# New Techniques for Identifying and Studying Microbial Natural Products

# NEW TECHNIQUES FACILITATE THE DISCOVERY AND STUDY OF MODULAR MICROBIAL NATURAL PRODUCTS

By CHAD WILLIAM JOHNSTON, HBSc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree: Doctor of Philosophy

McMaster University Doctor of Philosophy (2016) Hamilton, Ontario (Biochemistry and Biomedical Sciences)

TITLE: New Techniques Facilitate the Discovery and Study of Modular Microbial Natural Products. AUTHOR: Chad William Johnston, Hon. BSc. (McMaster University). SUPERVISOR: Associate Professor Nathan A. Magarvey.

Number of pages: xvi, 341.

## Abstract

Microbial natural products have been one of the most important sources of drugs for the last century. However, as increasing numbers of natural products were discovered, researchers increasingly found previously described compounds, leading to a decline in efficiency and perceived future prospects for natural products discovery. Now, we know from genome sequencing that massive numbers of natural products remain to be discovered, indicating that new strategies and techniques may be required to leverage new bio- and chemo-informatic data to reveal these previously undiscovered small molecules.

First, I developed an automated database search strategy that could reveal natural products in liquid chromatography coupled mass spectrometry (LCMS) data based on the *in silico* fragmentation of automated structure predictions based on genome sequence inputs. This Genomes-to-Natural Products pipeline (GNP) was used as an automated approach for uncovering new modular microbial natural products, identifying structures from cryptic biosynthetic gene clusters as well as from organisms which had not been known to produce natural products previously. Next, I generated a comprehensive library of microbial antibacterial natural products and used a retrobiosynthetic algorithm to identify chemical families with conserved mechanisms of action. This approach led to investigations of the telomycin family of antibiotics, revealing that this old molecule has a novel target. Finally, I led two targeted investigations of organisms that had been identified as potential natural product producers via bioinformatic analysis, revealing new chemical compounds from *Legionella* and *Delftia*. Throughout this thesis, bio- and chemo-informatics have been used to illustrate new approaches to natural products research,

generating novel platform technologies and investigating untouched organisms as a means of rediscovering the potential of microbial natural products.

**Acknowledgements**

First, I'd like to thank my co-workers for making these some of the most enjoyable and productive years of my life. To Morgan and Michael, in particular, I hope that we remain life-long friends. Thank you for all your support during our time both in the lab and beyond.

To Nathan, thank you for putting up with me for all these years. Your mentorship has been the most valuable part of this degree. Thank you for the great conferences, the great projects, and most of all, for your encouragement.

I'd like to thank my committee members, Tim Gilberger, Paul Harrison, and Brian Coombes for their insight and encouragement during my time at McMaster. To Brian Coombes and Gerry Wright, thank you for being so welcoming and helpful during my graduate studies – it really meant a lot to me.

Lastly, I wish to thank my family and friends for their encouragement during my graduate work. To Laura, thank you for your unending love and support, even when I wake you up at 5 in the morning and put my cold feet on you.

*We fill pre-existing forms and when we fill them we change them and are changed.*

– Frank Bidart on “Borges and I”

## Table of Contents

## List of Tables



## List of Figures

## Abbreviations

| | |
|---|---|
| DNA | deoxyribonucleic acid |
| RNA | ribonucleic acid |
| NRP | nonribosomal peptide |
| PK | polyketide |
| NRPS | nonribosomal peptide synthetase |
| PKS | polyketide synthase |
| A | adenylation (domain) |
| T | thiolation (domain) |
| C | condensation (domain) |
| TE | thioesterase (domain) |
| KS | ketosynthase (domain) |
| KR | ketoreductase (domain) |
| DH | dehydratase (domain) |
| ER | enoylreductase (domain) |
| MS | mass spectrometry |
| LC-MS | liquid chromatography coupled mass spectrometry |
| HRMS | high resolution mass spectrometry |
| NMR | nuclear magnetic resonance |

## Declaration of Academic Achievement

This thesis is formatted as a sandwich thesis, with specific details relating to each contribution described in the corresponding chapter preface.

# Chapter 1. Introduction

## 1.1 Thesis Context

Microorganisms interact with their environments through secreted products of metabolism, using small molecules to affect the world around them and facilitate their own growth.[1-3] In contrast to essential primary metabolites – such as amino or nucleic acids that are used to construct crucial biopolymers including proteins, RNA, and DNA – microorganisms use 'secondary' metabolites to interact with their surroundings and with each other.[4] Although these small molecules are non-essential, they assist in adaptation to specific environmental niches or life-styles, frequently making use of exotic chemistry to achieve complex results, such as antibiosis or antagonism.[5] It was the discovery of this antibiotic biological activity that, in the early-to-mid 1900's, led scientists to investigate small molecules – particularly natural products – as potential therapeutic agents.

The publication of Koch's postulates in 1890 highlighted a growing awareness of the etiology of diseases, as well as a desire to treat specific causes of disease. In keeping with this, advances in chemistry and biology demonstrated that certain organisms or chemical compounds could be used to treat human diseases specifically. Indigenous populations had used various plants and other organisms for centuries to help treat disease,[6] and as natural products chemistry techniques became established, scientists were able to reveal specific small molecules responsible for the therapeutic effects. The most well-known and arguably most important example of this reductionist approach was the isolation of the anti-malarial drug quinine from the bark of the *Cinchona* tree, providing a life-saving drug which is still among the most globally important therapeutics nearly 200 years later.[7]

The specific biological activity of quinine – in that it kills the *Plasmodium* parasite responsible for malaria without affecting the patient – inspired scientists to pursue the discovery of new biologically-active small molecules that could be used to treat diseases ranging from bacterial infections to cancer. In the first decade of the 1900's, Paul Erlich had started using bioactivity-based screens of synthetic compounds to discover bioactive molecules against trypanosomes, *Plasmodium*, and most notably, the *Spriochaetes* bacteria responsible for syphilis, which could be treated with the arsenic containing compound Salvarsan.[8] Over the next few decades, the hunt for biologically active small molecules benefitted tremendously from bacteriological studies, as scientists including Alexander Fleming and Selman Waksman demonstrated that, in addition to plants such as *Cinchona*, bacteria and fungi were a rich source of natural products.[5] Using bioactivity assays, microbes that demonstrated overt antibiotic potential could be isolated *en masse* and cultured extensively, yielding many of the well-known natural product producers known today, including the Bacilli and Actinomycetes.[9] From the 1950's to the 1970's, academic and industrial natural products chemists used bioactivity-guided isolation techniques to mine these specific microbial families, generating many of the antibiotics that are still used today, in a period now referred to as the 'Golden Age of Antibiotic Discovery'.[10]

Of the natural products discovered during this golden age, most could be grouped into a relatively small number of structural classes based on monomers used during their assembly, still visible in the final molecule.[11] These molecular superfamilies of secondary metabolites were often related to primary metabolites, and were constructed from similar monomers, including amino acids (for peptide natural products), acetate and malonate (for

polyketides and bioactive fatty acids), isoprene units (for terpenes), nucleic acids (for nucleosides), and sugars (for aminoglycosides), along with combinations thereof.[11] During the 1970's and 80's, feeding studies with isotopically-labelled precursors provided the first verification of the origins of these diverse molecules.[12] As molecular biology and DNA sequencing technologies came about, scientists could find and characterize the genes responsible for natural product production, leading to phylogenetically accurate classifications for these molecules.[11, 13-14] In nearly all cases, genes responsible for the creation of a given natural product were found to be tightly associated with one another in DNA sequences, if not being directly transcriptionally-linked in operons.[11, 15-16] These biosynthetic gene 'clusters' were critical in determining the genetic origins of many natural products, particularly the polyketides and nonribosomal peptides, which are amongst the most structurally and chemically diverse microbial natural products. Early studies into the creation of the polyketide erythromycin[16-17] and the (fungal) nonribosomal peptide cyclosporin[18] uncovered an important distinction for polyketides and nonribosomal peptides that set them apart from other natural products: they were created by massive assembly line-like enzymes.[15-18] This discovery, along with the observation that genes involved with the creation of secondary metabolites were clustered together in genomes, formed the basis for the first predictions of natural product structures and functions based exclusively on genetic information.[19] But before this could happen, biosynthetic studies were needed to establish how these complex molecules were constructed,[11, 15] fueling hope that these enzymatic processes might be harnessed later for directed engineering of natural

products,[20-21] generating molecules that were challenging for contemporary synthetic chemistry techniques.

Given their chemical complexity, important biological activities, and relatively straightforward (or at least easily identifiable) gene cluster architectures, polyketides and nonribosomal peptides received significant attention in studies relating to natural product biosynthesis. It is now known that the assembly line enzymes for nonribosomal peptides (referred to as nonribosomal peptide synthetases; NRPSs) and polyketides (polyketide synthases; PKSs) are composed of multiple multi-domain modules,[15-18] with each module containing domains which facilitate the addition and modification of a single monomer to the growing natural product polymer chain.[22-23] For NRPS systems, monomers are typically amino acids, selected and activated by an adenylation (A) domain and then tethered to the assembly line for chain elongation or further tailoring via a thiolation (T) or peptidyl / acyl carrier protein (PCP or ACP, respectively). Following any associated tailoring reactions, such as O- or N-methylation, the protein-tethered monomer is linked with an upstream monomer via the neighboring modules condensation (C) domain. For NRPSs, this iterative process of activation, tethering, and condensation grows an amino acid polymer on the assembly line enzyme which can be released as a nearly complete (or fully completed) natural product via a C-terminal chain-release domain.[24] As the thiolation domains utilize a post-translationally attached phosphopantathiene arm to tether activated monomers via a thioester, many of the chain release domains are thioesterase (TE) domains, which use a conserved serine residue to facilitate their own esterification, before utilizing water or a natural product-associated nucleophile to achieve release of a linear or macrocyclized

molecule.[24] Although this is largely considered the standard strategy for chain release, similar results can be achieved via NADPH-dependent reductase domains[25] or by domains with homology to condensation domains,[26] among other approaches. PKSs that are known as 'type 1' use a similar module-based activation, tethering, and chain elongation system, which results in a similar architecture to NRPSs, as well as compatibility for hybridization.[3, 26] Acyl transferase (AT) domains effectively replace A domains, selecting CoA-versions of small organic acids (such as malonate or methyl malonate) which are briefly attached to the AT domain via a conserved serine, before being transferred to an associated thiolation domain.[15, 27] Ketosynthase (KS) domains effectively replace C domains, using a cysteine residue to briefly capture the monomer loaded on an upstream T domain before catalyzing C-C bond formation with a KS-mediated decarboxylated malonate attached on the module-associated thiolation domain. The poly-ketone chain generated by this process is highly reactive (leading to intramolecular cyclizations observed in the iterative type 2 and 3 polyketides[27]), and is typically reduced to some extent by module-associated domains. Ketoreductase (KR) domains work on β-ketoacyl substrates, reducing them to β-hydroxyacyls. This is can be followed by dehydratase (DH) domains that reduce this alcohol to form an α,β-enoyl, which can then be reduced further by an enoylreductase (ER) domain to form a saturated acyl. Although NRPS modules can possess associated monomer-tailoring domains in addition to the standard C, A, and T domains, this is relatively uncommon, while it is the norm in type 1 PKS systems.[15, 27] In both cases, varying monomer selection, module organization, assembly line length, and tailoring allows for the creation of incredibly complex natural products with a variety of biological activities.

Despite the inherently bioactive nature of microbial products, rediscovery of known molecules from well-explored bacterial families limited the efficiency of discovery programs,[28-31] and led to many to believe that industrial natural products programs were becoming prohibitively expensive. Meanwhile, the rise of modern techniques for synthetic and analytical chemistry were used to promote a variety of alternative means for developing drug-like molecules, including strategies such as combinatorial chemistry[32] and diversity-oriented synthesis.[33] Although these strategies were useful for creating a variety of kinase inhibitors and G-protein coupled receptor directed agents,[34] they were an ineffective source of antimicrobial agents relative to natural products (yielding only the oxazolidinone class of antibacterials, which work through inhibition of the ribosome, analogous to many microbial natural products).[35] Although synthetic chemistry efforts had generally been ineffective in generating antibacterial drug scaffolds (with notable exceptions being the sulfonamides [1930's], oxazolidinones [1990's], and fluoroquinolones [1980's]),[36] large investments into synthetic chemical libraries and high-throughput screening facilities led to continued screening approaches for identifying antibiotics. During the 1990s, the first bacterial genome sequences raised hope that new antimicrobial targets could be identified and screened extensively to uncover new drugs and molecules, leading to an extensive period of target-based drug discovery, with tailored assays being used to uncover promising leads from libraries of synthetic compounds.[37] Unfortunately, this target-based approach proved to be even less productive than previous explorations of synthetic libraries for new antibiotics, in part because it diverged from the productive phenotypic screens[38] which had been used to identify potent antibiotics in the past, and did not yield molecules with efficacy

in whole-cell or infection-based assays (though well designed target-based strategies did typically prevent the re-isolation of known compounds). While much of this work proved fruitless – and may have contributed to the 'discovery gap' for antibiotics, leading to the current antibiotic resistance crisis[39-40] – there were silver linings for future research. Microbial natural products, as we now understand them, are privileged molecules,[41] having evolved over many thousands of years[42] to reach a potent and often highly selective chemical scaffold.[43-44] Chemically, the complexity of natural products is often emphasized by the number of $sp^3$-configured centers found in these molecules, in contrast to the $sp^2$-rich synthetic molecules developed by combinatorial chemistry efforts.[45] Referring specifically to lessons learned from target-based discovery efforts, we now understand that the evolutionary pressures involved in the development of antimicrobial natural products leads to highly refined target selection. As we now know (defined in part by my own work – highlighted in **Chapter 3**), nearly all antibacterial targets have associated natural products which have been developed and honed by evolution,[46-48] validating their usefulness for future drug development. Microbial natural product discovery still had many problems which would need to be overcome before industrial discovery programs were viable again, but in light of the results from a few decades of synthetic efforts, it had become apparent that natural products remained the most valuable source of inherently bioactive molecules, particularly for antimicrobial drugs and scaffolds.[36, 41, 49] While rediscovery rates may have indicated that very few new natural products remain to be found, the same genome sequencing efforts which were used to justify target-based discovery screens would soon demonstrate that natural products discovery may be approaching a second golden age.

Having access to DNA sequences is critical for modern biological sciences, as they provide the basis for understanding and manipulating genotypes and phenotypes in all living organisms. Before the year 2000, DNA sequencing had been both time consuming and expensive, and sequencing entire genomes was a colossal effort. Following the successful sequencing of the first DNA-encoded bacteriophage genome in 1977,[50] incremental technical advances led to the first sequenced bacterial genomes[51-52] nearly 20 years later. After this however, new 'next-generation' sequencing techniques[53] – including those implemented by sequencing companies like Illumina and 454 – began providing high quality bacterial genome sequences at a fraction of the cost and time, and whole genome sequences started an exponential climb that has continued to this day. One of the most important findings to emerge from these new genomes for the natural products community was that in most genomes there were many, many more gene clusters for producing natural products than there were molecules that had been isolated from the sequenced organism.[52, 54-55] Early studies of well-established producers (such as the *Streptomyces*) estimated that only 10% of natural products had been found thus far.[54] As microbial genomes became easy to obtain and process, scientists could now use these newly identified genes to predict and find their associated small molecules, an approach known as 'genome mining'.[19]

As a result of studies into the biosynthetic mechanisms underlying microbial natural product creation, scientists in the early 2000's discovered the codes for monomer-specificity for adenylation[56-57] and acyl transferase domains,[58] facilitating the first predictions of natural product structures based on gene sequences.[19] Although this process could be performed manually – using sequence alignments of adenylation or acyl

transferase domains that had been described previously to predict the likely substrate of an new domain[19] – this was rarely the case, as bioinformatic tools were quickly developed that could perform this function with an increasing degree of accuracy.[59-61] However, during the mid- to late-2000's, these predictions could still only be performed by persons with considerable insight into NRPS and PKS systems, as monomer predictions only provide some of the likely pieces to a hypothetical structure, and must be assembled manually with consideration to other detected biosynthetic domains. However, there were some indications that this situation could be resolved in time, as Ecopia Biosciences Inc. had demonstrated in 2003 that their in-house DECIPHER software could generate relatively accurate predictions from sequenced NRPS-PKS assembly lines.[62] These co-linear assembly line enzymes, from which predicted monomers were assembled in a linear sequence, represented the first examples of biosynthetic gene clusters that could be accurately predicted, but these are hardly the norm. Even in this simple case, deviations from standard biosynthesis hindered the successful implementation of genetic predictions from the very beginning, as coelichelin – the first nonribosomal peptide predicted and found by genome-mining – could not be predicted correctly due to NRPS functions that differed from typical systems.[63]

Despite the growing atmosphere of optimism surrounding the explosion in the number of sequenced bacterial genomes and the prospect of using this data to find new natural products, by 2010 there were no defined strategies for how to predict and find these molecules *en masse*. In 2011, Kersten et al. outlined a peptidogenomic approach for identifying amino acid-based natural products in LC-MS/MS data based on their predicted

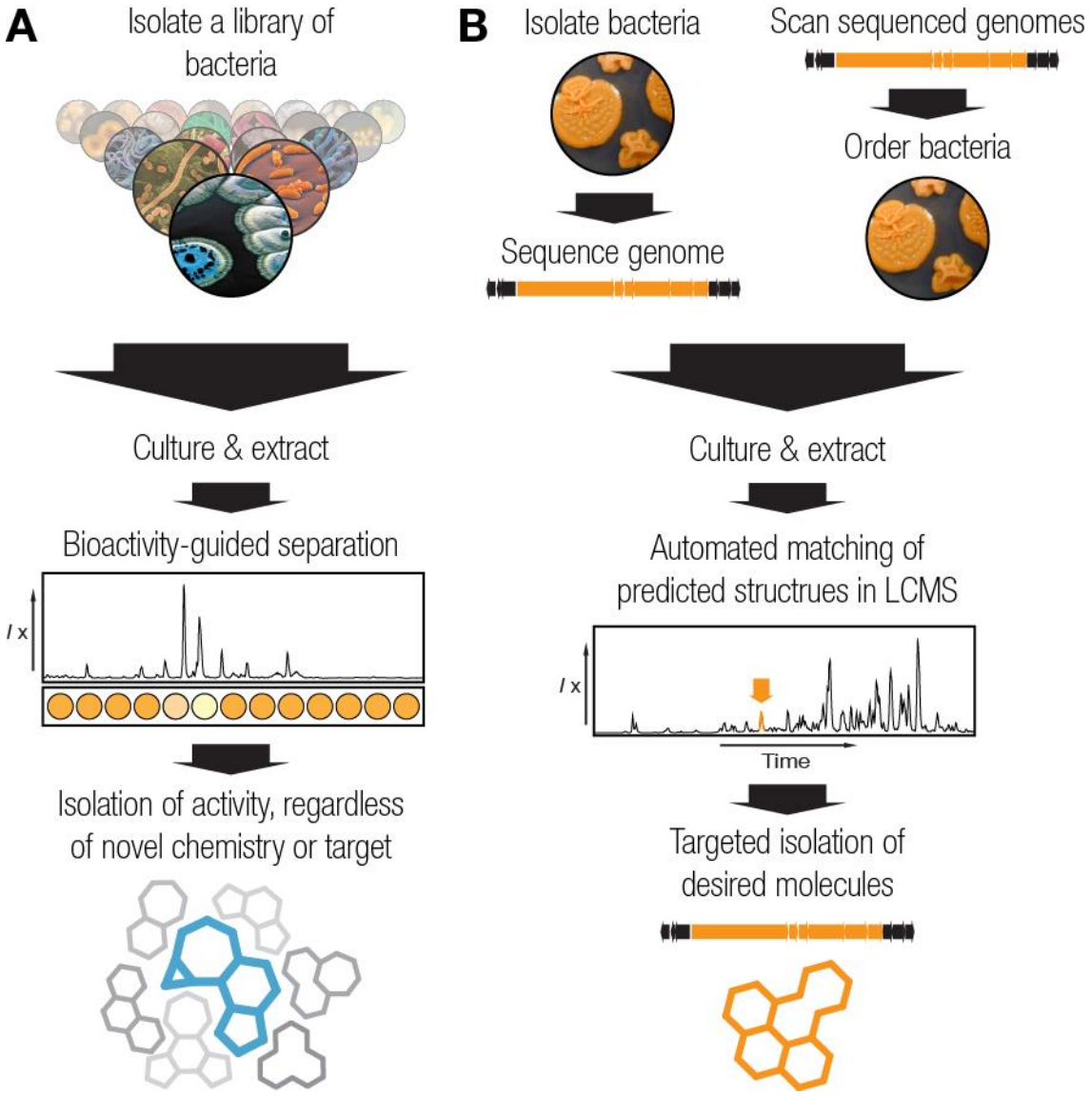



**Figure 1.1** Overview of traditional and modern strategies for microbial natural products. **A.** Traditional discovery programs culture large collections of bacteria and extract them to generate extract libraries. Bioactive extracts are serially fractionated to obtain the active components. **B.** Modern discovery platforms use small molecule predictions and genome sequencing to enrich for promising gene clusters, using chemoinformatic tools to facilitate automated detection and isolation of desired small molecules.

As a result of studies into the biosynthetic mechanisms underlying microbial natural product creation, scientists in the early 2000's discovered the codes for monomer-specificity for adenylation[56-57] and acyl transferase domains,[58] facilitating the first predictions of natural product structures based on gene sequences.[19] Although this process could be performed manually – using sequence alignments of adenylation or acyl transferase domains that had been described previously to predict the likely substrate of an new domain[19] – this was rarely the case, as bioinformatic tools were quickly developed that could perform this function with an increasing degree of accuracy.[59-61] These semi-automated programs still required that users input specific protein sequences for their PKS or NRPS queries, and rather than manually aligning them, these programs would create alignments in a similar manner to standard bioinformatic software. This necessity limited the applicability of these early programs, preventing the detection of NRPS-PKS monomers from genome sequences, and necessitating tedious BLAST exercises to identify biosynthetic gene clusters. Because of this alignment strategy, variations in primary protein sequence from the small number of experimentally validated adenylation or acyltransferase domains meant that misalignment of any of the key amino acids that determine monomer activation specificity (which are conserved in their position and function, but not in their structure or identity) could result in predictions being wildly inaccurate. A more general concern was species specificity or bias, as monomer activating domains were generally obtained from prolific and well-studied producers, like *Streptomyces*, Cyanobacteria, Myxobacteria, and *Bacillus*. This bias decreased the general accuracy of these relatively simplistic alignment-based, and limited the application of these predictions to these well-

studied bacterial families. These caveats, along with the poor appreciation of the biosynthetic potential of bacteria outside of the standard known producers, limited interest in these useful programs. During the mid- to late-2000's, these predictions could still only be performed by persons with considerable insight into NRPS and PKS systems, as monomer predictions only provide some of the likely pieces to a hypothetical structure, and must be assembled manually with consideration to other biosynthetic domains, which would have to be predicted and assessed for potential function manually. However, there were some indications that this situation could be resolved in time, as Ecopia Biosciences Inc. had demonstrated in 2003 that their in-house DECIPHER software could generate relatively accurate predictions from sequenced NRPS-PKS assembly lines.[62] These co-linear assembly line enzymes, from which predicted monomers were assembled in a linear sequence, represented the first examples of biosynthetic gene clusters that could be accurately predicted, but these are hardly the norm. Even in this simple case, deviations from standard biosynthesis hindered the successful implementation of genetic predictions from the very beginning, as coelichelin – the first nonribosomal peptide predicted and found by genome-mining – could not be predicted correctly due to NRPS functions that differed from typical systems.[63] Rearranged assembly line genes or shuffled gene clusters have continued to hinder the accuracy of automated predictions until very recently, with manual investigation often being required to set the order of natural product assembly. Aside from this however, new strategies and techniques for predicting natural products have considerably increased the accuracy of predictions, as well as the diversity of the organisms from which structure predictions can be obtained.

Despite the growing atmosphere of optimism surrounding the explosion in the number of sequenced bacterial genomes and the prospect of using this data to find new natural products, by 2010 there were no defined strategies for how to predict and find these molecules *en masse*. For instance, with coelichelin – the earliest example of a natural product discovered through genome mining – a knockout strategy was enacted in order to identify the true natural product; in effect, this eliminates the utility of creating the structure prediction in the first place. In contrast to these sorts of laborious genetic exercises, a number of groups began to converge of mass spectrometry based approaches, leveraging lessons from proteomics to identify natural products or structure predictions with amino acid components. In 2011, Kersten et al. outlined a peptidogenomic approach for identifying amino acid-based natural products in LC-MS/MS data based on their predicted monomers,[64] which marked the first defined strategy for prediction guided discovery of natural products. While scanning MS data, researchers could identify peptides by looking for clean spectra containing amino acid fragments that could be assembled in a linear sequence. By comparing this sequence to structure predictions of nonribosomal or ribosomal natural products in the genome (as well as to sequences of known molecules) these MS and MS/MS spectra could be associated with a given prediction. While this approach was certainly useful in a conceptual sense, it still relied entirely on manual user input, utilizing manual structure predictions, manual generation of hypothetical MS/MS fragment masses, and manual assessment of LC-MS/MS data. Most importantly, this manual intervention mandated that structures be assessed in both N-to-C and C-to-N fragmentation pathways, as ribosomal natural products often process complex cyclic or

partially-cyclic scaffolds, which complicate analysis of fragmentation patterns. Given that amino acid fragments were identified manually, these linear series could be matched with an accuracy never achieved before with early automated programs. However, new algorithms could soon provide similar accuracy without the need for excessive manual intervention. Shortly after this workflow was published, we introduced the iSNAP algorithm for identifying known nonribosomal peptides within LC-MS/MS data in an automated manner.[65] As peptides reliably fragment along amide bonds during MS/MS fragmentation, this fragmentation could be applied to known structures *in silico*, yielding a library of hypothetical parent and daughter ions that can be searched for and scored in LC-MS/MS data to find the associated known compound with unprecedented accuracy. Using this database-dependent search strategy, researchers could now detect known natural products within crude extracts without manual intervention. Moving forward, new fragment matching strategies could be developed for the detection of new, predicted molecules.[66]

Following technological and methodological advances in the last decade, we are now approaching a thorough understanding of natural product biosynthesis, facilitating increasingly accurate predictions from biosynthetic gene clusters.[67] Analytical chemistry techniques and instrumentation have also improved considerably, as mass spectrometers and other devices demonstrate higher sensitivity than ever before, making the detection of natural products increasingly straightforward. Most importantly, genome sequencing has advanced markedly in even the last 5 years, with over 30,000 publically available bacterial genomes sequenced at the time of writing, provided at record speeds and low prices.

However, in spite of these technical advances, many of the core problems of natural products research remain unsolved, particularly, how to make use of emergent chemical and genetic data to arrive at new or desired natural products. My thesis seeks to address this issue, using new informatics-based platforms and strategies to discover and study microbial natural products.

### 1.2 Scope and nature of this work

Natural products discovery has long been a crucial source of new bioactive chemical scaffolds and lead molecules,[41] but has suffered from diminishing returns after relying on tired, established techniques for the better part of a century.[28-30] As genome sequencing became widespread, it was apparent that many natural products had been overlooked, and raised the question of how new genetic and chemical data could be used to uncover molecules that had been missed in the past. In this thesis, I explore the use of bio- and chemo-informatic strategies and techniques to discover microbial natural products, both acutely, with automated molecule detection algorithms, and more broadly, with strategies to prioritize chemical or biological families for study. Four published scientific articles are used here to demonstrate my work towards the goal of developing informatic techniques and strategies to reveal new molecules.

### 1.3 Informatic platform development for the discovery and study of microbial natural products.

During the Golden Age of antibiotic discovery, thousands of microbial natural products were uncovered,[68] revealing chemical scaffolds and associated bioactivities, as well as organisms, culture conditions, and extraction techniques required to access these valuable molecules. Stemming from this success and the inherent value of microbial natural products, researchers have invested considerable effort in understanding the biosynthesis of these molecules, which are often produced by multi-modular assembly line-like enzymes.[11, 15] Now, with the emergence of advanced computational and 'big data' tools, we can use bio- and chemo-informatics to leverage extant data and get the most from past discoveries and emergent genome sequencing results. As the first aim of my thesis, I explore how developing new informatic pipelines can be used to direct efforts towards desired molecules, both from new genome sequences and from databases of known compounds.

With the advent of rapid and affordable bacterial genome sequencing, it has become apparent that many more natural products await discovery, both from previously described producers[14, 54] and from new sources.[69] Although structures can be predicted and used for targeted discovery efforts, this is a complicated process that requires considerable insight and often leads to predictions that differ from the true product due to promiscuous substrate selection or deviations from expected tailoring.[63] As such, new automated approaches are needed that can be accessible tools for genome-guided natural products discovery efforts. In **Chapter 2**, I describe the development of GNP, the Genomes-to-Natural Products pipeline, which can automatically identify NRPS and PKS gene clusters, predict their products, facilitate the construction of *in silico* chemical libraries related to these

predictions, and then search LC-MS/MS data for molecules related to the predicted structures. GNP was used to identify known molecules from previously undescribed gene clusters, to reveal new molecules from previously uninvestigated organisms, and to find glycosylated natural products – including polyketides. We also demonstrate that GNP can be used to identify cryptic natural products that had eluded detection and isolation previously by identifying and isolating thanamycin, a new syringomycin-like antifungal molecule. Collectively, we show that automated informatic platforms can be a valuable tool for revealing natural products from LC-MS/MS data, and hopefully, for supplying new leads for natural product drug development.

In the Golden Age of discovery, researchers revealed the structures and activities of thousands of microbial natural products,[68] leading to several valuable scaffolds that had been successful in the clinic until very recently.[40, 49] Now, as antibiotic resistance challenges the efficacy of these molecules, new scaffolds must be developed that can avoid cross-resistance.[70] Although over-looked organisms can be a source for these molecules,[71-73] previous work has already yielded thousands of compounds which had been described, but never developed, meaning that there should already be many valuable lead compounds available if only they can be correctly pulled out from their disparate sources. In **Chapter 3**, we compile a comprehensive collection of microbial antibacterial natural products and use a retrobiosynthetic algorithm to break-down many of them into chemical substructures that can be aligned and sorted into phylogenetically-related families through hierarchical clustering. We use this analysis of natural product families to describe the frequency at which natural product antibiotic scaffolds possess novel targets, and to define a number of

chemical families that may possess new molecular targets. From this list, we explored the mechanism of action of the telomycin-family of natural products using suppressor mutant genome sequencing along with *in vitro* assays. We found that telomycin possesses a novel antibacterial target – the phospholipid cardiolipin – demonstrating the value of this chemo-informatic clustering approach to enrich for promising scaffolds that can be mined for the next generation of antibacterial drugs.

### 1.4 Informatic directed investigations for novel sources of microbial natural products.

Following the discovery of penicillin in 1928, microbial natural products had received substantial attention as a source of bioactive molecules and drug-like chemical scaffolds, peaking between the late 1940's to the 1970's. During this time, academic and industrial researchers developed many of the techniques and methods that have defined modern natural products drug discovery. The identification of productive organisms such as actinomycetes[5, 9] and *Bacillus* (and later cyanobacteria[74] and myxobacteria[75]) led to selective culturing techniques and the generation of extensive, targeted microbial libraries that could probe the biosynthetic potential of these genera. This focused research and development strategy led to a wealth of natural product drugs, but these selective culturing techniques meant that many (if not most) bacteria were not included in natural product discovery efforts until very recently, indicating that many productive bacteria remain to be explored. In light of new advances in genome sequencing, we can now use informatic strategies to guide investigations of previously untouched organisms and direct efforts

towards gene clusters that are unrelated to described molecules, a goal which I set as the second aim of my thesis.

Even before genome sequencing became widespread, it was understood that certain natural product classes or chemical scaffolds were enriched in specific bacterial genera or families. This knowledge helped facilitate targeted discovery efforts to enrich for specific natural products, such as glycopeptides from actinomycetes. In a similar sense, this also implies that exploring new bacterial genera or families would be a useful strategy for uncovering new natural product chemical scaffolds – an idea which appears to be corroborated by emergent genome data. In **Chapter 4**, we explore the genus *Legionella* as a potential source of novel natural products using our in-house bio- and chemo-informatic software PRISM. *Legionella* genomes were found to contain a number of polyketide and nonribosomal peptide gene clusters that are not found in other bacteria, indicating that they would likely produce novel natural products. Following this, we initiated investigations into *L. pneumophila*, generating genetic knockouts that could be used for comparative metabolomic analysis. As a result of these investigations, we discovered legionellol, a unique natural product isolated from *L. pneumophila*. With this in hand, we are hopeful that continued study of *Legionella spp*. will yield more new natural products and chemical scaffolds for development.

Microbial natural products are often referred to as ‘secondary metabolites’, in that their functions are non-essential and largely assist in adaptation to specific environmental niches.[4] Now, as genome sequencing reveals previously unknown biosynthetic gene clusters from unstudied microbes, unique environmental niches are useful starting points

to find molecules that are unique, and may also provide insight into the life-styles of their producers.[76-77] In **Chapter 5**, I investigate the organism *Delftia acidovorans* – one of the few bacteria known to live on gold deposits[78] – and demonstrate that this unusual lifestyle is facilitated by a novel microbial natural product: delftibactin. Using genetic information, we predict and identify delftibactin in crude extracts, isolate it, and solve its structure using mass spectrometry and nuclear magnetic resonance spectroscopy (NMR). Following a number of in vitro assays, we discovered that delftibactin detoxifies ionic gold solutions at concentrations observed in liquid cultures, allowing *D. acidovorans* to survive exposure in the laboratory, and likely also in its native environment. Upon closer investigation with transmission electron microscopy, it was discovered that delftibactin facilitates the creation of gold nanoparticles with similar architecture to those observed in secondary gold deposits,[79] providing a new theory for the creation of these structures via biomineralization. By combining a variety of *in vitro* assays, analytical chemistry techniques, and new bioinformatic predictions, we could investigate a previously unstudied organism from a unique environmental niche and uncover a new natural product that facilitates a rare bacterial life-style.

### 1.5 Thesis overview

This thesis explores new means of finding and studying modular microbial natural products using bio- and chemo-informatic approaches. Using our knowledge of natural product biosynthesis and MS-fragmentation, I constructed a discovery platform that could reveal structures and locations of new natural products in an automated manner. Second,

after collecting and curating known antibacterial natural products, I used a retrobiosynthetic algorithm to reveal chemical families with potentially novel mechanisms of action, leading to investigations of telomycin and demonstration of its unique target. Third, by using bioinformatic platforms, I mapped out *Legionella spp*. as a new source of novel natural products, proving this by isolating a new molecule. Lastly, I investigated a previously unstudied bacteria from a unique environmental niche, and used informatic predictions to discover a new molecule that facilitated this bacteria's unusual life-style. This selection of projects that I executed during my doctoral studies showcases the potential that new informatic strategies possess when used to solve emergent challenges in the study and discovery of microbial natural products.

# Chapter 2. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products

## 2.1 Chapter preface

To develop an automated natural product identification program that could leverage extant genomes to guide the discovery of new molecules, we created the Genomes-to-Natural Products pipeline (GNP). This new program incorporated an early version of the bio- and chemo-informatic structure prediction algorithms which were later incorporated into PRISM, along with an interface where users can combinatorialize predicted scaffolds to build libraries of hypothetical products of a given NRPS or PKS gene cluster. LC-MS/MS data could be inspected for the presence of predicted molecules by comparing parent and fragment masses observed in MS/MS data to hypothetical fragments of combinatorial library compounds which had been fragmented *in silico*. We first demonstrated the efficacy of the GNP platform by investigating a novel NRPS gene cluster from *Streptomyces calvus*, using automated predictions of the product of this gene cluster to identify the known molecule WS9326. Next, we investigated organisms that had not previously been known to produce natural products, yielding three new NRPS-PKS molecules from *Acidovorax* and *Variovorax* bacteria. After incorporating a computational framework for the prediction of glycosylation events, we identified a glycosylated polyketide macrolide from *Nocardiopsis potens*. Although this was the first polyketide identified by our approach, we found that standards representing a large number of diverse modular natural products could be readily identified with GNP, indicating that future

improvements to the structure prediction engine will produce a more comprehensive natural product discovery tool. To show that GNP could be used to identify cryptic natural products that had evaded detection or isolation previously, we investigated a *Pseudomonas* bacterium which possessed a gene cluster for thanamycin – an established cryptic natural product. Using GNP, we could detect thanamycin in trace amounts, and could establish a nearly flawless structure prediction, guiding downstream spectroscopic investigations which provided the final structure of thanamycin. In contrast to our previous natural product discovery engines, GNP has a considerably expanded range of structural classes that can be considered, and it uniquely incorporates the wealth of available genome sequence data to discover completely new natural product structures.

The following chapter is a modified version of previously published journal article in which I was a lead author. For this work, I cultured bacteria, performed GNP-mediated compound identification and isolation, elucidated structures of potensimicin and thanamycin, contributed to study design, and wrote the manuscript. Michael Skinnider designed and created GNP, contributed to study design and wrote the manuscript. Morgan Wyatt cultured bacteria, identified and isolated compounds, and contributed to study design. Xiang Li elucidated the structures of acidobactin, vacidobactin, and variobactin. Michael Ranieri assisted with potensimicin isolation. Lian Yang assisted with GNP development. Prof. David Zechel contributed *Streptomyces calvus* and the elucidated WS9326 compound. Prof. Bin Ma assisted in GNP development, and Prof. Nathan Magarvey contributed to study design and wrote the manuscript. The citation for this publication is as follows:

**Johnston, C.W.***, Skinnider, M.A.*, Wyatt, M.A., Li, X., Ranieri, M.R.M., Yang, L., Zechel, D., Ma, B., & Magarvey, N.A. (2015) An Automated Genomes-to-Natural Products Platform (GNP) for the Discovery of Modular Natural Products. *Nature Communications* **6**, 8421.

## 2.2 Abstract

Bacterial natural products are a diverse and valuable group of small molecules, and genome sequencing indicates that the vast majority remain undiscovered. The prediction of natural product structures from biosynthetic assembly lines can facilitate their discovery, but highly automated, accurate, and integrated systems are required to mine the broad spectrum of sequenced bacterial genomes. Here, we present a genome-guided natural products discovery tool to automatically predict, combinatorialize, and identify polyketides and nonribosomal peptides from biosynthetic assembly-lines using LC-MS/MS data of crude extracts in a high-throughput manner. We detail the directed identification and isolation of six genetically predicted polyketides and nonribosomal peptides using our Genome-to-Natural Products platform (GNP). This highly automated, user-friendly program provides a means of realizing the potential of genetically encoded natural products.

## 2.3 Introduction

Natural products are valuable small molecules, whose unique and diverse chemical scaffolds have made them an important source of human therapeutics[1] and industrial

agents[2]. Polyketides and nonribosomal peptides are two of the most important[3] and diverse[4] classes of these secondary metabolites, and are constructed by assembly line-like enzymes known as polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs)[5]. With the advent of rapid and inexpensive bacterial genome sequencing, a wealth of orphan NRPS and PKS gene clusters have been uncovered in publicly accessible genomes (>25,000 c. 2015) of both well-[6] and under-studied[7, 8] microbes, prompting renewed enthusiasm for discovery of new natural products[9]. Chemical structures or key monomers of polyketides and nonribosomal peptides can be postulated from genetic information[10-15], but available computational tools for identifying compounds within complex mass spectral data generally require extensive knowledge and manual annotation of specific organisms[16-18], metabolites[19], and MS data[20-22]. Moreover, current tools available to partially automate these processes may require formal training in bio- or chemoinformatics or computer science to achieve results. The development of workflows to connect genomic to metabolomic data has significantly advanced the study of natural products, but now, highly automated and user-friendly software is required in order to access the wealth of genetically encoded natural products in both new and old microbial producers in a high-throughput context. Here, we present GNP, a Genomes-to-Natural Products platform, as an accessible and automated tool that can generate and utilize natural product predictions to directly identify desired small molecules in LC-MS/MS data, to facilitate the re-engagement of microbial libraries for discovering targeted molecules *en masse*, and for uncovering the remaining majority of genetically encoded polyketides and nonribosomal peptides. GNP is available as a web application and can be accessed at http://magarveylab.ca/gnp/.

## 2.4 Results

### Development of GNP

To expedite the discovery of genetically-encoded polyketides and nonribosomal peptides, we developed GNP to automatically predict and locate these metabolites within LC-MS/MS data of crude microbial extracts (**Fig. 2.1a** and **Supplementary Fig. 1-4**). GNP predicts chemical structures by identifying gene clusters, modules, domains, and domain substrate specificities with a series of hidden Markov models and curated BLAST databases (**Supplementary Fig. 1**), which are well-suited for defining biosynthetic gene clusters[23], adenylation domains[13], and other substrate-specific enzymes. Automatically generated predictions are forwarded to GNP's browser rendered structure combinatorialization interface[24, 25] where scaffolds may be modified to construct libraries of hypothetical products, in order to account for biosynthetic promiscuity, variation, or inaccurate predictions (**Supplementary Fig. 2**). Combinatorial libraries of predicted chemical scaffolds are loaded alongside our previously published structure library[26], and are fragmented *in silico* along a series of well documented fragmentation pathways, including water losses, amide cleavages, and ester cleavages. Natural product identification is achieved by matching *in silico* fragments of these known and predicted metabolites to real MS/MS fragments, using validated scoring algorithms[26] to locate molecules in LC-MS/MS chromatograms (**Fig. 2.1a**, and **Supplementary Fig. 3** and **4**). This profiling of parent and fragment ions from *in silico* and real LC-MS/MS data allows GNP to identify putative substructures and probability scores to directly locate the products of orphan NRPS and PKS gene clusters. In addition to a browser-rendered spreadsheet, a deconvoluted

prediction-guided discovery chart is provided with each GNP report, displaying hits for user-defined predicted structures and confidence scores for predicted structures alongside their retention times in a pseudo-chromatogram (**Supplementary Fig. 4**). To validate that this automated discovery tool could use genes to find natural products, we investigated orphan NRPS, PKS, and hybrid gene clusters from a diverse series of bacterial phyla.

As a test of our automated discovery pipeline, we chose to investigate a novel NRPS gene cluster identified within the genome of *Streptomyces calvus* (ATCC No. 13382; **Supplementary Table 1**). This unusual cluster was found to possess two trans-acting adenylation-thiolation didomains, along with an initial acylating condensation domain common to lipopeptides[27] (**Fig. 2.1b** and **Supplementary Fig. 5a**). GNP generated a predicted product scaffold that could be combinatorialized to automatically generate a library of 768 hypothetical molecules (**Supplementary Fig. 5b**). This predicted structure library was fragmented *in silico* and used to survey for potential matches in LC-MS/MS data of a *S. calvus* culture extract. GNP identified a series of metabolites eluting after 43 minutes in the *S. calvus* LC-MS chromatogram, corresponding to the predicted molecule calvus735 (**Fig. 2.1c**, **Supplementary Fig. 5c**, and **Supplementary Fig. 6**). Isolation of the indicated metabolites led to the identification of the nonribosomal peptides WS9326A and WS9326C[28], products of this hitherto undescribed NRPS gene cluster (**Fig. 2.1d, Supplementary Table 2-3**). Conveniently, because GNP uses MS/MS fragment matching to determine hits, it was capable of assigning large portions of the WS9326 prior to structure determination by NMR spectroscopy.

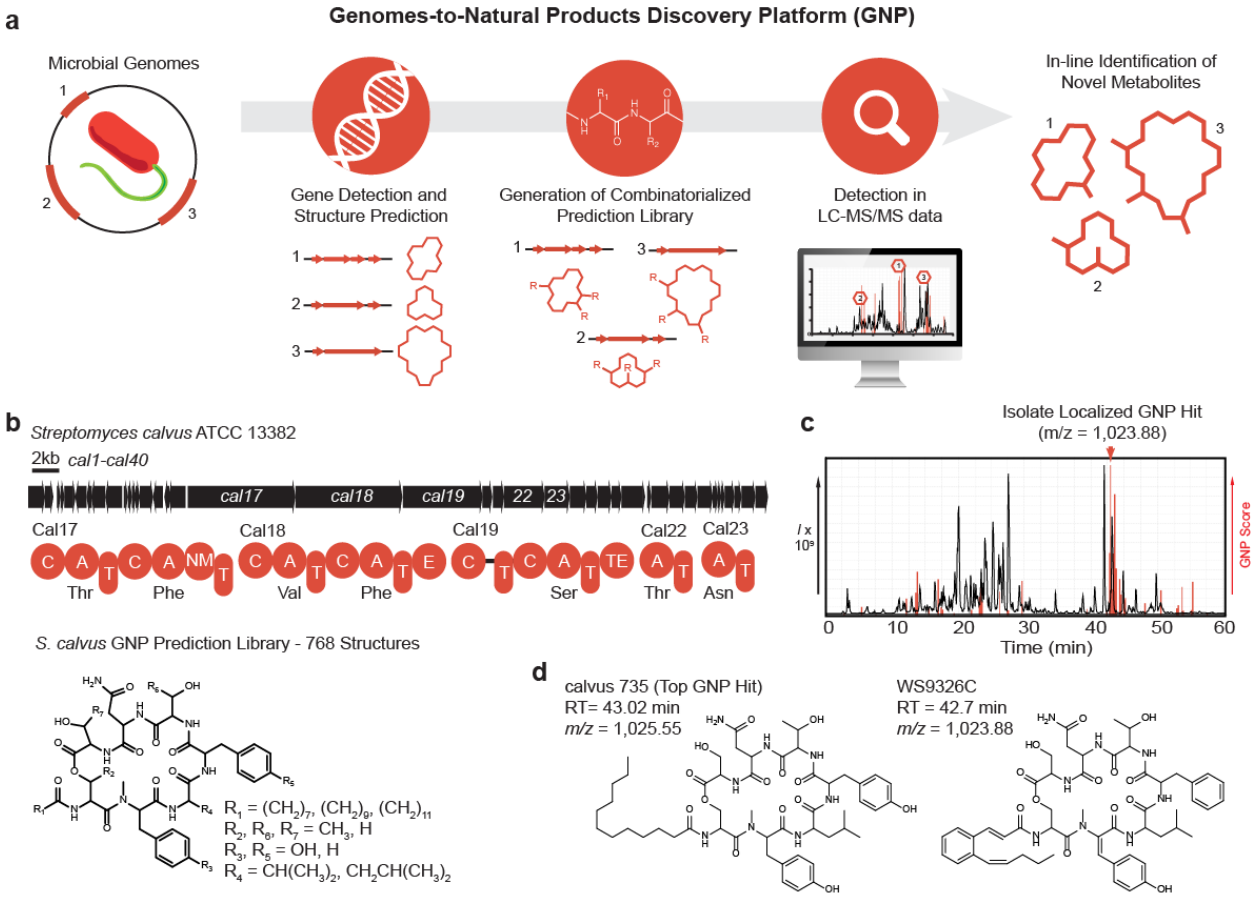



**Figure 2.1** The Genomes-to-Natural Products Discovery Platform (GNP). (**a**) The automated GNP pipeline processes submitted sequences to identify NRPS and PKS gene clusters and yield predicted structures. These predictions can be combinatorialized and elaborated to account for biosynthetic promiscuity, creating libraries of hypothetical structures that are used to search LC-MS/MS data and automatically reveal the true genetically-encoded natural product. (**b**) A novel nonribosomal peptide biosynthetic gene cluster identified within *S. calvus*, alongside the combinatorialized GNP-generated structure prediction. (**c**) LC-MS/MS chromatogram of an *S. calvus* culture extract with GNP result indicating a localized genetically-predicted structure. (**d**) Chemical structures

of the top scoring GNP-generated structure prediction (left) and the corresponding natural product, WS9326C (right), depicted with corresponding retention times (RT) and mass to charge ratios (m/z).

**New Nonribosomal Peptides from *Acidovorax* and *Variovorax***

While *Streptomyces* are well known producers of natural products, extensive microbial genome sequencing has revealed NRPS and PKS machinery in exotic, untouched branches of the microbial tree of life[29]. In light of recent genome-guided discoveries[7, 8, 30], we chose to investigate a series of previously unstudied Proteobacteria. A novel NRPS-PKS cluster was discovered in the genome of *Acidovorax citrulli* (DSM No. 17060), with GNP automatically generating a predicted scaffold molecule (**Supplementary Fig. 7a**, **Supplementary Table 4**). This prediction was combinatorialized for potential macrocyclization or amino acid substitutions, and to account for neighboring N-acetyl and N-formyltransferases (**Supplementary Fig. 7b**). The resulting library of 576 potential structures was used to automatically reveal targeted metabolites present in LC-MS/MS data, detecting a series of hits eluting after 12.1 minutes (**Supplementary Fig. 7c**). Isolation of the most substantial hits yielded novel polyketide/nonribosomal peptide natural products – the first from this genus – acidobactins A and B, whose final structures were solved by NMR (**Supplementary Fig. 7d**, **Supplementary Fig. 8- 11**, **Supplementary Table 5**). The only significant point of variation between the predicted and final structures stems from the unusual initiating adenylation domain, acting as a fatty acyl-AMP ligase (FAAL)[31], for which no specificity codes have yet been revealed. To correct this, we

constructed an extensive FAAL database of domains from established assembly lines to enable substrate prediction. As a test of this improved predictive capacity we investigated an acidobactin-like gene cluster in the related and similarly uninvestigated organism *Variovorax paradoxus* S110 (DSM No. 30034; **Fig. 2.2a**, **Supplementary Table 6**). Automated scaffold generation returned a prediction that was identical to acidobactin but with an enabled FAAL substrate prediction (**Supplementary Fig. 12a**). Following combinatorialization that was identical to that of the acidobactin scaffold (**Supplementary Fig. 12b**), GNP located two peaks from the *V. paradoxus* S110 LC-MS/MS sample (**Fig. 2.2a**). Isolation of these indicated metabolites provided the novel compounds vacidobactin A and B, which are methylmalonate-incorporating acidobactins, and represent the first natural products isolated from *Variovorax* (**Fig. 2.2a, Supplementary Fig. 12c**, **Supplementary Fig. 13-16**, **Supplementary Table 5**). A second *Variovorax* isolate (*V. paradoxus* P4B[32]) was found to possess another distinct NRPS/PKS gene cluster, corresponding to a putative lipopeptide (**Fig. 2.2b**, **Supplementary Fig. 17a**, **Supplementary Table 7**). By altering amino acid composition, macrocyclization status, and tailoring modifications, we generated a library of 32 structures (**Supplementary Fig. 17b**) for use in GNP analysis of the LC-MS/MS spectrum. The crude extract of *V. paradoxus* P4B was shown to contain a series of candidate hits (**Fig. 2.2b**) whose isolation yielded the novel natural products variobactin A and B, in spite of several unforeseen amino acid substitutions (**Fig. 2.2b**, **Supplementary Fig. 17c**, **Supplementary Fig. 18-20**), as confirmed by MS/MS annotation (**Supplementary Fig. 21**) and NMR spectroscopy (**Supplementary Fig. 18**, **Supplementary Table 8**, **Supplementary Note 1**).

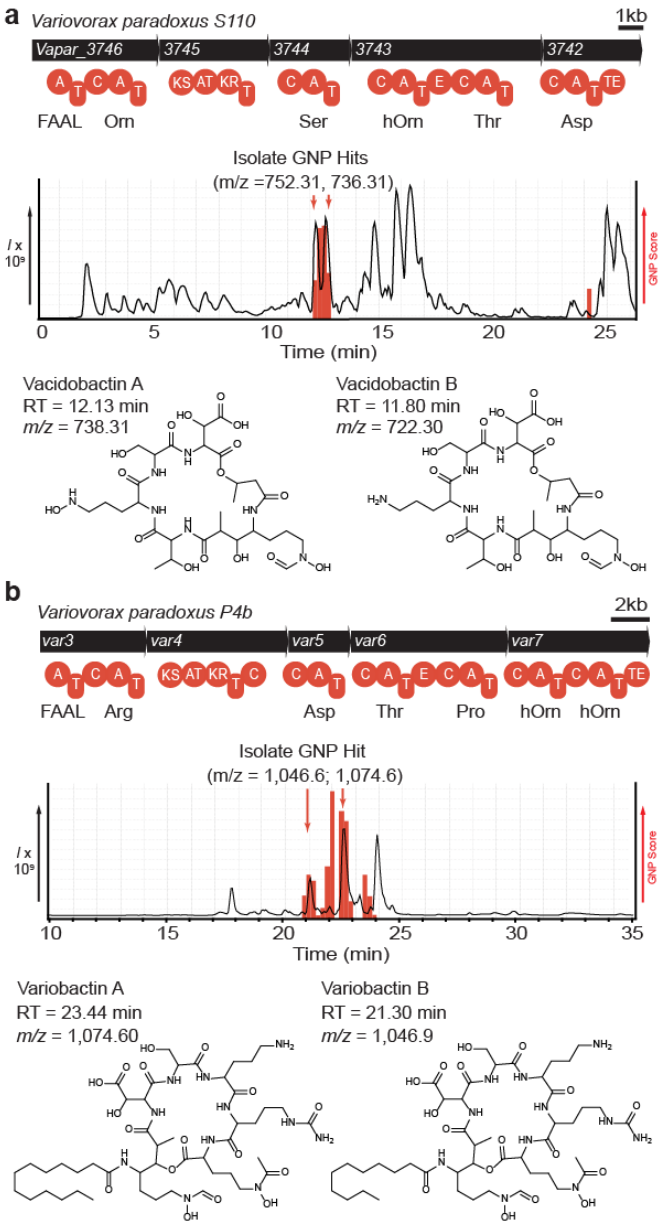



**Figure 2.2** Identifying novel hybrid nonribosomal depsipeptides from unexplored *Variovorax* strains. (**a**) GNP analysis of a novel NRPS-PKS gene cluster in *V. paradoxus* S110 provided a prediction that was combinatorialized and used to query corresponding LC-MS/MS data with GNP, revealing the true genetically-encoded natural products. Isolation of the indicated compounds yielded novel metabolites vacidobactin A and B,

depicted with corresponding retention times (RT) and mass to charge ratios (m/z). (**b**) GNP analysis of a novel NRPS-PKS gene cluster in *V. paradoxus* P4B provided a structure prediction that was combinatorialized and used to query LC-MS/MS data with GNP. Isolation of the indicated compounds yielded novel metabolites variobactin A and B, depicted with corresponding retention times (RT) and mass to charge ratios (m/z).

**Discovery of Glycosylated Natural Products and Polyketides**

In addition to the prediction of nonribosomal peptides, GNP is capable of predicting polyketide structures, as demonstrated by the prediction of ketide units within the acidobactins, vacidobactins, and variobactins. Although modular (type 1) polyketides are more readily predictable than other classes of microbial natural products[4, 5] such as terpenes, their MS/MS fragmentation patterns are less information-rich than those of peptidic natural products. However, polyketide biosynthetic gene clusters frequently possess enzymatic machinery for the biosynthesis of highly modified deoxysugar moieties[22, 33], whose characteristic masses can facilitate the identification of polyketides in a tandem mass spectrum[34]. In order to automate the prediction of deoxysugars from genomic information, we constructed a library of hidden Markov models corresponding to important families of deoxysugar biosynthesis genes[22, 35] (**Supplementary Table 9**). We revised and expanded the glycogenomic code developed by Kersten et al.[22], in particular by extending their logic to the biosynthesis of pentose deoxysugars[36] (**Supplementary Table 10;** See *Methods*). Novel logic was required to predict deoxysugars from sequence data, as polyketides often contain multiple sugars[37], enzymes are shared between

biosynthetic pathways[38], and these pathways may not be physically segregated within a given cluster[38]. We therefore developed an algorithm to predict sugar combinations. The number of deoxysugar glycosyltransferases is determined based on a homology search (**Supplementary Fig. 22**), and all possible sugar combinations of that size are iteratively evaluated based on an analysis of whether the identified deoxysugar pathway genes are both necessary and sufficient for the biosynthesis of a given combination (see **Methods**). Subsequent to user-directed combinatorialization, each scaffold in the combinatorial library is then glycosylated with each sugar combination at a random hydroxyl group to produce a glycosylated scaffold library. Glycosidic bond cleavage can then be enabled based on the identification of oxan-2-ol substructures.

To test these new models and assess whether GNP could also be used both to predict and detect glycosylated polyketides from crude extracts, we decided to search for the product of a unique deoxysugar and PKS gene cluster found in the genome of *Nocardiopsis potens* (DSM 45234; **Fig. 2.3a**, **Supplementary Fig. 23a**, **Supplementary Table 11**). Processing this biosynthetic gene cluster with GNP yielded a polyketide backbone which could be tailored with two predicted deoxysugars, mycaminose (or its diastereomer ravidosamine) and angolosamine. By combinatorializing cyclization, sugar appendages, and methylmalonate or malonate incorporation, we developed a library of 42 hypothetical structures (**Supplementary Fig. 23b**). To assess whether our hypothetical polyketides could be present in a *N. potens* extract LC-MS/MS data file we enabled fragmentation of potential macrocycle esters, losses of ketoreductase generated hydroxyls, as well as glycosidic cleavages. GNP revealed four related peaks within our extract, including a

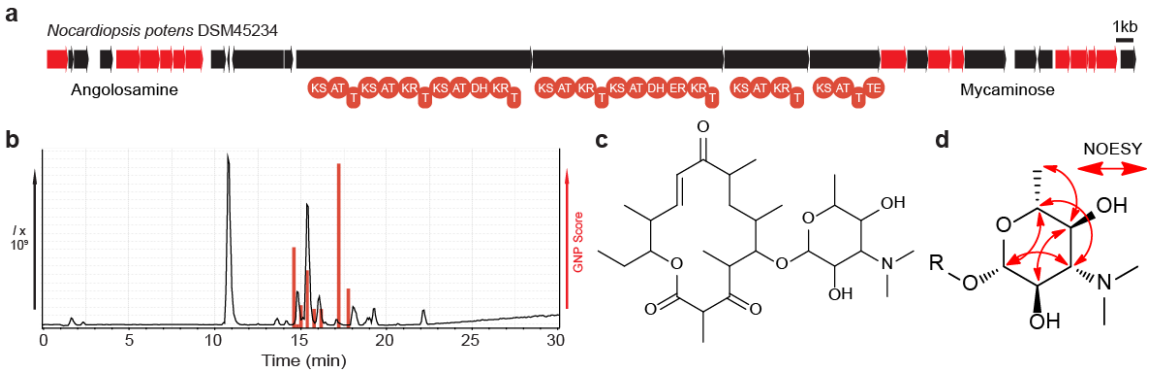



**Figure 2.3** Automated prediction, detection, and structure elucidation of a glycosylated polyketide from *Nocardiopsis potens*. (**a**) A biosynthetic gene cluster containing machinery for deoxysugar and polyketide biosynthesis was identified in the genome of *N. potens* (DSM 45234). The GNP generated polyketide and sugar predictions were combinatorialized to yield a library of 42 hypothetical products. (**b**) GNP analysis of LC-MS/MS data from a *N. potens* extract revealed four related candidate peaks. The most abundant of the detected hits - predicted polyketide 10 (**c**) – was isolated for structure elucidation by NMR. (**c**) Structure of the isolated glycosylated polyketide potensimicin. (**d**) NOESY NMR spectroscopy of the potensimicin deoxysugar demonstrates it is mycaminose.

consensus match for the most abundant hit, predicted polyketide 10 (**Fig. 2.3bc**, **Supplementary Fig. 23c**, **Supplementary Fig. 24-26**). Isolation and NMR-based structure elucidation of this putative polyketide revealed a planar structure identical to the matched prediction, including a narbonolide-type macrocyclic polyketide scaffold[39] and an O-linked mycaminose or ravidosamine deoxysugar (**Fig. 2.3c**, **Supplementary Table 12-13**). However, while LC-MS/MS-based techniques can provide useful information to assist in

structure elucidation, they cannot readily provide insight into stereochemsitry. Analysis of the potensimicin NOESY NMR spectrum and anomeric carbon coupling constant ($J_{hz}$ = 7.06)[40] demonstrated that the potensimicin deoxysugar was β-mycaminose (**Fig. 2.3d**). Potensimicin appears similar to mycaminose-modified narbonolide compounds that have been observed previously[41-42], and shares similar bioactivity profiles (**Supplementary Table 14**).

**Identification and Structure Elucidation of Thanamycin**

Biosynthetic gene clusters are being uncovered at an increasing rate, presenting new opportunities for the discovery of clinically relevant polyketides and nonribosomal peptides. However, low abundance or cryptic metabolites have proven to be a significant bottleneck in genome-guided efforts[43-45], as they are challenging to detect, and require inordinate processing to yield sufficient quantities of pure material for structure elucidation with NMR[45]. GNP has demonstrated itself to be a quick and sensitive means of detecting natural products, with routine natural product identification occurring with only nanograms of material reaching our mass spectrometer (**Supplementary Table 15**). In addition, MS-filters and fragment matching allow GNP to narrow in on predicted molecules that are highly similar to the final product, providing assistance for downstream structure elucidation efforts. To demonstrate the utility of this approach, we used GNP to investigate the cryptic lipopeptide thanamycin[43], whose production, isolation, and elucidation have proven elusive despite a series of genetic[43] and mass spectrometry-based studies[18]. We identified a thanamycin biosynthetic gene cluster in the recently sequenced genome of

*Pseudomonas fluorescens*[46] (DSM No. 11579; **Fig. 2.4a**, **Supplementary Fig. 27a**, **Supplementary Table 16**). To reveal the physical location of this cryptic metabolite in chromatograms, we used GNP to generate a structure library (**Fig. 2.4b** and **Supplementary Fig. 27b**) that would identify the most similar chemical structure from a complex extract. GNP utilized hypothetical MS-fragment matching to indicate one candidate structure from the library of 120 predictions, revealing a low abundance molecule, with *m/z* = 1,291 $[M+H]^+$ (**Fig. 2.4c** and **Supplementary Fig. 27c**). We first sought to verify our GNP match by manual tandem-MS annotation (**Supplementary Fig. 28**), confirming the partial amino acid sequence proposed previously[18], as well as an additional threonine or homoserine residue, as had been predicted by GNP. Next, $^{13}C$-amino acid incorporation was used to confirm the presence of the predicted ornithine (**Supplementary Fig. 29**) and threonine-derived residues (**Supplementary Fig. 30**), while providing further evidence for the presence of homoserine. As a conclusive means of assigning its structure, we undertook substantial efforts to obtain sufficient material for NMR experiments. Thanamycin had been initially identified at levels of < 0.1 mg $L^{-1}$, which was increased to 0.33 mg $L^{-1}$ through culture and purification optimization, reaching a final yield of 40 mg from 120 L of low volume cultures, sufficient for accessing the structure of thanamycin by NMR spectroscopy (**Fig. 2.4d**, **Supplementary Fig. 31-32**, **Supplementary Table 17**). Our observations demonstrated that GNP had incorrectly assigned the third amino acid as asparagine instead of the observed and originally predicted aspartic acid, though it did correctly detect the presence of ornithine as the second amino acid. Surprisingly, the source of the mass discrepancy between the predicted and observed

thanamycin structures appears to be an exceptionally rare hydroxylation at the ornithine α-carbon, a modification previously observed in various diketopiperazines[47-48] and the lipodepsipeptide skyllamycin[49]. This unusual modification may be associated with the potent antifungal activity of thanamycin, which is 32 times that of the related natural product syringomycin E. (**Supplementary Table 14**). This final example illustrates the manner in which GNP provides a rapid means of predicting, locating, and partially-solving the structures of assembly-line derived natural products, such as modular polyketides, nonribosomal peptides, and glycosylated natural products.

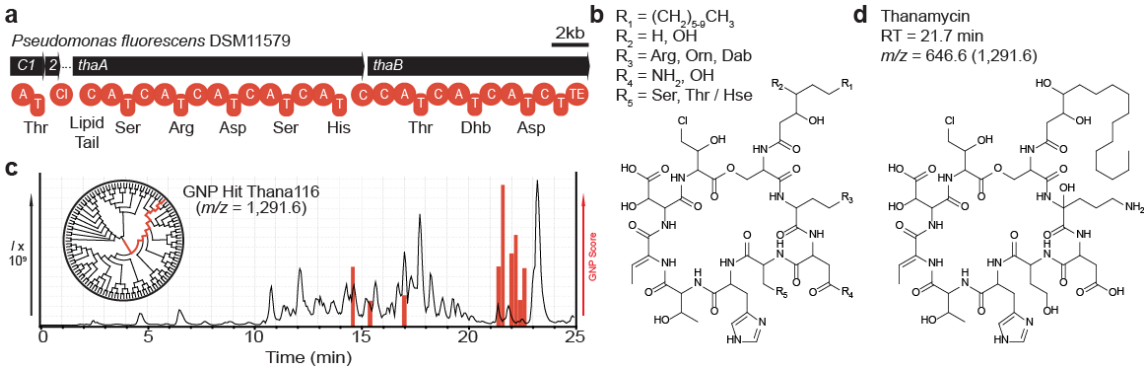



**Figure 2.4** GNP-facilitated detection and structure elucidation of the cryptic nonribosomal peptide thanamycin (**a**) Biosynthetic gene cluster for thanamycin identified within *P. fluorescens* DSM11579. (**b**) Combinatorialization of the GNP-generated structure prediction. (**c**) GNP prediction guided discovery chart indicates a series of related thanamycin-like ions from a *P. fluorescens* extract, including the main ion (1,291.6 m/z) which was found to possess the most structural similarity to predicted structure Thana116 from 120 hypothetical variants, shown as a chemoinformatic tree clustered by chemical similarity. (**d**) Structure of thanamycin with corresponding retention time (RT) and observed mass to charge ratio (m/z).

## 2.5 Discussion

The evolved bioactivities and unique chemical structures of microbial natural products have made them a valuable source of pharmaceutical and industrial agents[1, 2, 50]. The revelation that even well-studied producers, such as *Streptomyces coelicolor*[6, 51], possess gene clusters for many more natural products than they are known to produce has sparked renewed interest in natural product discovery[52]. A number of novel methods have therefore been developed in recent years in an attempt to bridge the gap between genomic potential and natural product discovery. Initial work to reveal and sequence complex nonribosomal peptides in mass spectral data built on advances in *de novo* sequencing techniques of proteomics, allowing researchers to reveal structural information about cyclic peptides[53]. Simultaneously, the development of computational approaches for predicting nonribosomal peptide adenylation domain specificities facilitated accurate prediction of NRPS small molecule products[10]. By bridging these advances in genomics-based structure prediction and MS-driven sequence annotation, Kersten et al. defined a manual peptidogenomic workflow to identify peptidic natural products – both ribosomal and nonribosomal – based on manually generated structure predictions and manually annotated mass tags, marking gaps between $MS^2$ fragment ions that could be annotated as amino acids[21]. A similar workflow was developed for the discovery of glycosylated natural products, combining manually annotated mass tags with the manual analysis of biosynthetic gene clusters[22]. Recently, various degrees of computer automation have been applied to the peptidogenomics workflow. In Pep2Path[20], a series of manually annotated mass tags from a chosen peak of interest can be queried against genomic data processed by

antiSMASH 2.0[54] and NRPSPredictor2[10] using a Bayesian algorithm, in order to associate user-annotated amino acid sequences with NRPS gene clusters. Recent advances such as RiPPQuest[55] and NRPQuest[56], developed by Mohimani *et al*., represent efforts to integrate lanthipeptide and nonribosomal peptide prediction and detection, based on genomic and MS/MS data, within a single software package. Molecular networking is a spectral alignment tool capable of clustering MS scans with closely related fragmentation patterns[18, 57], although determining the identity of clustered ions and the validity of their association requires a high degree of manual interrogation of results[16-19], including extensive annotation with known molecules[19], gene clusters[16, 17], or genetic knockouts[18] to identify ions of interest, typically through their extensive similarity to known compounds. For a summary and comparison of recent genomic- and metabolomic-based natural products discovery methods, see **Supplementary Figure 33**.

GNP's structure prediction engine expands the chemical search space relative to previously published methodologies (**Supplementary Figure 33**) by providing an integrated platform for the prediction of nonribosomal peptide, type I polyketide, and deoxysugar-containing natural products. Importantly, GNP does not rely on the manual annotation of a mass spectrum[16-18] or candidate peak[20], and thereby significantly increases throughput. While the aforementioned methodologies have represented important advances in linking genomic to metabolomic data in the context of natural product discovery, many are distributed as binaries[20] or with minimal user interfaces[55, 56] that may render them inaccessible to many users without formal bioinformatics or computer science training. In contrast, GNP offers a continuous workflow for genomic and metabolomic discovery,

integrated into a single web application with a user-friendly interface accessible to chemists or microbiologists without chemo- or bioinformatics training. Further, GNP provides access to data and matched fragments used to generate candidate scores and probable structures for detected natural products, available as easily interpreted, information-rich, and toggle-able report tables. Our library of hidden Markov models and BLAST databases represent an advance over support vector machine-based prediction methods[13], enabling the generation of accurate predictions across bacterial phyla and thereby facilitating the discovery of novel compounds from bacteria that have not previously been known to produce natural products (**Fig. 2.2**). As GNP reproduces common MS/MS fragmentation pathways such as amide-, ester-, and glycosidic cleavages *in silico* to generate appropriate fragments to match and detect candidate hits, the scope of natural products to which it can be applied is primarily limited by the availability of methods for reliable automated structure prediction. A recent manual workflow for hypothetical structure enumeration and fragment matching described by Zhang et al.[58] also appears to show promise in predicting correct structures of purified ribosomal peptide natural products using custom-tailored high resolution mass spectral techniques, and suggests the automated combinatorialization and identification program presented in GNP may be also applicable to these diverse molecules given an appropriate automated structure prediction methodology.

Although LC-MS based platforms have seen extensive development and use in the past decade, they are susceptible to limitations of MS-centric chemical detection. First, specific chemical configurations cannot be absolutely determined by mass spectrometry alone. LC-MS/MS based approaches can provide evidence for a given structure, but NMR

remains necessary to reveal bond order and stereochemistry. Second, the sensitivity of MS-based detection is based both on the abundance and ease of ionization of a given molecule, such that not all molecular species will be detected equally. Third, $MS^2$ rarely provides a completely comprehensive sampling of plausible fragmentation pathways. Although GNP is capable of generating and matching a diverse series of MS/MS fragments, only a small fraction of these are typically observed and complete coverage of a given molecule is rare, particularly for larger structures. Unlike approaches that utilize manually-annotated mass tags[20-22], GNP does not require user-intervention to identify MS/MS fragmentation pathways, and is capable of sampling all potential pathways simultaneously, providing results rapidly and facilitating a high-throughput workflow. Despite the limitations of LC-MS/MS, it remains the primary method for the rapid interrogation of complex mixtures of bacterial metabolites. While NMR spectroscopy will remain a necessity for structure elucidation for the foreseeable future, informatic MS/MS based platforms such as GNP represent a powerful means of linking metabolomic and genomic data for the discovery of novel assembly line-derived natural products.

In this work, we present a Genomes-to-Natural Products platform to predict and identify assembly-line derived natural products from a growing and diverse array of sequenced bacterial genomes. By providing an accessible and user-friendly interface, and automating the detection of natural products from biosynthetic assembly-lines, GNP represents a high-throughput tool for the discovery of uncovering novel small molecules without the need for genetic knockouts or engineering, bioactivity testing, or extensive manual MS annotation. This unique approach highlights the advantages of automated and

integrated informatic discovery tools by leveraging advances in genomics and metabolomics to access novel natural products.

## 2.6 Materials and Methods

### 2.6.1 General Experimental Procedures

1D ($^{1}H$ and $^{13}C$) and 2D ($^{1}H$-$^{13}C$ HMBC, HSQC, NOESY, and COSY) NMR spectra for acidobactins, vacidobactins, and variobactin were recorded on a Bruker AVIII 700 MHz NMR spectrometer in $D_2O$ ($D_2O$; Cambridge Isotope Laboratories). NMR spectra for WS9326A were recorded on Bruker Avance 500 and 600 MHz instruments in DMSO-$d_6$. High resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (*ThermoFisher Scientific, USA*) with an electrospray ionization source (ESI) and using CID with helium for fragmentation. LCMS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna C18 column (150 mm or 250 mm × 4.6 mm, *Phenomenex*) for analytical separations, running acetonitrile with 0.1% formic acid and dd$H_2O$ with 0.1% formic acid as the mobile phase. $MS^n$ measurements were made by direct infusion to a Bruker AmazonX mass spectrometer. All data required for GNP identification of the compounds identified in this study, including .FASTA files, formatted combinatorial structure libraries, and .mzXML files, are available for download from http://magarveylab.ca/data/gnp/.

### 2.6.2 Microbial Strains

*Acidovorax citrulli* AAC00-1, *Variovorax paradoxus* S110, and *Nocardiopsis potens* were ordered from the German Resource Centre for Biological Material (DSMZ, DSM No**.** 17060, 30034, and 45234) and cultured on Acidovorax Complex Media[30] (*A. citrulli*, *V. paradoxus*; ACM) or Bennett's media (*N. potens*). The environmental isolate *Variovorax paradoxus* strain P4B was found in a soil sample collected from McMaster University in 2010 and maintained on tryptic soy broth (TSB) media. *Streptomyces calvus* was obtained from the ATCC (No. 13382) and was cultured on mannitol soy agar. *Pseudomonas fluorescens* was obtained from DSMZ (DSM No. 11579) and was regularly maintained on LB agar. All bacteria were cultured at 30°C.

### 2.6.3 Fermentation and Small Molecule Isolation

***WS9326A.*** The production medium consisted of 10 g $L^{-1}$ dextrin, 10 g $L^{-1}$ tryptone, 2 g $L^{-1}$ NaCl, 2 g $L^{-1}$ $(NH_4)_2HPO_4$, 1.5 g $L^{-1}$ $KH_2PO_4$, 0.5 g $L^{-1}$ $K_2HPO_4$, 0.25 g $L^{-1}$ $MgSO_4 \cdot 7H_2O$. The pH of the medium was adjusted to 7.2 prior to autoclaving. After autoclaving 5 mL $L^{-1}$ of the following trace element solution (sterile filtered) was added: 2 g $L^{-1}$ $MgSO_4$, 2 g $L^{-1}$ $ZnSO_4 \cdot 7H_2O$, 2 g $L^{-1}$ $FeSO_4 \cdot 7H_2O$, 2 g $L^{-1}$ $MnCl_2 \cdot 4H_2O$, 2 g $L^{-1}$ $CaCl_2 \cdot 2H_2O$, 2 g $L^{-1}$ NaCl, 0.4 g $L^{-1}$ $CuCl_2 \cdot 2H_2O$, 0.4 g $L^{-1}$ $B(OH)_3$, 0.2 g $L^{-1}$ $Na_2MoO_4$, 0.2 g $L^{-1}$ $CoCl_2$, and 2.2 g $L^{-1}$ sodium citrate. Single colonies of *S. calvus* grown on mannitol soy agar were used to initiate 25 mL cultures of tryptic soy broth and grown to high cell density over three-to-four days at 28°C, 180 rpm. The starter culture was used to inoculate 4 × 500 mL production medium (1% v/v) in 2 L baffled shake flasks containing steel springs, which were then incubated for four days at 180 rpm, 28°C. The cells were removed by

centrifugation and the culture supernatant (1.8 L) was mixed with Diaion HP-20 resin (10 g $L^{-1}$) for 3 h. The resin was collected by vacuum filtration and washed with water (2 × 300 mL) then acetone (3 × 200 mL). The acetone washes were concentrated *in vacuo* and the resulting aqueous residue extracted 3 times with an equivalent volume of ethyl acetate. The organic fractions were dried over anhydrous $Na_2SO_4$ then concentrated to dryness *in vacuo*. The crude residue was dissolved in methanol (1 mL) then injected onto a Biotage system equipped with a Biotage Flash 25+M reverse phase column (KP-C18-HS, 25 mm × 150 mm, 48 mL column volume). WS9326A was resolved by eluting the column with 10% acetonitrile / water, (0.5% acetic acid) for 150 mL, followed by a gradient of 10% acetonitrile / water, (0.5% acetic acid) to 90% acetonitrile / water (0.5% acetic acid) over 350 mL, then holding the final solvent composition for 150 mL, all at a flow rate of 10 mL $min^{-1}$. Fractions containing WS9326A were pooled and concentrated to dryness, yielding 5 mg of pure compound.

***Acidobactins, vacidobactins, and variobactin.*** Fresh colonies of *A. citrulli* AAC00-1 or *V. paradoxus* S110 were inoculated into 2.8 L glass Fernbach flasks containing 1 L of Acidovorax Complete Media (ACM)[30]. Environmental strain *V. paradoxus* P4B colonies were used to inoculate 2.8 L glass Fernbach flasks containing 1 L water, 10 g casitone, 1 g $MgSO_4 \cdot 7\ H_2O$, 1 g $CaCl_2 \cdot 2\ H_2O$, 50 mM HEPES buffer, and 20 g $L^{-1}$ Diaion HP-20 resin with pH adjusted to 7.0[20]. Cultures were grown at 30°C, shaking at 190 rpm for three days, after which *A. citrulli* AAC00-1 and *V. paradoxus* S110 cells were pelleted by centrifugation at 7000 rpm for 15 min, adding HP-20 resin to the supernatant at 20 g $L^{-1}$,

and subsequently shaking for 2 h at 220 rpm. Resins were harvested by Büchner funnel filtration and washed with 400 mL of distilled water before eluting three times with 400 mL of methanol. The methanol eluents were evaporated to dryness under rotary vacuum. Acidobactins and vacidobactins were purified using a Luna 5 μm $C_{18}$ column (*Phenomenex*, 250 × 10 mm). The mobile phase was 2% acetonitrile with 0.1% formic acid, and 98% water with 0.1% formic acid at 2 minutes, increasing to 9% acetonitrile at 23 min at a flow rate of 6 mL $min^{-1}$. Acidobactin A eluted at 15.5 min, acidobactin B eluted at 15.9 min, vacidobactin A eluted at 15.7 min, and vacidobactin B eluted at 16.2 min. Variobactin was purified using a Luna 5 μm $C_{18}$ column (*Phenomenex*, 250 × 15 mm). The mobile phase was 5% acetonitrile with 0.1% formic acid, and 95% water with 0.1% formic acid at 0 min with a flow rate of 2.5 mL $min^{-1}$ increasing to 8 mL $min^{-1}$ at 1.5 min for an additional 3.5 min. The gradient increased in a linear fashion from 5 to 10 min to 10% acetonitrile then from 10-52 min the gradient was linear to 50% acetonitrile. Variobactin A eluted at 38 min.

***Potensimicin***. For production of potensimicin, *N. potens* was first grown in 50 mL of a media containing: 10 g $L^{-1}$ molasses, 10 g $L^{-1}$ meat extract, 10 g $L^{-1}$ peptone, and 10 g $L^{-1}$ soluble starch for 2 days at 250 rpm and 28°C. Following this, 10 mL of starter culture was used to inoculate 1 L of the same media. Cultures were grown for 3 days at 250 rpm and 28°C before being harvested by centrifugation at 7000 rpm, followed by methanol extraction of the cell pellet, and Diaion HP-20 (20 g $L^{-1}$) extraction of the supernatant. Methanol eluent of the HP-20 resin was pooled with the methanol extract of the cell pellet, dried under rotary vacuum, and re-dissolved in methanol. The extract was separated on an

open gravity column of LH-20 size exclusion resin (*Sephadex*) with methanol as a mobile phase. Fractions containing potensimicin were pooled, evaporated to dryness, and resuspended in methanol. Potensimicin was isolated by preparative scale LC-MS using a Luna 5 μm $C_{18}$ column (*Phenomenex*, 250 × 10 mm) with water (0.1% formic acid) and acetonitrile (0.1% formic acid) as the mobile phase, at a flow rate of 8 mL $min^{-1}$. After 3 min, acetonitrile was increased in a linear manner (curve 5) from 5% to 65% at 22 min, followed by a wash of 100% acetonitrile. Potensimicin eluted at 15 min and was isolated at a yield of ~9 mg $L^{-1}$.

***Thanamycin.*** For production of thanamycin, *P. fluorescens* was grown in 400 mL of YM media (3 g $L^{-1}$ malt extract, 3 g $L^{-1}$ yeast extract, 5 g $L^{-1}$ peptone, 10 g $L^{-1}$ glucose) per 2.8 L Fernbach flask for 72 h at 300 rpm and 28°C until 120 L of culture was accumulated. Cultures were harvested by centrifugation at 7000 rpm, followed by methanol extraction of the cell pellet, and Diaion HP-20 (20 g $L^{-1}$) extraction of the supernatant. Methanol eluent of the HP-20 resin was pooled with the methanol extract of the cell pellet and dried under rotary vacuum. This extract was dissolved and extracted with a 1:1 mixture of butanol and water. The butanol fraction was isolated, evaporated to dryness, resuspended in a minimal volume of methanol, and left to precipitate for 48 h at 4°C. The precipitant was collected by centrifugation at 4000 rpm and 4°C, washed once with methanol, and then dissolved in excess DMSO. Thanamycin was isolated using preparative scale LC-MS using a Luna 5 μm $C_{18}$ column (*Phenomenex*, 250 × 15 mm) with water (0.1% trifluoroacetic acid) and acetonitrile (0.1% trifluoroacetic acid) as the mobile phase, at a flow rate of 10 mL $min^{-1}$.

After 2 min, acetonitrile was gradually (curve 7) increased from 5% to 42% at 28 min, then ramping (curve 5) to 51% at 40 min, followed by a wash with 100% acetonitrile. Thanamycin eluted at 35 min.

### 2.6.4 Structure Elucidation

For methods, tables, figures, and spectra pertaining to the NMR and HRMS structure elucidation of isolated natural products, see **Supplementary Note 1** (in Appendix 1 of this thesis).

### 2.6.5 Incorporation of $^{13}C_4$-threonine and $^{13}C_5$-ornithine

A fresh colony of *P. fluorescens* was used to inoculate 40 mL of YM media containing 2 mM $^{13}C_4$-L-threonine or $^{13}C_5$-L-ornithine (*Cambridge Isotope Laboratories*) which had been added through sterile syringe filtering after autoclaving. Cultures were grown for 48 h before being harvested through centrifugation and Diaion HP-20 resin extraction of the supernatant. The SMILES structure of the top GNP thanamycin hit from the first round of GNP was modified with heavy atoms to query threonine incorporation at position 4 (two hypothetical structures) and ornithine incorporation at position 2 (two hypothetical structures). GNP settings identical to those for initial thanamycin detection were used to analyze stable $^{13}C$ incorporation experiment LC-MS/MS data.

### 2.6.6 Determination of Antimicrobial Activity

Minimum inhibitory concentrations for potensimicin and thanamycin were determined using broth microdilution. For potensimicin, activity was determined in cation-adjusted Mueller Hinton broth using *Bacillus subtilis* 168 and *Staphylococcus aureus* Newman as indicator organisms, grown at 28°C and 37°C respectively. Erythromycin was used as an internal control. For thanamycin, activity was determined in YPD media using *Saccharomyces cerevisiae* as an indicator organism, growing at 28°C, and using syringomycin E as an internal control. The MIC was determined as the lowest concentration of drug at which no growth was observed after 16 h.

### 2.6.7 Genome Sequencing

A single colony of *V. paradoxus* P4B was grown in 3 mL tryptic soy broth overnight at 30°C and 250 rpm. Genomic DNA was harvested using a GenElute Bacterial Genomic DNA Kit (Sigma). A single colony of *S. calvus* was used to inoculate a 50 mL culture of GYM media containing 0.5% glycine and grown for 96 h at 30°C and 250 rpm. 500 µL of culture was centrifuged at 12000 g for 5 min and resuspended in 500 µL SET buffer (75 mM NaCl, 25 mM EDTA pH 8.0, 20 mM Tris HCl pH 7.5, 2 mg $mL^{-1}$ lysozyme) to lyse for 2 h at 37°C. Proteinase K and SDS were added after lysis to final concentrations of 0.5 mg $mL^{-1}$ and 1%, respectively. The lysis mixture was incubated at 55°C for 2 h before adjusting the concentration of NaCl to 1.25 M and extracting twice with phenol-chloroform. Isopropanol was added (equivalent to 60% the volume of the solution) to precipitate genomic DNA, which was subsequently washed twice with 70% ethanol and resuspended in sterile water for sequencing. Genomic DNA was sent for library preparation

and Illumina sequencing at the Farncombe Metagenomics Facility at McMaster University, using an Illumina HiSeq DNA sequencer. Contigs were assembled using the ABySS genome assembly program[59] and with Geneious bioinformatic software.

### 2.6.8 Bioinformatic Methodology and Construction of GNP

#### File input and options

GNP has three input modes: genome search, scaffold library generation, and LC-MS/MS database-dependent search. Respectively, these modes allow the user to search a genome for natural product biosynthesis clusters, generate a combinatorialized library of predicted scaffold molecules, and identify compounds within a LC-MS/MS chromatogram. These three steps are intended to be performed in series to allow the user to isolate natural products based on genomic information.

GNP's genome scan mode searches a sequence file for natural product biosynthesis clusters, and uses its analysis of these clusters to predict nonribosomal peptide and polyketide molecules. The genome search mode can accept whole genomes, condensed DNA sequences (representing either individual clusters or sequence contigs), or translated protein sequences. All sequences must be submitted in FASTA format. In addition to allowing the user to submit a sequence and choose an operating mode, the genome search's web interface allows the user to adjust the maximum length between genes to consider them part of the same biosynthetic cluster, and to adjust the cutoffs for what is considered a statistically significant score for four classes of domains: adenylation, acyltransferase, thiolation or thioesterase, and all other results.

**Genome search**

In whole genome searches, Glimmer 3[60] is used to locate putative biosynthetic genes. Glimmer first identifies long, non-overlapping open reading frames within the genome, then uses these sequences to construct an interpolated context model of coding sequences. This model is then used to analyze the genome and predict biosynthetic sequences. For shorter DNA sequences (clusters or contigs), custom code was implemented to detect all possible open reading frames. Finally, translated protein sequences may be submitted as a multi-FASTA file; in the absence of any location information, all sequences will be considered part of the same biosynthetic cluster.

Once putative biosynthetic genes have been identified and translated to protein sequences, they are analyzed with a library of hidden Markov models created from multiple sequence alignments of experimentally characterized enzymes relevant to nonribosomal peptide or polyketide biosynthesis. HMMER[61] is used to perform hidden Markov model searches of the biosynthetic gene protein sequence database. Translated protein sequences are searched with adenylation, condensation, and thiolation domain models obtained from PFAM[62]; dehydratase, enoylreductase, ketoreductase, acyltransferase, and thioesterase domain models obtained from SMART[63]; and a N-methyltransferase domain model obtained from Ansari *et al.*[64] Putative acyltransferase and adenylation domains are reanalyzed using a library of profile hidden Markov models compiled by Khayatt *et al.*[13] to determine their substrates.

Next, the annotated genes are grouped into putative biosynthetic clusters using a simple greedy algorithm with an adjustable window (set by default, to 10,000 base pairs).

Putative clusters are discarded if they do not contain at least one adenylation or acyltransferase domain, and at least one condensation or ketosynthase domain. Within each cluster, biosynthetic modules are next defined. For adenylation modules, all domains between condensation and thiolation domains are considered a module if the intervening region contains an adenylation domain. For acyltransferase modules, all domains between ketosynthase and thiolation domains are considered a module if the intervening region contains an acyltransferase domain. A natural product scaffold is subsequently generated by considering the top scoring substrate for each biosynthetic module and assumed co-linearity. Finally, the chemical reactions of polyketide reductive loops and N-methyltransferase domains are simulated on the scaffold to generate a predicted molecule for each cluster. The predicted natural product is output as a SMILES string, and then converted to a MDL Molfile to be loaded into the scaffold combinatorialization mode.

**Scaffold library generation**

The scaffold library generation web interface allows the user to enumerate combinatorial libraries based on chemical input. The scaffold input mode allows the user to submit a scaffold molecule or database of molecules for combinatorialization directly. Alternatively, the scaffold library generation interface will be automatically loaded upon completion of a genome search, with each putative biosynthetic cluster's predicted natural product considered a scaffold. The interface allows users to define sites of variability and input R groups, and select at which sites their R groups may be combinatorialized. SmiLib[25] is used to enumerate the combinatorial library based on user-submitted input.

A complete guide to the scaffold library generator user interface is available at magarveylab.ca/gnp/#!/help.

**LC-MS/MS Database-dependent search**

The LC-MS/MS database-dependent search interface implements considerably updated software and options based on those originally reported in Ibrahim *et al*.[26] GNP includes several new cleavage options, including ester cleavage for depsipeptides, and inverse ester and thioether cleavages for abnormal MS/MS cleavage patterns[65]. GNP also has the ability to generate images of each matched fragment or structure to facilitate matched fragment verification.

The prediction guided discovery chart is generated by identifying all scans where GNP identifies a prediction library compound as the top ranked hit. This top ranked P1 value is then divided by the average P1 values of the dummy library (in-house NRP library of all other compounds) for that scan. Since eluting compounds will have multiple scans associated with them in a chromatogram, to identify true compounds, normalized P1 values are summed in 0.25 min buckets and plotted on the prediction guided discovery chart.

The GNP platform is an Apache Tomcat web application written in Java, and relies on sequence conversion by BioJava[66] and chemical abstractions developed by the CDK[67].

**Identification of sugar biosynthesis enzymes**

In order to connect genomic sequence information to natural product glycosylation patterns, we first developed a library of hidden Markov models corresponding to sugar

biosynthesis genes. Using the biosynthetic codes outlined by Kersten et al.[22] as a guide, protein sequences representing 20 families of TDP-sugar biosynthesis were manually collected based on homology to experimentally annotated sequences (**Supplementary Table 9-10**). These sequences were aligned using Clustal Omega (version 1.2.0), and hidden Markov models were generated from the resulting alignments using the hmmbuild program (version 3.1b1), part of the HMMER software package[61]. Appropriate bitscore cutoffs were determined for each hidden Markov model by manual analysis of the results of a hidden Markov model search against the UniProtKB database, using the HMMER web server[61]. The resulting library of hidden Markov models, and the sequences used to construct them, are available at http://magarveylab.ca/sugars/.

The biosynthetic pathways for natural product pentose sugars, including the deoxyaminosugars of calicheamicin and AT2433, and the madurose moiety of maduropeptin, lack a 4,6-dehydratase. Instead, a UDP-glucose dehydrogenase catalyzes the four-electron oxidation of glucose to glucuronic acid, followed by oxidative decarboxylation by a UDP-glucose decarboxylase[35-36]. In order to predict pentose sugar moieties from genomic DNA, hidden Markov models specific to these UDP-sugar pathways were additionally constructed.

Many natural products, including vancomycin, avilamycin, and BE-7585A, contain both hexose sugars and their deoxygenated derivatives. However, hexose sugars are primary metabolites, and consequently genes associated with hexose biosynthesis are scattered throughout microbial genomes. In order to develop a strategy to identify hexose sugars from genomic sequence information, we performed a phylogenetic analysis of

glycosyltransferase domains (**Supplementary Figure 17**). A database of glycosyltransferase domain sequences was manually curated, and each sequence was associated with a sugar substrate. The resulting 82 natural product glycosyltransferase sequences were aligned in MUSCLE (version 3.8.31) and manually refined and masked in Mesquite (version 2.75) at the amino acid level. A phylogenetic tree was created in RAxML (version 7.4.2), performing 100 bootstraps with a gamma distribution. Maximum likelihood analyses were based on the LG substitution model with empirical base frequencies. The resulting tree was rendered in Figtree (version 1.4.0). Phylogenetic analysis revealed that natural product glycosyltransferases specific to hexose sugars assort into distinct clades in a phylogenetic tree. The glycopeptide mannosyltransferases represent two distinct families of hexose glycosyltransferases, while BE-7585A type glucosyltransferases represent a third. One mixed clade, consisting of both hexose and deoxysugar glycosyltransferases, was observed. These are glycopeptide glycosyltransferases, which are position-specific but promiscuous in substrate selection[68].

These results supported the use of glycosyltransferase domain sequence homology to infer the presence of glucose, mannose, gulose, or N-acetylglucosamine within a natural product structures. A hidden Markov model was constructed from the glycosyltransferase sequence database using the methods described above. The unaligned sequences were also used to compile a BLAST database. Within GNP, glycosyltransferase domains identified using the hidden Markov model are queried against the BLAST database, and the highest-scoring result is used to determine whether the substrate of a given glycosyltransferase is a hexose sugar.

**Prediction of glycosylated natural products scaffolds**

In order to connect identified sugar biosynthesis genes to deoxy-sugar monomers, it was necessary to catalogue the biosynthetic pathways of deoxy-sugar biosynthesis. The biosynthetic logic developed Kersten et al.[22] was expanded to include UDP-sugar pathway genes, and modified in cases where an insufficient number of sequences were available to build a hidden Markov model. The structures of each of the resulting 67 sugars were compiled as a set of SMILES strings. The amended biosynthetic code, consisting of 67 deoxy sugars, their structures, and the genes associated with their biosynthesis, is presented in **Supplementary Table 10**.

Individual sugar biosynthesis genes are common to diverse biosynthetic pathways, and multiple pathways are often present in a cluster. Moreover, separate biosynthetic pathways may not be physically segregated within a cluster. It was therefore necessary to implement new algorithmic logic in order to predict multiple deoxy-sugars in a single biosynthetic gene cluster. Within GNP, the number of deoxy-sugars present in the natural product scaffold is calculated by subtracting the number of glycosyltransferases with hexose substrates from the total number of glycosyltransferases in the cluster. All combinations of the 67 sugars of this size are then computed. For each combination of sugars, both the number of genes present in the cluster but absent from the sugar biosynthesis pathways, and the number of genes present in the sugar biosynthesis pathways but absent from the cluster, are simultaneously minimized. The set of sugars which minimize both sums is retained and presented to the user as the genomically identified combinatorial sugar set. Hexose sugars identified based on glycosyltransferase homology

are then added to this set. The predicted sugars are then added to a random hydroxyl group within the predicted natural product scaffold subsequent to user-directed structure combinatorialization. The sugar prediction algorithm is represented by the following pseudocode:

```
n = number of glycosyltransferases in cluster
for glycosyltransferase in cluster glycosyltransferases
        if glycosyltransferase substrate == hexose
                n--

minimum remaining in pathway = 999;
minimum remaining in cluster = 999;

sugar combinations = get all combinations of sugars of size n

for sugar combination in sugar combinations
        remainingInCluster = number of cluster sugar genes
        remainingInPathway = number of pathway sugar genes

        for pathway sugar gene in pathway sugar genes
                if cluster genes contain pathway sugar gene
                        remainingInPathway--

        for cluster sugar gene in cluster sugar genes
                if pathway genes contain cluster sugar genes
                        remainingInPathway--

        if remaining in pathway <= minimum remaining in pathway
                        and remaining in cluster <= minimum remaining in cluster
                if remaining in pathway < minimum remaining in pathway
                                or remaining in cluster < minimum remaining in cluster
                        clear combinatorial sugar set
                minimum remaining in pathway = remaining in pathway
                minimum remaining in cluster = remaining in cluster
                add sugar combination to combinatorial sugar set

return combinatorial sugar set
```

### 2.6.9 GNP analysis of *S. calvus*

Mining of in-house sequenced *Streptomyces* genomes revealed a novel nonribosomal peptide biosynthetic gene cluster in *Streptomyces calvus* (ATCC No. 13382). The corresponding gene cluster was uploaded to the GNP server to generate an automated structure prediction. The presence of two unusual trans-acting adenylation domains led to position swapping for Asn6-Thr7 (and vice versa). Additionally, the presence of an initiating condensation domain indicated acylation of the N-terminal amine, indicating a cyclic lipodepsipeptide or linear lipopeptide structure. For combinatorialization, the initiating amine was modified with a $C_8$, $C_{10}$, or $C_{12}$ fatty acid. Serine and threonine, valine and leucine / isoleucine, and phenylalanine / tyrosine were considered to be interchangeable. The four combinatorialized scaffolds yielded a library of 768 hypothetical structures. LC-MS/MS data was probed with GNP using an 18 Da precursor window and minimal P-score cut-offs (P1 = 10; P2 = 10), with 1-2 amide cleavages, 0-1 ester cleavages, and 0-1 water losses. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds related to the novel biosynthetic gene cluster.

### 2.6.10 GNP analysis of *A. citrulli* AAC00-1

Bioinformatic analysis of publicly available genome sequences revealed a number of orphan biosynthetic gene clusters, including a novel nonribosomal peptide biosynthetic gene cluster in *Acidovorax citrulli* AAC00-1 (DSM No. 17060). To generate an automated structure prediction, the novel gene cluster was uploaded to the GNP server. Extensive similarity observed between the novel biosynthetic gene cluster and that of the delftibactin biosynthetic gene cluster allowed for the correction of monomer incorporation at position

2, as ornithine rather than serine. In addition, the presence of an aspartate β-hydroxylating enzyme led to the conclusion that the predicted position 6 aspartate was β-hydroxyaspartate. This structure was combinatorialized to generate the linear and peptide macrocycle forms, with serines and threonines being interchangeable, the polyketide portion incorporating either a malonate or methylmalonate, and ornithines with side chains either as free amines, *N*-hydroxylated, *N*-formylated, *N*-acetylated, or any combination thereof, yielding a library of 576 hypothetical structures. LC-MS/MS data was probed with GNP using a 50 Da precursor window and no P-score cut-offs (P1 = 0; P2 = 0), with 1-2 amide cleavages, 0-1 ammonia losses, and 0-1 water losses. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds related to the novel biosynthetic gene cluster.

### 2.6.11 GNP analysis of *V. paradoxus* S110

Bioinformatic analysis of the publicly available genome of *Variovorax paradoxus* S110 revealed an orphan biosynthetic gene cluster with considerable homology to that previously explored in *Acidovorax citrulli* AAC00-1. To test whether improvements in the structure-prediction abilities of GNP would improve detection of related scaffolds in a novel background, the *V. paradoxus* biosynthetic gene cluster was uploaded to the GNP server to generate an automated structure prediction. The scaffold was combinatorialized similarly to the scaffold of acidobactin, albeit with the new FAAL substrate prediction. This structure library included the linear and depsipeptide macrocycle forms, with serines and threonines being interchangeable, the polyketide portion incorporating either a malonate or

methylmalonate, and ornithines with side chains either as free amines, *N*-hydroxylated, *N*-formylated, *N*-acetylated, or any combination thereof, yielding a library of 576 hypothetical structures. Due to increased confidence in the prediction and possibility of the novel ester bond, LC-MS/MS data was probed with GNP using a 50 Da precursor window and standard P-score cut-offs (P1 = 20; P2 = 20), with 1-2 amide cleavages, 0-1 ester cleavages, and 0-1 water losses. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds related to the novel biosynthetic gene cluster.

### 2.6.12 GNP analysis of *V. paradoxus* P4B

Mining of in-house sequenced genomes of lytic environmental organisms revealed a novel nonribosomal peptide biosynthetic gene cluster in *Variovorax paradoxus* P4B. The corresponding gene cluster was uploaded to the GNP server to generate an automated structure prediction. The prediction was modified as the presence of an aspartate β-hydroxylating enzyme led to the conclusion that the predicted position 3 aspartate was β-hydroxyaspartate. This structure was combinatorialized to generate the linear lipopeptide and lipodepsipeptide macrocycle forms, with serines and threonines being interchangeable, the polyketide portion incorporating either a malonate or methylmalonate, and predicted *N*-hydroxyornithines being *N*-formylated, *N*-acetylated, or not further modified, yielding a concentrated library of 32 hypothetical structures. LC-MS/MS data was probed with GNP using an 18 Da precursor window and minimal P-score cut-offs (P1 = 10; P2 = 10), with 1-2 amide cleavages, 0-1 ester cleavages, and 0-1 water losses. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds

related to the novel biosynthetic gene cluster.

### 2.6.13 GNP analysis of *N. potens*

Analysis of the *Nocardiopsis potens* (DSM No. 45234) genome revealed a gene cluster for a glycosylated polyketide. The corresponding gene cluster was uploaded to the GNP server to generate an automated structure prediction. GNP indicated the presence of two deoxy sugar gene clusters, predicted to encode machinery for producing D-angolosamine and either D-mycaminose or D-ravidosamine (which are diastereomers). For combinatorialization, the polyketide backbone was either macrocyclized on the distal hydroxyl, or left linear. The linear polyketide was glycosylated with both angolosamine and mycaminose / ravidosamine, either sugar alone, or neither. The lone free hydroxyl of the predicted macrolide was glycosylated with angolosamine or mycaminose / ravidosamine, or left bare. All methylmalonates were also made interchangeable with malonate, resulting in a library of 42 hypothetical structures. Given the comprehensive coverage of hypothetical polyketides, LC-MS/MS data was probed with GNP using a 1 Da precursor window and a reduced P1 cut-off to account for the minimal fragmentation observed with polyketides (P1 = 5; P2 = 20), with 0-1 ester cleavages, 0-2 sugar cleavages, and 0-3 water losses to account for all plausibly predictable fragmentation. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds related to the novel biosynthetic gene cluster.

### 2.6.14 GNP analysis of *P. fluorescens*

Bioinformatic analysis of the in-house sequenced *Pseudomonas fluorescens* (DSM No. 11579) genome revealed a nonribosomal peptide biosynthetic gene cluster identical (~90% sequence identity) to one found in *Pseudomonas sp*. SH-C52[18, 43], responsible to producing the cryptic natural product thanamycin. Automated structure predictions of both compounds yielded identical predictions, albeit with a trans-acting adenylation domain responsible for incorporation of 4-chlorothreonine (inferred by the presence of the conserved halogenase *thaC2*) working either at the N- or C-terminal condensation domain, depending on the position of the free-standing adenylation domain relative to the remaining genes for the NRPS assembly line. This structure prediction was modified based on known similarity to the syringomycin type molecules, including incorporation of the chlorinated threonine, hydroxylation of the predicted aspartic acid, incorporation of the modified threonine (dehydrobutyrate) at position 7, modification with a 3-hydroxy fatty acid through the *N*-terminal acylating condensation domain, and macrocyclization on the hydroxyl of the conserved and predicted serine. This modified prediction was combinatorialized based on substitutions with similar monomers as well as similar monomers from corresponding positions in syringomycin-type compounds (syringotoxin, syringostatin, syringomycin, pseudomycin, and cormycin). Diaminobutyrate, ornithine, or arginine were incorporated at position 2, aspartate or asparagine were used at position 3, serine or threonine / homoserine were used at position 4, and fatty acids used ranged from C12-16 with one or two hydroxyl groups, as seen on syringostatins, cormycin, and the pseudomycins. The library of 120 hypothetical thanamycin structures was used to probe LC-MS/MS data with GNP using an 18 Da precursor window and reduced P-score cut-offs (P1 = 15; P2 = 15), with 1-2 amide

cleavages, 0-1 ester cleavages, and 0-1 water losses. The prediction guided discovery chart was overlaid with the base-peak chromatogram to indicate compounds originating from the thanamycin biosynthetic gene cluster. Identical parameters were used when scoring isotope-labelled thanamycin extracts.

### 2.6.15 Generation of chemoinformatic tree of thanamycin structures

To generate a chemoinformatic tree of predicted thanamycin structures, the combinatorialized library was assembled in SMILES format in a text file and submitted to the online interface of ChemMine Tools (http://chemmine.ucr.edu/tools/launch_job/Clustering/) for hierarchical clustering analysis. The resulting newick tree was rendered in Dendroscope and subsequently in Adobe Illustrator CS6.

### 2.6.16 Identification of natural product standards

Natural product 1000 × stock solutions were prepared by dissolving 1 mg of pure standard in 1 mL of water, methanol, or dimethyl sulfoxide. 10 μL of 1000 × stock was diluted into 1 mL of water to make 1 μg $mL^{-1}$ working concentration solutions. 10 μL of each stock was injected for LC-MS/MS analysis using standard conditions, including a 10:1 flow splitter for MS acquisition. Samples were routinely visible under these conditions, except for capreomycin, which was only visible at 10 μg $mL^{-1}$ working concentration. LC-MS/MS data files were converted to .mzXML files and analyzed with GNP. Minimal P-score cut-offs of 5 (P1) and 5 (P2) were implemented to account for minimal fragmentation and signal

from the low abundance standards. Fragmentation settings were as follows: daptomycin and thiostrepton - 1-2 amide cleavage, 0-1 ester cleavage, 0-1 water loss; lincomycin - 0-1 amide cleavage, 0-3 water loss; novobiocin - 0-1 amide cleavage, 0-1 ester cleavage, 0-1 sugar cleavage, 0-2 water loss; nystatin - 0-1 ester cleavage, 0-1 sugar cleavage, 0-3 water loss; erythromycin - 0-1 ester cleavage, 0-2 sugar cleavage, 0-3 water loss; capreomycin, bacitracin, gramicidin, polymyxin - 1-2 amide cleavage, 0-1 water loss; vancomycin - 0-1 amide cleavage, 0-2 sugar cleavage, 0-1 water loss; valinomycin - 0-2 amide cleavage, 0-2 ester cleavage, 0-1 water loss.

## 2.7 Supplementary Information

Supplementary information from the publication referred to in this chapter can be found in Appendix 1.

## 2.8 References

1. Newman, D.J. & Cragg, G.M. Natural products as sources of new drugs over the last 25 years. *J. Nat Prod.* **70**, 461-477 (2007)
2. Dayan, F.E., Cantrell, C.L. & Duke, S.O. Natural products in crop protection. *Bioorg. Med. Chem.* **17**, 4022-4034 (2009)
3. Felnagle, E.A. *et al*. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol. Pharm.* **5**, 191-211 (2008)

4. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48**, 4688-4716 (2009)
5. Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006)
6. Bentley, S.D. *et al*. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141-147 (2002)
7. Xu, Y. *et al*. Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *J. Am. Chem. Soc.* **134**, 8625-8632 (2012)
8. Freeman, M.F. *et al*. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338**, 387-390 (2012)
9. Li, J.W. & Vederas, J.C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161-165 (2009)
10. Röttig, M. *et al*. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic. Acids. Res.* **39**, W362-W367 (2011)
11. Medema, M.H. *et al*. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic. Acids. Res.* **39**, W339-W346 (2011)
12. Li, M.H.T. *et al*. Automated genome mining for natural products. BMC Bioinformatics **10**, (2009)

13. Khayatt, B.I. *et al*. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PloS one* **8**, e62136 (2013)
14. Prieto, C. *et al*. NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427 (2012)
15. Starcevic, A. *et al*. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic. Acids. Res.* **36**, 6882-6892 (2008)
16. Duncan, K.R. *et al*. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chem. Biol.* **22**, 460-471 (2015)
17. Nguyen, D.D. *et al*. MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. USA.* **110**, E2611-E2620 (2013)
18. Watrous, J. *et al*. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. USA.* **109**, E1743-1752 (2012)
19. Yang, J.Y. *et al*. Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686-1699 (2013)
20. Medema, M.H. *et al*. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* 10, e1003822 (2014)
21. Kersten, R.D. *et al*. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794-802 (2011)

22. Kersten, R.D. *et al*. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc. Natl. Acad. Sci. USA.* **110**, E4407-E4416 (2013)

23. Cimermancic, P. *et al*. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421 (2014)

24. Bienfait, B. & Ertl, P. JSME: a free molecule editor in JavaScript. *J. Cheminform.* **5**, 24 (2013)

25. Schüller, A., Hähnke, V. & Schneider, G. SmiLib v2. 0: A Java-Based Tool for Rapid Combinatorial Library Enumeration. *QSAR & Combinatorial Science* **26**, 407-410 (2007)

26. Ibrahim, A. *et al*. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. USA* **109**, 19196-19201 (2012)

27. Rausch, C. *et al*. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7**, 78 (2007)

28. Yu, Z. *et al*. New WS9326A congeners from *Streptomyces* sp. 9078 inhibiting Brugia malayi asparaginyl-tRNA synthetase. *Org. Lett.* **14**, 4946-4949 (2012)

29. Donadio, S., Monciardini, P. & Sosio, M. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.* **24**, 1073-1109 (2007)

30. Johnston, C.W. *et al*. Gold biomineralization by a metallophore from a gold-associated microbe. *Nat. Chem. Biol.* **9**, 241-243 (2013)

31. Trivedi, O.A. *et al*. Enzymic activation and transfer of fatty acids as acyl-adenylates in mycobacteria. *Nature* **428**, 441-445 (2004)

32. Wyatt, M.A. *et al*. Bioinformatic evaluation of the secondary metabolism of antistaphylococcal environmental bacterial isolates. *Can. J. Microbiol.* **59**, 465-471 (2013)

33. Walsh, C., Freel-Meyers, C.L. & Losey, H.C. Antibiotic glycosyltransferases: antibiotic maturation and prospects for reprogramming. *J. Med. Chem.* **46**, 3425-3436. (2003)

34. Gates, P.J. *et al*. Structural elucidation studies of erythromycins by electrospray tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **13**, 242-246 (1999)

35. Thibodeaux, C.J., Melançon, C.E. 3rd & Liu, H.W. Natural-product sugar biosynthesis and enzymatic glycodiversification. *Angew. Chem. Int. Ed. Engl.* **47**, 9814-9859 (2008)

36. Gao, Q. *et al*. Deciphering indolocarbazole and enediyne aminodideoxypentose biosynthesis through comparative genomics: insights from the AT2433 biosynthetic locus. *Chem. Biol*. **13**, 733-743 (2006)

37. Vara, J. *et al*. Cloning of genes governing the deoxysugar portion of the erythromycin biosynthesis pathway in *Saccharopolyspora erythraea* (*Streptomyces erythreus*). *J. Bacteriol.* 171, 5872-5881 (1989)

38. Ward, S.L. *et al*. Chalcomycin biosynthesis gene cluster from *Streptomyces bikiniensis*: novel features of an unusual ketolide produced through expression of the chm polyketide synthase in *Streptomyces fradiae*. *Antimicrob. Agents Chemother*. **48**, 4703-4712 (2004)

39. Xue, Y. *et al.* A gene cluster for macrolide antibiotic biosynthesis in *Streptomyces venezuelae*: architecture of metabolic diversity. *Proc. Natl. Acad. Sci. USA.* **95**, 12111-12116 (1998)

40. Bubb, W.A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. *Concept Magnetic Res.* **19**, 1-19 (2003)

41. Maezawa, I., Kinumaki, A. & Suzuki, M. Biological glycosidation of macrolide aglycones. I. Isolation and characterization of 5-O-mycaminosyl narbonolide and 9-dihydro-5-O-mycaminosyl narbonolide. *J. Antibiot*. **29**, 1203-1208 (1976)

42. Rengaraju *et al*. A new macrolide antibiotic kayamycin 10,11-dihydro-5-0-mycaminosyl narbonolide produced by Nocardiopsis. Meiji Seika Kenkyu Nenpo 24, 52-54 (1985)

43. Mendes, R. *et al*. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097-1100 (2011)

44. Behnken, S. & Hertweck, C. Cryptic polyketide synthase genes in non-pathogenic Clostridium SPP. *PloS One* **7**, e29609 (2012)

45. Cociancich, S. *et al*. The gyrase inhibitor albicidin consists of p-aminobenzoic acids and cyanoalanine. *Nat. Chem. Biol.* **11**, 195-197 (2015)

46. Johnston, C.W. *et al*. Nonribosomal assembly of natural lipocyclocarbamate lipoprotein-associated phospholipase inhibitors. *ChemBioChem* **14**, 431-435 (2013)

47. Pedras, M.S. & Biesenthal, C.J. Isolation, structure determination, and phytotoxicity of unusual dioxopiperazines from the phytopathogenic fungus Phoma lingam. *Phytochemistry* **58**, 905-909 (2001)

48. Healy, F.G. *et al*. Involvement of a cytochrome P450 monooxygenase in thaxtomin A biosynthesis by Streptomyces acidiscabies. *J. Bacteriol.* **184**, 2019-2029 (2002)

49. Pohle, S. *et al.* Biosynthetic gene cluster of the non-ribosomally synthesized cyclodepsipeptide skyllamycin: deciphering unprecedented ways of unusual hydroxylation reactions. *J. Am. Chem. Soc.* **133**, 6194-6205 (2011)

50. Clardy, J. & Walsh, C.T. Lessons from natural molecules. *Nature* **432**, 829-837 (2004)

51. Nett, M., Ikeda, H. & Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384 (2009)

52. Walsh, C.T. & Fischbach, M.A. Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.* **132**, 2469-2493 (2010)

53. Ng, J. *et al*. Dereplication and de novo sequencing of nonribosomal peptides. *Nat. Methods.* **6**, 596-599 (2009)

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

55. Mohimani, H. *et al*. Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545-1551 (2014)

56. Mohimani, H. *et al*. NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. *J. Nat. Prod.* **77**, 1902-1909 (2014)

57. Frank, A.M. *et al*. Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113-122 (2008)

58. Zhang, Q. *et al*. Structural investigation of ribosomally synthesized natural products by hypothetical structure enumeration and evaluation using tandem MS. *Proc. Natl. Acad. Sci. USA.* **111**, 12031-12036 (2014)

59. Simpson, J.T. *et al*. ABySS: a parallel assembler for short read sequence data. *Genome Res*. **19**, 1117-1123 (2009).

60. Delcher, A. L. *et al*. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673-679 (2007).

61. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

62. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res*. **42**, D222-230 (2014).

63. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res*. **40**, D302-305 (2012).

64. Ansari, M.Z. *et al*. In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. *BMC Bioinformatics*, **9**, 454 (2008).

65. Skinnider, M., Johnston, C. W., Zvanych, R., Magarvey, N. A. Automated identification of depsipeptide natural products by an informatic search algorithm. *ChemBioChem*, (2014).

66. Prlić, A. *et al*. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693-2695 (2012).

67. Steinbeck, C. *et al*. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**, 493-500 (2003)

68. Li, T.L. *et al*. Biosynthetic gene cluster of the glycopeptide antibiotic teicoplanin: characterization of two glycosyltransferases and the key acyltransferase. *Chem. Biol.* **11**, 107-119 (2004)

# Chapter 3. Assembly and Clustering of Natural Antibiotics Guides Target Identification.

## 3.1 Chapter preface

Antibiotic resistance is well recognized as one of the most serious threats to human health in coming years, prompting calls for the discovery and development of new antibiotics with new mechanisms of action. Most antibiotics we use today were discovered during the 'Golden Age' of natural products discovery, roughly encompassing the 1950's-1970's, wherein massive quantities of antibiotic microbial natural products were discovered by academic and industrial researchers. Despite their inherent value, natural products discovered during this time were often poorly documented, and many valuable leads were discarded in favour of highly attractive scaffolds such as macrolides, glycopeptides, or β-lactams. As many of these molecules were discarded for reasons that are not relevant in the current environment of antibiotic resistance, revisiting promising molecules may be a viable strategy for uncovering new leads for clinical development. To achieve this, we created a comprehensive database of microbial natural product antibiotics and clustered them into distinct families using a retrobiosynthetic algorithm (GRAPE). By analysing associated metadata, we could quantify that microbial natural product antibiotic scaffolds have a nearly 40% chance to affect a new target, demonstrating why microbial natural products have been such an effective source of clinically-relevant antibacterials. Using our clustering data we could identify unique antibacterial chemical scaffolds that did not possess known mechanisms of action and prioritize these for downstream resistance screening and mode-of-action studies. Analysis of the telomycin biosynthetic gene cluster

did not reveal known resistance genes, indicating that telomycin may possess a new mechanism of action, consistent with previous reports. Using suppressor-mutant genome sequencing and an extensive series of assays, we could demonstrate that telomycin targets the bacterial lipid cardiolipin, resulting in cell lysis. Using the approach demonstrated in this manuscript, we hope to continue to reveal over-looked chemical scaffolds with new mechanisms of action to address the current need for novel antibiotics that can be used to treat the increasing threat of multidrug-resistant bacteria.

The following chapter is a modified version of previously published journal article in which I was the lead author. I performed telomycin studies, assisted in antibacterial library curation and development of scoring strategies, performed Antibioticome analysis, contributed to study design, and wrote the manuscript. Michael Skinnider developed PRISM, developed the Antibioticome web application, assisted in resistance gene collection, and contributed to study design. Chris Dejong developed the retrobiosynthetic algorithm, devised scoring strategies for the Antibioticome, developed the Antibioticome web application, and contributed to study design. Philip Rees curated the antibacterial library. Gregory Chen developed the retrobiosynthetic algorithm and devised scoring strategies for the Antibioticome. Chelsea Walker collected resistance genes. Shawn French performed microscopy studies. Prof. Eric Brown edited the manuscript. Prof. János Bérdy curated the antibacterial library. Dennis Liu assisted in resistance gene collection and curation of the antibacterial library. Prof. Nathan Magarvey contributed to study design and wrote the manuscript. The citation for this publication is as follows:

**Johnston, C.W.**, Skinnider, M.A., Dejong, C.A., Rees, P.N., Chen, G.M., Walker, C.,

French, S., Brown, E.D., Bérdy, J., Liu, D.Y., & Magarvey, N.A. (2015) Charting the Antibioticome Uncovers an Antibiotic with a New Mode of Action. *Nature Chemical Biology* (Accepted).

## 3.2 Abstract

Antibiotics are essential for a number of medical procedures including the treatment of bacterial infections, but wide-spread use has led to the accumulation of resistance, prompting calls for the discovery of antibacterial agents with new targets. A majority of clinically approved antibacterial scaffolds are derived from microbial natural products, but these valuable molecules are not well annotated or organized, limiting the efficacy of modern informatic analyses. Here, we provide a comprehensive resource defining the targets, chemical origins, and families of the natural antibacterial collective via a retrobiosynthetic algorithm. From this we also detail the directed mining of biosynthetic scaffolds and resistance determinants to reveal structures with a high likelihood for new modes of action. Implementing this pipeline led to investigations of the telomycin family of natural products from *Streptomyces canus*, revealing that these bactericidal molecules possess a new antibacterial mode of action dependent on the bacterial phospholipid cardiolipin.

## 3.3 Introduction

The increasing prevalence of antibacterial resistance in the clinic has highlighted the need for new antibacterial drugs with divergent mechanisms[1, 2]. With few exceptions,

clinically relevant antibacterials are derived from microbial natural product antibiotics[3]. Although the scale and size of this natural antibacterial collective has been poorly defined to date, it is composed of valuable chemical structures that have seemingly been honed through natural selection to provide microbes with competitive advantages in their native environments[4]. Biosynthetic pathways, including modular assembly line-like enzymes, facilitate the synthesis of families of molecules that can undergo natural selection and yield privileged scaffolds[5]. Significant effort has been devoted to identifying clinically applicable antibacterial targets[6], but it is perhaps not surprising that nearly every known druggable target has a dedicated natural product antibiotic directed toward it[7–9] (**Fig. 3.1; Supplementary Results, Supplementary Table 1**). As such, new and known natural product antibacterials should provide innovation to derive the next generation of clinical antibiotics if they can be collected and charted adeptly using new chemo- and bioinformatic methods[10].

Natural products produced by modular biosynthetic gene clusters frequently possess bioactive chemical scaffolds that have been honed by natural selection, leading to families of molecules that represent variations on a conserved core directed towards a common molecular target[5]. Using new chemoinformatic techniques that can leverage our knowledge of biosynthesis, these families can be elaborated and charted in order to direct efforts towards exotic natural products with little chemical similarity to families with known modes of action. Bioinformatic analysis can also be used to direct efforts towards molecules that are unlikely to possess known targets. While many natural product biosynthesis genes are useful for inferring natural product structures, others can provide

clues on natural product function. Studies of antibacterial resistance genes that are naturally associated with biosynthetic gene clusters[11, 12] have revealed that resistance determinants typically have functions related to a specific class of small molecule, corresponding to a given antibiotic scaffold or its molecular target[13–16]. As such, resistance genes may provide a unique opportunity to reveal the target of a given antibiotic, independent of structural similarity to known molecules. Computational pipelines that could incorporate extensive accumulated knowledge of natural product structures and biosynthesis may create a unique means to identify orphaned antibacterial agents with unique mechanisms.

The golden age of antibiotic discovery (1940's–1970's) produced a massive quantity of data at a time when modern big data organization tools did not exist. Thousands of antibiotics were discovered, but only a small handful were developed and have persisted as drugs until recently. With the advent of widespread antibacterial resistance in the clinic[17], refactoring these old scaffolds is becoming more difficult and less effective, and new molecules will be needed to take their place. To capitalize on the valuable – but disparate – data uncovered from the golden age of discovery to today will require that it be collected, annotated, and processed by modern chemo- and bioinformatic methods[18]. Reconciling this massive quantity of chemical and biological data with modern techniques could provide an efficient, directed means to quickly identify chemical scaffolds with new modes of action to reinvigorate the antibiotic pipeline. Following the collection and organization of the known antibacterials, compelling orphan antibiotics can be identified and revisited, given that many of these valuable structures were discarded for reasons that are not necessarily relevant in the current climate of antibacterial resistance[19]. Revisiting promising orphaned

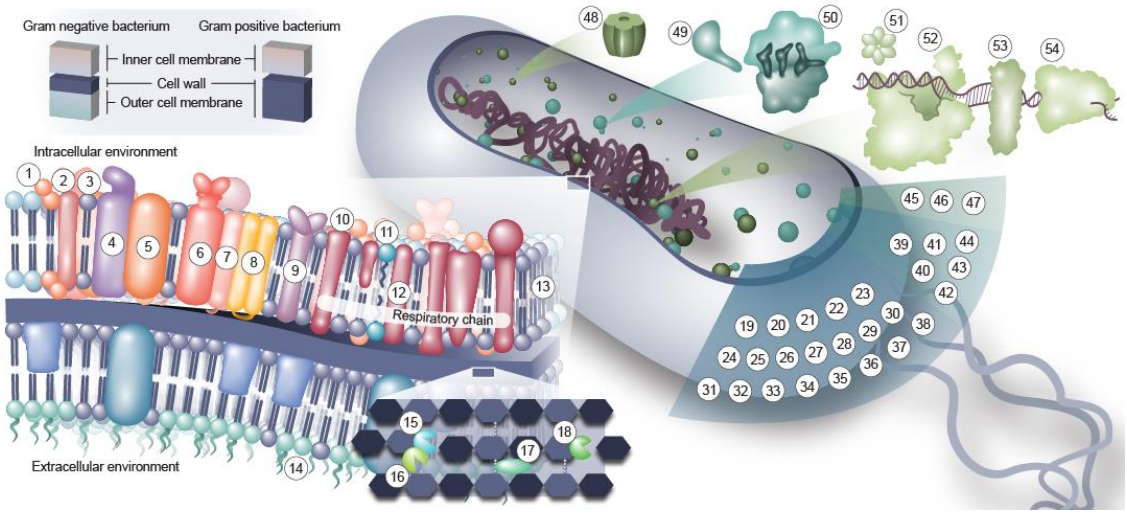



**Figure 3.1 Microbial natural products with specific antibacterial activity define a diverse range of antibacterial targets.** A comprehensive review of the antibacterial natural products identified 54 distinct molecular targets of known antibacterials, including membrane components and membrane-bound enzymes (1–14), cell wall associated enzymes (15–18), amino acid biosynthesis and metabolism (19–38), fatty acid biosynthesis (39–44), individual metabolic enzymes (45–47), and macromolecular machineries (48–54) such as the ribosome (50) or RNA polymerase (52). A full legend can be found in **Supplementary Table 1**. *Image credit: Sam Holmes.*

antibiotics at this time makes sense in light of these new clinical realities, in addition to the recent development of chemical and biosynthetic methods that provide the means to modify a wider range of scaffolds and improve clinical success[20, 21]. Orphaned antibiotics from microbes may also create new infection treatment strategies[22] and antibacterial mechanisms[7, 23–24] for development. As flagship antibiotic scaffolds succumb to resistance, new approaches are needed to find molecules with new modes of action. In this work, we

compile – for the first time – a comprehensive collection of antibacterial natural products and profile them with a retrobiosynthetic algorithm to define rare classes without known resistance mechanisms or known modes of action. This analysis led to investigation of the telomycin (**1**) family of natural products, which are shown to possess a new antibacterial target – the phospholipid cardiolipin.

## 3.4 Results

### Charting Natural Antibiotics

In contrast to growing genetic databases detailing the biosynthetic origins of natural products[25–27], there has been little progress in the development of a comprehensive annotated chemical database of extant microbial antibiotic natural products. Our systematic analysis of antibacterial natural products was initiated by incorporating an up-to-date version of the Handbook of Antibiotic Compounds[28, 29] and extensive reviews of published literature and patents into a single database. Chemical structures for each known antibiotic were generated and rendered in SMILES format (10,343 compounds). Manual annotation of the specificity of each compound identified that 7,184 molecules demonstrated non-specific antibacterial activity while 3,159 displayed specific antibacterial activity. This specificity analysis involved a comprehensive review of primary literature for each natural product described, assessing bioactivity towards bacteria as well as fungi, plants, human cell lines, and animals. Non-specific antibacterials were found from fungi (35%) as well as bacteria (66%), while specific antibacterial agents were significantly enriched in bacterial producers (96%) as opposed to fungi (4%), where specific antibacterials were most often

fusidic acids, mutilins, and penicillins. Meta-data related to the known or suspected mechanisms of action of these specific antibacterials was also incorporated into our analysis, revealing 54 described natural antibacterial targets (**Fig. 3.1, Supplementary Table 1**). Antibiotic classes that make use of modular biosynthetic pathways were segregated for analysis, including ribosomal and nonribosomal peptides, polyketides, β-lactams, and carbohydrate or aminoglycoside superfamilies. To generate high fidelity subgroupings of these molecules, we developed a retrobiosynthetic algorithm that could utilize biosynthetic and chemoinformatic logic to assign likeness, as opposed to classical chemoinformatic metrics such as Tanimoto similarity scoring. The limitations of chemoinformatic relationship scoring using Tanimoto have been reported previously, particularly with molecules that are as diverse as natural products[30, 31]. To facilitate high-fidelity clustering, our new algorithm was designed to integrate the distinct biosynthetic origins of the aforementioned modular natural product superfamilies, identifying substrates contained within these antibiotics, and the monomers used in their construction. We reasoned that this retrobiosynthetic algorithm could retrace the evolutionary histories of these selective antibacterial agents and provide a means of allocating them into defined sub-groups. This algorithm was incorporated as a publically available Java-based web application (**Supplementary Fig. 1**) that selectively deconstructs molecules from modular natural product superfamilies using a series of custom tailored rules to provide a proto-scaffold consisting of the originating building blocks. Each series of individual components are used as a fingerprint for their respective antibiotic, facilitating antibacterial fingerprint alignment via a Needleman-Wunsch algorithm (**Dejong, Chen, and Li, *submitted*; Fig.**

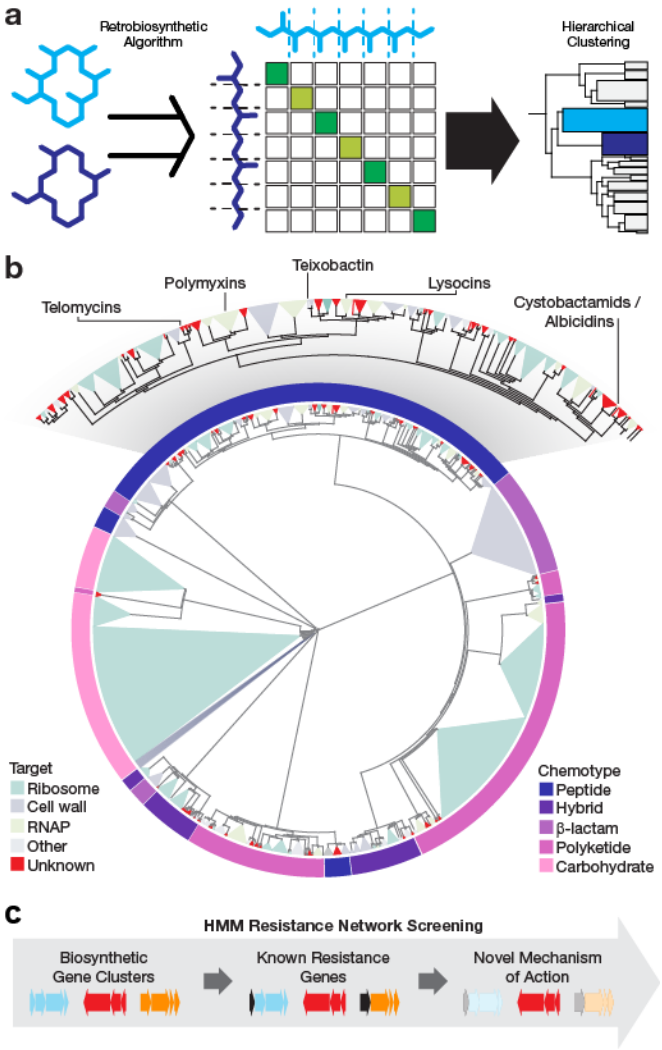



**Figure 3.2 A retrobiosynthetic strategy for charting antibacterial natural products and identifying rare scaffolds with new molecular targets.** (**a**) A retrobiosynthetic algorithm was devised to deconstruct antibiotics according to their modular biosynthetic origins and facilitate high fidelity clustering of related families. Products of modular assembly architectures are broken down with class-specific rules, providing biosynthetic monomers that can be aligned and scored using a similarity matrix that can be used for hierarchical molecular clustering. (**b**) A sampled collection of antibacteirals, including

1,908 modular natural products with specific antibacterial activity. Families are colored by known or unknown mechanism of action, and natural product chemotypes are highlighted by color (outer ring). A section of the peptide antibiotics is inset, demonstrating distinct branches for a number of antibacterial peptide subfamilies. (**c**) Schematic for PRISM-based HMM resistance network screening, outlining the automated detection of biosynthetic gene clusters and detection of known resistance determinants guiding focus towards gene clusters with potentially novel mechanisms of action.

**3.2a**). By processing this data through hierarchical clustering, we reasoned that we would have a protocol (see ***Methods***) sufficient to reveal the chemoinformatic relationships of all members of these classes and define groups of specific antibiotics with no known mode of action. Results of this analysis were readily organized into sub-families comprising 1,908 distinct molecules, with an exceptionally low rate of incorrect localization during clustering (<0.5%; **Fig. 3.2b**).

Parsing the accumulated antibiotic data revealed the distribution of molecules with specific antibacterial activity originating from modular biosynthetic assemblies. Within these groups, 668 molecules were of peptidic origin, followed in abundance by polyketides (571), carbohydrates (354), β-lactams (196), and hybrid molecules (119). The majority of the molecules in our analysis belong to subfamilies with known mechanisms of action (88.3%), spanning 32 established antibacterial mechanisms (**Fig. 3.1**), while molecules without established molecular targets are rare (11.7%; **Fig. 3.2b**). The discrepancy between all known targets and the targets identified in our analyzed subset of antibacterials largely

reflects small antimetabolites that interfere with amino acid metabolism, and which did not yield usable alignment data following deconstruction via our retrobiosynthetic algorithm. Nonribosomal peptides demonstrated the most variation in their modes of action, including 22 distinct molecular targets. Polyketides and hybrid assembly structures also demonstrated considerable diversity, with 10 and 7 antibacterial targets respectively. Collectively, these products were shown to affect 30 distinct targets. For the β-lactams (3), and carbohydrates (2) the list is much smaller, despite the fact that these groupings have a large number of individual members, the target diversity is relatively narrow. Inhibition of the ribosome was the most frequently observed antibacterial mechanism, including 45% (851 molecules) of the specific antibacterials from our modular antibacterial natural product sample, corresponding to 51% of those with known mechanisms (**Fig. 3.2b**). Inhibition of cell wall biosynthesis through various molecular targets was the second most frequently observed mechanism of the sampled antibacterials with known mechanisms (26%; 441 molecules) followed by inhibition of RNA polymerase (4.3%; 73 molecules). Analysis of higher order families (such as thiazolyl peptides, glycopeptides, or aminoglycosides) demonstrated the frequency of distinct, evolved scaffolds inhibiting a given target. The ribosome was again observed as the most frequent hit, and was the target of 19 distinct families from our sample of antibacterials with modular biosynthetic origins. RNA polymerase was also a frequent target with 9 distinct families of inhibitors, as well as DNA gyrase, which was the target of 5 distinct classes. Notably, of the 137 distinct families observed in our sample of 1,908 modular, specific antibacterials, 54 did not possess known mechanisms of action. This also indicates that the 83 molecular classes with known mechanisms in our sample covered 32

molecular targets, indicating a frequency for new scaffolds inhibiting new targets of nearly 40%. Closer inspection of well-established targets such as the ribosome or RNA polymerase shows that individual natural product scaffolds bind a wide variety of sites on these macromolecular assemblies[32–33], demonstrating that assembly line antibiotics will evolve to create chemical and biological diversity. Analysis regarding the specificity of these diverse natural antibiotics for their established targets is well characterized in primary literature and reviews (**Supplementary Dataset 1**).

This sampling of natural antibacterials demonstrates that certain structural classes are more effective at generating chemical and biological diversity than others. Peptide antibiotics were found to have the highest degree of chemical and mechanistic diversity, as well as the largest number of subfamilies without known mechanisms of action. Given the nature of their modular biosynthetic machinery and the degree to which monomers and scaffolds can be tailored, nonribosomal peptides offer the greatest number of combinations and permutations for natural selection. These peptides appear to be uniquely capable of interacting specifically with structural small molecules or membrane components, and can act as potent bactericidal agents[22, 24].

**Creating a Resistance Determinant Library**

A salient feature in the hierarchical clustering data reached through our retrobiosynthetic analysis is that subfamilies share a chemical scaffold and a common mechanism of action. With this unique analytical capacity, we could rapidly define all subfamilies from our sampling of the natural antibacterial collective, on the basis of their

chemical structures and shared biosynthetic components. To interrogate the genetic origins of promising subfamilies, we devised a bioinformatic web application for detecting and displaying biosynthetic gene clusters, PRISM[34] (**Supplementary Fig. 2**). By utilizing a series of hidden Markov models (HMMs), BLAST databases, and chemoinformatic algorithms, PRISM detects known and unknown gene clusters, provides similarity scores with known biosynthetic clusters, produces high fidelity predictions of chemical scaffolds, and identifies all known self-resistance determinants (***Methods***). Several key works have underscored that resistance is often disseminated from antibiotic producing organisms and is associated with biosynthetic gene clusters for antibacterial agents[12, 13, 16]. Classically, this collection of known resistance determinants has been used to expedite the tracking and spread of antibacterial resistance in a clinical setting. Now, by associating self-resistance determinants with known antibacterial mechanisms, PRISM can scan biosynthetic gene clusters for resistance genes that can predict antibacterial targets. By combining known cluster scoring, chemical predictions, and resistance screens, this pipeline can assist in validating antibiotics and prioritizing those with unique mechanisms that lack cross-resistance (**Fig. 3.2c**).

To facilitate optimal detection of resistance genes, we collected self-resistance determinants from all known antibiotic pathways and derived 108 HMMs that span mechanisms involved in drug efflux, drug modification, target alteration, and decoy targets (**Fig. 3.2c**; **Supplementary Dataset 2**). This database was subsequently paired with previously collected HMMs from clinically observed resistance genes to provide a comprehensive means of detecting known resistance genes[35]. To assess the use of our

pipeline for profiling self-resistance determinants that indicate mechanism, we loaded a series of biosynthetic gene cluster DNA sequences into PRISM. Biosynthetic genes were automatically detected and depicted, and known molecule identity was determined. PRISM generated predictions of the natural product structure, assessed similarity to gene clusters with known products, and identified a diverse series of scaffold- and target-specific resistance genes (**Supplementary Fig. 3–11**). Using the integrated HMMs, PRISM detected resistance determinants in each analyzed cluster that were telling of the activity for the chemical class produced. As unique structures are necessary, but not sufficient for a unique mechanism of action, this pipeline provides a crucial second measure to ensure that targeted rare scaffolds possess similarly uncommon modes of action without cross-resistance.

**Investigation of the Mode of Action of Telomycin**

Based on our defined retrobiosynthetic clustering of antibacterial natural products, we sought to investigate under-represented chemical scaffolds with the potential to possess uncharacterized molecular targets. Given their structural and mechanistic diversity, we chose to investigate nonribosomal peptides in particular. From a series of promising candidates including, among others, the pyloricidins[36] and griselimycins[37], we chose to investigate the telomycins[38–41], which possess potent bactericidal activity against a wide range of Gram positive organisms[39, 42]. The telomycin-producer, *Streptomyces canus* ATCC 12647, was purchased from the ATCC and submitted for genomic sequencing. Using PRISM, we were able to quickly identify a candidate telomycin biosynthetic gene

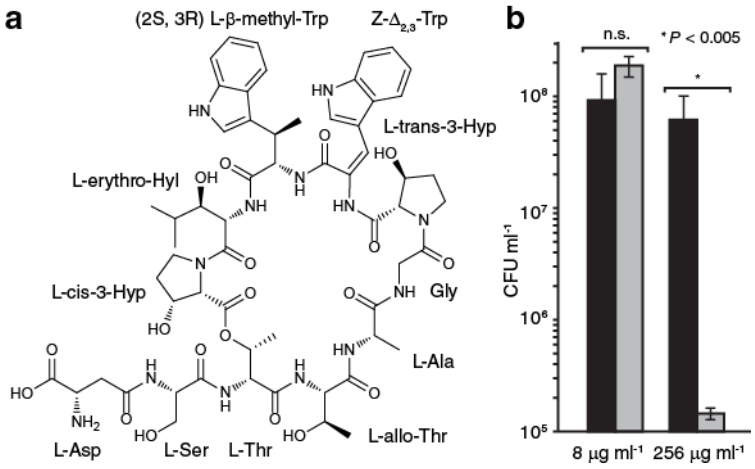



**Figure 3.3 Telomycin - a nonribosomal peptide that lyses bacteria through an unknown mechanism.** (**a**) Structure of telomycin. (**b**) Telomycin lyses *S. aureus* at high concentrations. Wild type *S. aureus* (OD 0.2) was exposed to 8 (1× MIC) or 256 μg $ml^{-1}$ (32× MIC) telomycin. Colony-forming units (CFUs) were taken before treatment (black) and after 90 min (white). Results are shown as ± s.d.; $n = 4$; Two-tailed student's *t*-test. $P = 0.0023$. Data are representative of three independent experiments.

cluster from the assembled contigs (**Fig. 3.3a**). The telomycin gene cluster was found to encode three nonribosomal peptide synthetases and a series of accessory enzymes consistent with formation of the telomycin undecapeptide core, precursor supply and tailoring (**Fig. 3.3ab**). When this biosynthetic gene cluster was passed through our resistance screening pipeline, no genes were identified with similarity to known specific self-resistance determinants (**Supplementary Fig. 12**), consistent with a unique mechanism of action. These results are consistent with previous observations that the telomycin-related molecule LL-AO341β does not demonstrate cross-resistance with other

antibacterial agents[42]. This automated assessment of potential resistance profiles supported additional efforts to be directed toward describing the mechanism of action of telomycin.

Telomycin is a lytic antibacterial agent, demonstrating rapid lysis of Gram positive bacteria including *S. aureus* and *B. subtilis* at concentrations above the minimum inhibitory concentration (MIC; **Fig. 3.3c**). Telomycin also does not display apparent cytotoxicity, as previous studies[43] with Chinese Hamster ovary (CHO) cells did not observe toxicity at concentrations up to 40 μg $mL^{-1}$, and we did not observe cytotoxicity with human kidney (HEK293) cells at concentrations up to 128 μg $mL^{-1}$. Further, telomycin has demonstrated efficacy in animal studies without apparent toxicity[40] and is not hemolytic (**Supplementary Fig. 13**). To determine the mechanism of specific bacterial lysis, *S. aureus* and *B. subtilis* strains were exposed to telomycin to generate and select for spontaneous suppressor mutants which were sequenced to reveal candidate target loci. Telomycin-resistant bacteria generated with limited passaging were refractory to lysis (**Fig. 3.4a**) and demonstrated a 16-fold increase in MIC (**Supplementary Table 3**). Comparison of spontaneously resistant mutant genomes with those of their sensitive parent strains revealed a series of inactivating mutations in cardiolipin synthase (*S. aureus cls2*; *B. subtilis clsA*; **Supplementary Fig. 14**). Cardiolipin is a phospholipid dimer formed from the condensation of two phosphatidylglycerol monomers, found in the lipid membranes of bacteria and the inner membranes of mitochondria. As telomycin-resistant mutants possessed truncating- or missense mutations near the active site of cardiolipin synthase and were refractory to telomycin-mediated lysis, cardiolipin was investigated as a potential target. Resistant mutants possessed levels of cardiolipin ranging from 2 to 10% of wild type concentrations,

with residual cardiolipin presumably produced by the remaining auxiliary cardiolipin synthase gene (*S. aureus cls1*; *B. subtilis clsB*; **Supplementary Table 3**). Cardiolipin levels are known to increase following growth in conditions of high salinity[44], and this was shown to correlate with improved telomycin activity (**Supplementary Table 4**). Having demonstrated that cardiolipin levels were strongly correlated with the antibacterial activity of telomycin, we sought to test whether telomycin specifically interacts with cardiolipin. In contrast to mixtures of telomycin with other lipids, telomycin and cardiolipin rapidly formed an opaque precipitant in a dose-dependent manner (**Fig. 3.4b**). This molecule-specific turbidity often occurs with peptide antibiotics that specifically bind membrane lipids, for instance cinnamycin binding to phosphatidylethanolamine[45], or lysocin binding bacterial menaquinone[24]. Consistent with this observation and the hypothesis that telomycin exerts lytic activity through a specific interaction with cardiolipin, excess cardiolipin is able to completely abolish the antibacterial action of telomycin, in contrast to other bacterial phospholipids (**Fig. 3.4cd**). This effect is dose-dependent (**Fig. 3.4c**) and is largely cardiolipin specific, although the monomeric cardiolipin analogue phosphatidylglycerol also demonstrates some protective effect when provided in 10× molar excess.

Cardiolipin possesses significant negative curvature relative to other phospholipids, which causes it to accumulate at poles and septa of bacterial cells[46–48]. Although this has previously been established with microscopy, the only dye available to image cardiolipin is a simple derivative of acridine orange (10-*N*-nonyl acridine orange; NAO) – which was recently found to promiscuously stain anionic lipids in general, with limited selectivity for

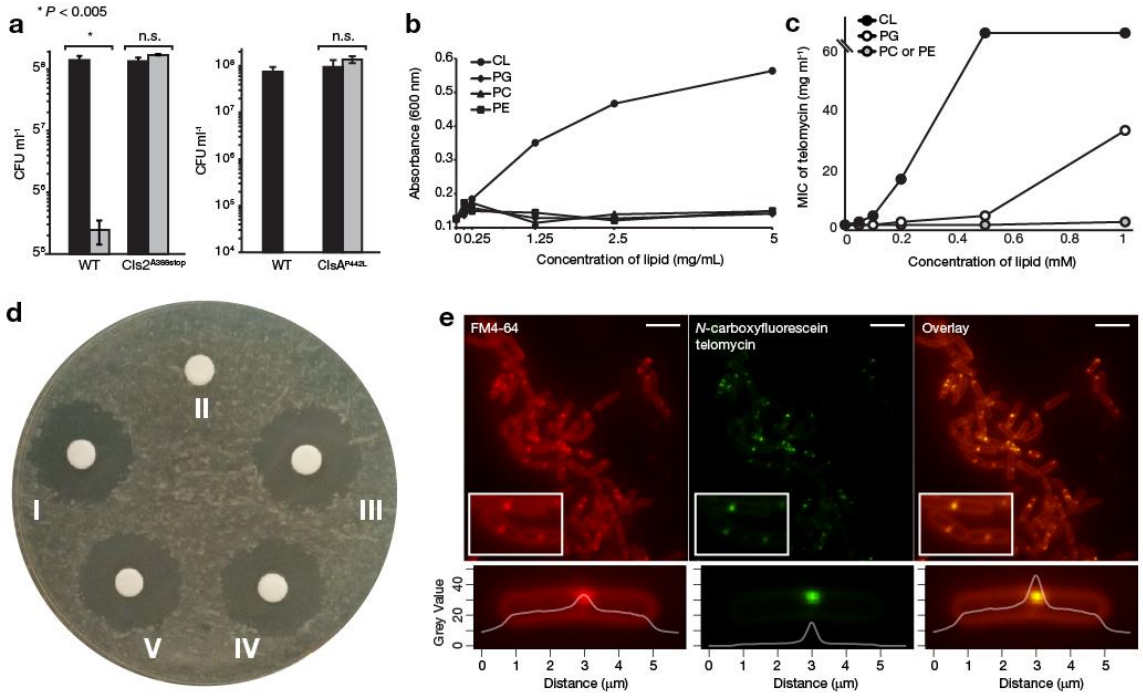



**Figure 3.4 Telomycin exerts bactericidal activity by interacting with cardiolipin.** (**a**) Mutating cardiolipin synthase provides spontaneous telomycin resistance. Wild type and spontaneously-resistant *S. aureus* (left) and *B. subtilis* (right) were exposed to 32× MIC concentrations of telomycin for 90 min. CFUs were taken before treatment (black) and after 90 min (white). Results are shown as ± s.d.; $n = 4$; Two-tailed student's *t*-test. $P = 0.0018$. Data are representative of three independent experiments. (**b**) Telomycin-cardiolipin mixtures uniquely and rapidly acquire turbidity. Telomycin (final concentration 5 mM; 6.3 mg $ml^{-1}$) was mixed with CL, PG, PE, or PC at concentrations ranging from 0.125 to 5 mg $ml^{-1}$. Mixtures were incubated for 5 min before turbidity was measured using absorbance at 600 nm. (**c**) Cardiolipin reduces telomycin's antibacterial activity in a dose-dependent manner. Telomycin (0.1 mM) was exposed to lipids at concentrations ranging from 1 to 0.05 mM, then used to determine MICs against wild type *B. subtilis*. Identical results were

observed for PE and PC (gray). (**d**) Cardiolipin abolishes telomycin's antibacterial activity. Telomycin (10 mM; I) was mixed with 2× molar equivalent of CL (II), PG (III), PE (IV), or PC (V). Mixtures were added to diffusion disks using *B. subtilis* as an indicator organism and grown overnight. (**e**) One-day cultures of *B. subtilis* were labelled with the membrane dye FM4-64 (left) and a fluorescent conjugate of telomycin (middle). Telomycin accumulated at poles and septa. Summed pixel intensities of a selected bacteria are depicted below. Scale bar, 5 μm.

cardiolipin[48]. Nevertheless, to investigate whether telomycin also localizes to these cardiolipin rich areas, we generated a fluorescein-labelled telomycin conjugate. Telomycin possesses a single free amine from the N-terminal aspartate residue, which is notably absent on the related molecule LL-AO341β and does not appear to impact activity. This amine was reacted with N-hydroxysuccinimide carboxyfluorescein to generate a fluorescent conjugate that demonstrated a slight decrease in antimicrobial activity. Imaging studies with live *B. subtilis* revealed that telomycin localizes to poles and septa (**Fig. 3.4e**) in a similar manner to the established cardiolipin dye NAO[46–48], albeit with less background signal from seemingly non-specific staining (**Supplementary Fig. 15**).

Although telomycin possesses a unique mechanism of action of and has demonstrated value *in vivo*[40], the emergence of spontaneous resistance could be problematic in a clinical setting. Interestingly, LL-AO341β demonstrates improved activity relative to telomycin[42], but lacks the N-terminal aspartate residue, as well as a hydroxylation on the C-terminal proline, suggesting that alterations to the telomycin

scaffold or tailoring may improve activity. To assess the impacts of structural alterations on antibacterial activity, we isolated six telomycin congeners and elucidated their structures using HRMS, MS/MS sequencing, and 2D NMR experiments (**Supplementary Note**). The planar structures of telomycin B–G (**2**–**7**) were similar to the telomycin A scaffold, but demonstrated incomplete tailoring, including loss of the cis-3-hydroxyproline hydroxylation (B–G), of the trans-3-hydroxyproline hydroxylation (C, E–G), loss of the erythro-hydroxyleucine hydroxylation (F–G), and saturation of the double bond in Z-$\Delta_{2,3}$-tryptophan. Although Z-$\Delta_{2,3}$-tryptophan saturation negatively affected bioactivity, the successive loss of each hydroxylation improved it, indicating that general hydrophobicity and tryptophan orientation may be important in driving the interaction with membranes in general and cardiolipin in particular. To assess the role of tryptophan in initiating this interaction we generated additional telomycin A analogues through directed biosynthesis, incorporating methyl-, methoxyl-, or hydroxyl groups at indole position 5 in both tryptophan residues (**Supplementary Note**). Consistent with a role for the indole rings in embedding in the membrane and initiating an interaction, hydroxylation completely abolished telomycin activity, methoxy groups caused a modest decrease in activity, and methylation improved activity (**Supplementary Table 5**). Importantly, telomycin F and di-5-methyltryptophan telomycin (**8**) possessed activity against resistant strains that was comparable with the activity of telomycin A against sensitive strains, suggesting that engineered variants with superior bioactivity can overcome potential resistance and take advantage of cardiolipin as a promising antibacterial target.

**3.5 Discussion**

The chemical and mechanistic diversity of antibacterial natural products has provided privileged scaffolds for drug development along with chemical probes of cellular processes. Here, we systematically profiled antibiotic scaffolds, using a new algorithm to sort these natural products according to their biosynthetic building blocks. Using this approach we could accurately group these diverse chemical structures into families and subfamilies for downstream analysis. Identified subfamilies could be profiled with extant metadata and passed through our comprehensive resistance pipeline to enrich for rare chemical families with unknown mechanisms of action. As an example of this approach, telomycin was identified as an uncommon scaffold with a potentially novel mechanism of action, which could be validated by *in silico* resistance profiling and subsequent experimental work. Substrate and tailoring promiscuity is a hallmark of these modular antibiotic assembly lines[5] and we leveraged this to produce improved telomycin analogues with superior activities against resistant isolates that were raised in the laboratory.

While this work was undergoing review, a biosynthetic study of telomycin was published[43], detailing the sequencing and analysis of the telomycin biosynthetic gene cluster, as well as the presentation of new telomycin congeners, including two which appear identical to telomycin B and C. This work delineated the boundaries of the telomycin biosynthetic gene cluster, indicating a number of elaborate tailoring modifications which take place to generate the mature telomycin scaffold. Most interestingly, this work demonstrated that telomycin undergoes transient acylation during biosynthesis, and noted that acylated variants had improved activity against antibiotic-resistant bacteria including vancomycin-resistant *Enterococci* (VRE). This finding is consistent with our results

demonstrating that hydrophobic telomycin variants have improved activity, likely due to an increased association with bacterial lipid membranes and cardiolipin in particular.

The selectivity of telomycin for cardiolipin is intriguing, as cardiolipin is a unique phospholipid dimer found in the membranes of nearly all bacteria and archaea[46–49]. Telomycin is not hemolytic and is not known to possess significant toxicity to mammals[40] and we define a lack of cytotoxicity at concentrations up to 128 μg $mL^{-1}$ in keeping with previous reports against cultured lines[43] where cardiolipin is exclusively sequestered to the inner mitochondrial membrane[46–49]. Unlike most other peptide antibiotics, it has been demonstrated that some members of the telomycin family of natural products may be orally-bioavailable[50], and it remains possible that future structural optimization may yield useful agents for combating bacterial infections.

In this work, we collected a comprehensive library of antibacterial natural products for the first time and have quantified the diversity that modular natural products possess, not just in their chemical scaffolds, but also in their evolved targets. By collecting and analyzing these molecules with a retrobiosynthetic algorithm, we quantified what natural products chemists have long suspected: that novel natural product scaffolds with specific activity often have novel targets. Assembly line-like enzymes can produce large numbers of unique scaffolds, and new scaffolds have nearly a 40% chance of inhibiting a new target. This unparalleled frequency of success in hitting distinct targets yields obvious conclusions for the future of antibiotic research. By charting and navigating this valuable collection of small molecules, we can readily identify unique chemical scaffolds that will have a high

likelihood of possessing distinct antibacterial targets. By taking up the approach, candidates (**Supplementary Dataset 1**), and resources (**Supplementary Fig.** 1) provided here, we hope the scientific community can meet the demand for antibiotics with uncharacterized mechanisms of action to overcome antibiotic resistance.

## 3.6 Materials and Methods

### 3.6.1 Cataloging the Natural Antibiotic Collective

We compiled a comprehensive database of all known antibiotics isolated from microbial sources. The Handbook of Antibiotic Compounds[28] was supplemented by an exhaustive review of published antibiotic literature and patents. For each antibiotic, chemical structures were generated in SMILES format, and both the known or suspected mechanism of action of the compound and relevant toxicity data were also recorded. In total, 10,343 microbially produced antibiotics were identified. Of these, 3,159 were selective to bacteria, meaning they had no broad cytotoxicity or activity against eukaryotic cells or organisms. A number of computational and manual curation processes were performed in order to ensure the high quality data. When chemical structures of the same molecule within two different data sources differed, the molecule was manually inspected and its SMILES redrawn. Publicly available SMILES were observed to frequently represent tautomers of their natural products, so an algorithm was implemented to redraw enol and iminol tautomers of ester and amide bonds, respectively. Finally, SMILES representing salts of their corresponding natural products were redrawn in their non-charged states.

### 3.6.2 Development of a retrobiosynthetic similarity scoring algorithm for natural products

In order to organize antibacterial natural products based on the biosynthetic history of these evolved small molecules, we developed a software package consisting of two algorithms. The first algorithm, termed GRAPE, performs *in silico* retrobiosynthesis of peptide, polyketide, and carbohydrate-containing compounds. GRAPE first cleaves defined biosynthetic bonds (chemical bridges [disulfide bonds, aromatic ethers, etc], heterocycles, core bonds [esters, amides, lactones, etc.], and tailoring modifications [glycosylation, methylation, sulfation, etc.]), then attempts to match remaining monomers with a comprehensive library of known biosynthetic units (amino acids, carbohydrates, etc.). Remaining units are parsed through a polyketide retrobiosynthetic module which scans the carbon backbone and assesses likelihood of polyketide origin, and predicts the architecture of the parent polyketide synthase (PKS). This collection of ordered monomers for each natural product deconstructed by GRAPE is then passed to a second algorithm for pairwise analysis based on order and composition. The second algorithm, termed GARLIC, aligns units of biosynthetic information identified by the retrobiosynthetic algorithm (i.e., proteinogenic and non-proteinogenic amino acids, ketide units, sugars, halogens, nonribosomal peptide and polyketide starter moieties, and selected tailoring modifications) to all molecules within the database of targeted antibiotics using a modified Needleman-Wunsch algorithm[51]. GRAPE and GARLIC implement chemical abstractions developed by the Chemistry Development Kit[52], version 1.4.19. A more comprehensive description of

the GRAPE and GARLIC algorithms can be found in Dejong, Chen, and Li *et al*. (*submitted*, 2015).

### 3.6.3 Development of PRISM

We recently described the development of PRISM[34] (PRediction Informatic for Secondary Metabolomes), a web application designed to identify biosynthetic gene clusters and predict the structures of genetically encoded nonribosomal peptides and type I and II polyketides. PRISM implements a library of nearly 500 hidden Markov models to identify conserved biosynthetic genes and the substrates of adenylation, acyltransferase, and acyl-adenylating enzymes. A simple greedy algorithm is used to identify plausible biosynthetic gene clusters, and rules specific to modular, *trans*-acyltransferase, and iterative type I polyketides, type II polyketides, and nonribosomal peptides are used to define true clusters. A library of 57 virtual tailoring reactions is leveraged in order to generate a combinatorial library of chemical structures when multiple potential substrates are biosynthetically plausible for one or more tailoring enzymes, including macrocyclization, heterocyclization, aromatization, halogenation, C- and O-glycosylation, O-, N-, and C-methylation, carbamoylation, amination, formylation, phosphorylation, sulfonation, oxidation and reduction, mono- and dioxygenation, Baeyer-Villager rearrangement, and acyl group transfer. A set of hidden Markov models for conserved deoxysugar biosynthesis genes and a BLAST database of natural product glycosyltransferases is used to predict potential combinations of hexose and deoxysugars which tailor the natural product scaffold, and a library of models for type II polyketide cyclases is used to predict the scaffolds of type II

polyketides. In addition to generating combinatorial libraries of chemical structures, biosynthetic information detected by PRISM can be aligned to biosynthetic information from other clusters, or to retrobiosynthetic information generated by GRAPE, within the GARLIC alignment package. The PRISM web application can be accessed at http://magarveylab.ca/prism.

### 3.6.4 Generation of the antibacterial tree

A subset of our antibacterial small molecule collection, consisting of 2,026 targeted antibiotics, was retrobiosynthetically decomposed using GRAPE, and the decomposed natural products were aligned to one another using GARLIC. The similarity matrices generated by GARLIC alignments of biosynthetic information were converted to dissimilarity matrices. The generated dissimilarity matrices were used to hierarchically cluster the antibiotics using the hclust function within the stats R package. A tree was generated with the ape R package[53]. The tree was visualized and clades were collapsed in Dendroscope based on observed families of natural products[54].

### 3.6.5 Development of a database of hidden Markov models for antibiotic resistance genes

In order to identify resistance determinants within biosynthetic gene clusters, we compiled a comprehensive database of 257 profile hidden Markov models for genes associated with antimicrobial resistance. 166 hidden Markov models were obtained from the Resfams antibiotic resistance profile hidden Markov model database[35]. A single model

(PFAM12847, a generic methyltransferase model) was removed as it was observed not to be specific to antibiotic resistance. The Resfams database was supplemented by the development of an additional 91 profile hidden Markov models associated with antibiotic resistance (**Supplementary Dataset 2**). Sequences were manually collected based on homology to experimentally annotated sequences, aligned using MUSCLE[55], and trimmed using trimAl[56] to remove gaps present in fewer than 50% of sequences. Hidden Markov models were generated from the resulting alignments using the hmmbuild program, version 3.1b1, from the HMMER3 software package[57]. Bitscore cutoffs for each hidden Markov model were determined by manual analysis of the results of a search of the UniProtKB database[58], using the HMMER web server[59]. The resulting library of hidden Markov models created in this study, and the aligned sequences used to construct them, are available at http://magarveylab.ca/resistance/.

### 3.6.6 Development of a web application to search the antibacterial chemical space

In order to make our results accessible to the broader scientific community, we developed a web application capable of querying our comprehensive antibacterial collection in order to predict the molecular target of real or predicted molecular structures. The web application accepts as input a single molecular structure, in SMILES format, or a line- or tab-delimited file of SMILES. User-submitted molecular structures undergo *in silico* retrobiosynthesis, and units of biosynthetic information are aligned to all molecules within a database of targeted antibiotics. The database implemented within the Antibioticome search web application is an annotated subset of the antibacterial library consisting of 1,868

compounds. Each compound is grouped into a larger family of natural products associated with a molecular target. For each user-submitted molecule, the web application reports the single highest-scoring targeted antibiotic as determined by retrobiosynthetic alignment, in addition to the natural product family to which that antibiotic belongs, its putative molecular target, and a score indicating the quality of the alignment.

A complete guide to the Antibioticome search web application is available at http://magarveylab.ca/antibioticome/#!/help. The Antibioticome search web application is written in Java 7 for the Tomcat 7 web server. To ensure that all results will remain confidential, discrete user logins will be made available in the near future.

### 3.6.7 General chemical procedures

1D ($^{1}H$ and $^{13}C$) and 2D ($^{1}H$-$^{13}C$ HMBC, HSQC, $^{1}H$-$^{1}H$ NOESY, TOCSY, and COSY) NMR spectra for telomycins were recorded on a Bruker AVIII 700 MHz NMR spectrometer in $d_6$-DMSO (*Sigma Aldrich*). High-resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (*ThermoFisher Scientific*) with an electrospray ionization source (ESI) and using CID with helium for fragmentation. LCMS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna C18 column (150 mm × 4.6 mm, *Phenomenex*) for analytical separations, running acetonitrile with 0.1% formic acid and dd$H_2O$ with 0.1% formic acid as the mobile phase.

### 3.6.8 Microbial strains and telomycin production

*Streptomyces canus* was purchased from the American Type Culture Collection (ATCC, ATCC no. 12647). *S. canus* was maintained on Bennett's agar at 28°C. *S. aureus* Newman was maintained on cation adjusted Mueller Hinton broth (CAMHB) agar at 37°C. *B. subtilis* 168 was maintained on CAMHB agar at 28°C.

Fresh colonies of *Streptomyces canus* were used to inoculate 50 mL cultures of GYM media containing 0.5% glycine (GGYM), and then grown for 72 h at 28°C and 250 rpm. 10 mL of starter culture was used to inoculate 500 mL of the same media, followed by growth for 72 h at 28°C and 250 rpm. Cultures were harvested by centrifugation, followed by a methanol extraction of the pellet and extraction of the supernatant with 2% absorbent HP-20 resin (*Diaion*). Resins were eluted with excess methanol, and the eluent was pooled with the pellet methanol extract and evaporated to dryness. The extract was resuspended in a small volume of methanol and separated on an open gravity column of LH-20 size exclusion resin (*Sephadex*) with methanol as a mobile phase. Fractions containing telomycin were pooled, evaporated to dryness, and resuspended in methanol. Telomycin was isolated by preparative scale LC-MS using a Luna 5 μm $C_{18}$ column (250 × 15 mm, *Phenomenex*) with water (0.1% formic acid) and acetonitrile (0.1% formic acid) as the mobile phase, at a flow rate of 10 mL/min. After 4 min, acetonitrile was increased in a linear manner (curve 5) from 5% to 30% at 14 min, then increased 31% by 20 min, then to 40% by 40 min, followed by a wash of 100% methanol. Telomycins eluted at the following retention times: A - 26 min, B - 28 min, C - 31.5 min, D - 32.5 min, E - 34 min, F - 35 min, G - 38 min.

### 3.6.9 Directed biosynthesis of new telomycins

Fresh colonies of *Streptomyces canus* were used to inoculate 50 mL cultures of GYM media containing 0.5% glycine (GGYM), and then grown at 28°C and 250 rpm. After 24 h, non-natural amino acids were added by sterile syringe filtration, to a final concentration of 4 mM. Cultures were grown for an additional 48 h and harvested, following the same isolation and purification procedure of natural telomycins. Unnatural amino acids were purchased from Sigma Aldrich. Individual analogue retention times were recorded and candidate peaks selected for purification and testing. Chemical structures were assigned using NMR and high-resolution mass spectrometry (see Supplementary information).

### 3.6.10 Determination of antibacterial activity

Minimum inhibitory concentrations (MICs) for telomycins were determined using broth microdilution in cation-adjusted Mueller Hinton broth (CAMHB). *Bacillus subtilis* strains were cultured at 28°C, and *S. aureus* strains were cultured at 37°C. The MIC was determined as the lowest concentration of drug at which no growth was observed after 16 h. For assessing the impact of treating strains with sodium chloride concentrations prior to telomycin, *S. aureus* was grown overnight in CAMHB containing 0, 0.5, 1, or 1.5 M NaCl and then MICs were recorded as described above. Broth microdilution assays were performed in fresh media with identical NaCl concentrations and again MIC was determined as the lowest concentration at which no growth was observed after 16 h.

### 3.6.11 Measuring turbidity of telomycin-lipid mixtures

To measure the turbidity of telomycin-lipid mixtures, telomycin was dissolved in methanol at a concentration of 12.8 mg/mL, and cardiolipin (*Sigma Aldrich*; C0563; ≥98% pure), phosphatidylglycerol (*Sigma Aldrich*; 63371; ≥98% pure), phosphatidylcholine (*Sigma Aldrich*; P3556; ; ≥99% pure), and phosphatidylethanolamine (*Sigma Aldrich*; P7943; ≥97% pure) were dissolved in methanol at a concentration of 10 mg/mL. 20 μL of telomycin was mixed with 20, 10, 5, 1, or 0.5 μL of each lipid (final reaction volume 40 μL) in a flat-bottom polystyrene 96-well plate. Optical density (OD = 600 nm) was measured after 10 min.

### 3.6.12 Colony forming unit (CFU) assays

Cultures of *S. aureus* and *B. subtilis* were grown overnight in CAMHB. Starter cultures were used to inoculate fresh media and were grown to OD = 0.2, after which they were transferred to 96-well plates. Cultures were serially diluted and plated on CAMHB agar to determine initial colony-forming units. Telomycin dissolved in DMSO was added to each well (1:100) and incubated for 90 min, followed by serial dilution and plating to determine colony-forming units. Results are shown as ± s.d.; $n = 4$; Two-tailed student's *t*-test.

### 3.6.13 Red blood cell (RBC) hemolysis assay

Hemolytic activity of telomycin was measured against a 0.25% sheep red blood cell (RBC, *Fisher Scientific*) suspension in phosphate-buffered saline. Telomycin was serially diluted from 256 to 2 μg/mL and incubated with the RBC suspension for 1 h at 37°C in a polypropylene 96-well plate with conical wells. After 1 h, RBCs were pelleted (1000 × g

for 5 min) and the supernatant was transferred to a flat-bottom polystyrene 96-well plate, measuring absorbance at 540 nm. 1% Triton X-100 was used as a positive control, while DMSO alone and RBCs alone were used as negative controls.

### 3.6.14 Measuring bioactivity of telomycin-lipid mixtures

To assess the impact of various lipids on telomycin's antibacterial activity on solid agar, 10 μL of 10 mM telomycin was mixed with 10 μL of 20 mM cardiolipin (*Sigma Aldrich*; C0563; ≥98% pure), phosphatidylglycerol (*Sigma Aldrich*; 63371; ≥98% pure), phosphatidylcholine (*Sigma Aldrich*; P3556; ≥99% pure), or phosphatidylethanolamine (*Sigma Aldrich*; P7943; ≥97% pure). All compounds were dissolved in methanol. Mixtures were incubated for 10 min, then added to diffusion disks, allowed to dry, then placed on a CAMHB agar plate with *B. subtilis* as an indicator organism. The plate was incubated overnight at 28°C.

To assess the impact of various lipids on telomycin's antibacterial activity in liquid media, telomycin was dissolved in methanol at 12.8 mg/mL (approx. 10 mM) and 1 μL was added to a 96-well plate. Lipids were dissolved in methanol and 10 μL of 10, 5, 2, 1, or 0.5 mM lipid was added to each telomycin-containing well. CAMHB inoculated with *B. subtilis* was added to each well and serially diluted to determine the minimum inhibitory concentration of each telomycin - lipid mixture. Plates were incubated shaking at 28°C for 16 h.

### 3.6.15 Preparation of N-fluorescein labelled telomycin

To generated a telomycin-fluorescein conjugate, 2.5 mg of telomycin A was dissolved in 200 μL DMSO and mixed with 10 mg 5(6)-carboxyfluorescein N-hydroxysuccinimide ester (*Sigma Aldrich*) dissolved in 180 μL DMSO, and 20 μL 0.5 M sodium bicarbonate. The reaction was allowed to proceed overnight at room temperature, after which the reaction was quenched by the addition of formic acid and N-labelled telomycin was purified by preparative scale LC-MS.

### 3.6.16 Measuring cardiolipin content of bacterial cells

Cardiolipin was extracted using an acidic Bligh Dyer method. A 50 mL culture of bacteria grown for 24 h was pelleted and resuspended in 1 mL 0.1 N HCl. 2.5 mL of methanol and 1.25 mL of chloroform was added to each sample, followed by 30 min incubation at room temperature. After this, 1.25 mL of 0.1 N HCl and 1.25 mL of chloroform was added to create a two-phase solution that was then centrifuged at 3000 × g for 10 min. The bottom phase was recovered, evaporated to dryness, and then resuspended in methanol. An established LC-MS method was used to quantify cardiolipin content, using a reverse-phase Luna C18 column (150 mm × 4.6 mm, *Phenomenex*). The mobile phases were A (90% acetonitrile, 10% water, 0.5% glacial acetic acid, 0.5% triethylamine) and B (90% isopropanol, 10% water, 0.5% glacial acetic acid, 0.5% triethylamine), running at 0.8 mL/min. After 3 min, solvent B was increased from 50% to 100% by 22 min, then held for 15 min, before returning to 50% by 38 min. Cardiolipin species eluted between 27 and 32 min. Relative cardiolipin content was assessed by summing the areas associated with cardiolipin species ions in simultaneously and identically extracted wild type and mutant

bacteria, presenting percentage of cardiolipin signal detected in the mutant strain, relative to wild type.

### 3.6.17 Genome sequencing and analysis of antibiotic biosynthetic gene clusters

A single colony of *Streptomyces canus* (ATCC 12647) was used to inoculate a 50 mL culture of GYM media containing 0.5% glycine (GGYM), and then grown for 96 h at 30°C and 250 rpm. 500 μL of culture was centrifuged at 12 × g for 5 min and resuspended in 500 μL SET buffer (75 mM NaCl, 25 mM EDTA pH 8.0, 20 mM Tris HCl pH 7.5, 2 mg/mL lysozyme) to lyse for 2 h at 37°C. Proteinase K and SDS were added after lysis to final concentrations of 0.5 mg/mL and 1%, respectively. Lysis mixtures were incubated at 55°C for 2 h before adjusting the concentration of NaCl to 1.25 M and extracting twice with phenol-chloroform. Isopropanol was added (equivalent to 60% the volume of the solution) to precipitate genomic DNA, which was subsequently washed twice with 70% ethanol and resuspended in sterile water for sequencing. For sequencing *Staphylococcus aureus* and *Bacillus subtilis* strains, single colonies were used to inoculate 3 mL overnight cultures in tryptic soy broth (TSB), incubated a 37°C and 30°C respectively. Genomic DNA was isolated using a GenElute Genomic DNA Extraction kit (*Sigma Alrich*). Genomic DNA for all strains was sent for library preparation and Illumina sequencing at the Farncombe Metagenomics Facility at McMaster University, using an Illumina HiSeq DNA sequencer. Contigs were assembled using the ABySS genome assembly program and with Geneious bioinformatic software.

### 3.6.18 Identification of telomycin-resistance mutations in sequenced isolate genomes

To identify mutations that conferred resistance to telomycin, we isolated colonies of *S. aureus* Newman that appeared inside the zone of inhibition observed during telomycin disk diffusion assays. Colonies were cultured overnight in TSB containing 50 μg/mL telomycin (~ 5× MIC), then used to inoculate a 96-well plates containing a serial dilution of telomycin, up to 128 μg/mL. Genomic DNA was extracted from highly resistant isolates and their sensitive parental strain, and sequenced with an Illumina MiSeq platform. Sequences were compared using BreSeq, which exclusively identified mutations in the predominant, house-keeping cardiolipin synthase (cls2). *Bacillus subtilis* telomycin-resistant mutants were identified in a similar manner, and mutations in cardiolipin synthase were confirmed by sequencing PCR products of the two cardiolipin synthase genes: *clsA* and *clsB*. A point mutation was observed next to the second HKD motif active site in clsA, which resulted in a 90% decrease in cardiolipin levels. PCR primers used for the amplification of *B. subtilis* cardiolipin synthase genes are as follows – clsAF: 5′-GTTTTAAAGAAATCTGCCCG-3′; clsAR: 5′-GCGAGACGGATTCTTTTATT-3′; clsBF: 5′-ATGAAGGTATTTATCGTGT-3′; clsBR: 5′-TTATAAGAAATAAGATAATG-3′.

### 3.6.19 Structure elucidation

The structure of telomycin A was confirmed by a series of 1D and 2D NMR spectroscopy experiments, high resolution mass measurement, and MS/MS fragmentation and annotation. Structures of new naturally occurring variants were elucidated by MS/MS fragmentation and annotation, high resolution mass measurements, and comparison of 1D

and 2D NMR experiments to those of telomycin A. Structures of telomycins generated by directed biosynthesis were confirmed by MS/MS fragmentation and annotation, and high resolution mass measurements. The structure of di-5-methyltryptophan telomycin was also confirmed by 1D and 2D NMR experiments.

Figures including structures, MS/MS fragmentation, and NMR spectra, as well as tables including high resolution mass measurements and NMR chemical shifts can be found in the Supplementary Note.

### 3.6.20 Cytotoxicity assay

HEK293 cells were obtained from the American Type Culture Collection (ATCC; ATCC CRL-1573) and maintained in Minimal Essesntial Media (MEM) Alpha modifications supplemented with 10% fetal bovine serum. Cellular identity was confirmed by Cells were cultured in 96-well plates containing 200 μL of media and 5000 cells per well. After 3 h of incubation, cells were treated with a serial dilution of telomycin A. Experiments were performed in duplicate with DMSO as a negative control and mitomycin C as a positive control. After 48 h incubation, 10% (22 μL) Alamar Blue (*Life Technologies*) was added and incubated with cells for 4 h at 37°C. Fluorescence was measured at an excitation wavelength of 530 nm and an emission wavelength of 590 nm and compared to a no-drug control.

### 3.6.21 Fluorescence microscopy

A stationary phase culture of *Bacillus subtilus* 168 was washed twice in LB medium to remove extracellular debris, centrifuging at 10,000 rpm for 1 min each wash. Both *N*-carboxyfluorescein telomycin and 10-*N*-nonyl acridine orange (NAO) were added to cells to compare probe localization. *N*-carboxyfluorescein telomycin was added to a final concentration of about 1 μM, and NAO was added at a final concentration of 0.5 μM. Both probes were incorporated for 20 min in the dark. The suspensions were washed again to remove extracellular probe, then added to a poly-L-lysine coated 0.17 μm glass bottom microplate (*Brooks Automation*). To these samples, the membrane dye FM 4-64 was added to a final concentration of 1 μg/mL. Cells were imaged using a Nikon Eclipse Ti inverted microscope at 1000× magnification with a Nikon Plan Apo λ 100× oil-immersion objective. Overlays were prepared using the Nikon Elements software suite.

Profile plots were prepared in ImageJ[60] by cropping regions of interest, and converting to 8-bit greyscale images. No background subtractions were made, and the grey values across the length of a cell were plotted for FM 4-64, *N*-carboxyfluorescein telomycin, and overlay, to highlight probe localization.

### 3.7 Supplementary Information

Supplementary information from the publication referred to in this chapter can be found in Appendix 2.

### 3.8 References

1. Bush, K. *et al*. Tackling antibiotic resistance. *Nature Rev. Microbiol.* **9**, 894–896 (2011).
2. Fischbach, M. A. & Walsh, C. T. Antibiotics for emerging pathogens. *Science* **325**, 1089–1093 (2009).
3. Newman, D.J. & Cragg, G.M. Natural products as a source of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75**, 311–335 (2012).
4. Vining, L.C. Roles of secondary metabolites from microbes. *Ciba Foundation Symposium 171 - Secondary Metabolites: their Function and Evolution*. (2007).
5. Fischbach, M.A. & Clardy, J. One pathway, many products. *Nature Chemical Biology* **3**, 353–355 (2007).
6. Payne, D.J. *et al*. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov.* **6**, 29–40 (2007).
7. Baumann, S. *et al*. Cystobactamids: Myxobacterial Topoisomerase Inhibitors Exhibiting Potent Antibacterial Activity. *Angew. Chem. Int. Edn. Engl.* **53**, 14605–14609 (2014).
8. Lin, A.H. *et al*. The oxazolidinone eperezolid binds to the 50S ribosomal subunit and competes with binding of chloramphenicol and lincomycin. *Antimicrob. Agents Chemother.* **41**, 2127–2131 (1997).
9. Keller, S. *et al*. Action of atrop-abyssomicin C as an inhibitor of 4-amino-4-deoxychorismate synthase PabB. *Angew Chem Int Ed Engl*. **46**, 8284–8286 (2007).

10. Li, J.W. & Vederas, J.C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161–50 (2009).
11. Cundliffe, E. & Demain, A.L. Avoidance of suicide in antibiotic-producing microbes. *J. Ind. Microbiol. Biotechnol*. **37**, 643–672 (2010).
12. D'Costa, V.M. *et al*. Sampling the antibiotic resistome. *Science* **311**, 374–377 (2006).
13. Thaker, M.N. *et al*. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31**, 922–927 (2013).
14. Bibikova, M.V., Ivanitskaia, L.P. & Singal, E.M. Directed screening of aminoglycoside antibiotic producers on selective media with gentamycin. (Original in Russian.) *Antibiotiki* **26**, 488–492 (1981).
15. Ivanitskaia, L.P., et al. Use of selective media with lincomycin for the directed screening of antibiotic producers. (Original in Russian.) *Antibiotiki* **26**, 83–86 (1981).
16. Forsberg, K.J. *et al*. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **337**, 1107–1111 (2012).
17. Boucher, H. W. *et al*. Bad bugs, no drugs: no ESKAPE! an update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
18. Doroghazi, J.R. *et al*. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol*. **10**, 963–968 (2014).
19. Donadio, S. *et al*. Antibiotic discovery in the twenty-first century: current trends and future perspectives. *J. Antibiot.* **63**, 423–430 (2010).
20. Walsh, C.T. & Wencewicz, T.A. Prospects for new antibiotics: a molecule-centered perspective. *J. Antibiot.* **67**, 7–22 (2014).

21. Goss, R.J.M., Shankar, S. & Fayad, A.A. The generation of ‘‘unNatural’’ products: Synthetic biology meets synthetic chemistry. *Nat. Prod. Rep*. **29**, 870–889 (2012).
22. Ling, L.L. *et al*. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
23. Cociancich, S. *et al*. The gyrase inhibitor albicidin consists of *p*-aminobenzoic acids and cyanoalanine. *Nat. Chem. Biol.* **11**, 195–197 (2015).
24. Hamamoto, H. *et al*. Lysocin E is a new antibiotic that targets menaquinone in the bacterial membrane. *Nat. Chem. Biol.* **11**, 127–133 (2015).
25. Weber, T. et al. antiSMASH 3.0 – a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucl. Acids Res. 43, W237–W243 (2015).
26. Medema, M.H. *et al*. Minimum Information about a Biosynthetic Gene Cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
27. Hadjithomas, M. *et al*. IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *mBio* 6, e00932-15 (2015).
28. Bérdy, J., *et al*. Handbook of Antibiotic Compounds, Vols I–X, CRC Press, Boca Raton, Florida, USA (1980–1982).
29. Bérdy, J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J. Antibiot.* **65**, 385–395 (2012).
30. Koch, M.A. *et al*. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. USA.* **102**, 17272–17277 (2005).

31. Over, B. *et al*. Natural-product-derived fragments for fragment-based ligand discovery. *Nat. Chem.* **5**, 21–28 (2013).
32. Wilson, D.N. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat. Rev. Microbiol.* **12**, 35–48 (2014).
33. Srivastava, A. *et al*. New target for inhibition of bacterial RNA polymerase: 'switch region'. *Curr. Opin. Microbiol*. **14**, 532–543 (2011).
34. Skinnider, M.A. *et al*. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res*. **43**, 9645-9662 (2015).
35. Gibson, M.K., Forsberg, K.J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. **9**, 207–216 (2015).
36. Nakao, M. *et al*. Pyloricidins, novel anti-*Helicobacter pylori* antibiotics produced by *Bacillus* sp. I. Taxonomy, fermentation and biological activity. *J. Antibiot.* **54**, 926–933 (2001).
37. Nakajima, N. *et al*. Mycoplanecins, novel antimycobacterial antibiotics from *Actinoplanes awajinensis* subsp. *mycoplanecinus* subsp. nov. II. Isolation, physico-chemical characterization and biological activities of mycoplanecin A. *J. Antibiot.* **36**, 961–966 (1983).
38. Misiek, M. *et al*. Telomycin a new antibiotic. *Antibiot. Annu.* **852**, (1957–1958).
39. Gourevitch, A. *et al*. Microbiological studies on telomycin. *Antibiot. Annu.* **856**, (1957–1958).

40. Tisch, D.E., Huftalen, J.B. & Dickson, H.L. Pharmacological studies with telomycin. *Antibiot. Annu.* **863**, (1957–1958).

41. Sheehan, J.C. *et al*. The structure of telomycin. *J. Am. Chem. Soc.* **90**, 462–470 (1968).

42. Oliva, B. et al. Mode of action of the cyclic depsipeptide antibiotic LL-AO341β, and partial characterization of a *Staphylococcus aureus* mutant resistant to the antibiotic. *J. Antimicrob. Chemother.* **32**, 817–830 (1993).

43. Fu, C. *et al*. Biosynthetic studies of telomycin reveal new lipopeptides with enhanced activity. *J. Am. Chem. Soc.* **137**, 7692–7705 (2015).

44. Tsai, M. *et al*. *Staphylococcus aureus* requires cardiolipin for survival under conditions of high salinity. *BMC Microbiol*. **11**, doi:10.1186/1471-2180-11-13 (2011).

45. Machaidze, G., Ziegler, A. & Seelig, J. Specific binding of Ro 09-0198 (cinnamycin) to phosphatidylethanoloamine: a thermodynamic analysis. *Biochemistry* **41**, 1965–1971 (2002).

46. Kawai, F. *et al*. Cardiolipin domains in *Bacillus subtilis* Marburg membranes. *J. Bacteriol*. **1788**, 2084–2091 (2009).

47. Mileykovskaya, E. & Dowhan, W. Cardiolipin membrane domains in prokaryotes and eukaryotes. *Biochim Biophys Acta.* **1817**, 1937–1949 (2012).

48. Oliver, P.M. *et al*. Localization of anionic phospholipids in *Escherichia coli* cells. *J. Bacteriol.* **196**, 3386–3398 (2014).

49. Arias-Cartin, R. *et al*. Cardiolipin binding in bacterial respiratory complexes: Structural and functional implications. *Biochim Biophys Acta.* **1817**, 1937–1949 (2012).

50. Stolpnik, V.G., Solovena, Y.V. & Antonenko, L.I. Neotelomycin blood levels in rabbits following intramuscular or oral administrations (in Russian). *Antibiotiki* **11**, 567–568.

51. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–453 (1970).

52. Steinbeck, C. *et al*. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **43**, 493–500 (2003).

53. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

54. Huson, D.H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).

55. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

56. Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

57. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195. (2011).

58. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, bar009. (2011)

59. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–37 (2011).

60. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Meth.* **9**, 671–675 (2012).

# Chapter 4. Informatic Analysis Reveals *Legionella* as a Source of Novel Natural Products.

## 4.1 Chapter preface

One of the most significant findings that has emerged from widespread bacterial genome sequencing is that biosynthetic potential is much more widespread than had previously been suspected. Aside from well-established prolific bacterial families including the Actinomycetes, Bacilli, Cyanobacteria, and Myxobacteria, a number of understudied bacterial genera demonstrate significant potential as a source of novel natural products and drug-like molecules. In this work, we used our established PRISM bio- and chemo-informatic platform to profile *Legionella* as a new source of natural products. *Legionella* possess a conserved biosynthetic potential, and while certain families of NRPS and PKS gene clusters are relatively conserved, there is also significant variability, indicating that continued sequencing efforts will reveal new gene clusters. Most importantly, the gene clusters found in *Legionella* are highly unusual, and do not appear in other bacterial genomes. To demonstrate that *Legionella* could serve as a source of novel natural products, I constructed a series of genetic knockouts to inactivate ketosynthases found in *Legionella pneumophila*. Metabolomic profiling revealed that inactivating one ketosynthase caused a loss of a series of molecules observed in wild type extract, coinciding with a loss of sliding motility. Isolation of this molecule yielded a unique polyketide and nonribosomal peptide hybrid compound, hence referred to as legionellol A. In addition to the conserved scaffold legionellol A, we observed a series of variably acylated variants, and demonstrated that this acyl legionellol complex was responsible for sliding motility in *Legionella*. As genome

sequencing continues to shine a light on previously unexplored bacteria, informatic strategies will be critical to prioritize investigations for novel natural products that could provide new drug-like chemical scaffolds.

The following chapter is a modified version of previously published journal article in which I was the lead author. I performed all experiments and bioinformatic analysis for this work, contributed to study design, and wrote the manuscript. Jonathan Plumb performed macrophage assays and microscopy analysis for an earlier version of this manuscript. Xiang Li performed NMR structure elucidation of legionellol A. Prof. Sergio Grinstein contributed to study design and microscopy data analysis. Prof. Nathan Magarvey contributed to study design and helped write the manuscript. The citation for this publication is as follows:

**Johnston, C.W.**, Plumb, J., Li, X., Grinstein, S., & Magarvey, N. (2015) Informatic Analysis Reveals *Legionella* as a Source of Novel Natural Products. *Synthetic and Systems Biotechnology* (Accepted).

## 4.2 Abstract

Microbial natural products are a crucial source of bioactive molecules and unique chemical scaffolds. Despite their importance, rediscovery of known natural products from established productive microbes has led to declining interest, even while emergent genomic data suggests that the majority of microbial natural products remain to be discovered. Now, new sources of microbial natural products must be defined in order to provide chemical scaffolds for the next generation of small molecules for therapeutic, agricultural, and

industrial purposes. In this work, we use specialized bioinformatic programs, genetic knockouts, and comparative metabolomics to define the genus *Legionella* as a new source of novel natural products. We show that *Legionella spp*. hold a diverse collection of biosynthetic gene clusters for the production of polyketide and nonribosomal peptide natural products. To confirm this bioinformatic survey, we create targeted mutants of *L. pneumophila* and use comparative metabolomics to identify a novel polyketide surfactant. Using spectroscopic techniques, we show that this polyketide possesses a new chemical scaffold, and firmly demonstrate that this unexplored genus is a source for novel natural products.

## 4.3 Introduction

Microbial natural products have been the most important source of chemical scaffolds and drugs for the last century [1]. Culturing random collections of bacteria and fungi isolated from the environment led to the identification of prolific natural product producing genera that then received extensive focus, including *Streptomyces* [2], *Bacillus* [3], Myxobacteria [4], and Cyanobacteria [5]. However, this tight focus, along with over-reliance on classical bioactivity-guided isolation approaches, has led to diminishing returns from natural products drug discovery efforts [6], resulting in the closure of most industrial natural products programs. In spite of this trend, untargeted sequencing of bacterial genomes has revealed that biosynthetic potential is much more widespread than was previously suspected [7], and that the majority of microbial natural products are still awaiting discovery [8-9]. In addition, it is now known that only a small fraction of bacteria

can be readily cultured in laboratory conditions [10], and that many talented natural products producers were likely missed in initial screening efforts. In light of this information, new genome-guided discovery efforts are a promising way to reveal valuable chemical scaffolds from previously uncharted bacteria.

Extensive studies on the biosynthesis of bacterial natural products have led to informatic strategies for their discovery based on extant genomic information [11-14]. Two of the most diverse and abundant classes of microbial natural products, polyketides and nonribosomal peptides, are produced by multi-enzyme assembly lines [15] – referred to as polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs) respectively – that can be readily detected in bacterial genomes. Genome sequencing has revealed a number of interesting bacterial families that could be candidates for second generation genome-guided natural products discovery efforts. Recent examples from *Clostridia* [16], *Eleftheria* [17], and *Entotheonella* [18] have demonstrated that exotic bacteria that had been challenging to culture can be a valuable resource of new molecules that were not discovered in previous random screening efforts.

In 2004, researchers described a unique polyketide fluorophore from a species of *Legionella* [19], a member of the diverse Gamma Proteobacteria that had not previously been investigated for natural products. In its native environment, *Legionella* reproduces by infecting environmental amoebae, replicating inside a specialized vacuole following phagocytosis [20]. *Legionella* is also able to infect some mammalian phagocytes, causing a pneumonia known as Legionnaire's disease in immunocompromised individuals. *Legionella* was first isolated following an outbreak in 1977, and was found to require highly

specialized media for growth [21-22]; relatively little remained known about the biology of this organism until the late 1990's. In part because of its relevance as a potential pathogen, a number of *Legionella* genomes are now sequenced, demonstrating a great deal of intra-genus differentiation, as well as a number of polyketide and nonribosomal peptide biosynthetic gene clusters. Given its niche culture requirements, relatively late discovery, and genetic diversity, *Legionella* is a strong candidate for the identification of new chemical scaffolds using genome-guided discovery efforts. In this work, we use bioinformatic tools to chart the diversity of natural product biosynthetic gene clusters in *Legionella*, and use targeted mutagenesis and comparative metabolomics to reveal a novel natural product chemical scaffold from this under-explored bacterium.

## 4.4 Results

### Bioinformatic Assessment of Biosynthetic Potential in *Legionella*

To assess the diversity of polyketide and nonribosomal peptide gene clusters present in *Legionella*, we profiled all sequenced genomes available for this genus through our software for PRediction Informatics for Secondary Metabolomes (PRISM [12]; Figure S1). This in-house web application is able to efficiently identify polyketide and nonribosomal peptide gene clusters from bacterial genomes, compare them to known biosynthetic gene clusters, and predict the structures of their small molecule products. We used PRISM to analyze 34 sequenced *Legionella* genomes – including 15 *L. pneumophila* strains – revealing a diverse array of NRPS, PKS, and hybrid biosynthetic gene clusters (Figure 4.1). From our sample of 34 genomes, we identified 141 biosynthetic gene clusters related to

nonribosomal peptides (46), polyketides (7), or hybrid (88) systems, with an average of 4 biosynthetic gene clusters per genome. In addition to the raw number of hybrid assembly systems found in *Legionella* genomes, these clusters were also the most diverse, with 18 distinct hybrid biosynthetic gene clusters from these 34 genomes, compared to 13 distinct nonribosomal peptide gene clusters and only 3 purely polyketide gene clusters. In sharp contrast to the massive canonical assembly line architectures [15], biosynthetic assembly lines in *Legionella* are highly fragmented and small, with many comprised of individual domains or lone multi-domain modules. This was particularly pronounced for nonribosomal peptide gene clusters, which nearly always possessed only a single monomer-activating adenylation domain (11 of 13). The majority of *Legionella* polyketide synthases, either in pure polyketide systems or in hybrid systems, were fully or partially *trans*-AT [25] (14 of 21), where the monomer-loading acyl transferase enzyme is sourced from outside the multi-domain assembly line. Although trans-AT polyketide synthases are becoming less rare with continued sequencing, *Legionella* has an unusual twist, in that nearly all of the trans-AT systems do not have an associated acyltransferase in the biosynthetic gene cluster, which is exceptionally rare [26]. In addition to the unusual organization, size, and monomer-loading strategies of these biosynthetic clusters, they also possess a large number of non-canonical termination mechanisms. In canonical multi-domain assembly line polyketide or nonribosomal peptide biosynthesis the small molecule chain is released from the enzyme assembly line via thioester hydrolysis through a C-terminal thioesterase domain [15]. Although thioesterases are still the most abundant chain release enzyme (12 of 33 complete clusters), most *Legionella* biosynthetic gene clusters do

not possess thioesterases, and instead appear to rely on NADH-dependent reductases [27] (5 of 33) or condensation domains (9 of 33), the latter of which are exceptionally rare in described bacterial polyketide and nonribosomal peptide biosynthetic gene clusters [28-30]. Most notably, many gene clusters did not possess analogues of any of these established chain release enzymes, suggesting alternative means of terminating thio-template natural product biosynthesis.

During this bioinformatic analysis, we did not uncover any biosynthetic gene clusters that were substantially similar to established or sequenced gene clusters from other bacteria, including from well-studied genera such as *Streptomyces* or relatively closely related natural product producers, such as *Pseudomonas*. Despite the unusual architecture of most of these biosynthetic gene clusters – which often lack conventional monomer-loading or chain-release enzymes – many of these assemblages can be found in different species and appear to be well-conserved. *Legionella* are known to have a patchwork genome [31], which rapidly sheds and acquires genes in accordance with survival needs [32], suggesting that these conserved unconventional biosynthetic gene clusters are likely still functional. Moreover, transcriptomics studies in *Legionella pneumophila* have demonstrated that these biosynthetic gene clusters are transcriptionally active in a variety of culture conditions [33-34]. To assess whether any of these unusual gene clusters produced new natural product scaffolds, we chose to pursue a genomic and metabolomic strategy with the genetically-tractable *L. pneumophila* strain LP02.

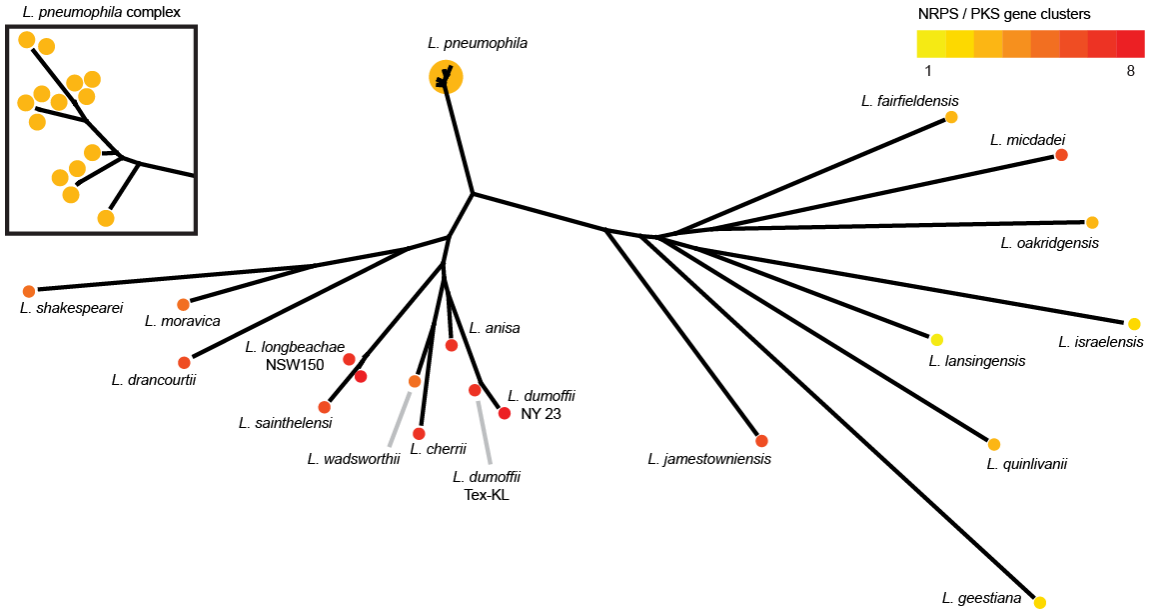



**Figure 4.1** PRISM analysis reveals that *Legionella* is a diverse genus with conserved biosynthetic potential. PRISM was used to identify polyketide and nonribosomal peptide gene clusters in 34 sequenced *Legionella* genomes, which was confirmed by manual inspection. PKS and NRPS gene clusters were sorted as a heat map and overlaid onto a phylogenetic tree constructed using 16S rRNA sequences to highlight the conserved biosynthetic potential of *Legionella*.

**Genetic and Metabolic Profiling of *L. pneumophila* uncovers a Novel Natural Product**

We constructed a series of genetic knockouts in *L. pneumophila* strain LP02, targeting ketosynthases in each of the 3 hybrid polyketide-nonribosomal peptide gene clusters visible in the genome (Figure 4.2, Table 1, Supplementary Table 1-2). Each of these gene clusters are well-conserved in *L. pneumophila*, represented in nearly all of the 15 complete or partially-sequenced strain genomes. However, each cluster has a highly

unusual architecture, as two (represented by Δlpg1939 and Δlpg2228) are composed exclusively of individual biosynthetic domains and lack chain terminating enzymes, while the third (represented by Δlpg2186) encodes a minimal NRPS and *trans*-AT PKS without an apparent *trans*-acting acyltransferase present in the genome. Following insertional activation, we observed that deletion of a ketosynthase (Δlpg2228) in one of these minimal gene clusters caused a loss of sliding motility, indicating the production of a surfactant, consistent with a previous microbiological study [35] (Figure 4.3A).

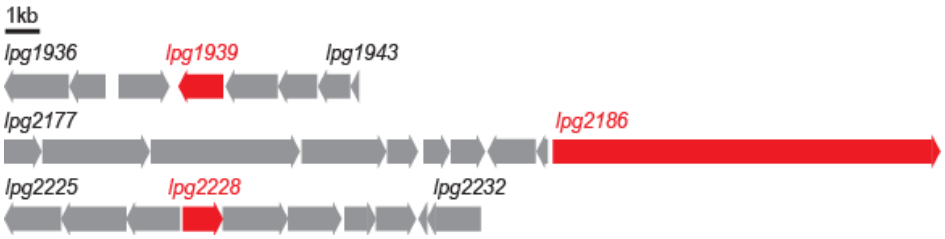



**Figure 4.2** Hybrid polyketide-nonribosomal peptide gene clusters of *L. pneumophila* targeted for mutagenesis. PRISM identified three hybrid PKS-NRPS gene clusters in the genome of *L. pneumophila* LP02, roughly defined as lpg1936-lpg1943, lpg2177-lpg2186, and lpg2225-lpg2232. Ketosynthases in each gene cluster were insertionally inactivated by insertion of a kanamycin resistance gene, disrupting expression of lpg1939, lpg2186, and lpg2228 (red).

To identify polyketides or nonribosomal peptides associated with the identified gene clusters, we used a comparative metabolomic strategy. All *L. pneumophila* LP02 wild type and knockout strains were grown in rich or chemically defined media until several days past maturity, before cell pellets were collected by centrifugation and extracted with

methanol, and supernatants were extracted with absorbent HP20 resin. Cell pellet and supernatant extracts were pooled and processed for small molecule contents using liquid chromatography paired with mass spectrometry (LCMS). Data files and chromatograms were analyzed to identify distinguishing molecular features between varying genetic backgrounds using Bruker MetaboliteDetect and ProfileAnalysis software, which enabled comprehensive chromatographic analysis and principal component analysis (PCA) respectively. The results of these analyses indicated no metabolomic differences between Δlpg2186, Δlpg1939, and wild type LP02. However, the Δlpg2228 mutant was shown to have significant metabolomic deviations from both wild type and the other mutants (Figure 4.3B, Figure S2). This strain had previously been observed to be deficient in sliding motility (Figure 4.3A), and appears to lack a series of metabolites which would correspond to the absent surfactant and biosynthetic intermediates. This series was present in all other biosynthetic mutants, which likewise showed no deviation from the wild type sliding colony phenotype, indicating that the unusual and minimal gene cluster associated with these metabolites likely does not rely on polyketide-derived precursors supplied in *trans*, at least not from the identified gene clusters. Although these surfactants ionize quite well, ionization intensity during mass spectrometry is not necessarily correlated with high abundance, and we found during initial isolation attempts that these molecules were present in extremely low quantities.

**Table 4.1** Genes of a *L. pneumophila* hybrid biosynthetic gene cluster (lpg2225-2234).

| Gene name | Locus tag | Predicted Function | Strand | Amino Acids |
|---|---|---|---|---|
| lol1 | lpg2225 | GH3-family auxin responsive protein | - | 509 |
| lol2 | lpg2226 | Isovaleryl CoA dehydrogenase | - | 563 |
| lol3 | lpg2227 | Propionyl-CoA carboxylase | - | 479 |
| lol4 | lpg2228 | 3-oxoacyl-(acyl carrier protein) synthase III | + | 353 |
| lol5 | lpg2229 | Acyl CoA synthetase | + | 581 |
| lol6 | lpg2230 | Acyl CoA ligase | + | 464 |
| lol7 | lpg2231 | 3-oxoacyl reductase | + | 250 |
| lol8 | lpg2232 | 3-oxoacyl-(acyl carrier protein) synthase III | + | 336 |
| lol9 | lpg2233 | Acyl carrier protein | - | 75 |
| lol10 | lpg2234 | Major facilitator superfamily efflux pump | - | 455 |

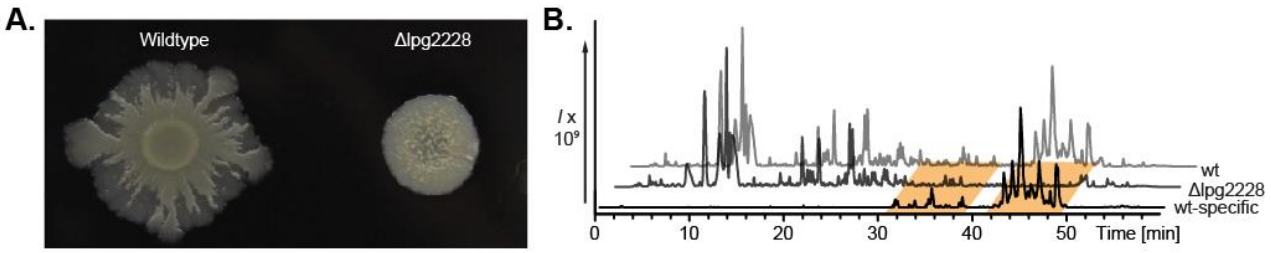



**Figure 4.3** Mutations in a hybrid polyketide gene cluster result in motility and metabolomic alterations. **A**. Extended growth of *L. pneumophila* LP02 on 0.5% agar plates results in pronounced sliding motility, which is absent in the Δlpg2228 strain. **B**. Comparative metabolomic analysis of wild type and Δlpg2228 cultures with LCMS highlights a series of molecules which are absent in Δlpg2228.

Close analysis of our metabolites detected by comparative metabolomics and PCA indicated two complexes of natural products (Figure 4.4A), likely corresponding to a series of core metabolites and acylated variants as indicated by an increased retention time and

mass, but highly similar MS/MS fragmentation (Figure S3). Although each unacylated compound was generally present as a single predominant peak for each mass, each acylated molecule was present as three to five individual peaks, indicating considerable heterogeneity. In hopes of obtaining a pure compound for NMR structure elucidation, we chose to isolate the most abundant core scaffold metabolite (459.2 $[M+H]^+$) through serial rounds of LCMS purification from 100 L of wild type LP02 culture. As a further demonstration of the remarkably low abundance of this molecule, preparative scale isolation yielded <500 μg of this metabolite. The structure of this pure compound was then elucidated by a combination of 1D and 2D NMR experiments and high resolution mass spectrometry, which provided a definitive molecular formula of $C_{23}H_{42}N_2O_7$ (0.327 Δppm; *see Material and Methods*). $^1$H, $^1$H-correlation spectroscopy (COSY) identified a short chain acyl unit, a number of secondary alcohols, one primary alcohol, and a system of four aromatic protons. Importantly, this also revealed that a prominent 133 m/z $[M+H]^+$ fragment observed during MS/MS was a 2,5-diaminopentane-1,3,4-triol. $^{13}$C-amino acid feeding experiments demonstrated that this hydroxylated diaminopentane was derived from ornithine (Figure S4). $^1$H, $^{13}$C-heteronuclear multiple-bond correlation spectroscopy (HMBC) revealed a gem-di-methyl group between the phenol and primary alcohol, and was used to link pieces identified with COSY, leading to a highly unusual molecule that we named legionellol A (Figure 4.4B; Supplementary Note – Structure characterization). Reanalysis of the legionellol series of metabolites identified through PCA indicated at least 10 distinct chemical entities, including smaller, more hydrophilic legionellol variants which appear to have shorter acyl tails or lack the gem-di-methyl. In addition, MS/MS revealed

that a minor series of legionellol variants likely possess an ornithine to arginine substitution in the modified amino acid portion of the molecule, and likewise possess a similar series of structural variations. Importantly, acylated forms of legionellol could also be observed, and MS/MS indicates that these are modified with a hydroxy-fatty acid, primarily $C_{12}$ and $C_{14}$ (Figure S3). To determine which molecular species was responsible for sliding motility, a 10 μg sample of an acyl legionellol or legionellol A was dried onto a 0.5% agar plate (as well as methanol alone), and the Δlpg2228 strain was deposited on top. The results of the corresponding outgrowth indicate that sliding motility was restored more effectively by supplementation with the acylated legionellol, rather than the scaffold alone, indicating that the variably acylated legionellol complex is likely the active surfactant (Figure 4.4C).

Legionellol is a novel natural product which arises from a minimal gene cluster composed of discrete domains often associated with polyketide biosynthesis (Figure 4.2; Table 4.1). From this gene cluster, enzymes can be identified which appear responsible fatty acid adenylation and CoA-ligation, and for the tethering, condensation, reduction, and dehydration of polyketide extending units such as malonate. However, genes required for several biosynthetic processes – including chain release and generation of the modified amino acid – are notably absent, suggesting that the enzymes required for these processes are located elsewhere in the genome, or that these functions are performed by new classes of biosynthetic enzymes. To assess whether unusual enzymes may account for some absent biosynthetic processes, we investigated an unusual GH3 family protein that was present in one of the two operons in the legionellol gene cluster, and which we suspected may be involved with attaching the modified ornithine residue. GH3 family proteins are known

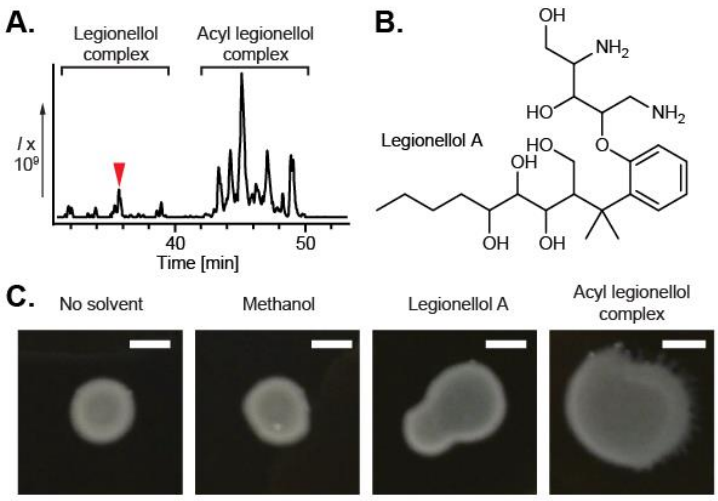



**Figure 4.4** An unusual PKS gene cluster in *L. pneumophila* encodes for legionellol, a novel surfactant scaffold. **A**. Metabolites absent from Δlpg2228 appear to separate as two complexes by LCMS, comprised of a smaller hydrophilic series of metabolites, and a larger hydrophobic series of metabolites. The most abundant of these hydrophilic scaffold molecules – legionellol A – is indicated with an arrow. **B**. Structure of the novel *L. pneumophila* surfactant legionellol A as deduced by NMR and MS experiments. **C**. To assess the impact of legionellol metabolites on sliding motility, 10 μL of Δlpg2228 overnight culture was added to a 0.5% agar BCYE plate alone, or over top of dried 10 μL drops of methanol, 1 μg/μL acyl legionellol (684 Da), or 1 μg/μL legionellol A. Plates were grown for one week at 30°C, and demonstrate that acyl legionellol is able to recapitulate sliding motility. Scale bars are equal to 5 mm.

from plants, where they act as indole-3-acetic acid amido-synthetases during auxin biosynthesis [36]. More specifically, these enzymes use ATP to adenylate a free carboxylic acid before catalysing the displacement of the adenosine monophosphate by a nucleophilic

amine from a free amino acid. While this mechanism is common in plants it has yet to be implicated in the biosynthesis of bacterial secondary metabolites, despite occasionally occurring in PKS biosynthetic gene clusters [37-38]. To assess whether this conspicuous enzyme was responsible for coupling the modified ornithine to the legionellol polyketide precursor, we generated an insertionally inactivated mutant using the same approach we had demonstrated previously. Despite its presence in the legionellol biosynthetic operon, insertional inactivation of this unusual enzyme did not alter legionellol production, although it remains to be seen whether other unusual biosynthetic enzymes may play a role in the construction of this novel natural product.

### 4.5 Discussion

Natural products have provided chemical scaffolds and molecular innovation that has driven drug discovery and development efforts for the last century [1]. Despite their inherent value, rediscovery of known natural products has led to declining discovery rates and a loss in industrial interest [6]. By pursuing challenging bacteria that had not been previously studied as natural products producers, we reasoned that we could identify new, exotic chemistry arising from novel biosynthetic gene clusters. In this work, we used bioinformatics and advanced natural products chemistry techniques to investigate *Legionella* as a source of novel natural product scaffolds, yielding the new molecule legionellol A.

*Legionella* are a diverse genus, and relatively limited genome sequencing has revealed species with significant biosynthetic potential. In this work, we chose to use the

genetically-tractable *L. pneumophila* strain LP02 to pursue a knockout and comparative metabolomics approach to discover new molecules. However, the *L. pneumophila* complex is not the most impressive candidate for natural product discovery, as species related to *L. dumoffii* and *L. longbeachae* typically possess at least twice the number of biosynthetic gene clusters present in *L. pneumophila*. Lone representative genomes of more exotic *Legionella spp*. such as *L. micdadei* and *L. shakespearei* also indicate that more extensive sequencing will reveal substantial numbers of new biosynthetic gene clusters from these unexplored organisms, as both species possess large and unique hybrid assembly lines. Using PRISM, we can now rapidly profile emergent genomes from this promising genus, identify biosynthetic gene clusters automatically, and prioritize candidates for discovery in the hopes of continuing to reveal interesting new natural products.

Although the rules for polyketide and nonribosomal peptide biosynthesis are well defined, exotic organisms often deviate from established systems [16, 18]. Legioliulin, the only example of a polyketide (or nonribosomal peptide) known from *Legionella* prior to this study, highlights several biosynthetic oddities that frequently occur in *Legionella* biosynthetic gene clusters, including a propensity for trans-AT PKS logic that does not include an associated acyl transferase [26]. This situation is more bizarre in *L. pneumophila*, which has three hybrid polyketide and nonribosomal peptide gene clusters, but only has a single malonyl-acyltransferase (lpg1394), which is associated with the fatty acid synthase operon, suggesting that there may be an alternative monomer loading strategy to create polyketide metabolites such as legionellol. *Legionella* also possesses a number of unusual chain termination strategies, including the use of C-terminal condensation domains

to facilitate trans-esterification and thioester cleavage, which had only been reported previously in a handful of bacterial natural products, including the Myxobacterial natural product crocacin [30], FK520 [28], and the C-1027 enediyne [29]. The frequency of these unusual biosynthetic domains, combined with the unorganized lone-domain structure of *Legionella*'s biosynthetic gene clusters limits the predictability of their polyketide and nonribosomal peptide products. This is well illustrated by legionellol, which is a remarkably complex natural product that arises from a seemingly simple biosynthetic gene cluster. Given its unusual structure and minimal biosynthetic gene cluster, the precise biosynthetic pathway of legionellol is still unclear, but the origins of some units can be speculated. Generation of the modified ornithine residue would require hydroxylation of the ornithine β- and γ-carbons, followed by activation of the carboxylic acid and iterative reduction to the observed alcohol. How the putative polyketide portion of legionellol is constructed is much less clear, and must include a number of exotic monomers and transformations. Construction of the polyketide likely begins with activation of a short-chain fatty acid of varying length, giving rise to the observed range of legionellol variants. Chain elongation with glycerate-derived hydroxymalonate followed by ketoreduction could then result in the hydroxylated acyl chain. Beyond this however, biosynthetic processes involved in constructing legionellol or its monomers become difficult to predict, such as the origins of the unusual primary alcohol and neighboring gem-dimethyl. Hopefully, further study of this promising genus may reveal innate biosynthetic logic involved in creating both legionellol and other unique natural products, facilitating prediction and discovery of new chemical scaffolds.

## 4.6 Materials and Methods

### 4.6.1 General experimental procedures

One-dimensional ($^{1}H$,$^{13}C$) and two-dimensional ($^{1}H$-$^{1}H$ and $^{13}C$-$^{1}H$ HMBC, HMQC, and COSY) NMR spectra were recorded on a Bruker AVIII 700 MHz NMR spectrometer in deuterated dimethyl sulfoxide (Cambridge Isotope Laboratories). High-resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (ThermoFisher Scientific, USA) with an electrospray ionization source (ESI). LCMS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna $C_{18}$ column (250 mm × 4.6 mm, Phenomenex) for analytical separations, running acetonitrile and dd$H_2O$ with 0.1% formic acid as the mobile phase.

### 4.6.2 Strains and culture conditions

All experiments and genetic manipulations were conducted using the genetically amenable *Legionella pneumophila* Philadelphia-1 variant LP02. Mutants were constructed in LP02 with or without a pBH6119 plasmid enabling GFP-expression through an upstream *icmR* promoter, maintained through complementation of LP02's natural thymidine auxotrophy [23]. *L. pneumophila* was grown at 37°C for all liquid cultures in either BYE (10 g/L ACES, 10 g/L yeast extract, 1 g/L monosodium α-ketoglutarate, 0.4 g/L L-cysteine, 0.25 g/L ferric pyrophosphate, 0.1 g/L thymidine, pH = 6.9) or chemically defined *Legionella* media (350 mg/L L-arginine, 510 mg/L L-aspartic acid, 400 mg/L L-cysteine, 600 mg/L L-glutamic acid, 150 mg/L L-histidine, 470 mg/L L-isoleucine, 640 mg/L L-leucine, 650 mg/L L-

lysine, 200 mg/L L-methionine, 350 mg/L L-phenylalanine, 115 mg/L L-proline, 650 mg/L L-serine, 330 mg/L L-threonine, 100 mg/L L-tryptophan, 400 mg/L L-tyrosine, 480 mg/L L-valine, 315 mg/L ammonium chloride, 50 mg/L sodium chloride, 20 mg/L calcium chloride, 1.18 g/L potassium phosphate monobasic, 70 mg/L magnesium sulfate, 250 mg/L ferric pyrophosphate, 100 mg/L thymidine, 10 g/L ACES). When visualizing sliding motility, plates were incubated at 30°C for roughly three weeks. All media were supplemented with 0.1 g/L thymidine to support the auxotrophy of LP02.

### 4.6.3 Comparative metabolomic analysis

To generate samples for LCMS analysis wild type, Δlpg1939, Δlpg2186, and Δlpg2228 LP02 strains were grown in 50 mL of chemically defined *Legionella* media at 37°C for one week. Following this, cultures were harvested by centrifugation, pellets were extracted with methanol, and supernatants were extracted with 20 g/L HP20 resin. Extracts were pooled and were subsequently dried by rotary evaporation and resuspended in methanol (2 mL). Samples were processed by LCMS with a 25 cm Luna $C_{18}$ column (250 mm × 4.6 mm), using water and acetonitrile with 0.1% formic acid as the mobile phase. Acetonitrile was held at 2% for the first 2 min, then steadily ramped to 100% by 45 min, held until 53 min, then reset to 2% and held until 60 min, at a flow rate of 1.2 mL/min. Principal component analysis of *Legionella* extracts was carried out using Bruker Daltonics Profile Analysis with the following parameters: Rt range: 3–58 min; mass range: *m/z* 200–1200; rectangular bucketing: 10 sec (Δ*m/z* of 2); normalized by using the sum of bucket values in the analysis. Chromatogram subtractions were performed using Bruker Daltonics MetaboliteDetect

software using the eXpose mode to reveal differences in excess of 5-fold, with Δ*m/z* of 0.5 and Δt of 0.5 min.

### 4.6.4 Isolation and purification of legionellol A

Wild type LP02 colonies from BCYE plates were inoculated into BYE cultures (5 mL) in sterile 50 mL Falcon tubes and grown for two days at 250 rpm and 37°C. These cultures were used to inoculate sterile 2.8 L Fernbach flasks containing BYE (1.5 L). Cultures were grown at 37°C with shaking at 200 rpm for roughly one week or until two days after peak melanin production. Following growth, cells were pelleted by centrifugation at 6000 rpm for 30 min. Supernatants were mixed with 20 g/L washed HP20 resin (Diaion) for 2 h at room temperature. Resins were harvested using Buchner funnel vacuum filtration, and washed with 10% methanol to remove highly polar melanins. Resin was eluted with excess 100% methanol which was then dried by rotary vacuum. Extracts were resuspended in methanol and separated by LCMS using a Luna $C_{18}$ column (250 mm × 10 mm) with HPLC grade water and acetonitrile with 0.1% formic acid as the mobile phase. To purify legionellol, acetonitrile began at 5% for the first 2 min, then ramped to 30% by 5 min and held until 27 min, followed by a shallow ramp to 45% by 40 min, followed by a wash of 100% from 42 to 52 min. Flow was maintained at 6 mL/min, and legionellol A eluted at 33 min.

### 4.6.5 Incorporation of $^{13}$C ornithine

To assess the origins of the modified amino acid present in the legionellol metabolites,

cultures of *L. pneumophila* in chemically defined media were grown for five days in the presence of $^{13}C$ ornithine (2 mM; Cambridge Isotope Laboratories) or $^{12}C$ ornithine (2 mM). Supernatants of cultures containing $^{12}C$ and $^{13}C$ ornithine were extracted with HP-20 resin and analyzed by LCMS.

### 4.6.6 Reconstitution of sliding motility

To assess whether various legionellol species were responsible for sliding motility, legionellol A or acylated legionellol (10 μg) was dissolved in methanol (10 μL) and added as a drop to a 0.5% agar plate of BCYE media. After the drops of legionellol, acyl legionellol, or methanol alone had dried, an overnight culture of LP02 Δlpg2228 was dropped (10 μL) over top of it and subsequently dried. Plates were incubated at 28°C for one week before imaging.

### 4.6.7 Insertional inactivation of genes in *Legionella*

Targeted insertional inactivation of biosynthetic genes in LP02 was performed as previously described [24], and will be summarized below. All primers and plasmids used in this process are described in Supplementary Table 3. If not stated explicitly, genetic manipulations and molecular biology techniques followed those from Cold Spring Harbor Protocols, available at http://www. molecularcloning.com/.

In short, two neighbouring 500 bp to 1000 bp fragments of target genes were cloned into pBlueScript KSII or a modified pBlueScript KSII bearing a chloramphenicol resistance cassette, swapped with the ampicillin resistance cassette through two flanking BspHI sites.

Plasmids were digested with Xba1 and Sac1, and gene fragments were digested with Xba1 and Kpn1, or with Sph1 and Sac1. Chloramphenicol or kanamycin resistance cassettes were amplified with flanking Kpn1 and Sph1 cut sites from pRE112 or pET-28b respectively. The digested plasmid, two digested gene fragments, and resistance cassette were ligated simultaneously and transformed into chemically competent DH5α (Invitrogen). All plasmids were confirmed by sequencing from either end of the homology arms and resistance cassettes. To transform into *Legionella*, fresh colonies of LP02 were re-streaked as a dime-sized patch on fresh BCYE and 10 μL of 100 ng/μL knockout vector solution was added on top, followed by incubation at 30°C for 48 h. Following incubation, colonies were re-streaked onto BCYE with kanamycin or chloramphenicol to select for transformants. Genomic integration was confirmed through PCR with primers for the resistance cassette and a region of the genome just outside the homology arm. PCR with primers for amplifying the original gene fragment was used to verify the exclusive presence of the insertionally inactivated allele.

### 4.6.8 High resolution mass spectrometry

Legionellol A was dissolved in a mixture of HPLC grade methanol and water with 0.1% formic acid providing a final concentration of roughly 10 μg/mL. The sample was infused directly into a Thermo LTQ OrbiTrap XL mass spectrometer running Xcaliber 2.07 and TunePlus 2.4 SP1 at a flow rate of 3 μL/min and ionized using an electrospray ionization source. The mass spectrometer was operated in positive mode with a maximum resolution of 100,000. The high resolution mass for legionellol A was an average of 29 scans.

| *Compound* | *Molecular Formula* | *Calculated m/z* | *Observed m/z* | *Δppm* |
|---|---|---|---|---|
| Legionellol A $[M+H]^+$ | $C_{23}H_{43}N_2O_7$ | 459.30650 | 459.30665 | 0.327 ppm |

### 4.6.9 PRISM analysis of *Legionella* genomes

*Legionella* genomes were downloaded from NCBI and loaded into PRISM [12] (http://www.magarveylab.ca/prism) using standard settings. Of the assembled genomes used in this analysis, 24 were draft genomes and 10 were fully assembled. Polyketide and nonribosomal peptide gene clusters were stored, annotated, and confirmed by secondary analysis using the BLAST function of Integrated Microbial Genomes (IMG; http://img.jgi.doe.gov/), which also checked for fragmentation resulting from incomplete genome assemblies. A phylogenetic tree of *Legionella spp*. was generated with 16S rRNA sequences collected from *Legionella* genomes, using the Geneious tree builder program with Tamura-Nei as the genetic distance model and neighbor-joining as the tree build method.

## 4.7 Supplementary Information

Supplementary information from the publication referred to in this chapter can be found in Appendix 3.

## 4.8 References

[1]. D.J. Newman, G.M. Cragg, Natural products as a source of new drugs over the 30 years from 1981 to 2010. J Nat Prod. 75 (2012) 311–335.

[2]. Hopwood DA. *Streptomyces* in Nature and Medicine: The Antibiotic Markers.NY, USA: Oxford University Press; 2007.

[3]. S.A. Cochrane, J.C. Vederas, Lipopeptides from *Bacillus* and *Paenibacillus spp*.: A Gold Mine of Antibiotic Candidates. Med Res Rev. (2014) 1-28.

[4]. K.J. Weissman, R. Müller, A brief tour of myxobacterial secondary metabolism. Bioorg Med Chem. 17 (2009) 2121-2136.

[5]. J.K. Nunnery, E. Meyers, W.H. Gerwick, Biologically active secondary metabolites from marine cyanobacteria. Curr Opin Biotechnol. 21 (2010) 787-793.

[6]. F.E. Koehn, G.T. Carter, The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 4 (2005) 206-220.

[7]. S. Donadio, P. Monciardini, M. Sosio, Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. 24 (2007) 1073-1109.

[8]. M. Nett, H. Ikeda, B.S. Moore, Genomic basis for natural product biosynthetic diversity in the actinomycetes. 26 (2009) 1362-1384.

[9]. M.A. Skinnider, C.W. Johnston, R. Zvanych, N.A. Magarvey, Automated identification of depsipeptide natural products by an informatic search algorithm. Chembiochem 16 (2015) 223-227.

[10]. M.S. Rappé, S.J. Giovannoni, The uncultured microbial majority. Annu Rev Microbiol. 57 (2003) 369-394.

[11]. C.W. Johnston, M.A. Skinnider, M.A. Wyatt, X. Li, M.R. Ranieri, L. Yang, et al., An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. Nat Commun. 6 (2015) 8421.

[12]. M.A. Skinnider, C.A. Dejong, P.N. Rees, C.W. Johnston, H. Li, A.L. Webster, et al., Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. (2015) 1-18.

[13]. H. Mohimani, W.T. Liu, R.D. Kersten, B.S. Moore, P.C. Dorrestein, P.A. Pevzner, NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. J Nat Prod. 77 (2014) 1902-1909.

[14]. T. Weber, K. Blin, S. Duddela, D. Krug, H.U. Kim, R. Bruccoleri, et al., antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 43 (2015) W237-W243.

[15]. M.A. Fischbach, C.T. Walsh, Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev. 106 (2006) 3468-3496.

[16]. T. Lincke, S. Behnken, K. Ishida, M. Roth, C. Hertweck, Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium *Clostridium cellulolyticum*. 49 (2010) 2011-2013.

[17]. L.L. Ling, T. Schneider, A.J. Peoples, A.L. Spoering, I. Engels, B.P. Conlon, et al., A new antibiotic kills pathogens without detectable resistance. Nature 517 (2015) 455-459.

[18]. M.C. Wilson, T. Mori, C. Rückert, A.R. Uria, M.J. Helf, K. Takada, et al., An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506 (2014) 58-62.

[19]. J. Amemura-Maekawa, Y. Hayakawa, H. Sugie, A. Moribayashi, F. Kura, B. Chang, et al. Legioliulin, a new isocoumarin compound responsible for blue-white autofluorescence in *Legionella* (*Fluoribacter*) *dumoffii* under long-wavelength UV light. Biochem Biophys Res Commun. 323 (2004) 954-959.

[20]. A. Hubber, C.R. Roy, Modulation of Host Cell Function by *Legionella pneumophila* Type IV Effectors. Annu. Rev. Cell Dev. Biol. 26 (2010) 261-283.

[21]. J.C. Feeley, R.J. Gibson, G.W. Gorman, N.C. Langford, J.K. Rasheed, D.C. Mackel, et al., Charcoal-yeast extract agar: primary isolation medium for *Legionella pneumophila*. J Clin Microbiol. 10 (1979) 437-441.

[22]. W.J. Warren, R.D Miller. Growth of Legionnaires disease bacterium (*Legionella pneumophila*) in chemically defined medium. J Clin Microbiol. 10 (1979) 50-55.

[23]. B.K. Hammer, M.S. Swanson. Co-ordination of *Legionella pneumophila* virulence with entry into stationary phase by ppGpp. Mol. Microbiol. 33 (1999) 721–731.

[24]. J.A. Sexton, J.P. Vogel. Regulation of hypercompetence in *Legionella pneumophila*. J. Bacteriol. 186 (2004) 3814-3825.

[25]. J. Piel. Biosynthesis of polyketides by trans-AT polyketide synthases. Nat Prod Rep. 27 (2010) 996-1047.

[26]. T. Ahrendt, M. Miltenberger, I. Haneburger, F. Kirchner, M. Kronenwerth, A.O. Brachmann, et al., Biosynthesis of the natural fluorophore legioliulin from legionella. Chembiochem 14 (2013) 1415-1418.

[27]. C. Reimmann, H.M. Patel, L. Serino, M. Barone, C.T. Walsh, D. Haas, Essential PchG-dependent reduction in pyochelin biosynthesis of *Pseudomonas aeruginosa*. J Bacteriol. 183 (2001) 813-820.

[28]. G.J. Jr Gatto, S.M. McLoughlin, N.L. Kelleher, C.T. Walsh, Elucidating the substrate specificity and condensation domain activity of FkbP, the FK520 pipecolate incorporating enzyme. Biochemistry 44 (2005) 5993-6002.

[29]. S. Lin, S.G. Van Lanen, B. Shen, A free-standing condensation enzyme catalyzing ester bond formation in C-1027 biosynthesis. Proc. Natl. Acad. Sci. USA. 106 (2009) 4183-4188.

[30]. S. Müller, S. Rachid, T. Hoffmann, F. Surup, C. Volz, N. Zaburannyi, et al., Biosynthesis of crocacin involves an unusual hydrolytic release domain showing similarity to condensation domains. Chem Biol. 21 (2014) 855-865.

[31]. C. Cazalet, C. Rusniok, H. Brüggemann, N. Zidane, A. Magnier, L. Ma, et al., Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet. 36 (2004) 1165-1173.

[32]. T.J. O'Connor, Y. Adepoju, D. Boyd, R.R. Isberg, Minimization of the *Legionella pneumophila* genome reveals chromosomal regions involved in host range expansion. Proc. Natl. Acad. Sci. USA. 108 (2011) 14733-14740.

[33]. T. Hindré, H. Brüggermann, C. Buchrieser, Y. Héchard. Transcriptional profiling of *Legionella pneumophila* biofilm cells and the influence of iron on biofilm formation. Microbiology 154 (2008) 30-41.

[34]. Z.D. Dalebroux, B.F. Yagi, T. Sahr, C. Buchrieser, M.S. Swanson, Distinct roles of ppGpp and DskA in *Legionella pneumophila* differentiation. Mol. Microbiol. 76 (2010) 200-219.

[35]. C.R. Stewart, D.M. Burnside, N.P. Cianciotto, The surfactant of *Legionella pneumophila* is secreted in a TolC-dependent manner and is antagonistic toward other Legionella species. J Bacteriol. 193 (2011) 5971-5984.

[36]. Q. Chen, C.S. Westfall, L.M. Hicks, S. Wang, J.M. Jez, Modulating plant hormones by enzyme action: the GH3 family of acyl acid amido synthetases. J. Biol. Chem. 285 (2010) 29780–29786.

[37]. C. Olano, B. Wilkinson, C. Sánchez, S.J. Moss, R. Sheridan, V. Math, et al., Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: cluster analysis and assignment of functions. Chem. Biol. 11 (2004) 87-97.

[38]. R. Kong, X. Liu, C. Su, C. Ma, R. Qiu, L. Tang, Elucidation of the biosynthetic gene cluster and the post-PKS modification mechanism for fostriecin in *Streptomyces pulveraceus*. Chem. Biol. 20 (2013) 45-54.

# Chapter 5. Gold Biomineralization by a Metallophore from a Gold-Associated Microbe

## 5.1 Chapter Preface

Microbial natural products are often referred to as secondary metabolites, in that their functions are non-essential (in contrast to primary metabolites), and serve to improve fitness in specific environments. In keeping with this, highly competitive environments like soil had been a well-spring of antimicrobial natural products which are believed to assist slow-growing producer organisms such as *Streptomyces* to inhibit the growth of neighbouring microbes and thus, thrive. The rise of inexpensive bacterial genome sequencing has revealed that there are a large number of unexpected secondary metabolite producers associated with a diverse range of environmental niches, where their metabolic versatility may play an integral role in their survival. Now, with modern bio- and chemo-informatic tools, these molecules can be predicted, identified, and interrogated to reveal new biological insight. In this work, I investigated the gold-associated organism *Delftia acidovorans*, which is one of the few organisms capable of living on gold deposits. Initial in vitro studies indicated that *D. acidovorans* secreted an agent that reduced toxic, ionic gold to the inert metal, indicating a potential mechanism by which *D. acidovorans* could survive on toxic gold deposits. *D. acidovorans* was found to possess a large NRPS-PKS gene cluster, and prediction-guided investigations revealed a novel molecule delftibactin as the secreted agent responsible for gold biomineralization. Further study revealed that delftibactin had a protective effect against exposure to Au (III) which could be verified at concentrations observed during in vitro culture. After defining the metal binding site of

delftibactin and partially-characterizing the events that transpire following gold-binding, transmission electron microscopy studies revealed that delftibactin-mediated gold reduction resulted in the generation of gold nanoparticles similar to those observed in secondary gold deposits (or nuggets), providing a new hypothesis for how these poorly understood deposits are created.

The following chapter is a modified version of previously published journal article in which I was a lead author. I isolated delftibactins, constructed the genetically modified strain, performed gold experiments and contributed to study design. Morgan A. Wyatt isolated delftibactins, performed gold experiments and contributed to study design. Xiang Li performed structural analysis. Ashraf Ibrahim performed structural and MS/MS analysis and isolated delftibactin B. Jeremiah Shuster obtained transmission electron microscopy (TEM) images. Gordon Southam performed TEM image analysis. Nathan A. Magarvey contributed to study design and wrote the manuscript. The citation for this publication is as follows:

**Johnston, C.W.***, Wyatt, M.A.*, Li, X., Ibrahim, A., Shuster, J., Southam, G., & Magarvey, N.A. (2013) Gold biomineralization by a metallophore from a gold-associated microbe. *Nature Chemical Biology* **9**, 241-243.

## 5.2 Abstract

Microorganisms produce and secrete secondary metabolites to assist in their survival. We report that the gold resident bacterium *Delftia acidovorans* produces a secondary metabolite that protects from soluble gold through the generation of solid gold

forms. This finding is the first demonstration that a secreted metabolite can protect against toxic gold and cause gold biomineralization.

## 5.3 Introduction

Microorganisms inhabit nearly all surfaces on the planet, an achievement typically attributed to their metabolic versatility. Frequently, secondary metabolic pathways and secreted products of these specialized branches of metabolism are complicit in an organisms' ability to capture niches, enhance fitness and overcome environmental stress and often have considerable industrial importance[1]. Metals represent a notable environmental condition for microbes, as some are required for growth (for example, $Fe^{3+}$), whereas others inhibit it (for example, $Au^{3+}$, $Ag^{+}$ and $Hg^{2+}$)[2]. Bacterial biofilms exist on the surface of gold nuggets[3, 4]; though soluble gold is inherently toxic[2], these bacteria are implicated in its accumulation and deposition[5, 6]. The existence of bacterial biofilms coating gold nuggets and the discovery of bacterioform gold suggest that bacteria and specialized bacterial metabolic processes are involved in gold biomineralization[3, 4, 5, 6]. Sequencing gold nugget microbiota has revealed that *Cupriavidus metallidurans* and *Delftia acidovorans* are dominant organisms within such communities and comprise over 90% of these populations[4]. Investigations into *C. metallidurans* have revealed that it bioaccumulates inert gold nanoparticles within its cytoplasm as a mechanism to protect itself from soluble gold[5].

## 5.4 Results & Discussion

We sought to test whether *D. acidovorans* has any appreciable differences with *C. metallidurans* with respect to such gold biomineralization. An assay was developed to define whether the mechanism of biomineralization was extracellular or intracellular and to reveal mechanisms of how *D. acidovorans* protects itself from toxic soluble $Au^{3+}$ and how these interactions may relate to gold deposition. We reasoned that if a cell-associated mechanism was predominant or exclusive, the bacteria would accumulate insoluble gold particles[7]; in contrast, if an extracellular gold reduction occurs at the cell surface or within the area surrounding microbial colonies, blackening would result because of gold reduction and the creation of solid gold particles. *D. acidovorans* and *C. metallidurans* were grown on agar plates and then flooded with solutions of Au(III), the dominant form of soluble gold found in terrestrial conditions[5, 8]. Following gold exposure, darkened zones developed around colonies of *D. acidovorans* but not *C. metallidurans* (Fig. 5.1a). These blackened zones suggested that *D. acidovorans* generated a diffusible metabolite that acts to generate reduced solid gold forms.

We investigated the *D. acidovorans* genome for genes that may be associated with a unique small-molecule biosynthesis pathway that is absent from *C. metallidurans*. For example, polyketides and nonribosomal peptides are classes of secondary metabolites that bacteria use to promote environmental fitness[1] and include members that function to bind metals (for example, iron and copper)[9, 10, 11]. Indeed, our analysis identified a candidate nonribosomal peptide synthetase/polyketide synthase (NRPS/PKS) gene cluster (*Daci_4753*, *Daci_4754*, *Daci_4755*, *Daci_4756*, *Daci_4757*, *Daci_4758*, *Daci_4759*); henceforth referred to as the Daci_4753–4759 or del cluster for an unknown secondary

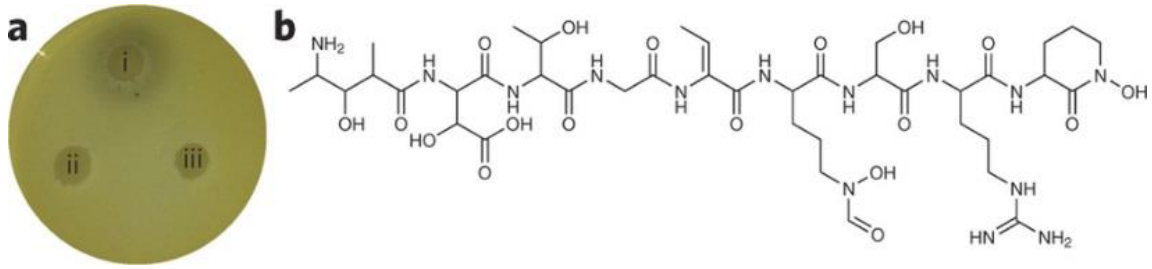



**Figure 5.1** (**a**) Gold resident bacteria *D. acidovorans* (wild type, i), the NRPS-null *D. acidovorans* mutant strain (Δ*delG*, ii) and *C. metallidurans* (iii) were grown for 3 d and overlaid with soft agarose containing 10 mM $AuCl_3$ for 2 h. Black halos are formations of gold nanoparticles. (**b**) Final structure of the gold-interacting nonribosomal peptide delftibactin.

metabolite that, according to bioinformatic analysis[12] and *in silico* predictions, was expected to be a polar peptidic small molecule (Supplementary Results, Supplementary Fig. 1a)). Upstream, flanking these biosynthetic genes is a tripartite heavy metal efflux pump[13, 14] (*Daci_4763*, *Daci_4764*, *Daci_4765*; 68% identical and 83% similar to the CzcA-like HmyA heavy metal efflux pump from *C. metallidurans* CH34 (*Rmet_4123*)), perhaps supporting a role of this cluster being associated with gold detoxification. Downstream genes were associated with metallophores that bind iron (siderophores) and, specifically, genes for their reception and regulation[9]. To reveal whether the Daci_4753–4759 (del) cluster (Supplementary Fig. 1a) was associated with the observed gold precipitation, we constructed an insertional inactivation of the nonribosomal peptide synthetase gene (*Daci_4754*; referred to here as *delG*), and the resulting mutant strain was compared to the wild-type in the soluble gold exposure agar plate assay (Fig. 5.1a

and Supplementary Fig. 1b). Unlike the wild-type colonies, colonies of the Δ*delG* strain were deficient in producing a blackening zone. To further reveal whether end products from this biosynthetic locus were solely responsible for the gold precipitation, we generated broth extracts of the entire *D. acidovorans* secreted metabolome, subjected the mixtures to chromatographic separations with LC/MS and eluted the separated contents into a 96-well plate. Within the water-soluble fractions, a select number of wells recapitulated the gold activity (Supplementary Fig. 2a). Well fractions capable of gold precipitation were analyzed further and were found to share a peptidic compound that closely matched the molecular weight of the predicted del nonribosomal peptide, which was absent in extracts from the Δ*delG* strain (Supplementary Fig. 2b) that lack gold-precipitating metabolites. This peptidic compound could be identified in supernatants in concentrations in excess of 200 μM (Supplementary Table 6), enabling its isolation and structure determination by high-resolution MS (Supplementary Fig. 3a) and NMR spectroscopy (Supplementary Fig. 3b), which revealed a linear polyketide-nonribosomal peptide consistent with the structural prediction, which was named delftibactin (**1**; Fig. 5.1b). *D. acidovorans* environmental isolates were also screened for their ability to produce delftibactin, resulting in its identification in all tested isolates (Supplementary Fig. 4a,b).

Purified delftibactin was observed to co-precipitate with gold from solution, recapitulating the original findings in end-point assays (Supplementary Fig. 5a–c). We sought to address whether the gold precipitation caused by delftibactin confers a protective advantage to *D. acidovorans* and assists in ameliorating gold toxicity. In initial assays, this question was addressed with an acute toxic exposure of the wild-type and the Δ*delG* strains,

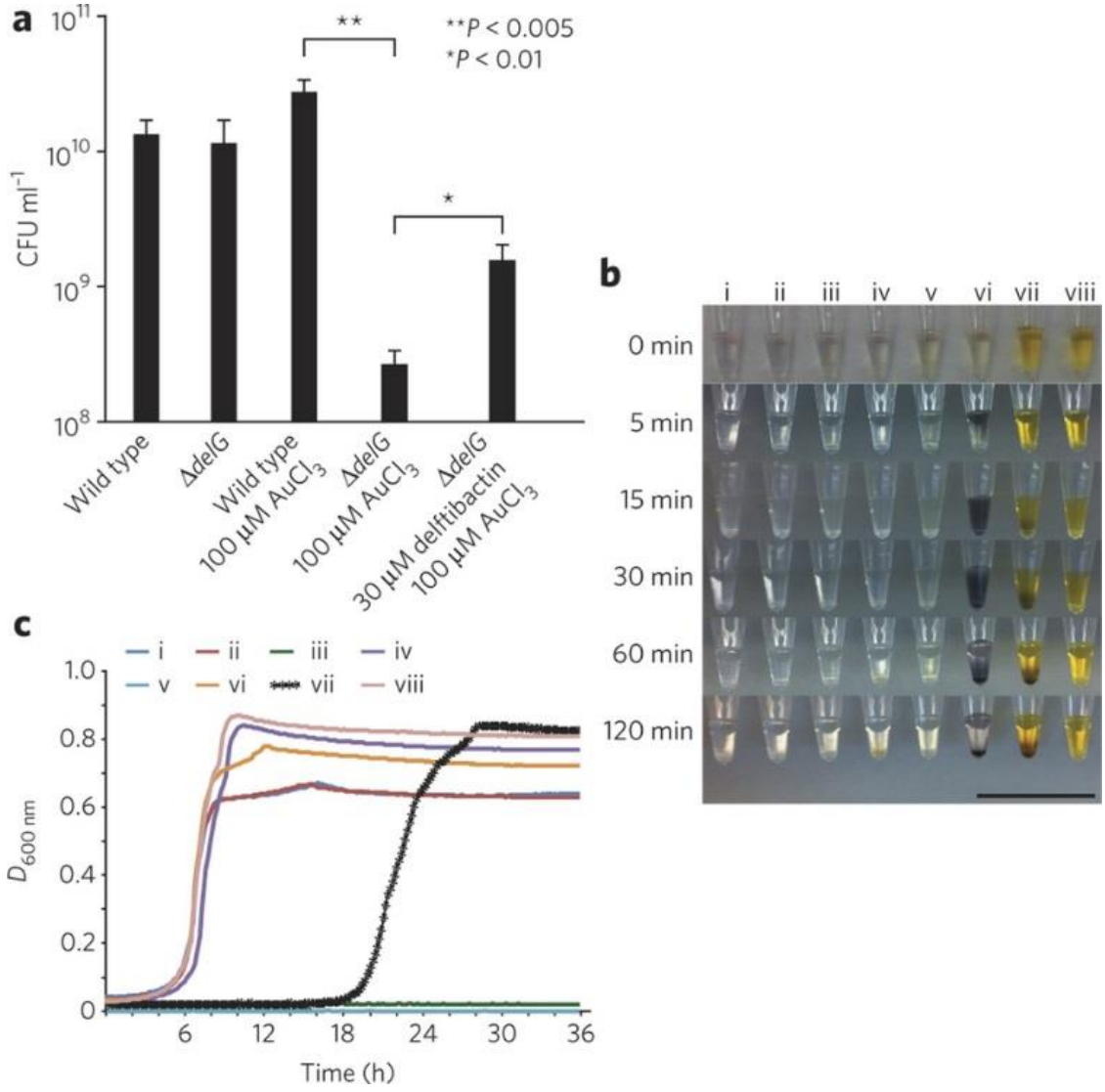



**Figure 5.2** (**a**) Delftibactin-null *D. acidovorans* shows increased sensitivity to gold toxicity and can be rescued by the addition of delftibactin. *D. acidovorans* wild-type and Δ*delG* cultures were grown for 48 h at 30 °C in deferrated Acidovorax complex medium (ACM). Cultures were incubated in the presence and absence of 100 μM$AuCl_3$ for 30 min, revealing increased sensitivity in the Δ*delG* strain. Addition of delftibactin (30 μM) to Δ*delG D. acidovorans* ameliorated gold toxicity. Results are shown as mean ± s.d.; $n = 4$; Two-tailed student's *t*-test. CFU, colony-forming units. (**b**) Delftibactin is capable of

precipitating gold in the presence of iron. Time course progression of 5 mM $AuCl_3$ reacted with: (i) water only, (ii) 5 mM delftibactin A, (iii) 5 mM $AuCl_3$, (iv) 5 mM $FeCl_3$, (v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$, (vi) 5 mM delftibactin A + 5 mM $AuCl_3$, (vii) delftibactin A + 5 mM $AuCl_3$ + 5 mM $FeCl_3$ and (viii) delftibactin A + 5 mM $FeCl_3$. Scale bar, 20 mm. (**c**) Growth curves of *D. acidovorans* Δ*delG* in the presence of each reaction mixture shown in **b** at a final concentration of 30 μM in ACM. Results are a mean of three growth curves for each condition from a single representative experiment.

whereby broth cultures of each were exposed for 30 min and the colony-forming units subsequently determined. The results of this assay showed that a 102.8-fold increase in sensitivity to gold toxicity could be observed in the Δ*delG* strain and that this increase could be rescued with exogenous addition of purified delftibactin (Fig. 5.2a). A detoxifying effect was observed in dose escalations of delftibactin to overtly toxic concentrations of $AuCl_3$ (Supplementary Fig. 6a), and subsequent examination revealed that although soluble gold is toxic at 10 μM, the blackened precipitate did not show any obvious toxicity when supplied in excess of 10 mM (data not shown). Metals found within secondary gold deposits have been outlined previously[4], specifically revealing that the concentration of iron is low relative to gold. However, we probed the fate of delftibactin when presented with equimolar concentrations of soluble gold and iron. This simultaneous exposure revealed that gold precipitation would proceed in the presence of high concentrations of iron (Fig. 5.2b). To assess what impacts this precipitation would have on *D. acidovorans* viability, we set up cultures of the Δ*delG* strain containing the resultant gold reactions. Growth curves

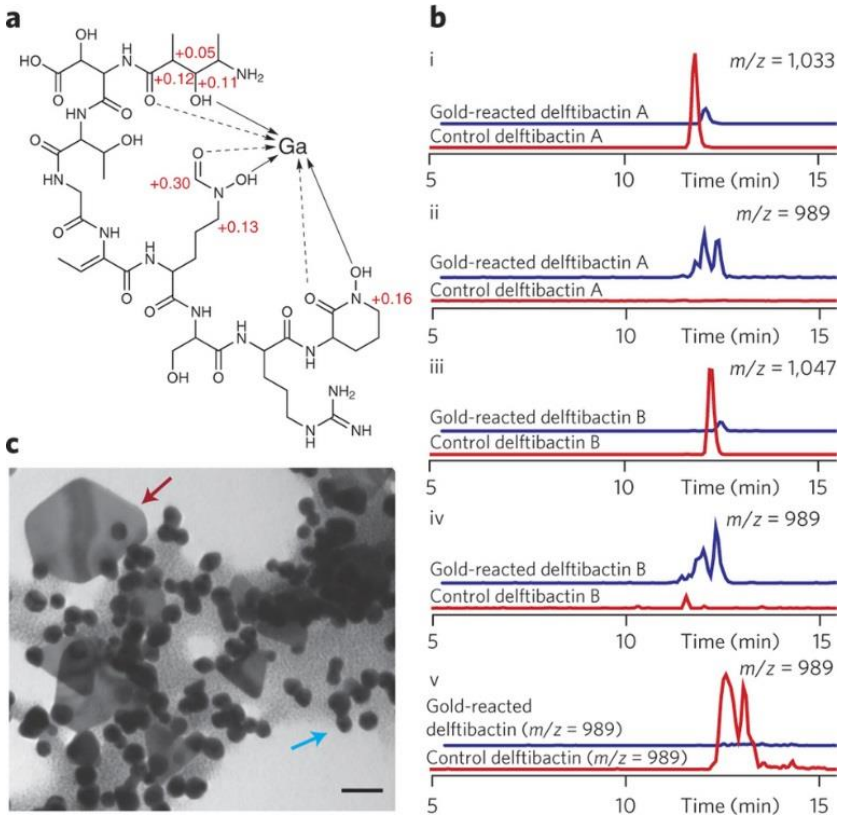



**Figure 5.3** (**a**) Gallium NMR confirms that delftibactin has a single metal-binding site. (**b**) Chelation core modification affects the rate of gold precipitation but does not affect the generation of a common transient intermediate. Delftibactin A and B were reacted with equimolar (5 mM) $AuCl_3$ for 2 h before reactions (blue) were analyzed by LC/MS and compared to an unreacted control (red). Extracted ion chromatograms show a depletion of delftibactin A (i) and B (iii) following incubation with $AuCl_3$, accompanied by the emergence of a new delftibactin species (*m*/*z* = 989; ii, iv) that could be further reacted with $AuCl_3$ and depleted from solution (v). (**c**) TEM of delftibactin–gold (2:1) complex after 10 min reveals the presence of colloidal and octahedral gold nanoparticles, reminiscent of those seen in natural deposits. Blue arrow, colloidal gold. Red arrow, octahedral gold. Scale bar, 50 nm.

demonstrate that although iron-free conditions are optimal for detoxification owing to metal competition, sufficient detoxification occurs in the presence of iron to support the growth of *D. acidovorans* (Fig. 5.2c). Chronic exposures were also tested, demonstrating that exogenous delftibactin addition to cultures of the Δ*delG* strain was sufficient to overcome chronic gold toxicity (Supplementary Fig. 6b). The results of these experiments, though not conducted in a natural context, may inform on the protective nature of delftibactin for *D. acidovorans*. We next performed an experiment aimed at revealing whether *D. acidovorans* may maintain protective extracellular concentrations of delftibactin by monitoring the loss of delftibactin by gold co-precipitation, leading to an increase in delftibactin production through a positive feedback mechanism[15]. *D. acidovorans* supernatants treated with 10 μM and 30 μM $AuCl_3$ caused delftibactin depletion, resulting in a compensatory increase in delftibactin concentrations compared to an untreated control, representing a form of reactive homeostasis (Supplementary Fig. 7). Delftibactin concentrations were also responsive to iron concentrations (Supplementary Table 6), indicating that delftibactin is most likely a siderophore that serves at least two purposes for this organism.

Metallophores are recognized to create complexes with metals, and we wished to reveal whether such complexation was part of the gold-delftibactin interaction. As the gold-delftibactin association leads to co-precipitation and formation of an insoluble material, we examined how delftibactin may bind metals using gallium. NMR analysis of the delftibactin–gallium complex showed the coordinating activity of delftibactin. These

results indicated that $N^{\delta}$-hydroxy-$N^{\delta}$-formylornithine, the polyketide-extended portion of the *N*-terminal alanine, and cyclic $N^{\delta}$-hydroxyornithine form ligands for metal binding (Fig. 5.3a). This complexation is relevant to gold, as purified gallium–delftibactin exposed to gold showed considerably decreased precipitation (Supplementary Fig. 8). Although the gallium-gold competition may inform on how gold interacts with delftibactin, we next sought direct evidence of gold binding by delftibactin. We identified an initial gold–delftibactin complex through MS and confirmed its identity through diagnostic MS/MS fragmentation (Supplementary Fig. 9). To reveal in more depth the mechanisms that lead to gold precipitation and which sites within delftibactin are associated with gold complexation, we made use of natural delftibactin variants. Several compounds were observed to elute at a similar time to delftibactin and had comparable fragmentation patterns and masses, indicating that they may be structural analogs that would be useful if they had modifications within the proposed chelation core. One promising candidate, bearing a hydroxylated and acetylated ornithine (delftibactin B (**2**) $m+/z$ = 1,047; Supplementary Fig. 10a), was identified and subsequently characterized. Subsequent examination of the complexing properties and protective nature of delftibactin B indicated that it was less efficient in gold reduction than delftibactin (renamed delftibactin A; Supplementary Fig. 10b), resulting in decreased detoxification (Supplementary Fig. 10c). Exposing delftibactin A and delftibactin B to $AuCl_3$ leads to their depletion; new peaks, however, emerge following the exposure, with the predominant one at $m+/z$ = 989 (Fig. 5.3b); the molecules at this molecular weight continue to react with gold and are also lost from solution (Fig. 5.3b and Supplementary Fig. 11). Structural characterization

indicates that this reaction product does not bear the ornithine modifications observed in delftibactin A and B (Supplementary Fig. 12), indicating that delftibactin can chelate gold and also react with it. This observation most likely explains why delftibactin B is less protective, as it has a ketone moiety that is less easily oxidized than the aldehyde found on delftibactin A, which may be partially responsible for gold reduction. Transmission electron microscopy was used to better assess the nature of the gold precipitate and revealed an abundance of colloidal gold nanoparticles and octahedral gold platelets (Fig. 5.3c). Such solid gold forms are authentic morphologies found in gold nuggets and bacterioform gold[6, 16]. These data show that pure delftibactin A is capable of creating naturally occurring complex gold structures from $Au^{3+}$ on short timescales (seconds) at room temperature and neutral pH and at rates that far exceed those observed for traditional gold nanoparticle–producing agents such as citrate[17](Supplementary Fig. 13), providing a potential mechanism for bacterial gold biomineralization. We propose that delftibactin facilitates this biomineralization and protects *D. acidovorans* by chelating soluble $Au^{3+}$ and directly precipitating it as a complex or by binding and reducing gold through oxidative decarboxylation before chelating a second $Au^{3+}$ ion and precipitating as a complex.

Collectively, these results indicate that although delftibactin is dispensable in culture, consistent with other secondary metabolites, it has an important role in protecting its gold-resident producer from toxic soluble gold. Further, we have shown that gold biomineralization can take place through the secretion of this metallophore from a gold resident bacterium. This phenomena echoes situations observed previously including boron chelation by vibroferrin[18] and copper chelation by yersiniabactin[11] and methanobactin[10],

wherein bacterial siderophores have dual physiological roles that are important in their environments. Delftibactin seems to be what is – to our knowledge – the first example of a co-opted metallophore that protects its producer from toxic soluble gold and provides a mechanism for bacterial gold biomineralization.

## 5.5 Materials and Methods

### 5.5.1 General experimental procedures.

One-dimensional ($^{1}H$ and $^{13}C$) and two-dimensional ($^{1}H$-$^{13}C$ and $^{1}H$-$^{15}N$ HMBC, HSQC, NOESY and COSY) NMR spectra were recorded on a Bruker AVIII 700 MHz NMR spectrometer in $D_2O$ ($D_2O$; Cambridge Isotope Laboratories). High-resolution MS spectra were collected on a Thermo LTQ OrbiTrap XL mass spectrometer (ThermoFisher Scientific, USA) with an electrospray ionization source (ESI) and using CID with helium for fragmentation. LC/MS data was collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna $C_{18}$ column (250 mm × 4.6 mm, Phenomenex) for analytical separations, running acetonitrile and $ddH_2O$ as the mobile phase.

### 5.5.2 Bacterial strains

*Delftia acidovorans* was ordered from the German Resource Centre for Biological Material (DSMZ, DSM no. 39). *Delftia acidovorans* was cultured on Acidovorax complex medium[19] (ACM) plates at 30 °C. The Δ*delG* strain was initially grown in the presence of 30 μg/mL tetracycline. Environmentally isolated *D. acidovorans* strains D27L and D126L

were found in soil samples collected around McMaster University from June–August 2010. Environmental isolates were identified as *D. acidovorans* strains based on 16S sequence alignment, using 16S sequences that were amplified from single colonies using the universal 16S primers[20]: 27f (AGAGTTTGATCMTGGCTCAG) and 1525r (AAGGAGGTGATCCAGCC).

### 5.5.3 Gold precipitation on agar plates.

Wild-type and Δ*delG D. acidovorans* were streaked onto a Chelex-treated (deferrated) ACM agar plate and grown for 3 d at 30°C. The plate was then overlaid with 10 mL of 0.5% agarose containing 10 mM $AuCl_3$. Gold complexing comparison to other bacteria was carried out as follows: 10 μL of an overnight culture of *D. acidovorans*, *D. acidovorans* Δ*delG* or *C. metallidurans* were placed onto deferrated ACM plates and grown for 3 d at 30 °C. The plates were overlaid with $AuCl_3$ as described above. Agar plate overlay images were taken after 2 h of incubation at room temperature.

### 5.5.4 *D. acidovorans* 96-well plate gold bioassay.

After brief centrifugation to remove particulates, 100 μL of the *D. acidovorans* HP20 extract was loaded onto a Waters Alliance 2695 separations module HPLC equipped with a photodiode array and fractionated into a 96-deep well plate, collecting 96 fractions starting at 2 min and finishing at 56 min. Fractions were obtained approximately every 30 s. The mobile phase was curved (curve 8) from 5% acetonitrile, 95% water at 2 min to 80% acetonitrile at 45 min at a flow rate of 3 mL/min. Plates were dried overnight in a

GeneVac HT4 series 2 and resuspended in 60 μL of dd$H_2O$, and 25 μL were placed in fresh plates along with 25 μL 10 mM $AuCl_3$ and left to react at room temperature for 30 min.

### 5.5.5 Identification of delftibactin biosynthetic gene cluster and adenylation domain specificity.

Delftibactin genes encoding NRPS and PKS were identified using the BLAST function of IMG (http://img.jgi.doe.gov/), using the sequence of *pksJ* as a query. Adenylation domain specificities were assessed using NRPS Predictor[21] or NRPS-PKS[22], and the ten residue codes[12] of each entry and its top scoring hit were recorded. For the alignment of the adenylation domains specific for hydroxylated ornithine, the delftibactin adenylation code and the vicibactin adenylation code were determined with NRPS Predictor[21] and aligned manually as neither database contained domains with homologous sequences.

### 5.5.6 Construction of the ΔdelG *D. acidovorans* strain.

All primers and plasmids used in this process are described in Supplementary Table 5. If not stated explicitly, genetic manipulations and molecular biology techniques followed those from Cold Spring Harbor Protocols, available at http://www.molecularcloning.com/.

A knockout plasmid for *D. acidovorans* was constructed by inserting a 2-kb PCR product of *delG*(primers 2kbNRPS2Xba2 and 2kbNRPS2Sac2) into pUC19 using Xba1 and Sac1 digest sites, ligating with T4 ligase, transforming into chemically competent DH5α (Invitrogen) and plating on LB medium with 100 μg/mL ampicillin. Positive clones were identified by colony PCR with 2kbNRPS2Xba2 and 2kbNRPS2Sac2 and verified

through digestion following an overnight growth and plasmid miniprep using a QIAprep Spin Miniprep Kit (Qiagen). A clone containing a 2-kb insert was digested with Not1 to cut in the middle of the 2-kb insert, treated with calf intestinal phosphatase (CIP) and gel extracted to remove remaining CIP. A tetracycline resistance cassette was amplified from pLLX13 (primers TetNotF and TetNotR), purified, digested with Not1 and ligated with the digested vector. This ligation was transformed into chemically competent DH5α (Invitrogen) and plated on LB medium with 100 μg/mL ampicillin and 10 μg/mL tetracycline. Positive colonies were confirmed by PCR with TetNotF and TetNotR and verified with digestion with both Xba1 and Sac1 or with Not1, following an overnight growth and plasmid miniprep. The resulting plasmid was modified further by digesting with HindIII, treating with CIP and gel extracting to remove remaining CIP. An oriT was amplified from pLLX13 with primers OriTF and OriTR, purified, digested with HindIII and ligated with the cut vector before transforming into chemically competent DH5α and plating on LB medium with 100 μg/mL ampicillin and 10 μg/mL tetracycline. Positive colonies were identified with PCR using OriTF and OriTR and verified by digestion following an overnight growth and plasmid miniprep. The final plasmid (pDEL19) was transformed into chemically competent *E. coli* ET12567 carrying the helper plasmid pUZ8002, plating on LB medium with 10 μg/mL tetracycline, 30 μg/mL chloramphenicol and 25 μg/mL kanamycin. Positive transformants were grown in 3 mL LB medium with antibiotics overnight at 37 °C, alongside a 3-mL ACM growth of *D. acidovorans* at 30 °C. One milliliter of the donor *E. coli* and 1 mL of the recipient *D. acidovorans* were centrifuged separately and washed twice with fresh LB medium. Cells

were resuspended in 1 mL LB medium and mixed 1:1, and 300 μL was dispensed on nutrient agar plates and left to grow overnight at 30 °C. Cells were scraped and resuspended in 3 mL LB medium, plating 50 μL on LB plates containing 30 μg/mL tetracycline and 100 μg/mL apramycin to remove *E. coli*. Colonies were observed and tested for growth in LB medium with antibiotics at 30 °C and 200 r.p.m., with viable cultures streaked on LB plates with 30 μg/mL tetracycline. Colony PCR with TetNotF and TetNotR was used to confirm the presence of the tetracycline cassette. Chromosomal integration of the tetracycline cassette was confirmed by PCR with TetNotR and NRPS2Seq2 primers.

### 5.5.7 16S alignment and delftibactin production in environmental strains.

Environmental isolates from around the McMaster University campus were identified as *D. acidovorans* strains on the basis of 16S sequence alignment, using 16S sequences that were PCR amplified from single colonies (see above). Using these sequences and the 16S sequence for the *D. acidovorans* genome strain SPH1 from GenBank, along with the sequence for a gold biofilm isolate of *D. acidovorans* (accession number: GU013673)[4] were aligned with Geneious software version 4.8.5 (http://www.geneious.com/), using a Tamura-Nei genetic distance model, a neighbor-joining tree building method featuring a global alignment with free end gaps. Isolated strains were grown for 3 d at 30 °C and 190 r.p.m. in 1 L of ACM that had been treated with Chelex100 resin to limit the iron concentration. Cultures were centrifuged at 7,000 r.p.m. to remove the cell mass, and supernatants were treated with 20 g/L washed HP20 resin (Dialon). After 1 h of shaking with the supernatant, HP20 was collected by Buchner

funnel vacuum filtration and eluted with 400 mL of methanol. The resin eluent was evaporated to dryness, resuspended in 50% methanol and water and injected into a Waters AutoPure LC/MS using a similar method as above. MassLynx software was used to generate the 1,033-*m/z* extracted ion chromatograms for each extract. Fragmentation of these compounds was carried out on a Bruker AmazonX ion trap mass spectrometer.

### 5.5.8 Delftibactin-Au(III) precipitation measurements.

The gold-delftibactin interaction was determined through two separate experiments. First, $AuCl_3$was held constant at 2.5 mM, and the interaction with delftibactin was monitored by measuring the absorption of $AuCl_3$ remaining in solution after precipitation by delftibactin through comparison with a standard curve. Briefly, 2.5 mM $AuCl_3$ was incubated with 5 mM, 2.5 mM, 1.25 mM, 0.6125 mM, 0.3063 mM and 0.1531 mM delftibactin for 1 h. Solutions were filtered with a 0.22-μM Acrodisc (Pall) to remove insoluble delftibactin-gold precipitate. One hundred microliters of the filtered reaction were placed in a 96-well plate, and the absorbance was read at 300 nm using a SpectraMax 384 Plus (Molecular Devices). Absorption was compared to a $AuCl_3$ standard curve to determine the concentration of $AuCl_3$ remaining in solution. To monitor the amount of delftibactin remaining in solution after reaction with gold, a similar experiment was conducted. Delftibactin (2.5 mM) was incubated with 5 mM, 2.5 mM, 1.25 mM, 0.6125 mM, 0.3063 mM and 0.1531 mM $AuCl_3$for 1 h. Solutions were filtered similar to above. Reaction mixtures were analyzed using a Waters Alliance 2695 RP-HPLC separations module, equipped with a Waters 2998 photodiode array and a Luna 5u $C_{18}$ column (250 ×

4.60 mm, Phenomenex). The mobile phase was linear from 2% acetonitrile, 98% water + 5 mM $(NH_4)_2CO_3$ at 2 min to 14% acetonitrile at 18 min at a flow rate of 1 mL/min. The UV peak associated with delftibactin ($T_r$ = 12.29 min) was integrated and compared to a standard curve.

### 5.5.9 Transmission electron microscopy of delftibactin-Au(III) complexes.

Delftibactin was reacted with $AuCl_3$ with a molar ratio equal to 2:1 for 10 min. Each separate reaction of delftibactin with $AuCl_3$ was examined using a Phillips CM-10 transmission electron microscope operating at 80 kV. The whole-mount sample was absorbed and dried on a formvar-carbon–coated 100-square mesh copper grid and rinsed with filter-sterilized, de-ionized water to remove any salt precipitates.

### 5.5.10 Gallium-delftibactin-gold interaction.

Gallium-bound delftibactin was adjusted to 10 mM and mixed 1:1 with an equimolar solution of $AuCl_3$, alongside purified delftibactin and water (control). The reaction mixture was monitored at room temperature for 30 min.

### 5.5.11 Gold detoxification by delftibactin.

The assay was set up as follows: 50 μL of $ddH_2O$ containing 3.2 mM, 1.6 mM, 1.4 mM, 1.2 mM, 1.0 mM, 0.8 mM, 0.6 mM, 0.4 mM and 0.2 mM delftibactin was added in quadruplicate to wells within a 96-well plate. A 1.6-mM stock solution of $AuCl_3$ was made, and 50 μL was added to each well containing delftibactin. No-delftibactin and no-

$AuCl_3$ controls containing only water were also added to the 96-well plate in quadruplicate. These were incubated for 30 min at room temperature, during which time 10 mL of an overnight culture of *D. acidovorans* grown in ACM was centrifuged and resuspended in 5 mL sterilized $ddH_2O$. After 30 min incubation of $AuCl_3$ with delftibactin, 100 μL of concentrated culture was added to each well. Final concentration of $AuCl_3$ was 400 μM, and the final concentrations of delftibactin were 800 μM, 400 μM, 350 μM, 300 μM, 250 μM, 200 μM, 150 μM, 100 μM and 50 μM. After 30 min of incubation at room temperature, mixtures were serially diluted and plated onto nutrient agar plates and incubated at 30 °C. Colonies were counted after 24 h of growth. Results are shown as mean ± s.d.; $n = 4$. To assess whether *D. acidovorans* could grow in the presence of the gold precipitate, several milligrams of delftibactin and $AuCl_3$ were reacted 1:1 overnight, centrifuged and washed once with water to concentrate the precipitate. The precipitate was resuspended in $ddH_2O$ at a final concentration of 100 mM, calculated using a molecular weight of 1,227 g/mol, corresponding to a gold-delftibactin species. *D. acidovorans* was grown overnight in a 96-well plate in 100 μL of ACM containing 20 μM–10 mM gold precipitate or $AuCl_3$. No growth was observed in any well containing $AuCl_3$, whereas full growth was observed in every well containing the corresponding amount of precipitate. We have determined the MIC of $AuCl_3$ to be roughly 10 μM.

### 5.5.12 Gold detoxification in chronic exposure by delftibactin in presence and absence of iron.

Twenty-microliter reactions were set up as follows: (i) water only, (ii) 5 mM delftibactin, (iii) 5 mM $AuCl_3$, (iv) 5 mM $FeCl_3$, (v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$, (vi) 5 mM delftibactin B + 5 mM $AuCl_3$, (vii) delftibactin B + 5 mM $AuCl_3$ + 5 mM $FeCl_3$, (viii) delftibactin B + 5 mM $FeCl_3$. Reactions were initiated by the addition of delftibactin, and images were taken at the time points as indicated in Figure 5.2b and Supplementary Figure 10. After 2 h, reactions were serially diluted to a 30-μM final concentration in ACM containing *D. acidovorans* Δ*delG* diluted 1:1,000 from an overnight culture. Optical density was monitored using a TECAN Sunrise microplate reader at 600 nm for 36 h. Results are a mean of three growth curves for each condition from a single representative experiment. As a second test of delftibactin protective capacity, *D. acidovorans* Δd*elG* cells from an overnight culture were inoculated 1:1,000 into 100 μL ACM in a 96-well plate containing 0 μM or 10 μM $AuCl_3$ and then provided 0 μM or 100 μM delftibactin. Cultures were grown for 84 h at 250 r.p.m. at 30° in a TECAN Sunrise microplate reader and measured at 600 nm to assess growth. Results are a mean of three growth curves for each condition from a single representative experiment.

### 5.5.13 Gold protective comparison of delftibactin A and B.

Twenty-microliter reactions were set up as follows: water only, 5 mM $AuCl_3$, 5 mM $AuCl_3$ + delftibactin A, and 5 mM $AuCl_3$ + delftibactin B. Reactions were initiated with the addition of delftibactin A or B. After 2 h, reactions were serially diluted to 125 μM in ACM containing *D. acidovorans* Δ*delG* diluted 1:1,000 from an overnight culture. Optical density was monitored using a TECAN Sunrise microplate reader at 600 nm for 36

h. Results are a mean of three growth curves for each condition from a single representative experiment.

### 5.5.14 Delftibactin-mediated protection against gold toxicity.

Cultures of *D. acidovorans* wild type and *D. acidovorans* Δ*delG* were grown in deferrated ACM for 2 d at 30 °C. In 96-well plates, 200 μL of wild-type or mutant grown culture were incubated in the presence and absence of 100 μM $AuCl_3$. At the same time, the mutant culture was also complemented with 30 μM delftibactin (biological concentration) in the presence of 100 μM $AuCl_3$. After 30 min, cultures were serially diluted in water and plated on LB agar to determine the colony-forming units of *D. acidovorans* after gold exposure. Results are shown as mean ± s.d.; $n = 4$; Two-tailed student's *t*-test.

### 5.5.15 Measuring delftibactin production following depletion by gold.

Cultures of *D. acidovorans* were grown in 10 mL ACM in 50-mL Falcon tubes for 48 h, at which point both growth and delftibactin production had ceased. Cultures were then pelleted by centrifugation at 4,500 r.p.m. for 30 min at 4 °C. Supernatants were kept separate and moved into labeled, sterile 50-mL Falcon tubes while cell pellets were kept on ice. Initial delftibactin concentrations were assessed by filter sterilizing and then placing 400 μL of each supernatant into a HPLC sample vial, storing at 4 °C. Supernatants were then adjusted to 0 μM, 10 μM or 30 μM $AuCl_3$ with the appropriate volume of 10 mM $AuCl_3$ and left to react at room temperature. After 12 h, supernatants were returned the appropriate cell pellets and resuspended by vortexing before taking a second 400-μL

sample to assess delftibactin depletion by gold treatment. Cultures were returned to the incubator and left shaking for another 48 h before a final sample was taken. Delftibactin concentrations were assessed by MRM-LC/MS for delftibactin, with washes between each sample. Values of integrated delftibactin peaks were normalized to the untreated control and represent the percent increase in delftibactin concentration (± propagated error) observed 48 h following gold precipitation; $n = 6$.

**5.5.16 MRM-LC/MS measurement of delftibactin production in response to iron.**

Cultures of *D. acidovorans* were grown for 48 h in 10-mL cultures of deferrated ACM resupplied with $FeCl_3$ in varying concentrations. Iron concentrations and corresponding delftibactin concentrations in the filter-sterilized supernatants are listed in Supplementary Table 6. Delftibactin concentrations were established using MRM-LC/MS. Results are shown as mean ± s.d.; $n = 3$.

**5.5.17 Citrate-gold and delftibactin-gold comparison.**

Stock 10-mM solutions of sodium citrate and purified delftibactin were mixed with an equimolar solution of $AuCl_3$ and allowed to react at room temperature in 1.5-mL Eppendorf tubes. Photographs were taken from the initial addition of gold to 1-h exposure. Similarly, TEM experiments were performed by mixing 10-mM stock solutions of $AuCl_3$ and either delftibactin or sodium citrate 1:1 on a formvar-carbon–coated 100-square mesh copper grid, imaging after 10 min of $AuCl_3$ exposure.

**Accession codes.**

Genbank: The accession code for *delG* is 5750346.

## 5.6 Supplementary Information

Supplementary information from the publication referred to in this chapter can be found in Appendix 4.

## 5.7 References

1. Vining, L.C. *Annu. Rev. Microbiol.* **44**, 395–427 (1990).
2. Nies, D.H. *Appl. Microbiol. Biotechnol.* **51**, 730–750 (1999).
3. Reith, F., Rogers, S.L., McPhail, D.C. & Webb, D. *Science* **313**, 233–236 (2006).
4. Reith, F. *et al. Geology* **38**, 843–846 (2010).
5. Reith, F. *et al. Proc. Natl. Acad. Sci. USA* **106**, 17757–17762 (2009).
6. Reith, F., Lengke, M.F., Falconer, D., Craw, D. & Southam, G. *ISME J.* **1**, 567–584 (2007).
7. Kashefi, K., Tor, J.M., Nevin, K.P. & Lovely, D.R. *Appl. Environ. Microbiol.* **67**, 3275–3279 (2001).
8. Usher, A., McPhail, D.C. & Brugger, *J. Geochim. Cosmochim. Acta* **73**, 3359–3380 (2009).
9. Hider, R.C. & Kong, X. *Nat. Prod. Rep.* **27**, 637–657 (2010).
10. Kim, H.J. *et al. Science* **305**, 1612–1615 (2004).
11. Chaturvedi, K.S. *et al. Nat. Chem. Biol.* **8**, 731–736 (2012).

12. Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. *Chem. Biol.* **6**, 493–505 (1999).

13. Diels, L., Dong, Q., van der Lelie, D., Baeyens, W. & Mergeay, M. *J. Ind. Microbiol.* **14**, 142–153 (1995).

14. Salem, I.B. *et al. Ann. Microbiol.* published online, doi:10.1007/s13213-012- 0462-3 (2012).

15. Miller, M.C. *et al. Microbiology* **156**, 2226–2238 (2010).

16. Hough, R.M. *et al. Geology* **36**, 571–574 (2008).

17. Ojea-Jiménez, I., Romero, F.M., Bastús, N.G. & Puntes, V. J. Phys. Chem. C 114, 1800–1804 (2010).

18. Amin, S.A. *et al. J. Am. Chem. Soc.* **129**, 478–479 (2007).

19. Pinel, N., Davidson, S.K. & Stahl, D.A. *Int. J. Syst. Evol. Microbiol.* **58**, 2147–2157 (2008).

20. Weisburg, W.G., Barns, S.M., Pelletier, D.A. & Lane, D.J. *J. Bacteriol.* **173**, 697–703 (1991).

21. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. *Nucleic Acids Res.* **33**, 5799–5808 (2005).

22. Ansari, M.Z., Yadav, G., Gokhale, R.S. & Mohanty, D. *Nucleic Acids Res.* **32**, W405–W413 (2004).

## Chapter 6. Significance and future prospective

Microbial natural products are valuable small molecules with immense importance as antibiotics and other therapeutics.[41] Despite this, over-reliance on classical discovery techniques and producers has resulted in stagnation and waning interest.[28-30] New techniques and approaches are now required to improve our understanding of microbial natural products and guide future discovery efforts to obtain these valuable molecules. In this thesis, I demonstrate new strategies to guide the study and discovery of microbial natural products, from databases of known compounds, from genome data, and from previously uninvestigated bacteria.

## 6.1 Informatic platform development for the discovery and study of microbial natural products.

Microbial natural products are one of the most important sources of bioactive small molecule leads, but decades of over-reliance on the same microorganisms and the same detection techniques has led to diminishing returns, bringing many to question the relevance of natural products in the modern era.[28] Widespread bacterial genome sequencing put many of these fears to rest, demonstrating that there are vast numbers of bacterial natural products left to discover,[14, 69] but the challenge has now become how to realize these nascent natural products that had been overlooked by previous studies. Moreover, now that natural product biosynthetic gene clusters could be identified in microbial genomes, how could these gene clusters be prioritized for focused discovery efforts? And how would these molecules be identified in a reliable and rapid fashion? In all cases, novel informatic approaches that could leverage our prior knowledge of these well-studied systems provided a useful solution. By using our knowledge of natural product biosynthesis, MS/MS fragmentation, and mechanism of action, we could construct and utilize new informatic programs to address these new challenges for modern natural products research, demonstrating that these promising molecules can still be targeted and identified to provide desired bioactive molecules and scaffolds.

At the beginning of my graduate studies, mining microbial genomes for natural products was possible, but it was a low fidelity process that required intense knowledge of biosynthesis and still rarely provided accurate predictions that could be used to find small molecules. Programs including NRPSPredictor[61] and NRPS-PKS[59] had been developed to

analyze ten key amino acids in adenylation domains to predict the monomers incorporated into nonribosomal peptides – a strategy that was then quickly be adapted to acyl transferases of PKS systems. Individuals could piece together predicted monomers along with inferred modifications to try and assemble a predicted natural product.[3] Using this predicted structure, researchers could search LC-MS data for closely related masses, and then inspect MS/MS fragmentation patterns of candidate ions to find features (e.g. mass gaps associated with amino acids) that are indicative of the predicted structure.[64] Alternatively, genetics knockouts of the targeted biosynthetic gene cluster would be required to identify molecules associated with the gene of interest. Whether using predictions or knockouts, both approaches require intimate knowledge of biosynthesis, or of genetics and molecular biology, both of which required a significant time investment. Although emergent programs such as AntiSmash[80] continued to focus on the detection of biosynthetic gene clusters and prediction of their products, there were no automated strategies for connecting genes to molecules, or for facilitating the rapid discovery of natural products from targeted gene clusters. In **Chapter 2**, we demonstrated the first broadly applicable and functional platform capable of using genetic inputs to predict and locate associated natural products in LC-MS/MS data.

The Genomes-to-Natural Products platform (GNP) represented the first fully realized automated pipeline for using genes to find small molecules, and did so using a unique combinatorialization strategy. The promiscuity inherent to natural product biosynthesis has long been a complicating factor in using predicted structures to find true natural products,[63] and even relatively small NRPS assembly lines can sometimes give rise

to as many as 50 variants in a single culture.[81] Although this complexity is a complicating factor for most natural products prediction or identification programs, GNP features a unique prediction combinatorialization package that elaborates predicted scaffolds to find all related natural products from a given extract simultaneously. Using fragment ion matching algorithms along with a generous window for parent ion matches, GNP can use libraries of thousands of predicted structures to profile LC-MS/MS data and identify natural products corresponding to the targeted gene cluster. Previously, methods of parent and fragment ion matching[65] (as well as the related concept of mass-gap matching[64]) had been limited to peptides, whose amino acid components would reliably fragment and provide a series of monomers that could be used 'sequence' the structure. Uniquely, GNP can also predict and detect polyketides and glycosylated natural products, wherein predicted parent masses, hydroxylation patterns, and glycosylation patterns provide similarly indicative data that can be used to identify small molecules in LC-MS/MS data and give insight into their structures. Importantly, we also demonstrated in this publication that this parent and fragment ion matching approach is also applicable to other classes of natural products, indicating that future work in developing more advanced and comprehensive prediction algorithms will facilitate the identification of nearly all microbial natural products. In light of this capability, GNP may be useful not only as a means of detecting the products of specific gene clusters, but also structures related to a desired scaffold. In this way, expansive microbial extract libraries could be screened to look for desired variants of known scaffolds, independent of genomic information.

To prove the efficacy of GNP for connecting genes and small molecules, we performed a number of important demonstrations that also became important milestones for informatic platform development. In addition to showing that GNP could predict, combinatorialize, and identify both nonribosomal peptides and polyketides, it was also important to demonstrate that GNP could function with bacteria that had not previously been known to produce natural products. Genome sequencing has shown that many families of bacteria may produce NRPS/PKS molecules, but aside from the obvious need to prove this assumption, it is also crucial to explore these organisms to show that actinomycete-centric homology models can function with exotic new protein and DNA sequences, both to identify new gene clusters and predict their small molecule products. In identifying the acidobactins, variobactins, and vacidobactins, we demonstrated for the first time that *Acidovorax* and *Variovorax spp*. could produce complex natural products, and that our Hidden Markov model collection was sufficiently comprehensive to detect molecules from previously unexplored organisms. Lastly, GNP was also the first automated or informatic strategy to identify a cryptic natural product,[82] which has been a long-standing challenge for the field. This success could largely be credited to the accurate prediction and fragment matching capabilities of GNP, the latter of which is not strongly dependent on abundance of a given natural product, and thus can facilitate identification of extremely low abundance molecules such as thanamycin. With the publication of the PRISM algorithm and web application,[67] our lab has demonstrated an even more complete and automated version of the GNP prediction and library generation software, and further establishes these methods as premier informatic tools for the prediction and identification of targeted natural products.

In this first manuscript, I used bio- and chemo-informatics to predict and identify natural products, making use of extant genome data to realize new molecules. In **Chapter 3**, I demonstrated how informatic techniques could be used to analyze and prioritize natural products for study, revealing valuable structures that had been overlooked in the past. Through years of effort, our laboratory was able to compile a comprehensive library of microbial natural products, including well known structures and obscure ones, collected in various databases including the Handbook of Antibiotic Compounds.[83] Simultaneously, the development of the GRAPE retrobiosynthetic algorithm provided a means to analyze these disparate structures, breaking them down into individual monomers that could be aligned, scored against one another, and used for hierarchical clustering. The end result of these two unique capabilities was an approach that could be applied to quantify various aspects of natural products, which are an extremely broad class of small molecules that have proven exceptionally challenging for chemoinformatic analysis.[84] As a first example of this, we analyzed microbial natural products that were known antibacterial agents, as these molecules are desperately needed for the development of new antibacterial drugs that can avoid or overcome known resistance.[70] In doing so, we learned a great deal about the diversity of natural product scaffolds, and of their targets, demonstrating that new scaffolds have nearly a 40% chance of affecting a completely new molecular target. Closer inspection of large targets such as the ribosome shows that divergent scaffolds affect different locations on these multi-component enzymes,[85] indicating that the frequency of new scaffolds possessing new mechanisms of action may be considerably higher still than those possessing new targets. Most importantly, this clustering analysis revealed a number of

antibiotic families that did not possess known targets or cross resistance that could then be prioritized for examination.

To demonstrate the value of this extensive chemoinformatic study, we undertook mechanistic studies of the telomycin family of nonribosomal peptide natural products. Our finding that telomycin possessed a new target (the phospholipid cardiolipin) validated our study, and was also a fascinating discovery in its own right. Although cardiolipin is present in mammalian cells, it is sequestered inside mitochondria,[86] and appears to be inaccessible to telomycin, as all studies to date have found no apparent toxicity in a number of different cell lines and organisms.[87-88] As such, cardiolipin is an essential lipid that is effectively unique to bacteria, and as such could prove to be a useful target for drug discovery efforts. Our own work into identifying or creating new telomycin derivatives also shows promise, as the loss of several post-NRPS hydroxylations appears to improve activity, and that incorporation of alkyl groups into the indole rings of telomycin's two tryptophans also improves activity. Combining this initial structure-activity relationship with the beneficial N-acylation observed by the Müller group,[88] or with the loss of the N-terminal aspartic acid as seen in the related molecule LL-AO341β,[89] may result in highly active molecules suitable as leads for development as clinically-relevant antibacterial drugs. Importantly, this success was dependent on the development of our unique informatic platform that used a novel algorithm and massive chemical database to identify these and numerous other molecules as candidate leads. Personally, I am excited to see whether other molecules we identified in this work are taken up by our laboratory or by others as potential therapeutics

that will further underscore the importance of using informatic platforms to find and study microbial natural products.

## 6.2 Informatics-directed investigations for novel sources of microbial natural products.

Microbial natural products are bioactive small molecules that have been evolved through natural selection over thousands of years to perform functions that benefit their producer.[4, 42-43] In this way, collections of natural products can be considered to be enriched in biological activity relative to synthetic compound libraries, which have not been honed through similar conditions, and have typically not been created with a specific biological activity or target in mind. Despite this, it is important to note that the specific biological activity or target of a natural product may not be relevant for a specific screen or desired phenotype, leading to libraries with a diverse range of activities that can be useful for screening a wide range of indications, even though hit rates for individual targets may be very low. As such, for natural products discovery to be commercially viable, it requires either targeted discovery approaches that can access desired pharmacophores or activities, or that their libraries be exceptionally large, such that a molecule with a desired activity must – statistically speaking – exist within it. During the early years of microbial natural products, researchers achieved scale and a sense of targeted discovery using selective culturing techniques to enrich for microorganisms that had previously been shown to produce natural products generally, or to produce a specific class of compounds which they wished to explore further. Organisms such as the actinomycetes have phenotypic features,

media preferences, and specific antimicrobial susceptibility profiles that allow for their enrichment from environmental samples, leading to expansive libraries of organisms that are often related to a select few 'talented' families that readily produce natural products in the laboratory. Although it has not been quantitatively studied, it is widely acknowledged that bacterial families and genera possess relatively distinct sets of natural product scaffolds,[74-75] in keeping with the evolution of these molecules for specific lifestyles or environmental niches. While some genera, such as *Streptomyces*, possess large numbers of biosynthetic gene clusters per genome,[14] and possess an exceptionally deep pool of associated natural product scaffolds, the number of scaffolds left to discover from these organisms still approaches zero as more organisms are fermented. In short, despite their significant endowment, continued sampling of even talented producers necessarily decreases the likelihood of finding something new. This unfortunate fact was illustrated in the rediscovery crisis that nearly killed industrial natural products programs in the late 1990s and early 2000s,[28] when identifying new molecules with useful (and profitable) activities became cost prohibitive.

As researchers hit diminishing returns from the reliable, readily culturable bacteria that had led the charge during the Golden Age of Antibiotics, productive gains were observed from less conventional bacteria that had been more challenging to grow and extract. For instance, myxobacteria[75] and some genera of cyanobacteria[74] were found to produce massive numbers of distinct bioactive natural products, but had not received extensive attention prior to the relative decline of actinomycetes and *Bacilli* due to their slow growth and exotic culture requirements. Some companies, including SmithKline-

Beecham and Bristol-Meyers Squibb, discovered valuable molecules from Gram-negative bacteria such as *Flavobacterium*,[90] *Burkholderia*,[91] and Pseudomonads,[92] although these bacteria were largely ignored by other groups. Now, genome sequencing has elaborated the ranks of natural product-generating bacteria even further, indicating that both exotic[76] and well-studied[93] bacteria could produce valuable small molecules, but may simply produce them infrequently relative to established genera. Finally, metagenomic sequencing has indicated that staggering numbers of bacteria have gone undiscovered, and that only 1% of bacteria can be readily cultured in laboratory conditions.[94] If this is true, then the current set of >10,000 microbial natural products, including hundreds of drugs and lead molecules, is perhaps 1% of the final number that could exist. A number of studies have shown the potential value of these 'unculturable' organisms. Work from the Piel group has shown that an unculturable family of bacteria associated with marine sponges produces (essentially) the entire formidable chemical arsenal of the sponge;[76] molecules that look nothing like natural products from other known organisms. In early 2015, the Lewis lab used a new approach to cultivate 'unculturable' bacteria for antibiotic screening, leading to the discovery of teixobactin, a lipid II binding nonribosomal peptide that does not appear to have any known relatives from previously studied bacteria.[71] As genome sequencing technologies and effort continue to improve our view of biosynthetic potential, new organisms must be explored to find new molecules. In **Chapters 4** and **5**, I demonstrated that new bio- and chemo-informatics strategies that can focus on new bacteria that possess unique biosynthetic gene clusters are some of the most promising means to access novel chemical scaffolds and biological activities.

As mentioned earlier, the explosion of sequenced bacterial genomes revealed that natural products might be produced by many more bacteria than had been previously suspected. In keeping with our understanding that bacterial families produce relatively distinct sets of small molecules, we chose to investigate an exotic organism which had not been previously studied for natural products, but had an established system for culture and genetic manipulation, settling on the genus *Legionella* (discussed in **Chapter 4**). Phylogenetically, *Legionella* (and the eponymous order Legionellales) are among the most obscure and early branching of the *Gammaproteobacteria*, and are more closely related to chemolithotrophic oceanic bacteria like *Beggiatoa*, *Thioploca*, and the Thiotrichales than they are to the well-known *Gammaproteobacteria* like *Escherichia coli* or *Pseudomonas aeruginosa*.[95] As such, it is perhaps not surprising that *Legionella spp*. possess a unique set of bizarre NRPS and PKS biosynthetic gene clusters, which are distinct from those observed in other related bacteria. Although our analysis of over 30 *Legionella* genomes revealed 34 different NRPS/PKS gene clusters (141 gene clusters were observed in total, including redundancies), it remains to be seen how many of these gene clusters are truly functional. In some cases, such as the large *trans*-AT polyketide legioliulin, relatively canonical type I PKS logic is observed and adhered to, presenting a pristine, massive assembly line similar to those observed in *Streptomyces*, *Bacillus*, or *Pseudomonas* (although the *trans*-acting acyl transferase [AT] is located on a distant end of the genome, which is unusual[96]). In other cases, however, several components that should be essential for natural product biosynthesis are missing, and many of these gene clusters seem too small to produce anything of value, and may function instead as bacterial versions of the

adenylate-forming reductases observed in fungi.[97] *Legionella* is known to possess a patchwork genome that acquires and loses DNA fairly frequently,[98] and so it is a live question whether the unusual architecture observed in these gene clusters is due to a unique biosynthetic strategy, or whether they are in the process of being lost. Our discovery and description of legionellol is fascinating in this respect, as it arises from a bizarre biosynthetic gene cluster composed to two operons that should not be capable of performing all the biosynthetic transformations involved in creating this unusual natural product. As such, it does seem that at least some of these unusual gene clusters likely still create natural products, and that more thorough genetic analysis could provide insight into new enzymes or mechanisms for creating NRPS-PKS molecules. In addition to the phenotypic effects noted following the interruption of legionellol biosynthesis, our group and others have noted the role of the larger PKS-NRPS gene cluster in promoting lysosome avoidance following phagocytosis.[99] At the moment, however, we have no leads regarding the nature of any potential small molecule products from this gene cluster, which also appears to lack several important enzymes, including an acyl transferase. While this last finding provides a place to go forward for future work, the identification of legionellol and the fascinating biosynthetic capabilities of *Legionella* serve to demonstrate the utility of our bioinformatics based approach to finding new sources of novel natural products chemistry.

In addition to new chemistry, the exploration of new bacteria for natural products should also reveal new biological activities. In an attempt to pursue these, we chose to investigate *Delftia acidovorans,* a bacterium that had not been investigated previously and

was known as one of the only bacteria capable of living on solid gold deposits.[78] Using a combination of prediction- and bioactivity-guided discovery, we identified the nonribosomal peptide delftibactin from *D. acidovorans* cultures, and demonstrated that this metallophore functioned to protect *Delftia* from toxic soluble gold. While another gold-resident, *Cupriavidus metallidurans*, is known to use enzymatic methods to encourage the intracellular precipitation of toxic gold ions,[100] this was the first example of small molecule mediated extracellular precipitation. As such, delftibactin production likely assists *D. acidovorans* thrive in this otherwise hostile environment, presenting one of the most eloquent demonstrations of how natural products assist in the colonization of unique environmental niches.

### 6.3 Concluding remarks

Despite the continuing advances made in synthetic chemistry and drug development, natural products remain some of the most important bioactive compounds available, thanks in large part to the selective pressure inherent in their creation. Following nearly a century of collecting these small molecules and studying their origins, we are finally beginning to develop targeted strategies that can reveal desired compounds in an automated way, enabling directed mining of Nature's most valuable chemical structures and bioactivities. As I have demonstrated in this thesis, new techniques that can leverage our knowledge of natural products chemistry and biosynthesis provide an opportunity to return these valuable small molecules to the forefront of therapeutic discovery and development.

**References**

1. George, K.M. *et al*. Mycolactone: a polyketide toxin from *Mycobacterium ulcerans* required for virulence. *Science* **283**, 854-857 (1999).
2. Keller, L., & Surette, M. G. Communication in bacteria: an ecological and evolutionary perspective. *Nat Rev Microbiol*. **4**, 249-258 (2006).
3. Johnston, C.W. *et al*. Gold biomineralization by a metallophore from a gold-associated microbe. *Nat Chem Biol*. **9**, 241-243 (2013).
4. Vining, C.L. Roles of secondary metabolites from microbes. CIBA Foundation Symposium 171 - Secondary Metabolites: Their Function and Evolution. *Wiley, Chichester*. 184-198 (1992).
5. Waksman, S.A. Antibiotics. *Biol Rev Camb Philos Soc*. **23**, 452-487 (1948).
6. Dias, D.A., Urban, S. & Roessner, U. A historical overview of natural products in drug discovery. *Metabol*. **2**, 303-336 (2012).
7. Greenwood, D. The quinine connection. *J Antimicrob Chemother*. **30**, 417-427 (1992).
8. Sepkowitz, K.A. One hundred years of Salvarsan. *N Engl J Med*. **365**, 291-293 (2011).
9. Waksman, S.A. Origin and nature of antibiotics. *Am J Med*. **7**, 85-99 (1949).
10. Katz, L. & Baltz, R.H. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol*. **in press** (2016).

11. Walsh, C.T. & Fischbach, M.A. Natural products version 2.0: Connecting genes to molecules. *J Am Chem Soc* **132**, 2469-2493 (2010).
12. Simpson, T.J. Application of isotopic methods to secondary metabolic pathways. *Topics in current chemistry* **195**, 1-48 (1998).
13. Bode, H.B. & Müller, R. The impact of bacterial genomics on natural product research. *Angew Chem Int Ed Engl* **44**, 6828-6846 (2005).
14. Doroghazi, J.R. *et al*. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* ***1*0**, 963-968 (2014).
15. Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. *Chem Rev* **106**, 3468-3496 (2006).
16. Donadio, S., Staver, M.J., McAlpine, J.B., Swanson, S.J., Katz, L. Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675-679 (1991).
17. Cortes, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J. & Leadlay, P.F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea*. *Nature* **348**, 176-178 (1990).
18. Lawen, A. & Zocher, R. Cyclosporin synthetase: The most complex peptide synthesizing multienzyme polypeptide so far described. *J Biol Chem* **265**, 11355-11360 (1990).

19. Challis, G.L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**, 111-114 (2000).
20. Alvarez, M.A., Fu, H., Khosla, C., Hopwood, D.A. & Bailey, J.E. Engineered biosynthesis of novel polyketides: Properties of the *whiE* aromatase/cyclase. *Nat Biotech* **14**, 335-338 (1996).
21. Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science* **291**, 1790-1792 (2001).
22. Dutta, S., *et al*. Structure of a modular polyketide synthase. *Nature* **510**, 512-517 (2014).
23. Whicher, J.R., *et al*. Structural rearrangements of a polyketide synthase module during its catalytic cycle. *Nature* **510**, 560-564 (2014).
24. Du, L. & Lou, L. PKS and NRPS release mechanisms. *Nat Prod Rep* **27**, 255–278. (2010).
25. Wyatt, M.A., Mok, M.C., Junop, M. & Magarvey, N.A. Heterologous expression and structural characterisation of a pyrazinone natural product assembly line. *Chembiochem* **13**, 2408-2415 (2012).
26. Müller, S., *et al*. Biosynthesis of crocacin involves an unusual hydrolytic release domain showing similarity to condensation domains. *Chem Biol* **21**, 855-865 (2014).
27. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl* **48**, 4688-4716 (2009).

28. Li, J.W. & Vederas, J.C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161-165 (2009).
29. Carter, G.T. Natural products and Pharma 2011: strategic changes spur new opportunities. *Nat Prod Rep* **28**, 1783-1789 (2011).
30. Koehn, F.E. & Carter, G.T. Rediscovering natural products as a source of new drugs. *Discov Med* **5**, 159-164 (2005).
31. Koehn, F.E. & Carter, G.T. The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* **4**, 206-220 (2005).
32. Terrett, N.K., Gardner, M., Gordon, D.W., Kobylecki, R.J. & Steele, J. Combinatorial synthesis – The design of compound libraries and their application to drug discovery. *Tetrahedron* **51**, 8135-8173 (1995).
33. Spandl, R.J., Díaz-Gavilán, M., O'Connell, K.M.G., Thomas, G.L. & Spring, D.R. Diversity-oriented synthesis. *Chem Rec* **8**, 129-142 (2008).
34. Drews, J. Drug discovery: A historical perspective. *Science* **287**, 1960-1964 (2000).
35. Lin, A.H., Murray, R.W., Vidmar, T.J. & Marotti, K.R. The oxazolidinone eperezolid binds to the 50s ribosomal subunit and competes with binding of chloramphenicol and lincomycin. *Antimicrob Agents Chemother* **41**, 2127-2131 (1997).
36. Walsh, C. Where will new antibiotics come from? *Nat Rev Microbiol* **1**, 65-70 (2003).
37. Rosamond, J. & Allsop, A. Harnessing the power of the genome in the search for new antibiotics. *Science* **287**, 1973-1976 (2000).

38. Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discov Today* **10**, 139-147 (2005).
39. Projan, S.J. Why is big Pharma getting out of antibacterial drug discovery? *Curr Opin Microbiol* **6**, 427-430 (2003).
40. Wright, G.D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* **5**, 175-186 (2007).
41. Newman, D.J. & Cragg, G.M. Natural products as sources of drugs over the last 30 years from 1981 to 2010. *J Nat Prod* **75**, 311-335 (2012).
42. D'Costa, V.M. *et al*. Antibiotic resistance is ancient. *Nature* **477**, 457-461 (2011).
43. Fischbach, M.A., Walsh, C.T. & Clardy, J. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci USA* ***105***, 4601-4608 (2008).
44. Fischbach, M.A. Antibiotics from microbes: Converging to kill. *Curr Opin Microbiol* **12**, 520-527 (2009).
45. Over, B. *et al*. Natural-product-derived fragments for fragment-based ligand discovery. *Nat Chem* **5**, 21-28 (2013).
46. Baumann, S. *et al*. Cystobactamids: Myxobacterial Topoisomerase Inhibitors Exhibiting Potent Antibacterial Activity. *Angew. Chem. Int. Edn. Engl.* **53**, 14605–14609 (2014).
47. Lin, A.H. *et al*. The oxazolidinone eperezolid binds to the 50S ribosomal subunit and competes with binding of chloramphenicol and lincomycin. *Antimicrob. Agents Chemother.* **41**, 2127–2131 (1997).

48. Keller, S. *et al*. Action of atrop-abyssomicin C as an inhibitor of 4-amino-4-deoxychorismate synthase PabB. *Angew Chem Int Ed Engl*. **46**, 8284–8286 (2007).
49. Walsh, C.T. & Wencewicz, T.A. Prospects for new antibiotics: a molecule-centered perspective. *J Antibiot* **67**, 7-22 (2014).
50. Sanger, F. *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695 (1977).
51. Fleischmann, R.D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* **269**, 496-512 (1995).
52. Bentley, S.D. *et al*. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141-147 (2002).
53. Loman, N.J. *et al*. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10**, 599-606 (2012).
54. Nett, M., Ikeda, H. & Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* **26**, 1362-1384 (2009).
55. Skinnider, M.A., Johnston, C.W., Zvanych, R. & Magarvey, N.A. Automated identification of depsipeptide natural products by an informatic search algorithm. *Chembiochem* **16**, 223-227 (2015).
56. Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol* **6**, 493–505 (1999).

57. Challis, G.L., Ravel, J. & Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* **7**, 211–224 (2000).

58. Reeves, C.D. *et al*. Alteration of substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. *Biochemistry* **40**, 15464-15470 (2001).

59. Ansari, M.Z., Yadav, G., Gokhale, R.S. & Mohanty, D. NRPS-PKS: a knowledge-based resource for analysis of NRPS-PKS megasynthases. *Nucleic Acids Res* **32**, W405-W413 (2004).

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

61. NRPSPredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* **39**, W362-W367 (2011).

62. Zazopoulos, E. *et al*. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat Biotech* **21**, 187-190 (2003).

63. Lantru, S., Deeth, R.J., Bailey, L.M. & Challis, G.L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat Chem Biol* **1**, 265-269 (2005).

64. Kersten, R.D. *et al*. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol* **7**, 794-802 (2011).

65. Ibrahim, A. *et al*. Dereplicating nonribosomal peptides using an informatic search algorithm for natural product (iSNAP) discovery. Proc Natl Acad Sci USA 109, 19196-19201 (2012).
66. Johnston, C.W. *et al*. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat Commun* **6**, 8421 (2015).
67. Skinnider, M.A. et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 43, 9645-9662 (2015).
68. Bérdy, J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J Antibiot* **65**, 385-395 (2012).
69. Cimermancic, P. *et al*. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421 (2014).
70. Bush, K. *et al*. Tackling antibiotic resistance. *Nature Rev. Microbiol.* **9**, 894–896 (2011).
71. Ling, L.L. *et al*. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
72. Cociancich, S. *et al*. The gyrase inhibitor albicidin consists of *p*-aminobenzoic acids and cyanoalanine. *Nat. Chem. Biol.* **11**, 195–197 (2015).
73. Hamamoto, H. *et al*. Lysocin E is a new antibiotic that targets menaquinone in the bacterial membrane. *Nat. Chem. Biol.* **11**, 127–133 (2015).
74. Nunnery, J.K., Mevers, E. & Gerwick, W.H. Biologically active secondary metabolites from marine cyanobacteria. *Curr Opin Biotechnol* **21**, 787-793 (2010).

75. Weissman, K.J. & Müller, R. A brief tour of myxobacterial secondary metabolism. *Bioorg Med Chem* **17**, 2121-2136 (2009).

76. Wilson, M.C. *et al*. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).

77. Kwan, J.C. *et al*. Genome streamlining and chemical defense in a coral reef symbiosis. *Proc Natl Acad Sci USA* **109**, 20655-20660 (2012).

78. Reith, F. *et al*. Nanoparticle factories: Biofilms hold the key to gold dispersion and nugget formation. *Geology* **38**, 843-846 (2010).

79. Reith, F., Lengke, M.F., Falconer, D., Craw, D. & Southam, G. The geomicrobiology of gold. *ISME J.* **1**, 567–584 (2007).

80. Medema, M.H. *et al*. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**, W339-W346 (2011).

81. Yang, L. *et al*. Exploration of nonribosomal peptide families with an automated informatic search algorithm. *Chem Biol* **22**, 1259-1269 (2015).

82. Mendes, R. *et al*. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097-1100 (2011).

83. Bérdy, J. *et al*. *Handbook of Antibiotic Compounds* Vols. I-X (CRC Press, Boca Raton, Florida, USA, 1980-1982).

84. Koch, M.A. *et al*. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* **102**, 17272-17277 (2005).

85. Wilson, D.N. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat Rev Microbiol* **12**, 35-48 (2014).
86. Mileykovskaya, E. & Dowhan, W. Cardiolipin membrane domains in prokaryotes and eukaryotes. *Biochim Biophys Acta* **1788**, 2084-2091 (2009).
87. Tisch, D.E., Huftalen, J.B. & Dickison, H.L. Pharmacological studies with telomycin. *Antibiot Annu* **5**, 863-868 (1957-1958).
88. Fu, C. et al. Biosynthetic studies of telomycin reveal new lipopeptides with enhanced activity. *J Am Chem Soc* **137**, 7692-7705 (2015).
89. Oliva, B., Maiese, W.M., Greenstein, M., Borders, D.B. & Chopra, I. Mode of action of the cyclic depsipeptide antibiotic LL-AO341β1 and partial characterization of a *Staphylococcus aureus* mutant resistant to the antibiotic. *J Antimicrob Chemother* **32**, 817-830 (1993).
90. Konishi, M. *et al*. Empedopeptin (BMY-28117), a new depsipeptide antibiotic. I. Production, isolation and properties. *J Antibiot* **37**, 949-957 (1984).
91. Meyers, E. *et al*. Xylocandin: a new complex of antifungal peptides. I. Taxonomy, isolation and biological activity. *J Antibiot* **40**, 1515-1519 (1987).
92. Thirkettle, J. *et al*. SB-253514 and analogues; novel inhibitors of lipoprotein-associated phospholipase A2 produced by *Pseudomonas fluorescens* DSM 11579. I. Fermentation of producing strain, isolation and biological activity. *J Antibiot* **53**, 664-669 (2000).
93. Wyatt, M.A. *et al*. *Staphylococcus aureus* nonribosomal peptide secondary metabolites regulate virulence. *Science* **329**, 294-296 (2010).

94. Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369–394 (2003).

95. Williams, K.P. *et al*. Phylogeny of *Gammaproteobacteria*. *J Bacteriol* **192**, 2305-2314 (2010).

96. Helfrich, E.J.N. & Piel, J. Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat Prod Rep* **in press** (2015).

97. Kalb, D., Lackner, G. & Hoffmeister, D. Functional and phylogenetic divergence of fungal adenylate-forming reductases. *Appl Environ Microbiol* **80**, 6175-6183 (2014).

98. Cazalet, C. *et al*. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* **36**, 1165-1173 (2004).

99. Shevchuk, O. *et al*. Polyketide synthase (PKS) reduces fusion of *Legionella pneumophila*-containing vacuoles with lysosomes and contributes to bacterial competitiveness during infection. *Int J Med Microbiol* **304**, 1169-1181 (2014).

100. Reith, F. *et al*. Mechanisms of gold biomineralization in the bacterium *Cupriavidus metallidurans*. *Proc Natl Acad Sci USA* **106**, 17757-17762 (2009).

## Appendix 1

# Supplementary Information

## An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products

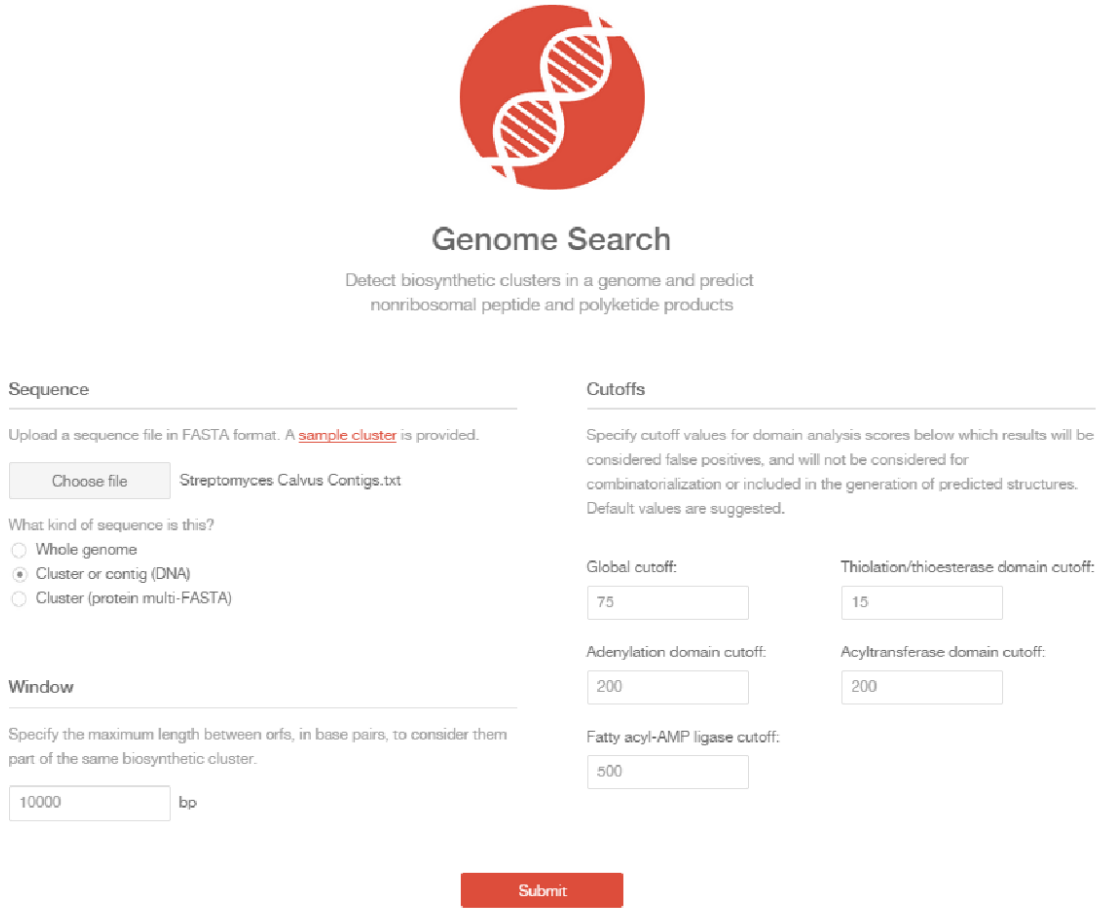
Chad W. Johnston[1]*, Michael A. Skinnider[1]*, Morgan A. Wyatt[1], Xiang Li[1], Michael R. M. Ranieri[1], Lian Yang[2], David L. Zechel[3], Bin Ma[2], & Nathan A. Magarvey[1]

[1] Department of Biochemistry & Biomedical Sciences; Chemistry & Chemical Biology; M. G. DeGroote Institute for Infectious Disease Research; McMaster University, Hamilton, ON, L8N 3Z5

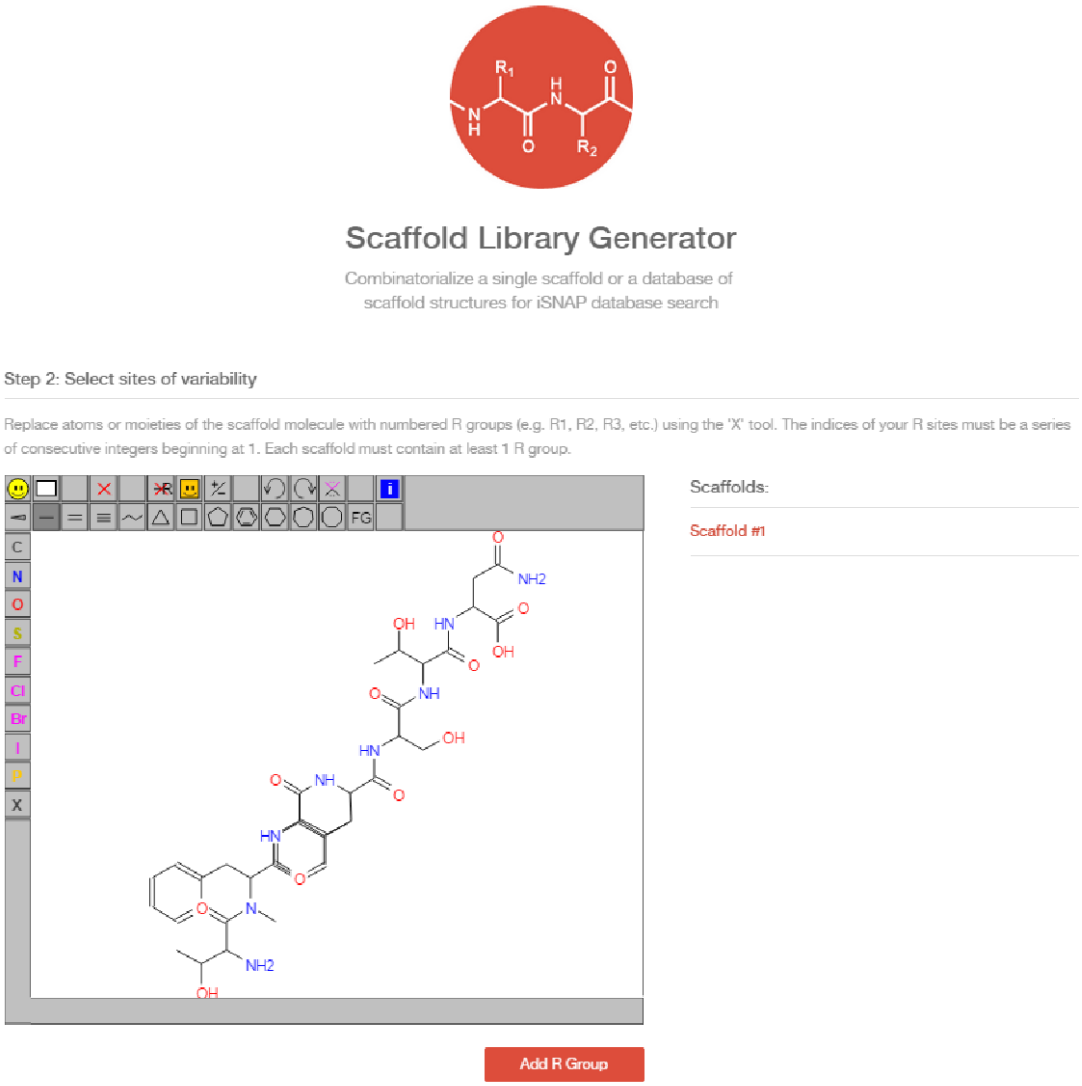
[2] The David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1

[3] Department of Chemistry; Queens University, Kingston, ON, K7L 3N6

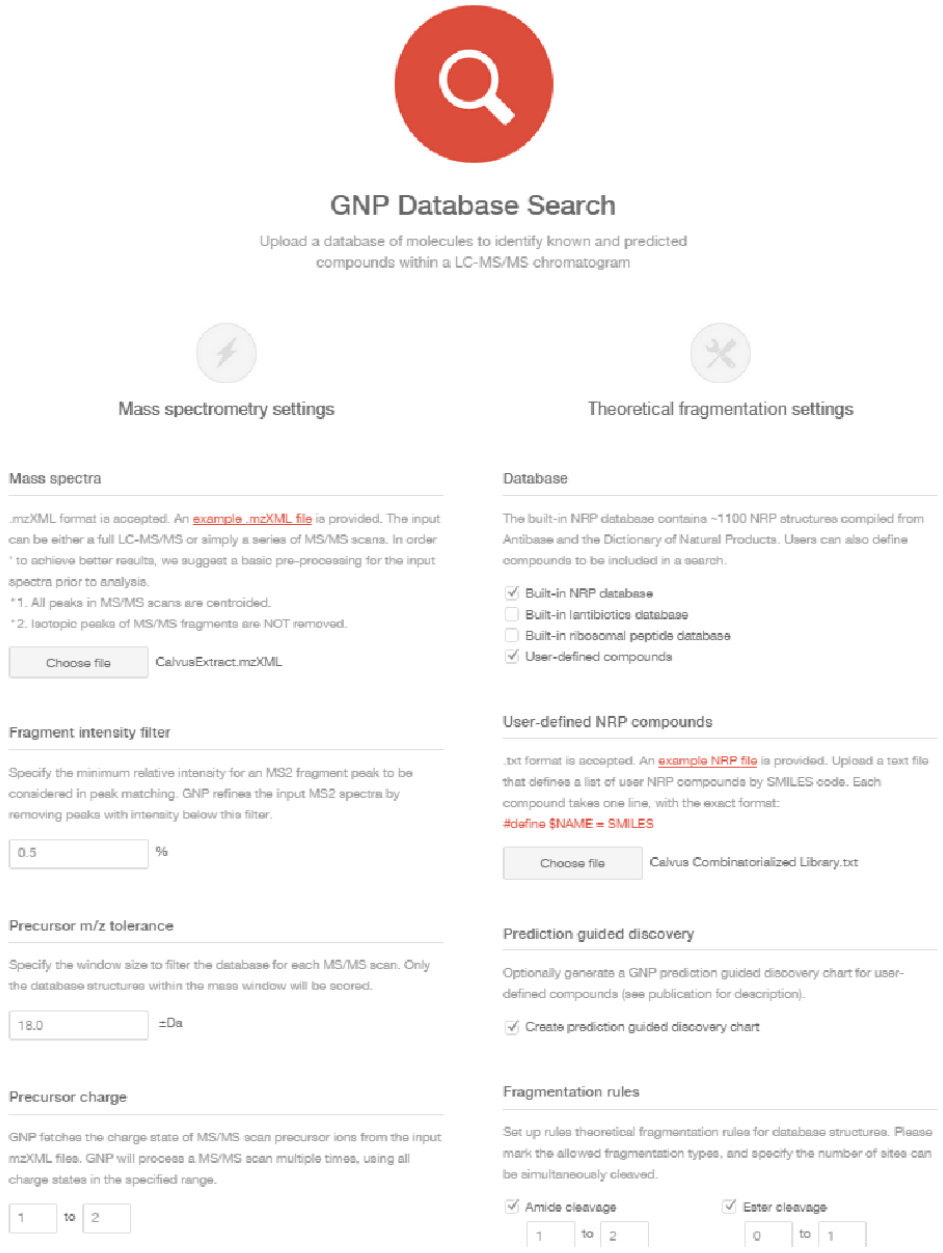
*These authors contributed equally to this work.




**Supplementary Figure 1** GNP's genome search functionality**.** As a standard initiating step for the Genomes-to-Natural products Platform, users are prompted to upload gene sequence files in FASTA format. These genomes are analyzed by GNP, which detects PKS and NRPS gene clusters and generates predicted structures automatically.




**Supplementary Figure 2** GNP's scaffold library generator. Following either the automated generation of predicted scaffolds, or the uploading of a user-defined scaffold, predicted structures are combinatorialized at a number of sites defined by R-groups. Combinatorialized libraries are forwarded onwards for use in detecting predicted structures within LC-MS/MS data.




**Supplementary Figure 3** GNP's database search functionality. Following the generation of a library of predicted structures, GNP can automatically detect compounds in relevant LC-MS/MS data. Uploaded mzXML files can be analyzed with predicted structure databases (user-defined compounds), using our established NRP database as a dummy to provide statistical scoring required for assessing hits. Various parameters can be adjusted based on the user confidence in the predicted structures, such as precursor m/z tolerance, which sets a window around a detected ion within which candidate predictions can be assessed for fragment ion matching.

Results

Your results will expire on December 21, 2014, and will be deleted.

Upload — Done.

Database Search — Generating report... Report #3070 of 3070

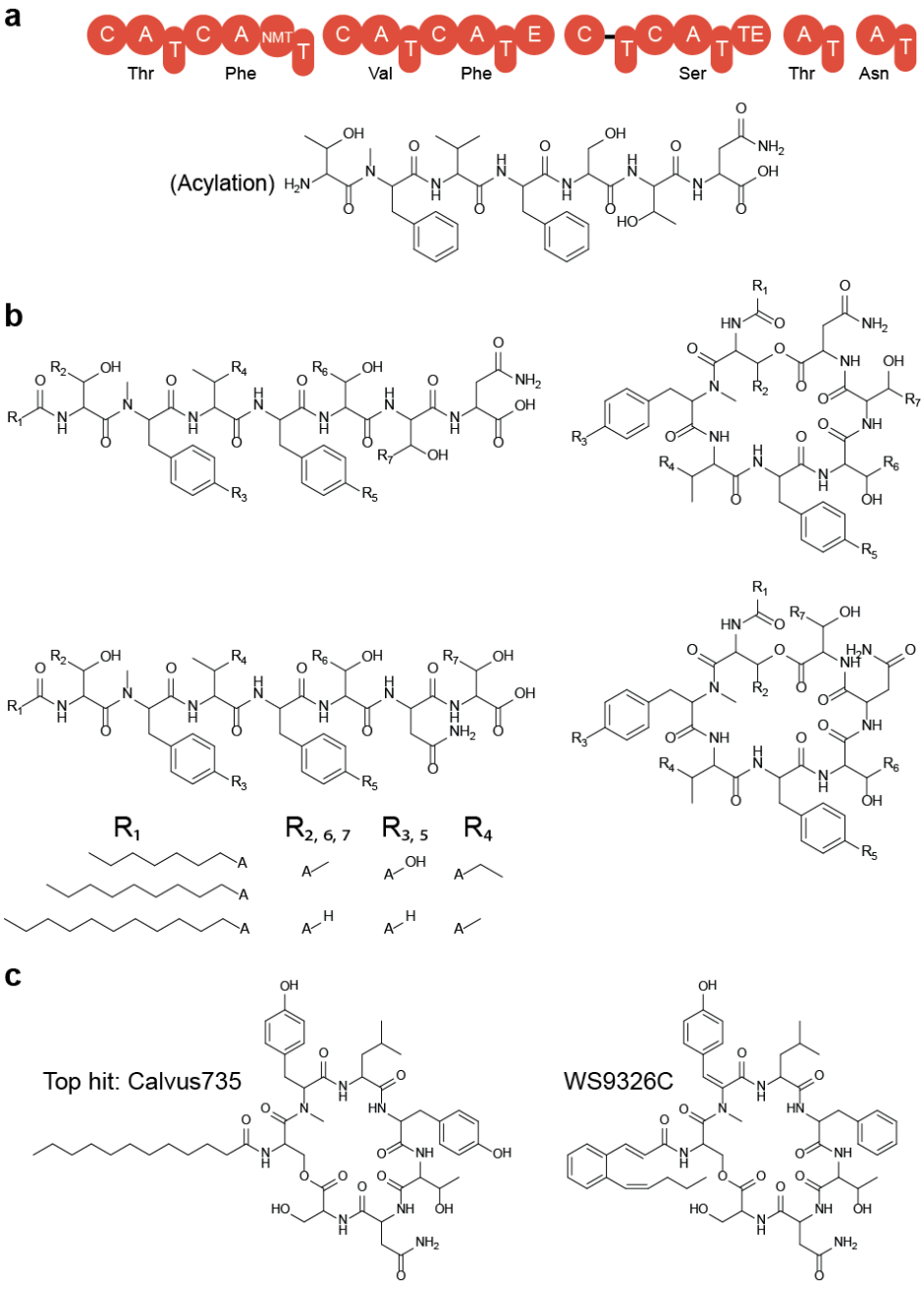
Reports — Download reports: Download NRP search report; Download matched fragment report



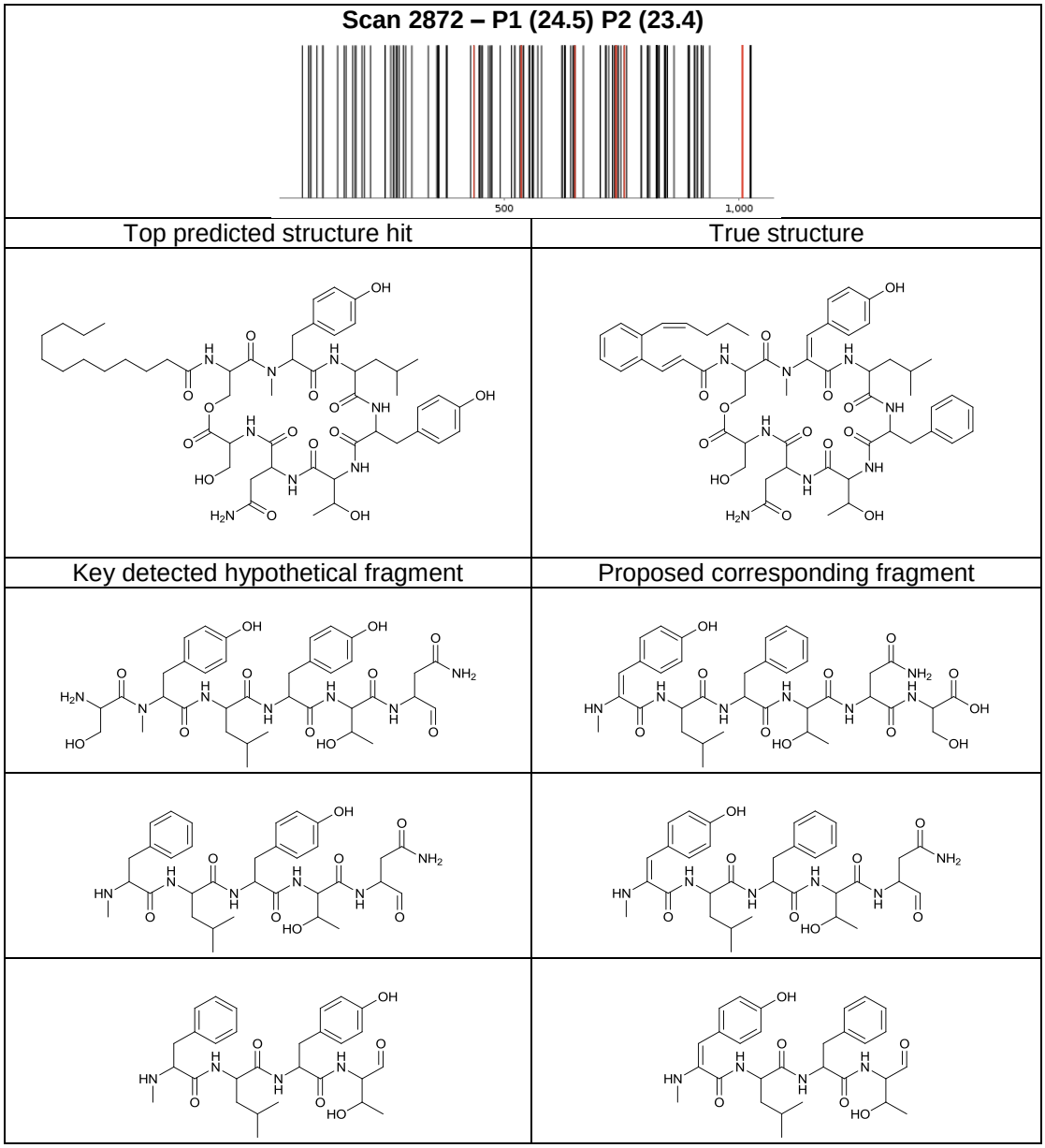
Click labels to sort.

| Scan No. | RT | Precursor M/Z | Precursor Charge | Precursor Mass | Result | SMILES | Mass | Matched Peaks | P1 | P2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1088 | 17.33 | 508.51 | 2 | 1015.01 | user_#define 1.1_2.1_4.1_6.1_7.1_4.1_5.1_5.1 | C%13(C(C(NC(C(NC(C(NC(C%16O)C(NC(C(O)=O)CC(N)=O)=O)=O)C%15O)=O)CC1=CC=C%14C=C1)=O)NC(C(N(C(C(C%11(O))NC%10=O)=O)C)CC2=CC=C%12C=C2)=O)C.C%10CCCCCCC.[H]%11.O%12.C%13C.[H] | 1012.55 | 10 | 13.4 | 25.6 |

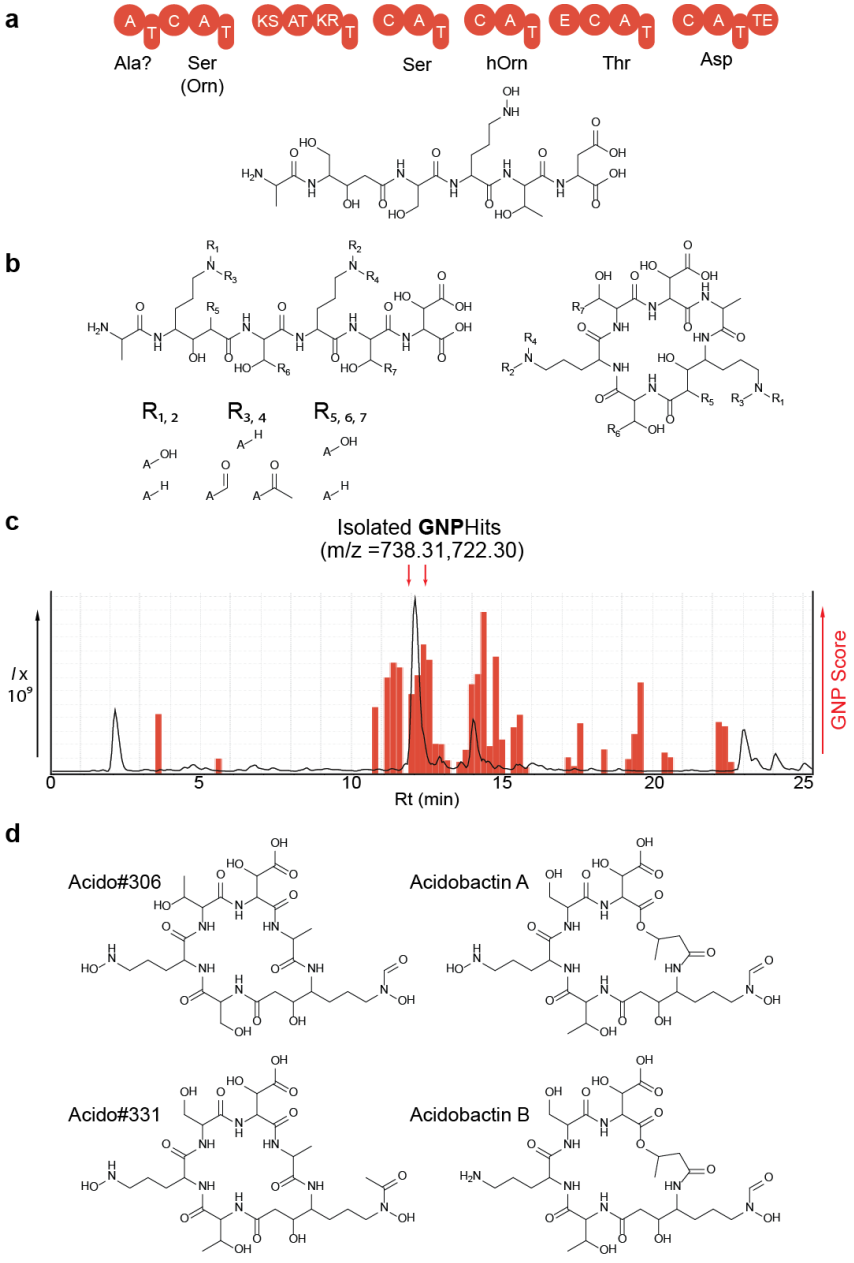
**Supplementary Figure 4** Results of GNP LC-MS/MS analysis**.** GNP analysis of LC-MS/MS data using a library of predicted structures yields a prediction guided discovery chart which indicates the frequency of user defined compounds identified as top-scoring hits over time. These compounds are also listed below in an Excel-compatible spreadsheet listing the relevant scan, retention time, precursor charge and mass, as well as the SMILES code and mass for the predicted structure, the number of matched fragment ions, and the statistical parameter scores. Structures of parent and fragment ions are also viewable online by moving the cursor over the relevant SMILES code, though this option must be selected in database search settings.




**Supplementary Figure 5** Prediction, combinatorialization, and detection of WS9326 metabolites. (**a**) GNP detected NRPS biosynthetic gene cluster from *Streptomyces calvus*, with predicted monomers listed and GNP-generated prediction shown. (**b**) Structures used for combinatorialization, including varied order of the two trans-acting A-T didomains. (**c**) Structure of the most frequent hit detected within the *S. calvus* LC-MS/MS data (*left*), as well as the true associated metabolite, WS9326C (*right*).

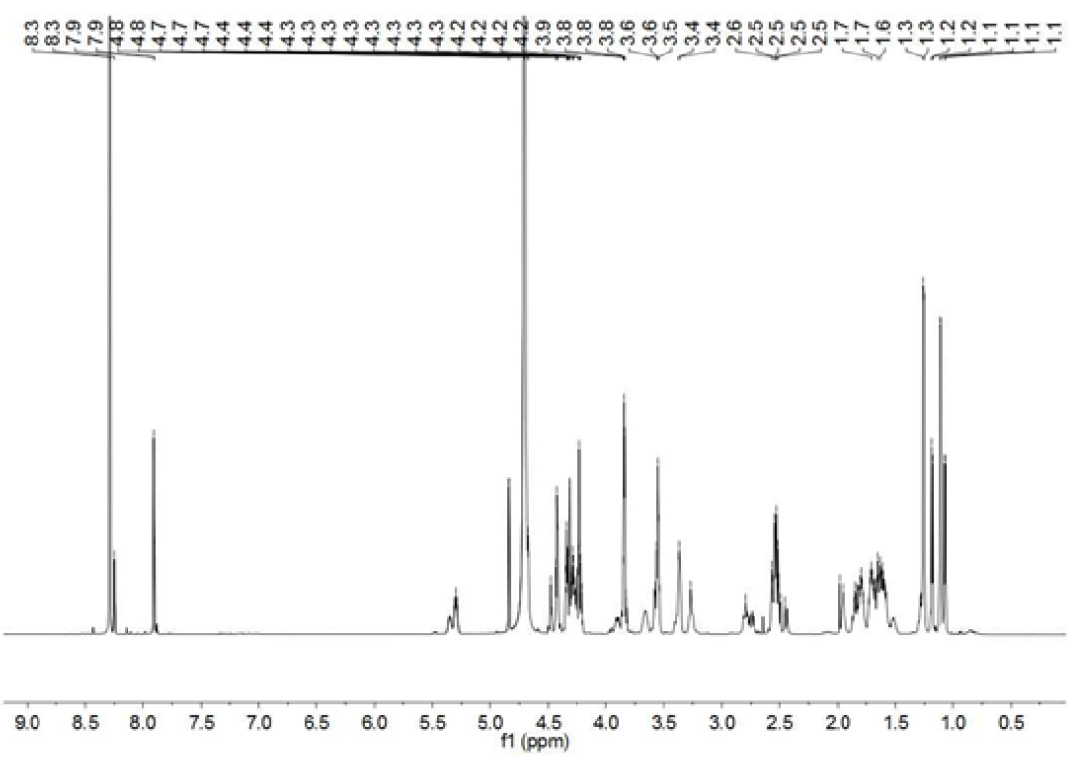



**Supplementary Figure 6** Sample matched fragments between a predicted structure hit and WS9326C. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of WS9326C were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.

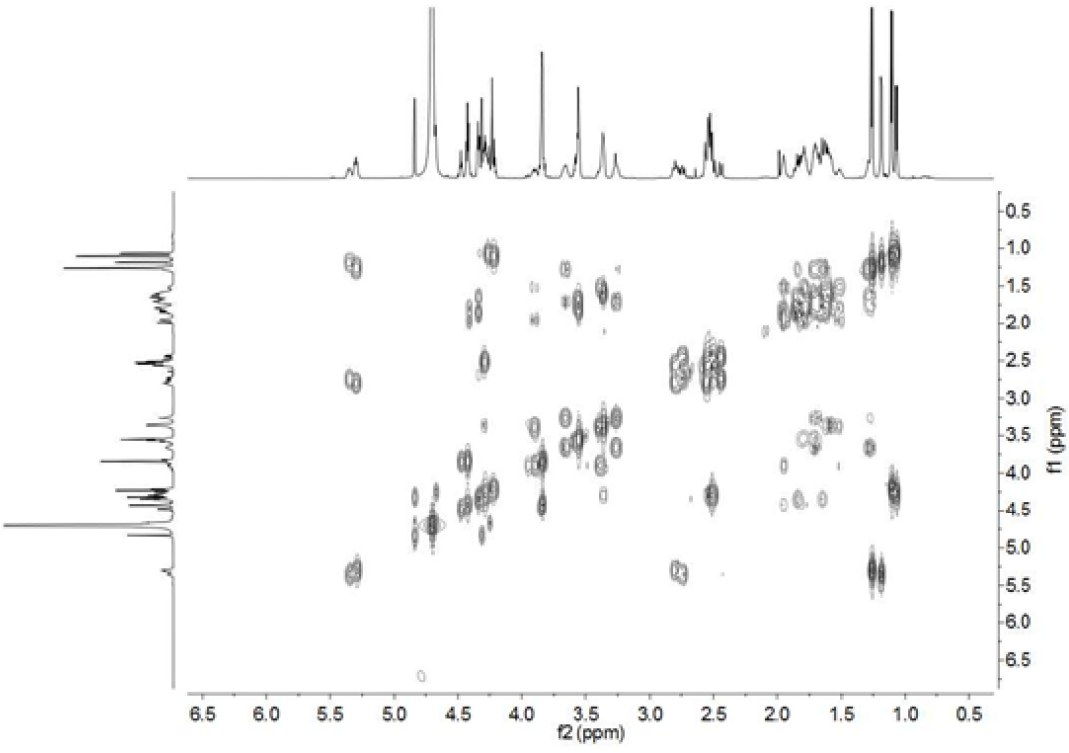


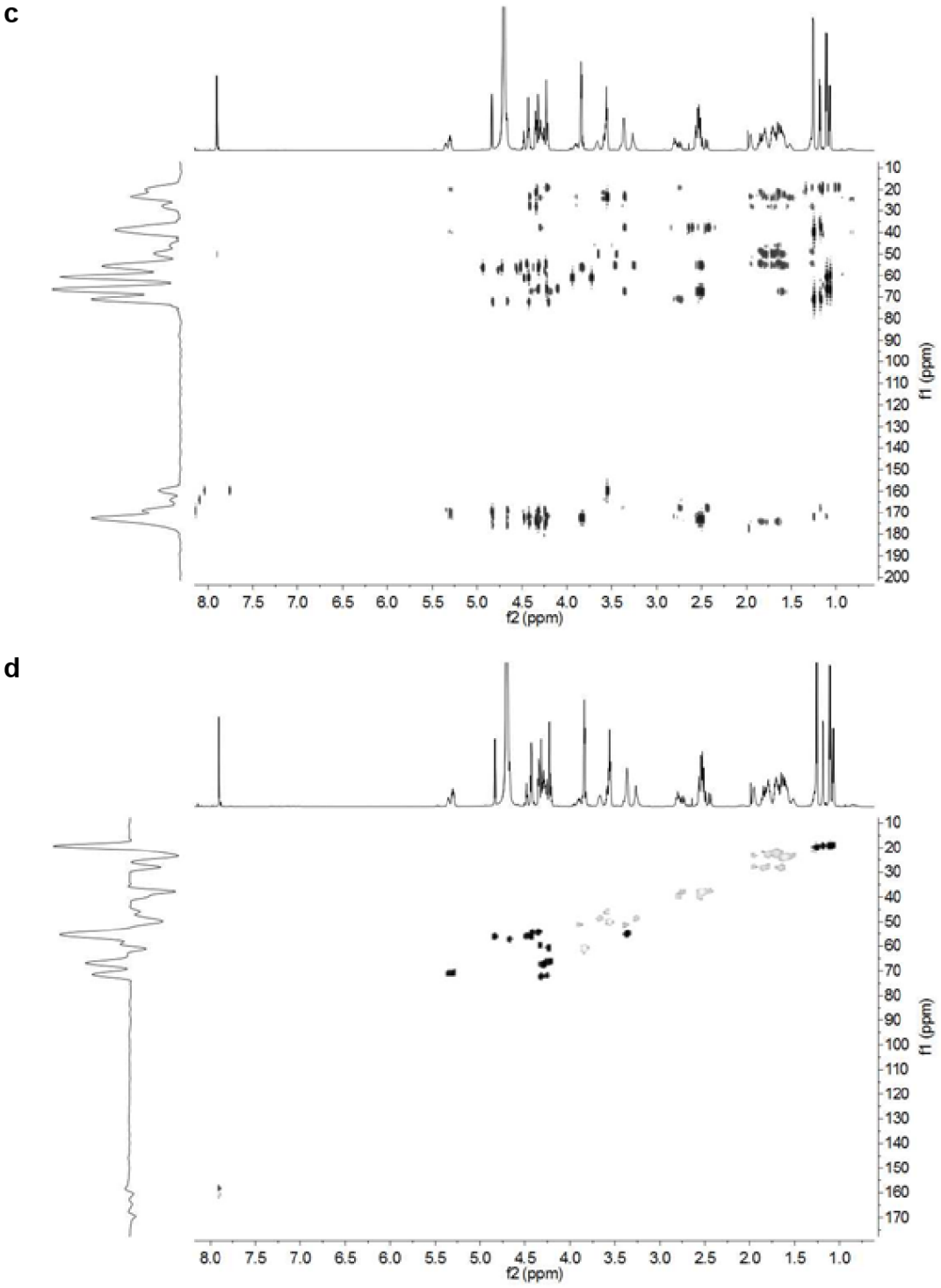

**Supplementary Figure 7** Prediction, combinatorialization, and detection of acidobactin metabolites. (**a**) GNP detected NRPS biosynthetic gene cluster from *Acidovorax citrulli*, with predicted monomers listed and GNP-generated prediction shown. (**b**) Structures used for combinatorialization, taking into account the anticipated ornithine incorporation at position 2. (**c**) Prediction guided discovery chart indicating detected natural products related to the combinatorialized structure library. (**d**) Structures of the most frequent hits associated with the major metabolites from the *A. citrulli* LC-MS/MS data (*left*), as well as the true associated metabolites, acidobactin A and B (*right*).

**a**

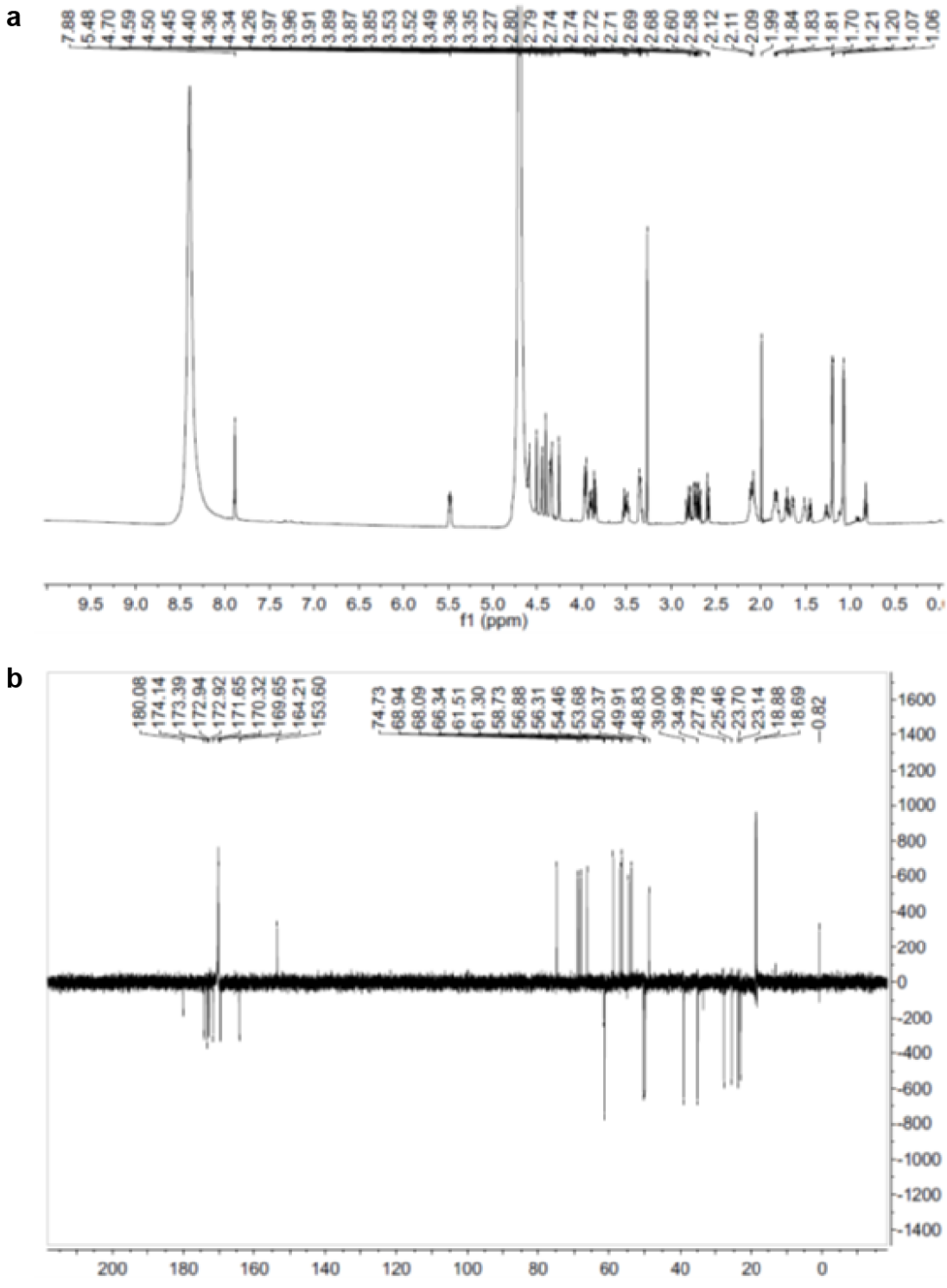



**b**

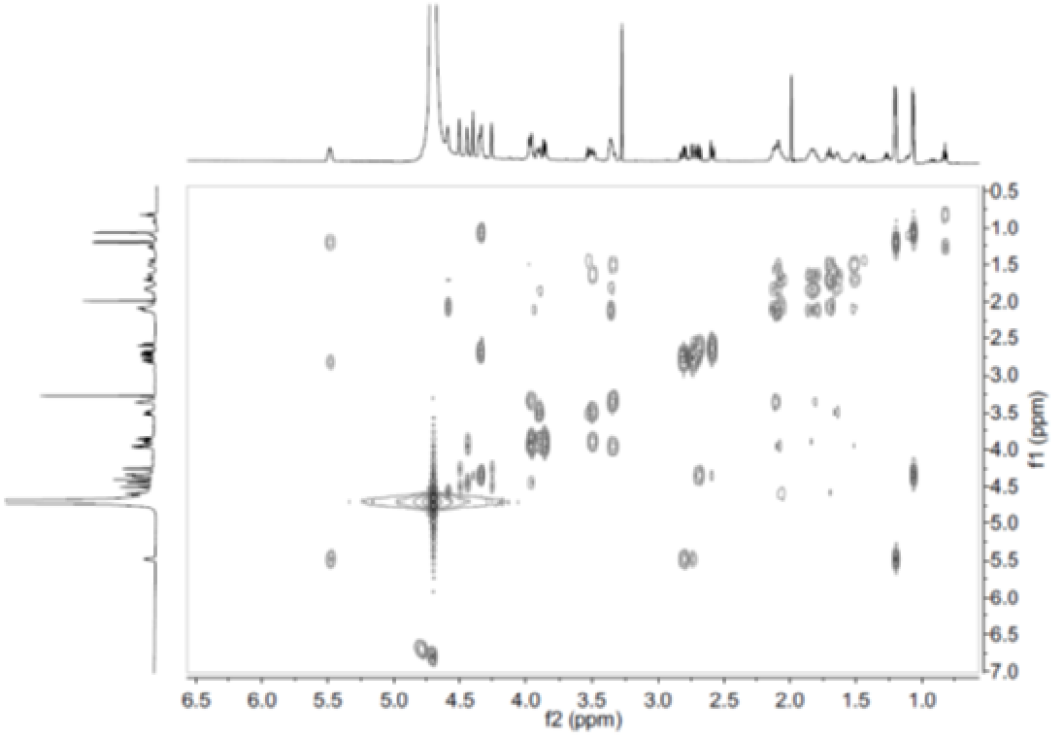




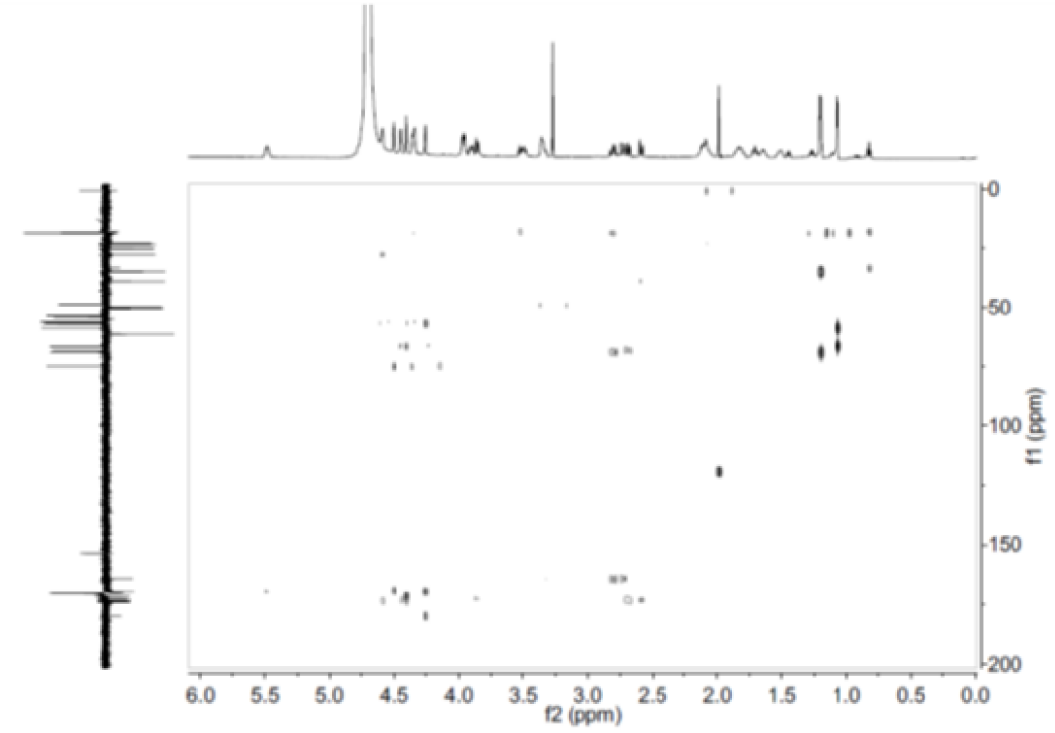

**Supplementary Figure 8** NMR analysis of acidobactin A (**2**) in $D_2O$. (**a**) $^{1}H$ NMR spectrum. (**b**) $^{1}H$-$^{1}H$ COSY spectrum. (**c**) $^{1}H$-$^{13}C$ HMBC spectrum. (**d**) $^{1}H$-$^{13}C$ HMQC spectrum.

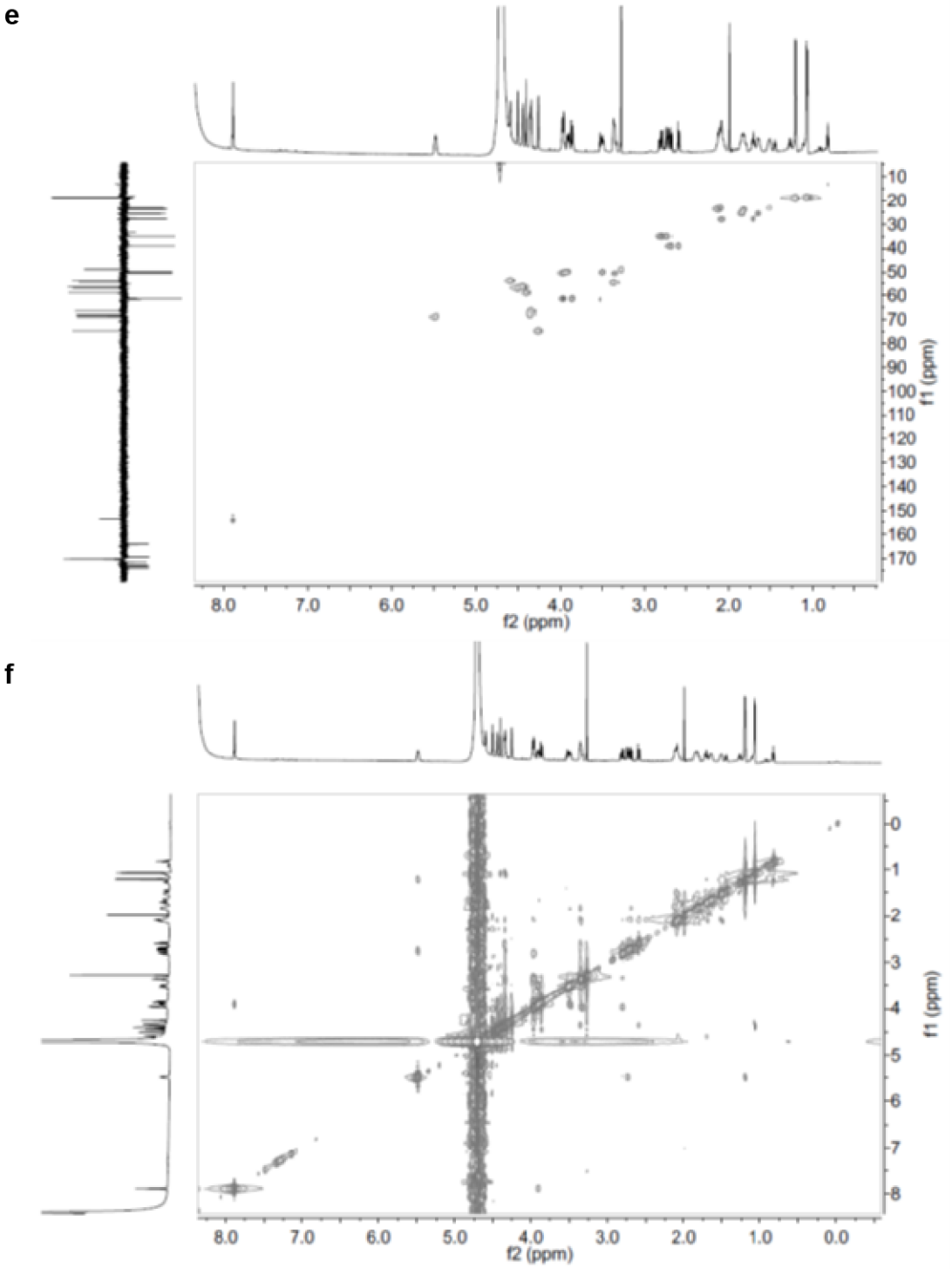

a
7.88
5.48
4.70
4.59
4.50
4.45
4.40
4.36
4.34
4.26
3.97
3.96
3.91
3.89
3.87
3.85
3.53
3.52
3.49
3.36
3.35
3.27
2.80
2.79
2.74
2.74
2.72
2.71
2.69
2.68
2.60
2.58
2.12
2.11
2.09
1.99
1.84
1.83
1.81
1.70
1.21
1.20
1.07
1.06
9.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5
f1 (ppm)
b
180.08
174.14
173.39
172.94
172.92
171.65
170.32
169.65
164.21
153.60
74.73
68.94
68.09
66.34
61.51
61.30
58.73
56.88
56.31
54.46
53.68
50.37
49.91
48.83
39.00
34.99
27.78
25.46
23.70
23.14
18.88
18.69
0.82
1600
1400
1200
1000
800
600
400
200
0
-200
-400
-600
-800
-1000
-1200
-1400
200 180 160 140 120 100 80 60 40 20 0

**c**

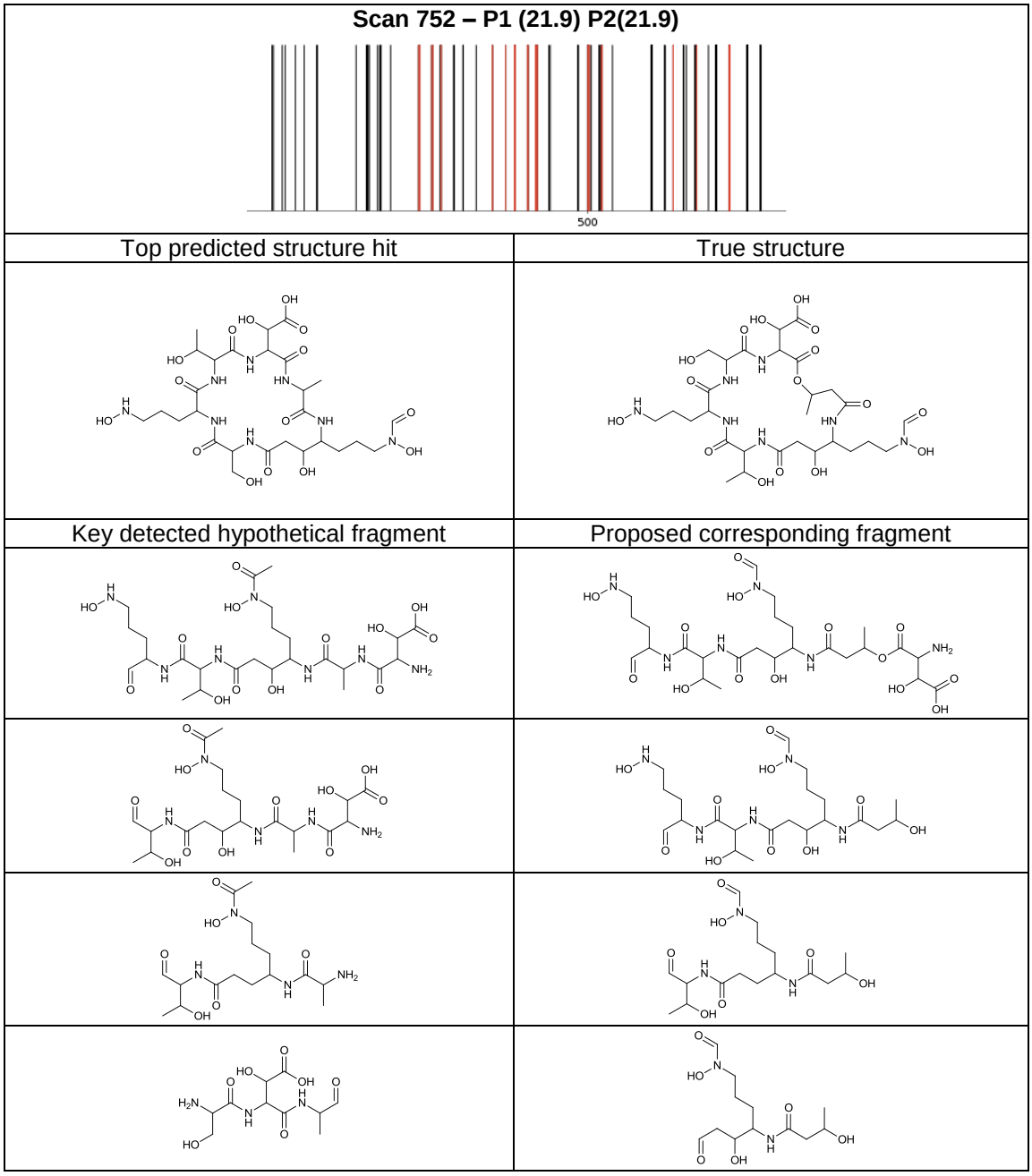



**d**

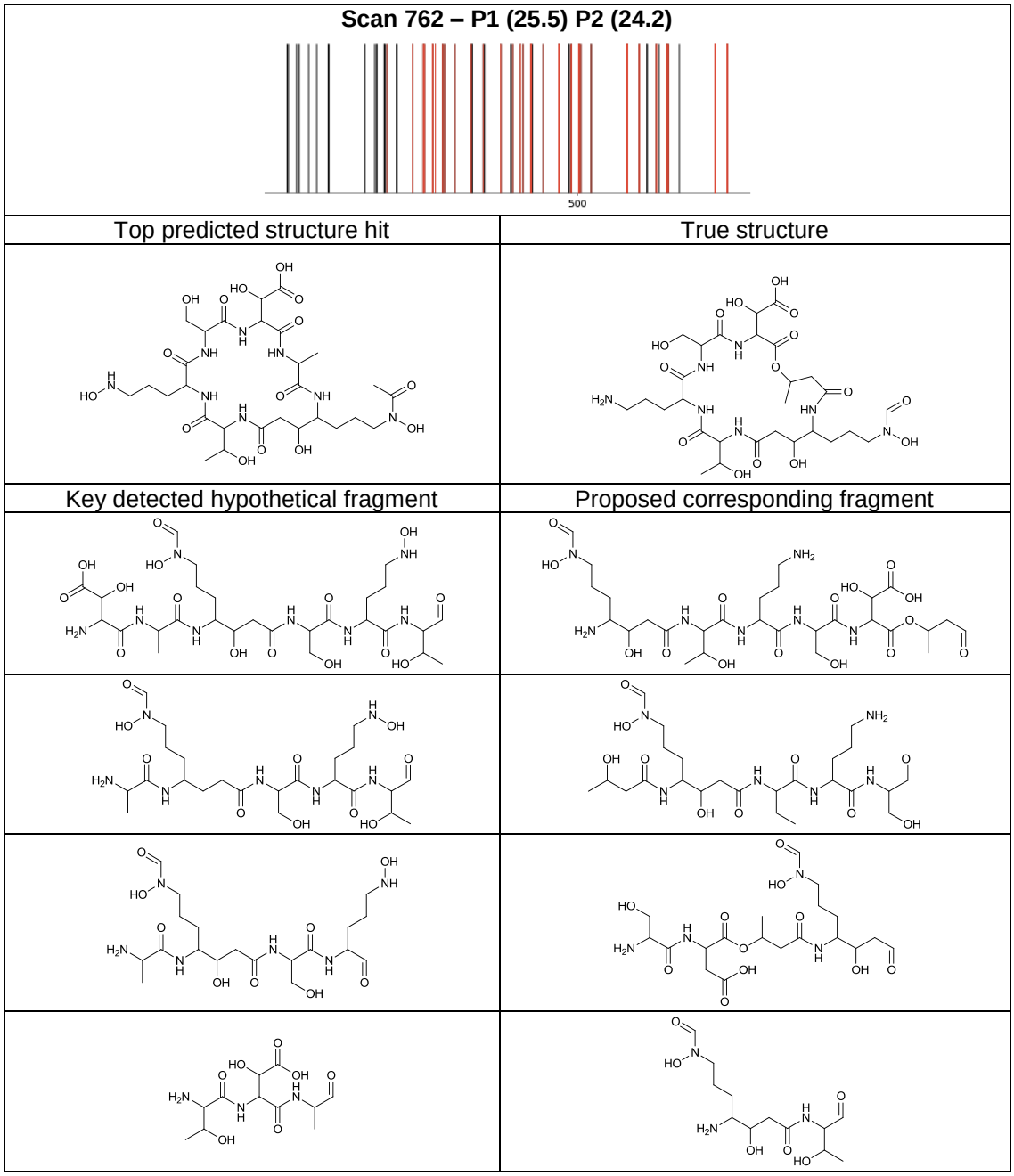




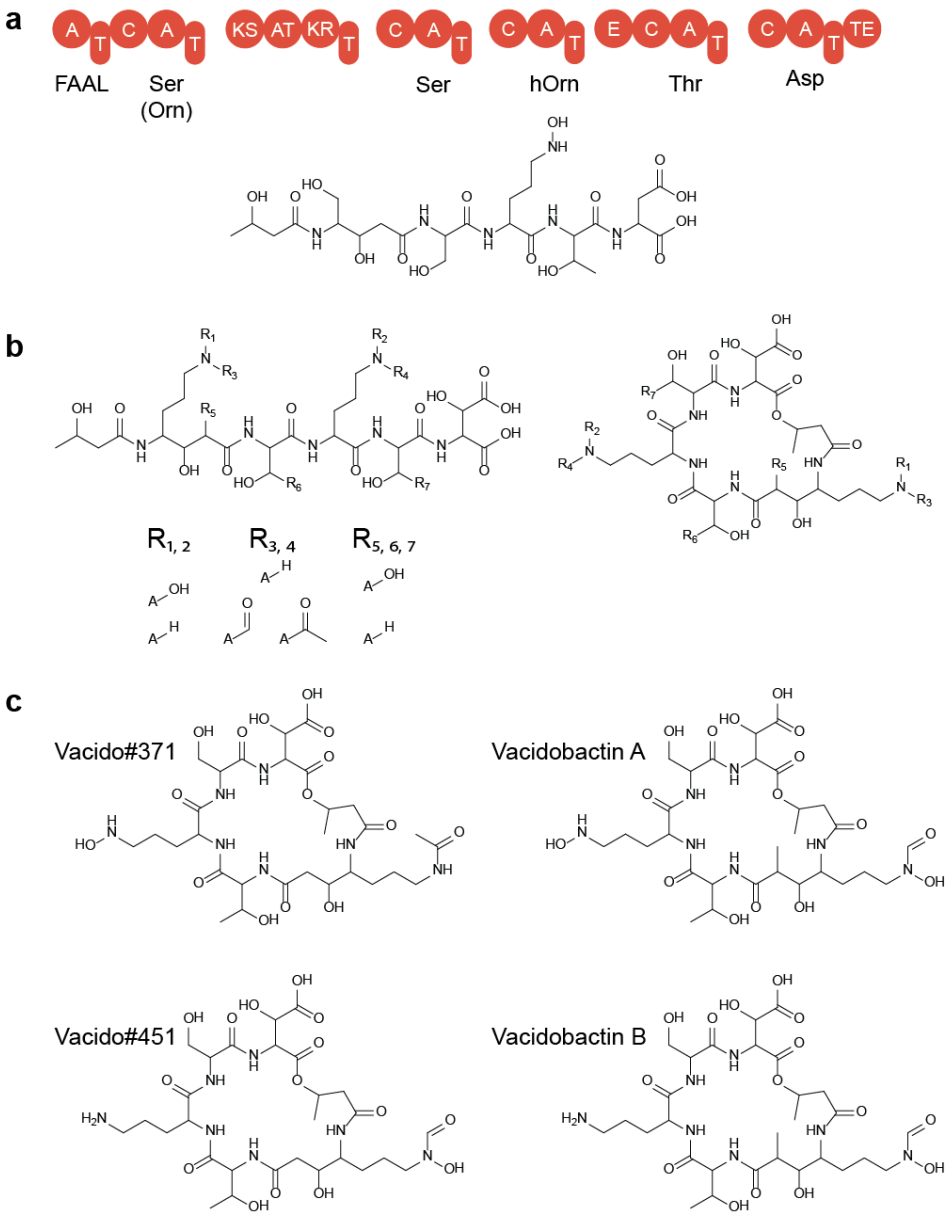

**Supplementary Figure 9** NMR analysis of acidobactin B (**3**) in $D_2O$. (**a**) $^{1}H$ NMR spectrum. (**b**) $^{13}C$ DEPTq spectrum. (**c**) $^{1}H$-$^{1}H$ COSY spectrum. (**d**) $^{1}H$-$^{13}C$ HMBC spectrum. (**e**) $^{1}H$-$^{13}C$ HMQC spectrum. (**f**) $^{1}H$-$^{1}H$ ROESY spectrum.



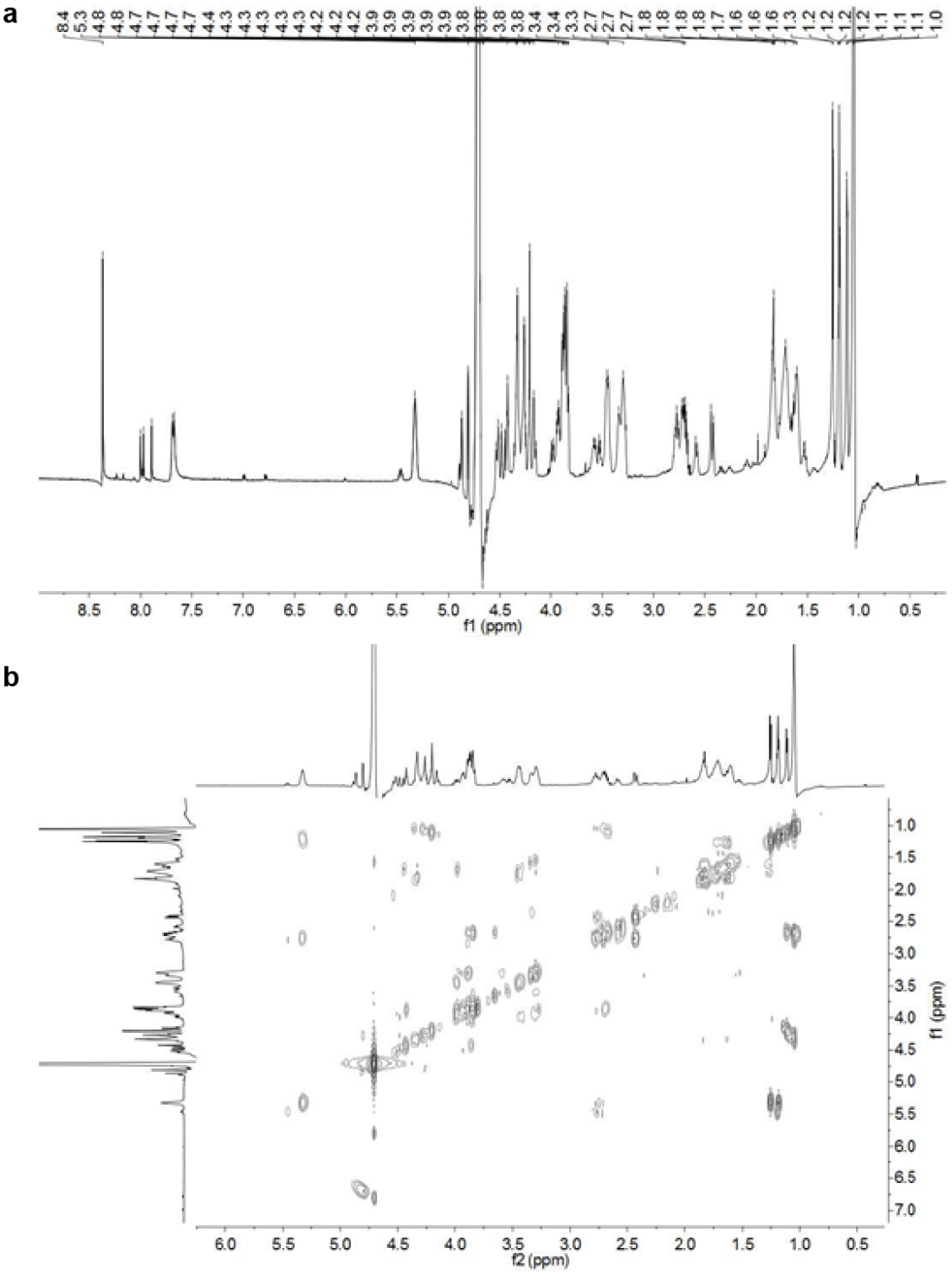

**Supplementary Figure 10** Sample matched fragments between a predicted structure hit and acidobactin A. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of acidobactin A were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.

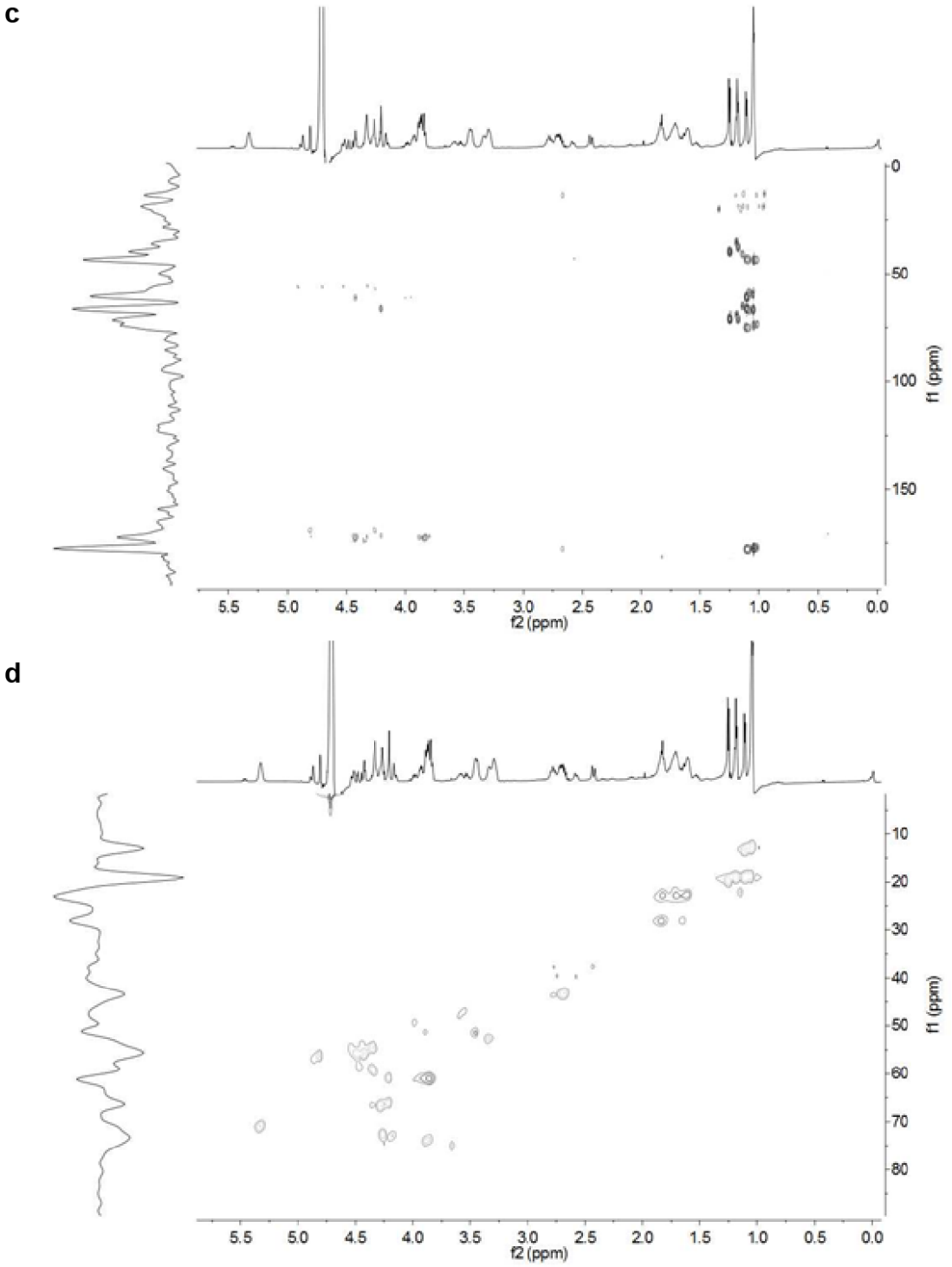



**Supplementary Figure 11** Sample matched fragments between a predicted structure hit and acidobactin B. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of acidobactin B were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.

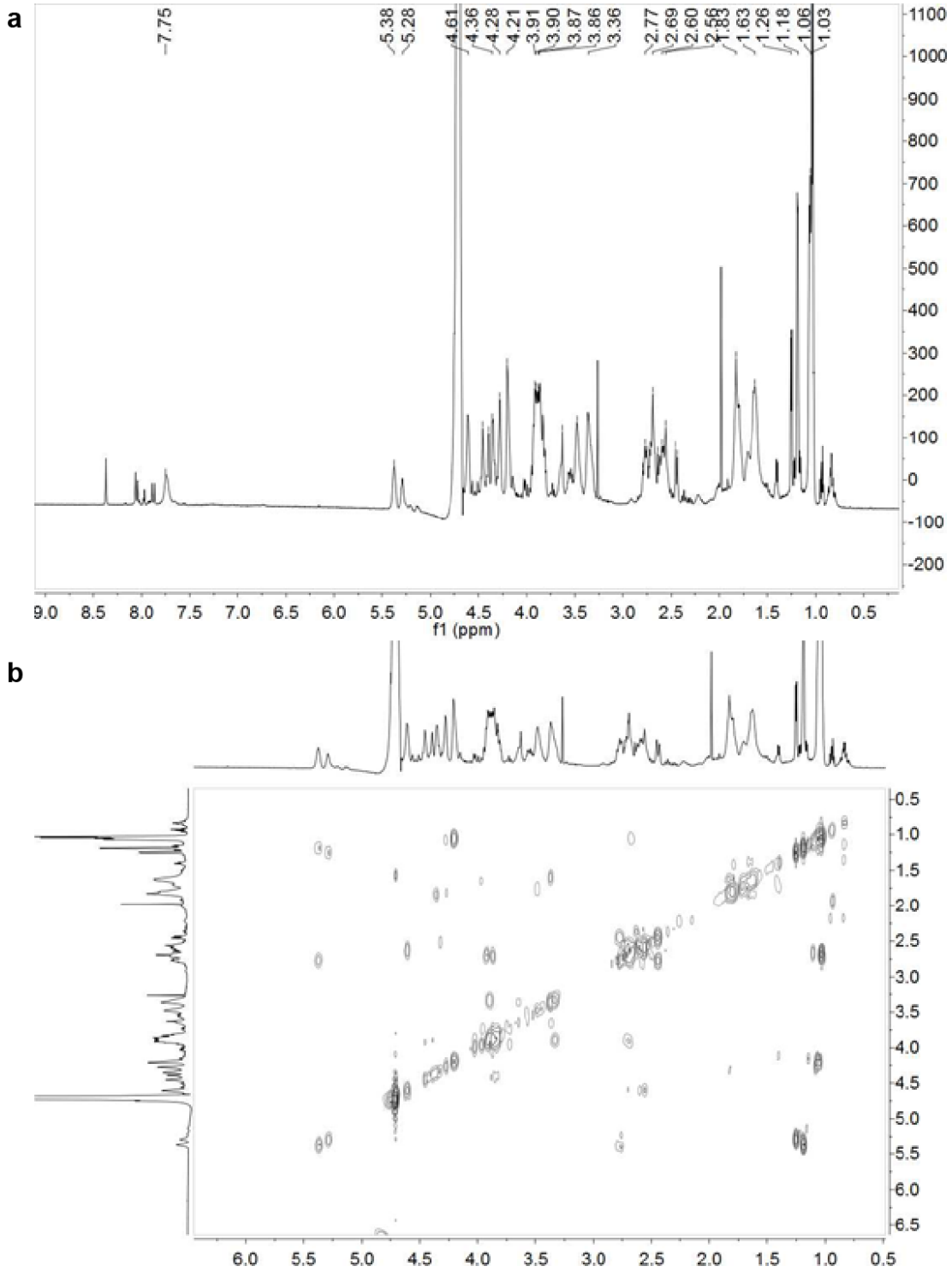


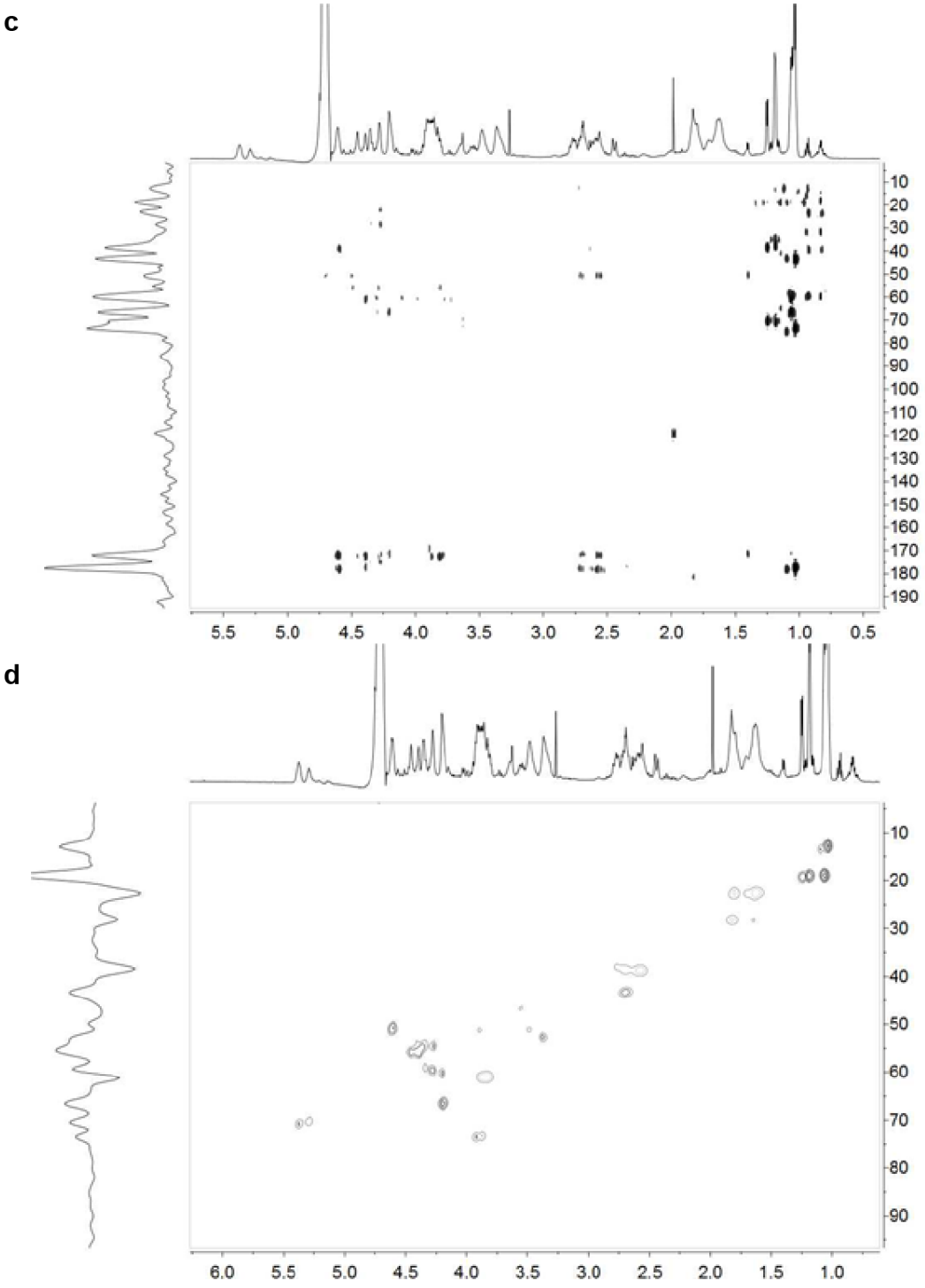

**Supplementary Figure 12** Prediction, combinatorialization, and detection of vacidobactin metabolites. (**a**) GNP detected NRPS biosynthetic gene cluster from *Variovorax paradoxus* S110, with predicted monomers listed and GNP-generated prediction shown. (**b**) Structures used for combinatorialization, taking into account the anticipated ornithine incorporation at position 2 (**c**) Structure of the most frequent hits detected within the *V. paradoxus* S110 LC-MS/MS data (*left*), as well as the true associated metabolites, vacidobactin A and B (*right*).


a
8.4 5.3 4.8 4.8 4.7 4.7 4.7 4.7 4.4 4.3 4.3 4.3 4.3 4.2 4.2 4.2 3.9 3.9 3.9 3.9 3.9 3.8 3.8 3.8 3.8 3.4 3.4 3.3 2.7 2.7 2.7 1.8 1.8 1.8 1.7 1.6 1.6 1.6 1.3 1.2 1.2 1.2 1.2 1.1 1.1 1.1 1.0
8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5
f1 (ppm)
b
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0
f1 (ppm)
6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5
f2 (ppm)

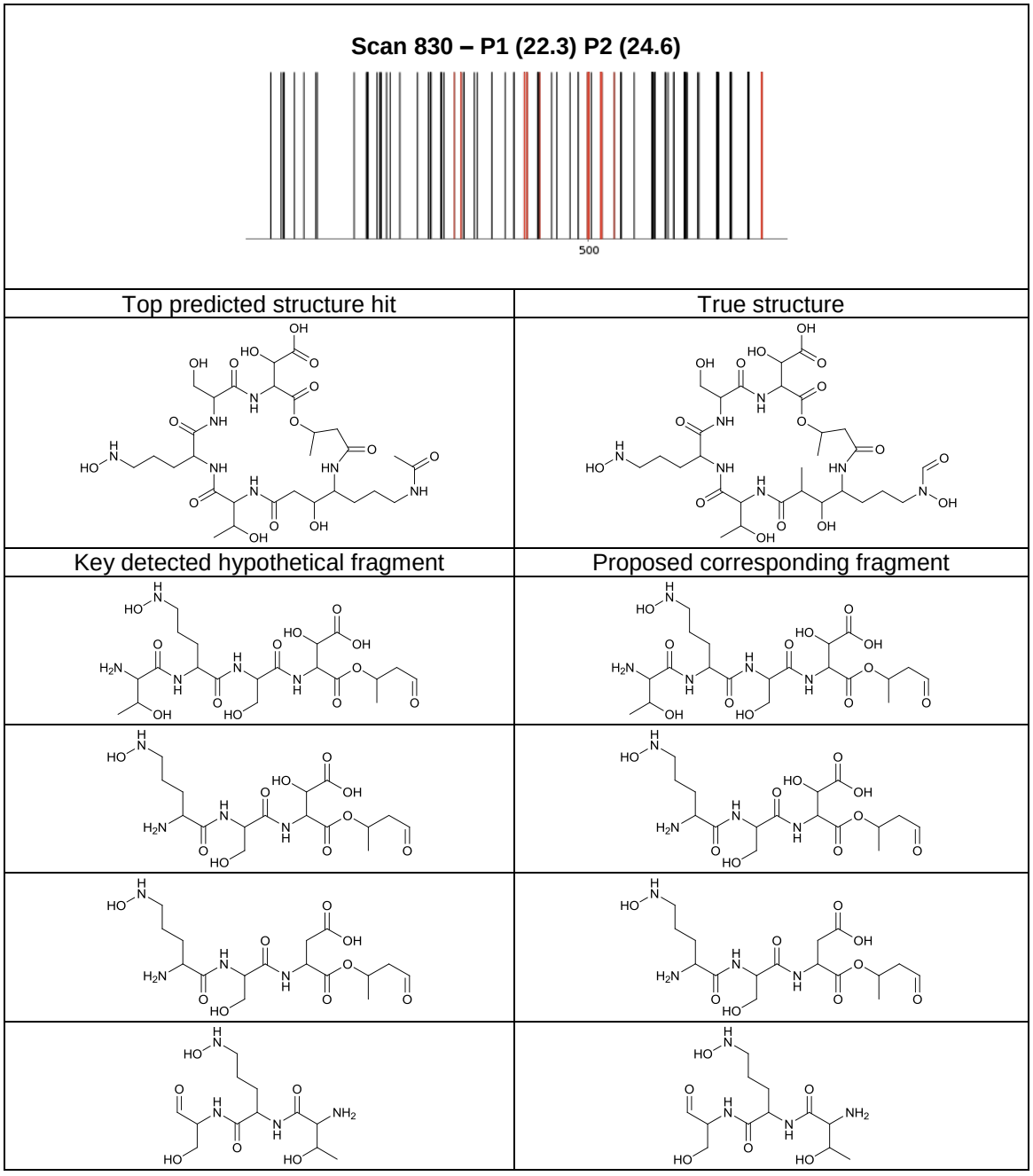



**Supplementary Figure 13** NMR analysis of vacidobactin A (**4**) in $D_2O$. (**a**) $^{1}H$ NMR spectrum. (b) $^{1}H$-$^{1}H$ COSY spectrum. (**c**) $^{1}H$-$^{13}C$ HMBC spectrum. (**d**) $^{1}H$-$^{13}C$ HMQC spectrum.

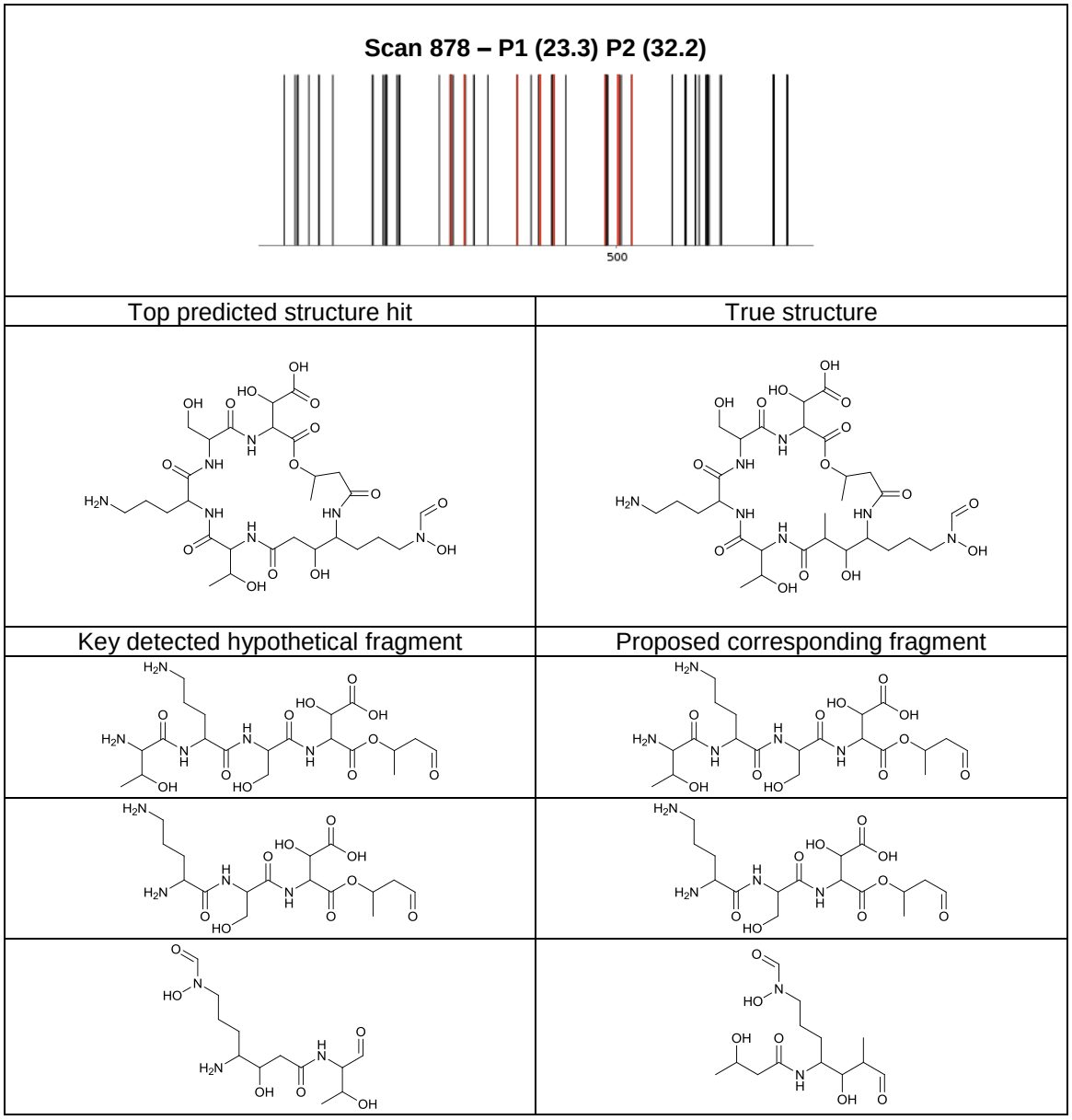

a
7.75
5.38
5.28
4.61
4.36
4.28
4.21
3.91
3.90
3.87
3.86
3.36
2.77
2.69
2.60
2.56
1.83
1.63
1.26
1.18
1.06
1.03
f1 (ppm)
b



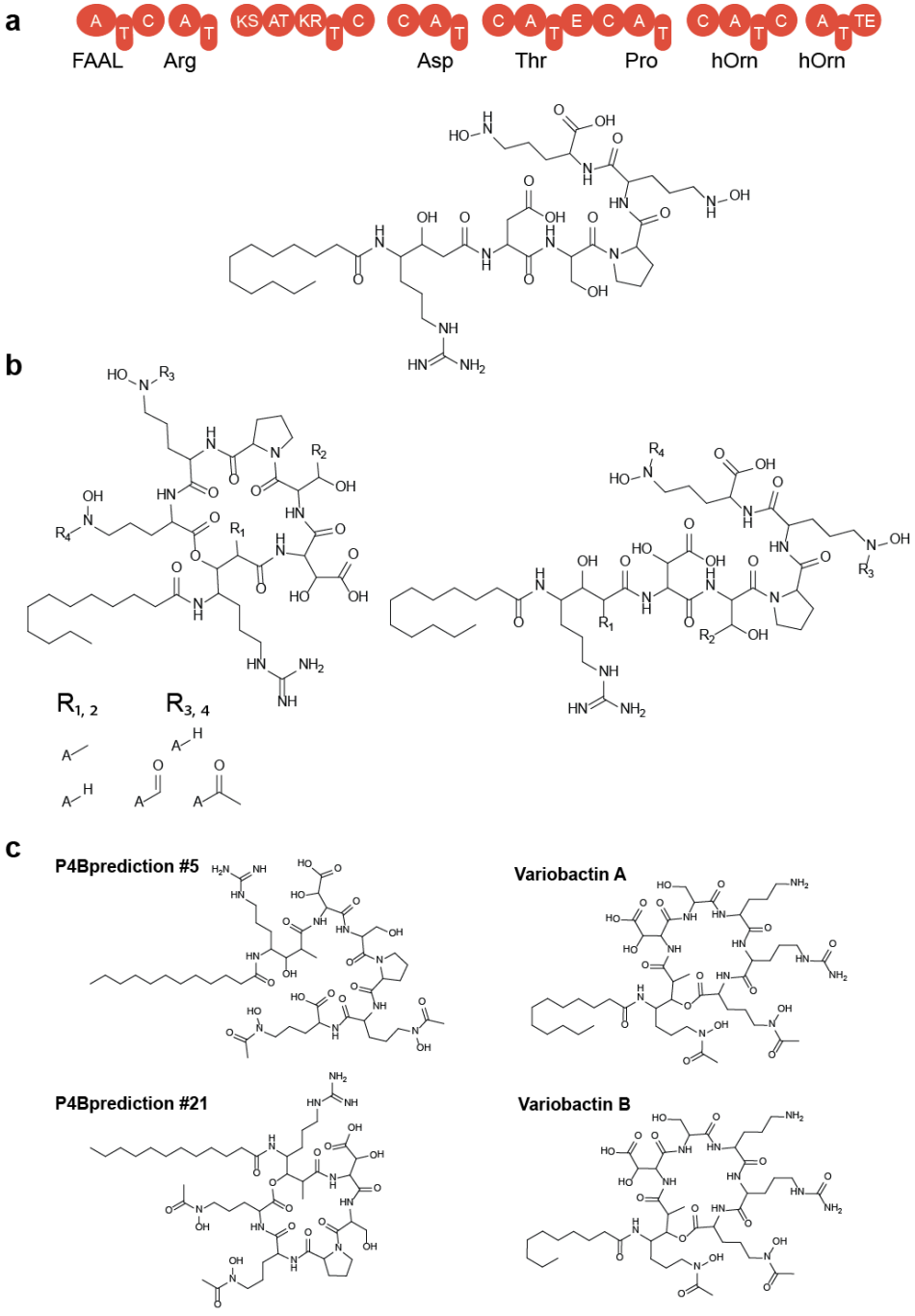

**Supplementary Figure 14** NMR analysis of vacidobactin B (**5**) in $D_2O$. (**a**) $^{1}H$ NMR spectrum. (**b**) $^{1}H$-$^{1}H$ COSY spectrum. (**c**) $^{1}H$-$^{13}C$ HMBC spectrum. (**d**) $^{1}H$-$^{13}C$ HMQC spectrum.



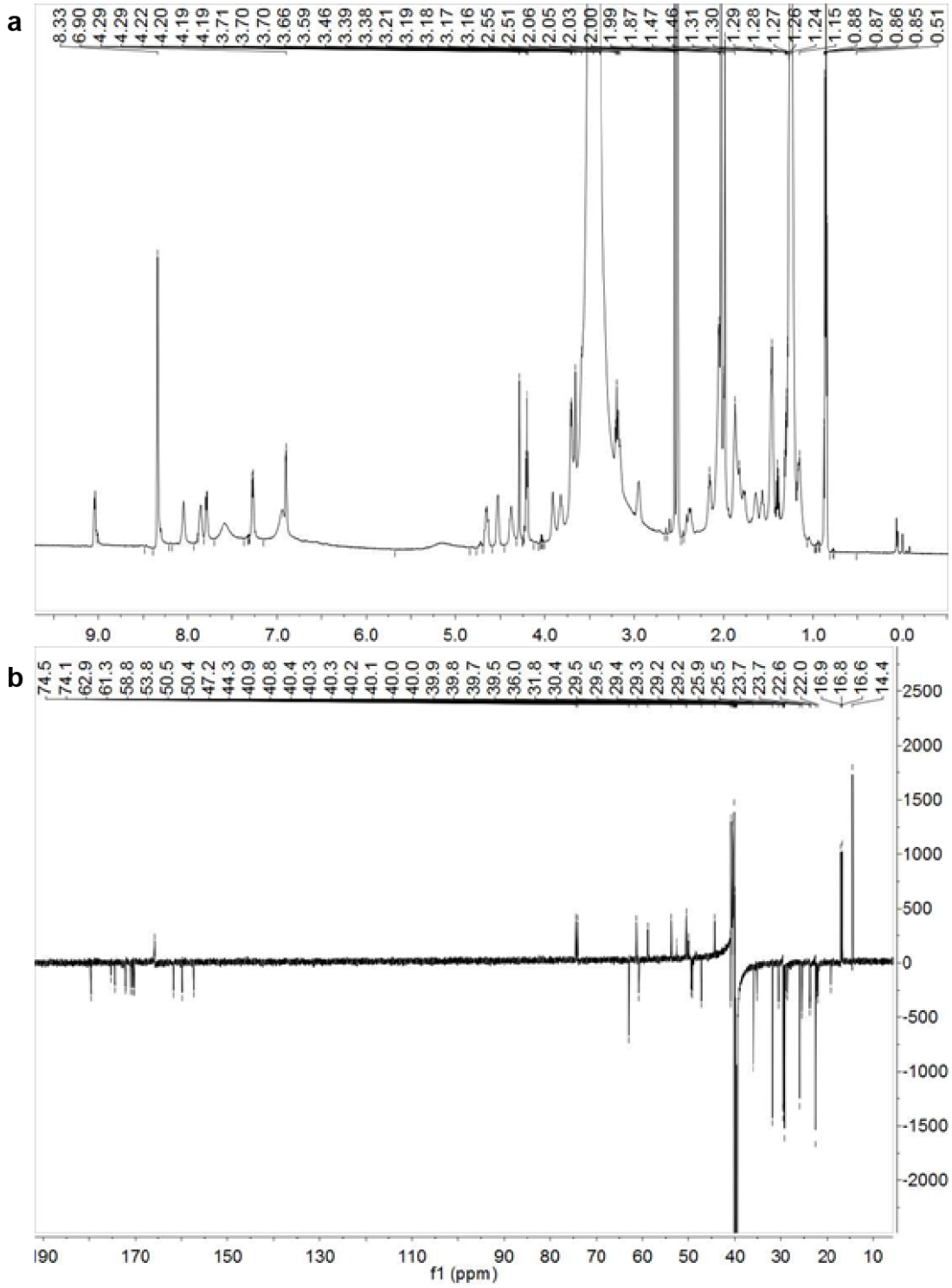

**Supplementary Figure 15** Sample matched fragments between a predicted structure hit and vacidobactin A. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of vacidobactin A were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.

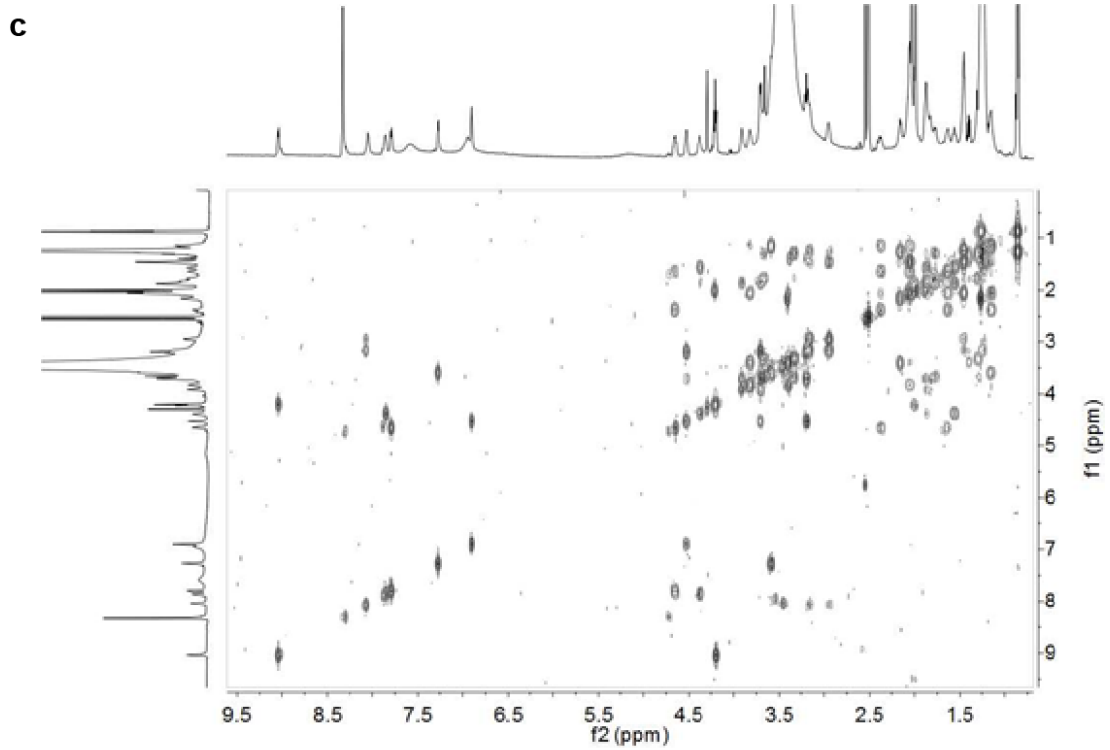


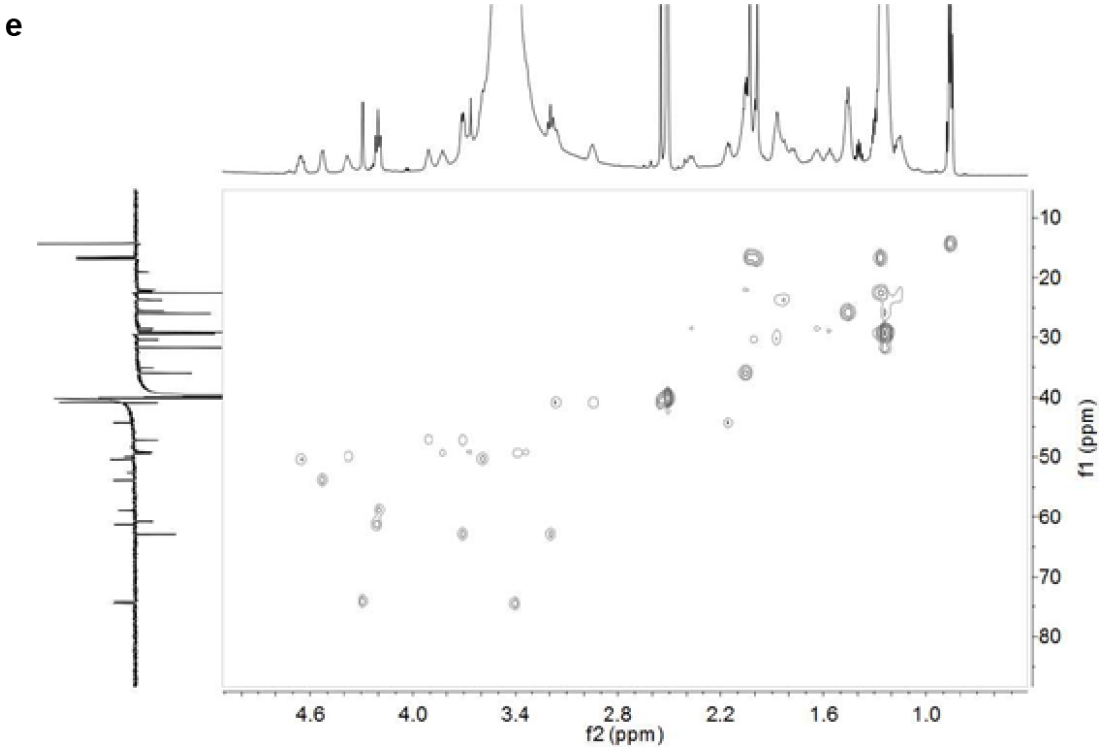

**Supplementary Figure 16** Sample matched fragments between a predicted structure hit and vacidobactin B. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of vacidobactin B were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.




**Supplementary Figure 17** Prediction, combinatorialization, and detection of variobactin metabolites. (**a**) GNP detected NRPS biosynthetic gene cluster from *Variovorax paradoxus* P4B, with predicted monomers listed and GNP-generated prediction shown. (**b**) Structures used for combinatorialization. (**c**) Structure of the most frequent hits detected within the *V. paradoxus* P4B LC-MS/MS data (*left*), as well as the true associated metabolites, variobactin A and B (*right*).

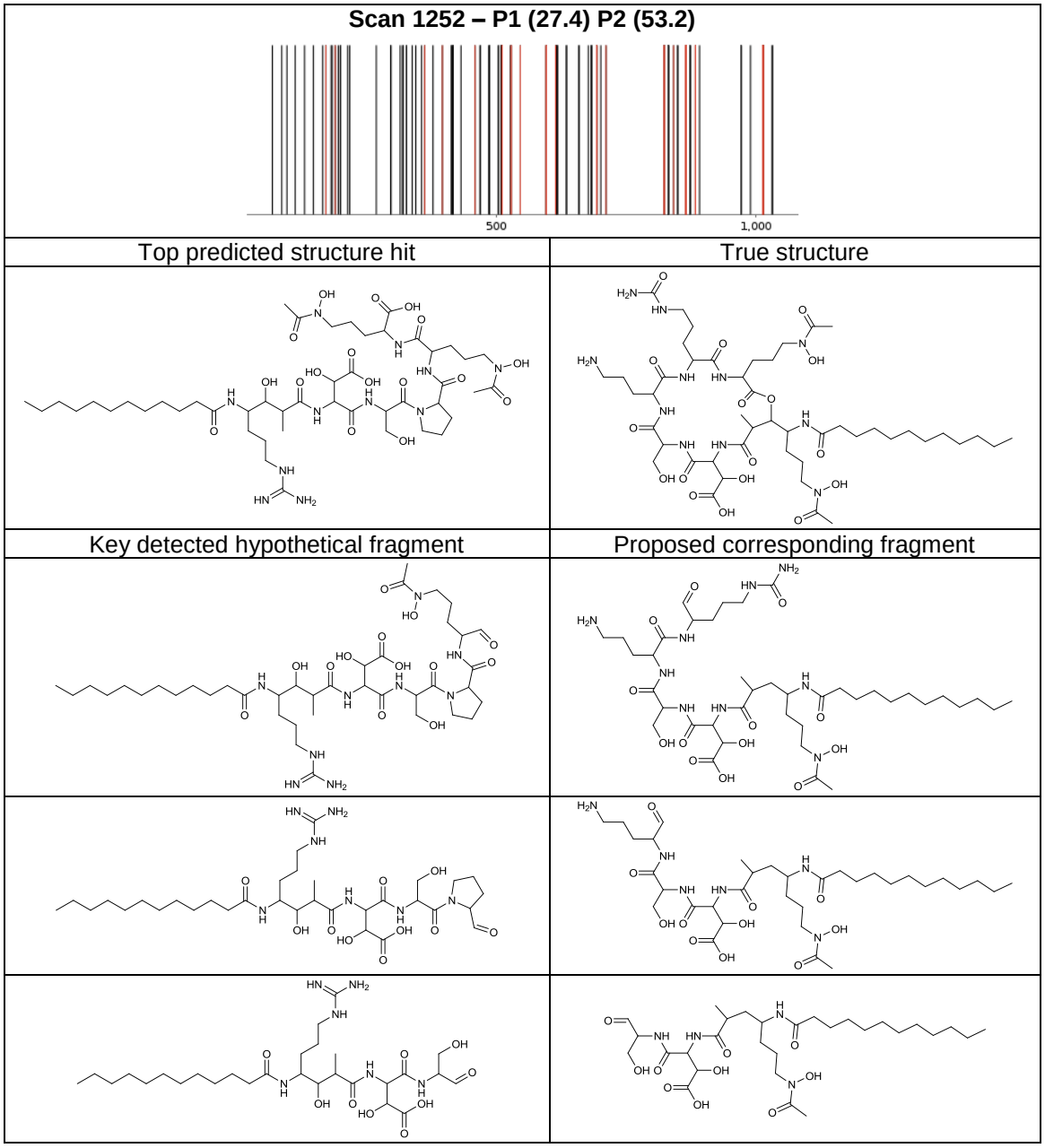

a
8.33
6.90
4.29
4.29
4.22
4.20
4.19
4.19
3.71
3.70
3.70
3.66
3.59
3.46
3.39
3.38
3.21
3.19
3.18
3.17
3.16
2.55
2.51
2.06
2.05
2.03
2.00
1.99
1.87
1.47
1.46
1.31
1.30
1.29
1.28
1.27
1.26
1.24
1.15
0.88
0.87
0.86
0.85
0.51
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
b
74.5
74.1
62.9
61.3
58.8
53.8
50.5
50.4
47.2
44.3
40.9
40.8
40.4
40.3
40.3
40.2
40.1
40.0
40.0
39.9
39.8
39.7
39.5
36.0
31.8
30.4
29.5
29.5
29.4
29.3
29.2
29.2
25.9
25.5
23.7
23.7
22.6
22.0
16.9
16.8
16.6
14.4
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
190
170
150
130
110
90
80
70
60
50
40
30
20
10
f1 (ppm)


c
f1 (ppm)
f2 (ppm)
d
f1 (ppm)
f2 (ppm)

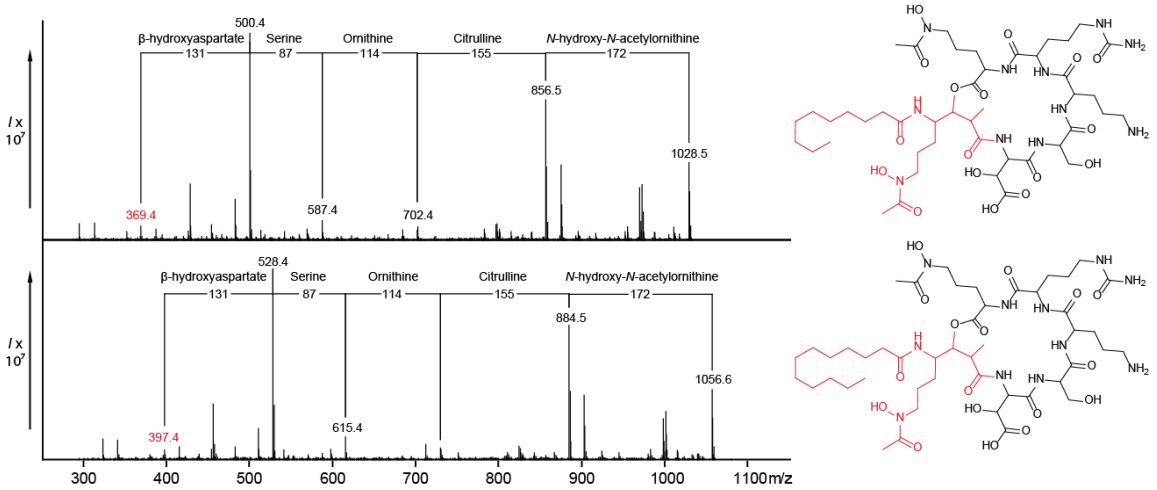



**Supplementary Figure 18** NMR analysis of variobactin A (**6**) in DMSO. (**a**) $^{1}$H NMR spectrum. (**b**) $^{13}$C DEPTq spectrum. (**c**) $^{1}$H-$^{1}$H COSY spectrum. (**d**) $^{1}$H-$^{13}$C HMBC spectrum. (**e**) $^{1}$H-$^{13}$C HMQC spectrum.

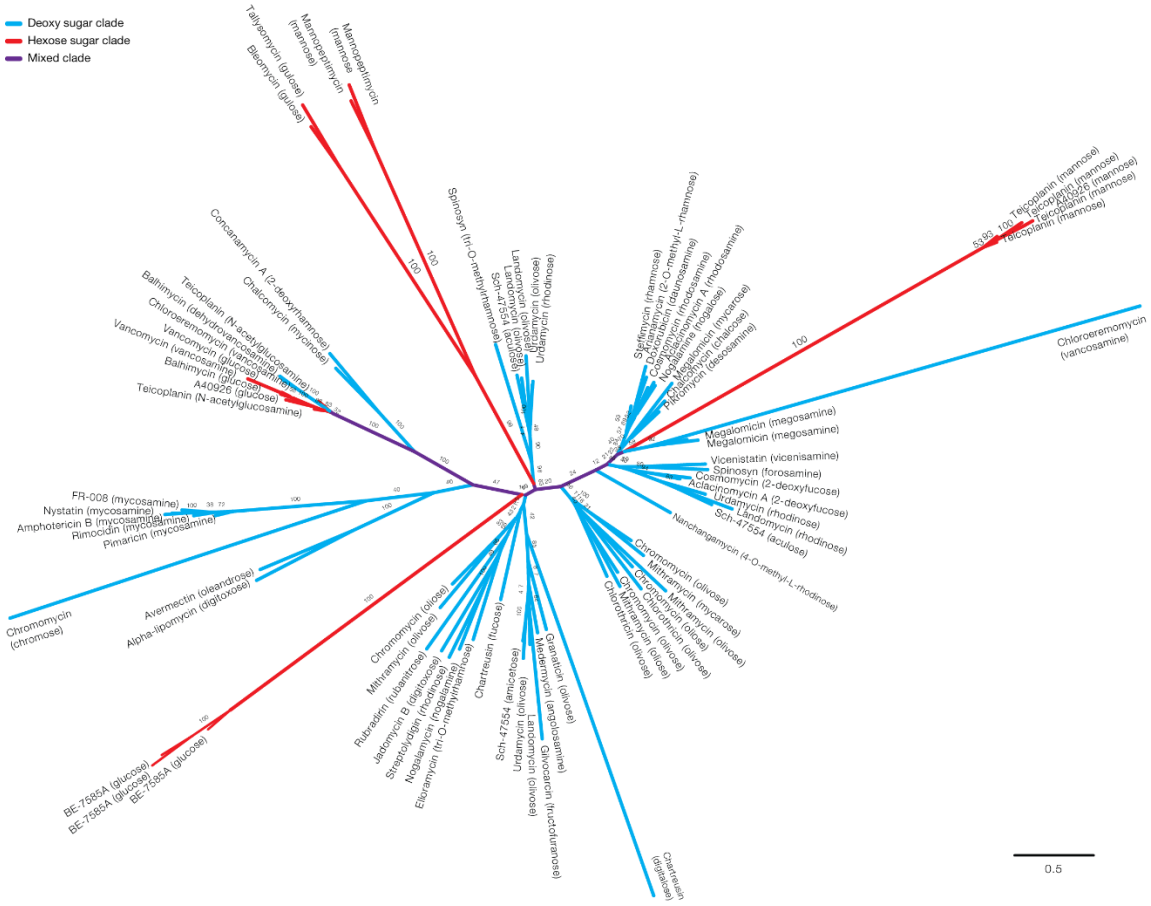
**Supplementary Figure 19** Sample matched fragments between a predicted structure hit and variobactin A. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of variobactin A were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.



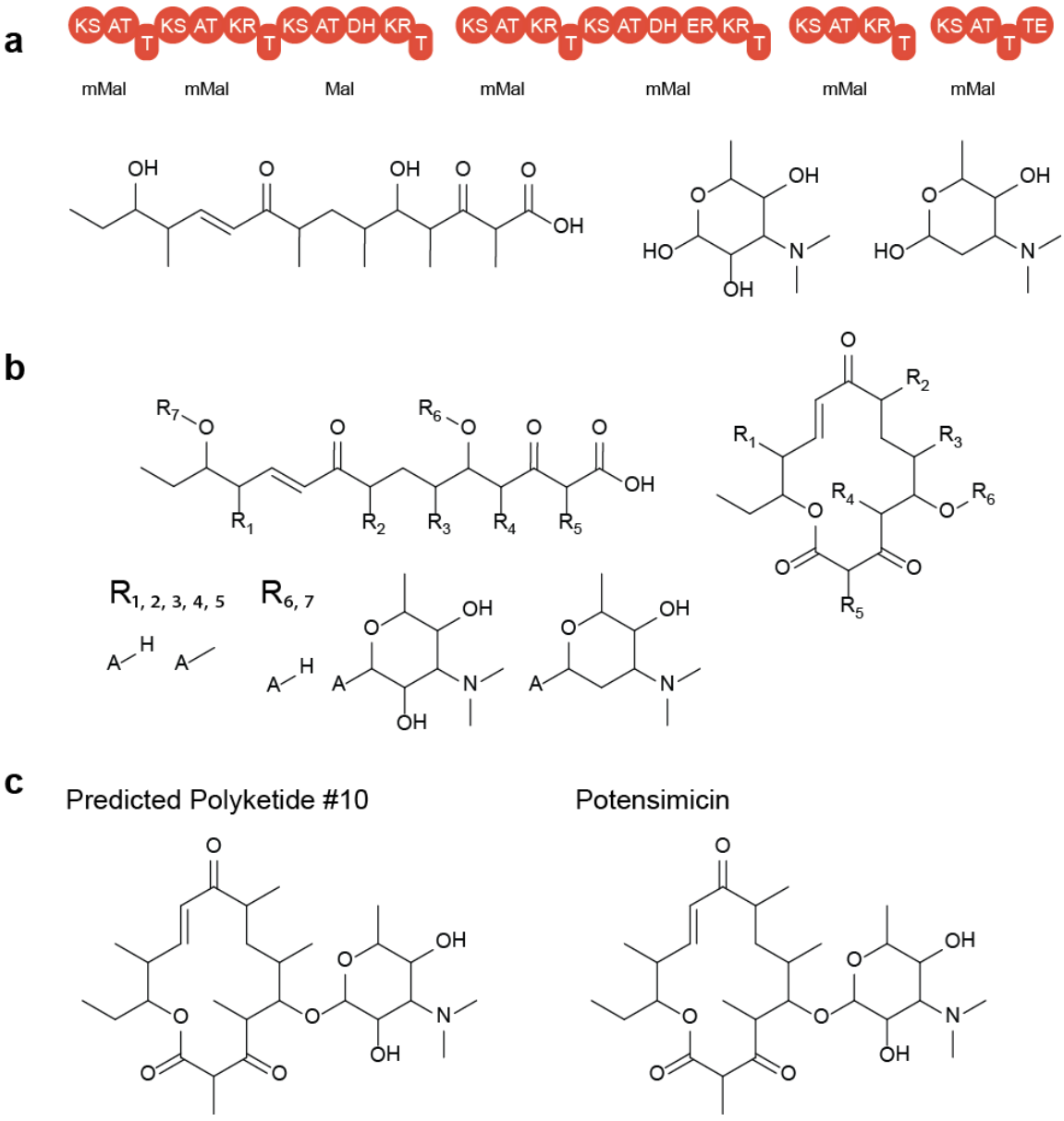

**Supplementary Figure 20** Sample matched fragments between a predicted structure hit and variobactin B. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of variobactin B were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.



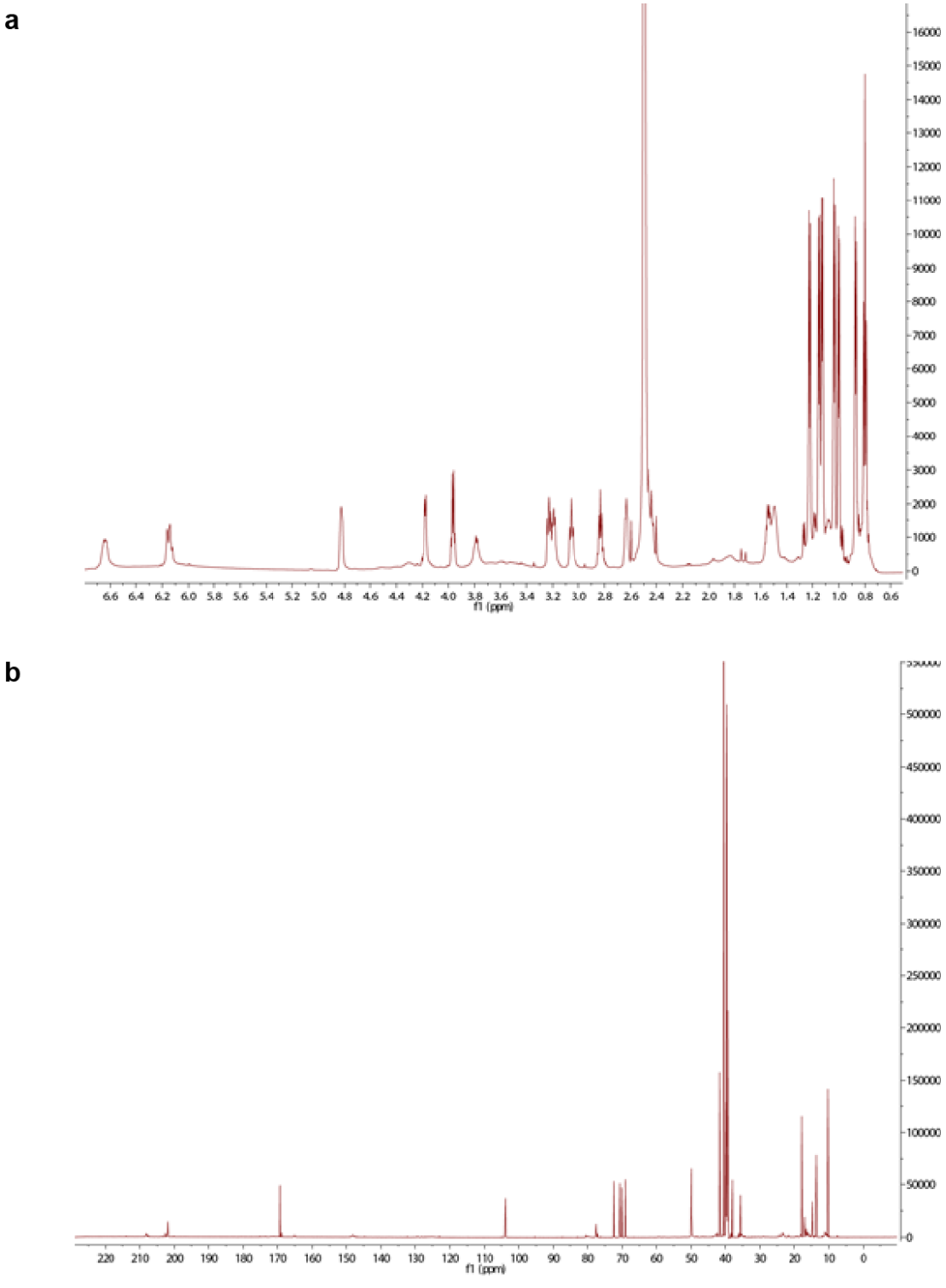

**Supplementary Figure 21** MS/MS comparison of variobactin A and B. LC-MS/MS fragments of variobactin B (*top*) and A (*bottom*) indicate the presented structures, with a 28 Da difference in fragment mass attributed to the addition of two $CH_2$ units to the acyl tail of variobactin A.




**Supplementary Figure 22** Phylogenetic analysis of natural product glycosyltransferases. A manually curated database of 82 natural product glycosyltransferase domains and substrates was aligned and used to construct a phylogenetic tree. Results indicate that natural product glycosyltransferases specific to hexose sugars form distinct clades, including glycopeptide mannosyltransferases, BE-7585A-type glucosyltransferases, and a fouth mixed mixed clade, consisting of both hexose and deoxysugar glycosyltransferases, consisting of position-specific glycopeptide glycosyltransferases.

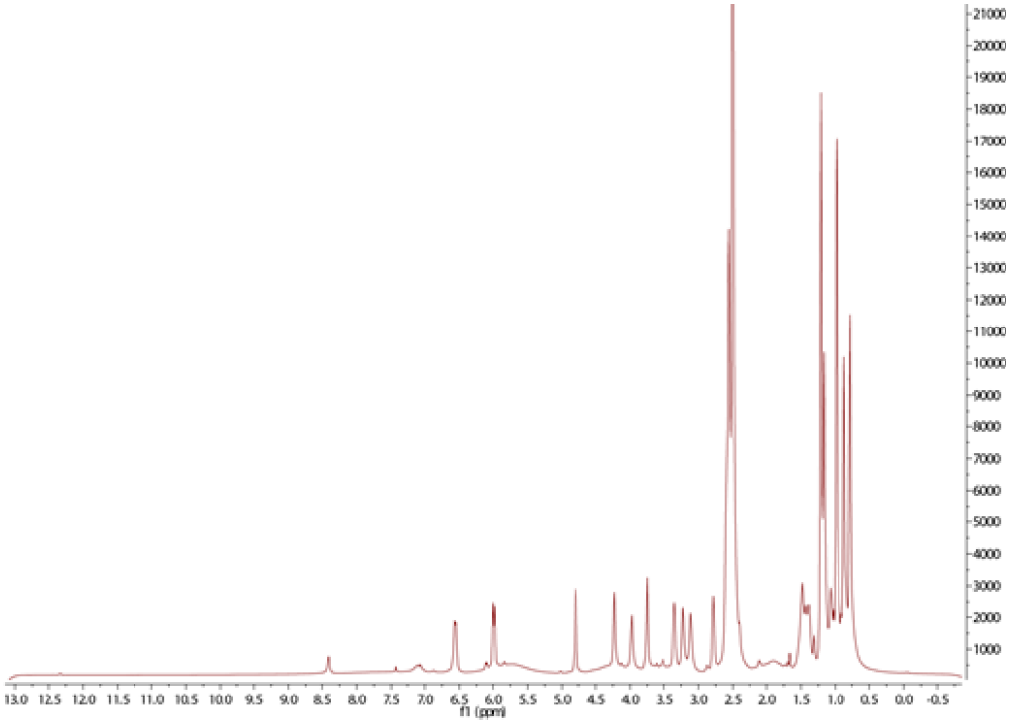



**Supplementary Figure 23** Prediction, combinatorialization, and detection of potensimicin. (**a**) GNP detected PKS biosynthetic gene cluster from *Nocardiopsis potens*, with predicted monomers listed and GNP-generated prediction shown. (**b**) Structures used for combinatorialization. (**c**) Structure of the most frequent hit detected within the *N. potens* LC-MS/MS data (*left*), as well as the true associated metabolite, potensimicin (*right*).

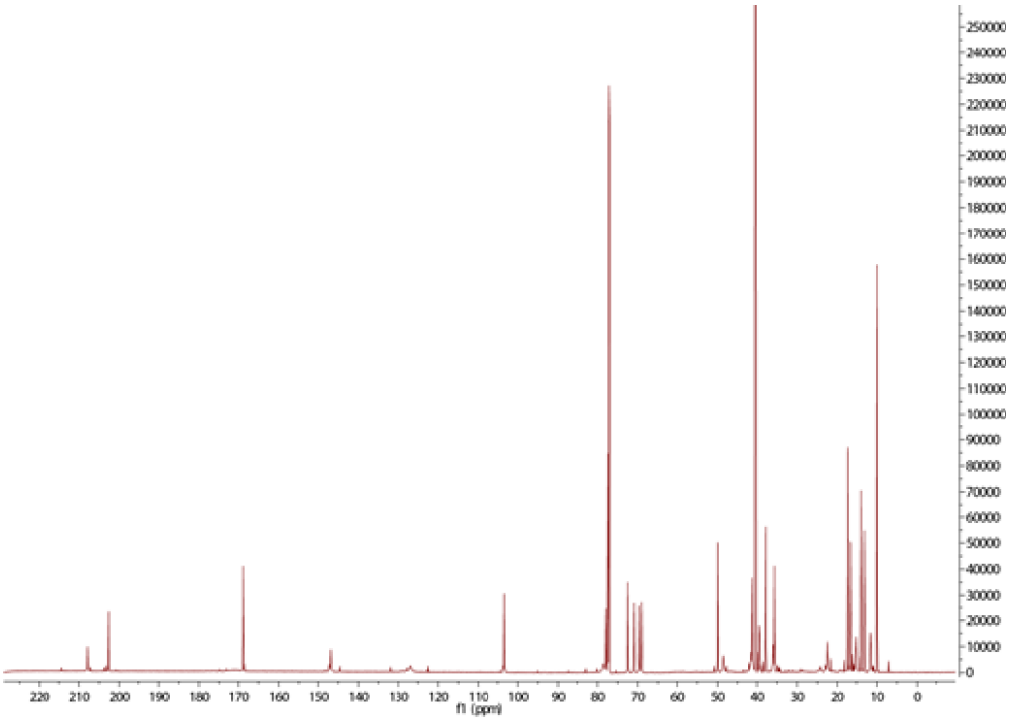



**Supplementary Figure 24** NMR analysis of potensimicin (**7**) in DMSO. (**a**) $^{1}H$ NMR spectrum. (**b**) $^{13}C$ NMR spectrum.

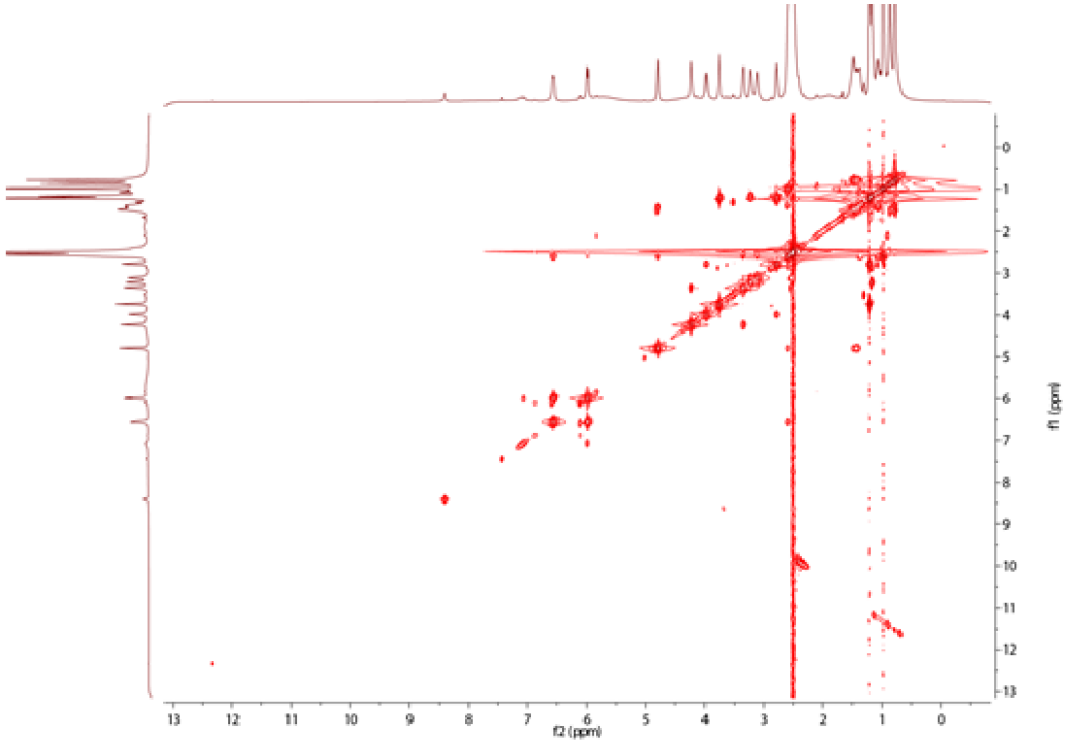
**a**



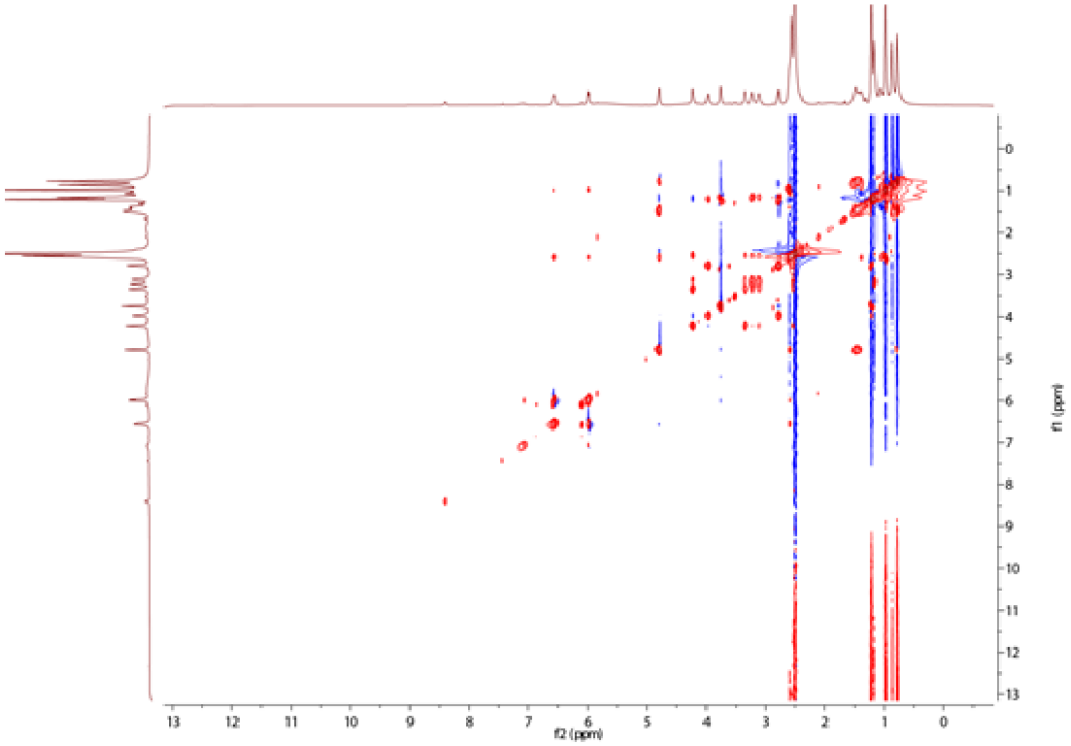

**b**



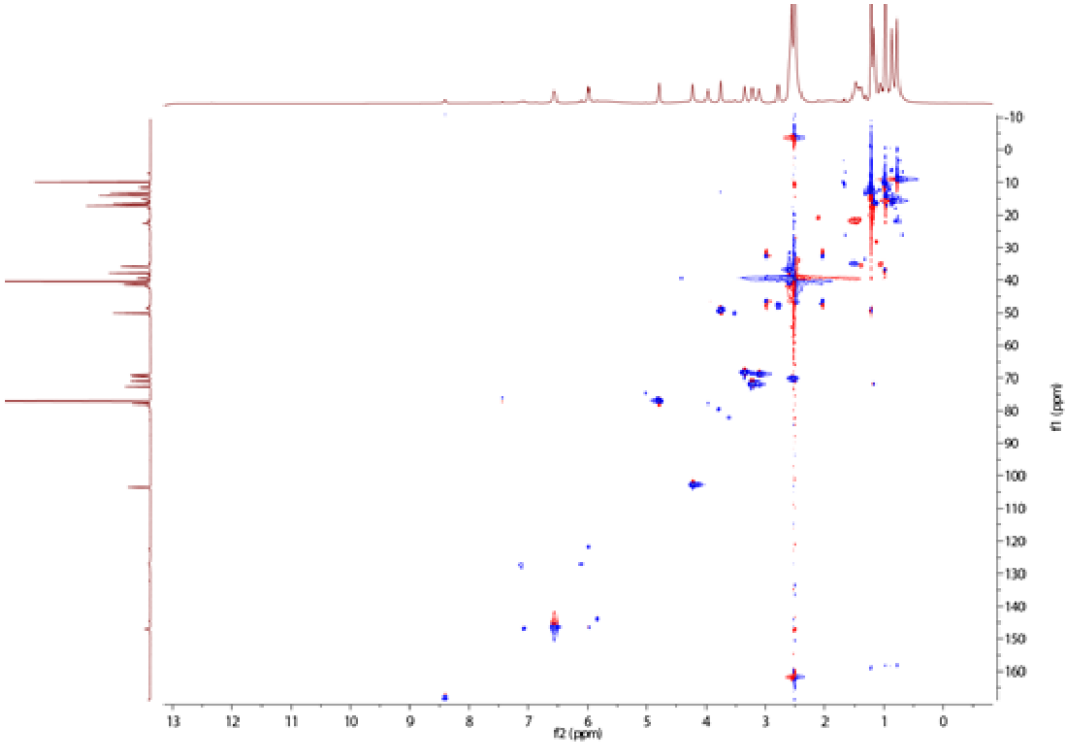
**c**



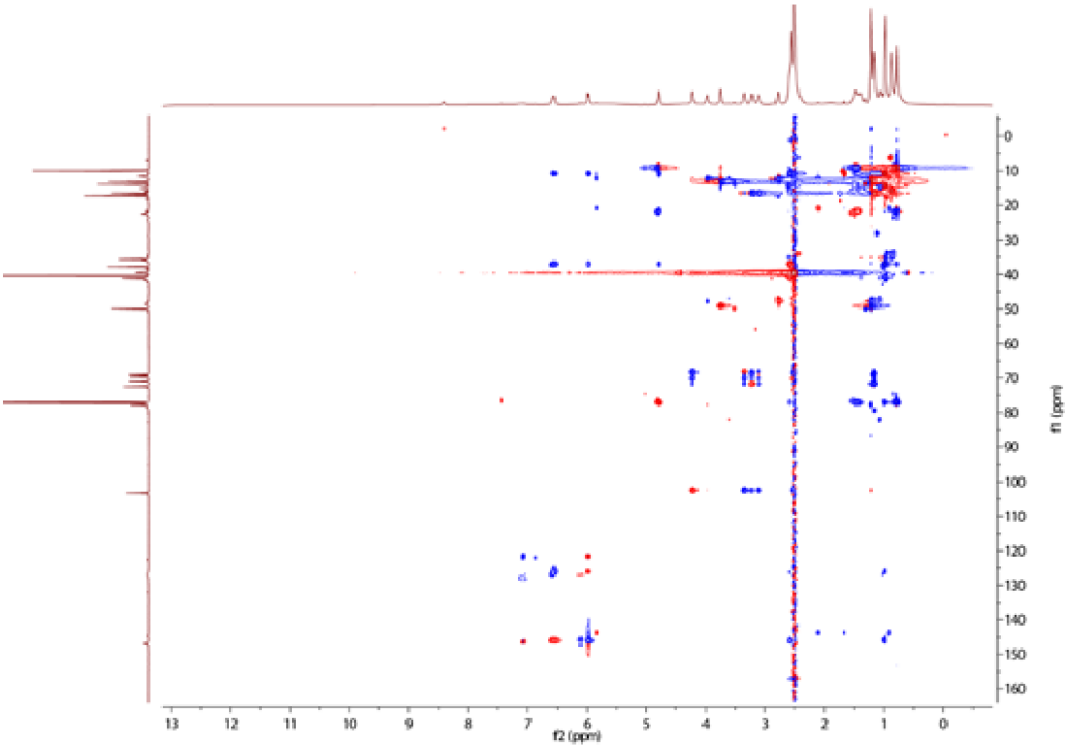

**d**

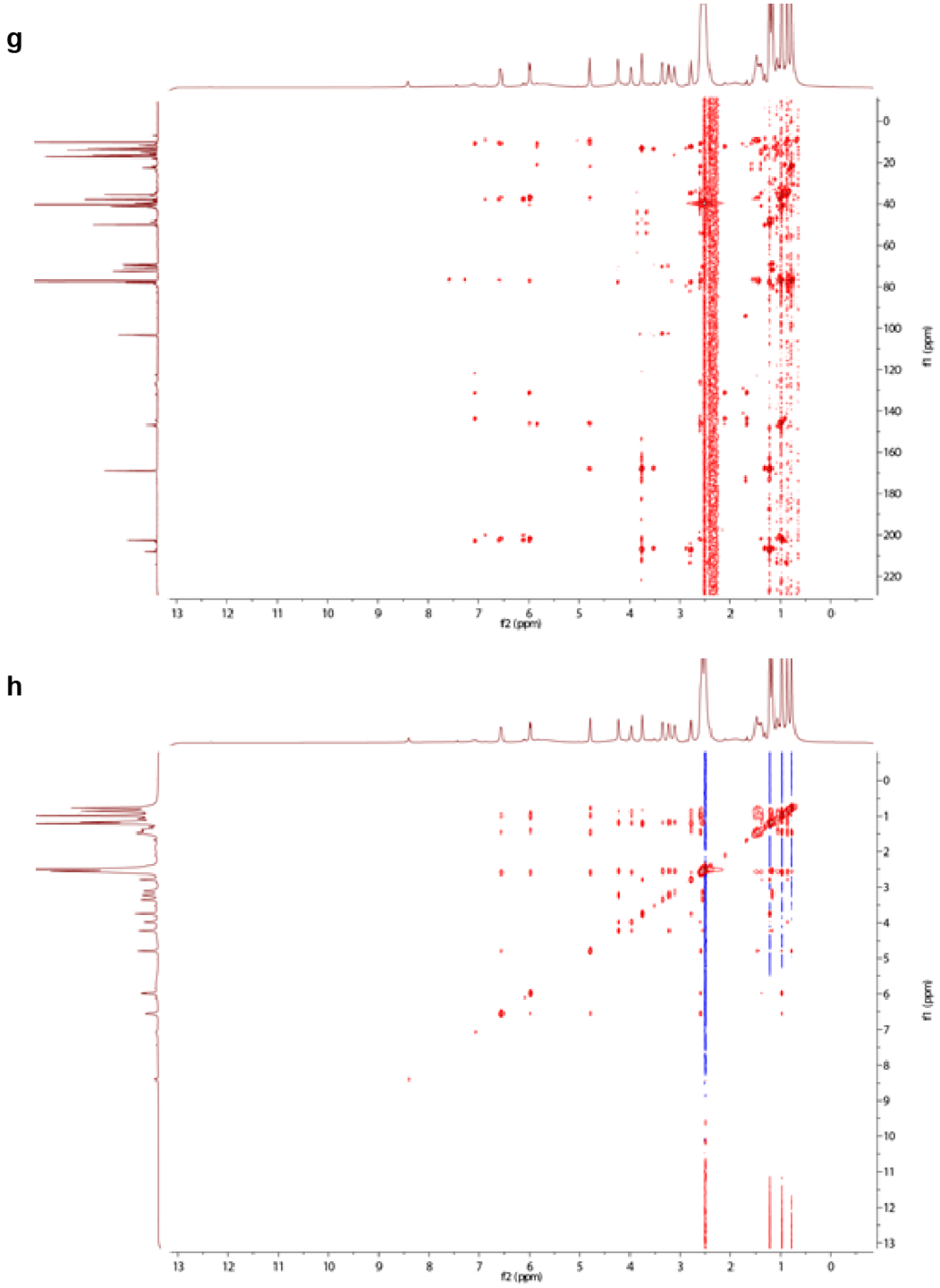


**e**

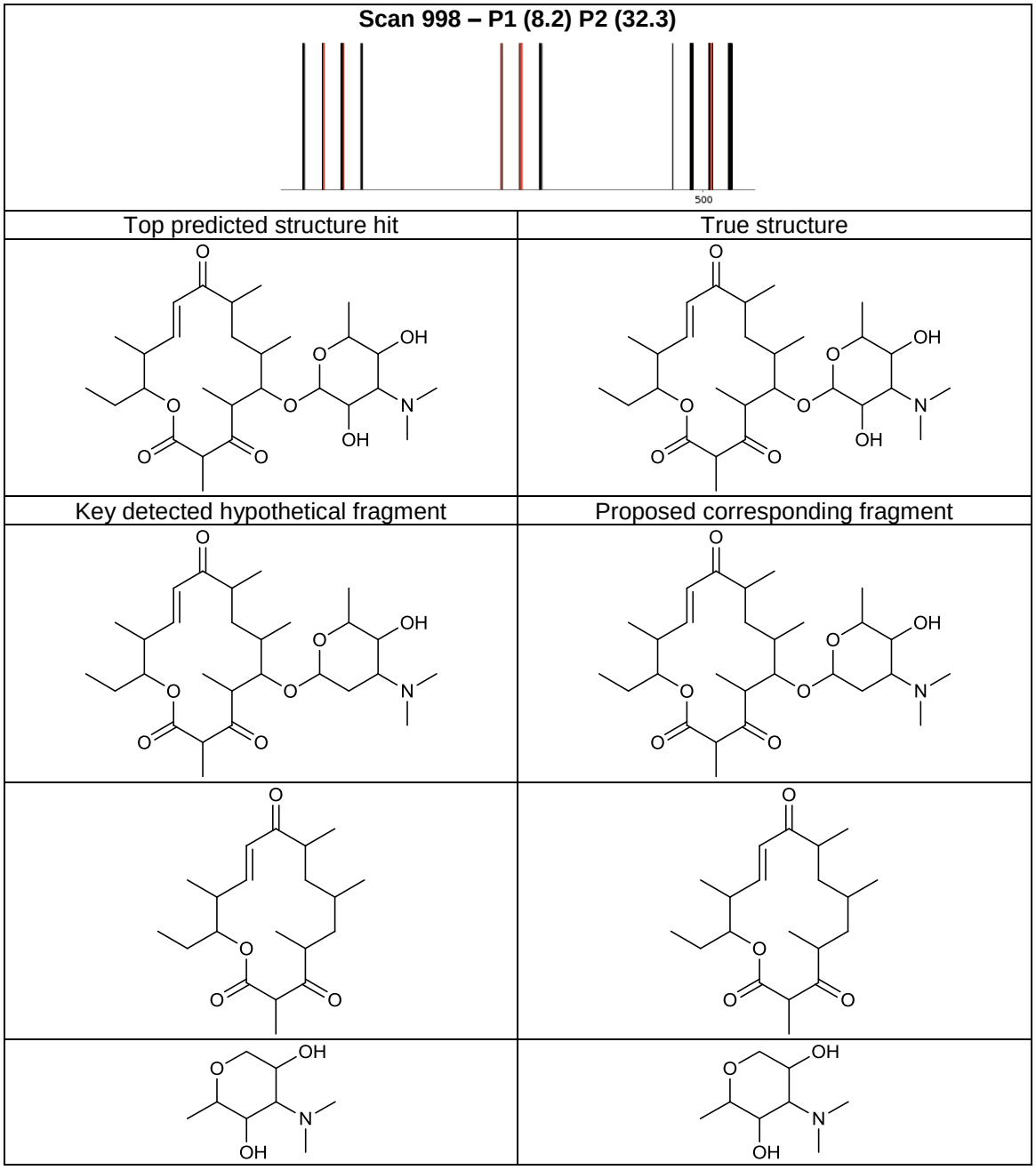



**f**

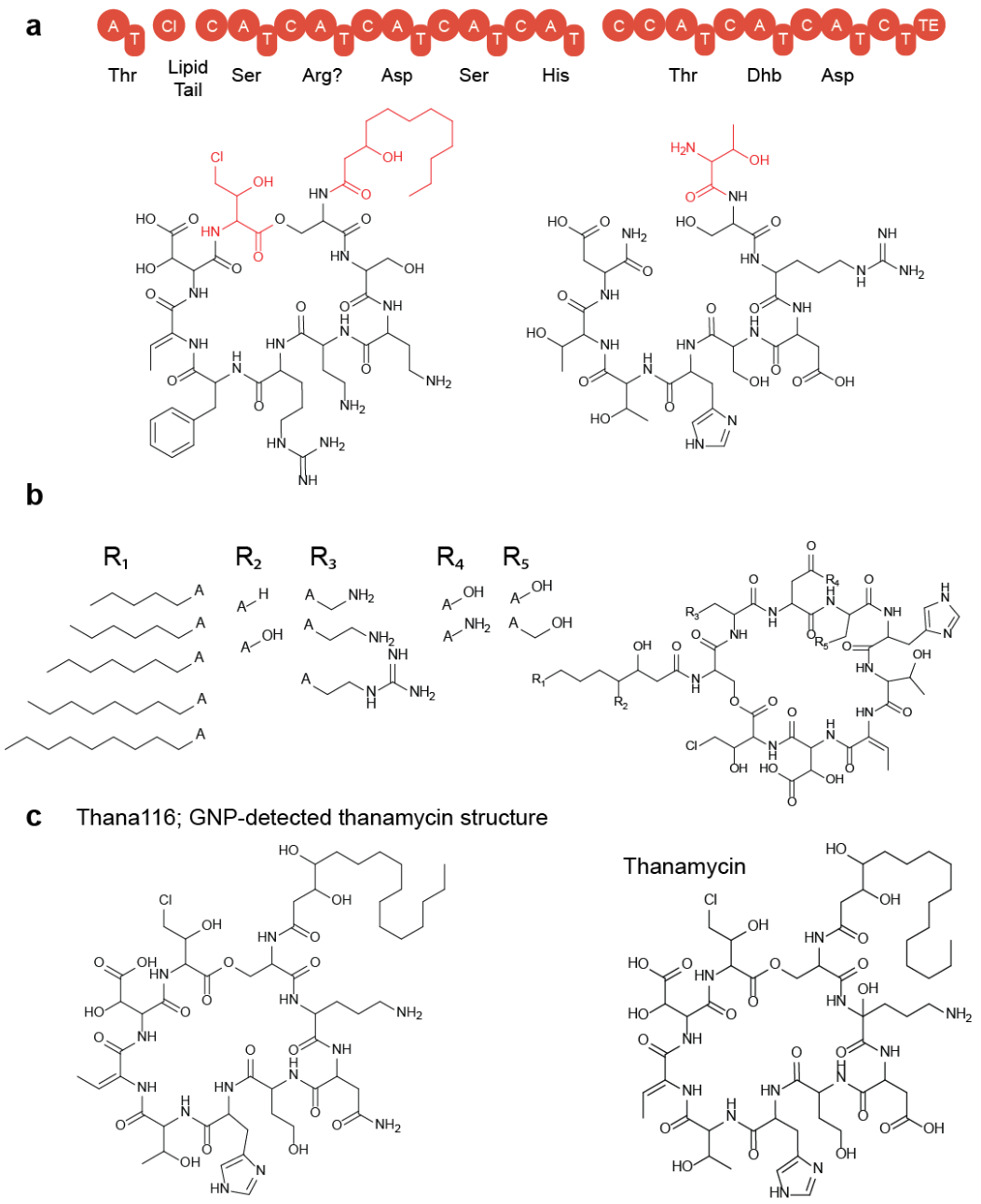




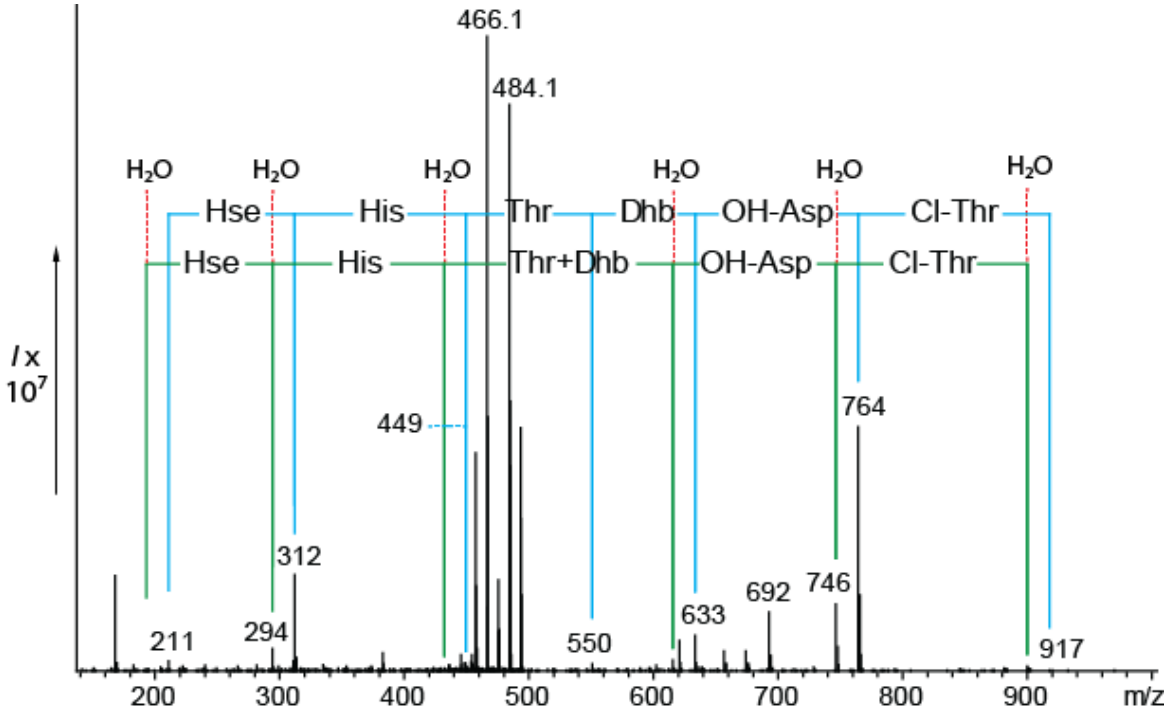

**Supplementary Figure 25** NMR analysis of potensimicin (**7**) in $CDCl_3$. (**a**) $^{1}H$ NMR spectrum. (**b**) $^{13}C$ NMR spectrum. (**c**) $^{1}H$-$^{1}H$ COSY spectrum. (**d**) $^{1}H$-$^{1}H$ TOCSY spectrum. (**e**) $^{1}H$-$^{13}C$ HSQC spectrum. (**f**) $^{1}H$-$^{13}C$ HSQC-TOCSY spectrum. (**g**) $^{1}H$-$^{13}C$ HMBC spectrum. (**h**) $^{1}H$-$^{1}H$ NOESY spectrum.



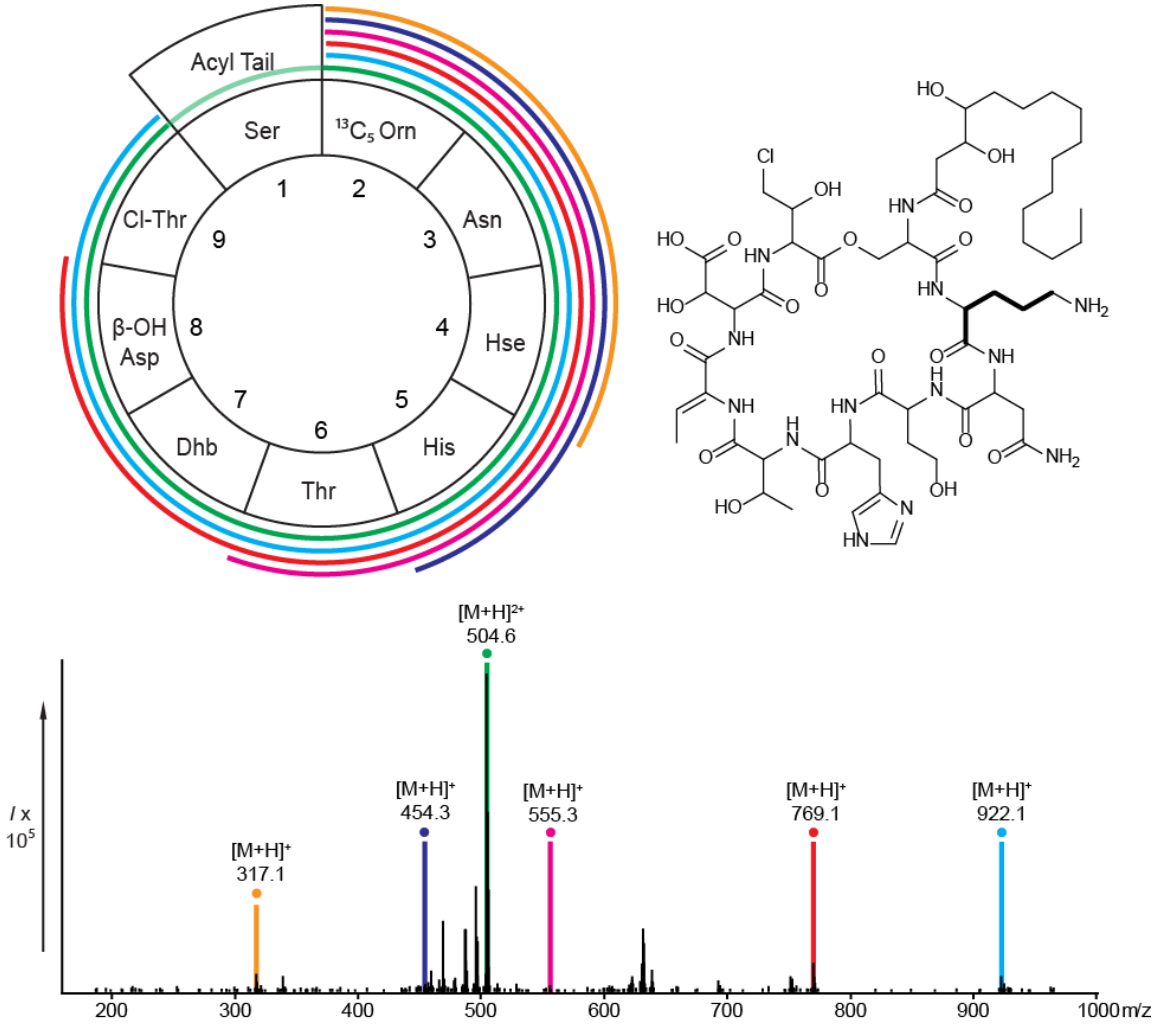

**Supplementary Figure 26** Sample matched fragments between a predicted structure hit and potensimicin. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of potensimicin were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.



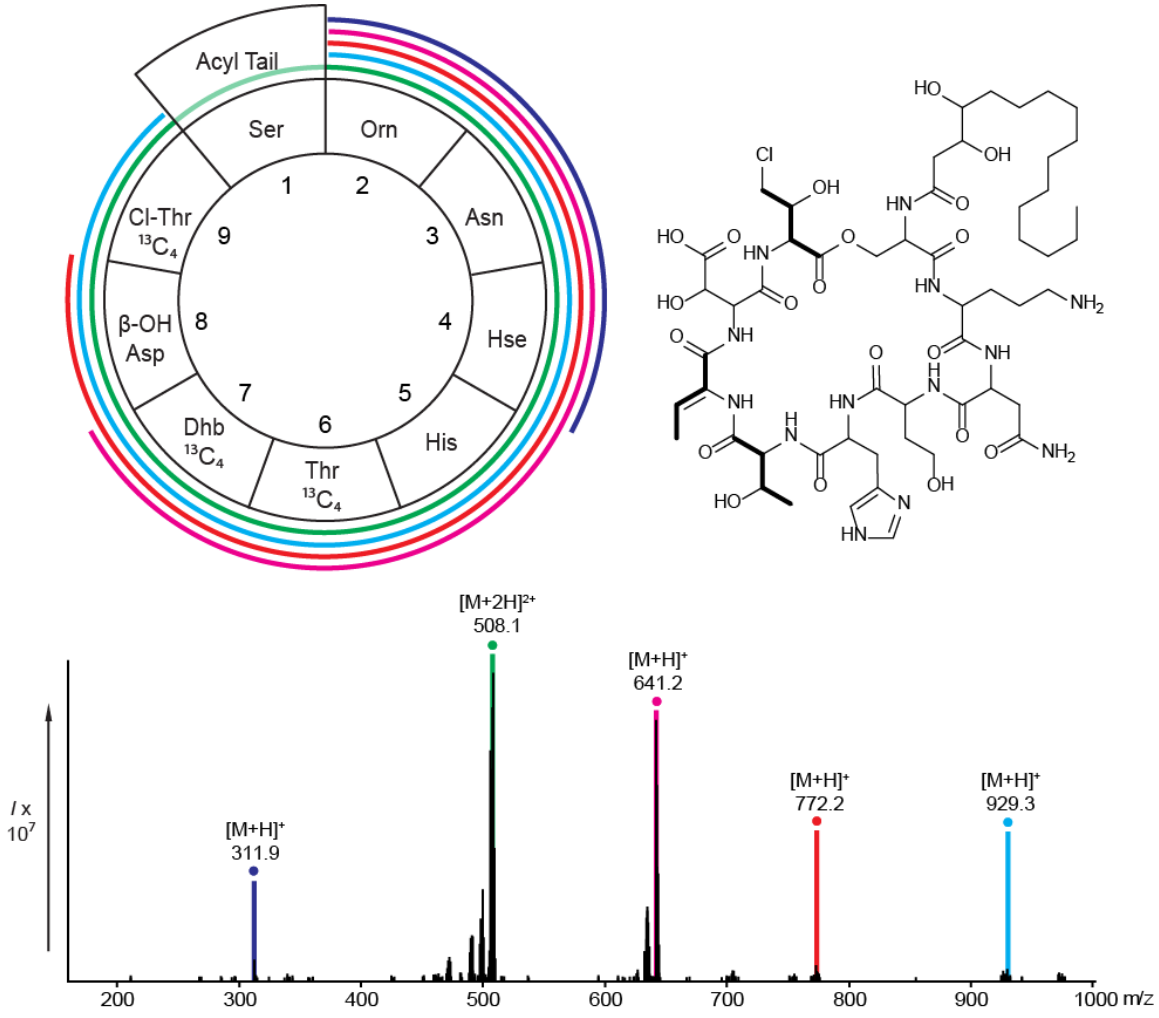

**Supplementary Figure 27** Prediction, combinatorialization, and detection of thanamycin metabolites. (**a**) GNP detected thanamycin biosynthetic gene cluster from *Pseudomonas fluorescens* DSM 11579 (90% pairwise identity with *Pseudomonas* SH-C52 ThaAB), with predicted monomers listed and GNP-generated prediction shown (*right*) beside the related syringomycin structure (*left*). (**b**) Structures used for combinatorialization, taking into account the initial acylating condensation domain and trans-acting chlorinated threonine adenylation domain, as well as conserved syringomycin-family modifications. (**c**) Structure of the most frequent hit detected within the *P. fluorescens* DSM 11579 LC-MS/MS data.




**Supplementary Figure 28** Thanamycin tandem MS. MS/MS of the deacylated and dehydrated 1004 $[M+H]^+$ thanamycin fragment with amino acid annotation.

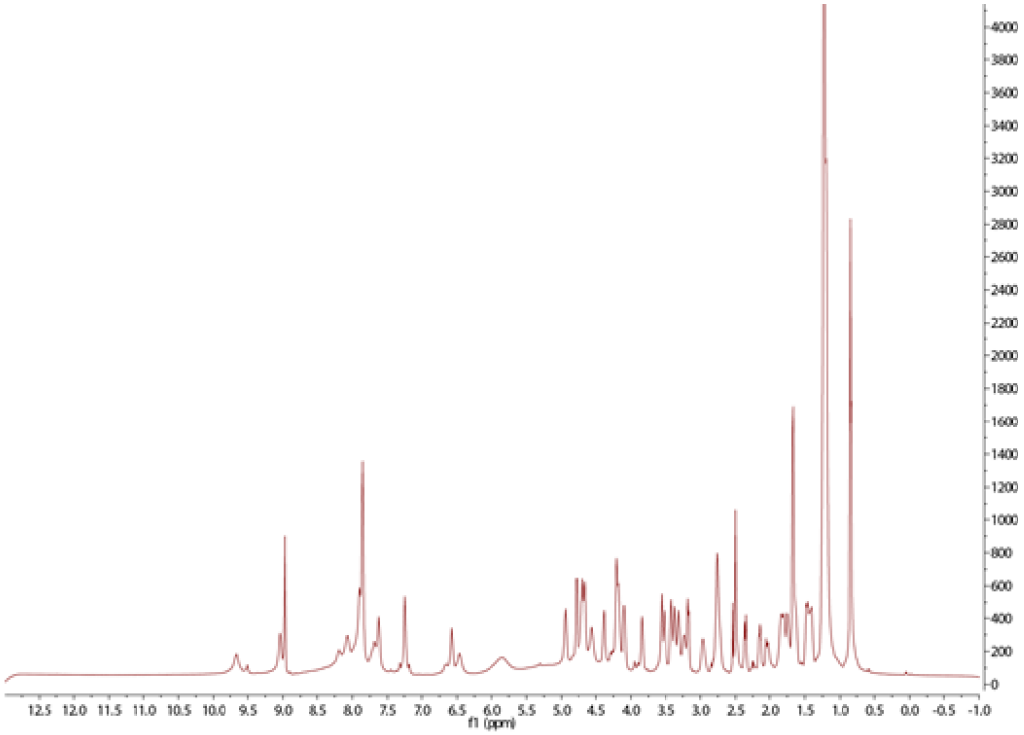



**Supplementary Figure 29** GNP analysis of $^{13}C_5$-ornithine incorporation indicates ornithine at position 2. GNP analysis of extracted cultures containing stable $^{13}C_5$-ornithine identified matched fragments corresponding thanamycin with heavy ornithine at position 2 (*top left*, *bottom*), resulting in the matched structure (*top right*). MS/MS spectra and matched fragments are from representative scan 913.

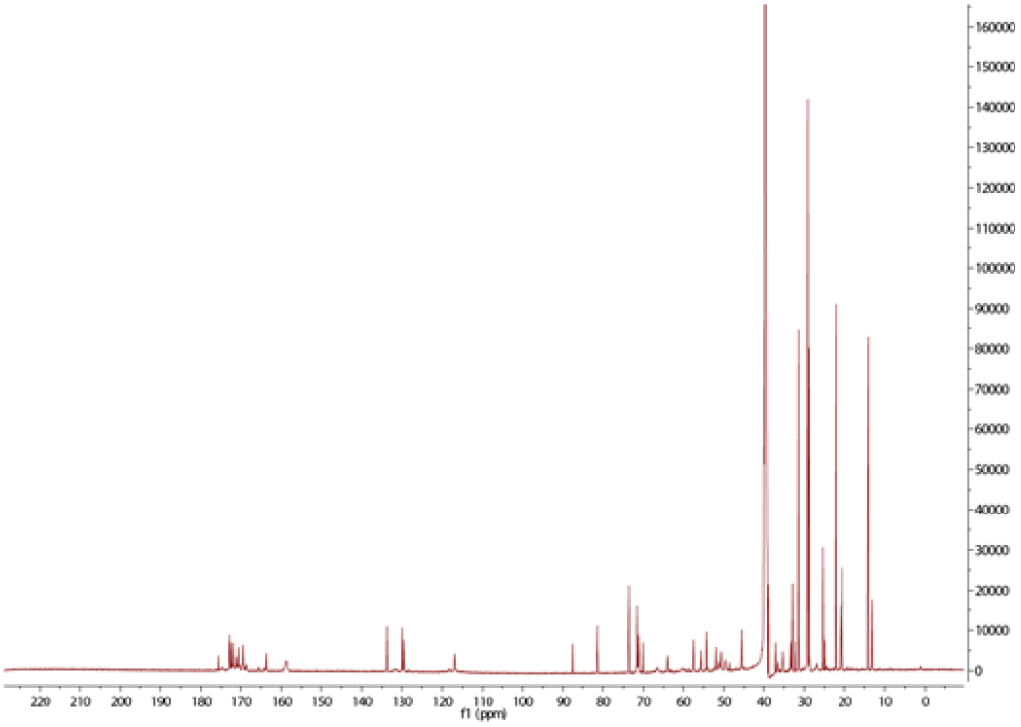



**Supplementary Figure 30** GNP analysis of $^{13}C_4$-threonine incorporation indicates homoserine at position 4. GNP analysis of extracted cultures containing stable $^{13}C_4$-threonine identified matched fragments corresponding thanamycin with heavy threonine or threonine derivatives (Dhb) at positions 6, 7, and 9 (*top left*, *bottom*). No fragments were detected that indicated incorporation at position 4, resulting in the matched structure with homoserine found at position 4 (*top right*). MS/MS spectra and matched fragments are from representative scan 1148.

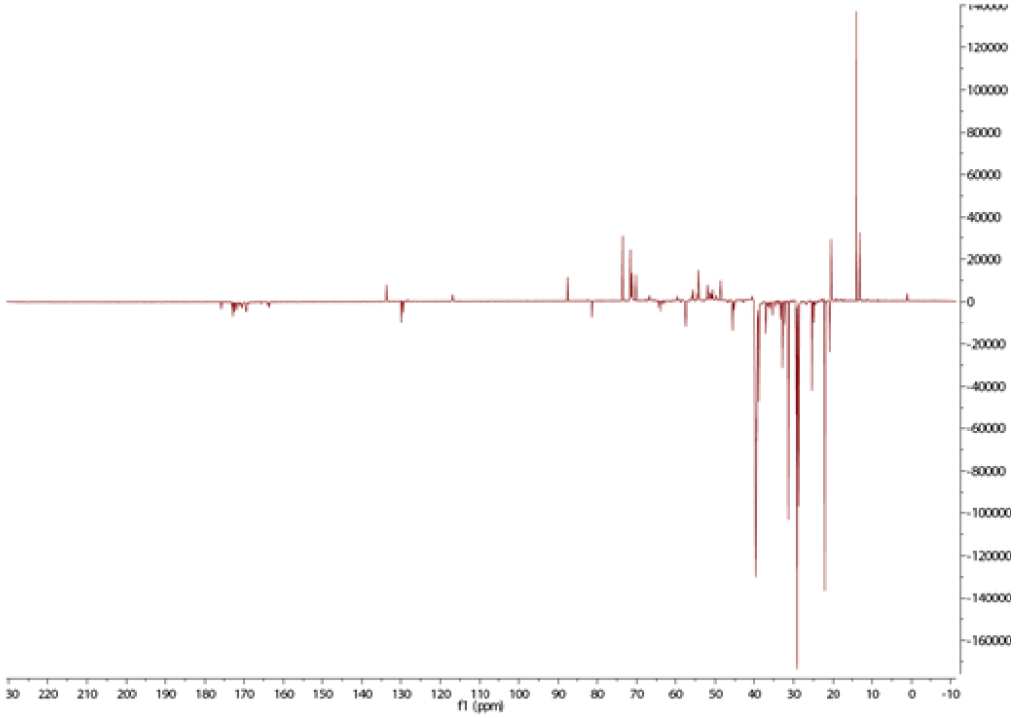
**a**



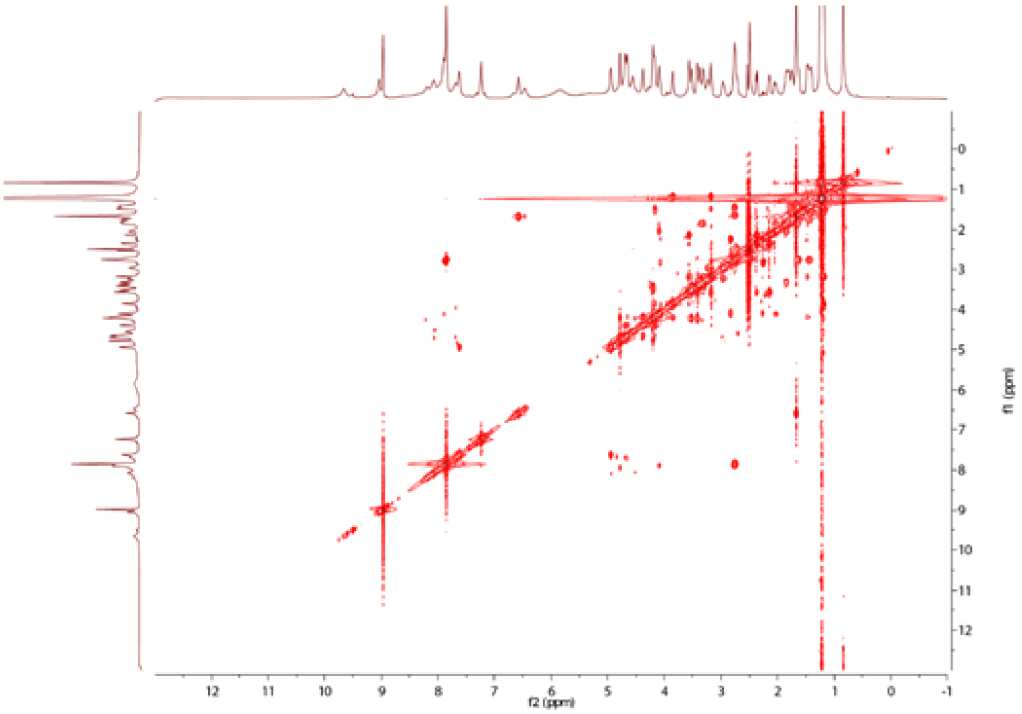

**b**



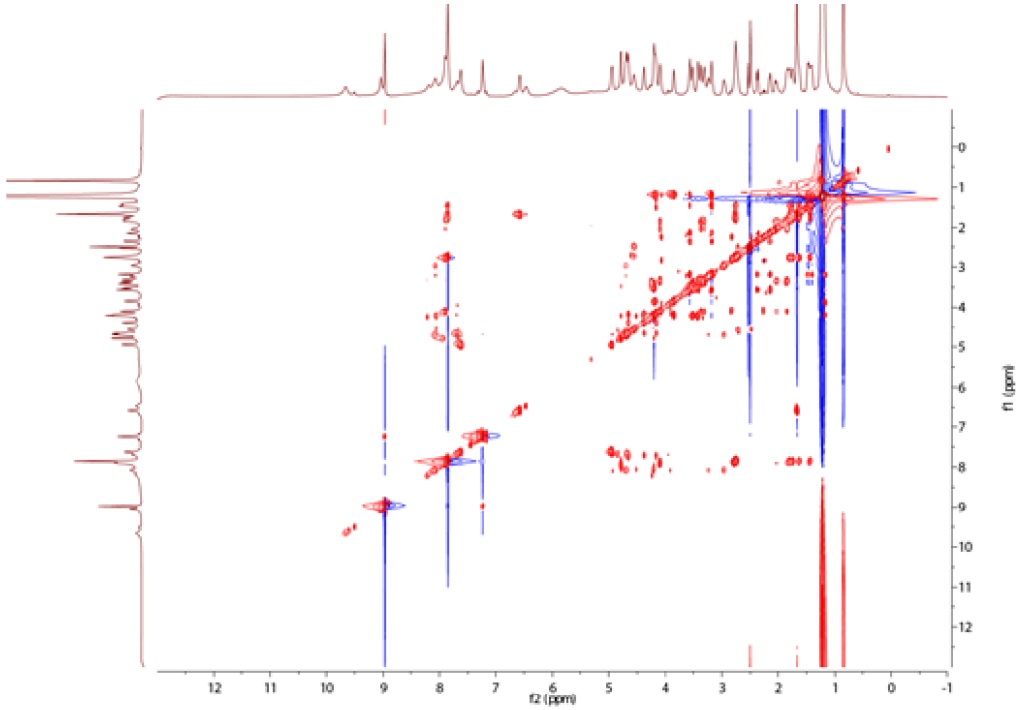
**c**


120000
100000
80000
60000
40000
20000
0
-20000
-40000
-60000
-80000
-100000
-120000
-140000
-160000
30 220 210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10
f1 (ppm)

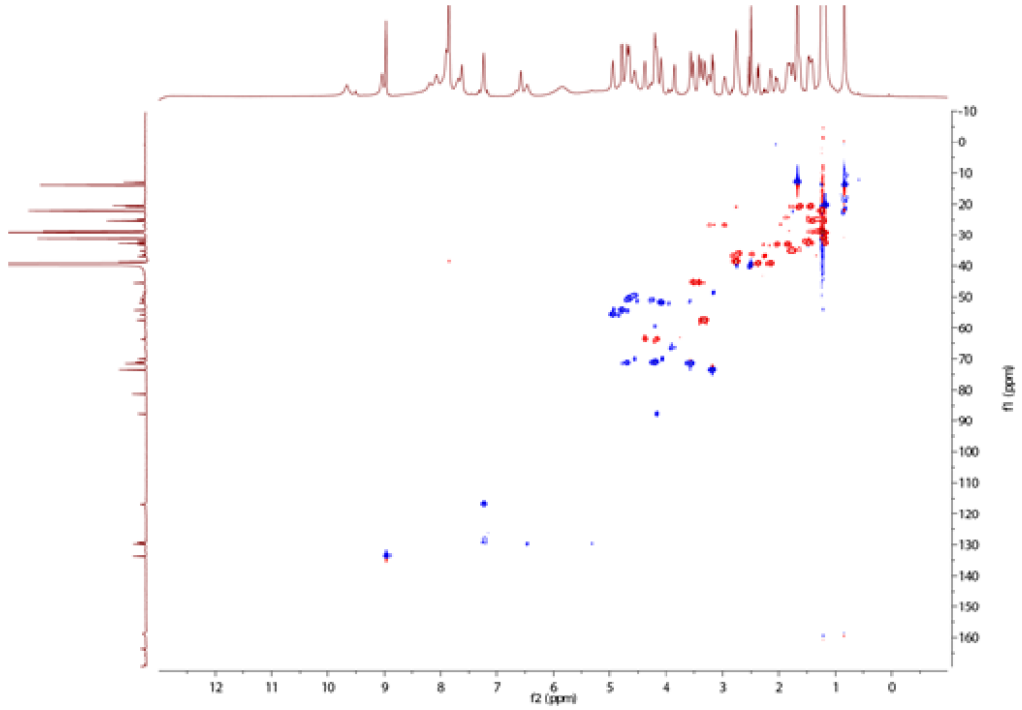
**d**


0
1
2
3
4
5
6
7
8
9
10
11
12
f1 (ppm)
12 11 10 9 8 7 6 5 4 3 2 1 0 -1
f2 (ppm)

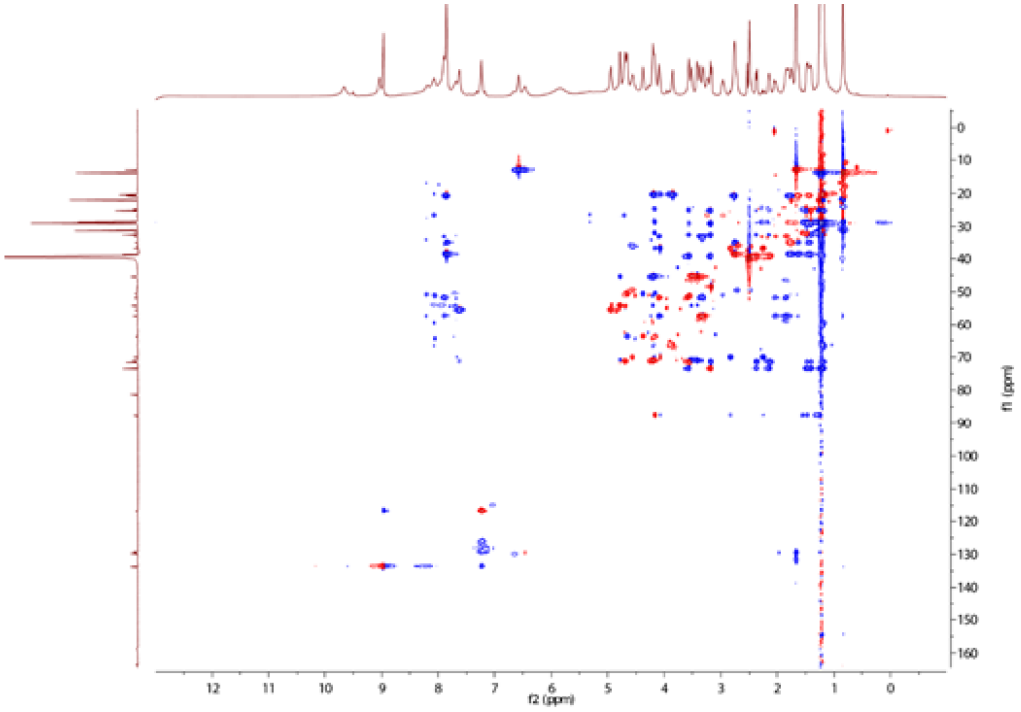
**e**



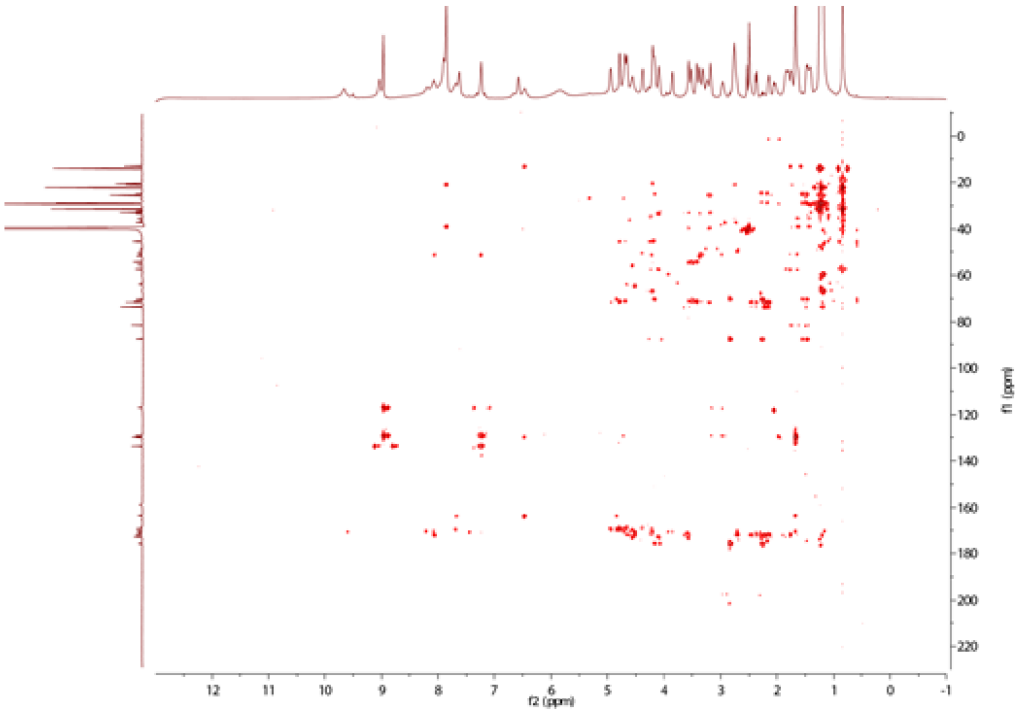

**f**



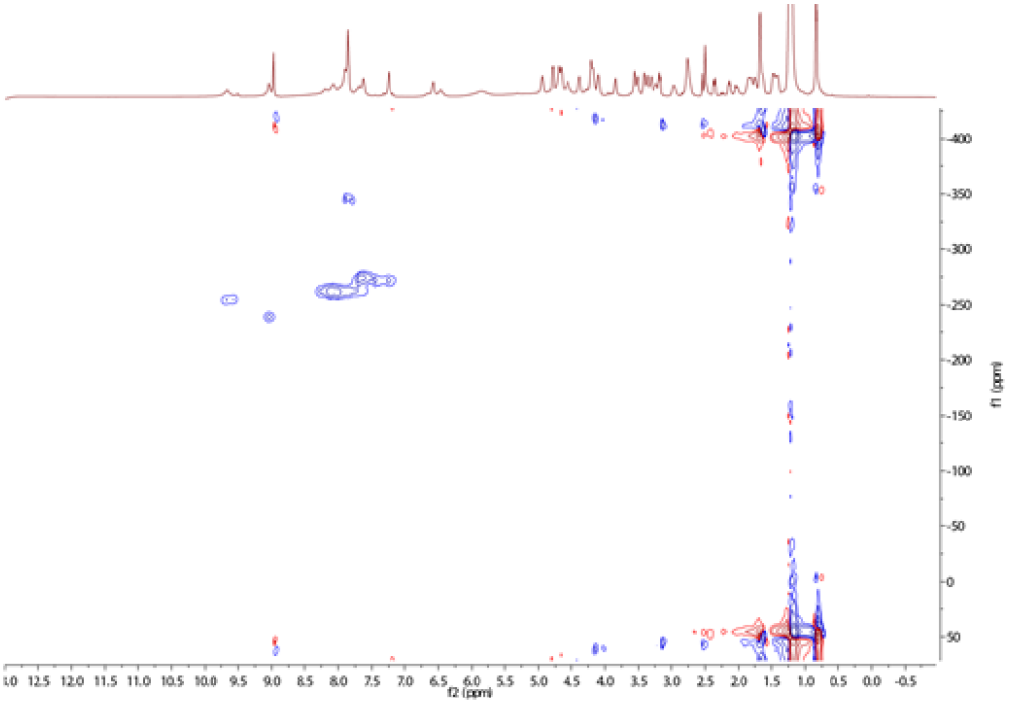
**g**



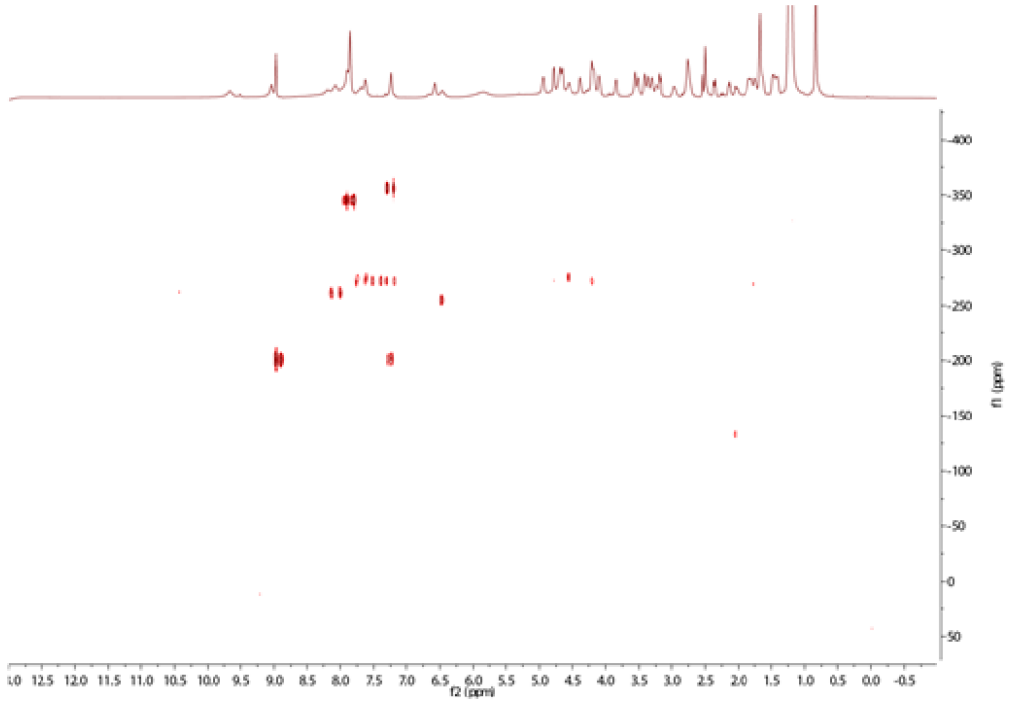

**h**

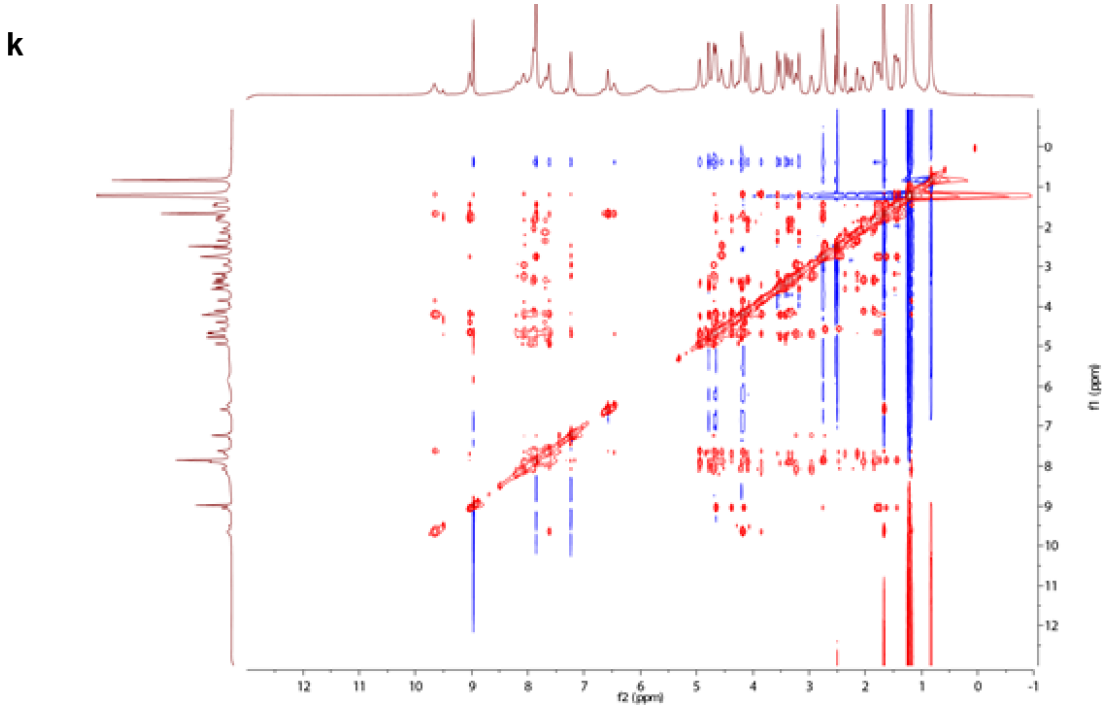


**i**

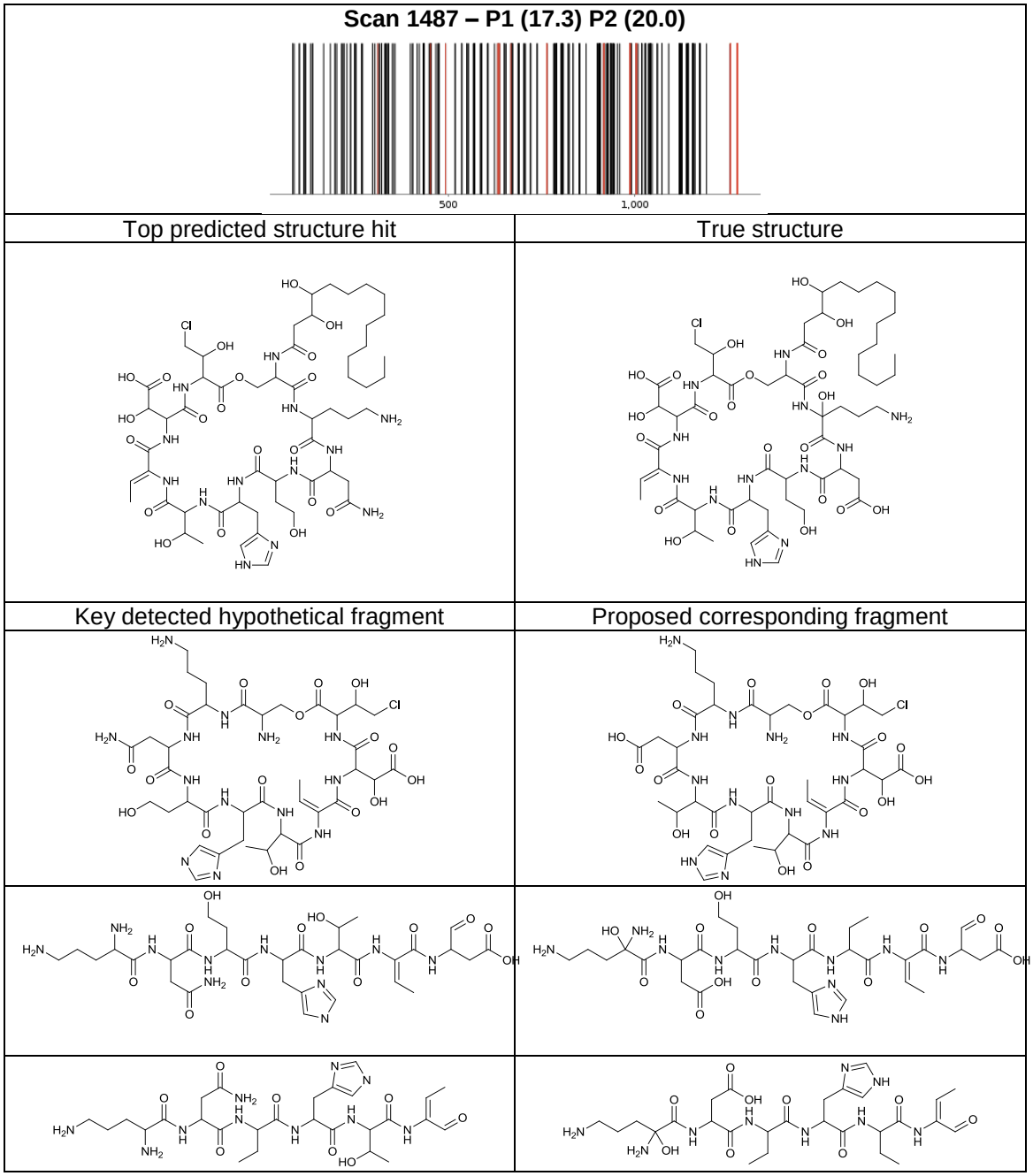



**j**






**Supplementary Figure 31** NMR analysis of thanamycin (**8**) in DMSO. (**a**) $^{1}$H NMR spectrum. (**b**) $^{13}$C NMR spectrum. (**c**) $^{13}$C DEPTq spectrum. (**d**) $^{1}$H-$^{1}$H COSY spectrum. (**e**) $^{1}$H-$^{1}$H TOCSY spectrum. (**f**) $^{1}$H-$^{13}$C HSQC spectrum. (**g**) $^{1}$H-$^{13}$C HSQC-TOCSY spectrum. (**h**) $^{1}$H-$^{13}$C HMBC spectrum. (**i**) $^{1}$H-$^{15}$N HSQC spectrum. (**j**) $^{1}$H-$^{15}$N HMBC spectrum. (**k**) $^{1}$H-$^{1}$H NOESY spectrum.




**Supplementary Figure 32** Sample matched fragments between a predicted structure hit and thanamycin. Several matched fragments between the hypothetical compound hit and the corresponding true structure are shown. Fragments of thanamycin were generated using the GNP Fragmenter application. Detected hypothetical spectral fragments are shown as lines in the diagram above, with detected masses shown in red. P-scores for the scan are listed above. Full results tables for this scan and experimental run are available at http://magarveylab.ca/data/gnp.

| Source | Program | Natural Product Classes | Automated NRP structure prediction? | Automated RiPP structure prediction? | Automated sugar structure prediction? | Automated polyketide structure prediction? | Automated MS analysis? | Statistical genome - metabolome matching? | Single software package? | User-friendly interface? |
|---|---|---|---|---|---|---|---|---|---|---|
| This paper | GNP | Glycosylated Polyketides, NRPs | **Yes** | No | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** |
| 55 | NRPQuest | NRPs | **Yes** | No | No | No | **Yes** | **Yes** | **Yes** | No |
| 54 | RiPPQuest | RiPPs | No | **Yes** | No | No | **Yes** | **Yes** | **Yes** | No |
| 20 | Pep2Path | NRPs, RiPPs | No | No | No | No | No | **Yes** | No | No |
| 21 | Peptidogenomics | NRPs, RiPPs | No | No | No | No | No | No | No | No |
| 22 | Glycogenomics | Glycosylated NPs | No | No | No | No | No | No | No | No |
| 57 | HSEE | RiPPs | No | No | No | No | No | **Yes** | No | No |

**Supplementary Figure 33** Comparison of GNP to several manual, automated, or semi-automated genomic and metabolomic natural products discovery methods. GNP expands the chemical search space of previously published methods by targeting a wider spectrum of natural product classes, and automates structure prediction for modular natural products. While previously published methods require the annotation of mass shifts within a tandem mass spectrum scan[20-22], or the selection of a single point on a high-resolution mass spectrum[57], GNP's integrated mass spectral analysis algorithms eliminate the need for manual mass spectrum annotation and therefore facilitates high-throughput discovery. Finally, as a continuous workflow integrated into a single web application with a user-friendly interface, GNP is accessible to natural products chemists without formal bioinformatics training. By automating the most labour-intensive steps associated with a combined genomic and metabolomic discovery approach, GNP has the potential to significantly accelerate natural product discovery. NRP, nonribosomal peptide; RiPP, ribosomally synthesized and post-translationally modified peptide; NP, natural product.

## Supplementary Tables

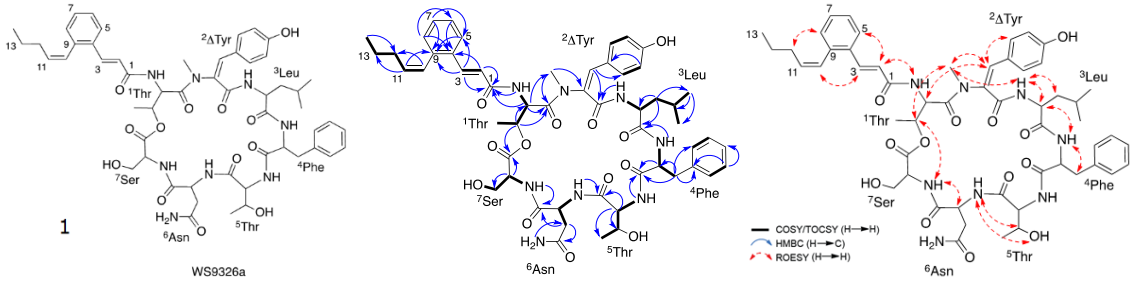
**Supplementary Table 1** WS9326 gene cluster analysis

| Gene name | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| cal1 | Aminotransferase | - | 414 |
| cal2 | LuxR transcriptional regulator | - | 215 |
| cal3 | LysR transcription factor | - | 80 |
| cal4 | MbtH domain protein | - | 73 |
| cal5 | Thioesterase | - | 263 |
| cal6 | Hypothetical protein | + | 319 |
| cal7 | 3-oxoacyl-ACP synthase | + | 335 |
| cal8 | 3-oxoacyl-ACP synthase | + | 414 |
| cal9 | Acyl-carrier protein | - | 84 |
| cal10 | Hypothetical protein | - | 124 |
| cal11 | Hypothetical protein | - | 137 |
| cal12 | Translation initiation factor IF-2 | - | 141 |
| cal13 | 3-hydroxylacyl-ACP dehydratase | - | 315 |
| cal14 | Hypothetical protein | + | 258 |
| cal15 | Ferredoxin | + | 66 |
| cal16 | Cytochrome P450 | + | 408 |
| cal17 | NRPS | - | 2568 |
| cal18 | NRPS | - | 2577 |
| cal19 | NRPS | - | 1892 |
| cal20 | Thioesterase | - | 234 |
| cal21 | Short chain dehydrogenase/reductase | - | 276 |
| cal22 | Adenylation-ACP didomain | - | 979 |
| cal23 | Adenylation-ACP didomain | - | 600 |
| cal24 | Oxidoreductase | - | 400 |
| cal25 | Hydrolase | - | 275 |
| cal26 | Putative thioesterase | - | 334 |
| cal27 | Isomerase | - | 242 |
| cal28 | Phytoene dehydrogenase | - | 575 |
| cal29 | Acyl-carrier protein | - | 87 |
| cal30 | 3-oxoacyl-ACP synthase | - | 417 |
| cal31 | 3-oxoacyl-ACP synthase | - | 380 |
| cal32 | 3-oxoacyl-ACP synthase | - | 315 |
| cal33 | 3-oxoacyl-ACP synthase | - | 372 |
| cal34 | Acyl-carrier protein | - | 88 |
| cal35 | 3-oxoacyl-ACP dehydratase | - | 134 |
| cal36 | 3-oxoacyl-ACP dehydratase | - | 160 |
| cal37 | 3-oxoacyl-ACP reductase | - | 249 |
| cal38 | YbaK/prolyl-tRNA synthetase | - | 179 |
| cal39 | ABC transporter ATPase | - | 319 |
| cal40 | Multidrug ABC transporter permease | - | 281 |

**Supplementary Table 2** Summary of $^{1}$H (600 MHz) and $^{13}$C (125 MHz) spectroscopic data for WS9326a (1) in DMSO-$d_6$[a]

| Position | | $d_H$ ($J$ in Hz) | $d_C$ | Position | | $d_H$ ($J$ in Hz) | $d_C$ |
|---|---|---|---|---|---|---|---|
| Acyl | 1 (C=O) | | 165.2, s | [4]Phe | NH | 9.16, 1H, d (7.7) | |
| | 2 | 6.68, 1H, d (15.6) | 122.7, d | | a | 4.33 m | 55.7, d |
| | 3 | 7.42, 1H, d (15.6) | 137.4, d | | b | 3.28, 1H, m | 36.3, t |
| | 4 | | 133.1, s | | b' | 2.73, 1H, o. t (12.9) | |
| | 5 | 7.53, 1H, d (7.6) | 126.0, d | | 1 | | 138.7, s |
| | 6 | 7.32, 1H, o. m | 127.3, d | | 2,6 | 7.32, 2H, o. d (8.4) | 128.9, d |
| | 7 | 7.36, 1H, o. t (7.6) | 128.9, d | | 3,5 | 7.27, 2H, o. t, (7.6) | 127.9, d |
| | 8 | 7.20, 1H, o. m | 129.6, d | | 4 | 7.27, 1H, o. m | 127.9, d |
| | 9 | | 137.0, s | | C=O | | 170.1, s |
| | 10 | 6.50, 1H, d (11.2) | 126.8, d | [5]Thr | NH | 7.59, 1H, d (9.5) | |
| | 11 | 5.83, 1H, dt (11.7, 7.4) | 134.0, d | | a | 4.36, 1H, =tho. t (9.8) | 57.2, d |
| | 12 | 1.99, 2H, m | 29.9, t | | b | 4.26, 1H, m | 68.1, d |
| | 13 | 1.36, 2H, m (7.8) | 21.9, t | | g | 0.64, 3H, o. d (6.4) | 22.0, q |
| | 14 | 0.79, 3H, t (7.4) | 13.53, q | | OH | 5.18, 1H, d (2.9) | |
| [1]Thr | NH | 8.70, 1H, d (9.3) | | | C=O | | 169.9, s |
| | a | 5.33, 1H, t (9.6) | 53.2, d | [6]Asn | NH | 8.33, 1H, d (7.3) | |
| | b | 5.03, 1H, dq (9.8, 6.2) | 73.24, d | | a | 4.44, 1H, m | 50.8, d |
| | g | 1.15, 3H, d, 6.0 | 16.56, q | | b | 2.46, 1H, m | 36.7, t |
| | C=O | | 169.0, s | | b' | 2.41, 1H, dd (15.3, 9.9) | |
| [2]DTyr | NMe | 2.98, 3H, s | 34.2, q | | g C=O | | 171.2, s |
| | a | | 128.5, s | | g $NH_2$ | 6.93, 1H, s | |
| | b | 6.13, 1H, s | 131.6, d | | g' $NH_2$ | 7.30, 1H, o | |
| | 1 | | 122.9, s | | C=O | | 171.6, s |
| | 2,6 | 7.39, 2H, d (8.6) | 131.5, d | [7]Ser | NH | 8.48, 1H, d (9.6) | |
| | 3,5 | 6.59, 2H, d (8.6) | 114.8, d | | a | 4.33, 1H, o | 56.0, d |
| | 4 | | 158.1, s | | b | 3.26, 1H, o | 60.8, d |
| | C=O | | 165.6, s | | b' | 3.16, 1H, br. t (~9) | |
| [3]Leu | NH | 9.23, 1H, br. s | | | OH | 4.78, 1H, br | |
| | a | 4.07, 1H, m | 53.7, d | | C=O | | 168.8, s |
| | b | 1.26, 2H, m | 38.8, t | | | | |
| | g | 0.89, 1H, m | 23.4, d | | | | |
| | d | 0.63, 3H, o. d | 22.8, q | | | | |
| | d' | 0.75, 3H, d (6.3) | 22.0, q | | | | |
| | C=O | | 172.1, s | | | | |

[a]Assignments based on HSQC, COSY, TOCSY, HMBC, and ROESY experiments. o = overlapping signal, br = broad signal, n/d = no data



**Supplementary Table 3** High Resolution Mass Data

| *Compound* | *Molecular Formula* | *Calculated m/z* | *Observed m/z* | *Δppm* |
|---|---|---|---|---|
| Acidobactin A [M+H] | $C_{28}H_{48}N_7O_{16}$ | 738.31575 | 738.31705 | 5.245 |
| Acidobactin B [M+H] | $C_{28}H_{48}N_7O_{15}$ | 722.32283 | 722.31768 | 2.609 |
| Variobactin A [M+H] | $C_{47}H_{84}N_{11}O_{17}$ | 1074.60462 | 1074.6093 | 4.1 |
| Potensimicin [M+H] | $C_{28}H_{48}NO_8$ | 526.33744 | 526.33559 | 4.565 |
| Thanamycin A [M+H] | $C_{54}H_{88}ClN_{12}O_{22}$ | 1291.58192 | 1291.58220 | 0.205 |

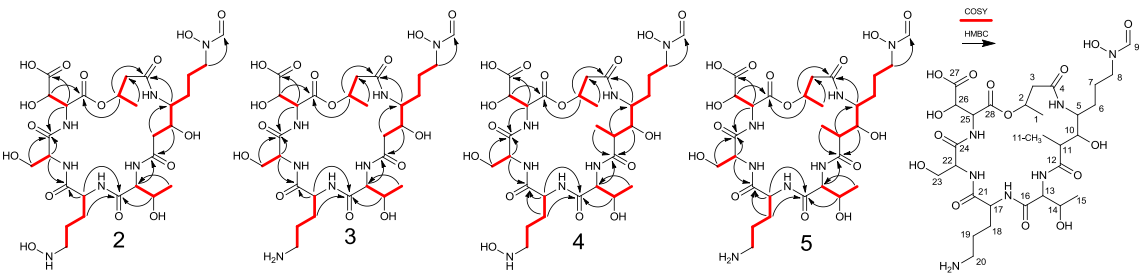
**Supplementary Table 4** Acidobactin gene cluster analysis

| Locus | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| Aave_3737 | RNA polymerase, sigma-24 subunit, ECF subfamily | - | 177 |
| Aave_3736 | MbtH domain protein | - | 95 |
| Aave_3735 | Thioesterase | - | 258 |
| Aave_3734 | TauD dioxygenase | - | 330 |
| Aave_3733 | NRPS | - | 1779 |
| Aave_3732 | PKS | - | 1535 |
| Aave_3731 | NRPS | - | 1130 |
| Aave_3730 | NRPS | - | 2651 |
| Aave_3729 | NRPS | - | 1368 |
| Aave_3728 | TonB siderophore receptor | + | 733 |
| Aave_3727 | N-acetyltransferase | + | 366 |
| Aave_3726 | N5-hydroxyornithine formyltransferase | + | 313 |
| Aave_3725 | Ferric iron reductase | + | 276 |
| Aave_3724 | Cyclic peptide transporter | - | 564 |
| Aave_3723 | L-lysine 6-monooxygenase | - | 452 |
| Aave_3722 | Phosphopantetheinyl transferase | - | 257 |

**Supplementary Table 5** NMR spectroscopic Data for acidobactin A (2), B (3), vacidobactin A (4), and B (5) (700 MHz, in $D_2O$)[a]

| position | $\delta_H$ mult. | | | | $\delta_C$ | | | |
|---|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **2** [a] | **3** | **4** [a] | **5** [a] |
| 1 | 1.25, *d*, 4.0 | 1.20, *d*, 4.2 | 1.19, *d*, 3.6 | 1.19, *d*, 3.6 | 18.6, $CH_3$ | 19.3, $CH_3$ | 19.0, $CH_3$ | 19.0, $CH_3$ |
| 2 | 5.34, *m* | 5.48, *m* | 5.38, *m* | 5.37, *m* | 68.9, CH | 70.7, CH | 70.6, CH | 70.7, CH |
| 3a | 2.51, *m* | 2.73, *m* | 2.58, *m* | 2.58, *m* | 34.9, $CH_2$ | 39.7, $CH_2$ | 38.4, $CH_2$ | 38.3, $CH_2$ |
| 3b | 2.80, *m* | 2.81, *m* | 2.77, *m* | 2.71, *m* | - | - | - | - |
| 4 | - | - | - | - | 164.2, C | 171.8, C | 164.6, C | 171.4, C |
| 5 | 3.36, *m* | 3.33, *m* | 3.31, *m* | 3.36, *m* | 54.4, CH | 54.9, CH | 52.8, CH | 52.9, CH |
| 6a | 1.61, *m* | 1.50, *m* | 1.63, *m* | 1.63, *m* | 23.1, $CH_2$ | 23.7, $CH_2$ | 22.4, $CH_2$ | 22.6, $CH_2$ |
| 6b | 1.59, *m* | 1.68, *m* | - | - | - | - | - | - |
| 7a | 1.71, *m* | 1.63, *m* | 1.83, *m* | 1.87, *m* | 25.4, $CH_2$ | 26.4, $CH_2$ | 28.6, $CH_2$ | 28.4, $CH_2$ |
| 7b | 1.80, *m* | 1.85, *m* | - | - | - | - | - | - |
| 8a | 3.55, *m* | 3.49, *m* | 3.38, *m* | 3.33, *m* | 49.9, $CH_2$ | 49.8, $CH_2$ | 52.1, $CH_2$ | 52.7, $CH_2$ |
| 8b | - | 3.90, *m* | - | 3.48, m | - | - | - | - |
| 9 | 7.90, *s* | 7.88, *s* | 7.74, *s* | 7.74, *s* | 153.6, CH | 160.0, CH | 155.4, CH | 155.5, CH |
| 10 | 4.22, *m* | 4.35, *m* | 4.30, m | 4.17, m | 68.0, CH | 71.9, CH | 72.3, CH | 66.9, CH |
| 11a | 2.79, *m* | 2.59, *m* | 2.71, *m* | 2.69, *m* | 39.0, $CH_2$ | 38.9, $CH_2$ | 43.2, CH | 43.1, CH |
| 11b | - | 2.69, *m* | - | - | - | - | - | - |
| 11-$CH_3$ | - | - | 1.03, *d* | 1.02, *d* | - | - | 12.8, $CH_3$ | 12.7, $CH_3$ |
| 12 | - | - | - | - | 173.4, C | 172.3, C | 177.0, C | 177.1, C |
| 13 | 4.23, *m* | 4.40, *d*, 1.8 | 4.41, *m* | 4.44, *m* | 58.7, CH | 60.8, CH | 56.2, CH | 55.6, CH |
| 14 | 4.22, *m* | 4.31, *m* | 4.26, *m* | 4.19, *m* | 66.3, CH | 66.5, CH | 66.8, CH | 66.4, CH |
| 15 | 1.10, *d*, 3.8 | 1.07, *d*, 4.0 | 1.10, *m* | 1.08, *m* | 18.8, $CH_3$ | 19.1, $CH_3$ | 19.2, $CH_3$ | 19.1, $CH_3$ |
| 16 | - | - | - | - | 174.1, C | 174.1, C | 172.3, C | 173.9, C |
| 17 | 4.36, *m* | 4.60, *t*, 3.6 | 4.40, *m* | 4.41, *m* | 53.6, CH | 54.1, CH | 55.3, CH | 55.1, CH |
| 18a | 1.62, *m* | 1.71, *m* | 1.60, *m* | 1.62, *m* | 27.7, $CH_2$ | 27.9, $CH_2$ | 28.4, $CH_2$ | 28.3, $CH_2$ |
| 18b | 1.84, *m* | 2.61, *m* | - | - | - | - | - | - |
| 19a | 1.71, *m* | 1.81, *m* | 1.85, *m* | 1.81, *m* | 23.7, $CH_2$ | 22.1, $CH_2$ | 22.6, $CH_2$ | 22.4, $CH_2$ |
| 19b | 1.87, *m* | 2.12, *m* | | | - | - | - | - |
| 20a | 3.55, *m* | 3.32, *m* | 3.41 | 3.37 | 50.3, $CH_2$ | 49.3, $CH_2$ | 51.4, $CH_2$ | 49.8, $CH_2$ |
| 20b | - | 3.96, *m* | - | 3.65 | - | - | - | - |
| 21 | - | - | - | - | 172.9, C | 174.2, C | 172.7, C | 174.9, C |
| 22 | 4.41, *m* | 4.44, *m* | 4.30, *m* | 4.27, *m* | 56.3, CH | 55.7, CH | 54.7, CH | 54.4, CH |
| 23a | 3.83, *d*, 8.5 | 3.84, *d*, 9.0 | 3.90, *m* | 3.85, *m* | 61.3, CH | 61.1, CH | 61.3, CH | 61.5, CH |
| 23b | - | 3.95, *d*, 9.0 | - | - | - | - | - | - |
| 24 | - | - | - | - | 171.6, C | 172.5, C | 169.3, C | 169.4, C |
| 25 | 4.82, *m* | 4.50, *m* | 4.80, *m* | 4.59, *m* | 56.8, CH | 56.2, CH | 56.4, CH | 51.9, CH |
| 26 | 4.31, *m* | 4.25, *m* | 3.90, *m* | 3.87, *m* | 74.7, CH | 72.4, CH | 73.9, CH | 73.6, CH |
| 27 | - | - | - | - | 180.0, C | 175.6, C | 179.6, C | 181.4, C |
| 28 | - | - | - | - | 169.6, C | 169.1, C | 177.9, C | 177.7, C |

[a] $^{13}C$ NMR were extracted from HMBC spectra.



**Supplementary Table 6** Vacidobactin gene cluster analysis

| Locus | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| Vapar_3752 | RNA polymerase, sigma-24 subunit, ECF subfamily | - | 181 |
| Vapar_3751 | anti-FecI sigma factor | - | 67 |
| Vapar_3750 | MbtH domain protein | - | 85 |
| Vapar_3749 | Thioesterase | - | 246 |
| Vapar_3748 | Phosphopantetheinyl transferase | - | 229 |
| Vapar_3747 | TauD dioxygenase | - | 330 |
| Vapar_3746 | NRPS | - | 1776 |
| Vapar_3745 | PKS | - | 1520 |
| Vapar_3744 | NRPS | - | 1110 |
| Vapar_3743 | NRPS | - | 2626 |
| Vapar_3742 | NRPS | - | 1358 |
| Vapar_3741 | TonB siderophore receptor | + | 723 |
| Vapar_3740 | L-lysine 6-monooxygenase | + | 439 |
| Vapar_3739 | N-acetyltransferase | + | 344 |
| Vapar_3738 | N5-hydroxyornithine formyltransferase | + | 281 |
| Vapar_3737 | Ferric iron reductase | + | 281 |
| Vapar_3736 | Cyclic peptide transporter | - | 563 |

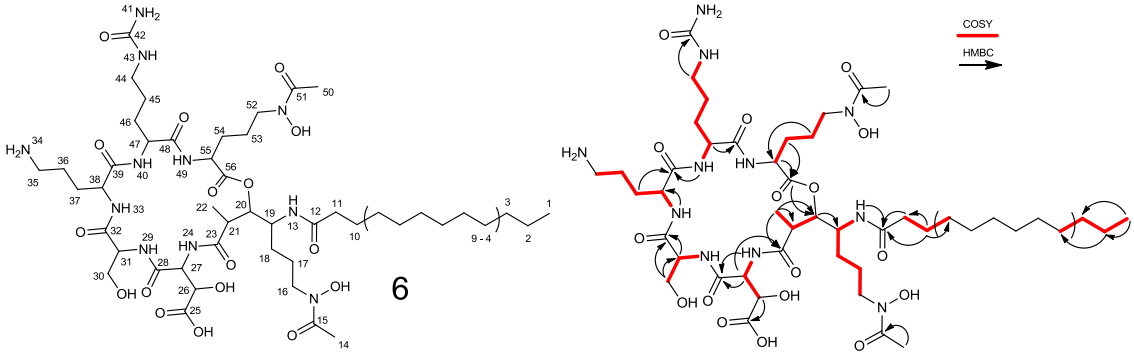
**Supplementary Table 7** Variobactin gene cluster analysis

| Locus | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| Var1 | Thioesterase | + | 278 |
| Var2 | Phosphopantetheinyl transferase | + | 234 |
| Var3 | NRPS | + | 1741 |
| Var4 | PKS | + | 2346 |
| Var5 | NRPS | + | 1038 |
| Var6 | NRPS | + | 2585 |
| Var7 | NRPS | + | 2431 |
| Var8 | TonB receptor | - | 816 |
| Var9 | FecR-like protein | - | 343 |
| Var10 | RNA polymerase subunit sigma-24 | - | 193 |
| Var11 | L-lysine 6-monooxygenase | - | 440 |
| Var12 | N-acetyltransferase | - | 369 |
| Var13 | Ferric iron reductase | - | 262 |
| Var14 | Cyclic peptide transporter | + | 560 |

**Supplementary Table 8** NMR spectroscopic data for variobactin A (6) (700 MHz, in DMSO-$d_6$)[a].

| Position | $\delta_H$ mult. | $\delta_C$ | Position | $\delta_H$ mult. | $\delta_C$ |
|---|---|---|---|---|---|
| 1 | 0.89 (*t*, 7.2) | 14.4, $CH_3$ | 34 | - | - |
| 2 | 1.23 (*ov*) | 22.5, $CH_2$ | 35a | 3.81 (*m*) | 49.4, $CH_2$ |
| 3 | 1.23 (*ov*) | 31.7, $CH_2$ | 35b | 3.36 (*m*) | |
| 4-9 | 1.23 (*ov*) | 28.5-29.5, $CH_2$ | 36 | 2.06 (*m*) | 23.7, $CH_2$ |
| 10 | 1.45 (*m*) | 25.9, $CH_2$ | 37a | 1.45 (m) | 29.3, $CH_2$ |
| 11 | 2.04 (*m*) | 36.0, $CH_2$ | 37b | 1.14 (m) | |
| 12 | - | 172.2, C | 38 | 3.59 (*ov*) | 50.4, CH |
| 13 | 7.84 (*broad*) | - | 39 | - | 171.1, C |
| 14 | 2.03 (*s*) | 16.6, $CH_3$ | 40 | 7.78 (*d*, 9.9) | - |
| 15 | - | 159.8, C | 41 | - | - |
| 16a | 3.90 (*m*) | 47.2, $CH_2$ | 42 | - | 157.3, C |
| 16b | 3.70 (*m*) | | 43 | 8.06 (*broad*) | - |
| 17 | 1.85 (*m*) | 23.7, $CH_2$ | 44a | 2.94 (*m*) | 40.9, $CH_2$ |
| 18a | 1.85 (*m*) | 29.0, CH2 | 44b | 3.15 (*m*) | |
| 18b | 1.56 (*m*) | | 45a | 1.22 (*m*) | 25.5, $CH_2$ |
| 19 | 4.36 (*m*) | 49.9, CH | 45b | 1.46 (*m*) | |
| 20 | 3.40 (*m*) | 74.5, CH | 46a | 2.37 (*m*) | 28.5, $CH_2$ |
| 21 | 2.14 (*m*) | 44.3, CH | 46b | 1.63 (*m*) | |
| 22 | 1.26 (*ov*) | 16.8, $CH_3$ | 47 | 4.64 (*m*) | 50.5, CH |
| 23 | - | 175.3, C | 48 | - | 172.1, C |
| 24 | 9.04 (*d*, 7.6) | - | 49 | - | - |
| 25 | - | 179.6, C | 50 | 1.99 (*s*) | 16.9, $CH_3$ |
| 26 | 4.29 (*d*, 2.1) | 74.1, CH | 51 | - | 161.7, C |
| 27 | 4.19 (*d*, 9.3) | 58.8, CH | 52a | 3.66 (*m*) | 48.8, $CH_2$ |
| 28 | - | 174.4, C | 52b | 3.32 (*m*) | |
| 29 | 6.89 (*d*, 4.9) | - | 53a | 1.74 (*m*) | 22.0, $CH_2$ |
| 30a | 3.70 (*d*, 5.7) | 62.9, $CH_2$ | 53b | 1.28 (*m*) | |
| 30b | 3.18 (*m*) | | 54a | 1.99 (*m*) | 30.1, $CH_2$ |
| 31 | 4.52 (*m*) | 53.8, CH | 54b | 1.86 (*m*) | |
| 32 | - | 170.3, C | 55 | 4.20 (*d*, 9.3) | 61.3, CH |
| 33 | 7.26 (*d*, 8.66) | - | 56 | - | 170.6, C |

[a]Chemical shift $\delta$ and (multiplicity, *J* in Hz).



**Supplementary Table 9** Hidden Markov models for sugar biosynthesis

| Domain | Abbrev. | Hidden Markov model | # Seq. | Cutoff |
|---|---|---|---|---|
| **2,3-dehydratase** | 2,3DH | 2_3_dehydratase.hmm | 23 | 100.0 |
| **3,4-dehydratase** | 3,4DH | 3_4_dehydratase.hmm | 10 | 600.0 |
| **3,4-ketoisomerase** | 3,4IM | 3_4_ketoisomerase.hmm | 8 | 150.0 |
| **3-aminotransferase** | 3-AmT | 3_aminotransferase.hmm | 13 | 451.0 |
| **3-ketoreductase** | 3KR | 3_ketoreductase.hmm | 15 | 200.0 |
| **4,6-dehydratase** | 4,6DH | 4_6_dehydratase.hmm | 21 | 400.0 |
| **4-aminotransferase** | 4-AmT | 4_aminotransferase.hmm | 9 | 350.0 |
| **4-ketoreductase** | 4KR | 4_ketoreductase.hmm | 26 | 190.0 |
| **Acetyltransferase** | AcT | acetyltransferase.hmm | 4 | 160.0 |
| **Carbamoyltransferase** | CarbT | carbamoyltransferase.hmm | 2 | 845.0 |
| **Decarboxylase** | UDP-DC | decarboxylase_epimerase.hmm | 15 | 400.0 |
| **Dehydrogenase** | UDP-DH | pentose_dehydrogenase.hmm | 15 | 400.0 |
| **Epimerase** | E | epimerase.hmm | 22 | 160.0 |
| **Glycosyltransferase** | GTr | glycosyltransferase.hmm | 82 | 125.0 |
| **N-ethyltransferase** | N-ET | n_ethyltransferase.hmm | 2 | 500.0 |
| **N,N-dimethyltransferase** | N,N-MT | n_n_dimethyltransferase.hmm | 10 | 147.0 |
| **Oxidative deaminase** | OxDA | oxidative_deaminase.hmm | 10 | 500.0 |
| **Oxidoreductase** | Ox | oxidoreductase.hmm | 6 | 500.0 |
| **Pyrrolyltransferase** | PyT | pyrrolyltransferase.hmm | 4 | 350.0 |

| Domain | Abbrev. | Hidden Markov model | # Seq. | Cutoff |
|---|---|---|---|---|
| **C-methyltransferase** | CMT | sugar_c_methyltransferase.hmm | 20 | 400.0 |
| **N-methyltransferase** | NMT | sugar_n_methyltransferase.hmm | 5 | 180.0 |
| **O-methyltransferase** | OMT | sugar_o_methyltransferase.hmm | 40 | 147.0 |
| **Thiosugar synthase** | ThiS | thiosugar_synthase.hmm | 3 | 300.0 |

**Supplementary Table 10** Rules for sugar biosynthesis

| Sugar | SMILES | Genes |
|---|---|---|
| **L-aculose** | CC1O[C@H](C=CC1=O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>Oxidoreductase |
| **L-cinerulose A** | CC1O[C@H](CCC1=O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |
| **L-rhodinose** | C[C@@H]1O[C@@H](CC[C@@H]1O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |
| **Rednose** | CC1OC(C(N)=CC1=O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |

| Sugar | SMILES | Genes |
|---|---|---|
| **L-cinerulose B** | CC1O[C@H]([C@@H](O)CC1=O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |
| **O-methyl-L-amicetose** | COC1CC[C@@H](O[C@H]1C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **4-O-methyl-L-rhodinose** | CO[C@H]1CC[C@@H](O[C@H]1C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **L-daunosamine** | C[C@@H]1OC(C[C@@H]([C@@H]1O)N)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>3-aminotransferase |
| **L-ristosamine** | CC1OC(CC(C1O)N)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>4-aminotransferase |
| **D-digitoxose** | CC1OC(CC(C1O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase |
| **L-digitoxose** | CC1OC(CC(C1O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |

| Sugar | SMILES | Genes |
|---|---|---|
| **2-deoxy-L-fucose** | CC1OC(CC(C1O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |
| **D-olivose** | CC1O[C@@H](C[C@H]([C@@H]1O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase |
| **D-oliose** | CC1O[C@H](C[C@H]([C@H]1O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase |
| **4-oxo-L-vancosamine** | C[C@@H]1OC(C[C@](N)(C1=O)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>Epimerase<br>4-aminotransferase<br>C-methyltransferase |
| **D-forosamine** | CC1OC(CC[C@@H]1N(C)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3,4-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>N,N-dimethyltransferase |
| **L-actinosamine** | COC1C(OC(CC1N)O)C | 4,6-dehydratase<br>2,3-dehydratase |
| **L-vancosamine** | OC1O[C@H]([C@@H](O)[C@](C1)(N)C)C | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>3-aminotransferase<br>C-methyltransferase |
| **L-vicenisamine** | CN[C@@H]1C(CC(OC1C)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>N-methyltransferase |

| Sugar | SMILES | Genes |
|---|---|---|
| **D-chalcose** | CO[C@H]1C[C@H](OC([C@@H]1O)O)C | 4,6-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>O-methyltransferase<br>Oxidative deaminase |
| **D-mycarose** | CC1OC(C[C@](O)([C@H]1O)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **L-oleandrose** | CO[C@H]1C[C@H](O[C@H]([C@@H]1O)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **Olivomose** | COC1C(CC(OC1C)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>O-methyltransferase |
| **D-mycosamine** | C[C@H]1O[C@H]([C@H]([C@H]([C@@H]1O)N)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>O-methyltransferase |
| **4-deoxy-4-thio-D-digitoxose** | CC1O[C@H](C[C@H](C1S)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>Epimerase<br>Thiosugar synthase |
| **D-fucofuranose** | C[C@H]([C@H]1O[C@H]([C@H](C1O)O)O)O | 4,6-dehydratase<br>4-ketoreductase |
| **D-fucose** | CC1OC([C@@H]([C@H]([C@H]1O)O)O)O | 4,6-dehydratase<br>4-ketoreductase |
| **D-rhamnose** | C[C@@H]1O[C@@H]([C@@H]([C@@H]([C@H]1O)O)O)O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase |

| Sugar | SMILES | Genes |
|---|---|---|
| **4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose** | CCN[C@H]1CO[C@H](C[C@H]1OC)O | Dehydrogenase<br>Decarboxylase<br>2,3-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>N-ethyltransferase |
| **D-3-N-methyl-4-O-methyl-L-ristosamine** | CN[C@H]1CC(OC([C@@H]1OC)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>4-aminotransferase<br>N-methyltransferase<br>O-methyltransferase |
| **N,N-dimethyl-L-pyrrolosamine** | CC1OC(CC(C1N(C)C)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>Epimerase<br>4-aminotransferase<br>N,N-dimethyltransferase |
| **D-desosamine** | C[C@@H]1C[C@H](N(C)C)[C@@H](O)[C@H](O)O1 | 4,6-dehydratase<br>3,4-dehydratase<br>3-aminotransferase<br>N,N-dimethyltransferase<br>Oxidative deaminase |
| **L-megosamine** | C[C@@H]1OC(C[C@@H](N(C)C)[C@H]1O)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>3-aminotransferase<br>N,N-dimethyltransferase |
| **Nogalamine** | OC1O[C@@H](C)[C@H](O)[C@@H](N(C)C)[C@@H]1O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>3-aminotransferase<br>N,N-dimethyltransferase |
| **L-rhodosamine** | CC1O[C@H](CC(N(C)C)[C@H]1O)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>3-aminotransferase<br>N,N-dimethyltransferase |

| Sugar | SMILES | Genes |
|---|---|---|
| **D-angolosamine** | C[C@@H]1OC(CC(N(C)C)[C@H]1O)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>3-aminotransferase<br>N,N-dimethyltransferase |
| **Kedarosamine** | OC1O[C@@H](C)[C@@H](N(C)C)[C@@H](O)C1 | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>N,N-dimethyltransferase |
| **L-noviose** | CC1(C)[C@H](O)[C@@H](O)[C@@H](O)C(O)O1 | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase |
| **L-cladinose** | C[C@H]1[C@H](O)[C@](C)(OC)CC(O)O1 | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>O-methyltransferase |
| **2-N-methyl-D-fucosamine** | CN[C@H]1C(O[C@@H]([C@@H]([C@@H]1O)O)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>N-methyltransferase |
| **D-digitalose** | CO[C@H]1[C@H]([C@H](O[C@H]([C@@H]1O)O)C)O | 4,6-dehydratase<br>4-ketoreductase<br>O-methyltransferase |
| **2-O-methyl-rhamnose** | COC1[C@H]([C@H](OC([C@H]1O)O)C)O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **3-O-methyl-rhamnose** | COC1[C@@H](OC([C@@H](C1O)O)C)O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **6-deoxy-3-C-methyl-L-mannose** | C[C@@H]1OC([C@@H]([C@](O)([C@H]1O)C)O)O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase |
| **4,6-dideoxy-4-hydroxylamino-D-glucose** | CC1OC(C(C(C1NO)O)O)O | 4,6-dehydratase<br>4-aminotransferase |

| Sugar | SMILES | Genes |
|---|---|---|
| **3-N,N-dimethyl-L-eremosamine** | OC1C[C@](C)(N(C)C)[C@@H](O)[C@H](C)O1 | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>N,N-dimethyltransferase |
| **Chromose** | CC1OC(CC(C1OC(C)=O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Acetyltransferase |
| **4-O-carbamoyl-D-olivose** | CC1OC(CC(C1OC(N)=O)O)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Carbamoyltransferase |
| **D-ravidosamine** | C[C@H]1O[C@@H]([C@@H]([C@@H](N(C)C)[C@H]1O)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>N,N-dimethyltransferase |
| **3-N,N-dimethyl-D-mycosamine** | C[C@H]1O[C@H]([C@@H]([C@@H](N(C)C)[C@@H]1O)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>N,N-dimethyltransferase |
| **2,3-O-dimethyl-L-rhamnose** | COC1[C@@H](OC([C@@H](C1OC)O)C)O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **2,4-O-dimethyl-L-rhamnose** | CO[C@H]1C(O[C@H](C(C1O)OC)O)C | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **3,4-O-dimethyl-L-rhamnose** | CO[C@H]1C(O[C@H](C(C1OC)O)O)C | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **2-thioglucose** | OC1[C@H](S)[C@@H](O)[C@H](O)[C@@H](CO)O1 | Thiosugar synthase |

| Sugar | SMILES | Genes |
|---|---|---|
| **Olivomycose** | C[C@@H]1O[C@H](C[C@](O)([C@H]1OC(C)=O)C)O | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>Acetyltransferase |
| **4-N,N-dimethylamino-4-deoxy-5-C-methyl-l-rhamnose** | CN(C1C(C(C(OC1(C)C)O)O)O)C | 4,6-dehydratase<br>Epimerase<br>4-aminotransferase<br>C-methyltransferase<br>N,N-dimethyltransferase<br>Acetyltransferase |
| **2,3,4-tri-O-methylrhamnose** | CO[C@@H]1[C@H](O[C@@H]([C@H]([C@H]1OC)OC)O)C | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>O-methyltransferase |
| **4-O-acetyl-L-arcanose** | OC1C[C@@](C)(OC)[C@H](OC(C)=O)[C@H](C)O1 | 4,6-dehydratase<br>2,3-dehydratase<br>3-ketoreductase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>Acetyltransferase |
| **3-N-acetyl-D-ravidosamine** | C[C@H]1O[C@@H]([C@@H]([C@@H](N(C(C)=O)C)[C@H]1OC(C)=O)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>N,N-dimethyltransferase<br>Acetyltransferase |
| **3-O-carbamoyl-L-noviose** | CC1([C@@H]([C@H]([C@H](C(O1)O)O)OC(N)=O)O)C | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>C-methyltransferase<br>Carbamoyltransferase |
| **L-nogalose** | COC1C(OC(C(C1(OC)C)OC)O)C | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>O-methyltransferase |

| Sugar | SMILES | Genes |
|---|---|---|
| **4-O-acetyl-D-ravidosamine** | C[C@H]1O[C@@H]([C@@H]([C@@H](N(C)C)[C@H]1OC(C)=O)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>N,N-dimethyltransferase<br>Acetyltransferase |
| **3-O-carbamoyl-4-O-methyl-L-noviose** | CC1(C)[C@H](OC)[C@@H](OC(N)=O)[C@@H](O)C(O)O1 | 4,6-dehydratase<br>2,3-dehydratase<br>4-ketoreductase<br>C-methyltransferase<br>O-methyltransferase<br>Carbamoyltransferase |
| **3-N-acetyl-4-O-acetyl-D-ravidosamine** | C[C@H]1O[C@@H]([C@@H]([C@@H](N(C(C)=O)C)[C@H]1OC(C)=O)O)O | 4,6-dehydratase<br>3,4-ketoisomerase<br>3-aminotransferase<br>N,N-dimethyltransferase<br>Acetyltransferase |
| **3-(5'-methyl-2'-pyrrolylcarbonyl-)4-O-methyl-L-noviose** | CO[C@@H]1[C@@H](C([C@@H](OC1(C)C)O)O)OC(C2=CC=C(N2)C)=O | 4,6-dehydratase<br>4-ketoreductase<br>Epimerase<br>C-methyltransferase<br>O-methyltransferase<br>Pyrrolyltransferase |
| **Madurose** | O[C@@H]1[C@H](O)[C@@](O)(C)[C@H](N)CO1 | Dehydrogenase<br>Decarboxylase<br>C-methyltransferase<br>4-aminotransferase |
| **4-N-methyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose** | O[C@@H]1C[C@H](OC)[C@@H](NC)CO1 | Dehydrogenase<br>Decarboxylase<br>2,3-dehydratase<br>3-ketoreductase<br>4-aminotransferase<br>N-methyltransferase |

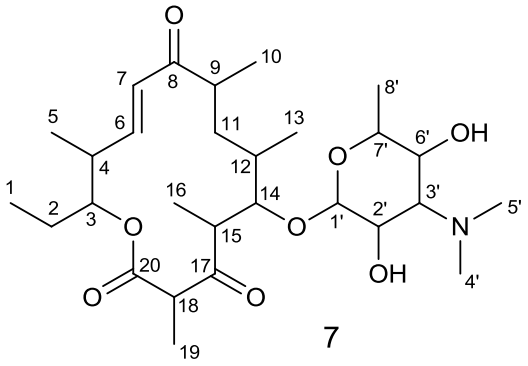
**Supplementary Table 11** Potensimicin gene cluster analysis

| Locus | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| D459DRAFT_04885 | Glycosyltransferase | + | 401 |
| D459DRAFT_04886 | Hypothetical protein | + | 99 |
| D459DRAFT_04887 | Hypothetical protein | + | 281 |
| D459DRAFT_04888 | Hypothetical protein | + | 244 |
| D459DRAFT_04889 | NDP-hexose-2,3-dehydratase | + | 449 |
| D459DRAFT_04890 | dTDP-4-amino-4,6-transaminase | + | 374 |
| D459DRAFT_04891 | N,N-dimethyltransferase | + | 247 |
| D459DRAFT_04892 | dTDP-4-dehydrorhamnose 3,5-epimerase | + | 228 |
| D459DRAFT_04893 | 4-ketoreductase | + | 341 |
| D459DRAFT_04894 | Hypothetical protein | + | 271 |
| D459DRAFT_04895 | AAA ATPase domain | - | 989 |
| D459DRAFT_04896 | Polyketide synthase | + | 4564 |
| D459DRAFT_04897 | Polyketide synthase | + | 3658 |
| D459DRAFT_05970 | Polyketide synthase | + | 1627 |
| D459DRAFT_05969 | Polyketide synthase | + | 1353 |
| D459DRAFT_04117 | Beta-glucosidase-related glycosidase | + | 485 |
| D459DRAFT_04116 | Cytochrome P450 | + | 399 |
| D459DRAFT_04115 | Glycosyltransferase | + | 424 |
| D459DRAFT_04114 | N,N-dimethyltransferase | + | 239 |
| D459DRAFT_04113 | Beta-glucosidase-related glycosidase | + | 793 |
| D459DRAFT_04112 | MFS / Sugar transport protein | + | 418 |
| D459DRAFT_04111 | Phosphohydrolase | - | 281 |
| D459DRAFT_04110 | Glucose-1-phosphate thymidylyltransferase | + | 285 |
| D459DRAFT_04109 | dTDP-4-amino-4,6-dehydratase | + | 341 |
| D459DRAFT_04108 | 3,5-ketoisomerase | + | 150 |
| D459DRAFT_04107 | dTDP-4-amino-4,6-transaminase | + | 389 |
| D459DRAFT_04106 | Adenine-N(6)-methyltransferase | + | 293 |

**Supplementary Table 12** NMR spectroscopic data for potensimicin (7) (700 MHz, in DMSO-$d_6$)[a].

| Position | $\delta_H$ (mult.) | $\delta_C$ | Position | $\delta_H$ (mult.) | $\delta_C$ |
|---|---|---|---|---|---|
| 1 | 0.80 (*t,* 7.30) | 9.36 | 14 | 3.78 (*m*) | 79.42 |
| 2a | 1.54 (*m*) | 22.50 | 15 | 2.83 (*quint.*, 7.19, 7.37) | 48.72 |
| 2b | 1.54 (*m*) | | 16 | 1.22 (*d*, 7.32) | 12.7 |
| 3 | 4.83 (*m*) | 76.5 | 17 | - | 208.17 |
| 4 | 2.63 (*m*) | 37.1 | 18 | 3.96 (*quart.*, 6.74, 6.89) | 49.1 |
| 5 | 0.98 (*d*, 6.02) | 10.03 | 19 | 1.15 (*d*, 6.84) | 13.84 |
| 6 | 6.64 (*m*) | 126.1 | 20 | - | 169.25 |
| 7 | 6.14 (*d*, 14.2) | 146.8 | 1' | 4.17 (*d*, 7.06) | 102.8 |
| 8 | - | 201.95 | 2' | 3.23 (*t*, 8.91) | 68.17 |
| 9 | 2.51 (*ov*) | 41.93 | 3' | 2.44 (*ov*) | 70.0 |
| 10 | 1.00 (*d*, 6.81) | 15.07 | 4' | 2.49 (*ov*) | 40.73 |
| 11a | 1.49 (*m*) | 34.37 | 5' | 2.49 (*ov*) | 40.73 |
| 11b | 1.08 (*m*) | | 6' | 3.05 (*t*, 9.31) | 69.1 |
| 12 | - | - | 7' | 3.19 (*quint.*, 5.64, 6.84) | 71.5 |
| 13 | 0.87 (*d*, 6.58) | 16.00 | 8' | 1.12 (*d*, 5.94) | 16.96 |

[a]Chemical shift $\delta$ and (multiplicity, $J$ in Hz).

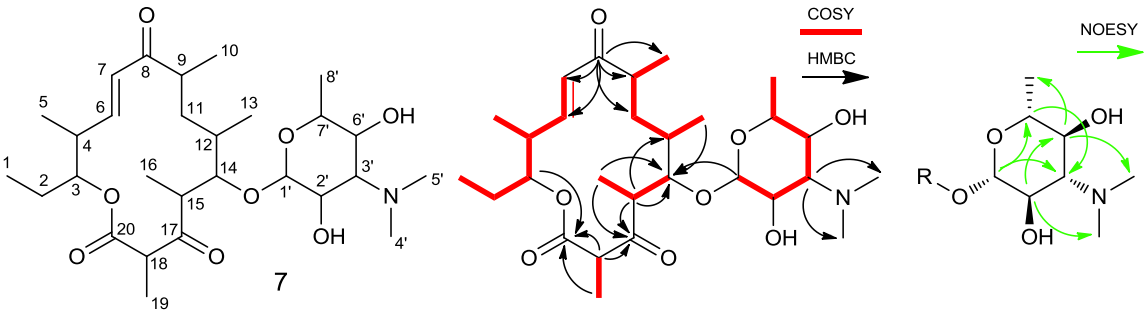



7

**Supplementary Table 13** NMR spectroscopic data for potensimicin (7) (700 MHz, in $CDCl_3$)[a].

| Position | $\delta_H$ (mult.) | $\delta_C$ | Position | $\delta_H$ (mult.) | $\delta_C$ |
|---|---|---|---|---|---|
| 1 | 0.79 (*m*) | 9.95 | 14 | 3.98 (*m*) | 78.56 |
| 2a | 1.43 (*ov*) | 22.48 | 15 | 2.78 (*m*) | 48.57 |
| 2b | 1.48 (*ov*) | | 16 | 1.21 (*ov*) | 13.1 |
| 3 | 4.79 (*m*) | 77.9 | 17 | - | 207.86 |
| 4 | 2.6 (*ov*) | 37.77 | 18 | 3.75 (*m*) | 49.93 |
| 5 | 0.98 (*m*) | 11.58 | 19 | 1.21 (*ov*) | 13.92 |
| 6 | 6.56 (*d*, 13.45) | 146.83 | 20 | - | 168.82 |
| 7 | 5.99 (*d*, 14.7) | 126.84 | 1' | 4.22 (*m*) | 103.38 |
| 8 | - | 202.59 | 2' | 3.35 (*m*) | 69.04 |
| 9 | 2.59 (*ov*) | 41.61 | 3' | 2.53 (*ov*) | 70.93 |
| 10 | 0.97 (*m*) | 15.12 | 4' | 2.56 (*ov*) | 41.21 |
| 11a | 1.38 (*ov*) | 35.94 | 5' | 2.56 (*ov*) | 41.21 |
| 11b | 1.07 (*m*) | | 6' | 3.11 (*m*) | 69.66 |
| 12 | 1.5 (*ov*) | 35.64 | 7' | 3.22 (*m*) | 72.5 |
| 13 | 0.87 (*m*) | 16.51 | 8' | 1.16 (*m*) | 17.22 |

[a]Chemical shift $\delta$ and (multiplicity, *J* in Hz).



**Supplementary Table 14** Minimum inhibitory concentrations (MICs) of potensimicin and thanamycin.

| | *B. subtilis* 168 | *S. aureus* Newman |
|---|---|---|
| Potensimicin | 2 μg/mL | 8 μg/mL |
| Erythromycin | 0.125 μg/mL | 0.5 μg/mL |

| | *S. cerevisiae* |
|---|---|
| Thanamycin | 0.0625 μg/mL |
| Syringomycin E | 2 μg/mL |

**Supplementary Table 15** Detection of natural product standards at low concentrations.

| Natural Product | Class | Concentration (μg/mL) | Detected Scans | Avg. P1 | Avg. P2 |
|---|---|---|---|---|---|
| Daptomycin | Lipodepsipeptide | 1 | 3 | 37.0 | 26.5 |
| Erythromycin | Glycosylated polyketide | 1 | 6 | 23.3 | 17.4 |
| Thiostrepton | Ribosomal thiopeptide | 1 | 3 | 7.1 | 12.6 |
| Lincomycin | Lincosamide | 1 | 7 | 8.8 | 18.3 |
| Novobiocin | Coumarin | 1 | 4 | 57.9 | 23.2 |
| Vancomycin | Glycopeptide | 1 | 7 | 45.8 | 49.0 |
| Capreomycin | Cyclic peptide | 10 | 5 | 7.3 | 9.5 |
| Nystatin | Glycosylated polyketide | 1 | 4 | 35 | 9.7 |
| Bacitracin | Branched cyclic peptide | 1 | 2 | 22.8 | 34.2 |
| Gramicidin A | Nonribosomal peptide | 1 | 8 | 40.9 | 40.4 |
| Polymyxin B | Branched cyclic peptide | 1 | 3 | 24.5 | 30.3 |
| Valinomycin | Cyclic depsipeptide | 1 | 7 | 24.8 | 16.0 |

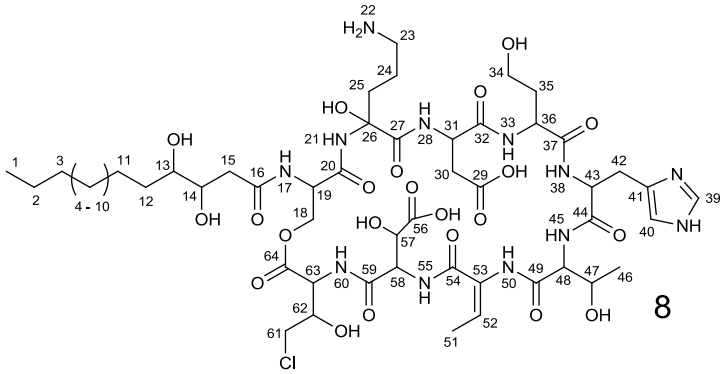
**Supplementary Table 16** Thanamycin gene cluster analysis

| Locus | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| ThaE | Cyclic peptide ABC transporter | - | 566 |
| ThaF | SyrP hydroxylase | - | 353 |
| ThaC1 | NRPS | + | 615 |
| ThaC2 | Chlorinating enzyme | + | 312 |
| ThaD | Thioesterase | + | 415 |
| ThaG | Branched chain amino acid permease | + | 233 |
| ThaH | Branched chain amino acid permease | + | 104 |
| ThaI | AraC family transcription factor | + | 172 |
| ThaJ | putative beta hydroxylase | - | 204 |
| ThaK | MFS transporter | - | 457 |
| ThaL | putative transcription factor | - | 335 |
| ThaA | NRPS | + | 5380 |
| ThaB | NRPS | + | 4236 |

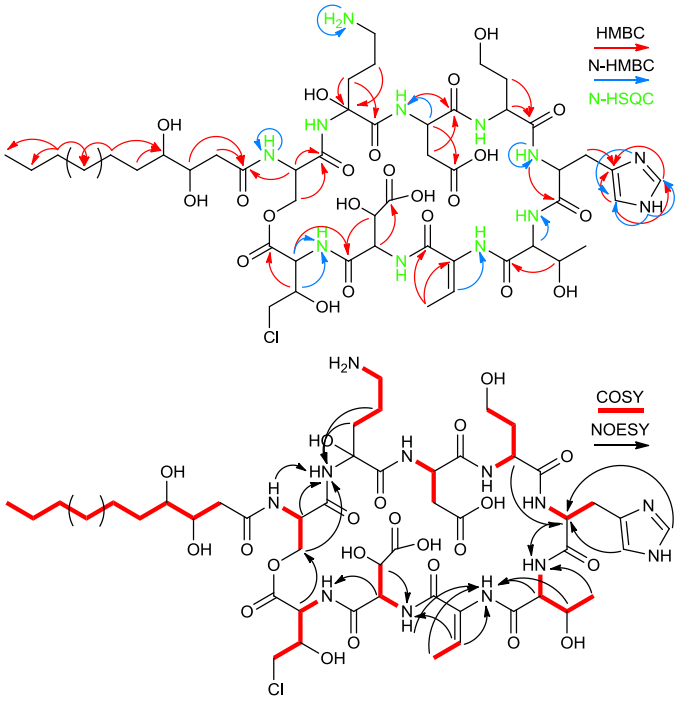
**Supplementary Table 17** NMR spectroscopic data for thanamycin (8) (700 MHz, in DMSO-$d_6$)[a].

| Position | $\delta_H$ mult. | $\delta_C$ | Position | $\delta_H$ mult. | $\delta_C$ |
|---|---|---|---|---|---|
| 1 | 0.84 (*t*, 6.1) | 14 | 34a | 3.36 (*m*) | 57.46 |
| 2 | 1.24 (*ov*) | 22.17 | 34b | 3.3 (*m*) | |
| 3 | 1.21 (*ov*) | 31.38 | 35a | 2.03 (*m*) | 33.33 |
| 4-10 | 1.23 (*ov*) | 29.17 | 35b | 1.84 (*ov*) | |
| 11a | 1.4 (*ov*) | 25.42 | 36 | 4.09 (*m*) | 51.92 |
| 11b | 1.2 (*ov*) | | 37 | - | 172.78 |
| 12a | 1.48 (*m*, 11.7) | 32.87 | 38 | 8.05 (*ov*) | - |
| 12b | 1.2 (*ov*) | | 39 | 8.97 (*s*) | 133.82 |
| 13 | 3.18 (*m*) | 73.57 | 40 | 7.24 (*s*) | 116.9 |
| 14 | 3.56 (*m*) | 71.63 | 41 | - | 129.4 |
| 15a | 2.36 (*d*, 13.8) | 39.15 | 42a | 3.23 (*m*) | 26.96 |
| 15b | 2.14 (*dd*, 10.6) | | 42b | 2.98 (*m*) | |
| 16 | - | 171.81 | 43 | 4.71 (*ov*) | 51.01 |
| 17 | 7.7 (*ov*) | - | 44 | - | 171.85 |
| 18a | 4.38 (*m*) | 63.89 | 45 | 8.07 (*ov*) | - |
| 18b | 4.16 (*ov*) | | 46 | 1.19 (*ov*) | 20.61 |
| 19 | 4.66 (*m*) | 50.68 | 47 | 3.84 (*m*) | 66.59 |
| 20 | - | 168.64 | 48 | 4.18 (*ov*) | 60.17 |
| 21 | 9.04 (*s*) | - | 49 | - | 170.42 |
| 22 | 7.85 (*ov*) | - | 50 | 9.65 (*s*) | - |
| 23a | 2.76 (*m*) | 38.85 | 51 | 1.67 (*d*, 5.6) | 13.07 |
| 23b | 2.76 (*m*) | | 52 | 6.47 (*m*) | 129.78 |
| 24a | 1.63 (*ov*) | 20.98 | 53 | - | 131.7 |
| 24b | 1.44 (*ov*) | | 54 | - | 163.69 |
| 25a | 1.8 (*m*) | 35.33 | 55 | 7.62 (*ov*) | - |
| 25b | 1.74 (*m*) | | 56 | - | 169.47 |
| 26 | - | 81.31 | 57 | 4.69 (*ov*) | 71.38 |
| 27 | - | 171.8 | 58 | 4.94 (*m*) | 55.61 |
| 28 | 8.21 (*ov*) | - | 59 | - | 169.33 |
| 29 | - | 173.03 | 60 | 7.94 (*ov*) | - |
| 30a | 2.71 (*ov*) | 36.43 | 61a | 3.51 (*m*) | 45.61 |
| 30b | 2.48 (*ov*) | | 61b | 3.41 (*m*) | |
| 31 | 4.56 (*m*) | 49.67 | 62 | 4.2 (*ov*) | 71.2 |
| 32 | - | 170.36 | 63 | 4.78 (*d*, 7.5) | 54.3 |
| 33 | 7.89 (*ov*) | - | 64 | - | 170.54 |

[a]Chemical shift $\delta$ and (multiplicity, *J* in Hz).



8



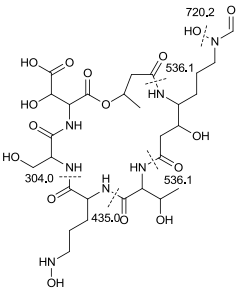

**Supplementary Note 1 Structure elucidation**

**High Resolution Mass Spectra**

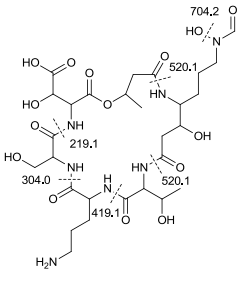
A stock solution of 20 mg/mL of each compound was diluted to a final concentration of 10 µg/mL in water with 0.1% formic acid. This solution was directly infused at a rate of ~3 µL per min into a Thermo Finnigan LTQ OrbiTrap XL mass spectrometer running Xcaliber 2.07 and TunePlus 2.4 SP1. High resolution MS was acquired using an electrospray ionization source and fragmentation was obtained through collision induced dissociation (CID). The instrument was operated in the positive mode using a maximum resolution of 100, 000. Data was acquired for approximately 1 min.

**Mass Spectra Fragmentation**

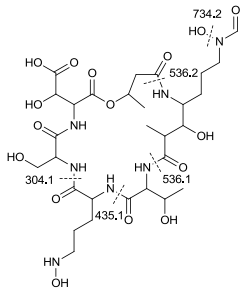
Mass spectral fragmentation patterns are shown for acidobactins, vacidobactins, and variobactin below.



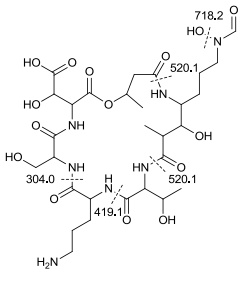

Acidobactin A (2)



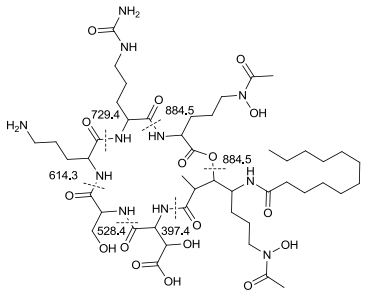

Acidobactin B (3)



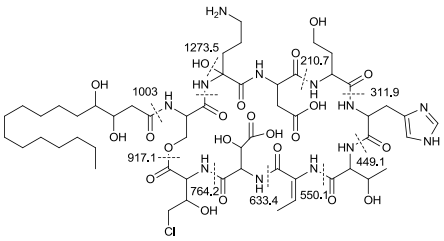

Vacidobactin A (4)




Vacidobactin B (5)




Variobactin A (6)




Thanamycin (8)

**NMR Methods and Structural Characterization**

NMR spectra were measured on a Bruker Avance 700 spectrometer equipped with a 5 mm inverse detection probe and using TMS as an internal standard. Lyophilized samples were dissolved in $D_2O$, $CDCl_3$, or $d_6$-DMSO (*Sigma Aldrich*) and spectra were recorded at 297 K. NMR experiments were processed and analyzed with Bruker TOPSPIN 2.1 or MestReNova 9.0. Chemical shifts ($\delta$) expressed in parts per million (ppm) and coupling constants ($J$) are reported in Hertz (Hz). Assembly of individual amino acids to form the final linear structure was accomplished by considering long-range $^{1}H$-$^{1}H$ NOESY and ROESY correlations, and $^{1}H$-$^{13}C$ HMBC correlations from protons adjacent carbonyl carbons, as well as by assignments of 2D $^{1}H$-$^{1}H$ COSY and 2D $^{1}H$-$^{13}C$ HSQC correlations.

# Appendix 2

## Supplementary Information

### Assembly and Clustering of Natural Antibiotics Guides Target Identification

Chad W. Johnston[1,2], Michael A. Skinnider[1,2], Chris A. Dejong[1,2], Philip N. Rees[1,2], Gregory M. Chen[1,2], Chelsea Walker[1,2], Shawn French[1], Eric D. Brown[1], Janos Berdy[3], Dennis Y. Liu[1,2] & Nathan A. Magarvey[1,2]*

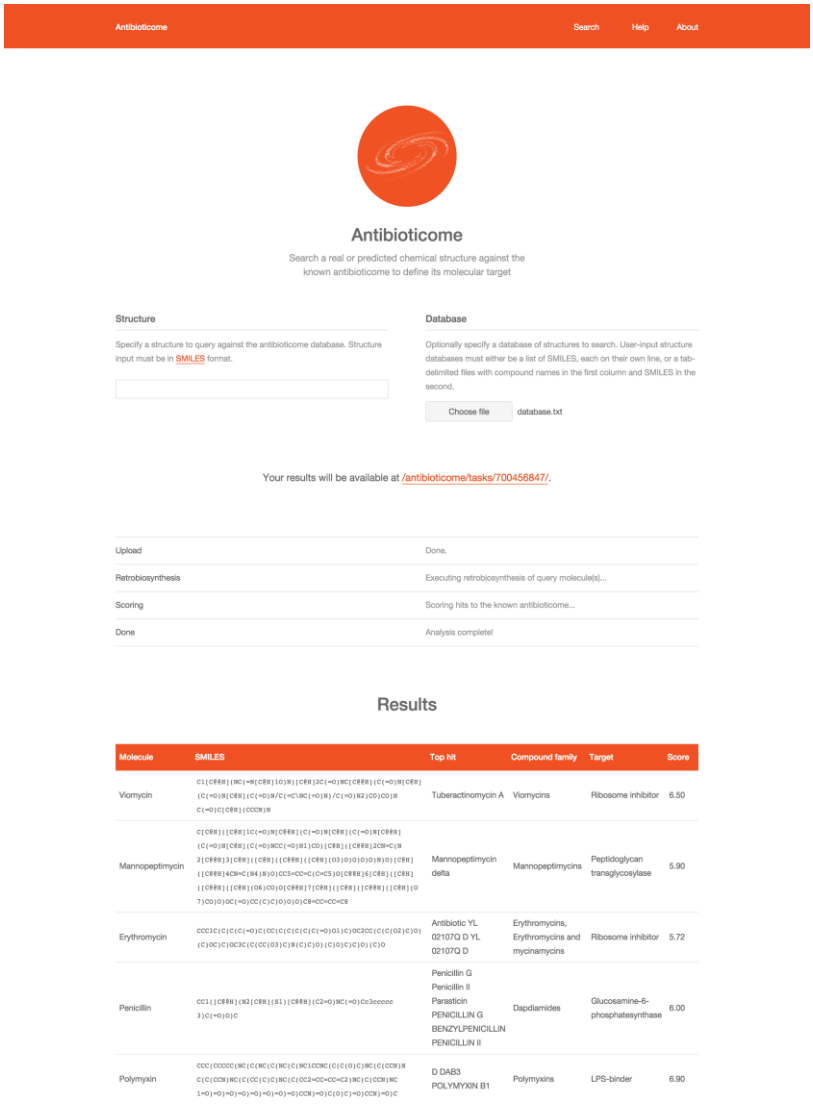
[1] Department of Biochemistry & Biomedical Sciences; M. G. DeGroote Institute for Infectious Disease Research;

[2] Department of Chemistry & Chemical Biology; McMaster University, Hamilton, Canada L8S 4K1

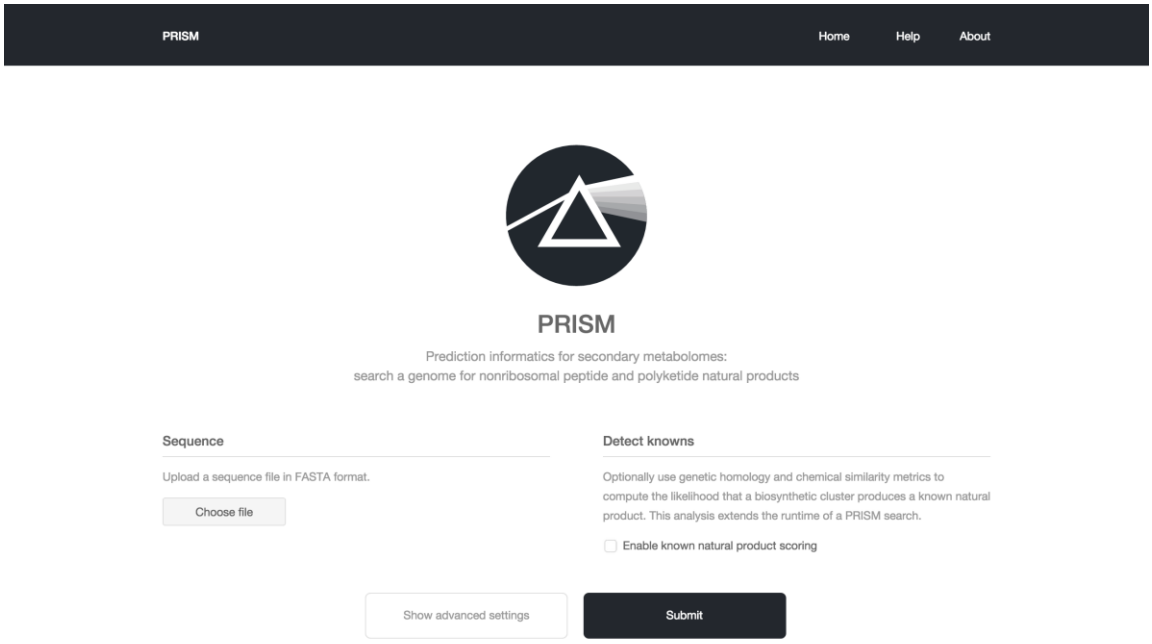
[3] Eötvös Loránd University, Budapest, Hungary

* Corresponding author, email address: magarv@mcmaster.ca

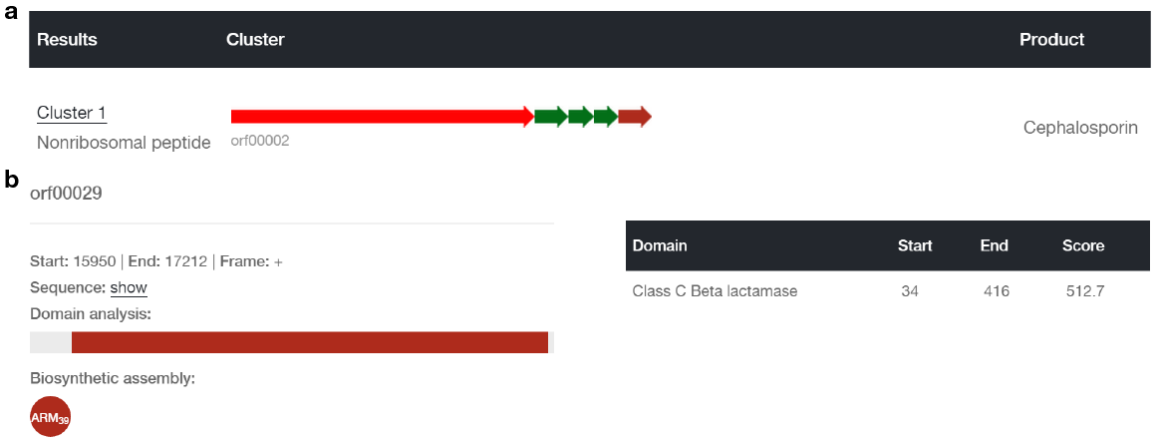
## Supplementary Results



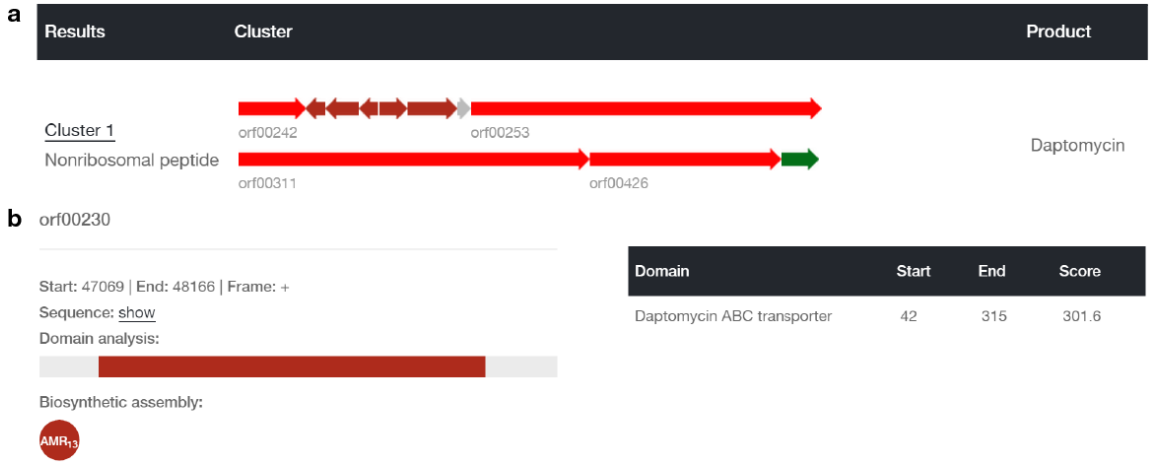

**Supplementary Figure 1. Antibioticome web application.** The antibioticome search web application (accessible at http://magarveylab.ca/antibioticome/) provides a user-friendly mechanism to query the targeted antibioticome, in order to define the putative molecular target of a user-submitted molecule. A single molecular structure or database of molecular structures are submitted in SMILES format, and undergo *in silico* retrobiosynthesis followed by hierarchical clustering. The single highest-scoring targeted antibiotic match identified by computational retrobiosynthesis is reported for each submitted molecule, as well as the family of natural products to which that targeted antibiotic belongs, the mechanism of action of that family, and the confidence score assigned to the match.



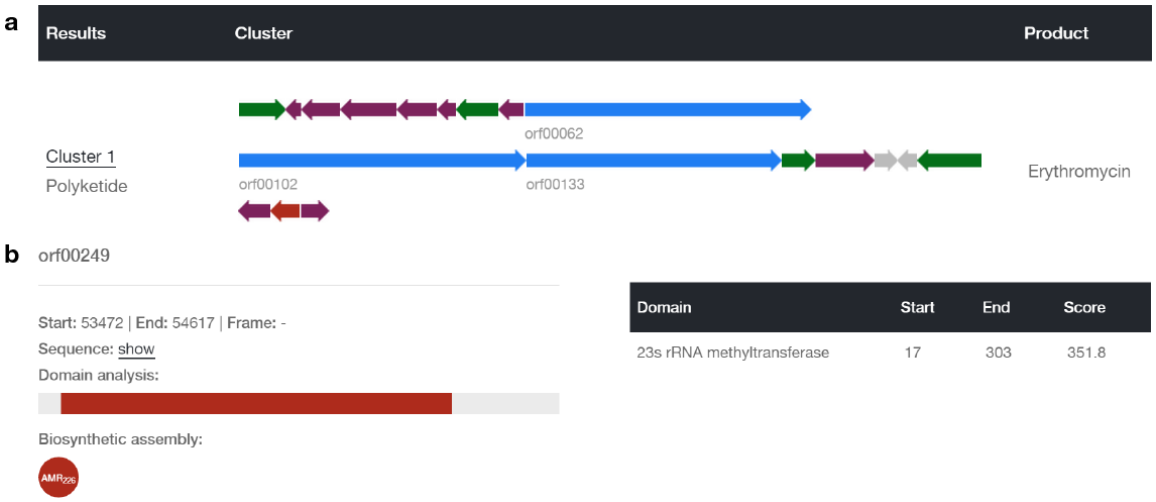

**Supplementary Figure 2. PRISM—a web application for identifying biosynthetic gene clusters.** Screenshot of the PRISM user interface.



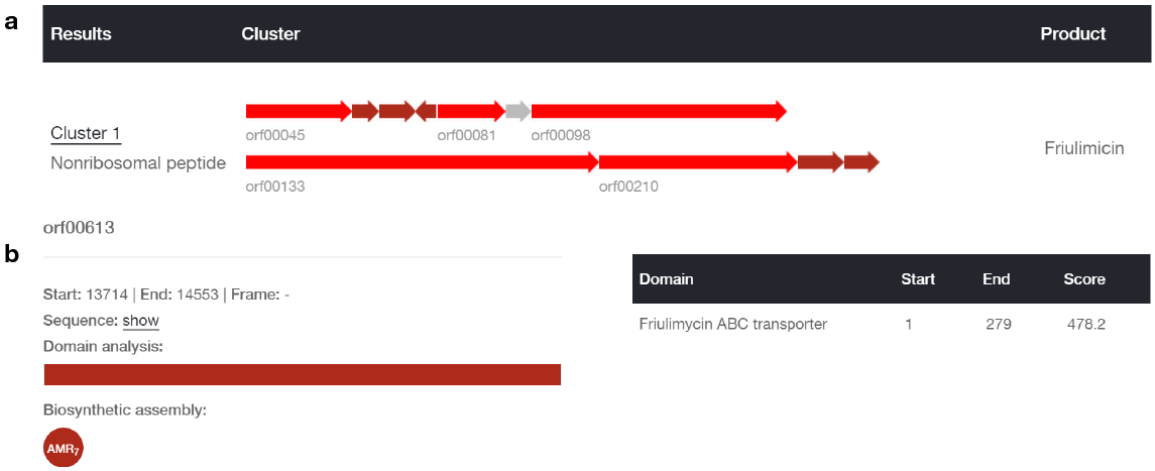

**Supplementary Figure 3. PRISM results for the cephalosporin biosynthetic gene cluster.** (a) Detection of the cephalosporin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, and green indicates tailoring enzymes. (b) Detection of an embedded resistance determinant related to the cephalosporin mode of action – Class C beta lactamase.



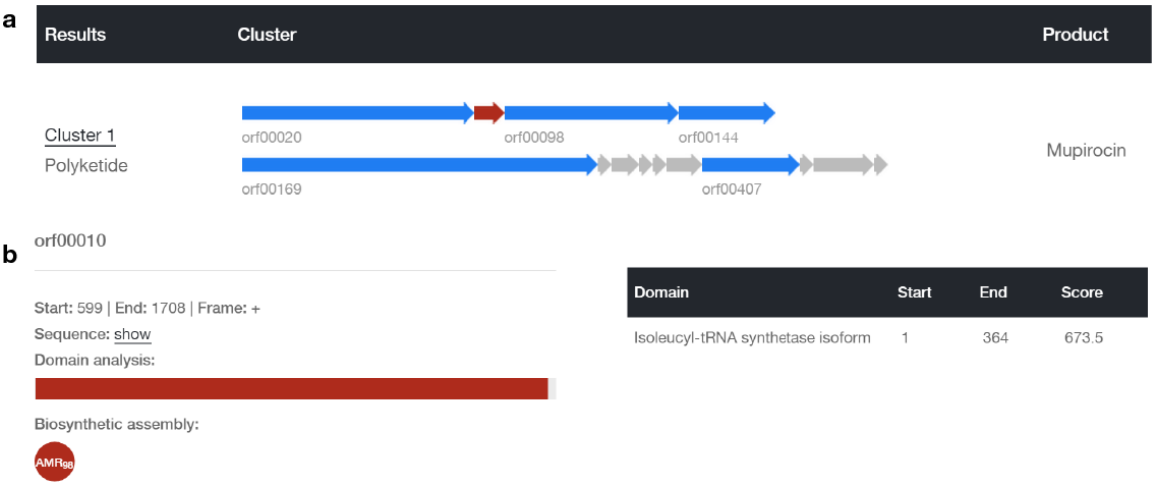

**Supplementary Figure 4. PRISM results for the daptomycin biosynthetic gene cluster.** (a) Detection of the daptomycin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, green indicates tailoring enzymes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the daptomycin chemical scaffold, a daptomycin ABC transporter.




**Supplementary Figure 5. PRISM results for the erythromycin biosynthetic gene cluster.** (a) Detection of the erythromycin biosynthetic gene cluster. Blue indicates polyketide synthase genes, brown indicates resistance genes, purple indicates sugar biosynthesis genes, green indicates tailoring enzymes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the erythromycin mode of action, the ribosome modifying rRNA methyltransferase.

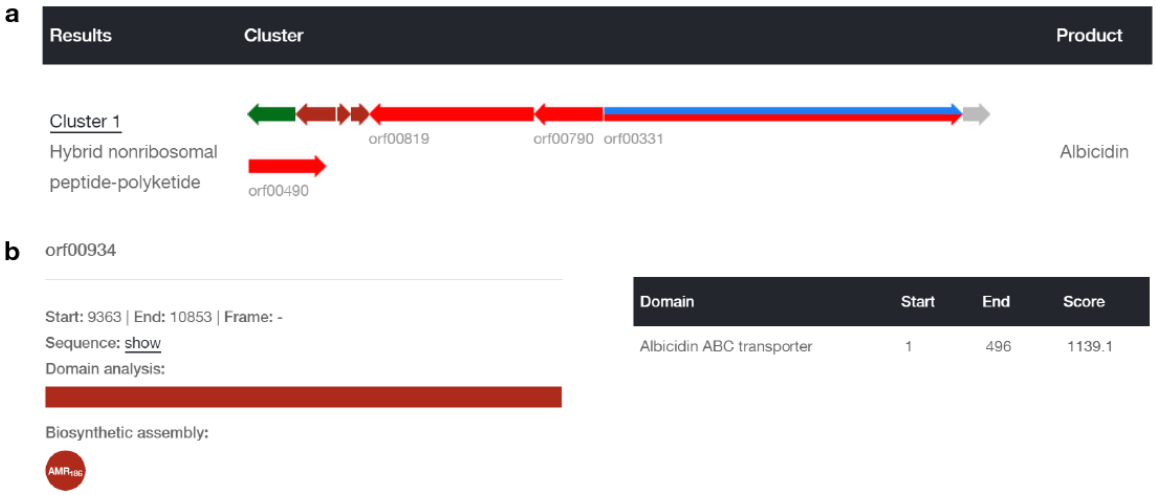



**Supplementary Figure 6. PRISM results for the friulimicin biosynthetic gene cluster.** (a) Detection of the friulimicin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the friulimicin chemical scaffold, a friulimicin ABC transporter.



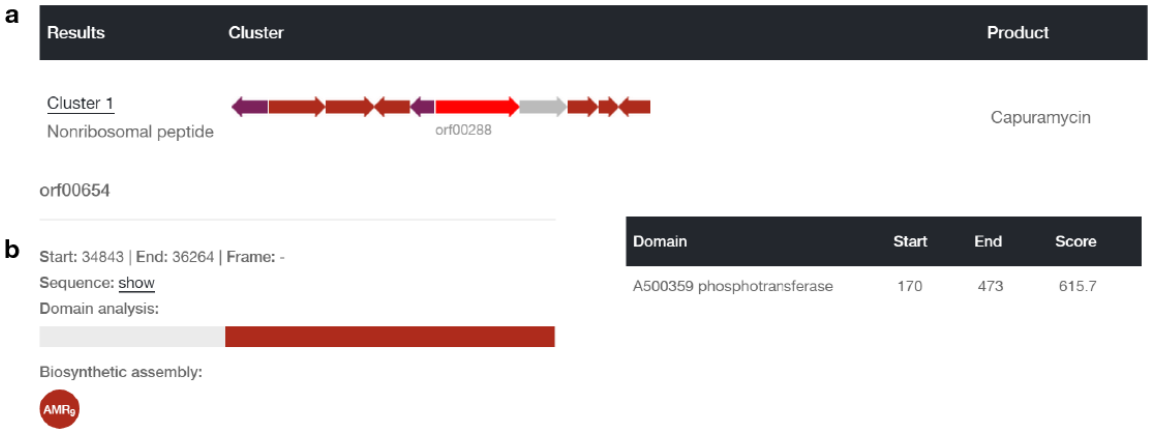

**Supplementary Figure 7. PRISM results for the mupirocin biosynthetic gene cluster.** (a) Detection of the mupirocin biosynthetic gene cluster. Blue indicates polyketide synthase genes, brown indicates resistance genes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the mupirocin mode of action, a mupirocin resistant Ile-tRNA synthetase.


**Supplementary Figure 8. PRISM results for the albomycin biosynthetic gene cluster.** (a) Detection of the albomycin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, and green indicates tailoring enzymes. (b) Detection of an embedded resistance determinant related to the albomycin mode of action, an albomycin resistant Ser-tRNA synthetase.

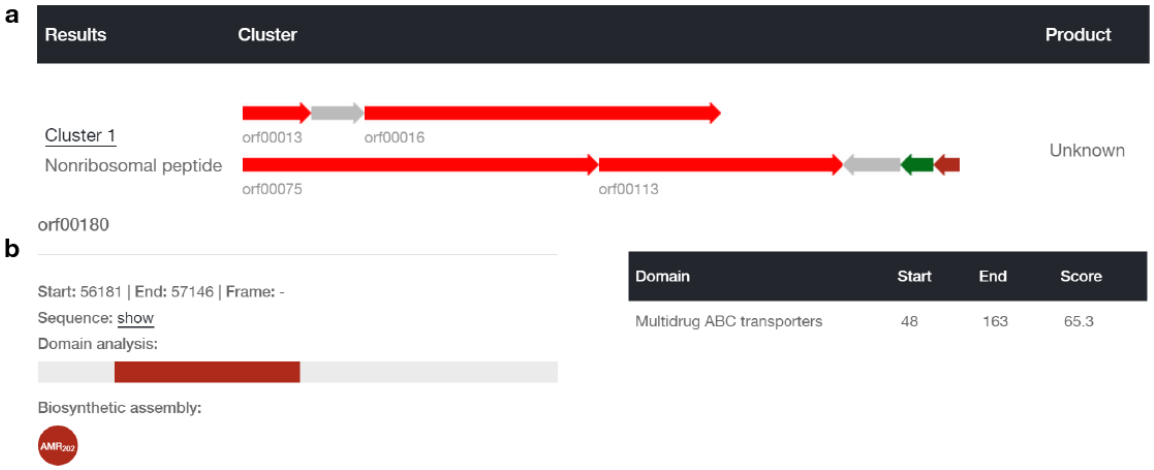



**Supplementary Figure 9. PRISM results for the albicidin biosynthetic gene cluster.** (a) Detection of the albicidin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, blue represents polyketide synthetase genes, brown indicates resistance genes, green indicates tailoring enzymes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the albicidin chemical scaffold, an albicidin ABC transporter.

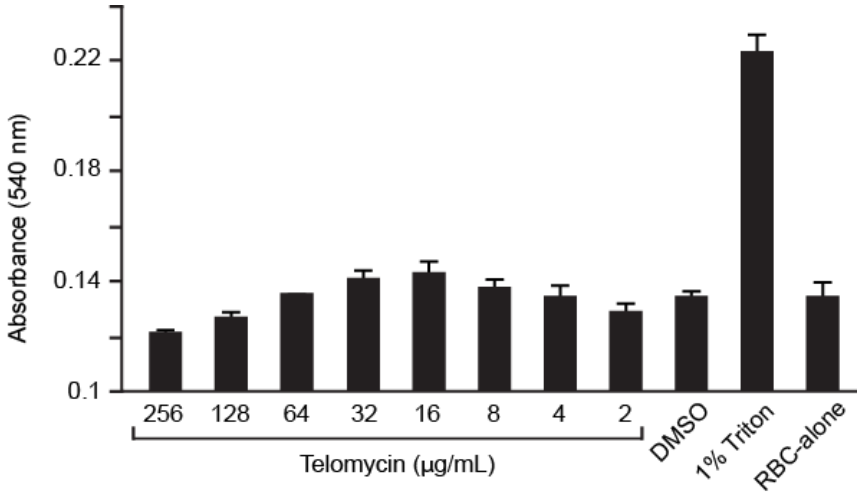

**Supplementary Figure 10. PRISM results for the teicoplanin biosynthetic gene cluster.** (a) Detection of the teicoplanin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, green indicates tailoring enzymes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant related to the teicoplanin mode of action – the D-ala-D-lac glycopeptide-resistant ligase.



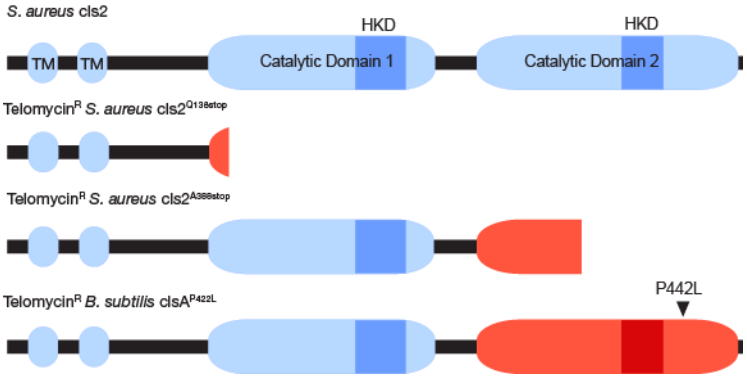

**Supplementary Figure 11. PRISM results for the capuramycin biosynthetic gene cluster.** (a) Detection of the capuramycin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, purple indicates sugar biosynthesis enzymes, and gray represents other biosynthetic enzymes. (b) Detection of an embedded resistance determinant specific to the capuramycin chemical scaffold – A500359 phosphotransferase.



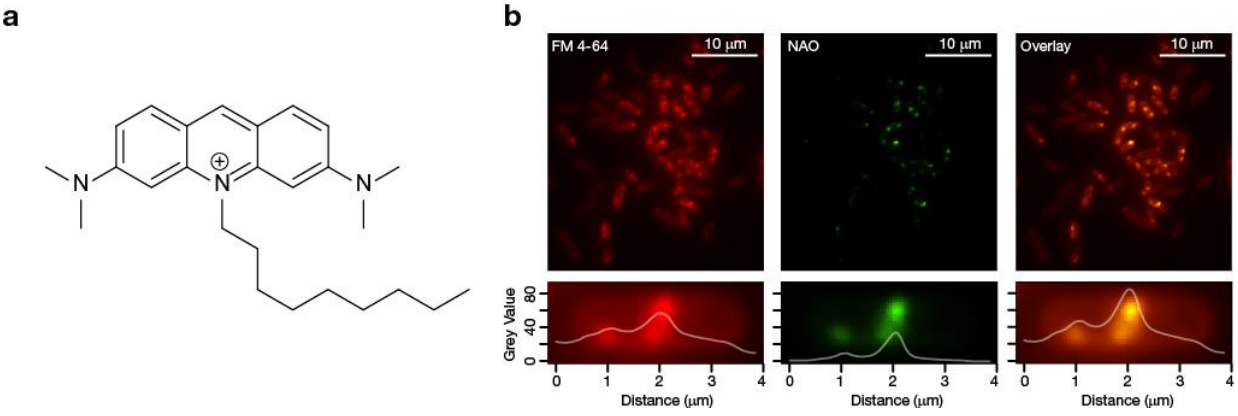

**Supplementary Figure 12. PRISM results for the telomycin biosynthetic gene cluster.** (a) Detection of the telomycin biosynthetic gene cluster. Red indicates nonribosomal peptide synthetase genes, brown indicates resistance genes, green indicates tailoring enzymes, and gray represents other biosynthetic enzymes. (b) Detection of a generic ABC transporter as the sole known resistance determinant detected in the telomycin gene cluster.




**Supplementary Figure 13. Telomycin does not cause hemolysis.** A solution of 0.25% sheep red blood cell suspension in phosphate-buffered saline was incubated for 1 h at 37°C with either a serial dilution of telomycin, 1% Triton X-100 (positive lysis control), DMSO alone (1%; negative lysis control), or without added content (negative lysis control). After 1 h RBCs were pelleted and lysis was assessed by measuring absorbance of the supernatant at 540 nm.




**Supplementary Figure 14. Summary of observed spontaneous telomycin-resistance mutations.** Sequenced genomes of telomycin-resistant *S. aureus* Newman and *B. subtilis* 168 were compared to sensitive parent strains, revealing mutations in the dominant, house-keeping cardiolipin synthase genes cls2 and clsA, respectively. Mutations occurred in the catalytic domains of cardiolipin synthase, and either resulted in truncations, or missense mutations near the active site HKD motifs.




**Supplementary Figure 15. Fluorescence microscopy of anionic-lipid dye NAO.** (**a**) Structure of the cardiolipin dye 10-*N*-nonyl-acridine orange (NAO). (**b**) Fluorescence microscopy images of *B. subtilis* with the standard membrane dye FM4-64 and the standard cardiolipin dye NAO. Summed pixel intensities over the length of a selected bacteria are depicted below.

## Supplementary Tables

| Gene | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| tlo1 | Alpha-mannosidase | - | 1049 |
| tlo2 | Hydrolase | - | 401 |
| tlo3 | ABC transporter permease | - | 271 |
| tlo4 | ABC transporter permease | - | 274 |
| tlo5 | Extracellular binding domain | - | 437 |
| tlo6 | Hypothetical protein | + | 572 |
| tlo7 | Putative glucokinase | + | 311 |
| tlo8 | LacI-family transcription regulator | + | 347 |
| tlo9 | Hypothetical protein | + | 434 |
| tlo10 | Hydrolase | + | 567 |
| tlo11 | Transcriptional regulator | + | 279 |
| tlo12 | Tryptophan dehydrogenase | - | 314 |
| tlo13 | Antitermination regulator | - | 313 |
| tlo14 | GntR-family transcription regulator | - | 477 |
| tlo15 | Anthranilate phosphoribosyltransferase | + | 352 |
| tlo16 | Phospholipase | + | 161 |
| tlo17 | HTH regulatory protein | + | 322 |
| tlo18 | Fatty acyl adenylate ligase | + | 583 |
| tlo19 | Acyl carrier protein | + | 90 |
| tlo20 | NRPS | + | 4769 |
| tlo21 | NRPS | + | 5830 |
| tlo22 | NRPS | + | 2400 |
| tlo23 | Cytochrome P450 | + | 416 |
| tlo24 | MbtH protein | + | 75 |
| tlo25 | Acylase | + | 825 |
| - | Transposase | - | 140 |
| - | Transposase | - | 128 |
| tlo26 | α/β Hydrolase | + | 269 |
| tlo27 | Aminotransferase-methyltransferase | - | 706 |
| tlo28 | Ferritin-like | - | 917 |
| tlo29 | Cytochrome P450 | - | 257 |
| tlo30 | ABC transporter membrane protein | - | 273 |
| tlo31 | ABC transporter ATPase | - | 315 |
| tlo32 | Proline hydroxylase | + | 287 |
| tlo33 | Hypothetical protein | + | 296 |
| tlo34 | Hypothetical protein | + | 669 |

**Supplementary Table 1.** Telomycin gene cluster analysis.

| Bacterial strain | Telomycin MIC | % Cardiolipin |
|---|---|---|
| *Staphylococcus aureus* Newman | 8 μg/mL | 100% |
| *S. aureus* Newman ClsA$^{Q148stop}$ | 128 μg/mL | 1.7% |
| *S. aureus* Newman ClsA$^{A388stop}$ | 128 μg/mL | 3.8% |
| *Bacillus subtilis* 168 | 1 μg/mL | 100% |
| *B. subtilis* 168 ClsA$^{P442L}$ | 16 μg/mL | 10.2% |

**Supplementary Table 2.** Cardiolipin levels present in spontaneously telomycin-resistant mutant strains.

| [NaCl] (M) | *S. aureus* wild type | *S. aureus* Telo$^R$ |
|---|---|---|
| 0 | 8 | 128 |
| 0.5 | 4 | 128 |
| 1 | 2 | 64 |
| 1.5 | 1 | 16 |

**Supplementary Table 3.** Increasing osmotic stress increases telomycin potency. MICs of sensitive and telomycin-resistant (cls2 A388stop) *S. aureus* cultured in CAMHB containing 0, 0.5, 1, or 1.5 M NaCl. MICs were determined after 16 h growth, shown in μg ml$^{-1}$.

| | *B. subtilis* 168 wildtype | *B. subtilis* 168 Telo$^R$ | *S. aureus* Newman wildtype | *S. aureus* Newman Telo$^R$ |
|---|---|---|---|---|
| (1) | 1 | 16 | 8 | 128 |
| (2) | 1 | 8 | 4 | 32 |
| (3) | 1 | 8 | 4 | 32 |
| (4) | 2 | 16 | 16 | 64 |
| (5) | 4 | 32 | 32 | 128 |
| (6) | 0.5 | 4 | 2 | 8 |
| (7) | 8 | 64 | 32 | >128 |
| (8) | 0.5 | 2 | 4 | 16 |
| di-5-hydroxytryptophan telomycin (9) | >128 | >128 | >128 | >128 |
| di-5-methoxytryptophan telomycin (10) | 8 | 128 | 32 | >128 |

**Supplementary Table 4.** MICs of telomycin antibiotics (1-10). Sensitive and resistant strains were exposed to telomycins and MICs were measured by microdilution in CAMHB. MICs were determined after 16 h growth, and are shown in μg ml$^{-1}$.

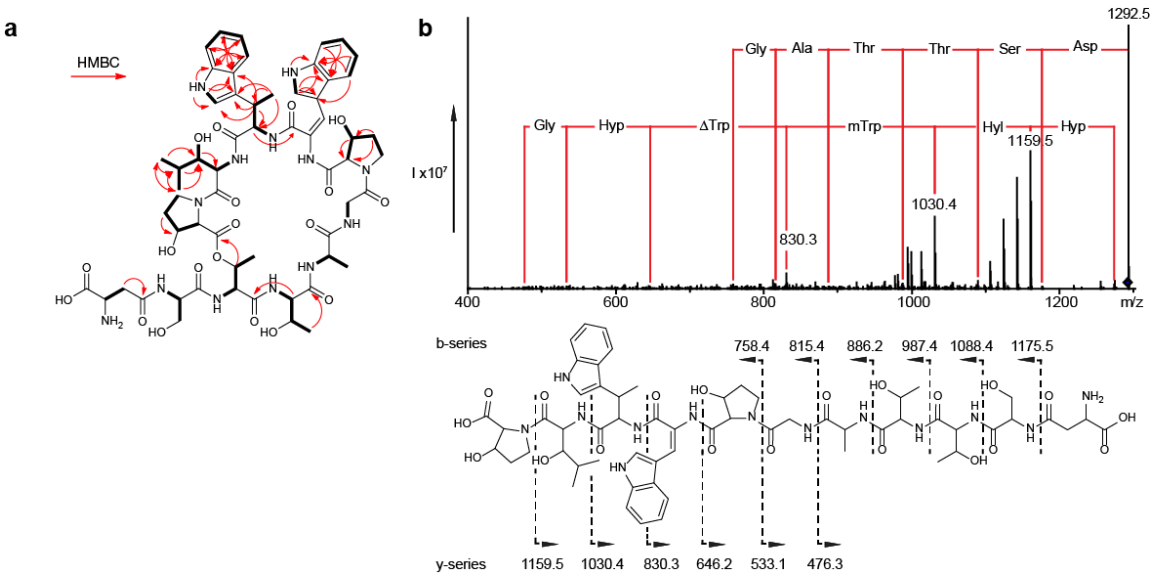
## Supplementary Datasets

**Supplementary Dataset 1. Results from retrobiosynthetic analysis of microbial modular natural product antibacterials.** A retrobiosynthetic algorithm was used to cluster specific antibacterial natural products by their chemical scaffolds. Clustering analysis yielded proposed classes based on conserved chemical scaffold, with the number of structures in each class is listed along with a representative chemical structure and statement on the mechanism of action or known cross resistance. Molecules which are not present in this dataset were not accepted as inputs by our retrobiosynthetic algorithm or were not specific antibacterial agents.

**Supplementary Dataset 2. Hidden Markov models used by PRISM to detect antibiotic resistance genes.**

## Supplementary Note – Structural Characterization

### Telomycin A (1)




**Structure confirmation of telomycin A (1).** (**a**) Structure of telomycin A confirmed by NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin A, including observed b- and y-ions resulting from amide bond cleavage.

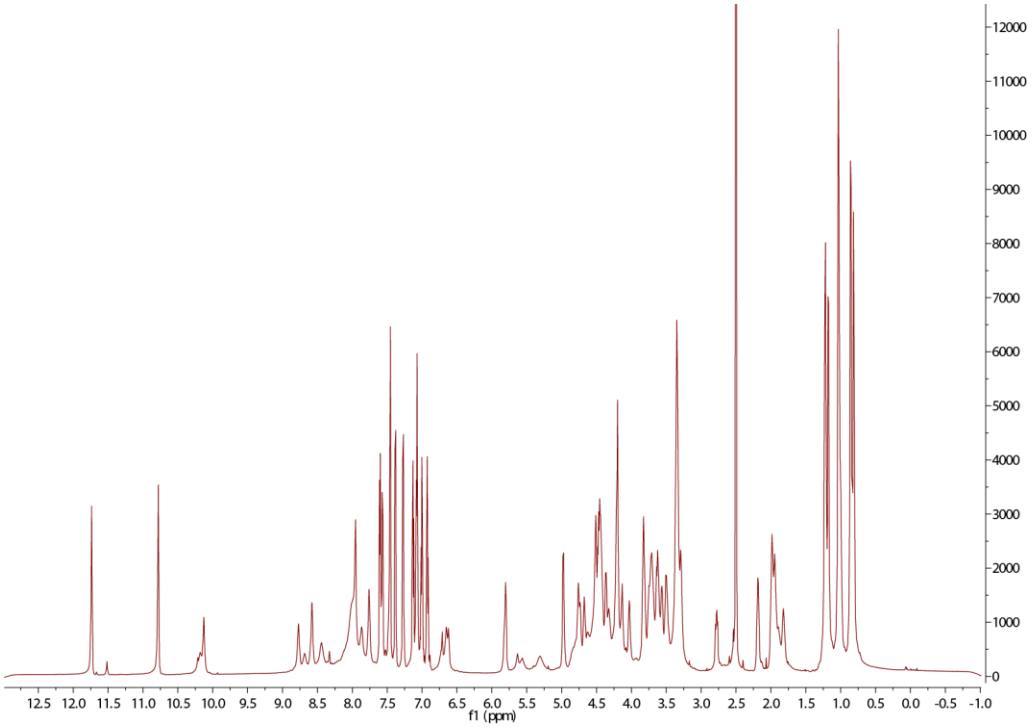
### High resolution mass data for telomycin A (1)

| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin A (1) | $C_{59}H_{78}N_{13}O_{19}$ [M+H] | 1272.55314 | 1272.55285 | 0.662 |

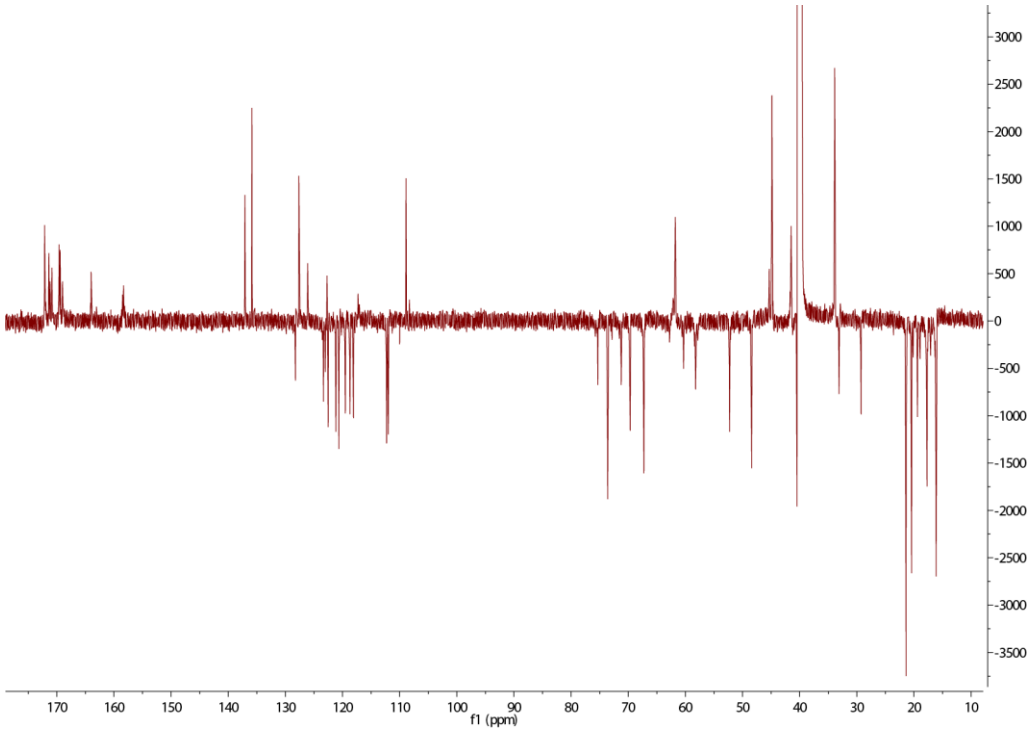
NMR spectroscopic data for telomycin A (1) (700 MHz, DMSO-$d_6$).

| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | - | - |
| 2 | - | - | 38 | - | 108.60 |
| 3 | 4.04 (*m*) | 50.37 | 39 | 7.96 (*br.*) | 128.42 |
| 4a | 2.78 (*ov.*) | 36.70 | 40 | 11.73 (*s*) | - |
| 4b | 2.54 (*ov.*) | | 41 | - | 135.60 |
| 5 | - | 170.82 | 42 | 7.4 (*d*, 7) | 112.40 |
| 6 | 8.57 (*br.*) | - | 43 | 7.14 (*t*, 7) | 122.70 |
| 7 | 4.33 (*dd*, 7, 7) | 60.29 | 44 | 7.08 (*t*, 7) | 120.80 |
| 8a | 3.72 (*d*, 7) | 61.74 | 45 | 7.61 (*d*, 7) | 118.30 |
| 8b | 3.57 (*d*, 7) | | 46 | - | 127.60 |
| 9 | - | - | 47 | - | 172.12 |
| 10 | - | - | 48 | 7.77 (*br.*) | - |
| 11 | 8.46 (*br.*) | - | 49 | 4.48 (*ov.*) | 59.85 |
| 12 | 4.38 (*d*, 7) | 58.10 | 50 | 3.63 (*m*) | 32.90 |
| 13 | 4.99 (*m*, 7) | 71.30 | 51 | 1.22 (*d*, 7) | 19.16 |
| 14 | 1.17 (*d*, 7) | 16.00 | 52 | - | 117.00 |
| 15 | - | 169.40 | 53 | 7.08 (*s*) | 123.40 |
| 16 | 7.46 (*br.*) | - | 54 | 10.78 (*ov.*) | - |
| 17 | 4.14 (*t*, 7) | 58.17 | 55 | - | 136.80 |
| 18 | 3.76 (*ov.*) | 67.31 | 56 | 7.28 (*d*, 7) | 112.15 |
| 19 | 1.03 (*d*, 7) | 21.31 | 57 | 7.00 (*t*, 7) | 121.40 |
| 20 | - | - | 58 | 6.94 (*t*, 7) | 118.90 |
| 21 | - | 172.03 | 59 | 7.57 (*d*, 7) | 119.70 |
| 22 | 7.87 (*br.*) | - | 60 | - | 125.90 |
| 23 | 4.45 (*ov.*) | 48.32 | 61 | - | - |
| 24 | 1.05 (*d*, 7) | 17.56 | 62 | 6.68 (*br.*) | - |
| 25 | - | - | 63 | 4.51 (*d*, 7) | 52.24 |
| 26 | 8.77 (*br.*) | - | 64a | 3.3 (*m*) | 75.42 |
| 27a | 4.43 (*m*) | 41.48 | 64b | - | |
| 27b | 3.83 (*m*) | | 65 | 4.76 (*br.*) | - |
| 28 | - | - | 66 | 1.96 (*m*) | 29.24 |
| 29 | 4.22 (*d*, 7) | 73.60 | 67 | 0.86 (*d*, 7) | 20.20 |
| 30a | 4.20 (*m*) | 67.26 | 68 | 0.81 (*d*, 7) | 16.90 |
| 30b | - | | 69 | - | - |
| 31 | 5.81 (*m*) | - | 70 | 4.75 (*m*) | 62.74 |
| 32a | 2.00 (*m*) | 33.75 | 71a | 4.69 (*m*) | 69.65 |
| 32b | 2.18 (*m*) | | 71b | - | |
| 33a | 3.83 (*d*, 7) | 44.86 | 72 | - | - |
| 33b | 3.51 (*d*, 7) | | 73a | 1.83 (*m*) | 33.09 |
| 34 | - | - | 73b | 1.99 (*m*) | |
| 35 | - | - | 74a | 3.71 (*m*) | 45.35 |
| 36 | - | - | 74b | - | |
| 37a | - | - | 75 | - | 168.99 |

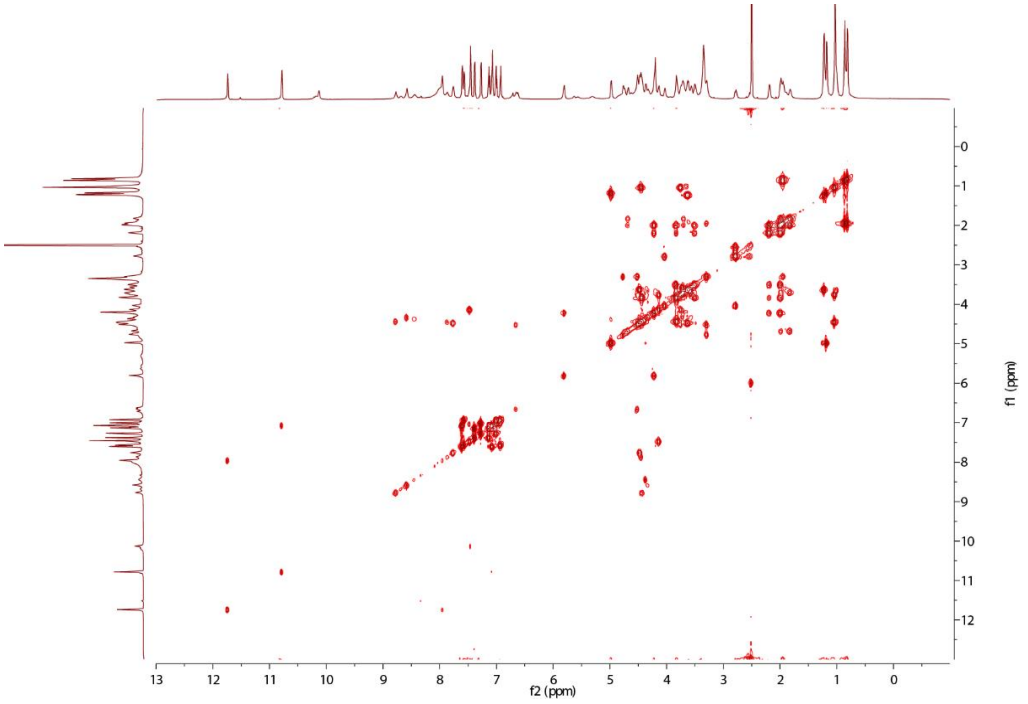
$^{1}$H NMR spectrum of telomycin A (1) in DMSO.



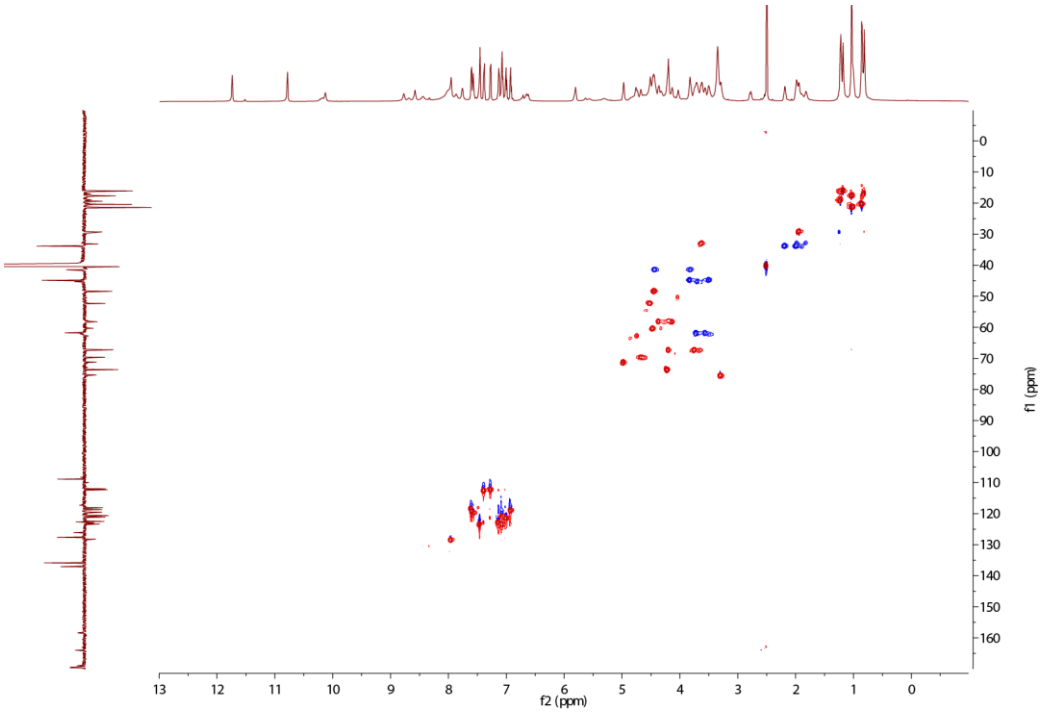

$^{13}$C NMR spectrum of telomycin A (1) in DMSO.



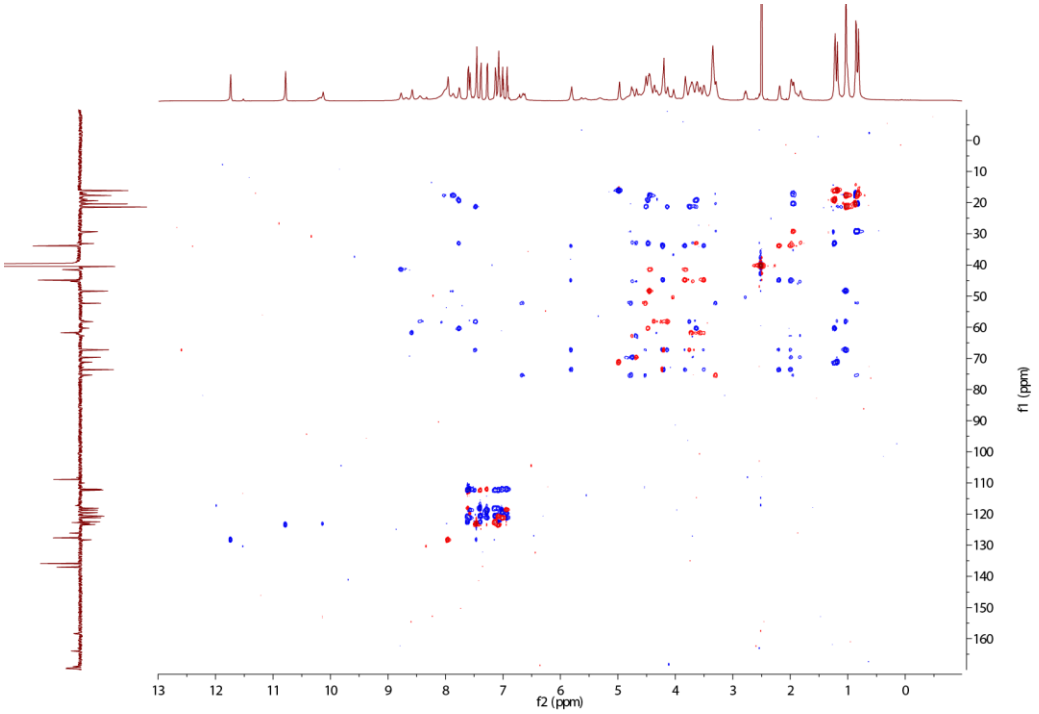
$^{1}H$-$^{1}H$ COSY spectrum of telomycin A (1) in DMSO.



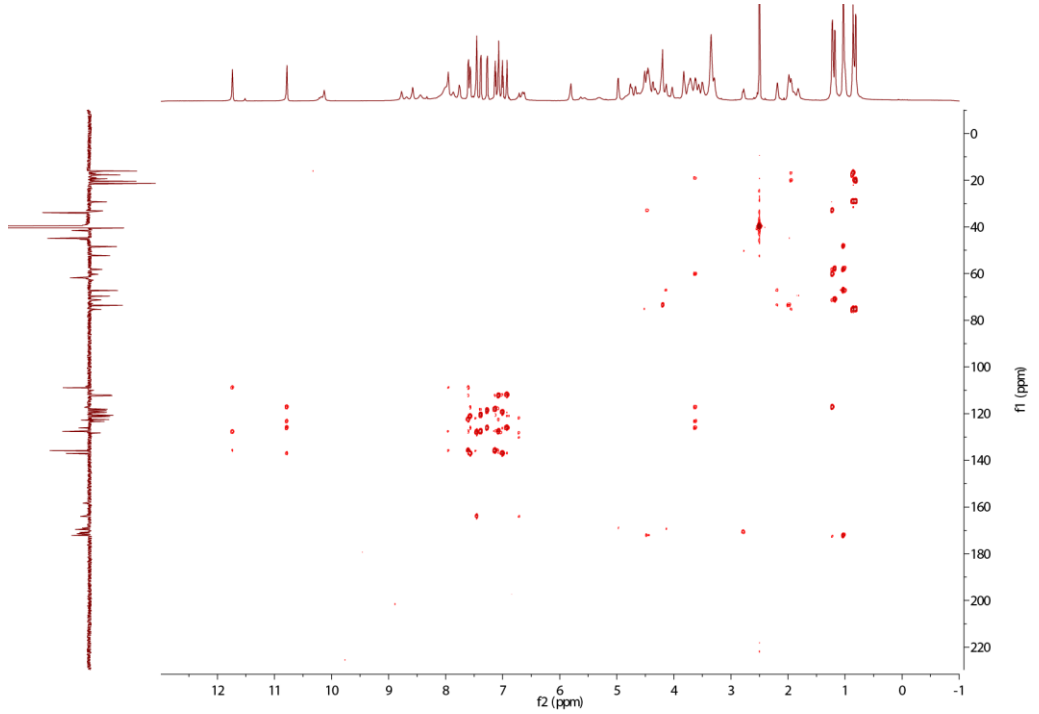

$^{1}H$-$^{13}C$ HSQC spectrum of telomycin A (1) in DMSO.



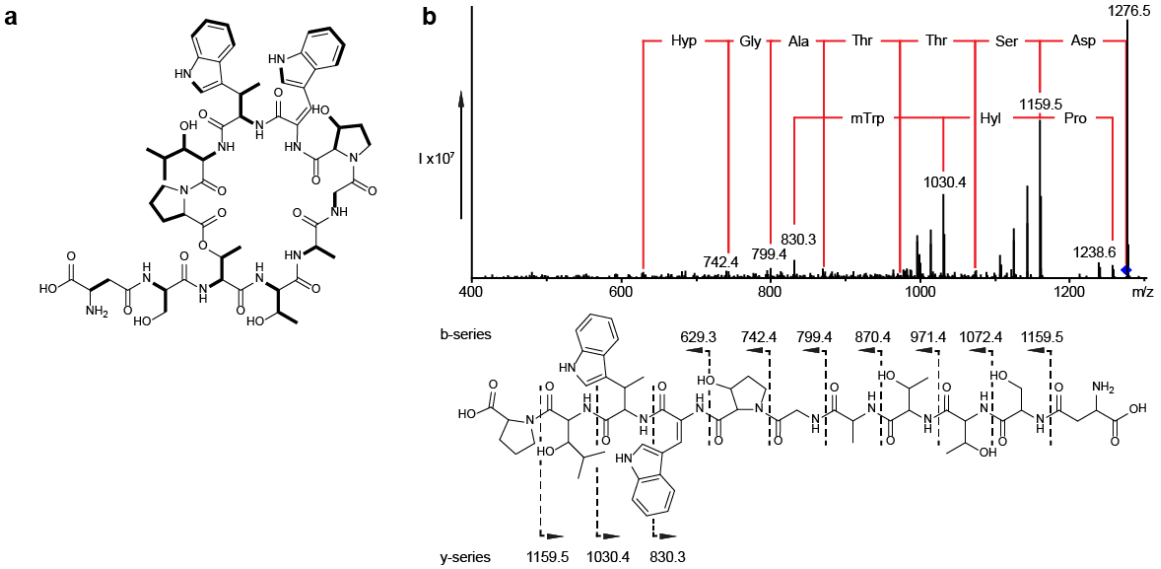
$^{1}H$-$^{13}C$ HSQC-TOCSY spectrum of telomycin A (1) in DMSO.




$^{1}H$-$^{13}C$ HMBC spectrum of telomycin A (1) in DMSO.



## Telomycin B (2)




**Structure elucidation of telomycin B (2).** (**a**) Structure of telomycin B elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin B, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin B (2)

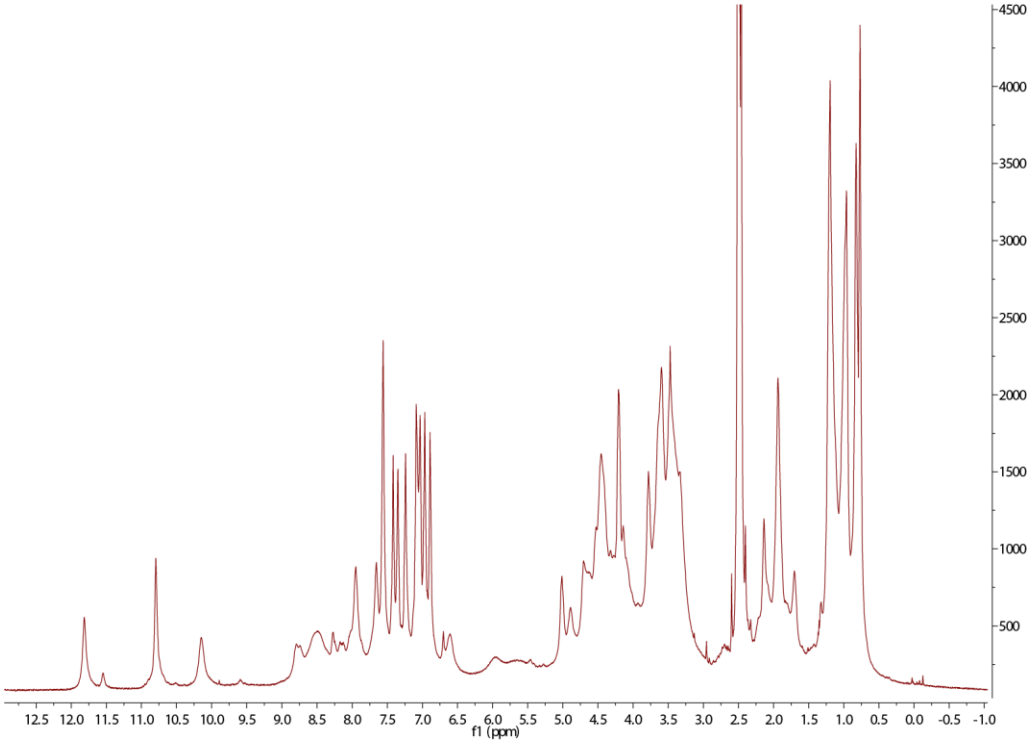
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin B (2) | $C_{59}H_{78}N_{13}O_{18}$ [M+H] | 1256.55823 | 1256.55857 | 0.165 |

NMR spectroscopic data for telomycin B (2) (700 MHz, DMSO-$d_6$).

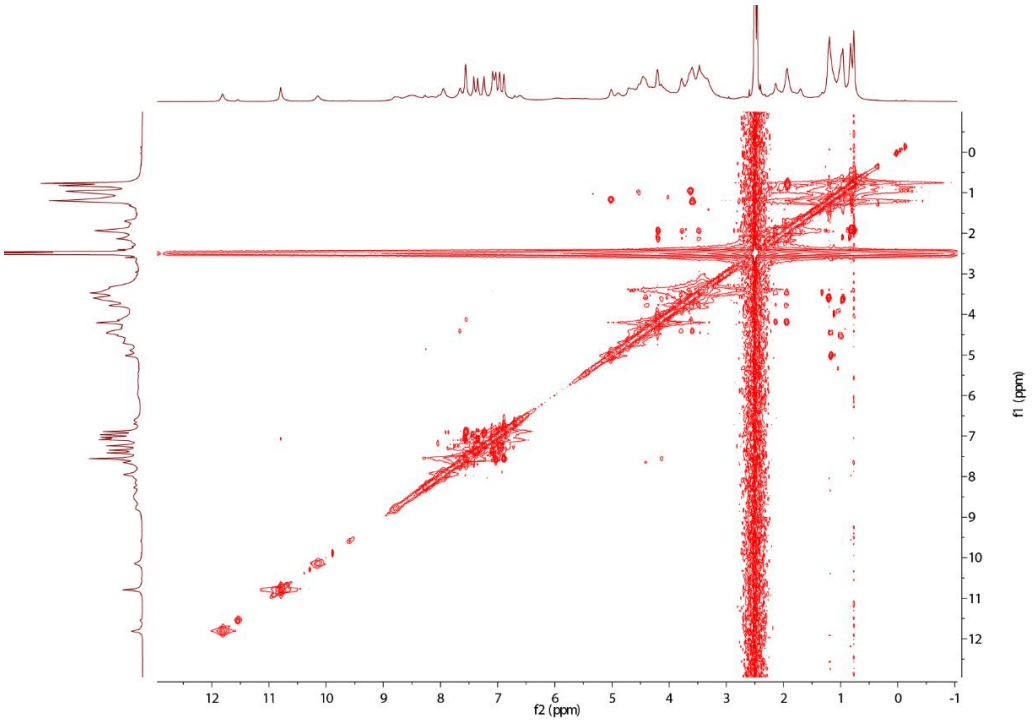
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | - | - |
| 2 | - | - | 38 | - | - |
| 3 | 4.22 (*ov.*) | 56.74 | 39 | 7.94 (*ov.*) | 127.27 |
| 4a | 2.98 (*br.*) | 39.63 | 40 | 11.81 (*s*) | - |
| 4b | 2.83 (*br.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.35 (*m*) | 111.28 |
| 6 | 8.13 (*br.*) | - | 43 | 7.08 (*m*) | 121.51 |
| 7 | 4.61 (*ov.*) | 53.46 | 44 | 7.03 (*m*) | 119.63 |
| 8a | 3.65 (*ov.*) | 60.97 | 45 | 7.56 (*m*) | 117.08 |
| 8b | 3.43 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.65 (*br.*) | - |
| 11 | 8.43 (*br.*) | - | 49 | 4.42 (*ov.*) | 59.17 |
| 12 | 4.28 (*ov.*) | 56.69 | 50 | 3.59 (*ov.*) | 31.89 |
| 13 | 5.01 (*m*) | 70.28 | 51 | 1.21 (*m*) | 17.52 |
| 14 | 1.16 (*m*) | 14.9 | 52 | - | - |
| 15 | - | - | 53 | 7.06 (*s*) | 122.35 |
| 16 | 7.55 (*br.*) | - | 54 | 10.79 (*s*) | - |
| 17 | 4.14 (*ov.*) | 56.86 | 55 | - | - |
| 18 | 3.63 (*ov.*) | 66.06 | 56 | 7.24 (*m*) | 110.97 |
| 19 | 0.96 (*m*) | 19.79 | 57 | 6.96 (*m*) | 120.16 |
| 20 | - | - | 58 | 6.89 (*m*) | 117.66 |
| 21 | - | - | 59 | 7.55 (*m*) | 118.58 |
| 22 | 7.96 (*br.*) | - | 60 | - | - |
| 23 | 4.53 (*ov.*) | 47.07 | 61 | - | - |
| 24 | 0.99 (*m*) | 16.43 | 62 | 6.83 (*br.*) | - |
| 25 | - | - | 63 | 4.46 (*ov.*) | 51.59 |
| 26 | 8.74 (*br.*) | - | 64a | 3.31 (*ov.*) | 73.86 |
| 27a | 4.4 (*ov.*) | 40.08 | 64b | - | |
| 27b | 3.79 (*ov.*) | | 65 | 4.7 (*ov.*) | - |
| 28 | - | - | 66 | 1.91 (*ov.*) | 27.68 |
| 29 | 4.19 (*ov.*) | 72.32 | 67 | 0.83 (*m*) | 19.26 |
| 30a | 4.2 (*ov.*) | 66.15 | 68 | 0.77 (*m*) | 14.98 |
| 30b | - | | 69 | - | - |
| 31 | 5.27 (*br.*) | - | 70 | 4.88 (*ov.*) | 56.95 |
| 32a | 2.13 (*ov.*) | 32.61 | 71a | 2.2 (*ov.*) | 27.89 |
| 32b | 1.94 (*ov.*) | | 71b | 1.98 (*ov.*) | |
| 33a | 3.77 (*ov.*) | 43.56 | 72 | - | - |
| 33b | 3.48 (*ov.*) | | 73a | 1.9 (*ov.*) | 22.84 |
| 34 | - | - | 73b | 1.7 (*ov.*) | |
| 35 | - | - | 74a | 4.05 (*ov.*) | 45.73 |
| 36 | - | - | 74b | 3.58 (*ov.*) | |
| 37a | - | - | 75 | - | - |

*Carbon signals were taken from HSQC spectra

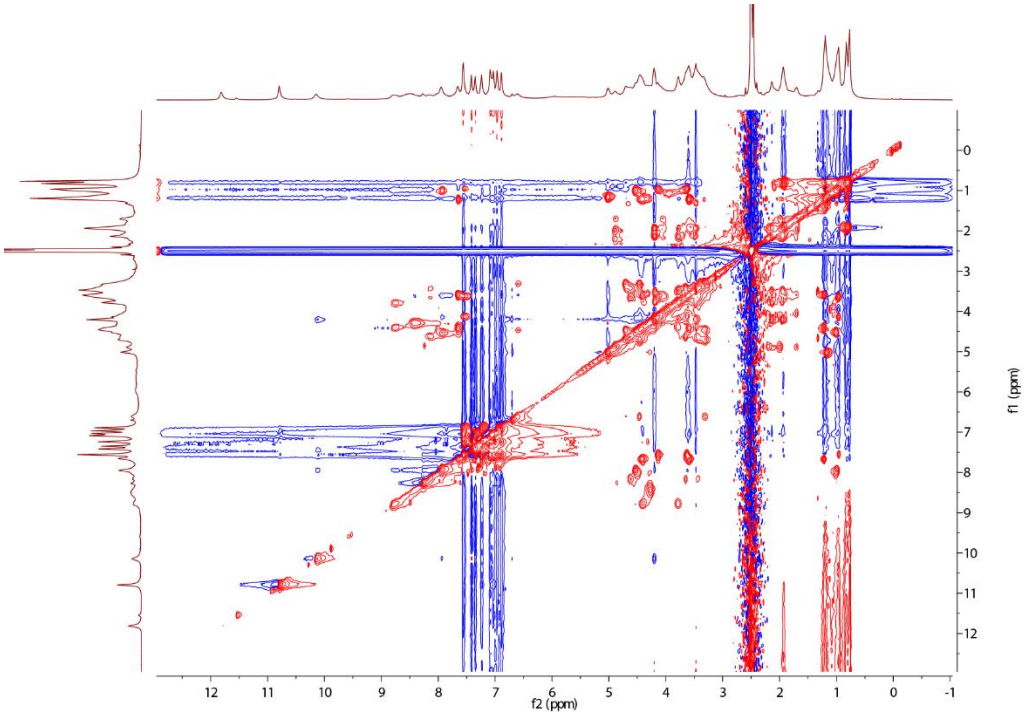
$^{1}$H NMR spectrum of telomycin B (2) in DMSO.



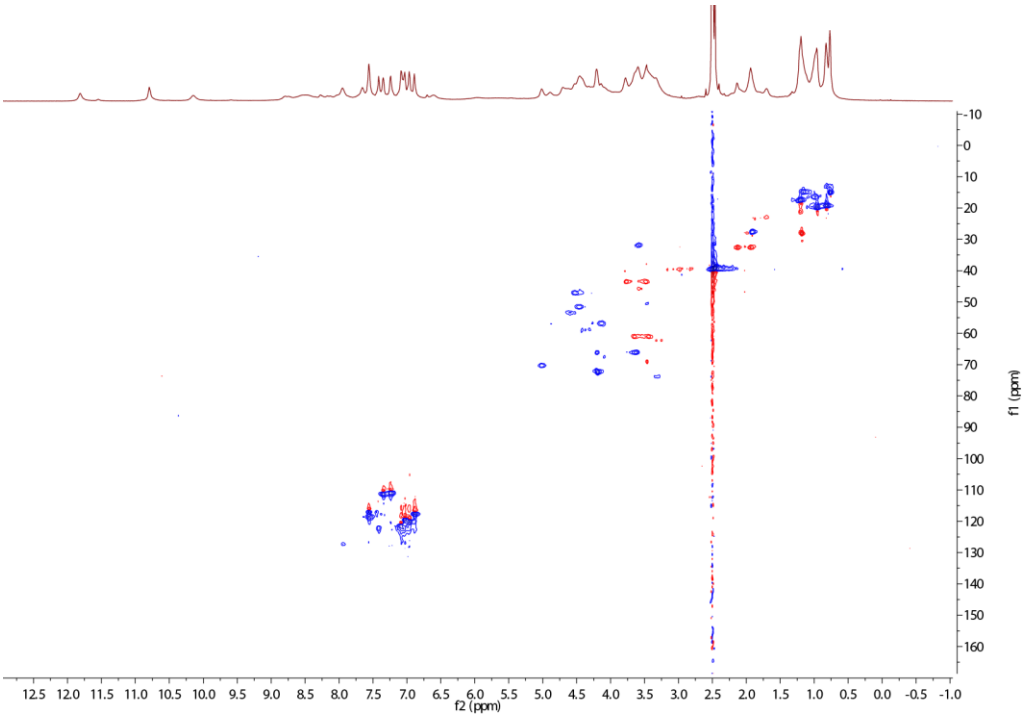

$^{1}$H-$^{1}$H COSY spectrum of telomycin B (2) in DMSO.



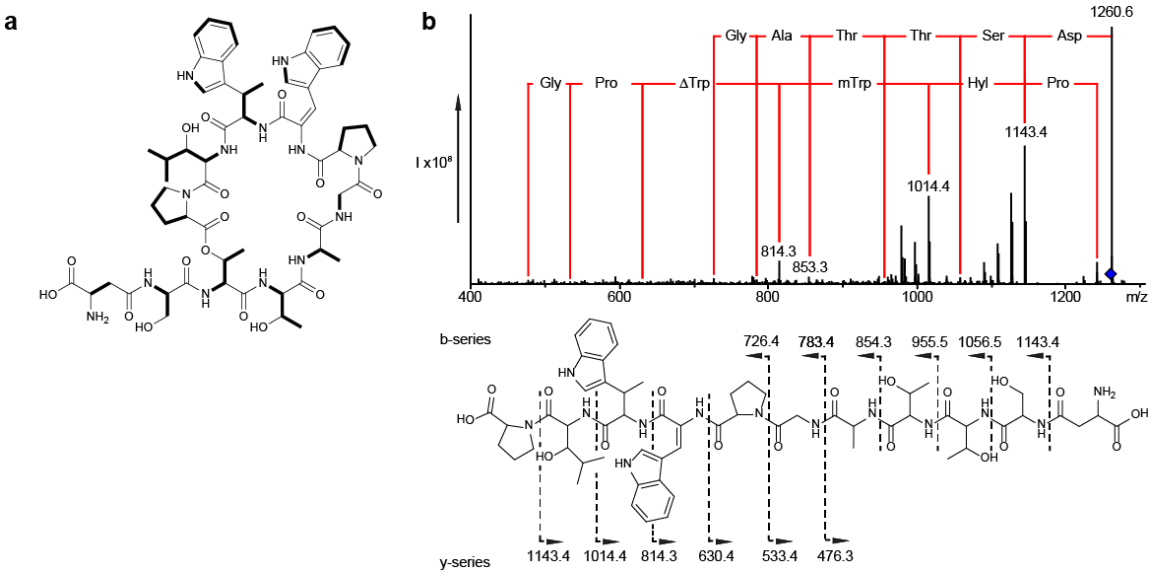
$^{1}H$-$^{1}H$ TOCSY spectrum of telomycin B (2) in DMSO.




$^{1}H$-$^{13}C$ HSQC spectrum of telomycin B (2) in DMSO.



## Telomycin C (3)




**Structure elucidation of telomycin C (3).** (**a**) Structure of telomycin C elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin C, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin C (3)

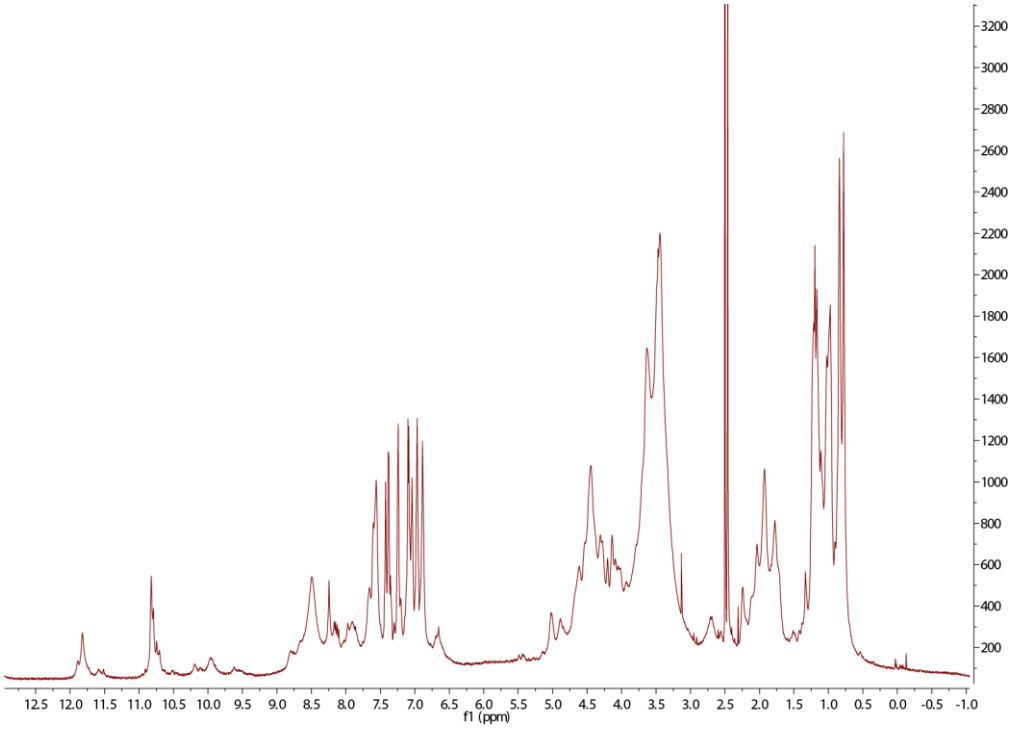
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin C (3) | $C_{59}H_{78}N_{13}O_{17}$ [M+H] | 1240.56331 | 1240.56227 | 1.284 |

NMR spectroscopic data for telomycin C (3) (700 MHz, DMSO-$d_6$).

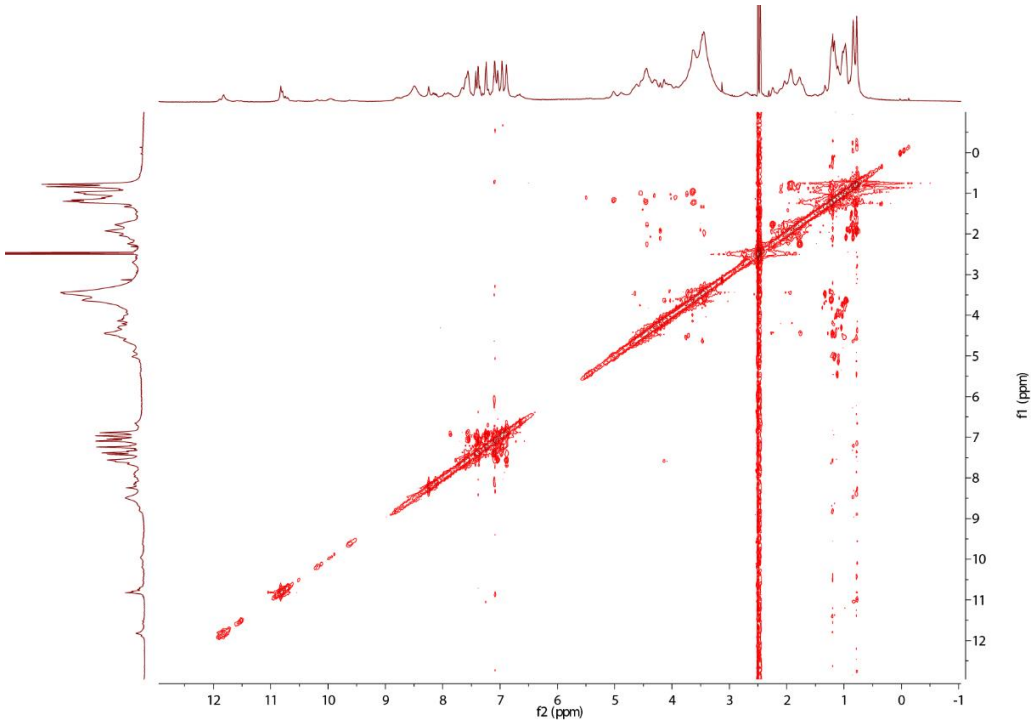
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | - | - |
| 2 | - | - | 38 | - | - |
| 3 | 4.19 (*ov.*) | 52.71 | 39 | 7.58 (*ov.*) | 126.41 |
| 4a | 2.70 (*br.*) | 36.80 | 40 | 11.82 (*s*) | - |
| 4b | 2.69 (*br.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.37 (*d*, 8) | 111.40 |
| 6 | 8.21 (*ov.*) | - | 43 | 7.10 (*t*, 7) | 121.60 |
| 7 | 4.61 (*ov.*) | 57.1 | 44 | 7.04 (*t*, 7) | 119.61 |
| 8a | 3.64 (*ov.*) | 61.00 | 45 | 7.55 (*ov.*) | 117.20 |
| 8b | 3.44 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.65 (*br.*) | - |
| 11 | 8.50 (*br.*) | - | 49 | 4.40 (*ov.*) | 58.71 |
| 12 | 4.29 (*ov.*) | 57.1 | 50 | 3.62 (*ov.*) | 31.71 |
| 13 | 5.01 (*br.*) | 70.01 | 51 | 1.23 (*ov.*) | 17.11 |
| 14 | 1.16 (*br.*) | 15.03 | 52 | - | - |
| 15 | - | - | 53 | 7.08 (*ov.*) | 122.40 |
| 16 | 7.58 (*ov.*) | - | 54 | 10.82 (*s*) | - |
| 17 | 4.13 (*ov.*) | 57.12 | 55 | - | - |
| 18 | 3.63 (*ov.*) | 67.81 | 56 | 7.24 (*d*, 7) | 110.90 |
| 19 | 0.96 (*ov.*) | 19.83 | 57 | 6.96 (*t*, 7) | 120.10 |
| 20 | - | - | 58 | 6.88 (*t*, 7) | 117.60 |
| 21 | - | - | 59 | 7.58 (*ov.*) | 118.40 |
| 22 | 7.91 (*br.*) | - | 60 | - | - |
| 23 | 4.53 (*ov.*) | 47.31 | 61 | - | - |
| 24 | 1.00 (*ov.*) | 19.22 | 62 | 8.19 (*br.*) | - |
| 25 | - | - | 63 | 4.63 (*ov.*) | 53.37 |
| 26 | 8.80 (*br.*) | - | 64a | 3.43 (*ov.*) | 74.91 |
| 27a | 4.40 (*ov.*) | 40.44 | 64b | - | |
| 27b | 3.79 (*ov.*) | | 65 | - | - |
| 28 | - | - | 66 | 1.92 (*ov.*) | 27.84 |
| 29 | 4.44 (*ov.*) | 51.60 | 67 | 0.84 (*br.*) | 19.23 |
| 30a | 2.24 (*br.*) | 27.81 | 68 | 0.78 (*br.*) | 15.11 |
| 30b | 1.77 (*ov.*) | | 69 | - | - |
| 31 | - | - | 70 | 4.20 (*ov.*) | 57.50 |
| 32a | 2.04 (*br.*) | 24.44 | 71a | 2.12 (*ov.*) | 24.14 |
| 32b | 1.94 (*ov.*) | | 71b | 1.92 (*br.*) | |
| 33a | 3.69 (*ov.*) | 45.60 | 72 | - | - |
| 33b | 3.44 (*ov.*) | | 73a | 1.98 (*ov.*) | 23.34 |
| 34 | - | - | 73b | 1.88 (*ov.*) | |
| 35 | - | - | 74a | 3.73 (*ov.*) | 45.31 |
| 36 | - | - | 74b | 3.49 (*ov.*) | |
| 37a | - | - | 75 | - | - |

*Carbon signals were taken from HSQC spectra

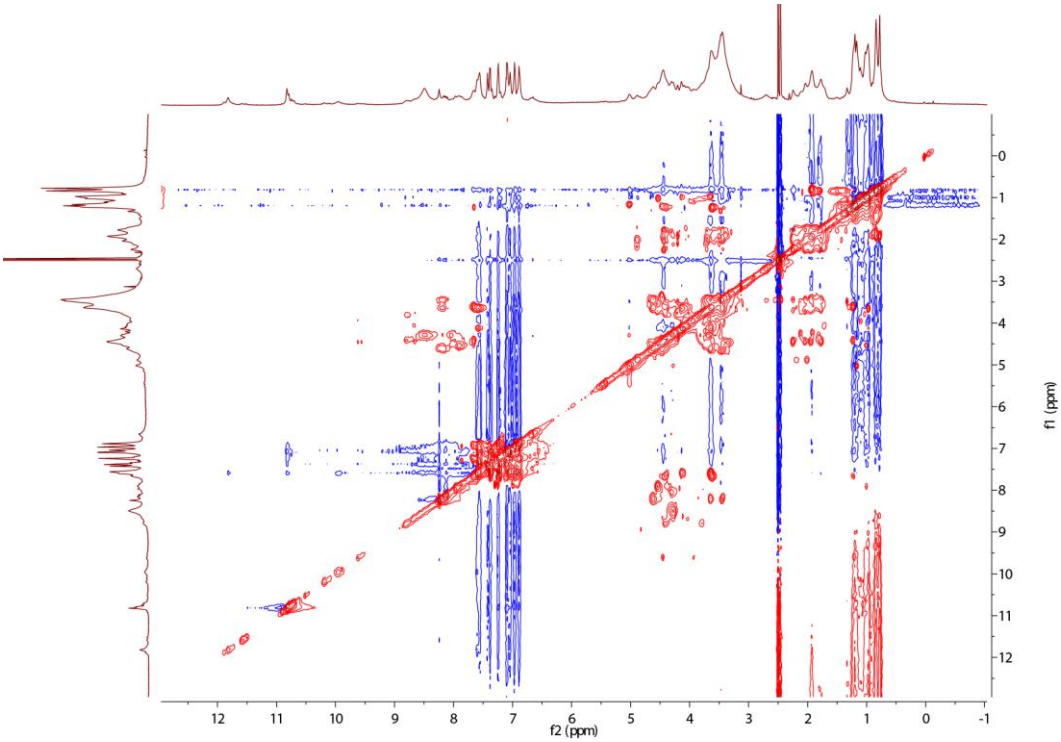
$^{1}$H NMR spectrum of telomycin C (3) in DMSO.



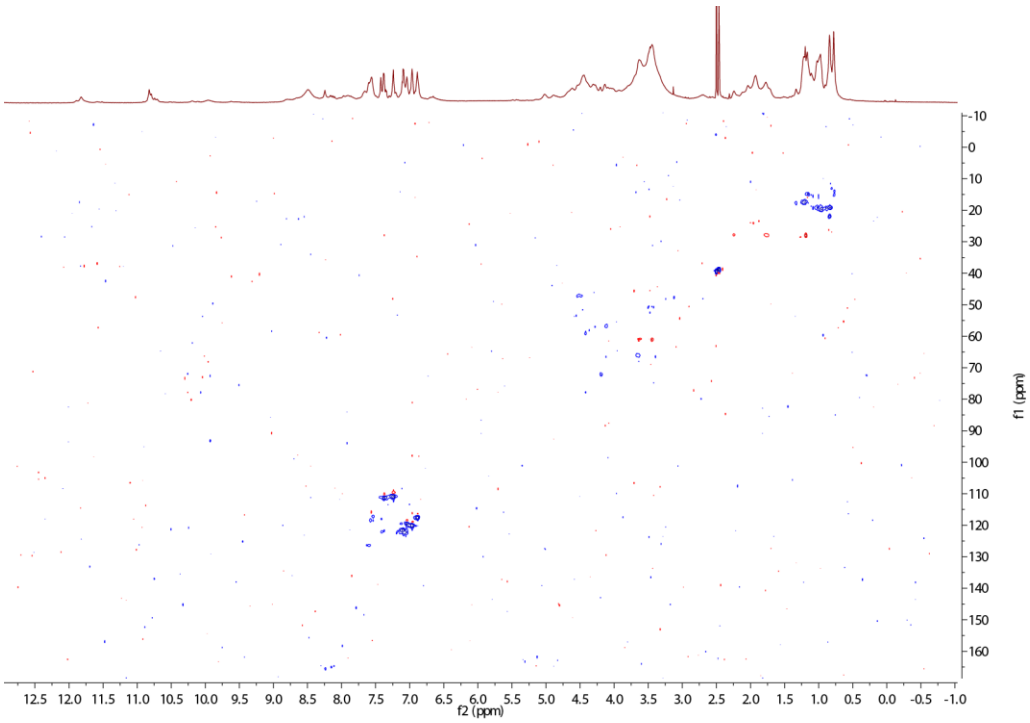

$^{1}$H-$^{1}$H COSY spectrum of telomycin C (3) in DMSO.



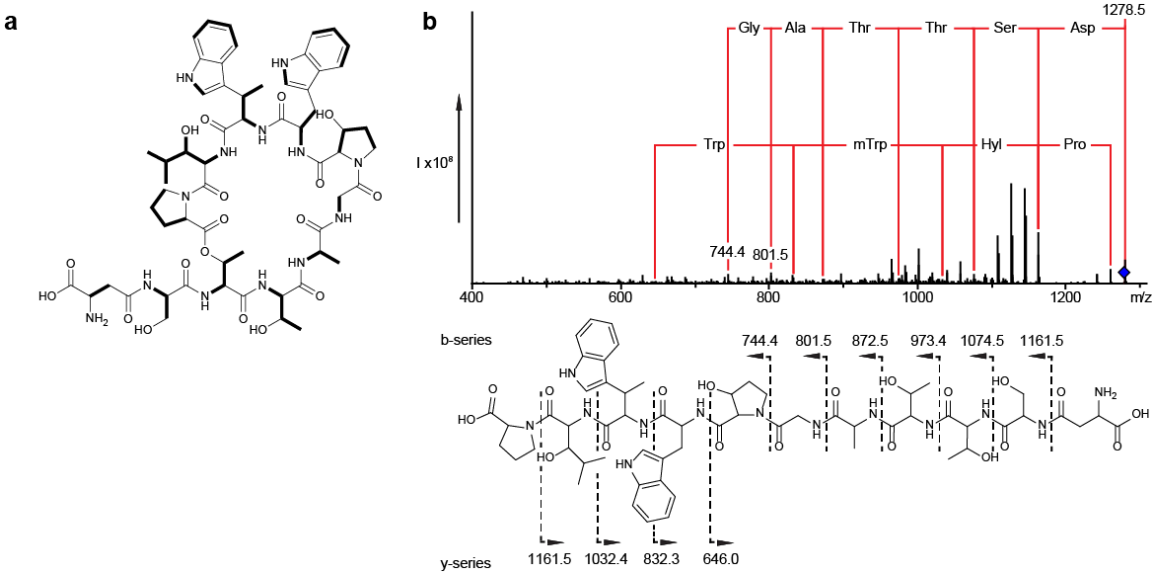
$^{1}H$-$^{1}H$ TOCSY spectrum of telomycin C (3) in DMSO.




$^{1}H$-$^{13}C$ HSQC spectrum of telomycin C (3) in DMSO.



## Telomycin D (4)




**Structure elucidation of telomycin D (4).** (**a**) Structure of telomycin D elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin D, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin D (4)

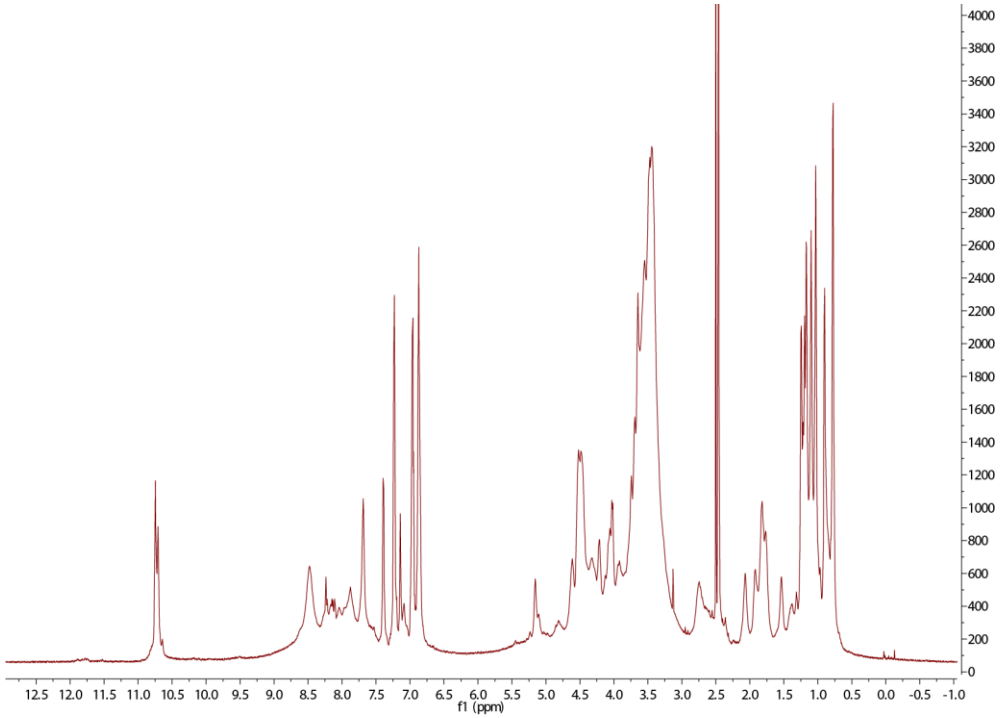
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin D (4) | $C_{59}H_{80}N_{13}O_{18}$ [M+H] | 1258.57388 | 1258.57210 | 1.85 |

NMR spectroscopic data for telomycin D (4) (700 MHz, DMSO-$d_6$).

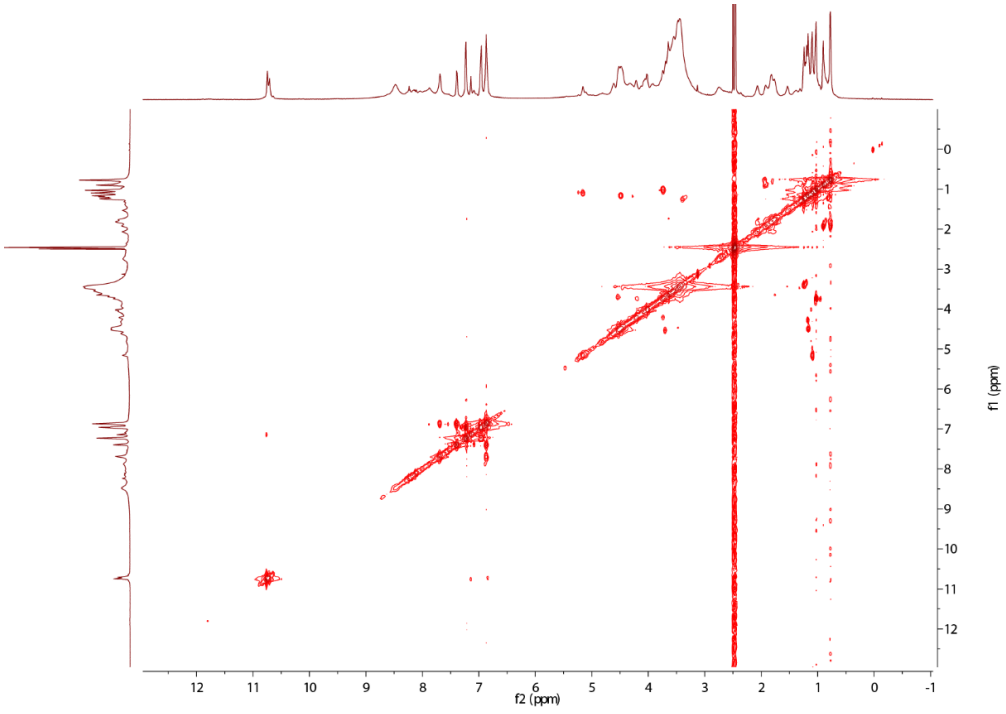
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | 2.43 (*br.*) | 36.82 |
| 2 | - | - | 38 | - | - |
| 3 | 4.28 (*ov.*) | 47.81 | 39 | 6.85 (*ov.*) | 126.40 |
| 4a | 2.73 (*br.*) | 36.92 | 40 | 10.71 (*s*) | - |
| 4b | 2.70 (*br.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.23 (*m.*) | 110.71 |
| 6 | 8.21 (*br.*) | - | 43 | 6.96 (*t*, 7) | 120.22 |
| 7 | 4.46 (*ov.*) | 53.80 | 44 | 6.87 (*m.*) | 117.60 |
| 8a | 3.56 (*ov.*) | 60.90 | 45 | 7.39 (*d*, 7) | 117.91 |
| 8b | 3.47 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.76 (*br.*) | - |
| 11 | 8.40 (*br.*) | - | 49 | 4.61 (*ov.*) | 57.23 |
| 12 | 4.52 (*ov.*) | 51.94 | 50 | 3.40 (*ov.*) | 32.59 |
| 13 | 5.15 (*br.*) | 70.41 | 51 | 1.25 (*d*, 6) | 17.31 |
| 14 | 1.09 (*m.*) | 15.00 | 52 | - | - |
| 15 | - | - | 53 | 7.14 (*s*) | 122.69 |
| 16 | 7.9 (*br.*) | - | 54 | 10.75 (*s*) | - |
| 17 | 4.18 (*ov.*) | 58.34 | 55 | - | - |
| 18 | 3.74 (*ov.*) | 65.93 | 56 | 7.23 (*m.*) | 110.60 |
| 19 | 1.02 (*m.*) | 19.51 | 57 | 6.96 (*t*, 7) | 120.11 |
| 20 | - | - | 58 | 6.87 (*m.*) | 117.42 |
| 21 | - | - | 59 | 7.70 (*m.*) | 119.02 |
| 22 | 8.18 (*br.*) | - | 60 | - | - |
| 23 | 4.48 (*ov.*) | 47.30 | 61 | - | - |
| 24 | 1.16 (*m.*) | 17.41 | 62 | 7.66 (*br.*) | - |
| 25 | - | - | 63 | 4.54 (*ov.*) | 51.79 |
| 26 | 7.85 (*br.*) | - | 64a | 3.7 (*ov.*) | 73.62 |
| 27a | 4.05 (*ov.*) | 40.91 | 64b | - | |
| 27b | 3.95 (*ov.*) | | 65 | 5.28 (*br.*) | - |
| 28 | - | - | 66 | 1.92 (*br.*) | 27.40 |
| 29 | 4.64 (*br.*) | 51.40 | 67 | 0.90 (*d*, 5) | 19.31 |
| 30a | 4.01 (*ov.*) | 69.10 | 68 | 0.77 (*d*, 5) | 13.79 |
| 30b | - | | 69 | - | - |
| 31 | - | - | 70 | 4.45 (*ov.*) | 57.61 |
| 32a | 2.07 (*br.*) | 28.20 | 71a | 2.07 (*br.*) | 27.93 |
| 32b | 1.75 (*br.*) | | 71b | 1.82 (*br.*) | |
| 33a | 3.62 (*ov.*) | 45.73 | 72 | - | - |
| 33b | 3.63 (*ov.*) | | 73a | 1.82 (*br.*) | 23.34 |
| 34 | - | - | 73b | 1.75 (*br.*) | |
| 35 | 8.25 (*br.*) | - | 74a | 3.55 (*ov.*) | 43.10 |
| 36 | 3.46 (*ov.*) | 50.51 | 74b | 3.43 (*ov.*) | |
| 37a | 2.68 (*br.*) | 36.82 | 75 | - | - |

*Carbon signals were taken from HSQC spectra

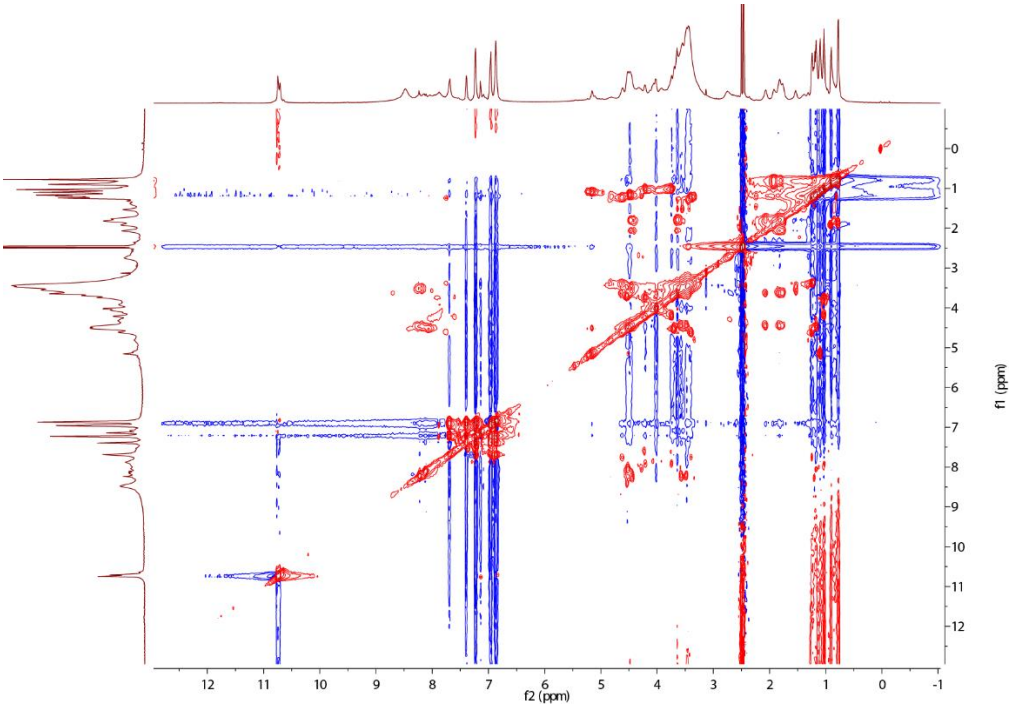
$^{1}$H NMR spectrum of telomycin D (4) in DMSO.



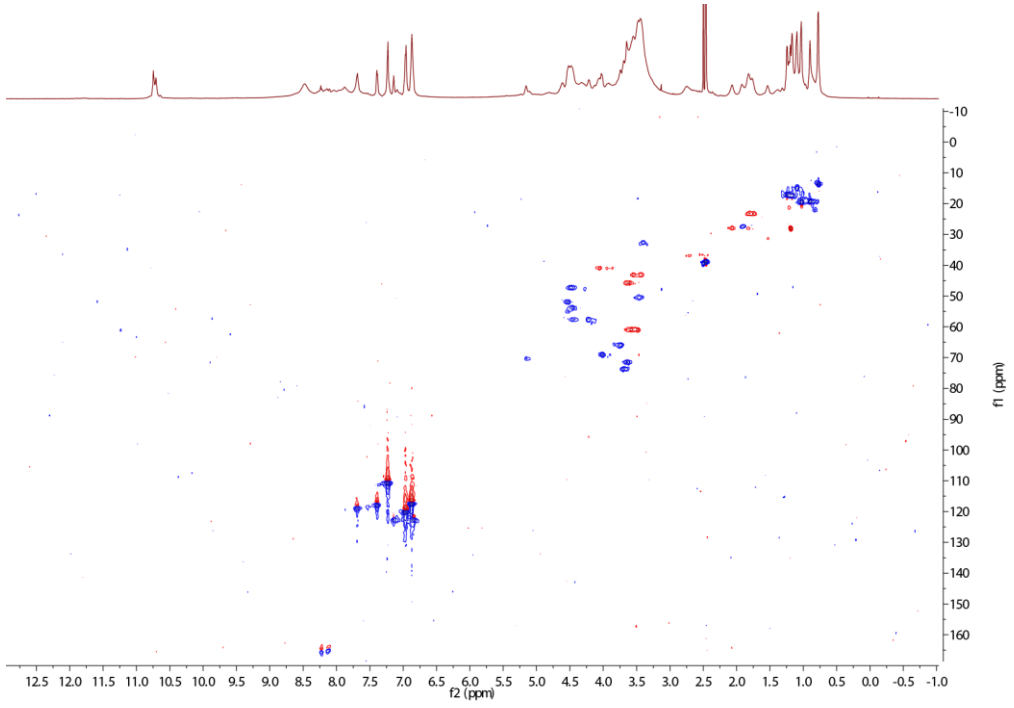

$^{1}$H-$^{1}$H COSY spectrum of telomycin D (4) in DMSO.



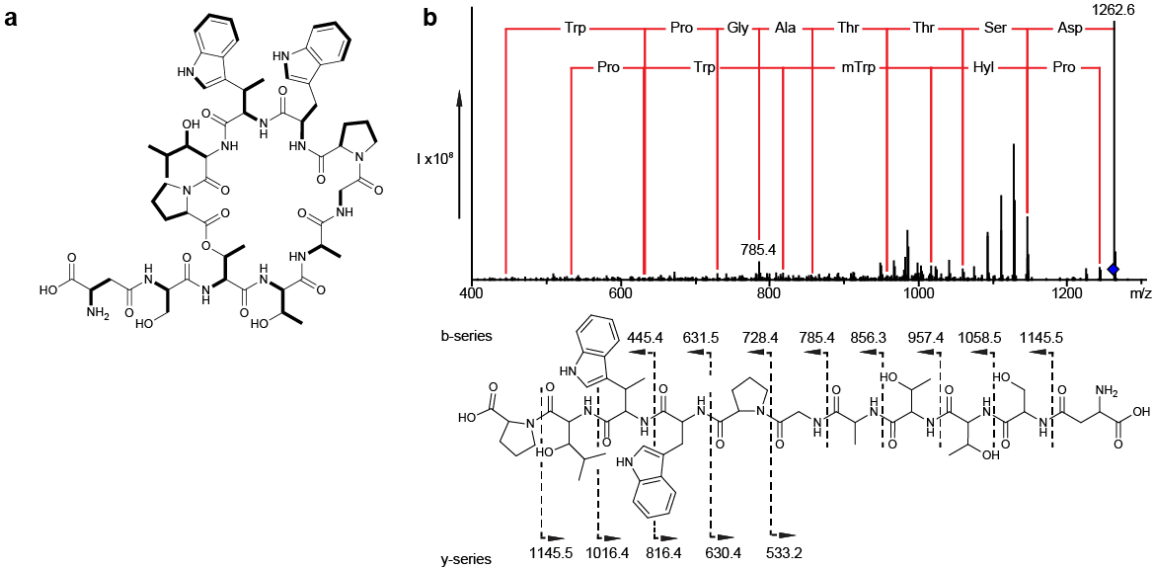
$^{1}H$-$^{1}H$ TOCSY spectrum of telomycin D (4) in DMSO.




$^{1}H$-$^{13}C$ HSQC spectrum of telomycin D (4) in DMSO.



## Telomycin E (5)




**Structure elucidation of telomycin E (5).** (**a**) Structure of telomycin E elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin E, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin E (5)

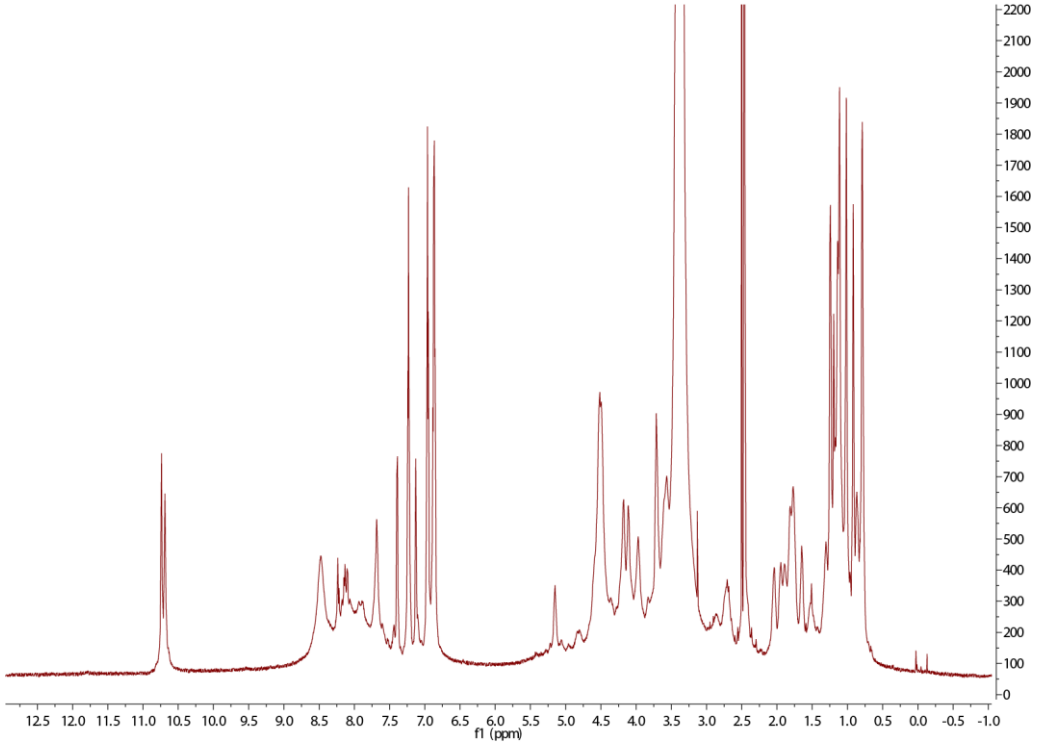
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin E (5) | $C_{59}H_{80}N_{13}O_{17}$ [M+H] | 1242.57896 | 1242.57716 | 1.894 |

NMR spectroscopic data for telomycin E (5) (700 MHz, DMSO-$d_6$).

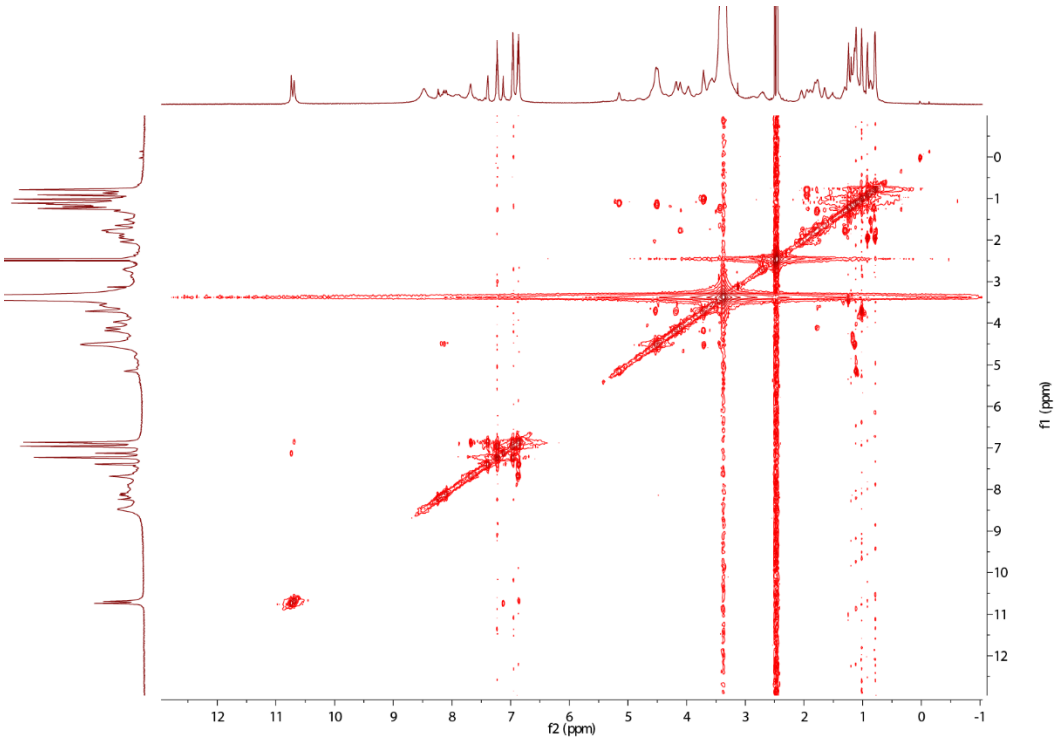
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | 2.44 (*br.*) | 36.76 |
| 2 | 7.80 (*br.*) | - | 38 | - | - |
| 3 | 4.21 (*ov.*) | 57.59 | 39 | 6.85 (*s*) | 122.59 |
| 4a | 2.73 (*br.*) | 37.01 | 40 | 10.69 (*s*) | - |
| 4b | 2.70 (*br.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.23 (*d*, 7) | 110.68 |
| 6 | 8.15 (*br.*) | - | 43 | 6.96 (*t*, 7) | 120.10 |
| 7 | 4.49 (*ov.*) | 47.32 | 44 | 6.86 (*t*, 7) | 117.70 |
| 8a | 3.56 (*ov.*) | 61.05 | 45 | 7.39 (*d*, 7) | 117.91 |
| 8b | 3.44 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.68 (*br.*) | - |
| 11 | 8.05 (*br.*) | - | 49 | 4.59 (*ov.*) | 57.72 |
| 12 | 4.47 (*ov.*) | 53.81 | 50 | 3.43 (*ov.*) | 32.62 |
| 13 | 5.15 (*br.*) | 70.61 | 51 | 1.23 (*d*, 6) | 17.33 |
| 14 | 1.11 (*m.*) | 15.20 | 52 | - | - |
| 15 | - | - | 53 | 7.13 (*s*) | 122.71 |
| 16 | 8.29 (*br.*) | - | 54 | 10.74 (*s*) | - |
| 17 | 4.17 (*br.*) | 57.80 | 55 | - | - |
| 18 | 3.72 (*ov.*) | 73.69 | 56 | 7.23 (*d*, 7) | 110.69 |
| 19 | 1.01 (*m.*) | 19.68 | 57 | 6.96 (*t*, 7) | 120.09 |
| 20 | - | - | 58 | 6.88 (*t*, 7) | 117.33 |
| 21 | - | - | 59 | 7.68 (*br.*) | 119.01 |
| 22 | 8.26 (*br.*) | - | 60 | - | - |
| 23 | 4.50 (*ov.*) | 53.94 | 61 | - | - |
| 24 | 1.14 (*m.*) | 17.55 | 62 | - | - |
| 25 | - | - | 63 | 5.03 (*br.*) | 55.62 |
| 26 | 7.96 (*br.*) | - | 64a | 3.38 (*ov.*) | 70.21 |
| 27a | 3.97 (*ov.*) | 41.03 | 64b | - | |
| 27b | 3.94 (*ov.*) | | 65 | 5.27 (*br.*) | - |
| 28 | - | - | 66 | 1.94 (*br.*) | 27.43 |
| 29 | 4.54 (*ov.*) | 52.01 | 67 | 0.92 (*d*, 5) | 19.34 |
| 30a | 2.03 (*br.*) | 27.89 | 68 | 0.79 (*d*, 5) | 13.91 |
| 30b | 1.89 (*br.*) | | 69 | - | - |
| 31 | - | - | 70 | 4.11 (*br.*) | 59.90 |
| 32a | 1.81 (*br.*) | 23.21 | 71a | 1.77 (*br.*) | 27.91 |
| 32b | 1.75 (*br.*) | | 71b | 1.29 (*br.*) | |
| 33a | 3.68 (*ov.*) | 45.74 | 72 | - | - |
| 33b | 3.61 (*ov.*) | | 73a | 1.77 (*br.*) | 23.32 |
| 34 | - | - | 73b | 1.64 (*br.*) | |
| 35 | 7.77 (*br.*) | - | 74a | 3.44 (*br.*) | 45.22 |
| 36 | 4.02 (*ov.*) | 60.02 | 74b | 3.38 (*ov.*) | |
| 37a | 2.68 (*br.*) | 36.79 | 75 | - | - |

*Carbon signals were taken from HSQC spectra

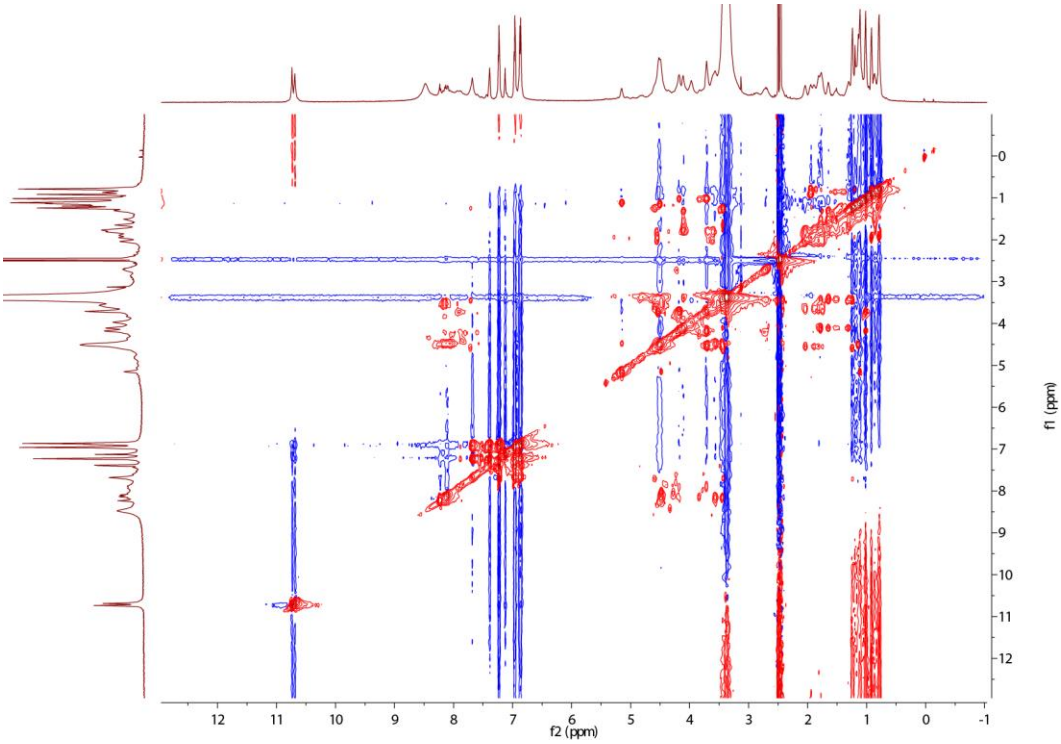
$^{1}$H NMR spectrum of telomycin E (5) in DMSO.



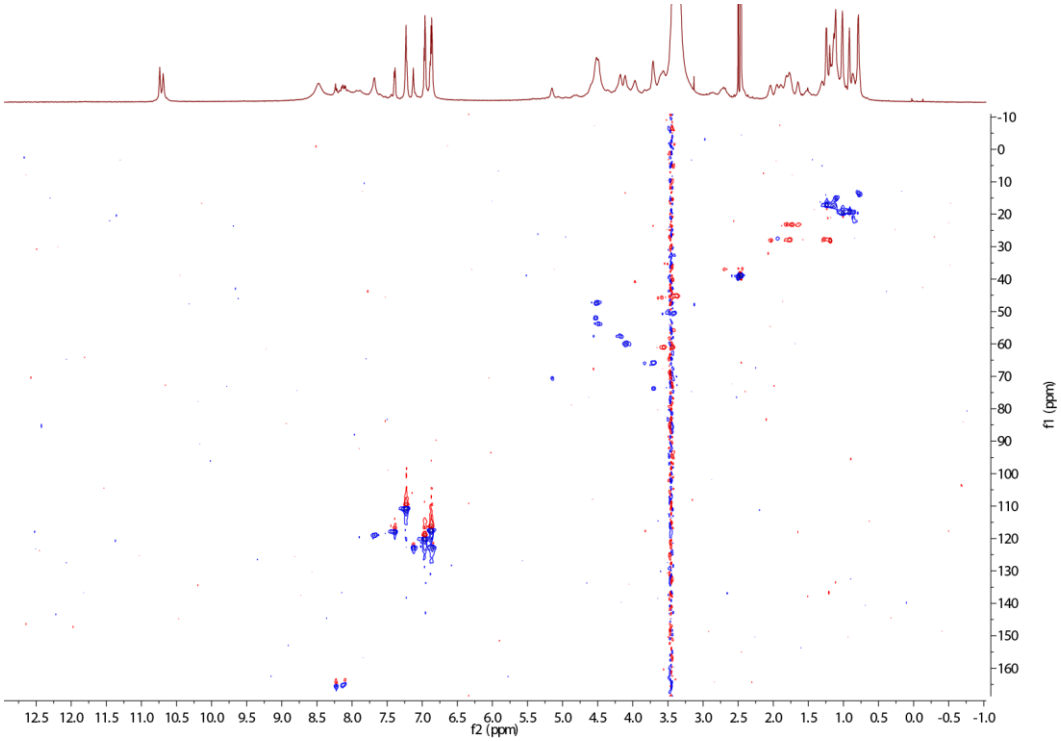

$^{1}$H-$^{1}$H COSY spectrum of telomycin E (5) in DMSO.



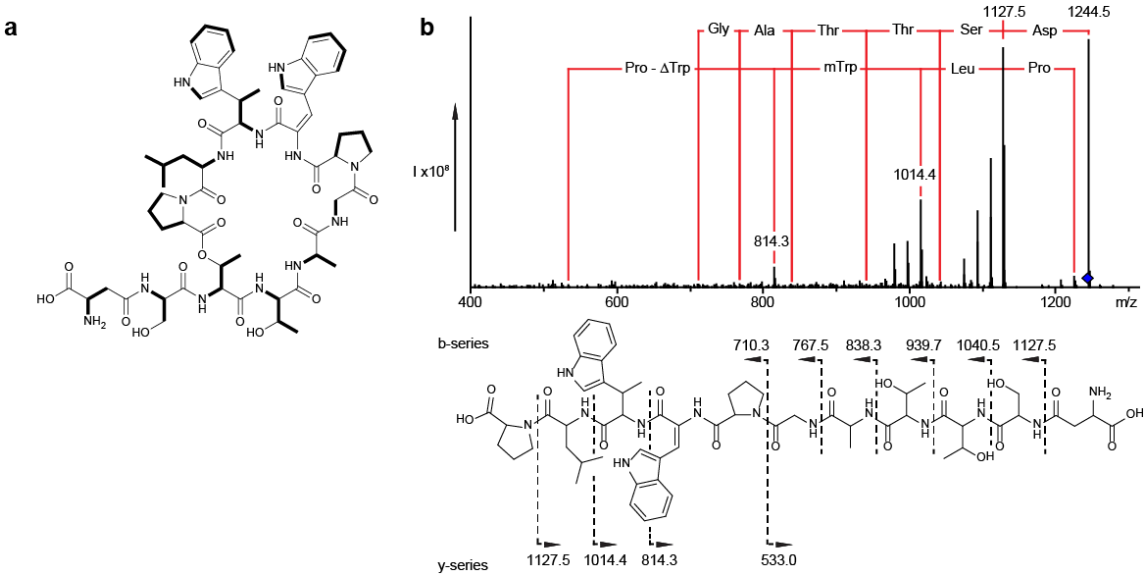
$^{1}$H-$^{1}$H TOCSY spectrum of telomycin E (5) in DMSO.




$^{1}$H-$^{13}$C HSQC spectrum of telomycin E (5) in DMSO.



## Telomycin F (6)




**Structure elucidation of telomycin F (6).** (**a**) Structure of telomycin F elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin F, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin F (6)

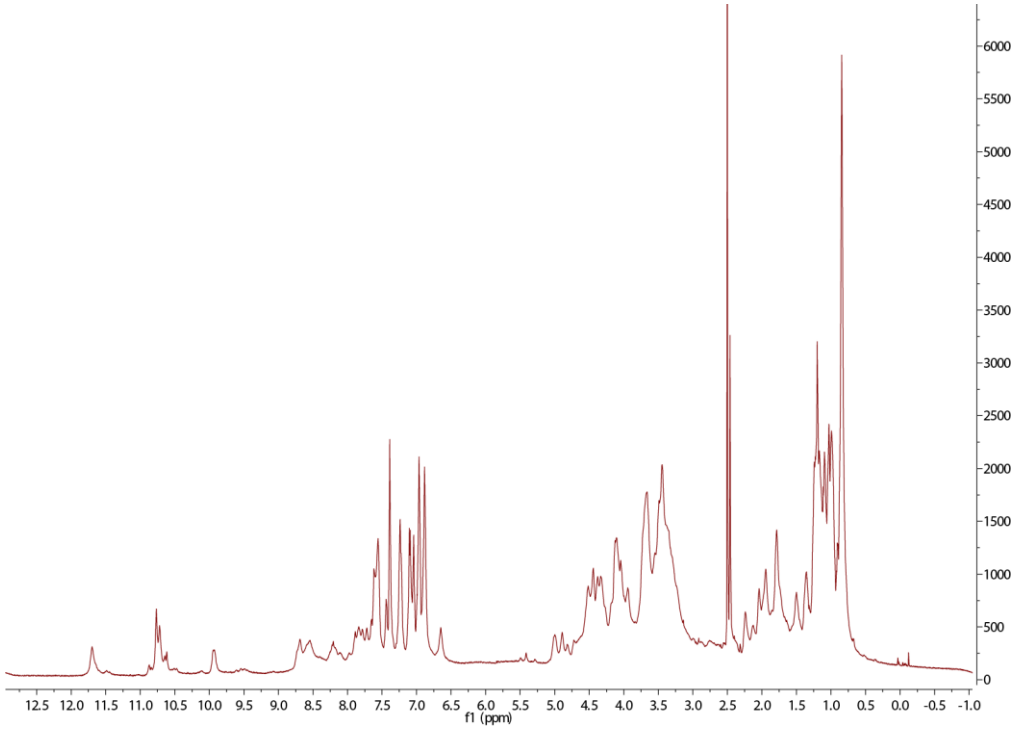
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin F (6) | $C_{59}H_{78}N_{13}O_{16}$ [M+H] | 1224.56840 | 1224.56768 | 1.036 |

NMR spectroscopic data for telomycin F (6) (700 MHz, DMSO-$d_6$).

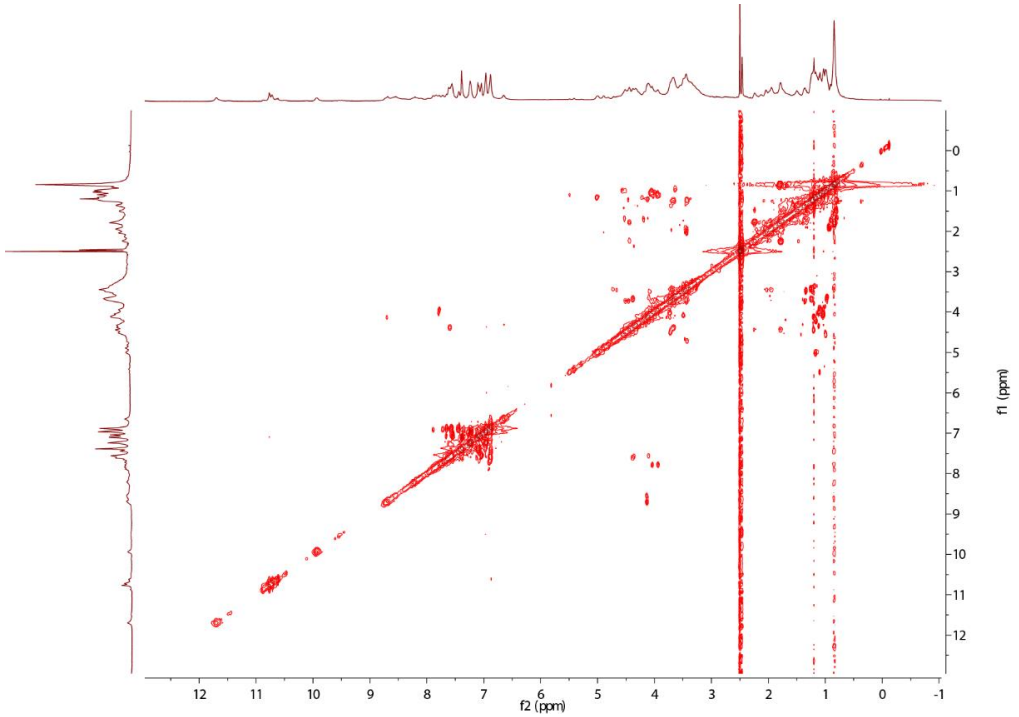
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | - | - |
| 2 | 7.86 (*br.*) | - | 38 | - | - |
| 3 | 4.34 (*ov.*) | 52.21 | 39 | 7.61 (*ov.*) | 126.50 |
| 4a | 2.98 (*br.*) | 25.72 | 40 | 11.70 (*s*) | - |
| 4b | 2.86 (*br.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.23 (*m.*) | 110.80 |
| 6 | 8.68 (*br.*) | - | 43 | 6.97 (*t*, 7) | 120.29 |
| 7 | 4.04 (*ov.*) | 58.10 | 44 | 6.87 (*t*, 7) | 117.70 |
| 8a | 3.73 (*ov.*) | 59.91 | 45 | 7.43 (*d*, 7) | 118.02 |
| 8b | 3.66 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.60 (*ov.*) | - |
| 11 | 8.38 (*br.*) | - | 49 | 4.37 (*ov.*) | 58.81 |
| 12 | 4.27 (*ov.*) | 56.80 | 50 | 3.66 (*ov.*) | 31.81 |
| 13 | 5.01 (*br.*) | 70.29 | 51 | 1.25 (*ov.*) | 18.01 |
| 14 | 1.17 (*ov.*) | 14.91 | 52 | - | - |
| 15 | - | - | 53 | 7.08 (*ov.*) | 122.30 |
| 16 | 7.56 (*br.*) | - | 54 | 10.76 (*s*) | - |
| 17 | 4.13 (*ov.*) | 59.09 | 55 | - | - |
| 18 | 3.64 (*ov.*) | 65.87 | 56 | 7.23 (*m.*) | 110.71 |
| 19 | 0.96 (*ov.*) | 19.92 | 57 | 6.98 (*t*, 7) | 120.04 |
| 20 | - | - | 58 | 6.88 (*t*, 7) | 117.60 |
| 21 | - | - | 59 | 7.57 (*m.*) | 118.52 |
| 22 | 7.89 (*br.*) | - | 60 | - | - |
| 23 | 4.52 (*ov.*) | 47.10 | 61 | - | - |
| 24 | 0.99 (*ov.*) | 16.61 | 62 | 6.64 (*br.*) | - |
| 25 | - | - | 63 | 4.31 (*ov.*) | 48.13 |
| 26 | 7.79 (*br.*) | - | 64a | 1.47 (*br.*) | 39.43 |
| 27a | 4.04 (*ov.*) | 45.42 | 64b | 1.37 (*br.*) | |
| 27b | 3.93 (*ov.*) | | 65 | - | - |
| 28 | - | - | 66 | 1.80 (*br.*) | 22.60 |
| 29 | 4.88 (*ov.*) | 56.88 | 67 | 0.85 (*m.*) | 22.00 |
| 30a | 2.20 (*br.*) | 27.81 | 68 | 0.83 (*m.*) | 20.71 |
| 30b | 2.04 (*br.*) | | 69 | - | - |
| 31 | - | - | 70 | 4.44 (*ov.*) | 59.21 |
| 32a | 1.95 (*br.*) | 23.80 | 71a | 2.24 (*br.*) | 27.92 |
| 32b | 1.76 (*br.*) | | 71b | 1.78 (*br.*) | |
| 33a | 3.49 (*ov.*) | 45.53 | 72 | - | - |
| 33b | 3.43 (*ov.*) | | 73a | 2.04 (*br.*) | 24.08 |
| 34 | - | - | 73b | 1.94 (*br.*) | |
| 35 | - | - | 74a | 3.69 (*ov.*) | 45.61 |
| 36 | - | - | 74b | 3.44 (*ov.*) | |
| 37a | - | - | 75 | - | - |

*Carbon signals were taken from HSQC spectra

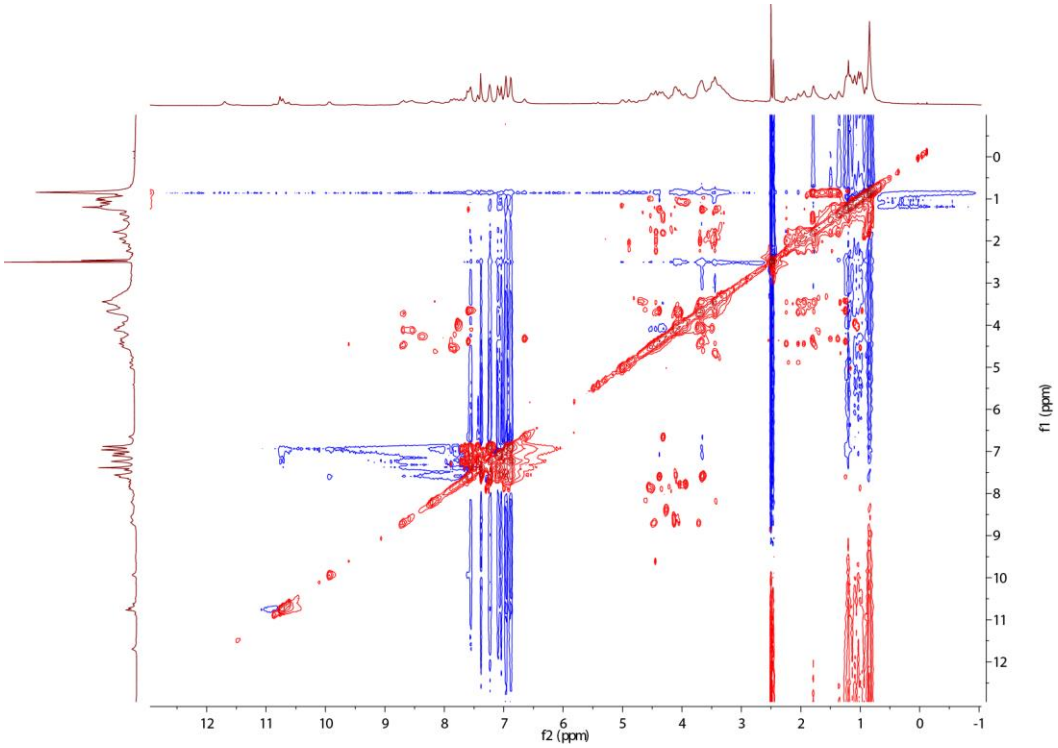
$^{1}$H NMR spectrum of telomycin F (6) in DMSO.



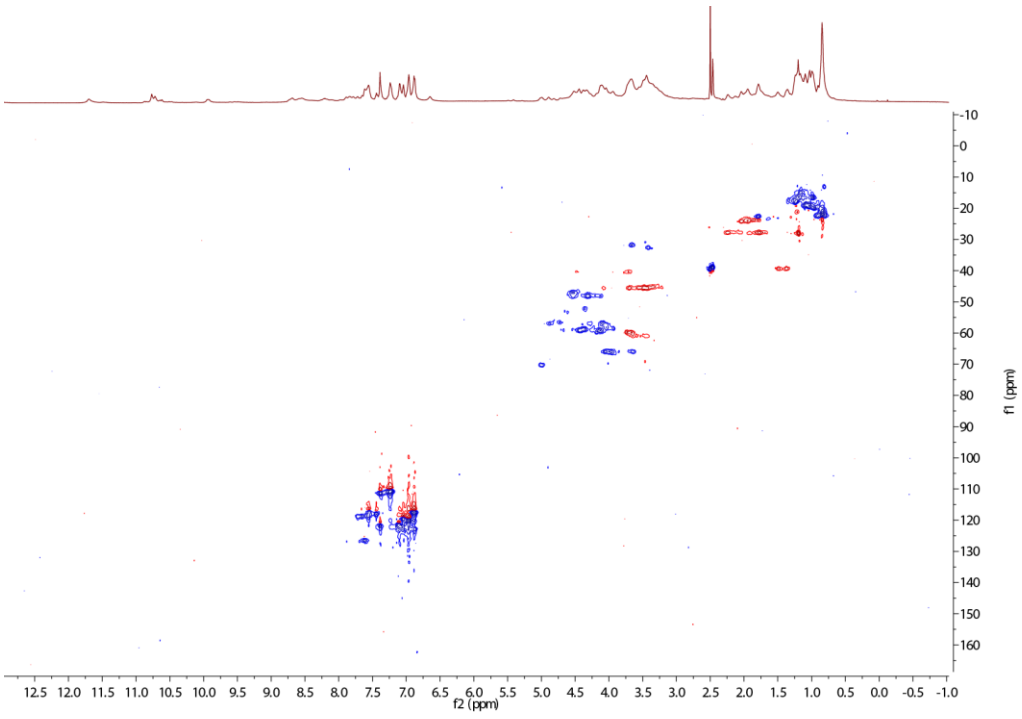

$^{1}$H-$^{1}$H COSY spectrum of telomycin F (6) in DMSO.



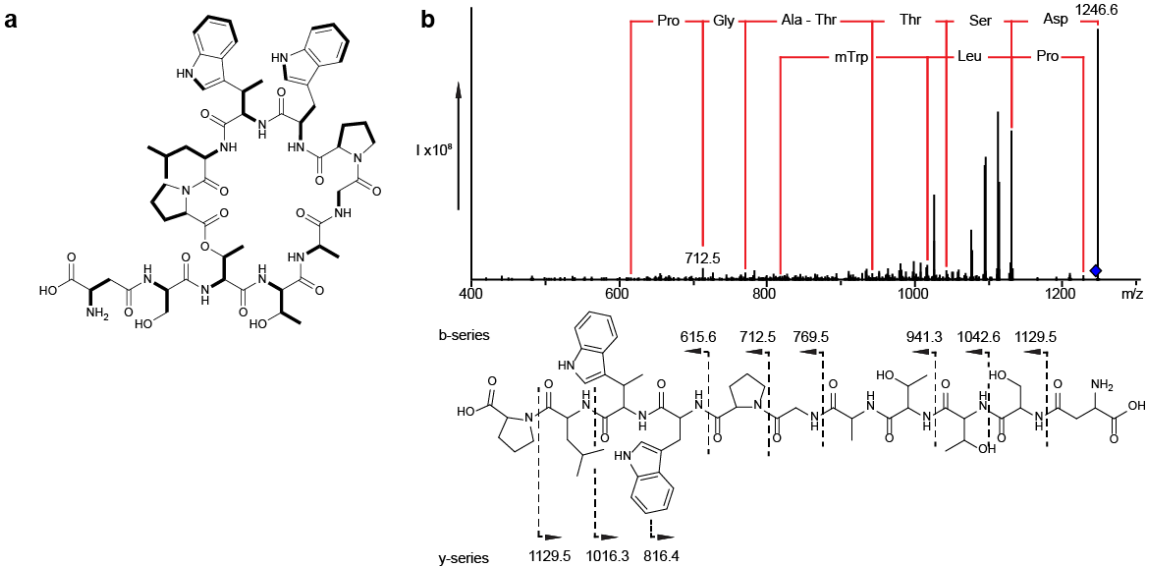
$^{1}H$-$^{1}H$ TOCSY spectrum of telomycin F (6) in DMSO.




$^{1}H$-$^{13}C$ HSQC spectrum of telomycin F (6) in DMSO.



## Telomycin G (7)




**Structure elucidation of telomycin G (7).** (**a**) Structure of telomycin G elucidated by comparative NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of base-hydrolyzed telomycin G, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for telomycin G (7)

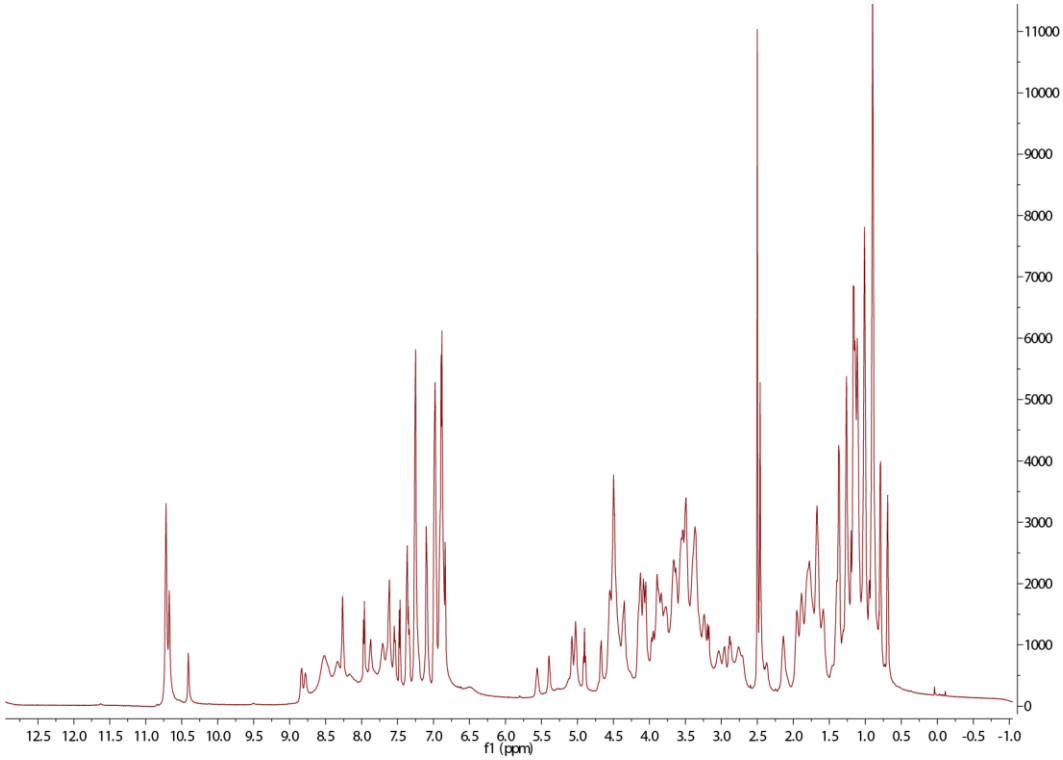
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Telomycin G (7) | $C_{59}H_{80}N_{13}O_{16}$ [M+H] | 1226.58405 | 1226.58187 | 2.225 |

NMR spectroscopic data for telomycin G (7) (700 MHz, DMSO-$d_6$).

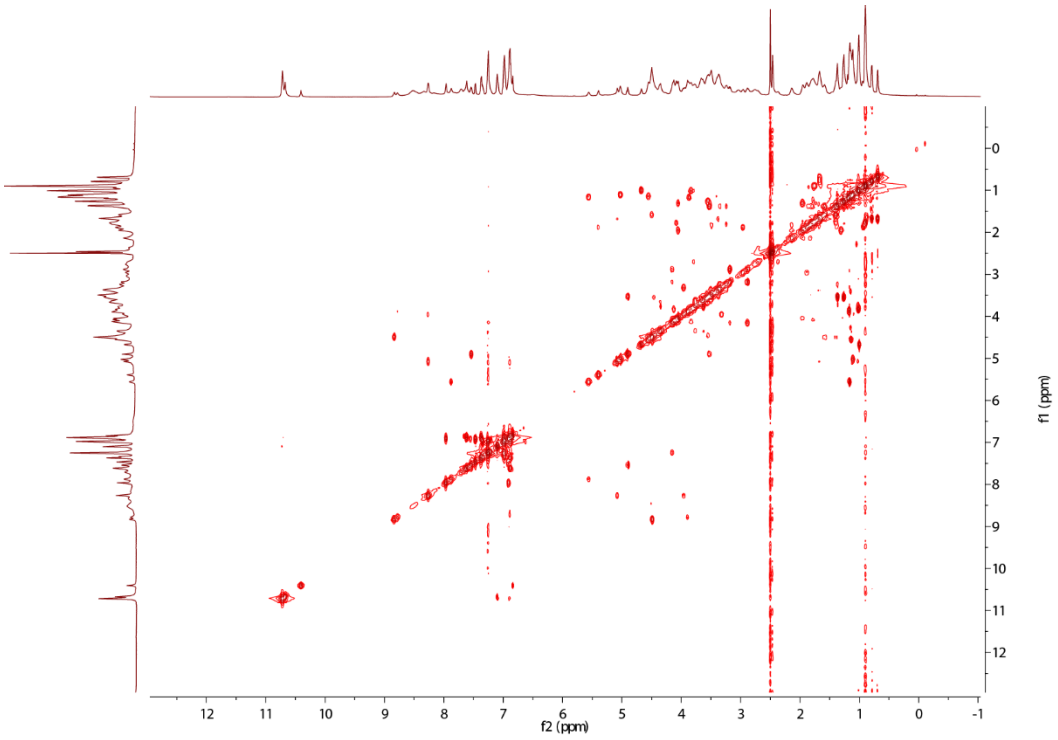
| Position | δH (mult.) | δC | Position | δH (mult.) | δC |
|---|---|---|---|---|---|
| 1 | - | - | 37b | 2.37 (*br.*) | 36.20 |
| 2 | 7.23 (*br.*) | - | 38 | - | - |
| 3 | 4.15 (*ov.*) | 52.59 | 39 | 7.10 (*s*) | 123.41 |
| 4a | 3.18 (*m.*) | 25.10 | 40 | 10.67 (*s*) | - |
| 4b | 2.88 (*m.*) | | 41 | - | - |
| 5 | - | - | 42 | 7.25 (*m.*) | 110.69 |
| 6 | 8.16 (*br.*) | - | 43 | 6.98 (*m.*) | 120.10 |
| 7 | 4.34 (*m.*) | 54.61 | 44 | 6.89 (*m.*) | 117.20 |
| 8a | 3.76 (*ov.*) | 60.90 | 45 | 7.37 (*d*, 7) | 117.71 |
| 8b | 3.64 (*ov.*) | | 46 | - | - |
| 9 | - | - | 47 | - | - |
| 10 | - | - | 48 | 7.56 (*br.*) | - |
| 11 | 8.27 (*br.*) | - | 49 | 4.89 (*t*, 9) | 55.81 |
| 12 | 4.50 (*ov.*) | 56.88 | 50 | 3.53 (*ov.*) | 32.90 |
| 13 | 5.02 (*m.*) | 70.10 | 51 | 1.36 (*d*, 7) | 15.82 |
| 14 | 1.11 (*m.*) | 15.10 | 52 | - | - |
| 15 | - | - | 53 | 7.09 (*s*) | 122.72 |
| 16 | 7.23 (*br.*) | - | 54 | 10.72 (*s*) | - |
| 17 | 4.12 (*ov.*) | 58.12 | 55 | - | - |
| 18 | 3.80 (*ov.*) | 65.70 | 56 | 7.25 (*m.*) | 110.70 |
| 19 | 1.01 (*m.*) | 19.31 | 57 | 6.98 (*m.*) | 120.11 |
| 20 | - | - | 58 | 6.90 (*m.*) | 117.41 |
| 21 | - | - | 59 | 7.62 (*m.*) | 118.70 |
| 22 | 8.16 (*br.*) | - | 60 | - | - |
| 23 | 4.55 (*ov.*) | 47.01 | 61 | - | - |
| 24 | 1.14 (*ov.*) | 16.77 | 62 | 7.63 (*br.*) | - |
| 25 | - | - | 63 | 4.49 (*m.*) | 47.91 |
| 26 | 8.27 (*br.*) | - | 64a | 1.57 (*br.*) | 39.20 |
| 27a | 3.95 (*m.*) | 41.62 | 64b | 1.38 (*ov.*) | |
| 27b | 3.31 (*ov.*) | | 65 | - | - |
| 28 | - | - | 66 | 1.76 (*br.*) | 22.79 |
| 29 | 4.51 (*br.*) | 52.27 | 67 | 0.90 (*d*, 6) | 22.31 |
| 130a | 2.15 (*br.*) | 27.41 | 68 | 0.89 (*d*, 6) | 20.69 |
| 30b | 2.12 (*br.*) | | 69 | - | - |
| 31 | - | - | 70 | 4.08 (*ov.*) | 59.90 |
| 32a | 1.89 (*br.*) | 23.51 | 71a | 1.80 (*br.*) | 27.71 |
| 32b | 1.82 (*br.*) | | 71b | 1.75 (*br.*) | |
| 33a | 3.67 (*br.*) | 45.31 | 72 | - | - |
| 33b | 3.5 (*br.*) | | 73a | 1.80 (*br.*) | 20.70 |
| 34 | - | - | 73b | 1.39 (*ov.*) | |
| 35 | 7.80 (*br.*) | - | 74a | 3.36 (*m.*) | 45.19 |
| 36 | 3.79 (*m.*) | 50.38 | 74b | 3.24 (*m.*) | |
| 37a | 2.70 (*br.*) | 36.18 | 75 | - | - |

*Carbon signals were taken from HSQC spectra

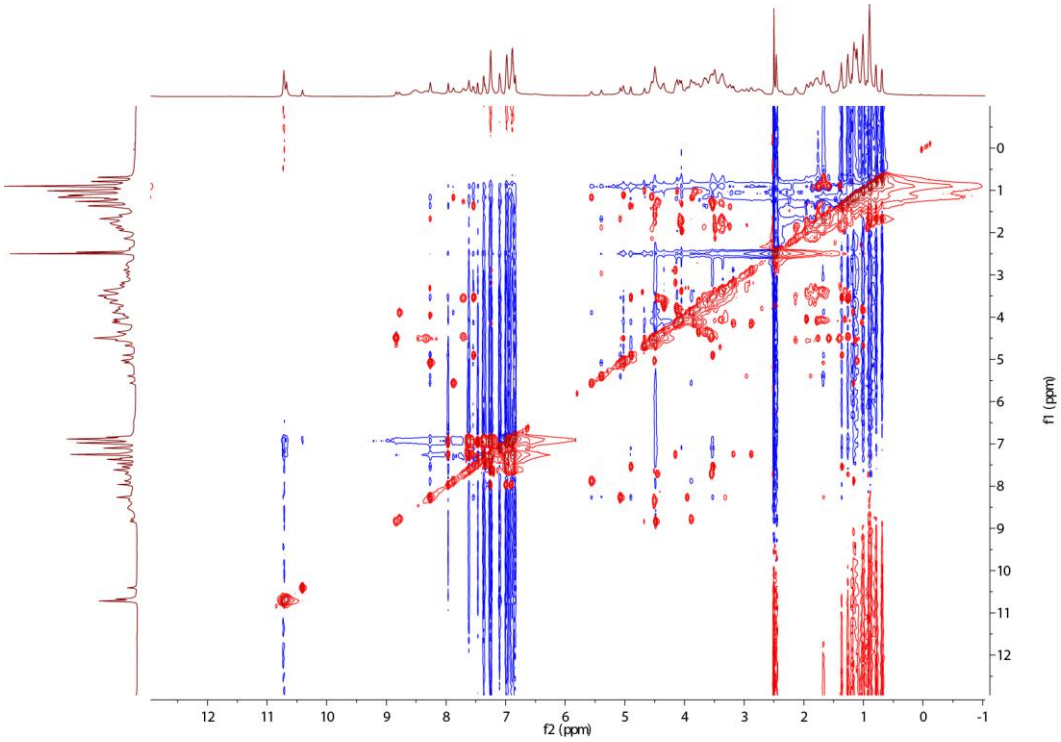
$^{1}$H NMR spectrum of telomycin G (7) in DMSO.



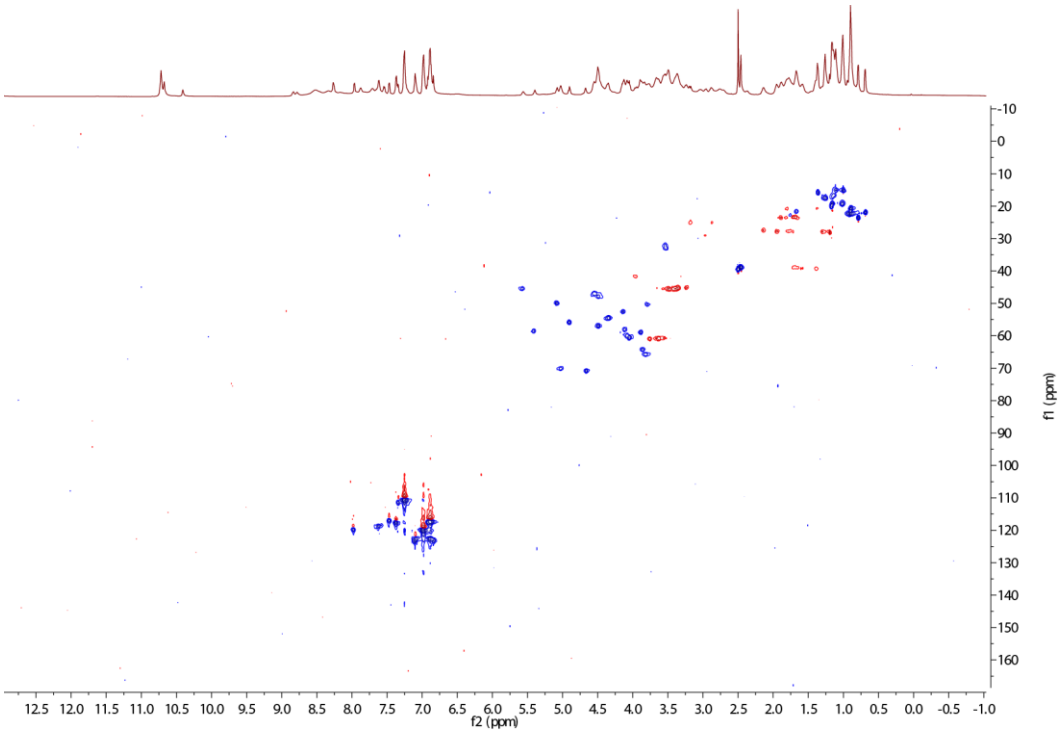

$^{1}$H-$^{1}$H COSY spectrum of telomycin G (7) in DMSO.



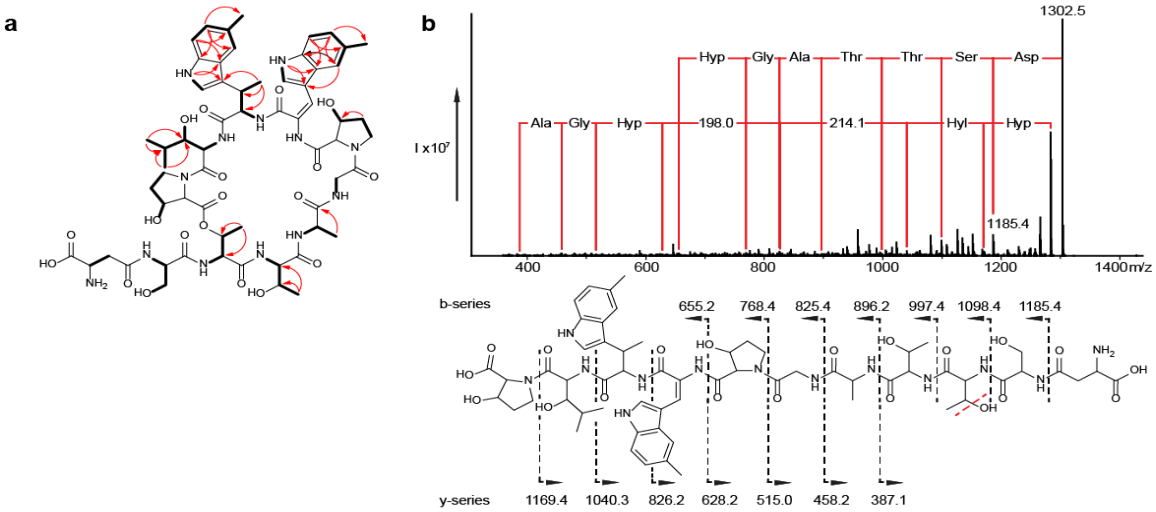
$^{1}H$-$^{1}H$ TOCSY spectrum of telomycin G (7) in DMSO.




$^{1}H$-$^{13}C$ HSQC spectrum of telomycin G (7) in DMSO.



## Di-5-methyltryptophan telomycin (8)




**Structure elucidation of di-5-methyltryptophan telomycin (8).** (**a**) Structure of di-5-methyltryptophan telomycin elucidated by NMR spectroscopy, HRMS, and MS/MS fragmentation. (**b**) Annotated MS2 spectra of di-5-methyltryptophan telomycin, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for di-5-methyltryptophan telomycin (8)

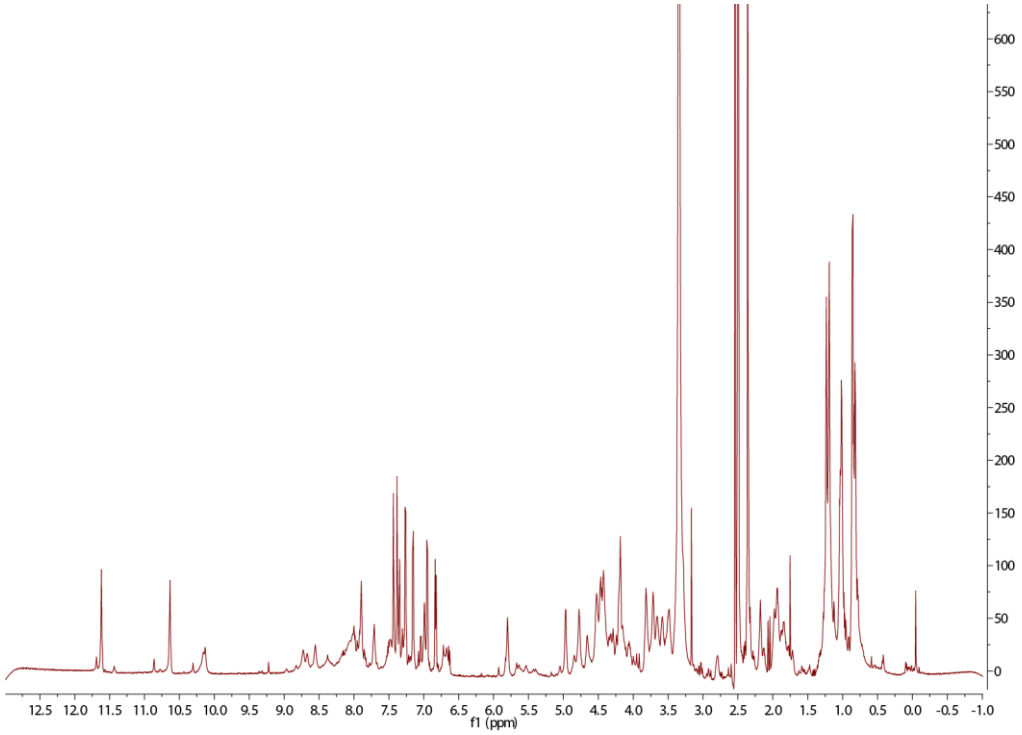
| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Di-5-methyltryptophan telomycin (8) | $C_{61}H_{82}N_{13}O_{19}$ [M+H] | 1300.58444 | 1300.58421 | 0.602 |

NMR spectroscopic data for di-5-methyltryptophan telomycin (8) (700 MHz, DMSO-$d_6$).

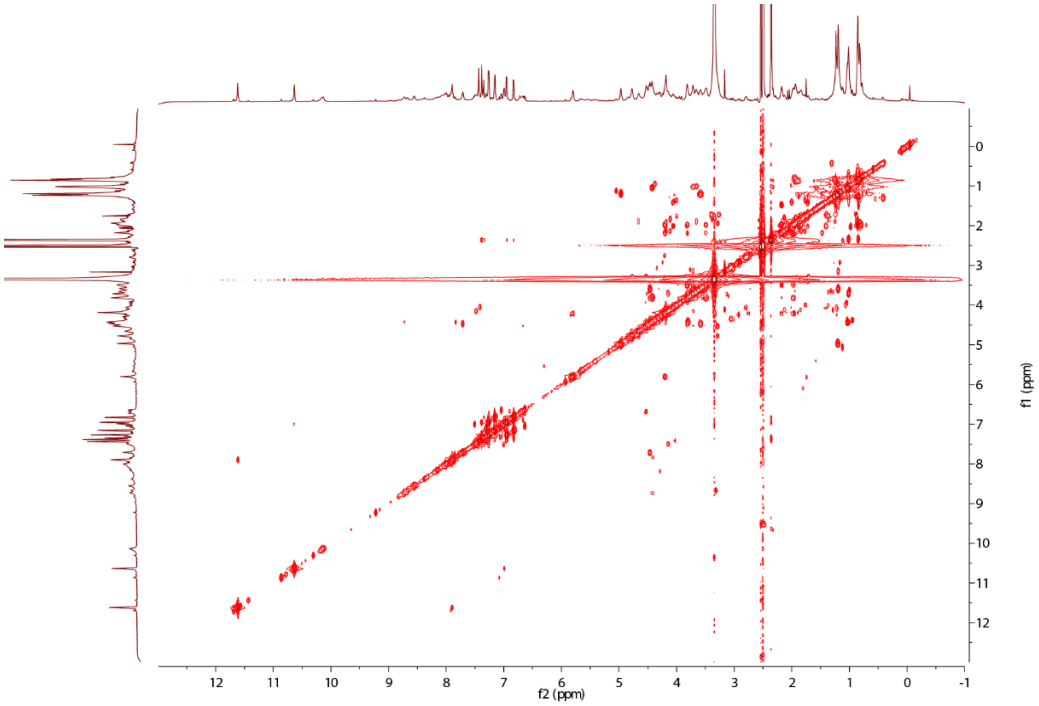
| Position | δH (mult.) | δC* | Position | δH (mult.) | δC* |
|---|---|---|---|---|---|
| 1 | - | - | 37b | - | - |
| 2 | 7.94 (*br.*) | - | 38 | - | 108 |
| 3 | 4.06 (*m.*) | 58.26 | 39 | 7.90 (*br.*) | 127.78 |
| 4a | 2.79 (*m.*) | - | 40 | 11.62 (*s*) | - |
| 4b | 2.80 (*m.*) | | 41 | - | 133.73 |
| 5 | - | - | 42 | 7.26 (*d*, 8) | 111.54 |
| 6 | 8.55 (*br.*) | - | 43 | 6.95 (*d*, 8) | 123.63 |
| 7 | 4.42 (*m.*) | 54.62 | 44 | - | 128.85 |
| 8a | 3.71 (*br.*) | 61.43 | 44-$CH_3$ | 2.36 (*s*) | 21.25 |
| 8b | 3.53 (*ov.*) | | 45 | 7.38 (*s*) | 117.25 |
| 9 | 5.53 (*br.*) | - | 46 | - | 127.45 |
| 10 | - | - | 47 | - | - |
| 11 | 8.36 (*br.*) | - | 48 | 7.71 (*br.*) | - |
| 12 | 4.31 (*ov.*) | 57.35 | 49 | 4.47 (*m.*) | 59.41 |
| 13 | 4.97 (*m.*) | 70.82 | 50 | 3.59 (*ov.*) | 32.36 |
| 14 | 1.20 (*ov.*) | 15.7 | 51 | 1.19 (*d*, 6) | 18.31 |
| 15 | - | - | 52 | - | 116.06 |
| 16 | 7.48 (*br.*) | - | 53 | 6.99 (*s*) | 122.94 |
| 17 | 4.15 (*ov.*) | 57.6 | 54 | 10.63 (*s*) | - |
| 18 | 3.71 (*br.*) | 66.79 | 55 | - | 135.07 |
| 19 | 1.01 (*ov.*) | 21.31 | 56 | 7.15 (*d*, 8) | 111.25 |
| 20 | - | - | 57 | 6.83 (*d*, 8) | 122.31 |
| 21 | - | - | 58 | - | 126.5 |
| 22 | 7.83 (*br.*) | - | 58-$CH_3$ | 2.36 (*s*) | 21.28 |
| 23 | 4.42 (*ov.*) | 47.91 | 59 | 7.34 (*s*) | 118.77 |
| 24 | 1.04 (*ov.*) | 17.15 | 60 | - | 125.7 |
| 25 | - | 171.67 | 61 | - | - |
| 26 | 8.7 (*br.*) | - | 62 | 6.67 (*br.*) | - |
| 27a | 4.41 (*ov.*) | 40.99 | 63 | 4.53 (*m.*) | 51.69 |
| 27b | 3.81 (*ov.*) | | 64a | 3.29 (*ov.*) | 74.86 |
| 28 | - | - | 64b | - | |
| 29 | 4.18 (*ov.*) | 66.67 | 65 | 4.79 (*br.*) | - |
| 30a | 4.20 (*ov.*) | 73.11 | 66 | 1.94 (*m.*) | 28.87 |
| 30b | - | | 67 | 0.86 (*d*, 7) | 19.84 |
| 31 | 5.80 (*br.*) | - | 68 | 0.81 (*d*, 7) | 17.01 |
| 32a | 1.98 (*m.*) | 33.39 | 69 | - | - |
| 32b | 2.17 (*m.*) | | 70 | - | - |
| 33a | 3.83 (*br.*) | 44.24 | 71a | 4.66 (*br.*) | 69.2 |
| 33b | 3.49 (*br.*) | | 71b | - | |
| 34 | - | - | 72 | 5.66 (*br.*) | - |
| 35 | - | - | 73a | 1.83 (*m.*) | 33.09 |
| 36 | - | - | 73b | 1.92 (*ov.*) | |
| 37a | - | - | 74a | 3.65 (*ov.*) | 45.05 |
| | | | 74b | 3.64 (*ov.*) | |
| | | | 75 | - | - |

*Carbon signals were taken from HSQC and HMBC spectra

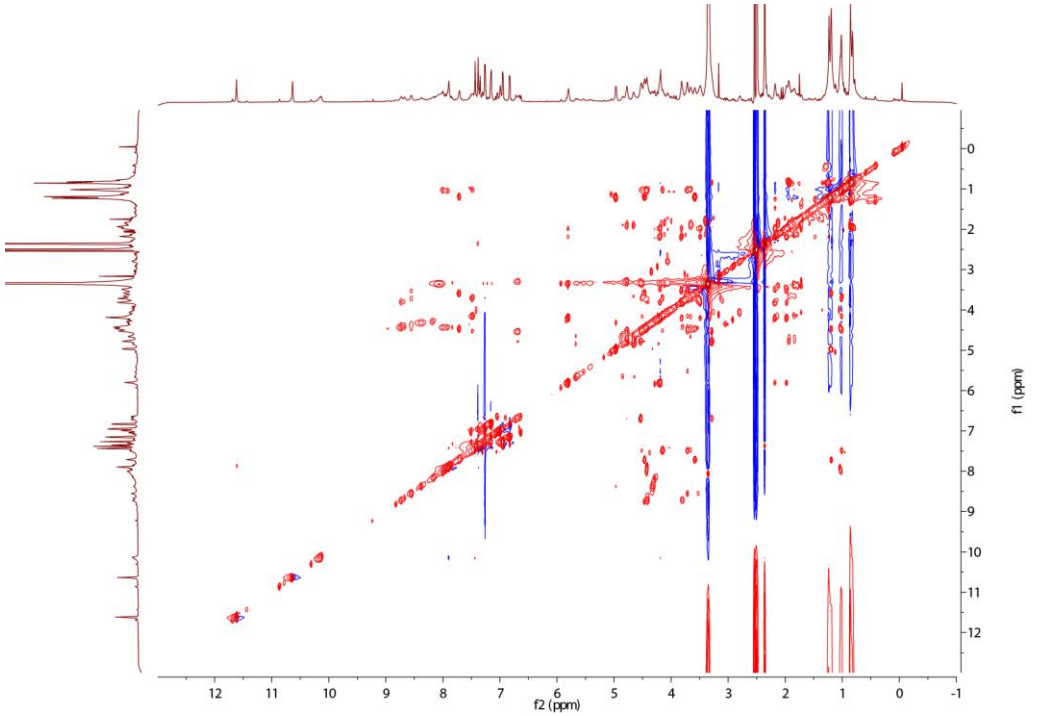
$^{1}$H NMR spectrum of di-5-methyltryptophan telomycin (8) in DMSO.



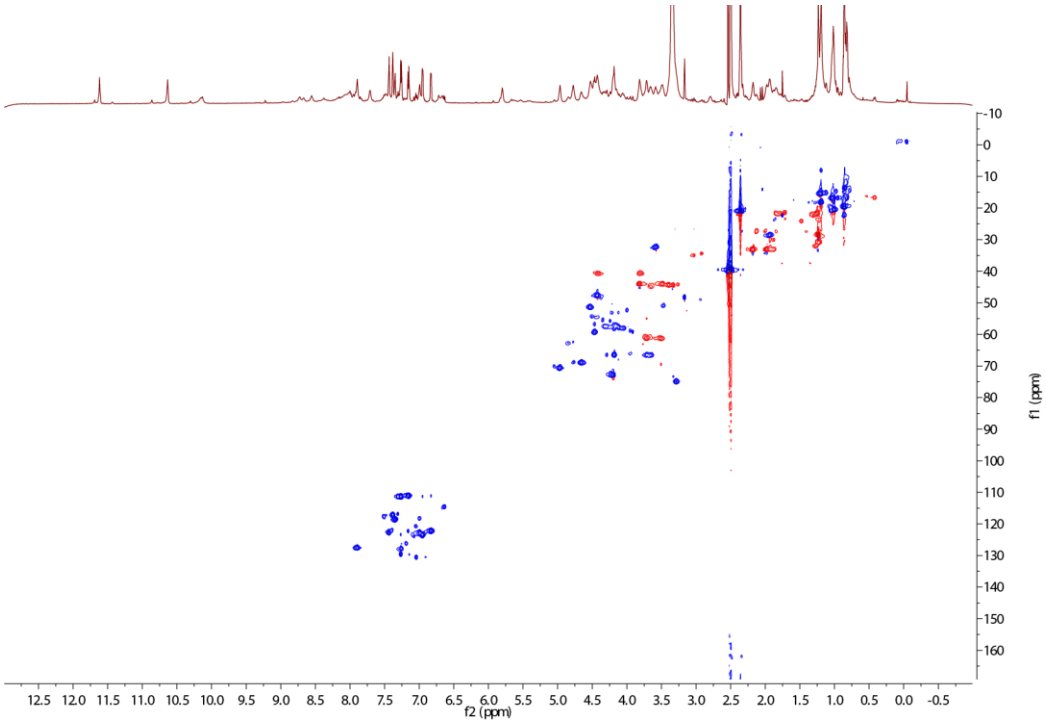

$^{1}$H-$^{1}$H COSY spectrum of di-5-methyltryptophan telomycin (8) in DMSO.



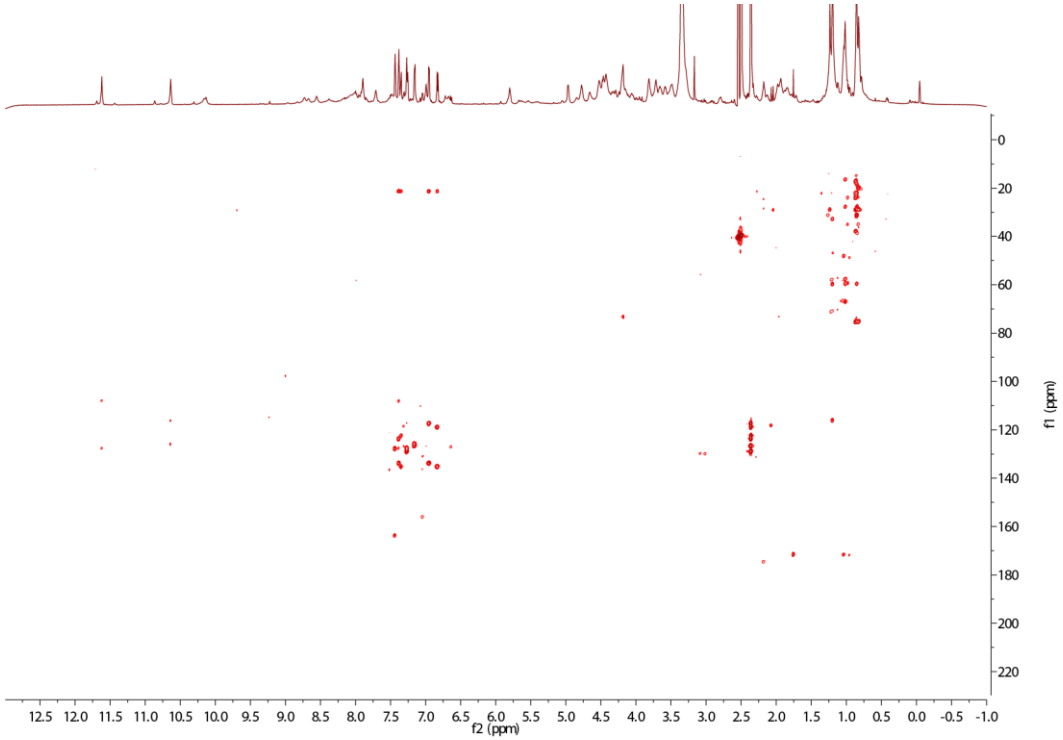
$^{1}H$-$^{1}H$ TOCSY spectrum of di-5-methyltryptophan telomycin (8) in DMSO.



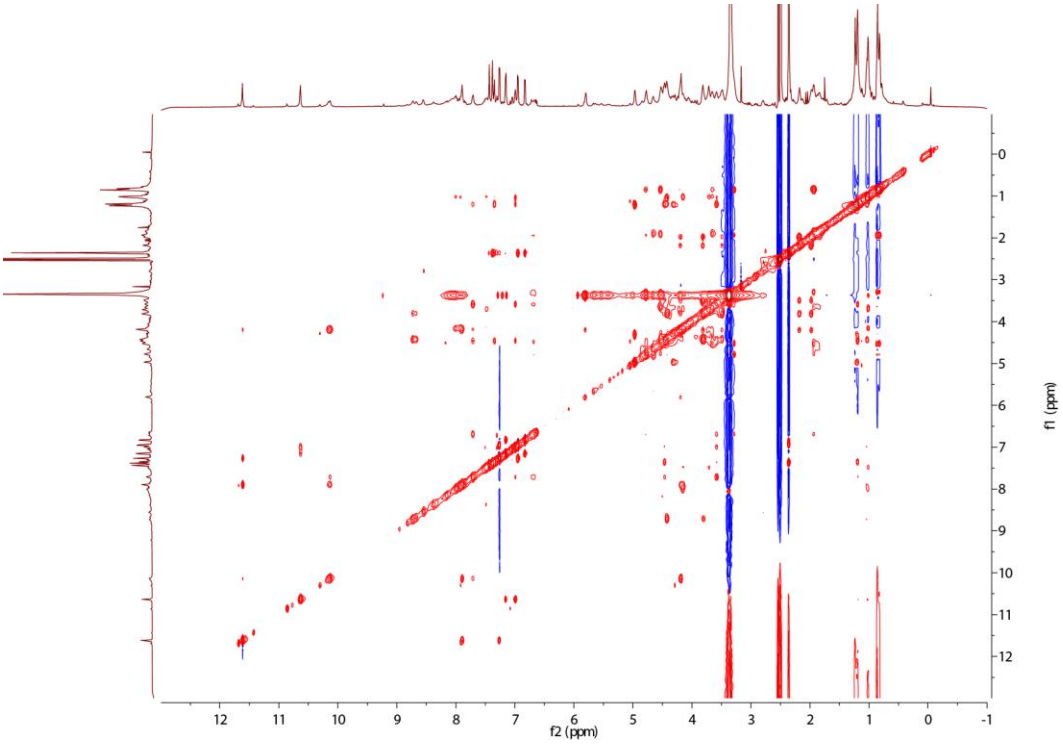

$^{1}H$-$^{13}C$ HSQC spectrum of di-5-methyltryptophan telomycin (8) in DMSO.



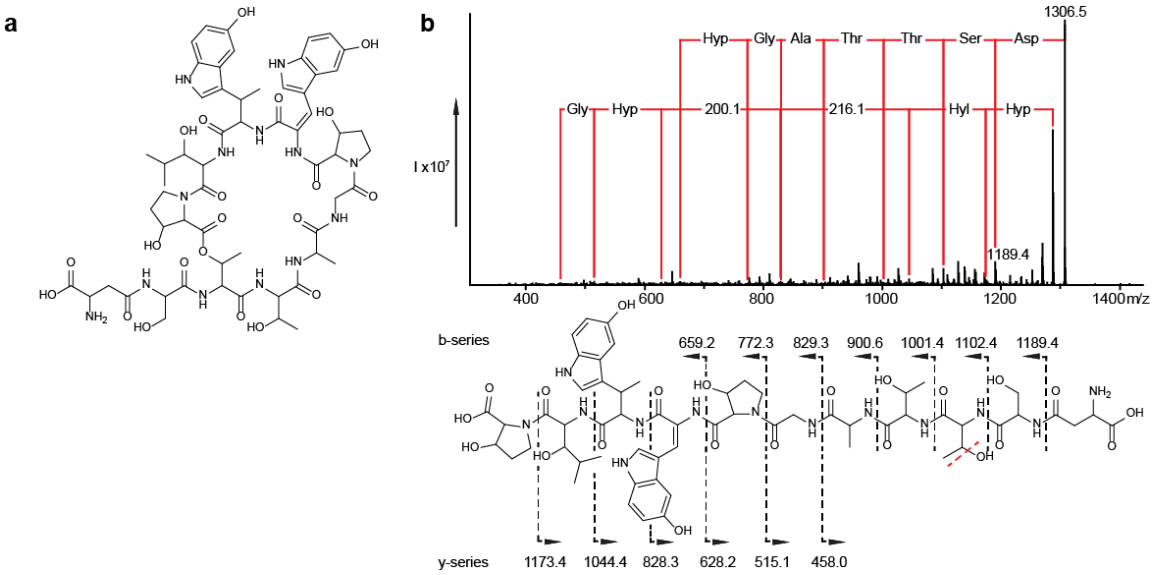
$^{1}H$-$^{13}C$ HMBC spectrum of di-5-methyltryptophan telomycin (8) in DMSO.




$^{1}H$-$^{1}H$ NOESY spectrum of di-5-methyltryptophan telomycin (8) in DMSO.



## Di-5-hydroxytryptophan telomycin (9)

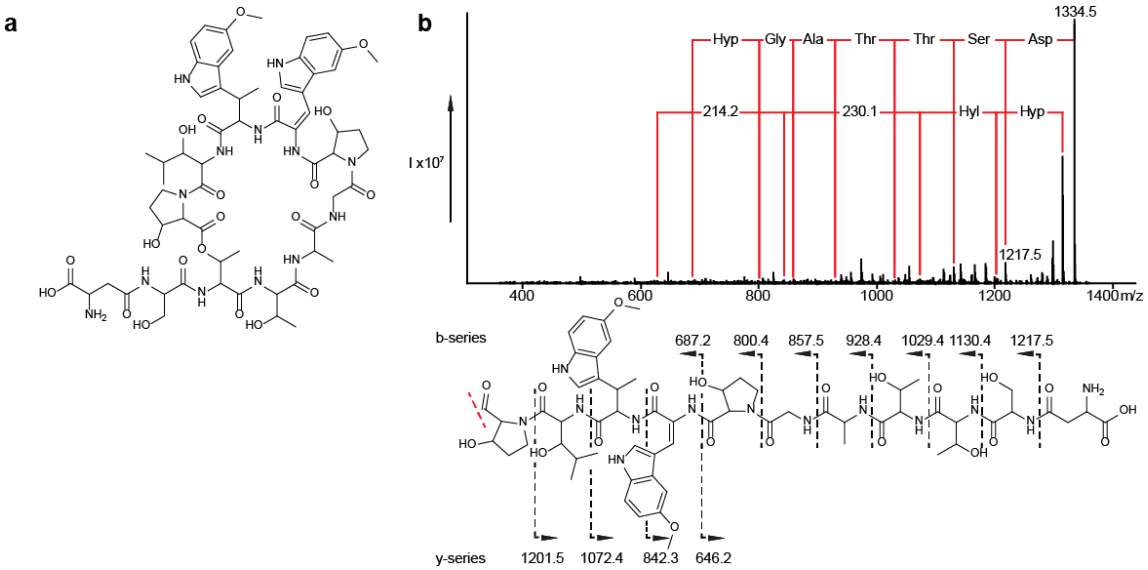



**Structure elucidation of di-5-hydroxytryptophan telomycin (9).** (**a**) Structure of di-5-hydroxytryptophan telomycin elucidated by HRMS and MS/MS fragmentation. (**b**) Annotated MS2 spectra of di-5-hydroxytryptophan telomycin, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for di-5-hydroxytryptophan telomycin (9)

| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Di-5-hydroxytryptophan telomycin (9) | $C_{59}H_{78}N_{13}O_{21}$ [M+H] | 1304.54297 | 1304.54333 | 0.147 |

## Di-5-methoxytryptophan telomycin (10)




**Structure elucidation of di-5-methoxytryptophan telomycin (10).** (**a**) Structure of di-5-methoxytryptophan telomycin elucidated by HRMS and MS/MS fragmentation. (**b**) Annotated MS2 spectra of di-5-methoxytryptophan telomycin, including observed b- and y-ions resulting from amide bond cleavage.

## High resolution mass data for di-5-methoxytryptophan telomycin (10)

| Compound | Formula | Calc. | Obs. | Δppm |
|---|---|---|---|---|
| Di-5-methoxytryptophan telomycin (10) | $C_{61}H_{82}N_{13}O_{21}$ [M+H] | 1332.57427 | 1332.57507 | 0.186 |

## Appendix 3

## Supplementary Information

## Informatic Analysis Reveals *Legionella* as a Source of Novel Natural Products

Chad W. Johnston[1], Jonathan Plumb[2], Xiang Li[1], Sergio Grinstein[2], Nathan A. Magarvey[1*]

[1]The Michael G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences; Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON, Canada L8N 3Z5.

[2] Cell Biology Program, Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada.

*Corresponding author:
Nathan Magarvey, magarv@mcmaster.ca

## Supplementary Figures


**Figure S1.** PRISM analysis of *Legionella* genomes. **A**. PRISM overview of a selection of polyketide and nonribosomal peptide gene clusters identified within the genome of *Legionella dumoffii* NY23 (also known as *Fluoribacter dumoffii* NY23). **B**. Domain and cluster analysis output for *L. dumoffii* NY23 polyketide cluster 1, encoding a novel trans-AT polyketide synthase. The trans-acting acyltransferase is predicted to specify malonate (Mal) as a substrate.

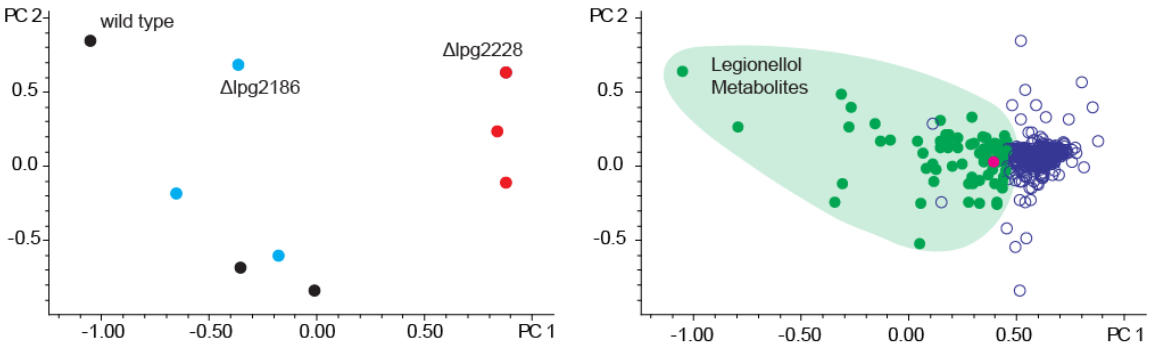



**Figure S2. Principal component analysis reveals the legionellol family of metabolites.** Principal component analysis of three whole culture extracts of wild type (black), Δlpg2186 (cyan), and Δlpg2228 (red) confirmed that while wild type and Δlpg2186 are indistinguishable, Δlpg2228 lacks a series of related metabolites (*right*, green) apparently originating from a common scaffold (458.2 Da; pink).

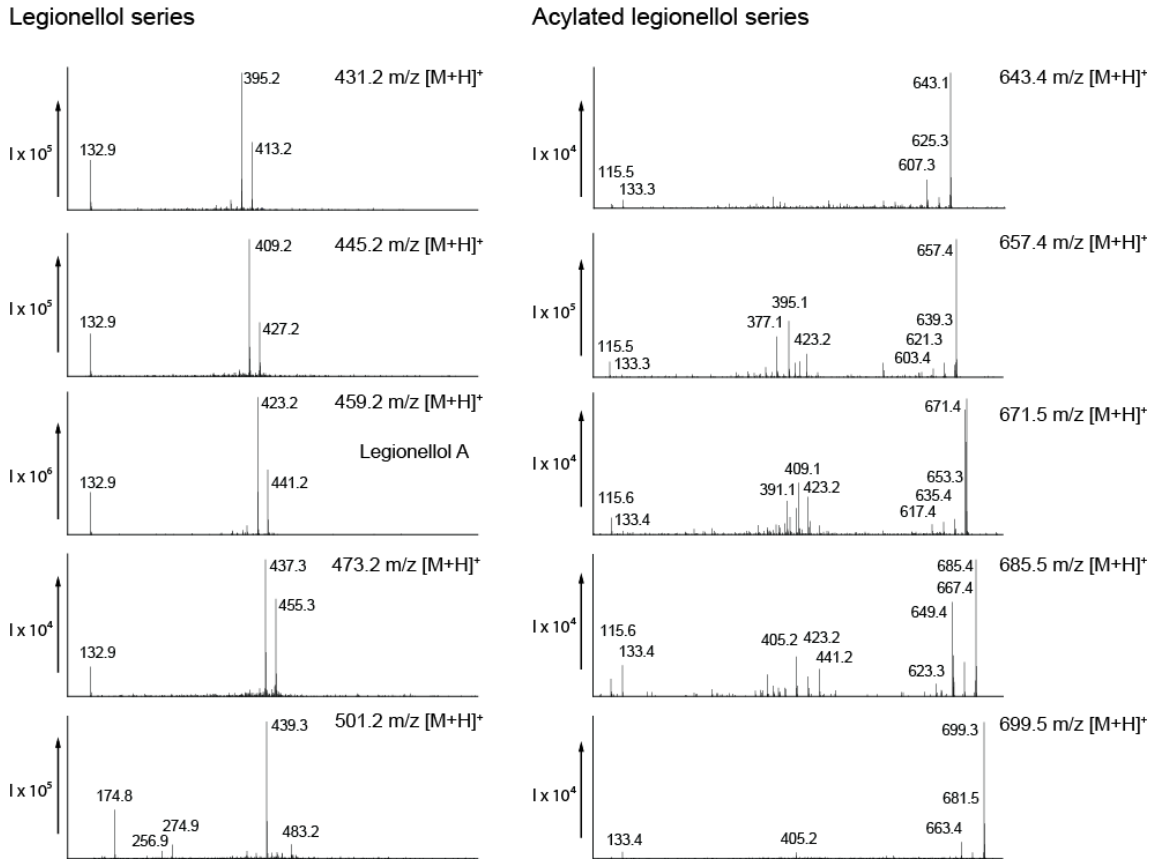



**Figure S3.** Legionellol series compounds identified through principal component analysis. *Left*: Product ion spectra of the legionellol series molecules are shown, including the conserved 133 *m/z* fragment for the liberated modified ornithine from the legionellol structure. A minor variant appears to possess an ornithine to arginine substitution (*bottom left*), as the 133 *m/z* fragment has been replaced by a 175 *m/z* fragment. *Right*: Product ion spectra of the acylated legionellol series molecules are shown, including the conserved 133 *m/z* fragment. Based on retention time and MS/MS analysis, these analogs presumably vary from the legionellol structure by the addition of a fatty acid.

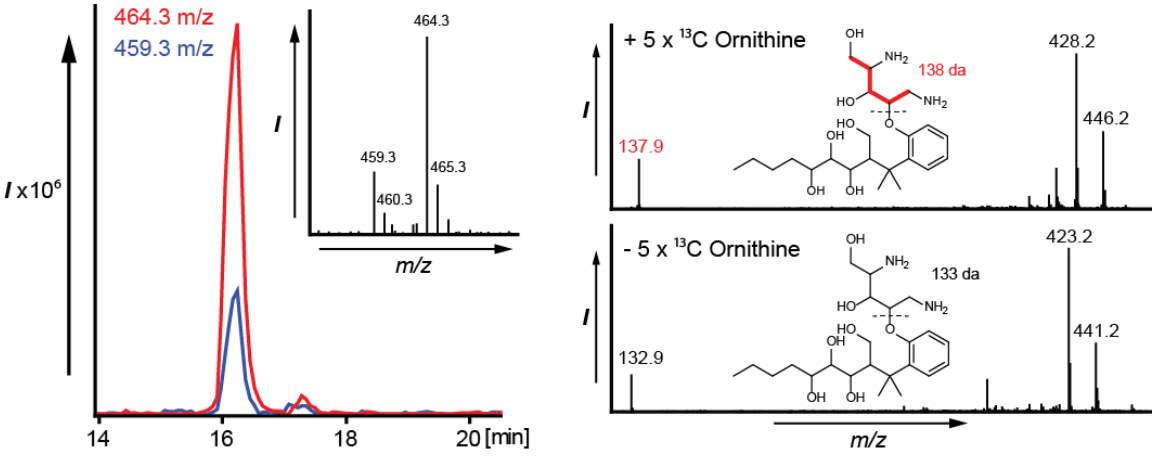



**Figure S4.** Verification of ornithine incorporation into legionellol. Cultures of *L. pneumophila* in chemically defined media were fed 2mM $^{13}$C ornithine to assess the origins of the putative modified amino acid in the legionellol structure. Incorporation of the five $^{13}$C carbons of the fed ornithine was observed in the parent (*left*), and fragment ions (*right*) of legionellol.

## Supplementary Tables

**Supplementary Table 1.** Genes of a *L. pneumophila* hybrid biosynthetic gene cluster (lpg1936-1943).

| Locus tag | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| lpg1936 | Methoxymalonate methyltransferase | - | 630 |
| lpg1937 | SyrP-like dioxygenase | - | 353 |
| lpg1938 | Coenzyme F390 synthetase, Adenylation domain | + | 493 |
| lpg1939 | Beta-ketoacyl synthase | - | 438 |
| lpg1940 | Adenylation domain | - | 506 |
| lpg1941 | Acyl CoA dehydrogenase | - | 375 |
| lpg1942 | 3-hydroxyacyl-CoA dehydrogenase | - | 284 |
| lpg1943 | Acyl carrier protein | - | 85 |

**Supplementary Table 2.** Genes of a *L. pneumophila* hybrid biosynthetic gene cluster (lpg2177-2186).

| Locus tag | Predicted Function | Strand | Amino Acids |
|---|---|---|---|
| lpg2177 | HlyD family secretion protein | + | 368 |
| lpg2178 | Multidrug resistance efflux pump | + | 1051 |
| lpg2179 | Nonribosomal peptide synthetase | + | 1453 |
| lpg2180 | Sensory box histidine kinase | + | 828 |
| lpg2181 | Response regulator | + | 302 |
| lpg2182 | Trans-aconitate methyltransferase | + | 259 |
| lpg2183 | SyrP-like dioxygenase | + | 350 |
| lpg2184 | Transposase | - | 468 |
| lpg2185 | Hypothetical protein | - | 102 |
| lpg2186 | Polyketide synthase | + | 3780 |

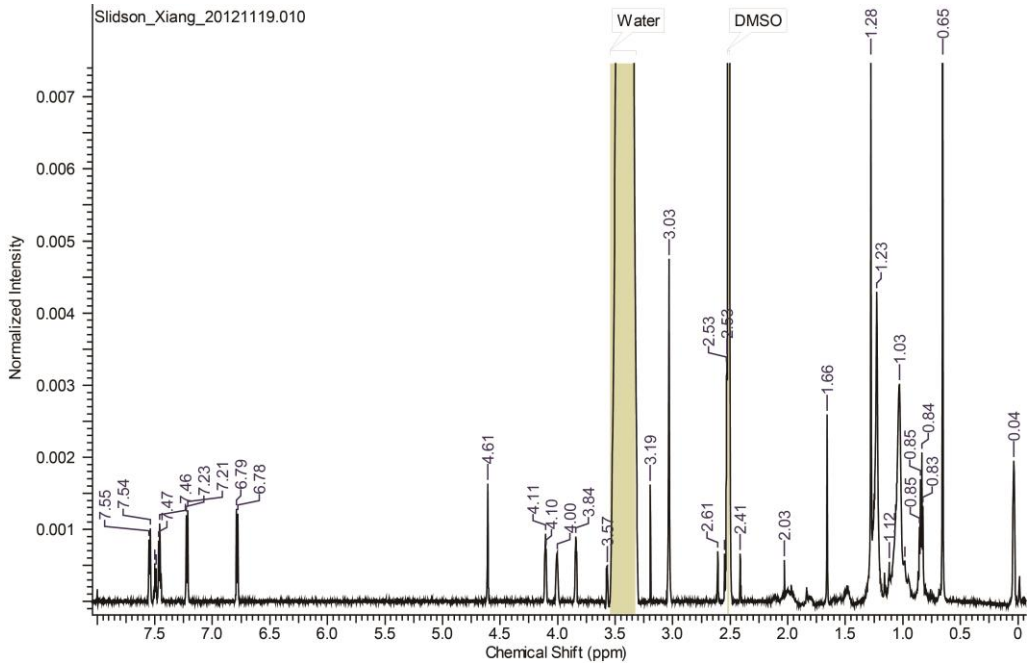
**Supplementary Table 3.** Primers used in the construction of *L. pneumophila* mutant strains.

| PCR Primer | Sequence (5'-3') | Purpose |
|---|---|---|
| ChlorBspHI F | TTTTCATGACTAAATACCTGTGACGGAAG | Amplification of a chloramphenicol resistance cassette from pRE112 to replace the pBlueScript KSII ampicillin resistance cassette. |
| ChlorBspHI R | TTTTCATGACTATCACTTATTCAGGCGTA | Amplification of a chloramphenicol resistance cassette from pRE112 to replace the pBlueScript KSII ampicillin resistance cassette. |
| KanKPN1 | TTTGGTACCGGTCTGACGCTCAGTGGAACG | Amplification of kanamycin resistance cassette from pET-28b to insertionally inactivate *Legionella* genes. |
| KanSPH1 | TTTGCATGCTTAGAAAAACTCATCGAGCATC | Amplification of kanamycin resistance cassette from pET-28b to insertionally inactivate *Legionella* genes. |
| 1939XBA1 | TTTTCTAGAGGCACTTTATACTCAAATGG | Amplification of lpg1939 gene fragment. |
| 1939KPN1 | TTTGGTACCTTTTTCAGGGCGTATATTTC | Amplification of lpg1939 gene fragment. |
| 1939SPH1 | TTTGCATGCAGCGTTCAAAGACGTTAAAA | Amplification of lpg1939 gene fragment. |
| 1939SAC1 | TTTGAGCTCCCAGATGTGTATTGCTGCCTC | Amplification of lpg1939 gene fragment. |
| 1939 Seq | GCTCTCTAACGAGCAATCC | Confirmation of genomic integration of the inactivated allele. |
| 2186XBA1 | TTTTCTAGACTTCTTCGTTCACTGCACTC | Amplification of lpg2186 gene fragment. |

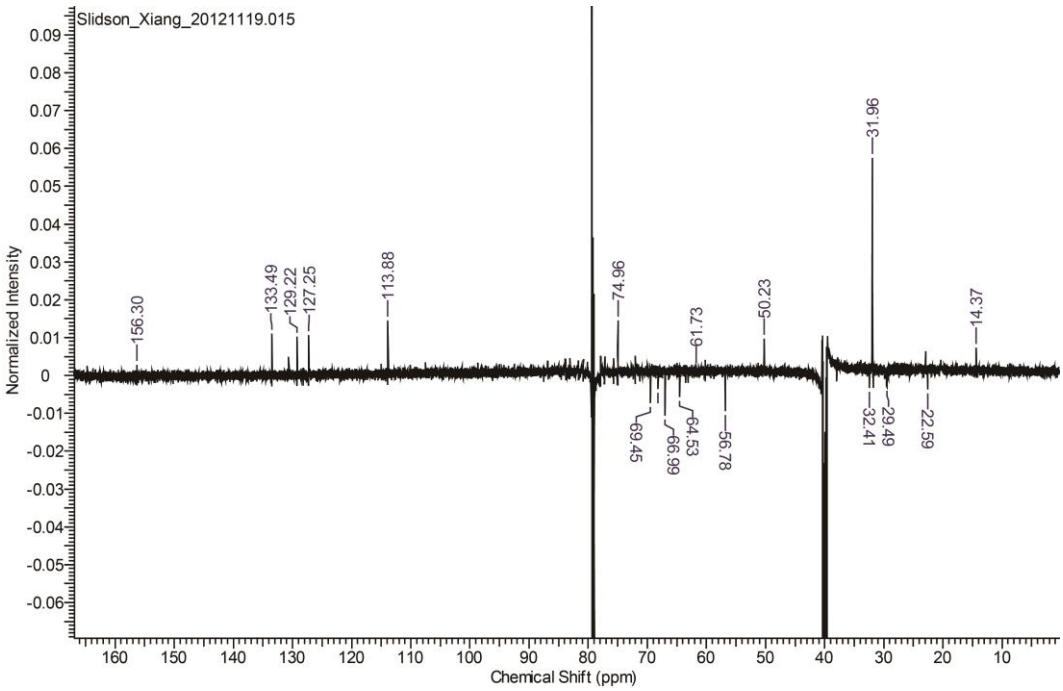
| | | |
|---|---|---|
| 2186KPN1 | TTTGGTACCAGGATCAATCAAGCTGTTCG | Amplification of lpg2186 gene fragment. |
| 2186SPH1 | TTTGCATGCGAAAGTATCCCTGCGCTTCT | Amplification of lpg2186 gene fragment. |
| 2186SAC1 | TTTGAGCTCTGTTTTAATGATTG | Amplification of lpg2186 gene fragment. |
| 2186 Seq | CCTACACACCCATTGGAATAGC | Confirmation of genomic integration of the inactivated allele. |
| 2225XBA1 | TTTTCTAGAGCCAGCTAATAATTTCTCCTTTG | Amplification of lpg2225 gene fragment. |
| 2225KPN1 | TTTGGTACCGACTAACAGCGTGGGTGGAA | Amplification of lpg2225 gene fragment. |
| 2225SPH1 | TTTGCATGCTCTACAGGAACTGTTAACCA | Amplification of lpg2225 gene fragment. |
| 2225SAC1 | TTTGAGCTCGTTGTATCTTGTCGCGGTAG | Amplification of lpg2225 gene fragment. |
| 2225 Seq | GTCACTCAGTCGATATGTTG | Confirmation of genomic integration of the inactivated allele. |
| 2228XBA1 | TTTTCTAGAATGAATATCTTAAAACCTAAC | Amplification of lpg2228 gene fragment. |
| 2228KPN1 | TTTGGTACCTTTCATTCTCAGGGTCAATC | Amplification of lpg2228 gene fragment. |
| 2228SPH1 | TTTGCATGCAAACCTGGATAGGTTTGGCTG | Amplification of lpg2228 gene fragment. |
| 2228SAC1 | TTTGAGCTCGCTATATGATGCTTTTTGATAG | Amplification of lpg2228 gene fragment. |
| 2228 Seq | CTGGCAGGATAGAAAGAGTG | Confirmation of genomic integration of the inactivated allele. |
| KanRev | GGTATTGATAATCCTGATATG | Confirmation of genomic integration of the inactivated allele. |
| KanNorm | GACGTTTCCCGTTGAATATGG | Confirmation of genomic integration of the inactivated allele. |

## Supplementary Note - Structure Elucidation

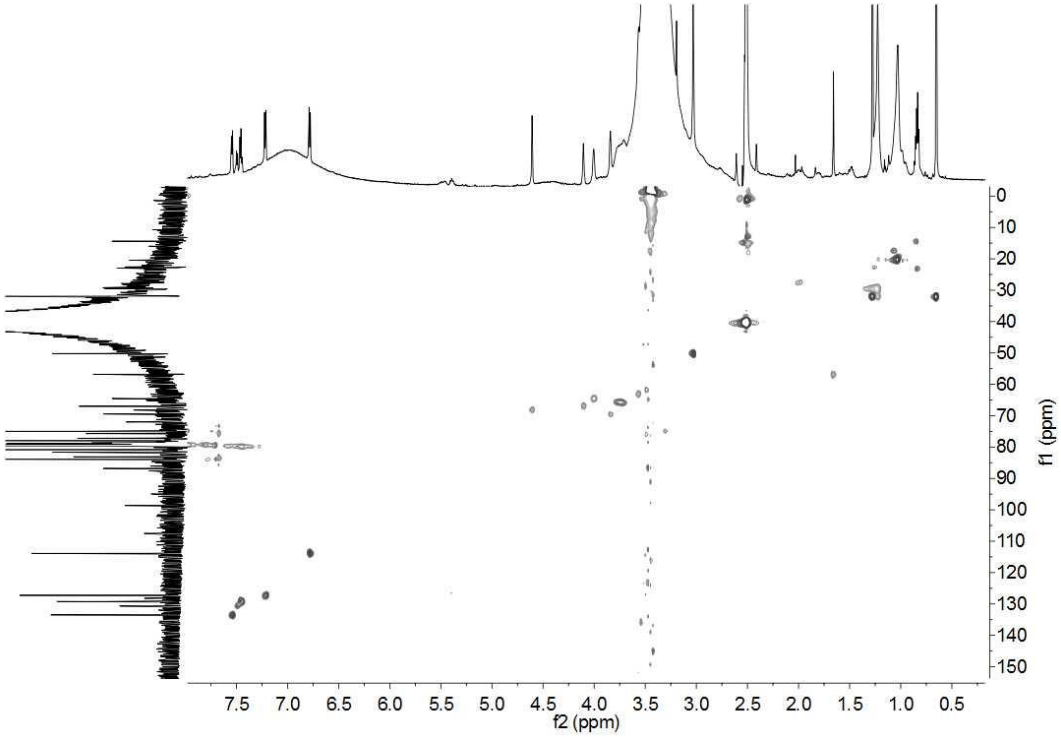
**Supplementary Figure 1** $^{1}$H NMR spectrum of legionellol A.



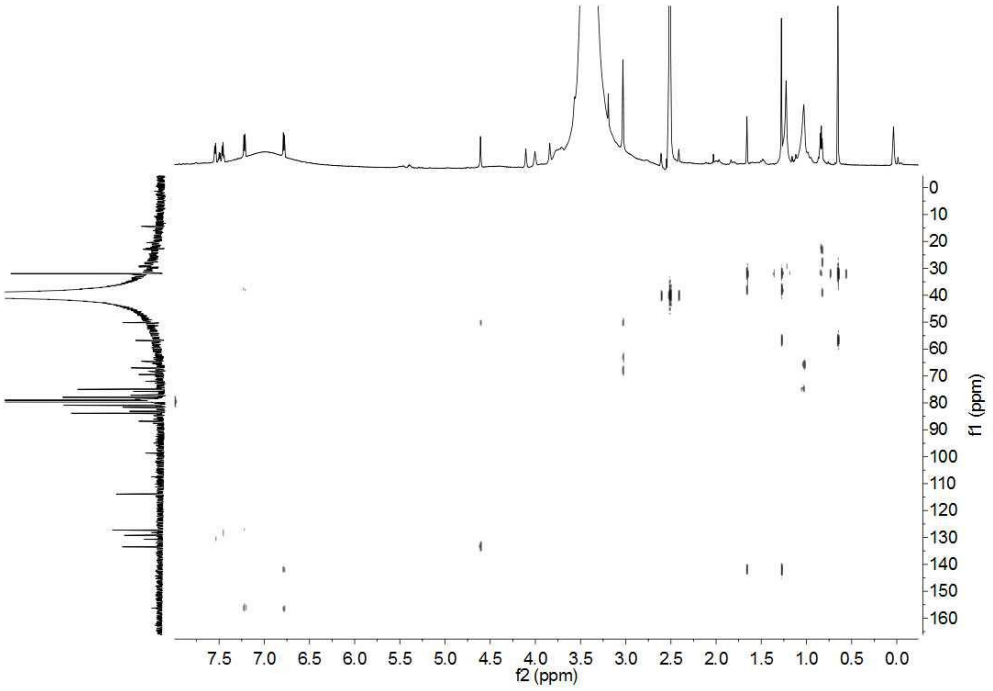

**Supplementary Figure 2** APT spectrum of legionellol A.



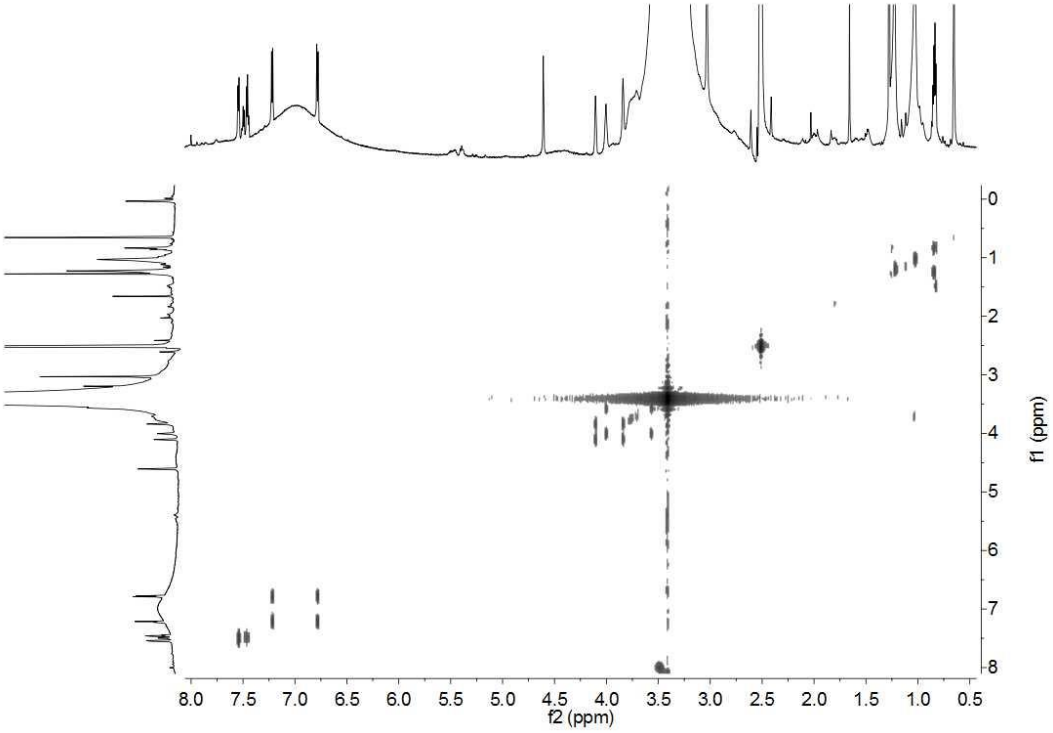
**Supplementary Figure 3** HMQC spectrum of legionellol A.



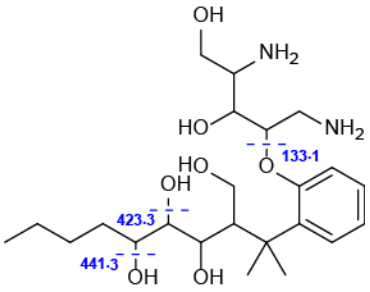

**Supplementary Figure 4** HMBC spectrum of legionellol A.



**Supplementary Figure 5** COSY spectrum of legionellol A




**Supplementary Figure 6** Important MS fragments of legionellol A




**Supplementary Figure 7** Structure of legionellol A deduced from NMR 1D and 2D experiments

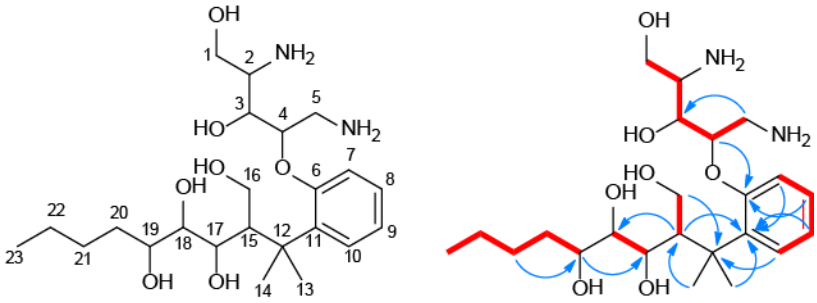


**Table 1** $^{1}H$ and $^{13}C$ NMR spectral data of legionellol in $d_6$-DMSO

| Position | $\delta_H$ mult. | $\delta_C$ | Position | $\delta_H$ mult. | $\delta_C$ |
|---|---|---|---|---|---|
| **1** | 3.48, *ov* | 61.7, *t* | **13** | 0.65, *s* | 31.9, *q* |
| **2** | 3.57, *ov* | 57.5*, *d* | **14** | 1.28, *s* | 32.0, *q* |
| **3** | 4.01, *m* | 64.5, *d* | **15** | 1.65, *m* | 56.8, *d* |
| **4** | 4.61, *m* | 68.2, *d* | **16** | 3.74, *m* | 65.6*, *t* |
| **5** | 3.02, *d*, 1.7 | 50.2, *t* | **17** | 4.10, *dd*, 2.8 | 66.9, *d* |
| **6** | - | 156.3, *s* | **18** | 3.84, *dd*, 3.2 | 69.5, *d* |
| **7** | 7.54, *d*, 8.5 | 133.5, *d* | **19** | 3.38, *m* | 75.0, *d* |
| **8** | 7.45, *t*, 2.1, 8.5 | 129.2, *d* | **20** | 1.23, *m* | 32.4, *t* |
| **9** | 6.78, *t*, 2.1, 8.5 | 113.9, *d* | **21** | 1.23, *m* | 29.5, *t* |
| **10** | 7.23, *d*, 8.5 | 127.2, *d* | **22** | 1.25, *m* | 22.6, *t* |
| **11** | - | 141.8*, *s* | **23** | 0.84, *t*, 3.6 | 14.3, *q* |
| **12** | - | 37.9*, s | | | |

*ov*, overlapped under the solvent peak. * signals extracted from HMBC and HMQC spectra.

## Appendix 4

## Supplementary Information

## Gold biomineralization by a metallophore from a gold-associated microbe.

Chad W. Johnston[1]∗, Morgan A. Wyatt[1]∗, Xiang Li[1], Ashraf Ibrahim[1], Jeremiah Shuster[2], Gordon Southam[2], & Nathan A. Magarvey[1]

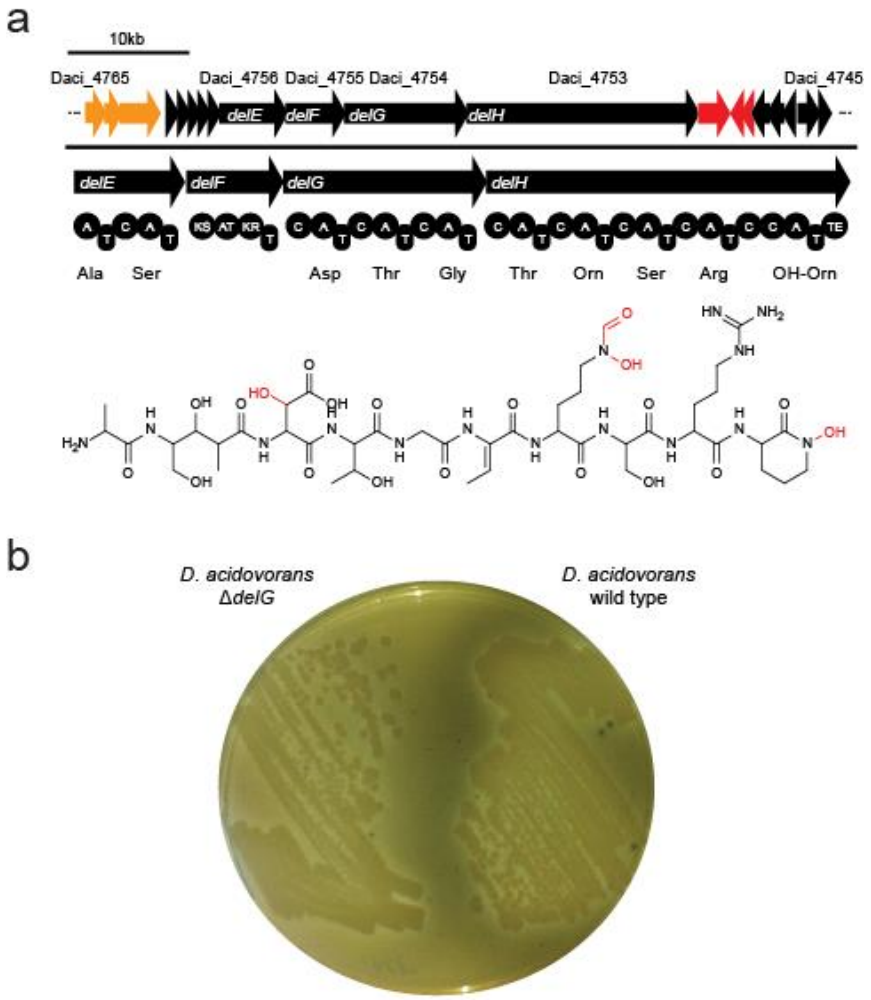
[1]Department of Biochemistry and Biomedical Sciences, Department of Chemistry and Chemical Biology, M.G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

[2]Department of Earth Sciences, Department of Biology, Western University, Ontario, London, Ontario N6A 5B7, Canada

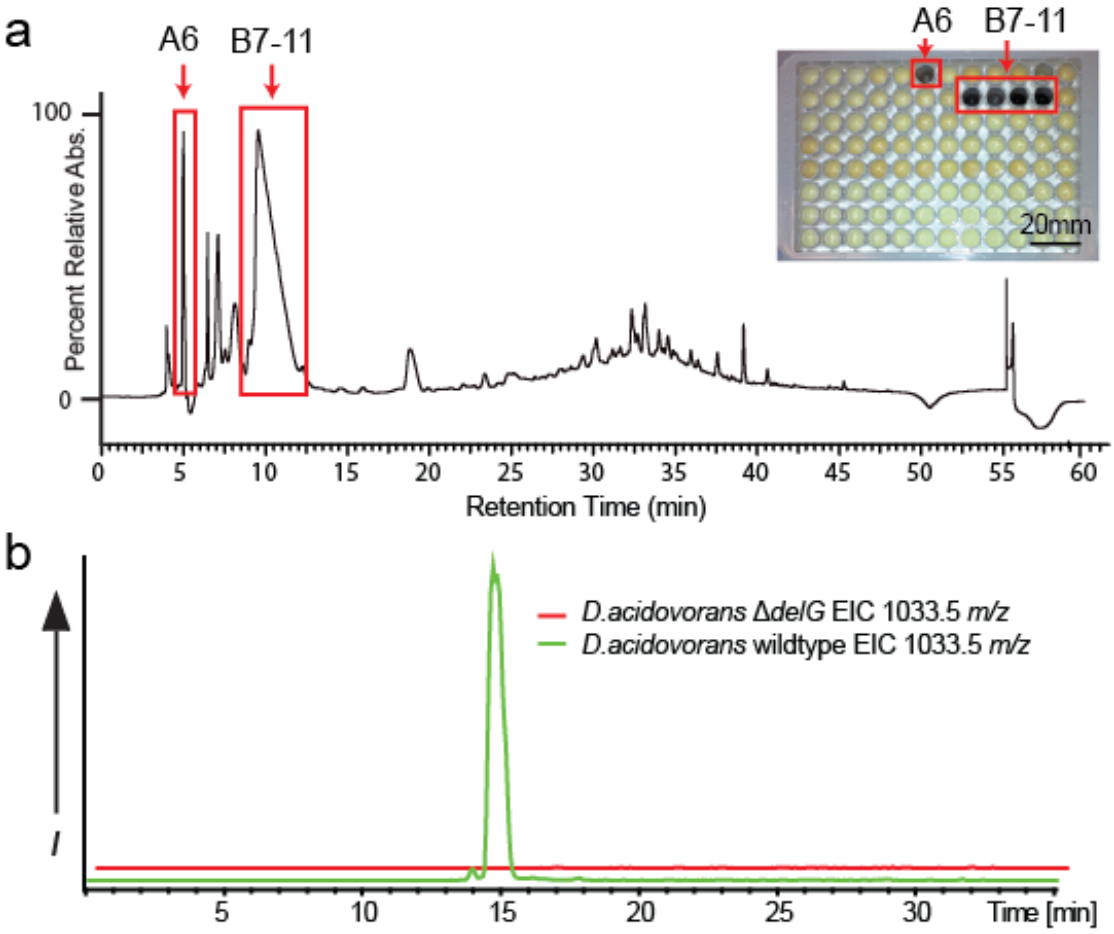
∗ These authors contributed equally to this work

## Supplementary Results

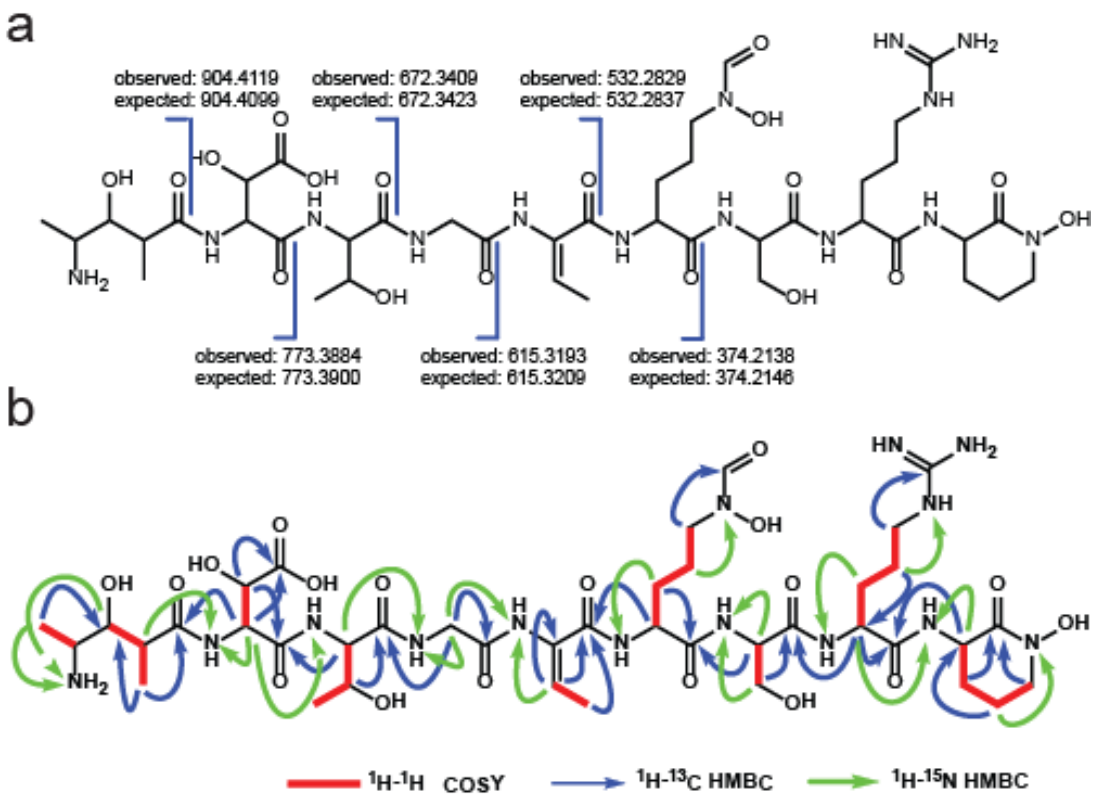
## Supplementary Figures



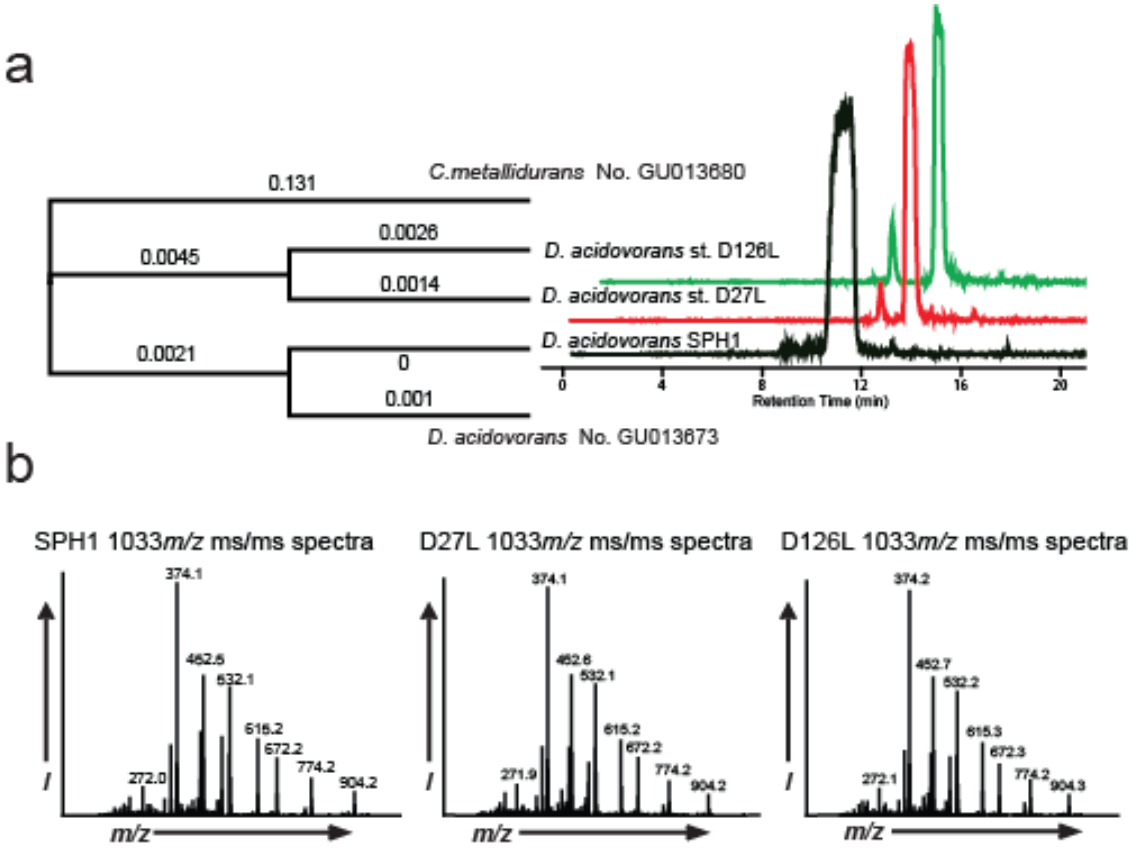

**Supplementary Figure 1**. ***Delftia acidovorans*** **possesses a PKS/NRPS gene cluster associated with extracellular gold precipitation.** a) *del* gene cluster and domain architecture of NRPS-PKS hybrid assembly-line is shown; consisting of adenylation (A), thiolation (T), condensation, (C), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR), and thioesterase domains. Flanking genes for heavy metal resistance (orange) and iron metabolism (red) are shown. Predicted activated amino acids are indicated below their respective A domains (see supplementary table 3). The final predicted structure of the unknown *del* metabolite is shown. b) *D. acidovorans* Δ*delG* does not produce the gold precipitate halo**.** *D.* acidovorans wildtype and Δ*delG* grown for 3 d at 30°C on ACM agar, overlaid with 0.5% agarose containing 10 mM $AuCl_3$ for 2 h.



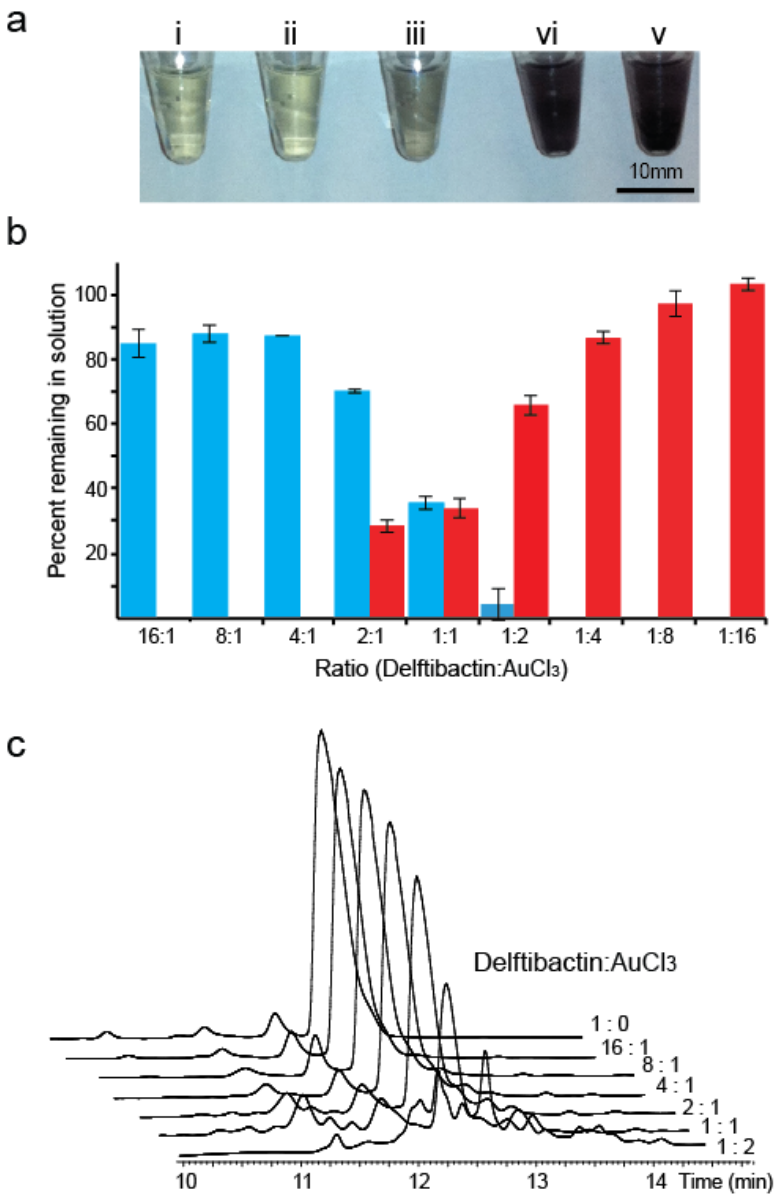

**Supplementary Figure 2. Gold precipitation is caused by a metabolite encoded by the *del* gene cluster.** a) Metabolites from *D. acidovorans* cultures were extracted, separated by HPLC into a 96-well plate, and reacted with 5 mM $AuCl_3$. Blackening indicates gold nanoparticle formation. Active wells with corresponding UV peaks are highlighted and were found to contain a common peptidic metabolite. b) Extracts of wildtype and Δ*delG* *D.acidovorans* analyzed by LCMS. The extracted ion chromatogram of the wildtype specific compound associated with gold precipitation is shown.



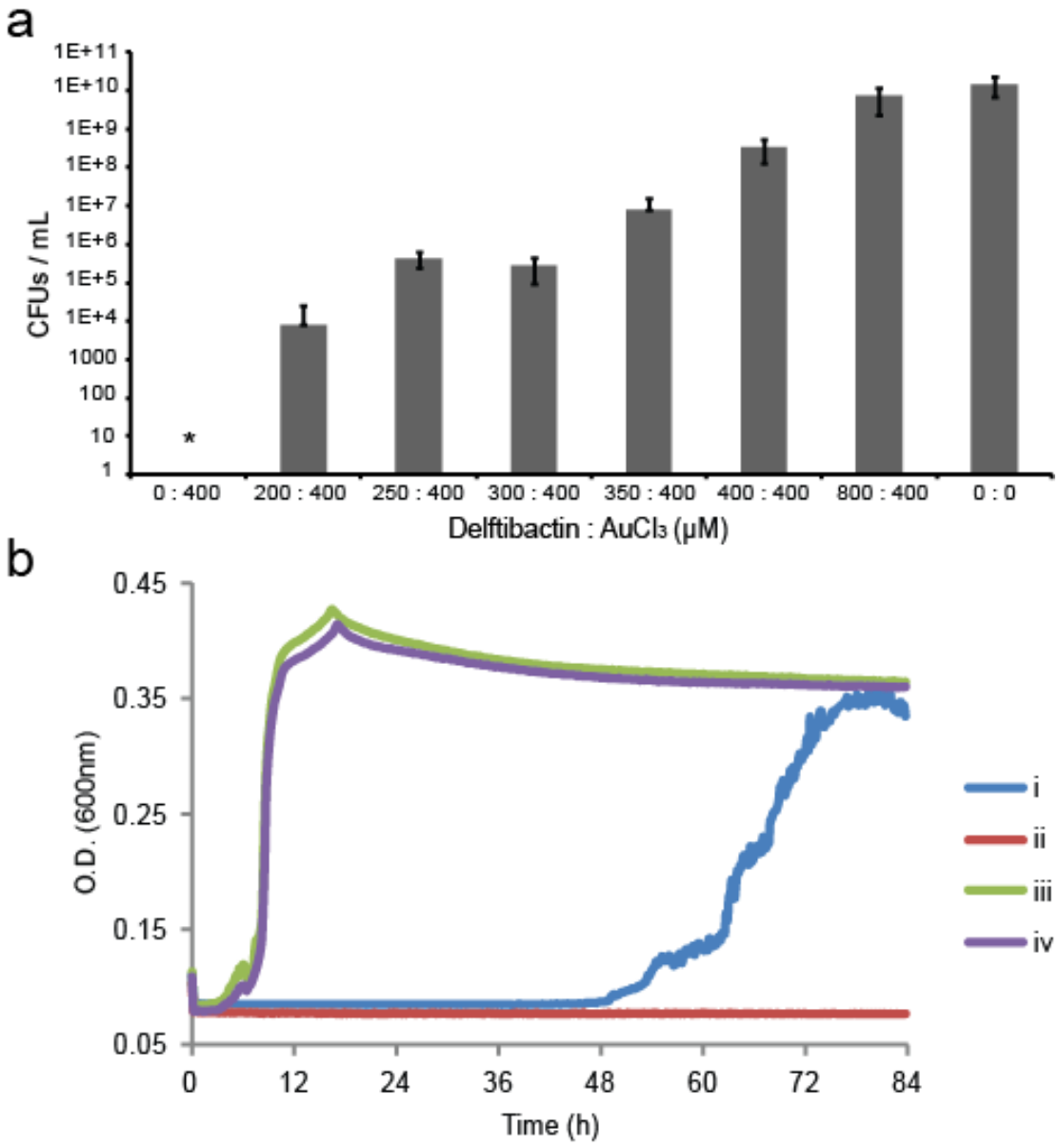

**Supplementary Figure 3. Structural characterization of delftibactin.** a) High resolution mass fragmentation and b) 2D NMR spin systems for $^{1}H$-$^{13}C$ HMBC, $^{1}H$-$^{1}H$ COSY and $^{1}H$-$^{15}N$ HMBC for delftibactin in $D_2O$.




**Supplementary Figure 4. Environmental isolates of *D. acidovorans* also produce delftibactin.** a) A 16S rDNA phylogenetic tree of *D. acidovorans* environmental isolates (D126L and D27L), genome strain (SPH-1), and sequences from samples associated with gold nugget biofilms, with respective extracted ion chromatogram ($m^+/z$ = 1033.5) from HP-20 extracts (see supplementary methods). b) Fragmentation pattern of $m^+/z$ = 1033.5 (delftibactin) from SPH-1, D27L, and D126L.




**Supplementary Figure 5 Delftibactin and $AuCl_3$ co-precipitate from solution.** a) Increasing concentrations of delftibactin in the presence of 2.5 mM $AuCl_3$ cause an increase in gold nanoparticle formation. Images were taken 30 min after the addition of delftibactin in the following concentrations: i) 0.3125 mM ii) 0.625 mM iii) 1.25 mM iv) 2.5 mM and 5 mM. b) Analysis of delftibactin-$AuCl_3$ reaction supernatants. Solutions of 2.5 mM $AuCl_3$ were reacted with 1:8, 1:4, 1:2, 1:1, and 2:1 equivalents of delftibactin for 30 minutes. Delftibactin remaining in solution was determined by integration at 220 nm (blue) by HPLC. The amount of gold remaining in solution was measured at 300 nm (red) by UV absorbance spectrometry. Results are shown as mean ± s.d; n = 3. c) Delftibactin depletion in representative HPLC chromatograms of delftibactin-$AuCl_3$ supernatants.

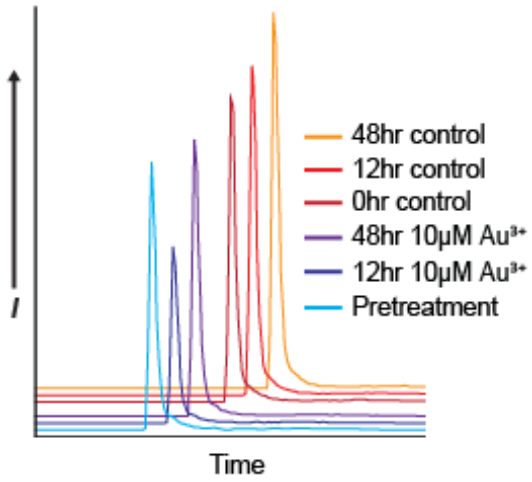



**Supplementary Figure 6. Delftibactin detoxifies $AuCl_3$ and enables growth under chronic gold exposure.** a) *Delftia acidovorans* colony forming units (CFUs) after exposure to 400 µM $AuCl_3$ for 30 minutes. Delftibactin was added to $AuCl_3$ solutions as indicated before exposing to *D. acidovorans.* *No survival was observed without the addition of delftibactin. Results are shown as mean ± s.d.; n = 4. b) Growth curves of *D. acidovorans* Δ*delG* cultures inoculated 1:1000 into ACM followed by the addition of delftibactin and/or gold as follows: i) 100 µM Delftibactin + 10 µM $AuCl_3$, ii) 10 µM $AuCl_3$ only, iii) 10 µM delftibactin only, and iv) water only. Results are a mean of three growth curves for each condition from a single representative experiment.

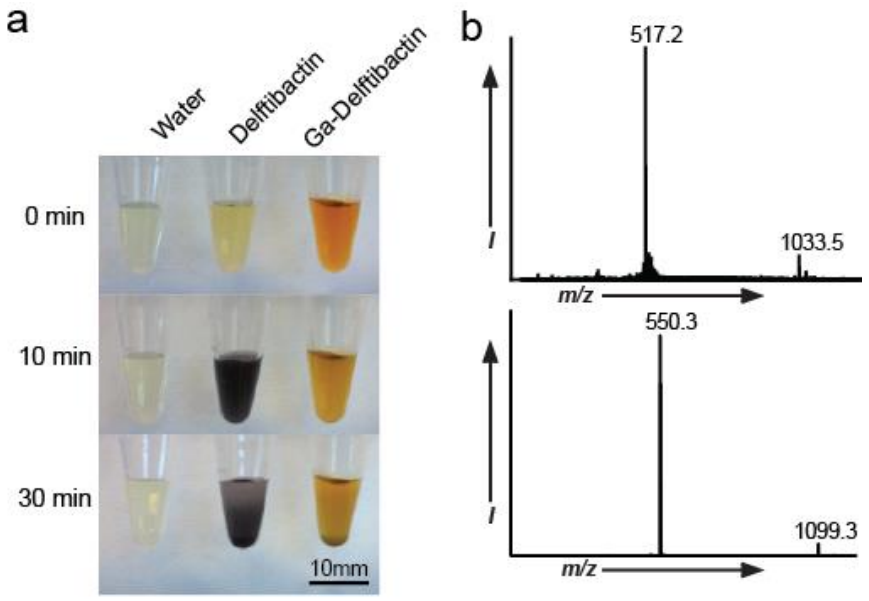
a

| [$Au^{3+}$] treatment | Normalized percent increase in [Delftibactin] following precipitation |
|---|---|
| 0 µM | -- |
| 10 µM | 65.6 ± 5.5% |
| 30 µM | 165.3 ± 5.1% |

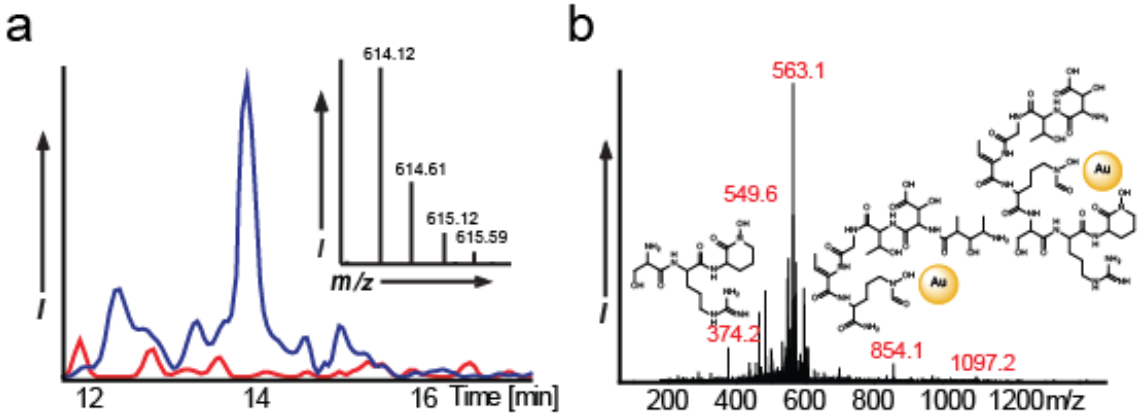
b



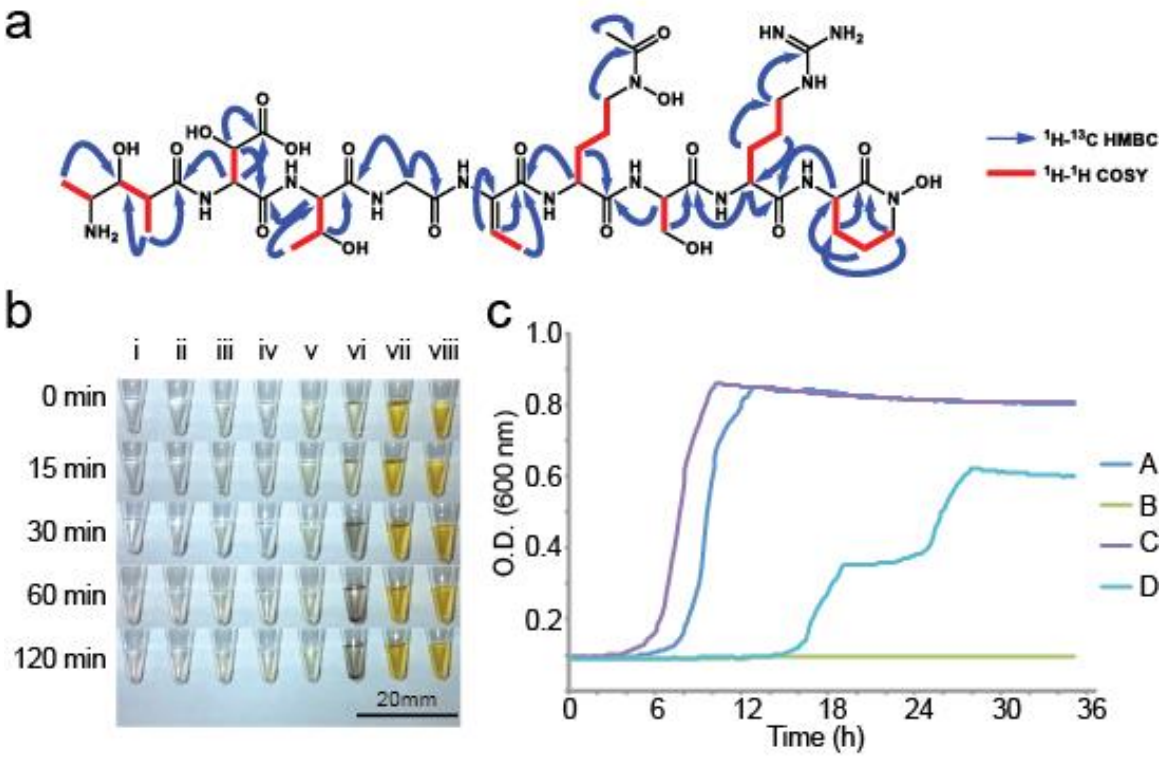

**Supplementary Figure 7. Delftibactin production is responsive to gold concentrations through reactive homeostasis.** a) MRM-LCMS quantification of delftibactin concentrations. Values represent the percent increase in delftibactin concentrations following precipitation by $AuCl_3$, normalized to an untreated control, ± propagated error; n = 6. b) Sample MRM-LCMS chromatogram of (a).



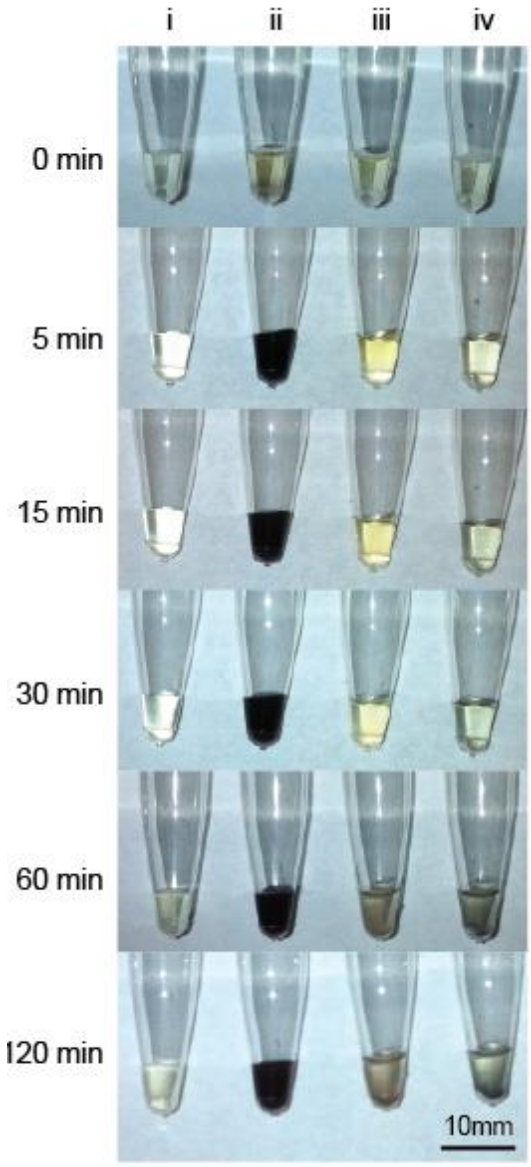

**Supplementary Figure 8. Delftibactin has a single metal binding site and occupancy by other metals impedes $AuCl_3$ precipitation.** a) Delftibactin (5 mM) and Ga-delftibactin (5 mM) reacted with equimolar concentrations of $AuCl_3$. b) Mass spectra of free (top) and gallium-bound delftibactin (bottom).



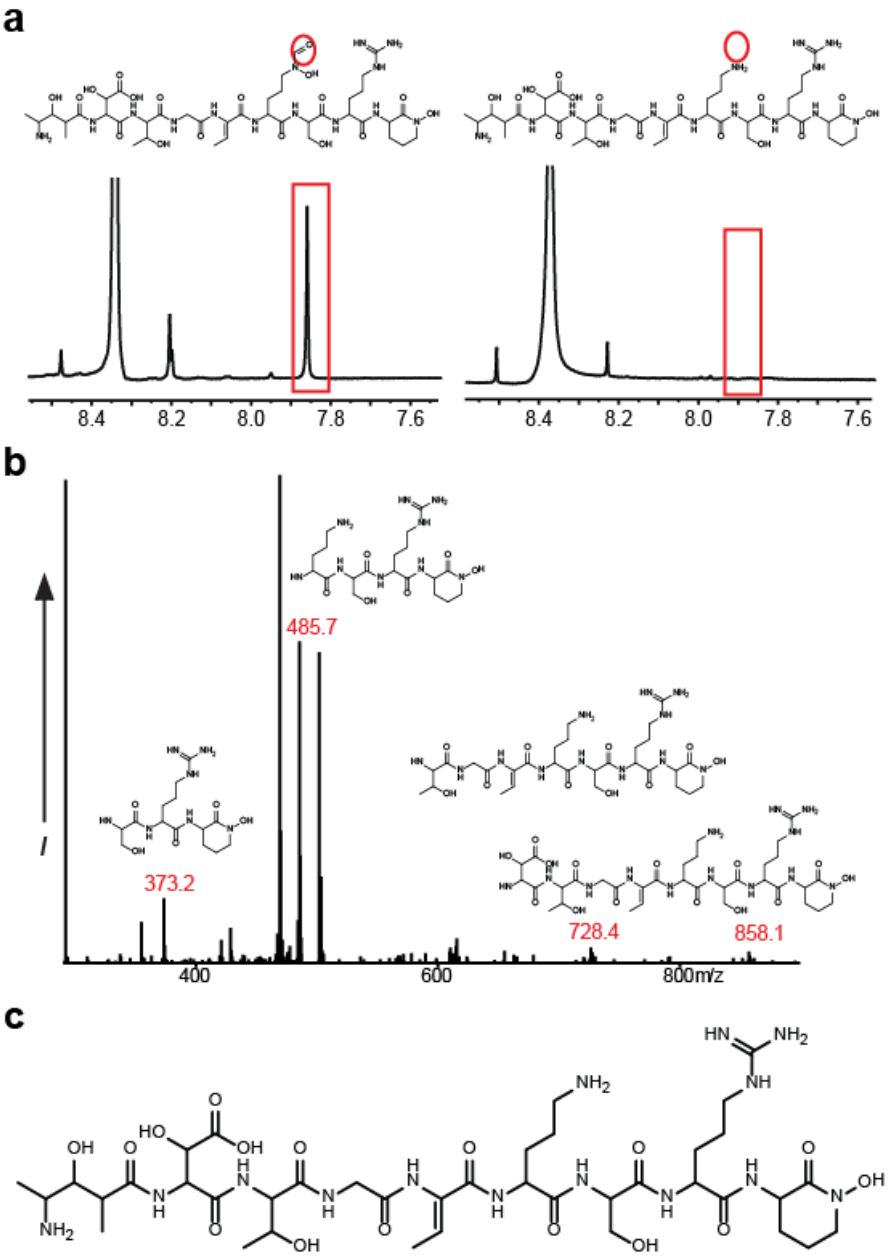

**Supplementary Figure 9. Mass spectral analysis of a gold-bound delftibactin species.** a) A mixture of 50 µM $AuCl_3$ and 100 µM delftibactin shows a double charged ion (blue) corresponding to a delftibactin-gold 1:1 complex not observed when gold is absent (red). b) The fragmentation pattern of the delftibactin-gold ion, showing the characteristic tripeptide loss and remaining delftibactin-gold complex (+374.2, +854.1 m/z), the double- and single-charged ions of the delftibactin-gold complex following the first amide cleavage (+549.6, +1097.2 m/z), and a double-charged delftibactin-gold ion following loss of one side chain (+563.1 m/z). Spectra averaged from a total of 96 scans.



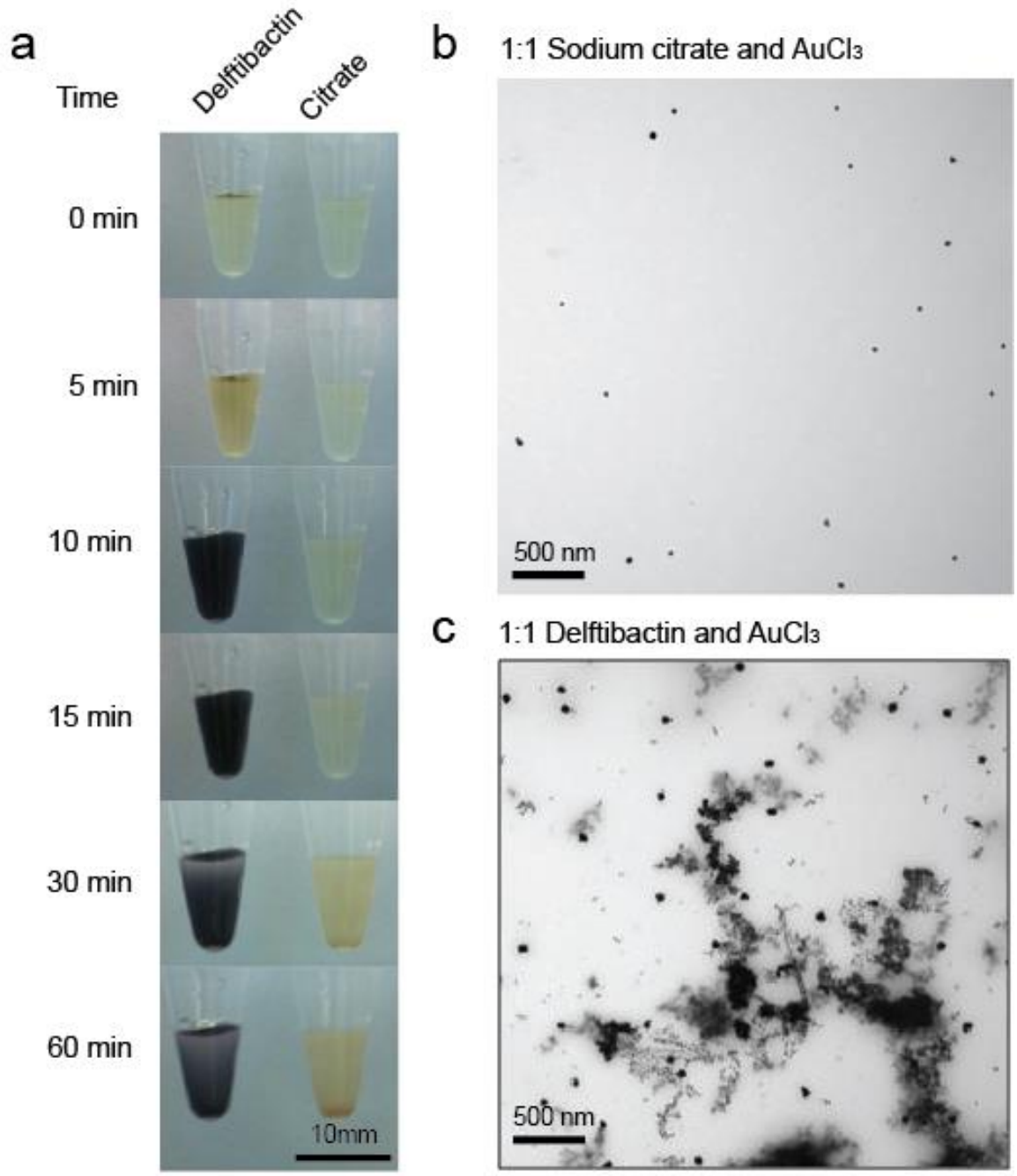

**Supplementary Figure 10. Delftibactin B (acetylated hydroxy-ornithine delftibactin A) is impaired in gold precipitation and is less protective against $AuCl_3$.** a) Structure of the acetylated analog and spin systems for $^1$H-$^{13}$C HMBC, $^1$H-$^1$H COSY and $^1$H-$^{15}$N HMBC in $D_2O$ for delftibactin B. b) $AuCl_3$ (5mM) precipitation by delftibactin B over time in the absence and presence of $FeCl_3$ as follows: i) water only ii) 5 mM delftibactin B iii) 5 mM $AuCl_3$ iv) 5 mM $FeCl_3$ v) 5 mM $AuCl_3$ + 5 mM $FeCl_3$ vi) 5 mM delftibactin B + 5 mM $AuCl_3$ vii) delftibactin B + 5 mM $AuCl_3$ + 5 mM $FeCl_3$ viii) delftibactin B + 5 mM $FeCl_3$. c) Delftibactin A (formylated) is more protective against $AuCl_3$ toxicity than delftibactin B (acetylated). Growth curves of *D. acidovorans* Δ*delG* in the presence of reaction mixture of water (A), 125 μM $AuCl_3$ (B), 125 μM $AuCl_3$ + 125 μM delftibactin A (C), and 125 μM $AuCl_3$ + 125 μM delftibactin B (D). Results are a mean of three growth curves for each condition from a single representative experiment.




**Supplementary Figure 11. All delftibactin species eventually co-precipitate with $AuCl_3$.** Timecourse of 5 mM $AuCl_3$ reacted with i) water ii) 5 mM delftibactin A iii) 5 mM delftibactin-$AuCl_3$ reaction product and iv) 5 mM delftibactin B were monitored over 2 h for $AuCl_3$ precipitation.




**Supplementary Figure 12. Determination of the delftibactin-$AuCl_3$ transient reaction product.** a) Aromatic region of 1D proton NMR of delftibactin A (left) and the reacted delftibactin intermediate (m+/z = 989) (right). The formyl hydrogen signal is absent in the m+/z = 989 delftibactin intermediate (red). b) Fragmentation of the double charged ion of the reacted delftibactin species (m+/z = 989) with diagnostic fragments shown. Mass loss of 44 is localized to ornithine functional group. c) Final structure of transient delftibactin-$AuCl_3$ reaction intermediate is shown. Structure and mass deviation is consistent with the loss of the formyl and hydroxyl group from ornithine.




**Supplementary Figure 13. Delftibactin rapidly forms complex precipitates of gold nanoparticles.** a) $AuCl_3$ (5 mM) was reacted with equimolar concentrations of delftibactin or sodium citrate. TEM images of sodium citrate-gold (b) and delftibactin-gold (c) reaction after 10 minutes.

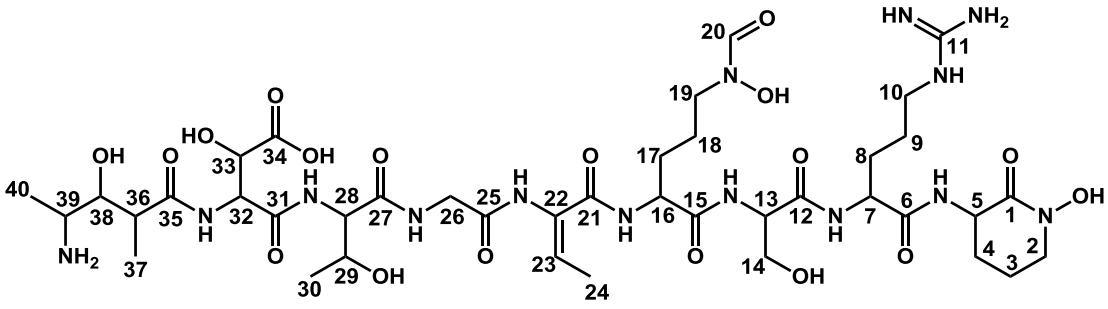
## Supplementary Tables

**Supplementary Table 1.** $^{1}H$ and $^{13}C$ NMR and NOESY spectral data of delftibactin A in $D_2O$[a,b]

| C/H | $^{1}H$ | $^{13}C$ | NOE | Ga | C/H | $^{1}H$ | $^{13}C$ | NOE | Ga |
|---|---|---|---|---|---|---|---|---|---|
| 1 | – | 165.9 (C) | – | | 19 | 3.49 | 49.4 (CH) | H-18 | 3.62 |
| 2a | 3.54 | 51.5 ($CH_2$) | H-3a | 3.67 | 20 | 7.86 | 159.1 (CH) | – | 8.14 |
| 2b | 3.59 | | H-3b | 3.70 | 21 | – | 166.4 (C) | – | |
| 3a | 1.87 | 19.4 ($CH_2$) | H-2a | 1.93 | 22 | – | 127.2 (CH) | – | |
| 3b | 1.92 | | H-2b | | 23 | 6.63 | 134.1 (CH) | H-24 | 6.62 |
| 4a | 1.69 | 26.0 ($CH_2$) | H-3b | 1.80 | 24 | 1.65 | 12.1 ($CH_3$) | H-23 | 1.66 |
| 4b | 1.94 | | H-5 | 2.01 | 25 | – | 172.4 (C) | – | |
| 5 | 4.36 | 49.8 (CH) | H-4b | 4.39 | 26 | 3.99 | 42.6 ($CH_2$) | – | 4.06 |
| 6 | – | 172.5 (C) | – | | 27 | – | 171.1 (C) | – | |
| 7 | 4.25 | 53.0 (CH) | H-8 | 4.22 | 28 | 4.32 | 58.5 (CH) | H-29 | 4.36 |
| 8 | 1.80 | 27.5 ($CH_2$) | H-7, H-9 | 1.76 | 29 | 4.32 | 66.1 (CH) | H-28, H-30 | 4.27 |
| 9 | 1.54 | 23.8 ($CH_2$) | H-8, H-10 | 1.54 | 30 | 1.13 | 18.3 ($CH_3$) | H-29 | 1.13 |
| 10 | 3.10 | 40.0 ($CH_2$) | H-9 | 3.12 | 31 | – | 171.0 (C) | – | |
| 11 | – | 156.2 (C) | – | | 32 | 4.68 | 55.8 (CH) | H-33 | 4.68 |
| 12 | – | 171.3 (C) | – | | 33 | 4.21 | 71.8 (CH) | H-32 | 4.24 |
| 13 | 4.32 | 55.3 (C) | H-14a | 4.37 | 34 | – | 176.4 (C) | – | |
| 14a | 3.76 | 60.3 ($CH_2$) | H-13 | 3.80 | 35 | – | 175.6 (C) | – | |
| 14b | 3.78 | | – | | 36 | 2.54 | 42.5 (CH) | H-37 | 2.67 |
| 15 | – | 173.5 (C) | – | | 37 | 1.18 | 10.3 ($CH_3$) | H-36, H-37 | 1.19 |
| 16 | 4.29 | 53.5 (CH) | H-17a | 4.24 | 38 | 3.79 | 71.5 (CH) | H-37, H-40 | 3.89 |
| 17a | 1.67 | 26.8 ($CH_2$) | H-16 | 1.78 | 39 | 3.33 | 48.6 (CH) | – | 3.44 |
| 17b | 1.79 | | – | | 40 | 1.17 | 14.1 ($CH_3$) | H-38 | 1.18 |
| 18a | 1.63 | 22.3 ($CH_2$) | H-17a | 1.67 | | | | | |
| 18b | 1.65 | | – | | | | | | |

[a] Chemical shift $\delta$ and (multiplicity, $J$ in Hz).

[b] Proton chemical shift changes after gallium binding are highlighted in yellow.



**Supplementary Table 2.** $^{1}H$ and $^{13}C$ NMR spectral data of delftibactin B in $D_2O$[a]

| C/H | $^{1}H$ NMR | $^{13}C$ NMR | C/H | $^{1}H$ NMR | $^{13}C$ NMR |
|---|---|---|---|---|---|
| 1 | – | 164.9 (C) | 19 | 3.49 | 49.4 (CH) |
| 2a | 3.54 | 51.7 ($CH_2$) | 20 | 7.86 | 160.0 (CH) |
| 2b | 3.59 | | 21 | – | 166.4 (C) |
| 3a | 1.87 | 19.6 ($CH_2$) | 22 | – | 127.2 (CH) |
| 3b | 1.92 | | 23 | 6.62 | 134.7 (CH) |
| 4a | 1.69 | 25.8 ($CH_2$) | 24 | 1.65 | 12.1 ($CH_3$) |
| 4b | 1.94 | | 25 | – | 172.4 (C) |
| 5 | 4.35 | 49.7 (CH) | 26 | 3.99 | 42.6 ($CH_2$) |
| 6 | – | 172.1 (C) | 27 | – | 171.1 (C) |
| 7 | 4.27 | 53.3 (CH) | 28 | 4.35 | 58.5 (CH) |
| 8 | 1.80 | 27.1 ($CH_2$) | 29 | 4.35 | 66.1 (CH) |
| 9 | 1.54 | 24.1 ($CH_2$) | 30 | 1.12 | 18.2 ($CH_3$) |
| 10 | 3.12 | 40.1 ($CH_2$) | 31 | – | 171.3 (C) |
| 11 | – | 156.7 (C) | 32 | 4.68 | 55.8 (CH) |
| 12 | – | 171.1 (C) | 33 | 4.21 | 71.8 (CH) |
| 13 | 4.31 | 55.1 (C) | 34 | – | 176.1 (C) |
| 14a | 3.76 | 60.0 ($CH_2$) | 35 | – | 175.6 (C) |
| 14b | 3.77 | | 36 | 2.56 | 42.1 (CH) |
| 15 | – | 173.5 (C) | 37 | 1.16 | 10.3 ($CH_3$) |
| 16 | 4.27 | 53.7 (CH) | 38 | 3.76 | 71.6 (CH) |
| 17a | 1.59 | 26.4 ($CH_2$) | 39 | 3.34 | 48.6 (CH) |
| 17b | 1.73 | | 40 | 1.17 | 14.1 ($CH_3$) |
| 18a | 1.66 | 22.1 ($CH_2$) | | | |
| 18b | 1.63 | | | | |

[a] Chemical shift $\delta$ and (multiplicity, $J$ in Hz).

**Supplementary Table 3.** Adenylation domain specificities**.** Prediction of incorporated amino acids was performed using NRPSPredictor and NRPS PKS, providing a probability sequence for the NRPS/PKS product.

| Adenylation Domain | Active Site Residues | Substrate | Product |
|---|---|---|---|
| DelE A1 | DMGGYGCLFK<br>DAGGCAMVAK | Alanine<br>Alanine | HC Toxin |
| DelE A2 | DIWHISLIEK<br>DVWHISLIDK | Inactive<br>Serine | Nostopeptolide |
| DelG A1 | DLTKVGHVGK<br>DLTKVGHIGK | Aspartic acid<br>Aspartic acid | Surfactin |
| DelG A2 | DFWNIGMVHK<br>DFWNIGMVHK | Threonine<br>Threonine | Syringopeptin |
| DelG A3 | DILQLGLIWK<br>DILQLGLIWK | Glycine<br>Glycine | Nostopeptolide |
| DelH A1 | DFWNIGMVHK<br>DFWNIGMVHK | Threonine<br>Threonine | Syringopeptin |
| DelH A2 | DVWNIGLIHK<br>DVGEIGSIDK | Ornithine<br>Ornithine | Fengycin |
| DelH A3 | DVWHLSLIDK<br>DVWHLSLIDK | Serine<br>Serine | Syringopeptin |
| DelH A4 | DGEDHGAVTK<br>DAEDIGAITK | Arginine<br>Arginine | Pederin |
| DelH A5 | DGEAVGGVTK<br>DGESSGGMTK | $N^{\delta}$-hydroxyornithine<br>$N^{\delta}$-hydroxyornithine | Vicibactin |

**Supplementary Table 4. Delftibactin gene cluster analysis**

| Locus | Gene | Predicted Function | Strand | Amino Acids |
|---|---|---|---|---|
| Daci_4765 | - | Heavy metal efflux outer membrane component | - | 485 |
| Daci_4764 | - | Heavy metal efflux periplasmic component | - | 405 |
| Daci_4763 | - | Heavy metal efflux inner membrane component | - | 1029 |
| Daci_4762 | - | Nitrogen regulatory protein P-II | - | 111 |
| Daci_4761 | - | RNA polymerase, sigma subunit | - | 174 |
| Daci_4760 | *delA* | MbtH domain protein | - | 121 |
| Daci_4759 | *delB* | Thioesterase | - | 247 |
| Daci_4758 | *delC* | Phosphopantetheinyl transferase | - | 229 |
| Daci_4757 | *delD* | Aspartic acid dioxygenase | - | 329 |
| Daci_4756 | *delE* | Nonribosomal peptide synthetase | - | 1789 |
| Daci_4755 | *delF* | Polyketide synthase | - | 1560 |
| Daci_4754 | *delG* | Nonribosomal peptide synthetase | - | 3331 |
| Daci_4753 | *delH* | Nonribosomal peptide synthetase | - | 6176 |
| Daci_4752 | *delI* | Siderophore receptor | - | 799 |
| Daci_4751 | *delJ* | anti-FecI sigma factor, FecR | + | 344 |
| Daci_4750 | *delK* | RNA polymerase, sigma subunit | + | 199 |
| Daci_4749 | *delL* | Lysine/Ornithine N-monooxygenase | + | 432 |
| Daci_4748 | *delM* | Acetyltransferase | + | 400 |
| Daci_4747 | *delN* | Esterase/Lipase | + | 321 |
| Daci_4746 | *delO* | Siderophore Export Pump | - | 571 |
| Daci_4745 | *delP* | N5-hydroxyornithine formyltransferase | - | 284 |

**Supplementary Table 5. Primers used in construction of the Δ*delG* strain.**

| PCR Primer | Sequence (5' – 3') | Purpose |
|---|---|---|
| TetNotF | TTT TGC GGC CGC TGC TGA ACC CCC AA | Amplification of tetracycline resistance cassette from pLLX13. |
| TetNotR | TTT TGC GGC CGC TAT CGT TTC CAC GA | Amplifcation of the tetracycline resistance cassette from pLLX13. |
| 2kbNRPS2Xba2 | TTT TTC TAG ACG CAT TGC TGA ACT ACC | Amplification of the 2kb *delG* fragment from *D.acidovorans*. |
| 2kbNRPS2Sac2 | TTT TGA GCT CAG CAG TTG CAC CAC CT | Amplification of the 2kb *delG* fragment from *D.acidovorans*. |
| OriTF | TTT TAA GCT TTT CCT CAA TCG CTC TTC | Amplification of an oriT from pLLX13. |
| OriTR | TTT TAA GCT TTT TTC GCA CGA TAT ACA | Amplification of an oriT from pLLX13. |
| NRPS2Seq2 | GGG GTG CGG AAA ATG TCC TG | Confirmation of genomic integration of the tetracycline resistance cassette. |

**Supplementary Table 6. Delftibactin production in response to [$Fe^{3+}$]**

| Media [Iron] | [Delftibactin] |
|---|---|
| 0 | 205.8 ± 28.0 μM |
| 100 nM | 196.1 ± 30.7 μM |
| 1 μM | 149.3 ± 21.0 μM |
| 10 μM | 21 ± 9.4 μM |
| 100 μM | 2.2 ± 1.4 μM |
| 1 mM | Growth Inhibitory |

**Supplementary Note 1**

Production, isolation, purification, and structure determination of delftibactin A & B.

**Culture and Isolation**

A colony from a fresh plate of *D. acidovorans* was inoculated into a 2.8 L glass Fernbach flask containing 1 L of Acidovorax Complete Media[19] (ACM) that had been treated with 4

grams of Chelex-100 resin (*Sigma*). Cultures were grown at 30°C, shaking at 190 rpm for three days, after which cells were pelleted by centrifugation at 7000 rpm for 15 min. HP20 resin (*Dialon*) was added to the supernatant at 20 g/L and shaken for ~2 h at 220 rpm. The resin was harvested by Buchner funnel filtration and washed with 400 mL of distilled water. The resin was washed with 400 mL of methanol. The methanol eluent was evaporated to dryness under rotary vacuum and resuspended in 2 mL of 50:50 dd$H_2O$:MeOH. Delftibactin A and B were purified using a Waters Alliance 2695 RP-HPLC separations module, equipped with a Waters 2998 photodiode array and a Luna 5 μm $C_{18}$ column (250 x 10.0 mm, *Phenomenex*). The mobile phase was linear from 2 % acetonitrile, 98 % water + 5 mM $(NH_4)_2CO_3$ at 2 minutes to 14 % acetonitrile at 18 min at a flow rate of 3 mL/min. Delftibactin A eluted at 14.20 min and delftibactin B eluted at 17.5 min.

**High Resolution Mass Spectra**

A stock solution of 20 mg/ml of delftibactin was diluted to a final concentration of 10 μg/ml in water with 0.1% formic acid. This solution was directly infused at a rate of ~3 μL per min into a Thermo Finnigan LTQ OrbiTrap XL mass spectrometer running Xcaliber 2.07 and TunePlus 2.4 SP1. High resolution MS was acquired using an electrospray ionization source and fragmentation was obtained through collision induced dissociation (CID). The instrument was operated in the positive mode using a maximum resolution of 100, 000. Data was acquired for approximately 1 min for a total of 32 scans. Bradykinin was used as an internal standard, and was premixed with delftibactin to a final concentration of 5 μg/ml. The lock mass feature was applied using the bradykinin standard at [M+H] = 1060.56922 *m/z*.

| ***Compound*** | ***Calculated m/z*** | ***Observed m/z*** | ***Δppm*** |
|---|---|---|---|
| Delftibactin [M+H] | 1033.49143 | 1033.49154 | 0.106 ppm |

**NMR Methods and Structural Characterization**

NMR spectra were measured on a Bruker Avance 700 spectrometer equipped with a 5 mm inverse detection probe and using TMS as an internal standard. Lyophilized samples were dissolved in $D_2O$ and spectra were recorded at 297 K. NMR experiments were processed and analyzed with Bruker TOPSPIN 2.1. Chemical shifts ($\delta$) expressed in parts per million (ppm) and coupling constants ($J$) are reported in Hertz (Hz). Assembly of individual amino acids to form the final linear structure was accomplished by considering long-range $^{1}H$-$^{13}C$ and $^{1}H$-$^{15}N$ HMBC correlations from both protons adjacent carbonyl carbons and nitrogens, as well as by assignments of 2D $^{1}H$-$^{1}H$ COSY and 2D $^{1}H$-$^{13}C$ HSQC correlations.

**Delftibactin A**

Comprehensive analysis of 2D NMR data, including the results of $^{1}H$-$^{1}H$ COSY, HSQC, and HMBC experiments have been used to elucidate the planar structure of delftibactin A, and the chemical shifts from these experiments are provided in Supplementary Table 1. The molecular formula of delftibactin was established as $C_{40}H_{68}N_{14}O_{18}$ based on positive HR-ESI-MS *m/z*: 1033.4915 $[M+H]^+$ (calculated 1033.4914) indicating 15 degrees of unsaturation. $^{13}C$ NMR (DEPT) and gHMBC spectra revealed ten amide carbons at 156.2, 165.9, 166.4, 171.0, 171.1, 171.3, 172.4, 172.4, 173.5 and 175.6 ppm along with a formyl singlet at 7.86 ppm ($^{13}C$ NMR, 159.1 ppm). Analysis of COSY cross peaks gave eight spin systems. On the basis of long range $^{1}H$-$^{13}C$, $^{1}H$-$^{15}N$ and COSY interactions and high resolution MS-MS fragmentation, we link the structure shown in Fig S3.

Ga-delftibactin A was obtained by reacting 1 mg/mL of delftibactin (20 mg in total) in a 100 mL flask with ~5-fold excess of solid $GaBr_3$, added slowly over ~5 min and stirred gently overnight at room temperature. Ga-delftiabactin A was purified using a C18 column (1.6 g, 20 x 0.5 cm), and eluted with 0 to 5% MeOH aqueous solution over 30 min. This procedure provided ~13 mg Ga-delftibactin.

**Delftibactin B**

Comprehensive analysis of 2D NMR data, including the results of $^{1}H$-$^{1}H$ COSY, HSQC, and HMBC experiments have been used to elucidate the planar structure of delftibactin B. The chemical shifts of the protons and carbons (**Supplementary Table 2**) of delftibactin B ($m^{+}/z$ = 1047.4) were similar to those of delftibactin A, the main differences between the two metabolites concerned a newly appeared acetyl group [$\delta_C$ 17.2, $\delta_H$ 2.06 (s)] in compound 2. The location of the methoxyl group was established taking into account the correlation observed between acetyl group $\delta_H$ 2.06 (s) and C-20 (δ160.0) in the HMBC experiment of 2.