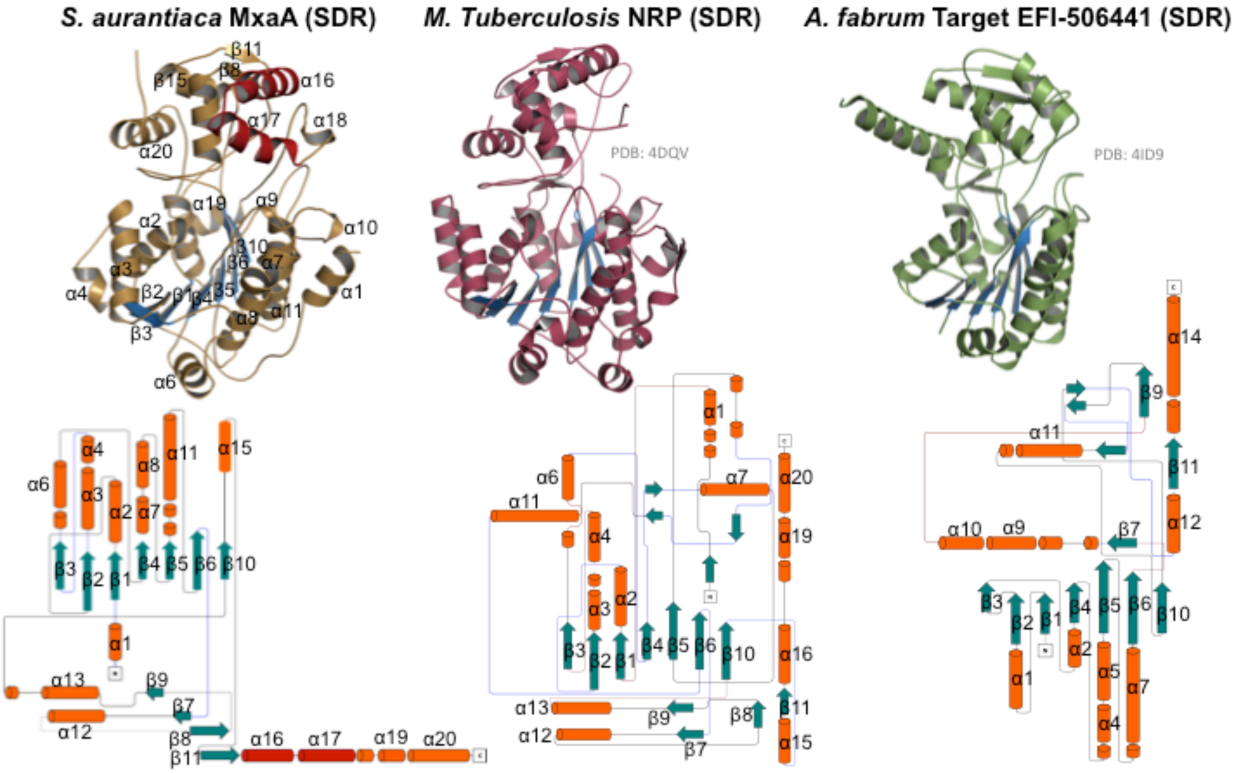

S. aurantiaca MxaA (SDR)
M. Tuberculosis NRP (SDR)
A. fabrum Target EFI-506441 (SDR)
PDB: 4DQV
PDB: 4ID9
β11 β8 β15 α16 α17 α18 α20 α19 α9 α2 α10 α3 β10 β6 α7 α4 β2 β1 β4 β5 α1 α8 α11 β3 α6
α6 α4 α8 α11 α15 α3 α2 α7 β3 β2 β1 β4 β5 β6 β10 α1 α13 β9 β7 α12 β8 β11 α16 α17 α19 α20
α1 α6 α7 α20 α11 α4 α19 α3 α2 β3 β2 β1 β4 β5 β6 β10 α16 α13 β9 β8 β11 α12 β7 α15
α14 β9 α11 β11 α10 α9 β7 α12 β3 β2 β1 β4 β5 β6 β10 α2 α1 α5 α7 α4

Figure 3-9. Topology analysis of the MxaA R, the Nrp R domain and other SDR family of reductases. Rossmann fold is colored blue.

Furthermore, evidence for its biological monomeric state was gathered from analytical size exclusion chromatography experiments, comparing the MxaA PCP-R didomain to known protein standards. Overall, these results suggest MxaA R exists in a biologically monomeric form rather than the crystallographically observed homodimeric state.

3.3.5 Structure analysis of NADPH-bound R domain

Currently, the Nrp terminal R domain (PDB: 4DQV) from *Mycobacteria smegmatis* involved in glycopeptide biosynthesis and the AusA R domain (PDB: 4F6C, 4F6L) from *Staphyloccoccus auresus* involved in pyrazinone biosynthesis are the sole PKS- or NRPS-associated R domains to have structures reported. While these monodomain structures have been solved with moderate resolution (2.30 Å for NRP and 2.81 Å for AusA), the lack of bound NADPH cofactor leaves key structural and mechanistic details rather unclear.[7, 34] In order to define residues required for cofactor binding in MxaA R, co-crystals of MxaA R complexed with NADPH were solved by molecular replacement of the apo structure to 1.90 Å (Table 3-1). NADPH binds to the well-known Rossmann fold that has a conserved nucleotide-binding motif TGxxGxxG, with the central diphosphate moiety hydrogen bonding to the peptide backbone of G1155, T1157, G1158, L1160, and G1161 (Fig. 3-2c). Further, the G1155 carbonyl forms a hydrogen bond with the adenosine 3'-hydroxyl group while the adenosine 2'-phosphate oxygen interacts with highly conserved T1157, R1181 and R1191. Both the 2'- and 3'-hydroxyl groups of the nicotinamide-containing ribose ring hydrogen bond with K1315 and Y1311. The nicotinamide amine hydrogen bonds with the G1338 carbonyl. Together, these interactions serve to tightly bind NADPH (K_d = 45 ± 3.7 μM, *vide supra*) and properly orient it in the active site for reduction of the PPant-bound intermediate to the terminal alcohol.

Several coordinated water molecules are present between the catalytic residues Y1311, T1283 and the non-catalytic S1285. One water molecule is positioned 2.7 Å from the hydroxyl of Y1311 and 2.8 Å from T1283, possibly occupying the oxyanion hole that these two residues create to assist in thioester and aldehyde reduction. T1283 and S1285 bind a second water molecule in the active site, although its positioning does not provide a clear role in catalysis. With respect to these observations, several SDR studies suggest that ordered water molecules in the active site might participate in a proton relay system involving the hydroxyl of Y1311, 2'-hydroxyl of the nicotinamide ribose and K1315.[35-36]

Structural comparison of the *apo* and NADPH-bound MxaA R domain show slight conformational changes with an overall RMSD of 0.63 Å. Out of these small differences, the C-terminal subdomain experiences a slightly higher conformational change upon NADPH binding, compared to the complete monomer, with a RMSD difference 0.73 Å.

3.3.6 MxaR, NRP R and AusA R structural comparison

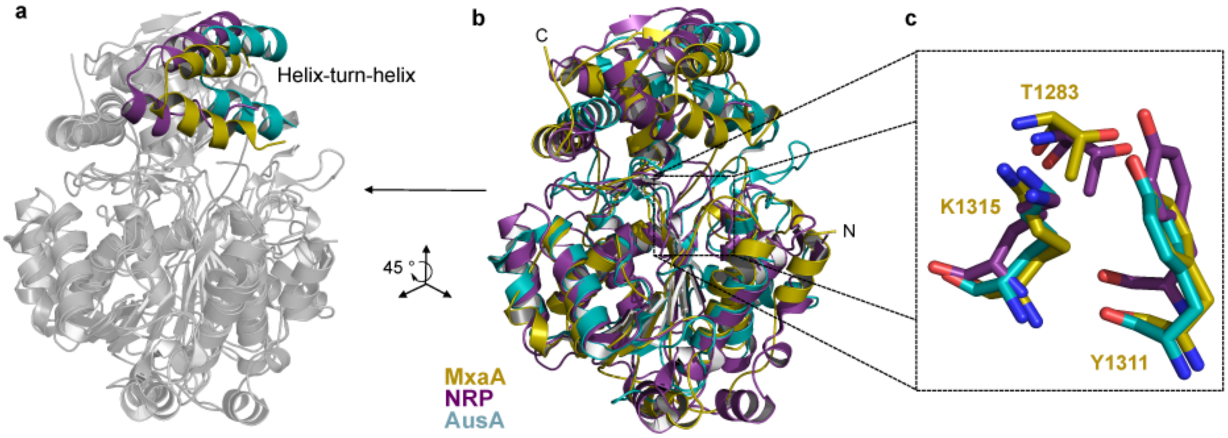


Figure 3-10. Structural comparison of three known, crystallized R domains: MxaA, Nrp and AusA. (a) All three R domain structure share a similar helix-turn-helix-motif at the C-terminal subdomain. (b) Structural alignment between MxaA, NRP and AusA. (c) The conserved catalytic T, Y, K triad.

The MxaA R domain from *Stigmatella aurantiaca* displays homology to the short chain dehydrogenase/reductase family of enzymes, which include the previously solved putative non-ribosomal peptide synthetase (NRP) R domain from *Mycobacterium tuberculosis*.[7] More specifically, both domains form part of the tyrosine-dependent class of oxidoreductase known for the common Thr, Tyr, Lys catalytic triad. Both MxaA, and NRP reductively release alcohol final products through a committed aldehyde intermediate, utilizing 2 NADPH molecules in the process. Mutants of the T-Y-K triad for both MxaA and NRP R domains significantly display lower affinity for the NADPH cofactor and are catalytically inactive. The MxaA shows 29.1 % sequence similarity with the NRP R domain respectively. Structural alignment between the MxaA and NRP R domains reveal a 2.19 Å RMSD of the alpha carbons through 279 residues. A close inspection of the T-Y catalytic residues in both MxaA and NRP display a highly coordinated water molecule, providing evidence for water displacement by the carboxylate group of the thioester moiety during the first reduction. This further supports the catalytic role of the T-Y residues of stabilizing the carboxylate through an oxyanion hole, typical of the SDR class of reductases.[24-26] Structural similarities include the NADPH binding Rossmann fold motif position at the N-terminal subdomain with the "helix-turn-helix" motif insertion. This includes the highly conserved "TGxxGxxG" motif for NADPH binding.

Substrate binding is proposed to occur on the C-terminal subdomain.[24-26] The C-terminal subdomain of the MxaA R domain is assembled by 5 helices and 2 parallel beta sheets. This feature is present in both the NRP R and the two electron AusA reductase from *Staphylococcus aureus* (PDB: 4F6L, 4F6C) (Fig. 3-10).[7, 34] All three reductase domains share a similar "helix-turn-helix" feature on the C-terminal subdomain. This HTH feature is observed in helix 16 and 17 of the MxaA R (Fig. 3-9). Docking analysis of both MxaA and NRP R domains with the decanal (for MxaA) or valeryl (NRP) suggest that the "helix-turn-helix" motif is important for substrate recognition and specificity.[7] This HTH motif in the NRP R domain, referred to as the helical insertion by Chhabra *et al.* , contain several hydrophobic residues that in conjunction with other non-polar residues form a large hydrophobic surface that could have a role in substrate binding.[7] Similarly, the HTH of the MxaA displays a comparable, yet distinct hydrophobic surface, which forms part of the myxalamid and decanal binding pocket in our *in silico* docking analysis. Such HTH structural features in the C-terminal subdomains of these currently known 2 and 4 e- reductases provide insights into the substrate specificity of terminal R domains.

3.3.7 Biochemical analysis of the NADPH and substrate-binding pocket

Moving beyond the structural and computational results pertaining to the reduction of **4** to **6**, and set in the context of advanced biofuel production **7** to **9** (Fig. 3-11), we aimed to provide biochemical support for residues involved in NADPH binding and the requirement of the Lys, Tyr, Thr catalytic triad. As the structural data revealed that NADPH contacts the amide backbone of G1155, G1158 and G1161, we generated single Gly to Ala mutants for each residue to examine the effect on NADPH binding capacity. Likewise, residues composing the putative catalytic triad were mutated one at a time (T1283A, K1315A and Y1311F) to similarly determine effects on both NADPH binding and catalytic activity. NADPH binding studies, as determined by intrinsic fluorescence measurements with the wild-type and mutant PCP-R didomains, clearly demonstrated disruption of NADPH binding by alanine mutation in the TGxxGxxG motif, as well as with mutants Y1311F and K1315A (Fig. 3-11, Table 3-2).[7, 37] Surprisingly, although not without precedent, the T1283A mutant possessed an approximate 50% increase in NADPH binding affinity ($K_{\mathrm{d(w.t.)}}$ = 45 ± 3.7 μM vs. $K_{\mathrm{d(T1283A)}}$ = 31.7 ± 1.6 μM).[7] This could be rationalized by the fact that T1283 has no direct contact with the cofactor and removal of its steric bulk likely improves access of NADPH to the binding site. These mutants were additionally investigated for activity toward the reduction of decanal. All mutants were catalytically inactive, including T1283A, which further verified requirement of the intact NADPH binding pocket and the complete catalytic triad for activity (Fig. 3-12).

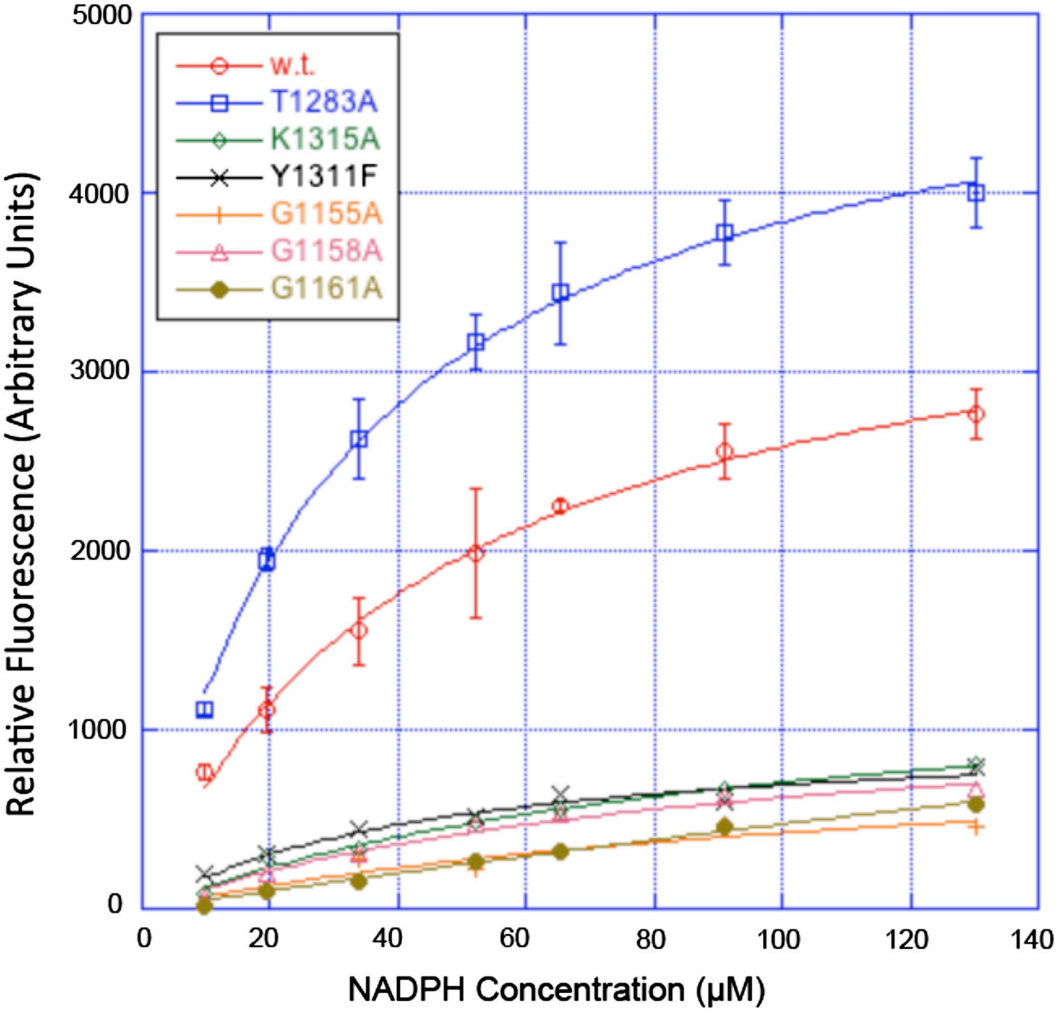


Figure 3-11. Determination of NADPH-Binding Constants for Wild-Type MxaA R and Select Mutants of NADPH-Interacting Residues

	w.t.	T1283A	K1315A	Y1311F	G1155A	G1158A	G1161A
Fl_{max}[a]	3750 ± 128	5059 ± 94	1408 ± 167	1007 ± 100	1063 ± 608	1183 ± 186	5087 ± 8387
K_d[b]	45 ± 3.7	31.7 ± 1.6	99.0 ± 21.1	45.4 ± 10.8	149 ± 133	89.9 ± 26.2	1125 ± 1767

[a] Fl_{max} is reported in arbitrary units; [b] K_d is reported in μM.

Table 3-2. Absolute values for maximum fluorescence (Fl_{max}) and K_d values for NADPH binding with MxaA and select mutants of the NADPH binding pocket. (Related to Figure 3-11)

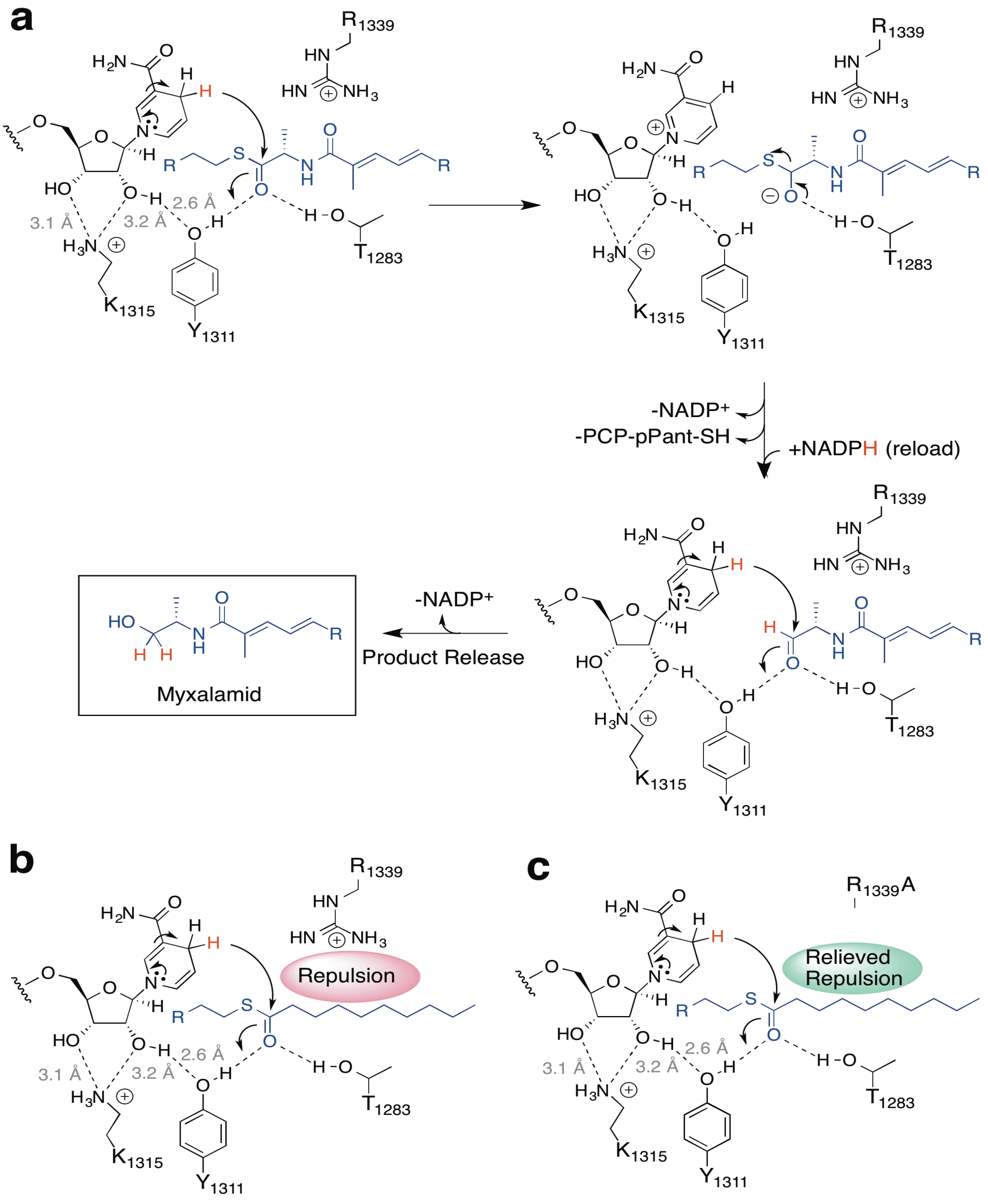


Figure 3-12. (a) Proposed two-step reduction of mxa. (b) Electrostatic interaction between decanoyl-PCP and R1339 and (c) relieved electrostatic interactions between R1339A and decanoyl-PCP.

Moving beyond characteristic analysis of the NADPH binding motif and catalytic residues, we sought to further investigate residues that may interact with the mxa substrate as determined through MD simulations and docking studies. We focused on five residues. The first, and perhaps most interesting, was R1339, which was determined by MD simulations to possess the highest probability of contact with the substrate – in particular the thioester-bound alanine moiety. Four additional residues with high probability of substrate interaction in the predominantly hydrophobic mxa-binding pocket were mutated to reverse their polarity or knockout key functional groups (F1248N, V1308T, Y1430F, R1468A). Because of the complexity of the substrates (PPant-mxa **4** and mxa aldehyde **5**) and concerns of their aqueous solubility, we opted to conduct the enzyme assay using simplified substrates: decanoyl-PCP and decanal. Moreover, as our ultimate goal is to use the information gained in these studies for the production of biologically derived replacement fuels and commodity chemicals, examination of the active site in the context of a target compound provides valuable knowledge to enable our desired goal. Studies additionally offer critical information pertaining to the mechanism of the R domain. Each mutant was assayed for the full reductive reaction and the second half reaction. Owing to the fact that aldehyde reduction is several orders of magnitude faster than PCP thioester reduction (Table 3-3), assays of this full reaction provide rates that are specific to the first half reaction. Therefore, we were able to obtain rates for both reductions: $k1$ and $k2$. Due to the fact that both decanoyl-CoA and decanoyl-loaded MxaA PCP monodomain were not turned over by MxaA R, we developed a single turnover assay by loading decanoyl-CoA to the PCP-R didomain with the promiscuous phosphopantethienyl transferase Sfp in order to obtain kinetic parameters for the first reduction.[38] Owing to low reactivity in the first-half reaction, multiple time points were taken within the first three hours without depleting the enzyme-substrate complex below 5% of the total concentration. This allowed the assays to be kept under pseudo-saturating (k_{cat}) conditions. While data concerning the first reduction provided turnover numbers that are significantly slower than those found with the second-half reaction, a clear dependence on residue identity was observed in the course of our studies. Moreover, as this system is a truncated portion of the complete MxaA module, changes in protein structure or substrate positioning—a parameter that may be altered by both protein truncation and use of a substrate lacking a PCP-bound amide bond—may have consequences that affect the upper limit of k_{cat}, but still clearly represent changes brought on by amino acid substitution. In contrast, observing NADPH consumption under saturating, multiple turnover conditions with the intermediate aldehyde **8** and yielded values similar to those obtained with the Nrp R domain (Table 3-3).

Table 3-3. Specific and Relative Activities for Wild-Type and Select Mutants with Respect to the First- and Second-Half Reactions

	Full Reaction		Second-Half Reaction	
	Enzyme Activity (pmol/min/mg MxaA)	Activity Relative to Wild-Type	Enzyme Activity (pmol/min/mg MxaA)	Activity Relative to Wild-Type
WT	3.69 ± 0.19	1.00	21.5 ± 1.7	1.00
F1248N	1.24 ± 0.03	0.34	27.2 ± 4.9	1.26
V1380T	1.86 ± 0.10	0.50	34.1 ± 0.7	1.58
R1339A	15.19 ± 0.46	4.11	134.41 ± 12.5	6.22
Y1430F	2.45 ± 1.56	0.66	37.4 ± 4.1	1.73
R1468A	1.10 ± 0.10	0.30	26.8 ± 3.1	1.24

With respect to the first reduction, we found mutations of the four residues that define the mxa-binding pocket to cause significant reductions in activity (Table 3-2). Mutation of residues closer to the NADPH binding site (F1248N, approximate 65% reduction in activity) caused a greater reduction in activity than those buried deeper in the pocket Y1430F and V1380T, (approximate 45% reduction in activity). This is likely due to reduced substrate-residue interactions, as indicated by our docking simulations with the non-native substrates. R1468, while buried deep in the binding pocket, still appeared to have an important role in the first-half reaction as demonstrated by the sharp reduction in activity with the R1468A mutant. Interestingly, for the second reduction, the same mutations moderately increased activity, compared to the wild-type. Aggregate results reveal a high probability that the first-half reaction is the rate-limiting step of this overall process and acutely sensitive to binding pocket mutations, while aldehyde reduction appears to be more robust. In our investigation, the second-half reaction turnover rate was actually improved by disruption of the binding pocket and active site entrance, supporting, that for the second half reduction, product release might be rate limiting.

In addition to exploring the mutational tolerance of the binding pocket we were interested to determine if R1339, as indicated by MD and docking simulations, interacted with the substrate. Computational data hinted to an electrostatic interaction between the R1339 guanidino group and the terminal alanine moiety contained within mxa. Therefore, we aimed to determine if R1339 in fact has an impact on catalysis. While kinetic analysis of substrates lacking a terminal alanyl thioester or alanal moiety, as found in mxa substrates, cannot definitively demonstrates the role of R1339 plays during catalysis with the native substrate, a comparison of the C_{10} substrate used in our studies with both w.t. and R1339A MxaA R provides a general understanding as to the nature of the residue-substrate interaction. Of significant importance, particularly in light of our goals to apply this enzyme in the production of fully reduced alcohols, we found that R1339A dramatically improved the ability of MxaA R to reduce C_{10} substrates with a 4.1- and 6.2-fold increase in activity for the first and second reduction, respectively. The large increases in activity can be rationalized by the fact that reduction of the thioester or aldehyde is guided by interactions between the PCP, R domain and PPant arm and R1339 appears to be poised to interact with incoming substrates (Fig. 3-5a,c). Both the first and second reductions with alternate substrates are improved by removal of the mismatched residue-substrate polarity (i.e., hydrocarbon-guanidino interaction) and, accordingly, are facilitated by an increase in the hydrophobicity of the active site tunnel (Fig. 3-12). In conclusion, the above biochemical findings support the combined crystal structure and computational data, and set the stage for future endeavors to further tune the active site to increase turnover of aliphatic substrates.

3.4 Materials and methods

3.4.1 Molecular dynamics

Molecular dynamics was carried out using Amber 14.[39] Both protein and ligand were prepared for docking by using the program Chimera.[17] Charges were calculated using the AMBER ff14SB force field. Selenomethionine (MSE) residues were converted to methionine (MET) residues, solvent was deleted and hydrogens were added. LEaP was used to neutralize the system by adding eight Na^+ ions, and solvating the apo enzyme in a 10 Å water buffer TIP3P truncated octahedron box. The fully solvated system contained 42,865 atoms. Minimization using SANDER was performed in two stages to remove any steric clashes present in the initial crystal structure. The initial stage was carried out over 2500 steps for the solvent and ions with the protein and cofactor restrained by a force constant of 500 kcal/mol/Å^2, followed by a second stage carried out over 5000 steps of the entire system. A short 20 ps simulation with weak restraints (force constant of 10 kcal/mol/Å^2 on the protein and cofactor) was used to heat up the system to a temperature of 300K using a langevin temperature equilibration scheme. Periodic boundary conditions were used, along with a non-bonded interaction cutoff of 10 Å. For the simulation, hydrogen atoms were constrained using the SHAKE algorithm, allowing for a 2 fs time step. The simulation was run over 100 ns (50,000,000 time steps). Simulation speeds of 4.0 ns/day were observed. A representative cluster ensemble was generated from MD using RMSD scoring as implemented in Chimera 1.9.[17] RMSD scoring reduced the initial set of 1000 frames generated to the 46 most unique frames. Molecular graphics and analyses were performed with the UCSF Chimera package. RMSD scoring was also used to calculate changes in the C-terminal and N-terminal domains. Highly mobile residues were identified in a similar approach.

3.4.2 *In silico* docking.

The docking program GOLD was used for docking between the MxaA R domain and the phosphopantetheine-tethered myxalamid intermediate.[16] Both protein and ligand were prepared for docking by removing waters, adding hydrogens, and converting the pdb files to Mol2 files using the program Chimera.[40] The MxaA R ligand-binding pocket was defined as residues within 20 Å of the hydrogen atom on the hydroxyl group of T1283. Docking was performed using the default settings with 100 docking trials performed. The docking solutions were ranked using the ChemPLP scoring functions. Molecular dynamic simulations generated 46 clusters with significant RMSD differences. A frame from each cluster was used to dock the phosphopantetheine-tethered myxalamid using the same docking parameters.

Prior to MxaA PCP-R domain docking, a PCP homology model was generated using the structure prediction HHpred [19]. The R domain monomer was docked with the PCP homology model using the protein-protein docking server Z-dock.[22] The ZDOCK 3.0/3.02 scoring function was used to identify the correct binding motif.[23]

3.4.3 Protein expression and purification

The recombinant wild-type and mutant MxaA R monodomains with an N-terminal His_{6x}-tag were expressed in BL21 (DE3) *E. coli* cells (Novagen). Cells containing the MxaA R domain plasmid were grown to OD_{600} = 0.6 at 37 °C in LB media containing 50 μg/mL kanamycin. The cell cultures were cooled to 18 °C and expression was induced using 0.5 mM IPTG. The cell cultures were incubated for an additional 16 hours at 18 °C and harvested by centrifugation at 5,525 r.c.f. for 15 minutes. The cell pellets were resuspended in 50mM tris-HCl pH 7.5, 10% glycerol, 10 mM imidazole, 300 mM NaCl and 1mg/ml lysozyme. Resuspended cells were cooled on ice for 30 min and the cells were disrupted using sonication. The cell debris was cleared by centrifugation at 21,036 r.c.f. for 1 hour. The supernatant was collected, and batch bound to HisPur™ Cobalt Resin (Thermo Scientific) for 1 hour at 4 °C. MxaA R was purified according to the manufacturer's instructions using an imidazole step-gradient. Fractions containing pure protein were determined by SDS-PAGE and fractions containing MxaA R were combined and dialyzed against 50mM tris-HCl pH 7.5, 10 % glycerol, 300 mM NaCl at 4 °C for 12 hours. Removal of the N-terminal His_6-tag was conducted by incubating the dialyzed MxaA R at 18 °C for 24 hours with thrombin from bovine plasma (Sigma-Aldrich, St. Louis MO) at a concentration of 2 U/mg of MxaA R protein and 3.5 mM $CaCl_{2.}$ Removal of thrombin and further purification of MxaA R was conducted by anion exchange chromatography using HiTrap Q FF (GE Healthcare) according to the manufacturer's instructions. Purified MxaA R was dialyzed against crystallization buffer, which consisted of 25 mM tris-HCl pH 7.5, 5 % glycerol, 1 mM dithiothreitol.

Selenomethionine-substituted (SeMet) MxaA R protein was produced in Bl21 (DE3) *E. coli* strain in M9 minimal medium using metabolic inhibition of the methionine biosynthetic pathway.[41] A 5 ml LB culture grown overnight was used to inoculate 2 X 1L of LB that were allowed to grow at 37 °C in the presence of 50 μg/mL kanamycin until OD_{600} = 0.6 was reached. The resulting cells were pelleted at 5,525 r.c.f. for 15 minutes and washed three times by suspension in 40ml of M9 medium and then transferred to 2 x 1L of M9 medium containing 50 μg/mL kanamycin and the following amino acids: lysine, phenylalanine and threonine (100 mg/L); isoleucine, leucine and valine (50 mg/L); and L-selenomethionine (40 mg/L) (sigma). The temperature was reduced to 18 °C, induced with 0.5 mM IPTG and was allowed to grow overnight for 16 hours. The cells were harvested and purified following the w.t. procedure. The incorporation of selenomethionine (10 residues total) was confirmed by MALDI-TOF mass spectrometry.

3.4.4 Crystallization, data processing, refinement and analysis.

Both native and SeMet crystals of the w.t. MxaA R domain (9 mg/ml) grew in 0.22 M ammonium acetate, 28 % PEG 3350 and 0.1 M Hepes pH 7.7 overnight at 25 °C using the hanging drop vapor diffusion method. NADPH bound MxaA R crystals formed similarly with the exception of incubating NADPH and MxaA R at a 5:1 molar ratio for 1 hr at 4 °C prior to crystal tray set up. Crystals were cryoprotected in well solution and flash frozen in liquid nitrogen prior to data collection. Data was collected at beamline 12-2 at the Stanford

Synchrotron Radiation Lightsource (SSRL) for SeMet crystals. Prior to data collection, initial frames were assessed for quality and redundancy using Mosflm and Web-ice [42-43]. Multiwavelength anomalous diffraction (MAD) data were collected to 1.70 Å for SeMet MxaA R at λ = 0.9792 Å (Selenium peak), λ = 0.9611 Å (inflection), λ = 0.9794 Å (remote). For MAD data collection, the exposure time was set to 0.2 s; 0.15° oscillation width for 1920 frames. All data was processed using Mosflm to the P21 space group (2). Native NADPH bound MxaA R data was collected at the Advance Light Source beamline 822 at the Lawrence Berkeley National Laboratory. Single monochromatic x-ray diffraction data (λ = 0.9775 Å, 700 frames at 0.5° oscillation width for 1 second exposure) was collected to 1.84 Å and processed with Mosflm using the $P2_1$ space group (2). Initial phases for MAD data set were obtained using PHENIX Autosol (site) and 9 out of the 10 heavy-atom derivatives were located. Initial model was constructed using PHENIX Autobuild. Refinement was done using PHENIX.REFINE and COOT.[44-45] Improved phases were used in COOT to model missing side residues manually and waters were added during the last refinement cycles. For the NADPH co-crystal structure, PHENIX LigandFit was used to model the NADPH upon obtaining initial model and phases from Phenix Autosol.[44] Both apo and NADPH bound structures were validated using PROCHECK and PDB_REDO.[46-47] Structural analysis such as structural superimposition, electrostatic potentials and figure generation used in the manuscript were made using PyMol.[48]

3.4.5 Circular dichroism (CD).

All samples, both mutant and w.t., were prepared by diluting protein to 0.2 mg/ml in 20mM Tris-HCl (pH 7.5). The CD data was collected using a Jasco J810 CD spectropolarimeter. Spectral scans were collected at 20 °C from 190 to 260 nm using 0.5nm steps with 5 repeats.

3.4.6 NADPH Consumption Time course.

Consumption of NADPH by MxaA, or variants thereof, was measured by decrease in absorption at 340 nm. 5.0 μM of the MxaA R was incubated in 100 mM potassium phosphate (pH = 7.0) buffer containing 200 μM NADPH and 1.0 mM decanal, in order to keep the substrate near saturating conditions. Measurements were recorded in triplicate and averaged, spontaneous NADPH degradation was accounted for in a control reaction lacking enzyme.

3.4.7 Fluorescence Titration of MxaA R and Mutants.

Assays were prepared by adding NADPH (1 mM stock solution, 1-130 μM final concentration) to 10 μM MxaA or mutant R in buffer containing 100 mM phosphate and 300 mM NaCl at pH 7.25. Fluorescence was measured on a Teacan Safire fluorometer (λ_{ex} = 340 nm, λ_{em} = 460 nm with excitation and emission slits set to 7.5 nm) and the relative increase in fluorescence was measured by subtracting autofluorescence of NADPH samples in the absence of enzyme from those interacting with the reductase domain. Plotting these data and fitting to the Michaelis-Menten equation determined the K_d and relative maximum fluorescence.

3.4.8 Determination of Enzyme Specific Activities with Decanal.

MxaA (w.t. or mutant) (20 μM final concentration) was added with NADPH (250 μM) and decanal (2 mM, saturating) to the reaction buffer (150 mM sodium phosphate, 200 mM

NaCl) at a total volume of 200 μL. These reactions were monitored at 340 nM for the depletion of NADPH over six minutes, corrected for background NADPH consumption and the resultant slope was used to calculate the specific activity. Conversion to decanol was verified by GCMS.

3.4.9 Single turnover assay for R domain reduction.

MxaA (w.t. or mutant) (50 μM final concentration) was combined with decanoyl-CoA (200 μM), $MgCl_2$ (10 mM), Sfp phosphopantetheinyl transferase (10 μM) in the reaction buffer (150 mM sodium phosphate, 200 mM NaCl) at a total volume of 300 μL.[38] Sfp-mediated PPant loading proceeded for 2 h at which point the extent of loading was determined by LC-MS/MS to provide 19 μM of decanoyl-loaded MxaA-PCP.[49] The reaction was initiated with NADPH (250 μM). Control reactions showed no reduction of decanoyl-CoA in the absence of being loaded to the MxaA PCP. Reactions, done in duplicate, were stopped at 1h, 2h and 3h with the addition of 30 μL 10 % (v/v) acetic acid and extracted with 2 x 300 μL hexanes containing an internal standard of 100 μM dodecanol. Combined extracts were concentrated ~10-fold and mixed with an equal volume of *N,O*-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) and analyzed on a Hewlett Packard 6890 series GC fitted with an Agilent 5973Network mass detector with a 30m x 0.25mm DB-5MS column (Agilent). Samples were injected at 80 °C, held at that temperature for 2.0 min and then ramped to 300 °C at 25 °C/min held at 300 °C for 1.0 min and returned to the initial temperature. Samples were compared to an authentic decanol standard curve and normalized to internal dodecanol.

3.4.10 Analytical Size Exclusion Chromatography.

Six milligrams of purified MxaA PCP-R didomain suspended in 500 µL of 50 mM sodium phosphate buffer pH 7.5 and 100 mM sodium chloride buffer was injected into a Superdex 200 10/300 GL column (GE Healthcare Life Sciences) using an Äkta explorer FPLC system (GE Healthcare Life Sciences). The sample was allowed to run over 1.3 column volumes at a flow rate of 0.5 mL/min using filtered 50 mM sodium phosphate buffer pH 7.5 and 100 mM sodium chloride buffer. MxaA PCP-R didomain sample was monitored at 280 nm. The molecular weight and multimeric state of the MxaA PCP-R didomain was assessed and compared to low and high molecular weight gel filtration calibration protein standards (GE Healthcare Life Sciences).

3.4.11 Cloning.

The MxaA PCP and R monodomains were amplified using the myxobacterium *S. aurantiaca* PCP-R didomain previously cloned into pET28a with *NdeI* tagged 5’ forward primer and the *HindIII* tagged 3’ flanking primer. The subsequent amplified product was inserted into the corresponding sites in pET28b (Novagen, Madison, WI) generating N-terminal thrombin cleavable His_{6x}-tagged constructs. The constructs were transformed into Rosetta Blue^{TM} (Millipore) Nova Blue competent cells for construct amplification. Sequence was confirmed through automated DNA sequencing.

Cloning Primer	Sequence
MxaA PCP Forward:	5’- ATATATCATATGCGCGCTGCTCTGCCG
MxaA PCP Reverse:	5’- ATATATAAGCTTTTAATCATGCGCCGGCAGAGAGC
MxaA R Forward:	5’- ATATATCATATGTCTCTGCCGGCGCATGATGT

MxaA R Reverse: 5’- ACATATATAAGCTTTTATTCTGGAGCCTTCAGGAAGCCAC

3.4.11 Site-directed-mutagenesis.

The online program PrimerX was used to design primers for all the R mutants using the previously described cloned pET28b w.t. R DNA template [50].The following primers were used for mutagenesis. All mutations were confirmed though automated DNA sequencing.

Mutant Primer		Sequence
MxaA R Y1311F	Forward:	5’- GCGGCTTTGCTCAGAGTAAATGGG
MxaA R Y1311F	Reverse:	5’- CTCTGAGCAAAGCCGCCCACC
MxaA R K1315A	Forward:	5’- GAGTGCGTGGGTCGCGGAAAAGCTGG
MxaA R K1315A	Reverse:	5’- GACCCACGCACTCTGAGCATAGCCGC
MxaA R T1283A	Forward:	5’- GTTAGCGCGGTCTCTGTGCTGCCGC
MxaA R T1283A	Reverse:	5’- CAGAGACCGCGCTAACATAGTGCAGCGG
MxaA R G1155A	Forward:	5’- ACCGCGGCTACGGGTTTTCTGGGC
MxaA R G1155A	Reverse:	5’- CCCGTAGCCGCGGTCAGCAGG
MxaA R G1158A	Forward:	5’- ACGGCGTTTCTGGGCGCGTTCCT
MxaA R G1158A	Reverse:	5’- CCCAGAAACGCCGTAGCACCGGTC
MxaA R G1161A	Forward:	5’- TTCTGGCGGCGTTCCTGCTGGAAG
MxaA R G1161A	Reverse:	5’- GGAACGCCGCCAGAAAACCCGTAG
MxaA R F1248N	Forward:	5’- TGGTCAATAATCTGTATCCGTACGAAAGC
MxaA R F1248N	Reverse:	5’- CGGATACAGATTATTGACCAGTGCACC
MxaA R V1308T	Forward:	5’- AGCCTGACGGGCGGCTATGCTCAGAG
MxaA R V1308T	Reverse:	5’- CGCCCGTCAGGCTGCTCGGAC

MxaA R R1339A	Forward:	5’- GGGTGCGGTGACCGGTCATTCACGC
MxaA R R1339A	Reverse:	5’- GGTCACCGCACCCGGACGCAGG
MxaA R Y1430F	Forward:	5’- CCGTTTGACCAGTGGCTGAGC
MxaA R Y1430F	Reverse:	5’- CACTGGTCAAACGGCAGAACG C
MxaA R R1468A	Reverse:	5’- CGGTCCGGCGATGGTGGTTTGCG
MxaA R R1468A	Forward:	5’- GCAACCGCCAGGCCGCTACCAC

3.4.12 Determining of the Extent of PCP Loading by Sfp.

Apo-MxaA(PCP-R) was purified in BL21(DE3) as reported above. Loading of acylated CoA (decanoyl CoA) was performed as described in the main text. Following the reaction of MxaA(PCP-R) (50 μM) with decanoyl CoA (200 μM) and Sfp (10 μM) the reaction was incubated with 10% (w/w) trypsin and allowed to digest for 2h at ambient temperature. Reactions lacking Sfp and decanoyl CoA were performed as controls.

Samples were analyzed on an Agilent 4640 Triple Quad LC/MS (Santa Clara, CA) mass spectrometer operating in SRM mode coupled to an Agilent 1260 Infinity HPLC (Santa Clara, CA). 5 μg of peptide was injected on a Supelco Acentis Express Peptide ES_C18 column (5cm x 2.1mm x 2.7 μm) (Sigma-Aldrich, St. Louis, MO) column with the following gradient applied: Buffer A = H_2O + 0.1% trifluoroacetic acid; Buffer B = 98% acetonitrile + 0.1 % trifluoroacetic acid; t = 0 min 95% A; t = 0.2 min 95% A ramped to 65% A at 5.7 min; 5.7 min to 6.0 min ramp to 10% A; 6.0 min to 8.0 min hold at 10% A; 8.0 min to 8.5 min ramp down to 5.0% A and hold to 11.0 min. The flow rate was held constant at 0.4 mL/min. The following parent (p) and daughter (d) ions were monitored to determine the relative amounts of apo, holo and C_{10} acylated PCP: apo (p: 1084.8685, d: 175.1190); holo (p:1084.8686, d: 261.1267) and decanoyl (p:1136.2472, d: 415.2625) with z=3. Data was analyzed using Skyline.

It was determined that unreacted apo MxaA(PCP-R) contained ~5% holo protein with the remainder as the apo form. Reaction with Sfp and no decanoyl CoA provided significantly increased amounts of the holo form (~2:1 holo:apo). This is presumably from co-purification of coenzyme A, despite dialysis of the Sfp protein itself. Similarly, reaction of apo MxaA(PCP-R) with Sfp and decanoyl CoA provided an approximate 2:1 ratio of holo to decanoyl phosphopantetheine. The final concentration of active MxaA (enzyme substrate complex) was adjusted to reflect these data.

3.5 Conclusions

Products generated by PKSs and NRPSs require release from PPant-tethered carrier proteins contained in mega-synthases. Both thioesterase and R domains mediate chain release to provide distinct terminal functional groups to enrich the chemical diversity of polyketide and non-ribosomal peptide natural products.[4] R domains are an NADPH dependent class of SDR-like enzymes capable of reductively releasing acyl and peptide intermediates from the PPant-tethered carrier protein. Prior to this study, no co-factor bound structure was available for modular enzyme-associated terminal R domains. Here, we report the crystal structure, with significantly increased resolution, of the myxalamid PKS-NRPS terminal R domain that catalyzes the non-processive four-electron reduction of **4** to **6** and decanoyl-PCP to 1-decanol. Computational MD and biochemical analysis support assertions that the C-terminal subdomain of the R domain is the most flexible region, responsible for substrate binding and selectivity. With respect to kinetic parameters the first reduction of decanoyl- PCP to yield the decanal intermediate is significantly slower than the second reduction of decanal to 1-decanol, thus providing insight to the rate-limiting step during R domain mediated product release. Structure-based mutations helped determine residues important for substrate binding and reduction. Furthermore, mutational analysis of the putative gatekeeping residue (R1339) improved reduction of both decanoyl-PCP and decanal. Combined, the mechanistic insights gained by our comprehensive investigation of MxaA R provide not only a deeper understanding of structural and catalytic features required for activity but set a foundation for future engineering efforts using modular catalyst associated R domains. Efforts in combining R domains with novel PKS- or NRPS-based assembly lines could produce alternate substrates that, for example, can be screened for new bioactivity or used in the production of biologically derived commodity chemicals.

3.6 Contributions and Acknowledgements

My contributions to this project included the majority of the computational work. The computational work included *in silico* docking and molecular dynamics, which provided the basis for biochemical investigations in this study. Dr. Jesus Barajas performed the X-ray crystallographic studies, with structural verification performed by Dr. David Jackson. Dr. Ryan Phelan performed the biochemical assays.

References

1. Silakowski, B.; Nordsiek, G.; Kunze, B.; Blocker, H.; Muller, R., Novel features in a combined polyketide synthase/non-ribosomal peptide synthetase: the myxalamid biosynthetic gene cluster of the myxobacterium Stigmatella aurantiaca Sga15. *Chem Biol* **2001,** *8* (1), 59-69.
2. Gerth, K.; Jansen, R.; Reifenstahl, G.; Hofle, G.; Irschik, H.; Kunze, B.; Reichenbach, H.; Thierbach, G., The myxalamids, new antibiotics from Myxococcus xanthus (Myxobacterales).

I. Production, physico-chemical and biological properties, and mechanism of action. *J Antibiot (Tokyo)* **1983,** *36* (9), 1150-6.

3. Gaitatzis, N.; Kunze, B.; Muller, R., In vitro reconstitution of the myxochelin biosynthetic machinery of Stigmatella aurantiaca Sg a15: Biochemical characterization of a reductive release mechanism from nonribosomal peptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (20), 11136-41.

4. Du, L.; Lou, L., PKS and NRPS release mechanisms. *Nat Prod Rep* **2010,** *27* (2), 255-78.

5. Silakowski, B.; Kunze, B.; Muller, R., Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium Stigmatella aurantiaca. *Gene* **2001,** *275* (2), 233-40.

6. Konz, D.; Marahiel, M. A., How do peptide synthetases generate structural diversity? *Chemistry & biology* **1999,** *6* (2), R39-48.

7. Chhabra, A.; Haque, A. S.; Pal, R. K.; Goyal, A.; Rai, R.; Joshi, S.; Panjikar, S.; Pasha, S.; Sankaranarayanan, R.; Gokhale, R. S., Nonprocessive [2 + 2]e- off-loading reductase domains from mycobacterial nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A* **2012,** *109* (15), 5681-6.

8. Rao, S. T.; Rossmann, M. G., Comparison of super-secondary structures in proteins. *J Mol Biol* **1973,** *76* (2), 241-56.

9. Kleiger, G.; Eisenberg, D., GXXXG and GXXXA Motifs Stabilize FAD and NAD(P)-binding Rossmann Folds Through Cα–H⋯O Hydrogen Bonds and van der Waals Interactions. *Journal of Molecular Biology* **2002,** *323* (1), 69-76.

10. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-74.

11. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics & modelling* **2006,** *25* (2), 247-60.

12. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of chemical theory and computation* **2012,** *8* (5), 1542-1555.

13. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.

14. Wickstrom, L.; Okur, A.; Simmerling, C., Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophysical journal* **2009,** *97* (3), 853-6.

15. David A. Case; Tom Darden; Thomas E. Cheatham III , C. S., Adrian Roitberg, ; Junmei Wang; Robert E. Duke; Ray Luo; Daniel R. Roe; Ross C. Walker; Scott LeGrand ; Jason Swails ; David Cerutti ; Joe Kaus; Robin Betz; Romain M. Wolf ; Kenneth M. Merz ; Gustavo Seabra; Pawel Janowski ; Andreas W. Götz ; István Kolossváry ; Francesco Paesani , J. L.; Xiongwu Wu ; Thomas Steinbrecher ; Holger Gohlke ; Nadine Homeyer; Qin Cai; Wes Smith; Dave Mathews; Romelia Salomon-Ferrer; Celeste Sagui; Volodymyr Babin ; Tyler Luchko; Sergey Gusarov ; Andriy Kovalenko ; Josh Berryman ; Peter A. Kollman Amber 14 Reference Manual. University of California, S. F., Ed. 2014.

16. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins* **2003,** *52* (4), 609-23.

17. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-12.

18. Liebeschuetz, J. W.; Cole, J. C.; Korb, O., Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *Journal of computer-aided molecular design* **2012,** *26* (6), 737-48.

19. Hildebrand, A.; Remmert, M.; Biegert, A.; Soding, J., Fast and accurate automatic structure prediction with HHpred. *Proteins* **2009,** *77 Suppl 9*, 128-32.

20. Wallace, A. C.; Laskowski, R. A.; Thornton, J. M., LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering* **1995,** *8* (2), 127-34.

21. Tanovic, A.; Samel, S. A.; Essen, L. O.; Marahiel, M. A., Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science* **2008,** *321* (5889), 659-63.

22. Pierce, B. G.; Hourai, Y.; Weng, Z., Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one* **2011,** *6* (9), e24657.

23. Chen, R.; Li, L.; Weng, Z., ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **2003,** *52* (1), 80-7.

24. Jornvall, H.; Persson, B.; Krook, M.; Atrian, S.; Gonzalez-Duarte, R.; Jeffery, J.; Ghosh, D., Short-chain dehydrogenases/reductases (SDR). *Biochemistry* **1995,** *34* (18), 6003-13.

25. Kallberg, Y.; Oppermann, U.; Persson, B., Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models. *The FEBS journal* **2010,** *277* (10), 2375-86.

26. Kavanagh, K. L.; Jornvall, H.; Persson, B.; Oppermann, U., Medium- and short-chain dehydrogenase/reductase gene and protein families : the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cellular and molecular life sciences : CMLS* **2008,** *65* (24), 3895-906.

27. Bergmann, S.; Schumann, J.; Scherlach, K.; Lange, C.; Brakhage, A. A.; Hertweck, C., Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. *Nature chemical biology* **2007,** *3* (4), 213-217.

28. Gomez-Escribano, J. P.; Song, L. J.; Fox, D. J.; Yeo, V.; Bibb, M. J.; Challis, G. L., Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the cpk gene cluster of Streptomyces coelicolor M145. *Chem Sci* **2012,** *3* (9), 2716-2720.

29. Masschelein, J.; Clauwers, C.; Awodi, U. R.; Stalmans, K.; Vermaelen, W.; Lescrinier, E.; Aertsen, A.; Michiels, C.; Challis, G. L.; Lavigne, R., A combination of polyunsaturated fatty acid, nonribosomal peptide and polyketide biosynthetic machinery is used to assemble the zeamine antibiotics. *Chem Sci* **2015,** *6* (2), 923-929.

30. Li, Y.; Weissman, K. J.; Muller, R., Myxochelin biosynthesis: direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. *Journal of the American Chemical Society* **2008,** *130* (24), 7554-5.

31. Duarte, J. M.; Srebniak, A.; Scharer, M. A.; Capitani, G., Protein interface classification by evolutionary analysis. *BMC bioinformatics* **2012,** *13*, 334.

32. Krissinel, E.; Henrick, K., Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **2007,** *372* (3), 774-97.

33. Jones, S.; Thornton, J. M., Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **1996,** *93* (1), 13-20.

34. Wyatt, M. A.; Mok, M. C.; Junop, M.; Magarvey, N. A., Heterologous expression and structural characterisation of a pyrazinone natural product assembly line. *Chembiochem* **2012,** *13* (16), 2408-15.

35. Eklund, H.; Plapp, B. V.; Samama, J. P.; Branden, C. I., Binding of substrate in a ternary complex of horse liver alcohol dehydrogenase. *The Journal of biological chemistry* **1982,** *257* (23), 14349-58.

36. Oppermann, U.; Filling, C.; Hult, M.; Shafqat, N.; Wu, X.; Lindh, M.; Shafqat, J.; Nordling, E.; Kallberg, Y.; Persson, B.; Jornvall, H., Short-chain dehydrogenases/reductases (SDR): the 2002 update. *Chemico-biological interactions* **2003,** *143-144*, 247-53.

37. Wilson, R. A.; Gibson, R. P.; Quispe, C. F.; Littlechild, J. A.; Talbot, N. J., An NADPH-dependent genetic switch regulates plant infection by the rice blast fungus. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** *107* (50), 21902-7.

38. Quadri, L. E.; Weinreb, P. H.; Lei, M.; Nakano, M. M.; Zuber, P.; Walsh, C. T., Characterization of Sfp, a Bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **1998,** *37* (6), 1585-95.

39. D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman, AMBER 14. University of California, San Francisco: 2014.

40. Huang, C. C., Couch, G.S., Pettersen, E.F., and Ferrin, T.E, Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components. *Pacific Symposium on Biocomputing* **1996,** *1*, 724.

41. Van Duyne, G. D.; Standaert, R. F.; Karplus, P. A.; Schreiber, S. L.; Clardy, J., Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J Mol Biol* **1993,** *229* (1), 105-24.

42. Powell, A. G. W. L. a. H. R., Evolving Methods for Macromolecular Crystallography. Randy J. Read, J. L. S., Ed. Springer: 2007. (accessed 06/12/14).

43. A. González, P. M., S. E. McPhillips, J. Song, K. Sharp, J. R. Taylor, P. D. Adams, N. K. Sauter and S. M. Soltis, Web-Ice: integrated data collection and analysis for macromolecular crystallography. *Journal of Applied Crystallography* **2008,** *41*, 176-184.

44. P. D. Adams, P. V. A., G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D* **2010,** *66* (2), 8.

45. Emsley, P.; Debreczeni, J. E., The use of molecular graphics in structure-based drug design. *Methods Mol Biol* **2012,** *841*, 143-59.

46. R. A. Laskowski, M. W. M., D. S. Moss and J. M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **1993,** *26* (2), 8.

47. Joosten, R. P.; Joosten, K.; Murshudov, G. N.; Perrakis, A., PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D Biol Crystallogr* **2012,** *68* (Pt 4), 484-96.

48. Schrödinger, L. *The PyMOL Molecular Graphics System*, 1.5.0.4; 2013.

49. Poust, S.; Phelan, R. M.; Deng, K.; Katz, L.; Petzold, C. J.; Keasling, J. D., Divergent mechanistic routes for the formation of gem-dimethyl groups in the biosynthesis of complex polyketides. *Angewandte Chemie* **2015,** *54* (8), 2370-3.

50. Lapid, C. PrimerX http://www.bioinformatics.org/primerx/.

51. Sherman, D. H., The Lego-ization of polyketide biosynthesis. *Nature biotechnology* **2005,** *23* (9), 1083-4.

52. Menzella, H. G.; Reid, R.; Carney, J. R.; Chandran, S. S.; Reisinger, S. J.; Patel, K. G.; Hopwood, D. A.; Santi, D. V., Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nature biotechnology* **2005,** *23* (9), 1171-6.

53. Hahn, M.; Stachelhaus, T., Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (2), 275-80.

54. Poust, S.; Hagen, A.; Katz, L.; Keasling, J. D., Narrowing the gap between the promise and reality of polyketide synthases as a synthetic biology platform. *Current opinion in biotechnology* **2014,** *30*, 32-9.

55. Weissman, K. J.; Leadlay, P. F., Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol* **2005,** *3* (12), 925-936.

56. Atsumi, S.; Hanai, T.; Liao, J. C., Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **2008,** *451* (7174), 86-9.

57. Peralta-Yahya, P. P.; Zhang, F.; del Cardayre, S. B.; Keasling, J. D., Microbial engineering for the production of advanced biofuels. *Nature* **2012,** *488* (7411), 320-8.

58. Chu, S.; Majumdar, A., Opportunities and challenges for a sustainable energy future. *Nature* **2012,** *488* (7411), 294-303.

59. Gouet, P.; Robert, X.; Courcelle, E., ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic acids research* **2003,** *31* (13), 3320-3.

60. Robert, X.; Gouet, P., Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research* **2014,** *42* (Web Server issue), W320-4.

CHAPTER 4

Structural Characterization of the Crucial Interaction Between the Acyl Carrier Protein and a Ketosynthase

4.1 Summary

Fatty acid synthases are dynamic ensembles of enzymes that can efficiently biosynthesize long hydrocarbon chains. Here we visualize the interaction between the *E. coli* acyl carrier protein (AcpP) and β-ketoacyl-ACP-synthase I (FabB) using X-ray crystallography, NMR, and MD simulations. We leveraged this structural information to alter lipid profiles *in vivo* and provide a molecular basis for how protein-protein interactions can regulate the fatty acid profile in *E. coli*.

4.2 Introduction

The *E. coli* fatty acid synthase (FAS) produces metabolites that comprise the cellular membrane through an iterative cycle via the activities of 13 discrete proteins that yield both saturated and unsaturated products, with each enzyme carrying out a single, simple transformation.[1-2] These crucial enzymes are targets for antibiotic drug discovery, and the ability of this pathway to efficiently form carbon-carbon bonds and generate long hydrocarbon chains has made it a target for biofuel development.[3-4] However, the rate limiting steps in this pathway remains poorly understood, and attempts to understand the fundamental phenomena of fatty acid metabolism have been hindered by a lack of information about the molecular interactions of the pathway enzymes.[5-7]

The acyl carrier protein, AcpP, at the center of *E. coli* fatty acid biosynthesis, shuttles cargo from one enzyme partner to another through iterative biosynthetic cycles (Fig. 4-1). As a small, dynamic, monomeric helical bundle, AcpP provides protection of the growing fatty acid from non-selective reactivity in the cytosol by sequestration within a central hydrophobic core.[8-10] Cargo is covalently tethered to AcpP via thioester linkage to the end of a post-translationally added phosphopantetheine (PPant) arm.[11] During interaction with partner proteins, the cargo translocates from the hydrophobic core of AcpP completely into the partner, a process called "chain flipping".[12-13]

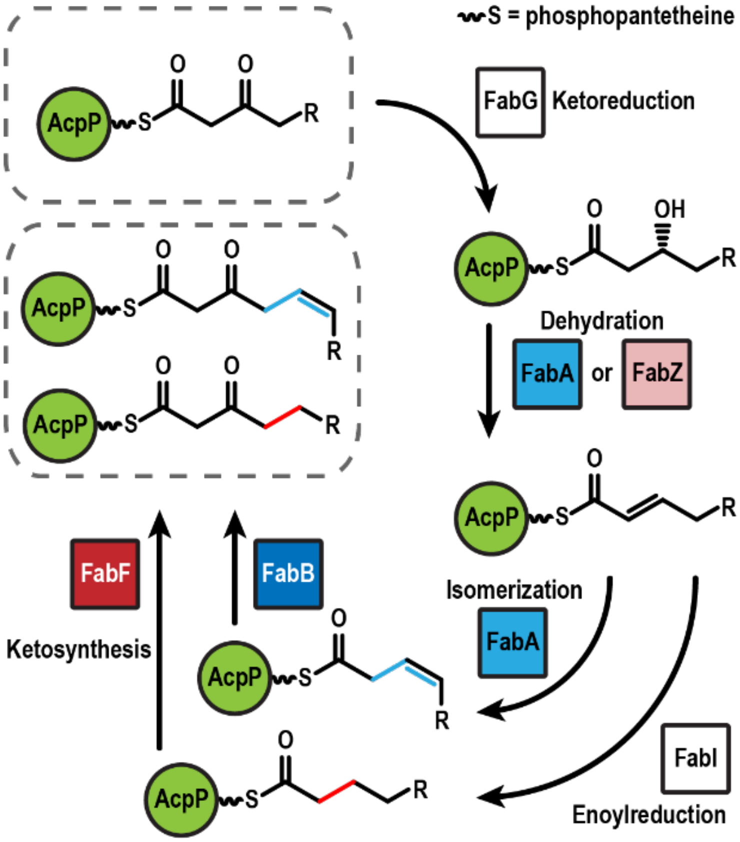


Figure 4-1. Biosynthesis of saturated and unsaturated fatty acids in *E. coli*. FabA (the dehydratase) and FabB (the ketosynthase), in blue and working in tandem, produce and elongate unsaturated fatty acids. FabZ (the second dehydratase) and FabF (the second ketosynthase), in red, working in tandem with FabG (the ketoreductase) and FabI (the enoylreductase), produce and elongate saturated fatty acids. These cycles are iterated to produce full-length fatty acid chains. R = acyl, varying in length based on how many upstream iterations of the cycle have occurred.

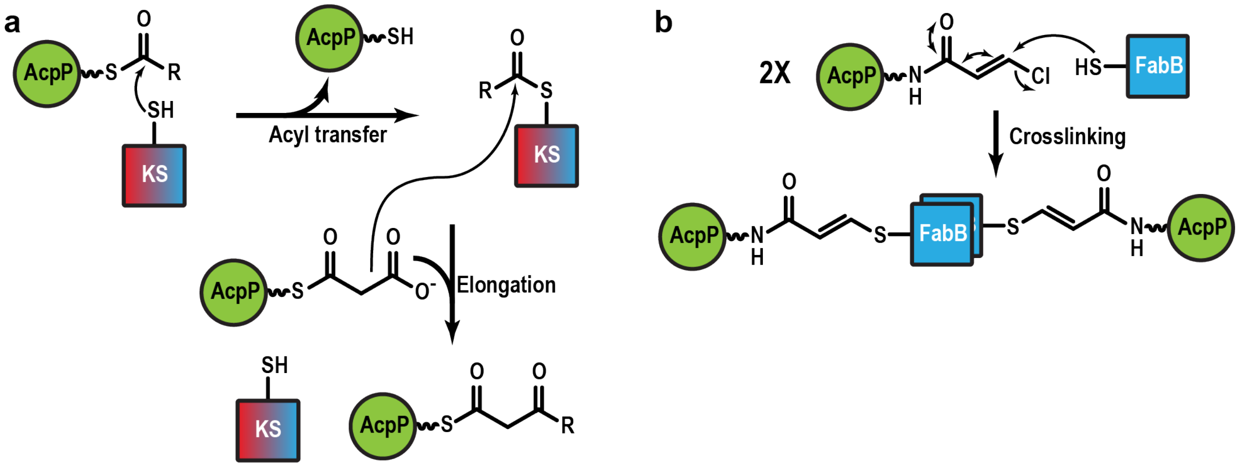


Figure 4-2. Enzyme mechanism and mechanism-based crosslinking. (a) mechanism for ketosynthase reaction and enzyme turnover. (b) mechanism-based crosslinking.

The β-ketoacyl-ACP-synthases (KSs) facilitate the key carbon-carbon bond formation step of fatty acid biosynthesis in three discrete steps (Fig. 4-2). First, an active site cysteine attacks the PPant thioester and releases the AcpP. Next, a malonyl-loaded AcpP associates and, by Claisen-like condensation, extends the chain. Finally, the AcpP dissociates, carrying with it the elongated β-keto-acyl chain. We recently applied a mechanism-based probe to crosslink AcpP with β-hydroxyacyl-AcpP-dehydratase (FabA), and the crystal structure of the complex was solved.[14-15] The FabA-AcpP complex structure reveals unprecedented information about how ACP interacts with its partner enzymes. However, to date, no information is available about how KS interacts with ACP. To address this knowledge gap, here, we report the investigation of KS-ACP interactions using a mechanism based inhibitor (Fig. 4-2B) using protein NMR, protein crystallography, and MD simulation. [16]

4.3 Results and discussions

4.3.1 The crystal structure of the AcpP-FabB complex

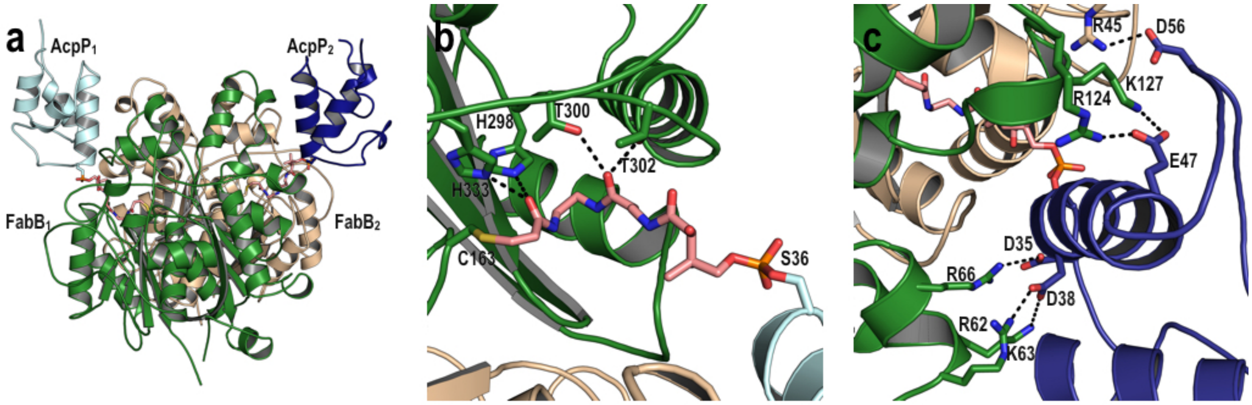


Figure 4-3. Crystal structure of *E. coli* AcpP-FabB complex. (a) Overall AcpP-FabB complex structure, with the FabB monomers shown in dark green and light tan, the AcpP monomers shown in dark blue and light cyan, and the crosslinker in pink. (b) FabB active site interactions with the crosslinker. (c) AcpP-FabB interface.

To prepare an AcpP-FabB complex, we developed a covalent chloroacrylyl probe to exploit the nucleophilicity of the FabB active site cysteine (Fig. 4-2a).[17] The chloroacrylyl-based probe was chemoenzymatically appended to AcpP followed by incubation with FabB to generate a covalent AcpP$_2$-FabB$_2$ complex (Fig. 4-2b).[16] The 2.4 Å resolution AcpP$_2$-FabB$_2$ crystal structure was solved by molecular replacement using a FabB structure (PDB code: 2VB9) as a search model followed by manual placement of the crosslinked AcpPs (Fig. 4-3).[18] The AcpP$_2$-FabB$_2$ complex consists of a core FabB dimer with each monomer crosslinked to a single AcpP (Fig. 4-3a), similar to the AcpP-FabA complex.[15] Each AcpP displays a four-helix bundle fold, and the probe extends from the conserved S36 sidechain at the bottom of AcpP helix II to the active site catalytic cysteine (C163) of FabB (Fig. 4-3b and Fig.4-4).

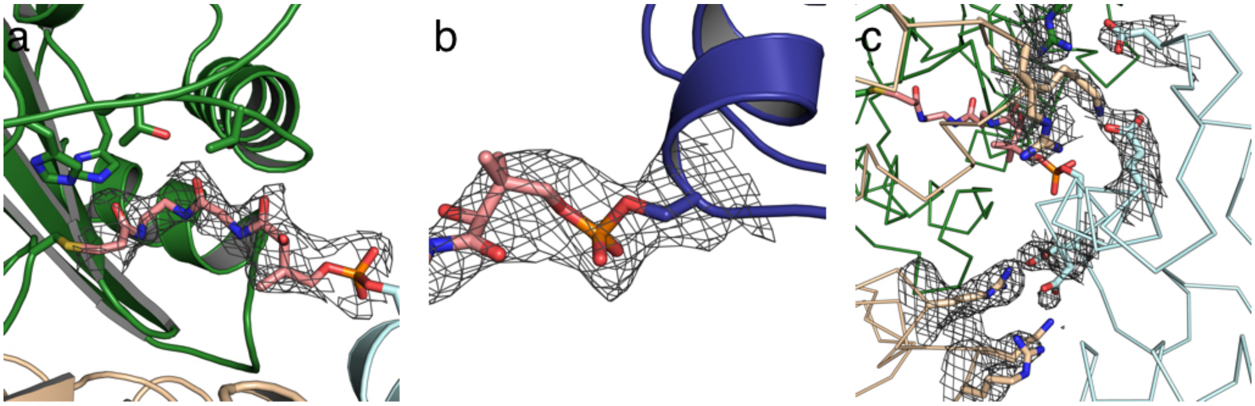


Figure 4-4. Composite omit maps at 1.0 sigma. (a) probe 2 in the FabB active site. (b) probe 2 connection with AcpP. (c) AcpP-FabB interface residues with AcpP shown in light cyan and FabB shown in dark green and light tan.

Each AcpP contacts FabB primarily through helix II, which is well conserved in carrier proteins (Fig. 4-3c).[19] Both AcpP-FabB interfaces share a set of common interactions. At the bottom of AcpP helix II, D35 and D38 interact with R62, K63, and R66 on FabB, while E47 at the top of helix II interacts with R124 and K127 on FabB. Additionally, D56 on helix III of each AcpP forms a salt bridge with R45 of its FabB partner. A comparison of the AcpP:FabB and AcpP:FabA interfaces reveals distinct AcpP docking motifs for each partner enzyme (Fig. 4-5); close contacts with FabB appear throughout helix II, whereas the interactions with FabA are predominantly at the top of helix II. In both structures, salt bridges are observed between helix III and the partner protein to facilitate chain flipping through a channel between helixes II and III (Fig. 4-5c,f).

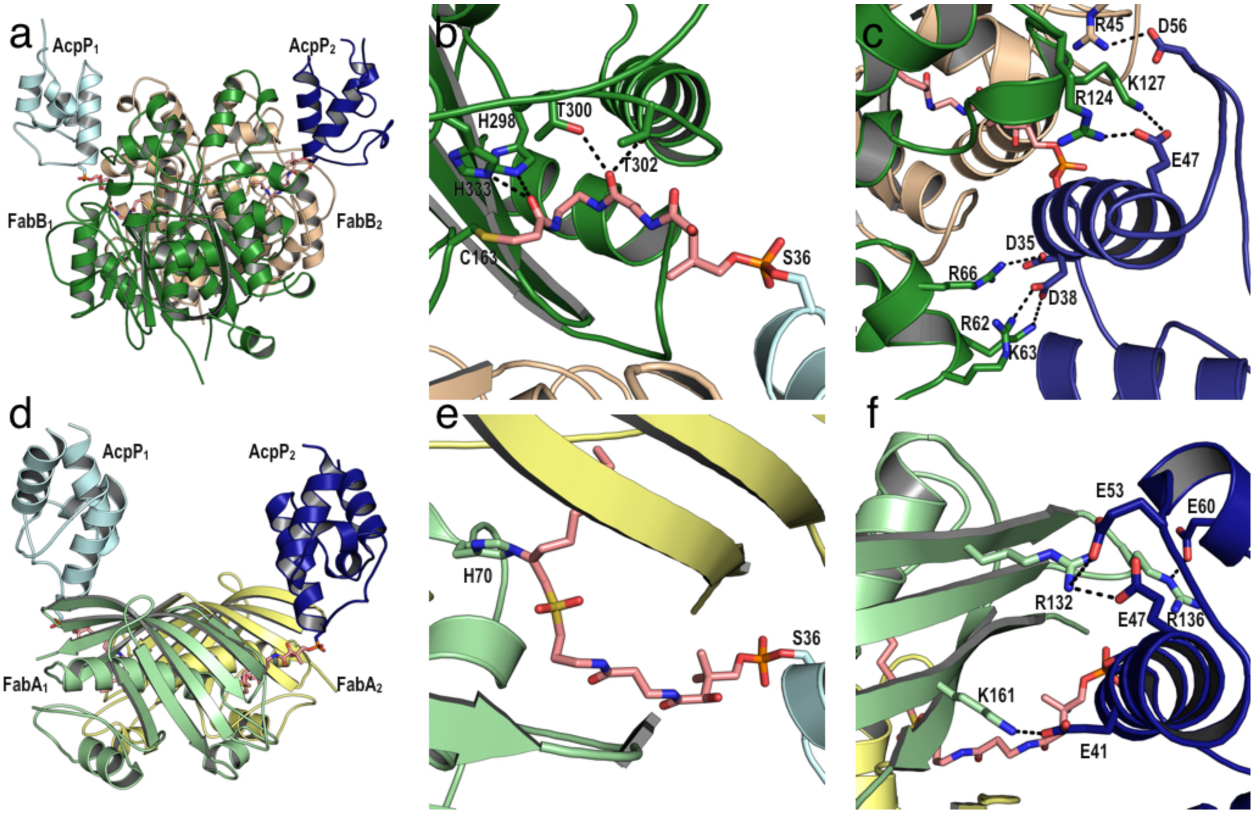


Figure 4-5. Comparison of AcpP-FabB and AcpP-FabA complex structures. (a) Overall AcpP-FabB complex structure, with the FabB monomers shown in dark green and light tan, the AcpP monomers shown in dark blue and light cyan, and the crosslinker in pink. (b) FabB active site interactions with the crosslinker. (c) AcpP-FabB interface. (d) Overall AcpP-FabA complex structure, with the FabA monomers shown in light green and light yellow, the AcpP monomers shown in dark blue and light cyan, and the crosslinker in pink. (e) FabA active site interactions with the crosslinker. (f) AcpP-FabA interface

A comparison of the individual AcpP-FabB pairs reveals that the pantetheine binding sites and protein-protein interfaces overlay well, but there are key differences. Most notably, divergence in the structure is observed in helix III of AcpP2 near D56; AcpP1 maintains its secondary structure at this location, while AcpP2 is completely disordered (Fig. 4-5b,e). Additionally, high crystallographic B-factors support the observation that helix III of AcpP2 is more disordered when compared to the ordered helix III on AcpP1 (Fig. 4-6). Similar differences were also observed in the AcpP monomers of the AcpP-FabA and AcpP-FabZ structures.[15, 20] Together, these suggest that AcpP interactions with dimeric partner enzymes may be influenced by allosteric regulation.

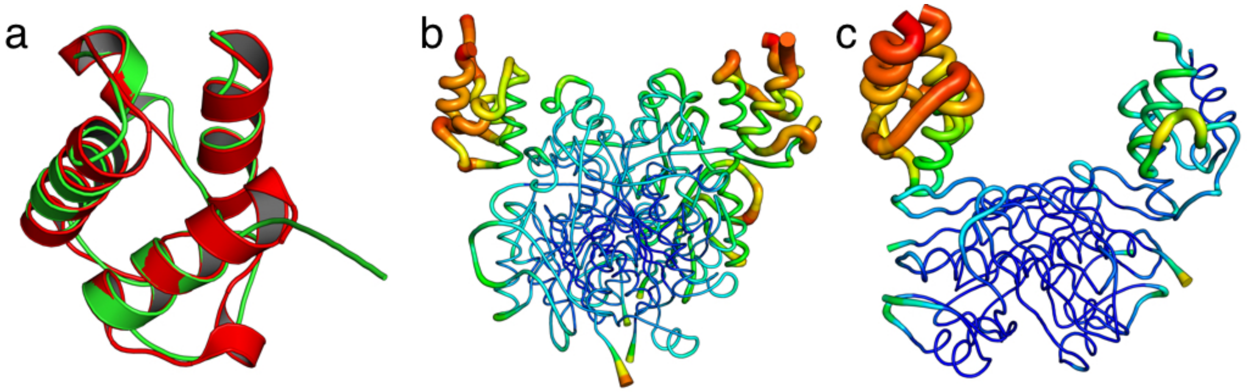


Figure 4-6. AcpP comparison between AcpP-FabB and AcpP-FabA. (a) Overlay of AcpP from AcpP-FabA (red) and from AcpP-FabB (green). (b)(c) B-factor putty representation of AcpP-FabB (b) and AcpP-FabA (c) where thin, blue tubes correspond to low B-factors and thicker, darker red tubes correspond to higher B-factors.

4.3.2 NMR titration studies

Solution-state protein NMR experiments were performed to characterize the AcpP-FabB interaction in vitro for comparison with crystallographic observations. $^{1}H,^{15}N$-HSQC titration experiments were performed in which octanoyl-AcpP was exposed to increasing molar equivalents of FabB (Fig. 4-7a and Fig. 4-8). Observed peak shifts due to the acyl-AcpP FabB interaction matched well with crystallographic observations. D35 and D38 exhibit dramatic chemical shift perturbations (CSPs) upon interaction, as well as E47 demonstrating a moderate CSP. Overall, the NMR CSP plots suggest the transient interaction between AcpP and FabB depends significantly on helix II, and to a lesser extent on helix III.

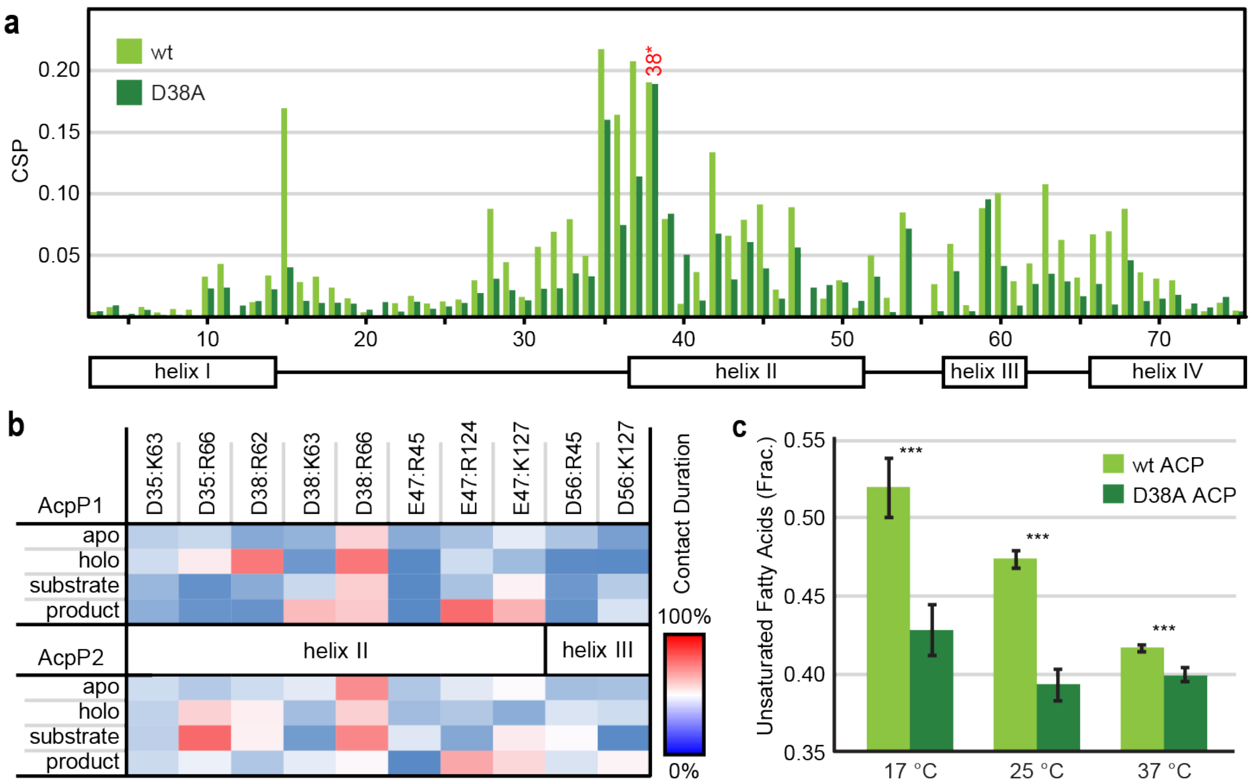


Figure 4-7. Protein-protein interactions and the fatty acid profile. (a) Chemical shift perturbation plot for $^{1}H,^{15}N$-HSQC titration experiments of WT AcpP and D38A AcpP against increasing concentrations of FabB. Residue number represented on the lower axis. (b) MD observed contact durations for helix II and III between different AcpP species FabB. (c) AcpP complementation temperature response. WT AcpP is shown in light green, D38A AcpP is shown in dark green. Four biological replicates were prepared. *** denotes statistical significance with P values < 0.001.

Due to the observed importance of D38, a D38A variant was prepared and subjected to the same NMR titration study as the wild-type AcpP (Fig. 4-7a and Fig. 4-8). D35A was not selected to avoid perturbing the DSL motif necessary for covalent attachment of the prosthetic arm. The D38A variant is still observed to interact with the FabB; however, the observed CSPs are smaller, suggesting a weaker interaction with FabB. Additionally, the top of helix II appears less perturbed in the D38A variant than in the wild type, confirming that the D38 (AcpP) – K63 (FabB) and D38 (AcpP) – R62 (FabB) salt bridges observed crystallographically are important in stabilizing the complex in solution.

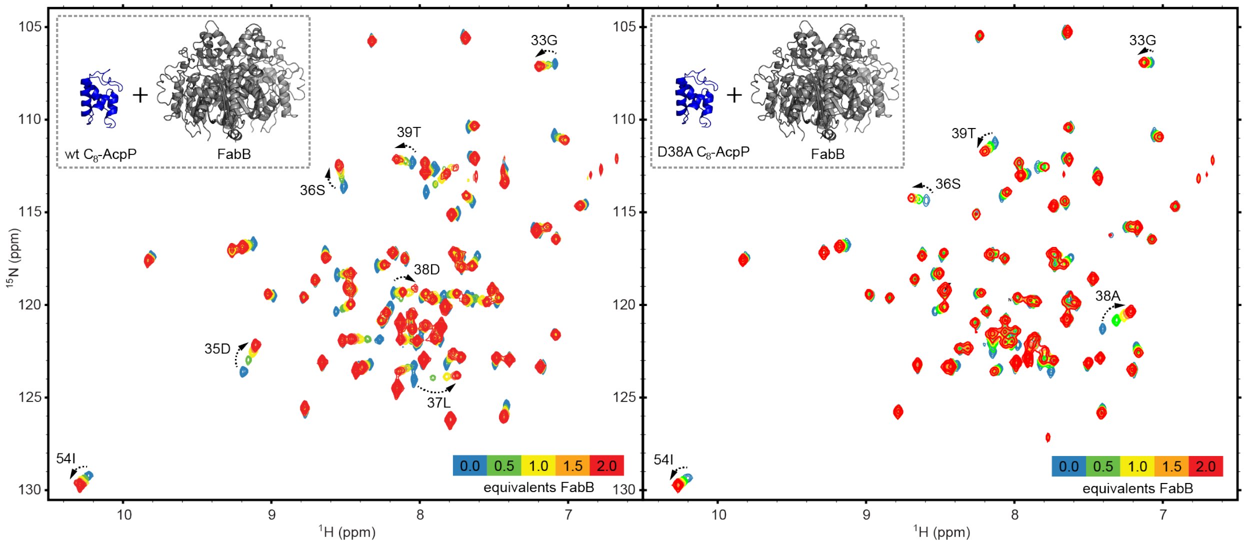


Figure 4-8. NMR studies of the AcpP-FabB complex. ^{1}H,^{15}N-HSQC titration experiments of wildtype AcpP (at left) and D38A AcpP (at right) against increasing concentrations of FabB.

4.3.3 Molecular dynamics simulations

Molecular dynamics (MD) was applied to the AcpP-FabB complex to explore the mechanism and dynamics involved in AcpP interactions with FabB. Several long simulations were performed on the complex with AcpP in different states to identify changes in protein-protein dynamics. The AcpPs were modeled as *apo* (no PPant), *holo* (empty PPant), C8 acyl-substrate loaded, or C10 acyl-product loaded. Comparisons of the average structures obtained from the various complex states revealed a possible mechanism for communication and recognition between AcpP and FabB (Fig. 4-7b, Fig.4-9, Fig.4-10). In the *apo* state, AcpP loses its ability to make specific salt bridge contacts observed in the crystal structure. However, upon being activated to its *holo* state, salt bridge contacts strengthen on helix II. Interestingly, upon actual loading of the PPant with a C8 FabB substrate or C10

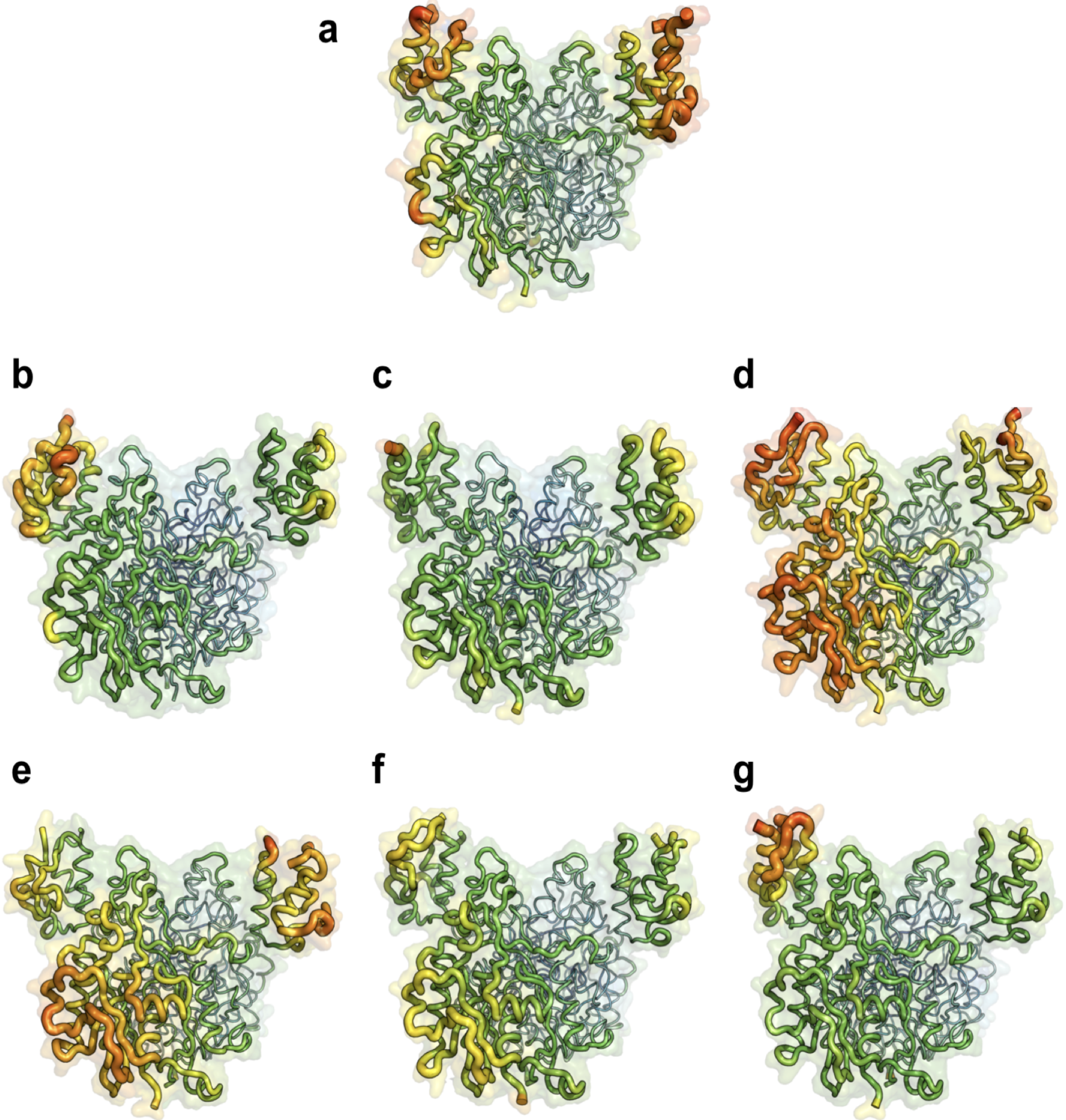


Figure 4-9. A comparison of MD derived RMSFs and crystallographic B-factors. All the above structures are displayed as putty, with higher RMSF/ B-factor values shown in red with increased thickness, and lower values shown in blue with decreased thickness. All MD structures are the average structure generated from the 1000 frames representing 140 to 150 ns. The RMSF MD structures are normalized with the crystallographic B-factors for comparison purposes. (a) Crystallographic structural for reference. Higher disorder is observed in FabB2 of the homodimer. (b) Both AcpPs in *apo* state. (c) Both AcpPs loaded with C8-substrate. (d) AcpP1 (left) and AcpP2 (right) loaded with C8-substrate and C8-product, respectively. (e) Both AcpPs in *holo* state. (f) Both AcpPs loaded with C8-product. (g) AcpP1 (left) and AcpP2 (right) loaded with C8-product and C8-substrate, respectively.

secondary structure reveals increased movement in the α6-α7 helix-turn-helix (HTH) motif present in FabB, which is in close proximity to interact with another highly flexible helix (α10) on the same monomer. Both motifs also are able to interact with substrates or products in the active sites; further, α10 is able to interact with helix III of AcpP. Therefore, binding of the substrate to monomer 1 may trigger the protein conformation/dynamic change of monomer 2, resulting in product release that completes the enzyme turnover of monomer 2. Together, these results highlight the importance of anchoring helix II interactions to facilitate correct orientation of the AcpP for productivity. Helix III becomes important when AcpP is loaded with the "cargo" (substrate or product of FabB), suggesting that chain translocation is facilitated by interactions between FabB and helix III of AcpP.

4.3.4 Root-mean square fluctuations (RMSF) analysis

RMSFs of the non-hydrogen atoms of each residue were computed on a total of 1000 frames, corresponding to 140 to 150 ns of each MD simulation using CPPTRAJ and a custom script.[21] Prior to performing the RMSF computations for each simulation, all frames were superimposed on the first frame, followed by stripping of all water molecules and sodium ions. Average structures with embedded RMSF values were generated for each simulation (Fig.4-9). Average structures for the wildtype complexes are shown with RMSF values normalized to crystallographic B-factors in the solved complex structure described in this paper. Interestingly, $FabB_2$ had increased RMSF values overall compared to $FabB_1$, consistent with the elevated B-factor values seen in $FabB_2$ of the crystallographic structure, strengthening the argument that MD and crystallographic data are able to structurally describe the FabB homodimer in two discrete states. RMSF values for every simulation were plotted for the FabB homodimer (Fig. 4-10), and both carrier proteins (Fig. 4-11). Difference plots were generated to compare the D38A, R62A and R124A variants to wild-type for the FabB homodimer (Fig. 4-12), and both carrier proteins (Fig. 4-13).

The D38A and R62A variants had much higher fluctuations compared to wild-type highlighting the sensitive protein-protein interactions. Interestingly, in all 24 simulations, $FabB_2$ had consistently higher fluctuations in RMSF values compared to FabB1 suggesting the homodimer is occupying two discrete states. The wildtype profile is relatively consistent amongst all loaded states, and the D38A, R62A and R124A variants showed a markedly decrease in consistency amongst states as a result of the disruption of key protein-protein interactions (Fig. 4-11). R124A showed the greatest increase in RMSFs. Comparisons between the variant and wild-type profiles reveal the D38A variant on AcpP to result in much higher RMSF values across both monomers (Fig. 4-12). Surprisingly, when a product is bound to either/both AcpPs there is no change, or a reduction in RMSFs for D38A, R62A, and R124A. This result implies the possibility that the specific salt bridges lose their importance after product turnover. R62A and R124A display the greatest change in RMSF on the AcpPs, as a decrease in specific protein-protein interactions between the pairs of proteins leads to an increase of atomic fluctuations on the relatively smaller AcpPs (Fig. 4-13).

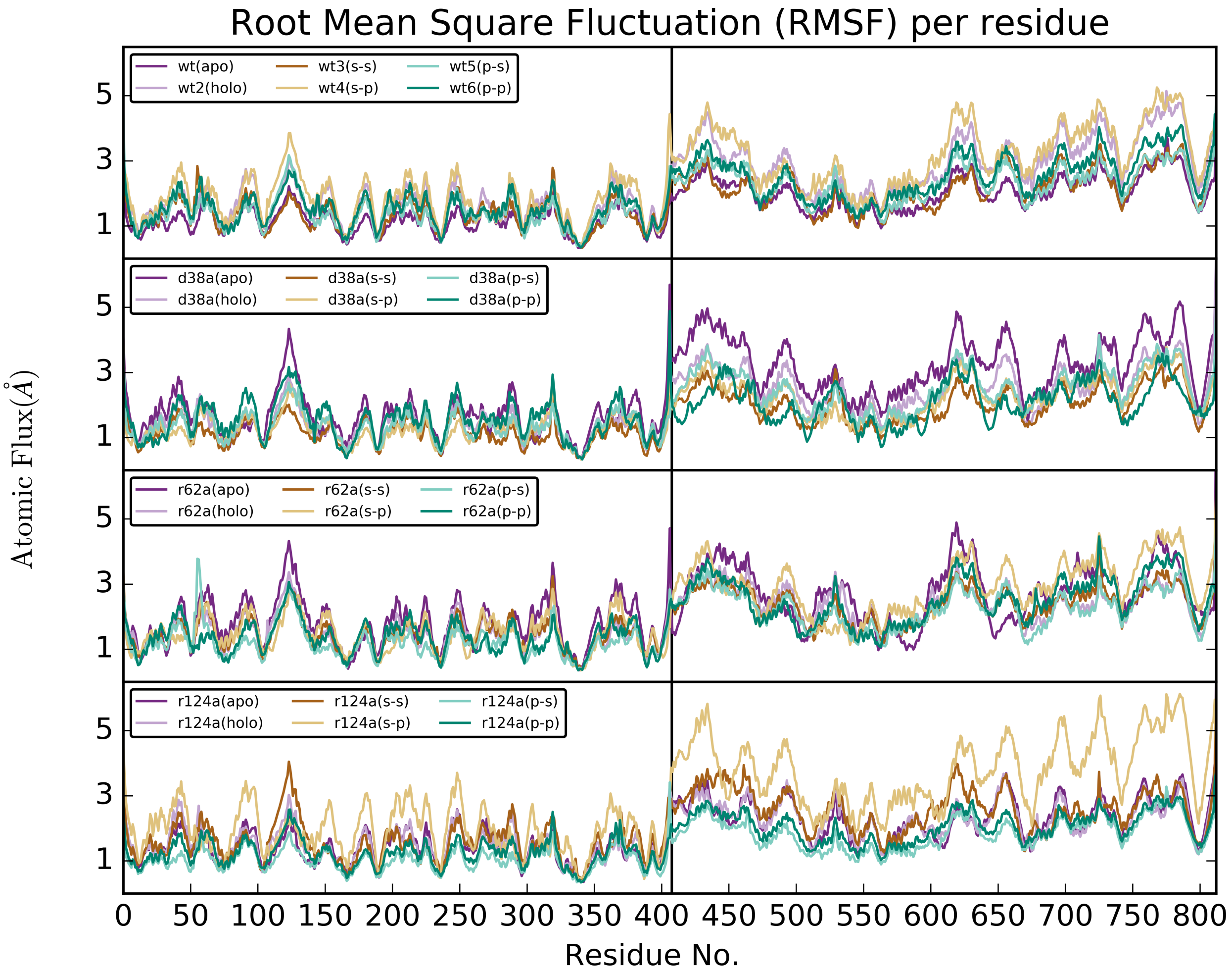


Figure 4-10. Root Mean Square Fluctuations (RMSF) per residue for the FabB homodimer were generated for wildtype, AcpP:D38A, FabB:R62A, and FabB:R124A over all 24 simulations.

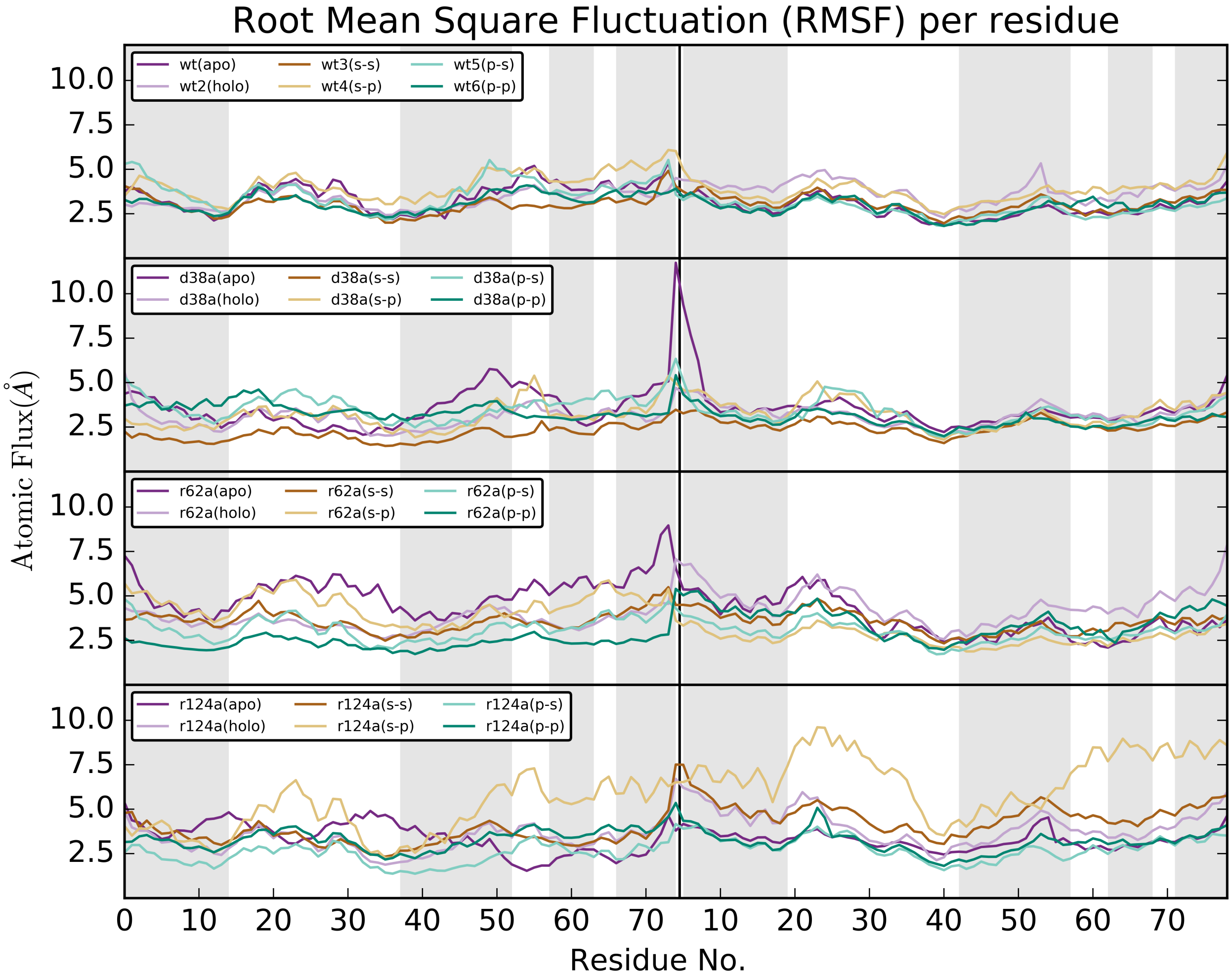


Figure 4-11. Root Mean Square Fluctuations (RMSF) per residue for both AcpPs were generated for wildtype, AcpP:D38A, FabB:R62A, and FabB:R124A over all 24 simulations. The grey highlighted portions in the background represent helical regions in AcpP.

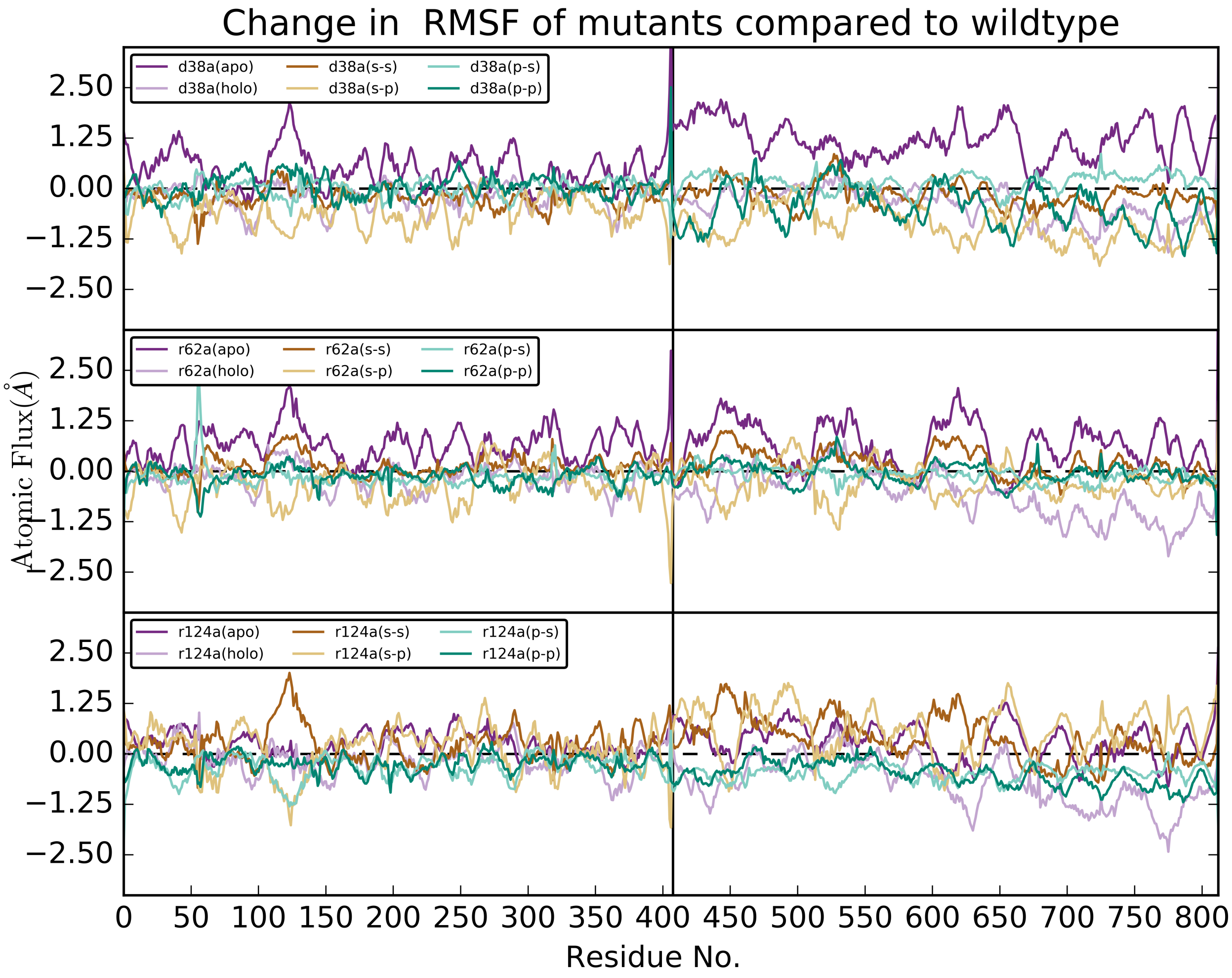


Figure 4-12. Comparisons between the variant and wild-type RMSF profiles for both FabB monomers.

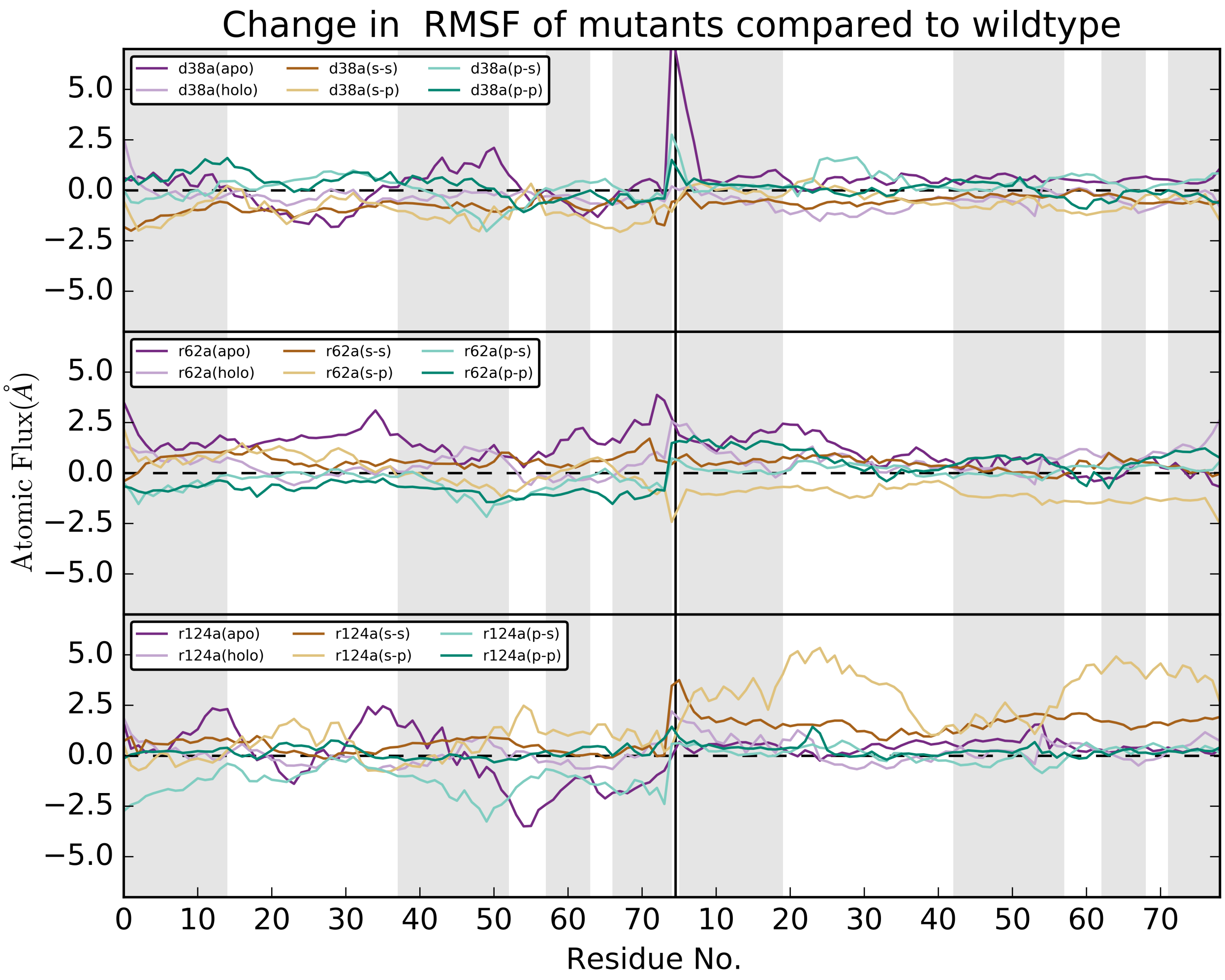


Figure 4-13. Comparisons between the variant and wild-type RMSF profiles for both AcpPs.

4.3.4 Free energy calculations

MM-PBSA calculations were performed to measure the ΔG of protein binding surface affinity for every AcpP present in the 24 simulations, for a total of 48 calculations (Fig. 4-14). These calculations were only performed using the frames from 140 to 150 ns, which were also the final set of frames for the variant simulations. When the AcpPs are in the *apo* state, the R62A and R124A variants had much lower binding affinity compared to wildtype, and slightly lower in the *holo* state compared to wildtype. Comparison of the ΔΔG values of the

D38A, R62A and R124A variants to the wildtype show a decrease in binding affinity for the R124A variant in the *apo* and *holo* states, with the measurements performed in the C8-substrate and C8-product states to be qualitatively decreased or equivalent (Fig. 4-15). The results of these calculations imply that the importance of the salt bridge interactions might only be important during recognition and initialization of the complex, and less important once the substrate is released via the switchblade mechanism into the active site.

Table 4-1. Results of MM-PBSA calculations.

	apo			**holo**		
	Average	Std. Dev	**ΔΔG vs wt**	Average	Std. Dev	**ΔΔG vs wt**
WT1	-32.182	11.019	**-**	-49.021	10.807	**-**
WT2	-19.290	14.346	**-**	-39.767	9.926	**-**
D38A1	-24.642	7.331	**7.541**	-52.019	9.556	**-2.998**
D38A2	-15.312	8.677	**3.977**	-37.382	9.188	**2.384**
R62A1	8.554	8.893	**4.554**	-22.507	8.070	**26.514**
R62A2	4.571	8.440	**23.861**	0.834	12.978	**40.600**
R124A1	-3.964	9.956	**28.218**	-14.524	9.443	**34.496**
R124A2	22.046	7.550	**41.335**	-1.124	8.311	**38.643**
	substrate			**product**		
	Average	Std. Dev	**ΔΔG vs wt**	Average	Std. Dev	**ΔΔG vs wt**
WT1	-54.494	10.444	**-**	-48.191	14.006	**-**
WT2	-26.413	16.917	**-**	-41.704	10.805	**-**
WT3	-36.733	15.971	**-**	-31.107	10.921	**-**
WT4	-49.100	10.304	**-**	-41.527	8.969	**-**
D38A1	-41.622	11.849	**12.872**	-48.634	16.302	**-0.442**
D38A2	-66.496	12.045	**-40.082**	-34.786	14.437	**6.918**
D38A3	-43.459	10.838	**-6.726**	-38.814	14.122	**-7.706**
D38A4	-52.578	14.110	**-3.478**	-47.527	11.605	**-6.000**
R62A1	-28.240	11.098	**26.253**	-27.971	12.310	**20.221**
R62A2	-44.623	10.575	**-18.210**	-28.068	13.985	**13.636**
R62A3	-42.327	10.661	**-5.593**	-21.964	10.759	**9.143**
R62A4	-40.060	12.611	**9.041**	-24.986	11.383	**16.542**
R124A1	-6.837	14.934	**47.657**	-10.614	9.363	**37.578**
R124A2	-27.210	12.313	**-0.797**	-9.296	11.265	**32.408**
R124A3	-14.090	7.598	**22.643**	-11.118	11.191	**19.989**
R124A4	-14.809	9.986	**34.292**	-43.781	11.650	**-2.254**

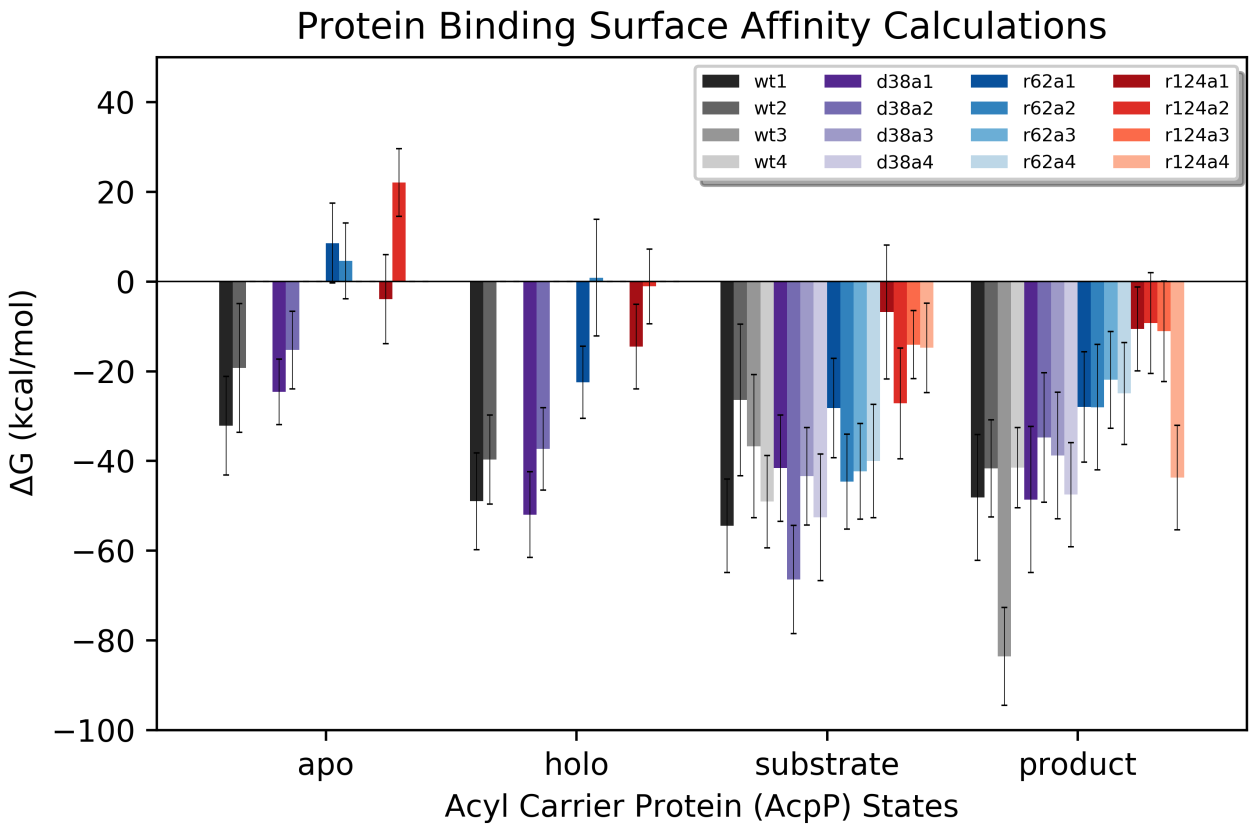


Figure 4-14. Protein binding surface affinity calculations**.** A total of 48 Molecular Mechanics-Protein Binding Surface Affinity (MM-PBSA) calculations were performed on the wild-type (wt), and three variants (D38A, R62A, and R124A) systems analyzing the ΔG of the second binding event between the carrier protein and the FabB homodimer. The complex encompassed both AcpPs and the FabB homodimer, with the ligand classified as an individual AcpP, and the remaining AcpP and FabB homodimer treated as the receptor. Two calculations were performed on each of the 24 simulations, measuring the ΔG of binding for both AcpPs in the complex. AcpPs were modeled in *apo*, *holo*, C8-substrate and C8-product forms. The wildtype, D38A, R62A and R124A values are colored in black, purple, blue and red, respectively. The first and third calculations in each set is measuring the the ΔG of binding for AcpP1, and the second and fourth calculations in each set are measuring the the ΔG of binding for AcpP2. For the first and second calculations in each set both carrier proteins are in identical states. To investigate cases of asymmetrical loaded states, both AcpPs were loaded with C8-substrate and C8-product for the third and fourth calculations.

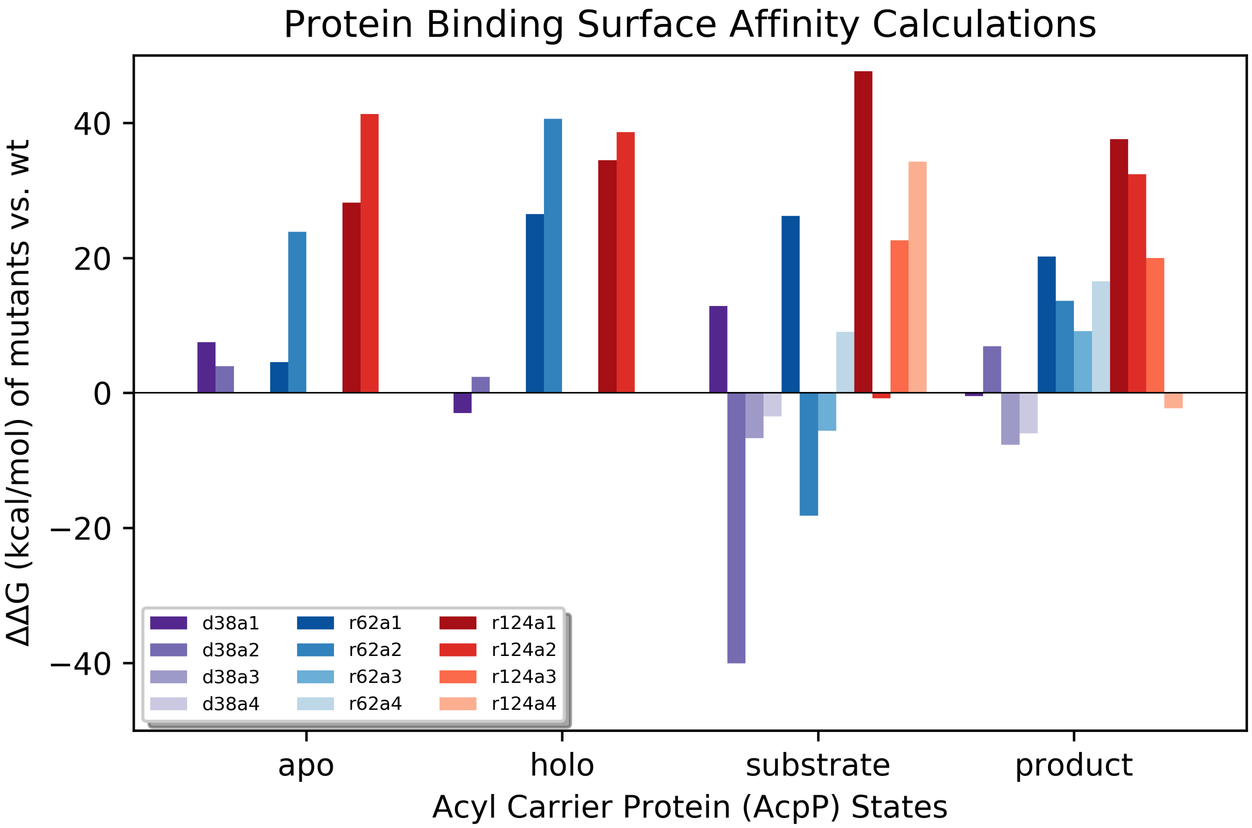


Figure 4-15. Protein binding surface affinity ΔΔG of variants compared to wildtype. Comparisons of ΔG values of the D38A, R62A and R124A variants with wildtype ΔG values is shown above, with the D38A, R62A and R124A variants colored in purple, blue and red, respectively. The first and third calculations in each set is measuring the the ΔG of binding for AcpP1, and the second and fourth calculations in each set are measuring the the ΔG of binding for AcpP2. For the first and second calculations in each set both carrier proteins are in identical states. To investigate cases of asymmetrical loaded states, the AcpPs were loaded with C8-substrate and C8-product for the third and fourth calculations. R124A has reduced binding affinity compared to wildtype for all states as shown by the increase in ΔG of binding. D38A, R62A and R124A have an observed reduction in binding affinity compared to wildtype for the apo state, suggesting that these specific residues involved in salt bridge formation are important for recognition.

4.3.5 FabB mutagenesis and lipid profile modulation in *E. coli*

A previous study showed the fatty acid product profile *in vitro* is subject to the relative concentration of FabB.[22] The specific parameters that the FabB concentration controls is chain length and level of saturation present in the fatty acids. As FabB is uniquely suited to extend 10-carbon cis-unsaturated fatty acids, it was initially hypothesized that a decrease in FabB will result in a decrease in the production of unsaturated fatty acids. FabF, another ketosynthase involved in fatty acid biosynthesis is unable to process unsaturated 10-carbon substrates. Surface variants were generated, and quite unexpectedly we did not observe a decrease in saturation as would have been initially predicted (Fig. 4-16).

Counter-intuitively, the majority of the variants actually increased unsaturation when FabB was overexpressed in respect to the native system. R124A saw a marked increase in C18:1 product, and even generated C16:1 product, which is not observed in wild-type. A new model is proposed which would support the above results, and that is a decrease in binding affinity as a result of the surface mutations, actually increases the turn-over rate between FabB and AcpP. Further kinetic studies would need to be performed to evaluate if the rate limiting step is actually in fact the dissociation step and enzyme turnover, instead of the previously assumed protein-association step.

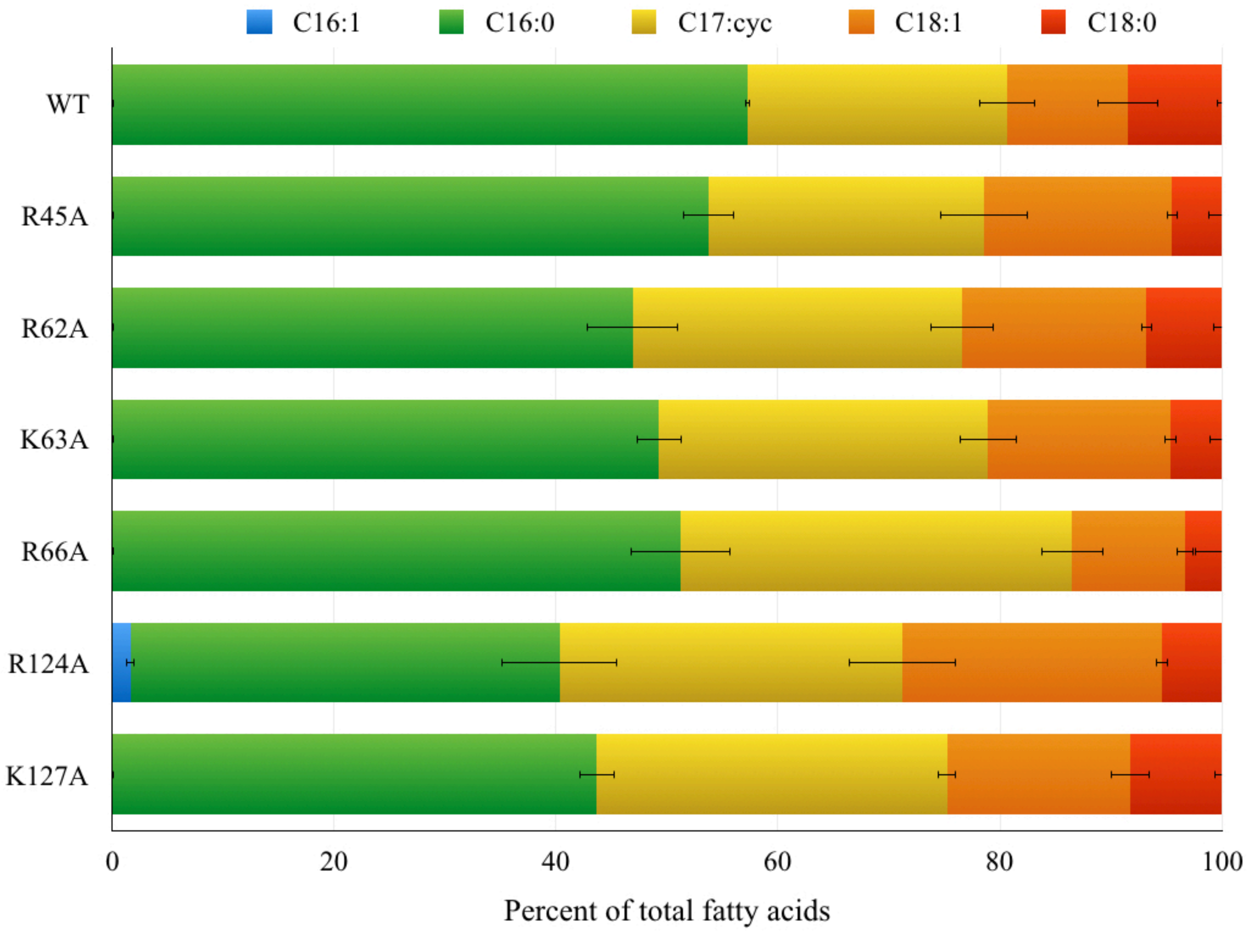


Figure 4-16. Product profile of the *E. coli* type II FAS bearing FabB surface mutations. The products are denoted by chain length and unsaturation level. All experiments were run in triplicate.

4.4 Materials and methods

4.4.1 Protein expression and purification

pET28b 6x His-tagged constructs containing the genes for *E. coli* AcpP and FabB were separately transformed into *E. coli* BL21 (DE3) cells by heat shocking at 42 °C for 45 seconds and plated onto LB agar plates supplemented with 50 μg/mL kanamycin. Colonies were transferred to a 10 mL LB starter culture supplemented with 50 μg/mL kanamycin and shaken overnight at 37 °C. The starter culture was then transferred to 1 L of LB supplemented with 50 μg/mL kanamycin and shaken at 37 °C. Once the OD600 reached 0.6, expression was induced by addition of 1 mM IPTG, and the cells were shaken overnight at 18 °C. Cells were harvested by centrifugation, resuspended in lysis buffer (50 mM Tris pH 7.5, 300 mM NaCl, 10 mM imidazole, 10 % glycerol), and flash frozen in liquid nitrogen for storage at -80 °C.

The resuspended cells were lysed by sonication, and the lysate was centrifuged at 21,000 rcf for 60 minutes to remove cellular debris. The lysate was then incubated with 5 mL of Ni-NTA resin (ThermoFisher Scientific) for 1 hour, and the resin was washed with lysis buffer to remove unbound protein. The proteins of interest were eluted in fractions using an imidazole gradient. SDS-PAGE was used to analyze the fractions, and fractions containing the protein of interest were combined and dialyzed into a storage buffer (25 mM Tris pH 7.5, 100 mM NaCl, 1 mM DTT, 5% glycerol). AcpP was further purified using a HiTrap Q FF anion exchange column (GE Healthcare Lifesciences).

4.4.2 Crosslinking, complex purification, and crystallization

The phosphopantetheine (PPT) prosthetic group was removed from AcpP by incubation with MBP-tagged AcpH, including 12.5 mM MgCl2 and 2.5 mM MnCl2. After removing the AcpH with amylose resin (New England Biolabs), apo-AcpP was chemoenzymatically loaded with the chloroacryl-pantetheine crosslinker to form crypto-AcpP using recombinant CoaA, CoaD, CoaE, Sfp, 200 mM ATP, 10 mM MgCl, and a 1.5x molar excess of the crosslinker.[17] The loading was confirmed using MALDI-TOF mass spectrometry. AcpP was purified away from the loading enzymes using a second HiTrap Q FF column.

100 μM FabB was incubated with 200 μM crypto AcpP in 25 mM Tris pH 7.5, 100 mM NaCl, 5 mM DTT, and 5% glycerol overnight at 37 °C. The complex was then purified using a Superdex 200 (GE Healthcare Lifesciences) size exclusion column, concentrated to 6 mg/mL, and flash frozen for storage. The complex was crystallized in 0.1 M sodium acetate pH 5.4, 0.2 M ammonium acetate, and 20% PEG 4000 using the sitting drop vapor diffusion method. The crystals were flash frozen in liquid nitrogen for storage. Diffraction data were measured using beamline 8.2.1 at the Advanced Light Source synchrotron facility and processed using Mosflm.[23] The structure was solved by molecular replacement using a FabB structure (PDB code: 2VB9) as a search model and refined using the Phenix suite of programs.[24]

4.4.3 NMR titration studies

Isotopically labeled wt and D38A AcpP for NMR studies was prepared as previously reported.[15, 17] In brief, BL21 (DE3) *E. coli* cells bearing the appropriate construct were first deuterium acclimated then grown at 1 L scale in 13C, 15N, and 2H M9 minimal media. A 5 mL culture of M9 minimal media prepared with 25% D2O and 75% H2O was inoculated and grown overnight at 37 °C. 100 µL of this dense culture was used to inoculate a 5 mL culture prepared with 50% D2O and 50% H2O, which was grown overnight at 37 °C. In turn, 100 µL of this dense culture was used to inoculate a 5 mL culture prepared with 75% D2O and 25% H2O, which was grown overnight at 37 °C. Finally, 100 µL of this dense culture was used to inoculate a 5 mL culture prepared with 100% D2O and grown overnight at 37 °C. This was used to inoculate a 1 L flask of M9 minimal media prepared with 1 L D2O, 1 g 15NH4Cl, and 4 g 13C glucose. The culture was grown (37 °C, 120 RPM shaking, baffled flask) to an OD600 of 0.8. Expression was induced with the addition of 1 mL of 1M IPTG (in D2O, 0.22 µM sterile filtered), and carried out for 4 hours at 37 °C. Cells were harvested by centrifugation at 600 rcf, 30 min, 6 °C.

Cells were resuspended in 40 mL of lysis buffer (25 mM HEPES pH 7.4, 250 mM NaCl, and 10% glycerol). They were lysed by French pressure cell at ~25,000 PSI, over three presses, with DNAse and RNAse. Lysate was clarified by centrifugation at 12,000 rcf, 45 min, 6 °C and subjected to Ni-NTA (ThermoFisher Scientific) purification. Clarified lysate was tumbled with loose resin at 4 °C for 30 minutes, then washed with 25 mL wash buffer (25 mM HEPES pH 7.4, 250 mM NaCl, and 10% glycerol, 25 mM Imidazole pH 7.4) to remove nonspecific binding. AcpP was eluted with 10 mL elution buffer (25 mM HEPES pH 7.4, 250 mM NaCl, and 10% glycerol, 250 mM Imidazole pH 7.4) then desalted (PD-10 desalting column, GE Healthcare) into 25 mM HEPES pH 7.4, 250 mM NaCl, and 10% glycerol.

NMR experiments were carried out in the Biomolecular NMR Facility at UCSD, managed by Dr. Xuemei Huang. Titration experiments were performed on a 600 MHz Bruker Avance III system equipped with a cryoprobe at 37 °C. Each 1H-15N HSQC was acquired with uniform sampling, 2048 (R+I) points in the direct dimension, 256 (R+I) points in the 15N dimension, 24 scans, and a 1.5 s recycle delay. Samples were prepared at 100 uM AcpP in 50 mM potassium phosphate pH 7.4, 0.01% sodium azide, 2.5 mM tris(2-carboxyethyl)phosphine, and 10% D_2O. Data was processed using nmrPipe5 and NMRFAM-SPARKY.[25] Backbone assignment of the D38A variant was achieved by comparison with the known wt AcpP and confirmation using an HNCACB. Titration increment points were achieved by preparing a 0.0 molar equivalent and 2.0 molar equivalent sample, acquiring the first and last points of the titration, then incrementally mixing them to achieve intermediate ratios. This approach was found to yield more accurate ratios and limit protein loss. Chemical Shift Perturbations (CSPs) were calculated using the formula: CSP = [0.5*((ΔδN/5)^2+(ΔδH)^2)]^(½) using the 0.0 and 2.0 molar equivalent titration points.[26]

4.4.4 Molecular dynamics (MD) simulations

The AcpP-FabB complex were modeled using the mechanism-based crosslinked structure described in this paper. Using the software UCSF Chimera, crystal waters were removed and hydrogens added using the Dock Prep tool, and the chloroacrylyl-based probe covalently linked to S36 was converted in silico into *holo*, C8-substrate and C8-product representations.[27-28] The phosphopantetheinyl serine residues were extracted and capped with N-methyl (NME) and acetyl (ACE) fragments to generate dipeptides for force field preparation and restricted electrostatic potential (RESP) charge fitting. Gaussian 09 Rev C was used for geometry optimization and electrostatic potential calculations using MP2/6-31G(d,p)//MP2/6-31G(d,p). RESP charges were calculated with intramolecular charge restraints and an overall charge of -1 for each non-standard residue and fit using the R.E.D.

Server.[29] When possible, bond parameters were assigned using the main parameter databases in the Amber ff14SB force field. Missing parameters were adopted from the general AMBER force field (GAFF). Six simulation types were setup by configuring S36 on chains C and D of the carrier proteins as: C0D0 (apo:apo), C1D1 (holo:holo), C2D2 (C8-substrate:C8-substrate), C2D3 (C8-substrate:C8-product), C3D2 (C8-product:C8-substrate), and C3D3 (C8-product:C8-product). The crystal structure served as the basis of the initial structure for all simulations. LEaP was used to neutralize the apo systems by adding 52 Na+ ions, and the non-apo systems by adding 54 Na+ ions and solvating the enzyme complexes in 12-Å water buffer TIP3P truncated octahedron boxes. The fully solvated systems contained between 82,554 and 83,278 atoms.

MD was carried out using AMBER 16.[30] Minimization was carried out in two stages using SANDER from AmberTools 16. The initial stage was carried out over 2,500 steps for the solvent, ions, and post-translationally modified S36 residues of the carrier proteins, with the remaining residues of the proteins restrained by a force constant of 500 kcal/mol/Å2, followed by a second stage carried out over 5,000 steps of the entire system. Heating was performed using SANDER over 100-ps allowing the system to heat up to a temperature of 300K using the Langevin temperature equilibration scheme. During heating the same set of atoms as the initial stage of minimization was restrained, but with a lower force constant of 10 kcal/mol/Å2. PME was used to compute the electrostatic interactions with a real space cutoff of 10 Å, which was also used for the van der Waals interactions. Time steps were set to 2 fs, with hydrogen atoms constrained using the SHAKE algorithm. Equilibration was carried out using SANDER for 10 ns on all simulations. The MD production simulations were carried out using PMEMD.CUDA. The 6 wildtype complex system simulations were run over

1 μs (500,000,000 time steps), and the 18 variant complex system simulations were run over 150 ns (75,000,000 time steps) for a total of 8.7 μs for all 24 simulations. Simulation speeds of 45 ns/day were observed.

4.4.5 Free energy calculations

The Molecular Mechanics – Protein Binding Surface Affinity (MM-PBSA) calculations were performed using MMPBSA.py in Amber 16 to measure the ΔG of the second binding event between the carrier protein and the FabB homodimer with one carrier protein attached in either the apo, holo, substrate or product states.[31] The total non-polar solvation free energy was modeled as a single term, which was linear and proportional to the solvent accessible surface area. The ionic strength was set at 0.1 mM for the PB equation. Atom-type/charge-based radii as described by Tan and Luo were used for the standard atom types, and radii from the parameter-topology files were used for the general amber force field (GAFF) atom types.[32] The calculations were performed twice on each of the 24 simulations, treating the ligand as either AcpP_1 or AcpP_2, and the remaining FabB homodimer and opposite AcpP as the receptor for the complexes. The wildtype, and three R38A, R62A and R124A variants were the four complex systems analyzed. A total of four calculation types were performed for each of the four states possible for each complex system, with the first type measuring the ΔG of binding in AcpP_1 in the represented state, to its respective FabB homodimer complex with AcpP_2 already bound. The second type measured the opposite carrier protein AcpP_2 in its represented state, to its respective FabB homodimer complex with AcpP_1 already bound. The third type for each system were calculations performed when AcpP_1 was in a substrate state or product state, and it's opposing AcpP was in the opposite substrate or product state. The fourth type was carried out like the 3rd type, but for $AcpP_2$. The calculations were performed using the frames corresponding to 140 to 150 ns of each MD simulation, and included a total of 1,000 evenly spaced frames. Convergence was measured (Fig.4-13) by assessing the cumulative average over the 1,000 frames.

4.4.6 Principal component analysis (PCA)

Principle component analysis was performed using CPPTRAJ.[21, 33-34] A total of 160 ns of simulation data was used for each PCA calculation, for a total of 16000 frames. Prior to performing PCA computations for each simulation, all frames were superimposed on the first frame, and water molecules and sodium ions were stripped. The coordinate covariance matrix used each residue's heavy atoms, for a total of 3,848 atoms and a coordinate covariance matrix size of 11,544 x 11,544 describing the Cartesian coordinates of each frame. The projection along these eigenvectors of each coordinate frame from the first simulation trajectory was calculated.

4.5 Conclusions

Combined with our previous studies on the AcpP-FabA interaction, the AcpP-FabB structure provides foundational progress towards understanding unsaturated fatty acid production in *E. coli*. Furthermore, the *in vitro*, *in silico*, and *in vivo* results together highlight the importance of AcpP helix II for anchoring, and AcpP helix III for chain translocation. Chain flipping may be required for the productive outcome of these protein-protein interactions, but this does not address the subtle differences in each interaction that are critical to pathway regulation, processivity, and cargo communication. The differences in the

AcpP-FabA and AcpP-FabB interactions reveal the different interaction networks that can act on the AcpP. Moving forward, these subtle interactions must be thoroughly understood to inform future drug discovery and pathway engineering efforts.

4.6 Acknowledgements

My contributions to this project included the molecular dynamic simulations, which provided the basis for biochemical investigations in this study. The X-ray crystallography studies and biological assays were performed by Dr. Jake Milligan, and the NMR experiments were performed by Dr. John Lee.

References

1. Finzel, K.; Lee, D. J.; Burkart, M. D., Using modern tools to probe the structure-function relationship of fatty acid synthases. *Chembiochem* **2015,** *16* (4), 528-547.
2. Feng, Y.; Cronan, J. E., Escherichia coli unsaturated fatty acid synthesis: complex transcription of the fabA gene and in vivo identification of the essential reaction catalyzed by FabB. *J Biol Chem* **2009,** *284* (43), 29526-35.
3. Wright, H. T.; Reynolds, K. A., Antibacterial targets in fatty acid biosynthesis. *Curr Opin Microbiol* **2007,** *10* (5), 447-53.
4. Yu, X.; Liu, T.; Zhu, F.; Khosla, C., In vitro reconstitution and steady-state analysis of the fatty acid synthase from Escherichia coli. *Proc Natl Acad Sci U S A* **2011,** *108* (46), 18643-8.
5. Chemler, J. A.; Tripathi, A.; Hansen, D. A.; O'Neil-Johnson, M.; Williams, R. B.; Starks, C.; Park, S. R.; Sherman, D. H., Evolution of Efficient Modular Polyketide Synthases by Homologous Recombination. *J Am Chem Soc* **2015,** *137* (33), 10603-9.
6. Zhu, L.; Cronan, J. E., The conserved modular elements of the acyl carrier proteins of lipid synthesis are only partially interchangeable. *J Biol Chem* **2015,** *290* (22), 13791-9.
7. Gajewski, J.; Buelens, F.; Serdjukow, S.; Janssen, M.; Cortina, N.; Grubmuller, H.; Grininger, M., Engineering fatty acid synthases for directed polyketide production. *Nat Chem Biol* **2017,** *13* (4), 363-365.
8. Rock, C. O.; Cronan, J. E., Jr., Acyl carrier protein from Escherichia coli. *Methods Enzymol* **1981,** *71 Pt C*, 341-51.
9. Kim, Y.; Prestegard, J. H., A dynamic model for the structure of acyl carrier protein in solution. *Biochemistry* **1989,** *28* (22), 8792-7.

10. Roujeinikova, A.; Simon, W. J.; Gilroy, J.; Rice, D. W.; Rafferty, J. B.; Slabas, A. R., Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J Mol Biol* **2007,** *365* (1), 135-45.

11. Mofid, M. R.; Finking, R.; Marahiel, M. A., Recognition of hybrid peptidyl carrier proteins/acyl carrier proteins in nonribosomal peptide synthetase modules by the 4'-phosphopantetheinyl transferases AcpS and Sfp. *J Biol Chem* **2002,** *277* (19), 17023-31.

12. Cronan, J. E., The chain-flipping mechanism of ACP (acyl carrier protein)-dependent enzymes appears universal. *Biochem J* **2014,** *460* (2), 157-63.

13. Beld, J.; Cang, H.; Burkart, M. D., Visualizing the chain-flipping mechanism in fatty-acid biosynthesis. *Angew Chem Int Ed Engl* **2014,** *53* (52), 14456-61.

14. Ishikawa, F.; Haushalter, R. W.; Lee, D. J.; Finzel, K.; Burkart, M. D., Sulfonyl 3-alkynyl pantetheinamides as mechanism-based cross-linkers of acyl carrier protein dehydratase. *J Am Chem Soc* **2013,** *135* (24), 8846-9.

15. Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R.; Bruegger, J.; Caldara-Festin, G.; Finzel, K.; Jackson, D. R.; Ishikawa, F.; O'Dowd, B.; McCammon, J. A.; Opella, S. J.; Tsai, S. C.; Burkart, M. D., Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **2014,** *505* (7483), 427-31.

16. Worthington, A. S.; Burkart, M. D., One-pot chemo-enzymatic synthesis of reporter-modified proteins. *Org Biomol Chem* **2006,** *4* (1), 44-6.

17. Worthington, A. S.; Rivera, H.; Torpey, J. W.; Alexander, M. D.; Burkart, M. D., Mechanism-based protein cross-linking probes to investigate carrier protein-mediated biosynthesis. *ACS Chem Biol* **2006,** *1* (11), 687-91.

18. Pappenberger, G.; Schulz-Gasch, T.; Kusznir, E.; Muller, F.; Hennig, M., Structure-assisted discovery of an aminothiazole derivative as a lead molecule for inhibition of bacterial fatty-acid synthesis. *Acta Crystallogr D Biol Crystallogr* **2007,** *63* (Pt 12), 1208-16.

19. Byers, D. M.; Gong, H., Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. *Biochem Cell Biol* **2007,** *85* (6), 649-62.

20. Zhang, L.; Xiao, J.; Xu, J.; Fu, T.; Cao, Z.; Zhu, L.; Chen, H. Z.; Shen, X.; Jiang, H.; Zhang, L., Crystal structure of FabZ-ACP complex reveals a dynamic seesaw-like catalytic mechanism of dehydratase in fatty acid biosynthesis. *Cell Res* **2016,** *26* (12), 1330-1344.

21. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013,** *9* (7), 3084-95.

22. Xiao, X.; Yu, X.; Khosla, C., Metabolic flux between unsaturated and saturated fatty acids is controlled by the FabA:FabB ratio in the fully reconstituted fatty acid biosynthetic pathway of Escherichia coli. *Biochemistry* **2013,** *52* (46), 8304-12.

23. Battye, T. G.; Kontogiannis, L.; Johnson, O.; Powell, H. R.; Leslie, A. G., iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D Biol Crystallogr* **2011,** *67* (Pt 4), 271-81.

24. 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.; Zwart, P. H., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **2010,** *66* (Pt 2), 213-21.

25. Lee, W.; Tonelli, M.; Markley, J. L., NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2015,** *31* (8), 1325-7.

26. Pellecchia, M.; Sebbel, P.; Hermanns, U.; Wuthrich, K.; Glockshuber, R., Pilus chaperone FimC-adhesin FimH interactions mapped by TROSY-NMR. *Nat Struct Biol* **1999,** *6* (4), 336-9.

27. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-12.

28. Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C. C.; Pettersen, E. F.; Goddard, T. D.; Meng, E. C.; Sali, A.; Ferrin, T. E., UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **2012,** *179* (3), 269-78.

29. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y., R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* **2011,** *39* (Web Server issue), W511-7.

30. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* **2012,** *8* (5), 1542-1555.

31. Miller, B. R., 3rd; McGee, T. D., Jr.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E., MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J Chem Theory Comput* **2012,** *8* (9), 3314-21.

32. Tan, C.; Yang, L.; Luo, R., How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *J Phys Chem B* **2006,** *110* (37), 18680-7.

33. Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E., 3rd, On the absence of intrahelical DNA dynamics on the mus to ms timescale. *Nat Commun* **2014,** *5*, 5152.

34. Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E., 3rd, Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim Biophys Acta* **2015,** *1850* (5), 1041-58.

CHAPTER 5

Identification of a Possible Allosterically Regulating Histidine pH sensor in a Type III Polyketide Synthase Benzalacetone Synthase

5.1 Summary

Type III polyketide synthases are stand-alone homodimers capable of carrying out multiple decarboxylative condensation reactions to generate complex polyketide products. In these systems, the acetate units are incorporated in the growing polyketide chain using malonyl-CoA extender units and a coumaroyl-CoA starter unit. The benzalacetone synthase RpBAS conducts a one-step decarboxylative condensation reaction to generate the diketide *p*-hydroxybenzalacetone (BA) as the dominant prouct at pH > 7, or it conducts two decarboxylative condensation reactions to generate the triketide *bis*-noryangonin (BNY) as the dominant product at pH < 7. While several structural and mutagenesis studies have been performed on RpBAS and its homologs, the precise mechanism of how pH affects the product outcome remains a mystery. In this chapter, using molecular dynamic simulations, we propose a framework that explains the pH dependence of product outcome via an allosteric regulation mechanism in type III polyketide synthases, involving a solvent exposed histidine on the surface of RpBAS. This *in silico* model resulted from the investigation of conformational dynamics and pKa calculations of titratable residues using MD simulations. The model was then explored experimentally via mutagenesis studies and biochemical assays. The MD-based mutagenesis successfully led to a change of product outcome at different pHs. The above result represents the first plausible model for pH dependence of type III PKS product outcome, and it paves the foundation for future engineering efforts of type III PKSs with a general approach towards harnessing pH for product specificity.

5.2 Introduction

The model enzyme of enzymology for type III polyketide synthases (PKSs) is chalcone synthase (CHS) from alfalfa (*Medicago sativa*), specifically MsCHS2, which is responsible for the generation of a tetraketide-derived naringenin chalcone (NAR), a precursor involved in the biosynthesis of flavonoid phytoalexins and anthocyanin pigments.[1] Crystal structures of MsCHS2 include an *apo* form, a CHS-CoA complex, a CHS-malonyl-CoA complex, CHS-hexanoyl-CoA complex, CHS-naringenin complex, and a CHS-resveratrol complex (PDB IDs: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, and 1CGZ, respectively).[2] This initial structural study provided important information about the three-dimensional structure of type III PKSs, as well as the substrate binding pockets and key residues for binding and catalysis. Another enzyme from the type III PKS family, Benzalacetone synthase (RpBAS), generates only a diketide-derived product, though its sequence similarity is 70% with MsCHS2. RpBAS was originally characterized by Abe *et al.*, who revealed that it produces diketide-derived *p*-hydroxybenzalacetone by a one-step decarboxylative condensation of 4-coumaroyl-CoA with one malonyl-CoA.[3] It is thought that *p*-hydroxybenzalacetone is important for the biosynthesis of phenylbutanoids.[4] Curiously, RpBAS biosynthesizes a diketide at pH 8, but it biosynthesizes a triketide at pH 6 (Fig. 5-1). It remains a mystery how pH affects the product outcome of RpBAS. We decided to undertake the task of exploring conformational dynamics of RpBAS at different pHs. It is the goal of this chapter to elucidate the structural basis of pH effects on the product outcome of RpBAS, thus providing increased detail in how dynamics and pH when coupled can regulate product generation in type III polyketide synthases.

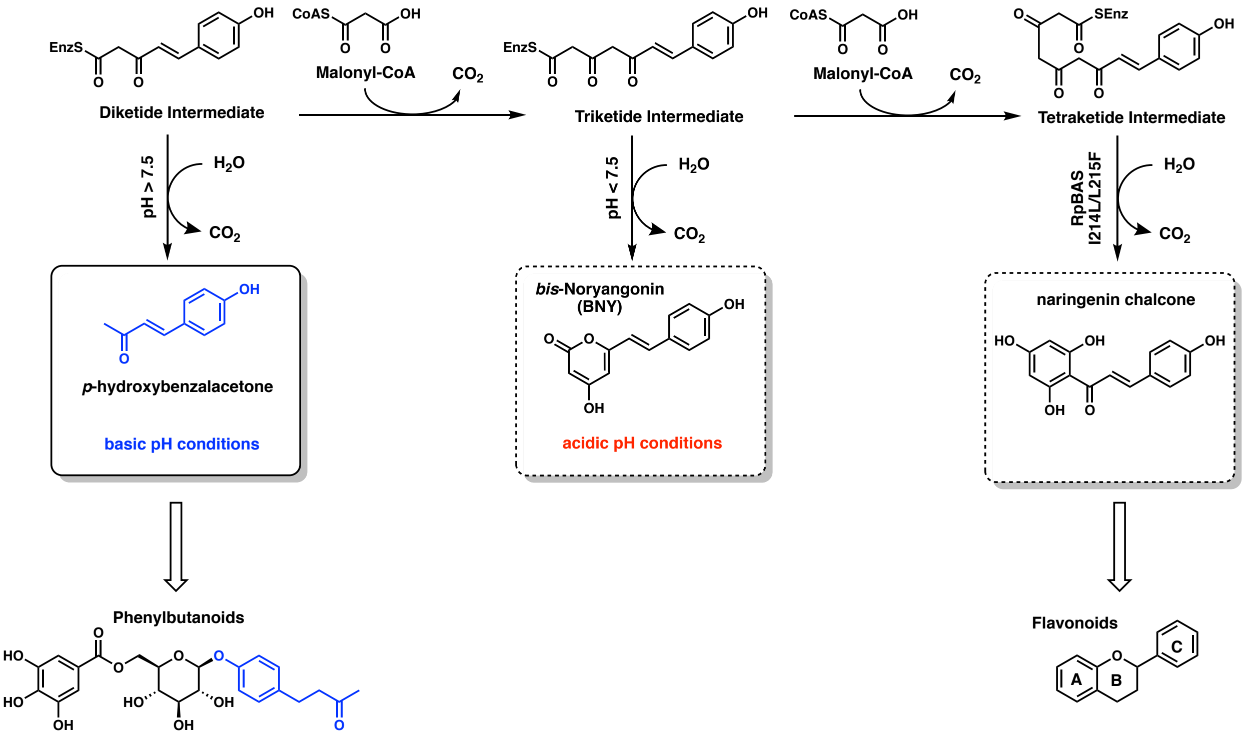


Figure 5-1. RpBAS is responsible for generation of the diketide *p*-hydroxybenzalacetone (BA) at an optimum pH of 8.0, which is provided to downstream enzymes responsible for generating phenylbutanoids. However, at lower pHs, RpBAS is capable of generating the triketide shunt product, *bis*-noryangonin (BNY). MsCHS is responsible for the generation of naringenin chalcone (NAR), a precursor involved in the biosynthesis of flavonoid phytoalexins and anthocyanin pigments. Previous mutagenesis studies have restored the NAR-forming behavior in RpBAS.

To investigate protein structures at a range of pHs structurally is challenging. X-ray crystallography is severely limited due to its difficulty in obtaining a structure at a specific pH, let alone obtaining structures at two or multiple pH values. Protein NMR is a technique that can investigate the pH effects on protein structure and dynamics, though is limited in the range of time scales it is capable of sampling. In this chapter, in the absence of a suitable experimental technique we utilize molecular dynamic simulations to understand how different pH affects the product outcome of RpBAS.

Theories had been abundant for the observed pH effects of RpBAS. For example, lysines and arginines are often involved in electrostatic stabilization of reaction intermediates.[5] The ionization state of lysine and arginine, either in the active site or on the surface, is important for the function and structure of a protein. Nonpolar environments, such as those found in the active site of RpBAS, result in an increase of pKas for acidic residues, and a decrease of pKas for basic residues. A driving force for ionization of charged residues in the active site is the formation of salt bridges.[6] Electrostatic interactions are also known to be an important factor controlling pKa in ionizable amino acids that serve a catalytic role, such as Cys150 in RpBAS. The pKa of the active site cysteine was measured at 5.5 +/- 0.1.[7] pH, along with conjugation and complexation is also important in setting pigment color in plants.[1] All of the above may be important for the observed pH effects on product outcome of RpBAS.

Noel proposed that Cys197, a second cysteine present in the active site of RpBAS might play a role in the decarboxylation reaction of the diketide intermediate, releasing *p*-hydroxybenzalcetone.[7] However, Abe *et al* showed that variants C197T and C197G resulted in no change in the product outcome, thus falsifying the initial Noel hypothesis.[8] Interestingly, this same study revealed a 2-fold decrease in product formation in G256L, and a 2-fold increase in S338V (both G256 and G338 are located in the active site), but both variants produced the triketide pyrone *bis*-noryangonin (BNY) product at pH 8, a reverse of the pH dependence. Another structural investigation on RpBAS was by Shimokawa *et al*, who investigated the role of Leu132 of RpBAS at the entrance of the coumaroyl-CoA binding pocket. This is because in MsCHS, Leu132 is replaced by Thr132, and he hypothesized that this difference may affect the product outcome.[9] Nine L132 variants of RpBAS were screened, and the MsCHS-mimic, L132T, indeed resulted in the production of chalcone by this variant RpBAS. Going further, the 132A, L132S, and L132C variants expanded their product outcome and were capable of biosynthesizing the tetraketide 4-coumaroyltriacetic acid lactone (CTAL). Homology modeling suggested that the expansion of product size in the variants was the result of restoration of the 'coumaroyl binding pocket' in the active site cavity of variant RpBAS. Another mutagenesis study was performed and found that the double variant I214L/L215F of RPBAS predominantly produced BNY, but also produced the tetraketide products naringenin chalcone (NAR) and CTAL.[4] The above structural and functional results are summarized in previous reviews, and are interesting; however, there remains a need to elucidate the molecular basis of pH effect on the product outcome of RpBAS.[10-11]

Here, we present the first MD simulation studies of a type III PKS, which elucidates how the product outcome of RpBAS is affected by pH. Based on the MD simulations, we further generated RpBAS variants that switched the product outcome. Our results pave a foundation for future protein engineering efforts of type III PKS to produce medically useful "unnatural" natural products.

5.3 Results and discussions

The initial hypothesis was that BNY, the triketide product formed at low pHs was due to a conformational change in the active site, which would not have been revealed in the crystal structures from the previously published work. It is also common in type III polyketide synthases to have 'hidden' tunnels, which can be activated or deactivated depending on certain conditions, and that active site volume size can control chain-length specificity.[12]

5.3.1 pH-dependent homology modeling and traditional MD simulations

Initial homology models were generated with the diketide intermediate manually modeled in using Chimera.[13-14] The pH dependence of homology models in the range of 4, 5, 6, 7, 8, 9, 10, and 11 was developed using the H++ server to obtain predicted pKa values and protonation state of titratable residues. The Asp and Glu residues have predicted pKa values near, but less than 4.122 and 4.822, respectively. Therefore, all aspartic and glutamic acid residues were set to deprotonated states above pH 6. Many histidine, cystine, lysine and tyrosine residues have predicted pKa values between 4 and 11, and we applied the predicted values to assign the initial protonation states for all homology models. Shown below in Table 5-1 are the predicted pKa values from the H++ servers, with calculations performed at each pH, and their average value shown. Residues with pKa values below 2.5 and above 11.5 are not shown in the interest of brevity.

The predicted pKa values from the H++ server uses an approach grounded in classical continuum electrostatics with statistical mechanics principles. The calculations did use several approximations to render them computationally feasible. The predicted pKa values of Lys residues are all relatively high, and Lys321 has the estimated pKa value of 9.255. Therefore, all six of the homology models generated at pH 4, 5, 6, 7, 8, and 9 had lysines in the charged form. MD simulations were performed for all homology models, with an RMSD of 1.56 Å observed between the pH 4 and pH 11 models. No major conformational changes were observed over the 500 ns simulations.

Table 5-1. pKa predictions from the H++ server

Res No.	Average pKa	Std Dev
CYS 164	10.840	0.031
CYS 197	**9.714**	**0.023**
HIS 48	6.377	0.006
HIS 71	**5.014**	**0.008**
HIS 95	6.439	0.006
HIS 126	4.629	0.009
HIS 161	**2.963**	**0.011**
HIS 205	5.979	0.013
HIS 251	6.661	0.009
HIS 257	**6.371**	**0.016**
HIS 266	**4.819**	**0.009**
TYR 32	11.179	0.013
TYR 142	9.999	0.010
TYR 268	**11.389**	**0.011**
LYS 55	**9.727**	**0.010**
LYS 57	**10.286**	**0.011**
LYS 78	11.130	0.008
LYS 96	10.453	0.011
LYS 100	10.965	0.006
LYS 107	11.098	0.009
LYS 115	10.741	0.009
LYS 123	11.064	0.007
LYS 146	11.167	0.012
LYS 155	10.551	0.007
LYS 182	10.634	0.012
LYS 269	**10.465**	**0.009**
LYS 280	9.706	0.008
LYS 316	10.518	0.006
LYS 321	9.255	0.010
LYS 325	10.056	0.017
LYS 352	11.376	0.010

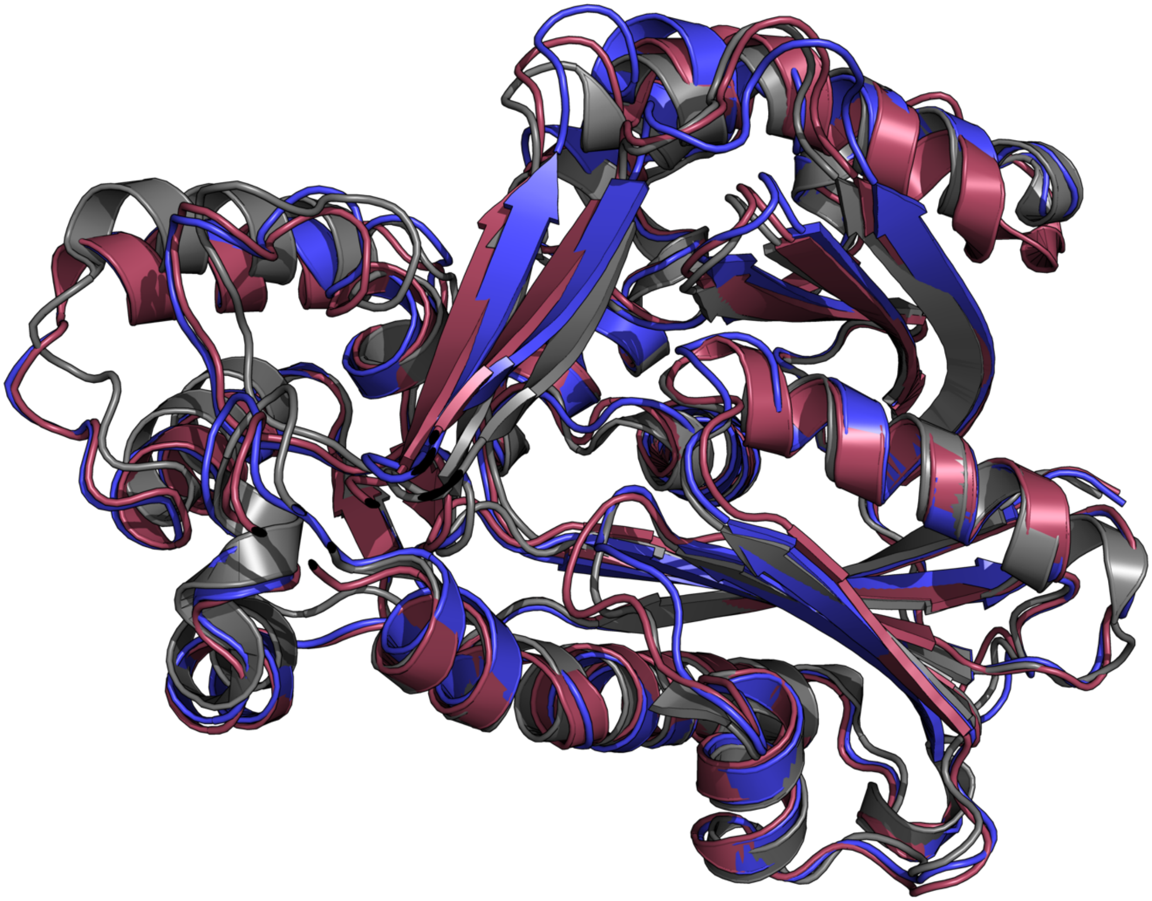


Figure 5-2. Traditional MD simulations were performed on the homology RpBAS models at pHs 4, 5, 6, 7, 8, 9, 10 and 11. Shown above is the initial crystal structure (shown in grey), compared to the average structures generated at pH 4 (red) and pH 11 (blue).

5.3.2 Constant pH simulations

Using the ends of previous trajectories from the initial MD simulations, constant pH simulations were performed on the diketide bound intermediate to calculate the pKa values for residues near the active site. During constant pH simulations that were implemented in AMBER, every 100 steps, a residue was randomly selected, and a change of protonation state was attempted. The constant pH molecular dynamics (CpHMD) method is limited to a total of 50 titratable residues, and RpBAS contains 98 such residues in total. In order to achieve consistent sampling, only 10 residues were selected to be titratable, and we allowed these residues to protonate/deprotonate over the course of the simulations (Fig. 5-3). Residues were selected based on their relative proximity to the active site, and predicted pKa values from the H++ server. In order to determine important regions in the active site, CoA was placed in the active site of RpBAS, using the CoA-bound cocrystal structure of MsCHS (PDB ID: 1BQ6). A total of 22 residues were within 8.00 Å of the monoketide intermediate and CoA (Table S5-1). Of the 22 residues, 10 were selected that had predicted pKas between 5 and 11, which per the H++ server was classified as titratable. These 10 residues included five histidines (H57, H147, H243, H252, and H289), three lysines (K41, K43, K255), one tyrosine (Y254), and one cysteine (C183). No aspartic or glutamic acid residues were selected as protonation of these residues occurs at a much lower pH range. Analysis shows the simulations were relatively stable, though the pH 10 simulation had an issue with the diketide bound intermediate that disrupted the lid region of the receptor.

RMSF calculations were performed for the final 50 ns of each constant pH simulation, and the more acidic pH simulations had greater fluctuations compared to the basic pH simulations (Fig. 5-5). Residues 240 to 270 represent the flexible beta-hairpin region that contains His257 and His266.

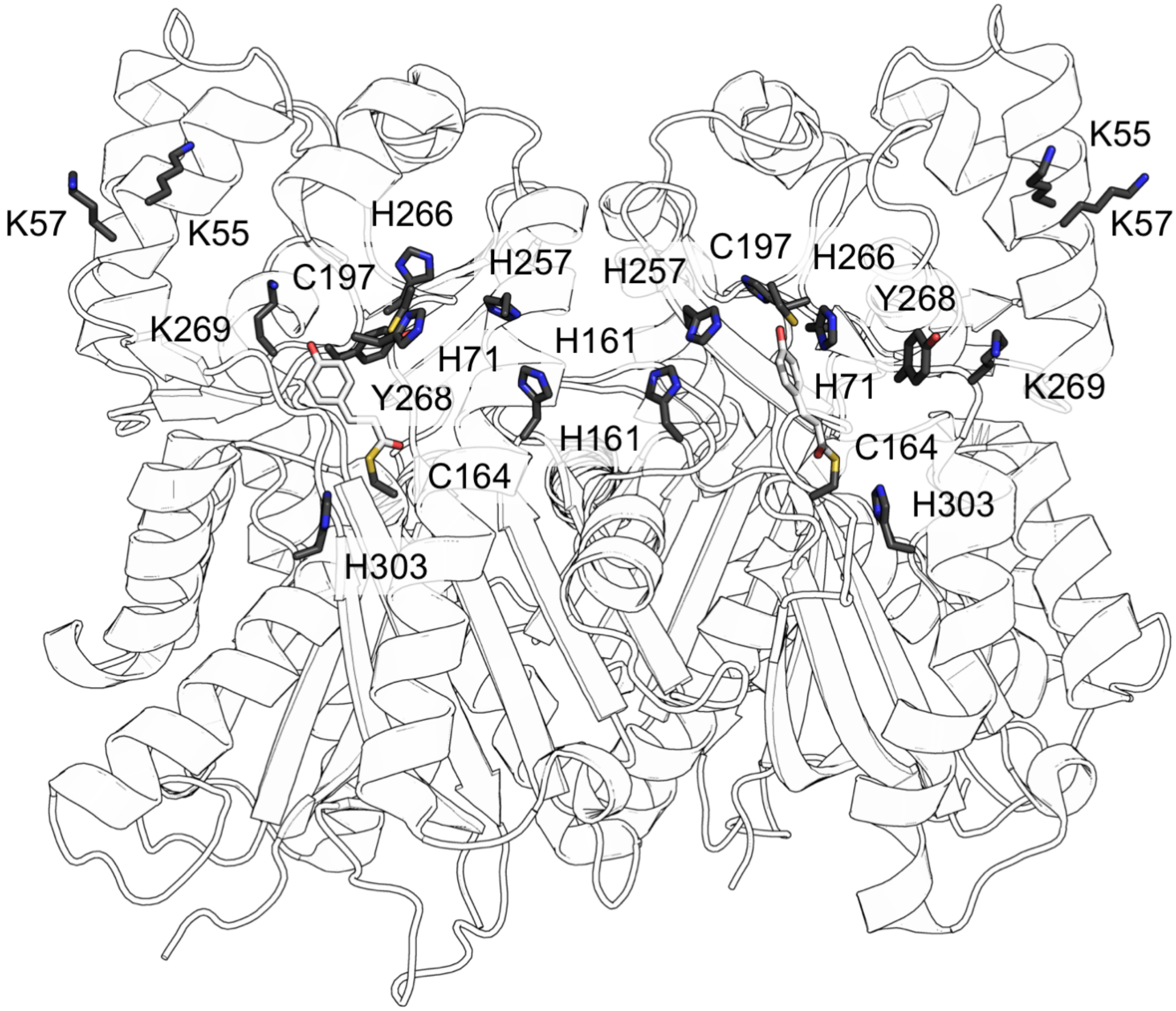


Figure 5-3. Ten residues were selected as titratable residues during the constant pH molecular dynamics simulations and include: K55, K57, H71, H161, C197, H257, H266, Y268, K269, H303. All residues are within 7.5 Å of the monoketide bound intermediate (C164) and Coenzyme A.

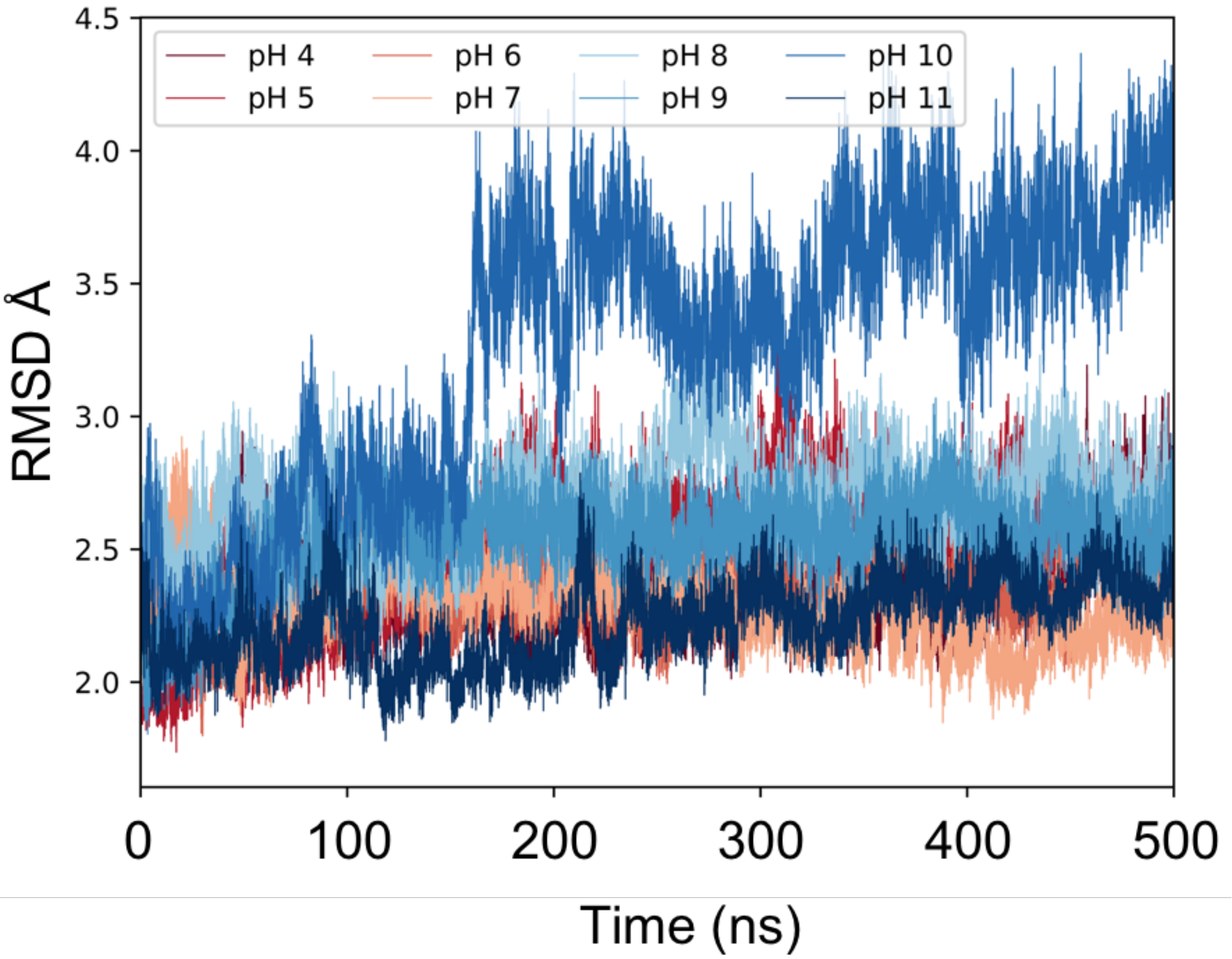


Figure 5-4. RMSD fluctuations over 500 ns of constant pH simulations, showing that the results were relatively stable, with all simulations (except pH 10) maintaining RMSD values below 3.5.

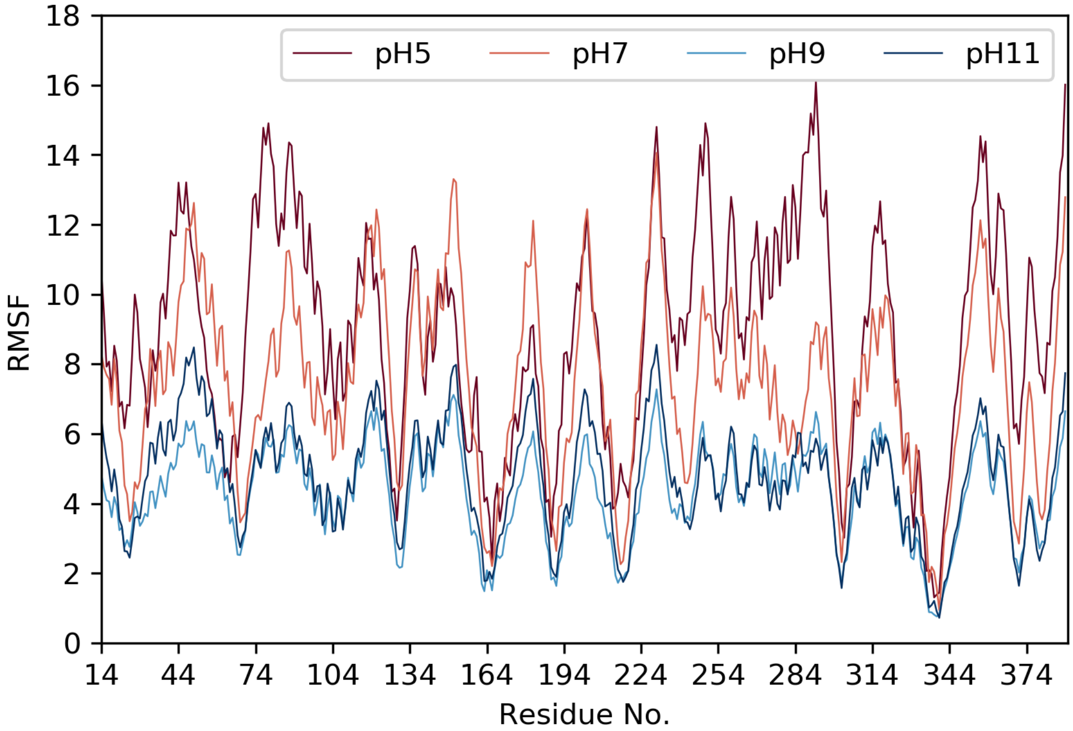


Figure 5-5. RMSF calculations were performed on the final 50 ns of each constant pH simulation.

5.3.3 pKa calculations

The protonation states of sidechains are influenced primarily by electrostatic contributions of protein-solvent and intra-protein interactions.[15] Using the results of the constant pH simulations, we obtained the ratio between protonated and deprotonated states in the constant pH simulations (Fig. 5-6). The second active site cysteine, Cys197, which had previously been predicted to be catalytic, showed a pKa shifted upfield, further falsifying the Noel hypothesis about the its catalytic role in the deprotonated form. Further, careful attention was given to Cys197, because another plausible model for diketide and triketide selection at different pHs could be the deprotonation of the phenol group of the bound intermediate (Fig. S5-1), which may be facilitated by Cys197, although this is doubtful, because our calculations shows that Cys197 stays protonated between pH 6-8 of the simulations.

Significantly, the expected pKa values of two solvent-exposed residues, His257 and His266, were shifted upfield with pKas calculated at 6.81, and 7.29, respectively. At pH 6, His257 is 70 % deprotonated, and at pH 9 it is almost 100% in the neutral protonated form. Similarly, at pH 6, His266 is 90 % deprotonated and charged, and at pH 9 it is almost 100 % in the neutral protonated form. Thus the pH dependence of RpBAS may reflect the protonation states of His257 and His266.

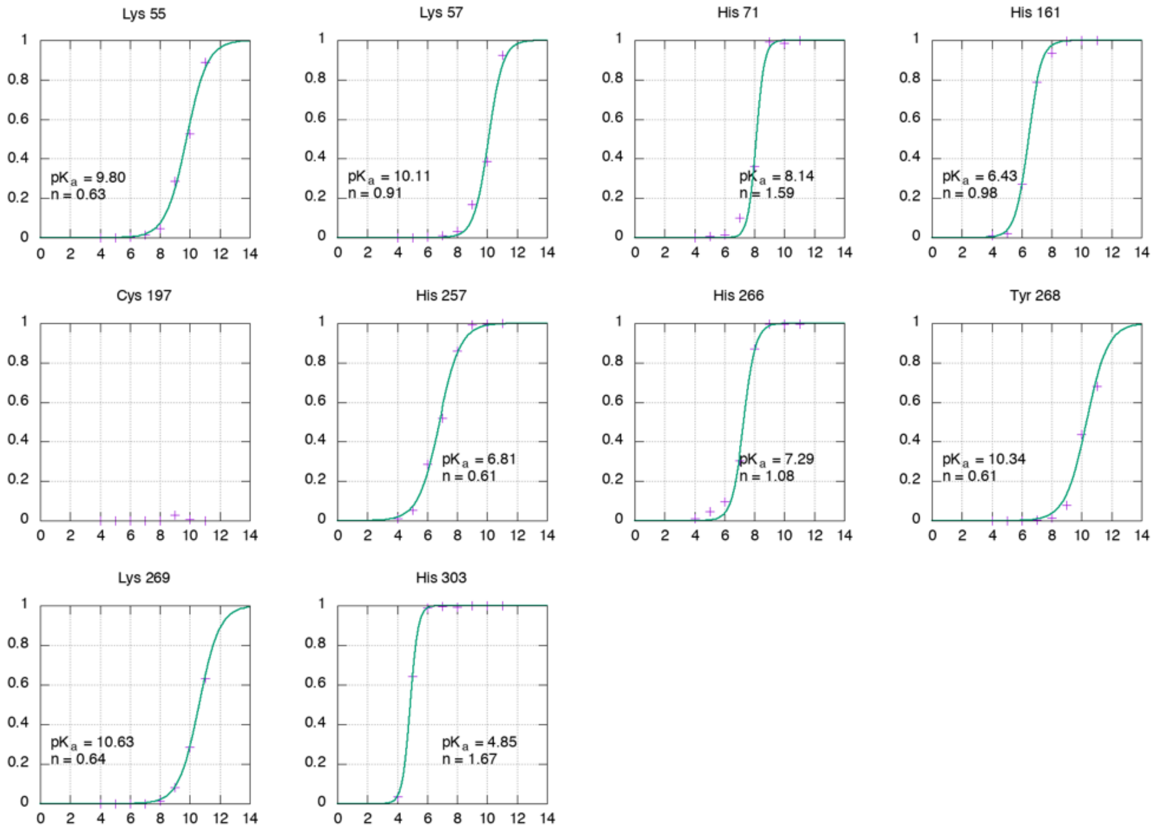


Figure 5-6. pKa calculations of titratable sidechains from constant pH MD simulations.

The above result is in stark contrast to the catalytic residue, His303, which is downshifted to a pKa value of 4.85 in the active site as a result of its interaction with Cys157. H++ server predicted that the pKa of His303 was below 1.5. Therefore, the pH dependence of RpBAS is unlikely linked to the active site His303.

In the original crystal structure of RpBAS, His266 is bound in two different binding motifs in monomers A and B (Fig. 5-7) (PDB ID: 5AQR).[16] Phe265 is known as the gatekeeper residue that is important for guiding the growing intermediate chain (hence selecting diketide versus triketide production). These structural observations, coupled with the above pKa calculations, suggest that His266 is responsible for orienting Phe265 during chain elongation, and this structural effect between His266 and Phe265 may be accomplished by hydrophilic interactions with His257. This explains how a change in the protonation states of His257 and His266 may disrupt this delicate balance of electrostatics.

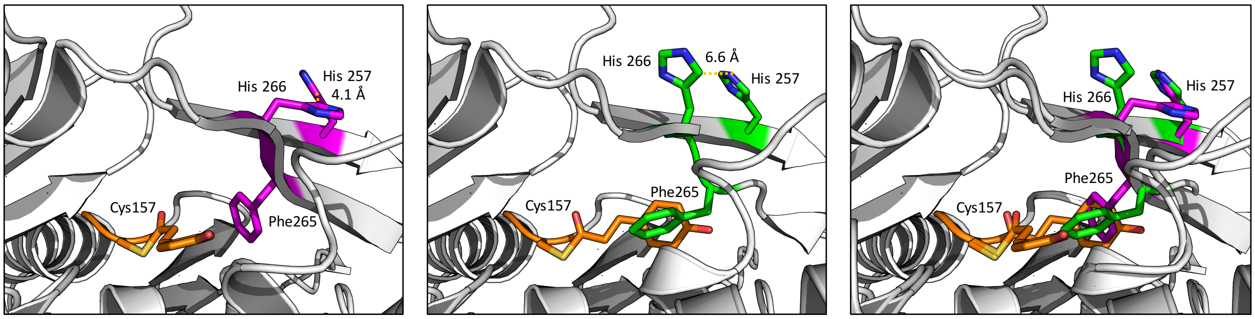


Figure 5-7. A comparison of active sites in RpBAS between chain A (magenta) and chain B (green) from the previously solved crystal structure. Cys157 is loaded with the monoketide intermediate, and Phe265 stabilizes the phenol moiety during chain elongation.

5.3.8 Structural Bioinformatics

To further investigate the role of His257 and His266 as possible pH sensors in RpBAS, bioinformatics analysis was performed. His 257 and His 266 are highly conserved in RpBAS and three other chalcone synthases (FhCHS1, OsCHS, and MsCHS2) (Fig. 5-8). In comparison, His257 is replaced by leucine in AhSTS and lysine in PsSTS, while His266 is replaced by tyrosine in AhSTS and glutamine in GhPYS and PsSTS. It is known that pH also has an effect on the product outcome of chalcone synthases; therefore, similar to the above analysis of RpBAS, the surface-exposed histidines of chalcone synthases (aligned to His257 and His266) may play a similar role in regulating the conformation of active site Phe265, which has been shown to be important for guiding the growing substrate in the active site cavity. The above analysis explains how pH, or protonation states of surface His, may affect the product outcome of type III PKS. To the best of our knowledge, such an observation has never been noted before.

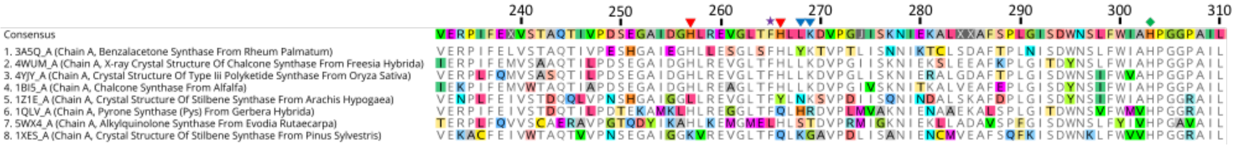

240
250
260
270
280
290
300
310
Consensus
VERPIFEXVSTAQTIVPDSEGAIDGHLREVGLTFHLLKDVPGXISKNIEKALXXAFSPLGISDWNSLFWIAHPGGPAIL
1. 3A5Q_A (Chain A, Benzalacetone Synthase From Rheum Palmatum)
VERPIFELVSTAQTIVPESHGAIEGHLLESGLSFHLYKTVPTLISNNIKTCLSDAFTPLNISDWNSLFWIAHPGGPAIL
2. 4WUM_A (Chain A, X-ray Crystal Structure Of Chalcone Synthase From Freesia Hybrida)
IERPIFEMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLEEAFKPLGITDYNSLFWIAHPGGPAIL
3. 4YJY_A (Chain A, Crystal Structure Of Type Iii Polyketide Synthase From Oryza Sativa)
VERPLFQMVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALGDAFTPLGISDWNSIFWVAHPGGPAIL
4. 1BI5_A (Chain A, Chalcone Synthase From Alfalfa)
IEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNITKALVEAFEPLGISDYNSIFWIAHPGGPAIL
5. 1Z1E_A (Chain A, Crystal Structure Of Stilbene Synthase From Arachis Hypogaea)
VENPLFEIVSTDQQLVPNSHGAIGGLLREVGLTFYLNKSVPDIISQNINDALSKAFDPLGISDYNSIFWIAHPGGRAIL
6. 1QLV_A (Chain A, Pyrone Synthase (Pys) From Gerbera Hybrida)
VERPIFEIVSTDQTILPDTEKAMKLHLREGGLTFQLHRDVPLMVAKNIENAAEKALSPLGITDWNSVFWMVHPGGRAIL
7. 5WX4_A (Chain A, Alkylquinolone Synthase From Evodia Rutaecarpa)
TERPLFQVVSCAERAVPGTQDYIKAHLKEMGMELHLSTDVPRMIGKNIEKLLADAVSPFGISDWNSLFYIVHPGAVAIL
8. 1XES_A (Chain A, Crystal Structure Of Stilbene Synthase From Pinus Sylvestris)
VEKACFEIVWTAQTVVPNSEGAIGGKVREVGLTFQLKGAVPDLISANIENCMVEAFSQFKISDWNKLFWVVHPGGRAIL

Figure 5-8. A multiple sequence alignment between RpBAS (PDB ID: 3A5Q) and homologs: FhCHS1, OsCHS, MsCHS2, AhSTS, GhPYS, ErAQS, and PsSTS (PDB IDs: 4WUM, 4YJY, 1BI5, 1Z1E, 1QLV, 5WX4, 1XES, respectively). H257 and H266 (red triangles) are relatively conserved among homologs, and were set as titratable residues in the constant pH simulations. . F265 is known to be important for stabilizing the growing intermediate (purple star). Y268 and K269 (blue arrows) are also shown and were set as titratable residues in the constant pH molecular dynamics simulations. H303 (green diamond) is known to function as one of the three catalytic residues of the catalytic triad.

5.3.9 Experimental investigations

In order to investigate the effect of protonation on His257 and His266, mutagenesis studies were performed. Several variants were generated, as well as the knockout mutations of active site Cys197 as the negative control. Assays were performed at pH 6, 7, 8 and 9. The product formation was measured as total ion count in the LC-MS elution spectra, with products reported as ratios between the diketide versus triketide products. Interestingly, H257A had shown nearly no production of diketide *p*-hydroxybenzalacetone, while it was able to biosynthesize the triketide *bis*-noryangonin at all pH values tested. The variants only have a small decrease in enzyme turnover as compared to the wildtype. These results support our hypothesis, that His257 is important for orienting Phe265 in the beta-loop region near the active site, such that the alanine mutation of His257 completely switched the product outcome of RpBAS. To the best of our knowledge, this is the first mutation of type III PKS that shows such a drastic switch of pH dependence for its product outcome. The above result supports that RpBAS has the ability to sense its external pH environment, and thereby allosterically modulate its substrate pocket size so that it can control the product outcome at different pHs.

5.4 Materials and methods

5.4.1 Protein expression and purification

The pET28B(+) vector encoding the wild type, full-length RpBAS with an N-terminal His$_{6x}$-tag (Genewiz) and variants were expressed in Top10 Competent Cells (ThermoFisher). After confirmation of DNA sequences, the expression plasmids were transformed and expressed in BL21 (DE3) *E. coli* cells (Novagen). Cells transformed with the RpBAS plasmid were grown to OD$_{600}$ = 0.8 at 310 K in LB media containing 50 μg/mL kanamycin. The cell cultures were cooled to 291 K and protein expression was induced using 100.0 ⍰M of IPTG. The cell cultures were incubated for an additional 16 hours at 291 K and harvested by centrifugation at 5,000*g* for 30 minutes. The cell pellets were resuspended in 50 mM tris-HCl pH 8.0, 5 %(*v/v*) glycerol, 7 mM imidazole, and 200 mM NaCl. Resuspended cells were flash frozen in liquid nitrogen and stored at -80 °C for 24 hours. Resuspended cells were thawed on ice for 30 min and the cells were disrupted using sonication. The cell debris was cleared by centrifugation at 14,100*g* for 1 hour. The supernatant was collected, and batch bound to HisPur™ Cobalt Resin (Thermo Scientific) for 30 minutes at 4°C. His-tag was cleaved using 1 unit thrombin / 2 mg RpBAS protein over 4 hours at 16 C. Successful cleavage measured by LC-MS at 1, 2 and 4 hour timepoints to ensure complete clevage. Pre-histag cleavage mass was 44,390.00 Da, and post-histag cleavage mass was 42,640.00, with no presence of his-tagged protein in mixture. of RpBAS was purified according to the manufacturer's instructions using an imidazole step-gradient. Fractions containing pure protein were determined by SDS-PAGE and fractions containing RpBAS were combined and dialyzed against 50mM tris-HCl pH 8.0, 5%(*v/v*) glycerol, and 2 mM dithiothreitol at 4 °C for 12 hours. Further purification of RpBAS was conducted by anion exchange chromatography using HiTrap Q FF (GE Healthcare) according to the manufacturer's instructions. Purified RpBAS was dialyzed against 20 mM HEPES-NaOH pH 8.0, 100 mM NaCL, and 2 mM dithiothreitol. The protein solution was concentrated to 20 mg/mL and filtered by 0.22 µM Ultrafree-MC filter (Millipore).

5.4.2 Site-directed mutagenesis

The online program NEBaseChanger (https://nebasechanger.neb.com/) was used to design primers for all the RpBAS variants using the previously described cloned pET28B(+) w.t. RpBAS DNA template. The following primers were used for mutagenesis. All mutations were confirmed though automated DNA sequencing (Genewiz/Retrogen).

Primer Description	Primer Sequence
H71 (general reverse)	5'- TTCTCAATACGGCTGTTC

H71A (forward)	5’- GCGCTATTTAgccGTGACCGAGGAAATTC
H71F (forward)	5’- GCGCTATTTAttcGTGACCGAGGAAATTC
H71I (forward)	5’- GCGCTATTTAatcGTGACCGAGGAAATTC
H71K (forward)	5’- GCGCTATTTAaagGTGACCGAGG
H71V (forward)	5’- GCGCTATTTAgtcGTGACCGAGGAAATTC
H71W (forward)	5’- GCGCTATTTAtggGTGACCGAGGAAATTC
H71Y (forward)	5’- GCGCTATTTAtacGTGACCGAGG
H71Q (special reverse)	5’- TTCTCAATACGGCTGTTCTC
H71Q (forward)	5’- GCGCTATTTAcaaGTGACCGAGG
C157 (reverse)	5’- TAGAACATGAAGCGCTTC
C157 (forward)	5’- TCATTTAGGTgctTATGCCGGCG
H161 (general reverse)	5’- AAGCGCTTCACGCTCGGA
H161A (forward)	5’- CATGTTCTATgccTTAGGTTGTTATGCCGGCGG
H161F (forward)	5’- CATGTTCTATttcTTAGGTTGTTATGCCGGCGG
H161I (forward)	5’- CATGTTCTATatcTTAGGTTGTTATGCCGGCGGTACAGTGC
H161K (forward)	5’- CATGTTCTATaagTTAGGTTGTTATGCCGGC
H161Q (forward)	5’- CATGTTCTATcagTTAGGTTGTTATGCCG
H161V (forward)	5’- CATGTTCTATgtcTTAGGTTGTTATGCCGGCGG
H161W (forward)	5’- CATGTTCTATtggTTAGGTTGTTATGCCGGCGG
H161Y (forward)	5’- CATGTTCTATtacTTAGGTTGTTATGCCGG
H257 (general reverse)	5’- GCGCCATGGCTTTCCGGA
H257A (forward)	5’- CATTGAGGGTgccCTGTTAGAGAGCGGTTTAAGTTTTC
H257V (forward)	5’- CATTGAGGGTgtcCTGTTAGAGAGCGGTTTAAGTTTTC

H257W (forward)	5’- CATTGAGGGTtggCTGTTAGAGAGCGGTTTAAGTTTTC	
H257Y (forward)	5’- CATTGAGGGTtacCTGTTAGAGAGCGGTTTAAG	
H266 (general reverse)	5’- CCGCTCTCTAACAGATGAC	
H266A (forward)	5’- TTTAAGTTTTgcgTTATACAAGACCGTGCCGACAC	"
H266V (forward)	5’- TTTAAGTTTTgttTTATACAAGACCGTGCC	
H266W (forward)	5’- TTTAAGTTTTtggTTATACAAGACCGTGCCG	

5.4.3 Enzyme Reaction

Reaction mixtures contained 27 μM 4-coumaroyl-CoA, 54 μM malonyl-CoA, 20 μg of purified enzyme in a 500 μL volume of 100 mM potassium phosphate buffer, and 1 mM EDTA. Reactions were performed at pHs 6.0, 7.0, 8.0 and 9.0. Reaction mixtures were incubated for 30 minutes and at 30°C and were quenched by the addition of 50 μL 20% HCl. Reaction products were extracted with 600 μL ethyl acetate, concentrated via Eppendorf Vacufuge, and dissolved in 200 μL UPLC-MS grade acetonitrile.

5.4.4 pH-dependent Receptor Modeling

Five receptors were modeled, including *apo* RpBAS, and RpBAS bound with one of the four intermediates (monoketide, diketide, triketide and tetraketide). The *apo* and monoketide-bound structures were obtained from the crystal structures (PDB code: 3A5Q and 3A5R). The diketide, triketide and tetraketide structures were modeled using the monoketide crystal structure and docking the substrates in the active site. To minimize initial steric clashes, the diketide, triketide and tetraketide intermediates were assigned AM1-BCC charges and minimized using Chimera. The bound intermediates were then excised and capped with acetyl and N-methyl caps, followed by geometry optimization, frequency calculations, single point energy calculations, and electrostatic potential (ESP) calculations using B3LYP/6-31G(d,p)//B3LYP/6-31G(d,p) by Gaussian 09 Rev E. Using the ESP derived from Gaussian, restricted ESP (RESP) charges were fit using the R.E.D. server. Initial protonation states of titratable residues were set using the H++ server at pHs: 4, 5, 6, 7, 8, 9, 10, and 11. The topology and coordinate files generated by H++ were converted to pdb using ambpdb from the AMBER suite.

5.4.5 Ligand Parameterization and Modeling

Six ligands were manually constructed using Chimera, which include 4-coumaroyl-CoA, malonyl-CoA, benzalacetone (BA), bis-noryangonin (BNY), 4-coumaroyltriacetic acid lactone (CTAL), and naringenin chalcone (NAR). Using these initial structures, we applied Gaussian 09 Rev E B3LYP/6-31G(d,p)//B3LYP/6-31G(d,p) to calculate geometry optimization, frequency calculations, single point energy calculations, and electrostatic potential (ESP). Because no RpBAS crystal structure in the PDB contained malonyl-CoA, the ligand was extracted from the malonyl-CoA bound MsCHS structure (PDB ID: 1CML). Coumaroyl-CoA was manually modeled using the CoA moiety from malonyl-CoA, because no type III PKSs in the PDB contain coumaroyl-CoA.

5.4.6 Constant pH Molecular Dynamics Simulations

Molecular dynamic simulations were carried out using Amber 14.[17] The Amber ff14SB force field with the constant pH library and constant pH modifications file were used.[18] K41, K43, H57, H147 C183, H243, H252, K255, Y254, and H289 were selected to be titratable residues, and were allowed to change protonation states during constant pH MD (cpHMD) simulations. LEaP was used to add hydrogens and neutralize the system by adding Na+ and Cl- ions dependent on pH. Systems were solvated in a 12 Å water buffer TIP3P box. Energy minimizations using SANDER were performed in two stages to remove any steric clashes present in the initial crystal structure. The initial stage was carried out over 2500 steps for the solvent and ions, with the protein and ligand restrained by a force constant of 500 kcal/mol/Å^2, followed by a second stage carried out over 5000 steps of the entire system. A short 100 ps simulation with weak restraints (force constant of 10 kcal/mol/Å2 on the protein and cofactor) was used to heat up the system to 300 K using a Langevin temperature equilibration scheme. Periodic boundary conditions were used, along with a non-bonded interaction cutoff of 10 Å. For the simulations, hydrogen atoms were constrained using the SHAKE algorithm, allowing for a 2 fs time step. Initial simulations were run over 500 ns (250,000,000 time steps) to equilibrate the systems prior to constant pH simulations. During constant pH simulations, the salt concentration was set to 100 mM, with 100 steps of relaxation dynamics, and 100 steps of in-between attempted protonation state changes. The evaluation of protonation attempts was performed with the generalized Born model (igb = 2 in AMBER), with a total of 2,500,000 attempted protonation state changes. Constant pH simulations were run over 500 ns (250,000,000 time steps).

5.4.7 Structural Bioinformatics

Using the FASTA sequence from *Rheum palmatum* BAS (UniProtKB accession number: Q94FV7), RpBAS was subjected to a BLAST search using Geneious with the blastp algorithm, with a BLOSUM62 matrix.[19]

5.4.8 LC-MS

The HPLC system utilized was a Waters Acquity H UPLC Class (Waters Corporation, Milford, MA, USA). The analytical column was a 1.7 µm x 2.1 mm x 50 mm Water Acquity UPLC BEH C18 column (Waters Corporation, Milford, MA, USA). The mobile phase was water:acetonitrile (v/v) with 0.1 % formic acid, and was delivered at a flow rate of 500 µL/min over five minutes. A single quadrupole Waters QDA Detector (Waters Corporation, Milford, MA, USA) was used in positive polarity ESI mode. Masses were detected on the 100-600 m/z range. Benzalacetone had a retention factor of 1.47 min, with the m/z of parent peak of 163.0. Nitrogen was used as a source gas and maintained at 8885 psi. Data acquisition and processing was performed using Waters MassLynx 4.0 (Waters Corporation, Milford, MA, USA).

4.4.9 Synthesis of p-coumaroyl-CoA

p-coumaroyl-CoA was chemically synthesized as previously described by Stöckigt *et al*.[20] The coumaroyl-ester intermediate was verified by 1H NMR and mass spec (m/z: 284.2 (parent peak, M + Na+), 285.2, 279.4). The final product, p-coumaroyl-CoA was verified by 1H NMR and mass spec (m/z: 914.14 (parent peak), 407.21, 476.69, 834.20, 768.0).

4.5 Conclusions

In this chapter, we analyzed the pH dependence of RpBAS using computational biology and molecular biology. This is new for PKS, but we had found precedents for enzymes unrelated with natural product biosynthesis. For example, the β-amyloid precursor protease (BACE1) has received much attention as a potential target against Alzheimer's disease.

Butler *et al.* developed a potent BACE1 inhibitor, which unfortunately had equal affinity for cathepsin D (CatD) at high pH, though at low pH it preferred to bind BACE1.[21] In order to investigate the effect of pH on inhibitor-BACE1 binding affinity, Harris *et al* performed constant pH simulations similar to those performed in this chapter.[22] They also identified a histidine residue, which is capable of allosterically modulating conformation changes in BACE1. They combined the constant pH approach with free-energy perturbation (FEP) to predict changes in binding free energy. Binding events between receptors and ligands can also shift pKas upward and downward, which has an effect on binding and stabilization at certain pH values. In BACE1, His45 modulated the conformation of Phe105 through pi-pi interactions in a pH-dependent manner.

Another study was performed by McDougal *et al* using ^{1}H NMR and constant pH MD simulations to determine pKa values of histidine residues in α-Conotoxin MII peptides.[23] NMR titration experiments were used to measure the pKa values experimentally, and were in consistent agreement with calculated pKa values from constant pH calculations, with an overall median absolute deviation (MAD) of 0.3 Å. We learned from these examples about the insights that constant pH MD simulations can offer about protein dynamics that govern the observed pKa values of enzyme residues.

In this Chapter, we found that His257 and His266 of RpBAS regulate the stability and protein dynamics of the beta-hairpin region. The protonation states of His257 and His266 help orient the gatekeeper residue, Phe265, which in term affects the product outcome to diketide versus triketide. A decrease in pH increases the flexibility in this region, as observed from RMSF fluctuations, and thereby alters the product profile as observed in the mutagenesis studies. In summary, this work paves the foundation for future engineering studies, which could combine the pH switching mechanism observed in RpBAS and other types III PKSs with their ability to accept un-natural starter units, so that new “unnatural” natural products can be biosynthesized.

4.7 Supplementary Tables and Figures

Table S5-1. Description of residue function from previous studies.

RpBAS		
Res No.	Res Name	Notes
55	Lys	Form h-bonds w/ phosphates of CoA
58	Arg	Form h-bonds w/ phosphates of CoA
62	Asn	Lys in MsCHS
71	His	Predicted pKa = 8.14
131 - 140	Loop: CLAGVDMPGA	Flexibility of the loop structure may be important for catalytic activity
132	Leu	L132G, L132A, L132S, L132C, L132T, L132F, L132Y, L132W, L132P mutants. L132T - restored chalcone-forming activity in BAS. L132A/S/C produced CTAL. Homology modeling suggested this was the result of restoration of the 'coumaroyl binding pocket' in the active-site cavity. Km value similar to double mutant, but decrease in kcat. this suggests no change in binding, but in catalysis. Triple mutant (L132T/I214L/L215F) showed no improvement in chalcone-forming activity, but resulted in loss of activity. L132T active site is large enough to support tetrakertide as was tested with other starter units.
133	Gly	Ser in MsCHS
137	Met	cyclization pocket. 4.0 Å from neighboring chain active site monoketide intermediate
161	His	Predicted pKa = 6.43.
164	Cys	Conserved in all CHS-like. Defines active site. (pKa H++ : 10.8)
197	Cys	C197T, C197G - no change in product pattern. Known Coumaroyl-CoA binding pocket. Was previously suggested to be a possible catalytic residue.
205	His	on outside, but highly conserved and close to His257/His 266 (neighbors). predicted pKa = 6.0
214	Ile	Critical role in diketide formation reaction
215	Leu	Critical role in diketide formation reaction
240-270	Beta hairpin VPESHGAIEGHLLESGLSFHLYKT	MD predicted this is pretty flexible. No trps here, but one proline (P233)
254	Ile	cyclization pocket (conserved)
256	Gly	cyclization pocket (conserved). G256L - reduced product formation by half
257	His	on outside, but highly conserved and close to His266. predicted pKa = 6.81
265	Phe	cyclization pocket (conserved)
266	His	on outside, but highly conserved and close to His257. predicted pKa = 7.29
303	His	Catalytic triad in CHS-like enzymes. His 303 most likely acts as a general base during the generation of a nucleophilic thiolate anion from Cys 164 (PTM). Predicted pKa = 4.85.
336	Asn	Conserved in all CHS-like. Defines active site. May function in the decarboxylation reaction.
338	Ser	Critical role in diketide formation reaction (conserved). S338V - increased activity 2-fold.
375	Pro	cyclization pocket (conserved)

Table S5-2. Table for residue selection

	Res	No.	Distance from Cys157	H++ pKa	cpH pKa	Expt. pKa	Notes
*	Cys	164	2.0 Å	10.8	*	5.5	The pKa of the catalytic cysteine in BAS must shift below 7.0 to serve as an effective nucleophile.
1	Lys	269	2.0 Å	10.5	10.630	-	K in seven sequences: RpBAS, FHCHS1, OsPKS, MsCHS, AsSTS, CmQNS, and AsPCS
	Glu	192	3.0 Å	0.0			Conserved in all 10 sequences. Present in Coumaroyl-CoA binding pocket. Known to bind Coumaroyl-CoA
	Asp	217	3.0 Å	0.0			Conserved in all 10 sequences
2	Cys	197	4.0 Å	9.7	>11.00	-	C in RpBAS, and T in FhCHS1, OsPKS, MsCHS, AsSTS, CmQNS, and PsSTS. Present in Coumaroyl-CoA binding pocket. Known to bind Coumaroyl-CoA
	Asp	207	4.0 Å	0.1			Conserved in first 9 sequences
	Glu	255	4.0 Å	2.4			D in FhCHS1 and MsCHS, E in RpBAS and CmQNS. 6 different residues among 10 sequences
3	Tyr	268	4.0 Å	11.4	10.340	-	T in RpBAS, L in RhCHS1, OsPKS, and MsCHS. This residue does not appear very conserved. 7 different residues.
4	Lys	55	5.0 Å	9.7	9.800	-	Conserved in all 10 sequences
	Tyr	165	5.0 Å	12.0			F in FhCHS1, OsPKS, MsCHS, AsSTS, PsSTS, OsCUS. Y in RpBAS, CmQNS, AsPCS
5	His	257	5.0 Å	6.4	6.810	-	Conserved in 7 sequences. L in AsSTS, K in PsSTS, and R in OsCUS. Interface residue
6	His	266	5.0 Å	4.8	7.290	-	H in RpBAS, FhCHS1, OsPKS, MsCHS, AsSTS, and CmQNS
7	His	303	6.0 Å	0.0	4.850	-	Catalytic Triad Residue. Conserved in all 10 sequences
	Asp	311	6.0 Å	0.8			Conserved in all 10 sequences
	Cys	60	7.0 Å	12.0			Conserved in all 10 sequences
	Glu	61	7.0 Å	3.5			E in RpBAS, FhCHS1, AsSTS, CmQNS, GhPYS, PsSTS. D in OsPKS and MsCHS
	Tyr	69	7.0 Å	11.6			Y in 8 sequences. H in AsSTS and CmQNS
8	His	161	7.0 Å	3.0	6.430	-	H in RpBAS and AsSTS. Q in FhCHS1, OsPKS, MsCHS, CmQNS, GhPYS, and PsSTS. AsSTS and PsSTS appear to have a HQ and QH flip at residues 154 & 155
	Cys	131	7.2 Å	12.0			C in RpBAS. T in FhCHS1, OsPKS, MsCHS, AsSTS, CmQNS, GhPYS
9	His	71	7.6 Å	5.0	8.140	-	H in RpBAS and OsPKS. Y in FhCHS1 and MsCHS
10	Lys	57	7.75 Å	10.3	10.110	-	K in 7 sequences including RpBAS. Q in MsCHS
	Tyr	334	7.9 Å	12.0			Conserved in all 10 sequences
	His	205	8.0 Å	6.0			Conserved in 8 sequences (D in AsSTS, C in OsCUS)
	Cys	190	8.75 Å	disulfide			disulfide
	Cys	130	8.8 Å	disulfide			disulfide
	Lys	53	9.0 Å	12.0			Conserved in all 10 sequences
	Tyr	160	10.0 Å	12.0			Conserved in 8 sequences. F in PsSTS, and H in OsCUS
	Glu	380	10.7 Å	0.0			Conserved in first 9 sequences
	Tyr	86	10.8Å	12.0			Conserved in 7 sequences (F in GhPYS and PsSTS, deleted in OsCUS)

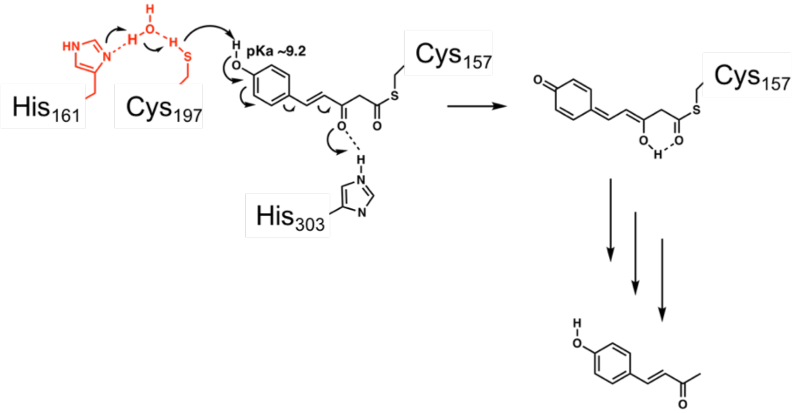


Figure S5-1. The second active site cysteine may deprotonate the diketide intermediate, though deprotonation was unobserved during constant pH simulations, casting doubt on this alternative model.

References

1. Austin, M. B.; Noel, J. P., The chalcone synthase superfamily of type III polyketide synthases. *Nat Prod Rep* **2003,** *20* (1), 79-110.

2. Ferrer, J. L.; Jez, J. M.; Bowman, M. E.; Dixon, R. A.; Noel, J. P., Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis. *Nat Struct Biol* **1999,** *6* (8), 775-84.

3. Abe, I.; Takahashi, Y.; Morita, H.; Noguchi, H., Benzalacetone synthase. A novel polyketide synthase that plays a crucial role in the biosynthesis of phenylbutanones in Rheum palmatum. *Eur J Biochem* **2001,** *268* (11), 3354-9.

4. Abe, I.; Sano, Y.; Takahashi, Y.; Noguchi, H., Site-directed mutagenesis of benzalacetone synthase. The role of the Phe215 in plant type III polyketide synthases. *J Biol Chem* **2003,** *278* (27), 25218-26.

5. Kessel, A.; Ben-Tal, N., *Introduction to proteins : structure, function, and motion*. CRC Press: Boca Raton, FL, 2011; p xxvii, 626 p.

6. Harris, T. K.; Turner, G. J., Structural basis of perturbed pKa values of catalytic groups in enzyme active sites. *IUBMB Life* **2002,** *53* (2), 85-98.

7. Jez, J. M.; Noel, J. P., Mechanism of chalcone synthase. pKa of the catalytic cysteine and the role of the conserved histidine in a plant polyketide synthase. *J Biol Chem* **2000,** *275* (50), 39640-6.

8. Abe, T.; Morita, H.; Noma, H.; Kohno, T.; Noguchi, H.; Abe, I., Structure function analysis of benzalacetone synthase from Rheum palmatum. *Bioorg Med Chem Lett* **2007,** *17* (11), 3161-6.

9. Shimokawa, Y.; Morita, H.; Abe, I., Structure-based engineering of benzalacetone synthase. *Bioorg Med Chem Lett* **2010,** *20* (17), 5099-103.

10. Shimokawa, Y.; Morita, H.; Abe, I., Benzalacetone synthase. *Front Plant Sci* **2012,** *3*, 57.

11. Abe, I.; Morita, H., Structure and function of the chalcone synthase superfamily of plant type III polyketide synthases. *Natural Product Reports* **2010,** *27* (6), 809-838.

12. Morita, H.; Kondo, S.; Oguro, S.; Noguchi, H.; Sugio, S.; Abe, I.; Kohno, T., Structural insight into chain-length control and product specificity of pentaketide chromone synthase from Aloe arborescens. *Chemistry & Biology* **2007,** *14* (4), 359-369.

13. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-12.

14. Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C. C.; Pettersen, E. F.; Goddard, T. D.; Meng, E. C.; Sali, A.; Ferrin, T. E., UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **2012,** *179* (3), 269-78.

15. Stanton, C. L.; Houk, K. N., Benchmarking pK(a) prediction methods for residues in proteins. *Journal of Chemical Theory and Computation* **2008,** *4* (6), 951-966.

16. Morita, H.; Shimokawa, Y.; Tanio, M.; Kato, R.; Noguchi, H.; Sugio, S.; Kohno, T.; Abe, I., A structure-based mechanism for benzalacetone synthase from Rheum palmatum. *Proc Natl Acad Sci U S A* **2010,** *107* (2), 669-73.

17. D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R.

Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman, AMBER 14. University of California, San Francisco: 2014.

18. Swails, J. M.; York, D. M.; Roitberg, A. E., Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *Journal of Chemical Theory and Computation* **2014,** *10* (3), 1341-1352.

19. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; Thierer, T.; Ashton, B.; Meintjes, P.; Drummond, A., Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012,** *28* (12), 1647-1649.

20. Stockigt, J.; Zenk, M. H., Chemical Syntheses and Properties of Hydroxycinnamoyl Coenzyme a Derivatives. *Z Naturforsch C* **1975,** *30* (3), 352-358.

21. Butler, C. R.; Brodney, M. A.; Beck, E. M.; Barreiro, G.; Nolan, C. E.; Pan, F.; Vajdos, F.; Parris, K.; Varghese, A. H.; Helal, C. J.; Lira, R.; Doran, S. D.; Riddell, D. R.; Buzon, L. M.; Dutra, J. K.; Martinez-Alsina, L. A.; Ogilvie, K.; Murray, J. C.; Young, J. M.; Atchison, K.; Robshaw, A.; Gonzales, C.; Wang, J. L.; Zhang, Y.; O'Neill, B. T., Discovery of a Series of Efficient, Centrally Efficacious BACE1 Inhibitors through Structure-Based Drug Design. *J Med Chem* **2015,** *58* (6), 2678-2702.

22. Harris, R. C.; Tsai, C. C.; Ellis, C. R.; Shen, J., Proton-Coupled Conformational Allostery Modulates the Inhibitor Selectivity for beta-Secretase. *J Phys Chem Lett* **2017,** *8* (19), 4832-4837.

23. McDougal, O. M.; Granum, D. M.; Swartz, M.; Rohleder, C.; Maupin, C. M., pK(a) Determination of Histidine Residues in alpha-Conotoxin MII Peptides by H-1 NMR and Constant pH Molecular Dynamics Simulation. *Journal of Physical Chemistry B* **2013,** *117* (9), 2653-2661.

24. Shen, Y. L.; Li, X.; Chai, T. Y.; Wang, H., Outer-sphere residues influence the catalytic activity of a chalcone synthase from Polygonum cuspidatum. *Febs Open Bio* **2016,** *6* (6), 610-618.

25. Guo, H. L.; Yang, Y. D.; Ma, Y. D.; Liu, W. B.; Feng, J.; Luo, Z. Q.; Lu, H. S.; Liu, C. M.; Yang, M. F.; Wang, Y. N.; Ma, L. Q., A bifunctional type III polyketide synthase from raspberry (Rubus idaeus L.) with both chalcone synthase and benzalacetone synthase activity. *J Plant Biochem Biot* **2017,** *26* (1), 80-90.

26. Abe, I.; Morita, H.; Nomura, A.; Noguchi, H., Substrate specificity of chalcone synthase: Enzymatic formation of unnatural polyketides from synthetic cinnamoyl-CoA analogues. *Journal of the American Chemical Society* **2000,** *122* (45), 11242-11243.

27. Morita, H.; Takahashi, Y.; Noguchi, H.; Abe, I., Enzymatic formation of unnatural aromatic polyketides by chalcone synthase. *Biochem Bioph Res Co* **2000,** *279* (1), 190-195.

28. Abe, I.; Takahashi, Y.; Noguchi, H., Enzymatic formation of an unnatural C-6-C-5 aromatic polyketide by plant type III polyketide synthases. *Org Lett* **2002,** *4* (21), 3623-3626.

29. Abe, I.; Abe, T.; Wanibuchi, K.; Noguchi, H., Enzymatic formation of quinolone alkaloids by a plant type III polyketide synthase. *Org Lett* **2006,** *8* (26), 6063-6065.

CHAPTER 6

Conclusions and Future Directions

6.1 Conclusions

Enzymes in the acetate and non-ribosomal peptide pathways biosynthesize valuable natural products in an assembly line fashion. These natural products are difficult to access through organic synthesis because of their chemical complexity, innate reactivity, and multiple stereocenters. Current and future engineering efforts have focused on the ability to add, delete or switch out starter and extender units in these pathways with minimal success. Genomics and bioinformatics have provided us with tools that can predict the product outcome of megasynthases, though we are still in the dark about the driving force for the protein-protein and protein-substrate interactions in these systems. Such a knowledge gap has hampered protein engineering efforts of megasynthases.[1] In order to address this gap in knowledge, I have developed a molecular mechanics force field for simulating enzymes in the acetate and non-ribosomal peptide pathways in which their intermediate products, often highly reactive, are chauffeured between the catalytic domains via carrier proteins. The force field, when implemented using MD, can perform simulations of megasynthases with atomic details, generate models of transient interactions in these megasynthases, and provide crucial information to engineers.

In Chapter 2, the methodology and validation of the pantetheine force field was described in detail. Primary force fields are limited to modeling standard monomeric units in macromolecules, such as amino acids, nucleic acids, sugars, as well as small molecules.[2-3] Specialized force fields, which are capable of modeling special cases, are available for NCAAs, phosphorylated amino acids, and more common PTMs.[4-8] The force field that I have developed is capable of modelling FASs, PKSs, and NRPSs in which starter units, extender units and intermediates are linked to a phosphopantetheine prosthetic group. Previously, in order to perform simulations on these megasynthases, intensive electronic structure calculations and optimizing parameters are needed. This new force field therefore provides accessibility to performing MD simulations of megasynthases, with the additional benefit of providing end-users with reproducible parameters. The force field was compared to experimental structural data published in the literature and showed high consistency with experimental data.

In Chapter 3, the force field was applied to the NRPS reductase domain (MxaR) from the slime bacteria *Stigmatella aurantiaca.* MxaR performs a four-electron reduction using two NADPH molecules to generate a hybrid polyketide-nonribosomal peptide product. Myxalamids serve as inhibitors of the electronic transport chain, with the potential to be used as an anti-cancer therapeutic, or as a biofuel additive.[9] MD based models identified residues important for protein-substrate interactions in the R domain. Biochemical engineering efforts, which were based on the MD models, resulted in the design of a mutant reductase domain capable of reducing the non-native substrate decanoyl-PCP to 1-decanol, a potential biofuel replacement. Specifically, a R1339A mutant accelerated the product turnover of decanoyl-PCP and decanal compared to the wildtype. This study combined crystallography, biochemistry and molecular dynamics, and it not only aided our understanding of a 4-electron reductase domain, but also provided a framework for future engineering efforts.

In Chapter 4, the interactions between the *E. coli* acyl carrier protein (AcpP) and β-ketoacyl-ACP-synthase I (FabB) were investigated using X-ray crystallography, NMR and MD simulations. The combination of structural biology and molecular dynamic simulation showed how the extension of fatty acid acyl chains is accomplished in the bacterial FASs, with a focus on the role of protein-protein interactions in the regulation of fatty acid composition. This study built upon a previous study where our group successfully cross-linked AcpP with β-hydroxyacyl-AcpP dehydratase (FabA), allowing the first comparison of protein-protein interactions in type II fatty acid biosynthesis. AcpP helix II is shown to be important in anchoring the CP when interacting with FabA and FabB. NMR and MD studies revealed helix III interactions with FabB which may function in product turnover and AcpP release.

In Chapter 5, pH modulated product generation in a type III polyketide synthase (the benzalacetone synthase, RpBAS) was investigated using *in silico* models and validated via mutagenesis and biological assays. At low pH, RpBAS generates the triketide bis-noryangonin (BNY). In comparison, at high pH, RpBAS generates the diketide p-hydroxybenzalacetone (BA). A previous study by Shimokawa *et al* revealed that by reducing active site occupancy through mutagenesis, additional chain extensions and additional products can be biosynthesized.[10] Therefore, initially, it was assumed that differences in pH probably alter the volume of the active site, hence selecting the diketide versus triketide product. Interestingly, MD simulations performed at low and high pH did not show any drastic conformational changes or change in active site cavity size. Rather, MD simulations shows that a surface-exposed histidine residue may function as a pH sensor, which is capable of restraining an active site phenylalanine that stabilizes the growing substrate via pi-pi interactions within the active site.

6.2 Future directions

6.2.1 Force field optimization and improvements

In Chapter 2, the development of an initial phosphopantetheine force field parameters was described in detail. It is common in force field development to improve force field parameters after it is utilized in several studies through parametrization. Below is an example. The initial LIPID11 framework parameter set was primarily based on the General Amber Force Field (GAFF), with additional parameters adopted from the Glycam force field.[3] Here, the authors purposely changed the names of the atom types, so that revisions in the future could be accomplished. Dickson *et al* would later optimize the original LIPID11 force field by updating the torsion parameters in the hydrocarbon tails using electronic structure calculated at a higher level with the cc-pVQZ basis set.[2] Furthermore, additional fitting to physical observables was performed by fitting the force field to reproduce the heat of vaporization for methyl acetate, which it performed poorly with the previous LIPID11 and GAFF parameters. This optimized force field was released as LIPID14.

A major future improvement of my phosphopantetheine force field would be to generate an all-inclusive force field, where extender units, starter units and intermediates could be generated with an as-needed basis. NORINE, a database of non-ribosomal peptide extender units and non-ribosomal peptides is available online, which contains 534 monomer units and 1187 peptides. In order to increase coverage, it is impractical to generate force field parameters for every *n*-mer, because this would be computationally infeasible. However, we could generate the force field of molecules on demand with minimal input and control by the end-user. This could be accomplished through a pipeline encoded in a software package, or a user-friendly web server. However, a foreseeable issue with such an approach would be the need to deal with unknown parameters, including torsion angles, this future direction is still worth pursuing if we wish to further expand the user coverage for the phosphopantetheine force field.

6.2.2 Domain docking engineering through free energy binding studies

A major engineering goal in these systems is to control the incorporation of "unnatural" building blocks through insertion, deletion or swapping of the modules. This is regulated in the assembly line through docking domains between modules. Vigorous efforts had been invested towards this goal, but they had failed or generated the desired product in low yield. This is most likely the result of our lack of understanding of protein dynamics and protein-protein interactions in these systems. In addition, more structures of the docking domains are needed. It would be incredibly useful if computational techniques could identify and predict binding patterns in the docking domains, which would allow docking domains to be swapped, or be altered to adjust their binding preference. A protocol to identify sequence motifs, known as the fingerprints, was described by Keatinge-Clay and used to define the regions of the ketoreductase domain that control the stereochemistry of the reduction product.[11] Wood *et al* had recently used a similar protocol on acyl carrier proteins (Figure 6-1). He identified contact patterns in

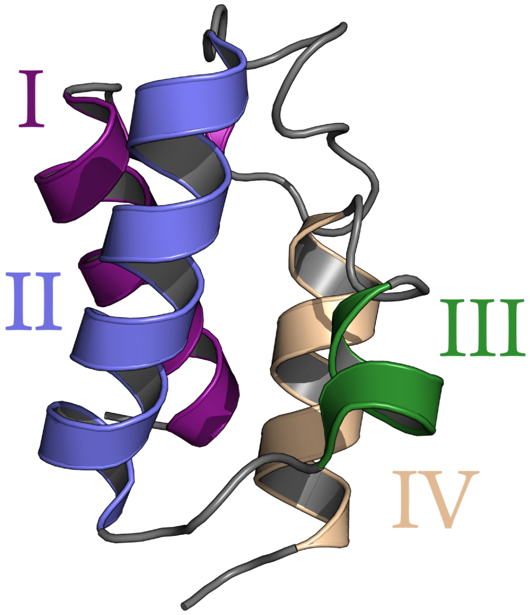


Figure 6-1. Helix II and helix III of ACP is known to be important for binding and interacting with its partner domains once loaded. (PDB ID: 1T8K)

different families of acyl carrier proteins through sequence and cladogram-based analysis in the trans-AT systems.[12] Wood's analysis on the co-evolution between ACP and its downstream partner provides evidence that simply engineering a single domain is not enough. This may also explain why previous engineering efforts have failed in these systems, and that a coupled approach between the interacting domains needs to be taken into account. While sequence comparison studies have been done on the carrier proteins and ketoreductase domains, none have been conducted on the docking domains, which is probably the result of a lack of structural information on these recently discovered domains.

One way this could address the above knowledge gap is to perform isothermal calorimetry (ITC) experiments between the docking domains of upstream and downstream modules to generate a database of binding energies. Initial free energy calculations performed *in silico* on the same systems could be used to fit the existing natural products force field to better represent the experimentally derived binding free energies. Such an approach, if setup correctly with a test set and validation set, could provide a force field capable of binding affinity between the docking domains in these systems.

6.2.3 Long-range interactions and allostery in polyketide synthases

In the type II FAS, it appears that interdomain communication may play a key role in product regulation. In Chapter 4, when FabB is bound to AcpP, MD simulations showed an increase in movement in the α6-α7 helix-turn-helix (HTH) motif of FabB, which is in close proximity to interact with another highly flexible helix (α10) of the same monomer. In order to analyze these dynamic regions in more depth, residue interaction network (RIN) analysis could be performed on these systems. RINs could be generated for FabB with no carrier protein, or with one carrier protein loaded in various states. Such an approach could be used to perturb FabB and cluster residues into communities. Furthermore, studies could be performed with carrier proteins present on both FabB monomers to further probe long-range interactions of FabB-AcpP complex so that we can better understand how the carrier proteins might communicate across homodimers of FabB.

In addition to network analysis of model type II FAS systems, network analysis can also be applied to long-range interactions that allosterically modulate active sites of enzymes in the acetate and non-ribosomal peptide pathways. Specifically, if lids are present in enzymes of these systems, the crystal structures are often obtained in the open or closed states. It is known that the lid regions can control preference for substrate selection and product generation. An increased knowledge in the dynamics of these hinge movements, or other conformational changes, would provide a blueprint for protein engineers who wish to modify these lid regions.

6.2.4 Molecular dynamics simulations on large enzyme complex over long timescales

The past five years have seen a proliferation of megasynthase structures, including the partner enzyme complexes described in Chapter 4, FabA:AcpP and FabB:AcpP. Juan Perilla and the late Klaus Schulten reported an MD simulation for a large protein complex with 64 million atoms over 1 μs to study the physical properties of the HIV-1 capsid.[13] Performing molecular dynamics simulations on complexes of this size reveals emergent properties that cannot be observed in simulations on a smaller scale. Physical properties such as electrostatics, vibrational and acoustic properties, and solvent effects can be elucidated via such large-scale simulations. Principal component analysis and normal-mode analysis can be used to analyze the collective motions of the viral capsid to study the hydrodynamic effects. Similar approaches can be utilized for FAS, PKS and NRPS.

While MD simulations have been performed on a homology model of hFAS, none have been performed on larger systems in these pathways. The computational power is available to perform all-atom simulations on system as large as the FAS multi-enzyme complexes found in fungi and bacteria (Fig. 6-2).[14] Anselmi *et al* performed a simulation on fungal FAS using a coarse-grained model, and showed that ACP shuttling is the result of molecular crowding effects.[15] All-atom simulations have the potential to explore long-standing questions in the field. Specifically, are the ACPs coupled and able to see each other, or are they independent? Because FASs from bacteria and fungi are different than those found in humans, they are vigorously-pursued targets for drug development. Additional details of protein dynamics in these systems could lead to a greater understanding of the complex interactions in these megasynthases.

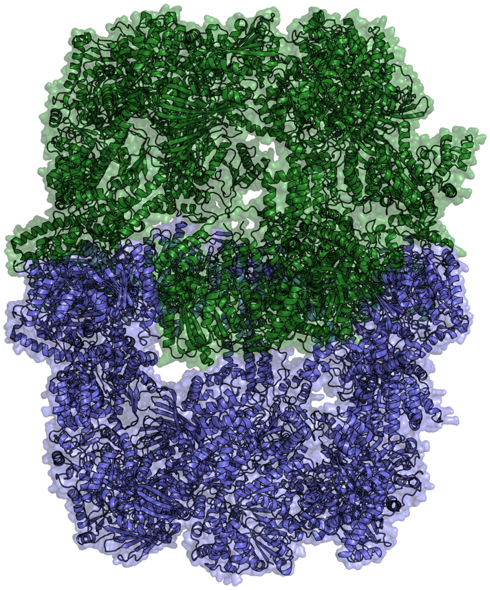


Figure 6-2. Crystal structure of the 1.9 Mda type I FAS from *Mycobacterium tuberculosis*.[14] The structure is similar to fungal FAS and consists of two reaction chambers on opposite sides of an internal wheel. (PDB ID: 4V8W)

References

1. Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R.; Bruegger, J.; Caldara-Festin, G.; Finzel, K.; Jackson, D. R.; Ishikawa, F.; O'Dowd, B.; McCammon, J. A.; Opella, S. J.; Tsai, S. C.; Burkart, M. D., Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **2014,** *505* (7483), 427-31.

2. Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: The Amber Lipid Force Field. *J Chem Theory Comput* **2014,** *10* (2), 865-879.

3. Skjevik, A. A.; Madej, B. D.; Walker, R. C.; Teigen, K., LIPID11: a modular framework for lipid simulations using amber. *J Phys Chem B* **2012,** *116* (36), 11124-36.

4. Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A., Forcefield_PTM: Charge and AMBER Forcefield Parameters for Frequently Occurring Post-Translational Modifications. *J Chem Theory Comput* **2013,** *9* (12), 5653-5674.

5. Khoury, G. A.; Smadbeck, J.; Tamamis, P.; Vandris, A. C.; Kieslich, C. A.; Floudas, C. A., Forcefield_NCAA: ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family. *ACS Synth Biol* **2014,** *3* (12), 855-69.

6. Homeyer, N.; Horn, A. H.; Lanig, H.; Sticht, H., AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J Mol Model* **2006,** *12* (3), 281-9.

7. Song, D.; Luo, R.; Chen, H. F., The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **2017,** *57* (5), 1166-1178.

8. Song, D.; Wang, W.; Ye, W.; Ji, D.; Luo, R.; Chen, H. F., ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chem Biol Drug Des* **2017,** *89* (1), 5-15.

9. Silakowski, B.; Nordsiek, G.; Kunze, B.; Blocker, H.; Muller, R., Novel features in a combined polyketide synthase/non-ribosomal peptide synthetase: the myxalamid biosynthetic gene cluster of the myxobacterium Stigmatella aurantiaca Sga15. *Chem Biol* **2001,** *8* (1), 59-69.

10. Shimokawa, Y.; Morita, H.; Abe, I., Structure-based engineering of benzalacetone synthase. *Bioorg Med Chem Lett* **2010,** *20* (17), 5099-103.

11. Keatinge-Clay, A. T., A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chemistry & Biology* **2007,** *14* (8), 898-908.

12. Wood, D. A. V.; Keatinge-Clay, A. T., The modules of trans-acyltransferase assembly lines redefined with a central acyl carrier protein. *Proteins* **2018,** *86* (6), 664-675.

13. Perilla, J. R.; Schulten, K., Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nat Commun* **2017,** *8*.

14. Ciccarelli, L.; Connell, S. R.; Enderle, M.; Mills, D. J.; Vonck, J.; Grininger, M., Structure and conformational variability of the mycobacterium tuberculosis fatty acid synthase multienzyme complex. *Structure* **2013,** *21* (7), 1251-7.

15. Anselmi, C.; Grininger, M.; Gipson, P.; Faraldo-Gomez, J. D., Mechanism of Substrate Shuttling by the Acyl-Carrier Protein within the Fatty Acid Mega-Synthase. *Journal of the American Chemical Society* **2010,** *132* (35), 12357-12364.