UNIVERSITY OF CALIFORNIA SAN DIEGO

Biophysical investigation of acyl-carrier and SARS-CoV-2 protein communication mechanisms

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Chemistry

by

Terra Sztain-Pedone

Committee in charge:

Professor Michael D. Burkart, Chair
Professor J. Andrew McCammon, Co-Chair
Professor Rommie E. Amaro
Professor Randolph Y. Hampton
Professor Alexis C. Komor
Professor Susan S. Taylor

2021

The dissertation of Terra Sztain-Pedone is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

For Terriel Cox, a relentless advocate for former foster youth, who gave more love than we deserved. Forever in our hearts.

For Alice, the most vibrant, hoppy, and sweet bunny, who lived a great life.

EPIGRAPH

"Shoot for the moon. Even if you miss, you'll land among the stars."

— Les Brown

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SCHEMES

ACKNOWLEDGEMENTS

There are so many people I am grateful to, who have guided, supported, or influenced me along this journey. For brevity, I will only mention a few. I would like to thank Gina Giacone, my independent living coordinator, who helped me begin my educational career, beginning with community college after aging out of foster care. I also thank my loving family. Next, the advisors, mentors, friends, and colleagues I've had throughout graduate school.

I appreciate my committee Chair, Mike Burkart, for the scientific ideas, discussions, and freedom to carry out exciting research. I am very grateful to my committee Co-Chair, Andy McCammon, for the instant feedback on manuscripts, nominations for awards and career development opportunities, brilliant discussion during group meetings, and continuous encouragement throughout my PhD studies. I am also grateful to my esteemed committee members, Rommie Amaro, Alexis Komor, and Susan Taylor, inspirations for women in science. Especially Rommie Amaro, who accepted me as a part of her team to simulate the SARS-CoV-2 viral proteins, and tirelessly mentored and supported me throughout all hours of the night on these urgent projects. My external department committee member, Randy Hampton, is also dear to my heart, as an old friend whose inspirational career path has helped me believe in my own.

Chapter 1, in full, is a reprint of the material as it appears: Sztain, T.; Patel, A.; Lee, D. J.; Davis, T. D.; McCammon, J. A.; Burkart, M. D. Modifying the Thioester Linkage Affects the Structure of the Acyl Carrier Protein. *Angew. Chemie Int. Ed.* **2019**, *58* (32), 10888–10892. The dissertation author was the primary author of this manuscript.

Chapter 2, in full, is a reprint of the material as it appears: Sztain, T.; Bartholow, T. G.; McCammon, J. A.; Burkart, M. D. Shifting the Hydrolysis Equilibrium of Substrate Loaded Acyl Carrier Proteins. *Biochemistry* **2019**, *58* (34), 3557–3560. The dissertation author was the primary author of this manuscript.

Chapter 3, in full, is a reprint of the material as appears: Sztain, T.; Bartholow, T. G.; Lee, D. J.; Casalino, L.; Mitchell, A.; Young, M. A.; Wang, J.; McCammon, J. A.; Burkart, M. D. Decoding Allosteric Regulation by the Acyl Carrier Protein. *Proc. Natl. Acad. Sci.* **2021**, *118* (16), e2025597118. The dissertation author was the primary author of the manuscript.

Chapter 4, in full, is currently being prepared for submission for publication as: Sztain, T, Corpuz, J.C., Bartholow, T.G., McCammon, J.A., Burkart, M.D. Metabolic Pathway Engineering via Protein-Protein Interface Design. The dissertation author is the primary co-author of this manuscript, along with Joshua Corpuz.

Chapter 5, in full, is currently being prepared for submission for publication as Sztain, T, McCammon, J.A., Burkart, M.D. Mechanism of acyl carrier protein chain flipping. The dissertation author is the primary author of this manuscript.

Chapter 6, in full, is a reprint of the material as it appears: Sztain, T.; Amaro, R.; McCammon, J. A. Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Protease. *bioRxiv*. **2020**. 2020.07.23.218784 The dissertation author is the primary investigator and author of this manuscript.

Chapter 7, in full, is a reprint of the material as it is appears: Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; McCool, R. S.; Kearns, F. L.; Andrew McCammon, J.; McLellan, J. S.; Chong, L. T.; Amaro, R. E. A Glycan Gate Controls Opening of the SARS-CoV-2 Spike Protein. *bioRxiv* **2021**, 2021.02.15.431212. The dissertation author is the primary co-author of the manuscript along with Surl-Hee Ahn.

VITA

2013	Rose Hills Research Fellow, Oak Crest Institute of Science Advisor: Mark Baum, Ph.D.
2014	National Cancer Institute Research Fellow, City of Hope Advisor: Dan Raz, M.D.
2016	Undergraduate Researcher, University of California, Los Angeles Advisor: Helena Chang, M.D., Ph.D.
2016	Bachelor of Science in Biochemistry, University of California, Los Angeles
2021	Doctor of Philosophy in Chemistry, University of California San Diego Advisors: Professors Michael D. Burkart and J. Andrew McCammon

PUBLICATIONS

Sztain, T., Corpuz, J., Bartholow, T.G., McCammon, J.A., Burkart, M.D. Metabolic Pathway Engineering via Protein-Protein Interface Design. (*In Preparation 2021*)

Sztain, T., McCammon, J.A., Burkart, M.D. Mechanism of acyl carrier protein chain flipping. (*In Preparation 2021*)

Bartholow, T.G., **Sztain, T**., Young, M.A., Lee, D.J., Davis, T.D., Abagyan, R., Burkart, M.D. Control of unsaturation in de novo fatty acid biosynthesis by FabA. *Biochemistry* (*In Press 2021*).

Bartholow, T.G., **Sztain, T**., Young, M.A., Davis, T.D., Abagyan, R., Burkart, M.D. Protein-protein Interaction based substrate control in the *E. coli* octanoic acid transferase, LipB *RSC Chem. Bio.* (*In Review 2021*).

Sztain, T., Ahn, S-H., Bogetti, A.T., Casalino, L, Goldsmith, J.A., McCool, R.S., Kearns, F.L., McCammon, J.A., McLellan, J.S., Chong, L.T., Amaro R.E., A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nature Chemistry* (*In Press 2021*).

Sztain, T., Amaro, R., McCammon, JA., Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease. *JCIM* (*In Press 2021*). ***Cover Art***

Sztain, T., Bartholow, T.G., Lee, D.J., Casalino, L., Mitchell, A., Young, M., Wang, J., McCammon J.A., Burkart, M.D. Decoding the allosteric nature of acyl carrier proteins. *PNAS*. **118**, (2021).

Casalino, L.; Dommer, A.; Gaieb, Z.; Barros, E. P.; **Sztain, T**.; Ahn, S.-H.; Trifan, A.; Brace, A.; Bogetti, A.; Ma, H.; Lee, H.; Turilli, M.; Khalid, S.; Chong, L.; Simmerling, C.; Hardy, D. J.; Maia, J. D. C.; Phillips, J. C.; Kurth, T.; Stern, A.; Huang, L.; McCalpin, J.; Tatineni, M.; Gibbs,

T.; Stone, J. E.; Jha, S.; Ramanathan, A.; Amaro, R. E. AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics. *Int. J. High Perform. Comput. Appl.* (2021). ***Gordon Bell Prize***

Bartholow, T.G., **Sztain, T**., Patel, A., Lee, D.J., Young, A.M., Abagayan, R., Burkart, M.D. Elucidation of transient protein-protein interactions within carrier protein-dependent biosynthesis *Commun. Biol.* **4**, 1–10 (2021).

Sztain, T., Bartholow, T.G., McCammon, J.A., Burkart, M. D. Shifting the Hydrolysis Equilibrium of Substrate Loaded Acyl Carrier Proteins. *Biochemistry* **58**, 3557–3560 (2019). ***Cover Art***

Sztain, T., Patel, A., Lee, D.J., Davis, T.D., McCammon, J. A., Burkart, M. D. Modifying the Thioester Linkage Affects the Structure of the Acyl Carrier Protein. *Angewandte Chemie International Edition* **58**, 10888–10892 (2019).

Nardi I., Reno T., Yun X., **Sztain T**., Wang J., Dai H., Zheng L., Shen B., Kim J., Raz D. Triptolide inhibits Wnt signaling in NSCLC through upregulation of multiple Wnt inhibitory factors via epigenetic modifications to Histone H3. *Int. J. Cancer* **143**, 2470–2478 (2018).

Gunawardana M., Chang S., Jimenez A., Holland-Moritz D., Holland-Moritz H., La Val T.P., Lund C., Mullen M., Olsen J., **Sztain T**., Yoo J., Moss J.A., Baum M.M. Isolation of PCR quality microbial community DNA from heavily contaminated environments. *J. Microbiol. Methods* **102,** 1–7 (2014).

ABSTRACT OF DISSERTATION

Biophysical investigation of acyl-carrier and SARS-CoV-2 protein communication mechanisms

by

Terra Sztain-Pedone

Doctor of Philosophy in Chemistry

University of California San Diego, 2021

Professor Michael D. Burkart, Chair

Professor J. Andrew McCammon, Co-Chair

Proteins are one of the primary macromolecules that carry out the many essential functions of life. They catalyze chemical reactions, provide structural integrity, and perform biological motor functions. Proteins dynamically interact with surrounding molecules by diffusing, binding, and changing conformations, signaling their stage in biochemical processes. Interpreting these signals is key to harnessing control over protein function for beneficial purposes, such as inhibiting disease related proteins and synthesizing desired molecules including biofuels and therapeutics. This dissertation investigates the communication mechanisms of two protein classes: acyl carrier proteins (ACPs) and SARS-CoV-2 viral proteins. The methods and fundamental properties uncovered herein are expected to be applicable to a broad range of protein systems.

Introduction

Primary and secondary metabolic pathways that produce fatty acids, polyketides, and non-ribosomal peptides depend on an ACP, or the homologous peptidyl carrier protein (PCP), to shuttle intermediates between enzymes. These biosynthetic pathways have evolved elegant communication mechanisms to regulate substrate specificity for efficient natural product formation. ACP carries a myriad of intermediate molecules in a hydrophobic core and delivers each to the appropriate enzyme via a large conformational change termed chain flipping. This process is mediated by a post-translational phosphopantetheine co-factor, which behaves as an "arm" holding onto substrates with a covalent thioester bond, translocating the substrate from the ACP pocket to the enzyme substrate pocket. Two conflicting hypotheses have been proposed by researchers regarding the ACP mechanism: 1) ACP stochastically delivers substrates to enzymes, sampling active sites until the appropriate enzyme is reached or 2) ACP communicates substrate identity to enzymes via protein-protein interactions, delivering substrates only upon interaction with the appropriate enzyme. The first part of this dissertation focuses on testing these hypotheses to uncover the decades long mystery of ACP communication.

Studying substrate-loaded ACP *in vitro* is challenging due to the liability of the thioester bond to hydrolysis. Chemoenzymatic methods have been developed to replace this with a more stable amide bond. However, as described in the following chapters, nuclear magnetic resonance (NMR) spectroscopy and molecular dynamics (MD) simulations revealed that even single atom changes to the thioester bond can impact the global structure of ACP. Therefore, an enzyme-based NMR method was developed to counteract thioester hydrolysis *in situ*. This method was next used, along with MD simulations, to characterize ACP carrying every acyl chain length from fatty acid

biosynthesis. Next, titration NMR was used to determine the binding interface of the enzyme LipB, which is specific for C8-AcpP, with the specific substrate and a non-cognate substrate, C12-AcpP. These experiments, combined with computational docking, confirmed a unique, more favorable interface with the specific substrate. To determine if the difference in binding regulates substrate delivery, an NMR experiment was designed using specific isotope labeling to detect chain flipping. These results confirmed that chain flipping occurred only with the specific substrate. Mutation of a residue at the protein-protein interface decreased specificity and allowed chain flipping, providing a platform for which specificity information can be leveraged to control metabolic pathways for desired outcomes. This inspired a metabolic pathway engineering framework relying on re-designing protein-protein interactions. An iterative experimental and computation protocol was developed, and a test case showed an increase from negligible wild type activity to > 90% desired activity, forming a new to nature product.

The next goal of this dissertation was to visualize atomic resolution details of the dynamic chain flipping mechanism using MD simulations. Large scale conformational changes occur between millisecond and second timescales and are not possible to detect with conventional MD simulations, which are limited to sampling up to microsecond processes. Therefore the enhanced sampling method, weighted ensemble, was used. This method relies on running many short simulations in parallel, focusing computational resources towards sampling new regions of conformational space by splitting trajectories, and avoiding redundant sampling by merging trajectories. Through rigorous tracking of each trajectory probability, this method is able to generate continuous, unbiased pathways, while maintaining both thermodynamic and kinetic properties.

Enhanced sampling MD was also applied to the SARS-CoV-2 main protease. The protease is responsible for cleaving the viral polypeptide chain into protein segments so replication can take place and viral functions can be performed. Therefore, the protease is a primary target for antiviral development. Identification of cryptic protein pocket conformations with enhanced sampling MD will provide greater opportunities for successful *in silico* inhibitor design.

Finally, the weighted ensemble approach was applied to resolve atomic details of the SARS-CoV-2 glycosylated spike activation mechanism. The spike proteins surrounding viral particles facilitate entry into human cells through a protein-protein interaction between the receptor binding motif of the spike and the ACE2 receptor on the surface of human cells. The spike proteins are cloaked by a shield of sugar molecules called glycans, which allow the virus to hide from the immune system. The receptor binding motif is also shielded by glycans, and a large conformational change of the receptor binding domain from a closed to open position is required for binding and subsequent infection. Simulations of this process set a new high watermark for the largest system simulated with the weighted ensemble method (~500,000 atoms), and contributed to a massive collaboration to simulate the entire viral envelope (305 million atoms) which won the Gordon Bell High Performance Computing Award. Analysis of the simulations revealed that the glycan at position N343 initiates spike opening through intercalating under the receptor binding domain through a “hand-jive” type motion. Abrogation of this glycan reduced ACE2 binding by 56% during biolayer interferometry experiments. Atomic details of the opening mechanism, and conformations along the entire opening pathway provide a foundation for understanding and preventing the mechanism of SARS-CoV-2 infection.

Chapter 1. Modifying the thioester linkage affects the structure of the acyl carrier protein

Terra Sztain, Dr. Ashay Patel, Dr. D. John Lee, Dr. Tony D. Davis, Prof. J. Andrew McCammon , Prof. Michael D. Burkart*

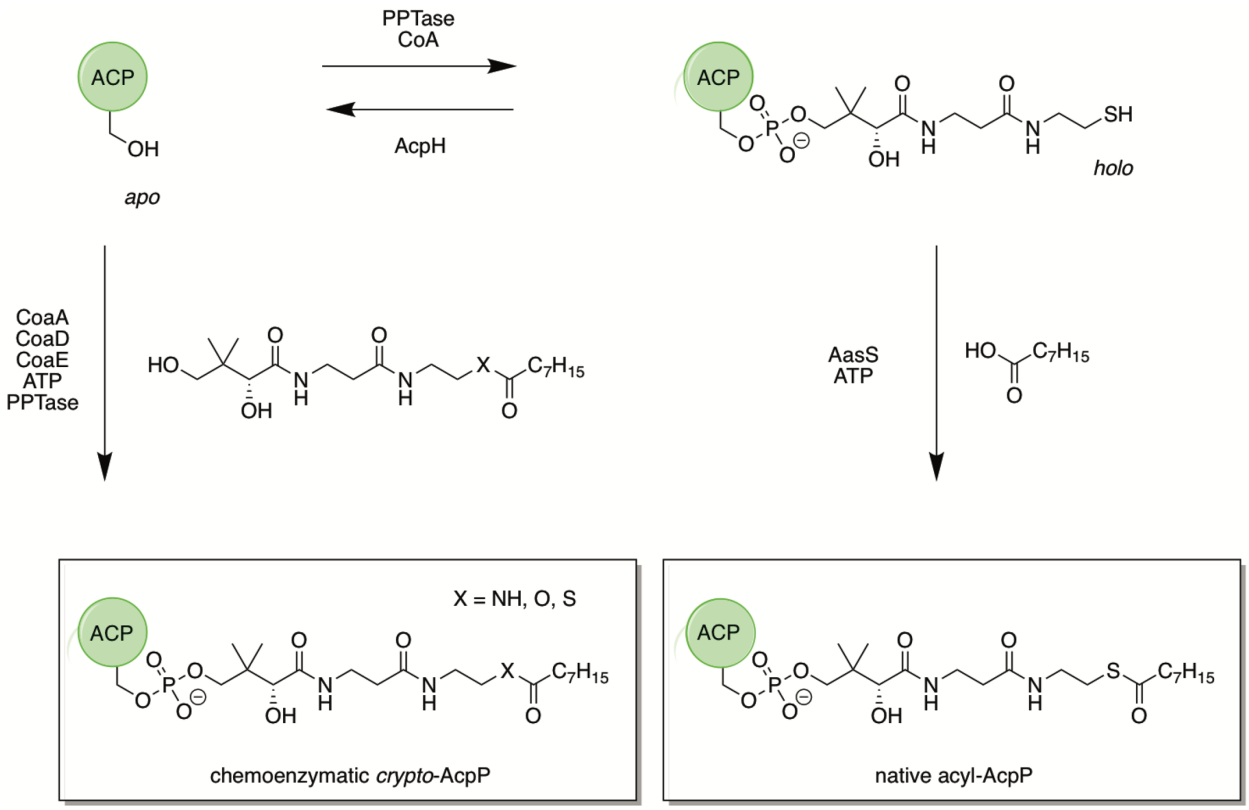
Abstract

At the center of many complex biosynthetic pathways, the acyl carrier protein (ACP) shuttles substrates to appropriate enzymatic partners to produce fatty acids and polyketides. Carrier proteins covalently tether their cargo via a thioester linkage to a phosphopantetheine cofactor. Due to the labile nature of this linkage, chemoenzymatic methods have been developed that involve replacement of the thioester with a more stable amide or ester bond. We explored the importance of the thioester bond to the structure of the carrier protein by using solution NMR spectroscopy and molecular dynamics simulations. Remarkably, the replacement of sulfur with other heteroatoms results in significant structural changes, thus suggesting more rigorous selections of isosteric substitutes is needed.

Acyl carrier proteins (ACPs) are involved in pathways with metabolic, pharmaceutical, and environmental significance.[1-3] A clear structural understanding of ACPs and their interactions with partner enzymes is essential to furthering metabolic engineering and drug discovery. Efforts in this regard have proven challenging due to transient interactions and the dynamic nature of the ACP, which transports intermediates to multiple enzymes via a covalent but labile thioester linkage. The archetypical carrier protein AcpP, which is involved in type II fatty acid biosynthesis in *Escherichia coli*, has been well studied,[4, 5] but key questions regarding conformational dynamics and interactions between AcpP and partner enzymes remain partially understood. Methods to capture ACP–partner proteins interactions have been developed, and often require the

preparation of acyl-ACPs in which the naturally occurring thioester linkage has been replaced with a more stable amide or ester bond. Chemoenzymatic preparations of ACPs bearing crosslinking probes have been used to covalently trap ACP–enzyme complexes.[6-8] Using this strategy, crystal structures of *E. coli* fatty acid ACP (AcpP) crosslinked to FabA,[9] FabB,[10] and FabZ[11] have been determined, which reveal discrete molecular interactions that mediate AcpP binding to these enzymes. What remains unknown, however, are the dynamics of these interactions throughout ACP–partner protein binding events, including how intermediates attached to the ACP influence protein dynamics and molecular recognition. *E. coli* AcpP interacts with at least 25 different proteins[9, 12, 13] and transports acyl intermediates with chain lengths ranging from four to eighteen carbons that assume one of four different β-oxidation states. Given the number of proteins with which AcpP interacts and the chemical diversity of the substrates it carries, it is likely that the process by which AcpP delivers its cargo (chain flipping) is not stochastic but rather regulated by substrate-influenced protein–protein interactions.[14]

Due the geometric and electronic differences between thioester, ester, and amide linkages, evaluation of such modifications is necessary. We sought to determine how the structure of AcpP is affected by the chemical nature of the substrate–AcpP linkage. In this study we used solution-state nuclear magnetic resonance (NMR) spectroscopy and molecular dynamics (MD) to evaluate the influence of thioester, ester, and amide linkages upon the structure of AcpP and its interactions with a partner enzyme.



Scheme 1.1 Chemoenzymatic synthesis of C8-NH-AcpP and C8-O-AcpP, and enzymatic synthesis of C8-S-AcpP. The synthesized phosphopantetheine (PPant) analogues are attached to *apo*-AcpP with a phosphopantetheinyl transferase (PPTase), whereas an acyl-acyl carrier protein synthetase (AasS) is used to attach octanoic acid to *holo*-AcpP. Conversion of *apo*- into *holo*-AcpP is achieved using PPTase and coenzyme A (CoA), and is reversed by AcpH.

First, solution NMR was used to compare the structures of variants of octanoyl-AcpP (C8-AcpP) featuring either a thioester (C8-S-AcpP), amide (C8-NH-AcpP), or ester (C8-O-AcpP) linkage. The thioacyl-AcpP analogues were synthesized using a chemoenzymatic "one-pot" method,[15] while the native thioacyl-AcpP was enzymatically synthesized with the enzyme VhAasS (Scheme **1**).[16] Substrate loading was monitored with conformationally sensitive urea polyacrylamide gel electrophoresis (PAGE). Linkage-based differences in gel migration were observed (Figure S1B in the Supporting Information). With this preliminary indication of conformational variation between the three ACPs, we further characterized the extent and nature of these differences using ^{1}H-^{15}N-heteronuclear single quantum correlation (HSQC) NMR experiments (Figure **1**). Perturbations were quantified by chemical-shift perturbation (CSP)

calculation,[17] in which the chemical shifts of each analogue were compared to naturally occurring thioacyl-AcpP. The spectra of the two non-natural AcpPs show significant chemical-shift perturbations of backbone amides (Figure **1**–**2**), thus indicating a change in the chemical environment of those residues. Notably, the residues that showed significant CSPs were all in regions of the carrier protein that are proximal to the phosphopantetheine (PPant) tethering the substrates. The presence of an amide linkage perturbs residues in helices II and III, while presence of the ester linkage induces these perturbations as well as additional CSPs of residues in helix IV (Figure **2**). We note that helix II is considered the "recognition helix,"[18] and that AcpP primarily interacts with partner proteins through helices II and III.[19]

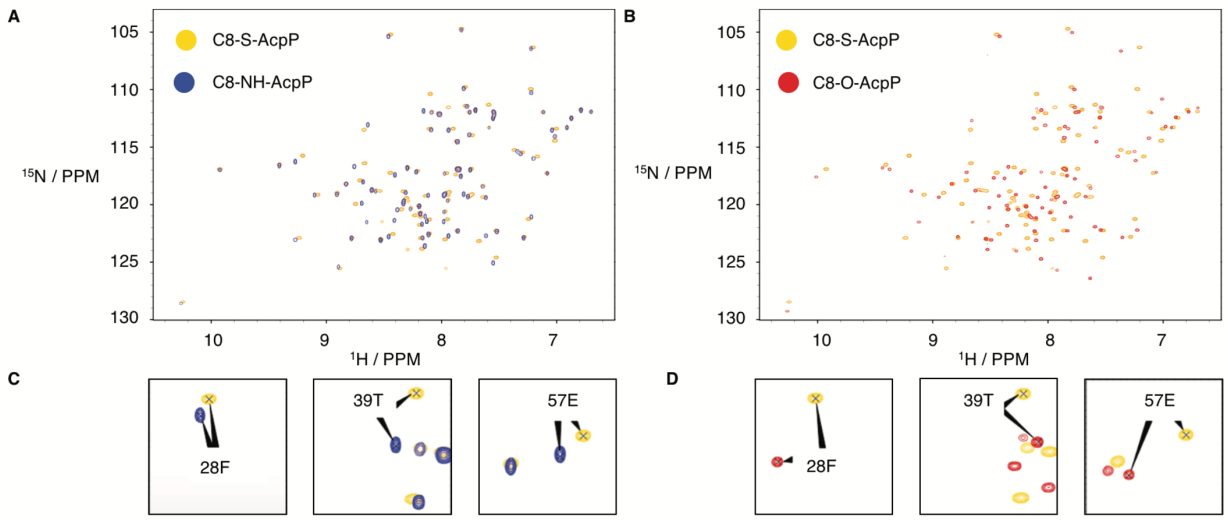


Figure 1.1 Comparison of the NMR spectra of three different acylated AcpP species. A) Overlay of ^{1}H-^{15}N HSQC spectra for C8-NH-AcpP and C8-S-AcpP. B) Overlay of ^{1}H-^{15}N HSQC spectra for C8-O-AcpP and C8-S-AcpP. C,D) Selected peaks that show significant chemical-shift perturbations in each spectrum are enlarged below.

These findings support the hypothesis that AcpP communicates substrate identity to partner proteins through changes in the recognition helix.[20, 21] To further investigate the structural consequences of the perturbed residues, we performed five independent 500 ns conventional MD simulations of each of the spectroscopically characterized AcpPs: C8-S-AcpP, C8-NH-AcpP, and

C8-O-AcpP. We analyzed how the different sterics, electrostatics, and hydrogen bonding of the linkage atoms affect the global structure of AcpP.

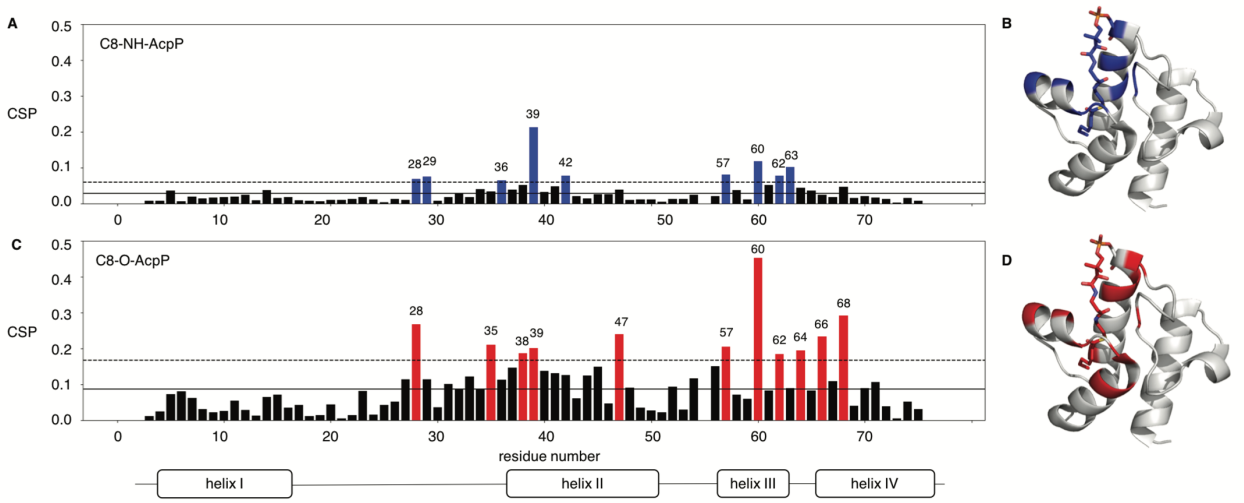


Figure 1.2 Plots of backbone amide chemical-shift perturbations (CSPs). A) CSP plots determined by comparing chemical shifts of C8-NH-AcpP and C8-S-AcpP. The mean and mean plus one standard deviation are illustrated using solid and dotted lines, respectively. Residues with CSPs larger than one standard deviation from the mean are colored blue. B) AcpP structure (PDB ID: 2FAD), illustrating the residues (in blue) that show significant CSPs. C) CSP plots determined by comparing chemical shifts of C8-O-AcpP and C8-S-AcpP. The mean and mean plus one standard deviation are illustrated using solid and dotted lines, respectively. Residues with perturbations above the standard deviation from the mean are colored red. D) AcpP structure (PDB ID: 2FAD), illustrating the residues (in red) that show significant CSPs.

The differences in the migration of each species on urea-PAGE indicated that the global shape and substrate sequestration were affected by substrate linkage in these acyl-AcpPs. MD simulations were consistent with the gel migration pattern, in which the sulfur atom of C8-S-AcpP is most sequestered, buried deeper in the pocket, and located further below helix III than the nitrogen atom of C8-NH-AcpP or the oxygen atom of C8-O-AcpP (Figure S1 A,D–F). The average calculated radius of gyration and solvent accessible surface area (SASA) of the entire protein are only slightly different, though they follow the same trend, increasing in order from C8-NH-AcpP to C8-O-AcpP to C8-S-AcpP (Figure S1 B,C).

There are substantial differences in the AcpP pocket volume, with average volumes calculated as 95.6 $Å^3$ in the native C8-S-AcpP species, compared to 102.6 $Å^3$ in the C8-NH-AcpP species and 119.2 $Å^3$ in the C8-O-AcpP species (Figure **3** A). These are more apparent than the differences contributed by electrostatic factors (Figure **3** B). The only difference between each species is the linkage between octanoic acid and PPant (S, O, or NH). Therefore, to investigate the steric effect of the decreased atomic radii of nitrogen and oxygen compared to sulfur, we calculated the average number of protein atoms within three differently sized shells surrounding the linkage atom throughout the simulation. No atoms occupy the sphere within 1.5 Å of the sulfur atom in the native species, while an average of three and two atoms occupy this sphere of the nitrogen and oxygen atoms, respectively. Considering a 2.5 Å sphere, four atoms are within the sphere in the native species, and double that number of atoms are present in both of the mimicking linkages. The 3.5 Å sphere contains the most similar numbers of atoms, with 15 for sulfur, 17 for nitrogen, and 16 for oxygen (Figure **3** C).

Hydrogen-bonding analysis revealed that the amide linkage in C8-NH-AcpP acts as a hydrogen bond donor that displaces native hydrogen bonding between the first proximal PPant amide (linking β-Ala and cysteamine) and AcpP. In the native C8-S-AcpP species, the first proximal amide, labeled α in Figure **3** D, donates a hydrogen bond to yhe backbone carbonyls of A34 (18 % of the simulation), A59 (18 %), D56 (3 %), and I62 (2 %), and the side-chain hydroxy group of T39 (7 %; Figure S8, Table S2). In C8-NH-AcpP, hydrogen bonding to the backbone oxygen of I62 and the α amide is completely displaced (0 % of simulation data features this contact) by the linkage amide, which spends 21 % of the simulation hydrogen bonded to I62. This results in the α amide of C8-NH-AcpP forming a new hydrogen bond, which is not observed in any of the other simulations, with the side-chain oxygen atoms of E60 for 4 % of the simulation

(Figure S8, Table S2). These findings can explain the chemical-shift differences observed between C8-S-AcpP and C8-NH-AcpP at residues E60 and I62 (Figure **2**). In C8-NH-AcpP, hydrogen bonding to the hydroxy group of T39 was with the α amide for 27 % of the simulation and was displaced by the amide linkage during 4 % of the simulation. The amide linkage also formed hydrogen bonds with the backbone oxygen atom of T39 for 7 % of the simulation (Figure **3** D, S8, Table S2). T39 is another residue that showed a significant chemical-shift perturbation when comparing the C8-S-AcpP and C8-NH-AcpP species (Figure **1**–**2**). On average, the phosphopantetheine of C8-NH-AcpP has one more hydrogen bond than C8-S-AcpP or C8-NH-AcpP (Figure S9). Calculating the SASA of just the ligand portion shows three distinct states in the native C8-S-AcpP and the C8-O-AcpP. The C8-NH-AcpP has only two distinct states, likely due to stable hydrogen bonding of the amide decreasing conformational transitions. The most occupied SASA state shifts from larger SASA in the native (and C8-O-AcpP) to a smaller SASA, thus indicating more sequestration inside the pocket, thereby shielding water accessibility. Due to the partial double-bond character of the amide bond, we sought to compare rotation about the linker atom and carbonyl carbon bond (Figure S10), but we found similar rigidity in the thioester and ester species. This is likely due to stereoelectronic effects and dipole minimization restricting rotation of the thioester and ester linkages.[22, 23]

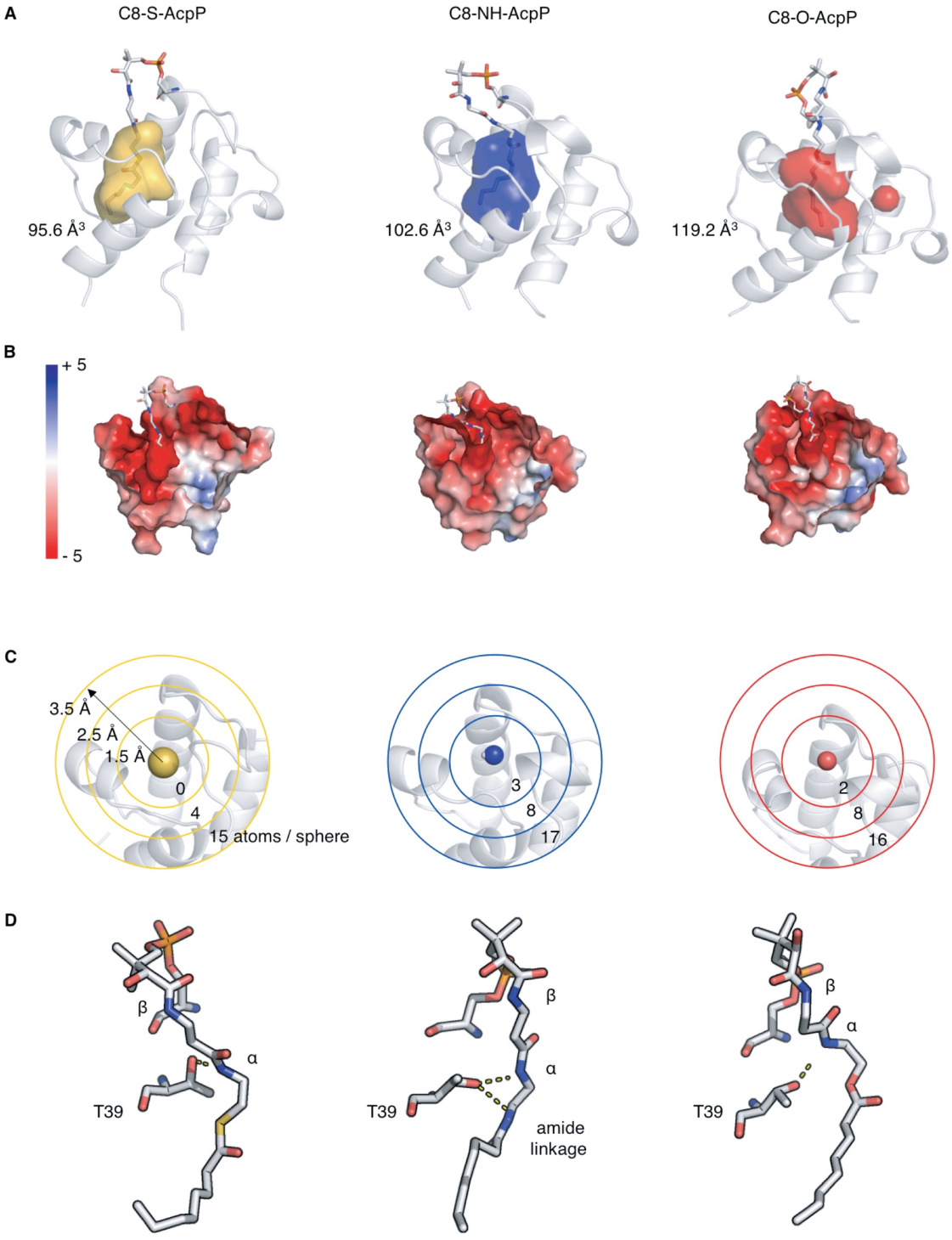


Figure 1.3 Analysis of AcpP pocket from molecular dynamics simulations. A) Average calculated pocket volume of each species. B) Comparison of the pocket electrostatic distribution of each species. C) Schematic showing average atom count within 1.5 Å (inner sphere), 2.5 Å (middle sphere), and 3.5 Å (outer sphere) from the center of mass of the linkage atom over 2.5 μs of simulation. D) Hydrogen bonding between the Ppant and residue T39 in each species.

Based on the differences in the position of PPant on the surface of these acyl-AcpPs and overall shape and substrate sequestration of each AcpP, we hypothesized that the thioester isosteres may lead to a functional difference in the binding of AcpP to partner proteins. To test this, we performed Cluspro[24-26] protein–protein docking simulations using representative structures from the MD data of either C8-S-AcpP, C8-NH-AcpP, or C8-O-AcpP with *apo*-FabA (PDB ID: 1MKB). The docking results confirm a canonical AcpP–partner protein binding mode that is primarily stabilized by electrostatic interactions between AcpP and FabA (Figure S11). The native C8-S- AcpP forms the most favorable interface with the lowest docking score (Table S5) and possesses the greatest number of total residues that participate in the protein–protein interface (Table S3, Table S4). The next most favorable score and greatest number of total interface residues is C8-NH-AcpP, followed by C8-O-AcpP.

Although the residues within 3.5 Å of the interface of each AcpP–FabA structure appear slightly different, there is only one key difference in the hydrogen bonding between FabA and AcpP. FabA residue K112 forms a hydrogen bond with Q14 in the native C8-S-AcpP docked structure, while with C8-NH-AcpP, the hydrogen bond is formed with E11. In the C8-O-AcpP structure, no hydrogen bonds are formed with K112 (Figure **4**). This particular interaction is unique because it is the only hydrogen bond between AcpP and the non-primary homodimer subunit of FabA. This interaction is not present in the AcpP–FabA crosslinked crystal structure (PDB ID: 4KEH), which captures the AcpP binding to FabA in a mechanistically relevant conformation, as opposed to the pre-chain flipping conformation detected by these docking simulations. Most of the interface residues elucidated by the crosslinked crystal structure are conserved between each of the docking simulations (Table S3, Table S4).

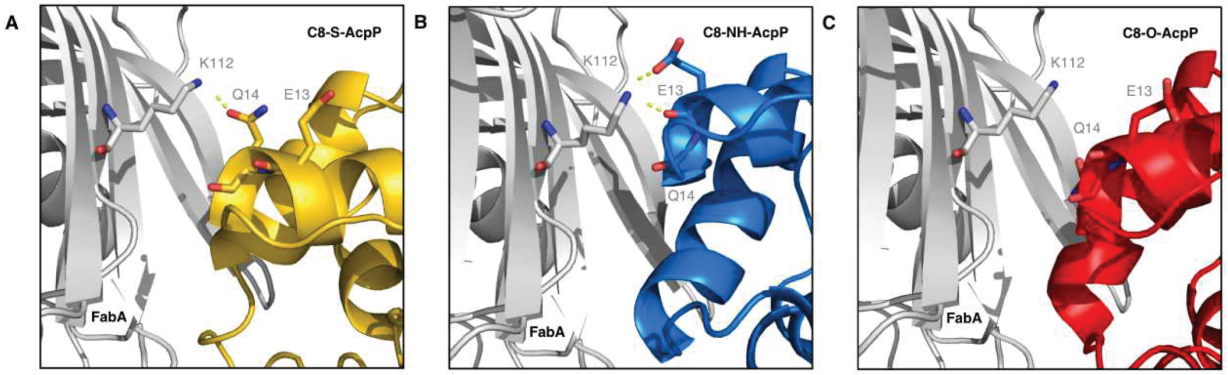


Figure 1.4 Differences in the docking interface in FabA–AcpP as a result of changing the acyl linker. The K112 salt bridge is shown as yellow dashes to either Q14 in C8-S-AcpP-FabA (A), E13 in C8-NH-AcpP-FabA (B), or neither in C8-O-AcpP-FabA (C).

This study demonstrates that minor chemical modifications of biomolecules can have dramatic effects on protein structure, which perhaps explains how the 77 amino acid AcpP protein can deliver the dizzying variety of possible intermediates to the appropriate partner enzymes in such a highly choreographed fashion. Although the amide linkage significantly alters the position of the fatty acid through hydrogen bonding, the oxoester linkage has a greater magnitude of chemical-shift values, significantly expands the pocket of AcpP, and results in poorer docking of AcpP to a partner enzyme. Therefore, we conclude the amide linkage is a more suitable substitute for the thioester linkage than the oxoester in this system and likely other systems involving sequestration of a chemical intermediate inside a protein cavity. Crosslinking experiments and the structural resolution of crosslinked AcpP–FabA, AcpP–FabZ, and AcpP–FabB suggest that modification of the thioester bond does not disrupt chain flipping and substrate localization into the active site of each enzyme. Although these differences do not seem to prohibit this primary function of AcpP, the finer details of the structure, dynamics, and interaction of AcpP with partner proteins are influenced by the chemical nature of the substrate–cofactor linkage. Such a fine understanding of AcpP will be required to further long-standing goals of metabolic engineering and drug discovery. Consequently, efforts should be directed towards developing more chemically

similar analogues as well as methods to study AcpP with substrates attached via the native thioester.

Acknowledgements:

We thank Prof. Stanley Opella for valuable NMR discussions, and Dr. Xuemei Huang for assistance with NMR facility use. This work was supported by NIH GM095970 to M.D.B. and NIH GM31749 to J.A.M. T.S. is an NSF GRFP fellow under grant number DGE-1650112. T.D.D. is a San Diego IRACDA Postdoctoral Fellow supported by the NIH K12 GM068524 award.

Chapter 1, in full, is a reprint of the material as it appears: Sztain, T.; Patel, A.; Lee, D. J.; Davis, T. D.; McCammon, J. A.; Burkart, M. D. Modifying the Thioester Linkage Affects the Structure of the Acyl Carrier Protein. *Angew. Chemie Int. Ed.* **2019**, *58* (32), 10888–10892. The dissertation author was the primary author of this manuscript.

1.1 References

[1] Finzel, K., Lee, D. J. & Burkart, M. D. Using modern tools to probe the structure-function relationship of fatty acid synthases. *Chembiochem Eur. J. Chem. Biol.* **2015**, 16, 528–547.

[2] Staunton, J. & Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **2001**, 18, 380–416.

[3] Marella, E. R., Holkenbrink, C., Siewers, V. & Borodina, I. Engineering microbial fatty acid metabolism for biofuels and biochemicals. *Curr. Opin. Biotechnol.* **2018**, 50, 39–46.

[4] Crosby, J. & Crump, M. P. The structural role of the carrier protein – active controller or passive carrier. *Nat. Prod. Rep.* **2012**, 29, 1111–1137.

[5] Beld, J., John Lee, D. & D. Burkart, M. Fatty acid biosynthesis revisited: structure elucidation and metabolic engineering. *Mol. Biosyst.* **2015,** 11, 38–59.

[6] Worthington, A. S., Porter, D. F. & Burkart, M. D. Mechanism-based crosslinking as a gauge for functional interaction of modular synthases. *Org. Biomol. Chem.* **2010**, 8, 1769–1772.

[7] Herbst, D. A.; Huitt-Roehl, C. R.; Jakob, R. P.; Kravetz, J. M.; Storm, P. A.; Alley, J. R.; Townsend, C. A.; Maier, T. The structural organization of substrate loading in iterative polyketide synthases. *Nature Chemical Biology* **2018**, 14, 474.

[8] Miyanaga, A.; Ouchi, R.; Ishikawa, F.; Goto, E.; Tanabe, G.; Kudo, F.; Eguchi, T. Structural Basis of Protein–Protein Interactions between a trans-Acting Acyltransferase and Acyl Carrier Protein in Polyketide Disorazole Biosynthesis. *J. Am. Chem. Soc.* **2018,** 140, 7970–7978.

[9] Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R. L.; 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, 427–431.

[10] Milligan, J. C., Jackson, D. R., Barajas, J. F. & Tsai, S. C. Crosslinked Crystal Structure of Type II Fatty Acid Synthase Ketosynthase, FabB, and Acyl Carrier Protein, AcpP. *BE Publ.* doi:10.2210/pdb5kof/pdb

[11] Dodge, G. J.; Patel, A.; Jaremko, K. L.; McCammon, J. A.; Smith, J. L.; Burkart, M. D. Structural and dynamical rationale for fatty acid unsaturation in Escherichia coli. *PNAS* **2019**, 201818686 doi:10.1073/pnas.1818686116

[12] De Lay, N. R. & Cronan, J. E. In vivo functional analyses of the type II acyl carrier proteins of fatty acid biosynthesis. *J. Biol. Chem.* **2007**, 282, 20319–20328.

[13] Gully, D. & Bouveret, E. A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in Escherichia coli. *PROTEOMICS* **2006**, 6, 282–293.

[14] Cronan, J. E. The chain-flipping mechanism of ACP (acyl carrier protein)-dependent enzymes appears universal. *Biochem. J.* **2014**, 460, 157–163.

[15] Worthington, A. S. & Burkart, M. D. One-pot chemo-enzymatic synthesis of reporter-modified proteins. *Org. Biomol. Chem.* **2006**, 4, 44–46.

[16] Beld, J., Finzel, K. & Burkart, M. D. Versatility of acyl-acyl carrier protein synthetases. *Chem. Biol.* **2014**, 21, 1293–1299.

[17] Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, 73, 1–16.

[18] Yadav, U., Arya, R., Kundu, S. & Sundd, M. The 'Recognition Helix' of the Type II Acyl Carrier Protein (ACP) Utilizes a 'Ubiquitin Interacting Motif (UIM)'-like Surface To Bind Its Partners. *Biochemistry* **2018**, 57, 3690–3701.

[19] 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, 13791–13799.

[20] 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, 135–145.

[21] Płoskoń, E.; Arthur, C. J.; Kanari, A. L. P.; Wattana-amorn, P.; Williams, C.; Crosby, J.; Simpson, T. J.; Willis, C. L.; Crump, M. P. Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. *Chem. Biol.* **2010**, 17, 776–785.

[22] Vajda, S.; Yueh, C.; Beglov, D.; Bohnuud, T.; Mottarella, S. E.; Xia, B.; Hall, D. R.; Kozakov, D. New additions to the ClusPro server motivated by CAPRI. *Proteins* **2017**, 85, 435–444.

[23] 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, 255–278.

[24] 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, 2159–2166.

Chapter 2. Shifting the hydrolysis equilibrium of substrate loaded acyl carrier proteins

Terra Sztain, Thomas G. Bartholow, J. Andrew McCammon, Michael D. Burkart*

Abstract

Acyl carrier proteins (ACP)s transport intermediates through many primary and secondary metabolic pathways. Studying the effect of substrate identity on ACP structure has been hindered by the lability of the thioester bond that attaches acyl substrates to the 4′-phosphopantetheine cofactor of ACP. Here we show that an acyl acyl-carrier protein synthetase (AasS) can be used in real time to shift the hydrolysis equilibrium toward favoring acyl-ACP during solution NMR spectroscopy. Only 0.005 molar equivalents of AasS enables 1 week of stability to palmitoyl-AcpP from *Escherichia coli*. 2D NMR spectra enabled with this method revealed that the tethered palmitic acid perturbs nearly every secondary structural region of AcpP. This technique will allow previously unachievable structural studies of unstable acyl-ACP species, contributing to the understanding of these complex biosynthetic pathways.

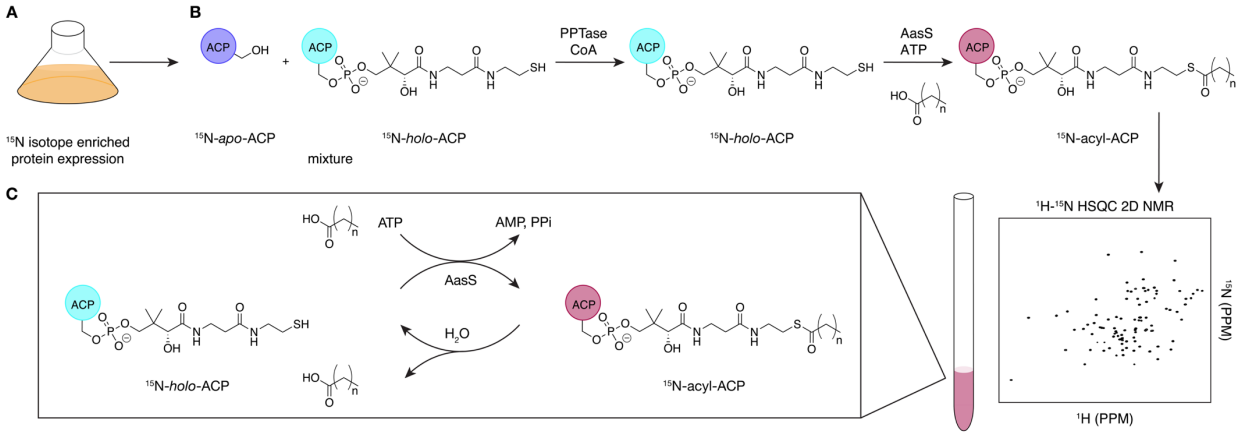
Acyl carrier proteins (ACP)s play a critical role by chaperoning intermediates between enzymes in many primary and secondary metabolic pathways.(1−3) Upon binding to the appropriate partner enzyme, the 4′-phosphopantetheine (PPant) “arm” of ACP facilitates exiting of the acyl substrate from within the ACP pocket and its entering into the active site of a partner enzyme in a process termed chain-flipping.(4) Many enzymes demonstrate specificity for particular acyl chain lengths or oxidation states. Since the acyl substrate is sequestered inside the hydrophobic pocket of ACP during partner enzyme binding,(5) protein–protein interactions must play a critical role in the processivity of these pathways.(6) However, little is understood about how the identity of the substrate is communicated by the ACP or how substrate identity effects the

overall structure of loaded ACP. We have sought to understand these processes using solution-phase NMR spectroscopy.

Acyl substrates are covalently attached to the terminal thiol of PPant through a thioester linkage. Sequestration of hydrophobic acyl chains inside the pocket of ACP is known to protect the labile thioester bond from hydrolysis, and longer chains have been shown to have increasingly greater hydrolysis rates,(7) a phenomenon that has impeded the study of long chain acyl-ACPs. Chemoenzymatic methods have been devised to replace the thioester with a more stable amide bond, however this has been shown to affect the structure of ACP.(8) Therefore, we sought to develop a method that would allow structural study of long chain acyl-ACP. Recently, a method was described to probe molecular influences of labile substrates in carrier proteins from nonribosomal peptide synthases (NRPS),(9) which allowed the first solution NMR structure of a substrate loaded peptidyl-carrier protein.(10) We realized that a similar strategy could be employed by leveraging the acyl acyl carrier protein synthetase (AasS) to shift the hydrolysis equilibrium toward favoring acyl-ACP during NMR experiments.

AasS is a promiscuous synthetase that attaches fatty acids to *holo*-ACP, forming acyl-ACP and consuming ATP.(11) This reaction opposes hydrolysis; therefore, we hypothesized that AasS and reaction components could be combined with purified acyl-ACP as a tool to reattach fatty acids to hydrolyzed *holo*-ACP (Scheme 1), thereby preserving the acylated form for long protein NMR experiments. As described herein, we have evaluated the stability of short, medium, and long acyl chain lengths at timescales relevant to solution NMR structural studies. We have also determined the optimal AasS concentration with which to collect NMR spectra. This method allowed the collection of 2D NMR spectra of palmitoyl-AcpP, a species which would have not otherwise been obtainable, leading to the discovery that addition of this long chain fatty acid

perturbs nearly every region of AcpP secondary structure. This concept can be further applied to uncover the effect of other labile intermediates upon the structure of ACP and its partner enzymes to further understand communication and processivity in these complex, carrier protein dependent, biosynthetic pathways.



Scheme 2.1. Shifting Equilibrium Towards Acyl-ACP Workflow (A) ACP overexpression typically yields two species, (B) *apo*-ACP, without PPant cofactor and *holo*-ACP with PPant covalently bound,a and (C) purified acyl-ACP added to NMR tube along with AasS and reaction components to counteract hydrolysis by reattaching fatty acids to *holo*-ACP.

[a]A PPTase attaches CoA to *apo*-ACP, producing *holo*-ACP to which AasS attaches a fatty acid.

Using ACP from *Escherichia coli* fatty acid synthase (AcpP), we first tested the stability of acyl-AcpP in the presence or absence of 0.005 molar equivalents of recombinant *Vibrio harveyi* AasS(12) in buffer conditions amenable to solution phase NMR. Using conformationally sensitive urea-polyacrylamide gel electrophoresis (PAGE), the acylation state of AcpP was monitored over the course of a week, an amount of time typically required to perform triple resonance NMR experiments and analyze protein structure.

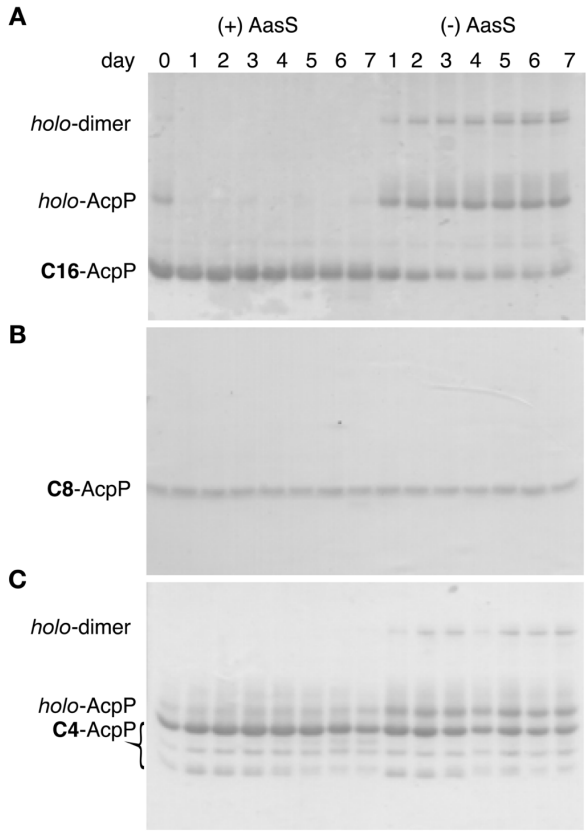


Figure 2.1. Acyl-AcpP stability time course of long, medium, and short chain lengths. C16-AcpP, C8-AcpP, and C4-AcpP are shown in A, B, and C, respectively. Purified acyl-AcpP and reaction components were monitored for a week in the presence or absence of AasS with 5 mM ATP, 5 mM $MgCl_2$, and 1 mM fatty acid in 50 mM phosphate buffer pH 7.4 by urea-PAGE.

Short-chain butyryl- (C4-), medium chain octanoyl (C8-), and long-chain palmitoyl- (C16-) fatty acids were evaluated, each showing significantly different stability profiles. C16-AcpP showed a dramatic enhancement by AasS. Without AasS, *holo*-AcpP and dimers of *holo*-AcpP immediately form, and their levels increase relative to C16-AcpP with each subsequent day. Excitingly, with AasS present, an entire week of C16-AcpP is afforded. C8-AcpP was expected to be the most stable, with the thioester protected by sequestration.(5,13,14) Accordingly, C8-AcpP was extremely stable both in the absence and presence of AasS for a week. C4-AcpP, which runs as multiple bands in urea-PAGE indicative of multiple sequestered states,(15) is also retained through AasS activity. In the absence of AasS, bands corresponding to *holo*-AcpP and *holo*-AcpP

dimers become clear after just 1 day, though about half of the population remains in the form of acyl-AcpP for an entire week (Figure 1).

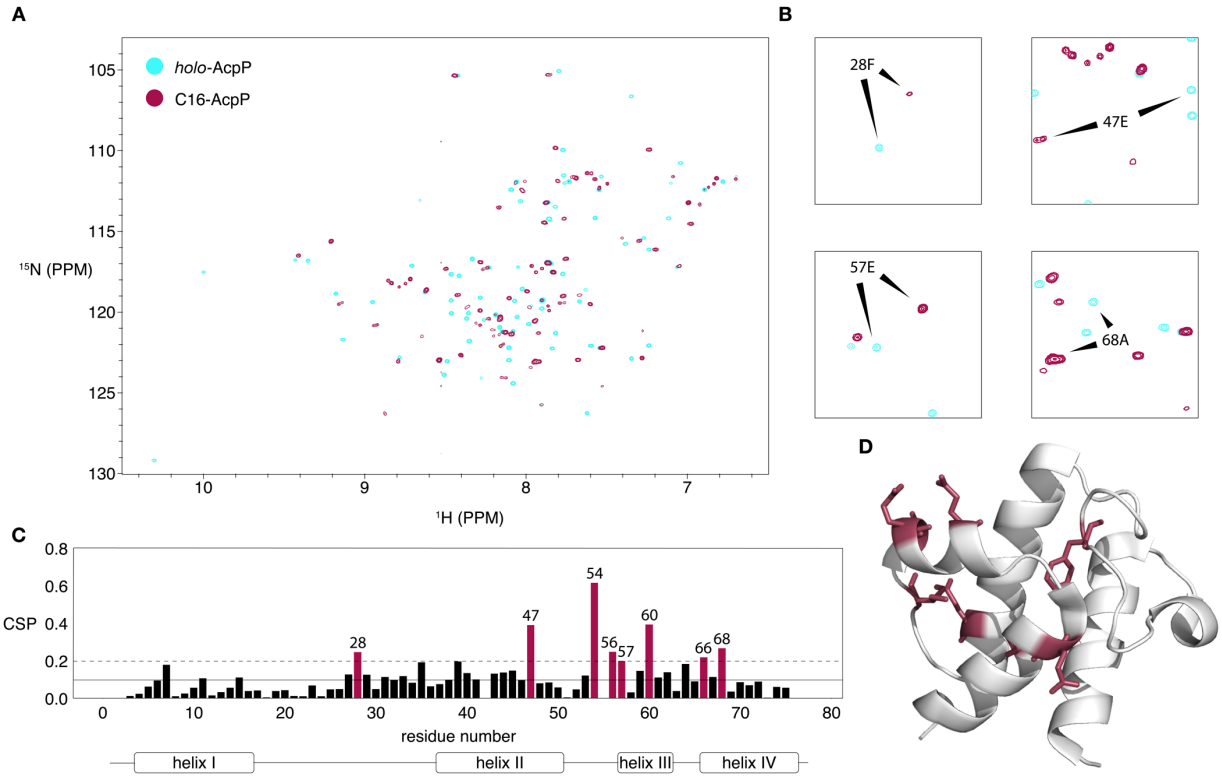


Figure 2.2. 2D NMR comparison of holo-AcpP and C16-AcpP. (A) Overlaid $^{1}H–^{15}N$ HSQC spectra of *holo*-AcpP and C16-AcpP in 50 mM phosphate buffer, pH 7.4. (B) Zoomed-in view of selected peaks with large chemical shift differences. (C) Chemical shift perturbation (CSP) plot comparing *holo*-AcpP to C16-AcpP. The solid line represents the mean CSP and the dashed line represents one standard deviation above the mean. Those CSP values greater than one standard deviation of the mean are colored in magenta. (D) Residues with CSPs larger than one standard deviation of the mean shown as sticks on AcpP crystal structure PDB ID 2FAD and colored magenta.

For NMR experiments, AcpP was expressed with ^{15}N isotope enrichment, while AasS was expressed with natural abundance and added to the NMR sample. Therefore, only the chemical shift of AcpP residues was detected as AasS served to counter the hydrolysis of acyl-AcpP to *holo*-AcpP. To first determine whether AasS would perturb the NMR spectra, we titrated increasing molar equivalents of AasS to the stable ^{15}N-C8-AcpP species, acquiring $^{1}H–^{15}N$ HSQC spectra at each titration point. The presence of AasS did not perturb the structure of AcpP; however,

significant signal loss preventing peak resolution was observed above AasS molar equivalents of 0.025 (Figure S1).

With an increase in stable acyl-AcpP, we sought to gain structural information about each species using solution NMR. $^{1}H–^{15}N$ HSQC spectra of *holo*-AcpP and C16-AcpP in the presence and absence of AasS were obtained over the course of 1 h and compared. Without AasS present, the spectra overlay exactly, indicating complete hydrolysis of the acylated species (Figure S2). In the presence of AasS, the C16-AcpP spectrum shown in Figure 2 was obtained. This data revealed that nearly every secondary structural region of AcpP is perturbed by addition of a palmitoyl group, including loop I, helix II, loop II, helix III, and helix IV (Figure 2).

The thioester bond attaching substrates to ACP is susceptible to hydrolysis, with longer acyl chains hydrolyzing more readily.(7) Molecular dynamics simulations have shown that the hydrophobic pocket of AcpP expands to accommodate growing acyl substrates up to 10 carbons in length before the acyl chains can no longer fit in the pocket.(16) This explains why longer chain lengths are less sequestered, leaving the thioester exposed and more readily susceptible to hydrolysis. Replacing the sulfur with more stable atoms has been used to stabilize this bond for NMR data collection, but these have been shown to perturb ACP structure, emphasizing that ACP is sensitive to even a one atom modification of the substrates.(8) The effect of substrate identity on ACP structure should provide further insight into the communication mechanisms which allow ACP to distinguish between over 25 enzymes in *E. coli*(17−19) and deliver substrates to the appropriate partner enzyme. The method described here to stabilize acyl-ACP by shifting the hydrolysis equilibrium through the use of AasS will allow for structural studies of substrate-loaded ACP that were previously inaccessible.

Acknowledgements:

We thank Prof. Stanley Opella for valuable NMR discussions and Dr. Xuemei Huang for assistance with the NMR facility use.

Chapter 2, in full, is a reprint of the material as it appears: Sztain, T.; Bartholow, T. G.; McCammon, J. A.; Burkart, M. D. Shifting the Hydrolysis Equilibrium of Substrate Loaded Acyl Carrier Proteins. *Biochemistry* **2019**, *58* (34), 3557–3560. The dissertation author was the primary author of this manuscript.

2.1 References

(1) Marrakchi, H., Zhang, Y. M., and Rock, C. O. (**2002**) Mechanistic diversity and regulation of Type II fatty acid synthesis. Biochem. Soc. Trans. 30, 1050−1055.

(2) Beld, J., Lee, D. J., and Burkart, M. D. (**2015**) Fatty acid biosynthesis revisited: structure elucidation and metabolic engineering. Mol. BioSyst. 11, 38−59.

(3) Staunton, J., and Weissman, K. J. (**2001**) Polyketide biosynthesis: a millennium review. Nat. Prod. Rep. 18, 380−416.

(4) Cronan, J. E. (**2014**) The chain-flipping mechanism of ACP (acyl carrier protein)-dependent enzymes appears universal. Biochem. J. 460, 157−163.

(5) Roujeinikova, A., Simon, W. J., Gilroy, J., Rice, D. W., Rafferty, J. B., and Slabas, A. R. (**2007**) Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. J. Mol. Biol. 365, 135−145.

(6) Crosby, J., and Crump, M. P. (**2012**) The structural role of the carrier protein − active controller or passive carrier. Nat. Prod. Rep. 29, 1111−1137.

(7) Zornetzer, G. A., Tanem, J., Fox, B. G., and Markley, J. L. (**2010**) The Length of the Bound Fatty Acid Influences the Dynamics of the Acyl Carrier Protein and the Stability of the Thioester Bond. Biochemistry 49, 470−477.

(8) Sztain, T., Patel, A., Lee, D. J., Davis, T. D., McCammon, J. A., and Burkart, M. D. (**2019**) Modifying the Thioester Linkage Affects the Structure of the Acyl Carrier Protein. Angew. Chem., Int. Ed. 58, 10888−10892.

(9) Goodrich, A. C., and Frueh, D. P. (**2015**) An NMR method to probe molecular influences of substrate loading in NRPS carrier proteins. Biochemistry 54, 1154−1156.

(10) Goodrich, A. C., Harden, B. J., and Frueh, D. P. (**2015**) Solution Structure of a Nonribosomal Peptide Synthetase Carrier Protein Loaded with Its Substrate Reveals Transient, Well-Defined Contacts. J. Am. Chem. Soc. 137, 12100−12109.

(11) Beld, J., Finzel, K., and Burkart, M. D. (**2014**) Versatility of acyl-acyl carrier protein synthetases. Chem. Biol. 21, 1293−1299.

(12) Jiang, Y., Chan, C. H., and Cronan, J. E. (**2006**) The soluble acyl-acyl carrier protein synthetase of Vibrio harveyi B392 is a member of the medium chain acyl-CoA synthetase family. Biochemistry 45, 10008−10019.

(13) Cronan, J. E. (**1982**) Molecular properties of short chain acyl thioesters of acyl carrier protein. J. Biol. Chem. 257, 5013−5017.

(14) Płoskon, E., Arthur, C. J., Kanari, A. L. P., Wattana-amorn, P.,́ Williams, C., Crosby, J., Simpson, T. J., Willis, C. L., and Crump, M. P. (**2010**) Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. Chem. Biol. 17, 776−785.

(15) Kosa, N. M., Haushalter, R. W., Smith, A. R., and Burkart, M. D. (**2012**) Reversible Chemoenzymatic Labeling of Native and Fusion Carrier Protein Motifs. Nat. Methods 9, 981−984.

(16) Chan, D. I., Stockner, T., Tieleman, D. P., and Vogel, H. J. (**2008**) Molecular Dynamics Simulations of the Apo-, Holo-, and Acyl-forms of Escherichia coli Acyl Carrier Protein. J. Biol. Chem. 283, 33620−33629.

(17) Nguyen, C., Haushalter, R. W., Lee, D. J., Markwick, P. R. L., Bruegger, J., Caldara-Festin, G., Finzel, K., Jackson, D. R., Ishikawa, F., O'Dowd, B., McCammon, J. A., Opella, S. J., Tsai, S.-C., and Burkart, M. D. (**2014**) Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. Nature 505, 427−431.

(18) De Lay, N. R., and Cronan, J. E. (**2007**) In vivo functional analyses of the type II acyl carrier proteins of fatty acid biosynthesis. J. Biol. Chem. 282, 20319−20328.

(19) Gully, D., and Bouveret, E. (**2006**) A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in Escherichia coli. Proteomics 6, 282−293.

Chapter 3. Decoding allosteric regulation by the acyl carrier protein

Terra Sztain, Thomas G. Bartholow, D. John Lee, Lorenzo Casalino, Andrew Mitchell, Megan A. Young, Jianing Wang, J. Andrew McCammon, and Michael D. Burkart*

Abstract

Enzymes in multistep metabolic pathways utilize an array of regulatory mechanisms to maintain a delicate homeostasis [K. Magnuson, S. Jackowski, C. O. Rock, J. E. Cronan, Jr, *Microbiol. Rev.* 57, 522–542 (1993)]. Carrier proteins in particular play an essential role in shuttling substrates between appropriate enzymes in metabolic pathways. Although hypothesized [E. Płoskoń et al., *Chem. Biol.* 17, 776–785 (2010)], allosteric regulation of substrate delivery has never before been demonstrated for any acyl carrier protein (ACP)-dependent pathway. Studying these mechanisms has remained challenging due to the transient and dynamic nature of protein–protein interactions, the vast diversity of substrates, and substrate instability [K. Finzel, D. J. Lee, M. D. Burkart, *ChemBioChem* 16, 528–547 (2015)]. Here we demonstrate a unique communication mechanism between the ACP and partner enzymes using solution NMR spectroscopy and molecular dynamics to elucidate allostery that is dependent on fatty acid chain length. We demonstrate that partner enzymes can allosterically distinguish between chain lengths via protein–protein interactions as structural features of substrate sequestration are translated from within the ACP four-helical bundle to the protein surface, without the need for stochastic chain flipping. These results illuminate details of cargo communication by the ACP that can serve as a foundation for engineering carrier protein-dependent pathways for specific, desired products.

Maintaining steady-state homeostasis of complex cellular processes involves propagation of signals across interconnected networks of molecular assemblies, and the organization of enzymes within metabolic pathways offers a substantive example of such regulatory control (**1**). A thorough molecular understanding of these processes can benefit modifications of these pathways, particularly for bottom-up approaches to metabolic pathway engineering (**2**). Fatty acid biosynthesis (FAB) has served as an archetypical system for interconnected metabolic enzymes that encompasses a broad range of fundamental reactions essential for all domains of life (**3**). Here, a central acyl carrier protein (ACP) delivers the correct intermediate to one of multiple enzymes, not only providing substrates for the acetate pathways of primary and secondary metabolism but also serving to regulate the cell cycle and other global processes (**4**). Within fatty acid synthase (FAS), a series of enzymes catalyze the iterative condensation of two-carbon ketogenic units to a growing acyl chain, where each elongation is followed by stepwise reduction of the resulting β-ketone to a saturated fatty acid precursor (**5**). The ACP from *Escherichia coli* (AcpP) carries acyl chain lengths from 4 to 18 carbons, each of which transitions through four oxidation states, and each intermediate must be delivered to the appropriate enzyme within the 27 AcpP-binding proteins of the known interactome (**6**–**8**). How these interactions are organized cannot be easily explained by stochastic sampling of each ACP-tethered substrate by catalytic or regulatory proteins. Here we evaluate the functioning of an allosteric regulatory framework, guided by substrate identity, that regulates the protein–protein interactions required for enzyme activity within the bacterial FAS.

Traditional models of protein allostery often refer to binding of effector molecules at distal binding sites which alter active site functionality (**9**). Rather than containing catalytic active sites, carrier proteins function upon binding to a partner enzyme, forming a protein–protein interaction.

Upon binding, substrate delivery to each enzyme within FAS requires a large conformational change by the ACP, termed "chain flipping." Here each acyl substrate, covalently tethered to a 4′-phosphopantetheine "arm" and sequestered within the ACP core, is translocated from the ACP pocket into the active site of the partner enzyme (**10**). Therefore, allostery described herein will refer to the classical definition which describes the phenomenon that indirect, specific protein sites are capable of modulating protein primary function (**11**). Here, the first site is within the ACP four-helical bundle, where cargos are sequestered, and the second site consists of the solvent exposed surface of this bundle protein, which interacts with catalytic partners.

Within FAB, many partner enzymes have active sites which perform specific chain elongation or reductive chemistries and are promiscuous to fatty acid chain length. However, some enzymes demonstrate chain length specificities that control the formation of key pathway products (*SI Appendix*, Fig. S1) (**12**–**17**). In these instances, two conflicting and still unproven hypotheses have been proposed concerning ACP-dependent enzyme regulation. One involves stochastic sampling, where each enzyme randomly binds ACP and samples ACP-tethered content via chain flipping into the active site (**18**). Another involves enzyme regulation via allosteric communication with the acyl-chain bearing ACP regulating enzyme reactivity, and ultimately controlling pathway processivity, without the need for chain flipping (**18**–**20**). Here we uncover the fundamental role of allostery in the transfer of octanoic acid from AcpP to the enzyme LipB, which re-directs the fatty acid from FAB to lipoic acid biosynthesis, demonstrating that the structural changes imparted by substrate sequestration are leveraged to control enzyme activity without the need for substrate sampling by the active site.

Characterization of ACPs bearing different chain lengths is inherently challenging to evaluate due to the lability of the 4′-phosphopantetheine thioester bond attachment of certain chain

lengths. Leveraging a recently described methodology to enzymatically reverse acyl hydrolysis in situ (**21**), we acquired ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR spectra of every physiological acyl chain length attached to AcpP. We additionally performed molecular dynamics (MD) simulations that provide a complete view of how AcpP structure is affected by sequestration of each chain length. Using solution NMR HSQC titration experiments, we discovered that the length of the acyl chain is communicated through differences in AcpP–partner protein interactions with LipB and the cognate (C8-AcpP) and noncognate (C12-AcpP) chain lengths. With selective ^{13}C labeling of the acyl chain termini, we observed that chain flipping occurs only when the specific cognate acyl chain is sequestered by AcpP. Next, we demonstrated that abrogation of this specific chain flipping could be achieved by mutation of a single surface residue discovered at the AcpP–LipB interface from MD simulations. These findings support the long-postulated hypothesis of allosteric control of ACP-dependent pathways by delineating each step in the process of acyl chain sequestration, partner-protein recognition, and chain flipping at the molecular level.

3.1 Solution NMR Elucidates Key Residues

We recently described a method that opposes hydrolysis by incorporating the acyl attachment enzyme AasS, which provided ^{1}H-^{15}N HSQC spectra of *holo*-, C4-, C8-, and C16-AcpP (**21**). In this study, we acquired remaining ^{1}H-^{15}N HSQC of AcpP carrying C6, C10, C12, C14, and C18 acyl chain lengths, allowing us to overlay spectra of every biological chain length of *E. coli* FAB (**Fig. 1*A***). Remarkably, overlay of these spectra revealed a distinct chemical shift pattern. Though many residues of AcpP maintained a fairly similar chemical environment regardless of chain length, a few residues were significantly perturbed and followed a visible

correlation between chain length and chemical shift (**Fig. 1 *B* and *C***). Linear regression analysis of the chemical shifts versus chain length revealed 10 residues with R^2 values >0.9 (Datasets S1 and S2). These 10 residues, V7, S27, F28, D38, L42, E47, I54, E57, I62, and T64 (**Fig. 1*D***), are highly conserved across multiple species (**22**), suggesting they may be critical for ACP communicating the length of the acyl chain to partner enzymes.

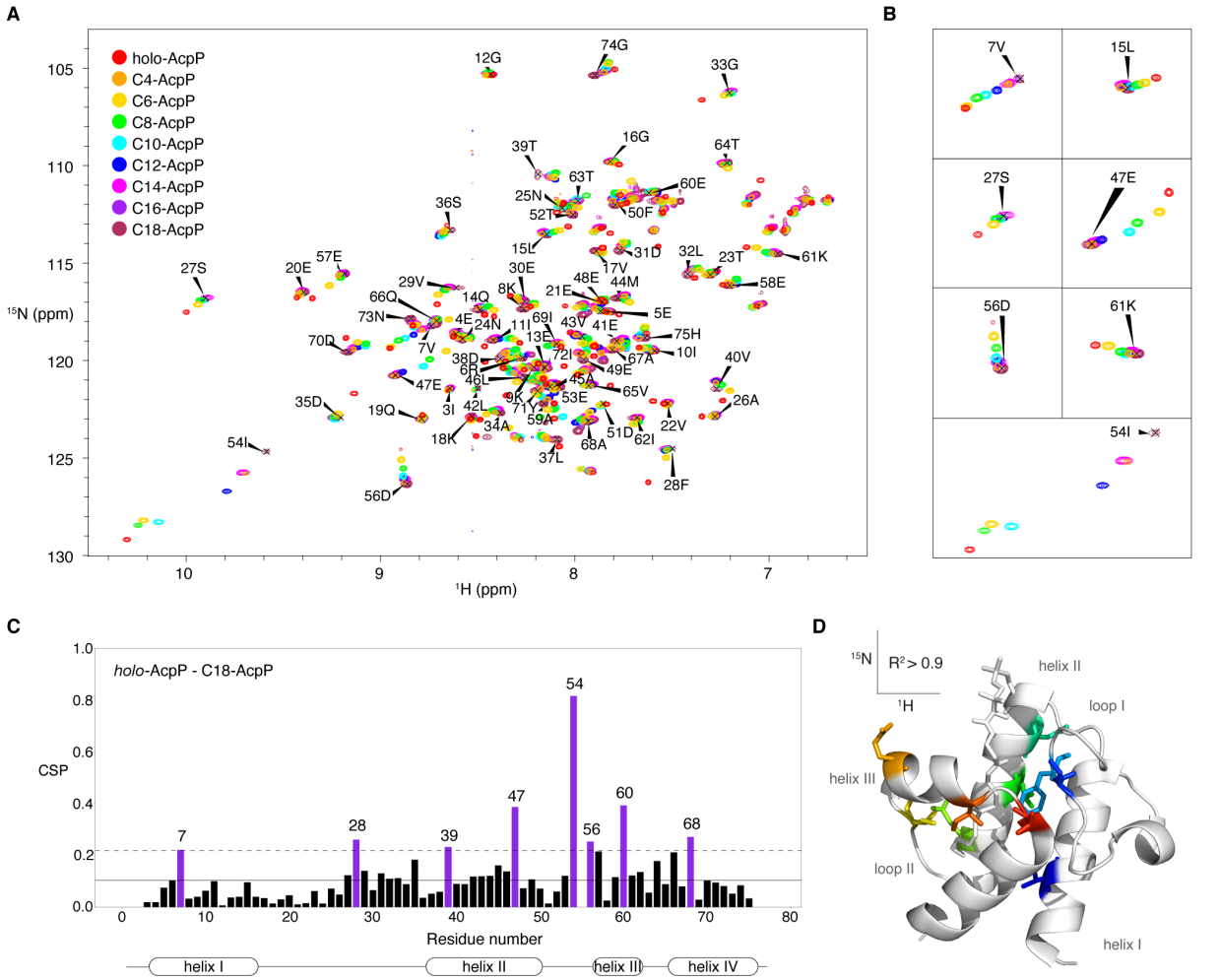


Figure 3.1. Solution NMR analysis of every chain length. (*A*) ^{1}H-^{15}N HSQC NMR spectra of AcpP carrying each biological chain length fatty acid. (*B*) Selected residues demonstrating chain length-dependent chemical shift migration are enlarged, with background peaks removed for clarity. (*C*) Chemical shift perturbation plot between *holo*-AcpP and C18-AcpP with the mean shown as a solid black line and the SD shown as a dashed line. Residues above 1 SD from the mean are colored in purple. (*D*) AcpP with residues with chemical-shift perturbations correlated to chain length with R^2 value >0.9 is shown as sticks and colored in rainbow on PDB ID code 2FAD. ppm, parts per million.

Isoleucine 54 shows the greatest magnitude chemical shift perturbation (CSP) between nearly every spectra, including between *holo*-AcpP and C4-AcpP, C4-AcpP and C6-AcpP, C10-AcpP and C12-AcpP, C12-AcpP and C14-AcpP, C16-AcpP and C18-AcpP, and *holo*-AcpP and C18-AcpP (**Fig. 1*C*** and *SI Appendix*, Fig. S2). I54 lines the bottom of the hydrophobic pocket in the loop between helices II and III. Mutations of I54 have shown it is critical for the structural integrity of AcpP (**20**) and binding to partner proteins (**23**). The CSPs between C6-AcpP and C8-AcpP, C8-AcpP and C10-AcpP, and C14-AcpP and C16-AcpP were not large in magnitude for I54, and the overall magnitude of the chemical shifts was small. This could help explain how some partner enzymes may recognize a class of medium (C6, C8, C10) or long (C14, C16) chain lengths, rather than specifically distinguishing the chain length.

Residues V7, F28, L42, I54, and I62 line the hydrophobic pocket of AcpP in crystal and NMR structures (**20**, **24**). Residues D38, E47, and E57 make up the acidic binding face of helices II and III. This suggests a communication mechanism by which the length of the fatty acid perturbs the residues in the pocket of AcpP, leading to perturbations in residues on the surface of AcpP which directly participate in partner enzyme binding.

3.2 Molecular Dynamics Simulations Reveal Structural Signatures and Correlated Motions

To gain a better understanding of the structural implications of the observed chemical shift trends, we performed MD simulations of AcpP carrying every biological chain length of *E. coli* FAB, except for the C8 simulation which we've previously reported (**25**) and further analyzed here. Five independent simulations of 500 ns for each acyl chain length were carried out, all modeled from the crystal structure of heptanoyl-AcpP (Protein Data Bank [PDB] ID code 2FAD) (**20**). For each of the 10 key residues identified during NMR, the distance between that residue and

the three main components of the ligand were monitored throughout the simulation and compared for each chain length. The main components were considered to be the phosphorus of 4′-phosphopantetheine, the linkage sulfur, or the terminal carbon of the acyl chain. For each of the AcpP residues, either one or two of the calculated distances was correlated to chain length (**Fig. 2*A*** and *SI Appendix*, Fig. S3). All of these distances increased with chain length, except for the distance between residues V7, I54, and E47 to the terminal carbon, which decreased with increasing chain length (*SI Appendix*, Fig. S3). This corresponded to a widening of helices II and III and loop III with increased chain length (**Fig. 2*B***), and a tightening of the bottom pocket including helices I and II and loop II. Overall, an increase in the radius of gyration was observed with increasing chain length (**Fig. 2*C***). The correlated motions of each residue were quantified, revealing high correlation in all simulations of the back face of AcpP, consisting of loops I and III and helix IV (*SI Appendix*, Fig. S4). Helices II and III were anticorrelated in all simulations, indicating a possible mechanism whereby the helices open apart in opposing directions for substrate delivery and close together for substrate sequestration. This is in agreement with the mechanism observed in a steered MD study where helix III dynamics facilitated substrate delivery from ACP to FabZ (**26**). Reordering the correlation matrix by internal versus external residues shows a strong positive correlation between the interior pocket and exterior surface of AcpP (**Fig. 2*D*** and *SI Appendix*, Fig. S5). This supports an allosteric substrate communication mechanism whereby the sequestered acyl chain affects the conformation of interior helix residues which is translated to the exterior, enzyme binding interface residues (**Fig. 2 *E* and *F***).

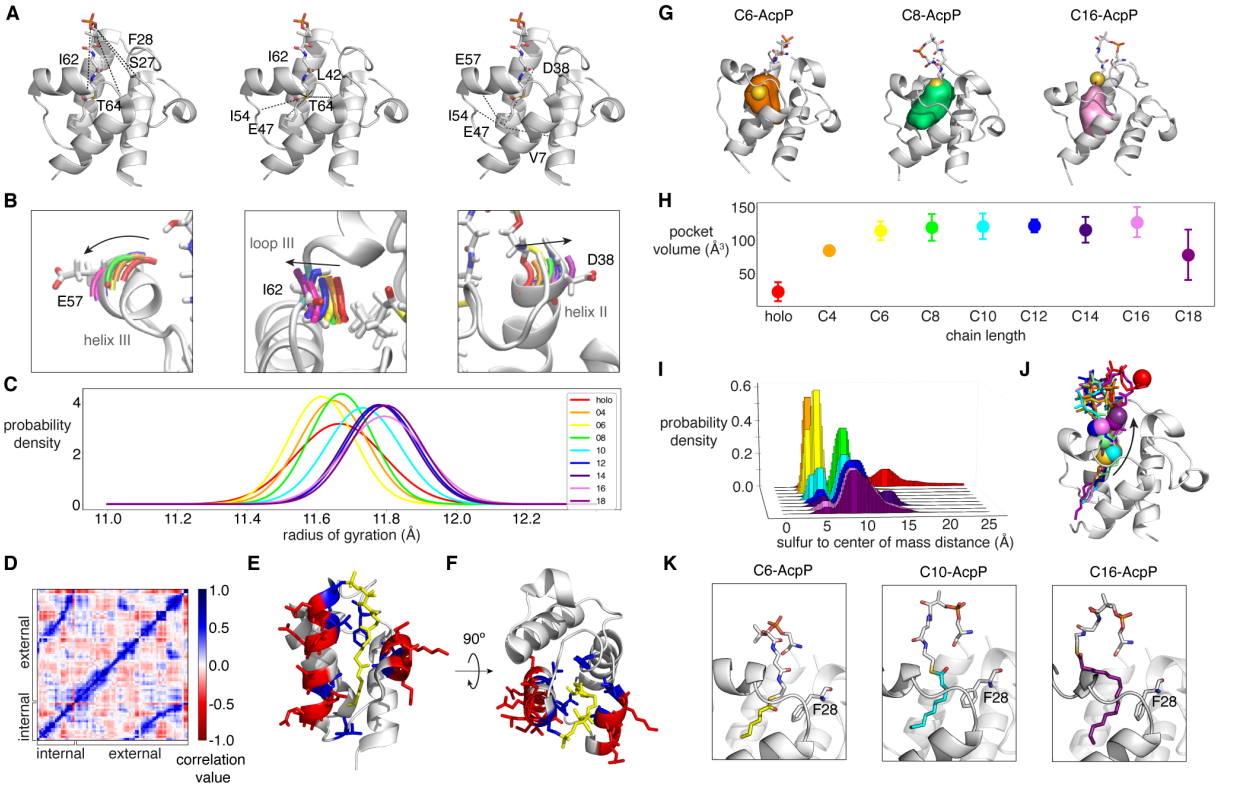


Figure 3.2. MD analysis of acyl chain length on the AcpP structure. (*A*) Distances which are correlated with chain length with R^2 >0.6 of residue to phosphorus, sulfur, or the terminal carbon shown on PDB ID code 2FAD. (*B*) Backbone residues shown in tube representation with increasing chain lengths colored and shown with black arrows. (*C*) Normalized probability density of the radius of gyration calculated from the total 2.5 μs simulated of each chain length. (*D*) Correlated motion map reordered to show correlations between internal and external residues. (*E*) AcpP from PDB ID code 2FAD with the acyl chain colored in yellow, internal residues colored in blue, and select external residues colored in red. (*F*) Top-down view of *E*. (*G*) Representative structures of C6-, C8-, and C16-AcpP with the pocket shown in orange, green, and pink, respectively. The sulfur atom is shown as a gold sphere. (*H*) Calculated pocket volume for each chain length. Error bars represent deviation from the mean from five independent simulations. (*I*) Histograms of the distance from the sulfur atom to the center of mass of acyl-AcpP calculated from the total 2.5 μs simulation of each acyl-AcpP. (*J*) Structure representing the distance calculated in *I* with sulfur shown as a gold sphere. The black arrow indicates the direction of the sulfur atom as the distance to center of mass is increased with increased chain length. (*K*) Representative structures of C6-, C10-, and C16-AcpP from MD simulations with acyl chains shown relative to residue F28.

3.3 Substrate Sequestration and the 4′-Phosphopantetheine Hairpin

Expansion of the hydrophobic pocket to accommodate increasing chain lengths has been observed in crystal structures (**27**) and MD simulations (**28**). The simulations performed by Chan et al. (**28**) also revealed that the pocket reaches a maximum expansion volume (**Fig. 2 *G* and *H***), helping to explain the increased solvent accessibility of the thioester and increased hydrolysis rates

of longer chain lengths (**29**). However, these simulations were performed for only 20 ns each. Our 500 ns simulations showed that long chain lengths built from medium chain length structures (heptanoyl-AcpP) required up to 250 ns to fully equilibrate to the ideal substrate sequestration (*SI Appendix*, Fig. S6).

This was determined by calculating the distance between the sulfur and center of mass of AcpP. Three distinct sequestration states were observed with distances of ∼5 Å (tightly sequestered), 6 to 8 Å (loosely sequestered), and 13 to 15 Å (unsequestered) (**Fig. 2 *I* and *J***). *Holo*-AcpP was extremely dynamic and sampled all three states throughout the simulation. Short chains C4- and C6-AcpP stayed tightly sequestered, while medium chains C8- and C10-AcpP spent some time in the tightly sequestered state, though they spent the majority of the simulation in the loosely sequestered state. C12-AcpP spent even less time tightly sequestered, most time loosely sequestered, and sampled the unsequestered state. Long chains C14-, C16-, and C18-AcpP spent the majority of the simulation in the loosely sequestered state and also sampled the unsequestered state. Histogram peaks corresponding to tight sequestration disappeared for C14-, C16-, and C18-AcpP when the beginning of the trajectories was not included in the calculation but did not decrease for C12-AcpP. Complete loss of these peaks required 250 ns (*SI Appendix*, Fig. S6). Zornetzer et al. (**29**) previously reported NMR experiments showing that chain length can affect dynamics of spinach ACP.

Trajectories representing principal component analysis of each simulation revealed structural and dynamic properties of the motion underlying the speculated “hairpin” formation of 4′-phosphopantetheine and corresponded to the distance between the sulfur and center of mass in many simulations (*SI Appendix*, Figs. S7 and S8 and Movie S1). The key residue F28 appears to facilitate the hairpin formation, acting as a ruler by which only around eight carbons can fit

securely in the hydrophobic pocket (**Fig. 2*K***). The acyl chain was observed sliding in and out of the hydrophobic pocket, equilibrating to, and then remaining dynamic within one of the three sequestration states in each of the trajectories.

3.4 AcpP Communicates Substrate Identity through Protein–Protein Interactions

In order to determine how AcpP uses these key residues during protein–protein interactions with a partner enzyme, we performed NMR titration experiments with the chain length-specific enzyme LipB. This enzyme selectively harnesses C8-AcpP from FAB for the biosynthesis of lipoic acid. We increased molar equivalents of inactive LipB (C169A active site mutation to prevent hydrolysis or covalent attachment) (**30**) with ^{15}N-labeled C8-AcpP or C12-AcpP and acquired ^{1}H-^{15}N HSQC spectra at each titration point up to 1.5 equivalents of LipB (**Fig. 3 *A–D***). Though a few peaks were lost due to signal dampening in the presence of LipB, a clear difference in AcpP perturbation by LipB was observed depending on chain length. The cognate, C8-AcpP was perturbed at helices II and III, whereas the noncognate C12-AcpP was only perturbed at helix II, and the magnitude of perturbations was greater than that of C8-AcpP (**Fig. 3 *E*, *F*, and *H***). Lineshape analysis was carried out using TITAN software (**31**) to determine the K_d and k_{off} from the titration data, showing larger values for each of C12-AcpP compared with C8-AcpP. This demonstrates both a kinetic and thermodynamic favoring toward binding to AcpP when the specific substrate is bound. The key residue, E57, was perturbed greater than 1 SD from the mean by LipB for C8-AcpP but not C12-AcpP. The more open position and outward rotation formed by C12-AcpP compared with C8-AcpP observed in MD simulations (**Fig. 2*C***) are likely not as

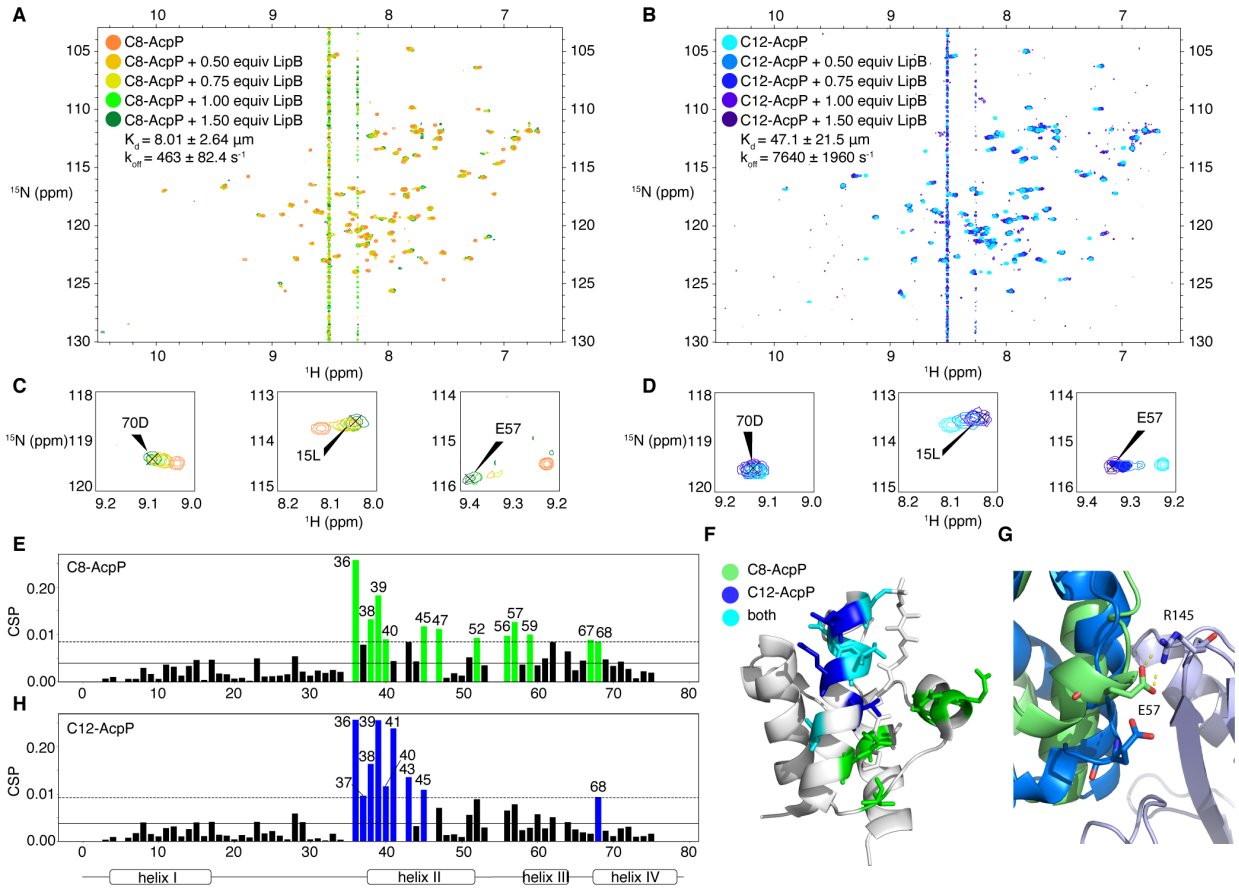


Figure 3.3. Protein–protein interaction analysis of cognate and noncognate chain lengths via NMR titration and docking. (*A*) ^{1}H-^{15}N HSQC spectrum overlay of C8-AcpP titrated with increasing equivalents of LipB. (*B*) ^{1}H-^{15}N HSQC spectrum overlay of C12-AcpP titrated with increasing equivalents of LipB. The interfering signals around 8.3 and 8.5 ppm are a result of t_1 noise due to the presence of adenosine triphosphate at high concentration within the buffer components (*SI Appendix*, Fig. S10), which was used for consistency with **Fig. 1** as detailed in *SI Appendix*, *NMR Methods*. (*C* and *D*) Enlarged view of selected peaks from *A* and *B*, respectively. (*E* and *H*) CSP plots quantifying *A* and *B*, respectively. The solid line represents the mean, and the dashed line represents 1 SD from the mean. Residues with CSP values above 1 SD from the mean are colored in green or blue. (*F*) CSP residues colored on PDB ID code 2FAD based on whether they are above 1 SD from the mean in C8-AcpP titration only (green), C12-AcpP titration only (blue), or both (cyan). (*G*) Docking interface of LipB (purple) to C8-AcpP (green) and C12-AcpP (blue).

suitable of a binding interface for LipB. This was supported by docking a homology model of *E. coli* LipB derived from *Thermus thermophilus* LipB PDB ID code 2QHS (**32**) (*SI Appendix*, Fig. S9) to MD-derived C8-AcpP and C12-AcpP structures, where hydrogen bonds were observed between R145 of LipB and E57 in C8-AcpP but not in the C12-AcpP structure. In the C12-AcpP

structure, helix III is shifted to accommodate the longer acyl chain through interaction of I54 (**Fig. 1 *A* and *B***), disallowing E57 from forming a salt bridge with R145 (**Fig. 3*G***).

3.5 Substrate-Specific Chain Flipping

To determine whether these differences in protein–protein interactions play a significant role in communicating chain length identity, we sought to determine whether chain flipping could still occur with a noncognate acyl chain length. Previous efforts to visualize chain flipping have involved attachment of nonphysiological cargo, such as fluorescent or solvatochromic adducts (**33**). Here we designed an NMR experiment to detect chain flipping of a native acyl-AcpP (**Fig. 4*A***). We obtained ^{13}C-labeled octanoic and dodecanoic acids selectively labeled at the terminal carbon (ω-1-^{13}C). Acquisition of ^{1}H-^{13}C HSQC NMR spectra of the fatty acids alone in solution gave peaks with distinct chemical shifts compared with those when covalently attached to AcpP (*SI Appendix*, Fig. S11). Addition of 1.5 equivalents of LipB (C169A) resulted in emergence of a second peak, apparently undergoing a slow exchange, with only C8-AcpP (**Fig. 4 *B* and *C***) and not C12-AcpP (**Fig. 4 *E* and *F***). The new peak, significantly downfield from the free or AcpP-sequestered peak, indicates a much less shielded environment upon chain flipping into the LipB pocket. Several smaller peaks are present in this spectrum, which we hypothesize may be metastable intermediates that make up less than 10% of the intensity of the upfield peak and were not included in quantifying the percentage of sequestered and flipped peak intensity.

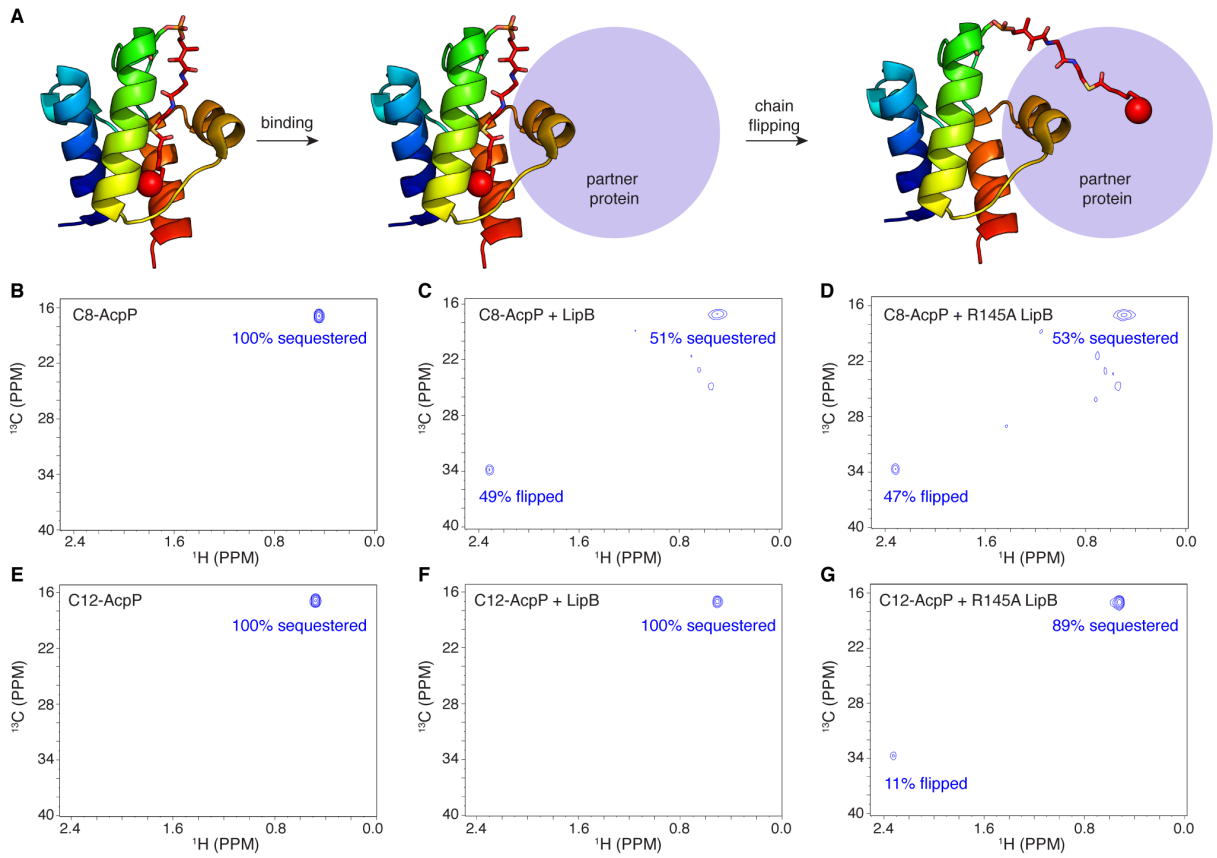


Figure 3.4. Chain flipping of cognate and noncognate chain lengths. (*A*) Schematic of chain flipping, with the terminal carbon which was selectively ^{13}C-labeled in the following experiments shown as a red sphere. (*B–D*) ^{1}H-^{13}C HSQC NMR spectra of C8-AcpP with the terminal C8 ^{13}C-labeled. The relative intensities of each peak are labeled as a percentage if they are >10% intensity compared with the sequestered peak. Spectra were acquired with C8-AcpP alone (*B*), in the presence of 1.5 equivalents of C169A LipB (*C*), and in the presence of 1.5 equivalents of R145A C169A LipB mutant (*D*). (*E–G*) ^{1}H-^{13}C HSQC NMR spectra of C12-AcpP with the terminal C12 ^{13}C-labeled. Spectra were acquired with C12-AcpP alone (*E*), in the presence of 1.5 equivalents of C169A LipB (*F*), and in the presence of 1.5 equivalents of R145A C179A LipB mutant (*G*).

To confirm the identity of the chain-flipped peak, we compared the spectra of both ω-1-^{13}C–labeled free fatty acids (C8 and C12) in the presence of only LipB. In the case of octanoic acid with LipB, there were no peaks resembling the free fatty acid, only the intense peak in the same location as the chain-flipped peak from the C8-AcpP study, along with similar minor downfield peaks (*SI Appendix*, Fig. S11*A*), indicative of binding to the hydrophobic pocket of the LipB active site. The spectrum of dodecanoic acid, however, showed little change upon addition of LipB (*SI Appendix*, Fig. S12*B*). Binding of free octanoic acid within the LipB active site has

been observed by X-ray crystallography following crystal soaking experiments of the *T. thermophilus* LipB with octanoic acid (**32**), confirming our assignment of fatty acid binding within the LipB active site. We therefore confidently assigned this downfield peak as "flipped" compared to the AcpP "sequestered."

The chain flipping detection experiment indicated that chain flipping is indeed avoided by the chain length-specific partner enzyme LipB when AcpP bears a noncanonical acyl chain length. The nature of this specific chain flipping is apparently due to the differences in protein–protein interactions of LipB with AcpP depending on whether AcpP carries the correct acyl chain length. To test this, we designed an interface mutant guided by the NMR and MD characterization of AcpP structure with each acyl chain length.

The residue E57 was one of the 10 key residues found to be important for chain length communication by NMR (**Fig. 1*D***) and MD (**Fig. 2 *C* and *D***). Residue E57 was also perturbed by LipB, with CSP greater than 1 SD from the mean, in C8-AcpP but not C12-AcpP (**Fig. 3 *E* and *H***), and participated in hydrogen bonding with R145 when docked to LipB in the C8-AcpP but not C12-AcpP structure (**Fig. 3*G***). Therefore, we generated an R145A mutant of LipB and evaluated chain flipping with selective ^{13}C-labeled C8-AcpP and C12-AcpP. This experiment identified chain flipping of both C8-AcpP and C12-AcpP species (**Fig. 4 *D* and *G***), indicating the acyl chain length specificity was disrupted by eliminating this interfacial interaction. We suspect that LipB R145 behaves as a gating residue that is opened upon salt bridge formation with AcpP E57 only when the proper binding orientation is formed, which occurs with C8-AcpP but not C12-AcpP. In this case, elimination of the gate leads to a more promiscuous LipB, which also induces chain flipping from C12-tethered AcpP.

3.6 Conclusion

Metabolic pathways in nature rely on an intricately choreographed series of reactions carried out by many enzymes. Advances in structural techniques can enable detailed accounts of the regulatory mechanisms governing these pathways. One central metabolic protein that has eluded full understanding is the ACP from fatty acid biosynthesis. This small, 77-residue protein interacts with dozens of enzymes and regulatory proteins, delivering each the appropriate acyl chain length and oxidation state substrate. The role of ACP as an active controller or passive carrier (**18**), and whether the universal chain flipping mechanism (**10**) is a stochastic event required for substrate recognition or triggered by a specific binding event, has been under discussion for decades. These results demonstrate that a specific pattern of structural changes in ACP can communicate the length of the acyl chain to a partner enzyme via protein–protein interactions, thereby allosterically regulating the chain flipping event. These findings represent a unique paradigm for understanding multistep, carrier protein-dependent biosynthetic pathways. The highly detailed evaluation of AcpP bearing each fatty acid chain length provided here will serve as a foundation for gaining more detailed control for engineering specific acyl chain length products or designing narrow-spectrum antibiotics targeting specific protein–protein interactions.

Acknowledgements:

We thank Prof. Stanley Opella and Prof. Lalit Deshmukh for helpful NMR discussions. We are also grateful for Dr. Xuemei Huang's assistance with NMR acquisition. T.S. is an NSF Graduate Research Fellowship Program recipient under Grant DGE-1650112. This work was supported by NIH GM031749 and GM095970.

Chapter 3, in full, is a reprint of the material as appears: Sztain, T.; Bartholow, T. G.; Lee, D. J.; Casalino, L.; Mitchell, A.; Young, M. A.; Wang, J.; McCammon, J. A.; Burkart, M. D. Decoding Allosteric Regulation by the Acyl Carrier Protein. *Proc. Natl. Acad. Sci.* **2021**, *118* (16), e2025597118. The dissertation author was the primary author of the manuscript.

3.7 References

1. M. Whiteley, S. P. Diggle, E. P. Greenberg, Progress in and promise of bacterial quorum sensing research. Nature **551**, 313–320 (2017).

2. J. Beld, D. J. Lee, M. D. Burkart, Fatty acid biosynthesis revisited: Structure elucidation and metabolic engineering. Mol. Biosyst. **11**, 38–59 (2015).

3. C. O. Rock, S. Jackowski, Forty years of bacterial fatty acid synthesis. Biochem. Biophys. Res. Commun. **292**, 1155–1166 (2002).

4. K. Magnuson, S. Jackowski, C. O. Rock, J. E. Cronan, Jr, Regulation of fatty acid biosynthesis in *Escherichia coli*. Microbiol. Rev. **57**, 522–542 (1993).

5. D. A. Hopwood, D. H. Sherman, Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. Annu. Rev. Genet. **24**, 37–66 (1990).

6. C. Nguyen, R. W. Haushalter, D. J. Lee, P. R. L. Markwick, J Bruegger, G. Caldara-Festin, K. Finzel, D. R. Jackson, F. Ishikawa, B. O'Dowd, J. A. McCammon, S. J. Opella, S.-C. Tsai, M. D. Burkart, Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. Nature **505**, 427–431 (2014).

7. N. R. De Lay, J. E. Cronan, In vivo functional analyses of the type II acyl carrier proteins of fatty acid biosynthesis. J. Biol. Chem. **282**, 20319–20328 (2007).

8. D. Gully, E. Bouveret, A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in *Escherichia coli*. Proteomics **6**, 282–293 (2006).

9. R. Nussinov, C.-J. Tsai, B. Ma, The underappreciated role of allostery in the cellular network. Annu. Rev. Biophys. **42**, 169–189 (2013).

10. J. E. Cronan, The chain-flipping mechanism of ACP (acyl carrier protein)-dependent enzymes appears universal. Biochem. J. **460**, 157–163 (2014).

11. J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: A plausible model. J. Mol. Biol. **12**, 88–118 (1965).

12. S. W. Jordan, J. E. Cronan, Jr, The *Escherichia* coli lipB gene encodes lipoyl (octanoyl)-acyl carrier protein:protein transferase. J. Bacteriol. **185**, 1582–1589 (2003).

13. C. R. H. Raetz, Z. Guan, B. O. Ingram, D. A. Six, F Song, X. Wang, J. Zhao, Discovery of new biosynthetic pathways: The lipid A story. J. Lipid Res. **50** (suppl.), S103–S108 (2009).

14. J. P. Issartel, V. Koronakis, C. Hughes, Activation of *Escherichia coli* prohaemolysin to the mature toxin by acyl carrier protein-dependent fatty acylation. Nature **351**, 759–761 (1991).

15. M. J. Grisewood N. J. Hernández-Lozada, J. B. Thoden, N. P. Gifford, D. Mendez-Perez, H. A. Schoenberger, M. F. Allan, M. E. Floy, R.-Y. Lai, H. M. Holden, B. F. Pfleger, C. D. Maranas, Computational redesign of acyl-ACP thioesterase with improved selectivity toward medium-chain-length fatty acids. ACS Catal. **7**, 3837–3849 (2017).

16. Y.-M. Zhang, C. O. Rock, Thematic review series: Glycerolipids. Acyltransferases in bacterial glycerophospholipid synthesis. J. Lipid Res. **49**, 1867–1874 (2008).

17. R. J. Heath, C. O. Rock, Roles of the FabA and FabZ β-hydroxyacyl-acyl carrier protein dehydratases in *Escherichia coli* fatty acid biosynthesis. J. Biol. Chem. **271**, 27795–27801 (1996).

18. J. Crosby, M. P. Crump, The structural role of the carrier protein—Active controller or passive carrier. Nat. Prod. Rep. **29**, 1111–1137 (2012).

19. E. Płoskoń, C. J. Arthur, A. L. P. Kanari, P. Wattana-amorn, C. Williams, J. Crosby, T. J. Simpson, C. L. Willis, M. P. Crump, Recognition of intermediate functionality by acyl carrier protein over a complete cycle of fatty acid biosynthesis. Chem. Biol. **17**, 776–785 (2010).

20. A. Roujeinikova, W. J. Simon, J. Gilroy, D. W. Rice, J. B. Rafferty, A. R. Slabas, Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. J. Mol. Biol. **365**, 135–145 (2007).

21. T. Sztain, T. G. Bartholow, J. A. McCammon, M. D. Burkart, Shifting the hydrolysis equilibrium of substrate loaded acyl carrier proteins. Biochemistry **58**, 3557–3560 (2019).

22. D. M. Byers, H. Gong, Acyl carrier protein: Structure-function relationships in a conserved multifunctional protein family. Biochem. Cell Biol. **85**, 649–662 (2007).

23. S. Angelini, L. My, E. Bouveret, Disrupting the acyl carrier protein/SpoT interaction in vivo: Identification of ACP residues involved in the interaction and consequence on growth. PLoS One **7**, e36111 (2012).

24. B.-N. Wu, Y.-M. Zhang, C. O. Rock, J. J. Zheng, Structural modification of acyl carrier protein by butyryl group. Protein Sci. **18**, 240–246 (2009).

25. T. Sztain, A. Patel, D. J. Lee, T. D. Davis, J. A. McCammon, M. D. Burkart, Modifying the thioester linkage affects the structure of the acyl carrier protein. Angew. Chem. Int. Ed. Engl. **58**, 10888–10892 (2019).

26. F. Colizzi, M. Masetti, M. Recanatini, A. Cavalli, Atomic-level characterization of the chain-flipping mechanism in fatty-acids biosynthesis. J. Phys. Chem. Lett. **7**, 2899–2904 (2016).

27. A. Roujeinikova, C. Baldock, W. J. Simon, J. Gilroy, P. J. Baker, A. R. Stuitje, D. W. Rice, J. B. Rafferty, A. R. Slabas, Crystallization and preliminary X-ray crystallographic studies on acyl-(acyl carrier protein) from *Escherichia coli*. Acta Crystallogr. D Biol. Crystallogr. **58**, 330–332 (2002).

28. D. I. Chan, T. Stockner, D. P. Tieleman, H. J. Vogel, Molecular dynamics simulations of the apo-, holo-, and acyl-forms of *Escherichia coli* acyl carrier protein. J. Biol. Chem. **283**, 33620–33629 (2008).

29. G. A. Zornetzer, J. Tanem, B. G. Fox, J. L. Markley, The length of the bound fatty acid influences the dynamics of the acyl carrier protein and the stability of the thioester bond. Biochemistry **49**, 470–477 (2010).

30. X. Zhao, J. R. Miller, J. E. Cronan, The reaction of LipB, the octanoyl-[acyl carrier protein]:protein *N*-octanoyltransferase of lipoic acid synthesis, proceeds through an acyl-enzyme intermediate. Biochemistry **44**, 16737–16746 (2005).

31. C. A. Waudby, A. Ramos, L. D. Cabrita, J. Christodoulou, Two-dimensional NMR lineshape analysis. Sci. Rep. **6**, 24826 (2016).

32. D. J. Kim S. J. Lee, H. S. Kim, K. H. Kim, H. H. Lee, H. J. Yoon, S. W. Suh, Structural basis of octanoic acid recognition by lipoate-protein ligase B. Proteins **70**, 1620–1625 (2008).

33. J. Beld, H. Cang, M. D. Burkart, Visualizing the chain-flipping mechanism in fatty-acid biosynthesis. Angew. Chem. Int. Ed. Engl. **53**, 14456–14461 (2014).

Chapter 4. Metabolic pathway engineering via protein-protein interface design

Terra Sztain[†], Joshua Corpuz[†], Thomas G. Bartholow, J. Andrew McCammon, Michael D. Burkart.

[†] these authors contributed equally.

Abstract

Carrier protein dependent pathways produce metabolites with pharmaceutical, environmental, and industrial properties. These pathways rely on an eloquent set of communication mechanisms involving protein-protein interactions (PPIs). Rational design of design of PPIs remains a challenge due to the complex nature of protein binding. Computational methods to design proteins are an encouraging avenue, though many score functions fail to predict experimental observables, leading to low success rates. Here we improve upon the Rosetta score function, leveraging experimental data through iterative rounds of computational prediction and mutagenesis, to design a hybrid non-ribosomal peptide synthetase — fatty acid biosynthesis pathway. By increasing the weight of electrostatic score term, the computational protocol was more predictive and required less rounds of iteration to identify mutants with high *in vitro* activity. This allowed efficient design of new PPIs between the adenylation domain Pltf, and the acyl carrier protein AcpP, leading to biosynthesis of new to nature natural products *in vivo*. This method provides a platform for improving upon Rosetta designs for customized purposes, and sets a new standard for carrier protein dependent pathway engineering.

4.1 Introduction

The ability to design and evolve catalysis between unrelated primary and secondary metabolic players remains a long-elusive goal for the carrier protein dependent pathways that include fatty acid synthase (FAS), polyketide synthase (PKS), and non-ribosomal peptide synthetase (NRPS). Although not broadly appreciated until recently, protein-protein interactions (PPI) have been demonstrated to be responsible for enzyme selectivity and pathway organization in these carrier protein dependent pathways.[1,] Within recent years, these PPIs have been structurally characterized in solution,[1,2]crystal structures,[3–6] and through microscopy;[7] predicted in silico;[2] abrogated by mutagenesis;[1] and improved by rational design.[8] What has remained uncharted, however, is the design of PPIs between non-reactive carrier protein–enzyme pairs for gain-of-function activity. The ability to rapidly engineer enzyme selectivity has the potential to revolutionize our ability to design and control these metabolic pathways. Here we demonstrate development of a computational platform for evolving PPIs between carrier proteins and enzymes from previously incompatible pathways. We show how iterating computational docking and mutagenesis with experimental feedback can be leveraged to accelerate PPI engineering and demonstrate production of a functional unnatural natural product through a chimeric FAS-NRPS pathway construction.

The biosynthesis of pyoluteorin in *Pseudomonas fluorescens* is initiated by the adenylation domain, PltF, which attaches proline to the phosphopantetheine of the peptidyl carrier protein, *holo*-PltL via an ATP dependent mechanism. (**Figure 1A**).[9] We sought to design a PltF mutant which could attach a proline to *holo*-AcpP from *Escherischia coli* fatty acid biosynthesis to produce hybrid amino-acyl products (**Figure 1B**).

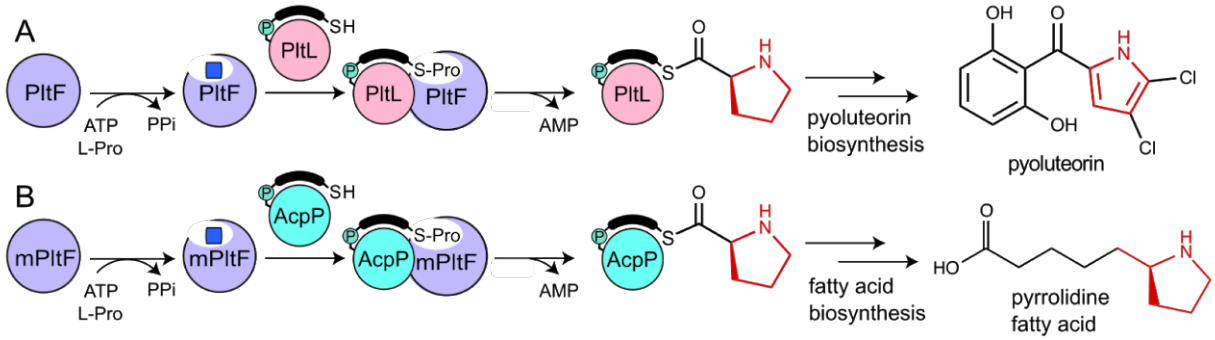


Figure 4.1 Adenylation and thiolation reactions by PltF. (A) Wild type reaction of PltF. PltF adenylates proline to create a prolyl-AMP intermediate (blue square), then transfers the prolyl moiety to *holo*-PltL. The pyrrolidine group is passed off and incorporated into pyoluteorin. (B) Designed PltF reaction. Mutant PltF (mPltF) instead transfers ATP-activated proline to *holo*-AcpP. *prolyl*-AcpP then shuttles the pyrrolidine group through the *E. coli* FAS for carbon chain elongation of the product. Depicted is a pyrrolidine fatty acid, a potential product of this hybrid metabolic pathway.

Results

4.2 Iterative computational and experimental design of PltF-AcpP interface

Increasing evidence has demonstrated the role of protein-protein interactions (PPIs) for controlling substrate specificity in carrier protein dependent pathways.[1,8] We therefore sought to design a new PPI between the non-cognate pairs, PltF and AcpP to facilitate a new activity for PltF, attaching proline to *holo*-AcpP. We leveraged information from the recently characterized, cognate PltF-PltL interface resolved by X-ray crystallography (PDB ID 6O6E).[10] AcpP shares 27% sequence identify with PltL, and aligns with 1.8 Å RMSD (**Figure S1**). To avoid deleterious pathway effects, we chose to focus mutations to PltF rather than AcpP.

The RosettaScripts interface[11] of the Rosetta Software Suite was used to determine initial PltF mutations to form a more favorable PltF-AcpP interface, using a protocol which iterates through 1,000 cycles of docking, design, minimization, and scoring with the REF15 score function[12] (**Figure 2**). Only PltF residues within 8 Å of AcpP were re-designed, and residues K402 and K486 were prevented from re-design due to importance for PltF catalysis. An additional bonus score was applied to favor native residues to avoid excess mutations. Six mutations significantly persisted throughout all cycles and were assayed for activity as point mutations in PltF: L235A,

S253A, M257Q, N436Y, Q438R, and K457I. Two mutants showed increased activity compared to the wild-type baseline, with N436Y prolylating 8% of the AcpP, and Q438R prolylating 38% of the AcpP (**Figure 3**).

Encouraged by these initial results, we tested whether combining these initial mutants would produce an additive or synergistic effect. Double and triple mutants did not show increased prolylation activity (**Figure S2**). Next, we tested whether subjecting single mutants through a subsequent round of *in silico* mutations would result in newer mutants with increased activity. Surprisingly, the majority of new mutants identified by Rosetta were unique compared to the first round of mutagenesis (**Table S1**). All of the second round of mutants identified for PltF N436Y showed decreased activity compared to the 8% single mutant (**Figure 3, Table S1**). Three of the second round mutants for Q438R showed similar or increased activity compared to the 38% single mutant, with the Q438R K472R and Q438R D263H double mutants prolylating > 60% AcpP (**Figure 3, Table S1**). A third round with each of these double mutants identified the triple mutant Q438R K472R D263N which prolylated 83% AcpP.

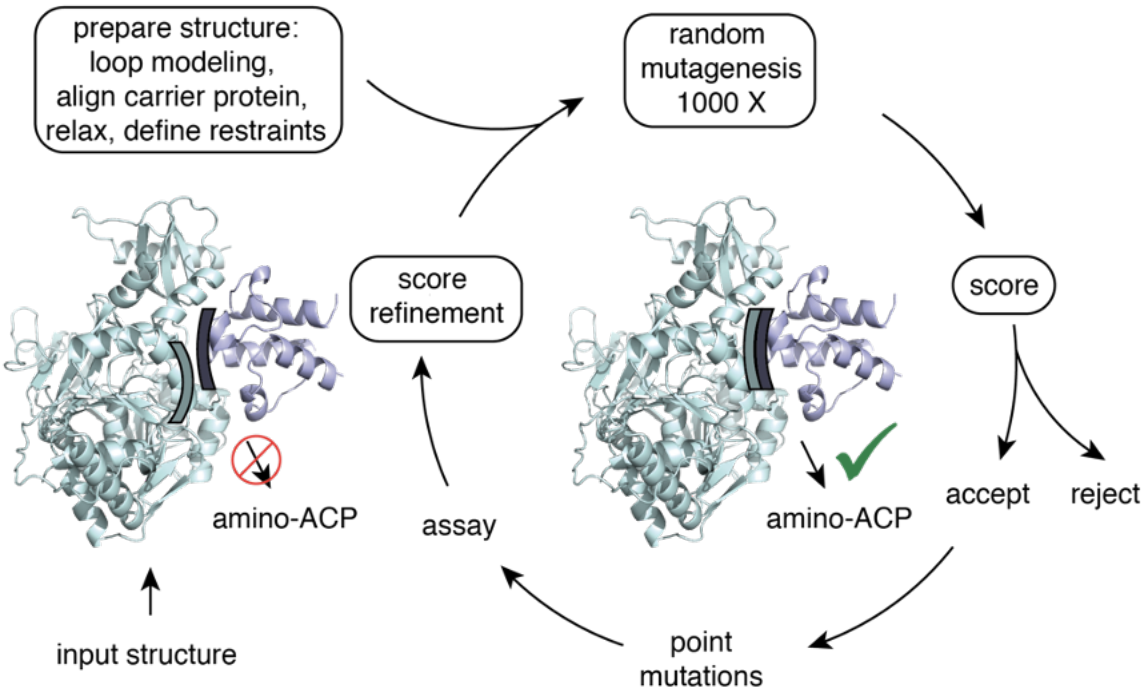


Figure 4.2. Schematic of computational and experimental interface design workflow. An initial interface is generated through alignment, and the structures are prepared by modeling missing loops, and performing relaxation. Residue or design restraints are selected followed by random mutagenesis of the interface to produce 1000 mutants which pass the score threshold. Mutants which significantly persist throughout the mutagenesis were selected for *in vitro* assay. Information from the experimental assay is then used to refine the Rosetta score function.

Application of the *in silico* Rosetta PPI design protocol in an iterative fashion identified a triple mutant of PltF with significantly increased activity compared to the wild type. However, testing each of the designs *in vitro* is a laborious process, and many of the mutants did not show any activity. Furthermore, the sequence profile identified by Rosetta was not correlated with mutant activity. Improvement of computational score functions to predict PPIs *in vitro* remains an ongoing challenge.[12,13] Therefore, we sought to refine the Rosetta score function, such that the PPI design protocol would be more predictive of desired activity. Examination of the individual score terms (**Data S1**) revealed that the electrostatic score was more predictive of activity order than the interface score. AcpP is a negatively charged protein, and many native partners have been found

to contain a basic binding patch, so electrostatic interactions were anticipated to play an important role.[14]

4.3 Refinement of Rosetta score function for improved PPI prediction

Several electrostatic coefficient weights were tested with the design protocol for the wild type PltF-AcpP interface. The standard weight for the electrostatic term is 0.875, but we found that increasing the weight to 1.500 gave the best prediction of *in vitro* activity (**Table S2**). We next repeated the computational protocol for three rounds, and found a remarkable increase in successful mutant predictability (**Figure 3**). Many fewer false positives (predicted mutants with no *in vitro* activity) and greater true positive results were generated. By the third round, all designs prolylated > 30%, five prolylated > 70% and the best design, Q438R D263K A230R, prolylated 92% of the *holo*-AcpP (**Figure 3**).

To better quantify and compare the mutant activities, the turnover numbers of select mutants were determined. The best triple mutant, Q438R D263K A230R, is able to convert 49 AcpP hr^{-1} (**Table 1**), compared to ~2 AcpP hr^{-1} for the best double mutants, Q438R D263K and Q438R K472R, and 0.7 AcpP hr^{-1} for the single Q438R mutant. Over three rounds of mutagenesis from the wild-type PltF to the Default Rosetta Score triple mutant, Q438R K472R D263N, the activity was improved 182-fold. Remarkably, the Refined Rosetta Score triple mutant, Q438R D263K A230R, exhibited an improved activity by 1690-fold relative to the wild-type enzyme. This improvement in activity, however, is still less than the cognate PltF-PltL reaction, where PltF can process about 4,000 PltL hr^{-1}.[15]

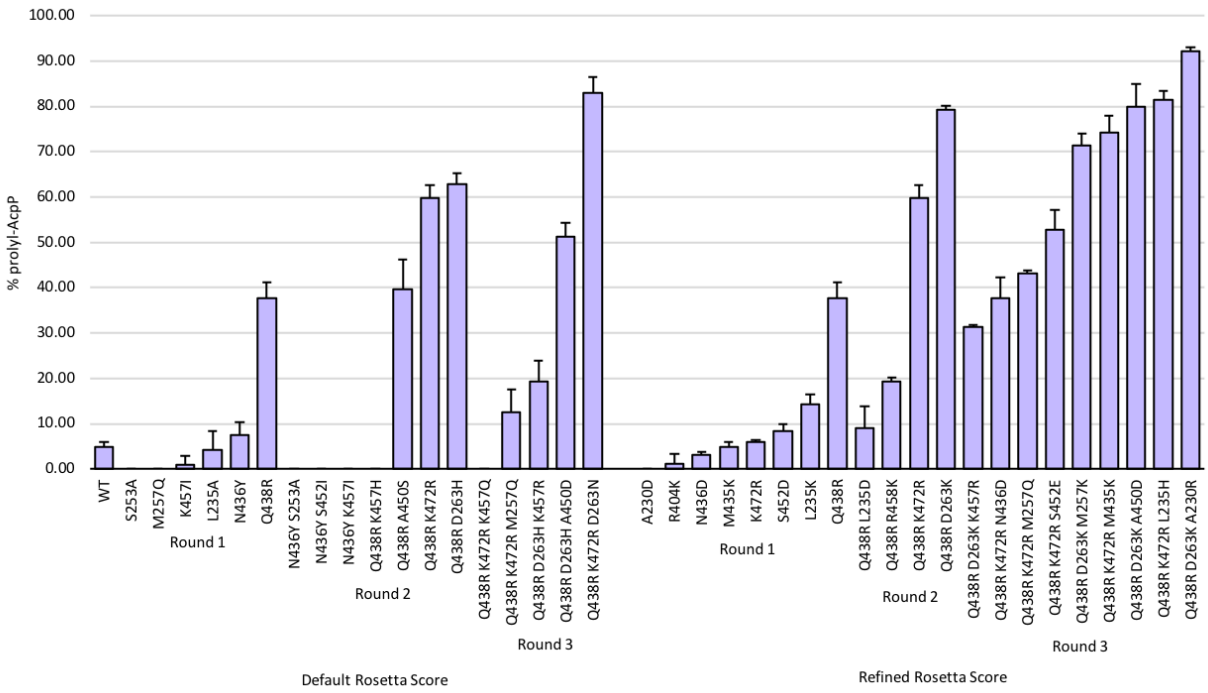


Figure 4.3 Prolylation activity of designed PltF mutants. Mutant PltF activity was monitored through HPLC chromatogram peak integration after incubation of 0.010 uM mPltF with 0.015 uM holo-AcpP, 5 mM ATP, 12.5 mM $MgCl_2$, and 5mM L-proline for 2 hours at 25C. Mutants were identified through iterative rounds of Rosetta with either the default or refined score function.

4.4 Structure of designed Pltf-AcpP interface

The cognate PltF-PltL interface was found to be dependent on hydrophobic interactions, where the X-ray crystal structure revealed only had one buried hydrogen bond interaction at the interface.[10] Thus, the initial superposition of AcpP onto PltL revealed a single electrostatic interaction between AcpP and PltF (**Figure 4A**). Despite a lack of initial interactions, three rounds with the refined Rosetta score introduced three positively charged residues while replacing a negatively charged residue, which created an area of positive electrostatic potential at the entrance to the PPant tunnel (**Figure 4B**). The refined Rosetta score was successful in designing an electrostatically-dependent interface from a hydrophobic interface to complement the negatively charged interface of AcpP.

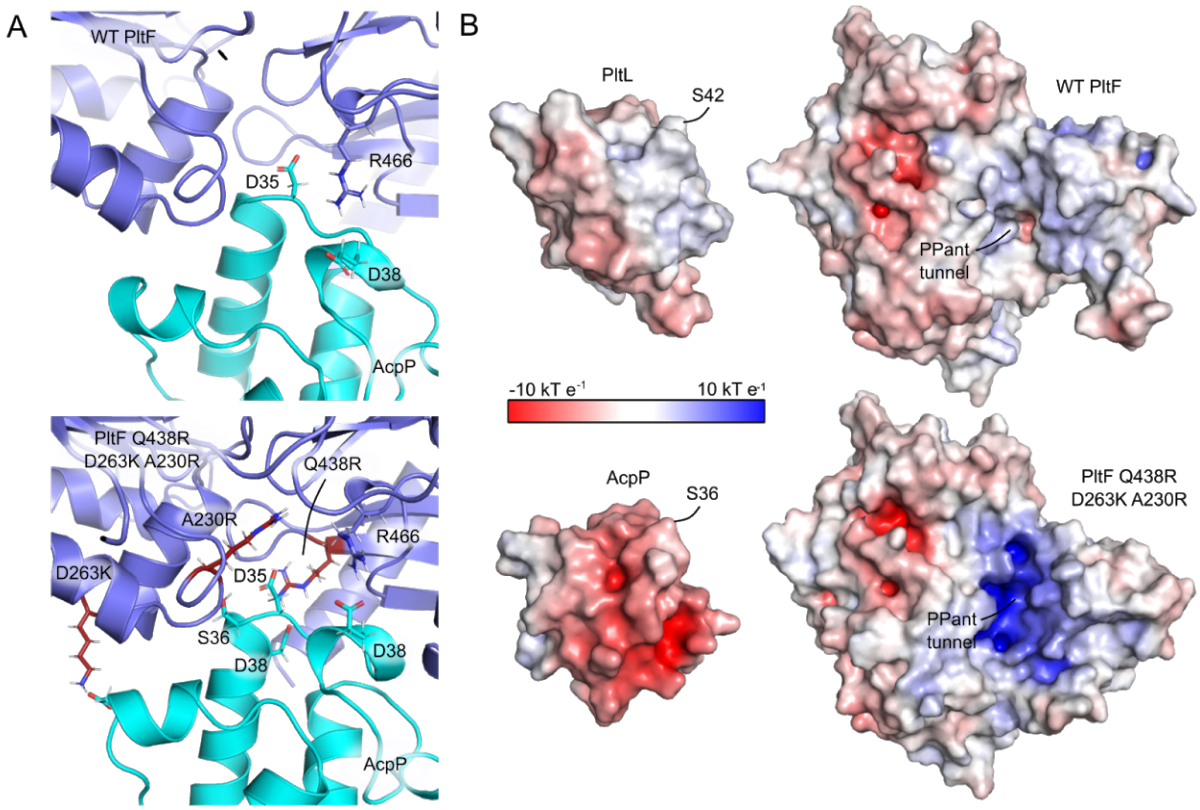


Figure 4.4 Comparison of wild-type and designed protein-protein interfaces. (A) Top: Relaxed model of wild-type PltF docked to AcpP. Bottom: Rosetta model of PltF Q438R D263K A230R bound to AcpP. (B) Electrostatic potentials mapped onto the surface of PltL (PDB ID 2N5H), PltF (PDB ID 6O6E), AcpP (PDB ID 2FAD), and a Rosetta model of PltF Q438R D263K A230R.

4.5 Conclusion

The ability to re-engineer carrier protein-dependent pathways to create custom products has been a long sought-after goal. Here we develop a computational and experimental workflow which incorporates both structure and function data to design mutations more efficiently than the current state of the art Rosetta ref2015 score function.[12] We engineered a hybrid FAS/NRPS pathway through interface mutations of the PltF adenylation domain from NRPS to interact with the AcpP acyl carrier protein from FAS. While the default Rosetta protocol generated 3 mutants with > 50% activity and 7 false positives, refining the score function based on the experimental assay resulted in 8 mutants with > 50% activity and only 1 false positive. Additionally, the most active mutant identified with the refined score function was almost 10 times faster than the mutant

identified with the default score function. This method advances the current state of interface engineering and provides a promising strategy for harnessing control over carrier protein-dependent pathways for designing new natural products.

Acknowledgements:

Chapter 4, in full, is currently being prepared for submission for publication as: Sztain, T, Corpuz, J.C., Bartholow, T.G., McCammon, J.A., Burkart, M.D. Metabolic Pathway Engineering via Protein-Protein Interface Design. The dissertation author is the primary co-author of this manuscript, along with Joshua Corpuz.

4.6 References

(1) Sztain, T.; Bartholow, T. G.; Lee, D. J.; Casalino, L.; Mitchell, A.; Young, M. A.; Wang, J.; McCammon, J. A.; Burkart, M. D. Decoding Allosteric Regulation by the Acyl Carrier Protein. *PNAS* **2021**, *118* (16). https://doi.org/10.1073/pnas.2025597118.

(2) Bartholow, T. G.; Sztain, T.; Patel, A.; Lee, D. J.; Young, M. A.; Abagyan, R.; Burkart, M. D. Elucidation of Transient Protein-Protein Interactions within Carrier Protein-Dependent Biosynthesis. *Communications Biology* **2021**, *4* (1), 1–10. https://doi.org/10.1038/s42003-021-01838-3.

(3) Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R. L.; 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), 427–431. https://doi.org/10.1038/nature12810.

(4) Milligan, J. C.; Lee, D. J.; Jackson, D. R.; Schaub, A. J.; Beld, J.; Barajas, J. F.; Hale, J. J.; Luo, R.; Burkart, M. D.; Tsai, S.-C. Molecular Basis for Interactions between an Acyl Carrier Protein and a Ketosynthase. *Nature Chemical Biology* **2019**, *15* (7), 669. https://doi.org/10.1038/s41589-019-0301-y.

(5) Dodge, G. J.; Patel, A.; Jaremko, K. L.; McCammon, J. A.; Smith, J. L.; Burkart, M. D. Structural and Dynamical Rationale for Fatty Acid Unsaturation in Escherichia Coli. *PNAS* **2019**, *116* (14), 6775–6783. https://doi.org/10.1073/pnas.1818686116.

(6) Mindrebo, J. T.; Patel, A.; Kim, W. E.; Davis, T. D.; Chen, A.; Bartholow, T. G.; La Clair, J. J.; McCammon, J. A.; Noel, J. P.; Burkart, M. D. Gating Mechanism of Elongating β-Ketoacyl-ACP Synthases. *Nat Commun* **2020**, *11*. https://doi.org/10.1038/s41467-020-15455-x.

(7) Lou, J. W.; Iyer, K. R.; Hasan, S. M. N.; Cowen, L. E.; Mazhab-Jafari, M. T. Electron Cryomicroscopy Observation of Acyl Carrier Protein Translocation in Type I Fungal Fatty Acid Synthase. *Sci Rep* **2019**, *9* (1), 12987. https://doi.org/10.1038/s41598-019-49261-3.

(8) Sarria, S.; Barthlow, T. G.; Verga, A.; Burkart, M. D.; Peralta-Yahya, P. Matching Protein Interfaces for Improved Medium-Chain Fatty Acid Production. *ACS Synth Biol* **2018**, *7* (5), 1179–1187. https://doi.org/10.1021/acssynbio.7b00334.

(9) Thomas, M. G.; Burkart, M. D.; Walsh, C. T. Conversion of L-Proline to Pyrrolyl-2-Carboxyl-S-PCP during Undecylprodigiosin and Pyoluteorin Biosynthesis. *Chemistry & Biology* **2002**, *9* (2), 171–184. https://doi.org/10.1016/S1074-5521(02)00100-X.

(10) Corpuz, J. C.; Podust, L. M.; Davis, T. D.; Jaremko, M. J.; Burkart, M. D. Dynamic Visualization of Type II Peptidyl Carrier Protein Recognition in Pyoluteorin Biosynthesis. *RSC Chem. Biol.* **2020**, *1* (1), 8–12. https://doi.org/10.1039/C9CB00015A.

(11) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E.-M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; Meiler, J.; Baker, D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLOS ONE* **2011**, *6* (6), e20161. https://doi.org/10.1371/journal.pone.0020161.

(12) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **2017**, *13* (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

Chapter 5. Mechanism of acyl carrier protein chain flipping

Terra Sztain, J. Andrew McCammon, Michael D. Burkart

5.1 Introduction

Biosynthetic pathways that produce fatty acids, polyketides, and non-ribosomal peptides depend on carrier proteins to shuttle intermediates between enzymes to produce primary and secondary metabolites. Acyl carrier proteins (ACPs) in particular, carry hydrophobic intermediates, which are sequestered inside a pocket to avoid water and prevent hydrolysis.[1] Substrate delivery requires a large conformational change, whereby the substrate translocates from inside the ACP pocket to inside the enzyme substrate pocket in a process termed "chain-flipping" (**Figure 1**).[2] X-ray crystallography has revealed the structures of the substrate sequestered ACP, and mechanism based crosslinking probes[3] have been developed to capture the post-chain flipped state.[4–6] The dynamic process has been monitored via fluorescence by appending a solvatochromic probe to ACP,[7,8] and via NMR by selective isotope labeling.[9] These dynamic experiments report whether a species is in the sequestered or chain-flipped conformation, but do not provide details of the mechanism. Molecular dynamics simulations performed by Colizzi et al[10] proposed a chain-flipping mechanism at the atomic-level, however these simulations applied an external biasing force to drive the catalytic atoms towards each other.

Therefore, we used the weighted ensemble (WE) enhanced sampling technique, which does not add any bias to the dynamics and maintains both thermodynamic and kinetic properties of the system, to elucidate the atomic mechanism of chain flipping from AcpP to FabA. The WE method relies on running many short simulations in parallel, replicating those which explore new areas of conformational space, and merging those which redundantly sample already visited regions.[11,12] The conformational space is divided into bins along a progress coordinate, and the probability of

simulations crossing bins is rigorously tracked, allowing accurate calculation of rate constants.[13,14] Using this method, we were able to generate continuous, unbiased pathways of chain flipping 3-(R)-hydroxydecanoic acid from AcpP to FabA.

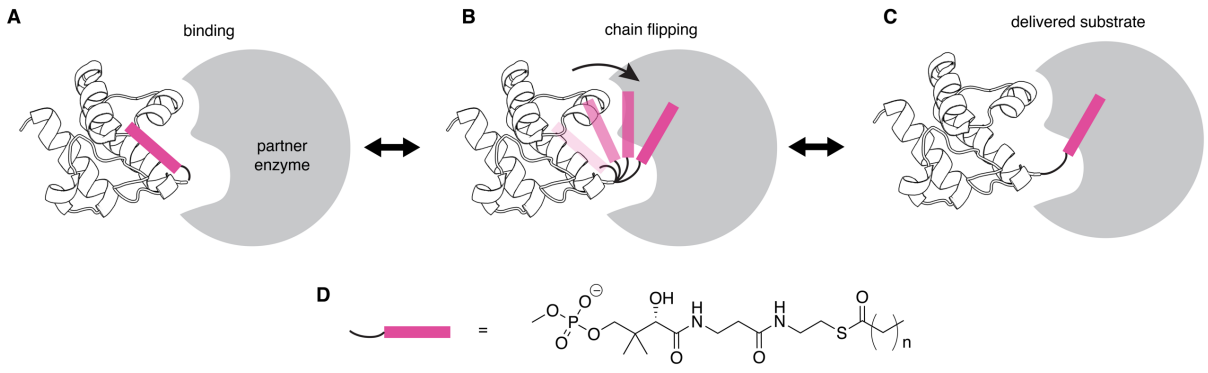


Figure 5.1 Schematic of ACP chain flipping. (A) Substrate sequestered ACP binds to a partner enzyme. (B) Hypothetical chain flipping of the substrate. (C) ACP bound to partner enzyme, with substrate delivered to active site. (D) Phosphopantetheine cofactor of ACP covalently attached to acyl chain via thioester bond.

5.2 Results

The first mechanism based crosslinked structure to trap ACP with a partner enzyme was AcpP from *Escherichia coli* with the enzyme FabA.[4] This provided an accurate starting point for our simulations without having to rely on docking (**Figure 2**). A two dimensional progress coordinate was chosen to balance focusing sampling to a region of conformational space, while still sampling pathways with multiple degrees of freedom. FabA catalyzes the dehydration of 3-(R)-hydroxydecanoyl-AcpP initiated when the Nτ of H70 acts as a catalytic base and abstracts a proton from the Cα of the carboxylic acid group (hereby referred to as Hα).[15] Therefore the distance between these atoms was used as one progress coordinate. Additionally, since both the initial and final structure are known, the RMSD of the ligand including the phosphopantetheine and (R)-hydroxydecanoyl group from the final structure was used as a second coordinate.

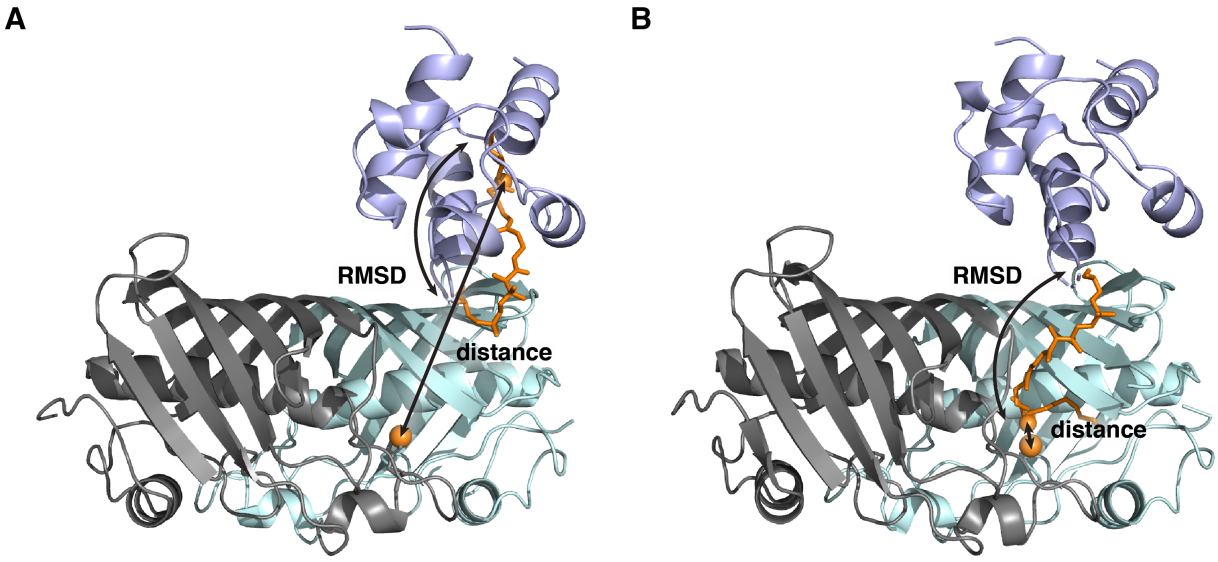


Figure 5.2 Chain flipping from AcpP to FabA. (A) Sequestered and (B) Chain flipped state of the phosphopantetheine tethered FabA substrate 3-(R)-hydroxydecanoic acid (orange) modeled from PDB: 4KEH[4] and 2FAD.[1] Catalytic atoms, Nτ of H70 and Hα of the carboxylic acid group of substrate, are shown as orange spheres. The two-dimensional progress coordinate is labeled and indicated with black arrows.

After 50 iterations, a nearly complete chain flipping pathway was observed, which will be analyzed here. The distance between catalytic residues decreased from a maximum of 35 Å to 10 Å, and the RMSD from the final structure decreased from 22 Å to 8 Å (**Figure 3A**). Structures along a continuous pathway revealed an unexpected mechanism. The acyl groups sequestered by ACPs are quite hydrophobic. Therefore, previous hypotheses described a mechanism in which the acyl chain is continuously shielded from solvent, and is pushed through helices of each protein during delivery.[10] Our simulations revealed the protein helices remaining well structured, while the acyl chain slowly hairpins out of the hydrophobic pocket, then rapidly flips once the terminal methyl group escapes (**Figure 3B-D**). The reverse of this mechanism was also observed by simulations of acyl-AcpP carried out by Chan et al. initiated with solvent exposed acyl chains.[16] A possible explanation for this discrepancy is that unlike the docked structure of *Plasmodium falciparum* ACP to FabZ, where ACP helix II and III lie against the partner protein, the structures

of AcpP bound to *E. coli* partner enzymes show the AcpP in a "hand-stand" position, balancing on S36, with helix II and III more solvent exposed.[4–6,17,18]

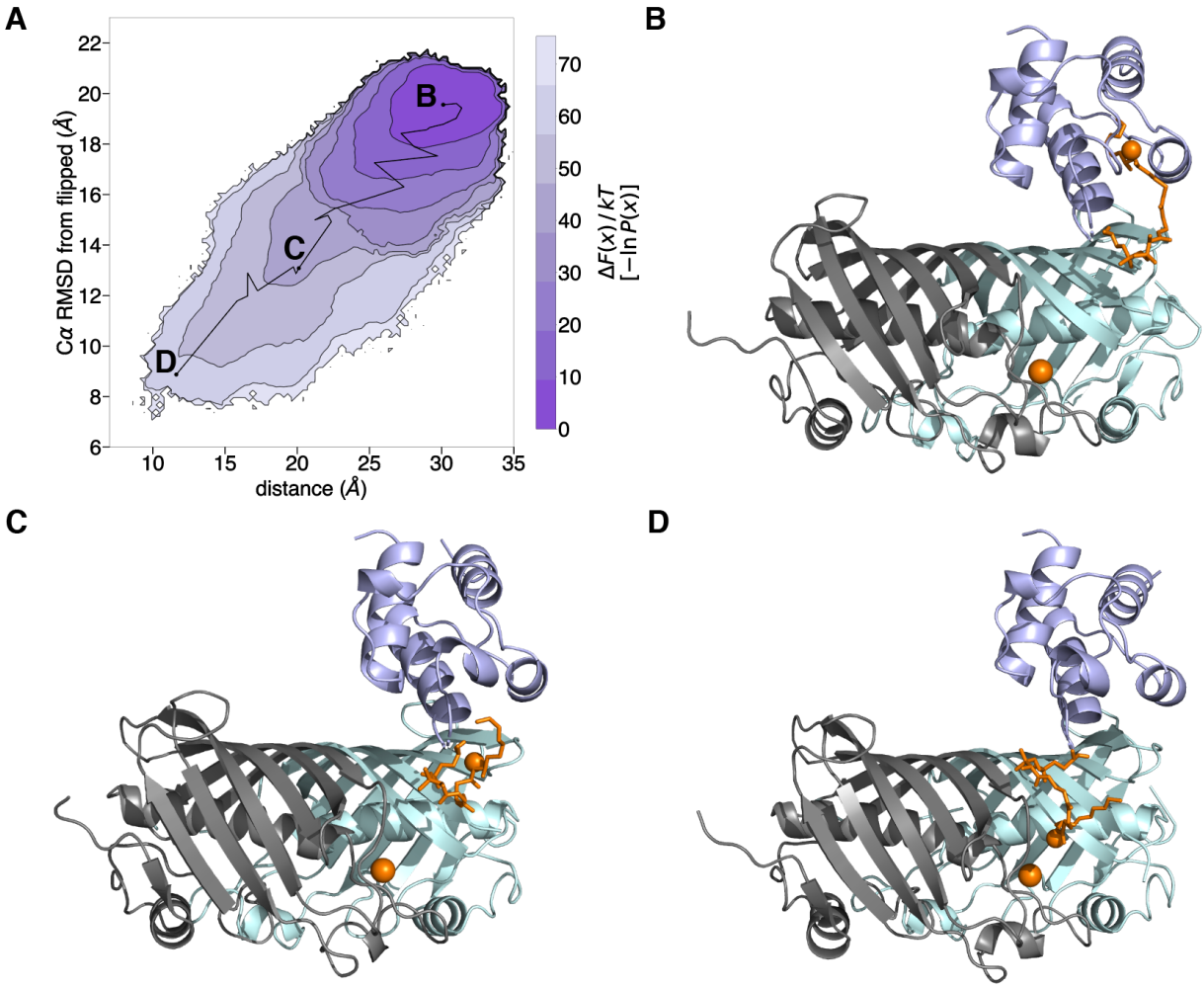


Figure 5.3 Chain flipping pathway. (A) Probability distribution of progress coordinates after 50 WE iterations with an independent pathway shown as a black line. (B-D) Structures along the chain flipping pathway with the phosphopantetheine tethered FabA substrate 3-(R)-hydroxydecanoic acid colored orange, as well as the catalytic atoms, Nτ of H70 and Hα of the carboxylic acid group of substrate, are shown as spheres.

To further investigate the plausibility of this mechanism, we analyzed the residues surrounding the acyl chain during chain flipping. There are indeed several hydrophobic residues on the surface of FabA interacting with the saturated portion of the acyl chain, and polar residues

interacting with the phosphopantetheine (**Figure 4**). Further investigation of interactions, solvent accessibility, kinetics, and experimental validation are currently underway.

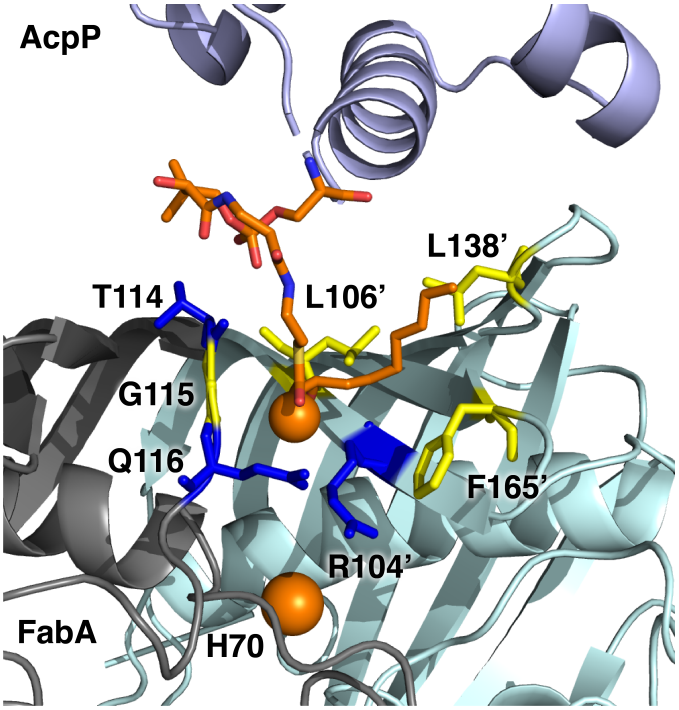


Figure 5.4 Acyl chain interactions during chain flipping. Carbon atoms of phosphopantetheine tethered FabA substrate 3-(R)-hydroxydecanoic acid are colored orange, nitrogen atoms in blue, oxygen atoms in red, phosphorous atom in orange, and sulfur atom in yellow. Catalytic atoms Nτ of H70 and Hα of the carboxylic acid group of substrate, are shown as orange spheres. FabA residues within 3.5 Å of the phosphopantetheine tethered substrate are shown as sticks and colored blue if polar and yellow if non-polar.

Acknowledgements

Chapter 5, in full, is currently being prepared for submission for publication as Sztain, T, McCammon, J.A., Burkart, M.D. Mechanism of acyl carrier protein chain flipping. The dissertation author is the primary author of this manuscript.

5.4 References

(1) 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), 135–145. https://doi.org/10.1016/j.jmb.2006.09.049.

(2) Cronan, J. E. The Chain-Flipping Mechanism of ACP (Acyl Carrier Protein)-Dependent Enzymes Appears Universal. *Biochem. J.* **2014**, *460* (2), 157–163. https://doi.org/10.1042/BJ20140239.

(3) Worthington, A. S.; Burkart, M. D. One-Pot Chemo-Enzymatic Synthesis of Reporter-Modified Proteins. *Org. Biomol. Chem.* **2006**, *4* (1), 44–46. https://doi.org/10.1039/b512735a.

(4) Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R. L.; 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), 427–431. https://doi.org/10.1038/nature12810.

(5) Milligan, J. C.; Lee, D. J.; Jackson, D. R.; Schaub, A. J.; Beld, J.; Barajas, J. F.; Hale, J. J.; Luo, R.; Burkart, M. D.; Tsai, S.-C. Molecular Basis for Interactions between an Acyl Carrier Protein and a Ketosynthase. *Nat. Chem. Biol.* **2019**, *15* (7), 669–671. https://doi.org/10.1038/s41589-019-0301-y.

(6) Dodge, G. J.; Patel, A.; Jaremko, K. L.; McCammon, J. A.; Smith, J. L.; Burkart, M. D. Structural and Dynamical Rationale for Fatty Acid Unsaturation in Escherichia Coli. *PNAS* **2019**, *116* (14), 6775–6783. https://doi.org/10.1073/pnas.1818686116.

(7) Beld, J.; Cang, H.; Burkart, M. D. Visualizing the Chain-Flipping Mechanism in Fatty-Acid Biosynthesis. *Angew. Chem. Int. Ed. Engl.* **2014**, *53* (52), 14456–14461. https://doi.org/10.1002/anie.201408576.

(8) Charov, K.; Burkart, M. D. A Single Tool to Monitor Multiple Protein–Protein Interactions of the Escherichia Coli Acyl Carrier Protein. *ACS Infect. Dis.* **2019**, *5* (9), 1518–1523. https://doi.org/10.1021/acsinfecdis.9b00150.

(9) Sztain, T.; Bartholow, T. G.; Lee, D. J.; Casalino, L.; Mitchell, A.; Young, M. A.; Wang, J.; McCammon, J. A.; Burkart, M. D. Decoding Allosteric Regulation by the Acyl Carrier Protein. *PNAS* **2021**, *118* (16). https://doi.org/10.1073/pnas.2025597118.

(10) Colizzi, F.; Masetti, M.; Recanatini, M.; Cavalli, A. Atomic-Level Characterization of the Chain-Flipping Mechanism in Fatty-Acids Biosynthesis. *J. Phys. Chem. Lett.* **2016**, *7* (15), 2899–2904. https://doi.org/10.1021/acs.jpclett.6b01230.

(11) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys J* **1996**, *70* (1), 97–110.

(12) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu Rev Biophys* **2017**, *46*, 43–57. https://doi.org/10.1146/annurev-biophys-070816-033834.

(13) Zwier, M. C.; Pratt, A. J.; Adelman, J. L.; Kaus, J. W.; Zuckerman, D. M.; Chong, L. T. Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein–Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered P53 Peptide. *J Phys Chem Lett* **2016**, *7* (17), 3440–3445. https://doi.org/10.1021/acs.jpclett.6b01502.

(14) Saglam, A. S.; Chong, L. T. Protein-Protein Binding Pathways and Calculations of Rate Constants Using Fully-Continuous, Explicit-Solvent Simulations. *Chemical Science* **2019**, *10* (8), 2360–2372. https://doi.org/10.1039/C8SC04811H.

(15) Moynié, L.; Leckie, S. M.; McMahon, S. A.; Duthie, F. G.; Koehnke, A.; Taylor, J. W.; Alphey, M. S.; Brenk, R.; Smith, A. D.; Naismith, J. H. Structural Insights into the Mechanism and Inhibition of the β-Hydroxydecanoyl-Acyl Carrier Protein Dehydratase from Pseudomonas Aeruginosa. *Journal of Molecular Biology* **2013**, *425* (2), 365–377. https://doi.org/10.1016/j.jmb.2012.11.017.

(16) 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), 33620–33629. https://doi.org/10.1074/jbc.M805323200.

(17) Masoudi, A.; Raetz, C. R. H.; Zhou, P.; Pemble, C. W. Chasing Acyl-Carrier-Protein Through a Catalytic Cycle of Lipid A Production. *Nature* **2014**, *505* (7483), 422–426. https://doi.org/10.1038/nature12679.

(18) Mindrebo, J. T.; Patel, A.; Kim, W. E.; Davis, T. D.; Chen, A.; Bartholow, T. G.; La Clair, J. J.; McCammon, J. A.; Noel, J. P.; Burkart, M. D. Gating Mechanism of Elongating β-Ketoacyl-ACP Synthases. *Nat Commun* **2020**, *11*. https://doi.org/10.1038/s41467-020-15455-x.

Chapter 6. Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease.

Terra Sztain *, Rommie Amaro[1], J. Andrew McCammon[1,2]

Abstract

The SARS-CoV-2 pandemic has rapidly spread across the globe, posing an urgent health concern. Many quests to computationally identify treatments against the virus rely on *in silico* small molecule docking to experimentally determined structures of viral proteins. One limit to these approaches is that protein dynamics are often unaccounted for, leading to overlooking transient, druggable conformational states. Using Gaussian accelerated molecular dynamics to enhance sampling of conformational space, we identified cryptic pockets within the SARS-CoV-2 main protease, including some within regions far from the active site. These simulations sampled comparable dynamics and pocket volumes to conventional brute force simulations carried out on two orders of magnitude greater timescales.

6.1 Introduction

In December 2019, the World Health Organization learned of a novel coronavirus which has since rapidly spread, leading to a global pandemic. The SARS-CoV-2 virus, causing the disease named COVID-19 has exceeded 100 million cases worldwide, taking over 2 million lives as of January 30, 2021.[1] The genetic and structural similarity of the SARS-CoV-2 virus to the agents of the severe acute respiratory syndrome coronavirus (SARS-CoV) epidemic in 2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) epidemic in 2012 provide a basis of information for understanding and ultimately treating or preventing this disease, however no highly effective antiviral drug exists for any human-infecting coronavirus.

Proteases are responsible for activating viral proteins for particle assembly and have proved successful targets for antiviral agents; most notable are the protease inhibitors used to treat HIV and Hepatitis C.[2,3] The main protease of SARS-CoV-2, called M^{pro} or $3CL^{pro}$ encoded by the nsp5 gene,[4] was the first SARS-CoV-2 protein deposited to the protein databank (PDB) on January 26th 2020.[5] This structure was crystalized with a covalent inhibitor (N3) identified from computer-aided drug design, and validated biochemically. Currently, over one hundred M^{pro} structures exist in the PDB and massive efforts to discover a successful inhibitor are underway.[5,6,7,8]

M^{pro} is highly similar to 3CL proteases from other coronaviruses in sequence, structure, and function.[9,10] First, M^{pro} autocatalytically cleaves itself from the SARS-CoV-2 polypeptide then forms a homodimer to subsequently cleave at 11 distinct sites, while the papain-like protease, PL^{pro} cleaves at 3 sites, allowing the viral proteins to fold and perform their functions.[11,12,13] Proteolytic cleavage is catalyzed by residues H41 and C145 of the catalytic cysteine dyad. These active site residues are located at the N-terminal globular domain of M^{pro}, which contains an anti-parallel beta barrel reminiscent of trypsin-like serine proteases, while the C-terminal domain consists of five alpha helices.[13] Upon dimerization, M^{pro} resembles the shape of a heart symbol (**Figure 1A**).

Many efforts to identify inhibitors involve high throughput virtual screening, which exploits the power of computational docking to screen millions of molecules *in silico* to narrow down a few "hits" for lead optimization. Incorporation of molecular dynamics (MD) has significantly improved the ability to identify promising protein inhibitors.[14] Docking to protein ensembles obtained from MD simulations is often employed to consider multiple target states that remain elusive in static crystal structures.[15] Conventional MD simulations are however limited in the amount of conformational space that can be sampled due to the amount of time required to

traverse energy barriers between stable conformational states.[16] In the present study, we used the enhanced sampling technique, Gaussian accelerated MD (GaMD),[17] to overcome such barriers. GaMD adds a harmonic boost to the potential energy up to a threshold, effectively "filling in the wells" creating a smoother potential energy surface.[17,18] This method allowed us to extensively sample conformations of the SARS-CoV-2 main protease at minimal computational expense and provide detailed characterization of several potentially druggable pockets which can serve as a basis for identifying an M^{pro} inhibitor for COVID-19 treatment.

6.2 Results

GaMD simulations were performed on the first published M^{pro} structure PDB 6LU7.[5] Six systems were simulated, including monomer and dimer simulation in the presence and absence of the co-crystalized N3 inhibitor either covalently or non-covalently bound (**Figure S1**). Five independent simulations of 200 ns, for an aggregate of 1 μs were carried out for each system. Three regions were defined to investigate potentially druggable pockets: the active site, which lies in the N-terminal top lobe of the heart shaped protease; the C-terminal region, as a potential allosteric site; and the dimer interface region (**Figure 1A**).

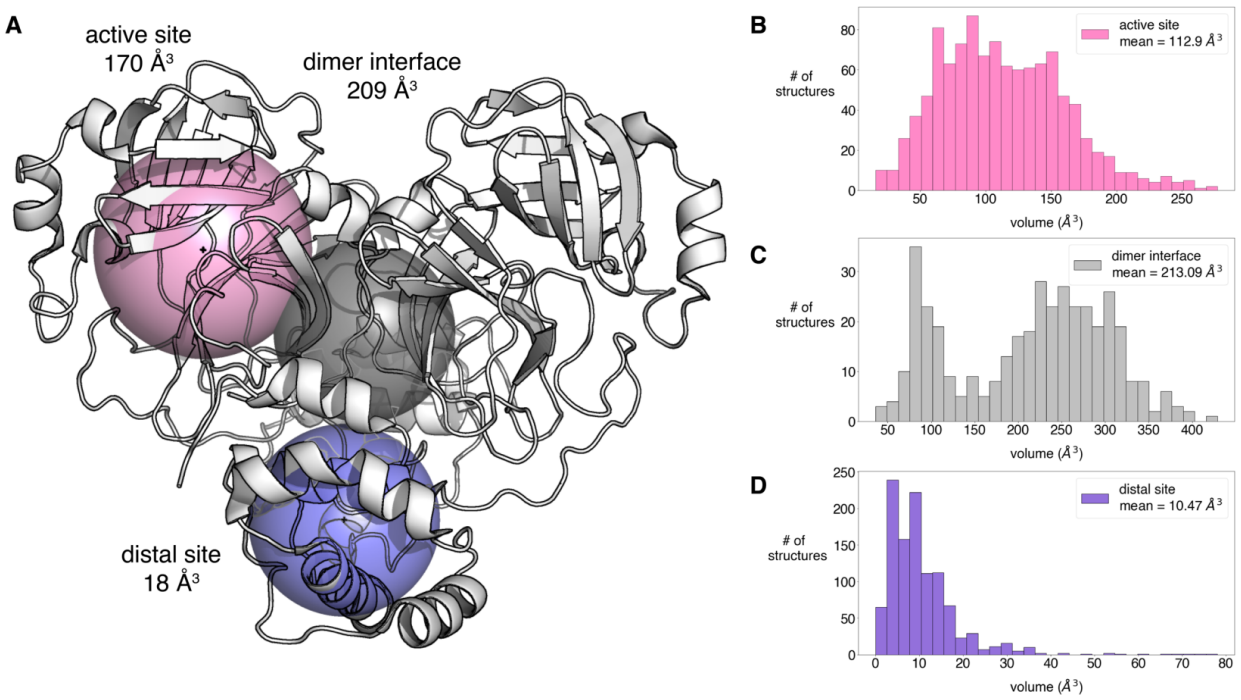


Figure 6.1 Pocket analysis of GaMD simulations of M^{pro}. (A) Active site, dimer interface, and distal site definitions for pocket calculations. The inclusion sphere was centered on the center of mass (COM) of the top and bottom protein domains, for active and distal site with 12 Å and 10 Å radii respectively. A 10 Å radii sphere was used for the dimer interface calculation. (B-D) Histograms showing the distribution of pocket volumes calculated from the aggregate 6 μs GaMD simulations.

6.3 Characterization of M^{pro} pockets

A variety of pocket volumes were sampled between 16 – 277 Å^3 for the active site pocket, starting from 170 Å^3 in the crystal structure. The interface spanned from 36 – 429 Å^3 starting from 209 Å^3, and the distal site from 0 – 78 Å^3 starting from 18 Å^3 (**Figure 1B-D**). Simulations with the N3 ligand present sampled greater mean volumes for the active site and distal region than the apo simulations without N3, however the average dimer interface region was greater in the apo simulation (**Figure S2**). Covalently bound N3 sampled greater mean volumes for the active site and dimer interface than non-covalent, which sampled greater mean volumes for the distal site (**Figure S3**). In a few simulations, the non-covalent ligand escaped the active site pocket (**Figure S4)**. In all simulations, the ligand re-arranges compared to the crystal structure. The non-covalent

ligands quickly decrease in total number of native contacts and increase in non-native contacts. The covalent ligands slightly decrease in native contacts but have a similar increase in non-native contacts. (**Figures S5-S6**). These simulations revealed significant loop dynamics (**Figure 2A-C**) leading to the distinct pocket conformations (**Figure 2D-F, Movie S1**) with varying active site water occupancy (**Figure S7**).

Several metrics were used to assess the properties of these pockets. The PockDrug[19] webserver predicts druggability probability based on a variety of factors including geometry, hydrophobicity, aromaticity, among others. The FTMap[20] computational solvent mapping webserver helps determine "hot spots" and can aid in fragment based drug design. For each site, the sampled pocket volumes were clustered into ten populations, and a frame corresponding to that volume was analyzed. The most populated volumes for the active site, dimer interface, and distal site were 114 $Å^3$, 220 $Å^3$, and 10 $Å^3$ respectively. The ten active site pockets gave predicted druggabilities between 0.18 - 0.86 and solvents mapped to seven of the pockets (**Figure S8**). The dimer interface pockets gave higher druggability prediction values, between 0.5 - 0.93, with solvent mapping to all ten structures (**Figure S9**). The distal site pockets gave the highest predicted druggability, between 0.37-1.00 with six out of ten pockets giving predicted druggability values greater then 0.90. Solvents mapped to seven out of the ten pockets (**Figure S10**). Notably, none of the solvent mapping to any of the sites showed overlapping results, indicating that each structure may provide distinct opportunities for designing inhibitors. The pocket volume, predicted druggability, and solvent mapping were not directly correlated, as expected, since each calculation considers different metrics. Overall, these analyses indicate that the simulations reveal a diverse set of conformations that are likely to serve as viable targets for in silico drug development.

Dimer association is necessary for catalytic activity of the protease,[5] therefore it is reasonable to assume binding of a molecule which prevents this association would inactivate the protease. The significance of the distal regions is less well understood, though computational investigations support a potential allosteric role.[21,22] To study this further, we examined the correlated motions of each residue throughout the simulations. In the monomer simulation, the active site region was positively correlated to the distal site region, and negatively correlated to the region that connects the two, whereas the inverse was observed in the dimer simulation (**Figure 3, Figures S11,12**). These correlations resulted in primary motions, detected by principal component analysis, comprising of hinge-like in the monomer versus twisting in the dimer (**Figure S13**). These motions were similar in both the apo and N3 bound simulations. In the dimer simulations, the active site of chain A was slightly correlated to the active site of chain B, and the distal regions were highly anticorrelated, twisting in opposite directions. (**Figure S13**). Regardless of positive or negative correlation values, the distal and active site show a dynamic interdependence on each other in both the apo and dimer forms.

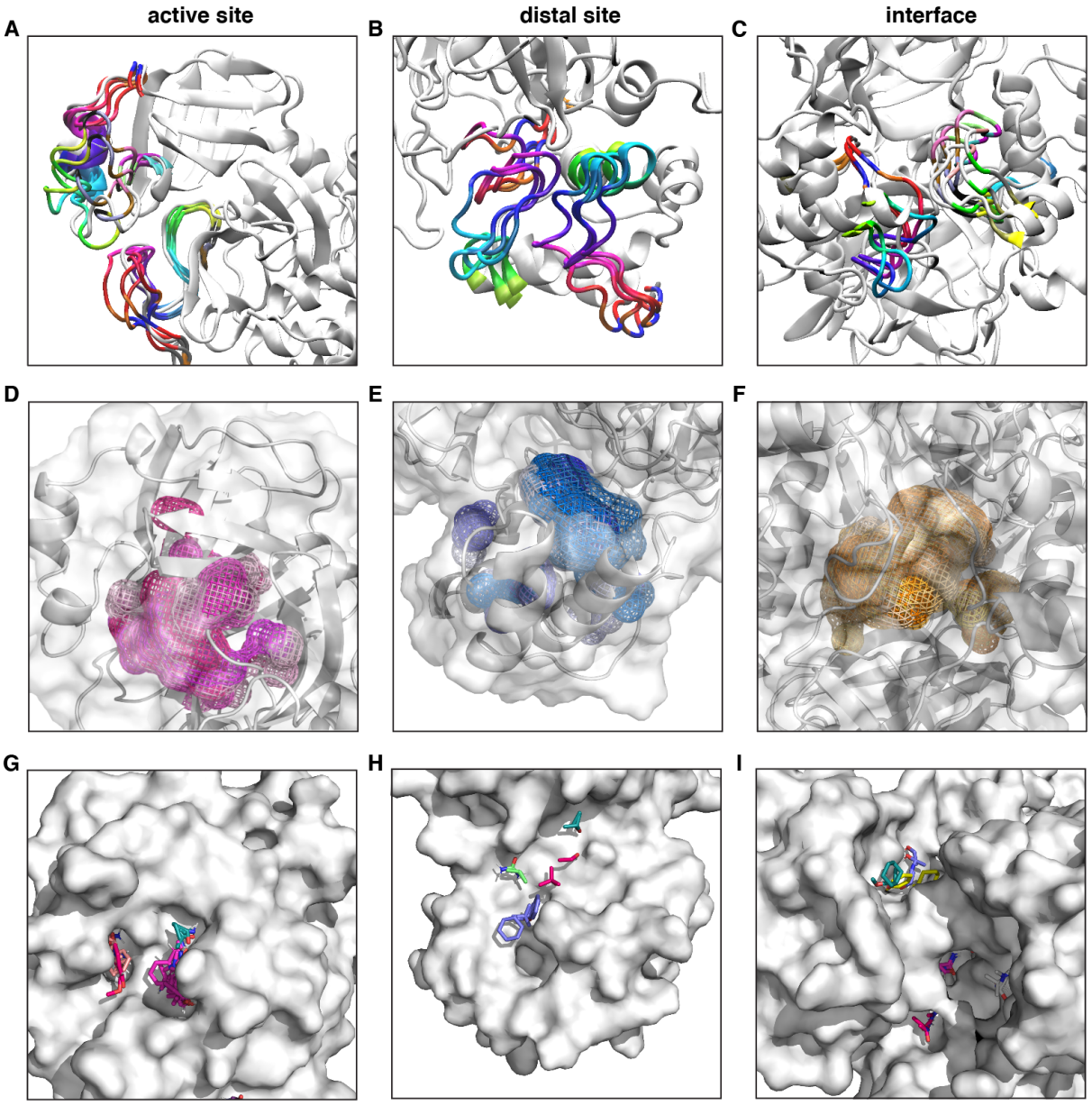


Figure 6.2 Cryptic pockets identified from GaMD simulations. (A-C) Overlay of several frames highlighting loop dynamics leading to various pocket conformations for each site. (D-F) Overlay of several pockets shown outlined in mesh, colored in varying shades of pink for the active site, blue for the distal site, and orange for the interface. (G-I) Computational solvent mapping of "hot spots" to pockets at each site.

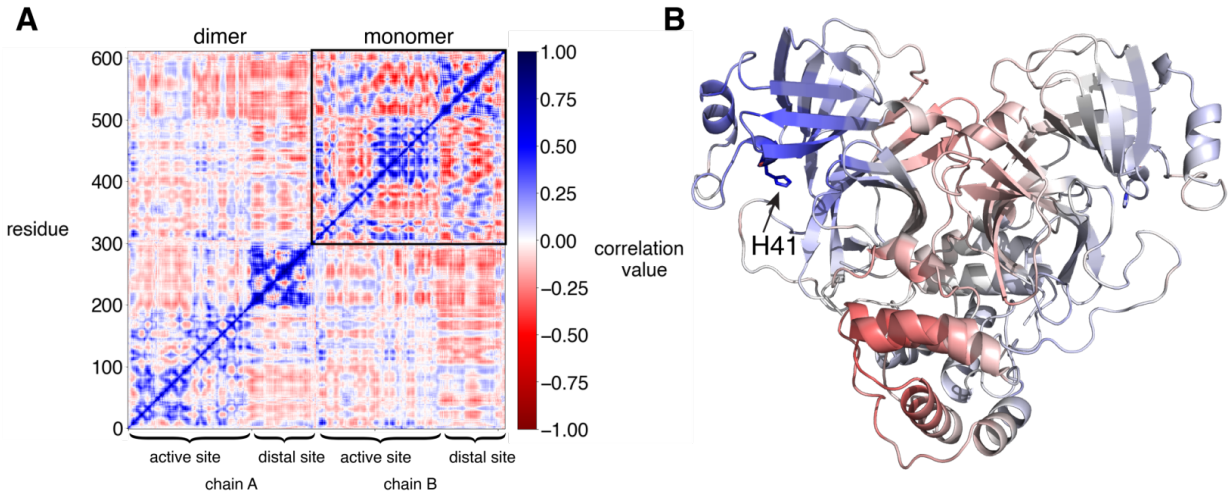


Figure 6.3 Correlated motions of each residue. (A) Correlation matrix calculated from aggregate 1 μs simulation of apo M^{pro}. Calculation from the monomer simulation is shown as an inset in the upper right corner of the calculation from the dimer simulation. (B) Structural representation of correlated motions, with each residue colored based on correlation to the catalytic histidine H41.

6.4 GaMD enhancement of M^{pro} conformational space

Previous benchmarks have demonstrated that GaMD can enhance sampling by multiple orders of magnitude.[23] To test the sampling enhancement compared to conventional MD simulations, we compared the 1 μs aggregate GaMD simulation from 5 replicates of 200 ns to the 10 μs simulations carried out at Riken[24] and 100 μs simulations D.E. Shaw Research (DESRES).[25] We additionally performed and compared those to 1 μs of GaMD continuous simulation and compared 200 ns cMD to 200 ns GaMD. The ranges of pocket volumes for each site were highly similar as detailed in **Figure S14**. Even the short conventional simulations sampled comparable pocket volumes, indicating this metric may not be the most effective measure of conformational sampling. Principal component analysis (PCA) is frequently used as a metric for conformational diversity as it decreases the dimensionality to maximize the dataset variance.[26,27] PCA analysis was indeed more consistent with expectations where increasing simulation time corresponded to larger regions of space covered by plotting the first two principal components (**Figure 4**).

While the GaMD simulations run continuously for 1 μs only covered 54% of the conformational space covered by the 100 μs DESRES, the 1 μs aggregate GaMD from 5 simulations run in parallel covered 80% of the DESRES space. The 10 μs Riken simulations covered 94% of that space (**Figure 4**). It is worth noting that each of the simulations with the most extensive sampling (DESRES, Riken, GaMD 5 x 200ns) reached areas in the PC1-PC2 subspace that neither of the other two reached. This just confirms that none of these simulations achieve comprehensive sampling which, in the ultimate case, would include extensively unfolded conformations. Each of these simulations reveals cryptic binding sites not seen in shorter, conventional MD simulations, let alone in the experimental structures, and some sites are seen in one but not the other simulations.

Though perhaps not intuitive at first glance, it is reasonable that many simulations in parallel would have a greater chance of sampling diverse space than one long simulation, which may settle into a favorable low energy conformation, even given the boost added by the GaMD, which is restricted by a threshold.[17] Representative structures from dense populations of the PCA plots were further analyzed structurally and through computational solvent mapping (**Figure S15**). Additionally, RMSD and RMSF comparison revealed the majority of the conformational deviations occur within the helices and loops lining the active site pocket (**Figure S16-S17, S21-S23**), which was also visualized in the representative structures from populations determined from PCA (**Figure S15**). Taken together, the range of pocket volumes, structural conformations, and dynamics were similar between short GaMD simulations and brute force conventional simulations. The plethora of available structures from all of these studies is expected to be useful for *in silico* inhibitor development for targeting M^{pro}.

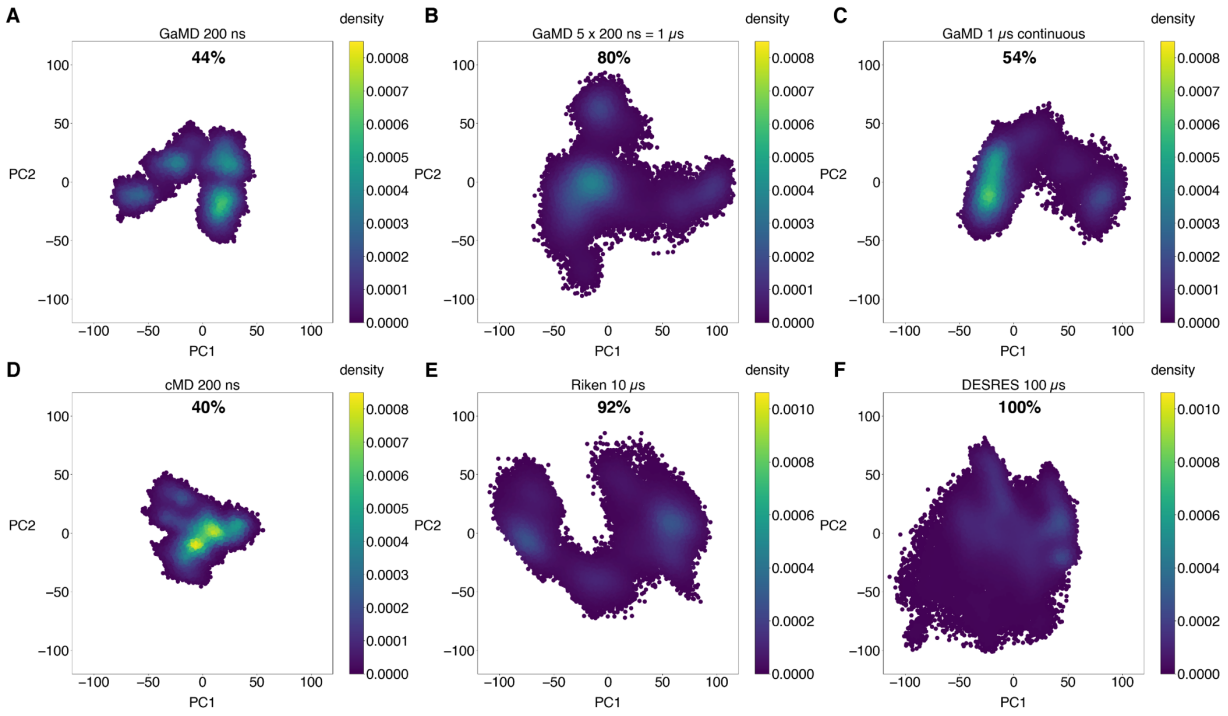


Figure 6.4 Conformational sampling of GaMD and conventional MD. Each plot shows the first and second principal components, projected from a mass weighted matrix of backbone atom positions of the apo dimer, colored by probability density. (A) 200 ns of GaMD simulation. (B) 1 μs of GaMD aggregated from 5 independent 200 μs simulation. (C) 1 μs of continuous GaMD simulation. (D) 200 ns of cMD. (E) 10 μs of cMD performed by Riken. (F) 100 μs of cMD performed by D.E. Shaw Research (DESRES). The area of a 2D histogram of the PC1 vs PC2 datapoints from each simulation is labeled as a percentage relative to the 100 μs DE. Shaw Research (DESRES) conventional MD simulation.

6.5 Conclusion

GaMD simulations of the SARS-CoV-2 main protease, M^{pro} revealed cryptic pockets not detectable from the crystal structure. In addition to characterizing the vast conformational landscape of the active site and dimer interface, a distal region was explored as a potential allosteric site. GaMD sampled conformational space more efficiently than brute force simulations, with 1 μs simulation data providing comparable results to 10 μs simulations carried out by Riken[24] and 100 μs simulations carried out by DE Shaw Research,[25] offering a more widely accessible simulation strategy. These structures, along with the millisecond simulations from Folding@home[28], can serve as a basis for *in silico* docking to identify M^{pro} inhibitors for COVID-19 treatment.

6.6 Methods

Simulation preparation

The first crystal structure of SARS-CoV-2 main protease, PDB ID 6LU7[5] was used as a starting point for all simulations. Protonation states of titratable residues were determined using the H++ webserver[29] with 0.15 M salinity, 10 internal dielectric constant, 80 external dielectric constant, and pH 7.4. Histidine 64, 80, and 164 were thus named HID to indicate delta nitrogen protonation before protonation of entire protein using tleap from AmberTools 18.[30] Six systems were simulated as shown in **Figure S1**. For apo simulations, the inhibitor was deleted from the PDB file. For dimer simulations, the PyMOL[31] symexp command was used to generate initial coordinates for the second chain based on symmetry. The ff14SB[32] forcefield was used for proteogenic residues. For simulations with the N3 bound, the antechamber package of AmberTools 18 was used to generate GAFF[33] forcefield parameters for N3. Gaussian 09 was used to calculate partial charges of atoms according to the RESP[34] method with HF/6-31G* level of theory. For covalent simulations, the cysteine 145 was treated as an unnatural amino acid and capped with N-methyl and C-acetyl groups for charge calculations. A TIP3P isometric water box was added with at least 10 Å buffer between solute and edges of the box. Enough Na+ was added to neutralize the system, then Na+ and Cl- ions were added to a final concentration of 150 mM. Minimization was carried out in two steps. First the solute was restrained with a 500 kcal / mol restraint force to minimize the solvent for 10,000 cycles, followed by unrestrained minimization for 300,000 cycles. Next the system was heated to 310 K over 350 ps using an isothermal-isovolumetric (NVT) ensemble, followed by isothermal-isobaric (NPT) equilibration for 1 ns.

Gaussian accelerated molecular dynamics

Gaussian accelerated molecular dynamics (GaMD)[17] pmemd.cuda implementation of Amber 18 was used to generate 5 independent trajectories of 200 ns, an aggregate of 1 μs for each system. The dual boost method was employed, adding a biasing force to both the total and dihedral potential energy. The threshold energy was set to the upper bound. GaMD production of equilibrated systems was carried out in 5 stages. First, 200 ps of preparatory conventional MD simulation were carried out, without any statistics collected. Second, 1 ns of conventional MD was carried out to collect potential statistics V_{max}, V_{min}, V_{avg}, and σV. Next, 800 ps of GaMD were carried out with a boost potential applied with fixed parameters. Then, 50 ns of GaMD were carried out with updated boost parameters and finally 150 ns of GaMD were carried out with fixed boost parameters. All production simulations were carried out using NVT, with periodic boundary conditions and 2 fs timesteps. The SHAKE[35] algorithm was used to restrain nonpolar hydrogen bonds and TIP3P water molecules. Particle Mesh Ewald[36] method was used for electrostatic interactions with a 10 Å cutoff for nonbonded interactions. Langevin thermostat was used for temperature regulation with a collision frequency of 5 ps^{-1}.

Pocket analysis

For each system, the aggregated simulation was clustered based on the RMSD from the first frame using cpptraj[37] hierarchical algorithm to obtain 100 distinct structures for pocket calculation. Three regions were defined for individual calculations: the active site pocket, distal site pocket, and dimer interface. Pocket volume was calculated using POVME 2[38] with 1 Å grid spacing. The point inclusion sphere was determined based on the average center of mass of

residues 7-198 with a 12 Å radius for the active site pocket, average center of mass of residues 198-306 with a 10 Å radius for the distal site pocket, and the average center of mass between residues within 3.5 Å of the other dimer with a 10 Å radius for the dimer interface. The pocket volumes were chosen to be underestimated as opposed to overestimated and were calculated with a minimal radius to avoid extensively calculating known solvent exposed regions. As shown in the FTMap solvent screening (**Figure S10, Figure S15**), fragments bound to various regions around the distal site crevices. Increasing the radius of the pocket calculation space would incorporate solvent exposed regions from many directions that wouldn't be useful and would involve other regions of the protein. Due to this volume dependence on chosen radius, these volumes are intended as an initial comparative metrics for identifying structures to analyze further, rather than concrete final values. For analysis of water occupancy, the GIST (Grid Inhomogeneous Solvation Theory Method)[39] function of cpptraj was used for the active site region, and the density of oxygen centers map was used to visualize results in VMD.[40]

To quantify druggability, we have used the PockDrug[19] webserver, which predicts druggability probability based on a variety of factors including geometry, hydrophobicity, aromaticity, among others. Additionally, we have used the FTMap[20] computational solvent mapping webserver to determine druggable "hot spots." Pocket volumes calculated by POVME[38] from each of the three regions were clustered into 10 populations and representative frames for each were selected for analysis. The fpocket estimation method was used for PockDrug, and the druggability probability of the pocket which aligned best to the POVME[38] calculation sphere was reported.

Principal component analysis

Principal motions were calculated using cpptraj, as described previously,[41] from projection of displacement vectors of each of the backbone atoms onto a diagonalized mass-weighted covariance matrix after rms fitting of every atom except protons to the first frame. Residue correlation matrices were also generated this way. For the graphical abstract, reweighting of the potential energy was carried out on one 200 ns simulation of the one N3 non-covalently bound dimer using PyReweighting toolkit.[42] Reweighting was carried out using the Maclaurin series expansion to the 10th order with a bin size of 6 and maximum energy of 100, based on the first two principal components. One conventional 200 ns MD simulation was performed for comparison and outline in the figure. For the conformational sampling comparison shown in **Figure 4**, the dimer-apo system was used for consistency with Riken and DESRES simulations. The colors of the histograms are normalized so the same number of datapoints (frames) is considered. Therefore, 40,000 frames were used for all of the in house simulations, and 50,000 frames were used for the Riken and DESRES simulations. The color bars were set to a maximum value of 0.0008 and 0.0010 respectively. This results in densely localized (yellow color) in the short simulations versus less density (blue color) and more spread out in the longer and enhanced simulations. Visualization of the first normal mode was carried out using the Normal Mode Wizard plugin of VMD.

Acknowledgments:

We thank Michael D. Burkart for advisement, and the NSF GRFP for supporting T.S. under grant number DGE-1650112. Simulations were carried out using GPU nodes of the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC). This work was supported by NIH GM031749.

Chapter 6, in full, is a reprint of the material as it appears: Sztain, T.; Amaro, R.; McCammon, J. A. Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Protease. *JCIM*. **2021**. The dissertation author is the primary investigator and author of this manuscript.

6.7 References

1. COVID-19 Map. *Johns Hopkins Coronavirus Resource Center* https://coronavirus.jhu.edu/map.html.

2. Patick, A. K.; Potts, K. E. Protease Inhibitors as Antiviral Agents. *Clin. Microbiol. Rev.* **1998**, *11* (4), 614–627.

3. De Clercq, E.; Li, G. Approved Antiviral Drugs over the Past 50 Years. *Clin. Microbiol. Rev.* **2016**, *29* (3), 695–747.

4. Gorbalenya, A. E.; Baker, S. C.; Baric, R. S.; de Groot, R. J.; Drosten, C.; Gulyaeva, A. A.; Haagmans, B. L.; Lauber, C.; Leontovich, A. M.; Neuman, B. W.; Penzar, D.; Perlman, S.; Poon, L. L. M.; Samborskiy, D. V.; Sidorov, I. A.; Sola, I.; Ziebuhr, J.; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The Species Severe Acute Respiratory Syndrome-Related Coronavirus : Classifying 2019-NCoV and Naming It SARS-CoV-2. *Nature Microbiology* **2020**, *5* (4), 536–544.

5. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582* (7811), 289–293.

6. Naik, B.; Gupta, N.; Ojha, R.; Singh, S.; Prajapati, V. K.; Prusty, D. High Throughput Virtual Screening Reveals SARS-CoV-2 Multi-Target Binding Natural Compounds to Lead Instant Therapy for COVID-19 Treatment. *Int J Biol Macromol* **2020**, *160*, 1–17.

7. Fischer, A.; Sellner, M.; Neranjan, S.; Smieško, M.; Lill, M. A. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. *Int J Mol Sci* **2020**, *21* (10), 3626.

8. Dai, W.; Zhang, B.; Jiang, X.-M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Peng, J.; Liu, F.; Li, C.; Li, Y.; Bai, F.; Wang, H.; Cheng, X.; Cen, X.; Hu, S.; Yang, X.; Wang, J.; Liu, X.; Xiao, G.; Jiang, H.; Rao, Z.; Zhang, L.-K.; Xu, Y.; Yang, H.; Liu, H. Structure-Based

Design of Antiviral Drug Candidates Targeting the SARS-CoV-2 Main Protease. *Science* **2020**, *368* (6497), 1331–1335.

9. Anand, K.; Palm, G. J.; Mesters, J. R.; Siddell, S. G.; Ziebuhr, J.; Hilgenfeld, R. Structure of Coronavirus Main Proteinase Reveals Combination of a Chymotrypsin Fold with an Extra α-Helical Domain. *EMBO J* **2002**, *21* (13), 3213–3224.

10. Wang, F.; Chen, C.; Tan, W.; Yang, K.; Yang, H. Structure of Main Protease from Human Coronavirus NL63: Insights for Wide Spectrum Anti-Coronavirus Drug Design. *Sci Rep* **2016**, *6*.

11. Chen, Y.; Liu, Q.; Guo, D. Emerging Coronaviruses: Genome Structure, Replication, and Pathogenesis. *J Med Virol* **2020**, *92* (4), 418–423.

12. Wu, C.; Liu, Y.; Yang, Y.; Zhang, P.; Zhong, W.; Wang, Y.; Wang, Q.; Xu, Y.; Li, M.; Li, X.; Zheng, M.; Chen, L.; Li, H. Analysis of Therapeutic Targets for SARS-CoV-2 and Discovery of Potential Drugs by Computational Methods. *Acta Pharm Sin B* **2020**, *10* (5), 766–788.

13. Ullrich, S.; Nitsche, C. The SARS-CoV-2 Main Protease as Drug Target. *Bioorg Med Chem Lett* **2020**, *30* (17), 127377.

14. Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **2000**, *43* (11), 2100–2114.

15. Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, *114* (10), 2271–2278.

16. Zwier, M. C.; Chong, L. T. Reaching Biological Timescales with All-Atom Molecular Dynamics Simulations. *Current Opinion in Pharmacology* **2010**, *10* (6), 745–752.

17. Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J Chem Theory Comput* **2015**, *11* (8), 3584–3595.

18. Pang, Y. T.; Miao, Y.; Wang, Y.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics in NAMD. *J Chem Theory Comput* **2017**, *13* (1), 9–19.

19. Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J Chem Inf Model* **2015**, *55* (4), 882–895.

20. Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand Binding Hot Spots of Proteins. *Nat Protoc* **2015**, *10* (5), 733–755.

21. Dubanevics, I.; McLeish, T. C. B. Computational Analysis of Dynamic Allostery and Control in the SARS-CoV-2 Main Protease. *Journal of The Royal Society Interface* **2021**, *18* (174), 20200591.

22. Sencanski, M.; Perovic, V.; Pajovic, S. B.; Adzic, M.; Paessler, S.; Glisic, S. Drug Repurposing for Candidate SARS-CoV-2 Main Protease Inhibitors by a Novel In Silico Method. *Molecules* **2020**, *25* (17).

23. Miao, Y.; McCammon, J. A. Chapter Six - Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. In *Annual Reports in Computational Chemistry*; Dixon, D. A., Ed.; Elsevier, **2017**; Vol. 13, pp 231–278.

24. Komatsu, T. S.; Koyama, Y.; Okimoto, N.; Morimoto, G.; Ohono, Y.; Taiji, M. COVID-19 related trajectory data of 10 microseconds all atom molecular dynamics simulation of SARS-CoV-2 dimeric main protease. **2020** v2 doi: 10.17632/vpps4vhryg.2.

25. D. E. Shaw Research, 'Molecular Dynamics Simulations Related to SARS-CoV-2,' D. E. Shaw Research Technical Data. **2020** http://www.deshawresearch.com/resources_sarscov2.html.

26. Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential Dynamics of Proteins. *Proteins* **1993**, *17* (4), 412–425.

27. Inizan, T. J.; Célerse, F.; Adjoua, O.; Ahdab, D. E.; Jolly, L.-H.; Liu, C.; Ren, P.; Montes, M.; Lagarde, N.; Lagardère, L.; Monmarché, P.; Piquemal, J.-P. High-Resolution Mining of the SARS-CoV-2 Main Protease Conformational Space: Supercomputer-Driven Unsupervised Adaptive Sampling. *Chem. Sci.* **2021**.

28. Zimmerman, M. I.; Porter, J. R.; Ward, M. D.; Singh, S.; Vithani, N.; Meller, A.; Mallimadugula, U. L.; Kuhn, C. E.; Borowsky, J. H.; Wiewiora, R. P.; Hurley, M. F. D.; Harbison, A. M.; Fogarty, C. A.; Coffland, J. E.; Fadda, E.; Voelz, V. A.; Chodera, J. D.; Bowman, G. R. SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome. *bioRxiv* **2020**, 2020.06.27.175430.

29. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating PKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W368-371.

30. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J Comput Chem* **2005**, *26* (16), 1668–1688.

31. The PyMOL Molecular Graphics System, Version 2.1.1 Schrödinger, LLC.

32. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713.

33. 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), 1157–1174.

34. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280.

35. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **1977**, *23* (3), 327–341.

36. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.

37. Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

38. Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput* **2014**, *10* (11), 5047–5056.

39. Ramsey, S.; Nguyen, C.; Salomon-Ferrer, R.; Walker, R. C.; Gilson, M. K.; Kurtzman, T. Solvation Thermodynamic Mapping of Molecular Surfaces in AmberTools: GIST. *Journal of Computational Chemistry* **2016**, *37* (21), 2029–2037.

40. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14* (1), 33–38, 27–28.

41. Casalino, L.; Palermo, G.; Spinello, A.; Rothlisberger, U.; Magistrato, A. All-Atom Simulations Disentangle the Functional Dynamics Underlying Gene Maturation in the Intron Lariat Spliceosome. *PNAS* **2018**, *115* (26), 6584–6589.

42. Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *J. Chem. Theory Comput.* **2014**, *10* (7), 2677–2689.

43. Amaro, R. E.; Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.* **2020**, *60* (6), 2653–2656.

Chapter 7. A glycan gate controls opening of the SARS-CoV-2 spike protein

Terra Sztain†, Surl-Hee Ahn†, Anthony T. Bogetti, Lorenzo Casalino, Jory A. Goldsmith, Ryan S. McCool, Fiona L. Kearns, J. Andrew McCammon, Jason S. McLellan, Lillian T. Chong, Rommie E. Amaro

† These authors contributed equally to this work.

Abstract

SARS-CoV-2 infection is controlled by the opening of the spike protein receptor binding domain (RBD), which transitions from a glycan-shielded (down) to an exposed (up) state in order to bind the human ACE2 receptor and infect cells. While snapshots of the up and down states have been obtained by cryoEM and cryoET, details of the RBD opening transition evade experimental characterization. Here, over 200 μs of weighted ensemble (WE) simulations of the fully glycosylated spike ectodomain allow us to characterize more than 300 continuous, kinetically unbiased RBD opening pathways. Together with biolayer interferometry experiments, we reveal a gating role for the N-glycan at position N343, which facilitates RBD opening. Residues D405, R408, and D427 also participate. The atomic-level characterization of the glycosylated spike activation mechanism provided herein achieves a new high-water mark for ensemble pathway simulations and offers a foundation for understanding the fundamental mechanisms of SARS-CoV-2 viral entry and infection.

7.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped RNA virus and the causative agent of COVID-19, a disease that has caused significant morbidity and mortality worldwide.[1,2] The main infection machinery of the virus is the spike (S) protein, which

sits on the outside of the virus, is the first point of contact that the virion makes with the host cell, and is a major viral antigen.[3] A significant number of cryoEM structures of the spike protein have been recently reported, collectively informing on structural states of the spike protein. The vast majority of resolved structures fall into either 'down' or 'up' states, as defined by the position of the receptor binding domain (RBD), which modulates interaction with the ACE2 receptor for cell entry.[4,5,6]

The RBDs must transition from a down to an up state for the receptor binding motif to be accessible for ACE2 binding (**Figure 1**), and therefore the activation mechanism is essential for cell entry. Mothes et al.[7] used single-molecule fluorescence (Förster) resonance energy transfer (smFRET) imaging to characterize spike dynamics in real-time. Their work showed that the spike dynamically visits four distinct conformational states, the population of which are modulated by the presence of the human ACE2 receptor and antibodies. However smFRET, as well as conventional structural biology techniques, are unable to inform on the atomic-level mechanisms underpinning such dynamical transitions. Recently, all-atom molecular dynamics (MD) simulations of the spike protein with experimentally accurate glycosylation together with corroborating experiments indicated the extensive shielding by spike glycans, as well as a mechanical role for glycans at positions N165 and N234 in supporting the RBD in the open conformation.[8] Conventional MD simulations as performed in Casalino et al.[8] also revealed microsecond-timescale dynamics to better characterize the spike dynamics but were limited to sampling configurations that were similar in energy to the cryoEM structures. Several enhanced sampling MD simulations have been performed to study this pathway, however, these simulations lacked glycosylation for the spike protein[9] or involve the addition of an external force[10] or did not provide mechanistic detail.[11]

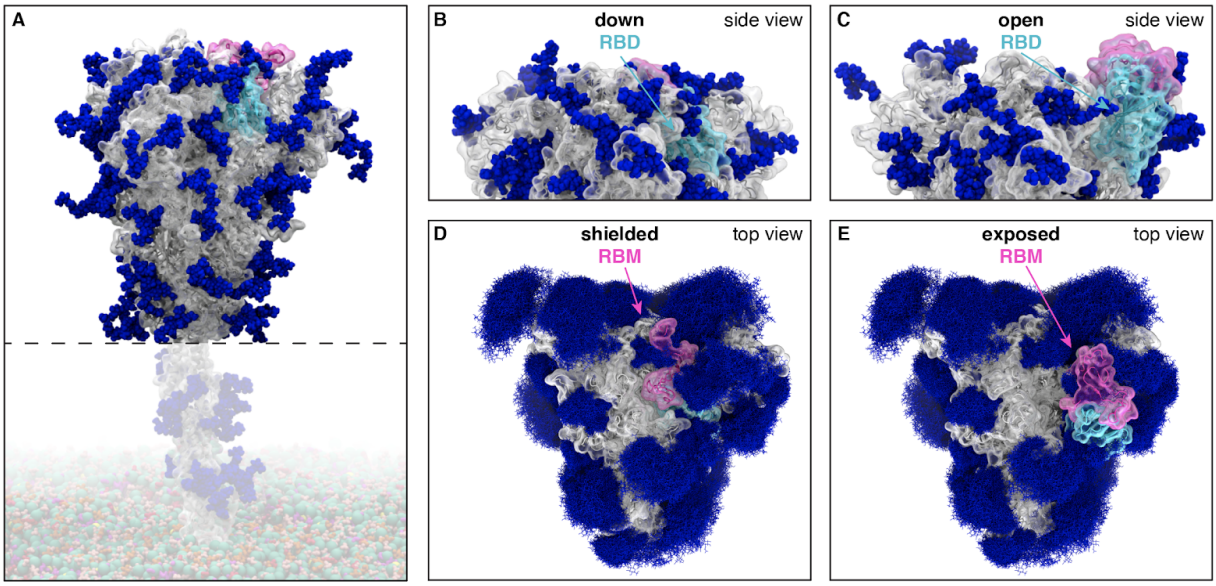


Figure 7.1 Glycosylated spike RBD down and open conformations. (A) The SARS-CoV-2 spike head (gray) with glycans (dark blue) as simulated, with the stalk domain and membrane (not simulated here, but shown in transparent for completeness). RBD shown in cyan, receptor binding motif (RBM) in pink. Side view of the down (shielded, B) and open (exposed, C) RBD. Top view of the closed (shielded, D) and open (exposed, E) RBM. Composite image of glycans (dark blue lines) shows many overlapping snapshots of the glycans over the microsecond simulations.

In this study, we characterized the spike RBD opening pathway for the fully glycosylated SARS-CoV-2 spike protein, in order to gain a detailed understanding of the activation mechanism. We used the weighted ensemble (WE) path sampling strategy[12,13] (**Figure S1**) to enable the simulation of atomistic pathways for the spike opening process. As a path sampling strategy, WE focuses computing power on the functional transitions between stable states rather than the stable states themselves.[14] This is achieved by running multiple trajectories in parallel and periodically replicating trajectories that have transitioned from previously visited to newly visited regions of configurational space[15], thus minimizing the time spent waiting in the initial stable state for "lucky" transitions over the free energy barrier. Given that these transitions are much faster than the waiting times,[16,17] the WE strategy can be orders of magnitude more efficient than conventional MD simulations in generating pathways for rare events such as protein folding and protein binding.[18,19] This efficiency is even higher for slower processes, increasing *exponentially* with the

effective free energy barrier.[20] Not only are dynamics carried out without any biasing force or modifications to the free energy landscape, but suitable assignment of statistical weights to trajectories provides an *unbiased* characterization of the system's time-dependent ensemble properties.[13] The WE strategy therefore generates continuous pathways with unbiased dynamics, yielding the most direct, atomistic views for analyzing the mechanism of functional transitions, including elucidation of transient states that are too fleeting to be captured by laboratory experiments. Furthermore, while the strategy requires a progress coordinate toward the target state, the definition of this target state need not be fixed in advance when applied under equilibrium conditions,[21] enabling us to refine the definition of the target open state of the spike protein based on the probability distribution of protein conformations sampled by the simulation.

Our work characterizes a series of transition pathways of the spike opening and identifies key residues, including a glycan at position N343, that participate in the opening mechanism. Our simulation findings are corroborated by biolayer interferometry experiments, which demonstrate a reduction in the ability of the spike to interact with ACE2 after mutation of these key residues.

Results and Discussion

7.2 Weighted ensemble simulations of spike opening

We used the WE strategy to generate continuous, atomistic pathways of the spike opening process with unbiased dynamics (**Figure 2A-E, Movie S1**); these pathways were hundreds of ns long, excluding the waiting times in the initial "down" state. The protein model was based on the head region (residues 16 to 1140) of the glycosylated SARS-CoV-2 spike from Casalino et al.[8] (**Figure 1**), which in turn was built on the cryoEM structure of the three-RBD-down spike (PDB

ID: 6VXX[5]). The entire simulation system, including explicit water and salt ions, reaches almost half a million atoms. We focused sampling along a two-dimensional progress coordinate to track RBD opening: the difference in the center of mass of the spike core to the RBD and the root mean squared deviation (RMSD) of the RBD (**Figure 2F-G**). On the SDSC Comet and TACC Longhorn supercomputers, 100 GPUs ran the WE simulations in parallel for over a month, generating over 200 μs of glycosylated spike trajectories and more than 200 TB of trajectory data. We simulated a total of 310 independent pathways, including 204 pathways from the RBD "down" conformation (PDB ID: 6VXX[5]) to the RBD "up" conformation (PDB ID: 6VSB[4]) and 106 pathways from the RBD "down" to RBD "open" state in which the RBD twists open beyond the cryoEM structure (PDB ID: 6VSB[4]). Remarkably, the open state that we sampled includes conformations that align closely with the ACE2-bound spike cryoEM structure (PDB ID: 7A95[6]) even though this structure was not a target state of our progress coordinate (**Figure 2F-G**, **Movie S1, Figures S2-3**). This result underscores the value of using (i) equilibrium WE simulations that do not require a fixed definition of the target state and (ii) a two-dimensional progress coordinate that allows the simulations to sample unexpected conformational space along multiple degrees of freedom. The ACE2-bound spike conformation has also been sampled by the Folding@home distributed computing project,[11] and RBD rotation has been detected in cryoEM experiments.[6]

7.3 The N343 glycan gates RBD opening

In the down state, the RBD of the SARS-CoV-2 spike is shielded by glycans at positions N165, N234, and N343.[22] Beyond shielding, a structural role for glycans at positions N165 and N234 has been recently reported, stabilizing the RBD in the up conformation.[8]

Our WE simulations reveal an even more specific, critical role of a glycan in the opening mechanism of the spike: the N343 glycan acts as a "glycan gate" pushing up the RBD by intercalating between residues F490, Y489, F456, and R457 of the ACE2 binding motif in a "hand-jive" motion (**Figure 2A-E**, **Figure 3**, **Movie S2**). This gating mechanism was initially visualized in several successful pathways of spike opening and then confirmed through analysis of all 310 successful pathways in which the N343 glycan was found to form contacts (within 3.5 Å) with each of the aforementioned residues in every successful pathway (**Figure S4**). The same mechanistic behavior of the N343 glycan was observed in two fully independent WE simulations, suggesting the result is robust despite potentially incomplete sampling that can challenge WE and other enhanced sampling simulation methods.[15]

To test the role of the N343 glycan as a key gating residue, we performed biolayer interferometry (BLI) experiments. BLI experiments assess the binding level of the spike receptor binding motif (RBM, residues 438 to 508) to ACE2, acting as a proxy for the relative proportion of RBDs in the up position for each spike variant. No residues directly involved in the binding were mutated (i.e., at the RBM-ACE2 interface), to ensure controlled detection of the impact of RBD opening in response to mutations. Although previous results have shown reduced binding levels for N165A and N234A variants in the SARS-CoV-2 S-2P protein,[8] the N343A variant displayed an even greater decrease in ACE2 binding, reducing spike binding level by ~56 % (**Figure 4, Table S1)**. As a negative control, the S383C/D985C variant,[23] which is expected to be locked by disulfides into the three-RBD-down conformation, showed no association with the ACE2 receptor. These results support the hypothesis that the RBD up conformation is significantly affected by glycosylation at position N343.

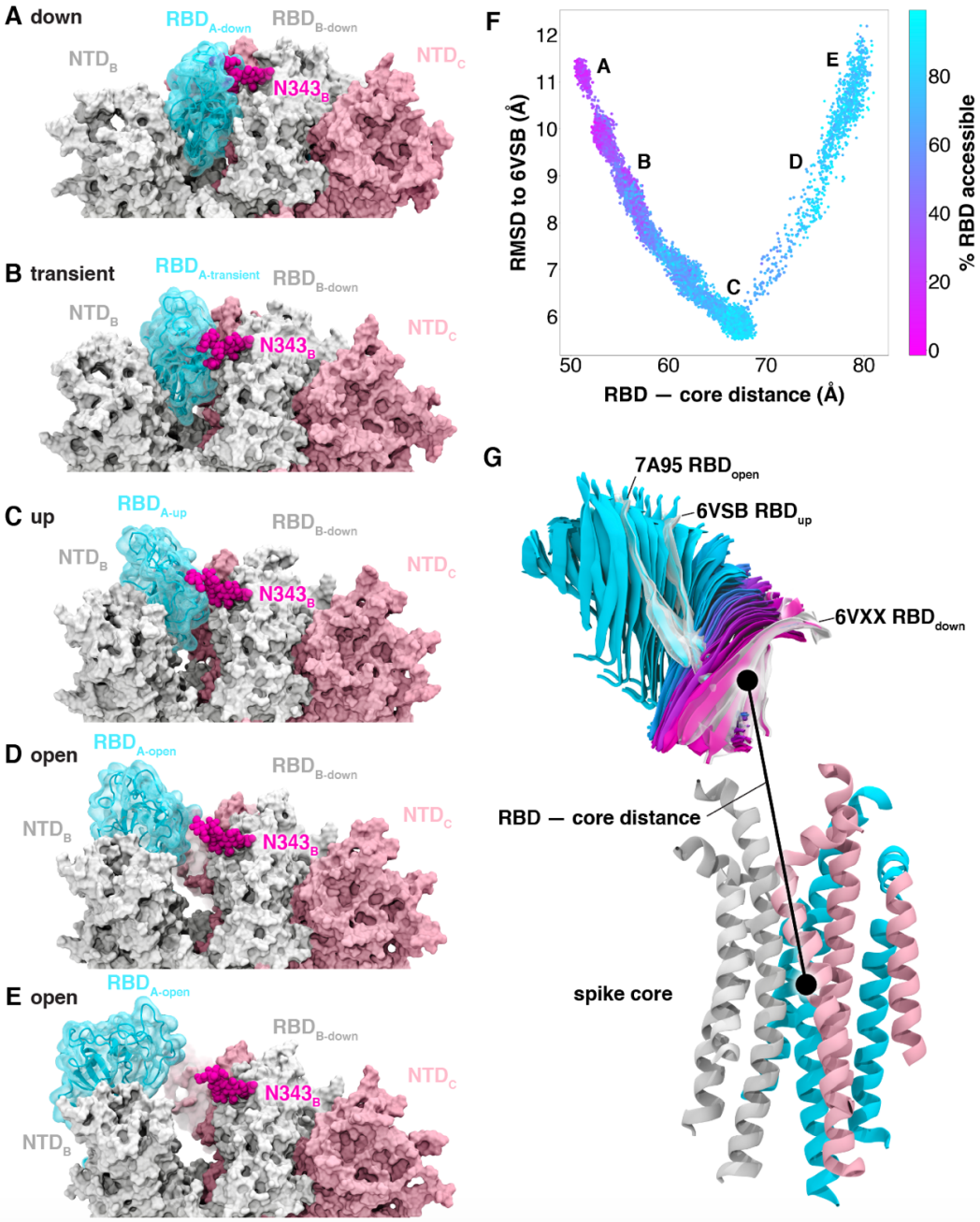


Figure 7.2 Atomically detailed pathways of spike opening. (A-E) Snapshot configurations along the opening pathway with chain A shown in cyan, chain B in gray, and chain C in pink and the glycan at position N343 is shown in magenta. (F) Scatter plot of data from the 310 continuous pathways with the RMSD of RBD to the 6VSB "up" state plotted against RBD — core distance. Data points are colored based on % RBD solvent accessible surface area compared to the RBD "down" state 6VXX. Location of snapshots shown in A-E are labeled. (G) Primary regions of spike defined for tracking progress of opening transition. The spike core is composed of three central helices per trimer, colored according to chain as in (A-E). The RBD contains a structured pair of antiparallel beta sheets and an overlay of snapshots from a continuous WE simulation are shown colored along a spectrum resembling the palette in (F). Overlayed cryoEM structures are highlighted and labeled including the initial RBD "down" state, 6VXX, the target RBD "up" state and the ACE2 bound "open" state, 7A95.

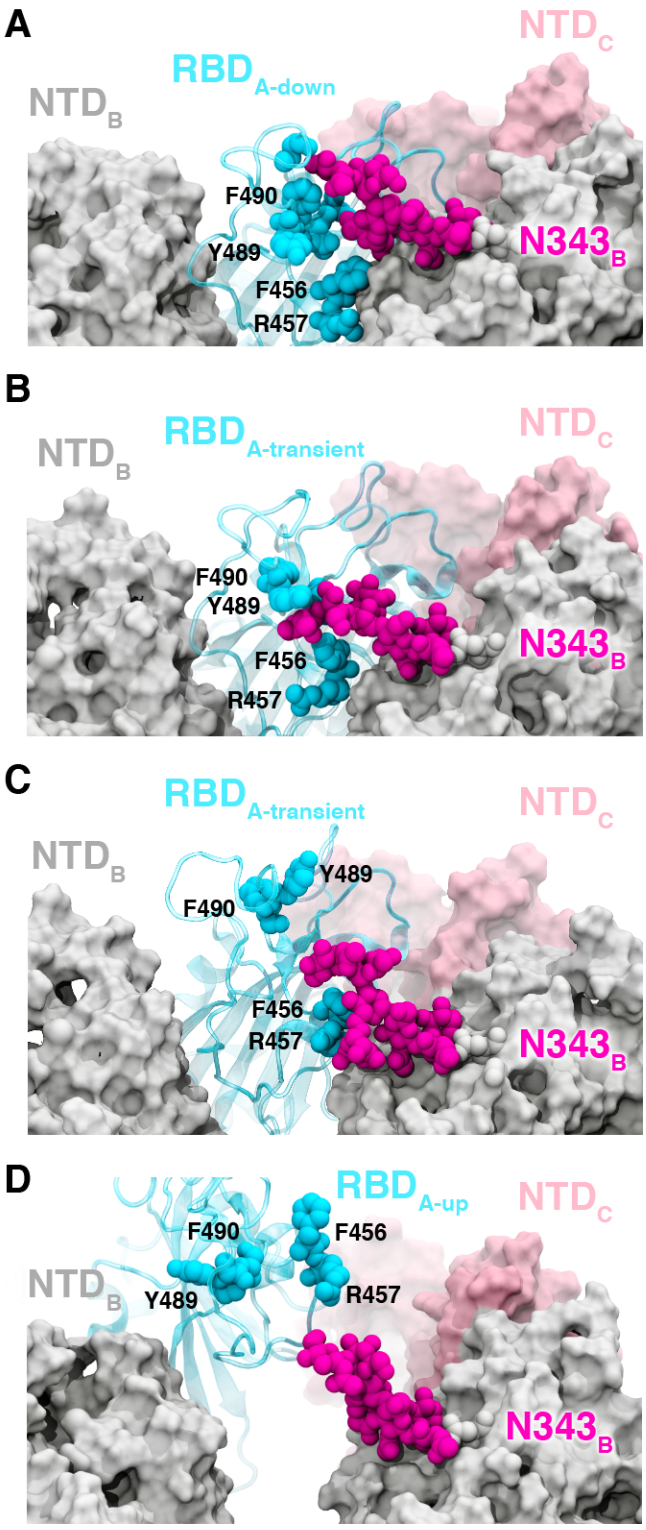


Figure 7.3 Glycan gating by N343. (A-D) Snapshot configurations along the opening pathway with chain A shown in cyan, chain B in gray, chain C in pink, and the glycan at position N343 is shown in magenta. (A) RBD A in the down conformation is shielded by the glycan at position N343 of the adjacent RBD B. (B-D) The N343 glycan intercalates between and underneath the residues F490, Y489, F456, F457 to push the RBD up and open (D).

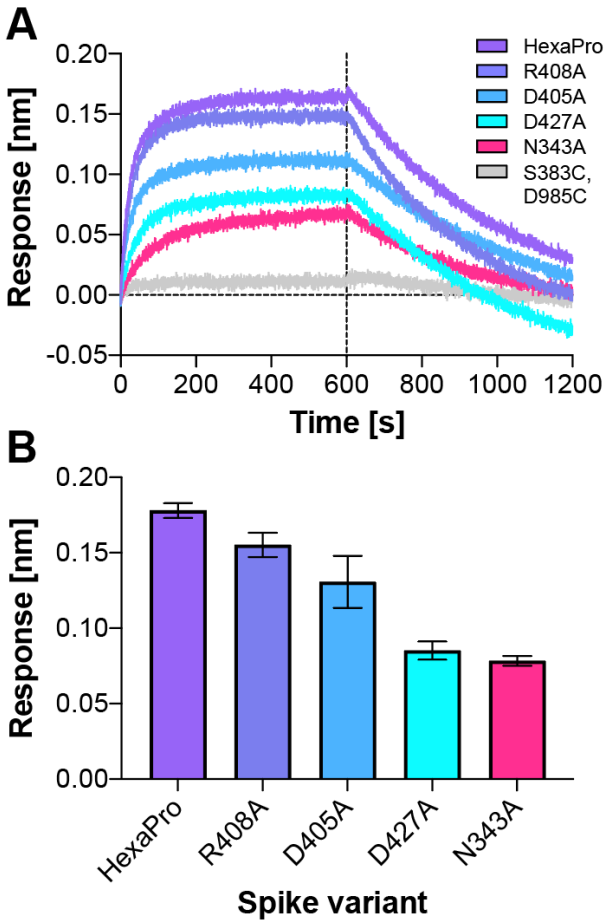


Figure 7.4 ACE2 binding is reduced by mutation of N343 glycosylation site and key salt bridge residues. (A) Biolayer interferometry sensorgrams of HexaPro spike variants binding to ACE2. For clarity, only the traces from the first replicate are shown. (B) Graph of binding response for BLI data collected in triplicate with error bars representing the standard deviation from the mean.

7.4 Atomic details of the opening mechanism

The RBD down state features a hydrogen bond between T415 of the RBD (chain A) and K986 of chain C, a salt bridge between R457 of RBD_A and D364 of RBD_B, and a salt bridge between K462 of RBD_A and D198 of NTD_C (**Figure 5A-C,E Figure S5**). The hydrogen bond $T415_A$ - $K986_C$ spends an average of 12% of the successful pathways to the up state before $K986_C$ makes a short lived (2% average duration to up state) salt bridge with RBD_A D427. (**Figure 5B,E Figure S3**) Next, $K986_C$ forms salt bridges with $E990_C$ and $E748_C$ as the RBD_A continues to open.

These contacts are formed in all 310 successful pathways (**Figure S5**). Mutation of K986 to proline has been used to stabilize the pre-fusion spike,[24,25] including in vaccine development,[26] and these simulations provide molecular context to an additional role of this residue in RBD opening.

Subsequently, at an average of 16% of the way through the successful pathways to the up state, the $R457_A$ - $D364_B$ salt bridge is broken, prompting the RBD_A to twist upward, away from RBD_B towards RBD_C and forms a salt bridge between R408 of RBD_A and D405 of RBD_C (**Figure 5C and 5E, Figure S5**). This salt bridge persists for 20% of the successful trajectories to the up state, and is also present in all 310 successful pathways.

A salt bridge between R466 of RBD_A and E132 from NTD_B is present in 189 out of 204 successful pathways to the up state, and all 106 pathways to the open state. This contact is most prevalent during the transition between the up and open state. Finally, the salt bridge between D428 of RBD_A and R454 of RBD_C is present only in all 106 pathways from the up to the open state and is the last salt bridge between the RBD and the spike in the open state before the S1 subunit begins to peel off (**Figure 5D and 5E, Figure S5**), at which point the last remaining contact to the RBD_A is the glycan at position N165 of NTB_B.

Additional BLI experiments of the key identified spike residues R408A, D405A, and D427A corroborate the pathways observed in our simulations. Each of these reduces the binding interactions of the spike with ACE2 by ~13%, ~27%, and ~52%, respectively (**Figure 4, Table S1**). We also note that identified residues D198, N343, D364, D405, R408, T415, D427, D428, R454, R457, R466, E748, K986 and E990 are conserved between SARS-CoV and SARS-CoV-2

spikes, supporting their significance in coordinating the primary spike function of RBD opening. The emerging mutant SARS-CoV-2 strains B.1: D614G; B.1.1.7: H69-V70 deletion, Y144-Y145 deletions, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H; B.1.351: L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, A701V; P1: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y ,D614G, H655Y, T1027I and CAL.20C: L452R, D614G[27] do not contain mutants in the residues we identified here to facilitate RBD opening. Analysis of neighboring residues and glycans to those mutated in the emerging strains along the opening pathway is detailed in **Table S2**, and distances between each residue and glycan to RBD_A is summarized in **Movie S3.**

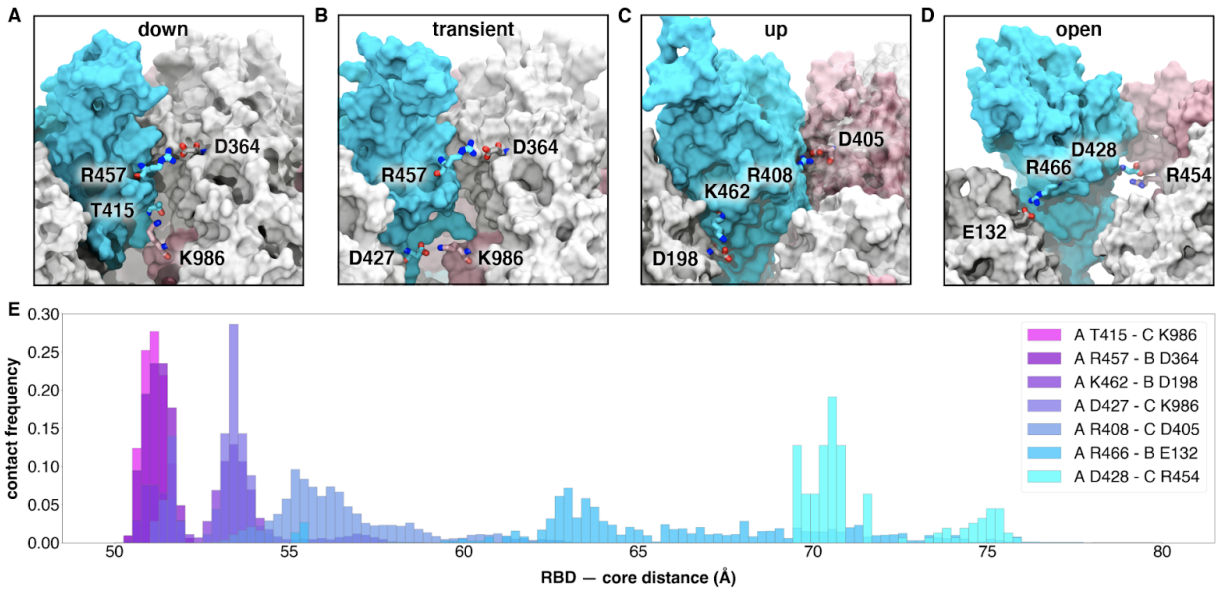


Figure 7.5 Salt bridges and hydrogen bonds along the opening pathway. (A-D) salt bridge or hydrogen bond contacts made between RBD A, shown in blue, with RBD B, shown in gray, or RBD C shown in pink within the down, transient, up, and open conformations, respectively. (E) Histogram showing the frequency at which residues from (A-D) are within 3.5 Å of each other relative to RBD — core distance. Frequencies are normalized to 1.

7.5 Conclusions

We report extensive weighted ensemble molecular dynamics simulations of the glycosylated SARS-CoV-2 spike head characterizing the transition from the down to up conformation of the RBD. Over 200 microseconds of simulation provide more than 300 independent RBD opening transition pathways. Analysis of these pathways from independent WE simulations indicates a clear gating role for the glycan at N343, which lifts and stabilizes the RBD throughout the opening transition. We also characterize an "open" state of the spike RBD, in which the N165 glycan of chain B is the last remaining contact with the RBD *en route* to further opening of S1. Biolayer interferometry experiments of residues identified as key in the opening transitions, including N343, D405, R408, and D427, broadly supported our computational findings. Notably, a 56% decrease in ACE2 binding of the N343A mutant, compared to a 40% decrease in N234A mutant, and 10% decrease in the N165A mutant reported previously[8] evidenced the key role of N343 in gating and assisting the RBD opening process, highlighting the importance of sampling functional transitions to fully understand mechanistic detail. Our work indicates a critical gating role of the N343 glycan in spike opening and provides new insights to mechanisms of viral infection for this important pathogen.

Acknowledgements:

We are grateful for the efforts of the Texas Advanced Computing Center (TACC) Longhorn team and for the compute time made available through a Director's Discretionary Allocation (made possible by the National Science Foundation award OAC-1818253). We thank Dr. Zied Gaieb for helpful discussions around system construction. We thank Mahidhar Tatineni for help with computing on SDSC Comet, as well as a COVID19 HPC Consortium Award for

compute time. We also thank Prof. Carlos Simmerling and his research group (SUNY Stony Brook), and Prof. Adrian Mulholland and his research group (University of Bristol) for helpful discussions related to the spike protein, as well as Prof. Daniel Zuckerman, Dr. Jeremy Copperman, Dr. Matthew Zwier, and Dr. Sinam Saglam for helpful methodological discussions.

Chapter 7, in full, is a reprint of the material as it is appears: Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; McCool, R. S.; Kearns, F. L.; Andrew McCammon, J.; McLellan, J. S.; Chong, L. T.; Amaro, R. E. A Glycan Gate Controls Opening of the SARS-CoV-2 Spike Protein. *bioRxiv* **2021**, 2021.02.15.431212. The dissertation author is the primary co-author of the manuscript along with Surl-Hee Ahn.

7.6 References

(1) Chan, J. F.-W.; Yuan, S.; Kok, K.-H.; To, K. K.-W.; Chu, H.; Yang, J.; Xing, F.; Liu, J.; Yip, C. C.-Y.; Poon, R. W.-S.; Tsoi, H.-W.; Lo, S. K.-F.; Chan, K.-H.; Poon, V. K.-M.; Chan, W.-M.; Ip, J. D.; Cai, J.-P.; Cheng, V. C.-C.; Chen, H.; Hui, C. K.-M.; Yuen, K.-Y. A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus Indicating Person-to-Person Transmission: A Study of a Family Cluster. *The Lancet* **2020**, *395* (10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

(2) Lu, R.;Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; Bi, Y.; Ma, X.; Zhan, F.; Wang, L.; Hu, T.; Zhou, H.; Hu, Z.; Zhou, W.; Zhao, L.; Chen, J.; Meng, Y.; Wang, J.; Lin, Y.; Yuan, J.; Xie, Z.; Ma, J.; Liu, W. J.; Wang, D.; Xu, W.; Holmes, E. C.; Gao, G. F.; Wu, G.; Chen, W.; Shi, W.; Tan. W. Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding. *Lancet* **2020**, *395* (10224), 565–574. https://doi.org/10.1016/s0140-6736(20)30251-8.

(3) Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol* **2016**, *3* (1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301.

(4) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation. *Science* **2020**, *367* (6483), 1260–1263. https://doi.org/10.1126/science.abb2507.

(5) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181* (2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058.

(6) Benton, D. J.; Wrobel, A. G.; Xu, P.; Roustan, C.; Martin, S. R.; Rosenthal, P. B.; Skehel, J. J.; Gamblin, S. J. Receptor Binding and Priming of the Spike Protein of SARS-CoV-2 for Membrane Fusion. *Nature* **2020**, *588* (7837), 327–330. https://doi.org/10.1038/s41586-020-2772-0.

(7) Lu, M.; Uchil, P. D.; Li, W.; Zheng, D.; Terry, D. S.; Gorman, J.; Shi, W.; Zhang, B.; Zhou, T.; Ding, S.; Gasser, R.; Prévost, J.; Beaudoin-Bussières, G.; Anand, S. P.; Laumaea, A.; Grover, J. R.; Liu, L.; Ho, D. D.; Mascola, J. R.; Finzi, A.; Kwong, P. D.; Blanchard, S. C.; Mothes, W. Real-Time Conformational Dynamics of SARS-CoV-2 Spikes on Virus Particles. *Cell Host Microbe* **2020**, *28* (6), 880-891.e8. https://doi.org/10.1016/j.chom.2020.11.001.

(8) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**. https://doi.org/10.1021/acscentsci.0c01056.

(9) Gur, M.; Taka, E.; Yilmaz, S. Z.; Kilinc, C.; Aktas, U.; Golcuk, M. Conformational Transition of SARS-CoV-2 Spike Glycoprotein between Its Closed and Open States. *J. Chem. Phys.* **2020**, *153* (7), 075101. https://doi.org/10.1063/5.0011141.

(10) Fallon, L.; Belfon, K.; Raguette, L.; Wang, Y.; Corbo, C.; Stepanenko, D.; Cuomo, A.; Guerra, J.; Budhan, S.; Varghese, S.; Rizzo, R.; Simmerling, C. Free Energy Landscapes for RBD Opening in SARS-CoV-2 Spike Glycoprotein Simulations Suggest Key Interactions and a Potentially Druggable Allosteric Pocket. **2020**. https://doi.org/10.26434/chemrxiv.13502646.v1.

(11) SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome | bioRxiv https://www.biorxiv.org/content/10.1101/2020.06.27.175430v3 (accessed Jan 13, 2021).

(12) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* **1996**, *70* (1), 97–110. https://doi.org/10.1016/S0006-3495(96)79552-8.

(13) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "Weighted Ensemble" Path Sampling Method Is Statistically Exact for a Broad Class of Stochastic Processes and Binning Procedures. *J. Chem. Phys.* **2010**, *132* (5), 054107. https://doi.org/10.1063/1.3306345.

(14) Chong, L. T.; Saglam, A. S.; Zuckerman, D. M. Path-Sampling Strategies for Simulating Rare Events in Biomolecular Systems. *Curr Opin Struct Biol* **2017**, *43*, 88–94. https://doi.org/10.1016/j.sbi.2016.11.019.

(15) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu Rev Biophys* **2017**, *46*, 43–57. https://doi.org/10.1146/annurev-biophys-070816-033834.

(16) Pratt, L. R. A Statistical Method for Identifying Transition States in High Dimensional Problems. *J. Chem. Phys.* **1986**, *85* (9), 5045–5048. https://doi.org/10.1063/1.451695.

(17) Zuckerman, D. M.; Woolf, T. B. Transition Events in Butane Simulations: Similarities across Models. *J. Chem. Phys.* **2002**, *116* (6), 2586–2591. https://doi.org/10.1063/1.1433501.

(18) Adhikari, U.; Mostofian, B.; Copperman, J.; Subramanian, S. R.; Petersen, A. A.; Zuckerman, D. M. Computational Estimation of Microsecond to Second Atomistic Folding Times. *J Am Chem Soc* **2019**, *141* (16), 6519–6526. https://doi.org/10.1021/jacs.8b10735.

(19) Saglam, A. S.; Chong, L. T. Protein–Protein Binding Pathways and Calculations of Rate Constants Using Fully-Continuous, Explicit-Solvent Simulations. *Chem. Sci.* **2019**, *10* (8), 2360–2372. https://doi.org/10.1039/C8SC04811H.

(20) DeGrave, A. J.; Ha, J.-H.; Loh, S. N.; Chong, L. T. Large Enhancement of Response Times of a Protein Conformational Switch by Computational Design. *Nature Communications* **2018**, *9* (1), 1013. https://doi.org/10.1038/s41467-018-03228-6.

(21) Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. *J. Chem. Theory Comput.* **2014**, *10* (7), 2658–2667. https://doi.org/10.1021/ct401065r.

(22) Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-Specific Glycan Analysis of the SARS-CoV-2 Spike. *Science* **2020**. https://doi.org/10.1126/science.abb9983.

(23) Henderson, R.; Edwards, R. J.; Mansouri, K.; Janowska, K.; Stalls, V.; Gobeil, S. M. C.; Kopp, M.; Li, D.; Parks, R.; Hsu, A. L.; Borgnia, M. J.; Haynes, B. F.; Acharya, P. Controlling the SARS-CoV-2 Spike Glycoprotein Conformation. *Nature Structural & Molecular Biology* **2020**, *27* (10), 925–933. https://doi.org/10.1038/s41594-020-0479-4.

(24) Pallesen, J.; Wang, N.; Corbett, K. S.; Wrapp, D.; Kirchdoerfer, R. N.; Turner, H. L.; Cottrell, C. A.; Becker, M. M.; Wang, L.; Shi, W.; Kong, W.-P.; Andres, E. L.; Kettenbach, A. N.; Denison, M. R.; Chappell, J. D.; Graham, B. S.; Ward, A. B.; McLellan, J. S. Immunogenicity and Structures of a Rationally Designed Prefusion MERS-CoV Spike Antigen. *PNAS* **2017**, *114* (35), E7348–E7357. https://doi.org/10.1073/pnas.1707304114.

(25) Hsieh, C.-L.; Goldsmith, J. A.; Schaub, J. M.; DiVenere, A. M.; Kuo, H.-C.; Javanmardi, K.; Le, K. C.; Wrapp, D.; Lee, A. G.; Liu, Y.; Chou, C.-W.; Byrne, P. O.; Hjorth, C. K.; Johnson, N. V.; Ludes-Meyers, J.; Nguyen, A. W.; Park, J.; Wang, N.; Amengor, D.; Lavinder, J. J.; Ippolito, G. C.; Maynard, J. A.; Finkelstein, I. J.; McLellan, J. S. Structure-Based Design of

Prefusion-Stabilized SARS-CoV-2 Spikes. *Science* **2020**, *369* (6510), 1501–1505. https://doi.org/10.1126/science.abd0826.

(26) Ryan Cross. The Tiny Tweak behind COVID-19 Vaccines. *Chemical & Engineering News*. September 29, 2020.

(27) Corum, J.; Zimmer, C. Coronavirus Variants and Mutations. *The New York Times*. February 10, 2021.

(28) Amaro, R. E.; Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.* **2020**, *60* (6), 2653–2656. https://doi.org/10.1021/acs.jcim.0c00319.