Biosynthetic Investigation, Synthesis, and Bioactivity Evaluation of Putative Peptide Aldehyde Natural Products From the Human Gut Microbiota

Biosynthetic Investigation, Synthesis, and Bioactivity Evaluation of Putative Peptide Aldehyde Natural Products from the Human Gut Microbiota

A dissertation presented

by

Benjamin Aaron Schneider

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, MA

April 2019

Dissertation Advisor: Professor Emily P. Balskus Benjamin Aaron Schneider

Biosynthetic Investigation, Synthesis, and Bioactivity Evaluation of Putative Peptide Aldehyde Natural Products from the Human Gut Microbiota

Abstract

Natural products produced by the human gut microbiota may mediate host health and disease. However, discovery of the biosynthetic gene clusters that generate these metabolites has far outpaced identification of the metabolites themselves. Employing traditional culturing and isolation-based natural product discovery techniques in this setting is difficult due to a variety of technical challenges. 'Cryptic' gene clusters identified in these microbes often remain unexpressed under standard laboratory culture conditions, and it can be cumbersome to cultivate the anaerobic bacteria on the scale required for isolation efforts. Several isolation-independent approaches have recently been used to access the products of gut microbial biosynthetic gene clusters, including functional metagenomics, heterologous expression, and synthesis of predicted natural product structures, and the continued development of such approaches may enable a greater understanding of the chemical ecology of the human gut.

In this work, we used an isolation-independent approach that combined biosynthesis and chemical synthesis to access putative products of a family of nonribosomal peptide synthetase(NRPS)-encoding gene clusters from the human gut microbiota. These NRPS gene clusters all contain terminal reductase domains, indicating that they may produce peptide aldehyde products. Such natural products are known to act as inhibitors of serine, cysteine, and threonine proteases. We initially targeted an NRPS gene cluster in this family from the abundant gut commensal microbe *Ruminococcus bromii* (the *rup* gene cluster) for small molecule

discovery. Using a combination of bioinformatic analyses and in vitro biochemical characterization of biosynthetic enzymes, we predicted that this gene cluster likely generates *N*-acylated dipeptide aldehyde(s), collectively referred to as the ruminopeptin(s), and gained insight about the biosynthetic building blocks likely incorporated by its main NRPS enzyme. These results demonstrate the utility of combining bioinformatic predictions with in vitro biochemical assays for predicting the structures of natural products.

We next used a short solution-phase synthesis to access 12 predicted peptide aldehyde structures of ruminopeptin(s). Many of these compounds contain a glutamyl aldehyde at their C-terminus. Therefore, we predicted that they may target the glutamyl endopeptidases, a family of serine proteases found in several human opportunistic pathogens including *Staphylococcus aureus* and *Enterococcus faecalis*. We found that several putative ruminopeptins inhibited the *S. aureus* glutamyl endopeptidase SspA (also known as endoproteinase GluC or V8 protease). We also identified homologs of this protease encoded in the genomes of gut commensals and opportunistic pathogens as well as metagenomes from the human gut environment. We hypothesize that inhibition of this family of proteases by ruminopeptin(s) may be important for mediating microbe-microbe interactions in the human gut.

Inspired by the success of this approach in discovering interesting small molecules with potentially physiologically relevant activity, we next expanded the scope of our study to investigate additional predicted peptide aldehydes from gut microbial genomic data. We selected four additional NRPS gene clusters of interest from gut commensal organisms and synthesized a comprehensive set of their bioinformatically predicted biosynthetic products to reach a total library size of 48 peptide aldehydes. We evaluated these compounds as potential inhibitors of human proteases, antibiotics against a set of prominent pathogens and commensals, and

inhibitors of gut microbial secreted protease activity. Finally, we designed and synthesized an activity-based probe based on the structure of a peptide aldehyde and attempted to identify potential protein targets of this compound in *Clostridioides difficile* 630Δerm using an untargeted chemoproteomics workflow. Small molecule target identification and validation in the complex environment of the human gut is an emerging area of interest, and the approaches we have employed to investigate the biological activities of peptide aldehydes illustrate the current challenges and opportunities in this field.

Table of Contents

Acknowledgments

First, I would like to thank my advisor, Professor Emily P. Balskus. I first met Professor Balskus at a seminar she gave at the University of Pennsylvania when I was an undergraduate, and I remember being excited about assays she described then that I ended up adapting as part of my own work here. Aside from my interest in the work itself, it has also been a great experience to have Professor Balskus as a mentor. I am so grateful for her insightful scientific questions, her eagerness to let me pursue my own research directions and focus on the techniques that I was most interested in learning, and the training she has provided in writing and presentation skills. I also appreciate her cultivation of this lab as an incredible scientific environment in which to work. Because our work is so interdisciplinary, members of this group are skilled in a diverse range of techniques, but equally important is that they are also extremely willing to share their time and expertise with their colleagues. I find Emily's vision for our work on the gut microbiota inspiring, and it has been a privilege to be a part of this lab's story and witness its growth for the past five years. I cannot wait to see the innovation and excellence that I have no doubt will continue to come in the future.

I would also like to express my gratitude to Professor Stuart L. Schreiber and Professor Matthew Shair for serving on my dissertation committee. They have provided helpful advice and direction to my work during my time at Harvard, and I have always looked forward to our meetings. I am grateful to Dr. Sunia A. Trauger, Jennifer Wang, Dr. Shaw Huang, and Claire Reardon for maintaining core facilities at Harvard, to Dr. Richard Novak and Dr. Rachelle Prantil-Baun for screening compounds in several interesting biological assays, and to Dr. Edward Mandell, Professor Harry Flint, and Dr. Sylvia Duncan for helpful conversations.

My journey towards chemistry research began with many impactful teachers at William H. Hall High School in West Hartford, CT, and I am especially grateful to Cindy Caruk and Joan Pease for my first exposures to biology and chemistry wet lab work. As an undergraduate, I was a Roy and Diana Vagelos Science Scholar at the University of Pennsylvania, and the opportunity that program granted me to pursue scientific research in a serious way was instrumental in my decision to pursue graduate studies in chemistry. As part of that program, I was also fortunate to be advised by Professor Ponzy Lu and Professor Jeffery G. Saven. Professor Jeffrey D. Winkler taught my first organic chemistry class in college, and I worked in his laboratory for two years. I am indebted to him for my foundational knowledge of organic chemistry, both on paper and at the bench. I am also grateful to my many mentors in the Winkler lab, including Dr. Barry Twenter, Dr. Mark Nilson, Dr. Ae Jaru, and Dr. Lindsay Leal.

I have learned so much from everyone I've worked with in the Balskus lab, but some specific individuals deserve special mention for the amount of time and attention they've devoted to helping me develop as a scientist. I am thankful to Dr. Hitomi Nakamura, Dr. Abraham Waldman, Dr. Erica Schultz, Dr. Kristen Seim, Dr. Li Zha, and Dr. Matthew Wilson for much assistance and advice in the early days of my graduate school career. Additionally, I am thankful to my colleagues Carina Chittim, Tai Ng, Samantha Cassell, Dr. Monica McCallum, Dr. Paul Boudreau, Dr. Yolanda Huang, Dr. Li Zha, Dr. Maud Bollenbach, Dr. Lihan Zhang, Dr. Spencer Peck, Dr. Lauren Rajakovich, Dr. Michael Luescher, Dr. Nitzan Koppel, Vayu Maini Rekdal, Matthew Volpe, and Nathaniel Braffman for helpful scientific discussions, assistance with editing and proofreading, and all that they do to help the lab run. I am particularly grateful to Doug Kenny for the significant time and effort he devoted to helping me screen some synthetic

compounds in mouse cells. I also wish to thank laboratory administrators Tracey Schaal, Linda Hill-White, and Rhonda Pautler.

I count myself lucky to have such wonderful friends in Cambridge, Somerville, and beyond, and my graduate school experience would not have been the same without them. Thank you to the Cambridge Minyan, Minyan Tehillah, and Tremont Street Shul communities, the Hadar alumni community, my Pardes crew, my Penn friends, and everyone else. I am extremely grateful for all the help and support you have provided me over the past five years. I offer special thanks to Joe-Ann Moser, Becca Goldstein, Jon Gould, Julia Goldberg, Carina Chittim, Li Zha, Doug Kenny, Abraham Waldman, Nathaniel Braffman, Ioana Moga, Ethan Schwartz, Rebecca Peretz-Lange, and Judy Gerstenblith, with whom I have shared the experience of pursuing a doctorate during this time period.

I could not have reached this occasion without the support of my family. I am grateful to my grandparents Linda Miller, Burton Miller, and Hollis Schneider. I am thinking of my grandfather Bernard Schneider, z”l, and wish that he could be here with us. Thank you to my siblings, Abby and Jacob, for providing me with essential support and perspective over the past five years. I am also grateful to my uncles Jeffrey Miller and Steven Miller, my aunt Lorena Miller, and my cousins Noah Miller, Cole Miller, and Freddie Aguilar. I love you all. Finally, I want to thank my parents, Bonnie and Eric Schneider, for their unwavering support and encouragement. You have always believed that I could accomplish my goals and done whatever you could to help me achieve them. I love you.

To my parents

List of Abbreviations

Ac	acetyl
Ala	alanine
Arg	arginine
Asp	aspartate
Asn	asparagine
BLAST	Basic Local Alignment Search Tool
Boc	*tert*-butoxycarbonyl
Ci	curie
d	deuterium
D	dextrorotatory
Da	dalton
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
equiv	equivalent
g	gram
Gln	glutamine
Glu	glutamate
Gly	glycine
h	hour
His	histidine
HPLC	high-performance liquid chromatography

*i*Bu	isobutyl
IC_{50}	half maximal inhibitory concentration
Ile	isoleucine
*i*Pr	isopropyl
IPTG	isopropyl β-D-1-thiogalactopyranoside
J	coupling constant
L	levorotatory
L	liter
LB	lysogeny broth
LC	liquid chromatography
Leu	leucine
Lys	lysine
M	molarity
Me	methyl
Met	methionine
min	minute
mol	mole
MS	mass spectrometry
MW	molecular weight
MWCO	molecular weight cut-off
NAD(P)H	nicotinamide adenine dinucleotide (phosphate)
Nle	norleucine
NMR	nuclear magnetic resonance

ºC	degree Celsius
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
Phe	phenylalanine
ppm	parts per million
Pro	proline
psi	pounds per square inch
psig	pounds per square inch gauge
RNA	ribonucleic acid
rpm	revolutions per minute
s	second
SDS	sodium dodecyl sulfate
Ser	serine
*t*Bu	*tert*-butyl
Thr	threonine
Tris	tris(hydroxymethyl)aminomethane
Trp	tryptophan
Tyr	tyrosine
UV	ultraviolet
v/v	volume/volume
Val	valine
Vis	visible

w/v	weight/volume

1. Introduction: Discovering novel bioactive small molecules from the human gut microbiota

1.1. The human gut microbiota in health and disease

It is an exciting time to study the gut microbiota, the population of microbes living in the human gut. Over the past few decades, many studies have revealed an expanding set of connections between this complex microbial community and various aspects of human health and disease. The commensal gut microbiota provides nutrients to the human host, produces essential vitamins and cofactors, and aids in development of the human immune system.[1–5] However, the gut microbiota can also be correlated with disease, as this community can promote inflammation and serve as a reservoir of pathogenic species that either produce toxins in the gut or invade and infect other body sites.[3,5,6]

In the past several years, a variety of techniques have advanced the questions asked in gut microbiota research, from "who is there?" to "what are they doing?". Several impactful recent research efforts include profiling the immune system effects of colonization with single microbial species in mice[7] to major longitudinal studies of gut microbiota composition under a variety of diets and disease states.[8] Emerging areas of interest include the gut microbiota's contribution to disorders such as autism and depression through the gut-brain axis,[9] the role of the gut microbiota in the immune system as it relates to allergies and auto-immune disorders,[10] and potential therapeutic interventions to alter this community's detrimental effects in conditions such as heart disease and non-alcoholic fatty liver disease.[11] Though correlations between these conditions and the gut microbiota are striking, we still understand very little about the actual mechanisms by which microbes contribute to these phenomena.

The "gut microbiota" is the collection of microbial species living in the gut, while the "gut microbiome" is the collection of genes encoded by these organisms. The development of modern methods of DNA sequencing and genomics have revolutionized the study of the gut microbiome.[5] Over the past decade, several large-scale efforts have been made to sequence whole genomes of large collections of gut microbial strains as well as metagenomic sequences from environmental DNA (eDNA) isolated from human fecal samples. In 2010, the metaHIT consortium (financed by the European Commission) generated a large gene set from sequencing fecal samples of 124 European individuals and identified 3.3 million non-redundant microbial genes from this cohort.[12] In 2012, the NIH Human Microbiome Project Consortium followed suit by assembling a large cohort with sampling across multiple body sites to study the ecology of human-associated microbial communities.[13] Their dataset contained 3.5 terabases of metagenomic sequence data obtained from 242 healthy adults sampled at 15 – 18 body sites. Additionally, whole genomes for 800 reference microbial strains isolated from the human body were sequenced and annotated as part of this effort.[14] A second generation of the HMP study, focusing in part on patients with inflammatory bowel disease (IBD), is now in progress.[15] These large, publicly funded initiatives have been truly enabling for the field, leading to a major increase in the quantity of resources and data available to study the microbiota.[16]

These projects have revealed that the richest human-associated microbial community is found in the gut, particularly the colon. The metaHIT study identified over 1000 distinct gut bacterial species and found that each person carries approximately 160 species in this environment.[12] The colon microbial community in healthy adults is dominated by Firmicutes and Bacteroidetes.[12,13] The initial phase of the HMP analyzed community diversity between samples from different individuals collected at the same body site and concluded that in general, human

microbiota communities cluster based on body site.[13] However, diversity in gut microbiota composition can still be observed among individuals, both within and among geographically distinct populations, and within individuals sampled over different time points.[17]

The physiology of the human gut is significant for understanding how it interacts with the gut microbiota. The human gastrointestinal tract is about 9 m long, and its physical properties vary greatly along this distance. Moving along the gastrointestinal tract, from the stomach to the colon, nutrients become more limited, the environment becomes increasingly anaerobic, the pH neutralizes, and the concentration of host-derived antimicrobial peptides decreases.[18,19] Human digestive enzymes are largely restricted to the stomach and small intestine, while the largest microbial communities are found in the colon.[18] In the colon, a collection of different cell types enforces a stratified separation of most human cells from resident microbes.[19] The innermost layer of the colon is lined with several types of epithelial cells (mainly enterocytes), which are shielded from the contents of the gut by a protective mucus layer (composed of polysaccharides and proteins) that is secreted by goblet cells.[19,20] Beneath this epithelial layer is the lamina propria, which contains connective tissue to support the epithelium and also harbors many human immune cells.[21] Together, the mucus, epithelium, and lamina propria comprise the mucosa, which is the primary site of interaction between the human body and the gut microbiota.[20]

Different bacterial populations are enriched at different locations within the colon environment.[19,22] For example, in mice, differences have been observed in the dominant bacterial species in interfold regions of the mucosa versus the lumen,[23] and certain bacterial populations are enriched in the privileged anatomical site of colonic crypts, which are segregated from the contents of the lumen.[24] There are limits to gaining a high resolution view of the spatial

organization of the gut microbiota in humans, and studies of this environment more often rely on fecal samples, which likely represent predominantly the luminal gut community, rather than colon biopsies.[22] However, even within the lumen, as measured from human fecal samples, significant differences have been found in the distribution of microbial species associated with either undigested, insoluble food particles or in the liquid phase.[25]

1.1.1. The human gut microbiota in health

The gut microbiota plays important roles in health and disease. One of the most well-studied mechanisms by which this community exerts its effects is through the production of short chain fatty acids (SCFAs). These metabolic end products are produced mainly by degradation of carbohydrates undigested by the host.[2] SCFAs are the primary source of energy for colonic epithelial cells[26] and have other interesting roles (for instance, interacting with host signaling pathways to strengthen the intestinal barrier).[27] Butyrate is of particular interest, as this specific SCFA appears to be a privileged scaffold for both anti-inflammatory activity and inhibiting histone deacetylase enzymes.[28,29] Gut microbes also biosynthesize essential vitamins and cofactors for the human host, such as vitamin B_{12} and vitamin K.[1]

As highlighted above, commensal gut bacteria also interact with pathogens to promote the phenomenon of colonization resistance (reviewed by Núñez and coworkers[3] and by Frankel and coworkers[30]). The precise mechanisms of colonization resistance are still being elucidated and debated. However, there is some evidence that antibiotic treatment creates favorable conditions for the colonization of pathogenic species in animal models and humans,[30–33] perhaps by depleting the population of gut commensal microbes.[34,35] Proposed mechanisms of colonization resistance by commensal microbes include direct niche exclusion of pathogenic bacteria (both by

nutrient consumption and geographic site occupation) and induction of anti-pathogen responses in the host.[30] There is also some evidence that antibiotic warfare between microbial species, well documented in the soil environment, also occurs in the gut, with commensal bacteria known to produce bacteriocins (proteinaceous toxins) that inhibit the growth of closely related species.[36]

A significant current question in this field is to what extent the human host selects for a beneficial microbiota. The human host is thought to have mechanisms to distinguish between commensal and pathogenic microbial species, selectively providing nutrients that benefit commensals and in the process excluding pathogenic species.[37,38] However, the host immune system faces a major problem in discriminating between commensal and pathogenic bacteria, which is that on a molecular level, these types of organisms generally look the same.[39] Therefore, commensal microbes have also evolved a variety of mechanisms to promote their tolerance in the gut environment (recently reviewed by Ayres[37]). One such strategy is the production of anti-inflammatory proteins (as in the case of *Faecalibacterium prausnitzii*[40]) or polysaccharides (as in the case of *Bacteroides fragilis*[41]). Another striking example, found among the Bacteroidetes, is the modification of lipopolysaccharide to be resistant to host-derived antimicrobial peptides.[42] Other mechanisms for the promotion of tolerance include modulation of the host immune response through the production of SCFAs, as discussed above, and degradation of host proinflammatory cytokines.[43]

1.1.2. The human gut microbiota in disease

The gut presents a high surface area of interaction between human cells and their "external" environment, and accumulating evidence suggests this body site and the microbes that live within it can also contribute to disease. The concept of "dysbiosis" is widely used to describe a

gut community that is "out of balance," but this concept has recently been criticized as non-scientific or too unspecific to be of practical use.[44] It is also misleading to define "health" and "disease" based on the presence or absence of particular microbial species.[17] A more useful way to think about this distinction may be in terms of functions of the community that could be contributed by a variety of species, as recently reviewed by Huttenhower and coworkers.[17] A healthy microbiota would contain functions that are necessary for microbial life and that lead to beneficial interactions with the host, while an unhealthy microbiota would lack some of the healthy community's functions and add detrimental functions.[17]

Many mechanistic hypotheses for how the gut microbiota contributes to disease are rooted in inflammation, an immune response that normally serves a protective function but that can also become dysregulated.[45] IBD, which is characterized by chronic inflammation in the gut, is one such disorder, and it can be further subcategorized as Crohn's disease (CD) and ulcerative colitis (UC). Both of these disorders are associated with alterations in the gut microbiota (recently reviewed by Nishida and coworkers[46]), particularly the decrease of oxygen sensitive bacteria (such as *F. prausnitzii*) and the increase of oxygen tolerant bacteria (such as *E. coli*).[46] Changes in the metabolite profile of the microbiota may also effect this disease state, either through the reduction of beneficial metabolites (such as SCFAs) or the production of harmful metabolites (such as hydrogen sulfide).[46] Genome-wide association studies have also identified genetic risk alleles for IBD, including genes involved with innate immune recognition of bacteria and autophagy.[47] Along with IBD, roles for inflammation associated with the microbiota have also been proposed in the development of diabetes and obesity.[48,49]

The gut microbiota is also associated with disease through the breakdown of colonization resistance, which allows for the invasion and colonization of pathogenic species in this

environment. A representative example is infection by *Clostridioides difficile*. *C. difficile* spores are found in the environment, particularly in environments such as hospitals, and are also present in the microbiotas of many individuals without inducing symptoms.[50] Antibiotic treatment, which eliminates a large portion of the gut microbial community, opens up niches that pathogens can then colonize, creating an opportunity for *C. difficile* infection (CDI).[51,52] *C. difficile* pathogenesis is mainly characterized by diarrhea that is likely caused by two toxins, toxin A and toxin B.[53] Once an individual has suffered one CDI, additional infections become more likely.[54] The first line therapeutics for recurrent *C. difficile* infection remain metronidazole and vancomycin, antibiotics of last resort that are intended to completely eradicate the problematic spores.[55] However, under the current standard of care, as many as 20–60% of patients with a *C. difficile* infection will experience a recurrence.[54]

1.1.3. Therapeutic interventions involving the gut microbiota

In general, therapeutic interventions that target the gut microbiota remain in their infancy. A significant recent development in this area has been the use of fecal microbial transplantation (FMT) to treat recurrent CDI.[56] This treatment involves recolonization of the infected gut with a healthy donor microbiota, re-establishing colonization resistance, and it is very effective in resolving infections for recurrent CDI.[56] The development of this treatment, from its origins in ancient Chinese medicine to its more recent application to recurrent CDI, is summarized in several recent reviews.[54,57] However, though FMT has recently been recommended in the clinical practice guidelines by the Infectious Disease Society of America for recurrent CDI,[58] this therapy currently faces a complicated regulatory environment.[59] The success of FMT as a remedy for recurrent CDI has inspired investigations in two separate areas: 1) methods to recolonize the gut

in more regulated and reproducible ways (such as the application of "synthetic" therapeutic bacterial cocktails[60]) and 2) using FMT to treat other disorders. Studies have been conducted to explore the efficacy of FMT in treating IBD,[61] and ongoing clinical trials are also investigating its potential to treat obesity, depression, and food allergies.[62] While FMT represents an extreme intervention in remodeling the gut ecosystem, narrow spectrum antibiotics and phage therapy are two additional emerging technologies with the potential to reshape the gut microbial community in a more selective way.[63,64]

Probiotics (live microbial strains) and prebiotics (substrates that promote the growth of specific bacteria) are other methods of modulating gastrointestinal health, but their efficacy is highly debated.[65] Popular probiotic species include various *Lactobacillus* and *Bifidobacterium* species, and some studies have shown efficacy for these treatments in IBD, particularly UC.[66] Probiotics have also been investigated as treatments for obesity, diabetes, and depression, but this work remains in its early stages, and most studies have been limited to very small sample sizes.[65] The use of synthetic biology to engineer probiotics (to produce, for example, useful recombinant proteins in situ) is also an emerging area of interest.[67] In a less definable way than investigations of individual probiotic species, some beneficial health effects have also been demonstrated from ingesting fermented food products, potentially in part because these foods contain live cultures.[68] Finally, prebiotics are hypothesized to function as a nutrient source for beneficial bacteria, and the prototypical members of this class are indigestible carbohydrates. There is strong circumstantial evidence for the benefits of consuming such substrates, as well as dietary fiber more broadly, but more work is needed to determine the mechanistic reasons for these effects.[69–71]

Along with these therapeutic modes, small molecule inhibitors that target gut microbial enzymes are an emerging area of interest, with the eventual goal of using small molecule drugs to prevent undesirable microbial metabolism. Inhibition of gut microbial β-glucuronidases has been investigated by Redinbo and coworkers.[72] These widespread bacterial enzymes liberate a sugar molecule from a glucuronidated, host-derived metabolite of the cancer drug irinotecan (presumably for nutritional purposes), but in the process release the active molecule in the incorrect body site, leading to dose-limiting diarrhea. Inhibitors of β-glucuronidases that could be dosed alongside the drug in order to target these gut microbial enzymes and prevent this undesirable side effect are therefore under active investigation.[73] Along with inhibiting microbial metabolism that leads to toxicity of exogenous drugs, another area of interest is the inhibition of microbial primary metabolism that generates potentially harmful metabolites. An example of this idea is provided by Hazen and coworkers in their work on inhibitors of choline trimethylamine (TMA) lyase.[74,75] Though there are no small molecule therapeutics currently used in the clinic with the goal of targeting specific gut bacterial enzymes, this is an exciting area for drug development.

1.1.4. Challenges associated with studying the gut microbiota

There are several challenges associated with the study of the gut microbiota. First is the difficulty of culturing organisms from this environment. A 2011 study by Gordon and coworkers showed that approximately half of the operational taxonomic units (OTUs, comprising microbes with >97% sequence ID of their 16S rRNA genes, which is a commonly used criterion to demarcate microbial species) could be isolated and grown in the laboratory.[76] Clavel and coworkers reached a similar conclusion in 2017 through a meta-analysis of the literature and

their own bioinformatics workflow, and they estimated that 35 – 65 % of molecular species in the human gut microbiota have been cultured.[77] In the past several years, many techniques for culturing the "unculturable" organisms from this population have been developed. A 2016 study by Lawley and coworkers developed a workflow to isolate and culture the spore-forming fraction of the gut microbiota through targeted phenotypic culturing and managed to isolate 137 distinct bacterial species, including 45 candidate novel species and 90 species which had previously been uncultured.[78] Another 2016 study by Raoult and coworkers relied on "culturomics" to prioritize the isolation of new bacteria. In this workflow, MALDI-TOF or 16S rRNA sequencing were used to identify growing colonies in a high-throughput way and prioritize previously uncultured organisms. These colonies were then screened in an exhaustive number of culture conditions, some containing unusual additives, to arrive at conditions for growing 385 previously uncultured organisms from the gut microbiota.[79]

Sequencing-based technologies can also be used to provide information about the gut microbiota, including organisms that have not been cultured (recently reviewed by Forbes-Blom and coworkers[5]). Classically, 16S ribosomal RNA sequencing has been used to profile the members of this community.[5] However, phylogeny does not necessarily correlate with microbial function, and significant genetic differences can be observed even among strains of the same species.[80] (Meta)genomic sequencing can therefore provide more detail about the potential functional roles of gut microbes. This technique has been instrumental in profiling the "core" functions of the healthy gut microbiota, as discussed above.[12,13] The reconstruction of genomes from metagenomic sequence information also continues to prove valuable for the identification of novel species in this environment, as Segata and coworkers recently demonstrated in their report on over 150,000 new human-associated microbial genomes (across many body sites).[81]

There are still major limitations in moving from metagenomic sequencing information to prediction of microbial functions, due to the challenge of annotating these genes. From the initial phase of the HMP, for example, Huttenhower and coworkers concluded that "roughly 50% of genes in the gut microbiomes of HMP participants … could not be characterized using standard annotation methods."[13,82] Moreover, the annotations that have been assigned are often not helpful for identifying the specific activities of individual enzymes.[13,82]

Recent studies in the Balskus laboratory, in collaboration with the Huttenhower group at the Harvard School of Public Health, have reported a method to glean additional information about functions of prominent microbial enzyme families from metagenomes.[83] Termed "chemically guided functional profiling," this workflow combines understanding of biochemically characterized enzyme functions, sequence similarity networks, and quantification of metagenomic short reads to highlight clusters of enzymes in a particular family that have yet to be characterized but are particularly abundant in metagenomes. This technique was successfully applied to the glycyl radical enzyme family, a group of enzymes that are highly similar in sequence but catalyze a remarkable and ever-expanding diversity of unique chemistries.[84] There has been a long-standing appreciation that these enzymes are enriched in the human gut,[85] but they have just started to be broadly characterized. CGFP was used to highlight a new glycyl radical enzyme in the human gut microbiota, 4-hydroxyproline dehydratase, which was biochemically characterized and is likely involved in metabolism of host-derived and dietary collagen.[83]

Functional metagenomics is another technique that has been exploited to discover novel functions of the gut microbiota from uncultured organisms.[86,87] In this technique, large amounts of environmental DNA are digested and cloned into a vector that is then transformed into a

suitable expression host to generate a metagenomic library.[86] These clone libraries can then be screened for a variety of phenotypes, and the plasmids or fosmids responsible for generating those phenotypes can be isolated and sequenced. Functional metagenomics in the gut environment has been used to discover many new activities of this community, including mechanisms of antibiotic resistance,[88] natural product production,[89] and bile salt hydrolase activity,[90] as well as for the attempted discovery of novel protease activity.[91] Limitations of this technique include the large number of clones that must be screened to identify a phenotype of interest and the need for the bacterial enzyme of interest to be properly expressed and folded in the non-native host.[92,93]

Recently, additional "omics" techniques have attracted interest as another way of uncovering functions of the gut microbiota. For example, in the second iteration of the Human Microbiome Project, the iHMP,[94] metatranscriptomics has been used to highlight particular species and pathways that are associated with the development of IBD.[15] In this study, gene expression was tracked over time in multiple samples from over 100 individuals (both IBD patients and non-IBD controls) over the course of one year. This work revealed transcriptional variation of specific bacteria between IBD and non-IBD patients, including global differences in transcription of *Ruminococcus gnavus* and differences in specific pathways in *Bacteroides vulgatus*.[15] Metaproteomics has also been used to study the gut microbiota in humans and model organisms,[95,96] and further integration of metatranscriptomics and metaproteomics should enable a greater understanding of abundance and fluctuations of microbial functions in this environment.[97]

Overall, though the roles of the gut microbiota in human health and disease have attracted major interest over the past several years, there remains much to be discovered about the

mechanistic features that make a particular microbial species useful or detrimental to the human host. There is a continued need to investigate not only the phylogenetic and metagenomic makeup of the gut microbial community but also its biochemical capabilities. The production of bioactive small molecule natural products by the gut microbiota has lately attracted much interest,[98–100] and in this dissertation, we have investigated the capacity of certain gut microbial species to produce such compounds. Our goal was not necessarily to discover compounds that would be useful as therapeutics, but to gain inspiration from molecules already produced in this community and the interactions they may mediate. As discussed in the next section, there is clearly an opportunity for more investigation of the small molecules produced in this environment, as diffusible small molecules provide one of the most obvious hypotheses for how this microbial community could exert effects on human health.

1.2. Small molecules from the gut microbiota and their potential physiological roles

A major way that microbes interact with surrounding environments is through the production of small molecules, and small molecules produced by the human gut microbiota are potential mediators of host health and disease.[99] Natural small molecules can be divided into two categories: primary metabolites, which are directly involved in an organism's life cycle, and secondary metabolites (natural products), which serve more specialized roles. The gut microbiota is already known to produce many bioactive primary metabolites that influence host health, including SCFAs, sphingolipids, enterolignans, tryptamine, trimethylamine, and polyamines.[101] Though genome and metagenome sequencing continue to reveal that human gut microbes have a rich biosynthetic potential, as evidenced by gene clusters encoding natural product biosynthetic pathways, discovering natural products from these organisms has proven challenging, in part

because many cannot be cultivated in the laboratory. Moreover, investigations to date have found that gut microbial natural products are often difficult or impossible to isolate or are not produced under standard laboratory conditions.[99,102,103] Therefore, a variety of alternative methods have also been explored to access these potentially bioactive molecules (Figure 1.1).

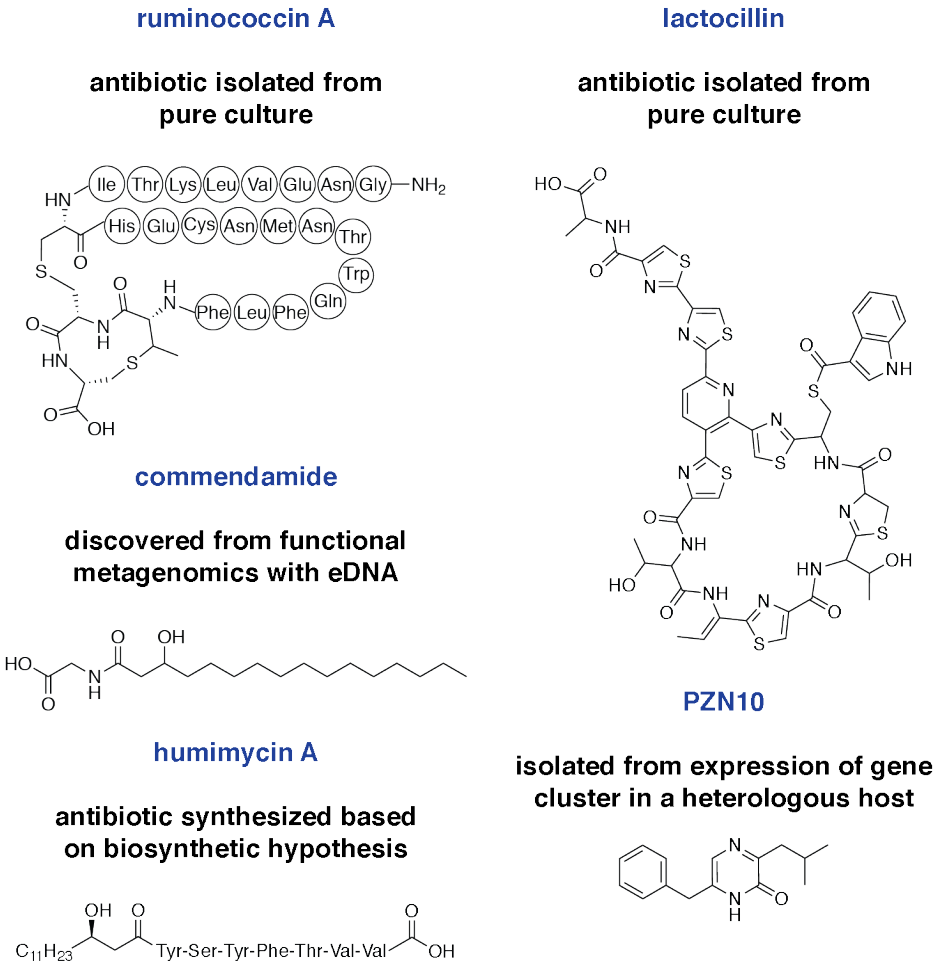


Figure 1.1: Small molecules discovered from human-associated microbial species by a variety of methods.

Selected natural products from human gut bacteria, including the proposed structure of the lanthionine-containing bacteriocin (lantibiotic) ruminococcin A, the thiopeptide antibiotic lactocillin, PZN10, the antibiotic humimycin, and the bacterial effector/GPCR ligand commendamide.

Several examples exist of isolated small molecules produced by gut microbes. Ruminococcin A is a lantibiotic (lanthionine-containing bacteriocin) that was isolated from a pure culture of *Ruminococcus gnavus*, an abundant gut commensal, and had its structure partially determined by LC-MS analysis by Fons and coworkers (Figure 1.1).[104] This compound is active as an antibiotic against a variety of pathogenic *Clostridium* species, as well as other *Ruminococcus* species. After discovery of this compound, a subsequent investigation aimed to determine if it was actually produced in vivo. This same group used RT-qPCR to demonstrate that the RumA precursor peptide was only present at low levels in rat caecal contents, and they therefore suggested that ruminococcin A was likely not responsible for the antibacterial activity of these samples. Therefore, they targeted rat caecal contents for small molecule discovery and isolated an additional compound, ruminococcin C, and also partially determined its structure by LC-MS.[102] This compound is active as an antibiotic against *Clostridium perfringens* and certain strains of *Bacillus cereus* and *Listeria monocytogenes*. Overall, these studies showed that it is possible to isolate antimicrobial peptides from species found in the gut and even directly from gut contents. However, they relied on the targeted search for compounds from particular species, and their method did not allow for complete structural determination due to the low quantities of material obtained.

In 2014, Fischbach and coworkers conducted a systematic survey of the biosynthetic potential of the human microbiome across various body sites.[105] Using the bioinformatic tools ClusterFinder[106] and antiSMASH,[107] they identified biosynthetic gene clusters in ~2,500 HMP sequenced genomes and then quantified the presence of ~3,000 of these clusters in ~750 HMP metatranscriptomes. One of the significant findings in this work was that gene clusters predicted to produce a subclass of ribosomally synthesized, posttranslationally modified peptides (RiPPs),

thiopeptides, were found in the microbiomes of all examined major body sites (skin, oral, gut, and urogenital). They subsequently isolated one of these thiopeptide compounds from a member of the vaginal microbiota. A 50 L culture of *Lactobacillus gasseri* was grown and extracted to yield lactocillin, a thiopeptide antibiotic which was subsequently purified and characterized by NMR (after modification of the free carboxylic acid with TMS-diazomethane to make the isolated compound more stable) (Figure 1.1). Purified lactocillin was tested as an antibiotic against common pathogens and commensals found in the vaginal microbiota, and it was shown to be effective against *Staphylococcus aureus*, *E. faecalis*, *Gardnerella vaginalis*, and *Corynebacterium aurimucosum*. The gene cluster producing this compound was actively transcribed in human samples, albeit in metagenomes from the oral environment and not the vaginal environment where the species it was isolated from normally resides.[105] This result indicates the potential physiological relevance of production of this antibiotic by the human microbiota.

Direct attempts to isolate compounds from the microbiota are laborious, and there is no guarantee that the isolated molecules will demonstrate physiologically relevant bioactivity. Therefore, other strategies, such as functional metagenomics, have been used to prioritize investigation of biosynthetic pathways in this environment.[89] In a 2015 study, Brady and coworkers generated a metagenomic library from 3,000 MB of environmental DNA isolated from human fecal samples and then screened this library for activators of NF-κB, a transcription factor involved in a wide variety of cellular processes. From this study, they identified several commensal bacterial effector genes that activated NF-κB. Based on bioinformatic analysis and metabolite extraction, they determined that one of these genes produced the biosynthetic product commendamide, an *N*-acylated amino acid (*N*-acyl-3-hydroxy-palmitoyl glycine) (Figure 1.1).

Based on structural homology, the authors hypothesized that this compound would be a GPCR ligand, and from a screen of human GPCRs they discovered that it activates GPCR132/G2A. A subsequent study from this group identified additional genes encoding enzymes that produce similar compounds and found them to be enriched in gastrointestinal bacteria, suggesting that these compounds may serve as ligands for a wide diversity of human GPCR's in vivo.[108]

Another method that has been used to accelerate the discovery of small molecules from the gut microbiota is the expression of biosynthetic gene clusters in heterologous hosts.[109] Fischbach and coworkers recently used this technique to investigate a family of nonribosomal peptide synthetase (NRPS) gene clusters from the gut microbiota. In a 2014 study, this group recognized the abundance of a particular family of NRPS-encoding gene clusters in the commensal gut microbiota,[105] and in a 2017 study they expanded this investigation and attempted to determine the biosynthetic products of these gene clusters.[109] Through heterologous expression of 14 of these gene clusters in two different heterologous hosts, the authors were able to identify candidate metabolite products from seven gene clusters. The observed metabolites were mostly cyclic pyrazinone products, which are presumably the cyclization and oxidation products of physiologically produced peptide aldehydes (Figure 1.1). The authors synthesized several compounds mimicking these potential precursors and demonstrated their activity as protease inhibitors towards several human proteases, including calpain and the cysteine cathepsins, which may be involved in host immune response.[109] The work described in this dissertation addresses this same family of NRPS-encoding gene clusters, and a more extensive discussion of this work by Fischbach and coworkers and comparison with our own work is presented in the conclusion of Chapter 4.

The explosion of gut microbial sequence data and high-quality tools for bioinformatic analysis and structural prediction have also allowed for even more high-throughput methods of small molecule discovery from the gut environment. Brady and coworkers recently demonstrated a novel strategy (the "synthetic-bioinformatic natural products", or syn-BNPs, approach) in their discovery of humimycin A (Figure 1.1).[110] By mining sequenced genomes from the human microbiota for NRPS gene clusters, predicting the structures of the likely gene cluster products using bioinformatics, and synthesizing 25 predicted nonribosomal peptides, they accessed a new antibiotic that is active against methicillin-resistant *S. aureus* clinical isolates.[110]

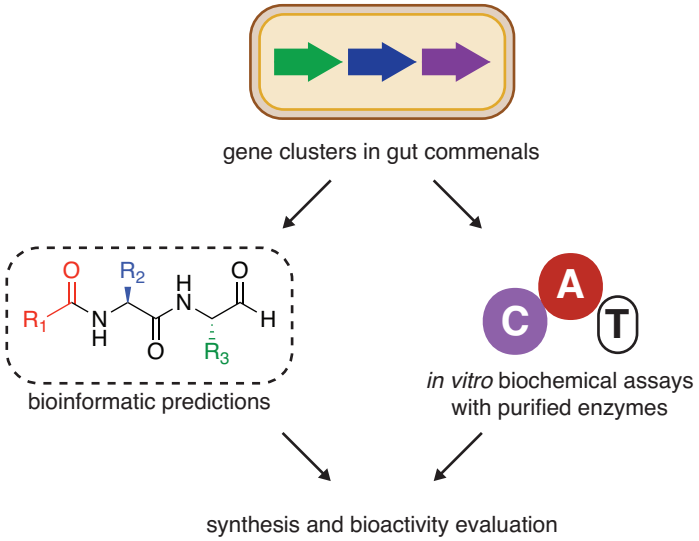


Figure 1.2: Our approach for isolation independent small molecule discovery in the gut microbiota.

Our isolation-independent workflow for characterizing small molecules produced by important gut commensals involves first selecting NRPS-encoding biosynthetic gene clusters of interest based on abundance in metagenomic sequencing data and microbial ecology. Bioinformatic predictions and in vitro biochemical assays then provide structural information that informs the chemical synthesis of candidate natural product structures. These focused small molecule libraries can then be evaluated for bioactivity.

Overall, there is a continued interest in approaches to provide rapid access to products of gut microbial gene clusters. In this work, we envisioned a strategy for accessing gut microbial secondary metabolites that would combine in vitro characterization of biosynthetic enzymes with chemical synthesis (Figure 1.2). By mining human gut metagenomic sequence data, we could identify small NRPS biosynthetic gene clusters of interest based on metagenomic sequencing data and microbial ecology. We could then test our predictions and identify key biosynthetic building blocks using in vitro biochemical assays with purified biosynthetic enzymes. Finally, we would access the candidate natural product structures using chemical synthesis and evaluate these focused small molecule libraries for bioactivity. A key advantage of this approach is that it could provide a more rapid way to access bioactive small molecules compared to traditional isolation- or heterologous expression-based natural product discovery. NRPS biosynthetic gene clusters are very amenable this approach, as these enzymes share a predictable chemical logic. An understanding of how these enzymes work provides the foundation for the bioinformatic analysis of gene clusters, protein biochemistry of NRPS enzymes, and structural prediction of gene cluster products that we discuss in this dissertation.

1.3. Biosynthetic chemistry of nonribosomal peptide synthetase (NRPS) enzymes

1.3.1. Canonical NRPS biosynthetic chemistry

NRPS enzymes are a major class of biosynthetic enzymes that are responsible for producing the molecules that are the focus of this work. These enzymes, along with the closely related polyketide synthetases, are often termed “assembly line enzymes.” NRPS enzymes produce peptidic natural products from amino acids in a process that is completely distinct from ribosomal peptide synthesis, and the essential details of how these enzymes work have been

determined over the past several decades (reviewed by Walsh and Fischbach[111]). Canonical NRPS chemistry is performed cooperatively by several different enzymatic domains, which can be organized into functional units called modules (Figure 1.4). These enzymes perform processive biosynthetic steps on intermediates that are tethered to the protein with thioester linkages. In NRPS pathways, the protein domains that carries these thioesters are known as peptidyl carrier protein (PCP) or thiolation (T) domains. Phosphopantetheine (ppant) transferases are enzymes that perform this post-translational modification, with Sfp from *Bacillus subtilis* being the canonical example.[112] These enzymes catalyze the nucleophilic attack of a conserved residue in the T domain onto coenzyme A in order to generate the ppant arm (Figure 1.3).

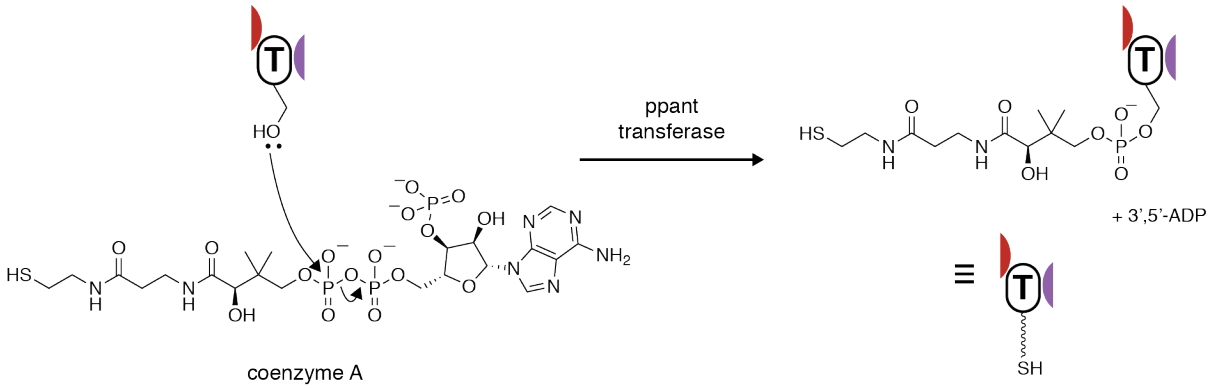


Figure 1.3: Phosphopantetheine (ppant) transferases install the ppant arm onto NRPS thiolation (T) domains.

A conserved serine residue in NRPS thiolation (T) domains is post-translationally modified with coenzyme A to generate the ppant arm, which contains the thioester used to tether intermediates to these carrier proteins in NRPS biosynthetic pathways.[111] The structure of this ppant arm is abbreviated as shown.

The basic mechanisms of some canonical NRPS domains are presented in Figure 1.4A–C. NRPS adenylation (A) domains use general base catalysis to activate specific amino acids with adenosine triphosphate (ATP), generating activated aminoacyl adenylates and pyrophosphate.

The activated aminoacyl-adenylate is then loaded onto a ppant-ylated T domain as a thioester (Figure 1.4A). Peptide bond formation on NRPS assembly lines is accomplished by condensation (C) domains, which use general base catalysis to effect nucleophilic attack of a downstream tethered amino acid on the thioester linkage of an upstream tethered amino acid, forming a peptide bond (Figure 1.4B). Termination of NRPS assembly lines is canonically accomplished by thioesterase (TE) domains, which catalyze the hydrolysis of thioesters, resulting in the release of free carboxylic acid products (Figure 1.4C). Alternatively, TE domains can also catalyze cyclization of longer peptide substrates with intramolecular nucleophiles to generate macrolactones, macrothiolactones, and macrolactams (for example, in the biosynthesis of surfactin,[113] thiocoraline,[114] and tyrocidine,[115] respectively).[116] The hypothetical NRPS assembly line shown in Figure 1.4D introduces a typical notation for these enzymes, where the tethered intermediate resulting from the action of each module is shown on each T domain.

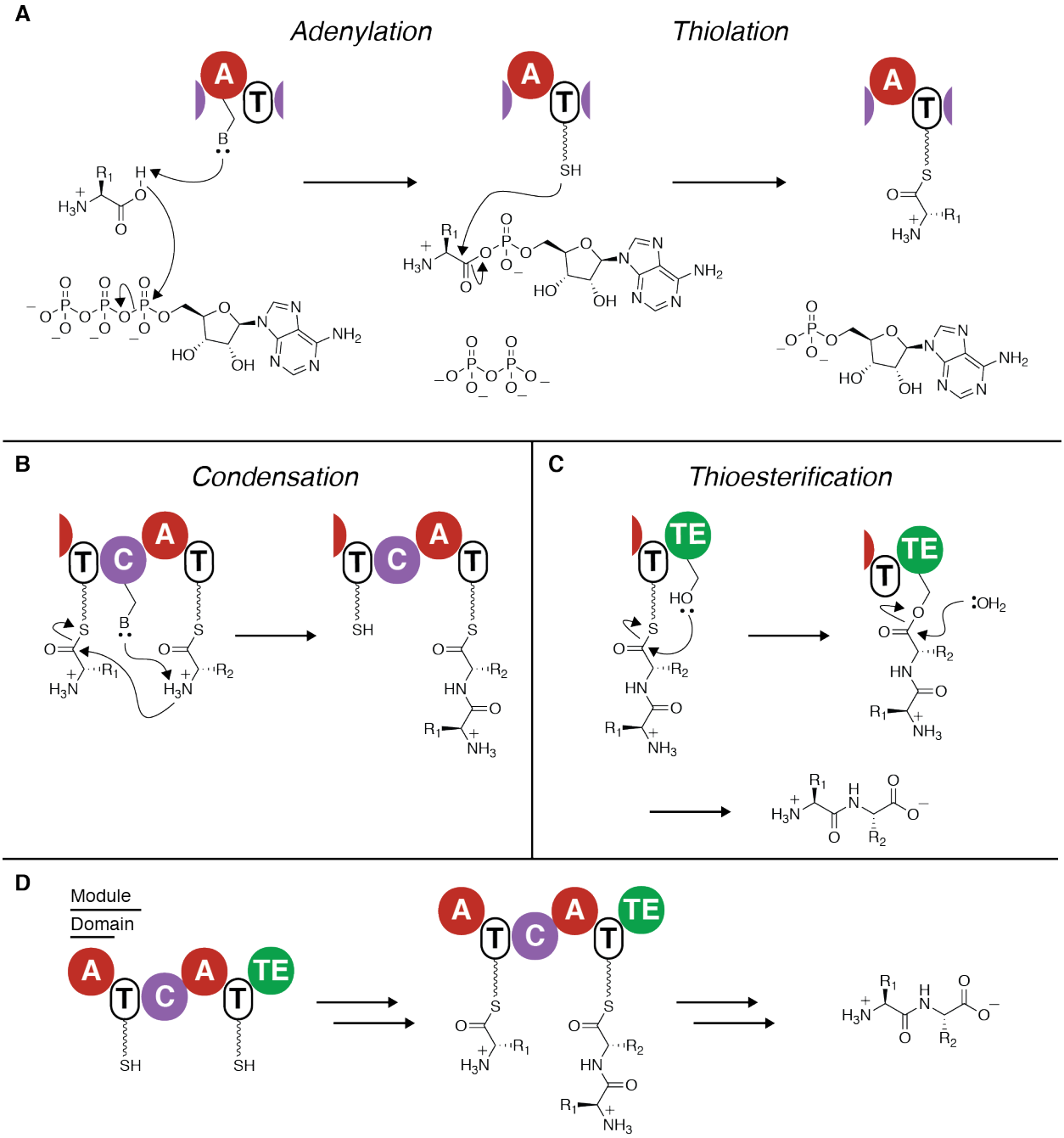


Figure 1.4: Representative canonical NRPS mechanisms, and a simple hypothetical assembly line producing a dipeptide.

(A–C) Basic mechanisms of the canonical domains of NRPS assembly line enzymes.[111] (A) Adenylation domains activate an amino acid by catalyzing the formation of an aminoacyl-adenylate. This activated intermediate is then tethered onto a thiolation domain. (B) Condensation domains catalyze nucleophilic attack by the downstream tethered amino acid on the upstream tethered intermediate to form a peptide bond. (C) Thioesterase domains catalyze hydrolysis of the tethered

Figure 1.4 (continued)

intermediate to release the free dipeptide. (D) A representative di-modular NRPS assembly line producing a simple dipeptide. (A = adenylation domain, T = thiolation domain / peptidyl carrier protein (PCP), C = condensation domain, TE = thioesterase domain.)

1.3.2. Mechanisms for generating structural diversity in NRPS biosynthesis

Structural diversity of NRPS products arises primarily from the ability of A domains to selectively incorporate different amino acids. A domain binding pockets selectively activate particular amino acids based on steric and electronic considerations, and 10 key residues that interact with amino acid substrates to confer this specify were identified by Stachelhaus and coworkers in 1999.[117] Since then, many bioinformatics tools have been developed that allow for the prediction of the specificity of these domains, including NRPSPredictor2,[118] Minowa,[119] and antiSMASH.[120] A domains are usually specific for one amino acid, but there are known examples of these domains activating multiple similar amino acids as well.[121,122] In addition to the proteogenic amino acids, A domains have been discovered that catalyze activation of a remarkable diversity of substrates, including β-amino acids,[123] halide-containing amino acids,[124] polyketide synthase-derived amino acids,[125] aryl acids,[126] and *S*-adenosylmethionine.[127] Several other mechanisms exist for generating structural diversity within the canonical NRPS framework, including additional types of domains that are incorporated into the assembly line. These domains include cyclases and oxidases, which generate oxazoles and thiazoles from serine, threonine and cysteine residues, as well as epimerases, which interconvert L- and D-amino acids. Additionally, many NRPS products are subject to post-assembly line tailoring by oxidative enzymes, ligases, glycosyltransferases, or others that can convert these products into more complicated scaffolds.[111]

Another major source of structural diversity in NRPS's are various modifications introduced at the N-terminus of the peptide, including by unusual C domains. A relatively common modification to the canonical pathway is the inclusion of a C-starter domain rather than an A-T loading module (Figure 1.5). C-starter domains use general base catalysis to effect *N*-acylation of tethered amino acids by activated fatty acids, and these domains can operate on several types of substrates: either by appending pre-activated fatty acyl CoA's onto tethered amino acids (as shown in Figure 1.5A) or accepting fatty acyl thioesters bound to trans-acting T domains.[128] Along with their role in producing *N*-acylated products (such as glidobactin), C-starter domains are also involved in more exotic modifications such as the ureido linkage found in syringolin A (Figure 1.5B).[129,130] Recent examples of notable natural products produced with C-starter domains include the candidate precolibactin prodrug motif,[131] which is involved in the biosynthesis of this genotoxin produced by *Escherichia coli*, as well as icosalide, which is produced by a unique incorporation of two β-hydroxy acids by two C-starter domains in the same multi-modular NRPS protein (Figure 1.5B).[132]

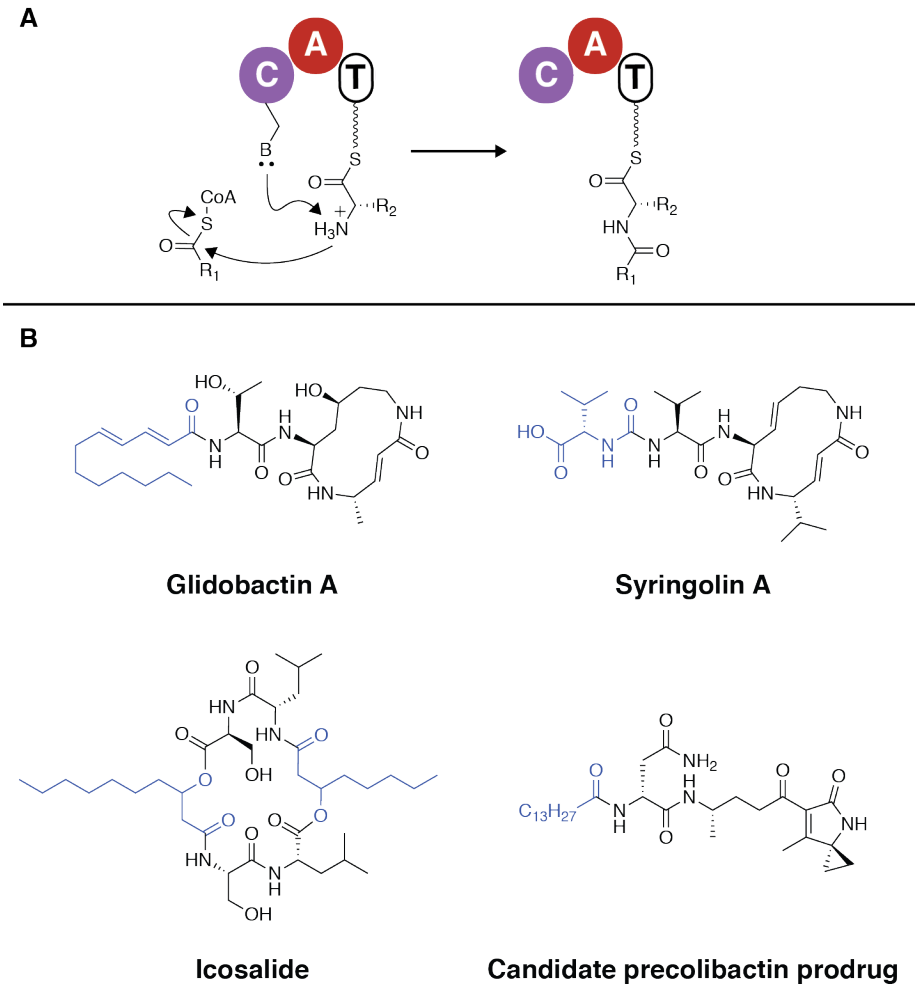


Figure 1.5: C-starter domains generate structural diversity in NRPS- and polyketide synthase(PKS)-derived natural products.

(A) Basic mechanism of C-starter domains.[133] (B) Natural products involving C-starter domains in their biosynthesis. Motif(s) installed by the action of C-starter domains are highlighted in blue.

1.3.3. Terminal reductase (R) domains in NRPS biosynthesis

Another major source of structural diversity in NRPS natural products comes from alternative assembly-line release mechanisms.[134] Terminal reductase (TR or R) domains use NADH or NADPH cofactors to generate terminal aldehydes and alcohols on nonribosomal peptide products (recently reviewed by Thomson and coworkers[135]). These domains are

evolutionarily related to the short-chain dehydrogenase/reductase (SDR) enzymes and share features important for NAD(P)H binding.[136] Reduction by these domains is dependent on a catalytic triad composed of tyrosine, threonine and lysine residues.[137] In the reduction step, hydride is transferred from NAD(P)H to the tyrosine-stabilized thioester carbonyl, forming a thiohemiacetal intermediate (Figure 1.6A). Collapse of this intermediate, with elimination of the thiol, leads to release of the aldehyde product (Figure 1.6A).[135,137,138]

Terminal R domains are also capable of performing a second $2e^-$ reduction on a free aldehydes to generate primary alcohol products.[135,137] This reduction proceeds by a similar mechanism, preceded by unbinding of the oxidized nicotinamide cofactor and rebinding of another reduced cofactor. In this second reduction, the resultant alcohol generated by hydride transfer retains the proton from tyrosine (Figure 1.6A).[135,137] In pathways where it occurs, the second reduction is as much as 15 times more efficient than the first reduction.[138]

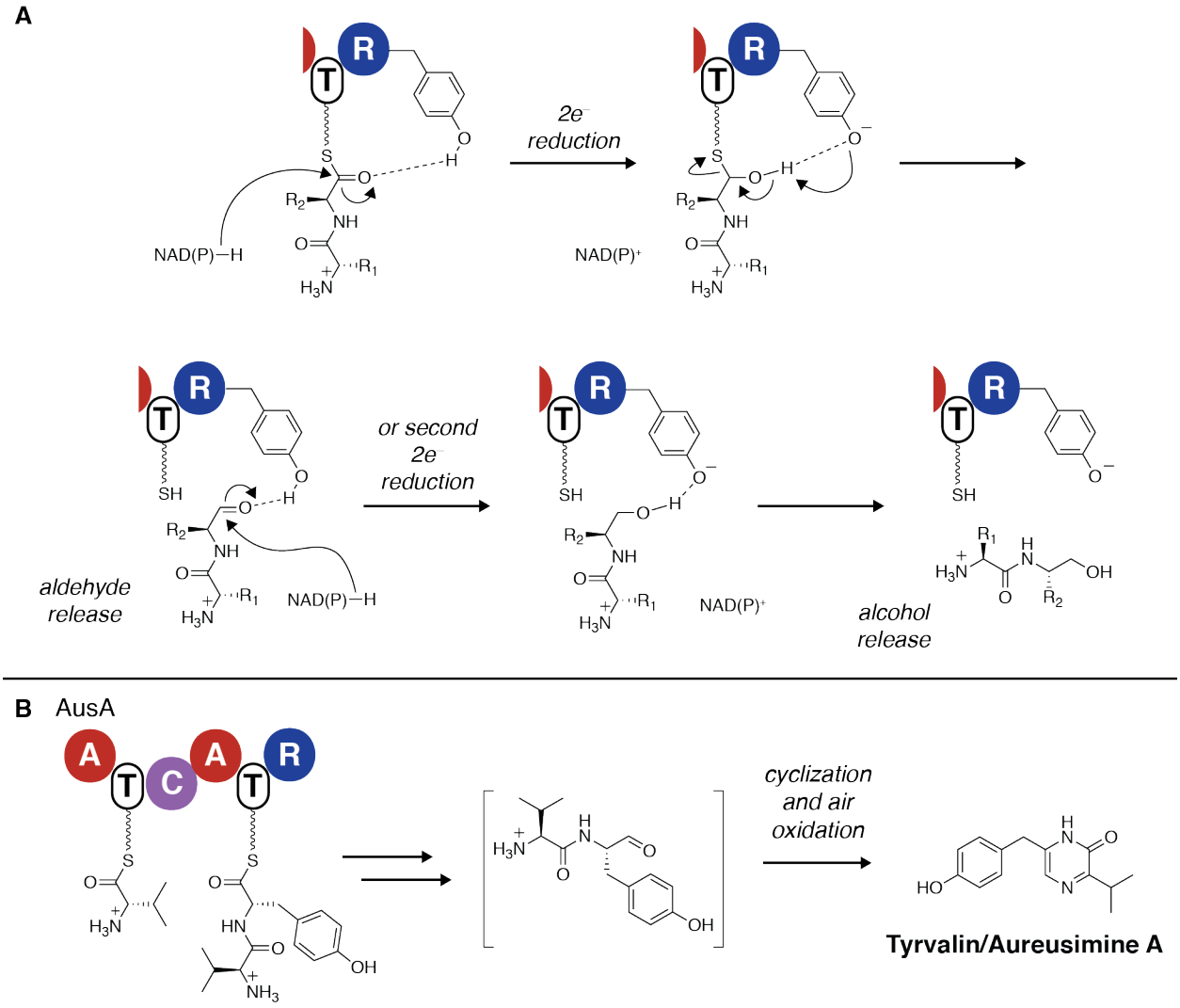


Figure 1.6: Thioester reductase (R) domains catalyze 2e⁻ and 4e⁻ reduction of NRPS- and PKS-tethered thioesters.

(A) Basic mechanism of a thioester reductase (R) domain performing a 2e⁻ reduction on a tethered thioesters.[135] Relying on a key tyrosine residue in the R domain, the domain uses NAD(P)H to reduce the tethered thioester to a thiohemiacetal intermediate. This intermediate then collapses to afford the free peptide aldehyde. In some cases, these domains perform a subsequent 2e⁻ reduction on the aldehyde product, again relying on NAD(P)H and the key tyronine residue to generate alcohol products. (B) Tyrvalin biosynthesis is similar to biosynthesis of the simple dipeptide shown in Figure 1.4. The free dipeptide aldehyde produced by AusA likely spontaneously cyclizes and oxidizes to form the pyrazinone product tyrvalin/aureusimine A (R = thioester reductase domain).[139]

Most of the biosynthetic steps in pathways involving an R domain remain the same as in canonical NRPS chemistry, as exhibited by the example of tyrvalin (aureusimine A) biosynthesis by the NRPS AusA (Figure 1.6B). This simple pyrazinone, the spontaneous cyclization and oxidation product of a precursor dipeptide aldehyde, was isolated from *S. aureus*[140,141] and may play a role in this organism's survival within epithelial cells and phagocytes.[142] Two related compounds that replace the tyrosine residue with phenylalanine (phevalin) and leucine (leuvalin) are also known.[140] In tyrvalin biosynthesis, the loading module activates L-valine with ATP, generating the tethered thioester intermediate. The extension module activates L-tyrosine and forms a peptide bond, leading to the tethered dipeptide. Finally, using NADPH as a cofactor, the terminal domain releases this compound as the aldehyde.[139] In the case of tyrvalin, the thermodynamically favorable formation of a cyclic imine, followed by spontaneous air oxidation of the compound to the pyrazinone, affords the product (Figure 1.6B).[139]

Many interesting natural products are produced through $2e^-$ and $4e^-$ reductions catalyzed by R domains (Figure 1.7).[135] Nostocyclopeptide M1 is a macrocyclic imine cytotoxin produced by *Nostoc* sp. ATCC53789, generated by the cyclization of a peptide aldehyde.[143,144] Anthramycin, an antitumor antibiotic isolated from *Streptomyces spp.*, is the product of nucleophilic attack of the nearby aniline moiety onto the aldehyde moiety of its precursor.[145,146] An interesting molecule with a potential relevance for human health comes from an impressive 2012 study by Khosla and coworkers.[147] Infectious *Norcardia* strains share a large multi-modular PKS gene cluster that terminates in an R domain, but no natural products have been isolated from these strains. The authors in this study used in vitro biosynthetic reconstitutions to generate partial structures of the putative natural product produced by this gene cluster, leading to hypothesized aromatic aldehyde compounds that may have a role in the pathogenesis of these

strains.[147] As mentioned above, R domains are also capable of performing two processive 2e⁻ reductions to generate primary alcohols. Some notable hybrid NRPS/PKS products containing alcohols generated by these domains include myxochelin A, a siderophore produced by the myxobacteria,[148–150] and myxalamid A, which is produced by the myxobacterium *Stigmatella aurantiaca*.[137,151]

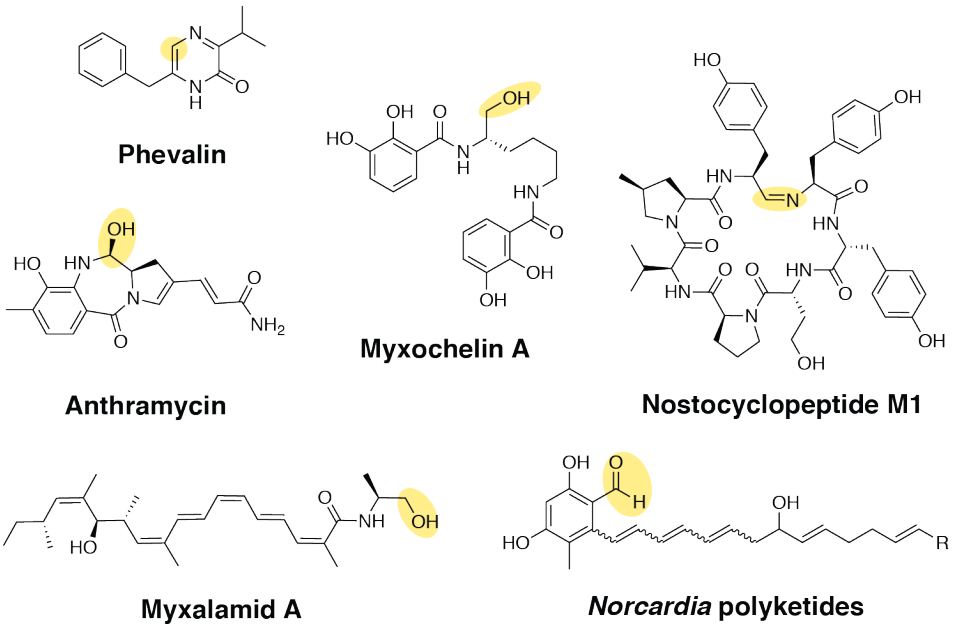


Figure 1.7: Structural diversity from R domains in NRPS- and PKS-derived natural products.

A diverse group of compounds is produced by the action of these domains and subsequent chemistry in which these reactive groups can engage. Motifs arising from action of an R domain are highlighted in yellow.

Installation of a C-terminal peptide aldehyde on an NRPS or PKS scaffold is one of Nature's simplest ways for introducing a reactive electrophile that may that covalently inhibit enzyme targets.[135] Natural products from this class have therefore attracted much interest and been widely evaluated for bioactivity. In particular, NRPS products containing terminal aldehydes

produced by R domains (peptide aldehydes) are a well-studied class of protease inhibitors.[152,153] We were initially drawn to investigate the molecules described in this work due to the ability to predict this potentially relevant biological function directly from primary genomic sequence information. In this next section, we discuss what is currently known about peptide aldehyde protease inhibitors.

1.4. Peptide aldehydes, an important class of natural product protease inhibitors

Serine and cysteine proteases are evolutionarily and structurally diverse, but they all catalyze hydrolysis of a peptide substrate by nucleophilic attack of a serine or cysteine residue that is activated as part of a catalytic triad or dyad.[154,155] For serine proteases, the catalytic triad of His, Asp, and Ser is well conserved, while for cysteine proteases, a greater diversity of catalytic mechanisms has been observed.[154,155] The catalytic mechanism of a typical serine protease is shown in Figure 1.8. In this mechanism, Asp102 activates His57 to deprotonate Ser197 (i), which can then perform a nucleophilic attack on the peptide substrate (ii). The resulting hemiacetal intermediate (iii) collapses to cleave the peptide bond, releasing the downstream portion of the peptide (iv). His57 activates a water molecule for nucleophilic attack on the enzyme-substrate complex (iv), generating a tetrahedral intermediate (v) which collapses to liberate Ser195 and the upstream portion of the peptide (vi).[154]

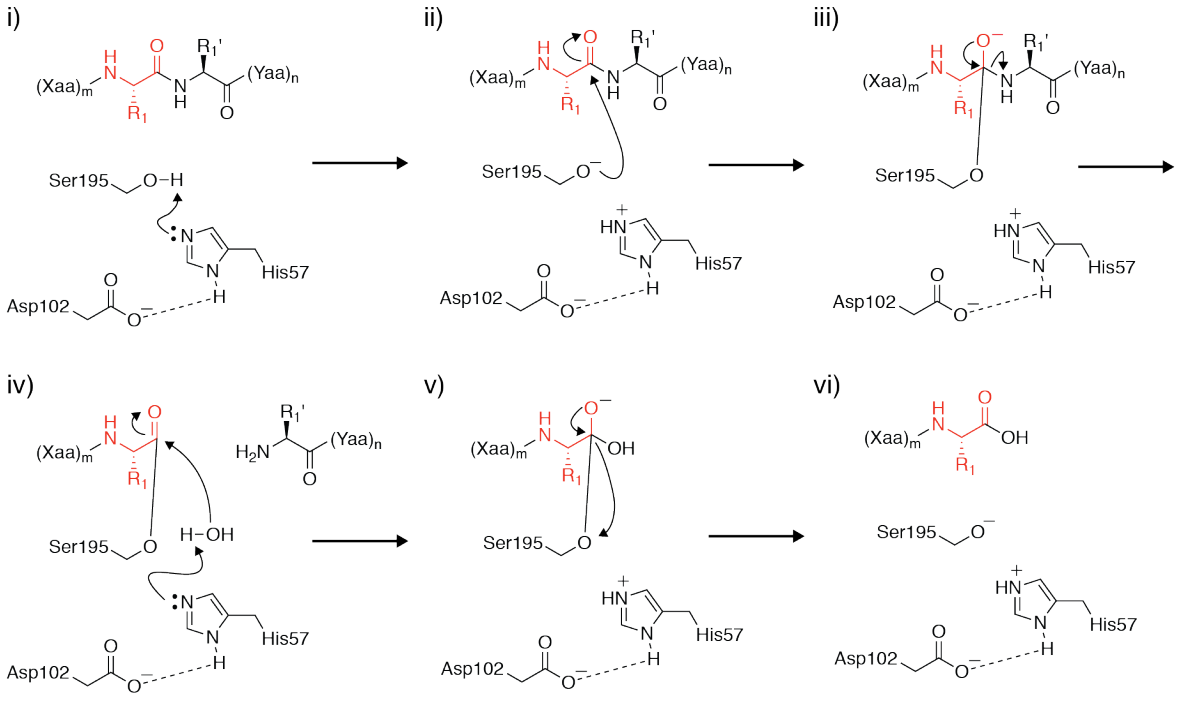


Figure 1.8: Mechanism of chymotrypsin-like serine proteases.

Basic mechanism of a chymotrypsin-like serine protease cleaving a peptide, showing the roles of catalytic triad residues His57, Asp102, and Ser195.[154]

Serine and cysteine proteases usually possess particular cleavage specificities, which are conferred in part by the residues lining the substrate-binding channels in their active sites.[154,155] A model serine protease active site is shown in Figure 1.9. The subsites where individual residues from the peptide substrate bind from N-terminus to C-terminus are labeled upstream (S1, S2, …) and downstream (S1', S2', …) of the scissile bond. The residues of the peptide that bind in this channel are correspondingly labeled upstream (P1, P2, …) and downstream (P1', P2', …) (Figure 1.9A).[154] Some common cleavage specificities of serine and cysteine proteases include trypsin-like (basic residues in the P1 position), caspase-like (acidic residues in the P1 position), and chymotrypsin-like (large hydrophobic residues in the P1 position) (Figure 1.9B). The ability of peptide aldehydes to inhibit these proteases is apparent by observing how these

compounds bind in the same channel. When the residues of the peptide aldehyde align with the protease's preferred cleavage specificity, the reactive aldehyde electrophile is oriented to react with the catalytic nucleophilic residue, forming a reversible hemiacetal (or thiohemiacetal) linkage (Figure 1.9A).[156,157]

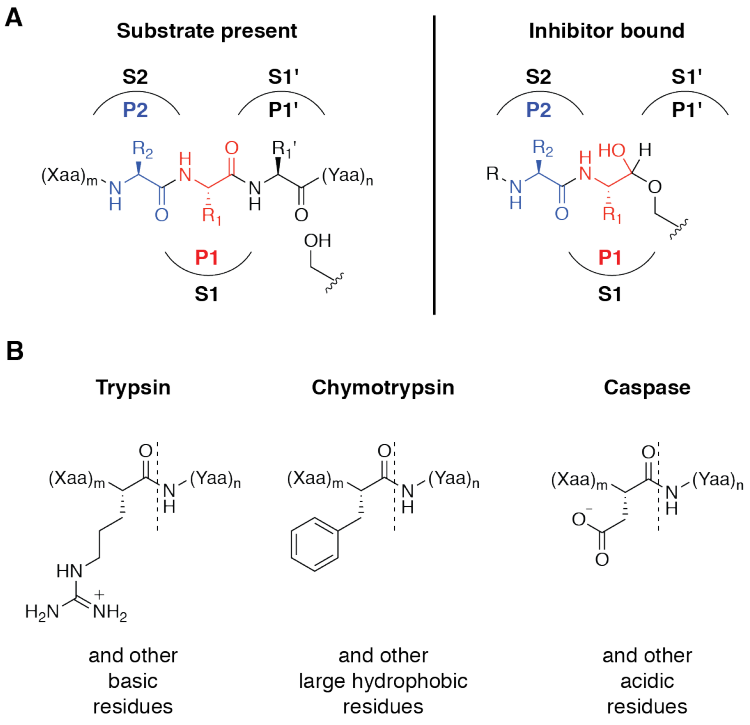


Figure 1.9: Serine and cysteine protease cleavage specificities allow for prediction of protease inhibitor specificity.

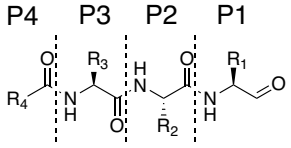
(A) The protease substrate binding pocket is responsible for generating cleavage specificity in proteases by orienting a particular peptide bond for attack by the serine or cysteine nucleophile. Peptide aldehyde inhibitors can bind in these pockets in an analagous fashion, leading to formation of the reversibly bound hemiacetal enzyme-inhibitor complex. (B) Some common specificities of human serine proteases are cleavage after basic residues (trypsin), large hydrophobic residues (chymotrypsin), and acidic residues (caspase).

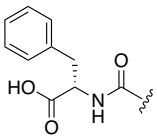
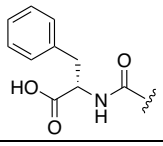
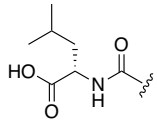
Since the 1970's, natural product peptide aldehydes have been isolated from many microbial species, particularly soil Actinomycetes (Table 1, Figure 1.10). In their original isolations or subsequent follow up studies, these compounds have often been evaluated as protease

inhibitors.[158–166] Some notable examples of peptide aldehyde natural product protease inhibitors that are commonly used in biochemical assays include leupeptin, chymostatin, and elastatinal. Leupeptin is commonly used as a trypsin inhibitor, while chymostatin inhibits chymotrypsin-like proteases and elastatinal is active against the immune system protease neutrophil elastase (all at effective concentrations of 10-100 μM). More recently, the flavopeptins were discovered using a proteomics-based workflow,[167] and the unique biosynthesis of livipeptin (with loading of L-arginine onto its assembly line by a tRNA synthetase rather than a standard A domain) was elucidated.[168]

Table 1: Some examples of natural product peptide aldehydes.

A wide variety of natural product peptide aldehydes have been isolated from microbial species. (Cap = capreomycidine.)



	Source	P4	P3	P2	P1
Leupeptin[158]	*Streptomyces roseus*, *Streptomyces exfoliatus*	Ac	L-Leu	L-Leu	L-Arg-CHO
Chymostatin A[159]	*Streptomyces hygroscopicus*		L-Cap	L-Leu	L-Phe-CHO
Antipain[160]	*Streptomyces* sp. MB 561-C2		L-Arg	L-Val	L-Arg-CHO
Elastatinal[161]	*Actinomycetes* MD469-CG8		L-Cap	L-Asn	L-Ala-CHO
Fellutamide B[162]	*Penicillium fellutanum*	β-hydroxydodecanoyl	L-Asn	L-Gln	L-Leu-CHO
Nerfilin I[163]	*S.halstedii* 2723-SV2	3-methylbutanoyl	L-Tyr	L-Val	L-Phe-CHO
Tyrostatin (Tyropeptin A)[164]	*Kitasatospora* sp. MK993-dF2	3-methylbutanoyl	L-Tyr	L-Val	L-Tyr-CHO
Tyropeptin B[165]	*Kitasatospora* sp. MK993-dF2	butanoyl	L-Tyr	L-Val	L-Tyr-CHO
Livipeptin[166]	*S. lividans* 66	–	Ac	L-Leu	L-Arg-CHO

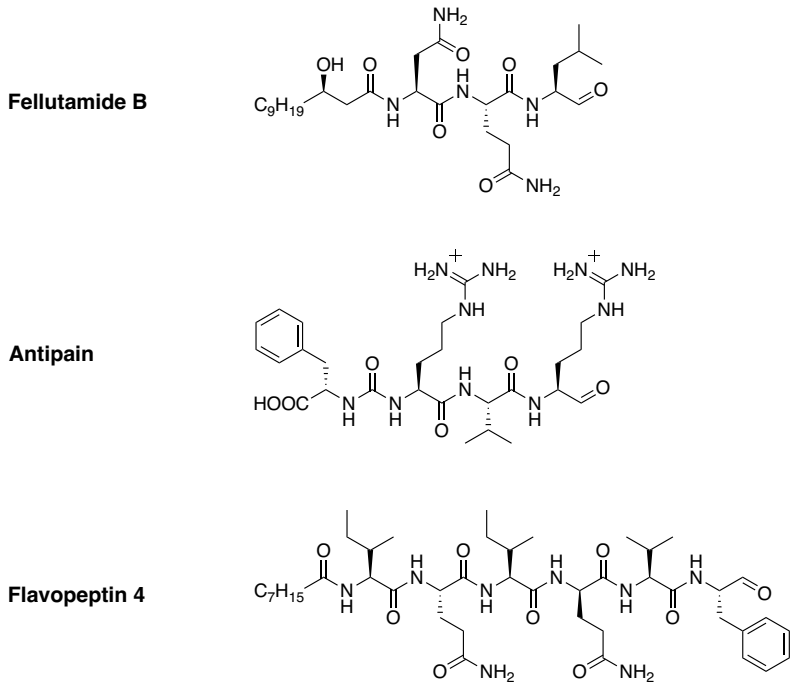


Figure 1.10: Structural diversity of natural product peptide aldehyde protease inhibitors.

Natural product peptide aldehydes investigated as protease inhibitors contain a variety of structural motifs, including β-hydroxy acids (fellutamide B) and ureido-linked amino acid residues (antipain) on their N-terminus.

Aside from its utility as a protease inhibitor in the laboratory, the natural physiological role of leupeptin has also been investigated in one of its producing organisms, *Streptomyces exfoliatus*, by Lee and coworkers (Figure 1.11).[169,170] Leupeptin is constitutively produced at very high concentrations in batch cultures of this organism (~2 mM), and it inhibits a trypsin-like protease involved in mycelial degradation, which is also constitutively produced. In nutrient-limited conditions, where mycelial growth is unfavorable and aerial hyphae are preferred, the bacteria produce and secrete a leupeptin-inactivating enzyme, which is a metalloprotease. This protease chemically degrades leupeptin, allowing the trypsin-like protease to degrade the mycelium and promote growth of aerial hyphae.[169,170] Interestingly, a nearly identical interaction network is observed in *Streptomyces coelicolor*, except that the peptide aldehyde protease

inhibitor leupeptin is replaced with a proteinaceous protease inhibitor.[171] This interaction network provides a fascinating natural example of a peptide aldehyde compound exerting a physiological role as a protease inhibitor.

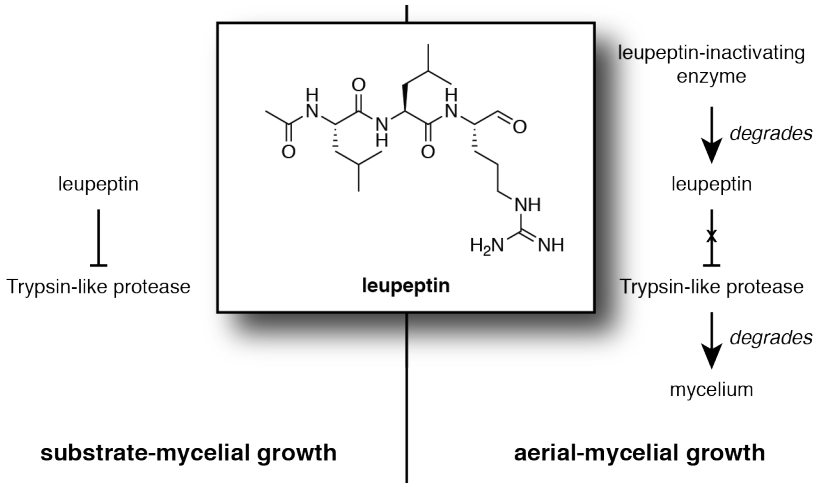


Figure 1.11: Physiological role of leupeptin in the life cycle of *S. exfoliatus*.

Leupeptin plays a role in life cycle regulation in one of its producing species. During substrate-mycelial growth, leupeptin inhibits the trypsin-like protease of this organism in batch cultures, allowing for substrate-mycelial growth. In aerial-mycelial growth, the organism produces leupeptin-inactiving protein, a metalloenzyme which degrades leupeptin and liberates the trypsin-like protease, which can then degrade the substrate-mycelium and promote formation of aerial hyphae.[170]

Along with these natural products, synthetic peptide aldehydes have also been exploited as protease inhibitors. Such molecules have been generally used as tool compounds along the path to developing drugs that incorporate more potent electrophiles, such as boronic acids (Figure 1.12). For example, MG-132 is a peptide aldehyde that was originally discovered in the quest to develop proteasome inhibitors by Myogenics, a biotechnology company. Research on inhibiting the proteasome, an N-terminal threonine protease, with this type of reactive electrophile eventually led to the development of bortezomib, a boronic acid-based leukemia therapeutic.[172] Peptide aldehydes have also attracted interest as inhibitors of caspases, which are potential

targets for modulating inflammation,[173] and the compound Ac-YVAD-H is still used as a tool compound in biochemical assays to inhibit caspase activity.[174]

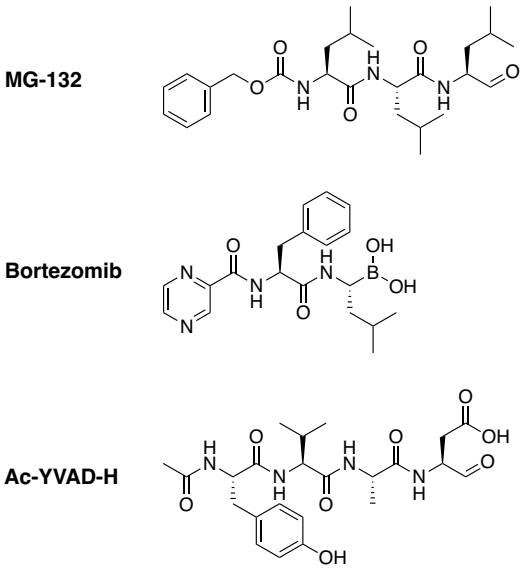


Figure 1.12: Synthetic peptide aldehydes used as inspiration for drug development and tool compounds.

Though peptide aldehydes are not typically considered potential drug candidates themselves, many of the properties that make them unsuitable for this role (relatively high IC50's, low selectivity between different types of proteases)[175] are likely not prohibitive for interesting activities that they might exert if they are produced in situ in the gut environment. In our work, we were initially drawn to investigate predicted peptide aldehyde production by gut microbes because we suspected that they could inhibit proteases in the intestinal environment. There are a vast number of proteases present in this body site, and though many have been studied for their roles in health and disease, there is much to learn about the precise mechanisms through which they exert their effects. In the next section, we discuss what is currently known about the roles of proteases in the human gut environment.

1.5. Targeting proteases in the gut microbiota to impact health and disease

Proteases from both host and microbial sources are present in the human colon and serve myriad roles there (for recent reviews about host proteases in this environment, see Vergnolle[176] and Edgington-Mitchell[177], and for microbial proteases, Haller and coworkers[178] and Maharshak and coworkers[179]). Host proteases in the gut are involved in gut barrier maintenance, immune system regulation, and pathogen defense.[176,177] Considering roles for microbial proteases, up to 25 g of undigested protein fragments enter the colon each day (both from food and shed from the host), and microbial proteases are likely involved in degrading this protein in order to access amino acids.[180,181] Potential human targets of gut microbial proteases include host receptors, the extracellular matrix, the epithelial barrier, and mucin.[178,179] Dysregulation of proteolytic activity in the intestine has been suggested to contribute to disease, but there remains much to be determined about the most significant proteases in this environment and their targets.[176–179]

1.5.1. Host proteases in the gut environment

Host proteases in the gut are produced by several different cell types. Mucosal mast cells, which are involved with innate immunity, make up 2–5% of the cells in the lamina propria, and they are known to release pro-inflammatory effectors upon stimulation with an antigen.[21] These effectors include proteases tryptase, chymase, and granzymes.[182] Upon release, these proteases can activate other inflammatory mediators or modify extracellular matrix components.[182] Intestinal macrophages, immune system cells resident to lamina propria, are another major cell type responsible for protease production in this environment. These cells are responsible for phagocytosis of bacteria and helping to maintain mucosal homeostasis in spite of the large bacterial load in this environment.[183] They produce several different types of proteases, including

matrix metalloproteases (MMPs), which are responsible for remodeling the ECM and may have roles in regulating inflammation and gut injury repair;[184] caspases, which can activate immune system components;[173] and cathepsins, which are important for antigen processing and presentation.[185] In the intestinal epithelial cells themselves, matriptase, a transmembrane serine protease, has been shown to be important for intestinal barrier function.[186] The immunoproteasome, which is derived from many of the same components as the constitutive proteasome but is expressed mainly in immune cells, may also play an important role in intestinal health.[187,188] This protease is involved in antigen presentation by major histocompatibility complex (MHC) I, and it has also been implicated in the maintenance of cellular homeostasis and stress response.[188]

Along with dysregulation of the proteases that are normally present in the environment, additional proteases that have been implicated in harmful inflammation include those that are produced by neutrophils, which are immune cells that are recruited to sites of microbial infection.[189] Though neutrophils are thought to kill bacteria mainly by phagocytosis, they also secrete three specific proteases (elastase, proteinase-3 and cathepsin G) that may be involved in degrading bacterial components and activating inflammation.[189]

Even where firm mechanistic links have not been drawn, other associations between proteases and gut disease states have been revealed by studies of protease inhibitors in model organisms. This technique has proven particularly fruitful in investigations of detergent-induced colitis. Inhibitors of cysteine cathepsins B, L and S (including some synthetic peptide aldehydes) have been used to ameliorate the effects of detergent-induced colitis in mouse and rat models.[190,191] The caspase-1 inhibitor pralnacasan has also been used to this effect in a mouse model.[192] Finally, strengthening the case for an important role of the immunoproteasome in the

gut, detergent-induced colitis is attenuated both by a selective inhibitor of the immunoproteasome and in LMP7 knockout mice (which lack the gene producing the β_{5i} subunit of the immunoproteasome).[193,194] However, despite extensive investigations of gut-localized human proteases and explorations of their inhibition as a therapeutic strategy for IBD, protease inhibitors are still not in use as therapeutics for this condition.[176]

Interestingly, while small molecule drugs are not currently used to modulate the activity of human gut proteases, several commensal bacterial strains produce proteinaceous protease inhibitors that likely participate in the protease interaction network in this environment. *B. longum* produces a serpin in the gut that inhibits eukaryotic elastin-like proteases, and a similar protein is produced by *Bifidobacterium breve* and *Bifidobacterium dentium*.[195] Two serpins produced by *Eubacterium sireaum* inhibit neutrophil elastase and proteinase 3, which are both associated with IBD.[196] Though the treatment of intestinal diseases with exogenous inhibitors of human proteases remains an unattained goal, the fact that this appears to happen in Nature suggests that this mode of therapeutic action is worthy of further investigation.

1.5.2. Microbial proteases in the gut environment

Aside from human proteases, microbial proteases contribute to the proteolytic potential of the intestinal environment.[179] Gut commensal proteolytic activity was first seriously investigated in the 1980's by Macfarlane and coworkers.[197–199] From their studies, it became clear that significant microbially-derived protease activity is found in the colon, with both trypsin- and chymotrypsin-like activity observed.[198] Though the major proteases contributing to this activity have not been identified, several specific gut species were identified as significant extracellular

protease producers, including *B. fragilis*, *Clostridium perfringens*, and *Clostridium sporogenes*.[197–199]

More recently, some specific proteases from commensal gut microbes have been characterized and linked with physiological roles. Mottram and coworkers[200] and Wolan and coworkers[201] have recently determined crystal structures and investigated substrate specificity for the C11 protease from *Parabacteroides merdae*. The C11 clostripain-like proteases are cysteine proteases that are found in many gut bacterial strains, and they exhibit a variety of potential biological roles in this environment.[201] Another significant gut microbial protease is lactocepin from *Lactobacillus casei*, a popular probiotic, which has been shown to degrade pro-inflammatory cytokines in vitro.[202] Microbial proteases have also attracted interest for their potential interaction with the gut-brain axis. A recent study by Lomax and coworkers showed that *F. prausnitzii* secreted serine protease(s) interact with neurons that are found in the mouse gut, but did not identify which protease(s) were responsible for this effect.[203]

Microbial proteases from gut pathogens have also been studied for their contributions to disease.[178,179] *Vibrio cholerae*, the causative agent of cholera, secretes three proteases with potential roles in virulence: hemagluttinin protease,[204] metalloprotease PrtV,[205] and a serine protease.[206] Serine protease autotransporters (SPATEs) are a family of proteases secreted by *E. coli* and *Shigella* which may have roles in infection and colonization of the gut.[207–209] It is hypothesized that these proteases may degrade mucin or other host components, and also that they may interact with the host immune system.[207,208] The opportunistic pathogen *E. faecalis* produces a secreted gelatinase, which can degrade host structural components and interact with protease activated receptors.[210,211]

In addition to studies on individual species and enzymes, meta'omics techniques have been used to investigate protease activity in the commensal gut microbiota.[95,212] Wolan and coworkers recently used an activity-based chloromethyl ketone probe to identify reactive cysteines in mouse fecal samples.[213] In this study, cysteine protease and hydrolase gene orthologies were enriched in *Rag1*$^{-/-}$ mice subjected to the T cell transfer model of colitis ("IBD mice"). However, these specific proteases were not identified. Another study attempted functional metagenomics to discover proteases in this environment but failed to identify any novel proteases.[214] Several significant microbial proteases were identified in a recent metaproteomics analysis of the gut microbiota in a healthy cohort (15 subjects), including serine protease HtrA and the ATP-dependent lon and Clp proteases.[215]

Overall, though much of this work suggests that inhibition of proteases may be a promising therapeutic direction for gastrointestinal disease, there are no drugs currently in use with this mechanism of action.[176] The incredible diversity of protease production in the gut and the relative lack of knowledge about how these proteins are contributing to ecological interactions was also an inspiration for the work described in this dissertation, as discovering native protease-inhibitor interactions in the commensal gut microbiota may highlight significant proteases in this environment that could be targeted for therapeutic effect.

1.6. Chapter preview

This dissertation describes my work in the Balskus laboratory to characterize the structures and activities of a family of putative peptide aldehydes produced by prominent members of the human gut microbiota. In Chapter 2, I describe our efforts to use our isolation-independent approach to access the putative products of an NRPS gene cluster from *Ruminococcus bromii*,

one of the most abundant commensal microbes in the human gut. We first employed bioinformatic analyses to predict the product of this conserved and widely distributed gene cluster (the *rup* gene cluster) as a reactive, *N*-acylated dipeptide aldehyde (ruminopeptin). We then used in vitro biochemical characterization of the NRPS assembly line enzymes to identify the building blocks of ruminopeptin and predict its most likely structure.

In Chapter 3, I describe how we used synthesis and bioinformatic analysis as tools to predict a physiologically relevant target for ruminopeptin(s) in the gut microbiota. Using a concise solution phase synthesis, I accessed a library of ruminopeptin analogues and evaluated their bioactivities as protease inhibitors. We found these molecules inhibit *S. aureus* endoproteinase GluC (also known as SspA/V8 protease), which has been implicated in virulence in a mouse abscess model.[216] The human gut microbe and opportunistic pathogen *Enterococcus faecalis* also produces a virulence-related glutamyl endopeptidase,[217] and further bioinformatics analyses revealed additional homologs of this enzyme in gut microbial genomes and metagenomes. We also attempted to observe phenotypes arising from mutations in glutamyl endopeptidase genes in two different species and several assays. Though we could not identify observable differences between these strains and their corresponding wild type strains, we hypothesize that protease inhibition of this protease family by ruminopeptin(s) may be important for mediating microbe-microbe interactions in the human gut.

We were inspired by the success of this strategy in discovering interesting small molecules with a potentially physiologically relevant activity, and in Chapter 4, I describe how we expanded the scope of our study to investigate additional predicted peptide aldehydes from gut microbial genomic data. We selected 4 additional gene clusters of interest and synthesized a comprehensive set of their predicted biosynthetic products to reach a total library size of 48

peptide aldehydes. We then pursued several strategies to identify putative targets of these compounds by evaluating them as inhibitors of human proteases, antibiotics against a set of prominent pathogens and commensals, and inhibitors of gut microbial secreted protease activity. Finally, we designed and synthesized an activity probe based on the structure of one of the compounds and used it to investigate potential protein targets in *C. difficile* 630Δerm using an untargeted chemoproteomics workflow.

1.7. References

1. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–336 (2011).

2. Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* **9,** 577–589 (2012).

3. Kamada, N., Chen, G. Y., Inohara, N. & Núñez, G. Control of pathogens and pathobionts by the gut microbiota. *Nat. Immunol.* **14,** 685–690 (2013).

4. Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol. Int.* **66,** 515–522 (2017).

5. van den Elsen, L. W., Poyntz, H. C., Weyrich, L. S., Young, W. & Forbes-Blom, E. E. Embracing the gut microbiota: the new frontier for inflammatory and infectious diseases. *Clin. Transl. Immunol.* **6,** e125 (2017).

6. Ribet, D. & Cossart, P. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect.* **17,** 173–183 (2015).

7. Geva-Zatorsky, N. *et al.* Mining the human gut microbiota for immunomodulatory organisms. *Cell* **168,** 928–943 (2017).

8. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell Host Microbe* **17,** 260–273 (2015).

9. Carabotti, M., Scirocco, A., Maselli, M. A. & Severi, C. The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems. *Ann. Gastroenterol.* **28,**

203–209 (2015).

10. Berni Canani, R., Gilbert, J. a. & Nagler, C. R. The role of the commensal microbiota in the regulation of tolerance to dietary allergens. *Curr. Opin. Allergy Clin. Immunol.* **15,** 243–249 (2015).

11. Ma, J., Zhou, Q. & Li, H. Gut microbiota and nonalcoholic fatty liver disease: Insights on mechanisms and therapy. *Nutrients* **9,** (2017).

12. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

13. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).

14. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).

15. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat. Microbiol.* **3,** 337–346 (2018).

16. Young, V. B. The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ* **356,** j831 (2017).

17. Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8,** 51 (2016).

18. Walter, J. & Ley, R. The human gut microbiome: Ecology and recent evolutionary changes. *Annu. Rev. Microbiol.* **65,** 411–429 (2011).

19. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14,** 20–32 (2015).

20. Yoo, B. B. & Mazmanian, S. K. The enteric network: Interactions between the immune and nervous systems of the gut. *Immunity* **46,** 910–926 (2017).

21. Ramsay, D. B., Stephen, S., Borum, M., Voltaggio, L. & Doman, D. B. Mast cells in gastrointestinal disease. *Gastroenterol. Hepatol. (N. Y).* **6,** 772–777 (2010).

22. Tropini, C., Earle, K. A., Huang, K. C. & Sonnenburg, J. L. The gut microbiome: Connecting spatial organization to function. *Cell Host Microbe* **21,** 433–442 (2017).

23. Nava, G. M., Friedrichsen, H. J. & Stappenbeck, T. S. Spatial organization of intestinal microbiota in the mouse ascending colon. *ISME J.* **5,** 627–638 (2011).

24. Morgan, X. C. *et al.* A Crypt-Specific Core Microbiota Resides in the Mouse Colon. *MBio* **3,** doi:10.1128/mBio.00116-12. (2012).

25. Walker, A. W. *et al.* The species composition of the human intestinal microbiota differs between particle-associated and liquid phase communities. *Environ. Microbiol.* **10,** 3275–3283 (2008).

26. Donohoe, D. R. *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.* **13,** 517–526 (2011).

27. Kelly, C. J. *et al.* Crosstalk between microbiota-derived short-chain fatty acids and intestinal epithelial HIF augments tissue barrier function. *Cell Host Microbe* **17,** 662–671 (2015).

28. Verma, M. S. *et al.* A common mechanism links activities of butyrate in the colon. *ACS Chem. Biol.* **13,** 1291–1298 (2018).

29. Chang, P. V., Hao, L., Offermanns, S. & Medzhitov, R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc. Natl. Acad. Sci.* **111,** 2247–2252 (2014).

30. Mullineaux-Sanders, C., Suez, J., Elinav, E. & Frankel, G. Sieving through gut models of colonization resistance. *Nat. Microbiol.* **3,** 132–140 (2018).

31. Pamer, E. G. *et al.* Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J. Clin. Invest.* **120,** 4332–4341 (2010).

32. Ayres, J. S., Trinidad, N. J. & Vance, R. E. Lethal inflammasome activation by a multidrug-resistant pathobiont upon antibiotic disruption of the microbiota. *Nat. Med.* **18,** 799–806 (2012).

33. Pavia, A. T. *et al.* Epidemiologic evidence that prior antimicrobial exposure decreases resistance to infection by antimicrobial-sensitive *Salmonella*. *J. Infect. Dis.* **161,** 255–260 (1990).

34. Modi, S. R. *et al.* Antibiotics and the gut microbiota. *J. Clin. Invest.* **124,** 4212–4218 (2014).

35. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome

throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8,** (2016).

36. Drissi, F., Buffet, S., Raoult, D. & Merhej, V. Common occurrence of antibacterial agents in human intestinal microbiota. *Front. Microbiol.* **6,** 441 (2015).

37. Ayres, J. S. Cooperative microbial tolerance behaviors in host-microbiota mutualism. *Cell* **165,** 1323–1331 (2016).

38. Ismagilov, R. F. *et al.* Rapid fucosylation of intestinal epithelium sustains host–commensal symbiosis in sickness. *Nature* **514,** 638–641 (2014).

39. Chu, H. & Mazmanian, S. K. Innate immune recognition of the microbiota promotes host-microbial symbiosis. *Nat. Immunol.* **14,** 668–675 (2013).

40. Quévrain, E. *et al.* Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. *Gut* **65,** 415–425 (2016).

41. Mazmanian, S. K., Cui, H. L., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122,** 107–118 (2005).

42. Cullen, T. W. *et al.* Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* **347,** 170–175 (2015).

43. Von Schillde, M. A. *et al.* Lactocepin secreted by *Lactobacillus* exerts anti-inflammatory effects by selectively degrading proinflammatory chemokines. *Cell Host Microbe* **11,** 387–396 (2012).

44. Hooks, K. B. & O'Malley, M. A. Dysbiosis and its discontents. *MBio* **8,** e01492-17 (2017).

45. Guo, H., Callaway, J. B. & Ting, J. P.-Y. Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nat. Med.* **21,** 677–687 (2015).

46. Nishida, A. *et al.* Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clin. J. Gastroenterol.* **11,** 1–10 (2018).

47. Turpin, W., Goethel, A., Bedrani, L. & Croitoru, K. Determinants of IBD heritability: genes, bugs, and more. *Inflamm. Bowel Dis.* **24,** 1133–1148 (2018).

48. Tilg, H. & Moschen, A. R. Microbiota and diabetes: An evolving relationship. *Gut* **63,** 1513–1521 (2014).

49. Khan, M. J., Gerasimidis, K., Edwards, C. A. & Shaikh, M. G. Role of gut microbiota in the aetiology of obesity: Proposed mechanisms and review of the literature. *J. Obes.* **2016,** 7353642 (2016).

50. Lessa, F. C., Gould, C. V. & Clifford McDonald, L. Current status of *Clostridium difficile* infection epidemiology. *Clin. Infect. Dis.* **55,** S65–S70 (2012).

51. Centers for Disease Control and Prevention (CDC). Severe *Clostridium difficile*-associated disease in populations previously at low risk – four states, 2005. *MMWR. Morb. Mortal. Wkly. Rep.* **54,** 1201–1205 (2005).

52. Kuijper, E. J., Coignard, B. & Tüll, P. Emergence of *Clostridium difficile*-associated disease in North America and Europe. *Clin. Microbiol. Infect.* **12,** 2–18 (2006).

53. Di Bella, S., Ascenzi, P., Siarakas, S., Petrosillo, N. & di Masi, A. *Clostridium difficile* toxins A and B: Insights into pathogenic properties and extraintestinal effects. *Toxins (Basel).* **8,** 1–25 (2016).

54. Leffler, D. A. & Lamont, J. T. *Clostridium difficile* infection. *N. Engl. J. Med.* **372,** 1539–1548 (2015).

55. Zhu, D., Sorg, J. A. & Sun, X. Clostridioides difficile biology: Sporulation, germination, and corresponding therapies for C. difficile infection. *Front. Cell. Infect. Microbiol.* **8,** 1–10 (2018).

56. Mattila, E. *et al.* Fecal transplantation, through colonoscopy, is effective therapy for recurrent *Clostridium difficile* infection. *Gastroenterology* **142,** 490–496 (2012).

57. Gupta, S., Allen-Vercoe, E. & Petrof, E. O. Fecal microbiota transplantation: In perspective. *Therap. Adv. Gastroenterol.* **9,** 229–239 (2016).

58. McDonald, L. C. *et al.* Clinical practice guidelines for *Clostridium difficile* infection in adults and children: 2017 update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clin. Infect. Dis. an Off. Publ. Infect. Dis. Soc. Am.* **31,** 431–455 (2018).

59. Hoffmann, D. *et al.* Improving regulation of microbiota transplants. *Science* **358,** 1390–1391 (2017).

60. Garber, K. Drugging the gut microbiome. *Nat. Biotechnol.* **33,** 228–231 (2015).

61. Anderson, J. L., Edney, R. J. & Whelan, K. Systematic review: Faecal microbiota transplantation in the management of inflammatory bowel disease. *Aliment. Pharmacol.*

Ther. **36,** 503–516 (2012).

62. OpenBiome. Current Studies. Available at: https://www.openbiome.org/current-studies/. (Accessed: 22nd January 2019)

63. Melander, R. J., Zurawski, D. V. & Melander, C. Narrow-spectrum antibacterial agents. *Medchemcomm* **9,** 12–21 (2018).

64. Paule, A., Frezza, D. & Edeas, M. Microbiota and phage therapy: Future challenges in medicine. *Med. Sci.* **6,** 86 (2018).

65. Schneiderhan, J., Master-Hunter, T. & Locke, A. Targeting gut flora to treat and prevent disease. *J. Fam. Pract.* **65,** 34–38 (2016).

66. Fujiya, M., Ueno, N. & Kohgo, Y. Probiotic treatments for induction and maintenance of remission in inflammatory bowel diseases: A meta-analysis of randomized controlled trials. *Clin. J. Gastroenterol.* **7,** 1–13 (2014).

67. Sola-Oladokun, B., Culligan, E. P. & Sleator, R. D. Engineered probiotics: Applications and biological containment. *Annu. Rev. Food Sci. Technol.* **8,** 353–370 (2017).

68. Veiga, P. *et al.* Changes of the human gut microbiome induced by a fermented milk product. *Sci. Rep.* **4,** 6328 (2015).

69. Deehan, E. C. *et al.* Modulation of the gastrointestinal microbiome with nondigestible fermentable carbohydrates to improve human health. *Microbiol. Spectr.* **5,** BAD-0019-2017 (2017).

70. Petrof, E. O., Claud, E. C., Gloor, G. B. & Allen-Vercoe, E. Microbial ecosystems therapeutics: A new paradigm in medicine? *Benef. Microbes* **4,** 53–65 (2013).

71. Makki, K., Deehan, E. C., Walter, J. & Bäckhed, F. The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host Microbe* 705–715 (2018). doi:10.1016/j.chom.2018.05.012

72. Wallace, B. D. *et al.* Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330,** 831–835 (2010).

73. Pellock, S. J. *et al.* Gut microbial β-glucuronidase inhibition via catalytic cycle interception. *ACS Cent. Sci.* **4,** 868–879 (2018).

74. Wang, Z. *et al.* Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* **163,** 1585–1595 (2015).

75. Roberts, A. B. *et al.* Development of a gut microbe–targeted nonlethal therapeutic to inhibit thrombosis potential. *Nat. Med.* **24,** 1407–1417 (2018).

76. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 6252–6257 (2011).

77. Lagkouvardos, I., Overmann, J. & Clavel, T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut Microbes* **8,** 493–503 (2017).

78. Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533,** 543–546 (2016).

79. Lagier, J.-C. *et al.* Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* **1,** 16203 (2016).

80. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160,** 583–594 (2015).

81. Rice, B. L. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176,** 649–662 (2019).

82. Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C. & Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* **20,** 731–741 (2014).

83. Levin, B. J. *et al.* A prominent glycyl radical enzyme in human gut microbiomes metabolizes *trans*-4-hydroxy-L-proline. *Science* **355,** eaai8386 (2017).

84. Backman, L. R. F., Funk, M. A., Dawson, C. D. & Drennan, C. L. New tricks for the glycyl radical enzyme family. *Crit. Rev. Biochem. Mol. Biol.* **52,** 674–695 (2017).

85. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).

86. Moore, A. M., Munck, C., Sommer, M. O. A. & Dantas, G. Functional Metagenomic Investigations of the Human Intestinal Microbiota. *Front. Microbiol.* **2,** 1–8 (2011).

87. Wang, W. L. *et al.* Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* **21,** 803–814 (2015).

88. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the

antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).

89. Cohen, L. J. *et al.* Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. *Proc. Natl. Acad. Sci.* **112,** E4825–E4834 (2015).

90. Jones, B. V., Begley, M., Hill, C., Gahan, C. G. M. & Marchesi, J. R. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 13580–13585 (2008).

91. Jones, B. V., Sun, F. & Marchesi, J. R. Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. *Lett. Appl. Microbiol.* **45,** 418–420 (2007).

92. Sabree, Z. L., Rondon, M. R. & Handelsman, J. in *Encyclopedia of Microbiology* **72,** 622–632 (Elsevier, 2009).

93. Liebl, W. *et al.* Alternative hosts for functional (meta)genome analysis. *Appl. Microbiol. Biotechnol.* **98,** 8099–8109 (2014).

94. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16,** 276–289 (2014).

95. Lee, P. Y., Chin, S.-F., Neoh, H. & Jamal, R. Metaproteomic analysis of human gut microbiota: where are we heading? *J. Biomed. Sci.* **24,** 36 (2017).

96. Petriz, B. A. & Franco, O. L. Metaproteomics as a complementary approach to gut microbiota in health and disease. *Front. Chem.* **5,** (2017).

97. Franzosa, E. A. *et al.* Sequencing and beyond: Integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13,** 360–372 (2015).

98. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349,** 1254766 (2015).

99. Garg, N. *et al.* Natural products as mediators of disease. *Nat. Prod. Rep.* **34,** 194–219 (2017).

100. Mousa, W. K., Athar, B., Merwin, N. J. & Magarvey, N. A. Antibiotics and specialized metabolites from the human microbiota. *Nat. Prod. Rep.* **34,** 1302–1331 (2017).

101. Turroni, S., Brigidi, P., Cavalli, A. & Candela, M. Microbiota–host transgenomic

metabolism, bioactive molecules from the inside. *J. Med. Chem.* **61,** 47–61 (2018).

102. Crost, E. H. *et al.* Ruminococcin C, a new anti-*Clostridium perfringens* bacteriocin produced in the gut by the commensal bacterium *Ruminococcus gnavus* E1. *Biochimie* **93,** 1487–1494 (2011).

103. Wilson, M. R., Zha, L. & Balskus, E. P. Natural product discovery from the human microbiome. *J. Biol. Chem.* **292,** 8546–8552 (2017).

104. Dabard, J. *et al.* Ruminococcin A, a new lantibiotic produced by a *Ruminococcus gnavus* strain isolated from human feces. *Appl. Environ. Microbiol.* **67,** 4111–4118 (2001).

105. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158,** 1402–1414 (2014).

106. Cimermancic, P. *et al.* Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **158,** 412–421 (2014).

107. Medema, M. H. *et al.* AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39,** 339–346 (2011).

108. Cohen, L. J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549,** 48–53 (2017).

109. Guo, C. *et al.* Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell* **168,** 517–526 (2017).

110. Chu, J. *et al.* Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat. Chem. Biol.* **12,** 1004–1006 (2016).

111. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic machinery, and mechanisms. *Chem. Rev.* **106,** 3468–3496 (2006).

112. Quadri, L. E. N. *et al.* Characterization of Sfp, a *Bacillus subtilis* phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **37,** 1585–1595 (1998).

113. Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A. & Walsh, C. T. Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry* **40,** 7099–7108 (2001).

114. Robbel, L., Hoyer, K. M. & Marahiel, M. A. TioS T-TE – A prototypical thioesterase responsible for cyclodimerization of the quinoline- and quinoxaline-type class of chromodepsipeptides. *FEBS J.* **276,** 1641–1653 (2009).

115. Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A. & Walsh, C. T. Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature* **407,** 215–218 (2000).

116. Sieber, S. A. & Marahiel, M. A. Learning from Nature's drug factories: Nonribosomal synthesis of macrocyclic peptides. *J. Bacteriol.* **185,** 7036–7043 (2003).

117. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).

118. Röttig, M. *et al.* NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** 362–367 (2011).

119. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **368,** 1500–1517 (2007).

120. Blin, K. *et al.* antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41,** W204–W212 (2013).

121. Qiao, K. *et al.* A fungal nonribosomal peptide synthetase module that can synthesize thiopyrazines. *Org. Lett.* **13,** 1758–1761 (2011).

122. Crawford, J. M., Portmann, C., Kontnik, R., Walsh, C. T. & Clardy, J. NRPS substrate promiscuity diversifies the xenematides. *Org. Lett.* **13,** 5144–5147 (2011).

123. Van Lanen, S. G., Lin, S., Dorrestein, P. C., Kelleher, N. L. & Shen, B. Substrate specificity of the adenylation enzyme SgcC1 involved in the biosynthesis of the enediyne antitumor antibiotic C-1027. *J. Biol. Chem.* **281,** 29633–29640 (2006).

124. Heemstra, J. R. & Walsh, C. T. Tandem action of the O_2- and $FADH_2$-dependent halogenases KtzQ and KtzR produce 6,7-dichlorotryptophan for kutzneride assembly. *J. Am. Chem. Soc.* **130,** 14024–14025 (2008).

125. Pfeifer, V. *et al.* A polyketide synthase in glycopeptide biosynthesis. The biosynthesis of the non-proteinogenic amino acid (S)-3,5-dihydroxyphenylglycine. *J. Biol. Chem.* **276,** 38370–38377 (2001).

126. May, J. J., Kessler, N., Marahiel, M. A. & Stubbs, M. T. Crystal structure of DhbE, an

archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc. Natl. Acad. Sci.* **99,** 12120–12125 (2002).

127. Zha, L. *et al.* Colibactin assembly line enzymes use *S*-adenosylmethionine to build a cyclopropane ring. *Nat. Chem. Biol.* **13,** 1063–1065 (2017).

128. Wittmann, M., Linne, U., Pohlmann, V. & Marahiel, M. A. Role of DptE and DptF in the lipidation reaction of daptomycin. *FEBS J.* **275,** 5343–5354 (2008).

129. Ramel, C. *et al.* Biosynthesis of the proteasome inhibitor syringolin A: the ureido group joining two amino acids originates from bicarbonate. *BMC Biochem.* **10,** 26 (2009).

130. Imker, H. J., Walsh, C. T. & Wuest, W. M. SylC catalyzes ureido-bond formation during biosynthesis of the proteasome inhibitor syringolin A. *J. Am. Chem. Soc.* **131,** 18263–18265 (2009).

131. Brotherton, C. A. & Balskus, E. P. A prodrug resistance mechanism is involved in colibactin biosynthesis and cytotoxicity. *J. Am. Chem. Soc.* **135,** 3359–3362 (2013).

132. Dose, B. *et al.* Unexpected bacterial origin of the antibiotic icosalide: Two-tailed depsipeptide assembly in multifarious *Burkholderia* symbionts. *ACS Chem. Biol.* **13,** 2414–2420 (2018).

133. Imker, H. J., Krahn, D., Clerc, J., Kaiser, M. & Walsh, C. T. *N*-Acylation during glidobactin biosynthesis by the tridomain nonribosomal peptide synthetase module GlbF. *Chem. Biol.* **17,** 1077–1083 (2010).

134. Du, L. & Lou, L. PKS and NRPS release mechanisms. *Nat. Prod. Rep.* **27,** 255–78 (2010).

135. Mullowney, M., McClure, R. A., Robey, M. T., Kelleher, N. L. & Thomson, R. J. Natural products from thioester reductase containing biosynthetic pathways. *Nat. Prod. Rep.* **35,** 847–878 (2018).

136. Kavanagh, K. L., Jörnvall, H., Persson, B. & Oppermann, U. Medium- and short-chain dehydrogenase/reductase gene and protein families: The SDR superfamily: Functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell. Mol. Life Sci.* **65,** 3895–3906 (2008).

137. Barajas, J. F. *et al.* Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem. Biol.* **22,** 1018–1029 (2015).

138. Chhabra, A. *et al.* Nonprocessive [2 + 2]e^- off-loading reductase domains from

mycobacterial nonribosomal peptide synthetases. *Proc. Natl. Acad. Sci.* **109,** 5681–5686 (2012).

139. Wilson, D. J., Shi, C., Teitelbaum, A. M., Gulick, A. M. & Aldrich, C. C. Characterization of AusA: A dimodular nonribosomal peptide synthetase responsible for the production of aureusimine pyrazinones. *Biochemistry* **52,** 926–937 (2013).

140. Zimmermann, M. & Fischbach, M. A. A family of pyrazinone natural products from a conserved nonribosomal peptide synthetase in *Staphylococcus aureus*. *Chem. Biol.* **17,** 925–930 (2010).

141. Wyatt, M. A. *et al. Staphylococcus aureus* nonribosomal peptide secondary metabolites regulate virulence. *Science* **329,** 294–296 (2010).

142. Blättner, S. *et al. Staphylococcus aureus* exploits a non-ribosomal cyclic dipeptide to modulate survival within epithelial cells and phagocytes. *PLOS Pathog.* **12,** e1005857 (2016).

143. Golakoti, T., Yoshida, W. Y., Chaganty, S. & Moore, R. E. Isolation and structure determination of nostocyclopeptides A1 and A2 from the terrestrial cyanobacterium *Nostoc* sp. ATCC53789. *J. Nat. Prod.* **64,** 54–59 (2001).

144. Kopp, F., Mahlert, C., Grünewald, J. & Marahiel, M. A. Peptide macrocyclization: the reductase of the nostocyclopeptide synthetase triggers the self-assembly of a macrocyclic imine. *J. Am. Chem. Soc.* **128,** 16478–16479 (2006).

145. Leimgruber, W., Batcho, A. D. & Schenker, F. The structure of anthramycin. *J. Am. Chem. Soc.* **87,** 5793–5795 (1965).

146. Hu, Y. *et al.* Benzodiazepine biosynthesis in *Streptomyces refuineus*. *Chem. Biol.* **14,** 691–701 (2007).

147. Kuo, J., Lynch, S. R., Liu, C. W., Xiao, X. & Khosla, C. Partial in vitro reconstitution of an orphan polyketide synthase associated with clinical cases of Nocardiosis. *ACS Chem. Biol.* **11,** 2636–2641 (2016).

148. Kunze, B., Bedorf, N., Kohl, W., Höfle, G. & Reichenbach, H. Myxochelin A, a new iron-chelating compound from *Angiococcus disciformis* (myxobacterales). Production, isolation, physico-chemical and biological properties. *J. Antibiot. (Tokyo).* **42,** 14–17 (1989).

149. Gaitatzis, N., Kunze, B. & Müller, R. In vitro reconstitution of the myxochelin biosynthetic machinery of *Stigmatella aurantiaca* Sg a15: Biochemical characterization of

a reductive release mechanism from nonribosomal peptide synthetases. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 11136–11141 (2001).

150. Li, Y., Weissman, K. J. & Müller, R. Myxochelin biosynthesis: direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. *J. Am. Chem. Soc.* **130,** 7554–7555 (2008).

151. Gerth, K. *et al.* The myxalamids, new antibiotics from *Myxococcus xanthus* (Myxobacterales) I. Production, physico-chemical and biological properties, and mechanism of action. *J. Antibiot. (Tokyo).* **36,** 1150–1156 (1983).

152. Thompson, R. C. [19] Peptide aldehydes: Potent inhibitors of serine and cysteine proteases. *Methods Enzymol.* **46,** 220–225 (1977).

153. Siklos, M., BenAissa, M. & Thatcher, G. R. J. Cysteine proteases as therapeutic targets: Does selectivity matter? A systematic review of calpain and cathepsin inhibitors. *Acta Pharm. Sin. B* **5,** 506–519 (2015).

154. Hedstrom, L. Serine protease mechanism and specificity. *Chem. Rev.* **102,** 4501–4523 (2002).

155. Otto, H.-H. & Schirmeister, T. Cysteine proteases and their inhibitors. *Chem. Rev.* **97,** 133–172 (1997).

156. Kurinov, I. V. & Harrison, R. W. Two crystal structures of the leupeptin-trypsin complex. *Protein Sci.* **5,** 752–758 (1996).

157. Margolin, N. *et al.* Substrate and inhibitor specificity of interleukin-1β-converting enzyme and related caspases. *J. Biol. Chem.* **272,** 7223–7228 (1997).

158. Kondo, S. *et al.* Isolation and characterization of leupeptins produced by Actinomycetes. *Chem. Pharmacetical Bull.* **17,** 1896–1901 (1969).

159. Umezawa, H. *et al.* Chymostatin, a new chymotrypsin inhibitor produced by Actinomycetes. *J. Antibiot. (Tokyo).* **23,** 425–427 (1970).

160. Suda, H., Aoyagi, T., Hamada, M., Takeuchi, T. & Umezawa, H. Antipain, a new protease inhibitor isolated from Actinomycetes. *J. Antibiot. (Tokyo).* **XXV,** 263–266 (1972).

161. Umezawa, H. *et al.* Elastatinal, a new elastase inhibitor produced by Actinomycetes. *J. Antibiot. (Tokyo).* **26,** 787–789 (1973).

162. Shigemori, H. *et al.* Fellutamides A and B, cytotoxic peptides from a marine fish-

possessing fungus *Penicillium fellutanum*. *Tetrahedron* **47,** 8529–8534 (1991).

163. Hirao, T., Tsuge, N., Imai, S., Shin-Ya, K. & Seto, H. Nerfilin I, a novel microbial metabolite inducing neurite outgrowth of PC12 cells. *J. Antibiot. (Tokyo).* **48,** 1494–1496 (1995).

164. Oda, K., Fukuda, Y., Murao, S., Uchida, K. & Kainosho, M. A novel proteinase inhibitor, tyrostatin, inhibiting some pepstatin-insensitive carboxyl proteinases. *Agric. Biol. Chem.* **53,** 405–415 (1989).

165. Momose, I. *et al.* Tyropeptins A and B, new proteasome inhibitors produced by *Kitasatospora* sp. MK993-dF2. I. Taxonomy, isolation, physico-chemical properties and biological activities. *J. Antibiot. (Tokyo).* **54,** 997–1003 (2001).

166. Cruz Morales, P., Barona Gómez, F. & Ramos Aboites, H. E. Genetic System for Producing a Proteases Inhibitor of a Small Peptide Aldehyde Type. WO/2016/097957 (2015).

167. Chen, Y., McClure, R. A., Zheng, Y., Thomson, R. J. & Kelleher, N. L. Proteomics guided discovery of flavopeptins: anti-proliferative aldehydes synthesized by a reductase domain-containing non-ribosomal peptide synthetase. *J. Am. Chem. Soc.* **135,** 10449–10456 (2013).

168. Cruz-Morales, P. *et al.* The genome sequence of *Streptomyces lividans* 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island. *Genome Biol. Evol.* **5,** 1165–75 (2013).

169. Kim, I. S. & Lee, K. J. Physiological roles of leupeptin and extracellular proteases in mycelium development of *Streptomyces exfoliatus* SMF13. *Microbiology* **141,** 1017–1025 (1995).

170. Kim, I. S., Kim, Y. B. & Lee, K. J. Characterization of the leupeptin-inactivating enzyme from *Streptomyces exfoliatus* SMF13 which produces leupeptin. *Biochem J* **545,** 539–545 (1998).

171. Kim, D. W. *et al.* Complex extracellular interactions of proteases and a protease inhibitor influence multicellular development of *Streptomyces coelicolor*. *Mol. Microbiol.* **70,** 1180–1193 (2008).

172. Sánchez-Serrano, I. Success in translational research: lessons from the development of bortezomib. *Nat. Rev. Drug Discov.* **5,** 107–114 (2006).

173. Becker, C., Watson, A. J. & Neurath, M. F. Complex roles of caspases in the pathogenesis

of inflammatory bowel disease. *Gastroenterology* **144,** 283–293 (2013).

174. Graybill, T. L., Dolle, R. E., Helaszek, C. T., Miller, R. E. & Ator, M. A. Preparation and evaluation of peptidic aspartyl hemiacetals as reversible inhibitors of interleukin-1ß converting enzyme (ICE). *Int. J. Pept. Protein Res.* **44,** 173–182 (1994).

175. Kisselev, A. F. & Goldberg, A. L. Proteasome inhibitors: from research tools to drug candidates. *Chem. Biol.* **8,** 739–758 (2001).

176. Vergnolle, N. Protease inhibition as new therapeutic strategy for GI diseases. *Gut* **65,** 1215–1224 (2016).

177. Edgington-Mitchell, L. E. Pathophysiological roles of proteases in gastrointestinal disease. *Am. J. Physiol. - Gastrointest. Liver Physiol.* **310,** G234–G239 (2016).

178. Steck, N., Mueller, K., Schemann, M. & Haller, D. Bacterial proteases in IBD and IBS. *Gut* **61,** 1610–1618 (2012).

179. Carroll, I. M. & Maharshak, N. Enteric bacterial proteases in inflammatory bowel disease: pathophysiology and clinical implications. *World J. Gastroenterol.* **19,** 7531–7543 (2013).

180. Macfarlane, G. T. & Macfarlane, S. Bacteria, colonic fermentation, and gastrointestinal health. *J. AOAC Int.* **95,** 50–60 (2012).

181. Smith, E. A. & MacFarlane, G. T. Enumeration of amino acid fermenting bacteria in the human large intestine: Effects of pH and starch on peptide metabolism and dissimilation of amino acids. *FEMS Microbiol. Ecol.* **25,** 355–368 (1998).

182. Dai, H. & Korthuis, R. J. Mast cell proteases and inflammation. *Drug Discov. Today Dis. Model.* **8,** 47–55 (2011).

183. Mowat, A. M. & Bain, C. C. Mucosal macrophages in intestinal homeostasis and inflammation. *J. Innate Immun.* **3,** 550–564 (2011).

184. Medina, C. Role of matrix metalloproteinases in intestinal inflammation. *J. Pharmacol. Exp. Ther.* **318,** 933–938 (2006).

185. Turk, V. *et al.* Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochim. Biophys. Acta - Proteins Proteomics* **1824,** 68–88 (2012).

186. Van Spaendonk, H. *et al.* Regulation of intestinal permeability: The role of proteases. *World J. Gastroenterol.* **23,** 2106–2123 (2017).

187. Fitzpatrick, L. R. Evidence that the ubiquitin proteasome system plays a prominent role in inflammatory bowel disease: Possible pharmacological approaches. *Pharm. Pharmacol. Int. J.* **4,** 308–309 (2016).

188. Ferrington, D. A. & Gregerson, D. S. Immunoproteasomes: structure, function, and antigen presentation. *Prog. Mol. Biol. Transl. Sci.* **109,** 75–112 (2012).

189. Meyer-Hoffert, U. & Wiedow, O. Neutrophil serine proteases: Mediators of innate immune responses. *Curr. Opin. Hematol.* **18,** 19–24 (2011).

190. Menzel, K. *et al.* Cathepsins B, L and D in inflammatory bowel disease macrophages and potential therapeutic effects of cathepsin inhibition in vivo. *Clin. Exp. Immunol.* **146,** 169–180 (2006).

191. Cuzzocrea, S. *et al.* Calpain inhibitor I reduces colon injury caused by dinitrobenzene sulphonic acid in the rat. *Gut* **48,** 478–488 (2001).

192. Loher, F. *et al.* The interleukin-1β-converting enzyme inhibitor pralnacasan reduces dextran sulfate sodium-induced murine colitis and T helper 1 T-cell activation. *J. Pharmacol. Exp. Ther.* **308,** 583–590 (2004).

193. Schmidt, N. *et al.* Targeting the proteasome: Partial inhibition of the proteasome by bortezomib or deletion of the immunosubunit LMP7 attenuates experimental colitis. *Gut* **59,** 896–906 (2010).

194. Basler, M., Dajee, M., Moll, C., Groettrup, M. & Kirk, C. J. Prevention of experimental colitis by a selective inhibitor of the immunoproteasome. *J. Immunol.* **185,** 634–641 (2010).

195. Ivanov, D. *et al.* A serpin from the gut bacterium *Bifidobacterium longum* inhibits eukaryotic elastase-like serine proteases. *J. Biol. Chem.* **281,** 17246–17252 (2006).

196. Mkaouar, H. *et al.* Siropins, novel serine protease inhibitors from gut microbiota acting on human proteases involved in inflammatory bowel diseases. *Microb. Cell Fact.* **15,** 201 (2016).

197. Macfarlane, G. T., Macfarlane, S. & Gibson, G. R. Synthesis and release of proteases by *Bacteroides fragilis*. *Curr. Microbiol.* **24,** 55–59 (1992).

198. Gibson, S. A. W., McFarlan, C., Hay, S. & MacFarlane, G. T. Significance of microflora in proteolysis in the colon. *Appl. Environ. Microbiol.* **55,** 679–683 (1989).

199. Allison, C. & Macfarlane, G. T. Physiological and nutritional determinants of protease

secretion by *Clostridium sporogenes*: characterization of six extracellular proteases. *Appl. Microbiol. Biotechnol.* **37,** 152–156 (1992).

200. McLuskey, K. *et al.* Crystal structure and activity studies of the C11 cysteine peptidase from *Parabacteroides merdae* in the human gut microbiome. *J. Biol. Chem.* **291,** 9482–9491 (2016).

201. Roncase, E. J. *et al.* Substrate profiling and high resolution co-complex crystal structure of a secreted C11 protease conserved across commensal bacteria. *ACS Chem. Biol.* **12,** 1556–1565 (2017).

202. Hörmannsperger, G., von Schillde, M. A. & Haller, D. Lactocepin as a protective microbial structure in the context of IBD. *Gut Microbes* **4,** 152–157 (2013).

203. Sessenwein, J. L. *et al.* Protease-mediated suppression of DRG neuron excitability by commensal bacteria. *J. Neurosci.* **37,** 11758–11768 (2017).

204. Finkelstein, R. A. & Hanne, L. F. Purification and characterization of the soluble hemagglutinin (cholera lectin) produced by *Vibrio cholerae*. *Infect. Immun.* **36,** 1199–1208 (1982).

205. Vaitkevicius, K. *et al.* A *Vibrio cholerae* protease needed for killing of *Caenorhabditis elegans* has a role in protection from natural predator grazing. *Proc. Natl. Acad. Sci.* **103,** 9280–9285 (2006).

206. Sikora, A. E., Zielke, R. a., Lawrence, D. a., Andrews, P. C. & Sandkvist, M. Proteomic analysis of the *Vibrio cholerae* type II secretome reveals new proteins, including three related serine proteases. *J. Biol. Chem.* **286,** 16555–16566 (2011).

207. Ruiz-Perez, F. *et al.* Serine protease autotransporters from *Shigella flexneri* and pathogenic *Escherichia coli* target a broad range of leukocyte glycoproteins. *Proc. Natl. Acad. Sci.* **108,** 12881–12886 (2011).

208. Dautin, N. Serine protease autotransporters of Enterobacteriaceae (SPATEs): Biogenesis and function. *Toxins (Basel).* **2,** 1179–1206 (2010).

209. In, J. *et al.* Serine protease EspP from enterohemorrhagic *Escherichia coli* is sufficient to induce Shiga toxin macropinocytosis in intestinal epithelium. *PLoS One* **8,** e69196 (2013).

210. Steck, N. *et al. Enterococcus faecalis* metalloprotease compromises epithelial barrier and contributes to intestinal inflammation. *Gastroenterology* **141,** 959–971 (2011).

211. Maharshak, N. *et al. Enterococcus faecalis* gelatinase mediates intestinal permeability via

protease-activated receptor 2. *Infect. Immun.* **83,** 2762–2770 (2015).

212. Klingler, D. & Hardt, M. Profiling protease activities by dynamic proteomics workflows. *Proteomics* **12,** 587–596 (2012).

213. Mayers, M. D., Moon, C., Stupp, G. S., Su, A. I. & Wolan, D. W. Quantitative metaproteomics and activity-based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. *J. Proteome Res.* **16,** 1014–1026 (2017).

214. Morris, L. S. & Marchesi, J. R. Current functional metagenomic approaches only expand the existing protease sequence space, but does not presently add any novelty to it. *Curr. Microbiol.* **70,** 19–26 (2014).

215. Tanca, A. *et al.* Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* **5,** 79 (2017).

216. Coulter, S. N. *et al. Staphylococcus aureus* genetic loci impacting growth and survival in multiple infection environments. *Mol. Microbiol.* **30,** 393–404 (1998).

217. Qin, X., Singh, K. V., Weinstock, G. M. & Murray, B. E. Effects of *Enterococcus faecalis fsr* genes on production of gelatinase and a serine protease and virulence. *Infect. Immun.* **68,** 2579–2586 (2000).

2. Using Bioinformatics and Protein Biochemistry to Predict the Most Likely Product(s) of the *rup* Gene Cluster, the Ruminopeptin(s)

This chapter is an unofficial adaptation of previously published work.[1]

2.1. Introduction

As discussed in Chapter 1, in this portion of my thesis I developed a workflow for characterizing the structures and bioactivities of predicted natural products produced by the human gut microbiota. The growing amount of primary sequence data available from this environment, combined with the bioinformatic tools available to predict the products of biosynthetic gene clusters, has revealed that this community has the capacity to produce a diversity of natural products. Though many methods have been developed for discovering small molecules from this environment, little is known about the important functional roles these natural products may play. Recognizing the difficulty of prioritizing and isolating natural products from the anaerobic organisms that live in the healthy gut community, we set out with the goal of combining bioinformatics with in vitro biochemical assays to predict the structures of putative bioactive natural products. Due to its abundance, well-studied ecological role, and the interesting compound(s) it is predicted to produce, we initially focused on an NRPS gene cluster from the prominent gut commensal *Ruminococcus bromii.*

R. bromii is one of the most abundant microbes in the human gut across a diversity of environments and diets. In a study of 46 healthy adults in Australia, phylotypes closely related (>97% ID) to *R. bromii* were found to comprise 4.4% of total 16S rRNA gene abundance.[2] In an isolation-based study, *R. bromii* made up from 1 – 5.2% of gut isolates in the geographically

diverse populations of Japanese-Hawaiians, North American Caucasians, rural native Japanese, and rural native Africans.[3] *R. bromii* is a member of Clostridium cluster IV, which contains organisms that are generally considered to be beneficial in the gut environment.[4,5] Clostridium cluster IV strains have been shown to induce regulatory T cells in the colon,[6] and this grouping includes *Faecalibacterium prausnitzii*, which has a well-studied anti-inflammatory role.[7] In a study of 59 patients in India measuring 16S rRNA sequence abundances, organisms from Clostridium cluster IV were found to be significantly less abundant in patients with inflammatory bowel disease (IBD) as compared with healthy subjects.[8]

R. bromii already has a known important ecological role in the colon as a keystone species in the degradation of resistant starch.[9] Resistant starch is produced when high-starch food sources (e.g. potatoes, wheat) are cooked and then cooled, leading to crystallization and reordering of a network of hydrogen bonds that make it "resistant" to dissolution or enzymatic degradation.[10] Several detailed studies relating *R. bromii* and resistant starch have been performed by Flint and coworkers.[9,11,12] In 2008, they used 16S rRNA sequencing to determine what species are associated with insoluble particulates in human fecal samples, and they found that *R. bromii* is among the species enriched in this phase.[11] In 2012, they demonstrated *R. bromii*'s "keystone" role in degrading resistant starch, by showing that it promotes starch utilization in co-cultures with other bacteria that normally cannot access this dietary substrate.[9] In 2015, they reported the unique organization of extracellular amylases in *R. bromii*, which are likely responsible for its resistant starch degrading capability.[12] The products of this initial starch degradation are simpler mono-, di-, and trisaccharides, some of which can be used by *R. bromii* itself as a nutrient source but others which are able to serve as nutrients for other gut microbial species.[9] As these sugars are an important source of nutrients for the microbes that exert beneficial health effects in the

colon, including those that produce short chain fatty acids (SCFAs), *R. bromii* likely plays an important role in gastrointestinal health.[12]

Though to our knowledge *R. bromii* has not yet been reported to produce natural products, we hypothesized that this could be another mechanism by which this organism exerts its beneficial effects or maintains its ecological niche in the human gut. In this chapter, I describe our efforts to investigate the biosynthetic product of an NRPS gene cluster that is conserved among strains of *R. bromii* isolated from the human gut. From bioinformatic predictions, we determined that this gene cluster would likely produce a peptide aldehyde product, and though we could not isolate aldehyde compounds from cultures of this organism, we have used partial in vitro biosynthetic reconstitutions to gain insights about the likely product(s) produced by this gene cluster.

2.2. Results and discussion

2.2.1. The *rup* gene cluster from *R. bromii* is abundant in the commensal human gut microbiota and is evolutionarily conserved.

The *rup* gene cluster (also known as bgc45) has been identified previously by Fischbach and co-workers in a large survey of biosynthetic gene clusters from the human microbiome and is part of a larger family of NRPS gene clusters found in gut microbial genomes and metagenomes.[13] This study also revealed the *rup* gene cluster to be one of the most abundant gene clusters found in human microbiome project (HMP) stool metagenomes (relative to the other clusters identified in this study). Moreover, a highly similar gene cluster (bgc71, 97.2% nucleotide sequence identity) from a closely related, unisolated *Ruminococcus* species was identified in several RNAseq datasets from stool samples of healthy subjects, indicating that this

biosynthetic pathway is likely expressed under physiological conditions.[14] Overall, these findings suggest the product of the *rup* gene cluster is likely produced under physiological conditions. Coupled with the established importance of *R. bromii*, this may indicate a particularly important role for this metabolite in the human gut microbiota.

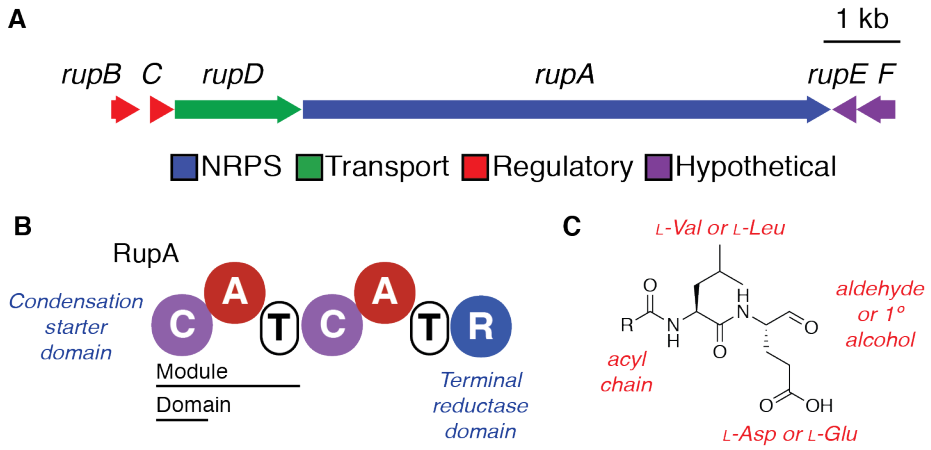


Figure 2.1: A biosynthetic gene cluster from the abundant gut commensal *Ruminococcus bromii*.

(A) The *rup* gene cluster from *R. bromii*. The gene cluster encodes a single multi-module NRPS, a transporter, two regulatory elements, and two hypothetical proteins. (B) The RupA NRPS contains a condensation-starter (C-starter) domain and a terminal reductase (R) domain (A = adenylation domain, T = thiolation domain). (C) Predicted scaffold of the *rup* cluster product, ruminopeptin.

Table 2.1: Predicted gene functions in the *rup* gene cluster.

Name	Locus tag	Annotation
rupB	RBR_11740 CBL15462	Anti-sigma factor
rupC	RBR_11730 CBL15461	Anti-anti-sigma factor
rupD	RBR_11720 CBL15460	Na+ driven multidrug efflux pump
rupA	RBR_11710 CBL15459	Nonribosomal peptide synthetase (C-A-T-C-A-T-R)
rupE	RBR_11700 CBL15458	Hypothetical protein
rupF	RBR_11690 CBL15457	Hypothetical protein

The 10.9 kb *rup* gene cluster encodes a single two-module NRPS, an efflux pump (ABC transporter), two regulatory elements, and two hypothetical proteins (Figure 2.1, Table 2.1). Based on gene content and NRPS biosynthetic logic, we predicted that the *rup* gene cluster would produce a peptide aldehyde natural product. The NRPS (RupA) features a condensation-starter (C-starter) domain in its first module, indicating that the N-terminus of the product nonribosomal peptide is likely *N*-acylated,[15] one complete NRPS module, and a terminal reductase (R) domain. This final domain should catalyze release of a nascent thioester intermediate from the NRPS enzyme, generating either an aldehyde or a primary alcohol-containing product.[16] As discussed in Chapter 1, a peptide aldehyde product would likely act as an inhibitor of serine, cysteine, or threonine proteases as has been demonstrated for other NRPS-derived peptide aldehydes produced by soil microbes (e.g. fellutamide B[17] and the flavopeptins[18]). Notably, Ruminococceae are negatively correlated with protease activity in fecal samples,[19] and production of small molecule protease inhibitors by these organisms is a potential mechanism by which this association could arise.

If the product of the *rup* gene cluster does play a crucial role in *R. bromii*'s ecology and evolutionary history, we might expect it to be highly conserved in this species. To assess the

presence of this gene cluster across *R. bromii* strains, we used PCR with specific primers to amplify a fragment of the first NRPS adenylation domain ($RupA_{A1}$) in three available human-derived *R. bromii* isolates (*R. bromii* L2-63, *R. bromii* ATCC 27255, and *R. bromii* 22-5-S 6 FAA NB).[9,20] We observed amplification in each strain (Figure 2.2). We then subsequently PCR-amplified and sequenced the full gene clusters to reveal greater than 96% identity on the nucleotide level (Table 2.2). Conservation of this biosynthetic gene cluster across these three human-derived *R. bromii* isolates provides evidence that this pathway may be important for the organism's native biological role.

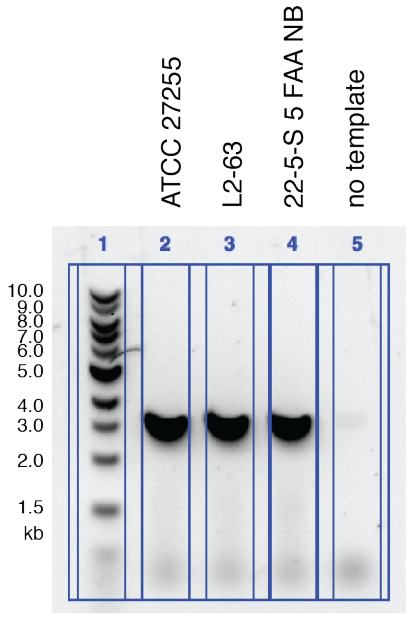


Figure 2.2: **Identification of the *rup* gene cluster in *R. bromii* strains.**

Agarose gel of the PCR amplification products of the RupA A_1 domain from *R. bromii* strains ATCC 27255, L2-63, and 22-5-S 6 FAA NB. Primers used in this experiment were rupDetect-1 and rupDetect-2 (Table 2.8).

Table 2.2: %ID among *rupA* nucleotide sequences.

	R. bromii 22-5-6 S FAA NB	*R. bromii* ATCC 27255	*R. bromii* L2-63
R. bromii 22-5-6 S FAA NB		97.1	97.8
R. bromii ATCC 27255	97.1		96.9
R. bromii L2-63	97.8	96.9	

2.2.2. Bioinformatic analysis of the *rup* gene cluster predicts that it produces an *N*-acylated dipeptide aldehyde

In order to gain information about the product of the rup gene cluster, we first used bioinformatic analyses to predict the activities and substrate specificities of each of the domains in this two-module NRPS assembly line. RupA lacks an adenylation-thiolation (A-T) didomain loading module and instead contains a predicted C-starter domain. C-starter domains catalyze *N*-acylation of initially loaded, assembly-line tethered amino acids with fatty acyl-CoAs. Multiple sequence alignments with biochemically and genetically characterized C-starter domains (ClbN, XcnA, and GlbF) revealed the RupA C-starter domain contains key conserved residues indicative of *N*-acylation activity (Figure 2.4).[15,21–23] We then used the University of Maryland's PKS/NRPS Analysis Web-site[24] to predict the substrate specificities of the two A domains of RupA. We found that the first NRPS module likely preferred L-leucine and the second NRPS module likely used either L-aspartate or L-glutamate (Figure 2.3). Finally, we generated a structure-based multiple sequence alignment of the final $RupA_R$ domain with other characterized NRPS terminal R domains using PROMALS3d.[25] From this alignment, we could identify all of the key conserved active site residues involved in NAD(P)H binding as well as the Thr/Tyr/Lys catalytic triad required for thioester reduction (Figure 2.5).[26]

Position	**235**	**236**	**239**	**278**	**299**	**301**	**322**	**330**	**Substrate/prediction**
GrsA	D	A	W	T	I	A	A	I	L-Phe
RupA A_1	**D**	**A**	**S**	**F**	**L**	**G**	**G**	**V**	**L-Leu (predicted)**
SrfAA A_3	D	A	W	F	L	G	N	V	L-Leu
RupA A_2	**D**	**M**	**K**	**N**	**L**	**G**	**T**	**V**	**L-Glu (predicted)**
SrfAA A_1	D	A	K	D	L	G	G	V	L-Glu

Figure 2.3: **Predicted A domain specificity-conferring residues (Stachelhaus codes) for RupA.**

Specificity-conferring residues were identified using the University of Maryland's PKS/NRPS Analysis Web-site.[24] Reference codes are from the initial identification of key A domain residues by Stachelhaus et al.[27] Numbering of positions references the sequence of phenylalanine-activating A domain GrsA.

Figure 2.4: ClustalW2 alignment of C domains shows residues that distinguish $^{L}C_{L}$ and C-starter activities.

C-domains were identified with the University of Maryland's PKS/NRPS Analysis Web-site and further trimmed from the beginning of conserved motif C1 to the end of motif C5.[15] The multiple sequence alignment was generated using ClustalW2 and visualized with Geneious. Included are the sequences of bgc35 C_1 (*Clostridium* sp. KLE 1755 (bgc35), ERI72059.1), ClbN C_1 (*E. coli*, Q0P7K4), GlbF C_1 (*[Polyangium] brachysporum*, CAL80824.1), RupA C_1 (*R. bromii*, YP_007781236.1), AsfA C_1 (*Clostridium* sp. ASF502, WP_004068886.1), SrfAA C_1 (*Bacillus subtilis*, NP_388230.1), SrfAB C_2 (*Bacillus subtilis*, Q04747), DptA C_4 (*Streptomyces filamentosus*, AAX31557.1), and RupA C_2 (*R. bromii*, YP_007781236.1). ClbN C_1, GlbF C_1, and SrfAA C_1 are C-starter domains, and DptA C_4 and SrfAB C_2 are standard $^{L}C_{L}$ domains. The four number positions used in this analysis to distinguish C-starter domains and $^{L}C_{L}$ domains were identified based on work by Huson and coworkers[15]: 1 (S/G/A vs. P), 2 (L/I/A/M vs. T), 3 (V/I/L vs. A), 4 (P vs. A/V).

Figure 2.4 (continued)

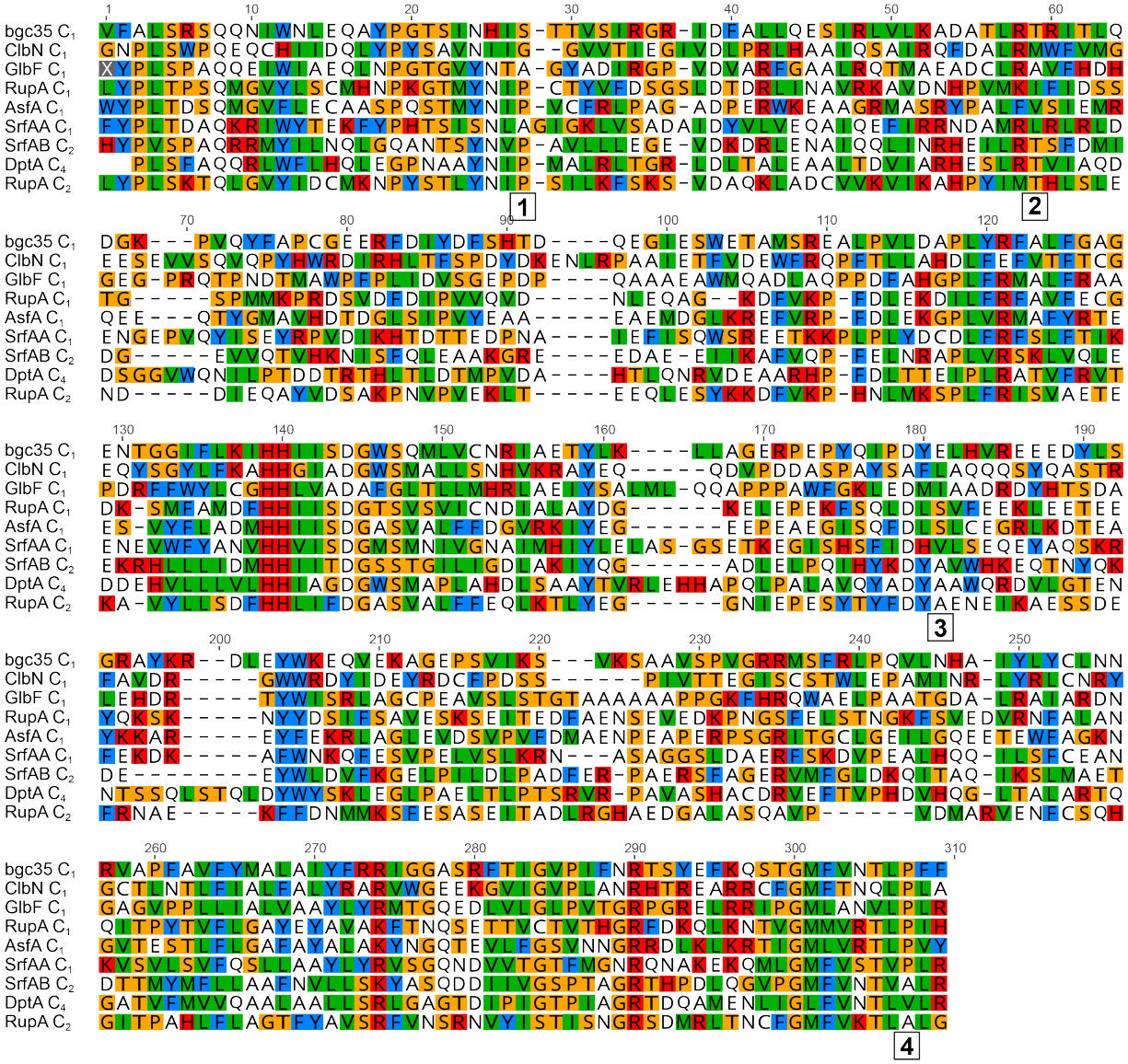
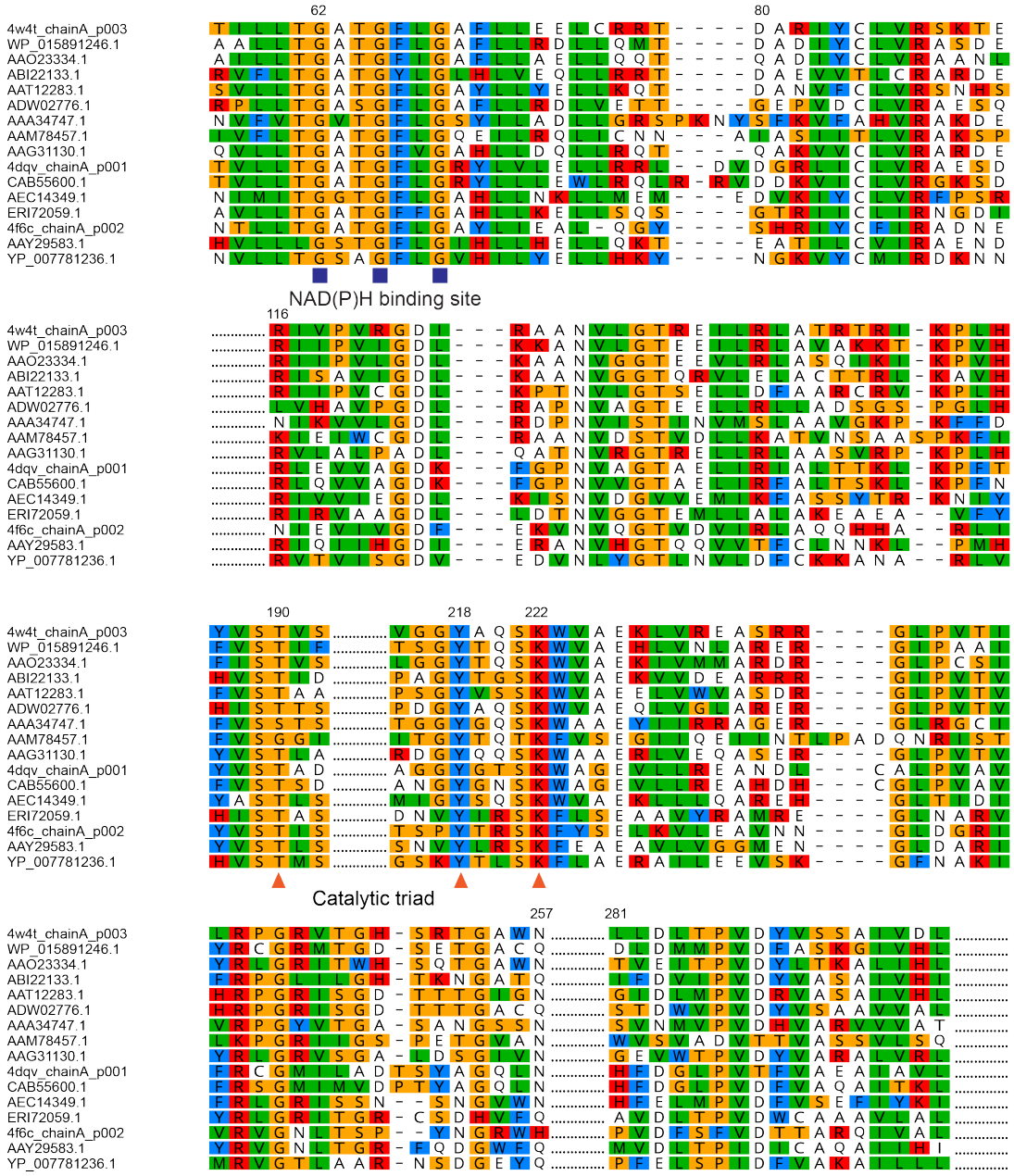


Figure 2.5: Structure-based alignment of terminal reductase domains from NRPS pathways shows conservation of residues important for catalysis.

The multiple sequence alignment was generated using PROMALS3d and visualized with Geneious. Included are the sequences from pathways producing myxalamid (*Stigmatella aurantiaca*, 4w4t), gramicidin (*Brevibacillus brevis*, WP_015891246.1), nostocyclopeptide (*Nostoc* sp. ATCC 53789, AAO23334.1), saframycin (*Streptomyces lavendulae*, ABI22133.1), lyngbyatoxin (*Lyngbya majuscula*, AAT12283.1), flavopeptin (*Streptomyces pratensis* ATCC 33331, ADW02776.1), lys2, (*Saccharomyces cerevisiae*, AAA34747.1), peptaibol (*Trichoderma virens*, AAM78457.1), myxochelin (*Stigmatella aurantiaca*, AAG31130.1), putative isonitrile lipopeptide[28] (product of *Rv0096–0101* gene cluster, *Mycobacterium tuberculosis*, AIR12822.1/4dqv), glycopeptidolipid (*Mycobacterium smegmatis* str. MC2 155, CAB55600.1), koranimide (*Bacillus* sp. NK2003, AEC14349.1), PZN2 (bgc35) (*Clostridium* sp. KLE 1755, ERI72059.1), aureusimine (*Staphylococcus aureus*, 4f6c), BT peptide (*Brevibacillus texasporus*, AAY29583.1), and ruminopeptin (*Ruminococcus bromii* L2-63, YP_007781236.1). Sequences were trimmed from 5 residues upstream of the beginning of conserved motif R1.[29] Conserved residues for NAD(P)H binding and the catalytic triad are indicated.[26]

Figure 2.5 (continued)



Together, these analyses allowed us to propose a biosynthetic hypothesis for the *rup* pathway and predict the structure of the final peptide aldehyde product(s), which we named ruminopeptin (Figure 2.6). After post-translational modification of the RupA T domains by a phosphopantetheinyl (ppant) transferase, initiation of biosynthesis occurs with the activation of L-leucine by the A domain of the first NRPS module and loading onto the ppant arm of the first T domain. The C-starter domain of the first module then acylates the amino group of the tethered L-leucine with a fatty acyl CoA. The resulting *N*-acylated aminoacyl thioester intermediate is then elongated by amide bond formation with the amino acid loaded by the second NRPS module, either L-aspartate or L-glutamate, to generate a nascent *N*-acylated dipeptide thioester intermediate. Finally, reductive offloading of this intermediate by the R domain will give a peptide aldehyde product, ruminopeptin.

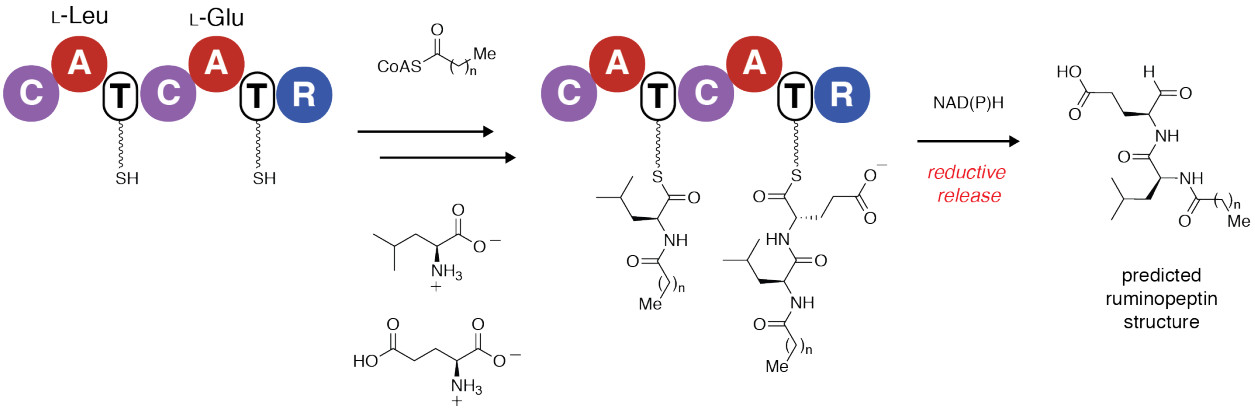
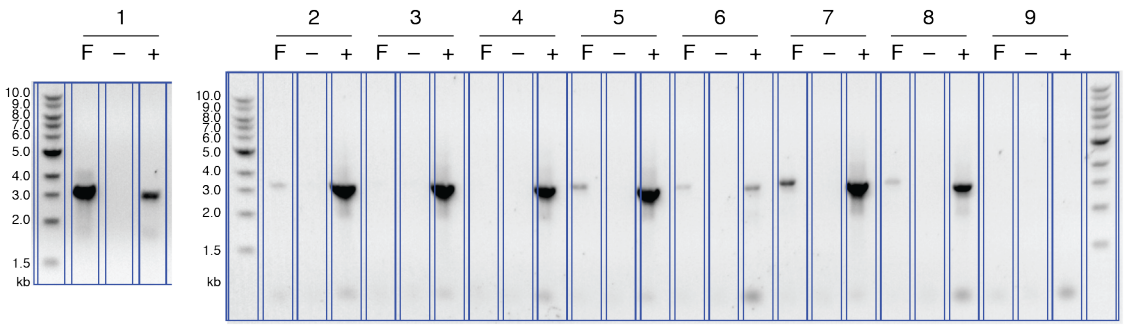


Figure 2.6: Biosynthetic hypothesis for RupA

Biosynthetic hypothesis for the production of ruminopeptin by the Rup pathway. We predict that RupA A_1 activates and loads L-leucine (or L-valine), which is then acylated with an acyl CoA by RupA C_1. The second module of RupA extends this tethered intermediate with L-glutamate (or L-aspartate), leading to a tethered *N*-acylated dipeptide. Finally, the terminal R domain uses NAD(P)H to perform a $2e^-$ reduction on this intermediate to form a peptide aldehyde. It is also possible that this domain catalyzes a second reduction on the aldehyde to ultimately produce an alcohol product.

2.2.3. The *rup* gene cluster is expressed under standard culture conditions, but no aldehydes can be isolated from *R. bromii* cultures

Having proposed a candidate structure for the *rup* gene cluster product(s), we wondered if it would be possible to isolate these secondary metabolites from *R. bromii*. We began by identifying culture conditions in which the *rup* pathway was expressed. We cultivated two strains of *R. bromii* using a variety of nutrient sources and several unusual culture additives (rumen fluid, chopped meat broth). We extracted RNA from saturated cultures and assessed gene cluster expression using specific primers with single-step RT-PCR. We observed that including fructose as a carbohydrate source in growth media was necessary for *rup* gene cluster expression and that inclusion of additives had no effect (Figure 2.7). However, in numerous attempts to extract metabolites from cultures grown under conditions where the *rup* genes were expressed (5 mL to 1 L scales), we could not identify candidate masses corresponding to any predicted ruminopeptin peptide aldehyde products by LC-MS.

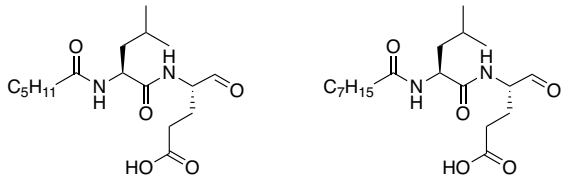


Condition	Strain	Carbohydrate	Additive	Band observed?
1	27255	fructose	–	yes
2	L2-63	fructose	–	yes
3	L2-63	maltose	–	no
4	27255	maltose	–	no
5	L2-63	fructose	10 mM hexanoic acid	yes
6	27255	fructose	10% rumen fluid	yes
7	L2-63	fructose	10% chopped meat broth	yes
8	27255	fructose	10% chopped meat broth	yes
9	No template	–	–	no

Figure 2.7: Detection of *rup* gene cluster expression by RT-PCR.

Agarose gel of RT-PCR amplification of the RupA A_1 domain from total RNA extracted from *R. bromii* strains. Primers used in this experiment were rupDetect-1 and rupDetect-2. Shown are the full reaction mixtures (F), reaction mixtures containing no reverse transcriptase (–), and reactions containing primers fD1 and rP2[30] for amplifying 16S rRNA (+).

Peptide aldehydes are prone to a variety of potential degradation pathways, including C–N bond cleavage by peptidases, reduction to alcohols, oxidation to carboxylic acids, and epimerization.[31] We therefore considered if our inability to isolate putative peptide aldehyde-based natural products from *R. bromii* was due to these compound(s) being unstable. Indeed, when we incubated a model synthetic ruminopeptin (*N*-hexanoyl-L-Leu- L-Glu-CHO, compound **2.1**, see Chapter 3 for synthesis) in cultures of *R. bromii*, we observed rapid disappearance of the compound's mass by LC-MS and could only detect it up to 30 minutes into the incubation. Therefore, we explored an alternative strategy to identify these compounds in cultures.



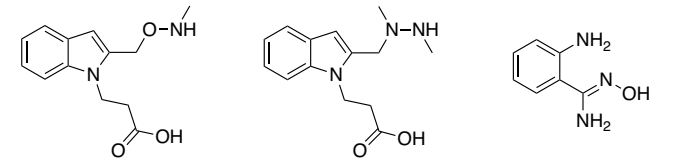
2.1 2.2

Figure 2.8: Structures of model synthetic ruminopeptins 2.1 and 2.2.
See Chapter 3 for synthesis of these compounds.

There are a variety of methods for installing aldehyde functionality on proteins,[32] and the rapid formation of hydrazines and oxime adducts has been exploited in the bioconjugation field for labeling these tagged proteins.[33] Classically, simple hydrazines and oximes that form reversible linkages have been used for this transformation,[33,34] but more recently several reagents that form hydrolytically stable adducts with these electrophiles have been developed (Figure 2.9). In 2013, Bertozzi and coworkers reported an indole-containing alkoxyamine reagent (**2.3**) for a Pictet-Spengler ligation with aldehydes and ketones.[35] Also in 2013, Rabuka and coworkers reported an indole-containing hydrazine reagent (**2.4**) and derivatives that perform a similar reaction.[36] In 2014, Derda and coworkers developed a third type of reagent, 2-aminobenzamide oxime (ABAO, **2.5**) and derivatives, which react faster than the Pictet-Spengler reagents and also have the unexpected benefit of producing products with a unique absorption spectrum.[37]

Taking inspiration from these reagents, which were primarily designed to label proteins, we wondered if we could instead label natural product peptide aldehydes produced by cultures of *R. bromii* in situ. We anticipated that this could have two advantages: both increasing the stability of these compounds and enabling their detection in a comparative fashion. Reacting cultures or extracts of *R. bromii* with ABAO and then monitoring them by HPLC for the appearance of novel peaks would be a rapid way to screen many culture conditions for aldehyde production.

Additionally, the unique mass shift generated by the reaction with an aldehyde compound would make it possible to identify new compounds by LC-MS using comparative metabolomics. A similar method was used by Mitchell and coworkers in 2016 to label unknown reactive aldehyde and ketone electrophiles with a hydroxylamine probe and prioritize natural products for isolation efforts.[38]



2.3 P. Agarwal, C. Bertozzi, et al. 2013

2.4 P. Agarwal, D. Rabuka, et al. 2013

2.5 P. Kitov, R. Derda, et al. 2014

Figure 2.9: Recently reported reagents for hydrolytically stable adduct formation with aldehydes.

The aldehyde adducts formed by ABAO contain a unique UV absorbance at 370 nm. To begin our experiments, we validated the reaction between ABAO and a model synthetic ruminopeptin (*N*-octanoyl- L-Leu- L-Glu-CHO, compound **2.2**, see Chapter 3 for synthesis) in acidic (pH 4.5) phosphate buffer and found that this reaction produced a specific peak with the desired UV absorbance (Figure 2.10A,B). We did not quantify conversion or the amount of remaining starting material in this reaction, but we did identify the product by HRMS. We subsequently confirmed that the reaction works in *R. bromii* growth media as well, both at neutral and acidic pH (conveniently, cultures of *R. bromii* naturally reach a pH of ~5 over the course of fermentation). Due to the hypothesis that ruminopeptin might be unstable in cultures of *R. bromii*, we included ABAO in the growth medium and allowed the derivatization reaction to occur over the course of growth. *R. bromii* was able to grow in cultures containing ABAO at concentrations of 10 mM and 1 mM, and we were able to identify the desired reaction product

(retention time 21.7 min) between compound **2.2** (130 µM) dosed into a growing *R. bromii* culture and ABAO (1 – 10 mM) (Figure 2.10C).

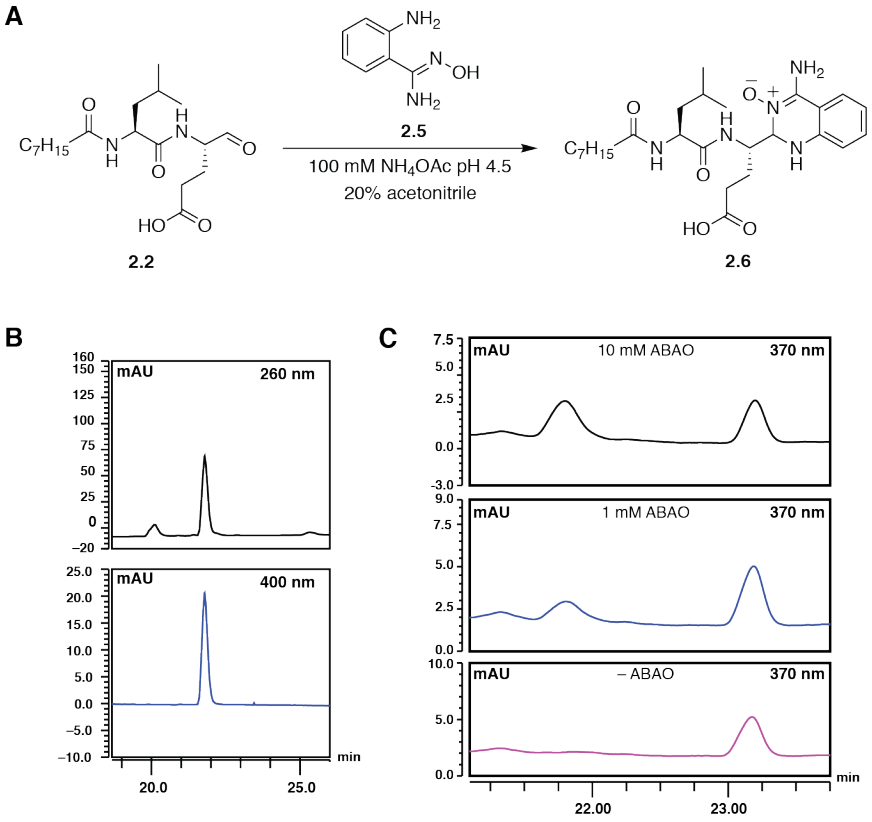


Figure 2.10: Derivatization reaction of model ruminopeptin 2.2 with ABAO.

(A) Reaction scheme. (B) The reaction in buffer yielded a product with a unique UV absorbance at 370 nm at ~21.7 min retention time (absorbance at 400 nm shown here). The identity of the peak was also confirmed by LC-MS (data not shown). The small peak at ~20.0 min retention time is likely an unidentified impurity. (C) Compound **2.2** reacted with ABAO in the presence of a growing *R. bromii* culture (peak at ~21.7 min) The large peak at ~23.1 min is likely an unidentified media component.

With this assay in place for detecting peptide aldehydes in *R. bromii* cultures, we were able to screen many different culture conditions for production of novel aldehyde compounds. Shannon Miller assisted with this work during her rotation in the Balskus Laboratory. We ran various experiments with different nutrient sources (fructose, maltose, glucose, cellobiose),

media additives (chopped meat broth, rumen fluid), and potential biosynthetic precursors (short and medium chain fatty acids, various combinations of amino acids). Aside from examining HPLC UV traces for the appearance of new UV400 peaks, we also performed LC-MS analyses on several experiments and then used XCMS[39] to compare the ±ABAO conditions and identify any specific ABAO adducts. Unfortunately, no candidate masses for ruminopeptin could be identified from these experiments.

2.2.4. In vitro biochemistry reveals the building blocks of the *rup* gene cluster product ruminopeptin

Since we could not readily isolate the predicted product(s) of the *rup* gene cluster, we sought to reconstitute this pathway in vitro to confirm our biosynthetic hypothesis and identify the preferred amino acid and acyl-CoA building blocks used by the NRPS assembly line. The individual modules of the RupA NRPS were expressed and purified in *Escherichia coli* as C-His_6-tagged constructs ($RupA_{C1\text{-}A1\text{-}T1}$ and $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$) (Figure 2.11). Attempts to express and purify the full RupA construct in any sort of reasonable yield were unsuccessful. To generate two functional modules by splitting the NRPS, we examined several different truncation points, relying on comparison of the domain architecture to various T- and C-domain containing constructs of other NRPS enzymes that had previously been expressed in soluble and active forms. Mapped onto the full length of RupA, these truncations overlap in 9 amino acid residues found both at the end of the first module and at the beginning of the second module.

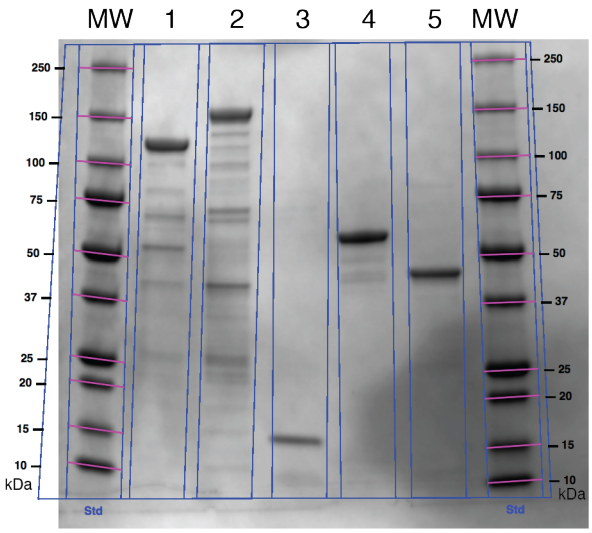


Figure 2.11: SDS-PAGE gel of purified proteins used in this study.
(1 = RupA$_{C1\text{-}A1\text{-}T1}$, 2 = RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$, 3 = RupA$_{T1}$, 4 = RupA$_{T2\text{-}R}$, 5= RupA$_{R}$, MW=Precision Plus Protein All Blue Prestained Protein Standards (BioRad)).

We used a set of standard biochemical assays to verify the activities of the two NRPS modules and determine the substrate specificities of their A domains. The ATP-[^{32}P]PP$_i$ exchange assay was used to assess amino acid activation by each individual module of RupA.[40] As discussed in Chapter 1, NRPS A domains catalyze the reaction of amino acids with ATP, generating aminoacyl adenylates and pyrophosphate. In the absence of a functional ppant arm on the downstream T domain, however, this reaction is reversible, and when run in the presence of radiolabeled pyrophosphate, this leads to the enrichment of radiolabeled ATP when the A domain is in the presence of its preferred substrate(s) (Figure 2.12A). These experiments revealed that RupA$_{C1\text{-}A1\text{-}T1}$ preferentially activates L-leucine but can also accept L-valine, while RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$ preferentially activates L-glutamate over L-aspartate (Figure 2.12B).

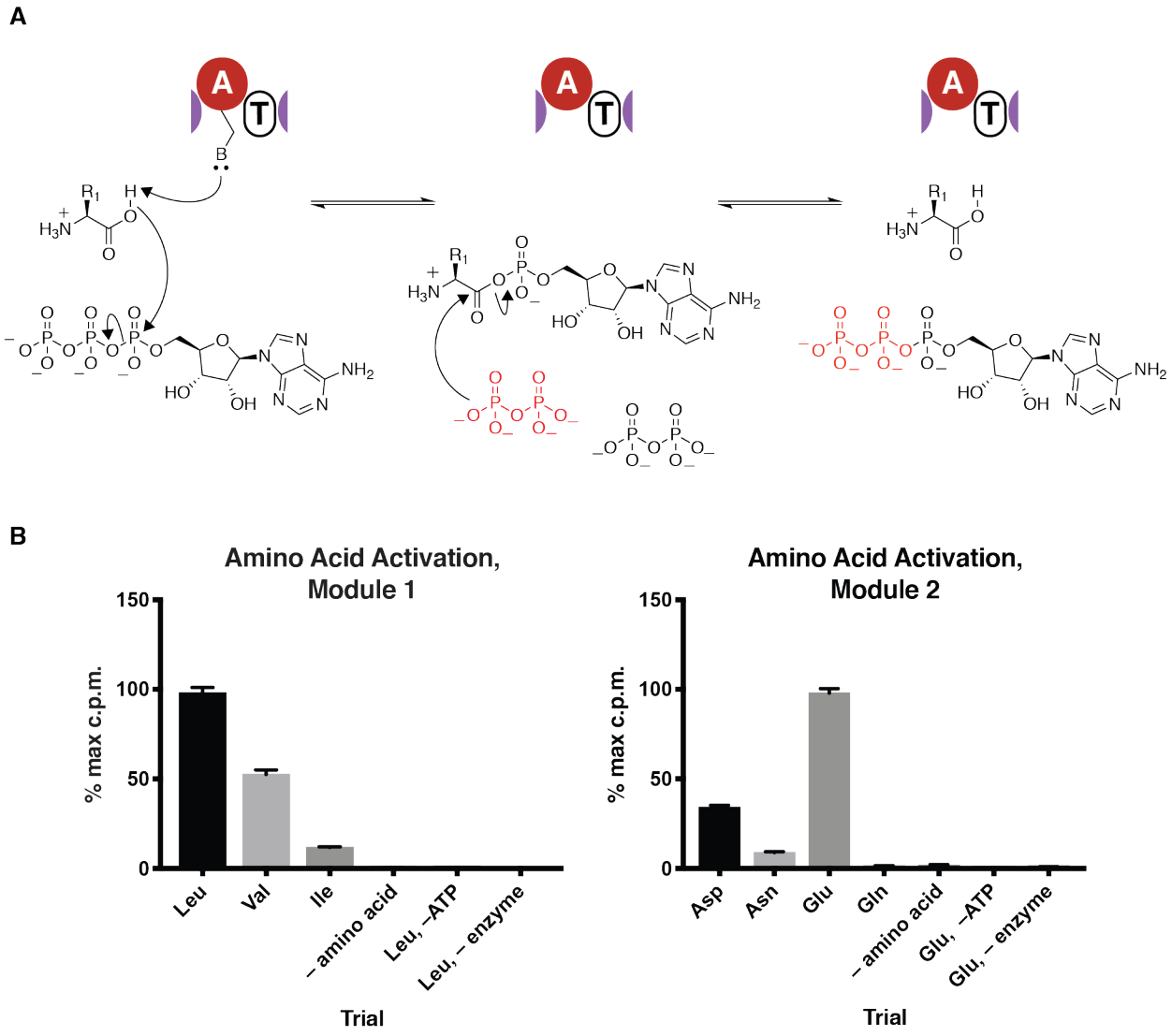


Figure 2.12: ATP-[^{32}P]PP$_i$ exchange assays for Rup$_{C1\text{-}A1\text{-}T1}$ and Rup$_{C2\text{-}A2\text{-}T2\text{-}R}$. (A) Assay scheme. (B) Results. Reactions were conducted in triplicate ("Leu, – enzyme" condition in duplicate). Error bars represent the standard deviation of three replicates (c.p.m. = counts per minute).

We then used the promiscuous ppant transferase Sfp to load BODIPY-CoA onto the T domain of each module.[41] Typically, NRPS T domains are post-translationally modified by a ppant transferase that utilizes coenzyme A as a substrate. In this assay, a fluorescently labeled BODIPY-CoA analog is used instead with the promiscuous ppant transferase Sfp (the *Bacillis*

subtilis enzyme heterologously expressed and purified from *E. coli*) to provide a fluorescently labeled ppant arm. Using this assay, we verified that Sfp can posttranslationally modify the RupA NRPS (Figure 2.13).[41] Finally, we used T domain loading assays with ^{14}C-labeled amino acids to confirm that amino acids matching our predicted specificities could be loaded on to these ppant arms.[40] In this assay, we exposed the enzyme with ATP and ^{14}C-labeled amino acids, precipitated the protein, and measured radioactivity using a liquid scintillation counter. To amplify the signal in this assay, the reactions with $RupA_{C1\text{-}A1\text{-}T1}$ were supplemented with an excess of the individually purified domain $RupA_{T1}$. From these assays, we confirmed that both L-leucine and L-valine could be loaded onto $RupA_{C1\text{-}A1\text{-}T1}$ and that both L-glutamate and L-aspartate could be loaded onto $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ (Figure 2.14).

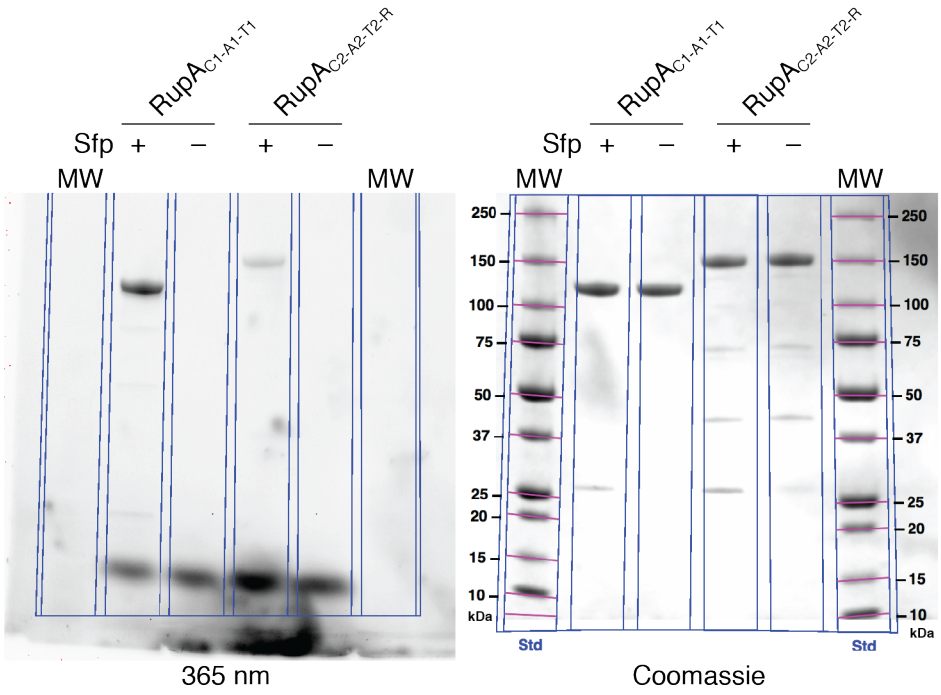


Figure 2.13: BODIPY-CoA loading assay for $Rup_{C1\text{-}A1\text{-}T1}$ and $Rup_{C2\text{-}A2\text{-}T2\text{-}R}$. (MW= Precision Plus Protein All Blue Prestained Protein Standards (BioRad)).

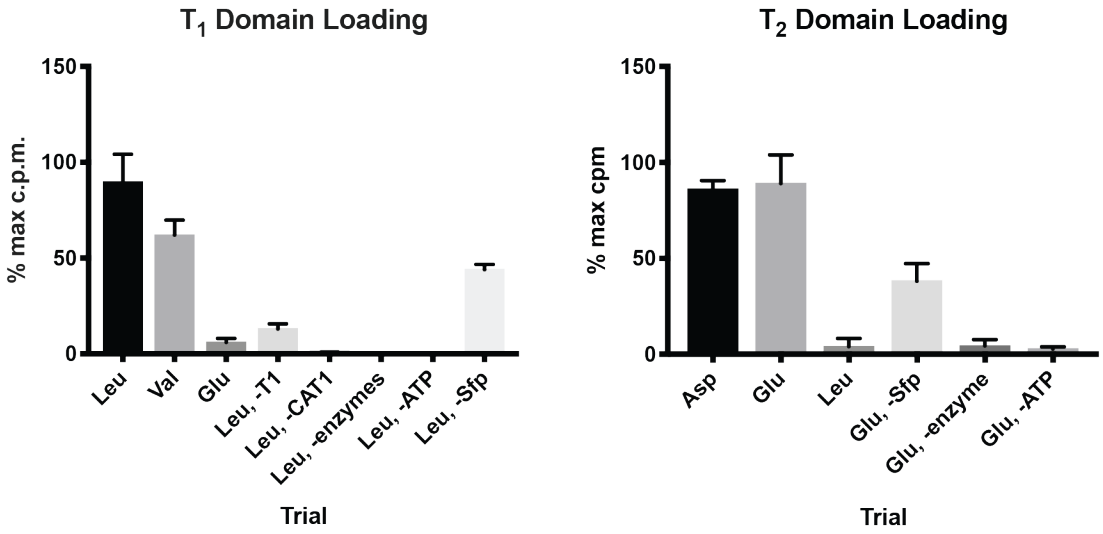


Figure 2.14: T_1 domain loading assays with $RupA_{C1\text{-}A1\text{-}T1}$ and with $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$. The assay with with $RupA_{C1\text{-}A1\text{-}T1}$ also includes excess $RupA_{T1}$. Reactions were performed in triplicate. Error bars represent the standard deviation of three replicates (c.p.m. = counts per minute).

With the amino acid building blocks established, we next set about identifying the specific acyl-CoA(s) that would be recognized and incorporated on the *N*-terminus of ruminopeptin. Though fatty acids can be incorporated into nascent polyketides and nonribosomal peptides by several mechanisms,[23,42,43] we predicted that the C-starter domain of the RupA NRPS would *N*-acylate L-leucine using a freely diffusible fatty acyl-CoA co-substrate. In order to determine this domain's preferred fatty acyl-CoA(s), we reconstituted the activity of $RupA_{C1\text{-}A1\text{-}T1}$. We incubated $RupA_{C1\text{-}A1\text{-}T1}$ with L-leucine, ATP, and a set of short, medium and long even-chain acyl-CoAs (C_2 to C_{14}) in a competition assay format. We then hydrolyzed the resulting *N*-acylated aminoacyl thioester intermediates from the NRPS for detection using LC-MS (Figure 2.15, Table 2.3, Figure 2.16).[21,23] Using this assay we identified *N*-hexanoyl-L-leucine as the most abundant product, though other medium-chain acyl-CoAs were also accepted. This result

indicates that hexanoyl-CoA and other medium-chain acyl-CoA's are likely preferred substrates of the RupA C-starter domain.

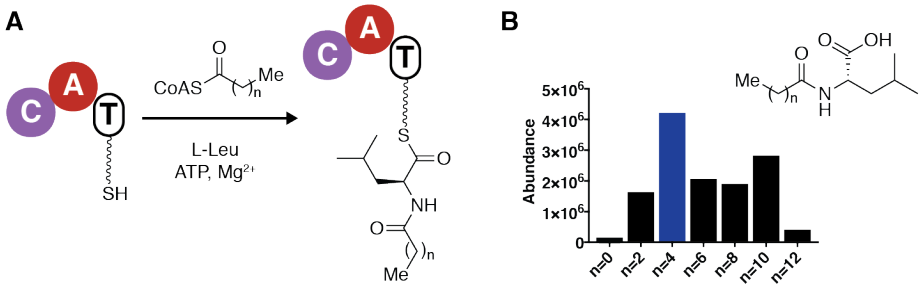


Figure 2.15: LC-MS assay for C-starter domain specificity

(A) Assay scheme. Post-translationally modified RupA$_{C1\text{-}A1\text{-}T1}$ was reacted with L-leucine, ATP, and equimolar amounts of C2−C14 fatty acyl-CoA substrates. The tethered thioesters are then hydrolyzed and detected by LC-MS. (B) Mass abundances (extracted ion chromatogram intensities) of possible products from this assay. The mass abundance of *N*-hexanoyl-L-leucine (compound **2.7**, Figure 2.16) is highlighted in blue. Representative results are shown from two independent experiments.

Table 2.3: LC-MS data for the condensation reactions of RupA$_{C1\text{-}A1\text{-}T1}$ with individual fatty acyl-CoA substrates and L-Leu.

Representative results are shown from two independent experiments.

Fatty acyl-CoA substrate	Acyl-Leu product formula	Expected mass $[M–H]^-$	Observed mass $[M–H]^-$	Error (ppm)	Area
Acetyl-CoA	$C_8H_{15}NO_3$	172.0979	172.0983	2.1	95641
Butyryl-CoA	$C_{10}H_{19}NO_3$	200.1292	200.1294	0.8	1575453
Hexanoyl-CoA	$C_{12}H_{23}NO_3$	228.1605	228.1608	1.2	4155921
Octanoyl-CoA	$C_{14}H_{27}NO_3$	256.1918	256.1921	1.0	2004235
Decanoyl-CoA	$C_{16}H_{31}NO_3$	284.2231	284.2233	0.5	1837544
Lauroyl-CoA	$C_{18}H_{35}NO_3$	312.2544	312.2544	0.1	2763249
Myristoyl-CoA	$C_{20}H_{39}NO_3$	340.2857	340.2857	0.0	359683

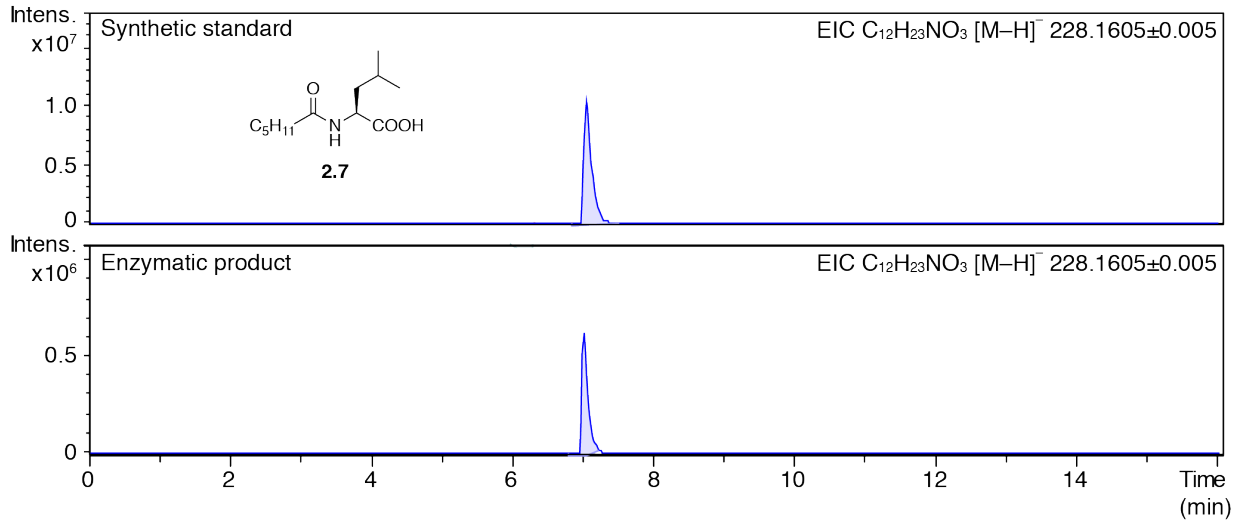


Figure 2.16: Extracted ion chromatograms for the synthetic standard 2.7 and enzymatic product from the reconstitution of RupA$_{C1\text{-}A1\text{-}T1}$.

We subsequently modified this assay to probe the activity of both NRPS modules. As we were unable to successfully express full-length RupA, we instead included the individually expressed and purified NRPS modules (RupA$_{C1\text{-}A1\text{-}T1}$ and RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$) in the reaction mixture along with amino acids, acyl-CoA substrates, and ATP. As the NAD(P)H cofactor required for the reductase domain was not provided in these initial attempts, we predicted this assay should generate T-domain-tethered *N*-acylated dipeptide thioesters which could be hydrolyzed from the enzyme and detected by LC-MS. We first performed a competition experiment to identify the preferred amino acid building blocks, including in the reaction mixture multiple amino acids (L-valine, L-leucine, L-aspartate, and L-glutamate) along with a single fatty acyl-CoA substrate (hexanoyl-CoA) (Figure 2.17, Table 2.4). The preferred product generated in this experiment incorporated L-leucine and L-glutamate. To verify the identity of the preferred acyl-CoA substrate, we next performed this assay using the preferred amino acid substrates (L-leucine and L-glutamate) and a mixture of even-chain acyl-CoA's (Figure 2.17, Table 2.5). In this experiment, *N*-hexanoyl-L-leucyl-L-glutamic acid was the most abundant product. From these

results, we concluded that RupA can produce a range of nascent T-domain tethered *N*-acylated dipeptide thioester intermediates, and may preferentially use hexanoyl-CoA, L-leucine, and L-glutamate building blocks.

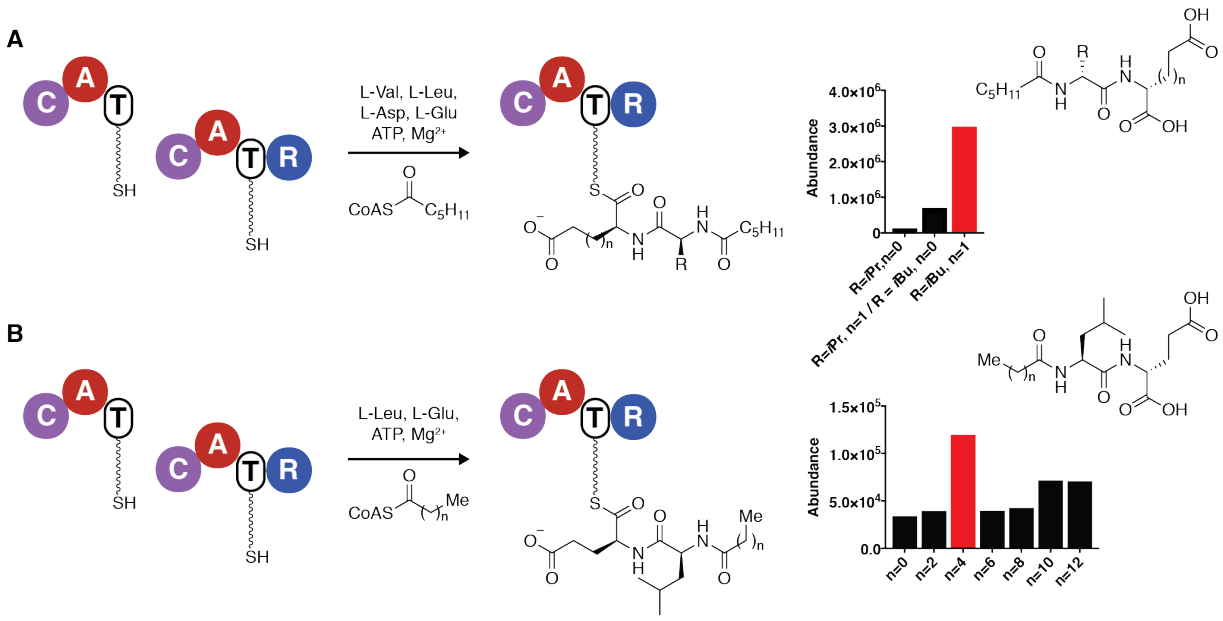


Figure 2.17: Dipeptide hydrolysis assay in two different competition formats.

LC-MS assay for tethered *N*-acylated dipeptide synthesis by RupA. Mass abundances (extracted ion chromatogram intensities) are shown for *N*-acylated dipeptide products produced in amino acid competition format (A) and acyl-CoA competition format (B). The mass abundance of *N*-hexanoyl-L-leucyl- L-glutamic acid (compound **2.8,** Figure 2.18) in each experiment is highlighted in red. Representative results are shown from two independent experiments.

Table 2.4: LC-MS data from the reconstitution of RupA$_{C1\text{-}A1\text{-}T1}$ and RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$ with hexanoyl-CoA, L-Val, L-Leu, L-Asp, and L-Glu.

Representative results are shown from two independent experiments.

Amino acid substrates	Hexanoyl-dipeptide product formula	Expected mass $[M–H]^-$	Observed mass $[M–H]^-$	Error (ppm)	Area
L-Val/ L-Asp	$C_{15}H_{26}N_2O_6$	329.1718	329.1718	0.1	20661
L-Leu/ L-Asp or L-Val/ L-Glu	$C_{16}H_{28}N_2O_6$	343.1875	343.1872	0.7	656658
L-Leu/ L-Glu	$C_{17}H_{30}N_2O_6$	357.2031	357.2029	0.6	2940200

Table 2.5: LC-MS data from the reconstitution of RupA$_{C1\text{-}A1\text{-}T1}$ and RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$ with acyl-CoAs, L-Leu, and L-Glu.

Representative results are shown from two independent experiments.

Fatty acyl-CoA substrate	Acyl-Leu-Glu dipeptide product formula	Expected mass $[M–H]^-$	Observed mass $[M–H]^-$	Error (ppm)	Area
Acetyl-CoA	$C_{13}H_{22}N_2O_6$	301.1405	301.1396	3.2	32151
Butyryl-CoA	$C_{15}H_{26}N_2O_6$	329.1718	329.1708	2.9	37857
Hexanoyl-CoA	$C_{17}H_{30}N_2O_6$	357.2031	357.2029	0.6	118228
Octanoyl-CoA	$C_{19}H_{34}N_2O_6$	385.2344	385.2330	3.7	38239
Decanoyl-CoA	$C_{21}H_{38}N_2O_6$	413.2657	413.2646	2.7	40931
Lauroyl-CoA	$C_{23}H_{42}N_2O_6$	441.2970	441.2961	2.1	70091
Myristoyl-CoA	$C_{25}H_{46}N_2O_6$	469.3283	469.3269	3.0	68932

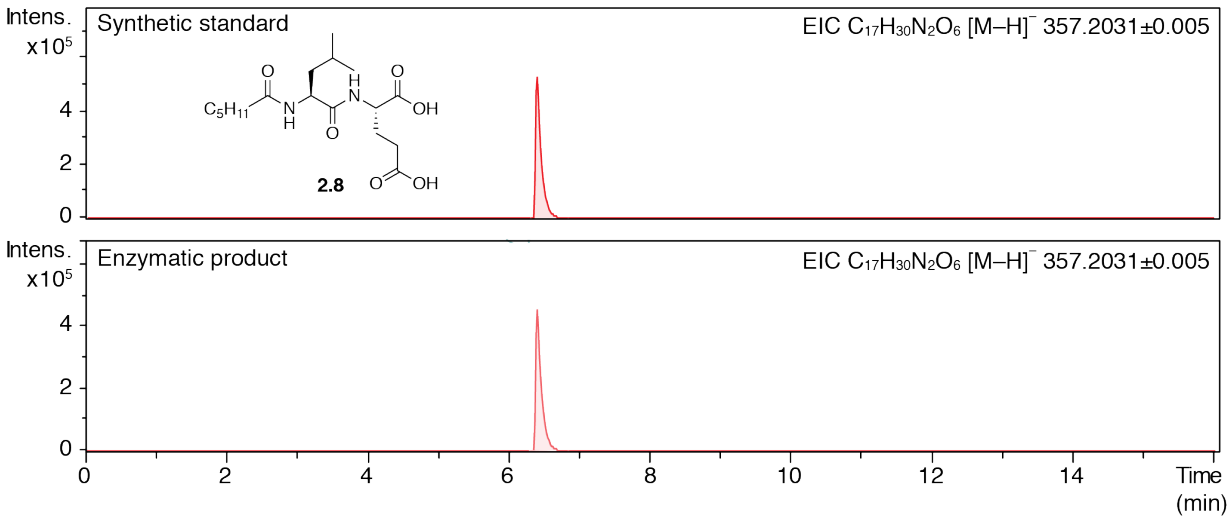


Figure 2.18: Extracted ion chromatograms for the synthetic standard 2.8 and enzymatic product from the reconstitution of RupA$_{C1\text{-}A1\text{-}T1}$ and RupA$_{C2\text{-}A2\text{-}T2\text{-}R}$.

Finally, we sought to completely reconstitute the RupA NRPS in vitro to access putative peptide aldehyde products. To accomplish this, we included either NADH or NADPH, the cofactors required for R domain activity, in reaction mixtures along with ATP and the preferred substrates hexanoyl-CoA, L-leucine, and L-glutamate. We analyzed the supernatants of reaction mixtures by LC-MS and attempted to detect the expected masses of peptide aldehydes, primary alcohols, truncated products, or molecules that could arise from degradation of the predicted structures. However, after extensive optimization, we were unable to detect any putative final products in this experiment. We were also unable to identify final products in the presence of any other combinations of building blocks that we had previously examined. We did observe formation of the hydrolysis products of tethered *N*-acylated dipeptide thioester intermediates, indicating that the NRPS modules were functional (data not shown). We also confirmed that synthetic standards of the predicted peptide aldehyde product **2.1** could be detected under these assay conditions (data not shown).

Suspecting that our $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ construct may have purified with an inactive R domain, we individually expressed and purified two additional constructs ($RupA_R$ single domain and $RupA_{T2\text{-}R}$ di-domain) and evaluated their reactivity toward a synthetic *N*-acetylcysteamine (SNAC) substrate **5** that mimics the preferred $RupA_{T2}$-tethered intermediate (Figure 2.19). This substrate was synthesized using standard peptide coupling chemistry: *N*-hexanoyl-L-leucine (**2.3**) was coupled to protected L-glutamic acid using DCC, and this *N*-acyl dipeptide (**2.5**) was then coupled to *N*-acetylcysteamine and deprotected to yield the desired compound (**2.7**). We used this substrate in a similar format to our previous reconstitution assay attempts to see if our purified proteins could act upon it. Monitoring consumption of NAD(P)H by the change in absorbance at 340 nm (A_{340}), we could detect activity of neither $RupA_R$, $RupA_{T2\text{-}R}$ nor the full module $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ toward the synthetic substrate (Figure 2.20). This suggests either that we have not successfully purified an active form of $RupA_R$ or that the enzyme natively performs some other unexpected activity.

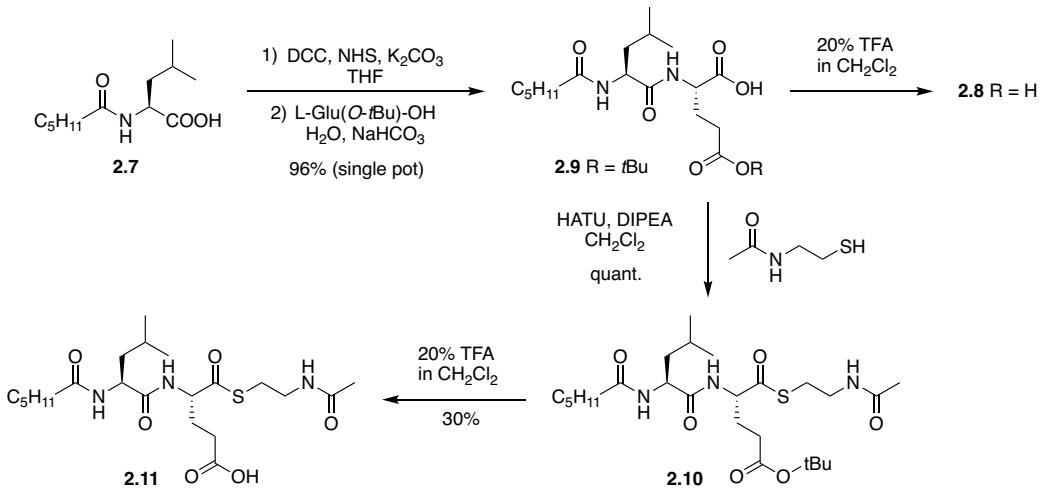


Figure 2.19: Synthesis of an *N*-acetylcysteamine (SNAC) substrate for $RupA_R$. (DCC = *N*,*N*'-dicyclohexylcarbodiimide, NHS = *N*-hydroxysuccinimide).

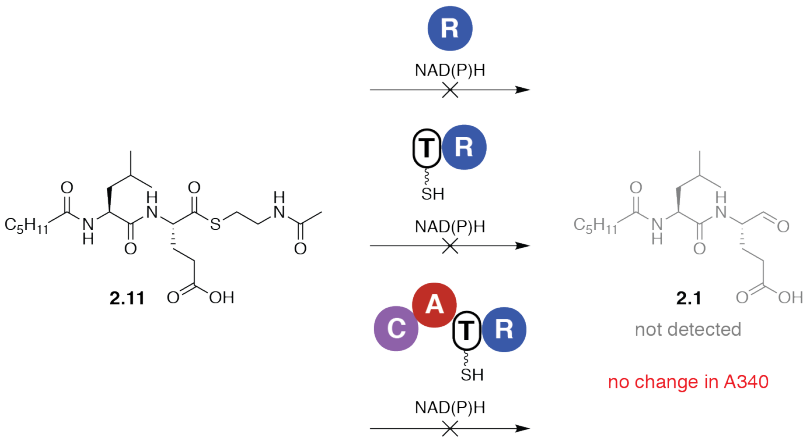


Figure 2.20: Assays with SNAC substrate.

The RupA reductase domain was inactive toward substrate mimic *N*-hexanoyl-L-Leu-L-Glu-SNAc **2.11** in vitro. By measuring decrease in absorbance at 340 nm, no consumption of NAD(P)H was observed when reacting **2.11** with purified $RupA_R$, $RupA_{T2\text{-}R}$, or $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$.

Though we were unable to reconstitute the activity of the RupA R domain in vitro, bioinformatic analyses suggest this domain should be active in vivo. Though $RupA_R$ shows only low amino acid sequence identity to other biochemically characterized R domains (e.g., 23.5% with $AusA_R$,[44] 18.7% with $MxcG_R$,[45] and 24.6% with the R domain from bgc35, which was previously reconstituted in vitro[14]), it does contain the conserved catalytic triad and NAD(P)H binding motifs (Figure 2.5). Among the diverse superfamily of short-chain dehydrogenases/reductases, which includes NRPS terminal R domains, proteins with sequence identities as low as 15-30% are reported to share similar three dimensional folds.[46] We generated a homology model of $RupA_R$ with the known two-electron reducing terminal R domain from AusA, using HHPred and MODELLER. This model suggests that the $RupA_R$ motif for NAD(P)H binding and catalytic triad for reduction chemistry are properly oriented (Figure 2.21). Therefore, we propose that this R domain is likely active in vivo and involved in producing the final product of the *rup* gene cluster.

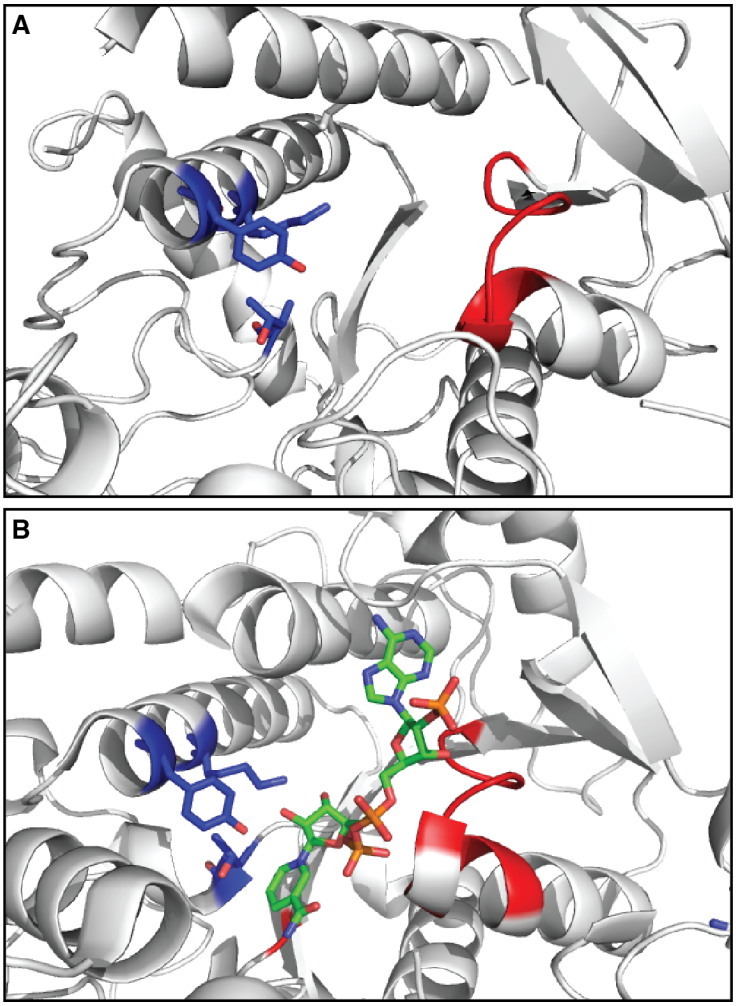


Figure 2.21: Homology model of RupA reductase domain.

The homology model (A) was generated with the AusA reductase domain (PDB: 4F6C) as a scaffold using HHPred and MODELLER (MPI Bioinformatics Toolkit, Max Planck Institute for Developmental Biology, Tübingen, Germany). It is shown in comparison to the MxaA reductase domain structure with NADPH bound (PDB: 4U7W) (B). Highlighted in red are the NAD(P)H binding motif and in blue the Thr/Tyr/Lys catalytic triad.[26]

2.2.5. Possible sources of acyl-CoAs in *R. bromii*

The results of biosynthetic reconstitution experiments led us to hypothesize that ruminopeptin selectively incorporates a medium chain acyl-CoA substrate, perhaps hexanoyl-CoA, on its N-terminus. Fatty acyl CoAs and fatty acids are an essential component of all known life, and fatty acid production is therefore a highly conserved biological function. However, the typical acyl-CoA pool of a bacterial cell is skewed towards short chain acyl-CoA's, which serve as a primer for fatty acid biosynthesis and also play other important metabolic roles. The majority of fatty acids in a microbial cell, in contrast, have 16-carbon to 18-carbon chains and are incorporated into biological membranes as phospholipids.[47] A medium chain acyl-CoA is therefore a somewhat unusual biosynthetic intermediate, and we were interested to interrogate the potential pathways that would lead to such a compound being available for ruminopeptin biosynthesis.

The basic steps of fatty acid biosynthesis in *E. coli* have been well elucidated (Figure 2.22A) (reviewed by Cronan and coworkers[48] and by Hirooka and coworkers[49]). Fatty acid biosynthetic enzymes are typically primed by acetyl-CoA or butyryl-CoA and use malonyl-CoA as an extender unit (using analogous logic to PKS enzymes). Biosynthesis begins with acyltransferase FabD loading malonyl-CoA onto acyl-carrier protein (ACP). The initial chain extension is performed by ketosynthase FabH, which catalyzes a decarboxylative chain extension on malonyl-ACP using acetyl-CoA as the electrophile. Ketoreductase FabG then reduces the β-ketone of this tethered intermediate with NADPH, dehydratase FabA/Z catalyzes elimination of hydroxyl to form the α,β-unsaturated thioester, and enoylreductase FabI uses NADH to reduce the double bond, forming a fatty acyl ACP. This cycle can then continue, with FabB/F extending this chain onto another acyl-ACP.[49] This pathway typically operates to produce fatty acids with

16-carbon to 18-carbon chains.[47] There are also some specific microbial fatty acid synthetases that produce medium chain fatty acids (e.g. hexanoate), but these are separate from primary metabolism and involved in the biosynthesis of particular metabolites.[47] From protein BLAST searches, we determined that *R. bromii* contains the necessary enzymes for fatty acid biosynthesis (Table 2.6). As noted, however, acyl CoA's are not intermediates in this pathway, and so this cannot directly explain the source of medium chain acyl-CoA's.

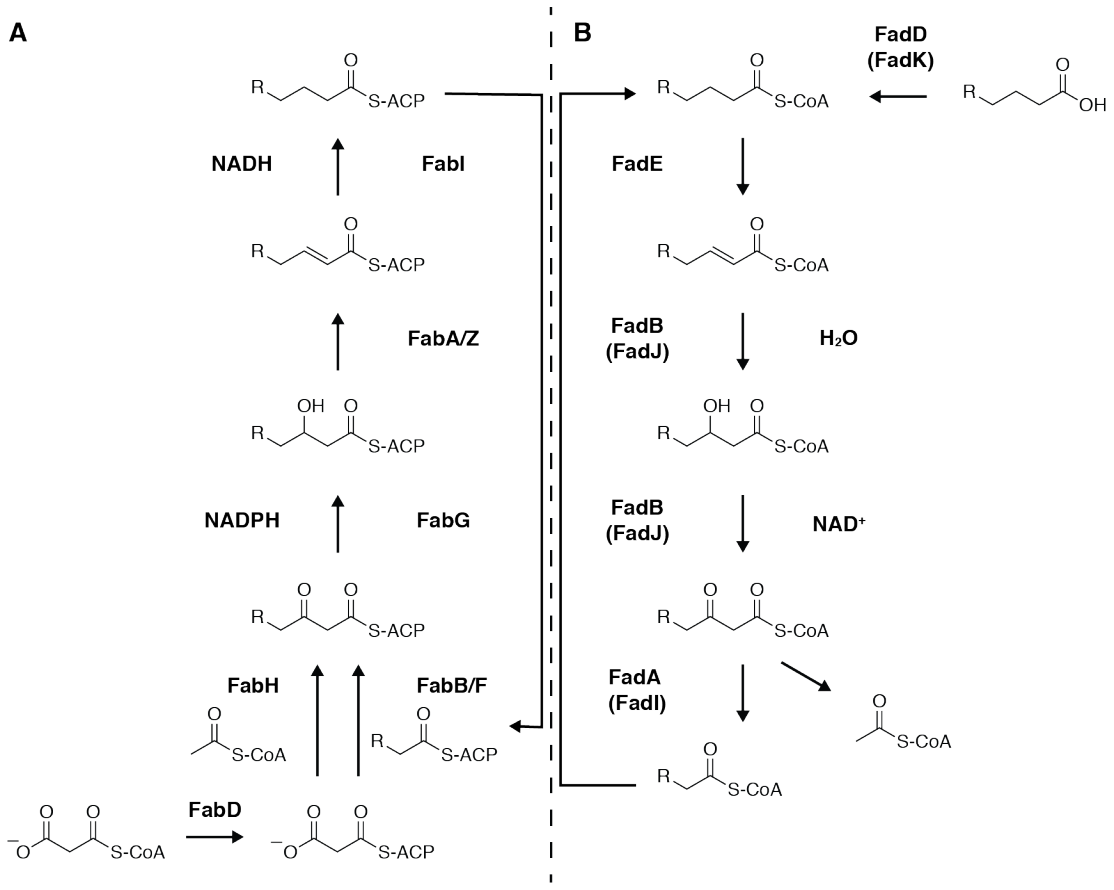


Figure 2.22: Fatty acid biosynthesis and degradation pathways

The canonical fatty acid biosynthesis (A) and degradation (B) pathways involve similar metabolic intermediates but proceed in opposite directions.[48,49] The enzyme names from the canonical pathways in *Escherichia coli* are indicated here. (A) In fatty acid biosynthesis, FabA/Z and FabB/F are each pairs of homologous enzymes that have different chain length preferences. (B) In fatty acid degradation (β-oxidation), there are distinct enzymes involved at some steps in the aerobic vs. anaerobic pathways, and the enzymes involved in the anaerobic pathway are shown in parentheses.

Table 2.6: Fatty acid biosynthesis enzymes in *R. bromii* L2-63

Reference protein	Locus tag	*R. bromii* protein
FabD	RBL263_RS00770	WP_015522652.1
FabH	RBL263_RS06105	WP_015523538.1
FabB	RBL263_RS00780	WP_015522654.1
FabG	RBL263_RS00775	WP_015522653.1
FabA	RBL263_RS06095	WP_015523536.1
FabI	RBL263_RS00765	WP_015522651.1

In contrast to the fatty acid biosynthesis pathway, the enzymes in fatty acid degradation pathway act directly on acyl-CoA substrates, providing a plausible hypothesis for the generation of a pool of medium chain fatty acyl CoA's. This pathway performs essentially the reverse chemistry to the biosynthetic pathway (Figure 2.22B).[49] The β-oxidation pathway exists in two versions (aerobic and anaerobic) in some species, such as *E. coli*, but the pathways contain homologous enzymes (anaerobic pathway enzymes are indicated in parentheses).[49,50] Catabolism begins with the activation of a fatty acid as a fatty acyl-CoA by fatty acid-CoA ligase FadD (FadK). Dehydrogenase FadE then catalyzes the oxidation of this substrate using FAD^{2+} as a cofactor. The α,β unsaturated acyl-CoA is hydrated and oxidized by hydratase/dehydrogenase FadB (FadJ), using NAD^{+} as a cofactor, and thiolase FadA (FadI) catalyzes the release of one unit of acetyl-CoA, also forming another acyl-CoA that has been shortened by two carbons.[49] The genome of *R. bromii* does not contain homologs of most of the enzymes in this pathway. However, it does contain two homologs of FadK (Table 2.7), suggesting a potential path for biosynthesis of acyl-CoA's by the organism.

Table 2.7: Fatty acid degradation enzymes in *R. bromii* L2-63

Reference protein	Locus tag	*R. bromii* protein
FadK	RBL263_RS05665	WP_015523460.1
FadK	RBL263_00633	SPE91331.1

Though a bioinformatics tool was previously developed for the classification of acyl-CoA ligase substrate specificities, it was unable to predict specific substrates for these *R. bromii* proteins (the tool referenced here is no longer maintained and available online).[51] Therefore, we turned to BLAST searches and homology modelling to explore the specificity of these enzymes. One (SPE91331.1, locus tag RBL263_00633) is highly homologous (45% amino acid sequence ID) to a characterized and crystallized medium-chain acyl-adenylate synthetase from *Methanosarcina acetivorans*, an archaeal species.[52,53] Uniquely, this FadK homolog preferentially accepts medium-length branched-chain fatty acids, with a preference for 2-methylbutyrate.[53] We constructed a homology model of this *R. bromii* acyl adenylate synthetase, using the *M. acetivorans* protein as a template, and found that it shows excellent conservation of the active site binding pocket (Figure 2.23).

Due to their limited commercial availability, we did not assay branched chain acyl-CoA's in any of our biosynthetic reconstitution assays for RupA. However, methods also exist for isolating acyl-CoAs from tissue and culture samples,[54] and we adapted these methods to analyze the fatty acyl-CoA pool in *R. bromii*. We grew cultures of *R. bromii* and attempted to isolate the amphipathic fatty acyl CoA's. We were unable to observe isolated fatty acyl-CoAs in this manner by HPLC and LC-MS, perhaps due to concentration or stability issues. We also assessed the ability of this concentrated fatty acyl-CoA extract to serve as a substrate in our biosynthetic reconstitution assay for C-starter domain specificity. From this experiment, we were able to observe the formation of acetyl-Leu, indicating that we likely successfully isolated acetyl-CoA

from a culture of *R. bromii* (data not shown). However, as our previous experiments with commercially available substrates showed (Figure 2.15), this is not a preferred substrate of this domain. Therefore, though it is possible that *R. bromii* possesses the capacity to produce hexanoyl-CoA or similar substrates from free fatty acids, we have not confirmed this hypothesis.

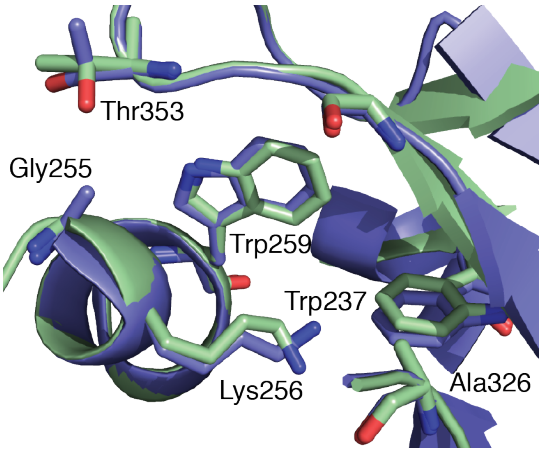


Figure 2.23: An *R. bromii* protein shows predicted strutural homlogy with a unique acyl-CoA ligase from *M. acetivorans*.

The homology model (blue) was generated with the FadK homolog from *M. acetivorans* (PDB:3ETC) as a scaffold using HHPred and MODELLER. The homology model is shown overlayed with the scaffold structure (green). Labeled residues in the acyl binding pocket of 3ETC were identified by Gulick and coworkers.[52]

2.2.6. The *asf* gene cluster from *Clostridium* sp. ASF502 may produce a similar product to ruminopeptin.

An additional reason that we selected the rup gene cluster for study was our prediction that it would produce a peptide aldehyde with an acidic residue in the P1 position. Though synthetic glutamyl and aspartyl aldehydes are known, to our knowledge there are no isolated natural product peptide aldehydes that contain an acidic residue in this position. In the family of NRPS clusters from the human gut microbiota identified by Fischbach and coworkers in 2014, there are

two additional clusters that are also predicted to load an acidic residue at P1. One of these is from the metagenomic species *Ruminococcus* sp. CAG:108 (bgc44) and likely produces the same product. (The main NRPS protein from this cluster has 98% amino acid sequence ID to RupA. No 16S rRNA sequence is available for this organism.) The other one is encoded by a member of the Altered Schaedler Flora (ASF), *Clostridium* sp. ASF502 (bcg42), which is part of Clostridium cluster XIV.[55] The ASF is an 8-species consortium of mouse gut microbial species that was originally developed to standardize mice for biological experiments.[56] As an extremely simplified community that mimics the behavior of the healthy mouse gut microbiota, it has recently attracted interest in its own right. The genome sequences for its 8 strains have recently been determined,[56] and it has been proposed as a useful model microbiota for determining important gene functions and metabolic products.[57,58]

The NRPS gene cluster from ASF502 (the *asf* gene cluster) includes a hypothetical protein, a predicted CoA ligase, and a large multi-module NRPS (AsfA) (Figure 2.24). Due to bioinformatic uncertainty about the initiation domain(s) in this NRPS, there are two major hypotheses to advance about the product of the *asf* gene cluster. One hypothesis is that AsfA may produce a tripeptide aldehyde that is not acylated. AsfA appears to contain a canonical starter A–T module, two full extension modules, and a terminal reductase domain, which would lead it to produce a tripeptide aldehyde. However, upon closer inspection, the first A domain of the cluster contains an unusual valine residue at position 235, rather than the standard (and highly conserved) aspartate residue (Figure 2.25). Based on homology, this domain is predicted bind AMP, and most bioinformatic tools (as exhibited by the consensus of NRPSPredictor2[59], Stachelhaus code[27], and Minowa[60] on antiSMASH[61]) do predict that this is an amino acid

adenylation domain that activates and loads L-leucine. Therefore, the first hypothesis is that AsfA may produce a tripeptide aldehyde with a free amino terminus.

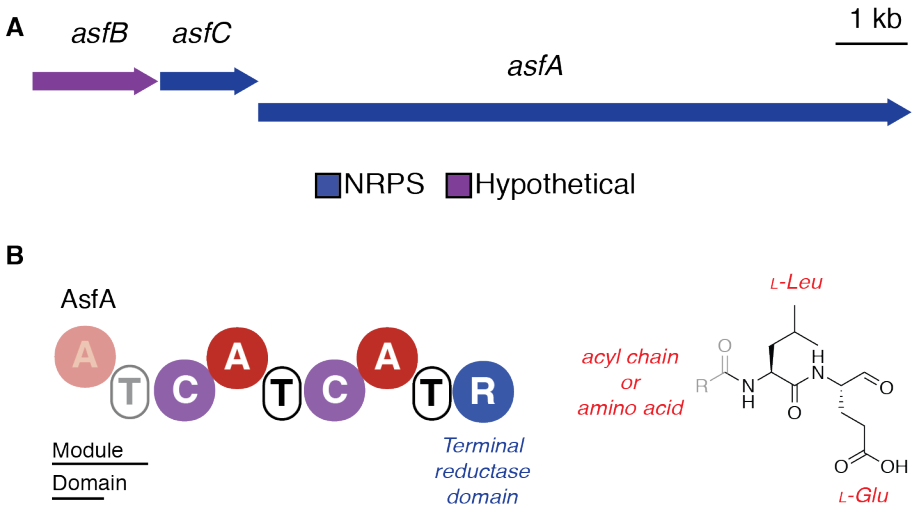


Figure 2.24: The *asf* gene cluster from *Clostridium* sp. ASF502 is predicted to produce a peptide aldehyde product.

(A) The *asf* gene cluster from *Clostridium* sp. ASF502. The gene cluster encodes a single multi-module NRPS, a putative CoA ligase, and a hypothetical protein. (B) The AsfA NRPS contains three modules and a terminal R domain. Though the A domain in the initial module appears to conserve important catalytic residues for amino acid adenylation, it also has an unusual A domain signature. Additionally, the C_1 domain contains some residues that are indicative of C-starter domain functionality.

Position	235	236	239	278	299	301	322	330	Substrate/prediction
AsfA A_1	**V**	**A**	**I**	**F**	**V**	**G**	**T**	**A**	**L-Leu?**
AsfA A_2	**D**	**A**	**M**	**F**	**L**	**G**	**C**	**I**	**L-Leu (predicted)**
RupA A_1	D	A	S	F	L	G	G	V	L-Leu (predicted)
SrfAA A_3	D	A	W	F	L	G	N	V	L-Leu
AsfA A_3	**D**	**M**	**K**	**N**	**M**	**G**	**T**	**V**	**L-Glu (predicted)**
RupA A_2	D	M	K	N	L	G	T	V	L-Glu (predicted)
SrfAA A_1	D	A	K	D	L	G	G	V	L-Glu

Figure 2.25: **Predicted A domain specificity-conferring residues (Stachelhaus codes) for AsfA.**

Specificity-conferring residues were identified using the University of Maryland's PKS/NRPS Analysis Web-site.[24] Reference codes are from the initial identification of key A domain residues by Marahiel and coworkers.[27] Numbering of positions references the sequence of phenylalanine-activating A domain GrsA. The residue at position 278 in AsfA A_2 was identified by manual inspection of a ClustalW2 alignment between the AsfA A domains and GrsA.

A second hypothesis comes from analysis of the other enzymes in the cluster and the first C domain of AsfA. The *asf* gene cluster contains an enzyme annotated as a CoA ligase (AsfC), which has homology to enzymes that form acyl-CoA's, and the first C domain of AsfA contains some of the residues that would lead us to predict that it functions as a C-starter domain (Figure 2.4). Based on this bioinformatic analysis, a second scenarios is possible: the first module of AsfA is inactive and the NRPS produces an *N*-acylated dipeptide aldehyde by action of this C-starter domain (perhaps with the CoA ligase in the cluster furnishing the acyl-CoA that it loads). These two possibilities could be distinguished by conducting in vitro biochemical assays on excised modules from this enzyme. Expression and purification of $AsfA_{A1\text{-}T1}$ could enable assays that would determine if these domains can activate and load amino acids, while expression and

purification of $AsfA_{C1–A2–T2}$ could enable assays that would determine if these domains can produce tethered *N*-acyl amino acids.

The subsequent A domains of AsfA are more canonical, and their specificity codes are nearly identical to those from the *rup* gene cluster (Figure 2.25). As we have biochemically verified the loading preferences of these motifs for RupA, we can say with some confidence that the peptide aldehyde produced by AsfA may contain L-glutamate in the P1 position and L-leucine in the P2 position. Due to these conserved residues, whether this product is actually a dipeptide aldehyde or a tripeptide aldehyde, it is likely to target proteases of a similar specificity as those targeted by ruminopeptin. Therefore, we have identified a predicted peptide aldehyde scaffold arising both from an abundant member of the healthy human gut microbiota and a member of the healthy mouse ASF. Genetic tools for Firmicutes genera such as *Clostridium* and *Ruminococcus* are not well developed, but there have been some recent encouraging results in the genetic manipulation of *Clostridium* species.[62] If a genetic knockout or disruption of the *asf* gene cluster in *Clostridium* sp. ASF502 could be obtained, it would open many additional avenues of inquiry for interrogating the ecological relevance for this organism of producing a ruminopeptin-type peptide aldehyde.

2.3. Conclusions

In summary, our efforts to reconstitute the activity of the *R. bromii* NRPS RupA strongly suggest that the peptide aldehyde *N*-hexanoyl-L-Leu-L-Glu-H (**2.1**) is a likely product of this enzymatic assembly line. This work does not rule out the possibility that RupA may produce additional metabolites in vivo. We observed some promiscuity in the *N*-acylation activity of the C-starter domain and A domain specificities of RupA. However, this type of promiscuity has

previously been observed for NRPS enzymes reconstituted in vitro.[21,23,63,64] In the case of AusA, which produces aureusimines A–C, the preferred products observed in this format correlate with the most abundant natural analogues.[44,65] It is also possible that there are unusual biosynthetic substrates available to *R. bromii*, particularly acyl-CoA's with unusual acyl-chain modifications, which we did not provide in our in vitro reconstitution experiments. Though RupA may produce additional molecules with different scaffolds, it is reasonable to propose that *N*-hexanoyl-L-Leu-L-Glu-H could be one major biosynthetic product.

Additionally, as we have never directly observed activity of the RupA R domain, we could not determine if this assembly line produces an aldehyde or an alcohol. In previous biosynthetic reconstitutions of NRPS terminal reductase domains, the aldehyde intermediate has not been detected in significant quantities when the final expected product is the peptide alcohol.[66,67] Though the activity of aldehyde-producing NRPS terminal reductase domains has been reconstituted in several cases,[44,68] the natural products generated by these pathways are cyclic imines or pyrazinones, so only trace amounts of free aldehyde intermediates were detected in these experiments. Given the difficulties encountered in resolving this biosynthetic step, we decided to move forward (in Chapter 3) to examine the biological activity of the putative peptide aldehydes we predict could be produced by RupA.

2.4. Materials and methods

2.4.1. General materials and methods

Oligonucleotide primers were synthesized by Integrated DNA Technologies (Coralville, IA) and Sigma Aldrich (Billerica, MA). Recombinant plasmid DNA was purified with the Qiaprep Kit from Qiagen (Germantown, MD) and the E.Z.N.A. Plasmid Mini Kit from OMEGA kit from

Omega Bio-Tek (Norcross, GA). 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), Genewiz (Cambridge, MA), and Eton Bioscience (Boston, MA). Restriction enzymes were purchased from New England BioLabs (Ipswich, MA). Nickel-nitrilotriacetic acid-agarose (Ni-NTA) resin was purchased from Qiagen. SDS-PAGE gels were purchased from BioRad. Protein concentrations were determined by quantifying protein A280 using a NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific) or by the Bradford assay. 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.

All chemicals were obtained from Sigma-Aldrich except where noted. Protected amino acids were obtained from Chem-Impex (Dale, IL) and Advanced ChemTech (Louisville, KY). HATU was purchased from Oakwood Chemical (Estill, SC). All NMR solvents were purchased from Cambridge Isotope Laboratories (Andover, MA). NMR spectra were visualized using iNMR version 5.5.7. and MestReNova version 12.0. 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, br s = broad singlet, d = doublet, t = triplet, q = quartet, m = multiplet), coupling constant, and integration.

High-resolution mass spectral (HRMS) data was obtained in the Small Molecule Mass Spectrometry Facility, FAS Division of Science. Enzyme assays were analyzed on a Bruker Impact II qTOF mass spectrometer in negative ion mode coupled to an Agilent 1290 uHPLC. Each LC-MS run was internally calibrated using sodium formate introduced at the end of the run.

For liquid chromatography, 5 μL of sample was injected onto a Phenomenex Kinetex C18 column (100 Å pore size, 150 mm x 2.1 mm, 2.6 μm particle size). Mobile phase A was 0.1% formic acid (v/v) in water, and mobile phase B was 0.1% formic acid (v/v) in acetonitrile. The mobile phase composition started at 1% B, which was maintained for 2 min. Samples were then subjected to a linear gradient over 8 min to 100% B. Flow of 100% B was maintained for 4 min, and the column was then re-equilibrated to 1% B over 1.9 min. HRMS data for synthetic compounds was obtained on an Agilent Technologies 6210 TOF coupled to an Agilent Technologies 1200 series LC. Liquid chromatography was performed with water/acetonitrile (1:1). The capillary voltage was 3.5 kV, the fragmentor voltage was 175 V, the drying gas temperature was 325 °C, the drying gas flow rate was 8 L/min, and the nebulizer pressure was 40 psig.

2.4.2. Cultivation of bacterial strains

R. bromii strains were cultivated using several different growth media: M2GSC (which is supplemented with 30% rumen fluid)[69] and RUM medium,[12] which were prepared as previously described with the following modifications: supplementary heat-sensitive vitamins were prepared as a 1000x aqueous stock (except for D-pantetheine, which was prepared as a 100x aqueous stock) and separately filtered and sparged with nitrogen to render anaerobic. Carbohydrates were also prepared as 100x aqueous stocks and treated with the same procedure. The media itself was boiled, sparged with nitrogen, dispensed in Hungate tubes under anaerobic conditions, and then autoclaved. Supplementary vitamins and carbohydrates were then added to individual aliquots of the growth media at the time of inoculation.

A lyophilized stock of *R. bromii* ATCC 27255 was purchased from the American Type Culture Collection, Manassas, VA. *R. bromii* L2-63 was provided as a glycerol stock by Harry Flint and coworkers (University of Aberdeen). *R. bromii* 22-5-S 6 FAA NB was provided as a glycerol stock by Emma Allen-Vercoe and coworkers (University of Guelph).

R. bromii L2-63, *R. bromii* ATCC 27255, and *R. bromii* 22-5-S 6 FAA NB were inoculated from frozen glycerol stocks as 5 mL cultures in RUM medium with fructose and allowed to grow in a 10% hydrogen/10% carbon dioxide/bal. nitrogen atmosphere for approximately 24 h until they reached saturation. These cultures were then passaged as 1:100 dilutions and allowed to reach saturation again before extraction of genomic DNA. Genomic DNA was extracted using the standard protocol of the UltraClean Microbial DNA Isolation Kit form MO BIO (Carlsbad, CA). To confirm strain identities, primers fD1 and rP2[30] were used to amplify and sequence 16S rRNA sequences.

2.4.3. PCR amplification and sequencing of *rup* cluster from *R. bromii* ATCC 27255 and *R. bromii* 5_1_S 6 FAA NB

PCR reactions for amplification of the *rup* gene cluster from each of the *R. bromii* strains were accomplished using Phusion PCR mix (ThermoFisher). Reactions were performed according to the manufacturer's instructions and contained 0.1 µL template DNA, 10 µM each of forward and reverse primers, half final volume of the 2x master mix, and water to total 25 µL. The PCR protocol began with heating at 98 ºC for 1 min. For 30 cycles, the following protocol was repeated: melting at 98 ºC for 10 s, annealing at 65 ºC for 30 s, extension at 72 ºC for 90 s. The reaction terminated with a final extension stage at 72 ºC for 10 min. The reactions were analyzed by agarose gel electrophoresis with SYBR Safe staining. Initially, the gene cluster was

detected by using primers repDetect1 and rupDetect2, designed to amplify the RupA A_1 domain from strain L2-63. Subsequently, the remainder of the gene cluster was sequenced by PCR-amplifying overlapping regions of the cluster, using primers designed for strain L2-63, and then assembling the resulting reads using the Geneious 9 assembler. The primers used for sequencing are indicated in Table 2.8.

Table 2.8: Oligonucleotides used for cloning and sequencing.

Where applicable, restriction sites are underlined.

Primer name	Sequence (5' to 3')	Cut by:
Universal T7	TAATACGACTCACTATAGGG	
T7 reverse	GCTAGTTATTGCTCAGCGG	
rupDetect-1	ATTAGCTAGCGTAAACGAAAATCAGCTTGCAG	
rupDetect-2 (rupSeq10)	GATCTCGAGTTAAAGGAATGTTAGTTCCCTCGG	
rupSeq1	GAAGAGGTTGTTTCAATTCTGATTCCGA	
rupSeq2	GAATATTGTAAAGTGTGCTGTACGGATTTT	
rupSeq3	GTTCTTCGACAATATGATGAAAAG	
rupSeq4	CTGTCGGCTGTATAGTAAACGCAGATTGCG	
rupSeq5	GATTAAGAATTGAACTCGGTGAAATTGAAA	
rupSeq6	ATTAGCTAGCGACGAACTTCAGGAGTTTATG	
rupSeq7	TGAATACGCAGTTGCAAAGTTTACAAA	
rupSeq8	GCTAGCTAGCAAAATCAGAAAGGTTTCACTCC	
rupSeq9	GCGCTCGAGTTACACAAATTTTGCAAGAGTC	
Ntermseq-f	TTCGGATAAGGTTGAATATTCAAG	
Ntermseq-r	GAAATGGCACTTTTTAACGGAACA	
Ctermseq-f	TCGGCAATGTTCTTCTGACAGGCT	
Ctermseq-r	CGGTGCTTTTTATGCTTGCCGTTA	
CAT1-f	GCCGAGATCTAATGAGTAATTTAATCAACTTGC	BglII
CAT1-r	ATATCTCGAGGCTGAACATAAACTCCTGAAGTTC	XhoI
CAT2R-f	ATATAGATCTAATGTTCAGCGGTACGGC	BglII
CAT2R-r (T-reductase-r)	ATATCTCGAGATCGAAGAAACCAAGACC	XhoI
reductase-f	ATTAGCTAGCGTGAGCAAGGATGATTCAG	NheI
reductase-r	ATTATCTCGAGTTAATCGAAGAAACCAAGACCT	XhoI
T-reductase-f	ATATAGATCTACTTGACGAAATGCCTCTCACAC	BglII
T1-f	ATATCTCGAGTTAGCTGAACATAAACTCCTGAAGTTC	XhoI
T1-r	ATTAGCTAGCGACAAAATTCCGCTCAATGTAAAC	NheI

2.4.4. RT-PCR for detection of *rup* gene cluster transcription

For detection of *rup* gene cluster expression in *R. bromii* strains, the organism was grown as described above and subjected to various culture conditions (Figure 2.7). Two mL of each culture was then mixed with an equal amount of bacterial RNAProtect solution (Qiagen) and processed according to the manufacturer's instructions. Protected cell pellets prepared in this way were stored at –80 ºC for up to 48 h before downstream processing. Total RNA was obtained with the TRIzol reagent using previously published procedures.[70] After extraction, RNA was air dried overnight and then digested with RNAse-free DNAse (Invitrogen) according to the manufacturer's procedure. RT-PCR was conducted using the SuperScript III one-step RT-PCR system with Platinum *Taq* DNA polymerase (ThermoFisher) according to the manufacturer's instructions. Reaction mixtures contained 0.3 µL template RNA, 10 µM each of forward and reverse primers, half final volume of the 2x master mix, 1 µL of SuperScript III RT/Platinum *Taq* enzyme mix, and water to total 25 µL. The PCR protocol began with heating at 55 ºC for 30 min (cDNA synthesis). The reaction was then heated at 94 ºC for 2 min. For 40 cycles, the following protocol was repeated: denaturing at 94 ºC for 15 s, annealing at 63 ºC for 30 s, extension at 68 ºC for 90 s. The reaction terminated with a final extension stage at 68 ºC for 5 min. Cluster expression was detected using primers rupDetect-1 and rupDetect-2. The reactions were analyzed by agarose gel electrophoresis with SYBR Safe staining. Bands were identified by imaging on a Gel Doc™ EZ Gel Documentation System (BioRad).

2.4.5. Attempts to isolate aldehydes from cultures of *R. bromii*.

Various attempts were made to isolate ruminopeptin-type compounds from cultures of *R. bromii* under a variety of conditions, with attempted identification of products by LC-MS. These

included extractions from culture volumes of 1 mL, 5 mL, 20 mL, 50 mL, 1L using ethyl acetate or methanol. Direct injection of culture supernatant was also attempted. These samples were analyzed on an Advion Expression CMS-L mass spectrometer in negative ion mode coupled to an Agilent 1200 Series HPLC. For liquid chromatography, 20 µL of sample was injected onto a Phenomenex Gemini C18 column (110 Å pore size, 50 mm x 4.6 mm, 5 µm particle size). Mobile phase A was 0.1% formic acid (v/v) in water, and mobile phase B was 0.1% formic acid (v/v) in acetonitrile. The flow rate was 0.3 mL/min. The mobile phase composition started at 5% B, which was maintained for 2.1 min. Samples were then subjected to a linear gradient over 3.9 min to 95% B. Flow of 95% B was maintained for 4 min, and the column was then re-equilibrated to 5% B over 2 min. The capillary temperature was 250ºC, the capillary voltage was 180 V, the source voltage offset was 30 V, the source voltage span was 30 V, the source gas temperature was 20ºC, and the ESI voltage was 2.5 kV. Under these conditions, which could be used to visualize ruminopeptin standards **2.1** and **2.2**, expected masses of the compounds could not be observed in extracts.

The stability of ruminopeptin-type compounds in *R. bromii* cultures was also interrogated. In one experiment, 5 mL saturated cultures of *R. bromii* were dosed with either 10 µM or 100 µM of **2.1**. After incubation for 5 hours, supernatants were extracted with ethyl acetate, resuspended in methanol (500 µL), and analyzed by LC-MS. Under these conditions, the compound was not observed. After an additional 10-fold concentration of this extract, the compound mass was observed, but at orders of magnitude lower abundance than would have been expected. In an additional experiment, six 5 mL cultures of *R. bromii* were treated with 100 µM **2.1** and extracted at various time points (0 – 24 h). After incubation, the cultures were flash frozen with liquid nitrogen, lyophilized, and the cell debris resuspended in 5 mL ethyl acetate and filtered.

The samples were concentrated in vacuo, resuspended in 500 µL MeOH, and analyzed by LC-MS using the above conditions. In this experiment, there was a clear time-dependent degradation of the compound over 0–24 h. The compound also degrades significantly in the uninoculated growth medium over this time, suggesting that the presence of the organism is not required for this degradation to occur.

2.4.6. Using 2-aminobenzamide oxime (ABAO) to derivatize aldehydes

For the initial reaction with a model synthetic aldehyde, ABAO (2 mg) was dissolved in 300 µL sodium acetate buffer (100 mM, pH 4.5). **2.2** (5 mg) was dissolved in acetonitrile (200 µL) and diluted with 400 µL sodium acetate buffer. The aldehyde solution was added dropwise to the ABAO solution and then stirred at room temperature overnight. The reaction mixture was diluted 1:100 and analyzed by HPLC (Figure 2.10).

For the model reaction in *R. bromii* media, a 44 mM stock of ABAO was prepared in RUM medium (0.5 mg in 75 µL). A separate 10.8 mM stock of peptide aldehyde **2.2** was prepared in 1:1 acetonitrile/water (0.16 mg in in 40 µL). The solutions were combined with an additional 50 µL medium for a final reaction volume of 165 µL. The reaction was incubated overnight with gentle shaking at rt. For analysis, the reaction mixture was diluted with 50 µL acetonitrile, incubated on ice for 10 min to precipitate solids, and centrifuged (13,000 rpm x 10 min). The reactions were then analyzed by HPLC (data not shown).

For the model reaction in a growing culture, a saturated culture of *R. bromii* was inoculated 1:100 in 5 mL RUM medium with added fructose and vitamins. **2.2** was added from a 100 mM DMSO stock to final concentration of 130 µM (6.5 µL). A 250 mM stock of ABAO was prepared in sodium acetate buffer (pH 4.5), and this stock was added to the cultures to result in a

final concentration of either 10 mM ABAO or 1 mM ABAO. A separate negative control reaction did not contain ABAO. The cultures were incubated at 37 ºC for 24 h, by which point all cultures had reached saturation. For analysis, a 165 µL of the culture was diluted with 50 µL acetonitrile, incubated on ice for 10 min to precipitate solids, and centrifuged (13,000 rpm x 10 min). The reactions were then analyzed by HPLC (Figure 2.10).

For liquid chromatography of the above reactions on a Dionex UltiMate 3000 HPLC, 20 µL of sample was injected onto a Phenomenex Gemini C18 column (120 Å pore size, 150 mm x 2.1 mm, 3 µm particle size). Mobile phase A was 0.1% formic acid (v/v) in water, and mobile phase B was 0.1% formic acid (v/v) in acetonitrile. The flow rate was 0.3 mL/min. The mobile phase composition started at 5% B, which was maintained for 10 min. Samples were then subjected to a linear gradient over 20 min to 95% B. A linear gradient was then employed to reach 5% B over 5 min and the column was re-equilibrated at 5% B for 10 min. UV absorbance was monitored at 260 nm and 370 nm.

The reaction in buffer was also analyzed on an Agilent Technologies 6210 TOF coupled to an Agilent Technologies 1200 series LC (Small Molecule Mass Spectrometry Facility, FAS Division of Science). Mobile phase A was 0.1% formic acid (v/v) in water, and mobile phase B was 0.1% formic acid (v/v) in acetonitrile. The flow rate was 0.4 mL/min. The mobile phase composition started at 5% B, which was maintained for 2 min. Samples were then subjected to a linear gradient over 8 min to 100% B. The flow was maintained at 100% B for 5 min, and flow was then returned to 5% B over 0.1 min. The capillary voltage was 3.5 kV, the fragmentor voltage was 100 V, the drying gas temperature was 325 °C, the drying gas flow rate was 10 L/min, and the nebulizer pressure was 10 psig. HRMS (ESI) for compound **2.6**: Calc'd for formula $C_{26}H_{40}N_5O_5^-$ $[M–H]^-$ 502.3035, found 502.3019.

2.4.7. Cloning, overexpression and purification of $RupA_{C1\text{-}A1\text{-}T1}$, $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$, $RupA_{T1}$, $RupA_{R}$, and $RupA_{T2\text{-}R}$

Protein expression constructs were PCR amplified from *R. bromii* L2-63 genomic DNA using the primers shown in Table 2.8. PCR amplification was performed using Phusion PCR mix (ThermoFisher). Reactions were performed according to the manufacturer's instructions and contained 0.1 μL template DNA, 10 μM each of forward and reverse primers, half final volume of the 2x master mix, and water to total 25 μL or 50 μL. Reaction mixtures were divided in 12.5 μL portions in order to assess annealing temperatures from 50–70 ºC. The PCR protocol began with heating at 98 ºC for 1 min. For 30 cycles, the following protocol was repeated: melting at 98 ºC for 10 s, annealing along a gradient of 50–70 ºC for 30 s, extension at 72 ºC for 30 s per 500 base pairs of the desired product length. The reaction terminated with a final extension stage at 72 ºC for 10 min. The reactions were analyzed by agarose gel electrophoresis with SYBR Safe staining. In each experiment, all reaction mixtures showing a band of the desired length by diagnostic PCR were pooled and purified.

Restriction digests were conducted according to the manufacturer's instructions, with the enzymes indicated in Table 2.8, and were purified directly using agarose gel electrophoresis with SYBR Safe staining. Gel fragments were further purified using the Illustra GFX PCR DNA and Gel Band Purification Kit. The digests were ligated into linearized expression vectors using T4 DNA ligase (New England Biolabs). $RupA_{C1\text{-}A1\text{-}T1}$, $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$, and $RupA_{T2\text{-}R}$ were ligated into the pET-29b vector to encode a C-terminal His_6-tagged construct. $RupA_{T1}$ and $RupA_{R}$ were ligated into the pET29a vector to encode a N-terminal His_6-tagged construct. Ligations were incubated at room temperature for 3 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). 10 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.

For protein expression, the vectors containing $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$, $RupA_{T1}$, $RupA_{R}$, $RupA_{T2\text{-}R}$ were transformed into chemically competent *E. coli* BL21 (DE3) cells. The vector containing $RupA_{C1\text{-}A1\text{-}T1}$ was co-transformed into *E. coli* BL21 GOLD (Agilent Technologies) with the addition of chaperone plasmid pGro7 (Takara Bio USA, Mountain View, CA). Cell stocks were stored at –80 °C in LB/glycerol.

The general procedure for protein large scale overexpression and purification was as follows. A 50 mL starter culture of BL21 or BL21+pGro7 *E. coli* was inoculated from a single colony and grown overnight at 37 ºC in LB medium supplemented with 50 µg/ml kanamycin (and 20 µg/mL chloramphenicol for BL21 + pGro7). Overnight cultures were diluted 1:100 into 2 L of LB medium containing 50 µg/mL kanamycin (and 20 µg/mL chloramphenicol for BL21+pGro7). Cultures were incubated at 37 ºC with shaking at 175 rpm, moved to 15 ºC at OD600 = 0.2-0.3, induced with 500 µM IPTG at OD600 = 0.5-0.6, and incubated at 15 ºC for 19 h. Cells from 2 L of culture were harvested by centrifugation (4,000 rpm x 10 min) and resuspended in 35 mL of lysis buffer (20 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, pH 7.5, supplemented with 1 mM DTT for purification of $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$). 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 (10,800 rpm x 30 min). The supernatant was supplemented with 1 M imidazole for a final concentration of 5 mM imidazole, treated with 20 µL DNAse I, and passed over 4 mL of Ni-NTA resin (pre-washed with 3 x 10 mL lysis buffer). The resin-bound protein was washed with 25 mL of 25 mM imidazole elution buffer. Protein was eluted from the column using a stepwise

imidazole gradient in elution buffer (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 used to determine which fractions contained the desired protein. Fractions were combined and dialyzed twice against 2 L of storage buffer (20 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 10% (v/v) glycerol, pH 7.5, supplemented with 1 mM DTT for purification of $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$). Solutions containing protein were frozen in liquid nitrogen and stored at –80 ºC. This procedure afforded yields of 8.6 mg/L for $RupA_{C1\text{-}A1\text{-}T1}$, 1.7 mg/L for $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$, 3.5 mg/L for $RupA_{T1}$, 2.4 mg/L $RupA_R$, and 7.5 mg/L $RupA_{T2\text{-}R}$.

2.4.8. ATP-$^{32}PP_i$ exchange assay for RupA

The reaction mixture (100 µL) contained 75 mM Tris-HCl pH 8.5, 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 (stock 1:1500 dilution prepared from Phosporous-32 radionuclide, PerkinElmer, ~6 mCi/mL, in 40 mM Na_4PP_i). Reaction mixtures were initiated by the addition of $RupA_{C1\text{-}A1\text{-}T1}$ or $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ (1 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.

2.4.9. BODIPY-CoA loading assay for RupA

BODIPY-CoA[41] and Sfp[71] were prepared using previously reported procedures. The reaction mixture (50 μL) contained 5 μM of either $Rup_{C1\text{-}A1\text{-}T1}$ or $Rup_{C2\text{-}A2\text{-}T2\text{-}R}$, 1.0 μM Sfp, 5 μM BODIPY-CoA, 10 mM $MgCl_2$, 25 mM Tris pH 8.5, and 50 mM NaCl. Reaction mixtures were incubated for 1 h in the dark at room temperature 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 Bio-Safe Coomassie Stain (BioRad) and imaged again.

2.4.10. T-domain loading assay for RupA

Reaction mixtures (50 μL) contained 25 mM Tris pH 8.5, 50 mM NaCl, 10 mM $MgCl_2$, 250 μM CoA tri-lithium salt, 500 μM DTT, 30 μM of the indicated amino acid, 3 μM of either $Rup_{C1\text{-}A1\text{-}T1}$ or $Rup_{C2\text{-}A2\text{-}T2\text{-}R}$. For the assay with $Rup_{C1\text{-}A1\text{-}T1}$, the reaction mixture was supplemented with 30 μM $RupA_{T1}$ to amplify signal. Amino acids used were ^{14}C-L-Leu (0.1 mCi/mL, 328 mCi/mmol), ^{14}C- L-Val (0.1 mCi/mL, 246 mCi/mmol), ^{14}C- L-Glu (0.1 mCi/mL, 260 mCi/mmol), and ^{14}C- L-Asp (0.1 mCi/mL, 201 mCi/mmol). Loading of the phosphopantetheinyl arm onto the T domains of $RupA_{C1\text{-}A1\text{-}T1}$ (and $RupA_{T1}$) or $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ was initiated by the addition of Sfp (1 μM) to the reaction mixture, followed by incubation at room temperature for 1 h. Loading of the T domain with amino acid was then initiated by the addition of ATP (3 mM). After incubation at room temperature for 1 h, the reaction was quenched by the addition of 100 μL of bovine serum albumin (1 mg/mL) followed by 500 μL of trichloroacetic acid (TCA) (10% (w/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% w/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.

2.4.11. LC-MS assay for C-domain substrate specificity

For the assay with the first module $RupA_{C1\text{-}A1\text{-}T1}$ (Figure 2.15), the reaction mixture (50 μL) contained 25 mM Tris buffer pH 8.5, 50 mM NaCl, 10 mM $MgCl_2$, 400 μM DTT, 4 mM L-Leu, 250 μM CoA-tri-lithium salt, 6.6 % (v/v) DMSO, and $RupA_{C1\text{-}A1\text{-}T1}$ (10 μM). Loading of the phosphopantetheinyl arm onto the T domain of $RupA_{C1\text{-}A1\text{-}T1}$ was initiated by the addition of Sfp (3 μM) to the reaction mixture, followed by incubation at room temperature for 1 h. 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 (1 mM). For the fatty acyl-CoA competition experiment, a stock solution containing all the fatty acyl-CoA substrates was added, each to a final concentration of 142 μM. This mixture was incubated at room temperature for 2 h and quenched by the addition of methanol (125 μL). After incubation on ice for 10 min, the samples were centrifuged (13,000 rpm x 10 min). The protein pellets were washed two times with 125 μL of methanol and dried under a stream of nitrogen gas. Products bound to the T domain were hydrolyzed by the addition of 0.1 M KOH (5 μL) followed by heating at 74 °C for 10 min. The samples were cooled on ice, and 0.1 M HCl (25 μL) was added to the solutions. Finally, methanol (60 μL) was added to the samples, which were then incubated at –80 °C for at least 2 h to precipitate protein. The samples were centrifuged (13,000 rpm x 15 min) and the supernatant was analyzed by LC-MS. Masses were not observed in reactions without ATP, without enzyme, or reactions containing boiled enzyme.

2.4.12. LC-MS assay for *N*-acyl dipeptide production

For assays including both modules $RupA_{C1\text{-}A1\text{-}T1}$ and $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ (Figure 2.17), the reaction mixture (50 µL) contained 25 mM Tris buffer pH 8.5, 50 mM NaCl, 10 mM $MgCl_2$, 400 µM DTT, 250 µM CoA-tri-lithium salt, 6.6 % (v/v) DMSO, $RupA_{C1\text{-}A1\text{-}T1}$ (10 µM), and $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ (10 µM). The amino acid competition experiment contained 4 mM each of L-valine, L-leucine, L-aspartate, and L-glutamate, and the fatty acyl-CoA competition experiment contained 4 mM each of L-leucine and L-glutamate. Loading of the phosphopantetheinyl arm onto the T domains of $RupA_{C1\text{-}A1\text{-}T1}$ and $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ was initiated by the addition of Sfp (3 µM) to the reaction mixture, followed by incubation at room temperature for 1 h. 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 (1 mM). For the fatty acyl-CoA competition experiment, a stock solution containing all the fatty acyl-CoA substrates was added, each to a final concentration of 142 µM. This mixture was incubated at room temperature for 19 h, and an identical workup procedure was followed. Product masses were not observed in reactions without ATP, without enzyme, or reactions containing boiled enzyme.

2.4.13. Synthesis of enzymatic assay standards

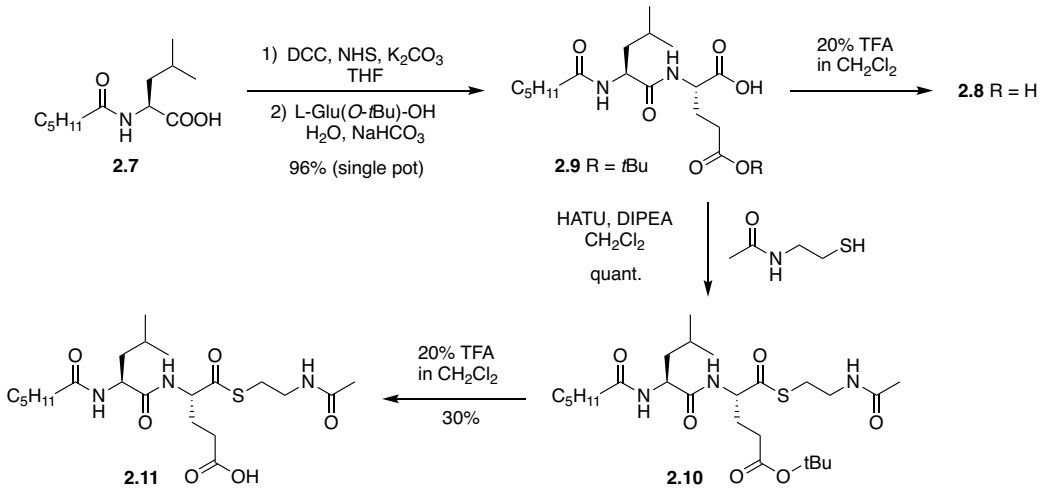
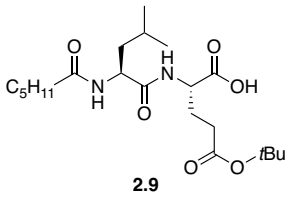


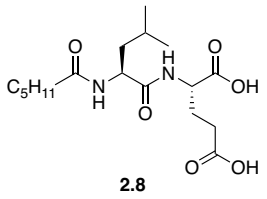
Figure 2.26: Synthesis of enzymatic assay standards.



2.4.13.1. (*S*)-5-(*tert*-Butoxy)-2-((*S*)-2-hexanamido-4-methylpentanamido)-5-oxopentanoic acid (**2.9**):

To an oven-dried flask containing *N*-hexanoyl-L-Leucine (**2.7**) (200 mg, 0.872 mmol, 1.00 equiv), was added DCC (1.01 equiv), NHS (1.01 equiv), anhydrous potassium carbonate (1.00 equiv) and anhydrous THF (5 mL). The reaction mixture was stirred at room temperature for 3 h. The mixture was filtered through glass wool into a suspension of L-Glu(*O*-*t*Bu)-OH (177 mg, 0.872 mmol, 1.00 equiv) in 10% aqueous sodium bicarbonate (20 mL). The glass wool was washed with THF (3 x 5 mL). The reaction mixture was stirred for an additional 2 h. The mixture was neutralized with 5% aqueous citric acid to pH = 7 and extracted with ethyl acetate (3 x 25 mL). The combined organic layers were washed with water (25 mL) and brine (25 mL), dried over Na_2SO_4, filtered, and concentrated in vacuo. The crude product was purified by flash

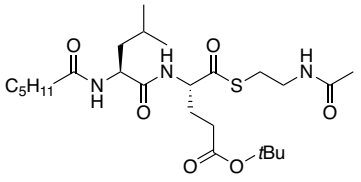
chromatography on silica gel using $CHCl_3$/MeOH/AcOH (93:5:2) to afford the product (362 mg, 96%) as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 7.34 (d, *J* = 5.9 Hz, 1H), 6.63 (m, 1H), 4.60 (t, *J* = 6.7 Hz, 1H), 4.50 (q, *J* = 6.2 Hz, 1H), 2.33 (m, 2H), 2.21 (m, 4H), 2.00 (m, 2H), 1.61 (m, 3H), 1.43 (s, 9H), 1.29 (m, 4H), 0.89 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 175.9, 174.5, 173.4, 172.4, 129.2, 125.4, 81.1, 52.0, 41.4, 36.4, 31.4, 25.5, 24.8, 22.9, 22.3, 21.0, 14.1. HRMS (ESI): Calc'd for formula $C_{21}H_{37}N_2O_6^-$ $[M–H]^-$ 413.2657, found 413.2674.



2.8

2.4.13.2. *N*-Hexanoyl-L-leucyl-L-glutamic acid (**2.8**):

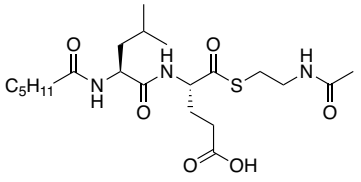
A solution of **2.9** in 20% trifluoroacetic acid in dichloromethane (4.28 mL) with 5 μL water was stirred for 1 h at room temperature. The reaction mixture was concentrated in vacuo, and residual TFA was removed by forming the azeotrope with anhydrous toluene. The resulting crude product was purified by flash chromatography on silica gel using $CHCl_3$/MeOH/AcOH (88:10:2) to afford the product (28 mg, 83%) as a colorless oil. ^{1}H NMR (500 MHz; 10% CD_3OD in $CDCl_3$, referenced to $CDCl_3$): δ 7.56 (d, *J* = 7.9 Hz, 1H), 6.87 (d, *J* = 8.7 Hz, 1H), 4.54 (m, 2H), 2.39 (m, 2H), 2.20 (m, 2H), 1.59 (m, 4H), 1.51 (m, 1H), 1.24 (m, 6H), 0.88 (m, 9H). ^{13}C NMR (126 MHz; 10% CD_3OD in $CDCl_3$, referenced to $CDCl_3$): δ 176.1, 174.5, 173.0, 129.2, 128.4, 41.4, 36.5, 31.5, 30.2, 27.0, 25.5, 24.9, 23.0, 22.5, 22.2, 20.8, 14.1. HRMS (ESI): Calc'd for formula $C_{17}H_{29}N_2O_6^-$ $[M–H]^-$ 357.2031, found 357.2053.



2.10

2.4.13.3. *tert*-Butyl (*S*)-5-((2-acetamidoethyl)thio)-4-((*R*)-2-hexanamido-4-methylpentanamido)-5-oxopentanoate (**2.10**):

To a solution of **2.9** (362 mg, 0.873 mmol) in dry dichloromethane (8.7 mL) were added HATU (1.5 equiv), DIPEA (3.0 equiv), and *N*-(2-mercaptoethyl)acetamide (1.2 equiv) under argon. The reaction mixture was stirred for 18 h at room temperature and then quenched with 10 mL saturated aqueous $NaHCO_3$. The aqueous layer was extracted with dichloromethane (3 x 10 mL). The combined organic layers were washed with 10 mL brine, dried over $MgSO_4$, filtered, and concentrated in vacuo. The crude product was purified by flash chromatography on silica gel using CH_2Cl_2/MeOH (9:1) to afford the product (466 mg, quant.) as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 4.50 (m, 1H) 3.71 (m, 2H), 3.45 (m, 1H), 3.17 (m, 2H), 3.17 (m, 1H), 3.02 (m, 4H), 2.90 (br s, 3H), 2.33 (m, 2H), 2.12 (m, 2H), 1.64 (m, 2H), 1.49 (m, 6H), 1.44 (s, 9H), 1.31 (m, 2H), 0.92 (m, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 196.7, 173.2, 172.6, 161.9, 81.3, 55.5, 45.1, 43.7, 40.2, 39.0, 36.5, 31.6, 30.9, 29.3, 28.3, 26.8, 25.6, 25.0, 23.2, 22.6, 18.8, 17.3, 14.1, 12.0. HRMS (ESI): Calc'd for formula $C_{25}H_{44}N_3O_6S^-$ $[M–H]^-$ 514.2956, found 514.2974.



2.11

2.4.13.4. *N*-hexanoyl-L-Leu-L-Glu-SNAC (**2.11**):

A solution of **2.10** in 20% trifluoroacetic acid in dichloromethane (20 mL) with 5 µL water was stirred for 30 min at 0 ºC. The reaction mixture was concentrated in vacuo, and residual TFA was removed by forming an azeotrope with anhydrous toluene. The crude material was purified using a 15.5 g RediSep Rf Gold C18Aq column on a Combiflash Rf Teledyne ISCO Purification System (mobile phase A: 0.1% TFA in water, mobile phase B:0.1% TFA in acetonitrile) to afford the product (31 mg, 30%) as a colorless oil. ^{1}H NMR (500 MHz; DMSO-d_6): δ 8.53 (s, 1H), 8.03 (m, 1H), 7.93 (m, 1H), 4.36 (m, 2H), 3.11 (m, 2H), 2.85 (m, 2H), 2.27 (m, 2H), 2.10 (m, 2H), 1.78 (s, 3H), 1.48 (m, 4H), 1.24 (m, 3H), 0.86 (m, 9H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 200.7, 173.7, 172.9, 172.2, 58.3, 50.8, 39.5, 38.1, 35.1, 30.8, 29.6, 27.6, 26.3, 26.2, 25.0, 24.2, 23.1, 22.9, 22.5, 21.9, 21.5, 13.9. HRMS (ESI): Calc'd for formula $C_{21}H_{36}N_3O_6S^-$ $[M–H]^-$ 459.2403, found 459.2414.

2.4.14. Monitoring consumption of NAD(P)H in reconstitution assays

Similar assay conditions to those used in the biosynthetic reconstitution experiments were used for monitoring NAD(P)H consumption by various RupA R domain-containing constructs. The reaction mixture (100 µL) contained 25 mM Tris buffer pH 8.5, 50 mM NaCl, 10 mM $MgCl_2$, 400 µM DTT, 250 µM CoA-tri-lithium salt, 3.3 % (v/v) DMSO. Different assays were set up to contain either $RupA_{T2\text{-}R}$ (10 µM), $RupA_R$ (10 µM), or $RupA_{C2\text{-}A2\text{-}T2\text{-}R}$ (10 µM). Loading of the ppant arm onto the T domain of these various constructs was initiated by the addition of

Sfp (1.5 μM) to the reaction mixture, followed by incubation at room temperature for 1 h. SNAC substrate **2.7** (1 mM) and NAD(P)H were then added to the reaction mixture, and the reaction mixture was transferred to a transparent 96 well microplate. The reaction mixture was incubated at room temperature and absorbance at 340 nm was monitored each minute for 1 h. In each iteration of this assay, no significant differences were observed between the full reactions and reactions either containing boiled enzyme or lacking the SNAC substrate.

2.4.15. Extraction of fatty acyl-CoAs from *R. bromii* cultures

For extraction of fatty acyl-CoA's from cultures of *R. bromii*, saturated cultures were grown and inoculated 1:100 in 20 mL RUM medium. After growth to saturation, the cells were pelleted by centrifugation and resuspended in 200 μL of 10 mM ammonium acetate buffer (pH 5.3). The cells were lysed by sonication on ice using a Branson Digital Sonifier equipped with a Double Stepped Microtip (10 s pulse, 30 s rest, 25% amplitude, 4 cycles). To the homogenized lysate was added 200 μL of ice cold chloroform/methanol (2:1). The sample was vortexed and centrifuged to separate layers (13,000 rpm x 30 min). The upper phase was collected and washed with ice cold chloroform (2 x 100 μL). To the upper phase was added 50 μL acetonitrile, and this mixture incubated at 0 ºC for 10 min to precipitate solids. The samples were centrifuged (13,000 rpm x 10 min), diluted with water, and lyophilized. For inclusion in the C-starter domain biosynthetic reconstitution assay, the lyophilized sample was resuspended in 25 μL water. In one experiment, a mass corresponding to formation of acetyl-Leu was observed at the expected retention time. However, these results were not confirmed by high resolution mass spectrometry.

These extracted fatty acyl CoA samples were also analyzed on an Advion Expression CMS-L mass spectrometer in negative ion mode coupled to an Agilent 1200 Series HPLC. For liquid

chromatography, 20 µL of sample was injected onto an Acclaim C8 column (120 Å pore size, 150 mm x 2.1 mm, 3 µm particle size). Mobile phase A was 10% acetonitrile in 10 mM ammonium acetate buffer (pH 5.3), and mobile phase B was acetonitrile. The flow rate was 0.3 mL/min. The mobile phase composition started at 50% B, which was maintained for 5 min. Samples were then subjected to a linear gradient over 15 min to 30% B. The column was then re-equilibrated to 5% B over 4.9 min. The capillary temperature was set to 275 ºC, the capillary voltage was set to 180 V, the voltage offset was 30 V, the voltage span was 20 V, the source gas temp was 200 ºC, and the ESI voltage was 2.5 kV. Under these conditions, which were successfully used to visualize acyl-CoA standards, expected masses of even chain fatty acyl CoA's could not be observed in extracts.

2.5. References

1. Schneider, B. A. & Balskus, E. P. Discovery of small molecule protease inhibitors by investigating a widespread human gut bacterial biosynthetic pathway. *Tetrahedron* **74,** 3215–3230 (2018).

2. Abell, G. C. J., Cooke, C. M., Bennett, C. N., Conlon, M. a. & McOrist, A. L. Phylotypes related to *Ruminococcus bromii* are abundant in the large bowel of humans and increase in response to a diet high in resistant starch. *FEMS Microbiol. Ecol.* **66,** 505–515 (2008).

3. Moore, W. E. C. & Moore, L. H. Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.* **61,** 3202–3207 (1995).

4. Collins, M. D. *et al.* The phylogeny of the genus *Clostridium*: Proposal of five new genera and eleven new species combinations. *Int. J. Syst. Bacteriol.* **44,** 812–826 (1994).

5. Lay, C. *et al.* Design and validation of 16S rRNA probes to enumerate members of the *Clostridium leptum* subgroup in human faecal microbiota. *Environ. Microbiol.* **7,** 933–946 (2005).

6. Atarashi, K. *et al.* Induction of colonic regulatory T cells by indigenous *Clostridium*

species. *Science* **331,** 337–341 (2011).

7. Sokol, H. *et al. Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 16731–16736 (2008).

8. Kabeerdoss, J., Sankaran, V., Pugazhendhi, S. & Ramakrishna, B. S. *Clostridium leptum* group bacteria abundance and diversity in the fecal microbiota of patients with inflammatory bowel disease: a case-control study in India. *BMC Gastroenterol.* **13,** 20 (2013).

9. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J.* **6,** 1535–1543 (2012).

10. Englyst, H. N. & Macfarlane, G. T. Breakdown of resistant and readily digestible starch by human gut bacteria. *J. Sci. Food Agric.* **37,** 699–706 (1986).

11. Walker, A. W. *et al.* The species composition of the human intestinal microbiota differs between particle-associated and liquid phase communities. *Environ. Microbiol.* **10,** 3275–3283 (2008).

12. Ze, X. *et al.* Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic Firmicutes bacterium *Ruminococcus bromii*. *MBio* **6,** e01058-15 (2015).

13. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158,** 1402–1414 (2014).

14. Guo, C. *et al.* Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell* **168,** 517–526 (2017).

15. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).

16. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic machinery, and mechanisms. *Chem. Rev.* **106,** 3468–3496 (2006).

17. Yeh, H.-H. *et al.* Resistance gene-guided genome mining: Serial promoter exchanges in *Aspergillus nidulans* reveal the biosynthetic pathway for fellutamide B, a proteasome inhibitor. *ACS Chem. Biol.* **11,** 2275–2284 (2016).

18. Chen, Y., McClure, R. A., Zheng, Y., Thomson, R. J. & Kelleher, N. L. Proteomics guided discovery of flavopeptins: anti-proliferative aldehydes synthesized by a reductase domain-containing non-ribosomal peptide synthetase. *J. Am. Chem. Soc.* **135,** 10449–10456 (2013).

19. Carroll, I. M. *et al.* Fecal protease activity is associated with compositional alterations in the intestinal microbiota. *PLoS One* **8,** e78017 (2013).

20. Moore, W. E. C., Cato, E. P. & Holdeman, L. V. *Ruminococcus bromii* sp. n. and emendation of the description of *Ruminococcus* Sijpestein. *Int. J. Syst. Bacteriol.* **22,** 78–80 (1972).

21. Brotherton, C. A. & Balskus, E. P. A prodrug resistance mechanism is involved in colibactin biosynthesis and cytotoxicity. *J. Am. Chem. Soc.* **135,** 3359–3362 (2013).

22. Reimer, D., Pos, K. M., Thines, M., Grün, P. & Bode, H. B. A natural prodrug activation mechanism in nonribosomal peptide synthesis. *Nat. Chem. Biol.* **7,** 888–890 (2011).

23. Imker, H. J., Krahn, D., Clerc, J., Kaiser, M. & Walsh, C. T. *N*-Acylation during glidobactin biosynthesis by the tridomain nonribosomal peptide synthetase module GlbF. *Chem. Biol.* **17,** 1077–1083 (2010).

24. Bachmann, B. O. & Ravel, J. Chapter 8: Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.* **458,** 181–217 (2009).

25. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36,** 2295–2300 (2008).

26. Barajas, J. F. *et al.* Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem. Biol.* **22,** 1018–1029 (2015).

27. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).

28. Harris, N. C. *et al.* Biosynthesis of isonitrile lipopeptides by conserved nonribosomal peptide synthetase gene clusters in Actinobacteria. *Proc. Natl. Acad. Sci.* **114,** 201705016 (2017).

29. Konz, D. & Marahiel, M. A. How do peptide synthetases generate structural diversity? *Chem. Biol.* **6,** R39–R48 (1999).

30. Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173,** 697–703 (1991).

31. Kisselev, A. F. & Goldberg, A. L. Proteasome inhibitors: from research tools to drug candidates. *Chem. Biol.* **8,** 739–758 (2001).

32. Spears, R. J. & Fascione, M. A. Site-selective incorporation and ligation of protein aldehydes. *Org. Biomol. Chem.* **14,** 7622–7638 (2016).

33. Sletten, E. M. & Bertozzi, C. R. Bioorthogonal chemistry: Fishing for selectivity in a sea of functionality. *Angew. Chemie - Int. Ed.* **48,** 6974–6998 (2009).

34. Stephanopoulos, N. & Francis, M. B. Choosing an effective protein bioconjugation strategy. *Nat. Chem. Biol.* **7,** 876–884 (2011).

35. Agarwal, P., van der Weijden, J., Sletten, E. M., Rabuka, D. & Bertozzi, C. R. *A Pictet-Spengler ligation for protein chemical modification*. *Proceedings of the National Academy of Sciences* **110,** (2013).

36. Agarwal, P. *et al.* Hydrazino-pictet-spengler ligation as a biocompatible method for the generation of stable protein conjugates. *Bioconjug. Chem.* **24,** 846–851 (2013).

37. Kitov, P. I., Vinals, D. F., Ng, S., Tjhung, K. F. & Derda, R. Rapid, hydrolytically stable modification of aldehyde-terminated proteins and phage libraries. *J. Am. Chem. Soc.* **136,** 8149–8152 (2014).

38. Maxson, T. *et al.* Targeting reactive carbonyls for identifying natural products and their biosynthetic origins. *J. Am. Chem. Soc.* **138,** 15157–15166 (2016).

39. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: A web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84,** 5035–5039 (2012).

40. Linne, U. & Marahiel, M. A. Reactions catalyzed by mature and recombinant nonribosomal peptide synthetases. *Methods Enzymol.* **388,** 293–315 (2004).

41. La Clair, J. J., Foley, T. L., Schegg, T. R., Regan, C. M. & Burkart, M. D. Manipulation of carrier proteins in antibiotic biosynthesis. *Chem. Biol.* **11,** 195–201 (2004).

42. Nakamura, H., Hamer, H. A., Sirasani, G. & Balskus, E. P. Cylindrocyclophane biosynthesis involves functionalization of an unactivated carbon center. *J. Am. Chem. Soc.* **134,** 18518–18521 (2012).

43. Watanabe, C. M. H. & Townsend, C. A. Initial characterization of a type I fatty acid

synthase and polyketide synthase multienzyme complex NorS in the biosynthesis of aflatoxin B1. *Chem. Biol.* **9,** 981–988 (2002).

44. Wilson, D. J., Shi, C., Teitelbaum, A. M., Gulick, A. M. & Aldrich, C. C. Characterization of AusA: A dimodular nonribosomal peptide synthetase responsible for the production of aureusimine pyrazinones. *Biochemistry* **52,** 926–937 (2013).

45. Gaitatzis, N., Kunze, B. & Müller, R. In vitro reconstitution of the myxochelin biosynthetic machinery of *Stigmatella aurantiaca* Sg a15: Biochemical characterization of a reductive release mechanism from nonribosomal peptide synthetases. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 11136–11141 (2001).

46. Kallberg, Y., Oppermann, U., Jörnvall, H. & Persson, B. Short-chain dehydrogenases/reductases (SDRs). Coenzyme-based functional assignments in completed genomes. *Eur. J. Biochem.* **269,** 4409–4417 (2002).

47. Schweizer, E. & Hofmann, J. Microbial type I fatty acid synthases (FAS): Major players in a network of cellular FAS systems. *Microbiol. Mol. Biol. Rev.* **68,** 501–517 (2004).

48. Magnuson, K., Jackowski, S., Rock, C. O. & Cronan, J. E. Regulation of fatty acid biosynthesis in *Escherichia coli*. *Microbiol. Rev.* **57,** 522–42 (1993).

49. Fujita, Y., Matsuoka, H. & Hirooka, K. Regulation of fatty acid metabolism in bacteria. *Mol. Microbiol.* **66,** 829–839 (2007).

50. Campbell, J. W., Morgan-Kiss, R. M. & Cronan, J. E. A new *Escherichia coli* metabolic competency: Growth on fatty acids by a novel anaerobic β-oxidation pathway. *Mol. Microbiol.* **47,** 793–805 (2003).

51. Khurana, P., Gokhale, R. S. & Mohanty, D. Genome scale prediction of substrate specificity for acyl adenylate superfamily of enzymes based on active site residue profiles. *BMC Bioinformatics* **11,** 57 (2010).

52. Shah, M. B. *et al.* The 2.1 Å crystal structure of an acyl-CoA synthetase from Methanosarcina acetivorans reveals an alternate acyl-binding pocket for small branched acyl substrates. *Proteins Struct. Funct. Bioinforma.* **77,** 685–698 (2009).

53. Meng, Y., Ingram-Smith, C., Cooper, L. L. & Smith, K. S. Characterization of an archaeal medium-chain acyl coenzyme A synthetase from *Methanosarcina acetivorans*. *J. Bacteriol.* **192,** 5982–5990 (2010).

54. Kasuya, F., Oti, Y., Tatsuki, T. & Igarashi, K. Analysis of medium-chain acyl-coenzyme A esters in mouse tissues by liquid chromatography-electrospray ionization mass

spectrometry. *Anal. Biochem.* **325,** 196–205 (2004).

55. Sarma-Rupavtarm, R. B., Ge, Z., Schauer, D. B., Fox, J. G. & Polz, M. F. Spatial distribution and stability of the eight microbial species of the altered Schaedler flora in the mouse gastrointestinal tract. *Appl. Environ. Microbiol.* **70,** 2791–2800 (2004).

56. Wannemuehler, M. J., Overstreet, A., Ward, D. V & Phillips, J. Draft genome sequences of the Altered Schaedler Flora, a defined bacterial community from gnotobiotic mice. *genomeA* **2,** e00287-14 (2014).

57. Brand, M. W. *et al.* The Altered Schaedler Flora: Continued applications of a defined murine microbial community. *ILAR J.* **56,** 169–178 (2015).

58. Biggs, M. B. *et al.* Systems-level metabolism of the altered Schaedler flora, a complete gut microbiota. *ISME J.* **11,** 426–438 (2017).

59. Röttig, M. *et al.* NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** 362–367 (2011).

60. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **368,** 1500–1517 (2007).

61. Blin, K. *et al.* antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41,** W204–W212 (2013).

62. Joseph, R. C., Kim, N. M. & Sandoval, N. R. Recent developments of the synthetic biology toolkit for Clostridium. *Front. Microbiol.* **9,** 154 (2018).

63. Christiansen, G., Philmus, B., Hemscheidt, T. & Kurmayer, R. Genetic variation of adenylation domains of the anabaenopeptin synthesis operon and evolution of substrate promiscuity. *J. Bacteriol.* **193,** 3822–3831 (2011).

64. Qiao, K. *et al.* A fungal nonribosomal peptide synthetase module that can synthesize thiopyrazines. *Org. Lett.* **13,** 1758–1761 (2011).

65. Wyatt, M. A. *et al. Staphylococcus aureus* nonribosomal peptide secondary metabolites regulate virulence. *Science* **329,** 294–296 (2010).

66. Read, J. A. & Walsh, C. T. The lyngbyatoxin biosynthetic assembly line: Chain release by four-electron reduction of a dipeptidyl thioester to the corresponding alcohol. *J. Am. Chem. Soc.* **129,** 15762–15763 (2007).

67. Li, Y., Weissman, K. J. & Müller, R. Myxochelin biosynthesis: direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. *J. Am. Chem. Soc.* **130,** 7554–7555 (2008).

68. Kopp, F., Mahlert, C., Grünewald, J. & Marahiel, M. A. Peptide macrocyclization: the reductase of the nostocyclopeptide synthetase triggers the self-assembly of a macrocyclic imine. *J. Am. Chem. Soc.* **128,** 16478–16479 (2006).

69. Miyazaki, K., Martin, J., Marinsek-Logar, R. & Flint, H. . Degradation and utilization of xylans by the rumen anaerobe *Prevotella bryantii* (formerly *P. ruminicola* subsp. *brevis*) $B_1$4. *Anaerobe* **3,** 373–381 (1997).

70. Rio, D. C., Ares, M., Hannon, G. J. & Nilsen, T. W. Purification of RNA using TRIzol (TRI Reagent). *Cold Spring Harb. Protoc.* (2010). doi:10.1101/pdb.prot5439

71. Mofid, M. R., Marahiel, M. A., Ficner, R. & Reuter, K. Crystallization and preliminary crystallographic studies of Sfp: A phosphopantetheinyl transferase of modular peptide synthetases. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55,** 1098–1100 (1999).

3. Synthesis and Bioactivity Evaluation of Ruminopeptin and Analogues

This chapter is an unofficial adaptation of previously published work.[1]

3.1. Introduction

As discussed in Chapter 2, our work to elucidate the predicted peptide aldehyde product(s) of the *rup* gene cluster in *R. bromii*, the ruminopeptins, revealed that this gene cluster is transcribed under standard culture conditions. However, we could not isolate any aldehyde compounds from these cultures, likely due to issues with stability. Therefore, we synthesized a series of ruminopeptin analogues in order to access them and study their bioactivity. Synthesis of putative natural products that have eluded isolation in order to evaluate their bioactivity is an emerging area of interest, and several compounds from the gut microbiota have been discovered in this way over the past several years.[2] The synthesis of such molecules enables access to large quantities of material for screening efforts. To determine biological effects, it is helpful to be able to predict what sort of phenotype might be expected for the compounds of interest. Our work was enabled by the hypothesis that ruminopeptins would likely target post-glutamyl hydrolyzing proteases.

In this chapter, we used a short solution phase synthesis to access 12 potential analogues of ruminopeptin. Relying on prediction of a specific microbial target for these compounds based on the presence of a glutamate residue in the P1 position, we then evaluated these compounds as inhibitors of the glutamyl endopeptidase from *Staphylococcus aureus*. Glutamyl endopeptidases are implicated in the life cycle and virulence of *Enterococcus faecalis* and *S. aureus*, and we evaluated the compounds in a number of phenotypic assays with these organisms. Finally, we

assessed the presence of glutamyl endopeptidases in the human gut microbiota and discuss an intriguing secreted serine protease from *Faecalibacterium prausnitzii* as a target for further study.

3.2. Results and discussion

3.2.1. Design and synthesis of ruminopeptin analogues

We used the information that we obtained from our bioinformatic and biochemical analyses of the *rup* biosynthetic pathway to inspire the chemical synthesis of a focused library of predicted ruminopeptin structures. We first designed 12 analogues of the predicted *N*-acyl dipeptide aldehyde scaffold that contained varied *N*-acyl substituent and amino acid components (Figure 3.1). We then accessed these compounds using a solution-phase synthetic route adapted from previous syntheses of aspartyl and glutamyl peptide aldehydes.[3,4] Synthesis has previously been used as a tool to access peptide aldehyde natural products when isolation efforts yielded insufficient quantities of pure material for activity screening.[5] In our case, we envisioned that accessing a small library of ruminopeptin analogs could not only provide compounds for assays but also enable structure-activity relationship studies.

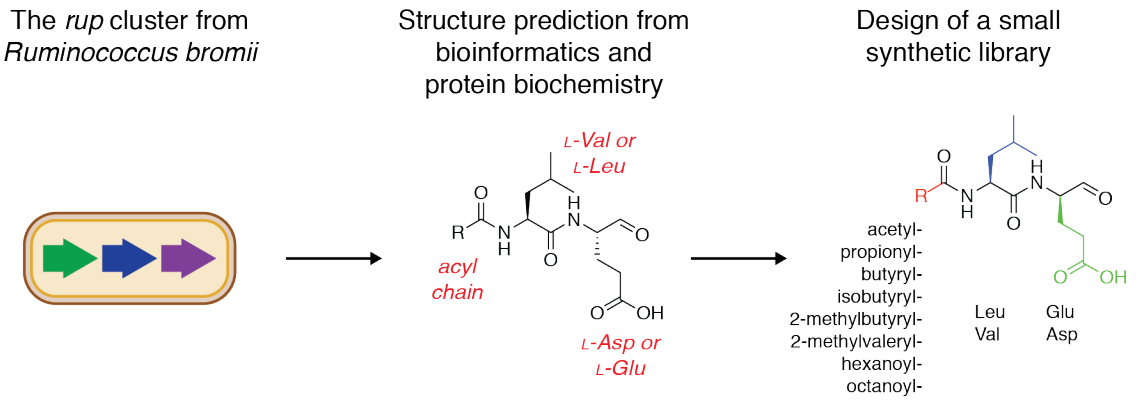


Figure 3.1: Design of a small library of ruminopeptin analogues.

Relying on the bioinformatic and biochemical analyses conducted in Chapter 2, we designed a small set of ruminopeptin analogues to synthesize in this chapter. The strutures were selected to incorporate acyl chains of varying lengths and several combinations of the amino acids that were activated by the RupA nonribosomal peptide synthetase (NRPS).

From *N*-Cbz- and *O*-*t*Bu-protected L-glutamate and L-aspartate precursors **3.1a–b**, we accessed key intermediates containing an aldehyde masked as a semicarbazone functional group using a previously reported reaction sequence (**3.4a**–**b**, Figure 3.2).[3,4] Briefly, we formed Weinreb amides **3.2a–b** and then generated the corresponding aldehydes by reduction with $LiAlH_4$. These aldehydes were reacted with semicarbazide hydrochloride to form the protected semicarbazone compounds **3.3a–b**, and the Cbz protecting groups were then removed by hydrogenation to afford the key semicarbazone-protected intermediates **3.4a–b**. To generate the library, these intermediates were then coupled to *N*-acylated L-leucine and L-valine derivatives (**3.5a**–**i**, Figure 3.3) using the peptide coupling reagent 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU) to yield **3.6a**–**l** (Table 3.1). From these coupling products, deprotection of the *O*-*t*Bu ester proceeded with 20% trifluoroacetic acid in dichloromethane. Finally, transfer of the semicarbazide functional group to formaldehyde under acidic conditions and corresponding regeneration of the aldehyde provided

the desired peptide aldehyde products **3.7a**–**l** (Table 3.2). Using this route, we accessed a small library of ruminopeptin analogues on a multi-milligram scale (6–42% overall yield, 7–36 mg obtained). Comparison of these structures with compounds indexed by the Chemical Abstracts Service (as accessed through SciFinder)[6] revealed that only **3.7a** is a previously reported compound.

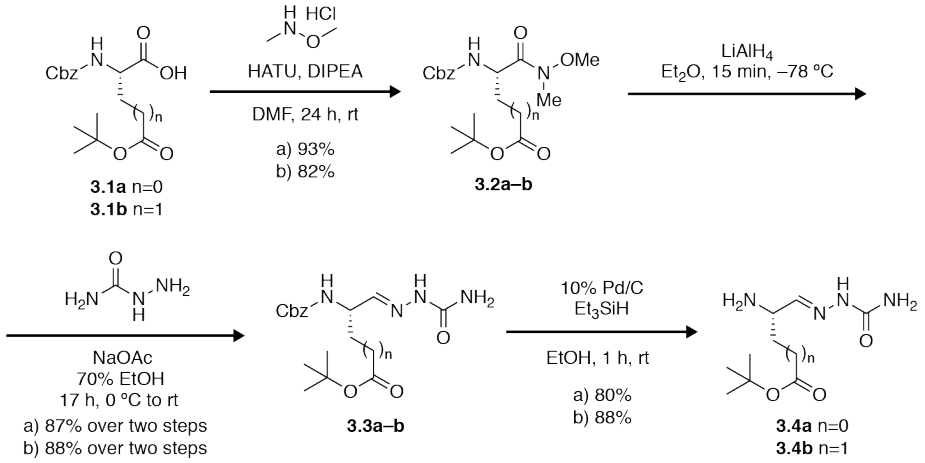


Figure 3.2: Synthesis of protected aldehyde building blocks 3.4a and 3.4b.

Reactions were performed according to previously published procedures.[3,4] (HATU = 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate, DIPEA = *N*,*N*-Diisopropylethylamine).

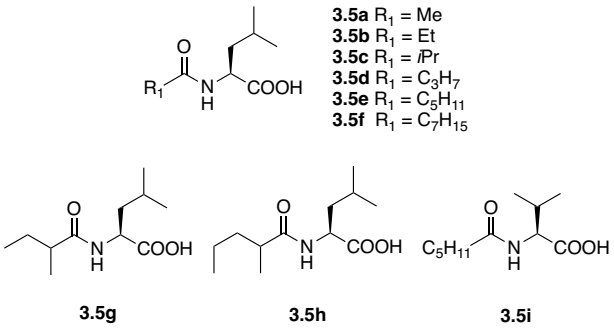
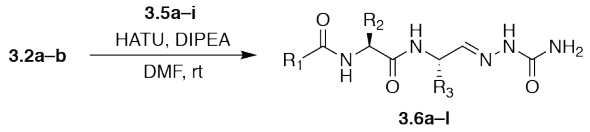


Figure 3.3: *N*-acyl amino acid synthetic precursors 3.5a–i used in this study.

N-acyl amino acids were purchased or prepared using the Schotten-Baumann reaction of acyl chlorides with amino acids under basic conditions.

Table 3.1: Synthesis of semicarbazone intermediates 3.6a–l.

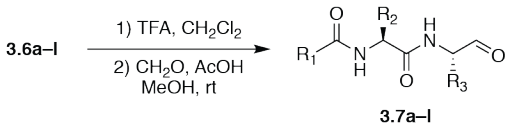
Coupling reaction between *O*-*t*Bu-protected semicarbazones **3.2a–b** and *N*-acyl amino acids **3.5a–i** gave semicarbazone intermediates **3.6a–l** (% Yield = isolated yield).



Entry	Product	R_1	R_2	R_3	% Yield
1	**3.6a**	Me	Leu	Asp (*O*-*t*Bu)	31%
2	**3.6b**	Me	Leu	Glu (*O*-*t*Bu)	28%
3	**3.6c**	C_2H_5	Leu	Glu (*O*-*t*Bu)	65%
4	**3.6d**	C_3H_7	Leu	Glu (*O*-*t*Bu)	63%
5	**3.6e**	*i*Bu	Leu	Glu (*O*-*t*Bu)	48%
6	**3.6f**		Leu	Glu (*O*-*t*Bu)	62%
7	**3.6g**		Leu	Glu (*O*-*t*Bu)	87%
8	**3.6h**	C_5H_{11}	Leu	Glu (*O*-*t*Bu)	88%
9	**3.6i**	C_5H_{11}	Val	Glu (*O*-*t*Bu)	67%
10	**3.6j**	C_5H_{11}	Leu	Asp (*O*-*t*Bu)	39%
11	**3.6k**	C_5H_{11}	Val	Asp (*O*-*t*Bu)	88%
12	**3.6l**	C_7H_{15}	Leu	Glu (*O*-tBu)	43%

Table 3.2: Synthesis of ruminopeptin analogues 3.7a–l.

Removal of *O*-*t*Bu protecting groups from **3.6a–l** and exchange of semicarbazones with formaldehyde afforded the desired ruminopeptin analogues **3.7a–l** (% Yield = isolated yield over 2 steps)



Entry	Product	R_1	R_2	R_3	% Yield*
1	**3.7a**	Me	Leu	Asp	66%
2	**3.7b**	Me	Leu	Glu	60%
3	**3.7c**	C_2H_5	Leu	Glu	64%
4	**3.7d**	C_3H_7	Leu	Glu	28%
5	**3.7e**	*i*Bu	Leu	Glu	55%
6	**3.7f**		Leu	Glu	48%
7	**3.7g**		Leu	Glu	16%
8	**3.7h**	C_5H_{11}	Leu	Glu	34%
9	**3.7i**	C_5H_{11}	Val	Glu	23%
10	**3.7j**	C_5H_{11}	Leu	Asp	24%
11	**3.7k**	C_5H_{11}	Val	Asp	17%
12	**3.7l**	C_7H_{15}	Leu	Glu	15%

3.2.2. Evaluation of ruminopeptins as inhibitors of glutamyl endopeptidases

With access to sufficient quantities of peptide aldehydes **3.7a-l**, we could begin to identify potential target(s) of these molecules. Our biosynthetic reconstitution experiments strongly suggested that ruminopeptin contains a glutamate residue in its P1 position. We thus gained insights into potential targets by comparing the predicted structures of the ruminopeptins to the known substrate specificities of secreted microbial serine and cysteine proteases, as well as host proteases. As specific post-glutamyl hydrolyzing activity is rare among microbial proteases and unknown among human proteases, this literature analysis revealed only one promising candidate class of targets: the glutamyl endopeptidases. Glutamyl endopeptidases are a class of secreted serine proteases found in several bacterial species, including the human pathogens *S. aureus*[7] (SspA, also known as endoproteinase GluC/V8 protease) and *E. faecalis*[8] (SprE). These proteases are regularly found encoded in quorum-sensing regulated operons alongside additional proteases (metalloprotease GelE in *E. faecalis* and cysteine protease SspB in *S. aureus*), and glutamyl endopeptidases may regulate the action of these enzymes.[9,10]

We screened our library of ruminopeptin analogues (**3.7a–l**) for their ability to inhibit the activity of SspA in vitro. In this assay, the protease was pre-incubated with a peptide aldehyde for 10 min. Protease activity was then quantified by measuring the increase in fluorescence corresponding to the release of the 7-amino-4-methylcoumarin (AMC) fluorophore from the fluorogenic peptide substrate Z-Leu-Leu-Glu-AMC (Figure 3.4A).[11] We found that several of the synthetic compounds inhibit SspA, with approximately 50% inhibition observed for the most potent compounds, medium-chain acyl analogues **3.7h** and **3.7l**, at 10 μM (Figure 3.4B, Table 3.3). Intriguingly, our in vitro reconstitution experiments described in Chapter 2 suggested that these compounds are also among the mostly likely products generated by the *rup* gene cluster.

We observed reduced SspA inhibition with the short-chain acyl analogues **3.7b**–**d** and insignificant inhibitory activity with the branched acyl chain analogues **3.7e**–**g** or aspartyl analogues **3.7a**, **3.7j** and **3.7k**. To further confirm these results, we determined IC50 values for compounds **3.7h** (34.6 ± 4.0 µM) and **3.7l** (51.4 ± 17.2 µM) in an identical assay (Figure 3.4C).

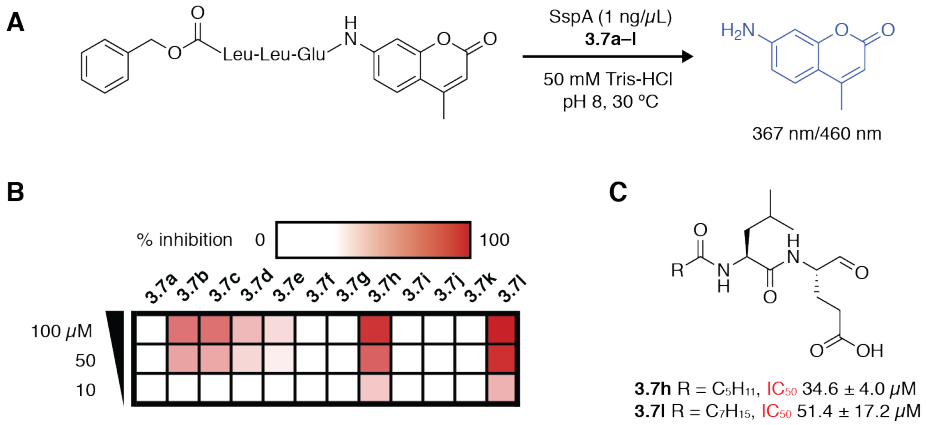


Figure 3.4: Ruminopeptin-like compounds inhibit SspA from *S. aureus*.

(A) In vitro assay setup. (B) Inhibition profile of synthetic ruminopeptin analogs against SspA. The assays were conducted by pre-incubating 1 ng/µL SspA with inhibitor for 10 min at room temperature followed by addition of fluorogenic peptide substrate Z-Leu-Leu-Glu-AMC to a final concentration of 75 µM. Fluorescence (367 nm excitation/460 nm emission) was then monitored for 20 min at 30 ºC and inhibitor efficiency was calculated by comparing the slope of the linear portion of the curve with the negative control (no inhibitor). Assays were performed in duplicate and inhibitor efficiency is reported as a mean of both trials. (C) IC50 values were determined for compounds **3.7h** and **3.7l**. The results are reported as the average ± standard deviation of IC50 values calculated for three independent series of serial dilutions.

Table 3.3: Inhibition profile of 12 peptide aldehydes against SspA.

Values are given as % inhibition and are a mean of duplicate trials.

	Inhibitor concentration (μM)		
	100	50	10
3.7a	68.7	55.7	19.7
3.7b	-20.3	8.6	18.3
3.7c	69.2	54.2	16.8
3.7d	49.7	41.4	16
3.7e	40.3	35.3	12.4
3.7f	20.3	15.5	23
3.7g	19.9	9.1	15.1
3.7h	85.7	73.3	45.6
3.7i	25.6	17.1	7.1
3.7j	16.9	20	22.6
3.7k	4.7	17.9	23.8
3.7l	94.4	86.7	51.6

To better understand the interaction between **3.7h** and SspA, we performed a docking analysis using Glide.[12] Substrate recognition by SspA is reported to rely on an electrostatic interaction between the negatively charged side chain of the glutamate residue in position P1 of the peptide substrate and the positively charged *N*-terminal amine of SspA.[13] We observed a similar interaction between this *N*-terminal amine and the glutamyl side chain of **3.7h** when we docked this inhibitor into the crystal structure of SspA (PDB: 1QY6). The electrophilic aldehyde warhead of **3.7h** was also located within reasonable proximity (4.6 Å) to the nucleophilic Ser residue (Figure 3.5), suggesting that this inhibitor binds the protease similarly to a model protein substrate and in an orientation that would facilitate formation of a reversible, covalent hemiacetal linkage.

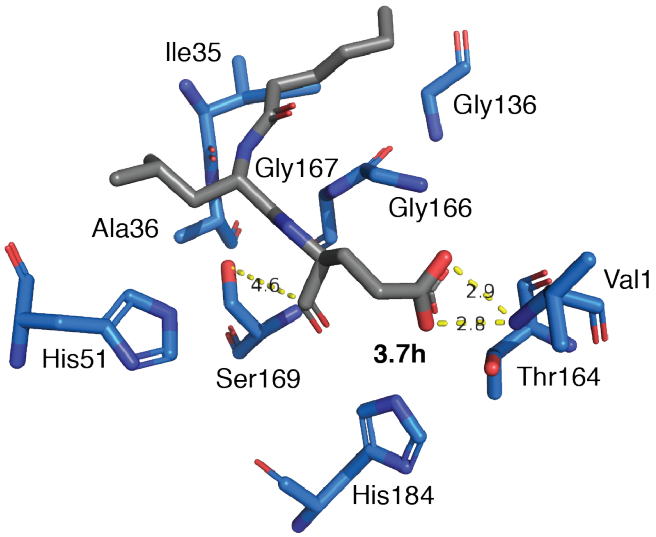
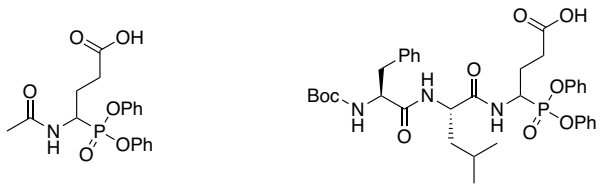


Figure 3.5: Docking of inhibitor 3.7h in the crystal structure of SspA.

Potential interaction between peptide aldehyde **3.7h** (grey) and SspA (blue). The inhibitor was docked into the crystal structure of SspA (PDB: 1QY6) using the induced fit docking algorithm in Glide.

Due to the unique substrate specificity of SspA, as well as general interest in the synthetic challenge of installing a reactive electrophile on an acidic amino acid mimetic, several inhibitors of SspA have previously been synthesized (Figure 3.6). In 1998, Walker and coworkers synthesized diphenyl phosphonate analogues of glutamic acid, including **3.8**, and evaluated them as inhibitors of SspA.[11] In 2013, Oleksyszyn and coworkers synthesized additional peptidyl derivatives of this phosphonic glutamic acid analogue and found an inhibitor, **3.9**, that is marginally more active than the one originally reported.[14] They also evaluated these compounds as inhibitors of V8 in a human IgG degradation gel electrophoresis assay.[14] Though there is no particular reason to think that this is an important target of SspA during *S. aureus* infection, a 1992 study had previously identified the ability of SspA to degrade this and other human antibodies.[15] Despite this precedent, the work described in this chapter provides the first

indication of endogenous inhibition of glutamyl endopeptidases by microbial natural products and is also the first report of peptide aldehyde inhibitors of this enzyme class.



3.8

k_2/K_i = 5.3 ± 0.5 x 10^3 $M^{-1}s^{-1}$

3.9

k_2/K_i = 8.5 x 10^3 $M^{-1}s^{-1}$

Figure 3.6: Known synthetic compounds that inhibit glutamyl endopeptidases.
Select diphenyl phosphonate inhibitors that have previously been evaluated as inhibitors of SspA in vitro.[11,14] Inhibitors were synthesized as mixtures of isomers at the undefined stereocenters. (k_2/K_i = apparent second-order rate constant for irreversible formation of the enzyme-inhibitor complex, i.e. inactivation of the protease.)

3.2.3. Potential biological implications of glutamyl endopeptidase inhibition

The observation that putative peptide aldehydes derived from a gut commensal could inhibit a bacterial glutamyl endopeptidase made us curious about the relevance of these proteases within the human gut. Of the species known to produce glutamyl endopeptidases, the two that are most obviously related to human gastrointestinal health and disease are *S. aureus* (SspA) and *E. faecalis* (SprE). Though *S. aureus* is more commonly associated with the nasal microbiota and can be detected from nasal swabs of approximately 40% of healthy individuals, studies have consistently detected this bacterium in the stool microbiomes of approximately 20% of healthy individuals.[16] Indeed, intestinal carriage of *S. aureus* is hypothesized to contribute to bacterial dissemination in the environment.[16] In comparison, the opportunistic pathogen *E. faecalis* can be detected in 47% of fecal samples from healthy individuals,[17] and its glutamyl endopeptidase SprE resembles SspA (49% amino acid similarity and 27% identity).[8] Therefore, the homologous

E. faecalis glutamyl endopeptidase represents an additional possible target that may be more relevant within the habitat of *R. bromii*.

E. faecalis has two major extracellular proteases, glutamyl endopeptidase SprE and gelatinase GelE, and they have been extensively studied for their roles in bacterial physiology and pathogenesis (reviewed by Shankar and coworkers[18] and by Murray and coworkers[19]). These proteases are regulated by the *fsr* operon, which consists of six genes: a quorum sensing system and two protease genes that it regulates. One of these genes (FsrD) encodes a peptide lactone, gelatinase biosynthesis-activating pheromone (GBAP). This gene is processed and secreted by FsrB, and the system senses and responds to extracellular accumulation of this molecule with FsrC, a histidine kinase. Upon activation, FsrC phosphorylates the response regulator and transcription factor FsrA, leading to transcription of GelE (a metalloprotease) and SprE (a glutamyl endopeptidase).[19] Disruption of the extracellular protease network in this organism may impact biofilm formation and autolytic activity, which may in turn affect *E. faecalis* virulence.[9,19,20]

The involvement of SprE and GelE in *E. faecalis* biofilm formation has been extensively investigated (Table 3.4). In 2004, Murray and coworkers showed that an *sprE* insertion mutant in OG1RF is not attenuated in biofilm formation, but that a Δ*gelE* mutant (which contains functional *sprE*) and a *gelE* insertion mutant (which has a polar effect deactivating *sprE*) are attenuated.[21] Inouye and coworkers also confirmed this effect for the *gelE* insertion mutant in this strain in 2004.[22] In a different strain, V583, Perego and coworkers also showed in 2004 that a *gelE* insertion mutant had very attenuated biofilm formation but that a *sprE* insertion mutant did not.[23] However, in 2008, Hancock and coworkers showed that a Δ*sprE* mutant in strain V583 had increased biofilm (though the effect was not statistically significant) and that this mutant also

has an increased rate of autolysis.[20] In 2009, they followed up on this work with a model for how SprE serves as an immunity protein, protecting cells that produce GelE from autolysis by regulating the local levels of GelE.[9] The increased biofilm mass and rate of autolysis observed in Δ*sprE* mutant strains could be explained by this hypothesis. As they lack the immunity protein SprE, mutants in *sprE* should have increased amounts of GelE, which would lead to increased rates of autolysis and an increased release of extracellular DNA (eDNA). This would provide favorable conditions for biofilm formation, as eDNA is an important early biofilm matrix component in this organism.[9,20]

Table 3.4: Studies of *sprE*/SprE on *E. faecalis* biofilm formation.

Authors	**Year**	**Strain investigated**	***gelE* mutant**	***sprE* mutant**	***gelE/sprE* mutant**
S. Pillai, R. Inouye, et al.[22]	2004	OG1RF	–	–	limits biofilm
J. Mohamed, B. Murray, et al.[21]	2004	OG1RF	limits biofilm	no change	limits biofilm
L. Hancock, M. Perego[23]	2004	V583[a]	limits biofilm	no change	–
V. Thomas, L. Hancock, et al.[20]	2008	V583[a]	limits biofilm	increases biofilm[b]	limits biofilm

[a] The mutant strains used in these two studies are distinct.
[b] This effect was not statistically significant.

It is not entirely clear how phenotypes of these mutants in monoculture translate into their effects on virulence in vivo. In 1998, Murray and coworkers showed that a *gelE* insertion mutant in strain OG1RF had reduced virulence in a mouse peritonitis model,[24] and in 2000 they observed a similar result in this strain for mutants in the *fsr* regulatory genes, *fsrA* and *fsrB*, as well as *sprE*.[25] Similarly, A 2002 study by Calderwood and coworkers showed that an insertion mutant in *sprE* and a Δ*gelE* mutant in strain OG1RF are both attenuated in *C. elegans* killing.[26] In 2004, Gilmore and coworkers showed in a rabbit model of eye infection that individual mutants lacking functional GelE or SprE (Δ*gelE* and *sprE* insertion) showed basically the same

course of disease as wild type OG1RF, but that a *gelE* insertion mutant (which also lacks functional *sprE*) was attenuated in disease. This result suggests that these proteases may have a synergistic effect contributing to virulence.[27] Similarly, in 2008, Suzuki and coworkers showed in another rabbit eye infection model that a mutant lacking gelatinase and serine protease activity demonstrated limited pathogenesis in comparison with the wild type.[28] This work was conducted with strain OG1S and its protease negative mutant OG1X, which was generated by chemical mutagenesis,[29] and it is not clear which protease was primarily responsible for the observed effect.[28] In 2010, Hancock and coworkers demonstrated that the Δ*gelE* and Δ*gelE* Δ*sprE* mutants of strain V583 led to lower bacterial burdens than the wild type strain and the Δ*sprE* mutant in a rabbit endocarditis model, suggesting that GelE is the most important protease for this effect.[30] Overall, while the *fsr* proteases are clearly involved in virulence, it has remained unclear how exactly these proteases interact in the many in vivo systems that have been studied. Although there are clear virulence effects observed for *sprE* mutant strains in some models, the lack of a robust phenotype of these strains in monocultures makes it difficult to hypothesize how this effect arises.

S. aureus has many secreted proteases, making an understanding of the precise effects of glutamyl endopeptidase in this species even more complicated. These proteases include glutamyl endopeptidase SspA, the chymotrypsin family serine proteases SplABCDEF, the cysteine protease staphopains SspB and SplA, the cysteine protease ScpA, and the metalloprotease aureolysin.[31] SspA and SspB are co-transcribed and regulated by the accessory gene regulator (*agr*) operon, which is very similar to the *fsr* operon of *E. faecalis* in both architecture and function.[32] SspA appears to cleave and activate SspB,[10] and these proteases are involved in biofilm formation and other related phenotypes.[33] Though the composition and developmental

path of *S. aureus* biofilms can vary substantially among strains, there are two basic paradigms, one that is dependent on polysaccharide intercellular adhesin (PIA) and one that is not.[34] PIA-independent *S. aureus* biofilms are held together by eDNA, which may be released through autolysis.[34,35]

The literature suggests that SspA may have various roles in the life cycle of a PIA-independent *S. aureus* biofilm, mainly through its interaction with other important proteins.[36–38] A model of this interaction network has been proposed by Schneewind and coworkers.[38] Initiation of biofilm formation involves secretion of certain cell wall-associated surface binding proteins, including biofilm-associated protein (Bap) and fibronectin-binding protein (FnBP).[36] SspA may interfere with this stage of biofilm formation, as it can degrade both of these proteins.[36,37] After *S. aureus* binds to the surface, autolysin (Atl), a peptidoglycan hydrolase responsible for autolysis, degrades the cell walls of some *S. aureus* cells, releasing eDNA that forms a component of the biofilm.[39] Secreted proteases appear to be involved in regulation of Atl, and SspA can degrade Atl, which suggests that SspA may help effect the transition from a developing to a mature biofilm by degrading this enzyme.[38,39] Finally, SspA may also contribute to biofilm detachment through the degradation of biofilm matrix components.[38]

In studies that have clearly isolated the effects of *sspA*/SspA, there is very inconsistent data on whether this gene/protein has a positive or negative effect on biofilm formation (Table 3.5). A 2008 study with strain SH1000 by Horswill and coworkers showed that an *sspA* insertion mutant was defective in forming biofilms.[40] In 2010 Penadés and coworkers assessed the Bap biofilm model in the genetic background of strain SH1000, and in this system the Δ*sspA* mutant resulted in increased biofilm formation.[36] In 2013, Schneewind and coworkers showed that a *sspA* insertion mutant in the Newman strain was attenuated in biofilm formation as compared with the

wild type. Complementation of the *sspA* mutant restored the biofilm phenotype.[38] Where direct genetic results have not been obtained, addition of purified SspA to cultures has also provided contradictory results. In the same 2013 study by Schneewind and coworkers, addition of SspA to Newman strain cultures impaired biofilm formation.[38] Interestingly, this addition of SspA could not complement the *sspA* mutation.[38] In 2008, O'Gara and coworkers demonstrated that purified SspA has no effect on biofilm formation in strain RN4220.[41] In 2014 Smeltzer and coworkers showed that addition of purified SspA to the USA 300 LAC strain could limit biofilm formation, but that SspA had little effect on promoting biofilm dispersal.[42] Overall, though the *S. aureus* literature also implicates many other proteases in biofilm formation, SspA does appear to be involved in this phenomenon in many different strains and assay systems.

Table 3.5: Studies of *sspA*/SspA on *S. aureus* biofilm formation.

Authors	Year	Strain investigated	*sspA* mutant	SspA addition
E. O'Neill, J. O'Gara, et al. [41]	2008	RN4220	–	no effect
B. Boles, A. Horswill[40]	2008	SH1000	eliminates biofilm	–
M. Martí, J. Penadés, et al.[36]	2010	SH1000 + Bap	increases biofilm	–
C. Chen, O. Schneewind, et al.[38]	2013	Newman	impairs biofilm	impairs biofilm
A. Loughran, M. Smeltzer, et al.[42]	2014	USA300 (LAC)	–	impairs biofilm

The literature also provides a complex picture of SspA and virulence. SspA was first identified as a potential virulence factor in a 1998 study by Stover and coworkers, which identified transposon mutants in strain RN6390 that were attenuated in abscess, systemic intravenous, and burn wound infections in mice.[43] However, a 2001 study by McGavin and coworkers showed that an *sspA* insertion mutant in strain RN4220 was not impaired in a mouse tissue abscess model.[44] In 2009, von Eiff and coworkers showed that serine protease inhibitors inhibit *S. aureus* supernatant-induced production of IL-8 and IL-6 in human nasal epithelial cells,

though the effect was not further isolated to either SspA or SplABCDEF.[45] Overall, while proteases are clearly important for *S. aureus* virulence, it is unclear how much of this effect is due to SspA.[46]

3.2.4. Evaluating glutamyl endopeptidase genetic disruption in *E. faecalis* and *S. aureus* strains

Above, we demonstrated inhibition of SspA by ruminopeptin analogues in vitro, and we hypothesized that inhibition of glutamyl endopeptidases by ruminopeptin(s) may also occur in the gut environment. In order to test this hypothesis, we first desired to observe a phenotype of glutamyl endopeptidase inhibition in either *E. faecalis* or *S. aureus*. As discussed above, the various phenotypes that have been observed for genetic disruption of glutamyl endopeptidase activity in these organisms vary substantially among strains and assay conditions. If we could observe a phenotypic difference in strains due to glutamyl endopeptidase genetic disruption, we imagined that we might then be able to replicate that effect by administration of synthetic ruminopeptin(s), or even by co-culturing the organism with *R. bromii* itself. Ultimately, we envisioned that this work could lead to experiments to determine the ecological relevance of ruminopeptin production by *R. bromii* in mice or another model organism. Therefore, we attempted to demonstrate robust phenotypes of these proteases by examining the corresponding mutant strains. For the experiments in *E. faecalis*, we worked with Δ*sprE*, Δ*gelE* and Δ*gelE* Δ*sprE* mutants of strain V583, which are in-frame deletions that have been complemented to confirm that only the indicated proteases are disrupted.[20] These strains were a gift from Michael Gilmore's laboratory at Massachusetts Ear and Eye Infirmary. For the experiments in *S. aureus*,

we used strain JE2 and its *sspA* transposon mutant, which are from the NARSA collection now hosted at BEI Resources.[47]

First, we sought to determine if genetic disruption or inhibition of the glutamyl endopeptidase in these strains could be observed by monitoring protease activity in culture supernatants.[48] To measure secreted protease activity, *E. faecalis* strain V583 and mutants were grown aerobically overnight in brain-heart infusion (BHI) broth. *S. aureus* JE2 was grown aerobically in tryptic soy broth (TSB) to mid-log phase. The clarified cell-free supernatants were evaluated for secreted glutamyl endopeptidase activity (hydrolyzing Z-Leu-Leu-Glu-AMC, Figure 3.4A) by comparison with commercial SspA (at 1 ng/μL in the final assay mixture). We were never able to observe secreted glutamyl endopeptidase activity from these strains in these experiments, suggesting that this assay may not be sensitive enough to reveal the proteases or that they were not expressed under these conditions. We also attempted the assay with concentrated supernatants from the *E. faecalis* strains, but this was similarly unsuccessful. (In Chapter 4, we discuss a secreted protease activity assay using FTC-casein that was successfully used to monitor secreted protease activity in a different strain of *E. faecalis* under similar conditions).

A common method for observing secreted protease activity of microbial strains is by plating supernatants from those strains on milk-agar plates.[49] These opaque plates demonstrate a zone of clearance where proteases are present that can hydrolyze the main milk protein, casein. Though this assay does not reveal what specific proteases are responsible for proteolysis, it is a useful first pass reporter on the various protease activities present in wild type and mutant strains. From the literature, it was expected that *E. faecalis* strain V583 would exhibit a clear zone of proteolysis on milk agar, while the Δ*gelE* and Δ*gelE* Δ*sprE* mutants would lack a zone of

proteolysis and the Δ*sprE* mutant would have a smaller zone compared with the wild type.[20] We generated cell free supernatants from these strains and added them to wells punched in milk agar plates. We observed similar results in this assay to what was reported in the literature. However, this assay format proved to be unamenable to studying the effect of inhibitors on these phenotypes. We observed large inter-plate variability in the phenotypes, even without inhibitor present, and the subtle differences in zones of clearance between the different mutants were not easily quantifiable. It was known from the literature on the *S. aureus* strain JE2 transposon library that the *sspA* insertion mutant is not deficient in caseinolysis as compared with the wild type.[47]

As differences in autolytic activity are one of the most robust observed phenotypes of *fsr* operon mutants in *E. faecalis* strains, we next adapted previously published procedures to evaluate this phenotype.[50–52] We grew overnight cultures of *E. faecalis* in BHI medium and *S. aureus* in TSB medium. The strains were then re-inoculated 1:100 in TSB medium and grown to mid-log phase. In the case of *E. faecalis*, this second culture also contained 3% glycine (w/v). The cells were pelleted by centrifugation, washed, resuspended in buffer, and dispensed into a 96 well plate. The plates were incubated at 37 ºC in a microplate reader and OD600 measured every 30 minutes for 16 hours to generate autolysis curves. After extensive optimization of autolytic conditions, we reliably observed autolysis of these strains over the course of overnight incubations (Figure 3.7). In our hands, the inclusion of 0.01% (w/v) Triton-X 100 in the resuspension buffer was the key factor in being able to observe autolysis, though this additive is only reported in some publications on the phenomenon.[50–52] However, contrary to what was reported in the literature for these *E. faecalis* strains, we did not observe the expected increased rate of autolysis in the Δ*sprE* mutant and decreased rate in the Δ*gelE* mutants.[20] Rather, the

mutants lacking *gelE* did not show the autolysis phenotype, and the phenotype was indistinguishable between the wild type strain and Δ*sprE* mutant. Autolysis was observed in *S. aureus* JE2, but there is no apparent difference in the autolytic behavior of the wild type and the *sspA* transposon mutant. In general, mutation of *sspA* is known to alter the autolysin profile in *S. aureus*, but to our knowledge, this particular mutant strain has not previously been evaluated for its autolytic behavior.[44]

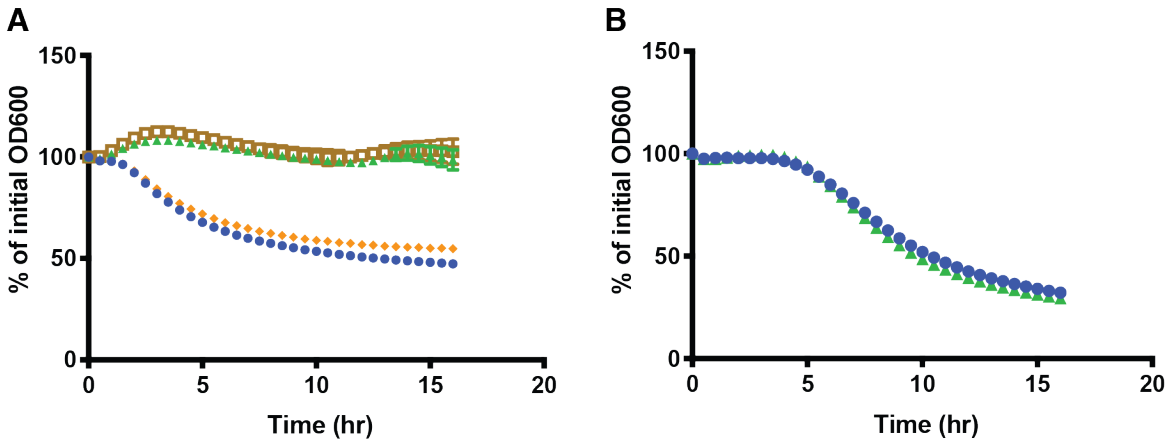


Figure 3.7: Single mutants of glutamyl endopeptidases do not exhibit differences in autolysis from the wild type strains in *E. faecalis* V583 and *S. aureus* JE2.

Autolysis experiments monitoring OD600 of bacterial suspensions over time. Cells were incubated at 37 ºC and OD600 measured every 30 minutes. Each point is the average of six replicates, individually referenced to the starting OD value of each well. Error bars, where observed, indicate standard deviations (in most cases the error bars are smaller than the points on the graph). (A) *E. faecalis* V583 (●) and extracellular protease mutants Δ*gelE* (▲), Δ*sprE* (◆), and Δ*gelE* Δ*sprE* (□). (B) *S. aureus* JE2 (●) and its transposon mutant in *sspA* (▲).

Finally, we analyzed the ability of different strains to form biofilms using crystal violet staining.[53,54] We grew overnight cultures of the strains in TSB, diluted them 1:40 in fresh medium, and inoculated them into 96-well plates. Biofilms were then allowed to form at 37 ºC for 24 – 48 h. After extensive optimization of growth and assay conditions (addition of salt and

glucose to the media, incubation for various time points, different plate washing and staining procedures), we could not observe differences between the wild type strains and their mutants in this assay. With *E. faecalis,* though we expected that the Δ*sprE* mutant might demonstrate increased biofilm formation and the Δ*gelE* mutant decreased biofilm formation, we observed little variation among the mutant strains (Figure 3.8A). These measured OD550 values are of the expected order of magnitude for *E. faecalis* biofilms. Though increasing the amount of glucose supplementation from 0.25% to 1.0% did increase biofilm formation in this strain, differences were still not observed between the various mutants (Figure 3.8B). In *S. aureus*, it was unclear what biofilm phenotype to expect from the *sspA* insertion mutant, and we could not observe a difference between the wild type and mutant strains in their capacity to form biofilms (Figure 3.8C). The OD550 values observed in this experiment are lower than expected for biofilms of this organism. A potential explanation for the lack of effect observed here is that we did not pre-treat these plates with human plasma, which can significantly increase the amount of biofilm formed by this organism.[55]

Overall, despite decades of work examining SspA in *S. aureus* and SprE in *E. faecalis*, disrupting these proteases does not lead to clear, conserved phenotypes in many different strains of these organisms. Regarding the specific case of *E. faecalis* biofilm formation, a recent book chapter by Shankar and coworkers offers a potential explanation for this phenomenon: "A cursory scan of 15–20 years of publications on enterococcal biofilms indicates that the number of different experimental conditions employed to measure biofilm formation is comparable to the number of published papers on the topic."[18] In various experiments designed to measure secreted protease activity, autolysis activity, and biofilm formation, we were unable to observe differences between strains deficient in these proteases and their wild type counterparts. This

does not necessarily indicate that these proteases are insignificant for these organisms. Rather, it may simply reflect the contrived nature of these assays attempting to measure physiological phenotypes in unnatural settings.

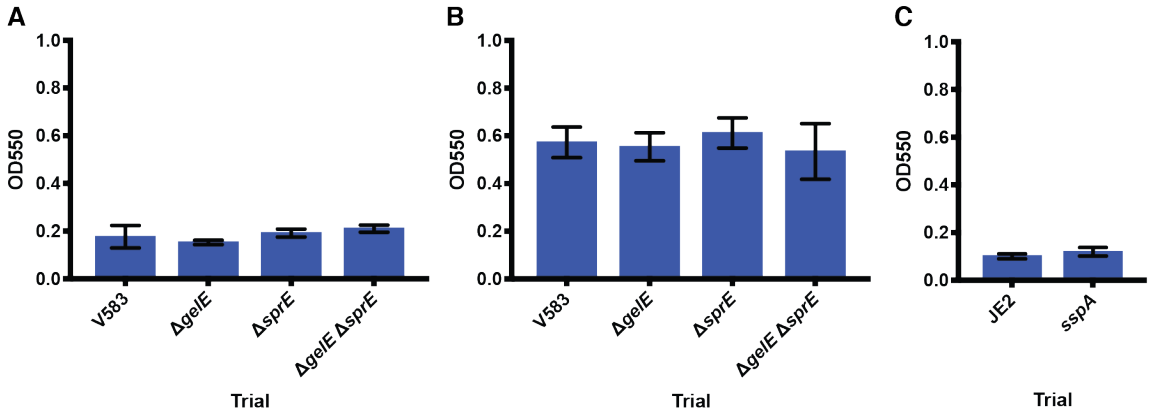


Figure 3.8: Secreted protease mutants do not exhibit differences in biofilm formation from the wild type strains in *E. faecalis* V583 and *S. aureus* JE2.

Optical density measurements of biofilm formation. Biofilm was stained with crystal violet and quantified by measurement of absorbance at 550 nm. The data for each condition is the mean of 8 replicates. Error bars represent standard deviation. (A) *E. faecalis* strains with 0.25% (w/v) glucose. (B) *E. faecalis* strains with 1.0% (w/v) glucose. (C) *S. aureus* strains with 0.25% (w/v) glucose.

A particularly notable discrepancy here was the clear increase in autolysis observed in published work for *E. faecalis* strain V583 Δ*sprE* that we were unable to replicate.[20] There are several possible reasons for this discrepancy. First, it is not clear if detergent was included in the assays in that publication, while this was required for us to observe autolysis. Therefore, the autolysis we observe may proceed through a distinct mechanism. It is also possible that our strains may have mutated or that there is some other required factor that we are missing in our assays. However, our interest was in the potential ecological roles of ruminopeptin production, eventually leading toward animal studies on the effects of small molecule production by *R.*

bromii, and we did not see how the increased autolysis of a particular strain of *E. faecalis* would be sufficiently interesting to advance the work in this direction, even if we could replicate that phenotype. Therefore, we decided to move on from investigating these proteases using biological assays to assessing their abundance and potential roles in the gut microbiota using bioinformatics.

3.2.5. Presence of glutamyl endopeptidases in the human gut microbiota

Though we could not observe phenotypic effects of glutamyl endopeptidase mutation in *S. aureus* and *E. faecalis*, precluding our efforts to phenocopy such an effect by inhibition with ruminopeptin(s), it is also possible that peptide aldehydes produced by *R. bromii* interact with related proteases found in other gut commensal microbes. In addition to the glutamyl endopeptidases in *S. aureus* and *E. faecalis*, these proteases have also been discovered in other *Staphylococcus*, *Bacillus*, and *Streptomyces* species. Many of these enzymes have been biochemically characterized, including glutamyl peptidase BL (from *Bacillus licheniformis*),[56] glutamyl peptidase BS (from *Bacillus subtilis*),[57] glutamyl peptidase BI (from *Bacillus intermedius*),[58] and glutamyl endopeptidase II (from *Streptomyces griseus*)[59], as well as SspA[7] and SprE.[8] These enzymes all belong to the structural chymotrypsin family,[60] and though they exhibit some differences in kinetic parameters and specificity, they all share a preference for cleavage after glutamyl residues (and would therefore likely be inhibited by a glutamyl aldehyde). To our knowledge, the presence and roles of glutamyl endopeptidases in the human gut microbiota have not previously been investigated. However, the diversity of previously biochemically characterized examples of these proteases provided a broad starting point for identifying additional potential targets of ruminopeptin in the human gut.

To explore whether additional glutamyl endopeptidases are present in the human gut microbiota, we used BLAST searches to locate members of this family in sequenced gut microbial genomes. Queries of the non-redundant (nr) protein database of NCBI with six representative glutamyl endopeptidase sequences resulted in hundreds of hits, from which we could identify homologs of these proteases in other residents of the human gut. Conserved residues Thr190 (or Ser190) and His213 (chymotrypsin numbering) in the S1 binding pocket of crystallized glutamyl endopeptidases have been identified as important for binding glutamate-containing substrates (Thr164 and His184 in SspA).[60] We were able to identify these residues in the BLAST hits from *Enterococcus faecium* (24.5% ID to SprE) and the pathogen *Listeria monocytogenes* (25.5% ID to glutamyl endopeptidase BL) (Figure 3.9). Additionally, we identified a sequence from *F. prausnitzii* that is annotated as a glutamyl endopeptidase precursor. This sequence maintains His213, but not Thr190, in the S1 site (Figure 3.9).

To what extent are these two conserved residues a good predictor of post-glutamyl hydrolyzing activity? The substrate specificity determinants of glutamyl endopeptidases were recently reviewed by Kostrov and coworkers.[60] In the chymotrypsin family of serine proteases, the major residues determining substrate specificity are found in a substrate binding pocket near the active site, as discussed in Chapter 1.[61] The first glutamyl endopeptidase structure determined in 1993 (in the presence of a tetrapeptide ligand) revealed several key residues that are important for binding the P1 glutamate residue: Ser190, His213, and Ser216 (Figure 3.10).[62] In this study, site directed mutagenesis confirmed the particular significance of Ser190 and His213 on conferring specificity.[62] In all subsequent crystal structures of this protease family, His213 and Thr/Ser190 have also been identified as important for selectivity.[60] In the subclass of glutamyl endopeptidases that matures by cleavage of a propeptide, the resulting free N-terminus also tends

to form a part of the active site and is involved with binding the P1 glutamate residue.[60] It should be noted that though we did not observe the expected hydrogen bonds with His213 and Ser190 in our docking experiment of ruminopeptin with SspA, we did observe this N-terminal hydrogen bonding interaction (Figure 3.5). Overall, these data suggest that presence of His213 and Ser/Thr 190 is a reasonable criterion for classifying proteases as glutamyl endopeptidase-like.

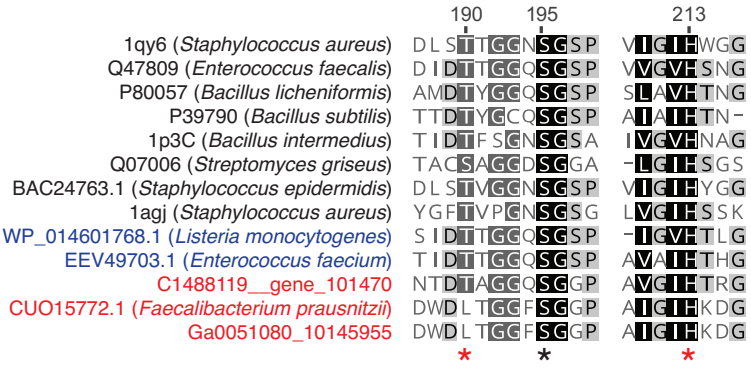


Figure 3.9: Glutamyl endopeptidase homologs are found in gut microbial genomes and human gut metagenomes.

ClustalW2 alignment of biochemically characterized glutamyl endopeptidases (black), homologs from sequenced organisms (blue), and homologs from human gut metagenomes (red). Included are the sequences of characterized glutamyl endopeptidases from *S. aureus* (1QY6),[13] *E. faecalis* (Q47809),[8] *B. licheniformis* (P80057),[57] *B. subtilis* (P39790),[63] *B. intermedius* (1P3C),[64] *S. griseus* (Q07006),[59] *Staphylococcus epidermidis* (BAC24763.1),[65] and epidermolytic toxin A from *S. aureus* (1AGJ),[66] with additional predicted glutamyl endopeptidases from *L. monocytogenes* (WP_014601768.1), *E. faecium* (EEV49703.1), and *F. prausnitzii* (CUO15772.1). Metagenomic sequences were retrieved using the BLAST tool at JGI Integrated Microbial Genomes & Microbiome Samples[67], as described below. Catalytic Ser195 is indicated with a black asterisk. Positions 190 and 213, which may be involved in conferring substrate specificity, are indicated with red asterisks.

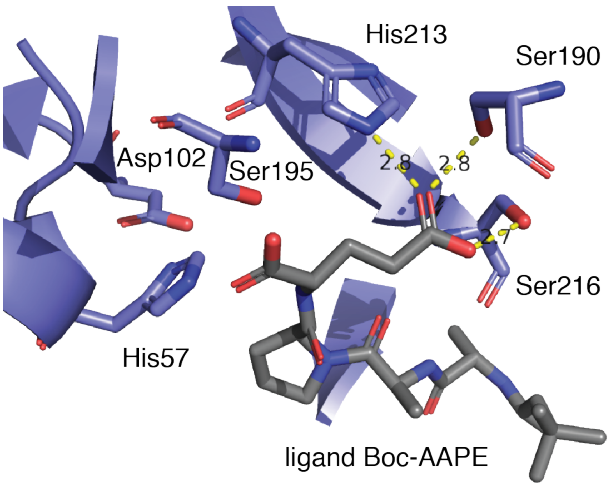


Figure 3.10: Structural determinants of glutamyl endopeptidase specificity.
Crystal structure of glutamyl endopeptidase from *S. griseus* (PDB: 1HPG, blue) with ligand Boc-AAPE (gray).[62]

The putative glutamyl endopeptidase from *F. prausnitzii*, which we identified bioinformatically, contains only His213. Position 190, where we would expect a serine or threonine residue, is occupied by a leucine residue. We used HHPred and Modeller to construct a homology model of the glutamyl endopeptidase from *F. prausnitzii* and confirmed that though the catalytic triad is well conserved, specificity conferring position 190 is indeed likely different (Figure 3.11). Aside from glutamyl endopeptidases, the only other characterized chymotrypsin fold proteases that contain histidine in position 213 are viral cysteine proteases that specifically cleave after glutamine residues.[60] Two examples of these include the human rhinovirus 3C protease and the tobacco etch virus (TEV) protease, which each recognize a very specific 7-residue cleavage motif.[68,69] Therefore, though the putative glutamyl endopeptidase from *F. prausnitzii* is missing one of the residues responsible for glutamate specificity in all characterized glutamyl endopeptidases, it is at least possible that it possesses either the same or a similar cleavage specificity. A secreted serine protease (or perhaps multiple secreted serine

proteases) from *F. prausnitzii* was recently reported to have the interesting phenotype of suppressing dorsal root ganglion (DRG) neuron excitability,[70] and an attractive hypothesis is that this uncharacterized protease could be responsible for this effect.

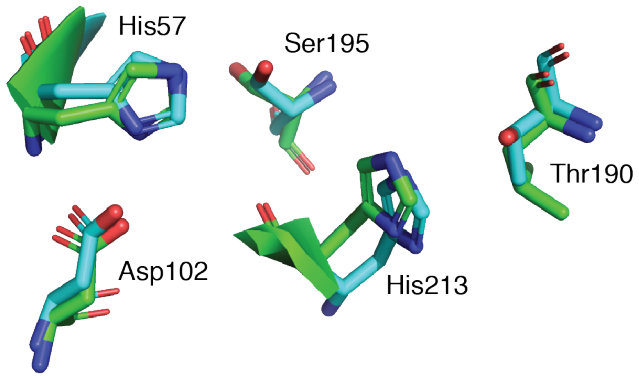


Figure 3.11: Homology model of glutamyl endopeptidase from *F. prausnitzii*.
F. prausnitzii serine protease (green) modelled using *B. intermedius* glutamyl endopeptidase as a template (PDB: 1P3C, blue). Chymotrypsin numbering.

In order to assess the presence of these bioinformatically identified proteases in human subjects and determine the distribution of glutamyl endopeptidases among unsequenced members of the gut microbiota, we also performed a BLAST search of representative glutamyl endopeptidase sequences against assembled stool metagenomes available through the Joint Genome Institute (JGI) (268 metagenomes) (Figure 3.12). After limiting the results based on an e-value cutoff (2e-10), length (188–400 residues to account for the diversity among characterized members of this protease family), and the presence of a candidate His213 residue in the S1 binding pocket, we identified 52 glutamyl endopeptidase homologs in 51 different samples. 47 of these sequences have ≥99% amino acid sequence ID to the putative glutamyl endopeptidase from *F. prausnitzii*. The remaining sequences do not map to sequenced genomes. This analysis suggests that these putative targets of the ruminopeptins may be present in the human gut.

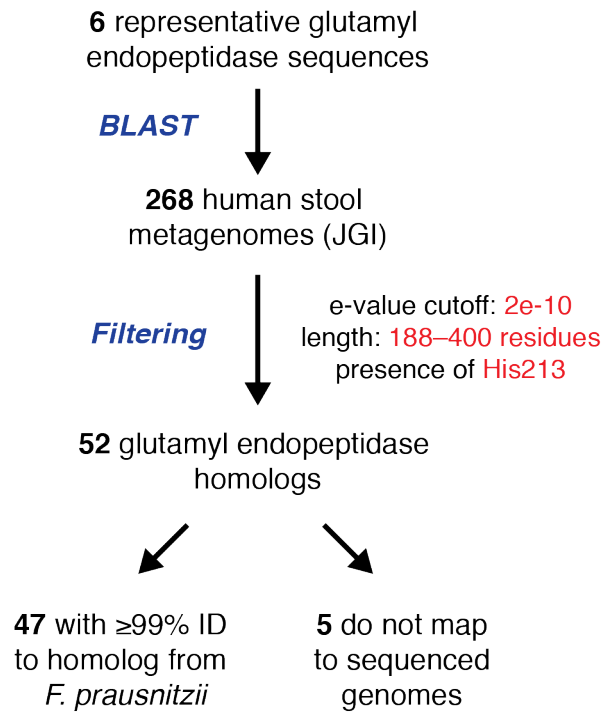


Figure 3.12: Searching for glutamyl endopeptidase homologs in human gut microbiota metagenomes.

Workflow for discovery of glutamyl endopeptidase homologs in the human gut microbiota. Representative sequences were used as a query in protein BLAST searches of the 268 human stool metagenomes at the JGI. After filtering based on three exclusion criteria, this search identified 52 glutamyl endopeptidase homologs (in 51 samples), 47 of which appear to be the glutamyl endopeptidase homolog from *F. prausnitzii*.

In other habitats, several physiological functions have been associated with glutamyl endopeptidases. In addition to SspA, the epidermolytic toxins of *S. aureus* also have glutamyl endopeptidase activity, and these are responsible for causing staphylococcal scaled skin syndrome (SSSS) on human skin.[31,66,71] A protease from *Staphylococcus epidermidis* (Esp), a member of the commensal nasal and skin microbiota, also has glutamyl endopeptidase activity.[65] This protease appears to inhibit *S. aureus* nasal colonization by cleaving *S. aureus* autolysin,

preventing *S. aureus* from releasing the eDNA that helps to form its early biofilm matrix.[38,72] Glutamyl endopeptidase activity was also found in *Bacillus pumilius* strains isolated from human endodontic lesions, and this protease is suggested to be a potential virulence factor for this oral pathogen.[73] Some viruses also produce glutamyl endopeptidases that are involved in proteolytic processing of viral precursor proteins.[74] Overall, though SspA and SprE are the best studied glutamyl endopeptidases from a human physiological standpoint, there are many more examples of proteases in this family which could also be potential targets of ruminopeptin-type compounds.

3.3. Conclusions

In this chapter, we synthesized a set of ruminopeptin analogues, encompassing many potential products of the *rup* gene cluster from *R. bromii*, and demonstrated that they inhibit a bacterial protease implicated in virulence in several human pathogens. Toward the goal of demonstrating the ecological relevance of this inhibitory interaction, we conducted a survey of potential phenotypes of *E. faecalis* and *S. aureus* glutamyl endopeptidase mutants but were unable to observe any significant phenotypes. Further study is needed to determine the precise roles of these proteases in basic microbial physiology and in their interactions with other species in the complex human gut environment.

Additionally, we found that homologs of glutamyl endopeptidases are present in commensal gut organisms and in human gut metagenomes, including a putative glutamyl endopeptidase from the abundant gut organism *F. prausnitzii*. It remains to be determined if these putative glutamyl endopeptidases from prominent gut commensals and human pathogens actually exhibit post-glutamyl hydrolyzing activity. Overall, these proteins may not only represent ecologically

relevant targets of the ruminopeptins but also provide a promising direction for investigating the biological roles of microbial proteases in the human gut.

Our work in this chapter has uncovered the first evidence that gut microbial natural products may be capable of modulating microbial protease activity. In Chapter 4, we discuss additional efforts to identify targets for the compounds synthesized in this chapter, including human targets. However, aside from the restriction of glutamyl endopeptidase activity to bacteria, the biogeography of *R. bromii* in the human gut and the potential instability of the ruminopeptins also suggest these natural products likely have a microbial target. In a study of gut microbes associated with insoluble, undigested polysaccharide particles in fecal samples, *R. bromii* was one of the three most enriched species in this phase as opposed to the soluble phase.[75] This observation may indicate that *R. bromii* is located distantly from host cells in comparison to other gut species. Moreover, a potential explanation for our inability to identify putative *rup* gene cluster products in *R. bromii* cultures in Chapter 2 could be the instability of these peptide aldehydes. As mentioned in Chapter 2, incubation of peptide aldehyde **3.7h** in an *R. bromii* culture resulted in almost complete degradation in just 15 min when incubated at 37 ºC. Overall, given their instability and the localization of *R. bromii*, we hypothesize that peptide aldehydes produced by this organism have evolved to target other microbial species living in close proximity.

The ecological details of the ruminopeptin-glutamyl endopeptidase interaction remain to be determined, as do the broader roles of gut microbial protease inhibitors and gut microbial proteases. Through the study of one particular family of proteases in this chapter, we became inspired to look more broadly at the potential biological roles of proteases and peptide aldehydes in the gut environment. In Chapter 4, we discuss our expanded approach for synthesizing

additional predicted compounds of other NRPS gene clusters from the commensal gut microbiota. With this larger and more diverse library, we were able to perform an expanded investigation of potentially physiologically relevant phenotypes of these compounds on human enzymes and gut microbial species.

3.4. Materials and methods

3.4.1. General materials and methods

Optical densities of bacterial cultures were determined with a DU 730 Life Sciences UV/Vis spectrophotometer (Beckman Coulter) or a GENESYS™ 20 Visible Spectrophotomer (Thermo Scientific™) by measuring absorbance at 600 nm. All chemicals were obtained from Sigma-Aldrich except where noted. Protected amino acids were obtained from Chem-Impex (Dale, IL) and Advanced ChemTech (Louisville, KY). HATU was purchased from Oakwood Chemical (Estill, SC). All NMR solvents were purchased from Cambridge Isotope Laboratories (Andover, MA). NMR spectra were visualized using iNMR version 5.5.7. and MestReNova version 12.0. 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, D_2O = 4.79 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, br s = broad singlet, d = doublet, t = triplet, q = quartet, qt = quintet, m = multiplet), coupling constant, and integration. High-resolution mass spectral data was obtained in the Small Molecule Mass Spectrometry Facility, FAS Division of Science. HRMS data for synthetic compounds was obtained on an Agilent Technologies 6210 TOF coupled to an Agilent Technologies 1200 series LC. Liquid chromatography was performed with water/acetonitrile (1:1). The capillary voltage was 3.5 kV, the fragmentor voltage was 175

V, the drying gas temperature was 325 °C, the drying gas flow rate was 8 L/min, and the nebulizer pressure was 40 psig.

3.4.2. Synthesis of *N*-acyl amino acids

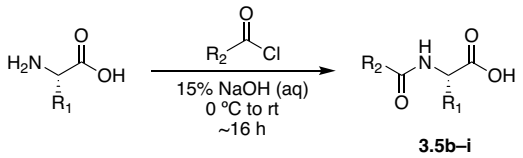
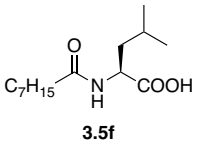


Figure 3.13: Synthesis of *N*-acyl amino acids.

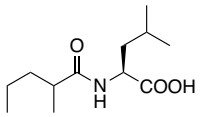
N-acetyl-L-Leucine (**3.5a**) was purchased from Chem-Impex. Other *N*-acyl amino acids were synthesized using the Schotten–Bauman reaction of acyl chlorides with amino acids in aqueous base. The amino acid (1.0 equiv) was dissolved in 15% aqueous NaOH (0.5 M) and cooled to 0 ºC. The acid chloride was added dropwise and the reaction mixture stirred overnight, allowing to warm to room temperature. 20% aqueous HCl was added to pH = 2 and the resulting solution was extracted with dichloromethane (three portions of 3x reaction volume). The combined organic layers were washed with saturated aqueous sodium chloride (one portion of 1x reaction volume). The solution was then dried over Na_2SO_4, filtered, and concentrated in vacuo. The characterization data for compounds **3.5b**–**3.5e** and **3.5g** matched previously reported results.[76,77]



3.4.2.1. Octanoyl-L-leucine (**3.5f**):

The product (701 mg, 54%) was isolated as a colorless solid (m.p. 121–123 ºC). ^{1}H NMR (500 MHz; $CDCl_3$): δ 10.63 (br s, 1H), 6.12 (d, J = 8.1 Hz, 1H), 4.62 (m, 1H), 2.24 (t, J = 7.6

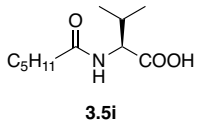
Hz, 2H), 1.71 (m, 2H), 1.63 (m, 2H), 1.60 (m, 1H), 1.28 (m, 8H), 0.94 (m**,** 6H), 0.87 (t, *J* = 7.0 Hz, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 176.8, 174.4, 51.1, 41.5, 36.7, 31.9, 29.4, 29.2, 25.9, 25.1, 23.1, 22.8, 21.1, 14.3. HRMS (ESI): Calc'd for formula $C_{14}H_{26}NO_3^-$ $[M–H]^-$ 256.1918, found 256.1927.



3.5h

3.4.2.2. (2-Methylpentanoyl)-L-leucine (**3.5h**):

The product (710 mg, 62%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 10.69 (br s, 1H), 6.10 (t, *J* = 9.1 Hz, 1H), 4.63 (m, 1H), 2.27 (q, *J* = 7.0 Hz, 1H), 1.70 (m, 2H), 1.61 (m, 2H), 1.36–1.29 (m, 4H), 1.16 (m, 1H), 1.13 (m, 3H), 0.94 (t, *J* = 8.2 Hz, 6H), 0.89 (m, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 177.6, 177.3, 50.9, 41.4, 36.5, 35.9, 25.2, 23.0, 22.1, 20.7, 17.9, 14.2. HRMS (ESI): Calc'd for formula $C_{12}H_{22}NO_3^-$ $[M–H]^-$ 228.1605, found 228.1614.



3.5i

3.4.2.3. Hexanoyl-L-valine (**3.5i**)

The product (2.0 g, 93%) was isolated as a colorless solid (m.p. 121–123 ºC). ^{1}H NMR (500 MHz; $CDCl_3$): δ 11.37 (s, 1H), 6.31 (d, *J* = 8.7 Hz, 1H), 4.59 (dd, *J* = 8.7, 4.7 Hz, 1H), 2.26 (t, *J* = 7.6 Hz, 2H), 2.22 (m, 1H), 1.63 (qt, *J* = 7.4 Hz, 2H), 1.30 (m, 4H), 0.95 (m, 6H), 0.87 (t, *J* = 6.87, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 175.5, 174.6, 57.2, 36.8, 31.5, 25.7, 22.5, 19.2, 17.9 14.1. HRMS (ESI): Calc'd for formula $C_{11}H_{20}NO_3^-$ $[M–H]^-$ 214.1499, found 214.1453.

3.4.3. Coupling of *N*-acyl amino acids to semicarbazone-protected aldehydes

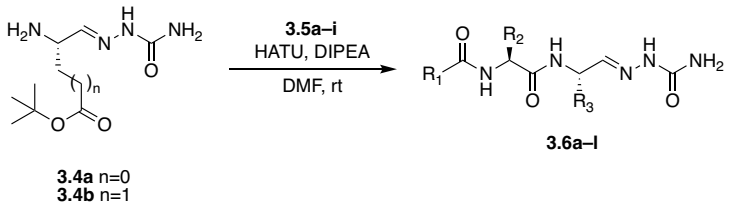
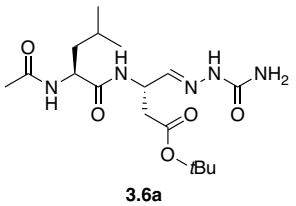


Figure 3.14: Synthesis of semicarbazone-protected aldehydes 3.6a–l.

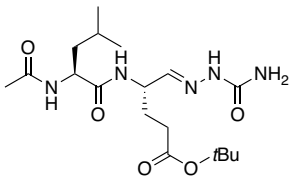
To a solution of semicarbazone-protected intermediate **3.4a** or **3.4b** (1.0 equiv) and acyl L-leucine **3.5a–i** (1.0 equiv) in DMF (0.6 M) was added HATU (1.1 equiv) and DIPEA (5.1 equiv) with stirring, under argon. After 3 h, the reaction mixture was diluted with ethyl acetate (to 10 x initial volume) and quenched by addition of 1M aqueous NaOH (10 x initial volume). The organic layer was collected, and the aqueous layer extracted with three portions of ethyl acetate (each 10x initial reaction volume). The combined organic layers were washed with water and brine (each 20x initial reaction volume), dried over Na_2SO_4, filtered, and concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator. Products were purified by flash chromatography on silica gel using dichloromethane/methanol (9:1).



3.4.3.1. *tert*-Butyl (*S*,*E*)-3-((*S*)-2-acetamido-4-methylpentanamido)-4-(2-carbamoylhydrazineylidene)butanoate (**3.6a**):

The product (68 mg, 31%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; DMSO-d_6): δ 10.00 (s, 1H), 8.18 (d, J = 8.2 Hz, 1H), 7.99 (d, J = 8.2 Hz, 1H), 7.08 (d, J = 3.2 Hz, 1H), 6.25 (br s, 2H), 4.69 (m, 1H), 4.27 (q, J = 7.7 Hz, 1H), 2.65 (m, 1H), 2.49 (m, 1H), 1.82 (s, 3H), 1.56

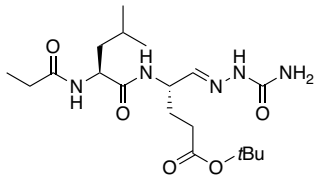
(m, 2H), 1.40 (m, 2H), 1.37 (m, 9H), 0.86 (m, 6H). ^{13}C NMR (126 MHz, DMSO-d_6): δ 171.9, 169.6, 169.1, 156.6, 140.3, 80.1, 51.0, 47.1, 41.1, 38.0, 27.6, 24.2, 22.9, 22.5, 21.7. HRMS (ESI): Calc'd for formula $C_{17}H_{31}N_5O_5Na^+$ $[M+Na]^+$ 408.2217, found 408.2226.



3.6b

3.4.3.2. *tert*-Butyl (*S*,*E*)-4-((*S*)-2-acetamido-4-methylpentanamido)-5-(2-carbamoylhydrazineylidene)pentanoate (**3.6b**):

The product (132 mg, 28%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; DMSO-d_6): δ 9.90 (s, 1H), 8.04 (d, J = 8.3 Hz, 1H), 7.96 (d, J = 8.1 Hz, 1H), 7.04 (d, J = 4.0 Hz, 1H), 6.25 (br s, 2H), 4.36 (m, 1H), 4.27 (m, 1H), 3.30 (s, 3H), 2.48 (m, 2H), 2.17 (m, 2H), 1.82 (s, 3H), 1.66 (m, 2H), 1.55 (m, 2H), 1.39 (m, 1H), 1.37 (s, 9H), 0.84 (m, 6H). ^{13}C NMR (126 MHz, DMSO-d_6): δ 171.9, 171.8, 169.1, 156.7, 141.3, 79.6, 51.1, 48.9, 41.0, 30.8, 27.7, 27.4, 24.2, 22.9, 22.5, 21.7. HRMS (ESI): Calc'd for formula $C_{18}H_{33}N_5O_5Na^+$ $[M+Na]^+$ 422.2374, found 422.2383.

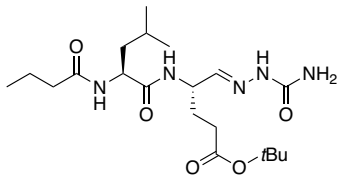


3.6c

3.4.3.3. *tert*-Butyl (*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((*S*)-4-methyl-2-propionamidopentanamido)pentanoate (**3.6c**):

The product (131 mg, 65%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.69 (m, 1H), 7.78 (m, 1H), 6.57 (m, 1H), 5.45 (br s, 2H), 4.55 (m, 1H), 4.50 (m, 1H), 2.25 (m,

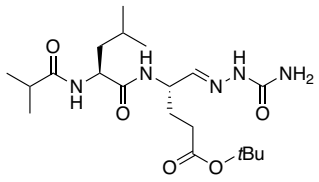
4H), 2.04 (m, 1H), 1.85 (m, 1H), 1.64 (m, 2H), 1.53 (m, 1H), 1.41 (m, 9H), 1.12 (m, 3H), 0.90 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 175.0, 172.7, 158.3, 80.9, 52.3, 50.2, 40.9, 31.3, 29.6, 28.3, 27.7, 25.1, 23.1, 22.9, 22.2, 10.0. HRMS (ESI): Calc'd for formula $C_{19}H_{34}N_5O_5^-$ $[M-H]^-$ 412.2565, found 412.2588.



3.6d

3.4.3.4. *tert*-Butyl (*S*,*E*)-4-((*S*)-2-butyramido-4-methylpentanamido)-5-(2-carbamoylhydrazineylidene)pentanoate (**3.6d**):

The product (160 mg, 63%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.68 (s, 1H), 7.78 (d, J = 7.0 Hz, 1H), 6.57 (br s, 1H), 4.54 (m, 1H), 4.50 (t, J = 7.0 Hz, 1H), 2.24 (m, 4H), 2.04 (m, 2H), 1.84 (m, 2H), 1.64 (m, 2H), 1.54 (m, 1H), 1.41 (m, 9H), 0.95 (m, 6H), 0.87 (m, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 174.3, 172.7, 170.6, 158.3, 80.9, 52.2, 50.2, 40.8, 38.5, 31.3, 28.3, 27.8, 25.0, 23.1, 22.2, 19.3, 13.9. HRMS (ESI): Calc'd for formula $C_{20}H_{37}N_5O_5Na^+$ $[M+Na]^+$ 450.2687, found 450.2675.

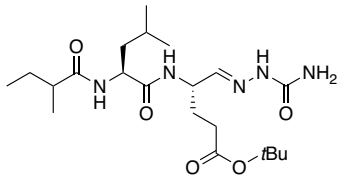


3.6e

3.4.3.5. *tert*-Butyl (*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((*S*)-2-isobutyramido-4-methylpentanamido)pentanoate (**3.6e**):

The product (124 mg, 48%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.71 (m, 1H), 7.81 (m, 1H), 7.21 (s, 2H), 6.47 (m, 1H), 4.56 (m, 1H), 4.51 (m, 1H), 2.43 (m,

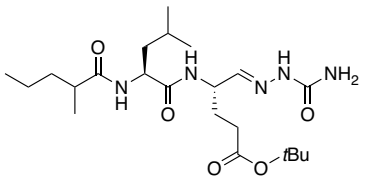
1H), 2.21 (m, 2H), 2.06 (m, 2H), 1.84 (m, 2H), 1.48 (m, 1H), 1.43 (m, 4H), 1.40 (s, 9H), 1.13 (m, 6H), 0.92 (d, J = 6.4 Hz, 3H), 0.88 (d, J = 6.3 Hz, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 178.2, 172.65, 158.3, 80.8, 54.7, 53.7, 51.9, 51.0, 50.1, 40.6, 35.5, 31.1, 28.3, 25.1, 23.1, 22.2, 19.9, 19.6, 18.8, 17.6. HRMS (ESI): Calc'd for formula $C_{20}H_{37}N_5O_5Na^+$ $[M+Na]^+$ 450.2687, found 450.2712.



3.6f

3.4.3.6. *tert*-Butyl (4*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((2*S*)-4-methyl-2-(2-methylbutanamido)pentanamido)pentanoate (**3.6f**):

The product (155 mg, 62%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.76 (s, 1H), 7.90 (m, 1H), 7.2 (m, 1H), 6.60 (m, 1H), 4.56 (m, 2H), 2.32 (m, 1H), 2.20 (m, 2H), 2.03 (m, 2H), 1.82 (m, 2H), 1.62 (m, 2H), 1.42 (m, 1H), 1.39 (m, 9H), 1.10 (m, 3H), 0.91 (m, 3H), 0.87 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 177.6, 172.6, 158.3, 142.2, 128.7, 80.8, 55.0, 51.93, 51.89, 42.9, 40.5, 31.2, 28.4, 27.5, 25.1, 23.1, 22.1, 18.8, 17.81, 17.66, 12.1. HRMS (ESI): Calc'd for formula $C_{21}H_{39}N_5O_5Na^+$ $[M+Na]^+$ 464.2843, found 464.2845.

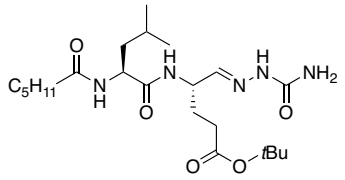


3.6g

3.4.3.7. *tert*-Butyl (4*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((2*R*)-4-methyl-2-(2-methylpentanamido)pentanamido)pentanoate (**3.6g**):

The product (217 mg, 87%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.76 (s, 1H), 7.84 (t, J = 7.9 Hz, 1H), 7.21 (s, 1H), 6.51 (br s, 1H), 4.55 (m, 2H), 2.28 (m, 1H)

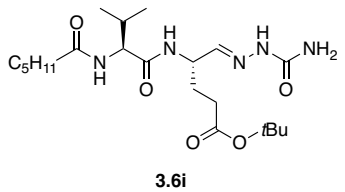
2.22 (m, 2H), 2.05 (m, 2H), 1.83 (m, 2H), 1.47 (m, 1H), 1.40 (m, 9H), 1.28 (m, 4H), 1.09 (m, 3H), 0.93 (m, 3H), 0.88 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 177.9, 172.6, 158.3, 80.8, 77.5, 77.2, 77.0, 54.9, 51.9, 41.2, 40.4, 36.6, 31.19, 31.07, 25.0, 23.1, 22.16, 22.07, 20.8, 18.8, 18.17, 18.01, 17.5, 14.2. HRMS (ESI): Calc'd for formula $C_{22}H_{41}N_5O_5Na^+$ $[M+Na]^+$ 478.3000, found 478.3016.



3.6h

3.4.3.8. *tert*-Butyl (*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((*S*)-2-hexanamido-4-methylpentanamido)pentanoate (**3.6h**):

The product (189 mg, 88%) was isolated as a colorless oil. ^{1}H NMR (399 MHz; $CDCl_3$): δ 9.94 (s, 1H), 7.94 (d, J = 6.8 Hz, 1H), 7.23 (s, 1H), 6.84 (d, J = 8.1 Hz, 1H), 4.54 (br s, 2H), 2.23–2.14 (m, 4H), 2.00 (m, 2H), 1.80 (m, 2H) 1.58 (m, 5H), 1.39 (s, 9H), 1.27 (m, 4H), 0.89 (m, 3H), 0.85 (s, 6H). ^{13}C NMR (100 MHz; $CDCl_3$): δ 174.2, 172.78, 172.58, 158.4, 142.2, 80.7, 52.0, 50.1, 46.5, 41.0, 36.5, 31.3, 28.3, 28.0, 25.56, 25.50, 25.0, 23.1, 22.5, 22.2, 14.1, 8.9. HRMS (ESI): Calc'd for formula $C_{22}H_{42}N_5O_5^+$ $[M+H]^+$ 456.318, found 456.3197.

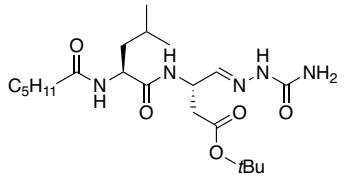


3.6i

3.4.3.9. *tert*-Butyl (*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((*S*)-2-hexanamido-3-methylbutanamido)pentanoate (**3.6i**):

The product (204 mg, 67%) was isolated as a yellow glass. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.76 (m, 1H), 7.73 (m, 1H), 7.33 (s, 1H), 7.18 (m, 2H), 6.23 (d, J = 6.3 Hz, 2H), 4.54 (m, 1H),

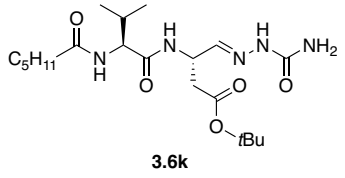
3.69 (m, 1H), 3.20–3.13 (m, 1H), 2.23 (m, 2H), 2.06 (m, 2H), 1.86 (m, 2H), 1.63 (m, 4H), 1.41 (m, 9H), 1.30 (m, 4H), 0.94 (m, 6H), 0.87 (m, 3H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 175.2, 174.0, 128.8, 57.35, 57.33, 55.2, 43.3, 36.8, 31.59, 31.54, 28.2, 25.66, 25.62, 22.57, 22.53, 19.3, 18.8, 18.0, 17.4, 14.1, 12.6. HRMS (ESI): Calc'd for formula $C_{21}H_{40}N_5O_5^+$ $[M+H]^+$ 442.3024, found 442.3041.



3.6j

3.4.3.10. *tert*-Butyl (*S*,*E*)-4-(2-carbamoylhydrazineylidene)-3-((*S*)-2-hexanamido-4-methylpentanamido)butanoate (**3.6j**):

The product (104 mg, 39%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.66 (s, 1H), 7.76 (d, J = 7.6 Hz, 1H), 6.58 (d, J = 8.1 Hz, 1H), 4.85 (m, J = 2.8 Hz, 1H), 4.49 (m, 1H), 2.70 (m, 1H), 2.58 (m, 1H), 2.20 (m, 3H), 1.60 (m, 2H), 1.40 (m, 9H), 1.28 (m, 4H), 0.88 (m, 9H). ^{13}C NMR (125 MHz, $CDCl_3$): δ 174.2, 172.6, 170.3, 158.1, 141.6, 81.6, 55.2, 52.0, 47.8, 43.2, 41.0, 38.1, 36.6, 31.6, 28.2, 25.5, 25.1, 23.2, 22.6, 22.1, 18.8, 17.5, 14.2, 12.6. HRMS (ESI): Calc'd for formula $C_{21}H_{39}N_5O_5Na^+$ $[M+Na]^+$ 464.2843, found 464.2845.

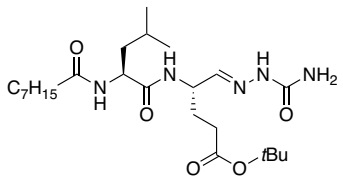


3.6k

3.4.3.11. *tert*-Butyl (*S*,*E*)-4-(2-carbamoylhydrazineylidene)-3-((*S*)-2-hexanamido-3-methylbutanamido)butanoate (**3.6k**):

The product (183 mg, 88%) was isolated as a pale yellow solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.72 (s, 1H), 7.77 (d, J = 7.7 Hz, 1H), 7.28 (s, 1H), 6.76 (m, 1H), 4.86 (m, 1H), 4.30 (m, 1H),

2.70 (m, 1H), 2.60 (m, 1H), 2.22 (m, 2H), 2.04 (m, 1H), 1.61 (m, 2H), 1.41 (m, 9H), 1.29 (m, 4H), 0.91 (m, 9H). ^{13}C NMR (125 MHz, $CDCl_3$): δ 174.1, 171.8, 170.4, 158.1, 141.6, 81.7, 58.9, 55.0, 47.8, 36.7, 31.6, 31.0, 28.3, 25.6, 22.6, 19.6, 18.7, 17.5, 14.2, 12.5. HRMS (ESI): Calc'd for formula $C_{20}H_{37}N_5O_5Na^+$ $[M+Na]^+$ 450.2687, found 450.2679.

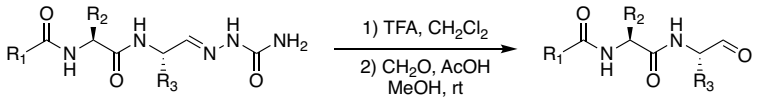


3.6l

3.4.3.12. *tert*-Butyl (*S*,*E*)-5-(2-carbamoylhydrazineylidene)-4-((*S*)-4-methyl-2-octanamidopentanamido)pentanoate (**3.6l**):

The product (59 mg, 43%) was isolated as a colorless solid. ^{1}H NMR (500 MHz, $CDCl_3$) δ 9.74 (s, 1H), 7.78 (d, *J* = 7.5 Hz, 1H), 7.19 (m, 1H), 6.55 (d, *J* = 8.2 Hz, 1H), 4.53 (m, 2H), 2.21 (m, 4H), 2.03 (m, 1H), 1.84 (m, 1H), 1.58 (m, 6H), 1.40 (s, 9H), 1.26 (m, 8H), 0.86 (m, 9H). ^{13}C NMR (125 MHz, $CDCl_3$): δ 174.4, 172.7, 158.1, 80.9, 77.5, 77.2, 77.0, 54.9, 52.2, 50.3, 43.0, 40.7, 36.7, 31.9, 31.3, 29.4, 29.2, 28.3, 27.7, 25.9, 25.1, 23.2, 22.8, 22.1, 18.9, 17.6, 14.3, 12.5. HRMS (ESI): Calc'd for formula $C_{24}H_{44}N_5O_5^-$ $[M-H]^-$ 482.3348, found 482.3367.

3.4.4. Removal of *tert*-butyl protecting groups and regeneration of aldehydes



3.6a–l **3.7a–l**

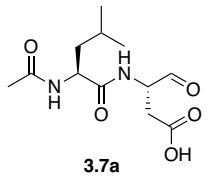
Figure 3.15: Synthesis of peptide aldehydes 3.7a–l.

The semicarbazone-protected N-acyl dipeptide 3.6a–l (1.0 equiv) was stirred in 20% trifluoroacetic acid in dichloromethane (0.02 M) under argon, with immediate addition of water

(3.0 equiv). The reaction mixtures were stirred for 1 h and concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator, and residual TFA was removed by forming the azeotrope with anhydrous toluene.

The deprotected semicarbazone intermediate (1.0 equiv) was dissolved in methanol/37% formaldehyde/acetic acid (5:1:1, 16 mM) and stirred for 30 min at room temperature. Water was added to the reaction mixture (to 2x initial reaction volume), and the reaction mixture was concentrated in vacuo to remove methanol. The reaction mixture was then diluted with water (1x initial reaction volume) and extracted with three portions of ethyl acetate (each 1x initial reaction volume). The combined organic layers were washed with two portions of water and one portion of brine (each 1x initial reaction volume). The combined organic layers were dried over Na_2SO_4, filtered, and concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator. The resulting products were further purified by trituration with two 2 mL volumes of diethyl ether.

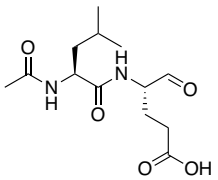
For several of the more hydrophilic compounds (**3.7a**–**e**), the final reaction step diverged from the above procedure. The deprotected semicarbazone intermediate (1.0 equiv) was dissolved in methanol/37% formaldehyde/acetic acid (5:1:1, 16 mM) and stirred for 30 min at room temperature. Water was added to the reaction mixture (to 2x initial reaction volume), the reaction mixture was concentrated in vacuo to remove methanol, and water was then removed by lyophilization. The resulting solid was re-dissolved in water (1x initial reaction volume) and filtered to remove insoluble particulates. The solution was lyophilized again, and the resulting products were further purified by trituration with two 2 mL volumes of diethyl ether.



3.7a

3.4.4.1. (*S*)-3-((*S*)-2-Acetamido-4-methylpentanamido)-4-oxobutanoic acid (**3.7a**):

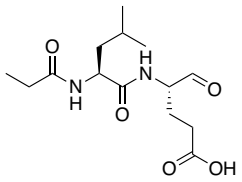
The product (24 mg, 66%) was isolated as a colorless solid. ^{1}H NMR (500 MHz, 1:1 CD_3OD/D_2O, referenced to D_2O): δ 4.92 (m, 1H), 4.59 (m, 2H), 4.24 (m, 1H), 3.39 (s, 1H), 3.31 (s, 1H), 2.60 (m, 1H), 2.50 (m, 1H), 2.00 (s, 3H), 1.60 (m, 2H), 1.37 (m, 1H), 0.92 (m, 6H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 174.9, 172.6, 169.1, 102.5, 87.6, 85.3, 81.9, 50.8, 40.9, 24.2, 24.2, 22.9, 22.5, 21.6. HRMS (ESI): Calc'd for $C_{12}H_{19}N_2O_5^-$ $[M–H]^-$ 271.1299, found 271.1298.



3.7b

3.4.4.2. (*S*)-4-((*S*)-2-Acetamido-4-methylpentanamido)-5-oxopentanoic acid (**3.7b**):

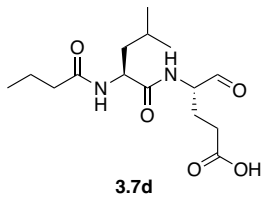
The product (23 mg, 60%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; DMSO-d_6): δ 9.33 (s, 1H), 8.06 (m, 1H), 7.36 (m, 1H), 4.32 (m, 1H), 3.97 (m, 1H), 2.09 (m, 2H), 1.69 (m, 2H), 1.59 (m, 1H), 1.46 (m, 2H), 1.23 (m, 3H), 0.84 (m, 6H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 200.7, 178.8, 173.9, 57.4, 50.8, 45.8, 29.6, 24.2, 22.9, 22.5, 21.6, 12.1, 8.6. HRMS (ESI): Calc'd for formula $C_{15}H_{25}N_2O_5^-$ $[M–H]^-$ 285.1456, found 285.1454.



3.7c

3.4.4.3. (*S*)-4-((*S*)-4-Methyl-2-propionamidopentanamido)-5-oxopentanoic acid (**3.7c**):

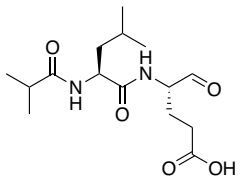
The product (36 mg, 64%) was isolated as an orange solid. ^{1}H NMR (500 MHz; DMSO-d_6): δ 9.56 (s, 1H), 9.36 (s, 1H), 8.38 (d, J = 7.1 Hz), 7.92 (m, 1H), 4.33 (m, 1H), 4.03 (m, 1H), 2.12 (m, 2H), 1.58 (m, 1H), 1.44 (m, 2H), 1.26 (m, 2H), 0.98 (m, 3H), 0.86 (m, 6H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 200.7, 173.9, 173.1, 125.5, 57.4, 50.7, 39.5, 29.5, 28.3, 24.3, 23.1, 23.0, 21.6, 18.1, 16.7, 9.9. HRMS (ESI): Calc'd for formula $C_{14}H_{23}N_2O_5^-$ $[M–H]^-$ 299.1612, found 299.1611.



3.7d

3.4.4.4. (*S*)-4-((*S*)-2-Butyramido-4-methylpentanamido)-5-oxopentanoic acid (**3.7d**):

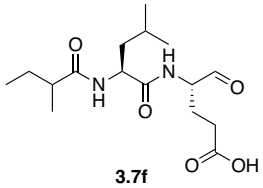
The product (12 mg, 28%) was isolated as an orange solid. ^{1}H NMR (500 MHz; DMSO-d_6): δ 9.56 (s, 1H), 9.35 (s, 1H), 8.18 (m, 1H), 7.95 (m, 1H), 4.59 (m, 1H), 4.33 (m, 1H), 4.03 (m, 1H), 2.48 (m, 2H), 2.07 (m, 2H), 1.49 (m, 3H), 1.27 (m, 1H), 0.84 (m, 9H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 200.7, 173.9, 173.1, 172.1, 57.4, 53.5, 50.7, 37.1, 29.5, 24.3, 23.1, 21.6, 18.7, 18.1, 16.7, 13.5. HRMS (ESI): Calc'd for formula $C_{15}H_{25}N_2O_5^-$ $[M–H]^-$ 313.1769, found 313.1775.



3.7e

3.4.4.5. (*S*)-4-((*S*)-2-Isobutyramido-4-methylpentanamido)-5-oxopentanoic acid (**3.7e**):

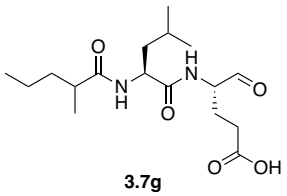
The product (7 mg, 55%) was isolated as a light brown solid. ^{1}H NMR (600 MHz; $CDCl_3$): δ 9.54 (s, 1H), 7.48 (s, 1H), 7.35 (s, 1H), 6.49 (d, *J* = 8.1 Hz, 1H), 4.60 (d, *J* = 6.1 Hz, 1H), 4.49 (m, 1H), 2.42 (m, 2H), 1.90 (m, 1H), 1.65 (m, 2H), 1.57 (m, 1H), 1.25 (m, 2H), 1.14 (m, 6H), 0.92 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): 198.7, 178.3, 176.2, 173.5, 110.2, 77.2, 58.2, 56.1, 55.5, 51.7, 51.0, 41.2, 35.7, 29.9, 25.1, 23.0, 22.4, 19.8, 19.8, 19.4. HRMS (ESI): Calc'd for formula $C_{15}H_{25}N_2O_5^-$ $[M–H]^-$ 313.1769, found 313.1768.



3.7f

3.4.4.6. (4*S*)-4-((2*S*)-4-Methyl-2-(2-methylbutanamido)pentanamido)-5-oxopentanoic acid (**3.7f**):

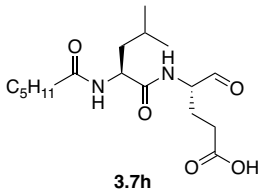
The product (9 mg, 48%) was isolated as a yellow solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.55 (s, 1H), 6.39 (m, 1H), 4.60 (m, 1H), 4.50 (m, 1H), 2.41 (m, 2H), 2.17 (m, 2H), 1.63 (m, 3H), 1.43 (m, 1H), 1.25 (m, 2H), 1.12 (m, 3H), 0.92 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 177.8, 173.5, 125.7, 53.7, 51.7, 30.6, 29.9, 27.5, 25.1, 23.0, 22.4, 22.3, 17.4, 12.0. HRMS (ESI): Calc'd for formula $C_{16}H_{27}N_2O_5^-$ $[M–H]^-$ 327.1925, found 327.1924.



3.7g

3.4.4.7. (4*S*)-4-((2*S*)-4-Methyl-2-(2-methylpentanamido)pentanamido)-5-oxopentanoic acid (**3.7g**):

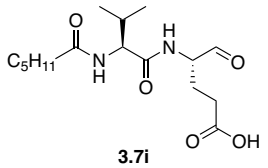
The product (7 mg, 16%) was isolated as a light brown solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.54 (s, 1H), 7.57 (m, 1H), 6.81 (m, 1H), 4.63 (m, 1H), 4.47 (m, 1H), 2.40 (m, 2H), 2.26 (m, 2H), 1.89 (m, 1H), 1.57 (m, 3H), 1.25 (m, 4H), 1.08 (m, 3H), 0.89 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 198.6, 178.2, 176.4, 175.8, 128.8, 110.2, 60.7, 58.2, 56.1, 51.7, 41.0, 36.5, 25.0, 22.3, 21.8, 20.7, 17.91, 17.72, 14.2. HRMS (ESI): Calc'd for formula $C_{17}H_{29}N_2O_5^-$ $[M–H]^-$ 341.2082, found 341.2082.



3.7h

3.4.4.8. (*S*)-4-((*S*)-2-Hexanamido-4-methylpentanamido)-5-oxopentanoic acid (**3.7h**):

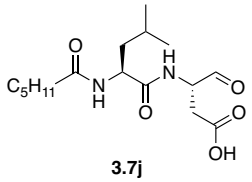
The product (11 mg, 34%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.53 (s, 1H), 7.67 (d, J = 7.0 Hz, 1H), 6.74 (d, J = 8.2 Hz, 1H), 4.68 (br s, 1H), 4.44 (br s, 1H), 2.39 (br s, 2H), 2.20 (m, 4H), 1.90 (m, 1H), 1.59 (m, 2H), 1.28 (m, 4H), 0.91 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 198.5, 176.0, 174.9, 171.5, 60.7, 58.3, 51.8, 41.2, 36.5, 31.5, 29.8, 25.5, 22.9, 22.5, 21.0, 14.4, 14.1. HRMS (ESI): Calc'd for formula $C_{17}H_{29}N_2O_5^-$ $[M–H]^-$, 327.1925; Found, 327.1924 .



3.7i

3.4.4.9. (*S*)-4-((*S*)-2-Hexanamido-3-methylbutanamido)-5-oxopentanoic acid (**3.7i**):

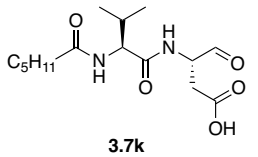
The product (8 mg, 23%) was isolated as an orange solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.55 (s, 1H), 6.60 (m, 1H), 6.11 (m, 1H), 4.35 (m, 1H), 4.13 (m, 1H), 2.42 (m, 2H), 2.24 (m, 4H), 1.61 (m, 3H), 1.28 (m, 4H), 0.89 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 176.6, 175.2, 77.2, 56.1, 36.5, 31.5, 25.6, 22.5, 21.0, 19.3, 18.7, 14.1. HRMS (ESI): Calc'd for formula $C_{16}H_{27}N_2O_5^-$ $[M–H]^-$ 327.1925, found 327.1924.



3.7j

3.4.4.10. (*S*)-3-((*S*)-2-Hexanamido-4-methylpentanamido)-4-oxobutanoic acid (**3.7j**):

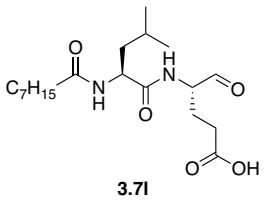
The product (10 mg, 24%) was isolated as a colorless glass. ^{1}H NMR (500 MHz; $CDCl_3$): δ 7.34 (m, 1H), 6.21 (m, 1H), 4.59 (m, 2H), 2.26 (m, 3H), 1.63 (m, 4H), 1.30 (m, 4H), 0.96 (m, 2H), 0.88 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 176.8, 176.1, 174.3, 57.2, 36.8, 31.6, 31.5, 31.2, 25.6, 22.56, 22.51, 21.0, 19.2, 17.9, 14.1. HRMS (ESI): Calc'd for formula $C_{16}H_{27}N_2O_5^-$ $[M–H]^-$ 327.1925, found 327.1926.



3.7k

3.4.4.11. (*S*)-3-((*S*)-2-Hexanamido-3-methylbutanamido)-4-oxobutanoic acid (**3.7k**):

The product (7 mg, 17%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.72 (s, 1H), 8.70 (br s, 1H), 8.18 (m, 1H), 7.14 (m, 1H), 4.59 (m, 1H), 4.50 (m, 1H), 2.20 (m, 2H), 1.57 (m, 5H), 1.26 (m, 6H), 0.88 (m, 9H). ^{13}C NMR (151 MHz; $CDCl_3$): δ 176.2, 175.1, 174.0, 77.2, 56.0, 53.6, 51.9, 41.1, 36.4, 31.5, 25.5, 25.0, 22.9, 22.5, 22.0, 21.0, 14.1. HRMS (ESI): Calc'd for formula $C_{15}H_{25}N_2O_5^-$ $[M–H]^-$ 313.1769, found 313.1769.



3.7l

3.4.4.12. (*S*)-4-((*S*)-4-Methyl-2-octanamidopentanamido)-5-oxopentanoic acid (**3.7l**):

The product (7 mg, 15%) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.54 (s, 1H), 7.45 (s, 1H), 6.29 (br s, 1H), 6.17 (m, 1H), 4.49 (m, 1H), 4.35 (m, 1H), 2.42 (m, 2H), 2.23 (m, 4H), 1.95 (m, 1H), 1.62 (m, 4H), 1.28 (m, 8H), 0.92 (m, 9H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 175.0, 174.6, 170.5, 128.7, 125.7, 66.0, 56.0, 53.6, 31.9, 29.3, 29.2, 25.9, 25.0, 23.5, 22.8, 15.4, 14.2. HRMS (ESI): Calc'd for formula $C_{19}H_{33}N_2O_5^-$ $[M–H]^-$ 369.2395, found 369.2395.

3.4.5. In vitro SspA inhibition assays

For SspA inhibition assays, the reaction mixture (50 μL) contained 50 mM Tris-HCl (pH 8), 1 ng/μL endoproteinase GluC (Worthington Biochemical Corporation, Lakewood, NJ), and 75 μM Z-LLE-AMC (Ubiquitin-Proteasome Biotechnologies, Aurora, CO). The assays were conducted in half-area white microplates (Enzo Life Sciences). To set up the reaction mixtures, assay buffer was added to each well, followed by the candidate inhibitor and then SspA. The protease was incubated with inhibitor at room temperature for 10 min in order to allow for protease/inhibitor interaction. Substrate was then added, and the reaction mixtures were monitored for fluorescence (367 nm excitation/460 nm emission, PMT medium, plate read height 0.81 mm) in a **Spectramax i3** Plate Reader once per minute for 20 min at 30 ºC. Reactions were performed in duplicate and inhibitor efficiency is calculated as a mean of both trials. The positive control inhibitor Ac-Glu^{P}(OPh)$_2$ was used to validate the assay.[14] For determination of IC50 values of compounds **3.7h** and **3.7l**, experiments were performed in triplicate over a concentration range of 0.01 – 800 μM, curves individually fit using GraphPad Prism 7, and error then calculated from IC50 values calculated for each separate series of serial dilutions.

3.4.6. Secreted protease activity assays in *E. faecalis* and *S. aureus*

A similar assay was used to measure proteolytic activity of *E. faecalis* and *S. aureus* supernatants on the Z-Leu-Leu-Glu-AMC substrate. *E. faecalis* V583 and mutants Δ*gelE*, Δs*prE*, and Δ*gelE*Δs*prE* were inoculated from single colonies in 5 mL BHI medium and grown to saturation. *S. aureus* JE2 and its *sspA* transposon mutant were inoculated from single colonies in TSB medium, grown to saturation, reinoculated 1:100 in TSB medium, and grown to mid-log phase. The cells were pelleted by centrifugation (4000 r.p.m. x 10 min) and the supernatants

were filtered through a 0.2 μM filter. The supernatants were then used at 80% final strength in the Z-Leu-Leu-Glu-AMC cleavage assay. For one experiment with the *E. faecalis* strains, the supernatants were concentrated with a 5K MWCO spin filter (4000 g x 2 h) to approximately 700 μL volume, and this concentrated supernatant was then used in the assay at 80% strength. 1 ng/μL endoproteinase GluC was used as a positive control. To set up the reaction mixtures (50 μL), the culture supernatant was added to each well, followed by substrate (75 μM Z-Leu-Leu-Glu-AMC). The reaction mixtures were monitored for fluorescence (367 nm excitation/460 nm emission, PMT medium, plate read height 0.81 mm) in a **Spectramax i3** Plate Reader once per minute for 20 min at 30 ºC. Hydrolysis of the peptide substrate could not be observed under any of these experimental conditions.

3.4.7. Milk agar clearance assay in *E. faecalis*

To monitor protease production on milk agar plates, *E. faecalis* strains were grown overnight, cells were pelleted, and the supernatants were filtered through a 0.2 μm filter. BHI medium was prepared at 0.9x the desired final volume, and a 100 g/L solution of skim milk powder was also prepared (10x). After autoclaving, the milk solution was added to the agar solution to give the final desired concentration of milk powder (10 g/L).[78] Wells were made in the plates using the wide end of a glass pipette, and 50 μL of the culture supernatants was added to each well. The plates were then incubated face up at 37 ºC overnight.

3.4.8. Autolysis assays in *E. faecalis* and *S. aureus*

The autolysis assay procedures were adapted from previously reported conditions.[50–52] For autolysis experiments, *E. faecalis* strains were grown overnight in 5 mL BHI medium at 37 ºC

and then reinoculated 1:100 in M17 medium with 3% glycine and grown to mid-log phase (~3 h) with shaking. For the experiments with *S. aureus*, the strains were grown overnight in TSB medium and then re-inoculated 1:100 in TSB medium with 3% glucose and grown to mid-log phase (~4 h) with shaking. For each test condition, 1.5 mL of the culture was centrifuged (13,000 g x 3 min) and the supernatant discarded. The pellet was washed three times with ice-cold sterile MQ water and resuspended in 1.4 mL sodium phosphate buffer (pH 6.8) with 0.01% (w/v) Triton X-100. 200 µL of each test condition were dispensed into a transparent 96-well plate, and OD600 was monitored at 37 ºC for 16 h at 30 min intervals in a BioTek SynergyHTX multi-mode microplate reader.

3.4.9. Biofilm formation assays in *E. faecalis* and *S. aureus*

Biofilm assay methods were adapted from previously published procedures.[53,54] Overnight cultures of *E. faecalis* strain V583 and mutants or *S. aureus* strain JE2 and its *sspA* transposon mutant were prepared by inoculating single colonies into 5 mL TSB medium with glucose (0.25% w/v) and grown at 37 ºC. OD was normalized across the strains and the cultures diluted 1:40 in TSB-glucose (0.25% or 1% w/v). 100 µL of this cell suspension was aliquoted into each desired well in a 96-well polystyrene microtiter plate (clear, tissue culture treated). The plates were incubated at 37 ºC for 24 – 48 h. Supernatant was shaken out of the plates, and the wells washed twice by gently submerging the entire plate in phosphate buffered saline and shaking out the liquid. To each well was added 125 µL of 0.1% (w/v) aqueous crystal violet stain and allowed to incubate at rt for 15 min. The stain was shaken out of the plate and the plate was washed three times with PBS. The microplate was then dried upside down overnight. The crystal violet in each well was solubilized by addition of 125 µL 30% (v/v) acetic acid in water. The

plate was incubated at rt for 15 min and then absorbance at 550 nm measured on a BioTek SynergyHTX multi-mode microplate reader.

3.5. References

1. Schneider, B. A. & Balskus, E. P. Discovery of small molecule protease inhibitors by investigating a widespread human gut bacterial biosynthetic pathway. *Tetrahedron* **74,** 3215–3230 (2018).

2. Chu, J. *et al.* Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat. Chem. Biol.* **12,** 1004–1006 (2016).

3. Cardozo, C., Chen, W. E. & Wilk, S. Cleavage of Pro-X and Glu-X bonds catalyzed by the branched chain amino acid preferring activity of the bovine pituitary multicatalytic proteinase complex (20S proteasome). *Arch. Biochem. Biophys.* **334,** 113–120 (1996).

4. Graybill, T. L., Dolle, R. E., Helaszek, C. T., Miller, R. E. & Ator, M. A. Preparation and evaluation of peptidic aspartyl hemiacetals as reversible inhibitors of interleukin-1ß converting enzyme (ICE). *Int. J. Pept. Protein Res.* **44,** 173–182 (1994).

5. Chen, Y., McClure, R. A., Zheng, Y., Thomson, R. J. & Kelleher, N. L. Proteomics guided discovery of flavopeptins: anti-proliferative aldehydes synthesized by a reductase domain-containing non-ribosomal peptide synthetase. *J. Am. Chem. Soc.* **135,** 10449–10456 (2013).

6. American Chemical Society; Chemical Abstracts Service. SciFinder – Explore. (2019). Available at: https://scifinder.cas.org/scifinder/view/scifinder/scifinderExplore.jsf. (Accessed: 7th March 2019)

7. Nemoto, T. K. *et al.* Characterization of the glutamyl endopeptidase from *Staphylococcus aureus* expressed in *Escherichia coli*. *FEBS J.* **275,** 573–587 (2008).

8. Kawalec, M., Potempa, J., Moon, J. L., Travis, J. & Murray, B. E. Molecular diversity of a putative virulence factor: Purification and characterization of isoforms of an extracellular serine glutamyl endopeptidase of *Enterococcus faecalis* with different enzymatic activities. *J. Bacteriol.* **187,** 266–275 (2005).

9. Thomas, V. C. *et al.* A fratricidal mechanism is responsible for eDNA release and contributes to biofilm development of *Enterococcus faecalis*. *Mol. Microbiol.* **72,** 1022–1036 (2009).

10. Massimi, I. *et al.* Identification of a novel maturation mechanism and restricted substrate specificity for the SspB cysteine protease of *Staphylococcus aureus*. *J. Biol. Chem.* **277,** 41770–41777 (2002).

11. Hamilton, R., Walker, B. & Walker, B. J. Synthesis and proteinase inhibitory properties of diphenyl phosphonate analogues of aspartic and glutamic acids. *Bioorg. Med. Chem. Lett.* **8,** 1655–1660 (1998).

12. Friesner, R. A. *et al.* Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).

13. Prasad, L., Leduc, Y., Hayakawa, K. & Delbaere, L. T. J. The structure of a universally employed enzyme: V8 protease from *Staphylococcus aureus*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60,** 256–259 (2004).

14. Burchacka, E., Skoreński, M., Sieńczyk, M. & Oleksyszyn, J. Phosphonic analogues of glutamic acid as irreversible inhibitors of *Staphylococcus aureus* endoproteinase GluC: An efficient synthesis and inhibition of the human IgG degradation. *Bioorg. Med. Chem. Lett.* **23,** 1412–1415 (2013).

15. Prokešová, L. *et al.* Cleavage of human immunoglobulins by serine proteinase from *Staphylococcus aureus*. *Immunol. Lett.* **31,** 259–265 (1992).

16. Acton, D. S., Tempelmans Plat-Sinnige, M. J., Van Wamel, W., De Groot, N. & Van Belkum, A. Intestinal carriage of *Staphylococcus aureus*: How does its frequency compare with that of nasal carriage and what is its clinical impact? *Eur. J. Clin. Microbiol. Infect. Dis.* **28,** 115–127 (2009).

17. Blaimont, B., Charlier, J. & Wauters, G. Comparative distribution of *Enterococcus* species in faeces and clinical samples. *Microb. Ecol. Health Dis.* **8,** 87–92 (1995).

18. Dunny, G. M., Hancock, L. E. & Shankar, N. "Enterococcal Biofilm Structure and Role in Colonization and Disease" in *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]* (eds. Gilmore M. S., Clewell D. B., Ike Y., et al.) (Massachusetts Ear and Eye Infirmary, 2014). Available from: https://www.ncbi.nlm.nih.gov/books/NBK190433/.

19. Garsin, D. A. *et al.* "Pathogenesis and Models of Enterococcal Infection" in *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]* (eds. Gilmore M. S., Clewell D. B., Ike Y., et al.) (Massachusetts Ear and Eye Infirmary, 2014). Available from: https://www.ncbi.nlm.nih.gov/books/NBK190426/.

20. Thomas, V. C., Thurlow, L. R., Boyle, D. & Hancock, L. E. Regulation of autolysis-

dependent extracellular DNA release by *Enterococcus faecalis* extracellular proteases influences biofilm development. *J. Bacteriol.* **190,** 5690–5698 (2008).

21. Mohamed, J. A., Huang, W., Nallapareddy, S. R., Teng, F. & Murray, B. E. Influence of origin of isolates, especially endocarditis isolates, and various genes on biofilm formation by *Enterococcus faecalis*. *Infect. Immun.* **72,** 3658–3663 (2004).

22. Pillai, S. K. *et al.* Effects of glucose on *fsr*-mediated biofilm formation in *Enterococcus faecalis*. *J. Infect. Dis.* **190,** 967–970 (2004).

23. Hancock, L. E. & Perego, M. The *Enterococcus faecalis fsr* two-component system controls biofilm development through production of gelatinase. *J. Bacteriol.* **186,** 5629–5639 (2004).

24. Singh, K. V., Qin, X., Weinstock, G. M. & Murray, B. E. Generation and Testing of Mutants of *Enterococcus faecalis* in a Mouse Peritonitis Model. *J. Infect. Dis.* **178,** 1416–1420 (1998).

25. Qin, X., Singh, K. V., Weinstock, G. M. & Murray, B. E. Effects of *Enterococcus faecalis fsr* genes on production of gelatinase and a serine protease and virulence. *Infect. Immun.* **68,** 2579–2586 (2000).

26. Sifri, C. D. *et al.* Virulence effect of *Enterococcus faecalis* protease genes and the quorum-sensing locus *fsr* in *Caenorhabditis elegans* and mice. *Infect. Immun.* **70,** 5647–5650 (2002).

27. Engelbert, M., Mylonakis, E., Ausubel, F. M., Calderwood, S. B. & Gilmore, M. S. Contribution of gelatinase, serine protease, and *fsr* to the pathogenesis of *Enterococcus faecalis* endophthalmitis. *Infect. Immun.* **72,** 3628–3633 (2004).

28. Suzuki, T. *et al.* Contribution of secreted proteases to the pathogenesis of postoperative *Enterococcus faecalis* endophthalmitis. *J. Cataract Refract. Surg.* **34,** 1776–1784 (2008).

29. Ike, Y., Craig, R. A., White, B. A., Yagi, Y. & Clewell, D. B. Modification of Streptococcus faecalis sex pheromones after acquisition of plasmid DNA. *Proc. Natl. Acad. Sci. U.S.A.* **80,** 5369–5373 (1983).

30. Thurlow, L. R. *et al.* Gelatinase contributes to the pathogenesis of endocarditis caused by *Enterococcus faecalis*. *Infect. Immun.* **78,** 4936–4943 (2010).

31. Dubin, G. Extracellular proteases of *Staphylococcus* spp. *Biol. Chem.* **383,** 1075–1086 (2002).

32. Novick, R. P. & Geisinger, E. Quorum sensing in Staphylococci. *Annu. Rev. Genet.* **42,** 541–564 (2008).

33. Lister, J. L. & Horswill, A. R. *Staphylococcus aureus* biofilms: recent developments in biofilm dispersal. *Front. Cell. Infect. Microbiol.* **4,** 178 (2014).

34. Archer, N. K. *et al. Staphylococcus aureus* biofilms. *Virulence* **2,** 445–459 (2011).

35. Boles, B. R., Thoendel, M., Roth, A. J. & Horswill, A. R. Identification of Genes Involved in Polysaccharide-Independent *Staphylococcus aureus* Biofilm Formation. *PLoS One* **5,** e10146 (2010).

36. Martí, M. *et al.* Extracellular proteases inhibit protein-dependent biofilm formation in *Staphylococcus aureus*. *Microbes Infect.* **12,** 55–64 (2010).

37. McGavin, M. J., Zahradka, C., Rice, K. & Scott, J. E. Modification of the *Staphylococcus aureus* fibronectin binding phenotype by V8 protease. *Infect. Immun.* **65,** 2621–2628 (1997).

38. Chen, C. *et al.* Secreted proteases control autolysin-mediated biofilm growth of *Staphylococcus aureus*. *J. Biol. Chem.* **288,** 29440–29452 (2013).

39. Rowe, S. E., O'Gara, J. P., Houston, P., Waters, E. M. & Pozzi, C. Essential role for the major autolysin in the fibronectin-binding protein-mediated *Staphylococcus aureus* biofilm phenotype. *Infect. Immun.* **79,** 1153–1165 (2010).

40. Boles, B. R. & Horswill, A. R. *agr*-mediated dispersal of *Staphylococcus aureus* biofilms. *PLoS Pathog.* **4,** (2008).

41. O'Neill, E. *et al.* A novel *Staphylococcus aureus* biofilm phenotype mediated by the fibronectin-binding proteins, FnBPA and FnBPB. *J. Bacteriol.* **190,** 3835–3850 (2008).

42. Loughran, A. J. *et al.* Impact of individual extracellular proteases on *Staphylococcus aureus* biofilm formation in diverse clinical isolates and their isogenic *sarA* mutants. *MicrobiologyOpen* **3,** 897–909 (2014).

43. Coulter, S. N. *et al. Staphylococcus aureus* genetic loci impacting growth and survival in multiple infection environments. *Mol. Microbiol.* **30,** 393–404 (1998).

44. Rice, K., Peralta, R., Bast, D., de Azavedo, J. & McGavin, M. J. Description of Staphylococcus Serine Protease (ssp) operon in *Staphylococcus aureus* and nonpolar inactivation of *sspA*-encoded serine protease. *Infect. Immun.* **69,** 159–169 (2001).

45. Rudack, C., Sachse, F., Albert, N., Becker, K. & von Eiff, C. Immunomodulation of nasal epithelial cells by *Staphylococcus aureus*-derived serine proteases. *J. Immunol.* **183,** 7592–7601 (2009).

46. Kolar, S. L. *et al.* Extracellular proteases are key mediators of *Staphylococcus aureus* virulence via the global modulation of virulence-determinant stability. *MicrobiologyOpen* **2,** 18–34 (2013).

47. Fey, P. D. *et al.* A genetic resource for rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus* genes. *MBio* **4,** e00537-12 (2013).

48. Zhang, G. "Protease Assays" in *Assay Guidance Manual [Internet]* (eds. Sittampalam G. S., Coussens N. P., Brimacombe K., et al.) (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2004). Available from: https://www.ncbi.nlm.nih.gov/books/NBK92006/.

49. Pailin, T., Kang, D. H., Schmidt, K. & Fung, D. Y. C. Detection of extracellular bound proteinase in EPS-producing lactic acid bacteria cultures on skim milk agar. *Lett. Appl. Microbiol.* **33,** 45–49 (2001).

50. Cornett, J. B. & Shockman, G. D. Cellular lysis of *Streptococcus faecalis* induced with Triton X-100. *J. Bacteriol.* **135,** 153–160 (1978).

51. Del Papa, M. F., Hancock, L. E., Thomas, V. C. & Perego, M. Full activation of *Enterococcus faecalis* gelatinase by a C-terminal proteolytic cleavage. *J. Bacteriol.* **189,** 8835–8843 (2007).

52. Bose, J. L., Lehman, M. K., Fey, P. D. & Bayles, K. W. Contribution of the *Staphylococcus aureus* Atl AM and GL Murein hydrolase activities in cell division, autolysis, and biofilm formation. *PLoS One* **7,** e42244 (2012).

53. O'Toole, G. A. *et al.* [6] Genetic approaches to study of biofilms. *Methods Enzymol.* **310,** 91–109 (1999).

54. O'Toole, G. A. Microtiter dish biofilm formation assay. *J. Vis. Exp.* e2437 (2011). doi:10.3791/2437

55. Mootz, J. M., Malone, C. L., Shaw, L. N. & Horswill, A. R. Staphopains modulate *Staphylococcus aureus* biofilm integrity. *Infect. Immun.* **81,** 3227–3238 (2013).

56. Svendsen, I. & Breddam, K. Isolation and amino acid sequence of a glutamic acid specific endopeptidase from *Bacillus licheniformis*. *Eur. J. Biochem.* **204,** 165–171 (1992).

57. Kakudo, S. *et al.* Purification, characterization, cloning, and expression of a glutamic acid-specific protease from *Bacillus licheniformis* ATCC 14580. *J. Biol. Chem.* **267,** 23782–23788 (1992).

58. Leshchinskaya, I. B. *et al.* Glutamyl endopeptidase of *Bacillus intermedius*, strain 3-19. *FEBS Lett.* **404,** 241–244 (1997).

59. Yoshida, N. *et al.* Purification and characterization of an acidic amino acid specific endopeptidase of *Streptomyces griseus* obtained from a commercial preparation (Pronase). *J. Biochem.* **104,** 451–456 (1988).

60. Demidyuk, I. V, Chukhontseva, K. N. & Kostrov, S. V. Glutamyl endopeptidases: The puzzle of substrate specificity. *Acta Naturae* **9,** 17–33 (2017).

61. Hedstrom, L. Serine protease mechanism and specificity. *Chem. Rev.* **102,** 4501–4523 (2002).

62. Stennicke, H. R., Birktoft, J. J. & Breddam, K. Characterization of the S1 binding site of the glutamic acid-specific protease from *Streptomyces griseus*. *Protein Sci.* **5,** 2266–2275 (1996).

63. Niidome, T., Yoshida, N., Ogata, F., Ito, A. & Noda, K. Purification and characterization of an acidic amino acid-specific endopeptidase of *Bacillus subtilis* obtained from a commercial preparation (Protease Type XVI, Sigma). *J. Biochem.* **108,** 965–970 (1990).

64. Meijers, R. *et al.* The crystal structure of glutamyl endopeptidase from *Bacillus intermedius* reveals a structural link between zymogen activation and charge compensation. *Biochemistry* **43,** 2784–2791 (2004).

65. Moon, J. L., Banbula, A., Oleksy, A., Mayo, J. A. & Travis, J. Isolation and characterization of a highly specific serine endopeptidase from an oral strain of *Staphylococcus epidermidis*. *Biol. Chem.* **382,** 1095–1099 (2001).

66. Cavarelli, J. *et al.* The structure of *Staphylococcus aureus* epidermolytic toxin A, an atypic serine protease, at 1.7 Å resolution. *Structure* **5,** 813–824 (1997).

67. Markowitz, V. M. *et al.* IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40,** 115–122 (2012).

68. Garsky, V. M., Colonno, R. J., Cordingleys, G., Colonno, J. & Callahan, P. L. Substrate requirements of human rhinovirus 3C protease for peptide cleavage in vitro. *J. Biol. Chem.* **265,** 9062–9065 (1990).

69. Adams, M. J., Antoniw, J. F. & Beaudoin, F. Overview and analysis of the polyprotein cleavage sites in the family *Potyviridae*. *Mol. Plant Pathol.* **6,** 471–487 (2005).

70. Sessenwein, J. L. *et al.* Protease-mediated suppression of DRG neuron excitability by commensal bacteria. *J. Neurosci.* **37,** 11758–11768 (2017).

71. Bukowski, M., Wladyka, B. & Dubin, G. Exfoliative toxins of *Staphylococcus aureus*. *Toxins (Basel).* **2,** 1148–1165 (2010).

72. Iwase, T. *et al. Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature* **465,** 346–349 (2010).

73. Johnson, B. T., Shaw, L. N., Nelson, D. C. & Mayo, J. A. Extracellular proteolytic activities expressed by *Bacillus pumilus* isolated from endodontic and periodontal lesions. *J. Med. Microbiol.* **57,** 643–651 (2008).

74. Snijder, E. J., Wassenaar, A. L. M., Van Dinten, L. C., Spaan, W. J. M. & Gorbalenya, A. E. The arterrvirus Nsp4 protease is the prototype of a novel group of chymotrypsin-like enzymes, the 3C-like serine proteases. *J. Biol. Chem.* **271,** 4864–4871 (1996).

75. Walker, A. W. *et al.* The species composition of the human intestinal microbiota differs between particle-associated and liquid phase communities. *Environ. Microbiol.* **10,** 3275–3283 (2008).

76. Grahl-Nielsen, O. & Solheim, E. Gas chromatography and mass spectrometry of derivatives of amino acids. *J. Chromatogr. A* **105,** 89–94 (1975).

77. Yang, X. *et al.* Flavour Modifying Compounds. US 2014/0127144 (2014).

78. Morris, L. S., Evans, J. & Marchesi, J. R. A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. *J. Microbiol. Methods* **91,** 144–146 (2012).

4. Structure prediction, synthesis, and biological investigations of a putative gut microbial peptide aldehyde library

4.1. Introduction

In Chapters 2 and 3 of this work, we discussed our efforts to discover the structure of ruminopeptin(s) from *Ruminococcus bromii* and demonstrate a putative functional role for these protease inhibitors. However, these compounds appear to be members of a much larger family of putative gut microbial natural products. The gut microbiota is a complex mixture of hundreds of species, and the work that initially identified the *rup* gene cluster from *R. bromii* also identified additional nonribosomal peptide synthetase (NRPS) gene clusters predicted to produce peptide aldehydes in this environment.[1] These gene clusters occur in approximately 30 different gut bacterial species. In order to gain a broader sense of the function(s) of this entire family of compounds, we expanded the scope of our studies to encompass some of these additional gene clusters. When we used bioinformatic tools to predict the core structures of these compounds and their likely amino acid building blocks, we found a great deal of structural diversity, leading us to hypothesize that while peptide aldehyde production seems to be a notable function of commensal gut microbial species, compounds from different biosynthetic pathways may have evolved for different purposes or interact with different targets. The gut bacterial species predicted to produce peptide aldehydes are diverse, but many are classified as members of Clostridium clusters IV and XIVa, which both have important roles in human health and the gut microbial community.[2]

In our work with the *rup* gene cluster, we initially cloned and heterologously expressed individual modules of the RupA NRPS and performed biochemical assays to validate our bioinformatic predictions of its function and substrate specificity. Though this work provided

useful information about the likely side chain length of the *N*-acyl group of ruminopeptin(s) and the specific amino acids incorporated in the P1 and P2 positions, it was also time and resource intensive. As our major interest in these NRPS enzymes was to discover bioactive small molecules that they may produce, we sought to eliminate this step in our expanded investigation. Therefore, in this chapter, we relied solely on bioinformatic analysis to predict the structures of small libraries of potential NRPS gene cluster products (Figure 4.1).

Due to ambiguities in bioinformatic prediction tools, the approach described in this chapter has led us to synthesize compounds that likely possess a range of structural resemblance to the actual natural products of these NRPS enzymes. Though our approach may identify bioactive compounds which are not actually produced in the gut environment, if bioactivity is the end goal then the ecological relevance of these compounds is irrelevant.

After synthesis of our peptide aldehyde library, we evaluated these compounds as inhibitors of a group of human proteases predicted to be relevant in the context of the human gut. We also screened this library of compounds for inhibitory activity against two microbial phenotypes (growth and secreted protease). Finally, we formulated and tested an activity-based protein profiling (ABPP) workflow to discover novel and unpredictable targets for these peptide aldehyde compounds.

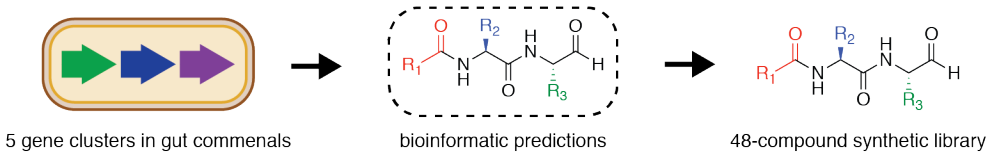


Figure 4.1: Discovery of bioactive small molecules from the gut microbiota by bioinformatic prediction and chemical synthesis.

In summary, our approach for accessing additional putative peptide aldheydes from the human gut microbiota was to bioinformatically predict the structures of these compounds and then synthesize small libraries encompassing their candidate products. A similar approach has recently been reported by Brady and coworkers.[3]

4.2. Results and discussion

4.2.1. Bioinformatic analysis and biosynthetic predictions for additional gut microbial NRPS gene clusters

We selected the additional gene clusters to study based on a combination of factors. Primarily, we were interested in gene clusters that had a demonstrated abundance in HMP gut metagenomes as reported in the initial 2014 investigation by Fischbach and coworkers.[1] Aside from the *rup* gene cluster (bgc45), the gene clusters they identified as most abundant in the HMP metagenomic stool samples were NRPS-encoding gene clusters bgc34, bgc35, bgc37, bgc38, bgc39, bgc41, and bgc52.[1] We also prioritized gene clusters from organisms that have been isolated as pure cultures, which would enable future in vivo studies if we could identify putative compounds of interest. Where available, we also considered the known ecological contexts and biological/metabolic functions of the particular bacterial species that contained these biosynthetic gene clusters.

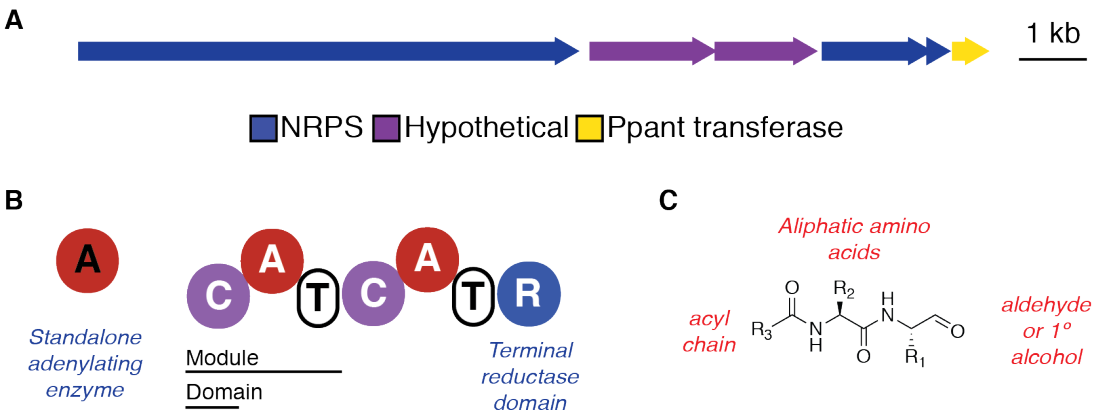


Figure 4.2: Gene cluster architecture and domain organization of nonribosomal peptide synthetase (NRPS) biosynthetic gene clusters bgc34, bgc37, and bgc38.

(A) bgc37 from *Clostridium leptum* ATCC 29065. The gene cluster contains three NRPS proteins, two hypothetical proteins, and a ppant transfersase. This same gene cluster architecture is shared with bgc34 and bgc38, though bgc38 does not contain a ppant transferase. (B) The main NRPS assembly line encoded by each of these gene clusters contains a condensation-starter (C-starter) domain and a terminal reductase (R) domain (A = adenylation domain, T = thiolation domain). They also each contain an additional stand-alone adenylating enzyme.

One group of compounds that we chose to focus on consists of NRPS-encoding gene clusters that we predicted would produce a variety of aliphatic *N*-acylated dipeptide aldehydes. These include bgc34 from *Lachnospiraceae* sp. 3_1_57FAA, bgc37 from *Clostridium leptum* DSM 753, and bgc38 from *Blautia producta* ATCC 27340. These three gene clusters have similar architectures, with a main two-module NRPS, an additional adenylating enzyme, an additional free-standing acyl carrier protein/T domain, and two regulatory proteins. Bgc34 and bgc37 also encode a putative phosphopantetheine (ppant) transferase for post-translational modification of the NRPS enzymes. The architecture and domain organization of bgc37 is shown as a representative example (Figure 4.2). Within this group of three gene clusters, the predicted A domain specificities differ (see below). All of the main NRPS enzymes from these pathways contain characteristic residues of C-starter domains, which catalyze amide bond formation

between pre-activated fatty acyl-CoAs and an initial assembly-line tethered amino acid (Figure 4.3).[4] The R domains in each of these clusters also contained the key conserved catalytic residues for producing peptide aldehydes (or alcohols) (Figure 4.4).[5]

Figure 4.3: ClustalW2 alignment of C domains shows residues that distinguish $^{L}C_{L}$ and C-starter activities.

C-domains were identified with the University of Maryland's PKS/NRPS Analysis Web-site and further trimmed from the beginning of conserved motif C1 to the end of motif C5.[4] The multiple sequence alignment was generated using ClustalW2 and visualized with Geneious. Included are the sequences of ClbN C_1 (*E. coli*, Q0P7K4), GlbF C_1 (*[Polyangium] brachysporum*, CAL80824.1), SrfAA C_1 (*Bacillus subtilis*, NP_388230.1), bgc34 C_1 (*Lachnospiraceae* sp. 3_1_57FAA, EGN48032.1), bgc37 C_1 (*Clostridium leptum* DSM 753, EDO60020.1), bgc38 C_1 (*Blautia producta*, WP_033141845.1), SrfAB C_2 (*Bacillus subtilis*, Q04747), DptA C_4 (*Streptomyces filamentosus*, AAX31557.1), and RupA C_2 (*R. bromii*, YP_007781236.1). ClbN C_1, GlbF C_1, and SrfAA C_1 are C-starter domains, and DptA C_4 and SrfAB C_2 are standard $^{L}C_{L}$ domains. The three numbered positions used in this analysis to distinguish C-starter domains and $^{L}C_{L}$ domains were identified based on work by Huson and coworkers[4]: 1 (S/G/A/C vs. P), 2 (V/I/L vs. A), and 3 (P vs. A/V).

Figure 4.3 (continued)

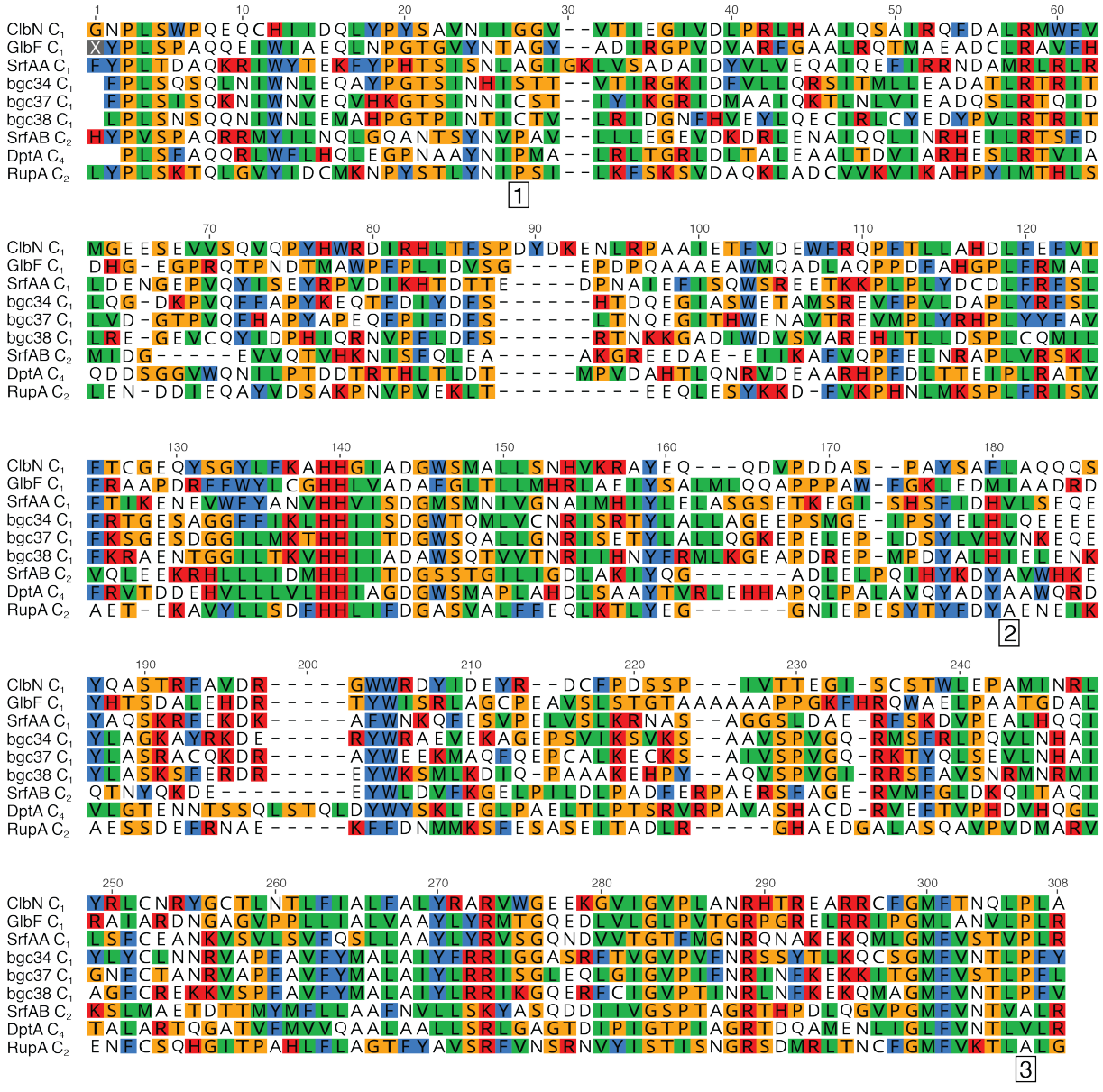
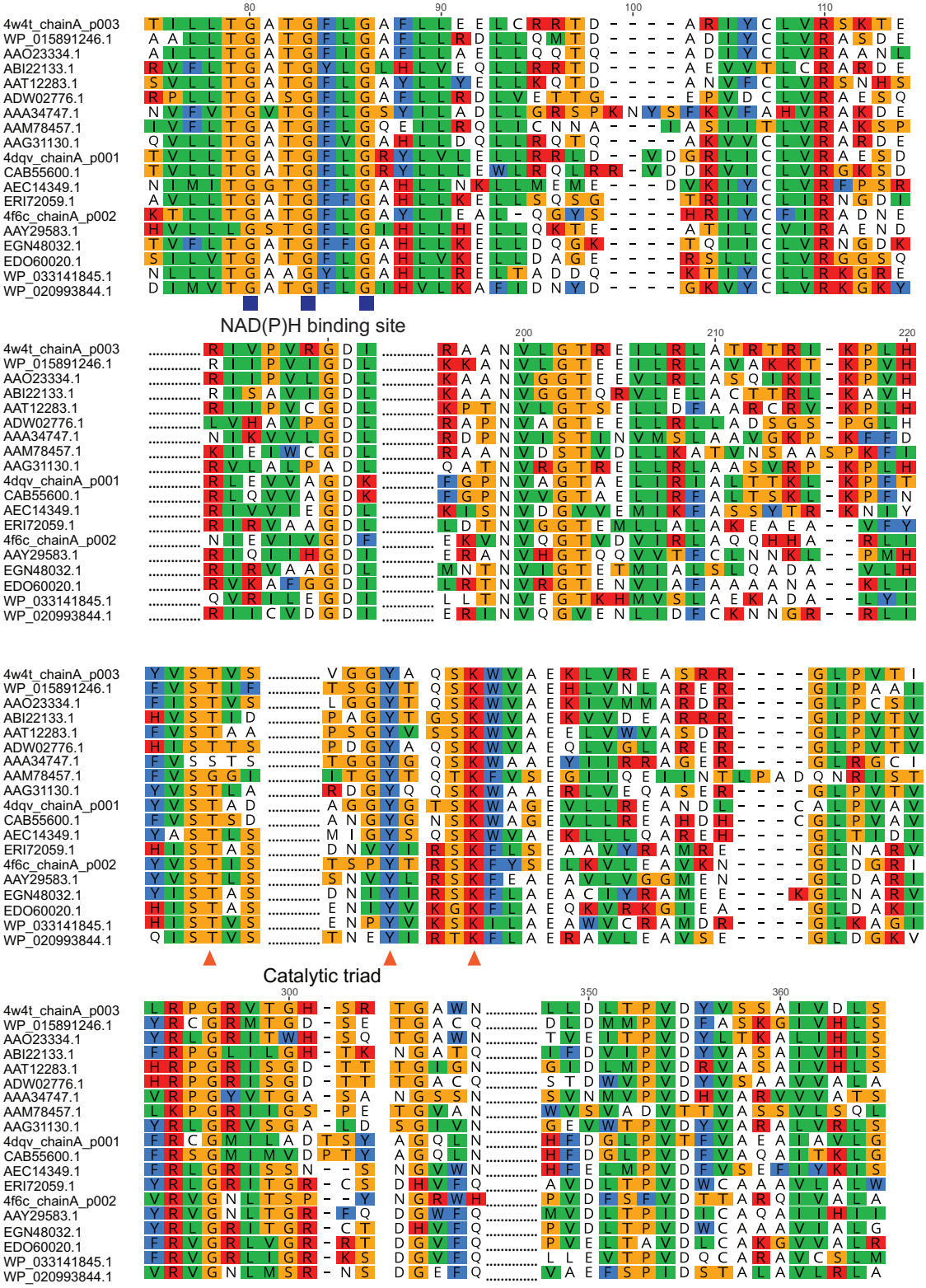


Figure 4.4: Structure-based alignment of terminal reductase domains from NRPS pathways shows conservation of residues important for catalysis.

The multiple sequence alignment was generated using PROMALS3d and visualized with Geneious. Included are the sequences of R domains from pathways producing myxalamid (*Stigmatella aurantiaca*, 4W4T), gramicidin (*Brevibacillus brevis*, WP_015891246.1), nostocyclopeptide (*Nostoc* sp. ATCC 53789, AAO23334.1), saframycin (*Streptomyces lavendulae*, ABI22133.1), lyngbyatoxin (*Lyngbya majuscula*, AAT12283.1), flavopeptin (*Streptomyces pratensis* ATCC 33331, ADW02776.1), lys2 (*Saccharomyces cerevisiae*, AAA34747.1), peptaibol (*Trichoderma virens*, AAM78457.1), myxochelin (*Stigmatella aurantiaca*, AAG31130.1), putative isonitrile lipopeptide[6] (product of *Rv0096-0101* gene cluster, *Mycobacterium tuberculosis*, AIR12822.1/4DQV), glycopeptidolipid (*Mycobacterium smegmatis* str. MC2 155, CAB55600.1), koranimide (*Bacillus* sp. NK2003, AEC14349.1), PZN2 (bgc35) (*Clostridium* sp. KLE 1755, ERI72059.1), aereusimine (*Staphylococcus aureus*, 4f6c), bgc34 (*Lachnospiraceae* sp. 3_1_57FAA, EGN48032.1), bgc37 (*Clostridium leptum* DSM 753, EDO60020.1), bgc38 (*Blautia producta*, WP_033141845.1), and bgc52 (*Ruminococcus* sp. 5_1_39BFAA, WP_020993844.1). Sequences were trimmed from 1 residue upstream of the beginning of conserved motif R1.[7] Conserved residues for NAD(P)H binding and the catalytic triad are indicated.[5]

Figure 4.4 (continued)



The set of enzymes encoded by these three gene clusters indicates that though they likely produce *N*-acyl dipeptide aldehydes, slightly different chemistry may be performed here than in the pathway encoded by the *rup* gene cluster. This hypothesized difference is due to the presence of an additional adenylating enzyme and an additional T domain in these clusters. Based on protein BLAST searches, we predicted that the additional adenylating enzyme in these gene clusters is not an NRPS A domain, but rather that it activates fatty acids. Therefore, these pathways may generate the activated acyl substrates that *N*-acylate their predicted dipeptide aldehyde products.

ATP-dependent, fatty acyl adenylating enzymes operate either as fatty acyl-CoA ligases (FACLs), which produce a freely diffusible acyl-CoA, or as fatty acyl-AMP ligases (FAALs), which produce acyl-adenylates that are directly loaded onto a T domain.[8,9] We predicted that either one of these paradigms for fatty acid activation could be operational in these biosynthetic pathways. If acting as an FACL, the A domain could produce a freely diffusible acyl-CoA that would be utilized as a substrate by the C-starter domain in the main NRPS. If acting as an FAAL, the A domain would activate a fatty acid and load it on to the extra T domain, which would then *N*-acylate the amino acid tethered to the first module of the NRPS using standard NRPS chemistry. This second possibility is also consistent with the observations that the stand-alone T domains in these gene clusters are likely of a suitable size to be functional and contain the necessary conserved serine residue for post-translational modification with a ppant arm (data not shown).[10]

In order to distinguish FAAL versus FACL activity, we relied on a previous analysis of the differences between these enzymes in *Mycobacterium tuberculosis* by Gokhale and coworkers.[8] Comparison of FACLs and FAALs in this species has revealed that FAALs contain an extra

insertion of ~20 residues,[8] and this insertion has also been observed in FAALs from *Escherichia coli* and *Legionella pneumophila*.[9] We generated a PROMALS3d structure based alignment of the predicted adenylating enzymes from bgc34, bgc37 and bgc38 along with a representative set of these adenylating enzymes from *M. tuberculosis*. This alignment clearly indicated that the stand-alone adenylating enzymes in these clusters are more similar to the FACLs in *M. tuberculosis* and are therefore likely producing freely diffusible acyl-CoAs (Figure 4.5). A bioinformatics tool for predicting acyl-CoA adenylating enzyme specificity predicted that these enzymes would selectively activate long chain fatty acids (the tool referenced here is no longer maintained and available online).[11] However, as there was some uncertainty about this prediction, for these three gene clusters we chose to synthesize potential products that would encompass a variety of acyl chain lengths (containing 6, 10, and 14 carbons). As these enzymes are predicted to be FACLs, the role(s) of the additional T domain in these NRPS gene clusters is not clear. They may be an evolutionary artifact, or they may serve some other role that we have not been able to predict.

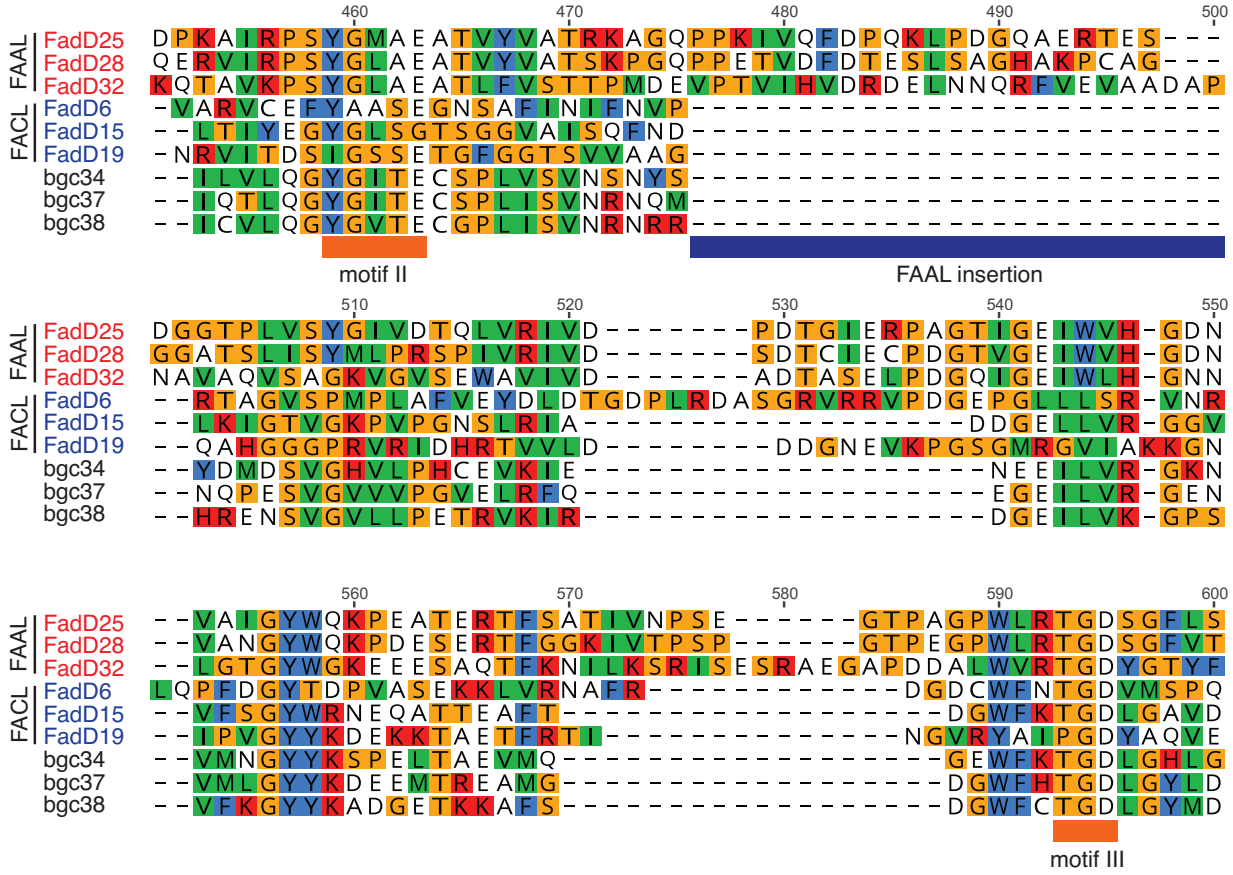


Figure 4.5: Structure-based alignment of stand-alone adenylating enzymes from NRPS gene clusters with FAALs and FACLs from *M. tuberculosis* shows that they lack the FAAL insertion motif.

The multiple sequence alignment was generated using Promals3D and visualized with Geneious. Included are the sequences of *M. tuberculosis* enzymes FadD25 (WP_003901187.1), FadD28 (CFE39237.1), FadD32 (WP_003899700.1), FadD6 (WP_003406240.1), FadD15 (WP_003906768.1), FadD19 (WP_003901660.1) and stand-alone adenylating enzymes from bgc34 (WP_009256183.1), bgc37 (WP_003532272.1), bgc38 (WP_033141779.1). FadD25, FadD28, and FadD32 are FAALs, while FadD6, FadD15, and FadD19 are FACLs.[8] The FAAL insertion was initially identified by Gokhale and coworkers.[8] Conserved motifs II and III for adenylate forming enzymes were initially identified by Dunaway-Mariano and coworkers.[12,13]

Having assigned a putative role for each of the biosynthetic enzymes, we next used A domain specificity prediction tools to identify the likely residues incorporated by each of these NRPS modules. In many cases, the Stachelhaus codes identified for these enzymes showed weak or no

matches to enzymes that have previously been characterized. Therefore, we relied on additional bioinformatic tools (NRPSPredictor2[14] and Minowa,[15] which are included in the analysis by antiSMASH[16]) to identify potential amino acids that may be activated by these domains.

The first gene cluster in this group, bgc34, was initially found in a species identified as *Lachnospiraceae* sp. 3_1_57FAA (EGN48032.1). More recently, the species *Eisenbergiella tayi* ATCC BAA-2558, with 16S ribosomal RNA sequence 99.4% ID to *Lachnospiraceae* sp. 3_1_57FAA, had its genome sequenced[17] and was found to contain a very close homolog of bgc34 (98.4% amino acid ID to the main NRPS protein). This species is a member of Clostridium cluster XIVa.[17] Using the bioinformatic tools mentioned above, we predicted that the NRPS from this species incorporates glycine in the P1 position and L-alanine or L-tyrosine in the P2 position (Figure 4.6). The A domain specificity conferring codes in this gene cluster are identical to those predicted for bgc35 from *Clostridium* sp. KLE 1755, an HMP isolate that has not been widely studied. Therefore, this NRPS may produce a similar suite of products.

The second gene cluster in this group, bgc37 from *C. leptum* DSM 753 (EDO60020.1) is from the type strain of this species. *C. leptum* is a member of Clostridium cluster IV,[18] and this cluster is sometimes known as the *C. leptum* group.[2] This is the only strain of this species with a publicly available sequenced genome. We predicted that this NRPS would incorporate L-alanine in the P1 position and either L-alanine or glycine in the P2 position (Figure 4.6). These specificities are identical to those that we predicted for bgc36 from *C. leptum* CAG:27, a metagenomic species.

The third gene cluster in this group, bgc38 from *B. producta* ATCC 27340 WP_033141845.1), is from the type strain of this species. This species is a member of Clostridium cluster XIVa.[19] We also identified this gene cluster in the genome of another strain

of this species (*B. producta* DSM 3507). We predicted that the NRPS encoded by this gene cluster would incorporate L-alanine in the P1 position and either L-leucine or L-tyrosine in the P2 position (Figure 4.6). These specificities are identical to those that we predicted for bgc39 from *Clostridium* sp. D5 and bgc40 from *Clostridium scindens* ATCC 35704.

Position	**235**	**236**	**239**	**278**	**299**	**301**	**322**	**330**	**Substrate/prediction**
NosC A_2[20]	D	I	L	Q	L	G	L	I	Gly
JamO[21]	D	L	F	N	N	A	L	T	L-Ala
CpbI A_2[22]	D	V	W	H	I	S	L	I	L-Ala
TycC A_6[23]	D	G	A	Y	T	G	E	V	L-Leu
MycB A_1[24]	D	A	L	S	V	G	E	V	L-Tyr
bgc34 A_1	**D**	**A**	**L**	**S**	**V**	**G**	**Q**	**V**	**L-Ala, L-Tyr (predicted)**
bgc34 A_2	**D**	**I**	**V**	**R**	**I**	**G**	**M**	**V**	**Gly (predicted)**
bgc37 A_1	**D**	**V**	**L**	**A**	**I**	**G**	**Q**	**I**	**Gly, L-Ala (predicted)**
bgc37 A_2	**D**	**T**	**T**	**Q**	**F**	**C**	**L**	**V**	**L-Ala (predicted)**
bgc38 A_1	**D**	**A**	**L**	**A**	**V**	**G**	**Q**	**V**	**L-Leu, L-Tyr (predicted)**
bgc38 A_2	**D**	**V**	**I**	**Q**	**I**	**M**	**I**	**V**	**L-Ala (predicted)**

Figure 4.6: **Predicted A domain specificity-conferring residues (Stachelhaus codes) for the NRPS enzymes encoded in bgc34, bgc37, and bgc38.**

Specificity-conferring residues were identified using the University of Maryland's PKS/NRPS Analysis Web-site.[25] Predictions were made based on that tool and AntiSMASH.[16] Numbering of positions references the sequence of phenylalanine-activating A domain GrsA.

In addition to the gene clusters predicted to produce *N*-acyl dipeptide aldehydes, we also targeted the bgc52 gene cluster from *Ruminococcus* sp. 5_1_39BFAA for small molecule discovery. The bgc52 gene cluster is simple, encoding only a single multi-module NRPS, a putative transport protein, a hypothetical protein, and a ppant transferase. In contrast to the other NRPS gene clusters previously discussed, the NRPS protein in this gene cluster contains an

additional adenylation domain as part of a canonical starter module. Therefore, we predicted that this gene cluster produces a tripeptide aldehyde. This NRPS also contains key conserved residues in its R domain that indicate it could generate a peptide aldehyde (or peptide alcohol) (Figure 4.4). Little is known about this species, but more recently *Blautia wexlerae* BAA-1564, which has a 16S ribosomal RNA sequence that is 98.8% ID to this species, was isolated[19] and found to encode a very close homolog of bgc52 (CUP17577.1, nucleotide sequence 98% ID). This species is a member of Clostridium cluster XIVa.[19] Another close homolog of this NRPS is encoded by *Blautia obeum* ATCC 29174 (99% nucleotide ID, CUO50064.1, CUN66576.1).

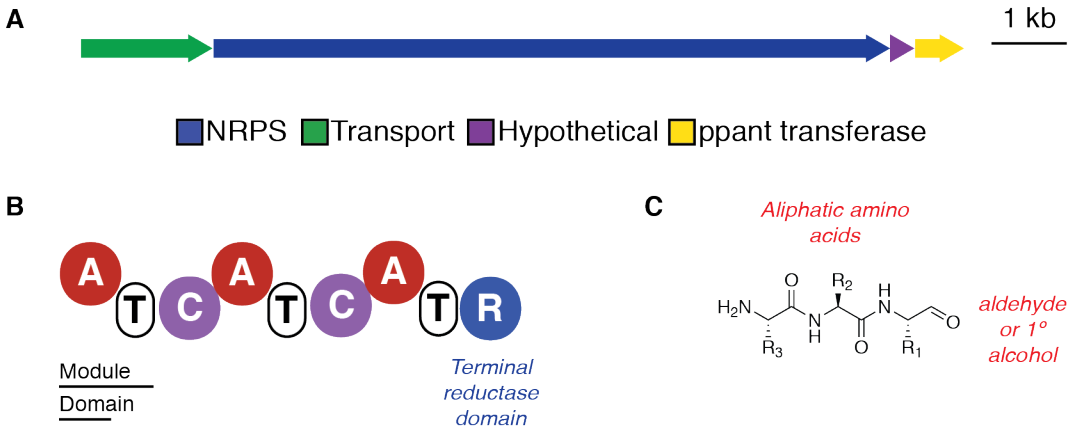


Figure 4.7: Gene cluster architecture and domain organization of NRPS biosynthetic gene cluster bgc52.

(A) bgc52 from *Ruminococcus* sp. 5_1_39BFAA (and *Blautia wexlerae* BAA-1564). This gene cluster encodes one NRPS, one putative transporter, one hypothetical protein, and a ppant transferase. (B) The main NRPS encoded by this gene cluster contains a canonical starter module, two full extension modules, and a terminal R domain.

As was the case with the other gene clusters discussed above, Stachelhaus codes provided weak or no matches for the A domains of the encoded NRPS, so we also relied on other tools to make these A domain specificity predictions (NRPSPredictor2[14] and Minowa,[15] which included

in the analysis by antiSMASH[16]). These analyses predicted that this NRPS should load L-phenylalanine, L-tyrosine or L-leucine in the P1 position, L-leucine or L-valine in the P2 position, and L-alanine, glycine, or L-leucine in the P3 position (Figure 4.8). The predicted A domain specificities of the bgc52 NRPS are the same as those predicted for bgc50, bgc51, and bgc53, which are from metagenomic species that have not been isolated.

Position	**235**	**236**	**239**	**278**	**299**	**301**	**322**	**330**	**Substrate/prediction**
NosC A_2[20]	D	I	L	Q	L	G	L	I	Gly
JamO[21]	D	L	F	N	N	A	L	T	L-Ala
CpbI A_2[22]	D	V	W	H	I	S	L	I	L-Ala
TycC A_4[23]	D	A	F	W	I	G	G	T	L-Val
TycC A_6[23]	D	G	A	Y	T	G	E	V	L-Leu
GrsA[26]	D	A	W	T	I	A	A	I	L-Phe
bgc52 A_1	**D**	**V**	**L**	**T**	**F**	**V**	**G**	**I**	**L-Ala, L-Gly, L-Leu (predicted)**
bgc52 A_2	**D**	**A**	**M**	**F**	**L**	**V**	**A**	**I**	**L-Leu, L-Val (predicted)**
bgc52 A_3	**D**	**A**	**I**	**T**	**V**	**L**	**G**	**V**	**L-Phe, L-Tyr, L-Leu (predicted)**

Figure 4.8: **Predicted A domain specificity-conferring residues (Stachelhaus codes) for the NRPS enzymes encoded in bgc52.**

Specificity-conferring residues were identified using the University of Maryland's PKS/NRPS Analysis Website[25] or from the intial analysis of A domain specifity-conferring residues by Marahiel and coworkers.[26] Position 278 of bgc52 A_1 and positions 299 and 301 from all bgc52 A domains were identified by manual inspection of a ClustalW2 alignment with GrsA. Predictions were made based on the results provided by that tool and AntiSMASH.[16] Numbering of positions references the sequence of phenylalanine-activating A domain GrsA.

The bacterial species harboring the gene clusters we selected for small molecule discovery have several potential and known biological roles in the commensal gut microbiota. All are

members of Clostridium clusters IV or XIVa.[2] As mentioned in Chapter 2, a 2013 study found that organisms in Clostridium cluster IV were significantly depleted in patients with IBD as compared with healthy controls.[27] *E. tayi*, which contains bgc34, is a strict anaerobe in a novel genus (within the Lachnospiraceae family) that has recently been isolated in two different studies from human blood, though it has been suggested that its natural environment may be in the gut.[17,28] This species is also part of a group of 17 Clostridia strains that were found to induce the production of $CD4^+FOXP3^+$ regulatory T (T_{reg}) cells, which serve an anti-inflammatory function, in mice.[29] *B. producta*, which contains bgc38, is part of the Simplified Human Microbiome (SIHUMIx), a gnotobiotic rat model which mimics the behavior of the healthy microbiota.[30] It is also a member of the same group of 17 Clostridia strains that induce T_{reg} cells.[29] *B. wexlerae*, which contains bgc52, was found to be particularly abundant in human fecal samples in a 2014 study.[31] Overall, the species that contain the gene clusters we selected for study may be associated with human health, and the small molecules produced by these NRPS pathways may be important for their beneficial effects.

Using the bioinformatic analyses discussed in this section, we predicted a set of potential products for bgc34 (6 products), bgc37 (6 products), bgc38 (6 products), and bgc52 (18 products) (Figure 4.9). As mentioned above, due to shared A domain specificity codes and gene cluster architectures, this set of predicted compounds may also encompass the potential products of bgc35, bgc36, bgc39, bgc40, bgc50, bgc51, and bgc53, leading to broad coverage of the potential peptide aldehydes produced by this family of gut microbial NRPS gene clusters. Additionally, as discussed in Chapters 2 and 3, we also predicted 12 possible products of bgc45 from *R. bromii*, and these structures may also represent the potential products of bgc44. We next

set about synthesizing all of these predicted structures so that we could evaluate their bioactivities.

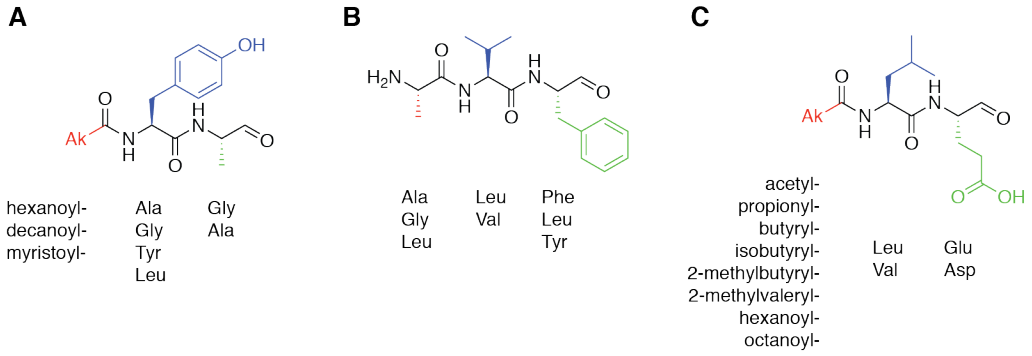


Figure 4.9: Structural features of the 48-compound peptide aldehyde library.
The structures predicted to be synthesized by the NRPS gene clusters can be divided into three major categories: aliphatic *N*-acyl dipeptide aldehydes (A), tripeptide aldehydes (B), and acidic *N*-acyl dipeptide aldehydes (ruminopeptin analogues) (C).

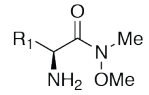
4.2.2. Synthesis of putative peptide aldehyde products of gut microbial NRPS gene clusters

Predicting the structures of gut microbial peptide aldehydes allowed us to proceed with their synthesis. Working in parallel using key building blocks and intermediates, we quickly accessed a small library of compounds. Though there are likely hundreds (or even thousands) of known compounds that could be broadly classified as peptide aldehydes, the Chemical Abstracts Service (as accessed through SciFinder) contains only ~40 *N*-acyl dipeptide aldehydes and ~70 tripeptide aldehydes that share the same core scaffolds as these compounds.[32] This analysis is limited to peptide aldehydes composed of the 20 canonical amino acids, capreomycidine, ornithine, and homoserine, and hydrocarbon acyl chains. It also excludes the 12 ruminopeptin analogues discussed in Chapter 3, as these compounds were previously reported by us.[33] Among these known compounds, two of the known *N*-acyl dipeptide aldehydes and three of the known

tripeptide aldehydes overlap with structures in our library. Additionally, one of the ruminopeptin analogues we synthesized in Chapter 3 is also a previously reported *N*-acyl dipeptide aldehyde. Therefore, though the synthetic chemistry to produce these sorts of compounds is very well established, the vast majority of the compounds that we synthesized are previously unreported and may therefore demonstrate novel bioactivities.

The synthesis of *N*-acyl dipeptide aldehydes **4.5a–r** was accomplished using standard solution-phase peptide coupling chemistry, with the synthetic strategy and reaction conditions inspired by prior syntheses of similar hydrophobic peptide aldehydes.[34,35] We first synthesized the key building blocks for these compounds: amino acid Weinreb amides **4.1a–e** and *N*-acyl amino acids **4.2a–m** (Figure 4.10A,B). Generally, the *N*-acyl amino acids were synthesized using the Schotten–Baumann reaction of acyl chlorides with amino acids under basic conditions, and the amino acid Weinreb amides were synthesized by coupling *N*,*O*-dimethylhydroxylamine hydrochloride to the corresponding Boc-protected amino acids and then performing Boc deprotections using 4 M HCl in dioxane. We next coupled these building blocks together using 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU) as the peptide coupling reagent to produce intermediates **4.4a–r** (Figure 4.11). Finally, these coupled products were reduced with $LiAlH_4$ to afford the product aldehydes **4.5a–r**. These two reaction steps afforded sufficiently clean products without requiring further purification, as assessed by NMR. The two-step yields from key precursors were in the range of 5–78%, affording 4–49 mg of each compound (Figure 4.11).

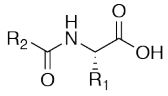
A



4.1a–e

Entry	Product	R_1
1	**4.1a**	Gly
2	**4.1b**	Ala
3	**4.1c**	Leu
4	**4.1d**	Phe
5	**4.1e**	Tyr

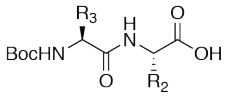
B



4.2a–l

Entry	Product	R_1	R_2
1	**4.2a**	C_5H_{11}	Gly
2	**4.2b**	C_9H_{19}	Gly
3	**4.2c**	$C_{13}H_{27}$	Gly
4	**4.2d**	C_5H_{11}	Ala
5	**4.2e**	C_9H_{19}	Ala
6	**4.2f**	$C_{13}H_{27}$	Ala
7	**4.1g**	C_5H_{11}	Leu
8	**4.2h**	C_9H_{19}	Leu
9	**4.2i**	$C_{13}H_{27}$	Leu
10	**4.2j**	C_5H_{11}	Tyr
11	**4.2k**	C_9H_{19}	Tyr
12	**4.2l**	$C_{13}H_{27}$	Tyr

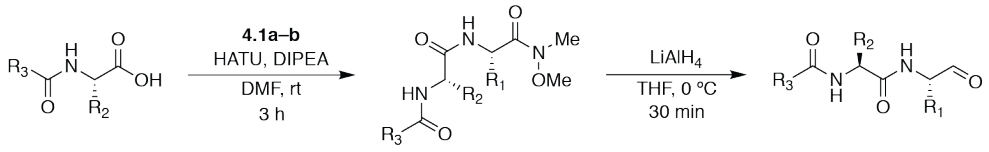
C



4.3a–f

Entry	Product	R_1	R_2
1	**4.3a**	Ala	Val
2	**4.3b**	Ala	Leu
3	**4.3c**	Gly	Val
4	**4.3d**	Gly	Leu
5	**4.3e**	Leu	Val
6	**4.3f**	Leu	Leu

Figure 4.10: Building blocks for the synthesis of *N*-acyl dipeptide aldehydes 4.5a–r and tripeptide aldehydes 4.8a–r.

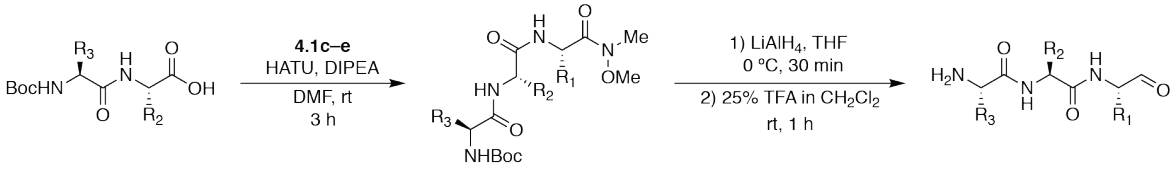


4.2a–l **4.4a–r** **4.5a–r**

Entry	Product	R_1	R_2	R_3	% Yield*	Entry	Product	R_1	R_2	R_3	% Yield*
1	**4.5a**	C_5H_{11}	Tyr	Gly	19%	10	**4.5j**	C_5H_{11}	Ala	Ala	32%
2	**4.5b**	C_9H_{19}	Tyr	Gly	40%	11	**4.5k**	C_9H_{19}	Ala	Ala	46%
3	**4.5c**	$C_{13}H_{27}$	Tyr	Gly	45%	12	**4.5l**	$C_{13}H_{27}$	Ala	Ala	12%
4	**4.5d**	C_5H_{11}	Ala	Gly	5%	13	**4.5m**	C_5H_{11}	Leu	Ala	78%
5	**4.5e**	C_9H_{19}	Ala	Gly	34%	14	**4.5n**	C_9H_{19}	Leu	Ala	43%
6	**4.5f**	$C_{13}H_{27}$	Ala	Gly	13%	15	**4.5o**	$C_{13}H_{27}$	Leu	Ala	61%
7	**4.5g**	C_5H_{11}	Gly	Ala	12%	16	**4.5p**	C_5H_{11}	Tyr	Ala	10%
8	**4.5h**	C_9H_{19}	Gly	Ala	25%	17	**4.5q**	C_9H_{19}	Tyr	Ala	15%
9	**4.5i**	$C_{13}H_{27}$	Gly	Ala	28%	18	**4.5r**	$C_{13}H_{27}$	Tyr	Ala	12%

Figure 4.11: Synthesis of *N*-acyl dipeptide aldehydes 4.5a–r.

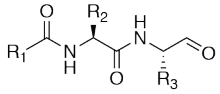
% Yield* = isolated yield for two reaction steps from key precursors (HATU = 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate, DIPEA = *N*,*N*-Diisopropylethylamine).



Entry	Product	R_1	R_2	R_3	% Yield*	Entry	Product	R_1	R_2	R_3	% Yield
1	**4.8a**	Ala	Val	Phe	48%	10	**4.8j**	Gly	Leu	Phe	58%
2	**4.8b**	Ala	Val	Tyr	50%	11	**4.8k**	Gly	Leu	Tyr	49%
3	**4.8c**	Ala	Val	Leu	50%	12	**4.8l**	Gly	Leu	Leu	26%
4	**4.8d**	Ala	Leu	Phe	34%	13	**4.8m**	Leu	Val	Phe	58%
5	**4.8e**	Ala	Leu	Tyr	48%	14	**4.8n**	Leu	Val	Tyr	49%
6	**4.8f**	Ala	Leu	Leu	49%	15	**4.8o**	Leu	Val	Leu	72%
7	**4.8g**	Gly	Val	Phe	83%	16	**4.8p**	Leu	Leu	Phe	44%
8	**4.8h**	Gly	Val	Tyr	66%	17	**4.8q**	Leu	Leu	Tyr	36%
9	**4.8i**	Gly	Val	Leu	55%	18	**4.8r**	Leu	Leu	Leu	29%

Figure 4.12: Synthesis of tripeptide aldehydes 4.8a–r.

% Yield* = isolated yield for three reaction steps from key precursors (TFA = Trifluoroacetic acid).



4.9a–l

Entry	Product	R_1	R_2	R_3
1	**4.9a**	Me	Leu	Asp
2	**4.9b**	Me	Leu	Glu
3	**4.9c**	C_2H_5	Leu	Glu
4	**4.9d**	C_3H_7	Leu	Glu
5	**4.9e**	*i*Bu	Leu	Glu
6	**4.9f**		Leu	Glu
7	**4.9g**		Leu	Glu
8	**4.9h**	C_5H_{11}	Leu	Glu
9	**4.9i**	C_5H_{11}	Val	Glu
10	**4.9j**	C_5H_{11}	Leu	Asp
11	**4.9k**	C_5H_{11}	Val	Asp
12	**4.9l**	C_7H_{15}	Leu	Glu

Figure 4.13: Ruminopeptin analogues synthesized in Chapter 3.

The syntheses of tripeptide aldehydes **4.8a–r** proceeded from Boc-protected dipeptides and amino acid Weinreb amides building blocks (Figure 4.10A,C). The synthetic strategy and reaction conditions were inspired by previous syntheses of hydrophobic peptide aldehydes with unprotected N-termini.[36] According to previously reported conditions, we first generated Boc-protected dipeptide methyl esters using 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) to couple Boc-protected amino acids and amino acid methyl esters under previously reported conditions.[37] These esters were then saponified with lithium hydroxide to afford Boc-protected dipeptides **4.3a–f** (Figure 4.10).[37] We then coupled these building blocks with the Weinreb amides **4.2c–e,** using HATU as the peptide coupling reagent, to afford Boc-protected tripeptide Weinreb amides **4.6a–r** (Figure 4.12). These compounds were reduced with $LiAlH_4$ to generate the aldehydes **4.7a–r**, and the Boc groups were then removed by treatment with 20% TFA in DCM to afford the final tripeptide aldehydes **4.8a–r**. Trituration of these compounds with diethyl ether afforded sufficiently clean products without further purification, as assessed by NMR. Overall, these target peptide aldehydes were synthesized with three step yields in the range of 26–83%, providing between 18–71 mg (Figure 4.12).

Overall, the 36 compounds described here, along with the 12 ruminopeptin analogues from Chapter 3 (Figure 4.13) comprise a 48-compound library of predicted gut microbial NRPS products. With these compounds in hand, we next evaluated their potential bioactivity in several different contexts.

4.2.3. Putative gut microbial peptide aldehydes inhibit human proteases

The chemical space theoretically accessible to NRPS biosynthetic enzymes contains a large number of peptide aldehydes, and there are a multitude of proteases in nature that could

potentially be inhibited by these compounds. Therefore, though inhibitory interactions between peptide aldehydes and proteases have been studied for decades,[38] there still remains an opportunity for discovering new physiological roles of this natural product class. Considering potential targets of peptide aldehyde compounds in the gut environment, we initially focused on human proteases. Peptidases are a major class of human enzymes, with 566 known protease encoding genes in the human genome.[39] Of these, 175 are serine proteases and 148 are cysteine proteases.[39] In addition to their catalytic strategies, these enzymes can also be divided by the substrate amino acids they recognize and whether they cleave large or small substrates. In thinking about putative targets of our peptide aldehydes, we focused on endopeptidases that cleave within large peptide chains (rather than exopeptidases that remove a single amino acid from the ends of proteins and di- and tripeptidyl peptidases that degrade these small signaling molecules).

In order to maximize our chances of identifying interactions between putative gut microbial peptide aldehydes and human proteases, we decided to work with two contract research organizations to conduct screens of the full 48 compound peptide aldehyde library. The initial screen of calpain 1 was performed by GenScript (Piscataway, NJ). All other human protease assays were performed by Reaction Biology Corporation (Malvern, PA). Out of a set of ~50 human cysteine and serine proteases available for screening, we conducted literature searches to prioritize 11 enzymes with known relevance in the gut context. Details about cleavage specificities and known inhibitors of the proteases we selected to screen are presented in Table 4.1.

Table 4.1: Human proteases screened with Reaction Biology Corporation and GenScript.

Primary specificities are indicated (with secondary specificities in parentheses) (φ = aromatic = Phe, Tyr, Trp; θ = aliphatic = Gly, Ala, Val, Leu, Ile). These published cleavage specificities have been determined using a variety of different methods, inckuding chromogenic substrate libraries and mass spectrometry analysis of peptide libraries.[40–47]

Table 4.1 (continued)

Protease	Nucleophile	Cleavage site specificity / enriched residues				Selected peptide aldehyde inhibitors
		P3	P2	P1	Ref.	
Calpain	Cysteine	Phe Leu Pro (Ile) (Met)	Leu Val (Ile) (Met)	Phe Leu (Tyr) (Val)	40	Ac-Leu-Leu-Nle-H; K_i 0.19 μM[48] Ac-Leu-Leu-Met-H; K_i 0.12 μM[48]
Cathepsin B	Cysteine		Ala Val Tyr Phe Ile	Gly (Ala) (Met) (Gln) (Thr)	41	Ac-Leu-Leu-Nle-H; K_i 0.15 μM[48] Ac-Leu-Leu-Met-H; K_i 0.10 μM[48]
Cathepsin L	Cysteine		φ Val Leu Ile	Gly Gln Thr Ala Asn	41,42	Ac-Leu-Leu-Nle-H; K_i 0.50 nM[48] Ac-Leu-Leu-Met-H; K_i 0.60 nM[48]
Cathepsin S	Cysteine		Val Leu Met Phe Ile	Gly Glu Gln Ala Thr	41,42	–
Cathepsin V	Cysteine		φ	Arg Lys (Gln) (Nle) (Met) (Thr)	42	–
Caspase 1	Cysteine	Glu	His (Thr)	Asp	43,44	Ac-Trp-Glu-His-Asp-H; K_i 56 pM[49]
Caspase 3	Cysteine	Glu	Val (Ile)	Asp	43,44	Ac-Tyr-Val-Ala-Asp-H; K_i 12 μM[50]
Caspase 8	Cysteine	Glu	Thr (Val) (Ile)	Asp	43,44	Ac-Ile-Glu-Thr-Asp-H; IC_{50} 50 nM[51]
Cathepsin G	Serine	θ	Leu Val (Phe)	Leu Phe Trp (Tyr)	45	Chymostatin; K_i = 0.15 μM[52]
Neutrophil elastase	Serine	Gln Leu Glu	Arg Nle Pro	Ile Val Thr	46	Elastatinal; K_i = 0.24 μM[53]
Trypsin	Serine			Lys Arg	47	Leupeptin; ID_{50} = 2 μg/mL[54] (corresponds to IC_{50} = 4.2 μM)

In our initial screens, we tested each compound at a single concentration against each protease in duplicate. Calpain and cathepsins B, L, S, and V were screened at 1 μM inhibitor concentration; caspases 1, 3, and 8 were screened at 10 μM inhibitor concentration; and serine proteases cathepsin G, neutrophil elastase, and trypsin were screened at 100 μM inhibitor concentration. We selected these initial concentrations based on previous knowledge of the potencies of peptide aldehyde inhibitors toward these or similar proteases, as well as several pilot experiments with sample cysteine, serine and threonine proteases (an in-house pilot experiment of the 20S proteasome and calpain, and a pilot screen of cathepsin G conducted by Reaction Biology Corporation, data not shown). Our goal was to identify an inhibitor concentration for the screen of each protease that would reveal specific activity of these compounds. Too low of a concentration would not tell us whether the compounds were active as protease inhibitors, while too high of a concentration would not reveal the structure activity relationships that we hoped to use to prioritize compounds for further investigation. The results of these single concentration screens showed that we selected these inhibitor concentrations with various degrees of success (Figure 4.14). For instance, we observed the desired range of activities against calpain at an inhibitor concentration of 1 μM but found that the peptide aldehyde compounds indiscriminately inhibited cathepsin L at an inhibitor concentration of 1 μM. The assays were conducted using fluorogenic protease substrates, and protease activity was quantified by measuring the increase in fluorescence over time (see Materials and Methods for details). The results of our screens are presented in Figure 4.14. Our discussion of the potential biological roles of these proteases and the results of our screens is presented in the following paragraphs.

Figure 4.14: Screening of 48 peptide aldehydes as inhibitors of human proteases.

(A) N-acyl dipeptide aldehydes. (B) Tripeptide aldehydes. (C) Ruminopeptin analogues. The assays were conducted by GenScript and Reaction Biology Corporation using fluorogenic protease substrates and measuring increase in fluorescence over time (see Materials and Methods for details). Assays were performed in duplicate and inhibitor efficiency is reported as a mean of both trials.

Figure 4.14 (continued)

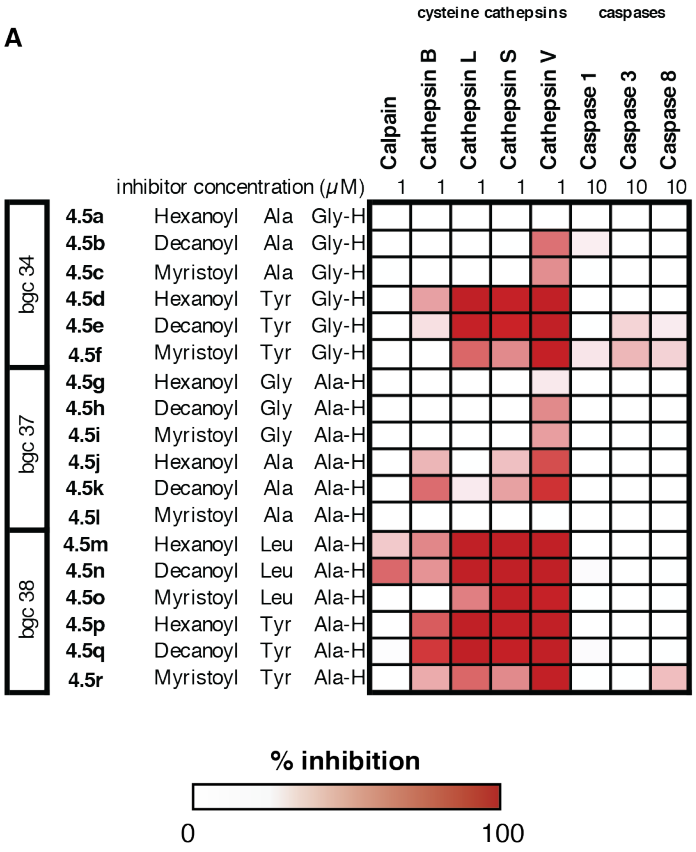
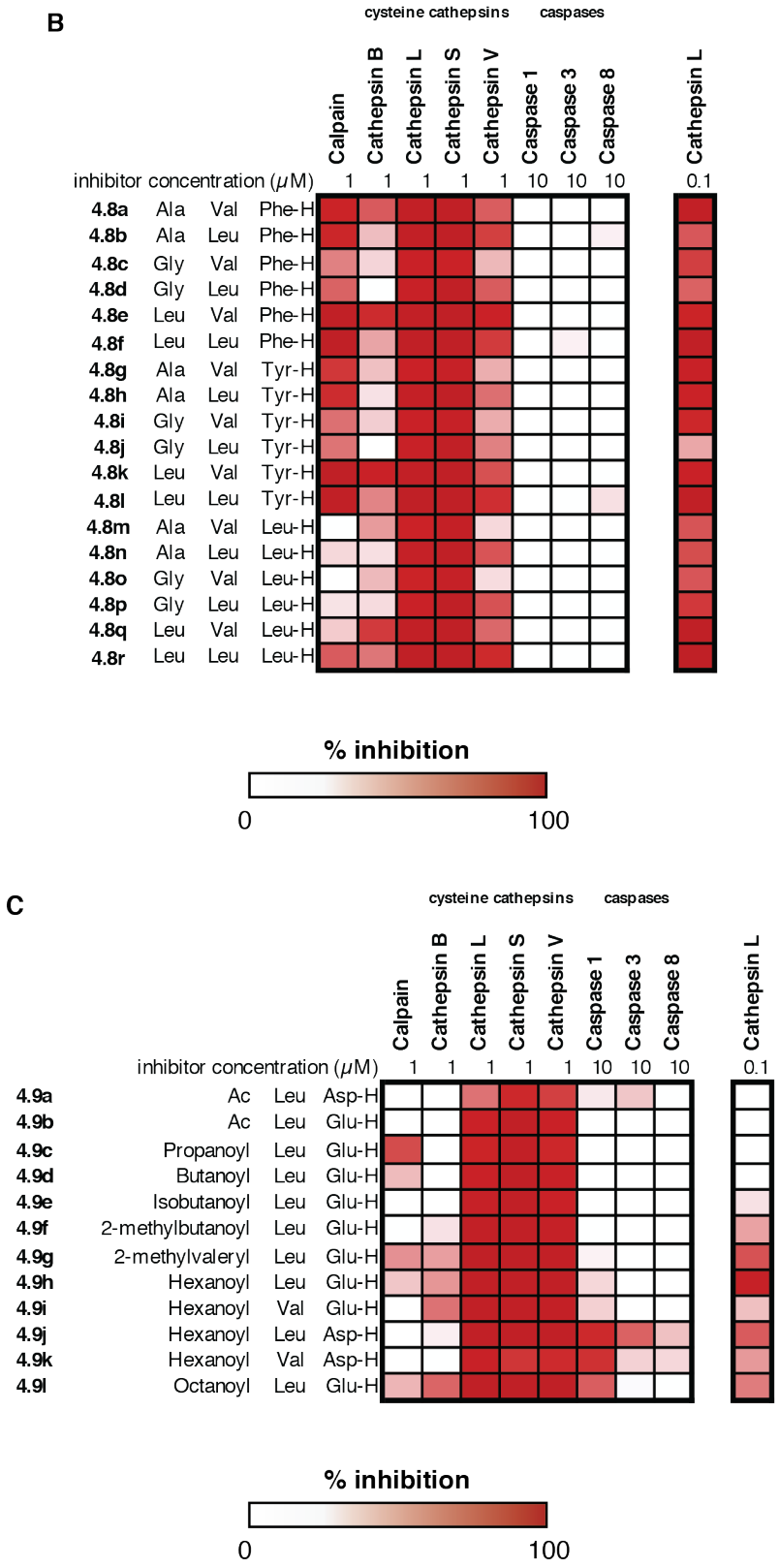


Figure 4.14 (continued)



The first protease we screened, calpain 1, is a calcium-dependent intracellular cysteine protease. Calpain 1 has a well-studied role in inflammation,[55] particularly through activation of the transcription factor NF-κB, which has led to its study in an inflammatory bowel disease (IBD) context.[56,57] Screening peptide aldehyde inhibitors against calpain revealed a group of compounds that were active against this protease at a concentration of 1 μM. These include an alanine aldehyde (**4.5n**), a glutamyl aldehyde (**4.9c**), and tripeptide aldehydes containing aromatic residues in the P1 position (**4.8a–l**).

We next screened for inhibition of the cysteine cathepsins B, L, S, and V. Broadly speaking, these lysosomal proteases participate in intracellular protein degradation and antigen presentation in the context of the immune system.[58] Some of these proteases are ubiquitously expressed, while others are confined to specific cell types.[58] Cathepsin B is ubiquitously expressed,[58] and simultaneous inhibition of cathepsins B and L has been evaluated in a mouse model of detergent-induced colitis by Rogler and coworkers.[59] Our screen of peptide aldehydes against cathepsin B at 1 μM inhibitor concentration revealed limited activity, with the most active compounds being alanine aldehyde **4.5q** and tripeptide aldehydes with a Leu-Val motif in positions P2-P3 (**4.8e**, **4.8k**, and **4.8q**).

Cathepsin L is ubiquitously expressed.[58] Its role in IBD (along with that of cathepsin B) was evaluated in the study referenced above, wherein Rogler and coworkers inhibited cathepsin L with a small molecule and also evaluated expression levels of the protease in macrophages isolated from normal and inflamed human tissue.[59] They found that this protease was upregulated in inflamed human mucosa and that simultaneous inhibition of cathepsins B and L improved histological scores of colon sections from dextran-sulphate-sodium (DSS)-treated mice.[59] In our screen of peptide aldehyde compounds against cathepsin L at 1 μM inhibitor concentration, we

observed some selective inhibition among the aliphatic *N*-acyl dipeptide aldehydes but very high inhibitory activity for almost all of the ruminopeptin analogues and tripeptide aldehydes. Therefore, we conducted a second experiment with those two compound families, screening at a lower concentration of 100 nM (Figure 4.14).

In this second screen, among the aliphatic *N*-acyl dipeptide aldehydes, we observed inhibition by glycine aldehydes such as **4.5d** and **4.5e**, which do align with the known specificity of this protease. We see no inhibition by compounds containing small residues in both P1 and P2 (**4.5g–l**), but compounds that contain alanine in P1 and leucine or tyrosine in P2 (**4.5m–r**) are also active. The observed inhibition by the ruminopeptin analogues (**4.9a–l**) is somewhat surprising but can perhaps be rationalized by the structural homology of glutamate to glutamine. The potent inhibition we observed of this protease with all the tripeptide aldehydes in this library (**4.8a–r**) is also somewhat surprising. However, though the structures of these compounds do not precisely align with the known P1 cleavage preferences of cathepsin L, they do resemble peptide aldehydes that are known to inhibit this protease (Table 4.1).

Cathepsin S is known to be expressed in antigen-presenting cells.[58] This protease may also be secreted under certain conditions,[60] and a 2011 study by Bunnett and coworkers suggested a potential association between this protease and IBD.[61] They used fluorescent activity-based probes to analyze the activation and localization of cathepsins in a genetic-pharmacological mouse model of IBD. They were able to detect cathepsin B and cathepsin S in luminal fluid from both the IBD model mice and healthy controls, but they observed that cathepsin S was activated in the IBD model state.[61] In our screen of the peptide aldehyde library against cathepsin S at 1 μM inhibitor concentration, the inhibition profile was very similar to that observed for cathepsin

L, with almost all compounds showing very strong inhibition. We did not screen the peptide aldehyde library against cathepsin S at a lower concentration.

Cathepsin V (also known as cathepsin L2) shares high sequence homology (78% ID) with cathepsin L.[62] This protease is not normally expressed in the colon, but it has been identified in colorectal carcinomas.[62] Our screen of cathepsin V against peptide aldehyde library at 1 μM inhibitor concentration revealed a pattern of inhibition generally similar to that observed for cathepsins L and S. There was some variation in activity among the tripeptide aldehydes, though there is no clearly discernible pattern for the structure-activity relationship (SAR) observed here.

We also chose to screen the peptide aldehydes for inhibition of caspases, which are intracellular cysteine proteases that serve a variety of functions.[63] Caspases can be divided into several categories: inflammatory caspases, which are involved with formation of the inflammasome and pyroptosis, and initiator caspases and executioner caspases, which are involved with apoptosis.[64] Based on the literature describing these proteases in the gut environment, we chose to limit our screen to caspases 1, 3, and 8. Caspase 1, also known as interleukin-converting enzyme, is an inflammatory caspase that is responsible for maturation of the pro-inflammatory cytokine IL-1β.[64] Inhibition of this protease has previously been suggested as a treatment for IBD.[65] We were particularly interested in this protease, as its inhibition by peptide aldehydes would explicitly link these compounds to an anti-inflammatory signaling pathway. In our screen of the peptide aldehyde library against this protease at 10 μM inhibitor concentration, we saw that several compounds containing an Asp residue inhibited this protease (**4.9j, 4.9k**). This was expected based on the known P1 specificity of this protease (Table 4.1). A long acyl chain substituent appears to be important for this inhibition by these Asp-containing dipeptide aldehydes, as an aspartyl aldehyde possessing an acetyl chain was a poor inhibitor

(**4.9a**). We also unexpectedly observed weaker inhibition of this protease by compound **4.9l**, which contains a glutamate residue in P1.

We next screened our peptide aldehydes for activity toward caspase 3, an initiator caspase involved in apoptosis.[64] This protease is a key activator of ATG16L1, which is known as the "essential autophagy gene", and a polymorphism in this gene is associated with Crohn's disease.[66] In our screen of the peptide aldehyde library against caspase 3 at 10 μM inhibitor concentration, we observed only weak inhibitory activity, with only compound **4.9j** giving partial inhibition at 10 μM.

We also screened our compound library for inhibition of caspase 8, which is mainly known as an executioner caspase involved in apoptosis.[64] However, this protease has more recently been identified as having a role in inflammasome activation under certain conditions.[67] Our screen of caspase 8 with the peptide aldehyde library at 10 μM inhibitor concentration revealed no significant inhibition. Though we have only discussed here the residue preferences for P1–P3, the various P4 residue preferences of these three caspases do exhibit some differences,[43] so the activity variation observed in our library and among known inhibitors of these proteases may be rationalized by invoking additional interactions (or lack thereof) with the substrate binding pockets of these proteases.

Finally, we also screened our peptide aldehyde library for inhibition of several serine proteases: cathepsin G, neutrophil elastase, and trypsin. Neutrophils, which respond to bacterial pathogens as part of the innate immune system, produce three major serine proteases (proteinase 3, cathepsin G, and neutrophil elastase).[68] Intracellularly, these proteases are involved with the degradation of phagocytized bacteria.[69] In an extracellular context they are also known to participate in inflammatory signaling pathways.[68,69] Trypsin, one of the major digestive

proteases, is secreted by the pancreas and found mainly in the small intestine.[70] Several studies have identified trypsin activity in the colon in human disease states, such as in irritable bowel syndrome[71] and colorectal cancer[72] patients, but it is likely absent from this environment in the healthy state.[73] Based on the known cleavage specificities of these serine proteases, we expected that at least some of the peptide aldehydes would be active against a subset of these proteases. However, in our screen of the peptide aldehyde library against these proteases at inhibitor concentration of 100 μM, there was little inhibition observed (data not shown). Relative to what is known about inhibition of these proteases by peptide aldehydes (Table 4.1), our selected concentration of 100 μM was already much higher than what we predicted would be necessary to observe specific inhibitory interactions. Therefore, we did not reexamine inhibition of these proteases at a higher peptide aldehyde concentration.

From these initial screens, we were able to identify the most active compounds against several human proteases. We subsequently worked with Reaction Biology Corporation to determine IC_{50} values for some of the identified interactions (Table 4.2). This effort confirmed that these compounds demonstrated dose-dependent effects and further highlighted which interactions might be most interesting for further study. Additionally, these experiments revealed new SAR information that could not be obtained from the single concentration screens. For instance, with cathepsin B, the compounds containing aliphatic or aromatic residues in P1 have an order of magnitude lower IC_{50} values than those containing glutamate at this position, confirming what was expected about the selectivity of this protease. Also, the established caspase 1 cleavage preference for valine in P2 is more apparent here, as the IC_{50} value for compound **4.9k**, which contains valine, is an order of magnitude lower than that of compound **4.9j**, which contains leucine in this position.

This IC_{50} experiment also establishes that cathepsin L is by far the most sensitive protease we screened, with IC_{50} values for peptide aldehydes in the low nM (rather than low μM) range. Similar activity differences in peptide aldehyde inhibition of cathepsin L versus other cathepsins have previously been shown.[74] If these compounds are actually produced, they could be produced at low concentrations in the gut, highlighting their potent inhibition of cathepsin L as a potentially interesting activity. Inhibition of cysteine proteases by peptide aldehydes is a well-known phenomenon,[75] and modulation of gastrointestinal protease activity as a therapeutic strategy has also been proposed.[76] However, a 2016 review by Vergnolle concluded that "[i]n the long term, there is a future need to characterize the proteolytic profiles associated with each intestinal disease."[76] If inhibition of any of the proteases we screened here could be identified as particularly important to a disease state, we could rely on their inhibition profiles by this peptide aldehyde library, along with much previous work on their cleavage specificity and known inhibitors (Table 4.1), to design more potent and selective analogues.

There are some important limitations to our identification of human protease targets for these predicted peptide aldehydes. Even if these compounds are actually produced in the gut, the concentrations at which they accumulate are unknown. Though there is some evidence that the cysteine cathepsins that we predict are targets of these molecules may be secreted under certain circumstances, they are mainly known as intracellular proteases.[58] This prompts additional questions of how readily these compounds would enter human cells and what their intercellular concentrations might be. Additionally, as discussed in Chapter 1, the immunoproteasome is another human protease that is potentially important in the gut environment.[77,78] Though these compounds share structural similarities with known 20S proteasome inhibitors (such as Bortezomib, Chapter 1),[79] we did not evaluate them as inhibitors of the immunoproteasome.

Overall, after identifying potential human targets of these compounds, we hypothesized that they might also interact with microbial targets in the gut environment, and we set about evaluating those potential interactions next.

Table 4.2: IC_{50} values for selected peptide aldehyde compounds against several human proteases.

The assays were conducted by Reaction Biology Corporation using fluorogenic protease substrates and measuring increase in fluorescence over time (see Materials and methods for details). Assays were performed in duplicate over threefold serial dilutions from either 0.00051–10 μM (calpain 1, cathepsins B and L) or 0.0051–100 μM (caspase 1). Curves were individually fit to determine IC_{50}. The reported values are the mean and standard deviation of values calculated from these independent series of serial dilutions.

					Calpain 1	Caspase 1	Cathepsin B	Cathepsin L
Compound	**Cluster**	**P3**	**P2**	**P1**	**IC_{50} (μM)**	**IC_{50} (μM)**	**IC_{50} (μM)**	**IC_{50} (nM)**
4.8r	bgc52	Leu	Leu	Leu-H				0.53 ± 0.04
4.8l	bgc52	Leu	Leu	Tyr-H	0.54 ± 0.04			0.67 ± 0.05
4.5p	bgc 38	Hexanoyl	Tyr	Ala-H			0.56 ± 0.06	0.80 ± 0.05
4.8a	bgc52	Ala	Val	Phe-H	0.28 ± 0.03		0.50 ± 0.06	1.1 ± 0.19
4.8f	bgc52	Leu	Leu	Phe-H	1.0 ± 0.09			1.6 ± 0.33
4.9h	bgc45	Hexanoyl	Leu	Glu-H			0.29	1.7 ± 0.03
4.8q	bgc52	Leu	Val	Leu-H				1.7 ± 0.34
4.5q	bgc 38	Decanoyl	Tyr	Ala-H			0.14 ± 0.01	1.9 ± 0.43
4.5n	bgc 38	Decanoyl	Leu	Ala-H	0.47 ± 0.03			5.7 ± 1.6
4.5m	bgc 38	Hexanoyl	Leu	Ala-H				6.4 ± 0.7
4.5d	bgc 34	Hexanoyl	Tyr	Gly-H				11 ± 0.04
4.5e	bgc 34	Decanoyl	Tyr	Gly-H				12 ± 4.8
4.8k	bgc52	Leu	Val	Tyr-H	0.63 ± 0.03		0.16	
4.8e	bgc52	Leu	Val	Phe-H	0.74 ± 0.1		0.21 ± 0.01	
4.9l	bgc45	Octanoyl	Leu	Glu-H		29 ± 0.81	2.6 ± 0.01	
4.9i	bgc45	Hexanoyl	Val	Glu-H			5.8 ± 0.06	
4.9k	bgc45	Hexanoyl	Val	Asp-H		5.3 ± 0.16		
4.9j	bgc45	Hexanoyl	Leu	Asp-H		51 ± 0.25		
4.8g	bgc52	Ala	Val	Tyr-H	0.92 ± 0.06			
4.8h	bgc52	Ala	Leu	Tyr-H	0.96 ± 0.06			
4.8b	bgc52	Ala	Leu	Phe-H	2.7 ± 0.05			

4.2.4. Evaluating antibiotic activity of putative gut microbial peptide aldehydes against gut commensals and pathogens

Over one thousand bacterial strains have been identified as members of the commensal human gut microbiota, and the gut microbiota contains over two orders of magnitude more genes than the human genome.[80–82] Therefore, discovery of the microbial proteases that are potentially targeted by our peptide aldehydes requires a different approach than simply screening against panels of microbial proteases. As discussed in Chapter 1, although the functional roles of some gut microbial proteases have been investigated, there remains a great deal to be discovered about this class of enzymes.[83,84] As a rapid way to evaluate the response of select gut microbial strains to peptide aldehydes, we decided to first screen these compounds for antibiotic activity. If an organism demonstrated a specific response to a peptide aldehyde at a physiologically relevant concentration, we would then seek to identify the molecular interaction responsible for that effect. In addition to discovering and prioritizing peptide aldehyde-gut microbe interactions, this method could also reveal new essential proteases in microbes that could be useful targets for antibiotic development.

The idea that modulating protease activity in microbes could serve a therapeutic purpose has recently attracted interest.[85,86] There are several known examples of antibiotics that interact with proteolytic complexes, but to our knowledge, bacterial growth inhibition would be a novel bioactivity for a peptide aldehyde compound. A classic example of this phenomenon can be found in studies of the Clp protease complex (reviewed by Rubin and coworkers[85] and by Wright and Culp[86]). This complex, which is widely conserved in bacteria, consists of ClpP subunits, which have serine protease activity, and ATPases, which unfold proteins and deliver them to ClpP for hydrolysis. Different bacterial species contain various ATPases as part of this system.

Notably, in *Mycobacterium*, the genes involved in forming this complex are essential for growth.[87] The natural products cyclomarin[88] and ecumicin[89] exhibit bactericidal activity by targeting the chaperone in this complex, ClpC1, in *M. tuberculosis*. A 2015 study by Cho and coworkers revealed that ecumicin prevents association of ClpC1 with ClpP, disrupting ClpP function and potentially leading to cell death due to the accumulation of toxic proteins.[89]

Similar compounds, known as the acyl depsipeptide antibiotics (ADEPs), have been investigated as antibiotics in *Escherichia coli* and *Bacillus subtilis* and were found to target the Clp-associated ATPases in these species.[90] Though the ClpP protease and its chaperones are not essential in these organisms, these compounds still exhibit antibiotic activity.[90] This phenomenon was explained in a 2009 study by Turgay and coworkers, which showed that ADEPs function both by preventing association of ClpP with its chaperones and by activating ClpP to carry out unregulated and destructive intracellular proteolysis.[91] Therefore, there are at least two hypothesized mechanisms by which interfering with proteases can bring about antibiosis: either loss of protease activity leads to the accumulation of toxic proteins, or uncontrolled protease activity destroys cellular components.[86] To our knowledge, there are no known examples of small molecules that kill bacteria directly by competitively inhibiting a protease, but there is at least one example of a proteinaceous plant protease inhibitor displaying antibiotic activity against several microbes.[92]

As an initial selection of species for antibiotic screening, we chose to focus on the ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumanii*, *Pseudomonas aeruginosa*, and *Enterobacter* sp.). This choice was inspired by Brady and coworkers, who screened bioinformatically predicted and synthesized NRPS products from a large variety of bacterial species for antibiotic activity against these organisms in a 2017 study.[93]

These six pathogens are a common cause of hospital acquired infections and have developed resistance to many antibiotics.[94] We worked with a contract research organization to screen our 48 compound peptide aldehyde library for antibiotic activity against these organisms. These experiments were performed by iFyber (Ithaca, NY) using the agar overlay method.[95] Briefly, soft agar plates inoculated with the pathogen of interest were prepared. These plates were then spotted with the dose series of each peptide aldehyde (10 mM, 3.33 mM, and 1.11 mM) and incubated overnight. After overnight growth, the plates were examined for zones of inhibition (ZOI).

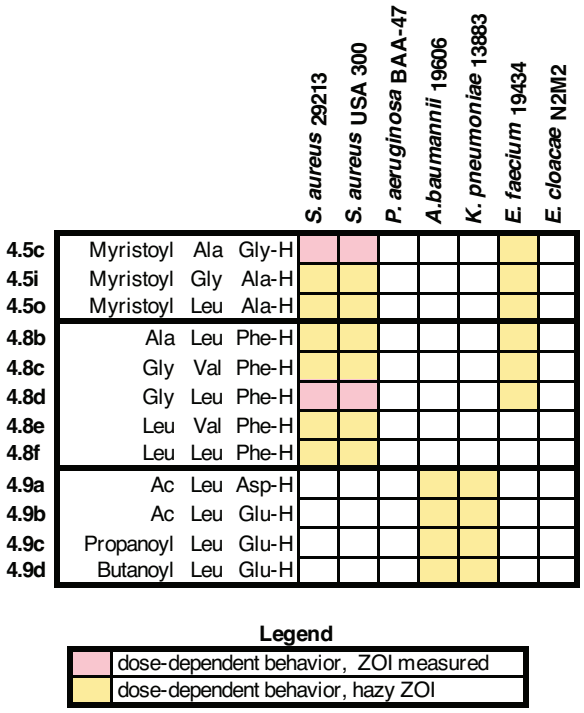


Figure 4.15: Peptide aldehydes with antibiotic activity against the ESKAPE pathogens.

Peptide aldehydes that exhibited either full or partial growth of at least one of the ESKAPE pathogens are shown.

In these experiments, only a fraction of compounds in the library (12/48) demonstrated any antibiotic activity against these pathogens. In most of these cases, only a hazy ZOI was observed

even at the highest concentration tested. This may indicate either that these compounds are not potent antibiotics or that they do not remain stable over time in this assay format. However, this screen did reveal a small subset of compounds with some antimicrobial activity and two compounds, **4.5c** and **4.8d**, that fully inhibited growth of two *S. aureus* strains at the highest concentration tested. An interesting SAR can be observed for compound **4.8d** and its close homologs: four of the five compounds that differ from this compound in the P2 residue, the P3 residue, or both (**4.8b, 4.8c, 4.8e, 4.8f**) demonstrate a less severe but still observable phenotype. This increases our confidence that this may be a specific interaction. Along with the compounds that are active against *S. aureus* strains, we also observed that a different set of compounds, acidic peptide aldehydes **4.9a–e**, demonstrated weak growth inhibition against the Gram-negative species *A. baumanii* and *K. pneumoniae*. Notably, this activity is restricted to acidic peptide aldehydes containing short *N*-acyl chains.

In addition to the screen of the ESKAPE pathogens, we also investigated the antibiotic activity of peptide aldehydes against common gut commensals and pathogens. Again inspired by Brady and coworkers, in their work on humimycin A (Chapter 1),[3] we selected a panel of species for screening with the goal of achieving broad representation of the gut microbiota on the phylum level (Table 4.3). We performed our own zone of inhibition screen against 7 obligate anaerobes and facultative anaerobes. In this experiment, the compounds were tested at a single concentration (10 mM) in duplicate. Surprisingly, a majority of compounds in our library (31/48) showed weak activity against at least one strain in these experiments (Figure 4.16). There was little discernible pattern to distinguish inhibition of growth of Gram-negative species vs. Gram-positive species or anaerobic species vs. aerobic species. However, compounds **4.5c** and **4.8d** (and its near structural homologs) were active against many different species. These are the same

two compounds that were identified as most active in the ESKAPE pathogen screen. Additionally, one of the compounds that showed activity against the Gram-negative ESKAPE pathogens (**4.9a**) also showed specific activity here against Gram-negative *E. coli* MS200-1.

Table 4.3: Characteristics of gut microbes screened for antibiotic activity.

Strain	Phylum	Gram stain	Atmosphere
Bacillus subtilis 168	Firmicutes	positive	aerobic
Clostridioides difficile 630Δerm	Firmicutes	positive	anaerobic
Staphylococcus epidermidis ATCC 12228	Firmicutes	positive	aerobic
Bacteroides dorei CL02T12C06	Bacteroidetes	negative	anaerobic
Bacteroides fragilis ATCC 25285	Bacteroidetes	negative	anaerobic
Parabacteroides merdae ATCC 43184	Bacteroidetes	negative	anaerobic
Escherichia coli MS 200-1	Proteobacteria	negative	aerobic

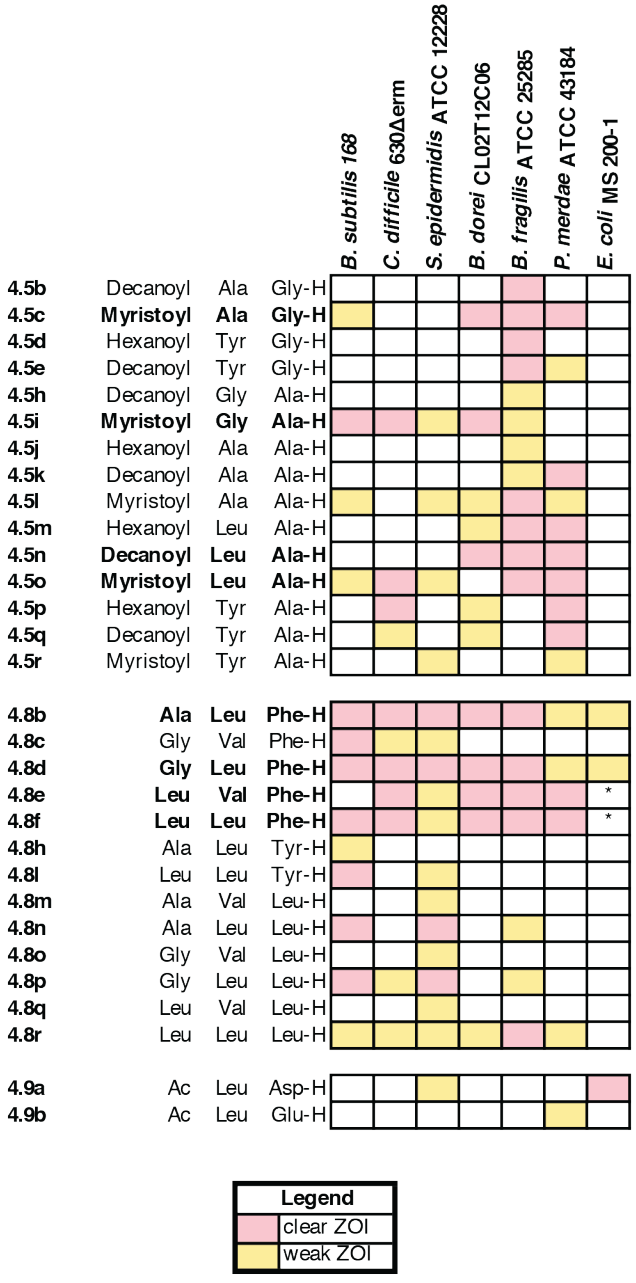


Figure 4.16: Antibiotic activity of peptide aldehydes against human gut bacteria.

(A) Aliphatic *N*-acyl dipeptide aldehydes. (B) Tripeptide aldehydes. (C) Ruminopeptin analogues. * = rather than a ZOI, a white spot on the plate was observed.

Though these results indicated a potential biological role for these compounds, the very high concentrations at which they were screened limit the conclusions we can draw from this experiment. In order to interrogate the potential physiological significance of these interactions, we next determined minimum inhibitory concentrations for select compounds and species. We used a standard laboratory protocol to determine MICs for some of the compounds that had shown clear zones of inhibition.[96] We evaluated eight compounds (**4.5c, 4.5i, 4.5n, 4.5o, 4.8b, 4.9d, 4.8e, 4.8f**), which had all shown clear ZOIs against at least three microbial strains, against one aerobic strains and three anaerobic strains, and we also evaluated compound **4.9a** against *E. coli* MS 200-1 (Figure 4.17). Overall, this screen revealed that most of these interactions had MIC values of 32 μg/mL or greater. However, one compound, **4.5o**, which is structurally related to the compound that was active against *S. aureus* in the ESKAPE pathogens screen, had an MIC of 8 μg/mL against *B. fragilis* and *P. merdae*.

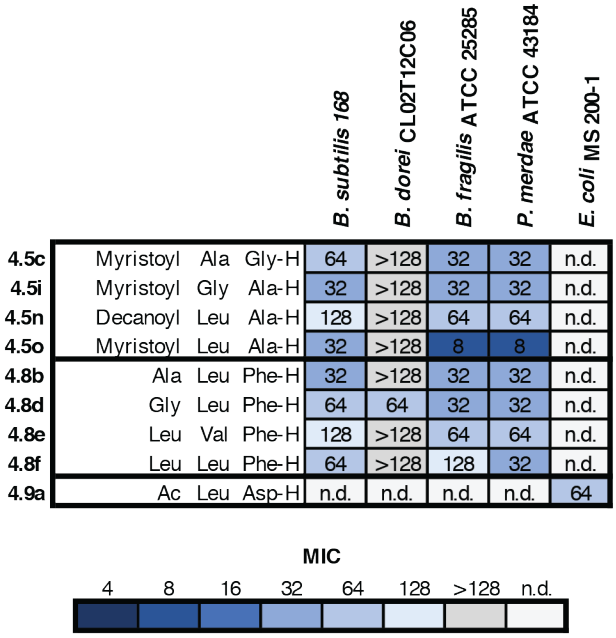


Figure 4.17: MIC determination of peptide aldehyde antibiotic activity.

The minimum inhibitory concentration (MIC) of eight compounds was determined against four strains, and the MIC of one compound was determined against *E. coli* MS 200-1 (n.d. = not determined).

We cannot predict the concentrations at which these compounds may be produced in the gut, and to our knowledge there is no established standard for what MICs are indicative of a physiological effect in a natural microbial environment. However, based on comparison with the literature, an MIC of 8 μg/mL may be low enough to warrant further investigation of compound **4.5o**. Common "useful" antibiotics typically have MIC values from less than 10 μg/mL to much less than 1 μg/mL.[97] For comparison, humimycin A is a known lipopeptide antibiotic with MIC as low as 8 μg/mL against certain methicillin-resistant *S. aureus* strains.[3] Lactocillin, which was isolated from the vaginal microbe *Lactobacillus gasseri*, has reported MIC values of 42–425 nM against a variety of oral and vaginal microbial species (corresponding to 0.051 – 0.514 μg/mL).[1] Ruminococcin A, a lantibiotic produced by gut microbe *Ruminococcus gnavus*, has reported

MIC values of 32.5–600 μg/mL against a variety of gut microbial species.[98] Based on the analogous structures present in our library, we have already revealed certain structural features that appear to be important for the activity of **4.5o**. Namely, this compound is more active than analogues that contain the same dipeptide scaffold with shorter acyl chains, **4.5m** and **4.5n**, and compounds that share only one of its amino acid residues, such as **4.5i**. However, it would be interesting to determine if the aldehyde functional group is also required for activity. We could also attempt to raise mutants resistant to this compound and then sequence them to determine what protein target(s) might be responsible for this antibiotic phenotype. Though antibiotic effects are one possible mechanism by which these putative natural products could interact with their environment, we were also interested in examining their non-antibiotic effects. We accomplished this by evaluating their ability to inhibit secreted protease activity from gut microbial strains.

4.2.5. Putative gut microbial peptide aldehydes inhibit gut microbial secreted protease activity

We hypothesized that peptide aldehydes might also interact with secreted proteases in the gut environment. As discussed in Chapter 1, there remains a lack of comprehensive knowledge about both the major contributors of proteolytic activity in the commensal gut microbiota and the physiological functions of known secreted gut microbial proteases.[84] We envisioned leveraging our peptide aldehyde library to simultaneously address two questions about proteases in the human gut microbiota. Firstly, we were interested in identifying potential secreted gut microbial protease targets for the peptide aldehydes. Secondly, we anticipated that the inhibition profile of

the peptide aldehyde library against a microbial strain could rapidly reveal information about the substrate specificities of its secreted proteases (Figure 4.18A).

We began by identifying several gut commensal and pathogenic species that had previously been reported to secrete proteases.[99–102] We used an assay based on fluorescein thiocarbamoyl-casein (FTC-casein) to measure secreted protease activity in cell-free culture supernatants from four strains: *Enterococcus faecalis* TX0104, *Clostridium sporogenes* ATCC 15579, *Klebsiella aerogenes* ATCC 13048, and *Bacillus cereus* ATCC 53522.[103] We assessed both overnight and mid-log phase cultures and determined conditions for which we could observe significant secreted protease activity (overnight cultures for *B. cereus*, *C. sporogenes*, and *E. faecalis*, and mid-log phase cultures for *K. aerogenes*.)

After this baseline determination, we screened the peptide aldehyde library for their ability to inhibit protease activity in culture supernatants from these four strains. Compounds were added to the supernatants at 100 μM and preincubated for 10 minutes at 37 ºC. The FTC-casein substrate was then added and fluorescence measured over 20 min at 37 ºC. Percent inhibition was calculated by comparing initial rates with a negative (no inhibitor) control. As this was merely a pilot experiment to identify potential hits for follow up study, replicate experiments were not performed. In these data, it is not meaningful to compare the magnitude of effects observed against different species, as we have not attempted to normalize initial protease activity across the different strains screened. *C. sporogenes*, for instance, demonstrated much greater raw secreted protease activity than any of the other strains.

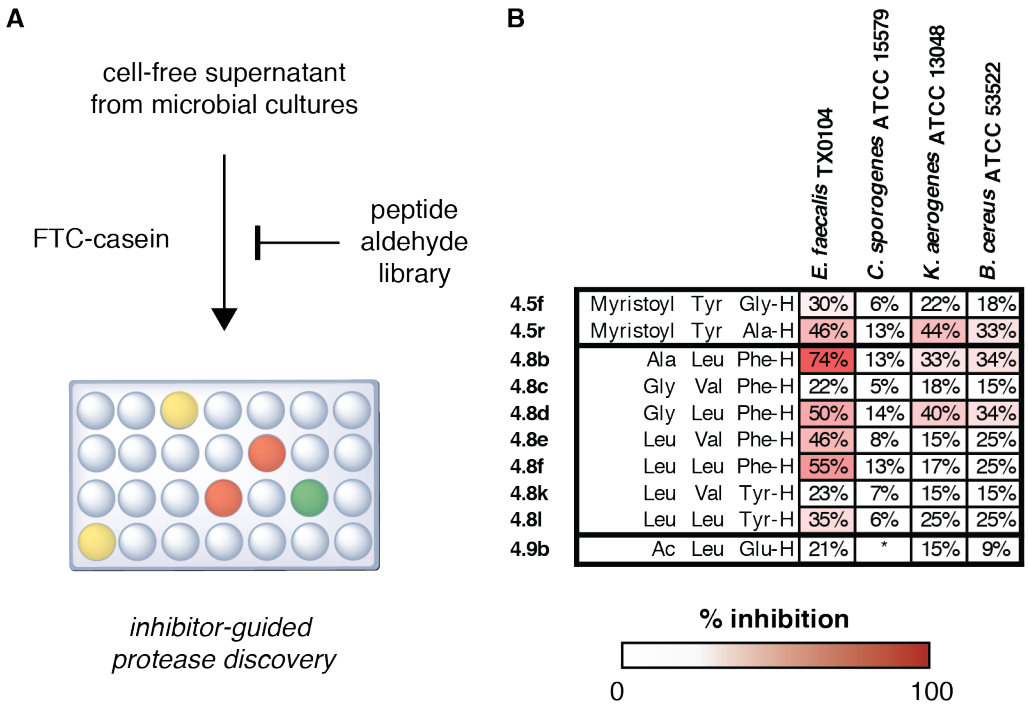


Figure 4.18: Peptide aldehydes inhibit the secreted protease activity of gut microbes.

(A) Experimental design. Microbial supernatants, preincubated with peptide aldehydes from the library, were exposed to a generic protease substrate (FTC-casein), generating a "heat map" of secreted protease-inhibitor interactions. (B) Ten compounds showed inhibiton of secreted protease activity in four microbial strains. * = calculated value was <0% inhibition. See Materials and methods for assay details.

From this screen, we observed that 10 compounds display significant inhibition of secreted protease activity (>20% inhibition) in at least one of the strains (Figure 4.18B). Interestingly, the best inhibitors appear to be active against many species. Several patterns can be discerned in these data. Among the aliphatic *N*-acyl dipeptide aldehydes, the only compounds meeting the activity threshold (**4.5f** and **4.5r**) contain the myristoyl-L-tyrosine motif, and these two compounds are the only ones in this library that contain this motif. Within this family, the next highest inhibition values are for compounds that contain a decanoyl-L-tyrosine motif, **4.5q** and **4.5e**, which showed 11% and 8% inhibition in *E. faecalis*, respectively (data not shown). Among the tripeptide aldehydes, most of the compounds with phenylalanine in position P1 demonstrate

strong inhibitory activity, while the only tyrosinyl aldehydes that meet the activity threshold are those with leucine in P3. Though we cannot conclude that these distinct compounds are interacting with the same proteases within each species or with homologous proteases among different species, these results suggest that these organisms may produce secreted proteases with cleavage preferences for aromatic residues either in P1 or P2.

In order to predict what proteases these peptide aldehydes could be inhibiting to yield these effects, we examined the literature on secreted proteases from these species. We have already discussed the roles of *E. faecalis* secreted proteases at length in Chapter 3. Though the TX0104 strain used in these experiments is distinct from the OG1RF and V583 strains that are discussed in Chapter 3, it does also contain SprE and GelE. To our knowledge, SprE is the only secreted serine protease of *E. faecalis* that has been reported, and there are no known secreted cysteine proteases from this organism.[99] It is possible that the effect observed in *E. faecalis* by **4.9b** is due to inhibition of SprE, and we could test this in vitro (either by using SspA as a proxy or by heterologously expressing and purifying SprE). It may also be worth evaluating the rest of these compounds as inhibitors of SprE, though it would be somewhat surprising if these aliphatic and aromatic aldehydes were found to inhibit this glutamyl endopeptidase.

The caseinolytic activity of *C. sporogenes* was first identified in 1975, and this is also one of the species that was studied by MacFarlane and coworkers.[104–106] In 1992, they identified at least 6 individual proteases in cell-free supernatants from this organism and used specific inhibitors to classify them. This analysis suggested that all the identified enzymes were at least partially dependent on metals, but that some may also have been cysteine and serine proteases.[106] Though this species is not itself considered a pathogen, it is often found in host infection sites, and it has been proposed that protease production by this species helps to mediate infection by other

bacteria.[107] Therefore, an interesting future direction here may be to determine which specific secreted protease(s) from this species are interacting with these peptide aldehydes. We could accomplish this by heterologously expressing and purifying the proteases that were identified by Macfarlane and coworkers[106] and evaluating their inhibition by peptide aldehydes in vitro.

In a 2011 study, *K. aerogenes* (previously known as *Enterobacter aerogenes*) was identified as producing a secreted protease that hydrolyzes casein,[100] but to our knowledge there have been no additional published investigations on the secreted proteases of this organism. Several secreted serine proteases have been identified from different strains of *B. cereus*, including human opportunistic pathogens.[102,108,109] The particular strain that we investigated in this work (UW85) has not been specifically studied in this context and is not predicted to be found in the human gut. However, it has recently had its genome sequenced,[110] and there are many annotated proteases in its genome. Identification of protease targets of these peptide aldehydes in these two strains may first require additional investigations to determine their most significant secreted proteases. This could be accomplished by using gel zymography to detect protease activity against various substrates (e.g. casein, gelatin).[111] Additionally, it would be interesting to analyze what catalytic types of proteases are present in each of these supernatants by treating them with specific generic reference inhibitors for individual protease classes.[112]

In summary, these experiments have revealed that several gut microbial strains secrete proteases that can be inhibited by compounds in our peptide aldehyde library, primarily those containing a myristoyl-L-Tyr motif or a phenylalanyl aldehyde. Discovery of the targets of small molecule inhibitors in the complex environment of the gut microbiota and elucidating the biological relevance of such interactions is complicated by the huge number of microbial species and genes present in this environment and the still very limited ways to model these interactions

in a laboratory setting.[113,114] Though we can show that our peptide aldehydes can inhibit some secreted proteases in these organisms, there is not a simple way to prioritize the most significant of these interactions for further study. In our final approach for identifying targets of these compounds, we approached this challenge in a different way, by synthesizing activity-based probes and using them to identify target proteins in a prominent gut microbial pathogen in an untargeted fashion.

4.2.6. Activity-based protein profiling (ABPP) in a gut microbial pathogen using a peptide aldehyde mimetic

Our screens for activities of peptide aldehydes described above were limited to obvious observable phenotypes. However, there may be additional types of proteins with which these compounds can interact and additional phenotypes that they may modulate. In order to interrogate how these compounds interact with microbes on a proteome-wide scale, we employed an activity-based protein profiling (ABPP) approach.[115] This technique works by using a reactive probe molecule to label and enrich target proteins from a proteome of interest. The proteins targeted by the probe can then be identified with LC-MS and this information used to generate hypotheses about additional activities that these compounds may demonstrate in vivo.[115] ABPP is a popular technique for studying protease activity in a variety of species,[116] and its application to microbial pathogens is a growing area of interest.[117–119] To start, we conducted these experiments with *Clostridioides difficile* 630Δerm, a common laboratory strain of this human pathogen.[120]

The design and execution of these experiment were performed in collaboration with Steven Carr, Sam Myers, and Deepak Mani (Broad Institute Proteomics Platform, Cambridge, MA). I

performed the probe synthesis, gel-based probe validation, and biotin labeling of samples described in the following paragraphs, and they performed the trypsin digestions and LC-MS analysis (see below). We began by designing and synthesizing two activity-based probes for labeling protein targets of peptide aldehydes. As electrophilic peptide aldehydes form reversible covalent adducts with nucleophilic residues in the active sites of their target proteins, they are therefore not suitable activity-based probes.[121,122] Instead, we used probes containing electrophiles that would form irreversible covalent adducts but mimic the structures of our peptide aldehydes. As we had previously shown that these peptide aldehydes were most potent as inhibitors of human cysteine proteases, we chose to incorporate a reactive halomethyl ketone electrophile. Halomethyl ketone moieties are commonly used to covalently label reactive cysteines,[74,123] and this functional group was relatively simple to install synthetically. One probe was designed to mimic the L-leucyl-L-alanyl aldehyde scaffold found in several compounds (such as **4.5m**) that showed broad activity against many human proteases and microbial strains. A second probe was designed to mimic the ruminopeptin analogue **4.9h**, which was the focus of the work described in Chapters 2 and 3 and is the structure we have the highest confidence in based on biochemical analysis of the *rup* cluster.

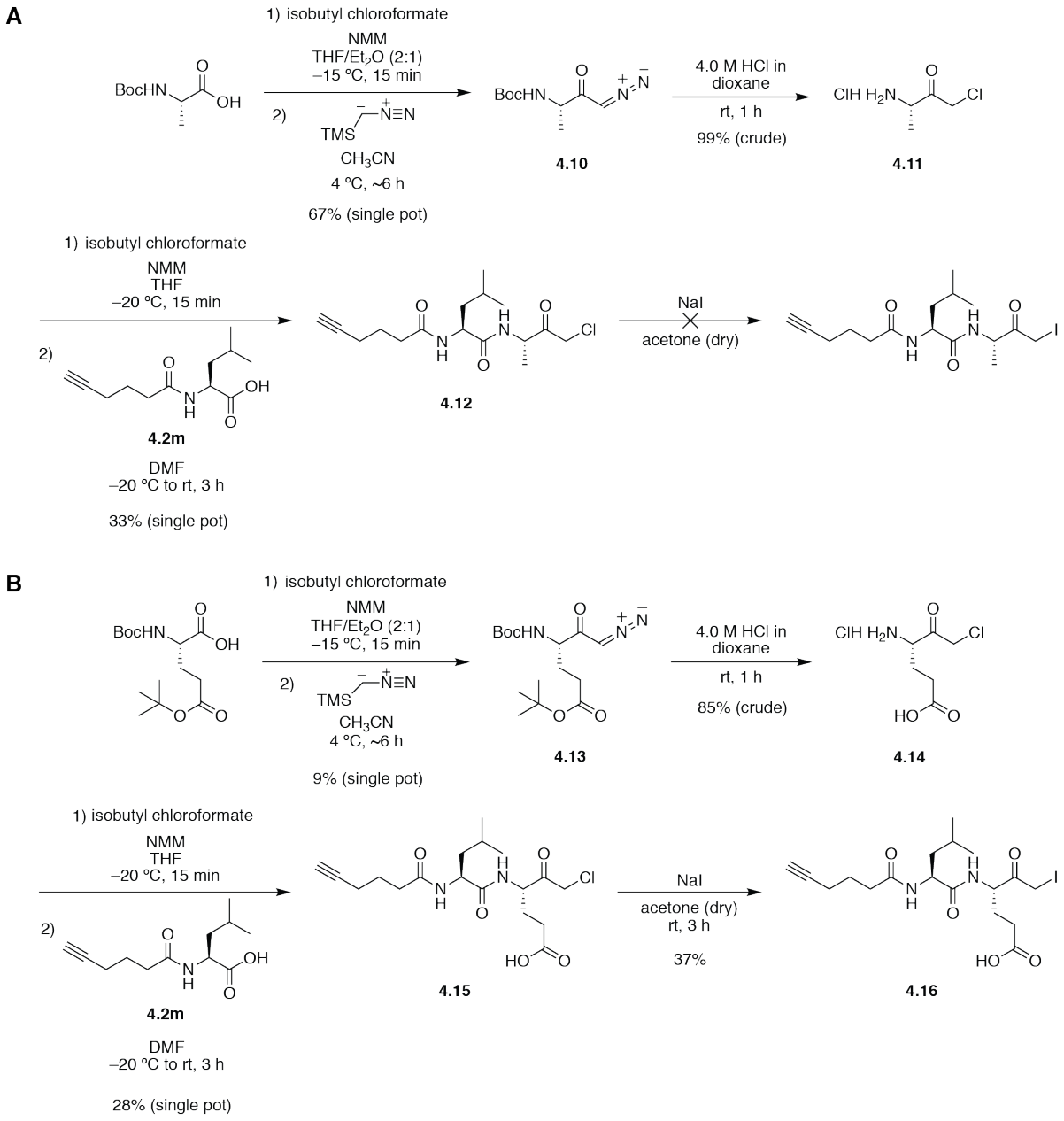
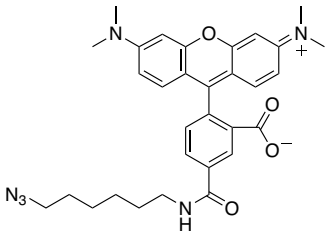


Figure 4.19: Synthesis of activity-based probes for peptide aldehyde target identification.

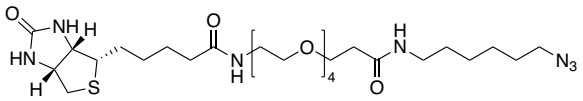
(A) Synthesis of Leu-Ala-chloromethyl ketone probe **4.12**. (B) Synthesis of Leu-Glu-iodomethyl ketone probe **4.16** (NMM = *N*-Methylmorpholine).

Synthesis of the probes relied upon the modified Nierenstein reaction of carboxylic acids with TMS-diazomethane to generate α-diazoketones **4.10** and **4.13** followed by conversion of

these intermediates to chloromethyl ketone analogues of L-alanine and L-glutamate (**4.11** and **4.14**).[124,125] These amino acid mimetics were then coupled to hex-5-ynyl-L-leucine using standard peptide coupling chemistry as previously reported[126] to afford the *N*-acylated dipeptide chloromethyl ketones **4.12** and **4.15**. We attempted to use the Finkelstein reaction with sodium iodide to convert the chloromethyl ketones (cmk) to the corresponding iodomethyl ketones (imk).[127] Oddly, while this transformation was successful for the glutamyl probe, generating **4.16**, we could not identify conditions to convert the alanyl probe **4.12** to its iodomethyl ketone analogue despite screening time and temperature and preparing dry reagents. Though we hypothesized that **4.12** might be a less efficient probe than the desired iodomethyl ketone, we decided to move forward with both compounds.



4.17



4.18

Figure 4.20: Azide reagents used for click chemistry.

The azide reagents used in this work were tetramethylrhodamine (TAMRA) azide (Invitrogen, **4.17**). and PEG4 carboxamide-6-azidohexanyl biotin (Biotin Azide, Invitrogen, **4.18**).

With the probes in hand, we performed an initial experiment to validate that they could label cysteine proteases in vitro. We incubated purified human calpain, which was weakly inhibited by both of the corresponding peptide aldehydes, with probes **4.12** and **4.16** at 100 μM concentration, performed a copper-catalyzed Huisgen azide-alkyne 1,3-dipolar cycloaddition (click reaction) with tetramethylrhodamine (TAMRA) azide, **4.17**, to append the alkyne-labeled

protease with a fluorescent tag, and then separated and visualized these fluorescent bands by SDS-PAGE. In this experiment, we observed weak labeling of calpain by **4.12** and strong labeling by **4.16** (Figure 4.21).

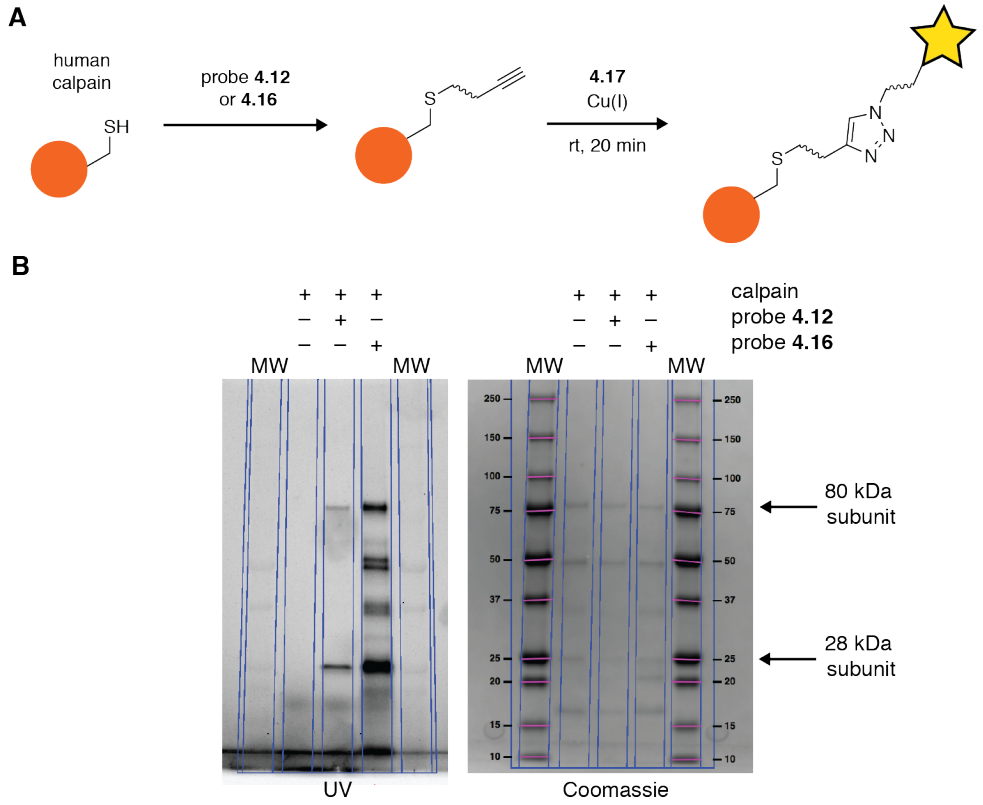


Figure 4.21: Validation of activity-based probes by labeling human calpain.

(A) Workflow for labeling human calpain with halomethyl ketone probes **4.12** and **4.16** and appending the TAMRA fluorophore for analysis by SDS-PAGE. Cu(I) was generated in situ from $CuSO_4$ using the proprietary reductant from the Click-iT Protein Reaction Buffer Kit (Invitrogen). See Materials and methods for details. (B) Results of the labeling experiment (left = UV image, right = Coomassie stain). Calpain contains an 80 kDa subunit and a 28 kDa subunit, which can be observed on the gel, along with other bands which may be impurities or products of the protease hydrolyzing itself.

Having validated that the two probes could label a purified cysteine protease, we next tested whether the probe could also effectively label proteins in the more complex setting of a bacterial

proteome. For our initial experiment, we generated lysates of *C. difficile* and treated these lysates with the probes (at 10 μM concentration). The labeled proteomes were then tagged with **4.17** using the click reaction and visualized by SDS-PAGE. Our initial negative control for this experiment was to pre-treat *C. difficile* proteomes with the peptide aldehyde inhibitors that had inspired the design of the halomethyl ketone probes (at 100 μM concentration) before labeling (Figure 4.22). Presumably this would identify non-specific interactions of the halomethyl ketone probe, as peptide aldehyde pretreatment would block labeling of targets that specifically recognized elements of the probe scaffold.[115] In this experiment, it did not appear that the pre-treatment strategy had succeeded in blocking any labeling by the probes (data not shown). Furthermore, we could not observe any bands labeled by probe **4.12** under these experimental conditions. Therefore, we decided to move forward with only probe **4.16** and to pursue an alternative negative control strategy.

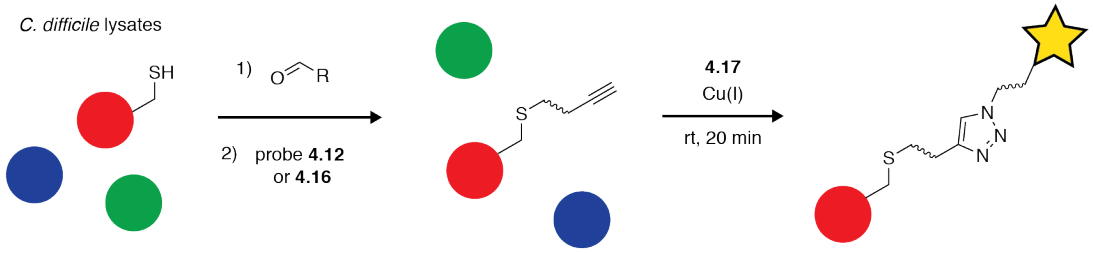


Figure 4.22: Initial design for activity-based protein profiling experiments with in-gel fluorescence profiling.

Workflow for pre-treatment of *C. difficile* lysates with aldehydes, labeling of pre-treated lysates with halomethyl ketone probes, and appending the TAMRA fluorophore for detection by SDS-PAGE.

We next employed an alternative negative control strategy. In this experiment, we treated independent samples with either **4.16** or a generic iodoacetamide(IA)-alkyne probe, **4.19**.[128] Our

goal was to identify gel bands that were preferentially labeled by **4.16**, indicating a specific interaction. We indeed observed different labeling patterns between the generic probe and the specific probe (Figure 4.23). It was not immediately obvious if there were any bands labeled exclusively with **4.16**, though there did appear to be bands which were labeled exclusively with **4.19**. We expected that biotin enrichment and LC-MS analysis might reveal additional differences that were not visible on the gel. In retrospect, these failures to identify abundant clear targets of the specific probe in *C. difficile* lysates by SDS-PAGE analysis may have indicated that we were also unlikely to discover interesting targets by LC-MS-based peptidomics.

To further analyze these labeled proteomes by tagging with biotin, enrichment, and LC-MS identification of targets, we used two different workflows for sample preparation and analysis (Figure 4.24). The first workflow (A) was designed to give information about the proteins enriched by the specific probe over the generic probe. In this workflow, proteins are labeled with the alkyne tagged iodomethyl ketone probes, a biotin azide reagent is appended to these probes using the click reaction, and tagged proteins of interest are enriched using streptavidin beads. The bead-bound protein samples are then digested with trypsin, individual channels are labeled with isotopic tags, and the combined samples are analyzed by LC-MS.[129] The second workflow (B) was intended to give information about the specific amino acid residues labeled by the probes. In this workflow, proteins are also labeled with the alkyne-tagged probes and then subjected to the click reaction to label them with a biotin azide reagent. These tagged samples are digested with trypsin, and the individual biotin-tagged peptides are then subjected to anti-biotin antibody enrichment. Analysis of these samples by LC-MS/MS was intended to reveal which sites within the proteins of interest were labeled by the probes.[130]

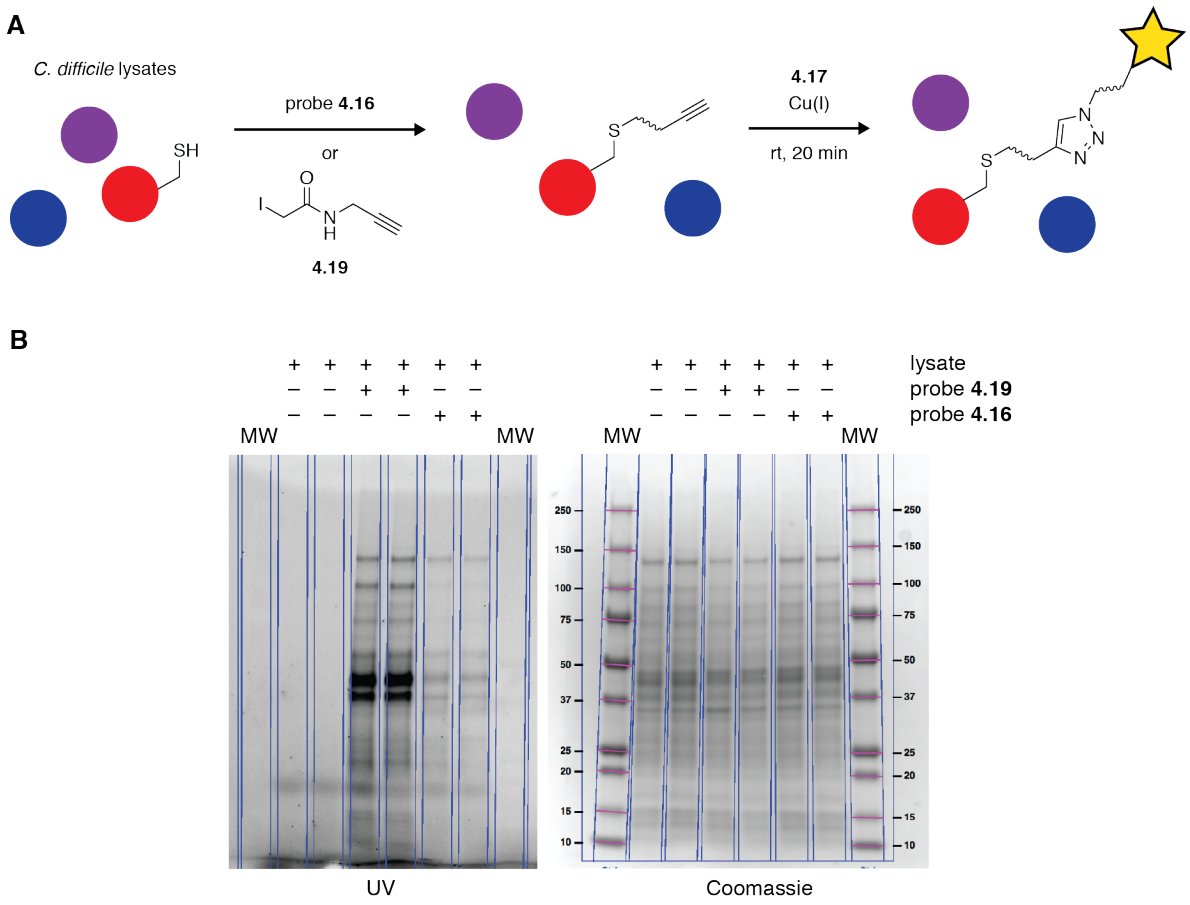


Figure 4.23: Alternative design for activity-based protein profiling experiments with in-gel fluorescence profiling.

(A) Workflow for labeling of *C. difficile* lysates with generic and specific iodomethyl ketone probes and appending the TAMRA fluorophore for detection by SDS-PAGE. (B) Results of the labeling experiment (left = UV image, right = Coomassie stain).

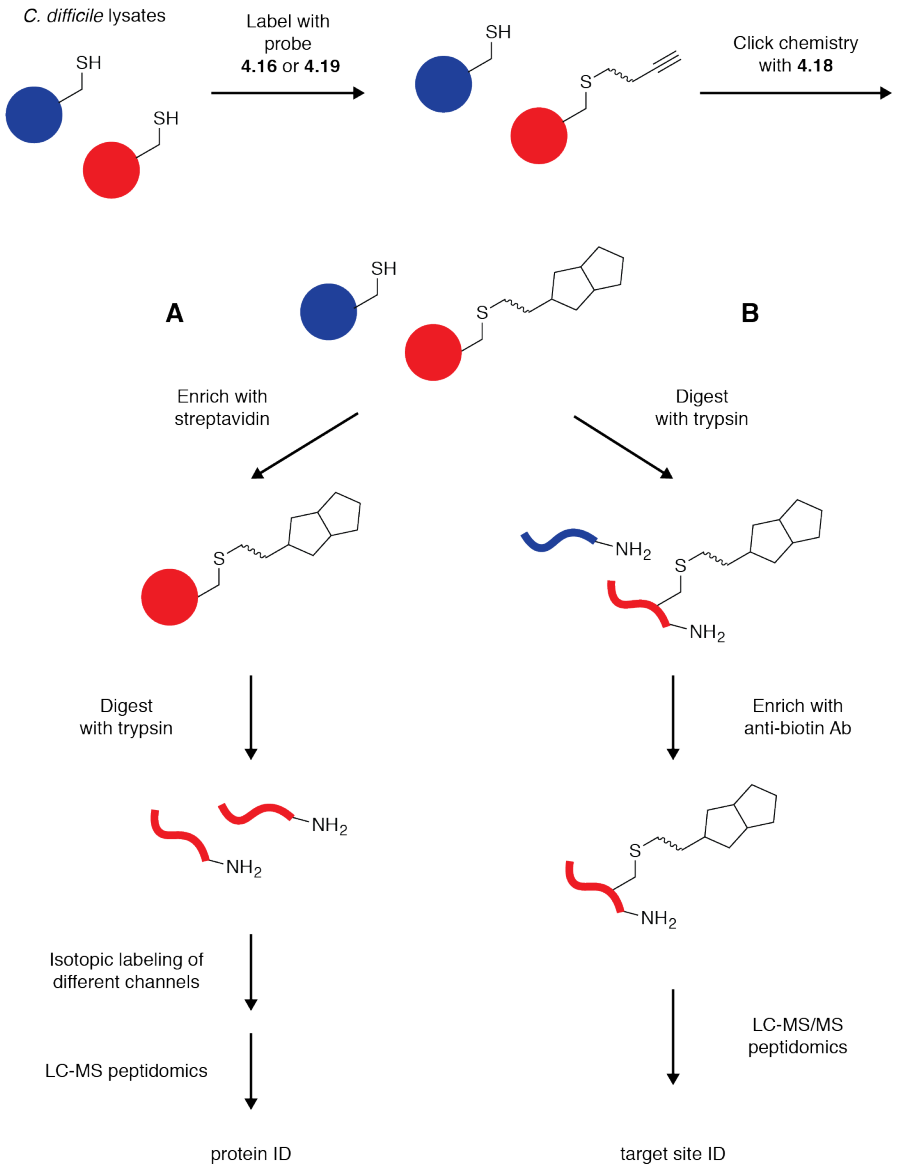


Figure 4.24: Workflows for preparation and analysis of labeled proteomes by LC-MS.

After appending the biotin tag to the alkyne-labeled proteomes, two different workflows were used to analyze these samples. (A) The goal of workflow A was to obtain protein-level information about targets of the probes. Biotinylated proteins were enriched on streptavidin beads and digested with trypsin. The resulting peptides from different samples were labeled with isobaric tags and pooled for analysis by LC-MS. (B) The goal of workflow B was to identify the specific amino acid residues modified by the probes. Biotinylated peptides were digested with trypsin and then subjected to anti-biotin antibody enrichment. These enriched peptides were directly identified by LC-MS.

To validate this experimental strategy and confirm that we would enrich sufficient quantities of protein in this experiment for visualization by LC-MS, I conducted a pilot experiment of workflow A by labeling proteomes with **4.16**, modifying them with biotin using the click reaction with **4.18**, enriching the biotin-tagged proteomes with streptavidin-agarose beads,[123] and observing them by SDS-PAGE with silver staining. In this experiment, we labeled 200 μg of each sample. This experiment revealed that we were indeed able to enrich protein bands on the order of 10 ng each, which we expected would be sufficient for observation by LC-MS (Figure 4.25).

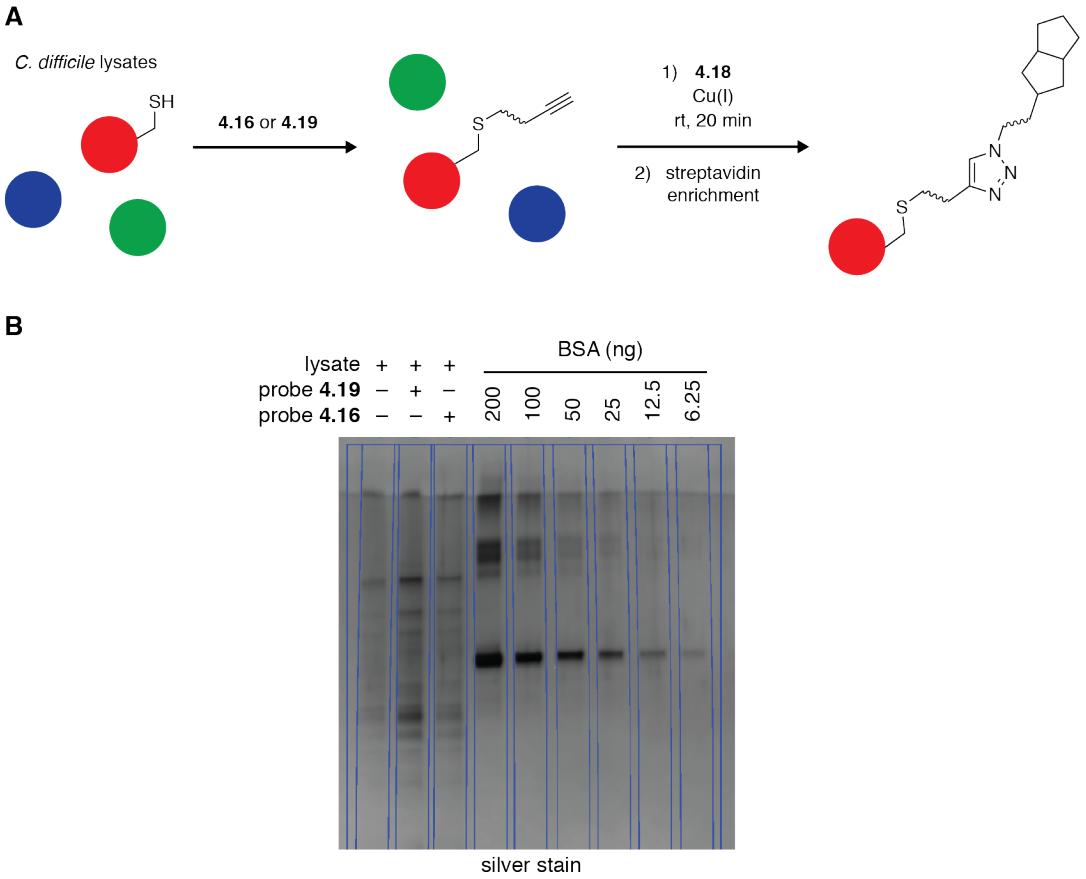


Figure 4.25: Enrichment of biotin-tagged proteins in *C. difficile* 630Δerm and visualization by SDS-PAGE and silver staining.

(A) Workflow for labeling of *C. difficile* lysates with generic and specific iodomethyl ketone probes, appending the biotin affinity tag, enrichment on streptavidin beads, and analysis by SDS-PAGE and silver staining. (B) Results of the enrichment experiment (silver stained gel, BSA = bovine serum albumin).

We then proceeded to analyze identically enriched samples by LC-MS. For workflow A, I performed the bead-bound enrichment of samples and then transferred them to Sam Myers and Deepak Mani, who performed the trypsin digestion and LC-MS analysis. The enriched protein samples were digested with trypsin, and the resulting peptide fragments were labeled with isobaric tags specific for each condition. The samples were then pooled and analyzed by LC-MS, and peptides identified by searching the *C difficile* 630 proteome database (UP000001978).[129] Sam Myers generated the plot of enriched targets, and we discussed the results together.

It must be noted that the proteome database used here does not precisely correspond with the strain that we used for these experiments. *C. difficile* 630 is a multidrug-resistant clinical isolate,[131] and *C. difficile* 630Δerm is a derivative of this strain that is sensitive to erythromycin.[132] A recent reannotation and comparison of the genome sequences of these two strains by Schomburg and coworkers found that 3762/3782 (99.5%) of the coding sequences in strain 630Δerm are also found in strain 630.[120] We used the *C. difficile* 630 proteome database for our peptide search due to its high quality. However, there are still at least 20 known proteins that we potentially missed by conducting our analysis in this way, as they could potentially be present in *C. difficile* 630Δerm but not identified in the *C. difficile* 630 proteome database.

The experiment in workflow A revealed a small set of targets which were significantly enriched in samples treated with the specific probe **4.16** over the generic probe (Figure 4.26). Notably, the top five hits in this analysis all contained cysteine residues (Table 4.4). None of these hits from this particular species have been characterized in the literature, and there is only sparse information available about their homologs (if any) from other species. However, based on this preliminary bioinformatic analysis, it appears unlikely that any of these proteins act as proteases.

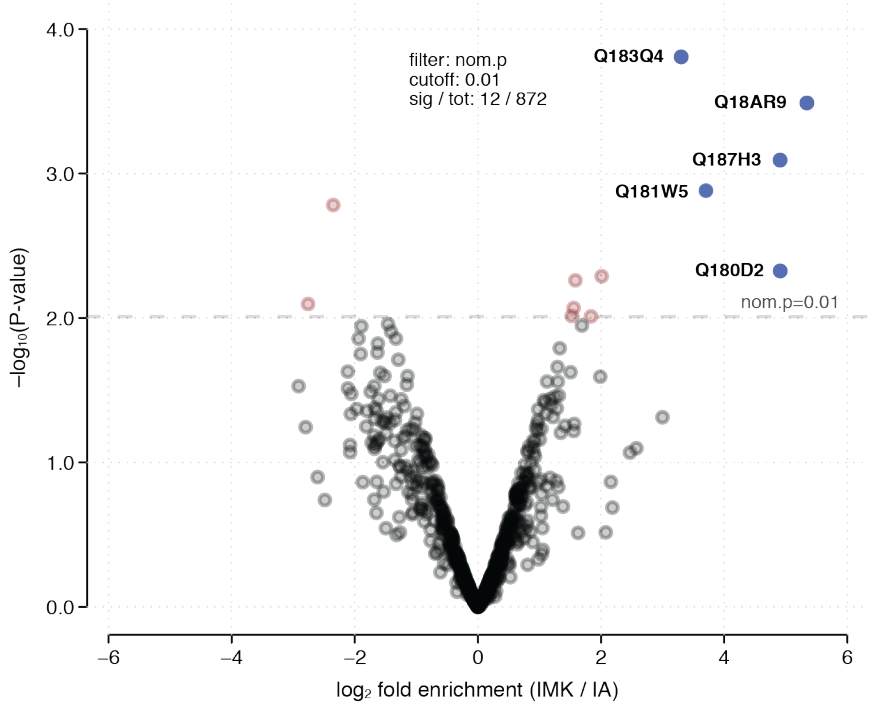


Figure 4.26: Protein-level enrichment of probe targets in *C. difficile* 630Δerm lysates (results from workflow A).

Plot of target enrichment by the specific probe versus calculated p-values (IMK = specific probe **4.16**, IA = generic probe **4.19**). Labeled blue dots represent enriched proteins that contain cysteine residues. Pink dots represent proteins that exceed the p=0.01 threshold but do not contain cysteine residues.

Table 4.4: Enriched targets from protein-level enrichment (workflow A).

UniProt ID	Annotation	log_2 fold enrichment (IMK / IA)	No. of cysteine residues
Q18AR9	Putative aminotransferase	5.3	3
Q180D2	Uncharacterized protein	4.9	1
Q187H3	Putative DNA/RNA helicase, Tn1549-like, CTn5-Orf21	4.9	8
Q181W5	Indolepyruvate oxidoreductase subunit IorA	3.7	20
Q183Q4	ABC-type transport system, multidrug-family ATP-binding/permease protein	3.3	2

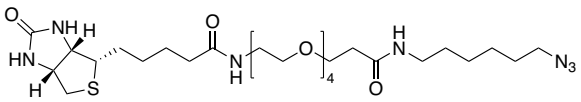
Concerned that perhaps the lack of proteases identified here indicated a problem with these probes labeling *C. difficile* proteases, I searched for annotated proteases in the full set of proteins identified from the LC-MS analysis of these samples. In this data set, I observed several common

microbial serine proteases (Clp, lon, HtrA) that did not significantly differ in abundance between the specific probe enriched and generic probe enriched samples. Though inhibition of serine proteases by chloromethyl ketones is the better studied phenomenon, the iodomethyl ketone electrophile is generally considered to be even more reactive, and it is possible that the probes are binding to the active site serine of these proteases.[133] It is also possible that the probes are reacting with other nucleophilic residues of these proteins.

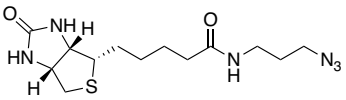
I continued our analysis of these results by comparing the hits from this experiment to all known peptidases and proteases in *C. difficile* 630Δerm. Of 122 proteins annotated as "peptidase" or "protease" in the genome, 19 could be easily identified as serine proteases, 44 as metalloproteases, 1 as an aspartic protease, and 5 as cysteine proteases. These cysteine proteases include the cell-surface associated protease Cwp84, which has been characterized,[134,135] and *C. difficile* toxin B, which has autoproteolytic activity.[136] Both of these proteins were identified in the proteomics data, but their abundances were not significantly different between the specific probe **4.16** and the generic probe **4.19**. Additionally, as with the serine proteases, we cannot confirm whether the labeling of these proteins resulted from the reaction of the probes with the active site nucleophiles of these proteases or with other residues.

For the experiment in workflow B, I performed the biotin labeling of samples and transferred them to Sam Myers and Deepak Mani, who performed the trypsin digestion, anti-biotin antibody enrichment, and LC-MS analysis. For these experiments, a larger amount of labeled proteomes (500 μg) were prepared. These samples were digested with trypsin, subjected to anti-biotin antibody enrichment, and analyzed by LC-MS/MS to attempt identification of probe target sites.[130] Unfortunately, our experiment with workflow B was unsuccessful. Sam Myers detected no enriched peptides from the specific probe and only a small number (13) from the generic

probe. There are several possible reasons for this failure, including an insufficient quantity of samples to detect low-abundance targets and poor retention and ionization efficiency of peptides on the LC-MS due to the presence of a PEG linker in biotin reagent **4.18**. Generally, a shorter linker between the biotin moiety and the peptide of interest is preferable for the enrichment of peptides to be analyzed by LC-MS (**4.20**, Figure 4.27).[137] In addition to products arising from the potential modification of cysteine by these probes, Sam Myers also searched these results for masses that could come from labeling of other nucleophilic residues (serine, threonine, and lysine). However, this search also did not reveal any peptides labeled by the specific probe.



4.18



4.20

Figure 4.27: Comparison of biotin azide reagents for click chemistry.
We used the reagent with a PEG_4 linker, **4.18**, in our experiments. A reagent with a shorter linker, such as **4.20**, may give better results.[137]

There are several obvious improvements to recommend if these experiments were to be repeated, as well as some lessons learned about using this technique for target ID in the gut microbiota. If we were to repeat the experiment in workflow B, we would use larger sample quantities and a probe containing a shorter biotin linker in order to maximize our chances of identifying specific target sites. It might also make sense to first establish that a model thiol containing peptide could react with this probe, that the resulting adduct would be chromatographically tractable, and that it would be visible and fragmentable by LC-MS.

In order to increase the likelihood of identifying peptidase targets for these probes in screens of additional species, it might make sense to focus on organisms that already have known or demonstrated secreted protease activity. However, as the goal of this method is to identify novel and/or low abundance interaction partners, pre-selecting species according to these criteria may neglect important potential targets. Based on the rarity of post-glutamyl hydrolyzing peptidases, it may not be surprising that we did not identify a protease target of ruminopeptin-type compounds *C. difficile* 630Δerm. Prior to this ABPP work, we could have first assessed if lysates from this organism could degrade a glutamyl endopeptidase substrate, which may have provided an early indication that we were unlikely to succeed in identifying specific targets of probe **4.16**.

We could also conduct similar ABPP experiments in other microbial strains, to see if we could identify any novel targets of these probes in a larger pool of proteins. However, this type of experiment may currently be too resource-intensive to be feasible. Though a generic activity-based probe was recently used by Wolan and coworkers to broadly examine differences in protease activity between *Rag1*$^{-/-}$ mice subjected to the T cell transfer model of colitis ("IBD mice") and healthy mice, the individual strains or species responsible for these differences were not identified.[123] These authors suggested that detection limits of tandem mass spectrometers are a current limitation of this work.[123] Development of higher throughput, lower cost, and more sensitive proteomics technologies should help to bridge the gap between the associations of protease activity and protease inhibition with disease and the actual microbial enzymes that are responsible for these effects.

4.3. Conclusions

In this chapter, we took a wide-ranging approach to access many potential products of NRPS-encoding gene clusters from human gut commensals and evaluate their bioactivities. Among this library of 48 compounds, we have discovered molecules that are potent inhibitors of human proteases relevant to the gut context, molecules that display antibiotic activity against gut commensals and pathogens, and molecules that inhibit secreted protease activity of gut bacteria. Additionally, we formulated a preliminary workflow to assess potential targets of a ruminopeptin-inspired halomethyl ketone probe in the gut bacterial pathogen *C. difficile*. Though this experiment did not reveal any targets of note for further study, this preliminary work may inform future efforts to identify the targets of small molecule probes in gut microbial species and more complex communities.

As the work described in this dissertation, and particularly in this chapter, relied heavily on bioinformatic predictions made with varying degrees of certainty, it is reasonable to ask if the active compounds identified here are produced in bacterial cultures or in the gut environment. This cannot be said with certainty without isolating the compounds produced by these NRPS-encoding gene clusters. However, our careful analysis of these biosynthetic assembly lines, relying on closely related characterized enzymes, gives us some confidence that the basic scaffolds of the *N*-acyl dipeptide aldehydes and tripeptide aldehydes predicted here are reasonable candidate products of these gene clusters. As there were some A domains for which we could not be predict specificity with certainty, we intentionally took a liberal approach and synthesized all the potential combinatorial products resulting from these predictions. Therefore, this 48 compound peptide aldehyde library should contain within it structure(s) comprised of the actual preferred building blocks of each assembly line. Heterologous expression and purification

of the biosynthetic enzymes in these gene clusters could enable assays to determine their actual A domain specificities, but the disadvantage of this approach is that it is potentially time-consuming.

4.3.1. Comparison of our findings with a previous study of these gene clusters

Portions of this section are adapted from our previously published work.[33]

As highlighted earlier, the gene clusters discussed in this work are part of a larger family of NRPS biosynthetic gene clusters found in human gut bacterial genomes and metagenomes.[1] In recent work, Fischbach and coworkers accessed the putative products of several of these gene NRPS gene clusters using a distinct workflow.[74] Relying principally on heterologous expression of these gene clusters in *E. coli* and *B. subtilis*, they were able to identify primarily cyclic pyrazinones and dihydropyrazinone compounds, along with one *N*-acylated dipeptide aldehyde (*N*-octanoyl-L-Met-L-Phe-H, a product of bgc33 from *Clostridium* sp. CAG:567) (Figure 4.28). This result indicates that the R domains of these NRPS assembly lines can produce aldehyde products. Their heterologous expression strategy was not universally effective, as products could be identified for only 7 of the 14 gene clusters investigated.[74] Bgc37 and bgc45 are among the gene clusters investigated for which no products could be isolated. As evidence that these products were not simply artifacts of heterologous expression, they also isolated one cyclic compound from its native producing organism and reconstituted of the biosynthesis of a cyclic pyrazinone *in vitro*.

Fischbach and coworkers hypothesized that the cyclic compounds observed in their study were derived from linear dipeptide aldehyde precursors and that these dipeptide aldehydes may be the relevant bioactive metabolites in vivo (Figure 4.28). To evaluate the inhibitory activity of these predicted compounds against human proteases, they synthesized several dipeptide aldehydes (containing both Boc protecting groups and unprotected N-termini, as well as one *N*-acylated dipeptide aldehyde) and evaluated them as inhibitors of human proteases. They found that free-amino dipeptide aldehydes had potent activity against calpain 1 and the cysteine cathepsins B, L, C and S. They then used chemoproteomics to measure the global interactions of a representative dipeptide aldehyde (L-phenylalanyl-L-phenylalanyl aldehyde) with the human proteome and concluded that the cathepsins are likely principal targets of this compound.[74]

There is some overlap between the gene clusters that were a focus of that work and this work, so it is useful to compare the predicted linear aldehyde precursors of the pyrazinones identified by Fischbach and coworkers[74] with our synthetic library (Figure 4.28). In their heterologous expression of bgc34 in *E. coli*, they observed pyrazinones with aromatic amino acids in positions P1 and P2. Though this is in line with our predictions for the P2 specificity of this NRPS, it contradicts our P1 prediction. In their heterologous expression of bgc38, they observed one major pyrazinone that incorporates α-aminobutyric acid in P1 and tryptophan in P2. This is reasonably similar to our A domain loading predictions for this cluster, though there are no scaffolds that exactly match this one in our synthetic library. Finally, from heterologous expression of bgc52, they observed many products, all of which incorporate an aromatic residue (phenylalanine, tyrosine, tryptophan) in position P1 and a large aliphatic residue (valine, leucine, isoleucine) in position P2. These are in good agreement with the P1 and P2 loading specificities that we predicted for this gene cluster.

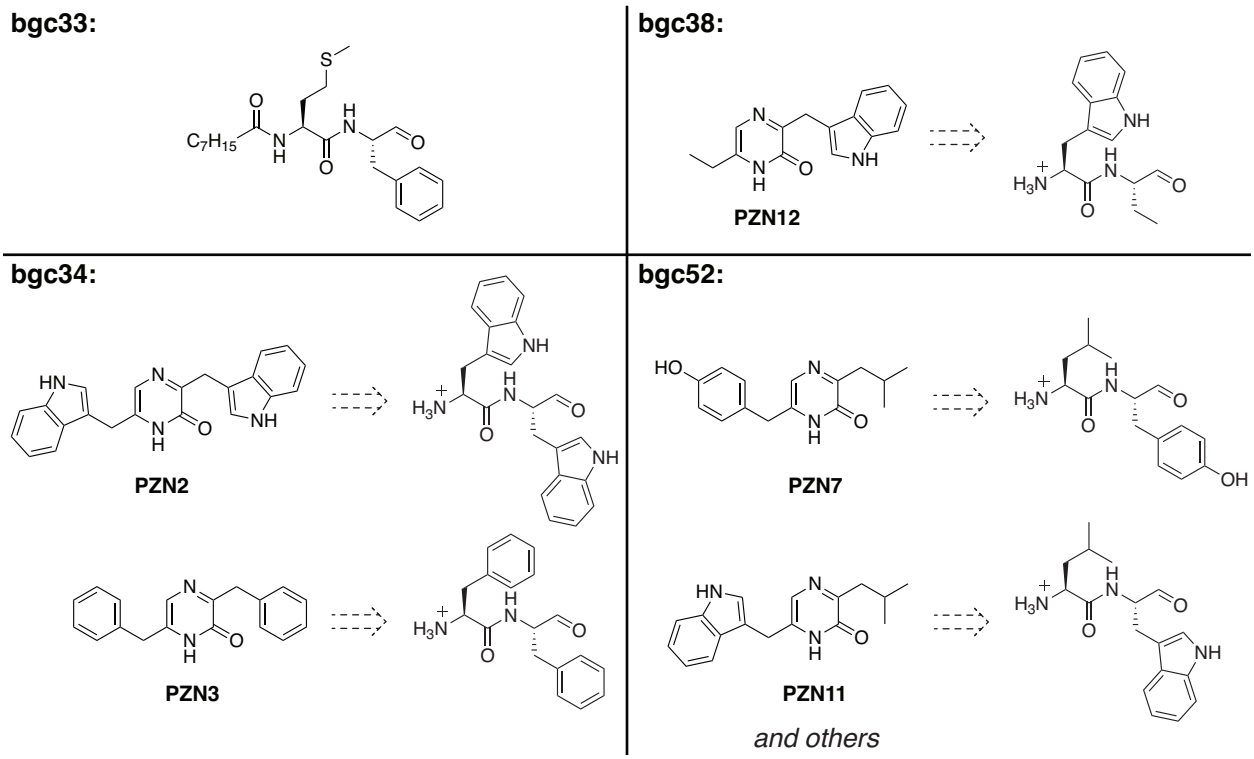


Figure 4.28: Compounds isolated by Fischbach and coworkers in their study of gut microbial NRPS gene clusters.

Fischbach and coworkers isolated one peptide aldehyde compound from heterologous expression of bgc33 in *E. coli* and many cyclic pyrazinone compounds from heterologous expression of other clusters in *E. coli* and *B. subtilis*. These compounds are hypothesized to be the cyclization and oxidation products of the linear precursors shown here, which were not directly observed.[74]

Though this work does corroborate some of our predictions for A domain specificities of the NRPS enzymes encoded by these gene clusters, the structures isolated in this study are not completely consistent with the biosynthetic machinery present in these pathways. As discussed above, based on domain architectures, we believe that bgc34 and bgc38 should produce *N*-acyl dipeptide aldehydes, while bgc52 should produce tripeptide aldehyde(s) with unprotected N-termini. The cyclic pyrazinones isolated in their study did not demonstrate protease inhibitory activity, and so it remains to be determined what compounds are actually produced in vivo and what their physiological roles are.[74]

The intermolecular cyclization of dipeptide aldehydes with unprotected N-termini to form cyclic dihydropyrazinones and the oxidation of these compounds to the corresponding aromatic pyrazinones are likely spontaneous processes, as has been suggested by in vitro reconstitution of the NRPS AusA to produce aureusimine B (phevalin).[138] Our *N*-acyl dipeptide aldehydes are prevented from undergoing cyclization by the "caps" on their N-termini. Though there has been one report of the cyclization of a tripeptide aldehyde with a free amine,[139] this should be a reversible and much less favorable process. Therefore, considering the potential bioactivity of these compounds, it is likely that the scaffolds we have predicted here could be available as free electrophiles in the gut environment.

Additionally, to our knowledge, we are the first to evaluate the bioactivity of the predicted products of this family of gut microbial NRPS gene clusters towards microbes. In Chapter 3, we showed that a subset of these compounds can inhibit the glutamyl endopeptidase SspA from *S. aureus*, and in this chapter, we showed that these compounds demonstrate antimicrobial activity against gut microbes and may inhibit secreted proteases from these organisms. Based on what is currently understood about the biogeography of the gut microbiota, these peptide aldehydes may be more likely have microbial targets than human targets. As discussed above, the major species encoding the gene clusters discussed in this chapter are *E. tayi*, *C. leptum*, *B. producta*, *R. bromii*, and *B. wexlerae*, which are all members of Clostridium clusters IV and XIVa. A stratified mucus layer prevents most microorganisms in the colon from direct contact with human cells, and though some bacteria are regularly observed in association with the colon mucus (such as *Akkermansia muciniphila*), these species are not hypothesized to share this niche.[140,141] Rather, it is likely that they are found mostly in the lumen. Though it is possible that these species have evolved to produce small molecules that diffuse to the host epithelial cell

layer to exert their effects, it is also reasonable to propose that these molecules could act locally against other luminal bacteria. The large number of microbial proteins in this environment poses a challenge in identifying the specific targets of small molecules, but in this chapter, we have demonstrated three distinct workflows to address this challenge.

4.3.2. Future directions

Several interesting future directions for this work have already been identified and discussed in the body of this chapter. We and others have now identified that human proteases involved in antigen presentation in the gut environment may be in vivo targets of putative gut microbial peptide aldehydes, providing a potential link between these compounds, inflammation, and immune response.[58,74] It is possible to envision complex cell-based assays for evaluating small molecules as modulators of antigen presentation.[142,143] However, a simpler experiment, which relies on the assumption that these compounds are actually produced in vivo, would be to compare the effects of bacterial deletion mutants missing these gene clusters with wild type strains in animal models of IBD. As several of these gene clusters are from organisms that are part of well-defined model microbial communities, the major current limitation here is the paucity of tools for genetically manipulating gut microbes and particularly anaerobic species. There has been some encouraging progress recently in this area,[144] and we believe that genetic manipulation of these gene clusters would be a transformative advance in our ability to reveal their biological role(s).

We have also found that some of the peptide aldehydes exhibit antibiotic activity against gut microbes and inhibit secreted protease activity of gut microbes. Identification of the specific targets leading to these phenotypes may reveal strategies for modulating the activity of certain

gut microbial strains or selectively killing them.[145] To address if these effects actually happen in the gut environment, we could again study the effect of deletion mutants on gut microbial composition in vivo. However, if our work or related efforts lead to the discovery of potent small molecules with useful bioactivities, it may not be important if the compounds that originally inspired this investigation are actually produced in the environment or not. We used distinct selections of species for these two sets of experiments, and it would also be useful to comprehensively evaluate the species screened for one activity in the other assay as well. For example, if a strain exhibiting antibiotic susceptibility to a peptide aldehyde compound also had demonstrable inhibition of secreted protease activity by a (similar) peptide aldehyde, we could prioritize that interaction for mechanistic study and attempt to design more potent compounds.

Our ABPP experiments did not identify specific targets in the single strain that we screened. It is possible that we may be able to identify novel targets of these compounds by applying this approach to additional species. However, it may be advisable to prioritize species for which we could first observe significant labeling of targets by fluorescence labeling and SDS-PAGE. This is an exciting future direction, and in general we expect that proteomics-based technologies for identifying the targets of small molecules in microbes will continue to develop in the coming years.[117]

Over the past several years, there have been many studies that identified potential roles for secondary metabolites generated by the human gut microbiota. These compounds may have roles as antibiotics (ruminococcins,[98,146] humimycin[3]), in signaling with GPCRs (commendamide and analogues[147,148]), in the development of colorectal cancer (colibactin[149,150]), and in the immune system (putative dipeptide aldehydes[74]). The study of secondary metabolite production by gut microbes and the elucidation of these compounds' biological roles remain active areas of

interest, as discussed in several recent reviews.[151–153] However, biosynthetic chemistry is only a small part of the fascinating and unique chemistry performed by these organisms. We believe that the study of synthetic compounds that inhibit gut microbial enzymes is equally exciting to explore. Considering the problem of "drugging" the gut microbiota,[154] similar techniques can be used in both of these research areas to validate targets, with the eventual goal of using small molecules to modulate gut microbial growth, competition, and virulence.

4.4. Materials and methods

4.4.1. General materials and methods

All chemicals were obtained from Sigma-Aldrich except where noted. Protected amino acids were obtained from Chem-Impex (Dale, IL) and Advanced ChemTech (Louisville, KY). HATU was purchased from Oakwood Chemical (Estill, SC). All NMR solvents were purchased from Cambridge Isotope Laboratories (Andover, MA). NMR spectra were visualized using iNMR version 5.5.7. and MestReNova version 12.0. 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, br s = broad singlet, d = doublet, t = triplet, q = quartet, m = multiplet), coupling constant, and integration. High-resolution mass spectral (HRMS) data was obtained in the Small Molecule Mass Spectrometry Facility, FAS Division of Science. HRMS data for synthetic compounds was obtained on an Agilent Technologies 6210 TOF coupled to an Agilent Technologies 1200 series LC. Liquid chromatography was performed with water/acetonitrile (1:1). The capillary voltage was 3.5 kV,

the fragmentor voltage was 175 V, the drying gas temperature was 325 °C, the drying gas flow rate was 8 L/min, and the nebulizer pressure was 40 psig.

4.4.2. Synthesis of amino acid Weinreb amides

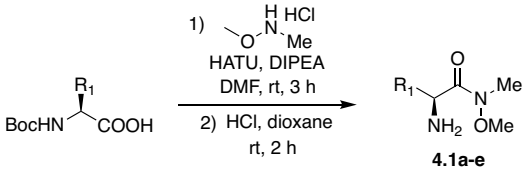
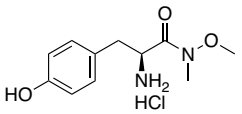


Figure 4.29: Synthesis of amino acid Weinreb amides 4.1a–e.
HATU = 1-[Bis(dimethylamino)methylene]-1H-1,2,3-striazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate, DIPEA= *N*,*N*-Diisopropylethylamine.

To a solution of Boc-protected amino acid (20 mmol) and *N*,*O*-dimethylhydroxylamine hydrochloride in DMF (0.6 M) was added HATU (1.1 equiv) and DIPEA (3.1 equiv) with stirring, under argon. After 3 h, the reaction mixture was diluted with ethyl acetate (200 mL) and quenched by adding to 1 M aqueous NaOH (200 mL). The organic layer was collected and the aqueous layer extracted with two portions of ethyl acetate (each 200 mL). The combined organic layers were washed twice with water and once with brine (each 200 mL), dried over Na_2SO_4, filtered, and concentrated in vacuo to afford crude Boc-protected amino Weinreb amides. These crude products (1.0 equiv) were dissolved in 4 M HCl in dioxane (0.66 M) and stirred at rt for 2 hours. The reaction mixture was concentrated in vacuo, and diethyl ether (100 mL) was added to the residue. The precipitated product **4.1a–e** was collected by filtration after standing at –20 ºC overnight. The characterization data for compounds **4.1a–d** matched previously reported results.[155–158]



4.1e

4.4.2.1. (*S*)-2-amino-3-(4-hydroxyphenyl)-*N*-methoxy-*N*-methylpropanamide (**4.1e**):

The product (5.17 g, 99%) was isolated as a colorless solid (m.p. 99 – 101 ºC). ^{1}H NMR (400 MHz; DMSO-d_6): δ 8.31 (s, 3H), 6.78 (d, J = 8.2 Hz, 2H), 6.55 (d, J = 8.4 Hz, 2H), 4.11 (s, 1H), 3.37 (s, 3H), 2.98 (s, 2H), 2.90 (s, 3H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 168.3, 156.7, 130.5, 124.5, 115.4, 61.5, 50.9, 40.1, 39.9, 39.7, 39.5, 39.3, 39.1, 38.9, 35.3, 34.0, 31.8. HRMS (ESI): Calc'd for formula $C_{11}H_{17}N_2O_3^+$ $[M+H]^+$ 225.1234, found 225.1245.

4.4.3. Synthesis of *N*-acyl amino acids

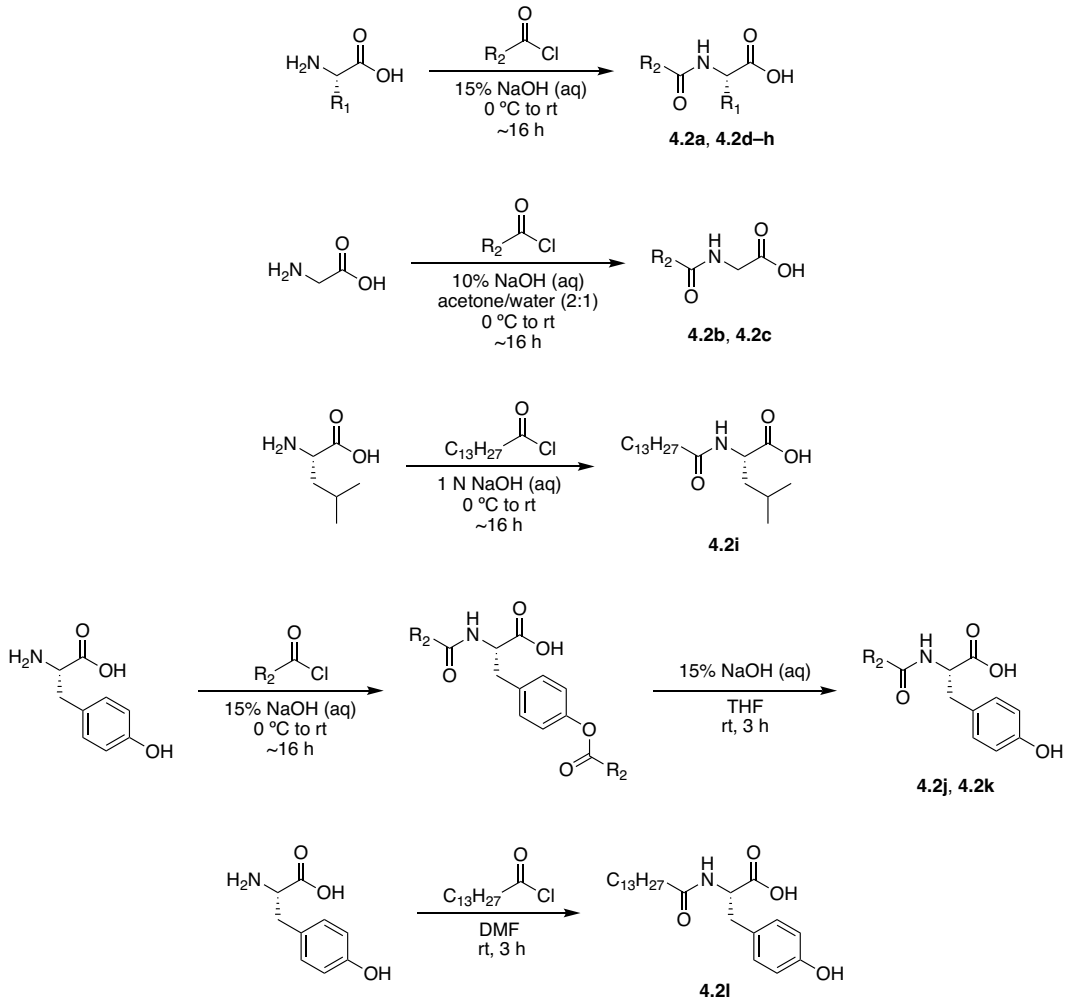
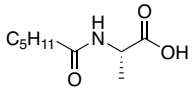


Figure 4.30: Synthesis of *N*-acyl amino acids 4.2a–l.

N-acyl amino acids were generally synthesized using the Schotten–Baumann reaction of acyl chlorides with amino acids in aqueous base. Unless otherwise noted, the procedure was as follows: the amino acid (1.0 equiv) was dissolved in 15% aqueous NaOH (0.5 M) and cooled to 0 °C. The acid chloride (1.1 equiv) was added dropwise and the reaction mixture stirred overnight, allowing to warm to rt. 20% aqueous HCl was added to pH = 2 and the resulting solution was extracted with dichloromethane (three portions of 3x reaction volume). The

combined organic layers were washed with saturated aqueous NaCl (one portion of 1x reaction volume). The solution was then dried over Na_2SO_4, filtered, and concentrated in vacuo.

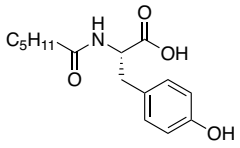
Compounds **4.2b** and **4.2c** were synthesized according to a previously reported procedure.[159] Compound **4.2i** was synthesized according to a previously reported procedure.[160] For compounds **4.2j** and **4.2k**, the Schotten-Bauman reaction conditions initially yielded the *N*,*O*-diacylated products. The *O*-acyl groups were removed as previously described.[161] Compound **4.2l** was synthesized according to a previously reported procedure.[162] The characterization data for compounds **4.2a**, **4.2b**, **4.2c**, **4.2e**, **4.2f**, **4.2g**, **4.2h**, **4.2i**, **4.2k**, and **4.2l** matched previously reported results.[159–166]



4.2d

4.4.3.1. Hexanoyl-L-alanine (**4.2d**):

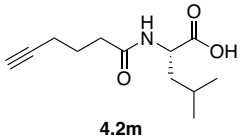
The product (1.24 g, 33%) was isolated as a colorless solid (m.p. 83 – 87 ºC). ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.83 (br s, 1H), 6.21 (d, J = 7.0 Hz, 1H), 4.58 (m, 1H), 2.24 (m, 2H), 1.63 (p, 2H), 1.46 (d, J = 7.1 Hz, 3H), 1.31 (m, 4H), 0.90 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 176.1, 174.3, 48.5, 36.6, 31.5, 25.5, 22.6, 18.3, 14.1. HRMS (ESI): Calc'd for formula $C_9H_{16}NO_3^-$ [M–H]$^-$ 186.1136, found 186.1143.



4.2j

4.4.3.2. Hexanoyl-L-tyrosine (**4.2j**):

The product (1.34 g, 43%) was isolated as an orange oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 6.97 (d, *J* = 8.1 Hz, 2H), 6.73 (d, *J* = 8.1 Hz, 2H), 6.08 (d, *J* = 7.4 Hz, 1H), 4.81 (d, *J* = 6.9 Hz, 1H), 3.46 (s, 2H), 3.13 – 3.00 (m, 2H), 2.17 (t, *J* = 7.6 Hz, 2H), 1.57 (t, *J* = 7.2 Hz, 2H), 1.25 (m, 4H), 0.86 (t, *J* = 6.9 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 177.3, 175.6, 175.0, 172.4, 155.5, 130.5, 127.2, 115.8, 53.6, 52.0, 36.8, 36.4, 31.3, 25.4, 20.9, 14.0. HRMS (ESI): Calc'd for formula $C_{15}H_{20}NO_4^-$ $[M–H]^-$ 278.1398, found 278.1417.



4.2m

4.4.3.3. Hex-5-ynoyl-L-leucine (**4.2m**)

The product (1.34 g, 79%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 11.16 – 11.02 (m, 1H), 6.05 (m, 1H), 4.61 (m, 1H), 2.50 (t, *J* = 7.4 Hz, 1H), 2.38 (t, *J* = 7.4 Hz, 2H), 2.26 (m, 4H), 1.97 (m, 2H), 1.84 (m, 4H), 1.74 – 1.66 (m, 2H), 1.57 (m, 1H), 0.93 (t, *J* = 8.2 Hz, 6H). ^{13}C NMR (101 MHz; $CDCl_3$) δ 179.1, 83.6, 69.6, 51.1, 41.4, 35.0, 32.8, 25.1, 24.2, 23.5, 23.0, 22.1, 18.0. Calc'd for formula $C_{12}H_{18}NO_3^-$ $[M–H]^-$ 224.1292, found 224.1308.

4.4.4. Synthesis of Boc-protected dipeptides

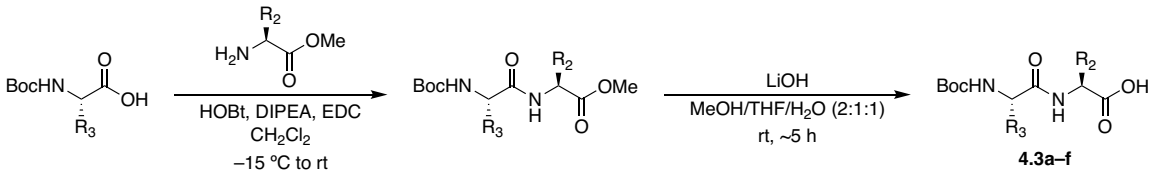
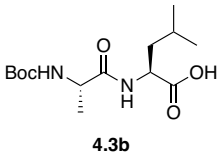


Figure 4.31: Synthesis of Boc-protected dipeptides 4.3a–f.

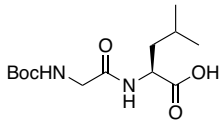
(HOBt = hydroxybenzotriazole, EDC = 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide)

Boc-protected dipeptides were prepared according to a previously published procedure.[37] The characterization data for compounds **4.3a** and **4.3c** matched previously reported results.[167,168]



4.4.4.1. (*tert*-butoxycarbonyl)-L-alanyl-L-leucine (**4.3b**):

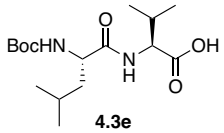
The product (1.46 g, 84% over 2 steps) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 10.39 (br s, 1H), 6.91 (br s, 1H), 5.37 (br s, 1H), 4.59 (m, 1H), 4.21 (br s, 1H), 1.65 (m, 3H), 1.31 (m, 12H), 0.88 (s, 6H). ^{13}C NMR (101 MHz; $CDCl_3$) δ 176.2, 173.4, 51.0, 41.4, 28.5, 25.0, 23.1, 21.9, 18.1. HRMS (ESI): Calc'd for formula $C_{14}H_{25}N_2O_5^-$ $[M–H]^-$ 301.1769, found 301.1787.



4.3d

4.4.4.2. (*tert*-butoxycarbonyl)-L-glycyl-L-leucine (**4.3d**):

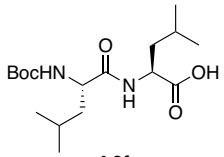
The product (1.66 g, 96% over 2 steps) was isolated as a colorless solid. ^{1}H NMR (400 MHz; Chloroform-*d*): δ 6.94 (s, 1H), 5.56 (s, 1H), 4.60 (m, 1H), 3.86 (m, 2H), 1.74 – 1.51 (m, 3H), 1.45 (m, 9H), 0.96 – 0.87 (m, 6H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 169.7, 129.0, 109.7, 52.6, 41.7, 28.5, 25.1, 25.0, 23.2, 23.0, 22.1, 22.0. HRMS (ESI): Calc'd for formula $C_{13}H_{23}N_2O_5^-$ [M–H]$^-$ 287.1612, found 287.1625.



4.3e

4.4.4.3. (*tert*-butoxycarbonyl)-L-leucyl-L-valine (**4.3e**):

The product (1.22 g, 64% over 2 steps) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 6.90 (m, 1H), 5.11 (m, 1H), 4.56 (dd, *J* = 8.8, 4.8 Hz, 1H), 4.16 (m, 1H), 2.25 (td, *J* = 6.8, 5.0 Hz, 1H), 1.65 (m, 2H), 1.48-1.46 (s, 9H), 0.94 (m, 12H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 174.9, 173.3, 80.7, 57.2, 53.3, 31.4, 28.5, 24.9, 22.9, 22.4, 19.2, 17.7. HRMS (ESI): Calc'd for formula $C_{16}H_{29}N_2O_5^-$ [M–H]$^-$ 329.2082, found 329.21.



4.3f

4.4.4.4. (*tert*-butoxycarbonyl)-L-leucyl-L-leucine (**4.3f**):

The product (1.57 g, 80% over 2 steps) was isolated as a colorless solid. ^{1}H NMR (500 MHz; $CDCl_3$): δ 6.93 (d, *J* = 7.6 Hz, 1H), 5.25 (d, *J* = 8.1 Hz, 1H), 4.63 (m, 1H), 4.21 – 4.14 (m, 1H),

1.70 – 1.55 (m, 5H), 1.44 (s, 9H), 0.89 (s, 12H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 176.0, 173.1, 80.7, 53.2, 50.8, 41.6, 40.9, 28.5, 24.8, 23.1, 22.5, 21.9. HRMS (ESI): Calc'd for formula $C_{17}H_{31}N_2O_5^-$ $[M–H]^-$ 343.2238, found 343.2259.

4.4.5. Coupling of *N*-acyl amino acids to Weinreb amides

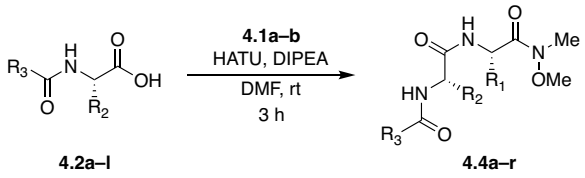
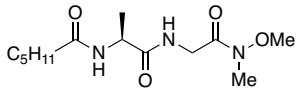


Figure 4.32: Coupling of *N*-acyl amino acids to Weinreb amides.

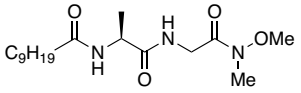
To a solution of Weinreb amide **4.1a–b** (1.0 equiv) and *N*-acyl amino acid **4.2a–l** (1.1 equiv) in DMF (0.6 M) was added HATU (1.1 equiv) and DIPEA (4.1 equiv) with stirring, under argon. After 3 h, the reaction mixture was diluted with ethyl acetate (to 4x initial volume) and quenched by addition of 1 M aqueous NaOH (4x initial volume). The organic layer was collected and the aqueous layer extracted with three portions of ethyl acetate (each 4x initial reaction volume). The combined organic layers were washed with water and brine (each 8x initial reaction volume), dried over Na_2SO_4, filtered, and concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator.



4.4a

4.4.5.1. (*S*)-*N*-(1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)hexanamide (**4.4a**):

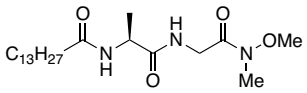
The product (44 mg, 31%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.87 (s, 1H), 6.26 (d, *J* = 7.5 Hz, 1H), 4.56 (m, 1H), 4.17 (d, *J* = 4.7 Hz, 2H), 3.75 (s, 1H), 3.72 (s, 3H), 3.20 (s, 3H), 2.20 (m, 3H), 1.62 (t, *J* = 7.5 Hz, 3H), 1.41 (m, 6H), 1.35 – 1.10 (m, 8H), 0.88 (m, 5H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.2, 172.9, 61.7, 48.9, 41.0, 31.6, 25.5, 22.6, 19.0, 14.1. HRMS (ESI): Calc'd for formula $C_{13}H_{26}N_3O_4^+$ $[M+H]^+$ 288.1918, found 288.192.



4.4b

4.4.5.2. (*S*)-*N*-(1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)decanamide (**4.4b**):

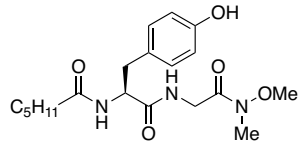
The product (101 mg, 59%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.99 (m, 1H), 6.38 (d, *J* = 7.4 Hz, 1H), 4.56 (p, *J* = 7.1 Hz, 1H), 4.16 (m, 2H), 3.70 (s, 3H), 3.19 (s, 3H), 2.18 (m, 2H), 1.59 (m, 2H), 1.38 (d, *J* = 6.7 Hz, 3H), 1.34 – 1.20 (m, 15H), 0.85 (t, *J* = 6.6 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.6, 173.3, 173.0, 61.7, 48.8, 40.9, 36.7, 32.0, 29.6, 29.5, 29.4, 25.8, 22.8, 18.9, 14.3. HRMS (ESI): Calc'd for formula $C_{17}H_{34}N_3O_4^+$ $[M+H]^+$ 344.2544, found 344.256.



4.4c

4.4.5.3. (*S*)-*N*-(1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)tetradecanamide (**4.4c**):

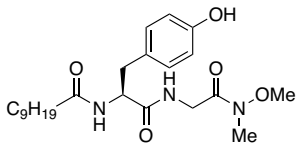
The product (170 mg, 85%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.07 (t, *J* = 4.8 Hz, 1H), 6.49 (d, *J* = 7.5 Hz, 1H), 4.55 (p, *J* = 7.1 Hz, 1H), 4.15 (m, 2H), 3.69 (s, 3H), 3.18 (s, 3H), 2.19 (m, 2H), 1.57 (q, *J* = 7.5 Hz, 3H), 1.36 (d, *J* = 6.8 Hz, 5H), 1.30 – 1.19 (m, 31H), 0.84 (t, *J* = 6.8 Hz, 4H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.6, 173.3, 173.0, 61.7, 54.9, 48.8, 43.0, 40.8, 38.7, 36.7, 29.8, 29.8, 29.6, 29.6, 29.5, 29.5, 29.5, 29.4, 25.8, 18.8, 14.2. HRMS (ESI): Calc'd for formula $C_{21}H_{41}N_3O_4Na^+$ $[M+Na]^+$ 422.2989, found 422.3006.



4.4d

4.4.5.4. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)hexanamide (**4.4d**):

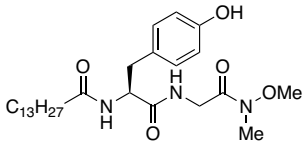
The product (68 mg, 36%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.99 (m, 2H), 6.72 (m, 2H), 6.24 (m, 1H), 4.72 (m, 1H), 4.22 (m, 1H), 4.04 (m, 1H), 3.70 (s, 2H), 3.18 (s, 2H), 2.97 (m, 1H), 2.80 (s, 12H), 2.16 (m, 2H), 1.64 – 1.38 (m, 4H), 1.34 – 1.14 (m, 6H), 0.86 (q, J = 7.0 Hz, 3H). 13C NMR (101 MHz; $CDCl_3$): δ 171.8, 155.8, 130.5, 127.7, 121.9, 115.8, 61.7, 55.9, 54.6, 40.9, 38.8, 36.7, 31.5, 25.4, 22.5, 18.8, 17.5, 14.1. HRMS (ESI): Calc'd for formula $C_{19}H_{30}N_3O_5^+$ $[M+H]^+$ 380.218, found 380.2161.



4.4e

4.4.5.5. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)decanamide (**4.4e**):

The product (187 mg, 86%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.96 (d, *J* = 8.3 Hz, 3H), 6.69 (d, *J* = 8.3 Hz, 2H), 6.48 (d, *J* = 7.9 Hz, 1H), 4.71 (q, *J* = 7.2 Hz, 1H), 4.19 (m, 1H), 4.01 (m, 1H), 3.68 (s, 3H), 3.15 (s, 3H), 2.93 (m, 3H), 2.79 (s, 3H), 2.13 (dd, *J* = 8.6, 6.6 Hz, 2H), 1.51 (m, 2H), 1.22 (m, 16H), 0.85 (m, 4H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 174.0, 172.1, 155.9, 130.4, 127.5, 115.8, 61.6, 54.6, 40.9, 40.8, 38.8, 37.8, 36.6, 32.0, 29.6, 29.5, 29.4, 29.4, 25.8, 25.7, 22.8, 14.3. HRMS (ESI): Calc'd for formula $C_{23}H_{37}N_3O_5Na^+$ $[M+Na]^+$ 458.2625, found 458.2639.

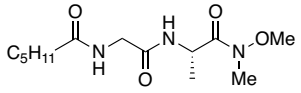


4.4f

4.4.5.6. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-((2-(methoxy(methyl)amino)-2-oxoethyl)amino)-1-oxopropan-2-yl)tetradecanamide (**4.4f**):

The product (164 mg, 67%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.95 (m, 3H), 6.69 (d, *J* = 8.1 Hz, 2H), 6.44 (d, *J* = 8.1 Hz, 1H), 4.72 (m, 1H), 4.20 (m, 2H), 4.02 (m, 1H), 3.69 (m, 5H), 3.18 (s, 4H), 3.03 – 2.90 (m, 2H), 2.80 (s, 2H), 2.14 (dd, *J* = 8.6, 6.6 Hz, 2H), 1.61 (m, 1H), 1.52 (t, *J* = 7.3 Hz, 2H), 1.22 (m, 32H), 0.86 (t, *J* = 6.7 Hz, 5H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.9, 173.9, 172.0, 155.9, 130.4, 127.6, 115.8, 61.7, 54.6, 40.9, 40.8,

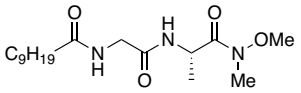
38.8, 36.7, 32.1, 29.9, 29.8, 29.8, 29.8, 29.7, 29.6, 29.5, 29.5, 29.4, 25.9, 25.7, 22.9, 14.3. HRMS (ESI): Calc'd for formula $C_{27}H_{45}N_3O_5Na^+$ $[M+Na]^+$ 514.3251, found 514.326.



4.4g

4.4.5.7. (*S*)-*N*-(2-((1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-2-oxoethyl)hexanamide (**4.4g**):

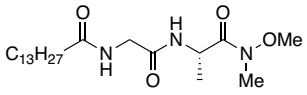
The product (70 mg, 49%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.04 (d, *J* = 7.6 Hz, 1H), 6.46 (m, 1H), 4.92 (m, 1H), 3.94 (m, 2H), 3.77 (s, 3H), 3.20 (s, 3H), 2.22 (m, 2H), 1.63 (m, 2H), 1.30 (m, 8H), 0.87 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.9, 168.7, 61.8, 45.9, 43.0, 36.6, 31.6, 25.5, 22.6, 18.3, 14.1. HRMS (ESI): Calc'd for formula $C_{13}H_{25}N_3O_4Na^+$ $[M+Na]^+$ 310.1737, found 310.1753.



4.4h

4.4.5.8. (*S*)-*N*-(2-((1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-2-oxoethyl)decanamide (**4.4h**):

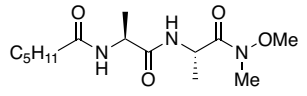
The product (167 mg, 97%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.24 (d, *J* = 7.5 Hz, 1H), 6.62 (m, 1H), 4.90 (m, 1H), 3.93 (m, 2H), 3.75 (s, 3H), 3.18 (s, 3H), 2.76 (s, 1H), 2.18 (m, 2H), 1.58 (p, *J* = 7.2 Hz, 2H), 1.30 (d, *J* = 6.9 Hz, 3H), 1.22 (m, 14H), 0.83 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 168.9, 61.7, 45.8, 42.9, 36.5, 32.0, 29.6, 29.5, 29.4, 29.4, 25.8, 22.8, 18.5, 18.1, 14.2. HRMS (ESI): Calc'd for formula $C_{17}H_{33}N_3O_4Na^+$ $[M+Na]^+$ 366.2363, found 366.2374.



4.4i

4.4.5.9. (*S*)-*N*-(2-((1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-2-oxoethyl)tetradecanamide (**4.4i**):

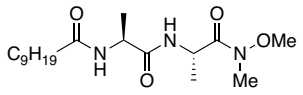
The product (187 mg, 94%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.24 (d, *J* = 7.5 Hz, 1H), 6.60 (m, 1H), 4.90 (p, *J* = 7.2 Hz, 1H), 3.93 (d, *J* = 5.2 Hz, 2H), 3.75 (s, 3H), 3.18 (s, 3H), 2.68 (s, 1H), 2.19 (m, 2H), 1.59 (q, *J* = 7.3 Hz, 2H), 1.34 (d, *J* = 6.8 Hz, 4H), 1.25 (d, *J* = 6.7 Hz, 2H), 1.21 (m, 15H), 0.84 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.9, 168.9, 61.7, 45.8, 42.9, 36.5, 32.1, 29.8, 29.8, 29.6, 29.5, 29.5, 29.5, 25.8, 22.8, 18.1, 14.3. HRMS (ESI): Calc'd for formula $C_{21}H_{42}N_3O_4Na^+$ $[M+Na]^+$ 422.2989, found 422.3006.



4.4j

4.4.5.10. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)hexanamide (**4.4j**):

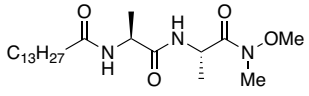
The product (92 mg, 61%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.03 (d, *J* = 7.7 Hz, 1H), 6.35 (d, *J* = 7.5 Hz, 1H), 4.88 (p, *J* = 7.1 Hz, 1H), 4.53 (p, *J* = 7.1 Hz, 1H), 3.76 (s, 3H), 3.20 (s, 3H), 2.17 (m, 2H), 1.61 (m, 2H), 1.3 (m, 10H), 1.23 (s, 1H), 0.86 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.1, 172.2, 61.8, 48.8, 45.9, 38.8, 36.7, 31.6, 25.5, 22.6, 19.0, 18.3, 14.1. HRMS (ESI): Calc'd for formula $C_{14}H_{27}N_3O_4Na^+$ $[M+Na]^+$ 324.1894, found 324.1906.



4.4k

4.4.5.11. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)decanamide (**4.4k**):

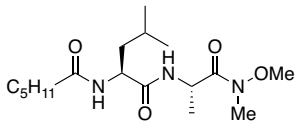
The product (131 mg, 73%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.14 (d, *J* = 7.8 Hz, 1H), 6.43 (d, *J* = 7.6 Hz, 1H), 4.89 (m, 1H), 4.54 (p, *J* = 7.1 Hz, 1H), 3.74 (s, 3H), 3.18 (s, 3H), 2.15 (m, 2H), 1.59 (q, *J* = 7.3 Hz, 2H), 1.27 (m, 19H), 0.83 (t, *J* = 6.8 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.1, 172.3, 61.7, 48.7, 45.8, 45.4, 36.7, 32.0, 29.6, 29.5, 29.4, 25.8, 22.8, 19.1, 18.5, 18.2, 14.2. HRMS (ESI): Calc'd for formula $C_{18}H_{35}N_3O_4Na^+$ $[M+Na]^+$ 380.252, found 380.2521.



4.4l

4.4.5.12. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)tetradecanamide (**4.4l**):

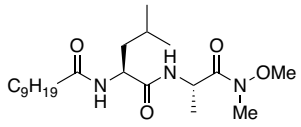
The product (187 mg, 90%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.11 (d, *J* = 7.8 Hz, 1H), 6.43 (d, *J* = 7.3 Hz, 1H), 4.91 (m, 1H), 4.54 (p, *J* = 7.1 Hz, 1H), 3.75 (s, 3H), 3.19 (s, 3H), 2.16 (t, *J* = 7.6 Hz, 2H), 1.59 (m, 2H), 1.31 (m, 5H), 1.21 (s, 21H), 0.84 (t, *J* = 6.6 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.1, 172.2, 61.7, 48.7, 45.8, 36.8, 32.1, 29.8, 29.8, 29.8, 29.6, 29.5, 29.4, 25.8, 22.8, 19.1, 18.6, 18.2, 14.3. HRMS (ESI): Calc'd for formula $C_{22}H_{43}N_3O_4Na_+$ $[M+Na]_+$ 436.3146, found 436.3149.



4.4m

4.4.5.13. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-4-methyl-1-oxopentan-2-yl)hexanamide (**4.4m**):

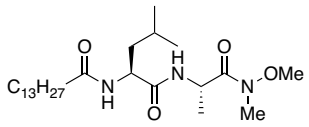
The product (161 mg, 94%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.06 (d, *J* = 7.8 Hz, 1H), 6.37 (d, *J* = 8.5 Hz, 1H), 4.86 (p, *J* = 7.1 Hz, 1H), 4.52 (m, 1H), 3.73 (s, 3H), 3.18 (s, 3H), 2.16 (m, 2H), 1.55 (m, 5H), 1.27 (m, 7H), 0.98 – 0.80 (m, 10H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.4, 172.2, 61.7, 51.6, 45.8, 41.9, 36.7, 31.6, 25.5, 24.9, 23.2, 22.5, 22.1, 18.1, 14.1. HRMS (ESI): Calc'd for formula $C_{17}H_{33}N_3O_4K_+$ $[M+K]_+$ 382.2103, found 382.2113.



4.4n

4.4.5.14. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-4-methyl-1-oxopentan-2-yl)decanamide (**4.4n**):

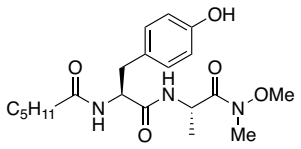
The product (197 mg, 99%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.08 (d, *J* = 7.7 Hz, 1H), 6.38 (d, *J* = 8.5 Hz, 1H), 4.85 (q, *J* = 7.1 Hz, 1H), 4.51 (m, 1H), 3.80 (s, 1H), 3.72 (s, 3H), 3.17 (s, 3H), 2.14 (t, *J* = 7.6 Hz, 2H), 1.69 – 1.43 (m, 3H), 1.31 – 1.17 (m, 21H), 0.95 – 0.78 (m, 12H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.4, 172.3, 61.7, 51.6, 45.8, 41.9, 38.8, 36.7, 32.0, 29.6, 29.5, 29.4, 25.8, 24.9, 23.2, 22.8, 22.1, 18.1, 14.2. HRMS (ESI): Calc'd for formula $C_{21}H_{41}N_3O_4Na^+$ $[M+Na]^+$ 422.2989, found 422.2989.



4.4o

4.4.5.15. *N*-((*S*)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-4-methyl-1-oxopentan-2-yl)tetradecanamide (**4.4o**):

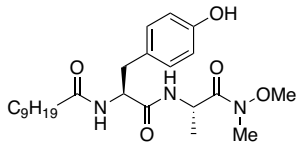
The product (228 mg, 99%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.15 (d, *J* = 7.7 Hz, 1H), 6.48 (d, *J* = 8.41, 1H), 4.84 (p, *J* = 7.1 Hz, 1H), 4.51 (m, 1H), 3.72 (s, 3H), 3.16 (s, 3H), 2.14 (t, *J* = 7.6 Hz, 2H), 1.67 – 1.44 (m, 5H), 1.43 – 1.17 (m, 30H), 0.91 – 0.78 (m, 11H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.4, 172.3, 61.7, 54.4, 51.6, 45.7, 41.8, 38.7, 36.7, 32.0, 29.8, 29.8, 29.7, 29.7, 29.5, 29.4, 25.8, 24.9, 23.2, 22.8, 22.1, 18.0, 14.2. HRMS (ESI): Calc'd for formula $C_{25}H_{50}N_3O_4^+$ $[M+H]^+$ 456.3796, found 456.381.



4.4p

4.4.5.16. *N*-((*S*)-3-(4-hydroxyphenyl)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)hexanamide (**4.4p**):

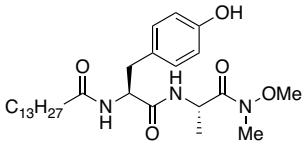
The product (114 mg, 58%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 7.01 (m, 2H), 6.71 (m, 2H), 6.18 (d, *J* = 8.0 Hz, 1H), 4.84 (m, 1H), 4.67 (m, 1H), 3.75 (m, 1H), 3.74 (s, 3H), 3.20 (s, 3H), 2.96 (d, *J* = 6.8 Hz, 2H), 2.80 (s, 4H), 2.15 (m, 2H), 1.56 (m, 1H), 1.43 (d, *J* = 7.0 Hz, 1H), 1.37 – 1.16 (m, 9H), 0.86 (t, *J* = 7.0 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 170.9, 155.8, 130.6, 127.8, 115.8, 54.5, 46.0, 38.9, 38.1, 36.8, 31.5, 25.5, 22.6, 18.3, 14.1. HRMS (ESI): Calc'd for formula $C_{20}H_{31}N_3O_5Na^+$ $[M+Na]^+$ 416.2156, found 416.2162.



4.4q

4.4.5.17. *N*-((*S*)-3-(4-hydroxyphenyl)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)decanamide (**4.4q**):

The product (191 mg, 85%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.98 (m, 2H), 6.71 (m, 2H), 6.35 (m, 1H), 4.85 (s, 1H), 4.68 (m, 1H), 3.74 (m, 3H), 3.19 (m, 3H), 3.19 (m, 3H), 2.80 (s, 2H), 2.23 – 2.08 (m, 2H), 1.55 (m, 3H), 1.46 – 1.20 (m, 20H), 1.18 (d, J = 7.0 Hz, 1H), 0.85 (t, J = 6.7 Hz, 4H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.7, 173.6, 171.0, 171.0, 156.0, 130.5, 130.5, 127.6, 115.8, 54.5, 45.8, 45.7, 45.6, 38.8, 38.1, 36.8, 32.0, 29.6, 29.5, 29.4, 29.4, 25.8, 25.7, 22.9, 18.6, 18.2, 17.9, 14.3. HRMS (ESI): Calc'd for formula $C_{24}H_{39}N_3O_5Na^+$ $[M+Na]^+$ 472.2782, found 472.2794.



4.4r

4.4.5.18. *N*-((*S*)-3-(4-hydroxyphenyl)-1-(((*S*)-1-(methoxy(methyl)amino)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)tetradecanamide (**4.4r**):

The product (220 mg, 87%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 6.95 (m, 2H), 6.69 (m, 2H), 4.84 (m, 1H), 4.69 (q, J = 7.1 Hz, 1H), 3.73 (m, 4H), 3.19 (d, J = 7.1 Hz, 4H), 3.00 – 2.84 (m, 3H), 2.79 (s, 2H), 2.22 – 2.08 (m, 3H), 1.65 – 1.49 (m, 3H), 1.40 – 1.13 (m, 36H), 0.86 (m, 5H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 173.7, 173.1, 171.1, 156.0, 130.5, 127.5, 115.8, 54.5, 45.8, 45.5, 38.8, 36.8, 32.1, 29.8, 29.8, 29.8, 29.7, 29.5, 29.5, 25.8, 22.9, 18.2, 14.3. HRMS (ESI): Calc'd for formula $C_{28}H_{48}N_3O_5^+$ $[M+H]^+$ 506.3588, found 506.3605.

4.4.6. Synthesis of *N*-acyl dipeptide aldehydes

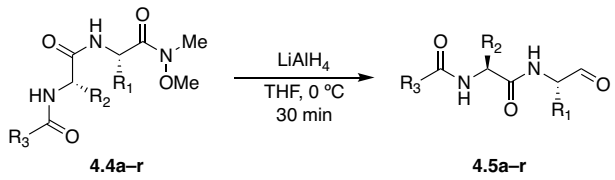
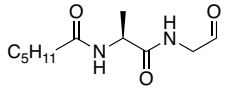


Figure 4.33: Synthesis of *N*-acyl dipeptide aldehydes 4.5a–r.

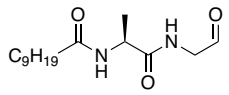
The *N*-acyl dipeptide Weinreb amide **4.4a–r** (1.0 equiv) was stirred in dry THF (0.15 M) at 0 ºC. To this solution was added $LiAlH_4$ (1 M in Et_2O, 1.1 equiv) and the reaction mixture was stirred at 0 ºC for 30 min. The reaction mixture was quenched by addition of 1 M HCl (1x initial reaction volume). The organic layer was removed in vacuo and the remaining aqueous layer then extracted with ethyl acetate (3 portions, each 2x initial reaction volume). The combined organic layers were washed with water and brine (each 2x initial reaction volume), dried over Na_2SO_4, and concentrated in vacuo to afford the products, which were judged to be suitably pure in their crude state by NMR.



4.5a

4.4.6.1. (*S*)-*N*-(1-oxo-1-((2-oxoethyl)amino)propan-2-yl)hexanamide (**4.5a**):

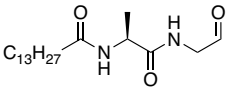
The product (10 mg, 60%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.63 (s, 1H), 7.08 (s, 1H), 6.20 (d, *J* = 7.4 Hz, 1H), 4.57 (m, 1H), 4.15 (m, 2H), 2.39 – 2.04 (m, 2H), 1.63 (m, 3H), 1.31 (m, 6H), 0.89 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 196.4, 173.8, 173.2, 50.3, 48.8, 36.7, 31.6, 25.5, 22.6, 18.2, 14.1. HRMS (ESI): Calc'd for formula $C_{11}H_{21}N_2O_3^+$ $[M+H]^+$ 229.1547, found 229.1548.



4.5b

4.4.6.2. (*S*)-*N*-(1-oxo-1-((2-oxoethyl)amino)propan-2-yl)decanamide (**4.5b**):

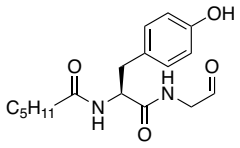
The product (26 mg, 67%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.59 (s, 1H), 7.38 (m, 1H), 6.43 (t, *J* = 7.0 Hz, 1H), 4.60 (m, 1H), 4.10 (m, 2H), 2.21 (m, 2H), 1.58 (m, 3H), 1.47 – 1.19 (m, 14H), 0.85 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 196.8, 173.7, 173.4, 50.2, 48.7, 36.7, 32.0, 29.6, 29.5, 29.5, 29.4, 25.8, 22.9, 18.5, 14.3. HRMS (ESI): Calc'd for formula $C_{15}H_{29}N_2O_3^+$ $[M+H]^+$ 285.2173, found 285.218.



4.5c

4.4.6.3. (*S*)-*N*-(1-oxo-1-((2-oxoethyl)amino)propan-2-yl)tetradecanamide (**4.5c**):

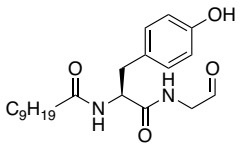
The product (39 mg, 53%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.61 (s, 1H), 7.26 (s, 1H), 6.32 (m, 1H), 4.60 (t, *J* = 7.2 Hz, 1H), 4.13 (d, *J* = 5.2 Hz, 2H), 2.21 (m, 2H), 1.61 (m, 3H), 1.48 – 1.12 (m, 22H), 0.87 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 196.7, 173.7, 173.3, 50.3, 48.7, 36.7, 32.1, 29.9, 29.9, 29.8, 29.8, 29.7, 29.7, 29.6, 29.5, 25.8, 25.8, 22.9, 18.4, 14.3. HRMS (ESI): Calc'd for formula $C_{19}H_{37}N_2O_3^+$ $[M+H]^+$ 341.2799, found 341.2834.



4.5d

4.4.6.4. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-oxo-1-((2-oxoethyl)amino)propan-2-yl)hexanamide (**4.5d**):

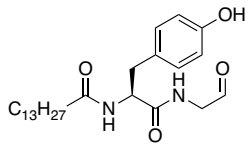
The product (4 mg, 15%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.56 (s, 1H), 7.07 (d, *J* = 8.5 Hz, 2H), 6.76 (d, *J* = 8.6 Hz, 2H), 6.46 (br s, 1H), 6.00 (d, *J* = 7.7 Hz, 1H), 5.17 (m, 1H), 4.67 (m, 1H), 4.09 (m, 2H), 3.01 (m, 2H), 2.18 (m, 2H), 1.68 – 1.46 (m, 2H), 1.36 – 1.17 (m, 4H), 0.88 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 172.6, 156.3, 130.5, 128.6, 115.3, 49.1, 35.7, 31.2, 29.6, 25.4, 22.4, 14.4. HRMS (ESI): Calc'd for formula $C_{17}H_{25}N_2O_4^+$ $[M+H]^+$ 321.1809, found 321.1818.



4.5e

4.4.6.5. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-oxo-1-((2-oxoethyl)amino)propan-2-yl)decanamide (**4.5e**):

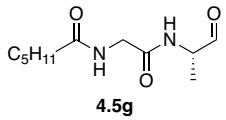
The product (25 mg, 39%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.4 (s, 1H), 7.88 (br s, 1H), 7.18 (m, 1H), 6.97 (m, 2H), 6.68 (m, 2H), 6.55 (d, *J* = 7.9 Hz, 1H), 4.73 (m, 1H), 4.10 (m, 2H), 2.93 (m, 2H), 2.14 (m, 2H), 1.61 – 1.46 (m, 2H), 1.32 – 1.12 (m, 12H), 0.85 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 197.3, 174.2, 172.4, 155.9, 130.5, 130.5, 127.5, 115.8, 115.8, 38.8, 32.0, 29.6, 29.5, 29.5, 29.4, 29.4, 25.8, 25.7, 22.8, 21.2, 14.4. HRMS (ESI): Calc'd for formula $C_{21}H_{33}N_2O_4^+$ $[M+H]^+$ 377.2435, found 377.2443.



4.5f

4.4.6.6. (*S*)-*N*-(3-(4-hydroxyphenyl)-1-oxo-1-((2-oxoethyl)amino)propan-2-yl)tetradecanamide (**4.5f**):

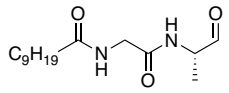
The product (13 mg, 19%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.44 (s, 1H), 6.99 (m, 2H), 6.71 (m, 2H), 6.45 (d, *J* = 7.9 Hz, 1H), 6.24 (br s, 1H), 4.72 (m, 1H), 4.22 (m, 1H), 4.01 (m, 2H), 3.05 – 2.84 (m, 2H), 2.38 – 2.08 (m, 2H), 1.59 (m, 2H), 1.24 (m, 20H), 0.87 (m, 3H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 199.9, 172.0, 155.7, 130.0, 128.2, 114.7, 61.1, 54.0, 49.2, 48.6, 44.0, 40.1, 36.9, 35.2, 35.2, 31.3, 29.1, 28.9, 28.8, 28.7, 28.5, 25.2, 22.1, 13.9. HRMS (ESI): Calc'd for formula $C_{25}H_{41}N_2O_4^+$ $[M+H]^+$ 433.3061, found 433.3068.



4.5g

4.4.6.7. (*S*)-*N*-(2-oxo-2-((1-oxopropan-2-yl)amino)ethyl)hexanamide (**4.5g**):

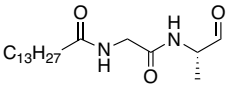
The product (6 mg, 24%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.53 (s, 1H), 7.04 (d, *J* = 6.6 Hz, 1H), 6.45 (m, 1H), 4.46 (m, 1H), 4.00 (m, 2H), 2.24 (m, 2H), 1.63 (m, 2H), 1.37 (d, *J*=7.4 Hz, 3H), 1.29 (m, 4H), 0.89 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.0, 174.3, 169.5, 54.9, 43.4, 36.5, 31.6, 25.5, 22.6, 14.5, 14.1. HRMS (ESI): Calc'd for formula $C_{11}H_{21}N_2O_3^+$ $[M+H]^+$ 229.1547, found 229.1546.



4.5h

4.4.6.8. (*S*)-*N*-(2-oxo-2-((1-oxopropan-2-yl)amino)ethyl)decanamide (**4.5h**):

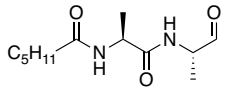
The product (17 mg, 26%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.50 (s, 1H), 7.45 (d, *J* = 6.6 Hz, 1H), 6.74 (t, *J* = 5.2 Hz, 1H), 4.39 (m, 1H), 4.02 (m, 2H), 2.22 (m, 2H), 1.60 (m, 2H), 1.33 (d, *J* = 7.4 Hz, 3H), 1.24 (d, *J* = 9.6 Hz, 12H), 0.85 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.3, 174.4, 169.6, 54.8, 43.3, 36.5, 32.0, 29.6, 29.5, 29.4, 29.4, 25.8, 22.8, 14.3, 14.3. HRMS (ESI): Calc'd for formula $C_{15}H_{28}N_2O_3Na^+$ $[M+Na]^+$ 307.1992, found 307.1994.



4.5i

4.4.6.9. (*S*)-*N*-(2-oxo-2-((1-oxopropan-2-yl)amino)ethyl)tetradecanamide (**4.5i**):

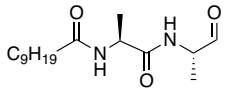
The product (23 mg, 30%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.52 (s, 1H), 7.25 (d, *J* = 7.8 Hz, 1H), 6.58 (t, *J* = 5.2 Hz, 1H), 4.42 (m, 1H), 4.02 (m, 2H), 2.24 (m, 2H), 1.60 (m, 3H), 1.44 – 1.07 (m, 22H), 0.86 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.2, 174.3, 169.5, 54.8, 43.3, 36.5, 32.1, 29.9, 29.9, 29.8, 29.7, 29.6, 29.5, 25.8, 22.9, 14.4, 14.3. HRMS (ESI): Calc'd for formula $C_{19}H_{36}N_2O_3Na^+$ $[M+Na]^+$ 363.2618, found 363.2597.



4.5j

4.4.6.10. *N*-((*S*)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)hexanamide (**4.5j**):

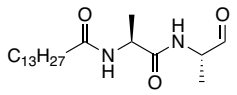
The product (21 mg, 52%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.52 (s, 1H), 7.25 (d, *J* = 6.6 Hz, 1H), 6.36 (d, *J* = 7.4 Hz, 1H), 4.61 (m, 1H), 4.39 (m, 1H), 2.19 (m, 2H), 1.61 (m, 2H), 1.40 (d, *J* = 7.0 Hz, 3H), 1.34 (d, *J* = 7.4 Hz, 3H), 1.29 (m, 4H), 0.88 (t, *J* = 6.9 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.2, 173.6, 173.0, 54.7, 48.8, 36.7, 31.6, 25.5, 18.8, 14.4, 14.1. HRMS (ESI): Calc'd for formula $C_{12}H_{23}N_2O_3^+$ $[M+H]^+$ 243.1703, found 243.171.



4.5k

4.4.6.11. *N*-((*S*)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)decanamide (**4.5k**):

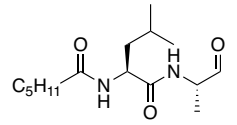
The product (33 mg, 63%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.52 (s, 1H), 7.28 (d, *J* = 6.7 Hz, 1H), 6.38 (d, *J* = 7.6 Hz, 1H), 4.62 (m, 1H), 4.38 (m, 1H), 2.19 (m, 2H), 1.61 (m, 2H), 1.40 (d, *J* = 7.0 Hz, 3H), 1.33 (d, *J* = 7.4 Hz, 3H), 1.30 – 1.20 (m, 12H), 0.86 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.3, 173.6, 173.0, 54.7, 48.7, 36.7, 36.6, 32.1, 29.6, 29.5, 29.5, 29.4, 25.8, 22.9, 18.9, 14.4, 14.3. HRMS (ESI): Calc'd for formula $C_{16}H_{30}N_2O_3Na^+$ $[M+Na]^+$ 321.2149, found 321.2158.



4.5l

4.4.6.12. *N*-((*S*)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)tetradecanamide (**4.5l**):

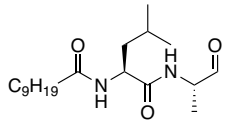
The product (10 mg, 13%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.53 (m, 1H), 7.19 (d, *J* = 6.7 Hz, 1H), 6.32 (d, *J* = 7.5 Hz, 1H), 4.58 (m, 1H), 4.39 (m, 1H), 2.21 (m, 2H), 1.61 (m, 2H), 1.39 (d, *J* = 7.0 Hz, 3H), 1.34 (d, *J* = 7.4 Hz, 3H), 1.31 – 1.19 (m, 20H), 0.87 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.2, 173.6, 172.9, 54.7, 48.8, 36.7, 32.1, 29. 9, 29.9, 29.8, 29.7, 29.7, 29.6, 29.5, 29.4, 25.8, 22.9, 18.7, 14.8, 14.4. HRMS (ESI): Calc'd for formula $C_{20}H_{39}N_2O_3^+$ $[M+H]^+$ 355.2955, found 355.2983.



4.5m

4.4.6.13. *N*-((*S*)-4-methyl-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)pentan-2-yl)hexanamide (**4.5m**):

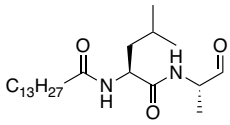
The product (49 mg, 83%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.48 (d, *J* = 11.1 Hz, 1H), 7.47 (d, *J* = 6.6 Hz, 1H), 6.53 (d, *J* = 8.3 Hz, 1H), 4.58 (m, 1H), 4.32 (m, 1H), 2.26 – 2.08 (m, 2H), 1.73 – 1.48 (m, 4H), 1.35 – 1.18 (m, 8H), 0.99 – 0.82 (m, 9H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.5, 173.8, 173.2, 54.6, 51.6, 41.5, 36.6, 31.5, 25.6, 25.0, 23.0, 22.5, 22.3, 14.4, 14.1. HRMS (ESI): Calc'd for formula $C_{15}H_{28}N_2O_3Na^+$ $[M+Na]^+$ 307.1992, found 307.2023.



4.5n

4.4.6.14. *N*-((*S*)-4-methyl-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)pentan-2-yl)decanamide (**4.5n**):

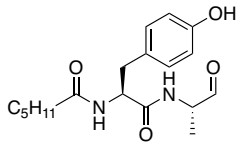
The product (36 mg, 43%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.48 (d, *J* = 10.9 Hz, 1H), 7.42 (m, 1H), 6.45 (m, 1H), 4.57 (m, 1H), 4.33 (m, 1H), 2.30 – 2.06 (m, 2H), 1.76 – 1.47 (m, 4H), 1.51 – 1.01 (m, 16H), 1.06 – 0.78 (m, 9H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.4, 173.8, 173.1, 54.7, 51.6, 41.5, 36.7, 32.1, 29.7, 29.5, 29.5, 29.4, 25.9, 25.0, 23.0, 22.3, 14.4, 14.3. HRMS (ESI): Calc'd for formula $C_{19}H_{36}N_2O_3Na^+$ $[M+Na]^+$ 363.2618, found 363.2632.



4.5o

4.4.6.15. *N*-((*S*)-4-methyl-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)pentan-2-yl)tetradecanamide (**4.5o**):

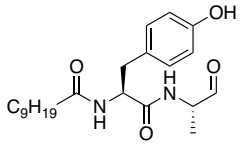
The product (45 mg, 62%) was isolated as an orange oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 9.50 (d, *J* = 12.1 Hz, 1H), 7.15 (d, *J* = 6.6 Hz, 1H), 6.19 (d, *J* = 8.2 Hz, 1H), 4.55 (m, 1H), 4.38 (m, 1H), 2.25 – 2.11 (m, 2H), 1.76 – 1.48 (m, 4H), 1.42 – 1.18 (m, 24H), 1.07 – 0.70 (m, 9H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.4, 173.9, 173.0, 54.7, 51.6, 50.8, 41.5, 41.2, 36.7, 32.1, 29.9, 29.8, 29.7, 29.5, 29.4, 25.9, 25.0, 23.0, 22.9, 22.3, 14.3. HRMS (ESI): Calc'd for formula $C_{23}H_{44}N_2O_3Na^+$ $[M+Na]^+$ 419.3244, found 419.3243.



4.5p

4.4.6.16. *N*-((*S*)-3-(4-hydroxyphenyl)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)hexanamide (**4.5p**):

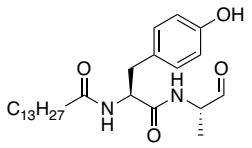
The product (4 mg, 17%) was isolated as a colorless oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.39 (s, 1H), 7.07 (d, *J* = 8.4 Hz, 2H), 6.76 (m, 2H), 6.38 (d, *J* = 6.6 Hz, 1H), 6.12 (d, *J* = 7.9 Hz, 1H), 4.65 (m, 1H), 4.34 (m, 1H), 3.00 (m, 2H), 2.19 (t, *J* = 7.6 Hz, 2H), 1.68 – 1.52 (m, 2H), 1.35 – 1.19 (m, 7H), 0.88 (t, *J* = 7.0 Hz, 3H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 172.0, 155.8, 130.0, 114.8, 97.7, 54.2, 48.6, 36.9, 35.1, 31.3, 30.7, 29.0, 24.9, 21.9, 13.8. HRMS (ESI): Calc'd for formula $C_{18}H_{27}N_2O_4^+$ $[M+H]^+$ 335.1965, found 335.1972.



4.5q

4.4.6.17. *N*-((*S*)-3-(4-hydroxyphenyl)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)decanamide (**4.5q**):

The product (13 mg, 18%) was isolated as an orange oil. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.41 (s, 1H), 9.33 (s, 1H), 7.01 (m, 2H), 6.83 (m, 1H), 6.74 (m, 2H), 6.43 (m, 1H), 4.71 (m, 1H), 4.31 (m, 1H), 3.10 – 2.78 (m, 2H), 2.34 – 2.02 (m, 2H), 1.68 – 1.43 (m, 2H), 1.33 – 1.05 (m, 15H), 0.92 (m, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.2, 173.9, 171.7, 155.7, 130.6, 127.9, 115.9, 54.8, 54.7, 36.8, 32.1, 32.1, 29.6, 29.5, 29.5, 29.4, 25.8, 22.9, 14.3, 14.3. HRMS (ESI): Calc'd for formula $C_{22}H_{33}N_2O_4^-$ $[M–H]^-$ 389.2446, found 389.2460.



4.5r

4.4.6.18. *N*-((*S*)-3-(4-hydroxyphenyl)-1-oxo-1-(((*S*)-1-oxopropan-2-yl)amino)propan-2-yl)tetradecanamide (**4.5r**):

The product (13 mg, 14%) was isolated as a colorless solid. ^{1}H NMR (400 MHz; $CDCl_3$): δ 9.54 (s, 1H), 9.40 (s, 1H), 9.31 (s, 1H), 6.99 (m, 2H), 6.73 (m, 2H), 6.48 (d, *J* = 8.2 Hz, 1H), 6.18 (d, *J* = 6.6 Hz, 1H), 4.70 (m, 1H), 4.30 (m, 1H), 2.96 (m, 2H), 2.40 – 1.96 (m, 2H), 1.72 – 1.47 (m, 2H), 1.45 – 1.04 (m, 23H), 0.87 (t, *J* = 6.6 Hz, 3H). ^{13}C NMR (101 MHz; $CDCl_3$): δ 199.4, 173.6, 130.6, 115.9, 54.7, 38.9, 36.7, 32.1, 30.5, 29.9, 29.0, 29.9, 29.8, 29.6, 29.5, 25.8, 22.9, 14.8, 14.3. HRMS (ESI): Calc'd for formula $C_{26}H_{43}N_2O_4^+$ $[M+H]^+$ 447.3217, found 447.3253.

4.4.7. Synthesis of tripeptide aldehydes

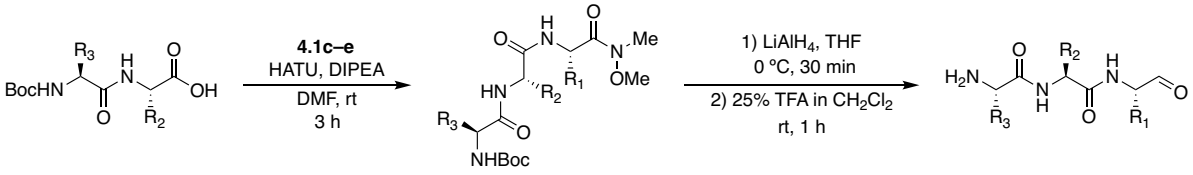


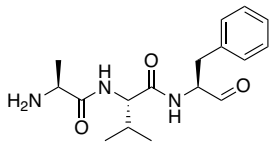
Figure 4.34: Synthesis of tripeptide aldehydes 4.8a–r.

To a solution of the Boc protected dipeptide **4.3a–f** (1.0 equiv) and the Weinreb amide **4.1c–e** (1.0 equiv) in DMF (0.6 M) was added HATU (1.1 equiv) and DIPEA (3.1 equiv) with stirring, under argon. After 3 h, the reaction mixture was diluted with ethyl acetate (4x initial volume) and quenched by addition of 1 M aqueous NaOH (4x initial volume). The organic layer was collected, and the aqueous layer was extracted with three portions of ethyl acetate (each 4x initial reaction volume). The combined organic layers were washed with water and brine (each 8x

initial reaction volume), dried over Na_2SO_4, filtered, and concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator.

The Boc-protected tripeptide Weinreb amide **4.6a–r** (1.0 equiv) was stirred in dry THF (0.15 M) at 0 ºC. To this solution was added $LiAlH_4$ (1 M in Et_2O, 1.1 equiv) and the reaction mixture was stirred at 0 ºC for 30 min. The reaction mixture was quenched by addition of 1 M HCl (1x initial reaction volume). The organic layer was removed in vacuo and the remaining aqueous layer then extracted with ethyl acetate (3 portions, each 2x initial reaction volume). The combined organic layers were washed with water and brine (each 2x initial reaction volume), dried over Na_2SO_4, and concentrated in vacuo to afford the crude products **4.7a–r**.

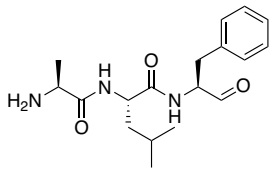
The Boc-protected tripeptide aldehyde **4.7a–r** was dissolved in anhydrous 25% TFA in DCM (0.05 M) and allowed to stir under argon for 1 h. The solvent was removed in vacuo using a Genevac EZ-2 Elite centrifugal evaporator. Anhydrous toluene (2x initial reaction volume) was added and evaporated to remove residual TFA. The resulting products were further purified by trituration with two 2 mL volumes of diethyl ether to afford the tripeptide aldehyde products **4.8a–r**, which were judged to be of suitable purity by NMR.



4.8a

4.4.7.1. (*S*)-2-((*S*)-2-aminopropanamido)-3-methyl-*N*-((*S*)-1-oxo-3-phenylpropan-2-yl)butanamide (**4.8a**):

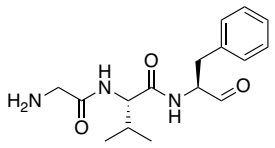
The product (44 mg, 48%) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.49 (d, J = 14.9 Hz, 1H), 8.51 (m, 1H), 8.16 (m, 4H), 7.46 – 6.98 (m, 5H), 4.40 (m, 1H), 4.22 (m, 1H), 3.94 (m, 2H), 3.26 – 2.64 (m, 2H), 1.92 (m, 1H), 1.47 – 1.16 (m, 6H), 0.93 – 0.63 (m, 3H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.2, 171.1, 158.8, 137.6, 129.1, 128.3, 126.4, 118.1, 115.2, 65.0, 48.0, 30.6, 19.1, 18.1, 17.8, 17.4, 15.2. HRMS (ESI): Calc'd for formula $C_{17}H_{25}N_3O_3Na^+$ $[M+Na]^+$ 342.1788, found 342.1795.



4.8b

4.4.7.2. (*S*)-2-((*S*)-2-aminopropanamido)-4-methyl-*N*-((*S*)-1-oxo-3-phenylpropan-2-yl)pentanamide (**4.8b**):

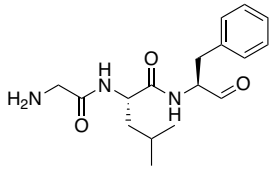
The product (33 mg, 50% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.47 (d, J = 17.3 Hz, 1H), 8.54 (m, 1H), 8.17 (m, 4H), 7.52 – 6.90 (m, 5H), 4.35 (m, 1H), 3.86 (m, 1H), 3.25 – 2.66 (m, 1H), 1.47 – 1.13 (m, 5H), 0.98 – 0.48 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.3, 172.2, 169.3, 137.7, 129.3, 129.2, 128.2, 128.2, 126.3, 65.0, 51.2, 48.0, 33.3, 24.0, 22.8, 21.7, 17.2, 15.2. HRMS (ESI): Calc'd for formula $C_{18}H_{28}N_3O_3^+$ $[M+H]^+$ 334.2125, found 334.2144.



4.8c

4.4.7.3. (*S*)-2-(2-aminoacetamido)-3-methyl-*N*-((*S*)-1-oxo-3-phenylpropan-2-yl)butanamide (**4.8c**):

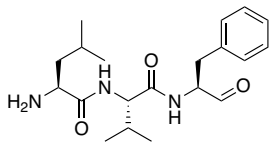
The product (49 mg, 50% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.49 (d, J = 18.8 Hz, 1H), 8.51 (m, 1H), 8.05 (br s, 4H), 7.29 – 7.08 (m, 5H), 4.41 (m, 1H), 4.30 (m, 1H), 3.62 (m, 2H), 3.24 – 2.57 (m, 2H), 1.94 (m, 1H), 1.83 (m, 1H), 0.90 – 0.67 (m, 6H). ^{13}C NMR (126 MHz; DMSO-d_6): δ 199.8, 170.7, 137.3, 128.9, 128.8, 128.0, 127.9, 126.1, 59.5, 57.2, 33.1, 30.5, 18.9, 18.7, 17.3. HRMS (ESI): Calc'd for formula $C_{16}H_{24}N_3O_3^+$ $[M+H]^+$ 306.1812, found 306.1818.



4.8d

4.4.7.4. (*S*)-2-(2-aminoacetamido)-4-methyl-*N*-((*S*)-1-oxo-3-phenylpropan-2-yl)pentanamide (**4.8d**):

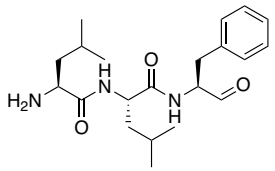
The product (53 mg, 34% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.47 (d, J = 21.0 Hz, 1H), 8.53 (m, 1H), 8.06 (m, 4H), 7.37 – 7.13 (m, 5H), 4.35 (m, 2H), 3.60 (m, 2H), 3.27 – 2.63 (m, 2H), 1.55 (m, 1H), 1.46 – 1.15 (m, 2H), 0.96 – 0.68 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.2, 172.1, 165.7, 158.5, 137.6, 129.3, 129.2, 128.2, 128.2, 64.9, 41.3, 38.9, 23.0, 22.8, 21.6, 21.5. HRMS (ESI): Calc'd for formula $C_{17}H_{26}N_3O_3^+$ $[M+H]^+$ 320.1969, found 320.1977.



4.8e

4.4.7.5. (*S*)-2-amino-4-methyl-*N*-((*S*)-3-methyl-1-oxo-1-(((*S*)-1-oxo-3-phenylpropan-2-yl)amino)butan-2-yl)pentanamide (**4.8e**):

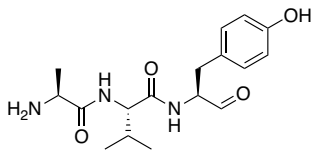
The product (39 mg, 48% over 3 steps) was isolated as an orange oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.47 (d, J = 9.5 Hz, 1H), 8.59 (m, 1H), 8.22 (m, 4H), 7.21 (m, 5H), 4.39 (m, 1H), 4.22 (m, 1H), 3.86 (m, 1H), 3.25 – 2.64 (m, 2H), 1.87 (m, 1H), 1.71 – 1.33 (m, 3H), 1.00 – 0.56 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.0, 168.9, 158.5, 137.7, 129.2, 129.0, 128.2, 118.7, 115.7, 50.7, 30.7, 23.5, 22.7, 22.7, 22.1, 21.9, 19.0, 18.3, 18.1, 15.2. HRMS (ESI): Calc'd for formula $C_{20}H_{32}N_3O_3^+$ $[M+H]^+$ 362.2438, found 362.244.



4.8f

4.4.7.6. (*S*)-2-amino-4-methyl-*N*-((*S*)-4-methyl-1-oxo-1-(((*S*)-1-oxo-3-phenylpropan-2-yl)amino)pentan-2-yl)pentanamide (**4.8f**):

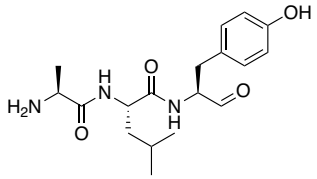
The product (53 mg, 49% over 3 steps) was isolated as an orange oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.46 (d, J = 14.1 Hz, 1H), 8.64 (m, 1H), 8.25 (m, 4H), 7.49 – 6.96 (m, 5H), 4.35 (m, 2H), 3.78 (br s, 1H), 3.28 – 2.63 (m, 2H), 1.76 – 1.24 (m, 6H), 1.19 – 0.48 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.6, 172.5, 169.2, 138.2, 129.7, 129.6, 128.7, 128.7, 126.8, 65.4, 60.3, 51.2, 24.0, 23.3, 23.2, 22.5, 22.4, 22.3, 22.3, 15.7. HRMS (ESI): Calc'd for formula $C_{21}H_{34}N_3O_3^+$ $[M+H]^+$ 376.2595, found 376.2605.



4.8g

4.4.7.7. (*S*)-2-((*S*)-2-aminopropanamido)-*N*-((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)-3-methylbutanamide (**4.8g**):

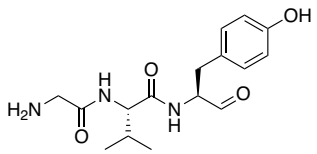
The product (71 mg, 83% over 3 steps) was isolated as brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.46 (d, J = 14.4 Hz, 1H), 8.45 (m, 1H), 8.12 (m, 4H), 7.00 (m, 2H), 6.64 (m, 2H), 4.24 (m, 1H), 3.95 (m, 2H), 3.17 – 2.58 (m, 2H), 1.93 (m, 1H), 1.40 – 1.18 (m, 3H), 0.95 – 0.68 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.3, 171.0, 169.4, 156.0, 130.0, 117.8, 115.1, 115.0, 114.9, 65.0, 48.0, 19.2, 19.1, 18.1, 17.8, 17.4, 15.2. HRMS (ESI): Calc'd for formula $C_{18}H_{26}N_3O_4^+$ $[M+H]^+$ 336.1918, found 336.1932.



4.8h

4.4.7.8. (*S*)-2-((*S*)-2-aminopropanamido)-*N*-((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)-4-methylpentanamide (**4.8h**):

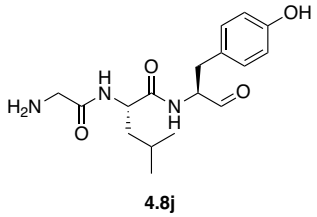
The product (52 mg, 66% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.45 (d, J = 17.0 Hz, 1H), 8.51 (m, 1H), 8.33 – 8.10 (m, 4H), 6.98 (m, 2H), 6.66 (m, 1H), 4.34 (m, 1H), 4.25 (m, 1H), 3.86 (m, 1H), 3.12 – 2.59 (m, 2H), 1.73 – 1.23 (m, 6H), 0.99 – 0.66 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.3, 171.9, 169.1, 155.8, 129.9, 127.1, 120.9, 117.9, 115.0, 64.7, 51.0, 47.8, 32.3, 23.9, 22.6, 21.3, 17.0, 14.9. HRMS (ESI): Calc'd for formula $C_{18}H_{28}N_3O_4^+$ $[M+H]^+$ 350.2074, found 350.2087.



4.8i

4.4.7.9. (*S*)-2-(2-aminoacetamido)-*N*-((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)-3-methylbutanamide (**4.8i**):

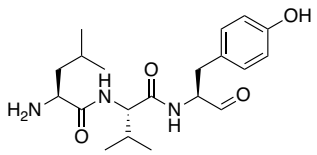
The product (50 mg, 55% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.46 (d, J = 16.3 Hz, 1H), 8.39 (m, 1H), 8.19 – 7.91 (m, 4H), 7.01 (m, 2H), 6.65 (m, 2H), 4.31 (m, 2H), 3.63 (m, 2H), 3.13 – 2.58 (m, 2H), 1.92 (m, 1H), 0.97 – 0.65 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.4, 170.6, 165.8, 158.6, 156.0, 130.0, 127.4, 118.0, 115.1, 65.0, 60.1, 32.7, 30.8, 19.1, 17.9, 17.6, 15.2. HRMS (ESI): Calc'd for formula $C_{16}H_{24}N_3O_4^+$ $[M+H]^+$ 322.1761, found 322.1755.



4.8j

4.4.7.10. (*S*)-2-(2-aminoacetamido)-*N*-((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)-4-methylpentanamide (**4.8j**):

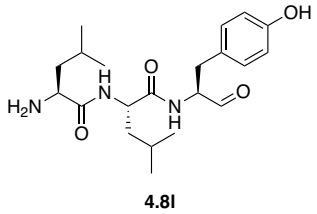
The product (52 mg, 58% over 3 steps) was isolated as a red oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 8.43 (d, J = 8.5 Hz, 1H), 8.36 (d, J = 8.0 Hz, 1H), 8.03 (m, 4H), 7.00 (m, 2H), 6.66 (m, 2H), 4.83 (br s, 1H), 4.43 (m, 1H), 3.74 – 3.45 (m, 2H), 2.88 – 2.62 (m, 2H), 1.57 (m, 1H), 1.47 – 1.31 (m, 2H), 0.96 – 0.78 (m, 6H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 171.7, 165.6, 158.2, 156.0, 130.0, 115.1, 65.0, 50.9, 41.4, 36.0, 24.0, 23.2, 21.6, 15.2. HRMS (ESI): Calc'd for formula $C_{17}H_{26}N_3O_4^+$ $[M+H]^+$ 336.1918, found 336.1934.



4.8k

4.4.7.11. (*S*)-2-amino-*N*-((*S*)-1-(((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)amino)-3-methyl-1-oxobutan-2-yl)-4-methylpentanamide (**4.8k**):

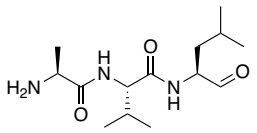
The product (45 mg, 49% over 3 steps) was isolated as an orange oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.45 (d, J = 10.3 Hz, 1H), 8.44 (d, J = 8.9 Hz, 1H), 8.29 (d, J = 7.6 Hz, 1H), 8.16 (m, 4H), 6.99 (m, 2H), 6.64 (m, 2H), 4.88 (br s, 1H), 4.25 (m, 1H), 3.88 (m, 1H), 2.87 – 2.61 (m, 2H), 1.95 (m, 1H), 1.68 – 1.42 (m, 3H), 0.97 – 0.80 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 170.9, 169.3, 158.7, 156.5, 130.3, 115.5, 58.1, 51.0, 39.4, 36.5, 31.4, 24.0, 23.2, 22.4, 19.6, 18.8. HRMS (ESI): Calc'd for formula $C_{20}H_{32}N_3O_4^+$ $[M+H]^+$ 378.2387, found 378.2405.



4.8l

4.4.7.12. (*S*)-2-amino-*N*-((*S*)-1-(((*S*)-1-(4-hydroxyphenyl)-3-oxopropan-2-yl)amino)-4-methyl-1-oxopentan-2-yl)-4-methylpentanamide (**4.8l**):

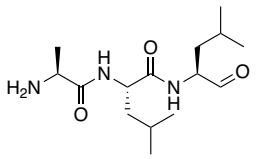
The product (23 mg, 26% over 3 steps) was isolated as an orange oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.44 (d, J = 15.4 Hz, 1H), 8.56 (m, 1H), 8.20 (m, 5H), 6.98 (m, 2H), 6.66 (m, 2H), 4.39 (m, 1H), 4.26 (m, 1H), 3.79 (m, 1H), 3.12 – 2.54 (m, 2H), 1.76 – 1.16 (m, 6H), 0.86 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 200.4, 172.0, 168.8, 156.0, 130.1, 115.0, 65.0, 50.8, 23.9, 23.5, 23.1, 22.8, 22.7, 21.9, 21.8, 21.7, 15.2. HRMS (ESI): Calc'd for formula $C_{21}H_{34}N_3O_4^+$ $[M+H]^+$ 392.2544, found 392.2561.



4.8m

4.4.7.13. (*S*)-2-((*S*)-2-aminopropanamido)-3-methyl-*N*-((*S*)-4-methyl-1-oxopentan-2-yl)butanamide (**4.8m**):

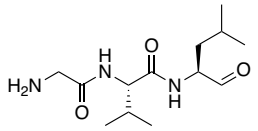
The product (31 mg, 58% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.39 (d, J = 1.4 Hz, 1H), 8.51 (m, 1H), 8.28 – 8.05 (m, 4H), 4.23 (m, 1H), 4.14 (m, 1H), 3.95 (m, 1H), 2.00 (m, 1H), 1.76 – 1.19 (m, 6H), 1.04 – 0.69 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.8, 171.6, 170.1, 65.4, 58.6, 57.1, 48.5, 36.6, 31.0, 30.9, 24.6, 24.6, 23.6, 23.5, 21.7, 21.5, 19.6, 18.7, 18.6, 17.8, 15.7. HRMS (ESI): Calc'd for formula $C_{14}H_{27}N_3O_3Na^+$ $[M+Na]^+$ 308.1945, found 308.1953.



4.8n

4.4.7.14. (*S*)-2-((*S*)-2-aminopropanamido)-4-methyl-*N*-((*S*)-4-methyl-1-oxopentan-2-yl)pentanamide (**4.8n**):

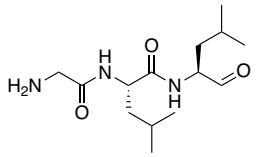
The product (37 mg, 49% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.38 (m, 1H), 8.54 (m, 1H), 8.27 – 8.01 (m, 4H), 4.0 (m, 1H), 4.12 (m, 1H), 3.87 (m, 1H), 1.72 – 1.56 (m, 2H), 1.59 – 1.40 (m, 3H), 1.39 – 1.23 (m, 4H), 1.00 – 0.76 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.3, 172.1, 169.4, 64.9, 56.6, 48.0, 24.1, 23.1, 22.9, 21.7, 21.2, 17.1, 15.2. HRMS (ESI): Calc'd for formula $C_{15}H_{30}N_3O_3^+$ $[M+H]^+$ 300.2282, found 300.2292.



4.8o

4.4.7.15. (*S*)-2-(2-aminoacetamido)-3-methyl-*N*-((*S*)-4-methyl-1-oxopentan-2-yl)butanamide (**4.8o**):

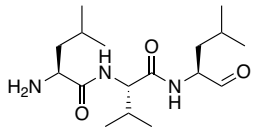
The product (46 mg, 71% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.40 (m, 1H), 8.52 (m, 1H), 8.07 (m, 4H), 4.34 (m, 1H), 4.13 (m, 1H), 3.69 (m, 2H), 2.01 (m, 1H), 1.73 – 1.33 (m, 3H), 0.98 – 0.72 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.3, 171.0, 166.0, 64.9, 57.6, 56.8, 56.6, 36.2, 30.9, 30.8, 24.1, 23.1, 21.3, 21.1, 19.1, 17.8, 15.2. HRMS (ESI): Calc'd for formula $C_{13}H_{26}N_3O_3^+$ $[M+H]^+$ 272.1969, found 272.1976.



4.8p

4.4.7.16. (*S*)-2-(2-aminoacetamido)-4-methyl-*N*-((*S*)-4-methyl-1-oxopentan-2-yl)pentanamide (**4.8p**):

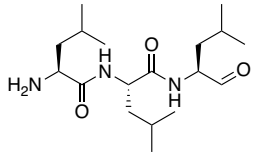
The product (29 mg, 44% over 3 steps) was isolated as a brown oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.38 (d, J = 6.7 Hz, 1H), 8.56 (m, 1H), 8.07 (m, 4H), 4.44 (m, 1H), 4.11 (m, 1H), 3.57 (m, 2H), 1.75 – 1.28 (m, 6H), 1.09 – 0.67 (m, 12H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.4, 172.2, 165.8, 56.6, 51.1, 41.4, 36.2, 24.1, 23.1, 22.9, 21.6, 21.3. HRMS (ESI): Calc'd for formula $C_{14}H_{28}N_3O_3^+$ $[M+H]^+$ 286.2125, found 286.2135.



4.8q

4.4.7.17. (*S*)-2-amino-4-methyl-*N*-((*S*)-3-methyl-1-(((*S*)-4-methyl-1-oxopentan-2-yl)amino)-1-oxobutan-2-yl)pentanamide (**4.8q**):

The product (31 mg, 36% over 3 steps) was isolated as a yellow oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.39 (m, 1H), 8.58 (m, 1H), 8.17 (m, 4H), 4.18 (m, 1H), 3.88 (m, 1H), 1.96 (m, 1H), 1.72 – 1.33 (m, 7H), 1.04 – 0.70 (m, 18H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.0, 171.0, 168.8, 56.7, 50.7, 36.1, 30.6, 24.0, 23.5, 23.1, 22.7, 22.5, 22.2, 22.0, 21.1, 21.0, 19.1, 18.6, 18.4. HRMS (ESI): Calc'd for formula $C_{17}H_{34}N_3O_3^+$ $[M+H]^+$ 328.2595, found 328.2587.



4.8r

4.4.7.18. (*S*)-2-amino-4-methyl-*N*-((*S*)-4-methyl-1-(((*S*)-4-methyl-1-oxopentan-2-yl)amino)-1-oxopentan-2-yl)pentanamide (**4.8r**):

The product (18 mg, 29% over 3 steps) was isolated as a yellow oil. ^{1}H NMR (400 MHz; DMSO-d_6): δ 9.37 (m, 1H), 8.70 (m, 1H), 8.32 (br s, 4H), 4.40 (m, 1H), 4.10 (m, 1H), 3.79 (m, 1H), 1.80 – 1.39 (m, 9H), 1.05 – 0.76 (m, 18H). ^{13}C NMR (101 MHz; DMSO-d_6): δ 201.4, 172.0, 168.7, 64.9, 56.5, 50.8, 44.0, 36.1, 24.0, 23.5, 23.1, 22.9, 22.6, 22.2, 22.1, 21.8, 21.1. HRMS (ESI): Calc'd for formula $C_{18}H_{35}N_3O_3Na^+$ $[M+Na]^+$ 364.2571, found 364.2581.

4.4.8. Synthesis of chloromethyl ketone probes

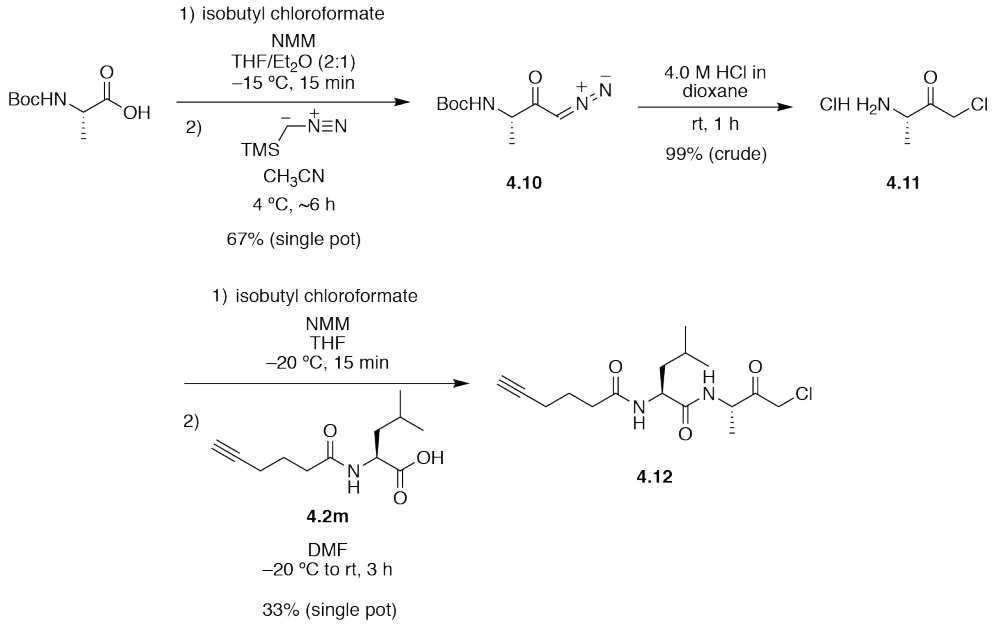


Figure 4.35: Synthesis of chloromethyl ketone probe 4.12.

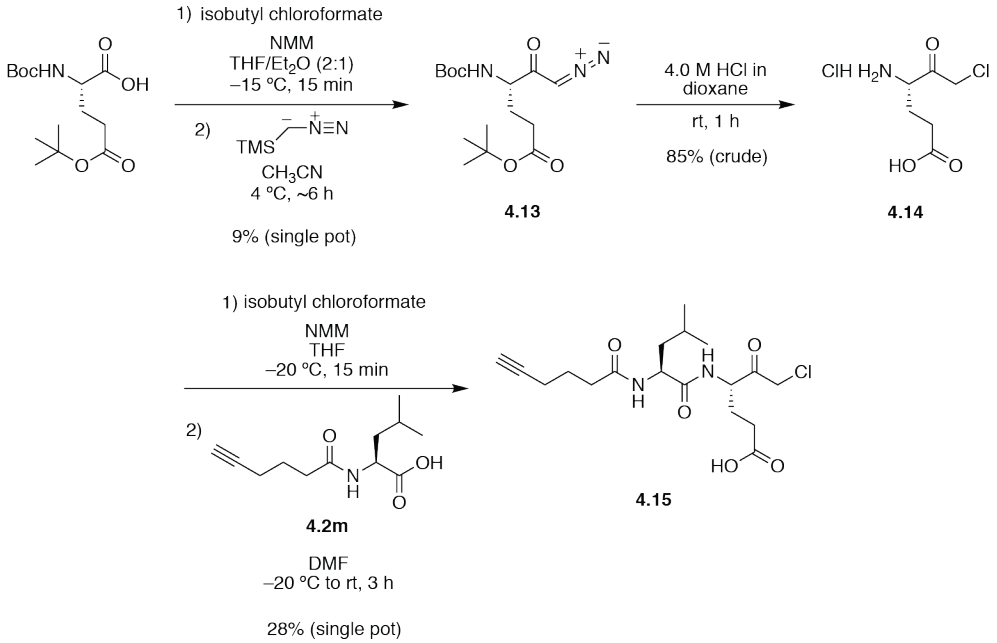
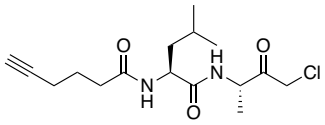


Figure 4.36: Synthesis of chloromethyl ketone probe 4.15.

4.10 (from Boc-L-Ala) and **4.13** from (Boc-L-Glu(*O*-*t*Bu)) were synthesized according to previously reported conditions,[124] with the following modifications: 1) TMS-diazomethane (2.0 M in hexanes, 2.0 equiv) was used rather than diazomethane, as previously reported;[125] 2) The reaction with TMS-diazomethane was conducted in acetonitrile (0.3 M), as previously reported;[125] 3) the reaction mixture was stirred until TLC (2:1 hexanes/ethyl acetate) indicated completion (~6 h); 4) excess TMS-diazomethane was quenched by addition of 10% aqueous citric acid (1x initial reaction volume); 5) the crude products were purified by flash column chromatography (with silica) using hexanes/ethyl acetate (3:1). The characterization data for **4.10** matched previously reported results.[169] **4.13** was not fully characterized here, though it has previously been reported in the literature.[170]

Boc deprotection and conversion of these compounds to the chloromethyl ketones **4.11** and **4.14** was accomplished using previously reported conditions,[124] except that the reaction mixture was stirred for 1 h at rt. The reaction mixtures were concentrated in vacuo using a Genevac EZ-2 Elite centrifugal evaporator to afford the crude deprotected products.

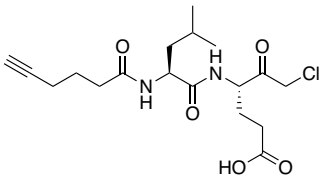
Coupling of **4.11** and **4.14** to **4.2m** to afford **4.12** and **4.15** was accomplished using previously reported conditions.[126] The crude products were preliminarily purified by trituration with hexanes, and the semi-pure compounds were then further purified by flash chromatography on silica gel using dichloromethane/methanol (gradient of 99:1 to 90:10).



4.12

4.4.8.1. *N*-((*S*)-1-(((*S*)-4-chloro-3-oxobutan-2-yl)amino)-4-methyl-1-oxopentan-2-yl)hex-5-ynamide (**4.12**)

The product (74 mg, 33%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 6.34 (d, *J* = 8.1 Hz, 1H), 4.57 (m, 1H), 3.44 (s, 2H), 2.36 (t, *J* = 7.4 Hz, 2H), 2.23 (m, 2H), 1.83 (quintet, *J* = 7.1 Hz, 2H), 1.67 (m, 2H), 1.55 (t, *J* = 8.6 Hz, 1H), 0.92 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 176.0, 173.3, 83.6, 69.5, 51.0, 50.6, 41.4, 35.0, 25.1, 24.3, 23.0, 22.0, 17.9. HRMS (ESI): Calc'd for formula $C_{16}H_{25}ClN_2O_3^+$ $[M+H]^+$ 329.1626, found 329.1636.



4.15

4.4.8.2. (*S*)-6-chloro-4-((*S*)-2-(hex-5-ynamido)-4-methylpentanamido)-5-oxohexanoic acid (**4.15**)

The product (11 mg, 28%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 7.13 (d, *J* = 7.1 Hz, 1H), 6.03 (d, *J* = 7.9 Hz, 1H), 4.76 (s, 1H), 4.46 (m, 1H), 4.30 (s, 2H), 2.42 (m, 4H), 2.25 (m, 3H), 1.98 (q, *J* = 2.6 Hz, 1H), 1.86 (m, 2H), 1.64 (m, 2H), 1.25 (m, 4H), 0.93 (m, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 200.5, 173.7, 172.9, 83.5, 69.6, 55.8, 52.3, 51.9, 46.7, 41.4, 41.1, 35.0, 29.9, 29.8, 25.9, 25.0, 24.2, 23.0, 22.3, 18.0. HRMS (ESI): Calc'd for formula $C_{20}H_{31}ClN_2O_7^-$ $[M–H]^-$ 445.1747, found 445.1786.

4.4.9. Synthesis of iodomethyl ketone probe

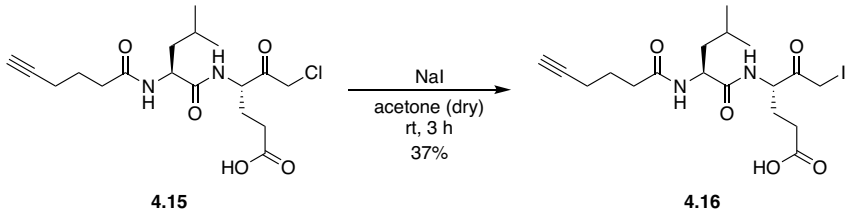


Figure 4.37: Synthesis of iodomethyl ketone probe 4.16.

4.16 was synthesized using previously reported conditions for the conversion of chloromethyl ketones to iodomethyl ketones,[127] except that the reaction only proceeded for 3 h. The crude product was purified by flash chromatography on silica gel using dichloromethane/methanol (gradient of 99:1 to 90:10).

4.4.9.1. (*S*)-4-((*S*)-2-(hex-5-ynamido)-4-methylpentanamido)-6-iodo-5-oxohexanoic acid (**4.16**)

The product (5 mg, 37%) was isolated as a colorless oil. ^{1}H NMR (500 MHz; $CDCl_3$): δ 6.95 (m, 1H), 5.90 (m, 1H), 4.84 (m, 1H), 4.45 (m, 1H), 2.39 (m, 6H), 2.27 (m, 4H), 2.00 (m, 2H), 1.86 (m, 3H), 1.65 (m, 4H), 1.26 (d, J = 8.4 Hz, 2H), 0.94 (q, J = 6.9 Hz, 6H). ^{13}C NMR (126 MHz; $CDCl_3$): δ 201.1, 173.7, 173.6, 172.8, 172.5, 172.2, 83.6, 69.6, 58.2, 55.6, 53.7, 52.8, 52.2, 52.1, 52.0, 51.8, 41.4, 41.2, 35.1, 30.1, 29.9, 29.7, 27.3, 27.0, 26.3, 25.1, 25.1, 24.2, 24.2, 23.1, 23.1, 22.4, 22.4, 22.3, 18.0, 3.0. HRMS (ESI): Calc'd for formula $C_{20}H_{31}IN_2O_7^-$ $[M–H]^-$ 537.1103, found 537.1121.

4.4.10. Human protease inhibitor assays

Screens of the 50-compound peptide aldehyde library against human proteases were conducted by GenScript (Piscataway, NJ) (initial single concentration screen for calpain 1) and

Reaction Biology Corporation (Malvern, PA) (all other assays). Solid stocks of compounds were sent to the organizations, where they were reconstituted as 100 mM stock solutions in DMSO and stored at –20 ºC.

For the calpain 1 inhibition assay performed by GenScript, the assay mixture (9 μL) contained buffer and protease as indicated in Table 4.5. The protease was incubated with peptide aldehyde compounds at rt for 10 min in order to allow for interaction between the protease and the inhibitor . The fluorogenic protease substrate (Table 4.5) was then added, and the assay mixtures were monitored for changes in fluorescence over time (485 nm excitation/520 nm emission) at rt for 20 min. Assays were performed in duplicate. Inhibitor efficiency was calculated from the slope of the linear portion of the curve as compared with the negative control (no inhibitor) and blank (buffer and substrate with no enzyme) and is reported as the mean of both trials. Positive control inhibitor B27-WT[171] was used to validate the assay.

For protease inhibition assays performed by Reaction Biology Corporation, each assay mixture (20 μL) contained buffer and protease as indicated in Table 4.5. The protease was incubated with peptide aldehyde compounds at rt for 20 min in order to allow for interaction between the protease and the inhibitor. The fluorogenic protease substrate (Table 4.5) was then added and the assay mixtures were monitored for fluorescence (355 nm excitation/460 nm emission) at rt over 2 h. Assays were performed in duplicate. Inhibitor efficiency was calculated from the slope of the linear portion of the curve as compared with the negative control (no inhibitor) and is reported as the mean of both trials. Positive control inhibitors were used to validate these assays as indicated in Table 4.5. For determination of IC_{50} values, experiments were performed in duplicate over threefold serial dilutions from either 0.00051–10 μM (calpain 1, cathepsins B and L) or 0.0051–100 μM (caspase 1). Curves were individually fit to determine

IC_{50}. The reported values are the mean of values calculated from these two series of serial dilutions.

Table 4.5: Conditions for human protease inhibition assays.

Protease	Amount of protease (nM)	Buffer	Substrate	Amount of substrate	Positive control inhibitor
Calpain 1	125.55	calpain assay buffer (SensoLyte® 520 Calpain Activity Assay Kit, AnaSpec)	5-FAM/QXL™	15 µM	B27-WT[171]
Cathepsin B	0.5	25 mM MES pH 6.0, 50 mM NaCl, 0.005% Brij35, 5 mM DTT	Z-FR-AMC	10 µM	E-64[172]
Cathepsin L	15.8	400 mM NaOAc pH 5.5, 4 mM EDTA, 8 mM DTT	Z-FR-AMC	10 µM	E-64
Cathepsin S	1.9	75 mM Tris pH 7.0, 1 mM EDTA, 0.005% Brij35, 3 mM DTT	Z-FR-AMC	10 µM	E-64
Cathepsin V	1.2	25 mM NaOAc pH 5.5, 0.1 M NaCl, 5 mM DTT	Z-FR-AMC	10 µM	E-64
Caspase 1	1	50 mM HEPES pH 7.4, 1 M sodium citrate, 100 mM NaCl, 0.01% CHAPS, 0.1 mM EDTA, 10 mM DTT	Ac-LEHD-AMC	5 µM	Ac-IETD-CHO[173]
Caspase 3	0.4	50 mM HEPES pH 7.4, 100 mM NaCl, 0.01% CHAPS, 0.1 mM EDTA, 10 mM DTT	Ac-DEVD-AMC	5 µM	Ac-DEVD-CHO[173]
Caspase 8	0.6	50 mM HEPES pH 7.4, 1 M sodium citrate, 100 mM NaCl, 0.01% CHAPS, 0.1 mM EDTA, 10 mM DTT	Ac-LEHD-AMC	5 µM	Ac-IETD-CHO[173]
Cathepsin G	430	100 mM Tris-HCl pH 8.0, 50 mM NaCl, 10 mM $CaCl_2$, 0.025% CHAPS, 1.5 mM DTT	Suc-AAPF-AMC	10 µM	chymostatin[174]
Elastase	0.5	100 mM Tris-HCl pH 8.0, 50 mM NaCl, 10 mM $CaCl_2$, 0.025% CHAPS, 1.5 mM DTT	AR-AMC	10 µM	Gabexate Mesylate[175]
Trypsin	0.1	25 mM Tris pH 8.0, 100 mM NaCl, 0.01% Brij35	Pefafluor® TH	12 µM	Gabexate Mesylate

4.4.11. Screens of peptide aldehydes for microbial growth inhibition

4.4.11.1. Zone of inhibition assays for the ESKAPE pathogens

Agar overlay zone of inhibition assays for assessing antibacterial activity of the peptide aldehydes against the ESKAPE pathogens were performed by iFyber (Ithaca, NY). Strains used for this experiment were *Enterococcus faecium* ATCC 19434, *Staphylococcus aureus* ATCC 29213, *Staphylococcus aureus* USA300, *Klebsiella pneumoniae* ATCC 13883, *Acinetobacter baumannii* ATCC 19404, *Pseudomonas aeruginosa* BAA-47, and *Enterobacter* sp. ATCC 27985. Overnight cultures (18 h) were grown in tryptic soy broth (TSB). Bacterial lawns were prepared by diluting the overnight cultures 20x in 0.75% tryptic soy agar (TSA) held at 55 °C, pouring 2 mL of the inoculated soft agar onto a 1.5% TSA plate, and allowing to solidify for ~30 min. 10 mM, 3.33 mM, and 1.11 mM stock solutions of the peptide aldehydes were prepared in 10% DMSO in water. Test compound solutions (10 µL each) were spotted on to the inoculated soft agar plates. The plates were incubated at 37 ºC for 24 h and then imaged and observed for ZOIs.

4.4.11.2. Zone of inhibition assays for common gut commensals and pathogens

A similar procedure was used to perform agar-overlay zone of inhibition assays for the screen of common gut commensals and pathogens in house. Strains were grown under the conditions indicated in Table 4.6 at 37 ºC. Rectangular plates containing the appropriate 1.5% agar medium for each strain were prepared as indicated in Table 4.6 and rendered anaerobic where indicated. Bacterial lawns were prepared by diluting the overnight cultures 20x in 0.75% TSA (8 mL) held at 50 °C, pouring the inoculated soft agar onto the prepared rectangular plates, and allowing them to solidify for ~30 min. 10 mM stock solutions of the peptide aldehydes were

prepared in 10% DMSO in water. Test compound solutions (5 μL each) were then spotted on to the inoculated soft agar plates, along with sterile 10% DMSO as a negative control. The plates were allowed to stand for 20 minutes and were then inverted and incubated at 37 ºC for 16 h. For anaerobic strains, the plates were incubated at rt in an anaerobic glovebox containing a 10% hydrogen/10% carbon dioxide/bal. nitrogen atmosphere for ~48 h. The plates were then inspected for full and partial zones of inhibition. Each experiment was performed in duplicate.

Table 4.6: Growth conditions for gut microbial zone of inhibition screening and MIC determination.

(TY = Tryptone Yeast, TS = Tryptic Soy, RCM = Reinforced Clostridial Medium.)

Strain	Agar/broth	Atmosphere
Bacillus subtilis 168	TY	aerobic
Escherichia coli MS 200-1	TS	aerobic
Staphylococcus epidermidis ATCC 12228	Nutrient	aerobic
Bacteroides dorei CL02T12C06	RCM	anaerobic
Bacteroides fragilis 168	RCM	anaerobic
Clostridioides difficile 630Δerm	RCM	anaerobic
Parabacteroides merdae ATCC 43184	RCM	anaerobic

4.4.11.3. MIC determination for gut commensals and pathogens

For MIC determination, an adaptation of the standard procedure was used.[96] For the aerobic strains, single colonies of the bacterial strain of interest were obtained by streaking frozen glycerol stocks on an agar plate. From this plate, a single colony was inoculated into the indicated liquid medium (Table 4.6). The anaerobic strains were directly inoculated from frozen glycerol stocks into an anaerobic aliquot of the indicated liquid medium (Table 4.6). The culture was incubated at 37 ºC until it contained ~1 x 10^8 colony forming units (cfu)/mL. The reference protocol recommends making this determination by comparing turbidity with a McFarland Standard 0.5 but also suggests that this corresponds with an OD625 nm reading of 0.08–0.13.[96] We made this determination by measuring OD625 on a GENESYS™ 20 Visible

Spectrophotomer (Thermo Scientific™). Cultures that exceeded the desired OD625 range were diluted with sterile water to bring them within the desired range. To prepare plates containing the desired range of concentration of the peptide aldehydes, stock antibiotic solutions were prepared aerobically (1280 μg/mL). From this point forward, the experiments for the strains that grow anaerobically were set up in a Coy glove box containing a 10% hydrogen/10% carbon dioxide/bal. nitrogen atmosphere. Antibiotic stock solutions were diluted to an initial concentration of 128 μg/mL, and two-fold serial dilutions were prepared across 10 wells of the plate to result in final concentrations of 128 – 0.25 μg/mL Each plate also contained a sterile broth negative control and a 'no added compound' positive control. The final volume of each experimental well was 100 μL. To initiate the experiment, the prepared culture was diluted 1:100 and then 50 μL was inoculated in each well, resulting in a final desired concentration of ~5 x 10^5 cfu/mL. For both the aerobic and anaerobic conditions, the plates were incubated overnight at 37 ºC. MIC was determined by reading off the lowest compound concentration at which no growth was observed.

4.4.12. Inhibition of secreted protease activity by peptide aldehydes

The Pierce Fluorescent Protease Assay Kit was used to assess secreted protease activity of gut microbial species and inhibition by peptide aldehydes employing a modified version of the manufacturer's instructions. The buffer (Tris-buffered saline, TBS) contained 25 mM Tris pH 7.2 and 0.15 M NaCl. The FTC-casein stock solution contained 5 mg/mL FTC-casein in DI water, and the FTC-casein working reagent was prepared by diluting the FTC-casein stock solution 1:500 in TBS.

For the assays, 5 mL cultures were grown according to the conditions indicated in Table 4.7. For all four strains, we initially examined the undiluted supernatants and three dilutions (1:2, 1:4, 1:8) for their secreted protease activity. Dilutions of the *E. faecalis* supernatant (1:4) and the *C. sporogenes* supernatant (1:2) demonstrated the highest signal in this initial screen (arbitrary units), so we used these dilutions in subsequent experiments, as indicated in Table 4.7. Each culture was pelleted by centrifugation, and the supernatant was filtered through a 0.2 μm filter and diluted with TBS where indicated. 50 μL of the (diluted) culture supernatant was added to the wells of a Corning white half area NBS 96-well assay plate. 10 mM stock solutions of peptide aldehydes were prepared in DMSO, and 1 μL of each stock solution was added to the desired wells for a final concentration of 100 μM in the final assay volume (100 μL). The plate was pre-incubated at 37 ºC for 10 min to allow for protease-inhibitor interaction. 50 μL of the FTC-casein working reagent was then added to each well, and fluorescence was measured in a BioTek SynergyHTX multi-mode microplate reader over 20 minutes at 1 minute intervals (485/528 nm excitation/emission, read height 1 mm, no shaking, 37 ºC). Percent inhibition was calculated by comparing initial rates of fluorescence change in the experimental wells with the mean signal from negative control (no inhibitor) wells.

Table 4.7: Growth conditions for gut microbial secreted protease activity assays.
(PYG = Peptone Yeast Glucose)

Strain	Agar/broth	Growth Conditions	Culture time	Dilution in assay
Bacillus cereus ATCC 53522	TS	30 ºC, aerobic	overnight	undiluted
Clostridium sporogenes ATCC 15579	PYG	37 ºC, anaerobic	overnight	1:2
Enterococcus faecalis TX0104	TS	37 ºC, aerobic	overnight	1:4
Klebsiella aerogenes ATCC 13048	Nutrient	30 ºC, aerobic	mid-log phase (~6 h)	undiluted

4.4.13. General procedure for the copper-catalyzed click reaction of alkyne probe-tagged proteases and proteomes with azides

Labeling of alkyne-tagged proteomes and preparation of precipitated protein pellets were conducted according to the Click-iT® Protein Reaction Buffer Kit protocols, using either tetramethylrhodamine (TAMRA) azide (Invitrogen, **4.17**) or PEG4 carboxamide-6-azidohexanyl biotin (Biotin Azide, Invitrogen, **4.18**) as the azide reagent. For SDS-PAGE analysis, the protein pellets were resolubilized in Lamelli sample buffer (12 μL) by heating the sample for 10 minutes at 70 °C. After performing SDS PAGE separation, the gel was imaged on a Gel Doc™ EZ Gel Documentation System (BioRad), stained with Coomassie, and imaged again.

4.4.14. Human calpain labeling by activity probes

Each reaction contained calpain (8 μL, 0.5 mg/mL, AnaSpec) and 8 μL calpain assay buffer (SensoLyte® 520 Calpain Activity Assay Kit, AnaSpec). Alkyne probes **4.12** and **4.16** were added from 10 mM DMSO stocks to a final concentration of 100 μM and the assay mixture was incubated at rt for 30 min. These samples were then labeled with TAMRA azide (**4.17**) and analyzed by SDS-PAGE according to the procedure described in 4.4.13.

4.4.15. Labeling of *C. difficile* lysates with tetramethylrhodamine

The protocol for preparation of labeled *C. difficile* proteomes was adapted from a previously published procedure, with many modifications.[176] *C. difficile* 630Δerm was inoculated 1:100 from an overnight culture in RCM and allowed to grow anaerobically to OD ~0.4. Cultures were normalized to the same OD600 (~0.4) and 4 mL of each normalized culture collected and pelleted by centrifugation (4,000 rpm x 10 min, 4 ºC). The cell pellets were washed with PBS

(pH 7.2, 2 x 600 µL) and resuspended in 600 µL of lysis buffer (10 mM Tris pH 8.0 with 0.1% (w/v) CHAPS). Cells were lysed by sonication on ice using a Branson Digital Sonifier equipped with a Double Stepped Microtip (10 s pulse, 30 s rest, 25% amplitude, 4 cycles). The alkyne-iodomethyl ketone probes were then added from 10 mM stocks in DMSO to a final concentration of 10 µM, and the reaction mixtures were incubated for 30 min at 30 ºC. Insoluble precipitates were pelleted by centrifugation (13,000 rpm x 10 min, 4 ºC), and 800 µL of ice cold acetone was added to each reaction mixture. The reaction mixtures were incubated overnight at –20 ºC. As this did not result in protein precipitation, an additional 1.6 mL of acetone was added to the reaction mixtures and they were incubated at –20 ºC for an additional 20 min. The precipitated protein was pelleted by centrifugation (13,000 rpm x 20 min, 4 ºC) and washed with 80% acetone in water (2 x 1 mL) and acetone (1 mL). The protein pellets were allowed to air dry and then resuspended in 50 µL resuspension buffer (50 mM TrisHCl pH 8.0, 8 M urea) by vortexing for 10 min. Resuspended protein concentrations were quantified using the Bradford assay. 20 µg of each sample was then labeled with TAMRA azide (**4.17**) and analyzed by SDS-PAGE according to the procedure described in 4.4.13.

4.4.16. Enrichment of *C. difficile* lysates with biotin tag

For the pilot enrichment of tagged *C. difficile* proteins with the biotin probe, alkyne-labeled proteomes were prepared according to the procedure described in 4.4.15, and 200 µg of each sample were labeled with **4.18** according to the procedure described in 4.4.13. The protein pellet generated from the click reaction was resuspended in 50 µL 2% SDS in PBS and heated at 70 ºC for 10 min. Biotin enrichment was performed according to a previously reported procedure,[123] using 10 µL of high capacity streptavidin agarose beads (20 µL slurry). The denatured protein

samples were diluted with 400 µL of PBS containing 0.2% SDS before addition to the washed beads. After overnight incubation, the beads were pelleted by centrifugation (800 rpm x 2 min, 4 ºC) and washed as described. The beads were denatured by heating in 20 µL SDS-PAGE loading buffer (10 min at 70 ºC) and proteins analyzed by SDS-PAGE gel and silver staining alongside a BSA standard curve (200–6.25 ng).

4.4.17. Proteomics analysis

For the preparation of samples for workflow A, the bead-bound proteomes were prepared as described in 4.4.16 and stored as a suspension in 50 µL buffer (50 mM Tris pH 8, 150 mM NaCl) at –20 ºC until analysis. Each experimental condition was prepared in duplicate, while the negative control (no probe) was a single trial. Trypsin digestion, isobaric labeling, and LC-MS analysis were performed according to previously published procedures.[129] The *C. difficile* strain 630 proteome database (UP000001978) was used to identify proteins.

For preparation of samples for workflow B, 50 mL cultures of *C. difficile* were grown to OD ~0.4, pelleted by centrifugation, washed with PBS (2 x 20 mL) and resuspended in 600 µL of lysis buffer. The samples were lysed, labeled, and precipitated as described in 4.4.15, except that each alkyne probe was used at a concentration of 100 µM. 500 µg of each of these samples were labeled with **4.18** as described in 4.4.13. Each condition was prepared in duplicate. Trypsin digestion, anti-biotin antibody enrichment, and LC-MS analysis were performed as previously described.[130] The anti-biotin antibody is a commercially available agarose-bound antibody (Biotin Antibody Agarose, ImmuneChem Pharmaceuticals Inc.). The data was analyzed using MaxQuant (Max Planck Institute of Biochemistry) and Xcalibr™ software (Thermo Fisher Scientific). The same proteome database as above was used to search for peptides.

4.5. References

1. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158,** 1402–1414 (2014).

2. Lopetuso, L. R., Scaldaferri, F., Petito, V. & Gasbarrini, A. Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* **5,** 23 (2013).

3. Chu, J. *et al.* Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat. Chem. Biol.* **12,** 1004–1006 (2016).

4. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).

5. Barajas, J. F. *et al.* Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem. Biol.* **22,** 1018–1029 (2015).

6. Harris, N. C. *et al.* Biosynthesis of isonitrile lipopeptides by conserved nonribosomal peptide synthetase gene clusters in Actinobacteria. *Proc. Natl. Acad. Sci.* **114,** 201705016 (2017).

7. Konz, D. & Marahiel, M. A. How do peptide synthetases generate structural diversity? *Chem. Biol.* **6,** R39–R48 (1999).

8. Arora, P. *et al.* Mechanistic and functional insights into fatty acid activation in *Mycobacterium tuberculosis*. *Nat. Chem. Biol.* **5,** 166–173 (2009).

9. Zhang, Z. *et al.* Structural and functional studies of fatty acyl adenylate ligases from *E. coli* and *L. pneumophila*. *J. Mol. Biol.* **406,** 313–324 (2011).

10. Walsh, C. T., Gehring, A. M., Weinreb, P. H., Quadri, L. E. N. & Flugel, R. S. Post-translational modification of polyketide and nonribosomal peptide synthases. *Curr. Opin. Chem. Biol.* **1,** 309–315 (1997).

11. Khurana, P., Gokhale, R. S. & Mohanty, D. Genome scale prediction of substrate specificity for acyl adenylate superfamily of enzymes based on active site residue profiles. *BMC Bioinformatics* **11,** 57 (2010).

12. Chang, K. H., Xiang, H. & Dunaway-Mariano, D. Acyl-adenylate motif of the acyl-adenylate/thioester-forming enzyme superfamily: a site-directed mutagenesis study with

the *Pseudomonas* sp. Strain CBS3 4-chlorobenzoate:coenzyme A ligase. *Biochemistry* **36,** 15650–15659 (1997).

13. Gulick, A. M. Conformational dynamics in the acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem. Biol.* **4,** 811–827 (2009).

14. Röttig, M. *et al.* NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** 362–367 (2011).

15. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **368,** 1500–1517 (2007).

16. Blin, K. *et al.* antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41,** W204–W212 (2013).

17. Bernard, K. *et al.* Characterization of isolates of *Eisenbergiella tayi*, a strictly anaerobic Gram-stain variable bacillus recovered from human clinical materials in Canada. *Anaerobe* **44,** 128–132 (2017).

18. Moore, W. E. C., Johnson, J. L. & Holdeman, L. V. Emendation of Bacteroidaceae and Butyrivibrio and Descriptions of Desulfornonas gen. nov. and Ten New Species in the Genera Desulfomonas, Butyrivibrio, Eubacterium, Clostridium, and Ruminococcus. *Int. J. Syst. Bacteriol.* **26,** 238–252 (1976).

19. Liu, C., Finegold, S. M., Song, Y. & Lawson, P. A. Reclassification of *Clostridium coccoides*, *Ruminococcus hansenii*, *Ruminococcus hydrogenotrophicus*, *Ruminococcus luti*, *Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydrogenotrophica* comb. nov., *Blautia luti* comb. nov., *Blautia producta* comb. nov., *Blautia schinkii* comb. nov. and description of *Blautia wexlerae* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **58,** 1896–1902 (2008).

20. Hoffmann, D., Hevel, J. M., Moore, R. E. & Moore, B. S. Sequence analysis and biochemical characterization of the nostopeptolide A biosynthetic gene cluster from Nostoc sp. GSV224. *Gene* **311,** 171–180 (2003).

21. Edwards, D. J. *et al.* Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chem. Biol.* **11,** 817–833 (2004).

22. Demirev, A. V., Lee, C. H., Jaishy, B. P., Nam, D. H. & Ryu, D. D. Y. Substrate

specificity of nonribosomal peptide synthetase modules responsible for the biosynthesis of the oligopeptide moiety of cephabacin in *Lysobacter lactamgenus*. *FEMS Microbiol. Lett.* **255,** 121–128 (2006).

23. Degen, A., Mayerthaler, F., Mootz, H. D. & Di Ventura, B. Context-dependent activity of A domains in the tyrocidine synthetase. *Sci. Rep.* **9,** 5119 (2019).

24. Duitman, E. H. *et al.* The mycosubtilin synthetase of *Bacillus subtilis* ATCC6633: a multifunctional hybrid between a peptide synthetase, an amino transferase, and a fatty acid synthase. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 13294–13299 (1999).

25. Bachmann, B. O. & Ravel, J. Chapter 8: Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.* **458,** 181–217 (2009).

26. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).

27. Kabeerdoss, J., Sankaran, V., Pugazhendhi, S. & Ramakrishna, B. S. *Clostridium leptum* group bacteria abundance and diversity in the fecal microbiota of patients with inflammatory bowel disease: a case-control study in India. *BMC Gastroenterol.* **13,** 20 (2013).

28. Amir, I., Bouvet, P., Legeay, C., Gophna, U. & Weinberger, A. *Eisenbergiella tayi* gen. nov., sp. nov., isolated from human blood. *Int. J. Syst. Evol. Microbiol.* **64,** 907–914 (2014).

29. Atarashi, K. *et al.* T_{reg} induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* **500,** 232–236 (2013).

30. Becker, N., Kunath, J., Loh, G. & Blaut, M. Human intestinal microbiota: Characterization of a simplified and stable gnotobiotic rat model. *Gut Microbes* **2,** 25–33 (2011).

31. Touyama, M., Jin, J. S., Kibe, R., Hayashi, H. & Benno, Y. Quantification of *Blautia wexlerae* and *Blautia luti* in human faeces by real-time PCR using specific primers. *Benef. Microbes* **6,** 583–590 (2015).

32. American Chemical Society; Chemical Abstracts Service. SciFinder – Explore. (2019). Available at: https://scifinder.cas.org/scifinder/view/scifinder/scifinderExplore.jsf. (Accessed: 7th March 2019)

33. Schneider, B. A. & Balskus, E. P. Discovery of small molecule protease inhibitors by

investigating a widespread human gut bacterial biosynthetic pathway. *Tetrahedron* **74,** 3215–3230 (2018).

34. Ganneau, C., Moulin, A., Demange, L., Martinez, J. & Fehrentz, J. A. The epimerization of peptide aldehydes – A systematic study. *J. Pept. Sci.* **12,** 497–501 (2006).

35. Billson, J. *et al.* The design and synthesis of inhibitors of the cysteinyl protease, *Der p* I. *Bioorganic Med. Chem. Lett.* **8,** 993–998 (1998).

36. Okada, Y., Taguchi, H. & Yokoi, T. Total synthesis of optically active deoxyaspergillic acid from dipeptidyl aldehyde. *Tetrahedron Lett.* **37,** 2249–2252 (1996).

37. Bo, J., Wang, L., Li, W., Zhang, X. & Zhang, A. Comb-like polymers pendanted with elastin-like peptides showing sharp and tunable thermoresponsiveness through dynamic covalent chemistry. *J. Polym. Sci. Part A Polym. Chem.* **54,** 3379–3387 (2016).

38. Thompson, R. C. [19] Peptide aldehydes: Potent inhibitors of serine and cysteine proteases. *Methods Enzymol.* **46,** 220–225 (1977).

39. Overall, C. M. & Blobel, C. P. In search of partners: linking extracellular proteases to substrates. *Nat. Rev. Mol. Cell Biol.* **8,** 245–257 (2007).

40. Cuerrier, D., Moldoveanu, T. & Davies, P. L. Determination of peptide substrate specificity for μ-calpain by a peptide library-based approach. *J. Biol. Chem.* **280,** 40632–40641 (2005).

41. Biniossek, M. L., Nägler, D. K., Becker-Pauly, C. & Schilling, O. Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *J. Proteome Res.* **10,** 5363–5373 (2011).

42. Choe, Y. *et al.* Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* **281,** 12824–12832 (2006).

43. Timmer, J. C. & Salvesen, G. S. Caspase substrates. *Cell Death Differ.* **14,** 66–72 (2007).

44. Thornberry, N. A. *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. *J. Biol. Chem.* **272,** 17907–17911 (1997).

45. Thorpe, M. *et al.* Extended cleavage specificity of human neutrophil cathepsin G: A low activity protease with dual chymase and tryptase-type specificities. *PLoS One* **13,** e0195077 (2018).

46. Perera, N. C. *et al.* Global substrate profiling of proteases in human neutrophil extracellular traps reveals consensus motif predominantly contributed by elastase. *PLoS One* **8,** e75141 (2013).

47. Olsen, J. V., Ong, S.-E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614 (2004).

48. Sasaki, T. *et al.* Inhibitory effect of di- and tripeptidyl aldehydes on calpains and cathepsins. *J. Enzyme Inhib.* **3,** 195–201 (1990).

49. Rano, T. A. *et al.* A combinatorial approach for determining protease specificities: Application to interleukin-1β converting enzyme (ICE). *Chem. Biol.* **4,** 149–155 (1997).

50. Ferrucci, A. *et al.* Ac-*t*Leu-Asp-H is the minimal and highly effective human caspase-3 inhibitor: biological and in silico studies. *Amino Acids* **47,** 153–162 (2014).

51. Watt, W. *et al.* The atomic-resolution structure of human caspase-8, a key activator of apoptosis. *Structure* **7,** 1135–1143 (1999).

52. Stein, R. L. & Strimpler, A. M. Slow-binding inhibition of chymotrypsin and cathepsin G by the peptide aldehyde chymostatin. *Biochemistry* **26,** 2611–2615 (1987).

53. Umezawa, H. Chapter 55: Structures and activities of protease inhibitors of microbial origin. *Methods Enzymol.* **45,** 678–695 (1976).

54. Aoyagi, T. *et al.* Leupeptins, new protease inhibitors from Actinomycetes. *J. Antibiot. (Tokyo).* **22,** 283–286 (1969).

55. Ji, J., Su, L. & Liu, Z. Critical role of calpain in inflammation. *Biomed. Reports* **5,** 647–652 (2016).

56. Cuzzocrea, S. *et al.* Calpain inhibitor I reduces colon injury caused by dinitrobenzene sulphonic acid in the rat. *Gut* **48,** 478–488 (2001).

57. Huang, Z. *et al.* Calpastatin prevents NF-κB-mediated hyperactivation of macrophages and attenuates colitis. *J. Immunol.* **191,** 3778–3788 (2013).

58. Turk, V. *et al.* Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochim. Biophys. Acta - Proteins Proteomics* **1824,** 68–88 (2012).

59. Menzel, K. *et al.* Cathepsins B, L and D in inflammatory bowel disease macrophages and potential therapeutic effects of cathepsin inhibition in vivo. *Clin. Exp. Immunol.* **146,** 169–180 (2006).

60. Kirschke, H. Cathepsin S. *Handb. Proteolytic Enzym.* **2,** 1824–1830 (2013).

61. Cattaruzza, F. *et al.* Cathepsin S is activated during colitis and causes visceral hyperalgesia by a PAR 2-dependent mechanism in mice. *Gastroenterology* **141,** 1864–1874 (2011).

62. Santamaría, I. *et al.* Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. *Cancer Res.* **58,** 1624–1630 (1998).

63. Becker, C., Watson, A. J. & Neurath, M. F. Complex roles of caspases in the pathogenesis of inflammatory bowel disease. *Gastroenterology* **144,** 283–293 (2013).

64. Galluzzi, L., López-Soto, A., Kumar, S. & Kroemer, G. Caspases connect cell-death signaling to organismal homeostasis. *Immunity* **44,** 221–231 (2016).

65. McAlindon, M. E., Hawkey, C. J. & Mahida, Y. R. Expression of interleukin 1β and interleukin 1β converting enzyme by intestinal macrophages in health and inflammatory bowel disease. *Gut* **42,** 214–219 (1998).

66. Lassen, K. G. *et al.* Atg16L1 T300A variant decreases selective autophagy resulting in altered cytokine signaling and decreased antibacterial defense. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 7741–7746 (2014).

67. Gringhuis, S. I. *et al.* Dectin-1 is an extracellular pathogen sensor for the induction and processing of IL-1β via a noncanonical caspase-8 inflammasome. *Nat. Immunol.* **13,** 246–254 (2012).

68. Meyer-Hoffert, U. & Wiedow, O. Neutrophil serine proteases: Mediators of innate immune responses. *Curr. Opin. Hematol.* **18,** 19–24 (2011).

69. Pham, C. T. N. Neutrophil serine proteases: Specific regulators of inflammation. *Nat. Rev. Immunol.* **6,** 541–550 (2006).

70. Chen, J. M., Radisky, E. S. & Férec, C. Human trypsins. *Handb. Proteolytic Enzym.* **3,** 2600–2609 (2013).

71. Rolland-Fourcade, C. *et al.* Epithelial expression and function of trypsin-3 in irritable bowel syndrome. *Gut* **66,** 1767–1778 (2017).

72. Soreide, K., Janssen, E. A., Körner, H. & Baak, J. P. A. Trypsin in colorectal cancer: Molecular biological mechanisms of proliferation, invasion, and metastasis. *J. Pathol.* **209,** 147–156 (2006).

73. Koshikawa, N. *et al.* Expression of trypsin by epithelial cells of various tissues, leukocytes, and neurons in human and mouse. *Am. J. Pathol.* **153,** 937–944 (1998).

74. Guo, C. *et al.* Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell* **168,** 517–526 (2017).

75. Otto, H.-H. & Schirmeister, T. Cysteine proteases and their inhibitors. *Chem. Rev.* **97,** 133–172 (1997).

76. Vergnolle, N. Protease inhibition as new therapeutic strategy for GI diseases. *Gut* **65,** 1215–1224 (2016).

77. Fitzpatrick, L. R. Evidence that the ubiquitin proteasome system plays a prominent role in inflammatory bowel disease: Possible pharmacological approaches. *Pharm. Pharmacol. Int. J.* **4,** 308–309 (2016).

78. Ferrington, D. A. & Gregerson, D. S. Immunoproteasomes: structure, function, and antigen presentation. *Prog. Mol. Biol. Transl. Sci.* **109,** 75–112 (2012).

79. Cromm, P. M. & Crews, C. M. The proteasome in modern drug discovery: Second life of a highly valuable drug target. *ACS Cent. Sci.* **3,** 830–838 (2017).

80. Rajilić-Stojanović, M. & de Vos, W. M. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* **38,** 996–1047 (2014).

81. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37,** 179–185 (2019).

82. Rice, B. L. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176,** 649–662 (2019).

83. Steck, N., Mueller, K., Schemann, M. & Haller, D. Bacterial proteases in IBD and IBS. *Gut* **61,** 1610–1618 (2012).

84. Carroll, I. M. & Maharshak, N. Enteric bacterial proteases in inflammatory bowel disease: pathophysiology and clinical implications. *World J. Gastroenterol.* **19,** 7531–7543 (2013).

85. Raju, R. M., Goldberg, A. L. & Rubin, E. J. Bacterial proteolytic complexes as therapeutic targets. *Nat. Rev. Drug Discov.* **11,** 777–789 (2012).

86. Culp, E. & Wright, G. D. Bacterial proteases, untapped antimicrobial drug targets. *J. Antibiot. (Tokyo).* **70,** 366–377 (2017).

87. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48,** 77–84 (2003).

88. Schmitt, E. K. *et al.* The natural product cyclomarin kills *Mycobacterium tuberculosis* by targeting the ClpC1 subunit of the caseinolytic protease. *Angew. Chemie - Int. Ed.* **50,** 5889–5891 (2011).

89. Gao, W. *et al.* The cyclic peptide ecumicin targeting ClpC1 is active against *Mycobacterium tuberculosis* in vivo. *Antimicrob. Agents Chemother.* **59,** 880–889 (2015).

90. Brötz-Oesterhelt, H. *et al.* Dysregulation of bacterial proteolytic machinery by a new class of antibiotics. *Nat. Med.* **11,** 1082–1087 (2005).

91. Hoffmann, A. *et al.* The antibiotic ADEP reprogrammes ClpP, switching it from a regulated to an uncontrolled protease. *EMBO Mol. Med.* **1,** 37–49 (2009).

92. Arulpandi, I. & Sangeetha, R. Antibacterial activity of fistulin: A arotease inhibitor purified from the leaves of *Cassia fistula*. *ISRN Pharm.* **2012,** 584073 (2012).

93. Vila-Farres, X. *et al.* Antimicrobials inspired by nonribosomal peptide synthetase gene clusters. *J. Am. Chem. Soc.* **139,** 1404–1407 (2017).

94. Santajit, S. & Indrawattana, N. Mechanisms of antimicrobial resistance in ESKAPE pathogens. *Biomed Res. Int.* **2016,** 2475067 (2016).

95. Hockett, K. L. & Baltrus, D. A. Use of the soft-agar overlay technique to screen for bacterially produced inhibitory compounds. *J. Vis. Exp.* e55064 (2017). doi:10.3791/55064

96. Wiegand, I., Hilpert, K. & Hancock, R. E. W. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3,** 163–175 (2008).

97. Andrews, J. M. Determination of minimum inhibitory concentrations. *J. Antimicrob. Chemother.* **48,** 5–16 (2001).

98. Dabard, J. *et al.* Ruminococcin A, a new lantibiotic produced by a *Ruminococcus gnavus* strain isolated from human feces. *Appl. Environ. Microbiol.* **67,** 4111–4118 (2001).

99. Garsin, D. A. *et al.* "Pathogenesis and Models of Enterococcal Infection" in *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]* (eds. Gilmore M. S., Clewell D. B., Ike Y., et al.) (Massachusetts Ear and Eye Infirmary, 2014). Available from: https://www.ncbi.nlm.nih.gov/books/NBK190426/.

100. Rodarte, M. P., Dias, D. R., Vilela, D. M. & Schwan, R. F. Proteolytic activities of bacteria, yeasts and filamentous fungi isolated from coffee fruit (*Coffea arabica* L.). *Acta Sci. Agron.* **33,** 457–464 (2011).

101. Allison, C. & Macfarlane, G. T. Regulation of protease production in *Clostridium sporogenes*. *Appl. Environ. Microbiol.* **56,** 3485–3490 (1990).

102. Prakash, M., Banik, R. M. & Koch-Brandt, C. Purification and characterization of *Bacillus cereus* protease suitable for detergent industry. *Appl. Biochem. Biotechnol.* **127,** 143–155 (2005).

103. Twining, S. S. Fluorescein isothiocyanate-labeled casein assay for proteolytic enzymes. *Anal. Biochem.* **143,** 30–34 (1984).

104. Nord, C.-E., Wadström, T., Dornbusch, K. & Wretlind, B. Extracellular proteins in five clostridial species from human infections. *Med. Microbiol. Immunol.* **161,** 145–154 (1975).

105. Allison, C. & Macfarlane, G. T. Regulation of protease production in *Clostridium sporogenes*. *Appl. Environ. Microbiol.* **56,** 3485–3490 (1990).

106. Allison, C. & Macfarlane, G. T. Physiological and nutritional determinants of protease secretion by *Clostridium sporogenes*: characterization of six extracellular proteases. *Appl. Microbiol. Biotechnol.* **37,** 152–156 (1992).

107. Vos, P. De & Bergey, D. H. "*Clostridium sporogenes*" in *Bergey's Manual of Systematic Bacteriology: Volume 3: The Firmicutes* (ed. Jones, D.) 817 (Springer, 2009).

108. Ouertani, A. *et al.* Two new secreted proteases generate a casein-derived antimicrobial peptide in *Bacillus cereus* food born isolate leading to bacterial competition in milk. *Front. Microbiol.* **9,** 1148 (2018).

109. Rose, M. *et al.* Sequence analysis of three *Bacillus cereus* loci carrying PlcR-regulated genes encoding degradative enzymes and enterotoxin. *Microbiology* **145,** 3129–3138 (1999).

110. Lozano, G. L. *et al.* Draft genome sequence of biocontrol agent *Bacillus cereus* UW85. *Genome Announc.* **4,** e00910–16 (2016).

111. Vandooren, J., Geurts, N., Martens, E., Van den Steen, P. E. & Opdenakker, G. Zymography methods for visualizing hydrolytic enzymes. *Nat. Methods* **10,** 211–220 (2013).

112. Macfarlane, G. T., Macfarlane, S. & Gibson, G. R. Synthesis and release of proteases by *Bacteroides fragilis*. *Curr. Microbiol.* **24,** 55–59 (1992).

113. McDonald, J. a. K. *et al.* Simulating distal gut mucosal and luminal communities using packed-column biofilm reactors and an in vitro chemostat model. *J. Microbiol. Methods* **108,** 36–44 (2015).

114. de Wiele, T., den Abbeele, P., Ossieur, W., Possemiers, S. & Marzorati, M. "The Simulator of the Human Intestinal Microbial Ecosystem (SHIME®)" in *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (eds. Verhoeckx, K. et al.) 305–317 (Springer International Publishing, 2015). doi:10.1007/978-3-319-16104-4_27

115. Cravatt, B. F., Wright, A. T. & Kozarich, J. W. Activity-based protein profiling: From enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.* **77,** 383–414 (2008).

116. Sanman, L. E. & Bogyo, M. Activity-based profiling of proteases. *Annu. Rev. Biochem.* **83,** 249–273 (2014).

117. Sadler, N. C. & Wright, A. T. Activity-based protein profiling of microbes. *Curr. Opin. Chem. Biol.* **24,** 139–144 (2015).

118. Ortega, C. *et al.* Systematic survey of serine hydrolase activity in *Mycobacterium tuberculosis* defines changes associated with persistence. *Cell Chem. Biol.* **23,** 290–298 (2016).

119. Lentz, C. S. *et al.* Identification of a *S. aureus* virulence factor by activity-based protein profiling (ABPP). *Nat. Chem. Biol.* **14,** 609–617 (2018).

120. Dannheim, H. *et al.* Manual curation and reannotation of the genomes of *Clostridium difficile* 630Δerm and *C. difficile* 630. *J. Med. Microbiol.* **66,** 286–293 (2017).

121. Kurinov, I. V. & Harrison, R. W. Two crystal structures of the leupeptin-trypsin complex. *Protein Sci.* **5,** 752–758 (1996).

122. Margolin, N. *et al.* Substrate and inhibitor specificity of interleukin-1β-converting enzyme and related caspases. *J. Biol. Chem.* **272,** 7223–7228 (1997).

123. Mayers, M. D., Moon, C., Stupp, G. S., Su, A. I. & Wolan, D. W. Quantitative metaproteomics and activity-based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. *J. Proteome Res.* **16,** 1014–1026 (2017).

124. Chambers, S. L., Ronald, R., Hanesworth, J. M., Kinder, D. H. & Harding, J. W. Solid-

phase synthesis of hydroxyethylamine angiotensin analogues. *Peptides* **18,** 505–512 (1997).

125. Cesar, J. & Sollner Dolenc, M. Trimethylsilyldiazomethane in the preparation of diazoketones via mixed anhydride and coupling reagent methods: A new approach to the Arndt-Eistert synthesis. *Tetrahedron Lett.* **42,** 7099–7102 (2001).

126. Sun, A., Shoji, M., Lu, Y. J., Liotta, D. C. & Snyder, J. P. Synthesis of EF24-tripeptide chloromethyl ketone: a novel curcumin-related anticancer drug delivery system. *J. Med. Chem.* **49,** 3153–3158 (2006).

127. Moriello, A. S. *et al.* Development of the first potential covalent inhibitors of anandamide cellular uptake. *J. Med. Chem.* **49,** 2320–2332 (2006).

128. Macpherson, L. J. *et al.* Noxious compounds activate TRPA1 ion channels through covalent modification of cysteines. *Nature* **445,** 541–545 (2007).

129. Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat. Methods* **15,** 437–439 (2018).

130. Udeshi, N. D. *et al.* Antibodies to biotin enable large-scale detection of biotinylation sites on proteins. *Nat. Methods* **14,** 1167–1170 (2017).

131. Wust, J. & Hardegger, U. Transferable resistance to clindamycin, erythromycin, and tetracycline in *Clostridium difficile*. *Antimicrob. Agents Chemother.* **23,** 784–786 (1983).

132. Hussain, H. A., Roberts, A. P. & Mullany, P. Generation of an erythromycin-sensitive derivative of *Clostridium difficile* strain 630 (630Δerm) and demonstration that the conjugative transposon Tn916ΔE enters the genome of this strain at multiple sites. *J. Med. Microbiol.* **54,** 137–141 (2005).

133. Powers, J. C., Asgian, J. L., Ekici, Ö. D. & James, K. E. Irreversible inhibitors of serine, cysteine, and threonine proteases. *Chem. Rev.* **102,** 4639–4750 (2002).

134. Janoir, C., Péchiné, S., Grosdidier, C. & Collignon, A. Cwp84, a surface-associated protein of *Clostridium difficile*, is a cysteine protease with degrading activity on extracellular matrix proteins. *J. Bacteriol.* **189,** 7174–7180 (2007).

135. Pantaléon, V. *et al.* The *Clostridium difficile* protease Cwp84 Modulates both biofilm formation and cell- surface properties. *PLoS One* **10,** e0124971 (2015).

136. Kreimeyer, I. *et al.* Autoproteolytic cleavage mediates cytotoxicity of *Clostridium difficile* toxin A. *Naunyn. Schmiedebergs. Arch. Pharmacol.* **383,** 253–262 (2011).

137. Sun, H. *et al.* Focusing on probe-modified peptides: A quick and effective method for target identification. *Chem. Commun.* **52,** 10225–10228 (2016).

138. Wilson, D. J., Shi, C., Teitelbaum, A. M., Gulick, A. M. & Aldrich, C. C. Characterization of AusA: A dimodular nonribosomal peptide synthetase responsible for the production of aureusimine pyrazinones. *Biochemistry* **52,** 926–937 (2013).

139. Bajusz, S. *et al.* Highly active and selective anticoagulants: D-Phe-Pro-Arg-H, a free tripeptide aldehyde prone to spontaneous inactivation, and its stable *N*-methyl derivative, D-MePhe-Pro-Arg-H. *J. Med. Chem.* **33,** 1729–35 (1990).

140. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14,** 20–32 (2015).

141. Tropini, C., Earle, K. A., Huang, K. C. & Sonnenburg, J. L. The gut microbiome: Connecting spatial organization to function. *Cell Host Microbe* **21,** 433–442 (2017).

142. Shen, Z., Reznikoff, G., Dranoff, G. & Rock, K. L. Cloned dendritic cells can present exogenous antigens on both MHC class I and class II molecules. *J. Immunol.* **158,** 2723–2730 (1997).

143. Raiber, E.-A. *et al.* Targeted delivery of antigen processing inhibitors to antigen presenting cells via mannose receptors. *ACS Chem. Biol.* **5,** 461–476 (2010).

144. Joseph, R. C., Kim, N. M. & Sandoval, N. R. Recent developments of the synthetic biology toolkit for Clostridium. *Front. Microbiol.* **9,** 154 (2018).

145. Melander, R. J., Zurawski, D. V. & Melander, C. Narrow-spectrum antibacterial agents. *Medchemcomm* **9,** 12–21 (2018).

146. Crost, E. H. *et al.* Ruminococcin C, a new anti-*Clostridium perfringens* bacteriocin produced in the gut by the commensal bacterium *Ruminococcus gnavus* E1. *Biochimie* **93,** 1487–1494 (2011).

147. Cohen, L. J. *et al.* Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. *Proc. Natl. Acad. Sci.* **112,** E4825–E4834 (2015).

148. Cohen, L. J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549,** 48–53 (2017).

149. Faïs, T., Delmas, J., Barnich, N., Bonnet, R. & Dalmasso, G. Colibactin: More than a new bacterial toxin. *Toxins (Basel).* **10,** 16–18 (2018).

150. Wassenaar, T. M. E. coli and colorectal cancer: a complex relationship that deserves a critical mindset. *Crit. Rev. Microbiol.* **44,** 619–632 (2018).

151. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349,** 1254766 (2015).

152. Garg, N. *et al.* Natural products as mediators of disease. *Nat. Prod. Rep.* **34,** 194–219 (2017).

153. Mousa, W. K., Athar, B., Merwin, N. J. & Magarvey, N. A. Antibiotics and specialized metabolites from the human microbiota. *Nat. Prod. Rep.* **34,** 1302–1331 (2017).

154. Garber, K. Drugging the gut microbiome. *Nat. Biotechnol.* **33,** 228–231 (2015).

155. Blaszczak, L. C. *et al.* Pyridine derivatives as dipeptidyl peptidase inhibitors. WO 2007/015767 (2007).

156. Morwick, T., Hrapchak, M., DeTuri, M. & Campbell, S. A practical approach to the synthesis of 2,4-disubstituted oxazoles from amino acids. *Org. Lett.* **4,** 2665–2668 (2002).

157. Okada, Y., Taguchi, H. & Yokoi, T. Amino acids and peptides. XLVII. Facile synthesis of flavacol, deoxymuta-aspergillic acid and optically active deoxyaspergillic acid from dipeptidyl aldehydes. *Chem. Pharm. Bull. (Tokyo).* **44,** 2259–2262 (1996).

158. Jorda, R. *et al.* Synthesis and antiproteasomal activity of novel *O*-benzyl salicylamide-based inhibitors built from leucine and phenylalanine. *Eur. J. Med. Chem.* **135,** 142–158 (2017).

159. Qiao, W. & Qiao, Y. The relationship between the structure and properties of amino acid surfactants based on glycine and serine. *J. Surfactants Deterg.* **16,** 821–828 (2013).

160. Miki, K. *et al.* Amphiphilic brush-like copolymers involving hydrophobic amino acid- and oligopeptide-side chains for optical tumor imaging in vivo. *Bull. Chem. Soc. Jpn.* **85,** 1277–1286 (2012).

161. Ueda, K. *et al.* Evaluation of inhibitory actions of flavonols and related substances on lysophospholipase D activity of serum autotaxin by a convenient assay using a chromogenic substrate. *J. Agric. Food Chem.* **58,** 6053–6063 (2010).

162. Brady, S. F. & Clardy, J. Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *J. Am. Chem. Soc.* **122,** 12903–12904 (2000).

163. Pal, A., Ghosh, Y. K. & Bhattacharya, S. Molecular mechanism of physical gelation of hydrocarbons by fatty acid amides of natural amino acids. *Tetrahedron* **63,** 7334–7348 (2007).

164. Luo, X., Li, C. & Liang, Y. Self-assembled organogels formed by monoalkyl derivatives of oxamide. *Chem. Commun.* **5,** 2091–2092 (2000).

165. Yang, X. *et al.* Flavour Modifying Compounds. US 2014/0127144 (2014).

166. Das, S., Maiti, S. & Ghosh, S. Synthesis of two biofriendly anionic surfactants (N-n-decanoyl-L-valine and N-n-decanoyl-L-leucine) and their mixed micellization with nonionic surfactant Mega-10 in Tris-buffer medium at pH 9. *RSC Adv.* **4,** 12275 (2014).

167. Zhu, J. *et al.* A mild method for the cleavage of the 4-picolyloxy group with magnesium under neutral conditions. *Synlett* 142–144 (2012). doi:10.1055/s-0031-1290092

168. Jahani, F., Tajbakhsh, M., Golchoubian, H. & Khaksar, S. Guanidine hydrochloride as an organocatalyst for N-Boc protection of amino groups. *Tetrahedron Lett.* **52,** 1260–1264 (2011).

169. Patil, B. S., Vasanthakumar, G. R. & Suresh Babu, V. V. Synthesis of β-amino acids: 2-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethyl-uronium tetrafluoroborate (TBTU) for activation of Fmoc-/Boc-/Z-α-amino acids. *Synth. Commun.* **33,** 3089–3096 (2003).

170. Rasool, C. G., Nicolaidis, S. & Akhtar, M. The asymmetric distribution of enzymic activity between the six subunits of bovine liver glutamate dehydrogenase. *Biochem. J.* **157,** 675–686 (1976).

171. Betts, R., Weinsheimer, S., Blouse, G. E. & Anagli, J. Structural determinants of the calpain inhibitory activity of calpastatin peptide B27-WT. *J. Biol. Chem.* **278,** 7800–7809 (2003).

172. Hanada, K. *et al.* Isolation and characterization of E-64, a new thiol protease inhibitor. *Agric. Biol. Chem.* **42,** 523–528 (1978).

173. Leiting, B. *et al.* Inhibition of Human Caspases by Peptide-based and Macromolecular Inhibitors. *J. Biol. Chem.* **273,** 32608–32613 (2002).

174. Umezawa, H. *et al.* Chymostatin, a new chymotrypsin inhibitor produced by Actinomycetes. *J. Antibiot. (Tokyo).* **23,** 425–427 (1970).

175. Tamura, Y., Hirado, M., Okamura, K., Minato, Y. & Fujii, S. Synthetic inhibitors of trypsin, plasmin, kallikrein, thrombin, C_1r, and C_1 esterase. *Biochim. Biophys. Acta* -

Enzymol. **484,** 417–422 (1977).

176. Sievers, S. *et al.* Comprehensive redox profiling of the thiol proteome of *Clostridium difficile*. *Mol. Cell. Proteomics* **17,** 1035–1046 (2018).