Title
Developing chemical biology tools to understand interactions in carrier protein dependent metabolic pathways

Permalink
https://escholarship.org/uc/item/6cf7r22t

Author
Kim, Woojoo Eunice

Publication Date
2019

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA SAN DIEGO

Developing Chemical Biology Tools to Understand Interactions in Carrier Protein Dependent Metabolic Pathways

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

in Chemistry

by

Woojoo Eunice Kim

Committee in charge:

Professor Michael Burkart, Chair
Professor Rommie Amaro
Professor William Gerwick
Professor Kamil Godula
Professor Tadeusz Molinski

2019
This dissertation of Woojoo Eunice Kim is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

______________________________________________________________________________

______________________________________________________________________________

______________________________________________________________________________

______________________________________________________________________________

______________________________________________________________________________

Chair

University of California San Diego

2019
DEDICATION

To my loving family, who have supported and encouraged me during my years in graduate school. Thank you for your love and prayers.
EPIGRAPH

Science without religion is lame, religion without science is blind.

— Albert Einstein
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signature Page................................................................................................................ iii</td>
</tr>
<tr>
<td>Dedication........................................................................................................................ iv</td>
</tr>
<tr>
<td>Epigraph.......................................................................................................................... v</td>
</tr>
<tr>
<td>Table of Contents............................................................................................................. vi</td>
</tr>
<tr>
<td>List of Abbreviations....................................................................................................... ix</td>
</tr>
<tr>
<td>List of Figures................................................................................................................ xi</td>
</tr>
<tr>
<td>List of Tables................................................................................................................... xvi</td>
</tr>
<tr>
<td>List of Schemes............................................................................................................... xvii</td>
</tr>
<tr>
<td>Acknowledgements.......................................................................................................... xviii</td>
</tr>
<tr>
<td>Vita.................................................................................................................................... xxi</td>
</tr>
<tr>
<td>Abstract of the Dissertation......................................................................................... xxii</td>
</tr>
</tbody>
</table>

**Chapter 1.** Developing chemical biology tools to understand interactions in carrier protein dependent metabolic pathways............................................................................................................ 1

References......................................................................................................................... 14

**Chapter 2.** Discovering *de novo* peptide substrates for enzymes using machine learning....................................................................................................................... 17

2.1 Introduction........................................................................................................... 17

2.2 Results and Discussion...................................................................................... 19

2.3 Conclusions....................................................................................................... 35

2.4 Methods.......................................................................................................... 36

2.5 Supplementary Information......................................................................... 42

2.6 References...................................................................................................... 55
**Chapter 3.** Activity-guided single cell genomics reveals a selective biosynthetic enzyme-substrate interaction........................................................................................................ 60

3.1 Introduction........................................................................................................ 60
3.2 Results and Discussion...................................................................................... 62
3.3 Conclusions....................................................................................................... 66
3.4 Methods............................................................................................................ 67
3.5 Supplementary Information............................................................................ 71
3.6 References....................................................................................................... 80

**Chapter 4.** Mechanistic probes for the epimerization domain of nonribosomal peptide synthetases........................................................................................................... 82

4.1 Introduction....................................................................................................... 82
4.2 Results and Discussion...................................................................................... 84
4.3 Conclusions..................................................................................................... 91
4.4 Methods.......................................................................................................... 92
4.5 Supplementary Information............................................................................ 95
4.6 References..................................................................................................... 109

**Chapter 5.** Developing crosslinkers specific for epimerization domain to understand its mechanism........................................................................................................... 113

5.1 Introduction..................................................................................................... 113
5.2 Results and Discussion.................................................................................... 114
5.3 Conclusions.................................................................................................... 121
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4 Methods</td>
<td>121</td>
</tr>
<tr>
<td>5.5 Supplementary Information</td>
<td>123</td>
</tr>
<tr>
<td>5.6 References</td>
<td>149</td>
</tr>
</tbody>
</table>
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRP</td>
<td>Non-ribosomal peptide</td>
</tr>
<tr>
<td>NRPS</td>
<td>Non-ribosomal peptide synthetase</td>
</tr>
<tr>
<td>PPTase</td>
<td>4’-Phosphopantetheinyl transferase</td>
</tr>
<tr>
<td>PPant</td>
<td>4’-Phosphopantetheine</td>
</tr>
<tr>
<td>PK</td>
<td>Polyketide</td>
</tr>
<tr>
<td>FAS</td>
<td>Fatty acid synthase</td>
</tr>
<tr>
<td>NRPS</td>
<td>Holo-ACP synthase</td>
</tr>
<tr>
<td>Sfp</td>
<td>Surfactin phosphopantetheinyl transferase</td>
</tr>
<tr>
<td>AcpS</td>
<td>Carrier protein</td>
</tr>
<tr>
<td>CP</td>
<td>Acyl carrier protein</td>
</tr>
<tr>
<td>ACP</td>
<td>E. coli fatty acid synthase acyl carrier protein</td>
</tr>
<tr>
<td>AcpP</td>
<td>Peptidyl carrier protein</td>
</tr>
<tr>
<td>PCP</td>
<td>Coenzyme A</td>
</tr>
<tr>
<td>CoA</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>ATP</td>
<td>Peptide Optimization with Optimal Learning</td>
</tr>
<tr>
<td>POOL</td>
<td>4-Dimethylnaphthalene</td>
</tr>
<tr>
<td>DMN-P</td>
<td>4-Dimethylnaphthalene-pantetheinamide</td>
</tr>
<tr>
<td>RSG</td>
<td>RedoxSensor Green</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>WC</td>
<td>Water column</td>
</tr>
<tr>
<td>Term</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Single amplified genomes</td>
<td>SAG</td>
</tr>
<tr>
<td>Communication-mediating</td>
<td>COM</td>
</tr>
<tr>
<td>Adenylation domain</td>
<td>A domain</td>
</tr>
<tr>
<td>Condensation domain</td>
<td>C domain</td>
</tr>
<tr>
<td>Epimerization domain</td>
<td>E domain</td>
</tr>
<tr>
<td>Pyridoxal 5’-phosphate</td>
<td>PLP</td>
</tr>
<tr>
<td>Chlorovinylglycine</td>
<td>CVG</td>
</tr>
<tr>
<td>Initiation module of Gramicidin synthetase</td>
<td>GrsA</td>
</tr>
<tr>
<td>Initiation module of Tyrocidine synthetase</td>
<td>TycA</td>
</tr>
<tr>
<td>Second module of Tyrocidine synthetase</td>
<td>TycB&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>Diketopiperazine</td>
<td>DKP</td>
</tr>
<tr>
<td>Molecular dynamics</td>
<td>MD</td>
</tr>
<tr>
<td><em>E. coli</em> Pantothenate Kinase</td>
<td>CoaA</td>
</tr>
<tr>
<td><em>E. coli</em> Phosphopantetheine adenylyl transferase</td>
<td>CoaD</td>
</tr>
<tr>
<td><em>E. coli</em> Dephosphocoenzyme A kinase</td>
<td>CoaE</td>
</tr>
<tr>
<td>5-Carboxytetramethylrhodamine</td>
<td>TAMRA</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>1.1</td>
<td>Post-translational modification and manipulation of carrier protein by “one-pot” chemoenzymatic method</td>
</tr>
<tr>
<td>1.2</td>
<td>Peptide Optimization with Optimal Learning (POOL) method outline</td>
</tr>
<tr>
<td>1.3</td>
<td>Sequence alignments between top recommended peptides and native PPTase substrates</td>
</tr>
<tr>
<td>1.4</td>
<td>The overview of enzyme mining strategy using fluorescent pantetheine analog and single-cell genomics</td>
</tr>
<tr>
<td>1.5</td>
<td>Tyrocidine biosynthetic pathway as an example of non-ribosomal peptide synthetase overview</td>
</tr>
<tr>
<td>1.6</td>
<td>Schematic diagram showing diketopiperazine (DKP) assay to evaluate epimerization domain inhibitor</td>
</tr>
<tr>
<td>1.7</td>
<td>Proposed mechanism of chlorovinylglycine (CVG) inhibition towards epimerization domain of GrsA</td>
</tr>
<tr>
<td>1.8</td>
<td>Epimerization domain inhibitors containing sulfonate moiety and its crosslinking validation</td>
</tr>
<tr>
<td>1.9</td>
<td>Mutational study showing the crosslinking site on catalytic residues of E domain</td>
</tr>
<tr>
<td>2.1</td>
<td>Overview of the iterative Peptide Optimization with Optimal Learning (POOL) method workflow</td>
</tr>
<tr>
<td>2.2</td>
<td>Specific peptide hits identified by POOL and model validation</td>
</tr>
<tr>
<td>2.3</td>
<td>Simulated method performance comparison and peptide diversity evaluation</td>
</tr>
<tr>
<td>2.4</td>
<td>Sequence alignment of hit peptides relative to native PPTase substrates</td>
</tr>
<tr>
<td>2.5</td>
<td>Validation of PPTase-specific peptide hits using GFP-peptide tags</td>
</tr>
</tbody>
</table>
Figure 2.6  Scheme illustrating selective Sfp and AcpS labeling on duplicate SPOT membranes................................................................. 34

Figure S2.1  Reaction scheme illustrating labeling by PPTase of carrier protein and POOL peptide substrate at conserved serine................................. 50

Figure S2.2  Twenty essential amino acids organized by 8 reduced classes........ 51

Figure S2.3  Distances in simulation study between training data and peptides recommended by POOL and Mutation............................................................. 52

Figure S2.4  Full SDS-PAGE images of GFP-peptide fusion proteins after PPTase labeling........................................................................................................ 53

Figure S2.5  LC-MS for characterizing labeling of AcpS peptides..................... 54

Figure 3.1  Activity-guided microbial single cell genomics............................. 62

Figure 3.2  Application of activity-guided single cell genomics to mine a selective CP and PPTase interactions from a colonial tunicate microbiome......................................................... 66

Figure S3.1  Enlarged super-resolution AiryScan images................................... 75

Figure S3.2  Discovery, synthesis and expression of Carrier Protein 1 (CP1)........ 76

Figure S3.3  Discovery, synthesis and expression of Carrier Protein 2 (CP2)........ 77

Figure S3.4  Discovery, synthesis and expression of 4’-Phosphopantetheinytransferase 1 (PPT1)................................................................................................. 78

Figure S3.5  Labeling experiments...................................................................... 79

Figure 4.1  Biosynthetic pathway of gramicidin S.............................................. 85

Figure 4.2  HPLC analysis of DKP formation after preincubation of GrsA with probes 1 and 2.............................................................. 87

Figure 4.3  Schematic diagram of probe 1 inhibition on epimerization domain.............................................................................................................. 89

Figure 4.4  Distribution of the substrate S-C(O)-C α-C β dihedral angles of (A) L-
Phe-PCP-E, (B) D-Phe-PCP-E, and (C) 1-PCP-E

Figure S4.1  HPLC analysis of % DKP formation after preincubation of probe 1 and 2 without ATP

Figure S4.2  Primary HPLC trace of %DKP formation of the control and with the probe 1

Figure S4.3  Schematic diagram showing DKP assay

Figure S4.4  GrsA PCP-E didomain

Figure S4.5  Root mean square (RMS) deviations of various aminoacyl-PCP-E didomain of GrsA

Figure S4.6  Root mean square (RMS) fluctuations of the various aminoacyl-PCP-E didomain of GrsA

Figure S4.7  Distances between catalytic residues and key substrate moieties sampled in cMD simulations

Figure S4.8  Distances between catalytic residues and key substrate moieties sampled in GaMD simulations

Figure S4.9  Distribution of C(O)-Cα dihedral angle of the aminoacyl group of the aminoacyl-PCP-E didomains sampled using cMD

Figure S4.10 Distribution of C(O)-Cα dihedral angle of the aminoacyl group of the aminoacyl-PCP-E didomains sampled using GaMD

Figure S4.11 His753-thioester hydrogen bonding in L-Phe-PCP, D-Phe-PCP, and 1-PCP rotamers sampled using cMD simulations

Figure S4.12 His753-thioester hydrogen bonding in L-Phe-PCP, D-Phe-PCP, and 1-PCP rotamers sampled using GaMD simulations

Figure 5.1  Crosslinker design and its validation by running a reaction on a SDS-PAGE gel

Figure 5.2  Different length crosslinkers and its crosslinking profile on a gel
Figure 5.3 Crosslinking pretease digestion mass spectrometry analysis compared with wild-type protein........................................................................... 118

Figure 5.4 Crosslinking reaction on alanine mutant library with the probe 3........ 120

Figure S5.1 Full SDS-PAGE gel image of TycA crosslinking cronstructs with different probes........................................................................................................ 134

Figure S5.2 Full SDS-PAGE gel image of TycA mutant crosslinking constructs with C5-Bsulfo probe........................................................................................................ 135

Figure S5.3 SDS-PAGE gel showing TycA alanine mutant constructs.............. 135

Figure S5.4 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S2 in CD$_3$OD........................................................................................................................................ 136

Figure S5.5 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S3 in CD$_3$OD........................................................................................................................................ 137

Figure S5.6 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C4-Ms in (CD)$_2$CO.................................................................................................................. 138

Figure S5.7 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S4 in CD$_3$OD........................................................................................................................................ 139

Figure S5.8 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S5 in CDCl$_3$........................................................................................................................................ 140

Figure S5.9 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C5-Ms in (CD)$_2$CO.................................................................................................................. 141

Figure S5.10 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S6 in CDCl$_3$........................................................................................................................................ 142

Figure S5.11 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C4-BSulfo in CDCl$_3$........................................................................................................................................ 143

Figure S5.12 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S7 in CDCl$_3$........................................................................................................................................ 144

Figure S5.13 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C5-BSulfo
Figure S5.14 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S8 in CDCl$_3$........................................................................................................ 146

Figure S5.15 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S9 in CDCl$_3$........................................................................................................ 147

Figure S5.16 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C6-BSulfo in CDCl$_3$........................................................................................................ 148
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table S2.1</th>
<th>Gel-based electrophoretic mobility shift kinetics for <em>B. subtilis</em> Sfp</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table S5.1</td>
<td>Primers used in site-directed mutagenesis</td>
<td>133</td>
</tr>
</tbody>
</table>
LIST OF SCHEMES

**Scheme 4.1**  Mechanism-based crosslinking probes chlorovinylglycine (1) and β-fluorophenylalanine (2) for the NRPS epimerization domain .......................... 87

**Scheme S4.1**  Distances between the catalytic groups of His$^{743}$, and Gly$^{748}$, Glu$^{892}$ and the reactive centers of the substrates........................................... 102

**Scheme S5.1**  Synthetic route to C4-Ms (1)................................................................. 123

**Scheme S5.2**  Synthetic route to C5-Ms (2)................................................................. 123

**Scheme S5.3**  Synthetic route to C4-Bsulfo (3).......................................................... 124

**Scheme S5.4**  Synthetic route to C5-Bsulfo (4).......................................................... 124

**Scheme S5.5**  Synthetic route to C6-Bsulfo (5).......................................................... 125
ACKNOWLEDGEMENTS

I would like to acknowledge my advisor and mentor, Professor Michael Burkart, for his guidance and support over the years in my graduate school. I remember listening to his lightning talk during orientation and got excited about modular synthases and developing chemical biology tools to study them. Mike supported my desire to explore and learn different research fields in his lab, which shaped me as an independent scientist. I really appreciate his guidance and advice that he gave me along the way when we met for our meetings and future career decisions. Also, I would like to thank my other mentor, Dr. James La Clair, for his guidance and support for me. I really enjoyed learning and discussing chemistry from him inside and outside of the lab.

I would like to also thank my committee members, Professor Kamil Godula, Professor Rommie Amaro, Professor Ted Molinski, and Professor William Gerwick. All their feedback and comments were essential and helpful in my research progress. Moreover, I would like to thank my collaborators, Professor Nathan Gianneschi, Professor Joseph Noel, Professor Andy McCammon, and Professor Fumihiro Ishikawa, for constructive feedback and discussions throughout my research progress.

I was fortunate to have met all my friends coming to San Diego. They helped me grow and gave me so many good memories. Coming to graduate school from Korea was definitely intimidating experience, however, with amazing friends that I met through graduate school, Allie, Naneki, Jeff, Ben, Thomas; you guys gave me amazing memories and helped me get through graduate school. John, Wonmin, Joohee and now Professor Dokyung Kim; you guys really made my graduate school life fun and enjoyable and it
almost reminded me of my times in undergraduate school. Also, Chanyoo and 언니들, Dayoung, Eunjeong, Jungmin, and Sunmi; it gave me a big relief that we came to San Diego together and you guys have been my huge support since day 1 starting graduate school.

My biggest thank you goes to the past and present lab members in the Burkart group. I still remember the family-like lab environment that made me consider joining the Burkart lab during my first rotation. All of the lab members have been really welcoming and helpful whenever I asked them for help or when we were engaged in scientific discussions. There are so many good memories with the lab from going to group lunch everyday, grabbing drinks to celebrate or to finish our grumpy days together, going to lab weddings and unforgettable lab retreats. Every one of the lab members became my dear friends and I really enjoyed my graduate school life with them.

Finally, I would like to express my sincere thank you to my family for their huge support. They always had me in their prayers and cheered me from Seoul. I am grateful that during my time in graduate school, it brought my family members in Korea and in America all close together. 엄마, 아빠, 미래 고맙고 사랑해요.

I appreciate my collaborators allowing me to use the following publications in my dissertation:

is the primary co-author of this manuscript along with Dr. Lori Tallorin and Dr. Jialei Wang.

Chapter 3, in full, is currently being prepared for submission for publication as: Kim, W. E., Charov, K., Becraft, E. D., Brown, J., La Clair, J. J., Stepanauskas, R., Burkart, M. D. “Activity-Guided Single Cell Genomics Reveals a Selective Biosynthetic Enzyme-Substrate Interaction” The dissertation author is the primary co-author of this manuscript along with Katherine Charov, Dr. Eric D. Becraft, and Dr. Julia Brown.

Chapter 4, in full, is a reprint of the material as it appears: Kim, W. E., Patel, A., Hur, G. H., Tufar, P., Wuo, M. G., McCammon, J. A., Burkart, M. D. “Mechanistic Probes for the Epimerization Domain of Nonribosomal Peptide Synthetases” *ChemBioChem* 2019 20(2), 147-152. The dissertation author is the primary co-author of this manuscript along with Dr. Ashay Patel.

Chapter 5, in full, is currently being prepared for submission for publication as: Kim, W. E.*, Ishikawa, F.*, Re, R. N., Suzuki, T., Dohmae, N., Kakeya, H., Tanabe, G., Burkart, M. D. “Developing Crosslinkers Specific for Epimerization Domain to Understand its Mechanism” The dissertation author is the primary author of this manuscript.
VITA

2019  Doctor of Philosophy, Chemistry, University of California San Diego

2016  Master of Science, University of California San Diego

2014  Bachelor of Science, Sogang University

PUBLICATIONS


ABSTRACT OF DISSERTATION

Developing Chemical Biology Tools to Understand Interactions in Carrier Protein Dependent Metabolic Pathways

by

Woojoo Eunice Kim

Doctor of Philosophy in Chemistry

University of California San Diego, 2019

Professor Michael D. Burkart, Chair

Natural products, such as fatty acids, polyketides, and non-ribosomal peptides, exhibit diverse biological functions toward human health and are constructed via many different pathways; however, these pathways often share the same synthetic logic. Central to these pathways is a carrier protein (CP). The role of a CP is to transfer elongating intermediates between catalytic domains and is mediated by protein-protein interactions between the CP and its partner enzymes. These interactions are transient, making it difficult to understand how they communicate with each other in recognizing a cargo or how the CP
meets the right partner enzymes. The Burkart laboratory has been developing fluorescent and mechanistic probes to understand these CP-partner protein interactions.

The CP requires post-translational modification to become an active form through the action of 4’-phosphopantetheinyl transferase (PPTase) loading 4’phosphopantetheine prosthetic (PPant) arm at the serine residue. The first part of this dissertation focuses on utilizing the PPTase function to search for (1) the minimum peptide substrate required for CP recognition via machine learning and (2) the biosynthetic pathway from unculturable microorganisms. Using the ability of some PPTases to recognize and transfer an unnatural PPant arm to a CP, we identified orthogonal peptide substrates that can be labeled by two different classes of PPTases. These peptide substrates can be appended to a protein and used as a peptide tag. Furthermore, we employed a PPTase to fluorescently label CPs in previously uncultured microorganisms. This enabled us to sequence single cells for the identification of a biosynthetic cluster with active PPTase-CP pairs.

In the second part of this dissertation, we developed two mechanistic probes to study protein-protein interactions between epimerization (E) domain and peptidyl carrier protein in non-ribosomal peptide synthetases. D-amino acids are incorporated into non-ribosomal peptides, which contribute to their unique conformation and bioactivity. The E domains convert L- to D-amino acids by deprotonating/reprotonating Cα-H. Despite the past research on the E domain, the mechanistic details remain unclear. Herein, with the help of molecular dynamics simulations and mutational studies, our research reveals more detailed evidence on which catalytic residues work as an acid/base in this mechanism.
Chapter 1. Developing chemical biology tools to understand interactions in carrier protein-dependent metabolic pathways

Nature has evolved elaborate machineries to provide a plethora of natural products with diverse biological functions and roles. In the past decades, natural products have gained much interest from pharmaceutical companies due to the variety of their beneficial bioactivities towards human health. One of the ways to produce these natural products is through iterative cycles of elongating small building blocks to assemble complex macromolecules. These biomachineries utilize multiple functional domains working together to produce natural products such as fatty acids, polyketides, and non-ribosomal peptides (NRPs). There are two types of synthases based on how the functional domains are arranged: type I contains functional domains in one polypeptide forming a megasynthase and type II has discrete functional domains that work together to elongate an intermediate.

Central to these natural product producing synthases is a small 8-kDa sized non-catalytic carrier protein (CP) that plays an essential role in transporting an elongating intermediate from one catalytic domain to the other. In nature, CPs pass through two different states, the inactive apo form and the active holo form. An apo-CP must undergo post-translational modification by 4’-phosphopantetheinyl transferase (PPTase) to become the biologically active holo form. PPTase transfers and covalently attaches 4’phosphopantetheine (PPant) prosthetic arm from coenzyme A (CoA) to a conserved serine residue of the CP.

In recent years, the Burkart laboratory has developed a “one-pot” chemoenzymatic method to load synthetic cargo onto a CP to study its activity, including interactions with
partner proteins in CP-dependent metabolic pathways. A variety of pantetheine analogs have been synthesized and converted to coenzyme A analogs in vitro using CoA biosynthetic enzymes from *Escherichia coli* (CoaA, CoaD, and CoaE). Once the CoA analogs have formed, PPTase transfers the synthetic PPant arm onto the CP. With the inherent promiscuity of these “one-pot” enzymes, pantetheine analogs of diverse functionalities and size can be loaded onto a CP for visualization and structural studies. (Figure 1.1) Of these, it has been shown that the fluorescent coumarin pantetheinamide probes are cell permeable and can be modified to produce a CoA analog and then loaded onto a CP in vivo.

![Figure 1.1](image)

Figure 1.1 Post-translational modification and manipulation of carrier protein by “one-pot” chemoenzymatic method. (A) Post-translational modification of apo-CP to an active holo-CP by appending 4'-phosphopantetheine arm from CoA. (B) List of pantetheine analogs of diverse functionalities and sizes. (C) “one-pot” chemoenzymatic method utilizing coenzyme A biosynthetic pathway enzymes in attaching synthetic pantetheine analogs onto a CP to make a crypto-CP.

There are two subtypes in the PPTase superfamily depending on the amino acid sequence, structural information, and target synthases. The AcpS-type (holo-acyl carrier
protein synthase) targets type II fatty acid synthase acyl carrier protein (AcpP) in bacteria and has a homo-trimeric quaternary structure. However, the Sfp-type (surfactin phosphopantetheinyl transferase) can load PPant arm onto a peptidyl carrier protein (PCP) of type I non-ribosomal peptide synthetases (NRPSs) and polyketide synthase (PKS). The Sfp-type has a pseudo-homodimeric structure and functions on a much broader substrate scope compared to the AcpS-type.

Previously, the search for the minimum AcpP sequence recognized by PPTase led to the discovery of 11-mer peptide ybbR as a CP surrogate. Phage display was used to find peptide substrates and showed that they can be modified by both PPTase subtypes. These small peptides can be appended to proteins at the C- or N-terminus and utilized as a chemical biology tool for site-selective protein modification by PPTases. With the two PPTase subtypes having different substrate scopes, we sought to search for small peptide substrates that can be orthogonally labeled by either Sfp or AcpS using machine learning.

We developed a hybrid computational and biochemical method called Peptide Optimization with Optimal Learning (POOL) to discover orthogonal peptide substrates for the two PPTase subtypes. This method employs an iterative machine learning process where empirical data is accumulated into a mathematical algorithm that selects potential peptide substrates to be tested experimentally. Using a high-throughput SPOT cellulose membrane screening system, we were able to identify ‘hit’ peptides that are orthogonally labeled by either Sfp or AcpS-type PPTases with fluorescent CoA. These ‘hit’ peptides were fed back into the algorithm to refine further selections. (Figure 1.2)
Figure 1.2 Peptide Optimization with Optimal Learning (POOL) method outline. The algorithm recommends peptide substrates based on known/confirmed peptide sequences that were put into the method. These recommendations were synthesized on a high-throughput membrane peptide array and tested for labeling with PPTases and fluorescent CoA. The fluorescently labeled “hit peptides” were then fed back into the algorithm and it goes through a couple of iterative cycles for better recommendations.

Three Sfp-type and four AcpS-type peptides top-ranked by POOL after four rounds of iterations were aligned with natural substrates to understand what induces the selectivity for one type of PPTase but not the other. The number of amino acids to be predicted was not limited nor was the position of the modification site, serine residue, fixed. We included *Bacillus subtilis* PCP (a known substrate for Sfp), *Streptomyces coelicolor* acyl carrier protein (ACP), and *E. coli* AcpP (both known substrates for AcpS). The ybbR peptide, a substrate for both Sfp and AcpS, was also aligned. All Sfp-type peptides contained a hydrophobic residue (either Ile or Leu) at positions 2 and 5 relative to the serine residue. This trend is consistent with the recent crystal structure of PCP bound to Sfp where the hydrophobic residues on PCP form favorable contacts with a nonpolar patch on the Sfp
surface. In contrast, AcpS-type peptides contained polar residues at positions 2 and 5, which can disrupt interactions important for binding to Sfp, thereby giving orthogonal selectivity. (Figure 1.3) Hence, the POOL method showed great ability to predict diverse substrates.

![Sequence alignment between top recommended peptides and native PPTase substrates.](image)

Figure 1.3 Sequence alignment between top recommended peptides and native PPTase substrates.

With these recommended peptides, we made a green fluorescent protein (GFP)-fusion peptide tag to validate peptide hits in vitro. Each peptide was fused at the C-terminal end of the super-folded GFP and was fluorescently labeled with tetramethylrhodamine (TAMRA)-CoA by either Sfp or AcpS. As predicted by the POOL algorithm, Sfp-specific peptides were only labeled by Sfp but not by AcpS. However, we did not see any labeling activity by AcpS nor Sfp for AcpS-specific peptides. We speculate that AcpS-specific peptides get modified by endogenous E. coli AcpS and CoA when GFP-peptide fusion proteins were expressed. We confirmed this hypothesis by hydrolyzing the PPant arm with a base and verifying the mass with liquid chromatography-mass-spectroscopy (LCMS).

This proof of concept experiment searches for both active and orthogonal peptide substrates for post-translational modification enzymes. The POOL method demonstrates
that machine learning can be leveraged to guide peptide optimization for specific biochemical functions not immediately accessible by biological screening techniques, such as phage display and random mutagenesis. We anticipate applying POOL to identify biologically active peptide substrates important for epitope mapping, antigen discovery, and receptors that are therapeutically relevant.

Next, we utilized the function of PPTase in the search for complex biosynthetic gene clusters from uncultured microorganisms. We targeted finding CP-dependent metabolic pathways, such as PKS or NRPS, using a fluorescent pantetheine probe to label CPs with the endogenous PPTase in a marine tunicate. Previously, the Burkart Laboratory utilized 4-dimethylnaphthalene pantetheinamide (DMN-P; a solvatochromic fluorophore) to study protein-protein interactions between AcpP and its partner proteins in *E. coli* fatty acid synthase.\(^{14,15}\) The fluorescence of a solvatochromic fluorophore responds differently depending on the hydrophobicity of the environment. When DMN-P is loaded onto a CP, the probe is sequestered into its hydrophobic core, thereby prompting a substantial increase in fluorescence intensity. We envisioned when DMN-P is put into a culture, it would cross the cell membrane and label a CP *in vivo* by endogenous PPTase. (Figure 1.4)
Figure 1.4 The overview of enzyme mining strategy using fluorescent pantetheine analog and single-cell genomics. We used PPTase to label the protein with the fluorescent probe in the marine bacterial community. Using fluorescence-activated cell sorting, we can sort out labeled single-cell and sequence them to find its biosynthetic gene clusters. AntiSMASH predicts for secondary metabolite pathways to guide the enzyme mining.

Using the fluorescent labeling of a CP as a guide, we applied fluorescence-activated cell sorting (FACS) to select microbes from a tunicate microbiome. Single-cell and single-aggregate sequencing revealed enhanced biosynthetic diversity in the enzyme-active cells. Compared to the cells that were labeled and sorted with RedoxSensor Green (RSG; a marker of bacterial cell viability), DMN-P sorted cells showed much more diversity in phylogenetic composition and predicted secondary metabolite clusters. AntiSMASH\textsuperscript{16} was used to analyze the sequenced genomic information and predicted secondary metabolite biosynthesis gene clusters. We mined cognate CP - PPTase from the antiSMASH prediction of the same contig, expressed, and purified to validate CP labeling \textit{in vitro}.

We demonstrated an approach to identify and mine microorganisms with targeted biosynthetic capabilities. Single-cell genomics enabled the mining of a unique enzyme-substrate pair, thus showing activity-guided single-cell genomics as a tool to facilitate
enzyme discovery. Considering the central role that CP-PPTase pairs play in the biosynthesis of polyketides and NRPs, the predicted secondary metabolite diversity shows great promise in the mining of therapeutically relevant natural products.

The next part of the thesis is focused on developing crosslinkers to probe protein-protein interactions in these pathways. Specifically, we set out to evaluate the interaction between PCP and epimerization (E) domain in NRPSs. NRPSs are responsible for the syntheses of a wide variety of clinically significant bioactive natural products, such as antibiotics, immunosuppressants, and anticancer agents. These synthases are multimodular biosynthetic machinery where sets of catalytic domains are arranged into modules. Each module incorporates single amino acid as a building block and elongates to form a peptidyl chain intermediate. Amino acids are recognized, activated via the adenylation (A) domain, and loaded onto a PCP by forming thioester linkage to the PPant arm. PPant-tethered upstream and downstream amino acids form peptide bonds at the condensation (C) domain. Besides these essential A-PCP-C domains, modules also contain auxiliary domains to give chemical diversity to the NRPs by modifying the building blocks to D-amino acids, heterocycles, or N-methylated side chains. (Figure 1.5)
One of the characteristics of NRPs is the presence of D-amino acids, which gives them unique conformations for bioactivities. D-amino acids can be directly recognized and incorporated by an A domain, as in the case of cyclosporin A biosynthesis. However, most of the D-amino acids are included as L-amino acids and converted to D-amino acids by E domain activity. The E domain only acts on amino acids that are tethered to PCP by the PPant arm and converts amino acids in both directions (from L- to D-amino acids and vice versa) until equilibrium is reached. Mutational studies have shown that epimerization occurs through two catalytic residues, histidine and glutamate, by deprotonating/reprotonating an alpha proton (Cα-H) of the substrate. Although the
function of the E domain has been studied extensively through mutational experiments and excised modules, the exact mechanistic process of the catalysis is still unclear.

Two crystal structures of the E domain have been published to date: the excised E domain (PDB: 2XHG) from the initiation module of tyrocidine synthetase (TycA) and the PCP-E didomain (PDB: 5ISX) from the initiation module of gramicidin synthetase (GrsA).\textsuperscript{21,22} There are two notable salt bridge pairs found on the PCP-E interface that provide anchor-like electrostatic interactions at the PCP-E didomain linker to help guide the PCP towards the E domain active site. Additionally, structure-based calculations for protonation states have revealed catalytic residue histidine is in a protonated state. This suggests that glutamate may act as a base to deprotonate C\textsubscript{α}-H and histidine may act as an acid to reprotonate in order to convert L- to D-amino acid.

Another way for nature to incorporate D-amino acid is through racemases, which convert L-amino acids to D-amino acids by using the cofactor pyridoxal 5’-phosphate (PLP).\textsuperscript{23} Once an L-amino acid is loaded onto the PLP, C\textsubscript{α}-H is deprotonated and reprotonated to form the D-amino acid. There is a mechanistic analogy between PLP-mediated racemase and the E domain in NRPSs whereby both enzymes require the substrate to be covalently tethered to an enzyme. In this way, an amino acid can be activated for C\textsubscript{α}-H to be deprotonated. Mechanism-based alanine racemase inhibitors, chlorovinylglycine (CVG) and β-fluorophenylalanine, have been previously reported.\textsuperscript{24,25} Due to the mechanistic similarity between a racemase and E domain, we sought to apply alanine racemase inhibitors to E domain inhibition.

There is no direct way to measure E domain activity due to the fast reaction timescale.\textsuperscript{20} Stachelhaus \textit{et al.} devised an indirect assay to measure the E domain activity
with GrsA (A-PCP-E) and TycB₁ (C-A-PCP, the second excised module of tyrocidine synthetase). The E domain is active on both L- and D-amino acids, but the C domain functions as a gatekeeper in accepting the amino acid with the correct stereocenter to form a peptide bond with a downstream amino acid. The C domain of TycB₁ only allows D-Phe - the outcome of GrsA E domain activity - to form the dipeptide, D-Phe-L-Pro. The dipeptide product gets spontaneously cyclized and hydrolyzed to form D-Phe-L-Pro diketopiperazine (DKP). The assay measures D-Phe-L-Pro DKP formation to indirectly confirm E domain activity. (Figure 1.6)

Alanine racemase inhibitors, CVG and β-fluorophenylalanine, were preincubated with GrsA to inhibit E domain activity. High-performance liquid chromatography (HPLC) analysis of %DKP formation showed that there was no E domain inhibition by β-fluorophenylalanine. We speculated that β-fluorophenylalanine is too similar in structure to
the natural substrate, L-Phe, and is taken up as a substrate by the E domain. On the other hand, CVG revealed 50% DKP formation but did not achieve full E domain inhibition even with prolonged incubation time. To interrogate the activity of CVG as an E domain inhibitor, we performed molecular dynamics (MD) simulations to better understand how CVG works as a substrate/inhibitor. The substrate dihedral angles in the PCP-E didomain were analyzed with L-Phe and CVG. Due to its small size, CVG was able to sample a broader range of dihedral angles compared to L-Phe. Based on the MD simulations study, we propose the mechanism of how CVG works as an inhibitor towards the E domain.

(Figure 1.7)

Figure 1.7 Proposed mechanism of chlorovinylglycine (CVG) inhibition towards epimerization domain of GrsA. CVG gets recognized, activated by A domain and loaded onto PCP. Cα-H is deprotonated by Glu882, which kicks out chlorine as a leaving group to form allene intermediate. The intermediate is then attacked by His743 to form covalent bond, inhibiting epimerization domain active site.

To design a better crosslinker towards the E domain, we applied ligand-directed tosyl chemistry (published by Hamachi et al. in 2009).\textsuperscript{28} Mesylate and benzene sulfonate warheads were synthesized as a pantetheinamide probe, which can be loaded onto the PCP domain of TycA. After the probe is loaded, the PPant arm directs the warhead towards the E domain catalytic pocket so that catalytic residues can attack the warhead, thereby
intermolecularly crosslinking PCP and the E domain. The crosslinking activity was confirmed by running the sample on a SDS-PAGE denaturing gel, on which the crosslinking band ran higher than apo-TycA. (Figure 1.8) To further validate the crosslinking site, we compared protease digestion analyzed via mass spectrometry between apo-TycA and crosslinked TycA. In addition, a mutational study was performed to reveal that the crosslinking activity happens on both of the catalytic residues (His743 and Glu882) of the TycA E domain. (Figure 1.9) This finding shows that Glu882 acts as a base/nucleophile more so than His743. However, since the E domain converts both L- and D-amino acids, both catalytic residues function as both acid and base when performing the reaction.

![Figure 1.8 Epimerization domain inhibitors containing sulfonate moiety and its crosslinking validation. (A) The inhibitors were synthesized with C5 linker to mimic the length of the natural substrate, L-Phe, so that the catalytic residues can attack on the same site. (colored in blue) (B) The gel shows that the crosslinking sample runs higher compared to apo protein in the presence of Sfp, which validates the crosslinking complex.](image)
Figure 1.9 Mutational study showing the crosslinking site on catalytic residues of E domain. The crosslinking happens on both of the catalytic residues, H743 and E882, with C5-benzenesulfonate-PPant probe. Red box figure shows crosslinking happening on both of the catalytic residues on wild-type TycA. Blue box figure shows crosslinking happening on E882 with H743A mutant, whereas green box shows crosslinking happening on H743A with E882A mutant. When both of the catalytic residues are mutated to alanine (HE), there is no crosslinking activity.

Developing a crosslinker for the E domain sets the stage for the next step in studying the molecular basis of the E domain in its intact module. Furthermore, these findings will serve as a starting point in further combinatorial biosynthetic endeavors to incorporate D-amino acids into novel compounds.

References


Chapter 2. Discovering de novo peptide substrates for enzymes using machine learning

Section 2.1 Introduction

Machine learning has garnered increased attention in recent years for success in applications ranging from internet commerce to autonomous vehicles\(^1\)-\(^5\) due in large part to advances in computing power and availability of data\(^6\). Machine learning methods enable the best selection among a set of diverse options\(^7,8\) by predicting their quality. They can also identify which additional experimental data would best improve prediction\(^9\). This process affords a more efficient and economical approach than unguided experimental evaluation. While application of machine learning towards the optimization and discovery of biochemical systems promises success similar to that observed in business and engineering design problems\(^9\)-\(^11\), there have been relatively few biochemical applications reported for optimizing multiple parameters in parallel\(^12\)-\(^15\). Here, we highlight a machine learning method that enables the discovery of selective substrates for a set of functionally related enzymes, where little orthogonality exists in nature.

Previous biochemical applications of machine learning used the pure exploitation approach\(^16\), in which confirmatory assays experimentally evaluate only those options with the best predicted performance\(^12,17,18\). More sophisticated Bayesian optimization\(^9,10\) and optimal learning\(^16\) approaches to designing assays given a machine learning-based predictive model have been shown to better assist in optimizing systems using a limited number of experimental evaluations\(^9,19,20\). These approaches experimentally evaluate options with the highest potential to improve over the best previously evaluated option\(^21\).
While previous work proposes Bayesian optimization for biological molecular design applications\textsuperscript{13,15,22}, there is a current interest for implementing machine learning toward macromolecular optimization and small molecule discovery\textsuperscript{12-15}.

Using this optimal learning approach, we developed a method entitled Peptide Optimization with Optimal Learning (POOL) to identify short (8-20 residues) peptides as selective substrates for enzymes. To validate our POOL method, we utilized a post-translational modifying enzyme, the 4’-phosphopantetheinyl transferase (PPTase)\textsuperscript{23}. PPTases covalently modify carrier proteins (CPs) involved in various biosynthetic pathways at a conserved serine residue by the addition of phosphopantetheine derived from coenzyme A (CoA) (Figure S2.1 A-C)\textsuperscript{23}. Previous work on PPTase using phage display led to the discovery of the 11-residue peptide, YbbR, that can act as a surrogate for the full-length CP and be used as a short peptide tag for such applications as biochemical protein labeling and affinity purification\textsuperscript{24}.

Here, we used POOL to guide our experimental high-throughput cellulose SPOT synthesis screening array\textsuperscript{25} to identify short peptides that meet the following criteria: 1) enzymatic activity by PPTases by covalent attachment (labeling) of CoA analog to peptide substrate; 2) orthogonal, in which a given peptide is a substrate for one class of PPTase, but not the other; these PPTase classes include the Sfp (surfactin phosphopantetheinyl transferase from \textit{B. subtilis}) and AcpS-type (\textit{holo}-acyl carrier protein synthase from \textit{S. coelicolor}) (Figure S2.1B). Due to their structural difference and nature of interaction between the various carrier proteins, the pseudodimer, Sfp, is known to be promiscuous toward a variety of carrier proteins while the homotrimer, AcpS, is more selective\textsuperscript{23}. In this study, we demonstrate how POOL addresses complex biochemical challenges that are
difficult or impossible to solve by conventional methods to identify active orthogonal peptide substrates for multiple post-translational modification enzymes.

We demonstrate a computationally-driven machine-learning system, POOL, that guides the evolution of optimized orthogonal peptide substrates by alternating between prediction and targeted experimentation. This targeted approach departs from methods that randomly generate and experimentally screen many peptides, and can optimize for multiple complex biochemical activities, such as the ability to be selective for specific enzymes and undergo chemical transformations that would normally require several tandem screens to achieve. We show that POOL efficiently identifies both active and orthogonal peptide substrates for post-translational modification enzymes that are uniquely diverse from its original training set. POOL increases its chance of identifying active substrates by: (1) combining information across enzymes with a predictive model; (2) diversifying selections against prediction uncertainty, using ideas from Bayesian optimization; (3) incorporating feedback iteratively. POOL can be applied generally to optimize peptides for specific and/or simultaneous biochemical activity.

Section 2.2 Results and Discussion

Hits identified by POOL

Truncated portions of the acyl carrier protein and peptides known to be active for either Sfp or AcpS\textsuperscript{24,26} along with additional truncated proteins and peptides known to be inactive were initially used to train the POOL model to generate predicted short (8-20 amino acid) peptide substrates. The POOL model produces a single top-scoring short peptide from this initial data. To decrease statistical overfitting, the natural 20 amino acid
residues were organized into 8 reduced residue classes by their chemical properties (Figure S2.2). Prior to any experimental confirmation, the algorithm calculates a new prediction model based on the same set of initial experimental data used to generate this top-scoring peptide, but with the assumption that the top-scoring peptide is not a hit; that is, it lacks the desired pattern of orthogonal activity. POOL subsequently selects a second top-scoring peptide based on this new model. Then, a third model is produced based on the second experimental data, with an assumption that neither of the first two peptides are hits and another top-scoring peptide from this third model is chosen. This process iterates so that each successive peptide is chosen based on a model that uses the existing experimental data and the assumption that all previously chosen peptides in the present set are not hits, until the desired number of peptides to be tested is reached (Figure 2.1). The assumption that previously recommended peptides are not hits in these successive models captures the real-world possibility that the predictions of any given model are not valid and causes subsequent newly recommended peptides to be different from previously recommended ones, diversifying the set of peptides tested while still including those likely to be a hit.

POOL was designed to find short hits in a small number of experimental testing rounds to satisfy our criteria. Experimental confirmation of predicted peptides was conducted by the synthesis of membrane arrays containing 600 peptides. The peptides were synthesized and displayed on a cellulose membrane and then screened for enzymatic activity. These SPOT membranes were evaluated for orthogonal labeling activity by using fluorescent CoA\(^{27}\) to identify peptides labeled by two different PPTases (Sfp and AcpS) (Figure 2.1). After one round, we identified peptides that were both hits and misses and inputted these to the POOL algorithm for analysis. The POOL algorithm then
recommended a new set of peptides to test in the subsequent round. This process was repeated for four rounds to identify *de novo* peptide substrates.

Figure 2.1 Overview of the iterative Peptide Optimization with Optimal Learning (POOL) method workflow. Top cycle illustrates the autonomous computer algorithm coupled with experimental testing of peptide recommendations. [1] An initial set of peptides from known substrates of PPTase (Section 2.6 Methods 1) are entered in the algorithm defined in the box below. [2] A set of peptides are recommended. [3] Next, peptides are synthesized on a cellulose membrane where peptides are [4] chemoenzymatically labeled by PPTase (Sfp or AcpS) (hit peptide labeled in pink) (PAP is 3',5'-phosphoadenosine phosphate). [5] Experimentally confirmed peptide hits are fed back into the algorithm and the process repeats. The details of the algorithm used to recommend peptides are described in Section 2.6, Methods 2.
Initial rounds of Sfp labeling yielded significantly more peptide substrates (hits) than AcpS, presumably because our preliminary experimental data came primarily from known protein targets of Sfp. Iterative cycles of experimental input from SPOT membrane treatments with orthogonal PPTases with POOL showed an increase in the number of peptide hits (i.e. orthogonal substrates for Sfp and AcpS, Figure 2.2 A-B) after several rounds.

Importantly, screening of the initial predicted Sfp-specific peptides garnered less than five hits, while the last round of Sfp peptides yielded a 10-fold increase in peptide hits (Figure 2.2A).
Figure 2.2 Specific peptide hits identified by POOL and model validation. The number of peptides with specific labeling for Sfp (purple bar graph) (A) and AcpS (cyan bar graph) (B) over four rounds. The receiver operating characteristic curve (ROC) plot for both Sfp (purple curve) (C) and AcpS (cyan curve) (D) selective labeling, which diagnoses POOL’s ability to predict hit peptide substrates. The dashed line is the baseline that indicates the performance of a hypothetical random model without prediction power.

**Model validation**

To evaluate the accuracy of the prediction model within POOL, receiver operating characteristic (ROC) curves (Figure 2.2 C-D), which are commonly used to evaluate predictive models for classification, were generated from peptide hits data. The dashed straight line in each plot is the baseline, indicating the performance of a hypothetical random model without prediction power. The further the solid curved line trends away from
the baseline toward the upper left corner, the higher the prediction accuracy exhibited by the model.

As expected, the trained prediction model within POOL provided predictions that were substantially better than random chance, but still imperfect. These imperfections are due to four factors. First, the reduced amino acid alphabet does not distinguish between amino acids in the same class. Second, POOL’s predictive model makes a conditional independence assumption. That is, that an amino acid’s contribution to activity at a position does not depend on the amino acids at other positions. This reduces the amount of data required to make predictions but introduces statistical bias. Third, predictive performance is evaluated on peptides recommended by POOL, which were selected in part because they were likely to be active. Ordering these by activity level is substantially more challenging than it would be for peptides with a wider range of predicted activity. Fourth and finally, the quantity of initial experimental data is small relative to the complexity of the space of peptides over which prediction was performed. These factors are discussed in more detail in Section 2.6 (Methods 2B).

A traditional pure-exploitation approach for selecting peptides based on a predictive model with such imperfect accuracy would have had limited success because of sensitivity to prediction error during the discovery phase (Section 2.6, Methods 3B). Unlike this traditional approach, POOL accounts for prediction uncertainty and actively diversifies its selections to increase the chance of finding hits, as discussed below.
**POOL performance relative to other methods evaluated via simulation**

The POOL method first analyzes existing experimental data to produce an initial machine learning model to predict peptide substrates that are selective for each PPTase type. The POOL approach then chooses experiments to perform that are likely to reveal hits, even when a lack of training data creates less than perfect predictive accuracy within an iterative scheme. Unlike a pure-exploitation or predict-then-optimize method (Section 2.6, Methods 3A) which does not explicitly account for prediction error, the POOL approach contains a built in contingency plan to combat prediction errors in previously added peptides, while also broadening the diversity of potential peptide substrates (Figure 2.1). This mirrors the importance of balancing exploitation (selecting options predicted to work well) and exploration (selecting items with uncertain predictions) found in other applications of Bayesian optimization\(^1\).

To further confirm this aspect of POOL’s behavior and performance relative to other methods, it was compared via a simulation study with the mutation method and the predict-then-optimize method (Figure 2.3, Section 2.6, Methods 3A).

The mutation method takes an evolutionary approach and mutates known hits, while the predict-then-optimize method trains POOL’s machine learning model and recommends peptides predicted to be hits, but does not use POOL’s iterative approach of re-training the model with an artificial dataset. In this simulation study, a peptide’s probability of being a hit was determined via the prediction model trained using data gathered from multiple rounds (2602 peptides). The three methods were then provided with only the initial training set and data from the first two rounds. Then, we asked POOL to recommend 100 new potential Sfp-specific hits and 100 potential AcpS-specific hits. The probability that the
recommended set contained at least one peptide hit was then calculated via the trained model for each recommended set. POOL had a higher probability of finding at least one peptide hit in its recommended batch compared to the alternative prediction methods (Figure 2.3). This analysis suggests that iterative machine learning of peptide hits over several rounds balancing exploitation and exploration is advantageous as POOL identified multiple, unique orthogonally labeled peptides. The significantly higher probability of finding Sfp-specific hits compared to the other methods in the simulation study suggests that POOL outperformed the other methods for Sfp selectivity (Figure 2.3A). Due to the lack of AcpS-specific hits in the training data, all three methods had a low probability of finding AcpS-specific hits, though the POOL method displayed a significantly higher chance of predicting AcpS selective peptides than mutation and predict-then-optimize (Figure 2.3B).

**Diversity of peptides recommended in the simulation study**

The simulation study (Figure 2.3) suggests that POOL is able to find peptide hits more effectively than the predict-then-optimize and mutation methods because it is able to better balance exploration with exploitation. To further confirm this behavior, we investigated the sequences of the peptides recommended by the different prediction methods in the simulation study, and of the training data. Figure 2.3 C-D shows a 2-dimensional representation of peptides recommended by those methods, where each point in the plot represents a distinct peptide (Section 2.6, Methods 3C). A large distance between any two points indicates dissimilarity in sequence between the corresponding peptides. Recommended peptides by the predict-then-optimize method clustered together
tightly, suggesting that the top-ranking peptides are expected to share similar sequences. Both POOL and the mutation method recommended more diversified peptides than predict-then-optimize. To support comparison between POOL and mutation, Figure 2.3 C-D includes a line between each of their recommended peptides and the closest peptide in the training data. Lines originating from mutation's peptides tend to be shorter than from POOL's: mean lengths in Figure 2.3C are 1.63 for mutation and 2.55 for POOL, and in Figure 2.3D are 0.97 and 1.67 respectively. This is also confirmed by a histogram and additional summary statistics available in Section 2.6 (Figure S2.3). This suggests that the process of mutating known hits generates peptides that are more similar to the training data than those generated by POOL. In contrast with mutation, peptides recommended by POOL spread more broadly in 2D space, filling more gaps between peptides in the training data. In addition to this diverse exploratory behavior, POOL chooses at least one peptide in the same region where the predict-then-optimize method recommended a peptide. This pattern indicates that POOL provides a balance between the exploitation of model predictions and diversification accounting for prediction errors.
**Figure 2.3** Simulated method performance comparison and peptide diversity evaluation. A comparison between the POOL (red circle), Mutation (blue triangle), and Predict-then-optimize (pale orange diamond) methods for Sfp (A) and AcpS (B) based on a simulation of 100 peptides recommended after observing results from two rounds, and evaluated using four rounds of data. Peptide homology in peptide space for Sfp (C) and AcpS (D) is shown for each of the methods POOL (red circle), Mutation (blue triangle), Predict-then-optimize (pale orange diamond) and overlaid with the peptide training data (grey square). Distances between points represents how similar or dissimilar the corresponding peptides are from each other. A line from each peptide recommended by POOL and Mutation indicates the closest peptide in the training data.

**Selective and diverse peptide hits identified by POOL**

To identify sequence elements that appear to confer PPTase selectivity, 3 Sfp-type and 4 AcpS-type peptides identified by POOL, were aligned to known PPTase substrates. We included in this alignment the *B. subtilis* peptidyl carrier protein (PCP) from the tyrocidine pathway, a known Sfp substrate; *S. coelicolor* ACP and *E.coli* ACP, which are native AcpS-type CPs; and YbbR, a peptide substrate that is recognized by both PPTases (Figure 2.4)\(^{23,24,28}\).
All substrates recognized by Sfp contain a large hydrophobic residue, here either Ile or Leu, at positions 2 and 5, relative to the catalytic serine (position 1) (Figure 2.4). This motif is consistent with a recent crystal structure of a bound complex of PCP with Sfp\textsuperscript{28}, where residues 46 and 49 of the PCP, which correspond to positions 2 and 5 here, are Leu and Met, respectively. These residues, located in the helix of the PCP, form favorable contacts with the extended nonpolar patch on Sfp's catalytic domain\textsuperscript{28}. Interestingly, mutation of Met49 in PCP to a polar residue impairs catalytic activity and binding to Sfp\textsuperscript{28}. In contrast, the AcpS-type peptides do not show a clear motif involving positions 2 and 5, with Leu, Ser and Val at position 2, and Gln, Arg, and Val, at position 5. These residues presumably disrupt interactions important for binding to Sfp, and thus help account for the AcpS selectivity of these peptides. Perhaps the most consistent feature of these four AcpS substrates is that they have a small hydroxyl residue (Ser or Thr) at position 4 (also present in one of the Sfp substrates).

Sfp is known to be promiscuous toward many CPs from various organisms and biosynthetic pathways, including CP substrates of AcpS\textsuperscript{21}. Additionally, CP substrates that are labeled by AcpS, and not Sfp, are rare. POOL successfully identified 4 AcpS-specific peptides that were not recognized by Sfp, which when compared to the native CP substrates possess highly divergent sequences. This highlights the advantage of POOL over traditional genetic and combinatorial methods. These sequences would not be typically accessible by traditional genetic and combinatorial methods as they are not designed to simultaneously access for orthogonality and diversity, while also employing rational design.
Figure 2.4 Sequence alignment of hit peptides relative to native PPTase substrates. The sequence alignment for *B. subtilis* PCP, *S. coelicolor* ACP, *E. coli* ACP, YbbR peptide, compared to Sfp-type peptide hits (4P28, 4N28, 4F01) and AcpS-type peptide hits (1F01, 1I04, 3K17, 4T25) to the secondary (α2) structure of *B. subtilis* PCP (PDB: 4MRT). The blue box and red residues show general conserved sequences across all the peptides. The majority of AcpS-type peptides have conserved polar residues in position 2 and 5 (highlighted in yellow). The peptide identification corresponds to the round number and location it was identified from (round number_letter row_spot number on membrane) during the iterative rounds of POOL.

**In vitro** validation of peptide hits with GFP peptide tags

Hit peptides were chosen after normalization (Section 2.6, Methods 4 A-B) of fluorescence intensities on the SPOT array membrane after treatment with each of the two PPTases. Peptides were selected as hits that displayed orthogonal characteristics, which exhibited a high fluorescence intensity when labeled by one PPTase and low intensity when treated with the other.

Previous studies demonstrated that YbbR can be genetically encoded at the *N*- and *C*-termini of eGFP. We sought to validate our hit peptide substrates identified on the membrane in order to confirm that no labeling artifacts were chosen as hits from the SPOT cellulose membrane. Hit peptides were encoded at the *C*-terminal end of super-folded GFP. Recombinant purified GFP-peptide tags were incubated individually with both Sfp
and AcpS and TAMRA CoA. As predicted by our POOL program, the Sfp specific GFP-peptides were only labeled by Sfp, and not by AcpS (Figure 2.5, Figure S2.4).

![Diagram](image)

**Figure 2.5** Validation of PPTase specific peptide hits using GFP peptide tags. (A) Modification of GFP peptide fusions by addition of TAMRA-CoA onto selective peptides appended on the C-terminus of GFP. (B) 12% SDS-PAGE gel corresponding to labeling ($\lambda_{ex} = 532$ nm/ $\lambda_{em} = 580$ nm) of GFP-peptide tag by Sfp and AcpS. PPTase (Sfp and AcpS) is not observed due to low abundance (Coomassie sensitivity). TAMRA-CoA was present in each enzymatic reaction.

We found that we were unable to label our AcpS type GFP-peptides with our TAMRA CoA analog after purification from *E. coli*. We hypothesized that these GFP peptides were modified by endogenous CoA and AcpS in *E. coli*. To test whether our AcpS type GFP-peptides were labeled by *E. coli* AcpS, *in cellulo* labeling with our previously published pantetheinamide coumarin probe confirmed labeling by characterization by
tandem mass spectrometry. Ions characteristic of 4’-phosphopantetheine and its coumarin CoA were observed (Figure S2.5)\textsuperscript{30,31}.

To evaluate the catalytic efficiency of the Sfp-type GFP peptides, a kinetic analysis was conducted (Table S2.1). The catalytic efficiency for labeling the various GFP-peptide substrates was comparable to that of the previously published YbbR-GFP\textsuperscript{24}.

Unlike previous phage display studies that identified similar peptide sequences, POOL was able to specifically hone in on important residues adjacent to the catalytic serine that increased the diversity of peptides recognized by Sfp and AcpS. POOL was trained to pick hit peptides, not necessarily catalytically efficient peptides. Future studies include incorporating kinetic data into the algorithm to be employed by an optimization strategy in subsequent rounds of POOL.

**Demonstrating orthogonal peptides on a membrane support**

To demonstrate peptide specificity for each of the two PPTases, the top lead peptide for each PPTase was written on the surface of two duplicate membranes as two letters, “A” and “S”. The circle on the left-hand side (A) is specific to AcpS labeling, while the circle on the right-hand side (S) is specific to Sfp-type (Figure 2.6). The peptide AVKMESLEYLDTM (4F01), specific to Sfp, was synthesized on the cellulose membrane to highlight “S” (imaged on the right). Conversely, peptide WPEEGIESFMSVPPP (4T25), specific to AcpS was synthesized on the cellulose membrane highlighting “A” (imaged on the left). All unfilled areas inside the circle were prepared with IHDGADSVVWLWSNC peptide; a sequence that is neither a substrate for Sfp nor AcpS. Duplicate membranes synthesized with the orthogonal and unspecific peptides were treated with TAMRA-CoA.
and either AcpS or Sfp. After treatment by AcpS with TAMRA-CoA, the peptides forming an “A” pattern were illuminated, while the duplicate membrane treated with Sfp illuminated the “S” peptides (Figure 2.6). In all treatments, the non-specific peptides synthesized in all unfilled areas within the circles were neither labeled by Sfp nor AcpS as expected. These results demonstrate the selectivity for each of these enzyme isoforms for the newly discovered peptides via POOL.
Figure 2.6 Scheme illustrating selective Sfp and AcpS labeling on duplicate SPOT membranes. Peptide AVKMESLEYLDTM (4F01), specific to Sfp, was synthesized on the cellulose membrane to highlight “S”, while peptide WPEEGIESFMSVPPP (4T25), specific to AcpS, highlighting “A”. All unfilled areas inside the circle were synthesized with a non-specific peptide, IHDGADSVWLWSNC, which was neither a substrate for Sfp nor AcpS.
Section 2.3 Conclusions

Here, we demonstrate POOL’s rapid convergence on biomolecules with a desired activity. Unlike classical selection-based approaches, such as phage display and mutagenesis, POOL is efficient at simultaneously (1) searching for orthogonal activity; (2) exploring diverse peptide scaffolds beyond known substrates; (3) and judiciously selecting small sets of peptides rather than screening large, randomly generated libraries. Coupled with the SPOT technology, POOL offers a mathematically driven approach for discovery of structurally diverse orthogonal peptide substrates that are selective for multiple enzymes catalyzing the same enzymatic reaction.

POOL enabled the discovery of short, selective oligopeptide substrates for Sfp and AcpS. While POOL was utilized herein for PPTases and the identification of peptide substrates for these enzymes, our approach is generalizable and can be applied for the discovery of peptide substrates of other enzymes, such as kinases, proteases, and glycosyltransferases, which are therapeutically relevant enzyme targets\(^{32-34}\). A key advantage of POOL over traditional peptide substrate discovery methods is the ability to screen for multiple biological properties simultaneously. Here, we demonstrated the discovery of orthogonal peptides that are specific for different classes of PPTases. Furthermore, the orthogonal peptides differ significantly from each other and from their initial training set (Figure 2.3 C-D), which includes natural substrates.\(^{26}\) That is, when compared to the mutation and predict-then-optimize methods, POOL enables an enhanced departure in peptide sequence space from the nearest training set, underscoring POOL’s ability to generate a wide diversity of evolved peptide substrates. Previously discovered orthogonal peptides maintained high sequence conservation toward the N-terminus when
compared to the natural CP substrate and a constant length of 12 amino acids.\textsuperscript{26} The orthogonal peptides found in POOL are beyond the initial peptide training set and differ in length (8-20 residues), position of the catalytic serine residue, and chemistry of residues surrounding the catalytic serine.

POOL is a powerful method that can be readily extended to a broad class of biochemical applications because the underlying machine learning model is a flexible classifier that can predict other properties from training data. By defining a hit as any biochemical property or outcome of interest in parallel with experimental data, POOL’s predictive model can be trained to guide the discovery process towards any application. Additionally, POOL is a flexible method that can be adapted beyond Naïve Bayes to a wide range of complementary machine learning models, while still maintaining its theoretically guaranteed performance (Section 2.6, Methods 2E). We envision applying and adapting POOL to identify biologically active peptide substrates important for epitope mapping, antigen discovery, receptors and natural ligands that are therapeutically relevant. Additionally, POOL can be applied toward identifying peptide substrates important for characterizing enzymatic functions, while simultaneously screening for multiple biochemical activities.

\textbf{Section 2.4 Methods}

\textit{Bayesian machine learning and POOL process}

The prediction model used in POOL is a Bayesian variant of Naïve Bayes, modified to incorporate the belief that amino acids far from the central serine at which modification occurs have less impact on activity. Despite the word naïve in its name, Naïve Bayes is a
simple but powerful classification method widely used within machine learning for text classification. POOL then selects peptides to add iteratively by maximizing the conditional probability of being a hit, conditioned on previously added peptides in the batch not being hits. It performs this maximization across all peptides with a target length. Details of the method are provided in Section 2.6, Methods 1A.

Preparation of super-folded GFP-peptide constructs

Super-folded GFP\textsuperscript{29} gBlock gene fragment (IDT) was subcloned into the NdeI-XhoI site of a pET28b vector (Novagen) containing an N-terminal His\textsuperscript{6}-tag. The Sfp-type (4P28, 4N28, 4F01) and AcpS-type (1F01, 1I04, 3K17 & 4T25) peptides for the various GFP constructs were ordered as single-stranded oligonucleotides (forward and reverse) with 5'-phosphorylation (IDT) and its corresponding XhoI overhangs. To prepare the double-stranded oligonucleotides, the forward and reverse single-stranded oligonucleotides were mixed together in equal molar amounts in 100 mM potassium acetate; 30 mM HEPES, pH 7.5. The equimolar forward and reverse oligos were heated to 94 °C for 2 minutes and slowly cooled by decreasing the temperature 1 °C every 30 seconds to a final temperature of 4 °C. The double-stranded oligomer was ligated into the corresponding XhoI (NEB) cut vector. These plasmids were transformed into an \textit{E. coli} BL21 \textit{ΔentD} strain, which was a gift from Professor D.F. Ackerley\textsuperscript{35}.

Recombinant protein purification

PPTases, \textit{B. subtilis} Sfp, \textit{S. coelicolor} AcpS, and GFP peptide proteins were expressed and purified\textsuperscript{29,36}. Cells were grown at 37 °C in terrific broth media containing 100 mg L\textsuperscript{-1} kanamycin sulfate to an OD\textsubscript{600} of 0.8. The cells were induced with a final concentration of 250 mM of IPTG and grown overnight at 16 °C. The cells were harvested
by centrifugation at 2,000 g for 30 minutes, resuspended in lysis buffer (50 mM Tris/HCl buffer pH 7.5 with 250 mM NaCl), and supplemented with 0.1 mg/mL lysozyme (Worthington Biochemical Corp), 5 µg mL⁻¹ DNAse I (Sigma), and 5 µg mL⁻¹ RNAse (Worthington Biochemical Corp.). The cells were lysed by a French pressure cell press between 500-1,000 psi. The lysate was subsequently centrifuged at 12,000 g for 45 minutes, and the supernatant was bound with Ni-NTA resin (Qiagen). The column was sequentially washed with the wash buffer (50 mM Tris/HCl, buffer pH 7.5, 250 mM NaCl, 25% glycerol, 10 mM buffered imidazole), followed by elution at 300 mM buffered imidazole. The purified protein was desalted into 50 mM Tris/HCl, buffer pH 7.5 with a PD-10 desalting column (GE Healthcare Life Sciences) and concentrated with a 10-kDa Amicon spin filter (Millipore Sigma). The concentrated sample was subsequently stored in 20% glycerol, 50 mM Tris/HCl buffer pH 7.4 with 150 mM NaCl at -80 °C after flash freezing in liquid nitrogen.

**SPOT synthesis of cellulose membrane peptides**

The peptides were synthesized on an amino-PEG-cellulose 10 x 15 cm membrane and synthesized automatically by MultiPep RSi (INTAVIS Bioanalytical Instruments AG). After Fmoc cleavage with 20% 4-methyl piperidine (Sigma) in dimethylformamide (DMF), activated with 1 equiv. hydroxybenzotriazole (HOBt) and 2 equiv. N-methylpyrrolidone (NMP), and directly spotted on the membrane; after 15 min this step was repeated. The membrane was washed with DMF (3 × 3 min). Solutions of the Fmoc amino acids in NMP were spotted on the membrane (0.6 M solutions; 0.8 M solutions for C, H, N, Q, and R; triple coupling, 15 min each). The Fmoc group was removed from the spots, and the
sequences of the peptides were completed using the standard SPOT synthesis protocol and followed by a N-terminal tag. Membranes were stored at -20 °C until treated.

Membrane treatment

Synthesized membranes were deprotected with cleavage cocktail consisting of 94% TFA, 2.5% water, 1% triisopropylsilane (Sigma), and 2.5% 1,2-ethanedithiol (Sigma) shaking at 2 hours at room temperature. The membranes were washed in triplicate with dichloromethane (Sigma), ethanol (Sigma), distilled deionized water, and Tween 20- Tris-buffered saline (TBST) (20 mM Tris/HCl buffer pH 7.4 with 150 mM NaCl) respectively. The membrane was incubated with 5% bovine serum albumin (BSA) in TBST for 1 hour at room temperature. Following blocking by BSA, the membranes were washed three times with TBST. The membranes were subjected to labeling by their respective PPTase, either B. subtilis Sfp or S. coelicolor AcpS. The final 20 mL reaction volume consisted of 10 µM PPTase enzyme, 40 µM TAMRA CoA, 0.01% Triton-X, and 15 µM BSA in 50 mM Tris/HCl buffer pH 8.0 and 10 mM MgCl2. The reaction was incubated for 16 hours at 37 °C. The membranes were rinsed liberally with TBST buffer, 0.1% TFA, and distilled deionized water before imaging on a Typhoon TRIO Variable Mode Imager (GE Healthcare BioSciences) at 50-µm resolution with 532 nm green laser excitation and 580 nm emission filter and a photomultiplier tube setting of 350 V. The membrane images were analyzed by ImageJ software37,38.

In vitro validation of peptide hits with GFP peptide tags

Three Sfp-type (4P28, 4N28, 4F01) and four AcpS-type (1F01, 1I04, 3K17, 4T25) peptides were fused to the C-terminal end of super-folded GFP29. To access orthogonal labeling, these specific GFP-peptide tags were incubated with either Sfp or AcpS in the
presence of TAMRA-CoA (4). The reactions were conducted in a 30 µL reaction containing a final concentration of 10 µM of the respective GFP-peptide tag, 5 µM TAMRA-CoA (4), and 0.05 µM of either Sfp or AcpS in 50 mM HEPES, 10 mM MgCl₂, pH 7.2. The mixture was gently shaken at 37 °C overnight. The resulting reactions were analyzed by 12% SDS-PAGE gel and imaged on a Typhoon TRIO Variable Mode Imager (GE Healthcare BioSciences) at 50-µm resolution with 532 nm green laser excitation and 580 nm emission filter and a photomultiplier tube setting of 350 V.

**PPTase gel-based electrophoretic mobility shift kinetics**

A concentration-dependent assay varying concentrations of GFP-peptides was conducted to characterize the kinetics of all the GFP-peptides with Sfp. Briefly, reaction was initiated by the addition of 0.1 µM Sfp into a 30 µL reaction containing 1 mM TAMRA-CoA, buffer (50 mM HEPES, 10 mM MgCl₂, pH 7.2), and with varying concentrations of GFP-peptide (0.1 - 10 µM). The reactions were quenched every 5 minutes sequentially by adding 100 mM EDTA. Resulting reactions were analyzed by 12% SDS-PAGE gel and imaged from Typhoon (GE Healthcare) scanner (λ_ex = 532 nm/ λ_em = 580 nm). The intensity of the fluorescence gel bands were measured and analyzed using GraphPad Prism and fitted to the Michaelis-Menten equation, \( Y = V_{\text{max}} \times X / (K_m + X) \), where \( K_m \) is the Michaelis-Menten constant, \( X \) is the substrate concentration, \( V_{\text{max}} \) is the maximum enzyme velocity, and \( Y \) is the velocity of the enzyme. The data was collected in triplicate (Table S2.1).

**In cellulo labeling and LC-MS analysis of AcpS GFP peptides**

GFP peptides encoded in a pET28b plasmid in *E. coli* BL21 (ΔentD) cells were grown for 5 h at 37 °C in 5 mL LB medium containing 100 mg L⁻¹ kanamycin sulfate to an
OD$_{600}$ of 0.6 – 0.8 and induced with a final concentration of 500 µM IPTG. Along with the addition of IPTG, 1 mM of coumarin CoA was added into the growth medium to investigate uptake and labeling$^{27}$. The cultures were grown for 16 h at 16 °C. The cells were centrifuged and resuspended in lysis buffer (50 mM sodium phosphate, 300 mM NaCl, pH 7.5), and lysed by incubation with lysozyme (1 mg ml$^{-1}$) for 1 h on ice followed by sonication (3 x 10 second pulses) and centrifugation. 50 µL of protein lysate of GFP-peptide at a concentration of 5 µM was precipitated by addition of 1 mL trichloroacetic acid solution (10% weight per volume percent) and pelleted by centrifugation at 4 °C at 14000 rpm for 20 min. The pellet was washed with 1 mL water twice and centrifuged at 4 °C at 14000 rpm for 20 min after each washing step. Subsequently, 100 µL of 0.1 M KOH was added and the sample was heated for 30 min at 75 °C. Then the protein was precipitated with 10 µL trifluoroacetic acid (50% volume per volume percent) followed by centrifugation as before, and the supernatant was analyzed by LC-MS under single-ion detection mode. Hydrolyzed coumarin CoA (m/z=626.64) was detected$^{30}$.

*Sequence alignment*

Sequence alignment was conducted using the open online software ESPript 3.0$^{40}$. 

41
Section 2.5 Supplementary Information

Supplementary Methods

1. Use of POOL methodology for discovery of peptides for orthogonal labeling

The POOL model was initially trained by three categories of peptides to generate short orthogonal active peptide substrates. First, we obtained peptides by truncating portions of acyl carrier proteins from various organisms known to be substrates for the following enzymes: B. subtilis Sfp [24], S. coelicolor AcpS [38], and F. thioreicsus AcpH [41]. (AcpH is an enzyme known to unlabel some substrates previously labeled by a PPTase, either AcpS or Sfp.) Second, we included short peptide substrates that were previously identified as substrates for PPTases. Third, we included truncated acyl carrier proteins from various organisms and short peptides known not be substrates (inactive) for the previous three enzymes [41]. Next, the POOL algorithm interactively queried the experimentalists to test a group of desired peptides (usually 300–500 peptides) to identify active short hits. We incorporated the test results as new data points into the dataset to update the learning model. We repeated this process iteratively for 5 rounds.

In each round, the POOL algorithm is told the type or types of hits to discover, and it allocates the available membrane space among them. For example, in one round, POOL may be told to search for both Sfp hits (peptides that are active for Sfp, regardless of their activity with AcpS and AcpH) and AcpS hits (peptides that are active for AcpS), in which case it would choose an equal number of peptides to test for each. Other types of hits to which POOL may be applied include Sfp-specific hits (active for Sfp and not active for AcpS), AcpS-specific hits (active for AcpS and not active for Sfp), Sfp-specific hits with unlabeling (active for Sfp and not active for AcpS), and AcpS-specific hits with unlabeling (active for AcpS and not active for Sfp).

In the first and second rounds, POOL searched for Sfp hits with unlabeling, while in rounds three through five POOL searched for Sfp-specific hits with unlabeling and AcpS-specific hits with unlabeling. In the first and second rounds, Sfp labeling and AcpH unlabeling activity was measured. Then at the start of round three, AcpS labeling and AcpH unlabeling activity was measured for the peptides tested in round 1. Sfp labeling with AcpH unlabeling, and AcpS labeling with AcpH unlabeling were tested for rounds 3 through 5. Thus, measurements required for assessing specific labeling were conducted for four rounds in total.

We originally had the goal of discovering substrates that were labeled by a PPTase and unlabelled by AcpH, optionally with specific activity for one PPTase, and directed POOL toward this goal. Later, we chose to focus on the specific labeling portion of this goal, without the extra complication of unlabeling. While ongoing studies in our laboratories are looking at subsequent processing steps, adding the extra requirement that modifications could be removed by hydrolases, AcpH unlabeling is not relevant for this manuscript.

As discussed in the main text and in more detail in the next section, the POOL algorithm comprises two pieces (see Figure 1). The first uses a machine learning method, Naive Bayes [42], to predict the probability that a given peptide sequence will be a hit based upon previously collected training data. The second uses value of information analysis [43, 44], together with the probabilities calculated by Naive Bayes, to find a set of peptides to test next that, considered together as a group, have maximum probability of containing at least one short hit. Both simulations and an analysis of previously obtained experimental data demonstrate that using POOL dramatically reduces the number of experiments required to find short hits, as compared with a baseline method of randomly mutating and truncating previously known substrates.

2. Description of the POOL methodology

We first introduce mathematical notation that allows a more precise discussion of the POOL methodology. Let $E$ be the set of peptides over which we would like to search, which we take to be the set of peptides of length less than 38 amino acids. For any peptide sequence $e \in E$, let $y(e) \in \{0, 1\}$ be a binary label indicating whether peptide $e$ is a hit, where the definition of hit is specific to the goal at hand (e.g., discovering Sfp hits, or Sfp-specific hits). This label can be learned through experiment, but is otherwise unknown $a$ priori.

We let $f(e)$ be equal to negative one times the length of the peptide. The negative one is present because we will want $f(e)$ to be large, but we want the length to be small. We will sometimes refer to $f(e)$ generically as the "fitness".

Our goal, in terms of the notation we have defined, is to find a peptide $e \in E$ for which $y(e) = 1$ and for which $f(e)$ is as large as possible. We can write this problem as,

$$\max_{e \in E} f(e).$$

POOL chooses a set of peptides $S$ to test in a single round, to best support solving Eq. (1), in the specific sense of maximizing the probability of finding a hit whose fitness is better than some target fitness $b$. It consists of three steps: first, a prediction step, in which a machine learning method (Naive Bayes, customized for peptide activity) provides a joint probability distribution over peptide activity; second, a probability-optimization step, in which we use value-of-information analysis to define the quality of a set of peptides to test; and third, a greedy optimization step, in which we use a greedy algorithm to find a set of peptides to test that provides a large probability of improvement.
Below, Section A describes the machine learning model POOL uses to predict \( y(e) \); Section C defines the probability of improvement criterion that POOL seeks to maximize in the set of peptides it recommends testing; Section D, F, and G describes the greedy approach that POOL uses to approximately maximize the probability of improvement; Section E provides a theoretical guarantee on the solution quality from this greedy approach; and Section II discusses how POOL can be used when only partial labeling information is available for some peptides in the training set.

### A. Bayesian Machine Learning Model

POOL’s prediction step uses a modified Naive Bayes classifier within a Bayesian framework to provide prediction of a joint probability distribution over peptide activity. Naive Bayes classifiers [43, 44] are popular and extensively studied in text classification problems because of their fast computation and ease of implementation. As peptides are sequences of amino acids just like documents are sequences of words, peptide classification is similar to text classification, and similar methods can be used. Rather than using the classical Naive Bayes directly, we modify it by putting a Bayesian prior distribution on the model’s parameters. This supports higher-quality inference with small amounts of data.

We now explain the approach in detail.

Since the chemoenzymatic reaction happens at the conserved Serine, we are looking for the peptides containing exactly one Serine in the sequence. We use Serine as the origin, and label positions of other amino acids in the sequence by their sequence-distance from the origin in the C-terminus or N-terminus direction. Each position contains one feature, which is determined by the amino acid residing in this position. To reduce dimensionality of the feature space, we group the 20 standard amino acids into 8 classes (shown in Supplementary Figure 2) according to chemical type, and represent the amino acid by its class. For example, a sequence X-X-X-X-S-X-X, where S represents Serine and X represents the reduced amino acid alphabet, can be written in this feature representation as N3-N2-N1-S-C1-C2, where \( N \) and \( C \) are the groups of the amino acids in the corresponding position.

We use a \( J \)-dimensional feature vector to represent a peptide, written as \( x = (x_1, \ldots, x_J) \), and the feature space is \( X \). In our application, each feature corresponds to a position in the peptide, and the feature vector indicates which the amino acid class is present at that position. Then we can predict \( y(e) \) by calculating \( P(y(e) = 1 \mid x) \). Using Bayes’ theorem, we have

\[
P(y(e) = 1 \mid x) = \frac{P(x \mid y(e) = 1) P(y(e) = 1)}{\sum_{y' \in \{0, 1\}} P(x \mid y(e) = y') P(y(e) = y')}
\]

Since Naive Bayes assumes the features are conditionally independent given \( y \), we can write Eq. (2) as

\[
P(y(e) = 1 \mid x) = \frac{\prod_{j=1}^J P(x_j \mid y(e) = 1) P(y(e) = 1)}{\sum_{y' \in \{0, 1\}} \prod_{j=1}^J P(x_j \mid y(e) = y') P(y(e) = y')}
\]

If we knew \( P(x_j \mid y(e) = y') \) and \( P(y(e) = y') \) for \( y' \in \{0, 1\} \), we could use Eq. (3) to compute \( P(y(e) = 1 \mid x) \). Our next task is to estimate both quantities from data.

We will adopt a Bayesian viewpoint in performing this estimation. Toward this end, we augment our model by adding a set of unknown latent random variables \( \theta^{i,j} \) where \( j \) ranges across positions, \( y' \in \{0, 1\} \), and \( i \) ranges across the \( K \) distinct groups of amino acids available for selection at position \( j \). \( \theta^{i,j} \) will represent the frequency with which an amino acid in class \( j \) arises at position \( i \) among peptides with label \( y' \). We let \( \Theta^{y'} = \{ \theta^{y',1}, \ldots, \theta^{y',K} \} \) and \( \Theta \) be the matrix comprised of \( \Theta^{y'} \) with \( y' \in \{0, 1\} \) and \( j \) across the set of positions. In this Bayesian context, Eq. (2) and Eq. (3) become

\[
\begin{align*}
P(y(e) = 1 \mid x, \Theta) &= \prod_{j=1}^J \prod_{y' \in \{0, 1\}} P(x_j \mid y(e) = y', \Theta) P(y(e) = y' \mid \Theta) \\
\sum_{y' \in \{0, 1\}} \prod_{j=1}^J P(x_j \mid y(e) = y', \Theta) P(y(e) = y' \mid \Theta)
\end{align*}
\]

We then define \( P(x_j \mid y(e) = y', \Theta) = \epsilon_{y',j} \theta_{y',j}^{y'} \) and \( P(y(e) = 1 \mid \Theta) = P(y(e) = 1) \).

We further put a Dirichlet prior over \( \theta^{y'} \), that is, \( \theta^{y'} \sim \text{Dirichlet}(\alpha^{y'}) \). We set \( \alpha^{y'} \) to be a vector that grows with the square root as we move further from the converged species: \( \alpha^{y'} = \alpha^{y'} \sqrt{|y'|} \). A Dirichlet prior on \( \theta^{y'} \) becomes more concentrated around its mean \( \alpha^{y'}/\sum \alpha^{y'} \) as we increase \( \sum \alpha^{y'} \), and so this causes us to put more weight in our prior on the belief that \( \theta^{y'} \) is close to \( \alpha^{y'}/\sum \alpha^{y'} \) as we move away from the converged species. We choose \( \alpha^{y'} \) so that \( \alpha^{y'}/\sum \alpha^{y'} \) is identical for \( y = 0 \) and \( y = 1 \), and so as we move away from the serine we become more confident that \( \theta^{y'} \) is similar across \( y \), which corresponds to this position having less influence on whether the peptide is active or not.

If \( \alpha^{y'} \) is proportional to number of amino acids in class \( i \), then the mean value \( \alpha^{y'}/\sum \alpha^{y'} \) corresponds to the amino acid class proportions we would expect at a position that has no influence on activity. Also, it would be appropriate to set the magnitude of \( \sum \alpha^{y'} \) to be larger for \( y = 0 \) than for \( y = 1 \), since most peptides are inactive causing the distribution over amino acid classes among inactive peptides to more closely mimic this distribution at all positions. We realized the importance of class size in this analysis after conducting experiments: \( \alpha^{y'} \) used to recommend peptides in our experiments was instead proportional to the unit vector with a larger proportionality constant for \( y = 1 \).

With this Dirichlet prior, the posterior distribution of \( \theta^{y'} \) given training data \( \{(x_1^{(1)}, y_1^{(1)}), \ldots, (x_N^{(1)}, y_N^{(1)})\} \) is also a Dirichlet distribution, with updated parameters \( \alpha_{\text{new}}^{y'} = \alpha^{y'} + \sum_{n=1}^N (y_n^{(n)} - \epsilon_{y',j}^{(n)}) \) for \( j = 1, \ldots, K \).

To estimate the conditional probability that \( y(e) = 1 \) given data, we sample \( \theta \) from its posterior distribution for all \( j \) and \( y' \), calculate the conditional probability that \( y(e) = 1 \) given \( \theta \) using Eq. (4), and then average across samples. We can calculate a joint conditional distribution on \( y(e) \) for a set of \( e \) given data using a similar technique.

### B. Accuracy of the Bayesian Machine Learning Model

While Figure 2 C-D show that the predictions from POOL’s prediction step are substantially better than random chance, they are imperfect. This imperfection is due to four factors.

First, the reduced amino acid alphabet does not distinguish between amino acids in the same class.

Second, Naive Bayes assumes conditional independence, but instead it may be that an amino acid at one position is more effective at generating activity when an adjacent position has a particular amino acid present.
Third, because most of the training data used in prediction was generated by POOL’s past recommendations, it consists largely of peptides that were likely to be a hit according to the predictive model at the time they were selected. However, hits are rare, and so only a fraction of these are actually hits. Moreover, this generation process causes more of the peptides in the training data to have similar representations in feature space than if the data had been generated uniformly at random. These characteristics cause the prediction problem with results pictured in Figure 2C-D to be particularly difficult.

In contrast, if the training data had been generated at random, the misses would be better separated from the hits in the model’s feature space, and it would be easier for POOL’s predictive model to provide accurate classification.

Fourth, the function POOL’s predictive model is trying to learn (the mapping from peptides onto activity) has a number of degrees of freedom substantially larger than the amount of training data available. With only the 10 amino acids closest to the Serine residue one may create $10^{10}$ (approximately $10^{15}$) peptides. The mapping from peptides onto activity would predict, for each of these peptides, what its activity is. Representing an arbitrary function of this form requires $2^{10^{10}}$ bits. This is many orders of magnitude larger than the amount of information present in POOL’s training data (activity for 2000 peptides, considering all rounds).

While one could address the first and second issues above through a more complex model with more parameters, the predictive model POOL uses already contains $2^8 = 256 = 14$ parameters at each location (the population of each class in the distributions of hits and misses respectively, minus 1 for the need for these prevalence to sum to 1), giving $14^2 = 196$ parameters to estimate over 20 locations. Moreover, POOL must estimate these parameters while avoiding overfitting not just after the final round, but before the start of each round. At the start first round, there were only 40 peptides in our training data, and at the start of the second round only 500. For this reason, not enough data is present to allow accurate estimation of the additional parameters in a substantially more flexible model.

Thus, POOL uses a reduced amino acid alphabet and assumes conditional independence because the reduction in the number of parameters required offers substantial advantages when making predictions with relatively little data. While a more complex model could have better predictive accuracy after substantially more data was collected, it would be likely to have worse accuracy in the data-poor regimes where peptide discovery typically operates.

C. Probability of improvement. POOL uses the joint probability distribution over $p(e) : e \in E$ provided by Naive Bayes to define an auxiliary function, called probability of improvement, which quantifies the probability that a set of peptides $S$, if tested, will reveal an active peptide whose fitness $f(e)$ improves on some benchmark value $b$, which is typically the best $f(e)$ of active peptides observed in the past. We define this probability-of-improvement, $P(I(S))$, to be

$$P(I(S)) = P\left( \max_{e \in S, f(e) > 0} f(e) > b \right).$$

[Equation 5]

In the event that all peptides tested are inactive, we define max over an empty set to be $-\infty$. POOL seeks to set $S$ by approximately solving

$$\arg \max_{S \subseteq E, |S| \leq k} \log P(I(S)),$$

where $k$ is the number of peptides that can be tested simultaneously in a single round of experiment.

After we find the set $S$ and tested the peptides in it, we use the results along with previously tested peptides as training data, and find a new set $S$ for the next round of experiment. We repeat this process until we find enough favored active peptides or resource is exhausted.

D. Maximizing probability of improvement using the greedy algorithm. Directly solving Eq. (6) is computationally challenging, first because the cost of computing $P(I(S))$ grows exponentially in $k$, and second because the number of sets $S$ satisfying $|S| \leq k$ also grows exponentially in $k$. In this section we describe an easily computed approximate solution to Eq. (6). The method we employ is a greedy algorithm, which starts with $S = \emptyset$, and then adds peptides one-by-one, each time choosing the one that will increase the probability of improvement by the greatest amount, that is, solving

$$\arg \max_{e \in E, |S| < k} P(I(S \cup \{e\})).$$

until $|S| = k$. A theoretical result provided in [47] provides a performance guarantee on solution quality in the form of a lower bound on the ratio between the probability of improvement provided by the greedy approach and solving Eq. (5) exactly.

E. Performance guarantee for the greedy algorithm. The following theorem provides a guarantee on the probability of improvement provided by the greedy algorithm relative to that provided by solving Eq. (6) exactly.

**Theorem 1.** The greedy algorithm always produces a solution whose probability-of-improvement is at least $1 - (1 + 1/e)^k \geq 1 - 1/e$ times the optimal objective value of Eq. (6).

The proof of the theorem follows directly from the following two lemmas. The first lemma is shown in [47] while we provide a proof of the second.

**Lemma 1.** If $F(S)$ is submodular, non-decreasing and $F(\emptyset) = 0$, the greedy heuristic always produces a solution whose value is at least $1 - (1 + 1/e)^k \geq 1 - 1/e$ times the optimal value, where $|S| \leq k$. This bound can be achieved for each $k$ and has a limiting value of $1 - 1/e$, which is the base of the natural logarithm.

**Lemma 2.** $P(I(S))$ is submodular, non-decreasing and $P(I(\emptyset)) = 0$.

**Proof.** First we show $P(I(\emptyset)) = 0$.

$$P(I(\emptyset)) = P(\max_{e \in S, f(e) > 0} f(e) > b) = P(-\infty > b) = 0.$$

To show $P(I(S))$ is non-decreasing, let $A \subseteq B \subseteq E$ where $E$ is a finite set. Then

$$P(I(B)) = P(\max_{e \in B, f(e) > 0} f(e) > b) \geq P(\max_{e \in A, f(e) > 0} f(e) > b) \geq P(I(A)).$$

44
Lastly, we show \( \text{PI}(S) \) is submodular. For \( e \in E \setminus B \),
\[
\text{PI}(A \cup \{e\}) - \text{PI}(A)
= \mathbb{P}(f^*(A \cup \{e\}) > b) - \mathbb{P}(f^*(A) > b)
= \mathbb{P}(f^*(A \cup \{e\}) > b) - \mathbb{P}(f^*(A) > b) + \mathbb{P}(f^*(A \cup \{e\}) > b | f^*(A) > b) \mathbb{P}(f^*(A) > b)
= \mathbb{P}(f^*(A \cup \{e\}) > b | f^*(A) > b) \mathbb{P}(f^*(A) > b) - \mathbb{P}(f^*(A) > b)
= \mathbb{P}(f(e) > b, y(e) = 1 | f^*(A \cup \{e\}) > b) \mathbb{P}(f^*(A) > b) - \mathbb{P}(f(e) > b, y(e) = 1 | f^*(A) > b).
\]
Using a similar argument,
\[
\text{PI}(B \cup \{e\}) - \text{PI}(B)
= \mathbb{P}(f(e) > b, y(e) = 1, f^*(B) \leq b)
= \mathbb{P}(f(e) > b, y(e) = 1, f^*(A \cup \{e\}) \leq b, f^*(B \setminus \{e\}) \leq b).
\]
Therefore, \( \text{PI}(A \cup \{e\}) - \text{PI}(A) \geq \text{PI}(B \cup \{e\}) - \text{PI}(B) \). Thus \( \text{PI}(S) \) is submodular.

\(\square\)

**F. Simplifying the greedy algorithm.** Although Eq. (7) is easier to solve than Eq. (6), it is nevertheless computationally infeasible to solve Eq. (7) using a naive approach in which we simply enumerate all peptides \( e \in E \setminus S \) and calculate \( \text{PI}(S \cup \{e\}) \) for each. This naive approach is infeasible because \( E \) is very large and calculating \( \text{PI}(S \cup \{e\}) \) can take significant computational effort. To overcome these challenges we first transform Eq. (7) into an easy-to-compute objective.

**Proposition 1.** The solution to Eq. (7) is equal to
\[
\arg \max_{e \in E \setminus S, f(e) > b} \mathbb{P}(y(e) = 1 \mid y(e') = 0, \forall e' \in S). \tag{8}
\]

**Proof.** Let \( f^*(S) = \max_{e \in E \setminus S, f(e) > b} f(e) \). Then,
\[
\text{PI}(S \cup \{e\}) = \mathbb{P}(f^*(S \cup \{e\}) > b)
= \mathbb{P}(f^*(S) > b) + \mathbb{P}(f^*(S) \leq b | f(e) > b, y(e) = 1 | f^*(S) \leq b).
\]
Since \( f^*(S) \leq b \) is strictly positive and does not depend on \( e \), Eq. (7) is equal to
\[
\arg \max_{e \in E \setminus S} \mathbb{P}(f(e) > b, y(e) = 1 | f^*(S) \leq b). \tag{9}
\]
Moreover, when \( f(e) \leq b \), \( \mathbb{P}(f(e) > b, y(e) = 1 | f^*(S) \leq b) = 0 \). Thus the solution to Eq. (9) is equal to
\[
\arg \max_{e \in E \setminus S, f(e) > b} \mathbb{P}(y(e) = 1 \mid y(e') = 0, \forall e' \in S). \tag{10}
\]

\(\square\)

**G. Implementing the greedy algorithm.** Although Eq. (8) avoids the challenge of computing the probability of improvement, it is nevertheless hard to solve directly through enumeration, because \( E \) is large. Here we provide an efficient approximate implementation. (While the performance guarantee we show in Supplementary Section E does not apply to this approximate implementation, it nevertheless suggests that this approximation should perform well, which is also supported by simulation experiments shown in the main paper in Figures 3A and 3B.)

This approximate implementation relies on approximating the Bayesian posterior on \( \theta \) from our Naive Bayes predictive model by a point estimator \( \hat{\theta} \) of \( \theta \), such as the maximum a posteriori (MAP) estimator [48], or the mean of the posterior distribution. With any such point estimator our Naive Bayes predictive model becomes
\[
\mathbb{P}(y(e) = 1 \mid x) \approx \frac{\prod_{j=1}^{J} \hat{\theta}^{1,y_j}}{\prod_{j=1}^{J} \hat{\theta}^{1,y_j} + \prod_{j=1}^{J} \hat{\theta}^{0,y_j}}.
\]
Now we can write Eq. (8) as
\[
\max_x \frac{\prod_{j=1}^{J} \hat{\theta}^{1,y_j}}{\prod_{j=1}^{J} \hat{\theta}^{1,y_j} + \prod_{j=1}^{J} \hat{\theta}^{0,y_j}} \tag{10}
\]
where \( \hat{\theta}^{1,y_j} = \frac{\hat{\theta}^{1,y_j}}{\hat{\theta}^{0,y_j}} \). Since \( \hat{\theta} \geq 0 \), the objective in Eq. (10) is monotone increasing with \( \hat{\theta}^{1,y_j} \), \( \forall j \). Thus, we can solve Eq. (10) by maximizing \( \prod_{j=1}^{J} \hat{\theta}^{1,y_j} \) over \( x \). Since \( \hat{\theta}^{1,y_j} \) are independent across \( j \) in the Naive Bayes model, we can decompose the objective across \( j \) and maximize \( \hat{\theta}^{1,y_j} \) by setting \( x_j = \arg \max \hat{\theta}^{1,y_j} \), \( \forall j \). The property of decomposition across \( j \) in this formulation greatly simplifies the optimization problem allowing an extremely efficient implementation.

This can be done to approximately optimize Eq. (8) among peptides of any fixed length. We find that the length for which Eq. (10) is largest among those with \( f(e) > b \) is length \( b - 1 \). This is because, for any fixed \( \hat{\theta} \), there is always an amino acid class for which \( \hat{\theta} \geq 1 \), and so adding one more amino acid from this class improves Eq. (8). To allow searching over hits of many different lengths, we instead use a probability distribution over lengths, which we take to be uniform, randomly sample a length, and construct of peptide with this length. This includes searching over peptides of length \( b - 1 \) as a special case. We summarize the algorithm below in Algorithm 1.

This algorithm produces a set \( S \) of sequences \( e \) of amino acid classes. For each position within each of these amino acid class sequences \( e \), we randomly sample an amino acid from that class according to a uniform distribution to finalize the set of peptides to test.

We also describe two modifications of Algorithm 1. In the first modification, we may replace the point estimation step (step 6 by sampling the \( \hat{\theta}^{1,y_j} \) and \( \hat{\theta}^{0,y_j} \) from their posterior distribution. We run this modified step along with steps 7 and 9 repeatedly within an inner loop to obtain a collection of amino acid class sequences \( e \). For each position, we then choose the most frequently represented class at that position within this collection. This is the sequence of amino acid classes that we then add to \( S \) in step 10. This approach is less computationally intensive that implementing the greedy algorithm exactly, but uses more distributional estimation than Algorithm 1. We refer to this modification as "sampling-based" POOL, and to the version setting \( \theta \) to the posterior mean as "mean-based" POOL.

In the second modification, instead of adding a single copy of \( e \) to \( S \) in step 10, we add multiple copies. This pushes POOL’s future recommendations further away from its past recommendations, and encourages more diversity. We refer to this modification as "add-1", where \( k \) refers to the number of copies added, and "add-1" corresponds to unmodified POOL.
Algorithm 1 Greedy algorithm implementation

1. procedure GREEDY\(\hat{k}, J, K\), dataset \(D = \{(x^{1}, y^{1}), \ldots, (x^{N}, y^{N})\}\) and prior hyperparameters \(\alpha^{d}\)
2. \(S \leftarrow \emptyset\)
3. Calculate posterior distribution of \(\theta^{i,1:d} \sim \text{Dirichlet}(\theta^{i,1:d} | \{x : x \in D, y(x) = 1\})\).
4. for \(m = 1 \text{ to } k\) do
5. Calculate the posterior distribution of \(\theta^{i,2:j} \text{ and } \theta^{i,1:j}\) given the dataset \(D\) and that \(y(x) = 0\) for all \(x \in S\).
6. Calculate point estimators (either the posterior mode or the posterior mean) \(\hat{\theta}^{i,1:j}\) and \(\hat{\theta}^{i,2:j}\) of \(\theta^{i,1:j}\) and \(\theta^{i,2:j}\) for all \(j\).
7. Let \(\hat{\theta}^{i,j}_{d} = \hat{\theta}^{i,j}\) for all \(i\) and \(j\).
8. Choose a desired length at random from a distribution over lengths smaller than \(\delta\).
9. Construct \(x\) by setting \(x_{j} = \arg \max_{i} \hat{\theta}^{i,j}_{d}\) for each \(j\) up to the desired length. Let \(c\) be the corresponding sequence of peptide classes.
10. \(S \leftarrow (S, c)\).

H. Using partial labeling information. When searching for specific hits, \(y(c)\) is determined by considering several separately observable labels. For example, when searching for Sfp-specific hits, to know whether a peptide is a hit, we must observe both whether a peptide is a substrate for Sfp and for AcpS. If we have data for only one enzyme, then we do not observe \(y(c)\).

To avoid simply discarding such observations, this is not an efficient use of information. Here we describe how to use the POOL methodology when searching for specific hits while using measurements of less than the full set of enzymes for some training data.

Let \(y_{i}(c)\) be a binary label for desired constituent property \(i\), so that \(y(c) = \prod_{i} y_{i}(c)\). For example, when searching for Sfp-specific hits, let \(y_{i}(c) = 1\) when \(c\) is a substrate for Sfp and 0 otherwise, and let \(y_{i}(c) = 1\) when \(c\) is not a substrate for AcpS and 0 otherwise. Then \(y(c) = 1\) can only happen when \(c\) is a substrate for Sfp and not for AcpS.

To utilize as much information as possible, instead of building Naive Bayes classifier that predicts \(y(c)\) directly, we build separate Naive Bayes classifiers that predict each \(y_{i}(c)\). Then, we may follow the same reasoning with which we derived the POOL algorithm (maximizing probability of improvement using a greedy approach), and see that we can create a set of peptides to test by iteratively adding to \(S\) the peptide that maximizes Eq. (8).

Using the property that \(y(c) = \prod_{i} y_{i}(c)\), we may rewrite Eq. (8) as

\[
\arg \max_{c \in 2^{S}, \theta^{1}, \ldots, \theta^{d}} \prod_{i} P(y_{i}(c) = 1, \forall i | \prod_{i} y_{i}(c') = 0, \forall c' \in S). \quad [11]
\]

This is hard to maximize exactly because the number of configurations of \((y_{i}(c') : i, c' \in S)\) that satisfy \(\prod y_{i} = 0\) for all \(c\) grows exponentially with the size of \(S\), and computing Eq. (11) quickly becomes infeasible as \(S\) grows. We instead solve Eq. (11) using a heuristic approach in which we approximate the objective by

\[
\prod (y_{i}(c) = 1, \forall i | y_{i}(c') = 0, \forall i, \forall c' \in S). \quad [12]
\]

This can be optimized efficiently using an approach analogous to the one in Algorithm 1, but where we condition on \(y_{i}(c) = 0, \forall i\) whenever that algorithm conditions on \(y(c) = 0\).

Training sets 1-3 were generated using a single machine learning model. Training sets 4 and 5 were generated using partial labeling information.

3. Simulation-based comparison of POOL with existing methods

Figure 3 shows results of a simulation study in which we compared POOL with two existing methods: the predict-then-optimize or pure-exploitation method; and the mutation method. Here we define these two methods precisely (Sections A), provide a more detailed discussion of the predict-then-optimize method (Section B) and describe the method used to embed peptide sequences in two-dimensional space in Figure 3 C-D (Section C).

A. Definitions of existing methods. In our simulation study we compare with the predict-then-optimize method and the mutation method:

- In the predict-then-optimize method, we rank peptides that have not been tested according to \(P(y(c) = 1)\), as calculated by the Naive Bayes predictive model. Then, if we are allowed to recommend \(k\) peptides to test, we recommend the \(k\) with the highest probabilities.

- In the mutation method, we choose a known hit at random and then mutate it as follows. We randomly select the number of positions to mutate from a uniform distribution over \([1, 2, 3, 4]\). We then randomly select this number of mutations to perturb the target peptide, uniformly and without replacement. Finally for each position chosen to mutate, we randomly choose an amino acid to replace the original amino acid, uniformly from among all 20 amino acids. (Note that with probability 1/20, this amino acid will be the same as the original.)

B. Discussion of predict-then-optimize. The predict-then-optimize approach does not consider how future activity measurements will change the machine learning model's predictions, and how in turn this should affect the order in which peptides are tested. We argue that this makes the approach unnecessarily brittle to inaccuracies in predictions, and causes the set of peptides tested to lack diversity. Indeed, if the first peptide tested by the predict-then-optimize approach is not a hit, then the second peptide is typically quite similar, and is likely to also not be a hit.

By considering how measuring one peptide's activity will change the prediction of another's activity, POOL performs a more complete accounting of the value of testing a particular set of peptides, which causes it to recommend a more diverse
set of peptides. We argue that this leads POOL to provide more robust performance, allowing it to find high-fit active peptides in fewer experiments than the predict-then-optimize approach.

In this section, we illustrate this with an example. Suppose there are three peptides, A, B and C; A and B are very similar to each other, and their predicted probabilities of being a hit are both 0.9, written as \( P(A \text{ is a hit}) = P(B \text{ is a hit}) = 0.9 \), while C is different from A or B, and C has a lower predicted probability being a hit, say \( P(C \text{ is a hit}) = 0.8 \). Also suppose that the events that A is a hit and B is a hit are perfectly correlated (since A and B are very similar), and are both independent from the event that C is a hit (since C is different from A and B). We also assume \( f(A), f(B), f(C) \) are all greater than h.

We wish to find a hit by testing only two peptides, Predict-then-optimize will test A and B, because they have higher predicted probabilities of being a hit. However, we will see that this choice does not provide the highest probability-of-improvement.

We compute the probability-of-improvement when choosing A and B as:

\[
P(I\{A, B\}) = P(A \text{ is a hit}, B \text{ is a hit}) + P(A \text{ is a hit}, B \text{ is not a hit}) + P(A \text{ is not a hit}, B \text{ is a hit}) = 0.9 + 0 + 0 = 0.9.
\]

However, the probability-of-improvement when choosing A and C is

\[
P(I\{A, C\}) = P(A \text{ is a hit}, C \text{ is a hit}) + P(A \text{ is a hit}, C \text{ is not a hit}) + P(A \text{ is not a hit}, C \text{ is a hit}) = 0.9 \times 0.8 + 0.9 \times 0.2 + 0.1 \times 0.8 = 0.98.
\]

Similarly, \( P(I\{B, C\}) = 0.98 \). Therefore maximizing probability-of-improvement will result in testing either \( \{A, C\} \) or \( \{B, C\} \).

To understand this intuitively, suppose that the first peptide tested, A, is not a hit. Then B is also unlikely to be a hit because A and B are very similar to each other. On the other hand, since C is different than A and B, C may be a hit even if A and B are not. Therefore, choosing diversified and good peptides to test, i.e., choosing \( \{A, C\} \) or \( \{B, C\} \) in this example, is a better strategy than choosing peptides with high predicted value regardless of diversity, like choosing \( \{A, B\} \). This diversity is the key to finding hits when the machine learning model prediction is not accurate.

C. Two-dimensional embedding of peptide sequences. Figure 3 C-D visualizes peptide sequences as points in two-dimensional space. To create this visualization, we first compute a distance between each pair of peptides included in the visualization as follows: First, we set three penalty parameters \( \alpha = 10, \beta = 1.5 \) and \( \gamma = 1 \). Then, we align the two sequences at the central Serine. For each position to the left and right of the Serine, we then compute a penalty. If the two sequences have the same amino acid the penalty is 0. If the amino acids are different but are from the same class according to Supplementary Figure 2, then a penalty \( \beta \) is incurred. If the amino acids are from different classes then a penalty \( \alpha \) is incurred. If one sequence has an amino acid at this position while the other does not, then a penalty \( \gamma \) is incurred. This penalty is then summed across all positions to create the distance.

This method creates a matrix containing a distance between each pair of peptides. We then pass this matrix to the t-Distributed Stochastic Neighbor Embedding (t-SNE) method [49], implemented in Python in the scikit-learn package (http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html). This produces a 2-dimensional point which we then plot for each peptide such that the distance between each pair of points approximates the distances provided in the matrix.

4. Normalization of peptide intensities between rounds and discussion of lead peptide determination

A. Analysis of peptide hits. After treatment by PPTases and AcpP, the fluorescent membranes were imaged by a Typhoon FLA 9500 variable mode laser scanner (GE Healthcare). The following settings were used to image the membranes: 520 nm resolution (50 px), 350V PMT, and Ex 532 nm/Em 580 nm.

B. Normalization calculation. Due to the change of light conditions and machine calibration, the readings of light intensity scale differently in different rounds and across membranes. We develop a normalization method that we apply to the raw data to allow comparing measurements across multiple rounds and membranes to identify peptide hits.

To support describing this normalization procedure, we define notation: We performed 5 rounds of experiments, and therefore \( j \) is one of 1,2,3,4, or 5. We also have 4 treatments: Sfp \( (k=1) \), Sfp + AcpP \( (k=2) \), AcpS \( (k=3) \), and AcpS + AcpP \( (k=4) \). For peptide \( i \) tested in round \( j \) for treatment \( k \), we indicate the raw reading by \( y_{ijk} \). We will indicate the corresponding "normalized" value by \( \tilde{y}_{ijk} \).

We assume the raw reading \( y_{ijk} \) is the result of scaling and shifting the normalized value \( \tilde{y}_{ijk} \) and adding noise:

\[
y_{ijk} = \sigma_{i,k} \tilde{y}_{ijk} + \epsilon_{i,k} + \mu_{i,k},
\]

where \( \sigma_{i,k} \) is the scaling factor, \( \mu_{i,k} \) is the shifting factor, and \( \epsilon_{i,k} \) is independent normally distributed noise with mean 0 and variance \( \sigma^2 \). In addition, we know that after AcpP treatment, the normalized light intensity is the same or lower than before treatment, therefore \( \tilde{y}_{i,1} \geq \tilde{y}_{i,2} \) and \( \tilde{y}_{i,3} \geq \tilde{y}_{i,4} \).

We fit the parameters \( \mu_{i,k}, \sigma_{i,k}, \tilde{y}_{i,k} \) by maximizing the log-likelihood of the data, where the log-likelihood is

\[
\log(L) = \frac{1}{2} \sum_{i,j,k} \left( \frac{y_{ijk} - \mu_{i,k} - \epsilon_{i,k}}{\sigma_{i,k}} \right)^2 + N \log(2\pi\sigma^2),
\]

and we enforce constraints when maximizing of \( \tilde{y}_{i,1} \geq \tilde{y}_{i,2} \geq 0 \) and \( \tilde{y}_{i,3} \leq \tilde{y}_{i,4} \geq 0 \) for all \( i \). In addition, membranes include control peptides that are known to not react with AcpP after reacting with Sfp or AcpS. We set \( \tilde{y}_{i,1} = \tilde{y}_{i,2} = 0 \) for all \( i \in I_{2} \), where \( I_{2} \) is the set of peptides that are known to react with Sfp and not with AcpP, and \( \tilde{y}_{i,3} = \tilde{y}_{i,4} = 0 \), for all \( i \in I_{2} \), where \( I_{2} \) is the set of peptides that are known to react with AcpS and not with AcpP. We additionally add the constraint that \( \sum \tilde{y}_{i,k} = 0 \) for all \( k \).
should be equal to the constant 1, which avoids degenerate solutions in which $\sigma_{j,k}$ grows to $\infty$ while $\theta_{i,j}$ shrinks to 0.

Using the notation $\beta$ to indicate the vector comprised of $\mu_{i,j}$ for all $j$ and $k$, and similarly for $\sigma$ and $\delta$, we can write this fitting procedure more compactly as solving the following optimization problem:

$$\min_{\beta, \sigma, \delta} \sum_{i,j,k} \left( \frac{y_{i,j,k} - \beta_{i,j,k}}{\sigma_{i,j,k}} - \theta_{i,j} \right)^2$$

subject to $\theta_{i,1} - \theta_{i,2} \geq 0$

$\theta_{i,3} - \theta_{i,4} \geq 0$

$\theta_{i,3} - \theta_{i,4} = 0$ for all $i \in I_3$

$\theta_{i,3} - \theta_{i,4} = 0$ for all $i \in I_4$

$\sum \theta_i = 1.$

The problem above is a quadratic program, and can be solved efficiently by quadratic programming software such as Gurobi [50]. The solution gives us the normalized value $\theta_{i,j}$ for every peptide $i$ and every treatment $k$.

5. Synthesis and Chemoenzymatic Synthesis

A. Synthesis of TAMRA-C6-Pantethenamide

**General Methods:** Unless otherwise noted, all reagents and chemical compounds were purchased from Alfa Aesar, Sigma-Aldrich, Fisher Scientific or AAPP Tec and used without further purification. Flash chromatography was carried out on 239-400 mesh grade 60 silica gel (Fisher Scientific). NMR spectra were recorded on Varian VX500 spectrometer. FID files were processed through using MestReNova (Mestrelab Research). Mass spectrometric analyses were conducted on the ACQUITY ultra performance LC (Waters) with 2.6 mm, C18 100 A, Kinetex LC column (150 x 3 mm).

**TAMRA-C6-Boc (1):** tert-butyl-(6-((3′,6′-bis(dimethylamino)-3-oxo-3H-spiro[isobenzofuran-1,9′-xanthene]-6-carboxamido)hexyl)carbamate (1). 5-TAMRA was made according to literature procedures [51]. To a stirred solution of 5-TAMRA (300 mg, 0.69 nmol) in dry DME (3 mL), HATU (300 mg, 0.84 mmol) and DIPEA (250 µL, 1.35mmol) were added. The solution was stirred under nitrogen at room temperature for 30 minutes. N-Boc-1,6-diminohexane (180mg, 0.84 mmol) was then added and the solution was left to stir overnight under nitrogen gas. The solution was then concentrated to dryness in vacuo then resuspended in 100mL of DCM and washed with aq NaHCO₃ (sat.) 3 x 50mL and brine 50 mL. The organic layer was dried over anhydrous Na₂SO₄ and then concentrated to dryness in vacuo. The residue was then run on silica gel in acetone to afford the product 247 mg (56.6%) as a purple solid. 1H NMR (500 MHz, CDCl₃) : δ = 8.40 (bs, 1H), 8.20 (d, 1H, J = 8.1 Hz), 7.25 (d, 1H, J = 8.1 Hz), 6.59 (d, 2H, J = 8.8 Hz), 6.49 (d, 2H, J = 2.6 Hz), 3.48 (q, 2H, J = 6.0 Hz), 3.13 (m, 2H), 3.00 (s, 12H), 1.66 (m, 2H), 1.49 (m, 4H), 1.44 (s, 9H), 1.56 (m, 2H) 13C NMR (125 MHz, CDCl₃): δ = 169.13, 165.94, 156.69, 153.01, 152.25, 136.32, 134.12, 128.70, 124.69, 122.86, 108.79, 106.26, 98.57, 78.65, 40.18, 40.69, 39.75, 29.97, 29.25, 28.57, 26.06, 25.82. HR-MS [M+H]+ Theor. 629.3334 Obs: 629.3330.

**TAMRA-C6-NH₂ (2):** 6-((3′,6′-bis(dimethylamino)-3-oxo-3H-spiro[isobenzofuran-1,9′-xanthene]-6-carboxamido)hexyl)ammonium TFA Salt (2). TAMRA-C6-Boc (1) (150 mg) was stirred in 10 mL of TFA for 2 hours. The solution was poured into 100 mL of diethyl ether and the precipitate was filtered and dried in vacuo to give 120 mg of the product as a purple solid (99%). 1H NMR (500 MHz, DMSO-d₆): δ = 9.00 (t, J = 5.6 Hz, 0H), 8.72 (d, J = 1.8 Hz, 0H), 8.34 (dd, J = 7.8, 1.8 Hz, 0H), 7.94 (s, 1H), 7.61 (d, J = 7.9 Hz, 0H), 7.68 (dt, J = 15.0, 7.5 Hz, 2H), 6.95 (d, J = 2.2 Hz, 0H), 3.77 (q, J = 6.6 Hz, 1H), 3.28 (s, 6H), 2.84 (h, J = 5.9 Hz, 1H), 1.61 (h, J = 6.8 Hz, 2H), 1.40 (p, J = 3.7 Hz, 2H). 13C NMR (125 MHz, DMSO-d₆): δ = 166.45, 165.04, 159.10, 158.85, 158.59, 158.34, 157.18, 157.08, 136.60, 131.77, 131.59, 131.01, 130.80, 129.95, 118.58, 116.24, 115.00, 115.10, 96.71, 40.98, 40.61, 39.24, 29.38, 27.50, 26.55, 26.08. HR-MS [M+H]+ Theor. 529.2804 Obs: 529.2810.

**TAMRA-C6-Pantethenamide (3):** (S)-N-(6-((3′,6′-bis(dimethylamino)-3-oxo-3H-spiro[isobenzofuran-1,9′-xanthene]-6-carboxamido)hexyl)(2,4-dihydroxy-3,3-dimethylbutanamido)propanamido)hexyl)-3′,6′-bis(dimethylamino)-3-oxo-3H-spiro[isobenzofuran-1,9′-xanthene]-6-carboxamide (3). TAMRA-C6-NH₂ (2) (100 mg) stirred in 2 mL of DME with 200 mg of HATU and 150 mg of PMB protected pantetheine acid. The solution was poured into 50 mL of DCM and washed with 3 x 25 mL of saturated sodium bicarbonate and 25 mL of brine. The organic layer was concentrated in vacuo. The residue was then dissolved in 5 mL of 80% acetic acid in water. The solvent was removed via rotary evaporator and the crude mixture was run on preparative HPLC 25% ACN/H₂O to 40% ACN/H₂O over 50 minutes retention time (26 minutes). Fractions containing the product were combined and lyophilized to give 60 mg (69%) of the product as a light purple solid. 1H NMR (500 MHz, DMSO-d₆): δ = 8.92 (t, J = 5.7 Hz, 0H), 8.69 (d, J = 1.8 Hz, 0H), 8.30 (d, J = 8.6, 1.8 Hz, 0H), 7.98 – 7.88 (m, 1H), 7.57 (d, J = 7.9 Hz, 0H), 7.10 – 6.99 (m, 2H), 6.95 (d, J = 2.1 Hz, 1H), 3.70 (s, 0H), 3.38 – 3.32 (m, 2H), 3.25 (s, 6H), 3.26 – 3.13 (m, 1H), 3.03 (qd, J = 6.9, 2.7 Hz, 1H), 2.26 (d,
J = 7.0, 2.2 Hz, 1H), 1.56 (q, J = 7.1 Hz, 1H), 1.40 (p, J = 7.1 Hz, 1H), 1.38 – 1.26 (m, 2H), 0.78 (d, J = 14.5 Hz, 3H).

$^{13}$C NMR (125 MHz, DMSO-d$_6$): $\delta$= 173.28, 170.72, 166.40, 164.95, 158.59, 158.33, 157.21, 157.10, 136.58, 136.26, 131.73, 131.54, 131.03, 130.84, 129.99, 115.04, 113.15, 96.67, 75.35, 75.34, 68.45, 59.96, 39.49, 38.85, 35.56, 35.28, 29.57, 29.47, 26.71, 26.67, 21.38, 20.71. HR-MS [M+H]$:^+$ Theo: 729.3738 Ori: 730.3809.

**B. Chemoenzymatic Synthesis of TAMRA Coenzyme A (CoA).**

![Diagram of TAMRA-CoA](image)

**TAMRA-CoA (4).** Compound 4 was synthesized in situ using CoA biosynthetic enzymes (CoaA, CoaD, CoaE) as previously reported [52, 53]. The 1 mL reaction volume was incubated at 37°C shaking for 8-12 hours and contained 5 mM of TAMRA-C6-Pantothenamide (3), 32 mM ATP pH 8, 0.01 μg μL$^{-1}$ *S. aureus* 6-His CoaA, 0.01 μg μL$^{-1}$ *E. coli* MBP CoaD, and 0.01 μg μL$^{-1}$ *E. coli* MBP-CoaE in 50 mM Tris/HCl buffer pH 7.5. The lyophilized reaction mixture was purified on a VyDAC protein & peptide C18 300 Å (4.6 x 250 mm) column using an Agilent 1100 semi-preparative HPLC using a reverse-phase protein peptide C18 column with a solvent system of (A) 0.1% TFA in water and (B) 0.1% TFA in acetonitrile at 1 ml min$^{-1}$. The run consisted of 95% of buffer A for 5 minutes, followed by a gradient of 95% buffer A to 50% buffer A over 30 minutes. TAMRA-CoA was monitored at 254 nm. Purified compound 4 was lyophilized and identity confirmed by HRMS: $C_{30}H_{32}N_4O_{20}P_3$ 1219.03.
Figure S2.1 (A) Reaction scheme illustrating labeling by PPTase of carrier protein with conserved serine. PPTase: 4'-phosphopantetheine transferase; CP: carrier protein. (B) Reaction scheme illustrating labeling by PPTase of POOL peptide substrate at conserved serine with TAMRA-CoA analog. (C) The apo-peptide is designed to act as a surrogate for the second helix (highlighted in orange) of the carrier protein containing the serine modification site.
**Figure S2.2** Twenty essential amino acids organized by 8 reduced classes.
Figure S2.3 Distances in simulation study between training data and peptides recommended by POOL and Mutation: As discussed in the main text (“Diversity of peptides recommended in the simulation study”) and our discussion of this simulation study (Section 3), peptides recommended by the Mutation method tend to be closer to the training data than those recommended by POOL. Recall that our simulation study represents peptides by 2-dimensional points. We calculated the Euclidean distance in 2-dimensions between the closest point in the training data and each peptide recommended by POOL and Mutation. The left plot shows the distributions of these distances for the two methods when searching for Sfp-specific peptides; the right plot shows the distributions when searching for AcpS-specific peptides. In each plot, two overlapping histograms are shown, where red indicates POOL, blue indicates Mutation, and purple indicates an overlap in the two histograms. We also provide summary statistics from these histograms. For the left (Sfp) plot: mean 1.63 (Mutation), 2.55 (POOL); median 1.58 (Mutation), 1.64 (POOL); 25% quartile 1.12 (Mutation), 1.13 (POOL); 75% quartile 2.08 (Mutation), 2.57 (POOL). For the right (AcpS) plot: mean 0.97 (Mutation), 1.67 (POOL); median 0.86 (Mutation), 1.01 (POOL); 25% quartile 0.40 (Mutation), 0.64 (POOL); 75% quartile 1.22 (Mutation), 1.37 (POOL).
Figure S2.4 Full SDS-PAGE images of GFP-peptide fusion proteins after PPTase labeling. (A) Repeat image for clarity of illustrating Sfp labeling of GFP-peptide fusion proteins from main text Figure 2.5A. (B-C) Full 12% SDS-PAGE gel images corresponding to the main text Figure 2.5B. (D) An additional Coomassie gel containing a molecular weight marker was run to confirm the correct size of each GFP-peptide fusion. PPTase (Sfp and AcpS) is not observed due to low abundance (Coomassie sensitivity). TAMRA-CoA was added to all reactions including no enzyme control.
**Figure S2.5** LC-MS for characterizing labeling of AcpS peptides: *In cellulo* labeling of coumarin CoA onto GFP-peptide by endogenous *E. coli* PPTase (AcpS) was confirmed by LC-MS analysis of coumarin phosphopantetheine after hydrolysis from the GFP-peptide [16]. (A) LC-MS analysis of hydrolyzed phosphopantetheine from AcpS-type peptide 1F01 was monitored by single ion detection mode set for measuring M+1=627, coumarin phosphopantetheine : m/z=626 (shown above). LC-MS trace showing total ion current of the same sample (shown below). (B) LC-MS analysis of hydrolyzed phosphopantetheine from AcpS-type peptide 1I04 was monitored by single ion detection mode (shown above). LC-MS trace showing total ion current of the same sample (shown below). (C) LC-MS analysis of hydrolyzed phosphopantetheine from AcpS-type peptide 3K17 was monitored by single ion detection mode (shown above). LC-MS trace showing total ion current of the same sample (shown below). (D) LC-MS analysis of hydrolyzed phosphopantetheine from AcpS-type peptide 4T25 was monitored by single ion detection mode (shown above). LC-MS trace showing total ion current of the same sample (shown below).
Table S2.1 Gel-based electrophoretic mobility shift kinetics for \textit{B. subtilis} Sfp.

<table>
<thead>
<tr>
<th>GFP FUSION PEPTIDE ID</th>
<th>Peptide Sequence</th>
<th>$k_{\text{cat}}$ min$^{-1}$</th>
<th>$K_{m}$ µM</th>
<th>$k_{\text{cat}} / K_{m}$ µM min$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YbbR-GFP</td>
<td>DSLEFIASKLA</td>
<td>2.8 ± 0.2</td>
<td>0.5 ± 0.1</td>
<td>5.9 ± 1.5</td>
</tr>
<tr>
<td>4P28</td>
<td>LIGIDSIETLKA</td>
<td>0.20 ± 0.02</td>
<td>0.4 ± 0.2</td>
<td>0.5 ± 0.1</td>
</tr>
<tr>
<td>4N28</td>
<td>YGDEIPAESLDFLE</td>
<td>0.04 ± 0.01</td>
<td>1.6 ± 0.6</td>
<td>0.03 ± 0.01</td>
</tr>
<tr>
<td>4F01</td>
<td>AVKMESLEYLDTM</td>
<td>1.3 ± 0.1</td>
<td>0.6 ± 0.1</td>
<td>2.2 ± 0.6</td>
</tr>
</tbody>
</table>

Acknowledgements: Chapter 2, in full, is a reprint of the material as it appears: Tallorin, L., Wang, J., Kim, W. E., Sahu, S., Kosa, N., Yang, P., Thompson, M., Gilson, M. K., Frazier, P. I., Burkart, M. D., Gianneschi, N. C. “Discovering de novo Peptide Substrates for Enzymes Using Machine Learning” \textit{Nature Communications} 2018 9(1) 5253. The dissertation author is the primary co-author of this manuscript along with Dr. Lori Tallorin and Dr. Jialei Wang.

Section 2.6 References


Chapter 3. Activity-guided single cell genomics reveals a selective biosynthetic enzyme-substrate interaction

Section 3.1 Introduction

Nearly 100 years ago in 1926, James B. Sumner obtained the first pure sample of an enzyme. These octahedral crystals of the jack bean urease provided the first evidence that enzymes were proteins and launched an ongoing exploration towards understanding the mechanisms guiding enzyme activity. Remarkably, 30 years earlier in 1896, the first commercial enzyme in the United States, diastase, a fungal amylase, gained its acceptance as a remedy for digestive disorders, and ultimately provided a spark to fuel the translation towards a global market that will reach $6.2 billion annually in 2020. Currently, a considerable part of this market is derived from microbial and microbial-produced enzymes, due to ease in availability, rapid growth and biotechnological optimization.

Currently, microbial genomic analyses are accomplished by accessing metagenomes, sequencing cultured organisms, or by single cell genomics. As illustrated in Figure 3.1a, one typically turns to bioinformatics analyses to identify a targeted enzymatic pathway within these datasets. Alternatively, one could integrate a fluorescence-based enzyme assay to guide the selection of individual microorganisms that express activities of interest directly in their natural environment. This activity-guided approach (Figure 3.1a) has the potential to identify targeted pathways in rare microorganisms without assumptions in the types of microorganisms and genes encoding these pathways. As shown in Figure 3.1a, it provides an excellent complement to current genomic sequencing...
approaches by applying a biochemical assay as the tool for organism selection, a process that can be directly integrated with single cell genomics.

To demonstrate this approach, we turned to carrier proteins (CPs) and their associated 4′- phosphopantetheinytransferases (PPTases), an enzyme-substrate pair that plays a key role in polyketide (PK) and non-ribosomal peptide (NRP) biosynthesis.\textsuperscript{10} As shown in Figure 3.1b, this system provides a robust model as we have previously shown that synthetic analogues of pantetheine (pantetheinamides) can cross the cell membrane.\textsuperscript{11} Once in the cell, they can highjack coenzyme A (CoA) biosynthesis and be converted to the corresponding CoA analog and ultimately be appended onto the CP by the action of the PPTase (bottom, Figure 3.1b).\textsuperscript{10} Overall, this process can be used to label CPs by attaching a fluorescent tag to the N-terminus of the pantetheinamide (Figure 3.1b).

Our study began by the choice of a fluorescent reporter. Advanced in 2008,\textsuperscript{12} 4-dimethylnaphthalene (4-DMN) is a solvatochromic tag whose fluorescence responds according to the hydrophobicity of its environment.\textsuperscript{13} In our studies, we have used a 4-DMN-pantetheinamide (DMN-P, Figure 3.1b) to probe protein-protein interactions between \textit{Escherichia coli} fatty acid synthase (FAS) acyl carrier protein (ACP) and its various partner enzymes.\textsuperscript{14} When an ACP is labeled with DMN-P, it sequesters the dye within its hydrophobic alpha helical core,\textsuperscript{14} leading to a large increase in fluorescence intensity relative to that unbound in solution. Based on our previous studies, we anticipated that DMN-P would cross the cell membranes and label CPs \textit{in vivo} (Figure 3.1b). The fact that the 4-DMN dye was not fluorescent in solution but was when protein-bound suggested that this strategy could be used to directly identify cells that contained the enzymatic machinery necessary to load a CP with DMN-P, as illustrated in Figure 3.1b.
Figure 3.1 Activity-guided microbial single cell genomics. (a) Multiple approaches exist to mine enzymes from microbial communities, including: sequencing of cultured microbial strains (C); metagenomic analyses (M); and, single cell genomics (S). Shown as in a cycle (blue lines and light blue shading), activity-guided single cell genomics (A) labels a targeted enzyme within the microbial community to guide the sorting of enzyme active microbes. While each pathway provides genomic data that can be searched for enzymatic pathways, the activity-guided approach enriches for the targeted enzyme prior to genome sequencing. (b) Schematic representation of the fluorescent CP labeling method used in this study. A fluorescent pantetheinamide DMN-P, designed to mimic pantetheine, is up taken into a microbial cell where it is converted to a fluorescent-CoA analog. An environmentally sensitive fluorescent tag was used to improve the detection of protein-conjugation in vivo, as the 4-DMN dye has been shown to undergo an increase and shift in fluorescence once appended to a CP by a PPTase.

Section 3.2 Results and Discussion

We began by testing the labeling strategy using a set of culturable marine microbes. Super-resolution imaging (see Online Methods) of Staphylococcus sp. CNJ924 indicated that specific cells (blue fluorescent cells, Figure 3.2a) became fluorescent upon incubation with 500 nM DMN-P. Counterstaining with 1 µM SYTO 9, a green-fluorescent nucleic acid live-cell stain, indicated that while some cells stained with DMN-P (green and blue fluorescent cells, Figure 3.2a) others did not (green fluorescent only cells, Figure 3.2a), therein suggesting that while some cells were biosynthetically active for CP labeling (Figure 3.1b), others were not. This observation suggested that DMN-P staining was able
not only to label the *Staphylococcus sp.*, but also to engage specific states of microbial growth. In other species, such as the rod-shaped *Bacillus sp.* CNJ803, consistent subcellular localization was observed. As evident in Figure 3.2b, *Bacillus sp.* cells incubated with 500 nM DMN-P displayed blue fluorescence throughout their cytosolic space, in contrast to DNA staining, which appeared within nanometer-sized nucleoid regions. From these examples, we were able to demonstrate cellular and intracellular selective labeling of microbial cells with DMN-P.

We then applied this labeling to an environmental sample. In August 2018, we collected tunicates *Ciona intestinalis* (Figure 3.2c) and ambient seawater (water column, WC) from a dock in Boothbay Harbor, Maine. Here, our goal was to test the activity-based single cell genomics concept using tunicate microbiomes. We began by incubating freshly caught tunicates in purified seawater with RedoxSensor Green (RSG), a marker of bacterial cell viability, or the enzyme-active DMN-P. A total of four tunicates were used for each probe. After homogenization of the tunicates, probe-positive cells were subjected to fluorescence-activated cell sorting (FACS) to deposit fluorescently-labeled cells and cell aggregates into a 384-well microplate, one per well (see Section 3.4). Of the ~12,000 µL⁻¹ viable (RSG-positive) microbial cells or cell aggregates in the tunicate homogenate, only ~13 µL⁻¹ cells (0.1%) were labeled with DMN-P, indicating high probe specificity. From each well, Illumina libraries were created, sequenced and assembled (see Section 3.4).

From this data, the phylogenetic composition of microbial cells in the adjacent water column (WC, Figure 3.2e) was typical of the Gulf of Maine prokaryoplankton, with a predominance of pelagic lineages SAR11 (Alphaproteobacteria) and Flavobacteriales (Bacteroidetes). The viable microorganisms in the tunicate microbiome were dominated by
Rhizobiales, Kordiimonadales (Alphaproteobacteria), Flavobacteriales (Bacteroidetes) and Campylobacterales (Epsilonproteobacteria), which have been shown to be associated with *C. intestinalis*<sup>17</sup>. Cell aggregates from the tunicate microbiomes labeled with the DMN-P were dominated by Campylobacterales; Flavobacterales; Alteromonadales, Cellvibrionales, Vibrionales and Oceanospirillales (Gammaproteobacteria). Genomic sequencing confirmed that, in contrast to cells labeled with SYTO-9 and RSG, many particles labeled with DMN-P contained multiple, phylogenetically divergent cells, resulting in mini-metagenomes. This potentially suggests that active CP labeling by PPTase may be greatest when the producer cell is in direct contact with a different microorganism.

As shown in Figure 3.2f, antiSMASH 5.0 identified 0.77 metabolite pathways per Mbp in the RSG probe-positive single amplified genomes (SAGs), and 1.17 metabolite pathways per Mbp in DMN-P positive SAGs, a level that was considerably higher than the 0.47 metabolite pathways per Mbp that were observed in the water column. This observation confirmed that activity-based probes, such as DMN-P, can select microbes with biosynthetic potential. From this data, we mined two CPs. The first CP1 (Figure S3.3) and its associated PPTase, PPT1 (Figure S3.4), were obtained from a putative NRPS/Type I hybrid gene cluster in an Oceanospirillales (Gammaproteobacteria) genome AH-491-C20. An *Amphritea spongicola* MEBiC0546<sup>18</sup> was found to be the most closely related cultured isolate in Genbank (97% 16S rRNA gene identity). This single-aggregate’s mini-metagenome also contained a Kordiimonadales (Alphaprotobacteria) member. A second, unrelated CP2 (Figure S3.5) was obtained from a predicted NPRS/NRPS-like/Type I PKS cluster in a Cellvibrionales (Gammaproteobacteria) genome AH-491-D14, with *Oceanicoccus sagamiensis* NBRC107125 as the most closely related cultured isolate in
Genbank (96% 16S rRNA gene identity).\textsuperscript{19} This single-aggregate mini-metagenome also contained a Flavobacteriales (Bacteroidetes) member.

Starting with the sequences obtained with antiSMASH,\textsuperscript{20} we designed genes for each of the three proteins that contained a fully translatable domain motif. For CP1, we selected a gene that contained a di-domain with two CPs (Figure S3.2). A similar strategy was used to include the adenylating domain with CP2 (Figure S3.3) as well as design PPT1 (Figure S3.4). \textit{E. coli} codon optimized genes were synthesized for CP1, CP2, and PPT1 into pET28a vectors and their associated proteins (Figure S3.2-S3.4) were prepared by recombinant expression in \textit{E. coli} followed by His\textsubscript{6}-tagged purification (Figure S3.5a).

Applying the method in Figure 3.1b \textit{in vitro}, samples of CP1 and CP2 were then screened for their ability to be fluorescently labeled \textit{in vitro}. Sfp,\textsuperscript{10} a member of the surfactin-type PPTase known to have a broad scope in CP labeling, was able to load DMN-P onto CP1 under conditions that were able to label the \textit{E. coli} FAS ACP, a positive control (Figure 3.2g, Figure S3.7). CP2 was not labeled under these conditions, nor was CP1 or CP2 labeled with the mined PPTase PPT1. Concerned that unfolding the proteins under SDS-PAGE would destroy the environmentally sensitive fluorescence from the 4-DMN tag, we repeated the labeling process (Figure 3.1b) with non-solvatochromic dye, TAMRA-CoA.\textsuperscript{21} We now observed labeling of CP1 with Sfp (Figure 3.2h, Figure S3.5b-d). Interestingly, the mined PPTase, PPT1, was only able to label CP1. Remarkably, while Sfp could label the ACP and CP1, the fact that PPT1 only labeled its native substrate CP1, and not the \textit{E. coli} FAS ACP, demonstrate the unique selectivity within PPTases.\textsuperscript{10}
Figure 3.2 Application of activity-guided single cell genomics to mine a selective CP and PPTase interactions from a clonal tunicate microbiome. Super-resolution images of a) Staphylococcus sp. CNJ924 or b) Bacillus sp. CN803 incubated for 3 h with 500 nM DMN-P (blue) and then stained with 1 μM nuclear SYTO-9 (green) prior to fixation and imaging. Bar denotes 10 μm. Enlarged images are provided in Figure S3.1. c) An image of a Ciona intestinalis specimen explored in this study. d) Cells (black dots) sorted and selected (blue regions) by flow cytometry with DMN-P staining for tunicate and seawater samples. Negative (Bacillus subtilis 168, a mutant deficient in Sfp) and positive (Bacillus subtilis 3610 DSM10, wild type) controls indicate DMN-P stained cells with active CP - PPTase pairs. e) Phylogenetic composition and f) predicted secondary metabolite clusters per megabase (Mbp⁻¹) of the total microbial community in the water column (WC) compared to the viable (R) and DMN-P-responsive (D) cells in the C. intestinalis microbiome. g)-h) In vitro CP-labeling analyses on the cloned and expressed CP1, CP2, PPT1 mined from the genomic data. SDS-PAGE gel depicting the fluorescence in CPs after labeling by Sfp or PPT1 with g) DMN-P or h) TAMRA-CoA. The E. coli FAS ACP was used as a positive control.

Section 3.3 Conclusions

Overall, this study demonstrates how activity-based fluorescent labeling can be used to guide single cell and cell aggregate selection and sequencing (Figure 3.1b). We demonstrated how the use of environmentally sensitive fluorescent probes designed to label CPs could be used to guide the selection of organisms whose genomes encode CPs and associated PPTase enzymes. Genomic data from the activity-selected organisms was enriched in biosynthetic activity, a finding that enabled the identification of an undiscovered carrier protein substrate (CP1) and PPTase enzyme partner (PPT1). The observed selective compatibility of PPT1 with a CP1 from the same cell demonstrates the enzymatic selectivity that can be harvested using activity-guided single cell genomics. Most
importantly, these enzymes can be realized without microbial cultivation. Herein, we illustrate an approach that suggests an important role for single cell genomics as a tool for the activity-guided discovery of enzymes of synthetic and biosynthetic utility. Conceptually, these studies suggest the need for expanded fusion between fluorescent labeling, high-content microscopy, cell sorting and single genomics.

Section 3.4 Methods

General materials

Unless described otherwise, supplies and materials were obtained from VWR or Fischer Scientific and used as is. The DMN-P$^{14}$ and TAMRA-CoA$^{21}$ were prepared by chemical synthesis. *E. coli* FAS ACP was prepared using established methods.$^{14}$

Microbial labeling studies

A single colony of each strain (*Staphylococcus sp. CNJ924* or *Bacillus sp. CNJ803*) was suspended in 200 µL of A1 media. Cells were then treated at 23 °C for 3 h with 500 nM DMN-P (Figure 3.1b) from a 100 µM stock solution of DMN-P dissolved in DMSO. Following incubation with DMN-P, nuclear stain, SYTO-9 (Thermo Fisher Scientific), was added to the cultures for 15 min at a final concentration of 1 µM. Cells were spun down at 2,000 rpm for 1 min and the supernatant was removed. Cell pellets were mixed with 200 µL of 3:1 EtOH/AcOH (fixing solution) and incubated for 10 min. Cells were centrifuged at 2,000 rpm for 1 min again and the supernatant was removed. The cells were re-hydrolyzed with 100 µL of water. A 20 µL aliquot of each final sample was loaded onto glass slides for super-resolution microscopy.
**Super-resolution microscopy**

Imaging was conducted on a Zeiss LSM 880 confocal laser scanning microscope with FAST Airyscan equipped with a Plan-Apochromat 63x/1.4 Oil DIC M27 objective. Blue fluorescence from the DMN-P was obtained using a 405 nm laser with beamsplitters (MBS 488/561/633, MBS_InVis MBS -405, DBS1 mirror, FW1 rear) and an additional 410-477 nm filter. Green fluorescence from the SYTO-9 control was obtained using a 488 nm laser with beamsplitters (MBS 488/561/633, MBS_InVis MBS-405, DBS1 mirror, FW1 rear). Pinhole sizes were kept between 40-90 µm. Images were collected in Zen (Zeiss) with recommended gains of 700-1100, digital gain of 1, depth of focus between 0.65-0.71 µm, and pixel time of 2-5 µs and image sizes at 2048 × 2048 pixels. Images were processed offline and rendered from Zen Blue (Zeiss). Copies of original CZI files can be provided upon request.

**Specimen collection**

Live *Ciona intestinalis* were collected from 20.4 °C seawater at a depth of ~0.5 m on the side of a floating dock in Boothbay Harbor Maine (Latitude 48.8, Longitude -69.6) between 8:00-9:00 AM on August 29, 2018. Immediately after collection, the specimens were placed in 50 mL centrifuge tubes containing ambient seawater. Additionally, ambient seawater samples were collected adjacent to tunicates in 50 mL centrifuge tubes. The samples were transported to Bigelow Laboratory for Ocean Sciences at *in situ* temperature in the dark.

*Activity-guided single cell genomics*
To target cells with CP activity, four tunicates were incubated in 80 mL of sterile Sargasso seawater amended with 10 µM DMN-P at *in situ* temperature in the dark for 2 h. To target all viable microbial cells, four other tunicates were incubated in 80 mL of ambient seawater amended with 1 µM RedoxSensor Green (RSG, Thermo Fisher Scientific) at *in situ* temperature in the dark for 20 min. Then the content of each tube was homogenized with a Ninja Professional 900 W blender until the majority of the biomass was visibly disintegrated. The homogenate was spun down at 2,000 rpm for 1 min and the supernatant was passed through a 100 µm mesh filter. To assess the composition of microorganisms in the seawater around tunicates, ambient seawater samples were labeled with a 5 µM SYTO-9 live nucleic acid stain (Thermo Fisher Scientific) for 20-40 min. Immediately before cell sorting, samples were diluted 10× in sterile Sargasso Sea water. Sort gates for probe-positive cells were defined along green fluorescence and forward scatter axes and adjusted for background noise using negative (*Bacillus subtilis* 168, a mutant deficient in Sfp) and positive (*Bacillus subtilis* 3610 DSM10, wild type) controls. Probe-positive cells were sorted into 384-well plates, lysed, their DNA amplified with WGA-X, and genomes sequenced and quality-controlled.\textsuperscript{14} Genome assemblies originating from multiple cells (mini-metagenomes) were identified and parsed using a combination of nucleotide tetramer principal component analysis and homology searches in the NCBI nr database.\textsuperscript{15} Biosynthetic pathways were identified with antiSMASH\textsuperscript{20} using KnownClusterBlast, ActiveSiteFinder and SubClusterBlast options. If antiSMASH predicted multiple metabolite pathways in the same coding region and overlapped then all possible products are reported.
Heterologous expression of select genes/domains

Selected domains and di-domains were chosen based upon evaluation of sequencing data with antiSMASH (versions 4.2 and 5.0 were used).\textsuperscript{20} In particular, we selected a PPTase (PPT1) and two CP (CP1 and CP2) containing sequences for recombinant evaluation. The predicted protein sequences of the full selected genes were further evaluated with BLAST alignment to identify homologues and better understand domain organization. In the case of CPs, disconnection locations for excised domains were determined by online domain organization tools\textsuperscript{22,23} and by comparison with recent crystal structures.\textsuperscript{24,25} The genes for the resulting proteins were synthesized (Twist Bioscience) and cloned into a pET28a vector. With the exception of the PPTase, all genes were cloned with a 3'- stop codon to provide an N-terminal His\textsubscript{6} tag for immobilized metal affinity chromatography purification. The PPTase gene was cloned without a stop codon to provide a C-terminal His\textsubscript{6} tag based upon our prior experience with loss of activity with N-terminal fusions.\textsuperscript{10} The identified protein sequences, gene sequences synthesized and protein sequences are provided in Figures S3.2-S3.4.

Expression of CP1, CP2, and PPT1

The His\textsubscript{6}-tagged proteins were expressed in \textit{E. coli} (BL21), and grown in Terrific Broth. Cells were grown in the presence of 50 mg/L kanamycin, induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (ITPG) at OD\textsubscript{600} = 0.8, and incubated at 16 °C for 16 h. The cell culture was spun down by centrifugation at 2,000 rpm for 30 min and the collected pellets were lysed by sonication, followed by another centrifugation at 10,000 rpm for 1 h to clear the lysate. The proteins were purified using Ni-NTA resin.
(ThermoFisher). Purified proteins were collected and concentrated to 2-4 mg/mL using 3 kDa (CP1 and CP2) or 10 kDa (PPT1) Amicon Ultra centrifuge filters (Millipore).

*Carrier protein labeling studies*

The CP labeling studies were conducted in a 30 µL reaction containing final concentration of 100 µM of the respective CP1 or CP2, 1 mM of TAMRA-CoA, and 1 µM of Sfp or PPT1 in 50 mM HEPES, 10 mM MgCl₂ pH 7.2. The mixture was gently shaken at 23 °C overnight (12 h). The resulting reactions were analyzed by 12% SDS-PAGE gel and imaged on Typhoon TRIO Variable Mode Imager (GE Healthcare BioSciences). Imaging was conducted using Cy2 (Excitation 473 nm, Emission 520 nm) for DMN-P and Cy3 (Excitation 532 nm, Emission 580 nm) for TAMRA-CoA.

**Section 3.5 Supplementary Information**

*DMN-P Guided Fluorescent Activated Cell Sorting (FACS)*

Bigelow Laboratory Single Cell Genomics Center’s (scgc.bigelow.org) standard workflow was employed to generate and sequence single amplified genomes. Prior to FACS, samples were diluted to below 10⁵ cells mL⁻¹ with filter-sterilized Sargasso Sea water and pre-screened through a 40 µm mesh filter (Becton Dickinson). FACS was performed using a BD InFlux Mariner flow cytometer equipped with 365 and 488 nm lasers for excitation and either a 70 µm nozzle orifice (Becton Dickinson, formerly Cytopeia). The cytometer was triggered on side scatter, and the “single-1 drop” mode was used for maximal sort purity. Gates for the sorting were defined based on particle blue
(DMN-P) or green (SYTO-9 and RSG probes) fluorescence, light side scatter, and the ratio of green versus red fluorescence (for improved discrimination of cells from detrital particles).

**Single Cell and Single Cell Aggregate DNA Amplification and Genome Sequencing**

Cells were lysed and their DNA was denatured by 2 freeze-thaw cycles, the addition of 700 nL of a lysis buffer consisting of 0.4 M KOH, 10 mM EDTA and 100 mM dithiothreitol, and a subsequent 10 min incubation at 20 °C. The lysis was terminated by the addition of 700 nL of 1 M Tris-HCl, pH 4. The 10 µLWGA-X reactions for genomic DNA amplification contained 0.2 U µL\(^{-1}\) Equiphi29 polymerase (Thermo Fisher Scientific), 1× Equiphi29 reaction buffer (Thermo Fisher Scientific), 0.4 µM each dNTP (New England BioLabs), 10 µM dithiothreitol (Thermo Fisher Scientific), 40 µM random heptamers with two 3'-terminal phosphorothioated nucleotide bonds (Integrated DNA Technologies), and 1 µM SYTO-9 (Thermo Fisher Scientific). These reactions were performed at 45 °C for 12-16 h, then inactivated by a 15 min incubation at 75 °C. In order to prevent WGA-X reactions from contamination with non-target DNA, all cell lysis and gDNA amplification reagents were treated with UV in a Stratalinker (Stratagene). An empirical optimization of the UV exposure was performed in order to determine the length of UV exposure that is necessary to cross-link all detectable contaminants without inactivating the reaction. Cell sorting, lysis and gDNA amplification setup were performed in a HEPA-filtered environment conforming to Class 1000 cleanroom specifications. Prior to cell sorting, the instrument, the reagents and the workspace were decontaminated for DNA using UV irradiation and sodium hypochlorite solution. To further reduce the risk of DNA
contamination, and to improve accuracy and throughput, Bravo (Agilent Technologies) and Freedom Evo (Tecan) robotic handlers were used for all liquid handling in 384-well plates.

Libraries for SAG genomic sequencing were created with Nextera XT (Illumina) reagents following the manufacturers instructions, except for purification steps, which were done with column cleanup kits (QIAGEN), and library size selection, which was done with BluePippin (Sage Science, Beverly, MA), with a target size of 500±50 bp. Libraries were sequenced with NextSeq 500 (Illumina) in 2×150 bp mode using v.2 reagents. The obtained sequence reads were quality-trimmed with Trimmomatic v0.32 using the following settings: -phred33 LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. Human DNA (≥95% identity to the *H. sapiens* reference assembly GRCh38) and low complexity reads (containing <5% of any nucleotide) were removed. For *de novo* assemblies, quality-filtered reads were digitally normalized with kmernorm 1.05 (http://sourceforge.net/projects/kmernorm) using settings -k 21 -t 30 -c 3 and then assembled with SPAdes v.3.0.0 using the following settings: --careful --sc --phred-offset. Each end of the obtained contigs was trimmed by 100 bp, and then only contigs longer than 2,000 bp were retained. This workflow was evaluated for assembly errors using three bacterial benchmark cultures with diverse genome complexity and %GC, indicating no non-target and undefined bases in the assemblies and average frequencies of mis-assemblies, indels and mismatches per 100 kbp: 1.5, 3.0 and 5.1.

Functional annotation was first performed using Prokka with default Swiss-Prot databases supplied by the software. The output of Prokka and the secondary annotation were joined into a single, tab-delimited table with headers identifying the origin of the assignment. The genome completion estimates were performed on conserved protein-
coding genes by CheckM. The 16S rRNA gene regions longer than 500 bp were identified using local alignments provided by BLAST against CREST's curated SILVA reference database SILVAMod v128; and taxonomic assignments were based on a reimplementaton of CREST's last common ancestor algorithm.

**Biosynthetic pathway analyses**

Biosynthetic pathways were identified with antiSMASH, using KnownClusterBlast, ActiveSiteFinder and SubClusterBlast options. Contigs with predicted metabolite pathways were categorized by metabolite classification type; if >1 metabolite pathway type was predicted all were reported. Total predicted metabolites for each probe type were divided by the total assembly sizes of all SAGs to calculate metabolites per Mbp.
Figure S3.1 Enlarged super-resolution AiryScan images. a) Image of *Staphylococcus sp.* CNJ924 incubated for 3 h with 500 nM DMN-P (blue) and then stained with 1 μM nuclear SYTO-9 (green) prior to fixation and imaging. b) Super-resolution images of *Bacillus sp.* CNJ803 incubated for 3 h with 500 nM DMN-P (blue) and then stained with 1 μM nuclear SYTO-9 (green) prior to fixation and imaging. Bars denote 10 μm. AiryScan imaging were collected under optical settings at which the blue fluorescence in unstained *Staphylococcus sp.* CNJ924 and *Bacillus sp.* CNJ803 was indistinguishable from background.
**Carrier Protein 1 (CP1)**, a putative PCP from well AH-491-C20, Cluster 23, Ctg 2642_6

Sequences identified antiSMASH search on Ctg 2642_6:

```
TFWERLLEQVPVTDPDSDFRLGGDSLLATRMVVELSKHLGQPVSLSQHVFEPVLPQAFCAALE
TLTGIWEDILEQPVNAQTDFFRAGGDSLSATRMVVAQQAGFSQISLQQVFEHPRFSDFCQVLTG
```

**Synthesized CP1 gene:**

```
GTCTCCGCTGATCTGCTGTTCACATTTCCAAACCTTCTGGGAACGCCTGCTGAGCAGCCTGTTAC
GCCGGATACGCATTTTTTCTGTTTAAGGTGGATTCTGGTTGTTGCAACCGCTATGGTAGCTGAAC
TGAGCAAAACATCTCGCCAGCAGGCTGTCTCTGCAACACGCACGTGGTTGAACAGCCAGCAGTTGCAAGC
ATTTTGCGCCGCCATTGGAGGCCCCCCTACCGCAGGCAGGGAGCGATGTCAGCCTTATCAGAA
ACCCTCAAGGAATTGGAGGATATCCTCGAAACAACAGTTAAACGCCGAAGCCGACTTCTCC
GCCGAAGCGGCGACTCGTCTCCCGCACCAGCCTGGTGGTCGCGCGCAGACAGCTGGGTATTC
GCAAATTAGCTTACAAACAGTATTCCGAACATCCGCGGGATTTCTGATTTCTGGCAGACCTTGGA
CAGCGGAAAATTTCCAGGTTGCAAGACCGGAGGGGAAGTCACTGCGACAGAAACCTAA
```

**Sequence of Expressed CP1 protein:**

```
MGSSHHHHHHSSGLVPRGSQMVSSLHLQLVEQTPQDFERLLEQVPVTDPDSDFRLGGDSLLATRMVVELSKHLGQPVSLSQHVFEPVLPQAFCAALEAPVTEAAAGDVLLESETTLTGIEDILEQPVNAQTDFFRAGGDSLSATRMVVAQQAGFSQISLQQVFEHPRFSDFCQVLTGAISELEAEPGESVETATET
```

**Figure S3.2** Discovery, synthesis and expression of Carrier Protein 1 (CP1). (top) Sequence identified from the Ctg 2642_6 were used to (middle) synthesize a gene and (bottom) prepare the CP1 protein. Protein sequences are colored with: (black) amino acid sequence added from the pET28a vector; (green) amino acid sequence of the gene not reported by antiSMASH but added from the sequenced gene to provide a complete protein; (blue) amino acid sequence of the first of the two carrier proteins identified by antiSMASH, and; (red) amino acid sequence of the second of the two carrier proteins identified by antiSMASH.
**Carrier Protein 2 (CP2)**, a putative ACP from c03335_AH-491, Cluster 28, Ctg 3232_22.

Sequences identified from Ctg 3232_22:

TYEQLHKSSSKLANFLSVHGVKPSDVICLDRSVELMICLIGILKMGAAYVPIDPNYPKERTNFILNSKLRLLLLESSEYSGLNFEDGETIYLDATCNYECADSSDFPGVSLSNALMYMIFTSGSTQPKGVRVS

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP2</td>
<td>Carrier Protein 2</td>
<td>MGSSHHHHHSSGLVPRGSHMLQKVSHPSVYAVEAGDQRTYTEQELHSSSKLANFLSVHGVKPSDVICLDRSVELMICLIGILKMGAAYVPIDPNYPKERTNFILNSKLRLLLLESSEYSGLNFEDGETIYLDATCNYECADSSDFPGVSLSNALMYMIFTSGSTQPKGVRVS</td>
</tr>
</tbody>
</table>

Figure S3.3 Discovery, synthesis and expression of Carrier Protein 2 (CP2). Protein sequences are colored with: (black) amino acid sequence added from the pET28a vector; (green) amino acid sequence of the gene not identified by antiSMASH but added from the sequenced gene to provide a complete protein; (blue) amino acid sequence of the adjacent adenylation domain identified by antiSMASH, and; (red) amino acid sequence of the carrier proteins identified by antiSMASH.
4'-Phosphopantetheinytransferase 2 (PPT1), a putative PPTASE from c02649_AH-491-C20, Cluster 23, Ctg 2642_10.

Sequences identified from Ctg 2642_10:
EWLLLQSYHCRRELLTVAFSVKEAFYKAIP

Synthesized PPT1 gene:
CAGTCCCGGGCGGATCTCCAAGGCAGCGTTCTTAACGTGCTAGTCTTAATTTTAGCAGGGGGGCTA CTGCGCGGTCCAAGCGATGGGTATGGCAGGGGAAATTAAATCTTGAAGCACCATTGCCTGCGCTCA GGTTGTCCTCCTAATCTGGCCGATGTCAGTGCCGCTGGGCTGCATGGGTATGGCCGGAATTTTAAATCCTGAGCCACCATTTGCGGGGCTTA CTCGCTATTTCTTTGCGTGACGGACACTGGCTGCGGACTGGGTGTGACATCGAGGGCCTTTAATT GCAGAGCATAGTCCTGCAACCAGTCATGCCCCTATGAGACAGCCAGATGTGGCTGTGCTGCTGC AGTCCTACCATGGCGCTGTCGAGTTATTAACGTGGGCCTTCAGCGTAAAGAGGCTCTTCTACAAA GCAAATCTATCCTATATCTGCCGCTTTGCCGATTTCCATAGGTGTTGTCATCATCAGCGTTTGGGCCA GGGACGCGACGCGTTCGAGCTGCGGACTGCAGCAGAGTTACAAAGAATATTCTCAGGCCAGTTACCGTATGGGTAGTTGCGTGTCGAAAGCCTTGTCGTTATGATGCGTCGTGCGGAAAAGGGAGAGGACAGCGGCCAAGAACCTTTTTGAGGGCTGCGCTGCCGCTTTAATCCAGACCGCATCGGCGGCT

CP2 protein:
MAVPADLQGAVLKRQCEYLAGRYCAVQAMGMAILNPEPPLRGSGGAPIWPAGIRGSISHSCQAV AILCRTLDTGCAGLGVHDIEGFIADDLTDVPLVLQDTEWLLLQSYHCRRELLTVAFSVKEAFYKAIP YICRFADFHEVITGLGKRAALTPRAGLAAELPKNILQASYRMGSCGVESRVMMRRAEKGGEVPE TFCEGSRLIQTQWAGLGH

Figure S3.4 Discovery, synthesis and expression of 4'-Phosphopantetheinytransferase 1 (PPT1). A putative PPTASE from c02649_AH-491-C20, Cluster 23, Ctg 2642_10. Protein sequences are colored with: (black) amino acid sequence added from the pET28a vector; (green) amino acid sequence of the gene not identified by antiSMASH but added from the sequenced gene to provide a complete protein; (blue) amino acid sequence of the PPTase domain identified by antiSMASH.
Figure S3.5 Labeling experiments. a) Gel depicting the expressed and purified CP1, CP2, and PPT1. PPT1 was observed as monomeric and dimeric form by SDS-PAGE. b) Total protein content for the labeling of CP1 and CP2 by Sfp or PPT1 with TAMRA-CoA (left) and DMN-P (right). c) TAMRA fluorescence observed in the gel in b) for the labeling of CP1 by Sfp or PPT1. This gel was also reported in Figure 3.2g. d) DMN fluorescence observed in the gel in b) for the labeling of CP1 by Sfp or PPT1. This gel was also reported in Figure 3.2g.

Acknowledgements. This study was supported by NIH R21 AI 134037 to M. D. B. and R. S. R. S. was also supported by NSF 826734, 1441717 and 133581. We thank Prof. William Fenical (Scripps Institution of Oceanography, UC San Diego) for samples of Staphylococcus sp. CNJ924 and Bacillus sp. CNJ803 used to develop the labeling conditions and Prof. Jörn Piel (ETH) for the Bacillus subtilis strains used as positive and negative controls. Confocal imaging was possible through the assistance of Jennifer Santini (UC San Diego) and supported by funding from the NIH P30 NS047101.

The material for this chapter is currently being prepared for submission for publication as: Kim, W. E., Charov, K., Becraft, E. D., Brown, J., La Clair, J. J., Stepanauskas, R., Burkart, M. D. “Activity-Guided Single Cell Genomics Reveals a
Selective Biosynthetic Enzyme-Substrate Interaction” The dissertation author is the primary co-author of this manuscript along with Katherine Charov, Dr. Eric D. Becraft, and Dr. Julia Brown.

Section 3.6 References

(1) Sumner, J. B. J. Biol. Chem. 1926, 69, 435-441.

(2) Sumner, J. B. Science 1933, 78, 335.


Chapter 4. Mechanistic probes for the epimerization domain of nonribosomal peptide synthetases

Section 4.1 Introduction

Natural products produced by microorganisms (i.e., bacteria and fungi) have had a tremendous impact on human health, even before Fleming’s discovery of penicillin’s activity as an antibiotic in 1928. Today, many of these microbial metabolites serve as antibiotics, antivirals, immunosuppressants, and anticancer agents. Prominent among these therapeutics are a large class of peptide-based natural products called nonribosomal peptides, which are assembled by large multimodular enzymes: the nonribosomal peptide synthetases (NRPSs). NRPSs are made up of modules that are in turn composed of catalytic domains that each perform a single chemical reaction either to incorporate or modify a monomer unit into the final natural product. Central in mediating NRP synthesis is the peptidyl carrier protein (PCP), which acts as a scaffold, tethering amino acid building blocks and peptidyl intermediates through a 4’-phosphopantetheine prosthetic arm as they are modified and condensed by other domains in the module. The past decade has seen remarkable progress in the structural and mechanistic elucidation of NRPS domains through the use of biochemical, structural, and genetic methods. Despite these achievements, new strategies to explore the enzymology of these complex, natural product-producing assembly lines are in constant development. One such approach, which has gained considerable attention, is the design, synthesis and application of small-molecule probes to investigate the functional workings of biological systems, a field called chemical biology. Such tools have been shown to have great promise in elucidating the
physiochemical mechanisms of type II fatty acid biosynthesis, a modular process that is similar in synthetic logic to NRP synthesis. The development of new chemical biology approaches will extend researchers’ ability to interrogate the mechanisms of NRPS catalytic domains as well as uncover the protein interactions regulating efficient processing of biosynthetic intermediates in NRP synthesis with the ultimate goal of reengineering their biosynthetic machinery for the biocombinatorial synthesis of novel peptides with improved potency.

Chemical probes have been developed to investigate the protein interactions that govern communication-mediating (COM) domains, determining the substrate specificities of the core thiolation (PCP), adenylation (A), and condensation (C) domains as well as a few tailoring domains in NRP systems. To continue with this approach, we have designed mechanism-based probes to investigate the epimerization (E) domains of NRPSs, which are responsible for the incorporation of D-amino acids into NRPs. The E domain producing D-amino acids occurs through the in situ epimerization of the Cα center of the S-aminoacyl-PCP. Few biochemical studies have been performed on the E domain since its discovery, and much regarding the mechanism of action of this enzyme has yet to elucidated.

Mutational studies of the E domain in the initiation module, GrsA, of the gramicidin S synthetase revealed active-site residues H753 and E892 to be vital for epimerization activity, thus suggesting a deprotonation/reprotonation mechanism. Sequence alignment and secondary-structure comparisons of the E domain of GrsA and the C domain of VibH, a condensation domain-like amide synthase involved vibriobactin biosynthesis, reveal the structural similarity of the two domains, including a similar channel present at the face of their active sites. Consequently, it was proposed that the 4’-phosphopantetheine prosthetic
arm of the PCP binds within this channel from the N-terminal face of the E domain, presenting the covalently attached L-amino acyl substrate at the active site in order for the catalytic residues to epimerize the C\textsubscript{\(\alpha\)} center of the amino acid. It was also shown that the E domain catalyzes racemization of the C\textsubscript{\(\alpha\)} center of the amino acid only when it is tethered to the 4’-phosphopantetheine prosthetic arm of the PCP.\textsuperscript{11}

Section 4.2 Results and Discussion

In contrast to the E domains of NRPSs, amino acid racemases, best exemplified by the bacterial alanine racemase, require a cofactor, pyridoxal 5’-phosphate (PLP), to catalyze racemization.\textsuperscript{12} PLP, a coenzyme form of vitamin B6, stabilizes the anionic intermediate formed when L- or D-alanine is deprotonated at its C\textsubscript{\(\alpha\)} center by capturing this intermediate covalently as a quinoid adduct.\textsuperscript{13} In addition, mutational studies of bacterial alanine racemase demonstrated that two catalytic residues (K39 and Y265) were responsible for proton abstraction and addition at the C\textsubscript{\(\alpha\)} center of the amino acid substrate.\textsuperscript{14} Haloalanines and halovinylglycines have been reported to be potent mechanism-based inhibitors for the bacterial alanine racemase.\textsuperscript{15,16} Given the mechanistic similarities between the alanine racemase and NRPS epimerization domain, we reasoned that chlorovinylglycine (CVG) \textsuperscript{1} and \(\beta\)-fluorophenylalanine \textsuperscript{2} (Scheme 4.1) would serve as promising candidates for the design of mechanism-based probes of the E domain. To accomplish this, we first synthesized probes \textsuperscript{1} and \textsuperscript{2} according to previously published procedures.\textsuperscript{16-18}
We used the tridomain (A-PCP$_{\text{Phe}}$-E) initiation module, GrsA, of the gramicidin S synthetase (Figure 4.1) to determine the efficacy of probes 1 and 2 as mechanism-based crosslinkers. We chose this module because it contained the necessary domains to activate (A), load (PCP), and undergo crosslinking (E) with the probes to show the inactivation of the E domain. To determine the activity of probes 1 and 2 toward the E domain, we employed a previously developed in vitro condensation assay to show the inhibition of the
E domain indirectly. This assay monitors dipeptide formation between the initiation module (GrsA) and the second tridomain (C-A-PCPPro) module of the tyrocidine system, TycB1. In this assay, L-Phe and L-Pro are first activated through ATP-dependent adenylation by the A domains of GrsA and TycB1, respectively. Once activated, they are loaded onto their corresponding PCPs, forming L-Phe-S-GrsA and L-Pro-S-TycB1. L-Phe-S-GrsA is modified to D-Phe-S-GrsA by the E domain. Due to the stereoselectivity of the TycB1 C domain, condensation only occurs once D-Phe is produced through epimerization by GrsA. The C domain of TycB1 then catalyzes peptide bond formation to give the dipeptidyl intermediate D-Phe-L-Pro-S-TycB1. This dipeptide undergoes intramolecular cyclization releasing the D-Phe-L-Pro diketopiperazone (DKP) product, which can be detected by HPLC. DKP formation cannot happen when the E domain is inactivated by using mechanism-based probes. Thus, DKP production can be used as a measure of probe activity.

For the condensation assay, we first preincubated GrsA at 37 °C with the probe and ATP for 1, 3, 6, 9, 12, and 24 h to prevent competitive adenylation of the natural substrate, L-Phe. This would activate and load the probe onto the PCP so that it could then react with and inactivate the E domain. Afterwards, we ran the condensation assay on the natural substrates L-Phe and L-Pro with the preincubated probe bearing GrsA and TycB1. HPLC analysis of the assay showed no decrease in DKP formation when GrsA was preincubated with β-fluorophenylalanine but showed a decrease in DKP formation with 1, thus suggesting that it was partially inactivating the E domain (Figure 4.2).
Scheme 4.1 Mechanism-based crosslinking probes chlorovinylglycine (1) and β-fluorophenylalanine (2) for the NRPS epimerization domain.

Figure 4.2 HPLC analysis of DKP formation after preincubation of GrsA with probes 1 and 2.

Based on these results, we next preincubated GrsA with 1 in the absence of ATP and then performed the condensation assay to verify that the inhibition of DKP formation was dependent on the A domain’s ability to recognize, activate, and load 1 onto the PCP domain prior to its reaction with the E domain. HPLC analysis of this assay showed no decrease in DKP formation when ATP was omitted in the preincubation period, thereby signifying that 1 alone could not inhibit the E domain, and thus DKP formation. Instead, the probe must be activated by the A domain and tethered onto the PCP in order to inactivate the E domain. With probe 2 there was still full DKP formation, thus no inhibition of the E domain. We presume that, as probe 2 is very similar to the natural substrate and
product, L- and D-phenylalanine, once loaded onto PCP, it undergoes E domain-catalyzed epimerization and is subsequently offloaded.

We propose a mechanism for the chemical crosslinking between 1-PCP and the GrsA E domain that is consistent with the mechanism proposed for the epimerization of L-Phe. The x-ray crystal structures of the TycA epimerization domain, a homologue of GrsA’s E domain, also responsible for the epimerization of L-Phe, suggests that – given the local environment – the catalytic histidine exists in its doubly protonated (imidazolium) form. Computational predictions of pKₐ values (using propKₐ as implemented on the H++webserver) of the catalytic histidines of both the TycA and GrsA E domains support this conclusion. Consequently, it is proposed that the histidine acts not as an acid or base, but rather promotes epimerization by stabilizing the enolate species that is generated as a reaction intermediate. Given His753’s inability to act as an acid or base, Glu892 (in the case of GrsA’s E domain) is proposed to assume this role. Thus, we propose that the Glu892 deprotonates 1 at Cₐ, generating an enolate intermediate that collapses to form the reactive allene intermediate (shown in Figure 4.3). The allene species is activated by protonation by His753, which simultaneously enhances the nucleophilicity of His753, thereby enabling it to undergo Michael addition with 1 (Figure 4.3). Further work is necessary to definitely demonstrate that the catalytic histidine is targeted by 1-PCP. Nonetheless, we contend that, of the two active-site residues, only His753 can react with 1-PCP to form a stable enzyme-probe adduct.
In order to further investigate the activity of the probe on the epimerization domain, we performed molecular dynamics (MD) simulations to better rationalize the behavior of 1 as probe and substrate. We generated models of L-phenylalanyl-PCP-E, D-phenylalanyl-PCP-E, and chlorovinylglycyl-PCP-E didomain by using the holo-PCP-E didomain crystallized by Bruner et al. (PDB ID: 5ISX). Our analysis of the MD simulations was performed with the aforementioned mechanistic information in mind.

Simulations suggest that the L-phenylalanyl substrate must undergo internal rotation in order to present the hydrogen at $C_{\alpha}$ to Glu892. For this to occur, the salt bridge between
Glu892 and the (protonated) amino group of the substrate must be disrupted. The reorganization of the substrate in this manner can be monitored by using the S-C(O)-Cα-Cβ dihedral angle (Figure 4.4). Analysis of this dihedral angle for L-phenylalanyl-PCP-E indicates that L-Phe-PCP assumes one of two rotamers: (-) anticlinal and (+) synclinal. The (+) synclinal conformer disposes the substrate’s phenyl substituent in tryptophan-rich region of the active site, while simultaneously maintaining a salt bridge between protonated amine of the substrate and Glu892. The (-) anticlinal conformation of L-Phe disposes the substrate alpha hydrogen in proximity to Glu892, but the substrate thioester no longer hydrogen bonds to His753. Simulations of D-phenylalanyl-PCP-E indicate that the D-phenylalanyl substrate adopts one of three conformations about its S-C(O)-Cα-Cβ dihedral in the E domain active site: (-) synclinal, (+) synclinal, or (+) anticlinal. Both the (-) and (+) synclinal substrate rotamers dispose the substrate towards Glu892, consistent with the E domain’s ability to convert D-phenylalanyl-PCP into the starting material. The difference in the S-C(O)-Cα-Cβ dihedral angles of the two epimeric phenylalanyl substrate samples can be attributed to the shape complementarity of the GrsA E domain towards L-phenylalanyl-PCP, namely that the substrate can dispose its aromatic side chain toward a region rich in tryptophan (aromatic) residues while forming salt bridge (Figure 4.4). As no such rotamer exists for the D substrate, transitions between S-C(O)-Cα-Cβ conformers are likely more facile.

Data collected from simulations of L- and D-phenylalanyl-PCP-E provide a context by which the reactivity of the probe 1 can be better understood. The chlorovinylglycyl substrate is able to sample a broad range of S-C(O)-Cα-Cβ dihedral angles similar to those sampled by the cognate L-Phe substrate. This presumably allows the probe to reorient itself
within the E domain active site so that its reactive warhead is presented to the active-site nucleophile. Further analysis and a complete description of the computational methodology employed to perform the simulations described herein can be found in the Supporting Information.

Section 4.3 Conclusions

In summary, we propose new insight into the mechanistic role of the catalytic residues in the epimerization domain of NRPSs. A mechanism-based crosslinker, chlorovinylglycine – previously shown to react with alanine racemase – demonstrates partial inhibition towards the GrsA’s epimerization domain. MD simulations suggest that substrate organization with the E domain active site is central to reactivity of both cognate and noncognate aminoacyl substrates.
Figure 4.4 Distribution of the substrate \(S\text{-C(O)}\text{-C}\alpha\text{-C}\beta\) dihedral angels of (A) L-Phe-PCP-E, (B) D-Phe-PCP-E, (C) 1-PCP-E. Newman projections and 3D representations of representative structures of each \(S\text{-C(O)}\text{-C}\alpha\text{-C}\beta\) rotamer sampled by conventional MD simulation. Catalytic residues His753 and Glu892 are shown in orange. The substrate and phosphopantetheine are shown as sticks and colored white. Trp632, Trp644, and Trp911 (aromatic binding pocket) are shown as spheres in light blue. Note that hydrogen, oxygen, nitrogen atoms are colored white, red, and blue.

Section 4.4 Methods

Expression and purification of GrsA and TycB₁

GrsA and TycB₁ were expressed and purified by using a modified version of previously published protocols:¹⁹,²² pQE60 plasmids (Qiagen) harboring the genes were transformed into BL21(DE3) cells. grsA was co-transformed with sfp (the gene for the phosphopantetheine transferase from the surfactin A NRPS) in pREP4, and tycB₁ with gsp (the gene for the phosphopantetheine transferase from the gramicidin S NRPS) in pREP4.

For protein production, two times yeast and tryptone medium (2xYT: 16 gL⁻¹ trypton, 10 gL⁻¹ yeast extract and 5 gL⁻¹ NaCl, 500 mL) with kanamycin sulfate (30 mgL⁻¹)
and ampicillin (100 mgL⁻¹) was inoculated with an overnight culture (5 mL) of the respective expression strain. The cultures were grown at 37 °C/120 rpm to an optical density at 600 nm (OD₆₀₀) of 0.3. At this point, the temperature was reduced to 30 °C, and expression was induced at OD₆₀₀ at 0.6-0.8 by adding isopropyl β-D-1-thiogalactopyranoside (IPTG; 0.4 mM), and the cells were subsequently grown for an additional 2 h.

Cells were harvested by centrifugation (2000 g, 30 min), resuspended in buffer A (25 mM HEPES, 200 mM NaCl, 10 mM imidazole, pH 8.0), and lysed by being passed through a French Press twice. The lysate, clarified by centrifugation, was loaded onto a NiNTA column (Qiagen), washed with 15 column volumes (CV) of buffer A, and eluted with 3 CV buffer B (as buffer A, but with 250 mM imidazole). The eluate was dialyzed against HEPES (50 mM), NaCl (300 mM), MgCl₂ (10 mM), and dithiothreitol (DTT; 5 mM, pH 7.5) overnight at 4 °C.

To ensure quantitative phosphopantetheinylation, the proteins were incubated with coenzyme A (CoA; 5 mM) and Sfp (0.05 equiv.) for 1 h at 20 °C before they were concentrated and loaded onto a Sephacryl S300 size-exclusion column (GE Healthcare) equilibrated with HEPES (50 mM), NaCl (300 mM), and MgCl₂ (10 mM, pH 7.5). Monomeric fractions were again concentrated, and aliquots were frozen in liquid nitrogen and stored at -80 °C.

Running the DKP formation assay with nonfrozen enzymes directly after the size-exclusion chromatography step and enzymes that were stored at -80 °C resulted in virtually identical yields, thus indicating no loss of activity upon freezing/thawing.
Diketopiperazine (DKP) formation assay (positive control)

The assay was performed according to a previously published protocol\textsuperscript{19} with some modifications: mixtures of GrsA (2.8 µM), L-Phe (8.0 mM), and ATP (8.0 mM), as well as TycB\textsubscript{1} (2.8 µM), L-Pro (8.0 mM), and ATP (8.0 mM), were made on a 200 µL scale in HEPES (50 mM, pH 7.5), NaCl (300 mM), MgCl\textsubscript{2} (10 mM). DKP formation was started by combining the two mixtures, which were allowed to react for 1 h at 37 °C. The reaction was stopped by extraction with butanol/chloroform (4:1; 2×400 µL). The combined organic phases were washed with NaCl (400 µL, 0.1 M) and dried under reduced pressure.

The dried extract was dissolved in methanol (50 µL, 10% v/v), and aliquots (20 µL) were loaded onto an Ascentis Express Peptide ES C\textsubscript{18} column (15 cm x 4.6 mm; 2.7 µm particle size; Sigma-Aldrich) for quantification. Solvent A was 0.05% trifluoroacetic acid (TFA) in water; solvent B was 0.05% TFA in acetonitrile. Prior to loading of a sample, the column was equilibrated with 2% B, and the following gradient was used at a flow rate of 1 mL min\textsuperscript{−1}: 0-2 min, 2% B; 2-10 min, 2-30% B; 10-12 min, 30-95% B; 12-15 min, 95% B; 15-17 min, 95-2% B; 17-20 min, 2% B (UV absorbance at 280, 254, 214 nm). Peaks were collected to screen for the mass of DKP (C\textsubscript{14}H\textsubscript{16}N\textsubscript{2}O\textsubscript{2} calcd: 245.13 [M+H]\textsuperscript{+}, obs.:245.18). The yield of DKP was determined by the peak intensity of its absorbance at 214 nm.

Diketopiperazine formation assay: Inhibition by mechanism-based crosslinkers

In 200 µL scale reactions in HEPES (50 mM), NaCl (300 mM), and MgCl\textsubscript{2} (10 mM, pH 7.5), GrsA (2.8 µM final concentration), ATP (8.0 mM), and probe (0.8 mM from a stock in DMSO) were pre-incubated for 1-24 h at 18 °C. Following the pre-incubation, DKP formation was started by adding a mixture (200 µL) of TycB\textsubscript{1} (2.8 µM), ATP (10.0
mM), L-Phe (8.0 mM), and L-Pro (8.0 mM). After 1 h at 37 °C, DKP was extracted and quantified as described for the positive control. In control reactions with 24 h pre-incubation, ATP was omitted in the pre-incubation step. All of the described reactions were also performed with DMSO substituting for chlorovinylglycine (CVG).

Section 4.5 Supplementary Information

Simulation preparation

Coordinates for the simulations were derived from the crystal structure of the holo-PCP-E didomain construct of the initiation module of the GrsA nonribosomal peptide synthetases (PDB code: 5ISX)\(^{21}\). The L-Phe-PCP-E, D-Phe-PCP-E, 1-PCP-E, 3R-2-PCP-E, 3S-2-PCP-E complexes were generated by modifying the phosphopantetheine moiety of the experimental structure using Avogardo 1.1.1,\(^{23,24}\) Gaussview 5.0,\(^{25}\) and Pymol v.1.8.6.\(^{26}\) The amino groups of the peptidyl substrates of each of these didomains were simulated in their protonated form. For each simulated didomain, the peptidyl substrate, phosphopantetheine cofactor (affixed to the Ser\(^{42}\)), and that serine residue of the PCP domain of each system were treated as a single nonstandard residue. AMBER (ff14SB)\(^{27}\) and GAFF\(^{28}\) type force field parameters were assigned to the atoms of these nonstandard residues using ANTECHAMBER.\(^{29}\) The RESP methodology\(^{30}\) (HF/6-31G*) was used to determine partial atomic charges for all nonstandard residues. All quantum calculations were performed using Gaussian 09.\(^{31}\) The protonation state of all titratable residues in these didomains were assigned using the H++ webserver.\(^{32-35}\) Proteins were solvated in a water box with TIP3P\(^{36}\) water molecules. Using TLEAP,\(^{37}\) the simulation cell was constructed such its edges were placed so that closest proteinogenic atom was 10 Å away. Counterions
(Na\(^+\) and Cl\(^-\)) were also added via TLEAP to neutralize and ‘salt’ the water box to mimic physiological conditions (0.15 M).

**Simulation details**

Amber\(^{14}\) was used to perform all conventional (cMD) and Gaussian accelerated molecular dynamics (GaMD)\(^{38-43}\) simulations. All simulations were performed using the ff14SB\(^{27}\) and GAFF\(^{28}\) force fields. A 2 fs time-step was utilized via the SHAKE algorithm, which constrains all nonpolar bonds involving hydrogen atoms.\(^{44}\) Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method with a 10 Å cutoff for all non-bonded interactions.\(^{45}\) Both solvated protein complexes were energy minimized in a two-step fashion. First, solvent and counterions were allowed to relax, while the synthetase didomain was restrained using a harmonic potential (k = 500 kcal mol\(^{-1}\) Å\(^{-2}\)). This geometry optimization was followed by an unrestrained energy minimization of the entire system. The thermal energy available at a physiological temperature of 305 K was slowly added to each system over the course of a 1.7 ns NVT ensemble simulation. The solvated complexes were then subjected to unbiased isobaric-isothermal (NPT) simulations for 2.5 ns in order to equilibrate the ‘heated’ structures. The Langevin thermostat (\(\tau = 1.0\) ps\(^{-1}\)) was used to maintain temperature control.\(^{46}\)

Three 250 ns cMD and GaMD simulations (NPT ensemble) of each didomain were performed. For GaMD simulations, the threshold energy was set to \(V_{\text{max}}\). The maximum (\(V_{\text{max}}\)), minimum (\(V_{\text{min}}\)), average (\(V_{\text{av}}\)), and the standard deviation (\(s_V\)) of the system potential was determined using a 1.8 ns unbiased (NPT) simulation. This data was not subjected to analysis.
**Analysis and visualization of simulation data**

Analysis was performed using PYTRAJ, a Python front-end for the CPPTRAJ analysis code.\(^{47}\) Trajectories were visualized using VMD 1.9.1\(^ {48}\) and Pymol v.1.8.6\(^ {26}\)

Coordinate data was written to disk every 0.5 ps. Structures rendered in Figure 4.4 were selected by clustering data based on the substrate S-C(O)-C\(_{\alpha}\)-C\(_{\beta}\) and identify the centroid of each cluster.

---

**Figure S4.1** HPLC analysis of %DKP formation after preincubation of probe 1 and 2 without ATP. This indicates that the probe is only active when it is loaded onto A domain with the presence of ATP. Probe in their free amino acid form do not inhibit the epimerization domain.
**Figure S4.2** Primary HPLC trace of % DKP formation of the control and with the probe 1. Raw HPLC trace of the DKP formation assay (A) without the probe and (B) with the probe 1. The peak at a retention time of 10.89 min corresponds to D-Phe-L-Pro diketopiperazine by mass spectrometry.

**Figure S4.3** Schematic diagram showing DKP assay. (Top) The diagram above shows the process by which holo-GrsA and holo-TycB₁ – in presence of L-Phe, L-Pro, and ATP – produces the D-Phe-L-Pro diketopiperazine. (Bottom) Illustration of the inactivation of holo-GrsA by preincubating of holo-GrsA with chlorovinylglycine. When inactivated, holo-GrsA is unable to epimerize the D-Phe, the substrate recognized by downstream condensation domain of holo-TycB₁, thus preventing the synthesis of the diketopiperazine dipeptide. Product formation is monitored by HPLC.
Figure S4.4 GrsA PCP-E didomain. The crystal structure of the GrsA PCP-E didomain (PDB ID: 5ISX) used to generate initial coordinates for molecular dynamics simulation. In orange, red, and ball and sticks are the epimerization domain, peptidyl carrier protein, and the phosphopantetheine cofactor, respectively.
Figure S4.5 Root mean square (RMS) deviations of various aminoacyl-PCP-E didomain of GrsA. In various blue hues, RMS deviations determined over the course of three independent conventional MD simulation. In various hues of red, RMS deviations determined over the course of three independent Gaussian accelerated molecular dynamics simulations. In all cases, RMS deviations were determined after aligning the frames of each simulation using the first frame as a reference.
Figure S4.6 Root mean square (RMS) fluctuations of the various aminoacyl-PCP-E didomain of GrsA. Analysis was performed on the conventional MD simulation data only. In orange, RMS fluctuations of the backbone atoms of each residue of the didomain. In blue, RMS fluctuations of the sidechain atoms of each residue of the didomain. Residues 536-607, 608-627, 628-802, and 803-1067 constitute the peptidyl carrier protein, linker region, C-N-terminal subdomains of the epimerization domain, respectively.
Scheme S4.1 Distances between the catalytic groups of His$^{743}$, and Gly$^{748}$, Glu$^{892}$ and the reactive centers of the substrates. Distance a was measured as the distance between the center of mass of the imidazole group of His$^{743}$ and the carbonyl oxygen of the amino acyl group. Distance b was measured as the distance between the backbone amide nitrogen of Gly$^{748}$ and the carbonyl oxygen of the amino acyl group. Distance c was measured as the distance between the center of mass of the carboxylate group of Glu$^{892}$ and the a carbon of the peptidyl substrate.
Figure S4.7 Distances between catalytic residues and key substrate moieties sampled in cMD simulations. 2D histogram of simulation data of the complexes of aminoacyl-PCP-EGsA. Simulation data is binned along two coordinates (dimensions). The first of which is the distance between the carbonyl oxygen of the aminoacyl substrate and the center of geometry of the imidazole moiety of His753; the second measures the distance between the carboxylate group of Glu892 and the α carbon of its substrate. Bin widths for both coordinates are 0.125 Å. Color bar indicate the absolute population of each histogram bin.
Figure S4.8 Distances between catalytic residues and key substrate moieties sampled in GaMD simulations. 2D histogram of simulation data of the complexes of aminoacyl-PCP-E$_{GrsA}$. Simulation data is binned along two coordinates (dimensions). The first of which is the distance between the carbonyl oxygen of the aminoacyl substrate and the center of geometry of the imidazole moiety of His$^{753}$; the second measures the distance between the carboxylate group of Glu$^{892}$ and the $\alpha$ carbon of its substrate. Bin widths for both coordinates are 0.125 Å. Color bar indicate the absolute population of each histogram bin.

Efficient catalysis requires the disposition of the active site residues (His$^{753}$ and Glu$^{892}$) in close spatial proximity to the substrate. The 2D histograms shown in Figure S4.5 and Figure S4.6 illustrate each substrates’ ability to simultaneously engage His$^{753}$ and Glu$^{892}$. Simulation data is consistent with experimental findings that the L-phenylalanyl substrate is highly complementary in shape to the epimerization domain’s substrate binding pocket as Glu$^{892}$ is within 4.5 Å of the C$_\alpha$ of the L-phenylalanyl substrate, while His$^{753}$ is within hydrogen-bonding distance of the substrate in most of the simulation data. Epimerization of L-phenylalanyl in the absence of a downstream condensation domain is reversible, suggesting that D-phenylalanyl-PCP must form contacts with active site residues similar to those formed by L-phenylalanyl-PCP. In fact, during the course of our simulations, the D-phenylalanyl-PCP is observed to form a hydrogen bond with the substrate (as measured by the distance between aminoacyl carbonyl oxygen and the center of His$^{753}$ imidazolium moiety). However, D-phenylalanyl-PCP “disengages” from His$^{753}$ throughout a portion of the simulation data. Lastly, chlorovinylglycyl probe, likely due to its smaller side chain is able to sample a broader array of distances along both coordinates. The probe is much less effective at maintaining a hydrogen-bonding interaction with the catalytic histidine. Nonetheless, its C$_\alpha$ is similar proximal to Glu$^{892}$ as are the corresponding C$_\alpha$’s of L- and D-phenylalanyl-PCP. The inability of chlorovinylglycyl-PCP to maintain a hydrogen bond with His$^{753}$ may explain its limited efficacy as a crosslinker.
Figure S4.9 Distribution of C(O)-Cα dihedral angle of the aminoacyl group of the aminoacyl-PCP-E didomains sampled using cMD. Histograms of the S-C(O)-Cα-Cβ dihedral angles of L-Phe-PCP, D-Phe-PCP, and 1-PCP sampled within the epimerization domain of TycA. Data was sorted into bins with widths of 5°. (Data also shown in Figure 4.5 of the main text.)
Figure S4.10 Distribution of C-(O)-C$_2$ dihedral angle of the aminoacyl group of the aminoacyl-PCP-E didomains sampled using GaMD. Histograms of the S-C(O)-C$_2$-C$_3$ dihedral angles of L-Phe-PCP, D-Phe-PCP, and L-PCP sampled within the epimerization domain of TycA. Data was sorted into bins with widths of 5°.

Aminoacyl substrate assumes the same rotameric states during the course of conventional and Gaussian accelerated molecular dynamics simulations.
Figure S4.11 His753-thioester hydrogen bonding in L-Phe-PCP, D-Phe-PCP, and I-PCP rotamers sampled using cMD simulations. 2D histogram of simulation data of the complexes of aminoacyl-PCP-EgrsA. Simulation data is binned along two coordinates (dimensions). The first of which is the S-C(O)-Cα-Cβ torsion and the second is the distance between the carbonyl oxygen of the aminoacyl substrate and the center of mass of the imidazole moiety of His753. Data was sorted into bins with widths of 5° and 0.125 Å. Color bar indicate the absolute population of each histogram bin.
Figure S4.12  His753-thioester hydrogen bonding in L-Phe-PCP, D-Phe-PCP, and I-PCP rotamers sampled using cMD simulations. 2D histogram of simulation data of the complexes of aminoacyl-PCP-E_{GrsA}. Simulation data is binned along two coordinates (dimensions). The first of which is the S-C(O)-Cα-Cβ torsion and the second is the distance between the carbonyl oxygen of the aminoacyl substrate and the center of mass of the imidazole moiety of His\textsuperscript{753}. Data was sorted into bins with widths of 5° and 0.125 Å. Color bar indicate the absolute population of each histogram bin.
Acknowledgements: Chapter 4, in full, is a reprint of the material as it appears:

Section 4.6 References

(1) Fleming, A. Br. J. Exp. Pathol. 1929, 10, 226-236.


Chapter 5. Developing crosslinkers specific for epimerization domain to understand its mechanism

Section 5.1 Introduction

In nature, microorganisms and plants produce secondary metabolites as a function of survival and protection against other organisms. Among these natural products include a peptide secondary metabolites called nonribosomal peptides (NRPs), which are synthesized by large, multi-modular enzyme machineries called nonribosomal peptide synthetases (NRPSs). NRPs exhibit a wide variety of clinical and potent bioactivities for human health, functioning as antibiotics, immunosuppressants, and anticancer reagents. NRPSs are composed of several modules, which contain catalytic domains that are responsible for recognizing, activating, and incorporating one amino acid at a time in a linear fashion towards the final peptide. Protein-protein interactions between the peptidyl carrier protein (PCP), which is central to shuttling substrates within the pathway, and its domain partners are crucial for modules in NRPSs to precisely assemble natural products.

Interestingly, NRPSs can incorporate nonproteinogenic D-amino acids into the final natural products, creating a unique conformation. This conformation is responsible for enabling bioactivity as well as slowing down the degradation of peptide natural products from L-amino acid specific proteases. D-amino acids can be directly recognized and activated by an adenylation (A) domain and serve as a building block, as in the case of cyclosporine A biosynthesis. However, the majority of D-amino acids are converted from L-amino acids through the use of auxiliary epimerization (E) domains. Mutational studies of the E domain of GrsA (initiation module of the gramicidin S synthetase) revealed that
His753 and Glu892 act as a base/acid to deprotonate and reprotonate Cα-H. The epimerization reaction can happen both ways by converting L- to D-amino acid and D- to L-amino acid until it reaches equilibrium. Crystal structures of excised E domain and PCP-E didomain have been solved and support the prediction that it is similar in structure with the condensation (C) domain.7,8 PCP-E didomain structure was solved in apo- and holo-form to show protein interface and how PCP was oriented to deliver the 4'-phosphopantetheine arm towards the active site of the E domain.8 Despite the previous biochemical and structural studies, several questions related to E domain substrate recognition, binding, and detailed chemical mechanism remain to be addressed.

Herein, we describe the design, synthesis, and application of a crosslinking probe that can target the active site residues of the E domain, which allows capturing a PCP domain with the E domain. Previously, chlorovinylglycine, a mechanism-based inhibitor for alanine racemase, was demonstrated as a crosslinker between PCP and E domain. However, due to its smaller size compared to the natural substrate, chlorovinylglycine only revealed a slight crosslinking activity.9 In this paper, we utilized a sulfonate warhead to elucidate the biochemical mechanism of the epimerization domain.

Section 5.2 Results and Discussion

Tsukiji et al. published ligand-directed tosyl chemistry as a way to introduce synthetic probes nongenetically onto a protein in vivo.10 This method utilizes a protein ligand and synthetic probe that are connected by a phenylsulfonate group. The ligand binds tightly onto a protein where it situates the electrophilic phenylsulfonate moiety in close proximity to a nucleophilic amino acid for a chemical reaction to occur. Inspired by this
approach, we sought to utilize this sulfonate warhead to target the active site residues in the E domain. We reasoned that synthesizing a pantetheinamide sulfonate warhead as a crosslinker would allow loading the crosslinker onto a PCP, which will situate the sulfonate warhead towards the active site residues of an E domain.

The probe with a mesylate warhead (1) was designed to mimic the length of L-Phe-pantetheine to lie in the attacking site at the same position where the protonation/deprotonation of the natural substrate for the initiation module of tyrocidine synthetase, TycA (A-PCP-E), occurs. (Figure 5.1) The mesylate probe (1) was chemoenzymatically loaded onto the PCP of TycA using recombinant CoaA, CoaD, CoaE and 4’phosphopantetheinyl transferase (PPTase), followed by intramolecular crosslinking. The crosslinking activity was verified by running the samples onto an 8% SDS-PAGE gel. When compared to apo TycA, a crosslinked protein ran higher on a denaturing gel. We noticed that the crosslinked protein lane contained two different bands, crosslinking product and residual apo TycA. Chasing experiments after the crosslinking reaction with TAMRA-CoA revealed that the bottom band was unreacted apo TycA. (Not shown here) We found that the mesylate moiety can often times be degraded to form sulfene in the presence of a non-nucleophilic base. In our experiments, some mesylate warheads were degraded before performing any crosslinking activity, which explained the presence of apo TycA along with the crosslinked product.
Figure 5.1 Crosslinker design and its validation by running a reaction on a SDS-PAGE gel. (a) The design of the mesylate probe, C5-Ms (1), as a crosslinker between the PCP and E domain on TycA. (b) The crosslinking reaction was validated by running the samples on an 8% SDS-PAGE gel.

Due to these unexpected crosslinking results, we designed another set of probes with various linkers containing a benzene sulfonate moiety, which better mimics the natural substrate L-Phe of TycA and prevents its degradation. Interestingly, the crosslinking results showed two different populations of bands (A and B) that are present above apo TycA in different ratios depending on the linker lengths, (Figure 5.2), suggesting two possible crosslinking sites. C5-Ms (1) and C5-Bnsulfo (3), which mimics the length of the natural substrate, displayed similar crosslinking profiles, depicting the majority of the crosslinking bands at the A position. Conversely, C4-BnSulfo (2) showed its majority of crosslinking bands at the B position. Additionally, C6-BnSulfo (4) displayed similar crosslinking populations at both A and B positions.
In order to locate the actual crosslinking site, protease digestion followed by MALDI-TOF mass spectrometry was performed. When compared with apo TycA, we were able to identify two fragments that contained covalent crosslinking sites. One of the fragments was located on PCP and one on the E domain, which were only found from the crosslinked protein digestion (Figure 5.3). The fragment from the PCP contained Ser563, where the pantetheine probe is covalently loaded by phosphopantetheinyl transferase. The other fragment on the E domain contained several nucleophilic residues along with the catalytic residue Glu882, all of which were potential crosslinking sites. We generated six alanine mutants out of eight nucleophilic residues as well as the catalytic residue alanine mutants of the E domain, H743A and E882A. In addition, the double alanine mutant of the
catalytic residues, H743AE882A (HE), was generated to examine if the crosslinking was happening on both of the residues.

**Figure 5.3** Crosslinking protease digestion mass spectrometry analysis compared with wild-type protein. The crosslinked proteins were treated with trypsin and AspN digestion and analyzed by MALDI-TOF mass spectrometry. (A) The diagram showing the PCP side of the fragment comparing apo-TycA and crosslinked TycA. (B) The diagram showing the epimerization domain side of the fragment comparing apo-TycA and crosslinked TycA.
Crosslinking experiment on the alanine mutant library was performed with 3 and was run on an 8% SDS-PAGE gel. Interestingly, crosslinking reactions were only happening on the catalytic residue mutants and double mutant (Figure 5.4). There were two histidine residues present in the E domain fragment; however, the catalytic H743A was the only histidine residue that was able to react on the crosslinker. In the case of the wild-type TycA crosslinking reaction with 3, the results uncovered two crosslinking bands in comparison to apo TycA, in which, band A was identified to show a larger population than band B. To our surprise, crosslinking reactions done on the catalytic residues that were mutated to alanine residues, revealed that each represented one of the crosslinking bands that were shown in the wild-type crosslinking reaction. The crosslinking band for H743A identified only band A in the same thickness as it was shown in the wild-type crosslinking band A. Though the crosslinking band for E882A displayed only band B, it showed the increase in thickness compared to the wild-type crosslinking band B. When both of the catalytic residues were mutated to alanine residue (HE), it showed no crosslinking activity, which indicated that the crosslinking is solely happening on the catalytic residues H743 and E882. It can be speculated that for the crosslinking reaction on the H743A mutant, E882 acts as a nucleophile and for the crosslinking reaction on the E882A mutant, H743 acts as a nucleophile.
Figure 5.4 Crosslinking reaction on alanine mutant library with the probe 3. Schematic depiction of the crosslinking reaction with C5-BnSulfo probe on wild-type TycA (red), H743A TycA mutant (blue), E882ATycA mutant (green).

We have developed and synthesized a pantetheine analog crosslinking probe containing a sulfonyl warhead that targets catalytic residues of the epimerization domain in nonribosomal peptide synthetases. The mechanism of the epimerization reaction in NRPSs is proposed to occur with both catalytic histidine and glutamate residues by deprotonating/reprotonating Ca-H of an amino acid. However, no direct assay exists to measure the epimerization reaction and the subtlety of the reaction made it difficult to study the mechanism of the reaction. Previously, from structure-based calculations for protonation states of the epimerization domain of gramicidin S synthetase, catalytic residue H753 was calculated to have pKa 7.8-9.0. This pKa measurement shows that histidine remains in its protonated state; thus, suggesting that glutamate may act as a base. [Samel 2014, Kim 2018] This correlates with our data that in the wild-type TycA crosslinking, band A shows a thicker crosslinking band than band B and displays the same thickness when crosslinking reaction occurs on the H743A mutant, where E882 acts as a nucleophile.
However, when E882 is mutated to alanine, H743 acts as a nucleophile and the thickness of the band increases in comparison to the wild-type crosslinking band B. Since the epimerization domain can convert amino acids in both ways, L- to D-amino acids and D- to L-amino acids, both of the catalytic residues have to function as both acid and base.

**Section 5.3 Conclusions**

From this research, we were able to load a sulfonate warhead pantetheinamide probe onto a PCP domain and directly aim the catalytic residues of the epimerization domain to intermolecularly crosslink TycA. We revealed that both of the catalytic residues act as a base/nucleophile towards the probe, where glutamate acts stronger than histidine. When glutamate is mutated to alanine, histidine by itself can function as a base/nucleophile towards the probe. The development of this crosslinker can set a stage for the next step in the studying molecular basis of the epimerization domain in its modular settings and helping further combinatorial biosynthetic endeavors to incorporate D-amino acids into novel compounds.

**Section 5.4 Methods**

*Protein expression and purification*

Stocks of CoaA, CoaD, CoaE, and Sfp were prepared as described previously and stored at -80 °C. [Clarke 2004] Briefly, *apo*-TycA *E. coli* BL21(DE3) cells were grown in Terrific Broth at 37 °C with 30 µg/mL kanamycin and 100 µg/mL ampicillin (*apo* TycA). Cells were induced with 1 mM β-D-1-thiogalactopyranoside (IPTG) at OD$_{600}$ of 0.8 and incubated at 16 °C for 16 h. Cells were spun down by centrifugation at 2,000 rpm for 30
min and the collected pellets were lysed by sonication, followed by another centrifugation at 10,000 rpm for 1 h to clear the lysate. The His<sub>6</sub>-tagged protein was purified using a combination of Ni-NTA and size exclusion chromatography (Superdex 200 column, GE Healthcare). Purified proteins were concentrated to 8-10 mg/ml using a 100-kDa cut-off Amicon Ultra Centrifugation Filter (Millipore).

**Preparation of chemoenzymatic loading, crosslinking complex and purification**

The final reaction conditions for chemoenzymatic loading and crosslinking were: 12.5 mM MgCl<sub>2</sub>, 4 mM ATP, 0.01 mg/ml maltose-binding protein-CoA biosynthetic enzyme A (MBP-CoaA), 0.01 mg/ml MBP-CoaD, 0.01 mg/ml MBP-CoaE, 0.04 mg/ml Sfp (native), 0.1 mM crosslinking probe, and 1 mg/ml TycA in reaction buffer (50 mM HEPES, 250 mM NaCl, 0.5 mM TCEP, pH 7.4). After the addition of all components, reactions were gently mixed and incubated at room temperature for 12 hr. The reaction was purified by size exclusion chromatography (Superdex S200 column, GE Healthcare). The eluted protein was collected and concentrated using 50-kDa cut-off Amicon Ultra Centrifuge Filters (Millipore). The loading and crosslinking activity was monitored running an 8% SDS-PAGE gel.

**Generation of mutant constructs**

The His<sub>6</sub>-tagged TycA in pQE60 plasmid was used as the DNA template for site-directed mutagenesis with the mutagenic primer listed in Table S1. PCRs were performed using varying annealing temperatures. TycA PCR products were digested with Dpn1 to eliminate template wildtype TycA. Digested PCR products were transformed into pREP4. Single colonies were isolated and their plasmids were extracted using GenElute miniprep kit (Sigma). Site-directed mutagenesis was verified by DNA sequencing (Genewiz).
Section 5.5 Supplementary Information

Chemical synthetic procedures

Scheme S5.1 Synthetic route to C4-Ms (1). Reagents and conditions: [a] 4-amino-1-butanol, HATU, HOAT, DIEA, CH₂Cl₂, rt, 97%; [b] Methanesulfonyl chloride, Et₃N, CH₂Cl₂, rt, 99%; [c] 1 M aq. HCl, THF, rt, 99%.

Scheme S5.2 Synthetic route to C5-Ms (2). Reagents and conditions: [a] 5-amino-1-pentanol, HATU, HOAT, DIEA, CH₂Cl₂, rt, 89%; [b] Methanesulfonyl chloride, Et₃N, CH₂Cl₂, rt, 93%; [c] 1 M aq. HCl, THF, rt 76%.
Scheme S5.3 Synthetic route to C4-Bsulfo (3). Reagents and conditions: [a] 4-amino-1-butanol, HATU, HOAT, DIEA, CH₂Cl₂, rt, 97%; [b] Benzenesulfonyl chloride, Et₃N, CH₂Cl₂, rt, 93%; [c] 80% aq. Acetic acid, rt, 46%.

Scheme S5.4 Synthetic route to C5-Bsulfo (4). Reagents and conditions: [a] 5-amino-1-pentanol, HATU, HOAT, DIEA, CH₂Cl₂, rt, 89%; [b] Benzenesulfonyl chloride, Et₃N, rt, 90%; [c] 80% aq. Acetic acid, rt, 63.
Scheme S5.5 Synthetic route to C6-Bsulfo (5). *Reagents and conditions:* [a] 6-amino-1-hexanol, HATU, HOAT, DIEA, CH₂Cl₂, rt, 84%; [b] Benzenesulfonyl chloride, Et₃N, CH₂Cl₂, rt, 93%; [c] 80% aq. Acetic acid, rt, 69%.

**General synthetic methods**

All commercial reagents were used as provided unless otherwise indicated. Compound S1 is known compound. This compound was prepared according to published literature procedures.  All of the reactions were carried out under an atmosphere of nitrogen in dry solvents with oven-dried glassware and constant magnetic stirring unless otherwise noted. ¹H-NMR spectra were recorded at 500 MHz. ¹³C-NMR spectra were recorded at 125 MHz on JEOL NMR spectrometers and standardized to the NMR solvent signal as reported by Gottlieb. Multiplicities are given as s = singlet, d = doublet, t = triplet, p = pentet, m = multiplet using integration and coupling constant in Hertz. TLC analysis was performed using Silica Gel 60 F254 plates (Merck) and visualization was accomplished with ultraviolet light (λ = 254 nm) and/or the appropriate stain [phosphomolybdic acid, iodine, ninhydrin, and potassium permanganate]. Silica gel chromatography was carried out with SiliaFlash F60 230-400 mesh (Silicycle), according
Mass spectral data were obtained using a LCMS-IT-TOF mass spectrometer (Shimadzu).

Chemical synthesis of C4-Ms (1)
Compound number in bold refers to the structures shown in Scheme S5.1

**Compound S2**

HATU (449 mg, 1.18 mmol), 1-hydroxybenzotriazole (180 mg, 1.18 mmol), and N-ethyldiisopropylamine (124 µL, 0.71 mmol) were added to a solution of compound S1 (200 mg, 0.59 mmol) in CH₂Cl₂ (20 mL). The solution was stirred at room temperature for 10 min and 4-amino-1-butanol (66 µL, 0.71 mmol) was then added. After 12h, the reaction mixture was washed with 5% citric acid, 5% NaHCO₃, and brine. The organic layer was dried over Na₂SO₄ and evaporated to dryness. The residue was purified by flash chromatography (EtOAc to 95:5 EtOAc/MeOH) to afford compound S2 as a colorless oil (242 mg, 97%). ¹H NMR (500 MHz, CD₃OD): δ 7.45 (d, J = 8.9 Hz, 2H), 6.92 (d, J = 8.9 Hz, 2H), 5.51 (s, 1H), 4.13 (s, 1H), 3.79 (s, 3H), 3.69 (q, J = 10.1 Hz, 2H), 3.56–3.40 (m, 4H), 3.19–3.11 (m, 2H), 2.40 (t, J = 6.6 Hz, 2H), 1.58–1.46 (m, 4H), 1.10 (s, 3H), 1.04 (s, 3H). ¹³C NMR (125 MHz, CD₃OD): δ 173.5, 171.6, 161.6, 131.8, 128.8, 114.5, 102.6, 85.0, 79.2, 62.4, 55.7, 40.2, 36.4, 36.3, 34.0, 30.9, 26.8, 22.2, 19.7. HRMS (ESI+): [M+H]⁺ calcd for C₂₁H₃₂N₂NaO₆, 431.2158; found, 431.2010.

**Compound S3**

Methanesulfonyl chloride (45 µL, 0.58 mmol) and triethylamine (101 µL, 0.73 mmol) were added to a solution of compound S2 (123 mg, 0.29 mmol) in CH₂Cl₂ (10 mL). The solution
was stirred at 0 °C. After 30 min, the reaction mixture was diluted with CH₂Cl₂. The mixture was washed with 5% citric acid, 5% NaHCO₃, and brine. The residue was purified by flash chromatography (EtOAc) to afford compound S3 as a white solid (155 mg, 99%).

1H NMR (500 MHz, CDCl₃): δ 7.40 (d, J = 8.6 Hz, 2H), 6.89 (d, J = 8.6 Hz, 2H), 5.43 (s, 1H), 4.16 (t, J = 6.3 Hz, 1H), 3.78 (s, 3H), 3.69–3.60 (m, 2H), 3.56–3.42 (m, 2H), 3.28–3.16 (m, 2H), 2.95 (s, 3H), 2.41 (t, J = 6.3 Hz, 2H), 1.78–1.66 (m, 2H), 1.64–1.50 (m, 2H), 1.06 (s, 3H), 1.04 (s, 3H).


C4-Ms (1)

Compound S3 (88 mg, 0.18 mmol) was dissolved in 1 M aq. HCl (3 mL) and THF (3 mL) and stirred until the starting material was consumed as shown by TLC. The reaction was then neutralized by addition of AG-1-X8 strong basic anionic exchange resin. After filtration, the solvent was evaporated under reduced pressure. The residue was purified by flash chromatography (95:5 EtOAc/acetone to acetone) to afford compound C4-Ms (1) as a colorless oil (67 mg, 99%).

1H NMR (500 MHz, (CD₃)₂CO): δ 4.25 (t, J = 6.3 Hz, 2H), 3.93 (s, 1H), 3.55–3.38 (m, 4H), 3.24 (q, J = 6.3 Hz, 2H), 3.09 (s, 3H), 2.41, (t, J = 6.3 Hz, 2H), 1.80–1.72 (m, 2H), 1.64–1.56 (m, 2H), 0.94 (s, 3H), 0.86 (s, 3H).


Chemical synthesis of C5-Ms (2)

Compound number in bold refers to the structures shown in Scheme S5.2

Compound S4
HATU (2 g, 4.74 mmol), 1-hydroxybenzotriazole (726 mg, 4.74 mmol), and N-ethyldiisopropylamine (825 µL, 4.74 mmol) were added to a solution of compound S1 (800 mg, 2.37 mmol) in CH₂Cl₂ (100 mL). The solution was stirred at room temperature for 10 min and 5-amino-1-pentanol (292 mg, 2.84 mmol) was then added. After 10 h, the reaction mixture was washed with 5% citric acid, 5% NaHCO₃, and brine. The organic layer was dried over Na₂SO₄ and evaporated to dryness. The residue was purified by flash chromatography (EtOAc to 90:10 EtOAc/MeOH) to afford compound S4 as a colorless oil (890 mg, 89%). ¹H NMR (500 MHz, CD₂OD): δ 7.45 (d, J = 8.6 Hz, 2H), 6.93 (d, J = 8.6 Hz, 2H), 5.53 (s, 1H), 4.14 (s, 1H), 3.80 (s, 3H), 3.75–3.65 (m, 2H), 3.56–3.40 (m, 4H), 3.18–3.10 (m, 2H), 2.40 (t, J = 6.6 Hz, 2H), 1.56–1.44 (m, 4H), 1.40–1.31 (m, 2H), 1.10 (s, 3H), 1.04 (s, 3H). ¹³C NMR (125 MHz, CD₂OD): δ 173.5, 171.6, 161.7, 131.8, 128.8, 114.5, 102.6, 85.1, 79.3, 62.8, 40.4, 36.4, 36.3, 34.0, 33.2, 30.2, 24.3, 22.2, 19.7. HRMS (ESI+): [M+H]⁺ calcd for C₂₂H₃₅N₂O₆, 423.2495.

**Compound S5**

Methanesulfonyl chloride (325 µL, 4.2 mmol) and triethylamine (731 µL, 5.25 mmol) were added to a solution of compound S4 (890 mg, 2.10 mmol) in CH₂Cl₂ (50 mL). The solution was stirred at 0 °C. After 2 h, the reaction mixture was diluted with CH₂Cl₂. The mixture was washed with 5% citric acid, 5% NaHCO₃, and brine. The residue was purified by flash chromatography (EtOAc) to afford compound S5 as a white solid (975 mg, 93%). ¹H NMR (500 MHz, CDCl₃): δ 7.40 (d, J = 8.6 Hz, 2H), 6.89 (d, J = 8.6 Hz, 2H), 5.43 (s, 1H), 4.17 (t, J = 6.3 Hz, 2H), 4.04 (s, 1H), 3.79 (s, 3H), 3.66 (q, J = 11.2 Hz, 2H), 3.59–3.41 (m, 2H), 3.18 (q, J = 6.7 Hz, 2H), 2.96 (s, 3H), 2.38 (t, J = 6.3 Hz, 2H), 1.70 (qui, J = 7.0 Hz, 2H), 1.48 (qui, J = 7.3 Hz, 2H), 1.38 (qui, J = 7.3 Hz, 2H), 1.06 (d, J = 5.7 Hz, 6H). ¹³C NMR (125 MHz, CDCl₃): δ 170.9, 169.6, 160.3, 130.2, 127.6, 113.8, 101.3, 83.8, 78.5, 69.9, 55.4, 39.1, 37.4, 36.1, 35.0, 33.1, 28.9, 28.7, 22.8, 21.9, 19.2. HRMS (ESI+): [M+H]⁺ calcd for C₂₃H₃₇N₂O₈S, 501.2271.
C5-Ms (2)

Compound S5 (975 mg, 1.95 mmol) was dissolved in 1 M aq. HCl (10 mL) and THF (10 mL) and stirred until the starting material was consumed as shown by TLC. The reaction was then neutralized by addition of AG-1-X8 strong basic anionic exchange resin. After filtration, the solvent was evaporated under reduced pressure. The residue was purified by flash chromatography (90:10 EtOAc) to afford compound C5-Ms (2) as a colorless oil (565 mg, 76%).

$^1$H NMR (500 MHz, (CD$_3$)$_2$CO): $\delta$ 4.27–4.17 (m, 3H), 3.61 (d, $J = 5.1$ Hz, 1H), 3.56–3.35 (m, 4H), 3.20 (q, $J = 6.5$ Hz, 2H), 2.39 (q, $J = 6.3$ Hz, 2H), 1.75 (p, $J = 7.0$ Hz, 2H), 1.53 (p, $J = 7.2$ Hz, 2H), 1.44 (p, $J = 7.2$ Hz, 2H), 0.94 (s, 3H), 0.86 (s, 3H).

$^{13}$C NMR (125 MHz, (CD$_3$)$_2$CO): $\delta$ 174.1, 171.6, 77.5, 71.1, 70.8, 40.2, 39.4, 37.0, 36.2, 35.9, 29.8, 29.5, 23.5, 21.9, 20.5. HRMS (ESI+): [M+H]$^+$ calcd for C$_{15}$H$_{30}$N$_2$NaO$_7$S, 405.1671; found, 405.1653.

Chemical synthesis of C4-Bsulfo (3)
Compound number in bold refers to the structures shown in Scheme S5.3

Compound S6

Benzenesulfonyl chloride (47 µL, 0.37 mmol) and triethylamine (102 µL, 0.73 mmol) were added to a solution of compound S2 (100 mg, 0.25 mmol) in CH$_2$Cl$_2$ (10 mL). The solution was stirred at rt overnight. The reaction mixture was diluted with CH$_2$Cl$_2$ and H$_2$O. The organic phase was separated, washed with brine, dried and concentrated in vacuo. The oil was purified by flash chromatography (CH$_2$Cl$_2$:EtOAc = 5:3) to afford compound S6 as an oil (125 mg, 93%).

$^1$H NMR (500 MHz, CDCl$_3$): $\delta$ 7.89 (dd, $J = 8.2$, 1.4 Hz, 2H), 7.69–7.62 (m, 1H), 7.59–7.52 (m, 2H), 7.45–7.39 (m, 2H), 7.05–6.98 (m, 1H), 6.96–6.88 (m, 2H), 6.12 (s, 1H), 5.46 (s, 1H), 4.08 (s, 1H), 4.03 (q, $J = 6.1$ Hz, 2H), 3.81 (d, $J = 0.9$ Hz, 3H), 3.74–3.63 (m, 2H), 3.52 (p, $J = 6.6$ Hz, 2H), 3.25–3.11 (m, 2H), 2.41 (q, $J = 7.3$, 6.2 Hz, 2H), 1.65 (dq, $J = 8.2$, 6.1 Hz, 2H), 1.52 (q, $J = 7.1$ Hz, 2H), 1.09 (d, $J = 4.2$ Hz, 3H).
Hz, 6H). $^{13}$C NMR (125 MHz, CDCl$_3$): $\delta$ 170.99, 169.85, 160.38, 133.93, 129.43, 127.95, 127.64, 113.88, 101.47, 83.92, 78.57, 70.33, 55.48, 38.75, 36.36, 35.05, 33.22, 26.41, 25.74, 21.97, 19.26. HRMS (ESI+): [M+H]$^+$ calcd for C$_{27}$H$_{36}$N$_2$O$_8$S, 549.2192; found, 549.2262.

**C4-Bsulfo (3)**

Compound S6 (39 mg, 0.07 mmol) was dissolved in 80% aq. Acetic acid (1.5 mL) stirred until the starting material was consumed as shown by TLC. The solvent was evaporated under reduced pressure. The residue was purified by flash chromatography (EtOAc:CH$_2$Cl$_2$ = 5:3) to afford compound **C4-Bsulfo (3)** as a colorless oil (14 mg, 46%). $^1$H NMR (500 MHz, CDCl$_3$): $\delta$ 7.95 – 7.85 (m, 2H), 7.67 (t, $J$ = 7.4 Hz, 1H), 7.57 (t, $J$ = 7.6 Hz, 2H), 7.38 (q, $J$ = 6.2 Hz, 1H), 6.26 (t, $J$ = 5.8 Hz, 1H), 4.06 (t, $J$ = 6.1 Hz, 2H), 3.99 (s, 1H), 3.55 (q, $J$ = 6.0 Hz, 2H), 3.48 (s, 2H), 3.23 (ddq, $J$ = 19.2, 13.6, 6.8 Hz, 2H), 2.43 (q, $J$ = 6.0 Hz, 2H), 1.67 (ddd, $J$ = 16.4, 9.9, 6.7 Hz, 2H), 1.56 (p, $J$ = 6.5, 6.0 Hz, 2H), 1.00 (s, 3H), 0.91 (s, 3H). $^{13}$C NMR (125 MHz, CDCl$_3$): $\delta$ 173.81, 171.68, 135.86, 134.08, 129.50, 127.94, 71.06, 70.56, 44.70, 39.47, 38.88, 35.96, 35.37, 29.90, 27.01, 26.46, 25.75, 21.58, 20.52. HRMS (ESI+): [M+H]$^+$ calcd for C$_{19}$H$_{30}$N$_2$O$_7$S, 431.1774; found, 431.1847.

**Chemical synthesis of C5-Bsulfo (4)**

*Compound number in bold refers to the structures shown in Scheme S5.4*

**Compound S7**

Benzensulfonyl chloride (45 µL, 0.36 mmol) and triethylamine (99 µL, 0.71 mmol) were added to a solution of compound S4 (100 mg, 0.24 mmol) in CH$_2$Cl$_2$ (10 mL). The solution was stirred at rt overnight. The reaction mixture was diluted with CH$_2$Cl$_2$ and H$_2$O. The organic phase was separated, washed with brine, dried and concentrated in vacuo. The oil was purified by flash chromatography (CH$_2$Cl$_2$:EtOAc = 5:3) to afford compound S7 as an oil (117 mg, 90%). $^1$H NMR (500 MHz, CDCl$_3$): $\delta$ 7.93 – 7.85 (m, 2H), 7.69 – 7.61 (m,
1H, 7.56 – 7.51 (m, 2H), 7.40 (d, J = 8.2 Hz, 2H), 7.04 (t, J = 6.3 Hz, 1H), 6.89 (d, J = 8.1 Hz, 2H), 6.20 (t, J = 5.7 Hz, 1H), 5.44 (s, 1H), 4.05 (s, 1H), 4.00 (t, J = 6.3 Hz, 2H), 3.79 (s, 3H), 3.71 – 3.61 (m, 2H), 3.50 (dp, J = 20.3, 6.9 Hz, 2H), 3.17 – 3.10 (m, 2H), 2.38 (t, J = 6.2 Hz, 2H), 1.61 (p, J = 6.6 Hz, 2H), 1.40 (p, J = 7.1 Hz, 2H), 1.31 (q, J = 6.9, 5.7 Hz, 2H), 1.06 (d, J = 3.2 Hz, 6H).


C5-Bsulfo (4)

Compound S7 (24 mg, 0.04 mmol) was dissolved in 80% aq. Acetic acid (1 mL) stirred until the starting material was consumed as shown by TLC. The solvent was evaporated under reduced pressure. The residue was purified by flash chromatography (EtOAc:CH₂Cl₂ = 5:3) to afford compound C5-Bsulfo (4) as a colorless oil (12 mg, 63%). ¹H NMR (500 MHz, (CDCl₃): δ 7.92 – 7.89 (m, 2H), 7.66 (d, J = 7.5 Hz, 1H), 7.57 (dd, J = 8.5, 7.2 Hz, 2H), 7.35 (d, J = 8.5 Hz, 1H), 5.95 (s, 1H), 4.05 (t, J = 6.2 Hz, 2H), 4.00 (s, 1H), 3.58 (q, J = 6.2 Hz, 2H), 3.53 – 3.47 (m, 2H), 3.22 (dq, J = 19.2, 6.8 Hz, 2H), 2.46 – 2.41 (m, 2H), 1.69 (d, J = 6.9 Hz, 2H), 1.52 – 1.45 (m, 2H), 1.38 (tt, J = 9.4, 5.6 Hz, 2H), 1.02 (s, 3H), 0.92 (s, 3H). ¹³C NMR (125 MHz, (CDCl₃): δ 173.57, 171.54, 154.08, 135.98, 134.08, 129.39, 127.97, 77.90, 77.73, 71.12, 70.77, 39.52, 39.31, 35.98, 35.35, 28.89, 28.55, 22.85, 21.75, 20.51. HRMS (ESI+): [M+H]+ calcd for C₂₀H₃₂N₂O₇S, 445.1930; found, 445.2002.

Chemical synthesis of C6-Bsulfo (5)

Compound number in bold refers to the structures shown in Scheme S5.5

Compound S8
HATU (733 mg, 1.93 mmol), and N-ethyldiisopropylamine (520 µL, 2.96 mmol) were added to a solution of compound S1 (500 mg, 1.48 mmol) in CH₂Cl₂ (25 mL). The solution was stirred at room temperature for 10 min and 6-amino-1-hexanol (174 mg, 1.48 mmol) was then added. After 12 h, the reaction mixture was washed with 5% NaHCO₃, and brine. The organic layer was dried over Na₂SO₄ and evaporated to dryness. The residue was purified by flash chromatography (EtOAc to 95:5 EtOAc/MeOH) to afford compound S8 as a colorless oil (543 mg, 84%). ¹H NMR (500 MHz, CDCl₃): δ 7.47 – 7.41 (m, 2H), 7.02 (s, 1H), 6.97 – 6.90 (m, 2H), 5.95 (s, 1H), 5.46 (s, 1H), 4.08 (s, 1H), 3.82 (d, J = 1.7 Hz, 3H), 3.70 (dd, J = 13.9, 7.9 Hz, 2H), 3.62 (q, J = 6.7, 6.1 Hz, 2H), 3.51 (ddt, J = 19.9, 13.1, 6.3 Hz, 1H), 3.20 (dq, J = 12.0, 6.9 Hz, 2H), 3.12 (s, 1H), 2.41 (t, J = 6.2 Hz, 2H), 1.64 (s, 2H), 1.59 – 1.28 (m, 10H), 1.09 (d, J = 2.2 Hz, 6H). ¹³C NMR (125 MHz, CDCl₃): δ 171.10, 169.81, 160.25, 130.12, 127.70, 113.86, 101.33, 83.77, 78.46, 62.51, 55.34, 43.29, 39.39, 36.18, 35.29, 33.14, 32.44, 29.33, 26.47, 25.26, 21.95, 19.26, 12.61. HRMS (ESI+): [M+Na]⁺ calcd for C₂₃H₃₆N₂O₆Na, 459.2471; found, 459.2462.

**Compound S9**

Benzenesulfonyl chloride (100 µL, 0.61 mmol) and triethylamine (170 µL, 1.22 mmol) were added to a solution of compound S8 (177 mg, 0.41 mmol) in CH₂Cl₂ (10 mL). The solution was stirred at rt overnight. The reaction mixture was diluted with CH₂Cl₂ and H₂O. The organic phase was separated, washed with brine, dried and concentrated in vacuo. The oil was purified by flash chromatography (CH₂Cl₂:EtOAc = 5:3) to afford compound S9 as an oil (217 mg, 93%). ¹H NMR (500 MHz, CDCl₃): δ 7.90 (d, J = 7.2 Hz, 2H), 7.68 – 7.62 (m, 1H), 7.55 (dd, J = 8.4, 7.0 Hz, 2H), 7.45 – 7.39 (m, 2H), 7.00 (s, 1H), 6.94 – 6.88 (m, 2H), 5.91 (s, 1H), 5.45 (s, 1H), 4.07 (s, 1H), 4.02 (t, J = 6.3 Hz, 2H), 3.81 (s, 3H), 3.68 (q, J = 11.4 Hz, 2H), 3.53 (dq, J = 19.8, 6.9 Hz, 1H), 3.16 (q, J = 6.7 Hz, 2H), 2.40 (t, J = 6.2 Hz, 2H), 1.68 (s, 2H), 1.65 – 1.59 (m, 2H), 1.41 (q, J = 7.2 Hz, 2H), 1.33 – 1.27 (m, 4H), 1.09 (s, 6H). ¹³C NMR (125 MHz, CDCl₃): δ 170.82, 169.73, 160.36, 133.85, 130.25, 129.38, 127.95, 127.63, 122.54, 113.87, 101.46, 83.95, 78.60, 70.78, 60.54, 55.47, 39.40, 36.37, 35.10, 33.22, 29.41, 28.81, 26.25, 25.10, 21.98, 21.20, 19.26, 14.33. HRMS (ESI+): [M+H]⁺ calcd for C₂₈H₃₆N₂O₈S, 577.2505; found, 577.2580.
Compound S9 (82 mg, 0.14 mmol) was dissolved in 80% aq. Acetic acid (2 mL) stirred until the starting material was consumed as shown by TLC. The solvent was evaporated under reduced pressure. The residue was purified by flash chromatography (EtOAc:CH₂Cl₂ = 5:3) to afford compound **C6-Bsulfo (5)** as a colorless oil (45 mg, 69%). **1H NMR (500 MHz, (CDCl₃): δ 7.90 (dt, J = 7.2, 1.4 Hz, 2H), 7.69 – 7.63 (m, 1H), 7.56 (dd, J = 8.4, 7.0 Hz, 2H), 7.42 (t, J = 6.2 Hz, 1H), 6.18 (t, J = 5.9 Hz, 1H), 4.19 (d, J = 19.0 Hz, 1H), 4.04 (t, J = 6.3 Hz, 2H), 3.98 (s, 1H), 3.55 (qt, J = 6.2, 2.6 Hz, 2H), 3.48 (s, 2H), 3.18 (q, J = 6.7 Hz, 2H), 2.42 (t, J = 6.0 Hz, 2H), 1.93 (s, 2H), 1.70 – 1.60 (m, 2H), 1.45 (p, J = 7.1 Hz, 2H), 1.38 – 1.22 (m, 5H), 0.99 (s, 3H), 0.91 (s, 3H). **13C NMR (125 MHz, (CDCl₃): δ 173.75, 171.60, 136.12, 133.95, 129.43, 127.93, 77.76, 71.09, 70.90, 39.46, 35.92, 35.40, 29.32, 28.78, 26.20, 25.11, 20.56.** HRMS (ESI+): [M+H]⁺ calcd for C₂₁H₃₄N₂O₇S, 459.2087; found, 459.2155.

**Table S5.1** Primers used in site-directed mutagenesis
*Nucleotides that corresponds to the mutated amino acid residues are bolded.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>H743A</td>
<td>AATTCATGCGTTGGTCGTG GATGGCATTTCT</td>
<td>CAACGCATGAATTTCGAGAAACAAATG ATCGCC</td>
</tr>
<tr>
<td>D858A</td>
<td>CAACGCCGCTGTGTGTGGCCGC</td>
<td>AGCGCGGTTGATTTCCGTTGGATAGGCCCT GATTG</td>
</tr>
<tr>
<td>E870A</td>
<td>TTGCGGCCGTGGAGCAAGCT TGCGCA</td>
<td>ACGGCCGAAAAGCCAAGCCGAGCG</td>
</tr>
<tr>
<td>S872A</td>
<td>GGGCAGAAGCTTGCGCAA ATCGTCATTTGCGAGGGG</td>
<td>GCTTCGCCACTCGCAGGAAAGCCAAGGC</td>
</tr>
<tr>
<td>K873A</td>
<td>CGCGCTTGCGCAAATCGTC ATTCATTTGCGAGGGG</td>
<td>GCAAGCGCGCTCCACTCAGCAGAAGG</td>
</tr>
<tr>
<td>Q876A</td>
<td>GGCGATCGTCATTTGCGAGGGCAGG</td>
<td>GACGATCGCGCAAGCTTGCTCCACTC</td>
</tr>
<tr>
<td>H880A</td>
<td>GTCTTGGAGAGGGGCACGAGCC</td>
<td>CAACGCATGAGAATTTCGAGAAACAAATGGCTCC</td>
</tr>
<tr>
<td>E882A</td>
<td>TCATTTGGCGCGGACAGGGGC</td>
<td>CGCGAAATGAAATGAGAACGAGTTTGC</td>
</tr>
<tr>
<td>H887A</td>
<td>GGCGACATCAGAAGCGAA</td>
<td>GTCCGCAGCGCCGAGTCC</td>
</tr>
</tbody>
</table>

133
Figure S5.1 Full SDS-PAGE gel image of TycA crosslinking constructs with different probes. Full 8% SDS-PAGE gel image corresponding to the main text Figure 5.2.
Figure S5.2 Full SDS-PAGE gel image of TycA mutant crosslinking constructs with C5-Bsulfo probe. Full 8% SDS-PAGE gel image corresponding to the main text Figure 5.4.

Figure S5.3 SDS-PAGE gel showing TycA alanine mutant constructs.
Figure S5.4 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S2 in CD$_3$OD
Figure S5.5 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S3 in CDCl$_3$
Figure S5.6 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C4-Ms in (CD$_3$)$_2$CO
Figure S5.7 ¹H-NMR (500 MHz) and ¹³C-NMR (125 MHz) spectra of S4 in CD₃OD
Figure S5.8 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S5 in CDCl$_3$
Figure S5.9 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C5-Ms in (CD$_3$)$_2$CO
Figure S5.10 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S6 in CDCl$_3$
Figure S5.11 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C4-BSulfo in CDCl$_3$
Figure S5.12 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S7 in CDCl$_3$
Figure S5.13 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C5-BSulfo in CDCl$_3$. 
Figure S5.14 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S8 in CDCl$_3$. 
Figure S5.15 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of S9 in CDCl$_3$
Figure S5.16 $^1$H-NMR (500 MHz) and $^{13}$C-NMR (125 MHz) spectra of C6-BSulfo in CDCl$_3$. 

C6-BSulfo
Acknowledgements: Chapter 5, in full, is currently being prepared for submission for publication as: Kim, W. E.*, Ishikawa, F.*, Re, R. N., Suzuki, T., Dohmae, N., Kakeya, H., Tanabe, G., Burkart, M. D. “Developing Crosslinkers Specific for Epimerization Domain to Understand its Mechanism” The dissertation author is the primary author of this manuscript.

Section 5.5 References


