Investigations of the Biosynthesis and Structure of Colibactin, a Cytotoxin Made by Human-Associated Escherichia Coli

Investigations of the Biosynthesis and Structure of Colibactin, a Cytotoxin Made by Human-associated *Escherichia coli*

A dissertation presented

by

Carolyn Adams Brotherton

to

The Department of Chemistry and Chemical Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemistry

Harvard University

Cambridge, Massachusetts

December 2015

Dissertation advisor: Professor Emily P. Balskus Carolyn Adams Brotherton

Investigations of the Biosynthesis and Structure of Colibactin, a Cytotoxin Made by Human-associated *Escherichia coli*

Abstract

Humans exist in symbiosis with trillions of bacteria that are collectively referred to as the human microbiota. While commensal microbes are essential for health, some resident microbes can promote disease. Certain strains of human-associated *Escherichia coli* cause double-strand breaks in host DNA through the production of colibactin, a genotoxin of unknown structure. To broaden our understanding of the chemistry of the human microbiota, we sought to elucidate the structure of colibactin and characterize its biosynthetic pathway.

We first characterized the self-resistance mechanism in colibactin biosynthesis (Chapter 2). We found that the enzyme that initiates the assembly line pathway is ClbN, a non-ribosomal peptide synthetase (NRPS). Biochemical assays showed that ClbN biosynthesizes an N-terminal prodrug motif consisting of an *N*-myristoyl-D-asparagine residue that is proposed to mask the reactivity of colibactin. We performed bioinformatic analyses to identify the enzyme that acts after ClbN in the assembly line. *In vitro* reconstitution assays revealed that ClbB, a NRPS/ polyketide synthase (PKS) hybrid, elongates the prodrug motif produced by ClbN. In addition, we demonstrated that the periplasmic peptidase ClbP cleaves the prodrug motif from synthesized model substrates and that the membrane domain of ClbP is required for full activity.

Next, we sought to isolate precolibactin, the prodrug-containing precursor to the active genotoxin (Chapter 3). Metabolite profiling of the extracted metabolome of Δ*clbP* and wild-type colibactin-producing strains led to the identification of several candidate precolibactins. We isolated one of these metabolites, which we named Metabolite B. This metabolite contains an unusual azaspiro[2.4] bicyclic ring system that has not been observed previously in a natural

product scaffold. The structure of Metabolite B suggested that the colibactin assembly line pathway may utilize novel biosynthetic logic and pointed to a possible mechanism through which colibactin damages DNA.

Finally, we describe the isolation of other *pks*-associated metabolites and provide biosynthetic hypotheses for these pathway intermediates (Chapter 4). We also present our attempts to characterize the next enzymatic steps in the colibactin biosynthetic pathway. As part of these efforts, the PKS module of ClbB was characterized *in vitro* and several other colibactin biosynthetic enzymes were examined using genetic studies. The challenges associated with studying a biosynthetic pathway for which the final product is unknown are highlighted. Overall, the work presented here contributes significantly to our knowledge of the biosynthesis and structure of a microbial genotoxin.

Acknowledgements

John Donne's words "No man is an island/ Entire of itself" reflect my sentiments about my dissertation research. While working in the laboratory has been a sometimes lonely activity, obtaining my doctoral degree has been by no means a solitary exercise. I owe so much to many people and I will try to thank all of these wonderful mentors, colleagues and friends here.

I want to thank Professor Scott Miller for encouraging me to pursue research as an undergraduate at Yale. My goal in pursuing my PhD was to become as thoughtful a scholar as Scott and the graduate students in his lab, like Peter Jordan and Phil Lichtor. I am greatly appreciative of the guidance that my dissertation advisor Professor Emily Balskus has provided to me over the last five years. Emily believed in me when I wanted to completely change my field of research and join her lab in my second year of graduate school. The first few months in the Balskus group were an extremely exciting time, as I dived into the natural products literature, learned new techniques and happily realized that I had found myself in the right area of study.

I am grateful for the assistance of Sunia Trauger and Gary Byrd in LC–MS studies described in Chapters 2–4 and that of Gregory Heffron in NMR studies described in Chapter 3.

The best part of graduate school has been the people I've been so lucky to know. First, I want to thank my colleagues and friends in the Balskus group. I have to thank Li Zha, Matthew Wilson, Yindi Jiang and Ethan Winter for continuing to study colibactin. I have enjoyed the sometimes strange but usually entertaining conversations with Li. Matt's kindness, patience and willingness to help has been invaluable in stressful times. Stephen Wallace, Yolanda Huang, Benjamin Schneider and Abraham Waldman: we really have been the best bay. I always felt I could swivel around in my chair and ask science questions or discuss various subjects including–but not

limited to–the merits of post-doctoral training and the best way to lift weights. Also, thanks for the gum. Thanks go to Stephen and Spencer Peck for reading and editing my thesis. A warm thanks to Abraham for all of the science advice and fun times, in and out of lab, Nitzan Koppel for her weird sense of humor and musical skills, both at Courtside and at Porchfest, and Kristen Seim for all of the griping sessions and ambitious projects outside of the lab.

I want to say thanks to my incredible friends outside of the lab. I'm grateful for the superb conversational skills and insights of Alexandra Cantley. I'm thankful for my friendship with Noam Prywes, a true science guy with an ability to talk science at any time of day or state of mind. Thanks to Kyle Strom for teaching me about the science of home brewing and for being the super cool dude who can play the saw that I found on the street. Thanks to Emily Ricq–a legitimate science nerd– for her friendship, which has been a precious resource of maturity and perspective over the last few years. A ridiculously big "thanks" goes to my best friend, David Westwood. From the drawings of ambiguous animals left at my desk, to his patience for an endless numbers of practice talks, to his generous heart and kind understanding at the end of the day, I can't imagine the last five years without him.

Finally, to the ultimate sources of inspiration, love and support: my family. I am so thankful for Elspeth and Lydia Brotherton, my amazingly multi-talented sisters. Thanks so much for long phone conversations, always giving me the most honest advice out there and generally challenging me to up my smarts and cool-ness in all aspects of life. Finally, I have a deep sense of gratitude for my wonderful parents, Abigail and Timothy Brotherton. Thank you for inspiring me to seek the truth through science and teaching me how to "put my nose to the grindstone" in order to do so. I couldn't have done any of this without your unflagging encouragement and guidance.

Table of Contents

Abstract **iii**

Acknowledgements **v**

Table of Contents **vii**

List of Abbreviations **ix**

Chapter 1: The *pks* island encodes for the biosynthesis of a small-molecule genotoxin **1**

1.1: The human microbiota and the discovery of colibactin 1

1.2: The phylogenetic distribution and genetic context of the *pks* island 4

1.3: The human-associated biology of colibactin 6

1.4: The biochemistry and logic of assembly line biosynthetic pathways 8

1.5: The *pks* island encodes for a non-canonical assembly line pathway 13

1.6: References 19

Chapter 2: The prodrug resistance strategy in colibactin biosynthesis and genotoxicity **23**

2.1: Self-resistance in antibiotic producers and the discovery of the prodrug resistance strategy 23

2.2: The *in vitro* biochemical characterization of ClbN 25

2.3: The *in vitro* biochemical characterization of $ClbB_{NRPS}$ 36

2.4: Characterization of the prodrug cleaving enzyme ClbP 41

2.5: Conclusions 55

2.6: Experimental section 57

2.7: References 86

Chapter 3: Isolation of a *pks*-associated metabolite **89**

3.1: The prodrug resistance strategy guides the discovery of *pks* metabolites 89

3.2: Global metabolite profiling identifies several candidate precolibactins 90

3.4: Optimization of precolibactin production 99

3.5: The isolation and characterization of Metabolite B 102

3.6: ClbP can hydrolytically process Metabolite B 109

3.7: Conclusions 111

3.8: Experimental section 113

3.9: References 139

Chapter 4: Toward a complete understanding of the colibactin biosynthetic pathway **141**

4.1: The isolation of additional *pks*-associated metabolites 141

4.2: Biosynthetic hypothesis for Metabolites A and B 145

4.3: Biosynthetic hypothesis for Metabolites C, D and E 150

4.4: $ClbB_{PKS}$ is a hybrid NRPS/PKS with an unusual domain organization 151

4.5: The *in vitro* biochemical characterization of $ClbB_{PKS}$ 155

4.6: Genetic studies on the role of ClbI in colibactin biosynthesis 160

4.7: Studying the role of fatty acid synthase enzymes in colibactin biosynthesis 165

4.8: Genetic studies on the role of ClbD and ClbG in colibactin biosynthesis 167

4.9: Conclusions 169

4.10: Experimental section 173

4.11: References 184

List of Abbreviations

Ala	alanine
aq	aqueous
Asn	asparagine
Asp	aspartate
Boc	*t*-butoxycarbonyl
Bn	benzyl
Bu	butyl
°C	degree Celsius
Ci	curie
cpm	counts per minute
D	dextrarotatory
Da	dalton
DMSO	dimethyl sulfoxide
equiv	equivalent
ES	electrospray
eV	electron volt
g	gram
Gly	glycine
h	hour
HPLC	high-performance liquid chromatography
IR	infrared
J	coupling constant
L	liter
L	levoratory

M	molar
Me	methyl
min	minute
mol	mole
MW	molecular weight
NMR	nuclear magnetic resonance
ppm	parts per million
psi	pounds per square inch
rpm	revolutions per minute
s	second
Ser	serine
Tris	2-amino-2-(hydroxymethyl)propane-1,3-diol
UV	ultra-violet
Vis	visible

Chapter 1: The *pks* island encodes for the biosynthesis of a small-molecule genotoxin

1.1: The human microbiota and the discovery of colibactin

Humans exist in symbiosis with trillions of bacteria that are collectively referred to as the human microbiota.[1] The microbiota –a collection of distinct communities located at multiple body sites– is central to health, leading some to refer to it as an organ unto itself.[2] The colon harbors the largest human-associated microbial population, and, with a density of about 3×10^{11} microbial cells/g organ weight, is one of the most densely populated ecosystems on the planet.[3] The gut microbiota performs many functions that support human health, such as producing vitamins and fermenting non-digestible carbohydrates.[4,5] In addition, the gut microbiota is integral to the development and proper functioning of the immune system.[6]

Bolstered by major funding initiatives,[7] the past decade of research has yielded fundamental insights into the functions and composition of this ecosystem. This research has predominantly relied on culture-independent sequencing approaches, in which microbial community composition, determined largely by small-unit (16S) ribosomal RNA (rRNA) sequences, is correlated to host phenotypes.[7] Major advances in the field have included a greater understanding of how the gut microbiota is established from birth and the impact of the diet on gut microbial community composition.[8,9] At the same time that connections are being made between gut microbes and health, the relationships between the microbiota and both gastrointestinal and systemic disease are being elucidated. For instance, a core gut microbial community of reduced bacterial diversity has been associated with obesity.[10]

One major area of research at the intersection of gut microbial community composition and human disease is the study of colorectal cancer (CRC). The progression of normal colonic

epithelial cells to transformed, cancerous cells involves the accumulation of several identified oncogenic mutations.[11] However, it is unknown what exact genetic or environmental events cause the initial mutation(s) or promote subsequent cancer progression.[11] Importantly, research suggests that specific gut microbes or microbial metabolites could contribute to CRC development or progression.[12] While some forms of CRC are heritable, there is a large environmental component to the disease: CRC has been shown to be influenced by diet and lifestyle, and is more common in individuals suffering from inflammatory bowel diseases, such as Crohn's disease and ulcerative colitis.[13] Furthermore, CRC is more commonly found in the colon than in the small intestine.[11] Incidentally, the large intestine harbors a much larger microbial population (~10^{11} microbial cells/g) than does the small intestine (~10^{4}-10^{8} microbial cells/g).[14] It has been hypothesized that the gut microbiota and microbial metabolites may affect CRC development through various mechanisms, such as increasing inflammation, causing cell death, or damaging DNA.[12] For instance, metabolites such as secondary bile acids[15] and protein fermentation products[16] may cause DNA damage through oxidative pathways.

In 2006, Oswald and co-workers observed that some human-associated bacteria were genotoxic to host cells through the action of an unidentified small-molecule, marking an important addition to the arsenal of metabolites relevant to CRC development.[17] When certain strains of *Escherichia coli* were co-incubated with mammalian cells, including HeLa, CHO and IEC-6 cell lines, the human cells stopped dividing at the G2 phase of the cell-cycle and became enlarged, a phenotype known as megalocytosis. Cell division was blocked due to the accumulation of double-strand DNA breaks (DSB), identified through various assays, including one in which the presence of phosphorylated histone H2AX (γH2AX) was assessed with immunofluorescence microscopy. Using transposon mutagenesis, genotoxicity was linked to a 54-kilobase (kb) stretch of genomic DNA that was named the *pks* island, because this gene cluster encoded multiple polyketide

synthase (PKS) enzymes and had hallmarks of horizontal gene transfer (Figure 1.1). The genomic context and biosynthetic pathway of the *pks* island are discussed in detail below. The *pks* island was necessary and sufficient to transform *E. coli* DH10B, a non-pathogenic, lab strain of *E. coli,* into a genotoxic strain. Due to the many biosynthetic enzymes encoded in the *pks* island, it was proposed that genotoxicity could arise from the action of a small molecule produced by the *pks* cluster, a putative metabolite the authors named colibactin. Colibactin was not isolated, nor was a candidate structure proposed. Interestingly, genotoxicity was only observed when live *pks*$^+$ bacteria were in contact with human cells. When the bacteria were killed through heat or antibiotic treatment, or separated from the human cells by a 0.2 μm membrane, genotoxicity was not observed. In addition, the cell-free, spent media of *pks*$^+$ bacteria was not genotoxic to human cells. Finally, all of the *pks*-encoded proteins, with the exception of a Na^+/drug antiporter (*clbM*) and a protein with unknown function (*clbS*), were required for genotoxicity.

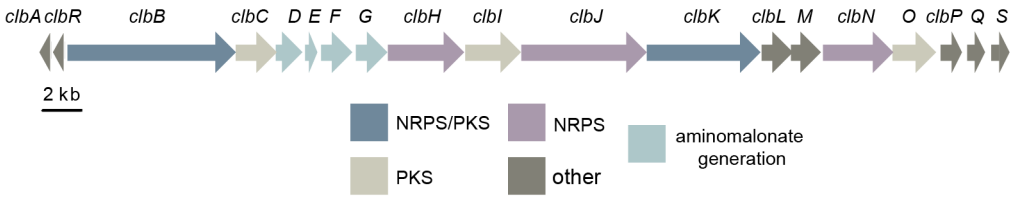


Figure 1.1: The *pks* island is a biosynthetic gene cluster containing PKS, non-ribosomal peptide synthetase (NRPS) and NRPS/PKS hybrid enzymes.

Since the discovery of the *pks* island nearly ten years ago, the biosynthesis, structure, mode of action and biological effects of colibactin have been the subject of intense research in multiple laboratories. Despite this broad interest, colibactin's structure is still unknown, and most questions regarding its biology remain unanswered. Here, investigations into the biosynthesis and structure of colibactin are reported, and these studies mark significant progress towards understanding this enigmatic and important small molecule.

1.2: The phylogenetic distribution and genetic context of the *pks* island

The initial report concerning the discovery of colibactin disclosed that the *pks* island was found in both commensal and pathogenic *E. coli* strains. A complete analysis of the phylogenetic distribution of the *pks* island showed that, within *E. coli,* the cluster is confined almost exclusively to the phylogenetic group B2, and about 73% of B2 strains harbor the gene cluster.[18,19] The B2 group is composed of extraintestinal pathogenic *E. coli* (ExPEC) such as *E. coli* CFT073, which is commonly implicated in urinary tract infections.[20] The B2 group also includes commensal strains such as the probiotic strain Nissle 1917, which is sold commercially in Europe as "Mutaflor" and has been used for over a century for the treatment of gastrointestinal diseases.[21]

Beyond *E. coli*, all known *pks*$^+$ strains are Gram-negative, in the phylum Proteobacteria, and symbionts of eukaryotes. Among the family Enterobacteriaceae, the *pks* island is found in *Klebsiella pneumoniae, Enterobacter aerogenes,* and *Citrobacter koseri* isolates.[18] The *pks* island is also found in the Alphaproteobacterial strain and sponge symbiont *Pseudovibrio denitrificans* FO-BEG1 as well as the Gammproteobacterial strain and honeybee symbiont *Frischella perrara* PEB0191.[22,23] All sequenced Enterobacteriaceae *pks* clusters share a high degree of nucleotide similarity (98%) with a G+C content of about 53%.[18] While nucleotide and amino acid composition diverges more significantly in the *pks* clusters of the sponge and honeybee symbionts, gene synteny and biosynthetic enzyme domain organization are conserved.[23] The one exception to this rule is that two genes, *clbG* and *clbH*, which are encoded as separate genes in other known *pks* clusters, are fused in the *Pseudovibrio denitrificans* FO-BEG1 *pks* cluster.[22]

Genetic features found within the *pks* island provide strong evidence that the cluster spread to *E. coli* and other Proteobacterial strains through horizontal gene transfer.[18,24] In *E. coli,* the *pks* cluster is found within the *asnW* tRNA locus.[18] tRNA loci are frequently the insertion site of

foreign DNA into the *E. coli* genome.[25] The cluster also contains an integrase gene and is flanked by 16 base pair (bp) direct repeats.[18] These elements suggest that the *pks* cluster could have been transferred by phage transduction.[24] Finally, the G+C content of the cluster is significantly higher than the *E. coli* genome: the G+C content is 53% in the *pks* island compared to 50% in *E. coli* genomic DNA. A disparity in G+C content compared to genomic DNA is often observed in foreign DNA acquired by horizontal gene transfer.[26]

The *pks* island is physically associated with the high pathogenicity island (HPI) in a subset of virulent *E. coli.*[18] The HPI is a ~45 kb stretch of chromosomal DNA found amongst many pathogenic members of the family Enterobacteriaceae, including certain strains of *Yersinia pestis, Y. enterolitica,* and *E. coli*.[27] Among other genes that are important for the virulence of ExPEC, the HPI encodes for the biosynthesis of the siderophore yersiniabactin, which is used for the acquisition and transport of iron, an essential and limited element.[28] The implications of the physical connection between the yersiniabactin and colibactin biosynthetic gene clusters are discussed in more detail below.

Finally, an analysis of the expression of the *pks* genes revealed the transcriptional organization of the cluster and the effects of various growth conditions on the expression levels of *pks* transcripts.[29] Reverse transcriptase polymerase chain reaction (RT-PCR) experiments revealed that the *pks* island is organized into at least seven transcriptional units, with two large transcripts covering a majority of the island: one transcript included six genes from *clbI* to *clbN* (23.3 kb) and another included five genes from *clbC* to *clbG* (6.2 kb). RT-PCR and luciferase reporter gene fusion assays showed that the *pks* cluster was expressed constitutively, albeit weakly, under all tested growth conditions, but transcript levels were affected by culture conditions, including aeration, carbon source and the richness of the culture media. It was also observed that co-

incubation with HeLa cells did not affect *pks* gene expression. Thus, while cell-cell contact is required for genotoxicity of *pks*$^{+}$ *E. coli* toward human cells, contact is not required for transcription of the *pks* genes, which suggests that colibactin is constitutively produced but may be highly unstable in solution.[17]

1.3: The human-associated biology of colibactin

About 90% of healthy people have *E. coli* as members of their gut microbiota and the proportion of B2 *E. coli* relative to other phylogenetic groups is higher in those individuals who consume Western (i.e. high-fat and protein) diets.[30,31] For instance, 2% of Malians' gut *E. coli* are in the B2 group, whereas Americans' *E. coli* are 48% B2.[31,32] These data suggest colibactin producers may be commonly found within healthy individuals' guts, at least in the United States and Europe. One study that relied on culture-based analyses suggested that 21% of healthy individuals, 40% of those with IBD, and 67% of CRC patients harbor *pks*$^{+}$ *E. coli*.[33] Inflammation in the gut favors the growth of facultative anaerobes and results in an increase in the ratio of Enterobacteriaceae to obligate anaerobes in the phyla Bacteroidetes and Firmicutes.[34] One mechanistic explanation for this shift in microbial community composition is that inflammation gives Enterobacteriaceae a competitive growth advantage over fermenters like the Bacteroidetes. Specifically, it has been shown that inflammation boosts the production of nitrate, which Enterobacteriaceae, such as *E. coli,* can use as a terminal electron acceptor in anaerobic respiration.[35]

Thus, the higher prevalence of *pks*$^{+}$ *E. coli* in people with inflammatory conditions like inflammatory bowel disease or CRC may reflect the competitive growth advantage of facultative anaerobes in an inflamed gut. However, given the genotoxicity of *pks*$^{+}$ *E. coli,* researchers asked if colibactin production could cause inflammation in the gut and if these bacteria could contribute to CRC development or progression. An initial study concerning the biological role of colibactin

in the human gut examined the ability of *pks*$^{+}$ *E. coli* to induce megalocytosis and DNA damage in human cells *in vitro* and *in vivo*.[36] In an intestinal loop model and in mice inoculated with colibactin-producing bacteria, *pks*$^{+}$ *E. coli* caused significant numbers of double-strand breaks in colonocytes. In addition, *in vitro* studies revealed that exposure of CHO cells to low doses of *pks*$^{+}$ *E. coli* resulted in the division of the CHO cells with DNA damage, increased gene mutation frequency, and anchorage-dependent growth.[36] These effects suggest that colibactin may act as a mutagen and promote cancer in the gut.

A 2012 report from Jobin and co-workers lent credence to the theory that colibactin may be linked to cancer.[33] This study utilized azoxymethane (AOM)-treated interleukin-10 knock-out (IL10$^{-/-}$) mice,[37] a model for colitis-associated CRC in which all IL10$^{-/-}$ mice develop colitis (inflammation of the large colon) and treatment with the carcinogen AOM results in tumors in 60-80% of the animals. To see whether colibactin could impact cancer development, germ-free mice exposed to AOM were monoassociated with the commensal *pks*$^{+}$ *E. coli* NC101 or a strain lacking the gene cluster, *Enterococcus faecalis* OG1RF. While eighty percent of *E. coli* NC101-treated mice developed invasive tumors, fewer than 20% of *E. faecalis*-treated mice developed invasive adenocarcinomas. Interestingly, inflammation and cytokine levels were not significantly different between the mice treated with *E. coli* NC101 or *E. faecalis*. When the *pks* island was knocked out in *E. coli* NC101, the number of tumors in the mice exposed to this *pks*$^{-}$ strain was decreased compared to treatment with the wild-type, *pks*$^{+}$ *E. coli* NC101 strain. Again, the levels of inflammation in the mouse gut were similar between the *pks*$^{+}$ and *pks*$^{-}$ strains. *E. coli* NC101 did not induce tumors in IL10$^{-/-}$ mice that were not treated with AOM or in wild-type mice exposed to AOM. Overall, their results indicated that, in a background of intestinal inflammation, *pks*$^{+}$ *E. coli* promote tumor progression significantly more than a *pks*$^{-}$ counterpart.

Recent studies have sought to elucidate the mechanism of this tumor-promoting phenotype. Two different studies have suggested that exposure to *pks*$^{+}$ *E. coli* induces cellular senescence and the secretion of tumor growth factors, including hepatocyte growth factor, in human cells.[38,39] Another study showed that transcripts from the *pks* island are much more abundant in CRC tumor samples than in the surrounding tissue, which demonstrates that the *pks* island is actively transcribed in the human gut.[40]

Another set of studies have asked questions about the biological role of colibactin production in contexts not directly related to CRC. A study of the effects of colibactin production in sepsis-associated lymphopenia found that mice injected with *pks*$^{+}$ ExPEC strain SP15 had significantly lower survival rates than those mice infected with a *clbA* knock-out (*pks*$^{-}$) SP15 strain.[41] Another report found that the probiotic and anti-inflammatory effects of Nissle 1917 were dependent upon the presence of the *pks* cluster, which suggests that colibactin may play a role in mediating the immune response to these probiotic bacteria in the gut.[42] Future work toward elucidating the biological impacts of colibactin production should focus on studying the effects of colibactin in natural gut communities, which will require targeted methods for eliminating or modulating colibactin production.

1.4: The biochemistry and logic of assembly line biosynthetic pathways

Assembly line biosynthetic pathways synthesize intricate and therapeutically useful metabolites, such as the well-known and clinically relevant antibiotics erythromycin and vancomycin, that have fascinated both the synthetic organic chemistry and biochemistry communities for decades.[43,44,45,46] An examination of assembly line enzymatic machinery and logic reveals the mechanisms by which these highly complex and useful metabolites are produced from small building blocks.[47]

PKSs, non-ribosomal peptide synthetases (NRPSs) and hybrids thereof are called assembly line enzymes because these pathways function in a way that is conceptually analogous to a factory assembly line, in which simple parts are installed one at a time by highly specialized workers to construct a final product. The term assembly line was first used in reference to PKS or NRPS enzymes in 1997.[48] Assembly line enzymes covalently tether, install and modify structural motifs on biosynthetic intermediates, and pass finished pieces off to the next enzyme in the pathway for the next round of bond formation or structural modification. Intermediates are tethered to the enzymes by thioester linkages, in which the sulfur atom in this linkage comes from a phosphopantetheine (ppant) arm attached to a conserved serine residue on the protein. The ppant arm is derived from coenzyme A (**1**) (Figure 1.2). This prosthetic group is installed as a post-translational modification, in which *apo* proteins are transformed to the corresponding *holo* forms by a ppant-transferase. While the gene for a dedicated ppant-transferase is often found within a given biosynthetic gene cluster, in some cases the ppant-transferase modifies *apo* enzymes from both primary and secondary metabolism and its gene is not clustered with the biosynthetic enzymes.[49,50]

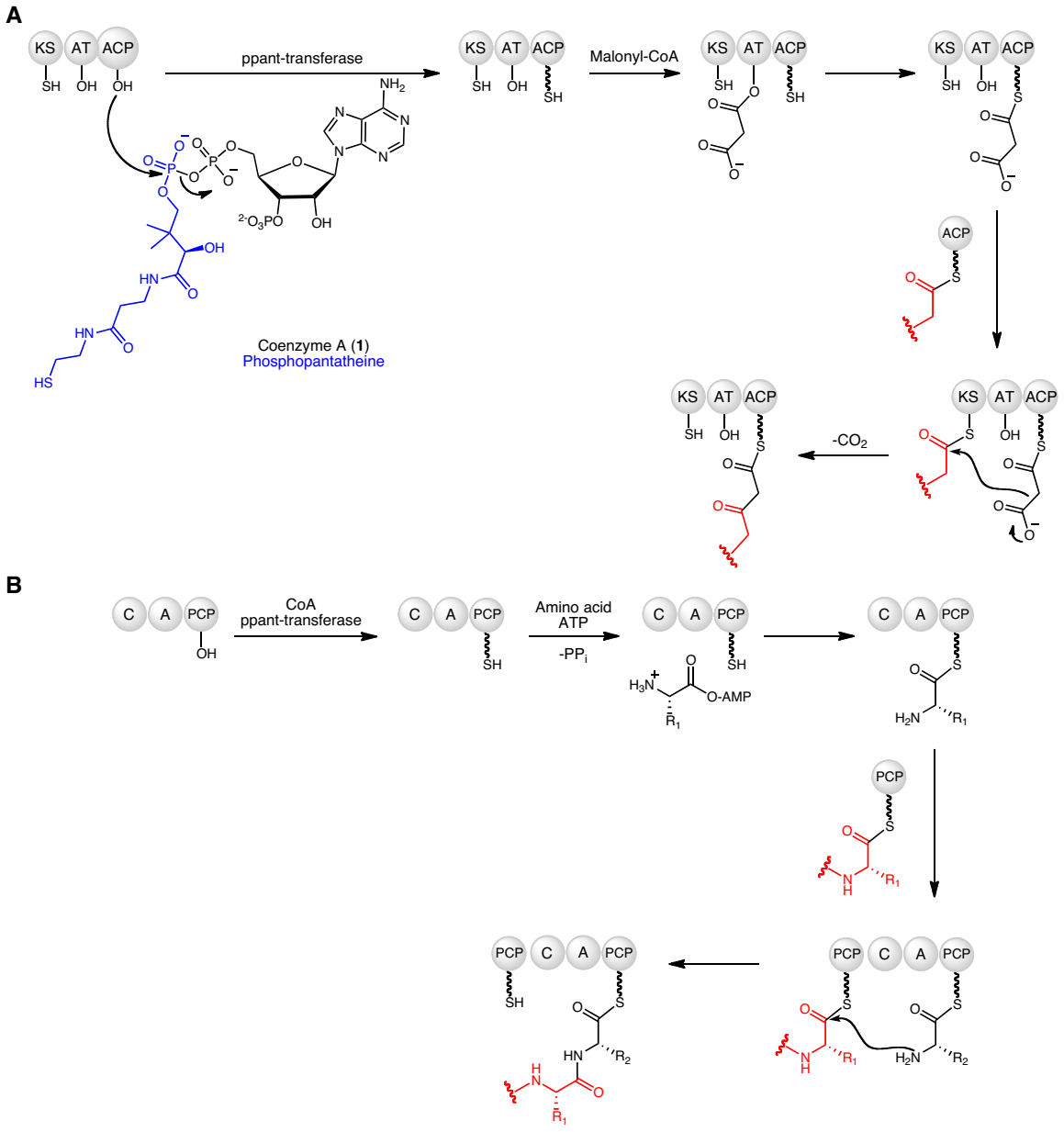


Figure 1.2: PKS and NRPS assembly line enzymes use parallel biosynthetic logic. First, the ppant arm, which is derived from the phosphopantetheine (blue) portion of Coenzyme A (**1**), is loaded by a dedicated ppant-transferase to convert enzymes to their *holo* forms. Building blocks are activated by AT and A domains, then loaded onto the ppant arm of the ACP or PCP domains. Bond formation with upstream intermediates (red) is catalyzed by the KS and C domains to provide ACP or PCP domain-bound elongated products. A) PKS assembly line logic. An upstream ACP domain-bound electrophile (red) is transferred to the KS domain cysteine residue before bond formation can occur. B) NRPS assembly line logic.

PKS and NRPS enzymes have discrete enzymatic domains that perform highly specialized functions (Figure 1.2). There are three types of domains that form the core catalytic machinery of assembly line enzymes. First, acyl-transferase (AT) and adenylation (A) domains select and

activate building blocks. The AT domain of PKS modules loads malonyl- or methylmalonyl-CoA monomers through a serine residue to form malonyl- or methylmalonyl-AT. The A domain of NRPS modules activates free amino acids by forming aminoacyl-adenylate (AMP) intermediates using adenosine triphosphate (ATP). For both AT and A domains, the monomer specificity can often be predicted with high levels of confidence. For AT domains, a conserved set of amino acid residues can be examined to predict whether that domain activates methylmalonyl-CoA or malonyl-CoA.[51] For A domains, monomer specificity can be predicted using bioinformatics programs that compare a set of conserved active-site amino acid motifs to those in characterized A domains with known substrate specificities. For instance, the amino acid specificities of most A domains can be predicted through the use of an online program like NRPSpredictor2,[52] which compares the identity of ten amino acids found in the substrate binding pocket – the so-called "Stachelhaus motif"[53] – with those in characterized A domains with known amino acid specificities. Next, the AT or A domain loads an activated monomer onto the acyl-carrier protein (ACP) or peptidyl-carrier protein (PCP) domains of PKSs and NRPSs, respectively. The ACP and PCP domains are also referred to as thiolation (T) domains. T domains tether intermediates through the ppant arm that is installed onto a conserved serine residue. All bond-forming reactions occur on ppant-bound intermediates, and the nucleophile in these reactions is always bound to the ACP or PCP domain. Finally, ketosynthase (KS) and condensation (C) domains are the bond-forming domains of PKSs and NRPSs. The KS domain of PKS modules tethers upstream intermediates through a cysteine residue, and catalyzes C-C bond formation by acting as a base to promote a decarboxylative Claisen condensation reaction. The C domain doesn't tether intermediates, but rather catalyzes peptide bond formation between two ppant-linked intermediates: one intermediate is bound to an upstream PCP domain and the other to a downstream PCP domain (Figure 1.2). Once bonds are formed, the ppant-bound intermediates

can be transferred to the next enzyme in the assembly line through nucleophilic attack by a downstream KS serine residue or a PCP-bound amine.

These core domains form the basis of most assembly line chemistry, but there are many additional catalytic domains found within NRPS and PKS modules. Together, these domains give rise to the chemical diversity that is associated with assembly line biosynthetic pathways. Type I modular PKSs, in which all catalytic domains within a module are found within the same polypeptide and each module is responsible for one round of bond-formation, typically contain three additional domains. These are the ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER) domains. The KR, DH and ER domains work to transform a β-ketone moiety into a methylene, through reduction of the ketone moiety by the KR domain using $NAD(P)H^+$, dehydration by the DH domain to form a *trans* alkene, and finally reduction by the ER domain using $NAD(P)H^+$ to form the fully reduced methylene. Intermediate levels of oxidation are seen when just the KR or KR and DH domains are present. Additional domains found in NRPSs include cyclization (Cy) domains, responsible for the cyclization of amino-acid side chains like cysteine and serine to form thiazoline and oxazoline heterocycles, respectively, and oxidation (Ox) domains, which oxidize these heterocycles to form aromatic thiazole and oxazole rings (Figure 1.3).

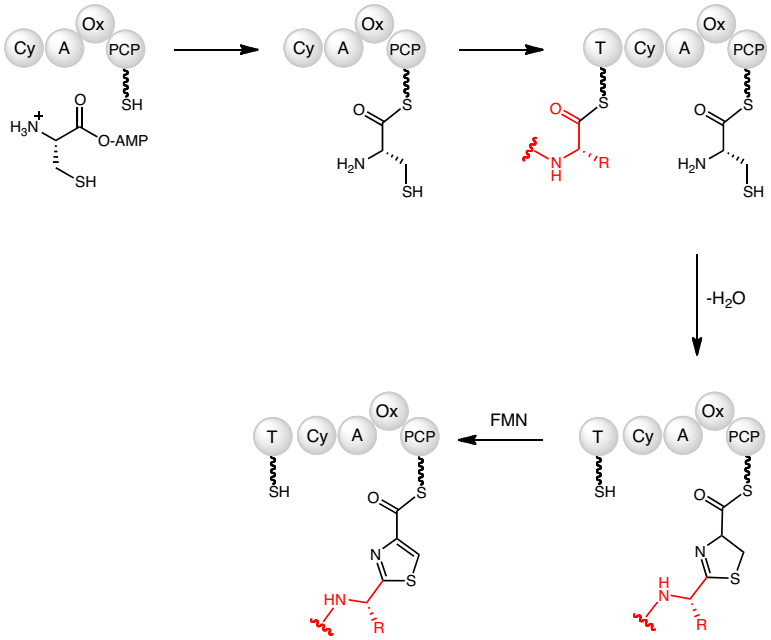


Figure 1.3: The NRPS cyclization (Cy) and oxidation (Ox) domains construct aromatic heterocycles. Here, an A domain activates L-cysteine, which is loaded onto the PCP domain. The Cy domain catalyzes amide bond formation with an upstream PCP domain-bound amino acid and condensation of the side-chain sulfhydryl group onto the upstream amide carbonyl. The Ox domain then dehydrogenates the thiazoline ring to form a thiazole.

1.5: The *pks* island encodes for a non-canonical assembly line pathway

Assembly line pathways are predominantly found in Gram-positive Actinomycetales bacteria. However, *E. coli* is known to biosynthesize a class of molecules named siderophores, which are iron-sequestering reagents that can serve as virulence factors in pathogenic strains. Examples of *E. coli*-derived siderophores include yersiniabactin and enterobactin.[27,54]

The *pks* island encodes PKS, NRPS, hybrid NRPS/PKS, transport and tailoring enzymes. Three NRPS enzymes are encoded in the cluster: ClbH, ClbN and ClbJ (Figure 1.4). ClbH has an unusual set of domains (A_1-C-A_2-PCP), in which a single A domain is followed by a complete set of catalytic domains. The ClbH A_1 domain is predicted to activate L-serine, while the ClbH A_2 domain lacks conserved Stachelhaus motif residues, which suggests that it does not activate an α-amino acid. ClbN is predicted to activate L-asparagine, and a thorough bioinformatic analysis of

its condensation and epimerization (E) domains will be left to the next chapter. ClbJ is a seemingly straightforward, two-module NRPS enzyme. Its A_1 and A_2 domains are predicted to activate glycine and L-cysteine, respectively. The second module of ClbJ contains a cyclization (Cy) domain, which likely catalyzes thiazoline ring formation.

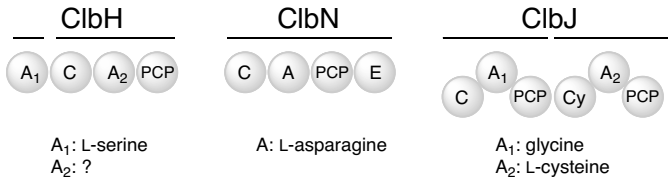


Figure 1.4: The *pks* island encodes three NRPS enzymes: ClbH, ClbN and ClbJ. The predicted amino acid specificity of each A domain is provided.

The *pks* cluster has multiple PKS and hybrid NRPS/PKS enzymes (Figure 1.5). Bioinformatics analyses reveal that none of the PKS modules found in the cluster resemble canonical type I modular PKS enzymes in the literature, such as the PKS assembly line involved in erythromycin biosynthesis, which suggests these enzymes may perform unusual biochemistry.[43] ClbI is a *cis*-AT type PKS module because its AT domain is contained within the same polypeptide as the KS and ACP domains. Its AT domain is predicted to activate malonyl-CoA based on a conserved signature found in characterized AT domains known to activate malonyl-CoA.[55] This signature includes a four-residue HAFH motif and an active-site serine residue. However, its KS domain lacks the conserved active-site cysteine that tethers intermediates and instead has a serine residue, which suggests that ClbI cannot accept an intermediate from an upstream assembly line enzyme. ClbI will be discussed in more detail in Chapter 4. ClbB is a two module, hybrid NRPS-PKS. The A domain of its NRPS module is predicted to activate L-alanine. The *cis*-AT PKS module of ClbB has an unusual domain organization of KS-AT-KR-DH-ER-ACP. The order normally seen is KS-AT-DH-ER-KR-ACP. The only other known example of a characterized type I PKS module with

the same domain organization as the PKS module of ClbB ($ClbB_{PKS}$) is TubD from the tubulysin biosynthetic pathway.[56] ClbB will be discussed in greater detail in Chapters 2 and 4.

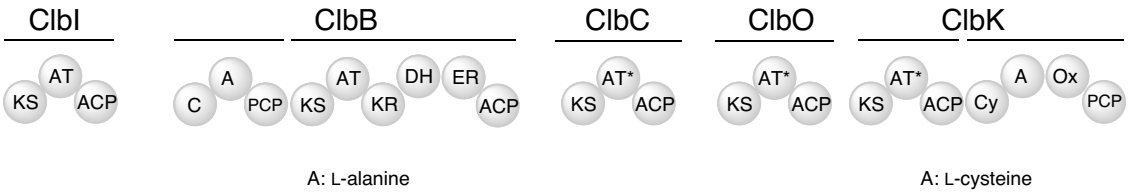


Figure 1.5: The *pks* island encodes both *cis*- and *trans*-AT PKS and hybrid NRPS/PKS enzymes, ClbI, ClbB, ClbC, ClbO and ClbK. AT* domains are predicted to be non-functional AT domains. The predicted amino acid specificity of each A domain is provided.

In addition to the unusual *cis*-AT PKS modules, the colibactin cluster also has three *trans*-AT type PKS modules: ClbC, ClbO and the first module of ClbK (Figure 1.5). In contrast to *cis*-AT PKSs, *trans*-AT PKSs utilize an AT that is not part of the same polypeptide as the KS and ACP domains.[57] In the colibactin pathway, the AT that predicted to be used in *trans* is the stand-alone AT domain ClbG. The *trans*-AT PKS in the colibactin cluster have what seem to be vestigial, inactive AT domains that lack conserved motifs found in active ATs, including the active-site serine and the HAFH motif.[23] These inactive AT domains are indicated as AT*. The *trans*-AT PKSs ClbC and ClbO have the same domain organization, while ClbK is a two-module NRPS/PKS hybrid with the domains KS-AT*-ACP-Cy-A-Ox-PCP. Like ClbJ, the NRPS module of ClbK contains a Cy domain, which is predicted to cyclize L-cysteine based on the A domain's Stachelhaus motif. The Ox domain likely utilizes an FMN cofactor to dehydrogenate the thiazoline formed by the Cy domain to provide a thiazole ring.[58] While there is a growing list of characterized all *trans*-AT PKS biosynthetic gene clusters, few characterized pathways have both *trans*- and *cis*-AT PKS enzymes.[59] One example is the zwittermicin gene cluster.[60]

Aminomalonate is an uncommon extender unit in PKS and NRPS assembly lines that is biosynthesized by a dedicated set of enzymes found within the *pks* island–ClbD, E, F and G– and is predicted to be utilized by the *trans*-AT PKS modules ClbC, O and K (Figure 1.6). Recently, some of the aminomalonate-forming enzymes were biochemically characterized *in vitro* by Piel and co-workers.[61] It was found that aminomalonate biosynthesis in the colibactin pathway parallels that in the zwittermicin pathway, which also possesses close homologues of ClbD-G and was characterized by the Thomas lab several years earlier.[62] The biosynthesis of the aminomalonate extender unit begins by activation of serine by the A_1 domain of ClbH and transfer of the serine-AMP to the ppant arm of the stand-alone ACP ClbE. The serine side-chain alcohol is then sequentially oxidized by the dehydrogenases ClbD and ClbF. Recently, Li Zha, a graduate student in the Balskus lab, characterized the transfer of aminomalonyl-ACP to the *trans*-acting AT ClbG, as well as transfer of aminomalonyl from ClbG to the ACP domains of ClbC, O and K. His work provides the first evidence that aminomalonate is used multiple times in the colibactin pathway.[63] However, *in vitro* characterization of bond formation involving an ACP-bound aminomalonate has been elusive, perhaps due to the inherent instability of this building block.

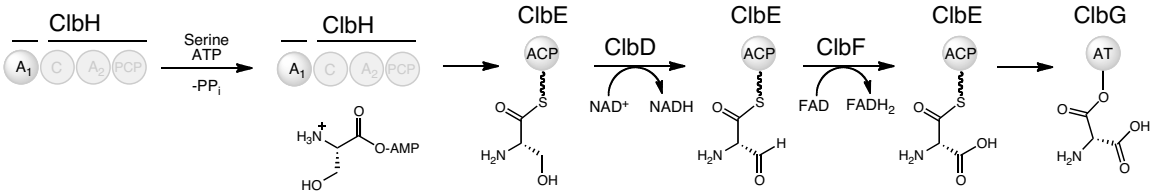


Figure 1.6: Aminomalonyl-AT is biosynthesized starting from serine by the enzymes ClbH, E, D,and F. ClbG act in *trans* to load the *trans*-AT PKS modules found within the assembly line.

In addition to the assembly line enzymes that are predicted to biosynthesize the colibactin scaffold, the *pks* cluster also encodes for proteins involved in tailoring and transport: ClbA, Q, M, L and P. ClbA is a ppant-transferase, and utilizes CoA (**1**) to transform the *pks* enzymes to their

corresponding *holo*, active forms. A study from the Oswald group demonstrated that ClbA could also modify enzymes in the yersiniabactin biosynthetic pathway in *E. coli,* which suggests that ClbA may play a role in virulence by helping to convert yersiniabactin assembly line modules from the *apo* to *holo* form.[64] Interestingly, *clbA* has often been the target of mutation in biological studies to provide *pks*$^-$ strains that serve as negative controls in assays with wild-type, *pks*$^+$ strains. Thus, the conclusions of studies that rely on the use of *clbA* mutants should be considered in light of these data that indicate *clbA* mutants may be deficient in both colibactin and siderophore biosynthesis *in vivo*.

ClbQ is homologous to type II thioesterases, which are stand-alone thioesterase domains that hydrolyze the thioester linkage of T domain-bound biosynthetic intermediates using a catalytic serine nucleophile.[65] Type II thioesterases have been proposed to increase the efficiency of assembly line pathways by hydrolyzing "incorrect" or stalled intermediates to increase flux toward the desired product(s).[65] Type II thioesterases have also been shown to regenerate free thiol groups on misacylated PCP domains, which can arise when the ppant-transferase loads the *apo* PCP domain with an acylated CoA, such as acetyl-CoA, from primary metabolism.[66] ClbL is homologous to the amidotransferase subunit of aspartate-tRNA amidotransferase, which transfers ammonia from glutamine to misacylated aspartate-tRNA^{Asn} to provide correctly charged asparagine-tRNA^{Asn}.[67] While the role of ClbL in the colibactin pathway has not yet been determined, it may modify a carboxylic acid or amide moiety on the colibactin scaffold. For example, ClbL may hydrolyze an amide bond to produce a carboxylic acid. ClbM is a Na^+/drug antiporter and a member of the multidrug and toxic compound extrusion (MATE) family of efflux pumps. ClbM is not required for *pks*-associated genotoxicity, suggesting that another transporter encoded in the genome may compensate for a loss of the activity of ClbM.

Finally, both *clbP* and *clbS* have been implicated in the self-resistance of colibactin producers. ClbP is a periplasmic peptidase that is integral to the prodrug resistance strategy, a self-resistance mechanism that is a focus of Chapter 2. Preliminary evidence has suggested that ClbS also plays a role in resistance. A 2012 report showed that a cysteine residue within ClbS formed covalent adducts with electrophilic α-alkylidene-γ-butyrolactones.[68] Recently, it was demonstrated that expression of ClbS in HeLa cells blocked the genotoxicity of *pks*$^+$ *E. coli*.[69] While *clbS* is not required for genotoxicity, the encoded protein may provide protection from colibactin within the cytoplasm of the producing bacteria, perhaps by sequestering the active molecule or modifying the structure to render it inactive. The molecular mechanism by which ClbS exerts its effects has yet to be elucidated.

Given the dearth of isolated metabolites associated with the *pks* island, the potential for non-canonical biochemistry performed by enzymes encoded in the cluster, and the intriguing biological activity of *pks*$^+$ bacteria, we initiated a biochemical investigation of the colibactin biosynthetic pathway. We believed that detailed *in vitro* studies of enzyme activity would allow us to generate structural hypotheses about colibactin that could guide isolation and further enzymatic characterization. Our unconventional approach to study the biosynthetic enzymes before isolation of the natural product has led to the discovery of novel biochemistry and the isolation of various *pks* metabolites. Our efforts are detailed in the following three chapters. Chapter 2 describes the characterization of the prodrug self-resistance strategy in colibactin biosynthesis. In Chapter 3, the isolation and characterization of a *pks* metabolite is described. Finally, Chapter 4 presents research towards understanding missing pieces of the colibactin biosynthetic pathway.

1.6: References

(1) Lederberg, J.; Mccray, A. *The Scientist* **2001**, *17*.

(2) Bäkhed, F.; Ley, R. E.; Sonnenburg, J. L.; Peterson, D. A.; Gordon, J. I. *Science* **2005**, *307*, 1915.

(3) Whitman, W. B.; Coleman, D. C.; Wiebe, W. J. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6578.

(4) Hill, M. J. *Eur. J. Cancer Prev.* **1997**, *6,* S43.

(5) Flint, H. J.; Scott, K. P.; Duncan, S. H.; Louis, P.; Forano, E. *Gut Microbes* **2012**, *3*, 289.

(6) Hooper, L. V.; Littman, D. R.; Macpherson, A. J. *Science* **2012**, *336*, 1268.

(7) Turnbaugh, P. J.; Ley, R. E.; Hamady, M.; Fraser-Liggett, C. M.; Knight, R.; Gordon, J. I. *Nature* **2007**, *449*, 804.

(8) *a)* Yatsunenko, T.; Rey, F. E.; Manary, M. J.; Trehan, I.; Dominguez-Bello, M. G.; Contreras, M.; Magris, M.; Hidalgo, G.; Baldassano, R. N.; Anokhin, A. P.; Heath, A. C.; Warner, B.; Reeder, J.; Kuczynski, J.; Caporaso, J. G.; Lozupone, C. A.; Lauber, C.; Clemente, J. C.; Knights, D.; Knight, R.; Gordon, J. I. *Nature* **2012**, *486*, 222. *b)* Costello, E. K.; Stagaman, K.; Dethlefsen, L.; Bohannan, B. J. M.; Relman, D. A. *Science* **2012**, *336*, 1255.

(9) David, L. A.; Maurice, C. F.; Carmody, R. N.; Gootenberg, D. B.; Button, J. E.; Wolfe, B. E.; Ling, A. V.; Devlin, A. S.; Varma, Y.; Fischbach, M. A.; Biddinger, S. B.; Dutton, R. J.; Turnbaugh, P. J. *Nature* **2014**, *505*, 559.

(10) Turnbaugh, P. J.; Hamady, M.; Yatsunenko, T.; Cantarel, B. L.; Duncan, A.; Ley, R. E.; Sogin, M. L.; Jones, W. J.; Roe, B. A.; Affourtit, J. P.; Egholm, M.; Henrissat, B.; Heath, A. C.; Knight, R.; Gordon, J. I. *Nature* **2009**, *457*, 480.

(11) Sears, C. L.; Garrett, W. S. *Cell Host Microbe* **2014**, *15*, 317.

(12) Louis, P.; Hold, G. L.; Flint, H. J. *Nat. Rev. Micro.* **2014**, *12*, 661.

(13) *(a)* Jess, T.;Gamborg, M.; Matzen, P., Munkholm, P.; Sørensen, T. I. A. *Am. J. Gastroenterol.***2005**, *100*, 2724. (*b)* Danese, S.; Malesci, A.; Vetrano, S. *Gut* **2011**, *60*, 1609.

(14) Walter, J.; Ley, R. *Annu. Rev. Microbiol.* **2011**, *65*, 411.

(15) Bernstein, H.; Bernstein, C.; Payne, C. M.; Dvorak, K. *World J. Gastroenterol.* **2009**, *15*, 3329.

(16) Windey, K.; De Preter, V.; Verbeke, K. *Mol. Nutr. Food Res.* **2012**, *56*, 184.

(17) Nougayrede, J.-P.; Homburg, S.; de ric Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J. R.; Dobrindt, U.; Oswald, E. *Science* **2006**, *313*, 848.

(18) Putze, J.; Hennequin, C.; Nougayrede, J. P.; Zhang, W.; Homburg, S.; Karch, H.; Bringer, M. A.; Fayolle, C.; Carniel, E.; Rabsch, W.; Oelschlaeger, T. A.; Oswald, E.; Forestier, C.; Hacker, J.; Dobrindt, U. *Infect. Immun.* **2009**, *77*, 4696.

(19) Johnson, J. R.; Johnston, B.; Kuskowski, M. A.; Nougayrede, J. P.; Oswald, E. *J. Clin. Microbiol.* **2008**, *46*, 3906.

(20) Snyder, J. A.; Haugen, B. J.; Buckles, E. L.; Lockatell, C. V.; Johnson, D. E.; Donnenberg, M. S.; Welch, R. A.; Mobley, H. L. T. *Infect, Immun.* **2004**, *72*, 6373.

(21) Olier, M.; Marcq, I.; Salvador-Cartier, C.; Secher, T.; Dobrindt, U.; Boury, M.; Bacquie, V.; Penary, M.; Gaultier, E.; Nougayrede, J.-P.; Fioramonti, J.; Oswald, E. *Gut Microbes* **2012**, *3*, 501.

(22) Bondarev, V.; Richter, M.; Romano, S. *Environ. Microbiol.* **2013**, *15*, 2095.

(23) Engel, P.; Vizcaino, M. I.; Crawford, J. M. *Appl. Environ. Microbiol.* **2015**, *81*, 1502.

(24) Lawrence, J. G.; Groisman, E. A. *Nature* **2000**, *405*, 299.

(25) Hacker, J.; Kaper, J. B. *Annu. Rev. Microbiol.* **2000**, *54*, 641.

(26) Schneider, G.; Dobrindt, U.; Middendorf, B.; Hochhut, B.; Szijártó, V.; Emődy, L.; Hacker, J. *BMC Microbiology* **2011**, *11*, 210.

(27) Schubert, S.; Rakin, A.; Karch, H.; Carniel, E.; Heeseman, J. *Infect. Immun.* **1998**, *66*, 480.

(28) Perry, R. D.; Balbo, P. B.; Jones, H. A.; Festerhston, J. D.; DeMoll, E. *Microbiology* **1999**, *145*, 1181.

(29) Homburg, S.; Oswald, E.; Hacker, J. R.; Dobrindt, U. *FEMS Microbiol. Lett.* **2007**, *275*, 255.

(30) Tenaillon, O.; Skurnik, D.; Picard, B.; Denamur, E. *Nat. Rev. Microbiol.* **2010**, *8*, 207.

(31) Penders, J.; Thijs, C.; Vink, C.; Stelma, F. F.; Snijders, B.; Kummeling, I.; van den Brandt, P. A.; Stobberingh, E. E. *Pediatrics* **2006**, *118*, 511.

(32) Duriez, P.; Clermont, O.; Bonacorsi, S.; Bingen, E.; Chaventré, A.; Elion, J.; Picard, B.; Denamur, E. *Microbiology* **2001**, *147*, 1671.

(33) Arthur, J. C.; Perez-Chanona, E.; Muhlbauer, M.; Tomkovich, S.; Uronis, J. M.; Fan, T. J.; Campbell, B. J.; Abujamel, T.; Dogan, B.; Rogers, A. B.; Rhodes, J. M.; Stintzi, A.; Simpson, K. W.; Hansen, J. J.; Keku, T. O.; Fodor, A. A.; Jobin, C. *Science* **2012**, *338*, 120.

(34) Lupp, C.; Robertson, M. L.; Wickham, M. E.; Sekirov, I. *Cell Host Microbe* **2007,** *2*, 119.

(35) Winter, S. E.; Winter, M. G.; Xavier, M. N.; Thiennimitr, P.; Poon, V.; Keestra, A. M.; Laughlin, R. C.; Gomez, G.; Wu, J.; Lawhon, S. D.; Popova, I. E.; Parikh, S. J.; Adams, L. G.; Tsolis, R. M.; Stewart, V. J.; Bäumler, A. J. *Science* **2013**, *339*, 708.

(36) Cuevas-Ramos, G.; Petit, C. R.; Marcq, I.; Boury, M.; Oswald, E.; Nougayrede, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 11537.

(37) Kühn, R.; Löhler, J.; Rennick, D.; Rajewsky, K.; Müller, W. *Cell* **1993**, *75*, 263.

(38) Secher, T.; Samba-Louaka, A.; Oswald, E.; Nougayrede, J.-P. *PLoS ONE* **2013**, *8*, e77157.

(39) Cougnoux, A.; Dalmasso, G.; Martinez, R.; Buc, E.; Delmas, J.; Gibold, L.; Sauvanet, P.; Darcha, C.; Dechelotte, P.; Bonnet, M.; Pezet, D.; Wodrich, H.; Darfeuille-Michaud, A.; Bonnet, R. *Gut* **2014**, *63*, 1932.

(40) Dutilh, B. E.; Backus, L.; van Hijum, S. A. F. T.; Tjalsma, H. *Best Pract. Res. Clin. Gastroenterol.* **2013**, *27*, 85.

(41) Marcq, I.; Martin, P.; Payros, D.; Cuevas-Ramos, G.; Boury, M.; Watrin, C.; Nougayrede, J.-P.; Olier, M.; Oswald, E. *J. Infect. Dis.* **2014**, *210*, 285.

(42) Olier, M.; Marcq, I.; Salvador-Cartier, C.; Secher, T.; Dobrindt, U.; Boury, M.; Bacquie, V.; Penary, M.; Gaultier, E.; Nougayrede, J.-P.; Fioramonti, J.; Oswald, E. *Gut Microbes* **2012**, *3*, 501.

(43) Boger, D. L. *Med Res Rev* **2001**, *21*, 356.

(44) Staunton, J.; Weissman, K. J. *Nat. Prod. Rep.* **2001**, *18*, 380.

(45) Walsh, C. T. *Nat. Prod. Rep.* **2015**. *Advance article*

(46) Khosla, C.; Herschlag, D.; Cane, D. E.; Walsh, C. T. *Biochemistry* **2014**, *53*, 2875.

(47) Fischbach, M. A.; Walsh, C. T. *Chem. Rev.* **2006**, *106*, 3468.

(48) Leadlay, P. F. *Curr. Opin. Chem. Biol.* **1997**, *1*, 162.

(49) Finking, R.; Solsbacher, J.; Konz, D.; Schobert, M.; Schafer, A.; Jahn, D.; Marahiel, M. A. *J. Bio. Chem.* **2002**, *277*, 50293.

(50) Beld, J.; Sonnenschein, E. C.; Vickery, C. R.; Noel, J. P.; Burkart, M. D. *Nat. Prod. Rep.* **2014**, *31*, 61.

(51)Del Vecchio, F.; Petkovic, H.; Kendrew, S. G.; Low, L.; Wilkinson, B.; Lill, R.; Cortes, J.; Rudd, B. A. M.; Staunton, J.; Leadlay, P. F. *J. Ind. Microbiol. Biotechnol.* **2003**, *30*, 489.

(52) Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. *Nucleic Acids Res.* **2011**, *39*, W362.

(53) Stachelhaus, T.; Mootz, H. D.; Marahiel, M. A. *Chem. Biol.* **1999**, *6*, 493.

(54) Ehmann, D. E.; Shaw-Reid, C. A.; Losey, H. C.; Walsh, C. T. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2509.

(55) Del Vecchio, F.; Petkovic, H.; Kendrew, S. G.; Low, L.; Wilkinson, B.; Lill, R.; Cortes, J.; Rudd, B. A. M.; Staunton, J.; Leadlay, P. F. *J. Ind. Microbiol. Biotechnol.* **2003**, *30*, 489.

(56) Chai, Y.; Pistorius, D.; Ullrich, A.; Weissman, K. J.; Kazmaier, U.; Müller, R. *Chem. Biol.* **2010**, *17*, 296.

(57) Piel, J. *Nat. Prod. Rep.* **2010**, *27*, 996.

(58) Schneider, T. L.; Shen, B.; Walsh, C. T. *Biochemistry* **2003**, *42*, 9722.

(59) Kampa, A.; Gagunashvili, A. N.; Gulder, T. A. M.; Morinaka, B. I.; Daolio, C.; Godejohann, M.; Miao, V. P. W.; Piel, J.; Andrésson, O. S. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E3129.

(60) Kevany, B. M.; Rasko, D. A.; Thomas, M. G. *Appl. Environ. Microbiol.* **2009**, *75*, 1144.

(61) Brachmann, A. O.; Garcie, C.; Wu, V.; Martin, P.; Ueoka, R.; Oswald, E.; Piel, J. *Chem. Commun.* **2015**, *51*, 13138.

(62) Chan, Y. A.; Thomas, M. G. *Biochemistry* **2010**, *49*, 3667.

(63) Zha, L.; Wilson, M.; Brotherton, C.A.; Balskus, E.P. *in preparation.*

(64) Martin, P.; Marcq, I.; Magistro, G.; Penary, M.; Garcie, C.; Payros, D.; Boury, M.; Olier, M.; Nougayrede, J.-P.; Audebert, M.; Chalut, C.; Schubert, S.; Oswald, E. *PLoS Pathog.* **2013**, *9*, e1003437.

(65) Yeh, E.; Kohli, R. M.; Bruner, S. D.; Walsh, C. T. *ChemBioChem* **2004**, *5*, 1290.

(66) Schwarzer, D.; Mootz, H. D.; Linne, U. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14083.

(67) Ibba, M.; Söll, D. *Annu. Rev. Biochem.* **2000**, *69*, 617.

(68) Kunzmann, M. H.; Sieber, S. A. *Mol. Biosyst.* **2012**, *8*, 3061.

(69) Bossuet-Greif, N.; Dubois, D.; Petit, C.; Tronnet, S.; Martin, P.; Bonnet, R.; Oswald, E.; Nougayrede, J.-P. *Mol Microbiol* **2015**.

Chapter 2: The prodrug resistance strategy in colibactin biosynthesis and genotoxicity

2.1: Self-resistance in antibiotic producers and the discovery of the prodrug resistance strategy

Antibiotic producers are known to use a variety of mechanisms to "avoid suicide" by the action of their own metabolites.[1] Common self-resistance mechanisms include modification of the active metabolite to render it inactive, removal or exclusion of the compound from the cell through the use of efflux pumps, and modification or replacement of the antibiotic's target protein.[1] Examples of these self-resistance strategies include the acetylation and phosphorylation of kanamycin and related aminoglycosides, which decreases the compounds' affinity for the target 16S rRNA binding site,[2] and the resistance mechanism of the thiostrepton producer *Streptomyces azureus*, which methylates the target of thiostrepton, the 23S rRNA subunit, to prevent binding of the oligopeptide antibiotic.[3] Less common self-resistance methods include the sequestration and stabilization of highly reactive molecules by specific, small-molecule binding proteins, as in bleomycin and mitomycin producers.[4]

In 2011, a new type of self-resistance mechanism was discovered and characterized *in vivo* by Bode and co-workers.[5] While studying the biosynthesis of xenocoumacin (**3**, Figure 2.1) by the Gram-negative bacterium *Xenorhabdus nematophilus*, it was noted that several proteins in the biosynthetic gene cluster had no clear role in the assembly of the natural product scaffold. Interestingly, strains lacking one of these uncharacterized proteins, XcnG, failed to produce xenocoumacin (**3**). XcnG was predicted to contain three transmembrane helices and a periplasmic domain, and is homologous to D-amino peptidases, which cleave peptide bonds that have a D-amino acid residue.[6] Instead of producing **3**, Δ*xcnG* strains accumulated a new set of compounds, named prexenocoumacins (**2**), that were xenocoumacin (**3**) derivatives containing an additional N-terminal *N*-acyl-D-asparagine residue (Figure 2.1). Unlike **3**, the

prexenocoumacins (**2**) completely lacked antibiotic activity, which prompted the authors to hypothesize that the N-terminal moiety could serve as a prodrug motif that masks the reactivity of the antibiotic while it is still in the cytoplasm of the producer cell. Feeding studies with *E. coli* strains expressing XcnG established that this enzyme could cleave the prodrug motif from the prexenocoumacins (**2**) to provide **3**, supporting the idea that the N-terminal group shields activity until the antibiotic has left the intracellular space. The authors also established that the full-length XcnG protein was required for activity: an *E. coli* strain expressing an XcnG construct lacking the transmembrane helices failed to process prexenocoumacins (**2**). This self-resistance mechanism involving the use of a prodrug motif was named the prodrug resistance strategy. It should be noted that, in their characterization and analysis of the zwittermicin biosynthetic pathway, Thomas and co-workers proposed that the pathway-encoded peptidase, ZmaM, would cleave an *N*-acyl-D-asparagine motif from a zwittermicin intermediate, but they did not provide experimental evidence for this post-assembly line step.[7]

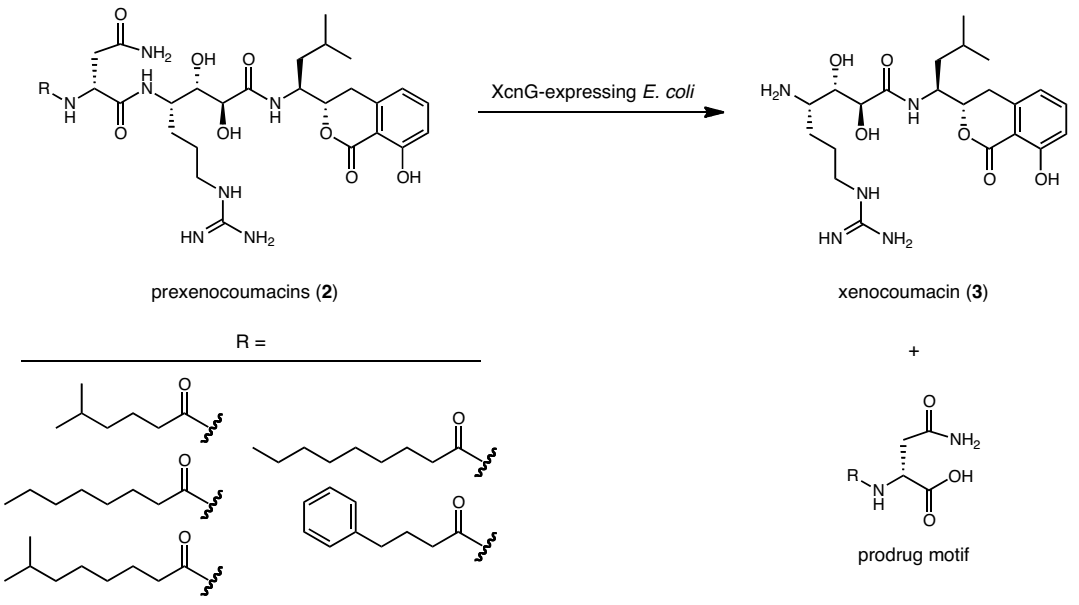


Figure 2.1: A Δ*xcnG Xenorhabdus nematophilus* strain accumulated prexenocoumacins (**2**), which have variable acyl chains. The prodrug motif was hydrolytically cleaved from prexenocoumacins by XcnG-expressing *E. coli* to provide the active molecule xenocoumacin (**3**).

Along with the peptidase XcnG, the authors also hypothesized that another uncharacterized protein found within the xenocoumacin cluster, XcnA, was responsible for the biosynthesis and incorporation of the prodrug motif. XcnA is an NRPS with two modules containing the domains C-A_1-PCP-E-C-A_2-PCP. The A_1 domain was predicted to activate L-asparagine and the E domain was predicted to epimerize a PCP domain-bound L-asparagine. These domain predictions closely matched the structure of the N-terminal prexenocoumacins. As the prodrug motif on the prexenocoumacins formed the N-terminal portion of the molecule, and the direction of NRPS assembly line chemistry is from N- to C-terminus, it was hypothesized that XcnA was the initiating module in the assembly line pathway.

While the authors did not characterize XcnA or XcnG *in vitro*, they performed a bioinformatics search to find other clusters that contained both a periplasmic peptidase like XcnG as well as an NRPS module with C-A-PCP-E domains, in which the A domain was predicted to activate either L-aspartate or L-asparagine. Surprisingly, this pair of enzymes was found in an array of otherwise unrelated biosynthetic gene clusters from both Gram-negative and Gram-positive bacteria. Gene clusters included in this list were the zwittermicin and colibactin clusters.

2.2: The *in vitro* biochemical characterization of ClbN

A thorough bioinformatic analysis of the *pks* cluster revealed that several components closely resembled those implicated in the self-resistance mechanism in xenocoumacin biosynthesis. Specifically, ClbN has the same C-A-PCP-E domain organization as the first module of XcnA, and ClbP is homologous to XcnG. When the colibactin project was initiated in our lab, there was *in vivo* evidence for the biochemical logic underlying the prodrug resistance mechanism in xenocoumacin biosynthesis, but no comprehensive *in vitro* characterization of the prodrug

synthesis and cleavage enzymes from any biosynthetic cluster. Furthermore, no structural information about colibactin or the prodrug-containing "precolibactin" was known. We decided to study the activities of the putative initiating NRPS (ClbN) and D-amino peptidase (ClbP) from the *pks* cluster, anticipating that an understanding of the prodrug resistance mechanism in colibactin biosynthesis would not only illuminate important structural features of the natural product but could also reveal strategies for inhibiting colibactin production in the human gut.

We began our biochemical characterization of the colibactin self-resistance mechanism by examining the activity of ClbN, the NRPS enzyme encoded in the *pks* island with homology to XcnA. ClbN was cloned from *E. coli* CFT073, expressed as an N-His_6-tagged fusion protein in BL21 (DE3) cells, and purified using standard Ni-affinity chromatography. A boron dipyrromethene–Coenzyme A (BODIPY–CoA) (**4**) loading assay confirmed that the PCP domain of *apo* ClbN could be post-translationally modified to the *holo* protein (Figure 2.2).[8] In this assay, the highly promiscuous ppant-transferase from the surfactin biosynthetic pathway in *Bacillus subtilis*, Sfp, is used with a fluorescent, modified CoA substrate (**4**) such that modification of the PCP domain can be observed by UV visualization of the SDS-PAGE gel.

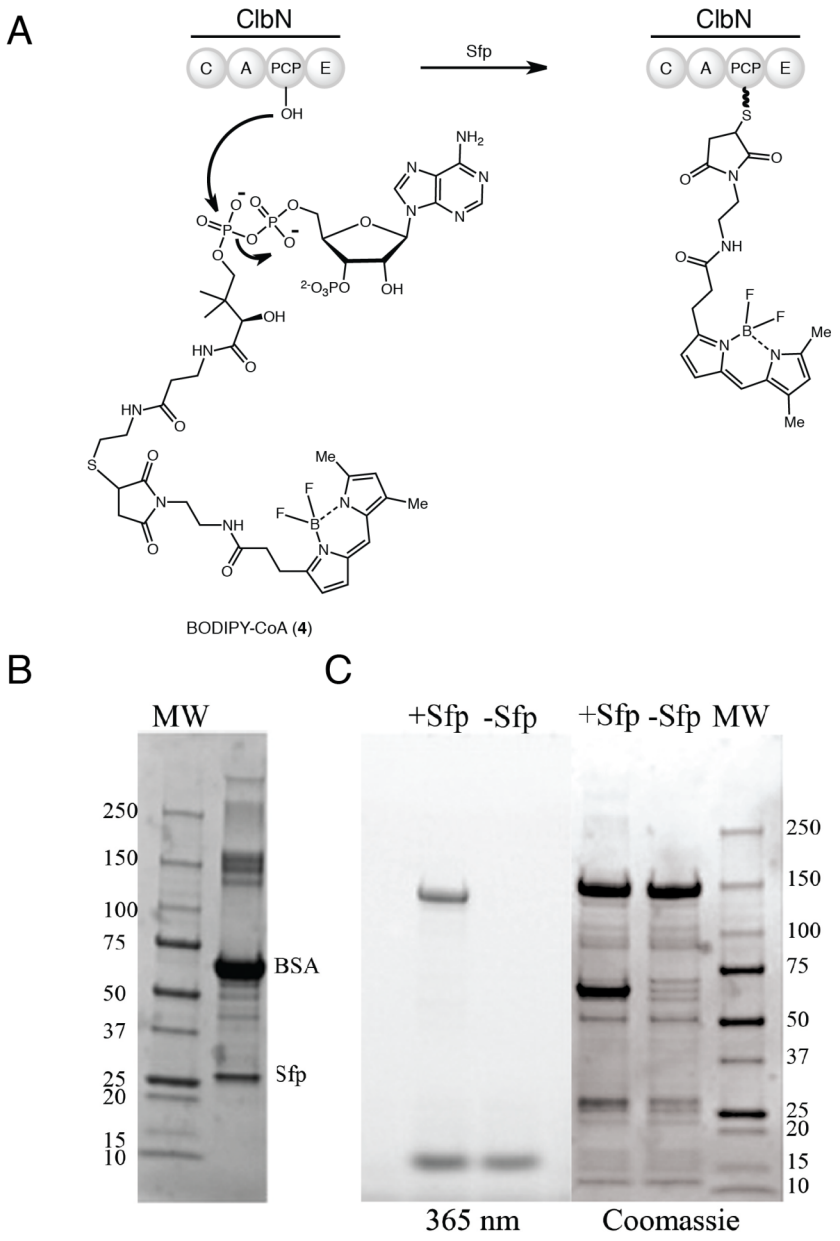


Figure 2.2: A) Reaction scheme for the BODIPY-CoA loading assay. The *apo* PCP domain of ClbN is charged with BODIPY-CoA (**4**) by the ppant-transferase Sfp. B) For this experiment, Sfp was diluted into a buffer that contained the protein Bovine Serum Albumin (BSA). C) In the presence of Sfp, ClbN was loaded with BODIPY-CoA (**4**). The molecular weight of ClbN is 165 kDa. MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

Next, we examined the substrate specificity of the ClbN A domain using an ATP–[^{32}P]PP$_i$ exchange assay. Activation of an amino acid by the A domain results in incorporation of [^{32}P]PP$_i$ into ATP, due to the equilibrium that exists between activated aminoacyl-AMP substrates, PP$_i$

and ATP. This assay utilizes the *apo* form of the protein, because use of the *holo* protein would result in stoichiometric transfer of the activated amino acid to the ppant arm of the PCP domain. The reaction is quenched with charcoal, which binds aromatic compounds including ATP. The charcoal is collected by centrifugation and subjected to scintillation counting. This assay revealed that the ClbN A domain was highly selective for L-asparagine, which matched the selectivity observed for ZmaO, the homologous NRPS module from the zwittermicin biosynthetic gene cluster as well as the structure of the prodrug motif of the prexenocoumacins (Figure 2.3).[5,7] No activity was seen with D-asparagine, L-aspartate or L-alanine.

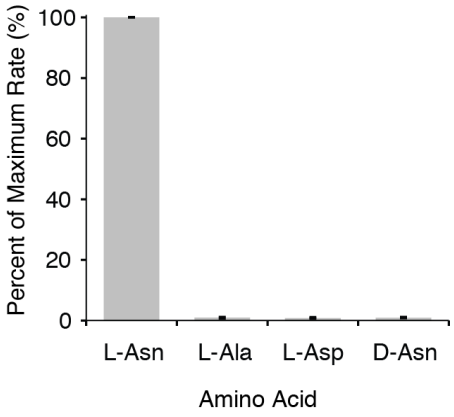


Figure 2.3: An ATP-[^{32}P]PP$_i$ assay showed the preference of the ClbN A domain for L-asparagine. Bar graphs represent the mean ± SD of three independent experiments.

After activation of L-asparagine, the A domain should transfer the aminoacyl-AMP to the ppant arm of the T domain of ClbN. This process was studied using a radiometric loading assay with ^{14}C-labeled L-asparagine. This assay demonstrated that the T domain of *holo* ClbN is charged with L-asparagine in the presence of Sfp and ATP (Figure 2.4). In addition, this assay revealed that about 18% of the total enzyme pool was loaded with ^{14}C-labeled L-asparagine, which is line with literature reports of other NRPS T domain loading assays.[9] No loading was seen in the absence of ATP and Sfp, demonstrating that these components were necessary for transfer to the T domain.

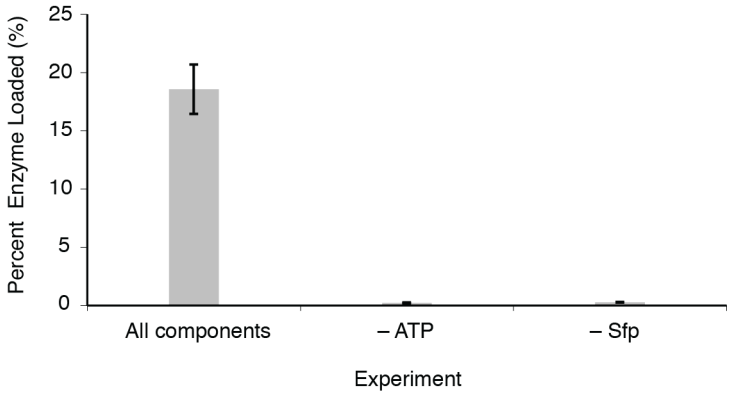


Figure 2.4: Loading of the T domain of ClbN with ^{14}C-L-asparagine. Bar graphs represent the mean ± SD of three independent experiments.

We next investigated the ClbN C domain, which was hypothesized to acylate the free amine of the PCP domain-bound asparagine. This hypothesis was based on the structure of the prodrug motif from xenocoumacin biosynthesis as well as a bioinformatic analysis. Multiple sequence alignments with characterized C domains revealed that the ClbN C domain possesses amino acid motifs characteristic of starter C domains involved in lipoinitiation (Figure 2.5).[10] Lipoinitiation is the initiation of an assembly line pathway by a module containing a starter C domain, A and PCP domains, which together construct an *N*-acylated, PCP domain-bound amino acid. More commonly, assembly line biosynthesis is initiated by a module containing only A and PCP domains, which provides a free amino acid residue at the N-terminus of the natural product scaffold.[11] Lipoinitiation is seen in the biosynthesis of lipopeptides such as surfactin.[12] Starter C domains that perform lipoinitiation must use an activated fatty acid to install the *N*-acyl moiety, and various strategies exist to generate the activated acyl building block. For instance, some starter C domains use fatty acyl thioesters bound to *trans*-acting ACP domains, which are usually encoded within the same biosynthetic cluster, while other starter C domains utilize fatty acyl-CoA thioesters, which may come from the primary metabolism pool.[12] We hypothesized that ClbN would accept fatty acyl-CoA substrates because the *pks* island does not encode any homologues

of fatty acid-activating enzymes. Furthermore, the isolation of multiple prexenocoumacins with variable acyl groups suggested that the initial *N*-acylation event may be promiscuous.

Figure 2.5: ClustalW2 alignment of the ClbN C domain with two $^{L}C_{L}$ domains and two starter C domains. C domain sequences were excised from the full length amino acid sequences using the University of Maryland's PKS/NRPS web-based tool.[13] The GenBank accession numbers and excised amino acid residues used for the analysis were: DptA C-4: AAX31557.1, 3677..4118; SrfAB C-2: NP_388231.1, 1055..1473; SrfAA C-1: NP_388230.1, 5..435; GlbF C-1: CAL80824.1, 10..438; ClbN: Q0P7K4, 1..432. DptA C-4[14] and SrfAB C-2[15] are standard $^{L}C_{L}$ domains, while SrfAA C-1[12] and GlbF[16] are starter C domains. The C domain sequences were aligned using the ClustalW2 alignment tool. Sequence motifs that were determined to differentiate various types of C domains are indicated with green boxes and specific amino acids within these motifs that distinguish $^{L}C_{L}$ domains and starter C domains are numbered.[10] The amino acid positions, and the key differences and similarities at each position that were used in our analysis, are: (**1**) P vs. A/G; (**2**) T vs. L/A/M; (**3**) D vs. A/R/H; (**4**) A vs. V/I/L; (**5**) V/A vs. P.

To address these questions, we examined the activity of ClbN toward a panel of fatty acyl-CoAs that could come from primary metabolism, specifically fatty acid catabolism, in *E. coli*. We chose to examine only straight chain, saturated, and even numbered fatty acyl-CoAs for the

reconstitution experiments with ClbN. Although xenocoumacin-producing *Xenorhabdus* strains are Gram-negative organisms that are related to *E. coli* and in the family *Enterobacteriaceae*, *Xenorhabdus* is known to synthesize a much wider range of fatty acids, including branched chain acids.[17] We hypothesized that *E. coli* would not synthesize the same diversity of prodrug scaffolds seen in xenocoumacin biosynthesis, because *E. coli* cannot access the same diversity of fatty acyl substrates.[18] ClbN was first reconstituted with Sfp and CoA to obtain the *holo* enzyme, and *holo*-ClbN was then incubated with L-asparagine, ATP, and individual fatty acyl-CoA substrates. Products were hydrolyzed from the ppant arm of ClbN using an aqueous solution of potassium hydroxide (0.1 M) and analyzed by high-resolution LC-MS. This analysis revealed that the enzyme accepted fatty acyl-CoAs with chain lengths ranging from six to fourteen carbons, to give compounds **6**-**10**. (Figure 2.6).

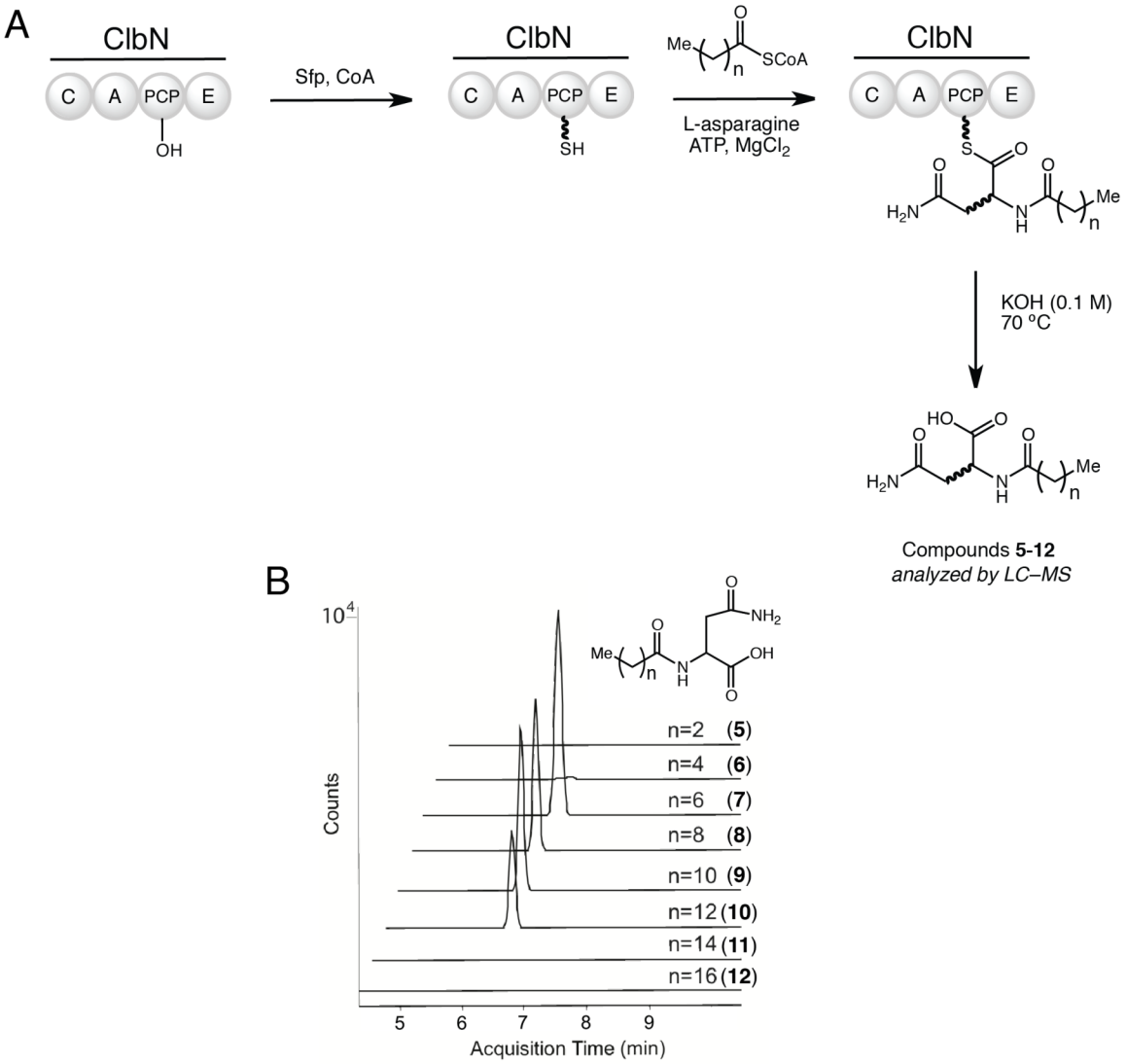


Figure 2.6: A) Reaction scheme for the *in vitro* reconstitution reaction of ClbN with individual fatty-acyl CoA substrates (n = 2-16). B) Extracted ion chromatograms (EICs) for the products (**5-12**) of the condensation reaction of ClbN with L-asparagine and individual fatty acyl-CoA substrates.

We then examined the specificity of the C domain through a competition assay, in which *holo* ClbN was reconstituted with L-asparagine, ATP, and an equal mixture of fatty acyl-CoAs each at a final concentration of 187 μM with chain lengths from four to eighteen carbons. As before, products were hydrolyzed from the ppant arm using an aqueous solution of potassium hydroxide (0.1 M) and analyzed by LC-MS. Strikingly, under these conditions ClbN preferentially utilized myristoyl-CoA for *N*-acylation (Figure 2.7) to give **10**. A small amount of the product (**9**) from incorporation of lauroyl-CoA was also observed in this assay.

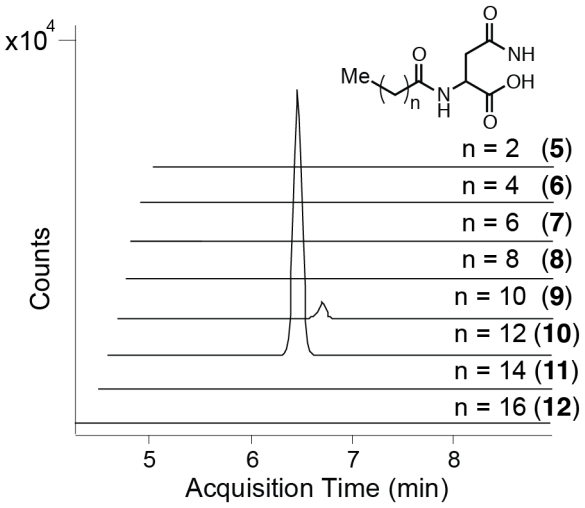


Figure 2.7: EICs for the products (**5-12**) of fatty acyl-CoA competition experiment involving the *in vitro* reconstitution of ClbN with L-asparagine and a panel of individual fatty acyl-CoA substrates.

To explore whether the substrate specificity observed in our *in vitro* reconstitution experiments represented the *in vivo* activity of ClbN, we compared metabolites produced by strains expressing either ClbN or an empty vector. For this assay we utilized the *E. coli* strain BAP1 to ensure that ClbN would be post-translationally modified to its *holo* form. The BAP1 strain is commonly used in metabolic engineering of NRPS and PKS pathways, and was obtained by Khosla and co-workers by integration of the *sfp* gene, under the control of the T7 RNA polymerase promoter, into the propionate catabolism operon (*prp*) of *E. coli* BL21 (DE3).[19] We identified both *N*-lauroyl-asparagine (**9**) and *N*-myristoyl-asparagine (**10**) in extracts of ClbN-expressing BAP-1 cells, and the myristoylated product (**10**) was much more abundant in these extracts (Figure 2.8). These metabolites were not found in the extracts of *E. coli* harboring an empty expression vector. Shorter chain *N*-acyl-asparagine products (**5-8**) were not detected in either strain. These results suggested that the selectivity observed in the *in vitro* competition assay reflects ClbN's substrate preference *in vivo*.

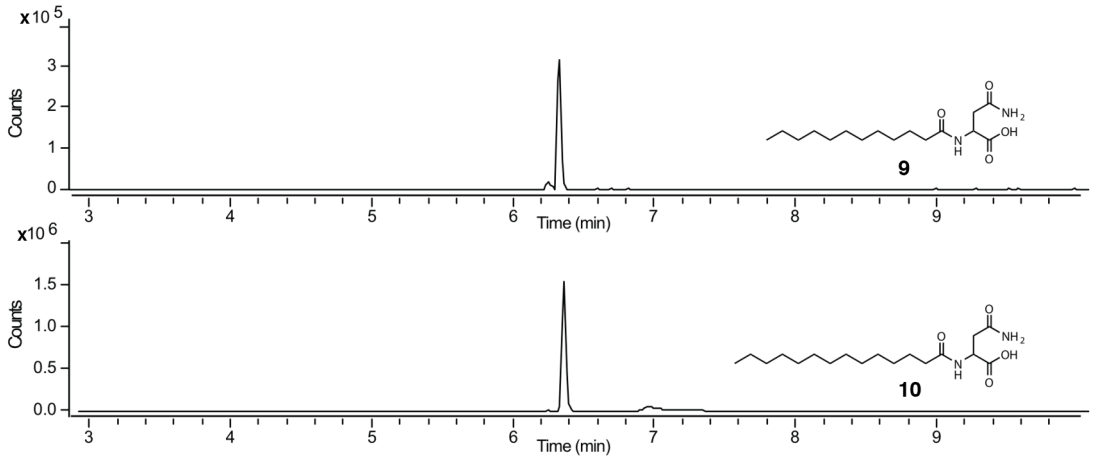


Figure 2.8: EICs of the products (**9**, **10**) from the *in vivo* experiment investigating the condensation activity of ClbN. Results shown are from the ClbN-expressing *E. coli* extracts, and are representative of three independent experiments. Please note that the scale of these plots is not the same.

Finally, we examined the activity of the E domain of ClbN. E domains epimerize the α-carbon of T domain-bound aminoacyl substrates. This is achieved by deprotonation of the α-carbon of an L-aminoacyl substrate by an enzyme-derived base to form an enolate, followed by reprotonation of the enolate to provide the D/L-aminoacyl product.[20] To examine the activity of the E domain, *holo* ClbN was reconstituted with L-asparagine, ATP and either myristoyl- or octanoyl-CoA in D_2O. We rationalized that if the asparagine α-carbon was subject to deprotonation, exchange of the enzyme active site with the solvent would result in incorporation of a single deuterium at the α-carbon. Products were hydrolyzed from the ppant arm using an aqueous solution of potassium hydroxide (0.1 M) and analyzed using LC-MS. This analysis confirmed the incorporation of single deuterium into these products resulting from exchange during epimerization (Figures 2.9-2.10). Incorporation of a deuterium into the products was only seen in samples that contained all of the reaction components and D_2O. The stereochemistry of the products was not determined.

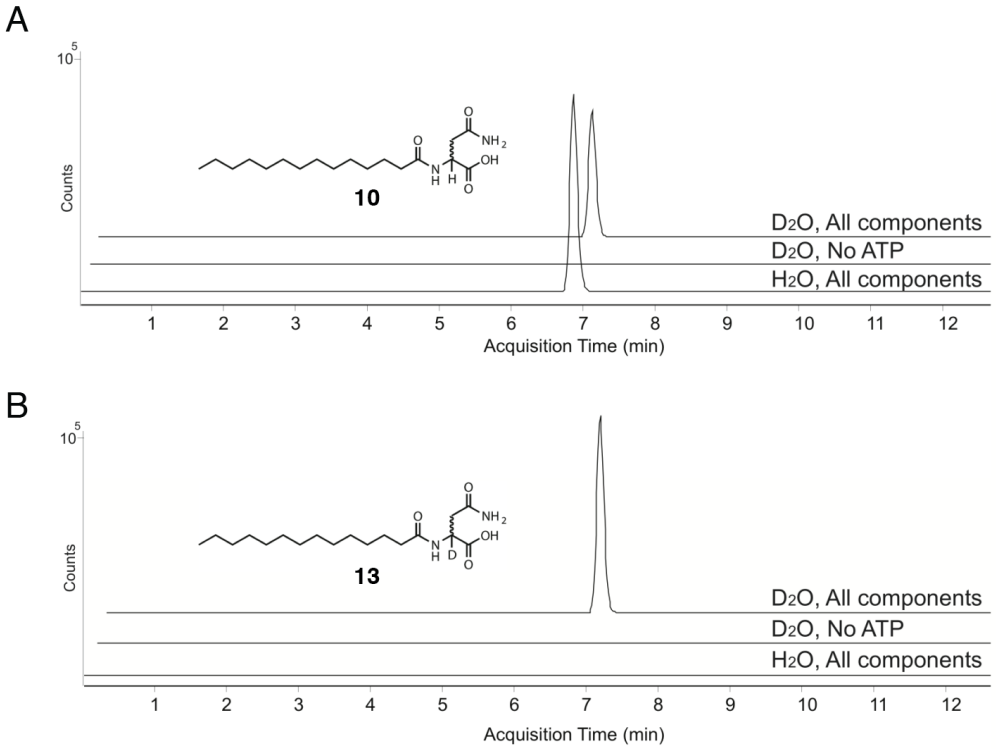


Figure 2.9: EICs of the products from the ClbN E domain assays with myristoyl-CoA and L-Asn in H_2O and D_2O. A) EICs of the protonated product (**10**). B) EICs of the deuterated product (**13**).

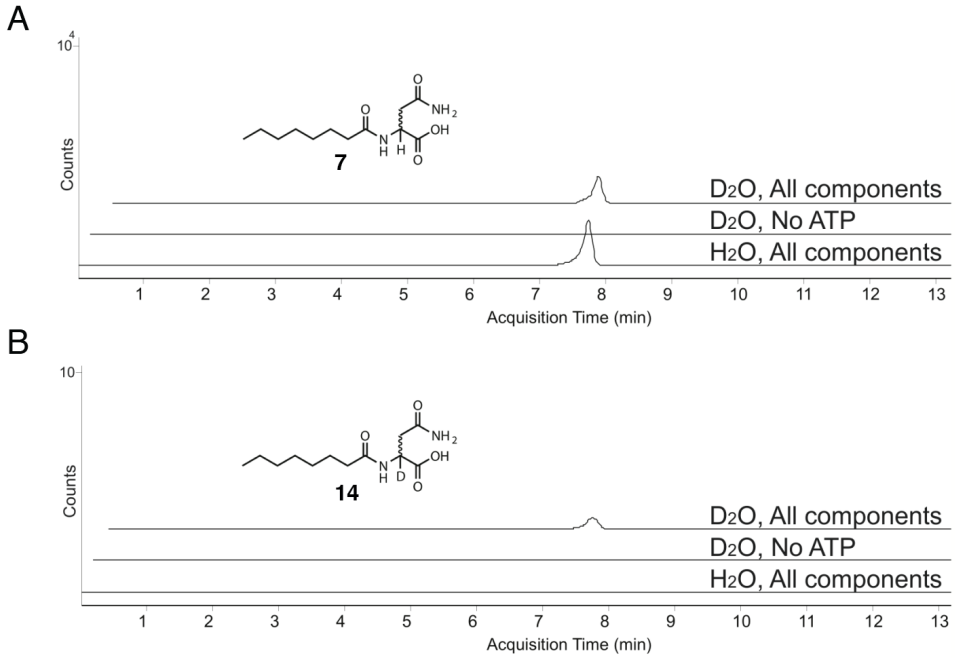


Figure 2.10: EICs of the products from the ClbN E domain assays with octanoyl-CoA and L-Asn in H_2O and D_2O. A) EICs of the protonated product (**7**). B) EICs of the deuterated product (**14**).

2.3: The *in vitro* biochemical characterization of $ClbB_{NRPS}$

Having established the activity of each of the domains of ClbN, we turned to identifying the enzyme responsible for elongating the prodrug motif biosynthesized by ClbN. Based on the underlying logic of the prodrug resistance strategy, we hypothesized that next module after ClbN was an NRPS. This NRPS could construct the amide bond that serves as an eventual peptidase substrate for the prodrug-cleaving peptidase. The fact that the last domain of ClbN is an E domain provided important information that aided in the identification of the next assembly line enzyme. Condensation domains that follow E domains and catalyze amide bond formation between an upstream D-aminoacyl thioester electrophile and a downstream T domain-bound L-aminoacyl nucleophile are known as $^{D}C_{L}$ domains.[21] These domains have signature amino acid motifs that differentiate them from $^{L}C_{L}$ and starter C domains.[9] A bioinformatic analysis of the NRPS C domains in the *pks* cluster identified the C domain of ClbB, a two-module NRPS/PKS hybrid, as the only $^{D}C_{L}$ domain (Figure 2.11). This analysis strongly suggested that ClbB follows ClbN in the assembly line.

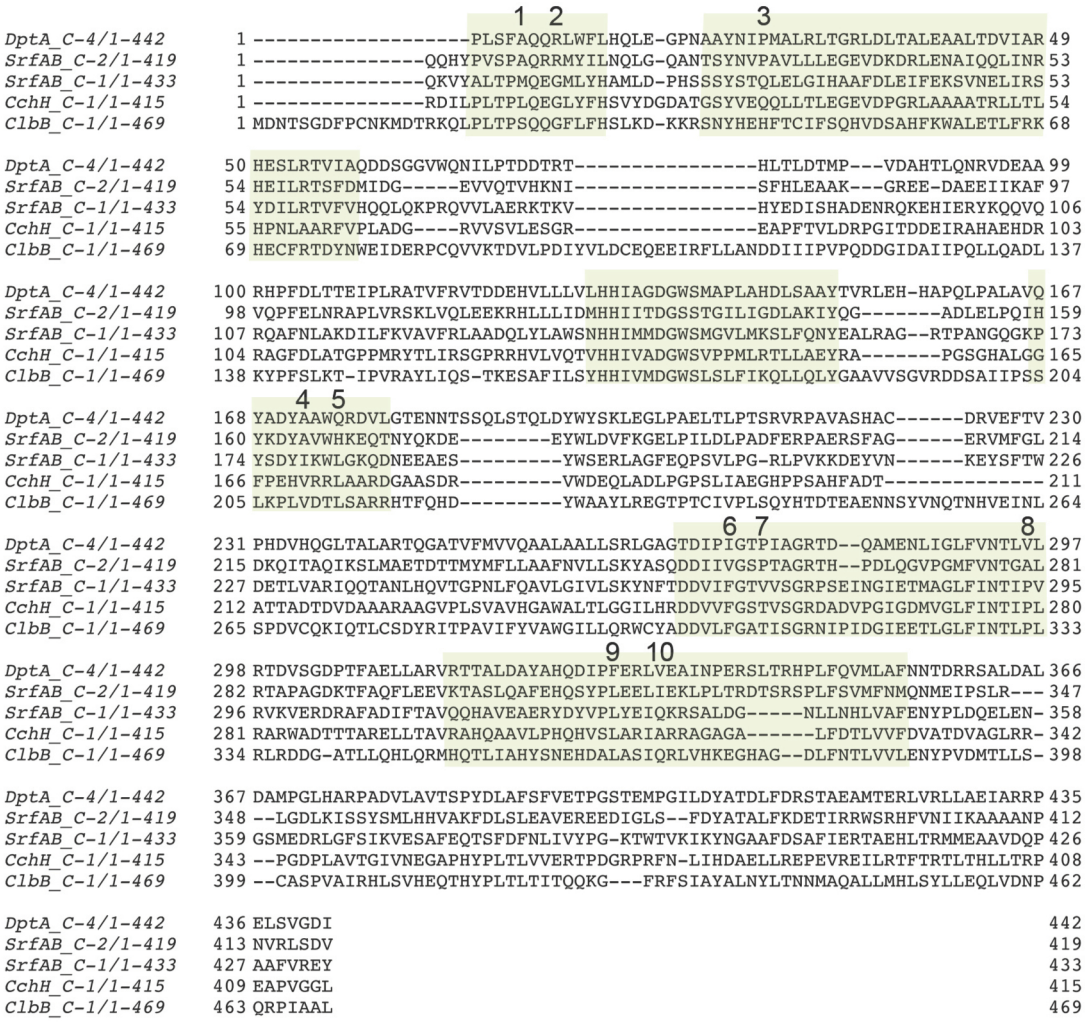


Figure 2.11: ClustalW2 alignment of the ClbB C domain with two $^{L}C_{L}$ domains and two $^{D}C_{L}$ domains. The C domains were excised and aligned as described above for ClbN. The GenBank accession numbers and excised amino acid residues used for the analysis were: DptA C-4: AAX31557.1, 3677..4118; SrfAB C-2: NP_388231.1, 1055..1473; CchH C-1: NP_624809.1, C-1 1192..1606; SrfAB C-1: NP_388231.1, 7..439; ClbB: Q0P7J2, 1..469. CchH C-1[22] and SrfAB C-1[15] are $^{D}C_{L}$ domains. Sequence motifs that differentiate various types of C domains are indicated with green boxes and specific amino acids within these motifs that distinguish $^{L}C_{L}$ domains and $^{D}C_{L}$ domains are numbered.[10] The amino acid positions, and the key differences and similarities at each position that were used in our analysis, are: (1) A vs. M/L/S; (2) R vs. G; (3) P vs. Q/H; (4) A vs. I/V; (5) Q/H vs. L; (6) I/V vs. F; (7) P vs. V/T; (8) V/A vs. P; (9) F vs. L; (10) V/I vs. Q/A.

The NRPS portion of ClbB ($ClbB_{NRPS}$), containing the first 1123 amino acids of the protein, was cloned from *E. coli* CFT073, overexpressed as an N-His_6-tagged fusion protein in BL21 (DE3) cells, and purified using Ni-affinity chromatography. A BODIPY−CoA loading assay confirmed

that the T domain of *apo* $ClbB_{NRPS}$ could be post-translationally modified to the *holo* protein in the presence of Sfp (Figure 2.12).

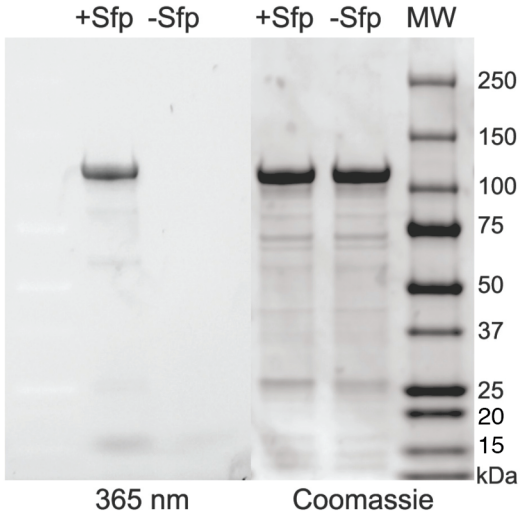


Figure 2.12: Loading of $ClbB_{NRPS}$ with BODIPY-CoA by Sfp. The molecular weight of $ClbB_{NRPS}$ is 124 kDa. MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

An ATP–[^{32}P]PP_i exchange assay revealed that the $ClbB_{NRPS}$ A domain accepts multiple amino acids (Figure 2.13A). $ClbB_{NRPS}$ displayed the highest activity with L-alanine, but also activated L-valine, L-serine, and glycine. Interestingly, the promiscuous activity of the A domain extended to T domain loading. Radiometric assays showed that both ^{14}C-labeled L-alanine and ^{14}C-labeled L-valine were loaded onto the T domain (Figure 2.13B). With ^{14}C-labeled L-alanine, the percent of enzyme in the reaciton mixture loaded with the amino acid was around 25%, whereas only about 10% of the T domains in solution were bound to ^{14}C-labeled L-valine.

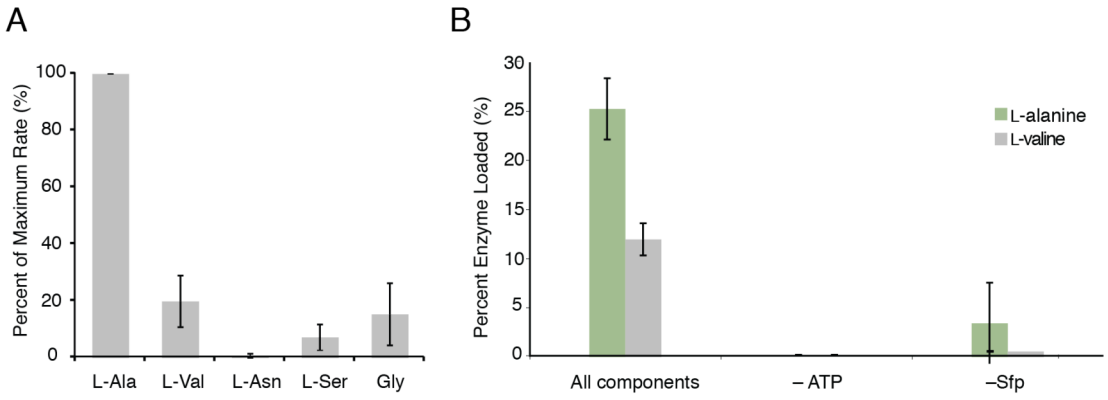


Figure 2.13: A) An ATP-[^{32}P]PP$_i$ assay showed the preference of the ClbB$_{NRPS}$ A domain for L-alanine. L-valine and glycine were also activated at lower levels. B) Loading of the T domain of ClbB with ^{14}C-L-alanine and ^{14}C-L-valine. Bar graphs represent the mean ± SD of three independent experiments.

The hypothesis that ClbB$_{NRPS}$ follows ClbN in the colibactin assembly line was tested by reconstituting ClbB$_{NRPS}$ and ClbN *in vitro* with all of the necessary biosynthetic building blocks (Figure 2.14A). ClbB$_{NRPS}$ and ClbN were incubated with CoA and Sfp to convert the enzymes to the *holo* form. Next, L-asparagine, L-alanine or L-valine, ATP and a variety of fatty acyl-CoA substrates were added to the *holo* enzymes. As described previously, products were hydrolyzed from the ppant arm of the T domains using an aqueous solution of potassium hydroxide (0.1 M) and were analyzed by LC-MS. *N*-acylated dipeptide products (**17-19**) were detected in the presence of octanoyl- and myristoyl-CoA (Figure 2.14B). No elongation products were detected in reactions containing butanoyl- and hexanoyl-CoA (**15-16**). These data confirmed the hypothesis that ClbB$_{NRPS}$ could elongate the prodrug motif constructed by ClbN and supported the theory that ClbB follows ClbN in the colibactin assembly line.

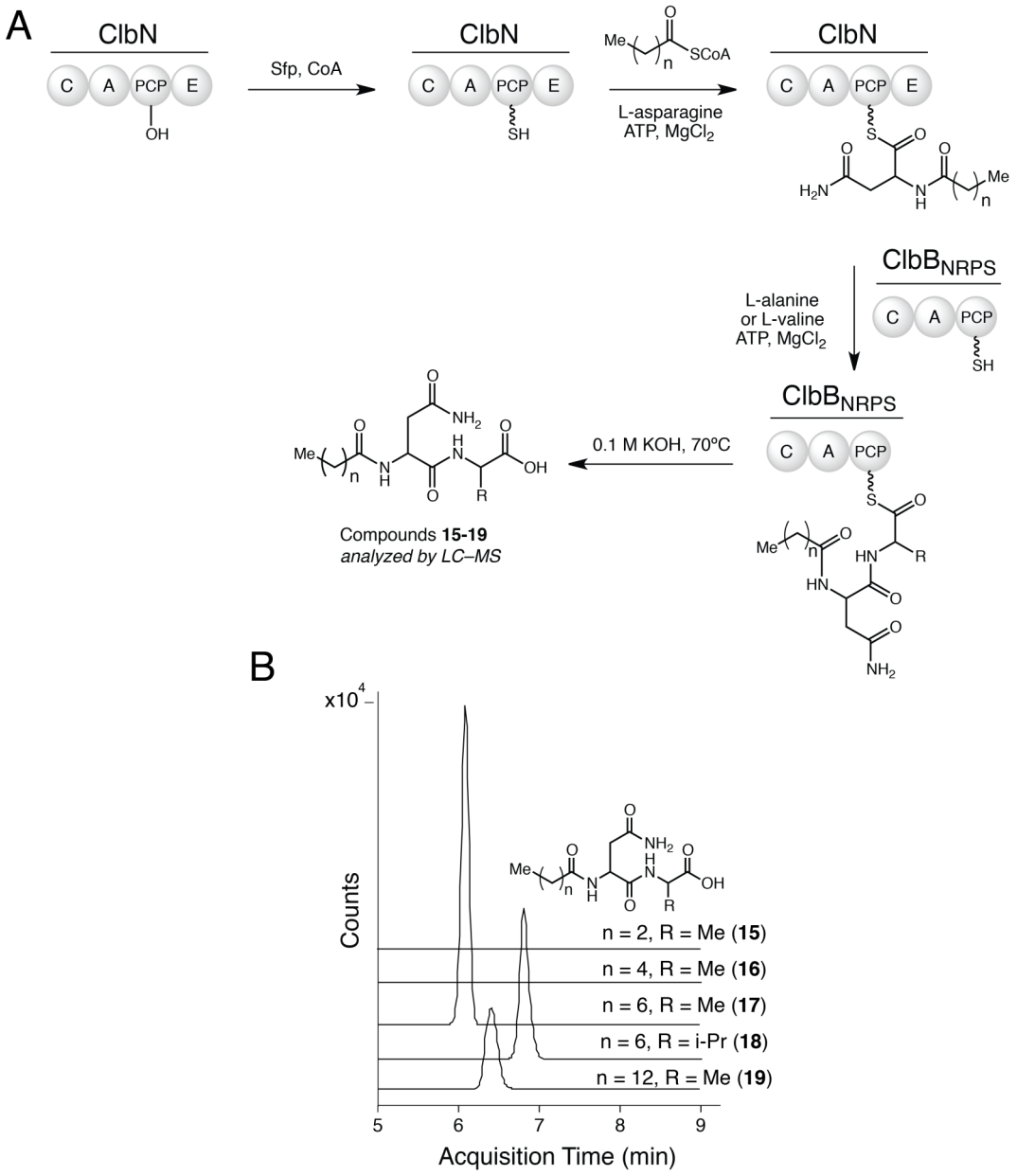


Figure 2.14: A) A reaction scheme showing the reconstitution of ClbN and $ClbB_{NRPS}$ with the required building blocks and cofactors. B) Extracted ion chromatograms of the reconstitution reaction shows that dipeptide products were observed when the acyl chain length is greater than six carbons. Both L-alanine and L-valine are incorporated by $ClbB_{NRPS}$.

To examine the preference for the fatty acyl-CoA chain length, a competition experiment, in which both enzymes were incubated with L-alanine and an equal concentration (375 μM) of octanoyl- and myristoyl-CoA, was performed. In this assay, the myristoylated product predominated, reflecting the preference of the ClbN C domain for the myristoyl-CoA substrate

(Figure 2.15). Together, these data indicated that although ClbB is promiscuous with respect to both the chain length of the *N*-acyl group on the intermediate it accepts from ClbN and the amino acid it uses for elongation, the selectivity of the initial *N*-acylation performed by ClbN controls the structure of the prodrug motif. Furthermore, the use of multiple amino acids by $ClbB_{NRPS}$ suggested the possibility that colibactin could be a set of related metabolites with different N-terminal amino acids.

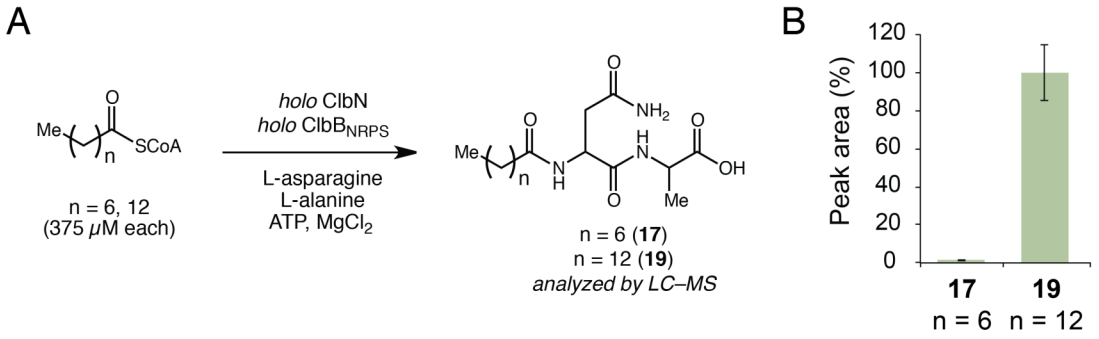


Figure 2.15: A) A reaction scheme showing the competition assay in which ClbN and $ClbB_{NRPS}$ were reconstituted with octanoyl- and myristoyl-CoA. B) A bar graph showing the relative peak areas of the EICs of the two possible dipeptide products (**17**, **19**) resulting from the competition reaction. Bar graphs represent the mean ± SD of three independent experiments.

2.4: Characterization of the prodrug cleaving enzyme ClbP

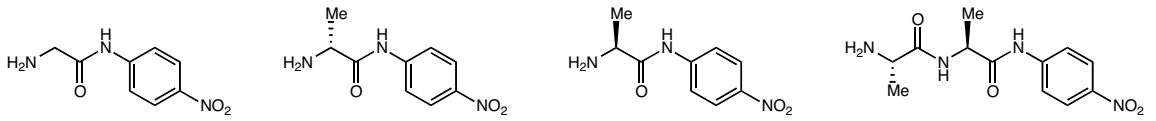
Initial *in silico* analyses revealed both the putative function and topology of ClbP. A bioinformatics analysis of ClbP using the Basic Local Alignment Tool and the Conserved Domains Database from the National Center for Biotechnology Information indicated that ClbP has a conserved domain that is found in the β-lactamase family of D-amino peptidases, which includes the well-studied peptidase AmpC. AmpC, a member of the Class C group of β-lactamases, hydrolyzes the β-lactam ring of cephalosporin and other β-lactam antibiotics using an active-site serine residue to produce biologically inactive compounds.[23] The homology to AmpC indicated that ClbP may cleave a peptide bond containing a D-amino acid residue.

Secondary structure prediction, performed using online programs such as Phobius[24] and PSORTb,[25] indicated that ClbP contains an N-terminal signal sequence that targets the protein to the inner membrane, a periplasmic domain, and three C-terminal transmembrane helices. This analysis suggested that ClbP has a similar topology to XcnG, the peptidase implicated in the hydrolysis of prexenocoumacins.

Genetic studies showed that ClbP is required for the genotoxicity of *pks*$^+$ *E. coli*.[26] In addition, complementation experiments demonstrated that genotoxicity could be restored in a *clbP* knock-out strain by a full-length ClbP construct, which contained the N-terminal signal sequence and all three transmembrane helices.[27] However, it was also shown in these complementation studies that a construct containing only the periplasmic peptidase domain of ClbP ($ClbP_{pep}$) was not able to restore genotoxicity, despite the fact that the catalytic residues of ClbP are located exclusively in the periplasmic domain.[27] Previously, $ClbP_{pep}$ was crystallized and its active site was observed to contain a serine residue (S95), predicted to serve as the nucleophile in peptide bond hydrolysis, as well as tyrosine (Y186) and lysine (K98) residues implicated in the active site hydrogen bonding network.[28] These active site residues (S95, Y186 and K98) were shown to be required for the activity of ClbP using genotoxicity complementation experiments with active-site point mutants of ClbP.[28]

The peptidase activity of $ClbP_{pep}$ was examined in a fluorogenic assay using *p*-nitroaniline-containing substrates (**20**-**23**), which were hydrolyzed with very low catalytic efficiencies by $ClbP_{pep}$ (Figure 2.16).[28] The low reactivity of $ClbP_{pep}$ toward these substrates could have arisen from the fact that $ClbP_{pep}$ is not a competent peptidase, as indicated by the complementation studies, or from the fact that these substrates are not similar enough to the natural precolibactin substrate. We also hypothesized that this low activity could have arisen from the activity of

endogenous protease or peptidase contaminants in their preparation of $ClbP_{pep}$. Beyond the complementation studies described above, there was no literature report that examined the ability of full-length ClbP to process any substrates *in vivo* or *in vitro*.



Glycine-*p*-nitroaniline (**20**) 0.005 μM^{-1} min^{-1} | D-alanine-*p*-nitroaniline (**21**) 0.018 μM^{-1} min^{-1} | L-alanine-*p*-nitroaniline (**22**) 0.020 μM^{-1} min^{-1} | L-alanine-L-alanine-*p*-nitroaniline (**23**) 0.040 μM^{-1} min^{-1}

Figure 2.16: The activity of $ClbP_{pep}$ was examined by Bonnet and co-workers using an *in vitro* fluorogenic assay that measured the release of the *p*-nitroaniline group upon amide bond hydrolysis. The k_{cat}/K_M values are given below each compound.

Thus, we were interested in characterizing the peptidase activity of both ClbP and $ClbP_{pep}$ more completely than in previous studies. Our initial efforts focused on $ClbP_{pep}$, which was cloned from *E. coli* CFT073 and expressed in the periplasm. The periplasmic expression was achieved using a construct containing the N-terminal signal sequence and a C-terminal His_6 tag. $ClbP_{pep}$ was purified from the periplasm using cold-osmotic shock to release periplasmic components followed by standard Ni-affinity chromatography. An active-site point mutant, $ClbP_{pep}$-S95A, was obtained using site-directed mutagenesis, and this mutant protein was also purified from the periplasm (Figure 2.17).

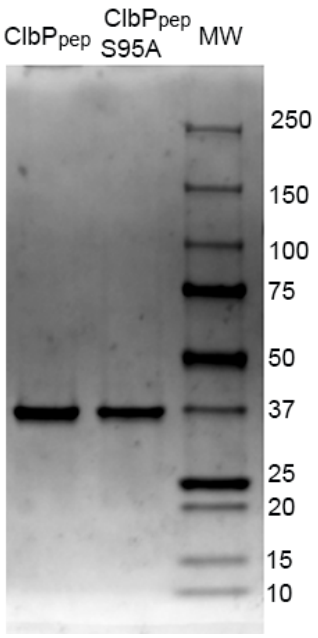


Figure 2.17: SDS-PAGE of $ClbP_{pep}$-C-His_6 and $ClbP_{pep}$-S95A-C-His_6. The molecular weight of these constructs is 42 kDa. (MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad)).

To begin the *in vitro* characterization of $ClbP_{pep}$, we first sought to replicate the literature results obtained with the *p*-nitroaniline-containing substrates. Under identical reaction conditions, but using different protein constructs and purification conditions as the literature report, we were unable to detect any cleavage of D-alanine and L-alanine-*p*-nitroanilides (**21** and **22**) by $ClbP_{pep}$ (Figure 2.18). These negative results led us to hypothesize that the inability of $ClbP_{pep}$ to process these substrates was due to the lack of structural similarity to the predicted N-terminal structure of precolibactin.

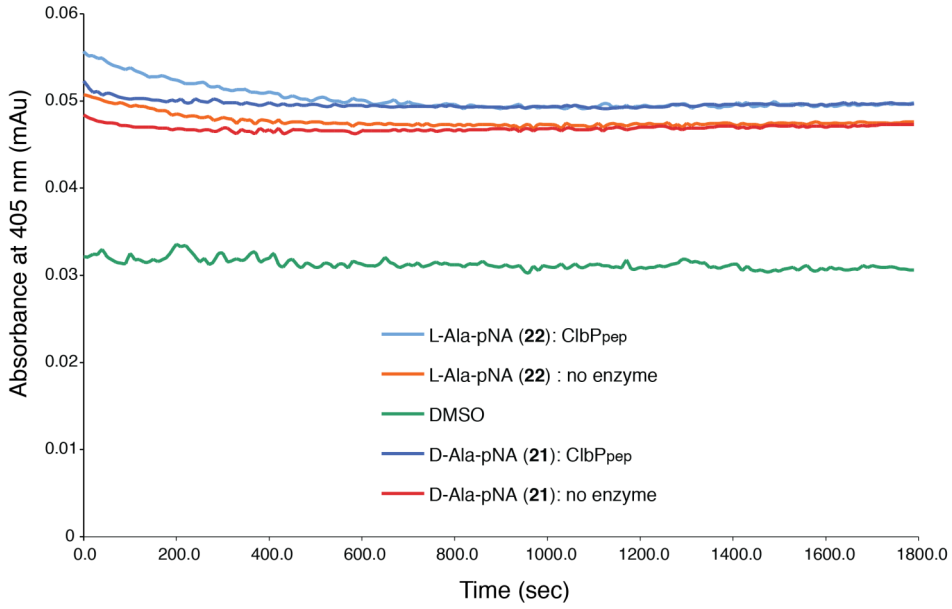


Figure 2.18: A *p*-nitroaniline release assay showed no activity of $ClbP_{pep}$ towards **21** and **22.**

We synthesized a panel of substrates (**24-30**) that resembled structures biosynthesized by ClbN and ClbB *in vitro* (Figure 2.19). To examine the activity of $ClbP_{pep}$ toward these dipeptide-containing substrates, a fluorenylmethyloxycarbonyl chloride (Fmoc-Cl) derivatization assay was used (Figure 2.20).[29] In this assay, cleavage of the amide bond releases a free amino acid that is derivatized with Fmoc-Cl under alkaline conditions. The products can be analyzed using high performance liquid chromatography (HPLC) with UV detection at 259 nm. $ClbP_{pep}$ or the S95A active-site mutant was incubated for 24 hours with the model substrates. This assay revealed that $ClbP_{pep}$ cleaves substrates containing *N*-octanoyl- or *N*–decanoyl-D-asparagine (**24, 25** and **27**) (Figures 2.21-2.26). $ClbP_{pep}$-S95A was not able to cleave these substrates, confirming that the observed activity was due to $ClbP_{pep}$. $ClbP_{pep}$ did not accept an *N*-octanoyl-L-asparagine-containing substrate (**26**). Surprisingly, $ClbP_{pep}$ also failed to process substrates with an acyl chain length greater than ten carbons (**28, 29**). Overall, this assay demonstrated that the substrate specificity displayed by $ClbP_{pep}$ did not match the product distribution observed in the ClbN

competition experiment. This discrepancy led to the hypothesis that the behavior of the truncated enzyme $ClbP_{pep}$ might not reflect the activity of the full-length enzyme *in vivo*.

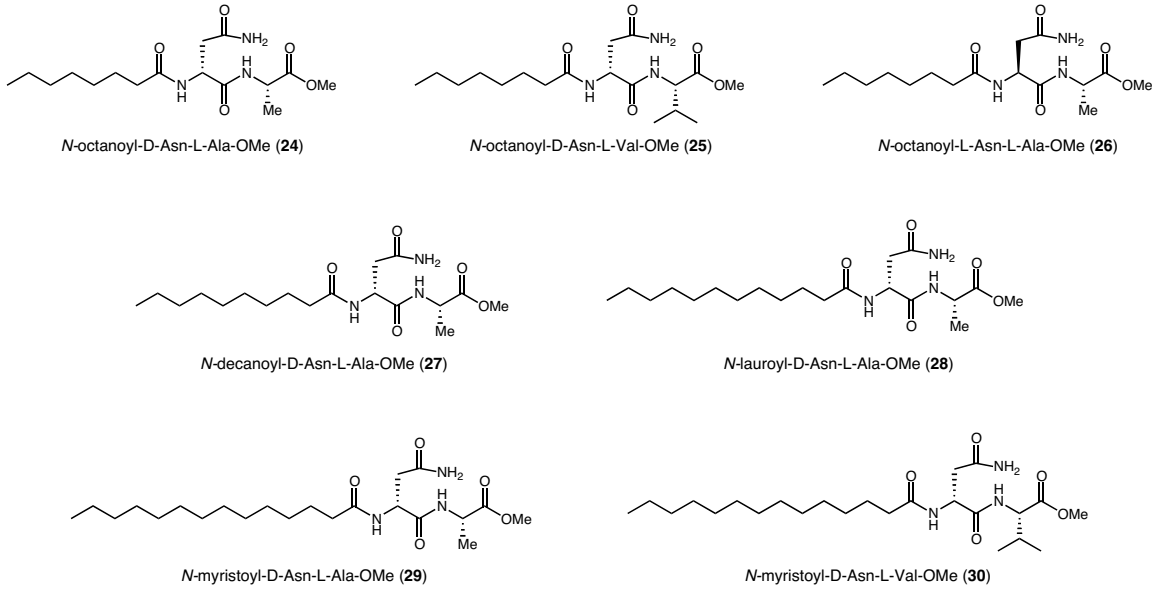


Figure 2.19: A panel of *N*-acylated dipeptides (**24-30**) that resemble the N-terminus of precolibactin were synthesized. Acyl chain lengths varied from eight to fourteen carbons and both L-alanine and L-valine were installed at the C-terminus.

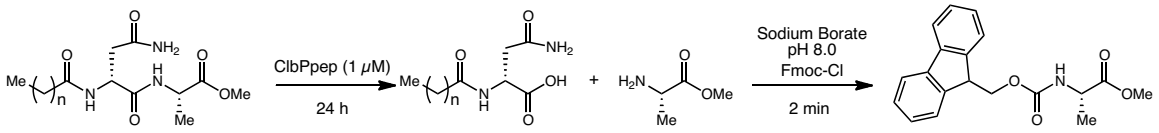


Figure 2.20: The products of amide hydrolysis are an *N*-acylated amino acid and an amino acid with a free primary amine, which is reacted with Fmoc-Cl in a pH 8.0 aqueous solution containing sodium borate.

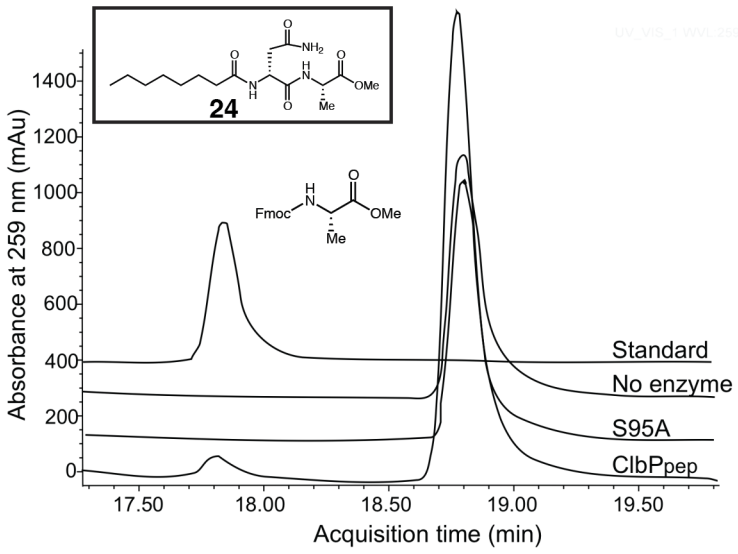


Figure 2.21: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **24**

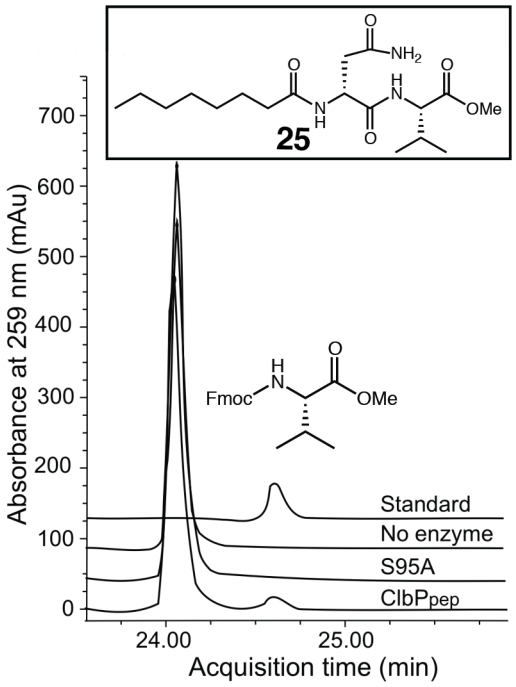


Figure 2.22: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **25**.

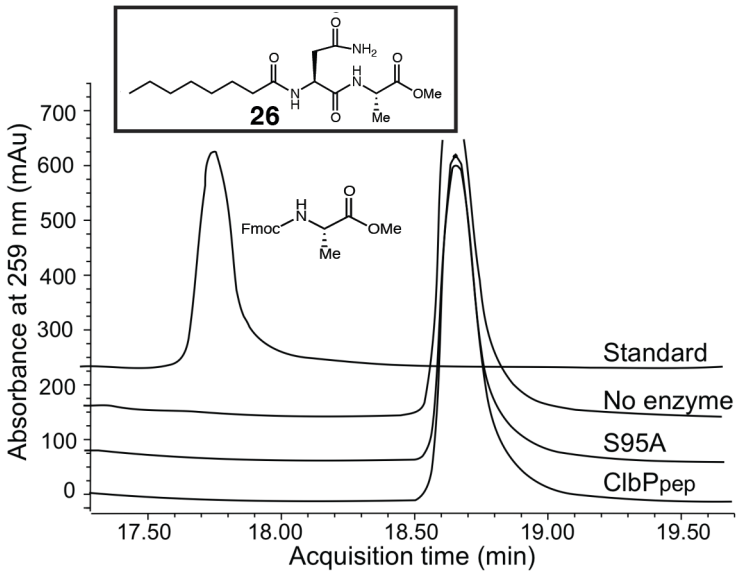


Figure 2.23: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **26**.

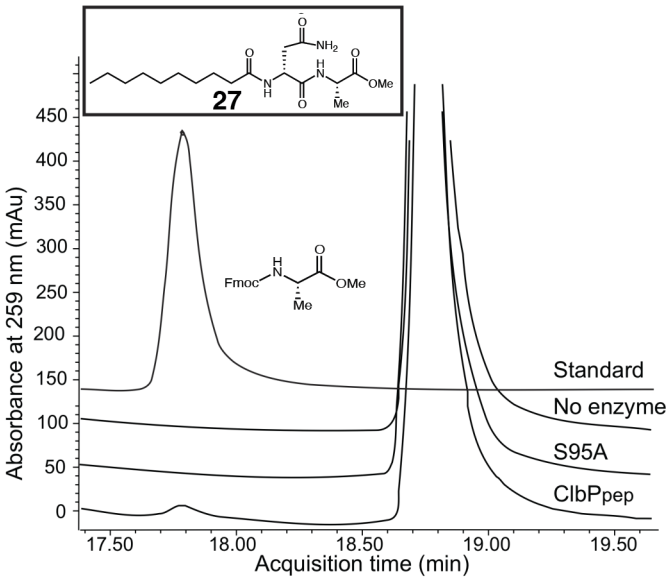


Figure 2.24: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **27**.

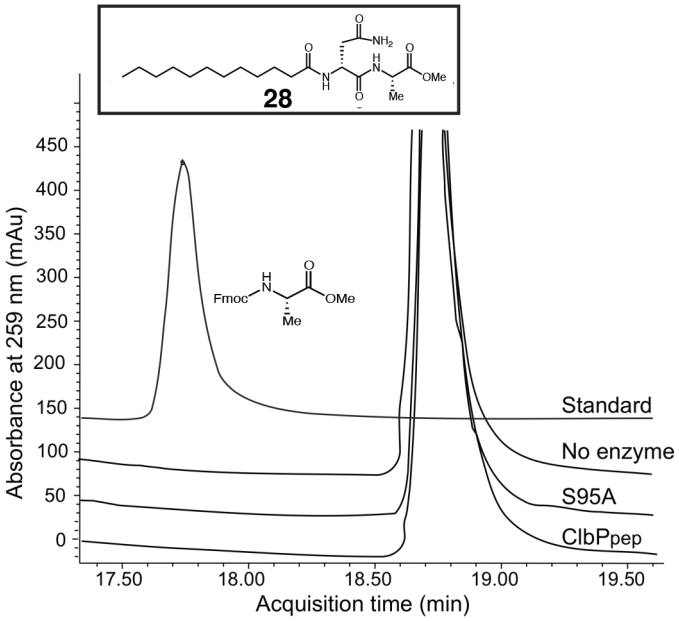


Figure 2.25: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **28**.

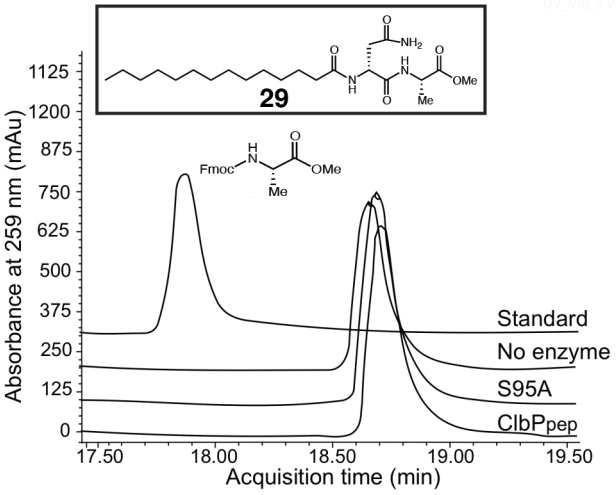


Figure 2.26: HPLC traces for the *in vitro* reaction of $ClbP_{pep}$ and $ClbP_{pep}$-S95A with **29**.

To examine this hypothesis, a whole-cell feeding assay with model substrates was performed. First, full-length ClbP and the S95A point mutant were cloned from *E. coli* CFT073. These constructs did not have an affinity tag, and were not purified for *in vitro* studies. In addition to ClbP, we were also interested in studying the activity of the peptidase ZmaM, which is encoded in

the zwittermicin biosynthetic pathway. Interestingly, complementation of a Δ*clbP* strain with ZmaM restored cytotoxicity, suggesting that ZmaM may be able to process substrates that resemble precolibactin.[28] A full-length, untagged construct of ZmaM was cloned from *Bacillus cereus* UW85.

To perform the whole-cell assay, cultures expressing each of the proteins ($ClbP_{pep}$, $ClbP_{pep}$-S95A, ClbP, ClbP-S95A and ZmaM) as well as an empty-vector control were incubated with individual model substrates (**24, 26, 29** and **30**) for eight hours. Methanol extracts of the lyophilized cultures were analyzed by LC-MS for the presence of the *N*-acylated cleavage product. Full-length ClbP processed both D-asparagine-containing substrates **24, 29** and **30**, while $ClbP_{pep}$ did not process either of these substrates in this assay (Figure 2.27). The ClbP-S95A mutant was inactive, indicating that the activity observed in the ClbP-expressing cultures was due to the peptidase activity of ClbP. Compared to **24, 29** and **30,** L-asparagine-containing substrate **26** was hydrolyzed at very low levels in all of the cultures. As the extent of hydrolysis of **26** was similar across all of the cultures, including the ClbP S95A mutant, this reactivity was attributed to endogenous peptidases. The strong preference of ClbP for substrates containing D-asparagine over L-asparagine further confirmed that ClbB elongates *N*-acyl-D-asparagine.

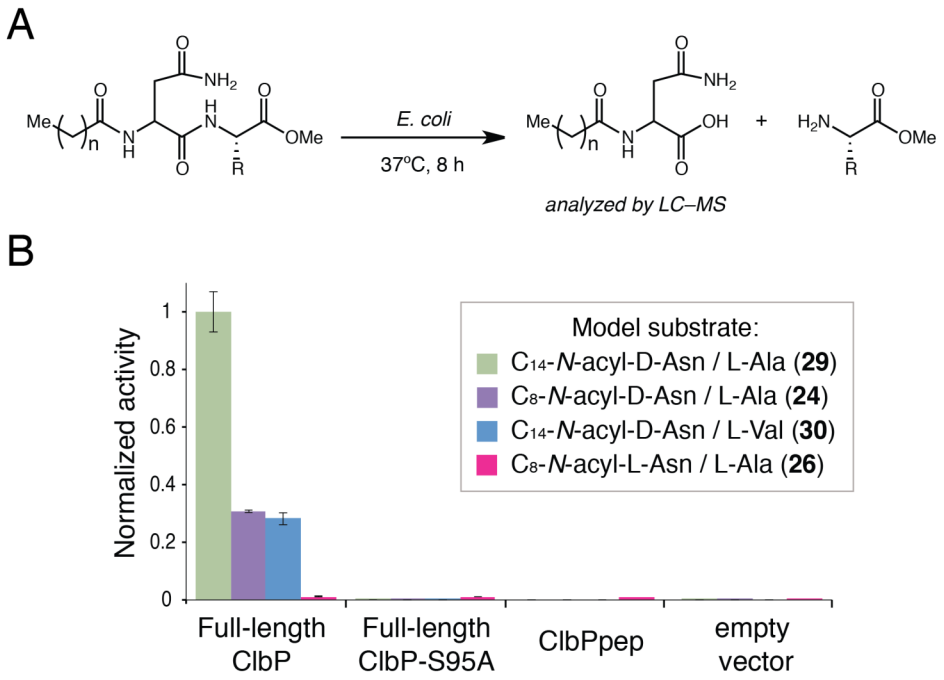


Figure 2.27: A) Scheme for the *in vivo* peptidase assay with model substrates. B) *E. coli* expressing full-length ClbP, but not its active site mutant or the periplasmic construct $ClbP_{pep}$, processed *N*-acyl-D-asparagine-containing substrates. The area under the curve for the extracted ion chromatograms for the *N*-acyl containing prodrug hydrolysis products was normalized relative to the ClbP-expressing strains. Each bar represents the mean ± SEM of three experiments.

The activity of ZmaM was examined with a subset of the model substrates. *E. coli* expressing ZmaM processed D-asparagine-containing substrates **24, 29** and **30** (Figure 2.28). The amount of prodrug hydrolysis products seen in ZmaM-expressing cultures was about ten times higher than in ClbP-expressing cultures. This could be due to more expressed protein or higher rates of catalysis. As the amount of protein was not quantified in these assays, the reasons for the higher activity in the ZmaM-expressing cultures could not be determined from these data. As noted above, ZmaM could restore genotoxicity to *ΔclbP pks*$^+$ *E. coli.* However, *E. coli* strains expressing ZmaM could not cleave prexenocoumacin.[5] Our results indicated that this discrepancy could arise from differences in metabolite structure, perhaps implying that the prodrug peptidases evolved to recognize specific substrates.

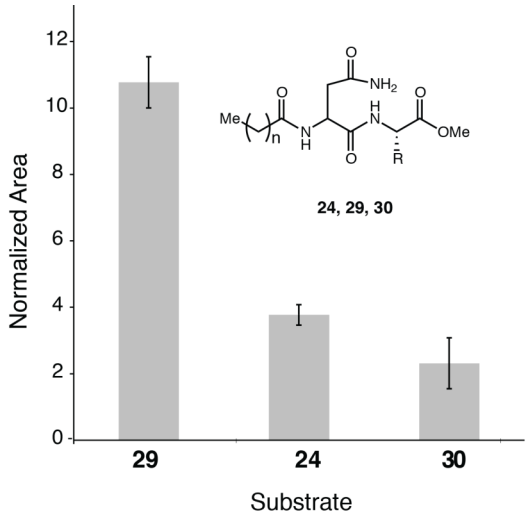
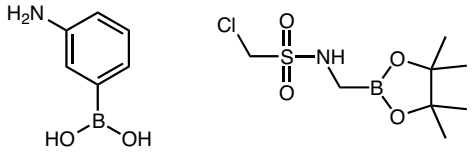


Figure 2.28: *E. coli* expressing full-length ZmaM processed *N*-acyl-D-asparagine containing substrates **29, 24** and **30**. The area under the curve for the extracted ion chromatograms for the *N*-acyl containing prodrug hydrolysis products was normalized relative to that obtained in the ClbP-expressing strain for compound **29**. Each bar represents the mean ± SEM of three independent experiments.

After demonstrating that *E. coli* expressing full-length ClbP could process model precolibactin substrates, ClbP was studied *in vitro*. Specifically, we were interested in characterizing the kinetics of the hydrolysis of the prodrug motif. Using *in silico* docking experiments based on the crystal structure of $ClbP_{pep}$, the Bonnet lab discovered two compounds (**31, 32**) that were able to inhibit the genotoxicity of *pks*$^+$ *E. coli* at relatively high concentrations (1 mM) (Figure 2.29).[30] While the compounds had been tested in a fluorogenic assay with $ClbP_{pep}$, the inhibition of full-length ClbP had not been tested *in vitro* or *in vivo*. In order to obtain purified enzyme for *in vitro* assays, the plasmids encoding full-length ClbP and ClbP-S95A were subjected to site-directed mutagenesis to remove the C-terminal stop codon. Removing the stop codon in the pET-29b inserts resulted in full-length constructs with a C-terminal His_6 tag. The tagged ClbP construct was expressed and purified from the inner membrane. Triton X-100 was used as the solubilizing detergent and standard Ni-affinity chromatography was performed to obtain the purified protein (Figure 2.30).



31 **32**

Figure 2.29: Putative ClbP inhibitors discovered by the Bonnet lab.

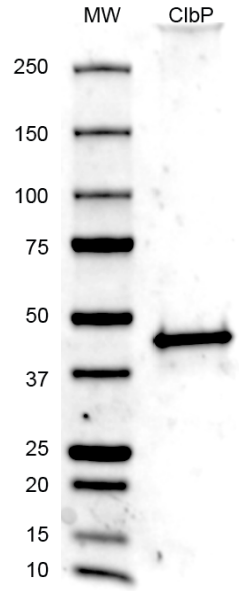


Figure 2.30: SDS-PAGE of purified ClbP-C-His_6. The molecular weight of this construct is 57 kDa. (MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad)).

Initial characterization of the prodrug cleavage reaction with purified ClbP was achieved using the Fmoc-derivatization assay described above. The substrate used for the *in vitro* assays was compound **29**. Preliminary experiments showed that with 100 μM of substrate and 1 μM of enzyme the hydrolysis reaction reached 50% conversion after five minutes and did not progress further at longer incubation times. The mechanism by which the reaction stopped progressing at early time points was not examined. Enzyme inactivation could have occurred due to product inhibition or enzyme denaturation. The kinetics of the reaction were examined using an LC-MS detection assay of the *N*-acyl-D-asparagine hydrolysis product (**10**). In the kinetic assays, 250 nM

enzyme was used and substrate concentration was varied from 0 to 250 μM. These assays revealed that the K_M was 67 μM and the k_{cat} was 24 min^{-1} (Figure 2.31).

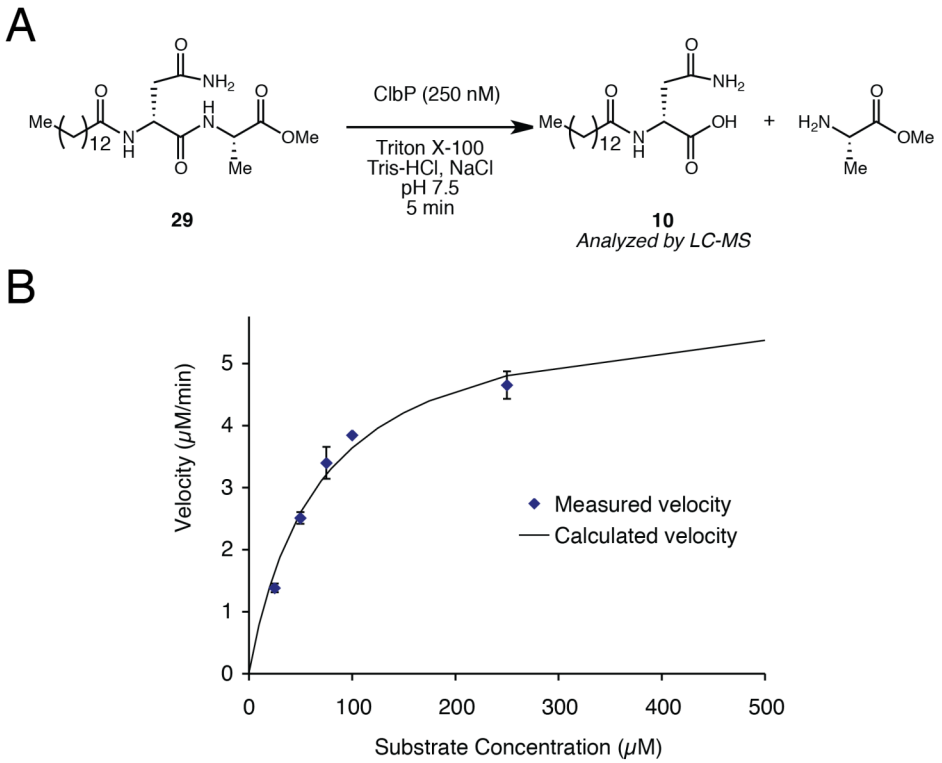


Figure 2.31: A) Reaction scheme for the *in vitro* kinetic analysis of ClbP. B) A plot of the measured reaction velocity versus substrate concentration. Error bars represent the ± SD from three independent experiments.

With the kinetic parameters in hand, we turned to studying the putative ClbP inhibitors discovered by the Bonnet lab. Compounds **31** or **32** were added simultaneously with **29** to ClbP and incubated for five minutes. LC-MS analysis showed that at a concentration of 1 mM neither **31** nor **32** were able to decrease the rate of the reaction compared to the DMSO control (Figure 2.32). These data suggest that there is still a great need to develop a selective inhibitor for ClbP, which could be used as a tool to study the effects of modulating colibactin production in natural gut communities.

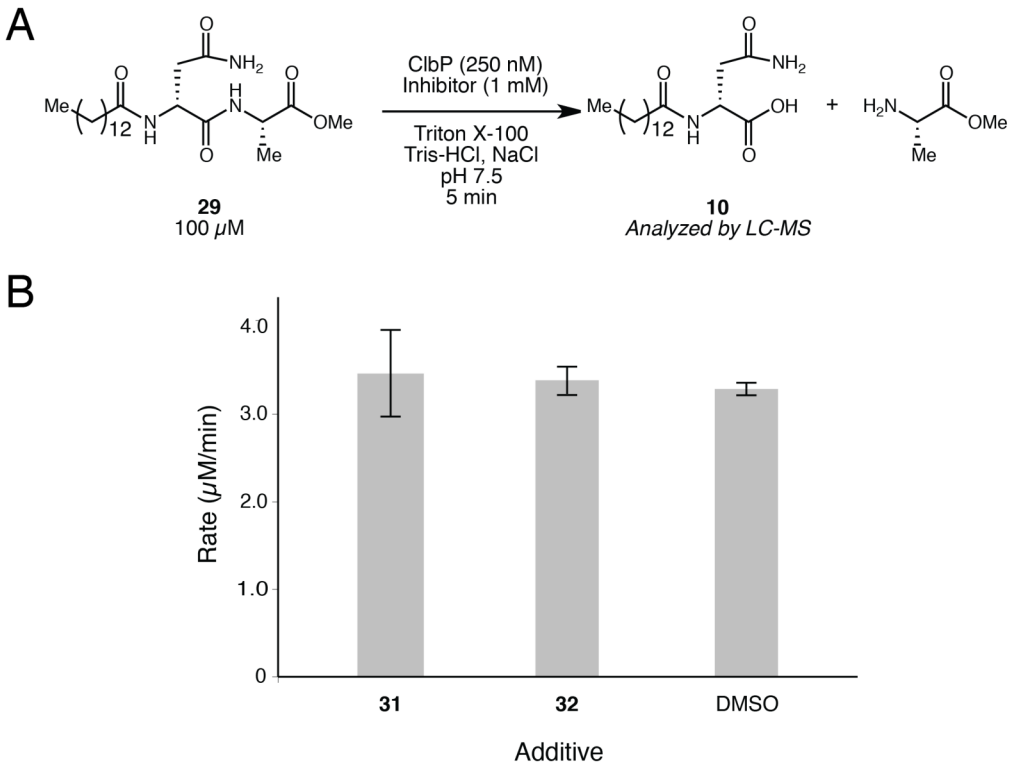


Figure 2.32: A) Reaction scheme for the *in vitro* inhibition assay with the model substrate **29**. B) A bar graph of the measured reaction rate in the inhibitor and DMSO treated reactions. Error bars represent the ± SD from three independent experiments.

2.5: Conclusions

In summary, detailed biochemical characterization of the prodrug enzymatic machinery illuminated the underlying logic employed in the colibactin self-resistance strategy. We uncovered the structure of the N-terminus of precolibactin and elucidated the functions of ClbN, $ClbB_{NRPS}$ and ClbP (Figure 2.33). The prodrug motif, composed of an *N*-myristoyl-D-asparagine residue, is biosynthesized by the initiating NRPS module ClbN and is elongated with both L-alanine and L-valine by the first module of ClbB, $ClbB_{NRPS}$. In addition, *in vivo* and *in vitro* peptidase assays provided evidence that colibactin formation involves amide bond hydrolysis by the periplasmic peptidase ClbP.

Gratifyingly, a study from the Müller group confirmed that the predominate structure biosynthesized by ClbN *in vitro* matches that produced *in vivo* by wild-type *pks*$^+$ bacteria.[31] When metabolite profiles of *E. coli* Nissle 1917 and a *clbP* knock-out strain were compared, a single compound was identified in the wild-type extracts that differentiated the two metabolite profiles. Isolation, purification and characterization of this metabolite revealed an *N*-myristoylated-D-asparagine carboxylic acid (**10**). These results demonstrated unequivocally that *pks*$^+$ *E. coli* produce a prodrug motif that matches the structure predicted from the *in vitro* reconstitution studies.

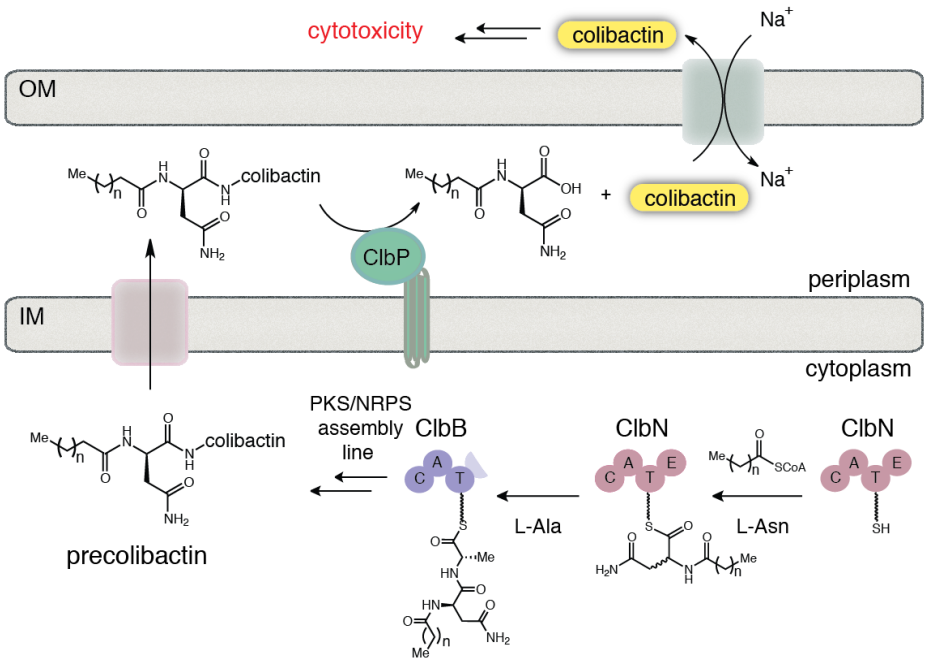


Figure 2.33: Proposed biochemical logic underlying the prodrug resistance mechanism in colibactin biosynthesis. IM = inner membrane, OM = outer membrane.

Importantly, the differential activity of ClbP and ClbP$_{pep}$ toward model substrates *in vivo* parallels their ability to complement Δ*clbP* strains. The difference in the activities of the full-length ClbP and ClbP$_{pep}$ could indicate that the transmembrane helices influence the catalytic activity or that

interactions of both substrate and enzyme with the inner membrane are important. In support of the latter hypothesis, the putative substrate binding pocket of ClbP has a negative electrostatic potential,[27] which suggests that precolibactin is positively charged. Positively charged amino acids are known to increase the binding of myristoylated proteins to cell membranes as a result of strong electrostatic interactions with acidic phospholipids.[32] If precolibactin contains a myristoylated prodrug scaffold attached to a positively charged core, it may be similarly membrane-associated. These results also support the hypothesis that the biochemical logic employed by the colibactin prodrug resistance mechanism is conserved across multiple biosynthetic pathways, as ClbN produced a prodrug scaffold resembling that of the prexenocoumacins and peptidase ZmaM processed model substrates containing the colibactin prodrug scaffold.

2.6: Experimental section

Oligonucleotide primers were synthesized by Integrated DNA Technologies (Coralville, IA). Recombinant plasmid DNA was purified with a Qiaprep Kit from Qiagen. Gel extraction of DNA fragments and restriction endonuclease clean up were performed using an Illustra GFX PCR DNA and Gel Band Purification Kit from GE Healthcare. DNA sequencing was performed by Genewiz (Boston, MA). Nickel-nitrilotriacetic acid-agarose (Ni-NTA) resin was purchased from Qiagen. SDS-PAGE gels were purchased from BioRad. Protein concentrations were determined according to the method of Bradford using bovine serum albumin (BSA) as a standard.[1] Optical densities of *E. coli* cultures were determined with a DU 730 Life Sciences UV/Vis spectrophotometer (Beckman Coulter) by measuring absorbance at 600 nm. Analytical HPLC was performed on a Dionex Ultimate 3000 instrument (Thermo Scientific). Nuclear magnetic resonance (NMR) spectroscopy was performed in the Department of Chemistry and Chemical Biology, Harvard University using Agilent DD2 600 and Varian Utility/Inova 500B spectrometers.

High-resolution mass spectral (HRMS) data for the synthetic compounds were obtained in the Magnetic Resonance Laboratory in Harvard University Department of Chemistry and Chemical Biology on a Bruker Micro QTOF-QII fitted with a dual-spray electrospray ionization (ESI) source. The capillary voltage was set to 4.5 kV and the end plate offset to -500 V, the drying gas temperature was maintained at 190 °C with a flow rate of 8 L/min and a nebulizer pressure of 21.8 psi. The liquid chromatography (LC) was performed using an Agilent Technologies 1100 series LC with 50% H_2O and 50% acetonitrile as solvent. Methanol and water used for LC-ESI-MS were B & J Brand High Purity Solvents (Honeywell Burdick & Jackson).

Chemical shifts are reported in parts per million downfield from tetramethylsilane using the solvent resonance as internal standard for ^{1}H ($CDCl_3$ = 7.26 ppm, DMSO-d_6 = 2.50 ppm) and ^{13}C ($CDCl_3$ = 77.3 ppm, DMSO-d_6 = 39.5 ppm). Data are reported as follows: chemical shift, integration multiplicity (s = singlet, d = doublet, t = triplet, m = multiplet, q=quartet, qt=quintet), coupling constant, integration, and assignment. All chemicals were obtained from Sigma-Aldrich except where noted. Solvents were obtained from Sigma-Aldrich except hexanes (Macron Fine Chemicals), ethyl acetate and isopropanol (VWR), methanol and diethyl ether (EMD Millipore), and ethanol (KOPTEC). All NMR solvents were purchased from Cambridge Isotope Laboratories. NMR spectra were visualized using iNMR Reader, version 4.0.2.

High-resolution LC-MS analyses of enzyme assays were performed in the Saghatelian research labs in the Department of Chemistry and Chemical Biology, Harvard University on an Agilent G3250AA LC/MS TOF Mass Spectrometer fitted with a dual-spray electrospray ionization (ESI) source, or in the Small Molecule Mass Spectrometry Facility at Harvard University on a Bruker Maxis Impact LC-q-TOF Mass Spectrometer. For the Agilent G3250AA LC/MS TOF, the capillary voltage was set to 3.5 kV and the fragmentor voltage to 100 V, and the drying gas

temperature was maintained at 350 °C with a flow rate of 10 L/min and a nebulizer pressure of 45 psi. For the Bruker Maxis Impact LC-q-TOF Mass Spectrometer, the countercurrent drying gas heater was set to 10 L/min and the temperature maintained at 220 °C ; the nebulizer pressure was set to 40 psi; a 20 µL sodium formate plug (10 mM) was introduced at 20 minutes, using a 6-port valve to externally calibrate the m/z scale with sodium formate clusters for each LC-MS run; and the collision energy was kept at 10 eV for efficient ion transmission. For all experiments, liquid chromatography was performed using an Agilent Technologies 1200 series LC using a Phenomenex Gemini C18 reverse phase column (50 x 460 mm). The following elution conditions were used for all experiments: 100% solvent A for 2.1 min, a gradient increasing to 100% solvent B in solvent A over 1.9 min, 100% solvent B for 10 min, a gradient decreasing to 0% solvent B in solvent A over 0.1 min, 100% A for 3.9 min (solvent A = 95:5 water/methanol; solvent B = 80:15:5 isopropanol/methanol/water).

Cloning, overexpression, and purification of ClbN, $ClbB_{NRPS}$, ClbP, $ClbP_{pep}$, ClbP-S95A, $ClbP_{pep}$-S95A, and ZmaM

Table 2.1. Oligonucleotides used for cloning. Restriction sites are underlined.

Primer Name	Target	Sequence (5' to 3')
ClbN-F	ClbN	CAGATCGAATTCATGATGTCGGGCAATCCG
ClbN-R	ClbN	ATTAGCGGCCGCTCATAGTGTCCACAAAGTC
ClbB-F	ClbB	TTAATTGCGGCCGCATGGATAATACCTCTG
ClbB-R	ClbB	AGTCATCTCGAGTCACGTTGGAATTGCATC
$ClbP_{pep}$-F	ClbP periplasmic domain	AATACATATGACAATAATGGAACACGTTAG
$ClbP_{pep}$-R	ClbP periplasmic domain	ATTACTCGAGATATTTGCCAATGCGCAG
ClbP-F	ClbP full length	CGAGCGCATATGACAATAATGGAACACGTTAGCATTAAAAC
ClbP-R	ClbP full length	TATTCTCGAGTTACTCATCGTCCCACTCCTTGTTG
ZmaM-F	ZmaM	AATTCCATGGGCAAGTTAAACATATGGTTG
ZmaM-R	ZmaM	AATTCTCGAGTCAAGTCAAAAGCGACTTTC

ClbN, *clbB*, *clbP*$_{pep}$ and *clbP* were PCR amplified from *E. coli* CFT073 genomic DNA (purchased from the American Type Culture Collection, Manassas, VA) using the primers shown in Table 2.1. *ClbP* is the full length sequence of the peptidase, whereas *clbP*$_{pep}$ encodes for the first 375 amino acids of the peptidase, which contains the soluble periplasmic peptidase domain.[2] *ZmaM* was PCR amplified from *Bacillus cereus* UW85 genomic DNA (*B. cereus* UW85 was obtained from the Handelsman lab, Yale University) using the primers shown in Table 2.1. *B. cereus* UW85 genomic DNA was isolated using the Ultra Clean Microbial DNA Isolation Kit from MoBio Laboratories (Carlsbad, CA). *ZmaM* encodes for the first 501 amino acids of the peptidase, which contains the soluble periplasmic peptidase domain and the first six transmembrane helices. *ClbN* was amplified using forward primer **ClbN-F** + reverse primer **ClbN-R**, *clbB* was amplified using forward primer **clbB-F** + reverse primer **cylB-R**, *clbP*$_{pep}$ was amplified using forward primer **ClbP$_{pep}$-F** + reverse primer **ClbP$_{pep}$-F**, *clbP* was amplified using forward primer **ClbP-F** and **ClbP-R**, and *ZmaM* was amplified using forward primer **ZmaM-F** + reverse primer **ZmaM-R**. All PCR reactions contained 25 µL of Phusion High-Fidelity PCR Master Mix (New England Biolabs), 2 ng of DNA template, and 500 pmoles of each primer in a total volume of 50 µL. Thermocycling was carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation for 1 min at 95 °C , followed by 50 cycles of 30 sec at 95 °C , 1 min at the annealing temperature, 5 min at 72 °C , and a final extension time of 10 min at 72 °C . The annealing temperatures used were as follows: *clbN*: 72 °C , *clbB*: 72 °C , *clbP*$_{pep}$: 56.9 °C , *clbP*: 65.7 °C , *zmaM*: 53.1 °C .

PCR reactions were analyzed by agarose gel electrophoresis with ethidium bromide staining, pooled, and purified. Amplified fragments were digested with the appropriate restriction enzymes (New England Biolabs) for 2.5 h at 37 °C . Digests contained 2 µL of water, 6 µL of NEB Buffer 4 (10x), 6 µL of BSA (10x), 40 µL of PCR product, and 3 µL of each restriction enzyme (20,000 U/µL). Restriction digests were purified directly using agarose gel electrophoresis. Gel

fragments were further purified using the Illustra GFX kit. The digests were ligated into linearized expression vectors using T4 DNA ligase (New England Biolabs). *ClbB* and *clbN* were ligated into the pET-28a vector to encode N-terminal His_6-tagged constructs, $clbP_{pep}$ was ligated into the pET-29b vector to encode a C-terminal His_6-tagged construct, *clbP* was ligated into the pET-29b vector to encode an untagged construct, and *zmaM* was ligated into the pET-28a vector to encode an untagged construct. Ligations were incubated at room temperature for 2 h and contained 3 µL of water, 1 µL of T4 Ligase Buffer (10x), 1 µL of digested vector, 3 µL of digested insert DNA, and 2 µL of T4 DNA Ligase (400 U/ µL). 5 µL of each ligation was used to transform a single tube of chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* BL21 (DE3) cells (Invitrogen) and stored at –80 °C as frozen LB/glycerol stocks.

ClbP and $ClbP_{pep}$ site-directed mutagenesis:

Site-directed mutagenesis of ClbP and $ClbP_{pep}$ was performed using 25 µL of Phusion High-Fidelity PCR Master Mix (New England Biolabs), 50 ng of pET-29b-ClbP or pET-29b-$ClbP_{pep}$ template, and 500 pmoles of each primer in a total volume of 50 µL. Thermocycling was carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters for $ClbP_{pep}$-pET-29b: denaturation for 1 min at 95 °C , followed by 18 cycles of 30 sec at 95 °C , 1 min at 62 °C , and 6.5 min at 72 °C . The following parameters were used for ClbP-S95A: denaturation for 1 min at 95 °C , followed by 18 cycles of 30 sec at 95 °C , 1 min at 57.8 °C , and 7.5 min at 72 °C . The ClbP-S95A and $ClbP_{pep}$-S95A mutants were made using primers 5'-GTTTACGAGCTGGGAGCCATGAGTAAGG-3' and 5'-CCTTACTCATGGCTCCCAGCTCGTAAAC-3', with the S95A nucleotides underlined. Digestion of the PCR products was performed by the addition of 1 µL of DpnI (20,000 U/mL, New

England Biolabs) per 50 µL PCR reaction and incubation at 37 °C for 1 h followed by the addition of another 1 µL of DpnI, followed by incubation at 37 °C for 1 h. 2 µL of each digestion reaction were used to transform a single tube of chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* BL21 (DE3) cells (Invitrogen) and stored at –80 °C as frozen LB/glycerol stocks.

ClbP and ClbP-S95A site-directed mutagenesis:

In order to obtain C-His_6 tagged constructs of ClbP and ClbP-S95A, we performed site-directed mutagenesis to remove the stop codon that was present in the insert of these expression vectors. Site-directed mutagenesis was performed using 25 µL of Phusion High-Fidelity PCR Master Mix (New England Biolabs), 50 ng of pET-29b-ClbP or pET-29b-ClbP-S95A template, and 500 pmoles of each primer in a total volume of 50 µL. Thermocycling was carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation for 30 sec at 98 °C , followed by 18 cycles of 30 sec at 98 °C , 1 min at 64 or 65 °C , and 7 min at 72 °C . The primers used were 5'-GTGGGACGATGAGCTCGAGCACCACCAC-3' and 5'-GTGGTGGTGCTCGA GCTCATCGTCCCAC-3'. Digestion of the PCR products was performed by the addition of 1 µL of DpnI (20,000 U/mL, New England Biolabs) per 50 µL PCR reaction and incubation at 37 °C for 1 h followed by the addition of another 1 µL of DpnI, followed by incubation at 37 °C for 1 h. 2 µL of each digestion reaction were used to transform a single tube of chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* BL21 (DE3) cells (Invitrogen) and stored at –80 °C as frozen LB/glycerol stocks.

Large scale overexpression and purification of ClbN and ClbB:

A 50 mL starter culture of pET-28a-ClbN or pET-28a-ClbB BL21 *E. coli* was inoculated from a frozen stock and grown overnight at 37 °C in LB medium supplemented with 50 μg/ml kanamycin. Overnight cultures were diluted 1:100 into 2 L of LB medium containing 50 μg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm, moved to 15 °C at OD_{600} = 0.2-0.3, induced with 500 μM IPTG at OD_{600} = 0.5-0.6, and incubated at 15 °C for ~ 16 h. Cells from 2 L of culture were harvested by centrifugation (6,000 rpm x 10 min) and resuspended in 80 mL of lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, pH 8). The cells were lysed by passage through a cell disruptor (Avestin EmulsiFlex-C3) twice at 10,000 psi, and the lysate was clarified by centrifugation (13,000 rpm x 30 min). The supernatant was incubated with 2 mL of Ni-NTA resin and 5 mM imidazole for 2 h at 4 °C . The mixture was centrifuged (3,000 rpm x 5 min) and the unbound fraction discarded. The Ni-NTA was resuspended in 10 mL of elution buffer (20 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, 5 mM imidazole, pH 8), loaded into a glass column, and washed with 10 mL of elution buffer. Protein was eluted from the column using a stepwise imidazole gradient in elution buffer (25 mM, 50 mM, 75 mM, 100 mM, 125 mM, 150 mM, 200 mM), collecting 2 mL fractions. SDS–PAGE analysis (4–15% Tris-HCl gel) was employed to ascertain the presence and purity of protein in each fraction. Fractions containing the desired protein were combined and dialyzed twice against 2 L of storage buffer (20 mM Tris-HCl, 50 mM NaCl, 10% (v/v) glycerol, pH 8). Solutions containing protein were frozen in liquid N_2 and stored at –80 °C . This procedure afforded yields of 8.3 mg/L for N-His_6-tagged ClbN, and 2.52 mg/L for N-His_6-tagged ClbB.

Large scale overexpression and purification of $ClbP_{pep}$ and $ClbP_{pep}$-S95A:

A 50 mL a starter culture of pET-29b-$ClbP_{pep}$ or pET-29b-$ClbP_{pep}$-S95A BL21 *E. coli* was inoculated from frozen stock and grown overnight at 37 °C in LB medium supplemented with 50

μg/ml kanamycin. Overnight cultures were diluted 1:100 into 2 L LB medium containing 50 μg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm, moved to 15 °C at OD_{600} = 0.2-0.3, induced with 500 μM IPTG at OD_{600} = 0.5-0.6, and incubated at 15 °C for ~ 16 h. Cells from the 2 L culture were harvested by centrifugation (4,000 rpm x 20 min) and resuspended in 400 mL (80 mL/g) Tris/Sucrose buffer (30 mM Tris-HCl pH 8 20% wt sucrose). To this mixture was added EDTA (1 mM) dropwise with stirring. This mixture was allowed to stir at 4 °C for 10 min and was then centrifuged (8,000 rpm x 20 min). The supernatant was removed, and the cell pellet was resuspended in 400 mL of 5 mM $MgSO_4$. The suspension was allowed to stir at 4 °C for 10 min, and was then centrifuged (8,000 rpm x 20 min). The supernatant was incubated with 3 mL of Ni-NTA resin and 5 mM imidazole for 2 h at 4 °C . This mixture was then poured onto a glass column and the solution eluted to give ~ 2 mL of Ni-NTA resin in a 5 mM imidazole solution. Protein was eluted from the column using a stepwise imidazole gradient in elution buffer (25 mM, 50 mM, 75 mM, 100 mM, 125 mM, 150 mM, 200 mM), collecting 2 mL fractions. SDS–PAGE analysis (4–15% Tris-HCl gel) was employed to ascertain the presence and purity of protein in each fraction. Fractions containing the desired protein were combined and dialyzed twice against 2 L of storage buffer (25 mM Tris-HCl, 50 mM NaCl, 10% (v/v) glycerol, pH 8.7). This procedure afforded yields of 0.27 mg/L for C-His_6-tagged $ClbP_{pep}$, and 0.21 mg/L for C-His_6-tagged $ClbP_{pep}$-S95A.

Large scale overexpression of ClbP and ClbP-S95A

A 50 mL starter culture of pET-29b-ClbP or pET-29b-ClbP-S95A BL21 *E. coli* was inoculated from a frozen stock and grown overnight at 37 °C in LB medium supplemented with 50 μg/ml kanamycin. Overnight cultures were diluted 1:100 into 1 L of LB medium containing 50 μg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm, moved to 28 °C at OD_{600} = 0.2-0.3, induced with 500 μM IPTG at OD_{600} = 0.5-0.6, and incubated at 28 °C for ~ 16 h. Cells

from 1 L of culture were harvested by centrifugation (6,000 rpm x 10 min) and resuspended in 30 mL of Buffer A (50 mM Tris HCl pH 7.0, 10 mM $MgCl_2$, 0.1 M NaCl, 10 μg/mL DNAseI). The cells were lysed by passage through a cell disruptor (Avestin EmulsiFlex-C3) twice at 10,000 psi, and the lysate was centrifuged (2000 g x 10 min; 3420 rpm) to remove unlysed cells. The volume of the cell lysate was adjusted 50 mL with Buffer A. The lysate was then ultracentrifuged (34 krpm x 1 h; 45Ti rotor), and the supernatant was removed. The cell pellet was then resuspended in 50 mL Buffer A, ultracentrifuged (34 krpm x 1 h; 45Ti rotor) and the supernatant was removed. The cell pellet was then resuspended in 50 mL Buffer B (50 mM Tris HCl pH 7.0, 0.1 M NaCl, 2% (v/v) Triton X-100) and was incubated with gentle mixing at 4 °C for 30 min. The sample was then ultracentrifuged (34 krpm x 1 h; 45Ti rotor). The supernatant was incubated with 1.5 mL Ni-NTA resin and 25 mM imidazole for 2 h and then loaded into a glass column. Protein was eluted from the column using a stepwise imidazole gradient (25 mM, 50 mM, 75 mM, 100 mM, 125 mM, 150 mM, 200 mM) in Buffer D (50 mM Tris HCl pH 7.0, 0.2 M NaCl, 1% (v/v) Triton X-100), collecting 2 mL fractions. SDS–PAGE analysis (4–15% Tris-HCl gel) was employed to ascertain the presence and purity of protein in each fraction. Fractions containing the desired protein were combined and dialyzed twice against 2 L of Buffer D. Solutions containing protein were frozen in liquid N_2 and stored at –80 °C . This procedure afforded yields of 1.3 mg/L for C-His_6-tagged ClbP and 0.7 mg/L for C-His_6-tagged ClbP-S95A.

In vitro assays for the biochemical characterization of ClbN

BODIPY-CoA fluorescent phosphopantetheinylation assay

BODIPY-CoA[3] and Sfp[4] were prepared using previously reported procedures. The reaction mixture (50 μL) contained 5 μM ClbN, 1.0 μM Sfp (prepared as a 4.8 μM solution in 10% (v/v) glycerol, 1 mg/mL BSA, 10 mM Tris-HCl pH 8), 5 μM BODIPY-CoA, 10 mM $MgCl_2$, and 50 mM

HEPES pH 7.5. Reaction mixtures were incubated in the dark at room temperature for 1 h and then diluted 1:1 in 2x Laemmli sample buffer (Bio-Rad), boiled for 10 min, and separated by SDS-PAGE (4-15% Tris-HCl gel). The gel was first imaged at λ=365 nm, then stained with Coomassie and imaged again.

ATP-[^{32}P]PP_i exchange assay

The reaction mixture (100 µL) contained 75 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 5 mM DTT, 5 mM ATP, 1 mM amino acid substrate, and 4 mM Na_4PP_i/[^{32}P]PP_i (~ 10^6 cpm/mL) (American Radiolabeled Chemicals). Reactions were initiated by the addition of ClbN (2 µM), and incubated at room temperature for 30 min. Reactions were quenched by the addition of 200 µL of charcoal suspension (16 g/L activated charcoal, 100 mM Na_4PP_i, 3.5 % (v/v) $HClO_4$). The samples were centrifuged (13,000 rpm x 3 min), and the supernatant was removed. The charcoal pellet was washed two times with 200 µL of wash buffer (100 mM Na_4PP_i, 3.5 % (v/v) $HClO_4$). The pellet was resuspended in 200 µL of wash buffer and added to 10 mL of scintillation fluid (Ultima Gold, Perkin Elmer). Radioactivity was measured on a Beckman LS 6500 scintillation counter.

ATP-[^{32}P]PP_i kinetics assay

Kinetic analysis of the adenylation domain of ClbN was conducted using the ATP-[^{32}P]PP_i assay as described above, with the following differences: the concentration of ClbN was 2.5 µM and the incubation time was 15 min. The concentration of L-Asn was varied from 10 µM to 2.5 mM. Kinetic parameters were calculated from three replicates at each substrate concentration (0 µM, 10 µM, 1 mM, 2.5 mM) using GraphPad.

T domain loading assay with ^{14}C-L-Asn

The reaction mixture (50 µL) contained 50 mM HEPES pH 7.5, 40 mM NaCl, 5 mM $MgCl_2$, 250 µM CoA tri-lithium salt, 500 µM DTT, 5% (v/v) DMSO, 200 nM Sfp, 7.5 µM ^{12}C-L-Asn, and 8 µM ^{14}C-L-Asn (American Radiochemicals, 0.1 mCi/mL, 208 mCi/mmol). Loading of the phosphopantetheinyl arm onto the T domain of *apo* ClbN was initiated by the addition of ClbN (5 µM) to the reaction mixture, followed by incubation at room temperature for 30 min. Loading of the T domain with amino acid was initiated by the addition of ATP (3 mM). After incubation at room temperature for 45 min, the reaction was quenched by the addition of 100 µL of BSA (1 mg/mL) followed by 500 µL of trichloroacetic acid (TCA) (10% (m/v) aqueous solution). The protein was pelleted by centrifugation (10,000 rpm x 8 min). After removal of the supernatant, the protein pellet was washed two times with 250 µL of TCA (10% wt aqueous solution). The pellet was resuspended in 200 µL of formic acid and added to 10 mL of scintillation fluid (Ultima Gold, Perkin Elmer). Radioactivity was measured on a Beckman LS 6500 scintillation counter.

LC-MS assay for C domain substrate specificity

The reaction mixture (300 µL) contained 40 mM HEPES buffer pH 7.5, 33 mM NaCl, 4 mM $MgCl_2$, 400 µM DTT, 4 mM L-Asn, 126 µM CoA-tri-lithium salt, 250 nM Sfp, and 6.6 % (v/v) DMSO. Loading of the phosphopantetheinyl arm onto the T domain of *apo* ClbN was initiated by the addition of ClbN (5 µM) to the reaction mixture, followed by incubation at room temperature for 30 min. ATP (5 mM) was then added to the reaction mixture, and the C domain loading reaction was initiated by the addition of the fatty acyl-CoA substrate (830 µM). This mixture was incubated at room temperature for 2 h and quenched by the addition of methanol (750 µL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The protein pellets were washed two times with 250 µL of methanol and dried under a stream of N_2. Products bound to the T domain were hydrolyzed by the addition of 0.1 M KOH (25 µL) followed by heating at 74 °C for

10 min. The samples were cooled on ice, and 0.1 M HCl (75 µL) was added to the solutions. Finally, methanol (200 µL) was added to the samples, which were then incubated at –20 °C overnight to precipitate protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS. The fatty acyl-CoA competition experiment was run using identical conditions, except that a stock solution containing all of the fatty acyl-CoA substrates was added instead of an individual substrate. Each fatty acyl-CoA in the mixture was added to a final concentration of 187 µM. The expected masses of the condensation reactions were found neither in the control reactions that did not contain ClbN nor in the reactions that contained boiled ClbN.

In vivo LC-MS assay for C domain substrate specificity

Starter cultures of BAP1[5] *E. coli* (50 mL) harboring pET-28a-ClbN-N-His_6 or empty pET28a vector were inoculated from a frozen cell stock and grown overnight at 37 °C in LB medium supplemented with 50 µg/mL kanamycin. The saturated cultures (10 mL) were used to inoculate 1 L LB medium containing 50 µg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm. At an OD_{600} of 0.3 the cultures were moved to 15 °C . At an OD_{600} of 0.6 the cultures were induced with 500 µM IPTG. The cultures were incubated at 15 °C for 18 h. The cells were harvested by centrifugation (6,000 rpm x 10 min) and resuspended in 40 mL lysis buffer (20 mM Tris-HCl, 300 mM NaCl, pH 7.98). The cells were lysed by passage through a cell disruptor (Avestin EmulsiFlex-C3) twice at 10,000 psi and the lysate was clarified by centrifugation (13,000 rpm x 30 min). A 20 mL aliquot of the soluble portion of the cell lysate was concentrated in a 30 kDa MWCO spin filter (Corning) to 12-15 mL. This concentrated aliquot was flash frozen and lyophilized. To the lyophilized powder was added 0.1 M KOH (2 mL). The mixtures were vortexed vigorously, and heated in a water bath at 74 °C for 10 min. The samples were cooled on ice, and 0.1 M HCl (3 mL) was added to the solutions, which were then flash frozen and

lyophilized. To the lyophilized powder was added 3 mL methanol. The samples were vortexed for 30 sec, and then centrifuged (4,000 rpm x 8 min). A portion of the supernatant (1 mL) was transferred to a microcentrifuge tube. The samples were incubated at –20 °C for 24 h. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS. None of the expected masses were found in three independent control reactions from BAP1 *E. coli* harboring an empty pET-28a vector or in three independent control reactions from BAP1 *E. coli* harboring pET-28a-ClbN-N-His_6 to which no IPTG was added during cell culture.

LC-MS assay for E domain activity

The reaction mixture (100 µL) contained 75 mM Tris-HCl pH 7.5, 33 mM NaCl, 4 mM $MgCl_2$, 400 µM DTT, 2 mM L-Asn, 126 µM CoA tri-lithium salt, 250 nM Sfp, and 5% (v/v) DMSO. All of these reagents, with the exception of Sfp and CoA, were prepared in D_2O (Alfa Aesar). For the control reaction that took place in H_2O, these reagents were prepared in MQ water. Loading of the phosphopantetheinyl arm onto the T domain of *apo* ClbN was initiated by the addition of ClbN (3 µM) to the reaction mixture. This mixture was incubated at room temperature for 30 min, at which point ATP (2 mM), prepared in either D_2O or MQ water, was added. The reaction was initiated by the addition of fatty acyl-CoA (750 µM, either octanoyl-CoA or myristoyl-CoA). This mixture was incubated at room temperature for 2 h and then the reaction was quenched by the addition of methanol (750 µL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The protein pellets were washed two times with methanol (250 µL) and dried under a stream of N_2. Products bound to the T domain were hydrolyzed by adding 0.1 M KOH (10 µL), prepared in MQ water, and heating at 74 °C for 10 min. The samples were cooled on ice, and 0.1 M HCl (30 µL), prepared in MQ water, was added to the solutions. Finally, methanol (80 µL) was added to the samples, which were then incubated at –20 °C overnight to precipitate

protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS.

In vitro assays for the biochemical characterization of $ClbB_{NRPS}$

BODIPY-CoA fluorescent phosphopantetheinylation assay

The reaction mixture (50 µL) contained 5 µM ClbB, 1.0 µM Sfp, 5 µM BODIPY-CoA, 10 mM $MgCl_2$, and 50 mM HEPES pH 7.5. Reactions were incubated in the dark at room temperature for 1 h. Reaction mixtures were diluted 1:1 in 2x Laemmli sample buffer (Bio-Rad), boiled for 10 min, and then separated by SDS-PAGE (4-15% Tris-HCl gel). The gel was first imaged at λ=365 nm, then stained with Coomassie and imaged again.

ATP-$[^{32}P]PP_i$ exchange assay

A typical reaction (100 µL) contained 75 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 5 mM DTT, 5 mM ATP, 1 mM amino acid substrate, and 4 mM $NaPP_i$/$[^{32}P]PP_i$ (~ 10^6 cpm/mL) (American Radiolabeled Chemicals). Reactions were initiated by the addition of ClbB (2 µM), and incubated at room temperature for 30 min. Reactions were quenched by the addition of 200 µL of charcoal suspension (16 g/L activated charcoal, 100 mM Na_4PP_i, 3.5 % (v/v) $HClO_4$). The samples were centrifuged (13,000 rpm x 3 min), and the supernatant was removed. The charcoal pellet was washed two times with 200 µL of wash buffer (100 mM Na_4PP_i, 3.5 % (v/v) $HClO_4$). The pellet was resuspended in 200 µL of wash buffer and added to 10 mL of scintillation fluid (Ultima Gold, Perkin Elmer). Radioactivity was measured on a Beckman LS 6500 scintillation counter.

T domain loading assay with ^{14}C-L-Ala and ^{14}C-L-Val

The reaction mixture (50 µL) contained 50 mM HEPES pH 7.5, 40 mM NaCl, 5 mM $MgCl_2$, 250 µM CoA tri-lithium salt, 500 µM DTT, 5% (v/v) DMSO, 200 nM Sfp, 7.5 µM ^{12}C-L-Asn, and either 8 µM ^{14}C-L-Ala (Moravek, 100 µCi/mL, 132 mCi/mmol) or 8 µM ^{14}C-L-Val (Moravek, 100 µCi/mL, 246 mCi/mmol). Loading of the phosphopantetheinyl arm onto the T domain of *apo* ClbB was initiated by the addition of ClbB (5 µM) to the reaction mixture, followed by incubation at room temperature for 30 min. Loading of the T domain with amino acid was initiated by the addition of ATP (3 mM). After incubation at room temperature for 45 min, the reaction was quenched by the addition of 100 µL of BSA (1 mg/mL) followed by 500 µL of TCA (10% (m/v) aqueous solution). The protein was pelleted by centrifugation (10,000 rpm x 8 min). After removal of the supernatant, the protein pellet was washed two times with 250 µL of TCA (10% m/v aqueous solution). The pellet was resuspended in 200 µL of formic acid and added to 10 mL of scintillation fluid (Ultima Gold, Perkin Elmer). Radioactivity was measured on a Beckman LS 6500 scintillation counter.

ClbN and ClbB reconstitution assay

The reaction mixture (300 µL) contained 40 mM Tris-HCl pH 7.5, 33 mM NaCl, 4 mM $MgCl_2$, 400 µM DTT, 1 mM L-Ala or 1 mM L-Val, 1 mM L-Asn, 126 µM CoA tri-lithium salt, 250 nM Sfp, and 1.67 % (v/v) DMSO. Loading of phospantetheinyl arms onto the T domains of *apo* ClbB and *apo* ClbN was initiated by the addition of ClbB (3 µM) and ClbN (3 µM) to the reaction mixture. This mixture was incubated at room temperature for 30 min, at which point ATP (4 mM) was added. The reaction was initiated by the addition of the fatty acyl-CoA substrate (750 µM). This mixture was incubated at room temperature for 2 h. The reaction was quenched by the addition of methanol (750 µL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The protein pellets were washed two times with methanol (250 µL) and dried under a

stream of N_2. Products bound to the T domains of both ClbB and ClbN were hydrolyzed by adding 0.1 M KOH (25 µL) and heating at 74 °C for 10 min. The samples were cooled on ice, and 0.1 M HCl (75 µL) was added to the solutions. Finally, methanol (200 µL) was added to the samples, which were then incubated at –20 °C overnight to precipitate protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS. The fatty acyl-CoA competition experiment was run using identical conditions as described above, except that the concentration of both ClbN and ClbB was raised to 15 µM, and a 1:1 mixture of octanoyl-CoA and myristoyl-CoA was added instead of an individual fatty acyl-CoA substrate. Octanoyl-CoA and myristoyl-CoA were each at a final concentration of 375 µM in the reaction mixture.

In vitro assays for the biochemical characterization of ClbP, $ClbP_{pep}$, ClbP-S95A, $ClbP_{pep}$-S95A, and ZmaM

Fluorogenic *p*-nitroaniline release assay

This assay was conducted as described previously.[28] The reaction mixture (50 µL) contained 100 mM Tris-HCl pH 7.5, 74 mM NaCl, and h1 µM ClbPpep or ClbPpep-S95A. The reaction was initiated by the addition of either **2** or **3** (Bachem). The reaction was monitored at 25 °C at 405 nm with reads every 10 s for a total of 30 min.

In vitro HPLC assay

The reaction mixture (50 µL) contained 50 mM Tris-HCl pH 7.5, 37.5 mM NaCl, and 500 µM substrate (added as a 5 mM solution in DMSO). The reaction was initiated by the addition of $ClbP_{pep}$ or $ClbP_{pep}$-S95A (1 µM). This mixture was incubated at room temperature for 24 h. The reaction was quenched by the addition of methanol (100 µL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The supernatant was removed, and to this was added sodium borate (40 µL of a 0.5 M aqueous solution, pH 8.0), followed by

fluorenylmethyloxycarbonyl chloride (20 µL of a 10 mM solution in 1,4-dioxane). After incubation at room temperature for 2 min, 1-aminoadamantane (40 µL of a 100 mM solution in DMSO) was added. After incubation at room temperature for 1 min, the samples were centrifuged (13,000 rpm x 1 min) and the supernatant was analyzed by HPLC. HPLC was performed using a Higgins analytical Clipeus C18 reverse phase column (5µm, 4.6 x 250 mm). The following elution conditions were used for all experiments except those with **25**: 90% solvent A for 5 min, a gradient increasing to 80% solvent B in solvent A over 1 min, 80% solvent B for 4 min, a gradient increasing to 100% solvent B in solvent A over 1 min, 100% B for 10 min, and a gradient decreasing to 10% solvent B in solvent A over 1 min. (solvent A = water with 0.1% (v/v) trifluoroacetic acid; solvent B = acetonitrile with 0.1% (v/v) trifluoroacetic acid). The following elution conditions were used for all experiments with **25**: 90% solvent A in solvent B for 5 min, a gradient increasing to 40% solvent B in solvent A over 3 min, 40% solvent B for 4 min, a gradient increasing to 80% solvent B in solvent A over 3 min, 80% solvent B for 3 min, and a gradient increasing to 100% solvent B over 7 min, 100% solvent B for 9 min, and a gradient decreasing to 10% solvent B in solvent A over 2 min. (solvent A = water with 0.1% (v/v) trifluoroacetic acid; solvent B = acetonitrile with 0.1% (v/v) trifluoroacetic acid).

In vivo LC-MS assay

Starter cultures of *E. coli* BL21(DE3) (5 mL) harboring pET-29b-ClbP, pET-29b-ClbP$_{pep}$, pET-29b-ClbP-S95A, pET-28a-ZmaM, and empty pET-29b vector were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 50 µg/ml kanamycin. These saturated cultures (100 µL) were used to inoculate 10 mL of LB medium containing 50 µg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm. At an OD_{600} of 0.5-0.6, substrate or DMSO was added to a final concentration of 100 µM. All substrates were prepared as either 50 mM or 25 mM stock solutions in DMSO. The cultures were then induced

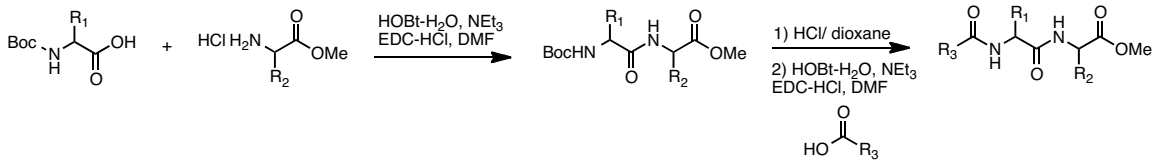
with 500 μM IPTG and incubated at 37 °C. 1 mL aliquots of the cultures were removed at 8 h. The aliquots were flash frozen in liquid N_2 and lyophilized. The lyophilized powder was extracted into 200 μL methanol by vortexing the mixture for twenty seconds and then heating at 40 °C for 5 min. The samples were then centrifuged (13,000 rpm x 15 min), and the supernatant was carefully transferred to a clean vial. These samples were stored at –20 °C overnight. Upon warming, these samples were centrifuged again (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS.

In vitro LC-MS assay

The reaction mixture (60 μL) contained 100 mM Tris-HCl pH 7.5, 75 mM NaCl, 0.8% Triton X-100, 100 μM compound **29** (added as a 25 mM solution in DMSO), and 1 mM compound **31** or **32** (added as a 30 mM solution in DMSO). The reaction was initiated by the addition of 250 nM ClbP. The mixture was incubated at room temperature for 5 min. The reaction was quenched by the addition of methanol (120 μL). LC-MS samples were prepared by a ten-fold dilution of 30 μL of the quenched reaction into 270 μL methanol.

Chemical synthesis procedures and characterization data

General procedure for preparation of *N*-acylated dipeptide peptidase substrates:



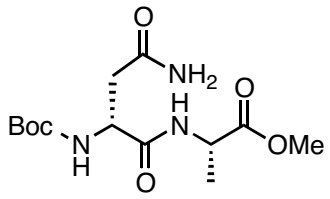
General procedure A: Boc-protected amino acid (Advanced Chem Tech) (1.2 equiv), the HCl-salt of the methyl ester-protected amino acid (1.0 equiv), and hydroxybenzotriazole (HOBt) monohydrate (TCI) (1.1 equiv) were dissolved in anhydrous *N,N*-dimethylformamide (DMF) (0.2 M). Triethylamine (2.2 equiv) was added via syringe, and the reaction mixture was stirred at

room temperature for 5 min. The reaction mixture was then cooled to 0 °C , and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) hydrochloride (Advanced Chem Tech) (1.1 equiv) was added. The reaction mixture was stirred at 0 °C for 15 min and was then warmed to room temperature and stirred overnight. The reaction mixture was diluted with water (1 x reaction volume) and ethyl acetate (5 x reaction volume), and the resulting organic layer was washed with a saturated aqueous NH_4Cl solution (5 x reaction volume). The aqueous layer was then extracted one more time with an equal portion of ethyl acetate. The organic layers were combined and washed with water, saturated aqueous $NaHCO_3$, water, and brine (10 x reaction volume of each). The organic layers were dried over $MgSO_4$, filtered, and concentrated *in vacuo.* The Boc-protected dipeptide was carried onto the next step without further purification unless otherwise noted.

General procedure B: To a solution of Boc-protected dipeptide in THF (0.2 M) under an atmosphere of argon was added 4 M HCl in 1,4-dioxane (0.4 M, 9.77 equiv) dropwise via syringe. The reaction mixture was stirred overnight at room temperature and then was sparged with argon for 20 min before being concentrated *in vacuo.* The resulting HCl salt of the dipeptide was carried onto the next step without further purification unless otherwise noted.

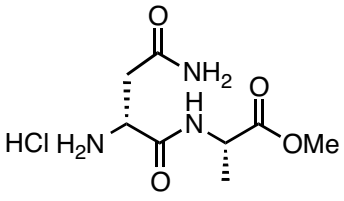
General procedure C: The HCl salt of the dipeptide (1.0 equiv), the carboxylic acid (1.2 equiv), and HOBt monohydrate (1.1 equiv) were dissolved in anhydrous DMF (0.2 M). Triethylamine (2.2 equiv) was added via syringe, and the reaction was stirred at room temperature for 5 min. The reaction mixture was then cooled to 0 °C , and EDC hydrochloride (1.1 equiv) was added. The reaction mixture was stirred at 0 °C for 15 min and was then warmed to room temperature and stirred overnight. The reaction mixture was diluted with water (1 x reaction volume) and ethyl acetate (5 x reaction volume), and the resulting organic layer was washed with an aqueous 1 M

HCl solution. The aqueous layer from the acidic wash was then extracted an additional time with an equal portion of ethyl acetate. The organic layers were then combined and washed with water, saturated aqueous $NaHCO_3$ solution, water, and brine (10 x reaction volume of each). The organic layer was dried over $MgSO_4$, filtered, and concentrated *in vacuo.* The crude material was purified by washing the solid obtained with three portions (2-3 mL) of ice-cold dichloromethane, unless otherwise noted.



33

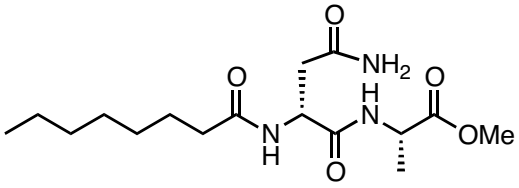
Dipeptide **33** was synthesized using general procedure A. The product was obtained as a white solid (200 mg, 30%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.06 (d, J = 7.4 Hz, 1H, N**H**), 7.25 (broad s, 1H, C(O)N**H₂**), 6.87 (broad s, 1H, C(O)N**H₂**), 6.84 (d, J = 8.3 Hz, 1H, N**H**), 4.24 (m, 2H, C**H**), 3.60 (s, 3H, OC**H₃**), 2.41 (dd, J = 8.7, 12 Hz, 1H, C**H₂**), 2.32 (dd, J = 8.7, 12 Hz, 1H, C**H₂**), 1.36 (s, 9H, C(C**H₃**)$_3$), 1.24 (d, J = 7 Hz, 3H, CHC**H₃**). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.5, 172.08, 171.97, 52.6, 51.8, 48.3, 36.7, 38.0, 35.9, 28.9 (3C), 17.8. HRMS (ESI): calcd for $C_{13}H_{24}N_3O_6^+$ $[M+H]^+$, 318.166; found, 318.1674.



34

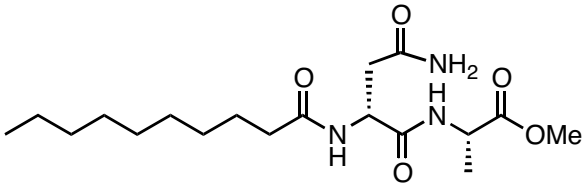
Dipeptide **34** was synthesized from dipeptide **33** using general procedure B and was purified by flash chromatography on silica gel eluting with 4:1:1 isopropanol: 0.1 M HCl: water to afford the

desired product as a white solid (130 mg, 90%). ^{1}H-NMR: (500 MHz, DMSO-d_6): δ 8.98 (d, J = 7 Hz, 1H, N**H**), 8.21 (broad s, 3H, N**H₃**), 7.73 (broad s, 1H, C(O)N**H₂**), 7.24 (broad s, 1H, C(O)N**H₂**), 4.30 (m, 1H, C**H**), 4.05 (dd, J = 4.5, 8 Hz, 1H, C**H**), 3.64 (s, 3H, OC**H₃**), 2.73 (dd, J = 5.1, 17.2 Hz, 1H, C**H₂**), 2.66 (dd, J = 8.5, 16.9 Hz, 1H, C**H₂**), 1.30 (d, J = 6.9 Hz, 3H, CHC**H₃**). ^{13}C NMR (100 MHz, DMSO-d_6): δ 173.1, 171.3, 168.6, 52.8, 49.8, 48.6, 36.2, 17.7. HRMS (ESI): calcd for $C_8H_{16}N_3O_4^+$ $[M+H]^+$, 218.1135; found, 218.1131.



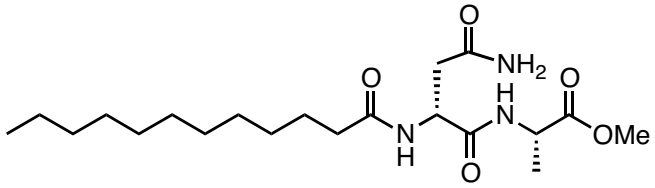
24

N-Acylated dipeptide **24** was synthesized from dipeptide **34** using general procedure C. The product was obtained as a white solid (20 mg, 23%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.06 (d, J = 7 Hz, 1H, N**H**), 7.92 (d, J = 7.1 Hz, 1H, N**H**), 7.25 (broad s, 1H, C(O)N**H₂**), 6.85 (broad s, 1H, C(O)N**H₂**), 4.57 (m, 1H, C**H**), 4.23 (m, 1H, C**H**), 3.59 (s, 3H, OC**H₃**), 2.46 (m, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.30 (dd, J = 7 Hz, 15.4, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.08 (t, J = 7.8 Hz, 2H, C(O)C**H₂**), 1.46 (m, 2H, $C(O)CH_2C\mathbf{H_2}$), 1.23 (m, 11H, CHC**H₃**, C**H₂**), 0.84 (t, J = 7.0 Hz, 3H, $CH_2C\mathbf{H_3}$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.5, 172.9, 171.9, 171.3, 52.5, 50.0, 48.3, 38.0, 35.9, 32.0, 29.23, 29.19, 25.9, 22.7, 17.8, 14.6. HRMS (ESI): calcd for $C_{16}H_{30}N_3O_5^+$ $[M+H]^+$, 344.2180; found, 344.2187.



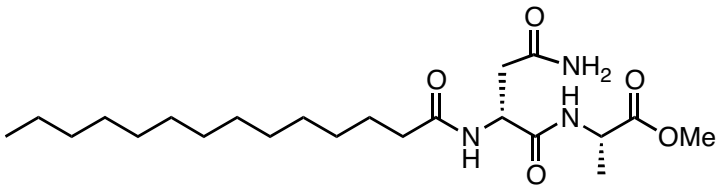
27

N-acylated dipeptide **27** was synthesized from dipeptide **34** using general procedure C, except the crude *N*-acylated dipeptide was washed with three portions (3 x 3 mL) of ice-cold acetone instead of dichloromethane. The product was obtained as a powdery white solid (18 mg, 30%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.08 (d, J = 7.4 Hz, 1H, N**H**), 7.95 (d, J = 8.1 Hz, 1H, N**H**), 7.26 (broad s, 1H, C(O)N**H**$_2$), 6.86 (broad s, 1H, C(O)N**H**$_2$), 4.58 (m 1H, C**H**), 4.24 (m, 1H, C**H**), 3.61 (s, 3H, OC**H**$_3$), 2.48 (m, 1H, NH$_2$C(O)C**H**$_2$), 2.32 (dd, J = 7.8, 15.3 Hz, 1H, NH$_2$C(O)C**H**$_2$), 2.09 (t, J = 8.2 Hz, 2H, C(O)C**H**$_2$), 1.45 (m, 2H, C(O)CH$_2$C**H**$_2$), 1.22 (m, 15H, CHC**H**$_3$, C**H**$_2$), 0.83 (m, 3H, CH$_2$C**H**$_3$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.5, 172.9, 171.9, 171.7, 52.5, 50.0, 48.3, 38.0, 35.9, 32.0, 29.58, 29.55, 29.39, 29.29, 25.9, 22.8, 17.8, 14.7. HRMS (ESI): calcd for $C_{18}H_{34}N_3O_5^+$ $[M+H]^+$, 394.2312; found, 394.2327.



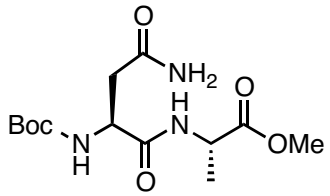
28

N-Acylated dipeptide **28** was synthesized from dipeptide **34** using general procedure C. The product was obtained as a white solid (31 mg, 37%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.06 (d, J = 8.2 Hz, 1H, N**H**), 7.93 (d, J = 8.2 Hz, 1H, N**H**), 7.25 (s, 1H, C(O)N**H**$_2$), 6.84 (s, 1H, C(O)N**H**$_2$), 4.56 (m, 1H, C**H**), 4.23 (m, 1H, C**H**), 3.59 (s, 3H, OC**H**$_3$), 2.46 (m, 1H, NH$_2$C(O)C**H**$_2$), 2.31 (dd, J = 8 Hz, 15.6, 1H, NH$_2$C(O)C**H**$_2$), 2.07 (t, J = 8 Hz, 2H, C(O)C**H**$_2$), 1.45 (m, 2H, C(O)CH$_2$C**H**$_2$), 1.22 (m, 19H, CHC**H**$_3$, C**H**$_2$), 0.84 (t, J = 6.8 Hz, 3H, CH$_2$C**H**$_3$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.5, 172.9, 171.9, 171.7, 52.5, 50.0, 48.3, 38.0, 35.7, 32.0, 29.74, 29.71, 29.64, 29.55, 29.4, 29.3, 25.9, 22.8, 17.8, 14.7. HRMS (ESI): calcd for $C_{20}H_{38}N_3O_5^+$ $[M+H]^+$, 400.2806; found, 400.2816.



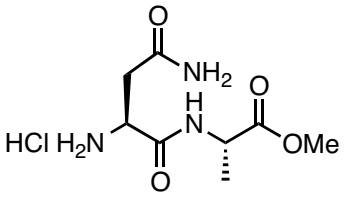
29

N-Acylated dipeptide **29** was synthesized from dipeptide **34** using general procedure C. The product was obtained as a white solid (83 mg, 60%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.05 (d, J = 7.4 Hz, 1H, N**H**), 7.92 (d, J = 8.3 Hz, 1H, N**H**), 7.24 (broad s, 1H, C(O)N**H₂**), 6.84 (broad s, 1H, C(O)N**H₂**), 4.56 (m, 1H, C**H**), 4.22 (m,1H, C**H**), 3.58 (s, 3H, OC**H₃**), 2.46 (m, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.29 (dd, J = 8.4, 15.8 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.15 (t, J = 7.4 Hz, 1H, C(O)C**H₂**), 2.65 (t, J = 7.9 Hz, 1H, C(O)C**H₂**), 1.45 (m, 2H, $C(O)CH_2C\mathbf{H_2}$), 1.22 (m, 23H, CHC**H₃**, C**H₂**), 0.83 (t, J = 6.9 Hz, 3H, $CH_2C\mathbf{H_3}$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.4, 172.9, 171.9, 171.7, 52.6, 50.0, 48.3, 38.0, 35.9, 34.3, 32.0, 29.74, 29.72, 29.6, 29.43, 29.41, 29.2, 26.0, 25.2, 22.8, 17.8, 14.7. HRMS (ESI): calcd for $C_{22}H_{42}N_3O_5^+$ $[M+H]^+$, 428.3119; found, 428.3134.



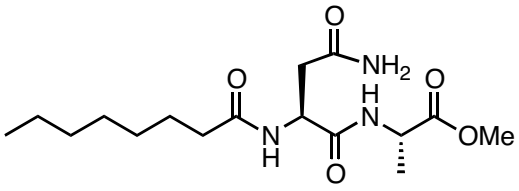
35

Dipeptide **35** was synthesized using general procedure A. The product was obtained as a white solid (180 mg, 35%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.16 (d, J = 6.8 Hz, 1H, N**H**), 7.21 (broad s, 1H, C(O)N**H₂**), 6.86 (m, 2H, C(O)N**H₂,** N**H**), 4.24 (m, 2H, C**H**), 3.59 (s, 3H, OC**H₃**), 2.33 (m, 2H, C**H₂**), 1.35 (s, 9H, $C(C\mathbf{H_3})_3$), 1.24 (d, J = 7.3 Hz, 3H, CHC**H₃**). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.6**,** 172.3, 172.1, 155.8, 78.9, 52.5, 51.6, 48.3, 38.0, 29.1 (3C), 17.6. HRMS (ESI): calcd for $C_{13}H_{24}N_3O_6^+$ $[M+H]^+$, 318.1660; found, 318.1674.



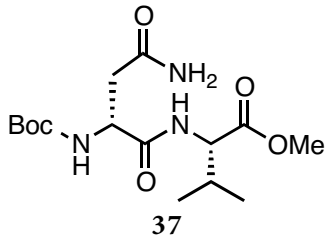
36

Dipeptide **36** was synthesized from **35** using general procedure B. The product was obtained in quantitative yield. ^{1}H-NMR (300 MHz, DMSO-d_6): δ 8.89 (d, J = 7.4 Hz, 1H, N**H**), 8.14 (broad s, 3H, N**H₃**), 7.73 (broad s, 1H, C(O)N**H₂**), 7.28 (broad s, 1H, C(O)N**H₂**), 4.31 (m, 1H, C**H**), 4.08 (m, 1H, C**H**), 3.64 (s, 3H, OC**H₃**), 2.72 (dd, J = 4.0, 17.3 Hz, 1H, C**H₂**), 2.56 (dd, J = 8.9, 17.2 Hz, 1H, C**H₂**), 1.31 (d, J = 7.3 Hz, 3H, CHC**H₃**).



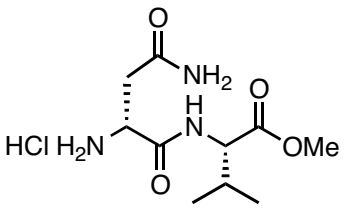
26

N-Acylated dipeptide **26** was synthesized from dipeptide **36** using general procedure C. The product was obtained as a white solid (57 mg, 34%). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.15 (d, J = 8.5 Hz, 1H, N**H**), 7.92 (d, J = 9.1 Hz, 1H, N**H**), 7.23 (s, 1H, C(O)N**H₂**), 6.86 (s, 1H, C(O)N**H₂**), 4.55 (m, 1H, C**H**), 4.22 (m, 1H, C**H**), 3.59 (s, 3H, OC**H₃**), 2.4 (dd, J = 4.7, 15.5 Hz, 1H, $NH_2C(O)C$**H₂**), 2.32 (dd, J =8, 17.4 Hz, 1H, $NH_2C(O)C$**H₂**), 2.06 (t, J = 7.4 Hz, 2H, C(O)C**H₂**), 1.44 (m, 2H, $C(O)CH_2C$**H₂**), 1.23 (m, 11H, CHC**H₃**, C**H₂**), 0.84 (t, J = 7.7 Hz, 3H, CH_2C**H₃**). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 172.9, 172.2, 171.34, 171.28, 51.9, 49.3, 47.6, 37.3, 35.2, 31.2, 29.6, 29.5, 25.2, 22.1, 16.9, 14.0. HRMS (ESI): calcd for $C_{16}H_{30}N_3O_6^+$ $[M+H]^+$, 344.2180; found, 344.2192.



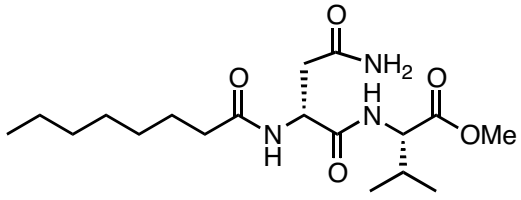
37

Dipeptide **37** was synthesized using general procedure A. The product was obtained as a white solid (362 mg, 58%). ^{1}H-NMR (500 MHz, $CDCl_3$): δ 7.35 (broad s, 1H, N**H**), 6.12 (broad s, 1H, N**H**), 5.96 (broad s, 1H, C(O)N**H₂**), 5.63 (broad s, 1H, C(O)N**H₂**), 4.50 (m, 1H, C**H**), 4.44 (dd, *J* = 5, 8.4 Hz, 1H, C**H**), 3.72 (s, 3H, OC**H₃**), 2.88 (d, *J* = 15.5 Hz, 1H, $NH_2C(O)CH_2$), 2.59 (dd, *J* = 6 Hz, 14.8 Hz, 1H, $NH_2C(O)CH_2$), 2.18 (m, 1H, C**H**$(CH_3)_2$), 1.45 (s, 9H, $C(CH_3)_3$), 0.93 (d, *J* = 7.4 Hz, 3H, $CH(CH_3)_2$), 0.89 (d, *J* = 6.4 Hz, 3H, $CH(CH_3)_2$). ^{13}C-NMR (100 MHz, $CDCl_3$): δ 173.7, 172.1, 171.57, 156.3, 80.6, 57.7, 52.5, 51.4, 37.3, 31.4, 28.6 (3C), 19.3, 17.9. HRMS (ESI): calcd for $C_{15}H_{28}N_3O_6^+$ $[M+H]^+$, 346.1978; found, 346.1997.



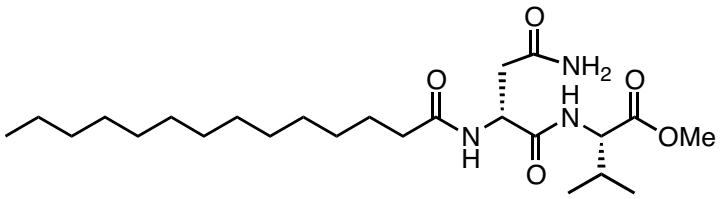
38

Dipeptide **38** was synthesized using general procedure B. The product was obtained as a white solid (132 mg, 99%). ^{1}H-NMR (500 MHz, DMSO-d_6: δ 8.76 (d, *J* = 8.3 Hz, 1H, N**H**), 8.26 (broad s, 3H, N**H₃**), 7.80 (broad s, 1H, C(O)N**H₂**), 7.28 (broad s, 1H, C(O)N**H₂**), 4.24 (m, 1H, C**H**), 4.14 (m, 1H, C**H**), 3.66 (s, 3H, OC**H₃**), 2.75 (dd, *J* = 6.4, 16.7 Hz, 1H, $NH_2C(O)CH_2$), 2.69 (dd, *J* = 8.4, 17.2 Hz, 1H, $NH_2C(O)CH_2$), 2.06 (m, 1H, C**H**$(CH_3)_2$), 0.86 (d, *J* = 2.5 Hz, 3H, $CH(CH_3)_2$), 0.84 (d, *J* = 4.2 Hz, 3H, $CH(CH_3)_2$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 172.1, 171.5, 169.0, 58.1, 52.6, 49.8, 36.5, 30.85, 19.6, 18.6. HRMS (ESI): calcd for $C_{10}H_{20}N_3O_4^+$ $[M+H]^+$, 246.1454; found, 246.1476.



25

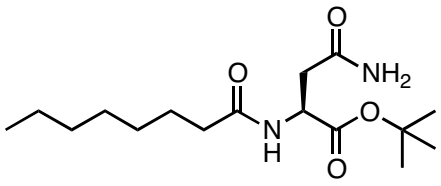
N-Acylated dipeptide **25** was synthesized from dipeptide **38** using general procedure C. The product was obtained as a white solid (82 mg, 14%). ^{1}H-NMR (500 MHz, $CDCl_3$): δ 7.59 (d, *J* = 8 Hz, 1H, N**H**), 7.32 (d, *J* = 6.8 Hz, 1H, N**H**), 6.25 (broad s, 1H, C(O)N**H₂**), 5.65 (broad s, 1H, C(O)N**H₂**), 4.82 (m, 1H, C**H**), 4.44 (dd, *J* = 5.2, 8.4 Hz, 1H, C**H**), 3.72 (s, 3H, OC**H₃**), 2.89 (dd, *J* = 4.4, 14.9 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.55 (dd, *J* = 6.3, 14.7 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.27 (t, *J* = 7.8 Hz, 2H, $C(O)C\mathbf{H_2}$), 2.19 (m, 1H, $C\mathbf{H}(CH_3)_2$), 1.65 (m, 2H, $C(O)CH_2C\mathbf{H_2}$), 1.29 (m, 8H, C**H₂**), 0.95 (d, *J* = 7.2 Hz, 3H, $CH(C\mathbf{H_3})_2$), 0.91 (d, *J* = 7.0 Hz, 3H, $CH(C\mathbf{H_3})_2$), 0.88 (m, 3H, $CH_2C\mathbf{H_3}$). ^{13}C-NMR (100 MHz, $CDCl_3$): δ 174.3, 174.1, 172.0, 171.1, 57.9, 52.4, 50.0, 36.82, 36.76, 31.9, 31.2, 29.5, 29.2, 25.9, 22.8, 19.4, 17.8, 14.3. HRMS (ESI): calcd for $C_{18}H_{34}N_3O_5^+$ $[M+H]^+$, 372.2493; found: 372.2497.



30

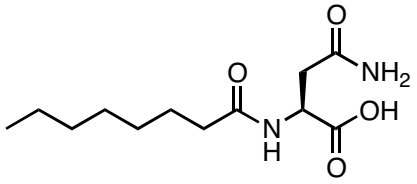
N-Acylated dipeptide **30** was synthesized from dipeptide **38** using general procedure C. The product (24 mg, 12%) was obtained as a white solid. ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.02 (d, *J* = 8.2 Hz, 1H, N**H**), 7.92 (d, *J* = 8.1 Hz, 1H, N**H**), 7.21 (s, 1H, C(O)N**H₂**), 6.89 (s, 1H, C(O)N**H₂**), 4.65 (m, 1H, C**H**), 4.17 (m, 1H, C**H**), 3.63 (s, 3H, OC**H₃**), 2.49 (m, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.34 (dd, *J* = 8.3, 15.9 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.09 (t, *J* = 7.8 Hz, 2H, $C(O)CH_2C\mathbf{H_2}$), 2.02 (m, 1H, $C\mathbf{H}(CH_3)_2$), 1.46 (m, 2H, $C(O)CH_2C\mathbf{H_2}$), 1.23 (m, 20H, C**H₂**), 0.84 (m, 9H, $CH_2C\mathbf{H_3}$, $CH(C\mathbf{H_3})_2$). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 172.4, 171.7, 171.4, 171.3, 57.1, 51.7, 49.4, 37.4, 35.2, 31.3, 30.1, 29.1 (2C), 29.0, 28.92, 28.86, 28.75, 28.71, 28.6, 25.2, 22.1, 18.9, 17.9, 13.9. HRMS (ESI): calcd for $C_{24}H_{46}N_3O_5^+$ $[M+H]^+$, 456.3432; found, 456.3431.

Synthesis of standards for LC-MS and HPLC assays



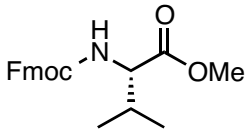
39

To L-asparagine-*O*(*t*-Bu)-HCl (Bachem) (200 mg, 0.89 mmol, 1.0 equiv) was added dry dimethylformamide (3.56 mL, 0.25 M) followed by octanoic acid (169 μL, 1.068 mmol, 1.2 equiv), HOBt monohydrate (TCI) (149.9 mg, 0.979 mmol, 1.1 equiv), and *N,N*-diisopropylethylamine (341 μL, 1.96 mmol, 2.2 equiv). The reaction mixture was stirred for 5 min and was then placed in an ice bath. EDC hydrochloride (187.7 mg, 0.979 mmol, 1.1 equiv) was added, and the reaction mixture was stirred at 0 °C for 15 min. The ice bath was then removed and the reaction mixture was stirred at room temperature overnight. The reaction mixture was diluted with water until the solution became cloudy (15 mL) and was diluted further with ethyl acetate (30 mL). The organic layer was washed with a saturated aqueous NH_4Cl solution (30 mL). The aqueous phase from the NH_4Cl wash was extracted with another two portions of ethyl acetate (2 x 20 mL), and the organics were then combined and washed with water (30 mL) and then brine (30 mL). The organics were dried over $MgSO_4$, filtered, and concentrated *in vacuo* to afford a white solid in quantitative yield. This material was used directly in the next reaction without further purification. ^{1}H-NMR (500 MHz, $CDCl_3$): δ 6.64 (d, J = 7.5 Hz, 1H, N**H**), 5.87 (broad s, 1H, C(O)N**H**$_2$), 5.67 (broad s, 1H, C(O)N**H**$_2$), 4.61 (m, 1H, NHC**H**), 2.89 (dd, J = 5.2, 16.6 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.77 (dd, J = 3.9, 15.8 Hz, 1H, $NH_2C(O)C\mathbf{H_2}$), 2.22 (t, J = 8.3 Hz, 2H, $C(O)C\mathbf{H_2}$), 1.62 (m, 2H, $C(O)CH_2C\mathbf{H_2}$), 1.47 (s, 9H, $C(C\mathbf{H_3})_3$), 1.29 (m, 8H, $C\mathbf{H_2}$), 0.88 (m, 3H, $CH_2C\mathbf{H_3}$). ^{13}C-NMR (100 MHz, $CDCl_3$): δ 173.6, 173.0, 170.2, 82.8, 49.6, 37.7, 36.9, 31.9, 29.4, 29.2, 28.2 (3C), 25.9, 22.8, 14.3.



40

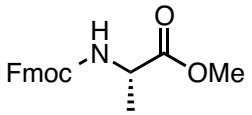
To a solution of **39** (0.89 mmol, 1 equiv) in dichloromethane (1.5 mL, 0.6 M) was added triisopropylsilane (2.67 mmol, 547 µL, 3 equiv) via syringe. The reaction mixture was placed in an ice bath and trifluoroacetic acid (9.345 mmol, 804 µL, 10.5 equiv) was added dropwise over the course of 5 min. The flask was then covered in tinfoil and removed from the ice bath. The reaction mixture was stirred for 15 h at room temperature. Toluene (2 x 5 mL) was added and the crude mixture was concentrated *in vacuo* two times to afford a quantitative yield of a white solid, which was used as a standard for LC-MS without further purification. ^{1}H-NMR (500 MHz, DMSO-d_6): δ 7.96 (d, J = 8.4 Hz, 1H, N**H**), 7.31 (s, 1H, C(O)N**H₂**), 6.87 (s, 1H, C(O)N**H₂**), 4.46 (q, J = 7.6 Hz, 1H, C**H**), 2.51 (m, 1H, $NH_2C(O)CH_2$), 2.41 (dd, J = 5.1, 8.2 Hz, 1H, $NH_2C(O)CH_2$), 2.05 (t, J = 7.2 Hz, 2H, C(O)C**H₂**), 1.45 (m, 2H, $C(O)CH_2CH_2$), 1.22 (m, 8H, C**H₂**), 0.84 (t, J = 7 Hz, 3H, CH_2CH_3). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.0, 172.0, 171.2, 48.7, 36.8, 35.1, 31.2, 28.5, 28.5, 25.2, 22.1, 14.0. HRMS (ESI): calcd for $C_{12}H_{23}N_2O_4^+$ $[M+H]^+$, 259.1652; found, 259.1660.



41

L-Valine methyl ester hydrochloride (100 mg, 0.597 mmol, 1.1 equiv) was dissolved in 1,4-dioxane (750 µL, 0.8 M), and the reaction flask was placed in an ice bath. To this was added an aqueous 10% Na_2CO_3 solution (1.5 mL), followed by fluorenylmethyloxycarbonyl chloride (0.542

mmol, 1 equiv) dissolved in 1,4-dioxane (1.4 mL, 0.4 M). This mixture was stirred for 5 min and was then allowed to warm to room temperature. After one hour the reaction was complete as judged by TLC (1:1 ethyl acetate: hexanes). The reaction mixture was diluted with water (5 mL) and extracted into dichloromethane (20 mL). The organic phase was washed with saturated aqueous NH_4Cl (20 mL), water (20 mL), and brine (20 mL) and was dried over $MgSO_4$, filtered, and concentrated *in vacuo.* The crude product was purified by flash chromatography on silica gel eluting with a gradient of 10% to 25% ethyl acetate in hexanes to afford the desired product as a white solid (55 mg, 26%). ^{1}H-NMR (500 MHz, $CDCl_3$): δ 7.77 (d, *J* = 8.1 Hz, 2H, ArC**H**), 7.61 (dd, *J* = 2.6, 7.4 Hz, 2H, ArC**H**), 7.41 (t, *J* = 7.3 Hz, 2H, ArC**H**), 7.32 (t, *J* = 7.3 Hz, 2H, ArC**H**), 5.34 (d, *J* = 9.4 Hz, 1H, N**H**), 4.41 (m, 2H, COC**H₂**), 4.33 (dd, *J* = 4.7, 9.8 Hz, 1H, C**H**), 4.24 (t, *J* = 7.4 Hz, 1H, C**H**CH_2O), 3.76 (s, 3H, OC**H₃**), 2.18 (m, 1H, C**H**$(CH_3)_2$), 0.98 (d, *J* = 7.3 Hz, 3H, CH(C**H₃**)$_2$), 0.92 (d, *J* = 5.3 Hz, 3H, CH(C**H₃**)**₂**). ^{13}C-NMR (100 MHz, $CDCl_3$): δ 172.7, 156.3, 143.9 (2C), 141.4 (2C), 127.8 (2C), 127.2 (2C), 125.2 (2C), 120.1 (2C), 67.1, 59.2, 52.3, 47.3, 31.4, 19.1, 17.8. HRMS (ESI): calcd for $C_{21}H_{23}NO_4Na^+$ $[M+Na]^+$, 376.1519; found, 376.1520.



42

L-Alanine methyl ester hydrochloride (200 mg, 1.43 mmol, 1.1 equiv) was dissolved in 1,4-dioxane (1.6 mL, 0.8 M), and the reaction flask was placed in an ice bath. To this was added an aqueous 10% Na_2CO_3 solution (1.5 mL), followed by fluorenylmethyloxycarbonyl chloride (337 mg, 1.3 mmol 1 equiv) dissolved in 1,4-dioxane (3.25 mL, 0.4 M). This mixture was stirred for 5 min and was then allowed to warm to room temperature. After 2 hours the reaction was complete

as judged by TLC (1:1 ethyl acetate:hexanes, R_f=0.45). The reaction mixture was diluted with water (5 mL), and extracted into dichloromethane (30 mL). The organic phase was washed with 1 M aqueous HCl (30 mL), water (30 mL), and brine (30 mL) and was dried over $MgSO_4$, filtered, and concentrated *in vacuo*. The crude product was purified by flash chromatography on silica gel eluting with a gradient of 0% to 30% ethyl acetate in hexanes to afford the desired product as a white solid (318 mg, 68%). ^{1}H-NMR (500 MHz, $CDCl_3$) and ^{13}C-NMR (100 MHz, $CDCl_3$) spectra match those reported previously.[33] HRMS (ESI): calcd for $C_{19}H_{21}NO_4^+$ (M+H)+, 326.1387; found, 326.1400.

2.7: References

(1) Cundliffe, E. *Annu. Rev. Microbiol.* **1989**, *43*, 207.

(2) Llano-Sotelo, B.; Azucena, E. F.; Kotra, L. P.; Mobashery, S.; Chow, C. S. *Chem. Biol.* **2002**, *9*, 455.

(3) Cundliffe, E. *Nature* **1978**, *272*, 792.

(4) Galm, U.; Hager, M. H.; Van Lanen, S. G.; Ju, J.; Thorson, J. S.; Shen, B. *Chem. Rev.* **2005**, *105*, 739.

(5) Reimer, D.; Pos, K. M.; Thines, M.; Grün, P.; Bode, H. B. *Nat. Chem. Biol.* **2011**, *7*, 888.

(6) Oefner, C.; D'Arcy, A.; Daly, J. J.; Gubernator, K.; Charnas, R. L.; Heinze, I.; Hubschwerlen, C.; Winkler, F. K. *Nature* **1990**, *343*, 284.

(7) Kevany, B. M.; Rasko, D. A.; Thomas, M. G. *Appl. Environ. Microbiol.* **2009**, *75*, 1144.

(8) La Clair, J. J.; Foley, T. L.; Schegg, T. R.; Regan, C. M.; Burkart, M. D. *Chem. Biol.* **2004**, *11*, 195.

(9) Balskus, E. P.; Walsh, C. T. *Science* **2010**, *329*, 1653.

(10) Rausch, C.; Hoof, I.; Weber, T.; Wohlleben, W.; Huson, D. H. *BMC Evol. Biol.* **2007**, *7*, 78.

(11) Tang, G.-L.; Cheng, Y.-Q.; Shen, B. *J. Bio. Chem.* **2007**, *282*, 20273.

(12) Kraas, F. I.; Helmetag, V.; Wittmann, M.; Strieker, M.; Marahiel, M. A. *Chem. Biol.* **2010**, *17*, 872.

(13) Bachmann, B. O.; Ravel, R. *Meth. Enzymol.* **2009**, *458,* 181.

(14) Miao, V. *Microbiology* **2005**, *151*, 1507.

(15) Galli, G.; Rodriguez, F.; Cosmina, P.; Pratesi, C.; Nogarotto, R.; de Ferra, F.; Grandi, G. *Biochim. Biophys.* **1994**, 19.

(16) Imker, H. J.; Krahn, D.; Clerc, J.; Kaiser, M.; Walsh, C. T. *Chem. Biol.* **2010**, *17*, 1077.

(17) Janse, J.D.; Smits, P.H. *Lett. Appl. Microbiol.* **1990**, *10*, 131.

(18) Marr, A. G.; Ingraham, J. L. *J. Bacteriol.***1962**, *84*, 1260.

(19) Pfeifer, B. A.; Admiraal, S. J.; Gramajo, H.; Cane, D. E.; Khosla, C. *Science* **2001**, *291*, 1790.

(20) Stachelhaus, T.; Walsh, C. T. *Biochemistry* **2000**, *39*, 5775.

(21) Clugston, S. L.; Sieber, S. A.; Marahiel, M. A.; Walsh, C. T. *Biochemistry* **2003**, *42*, 12095.

(22) Lautru, S.; Deeth, R. J.; Bailey, L. M.; Challis, G. L. *Nat. Chem. Biol.* **2005**, *1*, 265.

(23) Oefner, C.; D'Arcy, A.; Daly, J. J.; Gubernator, K.; Charnas, R. L.; Heinze, I.; Hubschwerlen, C.; Winkler, F. K. *Nature* **1990**, *343*, 284.

(24) Käll,L.; Krogh,A.; Sonnhammer, E.L.L. *J. Mol. Biol.* **2004**, *338*,1027.

(25) Yu, N.Y.; Wagner, J.R.; Laird, M.R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S.C.; Ester,M.; Foster, L.J.; Brinkman, F.S.L. *Bioinformatics* **2010**, *26*,1608.

(26) Nougayrede, J.-P.; Homburg, S.; de ric Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J. R.; Dobrindt, U.; Oswald, E. *Science* **2006**, *313*, 848.

(27) Cougnoux, A.; Gibold, L.; Robin, F.; Dubois, D.; Pradel, N.; Darfeuille-Michaud, A.; Dalmasso, G.; Delmas, J.; Bonnet, R. *J. Mol. Biol.* **2012**, *424*, 203.

(28) Dubois, D.; Baron, O.; Cougnoux, A.; Delmas, J.; Pradel, N.; Boury,M.; Bouchon,B.; Bringer,M.; Nougayrede,J.; Oswald,E.; Bonnet, R. *J. Biol. Chem.* **2011**, *286*, 35562.

(29) Jiang, W.; HeemstraJr, J. R.; Forseth, R. R.; Neumann, C. S.; Manaviazar, S.; Schroeder, F. C.; Hale, K. J.; Walsh, C. T. *Biochemistry* **2011**, *50*, 6063.

(30) Cougnoux, A.; Delmas, J.; Gibold, L.; Faïs, T.; Romagnoli, C.; Robin, F.; Cuevas-Ramos, G.; Oswald, E.; Darfeuille-Michaud, A.; Prati, F.; Dalmasso, G.; Bonnet, R. *Gut* **2015**, *0,* 1.

(31) Bian, X.; Fu, J.; Plaza, A.; Herrmann, J.; Pistorius, D.; Stewart, A. F.; Zhang, Y.; Müller, R. *ChemBioChem* **2013**, *14*, 1194.

(32) Sigal, C. T.; Zhou, W.; Buser, C. A.; McLaughlin, S.; Resh, M. D. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 12253.

(33) Hayashida, O. Sebo, L. Rebek, J. *J. Org. Chem*. **2002**, *67*, 8291.

Chapter 3: Isolation of a *pks*-associated metabolite

3.1: The prodrug resistance strategy guides the discovery of *pks* metabolites

The knowledge gained from our initial biochemical studies of the colibactin biosynthetic pathway provided a starting point for designing strategies to isolate *pks*-associated metabolites. In particular, the characterization of the colibactin self-resistance machinery uncovered the precise structural features of the prodrug motif. Since *clbP* is required for genotoxicity, we hypothesized that precolibactin, the precursor to the active colibactin, must contain an *N*-myristoylated-D-asparagine motif (Figure 3.1). Furthermore, we hypothesized that the N-terminal prodrug motif modulates both the localization and reactivity of precolibactin.

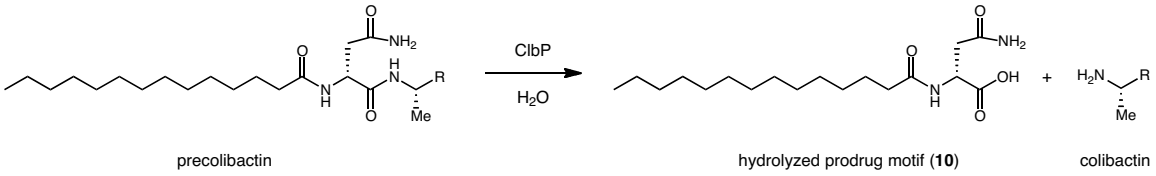


Figure 3.1: Periplasmic peptidase ClbP cleaves the prodrug motif from precolibactin to provide the hydrolyzed prodrug motif (**10**) and the active colibactin.

The presence of several *trans*-AT PKSs predicted to incorporate aminomalonate in the *pks* cluster suggested colibactin contained multiple positively charged amines. Positively charged amino acids increase the binding of myristoylated proteins to cell membranes due to strong electrostatic interactions with acidic phospholipids, a phenomenon named the myristoyl-electrostatic switch.[1] We hypothesized that interactions of the lipophilic myristoyl chain with membrane lipids combined with the predicted electrostatic interactions with membrane polar head groups would localize precolibactin in the membrane fraction of the producing organism.

In addition to the effects of the prodrug motif on localization, we hypothesized that hydrolysis of the prodrug motif by ClbP would unveil a reactive primary amine that could contribute to the instability of colibactin. The primary amine could react in intra- or inter- molecular degradation pathways, for instance through a condensation reaction with a ketone to form an iminium ion that could undergo further reactions. Thus, we decided to target precolibactins for isolation, which could be achieved by the extraction of Δ *clbP pks*$^+$ *E. coli* cultures. We believed precolibactin's cellular location could guide extraction techniques and its predicted lower reactivity in comparison to colibactin could aid in isolation and purification.

3.2: Global metabolite profiling identifies several candidate precolibactins

The classic approach to natural product isolation relies on activity-guided fractionation. Despite the potent bioactivity of colibactin, we concluded that activity-guided fractionation could prove problematic as an approach to identify precolibactin. This was due to the fact that the lack of genotoxicity of Δ *clbP pks*$^+$ *E. coli* could be explained through two different mechanisms. First, membrane localization of precolibactin could prevent diffusion of the molecule(s) out of the cell. Second, the prodrug motif could mask the reactivity of colibactin, such that even if precolibactin were able to reach the human cells, it would not cause double-strand breaks in DNA. Thus, in order to use activity-guided fraction to identify precolibactin, a method to cleave the prodrug motif from precolibactin in the presence of human cells could be required.

Instead of activity-guided fractionation, we relied on an unbiased LC–MS-based method called global metabolite profiling to identify precolibactin (Figure 3.2). This approach was pioneered in a study of the mammalian enzyme fatty acid amide hydrolase (FAAH).[2] While FAAH had been characterized *in vitro*, its substrate preferences *in vivo* were unknown. To address this problem, the metabolome of the brain tissue of wild-type mice was compared to that of FAAH knock-out

mice. The LC–MS spectra of tissue extracts from the two sets of samples were manually compared to find peaks enriched in the FAAH knock-out mice, which the authors considered to be candidate endogenous substrates of FAAH. This approach led to the identification of novel metabolites and FAAH substrates in the brain. More recently, global metabolite profiling has been made more accessible by the development XCMS, a web-based processing and analysis program, which performs the analysis that was done manually in earlier studies.[3] Researchers have utilized XCMS to connect the proteome to the metabolome in the study of human disease[4] and to discover novel microbial metabolites.[5]

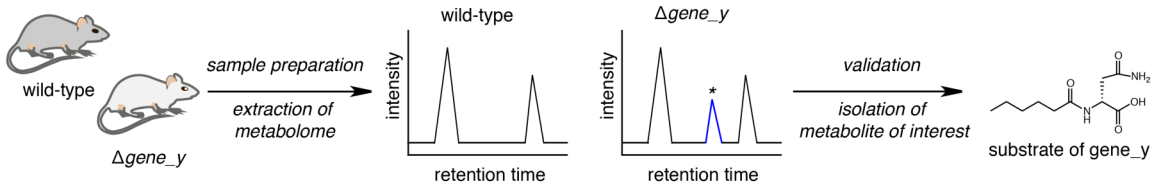


Figure 3.2: A general overview of global metabolite profiling. First, organisms expressing the wild-type gene or lacking a particular gene of interest (e.g. *gene_y*) are obtained. The pertinent samples are obtained, and the metabolome is extracted. The extracts from the two sample types are then analyzed using LC–MS. Peaks that are enriched in the mutant samples (blue peak) may represent substrates of the protein product of the mutated gene (e.g. gene_y). These peaks can be investigated further, for instance through LC-MS guided isolation of the metabolite of interest.

We began by comparing the metabolite profiles of extracts of *E. coli* DH10B expressing the wild-type *pks* cluster (BAC*pks*) or the *pks* cluster with *clbP* knocked out (BAC*pks*Δ *clbP*). These strains harbor a bacterial artificial chromosome (BAC) that encodes the complete *pks* cluster cloned from *E. coli* IHE3034.[6] The initial extraction method for precolibactin was guided by methods used to isolate a variety of membrane-associated and lipophilic compounds, such as isoprenoid quinones (**43**) and sulfonolipids (**44**) (Figure 3.3)[7,8] Large culture volumes (1 L) of *E. coli* DH10B harboring BAC*pks* or BAC*pks*Δ *clbP* were grown in triplicate in Luria–Bertani (LB) broth. The dried cell mass was extracted with a 2:1:0.8 mixture of chloroform, methanol and water. This extract was dried *in vacuo* and resuspended in methanol for LC–MS analysis. The raw LC–MS

data was analyzed using XCMS, which identified ten features that varied significantly between the wild-type and *clbP* knock-out extracts ($p < 0.01$), had a maximum intensity of greater than 1000 counts and were enriched by ten-fold or greater (Table 3.1). Of particular interest were features 2, 4 and 10 in Table 3.1. Feature 10 had a molecular weight that corresponded to the hydrolyzed prodrug motif (**10**). The presence of this molecular feature validated the general method used to produce and compare the metabolite extracts. Features 2 and 4 were hypothesized to be the $[M+H]^+$ and $[M+Na]^+$ adducts of the same parent molecule. The presence of multiple masses enriched in the *clbP* knock-out extracts suggested that there were multiple *pks*-associated metabolites and not just one precursor compound or precolibactin.

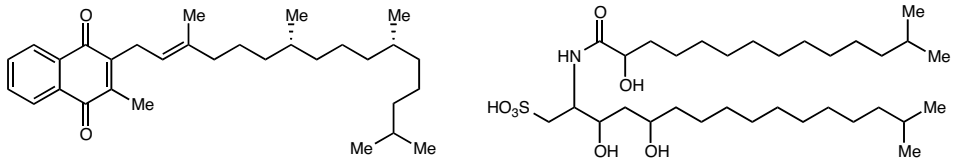


Figure 3.3: Lipophilic molecules **43** and **44** were extracted from bacteria using approaches based on Bligh and Dyer's method for the extraction of fatty acids.[7,8,9]

Table 3.1: Results from XCMS analysis. "DOWN" indicates enrichment in BAC*pks*Δ *clbP*, and "UP" indicates enrichment in BAC*pks*. Average intensity: the average integrated intensity of that feature across samples in each condition, either BAC*pks*Δ *clbP* or BAC*pks*. Fold change: the ratio of the two average intensities of that feature from each condition. P-value: the significance of the difference for a given feature between the two conditions (calculated in the XCMS program using a Welch t-test with unequal variances).

Feature	UP/ DOWN	*m/z*	Average intensity: Δ *clbP*	Average intensity: wild-type	Fold change	p-value	Retention time (min)
1	DOWN	599.3755	52,365	4,516	11.6	0.00003	26.56
2	DOWN	796.3524	80,418	7,081	11.4	0.00035	27.75
3	DOWN	1,240.94	22,029	1,719	10.8	0.00147	38.56
4	DOWN	818.3383	32,300	1,097	29.5	0.00352	27.76
5	DOWN	1,113.46	43,475	2,856	13.7	0.00623	38.85
6	DOWN	1,551.43	15,425	407	14.8	0.00741	38.61
7	DOWN	1,240.55	10,672	144	13.7	0.00871	38.62
8	UP	684.432	2,469	104,237	37.4	0.00017	43.21
9	UP	326.2421	4,864	54,353	11.2	0.0013	29.04
10	UP	343.2605	156,025	1,653,981	10.6	0.00093	29.04

To identify which of these features contained the prodrug motif and could correspond to precolibactin, and to test the reproducibility of our results from the first experiment, we sought to perform a feeding experiment with deuterated myristic acid (**45**). We rationalized that any features identified in the first round of experiments that contained the prodrug motif should be labeled in a feeding experiment with **45**, and undergo a +27 mass shift relative to the unlabeled compound. This mass shift would allow for identification of precolibactins from the set of ten features identified in the first comparative LC–MS experiment.

First, a feeding experiment was performed to assess the efficiency of incorporation of **45** into the natural product scaffold. In this assay, **45** (500 μM) was fed to *E. coli* DH10B harboring BAC*pks* at different stages of growth, from the beginning of growth (t = 0) to saturation (OD_{600} = 1.0). The area under the curve for the extracted ion chromatograms (EICs) of both the unlabeled (**10**) and labeled (**46**) hydrolyzed prodrug motif were compared. Compound **46** was seen in all cultures to

which **45** was added. The extent of incorporation did not vary greatly over different time points of addition of **45**, and was around 0.4% in all conditions (Figure 3.4).

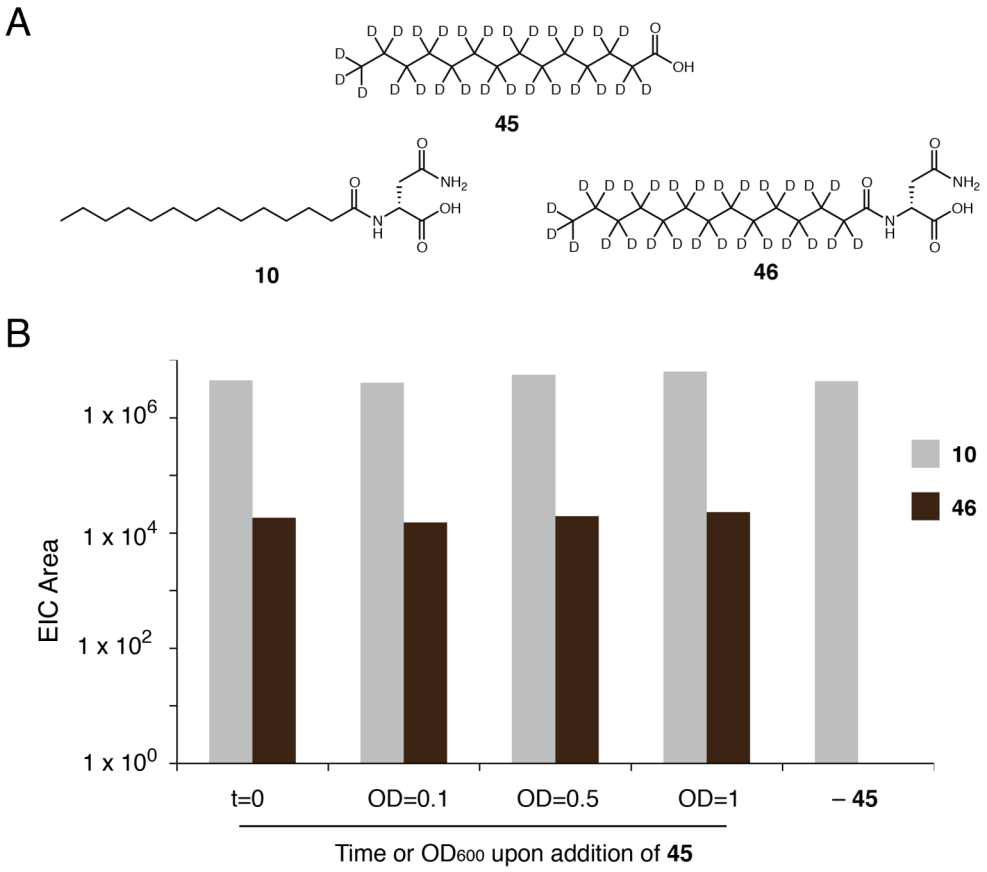


Figure 3.4: A feeding experiment with **45** resulted in ~0.4% incorporation of the labeled building block into the hydrolyzed prodrug motif. A) Structures of labeled myristic acid (**45**), as well as the hydrolyzed, unlabeled prodrug motif (**10**), and the labeled prodrug motif (**46**). B) The area under the curve of the EIC is provided for **10** and **46**.

The comparative LC–MS feeding experiment was performed analogously to that described above. Deuterated myristic acid (**45**) was added at the beginning of growth. In this experiment, greater than 50 features varied significantly between the extracts of the wild-type and Δ *clbP pks*$^+$ strains ($p < 0.01$), had an intensity of greater than 1000 and were enriched by ten-fold or greater. Interestingly, of those ten features that met these criteria in the earlier experiment (features 1-10, Table 3.1) only feature 2 (*m/z* 796.3524) also met these criteria in the second experiment.

Furthermore, a small set of features met stricter criteria, with $p < 0.01$, maximum intensity greater than 10,000 and enrichment greater than ten-fold (Table 3.2). We hypothesized that features 11-12 and 13-14 were the $[M+Na]^+$ and $[M+K]^+$ adducts of the same parent molecules with neutral masses of 439.3410 and 546.3781, respectively. While this feeding experiment identified additional metabolites that possibly contained the prodrug motif, the results also highlighted that the growth and extraction methods did not lead to a high degree of reproducibility between two separate experiments. In addition, perhaps because of the low degree of incorporation of **45**, this experiment failed to directly identify those metabolites from the first experiment that contained the prodrug motif.

Table 3.2: Results from XCMS analysis from the deuterated myristic acid feeding experiment. "DOWN" indicates enrichment in BAC*pks*Δ *clbP*, and "UP" indicates enrichment in BAC*pks*. Average intensity: the average integrated intensity of that feature across samples in each condition, either BAC*pks*Δ *clbP* or BAC*pks*. Fold change: the ratio of the two average intensities of that feature from each condition. p-value: the significance of the difference for a given feature between the two conditions (calculated in the XCMS program using a Welch t-test with unequal variances).

Feature	UP/ DOWN	*m/z*	Average intensity: Δ *clbP*	Average intensity: wild-type	Fold change	p-value	Retention time (min)
11	DOWN	462.3305	2,412,998	24,190	99.8	0.00434	34.21
12	DOWN	478.3053	569,718	22,110	25.8	0.00302	34.21
13	DOWN	569.3685	888,939	7,807	113.9	0.00343	33.61
14	DOWN	585.3421	134,886	7,193	18.8	0.00153	33.62
15	DOWN	713.3692	168,806	9,629	17.5	0.00759	33.01

We found that the some of the features identified in the first and second experiments could be seen when the culture volume was decreased and the extraction protocol was simplified. In a third experiment, small scale (5 mL) cultures were grown in triplicate, and a whole-culture aliquot (500 µL) was lyophilized and extracted with methanol (500 µL). XCMS analysis identified nine features that varied significantly between the wild-type and *clbP* knock-out extracts ($p < 0.01$), had a maximum intensity of greater than 1000 and were enriched by ten-fold or greater

(Table 3.3). Included in this set of features were the hydrolyzed prodrug motif (**10**) (features 22-24), as well as features that appeared to be adducts of the same parent molecules with neutral masses of 439.3410 (features 16, 18 and 19) and 546.3781 (feature 17). Adducts of these same neutral masses were also seen in previous experiments.

Table 3.3: Results from XCMS analysis from a smaller scale experiment. "DOWN" indicates enrichment in BAC*pks*Δ *clbP*, and "UP" indicates enrichment in BAC*pks*. Average intensity: the average integrated intensity of that feature across samples in each condition, either BAC*pks*Δ *clbP* or BAC*pks*. Fold change: the ratio of the two average intensities of that feature from each condition. P-value: the significance of the difference for a given feature between the two conditions (calculated in the XCMS program using a Welch t-test with unequal variances).

Feature	UP/ DOWN	*m/z*	Average intensity: Δ clbP	Average intensity: wild-type	Fold change	p-value	Retention time (min)
16	DOWN	440.3561	199,038	15,301	13	0.00112	35.04
17	DOWN	569.3775	214,197	0	214.6	0.00143	34.61
18	DOWN	462.3386	616,526	6,630	93	0.00195	35.07
19	DOWN	463.3415	173,195	1,535	112.8	0.00208	35.07
20	DOWN	414.303	144,223	8,457	17.1	0.00216	33.65
21	DOWN	662.4827	221,255	2,950	75	0.00706	41.83
22	UP	344.262	8,032	199,126	24.8	0.00328	34.37
23	UP	343.2549	9,930	936,052	94.3	0.0035	34.36
24	UP	365.2452	4,873	194,989	40	0.00436	34.35

With these results in hand, we chose to study compounds with protonated molecular ion masses of *m/z* 440.3483 (**47**), *m/z* 547.3854 (**48**) and *m/z* 796.3521 (**49**), as adducts of these particular compounds were identified as significantly enriched in the extracted metabolome of BAC*pks*Δ *clbP* cultures in multiple XCMS experiments. We were specifically interested in determining whether these metabolites contained the prodrug motif and could serve as candidates for isolation studies. When extracts from the feeding experiment described above were analyzed by LC–MS, masses corresponding to the deuterium labeled versions (+27) of **47** and **48** were observed, although at much lower intensities than the unlabeled compounds (Figure 3.5). We searched for loss of the prodrug motif and fragments thereof in the MS/MS spectra of these metabolites. Encouragingly,

the MS/MS fragmentation patterns strongly suggested that these **47**-**49** contained the prodrug motif (Figure 3.6).

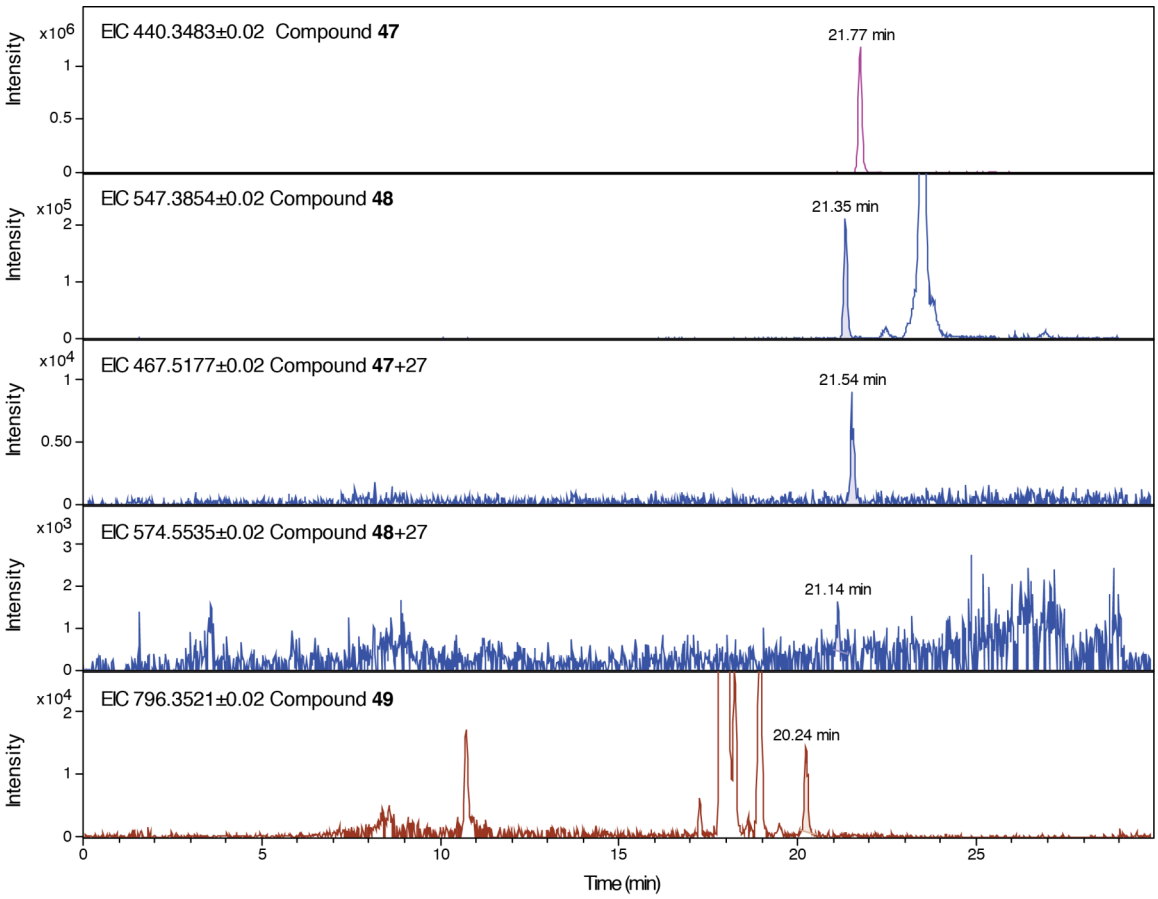


Figure 3.5: EIC corresponding to metabolites **47**-**49** identified using comparative LC–MS. The +27 labeled versions of **47** and **49** were also observed. The peaks labeled with a retention time were unique to BAC*pks*Δ *clbP* extracts and were not found in the BAC*pks* extracts.

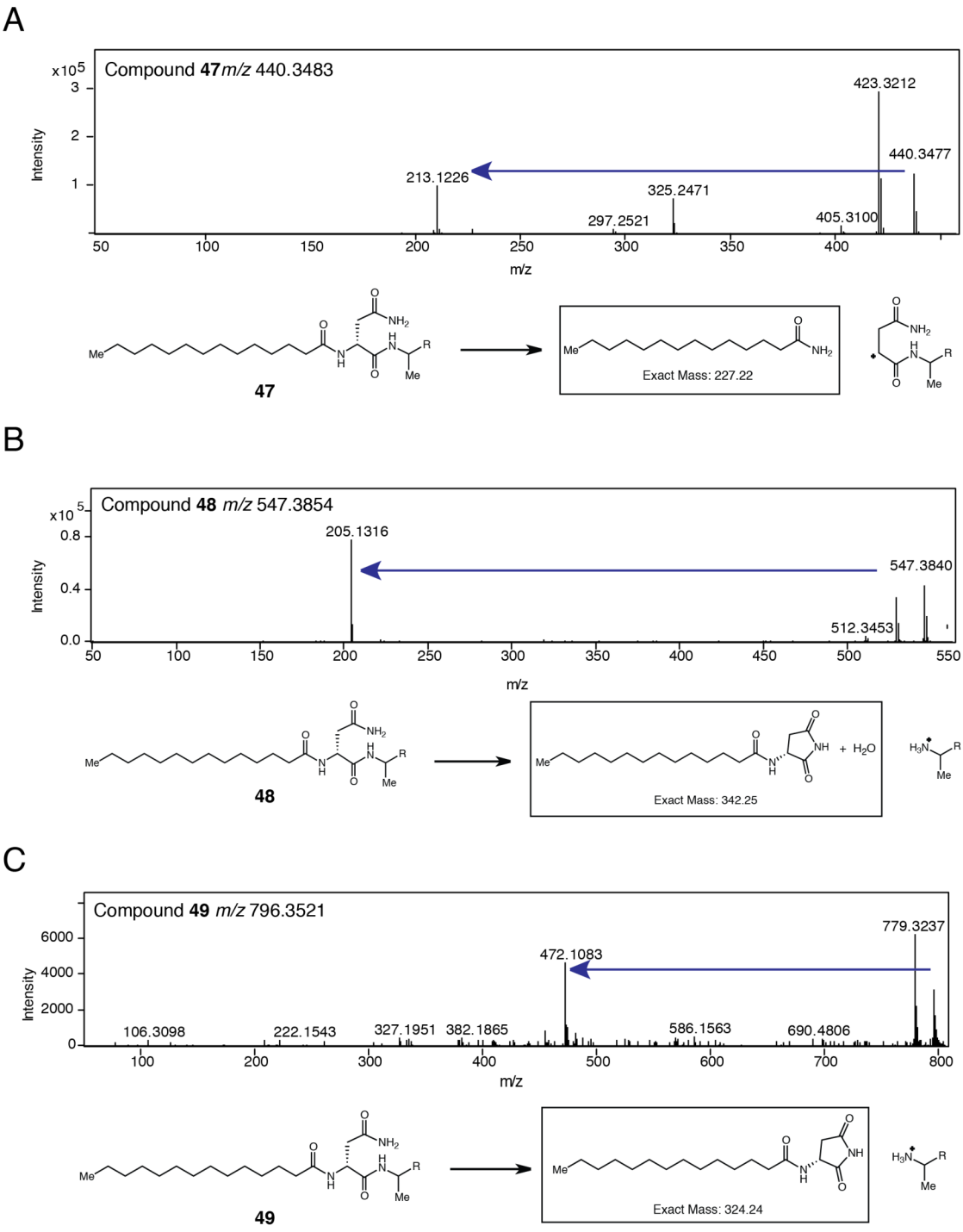
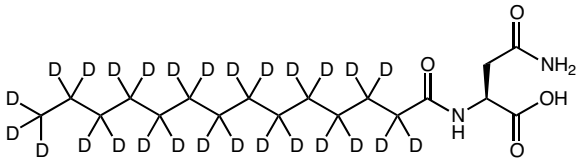


Figure 3.6: The MS/MS fragmentation patterns of metabolites **47**-**49** suggest loss of fragments from the prodrug motif. The proposed fragmentation for each transition (blue arrow) is provided below each spectrum, with the proposed fragments that are lost from the parent molecule outlined in a black box. A) Spectrum for compound **47** (*m/z* 440.3483). The ion collision energy was 15 eV. B) Spectrum for compound **48** (*m/z* 547.3854). The ion collision energy was 15 eV. C) Spectrum for compound **49** (*m/z* 796.3521). The ion collision energy was 25 eV.

3.4: Optimization of precolibactin production

Having identified several candidate precolibactin metabolites (**47**-**49**) using global metabolite profiling, we turned to optimizing production of the candidate precolibactins to facilitate isolation. We used the concentration of the hydrolyzed prodrug motif (**10)** as a proxy for flux through the assembly line pathway. In order to develop a LC–MS/MS method to quantify the concentration of **10** across different growth conditions, we synthesized a deuterated version of the hydrolyzed prodrug motif (**50**) with an L-asparagine residue to serve as an internal standard in these assays (Figure 3.7).



50

Figure 3.7: The structure of the internal standard (**50**) used for prodrug quantitation assays.

First, we measured the concentration of **10** produced by different *pks*$^{+}$ *E. coli* strains, including DH10B expressing either BAC*pks* or BAC*pks*Δ *clbP,* CCR20 (a clinical isolate obtained from the Bonnet lab), and CFT073 (a uropathogenic strain) (Figure 3.8). In this and all subsequent prodrug-quantitation assays, the same experimental protocol was used: cultures (5 mL) were grown at 37 °C for a total of 24 hours and a whole-culture aliquot (500 µL) from each culture was flash frozen, lyophilized, and extracted into methanol (500 µL) containing the internal standard (**50**) for LC–MS analysis. Interestingly, the two strains with chromosomal copies of the *pks* island (*E. coli* CCR20 and CFT073) produced much lower amounts (0.05-0.17 µM) of the hydrolyzed prodrug motif compared to DH10B with a BAC-encoded copy of the *pks* island (1.12 µM).

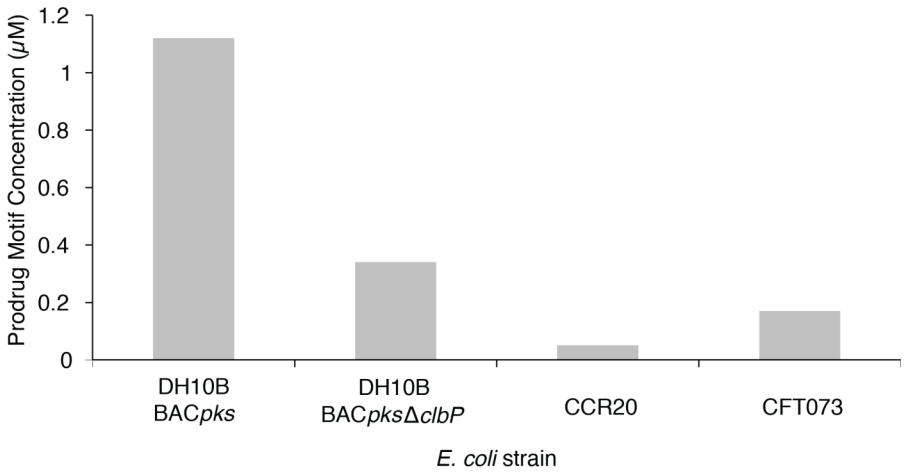


Figure 3.8: The concentration of the prodrug motif **10** in different *pks*$^+$ strains.

We considered whether the putative transcriptional regulator encoded in the *pks* cluster, *clbR*, could promote gene transcription and increase production of colibactin when overexpressed.[10] Bioinformatic analysis suggested that *clbR* encoded for a transcriptional regulator containing a C-terminal helix-turn-helix DNA binding domain with homology to LuxR. To investigate the effects of ClbR overexpression on the production of the prodrug motif, *clbR* was cloned into pTrcHisA (pTrc-ClbR), downstream of the *trc* promoter and *lac* operator. pTrcHisA encodes *lacI*, allowing for induction of overexpression by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG). We found that induction of ClbR expression with high levels of IPTG (500 μM) resulted in cell death and similar levels of prodrug motif as in the strain lacking pTrc-ClbR (Figure 3.9). However, when the amount of IPTG was lowered the concentration of the prodrug motif increased significantly.

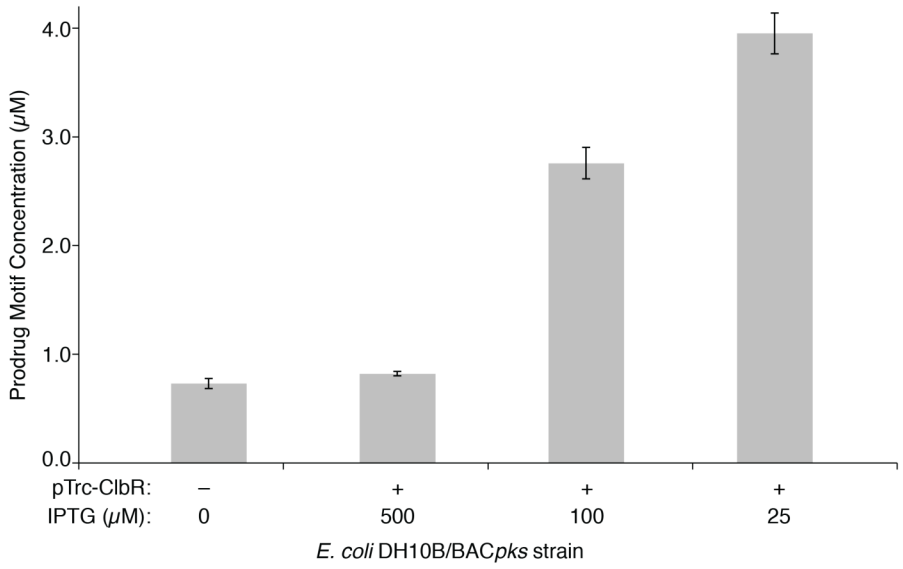


Figure 3.9: The concentration of the prodrug motif (**10**) in DH10B harboring BAC*pks* and pTrc-ClbR, where indicated. The concentration of IPTG is provided. Bar graphs represent the mean ± SD of three independent experiments.

We next examined the effects of growth media on prodrug motif production. A previous report demonstrated that rich media, such as Dulbecco's Modified Eagle Medium (DMEM), resulted in higher levels of transcription of the *pks* genes compared to growth in 3-(*N*-morpholino)propanesulfonic acid (MOPS)-based minimal media.[11] We found that growth in terrific broth (TB) led to a further four-fold increase in the concentration of the prodrug motif compared to growth in LB media (Figure 3.10). Growth in nutrient broth (NB) and DMEM gave similar levels of prodrug motif as growth in LB. The reasons for these effects could be due to the carbon source provided in these different culture medias, for instance, TB contains glycerol as a carbon source, whereas NB and LB lack glycerol and DMEM contains glucose, instead.

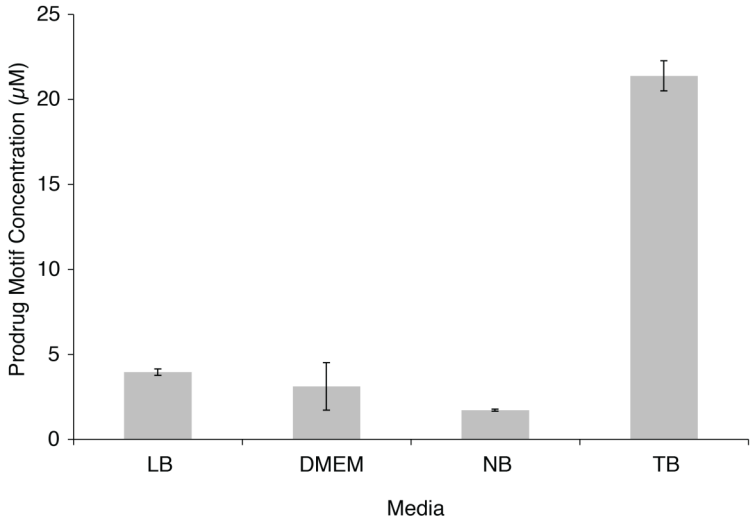


Figure 3.10: The concentration of the prodrug motif (**10**) in DH10B harboring BAC*pks* and pTrc-ClbR with induction by 25 μM IPTG. Bar graphs represent the mean ± SD of two independent experiments.

3.5: The isolation and characterization of Metabolite B

In the LC–MS/MS analysis of compound **48** we discovered that both the unlabeled and the +27 labeled **48** shared a fragment with the same molecular weight (*m/z* 205.1326) (Figure 3.11). From accurate mass measurement we determined that this fragment had a neutral molecular formula of $C_{12}H_{16}N_2O$ and a degree of unsaturation of six. From these data, we hypothesized this fragment could contain a heterocycle. These data suggested that **48** could provide an intriguing target for isolation. We named this candidate precolibactin "Metabolite B".

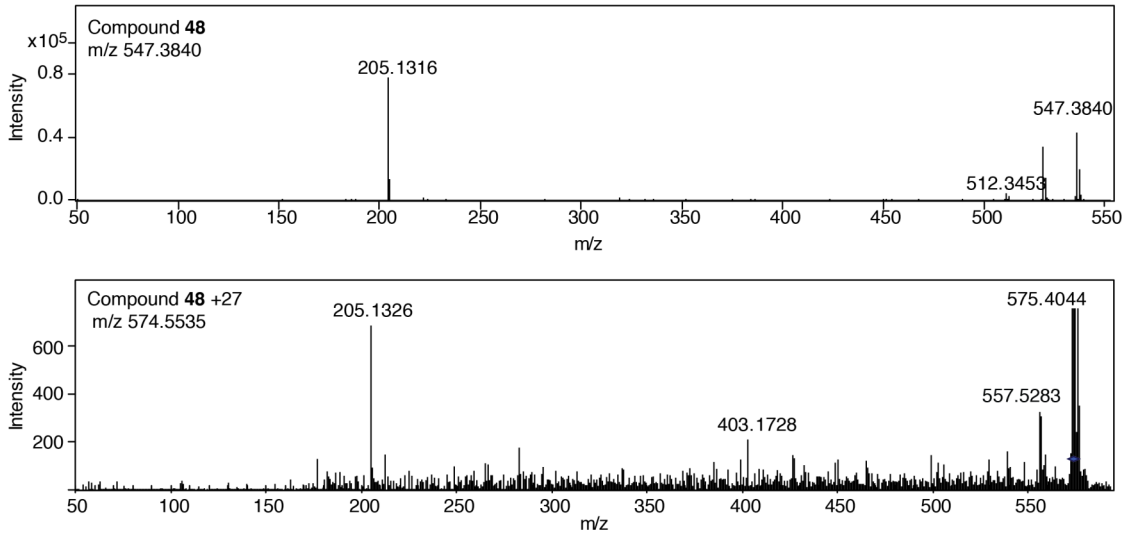


Figure 3.11: The MS/MS fragmentation pattern of **48** (*m/z* 547.3840) (top panel) and the labeled +27 version (*m/z* 574.5535) (bottom panel).

We isolated **48** using the optimized growth conditions described above. *E. coli* DH10B harboring both BAC*pks*Δ *clbP* and pTrcHisA-ClbR was grown in TB media at 37 °C for a total of 24 hours with ClbR overexpression induced by the addition of 25 μM IPTG. From 60 L of culture, 70 g of dried cell mass was obtained. Dr. Matthew Wilson, a post-doctoral researcher in the Balskus lab, assisted in obtaining the necessary quantities of dried cell mass for isolation. Extraction of the dried cell mass with methanol provided 10.6 g of crude extract that was further fractionated using silica gel flash column chromatography. Fractionation of the crude extract was guided by LC–MS. The 1:1 ethyl acetate/methanol fraction from this column was further purified using two rounds of reversed phase HPLC to obtain ~5 mg of **48**.

Complete analysis by both one-dimensional (1D) and two-dimensional (2D) NMR allowed for the complete structural assignment of Metabolite B (**48**) (Figure 3.11 and Table 3.4). The ^{1}H NMR spectrum of **48** in DMSO-d_6 (Figure 3.10) displayed resonances suggesting the presence of five NH protons at δ 8.51 (s), δ 7.88 (d, J = 8.1 Hz), δ 7.48 (d, J = 8.2 Hz), δ 7.24 (s), and δ 6.81 (s); two methines at δ 4.47 (m) and δ 3.71 (m); a methyl singlet at δ 1.96; a methyl doublet at δ 1.00 (J = 7.3

Hz); a methyl triplet at δ 0.85 (J = 7.3 Hz); seven methylenes at δ 2.92 (ddd, J = 6, 8.7, 17.3 Hz), δ 2.82 (ddd, J = 6, 8.7, 17.3 Hz), δ 2.45 (dd, J = 7.7, 15.1 Hz), δ 2.30 (dd, J = 7.7, 15.1 Hz), δ 2.08 (m), δ 1.60 (m), and δ 1.45 (m); and a broad multiplet corresponding to 20 protons at δ 1.30–1.13. ^{13}C NMR analysis in DMSO-d_6 (Figure 3.12) revealed resonances suggesting the presence of a carbonyl at δ 197.7; five amide or electron-deficient sp^2-hybridized carbons at δ 172.1, δ 171.4, δ 170.3, δ 169.6, and δ 169.2; an olefin carbon at δ 128.9; three carbons adjacent to amide nitrogens at δ 49.8, δ 45.5, and δ 44.1; three carbons adjacent to carbonyls at δ 38.4, δ 37.4, and δ 35.2; ten alkyl carbons at δ 31.3, δ 29.87, δ 29.10, δ 29.06, δ 28.99, δ 28.88, δ 28.75, δ 28.65, δ 25.2, and δ 22.1; and four methyl groups or upfield methylenes at δ 20.5, δ 14.0, δ 13.7, and δ 10.8.

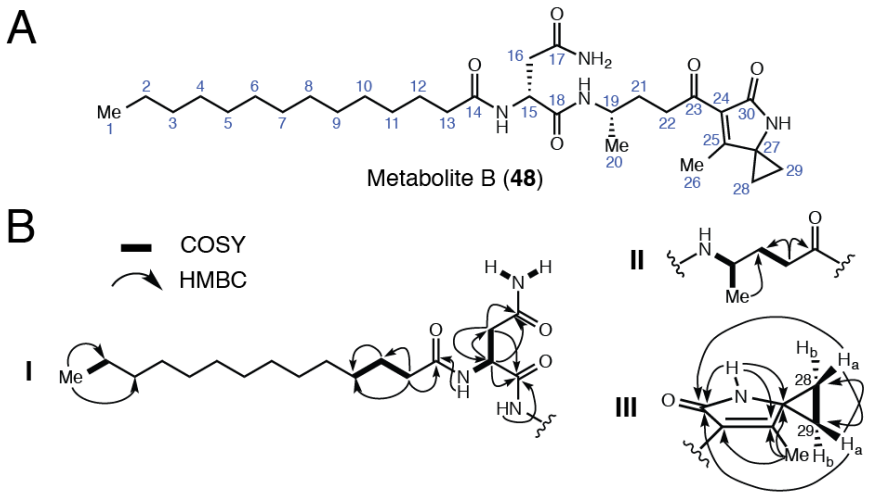


Figure 3.12: Metabolite B (**48**) possesses an unusual spirocyclopropane ring system. A) The deduced structure of **48**. B) Key COSY and HMBC correlations establishing the structure of **48**.

Table 3.4: NMR Data of Metabolite B (**48**) in DMSO-d_6

Fragment	Carbon	δ C (type)	δ H, multiplicity, *J* (Hz)	COSY	HMBC (^{1}H to ^{13}C)	ROESY
I	1	14.0 (CH_3)	0.85, t (7.3)	2	2, 3	
	2	22.1 (CH_2)	1.30–1.13, m	1		
	3	31.3 (CH_2)	1.30–1.13, m			
	4–10	29.10, 29.06, 28.99, 28.88, 28.75 (CH_2)	1.30–1.13, m			
	11	28.65 (CH_2)	1.30–1.13, m	12		
	12	25.2 (CH_2)	1.45, m	11, 13	11	13
	13	35.2 (CH_2)	2.08, m	12	11,12, 14	12, 14NH
	14	172.1 (C)	–	–		
			NH, 7.88, d (8.1)	15	14	13, 16b
	15	49.8 (CH)	4.47, m	14NH, 16a, 16b	16, 17, 18	18NH
	16	37.4 (CH_2)	a 2.45, dd (7.7, 15.1)	15, 16b	15, 17, 18	
			b 2.30, dd (7.7, 15.1)	15, 16a	15, 17, 18	14NH
	17	171.4 (C)	–			
			a NH, 7.24, s	17bNH		
			b NH, 6.81, s	17aNH		
	18	170.3 (C)	–			
			NH, 7.48, d (8.2)	19	18	15, 20, 21
II	19	44.1 (CH)	3.71, m	18NH, 20		20, 21
	20	20.5 (CH_3)	1.00, d (7.3)	19	21	18NH, 19, 21
	21	29.87 (CH_2)	1.60, m	22a, 22b		18NH, 19, 20, 22a, 22b
	22	38.4 (CH_2)	a 2.92, ddd (6, 8.7, 17.3)	21, 22b	21, 23	21
			b 2.82, ddd (6, 8.7, 17.3)	21, 22a	21, 23	21
III	23	197.7 (C)	–			
	24	128.9 (C)	–			
	25	169.2 (C)	–			
	26	10.8 (CH_3)	1.96, s		24, 25, 27	28a, 29a
	27	45.5 (C)	–			
	28	13.7 (CH_2)	a 1.49, m	28b, 29a, 29b	29, 30	26
			b 1.41, m	28a, 29a, 29b		30NH
	29	13.7 (CH_2)	a 1.49, m	28a, 28b, 29b	28, 30	26
			b 1.41, m	28a, 28b, 29a		30NH
	30	169.6 (C)	–			
			NH, 8.51, s		24, 25, 27	28b, 29b

From our analyses, we concluded that **48** is composed of an *N*-myristoyl-D-asparagine residue (**I**, Figure 3.12) joined by an amide bond to a saturated linker derived from L-alanine (**II**, Figure 3.12), which is connected by a carbonyl to an unsaturated heterocycle (**III**, Figure 3.12). We proposed the absolute stereochemistry of **48** based on analogy to prodrug motif **10**, as well as the known biosynthetic activities of ClbN and ClbB. Connectivity between and within fragments **I–III** were established by correlation spectroscopy (gCOSY) and homonuclear multiple bond correlation (gHMBCAD) spectroscopy (Table 3.4 and Figure 3.12). The left-hand portion of **48** (**I** and **II**) resembled those metabolites biosynthesized in our *in vitro* reconstitution of ClbN and ClbB. The right-hand fragment (**III**) appeared to be a novel structure composed of an unusual azaspiro[2.4]heptenone heterocycle.

The spirocyclopropane ring in fragment **III** was the most challenging structural assignment encountered in our analysis. The proton resonances at δ 1.49 (28Ha and 29Ha) and δ 1.41 (28Hb and 29Hb) showed COSY correlations only to each other. According to heteronuclear single quantum coherence spectroscopy (gHSQCAD), these protons were attached to a carbon with the same chemical shift (δ 13.7, 28C and 29C) and the same phase as the methine and methyl carbons. Analysis by distortionless enhancement by polarization transfer (DEPT) ^{13}C NMR unequivocally confirmed this resonance as a methylene (Figure 3.13-3.14). In the ^{1}H-decoupled DEPT spectrum, the carbon resonance at δ 13.7 matched the phase of the other methylene carbons. In the ^{1}H-coupled DEPT spectrum, the peak appeared as a negative triplet with a $^{1}J_{C-H}$ value of 168 Hz. The magnitude of the ^{13}C–^{1}H NMR coupling constant is proportional to the amount of carbon *s* character in the bond. Because cyclopropane rings have a high degree of *p* character in the hybrid orbitals used to form the C–C bonds, the C–H bonds in cyclopropane rings have high *s* character.[12] This results in a large $^{1}J_{C-H}$ value compared to that of most methylene C–H bonds. For instance, the $^{1}J_{C-H}$ value of a cyclohexane C-H bond is 124 Hz.[12]

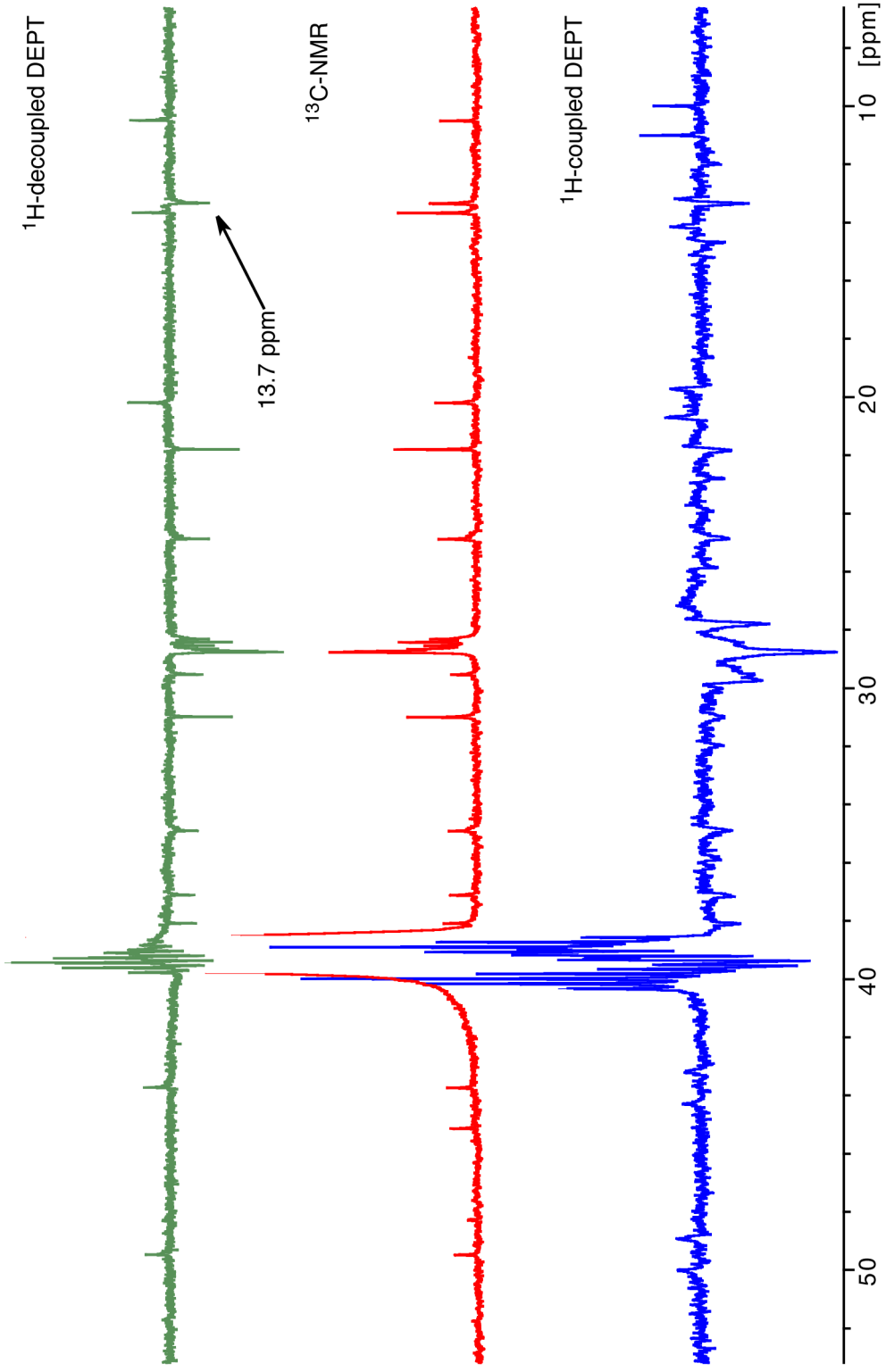


Figure 3.13: ^{1}H-decoupled DEPT, ^{13}C-NMR and ^{1}H-coupled DEPT spectra of **48** overlaid. The resonance of C28 and C29 at 13.7 ppm is indicated. Al sepctra were recorded in DMSO-d_6 at 125 MHx.

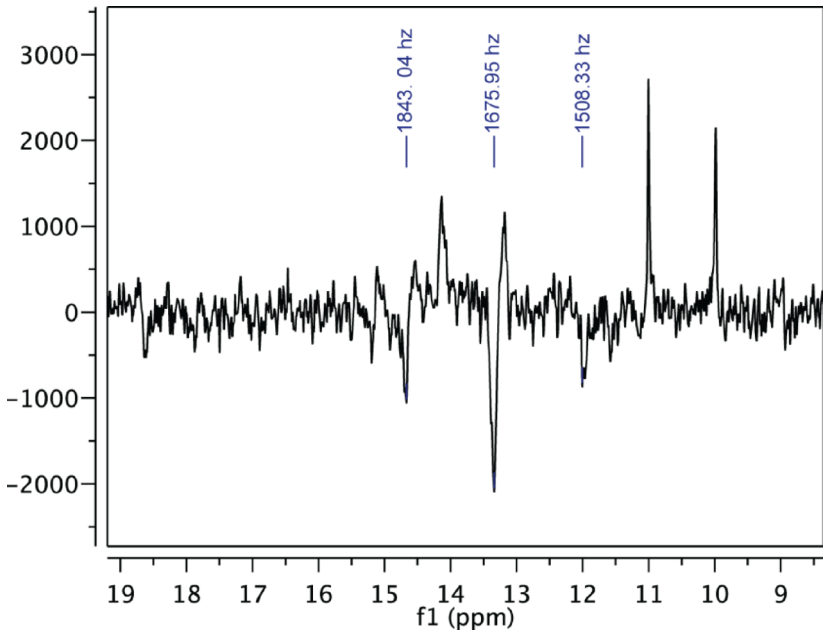
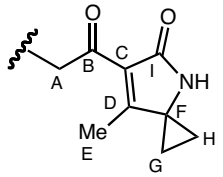


Figure 3.14: The ^{1}H-coupled DEPT spectra of **48** shows a negative triplet centered at 13.7 ppm with a $^{1}J_{C-H}$ value of 168 Hz. This spectrum was recorded in DMSO-d_6 at 125 MHz.

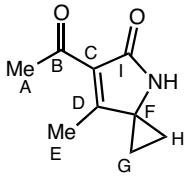
Distinguishing between potential constitutional isomers of fragment **III** was achieved through analysis of NOE correlations and chemical synthesis of a model compound. Nuclear Overhauser effect spectroscopy (ROESYAD) revealed multiple through-space correlations within fragment **III**, including correlations from δ 1.96 (26H) to δ 1.49 (28Ha and 29Ha) as well as correlations from δ 1.41 (28Hb and 29Hb) to δ 8.51 (30NH). These data suggested the lactam nitrogen and the allylic methyl carbon were both adjacent to the spiro carbon (27C). To provide additional support for this structural assignment, a model of fragment **III**, compound **51**, was synthesized by Matt Wilson, a post-doctoral researcher in the Balskus group. The ^{1}H- and ^{13}C-NMR spectra of this molecule closely matched those of fragment **III** (Table 3.5). From this comparison we concluded that **48** and model compound **51** very likely possess related connectivity.

Table 3.5: Comparison of ^{1}H- and ^{13}C-NMR chemical shifts of **48** and **51** recorded in DMSO-d_6.

48, fragment **III**



Compound **51**



Carbon	δ C, 48	δ C, 51	δ H, 48	δ H, 51
A	38.4	30.0	2.92, 2.82	2.43
B	197.7	195.4	–	–
C	169.2	169.4	–	–
D	128.9	129.0	–	–
E	10.8	10.9	1.96	1.98
F	45.5	45.4	–	–
G	13.7	13.8	1.49–1.47 (1.49), 1.41–1.39 (1.41)	1.52–1.49 (1.51), 1.43–1.40 (1.42)
H	13.7	13.8	1.49–1.47 (1.49), 1.41–1.39 (1.41)	1.52–1.49 (1.51), 1.43–1.40 (1.42)
I	169.6	170.0		
			NH, 8.51	NH, 8.52

3.6: ClbP can hydrolytically process Metabolite B

We next tested the ability of peptidase ClbP to cleave **48** in an *in vivo* LC–MS assay. We measured the amount of hydrolyzed prodrug motif (**10**) in culture extracts of ClbP-expressing *E. coli* strains incubated with 100 μM of **48**, 100 μM of known ClbP substrate *N*-myristoyl-D-asparagine-L-alanine-*O*-methyl ester (**29**) or an equal volume of DMSO. *E. coli* expressing full-length ClbP-C-His_6 hydrolyzed both **48** and **29**, as evidenced by the accumulation of **10** (Figure 3.15). In assays with **48**, we did not observe masses corresponding to the expected products, the primary amine **52** or imine **53**. Instead, we observed a product (**54)** that had a mass consistent with a chloride adduct (*m/z* 241.1108) (Figures 3.16-3.17). Neither **10** nor **54** was found in assays with cells expressing the inactive mutant ClbP-S95A-C-His_6 or an empty vector. The ability of ClbP to process Metabolite B (**48**) supports the hypothesis that the spirocyclic ring system is present in

precolibactin and/or other ClbP substrates produced by *pks*$^{+}$ *E. coli*.

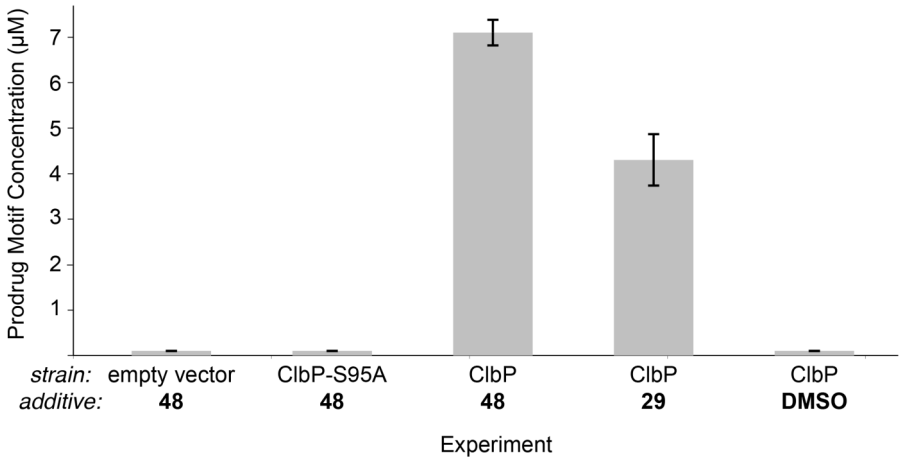


Figure 3.15: ClbP-expressing *E. coli* were incubated with **48** and model substrate **29.** The hydrolyzed product **10** was seen in cultures with ClbP-expressing *E. coli* fed **48** and **29**, but not in the DMSO control. Bar graphs represent the mean ± SD of two independent experiments

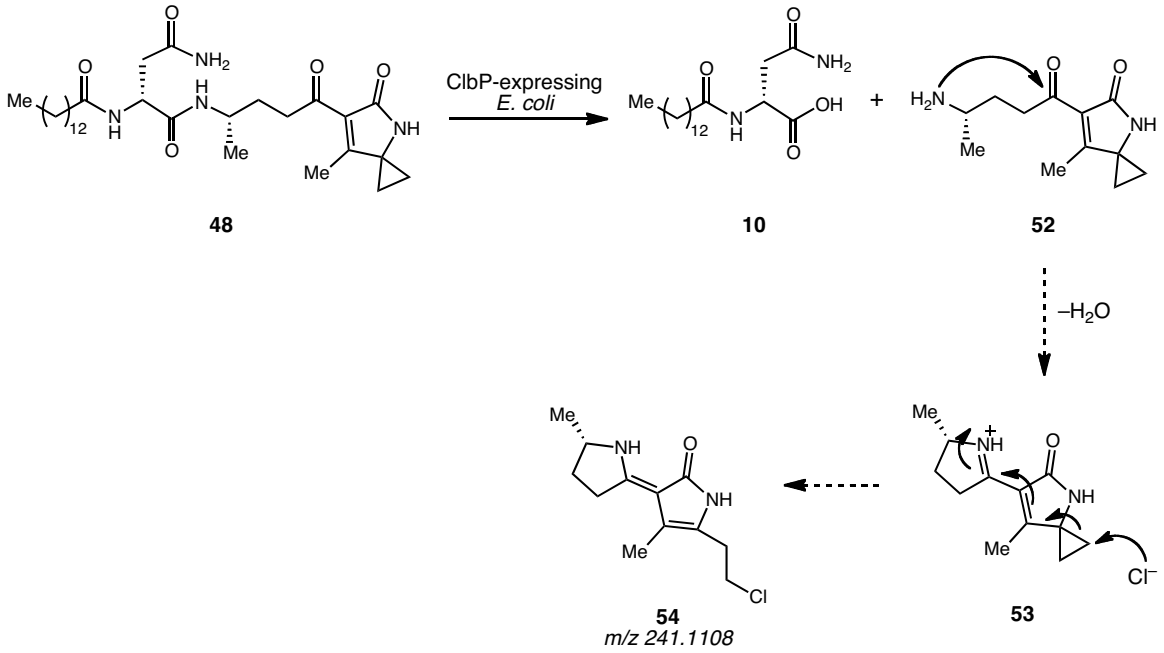


Figure 3.16: ClbP-expressing *E. coli* were incubated with **48.** While expected product **10** was detected, neither **52** nor **53** was seen using LC–MS analysis. However, the mass of a putative chloride adduct **54** was detected. The dotted arrows represent proposed and uncharacterized steps.

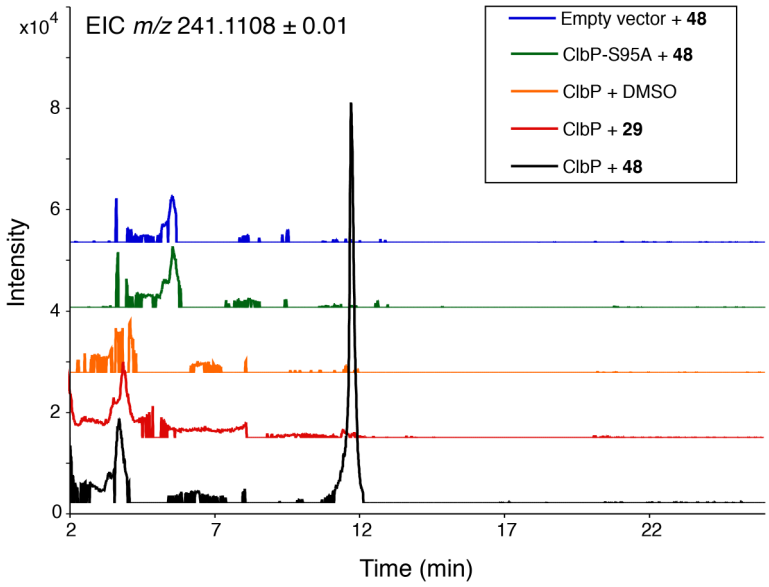


Figure 3.17: In those cultures fed **48** and expressing full-length ClbP, the mass of putative chloride adduct (*m/z* 241.1108) **54** was detected.

3.7: Conclusions

We successfully identified several candidate precolibactin metabolites using a comparative LC–MS approach. The concentration of the hydrolyzed prodrug motif **10** was used as a proxy for flux through the assembly line, and quantitation of **10** allowed for the optimization of precolibactin production. This work demonstrated that rich media and the overexpression of the putative *pks* transcriptional regulator ClbR increased the concentration of **10** by nearly 20-fold compared to the original growth conditions. The isolation and characterization of precolibactin Metabolite B was achieved using optimized growth conditions and extensive NMR analyses. The structure of Metabolite B demonstrates the unusual nature of the *pks* assembly line biochemistry, which will be discussed in more detail in Chapter 4.

In addition to questions regarding its biosynthesis, the structure of Metabolite B and the apparent instability of the expected cleavage product **52** raised many questions about the role of ClbP in

generating active genotoxin and the mechanism by which colibactin may cause DNA damage. We postulated that prodrug cleavage could enhance the activity of **48** and related *pks* metabolites in two ways. First, prodrug cleavage may release metabolites from the *E. coli* membrane through removal of the lipophilic *N*-acyl chain. Second, hydrolysis of the prodrug may generate a potent electrophile, an α,β-unsaturated iminium with an adjacent cyclopropyl group, that could alkylate DNA or a target protein. Iminium ion-formation is an activation strategy seen in both synthetic methodology and enzymatic chemistry; for instance synthetic amine catalysts are used to activate aldehyde- and ketone-containing substrates toward nucleophilic attack in a variety of reaction manifolds and iminium ion formation is a central part of the catalytic mechanism of pyridoxal-5′-phosphate-(PLP)-dependent enzymes, such as histone decarboxylase.[13,14]

Intriguingly, the potential colibactin activation strategy outlined above bears resemblance to, but is unique from, those involved in modulating the activities of other DNA-alkylating natural products that contain cyclopropanes (Figure 3.17).[15] For instance, CC-1065 and the related duocarmycins, such as duocarmycin A (**55**) undergo a conformational change upon DNA binding, which activates the cyclopropane ring toward nucleophilic attack by an adenine base.[16] In addition, (–)-acylfulvene (**56**) undergoes reductive activation within the cytoplasm of eukaryotic cells, providing a highly electrophilic ring system that alkylates DNA.[17] Based on this precedent, we tentatively proposed a cyclopropane-opened structure for compound **54**. Overall, additional studies are needed to better understand the reactivity of proposed products **52**-**54** and how this reactivity relates to the genotoxicity of colibactin *in vivo*.

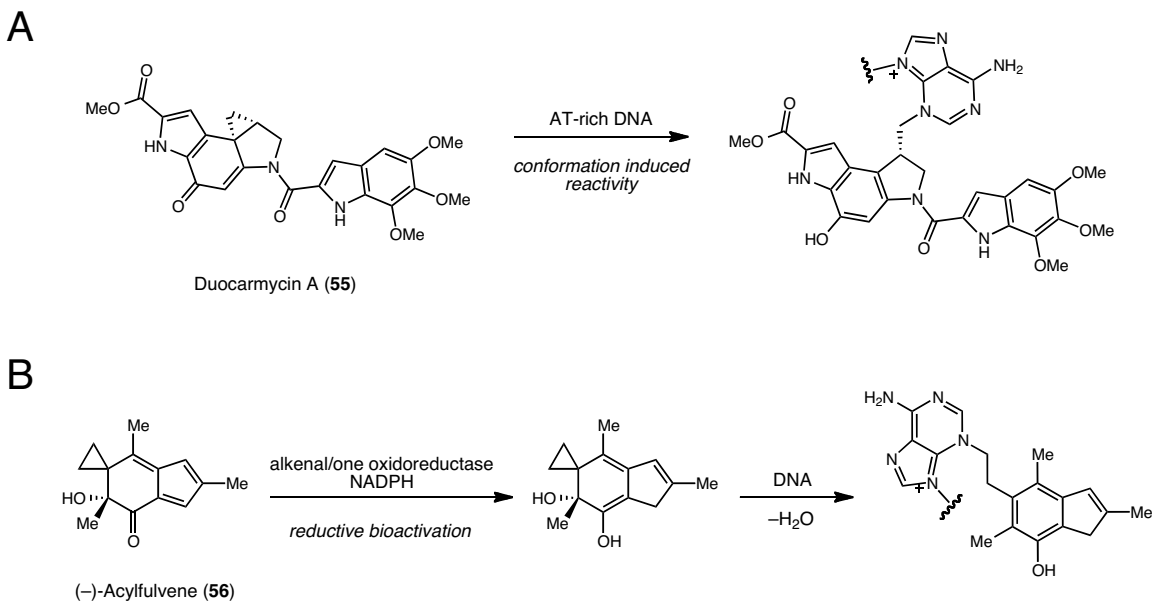


Figure 3.18: A) Duocarmycin A (**55**) undergoes conformation induced activation upon binding the minor-groove of AT-rich DNA sequences.[14] B) (–)-Acylfulvene (**56**) is activated toward attack by DNA through reduction by the cytosolic protein alkenal/one oxidoreductase.[15]

Finally, we wondered whether other *pks*-associated metabolites contained the azaspiro[2.4]-heptenone heterocycle found in Metabolite B, and whether the proposed activation strategy descibed above extended to larger *pks* metabolites. Answering this question would require the isolation of higher-molecular weight metabolites from the *pks* cluster, such as candidate precolibactin **29** identified in our comparative LC–MS studies.

3.8: Experimental section

General materials and methods

Oligonucleotide primers were synthesized by Integrated DNA Technologies (Coralville, IA). Recombinant plasmid DNA was purified with a Qiaprep Kit from Qiagen. Gel extraction of DNA fragments and restriction endonuclease clean up were performed using an Illustra GFX PCR DNA and Gel Band Purification Kit from GE Healthcare. DNA sequencing was performed by Beckman Coulter Genomics (Danvers, MA). Optical densities of *E. coli* cultures were determined

with a DU 730 Life Sciences UV/Vis spectrophotometer (Beckman Coulter) by measuring absorbance at 600 nm. HPLC was performed on a Dionex Ultimate 3000 instrument (Thermo Scientific). All chemicals and solvents were obtained from Sigma-Aldrich except where noted. Nuclear magnetic resonance (NMR) spectroscopy was performed in the Department of Chemistry and Chemical Biology, Harvard University using Agilent DD2 600 and Varian Utility/Inova 500B spectrometers. 1-D ^{13}C-NMR, DEPT-135 (CH,CH_3 positive, CH_2 negative) and ^{1}H-coupled DEPT-135 spectra were acquired at Harvard Medical School on a Bruker AVANCE I spectrometer equipped with a TXO cryoprobe (Bruker-BioSpin Corporation, Billerica, MA) designed for direct observation of ^{13}C. All data were collected at 25 °C. Chemical shifts are reported in parts per million downfield from tetramethylsilane using the solvent resonance as internal standard for ^{1}H ($CDCl_3$ = 7.26 ppm, DMSO-d_6 = 2.50 ppm) and ^{13}C ($CDCl_3$ = 77.25 ppm, DMSO-d_6 = 39.52 ppm). Data are reported as follows: chemical shift, integration multiplicity (s = singlet, d = doublet, t = triplet, m = multiplet, q = quartet, quint = quintet), coupling constant, integration, and assignment. All NMR solvents were purchased from Cambridge Isotope Laboratories. NMR spectra were visualized using MestReNova, version 10.0.0-14411.

High-resolution LC–MS analyses were performed in the Small Molecule Mass Spectrometry Facility at Harvard University. For the Agilent 6210 ESI-TOF, the capillary voltage was set to 3.5 kV and the fragmentor voltage to 100 V, and the drying gas temperature was maintained at 350 °C with a flow rate of 10 L/min and a nebulizer pressure of 45 psi. For the Bruker Maxis Impact q-TOF, the countercurrent drying gas heater was set to 10 L/min and the temperature maintained at 200 °C, the nebulizer pressure was set to 30 psi, the capillary was set to 4000 V in the positive-ion mode and 4500 V in the negative-ion mode, and the system was calibrated internally with a post-run injection of sodium formate solution. For metabolite quantitation, an Agilent 6460 Triple-Quad with a JetStream electrospray source was used. Source drying gas was set to 10 L/min at 350

°C and the nebulizer was set at 25 psi. Sheath gas was set to 11 L/min at 375 °C. The capillary was 3500 V with a nozzle voltage of 500 V. MassHunter quantitation software was used to process data. Strains of DH10B *E. coli* harboring pBelloBAC11-*pks* or pBelloBAC11-*pks*Δ *clbP* were obtained from the Bonnet lab, Laboratoire de Bacteriologie Clinique, Centre Hospitalier de Clermont-Ferrand, Clermont-Ferrand F-63003, France.

Comparative LC–MS

Large scale

Starter cultures of DH10B *E. coli* (50 mL) harboring pBelloBAC11-*pks* (BACpks) or pBelloBAC11-*pks*Δ *clbP* (BAC*pks*Δ *clbP*) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 1 L of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C on a rotary shaker for 24 h. At this point, the cells were harvested by centrifugation (6000 rpm x 15 min at 4 °C), flash frozen in liquid N_2, and lyophilized. The extraction solvent mixture (2:1:0.8 chloroform: methanol: 0.3% (m/v) NaCl (aq)) was added to the cells (70 mL/g dried biomass). The mixture was vortexed at high speed for 20 s and incubated on a nutating mixer at room temperature for 16 h. The mixture was then centrifuged (4000 rpm x 15 min), the supernatant was transferred to a round-bottom flask, and the extract was concentrated to dryness with slight heating to 30 °C. The dried extract was resuspended in 2 mL methanol, centrifuged (13 krpm x 15 min at 4 °C), and 350 μL of the supernatant was trasnsferred to a vial for LC–MS analysis. Experiments were performed in triplicate.

Small scale

Starter cultures of DH10B *E. coli* (5 mL) harboring pBelloBAC11-*pks* (BAC*pks*) or pBelloBAC11-*pks*Δ *clbP* (BAC*pks*Δ *clbP*) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 5 mL of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 2.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C on a rotary shaker for 24 h. At this point, 500 μL aliquots were removed from each culture. The aliquots were flash frozen in liquid N_2 and lyophilized. The lyophilized powder was extracted into 500 μL methanol by vortexing the mixture for 20 s. The samples were then centrifuged (13 krpm x 15 min at 4 °C) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis. Experiments were performed in triplicate.

LC–MS (Tables 3.1-3.2) was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 100% solvent A for 1.5 min, a linear gradient increasing to 100% solvent B over 43.5 min, 100% solvent B for 8 min, followed by re-equilibration in 100% solvent A for 10 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode. The chromatographic datasets were aligned by retention time and mass, and these aligned data were statistically analyzed using the XCMS software (https://xcmsonline.scripps.edu).[3]

d_{27}-myristic acid feeding study optimization

Starter cultures of DH10B *E. coli* (5 mL) harboring pBelloBAC11-*pks* (BAC*pks* were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 50 mL of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with 175 rpm shaking. There were a total of five cultures, and each was identical except for the timing of addition of d_{27}-myristic acid (**45)** (Cambridge Isotope Laboratories)**.** The d_{27}-myristic acid (**45**) was prepared as a 200 mM stock solution in DMSO. For addition of **45** at the beginning of growth (t=0), 500 μM **45** was added to the LB medium before inoculation with the overnight starter culture. For the other cultures, 500 μM **45** was added to cultures at OD_{600} ~0.1, 0.5, and 1.0. One culture was a negative control, to which no **45** was added. Cultures were grown for a total of 24 h. The cells were pelleted by centrifugation (6000 rpm x 15 min at 4 °C), flash frozen in liquid N_2, and lyophilized. methanol (700 μL) was added to the dried cells, and the mixture was vortexed for twenty seconds and incubated on a nutating mixer at room temperature for 2 h. The mixture was then centrifuged (4000 rpm x 15 min), the supernatant was transferred to a microcentrifuge tube and incubated at –20 °C overnight. The suspension was then centrifuged (13 krpm x 15 min at 4 °C), and 300 μL of the supernatant was transferred to a vial for LC–MS/MS analysis.

LC–MS/MS (Figure 3.3) was performed on a Bruker Maxis Impact q-TOF with Agilent 1290 HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 50 mm) with a C18 guard column. The following elution conditions were used for this experiment: 100% solvent A for 2 min, a linear gradient increasing to 100% solvent B over 2 min, 100% solvent B for 10 min, followed by re-equilibration in 100% A for 6 min (solvent A = 95:5 water:methanol + 0.03%

ammonium hydroxide; solvent B = 80:15:5 isopropanol:methanol:water; flow rate 0.25 mL/min for 4 min and increasing to 0.5 mL/min for the remainder of the run). Experiments were performed in negative ion mode.

d_{27}-myristic acid feeding study

Starter cultures of DH10B *E. coli* (50 mL) harboring pBelloBAC11-*pks* (BAC*pks*) or pBelloBAC11-*pks*Δ *clbP* (BAC*pks*Δ *clbP*) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 500 mL of LB medium containing 20 μg/mL chloramphenicol and 500 μM d_{27}-myristic acid (Cambridge Isotope Laboratories). The d_{27}-myristic acid was prepared as a 200 mM stock solution in DMSO. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with 175 rpm shaking for 24 h. At this point, the cells were harvested by centrifugation (6000 rpm x 15 min at 4 °C), flash frozen in liquid N_2, and lyophilized. The extraction solvent mixture (2:1:0.8 Chloroform: methanol: 0.3% (m/v) NaCl (aq)) was added to the cells (70 mL/g dried biomass). The mixture was vortexed at high speed for 20 s and incubated on a nutating mixer at room temperature for 18 h. The mixture was then centrifuged (4000 rpm x 15 min), the supernatant was transferred to a round-bottom flask, and the extract was concentrated to dryness with slight heating to 30 °C. The dried extract was resuspended in methanol (2 x 1 mL), centrifuged (13 krpm x 15 min at 4 °C), and 300 μL of the supernatant was transferred to a vial for LC–MS/MS analysis. The experiment was performed in triplicate.

LC–MS (Table 3.3) was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 100% solvent A for 1.5 min, a linear gradient

increasing to 100% solvent B over 43.5 min, 100% solvent B for 8 min, followed by re-equilibration in 100% solvent A for 10 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode. The chromatographic datasets were aligned by retention time and mass, and these aligned data were statistically analyzed using the XCMS software (https://xcmsonline.scripps.edu).[3]

LC–MS/MS (Figures 3.4, 3.5 and 3.10) was performed on a Bruker Maxis Impact q-TOF with Agilent 1290 HPLC using a Phenomenex Gemini C18 reverse phase column (5 µm, 4.6 x 50 mm) with a C18 guard column. The following elution conditions were used for this experiment: 2% solvent B in solvent A for 2 min, a linear gradient increasing to 90% solvent B in solvent A over 20 min, 90% solvent B in solvent A for 4 min, followed by re-equilibration in 2% solvent B in solvent A for 6 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode and product ion collision energies were varied from 15 to 25 eV.

Cloning of *clbR*

Table 3.6. Oligonucleotides used for cloning *clbR*. Restriction sites are underlined.

Primer Name	Target	Sequence (5' to 3')
ClbR-F	ClbR	AATTCTCGAGATGGGGGGAAACATGG
ClbR-R	ClbR	TGCTAAGCTTTTAGATAATCTCATTTCTG

ClbR was PCR amplified from *E. coli* CFT073 genomic DNA (purchased from the American Type Culture Collection, Manassas, VA) using the primers shown in Table 3.6. PCR reactions contained 25 µL Q5 High-Fidelity 2X Master Mix (New England Biolabs), 1 ng of DNA template, and 500 pmoles of each primer in a total volume of 50 µL. PCR reactions were carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation for 30 s at

98 °C, followed by 35 cycles of 10 s at 98 °C, 30 s at 64 °C, 15 s at 72 °C, and a final extension time of 5 min at 72 °C. PCR reactions were analyzed by agarose gel electrophoresis with ethidium bromide staining, pooled, and purified. Amplified fragments were digested with XhoI and HindIII (New England Biolabs) for 2.5 h at 37 °C. Digests contained 2 μL of water, 6 μL of NEB Buffer 4 (10x), 6 μL of BSA (10x), 40 μL of PCR product, and 3 μL of each restriction enzyme (20,000 U/μL). Restriction digests were purified directly using agarose gel electrophoresis. Gel fragments were further purified using the Illustra GFX kit. The digests were ligated into linearized pTrcHiA (Invitrogen) using T4 DNA ligase (New England Biolabs) to encode a N-terminal His_6-tagged construct. Ligations were incubated at room temperature for 2 h and contained 3 μL of water, 1 μL of T4 Ligase Buffer (10x), 1 μL of digested vector, 3 μL of digested insert DNA, and 2 μL of T4 DNA Ligase (400 U/μL). 5 μL of each ligation was used to transform 50 μL of chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* DH10B cells (Invitrogen) harboring pBelloBAC11-*pks* or pBelloBAC11-*pks*Δ *clbP* and stored at –80 °C as frozen LB/glycerol stocks.

Prodrug quantitation assay

Starter cultures of DH10B *E. coli* (5 mL) harboring pBelloBAC11-*pks* and ptrcHisA-*clbR* or just pBelloBAC11-*pks* were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol and 100 μg/mL ampicillin (if harboring ptrcHisA-ClbR). These saturated cultures were used to inoculate 5 mL of medium (including LB, Terrific Broth (TB), Dulbecco's Modified Eagle Medium (DMEM, Life Technologies), or Nutrient Broth (NB)) containing 20 μg/mL chloramphenicol and 100 μg/mL ampicillin (if harboring ptrcHisA- ClbR). All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were

incubated at 37 °C on a rotary shaker. At an OD_{600} of 0.4-0.5, overexpression of *clbR* was induced by the addition of IPTG, from 25 to 500 μM. The cultures were incubated at 37 °C on a rotary shaker for a total of 24 h. At this point, 500 μL aliquots were removed from each culture. The aliquots were flash frozen in liquid N_2 and lyophilized. The lyophilized powder was extracted into 500 μL methanol containing 1 μM of the synthesized internal standard (**50**) by vortexing the mixture for twenty seconds. The samples were then centrifuged (13 krpm x 15 min at 4 °C) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis. The amount of the prodrug motif was normalized against the internal standard. All experiments were perfomed in triplicate.

LC–MS/MS (Figures 3.7-3.9) was performed on an Agilent 6460 Triple Quad LC–MS with Agilent 1290 Infinity HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 50 mm) with a C18 guard column. The following elution conditions were used for this experiment: 10% solvent B in solvent A for 1 min, a linear gradient increasing to 90% solvent B in solvent A over 4 min, 90% solvent B in solvent A for 2.5 min, followed by re-equilibration in 2% solvent B in solvent A for 2.5 min (solvent A = 95:5 water/methanol + 0.03% ammonium hydroxide; solvent B = 80:15:5 isopropanol/methanol/water; flow rate 0.3 mL/min). Experiments were performed in negative ion mode. The mass spectrometer was operated in multiple reaction monitoring (MRM) mode with a fragmentor voltage of 154 V. The precursor-product ion pairs used in MRM mode were m/z 368→ m/z 253 (internal standard, **50**) (collision energy (CE) = 18 eV) and m/z 341→ m/z 226 (prodrug motif, **10**) (CE = 22 eV).

Isolation of and characterization of Metabolite B (48)

Starter cultures of DH10B *E. coli* (50 mL) harboring pBelloBAC11-*pks*Δ *clbP* and pTrcHisA-ClbR were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol and 100 μg/mL ampicillin. These saturated

cultures were used to inoculate 16 L of Terrific Broth medium containing 20 μg/mL chloramphenicol and 100 μg/mL ampicillin. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with 200 rpm shaking. At an OD_{600} of 0.4-0.5, overexpression of *clbR* was induced by the addition of 25 μM IPTG. The cultures were incubated at 37 °C with 200 rpm shaking for a total growth time of 24 h. The cells were pelleted by centrifugation (6 krpm x 15 min at 4 °C), flash frozen in liquid N_2 and lyophilized. This process was repeated to obtain 60 L of cell culture. The combined dried biomass (69.8 g) was ground into a fine powder and extracted with 3 portions of 1.4 L methanol to afford 10.58 g of crude extract. This extract was split into two equal portions, and these were fractionated by silica gel flash chromatography (190 g Si gel) using a stepwise gradient solvent system of increasing polarity starting from 1:1 hexanes/ethyl acetate and ending with 100% methanol (5 fractions, 1-5). The fraction eluting with 1:1 methanol/ethyl acetate (fraction 4, 770 mg) was separated further using RP-HPLC. The first round of RP-HPLC [Hypersil Gold aQ, 250 x 20 mm, 5 μm, 175 Å pore size, elution with a linear gradient of 65-100% solvent B in solvent A over 30 min, 100% solvent B for 1 min, then re-equilibration with 65% solvent B in solvent A for 10 min (solvent A = H_2O + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 8 mL/min)] yielded a partially purified extract. This sample was subjected to further purification by RP-HPLC [Kromasil 100 C18 250 x 10 mm, 5 μm, elution with a curved gradient 65-78% solvent B in solvent A over 26 min with curve of 5, 78% solvent B in solvent A to 100% B over 2 min with curve of 8, then 100% solvent B for 2 min, then re-equilibration with 65% solvent B in solvent A for 10 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 3 mL/min)] to yield 5 mg of a pure sample of Metabolite B (**48**) (yield 0.08 mg/L). NMR spectra were obtained in 200 μL DMSO-d_6 using a symmetrical NMR tube susceptibility matched to DMSO (Shigemi, Inc). HRMS (ESI): calcd for $C_{30}H_{51}N_4O_5^+$ $[M+H]^+$, 547.3859; found, 547.3851; $[\alpha]_D$ = –16 (*c* 0.1, DMSO-d_6); λ_{max}

(nm): 225, 285; IR (film), cm^{-1}: 2955 (alkyl CH stretch), 2928 (alkyl CH stretch), 1740 (C=O stretch), 1699 (amide C=O stretch), 1677 (C=C stretch).

LC–MS/MS (Figure 3.18) was performed on a Bruker Maxis Impact q-Tof with Agilent 1290 HPLC using a Phenomenex Gemini C18 reverse phase column (3 μm, 2 x 100 mm). The following elution conditions were used for this experiment: 10% solvent B in solvent A for 1 min, a linear gradient increasing to 90% solvent B in solvent A over 7 min, 90% solvent B in solvent A for 4 min, followed by re-equilibration in 2% solvent B in solvent A for 3 min (solvent A = 95:5 water/methanol + 0.03% ammonium hydroxide; solvent B = 80:15:5 isopropanol/methanol/water; flow rate 0.15 mL/min). Experiments were performed in positive ion mode and product ion collision energies were varied from 10 to 40 eV.

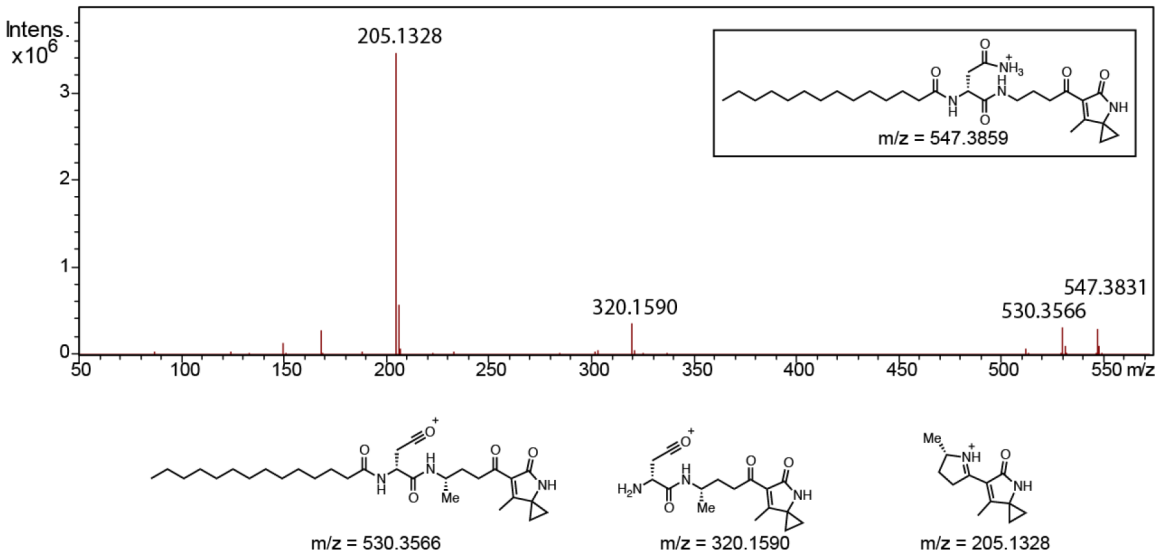


Figure 3.19: LC–MS/MS for the $[M+H]^+$ of **48** (*m/z* 547.3859).

NMR spectra for Metabolite B (**48**)

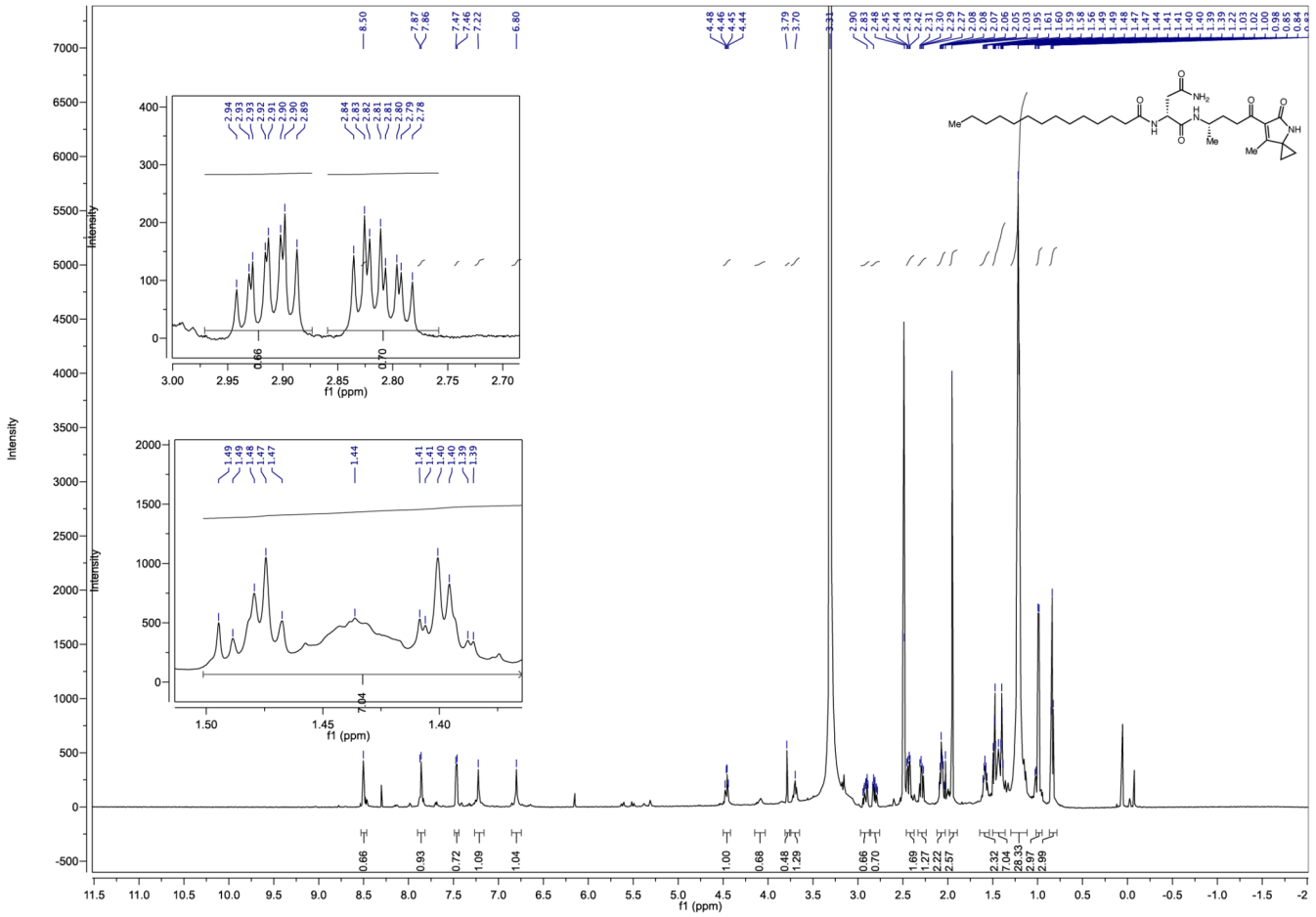


Figure 3.20: ^{1}H-NMR spectrum of 48 (recorded in DMSO-d_6 at 600 MHz).

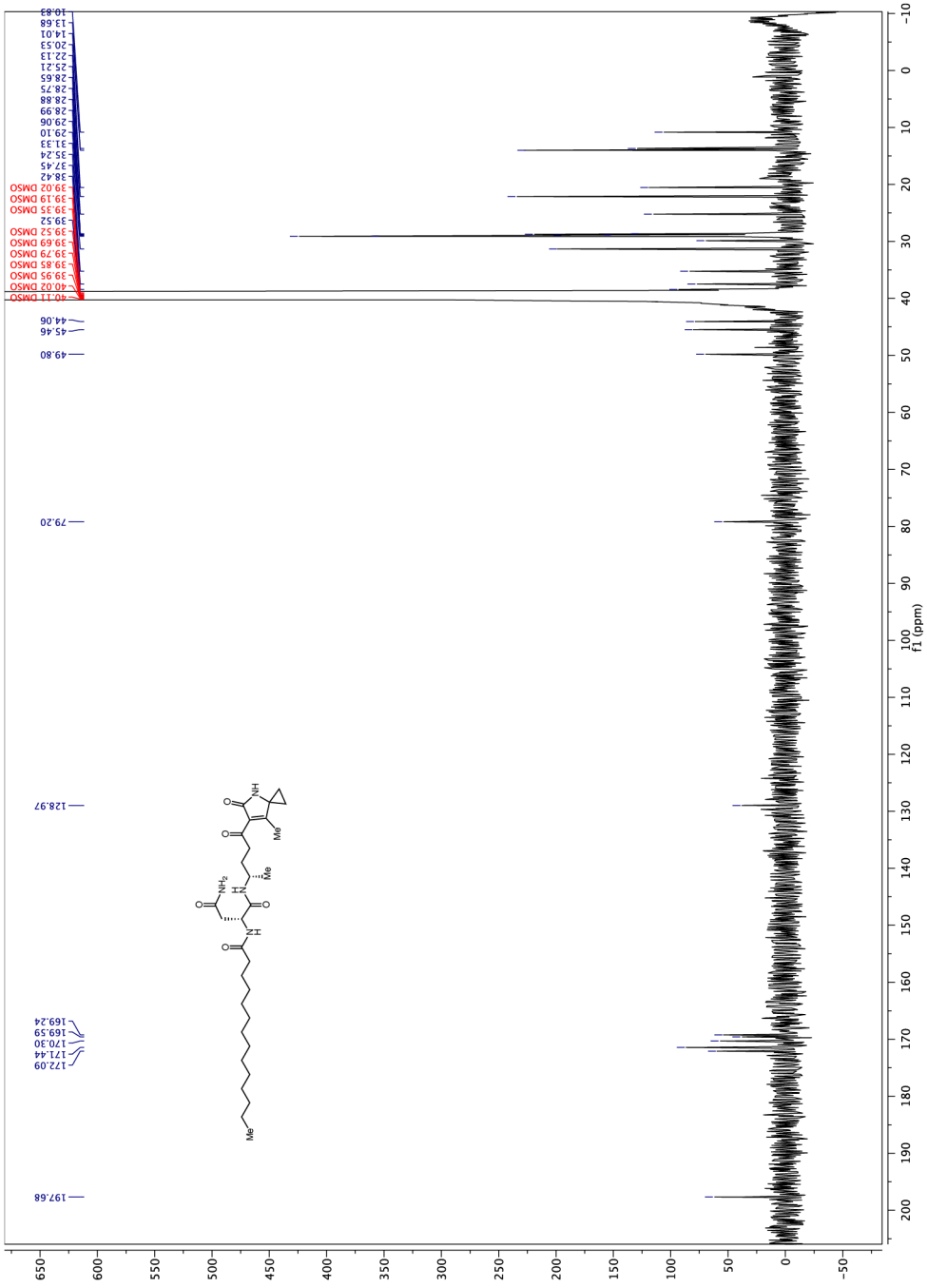


Figure 3.21: ^{13}C-NMR spectrum of 48 (recorded in DMSO-d_6 at 125 MHz).

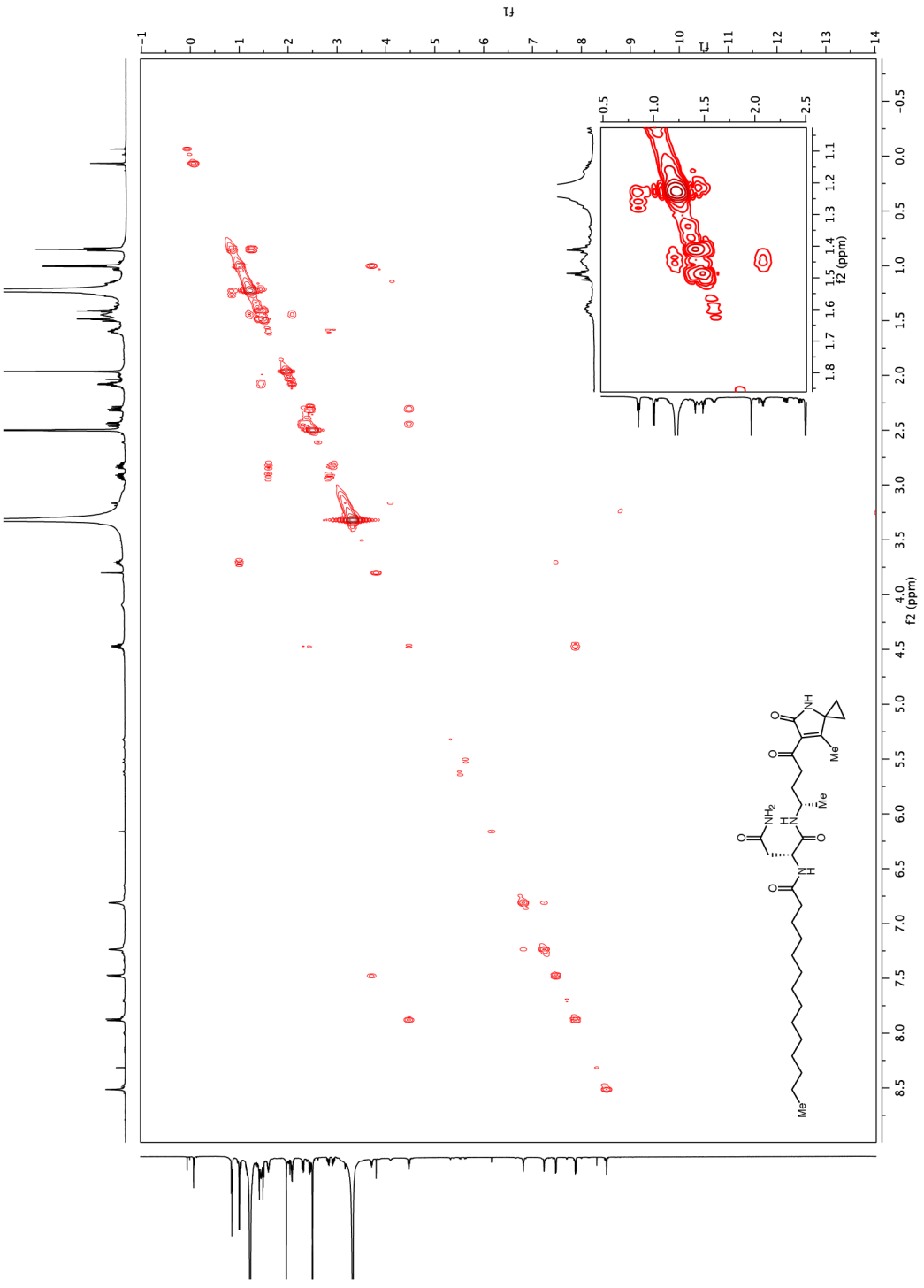


Figure 3.22: gCOSY spectrum of **48** (recorded in DMSO-d_6 at 600 MHz).

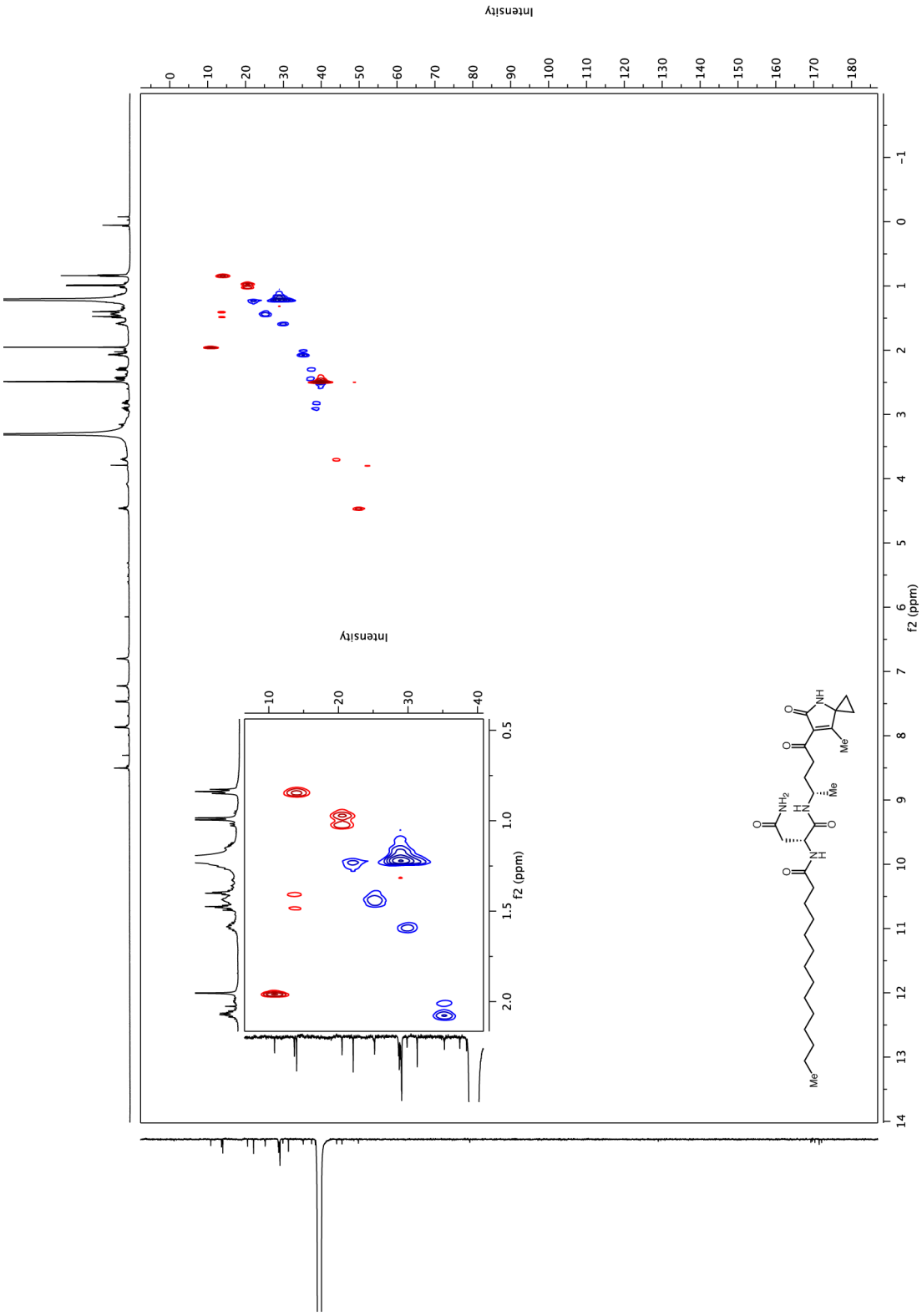


Figure 3.23: gHSQCAD spectrum of **48** (recorded in DMSO-d_6 at 600 MHz).

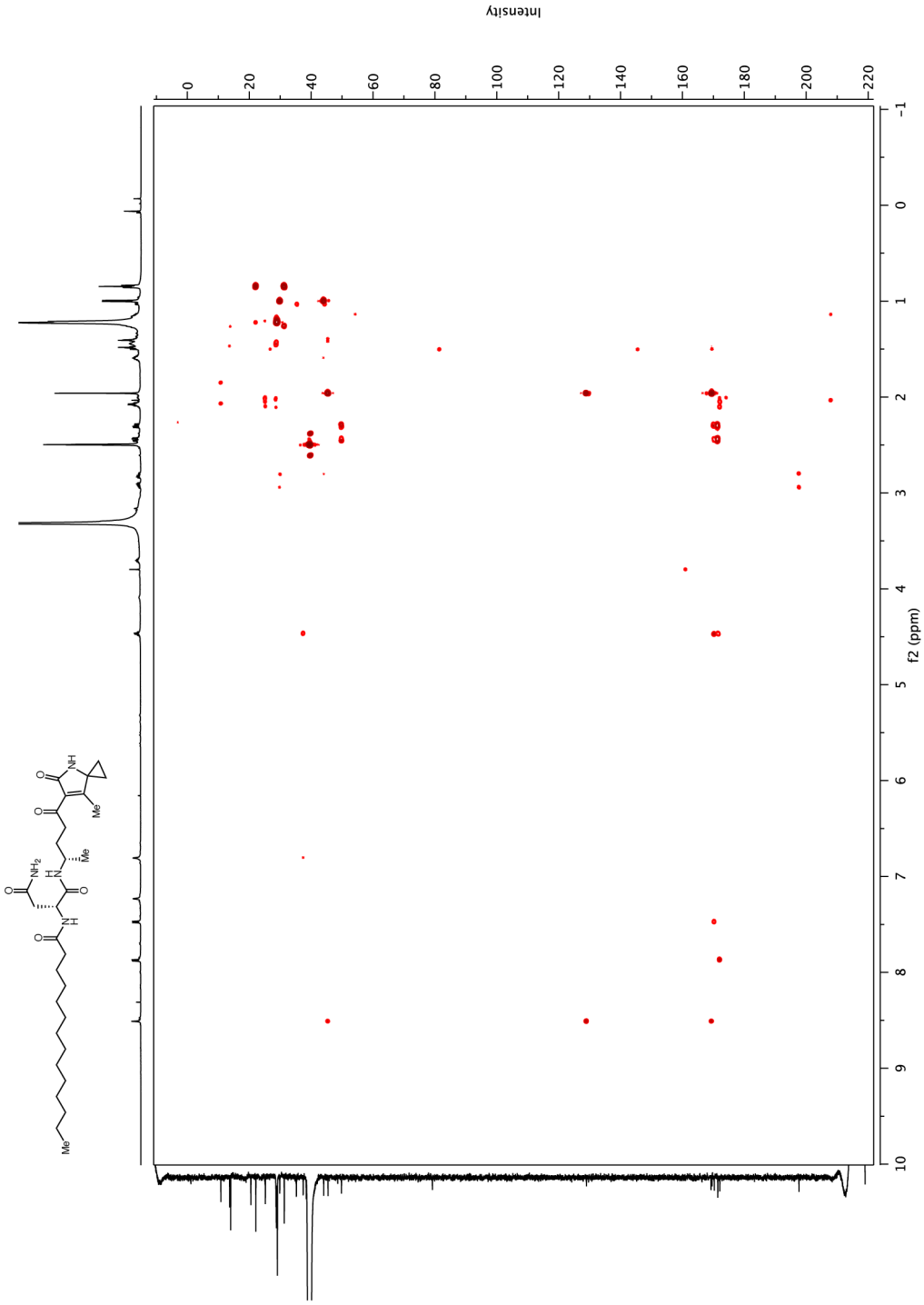


Figure 3.24: gHMBCAD spectrum of **48** (recorded in DMSO-d_6 at 600 MHz).

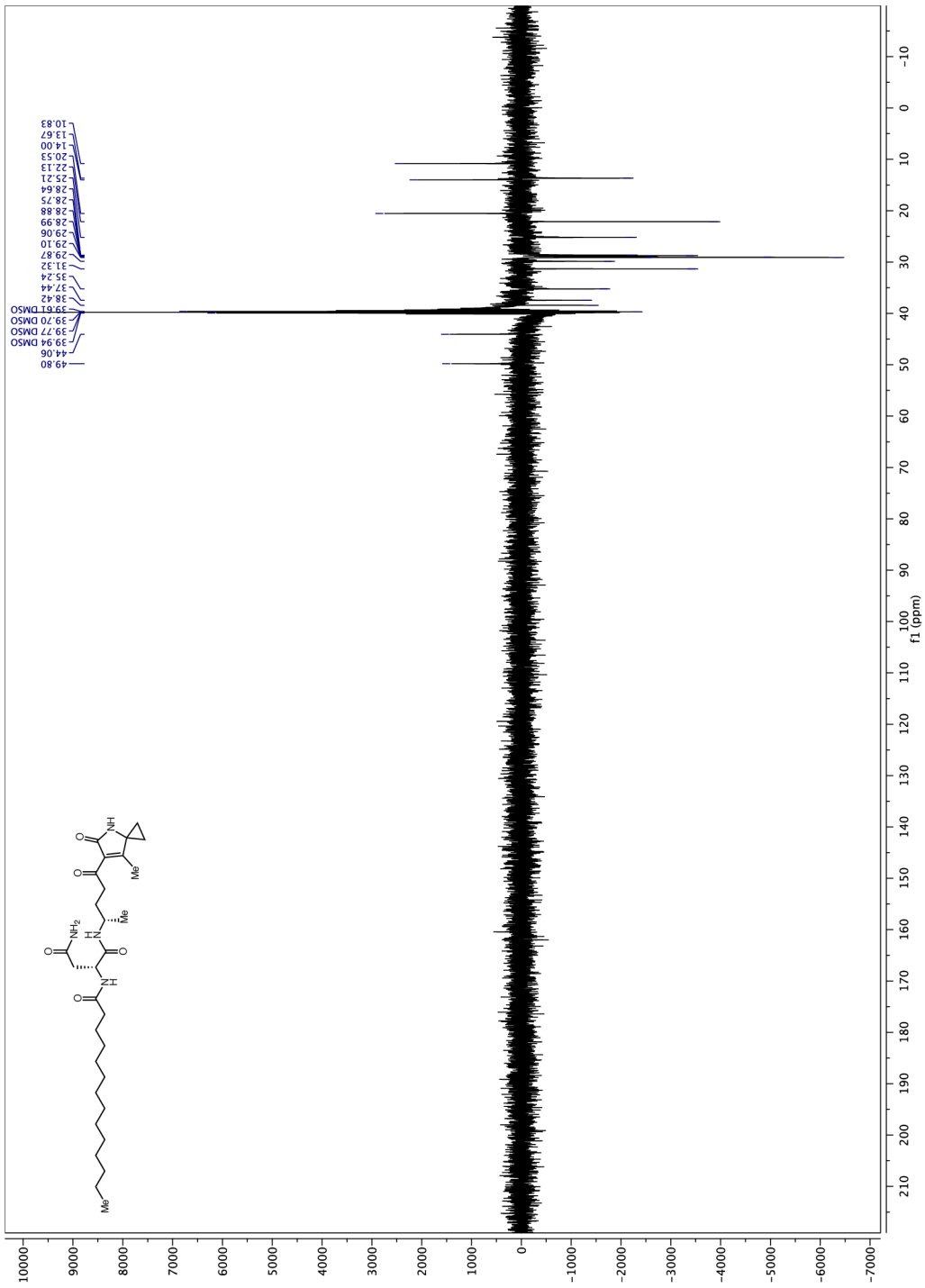


Figure 3.25: DEPT-135 with ^{1}H-decoupling spectrum of **48** (recorded in DMSO-d_6 at 125 MHz).

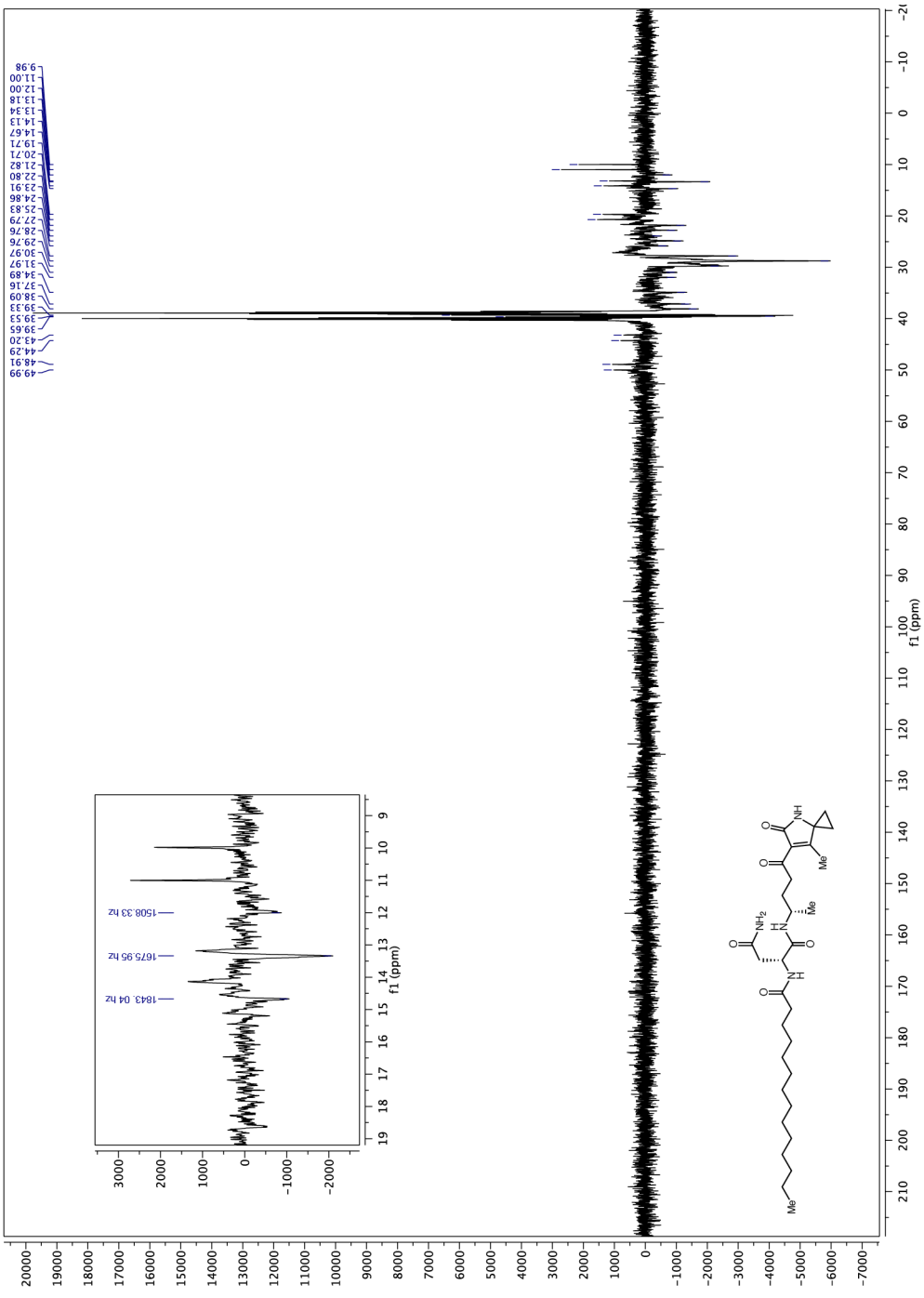


Figure 3.26: DEPT-135 with ^{1}H-coupling spectrum of **48** (recorded in DMSO-d_6 at 125 MHz).

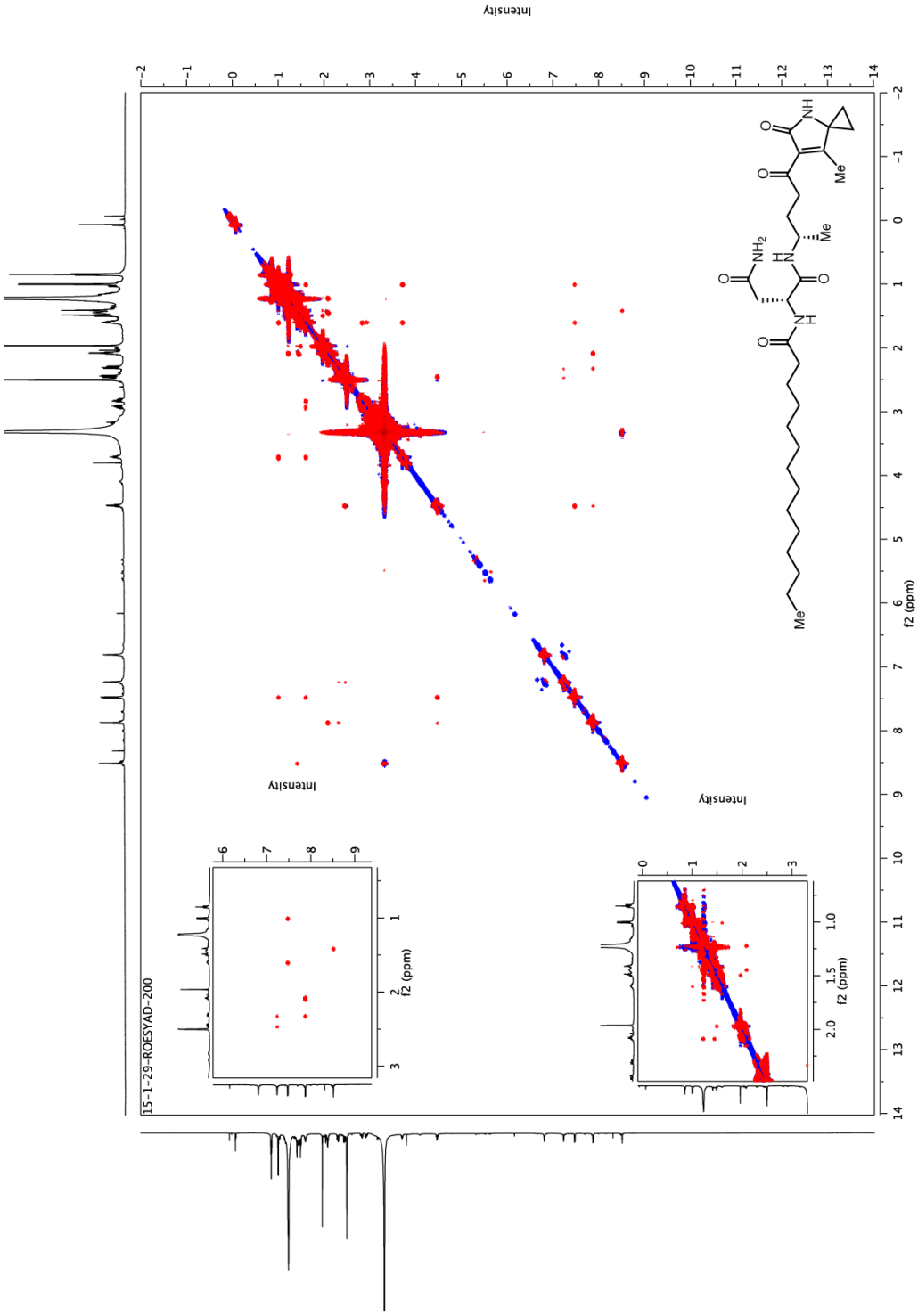


Figure 3.27: ROESYAD spectrum of **48** (recorded in DMSO-d_6 at 600 MHz).

In vivo peptidase assay

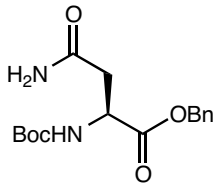
Starter cultures of BL21 *E. coli* (5 mL) harboring pET-29b-clbP-C-His_6, pET-29b-clbP-S95A-C-His_6, or pET-29b with no gene insert (empty vector control) were inoculated from single colonies and grown overnight at 37 °C in LB supplemented with 50 μg/mL kanamycin. These saturated cultures were used to inoculate 4 mL of LB supplemented with 50 μg/mL kanamycin. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C on a rotary shaker. At an OD_{600} of 0.4-0.5, overexpression was induced by the addition of 500 μM IPTG. The cultures were incubated at 37 °C on a rotary shaker for a total of 4.5 h. At this point, the cultures were centrifuged, the supernatant was decanted, and the cells were resuspended in LB medium supplemented with 50 μg/mL kanamycin such that the density of cells was 1 x 10^8 cells/mL. The resuspended cultures were aliquoted into 750 μL in microcentrifuge tubes and either 3 μL DMSO or Metabolite B (**48**) or *N*-myristoyl-D-asparagine-L-alanine-*O*-methyl-ester (**29**) were added to a final concentration of 100 μM. These substrates were prepared as 25 mM stock solutions in DMSO. The cultures were incubated at 37 °C for 1 h. A 500 μL aliquot from each sample was flash frozen in liquid N_2 and lyophilized. The lyophilized samples were extracted into 500 μL methanol containing 1 μM of the internal standard (**50**) by vortexing the mixture for 20 s. The samples were then centrifuged (13 krpm x 15 min) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis. The amount of the prodrug motif was normalized against the internal standard (**32**). All experiments were perfomed in duplicate.

LC–MS/MS (Figure 3.14) was performed on an Agilent 6460 Triple Quad LC–MS with Agilent 1290 Infinity HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 50 mm). The following elution conditions were used for this experiment: 10% solvent B in solvent A for 1 min, a linear gradient increasing to 90% solvent B in solvent A over 4 min, 90% solvent B in solvent

A for 2.5 min, followed by re-equilibration in 2% solvent B in solvent A for 2.5 min (solvent A = 95:5 water/methanol + 0.03% ammonium hydroxide; solvent B = 80:15:5 isopropanol/methanol/water; flow rate 0.3 mL/min). Experiments were performed in negative ion mode. The mass spectrometer was operated in multiple reaction monitoring (MRM) mode with a fragmentor voltage of 154 V. The precursor-product ion pairs used in MRM mode were m/z 368→ m/z 253 (internal standard) (collision energy (CE)= 18 eV) and m/z 341→ m/z 226 (prodrug motif) (CE = 22 eV).

LC–MS (Figure 3.16) was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 2% solvent B in solvent A for 2 min, a linear gradient increasing to 100% solvent B over 10 min, 100% solvent B for 5 min, followed by re-equilibration in 2% solvent B in solvent A for 8 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode.

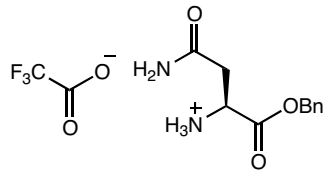
Synthesis of the analyte and internal standards for prodrug quantitation



57

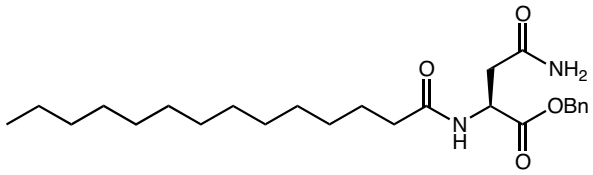
To N_{α}-(*tert*-Butoxycarbonyl)-L-asparagine (Advanced Chem Tech) (2.0 g, 8.64 mmol, 1.0 equiv) was added dry dimethylformamide (Sigma Aldrich) (17.3 mL, 0.5 M) and the reaction mixture was cooled to 0 °C in an ice bath. To this mixture was added Cs_2CO_3 (Sigma Aldrich) (2.82 g, 8.64 mmol, 1.0 equiv) and the reaction mixture was stirred at 0 °C for 45 min. To this mixture was

added benzyl bromide (Sigma Aldrich) (1.02 mL, 8.64 mmol, 1.0 equiv) by syringe and the reaction mixture was stirred at 0 °C for 30 min. The ice bath was then removed and the reaction mixture was stirred at room temperature overnight. The reaction mixture was poured into a separatory funnel containing 100 mL water. The reaction mixture was extracted with ethyl acetate (3 x 33 mL). The organic layers were combined and washed with a saturated aqueous NaCl solution (100 mL). The organic layer was collected and dried over $MgSO_4$, filtered, and concentrated *in vacuo*. The crude product was purified by flash chromatography on silica gel (50% ethyl acetate/hexanes followed by 100% ethyl acetate) to afford compound **57** as a solid (1.711 g, 61%). The ^{1}H-NMR (500 MHz, $CDCl_3$) spectrum matched reported literature values.[18]



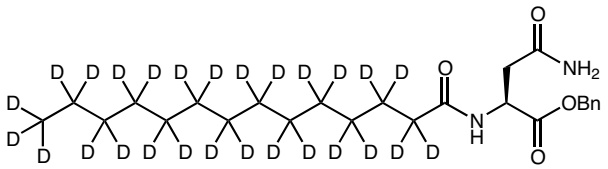
58

To compound **57** (750 mg, 2.33 mmol, 1.0 equiv) was added dry dichloromethane (6.0 mL). Trifluoroacetic acid (Alfa Aesar) (6.0 mL, 78.4 mmol, 33.6 equiv) was added dropwise. The reaction mixture was stirred at room temperature and monitored by thin-layer chromatography. After 45 min, the reaction was complete and the reaction mixture was concentrated *in vacuo* to obtain an oil. This crude product was purified by the addition 1:1 hexanes:diethyl ether (10 mL) with vigorous stirring until a white precipitate formed. The precipitate was collected by filtration and dried *in vacuo* to obtain compound **58** as a white solid (737 mg, 94% yield). TFA salt **58** was used in the next step without further purification. ^{1}H-NMR (500 MHz, DMSO-d_6): δ 8.43 (s, 3H, N**H₃**), 7.72 (s, 1H, C(O)N**H₂**), 7.39-7.38 (m, 5H, Aryl C**H**), 7.26 (s, 1H, C(O)N**H₂**), 5.21 (s, 2H, (C_6H_5)C**H₂**), 4.34 (t, J = 4.9 Hz, 1H, NH_3C**H**), 2.83-2.75 (m, 2H, C**H₂**C(O)NH_2). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 170.34, 168.84, 135.28, 128.41, 128.25, 127.90, 67.02, 48.76, 34.29.



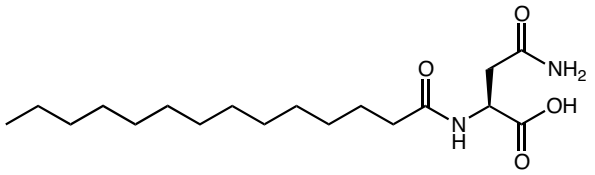
59

To compound **58** (450 mg, 1.33 mmol, 1.0 equiv) was added myristic acid (TCI) (365.4 mg, 1.6 mmol, 1.2 equiv), HOBt monohydrate (TCI) (224 mg, 1.46 mmol, 1.1 equiv), and dichloromethane (6.7 mL, 0.2 M). Triethylamine (Sigma Aldrich) (407 µL, 2.92 mmol, 2.2 equiv) was added by syringe, and the reaction mixture was stirred at room temperature for 5 min. The reaction mixture was then cooled to 0 °C in an ice-water bath, and EDC hydrochloride (Advanced Chem Tech) (280 mg, 1.46 mmol, 1.1 equiv) was added in one portion. The reaction mixture was removed from the ice bath after 30 min and was stirred at room temperature overnight. The reaction mixture was diluted with ethyl acetate (20 mL) and then the reaction mixture was washed with a 1 M aqueous hydrochloric acid solution (30 mL). The organic layer was collected, and the aqueous layer was extracted with ethyl acetate (2 x 20 mL). The organic fractions were combined and were washed with a saturated aqueous NaCl solution (60 mL). The organic layer was dried over $MgSO_4$, filtered, and concentrated *in vacuo* to obtain a white solid. The crude product was purified by flash chromatography on silica gel (gradient from 50% ethyl acetate/hexanes to 100% ethyl acetate). To remove trace HOBt monohydrate that remained after flash chromatography, the product was dissolved in dichloromethane and then washed with a saturated aqueous solution of sodium bicarbonate (2 x 60 mL) and a saturated aqueous NaCl solution (100 mL). The organic layer was dried over $MgSO_4$, filtered, and concentrated *in vacuo* to obtain compound **59** as a white solid (132 mg, 23% yield). The ^{1}H-NMR (500 MHz, $CDCl_3$) spectrum matched the reported literature values.[19]



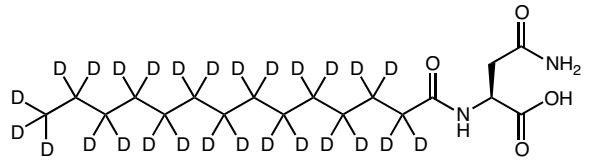
60

To compound **58** (200 mg, 0.595 mmol, 1.0 equiv) was added d_{27}-myristic acid (Cambridge Isotope Laboratories) (167 mg, 0.654 mmol, 1.1 equiv), HOBt monohydrate (TCI) (109.3 mg, 0.714 mmol, 1.1 equiv), and dichloromethane (3.0 mL, 0.2 M). Triethylamine (Sigma Aldrich) (182 µL, 1.31 mmol, 2.2 equiv) was added by syringe, and the reaction mixture was stirred at room temperature for 5 min. The reaction mixture was then cooled to 0 °C in an ice bath, and EDC hydrochloride (Advanced Chem Tech) (137 mg, 0.714 mmol, 1.1 equiv) was added in one portion. The reaction mixture was removed from the ice bath after 30 min and was stirred at room temperature overnight. The reaction mixture was diluted with ethyl acetate (10 mL) and the reaction mixture was washed with a 1 M aqueous hydrochloric acid solution (30 mL). The organic layer was collected, and the aqueous layer was extracted with ethyl acetate (2 x 10 mL). The organic fractions were combined and were washed with water (30 mL), a saturated aqueous $NaHCO_3$ solution (30 mL), and finally a saturated aqueous NaCl solution (30 mL). The organic layer was dried over $MgSO_4$, filtered, and concentrated *in vacuo* to obtain a white solid. The crude product was purified by flash chromatography on silica gel (gradient from 50% ethyl acetate/hexanes to 100% ethyl acetate) to obtain compound **60** as a white solid (164.7 mg, 60% yield). ^{1}H-NMR (500 MHz, $CDCl_3$): δ 7.36-7.32 (m, 5H, Aryl C**H**), 6.80 (d, J = 8 Hz, 1H, C(O)N**H**CH), 5.88 (br s, 1H, C(O)N**H$_2$**), 5.65 (br s, 1H, C(O)N**H$_2$**), 5.17 (s, 2H, (C_6H_5)C**H$_2$**), 4.86-4.82 (m, 1H, NHC**H**), 2.95 (dd, J = 4.7, 16.3 Hz, 1H, C**H$_2$**C(O)NH_2), 2.78 (dd, J = 4.4, 16.3 Hz, 1H, C**H$_2$**C(O)NH_2). ^{13}C-NMR (125 MHz, $CDCl_3$): δ 173.6, 172.5, 171.1, 135.4, 128.69, 128.52, 128.31, 77.4, 77.2, 76.9, 67.6, 48.9, 37.1. HRMS (ESI): calcd for $C_{25}H_{14}D_{27}N_2O_4^+$ $[M+H]^+$, 460.4761; found, 460.4773.



61

To compound **59** (100 mg, 0.23 mmol, 1.0 equiv) was added Pd/C (10 wt% Pd, Sigma Aldrich) (25 mg, 0.023 mmol, 0.1 equiv with respect to Pd), followed by methanol (Sigma Aldrich) (38.3 mL, 0.006 M). A septum was placed on the flask and a H_2 balloon was added. The reaction mixture was stirred vigorously at room temperature overnight. The reaction mixture was poured over Celite®, and the Celite® was then washed with methanol (2 x 10 mL). The eluant was concentrated *in vacuo* to obtain compound **61** as a white solid (72.8 mg, 92% yield). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 7.95 (d, J = 7.9 Hz, 1H, C(O)N**H**CH), 7.36 (s, 1H, C(O)N**H$_2$**), 6.87 (s, 1H, C(O)N**H$_2$**), 4.47-4.43 (m, 1H, NHC**H**), 2.52 (dd, J = 5.7, 15.5 Hz, 1H, C**H$_2$**C(O)NH$_2$), 2.41 (dd, J = 7.2, 15.5 Hz, 1H, C**H$_2$**C(O)NH$_2$), 2.07 (t, J = 7.3 Hz, 2H, C(O)C**H$_2$**CH$_2$), 1.48-1.42 (m, 2H, C(O)CH$_2$C**H$_2$**), 1.28-1.19 (m, 20H, myristoyl -C**H$_2$**), 0.85 (t, J = 6.8 Hz, 3H, CH$_2$C**H$_3$**). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.1, 172.0, 171.4, 48.9, 37.0, 35.1, 31.3, 29.11, 29.10, 29.07, 29.01, 28.89, 28.75, 28.62, 25.3, 22.1, 14.0. HRMS (ESI): calcd for $C_{18}H_{33}N_2O_4^-$ $[M-H]^-$, 341.2446; found, 341.2439.



50

To compound **60** (100 mg, 0.23 mmol, 1.0 equiv) was added Pd/C (10 wt% Pd, Sigma Aldrich) (25 mg, 0.023 mmol, 0.1 equiv with respect to Pd), followed by methanol (Sigma Aldrich) (38.3 mL, 0.006 M). A septum was placed on the flask and a H_2 balloon was added. The reaction mixture was stirred vigorously at room temperature overnight. The reaction mixture was poured over

Celite®, and the Celite® was then washed with methanol (2 x 10 mL). The eluant was concentrated *in vacuo* to obtain compound **50** as a white solid (81.3 mg, 96% yield). ^{1}H-NMR (500 MHz, DMSO-d_6): δ 7.95 (d, J = 7.9 Hz, 1H, C(O)N**H**CH), 7.35 (s, 1H, C(O)N**H₂**), 6.87 (s, 1H, C(O)N**H₂**), 4.48-4.44 (m, 1H, NHC**H**), 2.52 (dd, J = 5.7, 15.5 Hz, 1H, C**H₂**C(O)NH₂), 2.41 (dd, J = 7.2, 15.5 Hz, 1H, C**H₂**C(O)NH₂). ^{13}C-NMR (100 MHz, DMSO-d_6): δ 173.1, 172.1, 171.3, 48.8, 37.0. HRMS (ESI): calcd for $C_{18}H_6D_{27}N_2O_4^-$ [M–H]$^-$, 368.4135; found, 368.4131.

3.9: References

(1) *(a)* Sigal, C. T.; Zhou, W.; Buser, C. A.; McLaughlin, S.; Resh, M. D. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 12253. *(b)* McLaughlin, S.; Aderem, A. *Trends Biochem. Sci.* **1995**, *20*, 272.

(2) Saghatelian, A.; Trauger, S. A.; Want, E. J.; Hawkins, E. G.; Siuzdak, G.; Cravatt, B. F. *Biochemistry* **2004**, *43*, 14332.

(3) *(a)* Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779. *(b)* Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035.

(4) Patti, G. J.; Yanes, O.; Shriver, L. P.; Courade, J.-P.; Tautenhahn, R.; Manchester, M.; Siuzdak, G. *Nat. Chem. Biol.* **2012**, *8*, 232.

(5) Sidebottom, A. M.; Johnson, A. R.; Karty, J. A.; Trader, D. J.; Carlson, E. E. *ACS Chem. Biol.* **2013**, *8*, 2009.

(6) Nougayrede, J.-P.; Homburg, S.; de ric Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J. R.; Dobrindt, U.; Oswald, E. *Science* **2006**, *313*, 848.

(7) Minnikin, D. E.; O'Donnell, A. G.; Goodfellow, M.; Alderson, G.; Athalye, M.; Schaal, A.; Parlett, J. H. *J. Microbiol. Methods* **1984**, *2*, 233.

(8) Alegado, R. A.; Brown, L. W.; Cao, S.; Dermenjian, R. K.; Zuzow, R.; Fairclough, S. R.; Clardy, J.; King, N.; Greenberg, P. *eLife Sciences* **2012**, *1*.

(9) Bligh, E. G.; Dyer, W. J. *Can. J. Biochem. Physiol.* **1959**, *37*, 1.

(10) Krumbholz, G. Ph.D. Thesis, Julius-Maximilians-Universität Würzburg, 2010.

(11) Homburg, S.; Oswald, E.; Hacker, J. R.; Dobrindt, U. *FEMS Microbiol. Lett.* **2007**, *275*, 255.

(12) Anslyn, E.V., Dougherty, D. A. *Modern Physical Organic Chemistry;* University Science Books: California, 2006; p 10.

(13) Lelais, G.; MacMillan, D. *Aldrichimica Acta* **2006**. *39*, 79-87.

(14) Bach, D.R.; Canepa, C. *J. Am. Chem. Soc.*, **1997**, *119*, 11725.

(15) Wolkenberg, S. E.; Boger, D. L. *Chem. Rev.* **2002**, *102*, 2477.

(16) Boger, D. L.; Garbaccio, R. M. *Bioorg. Med. Chem.* **1997**, *5*, 263.

(17) Pietsch, K. E.; Neels, J. F.; Yu, X.; Gong, J.; Sturla, S. J. *Chem. Res. Toxicol.* **2011**, *24*, 2044.

(18) Nicolaou, K. C.; Truhillo, J. I.; Jandeleit, B.; Chibale, K.; Rosenfeld, M.; Diefenbaach, B.; Cheresh, D. A.; Goodman, S. L. *Bioorg. Med. Chem.* **1998,** *6*, 1185.

(19) Vizcaino, M. I.; Engel, P.; Trautman, E.; Crawford, J. M. *J. Am. Chem. Soc.* **2014**, *36,* 9244.

Chapter 4: Toward a complete understanding of the colibactin biosynthetic pathway

4.1: The isolation of additional *pks*-associated metabolites

The structure of Metabolite B (**48**), the identification and isolation of which were described in Chapter 3, could not have been predicted from bioinformatic analyses of the *pks* biosynthetic enzymes. The unique heterocyclic motif of Metabolite B (**48**) prompted us to ask whether higher molecular weight metabolites also contained this motif. In addition, we were also motivated to characterize the biosynthesis of **48**, as we believed that the construction of this and related metabolites would involve non-canonical assembly line biochemistry.

Before describing our efforts to identify higher molecular weight metabolites and further characterize the colibactin biosynthetic pathway, it is necessary to review the contributions of several other groups to the isolation of colibactin. Since we began the colibactin project, multiple research groups have published their work toward identifying, isolating and characterizing *pks*-associated metabolites. In general, these other groups also targeted candidate precolibactins produced by various strains of Δ*clbP pks*$^+$ *E. coli* for isolation. During our work toward the isolation of **48**, several *pks*-associated metabolites (**10, 62** and **47)** were disclosed by the Crawford group (Figure 4.1A).[1] Their approach involved the organic extraction of the metabolomes of both wild-type and Δ*clbP pks*$^+$ *E. coli*, including DH10B expressing the BAC*pks* and the probiotic strain *E. coli* Nissle 1917. A metabolomics analysis was performed to compare metabolites from different culture extracts and to create a network of related metabolites that shared MS/MS fragmentation profiles. One of the isolated compounds, Metabolite A (**47**), was also identified in our comparative LC–MS studies. Metabolite A contains the prodrug motif and shares the 4-aminopentanoic acid linker with Metabolite B. Metabolite **10** is the hydrolyzed prodrug motif containing the expected myristoyl acyl chain. Compound **62** was also proposed to

be a hydrolyzed prodrug motif, containing a fatty acyl chain derived from the β-oxidation of palmitoleic acid, a common unsaturated fatty acid in *E. coli*.

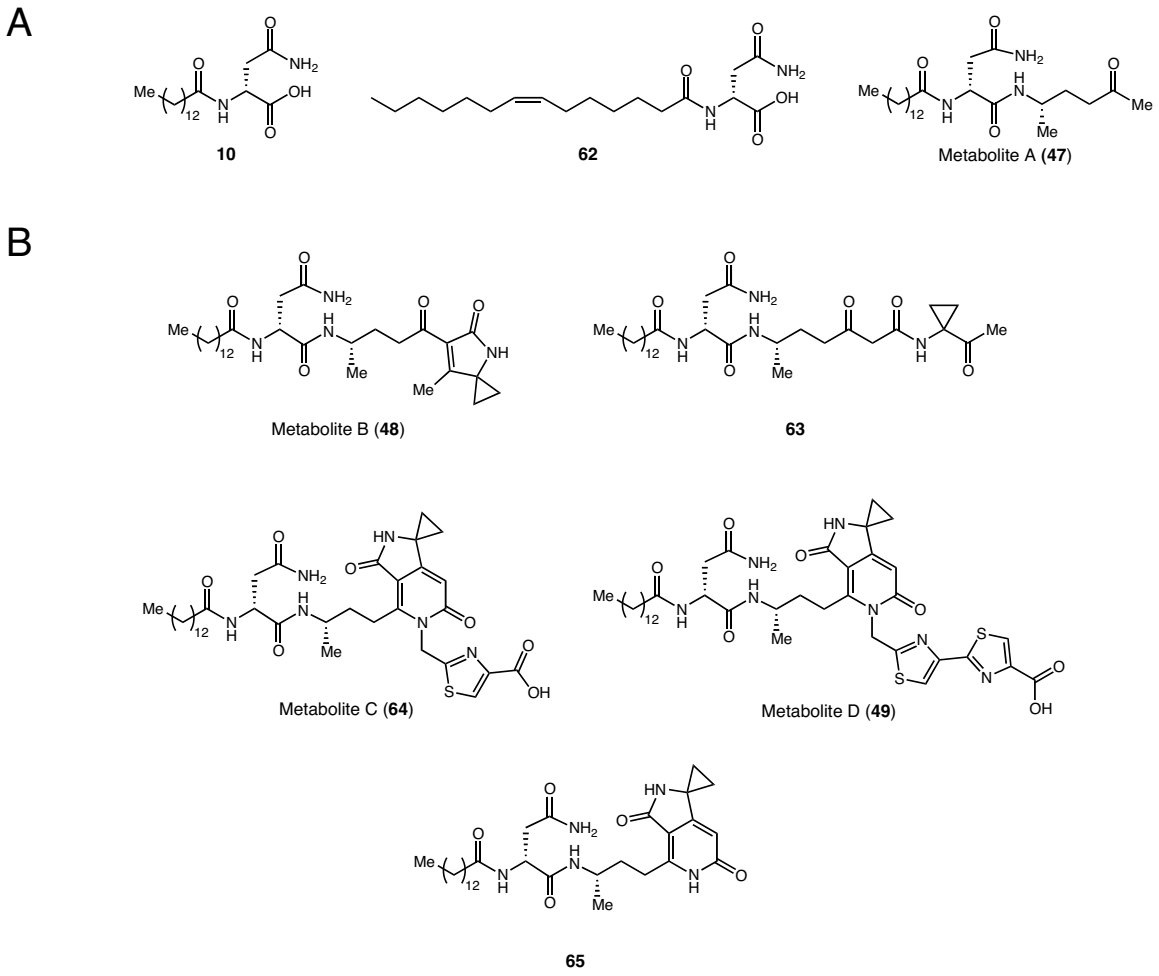


Figure 4.1: A) Compounds **10**, **62** and **47** were reported by Crawford and co-workers.[1] B) Metabolite B (**48**), C (**63**), D (**49**) and compounds **64** and **65** are related structures.[2-6]

In addition, the Müller, Crawford and Qian groups[2,3,4] reported the isolation of Metabolite B (**31**) soon after the publication of our own work describing this same molecule (Figure 4.1B).[5] Furthermore, after the publication of our own work concerning Metabolite B, we turned toward the isolation of higher molecular weight metabolites identified in our comparative LC–MS studies. The isolation and structural elucidation of both Metabolite C (**64**) and D (**49**), the details

of which are described below, was begun by myself and completed by Dr. Matthew Wilson, a post-doctoral researcher in the Balskus group. Soon after we had elucidated the structures of these metabolites, Qian and co-workers reported the structures of compounds **63-65**. In addition, the structure of Metabolite D (**49**) was proposed based on MS/MS fragmentation data.[4] Subsequent to this publication from Qian *et al,* Dr. Wilson isolated sufficient quantities of **49** to perform detailed NMR experiments and confirm the proposed structure.[6] Interestingly, all of the higher molecular weight compounds identified to date contain the azaspiro[2.4]heptenone heterocycle found in Metabolite B (**48**).

In addition to Metabolite B (**48**), Crawford and co-workers also proposed the structure of **66**, which we have named Metabolite E (Figure 4.2A). Despite multiple attempts, they could not isolate this molecule as it reproducibly degraded in organic extracts. The proposed structure of **66** was based on LC–MS analysis of metabolites from Δ*clbP* strains, MS/MS fragmentation data and feeding studies with labeled amino acids.[3] It was hypothesized that Metabolite E (**66**) is an advanced precolibactin intermediate and could represent the prodrug-containing precursor to the active colibactin. In addition to the identification of *pks* metabolites, they also performed *in vitro* DNA alkylation studies using isolated Metabolite B (**48)** and plasmid DNA. At high concentrations (0.5-1.0 mM) of **48**, *Eco*RI-linearized plasmid DNA underwent a gel shift in the presence and absence of reducing agents. They proposed that this gel-shifted DNA species corresponded to the production of interstrand cross-links with **48**, but this species was not characterized. The results of this *in vitro* assay with plasmid DNA led to the hypothesis that the azaspiro[2.4]heptenone heterocycle would serve as the "warhead" of colibactin and that the additional heterocycles found in Metabolite E (**66**) would increase the affinity of colibactin toward DNA. In addition, the authors proposed a mechanism for DNA alkylation by colibactin (Figure 4.2B). This proposal reflects what is known about the structure-activity relationships of

small molecules that cause DNA damage (Figure 4.3). For example, CC-1065 and the related duocarmycins, such as duocarmycin A (**55**), alkylate DNA through a cyclopropane ring, which is activated toward nucleophilic attack by an adenine base upon DNA binding.[7] Studies on the DNA-damaging agent bleomycin (**67**) have shown that the presence of both thiazole rings are important for DNA binding and site-selectivity.[8]

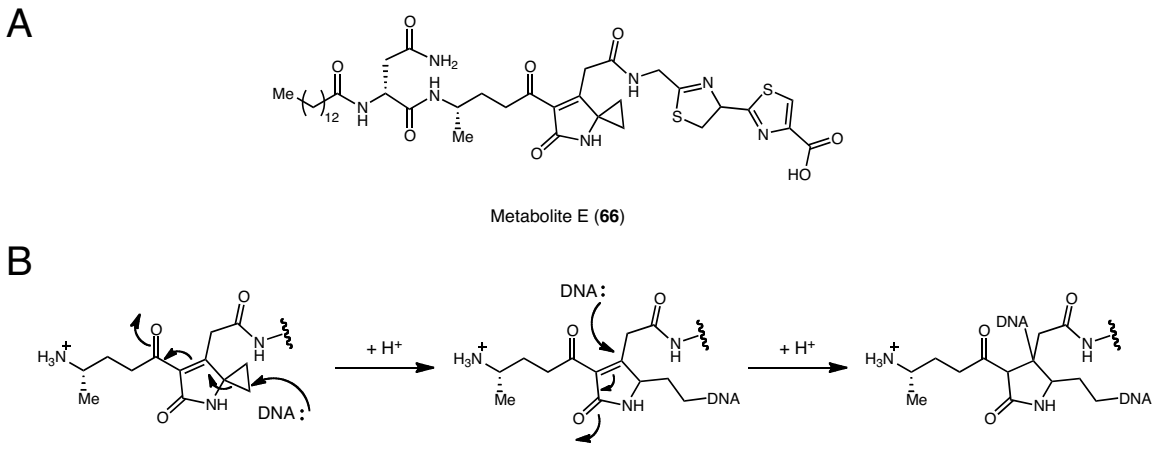


Figure 4.2: A) Metabolite E (**66**) was proposed as the precursor to the active colibactin by Crawford and co-workers. [3] B) A general mechanism for the cross-linking of DNA by metabolites containing the azaspiro[2.4]heptenone moiety found in the *pks*-associated metabolites described above was also proposed.[3]

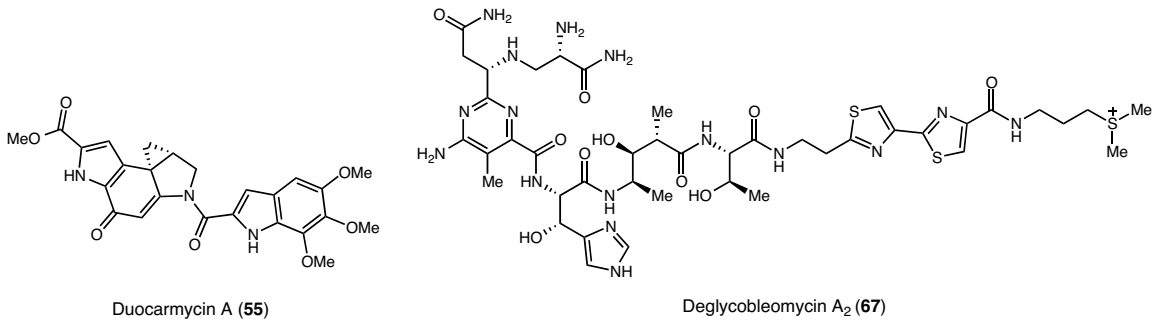


Figure 4.3: DNA-damaging small molecules Duocarmycin A (**55**) and Deglycobleomycin A_2 (**67**).

While the structure of the candidate precolibactin (**66**) proposed by Crawford and co-workers is compelling, other data suggest that the active colibactin or its corresponding precolibactin have yet to be identified. The most important piece of evidence is that provided by gene knock-out

studies: all of the biosynthetic genes, except for *clbM* and *clbS*, are required for genotoxicity.[9] These unaccounted for biosynthetic genes include aminomalonate forming enzymes *clbD, clbE, clbF* and *clbG*, as well as the *trans*-AT PKS *clbO*, thioesterase *clbQ* and predicted amidotransferase *clbL*. Recently, Li Zha, a graduate student in the Balskus lab, characterized the transfer of aminomalonyl-ACP to the *trans*-acting AT ClbG, as well as transfer of aminomalonyl from ClbG to the ACP domains of ClbC, ClbO and ClbK.[6] His work provides evidence that aminomalonate is used multiple times in the colibactin pathway, and supports the idea that additional *pks*-associated metabolites have yet to be discovered.

In addition to questions concerning the structure of the active genotoxin, the mechanism of action of colibactin remains uncharacterized. However, a recent paper from Nougayrede and co-workers that examined the role of *clbS* supported the hypothesis that colibactin directly damages DNA.[10] Previously, the ClbS protein had been shown to react with electrophilic α-alkylidene-γ-butyrolactones through a cysteine residue, and *clbS* knock-out strains still displayed genotoxicity toward human cells.[9, 11] Nougayrede *et al.* observed that Δ*clbS pks*$^+$ *E. coli* cells activated the SOS response and stopped dividing, and that a Δ*clbS* strain that also lacked the DNA repair enzyme *uvrB* displayed severe autotoxicity. Furthermore, it was found that that expression of ClbS in HeLa cells blocked the genotoxicity of *pks*$^+$ *E. coli.* These data demonstrated that without expression of ClbS, *pks*$^+$ *E. coli* are susceptible to the DNA-damaging effects of colibactin. These data suggested that ClbS is involved in self-resistance and that colibactin directly damages DNA. However, the precise mechanism through which colibactin causes DNA damage is still unknown.

4.2: Biosynthetic hypothesis for Metabolites A and B

The intriguing structure of Metabolite B (**48**), as well as the presence of the spirocyclopropane in higher molecule weight *pks* metabolites (**49, 63-66**) inspired us to seek a detailed understanding

of how the colibactin biosynthetic machinery constructs this metabolite. A general biosynthetic hypothesis is provided below (Figure 4.4). The first structure shown (**68)** is an enzyme-bound intermediate that would arise from elongation of the prodrug motif by incorporation of L-alanine followed by incorporation and reduction of malonyl-CoA. Another unit of malonyl-CoA is incorporated into intermediate **68** to form a β-ketothioester intermediate (**69**). Hydrolysis and decarboxylation of **69** would yield Metabolite A (**47**). Metabolite B (**48**) would arise from amide bond formation between **69** and T domain-bound cyclopropane-containing amino acid (**70**) to form **71**, followed by incorporation of another unit of malonyl-CoA to form **72**. Cyclization of **72** would occur spontaneously, either while still bound to an assembly line enzyme or as a free intermediate. Non-enzymatic hydrolysis and decarboxylation of T domain-bound intermediate **73** would provide **48**.

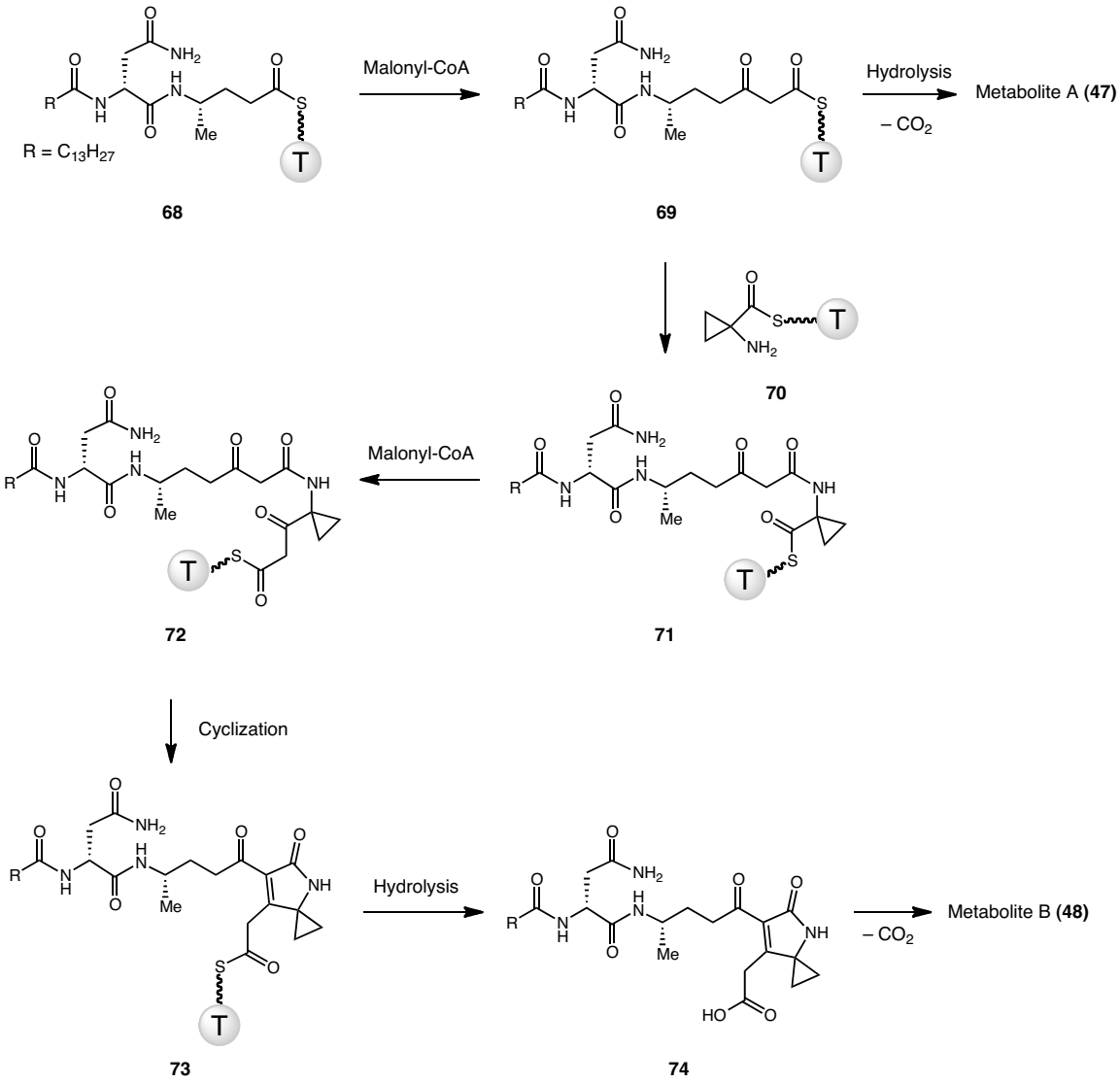


Figure 4.4: The biosynthetic hypothesis for the generation of Metabolite A (**48**) and B **(47**). Incorporation of malonyl-CoA, amide bond formation with a T domain-bound cyclopropane-containing amino acid (**70**) and another round of elongation with malonyl-CoA could yield an enzyme-bound intermediate **72** poised to undergo intramolecular cyclization to form **73**. Finally, hydrolysis, followed by decarboxylation could yield **48**. A T domain is shown to represent an unidentified ACP or PCP domain.

The unusual azaspiro[2.4]heptenone heterocycle found in Metabolite B (**48**) has not been observed in any other natural products to date. However, cyclopropanes, including spirocyclopropanes, are found in many metabolites produced by bacterial and eukaryotic organisms (Figure 4.5). For instance, a wide-array of bacteria, including *E. coli,* synthesize cyclopropane fatty acids from unsaturated fatty acids and *S*-adenosyl-L-methionine (SAM)

(Figure 4.5A).[12] In plants, the important signaling molecule ethylene is derived from 1-aminocyclopropane-1-carboxylic acid (ACC), which is biosynthesized from SAM by a pyridoxal-5′-phosphate-(PLP) dependent enzyme, ACC-synthase (Figure 4.5B).[13] The cyclopropane moiety of the phytotoxin coronatine (**81**) is biosynthesized by chlorination of a PCP domain-bound intermediate (**78**) by a Fe^{2+}-dependent enzyme (CmaB) to form the PCP domain-bound intermediate **79**, which then undergoes enzyme-catalyzed ring formation to form cyclopropane **80** (Figure 4.5C).[14] None of the *pks* enzymes have homology to characterized cyclopropane biosynthetic enzymes. Thus, we hypothesized that the biosynthesis of **48** may involve a novel strategy to construct the cyclopropane moiety.

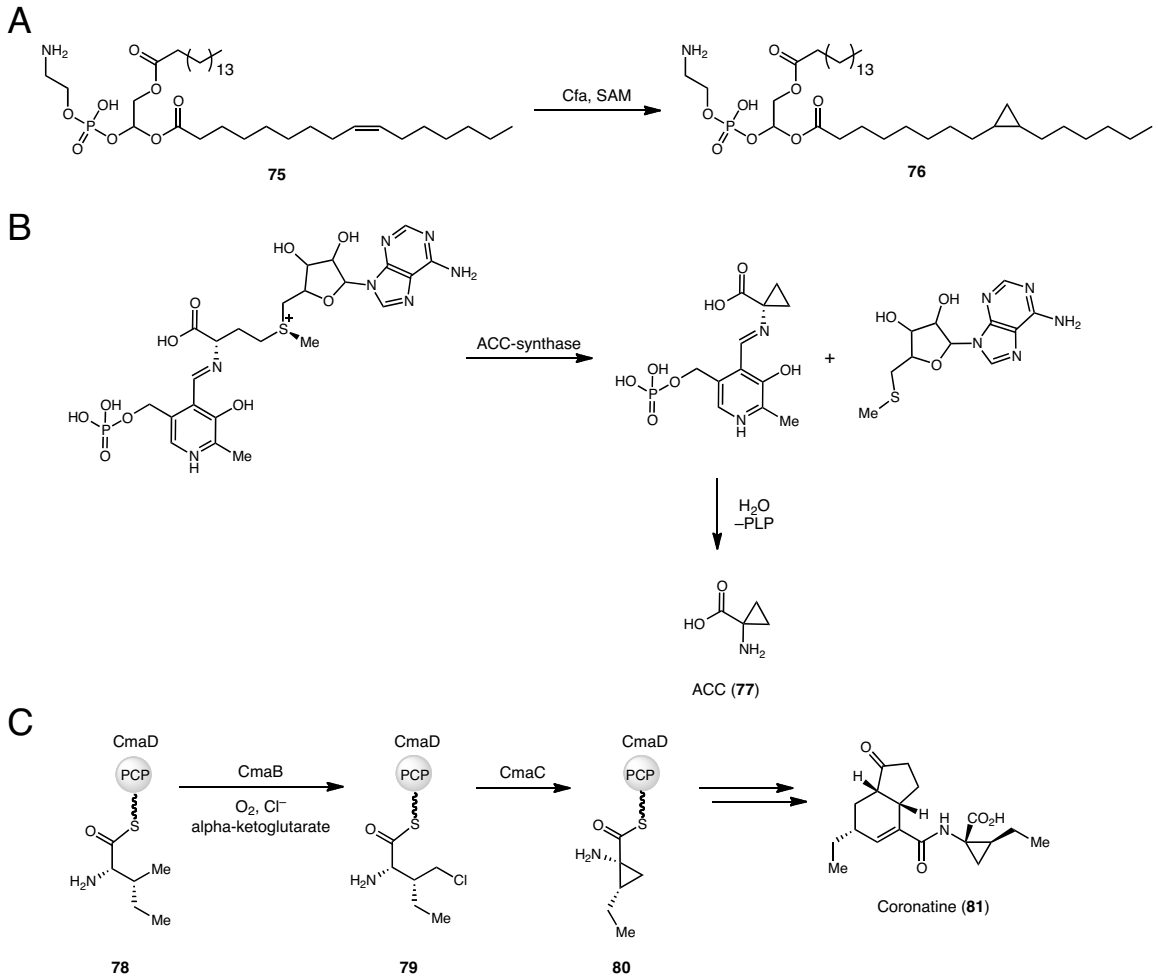


Figure 4.5: A) Unsaturated fatty acids (**75**) are transformed into cyclopropane fatty acids (**76**) by SAM-dependent enzyme Cfa. B) A PLP-dependent reaction transforms SAM into ACC (**77**). C) In coronatine (**81**) biosynthesis, a PCP domain-bound L-*allo*-isoleucine (**78**) is chlorinated by halogenase CmaB to form intermediate **79,** which is cyclized by CmaC to form **80**.

In their report concerning the isolation of Metabolite B (**48**), the Müller and Crawford labs also described feeding studies with labeled biosynthetic building blocks.[2,3] Their work revealed that the cyclopropane ring in **48** is derived from L-methionine. Moreover, the Müller lab found that feeding with L-[U-^{13}C, ^{15}N]methionine resulted in a +5 mass shift in **48**, while feeding L-[methyl-$^{2}H_3$]methionine failed to result in incorporation of a label.[2] These results suggested that cyclopropane formation involves the use of L-methionine, and may involve γ-cyclization of SAM as in ACC biosynthesis (Figure 4.3B).

4.3: Biosynthetic hypothesis for Metabolites C, D and E

The isolated *pks*-associated metabolites (Figures 4.2-4.3) share common structural motifs. Furthermore, results from feeding studies with labeled amino acids demonstrated that Metabolites C (**64**) and E (**66**) are derived from the same amino acid building blocks.[3] Our biosynthetic hypothesis for these colibactin pathway intermediates is provided below (Figure 4.6). The biosynthetic hypothesis for these metabolites begins with the intermediate (**73**) that gives rise to **48** upon hydrolysis from the T domain and decarboxylation. Amide bond formation with glycine would provide **82**. Amide bond formation with L-cysteine followed by ring formation would provide thiazoline-containing intermediate **83.** It is likely that a module containing a cyclization (Cy) domain, such as ClbJ or ClbK, would catalyze thiazoline and thiazole ring formation. It is not yet known whether cyclization of the glycine residue to form the pyridinone rings of Metabolites C (**64**) and D (**49**) occurs on the assembly line or after hydrolysis from the ppant arm. Here, we propose that cyclization occurs after thioester hydrolysis to give rise to **64** from **83**. From T domain-bound intermediate **83**, another round of amide bond formation with L-cysteine followed by ring formation and oxidation would provide thiazole intermediate **84.** Intermediate **84** would undergo hydrolysis to provide Metabolite E (**66**), or the thiazoline ring would be oxidized to provide T domain-bound intermediate **85**. Finally, hydrolysis and cyclization of the glycine amide nitrogen onto the ketone moiety would provide Metabolite D (**49**).

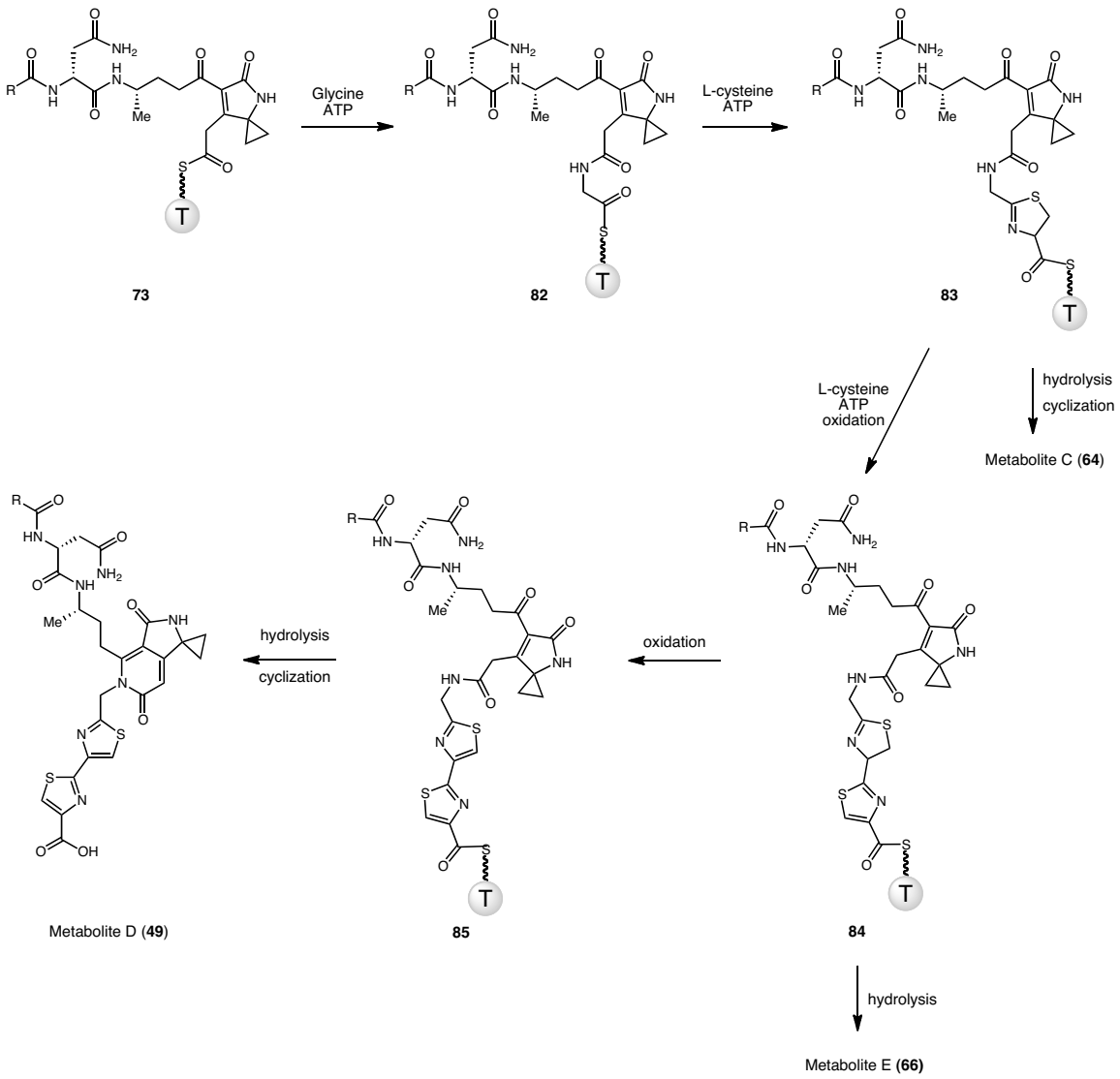


Figure 4.6 The biosynthetic hypothesis for the generation of Metabolites C (**64**), D (**49**) and E (**66**). Amide bond formation between glycine and T domain-bound intermediate **73**, followed by thiazoline ring formation with L-cysteine would give rise to **83**. Hydrolysis and cyclization of **83** would give rise to Metabolite C (**64**) whereas another round of ring formation with L-cysteine would give intermediate **84**. Hydrolysis of **84** would give Metabolite E (**66**), whereas oxidation of the thiazoline ring to a thiazole, followed by hydrolysis and cyclization would give Metabolite D (**49**). A T domain is shown to represent an unidentified ACP or PCP domain.

4.4: ClbB$_{PKS}$ is a hybrid NRPS/PKS with an unusual domain organization

We were interested in characterizing the biosynthesis of *pks* intermediates *in vitro* in order to learn more about the chemistry performed by the colibactin assembly line enzymes. We started our biochemical characterization by examining the first uncharacterized step in the proposed

biosynthetic pathway for the known *pks* metabolites. Specifically, we wondered which assembly line enzyme or enzymes incorporated two units of malonyl-CoA before the formation of an amide bond with a cyclopropane-containing amino acid (Figure 4.7). The first unit of malonyl-CoA (**87** to **68**) is reduced completely to the methylene, whereas the second (**68** to **69**) is kept at the ketone oxidation state to provide intermediate **69**. This intermediate would undergo hydrolysis and decarboxylation to provide **47**.

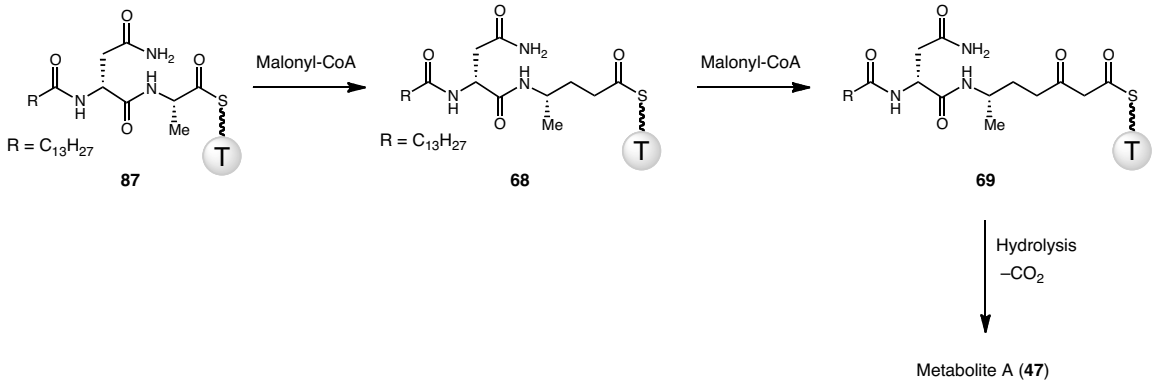


Figure 4.7: From the characterized intermediate **87,** two rounds of bond formation with malonyl-CoA, followed by hydrolysis and decarboxylation, could provide **47**.

In Chapter 2, the *in vitro* characterization of the beginning of the colibactin assembly line was described. First, NRPS module ClbN constructs an *N*-myristoylated-D-asparagine residue. This *N*-acylated amino acid is then extended with L-alanine by the NRPS module of ClbB, $ClbB_{NRPS}$, to provide a PCP-bound dipeptide intermediate (**87**). We hypothesized that the PKS module of ClbB, $ClbB_{PKS}$, which is part of the same polypeptide chain as $ClbB_{NRPS}$, would be responsible for the incorporation of at least one unit of malonyl-CoA (**87** to **68**). $ClbB_{PKS}$ has an unusual domain organization of KS-AT-KR-DH-ER-ACP. The more common order of domains is KS-AT-DH-ER-KR-ACP. The only example of a PKS module with the same domain organization as $ClbB_{PKS}$ – and from a pathway with a characterized product– is TubD, from the tubulysin biosynthetic pathway (Figure 4.8).[15] TubD is a hybrid NRPS/PKS and incorporates and completely reduces

one unit of malonyl-CoA in the construction of tubulysin.[8] TubD has not been shown to perform iterative bond formation.

Figure 4.8: $ClbB_{PKS}$ (the second module of ClbB) and the first module of TubD have the same domain organization.

We wondered whether the phylogeny of the ClbB KS domain would shed light on its biochemical activity. Previously, Crawford and co-workers constructed a maximum-likelihood tree of KS domains using 154 KS domain sequences and found that the KS domains of the colibactin PKS modules clustered together.[16] The closest relative of the colibactin KS domains was the KS domain of ZmaA from the zwittermicin pathway, which catalyzes C-C bond formation between an upstream aminomalonyl-ACP and a hydroxy-malonyl-ACP.[17] Two other close relatives were the KS domain of NspD, a *cis*-AT PKS module from nosperin biosynthesis that catalyzes bond formation between an upstream peptidyl-PCP intermediate and malonyl-ACP, and OzmK, a *trans*-AT PKS from oxazolomycin biosynthesis, which catalyzes bond formation between an upstream methoxy-malonyl-ACP and a malonyl-ACP.[18,19] These closely related KS domains come from biosynthetic pathways that share unusual features. First, the colibactin, nosperin, oxazolomycin and zwittermicin gene clusters encode for NRPS and PKS enzymes as well as *trans*-AT PKS modules. Second, the zwittermicin, oxazolomycin and colibactin clusters feature the biosynthesis and *in trans* use of rare building blocks. Finally, the zwittermicin, nosperin and colibactin clusters feature the unusual combination of both *cis* and *trans*-AT PKS modules. We also built a phylogenetic tree of 66 KS domains from a variety of PKS systems, such as type I modular PKS modules, including *cis*- and *trans*-AT PKS, hybrid NRPS/PKS modules, type II PKS pathways, and PKS modules from polyunsaturated fatty acid and enediyne biosynthetic pathways

(Figure 4.9). Similar to those results obtained by Crawford and co-workers, our tree showed that the colibactin KS domains (blue) cluster together, and that ZmaA (green) was a close relative of the colibactin KS domains. Importantly, the KS domain of ClbB did not cluster with iterative KS domains from polyunsaturated fatty acid or enediyne biosynthesis pathways (red).[20, 21] Thus, both phylogenetic and bioinformatics analyses suggested that $ClbB_{PKS}$ would catalyze only one round of bond formation to provide intermediate **68** (Figure 4.7).

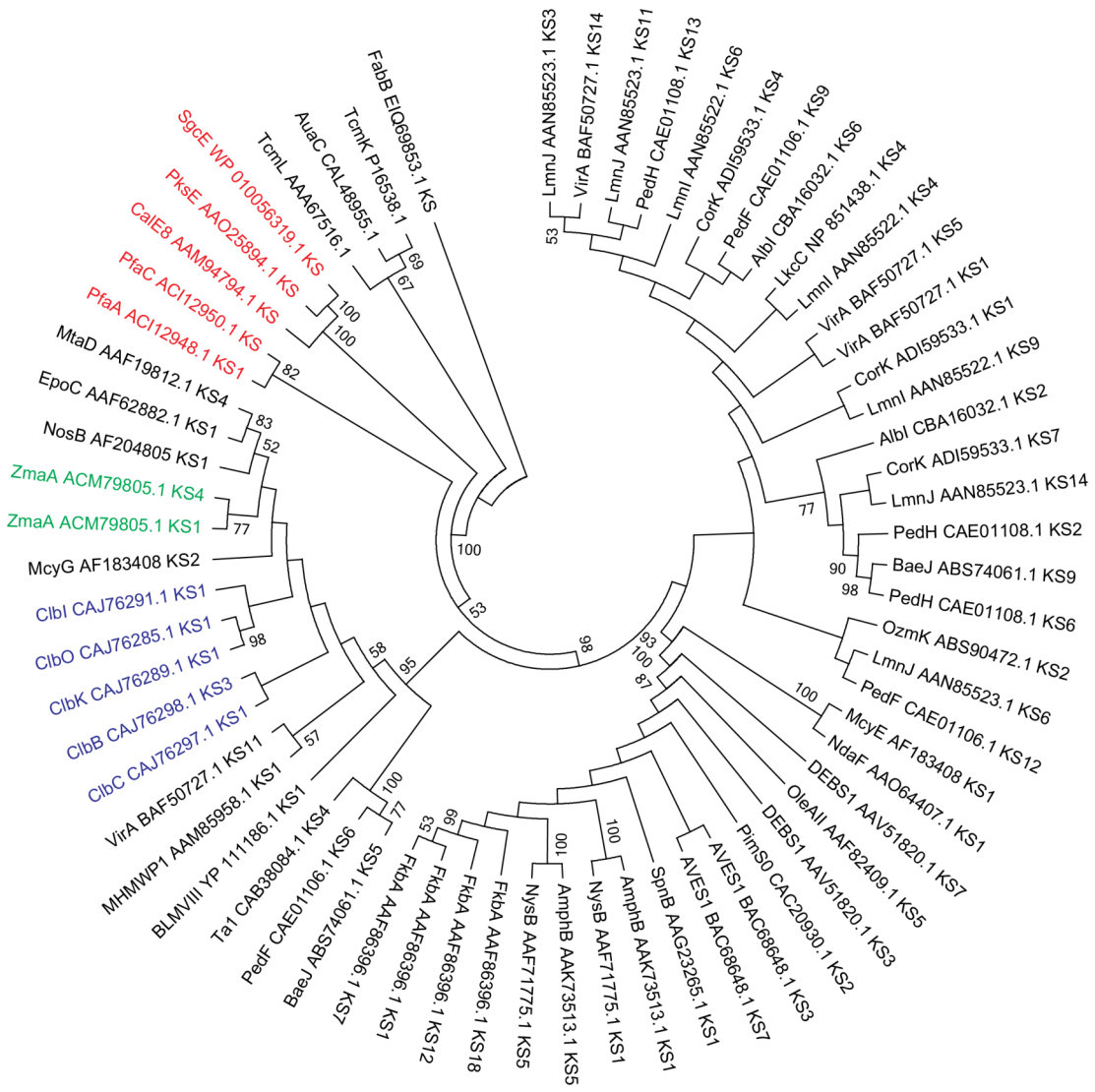


Figure 4.9: Molecular Phylogenetic analysis of excised KS domains using the Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model.[22] The tree with the highest log likelihood (-28441.3619) is shown. Initial trees for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. The analysis involved 66 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 304 positions in the final dataset. Evolutionary analyses were conducted in MEGA5.[23] The GenBank accession numbers of each protein are provided. The KS domains were excised using the PKS/NRPS analysis web-site from the University of Maryland.[24]

4.5: The *in vitro* biochemical characterization of $ClbB_{PKS}$

We sought to examine the bond-forming activity of $ClbB_{PKS}$ using *in vitro* reconstitution assays. $ClbB_{PKS}$ was cloned from *E. coli* CFT073, expressed as an N-His_6-tagged fusion protein in BL21 (DE3) cells, and purified using standard Ni-affinity chromatography (Figure 4.10). The truncation

point between the NRPS and PKS domains, the N-terminus of the encoded protein, was chosen based on the start site of the KS domain as predicted by the PKS/NRPS analysis web-site from the University of Maryland.[24] A boron BODIPY–CoA loading assay confirmed that the ACP domain of *apo*-ClbB$_{PKS}$ could be post-translationally modified to the *holo* protein by Sfp (Figure 4.11).

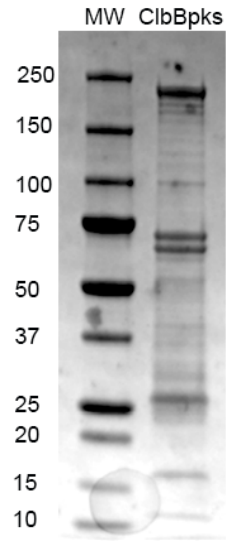


Figure 4.10: SDS-PAGE gel of purified ClbB$_{PKS}$-N-His$_6$. The molecular weight of ClbB$_{PKS}$ is 233 kDa. MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

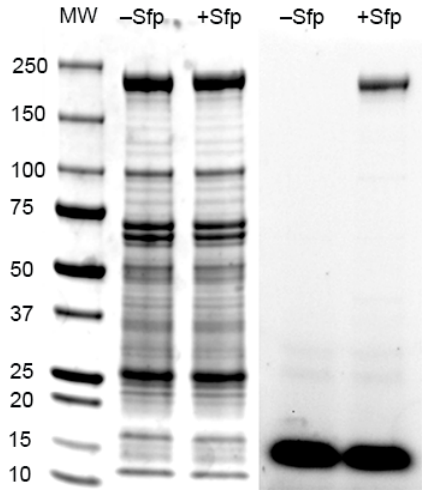


Figure 4.11: Loading of ClbB$_{PKS}$ with BODIPY-CoA by the ppant- transferase Sfp. In the presence of Sfp, ClbB$_{PKS}$ was loaded with BODIPY-CoA. MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

We began our biochemical characterization of $ClbB_{PKS}$ by examining the ability of $ClbB_{PKS}$ to elongate intermediates biosynthesized by ClbN and $ClbB_{NRPS}$ *in vitro* (Figure 4.12A). First, ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$ were incubated with ppant-transferase Sfp and CoA to convert each enzyme to the active, *holo* form. The *holo* enzymes were then incubated with the predicted cofactors and building blocks, including myristoyl-CoA, L-asparagine, L-alanine, ATP, $MgCl_2$, malonyl-CoA and nicotinamide adenine dinucleotide phosphate hydride (NADPH). The AT domain of $ClbB_{PKS}$ was predicted to utilize malonyl-CoA based on a conserved motif, which is described in more detail below. After incubation for 2 or 4 h, products were hydrolyzed from the ppant arm of the enzymes using an aqueous solution of potassium hydroxide (0.1 M) and analyzed by LC–MS. These reactions were analyzed for the presence of the expected product that would arise from elongation of the prodrug motif by $ClbB_{NRPS}$ (**19**) as well as possible products that could be biosynthesized by $ClbB_{PKS}$ (**88-92, 47**) (Figure 4.12B). Elongation of the $ClbB_{NRPS}$ PCP domain-bound *N*-acylated dipeptide intermediate by one unit of malonyl-CoA followed by complete reduction of the β-ketone would give rise to **89**. Incomplete or no reduction of the β-ketone would provide products **90-92**. We also searched for Metabolite A (**47**), which could be synthesized by elongation of the *N*-acylated dipeptide by two units of malonyl-CoA (Figure 4.7).

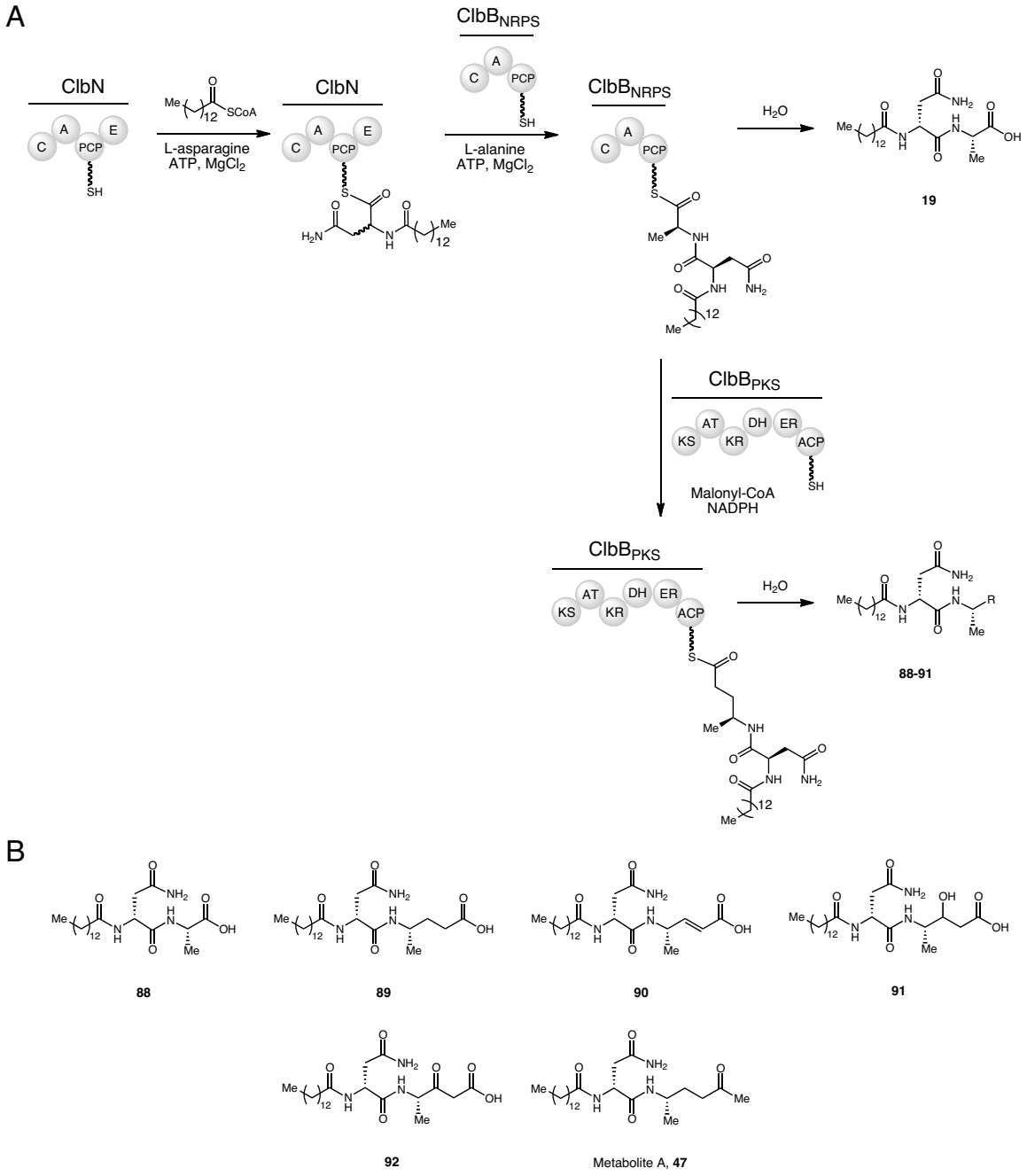


Figure 4.12: A) The *in vitro* reconstitution of ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$ with necessary cofactors. B) The production of compounds **88-92** and **47** was analyzed by LC–MS.

LC–MS analysis showed that the expected dipeptide product **19** was found in the complete reactions with both 2 and 4 h incubation, but was not found in the no ATP control (Figure 4.13), confirming that the activity of ClbN and $ClbB_{NRPS}$ could be reconstituted in the presence of

additional cofactors and $ClbB_{PKS}$. Interestingly, the only product observed from the action of $ClbB_{PKS}$ was the completely reduced product **89**, which was found at low levels in the full reaction with a 2 h incubation and at slightly higher levels in the full reaction with a 4 h incubation. The absence of products with higher oxidation states (**90**-**92**) as well as Metabolite A (**47**) suggested that $ClbB_{PKS}$ may perform just one round of bond formation with malonyl-CoA.

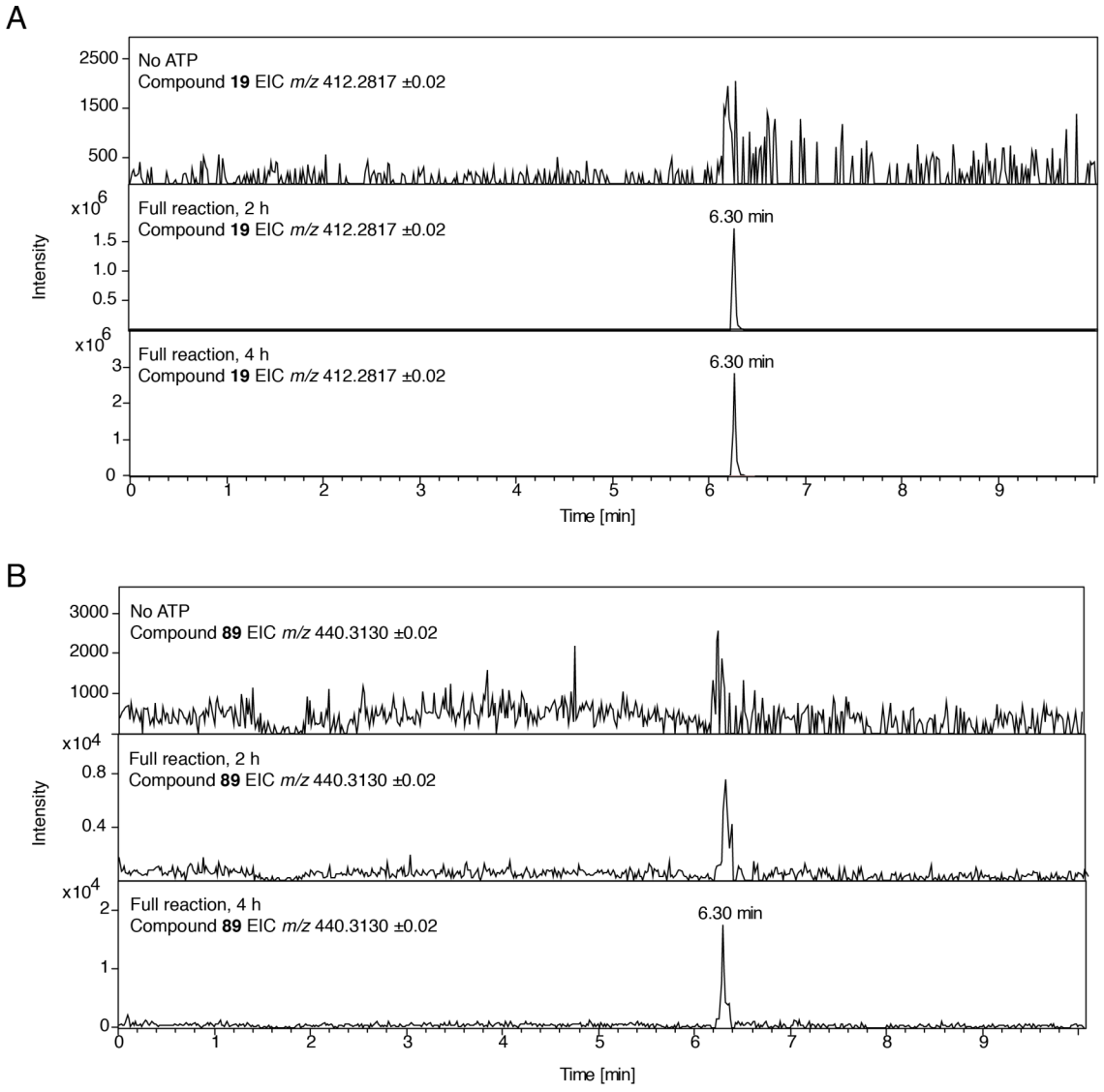


Figure 4.13: An LC–MS analysis of the *in vitro* reconstitution of ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$. A) EICs of product **19** (*m/z* 412.2817) in the no ATP control, the full reaction with 2 h incubation, and the full reaction with 4 h incubation. B) EICs of compound **89** (*m/z* 440.3130) in the no ATP control, the full reaction with 2 h incubation, and the full reaction with 4 h incubation.

We explored this reaction further by examining various factors that may have impacted the reaction outcome or efficiency. For instance, the C-His_6 construct of $ClbB_{PKS}$ overexpressed at higher yields, but displayed the same reactivity as the N-His_6 construct that was used in the above assay (Figure 4.14). In addition, we found that NADH inhibited the reaction, confirming that NADPH is the preferred cofactor for KR and ER domains of $ClbB_{PKS}$. Overall, these data suggested that $ClbB_{PKS}$ acts in a canonical fashion to install and completely reduce one unit of malonyl-CoA.

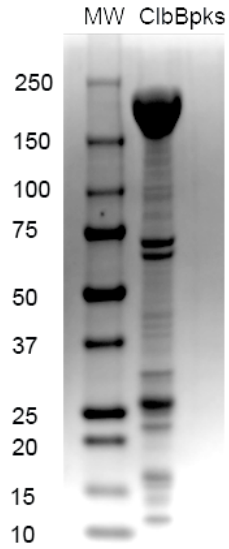


Figure 4.14: SDS-PAGE gel of purified $ClbB_{PKS}$-C-His_6. The molecular weight of $ClbB_{PKS}$ is 233 kDa. MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

4.6: Genetic studies on the role of ClbI in colibactin biosynthesis

We hypothesized that the only other PKS module found in the colibactin cluster that could be responsible for the incorporation of a second unit of malonyl-CoA was ClbI. Besides $ClbB_{PKS}$ and ClbI, the other PKS modules in the cluster were all predicted to be *trans*-AT PKSs that incorporate aminomalonate. ClbI is a *cis*-AT PKS module and its AT domain was predicted to activate malonyl-CoA based on a conserved signature found in characterized AT domains (Figure 4.15A).[25] Multiple sequence alignment with ClbB and the multimodular PKS from avermectin

biosynthesis (AVES1),[26] in which the first AT domain incorporates malonyl-CoA and the second AT domain incorporates methylmalonyl-CoA, revealed that ClbI, along with ClbB, contains the HAFH motif that signifies selectivity for malonyl-CoA. Along with this motif, the ClbI AT domain has the conserved active site residues GHS and R, which suggests that this domain is active.[23] Interestingly, a multiple sequence alignment with characterized KS domains demonstrates that the KS domain of ClbI does not have the active site cysteine that is required for tethering an upstream intermediate that serves as the electrophile in the decarboxylative Claisen condensation catalyzed by this domain (Figure 4.15B). Instead, ClbI has a serine residue, which is also seen in the first domain of nystatin biosynthetic pathway, NysA. NysA has been shown to catalyze decarboxylation of malonyl-CoA.[27]

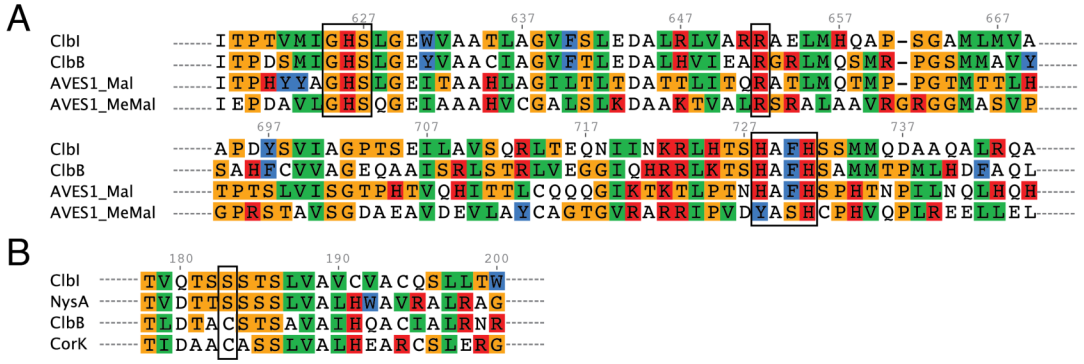


Figure 4.15: Multiple sequence alignment of ClbI with characterized PKS enzymes. The amino acid numbering refers to that of ClbI. A) The ClbI AT domain has the conserved active site residues GHS and R and the HAFH selectivity motif for malonyl-CoA. Accession number: AVES1: BAC68648. B) The conserved active cysteine residue is a serine in the KS domain of ClbI and NysA. Accession numbers: NysA: AAF71774; CorK: ADI59533[28]

In order to investigate the role of this enzyme in the biosynthesis of Metabolite A and B, we turned to genetic studies. The *clbI* gene encoded on BAC*pks* and BAC*pks*Δ*clbP* was knocked out using λ Red recombinase-mediated gene disruption.[29] In addition, a pTrc-HisA complementation vector harboring *clbI* (pTrc-ClbI) was cloned and transformed into DH10B-BAC*pks*Δ*clbI* and BAC*pks*Δ*clbP*Δ*clbI* strains such that we could determine whether the effects seen in the *clbI*

knockout strains were due to loss of this gene's function or polar effects. Small scale cultures (50 mL) of the Δ*clbI* and complemented strains were grown in LB media for 24 hours, and the cells were harvested, lyophilized and extracted with methanol. LC–MS analysis showed that the hydrolyzed prodrug motif (**10**) was found in the extracts of cultures expressing *clbP* and at much lower levels (ten-fold lower) in extracts of cultures lacking *clbP* (Figure 4.16A). Interestingly, Metabolite A (**47**) was found in extracts of DH10B-BAC*pks*Δ*clbP*Δ*clbI* cultures (Figure 4.16B). The levels of **47** were about ten-times lower in the extracts of DH10B-BAC*pks*Δ*clbP*Δ*clbI* / pTrc-ClbI cultures, which may indicate that overexpression of ClbI somehow diverts the flux of the biosynthetic pathway away from production of **47**. Metabolite B (**48**) was seen in culture extracts of Δ*clbP* strains also expressing *clbI*, including DH10B-BAC*pks*Δ*clbP*Δ*clbI* / pTrc-ClbI, but was not observed in extracts of DH10B-BAC*pks*Δ*clbP*Δ*clbI* (Figure 4.16C). Overall, these data strongly suggested that ClbI is not responsible for incorporation of the second unit of malonyl-CoA (**68** to **69**, Figure 4.4), but may be involved in a downstream biosynthetic step. We hypothesized that ClbI may be involved in the biosynthesis of Metabolite B, perhaps in the incorporation of malonyl-CoA to convert **71** to **72** (Figure 4.4).

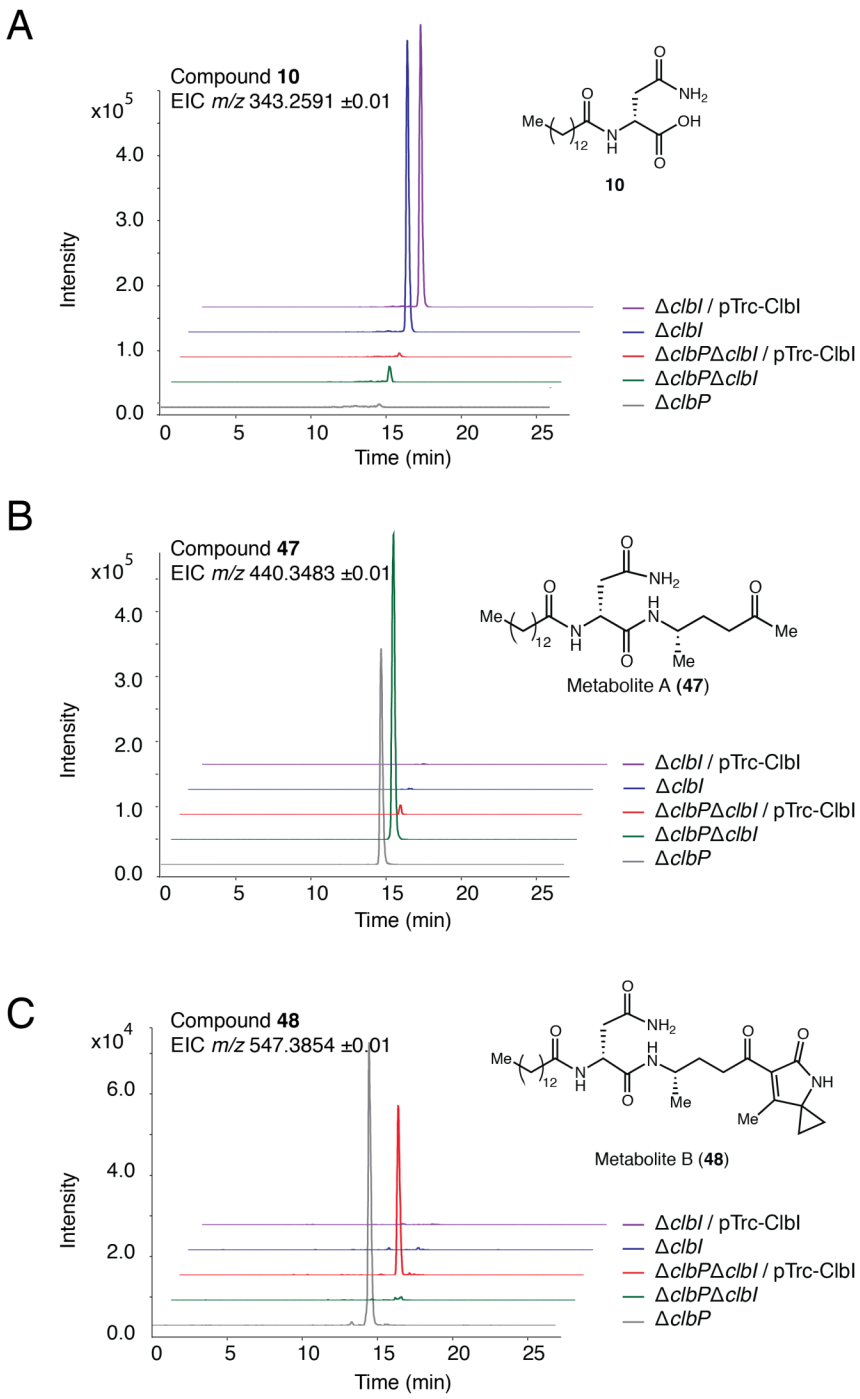


Figure 4.16: Extracted ion chromatograms for compounds of interest in the extracts of *pks*$^+$ strains. A) The hydrolyzed prodrug motif (**10**) was abundant in BAC*pks*Δ*clbI* and BAC*pks*Δ*clbI* / pTrc-ClbI strains. B) Metabolite A (**47**) was abundant in BAC*pks*Δ*clbP* and BAC*pks*Δ*clbP*Δ*clbI* strains. C) Metabolite B (**48**) was in BAC*pks*Δ*clbP* and complemented BAC*pks*Δ*clbP*Δ*clbI* / pTrc-ClbI strains.

What other *pks* enzymes could be responsible for this biochemical step? Qian and co-workers constructed gene knock-outs of biosynthetic genes in a DH10B strain harboring BAC*pks*Δ*clbP* and examined the metabolite profiles of the resultant double knock-out strains.[4] They observed that a Δ*clbP*Δ*clbB* strain was deficient in the production of the hydrolyzed prodrug motif (**10**) and higher molecular weight metabolites, such as Metabolite A (**47**). In contrast, a Δ*clbP*Δ*clbC* strain produced *N*-acylated dipeptide (**19**)–the product of ClbN and $ClbB_{NRPS}$– but failed to produce Metabolite A (**47**). These results prompted the authors to conclude that ClbC is responsible for this biochemical step to convert putative intermediate **68** to **69**. However, when we examined the metabolite profile of Δ*clbP*Δ*clbC E. coli* Nissle 1917, which was provided to us by the Müller lab, we observed that this strain still produced Metabolite A (**47**) (Figure 4.17). The reasons for this discrepancy between our results and those of the Qian group are unclear. One explanation could be that the Δ*clbP*Δ*clbC* strain examined by Qian and co-workers produced very low levels of **47** compared to the Δ*clbP*Δ*clbC E. coli* Nissle 1917 strain. The Qian lab utilized DH10B expressing a BAC containing the entire *pks* gene cluster for their studies, which may have different expression levels of metabolite concentrations compared to a strain with a chromosomal copy of the *pks* island, such as *E. coli* Nissle 1917.

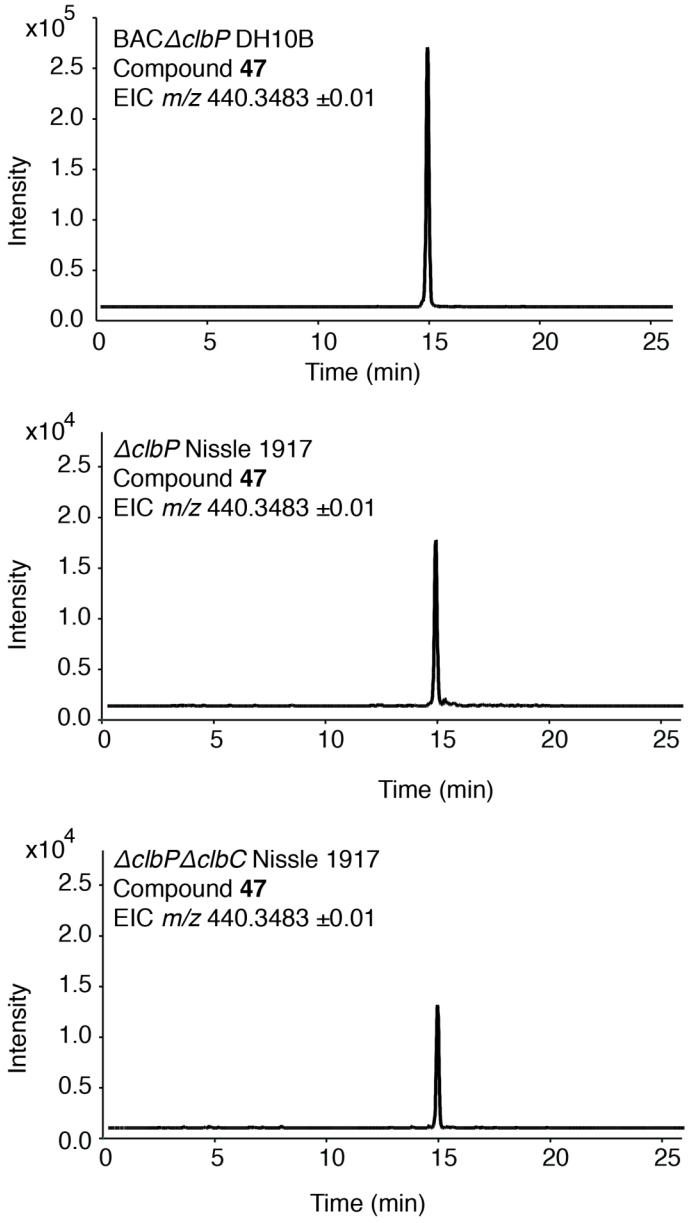


Figure 4.17: An LC–MS analysis of methanol extraction of whole-culture aliquots of DH10B harboring BAC*pks*Δ*clbP*, as well as Δ*clbP* and Δ*clbP*Δ*clbC E. coli* Nissle 1917. The EICs of compound **47** (*m/z* 440.3483) are shown. Please note that the chromatograms are not all at the same intensity scale.

4.7: Studying the role of fatty acid synthase enzymes in colibactin biosynthesis

The *in vitro* reconstitution assay with $ClbB_{PKS}$ and *in vivo* studies of *clbP*Δ*clbI* strains demonstrated that neither ClbB nor ClbI was responsible for the incorporation of the second unit of malonyl-CoA to convert **68** to **69** (Figure 4.4). One possible explanation for these observations

is that our *in vitro* assays failed to capture non-canonical chemistry that may be performed by ClbB *in vivo*. However, we also considered that an enzyme outside of the *pks* gene cluster could be responsible for this biochemical step. Specifically, we hypothesized that an enzyme from fatty acid synthase (FAS) could be involved. Components of FAS are known to participate in secondary metabolite biosynthetic pathways. For example, type II PKS gene clusters often do not harbor a gene for an AT domain, and instead use the equivalent domain from FAS, the malonyl-CoA acyltransferase (MCAT).[30] We hypothesized that the MCAT from *E. coli* FAS, FabD, could act in *trans* with $ClbB_{PKS}$ in a way that was analogous to the biochemistry seen in type II PKS pathways. In addition, we also thought that perhaps the elongating KS from FAS, FabB, or the ACP from FAS could be involved in the incorporation of the second unit of malonyl-CoA.

In order to test our hypothesis, FabD, FabB and the FAS ACP were overexpressed in BL21 (DE3) *E. coli* and purified using standard Ni-chromatography (Figure 4.18), as described previously by Khosla and co-workers.[31] The purified enzymes were reconstituted *in vitro* with ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$. The FAS enzymes were added both as single enzymes and as a mixture to the *pks* enzymes. Unfortunately, while we did see compound **19**, we failed to see the production of the expected $ClbB_{PKS}$ product **89** in our positive control reactions containing just the *pks* enzymes ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$. Furthermore, no additional products arising from incorporation of a second unit of malonyl-CoA were observed by LC–MS. These results are inconclusive as to the role of FAS enzymes in colibactin biosynthesis.

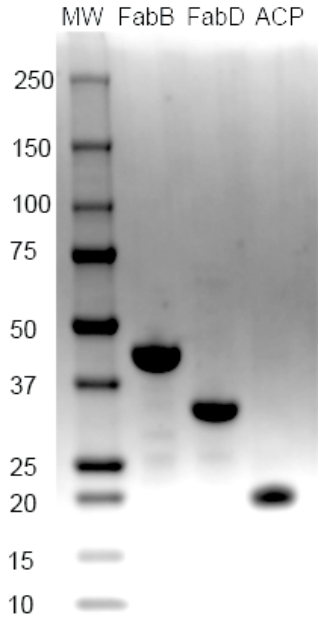


Figure 4.18: SDS-PAGE gel of purified FabB, FabD and ACP. The molecular weight of FabB is 43 kDa; FabD is 32 kDa; ACP is 9 kDa. The appearance of *holo* ACP at an apparent higher molecular weight on SDS-PAGE gels was noted by Khosla and co-workers.[23] MW = Precision Plus Protein All Blue Molecular Weight Standards (Bio-Rad).

4.8: Genetic studies on the role of ClbD and ClbG in colibactin biosynthesis

As stated above, aminomalonate-forming and utilizing enzymes are required for genotoxicity, but no known *pks* metabolites contain structural components that are derived from aminomalonate. Feeding studies using labeled building blocks have shown that the biosynthesis of Metabolites A-E (**47-49, 64** and **66**) does not incorporate L-serine, the amino acid from which aminomalonate is derived.[2,3] Furthermore, Δ*clbP* strains that also lack aminomalonate-forming enzymes *clbD*, *clbE*, *clbF* or *clbG* are still able to produce Metabolite B (**48**).[2] Given these data, we sought to identify amimomalonyl-derived *pks* metabolites.

We constructed Δ*clbP*Δ*clbG* and Δ*clbP*Δ*clbD* knock-out strains that we anticipated could be useful for comparative LC–MS studies. The *clbD* and *clbG* genes encoded on BAC*pks* and BAC*pks*Δ*clbP* were knocked out using λ Red recombinase-mediated gene disruption.[27] We first examined the ability of these strains to produce Metabolites C (**64**) and D (**49**). Interestingly, *clbD*

and *clbG* knock-out strains produced much higher levels of these metabolites compared to the control strain: over 35-times higher levels of Metabolite D (**49**) were seen in extracts of Δ*clbP*Δ*clbG* and Δ*clbP*Δ*clbD* strains compared to extracts of the control Δ*clb*P strain (Table 4.1). The significantly higher titers of **49** in Δ*clbP*Δ*clbG* cultures allowed Dr. Wilson to isolate reasonable quantities of this compound from extracts of this strain and complete the structural characterization of this metabolite.[6] Furthermore, these data unequivocally established that aminomalonate-forming machinery is not utilized in the biosynthesis of the known *pks* metabolites.

Table 4.1: Normalized areas of EICs for Metabolites C (**64**) and D (**49**) in DH10B harboring BAC*pks*Δ*clbP*, BAC*pks*Δ*clbP*Δ*clbD* or BAC*pks*Δ*clbP*Δ*clbG*.

Metabolite C (64)

Strain	Normalized Area
ΔclbP	1.00
Δ*clbP*Δ*clbD*	1.91
Δ*clbP*Δ*clbG*	2.10

Metabolite D (49)

Strain	Normalized Area
ΔclbP	1.00
Δ*clbP*Δ*clbD*	35.90
Δ*clbP*Δ*clbG*	38.50

Having established that DH10B harboring BAC*pks*Δ*clbP*Δ*clbD* or BAC*pks*Δ*clbP*Δ*clbG* produced **64** and **49**, we turned to comparative metabolomics to identify metabolites enriched in extracts of the metabolome of DH10B harboring BAC*pks*Δ*clbP* compared to extracts of the metabolome of a knock-out in the aminomalonate-utilization pathway, such as DH10B harboring BAC*pks*Δ*clbP*Δ*clbG*. We hypothesized that those metabolites enriched in extracts of Δ*clbP* strains would be likely to contain aminomalonate-derived moieties. To investigate this hypothesis, small culture volumes (4 mL) of *E. coli* DH10B harboring BAC*pks*Δ*clbP* or BAC*pks*Δ*clbP*Δ*clbG* were

grown in triplicate in Luria–Bertani (LB) broth for a total of 24 hours. A whole-culture aliquot was lyophilized and extracted with methanol for LC–MS analysis. The raw LC–MS data was analyzed using XCMS, which identified three features that varied significantly between *ΔclbP* and *ΔclbPΔclbG* extracts ($p < 0.01$), had a maximum intensity greater than 10,000 counts and were enriched by ten-fold or greater (Table 4.2). Interestingly, there were no features that were enriched in the extracts of the *ΔclbP* culture, even with less strict criteria applied (e.g. enriched by five-fold or greater). We did not pursue the characterization of the metabolites (Features 1-3, Table 4.2) identified in this experiment further, as we rationalized the biosynthesis of these metabolites could not include the use of the aminomalonate machinery.

Table 4.2: Results from XCMS analysis. "DOWN" indicates enrichment in BAC*pksΔclbP*, and "UP" indicates enrichment in BAC*pksΔclbPΔclbG*. Average intensity: the average integrated intensity of that feature across samples in each condition, either BAC*pksΔclbP* or BAC*pksΔclbPΔclbG*. Fold change: the ratio of the two average intensities of that feature from each condition. p-value: the significance of the difference for a given feature between the two conditions (calculated in the XCMS program using a Welch t-test with unequal variances).

Feature	UP/ DOWN	*m/z*	Average intensity: *ΔclbP*	Average intensity: *ΔclbPΔclbG*	Fold change	p-value	Retention time (min)
1	UP	662.4753	8.23E+02	6.50E+04	78.9	0.00043	55.69
2	UP	864.4641	1.27E+04	1.36E+05	10.7	0.00731	30.06
3	UP	549.3915	8.60E+03	1.04E+05	12.1	0.00992	38.16

4.9: Conclusions

The structures of Metabolite B (**47**) and related metabolites (**63**-**66**, **49**) suggested that the colibactin assembly line enzymes perform interesting biochemical transformations that merit greater study. Using our understanding of biosynthetic logic as well as information gained from bioinformatic analyses of the assembly line enzymes, we proposed biosynthetic hypotheses for these molecules that ultimately guided our biochemical and genetic experiments. Our attempts to characterize one proposed biochemical step (**68** to **69**, Figure 4.4) were described in this chapter. *In vitro* reconstitution assays demonstrated that $ClbB_{PKS}$ can extend the dipeptide intermediate

generated by ClbN and $ClbB_{NRPS}$ using malonyl-CoA. In these assays, one round of bond formation and complete reduction of the β-keto thioester were observed. We did not observe incorporation of a second unit of malonyl-CoA in the *in vitro* assays with $ClbB_{PKS}$. *In vitro* reconstitution with ClbN, $ClbB_{NRPS}$, $ClbB_{PKS}$ and FabB, FabB and the ACP from *E. coli* FAS gave inconclusive results. In addition to reconstitution of individual *pks* enzymes *in vitro,* we also undertook *in vivo* studies with various gene knock-out *pks*$^+$ strains. *ClbI* knock-out strains demonstrated that ClbI is not required for the incorporation of malonyl-CoA to extend the intermediate putatively biosynthesized by $ClbB_{PKS}$. However, our results indicate that ClbI is involved in the biosynthesis of Metabolite B (**48**). Experiments with *clbG* and *clbD* knock-out strains illustrated that these enzymes are not required for the biosynthesis of Metabolite C (**64**) or D (**49**). An attempt to discover aminomalonyl-derived metabolites using comparative metabolomics failed to identify compounds enriched in the culture extracts of DH10B harboring BAC*pksΔclbP* compared to those of DH10B expressing BAC*pksΔclbPΔclbG.*

Despite significant progress in our understanding of colibactin biosynthesis since its discovery in 2006, many important questions remain concerning the structure of the active genotoxin and how structural motifs in known *pks* metabolites are biosynthesized. First, as mentioned previously, several lines of evidence support the notion that the final active product (or products) of the *pks* assembly line contain structural units derived from aminomalonate, and to date, no known *pks* metabolites contain moieties derived from this unusual building block. We believe that the best way to discover aminomalonate-containing *pks* metabolites is to identify the final and active product of the pathway. To discover the active genotoxin, it may be necessary to use alternative discovery strategies that do not rely on Δ*clbP* strains or comparisons of the metabolite profiles of wild-type and *pks*$^+$ strains lacking individual biosynthetic genes. One strategy could be to identify DNA adducts in DNA isolated from human cells exposed to *pks*$^+$ *E. coli.* The mass of these

adducts could indicate the molecular weight of the active colibactin and could guide isolation efforts from wild-type *pks*$^{+}$ *E. coli.* Another approach could be to isolate compounds covalently attached to ClbS, which has been shown to react with electrophilic probe compounds through a cysteine residue.[10] While both of these approaches rely on hypotheses–one regarding the mechanism of action of colibactin and the other concerning the putative self-resistance protein ClbS– and could fail due to the possibly incorrect assumptions of these hypotheses, it is clear from our prior work that future strategies should target the active colibactin. Examination of metabolite profiles of strains heterologously expressing the *pks* island or pathways with specific biosynthetic genes knocked-out may lead to the discovery of abundant intermediates and shunt products, the structures of which may or may not be relevant to that of the final genotoxin.

Second, most of the steps in the biosynthesis of known *pks* metabolites are still uncharacterized. Interesting questions regarding the biochemical steps in the colibactin pathway include the step that was the focus of this chapter, the formation of the proposed ACC building block (**70,** Figure 4.4), and the function of ClbI in the biosynthesis of Metabolite B (**48**). The first step in answering these types of questions is to establish the minimal set of enzymes required to biosynthesize an intermediate of interest. To clearly establish this minimal enzyme set, genetic studies like those described here may prove useful. Another strategy could be to reconstitute the biosynthetic enzymes of interest *in vivo* individually or as sets of enzymes. Once the minimal set of biosynthetic enzymes is firmly established, the most interesting biochemical questions should be examined using *in vitro* biochemical assays with those enzymes.

Finally, the biological implications of colibactin production in natural systems have yet to be fully elucidated. First, the role of *pks*$^{+}$ *E. coli* in maintaining health or promoting disease in the human gut is unclear. For instance, how does colibactin production impact the efficacy of the widely

used probiotic strain *E. coli* Nissle 1917? Previous studies have examined this question by comparing wild-type and Δ*clbA* strains of *E. coli* Nissle 1917.[32] ClbA–the ppant-transferase encoded in the *pks* island– can modify the assembly line enzymes from yersiniabactin biosynthesis.[33] Thus, hypotheses regarding the role of colibactin production in the efficacy of *E. coli* Nissle 1917 should be tested again using a negative control strain with a knock-out of a *pks* gene that is not involved in siderophore production. In patients with CRC or inflammatory bowel disease, do *pks*$^+$ *E. coli* promote disease progression? This question has been studied using a mouse model for CRC that was inoculated with a single *E. coli* strain, which does not represent the gut microbiota of either a healthy individual or a person suffering from inflammatory bowel disease.[34] Model systems that more closely mimic the natural state of the gut microbiota may be required to study these types of issues. Beyond the human gut, what are the ecological effects of colibactin production? To date, the *pks* island has been found in bacteria in the phylum Proteobacteria that are also symbionts of eukaryotes. Does colibactin play a role in mediating host-symbiont interactions? Are other host-microbe systems that feature *pks*$^+$ symbionts more amenable to the study of colibactin biology than the human gut?

We believe that knowledge of colibactin's structure and the ability to isolate or synthesize this compound will prove invaluable in answering these types of complex questions. In addition, the ability to modulate colibactin production with a chemical probe could obviate the use of knock-out *pks*$^+$ strains and highly simplified gut communities. One strategy to modulate colibactin production could be the inhibition of the prodrug peptidase ClbP. A selective inhibitor of ClbP could serve as a powerful tool to help dissect the biological effects of colibactin in mediating host-microbe interactions and disease progression in the human gut.

Overall, the study of the biosynthesis and structure of colibactin has led to fascinating insights that

demonstrate that there is still much to learn about microbial natural products, from how the secondary metabolites are biosynthesized to why the microbes produce them. During the course of this project we have elucidated assembly line biochemistry involved in self-resistance and isolated metabolites with unusual structures. Importantly, the work described here has enabled future research into the biochemistry of colibactin biosynthesis and the biology of *pks*$^{+}$ *E. coli* in the human gut.

4.10: Experimental section

Oligonucleotide primers were synthesized by Integrated DNA Technologies (Coralville, IA). Recombinant plasmid DNA was purified with a Qiaprep Kit from Qiagen. Gel extraction of DNA fragments and restriction endonuclease clean up were performed using an Illustra GFX PCR DNA and Gel Band Purification Kit from GE Healthcare. DNA sequencing was performed by Beckman Coulter Genomics (Danvers, MA). Optical densities of *E. coli* cultures were determined with a DU 730 Life Sciences UV/Vis spectrophotometer (Beckman Coulter) by measuring absorbance at 600 nm. HPLC was performed on a Dionex Ultimate 3000 instrument (Thermo Scientific). All chemicals and solvents were obtained from Sigma-Aldrich except where noted. High-resolution LC–MS analyses were performed in the Small Molecule Mass Spectrometry Facility at Harvard University. For the Bruker Maxis Impact q-TOF, the countercurrent drying gas heater was set to 10 L/min and the temperature maintained at 200 °C, the nebulizer pressure was set to 30 psi, the capillary was set to 4000 V in the positive-ion mode and 4500 V in the negative-ion mode, and the system was calibrated internally with a post-run injection of sodium formate solution. For the Agilent 6210 ESI-TOF, the capillary voltage was set to 3.5 kV and the fragmentor voltage to 100 V, and the drying gas temperature was maintained at 350 °C with a flow rate of 10 L/min and a nebulizer pressure of 45 psi. MassHunter quantitation software was used to process data.

Cloning, overexpression, and purification of $ClbB_{PKS}$-N-His_6 and C-His_6, ClbI, FabB, FabD and ACP

Table 4.3: Oligonucleotides used for cloning. Restriction sites are underlined.

Primer Name	Target	Sequence (5' to 3')
$ClbB_{PKS}$-F	ClbB-PKS-N- and C-His_6	AATTCATATGCCGGTGGCGATTGTC
$ClbB_{PKS}$-R-C-His	ClbB-PKS-C-His_6	GCTATCCTCGAGATCCAAAGACGTGTG
$ClbB_{PKS}$-R-N-His	ClbB-PKS-N-His_6	GCTATCCTCGAGTTAATGCAAAGACGT
ClbI-F	ClbI-N-His_6	AATTGAATTCTCATGGCAGAGAATGATTTTG
ClbI-R	ClbI-N-His_6	TGCTAAGCTTTCACTCATTAATCATGTCG

$ClbB_{PKS}$ was PCR amplified from *E. coli* CFT073 genomic DNA (purchased from the American Type Culture Collection, Manassas, VA) using the primers shown in Table 4.3. *$ClbB_{PKS}$* encodes for the C-terminal 2122 amino acids of the protein, which contains the polyketide synthase domains. *$ClbB_{PKS}$* was amplified using the forward primer **$ClbB_{PKS}$-F** and the reverse primers **$ClbB_{PKS}$-R-N-His** or **$ClbB_{PKS}$-R-C-His**. All PCR reactions contained 25 µL of Q5 High Fidelity 2X Master Mix (New England Biolabs), 1 ng of DNA template, and 500 pmoles of each primer in a total volume of 50 µL. Thermocycling was carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation for 30 sec at 98 °C, followed by 40 cycles of 10 sec at 98 °C, 30 sec at the annealing temperature of 71°C, 4 min at 72 °C, and a final extension time of 5 min at 72 °C. PCR reactions were analyzed by agarose gel electrophoresis with ethidium bromide staining, pooled, and purified. Amplified fragments were digested with NdeI and XhoI (New England Biolabs) for 2.5 h at 37 °C. Digests contained 1 µL of water, 3 µL of NEB Buffer 4 (10x), 3 µL of BSA (10x), 20 µL of PCR product, and 1.5 µL of each restriction enzyme (20,000 U/µL). Restriction digests were purified directly using agarose gel electrophoresis. Gel fragments were further purified using the Illustra GFX kit. The digests were ligated into linearized expression vector pET-29b or pET-28a using T4 DNA ligase (New England Biolabs) to encode a C-terminal

or N-terminal His_6-tagged construct, respectively. Ligations were incubated at room temperature for 2 h and contained 3 µL of water, 1 µL of T4 Ligase Buffer (10x), 1 µL of digested vector, 3 µL of digested insert DNA, and 2 µL of T4 DNA Ligase (400 U/ µL). 5 µL of each ligation was used to transform a 50 µL chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of the purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* BL21 (DE3) cells (Invitrogen) and stored at –80 °C as frozen LB/glycerol stocks.

Large scale overexpression and purification of $ClbB_{PKS}$

A 50 mL starter culture of pET-29b- or pET-28a-$ClbB_{PKS}$ BL21 *E. coli* was inoculated from a frozen stock and grown overnight at 37 °C in LB medium supplemented with 50 µg/ml kanamycin. Overnight cultures were diluted 1:100 into 2 L of LB medium containing 50 µg/mL kanamycin. Cultures were incubated at 37 °C with shaking at 175 rpm, moved to 15 °C at OD_{600} = 0.2-0.3, induced with 500 µM IPTG at OD_{600} = 0.5-0.6, and incubated at 15 °C for ~ 16 h. Cells from 2 L of culture were harvested by centrifugation (6,000 rpm x 10 min) and resuspended in 80 mL of lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, pH 8). The cells were lysed by passage through a cell disruptor (Avestin EmulsiFlex-C3) twice at 10,000 psi, and the lysate was clarified by centrifugation (13,000 rpm x 30 min). The supernatant was incubated with 2 mL of Ni-NTA resin and 5 mM imidazole for 1 h at 4 °C. The mixture was loaded into a glass column, and the flow-through was discarded. Protein was eluted from the column using a stepwise imidazole gradient in elution buffer (20 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, pH 8, 25 mM, 50 mM, 75 mM, 100 mM, 125 mM, 150 mM, 200 mM), collecting 4 mL fractions. SDS–PAGE analysis (4–15% Tris-HCl gel) was employed to ascertain the presence and purity of protein in each fraction. Fractions containing the desired protein were combined and dialyzed twice against 2 L of storage buffer (20 mM Tris-HCl, 50 mM NaCl, 10% (v/v) glycerol, pH 8). Solutions containing protein

were frozen in liquid N_2 and stored at –80 °C. This procedure afforded yields of 1.6 mg/L for C-His_6-tagged $ClbB_{PKS}$ and 0.6 mg/L for N-His_6-tagged $ClbB_{PKS}$.

Cloning of pTrc-ClbI

ClbI was PCR amplified from *E. coli* CFT073 genomic DNA (purchased from the American Type Culture Collection, Manassas, VA) using the primers shown in Table 4.1. PCR reactions contained 25 µL Q5 High-Fidelity 2X Master Mix (New England Biolabs), 1 ng of DNA template, and 500 pmoles of each primer in a total volume of 50 µL. PCR reactions were carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation for 30 s at 98 °C, followed by 35 cycles of 10 s at 98 °C, 30 s at 68 °C, 105 s at 72 °C, and a final extension time of 10 min at 72 °C. PCR reactions were analyzed by agarose gel electrophoresis with ethidium bromide staining, pooled, and purified. Amplified fragments were digested with EcoRI and HindIII (New England Biolabs) for 2.5 h at 37 °C. Digests contained 1 µL of water, 3 µL Cut Smart Buffer (New England Biolabs), 20 µL of PCR product, and 1.5 µL of each restriction enzyme (20,000 U/µL). Restriction digests were purified directly using agarose gel electrophoresis. Gel fragments were further purified using the Illustra GFX kit. The digests were ligated into linearized pTrcHisA (Invitrogen) using T4 DNA ligase (New England Biolabs) to encode a N-terminal His_6-tagged construct. Ligations were incubated at room temperature for 2 h and contained 3 µL of water, 1 µL of T4 Ligase Buffer (10x), 1 µL of digested vector, 3 µL of digested insert DNA, and 2 µL of T4 DNA Ligase (400 U/µL). 5 µL of each ligation was used to transform 50 µL of chemically competent *E. coli* TOP10 cells (Invitrogen). The identities of the resulting constructs were confirmed by sequencing of purified plasmid DNA. These constructs were transformed into chemically competent *E. coli* DH10B cells (Invitrogen) harboring pBelloBAC11-*pks,* pBelloBAC11-*pks*Δ*clbP,* pBelloBAC11-*pks*Δ*clbI* and pBelloBAC11-*pks*Δ*clbP*Δ*clbI* and stored at –80 °C as frozen LB/glycerol stocks.

Large scale overexpression and purification of FabB, FabD and ACP

Expression vectors (described in Ref. 23) were obtained from the Khosla Lab at Stanford University. BL21 (DE3) *E. coli* were transformed with these vectors and overexpression of these proteins was performed as described previously.[23] This procedure afforded yields of 12 mg/L for FabB, 11 mg/L for FabD and 15 mg/L for ACP.

Inactivation of *clbI, clbD and clbG* using λ Red recombinase-mediated gene disruption

Table 4.4: Oligonucleotides used for generation of KO resistance cassette for targeted gene disruption. Regions homologous to the targeted gene are underlined.

Primer Name	Target	Sequence (5' to 3')
ClbI KO F	*clbI*	GGCGTTTCCCTCAAGCCGATACGGTACAGGCGTTTTGGGATTCCGGGGATCCGTCGACC
ClbI KO R	*clbI*	CATGTCGTTAACTAGCACGGCAAGTGCGGACCCTCCATCTGTAGGCTGGAGCTGCTTC
ClbD KO F	*clbD*	CGTTGTGGATATTTCTCAATCTCAGTTGGATAAATGCCGATTCCGGGGATCCGTCGACC
ClbD KO R	*clbD*	CGACCATTTTCTTTAATAAAAAGCTGGGGCGATACTTATTGTAGGCTGGAGCTGCTTC
ClbG KO F	*clbG*	ATGTTCCCTGGCTCCGGTTCGCAATATGTAGGCATGGCAATTCCGGGGATCCGTCGACC
ClbG KO F	*clbG*	TTCCGGATCGGTCTTCACCGCCATGTTATCCCCCAGCACTGTAGGCTGGAGCTGCTTC

The target genes *clbI, clbD* and *clbG* were disrupted by the apramycin resistance gene *aac(3)IV* using the PCR-targeting λ Red recombinase-mediated gene disruption system. The disruption cassette, which included the *aac(3)IV* gene flanked by 39 bp arms homologous to the target gene, was generated by PCR amplification using the primers shown in Table 4.4. PCR reactions contained 25 μL of Q5 High Fidelity 2X Master Mix (New England Biolabs), 0.5 μL of DNA template pIJ773 (digested with HindIII and EcoRI restriction enzymes, and gel purified prior to use in PCR), and 125 pmoles of each primer in a total volume of 50 μL. Thermocycling was carried out in a MyCycler gradient cycler (Bio-Rad) using the following parameters: denaturation

for 120 sec at 94 °C, followed by 10 cycles of 45 sec at 94 °C, 45 sec at 50 °C, 90 sec at 72 °C, followed by 29 cycles of 45 sec at 94 °C, 45 sec at 55 °C, 90 sec at 72 °C, and a final extension time of 5 min at 72 °C. For λ Red recombinase-mediated gene disruption, the purified PCR product was transformed by electroporation into *E. coli* BW25113 harboring the λ Red recombinase expression plasmid pkD46 and either BAC*pks* or BAC*pks*Δ*clbP*. Strains harobaring the apramycin resistance cassette were selected on LB plates containing 50 μg/mL apramycin and 20 μg/mL chloramphenicol. The BACs were isolated from these colonies and gene disruption was verified by PCR. DH10B *E. coli* were transformed by electroporation with these BACs and selected on LB plates containing 50 μg/mL apramycin and 20 μg/mL chloramphenicol.

Biochemical characterization of $ClbB_{PKS}$

BODIPY-CoA fluorescent phosphopantetheinylation assay

This assay was conducted as described in Chapter 2.

ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$ reconstitution assay

The reaction mixture (100 μL) contained 50 mM Tris-HCl pH 7.5, 135 mM NaCl, 10 mM $MgCl_2$, 500 μM DTT, 250 μM L-alanine, 250 μM L-asparagine, 125 μM CoA tri-lithium salt, 500 nM Sfp, and 1.67 % (v/v) DMSO. Loading of phosphopantetheinyl arms onto the T domains of the *apo* enzymes was initiated by the addition of ClbN (10 μM), $ClbB_{NRPS}$ (10 μM) and $ClbB_{PKS}$ (10 μM) to the reaction mixture. This mixture was incubated at room temperature for 30 min, at which point malonyl-CoA (500 μM), NADPH (2 mM), myristoyl-CoA (500 μM), and ATP (2 mM) were added. This mixture was incubated at room temperature for 2 or 4 h. The reaction was quenched by the addition of methanol (250 μL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The protein pellets were washed two times with methanol (250 μL) and dried under a stream of N_2. Products bound to the T domains of ClbN, $ClbB_{NRPS}$ and $ClbB_{PKS}$ were hydrolyzed by adding 0.1 M KOH (20 μL) and heating at 74 °C for 10 min. The

samples were cooled on ice and 0.1 M HCl (40 µL) was added to the solutions. Finally, methanol (200 µL) was added to the samples, which were then incubated at –20 °C overnight to precipitate protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS.

LC–MS (Figure 4.13) was performed on a Bruker Maxis Impact q-TOF using a Phenomenex Gemini C18 reverse phase column (5 µm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 100% solvent A for 2 min, a linear gradient increasing to 100% solvent B over 2 min, 100% solvent B for 10 min, followed by re-equilibration in 10% solvent A for 6 min (solvent A = 95:5 water:methanol + 0.03% ammonium hydroxide; solvent B = 80:15:5 isopropanol:methanol:water; flow rate 0.25 mL/min for the first 4 min and 0.5 mL/min for the remainder of the run). Experiments were performed in negative ion mode.

ClbN, $ClbB_{NRPS}$, $ClbB_{PKS}$, FabB, FabD and ACP reconstitution assay

The reaction mixture (50 µL) contained 50 mM Tris-HCl pH 7.5, 135 mM NaCl, 10 mM $MgCl_2$, 500 µM DTT, 250 µM L-alanine, 250 µM L-asparagine, 125 µM CoA tri-lithium salt, 500 nM Sfp, and 1.67 % (v/v) DMSO. Loading of phosphopantetheinyl arms onto the T domains of *apo* enzymes was initiated by the addition of ClbN (6 µM), $ClbB_{NRPS}$ (6 µM), $ClbB_{PKS}$ (4 µM), and either FabD (6 µM), FabB (6 µM) or ACP (6 µM) or an equal mixture (6 µM) of FabB, FabD and ACP to the reaction mixture. This mixture was incubated at room temperature for 30 min, at which point malonyl-CoA (500 µM), NADPH (2 mM), myristoyl-CoA (500 µM), and ATP (2 mM) were added. This mixture was incubated at room temperature for 2 h. The reaction was quenched by the addition of methanol (250 µL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 15 min). The protein pellets were washed two times with methanol (250 µL) and dried under a stream of N_2. Products bound to the T domains of the

enzymes were hydrolyzed by adding 0.1 M KOH (20 µL) and heating at 74 °C for 10 min. The samples were cooled on ice and 0.1 M HCl (40 µL) was added to the solutions. Finally, methanol (200 µL) was added to the samples, which were then incubated at –20 °C overnight to precipitate protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS.

LC–MS was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 µm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 2% solvent B in solvent A for 2 min, a linear gradient increasing to 100% solvent B over 10 min, 100% solvent B for 5 min, followed by re-equilibration in 2% solvent B in solvent A for 8 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode.

Metabolite analyses of gene knock-out *pks*$^+$ *E. coli* culture extracts

Δ*clbI*

Starter cultures of DH10B *E. coli* (4 mL) harboring pBelloBAC11-*pks* (BAC*pks*), pBelloBAC11-*pks*Δ*clbP* (BAC*pks*Δ*clbP*), BAC*pks*Δ*clbI,* BAC*pks*Δ*clbP*Δ*clbI* and pTrc-ClbI (where indicated in Figure 4.13) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 µg/mL chloramphenicol and 100 µg/mL ampicillin if harboring pTrc-ClbI. These saturated cultures were used to inoculate 50 mL of LB medium containing 20 µg/mL chloramphenicol and 100 µg/mL ampicillin if harboring pTrc-ClbI. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with shaking at 200 rpm. At an OD_{600} of 0.4-0.5 those cultures harboring pTrc-ClbI were induced by the addition of 25 µM IPTG. After a

total growth time of 24 h, the cells were harvested by centrifugation (6000 rpm x 20 min) and lyophilized overnight. The dried cell mass was extracted twice into 2.0 mL methanol by vortexing the mixture for 20 s. The samples were then centrifuged (4000 rpm x 10 min at 4 °C) and the supernatant was transferred to a scintillation vial. 400 μL of the combined methanol supernatant was centrifuged (13 krpm x 15 min at 4 °C) and 300 μL was transferred to a vial for LC–MS analysis.

<u>Δ*clbC*</u>

Starter cultures of DH10B *E. coli* (4 mL) harboring pBelloBAC11-*pks*Δ*clbP* (BAC*pks*Δ*clbP*), Δ*clbP* Nissle 1917 and Δ*clbP*Δ*clbC* Nissle 1917 were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 4 mL of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with shaking at 200 rpm. After a total growth time of 24 h, a 500 μL whole-culture aliquot from each sample was flash frozen in liquid N_2 and lyophilized. The samples was extracted into 500 μL methanol by vortexing the mixture for 20 s. The samples were then centrifuged (13 krpm x 15 min at 4 °C) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis. This experiment was performed with experimental replicates of each condition.

LC–MS (Figures 4.16 and 4.17) was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 2% solvent B in solvent A for 2 min, a linear gradient increasing to 100% solvent B over 10 min, 100% solvent B for 5 min, followed by re-equilibration in 2% solvent B in solvent A for 8 min (solvent A = water + 0.1% formic acid; solvent

B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode.

LC–MS/MS of Metabolite C and D

Starter cultures of DH10B *E. coli* (5 mL) harboring pBelloBAC11-*pks*Δ*clbP* (BAC*pks*Δ*clbP*), pBelloBAC11-*pks*Δ*clbP*Δ*clbD* (BAC*pks*Δ*clbP*Δ*clbD*) or pBelloBAC11-*pks*Δ*clbP*Δ*clbG* (BAC*pks*Δ*clbP*Δ*clbG*) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 50 mL of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 1.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C with shaking at 200 rpm 24 h. At this point, the cells were harvested by centrifugation (6000 rpm x 15 min at 4°C) and the cells were flash frozen in liquid N_2 and lyophilized. The dried cell mass was extracted into 2 mL methanol by vortexing the mixture for 20 s. The samples were then centrifuged (13 krpm x 15 min at 4 °C) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis.

LC–MS/MS (Table 4.1) was performed on a Agilent 6460 Triple Quad LC/MS with Agilent 1290 Infinity HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 50 mm). The following elution conditions were used for this experiment: 40% solvent B in solvent A for 1 min, a linear gradient increasing to 100% solvent B over 4 min, 100% solvent B for 2 min, followed by re-equilibration in 40% solvent B in solvent A for 3 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.3 mL/min). Experiments were performed in positive ion mode and product ion collision energies were varied from 10 to 40 eV. The mass spectrometer was operated in multiple reaction monitoring (MRM) mode with a fragmentor voltage of 155 V (Metabolite D) or 170 V (Metabolite C). The precursor-product ion pairs used in

MRM mode were *m/z* 713→ *m/z* 372 (Metabolite C) (collision energy (CE) = 41 eV) and *m/z* 713→ *m/z* 389 (Metabolite C) (CE = 25 eV); The precursor-product ion pairs used in MRM mode were *m/z* 796→ *m/z* 779 (Metabolite D) (CE) = 21 eV) and *m/z* 796→ *m/z* 455 (Metabolite D) (CE = 41 eV).

Comparative LC–MS

Starter cultures of DH10B *E. coli* (5 mL) harboring pBelloBAC11-*pks*Δ*clbP* (BAC*pks*Δ*clbP*) or pBelloBAC11-*pks*Δ*clbP*Δ*clbG* (BAC*pks*Δ*clbP*Δ*clbG*) were inoculated from frozen cell stocks and grown overnight at 37 °C in LB medium supplemented with 20 μg/mL chloramphenicol. These saturated cultures were used to inoculate 5 mL of LB medium containing 20 μg/mL chloramphenicol. All cultures were inoculated with a normalized number of cells, such that an OD_{600} of 2.0 of the overnight culture gave a 1:100 volume of inoculum. The cultures were incubated at 37 °C on a rotary shaker for 24 h. At this point, 500 μL aliquots were removed from each culture. The aliquots were flash frozen in liquid N_2 and lyophilized. The lyophilized powder was extracted into 500 μL methanol by vortexing the mixture for 20 s. The samples were then centrifuged (13 krpm x 15 min at 4 °C) and 300 μL of the supernatant was transferred to a vial for LC–MS analysis. Experiments were performed in triplicate.

LC–MS (Table 4.2) was performed on an Agilent 6210 ESI-TOF with an Agilent 1100 series HPLC using a Phenomenex Gemini C18 reverse phase column (5 μm, 4.6 x 250 mm). The following elution conditions were used for this experiment: 100% solvent A for 1.5 min, a linear gradient increasing to 100% solvent B over 43.5 min, 100% solvent B for 8 min, followed by re-equilibration in 100% solvent A for 10 min (solvent A = water + 0.1% formic acid; solvent B = acetonitrile + 0.1% formic acid; flow rate 0.4 mL/min). Experiments were performed in positive ion mode. The

chromatographic datasets were aligned by retention time and mass, and these aligned data were statistically analyzed using the XCMS software (https://xcmsonline.scripps.edu).[3]

4.11: References

(1) Vizcaino, M. I.; Engel, P.; Trautman, E.; Crawford, J. M. *J. Am. Chem. Soc.* **2014**, *136,* 9244.

(2) Bian, X.; Plaza, A.; Zhang, Y.; Müller, R. *Chem. Sci.* **2015**, *6*, 3154.

(3) Vizcaino, M. I.; Crawford, J. M. *Nat. Chem.* **2015**, *7*, 411.

(4) Li, Z.-R.; Li, Y.; Lai, J. Y. H.; Tang, J.; Wang, B.; Lu, L.; Zhu, G.; Wu, X.; Xu, Y.; Qian, P.-Y. *ChemBioChem* **2015**, *16*, 1715.

(5) Brotherton, C. A.; Wilson, M.; Byrd, G.; Balskus, E. P. *Org. Lett.* **2015**, *17*, 1545.

(6) Zha, L.; Wilson, M.; Brotherton, C.A.; Balskus, E.P. *in preparation.*

(7) Boger, D. L.; Garbaccio, R. M. *Bioorg. Med. Chem.* **1997**, *5*, 263.

(8) Hamamichi, N.; Natrajan, A.; Hecht, S. M. *J. Am. Chem. Soc.* **2001**, *114*, 6278.

(9) Nougayrede, J.-P.; Homburg, S.; de ric Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J. R.; Dobrindt, U.; Oswald, E. *Science* **2006**, *313*, 848.

(10) Bossuet-Greif, N.; Dubois, D.; Petit, C.; Tronnet, S.; Martin, P.; Bonnet, R.; Oswald, E.; Nougayrede, J.-P. *Mol Microbiol* **2015**.

(11) Kunzmann, M. H.; Sieber, S. A. *Mol. Biosyst.* **2012**, *8*, 3061.

(12) Grogan, D. W.; Cronan, J. E. *Microbiol. Mol. Biol. R.* **1997**, *61*, 429.

(13) Yip, W. K.; Dong, J. G.; Kenny, J. W.; Thompson, G. A.; Yang, S. F. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 7930.

(14) Vaillancourt, F. H.; Yeh, E.; Vosburg, D. A.; O'Connor, S. E.; Walsh, C. T. *Nature* **2005**, *436*, 1191.

(15) Chai, Y.; Pistorius, D.; Ullrich, A.; Weissman, K. J.; Kazmaier, U.; Müller, R. *Chem. Biol.* **2010**, *17*, 296.

(16) Engel, P.; Vizcaino, M. I.; Crawford, J. M. *Appl. Environ. Microbiol.* **2014**, *81*, 1502.

(17) Kevany, B. M.; Rasko, D. A.; Thomas, M. G. *Appl. Environ. Microbiol.* **2009**, *75*, 1144.

(18) Kampa, A.; Gagunashvili, A. N.; Gulder, T. A. M.; Morinaka, B. I.; Daolio, C.; Godejohann, M.; Miao, V. P. W.; Piel, J.; Andrésson, O. S. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E3129.

(19) Zhao, C.; Coughlin, J. M.; Ju, J.; Zhu, D.; Wendt-Pienkowski, E.; Zhou, X.; Wang, Z.; Shen, B.; Deng, Z. *J. Biol. Chem.* **2010**, *285*, 20097.

(20) Kaulmann, U.; Hertweck, C. *Angew. Chem. Int. Ed. Engl.* **2002**, *41*, 1866.

(21) Zhang, J.; Van Lanen, S. G.; Ju, J.; Liu, W.; Dorrestein, P. C.; Li, W.; Kelleher, N. L.; Shen, B. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1460.

(22) Jones, D. T.; Taylor, W. R.; Thornton, J. M. *Comput. Appl. Biosci.* **1992**, *8*, 275.

(23) Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. *Mol. Biol. Evol.* **2011**, *28*, 2731.

(24) Brian O. Bachmann, B.O.; Ravel, J. *Method. Enzymol.* **2009,** *458*, 181.

(25) Del Vecchio, F.; Petkovic, H.; Kendrew, S. G.; Low, L.; Wilkinson, B.; Lill, R.; Cortes, J.; Rudd, B. A. M.; Staunton, J.; Leadlay, P. F. *J. Ind. Microbiol. Biotechnol.* **2003**, *30*, 489.

(26) Ikeda, H.; Nonomiya, T.; Usami, M.; Ohta, T.; Omura, S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9509.

(27) Brautaset, T.; Borgos, S. E. F.; Sletta, H.; Ellingsen, T. E.; Zotchev, S. B. *J. Bio. Chem.* **2003**, *278*, 14913.

(28) Erol, O.; Schäberle, T. F.; Schmitz, A.; Rachid, S.; Gurgui, C.; Omari, El, M.; Lohr, F.; Kehraus, S.; Piel, J.; Müller, R.; König, G. M. *ChemBioChem* **2010**, *11*, 1253.

(29) Datsenko, K. A.; Wanner, B. L. *Proc. Natl. Acad. Sci. U.S.A.***2000**, *97*, 6640.

(30) Summers, R. G.; Ali, A.; Ben Shen; Wessel, W. A.; Hutchinson, C. R. *Biochemistry* **1995**, *34*, 9389.

(31) Yu, X.; Liu, T.; Zhu, F.; Khosla, C. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 18643.

(32) Olier, M.; Marcq, I.; Salvador-Cartier, C.; Secher, T.; Dobrindt, U.; Boury, M.; Bacquie, V.; Penary, M.; Gaultier, E.; Nougayrede, J.-P.; Fioramonti, J.; Oswald, E. *Gut Microbes* **2012**, *3*, 501.

(33)Martin, P.; Marcq, I.; Magistro, G.; Penary, M.; Garcie, C.; Payros, D.; Boury, M.; Olier, M.; Nougayrede, J.-P.; Audebert, M.; Chalut, C.; Schubert, S.; Oswald, E. *PLoS Pathog.* **2013**, *9*, e1003437.

(34) Arthur, J. C.; Perez-Chanona, E.; Muhlbauer, M.; Tomkovich, S.; Uronis, J. M.; Fan, T. J.; Campbell, B. J.; Abujamel, T.; Dogan, B.; Rogers, A. B.; Rhodes, J. M.; Stintzi, A.; Simpson, K. W.; Hansen, J. J.; Keku, T. O.; Fodor, A. A.; Jobin, C. *Science* **2012**, *338*, 120.