8-20-2024

Informatics-Based Discovery of New Natural Products, Their Biosynthesis, and Their Biological Roles

Ethan Older
University of South Carolina

INFORMATICS-BASED DISCOVERY OF NEW NATURAL PRODUCTS, THEIR BIOSYNTHESIS, AND THEIR BIOLOGICAL ROLES

by

Ethan Andrew Older

Bachelor of Science
University of California, San Diego, 2017

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Chemistry

College of Arts and Sciences

University of South Carolina

2024

Accepted by:

Jie Li, Major Professor

Thomas Makris, Committee Member

Susan Richardson, Committee Member

Guoshuai Cai, Committee Member

Ann Vail, Dean of the Graduate School

DEDICATION

To my family, who have made this all possible, especially my grandfathers, may you both rest in peace. To my partner, without whom this would not have been worthwhile. And to all those who have supported and encouraged me along the way.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Jie Li who met me as a young and curious undergraduate back in San Diego and took me under his guidance, shaping me to be the scientist I have become. I will continuously be grateful for his knowledge and encouragement over the years and in the years to come. I would also like to thank the members of the Li Lab, especially Dr. Dan Xue who has taught me so much over the years and helped me to become a better researcher. And thanks to the many undergraduate assistants that have helped me learn how to teach and inspire, including Michael Madden, Dorathea Lee, Emily Quinn, and Andrew Campbell.

I would like to thank the Department of Chemistry and Biochemistry, especially all the Biochemistry division professors who have answered my questions and shared their resources. I also thank my committee members, Dr. Tom Makris, Dr. Susan Richardson, and Dr. Guoshuai Cai, for their advice and guidance throughout my graduate school tenure.

Finally, I would like to thank my family, especially my grandparents and my parents Chris and Kate Older for their love, support, and encouragement. Thank you to my friends, old and new, that have motivated me to persevere, and importantly, thank you to Ashley Hamada, whose unconditional love and support has made finishing all this possible.

ABSTRACT

In this work, genome mining and biosynthetic knowledge are applied for the discovery of new natural products, their biosynthetic pathways, and their promising biological activities and roles through the development of two unique new informatics-based approaches.

In chapter one, an overview of microbial natural product biological functions and biosynthesis is provided. The use of bioinformatics leveraging microbial natural product biosynthetic knowledge is also introduced. This chapter highlights the importance and need for developing innovative approaches to accessing microbial natural products, a vastly untapped source of new molecules with promising biological activities.

In chapter two, a correlational network linking the gene-encoded precursor peptides of lanthipeptide natural products, sometimes referred to as lantibiotics due to their antimicrobial properties, to the hidden unclustered proteases required for their maturation is constructed. Interrogation of the network results in the discovery of new lanthipeptide-associated proteases. Particularly, a family of M16B metallopeptidases with previously unclear function is established as class III lanthipeptide proteases. Heterologous expression of lanthipeptide biosynthetic gene clusters alongside these proteases results in the discovery and characterization of several new class III lanthipeptides including unique new *N*,*N*-dimethylated lanthipeptides with antimicrobial activity.

In chapter three, the specific biosynthetic enzymes involved in the biosynthesis of human gut microbial sulfonolipids, microbial mimics of human endogenous sphingolipids,

are used in a large-scale genome mining effort across all available human gut microbial reference genomes to create clustered subfamilies of sulfonolipid biosynthetic enzyme homologs. These subfamilies are then used to construct a biosynthetic enzyme-disease correlation, linking the decreased biosynthesis of sulfonolipids to increased inflammatory parameters in human inflammatory bowel disease patient samples. This bioinformatic connection is then experimentally validated in a mouse model of colitis and foundational work is conducted demonstrating that naturally occurring human gut microbial sulfonolipids suppress toll-like receptor 4 (TLR4)-related inflammation through direct competition with lipopolysaccharide binding to the TLR4/myeloid differentiation factor-2 (MD-2) complex.

Taken together, these two examples demonstrate the application of biosynthetic knowledge in the development of informatics-based approaches for discovery of natural products with promising biological activities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D, 2D..... One- or two-dimensional

3OC6-HSL..... *N*-3-oxohexanoyl-L-homoserine lactone

3-oxoLCA..... 3-oxolithocholic acid

AA..... Amino acid

ACP..... Acyl carrier protein

AHL..... Acyl-homoserine lactone

Ala-Gly..... Dipeptide alanine-glycine

ANOVA..... Analysis of variance

B. fragilis..... *Bacteroides fragilis*

B. nakamurai..... *Bacillus nakamurai*

B. subtilis..... *Bacillus subtilis*

BGC..... Biosynthetic gene cluster

C. gleum..... *Chryseobacterium gleum*

CD..... Crohn's disease

CD(x)..... Cluster of differentiation (x)

CFAT..... Cysteate fatty acyl transferase

CID..... Collision induced dissociation

COSY..... Correlation spectroscopy NMR experiment

CXCL..... C-X-C motif chemokine ligand

CYS..... Cysteate synthase

D-FDAA..... 1-fluoro-2,4-dinitrophenyl-5-D-alanine amide

DH..... Dehydratase

Dhb..... Dehydrobutyrine

DMSO-d_6..... Dimethyl sulfoxide-d_6

DNA..... Deoxyribonucleic acid

DTT..... Dithiothreitol

E. coli..... *Escherichia coli*

EIC..... Extracted ion chromatogram

ER..... Enoyl reductase

ESI..... Electrospray ionization

FBMN..... Feature based molecular networking

FDA..... Food and Drug Administration

g..... G force, Relative centrifugal force

GNPS..... Global Natural Products Social Molecular Networking

H_2O..... Water

HCl..... Hydrochloric acid

HILIC..... Hydrophophic interaction liquid chromatography

HIV..... Human immunodeficiency virus

HMBC..... Heteronuclear Multiple Bond Correlation NMR experiment

HPLC..... High-performance liquid chromatography

HPLC-MS..... HPLC mass spectrometry

HRMS..... HPLC-high-resolution mass spectrometry

HRMS/MS..... HPLC-high-resolution tandem mass spectrometry

HRP..... Horseradish peroxidase

HSQC..... Heteronuclear Single Quantum Coherence NMR experiment

IBD..... Inflammatory bowel disease

IBDMDB..... Inflammatory bowel disease multiomics database

IFN-γ..... Interferon-γ

IL..... Interleukin

Il10$^{-/-}$..... *Il10* double knockout, *Il10*-deficient

Ile..... Isoleucine

ISF..... In source fragment

IsoalloLCA..... Isoallolithocholic acid

kb..... Kilobases

KR..... Ketoreductase

KS..... Ketosynthase

L-FDAA..... 1-fluoro-2,4-dinitrophenyl-5-L-alanine amide

LAL..... Limulus amoebocyte lysate

LAN..... Lanthionine

LAB..... Labionin

LC..... Liquid chromatography

LCA..... Lithocholic acid

LPS..... Lipopolysaccharide

m/z..... Mass-to-charge ratio

MAG..... Metagenome-assembled genome

MD-2..... Myeloid differentiation factor-2

MDR..... Multi-drug resistant

MeOD..... Deuterated methanol, Methanol-d4

MeOH..... Methanol

MeCN..... Acetonitrile

min..... Minute

mm..... Millimeter

mM..... Millimolar

MS..... Mass spectrometry

MS/MS..... Tandem mass spectrometry

MW..... Molecular weight

NCBI..... National Center for Biotechnology Information

NMR..... Nuclear magnetic resonance

NOS2..... Nitric oxide synthase 2

NP..... Natural product

NRPS..... Non-ribosomal peptide synthetase

P. polymyxa..... *Paenibacillus polymyxa*

P. thiaminolyticus..... *Paenibacillus thiaminolyticus*

PAMP..... Pathogen-associated molecular pattern

PAGE..... Polyacrylamide gel electrophoresis

PBA..... Primary bile acid

PCoA..... Principal coordinate analysis

PCR..... Polymerase chain reaction

PDB..... Protein Data Bank repository

PERMANOVA..... Permutational ANOVA

pHMMs..... Profile-Hidden Markov models

PKS..... Polyketide synthase

PQD..... Pulsed Q dissociation

PRR..... Pattern recognition receptor

qRT-PCR..... Quantitative reverse transcription PCR

RefSeq..... NCBI Reference Sequence Database

RiPPs..... Ribosomally synthesized and post-translationally modified peptides

RMSD..... Root mean square deviation

ROESY..... Rotating frame Overhauser enhancement spectroscopy

ROR-γt..... Retinoid-related orphan receptor-γt

RT..... Retention time

S. aureus..... *Staphylococcus aureus*

S. aureofaciens..... *Streptomyces aureofaciens*

S. griseus..... *Streptomyces griseus*

SAG..... Single amplified genome

SBA..... Secondary bile acid

SCFA..... Short chain fatty acid

SD..... Standard deviation

SDR..... Short chain dehydrogenase/reductase

SDS..... Sodium dodecyl sulfate

SDS-PAGE..... Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SL..... Sphingolipid

SoL..... Sulfonolipid

TGR5..... Transmembrane G protein-coupled receptor 5

T_H17..... T helper cells expressing IL-17

TIC..... Total ion current

TLR..... Toll-like receptor

TNFα..... Tumor necrosis factor-α

TOCSY..... Total correlation spectroscopy NMR experiment

T_{reg}..... Regulatory T cells

µg..... Microgram

µL..... Microliter

µM..... Micromolar

UC..... Ulcerative colitis

UV-vis..... Ultraviolet-visible spectroscopy

V. fischeri..... *Vibrio fischeri*

CHAPTER 1
INTRODUCTION

1.1 DIVERSE PHYSIOLOGICAL ROLES OF MICROBIAL NATURAL PRODUCTS

Natural products can be broadly defined as molecules that are biosynthesized by a living organism. Such molecules are divided into two classes: primary and secondary metabolites. Primary metabolites are involved in life-essential processes including basic metabolic pathways and are often referred to as the building blocks of life. On the other hand, secondary metabolites are molecules that are not absolutely required for life but impart competitive advantages to the producing organism. In natural products-focused fields such as medicinal chemistry and pharmacognosy, the term natural product refers more exclusively to secondary metabolites. Thus, the term natural products, as used in this work, is also used in reference to secondary metabolites.

Natural products are produced by nearly all living organisms including plants, animals, and microorganisms. For most of human history, it has been well-known that certain plants held medicinal properties, but only relatively recently have microbial natural products gained the spotlight[1]. The shorter generation time and complexity of microbes and microbial communities has resulted in the selective evolution of highly optimized molecules with complex and unique structures that fulfill a wide variety of biological functions such as chemical defense mechanisms, nutrient scavenging systems, and chemical communication signals[2–4].

Perhaps the most fundamental role of microbial natural products is as a chemical defense mechanism, providing the producing organism with a way to directly compete for survival in its ecological niche[4]. The most prominent example of antimicrobial natural products, penicillin (Figure 1.1a), is produced by some species of the *Penicillium* mold [5]. This β-lactam class antimicrobial inhibits cell wall synthesis in gram-positive bacteria by irreversibly binding to a catalytic serine in the active site of a terminal transpeptidase, blocking peptidoglycan cross-linking[6] (Figure 1.1b). Other microbes also produce a variety of other antimicrobial natural products, including polypeptides such as bacitracin produced by *Bacillus subtilis* (Figure 1.2) which also interferes with peptidoglycan synthesis[7,8], aminoglycosides such as streptomycin produced by *Streptomyces griseus* (Figure 1.2) which inhibits protein synthesis in both gram-negative and -positive bacteria[9,10], and tetracyclines such as chlortetracycline produced by *Streptomyces aureofaciens* (Figure 1.2) which also inhibits protein synthesis in bacteria[11].

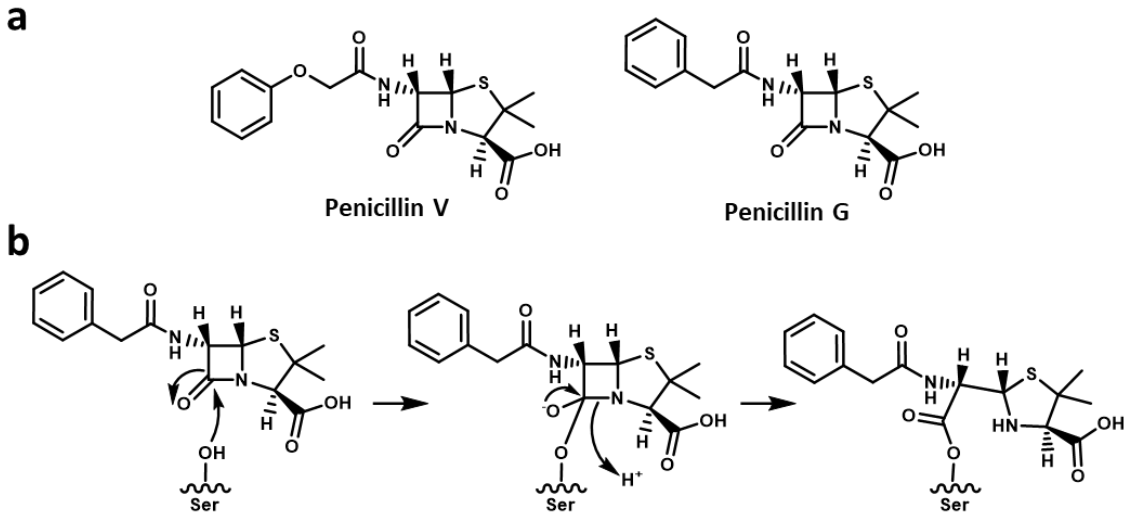


Figure 1.1 Major forms of penicillin and penicillin mechanism of inhibition. **a**, Structures of penicillin V and G, the two major forms in clinical use. Penicillin features a b-lactam ring that is susceptible to nucleophilic attack. **b**, Mechanism of penicillin G reaction with

a catalytic serine in the terminal peptidase involved in peptidoglycan synthesis, blocking enzyme function.

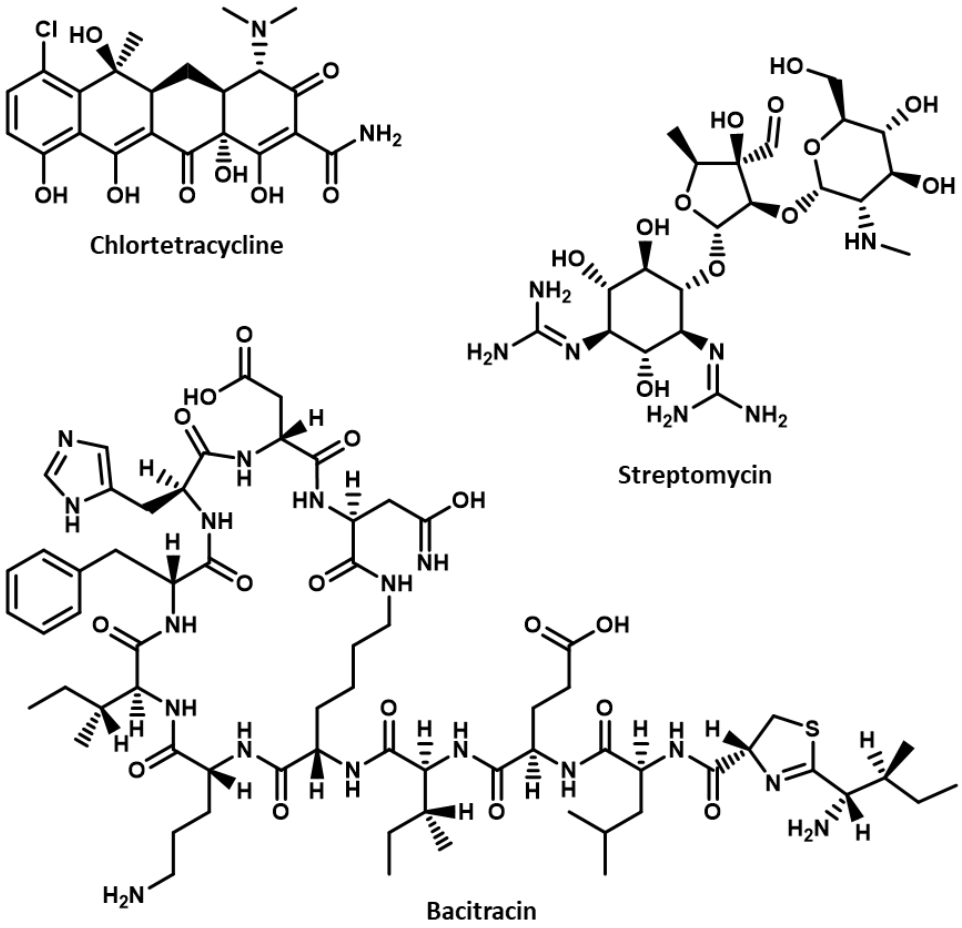


Figure 1.2 Structures of prominent antimicrobial natural products. Chlortetracyclin is a polyketide natural product produced by *S. aureofaciens*, streptomycin is an aminoglycoside natural product produced by *S. griseus*, and bacitracin is a peptidic natural product produced by *B. subtilis*

Besides such chemical defense mechanisms, microbes use natural products to scavenge for essential nutrients from their environment. Siderophores such as enterobactin, yersiniabactin, and desferrioxamine (Figure 1.3) are iron chelating natural products produced by a wide variety of bacteria including *Enterobacter*[12,13], *Yersinia*[14,15], and *Streptomyces*[16,17], respectively. These low molecular weight compounds bind iron with such high affinity that they can strip iron from iron-binding proteins and water-soluble iron

complexes[18]. Producing microbes have also developed specialized membrane proteins for recovering iron-bound siderophores from outside the cell to release iron inside the cell[18,19]. Other microbial nutrient scavenging natural products include staphylopine (Figure 1.3) which binds multiple metals including zinc, iron, cobalt, and copper and is produced by *Staphylococcus aureus*[20] and its analog pseudopaline which chelates zinc, nickel, and cobalt and is produced by *Pseudomonas aeruginosa*[21].

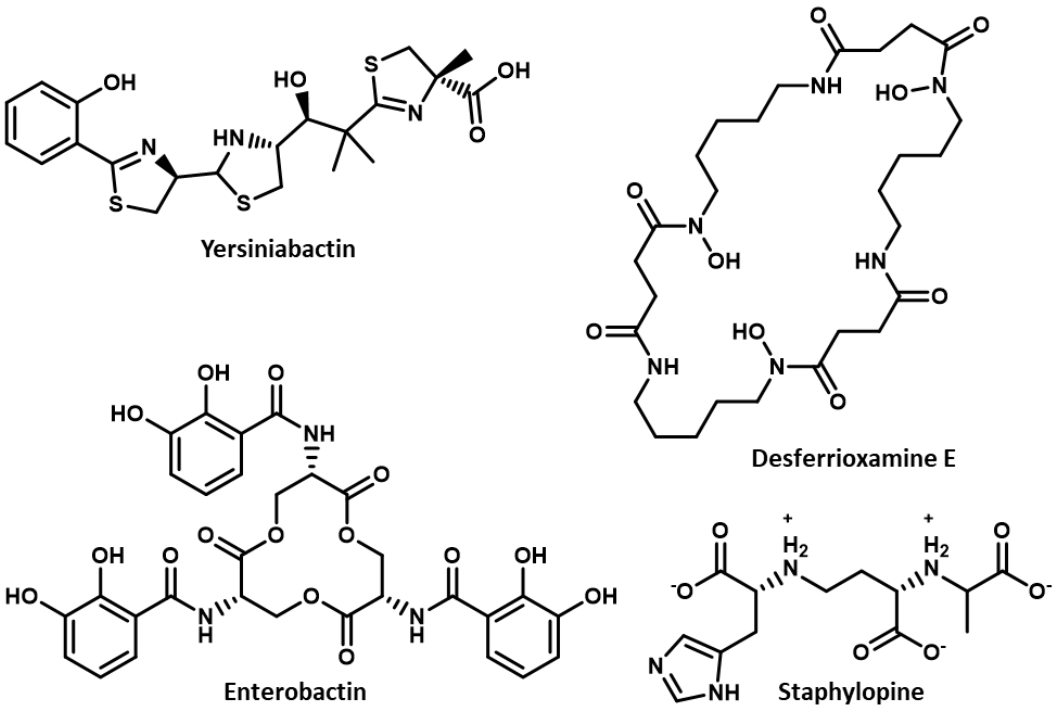


Figure 1.3 Structures of microbial nutrient scavenging natural products. Yersiniabactin is an iron chelator produced by species of *Yersinia*. Desferrioxamine is an iron chelator produced by several species of *Streptomyces*. Enterobactin is an iron chelator produced by primarily gram-negative microbes, primarily species of *Enterobacter*. Staphylopine is a broad spectrum metallophore produced by *S. aureus*.

Microbes also utilize natural products to communicate with neighboring cells such as through quorum sensing. Quorum sensing is a relatively newly studied aspect of microbial behavior and describes the ability of groups of microbial cells to coordinate their

activities through the production, accumulation, and detection of extracellular chemical signals[22]. A classic example of quorum sensing is the bioluminescence of *Vibrio fischeri* through the *lux* operon. In this unique interaction an acyl homoserine lactone (AHL), specifically *N*-3-oxohexanoyl-L-homoserine lactone (3OC6-HSL), produced and secreted by *V. fischeri*, binds to the transcriptional regulator protein encoded by *luxR*. This binding then turns on transcription of the lux operon, resulting in bioluminescence through the production of luciferase as well as transcription of *luxI*, an auto-inducer synthase which leads to increased production and secretion of 3OC6-HSL that goes on to activate *lux* transcription in other *V. fischeri* cells[23–25] (Figure 1.4). Other examples of microbial quorum sensing molecules include a heptadecapeptide, competence stimulating peptide (CSP), produced by *Streptococcus pneumoniae* which turns on genetic transformation competence after reaching a critical concentration proportional to its cell density[26,27] and the much smaller autoinducer-3 (Figure 1.4) produced by enterohemorrhagic *Escherichia coli* which activates virulence genes through the upregulation of the locus of enterocyte effacement[28,29].

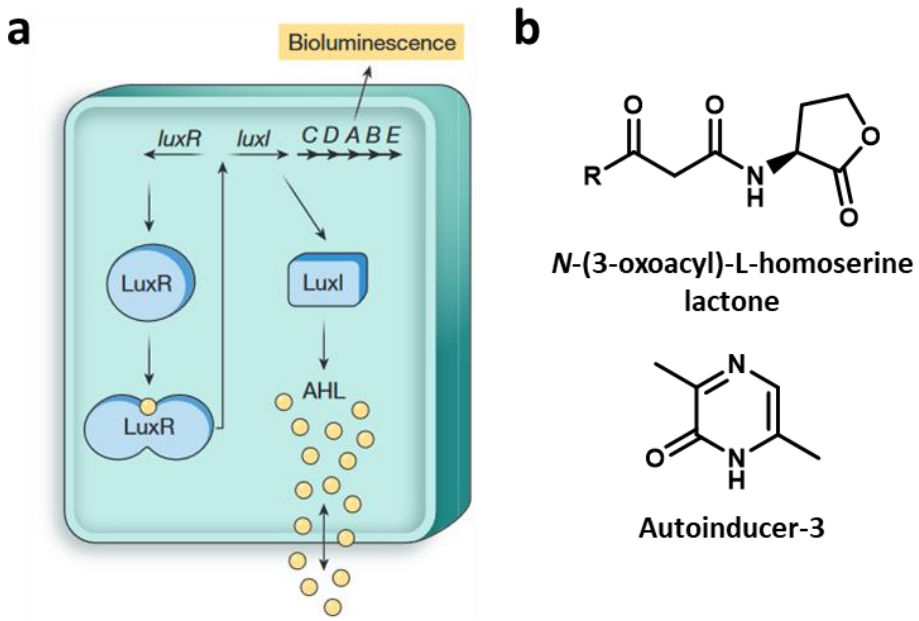


Figure 1.4 Microbial quorum sensing in the *lux* operon and other systems. **a**, Regulation of acyl-homoserine lactone (AHL) production by the *lux* operon. The lux operon is transcriptionally repressed by LuxR, encoded by *luxR*. AHLs produced by neighboring cells can be taken up and will bind LuxR, activating transcription of the *lux* operon. This results in activation of microbial luminescence as well as increased production of AHLs by LuxI. (Adapted from Whiteley, 2017). **b**, General structure of a representative AHL, *N*-(3-oxoacyl)-L-homoserine lactone. The structure of autoinducer-3, the quorum sensing signal involved in enterohemorrhagic *E. coli* virulence in the human gut, is also presented.

1.2 Molecular Drivers of Host-Microbe Interactions

Microbes residing within other organisms also produce natural products to manipulate their host environment. These so-called host-microbe interactions have gained significant interest owing to the growing understanding that complex human health conditions stemming from combinations of risk factors and genetic pre-dispositions, such as type 2 diabetes, cardiovascular disease, inflammatory bowel disease, and irritable bowel syndrome, are linked to alterations in the human gut microbiome[30,31]. Current genome and metagenome-wide association studies provide a strong foundation of correlational evidence between abundance of gut microbial species and corresponding disease states[32].

From a chemical perspective, the molecular mechanisms underlying host-microbe interactions remain largely unexplored and a recent report has identified more than 14,000 natural product-producing biosynthetic gene clusters (BGCs) from sequenced genomes of the human microbiota, many of which are uncharacterized[33,34].

A promising area of interest is in microbial mimics of host metabolites, for example, sphingolipids (SLs) are a class of bioactive lipids endogenously produced by eukaryotes and also by some microbes[35]. SLs are characterized by two long acyl chains linked to a polar headgroup (Figure 1.5). Endogenous SLs are involved in many complex eukaryotic signaling pathways with various effects ranging from regulating the cell cycle, endocytosis, and apoptosis to mediating inflammation and vesicular trafficking within the cell, but the full complexity of their role is still a very active field of research[36,37]. The phylum Bacteroidetes, which makes up 30-40% of the human gut microbiome, biosynthesize SLs that are nearly identical to endogenous SLs but have branched-end acyl chains[35,38]. These microbial SLs facilitate Bacteroidetes survival in the gastrointestinal tract by mediating cellular stress responses through interaction with host signaling pathways, and can also enter host cells and metabolism, affecting host sphingolipid levels, biosynthesis, and presumably impacting downstream host signaling pathways[34,35,39]. Like their endogenous counterparts, microbial SLs can also influence the host immune system. A glycosylated SL, α-galactosylceramide (Figure 1.5), produced by *Bacteroides fragilis* was found to bind to CD1d and activate iNKT cells in both humans and mice[40].

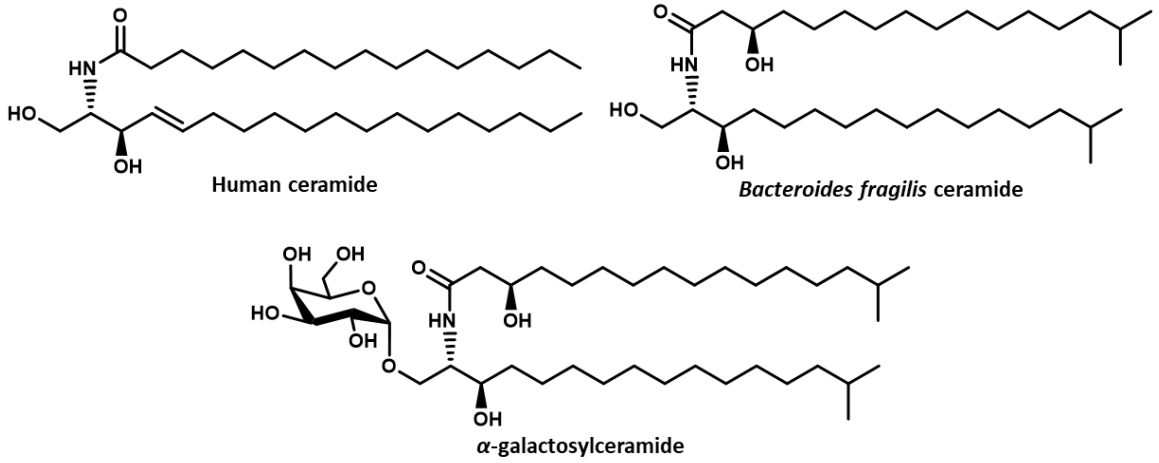


Figure 1.5 Structures of human sphingolipid and microbial mimics of host sphingolipids. Sphingolipids derived from microbial sources are characterized by their branched-end acyl chains which are not found in human sphingolipids. Both branched-chain sphingolipids presented here are produced by *B. fragilis*, a commensal member of the human gut microbiome.

Microbes in the gut can also chemically modify host metabolites to produce new biologically active natural products. Secondary bile acids (SBAs) are one such microbial biotransformation which have gained significant interest due to their various interactions with human cells[41,42]. SBAs are derived from endogenous primary bile acids (PBAs) but have undergone dehydroxylation at the 7α position (Figure 1.6)[42,43]. SBAs have been found to modulate gut-associated inflammation through multiple pathways, for example, lithocholic acid (LCA) (Figure 1.6) inhibits NOD-like receptor NLRP3 inflammasome activation through transmembrane G protein-coupled receptor-5 (TGR5) signaling[44], 3-oxoLCA (Figure 1.7) inhibits differentiation of T cells into T helper cells expressing IL-17a (T_H17) by binding to the key transcription factor retinoid-related orphan receptor-γt (RORγt)[45], and isoallolithocholic acid (isoalloLCA) (Figure 1.7) promotes regulatory T cell (T_{reg}) differentiation by increasing *Foxp3* gene expression through directly binding to the nuclear hormone receptor NR4A1[46].

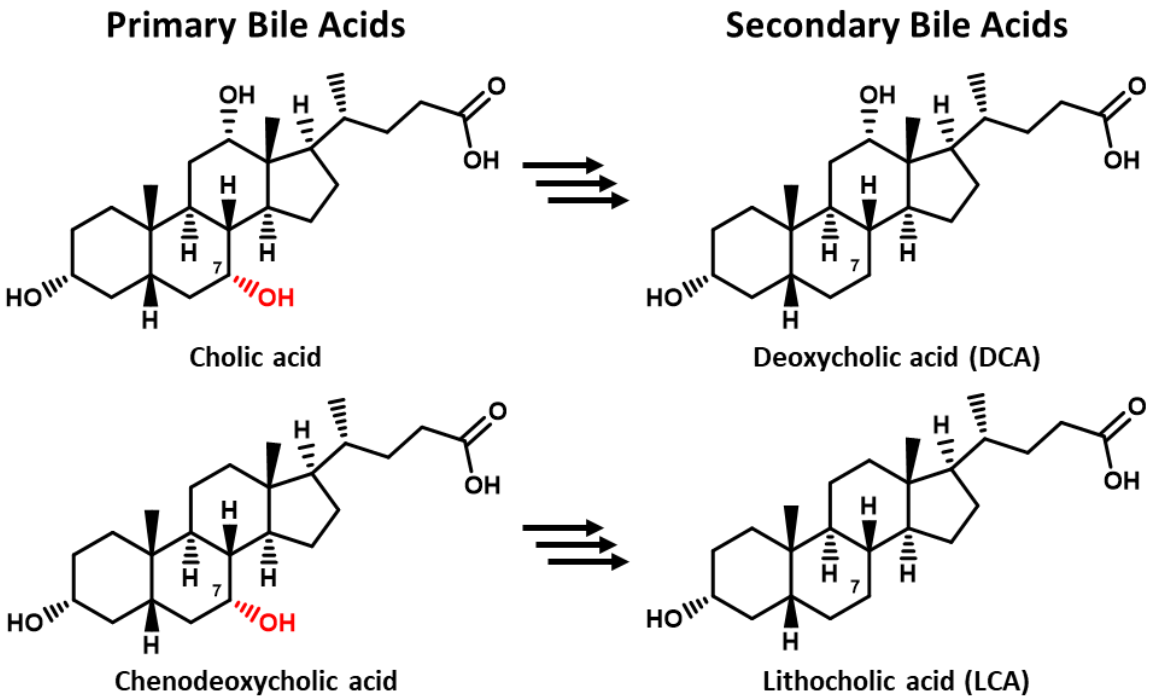


Figure 1.6 Microbial biotransformation of primary bile acids to secondary bile acids. Secreted primary bile acids (PBAs) are typically re-absorbed in the ileum and transported to the liver for recycling. PBAs that are not re-absorbed will pass into the colon where gut-residing microbes will convert them into secondary bile acids (SBAs) by dihydroxylation at the 7a position (marked and highlighted in red).

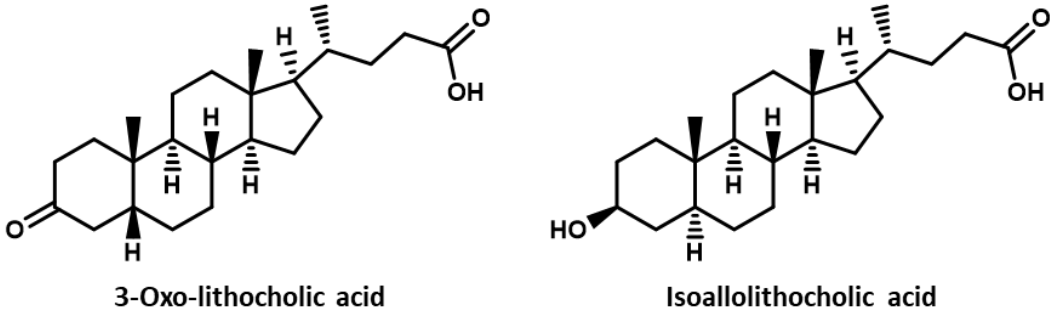


Figure 1.7 Secondary bile acids involved in regulating host inflammation. 3-Oxo-lithocholic acid (3-oxoLCA) suppresses inflammation by inhibiting T cell differentiation into T helper cells expressing the pro-inflammatory cytokine IL-17. Isoallolithocholic acid (isoalloLCA) regulates inflammation by promoting T cell differentiation into regulatory T cells with immunosuppressive effects.

1.3 REVITALIZING MICROBIAL NATURAL PRODUCTS AS DRUGS

Considering their myriad diverse functions, microbial natural products are a vast resource of chemically unique molecules with inherent biological activities. Accordingly, natural products have served as a significant source of medicines and therapies for millennia, dating back to the use of medicinal plants and herbs in traditional medicinal practices[47]. Over the last four decades, natural products have continued to contribute meaningful new drugs resulting in almost half all FDA approved small molecule drugs being either derived from natural products, inspired by natural product precursors, or natural products themselves[48].

Following the discovery of groundbreaking antimicrobial agents such as penicillin and streptomycin from microbes, the natural products research turned its focus towards microbially produced natural products. This paradigm shift resulted in the screening of countless readily accessible microbes, primarily soil actinomycetes such as *Streptomyces*, for potential new drugs[49,50]. The late 1940s to the early 1960s in particular saw the discovery of a large number of antimicrobial natural product classes and is often referred to as the "Golden Age" of natural product drug discovery (Figure 1.8). Following this period, however, the rate of microbial drug discovery has slowed significantly due to many factors but especially due to the inaccessibility of new microbial sources and high rates of compound rediscovery[47,51]. To address these challenges, natural products fields are embracing modern technological advancements as well as advancements in bioinformatics and chemoinformatics to revitalize this crucial outlet of new drugs and bring about a new "Golden Age" of natural product discovery[51,52].

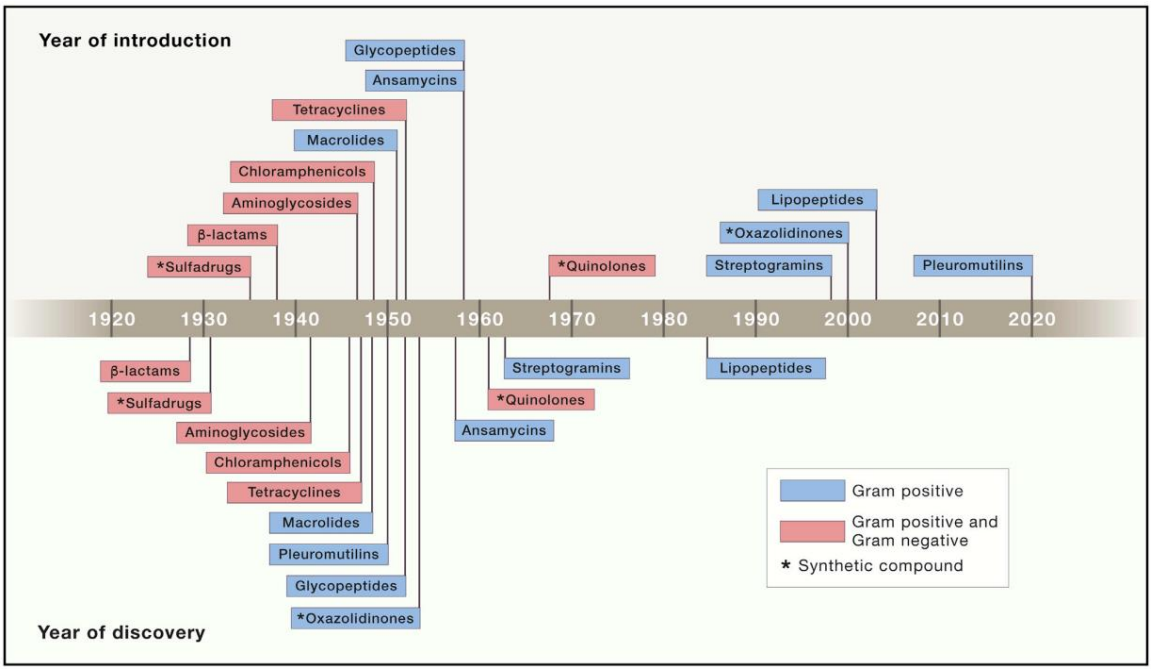


Figure 1.8 Timeline of natural product antibiotic discovery and clinical introduction. Broad-spectrum antibiotics are shown in red, narrow-spectrum antibiotics are shown in blue. *Indicates a synthetic compound. (Adapted from Lewis, 2020).

1.4 MICROBIAL NATURAL PRODUCTS BIOSYNTHESIS

Towards this new "Golden Age", significant effort has been invested into understanding the mechanisms of natural products biosynthesis. Natural products are genetically encoded molecules: the enzymes that catalyze the chemical reactions that make natural products are encoded in the genome of the producing organism. Frequently in microbes, these biosynthetic enzyme-encoding genes are found physically clustered in the genome in biosynthetic gene clusters (BGCs)[2,54]. This clustering phenomenon is believed to be derived from genetic acquisition events such as horizontal gene transfer driven by evolutionary natural selection[2,54]. Further evolution has given rise to significant BGC diversification which is represented by the wide variety of unique natural product structures that can be observed, but there are still fundamental similarities between BGCs which can be used to classify them and the natural products they produce into certain classes.

For example, polyketides are one class of natural products that is composed of repeating units of acyl-CoA precursors, commonly malonyl-CoA or methylmalonyl-CoA. Polyketide BGCs are characterized by one or multiple polyketide synthase (PKS) enzymes which are either singular multi-domain enzymes (type I) or multiple enzyme complexes consisting of individual domains (type II and III)[55]. The biosynthesis of polyketides resembles that of fatty acid biosynthesis and is often thought of as an assembly line involving two primary steps: 1) loading an acyl-CoA precursor onto an acyl-CoA carrier protein domain (ACP) and 2) condensation of the ACP-loaded unit with the growing chain by the ketosynthase domain (KS)[55,56] (Figure 1.9). These two steps are followed by up to three sequential tailoring steps including: reduction of the ketone to a hydroxyl group by a ketoreductase domain (KR), dehydration of the hydroxyl group by a dehydratase domain (DH) leaving behind a carbon-carbon double bond, and finally reduction of the double bond to a single bond by an enoyl reductase domain (ER)[55,56]. The completed chain is then cyclized and released from the ACP domain. Chemical diversity in the polyketide natural product class is generally derived from the inclusion or exclusion of the three tailoring domains, resulting in the presence of varying states of oxidation across the molecule as well as by non-PKS tailoring enzymes catalyzing various reactions including further reduction, glycosylation, methylation, prenylation, or halogenation[57].

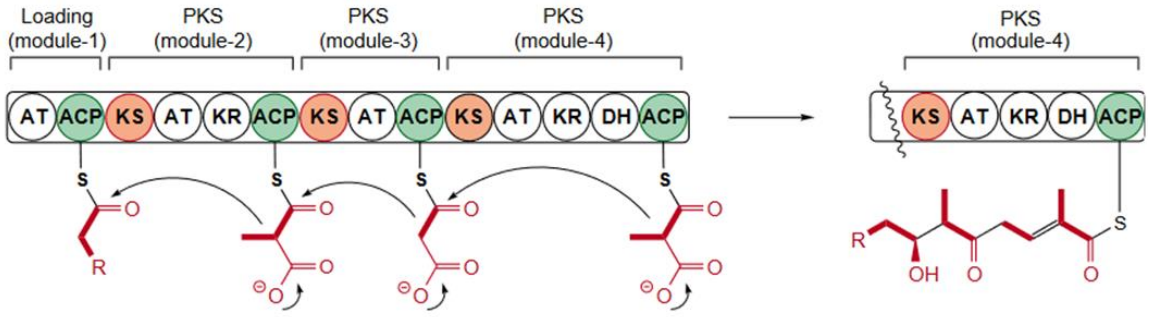


Figure 1.9 Overview of PKS biosynthetic assembly line. In this diagram of a typical type I PKS assembly line, module-1 first loads an acyl-CoA onto the ACP domain. Three more modules are then used to incorporate multiple units of methylmalonyl-CoA or malonyl-CoA with varying levels of oxidation. The final product is observed still tethered to the ACP domain of module-4. (Adapted from Shen, 2003).

1.5 Leveraging Biosynthetic Knowledge for Microbial Natural Product Discovery

Across all natural product classes, the biosynthetic enzymes encoded in BGCs are critically required for the biosynthesis of their corresponding natural products. This presents a unique opportunity to leverage the simplistic nature of DNA, being composed of combinations of only four nucleotides, to identify the unique DNA sequences representing important biosynthetic enzymes. These sequences can be used to search for other genes with similar DNA sequences encoding biosynthetic enzymes that are likely to perform similar functions in the biosynthesis of new natural products. Using this "genes-to-molecules" approach, known as genome mining, whole bacterial genomes can be searched for natural product BGCs which can then be rapidly identified on the basis of their similarity to other known BGCs, thus opening the door to finding new and unique natural products with promising biological activities[58–60]. Facilitating this approach, the introduction of rapid and cost-effective high-throughput genome sequencing has dramatically increased the availability of microbial genomic information[59,61]. This increase

in genome sequencing has led to the observation that microbial natural product biosynthetic diversity is vastly more than previously suspected[58,62,63]. For example, despite the significant number of known and characterized microbes, an estimated 99% of microbial species remain unstudied due to their inaccessibility or inability to be cultivated under laboratory conditions[62,64]. These uncultivable microbes represent a significant portion of the untapped biosynthetic potential. To access this new source for microbial natural products, researchers are turning to metagenomic sequencing and assembly which can capture the BGCs encoded in these inaccessible genomes[62,65,66]. With BGCs in hand, molecular biology approaches such as heterologous expression can then be employed to produce the BGC-encoded natural products[67]. This approach was used to discover a family of four new polyketide natural products from an uncultivable human oral microbe, named metamycins, which showed strong antimicrobial activity against gram-positive bacteria[68].

This work describes the development of two new informatics-based approaches towards the discovery of new natural products, their biosynthesis, and their biological activities and roles. In chapter two, biosynthetic knowledge of lanthipeptides is leveraged to construct a global correlational network between lanthipeptide precursors and proteases leading to the identification of a new family of class III lanthipeptide proteases as well as several new class III lanthipeptide structures with promising antimicrobial activity. In chapter three, critical biosynthetic enzymes involve in the biosynthesis of sulfonolipids by human gut microbiota are used to guide a biosynthetic enzyme -disease correlation revealing a negative correlation between sulfonolipid biosynthesis and inflammatory bowel disease status which is then supported by an in-depth investigation of sulfonolipids'

suppression of toll-like receptor 4-mediated inflammation by direct competition with the pro-inflammatory agent lipopolysaccharide.

CHAPTER 2
CORRELATIONAL NETWORKING GUIDES THE DISCOVERY OF UNCLUSTERED LANTHIPEPTIDE PROTEASE-ENCODING GENES

2.1 BACKGROUND AND INTRODUCTION

Many natural product biosynthetic gene clusters (BGCs) are found not to harbor a full set of necessary genes and must utilize enzymes encoded elsewhere in the genome for biosynthesis. These hidden enzymes are particularly important when the missing biosynthetic steps are responsible for installation of critical moieties. For example, the antibiotic gentamicin C complex and the antitumor reagent geldanamycin both rely on remote genes located apart from their BGCs for methylation steps that are essential for their biological activities[69,70]. Additionally, the antibiotic prodigiosin undergoes an oxidative cyclization performed by an enzyme encoded outside of its BGC to produce the more potent cycloprodigiosin[71] and the antibiotic activity of mature microcin C relies on a final proteolytic step by an evolutionarily conserved protease system located out of its BGC[72].

Besides individual cases, lanthipeptides as an entire class of natural products also frequently rely on unclustered biosynthetic genes for maturation. As one of the most common ribosomally synthesized and post-translationally modified peptides (RiPPs)[73,74], lanthipeptides have been shown to exhibit anti-fungal[75], anti-HIV[76], and antinociceptive activities[77], as well as broad antimicrobial activity against multi-drug-resistant (MDR) bacteria[78,79]. The biosynthesis of all five different classes of lanthipeptides involves a

crucial protease-mediated cleavage between the leader and core peptides for final maturation. However, only two types of proteases have been relatively well studied. The first is the subtilisin-like serine protease LanP, employed by class I and II lanthipeptides[80–82]. The second is the papain-like cysteine protease domain of the LanT transporter protein, involved exclusively in the biosynthesis of class II lanthipeptides[82–85]. Due to the absence of protease-encoding genes in most characterized class III and IV lanthipeptide BGCs, the maturation of these two classes is barely understood[86], with FlaP[87] and AplP[88] being only recently reported as potential proteases for class III lanthipeptides. With the recent explosion of sequenced microbial genomes, increasingly more lanthipeptide BGCs are being identified as lacking BGC-associated genes to encode proteases[86,89,90]. A missing link between these BGCs and their hidden proteases hinders the discovery, heterologous production, and bioengineering of these potentially bioactive lanthipeptides.

In this study, we hypothesize that lanthipeptide BGCs without any colocalized protease-encoding genes may rely on proteases encoded elsewhere in the genome. Here, we develop a genome mining workflow and use correlation analysis complemented by co-expression analysis to establish the first global correlation network between lanthipeptide precursor peptides and proteases from 161,954 bacterial genome sequences. This correlation network provides guidance for targeted discovery of hidden lanthipeptide proteases encoded by genes outside of the BGCs. As proof of principle, we select two representative correlations from the network for study, leading to a simultaneous discovery of previously unknown lanthipeptides and responsible hidden proteases. Particularly, a family of bacterial M16B metallopeptidases with previously unclear biological functions

is identified as being responsible for maturation of several previously unknown class III lanthipeptides.

2.2 Correlational Networking of Lanthipeptide Precursors and Proteases

2.2.1 Construction of a global correlational network for lanthipeptides

We established a global precursor-protease network for all class I-IV lanthipeptides using 161,954 bacterial genomes obtained from the NCBI RefSeq database. When we initiated our analysis, we had not noticed any reported class V lanthipeptides, thus, this new class was not included for analysis of this report. Analyzing a large number of genomes with antiSMASH 5.0[91], we identified 21,225 putative lanthipeptide BGCs widely distributed across all bacterial taxa. These BGCs harbor 29,489 highly diverse precursors (Supplementary Data 2.1). We analyzed genomic regions 10 kb up- and downstream of LanC-like proteins and observed that approximately one third of these genomic regions do not harbor any protease genes, especially the class III system (Figure 2.1). BGCs without any colocalized protease-encoding genes, may rely on proteases encoded outside of those BGCs for leader peptide removal[74,84,92]. In this scenario, we hypothesized that the specificity between precursors and corresponding proteases still exists, at least to some extent, based on two observations: (i) only certain homologs of a protease in a genome have proteolytic activity against a specific precursor[93,94], and (ii) the proteolytic activity is affected by core peptide modifications[87].

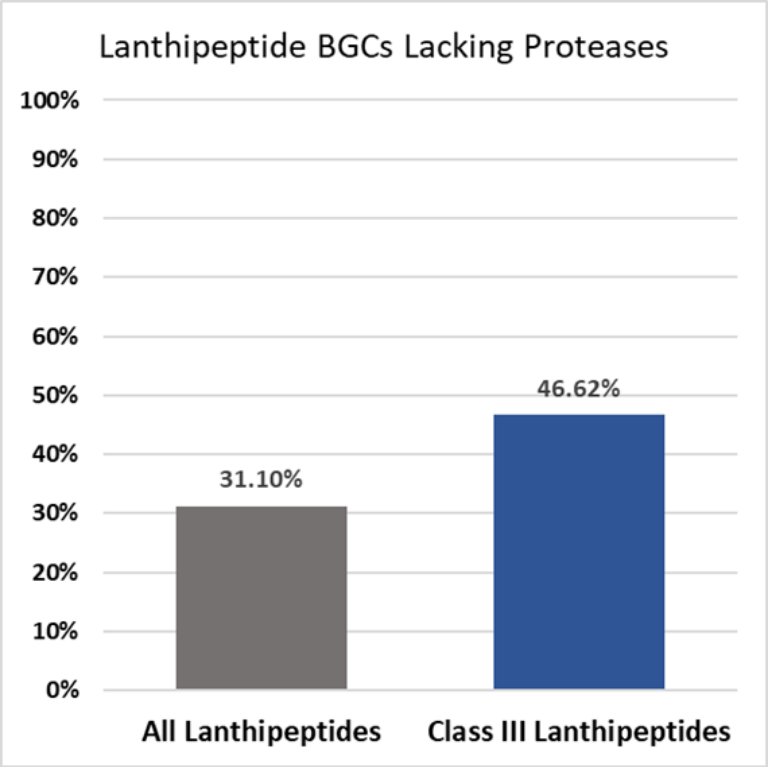


Figure 2.1 Lanthipeptide BGCs without proteases. 31.1% of lanthipeptide BGCs did not contain putative proteases while 46.6% of class III lanthipeptides BGCs did not contain putative proteases. Genomic regions 10 kb up- and downstream of LanC-like proteins were searched for putative lanthipeptide proteases. The number of proteases retrieved this way was likely inflated due to the wide genomic range searched, which possibly included non-lanthipeptide proteases. As a result, this may lead to an underestimation of the percentage of lanthipeptide BGCs that do not harbor any proteases.

Thus, based on the specificity between precursors and proteases, we performed global Spearman's rank-order correlation analysis to associate precursors with lanthipeptide proteases (regardless of being encoded inside or outside of a BGC), with an emphasis on the identification of hidden lanthipeptide proteases. Spearmen's correlation accounts for the ranking of data instead of exact values and does not require a normal distribution of the data, which matches the gene distribution in our data. Due to the large number of proteases contained in the genomes responsible for performing general

functions, it was not practical to directly correlate lanthipeptide precursors to all the proteases in the genomes. Thus, we started with pathway-specific proteases encoded by lanthipeptide BGCs. We hypothesized that pathway-specific proteases likely evolved from general proteases encoded elsewhere in the genome and, at the large scale of the dataset, pathway-specific proteases could collectively represent most functional domains of hidden proteases encoded outside of the BGCs. Indeed, by searching proteases from 21,225 putative lanthipeptide BGCs, we generated a library of 44,260 prospective lanthipeptide proteases, representing 120 unique Pfam[95] domains (Supplementary Table 2.1). In contrast, previously characterized proteases associated with lanthipeptides encompass only 6 Pfam domains (Supplementary Table 2.1). We used these 120 Pfam domains to search for proteases from the full set of 161,954 bacterial genomes, resulting in 23,777,967 putative lanthipeptide-related proteases. Grouping these proteases based on their sequence similarity using MMseqs2[96] led to 288,416 groups of proteases. Among these protease groups, 10,263 groups each containing 100 or more members were selected for downstream analysis. We also applied the same clustering approach to 29,489 lanthipeptide precursors, forming 4,527 groups. Among these precursor groups, 263 groups each containing ten or more precursors were selected for downstream analysis (Figure 2.2a). We then sought to identify links between the selected groups of proteases and precursors. To reduce the effect of phylogenetic relatedness, we performed the correlation analysis individually for each genus, leading to the identification of 5,209 significant correlations ($\rho>0.3$, pAdj<1E-5, Figure 2.2b) between precursors and proteases. These significant correlations ranged from 6 phyla and 114 genera, suggesting that widely distributed

proteases, even encoded by genes outside of the BGCs, may function against specific lanthipeptide precursors.

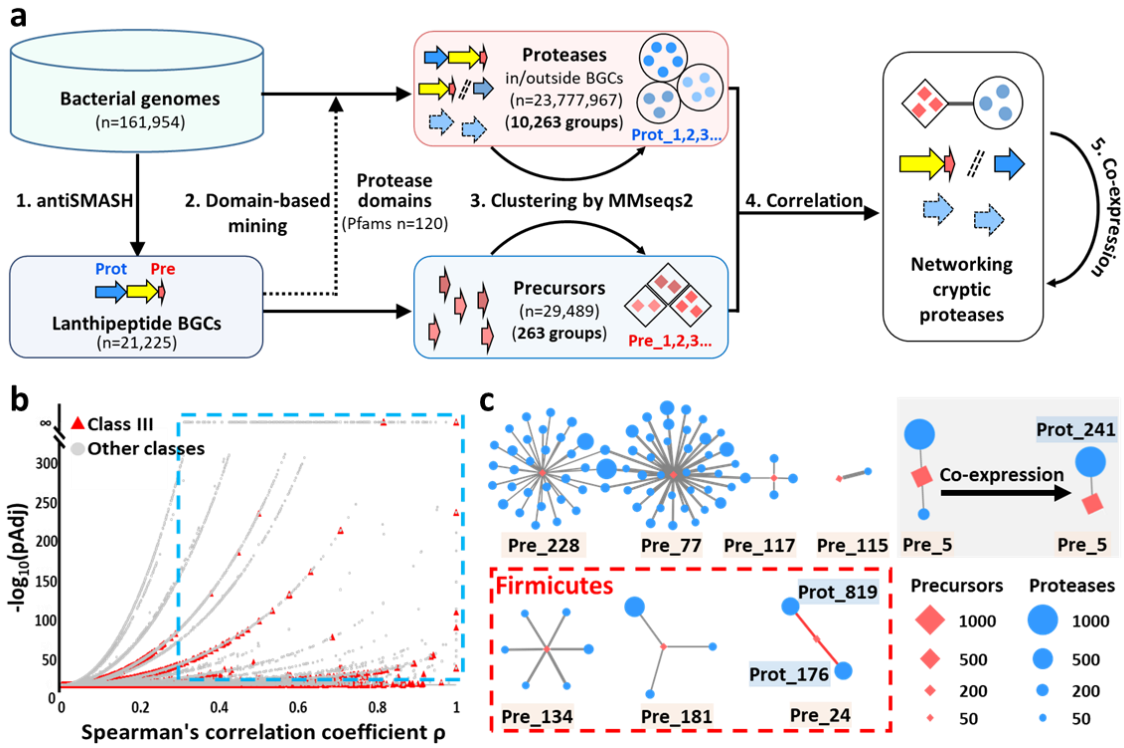


Figure 2.2 Overview of the precursor-protease correlational networking for discovery of hidden lanthipeptide proteases. **a**, A genome-mining workflow for discovering potential hidden lanthipeptide proteases, including four steps: (1). antiSMASH 5.0 was used to analyze 161,954 bacterial genomes for identification of lanthipeptide BGCs and BGC-associated proteases; (2). A Pfam domain-based mining of proteases (either inside or outside of BGCs) was performed against entire bacterial genomes based on the 120 Pfams of BGC-associated proteases identified in step (1); (3). Proteases and precursors were clustered by sequence similarity using MMseqs2; (4). Correlation networks at the genomic level were constructed to link precursors and hidden proteases that are not encoded by the BGCs; and (5). Co-expression analysis was integrated to refine the genomic-level correlation of interest. Pre: precursor peptides; Prot: proteases. **b**, Volcano plot of 81,323 correlations at the genus level. The blue box encompasses 5,209 significant correlations. Red triangles indicate correlations for class III lanthipeptides. P-values are calculated by one-sided *t*-test and adjusted by false-discovery rate. **c**, Prioritized significant correlations (ρ>0.3, pAdj<1E-5, one-sided *t*-test, adjusted by false-discovery rate) for class III lanthipeptides from eight genera. Only correlations identified in at least ten genomes (I>=10) were prioritized. Red diamonds represent groups of precursors and blue circles represent groups of proteases. Shape sizes are proportional to the number of precursors or proteases identified in a group at the genus level. Edges connecting diamonds and circles indicate a significant correlation, with the increasing thickness of the edge representing the

increasing strength of the correlation (Spearman's ρ). The 11 correlations in Firmicutes (indicated in red box) were formed between 3 groups of class III precursors (Pre_24, 134, and 181, totally 68 precursors) and 11 groups of peptidases. The remaining 80 correlations found in Actinobacteria were formed between 5 groups of class III precursors (Pre_5, 77, 115, 117, and 228, totally 796 precursors) and 76 groups of proteases. Specifically, Pre_5, the most abundant precursor group among the prioritized significant correlations, was strongly correlated with two groups of proteases, Prot_241 and Prot_1365. An integration of co-expression analysis reduced the two correlations to only one group, Prot_241.

We next focused on class III lanthipeptides due to their elusive maturation process[86]. We identified 1,833 putative precursors encoding class III lanthipeptides. These precursor peptides were classified into 217 groups based on sequence similarity, including 18 groups that each contained at least 10 members. We selected the correlations identified in at least ten genomes (I>=10) for further analysis, leading to the prioritization of 91 significant correlations (ρ>0.3, pAdj<1E-5, I>=10) between 8 groups of precursors (Figure 2.3) and 87 groups of proteases. These significant correlations were distributed in two phyla, including 80 correlations in Actinobacteria and 11 correlations in Firmicutes. The core information of these prioritized significant correlations is summarized in Figure 2.2c and Supplementary Table 2.2, with some representative discoveries described below.

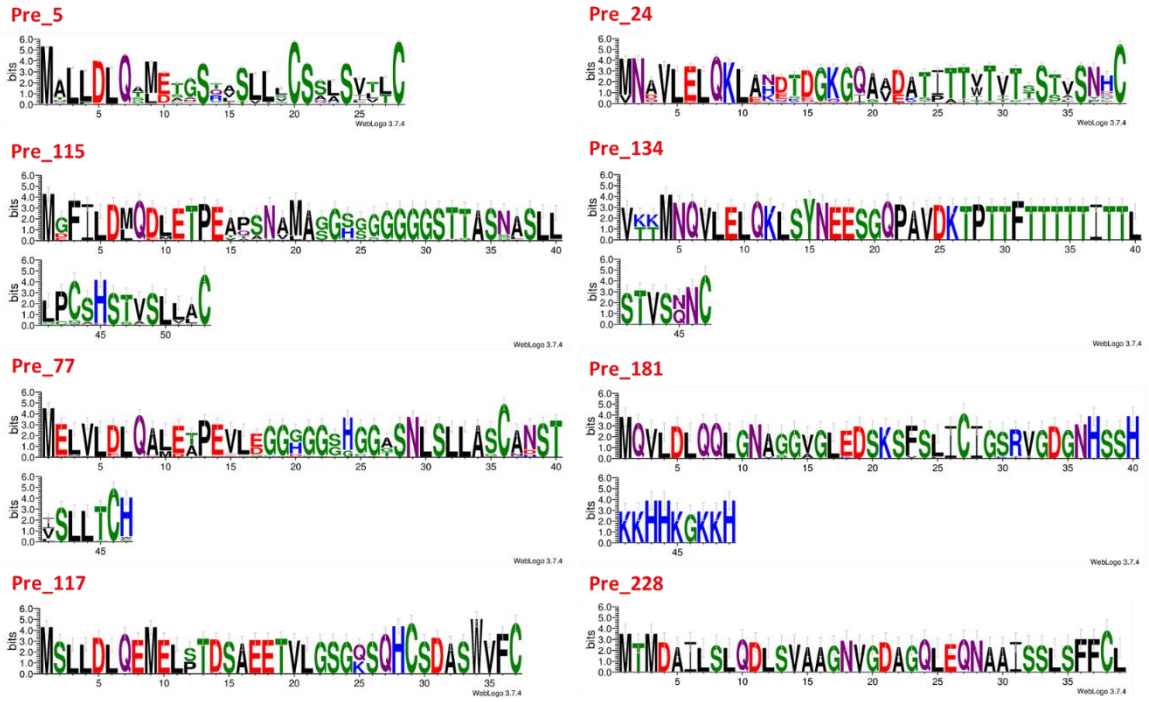


Figure 2.3 Precursor sequences of selected class III lanthipeptide precursor groups. Sequence logos representing the conserved motifs in prioritized class III lanthipeptide precursor groups. The 8 precursor groups depicted were all identified through the prioritization of significant correlations as described by Spearman's ρ, adjusted p-value, and number of genomes containing the correlation (ρ>0.3, pAdj<1E-5, one-sided *t*-test, adjusted by false-discovery rate, I>=10). These precursor groups correlate with 87 groups of proteases. Precursor sequences in each group were aligned and gaps were trimmed by trimAl before calculating logos using WebLogo v3.7.4. Error bars indicate sample correction, and the total height of the error bar is twice this correction.

Among the 91 significant correlations, metallopeptidases appeared in many significant correlations, suggesting that the metallopeptidase superfamily may play an important role in the maturation of class III lanthipeptides. In contrast, families of serine protease and cysteine protease have been reported for the maturation of class I and II lanthipeptides[82,85,92,97]. Remarkably, among the 91 significant correlations representing 864 precursors, 758 precursors (88%) were strongly correlated with only one or two groups of proteases. For example, the precursors of groups Pre_5, Pre_24, and Pre_115 were only correlated to two, two, and one group(s) of proteases (ρ>0.3, pAdj<1E-5, I>=10; Figure

2.2c and Supplementary Table 2.2) at the genus level of *Streptomyces*, *Paenibacillus*, and *Lentzea*, respectively. None of the 888 proteases corelated to Pre_5, Pre_24, or Pre_115 had been characterized and only 19 of them are encoded by genes within the corresponding BGCs. Thus, this result exhibited the potential of our correlation network in identifying hidden proteases for the maturation of class III lanthipeptides.

On the other hand, five groups of precursors, Pre_77, Pre_117, Pre_134, Pre_181, and Pre_228, were each correlated to multiple groups of proteases, forming five multiple-correlation clusters at the genus level with *Amycolatopsis*, *Streptomyces*, *Alkalihalobacillus*, *Lactobacillus*, and *Rhodoccocus*, respectively (Figure 2.2c). Together, these five groups only represent the remaining 12% of the precursors from the 91 significant correlations. Some groups of proteases, e.g., Prot_1169, Prot_2308, and Prot_9513, were observed to share similar functional domains, which may partially account for their simultaneous correlations with the same precursor group. At first glance, this multiple-correlation pattern presented a challenge to identify a responsible protease. However, the correlation strength between different groups of proteases and the same group of precursors appeared to be different based on Spearman's rank correlation coefficient (ρ). In addition, we noticed that multiple-correlation clusters could be distinguished by the classes of their correlated proteases. For example, proteases associated with Pre_117 belong primarily to the α/ß hydrolase class while those associated with Pre_134 mainly fall into the metallopeptidase class (Supplementary Table 2.2). For the multiple-correlation pattern, to completement correlational analysis at the genomic level, we also integrated co-expression analysis using whole genome transcriptomic data to prioritize target proteases at the intersection of correlational and co-expression analyses.

For example, regarding the aforementioned Pre_5 that was strongly correlated with two groups of proteases (Figure 2.2c), Prot_241 (metal-dependent hydrolase; PF10118) and Prot_1365 (carbon-nitrogen hydrolase; PF00795), we compiled publicly available transcriptomic data consisting of 80 samples from three *Streptomyces* strains that expressed Pre_5. Co-expression analysis using these data reduced the original two strong correlations to only one group, Prot_241 ($\rho>0.4$, $pAdj<0.05$). Thus, we demonstrate here the establishment of a global correlation network, integrating co-expression data to fortify lanthipeptide precursor-protease association predictions.

2.2.2 Validation of the precursor-protease correlation network

We used previously characterized representative lanthipeptide proteases to validate the correlation network by first examining the better understood class I and II lanthipeptide systems. In our network, we located the previously studied class I lanthipeptides epidermin[98] and gallidermin[98] in one group, Pre_1, due to their precursor similarity, while the class II lanthipeptides thuricin[99], cerecidin[100], and cytolysin[101] were found in four groups of precursors, Pre_11, Pre_12, Pre_10, and Pre_13 (cytolysin is comprised of two distinct precursors resulting in its representation in two precursor groups) (Figure 2.4). The BGCs producing these lanthipeptides also encode pathway-specific LanP and LanT homologs, two types of proteases known to be involved in class I and II lanthipeptide biosynthesis[82,84,102] (Figure 2.5). Indeed, these LanP and LanT homologs each were shown in our network to correlate with Pre_1, Pre_11, Pre_12, Pre_10, and Pre_13, respectively (Table 2.1**)**. Additionally, we used the previously characterized class III lanthipeptide protease FlaP for validation. Besides the genera of *Kribbella* and *Streptomyces* that were reported to harbor a FlaP-like protease colocalized with a class III precursor[87], we

identified FlaA homologs in additional genera, mainly *Amycolatopsis* and *Jiangella* (Figure 2.6). FlaP-like proteases were simultaneously identified in our network to significantly correlate with FlaA in *Amycolatopsis* and *Jiangella*, with ρ=0.77, pAdj=3E-16 and ρ=1, p=8E-78, respectively (Supplementary Data 2.1). Such strong correlations supported the effectiveness of applying our correlation network to class III lanthipeptides.

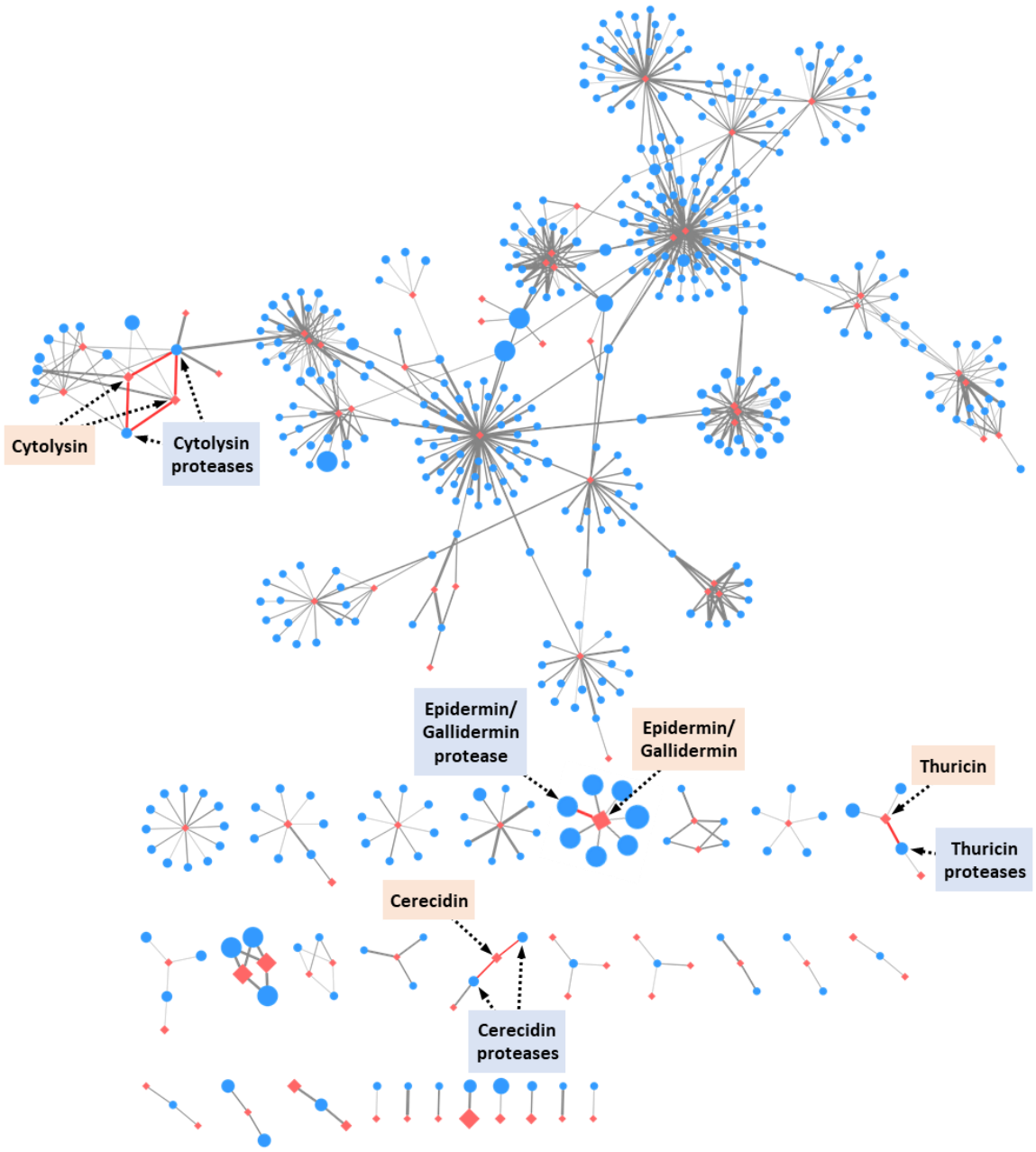


Figure 2.4 Prioritized correlation network of lanthipeptide precursor and protease groups. Correlational networking was performed between lanthipeptide precursor groups and protease groups. Network analysis was visualized using Cytoscape. Clusters were filtered for display using a set of thresholds: Spearman's correlation coefficient (ρ>0.5), false-discovery-rate adjusted p-value (pAdj<1E-5, one-sided *t*-test, adjusted by false-discovery rate), and number of genomes containing the correlation (I >= 10). Red diamonds represent precursor groups and blue circles represent protease groups. Shape sizes are scaled proportionally to the number of elements (proteases or precursors) contained in that group at the genus level. Groups containing more than 1000 elements were scaled individually. Edges between circles and diamonds indicate significant correlations as described above

with increasing thickness of the edge representing increasing strength of Spearman's correlation coefficient (ρ). Red edges connecting named precursor and protease groups represent literature-reported associations between the two groups.

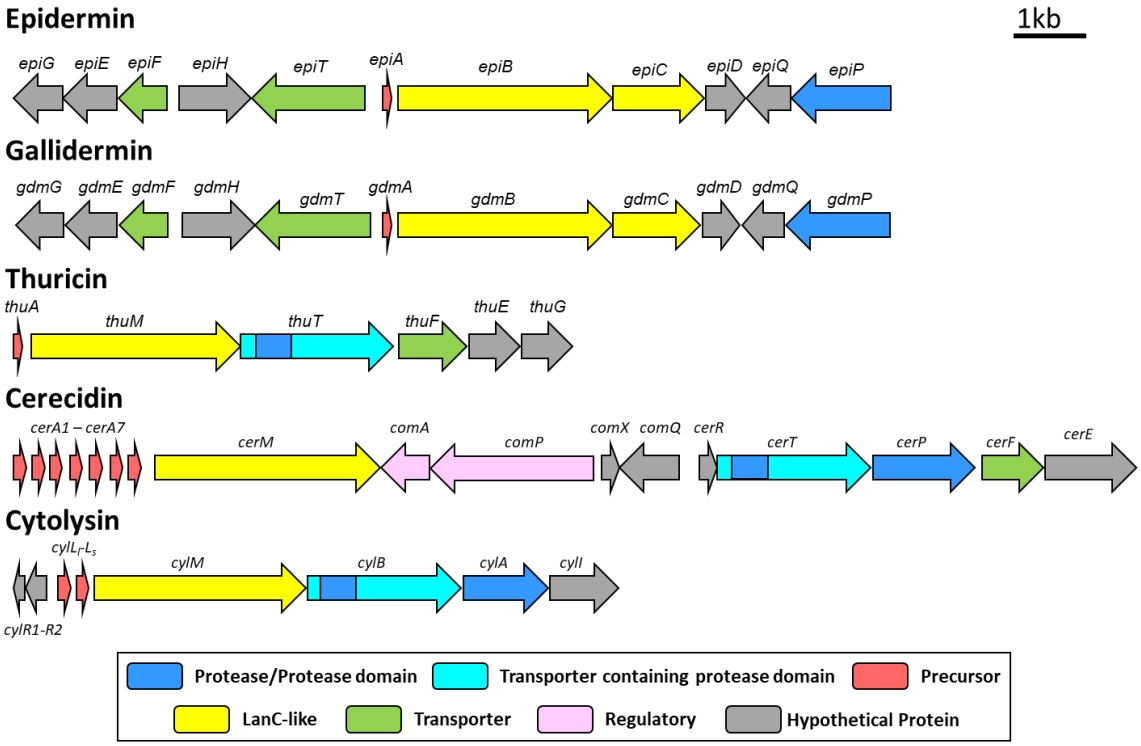


Figure 2.5 Identification of previously characterized lanthipeptide proteases to validate the correlational network. Five lanthipeptide BGCs producing epidermin, gallidermin, thuricin, cerecidin, and cytolysin, which were found to encode pathway-specific LanP and LanT-like proteases that are commonly known to be involved in class I and class II lanthipeptide biosynthesis. Genes encoding LanC-like proteins are represented in yellow, while genes encoding precursor peptides are represented in red, protease encoding genes in blue, transporter encoding genes in green, regulatory genes in pink, and genes encoding transporters which contain protease domains are represented in cyan with a small blue box to indicate the location of the protease domain.

Table 2.1 Correlations between selected lanthipeptide precursors and their proteases.

Genus	Peptide	Precursor	Precursor group	Protease	Protease group	Correlation
Staphylococcus	Epidermin	CAA44252.1	Pre_1	CAA44257.1	Prot_1211	ρ=0.97, pAdj=0, I=2955
	Gallidermin	ABC94902.1	Pre_1	ABC94907.1	Prot_1211	ρ=0.97, pAdj=0, I=2955
	Thuricin	AHX39582.1	Pre_11	AHX39584.1	Prot_4437	ρ=0.90, pAdj=0, I=297
Bacillus	Cerecidin	AHJ59543.1 AHJ59544.1 AHJ59545.1 AHJ59546.1 AHJ59547.1 AHJ59548.1 AHJ59549.1	Pre_12	AHJ59536.1	Prot_5982	ρ=0.72, pAdj=0, I=126
				AHJ59535.1	Prot_5818	ρ=0.85, pAdj=0, I=126
Enterococcus	Cytolysin	AAA62648.1	Pre_10	AAA62651.1	Prot_5642	ρ=0.90, pAdj=0, I=239
				AAA62652.1	Prot_5225	ρ=0.92, pAdj=0, I=250
		AAA62649.1	Pre_13	AAA62651.1	Prot_5642	ρ=0.90, pAdj=0, I=239
				AAA62652.1	Prot_5225	ρ=0.92, pAdj=0, I=250

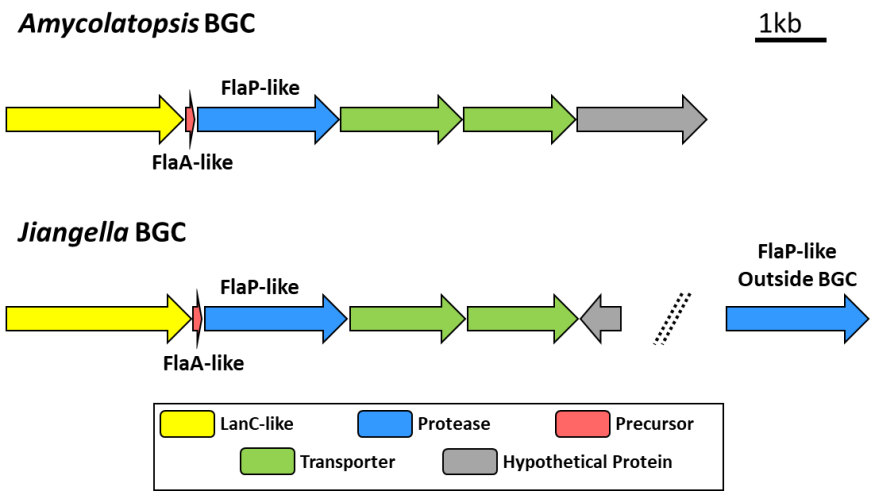


Figure 2.6 Class III lanthipeptide BGCs containing FlaA-like precursors. BGCs containing FlaA-like precursors and FlaP-like proteases were identified from the genera *Amycolatopsis* and *Jiangella*. An additional FlaP-like protease was detected outside of the BGC in the genus *Jiangella*. Genes encoding LanC-like (LanKC) proteins are represented in yellow, while genes encoding precursor peptides are represented in red, protease genes in blue, and transporter genes in green.

2.3 Interrogation of the Correlational Network

2.3.1 Identification of a protease for the maturation of a class I lanthipeptide

To identify lanthipeptide proteases that are not encoded by their respective BGCs, we first selected a group of class I lanthipeptide precursors, Pre_49 (Figure 2.7a), which presumably encodes paenilan as a representative product based on the precursor sequences[103] (Figure 2.7b). All 44 BGCs of this group do not harbor any protease-encoding genes within the BGCs and the responsible proteases for maturation of this group of lanthipeptides have not been reported[103], consistent with the recent finding that increasingly more class I lanthipeptide BGCs have been noticed to lack a pathway-specific protease[84,89,90]. In our network, 34 groups of proteases were shown to be strongly correlated with Pre_49 (ρ>0.3, pAdj<1E-5, I>=10). Additionally, integrating transcriptomic data of 18 samples from *Paenibacillus polymyxa* ATCC 842 expressing the precursor of Pre_49

(Supplementary Data 2.1), we looked for the intersection of proteases identified by both correlational and co-expression analyses and reduced the putative correlations from 34 groups to four groups, Prot_686, Prot_1570, Prot_2941, and Prot_8704 (Figure 2.7c). Thus, we expressed the corresponding proteases of these four groups from *P. polymyxa* ATCC 842 as recombinant proteins and tested their proteolytic activity *in vitro* against the PllB/PllC-modified precursor peptide from the same strain. High-performance liquid chromatography-mass spectrometry (HPLC-MS) analysis revealed that paenilan, the same product as the native strain, was produced in the assay with the member of Prot_686 (Figure 2.7d). Prot_686 belongs to the S8 family of peptidases. This family has been previously reported for the maturation of class I lanthipeptides[80,92]. Thus, we used the correlation of Prot_686 with a paenilan-producing precursor to demonstrate that proteases predicted by our network analysis could in fact be responsible for lanthipeptide maturation. This motivated us to further exploit the power of our network in the less understood class III lanthipeptide system.

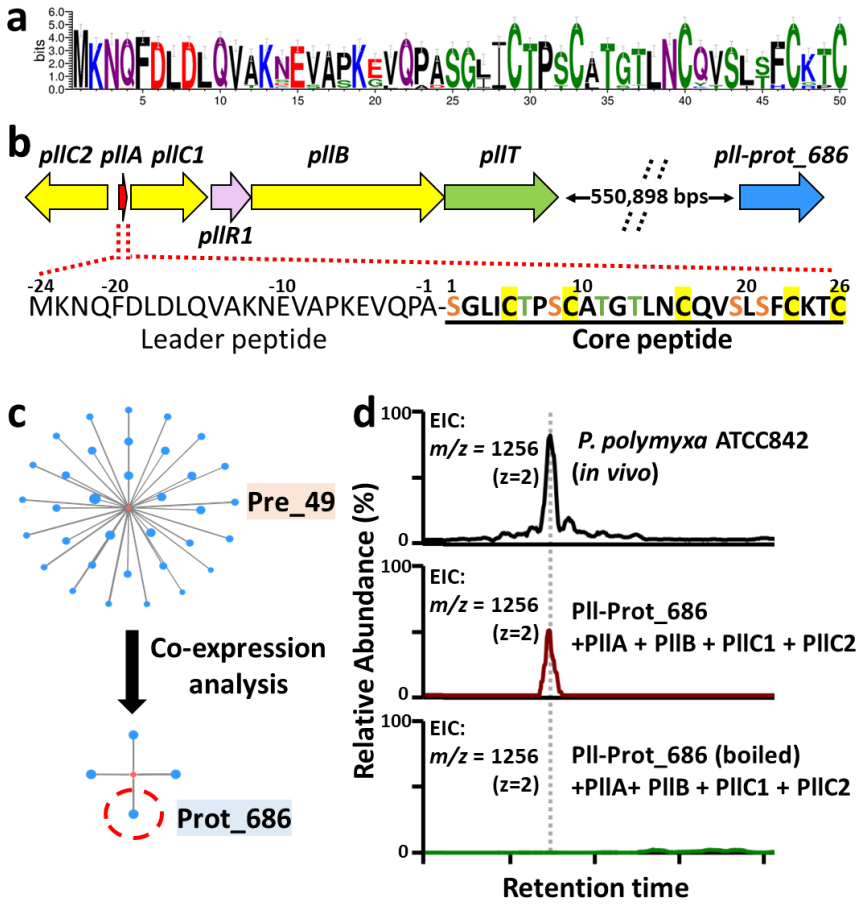


Figure 2.7 Identification of a hidden protease from Prot_686 for maturation of paenilan. **a**, Sequence logo of the group Pre_49. Sequence alignments were trimmed by trimAl to remove gap-rich regions before drawing logo. **b**, A representative BGC of Pre_49 in *P. polymyxa* ATCC 842 that presumably produces paenilan based on the precursor sequence. All 44 BGCs of this group do not harbor any protease-encoding genes, suggesting hidden proteases encoded elsewhere in the genome. Dashed lines separating *pllT* and *prot_686* indicate a gap in their respective genetic locations. This type of notation is also used in other figures below. The distance of this gap, 550,898 base pairs, is represented here to convey magnitude. **c**, Intersection of precursor-protease correlation (ρ>0.3, pAdj<1E-5, one-sided *t*-test, adjusted by false-discovery rate, I>=10) and co-expression (ρ>0.4, pAdj<0.05, one-sided *t*-test, adjusted by false-discovery rate) suggested candidate proteases(s). **d**, HPLC-MS analysis indicated the proteolytic activity of Prot_686 for the maturation of paenilan.

2.3.2 Discovery of a new family of lanthieptide proteases linked to the production of new class III lanthipeptides

Class III lanthipeptide BGCs are abundant in Firmicutes[73]. However, the lack of a protease-encoding gene in most class III BGCs has been a challenge toward exploiting these lanthipeptides[86], such as heterologous expression for the discovery of new class III lanthipeptides, pathway bioengineering for increasing production yield, chemical diversity, and biological activities, and leveraging the enzymology of proteases as a synthetic biology tool for general proteolytic and traceless tag removal applications[82]. Therefore, we focused on Firmicutes for a simultaneous discovery of previously unknown class III lanthipeptides and their associated hidden proteases. Our attention was drawn to two groups of proteases designated as Prot_176 and Prot_819, both of which had not been studied as lanthipeptide proteases and showed strong correlations (ρ=0.69, pAdj=4E-64) to a group of uncharacterized precursors, Pre_24 (Figure 2.2c). Pre_24, consisting of totally 115 precursors, is solely distributed in Firmicutes and represents the largest precursor group in Firmicutes (Figure 2.8a). These 115 precursors are distributed across 48 class III lanthipeptide BGCs from three genera, *Alkalihalobacillus*, *Bacillus*, and *Paenibacillus*. The strains harboring *pre_24* all possess one pair of *prot_176* and *prot_819* present side by side outside of the 48 BGCs (Figure 2.8b). Among these 48 BGCs, an extra pair of *prot_176* and *prot_819* was identified within 17 BGCs (Figure 2.8b). In addition, many other strains from *Alkalihalobacillus*, *Bacillus*, and *Paenibacillus* that do not encode Pre_24 also contain a pair of *prot_176* and *prot_819* in their genomes (Figure 2.8b).

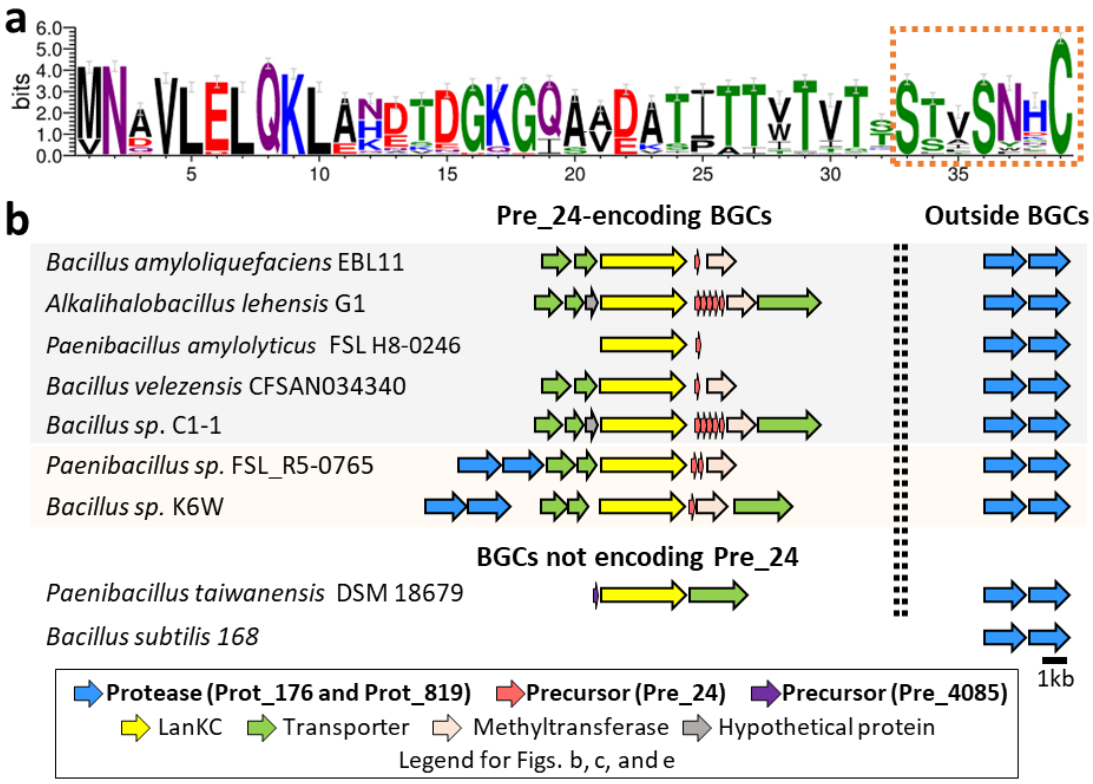


Figure 2.8 A family of previously unknown lanthipeptide proteases potentially linked to the production of unknown class III lanthipeptides. **a**, Sequence logo of Pre_24. Sequence alignments were trimmed by trimAl to remove gap-rich regions before drawing logo. The putative core peptides contain only one S-$(X)_2$-S-$(X)_{2\text{-}5}$-C motif (X stands for random amino acid) as opposed to two or three such motifs commonly seen in previously known class III lanthipeptides. **b**, Representative class III lanthipeptide BGCs encoding Pre_24, showing distribution of a pair of protease-encoding genes, *prot_176* and *prot_819*, outside of the BGCs and an extra pair in a minority of the BGCs. In addition, a pair of *prot_176* and *prot_819* is also present in the genomes of many strains that do not harbor Pre_24-encoding BGCs.

To investigate Prot_176 and Prot_819 as potential proteases for Pre_24, we first produced and characterized representative class III lanthipeptides from Pre_24. From the network, we selected two Pre_24-encoding BGCs for study. The first was from *Bacillus nakamurai* NRRL B-41092, designated as *bcn* (Figure 2.9a). To check for the production of lanthipeptides by this strain, *B. nakamurai* NRRL B-41092 was cultured and analyzed the *n*-butanol extract of the culture broth was analyzed by HPLC-MS. A doubly charged molecular ion at *m/z* 1113.0 $[M + 2H]^{2+}$ was detected, indicative of a molecular formula

$C_{106}H_{141}N_{27}O_{25}S$ (Figure 2.9b). Subsequent analysis using tandem mass spectrometry (MS/MS) with collision-induced dissociation (CID) suggested that the mature peptide, named bacinapeptin A, contains 23 amino acid residues including 11 dehydrated amino acids (e.g., Dhb, Dha) as well as a *N*,*N*-dimethylated alanine (Figure 2.10a). Based on the diagnostic fragment ions (b and y ions) generated by MS/MS, partial sequences of bacinapeptin A were established to be Dhb-Dhb-Dhb-Trp-Dhb-Val-Dhb-Dhb-Dhb and Ala-Phe, which are consistent with the predicted core peptide sequences prior to post-translational modifications, i.e. Thr-Thr-Thr-Trp-Thr-Val-Thr-Thr-Thr (TTTWTVTTT) and Ala-Phe (AF), respectively. Moreover, the lack of diagnostic fragments ions for the remaining sequence Ser-Thr-Val-Ser-Asn-His-Cys (STVSNHC) at the C-terminus suggested the formation of a lanthionine (LAN) or labionin (LAB) ring, which stabilizes the structure. Considering the nature of labionin formation, we propose that a C-terminal bicyclic labionin ring was generated through Michael addition cyclization among Dha17, Dha20, and Cys23 to give the final structure of bacinapeptin A (Figure 2.10b).

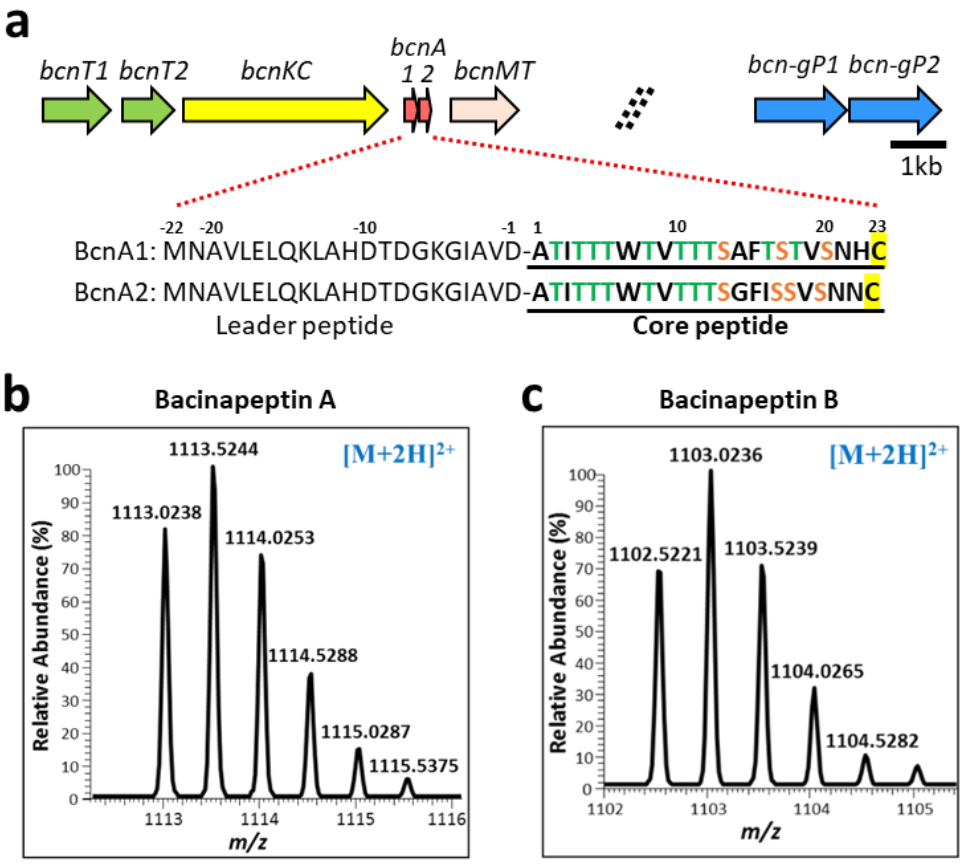


Figure 2.9 The bcn BGC and its two putative class III lanthipeptide products. **a**, A Pre_24-encoding BGC, *bcn*, selected for study. The *bcn* BGC putatively encodes a class III lanthipeptide synthetase (BcnKC), two precursors (BcnA1 and BcnA2), a methyltransferase (BcnMT), and a transporter (BcnT). A pair of protease-encoding genes belonging to *prot_819* and *prot_176*, respectively, is absent in the *bcn* BGC, but present elsewhere in the genome, designated as *bcn-genomeP1* (abbreviation *bcn-gP1*) and *bcn-genomeP2* (abbreviation *bcn-gP2*). **b**, The doubly charged state of bacinapeptin A is provided which was detected at 1113.02 m/z. **c**, The doubly charged state of bacinapeptin B is provided which was detected at 1102.52 m/z.

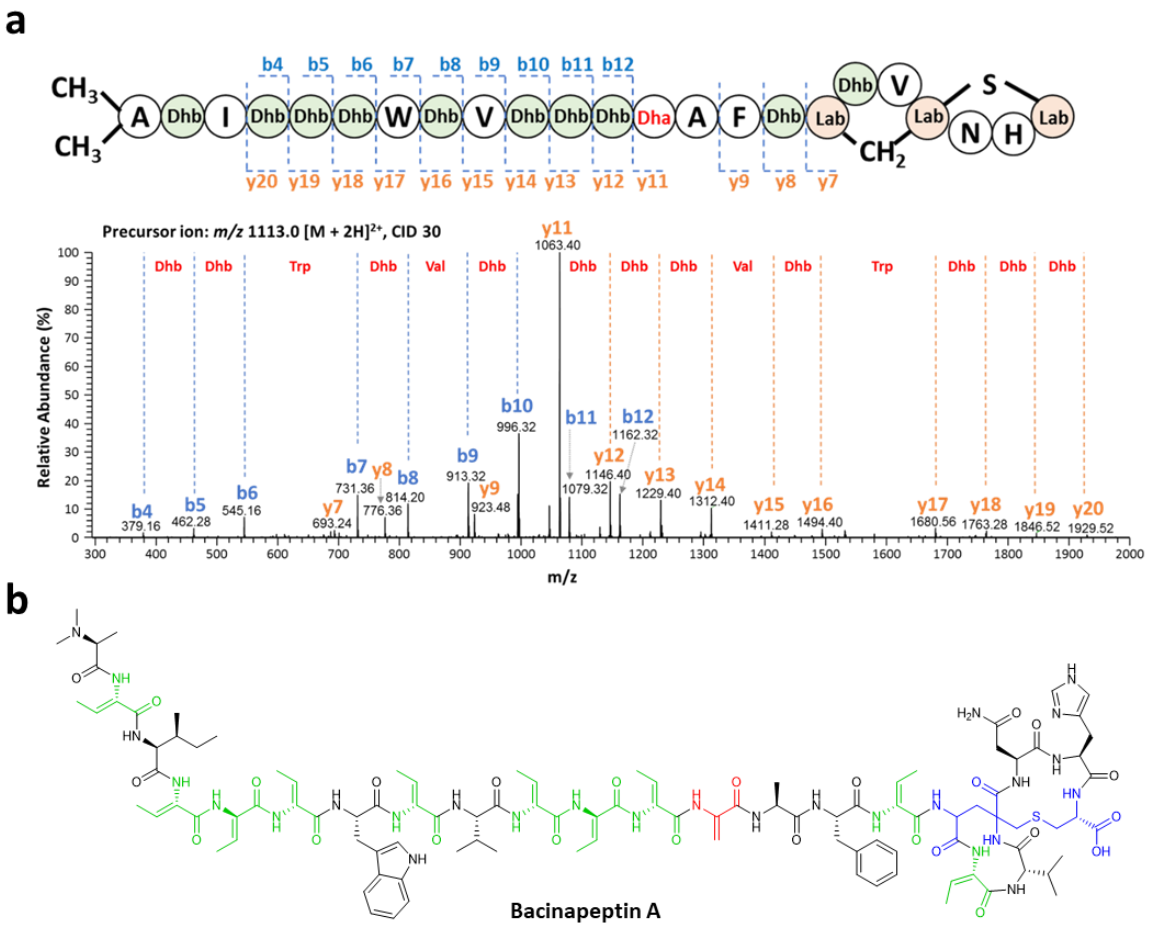


Figure 2.10 Structure elucidation of bacinapeptin A by MS/MS fragmentation analysis. **a**, The amino acid sequence of bacinapeptin A with b and y ions marked as well as the location of the labionin ring. MS/MS was used to fragment bacinapeptin A to confirm the amino acid sequence by examination of the fragmentation patterns. The doubly charged precursor ion, *m/z* 1113.0, was selected for collision induced dissociation (CID) at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **b,** Chemical structure of bacinapeptin A. Chemical formula: $C_{106}H_{141}N_{27}O_{25}S$. Dhb is in green, Dha is in red, and Ser and Cys involved in the labionin ring are in blue.

Besides bacinapeptin A, another minor peak at *m/z* 1102.6 $[M + 2H]^{2+}$ was detected in the *n*-butanol extract of *B. nakamurai* NRRL B-41092 (Figure 2.9c). It showed a similar MS/MS fragmentation pattern to bacinapeptin A, so it was named bacinapeptin B. Detailed interpretation of MS/MS spectrum of bacinapeptin B returned a partial peptide sequence Dhb-Dhb-Dhb-Dha-Gly-Phe-Ile, corresponding to a part of predicted core peptide

sequence Thr-Thr-Thr-Ser-Gly-Phe-Ile (TTTSGFI) prior to post-translational modifications, which is different from that for bacinapeptin A (TTTSAFT). Moreover, the y_7 ion in bacinapeptin B is 37 Da less than that in bacinapeptin A, suggesting a Dhb and a His in the labionin ring may be replaced by Dha and Asn in bacinapeptin B (Figure 2.11a). This hypothesis was further validated by the predicted core peptide sequence of bacinapeptin B, in which Thr18 was replaced by Ser18, and His22 was replaced by Asn22. The amino acid residues involved in labionin formation and *N*,*N*-dimethylation remain identical (Figure 2.11b).

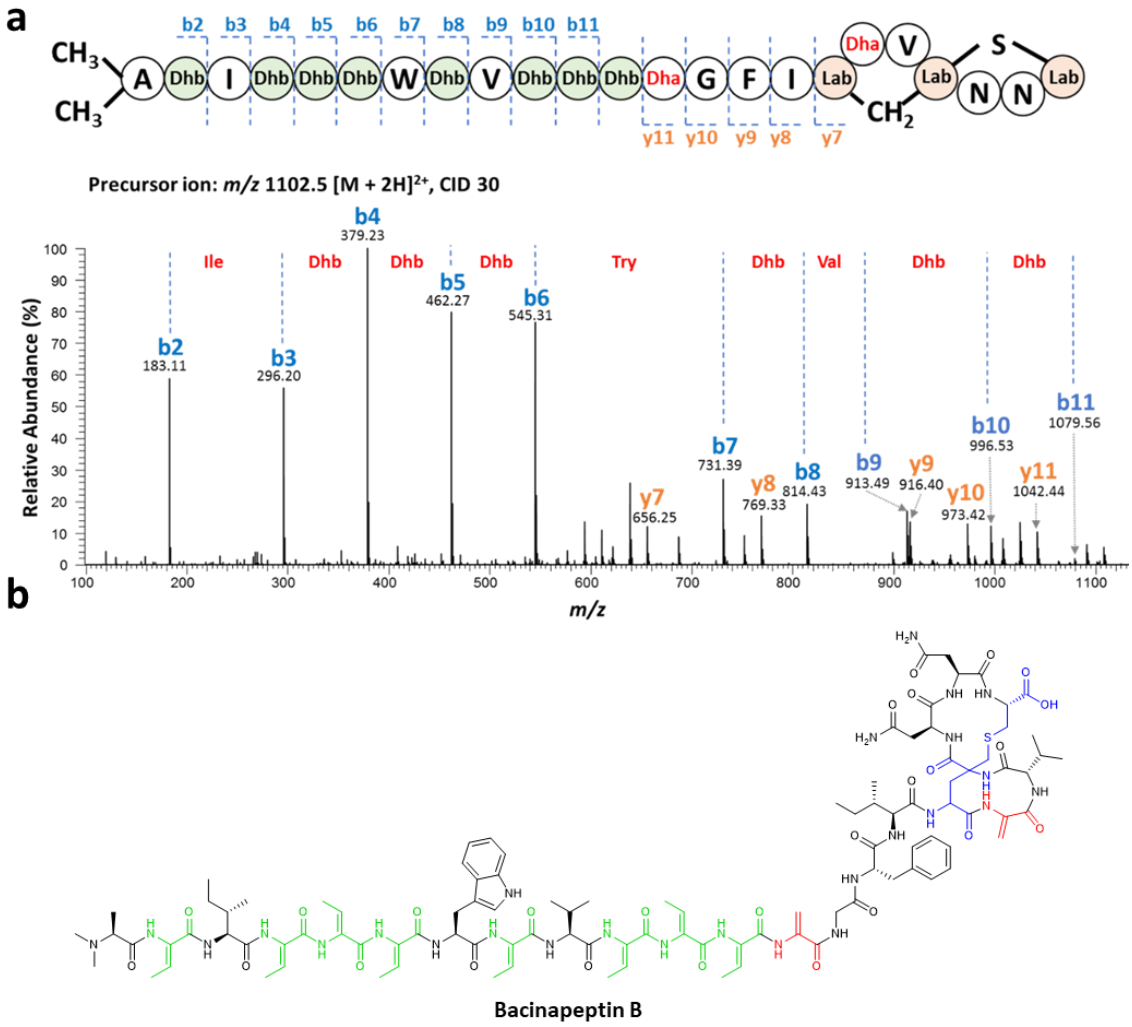


Figure 2.11 Structure elucidation of bacinapeptin B by MS/MS fragmentation analysis. **a**, The amino acid sequence of bacinapeptin B with b and y ions marked as well as the location of the labionin ring. MS/MS was used to fragment bacinapeptin B to confirm the amino acid sequence by examination of the fragmentation patterns. The doubly charged precursor ion, *m/z* 1102.5, was selected for collision induced dissociation (CID) at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **b**, Chemical structure of bacinapeptin B. Chemical formula: $C_{104}H_{142}N_{26}O_{26}S$. Dhb is in green, Dha is in red, and Ser and Cys involved in the labionin ring are in blue.

We next turned to confirm that bacinapeptins A and B are produced by the *bcn* BGC. The gene organization of the *bcn* BGC is shown in Figure 2.9a, with *bcnA1* and *bcnA2* corresponding to *pre_24*. The *bcn* BGC does not harbor any protease-encoding genes. Instead, outside of the *bcn* BGC, a pair of protease-encoding genes was identified

as belonging to *prot_819* and *prot_176*, respectively (Figures 2.2c and 2.8c). This pair of genes was designated as *bcn-genomeP1* (abbreviation *bcn-gP1* hereafter) and *bcn-genomeP2* (abbreviation *bcn-gP2* hereafter). The *bcn* BGC, together with *bcn-gP1* and *bcn-gP2*, was constructed into a pDR111-based integrative plasmid (Supplementary Table 2.4) and heterologously expressed in the host *Bacillus subtilis* 168, leading to the detection of bacinapeptins A and B (Figure 2.12).

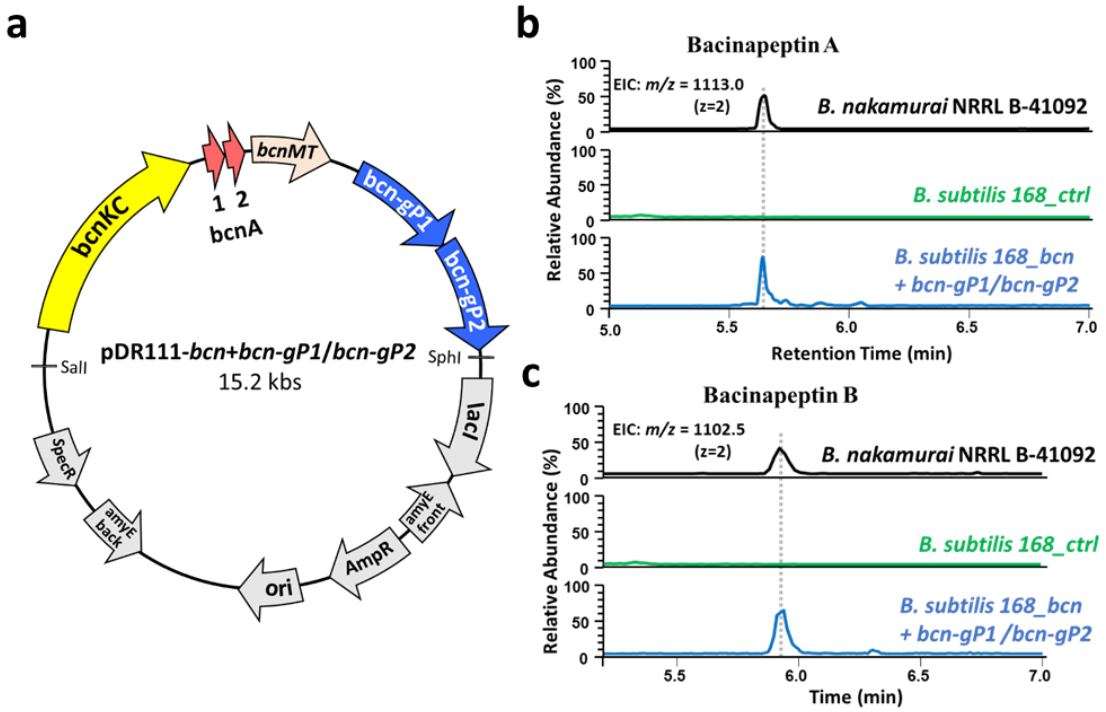


Figure 2.12 Heterologous production of bacinapeptins A and B. **a**, Plasmid map of pDR111-*bcn+bcn-gP1/bcn-gP2* that was used for the heterologous expression of the *bcn* BGC in *B. subtilis* 168. Grey arrows indicate elements in the vector. Arrows with other colors show genes of *bcn* BGC and *bcn-gP1/bcn-gP2*. **b**,**c**, Extracted ion chromatograms (EIC) of extracts from three samples: *B. nakamurai* NRRL B-41092, *B. subtilis* 168_ctrl, and *B. subtilis* 168_*bcn+bcn-gP1/bcn-gP2* are overlaid for comparison, showing the production of bacinapeptin A (**b**) and B (**c**) by the heterologous host.

The second BGC selected was from *Paenibacillus thiaminolyticus* NRRL B-4156, designated as *ptt*. In the *n*-butanol extract of the *P. thiaminolyticus* NRRL B-4156 culture

broth, we could detect several putative lanthipeptides, but their abundance was too low for isolation and structure elucidation. Encouraged by the success of the bacinapeptin heterologous expression system, we set out to construct a heterologous expression system for the *ptt* BGC. The gene organization of the *ptt* BGC is shown in Figure 2.13a, with *pttA1-A7* belonging to *pre_24*. Outside of the *ptt* BGC, there is a pair of protease-encoding genes, *ptt-genomeP1* (hereafter abbreviated as *ptt-gP1*) and *ptt-genomeP2* (hereafter abbreviated as *ptt-gP2*), that belongs to *prot_819* and *prot_176*, respectively (Figure 2.2c and Figure 2.13a). Within the *ptt* BGC, there is an additional pair of protease-encoding genes, *pttP1* and *pttP2*, also belonging to *prot_819* and *prot_176*, respectively. The *ptt* BGC, with *pttP1* and *pttP2*, or with *ptt-gP1* and *ptt-gP1* replacing *pttP1* and *pttP2*, was cloned into pDR111 (Figure 2.13b and Supplementary Table 2.4) and heterologously expressed in the host *B. subtilis* 168, respectively, both leading to the detection of paenithopeptins A-E (Figure 2.14), with the system using the in-BGC *pttP1*/*pttP2* protease pair showing a higher production yield. Using this system, we isolated the major product, paenithopeptin A, for detailed structure elucidation.

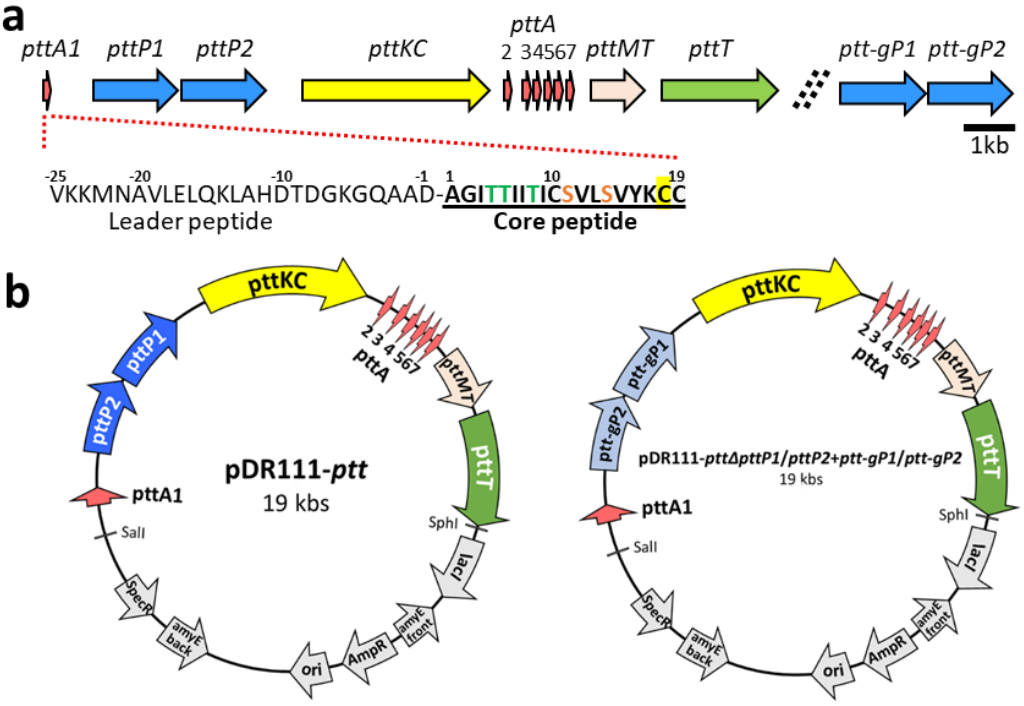


Figure 2.13 Construction of a heterologous expression system for the *ptt* BGC. **a**, The *ptt* BGC putatively encodes a characteristic class III lanthipeptide synthetase (PttKC) and seven precursors, PttA1-PttA7. While *pttA2-pttA7* are together, *pttA1* is located separately. The putative protease-encoding genes *pttP1* and *pttP2* belonging to *prot_819* and *prot_176*, respectively, are adjacent to each other downstream of *pttA1*. An extra pair of protease-encoding genes also belonging to *prot_819* and *prot_176* is located outside of the *ptt* BGC, designated as *ptt-genomeP1* (*ptt-gP1*) and *ptt-genomeP2* (*ptt-gP2*), respectively. *pttMT* and *pttT* are predicted to encode a methyltransferase and a transporter, respectively. **b**, Plasmid maps of pDR111-*ptt* for the heterologous expression of the *ptt* BGC in *B. subtilis* 168 and pDR111-*pttΔpttP1/pttP2+ptt-gP1/ptt-gP2* where *pttP1* and *pttP2* are replaced by *ptt-gP1* and *ptt-gP2* (light blue arrows). Grey arrows indicate elements in the vector.

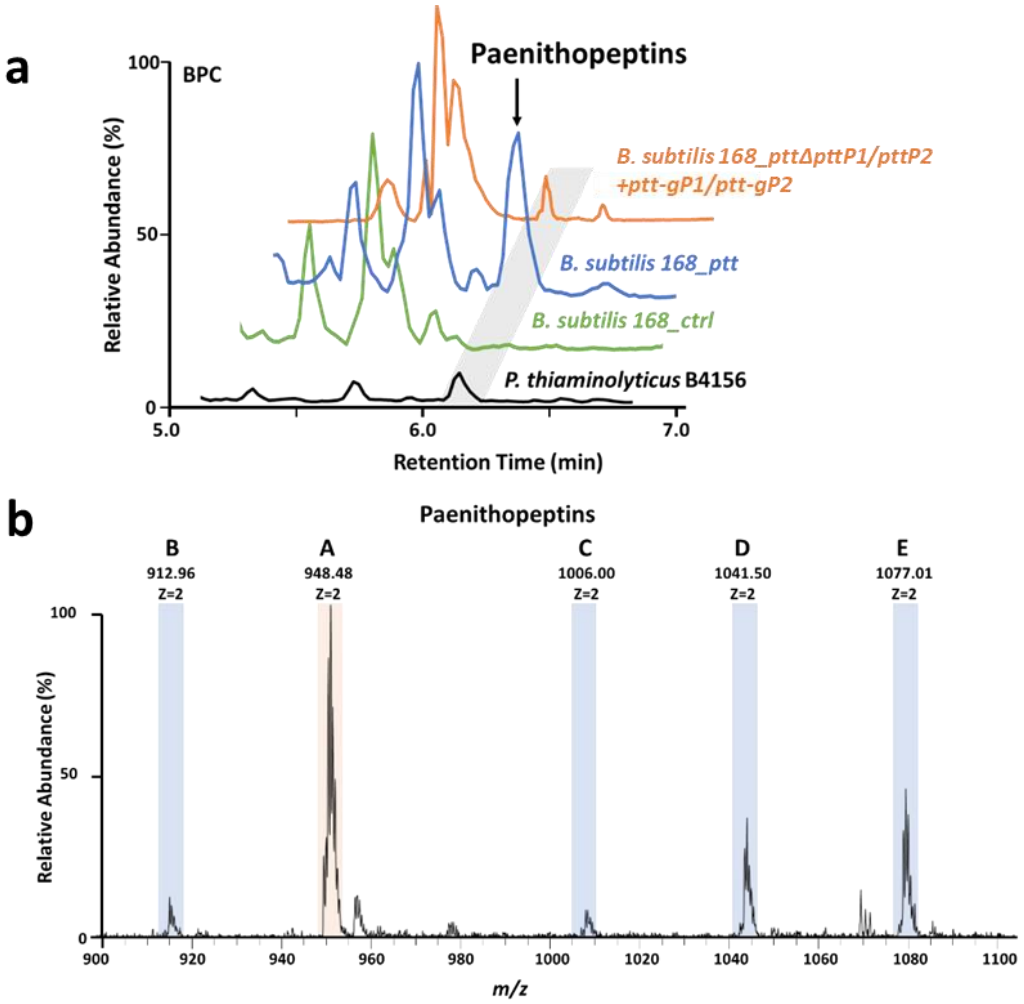


Figure 2.14 Heterologous production of paenithopeptins A through E. **a**, Base peak chromatograms of extracts from four samples: *B. subtilis* 168_*ptt*, *B. subtilis* 168_ *pttΔpttP1/pttP2+ptt-gP1/ptt-gP2*, *B. subtilis* 168_ctrl, and *Paenibacillus thiaminolyticus* NRRL B-4156 are overlaid for comparison with the paenithopeptins peak highlighted. **b**, A selected mass spectral scan of the indicated peak (retention time 5.75-6.75 min) displays the doubly charged states of all paenithopeptins A-E.

High-resolution mass spectrometry (HRMS) analysis of paenithopeptin A returned a doubly charged (*m/z* 948.4825) and a triply charged (*m/z* 632.6581) molecular ions, indicative of a molecular formula $C_{87}H_{138}N_{20}O_{21}S_3$ (Δppm -0.42) (Figure 2.15). MS/MS-based peptide sequencing of paenithopeptin A revealed diagnostic fragment ions (b and y ions), matching to the molecular weight of single amino acid residue or their combinations,

such as dehydrobutyrine (Dhb), isoleucine (Ile), and the dipeptide alanine-glycine (Ala-Gly) (Figure 2.16). Careful analysis of the fragment ions established the partial sequence of paenithopeptin A to be Ala-Gly-Ile-Dhb-Dhb-Ile-Ile-Dhb-Ile, which is consistent with the predicted core peptide sequence, Ala-Gly-Ile-Thr-Thr-Ile-Ile-Thr-Ile (AGITTIITI), prior to post-translational modification (e.g., dehydration of Thr to Dhb). However, the remaining core peptide sequence Cys-Ser-Val-Leu-Ser-Val-Tyr-Lys-Cys-Cys (CSVLSVYKCC) at the C-terminus cannot be deduced solely from MS/MS fragmentation analysis due to the lack of corresponding fragment ions, suggesting the formation of lanthionine (LAN), labionin (LAB), or disulfide ring that stabilizes the peptide structure. Co-incubation of paenithopeptin A with the redox reagent dithiothreitol (DTT) led to the yield of a reduced paenithopeptin A′ with the mass of 2 Da larger than that of paenithopeptin A, highly suggesting the reduction of a disulfide bond to form two sulfhydryl groups (Figure 2.17). Subsequent MS/MS analysis of paenithopeptin A′ revealed a similar fragmentation pattern to that of paenithopeptin A, but with significant differences being that (i) the molecular weights of all the y ions are 2 Da larger than those for paenithopeptin A; and (ii) a doubly charged b_{18} ion (*m/z* 889.0) deficient in the cysteine residue at the far end of C-terminus was observed (Figure 2.18). These observations indicated that a disulfide bridge is highly likely formed between Cys19 and Cys10. Moreover, the mutation of Cys19 to Ala19 abolished the formation of disulfide bridge as evidenced form the HPLC-MS and MS/MS analysis (Figure 2.19), reinforcing that Cys19 is involved in the formation of disulfide bond in paenithopeptin A. Analogous to labyrinthopeptin, we proposed that a labionin ring, characteristic of class III lanthipeptides, is formed among Cys18, Dha15 and Dha11 to generate a bicyclic ring system at the C-

terminus. To prove this hypothesis, the mutant C18A was established and showed the abolishment of peptide production (Figure 2.19), probably due to the instability of the peptide prior to labionin formation and thus supporting the labionin formation. 1D and 2D NMR analysis (DMSO-d_6) of paenithopeptin A (Figures 2.14-2.20) revealed two pairs of olefinic signals (δ_H 6.65, δ_C 115.3; δ_H 6.91, δ_C 130.5) representing the *para*-substituted phenolic ring in tyrosine residue, and three olefinic signals (δ_H 6.44, δ_C 129.9; δ_H 6.44, δ_C 129.9; δ_H 6.28, δ_C 127.6) for three Dhb residues. However, signals for Dha were not observed in the NMR spectra of paenithopeptin A, indicating all Dha residues are substituted and highly likely form labionin ring with Cys18. Marfey's analysis demonstrated all the proteinogenic amino acids in paenithopeptin A possess L-configuration (Table 2.2).

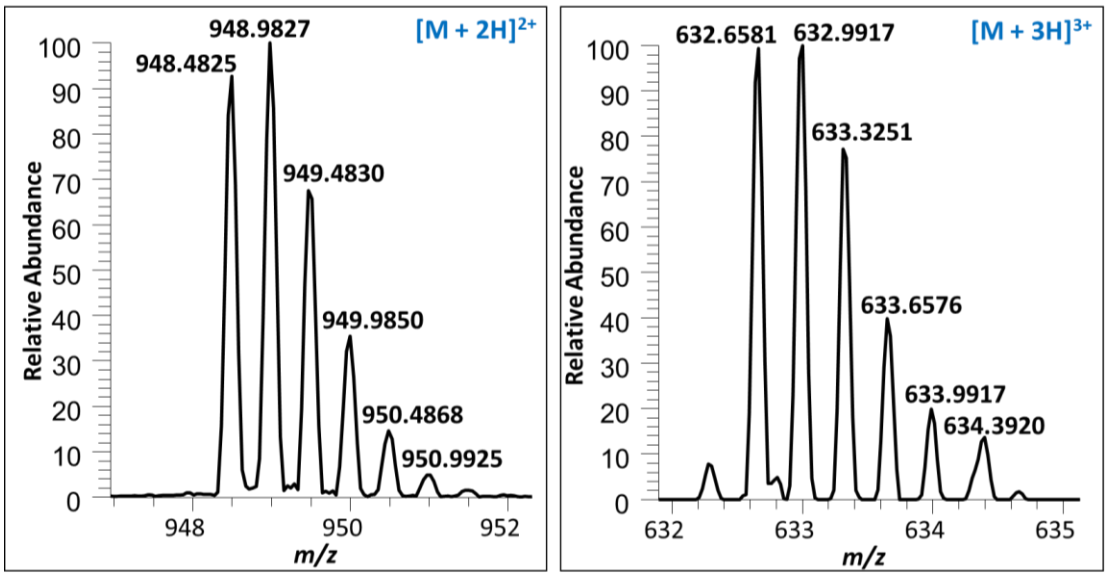


Figure 2.15 High resolution mass spectrometry of paenithopeptin A. Paenithopeptin A was subjected to high-resolution mass spectrometry for accurate mass determination. The monoisotopic molecular weight of paenithopeptin A was calculated at 1894.9650 Da. The calculated mass for its doubly ($[M+2H]^{2+}$) and triply charged ($[M+3H]^{3+}$) states are 948.5 m/z and 632.67 m/z, respectively. Shown here are MS spectra representing the doubly charged and triply charged states of paenithopeptin A, found at 948.4825 m/z and 632.6581 m/z, respectively.

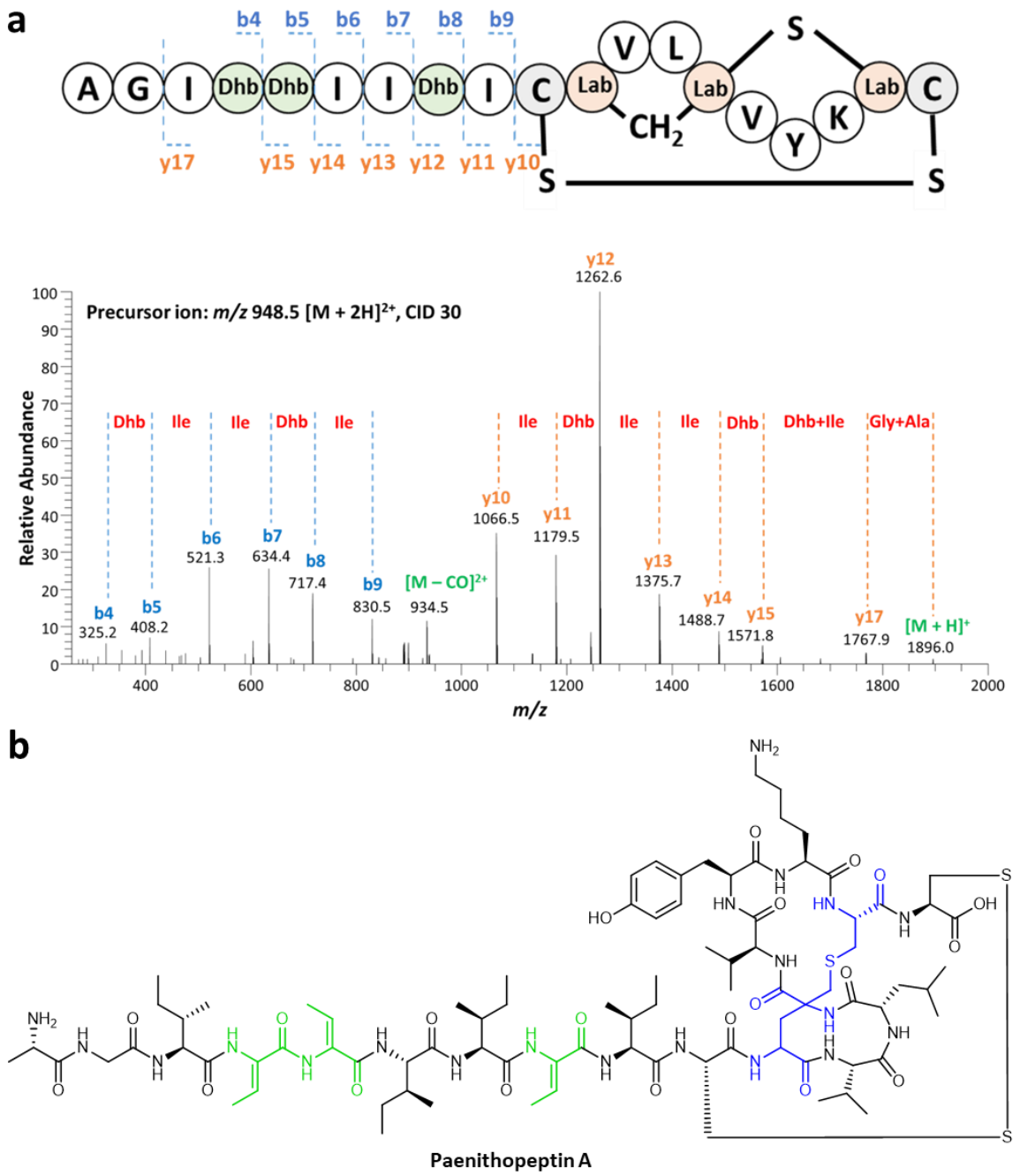


Figure 2.16 Structure elucidation of paenithopeptin A by MS/MS fragmentation analysis. **a**, The amino acid sequence of paenithopeptin A with b and y ions marked as well as the location of the labionin ring and disulfide bond. MS/MS was used to fragment paenithopeptin A to confirm the amino acid sequence by examination of the fragmentation patterns. The doubly charged precursor ion, *m/z* 948.5, was selected for collision induced dissociation (CID) at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **b**, Chemical structure of paenithopeptin A. Chemical formula: $C_{87}H_{138}N_{20}O_{21}S_3$. Dhb is in green, and Ser and Cys involved in the labionin ring are in blue.

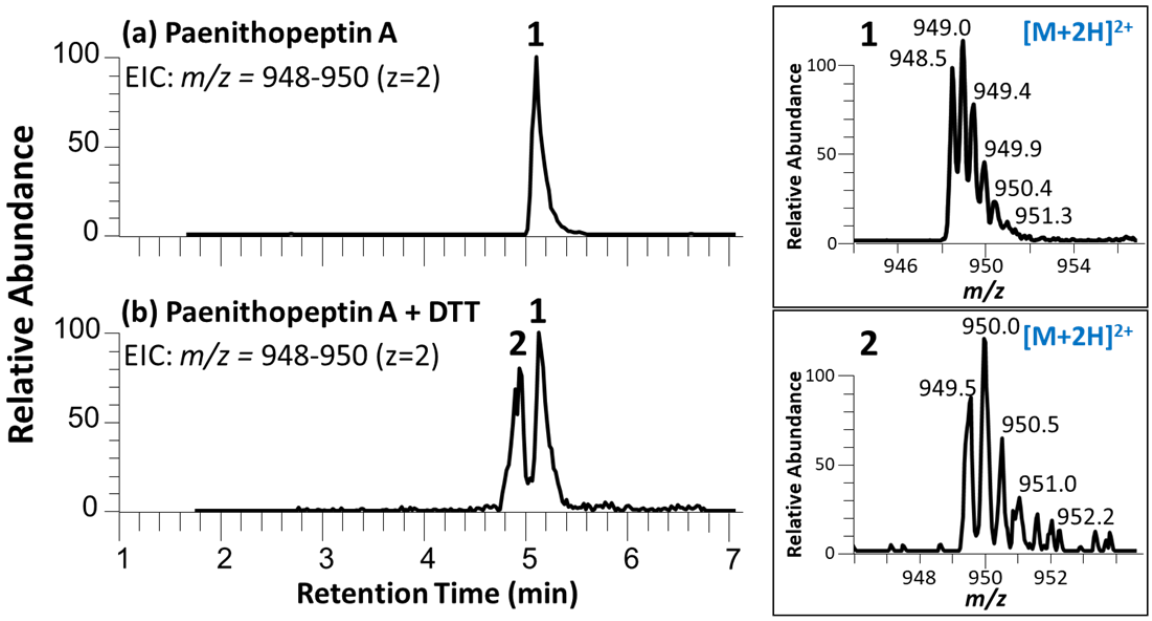


Figure 2.17 Chemical reduction of paenithopeptin A reveals presence of disulfide bond. Inconsistency between observed accurate mass and predicted mass of paenithopeptin A suggested the presence of an additional structural feature. Hypothesizing the presence of a disulfide bond as seen in some other class III lanthipeptides, paenithopeptin A was incubated with dithiothreitol (DTT) for 12 h. Extracted ion chromatograms of m/z range 948 - 950 were used to detect the doubly charged state of paenithopeptin A (peak 1) **(a)** before DTT treatment, and **(b)** after DTT treatment (peak 2). Zoomed isotopic masses for labelled peaks are presented at right.

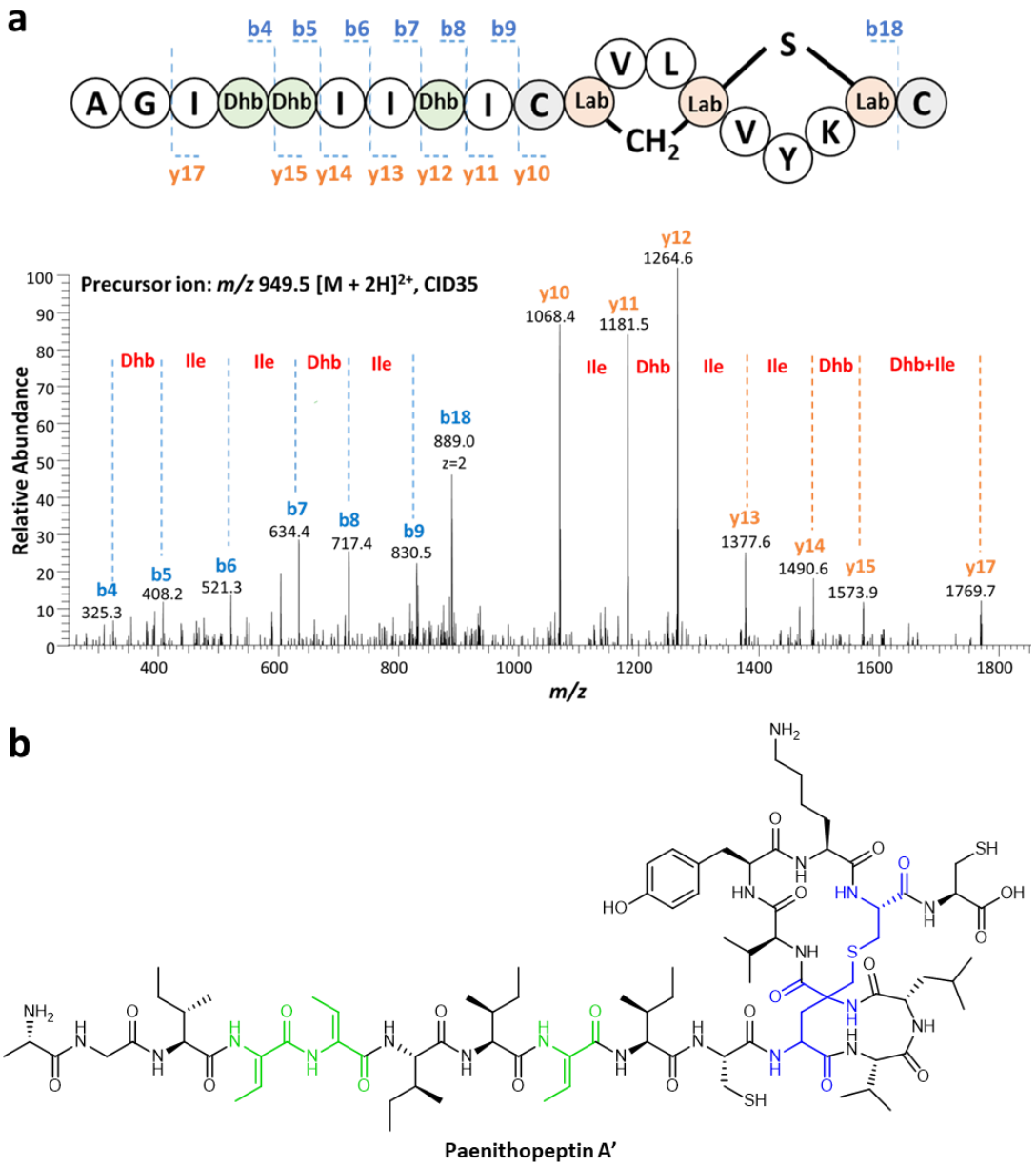


Figure 2.18 MS/MS fragmentation analysis of paenithopeptin A after DTT treatment. MS/MS was used to fragment paenithopeptin A after DTT treatment (paenithopeptin A') to confirm the presence of a disulfide bond by examination of the fragmentation patterns. **a**, The amino acid sequence of paenithopeptin A' with b and y ions marked as well as the location of the labionin ring is presented above. The doubly charged precursor ion, *m/z* 949.5, was selected for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. Appearance of b ion b18 compared to non-DTT treated paenithopeptin A indicated the reduction of the disulfide bond. **b**, Chemical structure of paenithopeptin A′. Chemical formula: $C_{87}H_{140}N_{20}O_{21}S_3$. Dhb is in green, and Ser and Cys involved in the labionin ring are in blue.

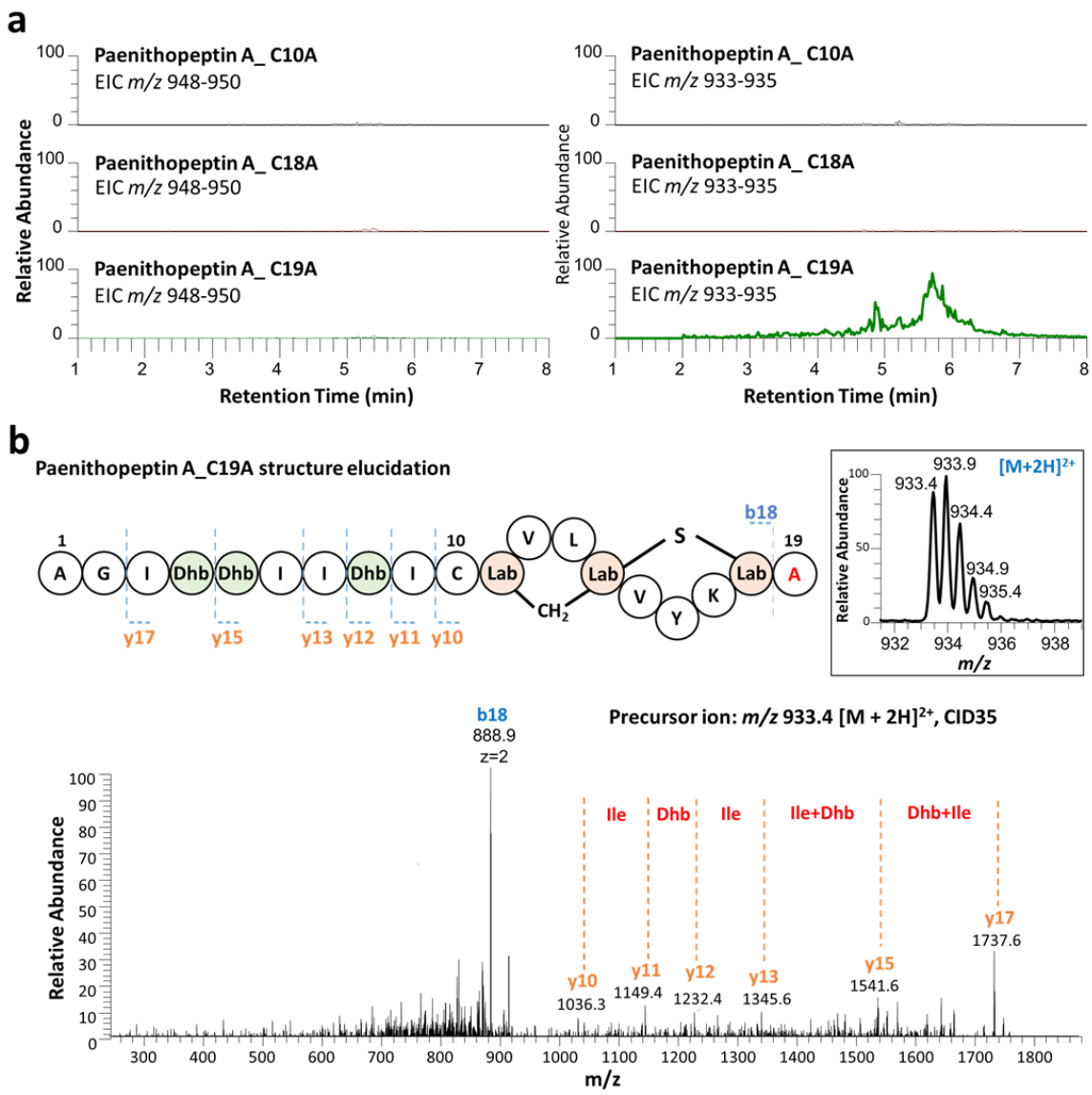


Figure 2.19 Extracted ion chromatograms (EICs) of paenithopeptin A mutations. Point mutation to alanine of each cysteine residue possibly involved in disulfide bond formation was performed to provide secondary support to the presence and location of the disulfide bond. **a**, EICs in the m/z ranges of 949 - 950 and 933 - 935 are presented for each point mutation with only C19A showing a peak, representing paenithopeptin A with no disulfide bond and with one cysteine changed to alanine. The presence of this peak indicated the position of the disulfide bond as being between Cys10 and Cys19. **b**, The structure of paenithopeptin A_C19A is presented below with marked b ions, y ions, and labionin ring alongside the zoomed mass spectrum of its corresponding doubly charged state. Tandem MS/MS was used to confirm the amino acid sequence and loss of the disulfide bond. The doubly charged precursor ion *m/z* 933.4 was selected for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and amino acid residues deduced from fragment ion analysis are labelled in red.

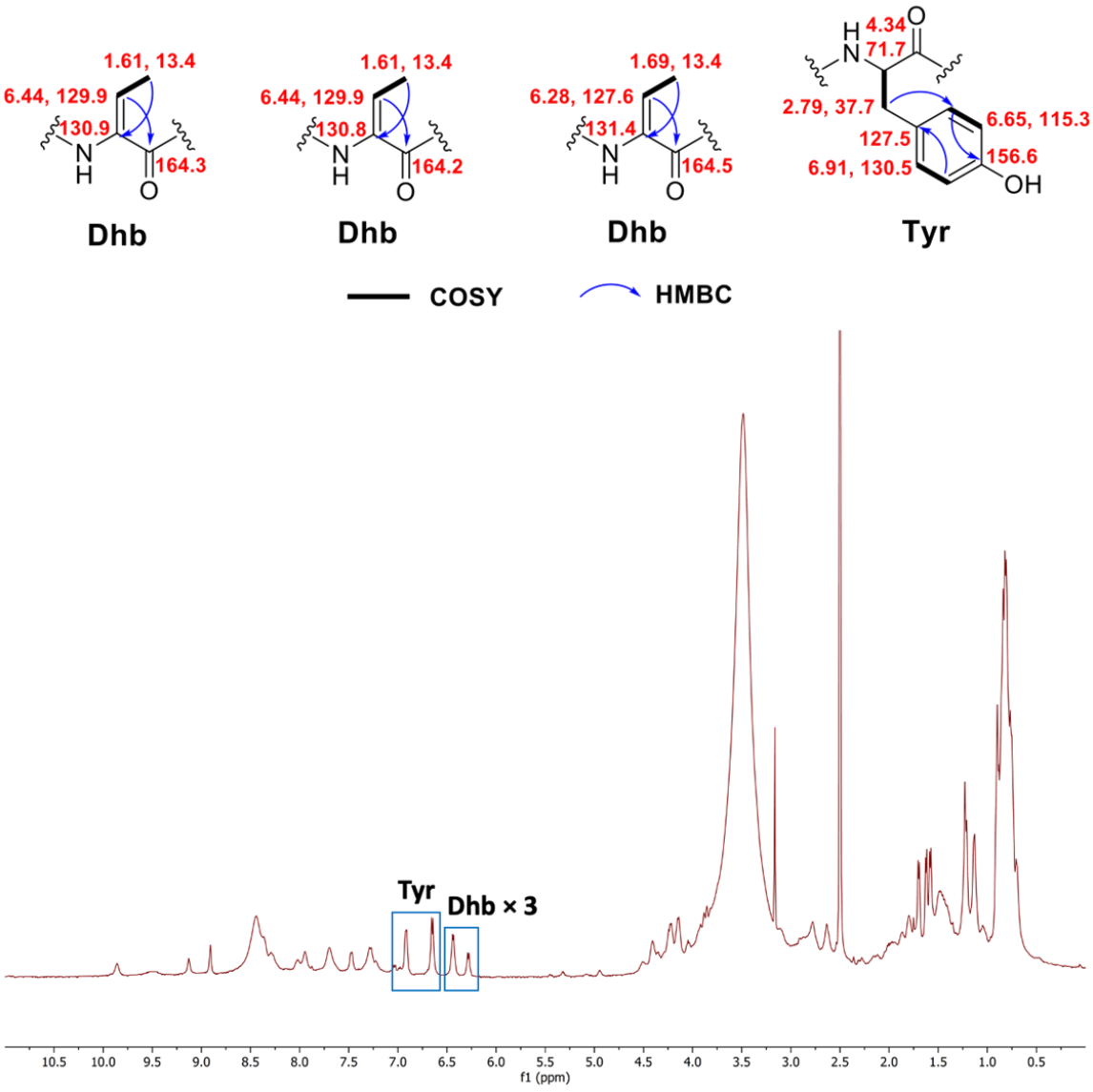


Figure 2.20 ^{1}H-NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

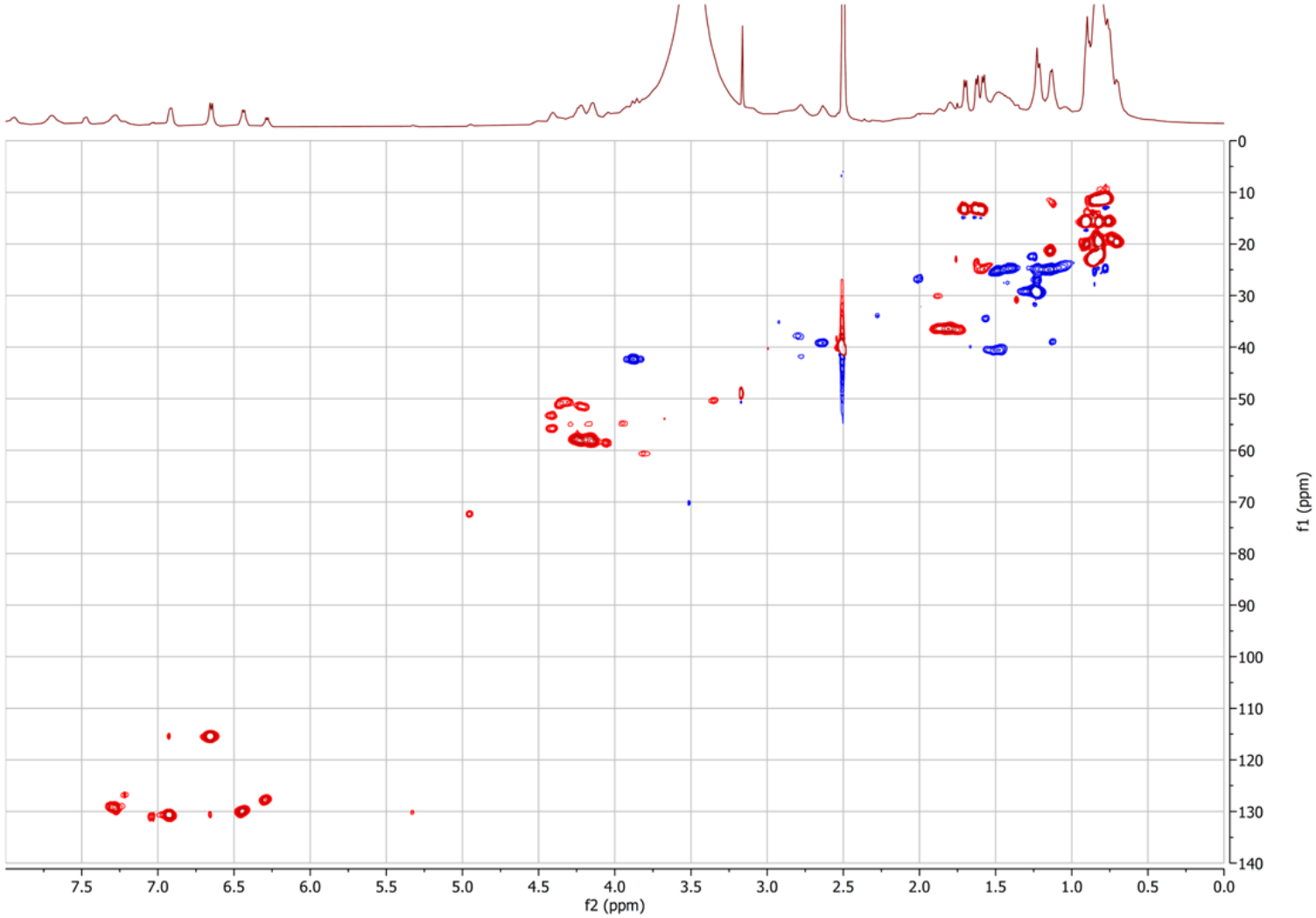


Figure 2.21 HSQC NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

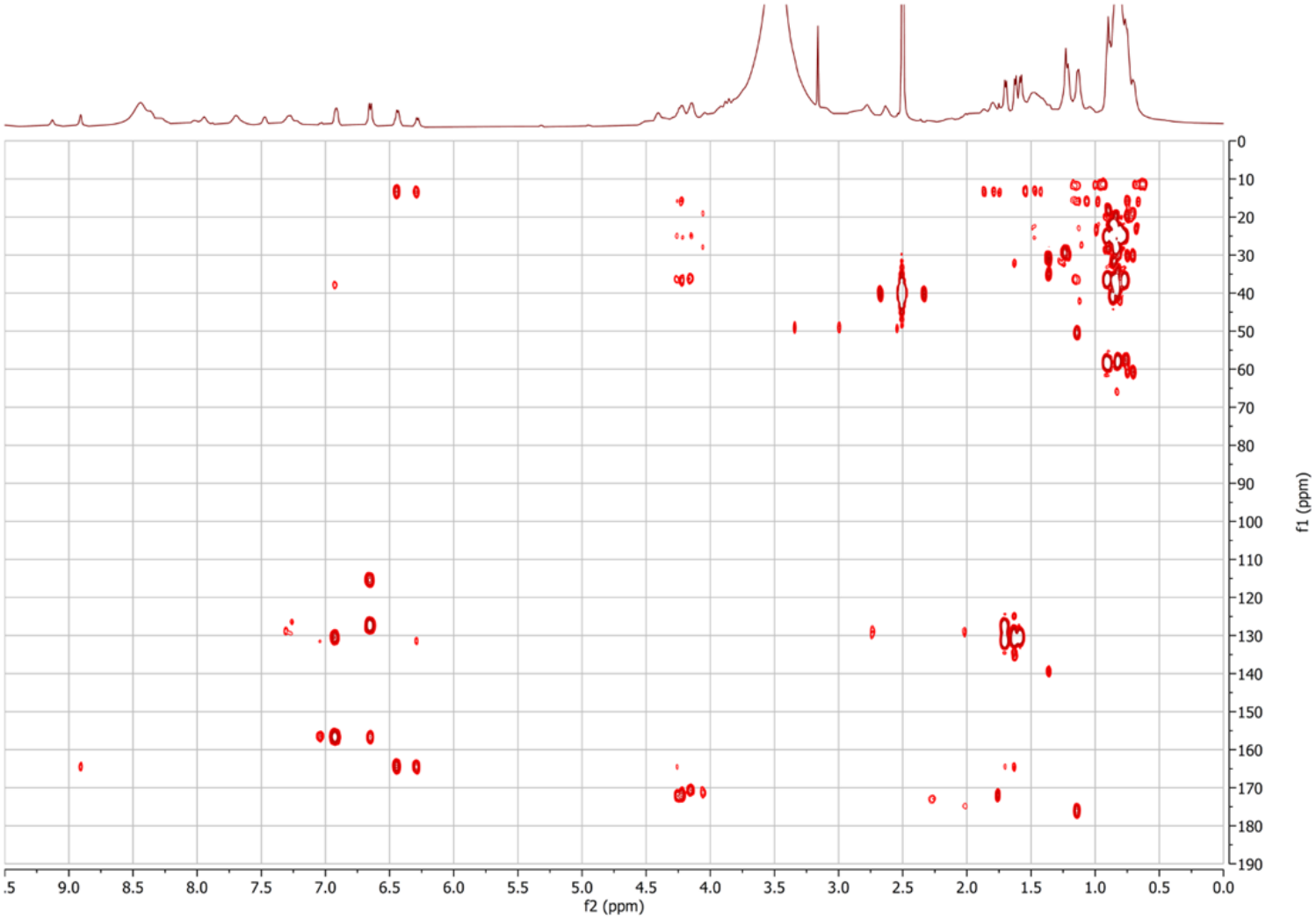


Figure 2.22 HMBC NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

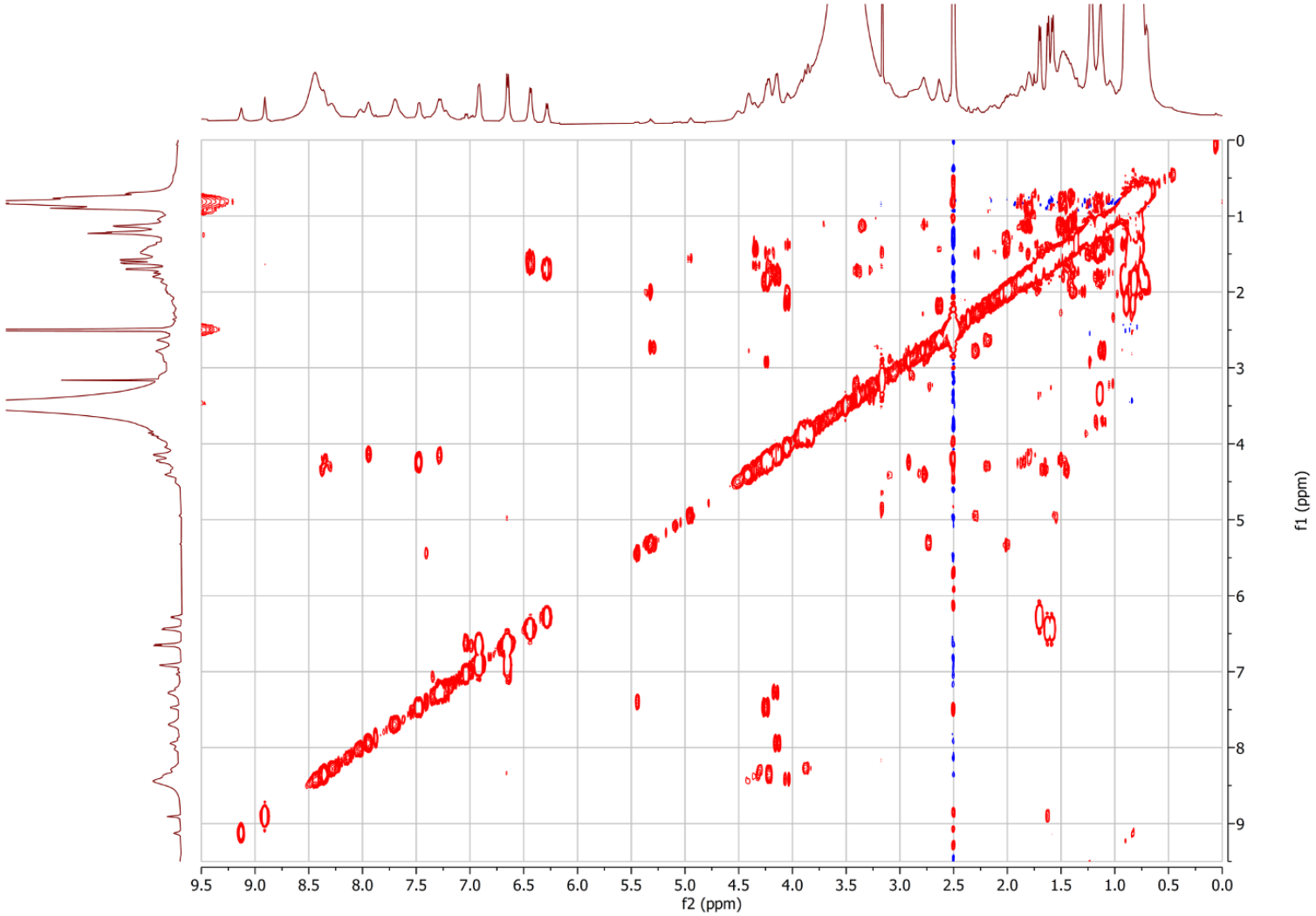


Figure 2.23 ^{1}H-^{1}H COSY NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

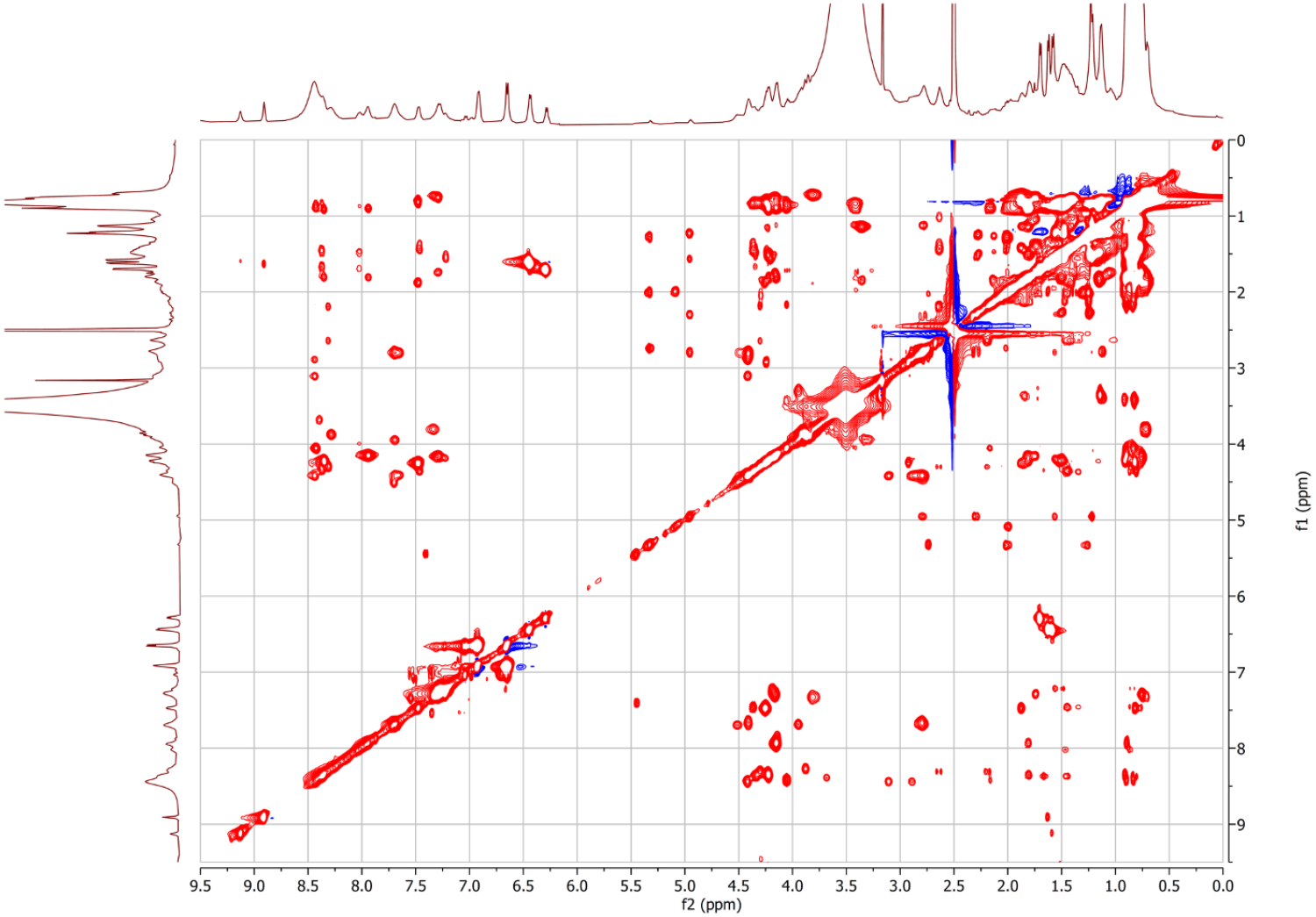


Figure 2.24 ^{1}H-^{1}H TOCSY NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

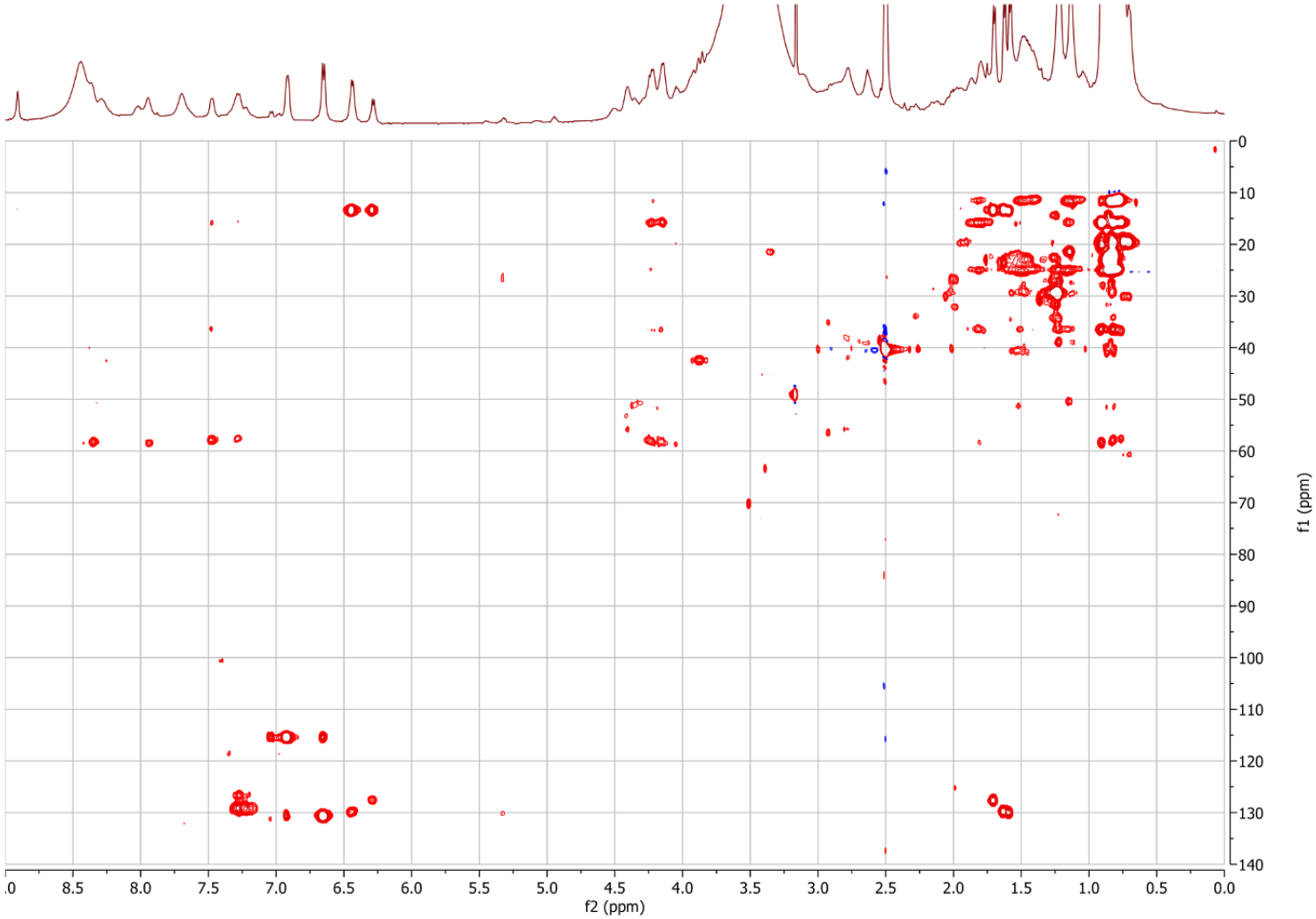


Figure 2.25 HSQC-TOCSY NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

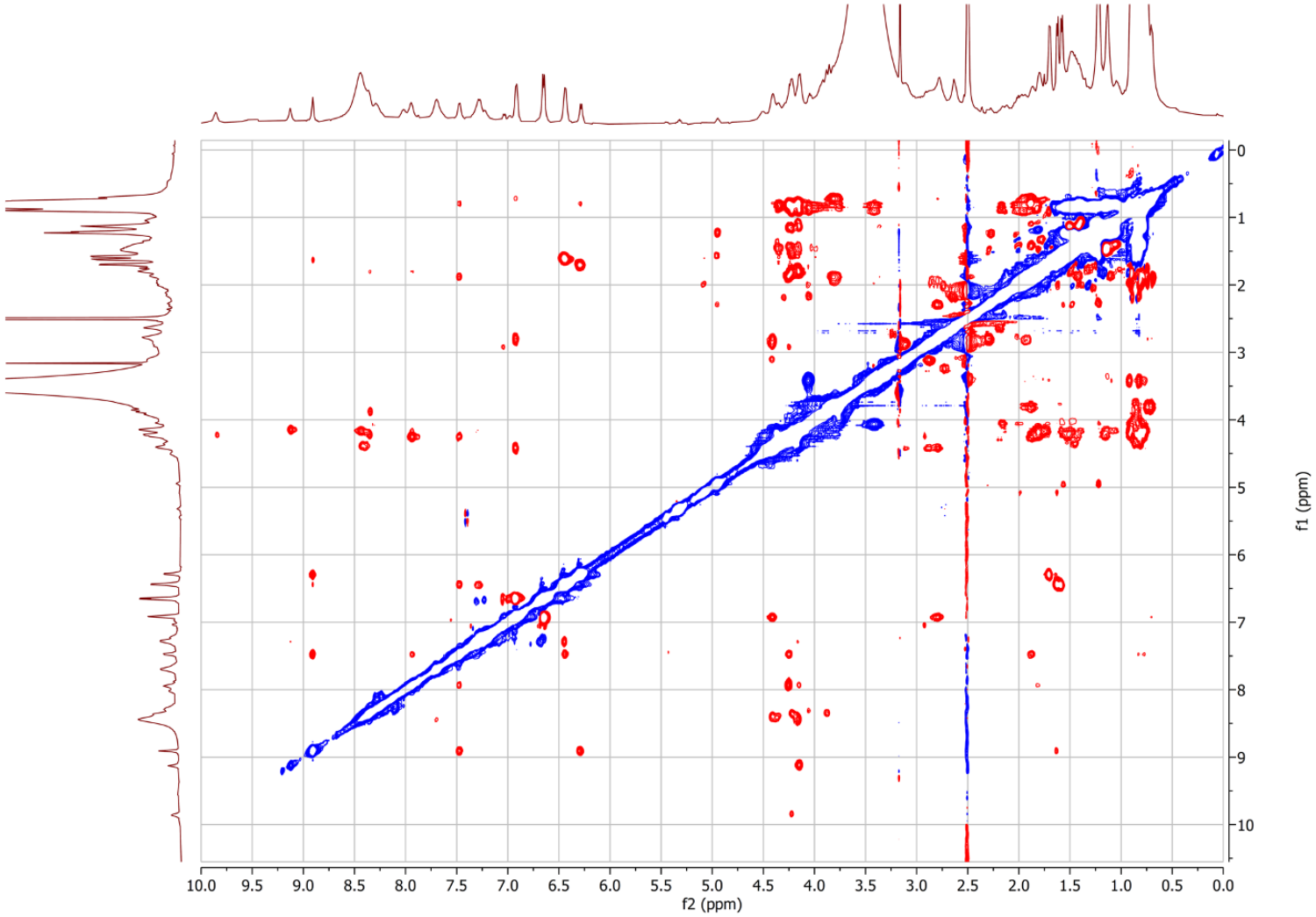


Figure 2.26 ROESY NMR (500 MHz, DMSO-d_6) spectrum of paenithopeptin A.

Table 2.2 Retention times from the Marfey's analysis of paenithopeptin A.

AA	MW (AA-FDAA)	AA standards		Hydrolysate of paenithopeptin A		Absolute configuration
		t_R (L-AA-L-FDAA) or t_R (D-AA-D-FDAA) (min)	t_R (L-AA-D-FDAA) or t_R (D-AA-L-FDAA) (min)	t_R (AA-L-FDAA) (min)	t_R (AA-D-FDAA) (min)	
Ala	341	13.9	15.6	13.8	15.6	L
Ile	383	19.0	21.5	19.0	21.5	L
Val	369	17.1	19.5	17.0	19.4	L
Leu	383	19.5	21.8	19.4	21.8	L
Tyr	433	22.9	24.9	23.0	25.0	L
Lys	398	20.1	21.1	20.2	21.1	L

Using the structure of paenithopeptin A as a guide, paenithopeptins B-E were identified by MS/MS fragmentation analysis (Figures 2.27-2.30) and were found to have different overhangs of amino acids at the N-terminus compared to paenithopeptin A. The chemical structures of paenithopeptins A-E suggest that they were all derived from the precursor PttA1. Notably, paenithopeptins A-E and bacinapeptins A and B represent previously unknown class III lanthipeptides. The tricyclic ring system (labionin in a larger disulfide-bridged ring) in paenithopeptins A-E has not been previously reported in Firmicutes.

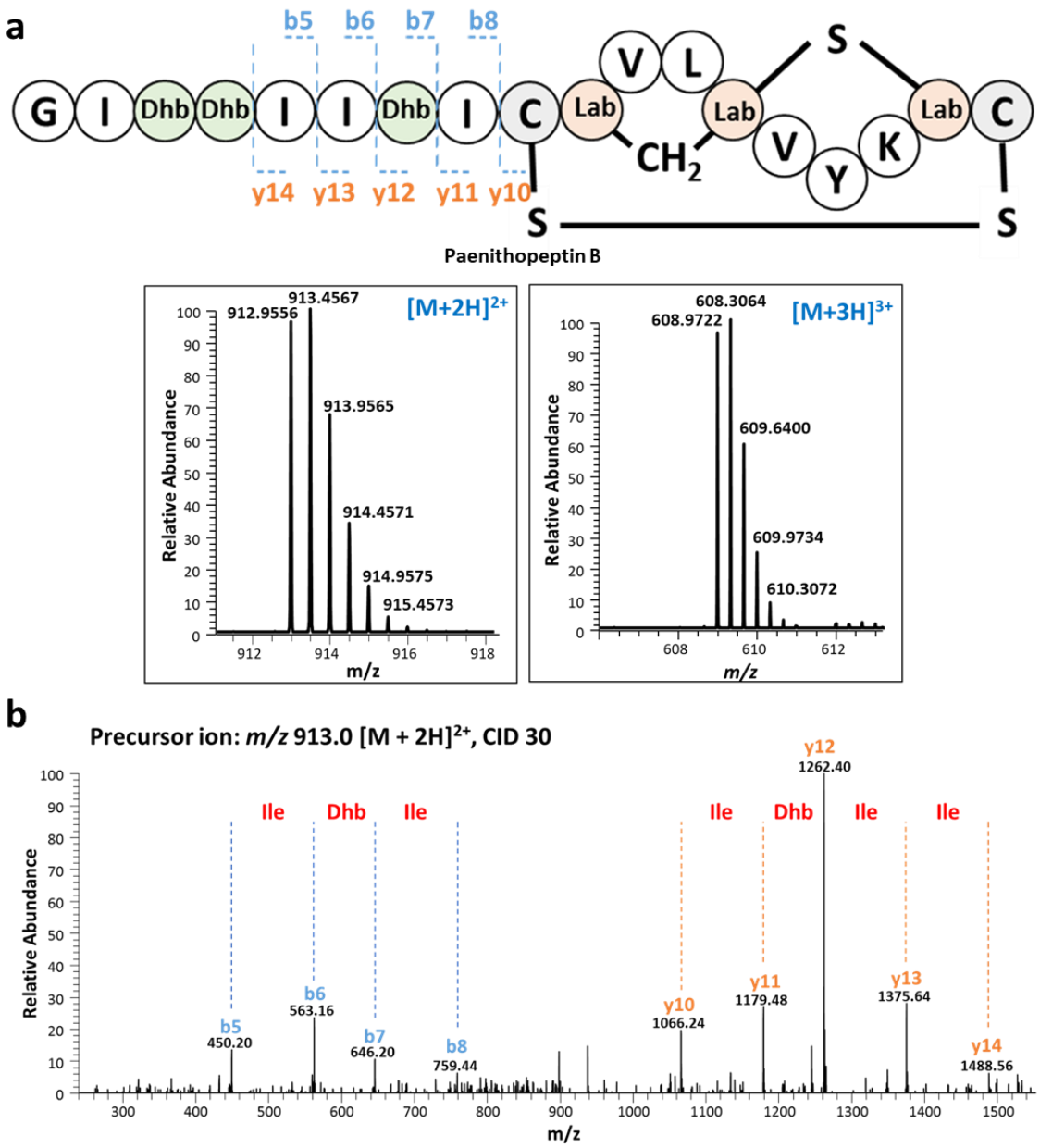


Figure 2.27 Structure elucidation of paenithopeptin B by MS/MS fragmentation analysis. **a,** The structure of paenithopeptin B with fragmentation points of corresponding b and y ions marked as well as the location of the labionin ring and disulfide bond. **b,** High resolution mass spectra representing the doubly and triply charged states of paenithopeptin B. **c,** The doubly charged precursor ion, *m/z* 913.0, was selected for CID at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red.

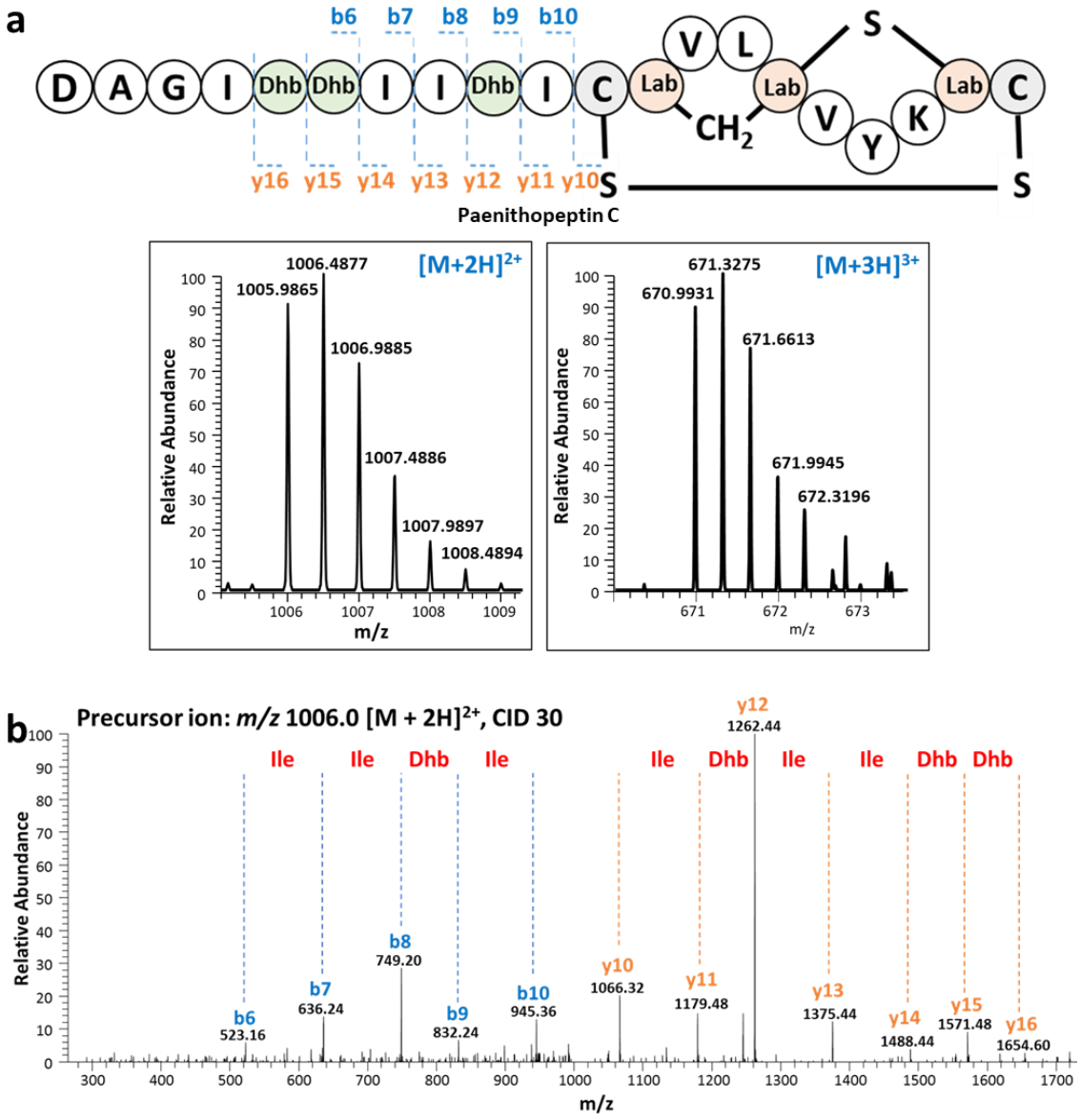


Figure 2.28 Structure elucidation of paenithopeptin C by MS/MS fragmentation analysis. **a,** The structure of paenithopeptin C with fragmentation points of corresponding b and y ions marked as well as the location of the labionin ring and disulfide bond. **b,** High resolution mass spectra representing the doubly and triply charged states of paenithopeptin C. **c,** The doubly charged precursor ion, *m/z* 1006.0, was selected for CID at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red.

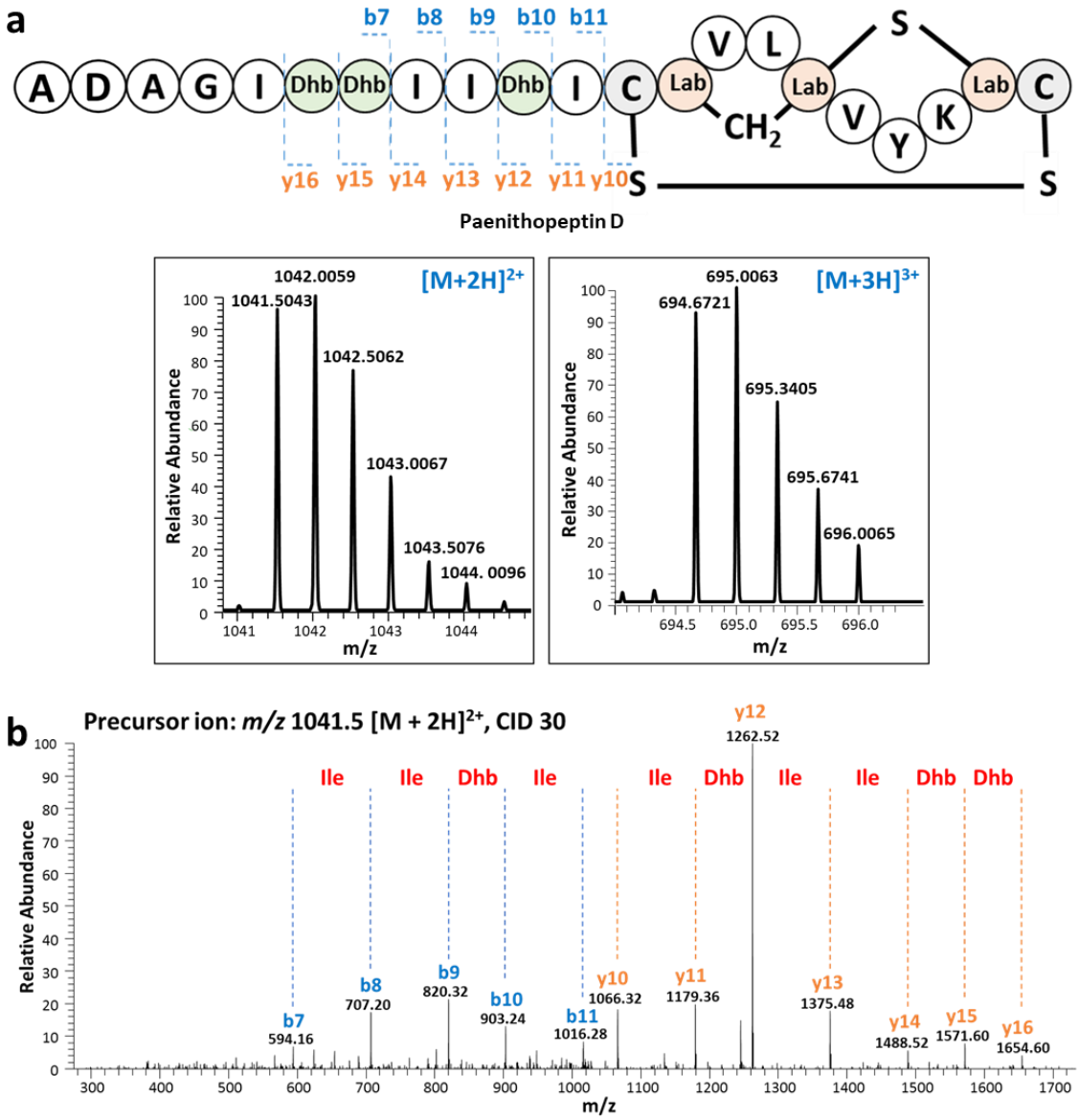


Figure 2.29 Structure elucidation of paenithopeptin D by MS/MS fragmentation analysis. **a,** The structure of paenithopeptin D with fragmentation points of corresponding b and y ions marked as well as the location of the labionin ring and disulfide bond. **b,** High resolution mass spectra representing the doubly and triply charged states of paenithopeptin D. **c,** The doubly charged precursor ion, *m/z* 1041.5, was selected for CID at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red.

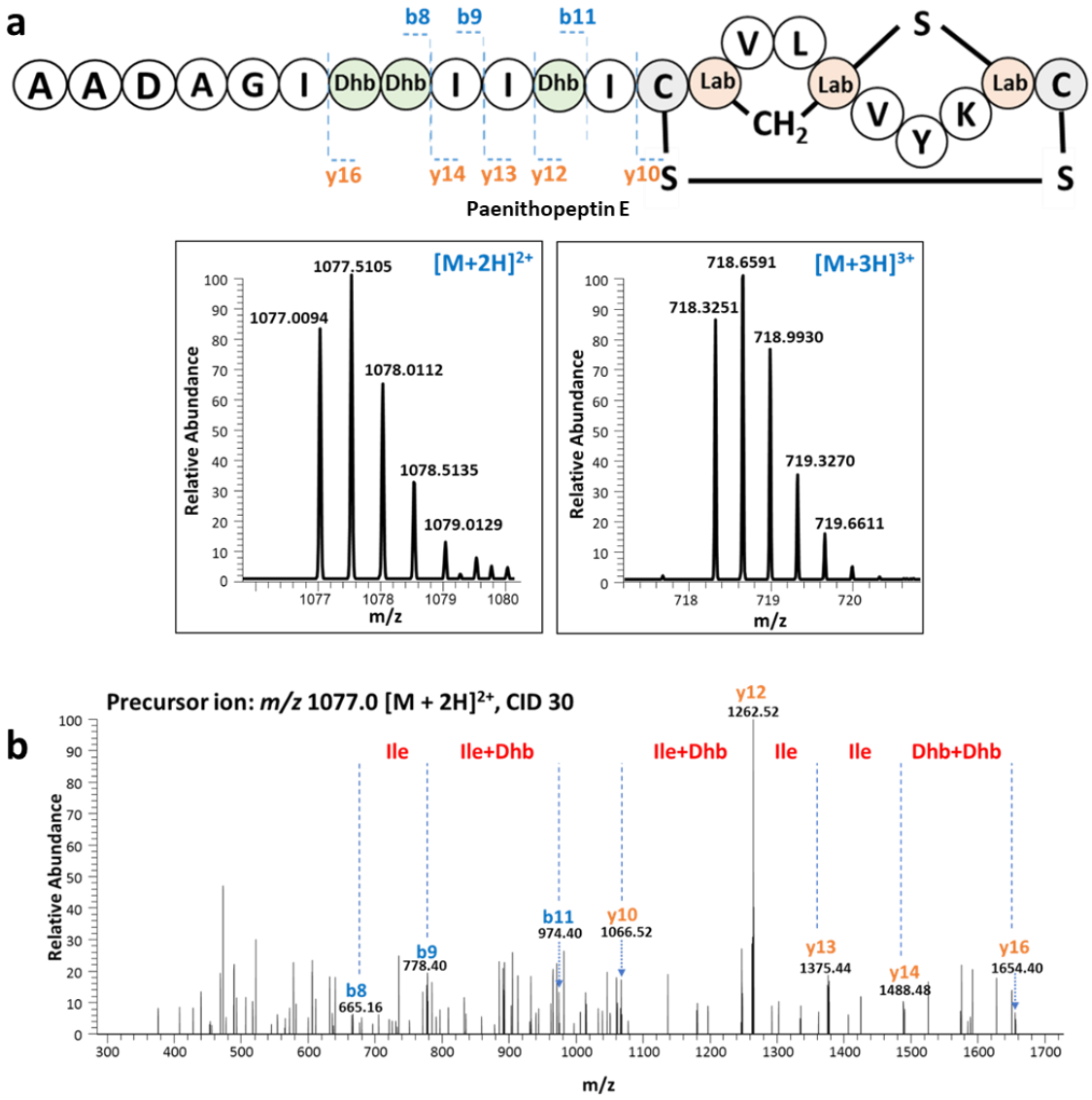


Figure 2.30 Structure elucidation of paenithopeptin E by MS/MS fragmentation analysis. **a,** The structure of paenithopeptin E with fragmentation points of corresponding b and y ions marked as well as the location of the labionin ring and disulfide bond. **b,** High resolution mass spectra representing the doubly and triply charged states of paenithopeptin E. **c,** The doubly charged precursor ion, *m/z* 1077.0, was selected for CID at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red.

2.3.3 Characterization of M16B metallopeptidases as new class III lanthipeptide proteases

Next, we used the pair of Bcn-gP1 and Bcn-gP2 and the pair of PttP1 and PttP2 as examples to confirm the proteolytic activity of Prot_819 and Prot_176 against Pre_24 as the precursor peptide. We selected these two pairs of proteases because: (i) they both belong to Prot_819 and Prot_176, and (ii) while Bcn-gP1 and Bcn-gP2 are encoded by protease genes outside of the *bcn* BGC, PttP1 and PttP2 are encoded by protease genes within another BGC *ptt*, representing both scenarios of protease gene distribution, outside or inside of BGCs. We began the investigation with bioinformatics analysis of Bcn-gP1 and Bcn-gP2, which suggested that both belong to the family of zinc-dependent M16 peptidases. Specifically, Bcn-gP2 possesses an HXXEH motif essential for Zn^{2+} binding and catalytic activity, while Bcn-gP1 contains an R/Y pair in the C-terminal domain to facilitate substrate binding (Figure 2.31a). The characteristic sequences of Bcn-gP1 and Bcn-gP2 are highly reminiscent of two known M16B peptidases from *Sphingomonas* sp. A1, Sph2681 and Sph2682, that form a heterodimer[104,105]. To investigate this, we performed homology modeling of Bcn-gP1 and Bcn-gP2, using the heterodimeric crystal structure of Sph2681/Sph2682 (PDB Accession: 3amj) as the template. The best model of the Bcn-gP1 and Bcn-gP2 heterodimer (hereafter called Bcn-gP1/Bcn-gP2) displayed a high degree of structural similarity toward the template structure (Figure 2.32). Likewise, bioinformatics analysis of PttP1 and PttP2 showed consistent results (Figure 2.33). We further performed pull-down assays and confirmed the respective protein-protein interaction between Bcn-gP1 and Bcn-gP2 (Figure 2.31b) as well as PttP1 and PttP2

(Figure 2.34). Taken together, these results suggested that Bcn-gP1/Bcn-gP2 and PttP1/PttP2 each functions as a heteromeric M16B metallopeptidase.

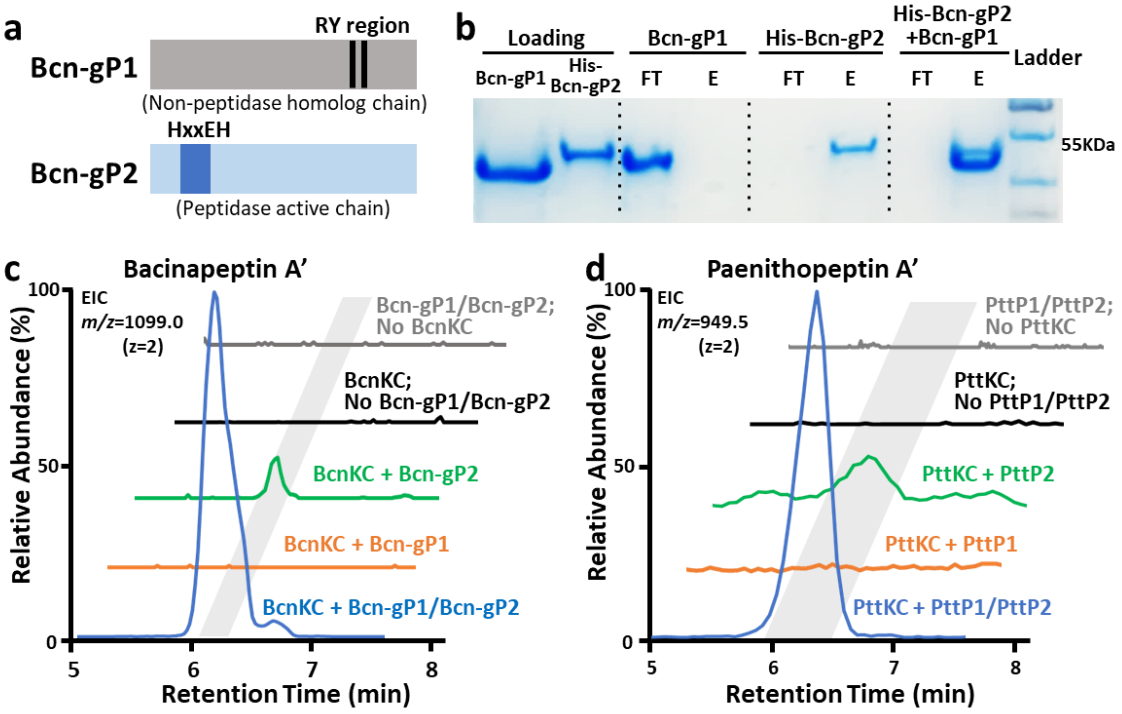


Figure 2.31 Bioinformatic analysis and enzymatic characterization of Bcn-gP1/Bcn-gP2 and PttP1/PttP2. **a,** Representation of the Zn^{2+}-binding HXXEH motif of Bcn-gP2 and PttP2 as well as the substrate-binding R/Y region of Bcn-gP1 and PttP1. These characteristic features suggested Bcn-gP1/Bcn-gP2 and PttP1/PttP2 as heteromeric M16B metallopeptidases. **b**, A pull-down assay showing protein-protein interaction between Bcn-gP1 and Bcn-gP2. His_8-tag-free Bcn-gP1 was immobilized on the nickel affinity column only when His_8-tagged Bcn-gP2 was present, as reflected by examining the flowthrough using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The experiment was repeated three times independently with similar results. FT: flowthrough; E: elute. **c**, EIC showing *in vitro* characterization of the proteolytic activity of Bcn-gP1/Bcn-gP2. Bcn-gP1 and Bcn-gP2 were required simultaneously to produce the highest yield of bacinapeptin A′. In addition, the lack of BcnKC in the assay completely abolished the production, suggesting that Bcn-gP1/Bcn-gP2 shows specificity towards BcnKC-modified precursor bearing a labionin ring. In the assay without BcnKC, *m/z* 1216.1 (z=2) was also used for EIC detection of potential unmodified core peptide. **d**, EIC showing *in vitro* characterization of the proteolytic activity of PttP1/PttP2. PttP1 and PttP2 were required simultaneously for the highest production yield of paenithopeptin A′ and the lack of PttKC abolished the production. In the assay without PttKC, *m/z* 994.5 (z=2) was also used for EIC detection of potential unmodified core peptide.

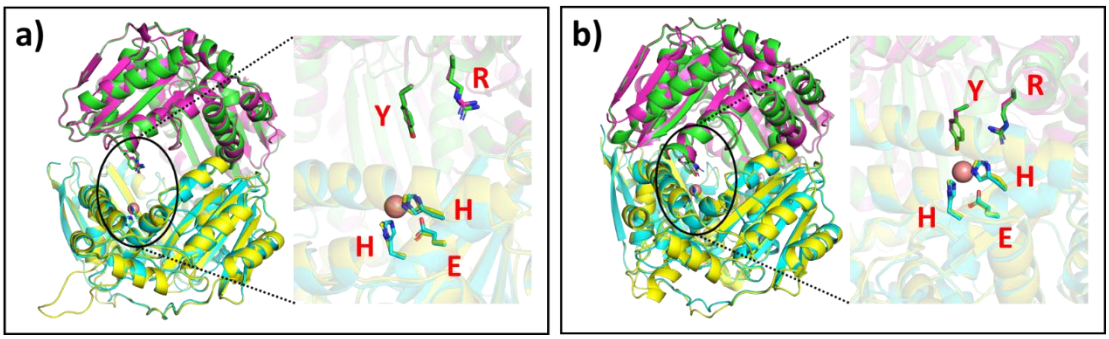


Figure 2.32 Homology modeling of Bcn-gP1 and Bcn-gP2. The sequences of Bcn-gP1 and Bcn-gP2 were used for homology modeling using SWISS-MODEL. Sequences were aligned to the heterodimeric template sequence of Sph2681/Sph2682 (PDB Accession: 3amj). Bcn-gP1 and Bcn-gP2 shared 28% and 30% sequence similarity with Sph2682 and Sph2681, respectively. **a**, Alignment of the heterodimeric model to the template structure in the open conformation revealed 0.43 RMSD for Bcn-gP2(yellow)/Sph2681(cyan) and 9.69 RMSD for Bcn-gP1(green)/Sph2682(pink). **b**, In the closed conformation, alignment revealed 0.40 RMSD for Bcn-gP2(yellow)/Sph2681(cyan) and 9.88 RMSD for Bcn-gP1(green)/Sph2682(pink). In both representations, inset zoomed views indicate residues of HxxEH motif and R/Y pair displayed in sticks and highlighted in red.

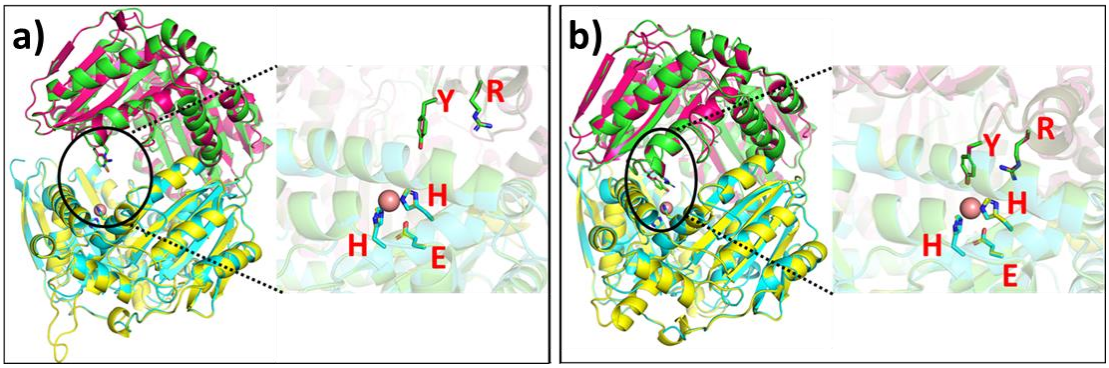


Figure 2.33 Homology modeling of PttP1 and PttP2. The sequences of PttP1 and PttP2 were used for homology modeling using SWISS-MODEL. Sequences were aligned to the heterodimeric template sequence of Sph2681/Sph2682 (PDB code 3amj). PttP1 and PttP2 shared 29% and 30% sequence similarity with Sph2682 and Sph2681, respectively. **a**, Alignment of the heterodimeric model to the template structure in the open conformation revealed 0.17 RMSD for PttP2(yellow)/Sph2681(cyan) and 0.55 RMSD for PttP1(green)/Sph2682(pink). **b**, In the closed conformation, alignment revealed 0.17 RMSD for PttP2(yellow)/Sph2681(cyan) and 0.49 RMSD for PttP1(green)/Sph2682(pink). In both representations, inset zoomed views indicate residues of HxxEH motif and R/Y pair displayed in sticks and highlighted in red.

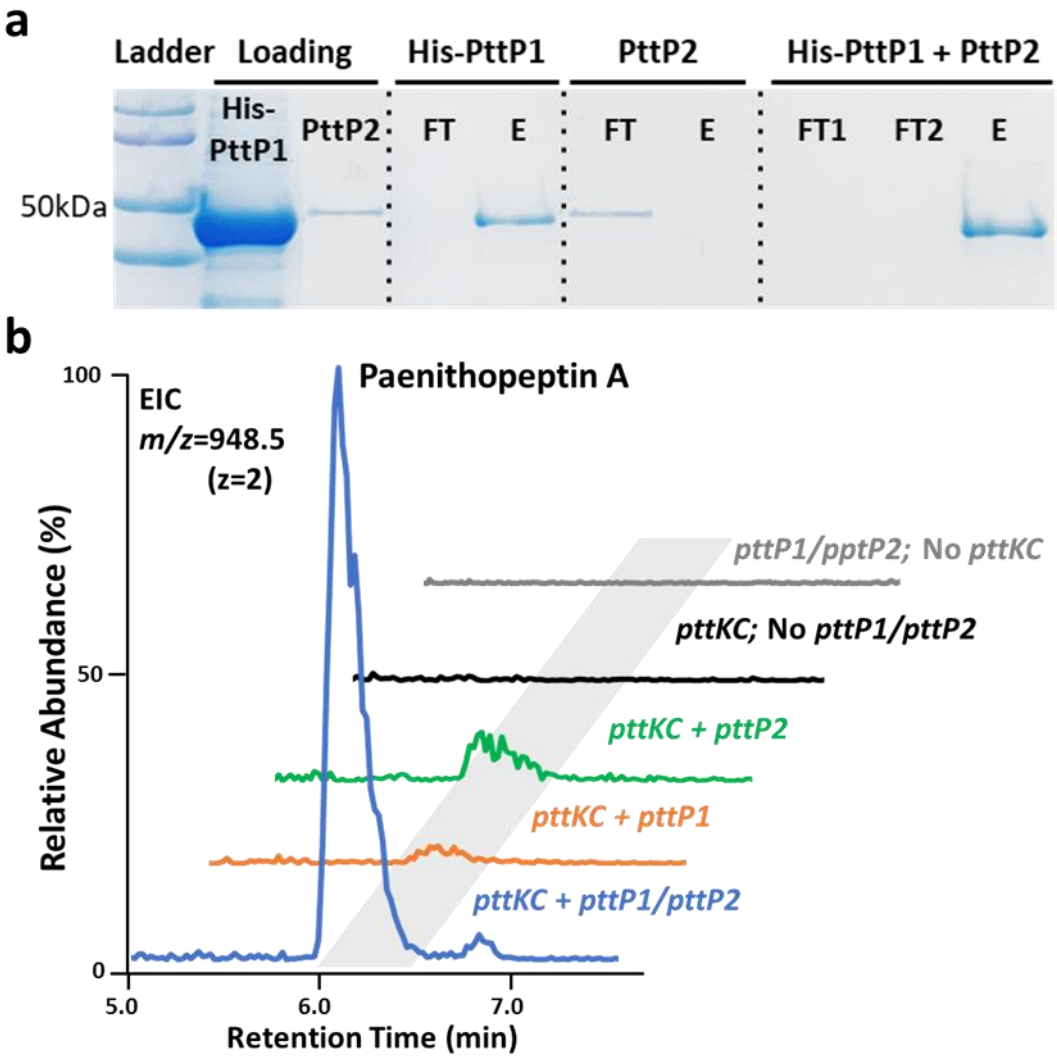


Figure 2.34 Pull-down assay and *in vivo* proteolytic activity of PttP1/PttP2. **a**, A pull-down assay showing protein-protein interaction between PttP1 and PttP2. His_8-tag-free PttP2 was immobilized in the nickel affinity column only when His_8-tagged PttP1 was present, as reflected by examining the flowthrough using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The experiment was repeated three times independently with similar results. FT: flowthrough; E: elute. **b**, Different combinations of four genes, *pttA1*, *pttKC*, *pttP1*, and *pttP2* were constructed into a series of pDR111-based integrative plasmids and expressed in an engineered *B. subtilis* 168 heterologous expression host lacking the native *pttP1/pttP2* homologs *ymfF*/*ymfH*. Extracted ion chromatograms (EIC) of m/z 948.5 ($[M+2H]^{2+}$), representing the doubly charged state of paenithopeptin A, are overlaid for comparison. In blue, *pttKC*, *pttP1*, and *pttP2* were all required for full production. In orange, expression of *pttKC* and *pttP1* resulted in a barely detectable amount of production. In green, *pttKC* and *pttP2* were enough for minor production. In black, lack of *pttP1/pttP2* protease pair resulted in no production. In grey, lack of *pttKC* resulted in no production.

We then performed an *in vitro* characterization of Bcn-gP1/Bcn-gP2 against BcnA1. We individually expressed and purified recombinant His-tagged BcnA1, BcnKC, Bcn-gP1, or Bcn-gP2. We then incubated BcnA1, with or without BcnKC, followed by adding Bcn-gP1, Bcn-gP2, or both, respectively, for an incubation of 12 hours. We observed that (i) while excluding Bcn-gP1 generated a detectable production presumably due to Bcn-gP2 possessing catalytic residues, Bcn-gP1 and Bcn-gP2 were simultaneously required to generate the highest yield of the product, demethylated bacinapeptin A (hereafter called bacinapeptin A′, produced due to not including the predicted methyltransferase BcnMT in the assay); and (ii) the Bcn-gP1/Bcn-gP2 complex has specificity against BcnKC-modified precursor bearing a labionin ring (Figure 2.31c). Likewise, we performed an *in vitro* characterization of PttP1/PttP2 against PttA1, with or without PttKC, using the same assay conditions mentioned above. We observed consistent results (Figure 2.31d) as described for Bcn-gP1/Bcn-gP2. Notably, the products of this *in vitro* assay, designated as paenithopeptins A′-E′, each lacked a disulfide bond compared to paenithopeptins A-E, presumably due to the absence of a disulfide bond-forming oxidoreductase[106,107] that is encoded by paenithopeptin-producing bacteria. We further performed an *in vivo* assay using *pttP1*/*pttP2*, aiming to produce intact paenithopeptins A-E to confirm the function of *pttP1*/*pttP2*. Different combinations of four genes, *pttA1*, *pttKC*, *pttP1*, and *pttP2*, were constructed into a series of pDR111-based integrative plasmids (Supplementary Table 2.4) and expressed in an engineered host *B. subtilis* 168 with *ymfF* and *ymfH* (homologs of *pttP1* and *pttP2*, respectively) deleted in advance. The in vivo assay indeed showed the production of fully modified paenithopeptins (Figure 2.34), in a production trend consistent with the *in vitro* assay. Taken together, these results

indicated Prot_819/Prot_176 as previously unknown lanthipeptide proteases and established their role as a heteromeric complex in the maturation of previously unknown class III lanthipeptides represented by bacinapeptins and paenithopeptins.

2.2.3 Protease efficiency suggests evolved specificity for class III lanthipeptide maturation

Due to the wide distribution of *prot_819*/*prot_176* members in bacterial genomes, we investigated the potential activity of different Prot_819/Prot_176 members for leader removal. The distribution of *prot_819*/*prot_176* pairs were classified into three scenarios for our comparison (Figure 2.8b): (i) outside of Pre_24-encoding class III lanthipeptide BGCs, e.g., *bcn-gP1*/*bcn-gP2* from *B. nakamurai* NRRL B-41092 and *ptt-genomeP1*/*ptt-genomeP2* (abbreviated as *ptt-gP1*/*ptt-gP2*) from *P. thiaminolyticus* NRRL B-4156, (ii) within the Pre_24-encoding class III lanthipeptide BGCs, e.g., *pttP1*/*pttP2* within the *ptt* BGC, and (iii) in the genomes that do not harbor any Pre_24-encoding class III lanthipeptide BGCs, e.g., *ymfF*/*ymfH* from *B. subtilis* 168. We compared the *in vitro* proteolytic activity of these Prot_819/Prot_176 members using the same assay conditions described above. The results show that YmfF/YmfH, Bcn-gP1/Bcn-gP2, and Ptt-gP1/Ptt-gP2 were able to cleave PttKC-modified PttA1 and produce paenithopeptins, but the efficiency, reflected as the production yield under the same conditions, was ~4%, ~13%, and ~15% compared to that of PttP1/PttP2, respectively (Figures 2.35 and 2.36). We further evaluated the activity of these Prot_819/Prot_176 proteases against an additional member of Pre_24, i.e. BcnA1 from *B. nakamurai* NRRL B-41092. As expected, we observed that YmfF/YmfH, Bcn-gP1/Bcn-gP2, Ptt-gP1/Ptt-gP2, and PttP1/PttP2 were all active against

BcnKC-modified BcnA1, with an efficiency trend (Figures 2.35 and 2.37) similar to that against PttKC-modified PttA1.

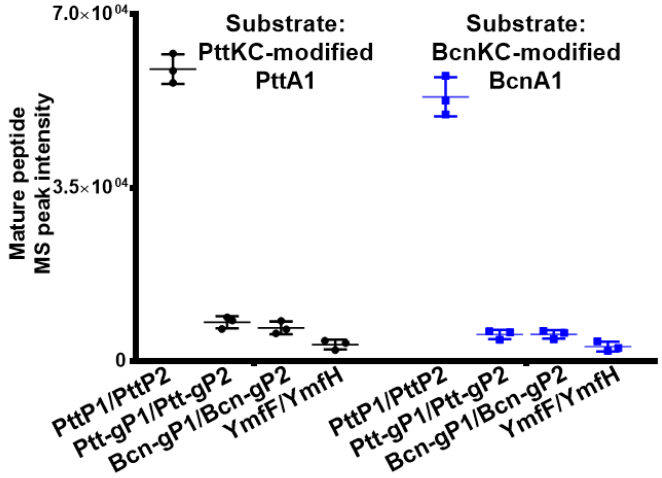


Figure 2.35 Efficiencies of Prot_819/Prot_176 members. PttP1/PttP2, Ptt-gP1/Ptt-gP2, Bcn-gP1/Bcn-gP2, and YmfF/YmfH against PttKC-modified PttA1 and BcnKC-modified BcnA1, respectively, with PttP1/PttP2 showing the highest efficiency in both cases. Error bars indicate standard deviation across triplicates. Data shown as mean ± SD (each group n = 3; SD: standard deviation).

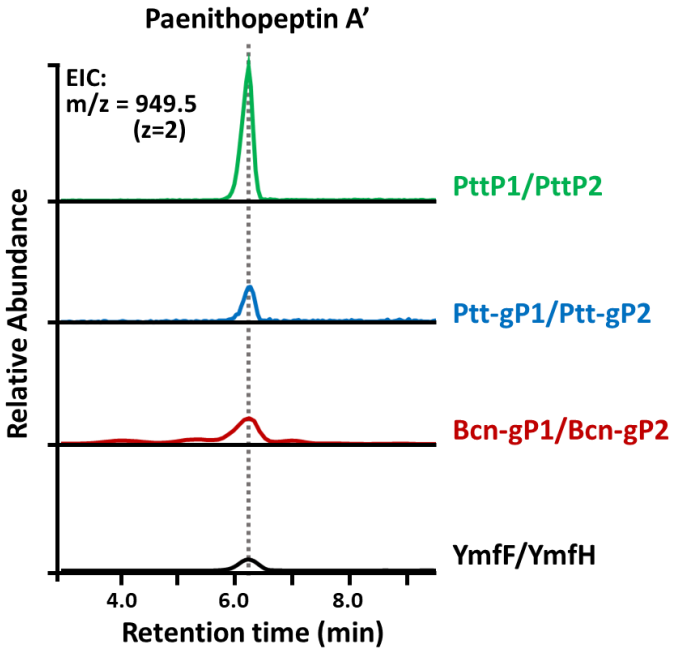


Figure 2.36 Efficiencies of Prot_819/ Prot_176 proteases against PttKC-modified PttA1. Proteolytic efficiency of Prot_819/ Prot_176 proteases was compared using *in vitro* proteolysis of PttKC-modified PttA1. EIC of m/z 949.5 ($[M+2H]^{2+}$), representing the doubly charged state of paenithopeptin A′, are overlaid for comparison. In green, production of paenithopeptin A′ with native PttP1/PttP2 proteases. In blue, Ptt-gP1/Ptt-gP2 found in the genome of *P. thiaminolyticus* NRRL B-4156, exhibited proteolysis with less efficiency compared to PttP1/PttP2. In red, Bcn-gP1/Bcn-gP2 detected in *B. nakamurai* NRRL B-41092, containing another Pre_24-encoding lanthipeptide BGC, demonstrated proteolytic activity towards PttA1 with less efficiency than native PttP1/PttP2. In black, YmfF/YmfH found in the *B. subtilis* 168 heterologous host, showed some proteolytic activity towards PttA1.

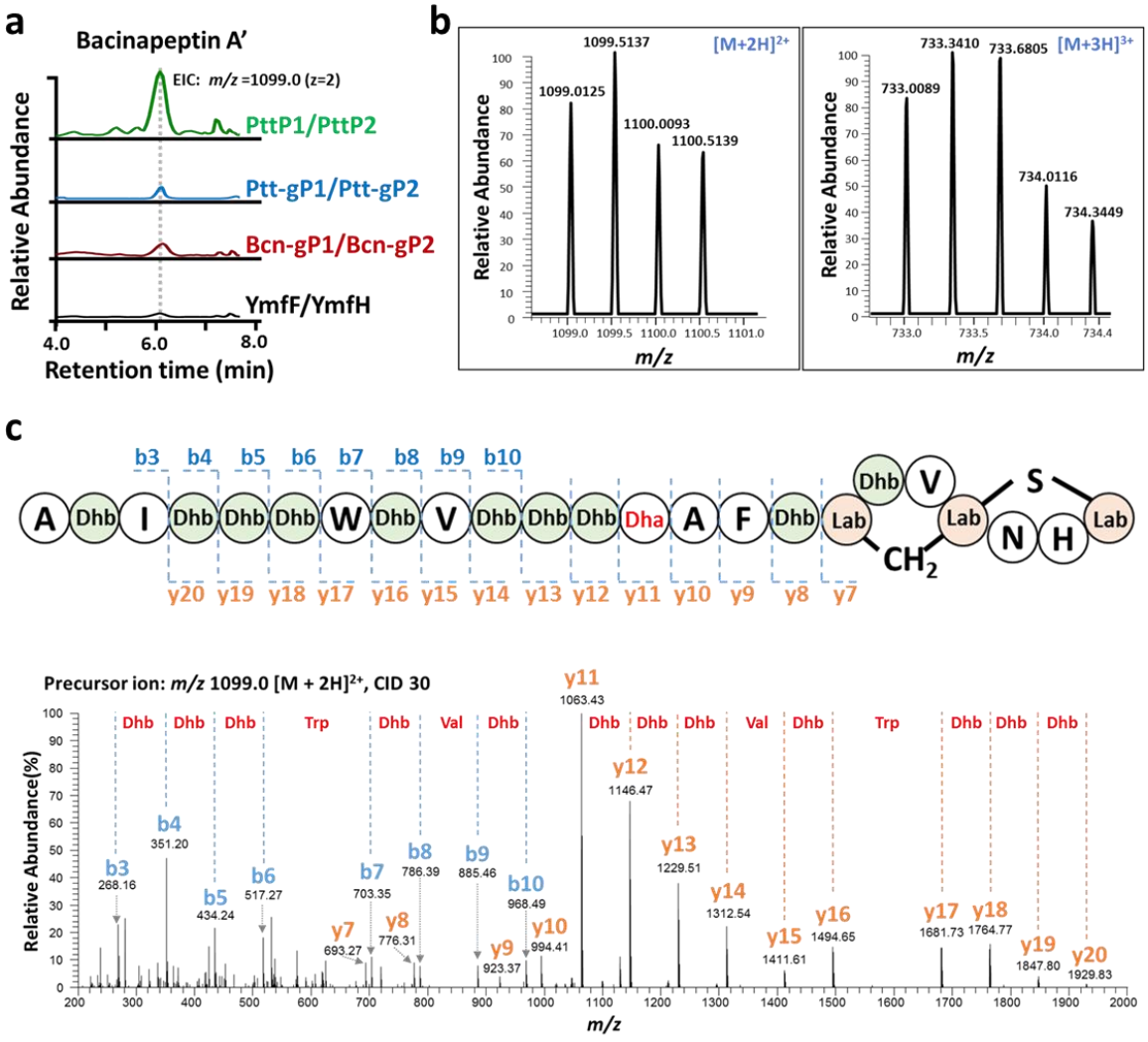


Figure 2.37 Efficiencies of Prot_819/Prot_176 proteases against BcnKC-modified BcnA1. Proteolytic efficiency of Prot_819/Prot_176 proteases was compared using *in vitro* proteolysis of BcnKC-modified BcnA1. **a**, EICs of m/z 1099.0 ($[M+2H]^{2+}$), representing the doubly charged state of bacinapeptin A′, are overlaid for comparison. In green, production of bacinapeptin A′ with PttP1/PttP2 proteases from the *ptt* BGC. In blue, Ptt-gP1/Ptt-gP2 exhibited proteolysis with less efficiency compared to PttP1/PttP2. In red, Bcn-gP1/Bcn-gP2 detected in *B. nakamurai* NRRL B-41092 demonstrated proteolytic activity towards BcnA1 with less efficiency than PttP1/PttP2. In black, YmfF/YmfH found in the *B. subtilis* 168 heterologous host, showed some proteolytic activity towards BcnA1. **b**, The high-resolution mass spectra of bacinapeptin A′ are included at right showing its doubly and triply charged states. **c**, The structure of bacinapeptin A′ with b or y ion fragment positions marked. The doubly charged precursor ion, *m/z* 1099.0, was selected for collision induced dissociation (CID) at 30 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red.

On the other hand, we tested whether the Prot_819/Prot_176 members could act on precursors from groups other than Pre_24. The precursor selected from a class III lanthipeptide BGC of *Paenibacillus taiwanensis* DSM 18679 was designated as PbtA (belonging to Pre_4085) (Figure 2.8b). We expressed PbtA and Pbt-genomeP1/Pbt-genomeP2 (abbreviation Pbt-gP1/Pbt-gP2) as recombinant proteins for *in vitro* proteolytic assays. Despite a variety of assay conditions used, we did not observe any obvious maturation products when PbtA, with or without PbtKC, was incubated with Pbt-gP1/Pbt-gP2, YmfF/YmfH, Bcn-gP1/Bcn-gP2, Ptt-gP1/Ptt-gP2, or PttP1/PttP2.

Taken together, the specificity between Pre_24 and Prot_819/Prot_176 supports our precursor-protease correlation network. Furthermore, considering the wide distribution of Prot_819/Prot_176, the different efficiencies of Prot_819/Prot_176 members appeared supportive of our hypothesis that certain proteases with general functions in the genome might have evolved more or less specific activity against class III lanthipeptides, and extra copies of them might have further evolved into pathway-specific proteases for enhanced activity and specificity. We therefore performed a phylogenetic analysis of Prot_176 members, the catalytic component of the Prot_819/Prot_176 pair. A phylogenetic tree was built at the level of genus *Paenibacillus* where paenithopeptins were isolated. Indeed, pathway-specific Prot_176 appeared in closely related lineages in the phylogenetic tree, forming an independent branch from other Prot_176 members encoded by genes outside of the BGCs, implying a process of gene divergence (Figure 2.38). Thus, the substrate specificity of Prot_819/Prot_176 likely gained during evolution provides evidence for our correlational networking that uses substrate specificity to look for hidden proteases.

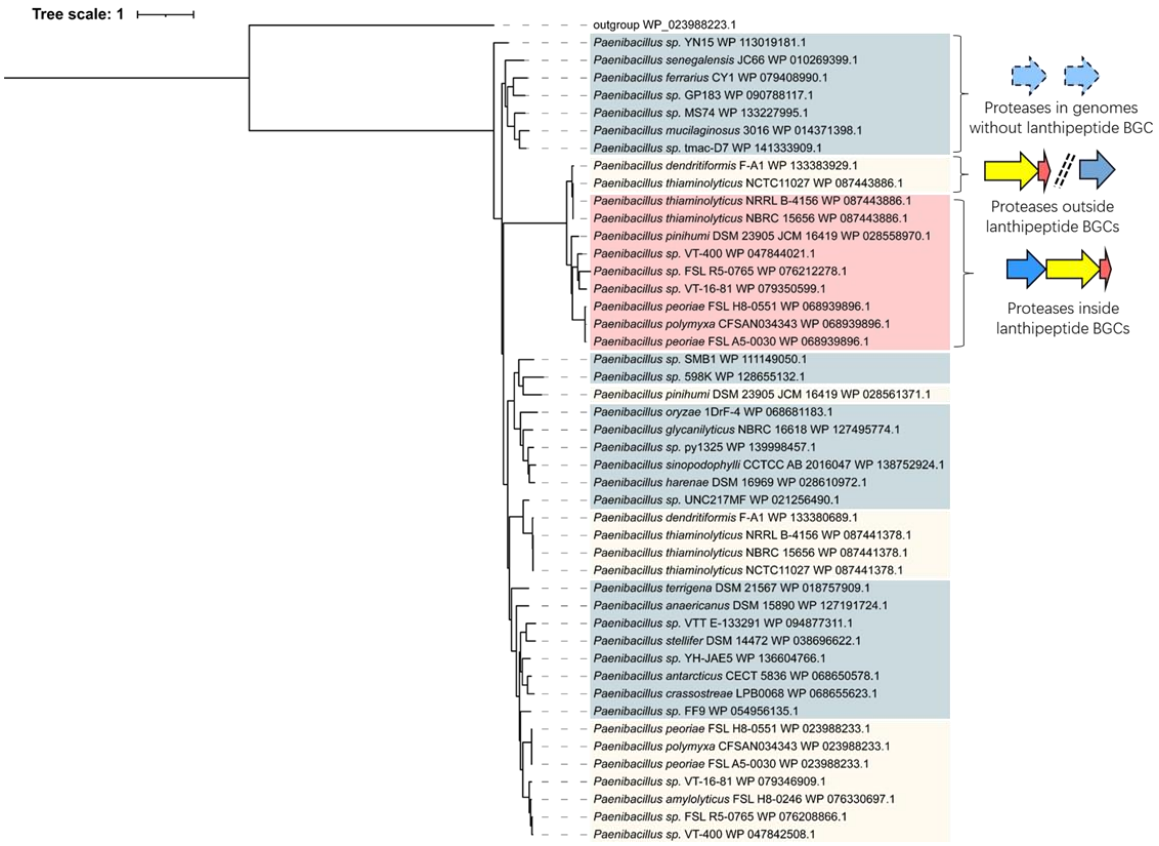


Figure 2.38 Phylogenetic tree of Prot_176 in *Paenibacillus*. A phylogenetic tree was constructed using Prot_176 sequences in *Paenibacillus*. An outgroup WP_023988223.1, representing an M16B metallopeptidase found in *Paenibacillus* that does not belong to the Prot_176 group, was added and set as the root point for the tree. Proteases were categorized into three groups based on their location in their respective genome as indicated at right: Proteases in genomes without a lanthipeptide BGC (light blue), proteases in genomes harboring lanthipeptide BGCs but with the protease encoded outside of the lanthipeptide BGC (light yellow), and proteases encoded within lanthipeptide BGCs (red).

2.2.4 Discovered proteases show specificity and unique activity for leader peptide removal

Due to the highest activity of PttP1/PttP2 shown above, we further characterized this protease pair in detail as a representative member of Prot_819/Prot_176 that are widely distributed outside of many class III lanthipeptide BGCs. First, we confirmed the importance of the Zn-binding HXXEH motif and the R/Y pair in the proteolytic activity of PttP1/PttP2 against PttKC-modified PttA1 precursor. We individually mutated H67, E70,

and H71 of the HXXEH motif and R298 and Y305 of the R/Y pair to Ala, leading to decreased activity of PttP1/PttP2 *in vitro* for each mutation (Figure 2.39). Simultaneous mutation of all these five residues into Ala residues completely abolished the production of paenithopeptins A′-E′ (Figure 2.39). In addition, the metal-chelating compound, *o*-phenanthroline, significantly inhibited the activity of PttP1/PttP2 (Figure 2.40).

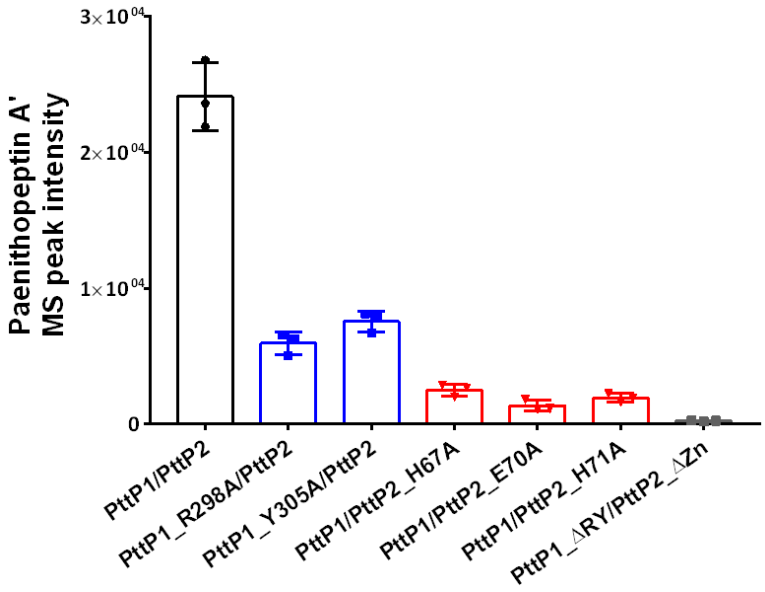


Figure 2.39 *In vitro* activity of PttP1/PttP2 mutations. The H67, E70, and H71 of the HXXEH motif and R298 and Y305 of the R/Y pair were mutated individually to Ala, leading to decreased activity of PttP1/PttP2 *in vitro* for each mutation. Simultaneous mutation of all five residues to Ala completely abolished the production of paenithopeptin A′. The first column (black) indicates the production of paenithopeptin A′ using wild type PttP1/PttP2. The next two columns (blue) represent the production with varying mutations to the R/Y pair of PttP1. The next three columns (red) indicate the production with mutations to the HXXEH motif of PttP2. The last column (grey, obscured by relative height) represents lost production with mutations to both conserved motifs of PttP1 and PttP2. Production of paenithopeptins B′-E′ showed the same trend. Data shown as mean ± SD (each group n = 3; SD: standard deviation).

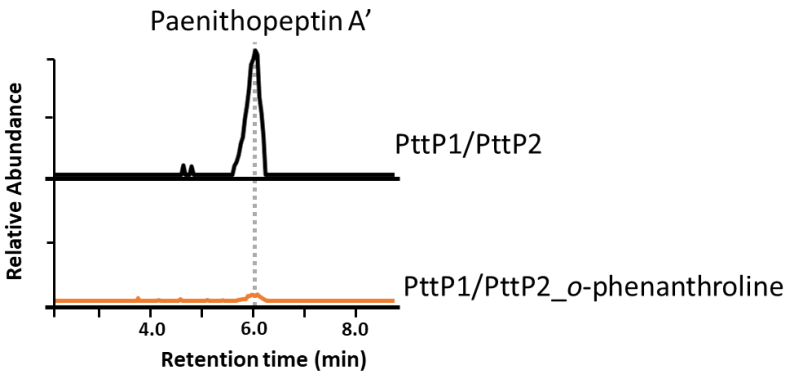


Figure 2.40 PttP1/PttP2 activity is metal dependent. Addition of the metal-chelating compound, *o*-phenanthroline, to an *in vitro* assay containing PttKC modified PttA1 resulted in significant inhibition of the proteolytic activity of PttP1/PttP2. Extracted ion chromatograms (EIC) of m/z 949.5 ($[M+2H]^{2+}$), representing the doubly charged state of paenithopeptin A′, are overlaid for comparison. In black, PttP1/PttP2 processes PttKC modified PttA1 to produce paenithopeptin A′. In orange, PttP2 was incubated with *o*-phenanthroline prior to addition to the *in vitro* assay, significantly reducing the production of paenithopeptin A′.

Next, we investigated how PttP1/PttP2 led to production of a series of compounds with different N-terminal overhangs from the same precursor peptide, PttA1 (Figure 2.41a). We first performed an *in vitro* enzymatic assay with an incubation of PttA1 and PttKC for four hours, followed by adding both PttP1 and PttP2 for different incubation lengths, ranging from 15 minutes to 36 hours. We observed that at the 15 minute-point, paenithipeptin E′ with the longest leader overhang was formed as the major product (Figure 2.41b), suggesting an endopeptidase activity of PttP1/PttP2. Over time, paenithipeptin E′ diminished while paenithipeptin A′ (with no leader overhang) was accumulated (Figure 2.41b), followed by removal of the N-terminal alanine of paenithipeptin A′ to form paenithipeptin B′ after an extended incubation of 36 hours (Figure 2.41b). This transformation process suggested an aminopeptidase activity of PttP1/PttP2. Next, we directly used paenithopeptin E′ as a substrate for incubation with PttP1/PttP2, leading to

the production of paenithipeptin A′ (Figure 2.41c). We also incubated paenithopeptin A with PttP1/PttP2, resulting in the generation of paenithopeptin B (Figure 2.41d). These experiments confirmed the aminopeptidase activity of PttP1 and PttP2. Taken together, our results show that the QAAD motif in the leader of PttA1 appeared to be specifically cleaved by PttP1/PttP2, through an initial endopeptidase activity to partially remove the leader followed by an aminopeptidase activity to successively remove the remaining overhangs. We then investigated whether PttP1 and PttP2 have proteolytic activity against PttA2-PttA7. Although *pttA2-pttA7* are located separately from *pttA1* in the *ptt* BGC (Figure 2.8c), they also belong to the *pre_24* group in our network. As expected, *in vitro* enzymatic assays using conditions described above revealed corresponding products cleaved from the conserved Q-A-(A/V)-(D/E) motif of PttA2-PttA7 (Figures 2.42-2.45). The products also showed different overhangs of amino acids at the N-terminus (Figures 2.46-2.49), suggesting again both endo- and aminopeptidase activities of PttP1/PttP2. In addition, formation of these products was also dependent on the precursor modification by PttKC (Figures 2.42-2.45).

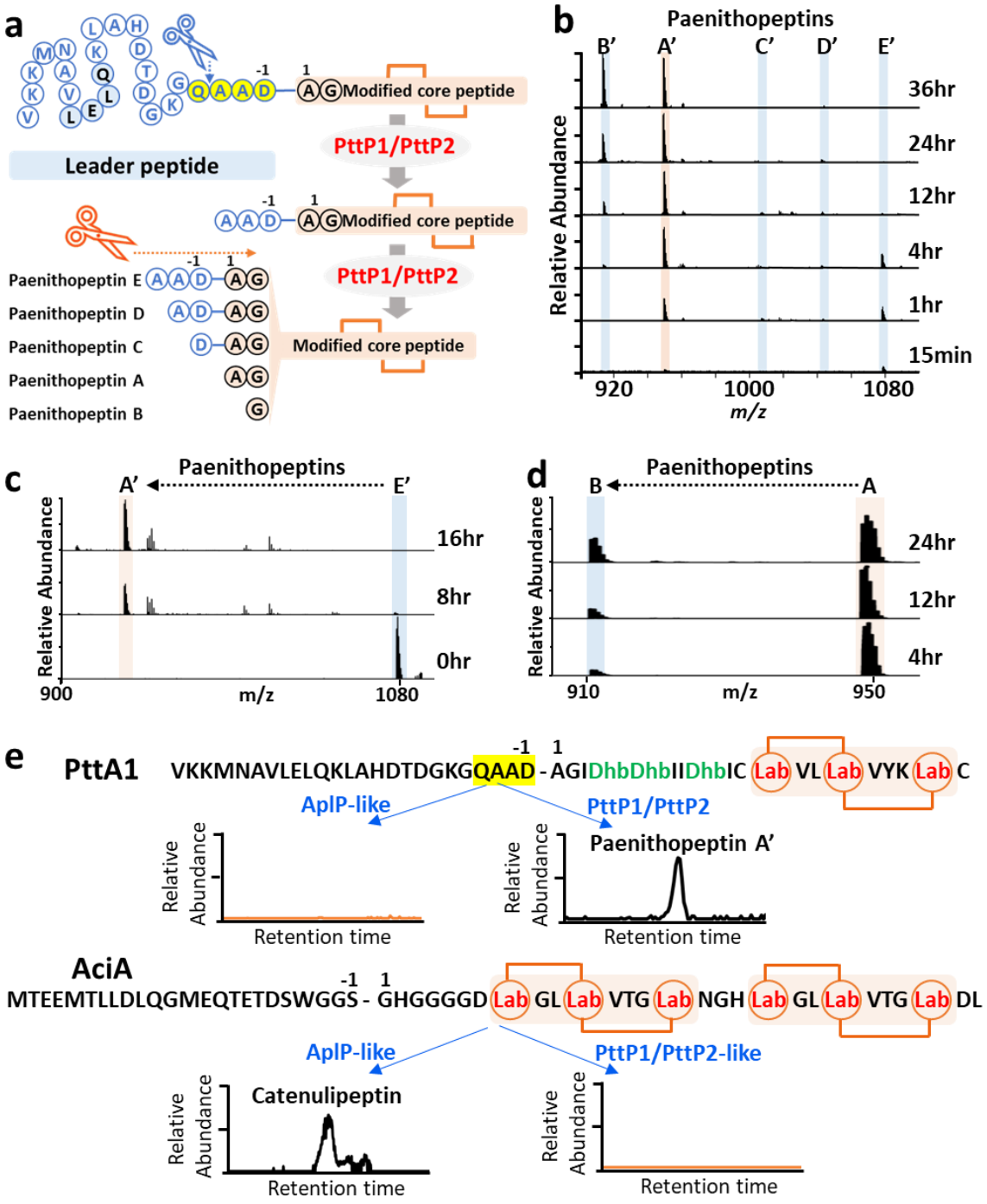


Figure 2.41 PttP1/PttP2 showed substrate specificity and bifunctional proteolytic activity for leader peptide processing. **a**, Representation of the leader peptide processing by PttP1/PttP2, through an initial endopeptidase activity to partially remove the leader followed by an aminopeptidase activity to successively remove the remaining overhangs. **b**, HPLC-MS analysis of an *in vitro* assay of PttP1/PttP2, showing production of paenithopeptins A′-E′ at different time points of the assay. Each spectrum shows the summed MS over the retention time of 5.75-6.75 min. **c**, HPLC-MS analysis showing transformation of paenithopeptin E′ to paenithopeptin A′ by aminopeptidase activity of

PttP1/PttP2. Each spectrum shows the summed MS over the retention time of 5.75-6.75 min. **d**, HPLC-MS analysis showing transformation of paenithopeptin A to paenithopeptin B by aminopeptidase activity of PttP1/PttP2. Each spectrum shows the summed MS over the retention time of 5.75-6.75 min. **e**, *In vitro* assay of PttP1/PttP2-like proteases in comparison with AplP-like proteases recently reported for class III lanthipeptide maturation, indicating that PttP1/PttP2-like and AplP-like proteases have respective substrate specificity. The conserved Q-A-(A/V)-(D/E) motif of Pre-24, absent in the precursor peptides for AplP-like proteases, was specifically cleaved by PttP1/PttP2-like proteases. The spectra shown were EIC at m/z = 949.5 (z = 2) for paenithopeptin A′ and EIC at m/z = 800.0 (z = 3) for catenulipeptin. The MS spectra in figures **b**-**e** were recorded at positive mode.

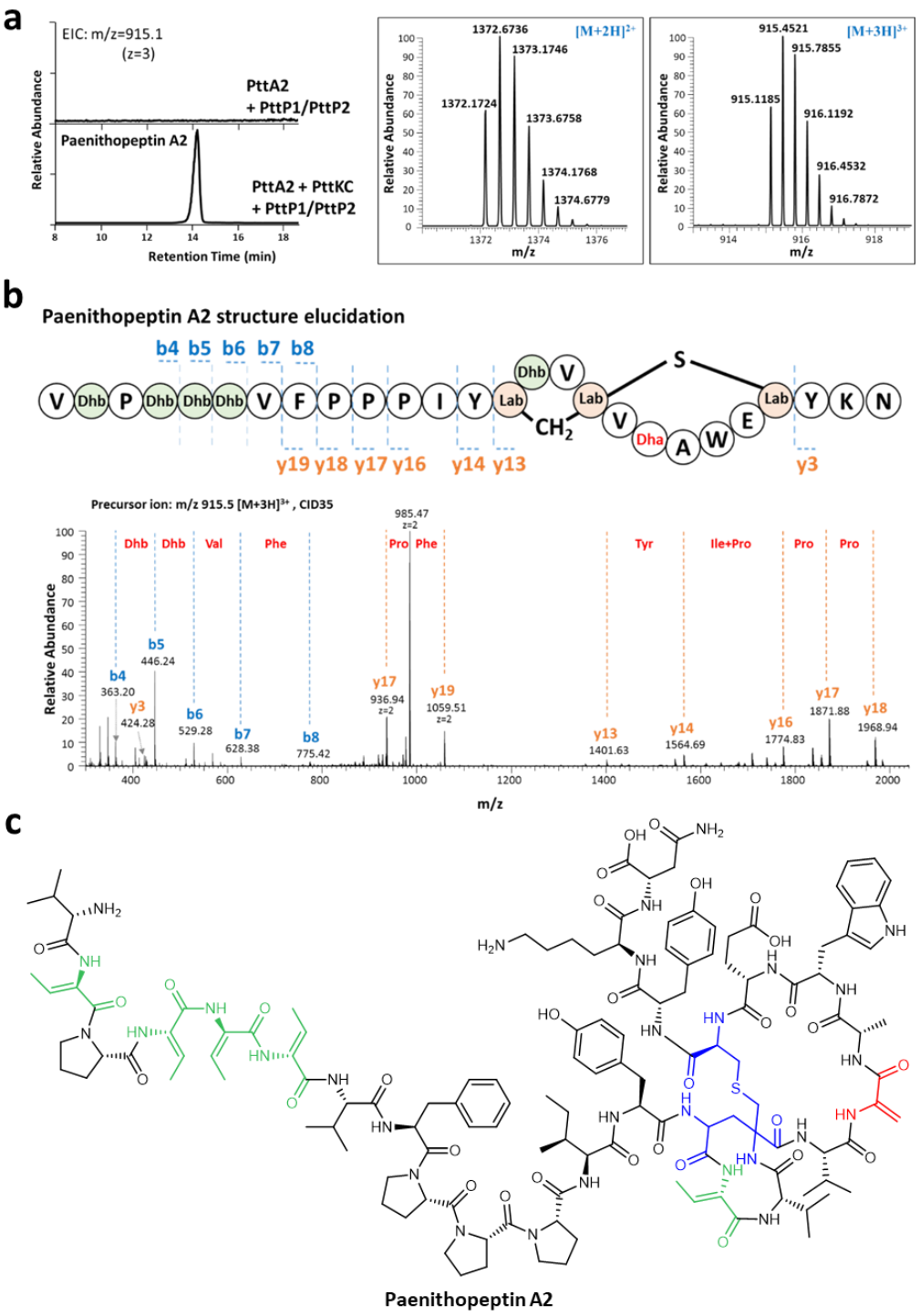


Figure 2.42 PttP1/PttP2 are responsible for the processing of paenithopeptin A2. *In vitro* enzymatic assays revealed that PttP1/PttP2 could also cleave precursor peptide PttA2 which contains the conserved Q-A-(A/I)-(D/E) motif in the leader peptide sequence. **a**, An EIC of m/z 915.1 representing the triply charged state ($[M+3H]^{3+}$) is overlaid against a control *in vitro* assay for comparison of paenithopepin A2 production. At right, doubly and triply

charged states are presented for paenithopeptin A2. **b**, The structure of paenithopeptin A2 is presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. The precursor ion m/z 915.5 representing $[M+3H]^{3+}$ was used for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **c**, Chemical structure of paenithopeptin A2. Chemical formula: $C_{134}H_{183}N_{29}O_{32}S$. Dhb is in green. Dha is in red. Ser and Cys involved in the labionin ring are in blue.

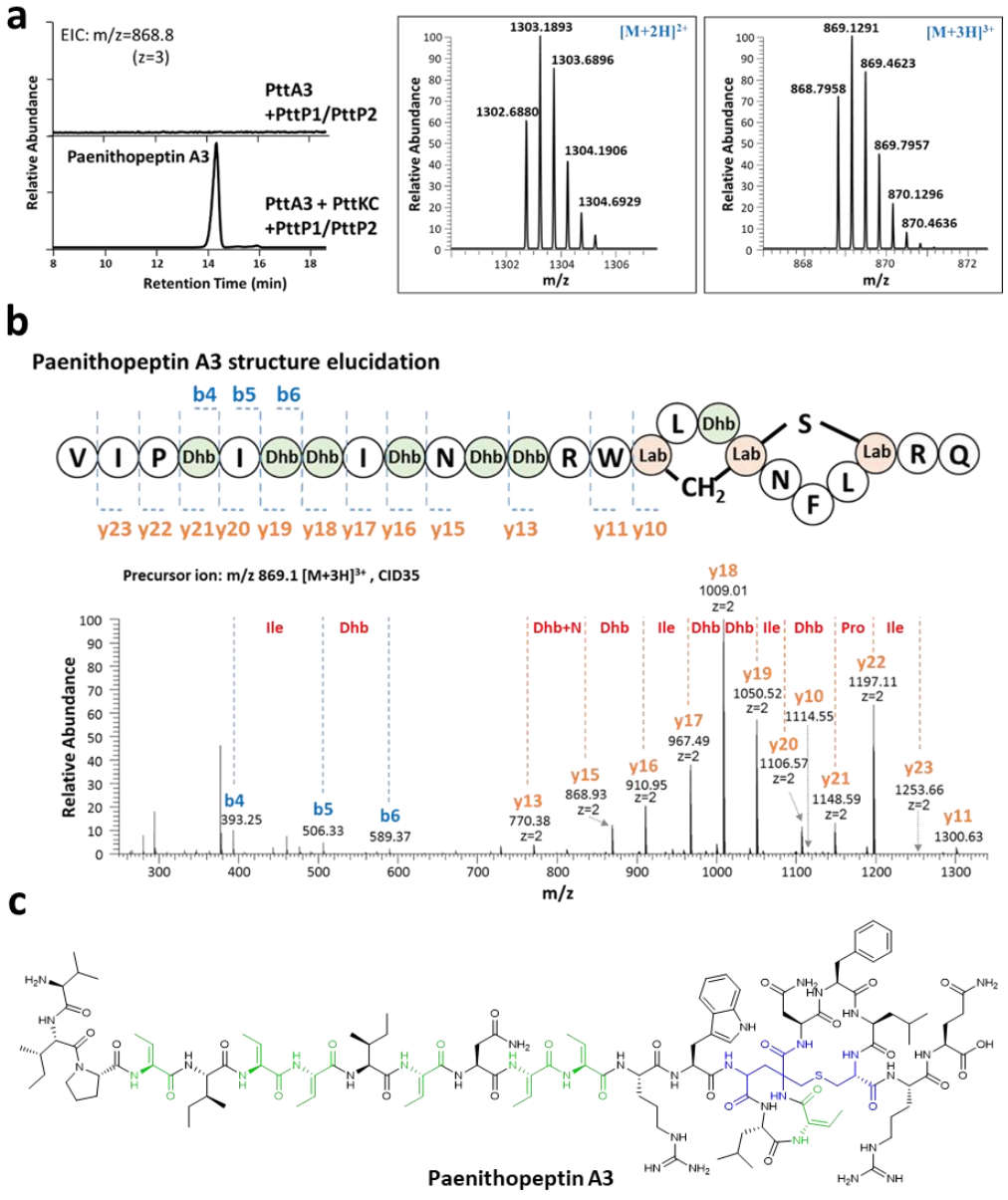


Figure 2.43 PttP1/PttP2 are responsible for the processing of paenithopeptin A3. **a**, An EIC of m/z 868.8 representing the triply charged state ($[M+3H]^{3+}$) is overlaid against a control *in vitro* assay for comparison of paenithopeptin A3

production. At right, doubly and triply charged states are presented for paenithopeptin A3. **b**, The structure of paenithopeptin A3 is presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. The precursor ion m/z 869.1 representing $[M+3H]^{3+}$ was used for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **c**, Chemical structure of paenithopeptin A3. Chemical formula: $C_{122}H_{182}N_{34}O_{28}S$. Dhb is in green. Ser and Cys involved in the labionin ring are in blue.

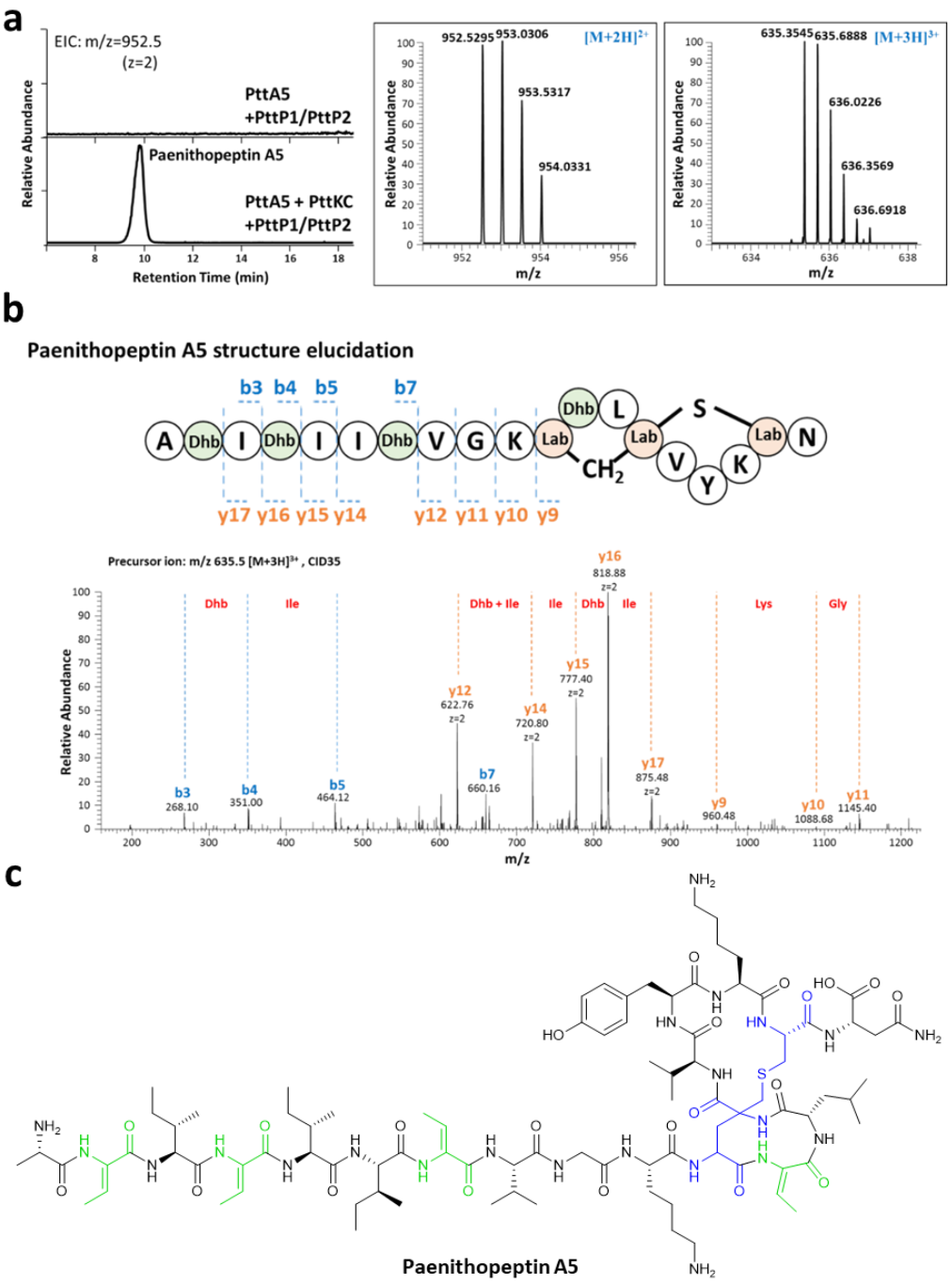


Figure 2.44 PttP1/PttP2 are responsible for the processing of paenithopeptin A5. **a**, An EIC of m/z 952.5 representing the doubly charged state ($[M+2H]^{2+}$) is overlaid against a control *in vitro* assay for comparison of paenithopeptin A5 production. At right, doubly and triply charged states are presented for paenithopeptin A5. **b**, The structure of paenithopeptin A5 is presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. The precursor ion m/z 635.5

representing the triply charged state ($[M+3H]^{3+}$) was used for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **c**, Chemical structure of paenithopeptin A5. Chemical formula: $C_{89}H_{142}N_{22}O_{22}S$. Dhb is in green. Ser and Cys involved in the labionin ring are in blue.

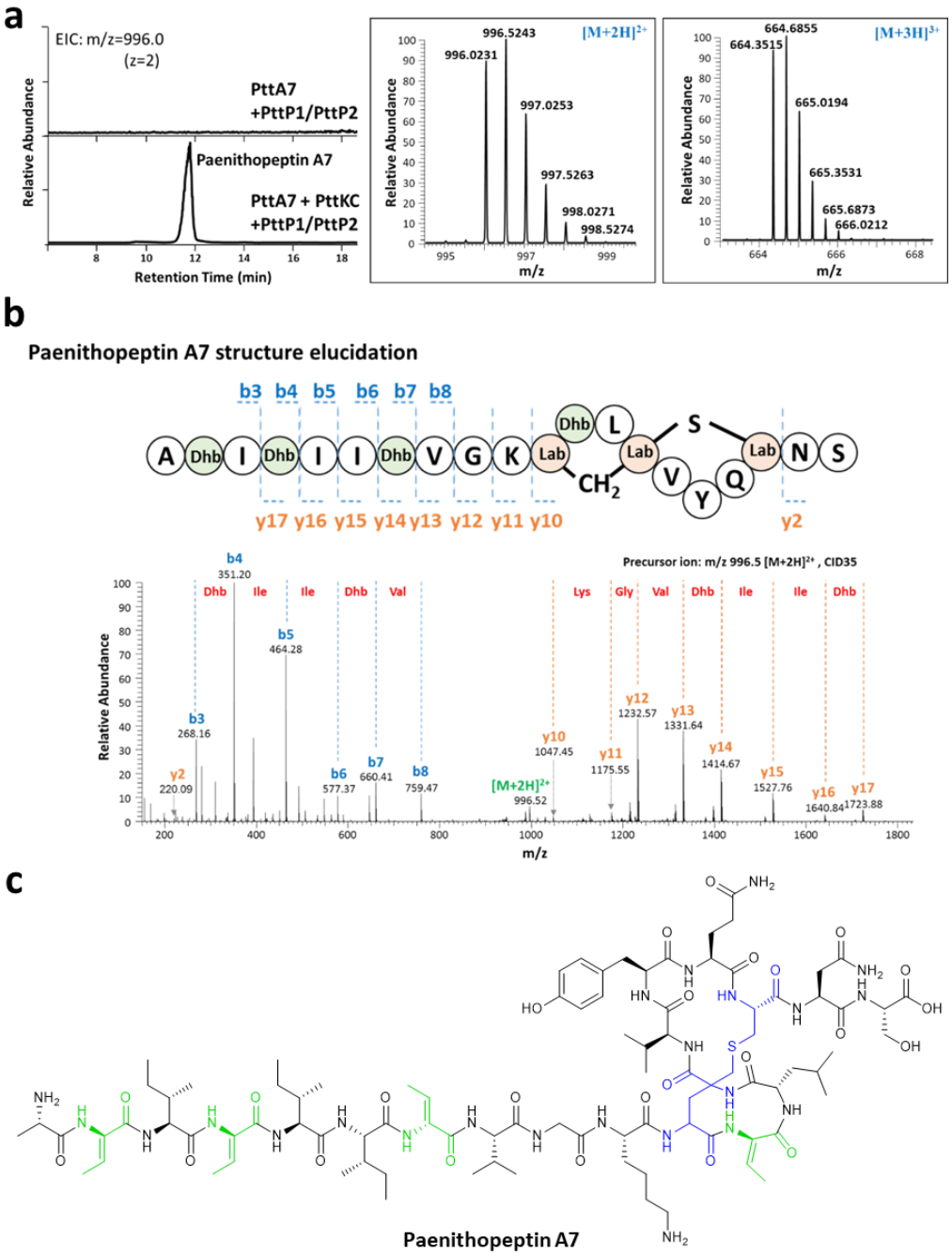


Figure 2.45 PttP1/PttP2 are responsible for the processing of paenithopeptin A7. **a**, An EIC of *m/z* 996.0 representing the doubly charged state ($[M+2H]^{2+}$) is overlaid against a control *in vitro* assay for comparison of paenithopeptin A7

production. At right, doubly and triply charged states are presented for paenithopeptin A7. **b**, The structure of paenithopeptin A7 is presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. The precursor ion *m/z* 996.5 representing $[M+2H]^{2+}$ was used for CID at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **c**, Chemical structure of paenithopeptin A7. Chemical formula: $C_{91}H_{143}N_{23}O_{25}S$. Dhb is in green. Ser and Cys involved in the labionin ring are in blue.

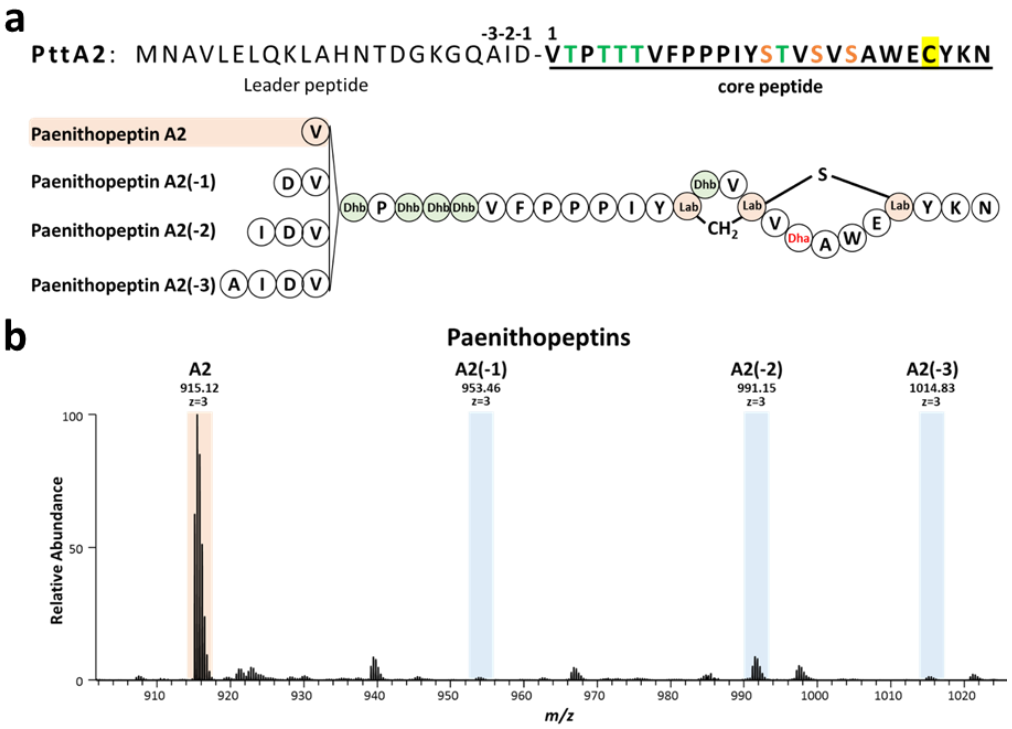


Figure 2.46 *In vitro* production and HPLC-MS detection of paenithopeptin A2 and analogs by PttP1/PttP2. PttA2 produced lanthipeptides with different amino acids overhangs by PttP1/PttP2 *in vitro*. **a**, The structure of PttA2 with the leader peptide and core peptide regions indicated and separated by a hyphen. The aminopeptidase activity of the PttP1/PttP2 resulted in the production of a series of compounds differing by one amino acid residue as represented by paenithopeptins A2 – A2(-3). Increasing numbers of additional residues are indicated with increasing x in (-x) denoting the distance from the cleavage site of the main product. **b**, High resolution mass spectrometry was used to analyze the resulting lanthipeptide products with triply charged states highlighted in the spectrum. Peaks are labeled corresponding to the structures presented in **(a)**.

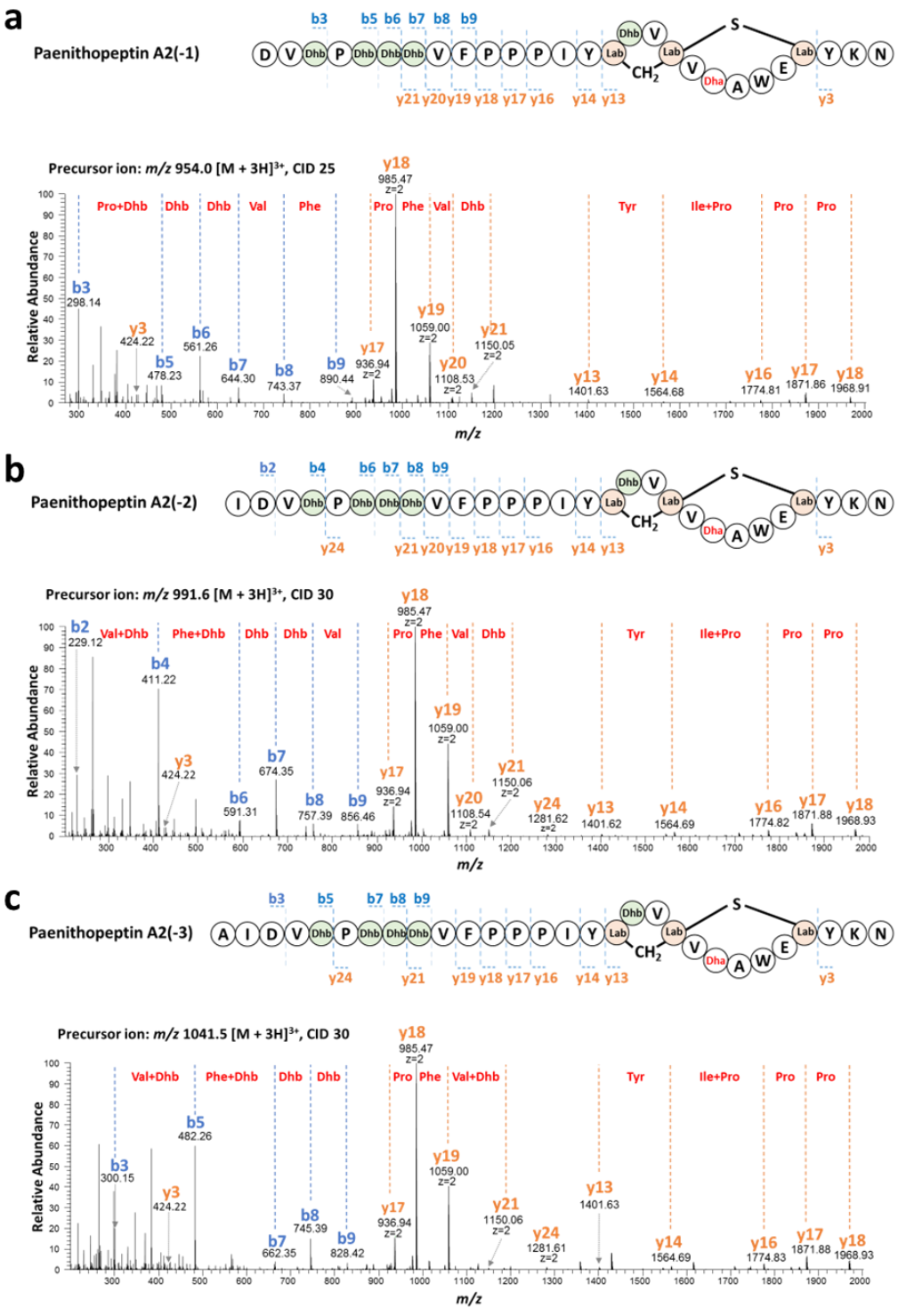


Figure 2.47 Structure elucidation of paenithopeptin A2 analogs by MS/MS fragmentation analysis. **a - c**, The structures of paenithopeptin A2(-1) **(a)**, paenithopeptin A2(-2) **(b)**, and paenithopeptin A2(-3) **(c)** are presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. Major fragment ions are annotated

with their b or y ion identity, and the amino acid residues deduced from fragment ions are labelled in red.

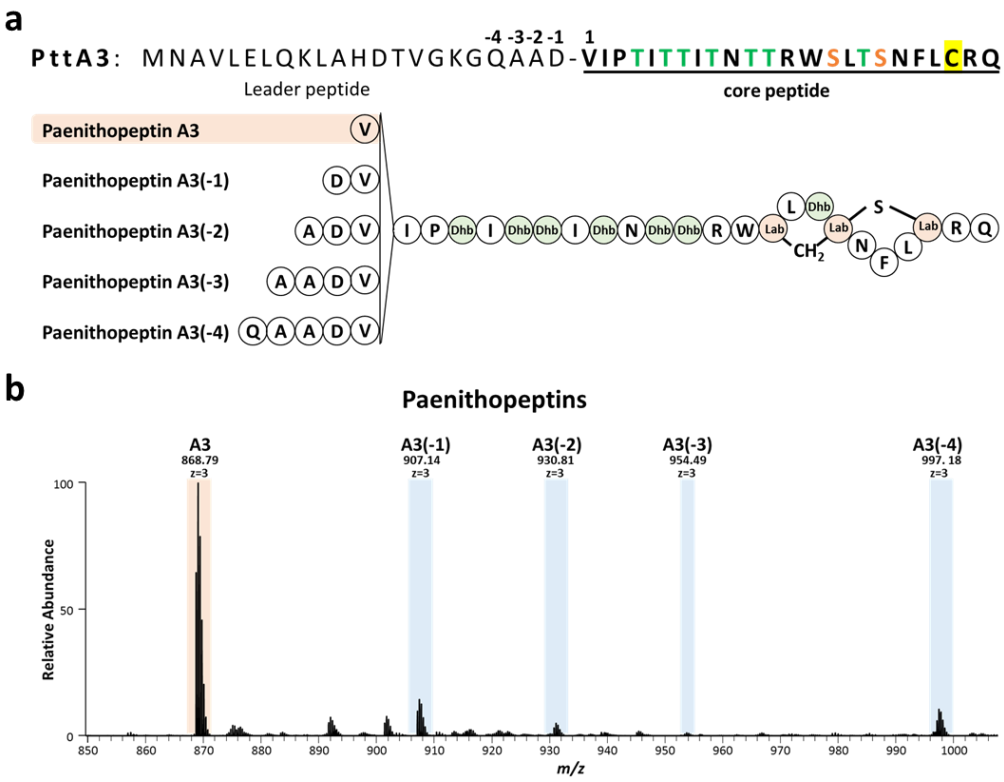


Figure 2.48 *In vitro* production and HPLC-MS detection of paenithopeptin A3 and analogs by PttP1/PttP2. PttA3 produced lanthipeptides with different amino acids overhangs by PttP1/PttP2 *in vitro*. **a**, The structure of PttA3 with the leader peptide and core peptide regions indicated and separated by a hyphen. The aminopeptidase activity of the PttP1/PttP2 should result in the production of a series of compounds differing by one amino acid residue as represented by paenithopeptins A3 – A3(-4). Increasing numbers of additional residues are indicated with increasing x in (-x) denoting the distance from the cleavage site of the main product. **b**, High resolution mass spectrometry was used to analyze the resulting lanthipeptide products with triply charged states highlighted in the spectrum. Peaks are labeled corresponding to the structures presented in **(a)**.

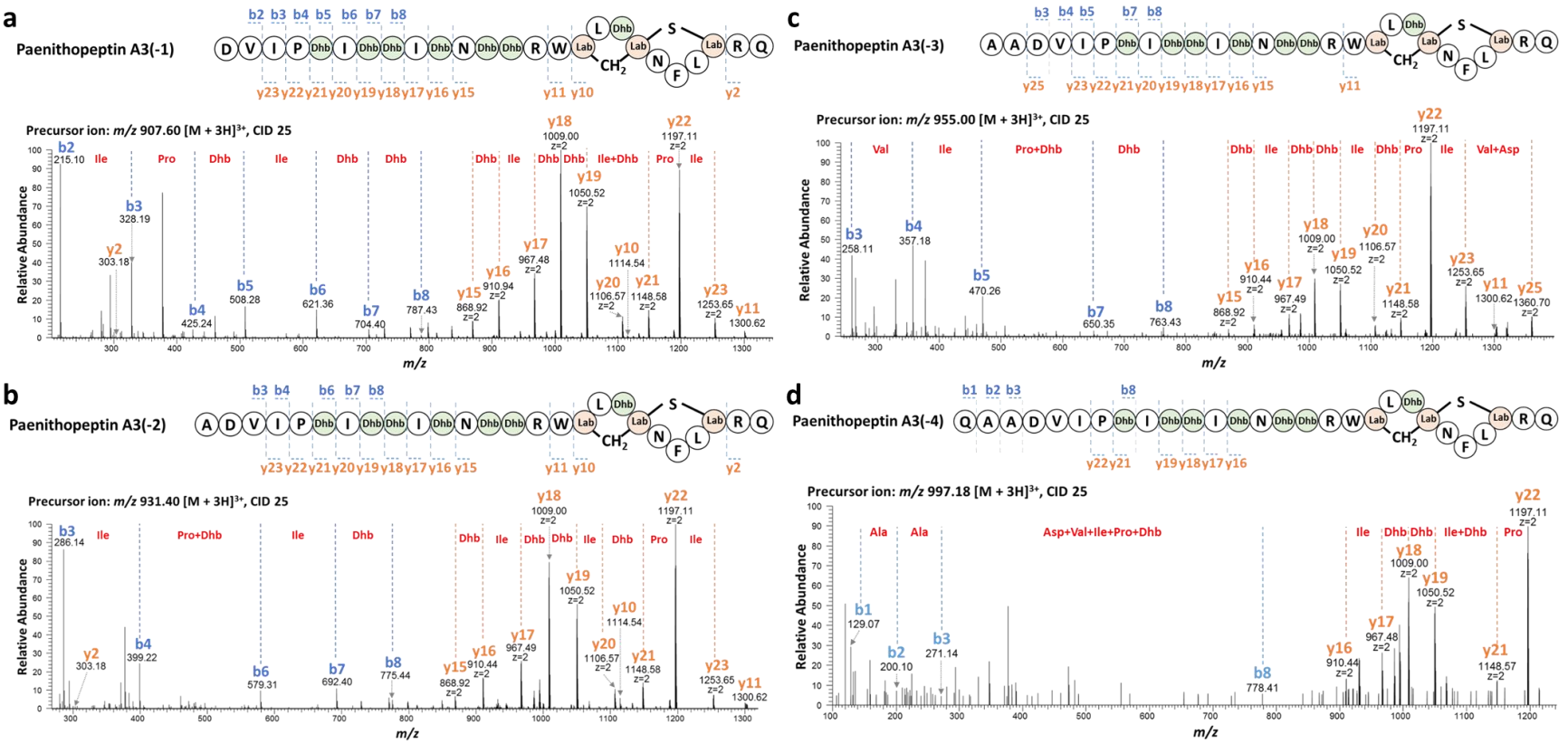


Figure 2.49 Structure elucidation of paenithopeptin A3 analogs by MS/MS fragmentation analysis. **a - d**, The structures of paenithopeptin A3(-1) **(a)**, paenithopeptin A3(-2) **(b)**, paenithopeptin A3(-3) **(c)**, and paenithopeptin A3(-4) **(d)** are presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. Major fragment ions are annotated with their b or y ion identity, and the amino acid residues deduced from fragment ions are labelled in red.

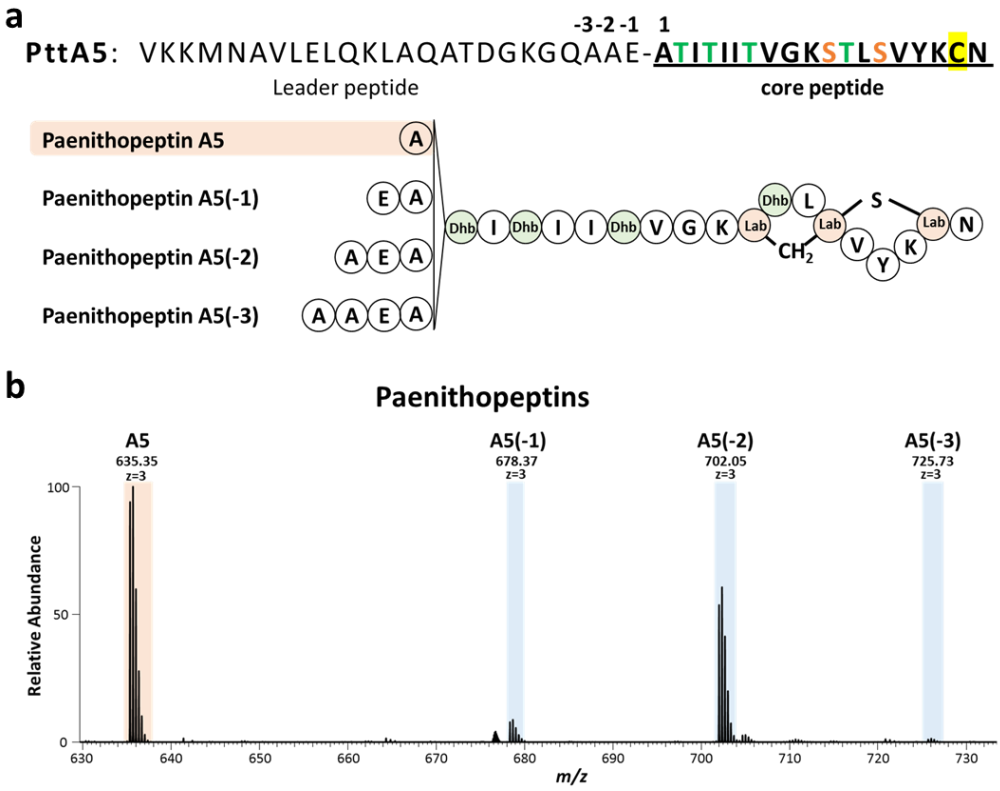


Figure 2.50 *In vitro* production and HPLC-MS analysis of paenithopeptin A5 and analogs by PttP1/PttP2. PttA5 produced lanthipeptides with different amino acids overhangs by PttP1/PttP2 *in vitro*. **a**, The structure of PttA5 with the leader peptide and core peptide regions indicated and separated by a hyphen. The aminopeptidase activity of the PttP1/PttP2 should result in the production of a series of compounds differing by one amino acid residue as represented by paenithopeptins A5 – A5(-3). Increasing numbers of additional residues are indicated with increasing x in (-x) denoting the distance from the cleavage site of the main product. **b**, High resolution mass spectrometry was used to analyze the resulting lanthipeptide products with triply charged states highlighted in the spectrum. Peaks are labeled corresponding to the structures presented in **(a)**.

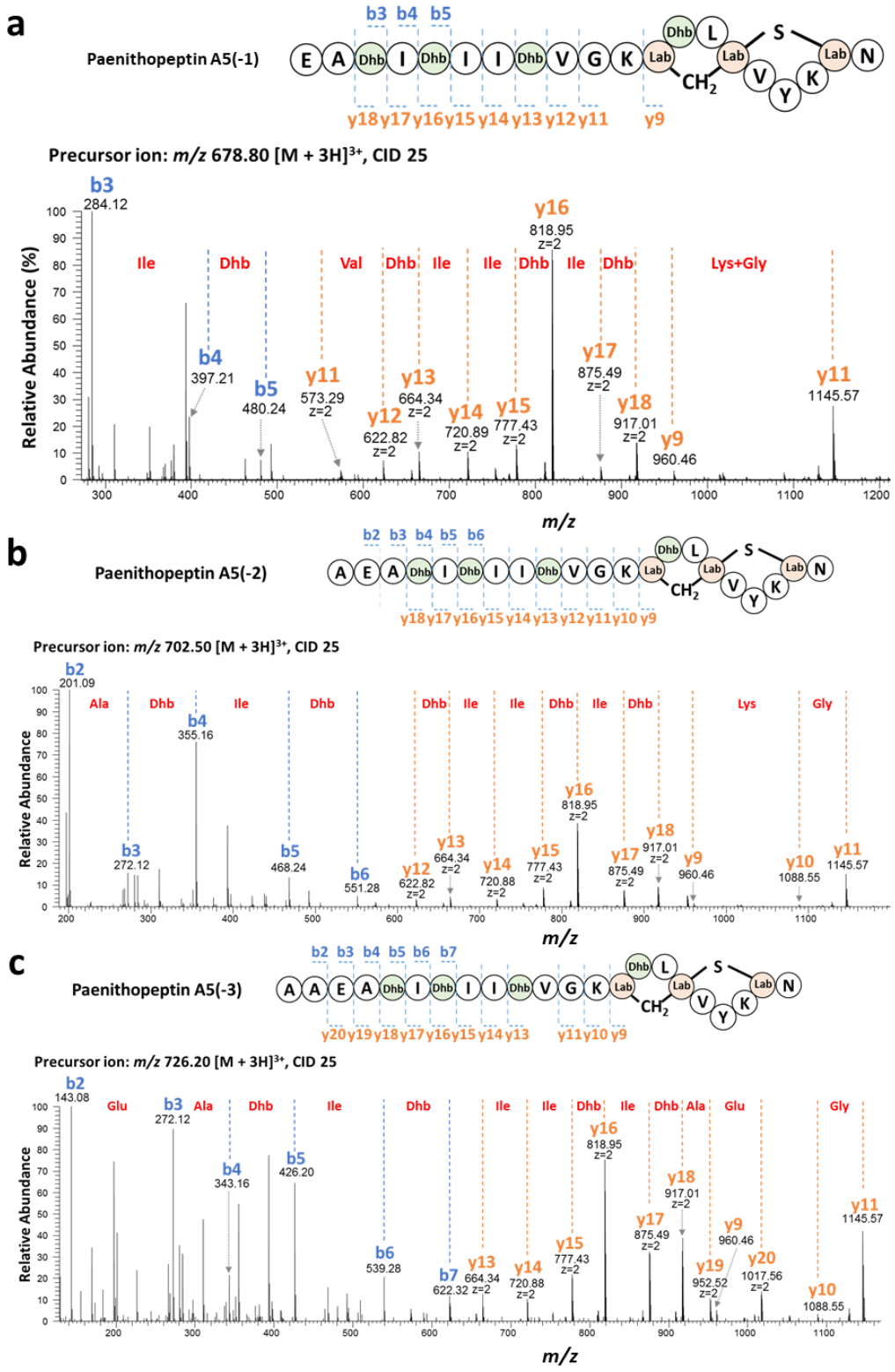


Figure 2.51 Structure elucidation of paenithopeptin A5 analogs by MS/MS fragmentation analysis. **a - c**, The structures of paenithopeptin A5(-1) **(a)**, paenithopeptin A5(-2) **(b)**, and paenithopeptin A5(-3) **(c)** are presented with

fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. Major fragment ions are annotated with their b or y ion identity, and the amino acid residues deduced from fragment ions are labelled in red.

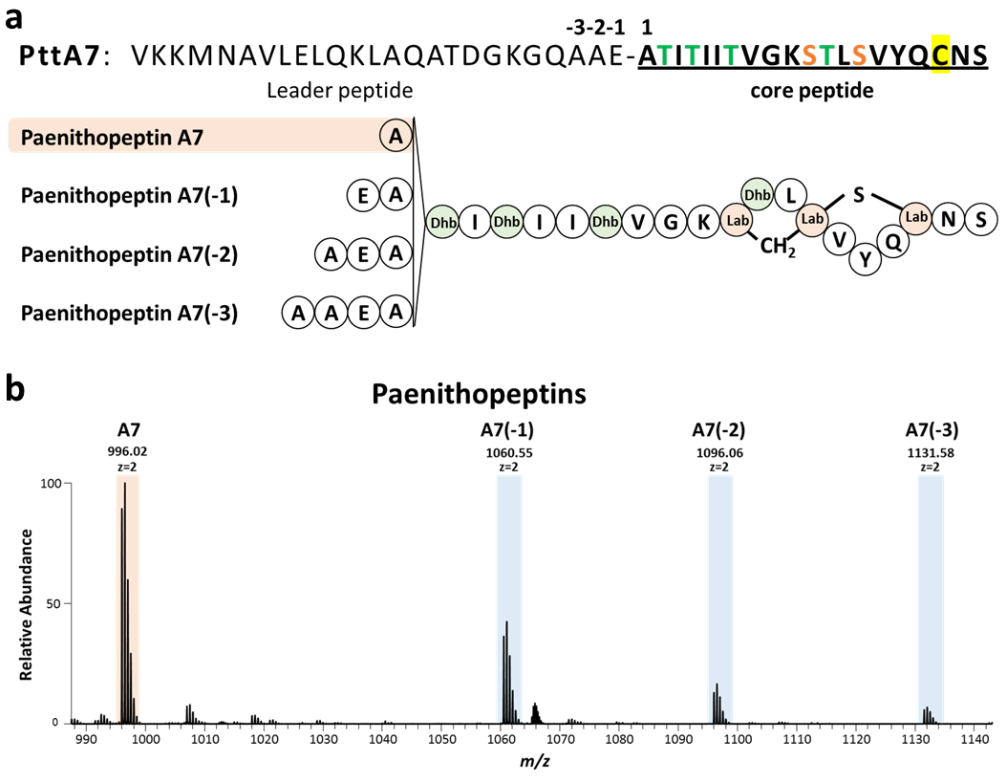


Figure 2.52 *In vitro* production and HPLC-MS analysis of paenithopeptin A7 and analogs by PttP1/PttP2. PttA7 produced lanthipeptides with different amino acids overhangs by PttP1/PttP2 *in vitro*. **a**, The structure of PttA7 with the leader peptide and core peptide regions indicated and separated by a hyphen. The aminopeptidase activity of the PttP1/PttP2 should result in the production of a series of compounds differing by one amino acid residue as represented by paenithopeptins A7 – A7(-3). Increasing numbers of additional residues are indicated with increasing x in (-x) denoting the distance from the cleavage site of the main product. **b**, High resolution mass spectrometry was used to analyze the resulting lanthipeptide products with triply charged states highlighted in the spectrum. Peaks are labeled corresponding to the structures presented in **(a)**.

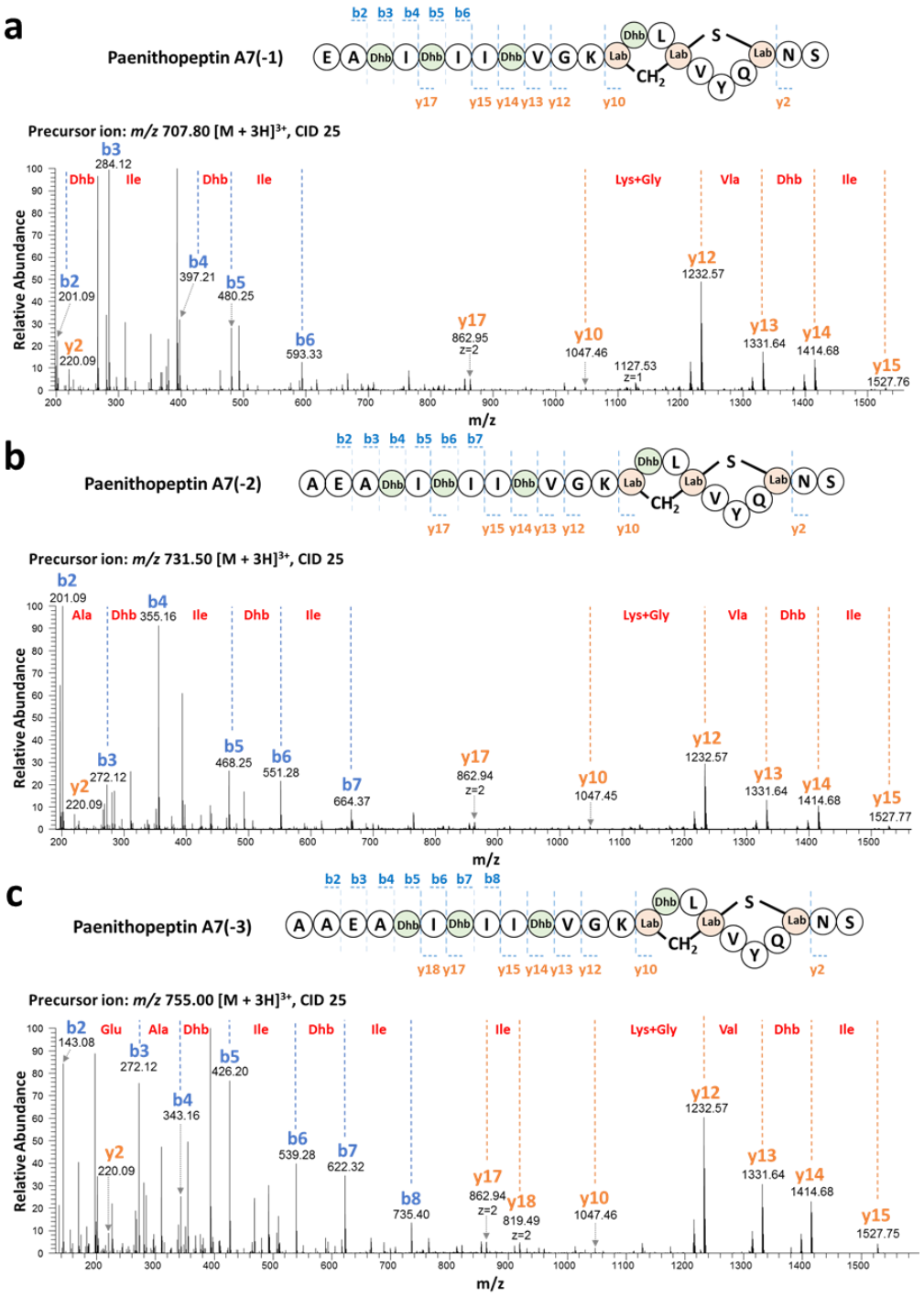


Figure 2.53 Structure elucidation of paenithopeptin A7 analogs by MS/MS fragmentation analysis. **a - c**, The structures of paenithopeptin A7(-1) **(a)**, paenithopeptin A7(-2) **(b)**, and paenithopeptin A7(-3) **(c)** are presented with fragmentation points of corresponding b and y ions as well as the location of the labionin ring marked. MS/MS was used to confirm the amino acid sequence by analysis of the fragmentation patterns. Major fragment ions are annotated

with their b or y ion identity, and the amino acid residues deduced from fragment ions are labelled in red.

We further studied substrate specificity of PttP1/PttP2 in comparison with AplP-like proteases recently discovered from Actinobacteria for the maturation of class III lanthipeptides[88]. Since the native producer of paenithopeptins A-E also harbors an AplP-like protease, we expressed this protease as a recombinant protein and tested whether it could cleave PttA1 (belonging to Pre_24), with or without modification by PttKC. However, although AplP-like proteases were hypothesized as a universal strategy for leader peptide removal of class III lanthipeptides[88], we could not detect any products, suggesting that the most abundant group of class III precursors in Firmicutes was not recognizable by Actinobacteria-derived AplP-like proteases (Figure 2.41e). On the other hand, we cloned AciA (an experimentally validated precursor for AplP-like proteases; belonging to Pre_2495) from *Catenulispora acidiphila* DSM 44928 for a comparison experiment. When AciA, with or without modification by AciKC, was incubated with a pair of PttP1/PttP2 homolog from *C. acidiphila* DSM 44928, no products were detected (Figure 2.41e). These results indicated that PttP1/PttP2 and AplP-like proteases have respective specificity against different groups of precursors and supported our protease-precursor correlation network.

2.2.5 Lanthipeptide *N*,*N*-dimethylation improves antimicrobial activity

N,N-dimethylation is an uncommon post-translational modification not only for lanthipeptides but across all RiPPs and has been linked to significantly increased antimicrobial activity such as in andalusicin, cypemicin, and plantazolicin[108–110]. Noting the presence of a methyltransferase-encoding gene in the *ptt* BGC (Figure 2.13a), we hypothesized that *N*,*N*-dimethylated paenithopeptins may also display antimicrobial

activity. We thus modified our established heterologous expression system for the *ptt* BGC by removing the precursor-encoding genes, *pttA1-A7*, and placing the post-translational enzyme-encoding genes under the control of the $P_{hyperspank}$ promoter (Figure 2.54). Precursor-encoding genes were then individually cloned into a high copy number plasmid, pBS0E, under the control of the P_{xylA} promoter and transformed into the *B. subtilis 168* host (Figure 2.54). Using this system, we produced and purified *N*,*N*-dimethylated version of paenithopeptin A5, named paenithopeptin mA5 (Figures 2.55). Indeed, we observed that the *N*,*N*-dimethylated paenithopeptin mA5 demonstrated increased antimicrobial activity compared to their non-methylated counterparts (Table 2.3).

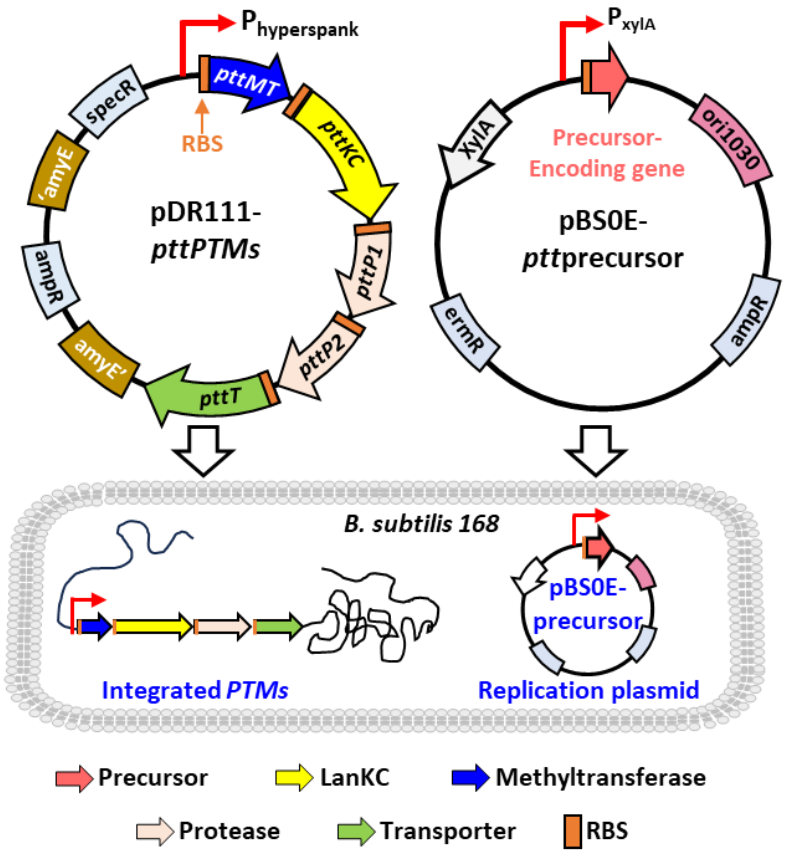


Figure 2.54 Heterologous overexpression system for the production of *N*,*N*-dimethylated paenithopeptins. Two central components were used in the system: pDR111-*pttPTMS*,

containing the *ptt* post-translational modification (PTM) enzyme-encoding genes (*pttMT* for methyltransferase, *pttKC* for lanthipeptide synthetase, *pttP1* and *pttP2* for proteases, and *ptt*T for transporter) and pBS0E-*ptt*precursor, containing one of the *ptt* precursor-encoding genes (*pttA1*-*pttA7*). pDR111-*pttPTMs* integrates the *ptt* PTM-encoding genes into the *Bacillus subtillis* 168 chromosome, while the pBS0E-precursor is subsequently transformed into the integrated host.

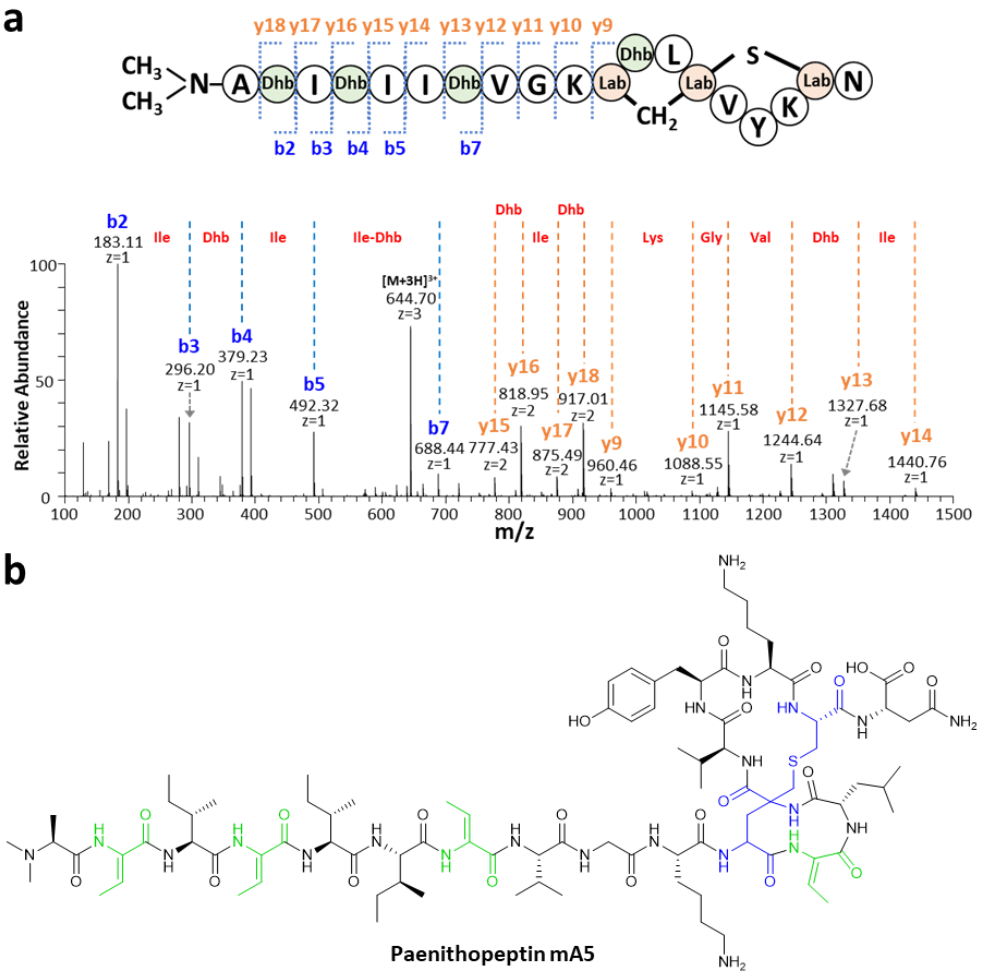


Figure 2.55 Structure elucidation of paenithopeptin mA5 by MS/MS fragmentation analysis. **a**, The amino acid sequence of paenithopeptin mA5 with b and y ions marked as well as the location of the labionin ring. The doubly charged precursor ion, *m/z* 966.54, was selected for collision induced dissociation (CID) at 35 eV. Major fragment ions are annotated with their b or y ion identity and the amino acid residues deduced from fragment ions are labelled in red. **b**, Chemical structure of paenithopeptin mA5. Chemical formula: $C_{91}H_{146}N_{22}O_{22}S$. Dhb is in green, and Ser and Cys involved in the labionin ring are in blue.

Table 2.3 Antimicrobial activity of *N*,*N*-dimethylated paenithopeptins.

Strain	Paenithopeptin	
	mA5	A5
Staphylococcus aureus RN4220	+	-
Enterococcus faecium EF16	+	-
Bacillus subtilis ATCC6633	+++	+
Paenibacillus durus DSM5976	+++	+
Paenibacillus odorifer DSM15391	+++	+
Escherichia coli DH5a	-	-

"+" corresponds to low inhibitory activity;
"+++" correspond to strong inhibitory activity;
"_" indicates no inhibitory activity.

2.4 CONCLUSIONS

Using a correlation-guided genome-mining strategy to search for unclustered natural product biosynthetic genes, we established a global network bridging the gap between bacterial lanthipeptide precursors and hidden proteases which are not encoded by their respective BGCs. Our results showcase the promise of precursor-protease correlational networking in targeted discovery of both missing proteases for previously known lanthipeptides (e.g., Prot_686 for the maturation of class I lanthipeptide paenilan) and new families of lanthipeptide proteases involved in the maturation of previously unknown products (e.g., M16B metallopeptidases leading to the production of previously unknown class III lanthipeptides paenithopeptins and bacinapeptins). This proof-of-principle study further suggests the potential of genomic-level correlational networking in discovering unclustered natural product biosynthetic genes.

Many natural product BGCs are found to lack certain key genes and must rely on enzymes encoded outside of the BGCs for biosynthesis. In lanthipeptides, a final protease performs an indispensable maturation step. Notably, with the advancement of sequencing and genome-mining technologies, increasingly more lanthipeptide BGCs lacking protease-encoding genes are being identified, consistent with our data that about one third of lanthipeptide BGCs do not harbor protease genes. Particularly, most reported class III and IV lanthipeptide BGCs lack protease-encoding genes[86]. Computational workflows have been developed to predict lanthipeptide BGCs with high accuracy by following canonical co-localized biosynthetic gene rules[91,111]. However, these approaches cannot account for hidden proteases located far outside of the BGC. We used functional domain-based genome mining to search for hidden proteases in 161,954 bacterial genomes, expanding the pool of lanthipeptide protease Pfam domains to 120 as opposed to the 6 domains (Supplementary Table 2.1) of previously reported lanthipeptide proteases LanP[82], LanT[84], FlaP[87], and AplP[88]. Furthermore, using Spearman's rank-order correlation analysis within each bacterial genus, our networking strategy linked hidden proteases to lanthipeptide precursors. As mentioned above, multiple-correlation clusters represent a challenge in determining candidate proteases. Additionally, the 5,209 precursor-protease correlations identified in this study may contain weak correlations which would require more data to refine. Addressing this, we integrated co-expression analysis to complement our approach. We envision that with increasing availability of sequenced bacterial genome and transcriptome data, we can continuously increase the power of our correlation network.

The biological function of widely distributed M16B metallopeptidases has been unclear in bacteria[112]. Our network guided the discovery of a new function of these proteins

as lanthipeptide proteases. We characterized representative Prot_819/Prot_176-belonging proteases of the M16B family and showed their high substrate specificity toward lanthipeptide precursors, despite being encoded outside of the respective BGCs. This result supports the basis of the correlational networking approach that hidden proteases exhibit a measurable degree of substrate specificity. Furthermore, varying efficiencies of Prot_819/Prot_176-belonging proteases from different genomic loci supports a process of functional evolution and provides a method to identify higher efficiency proteases for more productive bioengineering of lanthipeptides.

Consistent with the literature[73,110], our study revealed Firmicutes as a prolific source for the discovery of underexplored class III lanthipeptides that often lack pathway-specific proteases. Exploring this feature in the network, we identified previously unknown lanthipeptides paenithopeptins and bacinapeptins which represent the largest precursor group in Firmicutes. Our network further reveals the wide distribution of uncharacterized proteases in Firmicutes as exemplified by Prot_819/Prot_176-belonging proteases. Their discovery and characterization showcase the potential of our networking approach to link hidden proteases to corresponding BGCs as a means to unlock new structures with promising biological activities.

Taken together, our correlational networking approach takes advantage of readily available data and easily accessible computational platforms which can be developed into a streamlined process to discover hidden proteases for the exploitation of lanthipeptides. Furthermore, with exponentially increasing microbial genome sequences available, our approach is envisioned to be applicable to other natural product classes that lack necessary genes in their respective BGCs.

2.5 MATERIALS AND METHODS

Lanthipeptide BGCs, precursors, and proteases detection

Publicly available bacterial genomes were downloaded from NCBI RefSeq database (accessed Aug. 2019) and analyzed by antiSMASH 5.0 with default parameters. Lanthipeptide BGCs and precursors were detected automatically by antiSMASH 5.0. Lanthipeptide BGCs with a protein containing both LANC_Like domain (PF05147) and the Pkinase domain (PF00069) were classified as class III or class IV BGCs and further distinguished by profile-Hidden Markov models (pHMMs) developed previously[73].

Proteases in lanthipeptide BGCs were obtained by searching for keywords: "peptidase", "proteinase", "protease", "hydrolase", "beta-lactamase" from antiSMASH generated annotations. These proteases were further subjected to Pfam domain analysis by hmmsearch[113] (HMMER v3.3) with default parameters against the Pfam-A database[95] (v33.1). Pfam domains with hit score >0 with at least 5 occurrences in all lanthipeptide BGCs were selected and manually curated to generate a pool of 120 Pfam domains. These domains were considered as potential lanthipeptide protease Pfams (Supplementary Data 2.1). Proteases in bacterial genomes were obtained by searching for proteins in bacterial genomes and selecting proteases which contained at least one Pfam domain from the pool with hit score >0.

Protease and precursor clustering

Proteases in bacterial genomes were clustered using MMseqs2 (v11.e1a1c) using the easy-cluster workflow with the following parameters: easy-cluster --min-seq-id 0.45 --single-step-clustering --cluster-mode 1 (using "connected component" mode to cover more remote homologs). Lanthipeptide precursor sequences were clustered by MMseqs2

with the same parameters except for a larger similarity cutoff: easy-cluster --min-seq-id 0.6 --single-step-clustering --cluster-mode 1.

Correlation network construction

To develop this analysis occurrences of protease groups and precursor groups were counted in each bacterial genome and a co-occurrence networking analysis[114] using Fisher's exact test based on the presence or absence of protease and precursor groups in genomes was attempted. However, when only considering the presence or absence of proteases and precursors, information on their counts in each genome is lost. Therefore, Spearman's rank-order correlation was used to take into account the number of proteases and precursors present in a genome. For bacterial genomes that harbor multiple copies of proteases, the additional proteases might be used as the precursor-specific protease. Thus, the number of proteases may increase as precursors increase, and this association could be captured by Spearman's rank correlation.

Bacterial strains with close phylogenetic distance (e.g., different strains of the same species, different species of the same genus, etc.) have very similar genomic contents, thus these strains usually carry the same composition of protease groups and precursor groups. If correlation analysis were done in all bacterial genomes, phylogenetic relatedness, instead of precursor-protease specificity, would become the major factor, which would consequently generate uninformative correlations. Therefore, the analysis was performed on each phylogenetic level from species to phylum (i.e. only genomes within the same phylogenetic level were chosen to perform the analysis at one time; genomes outside this level were ignored) and the results were compared. In higher phylogenetic levels such as Phylum or Class, uninformative correlations due to phylogenetic relatedness instead of

precursor-protease specificity were still dominant, yet in lower phylogenetic levels such as Species, the sample sizes became too small for some species. Thus, the Genus level was chosen as a compromise between both phylogenetic relatedness and sample size. At the Genus level, genomes can be considered relatively independent of each other, thus reducing uninformative correlations.

Spearman correlation coefficients and corresponding one-sided p-values were calculated for each individual protease group to precursor group pair in each genus. P-values were further adjusted by Benjamini–Hochberg procedure[115] (false-discovery rate). All statistics were calculated in R (v.3.6). Python (v.3.7), pandas (v.0.24.2) and NumPy (v.1.17.4) were used for data processing. Calculated correlations were filtered by Spearman correlation coefficient (ρ), adjusted p-values (pAdj) and number of genomes containing the precursor-protease group pair (I). Networks were visualized by Cytoscape[116] (v.3.8.2), in which the size of node indicates the number of proteases or precursors in a genus, width of edge indicates the correlation strength (Spearman coefficient, ρ).

Transcriptomic analysis

Transcriptomic data from *Streptomyces albidoflavus* J1074, *Streptomyces coelicolor* A3(2) and *Streptomyces davaonensis* JCM 4913 were downloaded from the NCBI SRA database. SRA runs were filtered by paired library layout, random or cDNA library selection, and average length >=100. A full list of SRA runs used in this study are shown in Supplementary Data 2.1. BBMap[117] (v38.90) was used to remove base pairs with low quality as well as adapter sequences and PhiX sequences with the following parameters: qtrim=rl ktrim=r mink=11 trimq=10 minlen=45 corresponding to read quality cutoff of 10 and read length cutoff of 45 bp. rRNA sequences were filtered out using

SortMeRNA[118] (v.4.3.1) with default parameters with the smr_v4.3_sensitive_db_rfam_seeds database. Quality controlled transcriptomic data were mapped to corresponding genome sequences by BWA mem[119] (v0.7.17) with default parameters. Mapped reads were counted by featureCounts[120] (v2.0.1) with the following parameters: -M -O --fraction -p --primary --minOverlap 40 -Q 10 (minimum mapping quality score of 10, also counted multi-mapping reads). Transcripts per million (TPM) were calculated for each gene.

To calculate proteases that co-expressed with precursors, TPM of genes that belong to the same protease group (or precursor group) were summed to generate a protease and precursor TPM matrix. The matrix was further trimmed to contain only precursor groups and protease groups that exist in the correlation network calculated by genomic data. Spearman correlation coefficient between precursor groups and protease groups were calculated and one-sided p-values were adjusted by Benjamini–Hochberg procedure. Correlations presented in results of all 3 strains were taken.

For self-sequenced transcriptomic data from *Paenibacillus polymyxa* ATCC 842, the same quality control and statistical analysis workflow was performed.

Phylogenetic analysis

For proteases inside or outside BGCs at the genus *Paenibacillus*, protein sequences of all Prot_176 or Prot_819 within the Pre_24-containing BGCs, or in Pre_24-containing genomes (but outside BGCs) were selected. For proteases in genomes that didn't encode Lanthipeptide BGCs, protein sequences of all these members in Prot_176 or Prot_819 were clustered by CD-HIT[121] (v4.8.1) with default parameters to remove redundant sequences sharing >=90% sequence similarity, and representative sequences were randomly selected.

Sequences from proteases inside BGCs, outside BGCs, random representatives, as well as a manually added outgroup (WP_023988223.1) were aligned together using MAFFT[122] (v7.480) with the high accuracy linsi method. The resulting alignment was subjected to phylogenetic tree construction by IQ-TREE[123] (v2.1.3) using substitution model LG+G4. Bootstrap was calculated using the Ultrafast bootstrap[124] algorithm implemented in IQ-TREE with 1000 bootstrap replicates. The resulted tree was re-rooted at the manually added outgroup and visualized with Interactive Tree of Life[125].

Sequence logos

Full precursor peptide sequences were aligned by MAFFT (v7.480) using the linsi method and trimmed by trimAl[126] (v1.4.rev22) with the -automated1 option. The resulted alignment was used to generate sequence logos using WebLogo[127] (v3.7.4).

General materials, reagents, strains, and culture conditions

Biochemicals and media components for bacterial cultures were purchased from Thermo Fisher Scientific Co. Ltd. (USA) unless otherwise stated. Chemical reagents were purchased from standard commercial sources. Restriction endonucleases were purchased from New England Biolabs, Inc. (USA). PCR amplifications were carried out on an Eppendorf® Mastercycler® Nexus X2 Thermal Cycler (Eppendorf Co., Ltd. Germany) using PrimeSTAR HS DNA polymerase (Takara Biotechnology Co., Ltd. Japan). The E.Z.N.A.® Gel Extraction Kit (Omega Bio-tek, Inc., USA) was used for PCR products purification. The NEBuilder® HiFi DNA Assembly master mix (New England Biolabs, Inc., USA) was applied for Gibson assembly[128]. Oligonucleotide synthesis and DNA sequencing were performed by Eton Bioscience, Inc. (USA). The small ubiquitin-like modified (SUMO)-tag gene was synthesized by Bio Basic, Inc. (USA).

All strains used in this study are listed in Supplementary Table 2.3. *Paenibacillus thiaminolyticus* NRRL B4156, *Bacillus nakamurai* NRRL B41092, *P. taiwanensis* DSM 18679, *P. polymyxa* ATCC 842 and *B. subtilis* 168 strains were cultured in Tryptic Soy Broth (TSB) at 30°C. *Catenulispora acidiphila* DSM44928 was grown in ISP2 medium at 28°C. *E. coli* DH10B strains were grown in Luria-Bertani broth (LB) at 37°C. *E. coli* BL21(DE3) strains were cultured in Terrific Broth (TB) at 37°C for general growth and 22°C for protein expression.

Deletion of *prot*_819 and *prot*_176 homologs in the heterologous expression host *B. subtilis* 168

The *prot*_819 and *prot*_176 homologs in the genome of *B. subtilis* 168, *ymfF* and *ymfH*, were knocked out using the CRISPR-Cas9 system[129]. The sgRNA Designer tool provided by the Broad Institute[130] was used to check the *ymfF* and *ymfH* genes and a high-scoring 20-nucleotide (nt) sequence was identified. The candidate sgRNA sequence was synthesized as two complementary oligonucleotides and inserted between the BsaI sites of pJOE8999 to construct pJOE8999.1. Then two 700-bp fragments flanking the target region were amplified by PCR and inserted between the SfiI sites of pJOE8999.1 to generate pJOE8999.2. *B. subtilis* 168 was transformed with pJOE8999.2 and grown on LB agar plates containing 5 μg/mL kanamycin and 0.2% mannose. After incubation at 30 ℃ for 1 day, colonies were placed on LB plates without antibiotics and incubated at 50 ℃. On the next day, they were streaked on LB plates to obtain single colonies at 42°C. Finally, the colonies were tested again for plasmid loss by transferring single colonies to LB agar plates with kanamycin. Knock-out of the *ymfF-ymfH* sequence in selected colonies was confirmed by colony PCR using the outer primers from the 700-bp homology templates.

Heterologous expression of the *bcn* BGC

The *bcn* BGC does not contain *prot*_819 and *prot*_176 homologs. Thus, the *bcn* BGC and the *bcn-gP1/bcn-gP2* including their native promoters were amplified from the genomic DNA of *B. nakamurai* NRRL B41092 with appropriate overhangs for Gibson assembly using corresponding primer listed in Supplementary Table 2.4. The vector, pDR111, was linearized by digestion with SalI and SphI. The two PCR products (*bcn* BGC and *bcn-gP1/bcn-gP2*) and the linearized pDR111 were ligated via Gibson assembly to construct the heterologous expression plasmid, pDR111-*bcn+bcn-gP1/bcn-gP2,* with the *bcn-gP1/bcn-gP2* being downstream of the *bcn* BGC (Supplementary Table 2.4 and Figure 2.10a). The pDR111-*bcn+bcn-gP1/bcn-gP2* plasmid was integrated into the chromosome of *B. subtilis* 168[131], which was cultivated for production of bacinapeptins A and B as described below.

Bacinapeptins production and extraction

The heterologous expression host *B. subtilis* 168 containing plasmid pDR111-*bcn+bcn-gP1/bcn-gP2* and the native strain *B. nakamurai* NRRL B-41092 was individually inoculated in a culture tube containing 2 mL of TSB and shaken at 220 rpm, 30°C, overnight, as seed culture. The seed culture was inoculated in 30 mL of TSB in 150 mL Ultra Yield flasks (Thomson Scientific) (3×), which were subsequently shaken at 220 rpm at 30 °C for 3 days. The bacterial broth was extracted with 10 mL butanol, and the organic phase was evaporated under N_2. The extracts were redissolved in MeOH to the concentration of 10 mL/mg for HPLC-MS analysis.

Heterologous expression of the *ptt* BGC

The *ptt* BGC containing *pttP1* and *pttP2* was amplified from the genomic DNA of *P. thiaminolyticus* NRRL B4156 with appropriate overhangs for Gibson assembly. The PCR products was ligated with SalI and SphI digested pDR111 via Gibson assembly to generate plasmid pDR111_*ptt* for the heterologous expression of paenithopeptins (Figure 2.13b). Also, the *ptt-gP1* and *ptt-gP2* out of the *ptt* BGC were amplified with their native promoter and cloned into pDR111 with other *ptt* BGC genes except *pttP1* and *pttP2*, generating plasmid pDR111-*pttΔpttP1/pttP2+ptt-gP1/ptt-gP2* (Figure 2.13b). To construct this plasmid, three fragments were PCR amplified: *ptt-gP1*/*ptt-gP2*, *pttA1*, and a region containing *pttKC, pttA2-pttA7, pttMT* and *pttT*. These three PCR products were assembled with SalI and SphI digested pDR111 via Gibson assembly. Additionally, different combinations of *pttA1*, *pttKC*, *pttP1*, and *pttP2*, were constructed into a series of pDR111-based integrative plasmids using the same Gibson assembly procedure as constructing the plasmid pDR111_*ptt* (Supplementary Table 2.4). In all heterologous expression plasmids for expressing paenithopeptins, native promoters of genes were included. All these plasmids constructed above was individually integrated into the chromosome of *B. subtilis* 168 or *B. subtilis* 168*ΔymfF*H strain (Supplementary Table 2.4). The strains harboring these plasmids were cultured, and the culture broth were extracted and analyzed using the same method mentioned above for production and analysis of bacinapeptins.

Paenithopeptin isolation

The heterologous expression host *B. subtilis*168 containing pDR111_*ptt* and the native strain *P. thiaminolyticus* NRRL B4156 was individually inoculated in a 150 mL Ultra Yield flask containing 30 mL of TSB broth and shaken at 220 rpm, 30 °C, overnight,

as seed culture. The seed culture was used to inoculate 500 mL of TSB in 2.5 L Ultra Yield flasks (Thomson Scientific) (20×) for a total of 10 L TSB, which were subsequently shaken at 220 rpm at 30 °C for 3 days. The bacterial broth was extracted with an equal volume of butanol, and the combined organic phase was concentrated in vacuo to yield the extract, which was partitioned on a reversed-phase C18 open column with a 25% stepwise gradient elution from 50% H_2O/MeOH to 100% MeOH. Pure MeOH with 1% formic acid was applied for the final elution. The MeOH/formic acid fraction was further purified by HPLC (Thermo Dionex Ultimate 3000 HPLC system with Chromeleon 7.2.10) on a Phenomenex Luna RP-C18 column (250 mm × 10 mm, 5 μm, 100 Å), with 3.5 mL/min isocratic elution at 32% H_2O/acetonitrile (MeCN) over 30 min with constant 0.1% formic acid) to yield paenithopeptin A. The paenithopeptin A characterization data can be found in the Supplementary Information.

High-resolution ESI-MS, MS^n, and NMR characterization of paenithopeptins and bacinapeptins

High-resolution ESI-MS spectra and MS^n analysis of paenithopeptins A-E and bacinapeptins A and B were recorded on a Thermo Scientific Q-Exactive HF-X hybrid Quadrupole-Orbitrap mass spectrometer using electrospray ionization in positive mode. Liquid chromatography was performed on a Thermo Vanquish LC interfaced to the aforementioned mass spectrometer. The LC column was a Thermo ProSwift RP-4H with dimensions 1x250 mm. Solvent A was 0.1% formic acid in H_2O and Solvent B was 0.1% formic acid in ACN, with the column flow rate being 200 μL/min. The LC gradient started at 10% B for 1 min then went to 100% B by 10 min where it remained for 5 min. MS1 scans were obtained in the orbitrap analyzer which was scanned from 500-2000 *m/z* at a

resolution of 60,000 (at 200 *m/z*). For tandem mass spectrometry (MS/MS), the relevant parent ion was selected with a 2 m/z window and fragmented it in the HCD cell (collision induced dissociation), using normalized collision energies of 20, 25 & 30 ev (combined into one spectrum). Fragment ions were then sent to the orbitrap for mass analysis at a resolution of 30,000. The MS data was analyzed by Thermo Xcalibur (4.2.47). ^{1}H, ^{13}C, ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{1}H NOESY, ^{1}H-^{13}C HSQC, and ^{1}H-^{13}C HMBC NMR spectra for paenithopeptin A in DMSO-d_6 were acquired on a Bruker Avance III HD 400 MHz spectrometer with a 5 mm BBO ^{1}H/^{19}F-BB-Z-Gradient prodigy cryoprobe, a Bruker Avance III HD 500 MHz spectrometer with a PA BBO 500S2 BBF-H-D_05 Z SP probe, or a Bruker Avance III HD Ascend 700 MHz equipped with 5 mm triple-resonance Observe (TXO) cryoprobe with Z-gradients. Data were collected and reported as follows: chemical shift, integration multiplicity (s, singlet; d, doublet; t, triplet; m, multiplet), coupling constant. Chemical shifts were reported using the DMSO-d_6 resonance as the internal standard for ^{1}H-NMR DMSO-d_6: δ = 2.50 p.p.m. and ^{13}C-NMR DMSO-d_6: δ = 39.6 p.p.m. The NMR data was processed by MestReNova v12.0.0-20080.

Determination of absolute configuration of paenithopeptin A

Paenithopeptin A (1.0 mg) was hydrolyzed in 6 M HCl (700 μL) at 115 °C for 10 h with stirring in a sealed thick-walled reaction vessel, after which the hydrolysate was concentrated to dryness under N_2 gas. The resulting hydrolysate was resuspended in distilled H_2O (700 μL) and dried again. This process was repeated three times to remove the acid completely. The hydrolysate was divided into two portions (2 × 500 μg) for chemical derivatization with 1-fluoro-2,4-dinitrophenyl-5-L-alanine amide (L-FDAA) and 1-fluoro-2,4-dinitrophenyl-5-D-alanine amide (D-FDAA). Each hydrolysate sample was

treated with 1 M $NaHCO_3$ (100 µL) and either L- or D-FDAA (100 µL, 1% solution in acetone) at 40 °C for 1 h. The reaction was then neutralized with addition of 1 M HCl (150 µL) and diluted with MeCN (150 µL). The final sample was analyzed by HPLC-MS (Kinetex C18 HPLC column, 4.6 × 100 mm, 2.4 µm, 100 Å, 1.0 mL/min gradient elution from 95% to 38% H_2O/MeCN over 30 min with constant 0.1% formic acid; positive and negative ionization modes; UV at 340 nm). For the preparation of FDAA derivatives of amino acid standards, 50 µL of each amino acid (L-Ala, L-Ile, L-Val, L-Leu, L-Tyr, L-Lys) (50 mM in H_2O) reacted with L- or D-FDAA (100 µL, 1% solution in acetone) at 40 °C for 1 h in the presence of 1 M $NaHCO_3$ (20 µL). The reaction was quenched by 1 M HCl (20 µL) and diluted with MeCN (810 µL) followed by HPLC-MS analysis using the same column and elution condition as above. The L- and D-FDAA derivatives were detected by either UV-vis or EIC. Absolute configurations of amino acid residues in paenithopeptin A were established by comparing the retention times of FDAA-derivatized peptide hydrolysate with those of amino acid standards.

Plasmids construction for *in vitro* enzymatic assays

All genes encoding precursor peptides involved in this research were PCR amplified from corresponding genomic DNA with appropriate overhangs for Gibson assembly using primer listed in Supplementary Table 2.5. The SUMO-tag gene was synthesized with designed overhangs for Gibson assembly. The vector, pET-28a(+), was linearized by digestion with NdeI and EcoRI. Then, the linearized vector, the SUMO-tag gene, and the PCR amplified gene encoding a precursor peptide were ligated via Gibson assembly to generate corresponding plasmid, with the precursor gene downstream of the SUMO-tag gene, for the expression of His_6-SUMO tagged precursor peptide.

To construct plasmids for the expression of lanthipeptide modification enzymes, genes encoding modification enzymes were PCR amplified from corresponding genomic DNA with appropriate overhangs for Gibson assembly using primer listed in Supplementary Table 2.5. The vector, pHis8, was linearized by digestion with NcoI and EcoRI. The linearized vector and the PCR amplified products were ligated via Gibson assembly to construct the plasmids for the expression of target proteins with an 8xHis tag at the N-terminus.

Mutagenesis in plasmid pHis8-*pttP1* and pHis8-*pttP2* was performed by PCR-based site-directed mutagenesis[132]. Briefly, primers incorporating the desired base changes were designed and applied through PCR to amplify target genes containing desired mutations. Mutated PCR products were used for Gibson assembly with linearized pHis8 as mentioned above to generate plasmids containing mutations.

Protein expression and purification

Plasmid constructed with pHis8 or pET-28(a)+ were transferred into *E. coli* BL21(DE3) by electroporation for protein expression. Precursor peptides were produced in a form tagged with 6×His fused and a SUMO fusion partner at the N-terminus. Modification proteins were expressed with an 8×His tag at the N-terminus. *E. coli* BL21(DE3) cells were transformed with pHis8 or pET28 derivative plasmids containing genes encoding precursor peptides or modification enzymes (Supplementary Table 2.4). A single colony was used to inoculate a 10 mL culture of LB supplemented with 50 mg/L kanamycin. The culture was grown at 37 °C for 8 h and used to inoculate 1 L of LB with kanamycin. Cells were grown at 37 °C to OD600 ~0.6-0.8, then cooled to 16 °C before IPTG was added to a final concentration of 0.25 mM. The culture was incubated at 20 °C

for an additional 12 h. Cells were harvested by centrifugation at 5,000 ×g for 30 min at 4 °C. Cell pellets were resuspended in 30 mL of lysis buffer (20 mM Tris, pH 8.0, 300 mM NaCl, 25 mM imidazole, 5% glycerol) and the suspension was sonicated on ice for 20 min to lyse the cells. Cell debris was removed by centrifugation at 15,000 ×g for 60 min at 4 °C. The supernatant was loaded onto a 3ml HisSpinTrap^{TM} column (GE Healthcare) previously charged with Ni^{2+} and equilibrated in lysis buffer. The column was washed with 10 mL of wash buffer I (35 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl) and 10ml of wash buffer II (55 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl). The protein was eluted stepwise with elution buffer I (250 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl) and elution buffer II (500 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl). Resulting elution fractions were collected and analyzed by SDS-PAGE. Fractions containing target proteins were combined and concentrated using an Amicon Ultra-15 Centrifugal Filter Unit (10 kDa for precursor peptide, 30kDa for modification enzymes, MWCO, Millipore). The resulting protein sample was stored at −70 °C.

Pulldown assay

The interaction between Bcn-gP1 and Bcn-gP2 was studied by a pull-down assay using HisSpinTrap^{TM} columns (GE Healthcare) with a bed volume of 200 µl. Binding buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl, 5% (v/v) glycerol) was used to immobilize protein (75 nmol) on the column then the column was washed with wash buffer (binding buffer with 30 mM imidazole added) and eluted by elution buffer (binding buffer with 250 mM imidazole added). Bcn-gP1 and His_8-Bcn-gP2 were mixed in a ratio of 1:1 and incubated at 30 °C for 2hrs to induce complex formation before applying to the affinity nickel column. A column control was also run to identify and eliminate false positives

caused by nonspecific binding of Bcn-gP1, as well as a control with His_8-Bcn-gP2 but no Bcn-gP1. We also performed the pull-down assay to explore the interaction between PttP1 and PttP2 using the same procedure. All pull-down assays were repeated three times independently.

In vitro enzymatic assays

Assays were based on previous reports on class I and III lanthipeptide enzymes[87,88,97,133]. The class I precursor peptide was treated as previously described for the dehydration of nisin with some modification. Briefly, precursor peptide (pllA; 20 μM) was incubated with cell extracts obtained from BL21 (DE3) expressing pllB and pllC (450 μL each) in 285 μL reaction buffer (100 mM HEPES pH 7.5, 1 mM dithiothreitol (DTT), 10 mM L-glutamate, 10 mM $MgCl_2$, 10 mM KCl, 5 mM ATP) in a final volume of 1.5 mL. The assay was incubated at 30 °C for 5h after which Tris-HCl pH 8.0 was added to 50 mM and Prot_686 to 2 μM followed by another 2h incubation at 30 °C. The reaction was quenched and extracted with butanol. The butanol phase was dried under N_2 and redissolved in 50 μL 50% MeOH/H_2O for further LC-ESI-MS analysis.

Class III precursor peptides (100 μM) were incubated with corresponding modification enzymes (20 μM) in 200 μL reaction buffer (20 mM Tris, PH 8.0, 10 mM $MgCl_2$, 1 mM DTT, 5 mM ATP). After 4h incubation at 30°C, Prot_176 and/or Prot_819 protease(s) (10 μM each) and 2.5 mM $ZnSO_4$ were added into the reaction. All *in vitro* assays were repeated three times independently and the statistical analysis was performed by GraphPad Prism 7.00.

Protease homology modeling

Homology based modeling was performed using the SWISS-MODEL platform[134].The amino acid sequences of Bcn-gP1/Bcn-gP2 and PttP1/PttP2 were individually used to search for crystal structure templates in the SWISS-MODEL repository. Templates were evaluated according to global mean quality estimate (GMQE) as calculated by SWISS-MODEL, sequence identity and similarity, as well as literature investigation to ensure a heterodimer structure was used for the modeling of putative heterodimers of Bcn-gP1/Bcn-gP2 and PttP1/PttP2. Based on these criteria, the M16B metallopeptidase heterodimer from *Sphingomonas* sp. A1 with PDB code 3amj[105] was selected as the template with a GMQE score of 0.6. The two sequences in this dimer contain the corresponding HXXEH and R/Y motifs found in Bcn-gP1/Bcn-gP2 and PttP1/PttP2, further supporting their use as templates. Models were then visualized and RMSD calculations were performed using PyMol.

Antimicrobial Activity Assay

Paenithopeptins were tested for antimicrobial activity using an agar diffusion assay. The compounds were dissolved in dimethyl sulfoxide (DMSO) and diluted with H_2O to a final concentration of 1 mg/mL and 2% DMSO. Inoculated LB agar test plates were prepared by diluting overnight-cultured bacterial strains 5000-fold in molten 0.75% LB agar. Sterile filter paper discs (diameter 5 mm) were then placed onto the agar surface, and 2 μL of paenithopeptin solution was aliquoted onto the filter paper. Ampicillin (1 mg/mL) was used as a positive control and 2% DMSO in H_2O was used as a negative control. Plates were then incubated at 30 °C for 24 h. The diameter of the growth inhibition zone was measured and assigned a category ("–", "+", "++", or "+++"), corresponding to the strength

of inhibition: "–" indicates no inhibition zone, growth inhibition zones between 1 and 2 cm are represented by "+", growth inhibition zones between 2 and 3 cm are represented by "++", and growth inhibition zones greater than 4 cm are represented by "+++".

2.6 Data Availability

All genomes used in this research were obtained from NCBI Assembly ResSeq database (https://www.ncbi.nlm.nih.gov/assembly), with a full list of accession numbers provided in Supplementary Data 2.1. Transcriptomic data for *Streptomyces* were obtained from NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra), with a full list of run accessions provided in Supplementary Data 2.1. All precursor sequences and their clustering information are in Supplementary Data 2.1. All protease sequences, their clustering information, and editable Cytoscape files for Figures 2.2c, 2.4, and 2.7c are deposited in Zenodo (https://zenodo.org/record/5162875). The crystal structure of Sph2681/Sph2682 used as the template for homology modeling was obtained from Protein Data Bank (PDB code 3amj). Paenilan transcriptomic data are uploaded to NCBI (BioProject PRJNA777777: https://www.ncbi.nlm.nih.gov/bioproject/777777).

2.7 Code Availability

Codes used for computing correlation networks are available on GitHub (https://github.com/yxllab-hku/correlational-network) and Zenodo (https://zenodo.org/record/5842713).

2.8 Acknowledgements

The work presented in the chapter was completed in collaboration with several incredibly talented researchers. Specifically, thank you to Dr. Dan Xue and Dr. Lukuan Hou, and Mr. Nolan Dittenhauser for their guidance and assistance in molecular biology

techniques. Thank you to Dr. Zhuo Shang and Dr. Michael Walla for their help with structure elucidation of new lanthipeptides by NMR and MS/MS, respectively. Also, special thank you to Dr. Yong-Xin Li and Dr. Zheng Zhong for their assistance in developing the bioinformatic pipeline featured in this work.

CHAPTER 3
BIOSYNTHETIC ENZYME-DISEASE CORRELATION CONNECTS GUT MICROBIAL SULFONOLIPIDS TO INFLAMMATORY BOWEL DISEASE

3.1 BACKGROUND AND INTRODUCTION

The human gut microbiome, composed of trillions of microorganisms, is intricately linked to human health[33]. At the species abundance level, numerous human microbes have been rigorously correlated with disease phenotypes, however, the mechanisms by which these microbes influence human health remain largely unkown[135–137]. One important mechanism is through the biosynthesis of functional metabolites by microbes[138–142]. Microbial functional metabolites are in direct contact and constant exchange with human cells, granting them inherent biological activity in complex host-microbe interactions[138,139,141]. Accordingly, human microbiome research has begun advancing to the next level of revealing these microbial functional metabolites and their corresponding molecular mechanisms that drive specific disease phenotypes[143,144].

Many studies have used untargeted metabolomics to probe trends between functional metabolite abundance and disease. However, this approach alone may face challenges such as the complexity of the metabolome, the lack of reference databases for identification, trace metabolite amounts, and a high degree of variability between different metabolomes[145–147], all of which can lead to difficulty in revealing specific gut microbial functional metabolites as drivers of molecular mechanisms in disease. In contrast, disease-related sequencing datasets are widely available, high quality, less dimensional, and less

variable. Thus, we set out to develop a unique approach which leverages this data to connect the biosynthesis of functional metabolites directly to human health conditions. Our approach takes advantage of critical biosynthetic enzymes specifically required for the biosynthesis of their corresponding functional metabolites. Increased or decreased expression of these enzymes, reflecting the differential production of a specific functional metabolite, can be correlated with different disease states to reveal positive or negative associations. Through these correlations, our "biosynthetic enzyme-guided disease correlation" approach can reveal trends in functional metabolite biosynthesis in the context of human health conditions, enabling a more focused, targeted chemoinformatic analysis to rapidly fish out metabolites of interest and further confirm their association with disease. Furthermore, with the increasing availability of disease-related sequencing data[148,149] and the rapidly advancing investigation of microbial biosynthetic pathways[58,150,151], our approach can be broadly applied to uncover previously unknown human microbial metabolites with potentially important human health implications.

As proof of principle, we apply our approach to microbe-derived lipids, a remarkable class of functional metabolites. While many studies have focused on common microbe-derived lipids, such as short chain fatty acids (SCFAs) and phospholipids[152,153], there are a significant number of underexplored lipids which may be equally capable of influencing human health[154]. One such class of microbe-derived lipids is sulfonolipids (SoLs), unique lipid molecules that bear striking structural similarity to both bacterial and endogenous sphingolipids (SLs) which are known for their role in mediating immune signaling in humans[155]. SoL-producing bacteria do not produce SLs but instead produce SoLs in high abundance[156,157], suggesting that SoLs may replace SLs as functional

metabolites with similar but distinctly different functions. In fact, two SoL-producing genera, *Alistipes* and *Odoribacter,* have been negatively correlated with the two primary forms of IBD, ulcerative colitis (UC) and Crohn's disease (CD)[31,158–160], with some species shown to ameliorate IBD symptoms[159,160]. We have also found that sulfobacin A (SoL A), a representative SoL produced by *Chryseobacterium gleum* F93 DSM 16776, exhibits unusual immunoregulatory activity *in vitro* by modulating inflammatory cytokine production, especially through suppression of the lipopolysaccharide (LPS)-induced inflammatory response[161] which has been reported as a key contributor to the progression of IBD[162–165]. Whether SoLs produced by *Alistipes* and *Odoribacter,* two genera negatively associated with IBD[31,158], represent functional metabolites in this negative correlation is unknown. Furthermore, the molecular target(s) of SoLs as a whole class of unique and abundant lipids is also unknown.

3.2 Biosynthetic Enzyme-Guided Disease Correlation

3.2.1 Genome Mining the Human Gut Microbiome

We began by systematically investigating the biosynthetic potential of SoLs from 285,835 human gut bacterial reference genomes including single amplified genomes (SAGs) and metagenome-assembled genomes (MAGs)[166]. Based on sequence homology with experimentally verified SoL biosynthetic enzymes[161,167–169] (Figure 3.1, Table 3.1), we identified a total of 562,214 homologous enzyme sequences, including 469,012 cysteate synthases (CYS), 33,486 cysteate fatty acyltransferases (CFAT), and 59,716 short-chain dehydrogenases/reductases (SDR) (Figure 3.2a). Uncovering phylogenetic trends, we found that these three enzymes were widely distributed in 255,572 genomes (Figure 3.2a) across 21 phyla, with the majority belonging to Bacteroidota and Firmicutes_A (Figure

3.2b). A subset of 6.21% (15,863/255,572) of the genomes was found to encode all three putative SoL biosynthetic enzymes (Figure 3.2a). To prioritize them for further analysis, we filtered the homologs on the basis of three rules: (1) the homology of both CFAT and CYS must equal or exceed 50% sequence similarity with experimentally validated CFATs and CYSs (Table 3.1), as these enzymes are the first two specific enzymes in the biosynthetic pathway of SoLs that distinguish the biosynthesis between SoLs and SLs[161,167–169]; (2) the homologous regions of CYS, CFAT, and SDR should include protein domains with Pfam IDs PF00291, PF00155, and PF00106, respectively (hit score > 50); (3) a set of homologous enzymes, especially SDR enzymes that show variable sequence similarities, should come from the same genome encoding all three enzymes as all three are required for SoL biosynthesis, thus ensuring co-occurrence. Applying these rules, we prioritized 9,731 CYS (1,384 unique sequences), 9,740 CFAT (917 unique sequences), and 10,319 SDR enzymes (1,076 unique sequences) (Figure 3.3a, Supplementary Data 3.1) from 9,633 bacterial genomes. The prioritized enzymes were distributed among 42 species from Bacteroidota (99.99%, 9632/9633, 95% confidence interval: 99.94% ~ 100%) and one species from Firmicutes_A (0.01%, 1/9633, 95% confidence interval: 0.0018% ~ 0.059%) (Figure 3.3b, Supplementary Data 3.2). Of note, among the 42 species from Bacteroidota, 71% (30/42) of them belong to bacterial families that have been previously reported to produce SoLs (Figure 3.3b) including *Rikenellaceae* (containing genera *Alistipes* and *Alistipes*_A)[38,156], *Marinifilaceae* (containing genus *Odoribacter*)[156], and *Weeksellaceae* (containing genus *Chryseobacterium* B)[161,170].

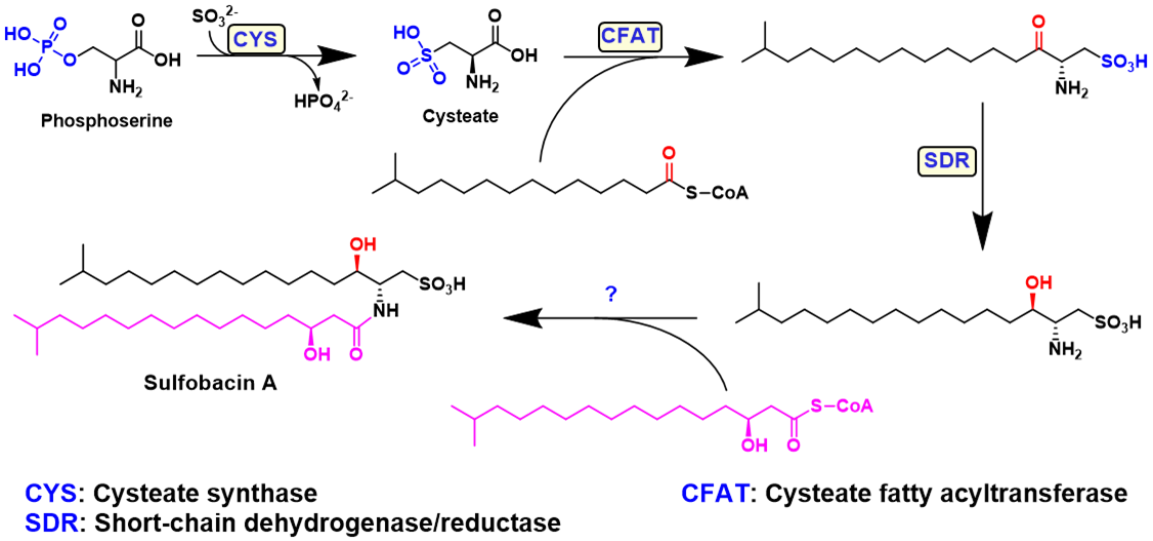


Figure 3.1 Biosynthetic pathway of sulfonolipids. Illustration of the biosynthetic pathway of sulfonolipids (SoLs). Currently, the first three genes have been reported and characterized: CYS, CFAT, and SDR[161,167–169]. The two first genes, CYS and CFAT have been shown to be specifically involved in SoL biosynthesis[161], allowing them to be used to distinguish between SoL and sphingolipid biosynthesis.

Table 3.1 Experimentally validated and literature reported sulfonolipid biosynthetic genes.

Gene Number	Gene Name	NCBI Accession #	Taxonomy	Sequences
Gene-1	CYS-1	EFK34534.1	*Chryseobacterium gleum* ATCC 35910	MSNVYDNILGLIGHTPMVKLNTVTKDIPATVYAKLESYNPGHSTKDRIALHIIENA EKKGLLKEDSVVVETTSGNTGFSIAMVCIIKGYKCILAVSDKTKPEKIAYLKALGA TVYICPANVPADDPRSYYEVAKRIAKETPNSIYINQYFNELNIDAHYQTTGPEIWE QTQGKITHLFACTGTGGTLSGSAKFLKEKNPDIKIIGVDADGSILKSYHETGEIHKE DVHPYQIEGMGKNLIPSALLFDKVDEFVRVNDEMSAYRTREIALKEAIMGGYTTG AVTQGLMQYAQSHELTENDLVVLIYPDHGSRYITKVYSDKWMAEQGFVNNCVH NYDEVFKTEFIK
Gene-1	CYS-3	SEA02653.1	*Alistipes timonensis* JC136 DSM25383	MKTKKFCLSENQMPTQWYNIVADMPTKPLPPLHPGTKQPVTKEQMSAIFAEELID QEMSTERYIDIPEEVQEIYKIWRPTPLVRATGLEKALGTPAKIYFKNESVSPAGSHK PNTAVPQAYYNYKQGIKHLTTETGAGQWGASIAFAAKHFGIDLQVFMVKVSYD QKPYRRLMMNTWGAECVASPSTLTNSGRAALERDPHCSGSLGLAISEAIEVALRR PEDTRYCLGSVLNHVVLHQTVIGQEAVTQMEMADAEPDMVIGCFGGGSNFAGIG FPFLQKNLTEGKHIRVIAVEPEGCPKLTRGEFQYDFGDVAGFTPLLPMYTLGHDF QPSDIHAGGLRYHGAGSIVSQLMKDGLIEAQSMPQVETLAAGILFAQTEGIIPAPES THAIAATIREALKAKEEGVSKTILFNLSGNGVIDLYAYEQYLAGALKDFSPSDEEI KKTVNQLEHLI
Gene-1	CYS-4	EAZ80562.1	*Algoriphagus machipongonensis*	MIYNSIIDTIGNTPMIRLNNLAKDIKGEVLVKVEYFNPGNSMKDRMAIKMVEDAE KAGLLKPGGTIIEGTSGNTGMGLALAAVAKGYKCIFTMADKQSKEKIDILKAVGA EVIVCPTNVSPEDPRSYYSVARKLNADIPNSFYPNQYDNLSNTAAHYETTGPEIWK DTEGKITHYAAGVGTGGSMCGTAQYLKEQNADVVTVGIDTYGSVFKKYKETGV FDEKEVYPYLTEGIGEDILPKNVNFDMIDHFVKVTDKDAAVMTRRLAKEEGLFV GWSCGSAVHGALEWAKENLKEGDQMVVILPDHGTRYLGKIYNDDWMRNHGFL EDRTYSKARDIIAQRSGSYTLVSTKKTDSVRDAIHLMNNTSVSQIPVMENGEVVG SLTDNKLLAKIIENPALKDAKVSEVMEESMHFVAMDSTLDVLSSMVDKEKAVLV RDVQDQIHIITKHDILDAFTK
Gene-1	CYS-2	ABQ05446.1	*Flavobacterium johnsoniae* UW101	MVKDLFERIQDNKGPLGKWASQAEGYYVFPKLEGELGPRMQFHGKNILNWSLN DYLGLANHPEVRKADTDAAAQFGAAYPMGARMMSGHTTYHEQLENELASFVM KESAYLLNFGYQGMVSIIDALVTKNDIIVYDVDSHACIIDGVRLHMGKRFTYKHN DLESMEKNLQRATKMAEETGGGILFITEGVFGMRGQQGKLKEIVALKQKYNFRL LVDDAHGFGTLGKTGAGAGEEQGVQADIDVYFSTFAKSMANIGAFVAADKTVID YLKYNLRSQMFAKALPMIQTIGSLKRLELLRQSSAIKDKLWENVNALQSGLKERG

				FNIGDTNTCITPVYLEGSIPEAMVMVNDLRENYGIFLSIVVYPVIPKGIILLRMIPTA SHTLEDIKETLAAFEAIREKLVNGTYKEIAERTTVDVS
	Gene-1	CYS-5	MBM7419833.1	*Chryseobacterium* JUb44	MSNVYDNILGLIGSTPLVKLNTVTKEIPATIYAKLESYNPGHSTKDRIALHIIENAE KKGLLKEDSVVVETTSGNTGFSLAMVCIIKGYKCILAVSDKTKPEKIAYLKALGA TVYVCPANVPANDPRSYYEVAKRIAAETPNSVYINQYFNELNIDAHYHTTGPEIW EQTQGKITHLFACTGTGGTLSGSAKFLKEKNPDIKIIGVDADGSILKSFHETGEIHK EDVHPYQIEGMGKNLIPSALLFDKIDEFVRVNDEMSAYRTREIALKEAIMGGYTT GAVTQGLMQYANSHQFSETDLVVLIFPDHGSRYITKVYSDKWMAEQGFINNCFH NYEEVFKTEIIK
	Gene-2	CFAT-1	EFK34533.1	*Chryseobacterium gleum* ATCC 35910	MLDIFERIKENPGPLGQFADYGEGYFIFPRLEGPIGPRMQFQGREVIFWSANDYLG LCNHPEVIEADAKAAAEYGMFYPMGARAMSGETDQHLQLERELADFVKKESAY LLNFGYQGMVSTIDALVSRNDVIVYDMDSHACIVDG VRLHSGKRFTYKHNDMASLEKNLQRATKVAEETGGGILVITEGVFGMRGQQGKI KEICDLKSKYQFRLLVDDAHGFGTLGKTGAGVGEEQDCNDQIDVYFSTFAKSMA GFGAFLAGDKEIIRYLKFNLRSQIFAKSLTMPMVIGGLKRLELLRSRPEIKAKLWE NVYKLQNGLKERGFNIGDTNTCVTPVMMQGTPVEATLLVKDLRENYGIFTSVVV YPVIPKGMILLRLIPTASHTDAEINETLAAFEAIHDKLVGGYYKEQEQKLLQEQGL SFKPI
	Gene-2	CFAT-2	SEA34915.1	*Alistipes timonensis* JC136 DSM25383	MVDIFSRLEKNAGGPIGQYMEYAHGYYAFPKLEGDIGPHMVFRGKKMLNWSLN NYLGLANHPEVRKADAEGAAQFGMAAPMGARMMSGQTVYHERLERELAEFVG KEDAFLLNFGYQGMISIIDCLLTPRDVVVYDAEAHACIIDGLRLHKGKRFVFGHN DMDSLRLQLQHATDLAEEQNGGVLVITEGVFGMKGDLGKLDEIVALKKDFQFRL LVDDAHGFGTMGPGGRGTAAHFGVADGVDVLFNTFAKSMAGIGAFVSGPRWLI NLLRYNMRSQLYAKSLPMPMVIGALKRLELIRNHPEFQQKLWENVRALQSSLKE NGFEIGVTNSPVTPVFLKGGIPEATNLVVDLRENHGIFCSMVVYPVIPKGEIILRIIP TAVHTLEDVKVTIEAFKAVREKLESGYYASLPIPVRADEGFKVR
	Gene-2	CFAT-3	EAZ83176.1	*Algoriphagus machipongonensis*	MGPLGKHSQFSDGYYMFPKLEGEIAPRMKFQGKEVLTWSLNNYLGLANHPEVR KTDAEAAAKWGAAYPMGARMMSGQTSLHEKLESELAKFVGKEKSYLLNYGYQ GIMSVIDALLDRKDVVVYDSECHACIIDALRMHMGKRYVFPHNDIENCEKQLER ATKLAQETGGGILVITEGVFGMTGDQGKLDEICALKEKFEFRLLVDDAHGFGTLG KTGAGTHEEQGVINEVDLYFSTFAKSMASIGAFIAGDEKVIHYLRFNMRSQIFAKS LPMILVEGALKRLELLQTQPELKDNLWKVVNALRDGLHREGFSTGQSNSPVTPV VLNGTVGEAAALSHDLRENFGIFCSVVIYPVVPKGMIILRLIPTAVHSLEDVEETIN AFATVKEKLSGGIYKNSELAISFGE
	Gene-2	CFAT-4	MBM7419832.1	*Chryseobacterium* JUb44	MLDIFERIKENPGPLGQFADYGEGYFIFPKLEGPIGPRMQFQGREVIFWSANDYLG MCNHPEVLEADAKAAAEYGMFYPMGARAMSGETHQHLQLEKELAEFVQKESA YLLNFGYQGMVSTIDALVSRNDVIVYDVDSHACIVDGVRLHAGKRFTYRHNDIE

Gene-2	CFAT-5	AFL77572.1	*Alistipes finegoldii* DSM17242	SLEKNLQRATKVAEETGGGILVITEGVFGMRGQQGKLKEICELKSKYQFRLLVDD AHGFGTLGETGAGAGEEQGCQDQIDVYFSTFAKSMAGFGAFIAGDKEIIRYLKFN LRSQIFAKSLTMPMVIGGLKRLELLRTRPEIKAKLWENTLKLQNGLSERGFNIGDT NTCVTPVMMQGSPVEATLLVKDLRENYGIFTSVVVYPVIPKGMILLRLIPTASHTD AEINETLAAFDAIHDKLKNGYYKEQEQKLLSEKGLSFKEI MVDIFARLEKNAGGPIGQYMSYAHGYFAFPKLEGEIGPHMVFRGKKMLNWSLN NYLGLANHPEVRKADAEGAAKFGMAAPMGARMMSGQTVYHEQLERELAEFVG KEDAFLLNFGYQGMISIIDCLLTPRDVVVYDAEAHACIIDGLRLHKGKRFVYGHN DMDSLRLQLQHATDLAEEQKGGVLVITEGVFGMKGDLGKLDEIVALKKDFQFRL LVDDAHGFGTMGEGGRGTASHFGVTDGVDVLFNTFAKSMAGIGAFVCGPRWLV NLLRYNMRSQLYAKSLPMPMVMGALKRLELIRNHPEYQQKLWEIVRALQNGLK ENGFEIGVTNSPVTPVFMKGGIPEATNLIVDLRENHGIFCSIVIYPVIPKGEIILRVIP TAAHTLDDVNYTIAAFKSVRDKLEGGIYAQMPIPVRADEGFKVR
Gene-3	SDR-1	QAR30703.1	*Ornithobacterium rhinotracheale*	MRYAVVTGVSSGIGKAICEKFLAKGLYVFGSVRKKEDAKYFEEKYPNTFHTLVF DTTDYPAVDKAVEEIHKVVGKKGLSVLVNNAGVAKYGPIQHVPIEELRQQYEVN VFASVYLTQKLLWLLGASKEAKWQGKVIQISSTAGVMTRPMLGPYSSSKHAVEA IYDALRRELMIYGVEVVLIEPGPIKTEIWGKAKSGGNPYKDTDYGEIFAQLDKAV DEIEKIGLPVEAVAGKAWEAFVAKKPKARYVVAPKKLMFKAAMYLIPDRMLDKI FYKDLKKLTQES

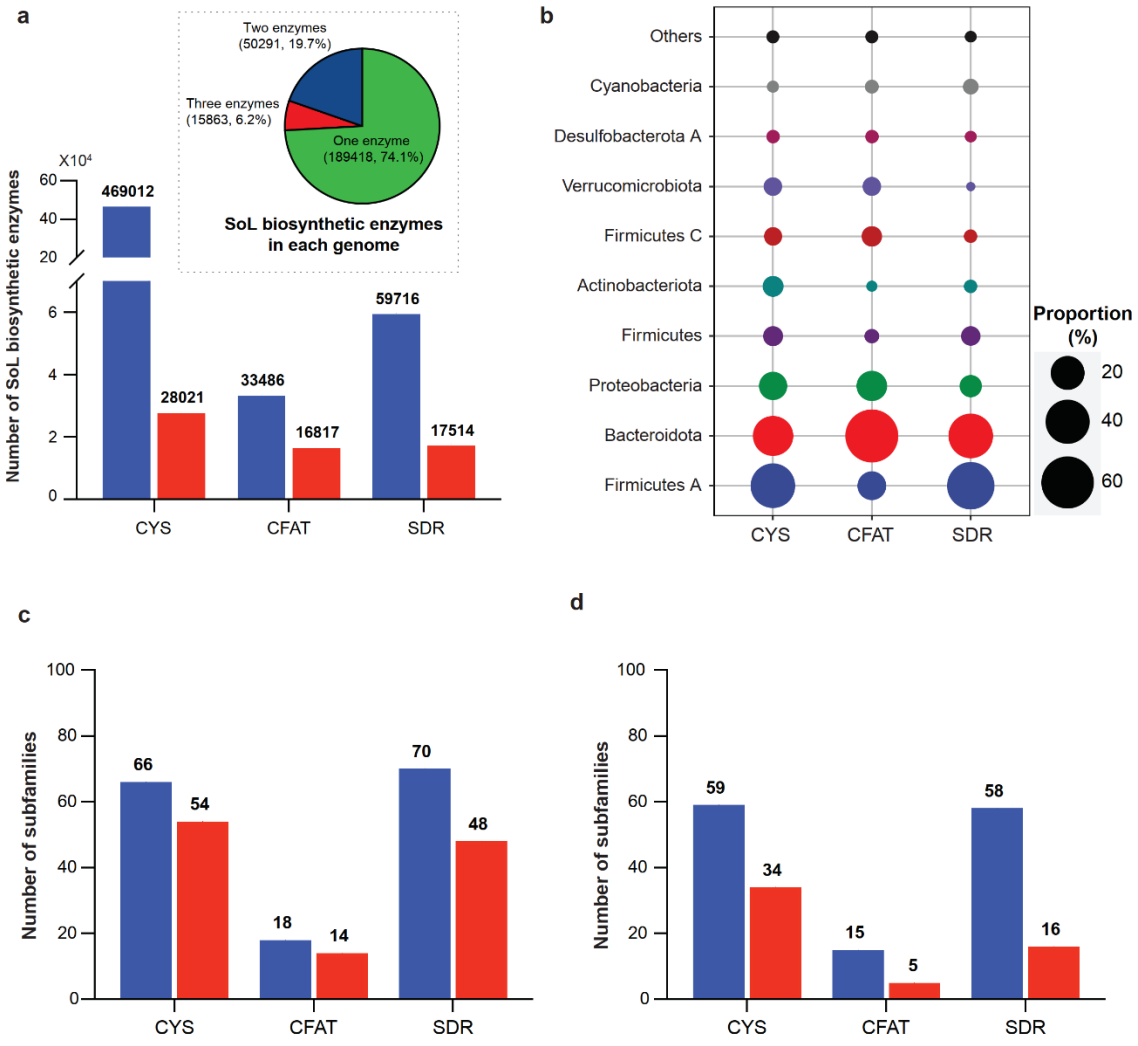


Figure 3.2 Genome mining and distribution of SoL biosynthetic genes in human gut microbial genomes. **a**, 562,214 Sulfonolipid (SoL) biosynthetic enzymes (colored in dark blue) were identified from 255,572 genomes. The pie chart indicates the proportion of genomes containing one (green), two (blue), or three (red) SoL biosynthetic enzymes. The bar chart highlighted in red shows the number of SoL biosynthetic enzymes identified from genomes encoding copies of all 3 known SoL biosynthetic enzymes. **b**, The proportion of each biosynthetic enzyme distributed in corresponding bacterial phyla. The proportions of enzymes encoded by phyla which were less than 1% were merged into the "Others" category. CYS and SDR enzymes are widely distributed in multiple phyla including Firmicutes, Bacteroidota, and Proteobacteria, while CFAT enzymes are predominantly encoded by Bacteroidota. **c**, 154 prioritized enzyme subfamilies could be identified in the metagenomic samples from the IBD cohort (colored in dark blue), and 116 subfamilies were detected in ≥ 5% of samples (colored in red). **d**, 132 prioritized biosynthetic enzyme subfamilies expressed in the metatranscriptomic samples collected from IBD cohorts (colored in dark blue), with 55 clusters detected in ≥ 5% of samples (colored in red)**.**

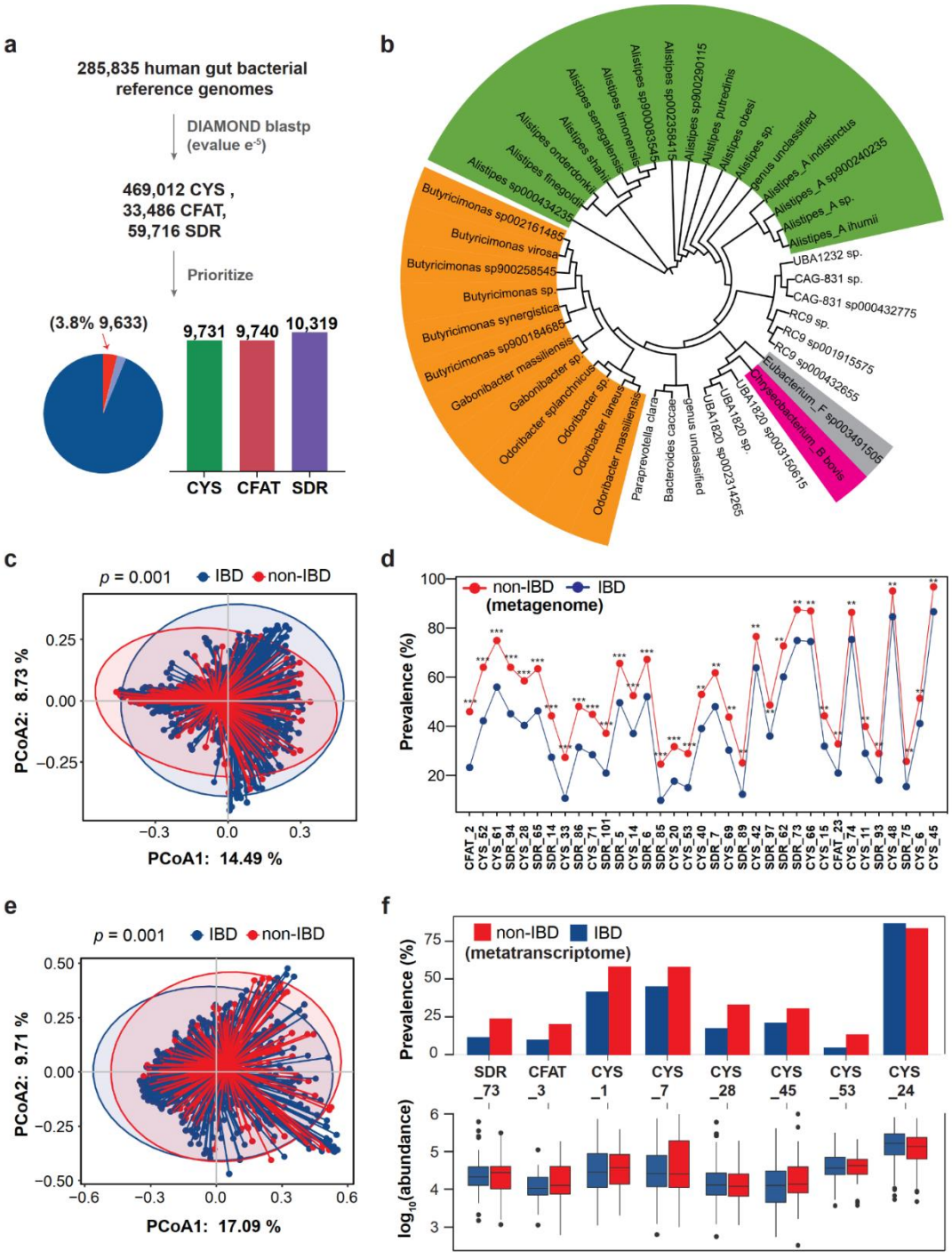


Figure 3.3 The presence and expression profiles of SoL biosynthetic enzymes and the production of SoLs differ in IBD subjects versus healthy controls. **a**, Overview of SoL biosynthetic enzymes identified in human gut bacteria. 562,214 putative SoL biosynthetic enzymes were identified across 21 bacterial phyla. 6.21% of genomes encode 3 types of SoL biosynthetic enzymes (Pie chart, sections in red and purple). Bar chart shows the number of prioritized SoL biosynthetic enzymes encoded by 9,633 genomes (highlighted in red in the pie chart). **b**, A circular phylogenetic tree shows the prioritized SoL

biosynthetic enzymes found primarily in species from Bacteroidota (highlighted in green and orange). The tree is annotated with species names and colored by taxonomic families (*Rikenellaceae*: green; *Marinifilaceae*: orange; *Weeksellaceae*: pink; *Lachnospiraceae*: grey). **c**, Principal Coordinate Analysis (PCoA) shows differences in the presence profile of overall SoL biosynthetic enzyme subfamilies between IBD and non-IBD groups based on Jaccard distance. **d**, 35 SoL biosynthetic enzyme subfamilies were significantly more prevalent in healthy individuals (red dots) than in IBD groups (blue dots) (two-sided Fisher's exact test, $p < 0.05$) with a difference of prevalence > 10%. For all p values: * $0.01 < p < 0.05$, ** $0.001 < p < 0.01$, and *** $p < 0.001$.) **e**, PCoA shows the differences in the expression profile of overall SoL biosynthetic enzyme subfamilies between IBD and non-IBD groups based on Bray-Curtis distances. **f**, Expression profiles of 8 differential SoL biosynthetic enzyme subfamilies (two-sided Mann-Whitney U test, adjusted $p < 0.05$). Upper panel: bar charts showing the prevalence of differential SoL biosynthetic enzyme subfamilies across non-IBD (red) and IBD individuals (dark blue). Statistical significance for prevalence was calculated using a two-sided Fisher's exact test. Except CYS subfamily24 (CYS_24, no significance), all were significantly higher in prevalence in non-IBD than IBD groups ($p < 0.05$). Lower panel: box plots displaying the abundance profiles of differential SoL biosynthetic enzyme subfamilies in non-IBD (red) and IBD individuals (dark blue). Significance was further determined by one-sided Mann-Whitney U test, with adjusted p-value < 0.05.

3.2.2 Multi-Omic Correlation of SoL Biosynthesis with IBD

To determine whether there is a link between gut microbial capacity of producing SoLs and IBD incidence, we conducted a comparative analysis of metagenomic and metatranscriptomic data obtained from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB, https://ibdmdb.org/)[158,171]. We began by generating sequence similarity networks with a 90% sequence identity threshold to group enzymes with similar functions. Consequently, we categorized the prioritized biosynthetic enzymes into 214 subfamilies (79 CYS subfamilies; 25 CFAT subfamilies; and 110 SDR subfamilies) for the subsequent analyses (Supplementary Table 3.2). Looking for the presence of the prioritized 214 subfamilies in IBD cohorts, we identified 154 subfamilies in 667 metagenome samples (182 healthy samples and 485 IBD disease samples), of which 116 subfamilies were detected in ≥ 5% of samples (Figure 3.2c). Beta diversity of the presence of these 116 subfamily biosynthetic enzymes indicated that the overall composition of SoL biosynthetic

enzyme subfamilies was significantly different between the healthy and IBD cohorts (Figure 3.3c, Jaccard distance, PERMANOVA $p = 0.001$). Of note, 57 subfamilies had a significantly higher prevalence (Fisher's exact test $p < 0.05$) in healthy individuals as compared to IBD cases (Supplementary Table 3.3), among which 35 subfamilies (18 CYS subfamilies, 2 CFAT subfamilies, and 15 SDR subfamilies) further show a difference of prevalence > 10% (Figure 3.3d).

To further examine the difference between the expression profiles of SoL biosynthetic enzymes between the IBD and healthy groups, we extended our comparative analysis to the metatranscriptomic level. We found that 132 SoL biosynthetic enzyme subfamilies were expressed in 777 metatranscriptomic samples (193 healthy samples and 584 IBD disease samples), with about 42% (55/132) detected in at least 5% of samples (Figure 3.2d). Beta diversity of the expression profiles of SoL biosynthetic enzymes suggested that the overall expression of these enzyme subfamilies is significantly different between the healthy and IBD cohorts (Fig. 3.3e, Bray-Curtis distance, PERMANOVA $p = 0.001$). To capture more detail, we compared the prevalence and abundance differences of each enzyme subfamily in the metatranscriptomic samples. Nine subfamilies had higher prevalence (Fisher's exact test $p < 0.05$, varying from 9% ~ 17%) in the non-IBD group than the IBD group (Supplementary Table 3.4). We further identified 8 subfamilies (6 CYS, 1 CFAT, and 1 SDR) as significantly different in abundance (expression) profiles between the healthy controls and IBD cases (Figure 3.3f, two-sided Mann-Whitney *U* test, adjusted $p < 0.05$). Notably, 7 of the 8 subfamilies had a higher prevalence (Figure 3.3f, upper panel, Fisher's exact test $p < 0.05$) and a higher abundance (Figure 3.3f, lower panel,

one-sided Mann-Whitney *U* test, adjusted $p < 0.05$) in the non-IBD group than in the IBD group.

We finally looked for metabolomic evidence in the differences of detectable SoLs, the products of the biosynthetic enzymes mentioned above, among IBD and non-IBD groups from publicly accessible metabolomics datasets. We expected that the increased expression of SoL biosynthetic enzymes would correspond with increased abundance of stool SoLs in non-IBD groups compared to IBD groups after possible uptake by the host. Using metabolomics data from two independent datasets (dataset 1: IBDMDB[158,171], corresponding to the same dataset used for metagenomics and metatranscriptomics analysis; dataset 2: PRISM[172]), we identified metabolite features putatively corresponding to specific sulfonolipids[156,173] (Supplementary Table 3.5), by exact mass comparison with mass error less than 5 ppm (Supplementary Data 3.3). Within each dataset individually, we indeed found that metabolomic features potentially corresponding to SoLs were decreased in stool samples of IBD groups compared to non-IBD groups, since there were no MS/MS data available in either dataset, we additionally utilized complementary approaches to confirm these features as SoLs including analysis of in-source fragmentation, correlation of co-eluting metabolomic features, and retention time matching between the dataset and data recreated using our own instrument, followed by experimental validation using targeted metabolomics with an additional set of independent IBD and non-IBD cohorts.

To validate the presence of SoLs in these datasets, we first tried to identify SoLs using in-source fragments (ISFs) of metabolites based on an established set of criteria[174]. We initially identified six groups of co-eluting metabolomic features as potential ISFs which showed peak-to-peak intensities highly correlated with putative SoL features (Figure

3.4, Pearson correlation coefficients ≥ 0.9, $p < 0.05$). We then examined the reference MS/MS spectra of our isolated and literature reported SoLs[161,173] matching putative SoL masses, which contained limited *m/z* values corresponding to potential ISFs we identified. However, their relatively low intensity was not conclusive enough to classify them as high-confidence ISFs[174]. Thus, we proceeded with a complementary correlational approach to identify the putative SoL features. Among the originally identified six groups of co-eluting metabolomic features, five members were detected in both datasets mentioned above (Figure 3.4c,d,e,f, features highlighted in bold). Based on exact mass matching, these features corresponded to SoL analogs: *SL 34:1;2O*, *SL 17:0;O/16:1;O*, *SL 33:1;2O|SL 17:0;O/16:1;O*, *SL 34:1;2O|SL 17:0;O/17:1;O*, and *SL 32:0;O|SL 17:0;O/15:0*[175]. In addition, these features had higher peak area-peak area correlation with each other (Pearson correlation ≥ 0.9, $p < 0.05$; Figure 3.4c,d,e,f). Notably, SoLs are often detected in metabolomics as a series of analogs with consecutive additions of CH_2 and H_2 moieties within the class and with different numbers of oxygens between classes[156,175]. Thus, these features likely represent a series of analogs chemically modified from a common parent metabolite, or co-produced by a specific microbe, which is consistent with SoL analogs. These metabolomic features were further positively correlated with species of the prolific SoL-producing genus *Alistipes*: *A. putredinis*, *A. finegoldii*, *A. indistinctus*, *A.* shahii, *A. onderdonkii*, and *Bacteroidales* bacterium ph8 (which belongs to *A. obesi*) with Spearman correlation coefficients ≥ 0.5 ($p < 0.05$; Figure 3.5a,b). This positive correlation indicated that the abundance of these metabolites increased with the increase in these species, supporting that these species likely produced these molecules. Furthermore, these metabolomic features had significantly higher abundance in non-IBD groups than IBD

groups in both IBD datasets (Figure 3.5c and Figure 3.6, Wilcoxon rank sum test, $p < 0.05$, one-sided), consistent with our exact mass matching analysis.

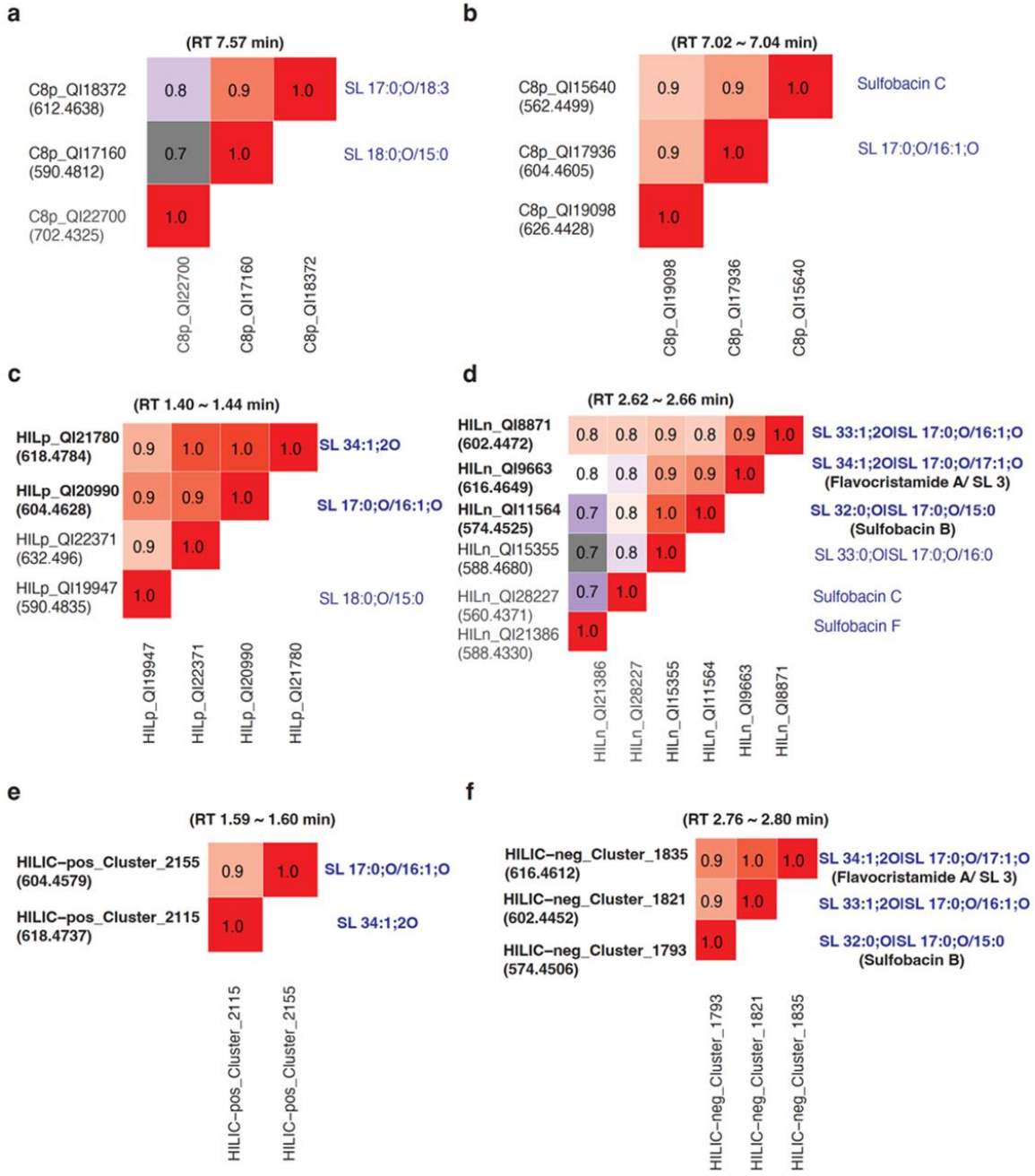


Figure 3.4 Feature similarity networks correlating SoL candidates within co-eluting MS1 groups. Associations shown all had Pearson correlation coefficients ≥ 0.9 and $p < 0.05$. Metabolomic features (left, bold) are listed with mass-to-charge ratios noted in parentheses and corresponding SoL annotations (right, blue; mass error ≤ 5 ppm, except for

HILn_QI9663 assigned as SL 34:1;2O|SL 17:0;O/17:1;O with 5.33 ppm). Feature similarity networks shown in (**a**), (**b**), (**c**), and d were derived from dataset 1 (IBDMDB), and the networks shown in (**e**) and (**f**) are from dataset 2 (PRISM). Each metabolomic feature was analyzed individually in each cohort. RT: retention time. The prefixes of metabolic feature names correspond to the detection method used: In dataset 1, C8p: C8-positive, HILp: HILIC-positive, and HILn: HILIC-negative. In dataset 2, HILIC-pos: HILIC-positive and HILIC-neg: HILIC-negative. Detailed descriptions of HPLC-MS methods can be found in the original studies. The feature corresponding to SL 34:1;2O correlated with SL 17:0;O/16:1;O and SL 34:1;2O|SL 17:0;O/17:1;O correlated with SL 33:1;2O|SL 17:0;O/16:1;O and SL 32:0;O|SL 17:0;O/15:0.

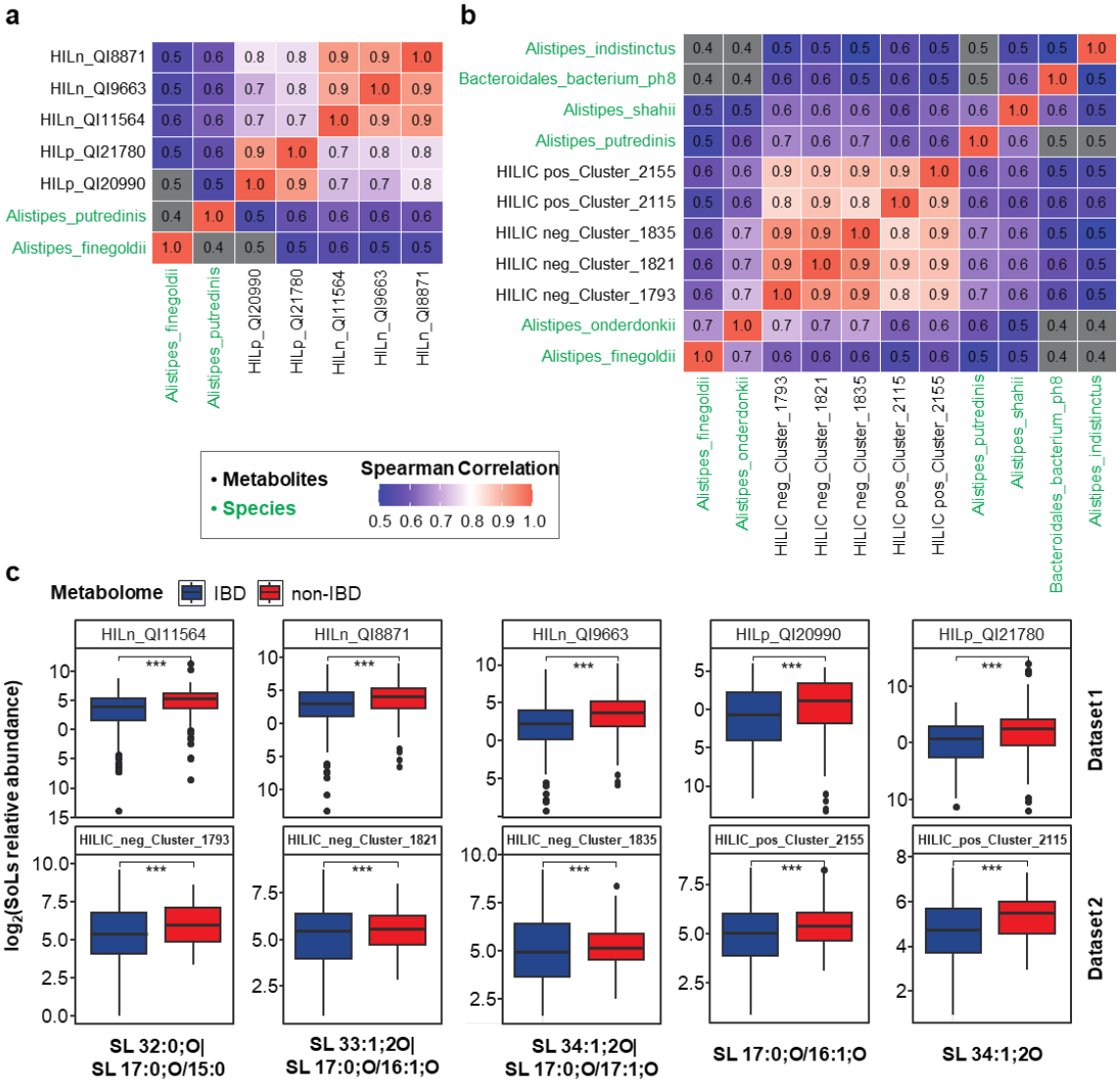


Figure 3.5 Putative SoLs candidates in IBD cohorts. Potential associations between microbes and putative SoL candidates in datasets 1 (**a**, IBDMDB) and 2 (**b**, PRISM). Spearman correlation coefficients were calculated using the relative abundance of microbes and metabolomic features. All Spearman correlations had $p < 0.05$. Species were

colored in green. **c**, Box plots showing the relative abundance of SoL candidates detected in non-IBD and IBD individuals collected from dataset 1 (upper) and dataset 2 (lower). Box plots include center lines representing the median, box limits representing upper and lower quartiles, whiskers representing the 1.5x interquartile range, and points representing outliers. Significance was determined by the one-sided Wilcoxon rank sum test with the hypothesis that the abundance of SoL was higher in the non-IBD than in IBD group. Significance shown with * $0.01 < p < 0.05$, ** $0.001 < p < 0$ .01, and *** $p < 0.001$. For each feature, the corresponding SoL is noted in the bottom label. The prefixes of metabolomic feature names correspond to the detection method used: In dataset 1, HILp indicates HILIC-positive method and HILn indicates HILIC-negative method. In dataset 2, HILIC-pos indicates HILIC-positive method and HILIC-neg indicates HILIC-negative method. Details for the corresponding HPLC-MS methods can be found in the original studies.

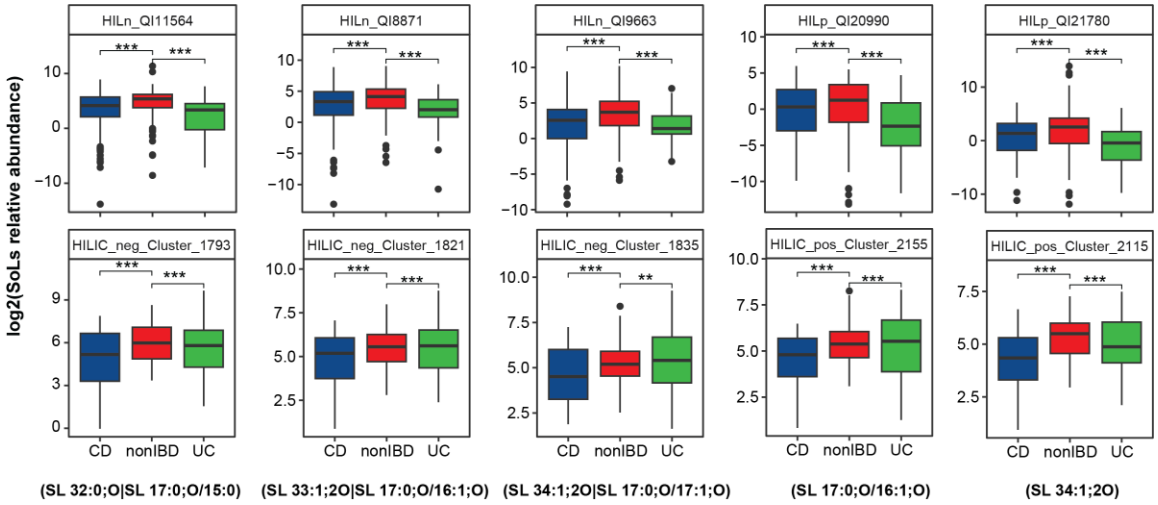
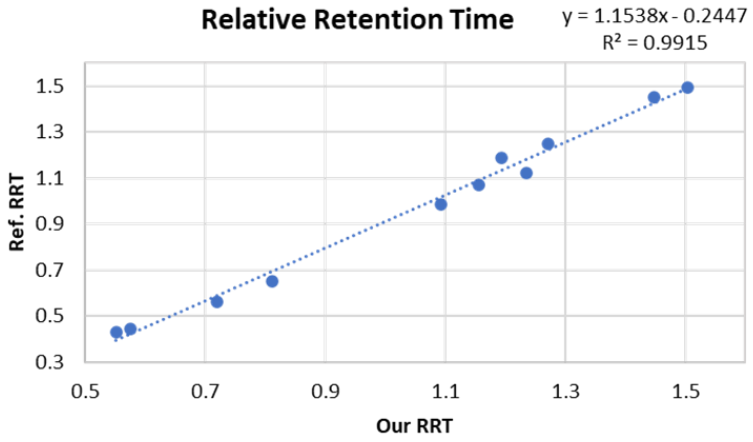


Figure 3.6 Abundance of SoL analogs in IBD patients vs. non-IBD cohorts from two independent metabolomic datasets. Box plots showing the relative intensity of SoLs candidates detected in non-IBD and IBD subtype (ulcerative colitis, UC; and Crohn's disease, CD) individuals collected from dataset 1 (IBDMDB, upper) and dataset 2 (PRISM, lower). Significance was determined by Wilcoxon rank sum test: * $0.01 < p < 0.05$, ** $0.001 < p < 0$ .01, and *** $p < 0.001$.

To further validate our identification of SoLs in these datasets, we acquired one of the columns used to generate the original data[158,172]. We then selected several standard compounds used in dataset 2 and the candidate SoL B feature that we identified by exact mass matching, and subsequently analyzed the retention times of the standard compounds alongside our own standard SoL B using our in-house high-performance liquid

chromatography-mass spectrometry (HPLC-MS) instrument. Due to the inherent variability of retention time between instruments[176,177], we calculated the relative retention time (RRT)[178,179] using each of the standard's retention time relative to that of our SoL B standard and compared these values to the RRTs calculated using the corresponding dataset standards and candidate SoL B. Indeed, we found that the RRT values using our SoL B standard and the RRT values using the candidate SoL B shared a linear relationship (Figure 3.7, $R^2 = 0.9915$), supporting that the shift in retention time was linear and thus suggesting that the candidate SoL B feature was SoL B.



Standard	Ref RRT	Our RRT
Uracil	0.552	0.428
Riboflavin	0.577	0.444
Sorbitol	0.719	0.562
Adenine	0.811	0.653
Proline	1.093	0.988
Phenylalanine	1.157	1.069
Pentadecanoate	1.194	1.187
Tryptophan	1.235	1.123
TCDCA	1.272	1.25
DCA	1.448	1.451
CA	1.504	1.496

Figure 3.7 Comparison of relative retention time for SoL B identity validation. Standard compounds were pooled together and the mixture analyzed by HPLC-MS using the same HPLC method and column as described in the original dataset publication[172]. MS exact mass matching and MS/MS fragmentation was used to identify each standard compound to determine the retention time (RT) as described in the original dataset[172]. Relative retention time (RRT) was calculated by dividing the standard compound RT to either the candidate SoL B RT in the original dataset or our in-house standard SoL B RT.

3.2.3 Validation in Human IBD Samples

Towards experimentally validating our informatic analysis, we obtained deidentified stool samples collected from an independent cohort of IBD patients (both UC and CD) and healthy controls, followed by analysis of their SoL abundance by HPLC-MS.

We detected SoLs B, C, and F as major SoLs and found all of their abundances decreased in IBD samples compared to non-IBD samples (Figure 3.8), consistent with our bioinformatic analysis which also showed that major SoLs including SoL B were significantly decreased in IBD metabolomes. This independent validation further supported our identification of SoLs in the metabolomics datasets and our chemoinformatic analysis showing decreased abundance of SoLs in stool samples of IBD.

Thus, our metagenomic analysis reflected that SoL biosynthetic enzymes were more prevalent in the non-IBD group than the IBD group, metatranscriptomics suggested that genes encoding these enzymes are more actively transcribed in the non-IBD group, and chemoinformatics and metabolomics indicated that representative SoLs are in higher abundance in stool samples from the non-IBD group. We further validated the metabolomics data in an independent cohort of IBD patient samples which showed that SoL abundance was indeed significantly decreased in IBD compared to non-IBD samples. Altogether, our findings establish a negative correlation directly between SoLs biosynthesis and IBD, consistent with the previously reported negative association between SoL-producers, namely *Alistipes* and *Odoribacter*, and IBD[22,23].

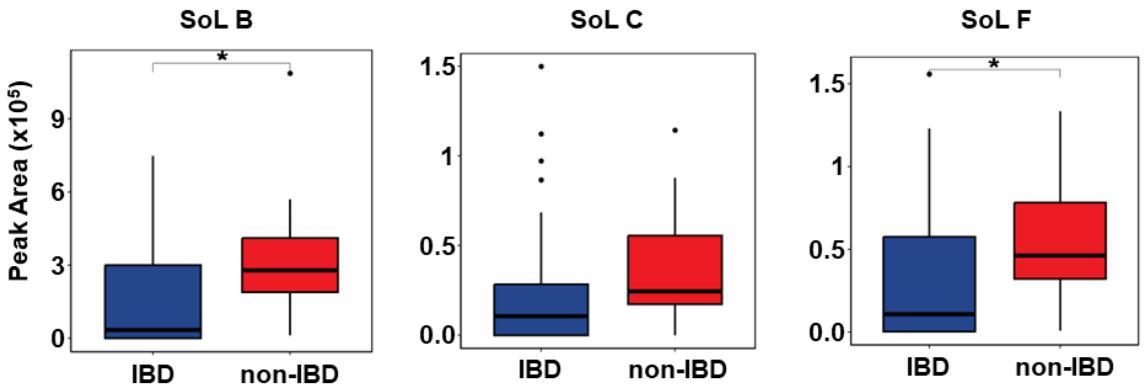


Figure 3.8 Analysis of SoL abundance in an independent cohort of IBD patient samples. Fecal samples were obtained from de-identified IBD patients representing both UC and CD, here unified as one IBD group, as well as healthy non-IBD patients. Lipids were extracted from the samples and prepared at a constant 200 mg/mL for analysis by HPLC-MS. SoLs B, C, and F were identified in the MS data by MS1 exact mass matching and MS/MS fragmentation comparison with reference spectra and in-house standards. Peak areas were used to quantify the change in SoL abundance between IBD and non-IBD groups showing significant decreases in SoL B and F in IBD patients compared to non-IBD patients, consistent with our chemoinformatic analysis of two independent IBD datasets.

3.2.4 Validation in a Mouse Model of Colitis

We further experimentally validated the link between SoL biosynthesis and IBD using a well-established mouse model of IBD. We used *Il10*-deficient (*Il10*$^{-/-}$) mice that are genetically susceptible to developing intestinal inflammation and chronically treated them with the non-steroidal anti-inflammatory drug piroxicam, which induces the development of colitis through the disruption of the gut mucosal barrier in inflammation susceptible hosts[180,181]. We selected this model due to its stability, as *Il10*$^{-/-}$ mice generally will not develop colitis when born and raised under specific pathogen free conditions unless induced by external stimuli such as piroxicam treatment. This allowed us to more confidently ensure that the effects observed were dependent on the induction of colitis and not due to the *Il10* deficiency. Stimulation of mucosal Toll-like receptors (TLRs) stemming

from mucosal barrier breakdown was another factor in our selection of this model, as we have previously shown that SoL A suppresses LPS-induced inflammation and LPS is well-known to activate TLR signaling[161,182]. As has been previously reported[180,181], we observed that the colonic tissues were inflamed in the piroxicam-treated (IBD) group of *Il10*$^{-/-}$ mice when compared to the control (pre-IBD) group as indicated by gross pathology and blinded histopathology analyses (Figure 3.9a,b,c, Tables 3.2 and 3.3).

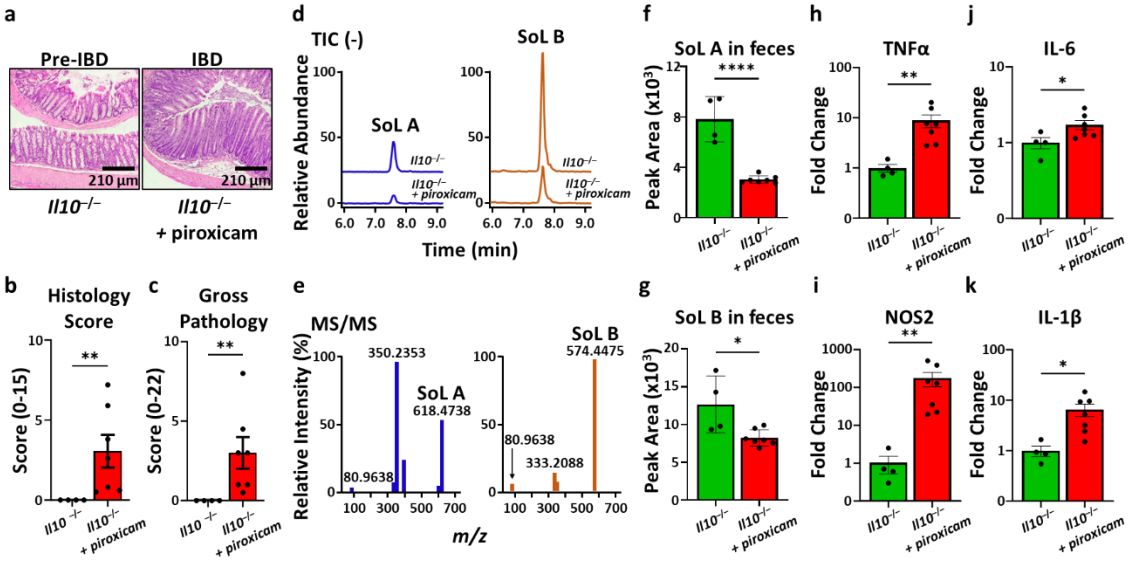


Figure 3.9 SoLs are decreased in a female mouse model of colitis concurrent with increased expression of inflammatory markers. **a**, Histological analysis of the mouse distal colon reveals that piroxicam treatment induced intestinal inflammation in *Il10*$^{-/-}$ mice. **b** and **c**, Histology and gross pathology scores indicate induction of colitis in *Il10*$^{-/-}$ mice treated with piroxicam (n = 7, female) compared to pre-IBD control *Il10*$^{-/-}$ mice (n = 4, female), confirming the successful establishment of the IBD model. The trends were consistent in male mice in another independent cohort using the same IBD model (Figure 3.10). **d**, Total ion chromatograms (TICs) obtained from HRMS analysis of fecal pellet extracts from control *Il10*$^{-/-}$ mice and *Il10*$^{-/-}$ + piroxicam mice reveal the presence of SoL A and SoL B. SoL abundances appear to be decreased in *Il10*$^{-/-}$ + piroxicam mice fecal pellets. **e**, MS/MS spectra of SoLs A and B confirm their identities based on the presence of the 80 *m/z* fragment characteristic of sulfonate-containing compounds as well as other characteristic fragments (Figure 3.11a,b) and compared to literature fragmentation patterns[156,161]. **f** and **g**, Peak areas were calculated using TICs obtained after MS/MS fragmentation and used to measure the abundance of SoLs A and B. Both SoLs A and B were significantly decreased in feces from inflamed mice. Significance was determined using Student's *t*-test. **h**–**k**, Gene expression of inflammatory markers TNFα, NOS2, IL-6,

and IL-1β were significantly increased in the ceca of *Il10*$^{-/-}$ + piroxicam mice. Significance was determined using Mann-Whitney *U* test. Error bars represent mean ± standard error. For all *p* values: ** $0.001 < p < 0.01$ and **** $p < 0.0001$.

Table 3.2 Histology scores for female mice in IBD mouse model.

Cage ID	Mouse ID	Mouse strain	Mouse sex	Diet	Crypt hyperplasia (0-4)	Crypt abscesses (0-4)	Edema (0-3)	Goblet cell loss (0-4)	Total score (0-15)
1	1	Il-10	F	Control	0.0	0.0	0.0	0.0	0.0
	2	Il-10	F		0.0	0.0	0.0	0.0	0.0
	3	Il-10	F		0.0	0.0	0.0	0.0	0.0
	4	Il-10	F		0.0	0.0	0.0	0.0	0.0
2	5	Il-10	F	Piroxicam 100 ppm	1.0	1.0	0.4	2.0	4.4
	6	Il-10	F		2.0	0.0	0.0	1.8	3.8
	7	Il-10	F		0.0	0.0	0.0	0.6	0.6
	8	Il-10	F		2.0	1.0	0.0	2.9	5.9
3	9	Il-10	F	Piroxicam 100 ppm	2.5	1.0	0.8	2.9	7.2
	10	Il-10	F		1.5	0.0	0.0	1.1	2.6
	11	Il-10	F		0.0	0.0	0.0	0.5	0.5
	12	Il-10	F		0.0	0.0	0.0	0.8	0.8

Table 3.3 Gross pathology scores for female and male mice in IBD mouse model.

Cohort	Cage ID	Mouse ID	Mouse Strain	Mouse Sex	Diet	CECUM				COLON			Gross Path Score
						Content Loss	Inflammation	Atrophy	Total	Stool Consistency	Inflammation	Total	
#1	1	1	Il-10	F	Control	0	0	0	0	0	0	0	0.0
		2	Il-10	F		0	0	0	0	0	0	0	0.0
		3	Il-10	F		0	0	0	0	0	0	0	0.0
		4	Il-10	F		0	0	0	0	0	0	0	0.0
	2	5	Il-10	F	Piroxicam 100 ppm	1	1	1	3	3	0.5	3.5	6.5
		6	Il-10	F		0	1	0	1	1	1	2	3.0
		7	Il-10	F		0	0	0.5	0.5	0	0	0	0.5
		8	Il-10	F		1	1	1.5	3.5	0	1	1	4.5
	3	9	Il-10	F	Piroxicam 100 ppm	2	1	2	5	1	2	3	8.0
		10	Il-10	F		0.5	0.5	1	2	0	1	1	3.0
		11	Il-10	F		0	0	0.5	0.5	0	0.5	0.5	1.0
		12	Il-10	F		0	0.5	0.5	1	0	0	0	1.0
#2	1	1	Il-10	M	Control	0	0	0	0	0	0	0	0.0
		2	Il-10	M		0	0	0	0	0	0	0	0.0
		3	Il-10	M		0	0	0	0	0	0	0	0.0
	2	4	Il-10	M	Piroxicam 100 ppm	0	1	0	1	1	3	4	5.0
		5	Il-10	M		1	1.5	2	4.5	2	4	6	10.5
		6	Il-10	M		1	1	1	3	1	4	5	8.0
	3	7	Il-10	M	Piroxicam 100 ppm	2	1	2	5	1	4	5	10.0
		8	Il-10	M		1	1	1	3	0	1	1	4.0

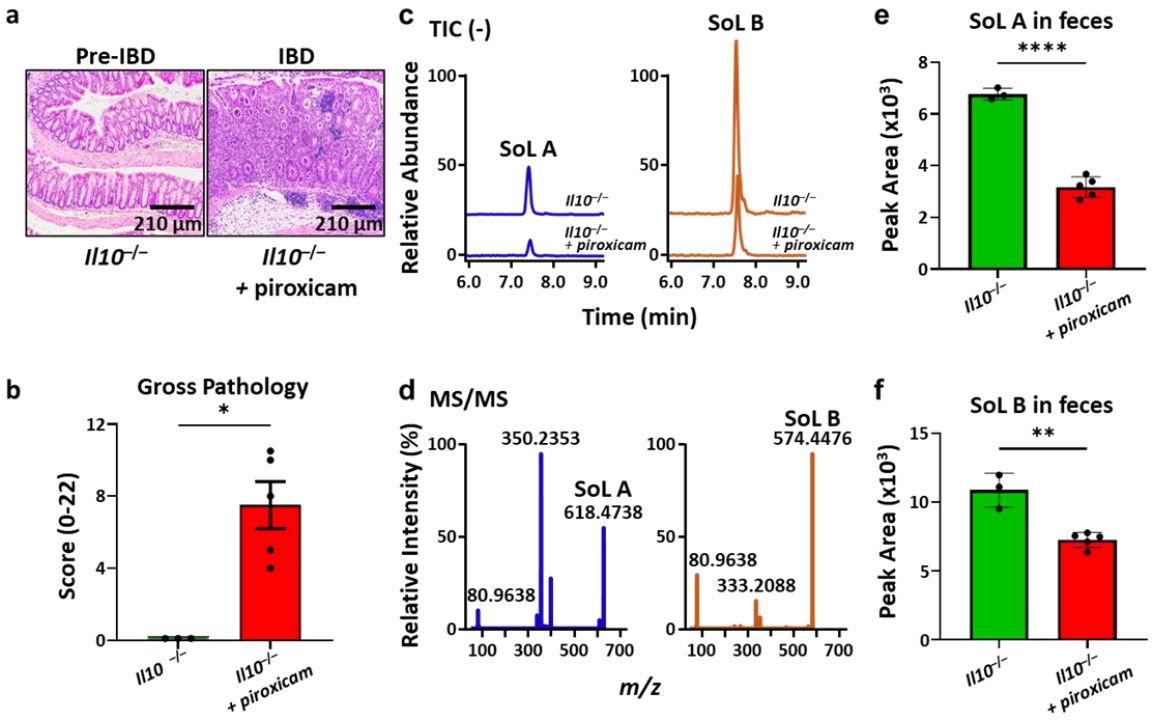


Figure 3.10 SoL abundance is decreased in a male mouse model of colitis. **a**, Histological analysis of the mouse distal colon shows piroxicam treatment induced intestinal inflammation. **b**, Overall gross pathology score further indicates induction of IBD in *Il10*$^{-/-}$ mice treated with piroxicam (n = 5, male) compared to pre-IBD control *Il10*$^{-/-}$ mice (n = 3, male). **c**, TICs obtained from HRMS analysis of fecal pellet extracts from *Il10*$^{-/-}$ and *Il10*$^{-/-}$ + piroxicam mice reveal the presence of SoL A and SoL B. SoL abundances were decreased in *Il10–/–* + piroxicam mice fecal pellets, consistent with female mice. **d**, MS/MS spectra of SoLs A and B confirm their identities based on the presence of the 80 *m/z* sulfonic acid fragment. **e** and **f**, Peak areas measuring the abundance of SoLs A and B. Both compounds were significantly decreased in IBD mice samples. Significance was determined using Student's t-test. Error bars represent mean ± standard error. For all p values: ** $0.001 < p < 0.01$ and **** $p < 0.0001$.

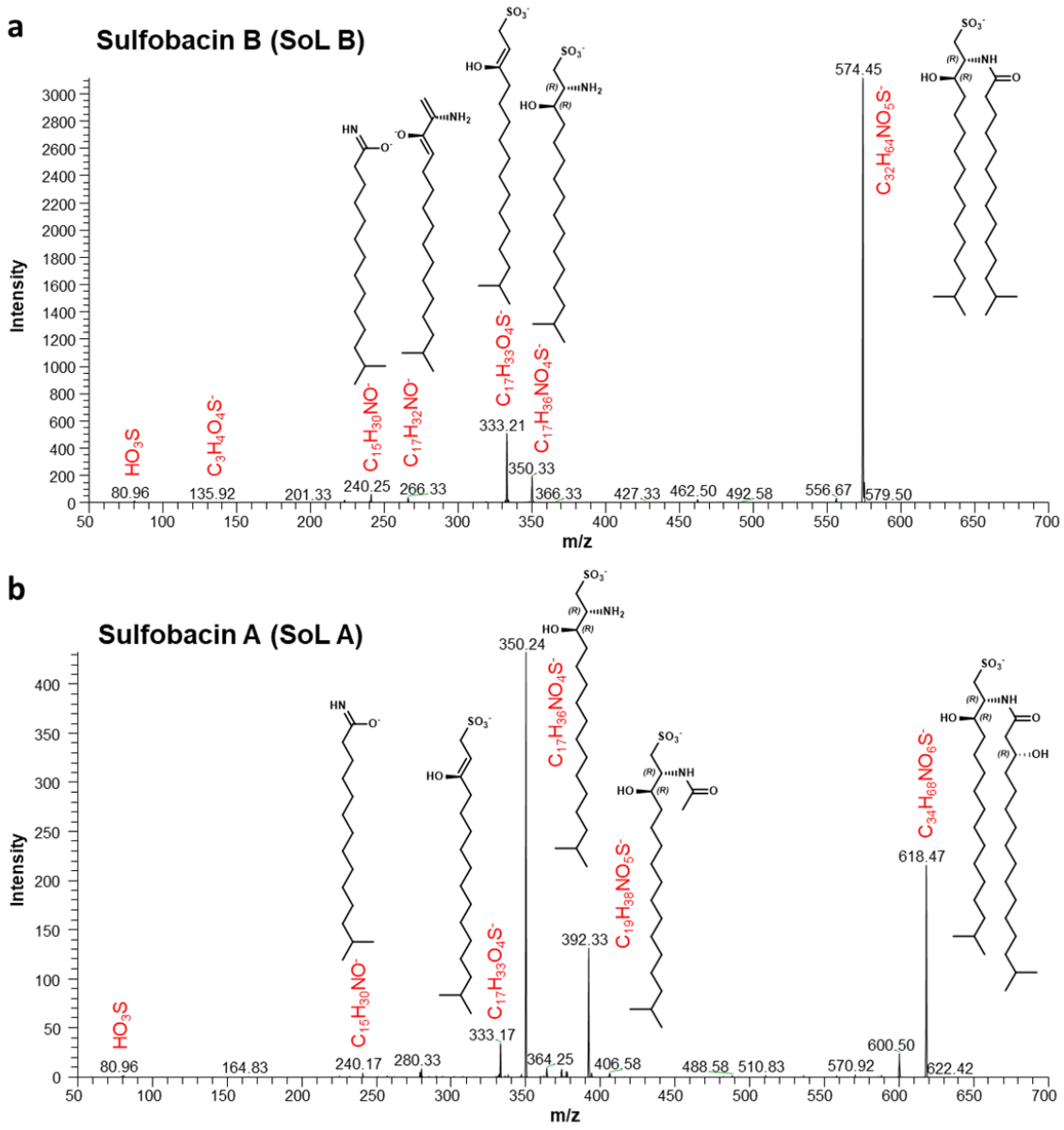


Figure 3.11 Fragmentation pattern of SoLs B and A. **a,b,** Fragmentation patterns for SoL B (**a**) and SoL A (**b**) were collected on a ThermoFisher Scientific LTQ XL using pulsed Q dissociation (PQD) at 35V. Major fragments of SoLs are labeled with their corresponding fragment structures and chemical formulas. All fragmentation patterns are consistent with previously reported fragmentations for SoLs[156].

To explore the link between SoLs production and IBD, we collected fecal material from piroxicam-treated *Il10*$^{-/-}$ (IBD) mice (n = 7, female) and pre-IBD control *Il10*$^{-/-}$ mice (n = 4, female)**,** extracted metabolites, and measured the abundance of SoLs by targeted

metabolomics using high resolution mass spectrometry (HRMS) (Figure 3.9d). We detected metabolites with *m/z* corresponding to major SoLs, specifically SoLs A and B, in all fecal samples tested and unambiguously determined their identities by HPLC-MS/MS (Figures 3.9e and 3.11a,b). We then determined that the abundances of both SoLs A and B were significantly decreased in feces from piroxicam treated mice compared to control (Figure 3.9f,g). This result confirms our above-described informatic analysis and directly establishes a negative correlation between SoLs production and colitis progression in the mouse model. In addition, we also observed significantly increased expression of the NF-κB-regulated inflammatory markers TNFα, NOS2, IL-6, and IL-1β in the IBD mouse group (Mann-Whitney *U* test, $p \leq 0.005$; Figure 3.9h-k), further indicating a negative correlation between SoL production and these inflammatory markers. Given our previously observed anti-inflammatory activity of SoL A against LPS[161], a natural ligand of TLR4, this negative correlation suggests a potential role of SoLs in regulating IBD that may involve suppressing TLR4-mediated NF-κB activation. To exclude any differences caused by sex, we performed another independent study with male mice using the same model and observed the same negative correlation between SoLs production and IBD progression (Figure 3.10).

3.3 Exploring the Biological Role of Sulfonolipids

3.3.1 SoL producers consistently contribute to anti-inflammatory activity

We next examined the production of SoLs and their contribution to immunomodulatory activity in different human gut commensals. Unlike *C. gleum* F93 DSM 16776, which we experimentally investigated for its functional metabolites in relation to inflammatory activity[161], the prolific SoL-producers *Alistipes* and *Odoribacter*

had not yet been thoroughly chemically investigated to identify the biologically active components associated with remediation of IBD. In addition, *Alistipes* and *Odoribacter* produce a mixture of other SoLs[156] and are likely to produce a multitude of other functional metabolites, both of which may complicate the potential immunomodulatory activity of these genera's metabolites with respect to their bioinformatically predicted negative association with IBD. Thus, we conducted bioactive molecular networking of three *Alistipes* and two *Odoribacter* strains (Table 3.4) to identify the constant contributor(s) to biological activity. We fractionated crude extracts of the *Alistipes* and *Odoribacter* strains and determined the biological activity of each fraction using a cell-based assay that measured the suppression of LPS-indued TNFα production (Figure 3.12a). We simultaneously analyzed each fraction by untargeted high resolution tandem mass spectrometry (HRMS/MS) to generate molecular networks using the Global Natural Products Social (GNPS) feature-based molecular networking (FBMN) pipeline[183]. We then correlated the relative expression of TNFα in each fraction with the relative peak area of molecular features across all fractions to generate a bioactivity score reflecting the contribution of specific features to the activity of the fractions. Bioactivity scores and relative peak areas were then mapped onto the molecular network to visualize these contributions. A representative bioactive molecular network generated from *Alistipes timonensis* DSM 27924 is presented in Figure 3.12b. The SoL-containing cluster contained the most abundant and most active molecular features, as indicated by the node size and color intensity compared to other clusters in the network. Additionally, this cluster contained several known SoLs but many more unannotated SoLs, suggesting that the family of biologically active SoLs is larger than what is currently known. In all other SoL-

producers tested, we consistently identified SoLs as a major contributor in the active fractions of each strain (Figure 3.13). To exclude the possibility of observed SoL activity being influenced by LPS contamination, we confirmed the absence of leftover LPS in the SoL samples using a chromogenic limulus amebocyte lysate (LAL) assay (Figure 3.12). Narrowing down the immunosuppressive activity of each of the strains to SoLs guided us to isolate pure SoLs A and B from *A. timonensis* DSM 27924 (structures confirmed by NMR spectroscopy; Tables 3.5 and 3.6, Figures 3.16–3.27), as well as from each of the other *Alistipes* and *Odoribacter* strains tested. We thus reinforced the contribution of this class of lipids to the observed biological activity of *Alistipes* and *Odoribacter*.

Table 3.4 SoL-producing strains used for bioactive molecular networking.

Bacterium	Strain	Strain Accession
Alistipes putredinis	Carlier 10204	DSM 17216
Alistipes timonensis	JC136	DSM 25383
Alistipes timonensis	CC-5826-WT-bac	DSM 27924
Odoribacter laneus	YIT 12061	DSM 22474
Odoribacter splanchnicus	1651/6	DSM 20712

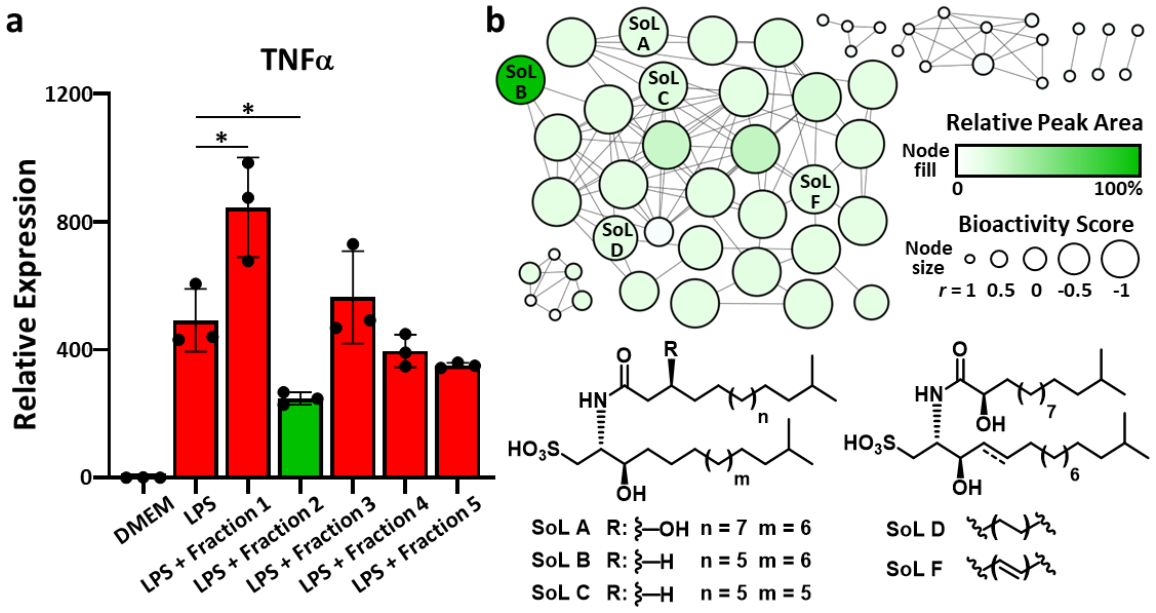


Figure 3.12 Bioactive molecular networking leads to the identification of SoLs as major bioactive components of a SoL-producer. **a**, A crude extract of *A. timonensis* DSM 27924 was separated based on polarity into 5 fractions. Each fraction was used in an *in vitro* cell-based assay to measure its respective capacity to suppress LPS-induced expression of TNFα. Fraction 2 was found to have the most significant anti-inflammatory effect compared to LPS. All fractions were compared to LPS for statistical significance with only fractions 1 and 2 showing significant change. Fractions 3, 4, and 5 showed no significant change. Statistical significance was determined using Student's t-test. Error bars represent mean ± standard error. For all p values: * $0.01 < p < 0.05$. **b**, Untargeted HRMS/MS was used to construct a molecular network for each fraction through GNPS FBMN. Relative peak area of each molecular feature in fraction 2 was mapped to the color of the nodes with more abundant features increasing from white to green. Bioactivity score was mapped to the node size with larger nodes indicating stronger negative correlations. Several known SoLs were annotated in this cluster and their structural variations are illustrated, further demonstrating that SoLs as a family of molecules contribute to the observed suppression of LPS-induced TNFα expression.

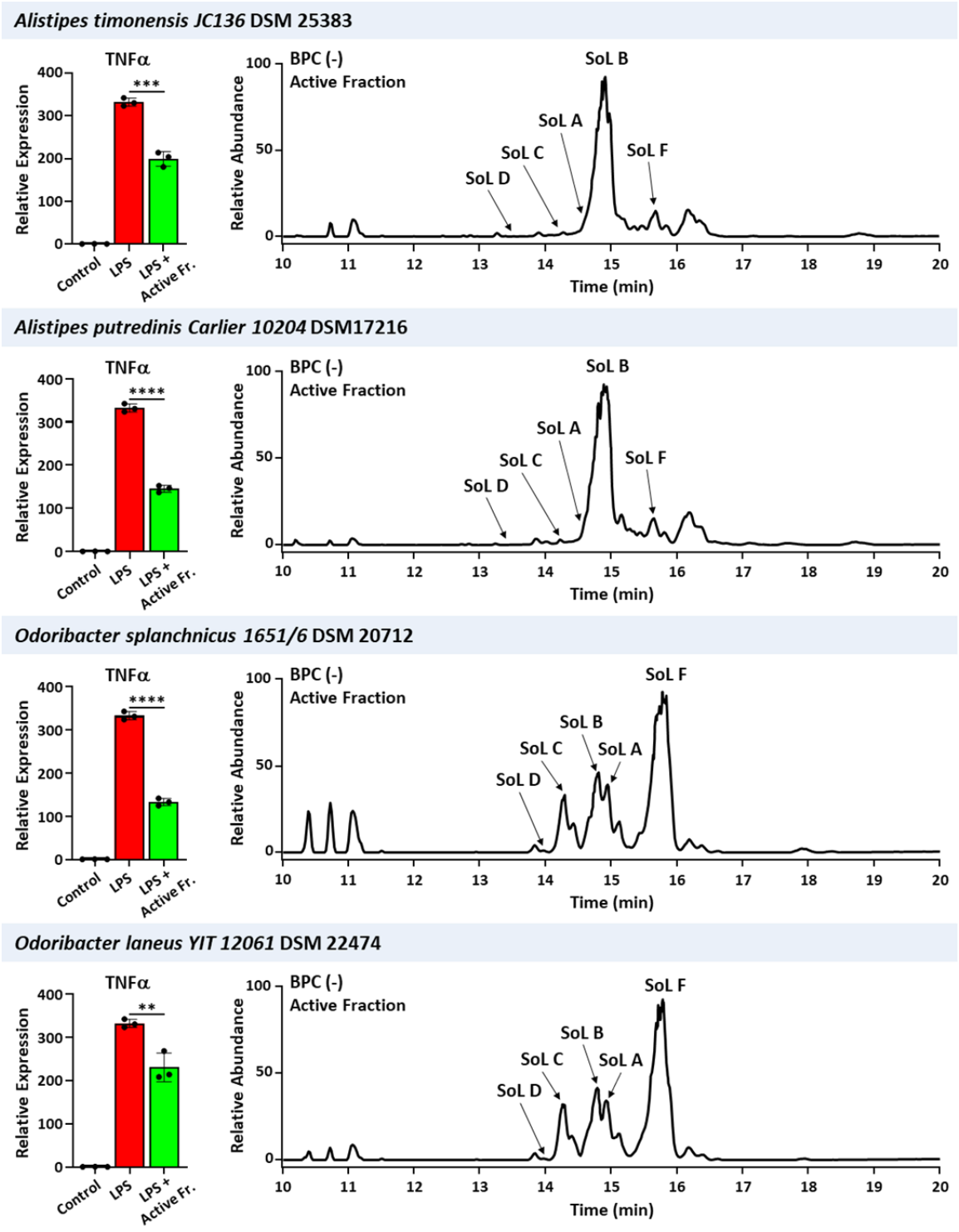
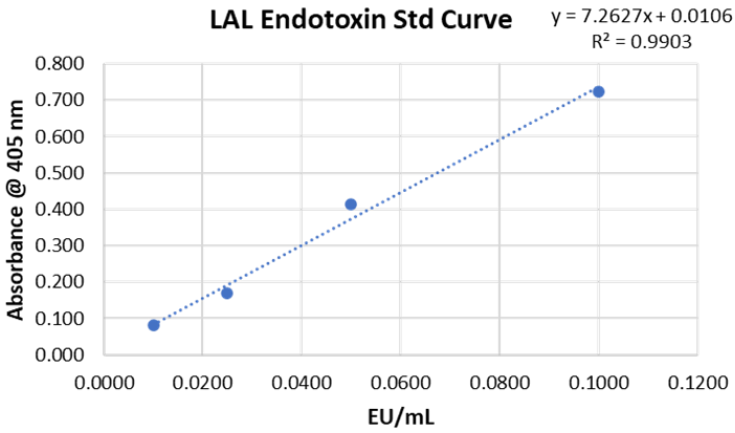


Figure 3.13 Biological activity and metabolomics screening of fractions from other SoL-producing strains. Four other SoL-producing strains were cultured, fractionated, and analyzed for biological activity using mouse macrophages and chemical composition by HRMS. In all strains, the active fraction significantly suppressed the relative expression of LPS-induced TNFa. Each active fraction was found to contain SoLs with fractions from

the two *Alistipes* species having SoL B as the major SoL and *Odoribacter* species having SoL F as the major SoL. Significance was determined using Student's t-test. Error bars represent mean ± standard error. For all *p* values: ** 0.001 < *p* < 0 .01 and **** *p* < 0.0001.



Sample	Abs - Blank	EU/mL
Std 1	0.723	0.1000
Std 2	0.414	0.0500
Std 3	0.168	0.0250
Std 4	0.081	0.0100
SoL A	**0.032**	**ND***
SoL B	**0.035**	**ND***

Figure 3.14 Chromogenic LAL assay of purified SoL A and B samples. A chromogenic limulus amebocyte lysate (LAL) assay was used to quantify endotoxin unit (EU) concentrations in SoLs A and B samples. SoL A and B were determined to be below the limit of detection for the assay with EU/mL < 0.01. Based on existing guidelines, this level of endotoxin was considered not to influence downstream assays[184].

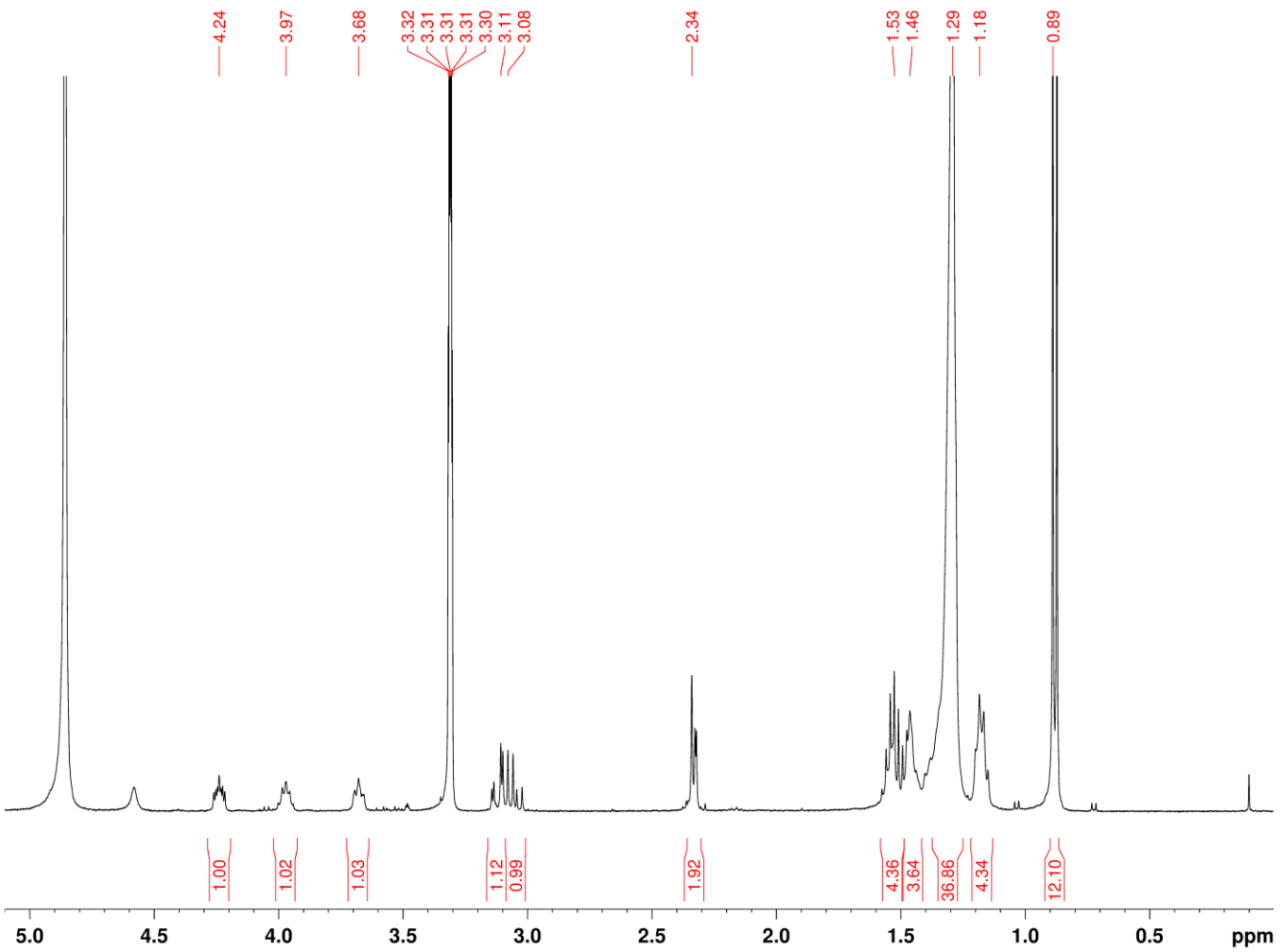


Figure 3.15 ^{1}H NMR (400 MHz, MeOD) spectrum of SoL A.

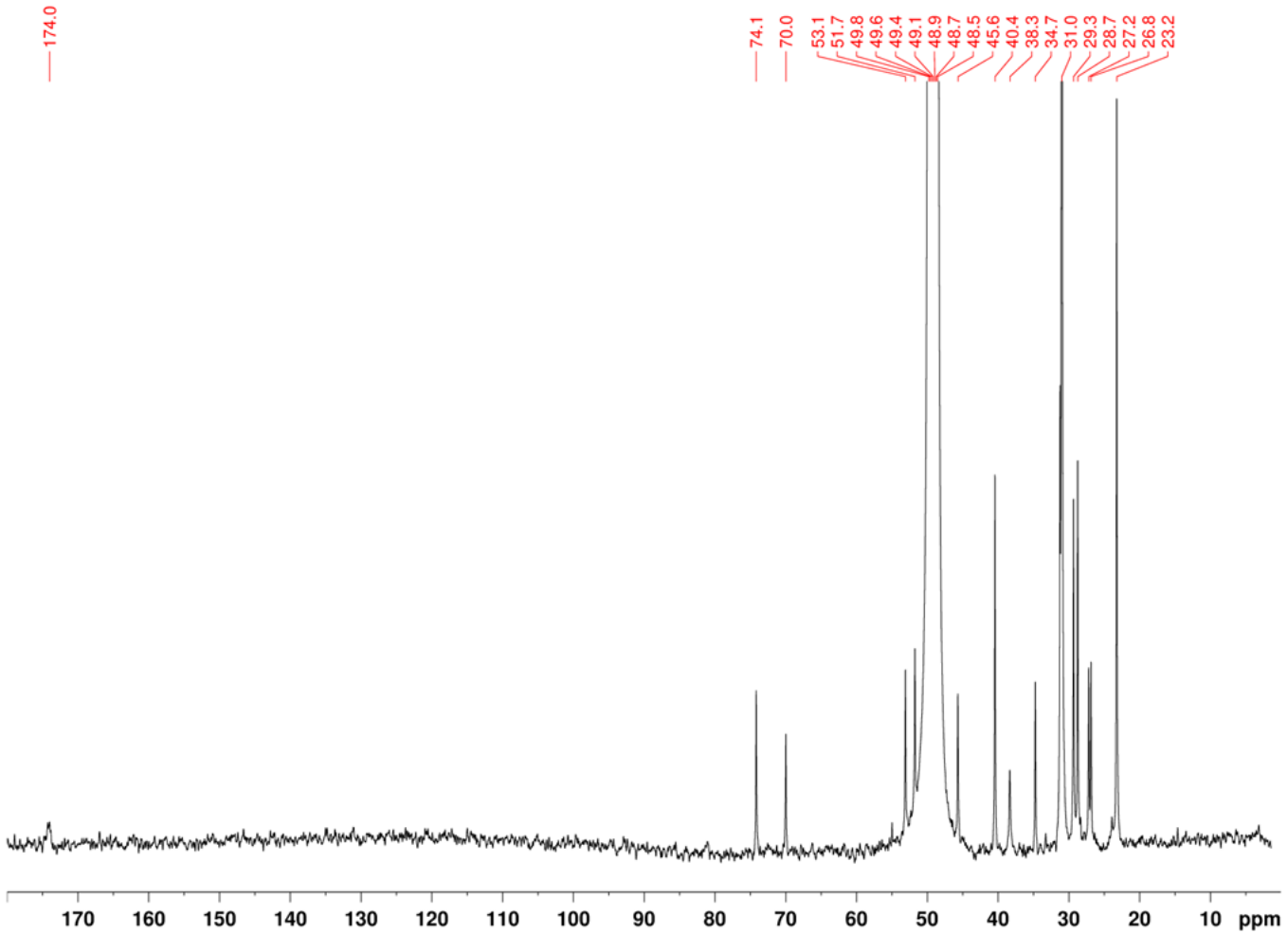


Figure 3.16 ^{13}C NMR (100 MHz, MeOD) spectrum of SoL A.

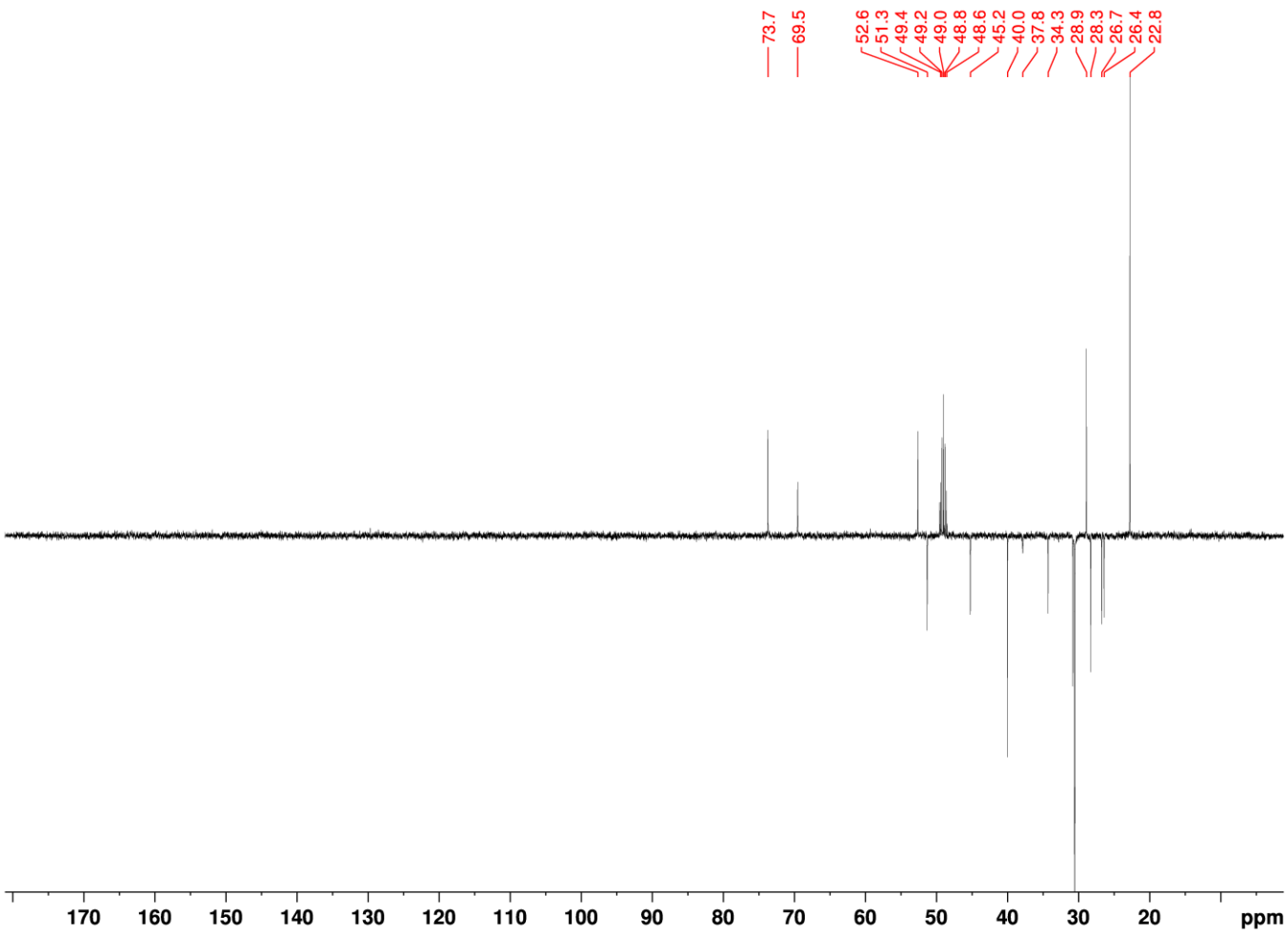


Figure 3.17 DEPT 135 NMR (100 MHz, MeOD) spectrum of SoL A.

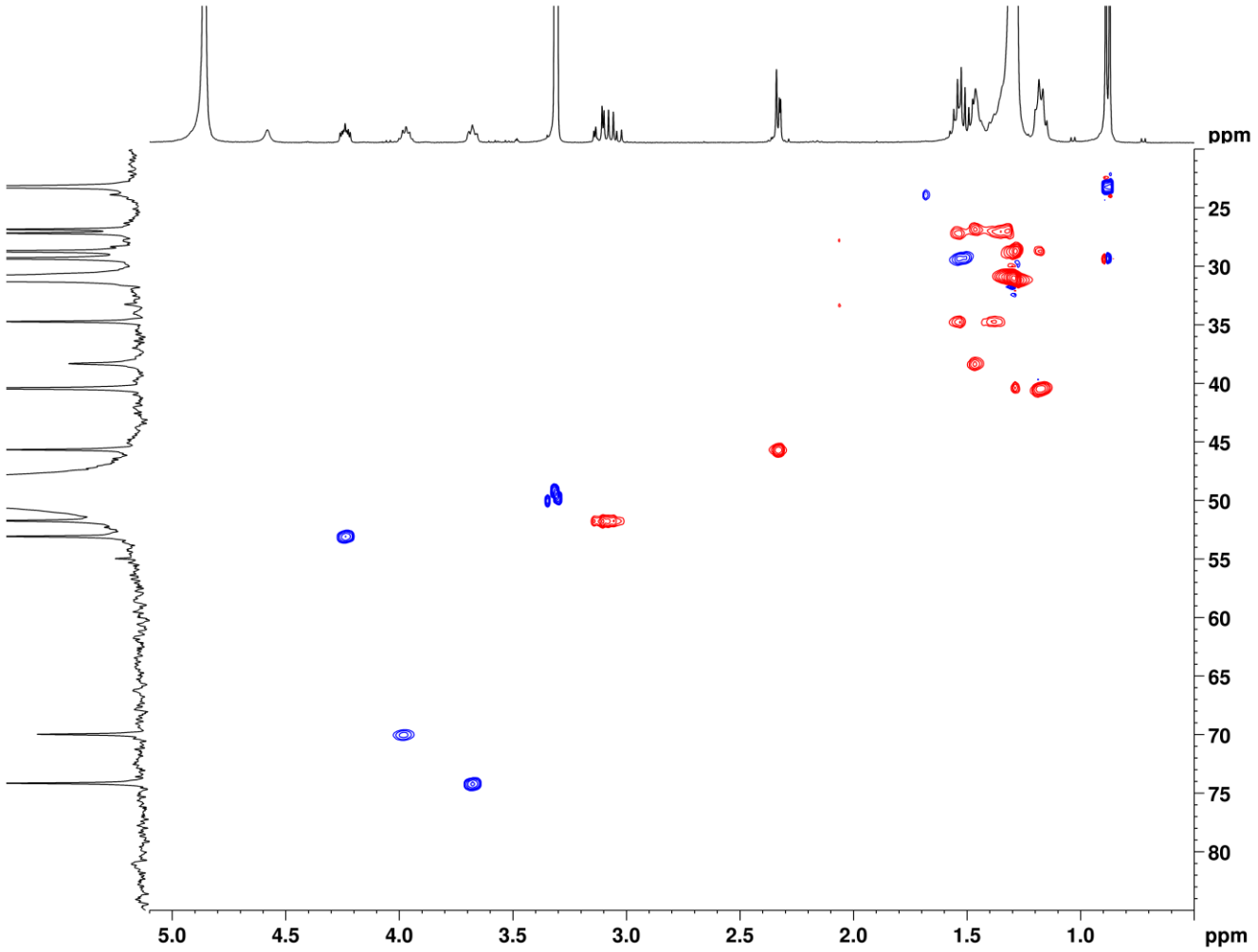


Figure 3.18 HSQC NMR (400 MHz, MeOD) spectrum of SoL A.

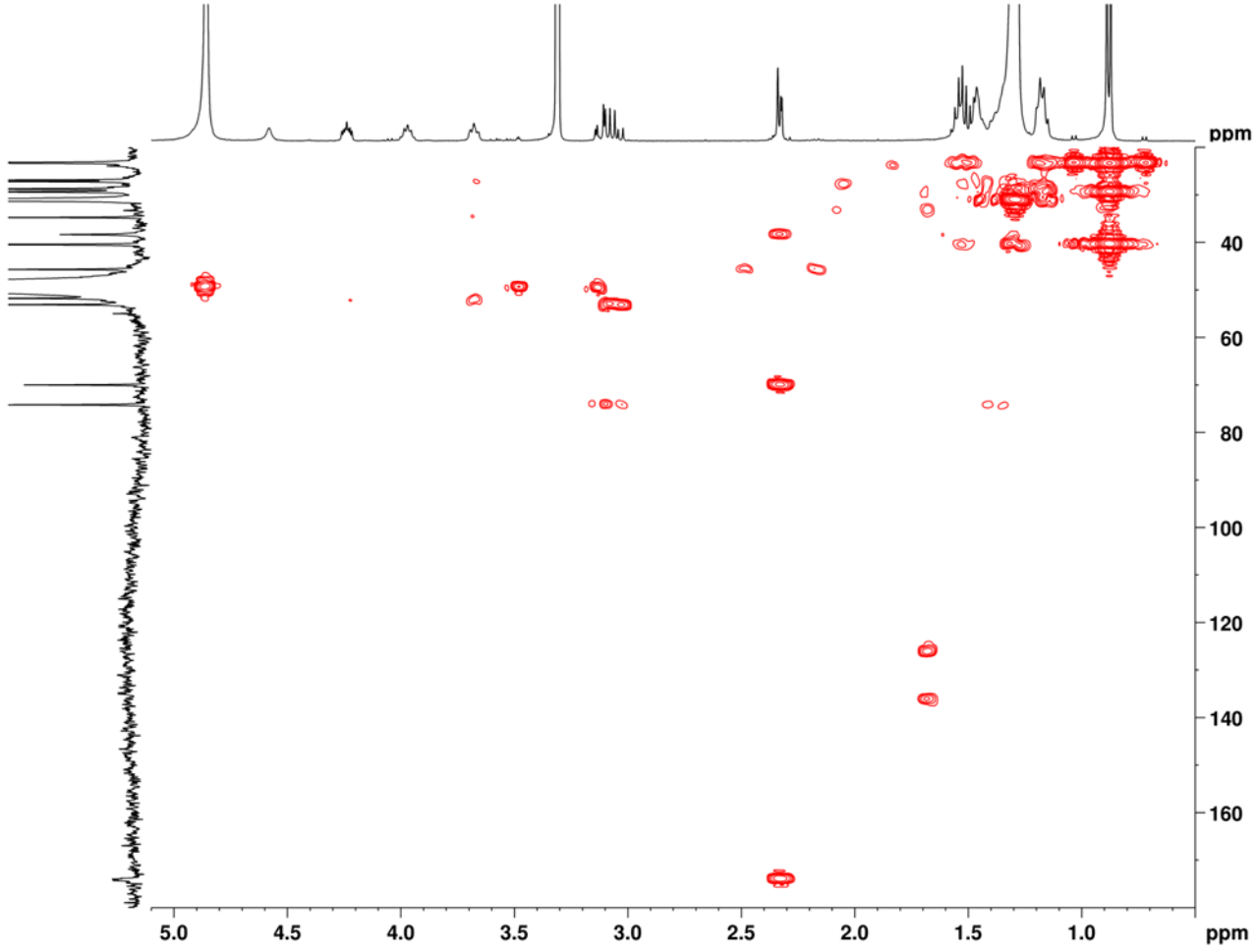


Figure 3.19 HMBC NMR (400 MHz, MeOD) spectrum of SoL A.

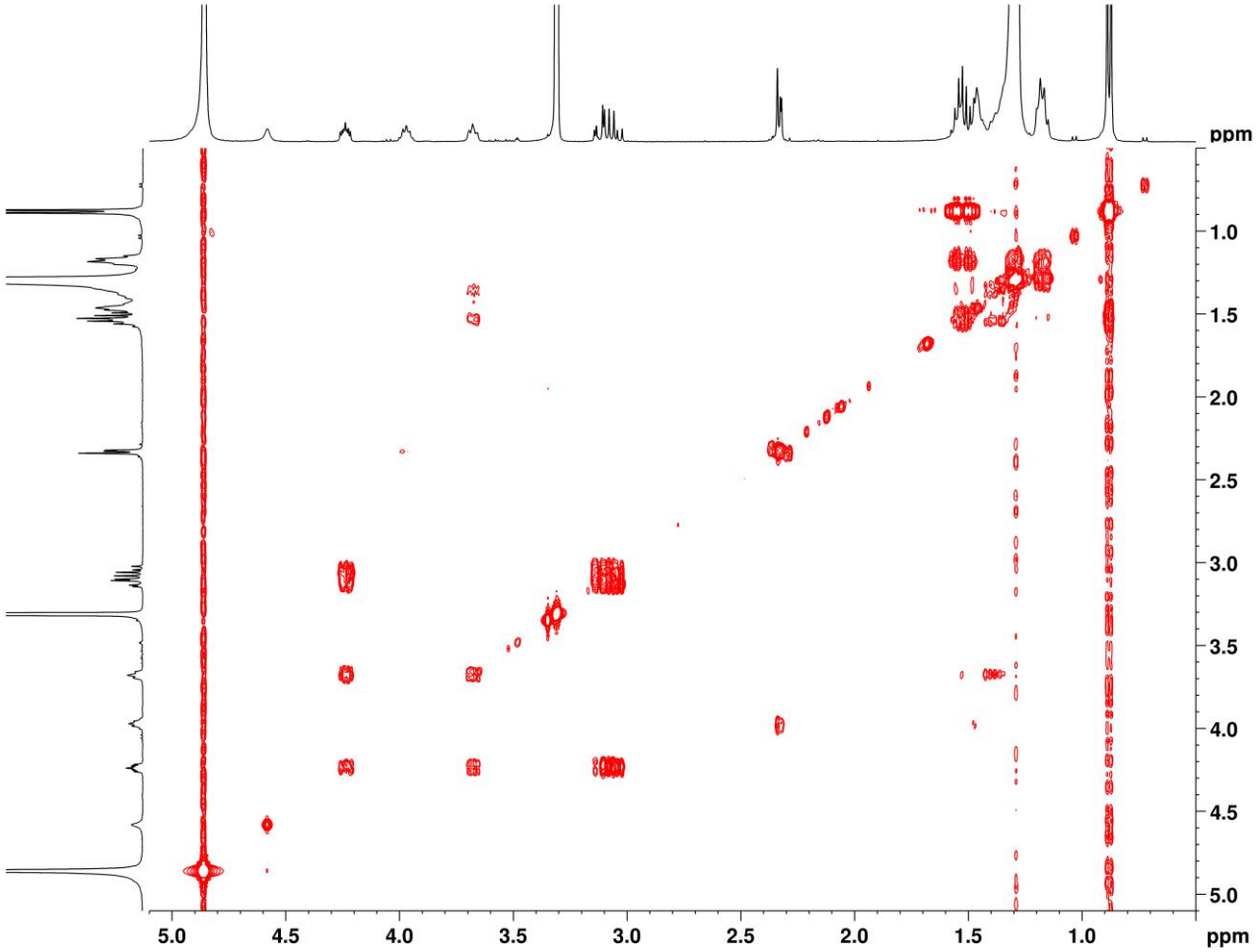


Figure 3.20 ^{1}H-^{1}H COSY NMR (400 MHz, MeOD) spectrum of SoL A.

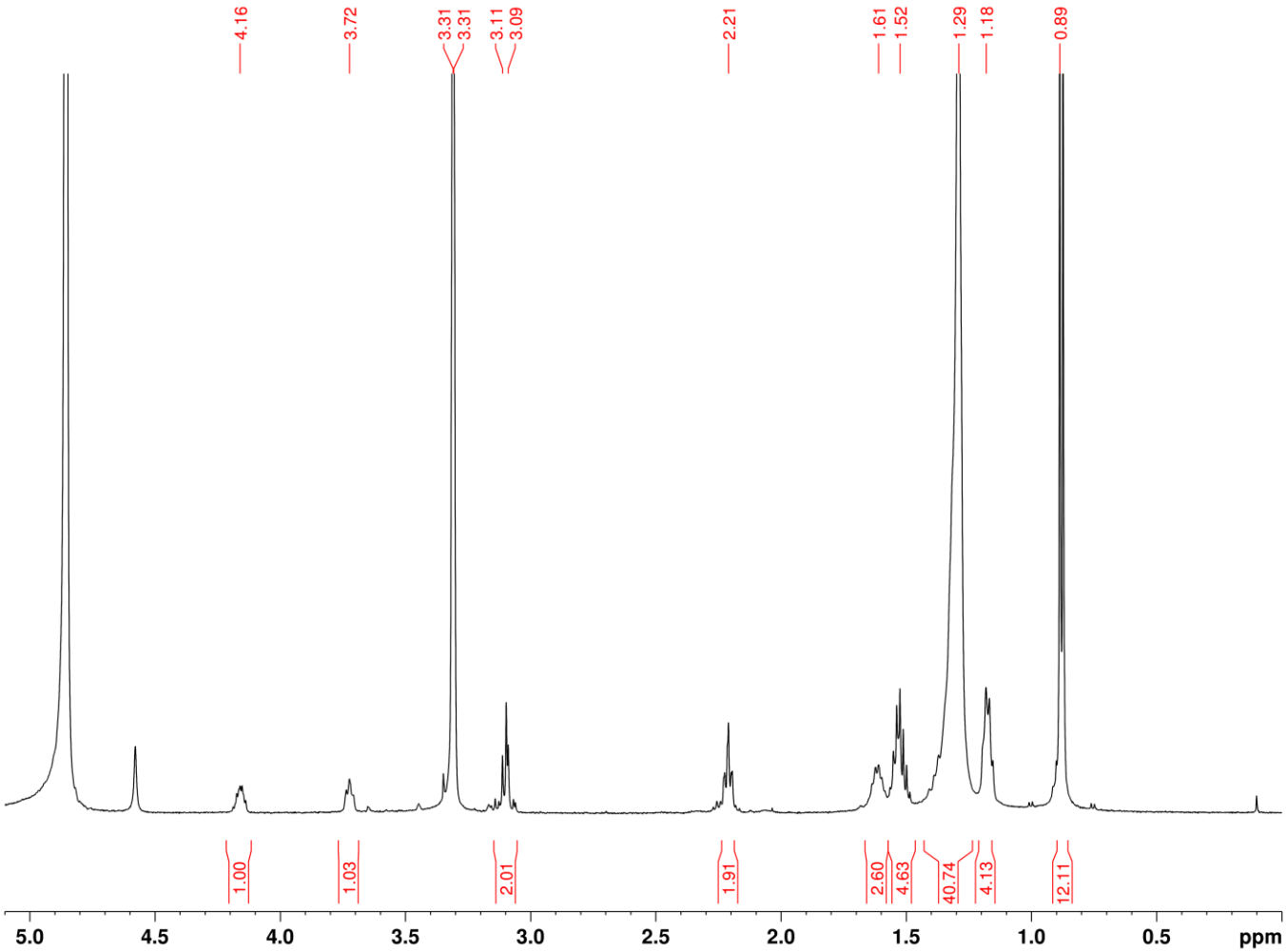


Figure 3.21 ^{1}H NMR (400 MHz, MeOD) spectrum of SoL B.

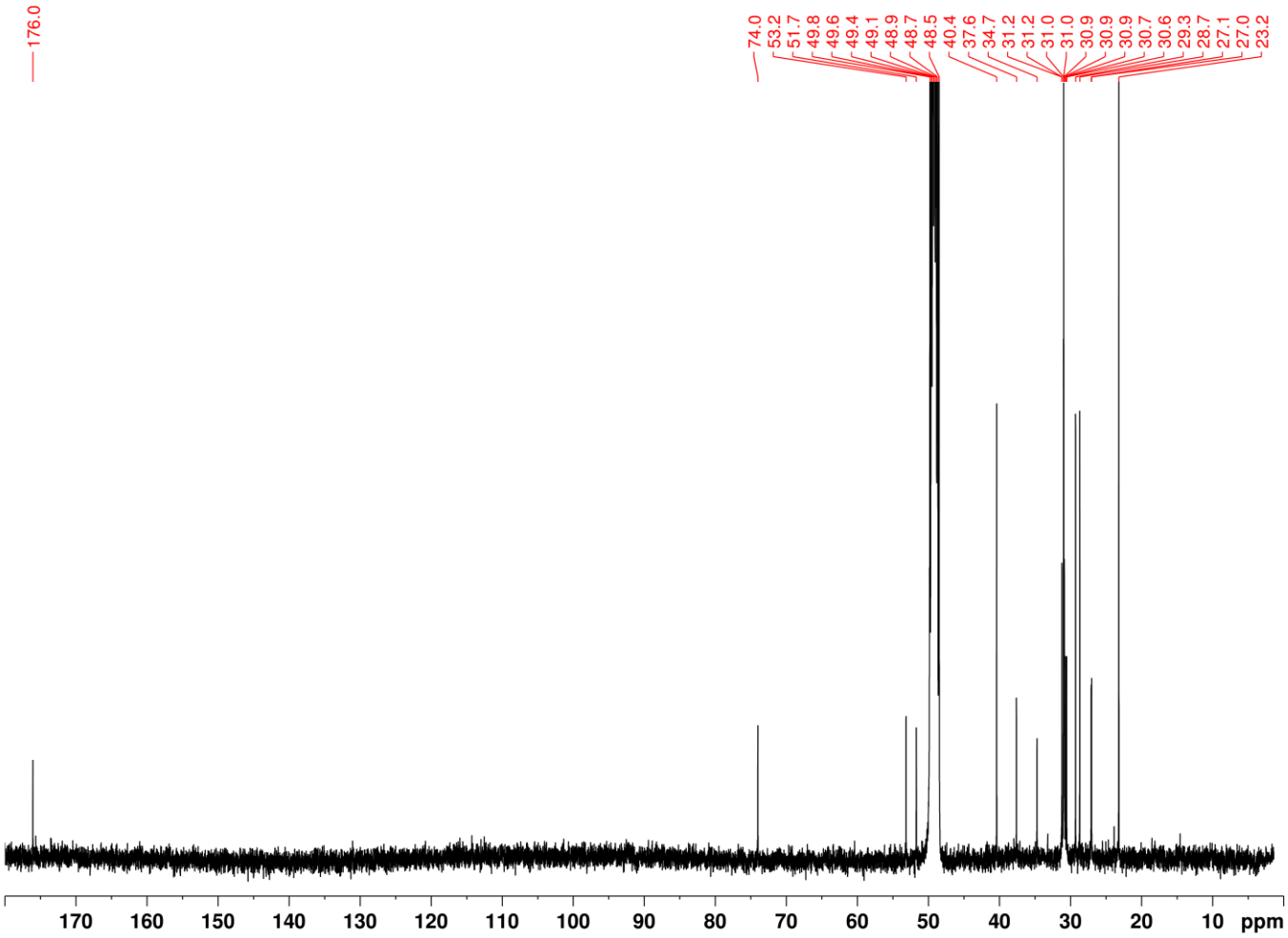


Figure 3.22 ^{13}C NMR (100 MHz, MeOD) spectrum of SoL B.

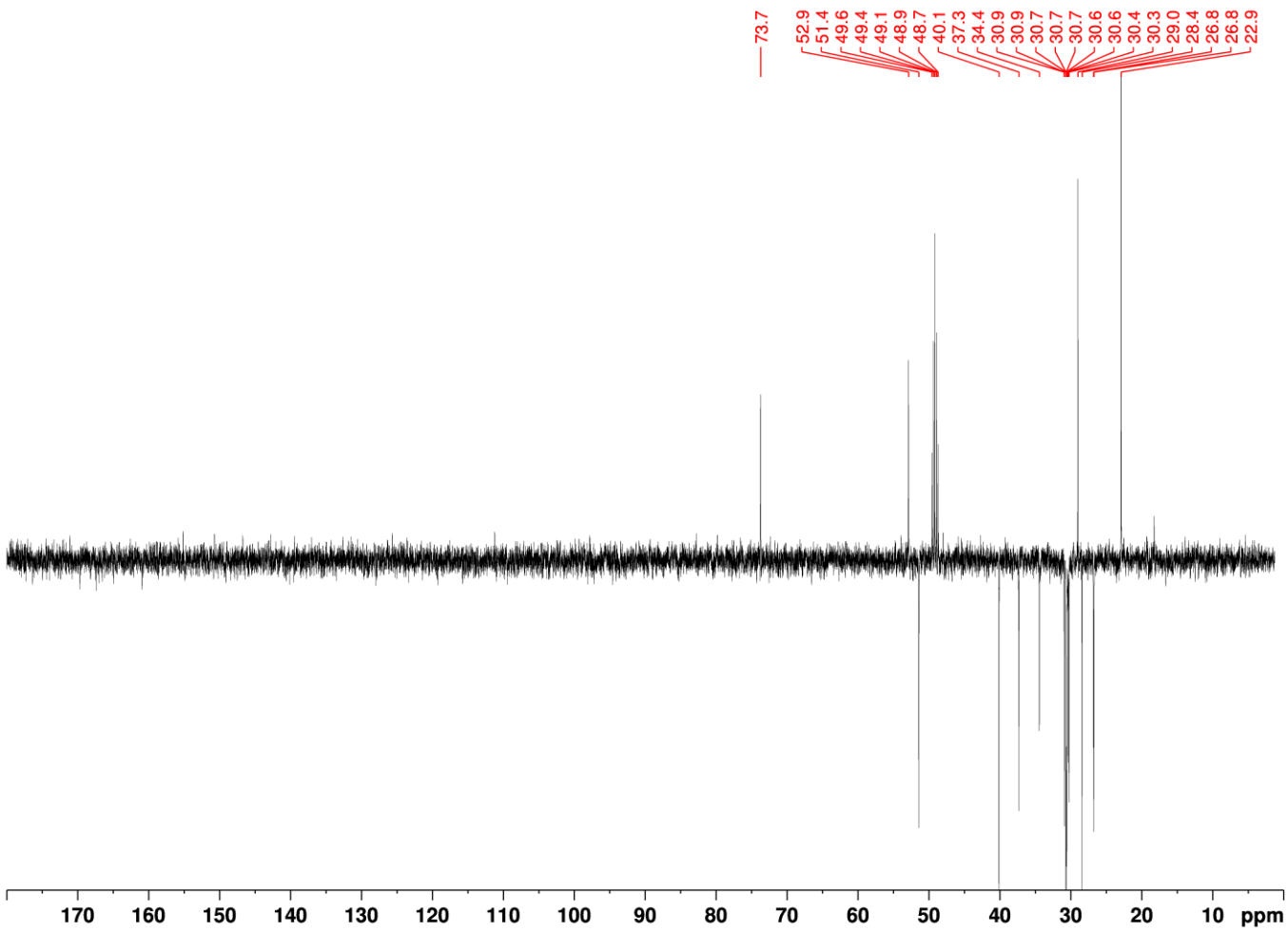


Figure 3.23 DEPT NMR (100 MHz, MeOD) spectrum of SoL B.

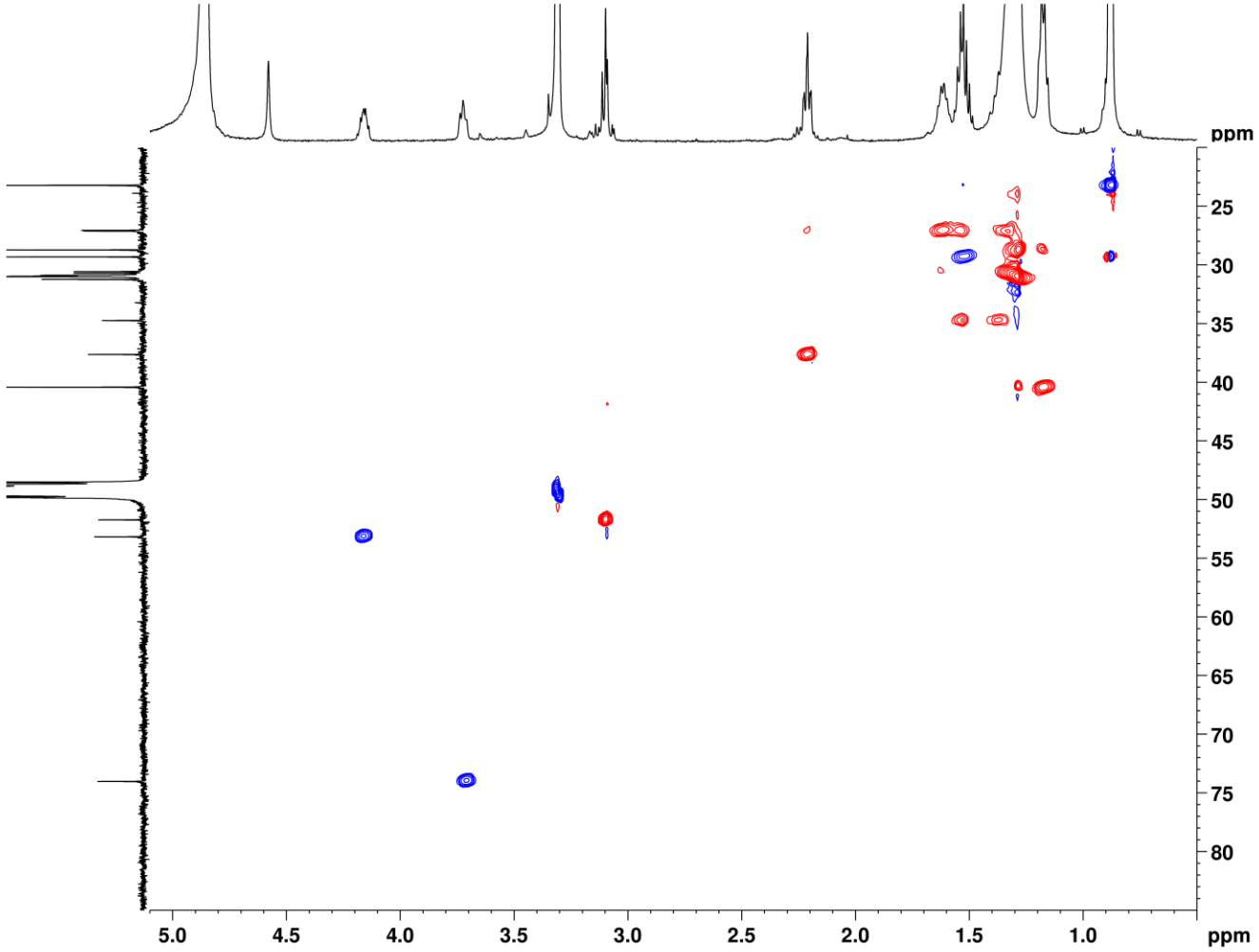


Figure 3.24 HSQC NMR (400 MHz, MeOD) spectrum of SoL B.

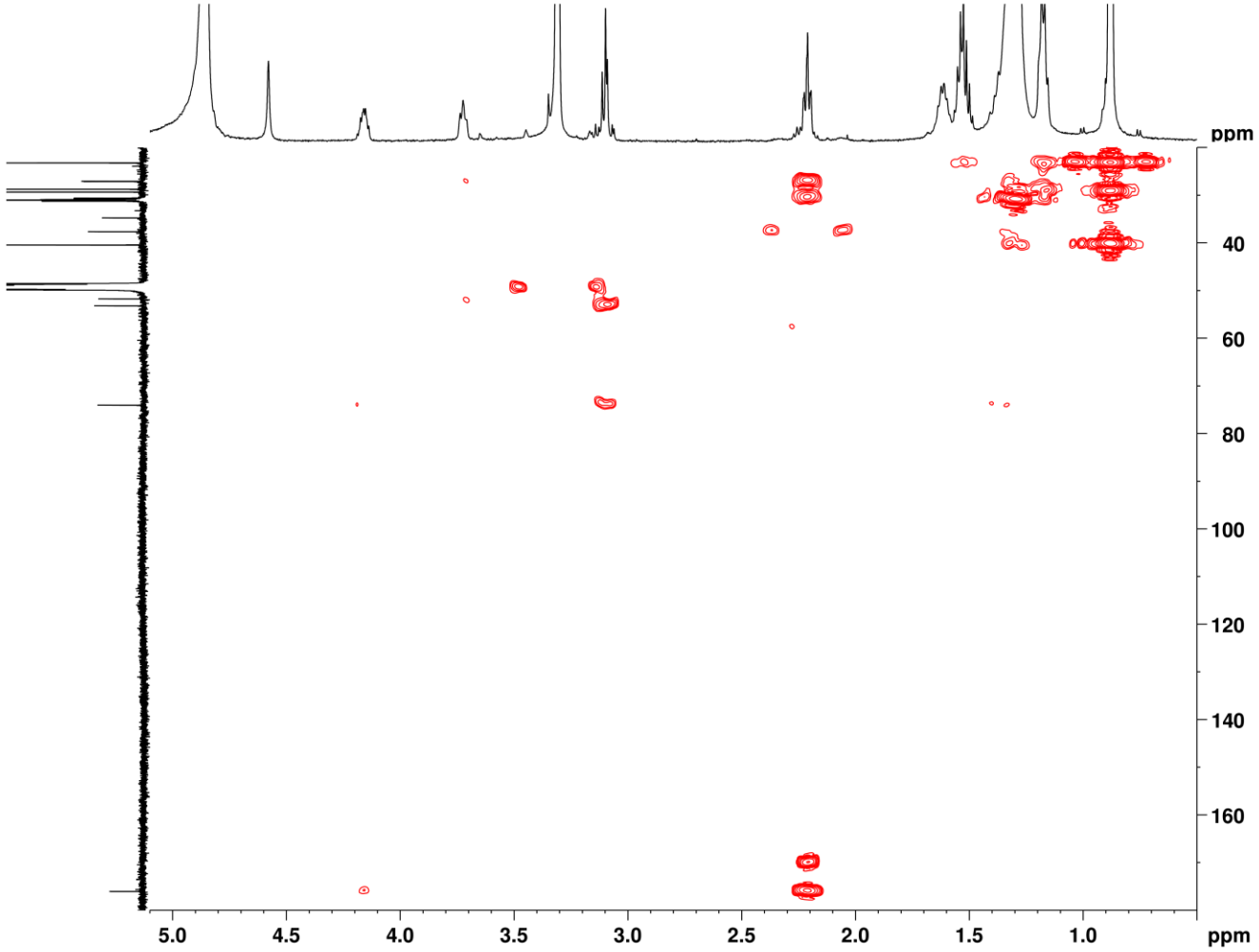


Figure 3.25 HMBC NMR (400 MHz, MeOD) spectrum of SoL B.

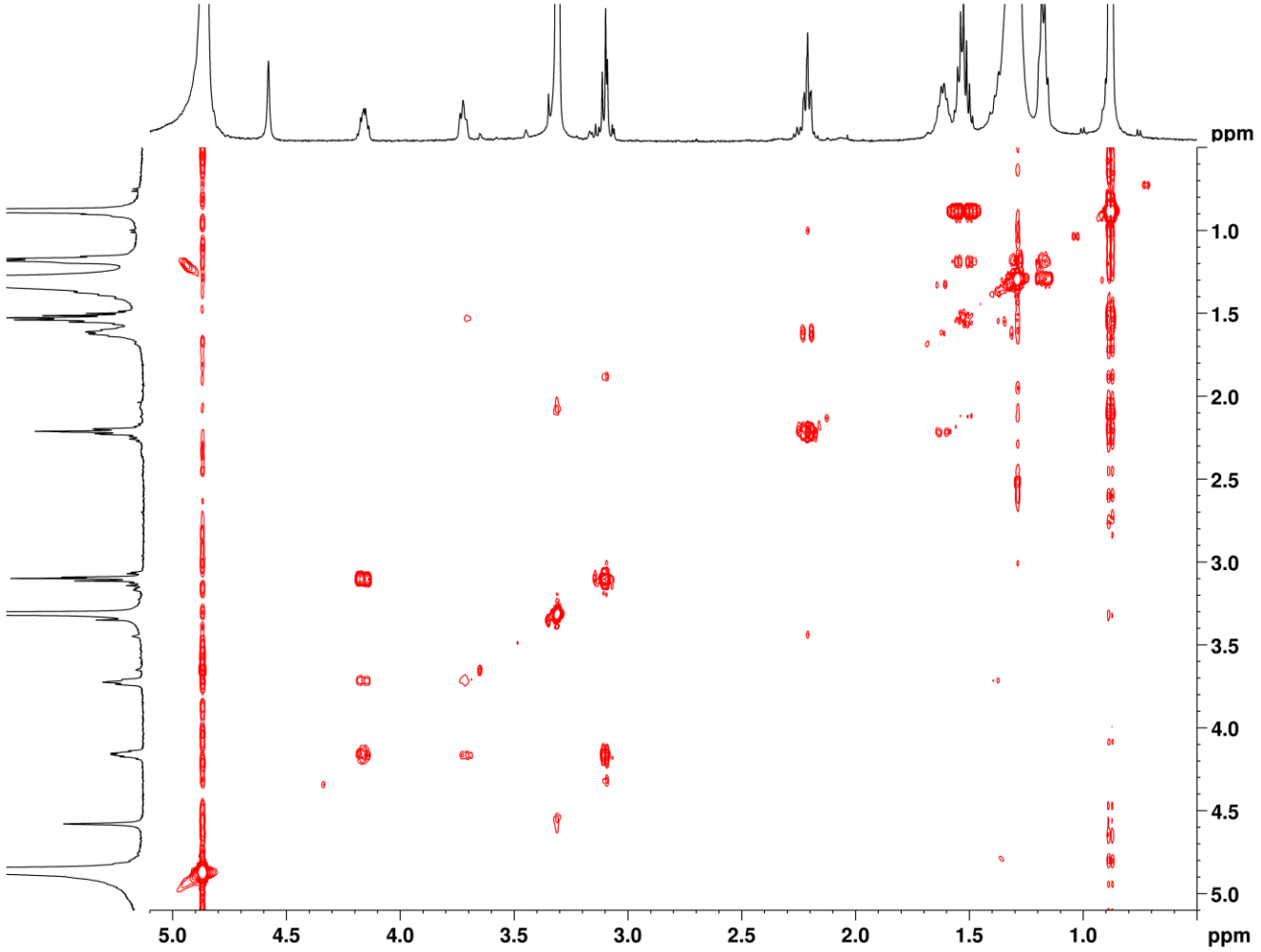


Figure 3.26 ^{1}H-^{1}H COSY NMR (400 MHz, MeOD) spectrum of SoL B.

3.3.2 SoLs preference for TLR4 *in vitro*

While the causes of IBD remain largely unknown, IBD progression has been linked to aberrant TLR signaling[182]. TLRs are pattern recognition receptors (PRRs) which initiate a variety of host processes, especially inflammatory responses, through the recognition of pathogen-associated molecular patterns (PAMPs) and other non-pathogenic microbial factors[182,185,186]. Specifically, TLR2 and TLR4 are well-known to recognize PAMPs in the gut microbiome[182]. In addition, their expression is significantly increased in IBD pathogenesis, reflecting a state of aberrant activation[182,185]. Thus, we expected that SoLs may interact with TLR4 or TLR2 to mediate their immunomodulatory activity. We treated primary mouse macrophages collected from wild-type C57BL/6 mice with SoL A (as a representative of SoLs) either alone or together with LPS (an agonist of TLR4) or Pam3CSK4 (an agonist of TLR1/2) and measured the expression of three inflammatory cytokines (IL-6, TNFα, and IL-1β). By itself, SoL A exhibited a mild to moderate effect on the expression of pro-inflammatory cytokines compared to control (Figure 3.27), generally consistent with our previous finding[161]. As expected, the TLR ligands, LPS and Pam3CSK4, both showed significant induction of all three cytokines compared to control (Student's *t* test, $p \leq 0.0001$) (Figure 3.27). Notably, SoL A was found to significantly suppress the expression of all three cytokines induced by LPS ($p \leq 0.0001$) (Figure 3.27). Together, SoL A's mild pro-inflammatory activity by itself and primarily strong inhibition against LPS-induced inflammation constitute its dual immunomodulatory activity. SoL A also inhibited Pam3CSK4-induced IL-6 and TNFα to a smaller extent while increasing IL-1β expression induced by Pam3CSK4 ($p \leq 0.05$) (Figure 3.27). This result indicates that SoL A primarily affects LPS-induced inflammation and implies that interaction with TLR4

may be involved in SoL A's mechanism of action. Interestingly, SoL A's partial suppression of Pam3CSK4-induced inflammation suggests that SoL A-related anti-inflammatory activity may also extend to the TLR1/2 pathway, albeit to a lesser extent, and warrants further investigation. After identifying that SoL A's primary effect is through TLR4, we further examined the biological activity of SoL B against LPS-induced inflammation. We found that SoL B also inhibited LPS-induced inflammation albeit to a lesser extent than SoL A (Figure 3.28). This activity is consistent with previous reports that SoL B exhibits anti-inflammatory activity both *in vitro* and *in vivo* in mice[187].

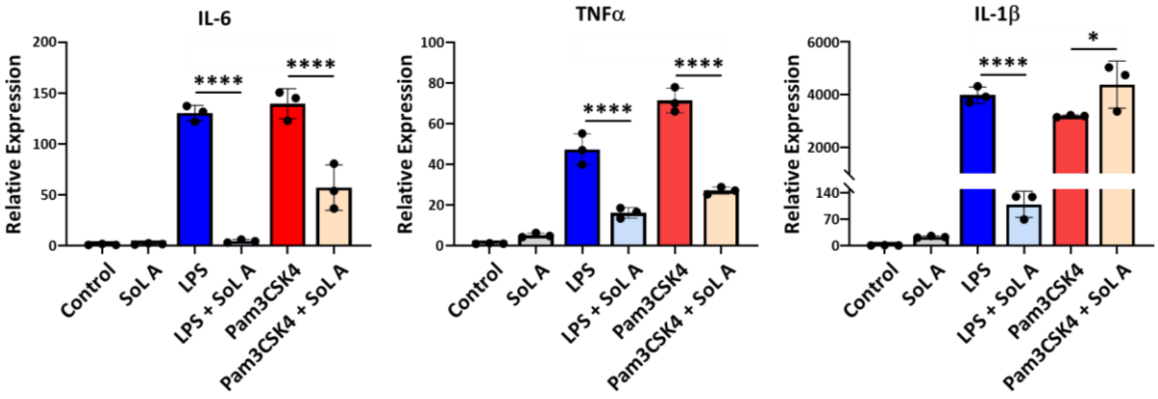


Figure 3.27 SoL A primarily suppresses LPS-induced TLR4 activation. Mouse peritoneal macrophages were treated with SoL A (10 μM), LPS (100 ng/mL), and Pam3CSK4 (500 ng/mL), either alone or in combination for 6 hours. RT-qPCR analysis revealed that SoL A induces a mild pro-inflammatory effect compared to control but significantly suppresses LPS-induced cytokine expression levels and only partially suppresses Pam3CSK4-induced cytokine expression. Error bars represent mean ± standard error. Experiments were independently repeated three times. For each treatment, $n = 3$. Significance was determined using Student's *t* test: * $p < 0.05$, **** $p < 0.0001$.

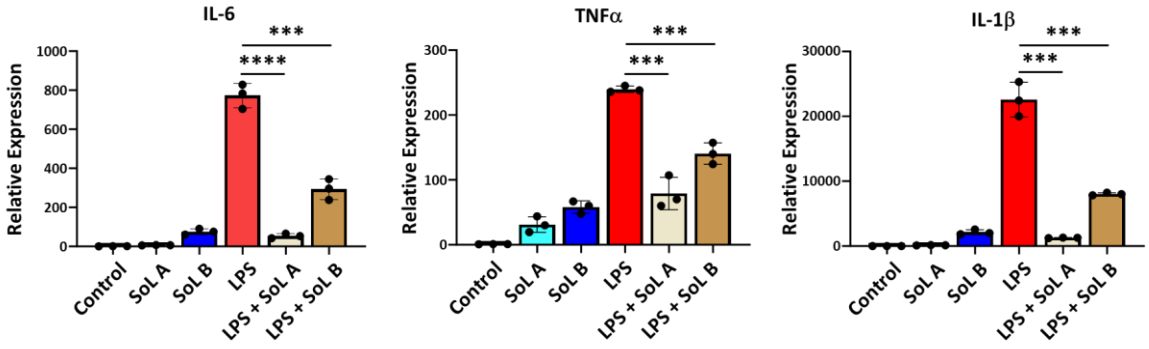


Figure 3.28 Dual immunomodulatory activity of SoLs A and B. Mouse peritoneal macrophages collected from C57BL/6 mice were treated with SoL A (10 uM), SoL B (10 uM), and LPS (50 ng/mL), either alone or in combination for 6 hours. RT-qPCR analysis of the expression of pro-inflammatory cytokines IL-6, TNFα, and IL-1β showed that SoL A and B both induced a weak pro-inflammatory effect and significantly decreased LPS-induced expression of pro-inflammatory cytokines although the effect of SoL B was slightly lesser in magnitude compared to SoL A. Error bars represent mean ± standard error. Experiments were independently repeated three times. For each treatment, n = 3. Significance was determined using Student's t test: * $p < 0.05$, **** $p < 0.0001$.

3.3.3 SoLs interact with TLR4 through binding to MD-2

LPS stimulation of TLR4 occurs through a series of interactions ultimately resulting in LPS binding to myeloid differentiation factor-2 (MD-2), which forms a complex with TLR4 and induces dimerization to initiate signaling[188–190]. The TLR4/MD-2 heterodimer recognizes structurally diverse LPS molecules, giving it flexibility to detect different LPS-related PAMPs in the human gut microbiome[190]. Interestingly, the TLR4/MD-2 complex was recently found to recognize human sulfatides, sphingolipid derivatives which bear a sulfated saccharide head group and dual acyl chains, presumably mimicking the disaccharide core and multiple acyl chains of LPS[191]. Comparing the chemical structure of SoL A to those of sulfatides and lipid A (the immunogenic portion of LPS) (Figure 3.29a), we noted structural similarity in the negatively charged head groups and multiple acyl chains. We thus considered if multiple molecules of SoL A might bind to MD-2 in a similar configuration as sulfatides and lipid A. Inspired by sulfatides that bind

in triplicate to MD-2[191], we used molecular docking to model the binding of three molecules of SoL A to MD-2. Our analysis predicted three molecules of SoL A indeed bind in the hydrophobic pocket of MD-2 (Figure 3.29b), where lipid A is known to bind, with a docking score of -8.9 kcal/mol, better than that of lipid A which had a docking score of -6.2 kcal/mol. Additionally, SoL A was predicted to make hydrophobic contacts with several amino acids including I117, F119, I52, and F121 (Supplementary Table 3.6), all of which are also reported to contact the acyl chains of lipid A[190]. Notably, SoL A is also predicted to contact residues including R264 and R90 (Supplementary Table 3.6), consistent with contacts between these residues and the phosphate groups of lipid A[190]. This suggests that SoL A may bind directly to the TLR4/MD-2 complex and possibly compete for binding with LPS, allowing it to suppress LPS-induced activation of the TLR4 pathway. A critical aspect of lipid A binding to MD-2 is the exclusion of one acyl chain from the hydrophobic pocket of MD-2 which forms a bridge with TLR4 and is involved in inducing dimerization[190]. Likewise, we observed one acyl chain of SoL A excluded from the hydrophobic pocket in our docking analysis (Figure 3.29b), further suggesting that SoL A mimics LPS as a ligand for TLR4. After successfully docking of SoL A, we tested SoL B which lacks an extra hydroxy group (Figure 3.29a) that may increase its interactions with the hydrophobic binding pocket of MD-2. Our docking analysis indeed showed that SoL B also binds to MD-2 (Figure 3.29b), with similar contacts as SoL A (Supplementary Table 3.6) but higher affinity (docking score of -9.6 kcal/mol) as predicted.

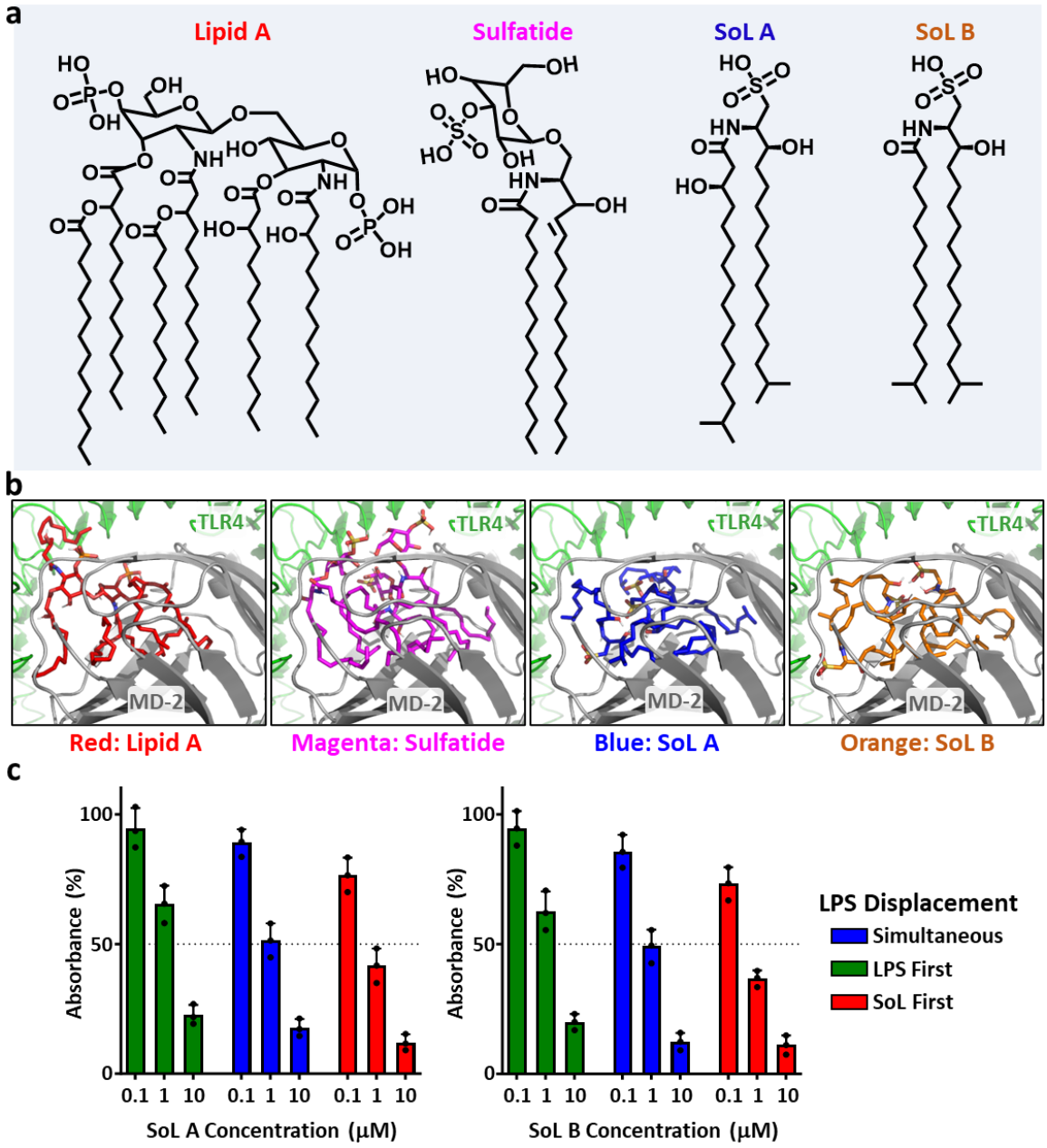


Figure 3.29 SoLs bear structural similarity to both lipid A and sulfatide and bind with MD-2 to block LPS binding. **a**, Chemical structures of immunogenic lipid A (derived from LPS), sulfatide, SoL A, and SoL B illustrating structural similarity in multiple acyl chains and negatively charged head groups. **b**, Molecular docking of lipid A (red), sulfatide (magenta), SoL A (blue), and SoL B (orange) into the hydrophobic pocket of MD-2 in complex with TLR4. Three molecules of SoLs A and B were used in molecular docking experiments to mimic the six acyl chains of lipid A as inspired by sulfatides[191]. **c**, ELISA displacement assay used to measure the binding behavior of SoLs A and B in competition with 1 ng/mL LPS, a natural ligand of the TLR4/MD-2 complex. Compounds were added either simultaneously (blue), LPS first (green), or SoL first (red). Experiments were

independently repeated three times. For each treatment, n = 3. Bars indicate mean ± standard deviation.

To experimentally determine if SoLs A and B bind to MD-2 and to what extent the SoLs compete with LPS for binding to MD-2, we conducted an ELISA-based displacement assay. Taking advantage of biotinylated LPS, which retains the activity of unconjugated LPS[192], we measured absorbance generated by an HRP-linked streptavidin probe to measure the relative amount of MD-2 which was bound with biotin-LPS as opposed to MD-2 bound with SoL A or B. We administered 0.1, 1.0, and 10 μM concentrations of SoL A or B and 1 ng/mL biotin-LPS to MD-2 in three sequences: 1) SoL first followed by LPS 1 hour later, 2) LPS first followed by SoL 1 hour later, and 3) both SoL A or B and LPS at the same time. After 1 hour of incubation, we found that at all concentrations, when SoL A or B was added first, there was a marked decrease in percent absorbance as compared to when LPS was added first and when the two compounds were added together (Figure 3.29c). This suggests that SoLs A and B both bind and occupy some sites of MD-2, preventing LPS from fully binding when it is added 1 hour after SoL A or B. Furthermore, when moving from low to high concentrations of SoL A or B, we observed that the percent absorbance decreased dramatically. This indicates that with increasing concentration of SoL A or B, less LPS binds to MD-2, implying that SoLs indeed compete with LPS for binding to MD-2. Taken together, these results indicate that SoLs A and B can bind directly to MD-2 and more importantly compete with LPS for binding to this target, thus providing a potential molecular mechanism underlying SoL A's pro-inflammatory activity by itself as well as its strong activity in suppressing LPS-induced inflammation which likely also expands to other members of the SoL family.

3.3.4 SoLs interfere with TLR4 signaling

Upon LPS binding, TLR4 initiates downstream signaling, such as through the NF-kB and MAPK pathways, resulting in the induction of inflammatory cytokine expression[185]. If a SoL binds to MD-2 to suppress LPS-induced inflammation, this would block activation of the TLR4 pathway. Therefore, we investigated whether addition of SoL A or B affected the phosphorylation of TLR4-downsteam signaling molecules, ERK1/2 and p38, and the degradation of IκBα, which are critical for LPS-induced cytokine expression[185] (Figure 3.30a). We treated macrophages with LPS in the presence of increasing concentrations of SoL A or B (from 0 to 20 μM), then performed western blot analysis to examine the TLR4-downstream signaling pathways. We found that both SoLs reduced LPS-induced phosphorylation of p38 and ERK1/2 in a concentration-dependent manner. At the concentration of 20 μM, SoLs A and B almost completely blocked LPS-induced phosphorylation of p38 and ERK1/2 (Figure 3.30b). Western blot also showed that SoLs concentration-dependently suppressed LPS-induced IκBα degradation (Figure 3.30b). These results support that SoLs exert their anti-inflammatory effect by blocking LPS-mediated phosphorylation of downstream TLR4 proteins, effectively negating LPS activation of the TLR4 pathway. Also, SoL A or B alone at the concentration of 20 μM slightly enhanced the phosphorylation of certain signaling molecules (e.g., ERK1/2) compared to control, consistent with our observations that SoLs alone induced a mild pro-inflammatory effect on cytokine expression (Figure 3.27) and supporting the dual immunomodulatory activity of SoLs which provides further opportunities to regulate homeostatic immune responses.

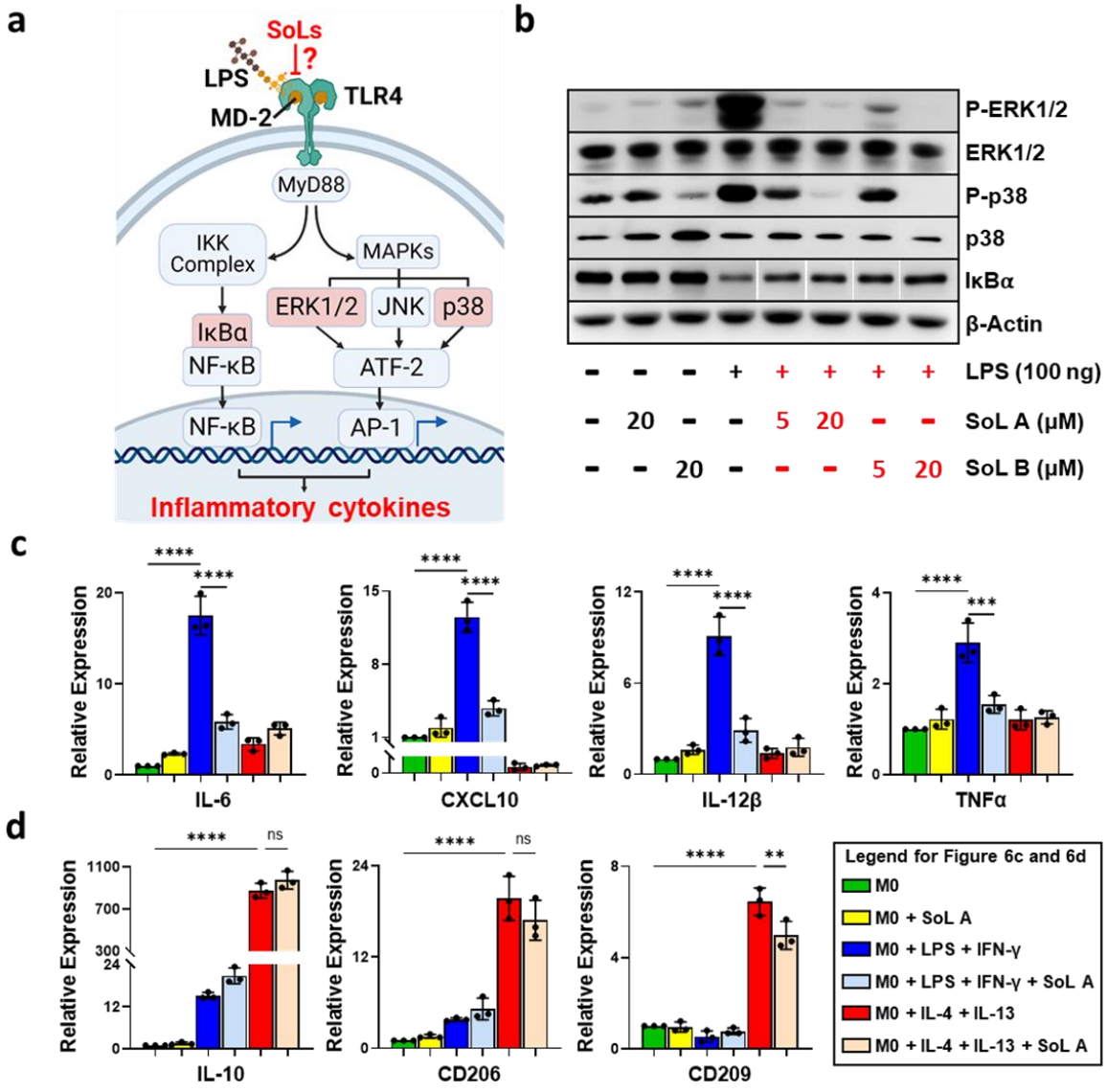


Figure 3.30 SoLs suppress LPS-induced activation of TLR4 signaling pathway and macrophage M1 polarization. **a**, Simplified pathway of TLR4 activation by LPS, highlighting proposed inhibition by SoLs competing for LPS binding. Proteins IκBα, ERK1/2, and p38 downstream of TLR4 which were selected for analysis are highlighted in red rectangles. **b**, Western blot analysis of protein levels of IκBα as well as total and phosphorylated ERK1/2 and p38, after treatment with LPS (100 ng/mL) with or without various concentrations of SoL A or B. The housekeeping gene β-Actin was used as a loading control. The gel for IκBα was spliced to remove extra lanes as indicated by white spaces between bands. **c** and **d**, THP-1-derived macrophages were treated with LPS + IFN-γ or IL-4 + IL-13 to polarize to M1 or M2 macrophages, respectively. Relative expression of markers IL-6, CXCL10, IL-12β, and TNFα (compared to M0) indicate that SoL A (10 μM) had a significant effect on suppressing M1 polarization (**c**); and relative expression of markers IL-10, CD206, and CD209 (compared to M0) indicate that SoL A had no significant effect on M2 polarization (**d**). Experiments were independently repeated three

times. For all treatments, $n = 3$ and significance was determined using one-way ANOVA: ** $0.001 < p < 0.01$, *** $0.0001 < p < 0.001$, and **** $p < 0.0001$.

Because TLR4 signaling leads to macrophage polarization which has been shown to contribute to IBD[193–195], we also examined the effects of SoL A on macrophage polarization. We treated THP-1 monocytes with IFN-γ and LPS to induce M1 polarization or IL-4 and IL-13 to induce M2 polarization. Successful induction of M1 and M2 polarization was confirmed by morphology changes and subsequent RT-qPCR quantification of cytokine profiles. When 10 μM of SoL A was added alongside the respective inducing agents, our relative cytokine expression results showed that SoL A significantly reduced the production of M1-polarized macrophage markers IL-6, CXCL10, IL-12β, and TNFα, compared to macrophages treated without SoL A (Figure 3.30c) but had a mostly non-significant effect on M2 polarization (Figure 3.30d). This suggested that SoL A suppresses macrophage M1 polarization, which supports our aforementioned result that SoLs interfere with TLR4 signaling potentially leading to inhibition of TLR4-mediated IBD.

3.4 CONCLUSIONS

Taking advantage of our unique biosynthetic enzyme-guided disease correlation approach, the results of this study have described two critical points, summarized here and discussed in further detail below. First, we have directly connected the biosynthesis and production of a class of abundant yet underexplored human microbial metabolites, SoLs, to IBD, an existing human health condition with complex and poorly understood etiology, followed by an independent IBD patient cohort and mouse model of IBD to validate this informatically predicted negative correlation. Second, we have revealed that SoLs A and

B, two representative gut microbial SoLs, modulate host immune responses through the TLR4/MD-2 complex and inhibit LPS-induced TLR4 activation and macrophage M1 polarization, which provides a mechanistic explanation for SoLs' potential protective activity against IBD.

Through both bioinformatic and chemoinformatic analyses, we revealed that the expression of SoL biosynthetic enzymes and abundance of SoLs in the gut metabolome are negatively correlated with IBD incidence in humans. Our IBD model with male and female piroxicam treated *Il10*$^{-/-}$ mice supported this sex-independent negative correlation with a concurrent negative association between SoL production and TLR-4-related inflammatory markers TNFα, NOS2, IL-6, and IL-1β. Our findings were consistent with literature reports which have shown that the bacterial genera of abundant SoLs production, *Alistipes* and *Odoribacter*, are associated with the remediation of IBD symptoms[159,160]. Considering that these SoL-producers are commensal members of the human gut microbiome[156,196], their constant production of SoLs in the human gut may help to maintain intestinal immune homeostasis, thereby preventing IBD. Both *Alistipes* and *Odoribacter* are also known to produce short chain fatty acids (SCFAs), which have been implicated in reducing intestinal inflammation[152,160,196–198]. Besides SCFAs, there are other gut microbial functional metabolites such as secondary bile acids and indole derivatives which have also been linked to modulating host inflammation and immunity[46,199–202]. Thus, the role SoLs play individually and/or synergistically with other factors in IBD pathogenesis remains interesting and awaits further investigation.

Towards understanding the mechanism of the immunomodulatory activity of SoLs, we showed that the representative SoLs A and B primarily target the TLR4 pathway and

that they both block LPS binding to MD-2 to suppress TLR4 signaling. Our analysis indicated that SoL A preferentially suppresses LPS-induced inflammation as compared to Pam3CSK4-induced inflammation, suggesting that TLR4 activation is more strongly affected by SoL A than TLR1/2 activation. The selectivity for TLR4 may stem from SoL A's structural similarity to LPS, a hypothesis supported by the recently reported human TLR4 ligands, sulfatides, which share highly similar structural features to SoLs[191]. As microbial functional metabolites, it is reasonable that SoLs are directly recognized by TLR4, which has evolved specifically to recognize PAMPs[203]. Our molecular docking also suggested SoLs A and B's recognition by TLR4, indicating that three molecules of either SoL indeed bind with stronger predicted binding score compared to lipid A and make contacts with important amino acids in the pocket of MD-2, consistent with lipid A binding as well as structurally-related sulfatides[190,191]. Our ELISA-based displacement analysis then confirmed that both SoLs A and B directly bind to MD-2, block LPS binding when added prior to LPS, and could displace bound LPS from MD-2 at higher concentrations. We further demonstrated that increasing concentration of SoLs A and B suppressed LPS-induced TLR4 signaling pathways in a dose-dependent manner and SoL A significantly suppressed M1 polarization of macrophages, further indicating their capacity to reduce downstream TLR4 signaling responses. Notably, increased numbers of M1 macrophages is a characteristic feature of IBD[195] and this suppressive effect may represent one explanation how SoL-producing bacteria are able to remediate symptoms of IBD. Taken together, our results represent the first report of SoLs A and B's binding to MD-2 and establish that SoLs likely mediate their dual immunomodulatory activity by occupying the hydrophobic binding pocket of MD-2 (pro-inflammatory by SoLs alone) but primarily

blocking LPS binding to the TLR4/MD-2 complex (anti-inflammatory against LPS-induced inflammation). This discovery suggests SoLs' mechanistic role in regulating a multitude of TLR4-related inflammatory conditions, most notably IBD which is associated with dysregulation of TLRs, especially TLR4[162–165], leading to aberrant macrophage activation[195,204].

While many studies have focused on the association of certain microbial strains with specific diseases by analyzing the abundance and distribution of strains[31,158,205–208], our biosynthetic enzyme-guided disease correlation approach has shown that the presence and expression of biosynthetic enzymes corresponding to functional metabolites can be directly correlated with human health conditions, effectively shifting the microbe-focused perspective to a functional metabolite-based molecular perspective. By applying this approach, we have identified a negative relationship between SoL biosynthesis and IBD directly from patients' data followed by verification using an IBD mouse model, which has guided us to further reveal a molecular target and a potential mechanistic explanation for the protective effect that SoLs and SoL-producing bacteria exert against IBD. We are now further characterizing the effect of SoL biosynthesis on IBD pathogenesis through in-depth *in vivo* studies using purified SoLs as well as developing isogenic mutants of SoL-producing bacteria for mouse colonization studies. With the exponentially increasing availability of human disease omics data, we expect our approach described in this study to be widely applicable to uncovering the molecular mechanisms of other intricate host-microbe interactions.

3.5 MATERIALS AND METHODS

Identification of SoL biosynthetic enzymes from human gut bacterial reference genomes

We collected experimentally validated enzymes involved in SoL biosynthesis as reference amino acid sequences of CYS, CFAT, and SDR (Supplementary Table 3.1). Reference amino acid sequences were used as seed sequences to search for homologs in the human gut bacterial reference genomes using the DIAMOND blastp model[209] with an e-value threshold of 10^{-5}. We then investigated the taxonomic distribution of the resulting homolog sets based on the taxonomy annotation for each genome[166]. To prioritize SoL biosynthetic enzymes, the homologs of CFATs and CYSs from genomes which encode copies of CFAT, CYS and SDR were first used to generate sequence similarity networks with experimentally validated CFAT and CYS at a threshold of 50% similarity using MMseqs2[210]. The prioritized CFATs and CYSs meet co-occurrence with SDRs in the same genome. The filtered enzymes were further subjected to Pfam domain analysis by hmmsearch (HMMER v3.3) with default parameters against the Pfam-A database (v33.1). Enzymes containing corresponding Pfam domains with hit score > 50 were selected as prioritized SoL biosynthetic enzymes. Prioritized enzymes were used to generate CYS, CFAT, and SDR subfamilies using sequence similarity networks with at least 90% similarity by MMseqs2 clustering[210]. Maximum-likelihood trees were generated using the representative genome (Supplementary Data 3.2) of each species by GTDB-Tk (v2)[211] with the following parameters (refer to the GTDB-Tk user guide: https://ecogenomics.github.io/GTDBTk/commands/index.html): 1. GTDB-Tk reference data release 207 was used. 2. Parameters of classify: gtdbtk classify_wf --genome_dir --

out_dir --force. 3. Parameters of infer tree from multiple sequence alignment: gtdbtk infer --msa_file MSA_FILE --out_dir OUT_DIR. 4. Using the convert_to_itol command to make the tree suitable for visualization in iTOL: gtdbtk convert_to_itol --input_tree --output_tree. Finally, we annotate their phylogenetic trends using iTOL[125].

Quantification of enzyme abundances in metagenomic and metatranscriptomic samples

Metagenomic and metatranscriptomic whole-genome sequencing datasets of human gut microbiomes related to IBD were downloaded from the Inflammatory Bowel Disease Multi'omics Database[158,171] (IBDMDB, https://ibdmdb.org/tunnel/public/summary.html) For both metagenomic and metatranscriptomic samples, reads were quality filtered and adapter removed using *bbduk.sh* with the following parameters: qtrim=rl ktrim=r mink=11 trimq=10 minlen=40 (read quality cutoff is 10, read length cutoff is 40). High-quality reads were mapped to the nucleotide sequences of corresponding contigs containing the SoL biosynthetic genes using the BWA mem algorithm[212] with default parameters. The reads mapped to genes were counted by featurecounts[120] with the following parameters: -f -t CDS -M -O -g transcript_id -F GTF -s 0 -p --fracOverlap 0.25 -Q 10 –primary. Enzymes encoded or expressed in at least 5% of samples were considered as common distribution in humans and were included in the comparative analysis. Transcripts per million (TPM) were calculated for each SoL biosynthetic enzyme-encoding gene. The abundance of each subfamily was calculated by the sum of the relative abundances of all genes in the subfamily. Beta diversity was performed to quantify the prevalence and relative abundance differences in the overall composition of SoL biosynthetic enzymes between the IBD and the control groups.

PERMANOVA was performed to show the encoding and expression profile differences of SoL biosynthetic enzymes between IBD and control groups. Both beta diversity and PERMANOVA were performed using the R package vegan[213,214]. To explore the differences between IBD and control groups of single SoL biosynthetic enzyme subfamilies, we used the Shapiro–Wilk test to evaluate the normality of a specific gene family's relative abundance. We then calculated the significance of relative abundance between healthy and IBD individuals using either two-sample Student's t-test (for normally distributed data) or two-sample Wilcoxon rank sum test (for not normally distributed data). Significance tests were performed in Python using packages Pandas and SciPy[215–217]. The Benjamini-Hochberg method was used to adjust p-values to correct for multiple testing[115]. SoL biosynthetic enzyme subfamilies were considered differential if the adjusted p-value was less than 0.05. For differential Sol biosynthetic enzymes, we performed a two-sided Fisher's exact test to explore their difference in prevalence across IBD and non-IBD groups.

Analysis of publicly available metabolomics datasets

Processed per-subject metabolomic feature tables and microbial species relative abundance profiles were collected from two publicly available IBD datasets: IBDMDB[158,171] (https://ibdmdb.org/tunnel/public/summary.html) and PRISM[172] (https://www.nature.com/articles/s41564-018-0306-4#Sec31). Known SoL-related metabolites were collected from MS-DIAL4[173] and the literature[156] (Supplementary Table 3.5). Search parameters were set to the exact mass of SoLs using a 5-ppm match tolerance for parent ions. Identified metabolomic features were considered SoLs candidates. Next, features co-eluting with SoL candidates (within ± 0.05 min retention time window) were

selected to build a feature similarity network (Pearson correlation). Correlations and significance were calculated in R using the "rcorr" function from the "Hmisc" package[218]. Relative abundance was normalized using the equation below following the calculation described in dataset 2:

$$\text{Metabolomic feature abundance } = \frac{Ni}{Sum(Ni)} * 10^6$$

Here, Ni represents the absolute intensity of a metabolomic feature, Sum(Ni) represents the total absolute intensity of all features in the sample. In this step, all the metabolomic samples were included (dataset 1: 546 samples; dataset 2: 220 samples). The correlation coefficients ≥ 0.9 and $p < 0.05$ between SoL candidates and other metabolomic features were subjected to further analysis hereafter. If the co-eluting features could be identified in SoL candidates' MS/MS spectra in MS-DIAL4[173] and a previous in-depth study[156], these features were considered SoL candidates' in-source fragments. If not, these metabolites might be co-produced by a specific microbe, or chemically modified from a common parent metabolite, and these metabolites were considered as SoL analog candidates. Of note, each data type, detected using the same HPLC-MS method[158,171,172], was analyzed separately in each cohort.

To further examine whether SoL candidates correlated with the abundance of any taxa, the Spearman correlation between species-level relative abundances and SoL candidates' relative abundance was calculated. To do so, a series of SoL analog candidates (SoL candidates correlated with other SoL candidates) were selected for further analysis. In total, 472 samples collected from dataset 1 and 220 samples collected from dataset 2 were included. Of note, corresponding microbial species' relative abundance profiles and metabolomic features data were paired in these samples. Species that were present in less

than 5% of samples were excluded. Finally, the relative abundance of SoL analogs was used to calculate differential abundances between non-IBD and IBD samples. The one-sided Wilcoxon rank sum test was used to measure statistical significance with the hypothesis that the abundance of SoL was higher in the non-IBD group than in UC or CD.

Animals

Il10$^{-/-}$ mice on the C57BL/6 background were originally purchased from The Jackson Laboratory. Mice were bred and maintained in specific pathogen free conditions at the University of South Carolina and were maintained on a 12-hour light/dark cycle with unlimited access to water and food (Inotiv Teklad, 8604, https://insights.envigo.com/hubfs/resources/data-sheets/8604-datasheet-0915.pdf). All animal protocols were approved by the University of South Carolina Institutional Animal Care and Use Committee.

Piroxicam-accelerated *Il10*$^{-/-}$ mouse IBD model and analysis

At 8-12 weeks of age, male (n = 5 distributed in 2 cages) and female (n = 7 distributed in 2 cages) *Il10*$^{-/-}$ mice were switched to a normal chow diet (Inotiv Teklad, 8604) supplemented with 100 ppm of piroxicam (Cayman Chemical, 13368; Inotiv Teklad, TD.210442) to induce colitis development as previously described[181]. In parallel to these mice, male (n = 3 in 1 cage) and female (n = 4 in 1 cage) *Il10*$^{-/-}$ mice were maintained on the control diet for the duration of the experiment to serve as pre-colitic controls. Mice were euthanized at 18 days to collect tissues for inflammation assessment and intestinal contents for quantification of SoLs. At necropsy, colitis severity was first grossly assessed, which included qualitative evaluations of cecal atrophy (0–5), thickening of cecal (0–5) and colon tissues (0–5), extent of content loss in the cecum (0–4) and stool consistency /

diarrhea (0–3). For histopathology, segments of the colon were first washed in PBS and then fixed in 10% neutral buffered formalin. The tissues were embedded in paraffin, cut into 5-mm sections, and stained with hematoxylin and eosin (H&E) at the Instrumentation Resource Facility at the University of South Carolina School of Medicine. Inflammation scores of colon sections were blindly assessed as previously described[219] using an Echo Revolve light microscope and accompanying software. Briefly, colitis severity was assessed based on the following histopathological features: length measurements in microns of crypt hyperplasia converted to a score from 0-4, qualitative assessment of goblet cell loss (0–5), crypt abscesses per 10X field counts converted to a score from 0-4, and qualitative assessment of submucosal edema (0-3). RNA isolations and RT-qPCR were performed as previously described[220]. Briefly, RNA was isolated from snap frozen cecal tissues using the TriZol method (Thermo Fisher Scientific). cDNA was synthesized using SuperScript III reverse transcriptase (ThermoFisher Scientific). qRT-PCR was performed at the Functional Genomics Core at the University of South Carolina. The relative abundance of mammalian mRNA transcripts was calculated using the $\Delta\Delta C_T$ method and normalized to *Eef2* levels. The oligonucleotides used for qRT-PCR were: *Eef2* forward: TGTCAGTCATCGCCCATGTG, reverse: CATCCTTGCGAGTGTCAGTGA; *Tnfa* forward: AGCCAGGAGGGAGAACAGAAAC, reverse: CCAGTGAGTGAAAGGGACAGAACC; *Nos2* forward TTGGGTCTTGTTCACTCCACGG, reverse: CCTCTTTCAGGTCACTTTGGTAGG; *Il6* forward: GAAATGATGGATGCTACCAAACTG, reverse: CTCTCTGAAGGACTCTGGCTTTG; *Il1b* forward: CTCAATGGACAGAATATCAACCAAC, reverse: GGCTGTGCCGTCTTTCATTAC.

Fecal samples were collected and immediately flash frozen and stored at -80 C. For SoL extraction and quantification, frozen fecal samples were lyophilized to remove remaining H_2O and subsequently resuspended in MeOH (ThermoFisher Scientific). Fecal sample suspensions were vortexed for 1 minute prior to sonication for 10 minutes. The MeOH extract was collected by centrifugation at 20,000 x *g* for 10 minutes and dried under a gentle stream of nitrogen. The resulting residue was then redissolved in MeOH + 0.1% ammonium hydroxide (ThermoFisher Scientific) and filtered through a 0.22 μm filter prior to analysis. High-resolution mass spectra were collected using a ThermoFisher Scientific Q-Exactive HF-X hybrid Quadrupole-Orbitrap mass spectrometer using electrospray ionization in negative mode. Liquid chromatography (LC) used a ThermoFisher Scientific Vanquish HPLC coupled to the aforementioned mass spectrometer. LC was performed using a Waters Xbridge BEH C18 XP column (2.1 x 100 mm) with alkaline mobile phase (pH 10.7) consisting of solvent A (H_2O + 0.1% ammonium hydroxide) and solvent B (ACN + 0.1% ammonium hydroxide) in a gradient starting from 35% B and increasing to 100% B over 10 minutes, hold at 100% B for 5 minutes, then re-equilibration at 35% B for 5 minutes at a flow rate of 0.4 mL/min. MS scans were obtained in the orbitrap analyzer which was scanned from 500 to 2000 *m/z* at a resolution of 60,000 (at 200 *m/z*). Targeted MS/MS fragmentation was conducted for SoLs in a ± 0.5 Da window around their expected *m/z*. MS data was analyzed by Thermo Xcalibur (4.2.47).

Anaerobic culture and bioactive molecular networking

Three *Alistipes* (*A. putredinis* DSM 17216, *A. timonensis* DSM 25383, and *A. timonensis* DSM 27924) and two *Odoribacter* strains (*O. laneus* DSM 22474 and *O. splanchnicus* DSM 20712) were cultured in Reinforced Clostridial Medium (RCM, BD

Biosciences) under anaerobic conditions at 37°C. After three days of growth, cultures were harvested by centrifugation at 12,000 x *g* for 30 minutes. The resulting cell-free supernatant was extracted with an equal volume of methyl ethyl ketone and the cell pellets were extracted by resuspension in MeOH and sonication before both extracts were combined and concentrated *in vacuo*. The combined crude extract was then fractionated on a silica gel column using a stepwise gradient of DCM and MeOH (DCM:MeOH; 15:1, 7:1, 5:1, 3:1, 1:1). Each fraction was then used in an *in vitro* cell-based assay measuring the suppression of LPS-induced TNFα expression. Simultaneously, samples of the fractions were subjected to untargeted HRMS/MS as described above. MS/MS was conducted using data-dependent acquisition with a resolution of 30,000, isolation window of 2.0 *m/z*, and dynamic exclusion time of 15 seconds. HRMS/MS data was processed using MZmine3 following the GNPS FBMN workflow with minimal changes[221]. Molecular networks were constructed using the quickstart GNPS FBMN setting with no changes[183]. Raw data used for this analysis was deposited in the University of California, San Diego Center for Computational Mass Spectrometry MassIVE database (ftp://massive.ucsd.edu/MSV000091884/). Bioactivity scores were assigned using a custom R script which calculated Pearson correlation coefficients between each molecular feature and the activity of each fraction[222]. Finally, bioactive molecular networks were visualized in Cytoscape v3.9.1[116].

Purification of SoLs A and B

Fractions containing SoLs from above were further purified by Sephadex LH-20 run in 1:1 DCM:MeOH. Finally, pure SoLs A and B were isolated by semi-preparative scale HPLC running an isocratic solvent composition of 47% H_2O + 0.1% ammonium

hydroxide and 53% acetonitrile + 0.1% ammonium hydroxide on a ThermoFisher Scientific Ultimate 3000 semi-preparative scale HPLC equipped with a Waters Xbridge Prep C18 5 μm OBD column (19 x 100 mm) with a flow rate of 5.0 mL/min. ^{1}H, ^{13}C, ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC, and ^{1}H-^{1}H COSY NMR spectra for SoLs A and B were acquired in MeOD on a Bruker Avance III HD 400 MHz spectrometer with a 5 mm BBO 1H/19F-BB-Z-Gradient prodigy cryoprobe. Data were collected and reported as follows: chemical shift, integration multiplicity (s, singlet; d, doublet; t, triplet; m, multiplet), coupling constant. Chemical shifts were reported using the MeOD resonance as the internal standard for ^{1}H-NMR MeOD: $\delta = 3.31$ ppm and ^{13}C-NMR MeOD: $\delta = 49.0$ ppm. Pure SoLs A and B were confirmed to be free of LPS using a Chromogenic Endotoxin Quant Kit (Pierce).

Preparation and treatment of macrophages

Primary mouse macrophages were prepared by first introducing 3 mL of 3% thioglycolate to mice via intraperitoneal injection. After 3 days, 10 mL of chilled PBS was introduced intraperitoneally to flush out macrophages. The cell suspension was then separated by centrifugation at 300 x *g* for 5 minutes. Cells were seeded in culture dishes containing DMEM with 10% FBS for 1 hour before being washed with serum-free DMEM two times to remove unattached cells. The cells were incubated in serum-free DMEM for 16 hours before treatment. To treat the macrophages, cells were incubated for 6 to 24 hours in DMEM without FBS with addition of LPS (Sigma-Aldrich), Pam3CSK4 (Invivogen), or SoL A. Cells were finally washed twice with Dulbecco's phosphate-buffered saline before being lysed for total RNA or protein extraction.

mRNA extraction and RT-qPCR in macrophage-based assays

Treated mouse macrophage cells were lysed with TriZol (Invitrogen) and total RNA was extracted from the cell lysate using a Direct-zol RNA miniprep kit (Zymo Research) according to the manufacturer's protocol. The quality and quantity of RNA was then determined using a nanodrop and 1000 ng of mRNA from each sample was used for cDNA synthesis using a First-strand cDNA Synthesis System (Marligen Bioscience). qPCRs reactions were prepared in a 20 μL final volume containing Fast Start Universal SYBR Green Master (Rox) (Roche Applied Science), cDNA template, deionized H_2O, and primers and probes for IL-1β, TNFα, IL-6, and the 18S rRNA which was used as a housekeeping gene Cycling conditions were 95 °C for 10 min followed by 40 cycles of 95 °C for 10 seconds, 60 °C for 15 seconds, and 68 °C for 20 seconds, then a melting curve analysis from 60 °C to 95°C every 0.2 °C was obtained. Amplifications were performed on an Eppendorf Realplex Mastercycler (Eppendorf). Relative gene expression levels were calculated using the $\Delta\Delta C_T$ method and expression levels of 18S were used to normalize the results.

Molecular docking

The crystal structure of the TLR4/MD-2 comlpex was retrieved from the Protein Data Bank (PDB ID: 3FXI)[190] and prepared using AutoDock Tools[223]. Molecular structures of SoL A, SoL B, sulfatide, and lipid A were constructed, and energy minimized using Marvin version 21.17.0, ChemAxon (https://www.chemaxon.com). Models of SoL A, SoL B, sulfatide, and lipid A were also prepared using AutoDock Tools and docked against the TLR4/MD-2 complex using AutoDock Vina[224,225] in a 32x32x32 angstrom box surrounding the MD-2 monomer. Docking results were visualized using PyMol[226].

ELISA displacement assay

Solid-phase sandwich ELISA kits were purchased from Invitrogen. The ELISA experiments were performed according to the kit instructions, using 50 nM hMD-2 (Novus Biologicals), 1 ng/mL LPS-EB Biotin (Invivogen, Cat. No.: tlrl-lpsbiot, derived from *E. coli* 0111:B4), and 0.1, 1.0, and 10 μM purified SoL A. SoL A was added to the assay 1 hour before, 1 hour after, or simultaneously with LPS-EB Biotin. Absorbance was measured at 450 nm using a BioTek microplate reader.

Western blot

Macrophages were treated with 100 ng/mL of LPS and 5 or 20 μM SoL A for 30 minutes. Following treatment, all cellular protein was extracted using MPER lysis buffer (Thermo Scientific). Protein samples were loaded onto SDS-PAGE gels for separation, then transferred to nitrocellulose membranes (Amersham Biosciences). Primary antibodies and HRP-conjugated secondary antibodies (Cell Signaling Technology) were used to detect target proteins. Signal was detected using an ECL kit (Thermo Scientific).

Macrophage polarization

THP-1 monocytes were maintained in RPMI1640 with 10% heat inactivated FBS, 1% Penicillin-Streptomycin-Amphotericin B, and 50 μM 2-mercaptoethanol prior to differentiation. The cells were differentiated into macrophages with 150 nM PMA for 48 hours. M1 polarization was induced by adding 20 ng/mL IFN-γ and 100 pg/mL LPS for 24 hours. M2 polarization was induced by adding 20 ng/mL IL-4 and 20 ng/mL IL-13 for 24 hours. In all tests, 10 μM SoL A was added at the same time as M1 and M2 differentiation agents. After 24 hours of treatment, total RNA was collected, and RT-qPCR was performed as described above.

3.6 DATA AVAILABILITY

285,385 human microbial reference genomes were obtained from a previously published collection[31]. Whole-genome sequencing datasets of human gut microbiomes related to IBD used for biosynthetic enzyme-guided disease correlation were downloaded from the NCBI SRA (SRA Accessions: PRJNA398089 and PRJNA389280). Targeted metabolomic analysis performed in this work was based on the data obtained from two publicly available metabolomics datasets downloaded from IBDMDB (https://ibdmdb.org/tunnel/public/summary.html) and Metabolomics Workbench (http://www.metabolomicsworkbench.org, accession PR000677). Tables for the processed biosynthetic enzymes and metabolite abundance are available in Supplementary Data 3.1, 3.2 and 3.3. Data used for the bioactive molecular networking analysis was deposited in the UCSD CCMS MassIVE database (https://doi.org/doi:10.25345/C5028PP9T; ftp://massive.ucsd.edu/MSV000091884/).

3.7 CODE AVAILABILITY

The code used for analysis in support of our findings are available in the GitHub repository: https://github.com/ZHANGJianArya/SoL.

3.8 ACKNOWLEDGEMENTS

The work presented in the chapter was completed in collaboration with multiple highly talented researchers. Specifically, thank you to Dr. Dan Xue, Mr. Michael Madden, Ms. Dorathea Lee, Mr. Andrew Campbell, Ms. Emily Quinn, Ms. Josie Sobecks, Ms. Sarah Zaw, and Ms. Anna Goshow for their assistance in isolation and purification of sulfonolipids. Thank you to Dr. Melissa Ellermann and Ms. Mary Mitchell for their help with the animal studies conducted in this work. Thank you to Dr. Daping Fan and Dr.

Yuzhen Wang for their assistance with *in vitro* cell-based assays. Thank you to Dr. Michael Walla and Dr. William Cotham for their assistance in high resolution mass MS and MS/MS. Also, thank you to Dr. Yong-Xin Li and Ms. Jian Zhang for their assistance in developing the bioinformatic pipeline featured in this work.

BIBLIOGRAPHY

1. Pearce, C., Eckard, P., Gruen-wollny, I. & Hansske, F. G. Microorganisms: Their role in the discovery and development of medicines. in *Natural product chemistry for drug discovery* (eds. Buss, A. D. & Butler, M. S.) 215–241 (Royal Society of Chemistry, 2010).
2. Maplestone, R. A., Stone, M. J. & Williams, D. H. The evolutionary role of secondary metabolites — a review. *Gene* **115**, 151–157 (1992).
3. Traxler, M. F. & Kolter, R. Natural products in soil microbe interactions and evolution. *Natural Product Reports* **32**, 956–970 (2015).
4. Fischbach, M. A. Antibiotics from microbes: Converging to kill. *Current Opinion in Microbiology* **12**, 520–527 (2009).
5. Fleming, A. On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of *B. influenzae*. *British Journal of Experimental Pathology* **10**, (1929).
6. Edoo, Z., Arthur, M. & Hugonnet, J.-E. Reversible inactivation of a peptidoglycan transpeptidase by a β-lactam antibiotic mediated by β-lactam-ring recyclization in the enzyme active site. *Scientific Reports* **7**, 9136 (2017).
7. Johnson, B. A., Anker, H. & Meleney, F. L. Bacitracin: A new antibiotic produced by a member of the *B. subtilis* group. *Science* **102**, 376–377 (1945).

8. Stone, K. J. & Strominger, J. L. Mechanism of action of bacitracin: Complexation with metal ion and C55-isoprenyl pyrophosphate. *Proceedings of the National Academy of Sciences* **68**, 3223–3227 (1971).

9. Waksman, S. A. Streptomycin: Background, isolation, properties, and utilization. *Science* **118**, 259–266 (1953).

10. Spotts, C. R. & Stanier, R. Y. Mechanism of streptomycin action on bacteria: A unitary hypothesis. *Nature* **192**, 633–637 (1961).

11. Chopra, I. & Roberts, M. Tetracycline antibiotics: Mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiology and Molecular Biology Reviews* **65**, 232–260 (2001).

12. Pollack, J. R. & Neilands, J. B. Enterobactin, an iron transport compound from *Salmonella typhimurium*. *Biochemical and Biophysical Research Communications* **38**, 989–992 (1970).

13. Raymond, K. N., Dertz, E. A. & Kim, S. S. Enterobactin: An archetype for microbial iron transport. *Proceedings of the National Academy of Sciences* **100**, 3584–3588 (2003).

14. Heesemann, J. *et al.* Virulence of *Yersinia enterocolitica* is closely associated with siderophore production, expression of an iron-repressible outer membrane polypeptide of 65,000 Da and pesticin sensitivity. *Molecular Microbiology* **8**, 397–408 (1993).

15. Pelludat, C., Rakin, A., Jacobi, C. A., Schubert, S. & Heesemann, J. The yersiniabactin biosynthetic gene cluster of *Yersinia enterocolitica*: Organization and siderophore-dependent regulation. *Journal of Bacteriology* **180**, 538–546 (1998).

16. Schupp, T., Waldmeier, U. & Divers, M. Biosynthesis of desferrioxamine B in *Streptomyces pilosus*: Evidence for the involvement of lysine decarboxylase. *FEMS Microbiology Letters* **42**, 135–139 (1987).

17. Barona-Gómez, F., Wong, U., Giannakopulos, A. E., Derrick, P. J. & Challis, G. L. Identification of a cluster of genes that directs desferrioxamine biosynthesis in *Streptomyces coelicolor* M145. *Journal of the American Chemical Society* **126**, 16282–16283 (2004).

18. Timofeeva, A. M., Galyamova, M. R. & Sedykh, S. E. Bacterial siderophores: Classification, biosynthesis, perspectives of use in agriculture. *Plants* **11**, 3065 (2022).

19. Murdoch, C. C. & Skaar, E. P. Nutritional immunity: The battle for nutrient metals at the host–pathogen interface. *Nature Reviews Microbiology* **20**, 657–670 (2022).

20. Ghssein, G. *et al.* Biosynthesis of a broad-spectrum nicotianamine-like metallophore in *Staphylococcus aureus*. *Science* **352**, 1105–1109 (2016).

21. Ghssein, G. & Ezzeddine, Z. A review of *Pseudomonas aeruginosa* metallophores: Pyoverdine, pyochelin and pseudopaline. *Biology* **11**, 1711 (2022).

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

23. Engebrecht, J., Nealson, K. & Silverman, M. Bacterial bioluminescence: Isolation and genetic analysis of functions from *Vibrio fischeri*. *Cell* **32**, 773–781 (1983).

24. Engebrecht, J. & Silverman, M. Identification of genes and gene products necessary for bacterial bioluminescence. *Proceedings of the National Academy of Sciences* **81**, 4154–4158 (1984).

25. Eberhard, A. *et al.* Structural identification of autoinducer of *Photobacterium fischeri* luciferase. *Biochemistry* **20**, 2444–2449 (1981).

26. Håvarstein, L. S., Coomaraswamy, G. & Morrison, D. A. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences* **92**, 11140–11144 (1995).

27. Yang, Y., Cornilescu, G. & Tal-Gan, Y. Structural characterization of competence-stimulating peptide analogues reveals key features for ComD1 and ComD2 receptor binding in *Streptococcus pneumoniae*. *Biochemistry* **57**, 5359–5369 (2018).

28. Chen, X. *et al.* Structural identification of a bacterial quorum-sensing signal containing boron. *Nature* **415**, 545–549 (2002).

29. Kim, C. S. *et al.* Characterization of autoinducer-3 structure and biosynthesis in *E. coli*. *ACS Central Science* **6**, 197–206 (2020).

30. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).

31. Durack, J. & Lynch, S. V. The gut microbiome: Relationships with disease and opportunities for therapy. *Journal of Experimental Medicine* **216**, 20–40 (2018).

32. Wang, J. & Jia, H. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology* **14**, 508–522 (2016).

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

34. Johnson, E. L. *et al.* Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. *Nature Communications* **11**, 1–11 (2020).

35. An, D., Na, C., Bielawski, J., Hannun, Y. A. & Kasper, D. L. Membrane sphingolipids as essential molecular signals for *Bacteroides* survival in the intestine. *Proceedings of the National Academy of Sciences* **108**, 4666–4671 (2011).

36. Norris, G. H. & Blesso, C. N. Dietary and endogenous sphingolipid metabolism in chronic inflammation. *Nutrients* **9**, 1180–1204 (2017).

37. Hannun, Y. A. & Obeid, L. M. Principles of bioactive lipid signalling: Lessons from sphingolipids. *Nature Reviews Molecular Cell Biology* **9**, 139–150 (2008).

38. Radka, C. D., Frank, M. W., Rock, C. O. & Yao, J. Fatty acid activation and utilization by *Alistipes finegoldii*, a representative Bacteroidetes resident of the human gut microbiome. *Molecular Microbiology* **113**, 807–825 (2020).

39. Brown, E. M. *et al. Bacteroides*-derived sphingolipids are critical for maintaining intestinal homeostasis and symbiosis. *Cell Host and Microbe* **25**, 668-680.e7 (2019).

40. Wieland Brown, L. C. *et al.* Production of α-galactosylceramide by a prominent member of the human gut microbiota. *PLoS Biology* **11**, e1001610 (2013).

41. Ridlon, J. M., Kang, D.-J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *Journal of Lipid Research* **47**, 241–259 (2006).

42. Devlin, A. S. & Fischbach, M. A. A biosynthetic pathway for a prominent class of microbiota-derived bile acids. *Nature Chemical Biology* **11**, 685–690 (2015).

43. Funabashi, M. *et al.* A metabolic pathway for bile acid dehydroxylation by the gut microbiome. *Nature* **582**, 566–570 (2020).

44. Guo, C. *et al.* Bile acids control inflammation and metabolic disorder through inhibition of NLRP3 inflammasome. *Immunity* **45**, 802–816 (2016).

45. Hang, S. *et al.* Bile acid metabolites control T_H17 and T_{reg} cell differentiation. *Nature* **576**, 143–148 (2019).

46. Li, W. *et al.* A bacterial bile acid metabolite modulates T_{reg} activity through the nuclear hormone receptor NR4A1. *Cell Host & Microbe* **29**, 1366-1377.e9 (2021).

47. Ribeiro da Cunha, B., Fonseca, L. P. & Calado, C. R. C. Antibiotic discovery: Where have we come from, where do we go? *Antibiotics* **8**, 45 (2019).

48. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products* **83**, 770–803 (2020).

49. Demain, A. L. Importance of microbial natural products and the need to revitalize their discovery. *Journal of Industrial Microbiology and Biotechnology* **41**, 185–201 (2014).

50. Demain, A. L. & Adrio, J. L. Contributions of microorganisms to industrial biology. *Molecular Biotechnology* **38**, 41–55 (2008).

51. Thomford, N. E. *et al.* Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. *International Journal of Molecular Sciences* **19**, 1578 (2018).

52. Shen, B. A new golden age of natural products drug discovery. *Cell* **163**, 1297–1300 (2015).

53. Lewis, K. The science of antibiotic discovery. *Cell* **181**, 29–45 (2020).

54. Machado, H., Tuttle, R. N. & Jensen, P. R. Omics-based natural product discovery and the lexicon of genome mining. *Current Opinion in Microbiology* **39**, 136–142 (2017).

55. Shen, B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* **7**, 285–295 (2003).

56. Weissman, K. J. Introduction to polyketide biosynthesis. in *Methods in Enzymology* vol. 459 3–16 (Elsevier, 2009).

57. Olano, C., Méndez, C. & Salas, J. A. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. *Natural Product Reports* **27**, 571–616 (2010).

58. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29–W35 (2021).

59. Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making? *Journal of Industrial Microbiology and Biotechnology* **41**, 175–184 (2014).

60. Kenshole, E., Herisse, M., Michael, M. & Pidot, S. J. Natural product discovery through microbial genome mining. *Current Opinion in Chemical Biology* **60**, 47–54 (2021).

61. Baltz, R. H. Natural product drug discovery in the genomic era: Realities, conjectures, misconceptions, and opportunities. *Journal of Industrial Microbiology and Biotechnology* **46**, 281–299 (2019).

62. Miao, V. & Davies, J. Metagenomics and antibiotic discovery from uncultivated bacteria. in *Uncultivated Microorganisms* (ed. Epstein, S. S.) vol. 10 161–180 (Springer Berlin Heidelberg, 2008).

63. 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).

64. Lewis, K. New approaches to antimicrobial discovery. *Biochemical Pharmacology* **134**, 87–98 (2017).

65. Katz, M., Hover, B. M. & Brady, S. F. Culture-independent discovery of natural products from soil metagenomes. *Journal of Industrial Microbiology and Biotechnology* **43**, 129–141 (2016).

66. Garza, D. R. & Dutilh, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences* **72**, 4287–4308 (2015).

67. Ongley, S. E., Bian, X., Neilan, B. A. & Müller, R. Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Natural Product Reports* **30**, 1121–1138 (2013).

68. Sugimoto, Y. *et al.* A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science* **366**, eaax9176 (2019).

69. Li, S. *et al.* Methyltransferases of gentamicin biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 1340–1345 (2018).

70. Yin, M. *et al.* The missing C-17 O-methyltransferase in geldanamycin biosynthesis. *Organic Letters* **13**, 3726–3729 (2011).

71. De Rond, T. *et al.* Oxidative cyclization of prodigiosin by an alkylglycerol monooxygenase-like enzyme. *Nature Chemical Biology* **13**, 1155–1157 (2017).

72. Tsibulskaya, D. *et al.* The product of *Yersinia pseudotuberculosis mcc* operon is a peptide-cytidine antibiotic activated inside producing cells by the TldD/E protease. *Journal of the American Chemical Society* **139**, 16178–16187 (2017).

73. Walker, M. C. *et al.* Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* **21**, 387 (2020).

74. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Natural Product Reports* **38**, 130–239 (2021).

75. Mohr, K. I. *et al.* Pinensins: The first antifungal lantibiotics. *Angewandte Chemie - International Edition* **54**, 11254–11258 (2015).

76. Férir, G. *et al.* The lantibiotic peptide labyrinthopeptin A1 demonstrates broad anti-HIV and anti-HSV activity with potential for microbicidal applications. *PLoS ONE* **8**, (2013).

77. Iorio, M. *et al.* A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity. *ACS Chemical Biology* **9**, 398–404 (2014).

78. Gomes, K. M., Duarte, R. S. & Bastos, M. do C. de F. Lantibiotics produced by *Actinobacteria* and their potential applications (a review). *Microbiology* **163**, 109–121 (2017).

79. Jabés, D. *et al.* Efficacy of the new lantibiotic NAI-107 in experimental infections induced by multidrug-resistant gram-positive pathogens. *Antimicrobial Agents and Chemotherapy* **55**, 1671–1676 (2011).

80. Xu, Y. *et al.* Structure of the nisin leader peptidase NisP revealing a C-terminal autocleavage activity. *Acta Crystallographica Section D: Biological Crystallography* **70**, 1499–1505 (2014).

81. Montalbán-López, M., Deng, J., van Heel, A. J. & Kuipers, O. P. Specificity and application of the lantibiotic protease NisP. *Frontiers in Microbiology* **9**, 1–16 (2018).

82. Tang, W. *et al.* Applications of the class II lanthipeptide protease LicP for sequence-specific, traceless peptide bond cleavage. *Chemical Science* **6**, 6270–6279 (2015).

83. Caetano, T., van der Donk, W. & Mendo, S. Bacteroidetes can be a rich source of novel lanthipeptides: The case study of Pedobacter lusitanus. *Microbiological Research* **235**, 126441 (2020).

84. Repka, L. M., Chekan, J. R., Nair, S. K. & Van Der Donk, W. A. Mechanistic understanding of lanthipeptide biosynthetic enzymes. *Chemical Reviews* **117**, 5457–5520 (2017).

85. Wang, J., Ge, X., Zhang, L., Teng, K. & Zhong, J. One-pot synthesis of class II lanthipeptide bovicin HJ50 via an engineered lanthipeptide synthetase. *Scientific Reports* **6**, (2016).

86. Hegemann, J. D. & Süssmuth, R. D. Matters of class: Coming of age of class III and IV lanthipeptides. *RSC Chemical Biology* **1**, 110–127 (2020).

87. Völler, G. H., Krawczyk, B., Ensle, P. & Süssmuth, R. D. Involvement and unusual substrate specificity of a prolyl oligopeptidase in class III lanthipeptide maturation. *Journal of the American Chemical Society* **135**, 7426–7429 (2013).

88. Chen, S. *et al.* Zn-dependent bifunctional proteases are responsible for leader peptide processing of class III lanthipeptides. *Proceedings of the National Academy of Sciences* **116**, 2533–2538 (2019).

89. Zhang, Q., Doroghazi, J. R., Zhao, X., Walker, M. C. & van der Donk, W. A. Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in *Actinobacteria*. *Applied and Environmental Microbiology* **81**, 4339–4350 (2015).

90. Barbosa, J., Caetano, T. & Mendo, S. Class I and class II lanthipeptides produced by bacillus spp. *Journal of Natural Products* **78**, 2850–2866 (2015).

91. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* **47**, W81–W87 (2019).

92. Corvey, C., Stein, T., Düsterhus, S., Karas, M. & Entian, K. D. Activation of subtilin precursors by *Bacillus subtilis* extracellular serine proteases subtilisin (AprE), WprA, and Vpr. *Biochemical and Biophysical Research Communications* **304**, 48–54 (2003).

93. Ren, H. *et al.* Discovery and characterization of a class IV lanthipeptide with a nonoverlapping ring pattern. *ACS Chemical Biology* **15**, 1642–1649 (2020).

94. Wiebach, V. *et al.* An amphipathic alpha-helix guides maturation of the ribosomally-synthesized lipolanthines. *Angewandte Chemie - International Edition* **59**, 16777–16785 (2020).

95. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

96. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).

97. Ortega, M. A. *et al.* Substrate specificity of the lanthipeptide peptidase ElxP and the oxidoreductase ElxO. *ACS Chemical Biology* **9**, 1718–1725 (2014).

98. Bierbaum, G. The biosynthesis of the lantibiotics epidermin, gallidermin, Pep5 and epilancin K7. *Antonie van Leeuwenhoek* **69**, 119–127 (1996).

99. Favret, M. E. & Yousten, A. A. Thuricin: The bacteriocin produced by *Bacillus thuringiensis*. *Journal of Invertebrate Pathology* **53**, 206–216 (1989).

100. Wang, J. *et al.* Cerecidins, novel lantibiotics from *Bacillus cereus* with potent antimicrobial activity. *Applied and Environmental Microbiology* **80**, 2633–2643 (2014).

101. Tang, W. & Van Der Donk, W. A. The sequence of the enterococcal cytolysin imparts unusual lanthionine stereochemistry. *Nature Chemical Biology* **9**, 157–159 (2013).

102. Skaugen, M., Andersen, E. L., Christie, V. H. & Nes, I. F. Identification, characterization, and expression of a second, bicistronic, operon involved in the production of lactocin S in *Lactobacillus sakei* L45. *Applied and Environmental Microbiology* **68**, 720–727 (2002).

103. Park, J.-E., Kim, H.-R., Park, S.-Y., Choi, S.-K. & Park, S.-H. Identification of the biosynthesis gene cluster for the novel lantibiotic paenilan from *Paenibacillus polymyxa* E681 and characterization of its product. *Journal of Applied Microbiology* **123**, 1133–1147 (2017).

104. Martins, A. M. *et al.* A two-component protease in *Methylorubrum extorquens* with high activity toward the peptide precursor of the redox cofactor pyrroloquinoline quinone. *Journal of Biological Chemistry* **294**, 15025–15036 (2019).

105. Maruyama, Y., Chuma, A., Mikami, B., Hashimoto, W. & Murata, K. Heterosubunit composition and crystal structures of a novel bacterial M16B metallopeptidase. *Journal of Molecular Biology* **407**, 180–192 (2011).

106. Meindl, K. *et al.* Labyrinthopeptins: A new class of carbacyclic lantibiotics. *Angewandte Chemie - International Edition* **49**, 1151–1154 (2010).

107. Dorenbos, R. *et al.* Thiol-disulfide oxidoreductases are essential for the production of the lantibiotic sublancin 168. *Journal of Biological Chemistry* **277**, 16682–16688 (2002).

108. Claesen, J. & Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proceedings of the National Academy of Sciences* **107**, 16297–16302 (2010).

109. Lee, J. *et al.* Structural and functional insight into an unexpectedly selective *N*-methyltransferase involved in plantazolicin biosynthesis. *Proceedings of the National Academy of Sciences* **110**, 12954–12959 (2013).

110. Grigoreva, A. *et al.* Identification and characterization of andalusicin: N-terminally dimethylated class III lantibiotic from *Bacillus thuringiensis* sv. *andalousiensis*. *iScience* **24**, 102480 (2021).

111. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology* **13**, 470–478 (2017).

112. Dabonné, S. *et al.* Cloning, expression and characterization of a 46.5-kDa metallopeptidase from *Bacillus halodurans* H4 sharing properties with the pitrilysin family. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1725**, 136–143 (2005).

113. Eddy, S. R. Accelerated profile HMM searches. *PLoS computational biology* **7**, e1002195 (2011).

114. Li, Y.-X., Zhong, Z., Hou, P., Zhang, W.-P. & Qian, P.-Y. Resistance to nonribosomal peptide antibiotics mediated by D-stereospecific peptidases. *Nature Chemical Biology* **14**, 381–387 (2018).

115. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

116. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498–2504 (2003).

117. Bushnell, B. BBMap: A fast, accurate, splice-aware aligner. in *9th Annual Genomics of Energy & Environment Meeting* (2014).

118. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

119. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

120. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

121. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

122. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

123. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

124. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).

125. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

126. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

127. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Research* **14**, 1188–1190 (2004).

128. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343–345 (2009).

129. Altenbuchner, J. Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 system. *Applied and Environmental Microbiology* **82**, 5421–5427 (2016).

130. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**, 184–191 (2016).

131. Radeck, J. *et al.* The *Bacillus* BioBrick Box: Generation and evaluation of essential genetic building blocks for standardized work with *Bacillus subtilis*. *Journal of Biological Engineering* **7**, 29 (2013).

132. Bachman, J. Site-directed mutagenesis. *Methods in Enzymology* **529**, 241–248 (2013).

133. Ortega, M. A. *et al.* Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* **517**, 509–512 (2015).

134. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, W296–W303 (2018).

135. Haiser, H. J. & Turnbaugh, P. J. Developing a metagenomic view of xenobiotic metabolism. *Pharmacological Research* **69**, 21–31 (2013).

136. Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289–306 (2012).

137. Dai, H. *et al.* Recent advances in gut microbiota-associated natural products: structures, bioactivities, and mechanisms. *Natural Product Reports* **40**, 1078–1093 (2023).

138. Lavelle, A. & Sokol, H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology* **17**, 223–237 (2020).

139. Skelly, A. N., Sato, Y., Kearney, S. & Honda, K. Mining the microbiota for microbial and metabolite-based immunotherapies. *Nature Reviews Immunology* **19**, 305–323 (2019).

140. Spanogiannopoulos, P., Bess, E. N., Carmody, R. N. & Turnbaugh, P. J. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nature Reviews Microbiology* **14**, 273–287 (2016).

141. Cao, Y. *et al.* Commensal microbiota from patients with inflammatory bowel disease produce genotoxic metabolites. *Science* **378**, eabm3233 (2022).

142. Yao, L. *et al.* A biosynthetic pathway for the selective sulfonation of steroidal metabolites by human gut bacteria. *Nature Microbiology* **7**, 1404–1418 (2022).

143. Fischbach, M. A. Microbiome: Focus on causation and mechanism. *Cell* **174**, 785–790 (2018).

144. Chaudhari, S. N., McCurry, M. D. & Devlin, A. S. Chains of evidence from correlations to causal molecules in microbiome-linked diseases. *Nature Chemical Biology* **17**, 1046–1056 (2021).

145. Chaleckis, R., Meister, I., Zhang, P. & Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology* **55**, 44–50 (2019).

146. Cui, L., Lu, H. & Lee, Y. H. Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases. *Mass Spectrometry Reviews* **37**, 772–792 (2018).

147. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry* **27**, 1897–1905 (2016).

148. Katz, K. *et al.* The sequence read archive: a decade more of explosive growth. *Nucleic Acids Research* **50**, D387–D390 (2022).

149. Mukherjee, S. *et al.* Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Research* **51**, D957–D963 (2023).

150. Blin, K. *et al.* The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research* **47**, D625–D630 (2019).

151. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603–D610 (2023).

152. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).

153. Chiurchiù, V., Leuti, A. & Maccarrone, M. Bioactive lipids and chronic inflammation: Managing the fire within. *Frontiers in Immunology* **9**, 38 (2018).

154. Bae, M. *et al. Akkermansia muciniphila* phospholipid induces homeostatic immune responses. *Nature* **608**, 168–173 (2022).

155. MacEyka, M. & Spiegel, S. Sphingolipid metabolites in inflammatory disease. *Nature* **510**, 58–67 (2014).

156. Walker, A. *et al.* Sulfonolipids as novel metabolite markers of *Alistipes* and *Odoribacter* affected by high-fat diets. *Scientific Reports* **7**, 11047 (2017).

157. Pitta, T. P., Leadbetter, E. R. & Godchaux, W. Increase of ornithine amino lipid content in a sulfonolipid-deficient mutant of *Cytophaga johnsonae*. *Journal of Bacteriology* **171**, 952–957 (1989).

158. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

159. Dziarski, R., Park, S. Y., Kashyap, D. R., Dowd, S. E. & Gupta, D. Pglyrp-regulated gut microflora *Prevotella falsenii*, *Parabacteroides distasonis* and *Bacteroides eggerthii* enhance and *Alistipes finegoldii* attenuates colitis in mice. *PLoS ONE* **11**, e0146162 (2016).

160. Lima, S. F. *et al.* Transferable immunoglobulin A–coated *Odoribacter splanchnicus* in responders to fecal microbiota transplantation for ulcerative colitis limits colonic inflammation. *Gastroenterology* **162**, 166–178 (2022).

161. Hou, L. *et al.* Identification and biosynthesis of pro-inflammatory sulfonolipids from an opportunistic pathogen *Chryseobacterium gleum*. *ACS Chemical Biology* **17**, 1197–1206 (2022).

162. Pasternak, B. A. *et al.* Lipopolysaccharide exposure is linked to activation of the acute phase response and growth failure in pediatric Crohn's disease and murine colitis: *Inflammatory Bowel Diseases* **16**, 856–869 (2010).

163. Pastor Rojo, O. *et al.* Serum lipopolysaccharide-binding protein in endotoxemic patients with inflammatory bowel disease. *Inflammatory Bowel Diseases* **13**, 269–277 (2007).

164. Im, E., Riegler, F. M., Pothoulakis, C. & Rhee, S. H. Elevated lipopolysaccharide in the colon evokes intestinal inflammation, aggravated in immune modulator-impaired mice. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **303**, G490–G497 (2012).

165. Stephens, M. & von der Weid, P.-Y. Lipopolysaccharides modulate intestinal epithelial permeability and inflammation in a species-specific manner. *Gut Microbes* **11**, 421–432 (2020).

166. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105–114 (2021).

167. Liu, Y. *et al.* Identification and characterization of the biosynthetic pathway of the sulfonolipid capnine. *Biochemistry* **61**, 2861–2869 (2022).

168. Vences-Guzmán, M. Á. *et al.* Identification of the *Flavobacterium johnsoniae* cysteate-fatty acyl transferase required for capnine synthesis and for efficient gliding motility. *Environmental Microbiology* **23**, 2448–2460 (2021).

169. Radka, C. D., Miller, D. J., Frank, M. W. & Rock, C. O. Biochemical characterization of the first step in sulfonolipid biosynthesis in *Alistipes finegoldii*. *The Journal of biological chemistry* **298**, 1–13 (2022).

170. Kamiyama, T. *et al.* Sulfobacins A and B, novel von Willebrand factor receptor antagonists: I. Production, isolation, characterization and biological activities. *The Journal of Antibiotics* **48**, 924–928 (1995).

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

172. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* **4**, 293–305 (2019).

173. Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nature Biotechnology* **38**, 1159–1163 (2020).

174. Guo, J., Shen, S., Xing, S., Yu, H. & Huan, T. ISFrag: De novo recognition of in-source fragments for liquid chromatography–mass spectrometry data. *Analytical Chemistry* **93**, 10243–10250 (2021).

175. Folz, J. *et al.* Human metabolome variation along the upper intestinal tract. *Nature Metabolism* **5**, 777–788 (2023).

176. Witting, M. & Böcker, S. Current status of retention time prediction in metabolite identification. *Journal of Separation Science* **43**, 1746–1754 (2020).

177. García, C. A., Gil-de-la-Fuente, A., Barbas, C. & Otero, A. Probabilistic metabolite annotation using retention time prediction and meta-learned projections. *Journal of Cheminformatics* **14**, 33 (2022).

178. Sun, L. *et al.* A simple method for HPLC retention time prediction: linear calibration using two reference substances. *Chinese Medicine* **12**, 16 (2017).

179. Wang, Y. *et al.* A simple method for peak alignment using relative retention time related to an inherent peak in liquid chromatography-mass spectrometry-based metabolomics. *Journal of Chromatographic Science* **57**, 9–16 (2019).

180. Hale, L. P., Gottfried, M. R. & Swidsinski, A. Piroxicam treatment of IL-10-deficient mice enhances colonic epithelial apoptosis and mucosal exposure to intestinal bacteria. *Inflammatory Bowel Diseases* **11**, 1060–1069 (2005).

181. Berg, D. J. *et al.* Rapid development of colitis in NSAID-treated IL-10-deficient mice. *Gastroenterology* **123**, 1527–1542 (2002).

182. Cario, E. Toll-like receptors in inflammatory bowel diseases: A decade later. *Inflammatory Bowel Diseases* **16**, 1583–1597 (2010).

183. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nature Methods* **17**, 905–908 (2020).

184. United States Pharmacopeial Convention. <85> Bacterial endotoxins test. *Pharmacopoeia 44th edn* https://www.usp.org/harmonization-standards/pdg/general-methods/bacterial-endotoxins (2021).

185. Kawasaki, T. & Kawai, T. Toll-like receptor signaling pathways. *Frontiers in Immunology* **5**, 1–8 (2014).

186. Kawai, T. & Akira, S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nature Immunology* **11**, 373–384 (2010).

187. Maeda, J. *et al.* Inhibitory effects of sulfobacin B on DNA polymerase and inflammation. *International Journal of Molecular Medicine* **26**, 521–527 (2010).

188. Shimazu, R. *et al.* MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *Journal of Experimental Medicine* **189**, 1777–1782 (1999).

189. Miyake, K. Roles for accessory molecules in microbial recognition by Toll-like receptors. *Journal of Endotoxin Research* **12**, 195–204 (2006).

190. Park, B. S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature* **458**, 1191–1195 (2009).

191. Su, L. *et al.* Sulfatides are endogenous ligands for the TLR4–MD-2 complex. *Proceedings of the National Academy of Sciences* **118**, 1–12 (2021).

192. Luk, J. M., Kumar, A., Tsang, R. & Staunton, D. Biotinylated lipopolysaccharide binds to endotoxin receptor in endothelial and monocytic cells. *Analytical Biochemistry* **232**, 217–224 (1995).

193. Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. *Nature Reviews Immunology* **8**, 958–969 (2008).

194. Zhang, Y. *et al.* ECM1 is an essential factor for the determination of M1 macrophage polarization in IBD in response to LPS stimulation. *Proceedings of the National Academy of Sciences* **117**, 3083–3092 (2020).

195. Zhang, X. & Mosser, D. Macrophage activation by endogenous danger signals. *The Journal of Pathology* **214**, 161–178 (2008).

196. Parker, B. J., Wearsch, P. A., Veloo, A. C. M. & Rodriguez-Palacios, A. The genus *Alistipes*: Gut bacteria with emerging implications to inflammation, cancer, and mental health. *Frontiers in Immunology* **11**, 906 (2020).

197. Chang, P. V., Hao, L., Offermanns, S. & Medzhitov, R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proceedings of the National Academy of Sciences* **111**, 2247–2252 (2014).

198. Trapecar, M. *et al.* Gut-liver physiomimetics reveal paradoxical modulation of IBD-related inflammation by short-chain fatty acids. *Cell Systems* **10**, 223-239.e9 (2020).

199. Bhattarai, Y. *et al.* Bacterially derived tryptamine increases mucus release by activating a host receptor in a mouse model of inflammatory bowel disease. *iScience* **23**, 101798 (2020).

200. Dodd, D. *et al.* A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. *Nature* **551**, 648–652 (2017).

201. Sato, Y. *et al.* Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians. *Nature* **599**, 458–464 (2021).

202. Paik, D. *et al.* Human gut bacteria produce T_H17-modulating bile acid metabolites. *Nature* **603**, 907–912 (2022).

203. Rakoff-Nahoum, S. & Medzhitov, R. Toll-like receptors and cancer. *Nature Reviews Cancer* **9**, 57–63 (2009).

204. Moreira Lopes, T. C., Mosser, D. M. & Gonçalves, R. Macrophage polarization in intestinal inflammation and gut homeostasis. *Inflammation Research* **69**, 1163–1172 (2020).

205. Bhattarai, Y., Muniz Pedrogo, D. A. & Kashyap, P. C. Irritable bowel syndrome: A gut microbiota-related disorder? *American Journal of Physiology-Gastrointestinal and Liver Physiology* **312**, G52–G62 (2016).

206. Wang, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).

207. Thomann, A. K. *et al.* Depression and fatigue in active IBD from a microbiome perspective-a Bayesian approach to faecal metagenomics. *BMC medicine* **20**, 366 (2022).

208. Mars, R. A. T. *et al.* Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* **182**, 1460-1473.e17 (2020).

209. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368 (2021).

210. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).

211. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

212. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

213. R Core Team. *R: A language and environment for statistical computing*. https://www.R-project.org/ (2021).

214. Oksanen, J. *et al. vegan: Community ecology package*. https://CRAN.R-project.org/package=vegan (2020).

215. Van Rossum, G. & Drake, F. L. *Python 3 reference manual*. (CreateSpace, 2009).

216. McKinney, W. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* 51–56 (2010). doi:10.25080/Majora-92bf1922-00a.

217. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* **17**, 261–272 (2020).

218. Harrell Jr, F. E. Hmisc: A package of miscellaneous R functions. (2014).

219. Erben, U. *et al.* A guide to histomorphological evaluation of intestinal inflammation in mouse models. *International Journal of Clinical and Experimental Pathology* **7**, 4557–4576 (2014).

220. Ellermann, M. *et al.* Endocannabinoids inhibit the induction of virulence in enteric pathogens. *Cell* **183**, 650-665.e15 (2020).

221. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology* **41**, 447–449 (2023).

222. Nothias, L.-F. *et al.* Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *Journal of Natural Products* **81**, 758–767 (2018).

223. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **30**, 2785–2791 (2009).

224. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31**, 455–461 (2010).

225. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling* **61**, 3891–3898 (2021).

226. Schrödinger, LLC. The PyMOL molecular graphics system, version 2.0. (2015).