A System-Wide Stable Isotope Labeling Approach for Connecting Natural Products to Their Biosynthetic Gene Clusters

by

Catherine S. McCaughey

BSc (Marine Biology), University of Rhode Island, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Chemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Catherine S. McCaughey**
Degree:	**Doctor of Philosophy**
Title:	**A System-Wide Stable Isotope Labelling Approach for Connecting Natural Products to Their Biosynthetic Gene Clusters**
Committee:	**Chair:** Hua-Zhong Yu Professor, Chemistry **Roger G. Linington** Supervisor Professor, Chemistry **David Vocadlo** Committee Member Professor, Chemistry **Erika Plettner** Committee Member Professor, Chemistry **Robert Young** Examiner Professor Department of Chemistry **Katherine Ryan** External Examiner Associate Professor Department of Chemistry University of British Columbia

Abstract

Although the first bacterial genome sequence was published almost 20 years ago, there is still no generalizable method for automatically assigning natural products to their cognate biosynthetic gene clusters (BGCs). This thesis describes the development of a mass spectrometry-based parallel stable isotope labeling (SIL) platform, termed IsoAnalyst, which automatically associates metabolite stable isotope labeling patterns with BGC structure prediction in order to connect natural products to their cognate BGCs. The parallel SIL experiments were optimized for small scale and a custom tool written in Python was developed for the untargeted detection and interpretation of SIL labeling patterns. This approach was validated in the industrial production strains *Saccharopolyspora erythraea* and *Amycolatopsis mediterranei* demonstrating that the compounds erythromycin A and rifamycin SV respectively, could be associated with the proper BGCs based on the distribution of isotopomer labeling patterns. The method was further validated by connecting known biosynthetic intermediates of these compounds to their associated BGCs and the identification of various siderophores through a combination of SIL labeling patterns and MS/MS fragmentation data. Extension to environmental organisms using a sequenced *Micromonospora sp.* from our Actinobacterial isolate library led to the discovery of lobosamide D, a new member of the lobosamide family of natural products, and an update to the lobosamide BGC to include relevant tailoring enzymes. This discovery illustrates the power of the IsoAnalyst platform for identifying new compounds, linking molecules to BGCs, and generating new knowledge about biosynthesis.

Keywords: Metabolomics; Mass Spectrometry; Stable Isotopes; Biosynthetic Gene Clusters; Natural Products

Dedication

This thesis is dedicated to every autistic person in academia suffering under florescent lights and open office plans out of dedication to our special interests.

We deserve to be here.

Acknowledgements

I respectfully acknowledge that the work presented in this thesis was completed in the unceded territory of the Squamish, Tsleil-Waututh, Kwikwetlem and Musqueam peoples. I am grateful to them for their stewardship of these lands, which I have been blessed to live and work on for the past six years.

I am profoundly indebted to Dr. Kenji Kurita, who dedicated endless hours in the trenches to training me to be a better scientist, leader, and human being. I am grateful to him for being an exceptional role model in and out of the lab and for his enthusiasm while working through all manner of difficult problems with me. I would like to express my sincere gratitude to my supervisor Dr. Roger Linington whose 30,000 ft point of view inspires us all to shoot for the stars. At the heart of this thesis there are many ideas he has put out into the lab, waiting for someone brave or foolish enough to try them, and I am grateful that I have had the opportunity to challenge myself in this way. Dr. Erika Plettner and Dr. David Vocadlo have supported me with enthusiasm and consistently challenged me to look at different perspectives and improve the overall quality of this thesis. I can't imagine having a more supportive team advising me through this program. Dr. Eric Ye provided essential guidance in the final steps of solving my new compound structure for which I am very grateful. I'd also like to thank Jamie and everyone in the machine shop for their work on the well plate covers and clamps.

I must acknowledge Jake Haeckl and Jamie Yin who joined myself, Roger, and Kenji on the exciting and unpredictable journey of moving a laboratory across the border. This experience goes beyond anything I would have expected to learn in an academic program and I think it left us all feeling more like family than colleagues. I also acknowledge Dr. Hamel Tailor who greeted us at science receiving on day one and has been and indispensable friend, resource, and advocate throughout our transition and during my entire time at SFU.

Beyond my academic achievements I am continually inspired by the women I have had the pleasure to meet and work with in this department. First and foremost, I would be no where without the continual support and creative escapism I shared with Yumeela throughout my time here. I am grateful to Nas, who worked tirelessly by my side on caucus and to start the momentum for the Women in Chemistry group. I would also like to express my gratitude to Dr. Corina Andreoiu for uplifting us and facilitating our efforts to come together as women in chemistry. To Claire, who has been a continual source of support and humor during the hardest times of my progress in this program. To Julienne and Dasha, who were a delight to know and teach and who contributed to this work during their summer research with me in 2019 and 2020. And to the many women who are daily motivation for me to live to my fullest potential including Sherry Young, Mercia Valentine, Nicole LeGrow, Saeideh Shamsi, Rhea Bains, and Jackie von Salm. Outside of SFU, I am especially grateful for my dear friend, Christina Palka, who started on day one with me at UC Santa Cruz, scared, confused, awed, and dead certain that we were both imposters. I never would have made it through that first semester without her friendship, dedication, and co-localized terror.

There are so many Linington lab members who have taught me, pushed me, worked side by side with me through difficult problems, and celebrated successes year after year. Some of them have already been named but I would like to additionally acknowledge Jessie Ochoa, Darryl Wilson, Trevor Clark, Sanghoon Lee, Fausto Neto-Carnavale, Mirelle Takaki, Dennis Liu, Joe Egan, Yuniel Rodriguez Rivero, Michael Recchia, Gregoire Jacob, Laura Sanchez, Bob James, and Weng Ru Wong. A special thanks to Jeff van Stanten for working through my code, turning it into something functional, and always teaching me with patience and genuine encouragement.

On a personal note I would like to thank my father for always believing fully in my abilities but especially for tutoring me in chemistry before I even knew I had any interest in it. I may have taken a very different path without those hours of problem solving and discovering my passion for this field. To my mother for her endless support, spiritual insight, and for being my biggest fan, thank you for handling me with steadiness. I am also deeply grateful to my grandmother who poured her entire life into uplifting me for 20 years. I have elevated to this level on the energy she gave selflessly to me throughout my youth.

Lastly I thank Jacob Harris who has been by my side for every single personal and professional step in this journey, from the day I arrived in Santa Cruz, to the day I asked him to move to Canada with me, to navigating life in a global pandemic. On a satellite phone from the middle of the Pacific Ocean, or outside of TASC 2 bringing me a lunch, he never missed a beat. I wouldn’t rather do this with anyone else by my side.

Contributions

Parts of this thesis have been adapted from the following manuscript:

McCaughey, Catherine S., van Santen, Jeffrey A., van der Hooft, Justin J.J., Medema, Marnix H., and Linington, Roger G. IsoAnalyst: A System-wide Stable Isotopic Labeling Approach for Connecting Natural Products to Their Cognate Biosynthetic Gene Clusters. *Nature Chemical Biology*. (In Review) (2021)

Authors from this manuscript contributed to the work presented in this thesis. I designed the data processing and statistical analysis steps that make up the core steps of IsoAnalyst, and wrote code using Python 3 to implement these data analysis steps, as described in Chapter 3. Jeffery van Santen developed and optimized the Python 3 environment and computational workflow so that the steps described in Chapter 3 are generalizable to different data formats and can be used by other researchers. The complete IsoAnalyst code, workflow tutorial, and descriptions of data input and output were developed collaboratively by myself and Jeffery van Santen, and can be found on the GitHub page (https://github.com/liningtonlab/isoanalyst)

Dr. Marnix Medema and Dr. Justin van der Hooft performed the automated antiSMASH analysis, and developed a protocol for manually curating the antiSMASH analysis output described in Chapter 5. Tables 5.1 and 5.2 contain the data generated by Dr. Medema and Dr. van der Hooft.

Table of Contents

List of Tables

List of Figures

List of Acronyms

SFU	Simon Fraser University
MS	Mass Spectrometry
SIL	Stable Isotope Labelling
BGC	Biosynthetic Gene Cluster
PKS	Polyketide Synthase
NRPS	Non-Ribosomal Peptide Synthase
UPLC	Ultra-Performance Liquid Chromatography
HPLC	High-Pressure Liquid Chromatography
SAM	S-Adenosyl Methionine
CoA	Co-enzyme A
qTOF-MS	Quadrupole Time-of-Flight Mass Spectrometry
ESI	Electrospray Ionization
DIA	Data Independent Acquisition
SER	Standard Error
NMR	Nuclear Magnetic Resonance
COSY	Correlation Spectroscopy

HSQC	Heteronuclear Single-Quantum Correlation Spectroscopy
HMBC	Heteronuclear Multiple-Bond Correlation Spectroscopy
ROESY	Rotating-frame Nuclear Overhauser Effect Spectroscopy

Chapter 1.

Modern Natural Products: Interdisciplinary Approaches in the Multi-Omics Era

1.1. Introduction

The relationship humanity shares with chemistry produced in nature has been central to not only to our health but also to our indulgence in food, color, and aromas for as long as humans have had the capacity to value such pleasures.[1,2] The oldest recorded uses of natural products come from plants, as they are a plentiful and accessible source of chemistry. Records of humans using plants for medicinal purposes dates back as far 2900 B.C., with the Ebers Papyrus, an Egyptian pharmaceutical record documenting over 700 plant-based medicines.[1] Willow bark, for example, has been utilized as a medicinal treatment for pain and fever in traditional cultures around the world for thousands of years, due to the presence of a compound salicin, which is metabolized to the active component, salicylic acid, in the human body.[1,3] In the nineteenth century, salicylic acid was further developed into acetylsalicylic acid, more commonly known as aspirin, and is still one of the most popular drugs for treating common maladies.[3]

Since the advent of natural product chemistry as a modern scientific discipline, we have been continually inspired by the complexity and specificity of natural product structures. Compounds evolved to solve problems encountered by plants, fungi, and bacteria in the natural world have had an undeniably massive effect on human health.[4,5] Biological sources from every corner of the planet have been studied for their capacity to produce bioactive chemistry.[1,6] Single-agent approaches to drug discovery have not only advanced therapeutic applications, but also provide researchers with chemical probes and tools for studying mechanism of action.[7,8] While the question of how many truly novel natural product scaffolds remain to be discovered in nature is still contended,[9–11] the ecological functions of many known natural products are still not understood.[5,6] In the last century there have been significant advances in our ability to study these compounds, which are often produced in incredibly small quantities due to their potency and the energy required to produce them.[12] Genomics has further changed the face of natural products by revealing immense metabolic diversity in the genomes of microorganisms.[13,14] Genetic and environmental factors contribute to the chemical phenotype of plant and microbial sources,[15] suggesting more complex questions surrounding the biological production and ecological functions of natural products.

In the post-genomic era, natural products research has moved beyond drug discovery to other applications in human health such as the complex interactions and immunity conferred by the microbiome.[16,17] There is no doubt that natural product research has had a massive impact on modern medicine,[4] but historical, traditional, and cultural knowledge of natural sources of medicines are still making their mark on human health.[18] The Nobel laureate, Tu Youyou, was acknowledged in 2015 for her work towards the discovery of the anti-malarial natural product artemisinin, which she famously extracted by interpreting ancient Chinese medical texts.[19,20] According to the World Health Organization (WHO), traditional and complementary medicines are a significant source of primary and complementary health care globally.[21] WHO emphasizes the importance of researching of naturally derived medicines which will continue to play a significant role in global health.[21] Developing a more holistic understanding of natural products and their roles in nature and human health is finally possible under the culmination of decades of biological and chemical research in multi-omics disciplines.[18]

1.1.1. The Omics Family

Twenty years ago, William Bains wrote in regards to the field of genomics:[22]

> The genome has turned out the be a relatively poor source of explanation for the differences between cells or between people.

This was something of a bitter revelation that took the scientific community decades to come to understand. The discovery of the genetic code led to ever growing excitement over promises the ‘blueprint for life’ offered to revolutionize medicine by revealing a genetic source for every disease. Over time, the sequencing of genes became faster, cheaper, and more efficient, and the field of genetics grew into genomics. Certainly if genetic sequences could provide the instructions for cellular functions, whole genome sequences would reveal every gene and its function in the diversity of life. However, the one gene one enzyme model began to unravel as full genome sequencing revealed far fewer genes than could account for all characterized proteins.

At the turn of the century, the field of genomics had reached a pivotal moment. The genetic determinism that had dominated the past 50 years of biology was no longer sufficient to describe biological variation. In his 2001 commentary “The parts list of life”, Bains underpins the significant need to transition towards a systems biology approach in order to apply different levels of explanation, such as genomics or proteomics, to higher levels of understanding disease, human health, and basic phenotypic variations.[22] In an earlier commentary, he compares this systematic thinking to a more obvious scenario:[23]

> Saying that a gene ‘causes’ hypertension or depression is similar to saying that a flat tire ‘causes’ a car to slow down. In a few pathological situations, the two are causally linked, but most cars traveling slowly do not have flat tires. A ‘tire knockout car’ would tell us little about traffic lights or driving skills or speed cameras.

It seems a bit ridiculous to take the tires off of a car in order to answer questions about the car’s speed, but this does demonstrate the nature of genetic knock-out experiment with a narrow outlook on how the genetic knock-out actually changes the system. The speed of the car is not only determined by the tires but also inherently by the engine and circumstantially by the driver, traffic, weather, and route. Likewise, a gene’s function is determined by many more factors than just its sequence. A researcher interested in applying the knowledge of genomics to a biological system needs a toolbox of approaches at many explanatory levels, or else risk mistaking a traffic light stuck on red with a nail in the tire. Proteomics, transcriptomics, and metabolomics were young fields twenty years ago, but they were already being applied in innovative ways to explain phenotypes and metabolic processes.[24] Genomics, and the daughter technologies that have come subsequently, arose from a need to functionally understand the connection between genotype and phenotype.

Transcriptomics looks at a snapshot of mRNA transcripts present in the cell as evidence to account for genes that are actively being transcribed, while proteomics provides evidence for enzymes and other proteins that are actively being translated. All stages, including genomics, are subject to a variety of factors that regulate and control the flow of information through each level.[25] Epigenomics is a growing area looking at factors influencing gene expression that has added yet more complexity to the whole picture of gene expression.[26] Metabolomics is the only omics technology that looks directly at output of all the aforementioned biochemical processes. Metabolomics is the most direct way to study phenotypic output and has proven to be much more powerful in clinical diagnostics than genomics.[25,27] The main bottleneck in metabolomics studies is still the difficulty in quickly and definitively identifying compounds, as well as differentiating unknown from known metabolites.[12,25,28]

Today, multi-omics approaches are advancing quickly and becoming the standard for understanding biological systems. As our knowledge of biological systems deepens, there is an ever greater need to incorporate large datasets together to draw meaningful conclusions and many reviews are now available on technological advances allowing for the combination of this information.[26,29–31] In this thesis I focus on the development of a mass-spectrometry based metabolomics tool to connect chemical phenotypes to genomic information. The motivation for this work is grounded in the unique position of metabolomics to galvanize a more holistic view of systems biology through its relationship to other omics technologies.

1.1.2. The Genomic Era of Natural Product Discovery

Genome sequencing has fundamentally changed the paradigm of natural products research and discovery. By the 1980s, natural product chemistry as a field was already focused on the chemical phenotype of cells, but had yet to fully dive below the surface of the genomic origin of these metabolites.[14] Classical genetic techniques were used to establish biosynthetic logic and understand the genetic basis for specialized metabolism, but most natural product discoveries in bacteria and fungi were applied in a chemistry-focused manner.[14] Bio-assay guided fractionation was the dominant technique that focused on identifying chemistry in terms of its structure and function.[1] Bacterial natural products had been a fruitful area of research for a number of decades but around the 1990s many industrial research programs were turning away from natural products research as it was believed we were reaching a plateau of discovery.[1,18] Model actinobacterial species such as *Streptomyces coelicolor* were known to produce a few compounds and an organism that produced 5 or 6 compounds was considered a privileged producer (Figure 1.1). However, when *S. coelicolor* was sequenced in 2002 it was found to have a plethora of identifiable BGCs without distinct products known to be produced.[14,32] This observation became a trend as more bacterial genomes were

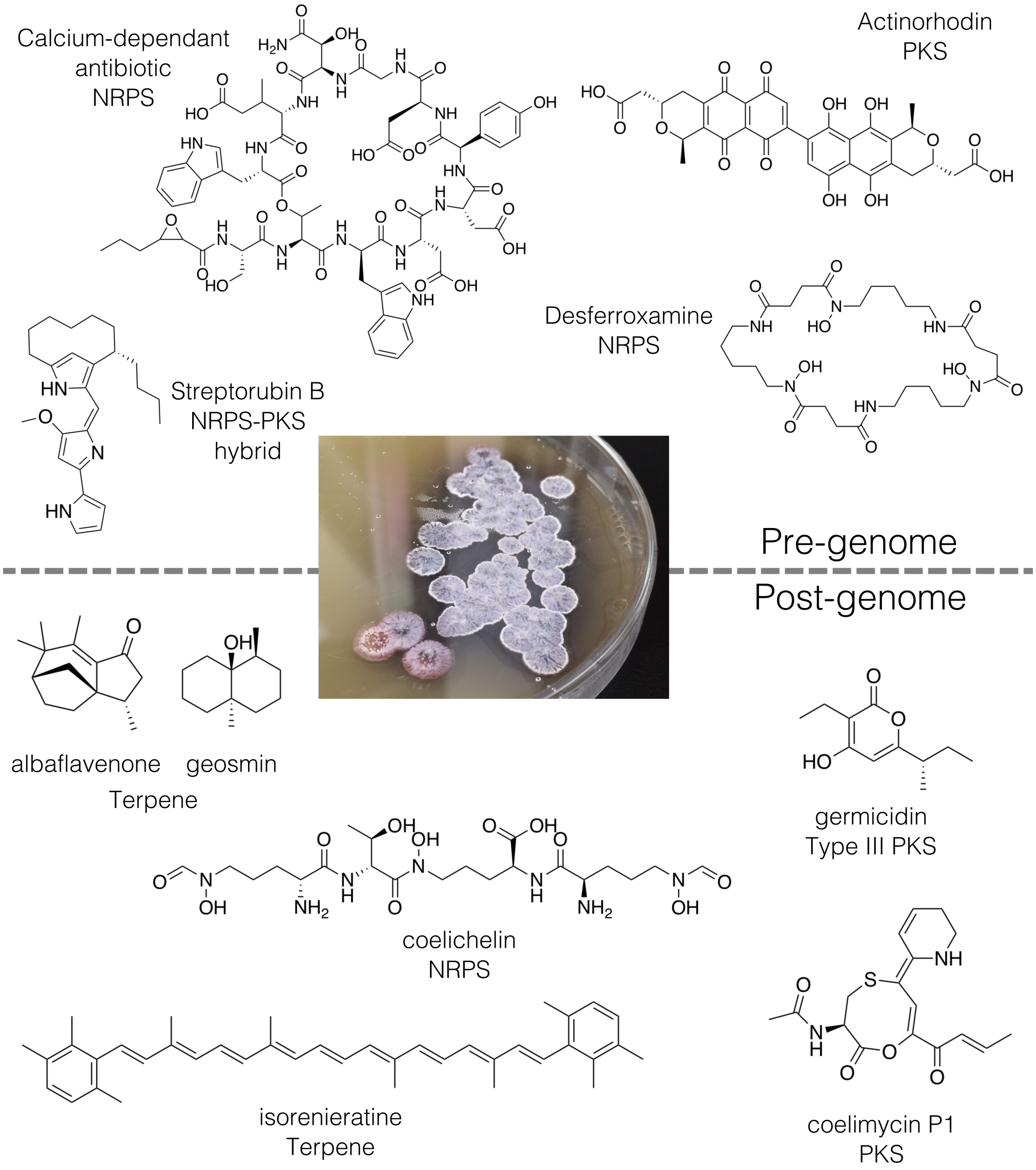


Figure 1.1 Natural Products Produced by *Streptomyces coelicolor*
Structures of natural products produced by *S. coelicolor* which were discovered prior to the publication of the full genome in 2002 (top), and those products discovered after the genome was sequenced (bottom).

sequenced and initiated a genomics revolution in natural products research. Like genetic determinism, genomics-driven natural products research at first promised to generate fast and high-throughput workflows to discover novel chemistry by drawing direct connections between BGCs and compounds.[13] As databases fill with BGC sequences, the influx of novel chemistry discovery has not followed, and the idea that every new BGC would also produce novel chemistry was soon shown to be flawed.[33] Just as one gene does not produce one enzyme, BGCs likewise do not produce singular chemical entities.[34,35] The levels of explanation of BGC diversity and evolution are still being built through the development of computational and analytical tools discussed throughout this chapter.

1.2. Advances and Challenges in Genomic Discovery of Natural Products

1.2.1. Tools for Sequencing and Analyzing Biosynthetic Gene Clusters

With advancing technology and decreasing costs, the accessibility of high-quality genome information has increased exponentially.[36,37] Over the last decade, this accumulation of genome sequence data has inspired the development and refinement of computational tools for identifying biosynthetic gene clusters (BGCs),[38] predicting structural elements of the compounds they produce,[39,40] prioritizing novel BGCs,[41] or a combination thereof.[42] Technological advances in sequencing have been a major driver in that ability of researchers to access genetic sequences of BGCs. BGCs are made up of large multi-modular domains resulting in long repetitive genetic sequences that are difficult to align due to high overlap.[43] Second and third generation sequencing technologies have complementary strengths, in terms of read length compared to error rate.[44] Third generation single molecule real time (SMRT) sequencing produces average read length of over 10 kb, compared to the 250 bp maximum read length of Illumina HiSeq, but suffers from a higher error rate.[44] For the application of assembling the large multi-domain BGCs that produce natural products, the tradeoff of a higher error rate for longer read lengths offered by SMRT is necessary to align large BGC sequences. Even laboratories that are not equipped to work with genomic sequencing technology can access genomic data and analysis of published BGC sequences through public databases such as the atlas of biosynthetic gene clusters within the Integrated Microbial Genomes system (IMG-ABC),[45] the Minimum Information about a BGC database,[46] and the antiSMASH database[47] (Figure 1.2). These databases will continue to grow in both size and value to the scientific community as annotations and experimental validation of chemical structures are updated. Community efforts such as the Natural Products Atlas[48] and Lotus[49] databases also enhance our understanding of the "specialized metabolome" by providing curated records linking structures to the taxonomy of their producing organisms.

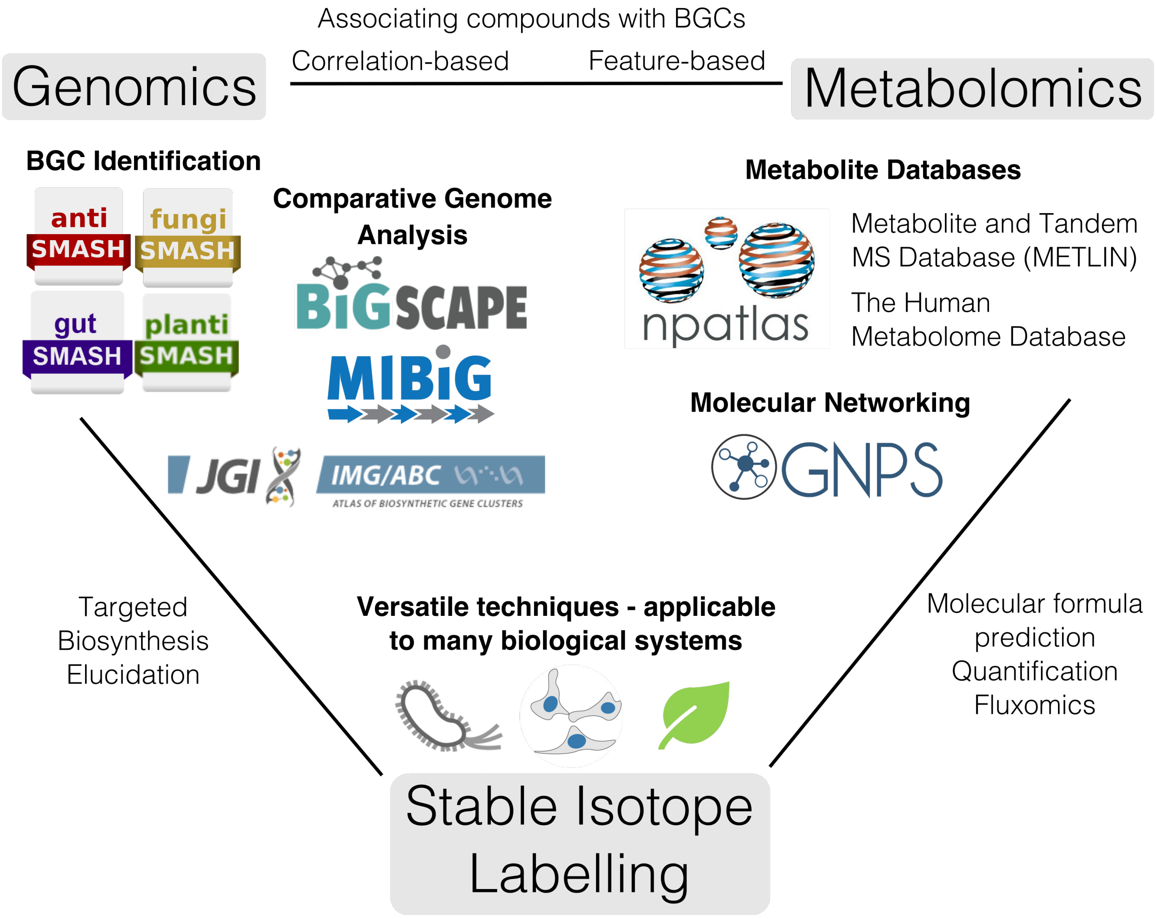

Associating compounds with BGCs
Correlation-based
Feature-based
Genomics
Metabolomics
BGC Identification
anti SMASH
fungi SMASH
gut SMASH
planti SMASH
Comparative Genome Analysis
BiGSCAPE
MIBiG
JGI
IMG/ABC
ATLAS OF BIOSYNTHETIC GENE CLUSTERS
Metabolite Databases
npatlas
Metabolite and Tandem MS Database (METLIN)
The Human Metabolome Database
Molecular Networking
GNPS
Targeted Biosynthesis Elucidation
Versatile techniques - applicable to many biological systems
Molecular formula prediction
Quantification
Fluxomics
Stable Isotope Labelling

Figure 1.2 Overview of Genomics, Metabolomics, and SIL Tools
This thesis focuses on the intersection of genomics, metabolomics, and SIL, which have complementary technologies and have been integrated in various ways. The center of this figure shows the technologies and advantages of each technique and the lines connecting genomics, metabolomics, and SIL show current integration of each technique. Genomics tools for BGC identification include antiSMASH for bacterial genomes, as well as related programs for fungi, plants, and intestinal microbiome sources, large-scale comparative genome analyses (BiG-SCAPE[52]), and open source BGC databases such as MIBiG[46] and IMG/ABC.[45] In metabolomics, growing metabolite databases facilitate the extension of both targeted and untargeted studies. SIL has proven to be a versatile technique, which is applied in different ways to both genomics and metabolomics.

Because of the increasing availability of full genome data, there have also been important advances in computational tools for the identification of BGCs, and for prioritizing them based on novelty or relatedness.[50] High confidence/low novelty tools that utilize hidden Markov models (HMM) and training sets to identify BGC Pfams are now standard for identifying non-ribosomal peptide synthase (NRPS) and polyketide synthase (PKS) BGCs.[36] Computational tools such as antiSMASH[42] and PRISM[37] apply HMMs to identify BGCs and represent a significant advance in the accessibility of BGC analysis to natural product chemists who do not typically have the background or resources to work with large genomic datasets. These tools rely on training sets of sequence alignments from well-characterized biosynthetic domains and therefore have a bias towards finding BGCs related to known classes. Although this is very effective for identifying new BGCs of known classes, two novel algorithms that prioritize novelty over confidence have been developed. The ClusterFinder algorithm uses an HMM to identify regions of the genome containing Pfam domains that resemble the overall organization and frequency of a BGC, as opposed to specific domains related to biosynthetic transformations.[38] ClusterFinder is available within the antiSMASH platform and can be run alongside the more traditional profile HMM algorithm for detecting known biosynthetic classes. Another approach, termed EvoMining, searches for additional copies of genes for primary metabolic enzymes, on the basis that specialized metabolism evolves from divergent copies of primary metabolic genes.[41] EvoMining prioritizes novelty by searching for genes known to derive from primary metabolism, but which may have novel biosynthetic activity compared to relying on training sets derived from known biosynthesis.[41]

The growing accessibility and versatility of these BGC analysis tools has transformed genome mining towards a more global approach. Large-scale comparison of genomic datasets and sequence alignment has enabled BGC discovery and classification across phyla, and deepened our understanding of gene cluster families.[14,33,36] The sequencing and characterization of hundreds of BGCs over that past ten years has opened new opportunities to look at the global diversity of BGCs across phyla. Comparisons of BGC families across multiple genomes have proven to be powerful tools in leveraging these large dataset for natural product discovery.[38,51] Bacterial genome analysis is rapidly changing from an effort directed towards a handful of strains, to one focused on thousands of sequences. Large-scale comparisons of BGC families have proven to be powerful tools in leveraging these large datasets for natural product discovery and for associating shared biosynthetic capacity with natural product production.[31,52]

1.2.2. Connecting Molecules to Biosynthetic Gene Clusters

The highest standard of evidence for connecting a BGC to the metabolite(s) produced is through heterologous expression, or a combination of promoter and knock-out studies to validate compound production when the BGC is being expressed. Pre-genomic era approaches relied on reverse genetics and heterologous expression to painstakingly assign BGCs to known chemistry.[14] Modern techniques for heterologous expression[53,54] and silent BGC activation[55] have re-invigorated biochemical approaches for natural product discovery. However, the massive accumulation of genomic data is further enabling analytical techniques for making connections between BGCs and compounds on a larger scale.[31] New BGCs are discovered with a higher frequency than new natural products due to a lack of tools for systematically connecting chemistry with BGCs.[31,56] It is a much more standard procedure to detect and compare BGCs across genomic datasets than it is to fully characterize the chemical potential of a single organism.

Two major bottlenecks in identifying chemistry from BGCs are the inaccuracy of compound structure prediction from BGC sequences, and the detection and dereplication of known and unknown compounds in metabolomics datasets.[31] Structure prediction from BGC sequences does not currently have a high enough accuracy to be fully automated.[31] Although domain sequences can be leveraged to predict enzyme substrates, especially for PKS and NRPS gene clusters, tailoring reactions, modifications, and cyclization reactions are not as predictable.[31] This precludes the ability to predict the exact structure, fragmentation, and molecular formula of a natural product from its BGC. Research on co-evolving sub-clusters are advancing our ability to predict shared structural features among gene cluster families,[40] however literature review remains necessary for accurate BGC substrate analysis.

One of the most significant challenges is that not all BGCs are expressed under laboratory conditions.[33] There are many techniques that can assist in the expression of cryptic or silent BGCs, including media experiments (one strain many compounds), challenge agents such as antibiotics, and co-culturing have all been used to encourage expression of BGCs.[57] Many innovative approaches have been developed to elicit compound production including high-throughput elicitor screening[58] and CRISPR-Cas9 strategies for activating BGCs in the native host.[55] However, these powerful tools for compound elicitation still require advanced analytical techniques to identify target compounds.

To address analytical challenges, there are two main types of approaches for connecting molecules to BGCs using MS metabolomics data. Feature-based methods use pattern matching of molecular networks generated from MS/MS fragmentation data to structural predictions from BGCs. The molecular networking approach employed by the Global Natural Products Social Molecular Networking (GNPS) knowledge base uses MS/MS spectra to visualize MS features with related fragmentation spectra[59] (Figure 1.2) Algorithms that predict BGC substrates, particularly for NRPS gene clusters, can be associated with molecular families identified by MS/MS networks.[34,60,61] This is effective for discovering compounds of specific classes, especially those that have distinct MS/MS fragmentation patterns. Correlation-based methods look at links between gene cluster families and molecular families across many strains harboring related BGCs.[31,35,62] Correlation-based methods can associate shared chemistry produced across many strains without MS/MS data, although the fragmentation spectra often plays an important role. A recent preprint by Eldjárn et al. details a novel computational approach which applies both feature and correlation based linking of compounds to BGCs systematically.[63] It is clear that all of these approaches have powerful ongoing applications in natural product discovery by genome mining. However, the over-reliance on MS/MS fragmentation for identifying and characterizing chemistry is a weakness that biases these approaches towards compounds that fragment well and BGCs that have predictable structures. Indeed, substrate predication is not nearly as automated as MS/MS fragmentation or BGC correlation across strain libraries, so orthogonal analytical methods to identify products of BGCs would greatly enhance these approaches.

In this chapter I describe the context of the development of such an analytical method which relies on genomics, metabolomics, and SIL tracers to identify natural products in association with the producing BGC. This method relies on the strengths of each area, and the modern technological advances that have galvanized the opportunity to draw connections between BGCs and chemistry in a systematic manner.

1.3. Natural Product Characterization in Metabolomics

1.3.1. Primary and Specialized Metabolomics

Natural products are generally understood to be small compounds (< 1500 Da) produced in nature that do not serve any primary biological function, but may serve important biological roles under environmental pressures. For this reason, natural products have been referred to as ‘secondary metabolites’, implying that they are not required for the organism’s survival.[64,65] Our understanding of what constitutes a secondary metabolite remains unclear as many molecules traditionally categorized as natural products have been shown to serve essential biological functions in niche environments. Consequently, the term ‘specialized metabolite’ is often used to indicate that these molecules serve specialized but still essential functions. While the monomeric units of proteins, sugars, DNA, and lipids remain nearly ubiquitous across domains of life, specialized metabolites are distributed according to varying ecological and evolutionary pressures.[65]

In reality, all structural classes of compounds are differentially subject to metabolomics-based investigation, and do not divide neatly into primary and specialized metabolites based on physical and chemical characteristics alone.[28] Primary and specialized metabolism also share pools of monomers and are biochemically structurally related evolutionary relatedness between metabolic processes.[66] For example, even though they are often structurally modified and contain non-canonical amino acids, NRPS compounds can be analyzed by fragmentation patterns using the same physical and chemical principles that define how protein sequences are derived from predictable fragmentation in proteomics.[67] Similarly, terpenes and fatty acids come from completely different biosynthetic classes, but both are characterized by hydrocarbon backbones. Although derived from different biosynthetic pathways and serving different biological roles, these classes of compound can both be detected and quantified by standard GCMS methods due to their volatility, availability of standards, and abundance of published protocols.[68,69]

Within this body of work I will henceforth use the terms specialized metabolite and natural product synonymously. This is done with the intention of distinguishing primary and specialized metabolomics studies, not by the compounds detected in the experimental system, but rather by the experimental design and proposed data analysis. Primary metabolic pathways that are central to proliferation and growth have been studied through metabolomics to establish pathway sequences, compare altered metabolism is different disease states, and as disease markers.[25,28] Primary metabolomics often goes beyond looking at a snapshot of metabolite levels, towards the study of fluxomics, which aims to observe metabolic flux in biological systems.[70] Systems biology and fluxomics studies can inform natural product applications, especially in chassis optimization for production of the primary building blocks necessary for heterologous expression. However, secondary metabolomics studies typically aim to characterize the final product of a pathway, rather than assess the overall metabolic processes at play.[12]

Both primary and specialized metabolomics aim to characterize the final products of gene expression however the prioritization of detected compounds differs significantly. Primary metabolites can be studied in a very targeted manner, focusing only on compounds with known identities, to derive biological insight.[12] For example, routine metabolomics methods targeting central metabolites are used to characterize glycolysis, pentose phosphate pathway, and the TCA cycle as well as amino acid metabolism.[25] Many specialized metabolomics experiments aim to discover unknown or novel metabolites and identify known compounds only in order to deprioritize them.[12] The value of natural products in human medicine drives this field to focus heavily on the discovery of new chemistry, for applications in drug discovery. The differences in aims is what sets apart primary and specialized metabolomics studies more than the target metabolites, as there is often overlap between primary and specialized metabolite pools. Annotation of chemistry associated with BGCs has implications across many disciplines in both medicine and chemical ecology.[13]

1.3.2. Traditional Compound Isolation and Characterization

Compound identification and characterization techniques have changed significantly throughout the history of studying natural products. In the earliest period of natural product discovery, from the 1940s through the 1970s, relatively simple analytical techniques were frequently rewarded with the discovery of novel compounds with powerful biological activities.[64] Extracts were typically profiled using a simple but powerful screening technique known as TLC-bioautography. Biological extracts were applied as mixtures to agar plates inoculated with target organisms as whole-cell live/dead assays. Hits were determined by a zone of inhibition around the spot where the extract was applied. Extracts demonstrating activity were then further separated using a TLC solvent system, and applying TLC-separated extracts to the same agar assays in order to track the active component in the mixture. In this iterative manner, the active component was identified and purified from large scale fermentations for further characterization.[64,71] Prior to HPLC, isolation methods were less precise and often resulted in mixtures of isomeric compounds. Similarly, chemical characterization methods such as IR, NMR, and degradation analyses were less precise and required large quantities of compound. Because of these limitations, this period is defined by the discovery of compounds often produced in high titer under laboratory conditions, and compounds which have strong identifiable activity in whole-cell antibacterial assays.

In the 1980s, HPLC revolutionized natural product research through the development of the bio-assay guided fractionation approach.[1,64] Bio-assay guided HPLC fraction workflows were essentially a result of multiple technological advances in both chromatography and biological activity determination. HPLC separations of extracts are slow due to their time-consuming and serial nature compared to TLC which is relatively quick and easily run in parallel conditions.[71] Despite this, the improvement in chromatographic precision achieved with HPLC allowed for more complete separation of isomers, confidently purified compounds on a micro-scale, and all around more efficient workflows. Automated fraction collection systems were a major driver in the throughput of this fractionation process, which was necessary to keep up with the advances in biological screening. By the mid-2000s, high throughput screening campaigns were becoming more available for applying large libraries of compounds made from HPLC separated extracts from natural sources.[1] Developments in NMR and MS technologies further drove this field to isolate and characterize more compounds.[72]

Bioactivity guided fractionation by HPLC quickly became a dominant workflow in natural products and it is still relied on heavily today. However, even carefully planned HTS and metabolomics experiments often lead to the rediscovery of compounds that were just as easily discovered in the TLC-bioautography workflows of the 60s and 70s.[1] Technological advances certainly drove natural products discovery over the last several decades, by improving the purification and analysis of natural products, but the challenge of deconvolution remains as a complex problem in metabolomics today as it was in early discovery workflows. Indeed, databases with Rf values in numerous solvent systems and bioactivity profiles of natural products were used to dereplicate TLC-bioautography results much in the same way MS/MS fragmentation comparison has become essential to the dereplication toolbox of modern natural products researchers.[59,71] Advancing technology is continually met with the infinite complexity of nature.

Modern metabolomics is done using MS or NMR based approaches and both have advantages and challenges associated with them. NMR is inherently quantitative but this advantage also biases the technique towards compounds produced in higher concentrations.[65] Various techniques have been developed for compound identification by NMR metabolomics of complex extracts,[73] however, the work presented in this thesis focuses on MS-based metabolomics approaches.

1.3.3. Mass Spectrometry Metabolomics

Technological Advances in MS have motivated the application of metabolomics to many different biological systems. Although it is sample destructive, MS is a highly sensitive technique for many compound classes and for this reason is often preferred over NMR for metabolomics applications. Improvements in MS resolution have culminated in the ultra-high resolution of the Fourier transform ion cyclotron resonance (FTICR) (resolution >500,000, accuracy <1ppm).[65] The increase in mass accuracy has greatly improved our ability to predict molecular formulae and identify compounds by database matching. Still, a sub-1 ppm mass error is not sufficient to eliminate all possible molecular formulae.[74] Advances in ultra-high-pressure liquid chromatography (UPLC) technology has allowed for sub-2 μm particle size, pressures above 500 bar, and effective runtimes under ten minutes. UPLC-MS is ideal for high throughput of metabolomics samples due to the short runtimes and excellent reproducibility in peak retention. No single hyphenated MS technique is ideal for complete coverage of all types of metabolites, however orthogonal chromatography methods (reverse phase, HILIC, GC, super critical fluid chromatography) and ionization techniques (ie EI, ESI, APCI, MALDI, DESI, etc) facilitate customized workflows for diverse metabolite coverage. Protocols for specific classes of compounds have become standardized for targeted metabolomics studies, especially of primary metabolism.[25]

In terms of natural products research and discovery, MS/MS molecular networking has become the gold standard for dereplication and has invigorated many creative MS metabolomics applications.[75] The GNPS community has contributed over 70,000 annotated MS/MS spectra for known compounds, allowing a more comprehensive and streamlined process for dereplication than ever before. The strength and sensitivity of modern MS technologies combined with the accessibility of the GNPS database has changed the face of natural products discovery. Beyond dereplication, the layering of multi-informational data such as bioactivity, taxonomy, geographical location, genomics, epigenetics, and SIL have greatly enhanced opportunities to apply molecular networking to complex problems and large datasets.[75] Imaging MS has also invigorated new natural products discovery approaches by allowing for spatial resolving power and *in situ* detection of metabolites. There are many different ionization techniques for imaging MS that can be applied to different types of samples and target compounds.[76] Imaging MS can also be paired with different detection systems to allow for high mass accuracy, ion mobility separation, and MS/MS fragmentation.[76]

These developments have led researchers to move towards a more global outlook on metabolomics. The sensitivity and throughput of modern MS technology allows for the acquisition and analysis of hundreds of thousands of mass features, many of which are not able to be identified by database searching.[25,28] As our experimental snapshots of the metabolome become more complete, it is clear that our current view of metabolism is insufficient to account for the complexity observed in the metabolome. These observations demonstrate that targeted metabolomics studies will only take us so far in advancing our understanding of metabolism, and consequently untargeted metabolomics workflows have become more popular and standardized.[25] While targeted metabolomics studies can be said to be 'hypothesis-driven', untargeted metabolomics studies are better viewed as 'hypothesis-generating.' A more global view of the metabolome promises an incredible potential for discovery, but is accompanied by new experimental and computational challenges. Untargeted MS metabolomics approaches have been mainly limited by compound identification which precludes the ability to draw significant biological conclusions.[25] Untargeted approaches are still limited by compound identification, however there are a growing number of tools available to analyze MS metabolomics data in an untargeted manner, and databases.

1.4. Stable Isotope Labeling

There is no doubt that isotopic labeling has shaped and driven our understanding of metabolism. The radioactive isotope ^{14}C was used by Krebs to elucidate the TCA cycle[77] and soon became established as the primary technique for following the fate of carbon atoms through biological transformations. Even stable isotopes such as ^{13}C and ^{2}H were detected by chemical degradation prior to the availability of MS to elucidate the biosynthesis of cholesterol.[78] Since heavy isotopes were first used in substrates to be fed into a biological system and traced to determine their metabolic fate, many advances have been made in both applications of SIL feedings, and detection of downstream metabolites. MS in particular has allowed for the efficient detection of stable isotopes, which have effectively replaced radioactive isotopic tracing in biosynthesis studies due to the considerably safer and more sensitive detection of stable isotopic tracers. SIL is the cornerstone of natural product biosynthesis, however, SIL applications are far from old-fashioned, and are continually being adapted for use with innovative applications.[70]

1.4.1. SIL Approaches for Genome Guided Natural Product Discovery

Because SIL precursor feeding has been so successful in the area of elucidating biosynthesis, it is a natural addition to the toolbox of genome-based discovery. Information about the domain specificities within BGCs allow for substrate predictions, which are particularly well-studied for NRPS and PKS gene clusters. The genomisotopic approach coined in 2007, led to the discovery of the lipopetides, orfamides A-C.[79] In this study the authors identified an NRPS BGC and used computational tools to predict the amino acid sequence of the peptide product[79] (Figure 1.3) They used ^{15}N labeled leucine, which was selected because it was predicted to occur in the product structure four times, ensuring more robust MS and NMR signals. This was a highly targeted but effective approach, which relied on selecting SIL precursors specific to a single target BGC.

Similar approaches have been successful in targeting compound discovery based on substrate specificity analysis of enzymatic domains of cryptic BGCs.[80,81] This however often requires the use of advanced precursors which target the BGC of interest selectively compared to other BGCs present in the genome. Compounds that are tailored to predicted enzyme specificity are more expensive and it is common to use precursors containing multiple labeled positions to clarify instances of direct incorporation from labeling derived from metabolic processing of the SIL precursor prior to incorporation[80] (Figure 1.3) Although the genomisotopic approach is powerful for targeted discovery, these limitations impede the development of an easily generalizable approach for isotopically labeling natural products globally.

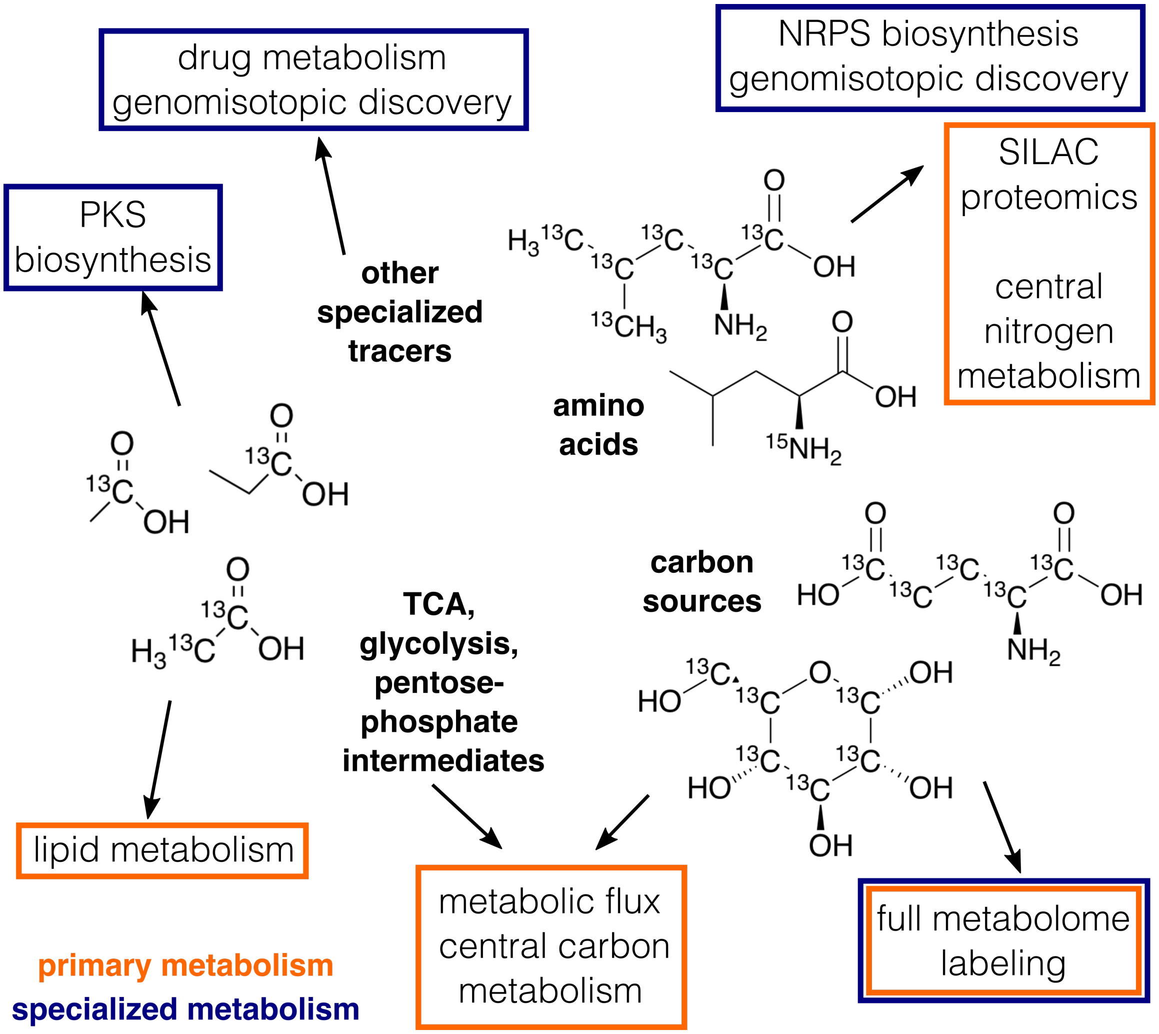


Figure 1.3 SIL Tracers and Common Biological Applications
Examples of SIL tracers that are commonly used in SIL feeding experiments to study primary (orange boxes) and specialized (blue boxes) metabolism. Most groups of metabolites used as SIL tracers may be applied to both primary and specialized metabolism. Because metabolic flux experiments look at the full metabolic landscape, they tend to rely on central primary pathways while specialized SIL metabolomics tend to use more targeted SIL tracers.

1.4.2. Stable Isotope Labeling in MS Metabolomics

SIL has many versatile applications in MS metabolomics. In particular, SIL assists in molecular formula assignment[74] and in differentiating biologically derived compounds from background ions.[70] Fully labeled extracts can be paired with unlabeled extracts as reference standards to assist in compound identification.[70] SIL MS metabolomics is also applied to detect the biotransformations of xenobiotics[82] as well as in metabolic flux analysis to determine metabolic fluxes under steady-state or dynamic conditions.[83] Metabolic flux is a measure of metabolite turnover over time and provides an orthogonal view of cellular metabolic phenotype to metabolomics (Figure 1.3).

Metabolomics is an attempt to observe the complete metabolic phenotype of an organism at a given moment in time.[70] Although this static view of the metabolome is rich and informative, it is not sufficient to gain a complete biological understanding of metabolism. Fluxomics is an adjacent field to metabolomics that is often used in conjunction with other ‘omics’ techniques to build systems level models of biological systems. SIL compounds, typically labeled by ^{13}C, are used to trace the flux through different metabolic pathways over time. Advanced computational analysis of SIL incorporation across the metabolome can determine pools of metabolites that are enlarged or depleted, and illuminate the metabolic pathways responsible based on layered omics information that underlies the constructed metabolic network.[83] Metabolite levels and fluxes provide complimentary and essential pieces of information in understanding the underlying metabolic system.[70] For example, a metabolomics experiment may detect a significant increase of a certain metabolite in a disease state model, but the metabolite level alone will not tell you if the pathway that consumes it has been downregulated, or if a reaction that produces it has been upregulated. Complex computational models are used to interpret these data but intuitive interpretation of untargeted SIL metabolomics data can be used to complement fluxomics. Because metabolic flux analysis is constraint based and requires exact inputs to the system, it is not ideal for natural product discovery which relies on changing environmental conditions and genetic manipulations to encourage compound production. In natural products, fluxomics analyses are most often applied to microbial factory and chassis optimization as opposed to compound discovery. However, because fluxomics aims to look at the full metabolic system of an organism, many powerful orthogonal data analysis tools have been developed for untargeted SIL detection across the metabolome.[84]

1.4.3. Untargeted Metabolomics

Untargeted SIL metabolomics present a very exciting outlook on the future of metabolic studies. These approaches have great potential for discovery of new metabolites, pathways, and genes.[84] As I have covered throughout this chapter there have been significant advances in MS technologies, and SIL approaches that facilitate researchers' ability to access massive amounts of data about biological systems. Despite the exciting perspective these approaches can bring to studying metabolism, untargeted SIL metabolomics is still quite intractable due to the great difficulty in analyzing the data.[84] Without knowing ahead of time which compounds to look for, MS-metabolomics datasets are intimidating in the sheer number of ions to interrogate. Advances in automated processing and database information about known compounds are assisting in the progress of dereplicating known compounds in both primary and secondary metabolism. Most of these methods are focused on fluxomics, and primary metabolism, but there are powerful implications these approaches can have in specialized metabolite discovery.

In natural products discovery, truly untargeted SIL MS metabolomics approaches are not common. Although it is true that all metabolomics studies aim to find the metabolic phenotype of a particular genotype, natural product metabolomics has a particularly strong focus on final metabolite discovery. Many untargeted approaches aim to identify metabolites in the context of their biological role in metabolism and often generate hypotheses about the biological system being studied.[84] In these kinds of metabolomics approaches statistical analyses are used to determine differences in metabolite flux and concentration under a biological context, while novel compound isolation is de-prioritized. In natural products, metabolomics is most often used as a discovery tool to identify compounds of interest for further isolation and characterization and application to an unrelated system.[12] Natural product metabolomics innovation is marked by the integration of orthogonal biological data from high-throughput screening with MS metabolomics data to facilitate compound discovery. This distinction in aims is likely why SIL MS metabolomics studies involving natural products are nearly always targeted.

The culmination of knowledge in the areas of genomics, BGC analysis, MS metabolomics, and systems biology has created a unique opportunity to better understand the relationship between genes and molecules. I have developed an untargeted SIL metabolomics platform which aims to facilitate specialized metabolite discovery by genome mining and is flexible enough to apply to any microbial culture and compound class. The central principle of this technique it to bring together knowledge of both the genome and metabolome to characterize the complex chemical phenotypes of BGCs.

1.5. Overview of IsoAnalyst Approach

The research described in this thesis brings together the strengths and weakness of current approaches in genomics, MS metabolomics, and SIL tracers to produce a systematic method for the untargeted discovery of compounds produced by BGCs. This platform employs parallel stable isotope labeling (SIL) to categorize specialized metabolites in liquid fermentations based on biosynthetic precursor incorporation, and connects these molecules to candidate BGCs using annotated genome sequence data. Parallel fermentation of the test organism is performed in the presence of either an isotopically labeled precursor or a control culture containing the corresponding unlabeled precursor for a panel of SIL tracers (Figure 1.4a). In Chapter 2 I describe the development of this fermentation protocol and describe the results of optimizing the growth media and SIL tracer selection. Following workup and UPLC-MS analysis of all samples, the IsoAnalyst pipeline identifies unique MS features present in each condition, and compares mass isotopologue distributions between unlabeled and labeled conditions to determine the degree of labeling by each precursor (Figure 1.4b). In Chapter 3 I describe the IsoAnalyst program itself, including the input data requirements, statistical analysis, and validation. I further validate the application of IsoAnalyst in Chapter 4 using model organisms with well-characterized biosynthetic pathways. The experimental precursor incorporation patterns determined by IsoAnalyst are then manually compared against the theoretical precursor incorporation rates derived from BGC annotations (Figure 1.4c) to yield candidate BGC(s) responsible for the production of each labeled metabolite. This approach streamlines the categorization of analytes by their biosynthetic origin and reduces MS metabolic profiles to quickly delineate the complex phenotypes of BGCs. In Chapter 5, I demonstrate the full application of the IsoAnalyst platform and BGC analysis on a sequenced environmental isolate, *Micromonospora sp.*, from our Actinobacterial isolate library. This complete workflow culminated in the association of both known and unknown compounds originating from the same BGC and the discovery of a new analogue belonging to the lobosamide family of macrolactams. There are many exciting and innovative applications for the IsoAnalyst platform to facilitate novel compound discovery and complete characterization of complex chemical phenotypes. In this thesis, I lay the foundation for this technique and demonstrate the utility of IsoAnalyst to understanding biosynthetic potential in microorganisms.

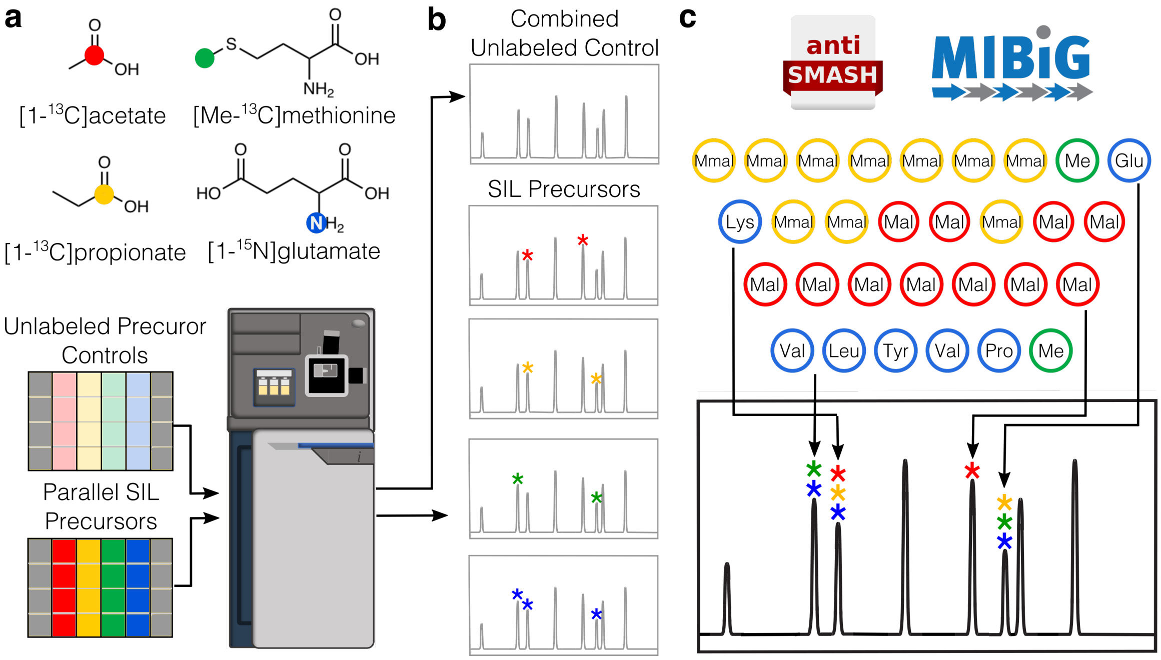


Figure 1.4 Overview of the IsoAnalyst Workflow
(a) SIL and unlabeled extracts are grown in parallel for four days, extracted, and analyzed by UPLC-MS. (b) IsoAnalyst program identifies SIL incorporation in MS features aligned across all samples. (c) IsoAnalyst profiles generated for all MS features are compared to curated genomic information from antiSMASH and MIBiG.

References

1. Dias, D. A., Urban, S. & Roessner, U. A Historical overview of natural products in drug discovery. *Metabolites* **2**, 303–336 (2012).

2. Müller, K. Pharmaceutically relevant metabolites from lichens. *Appl. Microbiol. Biotechnol.* **56**, 9–16 (2001).

3. Li, S., Gosslau, A., Lange, K. & Ho, C.-T. Profiled tea extracts exemplifying the importance of characterizing food bioactives: opinion piece. *J. Food Bioact.* **5**, 1–5 (2019).

4. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).

5. Behie, S. W., Bonet, B., Zacharia, V. M., McClung, D. J. & Traxler, M. F. Molecules to ecosystems: Actinomycete natural products in situ. *Front. Microbiol.* **7**, 381-11 (2017).

6. Smanski, M. J., Schlatter, D. C. & Kinkel, L. L. Leveraging ecological theory to guide natural product discovery. *J. Ind. Microbiol. Biotechnol.* **43**, 115–128 (2016).

7. Schulze, C. J. *et al.* 'Function-first' lead discovery: Mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20**, 285–295 (2013).

8. McMillan, E. A. *et al.* A Genome-wide Functional Signature Ontology Map and Applications to Natural Product Mechanism of Action Discovery. *Cell Chem. Biol.* **26**, 1380-1392 (2019).

9. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci.* **114**, 5601-5606 (2017).

10. Walsh, C. T. & Wencewicz, T. A. Prospects for new antibiotics: A molecule-centered perspective. *J. Antibiot. (Tokyo).* **67**, 7–22 (2014).

11. Wright, G. D. Something old, something new: Revisiting natural products in Antibiotic drug discovery. *Can. J. Microbiol.* **60**, 147–154 (2014).

12. Krug, D. & Müller, R. Secondary metabolomics: The impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products. *Nat. Prod. Rep.* **31**, 768–783 (2014).

13. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends Microbiol.* **24**, 968–977 (2016).

14. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).

15. Van Wezel, G. P. & McDowall, K. J. The regulation of the secondary metabolism of Streptomyces: New links and experimental advances. *Nat. Prod. Rep.* **28**, 1311–1333 (2011).

16. Alavi, S. *et al.* Interpersonal Gut Microbiome Variation Drives Susceptibility and Resistance to Cholera Infection. *Cell* **181**, 1533-1546 (2020).

17. Claesen, J. *et al.* A cutibacterium acnes antibiotic modulates human skin microbiota composition in hair follicles. *Sci. Transl. Med.* **12**, eaay5445 (2020).

18. David, B., Wolfender, J.-L. & Dias, D. A. The pharmaceutical industry and natural products: historical status and new trends. *Phytochem. Rev.* **14**, 299–315 (2015).

19. Miller, L. H. & Su, X. Artemisinin: Discovery from the Chinese Herbal Garden The Promise of Project 523. *Cell* **146**, 855–856 (2011).

20. Youyou, T. Artemisinin — A Gift from Traditional Chinese Medicine to the World. *Nobel Prize.* 283–313 (2015).

21. World Health Organization (WHO). WHO Traditional Medicine Strategy 2014-2023. *World Heal. Organ.* 1–76 (2013).

22. Bains, W. The parts list of life. *Nat. Biotechnol.* **19**, 401–402 (2001).

23. Bains, W. Epigenesis and Complexity: Should You Hire An Epistemologist? *Nat. Biotechnol.* **15**, 396 (1997).

24. Cornish-Bowden, A. & Cárdenas, M. L. Complex networks of interactions connect genes to phenotypes. *Trends Biochem. Sci.* **26**, 463–465 (2001).

25. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).

26. Wörheide, M. A., Krumsiek, J., Kastenmüller, G. & Arnold, M. Multi-omics integration in biomedical research – A metabolomics-centric review. *Anal. Chim. Acta* **1141**, 144–162 (2021).

27. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).

28. Milne, S. B., Mathews, T. P., Myers, D. S., Ivanova, P. T. & Brown, H. A. Sum of the Parts: Mass Spectrometry-Based Metabolomics. *Biochemistry* **52**, 3829–3840 (2013).

29. Schorn, M. A. *et al.* A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).

30. Roume, H. *et al.* Comparative integrated omics: Identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms Microbiomes* **1**, 15007 (2015).

31. Van Der Hooft, J. J. J. *et al.* Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).

32. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147 (2002).

33. Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinform.* **20**, 1103–1113 (2018).

34. 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).

35. Goering, A. W. *et al.* Metabologenomics: Correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent. Sci.* **2**, 99–108 (2016).

36. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).

37. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).

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

39. Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).

40. Del Carratore, F. *et al.* Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Commun. Biol.* **2**, 83 (2019).

41. Séelem-Mojica, N., Aguilar, C., Gutiéerrez-García, K., Martínez-Guerrero, C. E. & Barona-Gómez, F. Evomining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genomics* **5**, e000260 (2019).

42. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

43. Lee, N. *et al.* Systems and synthetic biology to elucidate secondary metabolite biosynthetic gene clusters encoded in Streptomyces genomes . *Nat. Prod. Rep.* (2021) doi:10.1039/d0np00071j.

44. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13**, 278–289 (2015).

45. Palaniappan, K. *et al.* IMG-ABC v.5.0: An update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* **48**, D422–D430 (2020).

46. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

47. Blin, K., Shaw, S., Kautsar, S. A., Medema, M. H. & Weber, T. The antiSMASH database version 3: Increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.* **49**, D639–D643 (2021).

48. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).

49. Rutz, A. *et al.* Open Natural Products Research: Curation and Dissemination of Biological Occurrences of Chemical Structures through Wikidata. *bioRxiv* 2021.02.28.433265 (2021) doi:https://doi.org/10.1101/2021.02.28.433265.

50. Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 203–209 (2014).

51. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *bioRxiv* 2020.08.17.240838 (2020) doi:10.1101/2020.08.17.240838.

52. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).

53. Harvey, C. J. B. *et al.* HEx: A heterologous expression platform for the discovery of fungal natural products. *Sci. Adv.* **4**, eaar5459 (2018).

54. Zhang, M. M., Wang, Y., Ang, E. L. & Zhao, H. Engineering microbial hosts for production of bacterial natural products. *Nat. Prod. Rep.* **33**, 963–987 (2016).

55. Zhang, M. M. *et al.* CRISPR-Cas9 strategy for activation of silent Streptomyces biosynthetic gene clusters. *Nat. Chem. Biol.* **13**, 607–609 (2017).

56. Kenshole, E., Herisse, M., Michael, M. & Pidot, S. J. Natural product discovery through microbial genome mining. *Curr. Opin. Chem. Biol.* **60**, 47–54 (2021).

57. Ren, H., Wang, B. & Zhao, H. Breaking the silence: new strategies for discovering novel natural products. *Curr. Opin. Biotechnol.* **48**, 21–27 (2017).

58. Xu, F. *et al.* A genetics-free method for high-throughput discovery of cryptic microbial metabolites. *Nat. Chem. Biol.* **15**, 161–168 (2019).

59. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

60. Liu, W. T. *et al.* MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in *Streptomyces roseosporus*. *J. Antibiot. (Tokyo).* **67**, 99–104 (2014).

61. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).

62. Doroghazi, J. R. *et al.* Aroadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).

63. Eldjárn, G. H. *et al.* Ranking microbial metabolomic and genomic links using correlation-based and feature-based scoring functions. *bioRxiv* 2020.06.12.148205 (2020) doi:10.1101/2020.06.12.148205.

64. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).

65. Wolfender, J. L., Nuzillard, J. M., Van Der Hooft, J. J. J., Renault, J. H. & Bertrand, S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography-High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics. *Anal. Chem.* **91**, 704–742 (2019).

66. Craney, A., Ozimok, C., Pimentel-Elardo, S. M., Capretta, A. & Nodwell, J. R. Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* **19**, 1020–1027 (2012).

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

68. Jiang, Z., Kempinski, C. & Chappell, J. Extraction and Analysis of Terpenes/Terpenoids. *Curr. Protoc. Plant Biol.* **1**, 345–358 (2016).

69. Wallace, K. K., Zhao, B., McArthur, H. A. I. & Reynolds, K. A. In vivo analysis of straight-chain and branched-chain fatty acid biosynthesis in three actinomycetes. *FEMS Microbiol. Lett.* **131**, 227–234 (1995).

70. Jang, C., Chen, L. & Rabinowitz, J. D. Metabolomics and Isotope Tracing. *Cell* **173**, 822–837 (2018).

71. Hook, D. J., Pack, E. J., Yacobucci, J. J. & Guss, J. Approaches to automating the dereplication of bioactive natural products - The key step in high throughput screening of bioactive materials from natural sources. *J. Biomol. Screen.* **2**, 145–152 (1997).

72. Hoffmann, T., Krug, D., Hüttel, S. & Müller, R. Improving natural products identification through targeted LC-MS/MS in an untargeted secondary metabolomics workflow. *Anal. Chem.* **86**, 10780–10788 (2014).

73. Emwas, A. H. *et al.* Nmr spectroscopy for metabolomics research. *Metabolites* **9**, 123 (2019).

74. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**, 105 (2007).

75. Fox Ramos, A. E., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: Different manners, one goal. *Nat. Prod. Rep.* **36**, 960–980 (2019).

76. Spraker, J. E., Luu, G. T. & Sanchez, L. M. Imaging mass spectrometry for natural products discovery: A review of ionization methods. *Nat. Prod. Rep.* **37**, 150–162 (2020).

77. Krebs, H. A., Gurin, S. & Eggleston, L. V. The pathway of oxidation of acetate in baker's yeast. *Biochem. J.* **51**, 614–628 (1952).

78. Rittenberg, D. & Bloch, K. THE UTILIZATION OF ACETIC ACID FOR THE SYNTHESIS OF FATTY ACIDS. *J. Biol. Chem.* **160**, 417–424 (1945).

79. Gross, H. *et al.* The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. *Chem. Biol.* **14**, 53–63 (2007).

80. Klitgaard, A. *et al.* Combining UHPLC-High Resolution MS and Feeding of Stable Isotope Labeled Polyketide Intermediates for Linking Precursors to End Products. *J. Nat. Prod.* **78**, 1518–1525 (2015).

81. Klitgaard, A., Nielsen, J. B., Frandsen, R. J. N., Andersen, M. R. & Nielsen, K. F. Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization. *Anal. Chem.* **87**, 6520–6526 (2015).

82. Kluger, B. *et al.* Stable isotopic labelling-assisted untargeted metabolic profiling reveals novel conjugates of the mycotoxin deoxynivalenol in wheat. *Anal. Bioanal. Chem.* **405**, 5031–5036 (2013).

83. Niedenführ, S., Wiechert, W. & Nöh, K. How to measure metabolic fluxes: A taxonomic guide for 13C fluxomics. *Curr. Opin. Biotechnol.* **34**, 82–90 (2015).

84. Weindl, D., Wegner, A. & Hiller, K. Metabolome-wide analysis of stable isotope labeling-Is it worth the effort? *Front. Physiol.* **6**, 189-3 (2015).

Chapter 2.

Design and Optimization of Parallel SIL Fermentation Method

2.1. Introduction

Stable isotope labeling (SIL) is a versatile technique that is utilized in many different biological and chemical applications. There are two main ways to apply SIL treatments to biological samples. You could replace an essential nutrient completely with a labeled version to generate a fully labeled metabolome. For example, replacing the sole carbon source in a minimal media with U-$^{13}C_6$ glucose for multiple generations results in a metabolome fully labeled by ^{13}C. Replacing an essential nutrient entirely with an SIL compound is powerful in that it allows the researcher to trace all metabolic processes involving that nutrient. This approach is used for metabolic flux measurements, quantification of known compounds, and untargeted discovery of novel metabolic pathways, however it is expensive to fully replace an essential nutrient with an SIL version.[1] It is also not necessary to fully label a pool of metabolites in order to gain insights from an SIL experiment.[1] Another approach is to add an SIL tracer molecule as a supplement to a culture medium which already contains all essential nutrients. This allows the researcher to add more specialized or expensive SIL precursors to target specific pathways as opposed to fully replacing a light element for a heavy one throughout the metabolome. Stable isotope labeling using amino acids in cell culture (SILAC) is commonly used where labeled amino acids are supplemented to the growth medium for applications in both metabolomics and proteomics.[2,3] Whether fully replacing an essential nutrient, or supplementing the culture with an SIL tracer, carefully controlled media conditions are used to ensure proper calculations of isotopic incorporation and reduce dilution of the SIL precursor by complex media constituents.

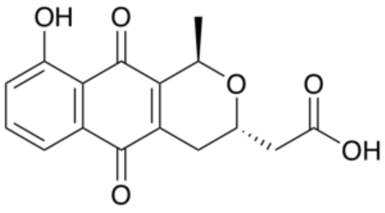
Although these studies do often use minimal media, not all SIL studies require knowledge of the exact isotopic ratios to draw useful biological conclusions. SIL has been used in the elucidation of specialized biosynthetic pathways that produce natural products by supplementing an excess of SIL tracer in complex growth media.[4] There is a balance between providing a medium that is rich enough to elicit compound production but simple enough to not interfere with the incorporation of the SIL precursor. Studies in polyketide biosynthesis for example often use a semi-rich media with aliquots of [1-^{13}C]acetate or [1,2-^{13}C]acetate supplemented into the media throughout the growth phase.[5] This is necessary to produce a sufficient amount of labeled material to fully purify the target compound and determine the positions of ^{13}C incorporation by NMR. There is no one set of media conditions and SIL precursors that will work for all target organisms or all BGCs. Metabolism by definition is highly flexible, and chemical phenotypes are known to be influenced more by small environmental changes than by genomic differences. In this chapter I discuss the preliminary experiments I used to develop the experimental design of the IsoAnalyst approach. This experimental design may be generally applicable to Actinobacteria, however, these experiments also demonstrate the overall considerations needed to optimize the IsoAnalyst method for any microorganism.

2.1.1. Considerations for Experimental Design

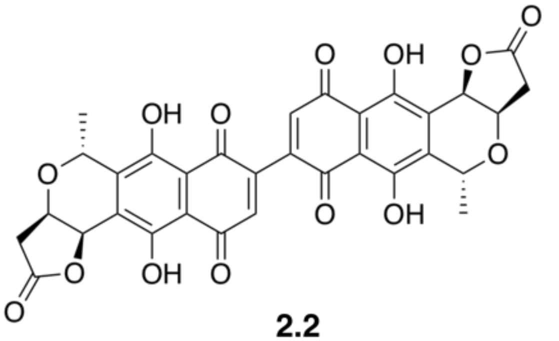
The purpose of this method is to label natural products on the basis of their biosynthetic pathways. Natural product structures are highly complex and are often derived from multiple specialized biosynthetic pathways. Microorganisms can harbor many different classes of BGCs, but most draw from the same pool of metabolites as primary metabolism for biosynthetic building blocks. The substrates of common biosynthetic classes such as NRPS, PKS, and terpenes can be partially deduced from the sequence of the BGC. I aimed to develop a generalizable method that could be adapted to cover many biosynthetic pathways, without expensive or multiply labeled tracer compounds. The main factors that went into developing this method were the media conditions, selection of SIL precursors, and selection of microorganisms to use for the optimization of this technique.

Three microorganisms are used throughout this chapter to test different media conditions and the SIL precursors. Two of these are model organisms that have been studied extensively for their natural product biosynthesis and one is a sequenced environmental isolate from our marine Actinobacteria isolate library. I performed SIL experiments in a variety of additional type strain organisms during the development of this fermentation protocol. The examples selected here generally demonstrate the factors that need to be considered when applying the IsoAnalyst technique to any target organism, and the basis for the optimized protocol used throughout the rest of this thesis.

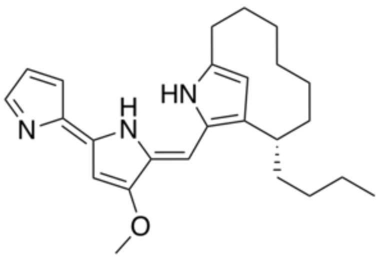
Streptomyces coelicolor M145 was selected due to it's well-studied metabolism and known production of a variety of natural products. *S. coelicolor* is a popular organism to test different environmental conditions due to the production of the colorful pigments actinorhodin and prodigiosin, which allow for the easy visual evaluation of metabolic changes. Four compounds produced by *S. coelicolor* are described in this chapter. Dihydrokalafungin (**2.1**) is the biosynthetic precursor to the polyketide actinorhodin including γ-actinorhodin (**2.2**), streptorubin B (**2.3**) is a PKS-NRPS hybrid, and desferrioxamine (**2.4**) is a common hydroxymate siderophore produced by many Actinobacteria (Figure 2.1). Other known compounds produced by *S. coelicolor* such as calcium dependant antibiotic, and coelimycin were not observed under any of the media conditions tested. *Saccharoplyspora erythraea* NRRL 23338 was used to test SIL incorporation into the macrolide antibiotic erythromycin A, and this model organism is used throughout this thesis as a proof-of-concept for the entire IsoAnlayst workflow. *S. erythraea* produces the polyketide antibiotic, erythromycin A (**2.5**, Figure 2.1), which has been studied extensively as a model for polyketide biosynthesis[6] and antibiotic drug discovery.[7] Finally, a *Micromonospora* sp. from the Linington Lab's marine Actinobacteria isolate library was used in the development of this method. This isolate was selected due to the fact that it produced a family of polyketides previously discovered in our lab, and this had prompted the complete sequencing of it's genome.[8] Having the full genome sequence of an organism is an important prerequisite for the development of this tool and offers discovery potential through genome mining. In this chapter I will only discuss the previously characterized lobosamide A (**2.6**, Figure 2.1) produced by *Micromonospora* sp. The complete strain designations for all microbial strains are indicated in the methods at the end of this chapter.



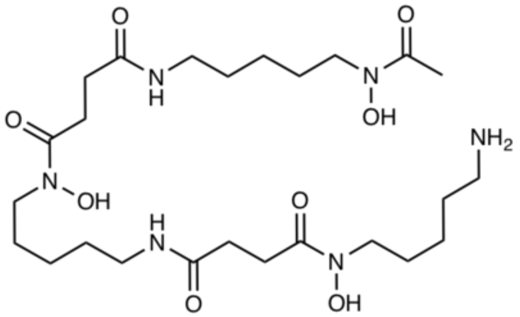
2.1
S. coelicolor



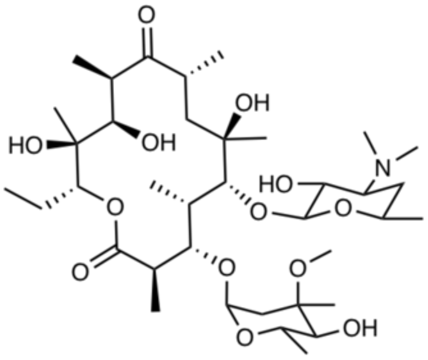
2.2
S. coelicolor



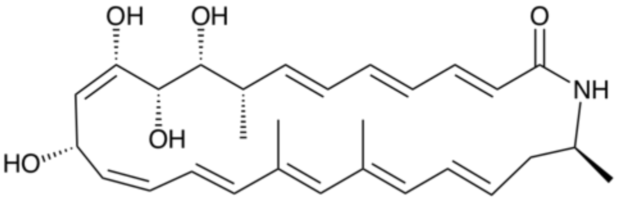
2.3
S. coelicolor



2.4
S. coelicolor



2.5
S. erythraea



2.6
Micromonospora sp.

Figure 2.1 Natural Products Produced by Model Organisms
Structures of natural products discussed throughout this chapter and names of the producing organisms. Dyhidrokalafungin (**2.1**), γ-actinorhodin (**2.2**), streptorubin B (**2.3**), and desferrioxamine B (**2.4**) are produced by *S. coelicolor*. Erythromycin A (**2.5**) is produced by *S. erythraea*. Lobosamide A (**2.6**) is produced by *Micromonospora* sp.

2.2. Media Optimization

Media conditions for growing bacteria isolated from marine sources have been tested extensively since the 1960s, including both complex and minimal media.[9–13] The main objectives of these experiments were to selectively isolate marine bacteria that had a high likelihood of producing natural products, and to optimize for the production of natural products. I am also aiming to enhance the production of different natural products, however, my central objective is to optimize the incorporation of SIL tracers into the natural products that are produced. Because there is no one growth medium that will optimize the production of all natural products an organism is capable of making, it is important to test a variety of conditions both for compound production and SIL incorporation. Specialized and primary metabolism are intricately connected, and those media constituents which are known to modify specialized metabolism also affect the central metabolic processes which govern transformations of the SIL tracer prior to incorporation into a natural product.[14,15]

The three major essential nutrients that need to be considered in a minimal media are carbon, nitrogen, and phosphate. Other essential nutrients include magnesium ($MgSO_4$) and iron ($FeSO_4$), but additional heavy metals and salts are often included as well. The Actinobacteria strains in our strain library were all isolated from marine sediments and were initially grown in a media made with a high salt concentration from a product called 'Instant Ocean' which contains the approximate salt mixture found in ocean water. This was intended for the isolation of marine obligate bacteria, however we cultivate many strains that are adaptable to more standard media without marine salts. The *Micromonospora sp.* isolated from our library was known to grow in GNZ media which does not contain Instant Ocean, so I aimed to develop a standard minimal media recipe with low salinity that could be generalized for all of the organisms discussed in this thesis. The ingredients for all media discussed are described in the last section of this chapter.

All of the experiments discussed in this section test the incorporation of [1-^{13}C]acetate into polyketide and NRPS compounds. I used this simple SIL tracer for the development of this technique due to its high applicability to a variety of natural products and its affordability as a simple and commonly used SIL tracer. [1-^{13}C]Acetate can be incorporated directly into polyketide products via malonyl-CoA, but can also be incorporated into other biosynthetic monomers such as amino acids through primary metabolism recycling (Figure 2.2). This makes [1-^{13}C]acetate a valuable feedstock for labeling a wide cross-section of natural product classes but it also complicates data interpretation, as I will discuss further in Chapter 4. Briefly, [1-^{13}C]acetate enters the TCA cycle as [1-^{13}C]acetyl-CoA resulting in ^{13}C incorporation into all TCA intermediates (Figure 2.2). When a ^{13}C labeled succinyl-CoA is converted to succinate, the position of the ^{13}C label becomes ambiguous due to the symmetry of succinate. Because of this, two positions have an equal chance of being labeled in the subsequent TCA cycle intermediates (Figure 2.2). Labeled oxaloacetate re-enters the TCA cycle resulting in up to two ^{13}C incorporation events in citrate, isocitrate, and α-ketoglutarate. The interconversion of succinate and succinyl-CoA leads to the indirect incorporation of ^{13}C into the C1 position of methylmalonyl-CoA as indicated by an open red circle while the direct labeling of the C4 position of methylmalonyl-CoA is lost to decarboxylation during the PKS condensation of methylmalonyl-CoA units (Figure 2.2). Some amino acids are derived from TCA cycle biosynthetic precursors, and therefore are able to be labeled by [1-^{13}C]acetate. [1-^{13}C]Acetate incorporation into amino acids was not detected in any of the experiments described in this chapter, and will therefore be discussed in more depth in Chapter 4.

Initially I optimized the media conditions to promote [1-^{13}C]acetate incorporation into polyketides due to the expected promiscuity of this SIL precursor incorporation into central pathways connected to the TCA cycle. I expected that the media conditions that significantly altered central metabolism would result not only in different natural product biosynthesis, but also different amounts of SIL tracer incorporation. The experiments presented in this chapter are not likely to produce the same results for every organism in every media. Media selection is an important part of optimizing and SIL tracer study to the test organism's metabolism, and the media conditions tested in this section represent a good starting place for applying SIL feeding to any microorganism.

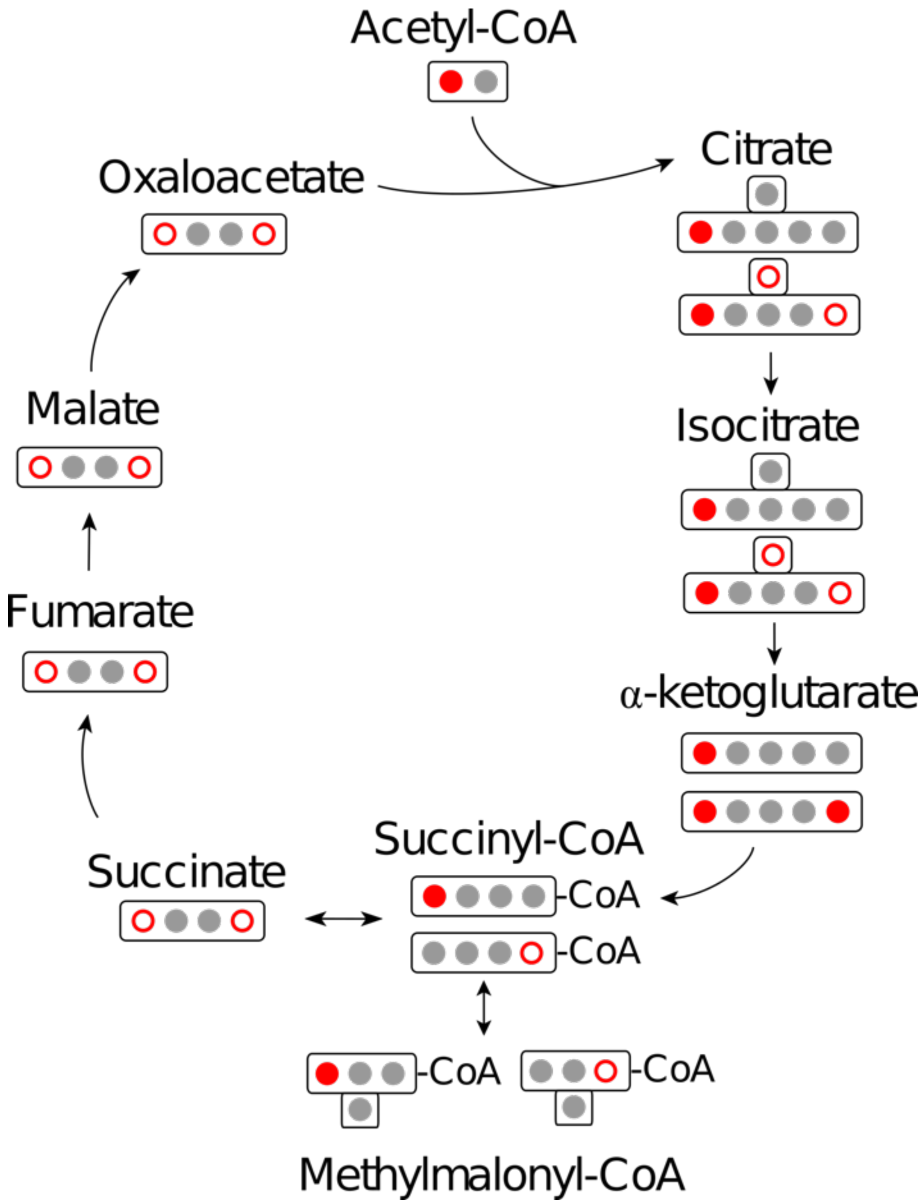


Figure 2.2 TCA Cycle Labeling by [1-^{13}C]Acetate
Red filled circles represent ^{13}C derived from [1-^{13}C]acetate following its direct transformation to acetyl-CoA. Open circles represent labeled carbon positions which are ambiguous due to compound symmetry. Molecules depicted with two open circles may have labeling in either position but not both. Groups of amino acids that are derived from TCA cycle biosynthetic precursors are indicated.

2.2.1. Carbon and Nitrogen Sources

Many members of the Actinobacteria phylum that are known for producing diverse natural products are found in soil environments. These environments are typically rich in different carbon sources from plant debris but are limited in nitrogen and phosphate.[14] Actinobacteria from both terrestrial soil and marine sediment are known to utilize a wide variety of simple and complex carbohydrates.[14] Carbon source influences antibiotic production, most notably by carbon catabolite repression which is known to repress the production of many natural products while the microorganism is consuming glucose as a preferred carbon source.[16] Many papers have been published looking at the selective use of carbon sources by microorganisms, and the influence of carbon source on the production of natural products.[10,17]

Nitrogen source also has a significant impact on natural product production and overall metabolism. Unlike carbon however, nitrogen is a limited resource in the natural soil environment of Actinobacteria. *Streptomyces* in particular are known to lack repression of amino acid biosynthesis pathways, such that most amino acids are being produced at all times in the bacterial lifecycle.[14] It has been suggested that the tight regulation of specialized metabolism allows for the catabolism of amino acids when a natural source of amino acids or other nitrogen source becomes available.[14] Because these bacteria have lost their ability to regulate much of the amino acid production, a sudden influx of a nitrogen source can result in the overproduction of certain amino acids. Without the ability to repress some of these pathways, natural product biosynthesis can act as a metabolic sink for the excess of specific amino acids. The tight regulation of specialized metabolism makes up for the lack of regulation of the primary metabolic pathways, and therefore can be heavily influenced by the availability of specific nitrogen sources.

2.2.2. Effects of Carbon and Nitrogen Sources on *Micromonospora sp.* Metabolism

The very first question I pursued as a graduate student was, ‘what minimal media conditions induce compound production and facilitate [1-^{13}C]acetate incorporation into polyketides produced by *Micromonospora sp.*?’ I began with the assumption that *Micromonospora sp.* may not produce lobosamides or any other natural product in most media, and so I aimed to test nutrients that were shown to influence natural product production. I grew *Micromonospora sp.* in a variety of media conditions with different carbon and nitrogen sources and screened for lobosamide A production using LCMS. I selected the media that were exhibiting the production of lobosamide A in *Micromonospora sp.*, and performed an experiment testing how these media conditions influenced [1-^{13}C]acetate incorporation into lobosamide A. Five types of media were tested using [1-^{13}C]acetate, and the full list of ingredients is given in Table 2.1 in the methods section at the end of this chapter. Four types of carbon sources were tested including glucose, maltose, sucrose, and soluble starch. All of these carbon sources were paired with $(NH_4)_2SO_4$ as a nitrogen source, while starch was tested with both $(NH_4)_2SO_4$ and glutamate due to the fact that both combinations produced lobosamide A. The basal media in all conditions was the same (Table 2.1). Observationally, I noticed that the starch media enhanced the growth of *Micromonospora sp.* regardless of which nitrogen source it was paired with, while the other carbon sources seemed to induce growth and lobosamide A production more when paired with the ammonium nitrogen source.

In this first experiment, I used the Linington Lab's standard extract library building protocol to extract and prepare the SIL samples for UPLC-MS analysis. Briefly, the bacterial cultures were grown in 60 mL of liquid minimal media and supplemented with 12 mM of either unlabeled acetate or [1-^{13}C]acetate. Unlabeled control cultures were grown in parallel with SIL supplemented cultures for each media condition. Cultures were grown for seven days while shaking at 200 rpm and extracted on the seventh day. A standard C18 column chromatography protocol was used to generate prefractionated extracts on the basis of polarity as described in the methods at the end of this chapter. Previous work done on lobosamide A production by this *Micromonospora sp.* strain had demonstrated that lobosamide A was present in the 80% and 100% methanol fraction and so these fractions were further processed for UPLC-MS analysis.

Figure 2.3 shows the MS results for lobosamide A in an unlabeled sample and in each of the five media conditions when labeled by [1-^{13}C]acetate. Clearly observable SIL incorporation was detected in the media with glucose and maltose carbon sources paired with $(NH_4)_2SO_4$ (Figure 2.3b,c) as well as in starch paired with glutamate (Figure 2.3f). The sucrose and starch carbon sources paired with $(NH_4)_2SO_4$ appeared to have very minor SIL incorporation when compared to the unlabeled condition (Figure 2.3d,e).

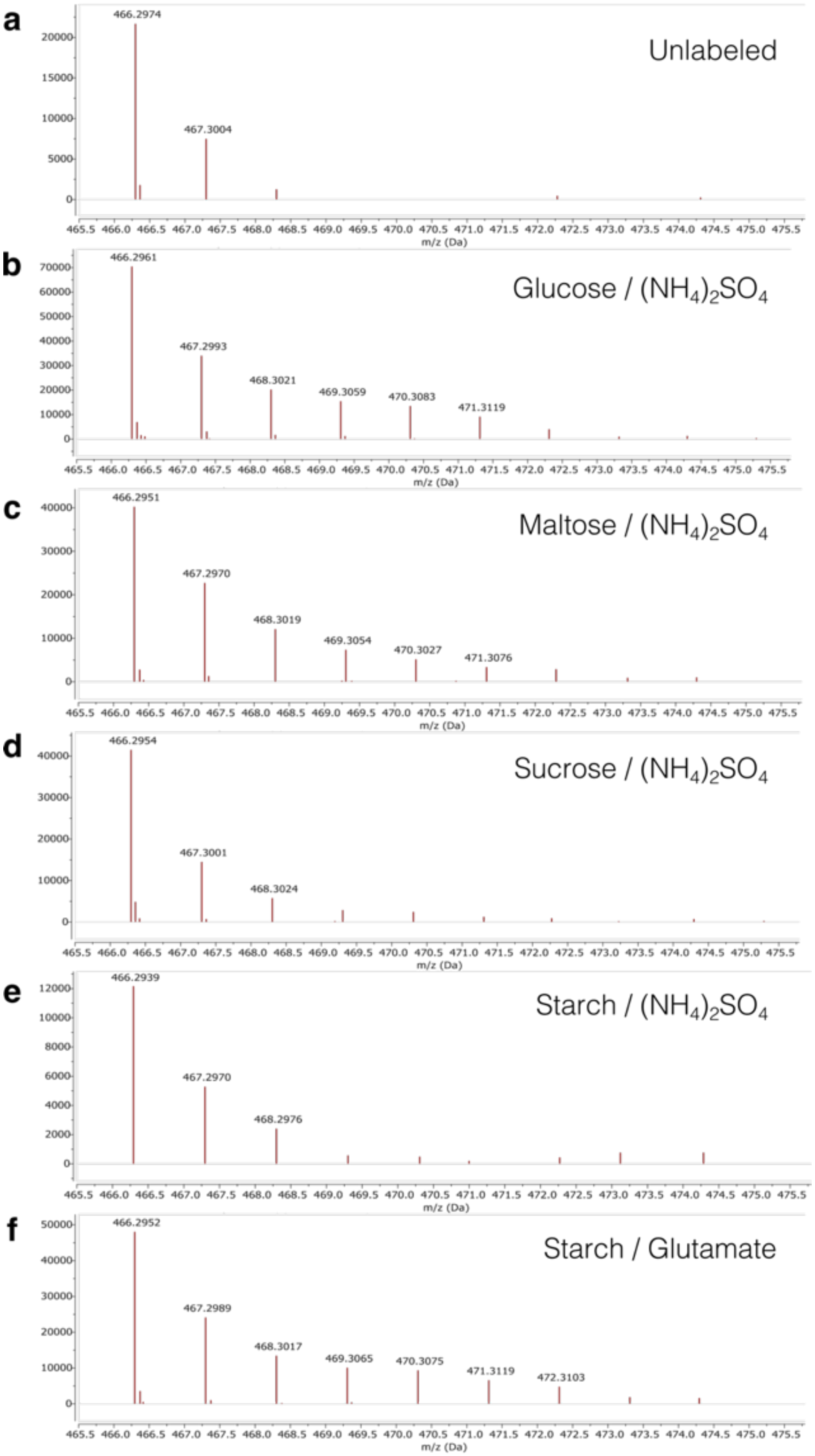


Figure 2.3 [1-^{13}C]Acetate Incorporation in Lobosamide A (2.6)
MS spectra of the $[M+H-H_2O]^+$ ion (*m/z* 466.29) of **2.6** in media containing different carbon and nitrogen sources. (a) Unlabeled MS spectrum of **2.6**. MS spectra of **2.6** labeled by [1-^{13}C]acetate in media containing $(NH_4)_2SO_4$ and one of four carbon sources (glucose (b), maltose (c), sucrose (d), starch (e)), and an additional media containing glutamate and starch (f).

These results show that carbon source affects the incorporation of [1-^{13}C]acetate into lobosamide A, but the difference between starch media containing glutamate or $(NH_4)_2SO_4$ indicates that nitrogen source influences [1-^{13}C]acetate incorporation as well. [1-^{13}C]Acetate incorporation was higher in the starch media containing glutamate compared to the starch media containing $(NH_4)_2SO_4$ (Figure 2.3e,f). Like other bacteria, Actinobacteria are known to prioritize nutrient consumption based on availability. This is central to carbon catabolite repression, which prioritizes glucose and other accessible carbon sources over more complex carbohydrate sources. Because starch is broken down into glucose, starch is also known to induce carbon catabolite repression, which may not be ideal for the production of many natural products. However, starch is used in many of the common growth media used for Actinobacteria and induced observably better growth in *Micromonospora sp.* and SIL incorporation in lobosamide A. Glutamate is readily used as a carbon source as well as a nitrogen source, and feeds directly into the TCA cycle by conversion to α-ketoglutarate. Media containing glutamate has been used to elicit natural product production, and the supply of two carbon sources that feed into the TCA cycle from different pathways improves overall growth and natural product production.[18] Labeling studies using *S. coelicolor* grown on media containing glucose and glutamate indicated that glutamate is a preferred carbon source, which heavily dominates the TCA cycle, while glucose-derived carbon dominates the pentose phosphate and glycolysis intermediates.[19] Because the starch-glutamate growth medium feeds carbon sources into the TCA cycle from two different pathways, a surplus of TCA cycle intermediates are likely generated. This surplus of TCA cycle intermediates may modify the central metabolic processes such that the [1-^{13}C]acetate tracer remains available for specialized metabolism.

Relying only on qualitative analysis, the glucose-ammonium, maltose-ammonium, and starch-glutamate media had essentially the same results, but the starch-glutamate medium may have slightly more SIL incorporation due to the fact that heavier isotopologues were detected with a relatively higher abundance (Figure 2.3f). I selected the starch-glucose medium moving forward because Actinobacteria are known for growing well and producing natural products in media containing complex carbohydrates such as starch[10] and because I observed more robust and consistent growth in the cultures containing starch. Starch is also present in both low salinity (GNZ) and high salinity media recipes used in our lab, indicating that starch is often an important nutrient for encouraging both bacterial growth and natural product biosynthesis.

2.2.3. Effects of Nitrogen Source on *S. coelicolor* Metabolism

The initial experiment performed with *Micromonospora sp.* indicated that a starch-glutamate minimal medium produced sufficient [1-^{13}C]acetate incorporation in lobosamide A, but changing the nitrogen source to $(NH_4)_2SO_4$ decreased [1-^{13}C]acetate incorporation. I aimed to test a starch-based minimal medium with different nitrogen sources in another microorganism to determine whether this growth medium was generalizable for [1-^{13}C]acetate incorporation in other Actinobacteria. *Streptomyces coelicolor* is a model organism that is known to produce a variety of natural products including blue and red pigments, actinorhodin and prodigiosin respectively.

In the subsequent experiments, I adopted a 24-well plate fermentation protocol that was optimized for Streptomyces strains to allow for screening of more conditions,[20] and simplified the extraction protocol. The well plates contain 2 mL of growth medium in each well, effectively decreasing the amount of SIL required and allowing for more media conditions and replicates to be tested. The complete optimized fermentation protocol is described in methods at the end of this chapter. For the extraction of the 24-well plate cultures, after five days of fermentation, I added an equal volume of methanol to each well, resulting in a 1:1 water:methanol extract. Each sample was sonicated and centrifuged prior to analysis by MS. This simplified protocol allows for a higher throughout of samples, at the cost of less compound coverage. As with any metabolomics experiment there will always be some bias in the compounds that are extracted and detected. I opted for a simple workup that would retain compounds of relatively low and high polarity, from the cells and the supernatant. It is important to note that although this general extraction method will work for many natural products, it should be optimized on the basis of target compound classes.

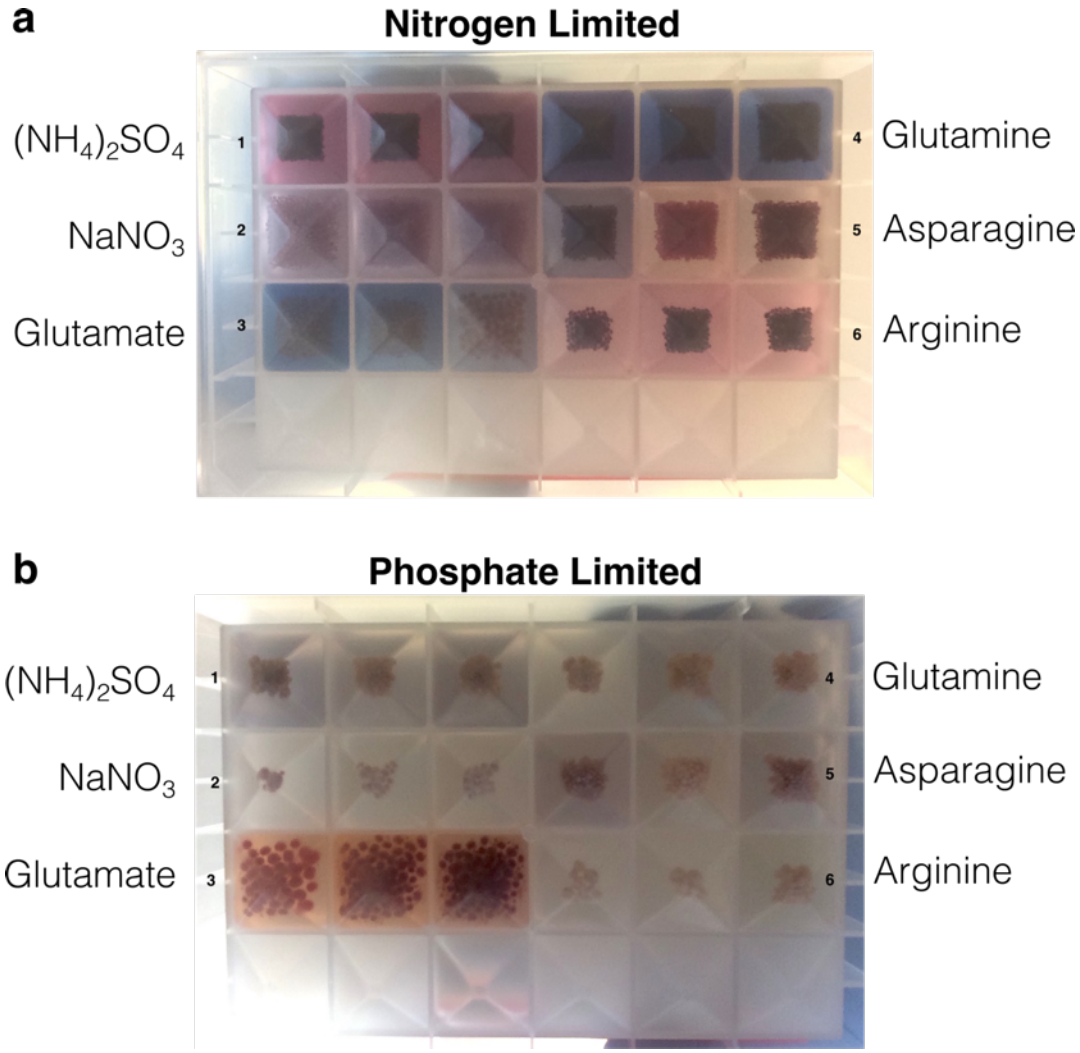


Figure 2.4 Photos of *S. coelicolor* Cultures in Different Minimal Media
S. coelicolor cultures grown in minimal media containing the same core ingredients but varying nitrogen sources. Three replicates of each media are shown. (a) Media containing limited nitrogen but excess phosphate. (b) Media containing excess nitrogen but limited phosphate.

Initially, I performed a qualitative experiment growing *S. coelicolor* in a starch based media with different nitrogen sources. *S. coelicolor* growth and natural product production is visually modified in different growth conditions, allowing for facile screening of media or other environmental changes. I tested *S. coelicolor* in 12 different minimal media with six different sources of nitrogen and either nitrogen or phosphate limitation. All of the minimal media contained starch as a carbon source (Table 2.2). Both nitrogen and phosphate limitation have both been shown to elicit different natural product production in various organisms.[9,11] This experiment was evaluated visually in order to narrow down ideal conditions to test [1-^{13}C]acetate incorporation. Figure 2.4 is a photograph showing the visual changes in *S. coelicolor* growth in response to both the nitrogen source, and the limiting nutrient in the media. Figure 2.4a shows the six different media when nitrogen is a limited nutrient, meaning the growth medium will run out of nitrogen before running out of phosphate. Figure 2.4b shows media containing the same ingredients, but with phosphate limitation and an excess of nitrogen. Generally, the nitrogen-limited media all induced varying pigmentation and sufficient growth, while most of the phosphate-limited conditions resulted in lower growth and little to no pigmentation (Figure 2.4). The phosphate-limited medium containing glutamate, however, did result in observable red pigmentation and significantly more biomass than the other phosphate-limited conditions (Figure 2.4). I selected three nitrogen-limited media from this panel to move forward and test [1-13C]acetate incorporation with *S. coelicolor*. I selected the $(NH_4)_2SO_4$, glutamate, and asparagine media, due to the fact that these three media resulted in visual differences in pigment production in the initial screening (Figure 2.4a).

I performed an experiment using the 24-well plate protocol and addition of 12 mM of [1-^{13}C]acetate in these three media and analyzed the samples by UPLC-MS according to the protocol described at the end of this chapter. The glutamate media again showed the best results across all the compounds detected in this experiment. Figure 2.5 shows the [1-^{13}C]acetate incorporation into the three putatively identified products γ-actinorhodin(**2.2**), streptorubin B (**2.3**), and desferrioxamine B (**2.4**) in the nitrogen-limited glutamate medium. Although the nitrogen-limited glutamate medium in Figure 2.4a was clearly dominated by blue pigmentation, both prodigiosin and actinorhodin were putatively identified with substantial labeling by [1-^{13}C]acetate (Figure 2.5).

The actinorhodins are dimeric benzoisochromanequinone polyketide antibiotics which are made up of 16 subunits of malonyl-CoA. There are various related actinorhodin analogues, but they are all biosynthesized by the dimerization of the octaketide dihydrokalafungin (**2.1**).[21] The only actinorhodin compound I was able to putatively identify in the nitrogen-limited glutamate media was γ-actinorhodin (**2.2**, Figure 2.5a). Although substantial incorporation of [1-^{13}C]acetate was detected (Figure 2.5a), the heaviest detectable isotopologue peak was M_9, indicating that not all sixteen available positions had SIL incorporation. Streptorubin B (**2.3**, Figure 2.5b) is an NRPS-

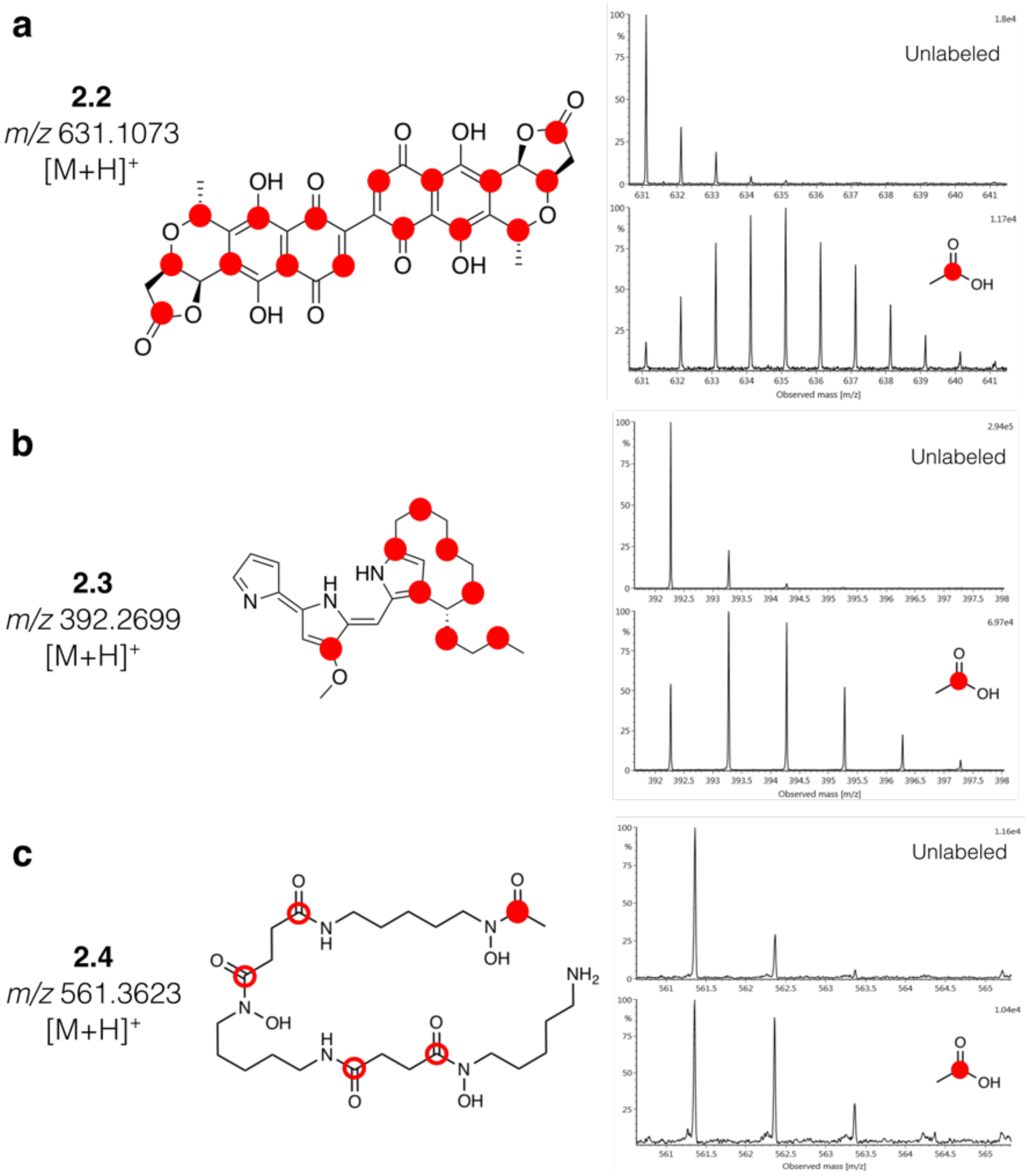


Figure 2.5 Mass Spectra of *S. coelicolor* Natural Products
Structures and mass spectra of (a) γ-actinorhodin (**2.2**), (b) streptorubin B (**2.3**), and (c) desferrioxamine B (**2.4**) in unlabeled and [1-^{13}C]acetate labeled extracts. Red circles represent positions which are expected to be labeled by [1-^{13}C]acetate. The open circles in **2.4** (c) represent alternate positions which can be labeled due to the symmetry of succinate (Figure 2.2).

PKS hybrid molecule which contains seven units of malonyl-CoA in the cyclized alkyl tail and one in the backbone.[22] There are eight carbon positions in streptorubin B (**2.3**) that are derived from malonyl-CoA and may be labeled by [1-^{13}C]acetate, but M_5 was the heaviest isotopologue detected (Figure 2.5b). Like γ-actinorhodin, streptorubin B had substantial incorporation of [1-^{13}C]acetate, but did have complete SIL incorporation in every available position. Desferrioxamine B (**2.4**) is a hydroxamate siderophore which also had detectable labeling by [1-^{13}C]acetate in one position (Figure 2.5c), likely corresponding to the acetyl-CoA subunit. Desferrioxamine B also contains succinyl-CoA subunits which can be labeled by [1-^{13}C]acetate indirectly through the TCA cycle, however these positions did not appear to have detectable SIL incorporation (Figure 2.5c).

Surprisingly, no streptorubin B or any other related prodigiosin compound was identified in the nitrogen-limited $(NH_4)_2SO_4$ medium despite the fact that this medium did clearly produce red pigmentation which can be observed visually (Figure 2.4a). Similarly, the nitrogen-limited asparagine medium did not produce detectable amounts of any desferrioxamines, prodigiosins, or actinorhodins. This was less surprising due to the overall less consistent growth and pigmentation observed in these cultures (Figure 2.4a), but does complicate the ultimate goal of comparing SIL incorporation across the different media conditions.

One compound was identified in all three media conditions (Figure 2.6). Dihydrokalafungin (**2.1**) is a biosynthetic precursor to the polyketide actinorhodin. Although γ-actinorhodin was only putatively identified in the glutamate medium (Figure 2.5a), other actinorhodin analogues were likely produced in the $(NH_4)_2SO_4$ and asparagine media due to the fact that the precursor **2.1** was produced. I was not able to detect any actinorhodin compounds in the $(NH_4)_2SO_4$ or asparagine media, which could be accounted for by multiple factors. A single extraction technique was used on all samples which may not be ideal for all actinorhodin analogues and the overall titer of actinorhodin production in the $(NH_4)_2SO_4$ and asparagine media may not be sufficient for detection by MS. There are additional biological factors such as the regulation of the biosynthetic pathway or degradation of the final product. Due to these inconsistencies, **2.1** was the only compound for which I was able to directly compare [1-^{13}C]acetate

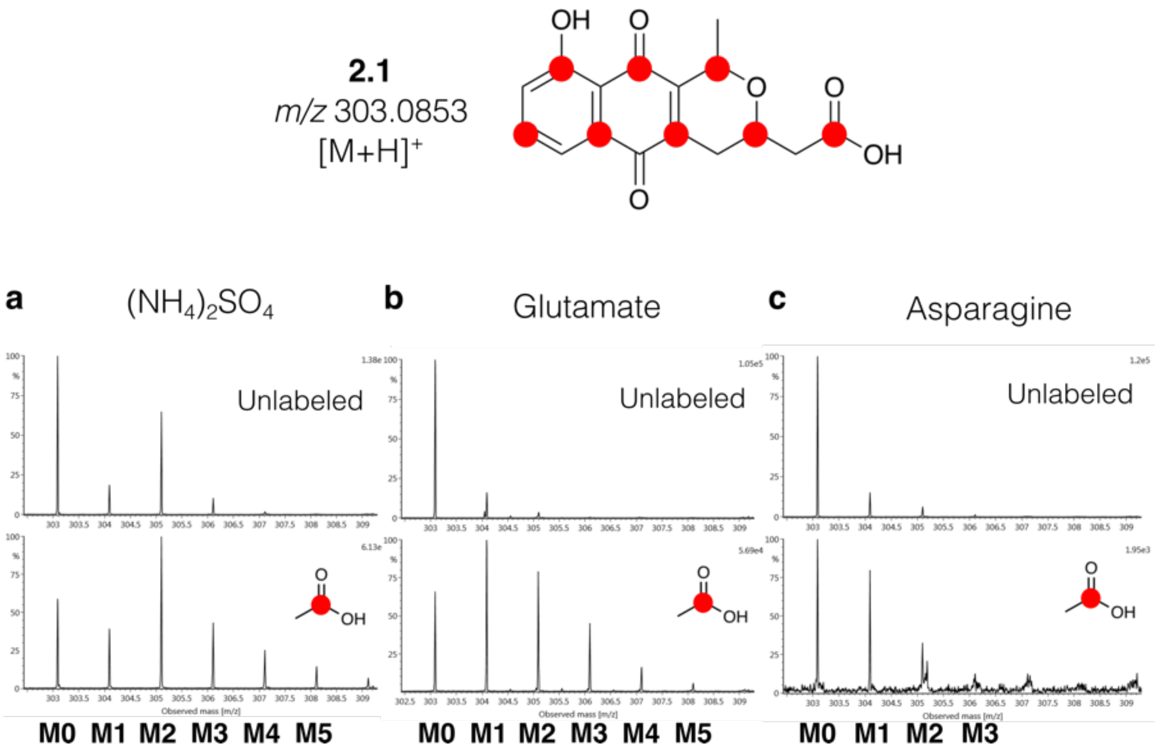


Figure 2.6 Mass Spectra of 2.1 in Minimal Media

Mass spectra of dihydrokalafungin (**2.1**) in minimal media containing $(NH_4)_2SO_4$ (a), glutamate (b), or asparagine (c) as a nitrogen source. The red circles represent positions that are expected to be labeled by [1-^{13}C]acetate through malonyl-CoA or acetyl-CoA.

incorporation (Figure 2.6). More SIL incorporation was detected in **2.1** in the glutamate medium than either the $(NH_4)_2SO_4$ or asparagine media (Figure 2.6). This is clear from two observations. The M_1 isotopologue peak is the base peak in the labeled spectrum of **2.1** in the glutamate medium, compared to the other media which maintain a higher relative intensity of the monoisotopic M_0 peak compared to the M_1 isotopologue peak (Figure 2.6). Additionally, isotopologue peaks up to M_5 were detected in the glutamate medium, indicating a higher degree of labeling than the other conditions. Interestingly, the $(NH_4)_2SO_4$ medium produced a DHKF spectrum with the M_2 isotopologue as the base peak. This highlights the importance of observing the unlabeled control MS spectrum side-by-side with the labeled spectrum. The M_2 isotopologue peak also had a higher intensity in the unlabeled control (Figure 2.6), indicating that this isotopologue peak is overlapped with the monoisotopic M_0 peak of a related analogue, likely the reduced form of **2.1**. The overall isotopologue distribution in the labeled spectrum of **2.1** in the $(NH_4)_2SO_4$ medium indicates that both the oxidized and reduced forms of **2.1** have some [1-^{13}C]acetate incorporation. However, the overlap of these signals complicates the interpretation of the SIL incorporation. It is likely that the mixture of oxidation states in this medium results from the ammonium content in the media, making this medium less ideal for interpreting [1-^{13}C]acetate in this class of compounds. [1-^{13}C]acetate incorporation was also observable in **2.1** in the asparagine condition, but the lower signal to noise and smaller distribution of isotopologue peaks indicates that this medium is also not ideal for [1-^{13}C]acetate detection in **2.1**. The incorporation of [1-^{13}C]acetate into three additional compounds in the glutamate condition (Figure 2.5) supports the selection of this media because of the consistency of SIL incorporation across three different natural products. Although none of these compounds exhibited full [1-^{13}C]acetate incorporation into all available positions, the consistency of these results was encouraging evidence that the nitrogen-limited media containing glutamate as the sole nitrogen source was a good general medium for conducting SIL studies in other Actinobacteria besides *Micromonospora sp*.

These results also indicated that *S. coelicolor* is not an ideal model organism for use in this study because its growth in the minimal media was highly inconsistent in follow-up experiments. Following these experiments, I struggled to maintain consistent growth with the *S. coelicolor* cultures. Figure 2.4 shows the clumpy sort of growth observed across these cultures, which is common for *S. coelicolor* but appeared to be exaggerated in the minimal media recipes compared to rich media. This observation paired with the inconsistent production of the target compounds led me to move away from using *S. coelicolor* as a model organism. Still, the results here support the notion that media selection has a significant influence on the experimental outcome, and that the nitrogen-limited media containing starch and glutamate enhanced [1-^{13}C]acetate incorporation into polyketide compounds.

2.2.4. Phosphate and Nitrogen Limitation in *Micromonospora* sp.

In the preliminary media experiment using *S. coelicolor*, I observed significant differences in metabolism and growth between nitrogen and phosphate limited conditions (Figure 2.4). I selected the nitrogen-limited media to test [1-^{13}C]acetate incorporation because nitrogen limitation induced more consistent production of observable natural products in *S. coelicolor*, however phosphate limitation has also been shown to improve natural product production in other cases.[11] Excess of either nitrogen or phosphate tends to delay and decrease the production of most natural products.[9] Many different combination and concentrations of nutrients have been tested in terms of optimizing natural product production, and limiting nitrogen or phosphate is commonly applied to influence specialized metabolism. Phosphate in particular controls many aspects of central metabolism including RNA, DNA, and protein synthesis as well as cellular respiration and ATP levels.[11] Testing different combinations of nitrogen sources with both nitrogen and phosphate as the limiting nutrient can not only induce different natural product biosynthesis as shown with *S. coelicolor*, but also influences central metabolism and therefore the efficiency of [1-^{13}C]acetate incorporation.

I tested the effect of phosphate and nitrogen limitation on [1-^{13}C]acetate incorporation into lobosamide A. I also tested glutamate and glutamine as different nitrogen sources. Glutamine is metabolically related to glutamate, but keeping the molar amount of nitrogen equal in the media results in a smaller molar amount of carbon coming from glutamine as opposed to glutamate. I qualitatively tested how the combination of nitrogen source and nitrogen or phosphate limitation affected [1-^{13}C]acetate incorporation into lobosamide A. I also increased the concentration of [1-^{13}C]acetate in the cultures to 30 mM in order to increase the SIL incorporation compared to the previous experiment shown in Figure 2.3. The basal media for this experiment contained starch as the primary carbon source and the same salt and metal content as the previous experiment (Table 2.2).

Both of the media containing glutamate as the nitrogen source demonstrated a slightly higher degree of [1-^{13}C]acetate labeling compared to the media containing glutamine (Figure 2.7). This observation is based on the general shift of the isotopologue distribution towards heavier isotopologues in the glutamate condition compared to glutamine, although all four media conditions had detectable intensity for the M_9 isotopologue peak of lobosamide A. The phosphate limited conditions generally had less SIL incorporation that the nitrogen limited media containing the same nitrogen source (Figure 2.7). However, the phosphate limited media containing glutamate has a more intense M_0 peak than the nitrogen limited media containing glutamine, indicating that the total SIL incorporation is actually less in the nitrogen limited - glutamate media. Even though the whole pool of lobosamide A ions in the phosphate-limited glutamine media have more SIL incorporation, the extent of [1-^{13}C]acetate incorporation is lower in the heavy isotopologues M_6, M_7, M_8, and M_9 (Figure 2.7). This is an important observation to consider in terms of the aims of the experiments presented here. Many SIL studies focus on the flux of an SIL tracer through a system or the fraction of labeled carbon present in a pool of labeled compounds. However, in this work I aim to use SIL precursors to infer the number of biosynthetic building blocks derived from that precursor. The optimization of this experiment aims to detect the heaviest labeled isotopologue peak, regardless of the overall efficiency of SIL incorporation. When comparing the phosphate-limited glutamate condition and the nitrogen-limited glutamine condition, the phosphate-limited glutamate condition is more optimal because it results in a high extent of labeling even though the pool of ions has a lower overall percentage of ^{13}C incorporation.

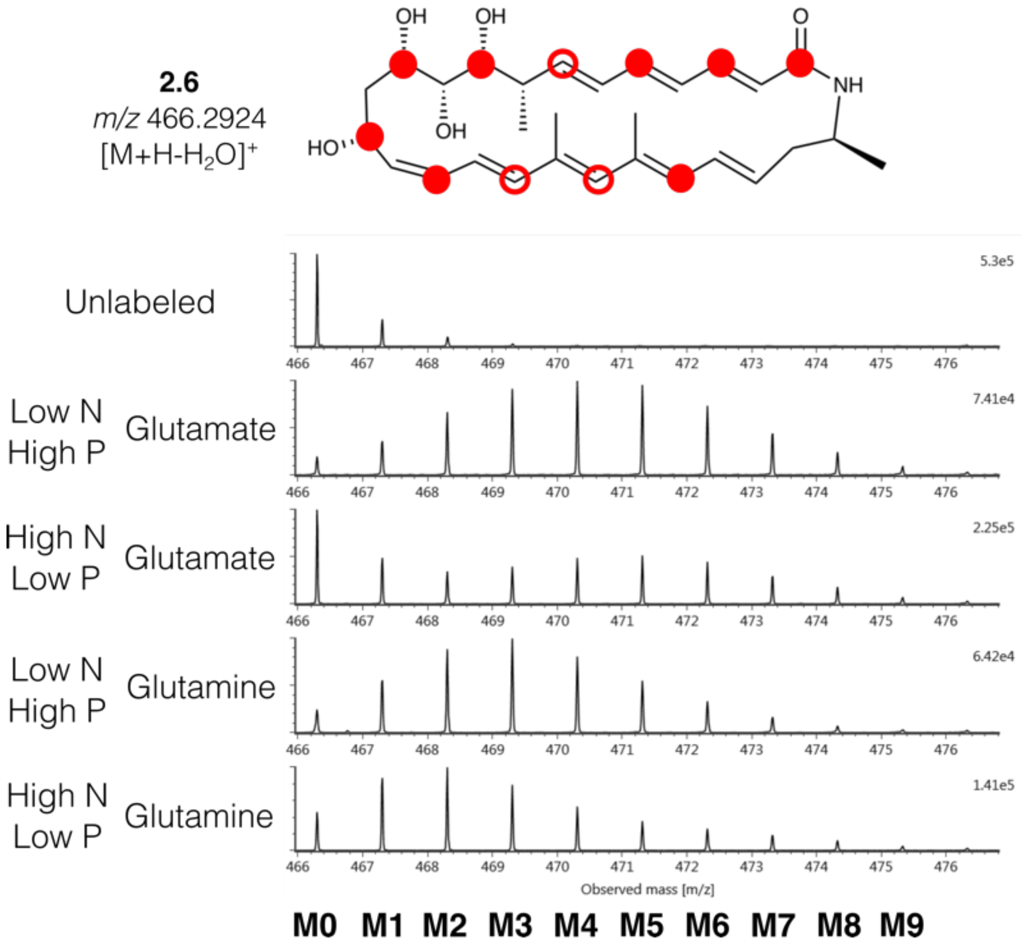


Figure 2.7 [1-^{13}C]Acetate Incorporation in 2.6 in Nutrient Limited Media
Mass Spectra of lobosamide A (**2.6**) in minimal media containing either glutamate or glutamine as a nitrogen source, and limitation in either nitrogen or phosphate. The filled red circles represent positions that are expected to be labeled by [1-^{13}C]acetate through malonyl-CoA or acetyl-CoA. The open red circles represent positions that may be indirectly labeled by [1-^{13}C]acetate through conversion to methylmalonyl-CoA (Figure 2.2).

Overall, these experiments show that the nitrogen-limited starch-glutamate media was optimal for both *Micromonospora sp.* and *S. coelicolor* in terms of compound production and [1-^{13}C]acetate incorporation into target polyketide products. I use this minimal growth medium in the subsequent SIL experiments described throughout this thesis, further demonstrating that this media recipe is generally effective for SIL incorporation in a variety of Actinobacteria. Although this medium is a good option, generally speaking, for the Actinobacteria presented in this thesis, the data presented here indicate that optimizing minimal media conditions for the test organism in terms of both compound production and SIL incorporation is crucial to success.

2.3. SIL Precursor Selection

In addition to [1-^{13}C]acetate, I selected three other SIL precursors to allow for association of MS signals with BGCs. IsoAnalyst was developed to permit flexibility in terms of SIL precursor selection. Any type of SIL tracer can be used in the IsoAnalyst method, but the goal of this approach is to minimize the cost of SIL tracers by using simple building blocks that are isotopically labeled in a single position. Theoretically, multiply labeled precursors can be used in IsoAnalyst, however, algorithms for detecting multiply labeled precursors already exist and are well-established. One novel aspect of IsoAnalyst is to detect the number of SIL precursors incorporated into a natural product structure, rather than identify SIL incorporation by large mass shifts associated with specific functional groups. I will discuss this aspect of the data analysis more in Chapter 3, but the SIL precursors selected here reflect this objective. With the ultimate goal of detecting the number of biosynthetic units in a compound, ideal SIL tracer molecules are those which are incorporated directly into natural product pathways but are not so specific that they only target a few BGCs. The SIL tracers described here were selected based on their general applicability to the model organisms used and can be used for first-pass screening of biosynthetic potential in any organism, but are by no means optimized for the full coverage of all classes of BGCs.

To demonstrate the scope of IsoAnalyst across a broad cross section of biosynthetic classes, I selected four SIL precursors; [1-^{13}C]acetate, [1-^{13}C]propionate, [methyl-^{13}C]methionine, and [1-^{15}N]glutamate. [1-^{13}C] Propionate is predominantly used for identifying polyketide pathways, due to its conversion to methylmalonyl-CoA, a common building block in type I and type II polyketide biosynthesis. [methyl-^{13}C]Methionine was selected in order to label compounds methylated by *S*-adenosyl methionine (SAM). Methylation by SAM is encountered relatively frequently in natural product biosynthesis and therefore assists in prioritizing compounds which have a high likelihood of being natural products, especially if they are also labeled by one or more of the other SIL precursors. Finally, [1-^{15}N] glutamate was selected to constitutively label all products containing nitrogen atoms. To accomplish this, 50% of the total available nitrogen was labeled in the form of [1-^{15}N] glutamate, with the remainder deriving from standard unenriched glutamate. The growth medium and procedures used in this experiment are described in the methods section at the end of this chapter.

2.3.1. Parallel SIL Incorporation in Erythromycin A

The biosynthesis of erythromycin A has been studied extensively as a model system for modular polyketide synthases.[23] The macrolide core is formed by the condensation of one propionyl-CoA and six methylmalonyl-CoA units followed by glycosylation with the saccharides desosamine and mycarose (Figure 2.8). [1-^{13}C]Propionate is directly converted to the substrates propionyl-CoA and methylmalonyl-CoA, however the latter may also be labeled by [1-^{13}C] acetate through conversion to the TCA cycle intermediate succinyl-CoA (Figure 2.2). The indirect labeling of methylmalonyl-CoA by [1-^{13}C] acetate is included in the labeling prediction as we cannot differentiate analytically between direct and indirect SIL incorporation (Figure 2.8a). Desosamine contains a tertiary dimethylamino group which is expected to be labeled by a single position in the [1-^{15}N]glutamate condition, and two positions in the [methyl-^{13}C] methionine condition. The mycarose unit has a single methylation position and is later methylated by an O-methyltransferase following attachment to the macrolide core (Figure 2.8a).

I optimized the concentration of each SIL precursor by testing a range of concentrations as described in the methods section at the end of this chapter. The final concentrations used for precursor were 30 mM [1-^{13}C]acetate, 30 mM [1-^{13}C]propionate, 5 mM [methyl-^{13}C]methionine, and 10 mM [1-^{15}N]glutamate. The expected labeled positions of erythromycin A and the MS spectra in each SIL condition are shown in Figure 2.8b. Substantial labeling was observed in all four conditions. In particular, [1-^{13}C]propionate, [methyl-^{13}C]methionine, and [1-^{15}N]glutamate qualitatively appear to have complete SIL incorporation in every available position (Figure 2.8b). [1-^{13}C]Acetate did not label erythromycin A as extensively as anticipated, however [1-^{13}C]acetate is only incorporated into erythromycin A indirectly through the conversion of succinyl-CoA to methylmalonyl-CoA (Figure 2.2).

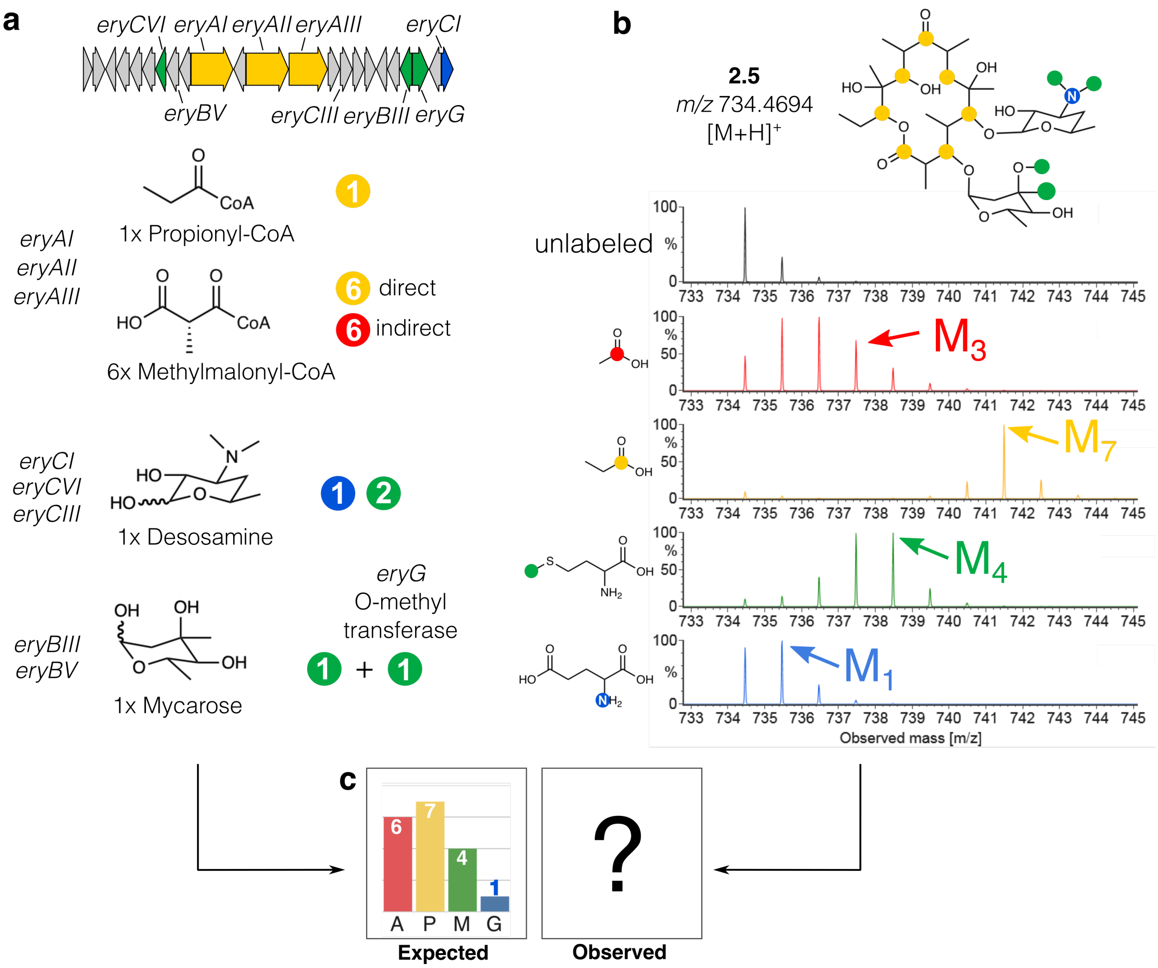


Figure 2.8 Biosynthesis and SIL Tracer Incorporation in 2.5

(a) The erythromycin BGC, expected biosynthetic precursors, and the expected SIL incorporation in those precursors. (b) Structure of erythromycin A (**2.5**) with positions of SIL incorporation indicated. Mass Spectra of **2.5** showing labeling by [1-^{13}C]acetate, [1-^{13}C]propionate, [methyl-^{13}C]methionine, and [1-^{15}N]glutamate. Arrows indicate the heaviest isotopologue peak that visually appears to have SIL enrichment. (c) Comparison of expected SIL incorporation on the basis of the BGC to observed SIL incorporation. Although visual inspection implies sufficient labeling, a statistical technique for determining the true SIL incorporation will be described in the next chapter.

The heaviest isotopologue peaks that are clearly enriched are indicated in Figure 2.8b by arrows, however it is not possible to tell by observation alone if these peaks are enriched with heavy isotopes derived from the SIL precursor, or from naturally occurring ^{13}C. Figure 2.8c shows the expected labeling on the basis of interpreting the BGC, but comparable information cannot be derived from observing the MS spectra alone. In Chapter 3 I will discuss the statistical approach used to interpret these data, however this fermentation and SIL precursor addition protocol were successfully optimized on the basis of this qualitative data analysis.

2.4. Optimized Parallel SIL Feeding Experiments in Minimal and Rich Media

Having optimized this experiment for the labeling of erythromycin A, and confirmed that the SIL incorporation qualitatively matched what was expected on the basis of the BGC, I aimed to test the full SIL panel in *Micromonospora sp.* I performed the same SIL experiment described previously for *S. erythraea* using *Micromonospora sp.* in the nitrogen-limited starch-glutamate minimal media. I looked at lobosamide A again to assess the outcome of this SIL experiment. Lobosamide A is a polyene macrolactam polyketide, which is biosynthesized from six units of malonyl-CoA, three methylmalonyl-CoA, and a 3-aminobutryate starter unit which is derived from glutamate.[8] I performed an additional experiment with *Micromonospora sp.* using the rich media GNZ which was previously used for the large-scale fermentation of this organism in order to isolate lobosamide A and related analogues.[8] The ingredients for the GNZ media are described in the last section of this chapter. I used higher concentrations of each SIL tracer in the GNZ media to ensure sufficient incorporation (100 mM [1-^{13}C]acetate, 100 mM [1-^{13}C]propionate, 16.5 mM [methyl-^{13}C]methionine, and 100 mM [1-^{15}N]glutamate).

Both the minimal medium and rich medium resulted in substantial labeling in the [1-^{13}C]acetate, [1-^{13}C]propionate, and [1-^{15}N]glutamate conditions. Surprisingly, the [1-^{13}C]acetate incorporation was lower in the rich medium, while [1-^{13}C]propionate incorporation was higher in the rich medium. These results demonstrate that this parallel SIL feeding approach can not only be applied in rich media, but in some cases may result in more complete incorporation of certain SIL precursors. This underscores the importance of testing a variety of media conditions and the utility of applying this technique in combination with different environmental stimuli to induce natural product biosynthesis.

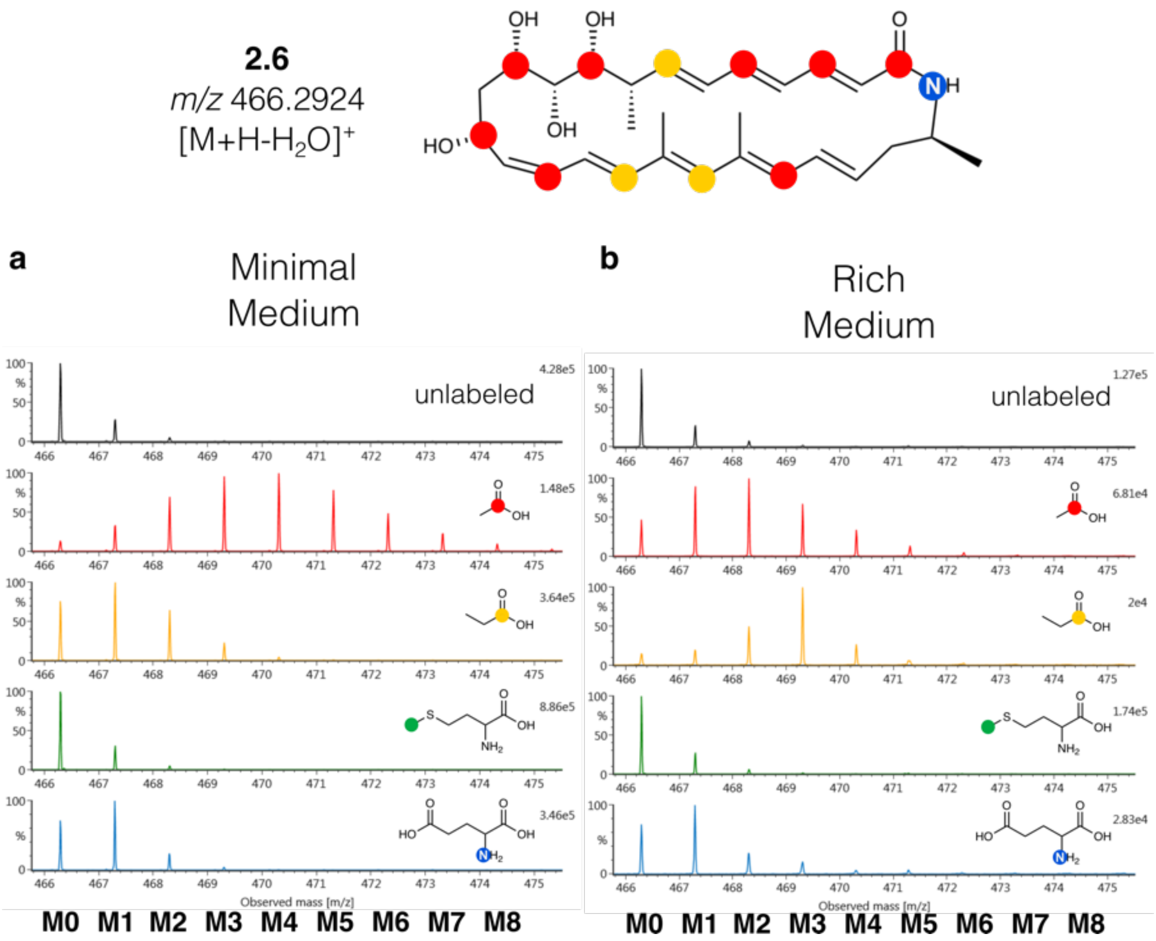


Figure 2.9 SIL Incorporation in 2.6 in Mimimal and Rich Media
Mass Spectra showing SIL Incorporation in lobosamide A (**2.6**) in nitrogen-limited starch-glutamate minimal medium (a) and the rich growth medium GNZ (b). More SIL incorporation was observed for [1-^{13}C]acetate in the minimal media, while more SIL incorporation was observed for [1-^{13}C]propionate in the rich media, indicating that there is not likely a single media condition that will be optimal for every SIL tracer used.

2.4.1. Significance of Fermentation Optimization

In this chapter I showed the importance of testing how major nutrients affect target compound production, as well as influence SIL incorporation into target compounds as an initial optimization step. The results of these experiments indicate that the minimal medium and SIL conditions developed in this chapter are generally applicable to my model organisms. Because of these promising results, I used the nitrogen-limited starch-glutamate minimal media and the SIL tracer concentrations described in this chapter throughout the remainder of this thesis. However, from working with minimal media and testing various conditions it was clear to me that the process of making and testing media conditions is tedious and often leads to discouraging results.

Particularly in the case of *S. coelicolor*, I was not able to induce detectable natural product biosynthesis in many of the media conditions. Even in the model organisms used, minimal media does not result in the production of more than one or two compound families. Furthermore, screening fermentation conditions typically involves rich media to induce compound production, and a method which relies solely on minimal media may not have the same potential for discovery due to this limitation. Still, the flexibility of this approach allows for many fermentation systems to be tested for compound discovery in parallel, and further optimized following identification of labeled MS signals.

When screening for unknown compounds, a variety of media conditions may be necessary in order to find the optimal conditions for the organism, compound, and selected SIL precursors. Alternatively, a starch-glutamate based media may be a good starting point, especially for labeled acetate incorporation, which is involved in many central metabolic processes. Beginning with a simple medium one can also perform a series of challenge experiments using heavy metals or antibiotics to induce different natural product biosynthesis while ensuring observable SIL incorporation. Complex media can be used as well however this requires a higher concentration of SIL tracers due to the high availability of different nutrients including amino acids, nucleic acids, and carbohydrates. Complex media are also likely to result in less reproducibility between batches because the ingredients are not exactly specified. Despite these caveats, a combination of minimal and rich media can easily be tested in parallel to optimize this method for any microorganism.

The data in this chapter were analyzed qualitatively for the purpose optimizing the fermentation conditions and SIL tracers. This was adequate for screening media and SIL conditions, however, in Chapter 3 I will describe the analytical platform I developed using Python 3 to statistically determine the number of SIL precursors incorporated into every MS feature in the parallel SIL metabolomics experiments presented here.

2.5. Methods

2.5.1. Strain Designations

Saccharopolyspora erythraea ATCC 11635 (NRRL 2338) and *Streptomyces coelicolor* ATCC BAA-471 (A3(2) / M145) were purchased from ATCC (USA). *Micromonospora* sp. RL09-050-HVF-A was isolated and sequenced as described in Schulze et al.[8] The *Micromonospora* sp. RL09-050-HVF-A genome was uploaded to NCBI under the accession number JAGKQP000000000 and the BioProject ID PRJNA718589.

2.5.2. Media Recipes

The minimal media recipes used for SIL experiments in this chapter are shown in Table 2.1 and Table 2.2. Variation in phosphate, nitrogen, and carbon sources and concentrations are also indicated. Table 2.1 shows the media recipes used in section 2.2.1 for preliminary experiments with *Micromonospora* sp. The basal medium in Table 2.2 was used with the microtiter plate protocol for the remainder of the experiments in this chapter.

Table 2.1 Minimal Media for Nitrogen and Carbon Sources

Basal Medium	Concentration
$MgSO_4$	2.0 g/L
$CaCO_3$	1.5 g/L
$FeSO_4$	1.5 mg/L
$CuSO_4$	1.5 mg/L
$ZnSO_4$	1.6 mg/L
Phosphate	
K_2HPO_4	1.5 g/L
Carbon	
Starch	40 g/L
Glucose	40 g/L
Maltose	40 g/L
Sucrose	40 g/L
Nitrogen	
Glutamate	5 g/L
$(NH_4)_2SO_4$	5 g/L

The complex medium GNZ (10 g glucose, 20 g starch, 5 g N-Z-amine, 5 g yeast extract, 1 g CaCO3, and 14 g agar per liter of water) was used to grow *Micromonospora* sp. seed cultures in all experiments, and as the primary growth medium in the SIL experiment in section 2.4. The complex medium ISP (3 g yeast extract, 5 g acid hydrolyzed casein, and 14 g of agar per liter of water) was used to grow seed cultures of *S. coelicolor* and *S. erythraea*. Both GNZ and ISP culture plates were made using 15 g/L of agar for streaking frozen stocks of all bacterial strains.

Table 2.2 Minimal Media for Nitrogen and Phosphate Limitation

Basal Medium	Concentration	
NaCl	2.0 g/L	
$MgSO_4$	1.0 g/L	
$CaCO_3$	1.0 g/L	
$FeSO_4$	0.01 g/L	
$CuSO_4$	1.5 mg/L	
$ZnSO_4$	3.0 mg/L	
$CoSO_4$	0.15 mg/L	
$MnSO_4$	0.15 mg/L	
Na_2MoO_4	1.0 mg/L	
Carbon		
Starch	10 g/L	
Phosphate	**4 mM**	**10 mM**
KH_2PO_4	0.16 g/L	0.41 g/L
K_2HPO_4	0.49 g/L	1.22 g/L
Nitrogen	**60 mM**	**20 mM**
Glutamate	10.2 g/L	3.38 g/L
Asparagine	7.93 g/L	2.64 g/L
$(NH_4)_2SO_4$	3.96 g/L	1.32 g/L

2.5.3. Medium Scale Fermentation and Extraction

Initial experiments testing nitrogen and carbon source with the *Micromonospora sp.* were carried out using the standard growth and extraction protocol used for building the Linington Lab's marine Actinobaterial natural product extract library. *Micromonospora* sp. was grown in culture flasks containing a metal spring and 60 mL of each type of media indicated in Table 2.1. Every carbon source was paired with $(NH_4)_2SO_4$, but only starch was paired with glutamate as a nitrogen source. The selection of these media for fermentation with [1-^{13}C]acetate was made on the basis of preliminary time-course experiments done in small scale for the detection of lobosamide A. Each culture flask also contained 1.2 g of Amberlite XAD-16 adsorbant resin to assist with the extraction of natural products from the culture medium. Each 60 mL culture was inoculated with 3 mL of seed culture grown in GNZ media for three days. An unlabeled control culture containing 12 mM of unlabeled acetate was grown in parallel to experimental cultures containing 12 mM of [1-^{13}C]acetate. Cultures were shaken at 200 rpm for seven days and then extracted and fractionated by polarity.

Each 60 mL culture was first filtered by vacuum filtration and the cell debris and resin were then extracted in a 1:1 mixture of methanol and dichloromethane for one hour. The organic extract was collected by vacuum filtration and dried by rotary evaporation. The crude organic extract was initially separated into seven fractions by a stepwise methanol/water elution (10, 20, 40, 60, 80, 100 vol/vol) and an additional ethyl acetate wash step on a RediSep Rf C18 cartridge (Teledyne Isco) using a Teledyne Isco CombiFlash Rf flash chromatography system. Lobosamide A is found in the 80% methanol and 100% methanol fractions, so these fractions were prepared for UPLC-MS analysis for each media condition control and [1-^{13}C]acetate culture. Samples were prepared by first dissolving each pre-fractionated extract in 1 mL of methanol and diluting the sample by 1:200 into 50% (vol.vol) methanol/water.

2.5.4. Microtiter Plate Fermentation with SIL Tracers

The SIL feedstock compounds, [99% 1-^{13}C]acetate, [99% 1-^{13}C]propionate, [99% methyl-^{13}C]methionine, and [98% 1-^{15}N]glutamate, and unlabeled version of each compound were prepared as stock solutions in Milli-Q water and sterilized by filtration (0.2 uM filter). Bacterial inoculum was prepared by first streaking a frozen stock on a GNZ agar plate (10 g glucose, 20 g starch, 5 g N-Z-amine, 5 g yeast extract, 1 g CaCO3, and 14 g agar per liter of water) for *Micromonospora* sp. and ISP agar plate (3 g yeast extract, 5 g acid hydrolyzed casein, and 14 g of agar per liter of water) for *S. erythraea.* Single colonies were selected to inoculate a 7 mL liquid culture of either GNZ or ISP media. Once turbid growth was observed in rich media, 50 μL of this culture was used to inoculate a 7 mL culture of the same minimal media to be used in the SIL experiment. After 24 hours of growth, this culture was used for the inoculation of the microtiter plates. The same culture was used as inoculum for all replicate wells of every feedstock condition in a given experiment. The 24-well microtiter plates and sandwich covers used for micro-scale bacterial cultures were purchased from Enzyscreen B.V. (The Netherlands) and the protocol for microtiter well plate fermentations was adapted from Duetz et al.[20]

The 24-well microtiter plates were cleaned and sterilized according to Duetz et al.,[20] and 2 mL of minimal media was added to each well. The first and last columns in each 24-well plate were left with sterile media and the inner 16 wells were inoculated with 80 μL of bacterial inoculum. Following inoculation, either an SIL compound or the corresponding unlabeled compound was added to each well by sterile filtration. Four replicate wells were prepared and inoculated for each condition, including unlabeled controls. Unlabeled control cultures were included for each feedstock condition to account for metabolic changes that may occur as a result of adding the precursor compound. The SIL feedstock compounds were tested at various concentrations as indicated in the following sections. Stock solution concentrations were adjusted according to the final desired concentration of each SIL or unlabeled precursor in the culture so that a minimal volume of 20-100 µL of stock solution was added to each well. The replicate cultures were fermented and analyzed separately and therefore account for technical variation in both the fermentation experiment as well as the analytical variation in the MS data.

Microtiter plates containing SIL supplemented bacterial cultures were shaken at 200 rpm and maintained at 23.0 °C for five days. On the fifth day the cultures were extracted by adding 2 mL of Optima methanol to each well. The contents of each well were then transferred to Eppendorf tubes, sonicated for 5 minutes, and centrifuged for 1 minute at 16,000 g. Methanol/water extracts were injected directly onto the UPLC-qTOF system, or diluted to maintain the most intense signals in the chromatogram in an optimal range for both sensitivity and mass accuracy.

2.5.5. Optimization of SIL Tracer Incorporation in Erythromycin A

In order to determine the optimal concentration for each of our SIL precursors, I performed an experiment using *S. erythraea* as described above with replicates of the following concentrations of each SIL: (100 mM, 10 mM, 1 mM, 0.1 mM) [1-^{13}C]acetate, (100 mM, 10 mM, 1 mM, 0.1 mM) [1-^{13}C]propionate, (16.5 mM, 10 mM, 1 mM, 0.1 mM) [methyl-^{13}C]methionine, and (20 mM, 10 mM, 2 mM) [1-^{15}N]glutamate. In all conditions the total concentration of glutamate is 20 mM, as glutamate is the only nitrogen source in the minimal media. For the [1-^{15}N]glutamate labeling condition we tested three ratios of [1-^{15}N]glutamate and unlabeled glutamate, so that the total available nitrogen was the same in every condition. Unlabeled controls were performed as four replicates with matching concentrations of unlabeled acetate, propionate, methionine, and glutamate respectively. Erythromycin A (**2.5**) was used as a test case to determine the optimal concentrations of each SIL precursor (Figure 2.10).

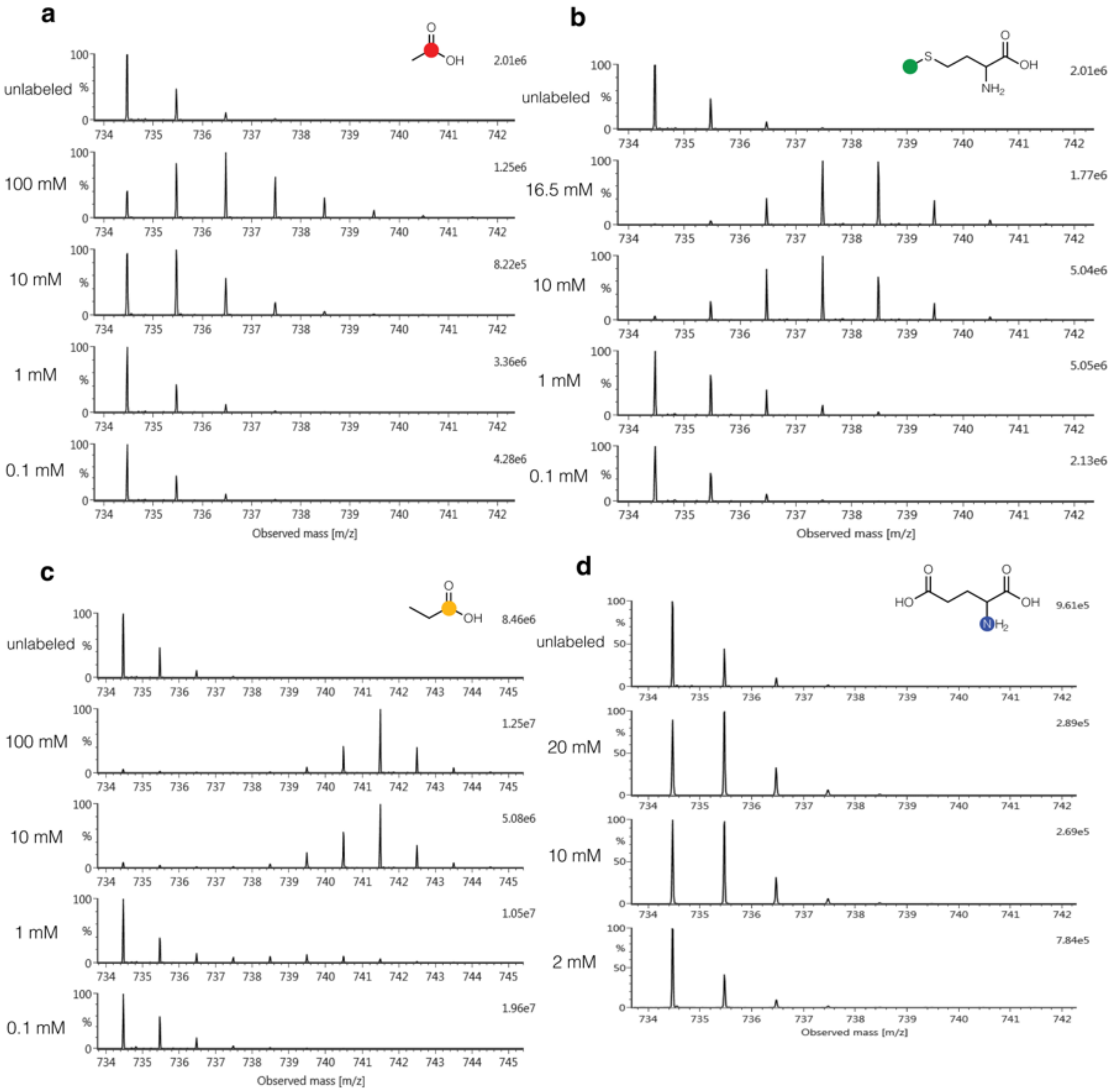


Figure 2.10 Mass Spectra of 2.5 with Varying Concentrations of SIL Tracers
Mass spectra of erythromycin A (**2.5**) from *S. erythraea* extracts supplemented with [1-^{13}C]acetate (a), [methyl-^{13}C]methionine (b), [1-^{13}C]propionate (c), and [1-^{15}N]glutamate (d) at the concentrations indicated.

2.5.6. UPLC-MS Methods

The samples that were prepared from the pre-fractionated extracts of 60 mL bacterial cultures were processed and ran at UC Santa Cruz using an Agilent 1260 binary pump and an Agilent 6230 time-of-flight mass spectrometer with a Jetstream ESI source. A 2 uL sample was injected onto a 1.8-um particle size, 50 x 2.3 mm i.d., ZORBAX RRHT C18 column. A gradient from 10 to 90% acetonitrile over 4 min with a

1.5 min hold at 90% and 3 min re-equilibration. Flow rate 08 mL/min. 0.02% formic acid in solvents. Positive mode, 100 – 1700 m/z, source temp 350 C, 11L/min drying gas.

All other samples were prepared and analyzed by the following procedure. Biological samples were diluted with an equal volume of methanol and the supernatants were subjected to chromatographic separation and mass spectrometric analysis. Chromatography was performed on a Waters I-Class Acquity UPLC system (Acquity HSS T3 1.8 µm, 2.1 x 100 mm) using a linear gradient (solvent A: H_2O + 0.01% formic acid, solvent B: acetonitrile + 0.01% formic acid) of 5-98% B over 5.8 minutes, a hold a 98% B for 0.3 min followed by a 1.8 minute re-equilibration at 5% B. All mass spectra were acquired using a Waters Synapt G2Si qTOF MS run in data-independent acquisition (DIA) mode. The MS detector range was set to 50-1500 *m/z* in positive mode, with a capillary voltage of 3.5 kV, and a desolvation temperature of 200 ºC.

References

1. Chokkathukalam, A., Kim, D.-H., Barrett, M. P., Breitling, R. & Creek, D. J. Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis* **6**, 511–524 (2014).

2. Ong, S. E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1**, 2650–2660 (2007).

3. Gross, H. *et al.* The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. *Chem. Biol.* **14**, 53–63 (2007).

4. Cane, D. E., Hasler, H., Taylor, P. B. & Liang, T. C. Macrolide biosynthesis-II. Origin of the carbon skeleton and oxygen atoms of the erythromycins. *Tetrahedron* **39**, 3449–3455 (1983).

5. Rinkel, J. & Dickschat, J. S. Recent highlights in biosynthesis research using stable isotopes. *Beilstein J. Org. Chem.* **11**, 2493–2508 (2015).

6. 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).

7. Jelić, D. & Antolović, R. From Erythromycin to Azithromycin and New Potential Ribosome-Binding Antimicrobials. *Antibiotics* **5**, 29 (2016).

8. Schulze, C. J. *et al.* Genome-Directed Lead Discovery: Biosynthesis, Structure Elucidation, and Biological Evaluation of Two Families of Polyene Macrolactams against Trypanosoma brucei. *ACS Chem. Biol.* **10**, 2373–2381 (2015).

9. Doull, J. L. & Vining, L. C. Nutritional control of actinorhodin production by *Streptomyces coelicolor* A3(2): suppressive effects of nitrogen and phosphate. *Appl. Microbiol. Biotechnol.* **32**, 449–454 (1990).

10. Sánchez, S. *et al.* Carbon source regulation of antibiotic production. *J. Antibiot. (Tokyo).* **63**, 442–459 (2010).

11. Martín, J. F. Control of antibiotic synthesis by phosphate. in vol. 6 105–127 (Springer Berlin Heidelberg, 1977).

12. Himabindu, M. & Jetty, A. Optimization of nutritional requirements for gentamicin production by *Micromonospora echinospora*. *Indian J. Exp. Biol.* **44**, 842–848 (2006).

13. Kawamoto, I., Oka, T. & Nara, T. Carbon and nitrogen utilization by *Micromonospora* strains. *Agric. Biol. Chem.* **47**, 203–215 (1983).

14. Hodgson, D. A. Primary metabolism and its control in Streptomycetes: A most unusual group of bacteria. *Adv. Microb. Physiol.* **42**, 47–238 (2000).

15. Craney, A., Ozimok, C., Pimentel-Elardo, S. M., Capretta, A. & Nodwell, J. R. Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* **19**, 1020–1027 (2012).

16. Romero-Rodríguez, A. *et al.* Carbon catabolite regulation in Streptomyces: new insights and lessons learned. *World J. Microbiol. Biotechnol.* **33**, 162 (2017).

17. Zhang, J. & Greasham, R. Chemically defined media for commercial fermentations. *Appl. Microbiol. Biotechnol.* **51**, 407–421 (1999).

18. Wentzel, A., Sletta, H., Consortium, S., Ellingsen, T. E. & Bruheim, P. Intracellular metabolite pool changes in response to nutrient depletion induced metabolic switching in *Streptomyces coelicolor*. *Metabolites* **2**, 178–194 (2012).

19. Wentzel, A. *et al.* Optimized submerged batch fermentation strategy for systems scale studies of metabolic switching in *Streptomyces coelicolor* A3(2). *BMC Syst. Biol.* **6**, 59 (2012).

20. Duetz, W. A. *et al.* Methods for intense aeration, growth, storage, and replication of bacterial strains in microtiter plates. *Appl. Environ. Microbiol.* **66**, 2641–2646 (2000).

21. Okamoto, S., Taguchi, T., Ochi, K. & Ichinose, K. Biosynthesis of Actinorhodin and Related Antibiotics: Discovery of Alternative Routes for Quinone Formation Encoded in the act Gene Cluster. *Chem. Biol.* **16**, 226–236 (2009).

22. Cerdeño, A. M., Bibb, M. J. & Challis, G. L. Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): New mechanisms for chain initiation and termination in modular multienzymes. *Chem. Biol.* **8**, 817–829 (2001).

23. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem. Rev.* **119**, 12524–12547 (2019).

Chapter 3.

IsoAnalyst: An MS Metabolomic Platform for the Detection of Stable Isotopic Label Incorporation in Natural Products

3.1. Introduction

Advances in MS technology have driven a surge in the development of innovative metabolomics approaches. As discussed in Chapter 1, targeted metabolomics methods are now streamlined for many general classes of metabolites such as amino acids, and central carbon metabolism. This was achievable due to an increase in accessible methods to quickly acquire high resolution MS datasets and advances in compound identification databases, especially those that use fragmentation data such as GNPS.[1,2] Untargeted metabolomics is an appealing approach to expand upon metabolomics for novel pathway and compound discovery. Coupling untargeted metabolomics studies with SIL precursors has powerful implications but there are challenges associated with such an approach. Many biologically relevant SIL tracers are commercially available, and MS facilitates the sensitive detection of their incorporation into metabolites, however dedicated and intuitive software for the interpretation of untargeted SIL MS data is lacking.[3] Still, there have been some exciting developments in this area as there are wide reaching applications of untargeted SIL MS metabolomics.

A recent study demonstrated how the complete ^{15}N labeling of a cyanobacterial culture using $Na^{15}NO_3$ can be employed to associate peptide natural products with their NRPS BGCs.[4] The authors focused on NRPS BGCs because adenylation domains can be identified in the genome sequence to predict the amino acid sequence and therefore the total number of expected nitrogen atoms in the product. By providing $Na^{15}NO_3$ as the sole nitrogen source, any nitrogen-containing natural product that the cyanobacteria produced may be labeled and detected. This was a rare untargeted study that aimed to detect all natural products in a given extract that demonstrated ^{15}N incorporation, however the authors did not use any statistical analysis.[4] Rather, they manually compared SIL incorporation in the experimental samples to the unlabeled controls to determine the number of ^{15}N atoms present in each compound. They successfully identified four compounds from four different BGCs by matching the mass difference between unlabeled and labeled isotopologue peaks with the expected number of nitrogen atoms in the predicted BGC products.[4] Two of the compounds identified were new natural product structures, highlighting the ability to discover novel compounds and quickly generate hypotheses about their biosynthetic origin. This promising example of untargeted SIL metabolomics demonstrates the need for powerful and flexible tools for the detection of SIL tracers in complex datasets.

Another recent study used [$^{13}C_6$]-phenylalanine in plants to detect compounds derived from this precursor, termed the ‘phenylalanine derived metabolome’. This study aimed to detect the entire phenylalanine derived metabolome of a wild type *Arabidopsis thaliana* and various mutants to identity genes associated with variation in phenylalanine derived metabolites.[5] Like many other SIL tracer studies, these authors used an amino acid with multiple labeled atoms because this makes the detection of incorporation much more straightforward. In paired samples, the mass shift of 6 Da can be detected by mass searching and confirmed as a full incorporation of the entire SIL tracer. The use of a tracer like [$^{13}C_6$]-phenylalanine increases the ease of detecting target compounds but at a high financial cost. Cheaper SIL tracers are preferable when designing experiments with parallel labeling conditions. Interpretation of singly labeled tracers is not as straightforward due to isotopic contributions from natural ^{13}C and the complex distributions of isotopologue peaks that occur when an SIL precursor is incorporated into the same compound multiple times. These recent publications show that there is a strong need across different biological systems to identify compounds associated with BGCs in a systematic way. Detection of SIL tracer incorporation in MS metabolomics is highly applicable across different systems and computational tools that allow for flexible SIL experimental designs are in high demand.

3.1.1. Computational Tools For Untargeted SIL MS Metabolomics

Untargeted SIL analyses have grown in popularity as they are better suited to novel pathway discovery and a number of tools have been developed for this purpose. Many of these tools can be used in conjunction with fluxomics analyses to discover novel pathways and metabolites, or to associate metabolites that are co-regulated under drug pressure or disease conditions.[2,3] XCMS is a common MS data processing platform that was originally developed in R but now in its online version has become highly accessible to researchers around the world.[6] Supplemental programs that work with the XCMS output to further process metabolomics data are used for many different applications, including untargeted SIL analysis. In particular, $X^{13}CMS$ has become a useful tool for comparing isotopic incorporation in *m/z* features across different conditions.[7,8] Other algorithms such as mzMatch-ISO[9] and geoRge[10] also rely on XCMS and mzMatch.R[11] for data pre-processing. XCMS is most often used to process LCMS data although it can handle GCMS data as well. Dedicated algorithms such as non-targeted tracer fate detection algorithm (NTFD)[12,13] and mass isotopolome analyzer (MIA)[14] were designed specifically for low resolution GCMS data. These tools have demonstrated effective deconvolution of low resolution SIL MS data and the ability to account for natural SIL abundances in bulky derivatization groups used for GCMS. The datasets generated by these tools have high potential for novel pathway discovery in global metabolic studies and in conjunction with fluxomics or other systems level modeling.[15–17] Although these methods are all applicable to novel natural product discovery, they are limited in that they typically require a high degree of labeling such as from a fully labeled carbon source and aim to compare changes in labeling across the metabolome as a result of a perturbation. Because they aim to look globally at the central metabolome, these tools can be applied to synthetic biology and optimization of compound production, but are too promiscuous in the SIL design for easy identification of novel compounds.

An R package, Miso, was recently developed specifically to handle the detection multiple SIL conditions in parallel experiments.[18] Miso automatically detects SIL incorporation in isotopologues across different SIL precursor conditions and aligns them with an unlabeled dataset. The output of Miso contains a complete list of all isotopologue pairs detected, and the number of labeled atoms in each condition. Miso was originally created for a particular experimental design published by the same group termed Dual Labeling of Metabolites for Metabolome Analysis (DLEMMA).[19,20] This approach requires parallel isotopic labeling conditions containing the same metabolite precursor with different SIL compositions (ie tyrosine-$^{2}H_4$ and tyrosine-$^{13}C_9{}^{15}N_1$). The aim of this analysis is to assist in sub-structure searching and molecular formula assignment for metabolite identification, but requires multiply labeled tracers to target a specific pathway. The novelty of this approach is in its ability to align labeling data from parallel SIL conditions, however the particular requirements of the SIL selection make these experiments somewhat inflexible. The Miso package has been developed to allow for additional flexibility in SIL precursor selection outside of the DLEMMA workflow.[18] The Miso algorithm and workflow are a powerful new tool for investigating natural product biosynthesis and facilitating discovery in microbial natural products, although it was developed initially as an application in plant metabolomics.

Like the untargeted SIL data analysis approaches described above, Miso compares isotopologue intensities between labeled and unlabeled conditions to statistically determine the enrichment of heavy isotopologue peaks in the labeled sample. Many of these methods give relative isotopologue ratios as outputs, but Miso gives a very specific output indicating the number of labeled atoms present in each detected molecule. This is possible because the published examples of the Miso analysis involve SIL precursors with multiple labeled atoms, and assumes that for biological relevance, the precursor will be fully or partially incorporated into the compounds of interest with two or more labeled atoms. This works well for the application of substructure searching of a specific metabolite or group of metabolites that are derived from a common precursor. Like Miso, IsoAnalyst aims to identify the number of SIL tracer molecules incorporated into metabolites across a set of SIL conditions. However, a method for interpreting isotopologue ion intensities that accurately detects iterative incorporations of a singly labeled SIL tracer is beyond what is currently available in untargeted MS metabolomics.

All of the algorithms discussed so far apply similar processing and statistical analyses of isotopologue peaks by comparing relative isotopologue intensities between unlabeled and labeled samples. Peaks that are determined to be labeled are typically displayed as mass isotopomer distributions (MID), or isotopologue distributions depending the resolving strength of the mass spectrometer. MID patterns can be used for statistical association between features showing similar labeling patterns or visualized for manual comparison. Miso differs from this type of output by providing the number of heavy isotopes detected in each compound. IsoAnalyst aims to generate an output similar to Miso, but uses relative isotopologue ratios within a single labeled sample to determine the number of SIL tracers incorporated into the structure. This approach allows for the more sensitive detection of SIL incorporation beyond the most intense isotopologue peak in the mass spectra.

3.1.2. The IsoAnalyst Approach to Untargeted SIL MS Metabolomics Analysis

While other approaches aim to interpret the SIL incorporation information directly for structure determination, IsoAnalyst focuses on the biosynthetic profile of molecules labeled under different conditions to generate hypotheses about the genomic origin of that molecule. To facilitate the use of various singly labeled precursors I developed the novel algorithm, IsoAnalyst, to determine the number of detectable SIL incorporation events in each feature under every SIL conditon. In Chapter 2 I covered how the IsoAnalyst approach is based on the flexibility of SIL precursor selection to identify different pathways. Here I will describe how IsoAnalyst determines the number SIL atoms in every mass feature across all experimental conditions.

IsoAnalyst processes mass spectrometry data in three stages. Initially, the raw data are pre-processed using third party software to generate peak and feature lists as input files. MS features are then aligned across all unlabeled control samples in the first processing step to generate a ground truth feature list containing the monoisotopic *m/z* value (M_0) for every MS feature detected across the entire experiment (Figure 3.1a). The next step uses the M_0 peak of each feature in the master list to identify and scrape the associated isotopologue peak data (M_1, M_2, M_3, etc.) for every feature in each SIL condition (Figure 3.1b) Finally, IsoAnalyst determines the degree of isotopic labeling for each analyte by comparing the relative isotopologue ratio M_1:M_0 of a feature in the unlabeled dataset with sequential pairs of peaks (e.g. M_1:M_0, M_2:M_1 etc.) in the labeled dataset. For every position where this ratio is significantly greater than the M_1:M_0 ratio of the unlabeled feature, we assume that at least that many positions are enriched in that given condition. The final output of IsoAnalyst is a summary of the number of SIL incorporations detected for each feature under every SIL condition (Figure 3.1b).

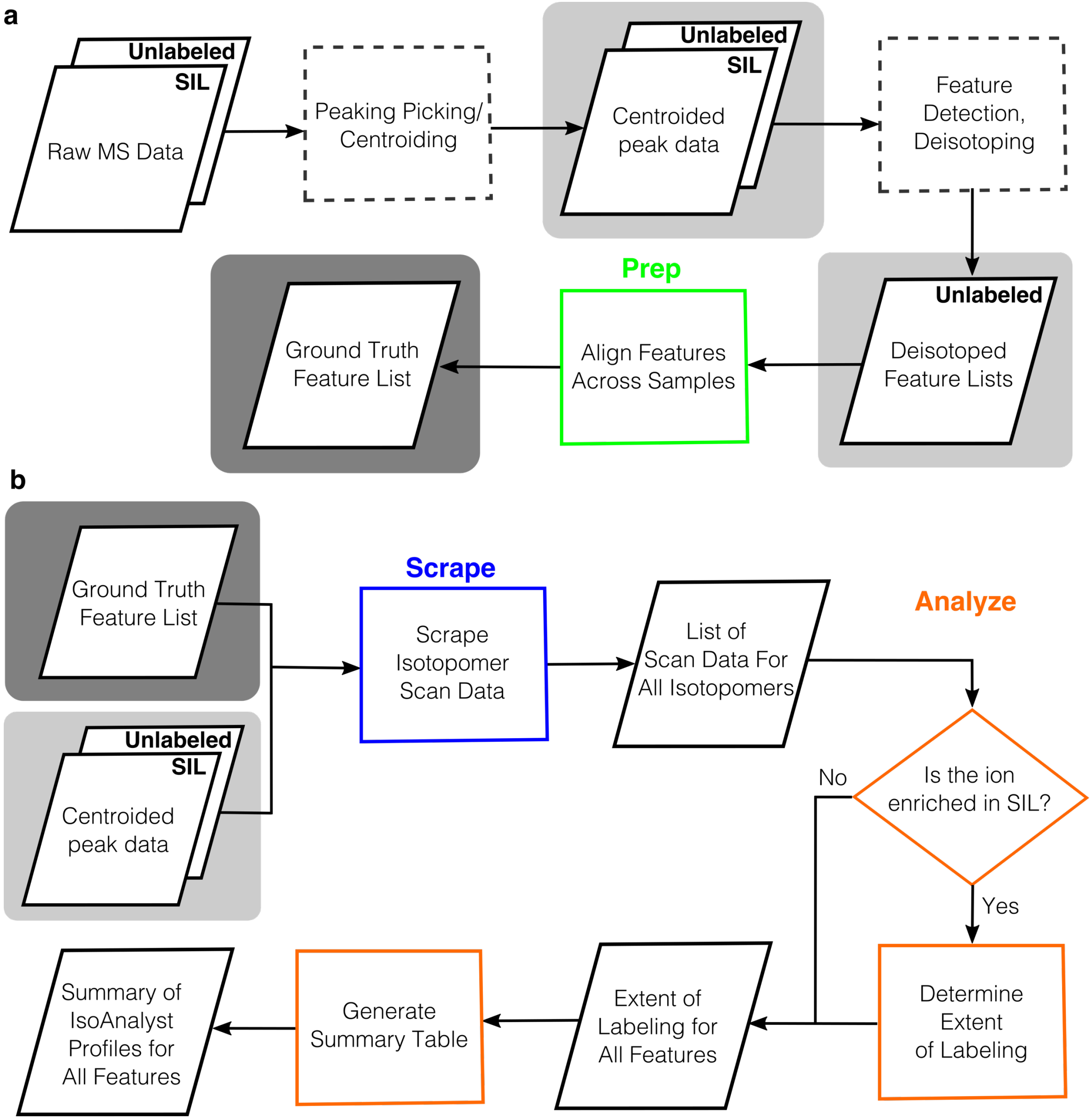


Figure 3.1 IsoAnalyst MS Data Processing Workflow

(a) Required data pre-processing steps to generate input files for IsoAnalyst. Files are indicated as tilted rectangles and those highlighted with a light gray box are required input files. Requirements described in the software documentation are available in the GitHub repository. The ground truth feature list of features aligned across samples is highlighted in a dark gray box and may be generated by the 'Prep' step of the IsoAnalyst program (highlighted in green) or by third party tools. (b) IsoAnalyst performs the following steps: all isotopologue peak information for every feature is first scraped from the centroided peak lists in the 'Scrape' step (highlighted in blue). In the 'Analyze' step (highlighted in orange), the isotopologue ratios are compared for every feature in each SIL condition to determine the extent of labeling. Finally, a summary file is generated containing all of the SIL incorporation profiles for every feature that contains labeling in two or more conditions.

3.2. Standard Processing: Feature Detection and Alignment

3.2.1. Peak Picking and Feature Detection

Mass spectrometry generates complex multi-dimensional datasets, and pre-processing of these datasets greatly affects the quality of downstream statistical results. There are many freely available tools and vendor specific software packages for this purpose, which typically involve processes such as noise filtering, baseline correction, peak picking, and retention time alignment.[6,21,22] Centroiding, sometimes referred to as 'peak picking', determines the *m/z* peak's centroid (or center) and collapses it into a single data point representing the intensity of one peak in one scan.[23] This is not to be confused with the subsequent pre-processing step, feature detection, or the chromatographic alignment of *m/z* peaks. Because the extracted ion chromatogram of an *m/z* peak is also often referred to as a 'peak', representing compound elution over time, there is some overlap in the terminology between these two pre-processing steps. Feature detection often involves various other processes for cleaning up the data such as smoothing, background subtraction, and noise filtering.[24] A chromatographic mass feature is a determined *m/z* value and retention time pair which represents a unique ion detected in an experiment. Isobars have the same mass but can elute at different times, highlighting the importance of accurate feature detection in order to determine individual analytes detected in an experiment even if they have the same mass.

The confusion around this terminology is often disregarded because the majority of MS metabolomics studies apply statistical analyses to only the deconvoluted MS features. The centroided peak lists are typically considered as raw data files and are disregarded in the downstream analysis, except to follow up on specific analytes and to check data quality parameters such as peak shape and overlap. IsoAnalyst differs from other statistical analyses in that it requires separate input files with centroided *m/z* peak information for every sample and deconvoluted MS features for the unlabeled control samples only (Figure 3.1). Throughout this chapter I use the term 'peak' to refer to the centroided *m/z* peaks detected in each scan performed by the MS detector, and the term 'feature' to refer to unique *m/z* and retention time pairs that were deconvoluted from the total ion chromatogram of the sample. When discussing the elution profile of an MS feature I will specify it as a 'chromatographic peak'.

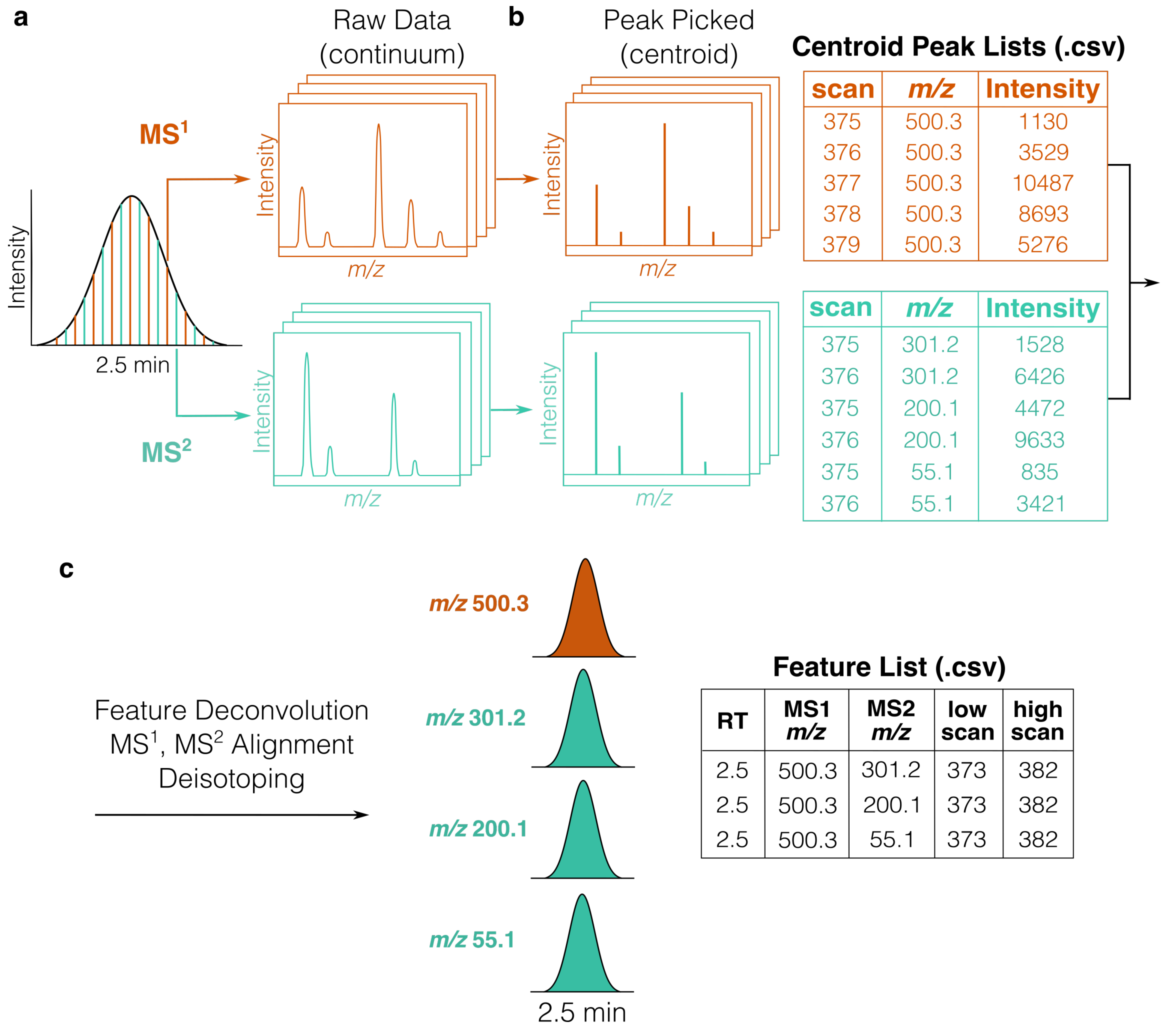
3.2.2. Considerations for Data Independent Acquisition

All of the MS data in this thesis were acquired on a Waters Synapt G2Si qTOF mass spectrometer in a data independent acquisition (DIA) mode, termed MS^E. Generally speaking, DIA of MS data means that all ions detected in the MS1 spectra are equally subjected to fragmentation. This is in contrast to data dependant acquisition (DDA) where only the highest intensity ions, or ions in an inclusion list will be fragmented. While more fragmentation information is retained in DIA methods, greatly reducing bias towards more abundant ions, the deconvolution of matching fragment ions to the corresponding parent mass is more complicated than in DDA. Waters MS^E technology uses alternating low and high energy scans throughout the chromatogram to acquire full profile data across the entire dynamic range in both MS^1 and MS^2 mode for precursor and fragment ion detection respectively (Figure 3.2a)

An in-house pre-processing pipeline which is specialized to handle MS^E data from our Waters Syanpt G2Si was used for peak picking and feature detection. First, a centroiding algorithm is applied, to produce scan-by-scan peak lists for every sample in a csv format (Figure 3.2b). This centroiding algorithm is applied to the high and low energy scan separately to create two centroided peak lists for each sample. These peak lists are then deconvoluted in the pre-processing pipeline to produce feature lists for each sample. The analysis takes into account the chromatographic peak shape of each feature and determines the center (retention time) as well as the low and high scans indicating the beginning and end of the chromatographic peak. Importantly, this step includes de-isotoping, which is essential to the downstream processes because the feature lists are used to identify the monoisotopic M_0 mass for every feature. The final step of feature detection that is specific to DIA MS data is to align the MS^2 features with their parent MS^1 ions and generate one file per sample which contains all parents and fragment ions (Figure 3.2c).

IsoAnalyst is currently only designed to handle MS^1 data, so although I perform the entire process shown in Figure 3.2 on every sample, I only keep the MS^1 centroided peak list and disregard the MS^2 peaks lists. I also disregarded the MS^2 data in the final feature lists in order to simplify the initial development of the algorithm. Fragmentation data are investigated for SIL incorporation on a case-by-case basis to aid in structural assignment in the following chapters. Future development of this work could retain the

MS^2 data in the processing pipeline however significant work remains in developing tools to automate the comparison of SIL labeled MS fragment ions across large datasets.



Centroid Peak Lists (.csv)

scan	*m/z*	Intensity
375	500.3	1130
376	500.3	3529
377	500.3	10487
378	500.3	8693
379	500.3	5276

scan	*m/z*	Intensity
375	301.2	1528
376	301.2	6426
375	200.1	4472
376	200.1	9633
375	55.1	835
376	55.1	3421

Feature List (.csv)

RT	MS1 *m/z*	MS2 *m/z*	low scan	high scan
2.5	500.3	301.2	373	382
2.5	500.3	200.1	373	382
2.5	500.3	55.1	373	382

Figure 3.2 Overview of DIA MS data acquisition and pre-processing
(a) Cartoon diagram of DIA MS data acquisition, where alternating MS^1 and MS^2 continuum scans are acquired across the chromatographic run. (b) Centroiding of continuum data collapses peaks into a single intensity in each scan, which is then written to a .csv file containing the full centroided peak lists. (c) Feature detection aligns the MS^1 and MS^2 data, and writes the data to a .csv file containing all MS^1 features and their aligned MS^2 features.

3.2.3. Generating the MS Feature Master List

The first step of the IsoAnalyst workflow is to align the features detected in the unlabeled control samples by retention time (0.03 min window) and *m/z* (15 ppm error) to generate a master list of all features present across the experiment (Figure 3.1a). IsoAnalyst contains a module which can perform this step, however other programs such as XCMS[6] and MZmine[21] can also align features across samples. This step may be performed within IsoAnalyst, or by a third party tool (Figure 3.1a) The final ground truth list of features for the experiment contains the *m/z* value, retention time, charge state, low and high scan numbers, and the conditions the feature occurs in. This information is used in the following steps to guide the isotopologue data extraction. A unique identifier is assigned to each feature and the *m/z* value recorded in the master list is assumed to be the monoisotopic M_0 mass for that analyte due to the de-isotoping algorithm applied during feature detection. The scan ranges vary slightly between replicates, but are highly reproducible in retention time as expected from a UPLC system. IsoAnalyst takes the highest and lowest scan numbers from the list of replicates of each feature and uses this scan range to scrape isotopologue peaks from all of the experimental scan-by-scan data for that feature in the next step.

The ground truth list of MS features is filtered to include only those features present in at least three of four replicate samples, and to exclude features present in solvent blanks. Doubly charged ions (ie $[M+2H]^{2+}$) are observed in our experiment and retained in the ground truth feature list, however these ions are not analyzed in the subsequent steps of IsoAnalyst. The doubly charged ions observed in these experiments were all found to correspond to singly charged adducts and so I opted to focus on the statistical analysis of only the singly charged species for the development of this tool. Additional considerations should be made for interpreting SIL incorporation into high molecular weight doubly charged ions, which often correspond to large peptides. IsoAnalyst can easily be modified to detect and analyze these isotopologues, however the statistical approach presented here may not be ideal for the interpretation of SIL enrichment in such ions. Software currently exists that is more suited for the analysis of SIL incorporation into large peptides for proteomics applications.[25]

3.3. Isotopologue Detection and Determination of SIL Incorporation

3.3.1. Interrogation of SIL Experimental Data for All Isotopologue Peaks

Isotopologues are isotopic homologs that contain the same number of light and heavy isotopic elements, but the isotopically enriched positions may differ.[26] A mass spectrometer cannot differentiate the position of isotopic enrichment, but identifies isotopologue peaks by their mass.[27] The qTOF-MS system used in this study operates at about 25,000 FWHM resolving power, allowing for accurate mass determination. However, the resolving power of the qTOF mass spectrometer is not sufficient to differentiate between the natural occurrence of ^{13}C and the enrichment of ^{15}N in the [1-^{15}N]glutamate condition used in my experiment. These isotopologues are detected together as a single mass isotopomer peak[27] as opposed to a pure isotopologue peak representing ions corresponding to the incorporation of an individual isotope. This complicates the detection of the isotopologue peaks in the [1-^{15}N]glutamate condition slightly because the combined mass contribution of ^{15}N and ^{13}C skews the detected mass of the heavy isotopologue peaks in comparison to SIL conditions supplemented only with ^{13}C.

For identifying isotopologues in the unlabeled datasets, and in the SIL conditions containing only ^{13}C enrichment, the mass difference of 1.00335 *m/z* is used to identify isotopologue peaks. For labeled features in the [1-^{15}N]glutamate condition, the mass isotopomer peaks M_1, M_2, etc each represent a mixture of isotopologues containing either ^{15}N, ^{13}C, or some combination of both isotopes. For clarity, throughout this text I refer to all isotopically enriched peaks as 'isotopologues,' with the understanding that these peaks represent mixtures of isotopologues only in the [1-^{15}N] glutamate condition. IsoAnalyst uses the mass difference between ^{14}N and ^{15}N (0.99704 *m/z*) to search for isotopologue peaks in the [1-^{15}N] glutamate condition. The contribution of ^{13}C to the isotopologue peaks in the [1-^{15}N] glutamate condition is accounted for in the next step, where isotopologue ratios are compared directly between labeled and unlabeled conditions.

The first challenge in analyzing the isotopologue peaks in this experiment was to attain the complete group of isotopologue peaks for every MS feature in the labeled samples. Because isotope distribution patterns vary significantly between analytes, I found that existing MS analysis software could not correctly identify the M_0 peak of extensively labeled features in the SIL conditions. With the iterative incorporation of singly labeled precursors, my aim is to identify the heaviest isotopologue mass with detectable SIL incorporation even if it is not the most abundant isotopologue peak. The accurate and complete detection of the isotopologue distributions of every feature in the dataset is critical to the sensitivity and accuracy of the downstream analysis. The scrape step in IsoAnalyst looks for all of the isotopologue peaks for every feature in the master list of feature aligned across the experiment in every sample (Figure 3.1b).

To overcome the challenge of accurately aggregating this isotopologue data, IsoAnalyst uses the monoisotopic mass (M_0) of each MS feature from the ground truth feature list as an anchor point to interrogate the labeled MS data for relevant MID patterns. This was accomplished using a custom data processing script that interrogated the centroided MS data from each experimental condition for the presence of each isotopologue peak (M_0, M_1, M_2, etc) for every feature in the master list (Figure 3.1b). Starting from the M_0 mass of each feature, the mass difference between the heavy and light isotope is iteratively added to obtain the theoretical masses of the isotopologue peaks M_1, M_2, M_3, etc. IsoAnalyst then scrapes the scan-by-scan centroided data for every peak corresponding to the calculated isotopologue mass using a 15 ppm error window. The scan range used for obtaining the isotopologue peak data is determined by the initial feature detection step. Isotopologue data is interrogated this way for both unlabeled and labeled samples and written to a single file containing the full scan-by-scan isotopologue data for all replicate samples in each feedstock condition (Figure 3.1b).

3.3.2. Calculating the Isotopologue Ratio

The isotopologue distribution of a molecule can be theoretically calculated on the basis of the charge, molecular formula, and relative isotopic abundances of the elements present in the molecule.[28] It is determined primarily by the natural abundance of ^{13}C (1.07%) and the number of carbons present in the molecule, but smaller isotopic contributions from other elements such as nitrogen and oxygen are often incorporated into theoretical isotopologue distribution calculations.[28] The measured isotopologue distribution of an analyte can be leveraged for molecular formula prediction and assist in compound identification.[29] Complete isotopologue distributions can reduce candidate molecular formulas by >95% compared to accurate mass alone.[28,30] Although this information is useful in compound identification, it is computationally time consuming to predict molecular formulas for hundreds to thousands of features. The ratio of the first isotopologue (one ^{13}C; M_1) to the monoisotopic mass (all ^{12}C; M_0) provides a more straightforward if less information rich way to measure isotopic abundance. This ratio (M_1:M_0) is not always sufficient for complete molecular formula predication but is a molecule-specific measurement of natural isotopic abundance.[31] The natural M_1:M_0 isotopologue ratio for every feature in the unlabeled control data plays a special role in the IsoAnalyst workflow as this calculated ratio is the basis for statistically determining whether that feature is labeled in the SIL conditions.

Mass accuracy has been a major focus in modern HRMS technological developments, however spectral accuracy is a companion concept that is often overlooked.[32] In MS, spectral accuracy typically refers to the ability of the MS to accurately measure isotopic distributions. A common measurement of spectral accuracy is to calculate the percent error of the measured M_1:M_0 isotopologue ratio of a standard from its theoretical calculated M_1:M_0 isotopologue ratio.[31] The M_1:M_0 ratio is usually calculated as a ratio of the integrated intensities of the M_1 and M_0 peaks across the chromatographic region.[29,31,33] In IsoAnalyst I took a different approach to determine the M_1:M_0 isotopologue ratio by plotting the intensity values of the M_1 and M_0 peaks in each scan and plotting a linear regression function to determine the slope (Figure 3.3). The slope of this function is used as the experimentally determined M_1:M_0 ratio for the natural isotopic distribution of each MS feature detected in the unlabeled control data. These isotopologue intensity values are compared for a single analyte in a sample and the slopes of each intensity ratio plot are then averaged across replicates. The mean M_1:M_0 ratio calculated for an MS feature in the unlabeled control samples is later used for statistical comparison to the isotopologue ratios in the labeled condition to determine the number of SIL precursors incorporated in the analyte. Centroid data are not preferable for

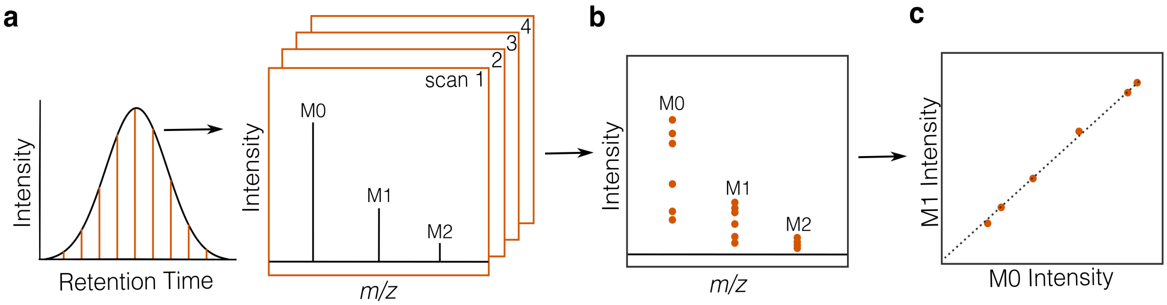


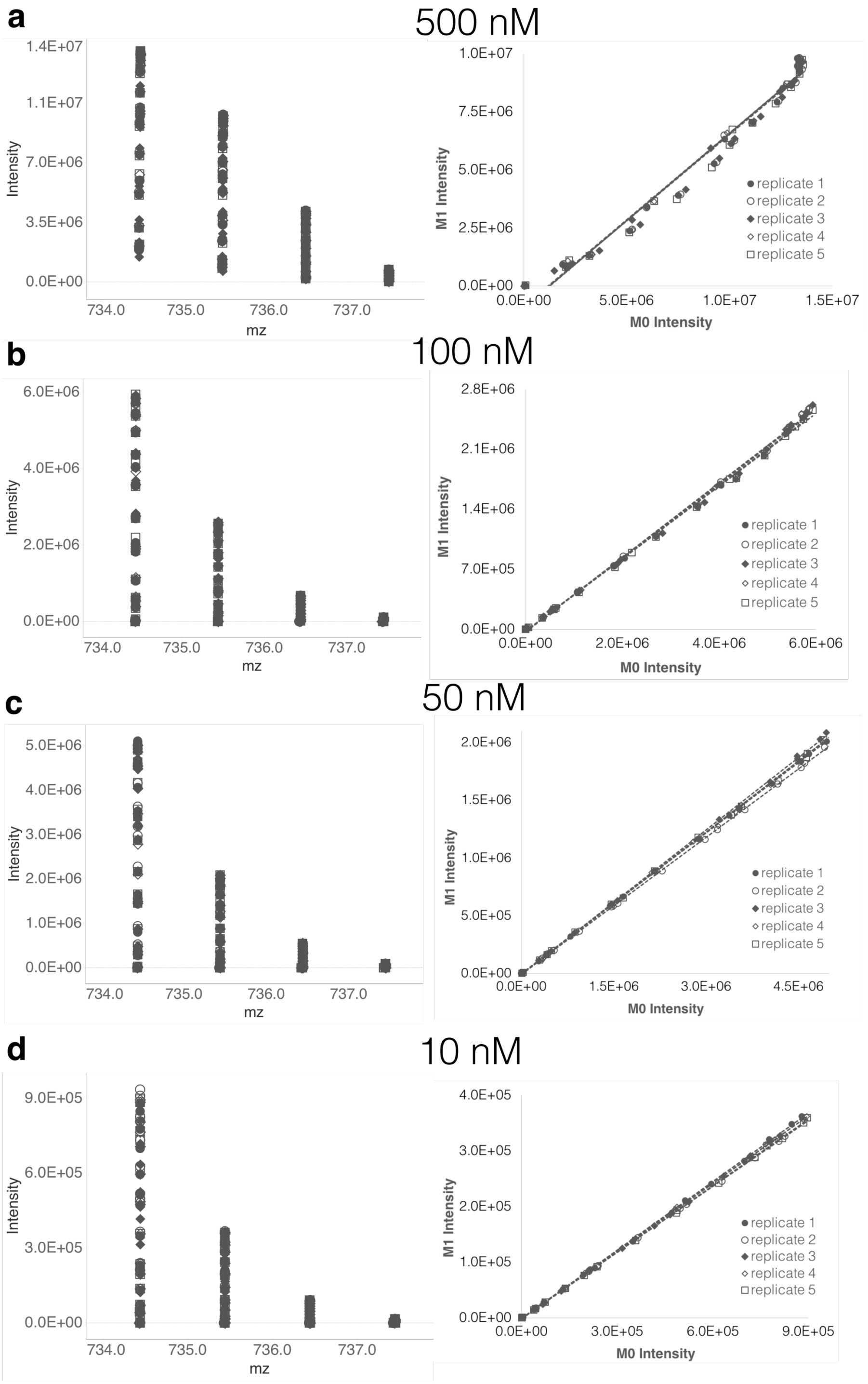
Figure 3.3 MS data centroiding and isotopologue ratio plotting
(a) Diagram of LC-MS data acquisition showing mass spectra collected at regular time intervals across a chromatographic peak. Each orange line represents a single mass spectrum, or scan containing m/z values across the instrument's range. Data are first centroided, or peak picked, to give a single data point corresponding to the intensity of every *m/z* value in a scan. (b) All centroided scan data for a given feature plotted together as single points. (c) All centroided scan data for the

first two isotopologue peaks, M1 and M0, plotted by matching scans. The slope of the linear trend line in (c) is the isotopologue ratio used in the analysis step of the IsoAnayst workflow.

spectral accuracy measurements as much of the spectral peak information is lost in the centroiding process.[30,32] However, by plotting the centroided peak intensities by each scan I can assess the quality of the observed M_1:M_0 ratio in terms of linearity as well as spectral accuracy.

To assess the variability of the naturally occurring M_1:M_0 isotopologue ratio for known analytes in our qTOF system, I analyzed erythromycin A at different concentrations (Figures 3.4, 3.5). Eight concentrations of erythromycin A, 500 nM, 100 nM, 50 nM, 10 nM, 5 nM, 1 nM, 0.5 nM, and 0.1 nM, were prepared and analyzed according to the methods at the end of this chapter. An additional higher concentration of 1 µM was tested, however the signal was highly saturated resulting in such a low mass accuracy that the isotopologue peaks could not be accurately detected and this sample was removed from the statistical analysis. Data was acquired for five replicate injections of each sample. Figures 3.4 and 3.5 show the calculated slope of the M_1:M_0 intensity ratio for the $[M+H]^+$ ion (*m/z* 734.4694) in the replicates for all eight concentrations. The linear regression statistics were computed using the linear regression function in the scipy.stats package. The slopes calculated from the linear regression of the M_1:M_0 intensity plot were then averaged to get the mean slope and standard deviation calculated across the five replicate sample injections (Figures 3.4, 3.5). I refer to the mean slope and the M_1:M_0 ratio interchangeably throughout this section depending on the context. This value is used as the experimentally determined M_1:M_0 isotopologue ratio in comparison to the theoretical M_1:M_0 isotopologue ratio as calculated from a molecular formula.

The standard error of regression (SER) values shown in figures 3.4 and 3.5 are similar to the commonly used R^2 value indicating the linear fit of the data, but provide the error in the same numerical scale as the slope of the linear regression trend line rather than as a percentage. These SER values represent how well the data points in a single sample fit the linear regression model. The overall linear fit of these data is good across all concentrations, however the standard errors for the 500 nM replicates (Figure 3.4a) are notably an order of magnitude higher than the other concentrations. The lowest SER values were found for the intermediate concentrations 50 nM, 10 nM, and 5 nM (Figures 3.4c,d, 3.5a). These statistics alone indicate that the data fit a linear model relatively well across all concentrations, except 500 nM which demonstrated other systematic errors as discussed below.



Rep	Slope	SER
1	0.687	0.014
2	0.687	0.013
3	0.688	0.014
4	0.686	0.014
5	0.682	0.013

Mean	0.686
StdDev	0.003

Rep	Slope	SER
1	0.422	0.002
2	0.428	0.002
3	0.428	0.003
4	0.431	0.002
5	0.420	0.002

Mean	0.426
StdDev	0.004

Rep	Slope	SER
1	0.406	3.0e-4
2	0.395	5.6e-4
3	0.416	0.001
4	0.410	0.001
5	0.409	0.001

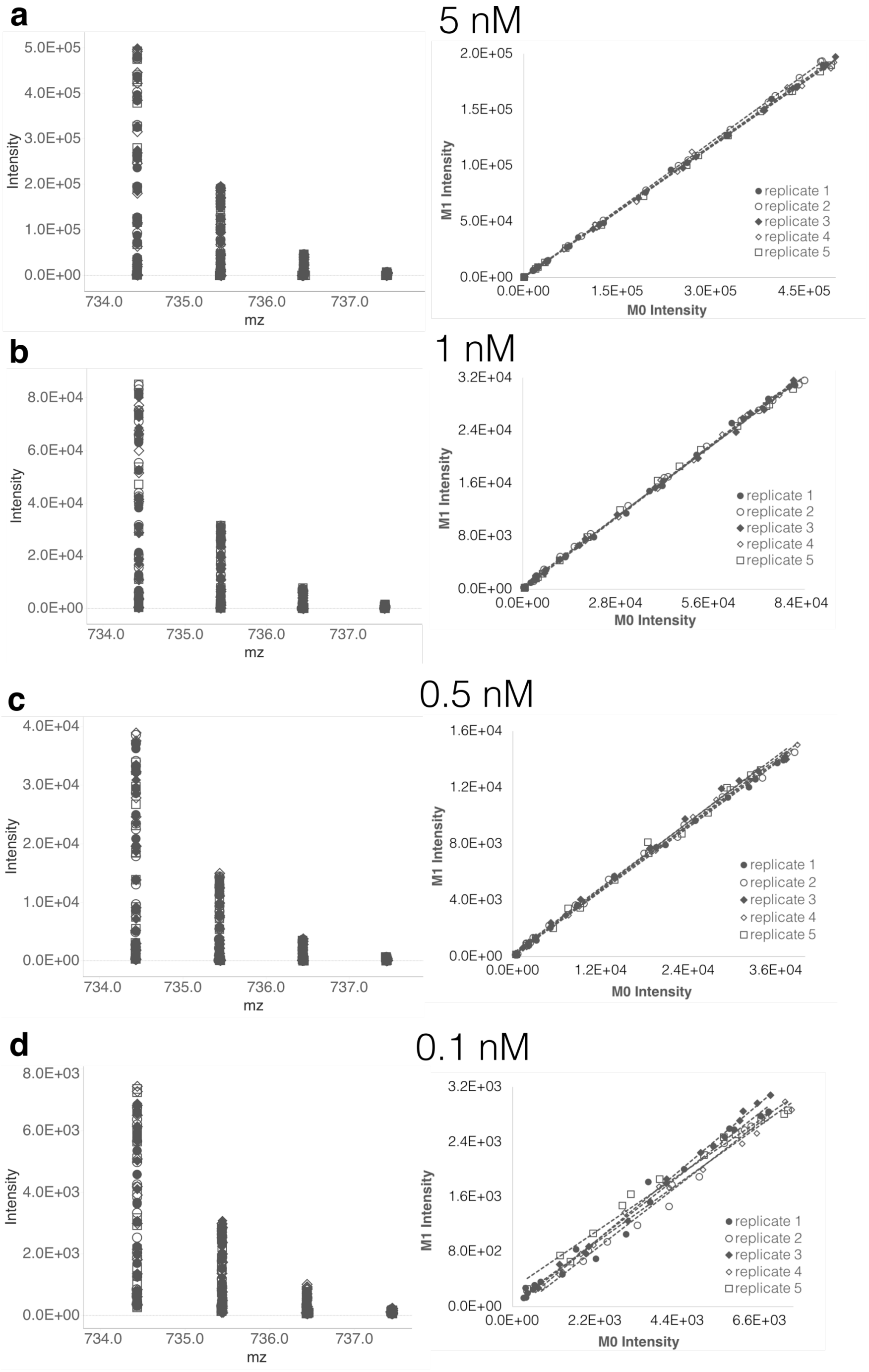
Mean	0.407
StdDev	0.008

Rep	Slope	SER
1	0.409	8.9e-4
2	0.394	2.8e-4
3	0.403	6.7e-4
4	0.404	6.2e-4
5	0.397	7.8e-4

Mean	0.402
StdDev	0.006

Figure 3.4 M_1:M_0 Isotopologue ratio plotting for commercial standard of erythromycin A (10 nM – 500 nM)

Mass isotopologue distributions plotted by *m/z* versus intensity (left), linear plots of the M1:M0 isotopologues of erythromycin A (center), and the individual and mean slopes of each replicate linear graph, standard error (SER) of each linear model, and the standard deviation of the mean slope (right). Data shown for four concentrations 500 nM (a), 100 nM (b), 50 nM (c), 10 nM (d).



Rep	Slope	SER
1	0.393	0.001
2	0.403	5.4e-4
3	0.390	6.6e-4
4	0.389	0.002
5	0.388	5.6e-4

Mean	0.392
StdDev	0.006

Rep	Slope	SER
1	0.383	0.002
2	0.378	0.002
3	0.380	0.001
4	0.384	0.001
5	0.378	0.002

Mean	0.381
StdDev	0.003

Rep	Slope	SER
1	0.379	0.002
2	0.383	0.003
3	0.393	0.003
4	0.387	0.002
5	0.394	0.003

Mean	0.387
StdDev	0.007

Rep	Slope	SER
1	0.421	0.007
2	0.400	0.006
3	0.452	0.004
4	0.397	0.005
5	0.408	0.007

Mean	0.416
StdDev	0.023

Figure 3.5 M_1:M_0 Isotopologue ratio plotting for commercial standard of erythromycin A (0.1 nM – 5 nM)

Mass isotopologue distributions plotted by m/z versus intensity (left), linear plots of the M1:M0 isotopologues of erythromycin A (center), and the individual and mean slopes of each replicate linear graph, standard error (SER) of each linear model, and the standard deviation of the mean slope (right). Data shown for four concentrations 5 nM (a), 1 nM (b), 0.5 nM (c), 0.1 nM (d).

The mean slope and standard deviation were also calculated for each concentration. The standard deviation represents the variability of the mean slope calculated across the five replicates, whereas the SER values indicate how well the M_1:M_0 isotopologue intensity ratio plot of each replicate fits a linear model. The standard deviations of the mean slope are low across all concentrations, except the lowest concentration (0.1 nM) which was an order of magnitude higher than the other concentrations (Figure 3.5d). The SER values for the 0.1 nM data were also slightly higher than the higher concentrations. This indicates that although the overall precision of determining the M_1:M_0 ratio by linear regression is high, there is a lower limit in signal intensity where it is not ideal for measuring the isotopologue intensity ratio.

3.3.3. Accuracy and Variability of the Isotopologue Ratio

The theoretical M_1:M_0 ratio for the $[M+H]^+$ ion of erythromycin A was calculated as 0.417 using the online tool enviPat Web 2.4.[34] Table 3.1 indicates the relative percent error of the experimentally determined M_1:M_0 ratio for each concentration, from the theoretical ratio M_1:M_0 calculated by the equation 3.1:

$$M1{:}\,M0\ error\ (\%) = \frac{M1{:}\,M0(\mathrm{exp.}) - M1{:}\,M0(theor.)}{M1{:}\,M0\ (theor.)} \times\ 100 \tag{3.1}$$

This calculation has been used previously to assess spectral accuracy of MS systems, or how faithfully the MS signals demonstrate the true isotopic abundance.[31–33]

A percent error of 2-5% is often cited as the optimal range for spectral accuracy as measured by the M_1:M_0 ratio of known compounds.[28,30,31] All of the experimentally derived M_1:M_0 ratios for the erythromycin A standard were within 10% of the true M_1:M_0 ratio, except 500 nM which was much higher (64.5%) (Table 3.1). The mean M_1:M_0 ratios and standard deviations for all eight concentrations of the erythromycin A standard are plotted together in Figure 3.6a. The dotted line in Figure 3.6 is drawn at the theoretical value 0.417. There is an apparent trend in these data that the M_1:M_0 ratio tends to decrease as concentration, and therefore signal intensity, decreases. This is true up to a point, as the lowest concentration (0.1 nM) demonstrated a higher average and larger standard deviation (Figure 3.6a).

Table 3.1 Statistics for the natural M_1:M_0 isotopologue ratio of erythromycin A

Sample		Mean Slope (M1:M0)	Slope RSD (%)	Percent Error ± stdev (Spectral Accuracy)*	Average High Intensity	Intensity RSD (%)	Shapiro-Wilk p-value
erythromycin A commercial standard	500 nM	0.686	0.44	64.5 ± 0.6	1.3E+07	0.52	0.183 (n=5)
	100 nM	0.426	0.94	2.08 ± 1.06	5.9E+06	0.89	0.476 (n=5)
	50 nM	0.407	1.97	-2.34 ± 1.88	5.0E+06	1.14	0.473 (n=5)
	10 nM	0.402	1.49	-3.70 ± 1.44	8.8E+05	5.11	0.780 (n=5)
	5 nM	0.392	1.53	-5.79 ± 1.46	4.9E+05	2.07	0.102 (n=5)
	1 nM	0.381	0.79	-8.69 ± 0.66	8.2E+04	3.88	0.531 (n=5)
	0.5 nM	0.387	1.81	-7.11 ± 1.55	3.7E+04	5.38	0.752 (n=5)
	0.1 nM	0.416	5.53	-0.295 ± 5.41	7.1E+03	4.54	0.225 (n=5)
S. erythraea extract	unlabeled acetate	0.356	2.81	-14.6 ± 2.38	8.2E+05	47.7	0.046 (n=4)
	unlabeled propionate	0.346	1.45	-17.0 ± 1.28	4.7E+05	17.3	0.668 (n=4)
	unlabeled methionine	0.352	1.14	-15.5 ± 0.86	4.5E+05	6.14	0.786 (n=4)
	unlabeled glutamate	0.345	2.03	-17.2 ± 1.77	3.6E+05	26.4	0.223 (n=4)
	All unlabeled	0.350	2.29	-16.1 ± 1.86	5.2E+05	49.8	0.238 (n =16)

*Based on theoretical M_1:M_0 0.417

Most of the percent errors calculated were negative, which aligns with published work showing that most HRMS systems including qTOF tend to underestimate the M_1:M_0 ratio.[31,32] The exception being the two highest concentrations, 500 nM and 100 nM which both have M_1:M_0 ratios with positive percent error. The 500 nM concentration in particular has a very high percent error despite having low standard deviation and SER values (Table 3.1, Figure 3.4a). This strongly indicates that the signal intensity in this condition has saturated the detector. When the monoisotopic mass (M_0) is saturated, the first isotopologue (M_1) may continue to show increasing signal intensity before it too reaches saturation at a high enough concentration. Because of this, saturated signals give erroneous M_1:M_0 values that cannot be detected by linear regression or variation statistics alone. The 100 nM concentration also has a positive percent error, however it is within the expected 5% error[28] from the theoretical value (Table 3.1).

The ideal signal range for the Synapt G2Si is an ion intensity of approximately 1.0e3 – 1.0e6. Signal intensities of 1.0e6 and higher are near saturation and often result in a decrease in mass accuracy, while signals higher than 8.0e6 are highly saturated and not guaranteed to meet the specifications for mass accuracy. The average signal intensity for the highest intensity scan is shown for each concentration of erythromycin A in Table 3.1. The signal intensities in the standard erythromycin A samples were highly consistent due to the fact that the five replicates were sample injections from the same vials. In the fermentation experimental design, I include unlabeled control samples that are matched to every SIL precursor condition. This means that for each SIL precursor there are four control wells to which an unlabeled version of that biosynthetic precursor is added. These four replicates are inoculated from the same starter culture, but are fermented, worked up, and analyzed separately. These are technical replicates that account for variation in the wells of the microtiter plate, extraction technique, and MS analysis and so the variability in signal intensity was expected to be higher in these samples.

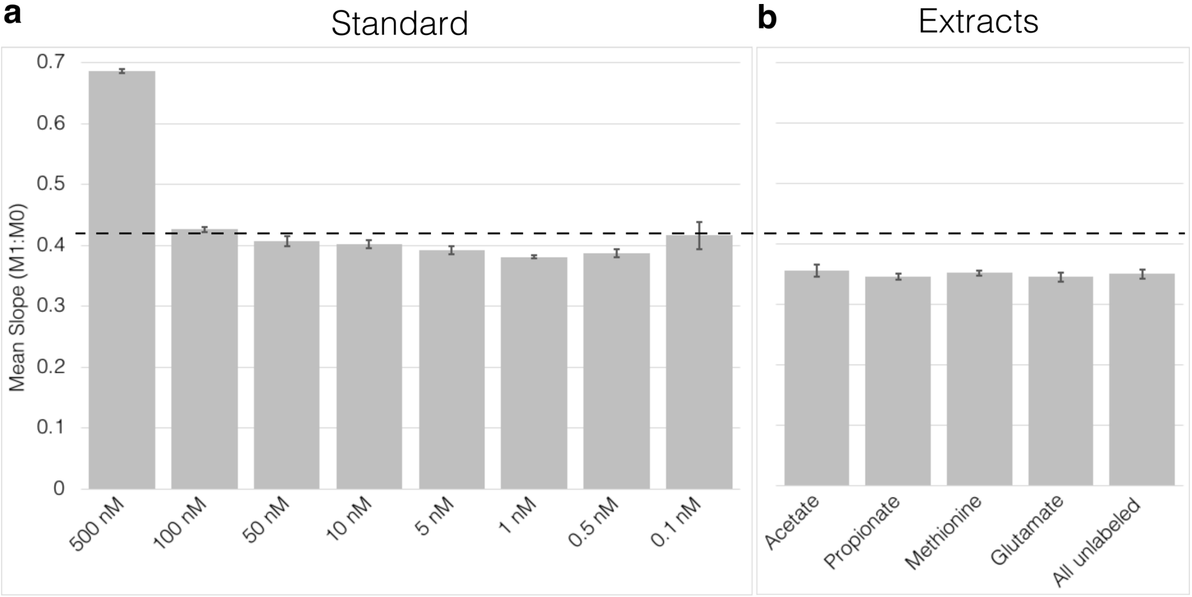


Figure 3.6 Average natural M1:M0 ratios measure for erythromycin A
Dotted line at 0.417 represents the theoretical M_1:M^0 ratio of erythromycin A as calculated using enviPat Web 2.4.[34] Error bars are standard deviation of the mean where n = 5 for the commercial standard of erythromycin A and n = 4 for each unlabeled extract condition.

I calculated the same statistics for the variability and accuracy of the M_1:M_0 ratios in the unlabeled control samples for a full *S. erythraea* experiment including acetate, propionate, methionine, and glutamate. The full experimental conditions can be found at the end of this chapter. Table 3.1 and Figure 3.6b include the statistics for the M_1:M_0 ratios of the $[M+H]^+$ ion of erythromycin A in the unlabeled control extracts of *S. erythraea*. The mean slopes in this table are generated from four replicate sample extracts of cultures that were supplemented with the unlabeled precursor. Each unlabeled precursor control condition is processed separately, however the M_1:M_0 ratio is expected to be identical in these samples as they are all subject to the natural abundance of ^{13}C and analyzed by MS on the same day. The variability (Slope RSD; Table 3.1) is low within these conditions and remains low when the M_1:M_0 ratio is averaged across all sixteen control samples ('All unlabeled'). It is notable however that the percent errors calculated for the M_1:M_0 ratio of erythromycin A in the extracts were considerably larger in magnitude than in the standard samples (Table 3.1). The fact that the M_1:M_0 ratio was consistently smaller in the unlabeled extracts compared to the standard samples suggested that the signal intensity may be lower and potentially biasing the measurement towards a lower M_1:M_0 ratio. The average upper signal intensities for these samples were in the 3.0e5 - 9.0e5 range, which is within the ideal signal range according to Waters and based on the results using the standard of erythromycin A (Table 3.1). As expected, the relative standard deviation (RSD) of these signal intensities were much higher than in the standard samples (Table 3.1), however this did not translate to a higher variability in the M_1:M_0 ratio itself (Figure 3.6b).

Although the M_1:M_0 ratio measured by qTOF-MS and calculated by the approach presented here is not reliably accurate to the literature standard[28], it is clear that within a single experiment we can expect the detected isotopologue ratio to be precise if not very accurate. Because there was low variability in the M_1:M_0 ratio across the full set of sixteen unlabeled control samples, I expect that the factors that contributed to a larger deviance from the theoretical M_1:M_0 value were systematic and affected the entire sample set, including the labeled samples, with the same bias. The 2-5% accuracy suggested by Kind and Fiehn[28] was intended for using the M_1:M_0 ratio to assist in the reduction of potential molecular formulas. This application requires objective accuracy of the M_1:M_0 ratio calculated for each molecular species, however my experiments use the M_1:M_0 ratio for internal comparisons so the overall precision is more important than the accuracy.

The subsequent statistical analysis described in the next section uses a Welch's t-test which assumes that the sample sets being compared have a normal distribution. One informal way of assessing normality is to plot the data in a frequency distribution histogram or to use other methods of visual plotting. I am working with relatively small sample sizes that are not amenable to this assessment. There are also a variety of test statistics used to determine is a sample set is normally distributed but the Shapirio-Wilk test is often cited as providing better power than other commonly used tests of normality.[35] A Shapiro-Wilks test was employed using the scipy.stats package to test the normality of the unlabeled M_1:M_0 ratios for the $[M+H]^+$ ion of erythromycin A in the standard samples and the unlabeled extracts. A significant p-value for a Shapiro-Wilk test (< 0.05) indicates that the data are not normally distributed. All of the M_1:M_0 ratios derived from standard and extract samples showed normal distribution according to this test except for the *S. erythraea* extract supplemented with unlabeled acetate (Table 3.1). This dataset was determined to include an outlier, as further indicated by the higher standard deviation in this condition. The Shapiro-Wilk p-value calculated for the full sixteen unlabeled control sample was also insignificant indicating an overall normal distribution of the M_1:M_0 ratios derived from linear regression slopes of these samples.

3.3.4. Using Isotopologue Ratios to Determine the Extent of SIL Incorporation in MS Features

Because all SIL precursors used in this method are singly labeled, incorporation of any SIL precursor will increase the M_1:M_0 intensity ratio for that condition. Starting with the M_1:M_0 ratio in the SIL condition, IsoAnalyst asks the question, *'is this isotopologue ratio significantly greater than the M_1 to M_0 ratio of the same chemical species having a natural isotopic distribution?'* If the isotopologue ratio of M_1:M_0 for the feature in the SIL condition is statistically indistinguishable from the unlabeled feature, it is determined to have no isotopic enrichment in that condition. However, if it is significantly greater, the feature is determined to be labeled. IsoAnalyst then iteratively compares the intensity ratios for subsequent isotopologue pairs (e.g. M_2:M_1, M_3:M_2, etc) until the isotopologue ratio is no longer statistically distinguishable from the natural M_1:M_0 ratio.

The heavy isotopologue ratios detected in the SIL conditions are expected to demonstrate a higher variability than the natural abundance M_1:M_0 ratios of features in the unlabeled conditions. The replicates in this experiment are derived from individual biological cultures maintained in separate sterile wells in 24-well plates. Although the media conditions are identical, and the culture inoculum for each well is derived from the same original colony, there will always be small variations that affect the metabolism of the SIL incorporation in each well. I looked at the variability of the heavy isotopologue ratios of erythromycin A in the [1-^{13}C]acetate condition because this precursor is not directly incorporated into the compound and therefore may demonstrate higher variability in SIL incorporation under minor environmental changes. Table 3.2 shows the mean slopes for each isotopologue ratio calculated for the $[M+H]^+$ ion of erythromycin A in the [1-^{13}C] acetate samples, with the RSD expressed as a percentage. The RSD values for the first three isotopologue ratios (M_1:M_0, M_2:M_1, M_3:M_2) were between 2-4% which is similar to the RSD of the M_1:M_0 isotopologue ratios calculated for the unlabeled conditions (Table 3.2). The heavier mass isotopologue ratios had increasing RSD values which could be explained not only by biological variability but the lower intensities of these heavy mass isotopologue peaks (M_7:M_6, M_8:M_7 in Table 3.2). Overall, the variability of the isotopologue ratios in the SIL conditions is slightly higher than the M_1:M_0 isotopologue ratios in the unlabeled conditions.

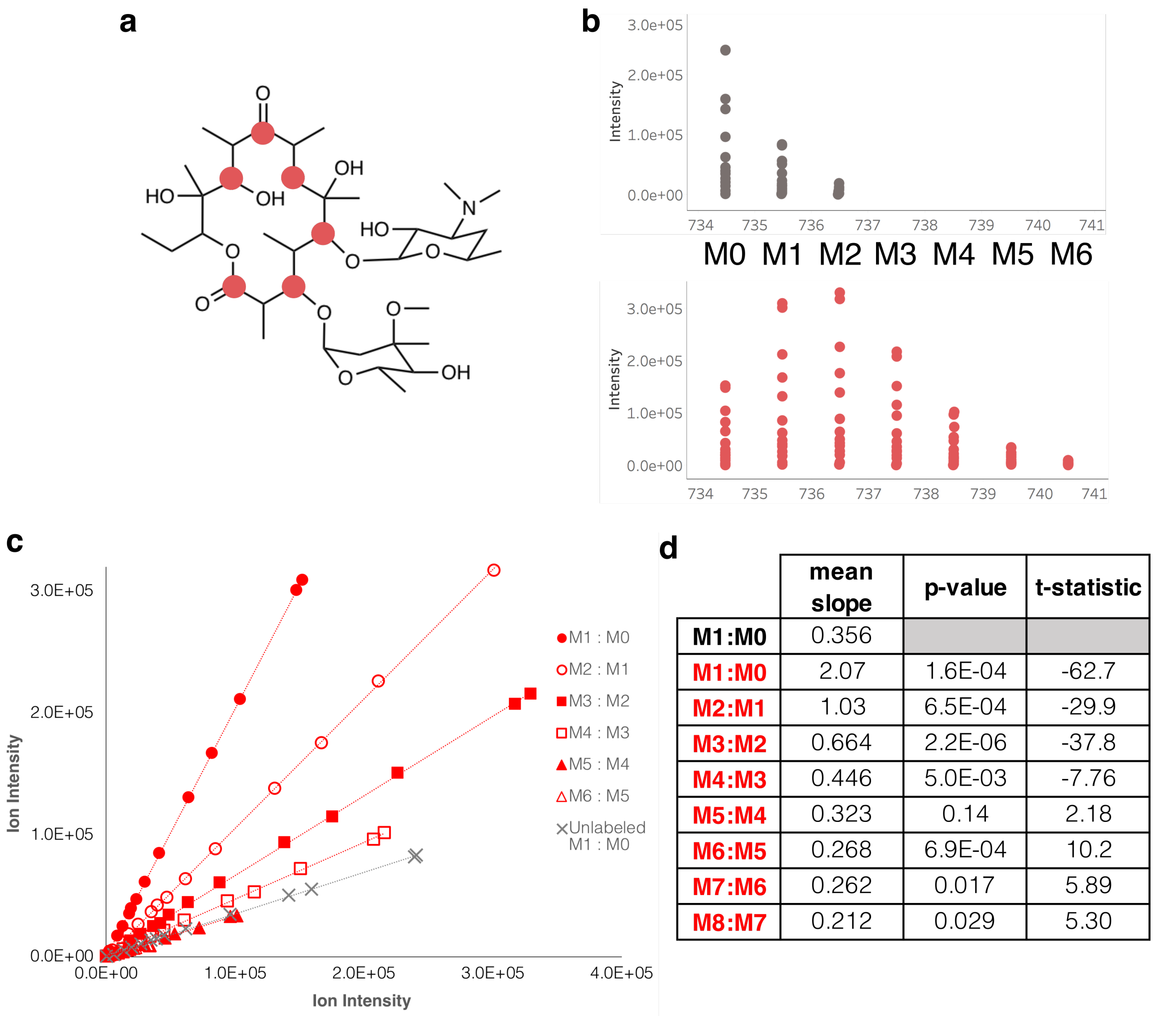
Table 3.2 Isotopologue ratio statistics for erythromycin A labeled by [1-^{13}C]acetate

Isotopologue Ratio	mean slope	Slope RSD (%)	Average High Intensity	Intensity RSD (%)	Shapiro-Wilk p-value
M_1:M_0	2.074	2.25	3.0E+05	5.15	0.478
M_2:M_1	1.032	3.71	3.1E+05	5.89	0.444
M_3:M_2	0.664	1.68	2.0E+05	5.48	0.643
M_4:M_3	0.446	4.06	9.2E+04	9.05	0.078
M_5:M_4	0.323	7.57	3.0E+04	13.1	0.187
M_6:M_5	0.268	4.65	8.2E+03	12.0	0.515
M_7:M_6	0.262	10.0	2.1E+03	6.11	0.106
M_8:M_7	0.212	21.8	4.9E+02	3.34	0.129

[1-^{13}C]Acetate can be incorporated into erythromycin A in six positions, through the indirect transformation of methylmalonyl-CoA from the TCA cycle intermediate succinyl-CoA, as discussed in Chapter 2. Six carbon positions of erythromycin A can be labeled indirectly by [1-^{13}C]acetate through malonyl-CoA (Figure 3.7a). The intensities of the isotopologue peaks of erythromycin A in each scan of the chromatographic peak are shown for one control replicate and one [1-^{13}C]acetate experimental replicate (Figure 3.7b). The M_1:M_0 ratio of the unlabeled control and all detectable isotopologue ratios in the [1-^{13}C]acetate condition are then plotted as intensity ratios (Figure 3.7c). A single replicate is plotted for each isotopologue ratio for comparison (Figure 3.7c), however mean slope data (n=3) were used in the statistical analyses shown in Figure 3.7d.

The statistical comparison between isotopologue ratios of a compound in the SIL condition and the M_1:M_0 ratio in the unlabeled condition is done using a two-tailed Welch's t-test with a p-value cut off of 0.05. Although IsoAnalyst performs a two-tailed t-test in this analysis, it only determines an isotopologue peak to be enriched in an SIL precursor if the M_1:M_0 ratio is significantly *greater* than the natural M_1:M_0 ratio. This is effectively a one-tailed t-test (p-value cut off of 0.025), using a combination of the two-tailed p-value (< 0.05) and the test statistic (< 0) to decide if a isotopologue is enriched with an SIL precursor. For erythromycin A labeled by [1-^{13}C]acetate, the first four isotopologue ratios (M_1:M_0, M_2:M_1, M_3:M_2, and M_4:M_3) all have significant p-values and negative t-statistic values, indicating they are all significantly larger than the M_1:M_0 ratio in the unlabeled condition (Figure 3.7d). When plotted together, it is clear that these isotopologue ratios have larger slopes than the unlabeled M_1:M_0 ratio (Figure 3.7c). The M_5:M_4 isotopologue ratio does not have a significant p-value (Figure 3.7d), indicating that is statistically indistinguishable from the unlabeled M_1:M_0 ratio. Enrichment of SIL in the M_5 isotopologue is therefore not detectable, as this isotopologue may have four positions enriched by ^{13}C derived from the [1-^{13}C]acetate tracer and one ^{13}C position deriving from the natural abundance of ^{13}C. Based on the statistical analysis shown in Figure 3.7d, the M_4 isotopologue is the heaviest isotopologue that has statistically significant enrichment from the [1-^{13}C]acetate tracer. The remaining isotopologues (M_6:M_5, M_7:M_6, and M_8:M_7) have significant p-values (< 0.05), but positive t-statistic values, indicating that they are all significantly smaller than the unlabeled M_1:M_0 ratio (Figure 3.7d).

Although various elements make small contributions to isotopologue abundance, the theoretical M_1:M_0 ratio for a compound depends primarily on the number of carbons in a compound's structure. Larger organic molecules have a higher M_1:M_0 ratio because there are more carbon positions available in the molecule where ^{13}C may occur. The M_1 isotopologue peak represents the same compound as the M_0 peak, except with one ^{13}C present somewhere in the structure. A molecule with more carbons will naturally have a higher probability of having a ^{13}C present in its structure, consequently increasing the ratio of the M_1 isotopologue peak in comparison to the monoisotopic peak. Likewise, when a molecule is enriched with an SIL precursor incorporated into its structure, the probability of the remaining carbons having a naturally occurring ^{13}C is less than that of the whole molecule produced in unlabeled conditions. This is why we see significance in



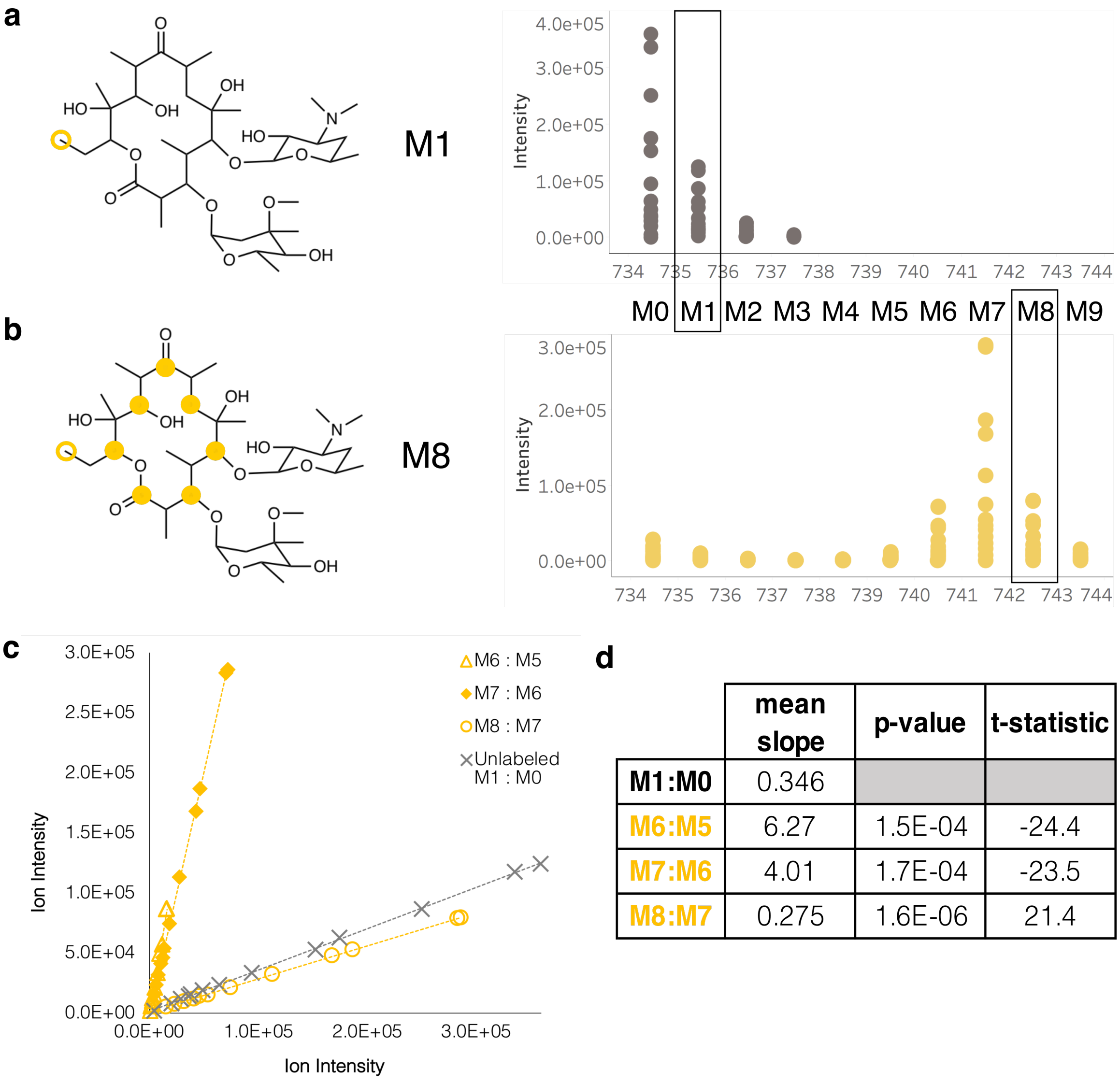
	mean slope	p-value	t-statistic
M1:M0	0.356		
M1:M0	2.07	1.6E-04	-62.7
M2:M1	1.03	6.5E-04	-29.9
M3:M2	0.664	2.2E-06	-37.8
M4:M3	0.446	5.0E-03	-7.76
M5:M4	0.323	0.14	2.18
M6:M5	0.268	6.9E-04	10.2
M7:M6	0.262	0.017	5.89
M8:M7	0.212	0.029	5.30

Figure 3.7 Isotopologue ratios for erythromycin A labeled by [1-^{13}C]acetate
(a) Structure of erythromycin A with red circles representing positions where ^{13}C can be incorporated into the structure via [1-^{13}C]acetate. (b) Plot of the isotopologue peak intensities for erythromycin A in an unlabeled control (top) and an [1-^{13}C]acetate labeled sample (bottom). (c) Isotopologue intensity ratios plots for M_1:M_0 ratio in the unlabeled control, and the first five isotopologue pairs in the labeled sample. (d) Table showing the mean slope of every isotopologue pair and the Welch's two-tailed t-test results for each labeled isotopologue pair.

the two tailed t-test beyond the isotopologue peak containing the most SIL atoms derived from the SIL precursor.

This phenomenon is best illustrated when a compound has complete SIL enrichment in every available position. [1-^{13}C]Propionate is a highly efficient SIL precursor for labeling erythromycin A, which uses six methylmalonyl-CoA units and one propionyl-CoA units in the biosynthesis of its polyketide backbone. In this experiment we observed complete labeling by [1-^{13}C]propionate, which appears as a clear 7 Da shift from M_0 to M_7 (Figure 3.8b). Figure 3.8a shows the structure of erythromycin A corresponding to the M_1 peak in the unlabeled condition, with a single naturally occurring ^{13}C in its structure (open circle). The structure in Figure 3.8b represents the M_8 isotopologue in the [1-^{13}C]propionate condition, which has seven ^{13}C atoms derived from enrichment by the [1-^{13}C]propionate (filled circles) and one position deriving from naturally occurring ^{13}C (open circle). Erythromycin A has 37 carbon positions that may contain ^{13}C derived from natural abundance to generate the M_1 isotopologue structure shown in Figure 3.8a. The isotopologue of erythromycin A which has complete labeling by [1-^{13}C]propionate (M_7) has 30 remaining carbon positions that may contain ^{13}C derived from natural abundance to generate the M_8 isotopologue structure shown in Figure 3.8b. The ratio of M_8:M_7 is significant in the two tailed t-test ($p = 1.6 \times 10^6$), but the graph in Figure 3.8c clearly shows that the M_8:M_7 ratio is significantly smaller than the M_1:M_0 ratio of the unlabeled feature. The M_8:M_7 ratio of erythromycin A in the [1-^{13}C] propionate condition is significantly smaller than the M_1:M_0 ratio of erythromycin A in the unlabeled condition because there are less carbon positions remaining in the structure where natural ^{13}C may occur. This trend of isotopologue ratios being a significantly smaller than the natural M_1:M_0 ratio was common for ions that had strong signal intensity for the heavy isotopologue peaks.

This statistical analysis therefore assigns the degree of isotopic labeling detected for a given SIL precursor in every MS feature aligned across the experiment. IsoAnalyst evaluates every MS feature from the ground truth list in each SIL condition to determine whether or not the feature is isotopically enriched in that condition, and how many positions can confidently be assigned as labeled. These data create a profile for each MS feature indicating the extent of labeling in each SIL precursor tested. The final step is to combine the labeling information into a summary file with every feature from the ground truth feature list and the number of SIL precursors detected in each condition.



	mean slope	p-value	t-statistic
M1:M0	0.346		
M6:M5	6.27	1.5E-04	-24.4
M7:M6	4.01	1.7E-04	-23.5
M8:M7	0.275	1.6E-06	21.4

Figure 3.8 Isotopologue ratios for erythromycin A labeled by [1-^{13}C]propionate (a) Structure of the M_1 isotopologue of erythromycin A in the unlabeled sample. The open yellow circle represents ^{13}C derived from the natural abundance of ^{13}C. (b) Structure of the M_8 isotopologue of erythromycin A. Filled yellow circles represent ^{13}C derived from [1-^{13}C]propionate and the open yellow circle represents ^{13}C from the natural abundance of ^{13}C. (c) Isotopologue ratio plots from an unlabeled control and [1-13C]propionate labeled sample. (d) Table showing the mean slope and Welch's two-tailed t-test results for the three heaviest isotopologue ratios of erythromycin A labeled by [1-^{13}C]propionate.

3.3.5. Complete SIL Incorporation in Erythromycin A

The statistical approach presented in this chapter fulfils the need outlined in Chapter 2 to identify the heaviest isotopologue peak of a compound that is enriched by an SIL precursor. This analysis therefore identifies the number of singly labeled SIL precursors incorporated into a given ion, and aligns these data across parallel SIL conditions. I have already shown that four ^{13}C atoms from [1-^{13}C]acetate and seven from [1-^{13}C]propionate were detected in erythromycin A. This aligns well with the predicted incorporation of these SIL precursors (Figure 3.9a) because [1-^{13}C]propionate is incorporated directly into all of the polyketide core substrates, while [1-^{13}C]acetate only labels the methylmalonyl-CoA extender units indirectly through the TCA cycle. Four labeled methylation positions were detected in the [methyl-^{13}C]methionine condition, corresponding to the dimethylamino group of desosamine, the methylation position of

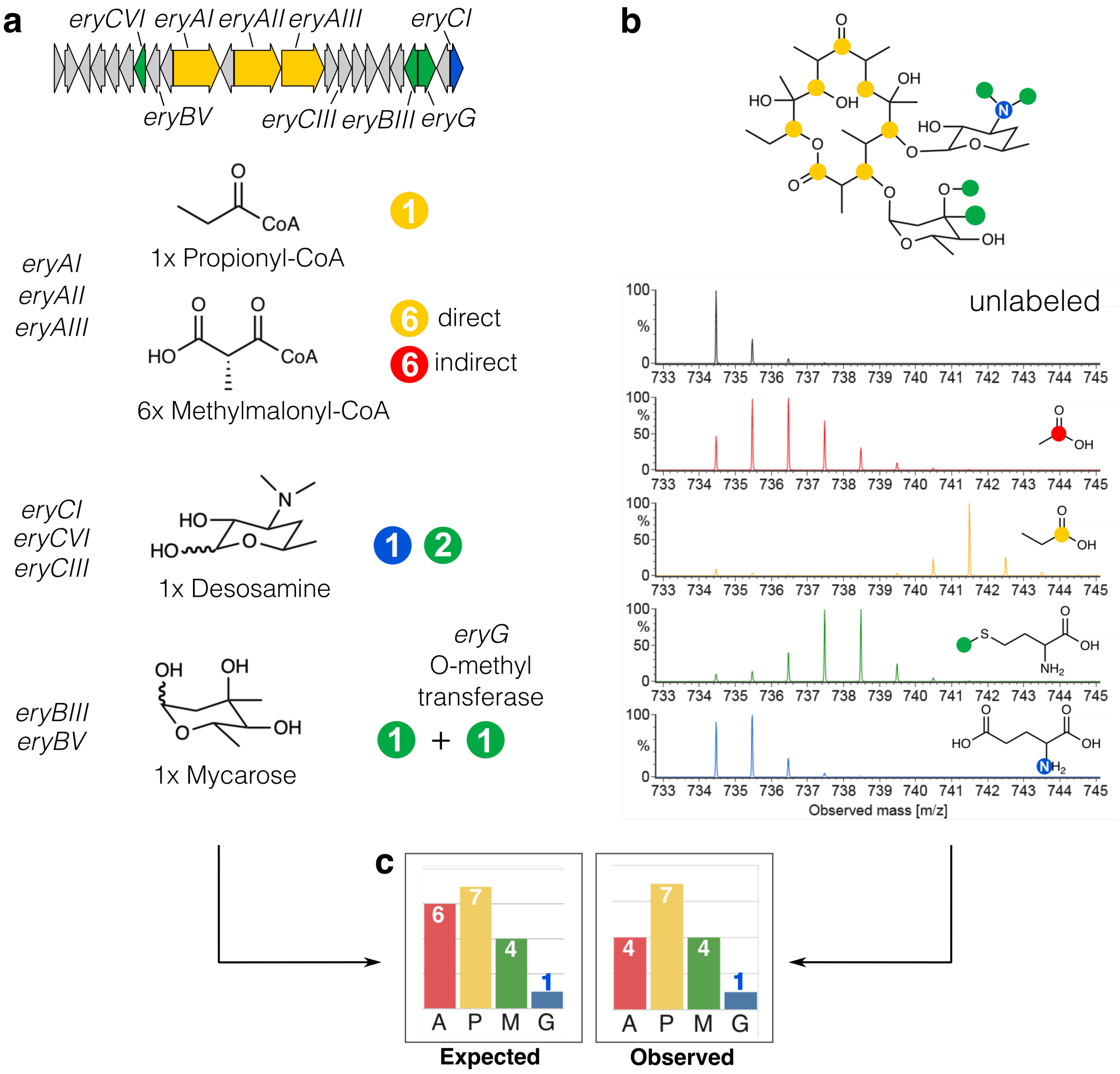


Figure 3.9 Full labeling prifle of erythromycin A by [1-^{13}C]acetate, [1-^{13}C]propionate, [methyl-^{13}C]methionine, [1-^{15}N]glutamate

(a) Erythromycin biosynthetic gene cluster, substrates used in the biosynthesis of erythromycin A, and the expected labeling of these substrates. (b) Structure of erythromycin A and the atoms that are expected to be labeled in the four SIL conditions, and MS data of erythromycin A in the four SIL conditons. (c) Comparison of expected and observed labeling of erythromycin A.

mycarose, and the O-methyl group of mycarose (Figure 3.9). Labeling of the single nitrogen present in desosamine was also detected in the [1-^{15}N]-glutamate condition (Figure 3.9). This demonstrates that the IsoAnalyst MS data processing platform can accurately detect complete iterative SIL incorporation in erythromycin A, a compound with a well-established biosynthetic framework.

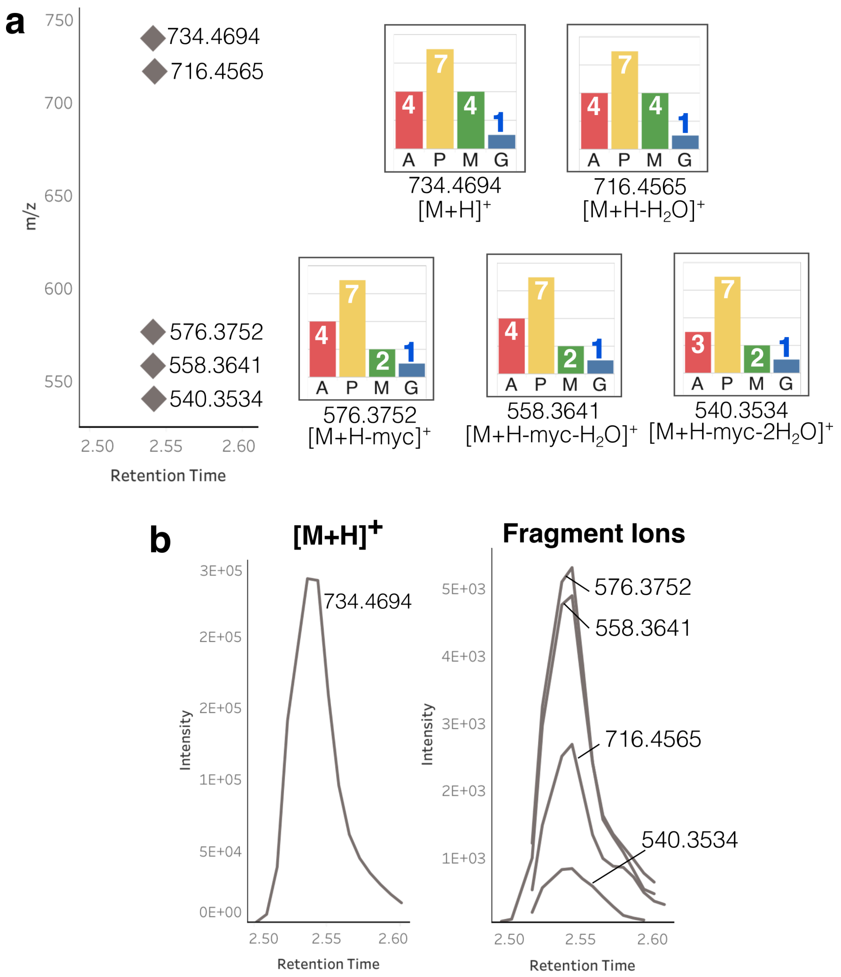


Figure 3.10 IsoAnalyst labeling profiles for adducts and in-source fragments of erythromycin A

(a) Diamonds represent MS features associated with erythromycin A plotted by m/z and retention time. Bar graphs show IsoAnalyst results for each features. (b) Chromatograms showing the peak shape and retention time of the MS features from (a).

A common difficulty in MS metabolomics is the high number of mass adducts and in-source fragments that can inflate the number of detected features compared to real compounds present in a sample.[36] A total of five features were detected for erythromycin A, corresponding to different fragments of the same compound (Figure 3.10). The dehydrated fragment ion *m/z* 716.4565 had labeling patterns that matched the [M+H]$^+$

ion, while smaller fragments showed decreased labeling from [methyl-^{13}C]methionine due to the loss of the mycarose sugar subunit (Figure 3.10a). IsoAnalyst therefore not only accurately detects SIL incorporation in MS features across multiple conditions, but can also group adducts and fragments that derive from the same compound, simplifying interpretation of labeling data for complex samples. In the following chapters I will demonstrate how correlating labeling patterns between ions can not only relate fragments and adducts of the same molecule, but assists in the association of biosynthetically related molecules across the entire metabolome.

3.4. Limitations of IsoAnalyst

3.4.1. Signal Intensity and Variability

Signal intensity is one of the main limiting factors in detecting compounds that are present in the extract, and has a known influence on accurate measurement of isotopologue ratios.[37] The analysis of erythromycin A at different concentrations shows that the accuracy and variability of the M_1:M_0 isotopologue is affected at both low and high signal intensities. Signal saturation has a clear detrimental impact on the overall measurement of isotopologue ratios, that would significantly affect the downstream analysis. Because saturation overestimates the M_1:M_0 ratio, an ion that is saturated in the unlabeled condition would result in higher false negatives in SIL detection. Signal saturation is a common problem and optimizing sample concentration is always an important aspect of MS metabolomics. Signals in the upper intensity ranges may still be peak picked and processed in standard MS metabolomics workflows, however saturated signals may cause significant downstream problems in the statistical analysis employed in IsoAnalyst. For this reason, it is especially important to test sample concentrations ahead of time and maintain consistency in sample preparation techniques.

At the lower end of signal intensity there are also limitations in data analysis. The main limiting factor is low concentrations of the compound in the sample either due to low production by the organism, ineffective extraction techniques, or inefficient ionization. The incorporation of singly labeled SIL compounds can exacerbate this by spreading the signal intensity across a larger number of isotopologue peaks, effectively lowering the signal intensity of individual isotopologue peaks compared to the unlabeled sample. In the case of erythromycin A, the ion intensity was consistently strong, all the way up to the M_8 isotopologue in the case of labeling by [1-^{13}C] propionate. In this ideal case, I showed how t-test statistics beyond the heaviest labeled isotopologue can have results which show that the isotopologue ratio is significantly smaller than the M_1:M_0 ratio of the unlabeled compounds. While the statistics calculated beyond the heaviest labeled isotopologue help verify the SIL incorporation in ions with sufficient signal intensity, it is not required for the identification of labeling. Lower abundance ions may have more SIL positions that are labeled than are able to be detected, however I have shown that accurate detection of SIL incorporation is possible as low as 8.0e3 ion intensity (Table 3.2, Figure 3.7d).

Another cause of low signal intensity that I have observed in these experiments is the inconsistent production of compounds across the panel of labeling conditions. The four SIL media conditions are fermented and analyzed separately, and this may influence the metabolism of the organism such that a compound is produced sufficiently in one SIL condition, but is not present or present in very small quantities in another. The purpose of growing paired controls with an unlabeled version of the precursor is to maintain similar nutrient environments so that the samples can be directly compared. Although this worked quite well for the compounds I observed and will discuss in the next chapter, this experimental design may be modified such that the medium used always contains unlabeled versions of every precursor. Doing so would potentially make the media conditions more consistent for compound production but, adds to the complication of setting up the experiment. This is a tradeoff that should be considered depending on the test organism.

3.4.2. Incomplete SIL Incorporation

A limitation of many studies that use SIL precursors is the interference of unlabeled nutrients which dilute the target biosynthetic precursor pools, and prevent complete labeling of target molecules. One approach to achieving compete labeling in the metabolome is to grow the organism in media containing a single carbon source of U-^{13}C glucose, as is common in metabolic fluxomics studies. Since the aim of IsoAnalyst is to selectively label certain classes of molecules, it was not practical to supplement the media with fully labeled carbon sources. Overall the IsoAnalyst approach is limited by the efficiency of SIL incorporation. As discussed in Chapter 2, these challenges are mitigated mostly in the development of media conditions and SIL compounds as the rate of SIL incorporation relies on the metabolic processes at play. This inherent limitation prevents IsoAnalyst from determining the true number of SIL precursors incorporated into any feature, but rather it identifies the number of SIL precursors that can be confidently detected in any given MS feature. This is important to consider when interpreting labeling data in terms of BGC information, as the absence of labeling cannot be conflated with the lack of particular substrate in the BGC structure prediction.

3.4.3. BGC regulation

IsoAnalyst only detects compounds which are produced under the media conditions used, and does not inherently elicit natural product production. The common challenges associated with silent BGCs apply to the IsoAnalyst experiments and workflow. One advantage of IsoAnalyst is that it is flexible and can be used in combination with other common methods for assessing genomic potential of micro-organisms. IsoAnalyst can be applied to many different media conditions, co-cultures, or in combination with other elicitors such as antibiotics and heavy metals. It can also be applied to genetically manipulated organisms such as knock-outs or constitutive promoters, as well as to heterologous hosts. The flexibility of IsoAnalyst allows for its application to any experiment that aims to find chemistry produced in a biological fermentation, where the genome of the organism is known.

3.5. Conclusion

Many of the SIL MS metabolomics algorithms described in the introduction of this chapter retain only the most abundant isotopologue peaks in their analysis. This is why there is a great advantage to targeted approaches that utilize SIL precursors with specific biosynthetic targets and distinct isotopic combinations. In the IsoAnalyst approach, I aim to identify the iterative incorporation of multiple SIL precursors that are labeled in a single position in both known and unknown compounds. Rather than targeting a specific compound or group of compounds, IsoAnalyst allows for the efficient association of MS features by their biosynthetic origin. To achieve this, I aimed to develop a method that detects SIL incorporation, the number of SIL precursors present, and aligns the data across parallel labeling conditions. IsoAnalyst compares isotopologue ratios between unlabeled controls and labeled samples to leverage information from the full isotopologue distribution, including highly labeled but low abundance isotopologue peaks. From this analysis it is possible to determine the minimum number of isotopic incorporation events for a given isotopically enriched feature. This analysis does not define the precise number of biosynthetic precursors in the structure, but rather the minimum number of detectable incorporation events, to provide overall labeling patterns for every feature under each SIL condition. In addition, it does not differentiate between direct biosynthetic incorporation and metabolic transformation of the feedstock prior to incorporation. Still, IsoAnalyst accurately detected SIL incorporation in all detected adducts and in-source fragment ions of erythromycin A across all four SIL conditions. In the subsequent chapters I will apply this method to the full metabolome of sequenced organisms to demonstrate how IsoAnalyst results relate to known biosynthetic pathways (Chapter 4) and discover novel chemistry from known biosynthetic pathways (Chapter 5).

3.6. Methods

3.6.1. UPLC-MS methods and data acquisition

All solvents used for UPLC and HPLC were Optima grade, and water used for chromatography was purified by a Milli-Q water purification system. All standard solutions and biological extracts were analyzed using a Waters Acquity I-class UPLC system coupled to a Waters Synapt G2Si qTOF mass spectrometer. Sample injections (5 µL) were subjected to chromatographic separation and mass spectrometric analysis. Chromatography was performed (Acquity HSS T3 1.8 µm, 2.1 x 100 mm) using a linear gradient (solvent A: H_2O + 0.01% formic acid, solvent B: acetonitrile + 0.01% formic acid) of 5-98% B over 5.8 minutes, a hold a 98% B for 0.3 min followed by a 1.8 minute re-equilibration at 5% B. All mass spectra were acquired in MS^E mode which is a data-independent acquisition (DIA) mode (Figure 3.2). The MS detector range was set to 50-1500 *m/z* in positive mode, with a capillary voltage of 3.5 kV, and a desolvation temperature of 200 ºC. A 0.4 second scan rate was used, with an alternating MS^1 and MS^2 scan acquisition. The MS^2 data is acquired and processed by the method shown in Figure 3.2, however only the MS^1 data is used in the IsoAnalyst processing workflow.

3.6.2. MS analysis of erythromycin A standard at varying concentrations

Solutions containing a commercial standard of erythromycin A (Sigma-Aldrich) were prepared at the concentrations 500 nM, 100 nM, 50 nM, 10 nM, 5 nM, 1 nM, 0.5 nM, and 0.1 nM in a 1:1 mixture of optima methanol and Milli-Q purified water. Five replicates of each concentration were analyzed. These data were pre-processed for peaking picking and feature detection by our in-house processing pipeline to produce the centroid peak lists and feature lists as shown in Figure 3.2. I then analyzed these data using the isotopologue scraping step of IsoAnalyst to detect all the naturally occurring isotopologues of the $[M+H]^+$ ion of erythromycin A. I wrote a custom Python 3 script to calculate the linear regression of the M_1:M_0 isotopologue ratio and Shapiro-Wilk statistics for the M_1:M_0 ratio of the erythromycin A in the commercial standard samples. Linear regression and Shapiro-Wilk statistics were both calculated using the scipy.stats package in Python 3.

3.6.3. IsoAnalyst program requirements

IsoAnlalyst is a custom script written in Python 3. Additional information regarding input requirements and user guides for IsoAnalyst are available on the GitHub repository (www.github.com/liningtonlab/isoanalyst). The data presented throughout this thesis were pre-processed using an in-house custom tool, designed to handle Waters MS^E data. This was done according to the workflow in Figure 3.2 to generate centroided peak lists and chromatographically aligned feature lists as .csv files for direct input to IsoAnalyst (Figure 3.1). IsoAnalyst also accepts generic data inputs that can be converted from most major vendor data types as described on the GitHub repository.

3.6.4. Bacterial strain and inoculum preparation

Saccharopolyspora erythraea ATCC 11635 (NRRL 2338) was purchased from ATCC (USA). Bacterial inoculum was prepared by first streaking a frozen glycerol stock on an ISP agar plate (3 g yeast extract, 5 g acid hydrolyzed casein, and 14 g of agar per liter of water). Single colonies were then selected to inoculate a 7 mL liquid culture of ISP media. Once turbid growth was observed in rich media, 50 µL of this culture was used to inoculate a 7 mL culture of the same minimal media to be used in the SIL experiment. After 24 hours of growth, this culture was used for the inoculation of the microtiter plates.

3.6.5. Parallel SIL experiment

A minimal medium was used for all SIL fermentation experiments (10 g of starch, 3.4 g of sodium glutamate, 0.4 g of KH_2PO_4, 1.2 g of K_2HPO_4, 1.0 g of $MgSO_4 \cdot 7H_2O$, 2.0 g of NaCl, 1.0 g of $CaCO_3$, 0.01 g of $FeSO_4 \cdot 7H_2O$, 1.5 mg $CuSO_4 \cdot 5H_2O$, 3.0 mg $ZnSO_4 \cdot 7H_2O$, 1.5 mg $CoSO_4 \cdot 7H_2O$, 1.5 mg of $MnSO_4 \cdot H_2O$, and 1.0 mg of $NaMoO_4 \cdot 2H_2O$ per liter of water). This medium was used for all SIL precursor conditions, except for the [1-^{15}N] glutamate and the corresponding unlabeled glutamate control. For these conditions the same minimal medium was prepared with 50% of the standard amount of unlabeled sodium glutamate (1.7 g/L instead of 3.4 g/L), and the remaining 50% was replaced with either [1-^{15}N] glutamate or unlabeled glutamate by sterile filtration at the time of inoculation.

Stable isotopically labeled compounds were purchased from Cambridge Isotope Laboratories, Inc. The corresponding unlabeled compounds were purchased from ThermoFisher Scientific. The SIL feedstock compounds, [99% 1-^{13}C]acetate, [99% 1-^{13}C]propionate, [99% methyl-^{13}C]methionine, and [98% 1-^{15}N]glutamate, and unlabeled version of each compound were prepared as stock solutions in Milli-Q water and sterilized by filtration (0.2 uM filter). The same culture was used as inoculum for all replicate wells of every feedstock condition in a given experiment. The 24-well microtiter plates and sandwich covers used for micro-scale bacterial cultures were purchased from Enzyscreen B.V. (The Netherlands) and the protocol for microtiter well plate fermentations was adapted from Duetz et al.[38]

The 24-well microtiter plates were cleaned and sterilized according to Duetz et al.,[38] and 2 mL of minimal media was added to each well. The first and last columns in each 24-well plate were left with sterile media and the inner 16 wells were inoculated with 80 µL of bacterial inoculum. Following inoculation, either an SIL compound or the corresponding unlabeled compound was added to each well by sterile filtration. Four replicate wells were prepared and inoculated for each condition, including unlabeled controls. Unlabeled control cultures were included for each feedstock condition to account for metabolic changes that may occur as a result of adding the precursor compound. Stock solution concentrations were adjusted according to the final desired concentration of each SIL or unlabeled precursor in the culture so that a minimal volume of 20-100 µL of stock solution was added to each well. The replicate cultures were fermented and analyzed separately and therefore account for technical variation in both the fermentation experiment as well as the analytical variation in the MS data.

Microtiter plates containing SIL supplemented bacterial cultures were shaken at 200 rpm and maintained at 23.0 °C for five days. On the fifth day the cultures were extracted by adding 2 mL of Optima methanol to each well. The contents of each well were then transferred to Eppendorf tubes, sonicated for 5 minutes, and centrifuged for 1 minute at 16,000 g. Methanol/water extracts were injected directly onto the UPLC-qTOF system, or diluted to maintain the most intense signals in the chromatogram in an optimal range for both sensitivity and mass accuracy.

References

1. Krug, D. & Müller, R. Secondary metabolomics: The impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products. *Nat. Prod. Rep.* **31**, 768–783 (2014).

2. Chokkathukalam, A., Kim, D. H., Barrett, M. P., Breitling, R. & Creek, D. J. Stable isotope-labeling studies in metabolomics: New insights into structure and dynamics of metabolic networks. *Bioanalysis* **6**, 511–524 (2014).

3. Weindl, D., Wegner, A. & Hiller, K. Metabolome-wide analysis of stable isotope labeling-Is it worth the effort? *Front. Physiol.* **6**, 189-- 3 (2015).

4. May, D. S. *et al.* 15N Stable Isotope Labeling and Comparative Metabolomics Facilitates Genome Mining in Cultured Cyanobacteria. *ACS Chem. Biol.* **15**, 758–765 (2020).

5. Simpson, J. P. *et al.* Metabolic Source Isotopic Pair Labeling and Genome-Wide Association Are Complementary Tools for the Identification of Metabolite-Gene Associations in Plants. *Plant Cell*. **33**, 492-510 (2021).

6. Gowda, H. *et al.* Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal. Chem.* **86**, 6931–6939 (2014).

7. Huang, X. *et al.* X13CMS: Global tracking of isotopic labels in untargeted metabolomics. *Anal. Chem.* **86**, 1632–1639 (2014).

8. Llufrio, E. M., Cho, K. & Patti, G. J. Systems-level analysis of isotopic labeling in untargeted metabolomic data by X13CMS. *Nat. Protoc.* **14**, 1970–1990 (2019).

9. Chokkathukalam, A. *et al.* mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics* **29**, 281–283 (2013).

10. Capellades, J. *et al.* geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics. *Anal. Chem.* **88**, 621–628 (2016).

11. Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A. & Breitling, R. PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Anal. Chem.* **83**, 2786–2793 (2011).

12. Hiller, K., Metallo, C. M., Kelleher, J. K. & Stephanopoulos, G. Nontargeted Elucidation of Metabolic Pathways Using Stable-Isotope Tracers and Mass Spectrometry. *Anal. Chem.* **82**, 6621–6628 (2010).

13. Hiller, K. *et al.* NTFD--a stand-alone application for the non-targeted detection of stable isotope-labeled compounds in GC/MS data. *Bioinformatics* **29**, 1226–1228 (2013).

14. Weindl, D., Wegner, A. & Hiller, K. MIA: Non-targeted mass isotopolome analysis. *Bioinformatics* **32**, 2875–2876 (2016).

15. Hillyer, K. E., Dias, D., Lutz, A., Roessner, U. & Davy, S. K. 13C metabolomics reveals widespread change in carbon fate during coral bleaching. *Metabolomics* **14**, 12 (2018).

16. Cobbold, S. A. *et al.* Non-canonical metabolic pathways in the malaria parasite detected by isotope-tracing metabolomics. *Mol. Syst. Biol.* **17**, 1–20 (2021).

17. Ćeranić, A. *et al.* Enhanced Metabolome Coverage and Evaluation of Matrix Effects by the Use of Experimental-Condition-Matched 13C-Labeled Biological Samples in Isotope-Assisted LC-HRMS Metabolomics. *Metabolites* **10**, 434 (2020).

18. Dong, Y., Feldberg, L. & Aharoni, A. Miso: an R package for multiple isotope labeling assisted metabolomics data analysis. *Bioinformatics* **35**, 3524–3526 (2019).

19. Feldberg, L., Venger, I., Malitsky, S., Rogachev, I. & Aharoni, A. Dual Labeling of Metabolites for Metabolome Analysis (DLEMMA): A New Approach for the Identification and Relative Quantification of Metabolites by Means of Dual Isotope Labeling and Liquid Chromatography−Mass Spectrometry. *Anal. Chem.* **81**, 9257–9266 (2009).

20. Feldberg, L., Dong, Y., Heinig, U., Rogachev, I. & Aharoni, A. DLEMMA-MS-Imaging for Identification of Spatially Localized Metabolites and Metabolic Network Map Reconstruction. *Anal. Chem.* **90**, 10231–10238 (2018).

21. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).

22. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).

23. Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr. Protoc. Bioinforma.* **46**, 13.24.1--13.24.9 (2014).

24. Theodoridis, G. A., Gika, H. G., Want, E. J. & Wilson, I. D. Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Anal. Chim. Acta* **711**, 7–16 (2012).

25. Slawski, M. *et al.* Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics* **13**, 291 (2012).

26. Murray, K. K. *et al.* Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure Appl. Chem.* **85**, 1515–1609 (2013).

27. Hellerstein, M. K. & Neese, R. A. Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. *Am. J. Physiol. Metab.* **276**, E1146–E1170 (1999).

28. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**, 105 (2007).

29. Xu, Y. *et al.* Evaluation of Accurate Mass and Relative Isotopic Abundance Measurements in the LTQ-Orbitrap Mass Spectrometer for Further Metabolomics Database Building. *Anal. Chem.* **82**, 5490–5501 (2010).

30. Erve, J. C. L., Gu, M., Wang, Y., DeMaio, W. & Talaat, R. E. Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *J. Am. Soc. Mass Spectrom.* **20**, 2058–2069 (2009).

31. Feith, A., Teleki, A., Graf, M., Favilli, L. & Takors, R. HILIC-Enabled 13C Metabolomics Strategies: Comparing Quantitative Precision and Spectral Accuracy of QTOF High- and QQQ Low-Resolution Mass Spectrometry. *Metabolites* **9**, 63 (2019).

32. Wang, Y. & Gu, M. The Concept of Spectral Accuracy for MS. *Anal. Chem.* **82**, 7055–7062 (2010).

33. Knolhoff, A. M., Callahan, J. H. & Croley, T. R. Mass Accuracy and Isotopic Abundance Measurements for HR-MS Instrumentation: Capabilities for Non-Targeted Analyses. *J. Am. Soc. Mass Spectrom.* **25**, 1285–1294 (2014).

34. Loos, M., Gerber, C., Corona, F., Hollender, J. & Singer, H. Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Anal. Chem.* **87**, 5738–5744 (2015).

35. Ghasemi, A. & Zahediasl, S. Normality tests for statistical analysis: A guide for non-statisticians. *Int. J. Endocrinol. Metab.* **10**, 486–489 (2012).

36. Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *Journal of the American Chemical Society*. **142**, 9097–9105 (2020).

37. González-Antuña, A., Rodríguez-González, P. & Alonso, J. I. G. Determination of the enrichment of isotopically labelled molecules by mass spectrometry. *J. Mass Spectrom.* **49**, 681–691 (2014).

38. Duetz, W. A. *et al.* Methods for Intense Aeration, Growth, Storage, and Replication of Bacterial Strains in Microtiter Plates. *Appl. Environ. Microbiol.* **66**, 2641–2646 (2000).

Chapter 4.

IsoAnalyst Case Studies: Connecting Chemical Phenotypes to Biosynthetic Gene Clusters of Model Organisms

4.1. Introduction

In this chapter I will demonstrate how the accurate detection of SIL incorporation described in Chapter 3 aligns with the biosynthesis of known compounds. IsoAnalyst enables the categorization of compounds detected in MS metabolomics experiments by their biosynthetic origin. I initially tested the IsoAnalyst platform on the type strain organisms *Saccharopolyspora erythraea* and *Amycolotopsis mediterranei*, which I selected for their efficient production of erythromycin A and rifamycin SV respectively.[1,2] The BGCs and biosynthetic machinery that produce these two compounds have served as model systems for understanding polyketide biosynthesis for decades.[3] I selected these systems in order to apply IsoAnalyst to a full MS metabolomics dataset where I was confident that sufficient SIL incorporation would occur. These type strains are optimized for production of their respective polyketide antibiotics, although their genomes both reveal various other BGCs that have not been investigated extensively.[1,2] Using IsoAnalyst I identified families of biosynthetic intermediates and analogues of both erythromycin A and rifamycin SV, as well as siderophore production in both organisms.

4.1.1. Differentiating Direct and Indirect SIL Incorporation

As I discussed in Chapter 2, a main challenge in using simple SIL precursors is that they can be transformed as substrates in primary metabolism prior to incorporation into natural products. [1-^{13}C]Acetate is particularly promiscuous due to its direct incorporation into the TCA cycle through acetyl-CoA (Figure 4.1). [1-^{13}C]Acetate has been added as a supplement to fermentations for the detection and analysis of polyketide structures for decades, however these studies often focus on specific compounds rather than looking at incorporation across the entire metabolome.[4–7] Indirect incorporation of [1-^{13}C]acetate into amino acids and other TCA-derived substrates is significant but variable under different metabolic conditions and therefore requires careful consideration. To account for common SIL precursor turnover in primary metabolism, I created a Substrate Labeling Table (Table 4.1) by considering the metabolic fate of [1-^{13}C]acetate, [1-^{13}C]propionate, [methyl-^{13}C]methionine, and [1-^{15}N]glutamate in common central metabolic pathways. The rows correspond to a list of common biosynthetic substrates in natural product biosynthesis and the columns to each SIL condition used in this study (Table 4.1). The Substrate Labeling Table designates the maximum theoretical incorporation events of each SIL precursor into each substrate.

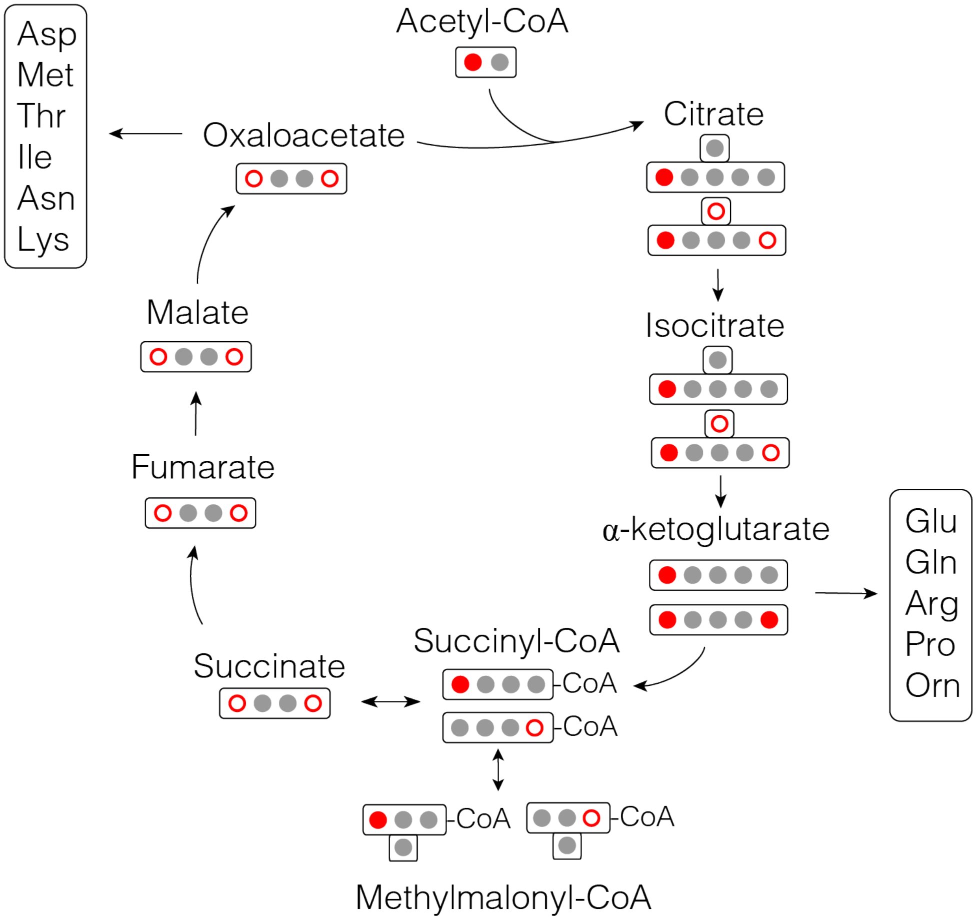


Figure 4.1 [1-^{13}C]Acetate Incorporation in the TCA Cycle
Red filled circles represent ^{13}C derived from [1-^{13}C]acetate following its direct transformation to acetyl-CoA. Open circles represent positions where the location of the ^{13}C atom is ambiguous due to the symmetry of succinate. Only one position represented by an open circle may be labelled in a molecule. Amino acids derived from TCA cycle intermediates are indicated.

The amino acids that are derived from TCA cycle intermediates (Figure 4.1) are designated in Table 4.1 to be labeled by [1-^{13}C]acetate a maximum of one or two times. Terpenes may be labeled by [1-^{13}C]acetate only in organisms with the mevalonate pathway[8] (Table 4.1). The [1-^{15}N]glutamate SIL condition is intentionally designed to be promiscuous and label all amino acids and other subunits derived from amino acids (Table 4.1). Although the [1-^{15}N]glutamate condition is designed to incorporate ^{15}N into any substrate, in theory some amino substrate acids will be much more easily labeled than others. Amino acids that have an α-nitrogen that is derived directly from transamination by glutamate, are more likely to be labeled by [1-^{15}N]glutamate. The biosynthesis of some amino acids are more tightly regulated than others such as histidine, arginine, and tryptophan due to their higher ATP demand.[9] These amino acids are more likely to be derived from recycled sources and will potentially have less ^{15}N incorporation from [1-^{15}N]glutamate than other amino acids. The growth medium used in this experiment is quite limited in terms of nitrogen metabolism, as either unlabeled or [1-^{15}N]glutamate are the only nitrogen sources. Although this was designed with the intention of labeling any nitrogen position with ^{15}N, the fact that glutamate is the sole nitrogen source in the media restricts the metabolism such that not all amino acids will be supplied in high quantities. This medium was selected for the optimization of [1-^{13}C]acetate incorporation, as described in Chapter 2, but may not be ideal for optimal ^{15}N incorporation into all amino acids. It is likely that under these conditions the organism is limited by amino acid supply, and therefore this media is probably not optimal for eliciting biosynthesis of amino acid-containing natural products such as NRPSs. These limitations highlight the importance of considering the test organism's ability to synthesize essential building blocks when optimizing the IsoAnalyst approach to that particular organism's metabolism.

Unlike [1-^{13}C]acetate and [1-^{15}N]glutamate, [1-^{13}C]propionate and [methyl-^{13}C]methionine are expected to label only specific substrates (Table 4.1). [1-^{13}C]Propionate is directly incorporated into propionyl-CoA and methylmalonyl-CoA, which are both common PKS substrates. [1-^{13}C]Propionate may also label succinyl-CoA by the conversion of methylmalonyl-CoA to succinyl-CoA.[10] [methyl-^{13}C]Methionine labels compounds which have a methyl group derived from S-adenosyl methionine (SAM). The SIL incorporation events referenced in this table are based on current knowledge of the metabolism of the four SIL precursors used here, but they are by no means a comprehensive overview of all the potential metabolic pathways that these four compounds can undergo. There is long standing evidence that alternative pathways for

Table 4.1 Substrate Labelling Table

Substrate	A	P	M	G
Glu	2			1
Gln	2			2
Arg	2			4
Pro	2			1
Orn	2			2
Asp	1			1
Met	1		1	1
Thr	1			1
Ile	1			1
Asn	1			2
Lys	1			2
Ala				1
Leu				1
Val				1
Phe				1
Tyr				1
Trp				2
Ser				1
Gly				1
Cys				1
His				3
Unknown amino acid				1+
Acetyl or Malonyl-CoA	1			
Propionyl-CoA		1		
Methylmalonyl-CoA	1	1		
Methoxymalonate			1	
Hydroxymalonate				
Methyl			1	
IPP[a]	2			
DMAPP[a]	2			
GPP[a]	4			
FPP[a]	6			
Succinyl-CoA	1	1		
Amino-saccharide				1
Amino group				1

[a]mevalonate pathway only

amino acid and PKS monomer biosynthesis can not only be influenced by genetics but environmental factors.[11] Other common pathways such as the glycoxylate cycle, and gluconeogenesis should be considered under metabolic circumstances where they are relevant. The Substrate Labeling Table (Table 4.1) can be modified to account for other such pathways or to include more SIL precursors.

4.1.2. Application of IsoAnalyst to the Full Metabolome

The summary output file of IsoAnalyst described in Chapter 3 contains every feature from the ground truth feature list, and the corresponding SIL incorporation for each condition. This summary file was filtered automatically to contain only those MS features that had detected isotope incorporation in two or more of the SIL conditions. This initial filtering reduces the number of primary metabolites present in the output. Although a compound with labeling in only one SIL condition may be a natural product, this is not sufficient to hypothesize which BGC is responsible for producing the labeled compound. There is an inherent bias in the selection of SIL precursors as to which types of natural products will be able to be identified. This element of the experimental design is flexible and can be adapted to different BGC classes. I then manually interrogated the summary file to filter features based on chromatographic peak shape and signal intensity. Features were eliminated if either of these factors interfered with the accuracy of the SIL detection. I manually grouped features that represented adducts or fragments of the same compound, and further compared SIL incorporation patterns between different compounds to relate compounds originating from the same BGC. Compound identities were confirmed by a combination of SIL incorporation, MS/MS fragmentation, and NMR when applicable. I did this for a full parallel SIL experiment using [1-^{13}C]acetate, [1-^{13}C]propionate, [methyl-^{13}C]methionine, and [1-^{15}N]glutamate in *S. erythraea* and *A. mediterranei*.

4.2. Saccharopolyspora erythraea

S. erythraea has had a massive impact on biosynthesis research.[12] Much of the basis for what we know about polyketide synthases today began with preliminary work on the 6-deoxyerythronolide B synthase (DEBS). Early studies on the biosynthesis of erythromycin A used [^{14}C]-, [^{13}C]-, [^{18}O]-, and [^{2}H]- labeled substrates to elucidate the steps of the biosynthetic pathway, before the BGC had even been discovered.[13] It was not until the genetic characterization of DEBS that the complex biochemical basis of polyketide formation was beginning to be understood.[14,15] Complete heterologous expression of erythromycin A has been accomplished[16] and this well-understood pathway has been manipulated countless times for drug development.[17] Furthermore, due to the extensive studies of this biosynthetic pathway, each biosynthetic intermediate has been characterized and described in the literature.[16] I have used the example of erythromycin A throughout the last two chapters because it is so often used as a model system for biosynthesis and there is strong evidence for the expected SIL incorporation pattern. Here I will focus on the overall metabolome of *S. erythraea.*

The SIL incorporation into erythromycin A was used as the primary example in Chapter 3 to confirm the accurate detection of SIL incorporation into a known compound. This example demonstrated that the full expected labeling of each SIL precursor could be detected in all of the ion adducts and in-source fragments of erythromycin A. In this section I will look at the full MS metabolomic dataset for *S. erythraea* as generated by the SIL experiment and data analysis described Chapter 3. The ground truth feature list for this *S. erythraea* experiment contained 786 unique features, and IsoAnalyst identified 147 which had SIL incorporation in two or more conditions. I further filtered this to 94 features with reliable SIL detection on the basis of chromatographic peak shape. I identified 71 of these features corresponding to erythromycin A and five additional compounds with SIL incorporation patterns related to erythromycin A (Figure 4.2). The differential labeling in the biosynthetic intermediates of the erythromycin pathway demonstrate the strength of IsoAnalyst to efficiently relate MS features corresponding to the same BGC. Furthermore, seven of the labeled features detected by IsoAnalyst corresponded to a siderophore, erythrochelin, which has previously been isolated from *S. erythraea*[18] (Figure 4.2).

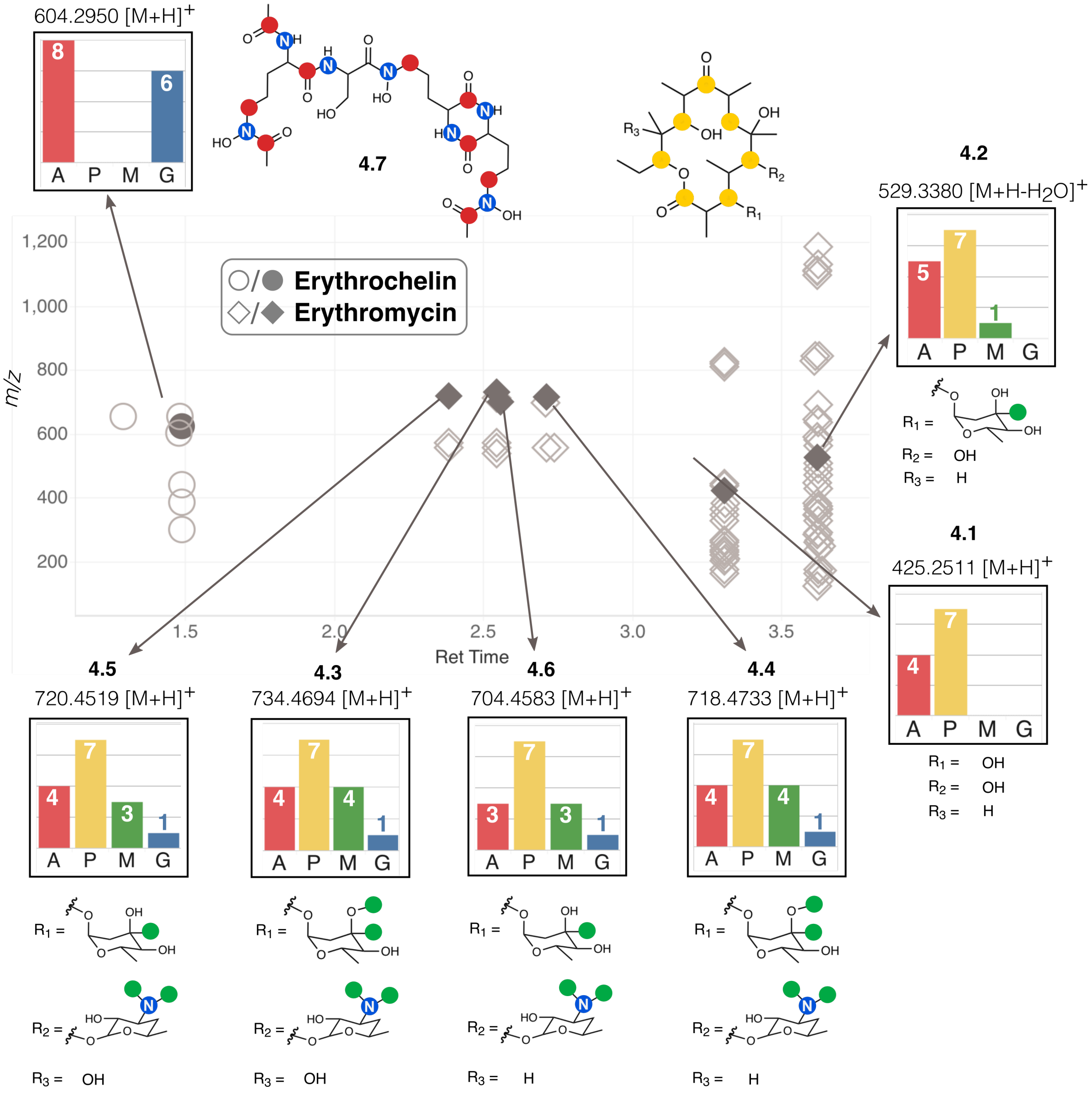


Figure 4.2 Overview of Labelled Ions Detected from *S. erythraea*
IsoAnalyst profiles and structures for selected labeled ions detected in the *S. erythraea* metabolome. Seventy-seven *m/z* features are shown which were identified as having SIL incorporation in two or more conditions, and having realistic ion intensity and chromatographic peak shape. Seventy-one features indicated as diamonds were identified as erythromycin A or related structures, and seven features indicated as circles were identified as erythrochelin. The IsoAnalyst results for selected ions are shown.

4.2.1. Erythromycin Family

The 14-member polyketide macrolide core, 6-deoxyerthronolide B, is biosynthesized from a propionyl-CoA starter unit and six methylmalonyl-CoA extender units by the large multi-domain PKS, 6-Deoxyerythonolide B Synthase (DEBS)[12] (Figure

4.3). All of the erythromycin compounds identified using IsoAnalyst had identical SIL incorporation in the [1-^{13}C]acetate and [1-^{13}C]propionate conditions, indicating this shared polyketide core. By contrast, labeling in [methyl-^{13}C]methionine and [1-^{15}N]glutamate varied between the five products, suggesting different degrees of decoration of the polyketide core (Figure 4.2). The polyketide product produced by DEBS, 6-deoxyerythronolide B (Figure 4.3), was not observed in this experiment. However, using IsoAnalyst I putatively identified the subsequent six biosynthetic intermediates, erythronolide B (**4.1**), O-α-mycarosylerythronolide B (**4.2**), erythromycin D (**4.3**), B (**4.4**), C (**4.5**), and A (**4.6**) (Figures 4.2, 4.3).

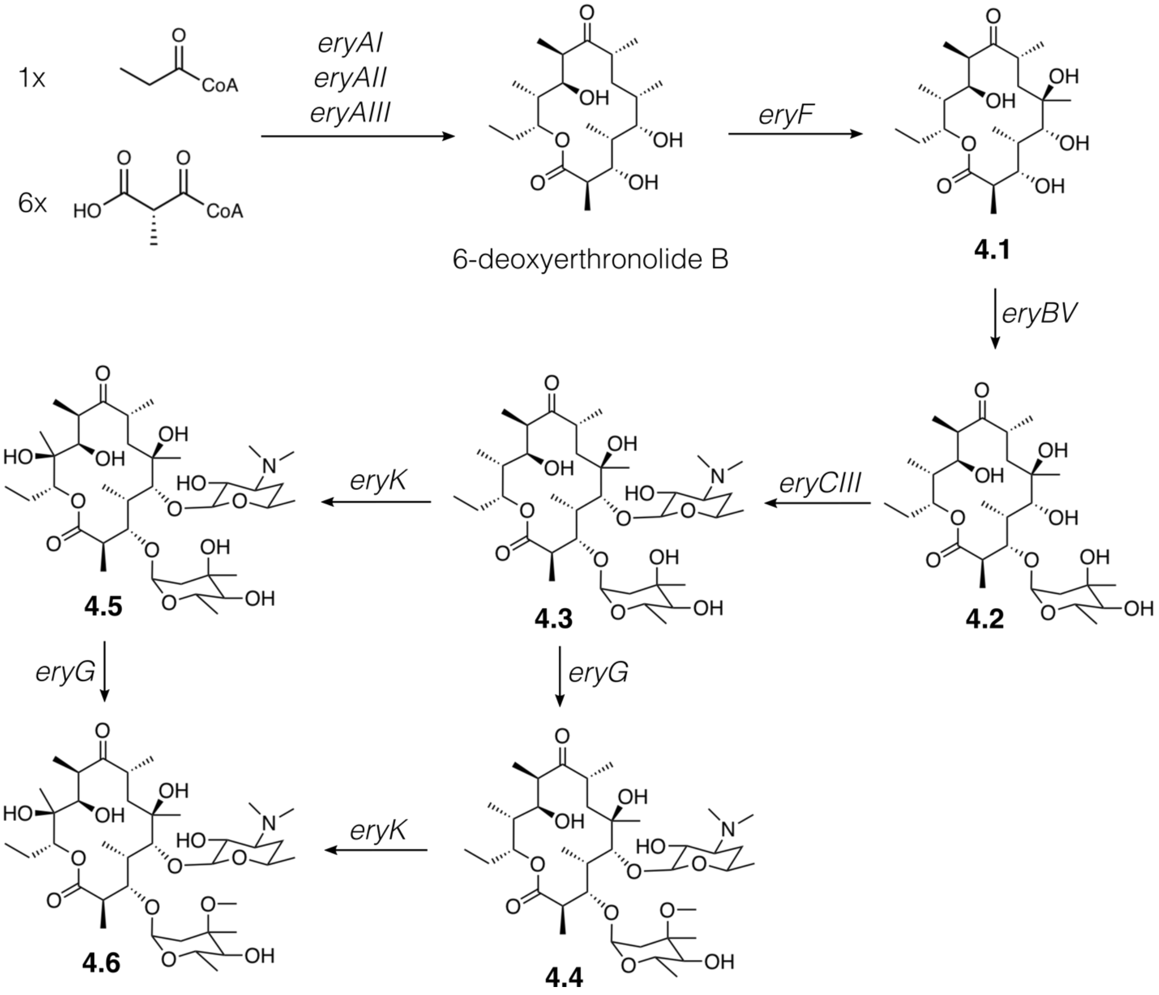


Figure 4.3 Biosynthesis of Erythromycin A

Biosynthetic scheme showing the steps in the biosynthesis of erythromycin A. The direct product of DEBS, 6-deoxyerythronolide B, is the only intermediate shown which was not detected in this experiment. All of the subsequent biosynthetic products have varying degrees of hydroxylation, methylation, and glycosylation, but contain the same polyketide core.

Erythronolide B (**4.1**) was not labeled by either [methyl-^{13}C]methionine or [1-^{15}N]glutamate, consistent with the proposed structure containing only the core macrocycle (Figure 4.2). 3-O-α-Mycarosylerythronolide B (**4.2**) contained a single labeled position in [methyl-^{13}C] methionine, consistent with the addition of the mycarose sugar, which contains a single site of SAM methylation. I isolated and confirmed the identity of **4.1** and **4.2** by NMR and UPLC-MS co-injection (Appendix A, Figures A1-A6). Erythromycin D (**4.3**) was putatively identified having three labeled positions in the [methyl-^{13}C]methionine condition and one position in the [1-^{15}N]glutamate condition, due to the addition of the desosamine sugar (Figures 4.2, 4.3). The final two reactions in the erythromycin pathway consist of a hydroxylation at C-12 of the macrolactone core and 3”-O-methylation on the mycarose sugar (Figure 4.3). These reactions may occur in a variable order resulting in the products **4.4** and **4.6**, which have one additional position labeled by [methyl-^{13}C]methionine compared to **4.3**, and **4.5**, which has the same SIL incorporation as **4.3** in all four conditions (Figure 4.2). Erythromycin A (**4.6**) was also discussed in detail in Chapter 3, the same SIL incorporation results are shown here (Figure 4.2). The identity of **4.6** was confirmed by UPLC-MS co-injection with an authentic standard (Appendix Figure A7). These data align with what is known about erythromycin biosynthesis and support the putative identification of **4.3**, **4.4**, and **4.5**.

4.2.2. Erythrochelin

Erythrochelin (**4.7**) is a hydroxomate siderophore which has previously been isolated from *S. erythraea*.[18] Erythrochelin is produced by the tetramodular nonribosomal peptide synthase (NRPS) ErcD.[18,19] The core substrates used in the biosynthesis of erythrochelin are L-ornithine (3), L-serine (1), and acetyl-CoA (3) (Figure 4.4). L-Ornithine is first hydroxylated by the δ-N-ornithine monooxygenase, ErcB. L-δ-N-hydroxyornithine is acetylated twice in the starter unit of erythrochelin, followed by the condensation of serine, L-δ-N-hydroxyornithine (hOrn) and L-δ-N-acetyl-δ-N-hydroxyornithine (haOrn). Direct ^{13}C incorporation into acetyl-CoA by [1-^{13}C]acetate, and ^{15}N incorporation into ornithine and serine by [1-^{15}N]glutamate, would result in an expected incorporation of three ^{13}C atoms from [1-^{13}C]acetate, and seven ^{15}N atoms from [1-^{15}N]glutamate (Figure 4.4). However, Table 4.1 indicates that ornithine is derived from the TCA intermediate α-ketoglutarate, which can also be labeled in two positions by [1-^{13}C]acetate. This results in a theoretical maximum labeling of erythrochelin by nine ^{13}C positions derived from [1-^{13}C]acetate, with six of these incorporation events occurring via labeling of the ornithine carbon skeleton (Figure 4.4).

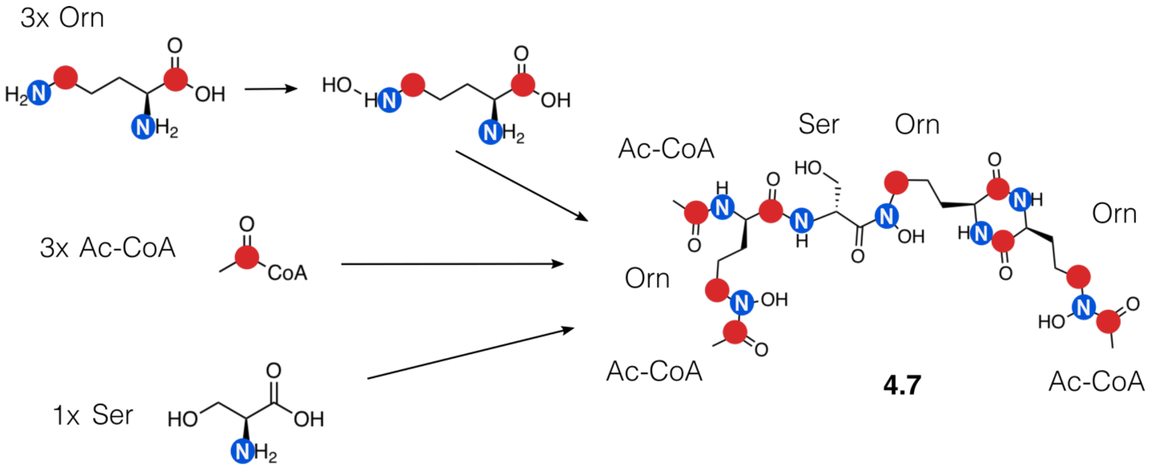


Figure 4.4 Biosynthesis of Erythrochelin
Scheme showing the biosynthetic substrates that make up erythrochelin. Expected positions of SIL tracer incorporation are indicated by colored circles. Ornithine is first hydroxylated to produce hOrn and three units of hOrn are later acetylated to produce haOrn. The NRPS condenses three units of haOrn and one serine to generate the final product, erythrochelin.

Both the iron-chelated and protonated adduct ions of **4.7** were detected by IsoAnalyst to have eight positions labeled by [1-^{13}C]acetate, and six positions labeled by [1-^{15}N]glutamate (Figure 4.5). A more detailed interrogation of the SIL data confirmed the identity of compound **4.7** through a combination of fragment *m/z* values, which matching those previously reported for erythrochelin[18] (Figure 4.6). The two larger fragment ions differ by the loss of a serine residue[18] and each fragment had detectable ^{15}N enrichment in 4 out of 5 positions and 3 out of 4 positions respectively (Figure 4.6a,b). This indicates that relative differences in SIL incorporation are detectable using IsoAnalyst, even when the SIL incorporation is not complete for every available position in the compound. The expected [1-^{13}C]acetate labeling was for 5 positions in both of these fragment ions, as serine does not have expected labeling by [1-^{13}C]acetate (Table 4.1). Four out of five positions were detected as having enrichment in [1-^{13}C]acetate for both fragment ions in Figure 4.6a and Figure 4.6b. The smaller fragment ions are the L-δ-N-acetyl-δ-N-hydroxyornithine subunit (Figure 4.6c) and δ-N-hydroxyornithine (Figure 4.6d).[18] These fragments show ^{13}C enrichment in three positions for L-δ-N-acetyl-δ-N-hydroxyornithine, and two positions for the δ-N-hydroxyornithine fragment (Figure 4.6c,d). Although the M_1 isotopologue is the most intense isotopologue peak in the [1-^{13}C]acetate mass spectrum in Figure 4.6d, the ^{13}C enrichment in the M_2 isotopologue peak was determined to be statistically significant for ^{13}C enrichment using IsoAnalyst. These data confirm that two positions of ornithine are labeled by [1-^{13}C]acetate as predicted in the Substrate Labeling Table (Table 4.1), even though the M_2 isotopologue peak has a lower relative intensity than the M_0 and M_1 peaks (Figure 4.6d).

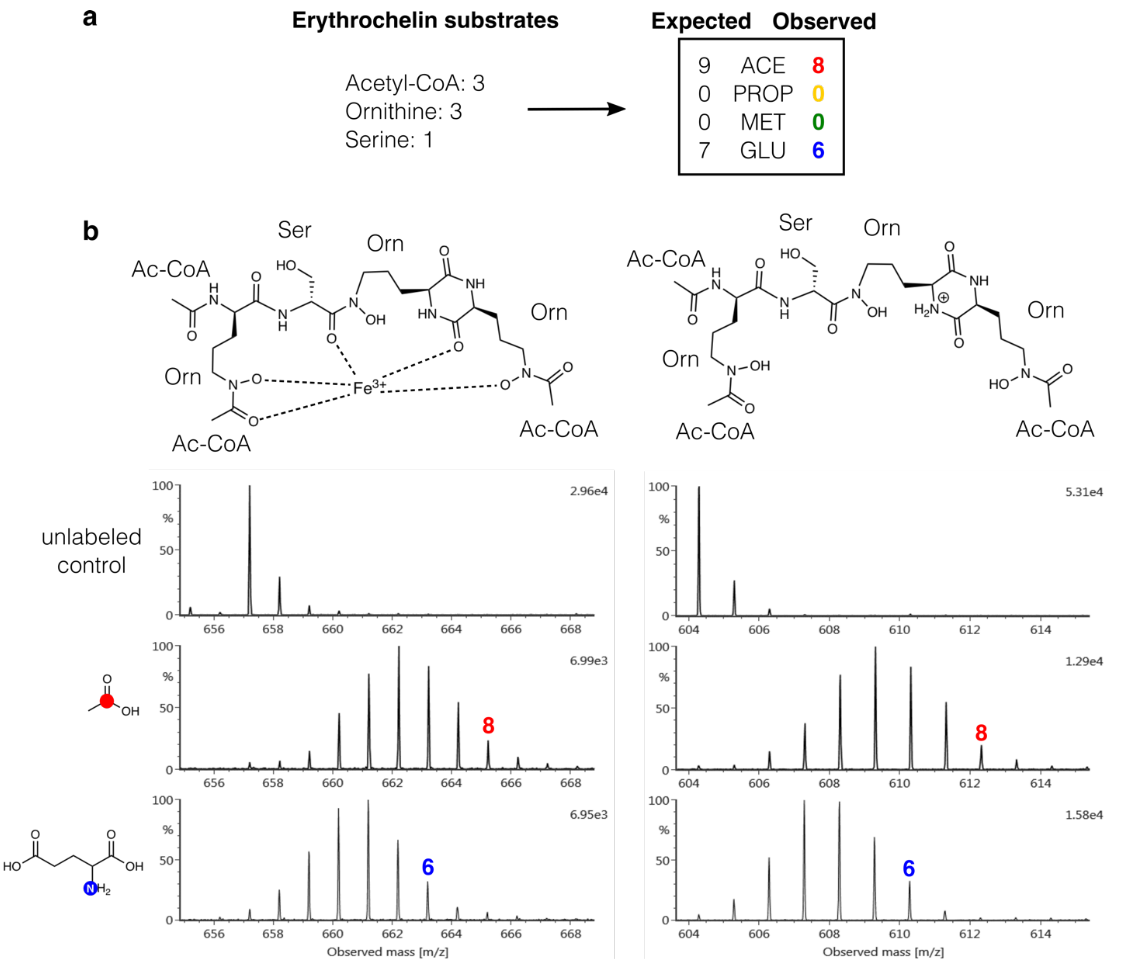


Figure 4.5 SIL Incorporation in Erythrochelin

(a) Substrates used in the biosynthesis of **7**, and both the expected and observed labeling of erythrochelin. (b) Structures and mass spectra for the iron adduct (*m/z* 657.2064) and the protonated adduct (*m/z* 604.2950) of **7** in the [1-^{13}C]acetate and [1-^{15}N]glutamate conditions. [1-^{13}C]Propionate and [methyl-^{13}C] methionine conditions are not shown as **7** was not produced under the [1-^{13}C] propionate condition and no SIL incorporation occurred under the [methyl-^{13}C]methionine condition.

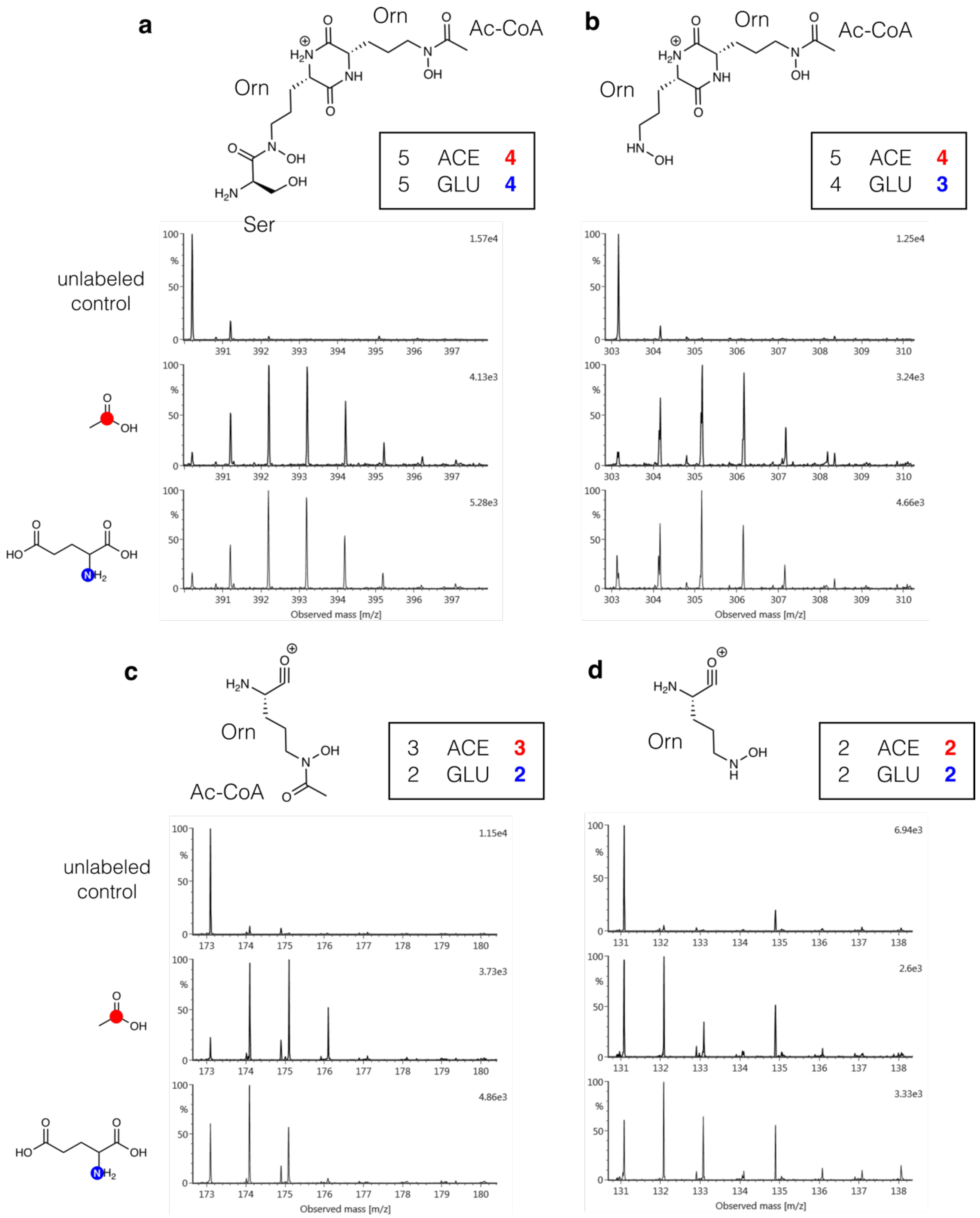


Figure 4.6 SIL Tracer Incorporation in Erythrochelin Fragment Ions
Mass spectra of fragments *m/z* 390.1990 (a), *m/z* 303.1666 (b), *m/z* 173.0932 (c) and *m/z* 131.0827 (d) under [1-^{13}C]acetate and [1-^{15}N]glutamate conditions indicate that the detected SIL incorporation is within the expected labeling maximums for the biosynthetic subunits derived from ornithine and acetyl-CoA.

4.3. Amycolatopsis meditteranei

A. meditteranei produces the antibiotic rifamycin SV, which was first isolated with a mixture of rifamycins in 1959 when the organism was classified as *Streptomyces meditteranei*.[20] Although rifamycin SV is the more potent antibiotic, it is a biosynthetic precursor of rifamycin B, and two *A. meditteranei* strains have been sequenced for specialization in producing either rifamycin B or SV.[2,21] *A. meditteranei* U32 is the commercial producer of rifamycin SV, as it contains a mutation in the P450 gene responsible for converting rifamycin SV to rifamycin B.[2] The biosynthesis of the rifamycins have been well-studied as a model PKS system making this organism a suitable test case for assessing SIL incorporation using IsoAnalyst

4.3.1. Rifamycin

Rifamycin SV is a macrocyclic polyketide which is biosynthesized from the starter unit 3-amino-5-hydroxybenzoic acid (AHBA), followed by chain extension with eight units of methylmalonyl-CoA and two units of malonyl-CoA.[22] The rifamycin BGC contains genes for the biosynthesis of AHBA[22] and this starter unit is only expected to be labeled by [1-^{15}N]glutamate. The rifamycin backbone undergoes a rearrangement to form a five-membered ring and ketal structure.[23] In the process of this transformation, the methyl (C-3) of one propionate unit is lost,[23] however, this does not affect the number of expected positions labeled in my experiment, as the SIL tracer [1-^{13}C]propionate is labeled in the carbonyl (C-1) position (Figure 4.7). Finally, the core polyketide is methylated by SAM, and acetylated to form rifamycin SV (Figure 4.7). Using Table 4.1 to account for both direct and indirect SIL incorporation, I determined the expected labeling of rifamycin SV to be 11 x [1-^{13}C]acetate, 8 x [1-^{13}C]propionate, 1x [methyl-^{13}Cmethionine], and 1 x [1-^{15}N]glutamate positions (Figure 4.7). While rifamycin SV is most commonly discussed in the literature, I only observed the oxidized form, rifamycin S in the following experiments. I confirmed the presence of rifamycin S by co-injection with a commercial standard (Appendix Figure A8).

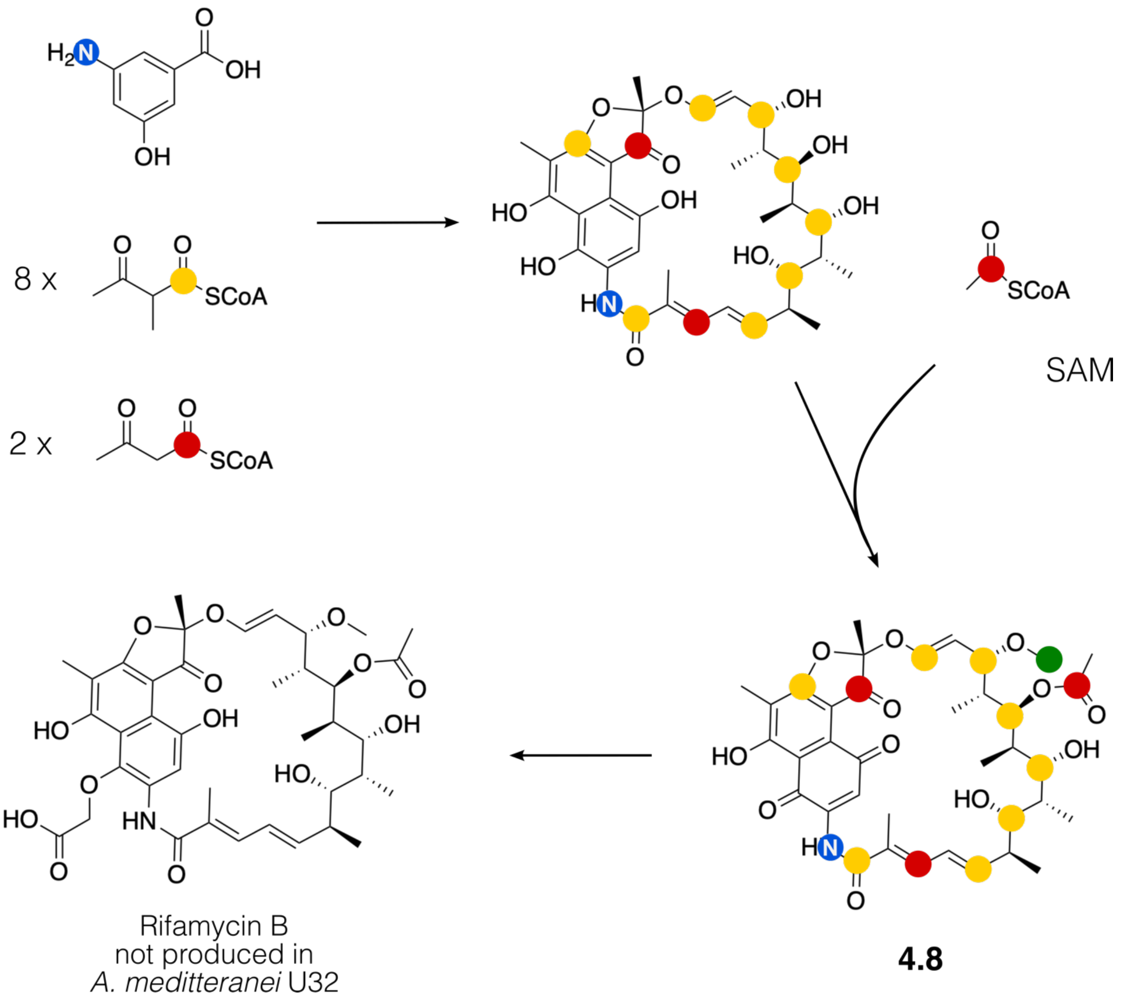


Figure 4.7 Biosynthesis and Expected SIL Incorporation in Rifamycin S
Biosynthetic scheme showing the building blocks and positions which are expected to have SIL incorporation in rifamycin S (**4.8**). Eight units of methylmalonyl-CoA, two of malonyl-CoA and one AHBA are used to make the polyketide core, followed by an acetylation and methylation by SAM to form **4.8**.

Unlike the *S. erythraea* experiment, I was unable to detect different biosynthetic precursors or analogues to rifamycin S. However, I did detect a series of in-source fragments of rifamycin S which demonstrate how IsoAnalyst can assist in interpreting the positions of SIL incorporation by analyzing fragment ions. In addition to the [M+H]$^+$ adduct (*m/z* 696.2994), I observed the subsequent loss of the O-methyl group (*m/z* 664.2709), the acetyl group (*m/z* 604.2562), and H_2O (*m/z* 586.2454) (Figure 4.8). The [M+H]$^+$ adduct had the exact expected SIL tracer incorporation in the [1-^{13}C]propionate, [methyl-13C]methionine, and [1-^{15}N]glutamate conditions, but slightly less labelling (9 out of 11 expected positions) in the [1-^{13}C]acetate condition Figure 4.8). As was the case with the erythromycins, this discrepancy is likely due to the indirect incorporation of [1-

^{13}C]acetate into the methylmalonyl-CoA subunits of rifamycin S. The fragment ions had decreased SIL incorporation in the [1-^{13}C]acetate and [methyl-^{13}C]methionine conditions, as expected from the loss of the O-methyl and acetyl groups (Figure 4.8).

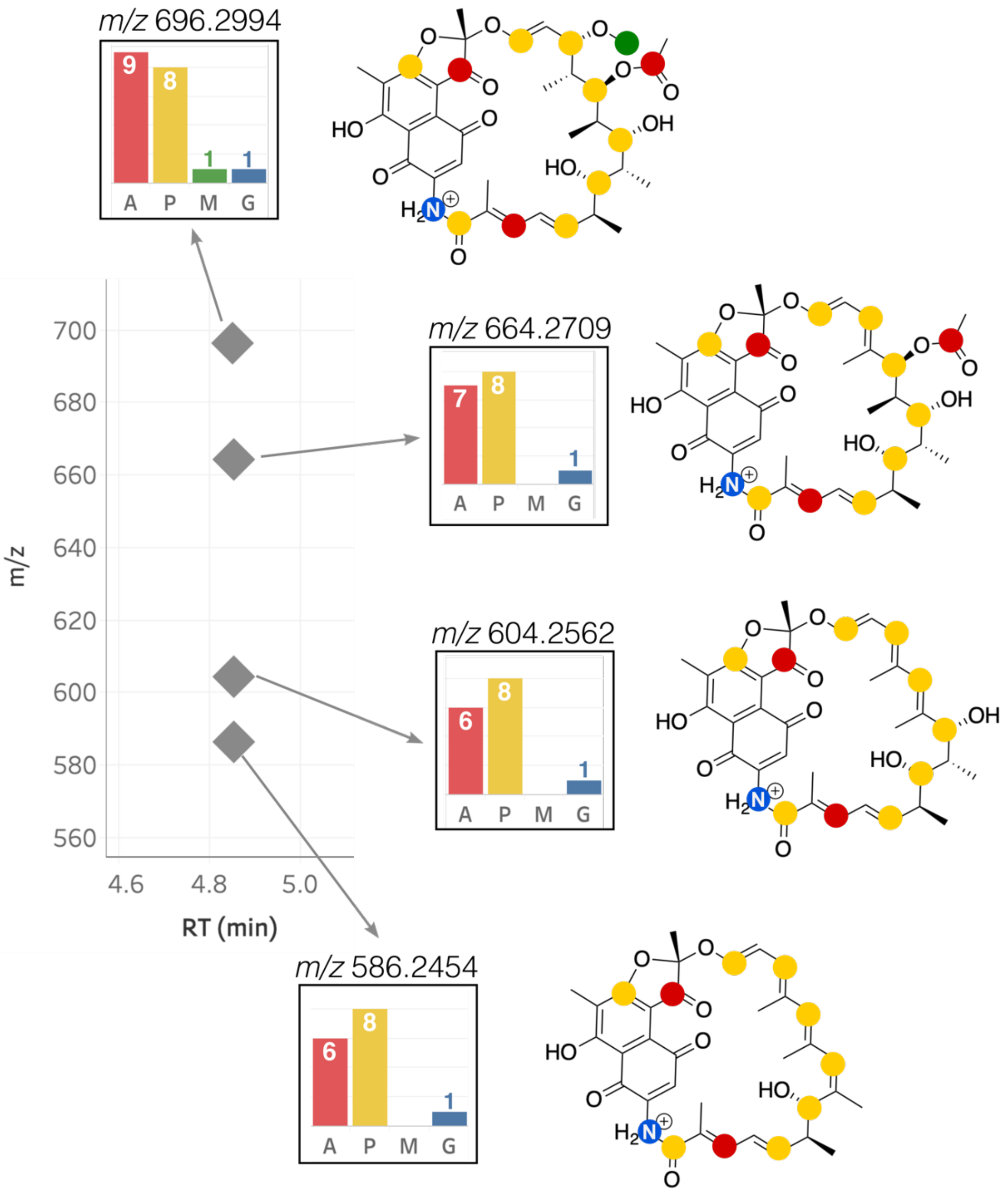


Figure 4.8 SIL Tracer Incorporation in Rifamycin S and Fragment Ions
Four MS features were detected with SIL tracer incorporation patterns corresponding to the in-source fragment ions of rifamycin S (**4.8**). The *m/z* 696.2994 is the [M+H]$^+$ ion. The *m/z* 664.2709 fragment has a loss of the O-methyl group. The *m/z* 604.2562 fragment has an additional loss of the acetate group, indicated by the decrease from 7 to 6 detected ^{13}C positions in the [1-^{13}C]acetate condition. The *m/z* 586.2454 has a further dehydration.

4.3.2. Unknown Siderophore

I detected a group of interesting MS features in the *A. meditteranei* experiment which had related labeling patterns and an iron adduct, similar to erythrochelin (Figure 4.9). In addition to the iron (*m/z* 700.2448) and protonated (*m/z* 647.3337) adducts, I detected three smaller in-source fragments at the same retention time with similar SIL incorporation patterns (Figure 4.9). SIL incorporation was detected for 5 x [1-^{13}C]acetate, 2 x [1-^{13}C]propionate, 3 x [methyl-^{13}C]methionine, and 6 x [1-^{15}N]glutamate units in the protonated adduct of this unknown siderophore (Figure 4.9). The iron adduct had more SIL tracer detected in the [1-^{13}C]acetate and [1-^{15}N]glutamate conditions, however, this method does not account for the contribution of ^{54}Fe to the isotopologue ratios, so the SIL tracer incorporation in the protonated adduct is likely more reliable. Notably, this siderophore has three methylated positions, only one of which is lost in the smallest fragment ion detected (*m/z* 475.2508, Figure 4.9).

The current antiSMASH database contains the complete BGC analysis output for *A. meditteranei* U32 performed in antiSMASH 5.0.[24,25] I identified a candidate BGC for the production of the unknown siderophore, of which 80% of the genes present share similarity to genes in the BGC for the known siderophore, scabichelin (**4.9**, Figure 4.10). These similarity metrics are automatically generated in antiSMASH, to link to related BGCs in the MIBiG database.[26] The scabichelin BGC contains five NRPS modules, two of which contain N-methyltransferase domain[27] (Figure 4.10a). The scabilchelin BGC has been previously characterized and associated with scabilchelin through mutation studies.[27] The diagram in Figure 4.10a is of the scabilchelin BGC modules which are provided in the MIBiG database.[26] The BGC located in ‘region 18’ of the *A. meditteranei* U32 genome also contains five modules, although the predicted amino acids are slightly different. Three N-methyltransferase domains were detected in region 18, consistent with the [methyl-^{13}C]methionine incorporation detected in the unknown siderophore (Figure 4.10). Additionally, two units of [1-^{13}C]propionate were consistently

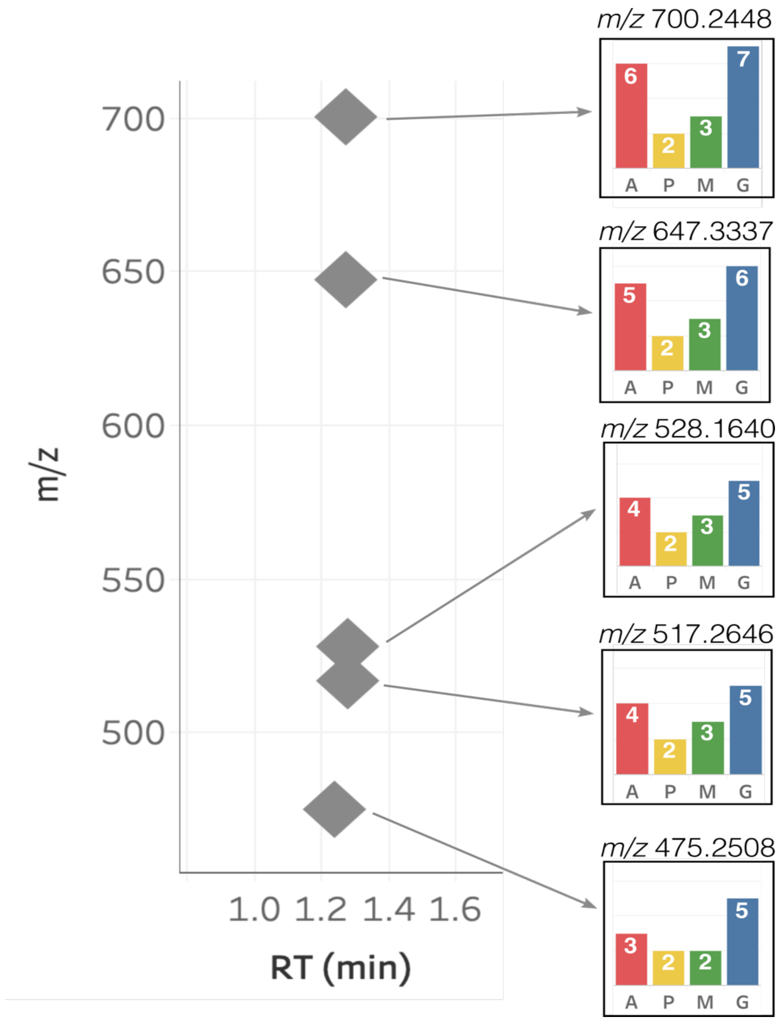


Figure 4.9 SIL Tracer Incorporation into an Unknown Siderophore
Five MS features were detected with SIL tracer incorporation patterns corresponding to an unknown siderophore which eluted at 1.27 min. The *m/z* 700.2448 feature is an $[M\text{-}2H\text{+}Fe]^+$ ion and the *m/z* 647.3337 feature is an $[M\text{+}H]^+$ ion. The remaining *m/z* features are unidentified in-source fragment ions.

detected in all of the fragment ions associated with the unknown siderophore, suggesting the presence of an amino acid that can be biosynthetically derived from [1-^{13}C]propionate. Observation of the raw MS data indicates that the M_0 peak of the protonated adduct (*m/z* 647.3337) was the base peak in the [1-^{13}C]propionate condition, despite the fact that the M_2 isotopologue peak was detected as having statistically significant SIL incorporation (Figure 4.11). This highlights both an advantage of IsoAnalyst, in that is it is sensitive to detecting SIL incorporation in higher isotopologues, as well as a limitation in the interpretation of the IsoAnalyst output. Observation of the raw data in this case clearly indicates that the 2 x [1-^{13}C]propionate incorporation occurs at a much lower ratio than the 3 x [methyl-^{13}C]methionine incorporation into the same molecule (Figure 4.11). These data suggest that the [1-^{13}C]propionate labelling in this siderophore originates from an indirect source, and orthogonal metabolomic approaches such as fluxomics or transcriptomic methods would assist in identifying the metabolic pathways involved. I was not able to determine the exact structural identity of this siderophore, however it is likely a hydroxamate siderophore containing two units of L-δ-N-acetyl-δ-N-hydroxyornithine, one N-methylated L-serine, and two unknown amino acids that may be indirectly labelled by [1-^{13}C]propionate. Further investigation of this compound may reveal a novel structure, and would provide evidence for the association of this siderophore with the BGC identified in region 18 of the *A. meditteranei* U32 genome.

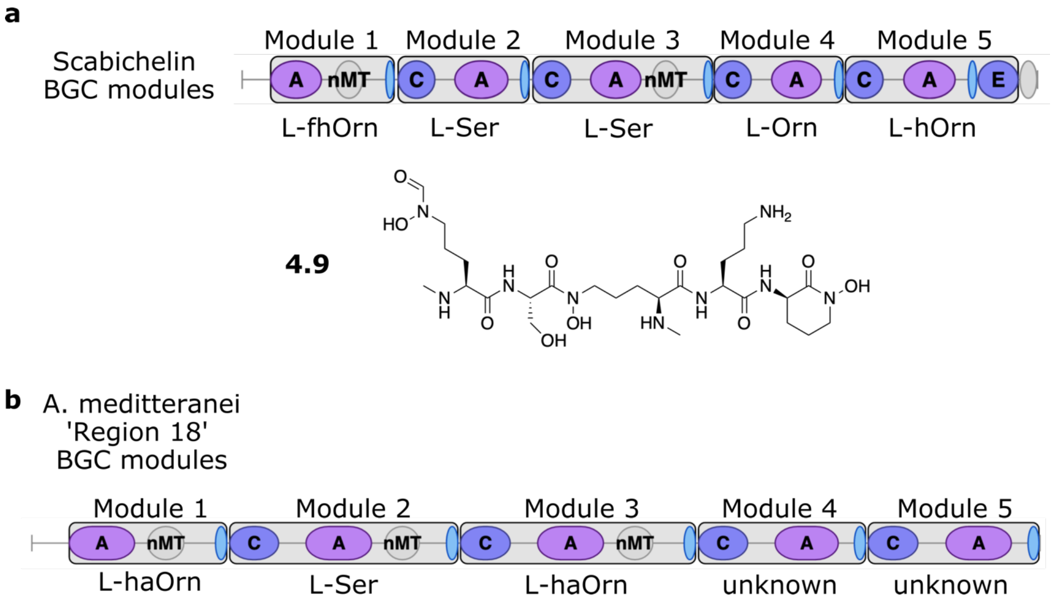


Figure 4.10 Comparison of Scabichelin BGC and Region 18 in A. meditteranei
(a) Modules of the scabilchelin BGC taken from the MIBiG database.[26] Substrates of each enzymatic module are indicated. Structure of scabichelin has been identified and associated with the BGC through genetic mutation studies.[27] (b) Modules and enzymatic substrate predictions from BGC region 18 in the *A. meditteranei* genome taken from the antiSMASH database.[25] Subatrate abbreviations: L-N5-formyl-N5-hydroxyornithine (L-fhOrn), L-N5-hydroxyornithine (L-hOrn), L-N5-acetyl-N5-hydroxyornithine (L-haOrn)

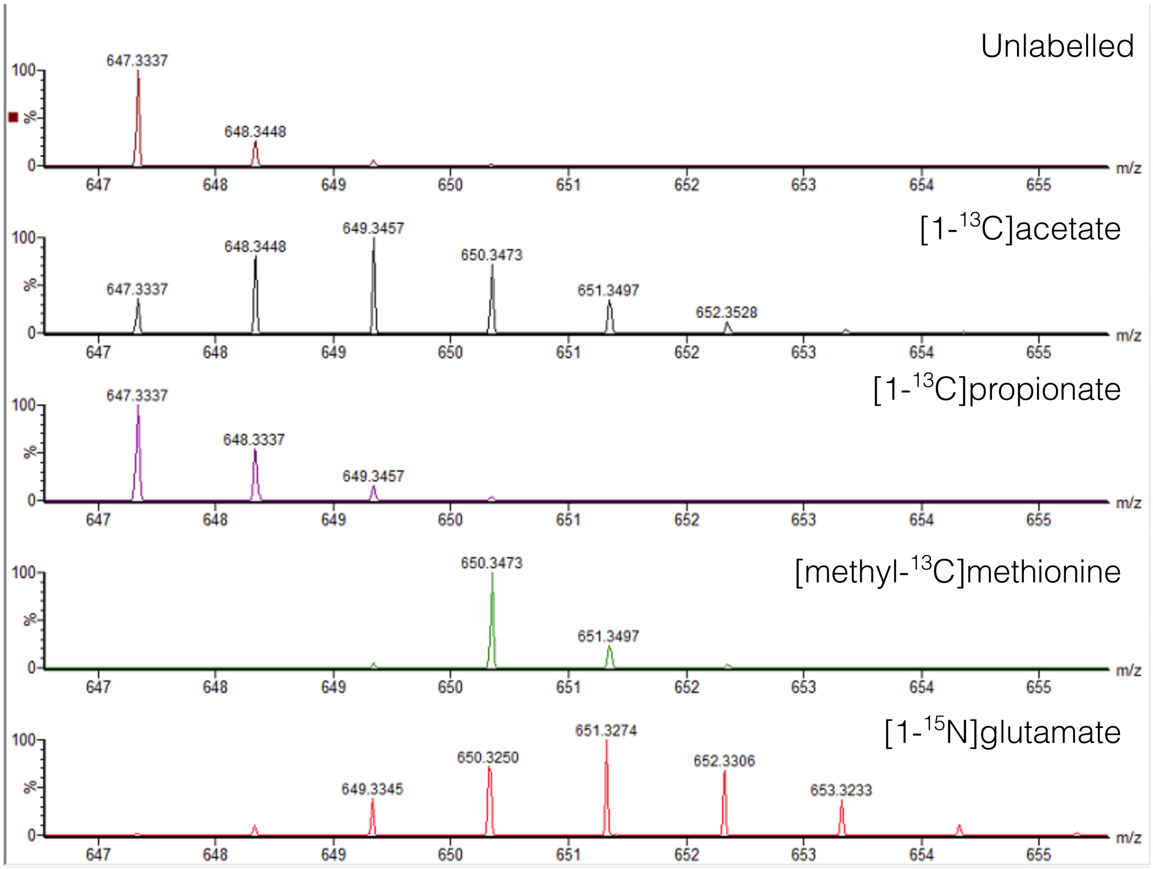


Figure 4.11 Unlabeled and SIL MS Spectra for *m/z* 647.3337
Unlabeled and SIL MS data for protonated adduct (m/z 647.3337) of the unknown siderophore from *A. meditteranei*. SIL incorporation in [1-^{13}C]acetate and [1-^{15}N]glutamate resemble the isotopologue distributions I observed for other siderophores such as erythrochelin. However, SIL incorporation for [methyl-^{13}C]methionine was distinctly efficient, showing very little isotopologue peak intensity for isotopologues below M_3.

4.4. Conclusion

These case studies in commercially available type strain bacteria demonstrated promising results for the application of IsoAnalyst to PKS and NRPS biosynthetic pathways. In *S. erythraea*, IsoAnalyst identified nearly all of the known biosynthetic intermediates of erythromycin A and accurately detected variable SIL incorporation across these compounds. Furthermore I identified the siderophore erythrochelin, and IsoAnalyst accurately detected SIL incorporation into the fragments ions, supporting the putative identity of these molecules. The ability to groups related MS features together and to interpret the positional SIL incorporation through fragmentation are currently manual processes which I have implemented here. However, both of these approaches to working with the IsoAnalyst output may be automated in the future, as adduct matching and fragmentation pattern association are already applied to unlabeled MS metabolomics data.[28,29] Overall, the SIL incorporation patterns detected by IsoAnalyst could assist greatly in associating related adducts and fragments as well as facilitate structural interpretations of MS metabolomics data. I also detected SIL incorporation in the polyketide antibiotic rifamycin S, and related in-source fragment adducts. Surprisingly, I was able to identify an unknown labeled compound in the metabolome data for *A. meditteranei*, although this organism has been exploited for natural product production for many years. This compound appears to be a hydroxymate siderophore which is related to scabeichelin, on the basis of the SIL incorporation pattern and BGC, further validating the utility of IsoAnalyst in indicating the biosynthetic origin of compounds prior to compound isolation and structure elucidation.

The examples shown here are limited as they represent only a narrow window of what is currently known about natural product biosynthesis. I selected these type strains which produce well-known polyketides and also identified siderophore NRPS compounds. While I have shown that the IsoAnalyst experimental design applied here works well for identifying these molecules, additional optimization would be needed to include a wider range of SIL tracers and biosynthetic pathways. Nonetheless, the generalizable SIL tracers used in combination with the IsoAnalyst data analysis platform have great potential for compound identification and biosynthetic grouping of MS features.

4.5. Methods

4.5.1. Parallel SIL Fermentation Experiments

Saccharopolyspora erythraea ATCC 11635 (NRRL 2338) and *Amycolatopsis meditteranei* ATCC 13685 were purchased from ATCC (USA). The parallel SIL fermentation protocols used in this chapter for *S. erythraea* and *A. meditteranei* were performed exactly as described in Chapter 3. The sample extraction, preparation, and UPLC-MS analysis were also performed as described in Chapter 3. Each experiment was analyzed using IsoAnalyst to generate a summary file of all ions with SIL incorporation detected in two or more conditions. The data were interrogated manually to identify the compounds described in this chapter.

4.5.2. Large Scale Fermentation and Extraction of 4.1 and 4.2

Four large-scale *S. erythraea* cultures were grown in 2.8 L Fernbach flasks containing 20.0 g of Amberlite XAD-16 adsorbent resin, a stainless steel spring, and 1 L of the same minimal media used in the SIL experiments. The large scale fermentation was done without the supplementation of SIL precursors. Cultures were shaken at 200 rpm for 6 days, at which time the cultures were filtered by vacuum filtration on Whatman glass microfiber filters. The cells and resin were collected and extracted with 250 mL of 1:1 methanol/dicholormethane. The organic extract was collected by vacuum filtration and dried by rotary evaporation. The crude organic extract was initially separated into seven fractions by a stepwise methanol/water elution (10, 20, 40, 60, 80, 100 vol/vol) and an additional ethyl acetate wash step on a RediSep Rf C18 cartridge (Teledyne Isco) using a Teledyne Isco CombiFlash Rf flash chromatography system.

4.5.3. Isolation and Characterization of 4.1 and 4.2

Purification of compounds **4.1** and **4.2** was performed on a Waters autopurification system with a SQ Detector 2 quadrupole MS detector. Both compounds were purified from the 60% methanol extract fraction. For all HPLC purification of **4.1** and **4.2** solvent A was water with 0.02% formic acid and solvent B was acetonitrile with 0.02% formic acid. The 60% methanol pre-fraction was separated by HPLC (Waters Atlantis T3 prep OBO column 5 μm, 19 x 250 mm) using an elution gradient of 45-83% B over 21 minutes, at a flow rate of 20 mL/min. Erythronolide B (**4.1**) was collected at 6.5 minutes by mass detection for the ion *m/z* 425.4 and **3** was collected at 9.5 minutes by mass detection for the ion *m/z* 529.4. 3-O-α-mycarosylerythronolide B (**4.2**) was further purified using an isocratic gradient of 42% B, with MEB eluting from the column at 11.8 minutes. 2.5 mg of **4.1** and 4.5 mg of **4.2** were isolated in total and analyzed by NMR for comparison to authentic standards (Appendix A).

References

1. Oliynyk, M. *et al.* Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat. Biotechnol.* **25**, 447–453 (2007).

2. Zhao, W. *et al.* Complete genome sequence of the rifamycin SV-producing *Amycolatopsis mediterranei* U32 revealed its genetic characteristics in phylogeny and metabolism. *Cell Res.* **20**, 1096–1108 (2010).

3. Staunton, J. & Weissman, K. J. Polyketide biosynthesis: A millennium review. *Nat. Prod. Rep.* **18**, 380–416 (2001).

4. Yoshizawa, Y., Li, Z., Vederas, J. C. & Reese, P. B. Intact Incorporation of Acetate-Derived Di- and Tetraketides during Biosynthesis of Dehydrocurvularin, a Macrolide Phytotoxin from *Alternaria cinerariae*. *J. Am. Chem. Soc.* **112**, 3212–3213 (1990).

5. Shindo, K., Sakakibara, M., Kawai, H. & Seto, H. Studies on cochleamycins, novel antitumor antibiotics III. Biosyntheses of Cochleamycins: Incorporation of ^{13}C-and ^{2}H-Labeled Compounds into Cochleamycins. *J. Antibiot. (Tokyo).* **49**, 249–252 (1996).

6. Wang, I. K., Vining, L. C., Walter, J. A. & Mcinnes, A. G. Use of carbon-13 in biosynthetic studies: Origin of the malonyl coenzyme a incorporated into tetracycline by *Streptomyces aureofaciens*. *J. Antibiot. (Tokyo).* **39**, 1281–1287 (1986).

7. Ferreira, E. L. F. *et al.* Structure and Biogenesis of Roussoellatide, a Dichlorinated Polyketide from the Marine-Derived Fungus *Roussoella* sp. DLM33. *Org. Lett.* **17**, 5152–5155 (2015).

8. Paseshnichenko, V. A. A new alternative non-mevalonate pathway for isoprenoid biosynthesis in eubacteria and plants. *Biochemistry. (Mosc).* **63**, 139–48 (1998).

9. Kulis-Horn, R. K., Persicke, M. & Kalinowski, J. Histidine biosynthesis, its regulation and biotechnological application in C orynebacterium glutamicum. *Microb. Biotechnol.* **7**, 5–25 (2014).

10. Reeves, A. R. *et al.* Effects of methylmalonyl-CoA mutase gene knockouts on erythromycin production in carbohydrate-based and oil-based fermentations of Saccharopolyspora erythraea. *J. Ind. Microbiol. Biotechnol.* **33**, 600–609 (2006).

11. Bode, H. B., Bethe, B., Höfs, R. & Zeeck, A. Big Effects from Small Changes: Possible Ways to Explore Nature’s Chemical Diversity. *ChemBioChem* **3**, 619 (2002).

12. Khosla, C., Tang, Y., Chen, A. Y., Schnarr, N. A. & Cane, D. E. Structure and Mechanism of the 6-Deoxyerythronolide B Synthase. *Annu. Rev. Biochem.* **76**, 195–221 (2007).

13. Cane, D. E., Hasler, H., Taylor, P. B. & Liang, T. C. Macrolide biosynthesis-II. Origin of the carbon skeleton and oxygen atoms of the erythromycins. *Tetrahedron* **39**, 3449–3455 (1983).

14. 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).

15. 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).

16. Zhang, H., Wang, Y., Wu, J., Skalina, K. & Pfeifer, B. A. Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chemistry and Biology*. **17**, 1232–1240 (2010).

17. Weissman, K. J. & Leadlay, P. F. Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.* **3**, 925–936 (2005).

18. Robbel, L., Knappe, T. A., Linne, U., Xie, X. & Marahiel, M. A. Erythrochelin - a hydroxamate-type siderophore predicted from the genome of *Saccharopolyspora erythraea*. *FEBS J.* **277**, 663–676 (2009).

19. Lazos, O. *et al.* Biosynthesis of the Putative Siderophore Erythrochelin Requires Unprecedented Crosstalk between Separate Nonribosomal Peptide Gene Clusters. *Chem.& Biol.* **17**, 160–173 (2010).

20. Sensi, P., Margalith P. & Timbal, M. T. Rifomycin, a new antibiotic; preliminary report. *Farmaco. Sci.* **14**, 146–147 (1959).

21. Verma, M. *et al.* Whole genome sequence of the rifamycin B-producing strain *Amycolatopsis mediterranei* S699. *J. Bacteriol.* **193**, 5562–5563 (2011).

22. Floss, H. G. & Yu, T. W. Rifamycin - Mode of action, resistance, and biosynthesis. *Chem. Rev.* **105**, 621–632 (2005).

23. Xu, J., Wan, E., Kim, C. J., Floss, H. G. & Mahmud, T. Identification of tailoring genes involved in the modification of the polyketide backbone of rifamycin B by *Amycolatopsis mediterranei* S699. *Microbiology* **151**, 2515–2528 (2005).

24. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

25. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45**, D555–D559 (2017).

26. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

27. Kodani, S. *et al.* Structure and biosynthesis of scabichelin, a novel tris-hydroxamate siderophore produced by the plant pathogen *Streptomyces scabies* 87.22. *Org. Biomol. Chem.* **11**, 4686–4694 (2013).

28. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

29. Uppal, K., Walker, D. I. & Jones, D. P. xMSannotator: An R package for network-based annotation of high-resolution metabolomics data. *Anal. Chem.* **89**, 1063–1067 (2017).

Chapter 5.

IsoAnalyst Coupled to Whole Genome Analysis as a Tool for Natural Product Discovery

5.1. Introduction

The ultimate goal of the IsoAnalyst platform is to facilitate natural product discovery through the characterization of complex chemical phenotypes produced by BGCs. In Chapter 4, I showed how IsoAnalyst can be used to accurately interpret the biosynthetic pathway of known compounds and how the various analogues and biosynthetic precursors of erythromycin A can be associated on the basis of related SIL incorporation patterns. In this chapter, I apply this workflow to an environmental strain to demonstrate how IsoAnalyst can reduce a complete untargeted MS metabolomics dataset to a manageable number of labeled MS features. Whole genome sequences can be analyzed for BGC presence by open source online tools and the growing availability of genome data further allows for large-scale comparisons between BGCs across thousands of strains.[1,2] I have developed a generalizable method for interpreting how BGC products will be labeled by specific SIL tracers on the basis of the enzymatic substrate predictions generated in antiSMASH[1] and investigation of literature about related BGCs from MIBiG.[2] I collaborated with the developers of these platforms to generate the manually curated substrate predictions, but I created the process for interpreting the data and associating the substrate prediction with the SIL incorporation profiles from IsoAnalyst. Using this complete workflow, I show how IsoAnalyst can quickly associate both known metabolites and unknown analytes to elucidate the complex chemical phenotypes of known BGCs.

5.2. Whole Genome BGC Analysis of *Micromonospora* sp.

In order to test the capacity of IsoAnalyst to identify natural products in a sequenced environmental strain, I applied the complete workflow to *Micromonospora* sp. RL09-050-HVF-A. This *Micromonospora* sp. was isolated from marine sediments in Point Lobos, California and produced the macrolactam, lobosamide A, which demonstrated submicromolar antitryposomal activity against *Trypanosoma brucei* and the related analogues lobosamide B and C.[3] In order to fully solve the configurational analysis of the hydroxyl groups present in the lobosamides, our laboratory had this *Micromonospora* sp. sequenced at the Institute of Genome Science sequencing facility (IGS, University of Maryland, Baltimore) using a Pacific Bioscience Sequencing machine and a 10KB insert library. The *Micromonospora sp.* RL09-050-HVF-A genome was uploaded to NCBI under the accession number JAGKQP000000000 and the BioProject ID PRJNA718589.

5.2.1. Curated antiSMASH Output

In order to apply the IsoAnalyst workflow to the full metabolome of *Micromonospora* sp., BGC mining was performed on the full genome using antiSMASH version 5.2.0-8ecc354. Our collaborators used the antiSMASH to perform sequence analysis of the BGCs, including automated substrate predictions. They further developed a protocol using MIBiG to manually interpret these substrate predictions using both sequence analysis and literature review. The antiSMASH output offers substrate predictions for a broad cross-section of biosynthetic classes using a suite of prediction algorithms including SANDPUMA,[4] NRPSPredictor2,[5] and RODEO.[6] These substrate predictions have a fair degree of accuracy, but manual interpretation is nearly always required, as not all enzymes yield equally likely substrate predictions. For example, type I and II PKS and NRPS enzymes have more predictable substrates while special algorithms have been developed for other classes such as type III PKS.[7] It is important to manually inspect every BGC substrate prediction in antiSMASH, as these predictions are potentially less accurate for more novel BGCs, and some BGCs have no substrate predictions at all. The antiSMASH output also indicates the most closely related BGC sequences that are available in the MIBiG database.[2] Using the antiSMASH substrate prediction for the detected enzymatic domains, in combination with literature information from related BGCs in MIBiG, our collaborators developed a protocol for curating enzymatic substrate prediction information with the highest possible accuracy. This protocol is described in more detail in the methods section at the end of this chapter. The curated antiSMASH output for *Micromonospora* sp. is shown in Tables 5.1 and 5.2. Table 5.1 contains metadata and comments about each BGC. The BGC type is determined automatically in antiSMASH, and the comments are based on manual interpretations of related BGCs by comparison with the MIBiG database. The columns of

Table 5.2 contain predicted substrate counts for a single gene cluster from a list of commonly encountered biosynthetic substrates that are expected to be labeled by one or more SIL precursors.

Table 5.1 BGCs Detected in *Micromonospora* sp. Genome

Cluster	Type	Product #	Comments
1a	NRPS	1	
1b	RiPP	1	
1b	RiPP	2	
2	PKS	1	galbonolides
3	PKS	1	
4	PKS-NRPS	1	
5	RiPP	1	
6	NRPS	1+	desferrioaxamines
7 & 21	terpene	1	sioxanthin - cluster 7 & 21 work together
8	lantipeptide	2	very similar to SapB
9	type II PKS	1	core PKS similar to frankiamycin type II PKS
10	type I PKS	1	similar to maduropeptin
11a	PKS-NRPS	1	
11b	lasso peptide	1	
12a	NRPS	1	
12b	PKS	1	
13	NRPS	1	indigoidine
14	PKS	1	
15	phenazine	1	
16	NRPS	1	similar to Fruilimicin
17	RiPP-like	1	no relevant predictions to be made
18	terpene	1	
19	NRPS-like	1	
20	type III PKS	2	alkyl-O-dihydrogeranyl-methoxyhydroquinones
21			see 7
22	PKS-like	1	paulomycin without paulomycose
23	indole	1	prenylated indole
24	terpene	1	
25	NRPS	1	
26	NRPS	1	one AA is probably aromatic, possibly Phe
27	NAGGN	1	N-acetylglutaminylglutamine amide
28	lantipeptide	1	
29	RiPP-like	1	no relevant predictions to be made
30a	PKS-NRPS	1	similar to Tirandamycin
30b	terpene	2	
30c	PKS	3	lobosamide

Table 5.2 **Curated AntiSMASH Output for *Micromonospora* sp.**

Cluster	1a	1b	1b	2	3	4	5	6	7 & 21	8	8	9	10	11a	11b	12a	12b	13	14	15	16	17	18	19	20	20	22	23	24	25	26	27	28	29	30a	30b	30c
Product #	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1+**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**3**
Glu																																					1
Gln											1							2														2					
Arg		2	2							1																											
Pro		3	3												1						2																
Orn																																					
Asp										1	1				3						2												5				
Met			2																																		
Thr		4	3							3	3				1																		2				
Ile		1																																			
Asn						1				1	1				1	1																					
Lys								3																									1				
Ala		8	9							1	1				1																		2				
Leu		3	1							4	4				3																		2				
Val		7	5			1				1	1				1						1												1				
Phe		1	1																																		
Tyr																					1																
Trp																												1									
Ser		1				1				5	5				1																		1				
Gly		8	5							4	4				7	1					6												2		1		
Cys										2	2														1								2				
His			1																																		
Unk. Amino Acid	1					1	1+																							1	2						
Acetyl-CoA								1					2																			1					
Malonyl-CoA					3+	1						12	12+	1+			1		8						3	3									5		8
Methylmalonyl-CoA				4	3+												1		1																5		3
Methoxymalonate				4																																	
Hydroxymalonate																																					
Methyl group					1	1														1					1	1										1	
IPP									8														3						1+							2	
DMAPP																												1									
GPP																									2												
FPP																										2											
Succinyl-CoA								2																													
Unknown fatty acid												1	1			1					1		1	1												1	
Amino saccharide													1				1			1	1																
Amino group																																					

5.2.2. BGC Labeling Prediction

In Chapter 4 I introduced the Substrate Labeling Table (Table 5.3) to show how the SIL tracer incorporation detected by IsoAnalyst can be aligned with the substrates used in the biosynthesis of compounds in *S. erythraea* and *A. medietteranei*. This table can further be used to predict SIL tracer incorporation into the products of the BGCs shown in Table 5.2. To predict the labeling patterns for each BGC, I integrated data from the curated antiSMASH output (Table 5.2) and the Substrate Labeling Table (Table 5.3) to create a table of predicted precursor labeling for all BGCs in the genome (Table 5.4).

Not all of the BGCs detected in the genome of *Micromonospora* sp. are equally amenable to SIL incorporation in our experiment. Some well-characterized BGC classes, such as terpenes, do not have sufficient SIL incorporation by the SIL tracers used in this experiment. On the other hand, some natural products may be labeled by the SIL precursors used, but cannot be easily connected to the BGC due to the lack of biochemical knowledge about the BGC. I used two criteria to determine the likelihood of identifying the product of a BGC with the SIL precursors used in this study. Firstly, the predicted product of a BGC should be labeled in two or more conditions, as it is not possible to differentiate products of related BGCs on the basis of labeling in a single condition. Additionally, the product must be labeled three or more times in at least one of the conditions, as minor SIL incorporation is likely to occur in off-target compounds involved in primary metabolism. In Table 5.4 BGCs that meet both criteria are highlighted in dark gray, while BGCs that meet only one of the criteria are highlighted in light gray.

The BGC substrate analysis indicates that NRPS and PKS BGCs are most well suited to discovery using this approach. There are several reasons for this bias. BGC identification and annotation tends to be most reliable for these two well-studied biosynthetic classes. Future improvements in BGC informatics, particularly in the area of substrate prediction for less well studied BGC classes, will increase the coverage of the IsoAnalyst method. Likewise, SIL precursor selection has a significant impact on BGC class coverage. Additional SIL precursors could be included to provide labeling for specific biosynthetic classes. For example, terpenes will only have robust SIL incorporation with [1-^{13}C] acetate if the organism possesses the mevalonate pathway. Inclusion of ^{13}C labeled IPP as an additional SIL precursor would significantly improve coverage of compounds in this class. Similarly, chorismate-derived precursors are also common and ^{13}C labeled chorismate or shikimate could help to prioritize specific BGC products containing these substrates. In applying this tool to environmental organisms, substrate analysis of the BGCs present ideally should be done in advance to guide the precursor selection process and optimize coverage of that particular organism’s metabolism.

Table 5.3 Substrate Labelling Table

Substrate	A	P	M	G
Glu	2			1
Gln	2			2
Arg	2			4
Pro	2			1
Orn	2			2
Asp	1			1
Met	1		1	1
Thr	1			1
Ile	1			1
Asn	1			2
Lys	1			2
Ala				1
Leu				1
Val				1
Phe				1
Tyr				1
Trp				2
Ser				1
Gly				1
Cys				1
His				3
Unknown amino acid				1+
Acetyl or Malonyl-CoA	1			
Propionyl-CoA		1		
Methylmalonyl-CoA	1	1		
Methoxymalonate			1	
Hydroxymalonate				
Methyl			1	
IPP[a]	2			
DMAPP[a]	2			
GPP[a]	4			
FPP[a]	6			
Succinyl-CoA	1	1		
Amino-saccharide				1
Amino group				1

Table 5.4 BGC Labelling Prediction for *Micromonospora* sp.

Cluster	Type	A	P	M	G
8	lantipeptide	7			27
28	lantipeptide	8			19
11b	lasso peptide	7			20
1b	RiPP	16			44
5	RiPP				1+
17	RiPP-like				
29	RiPP-like				
6	**NRPS**	**6**	**2**		**6**
12a	NRPS	1+			3
13	NRPS	4			4
16	NRPS	7+			13
1a	NRPS				1+
25	NRPS				1+
26	NRPS				2+
19	NRPS-like	1+			
4	PKS-NRPS	2		1	4
30a	PKS-NRPS	10	5		1
11a	PKS-NRPS	1+			
10	type I PKS	14+			1
9	type II PKS	12+			
20	type III PKS	15		1	1
2	PKS	4	4	4	
3	PKS	6+	3+	1	
12b	PKS	2	1		1
14	PKS	9	1		
30c	**PKS**	**11**	**3**		**1**
22	PKS-like				
7 & 21	terpene	16			
18	terpene	6+			
24	terpene	2+			
30b	terpene	2+			
23	indole	3			1
15	phenazine			1	3
27	NAGGN	5			5

5.2.3. *Micromonospora* sp. Complete IsoAnalyst Results

I performed parallel SIL culture, UPLC-MS analysis and data processing with the IsoAnalyst pipeline on *Micromonospora* sp. The summary file from IsoAnalyst initially contained 246 *m/z* features that were labeled in two or more SIL conditions. I then manually interrogated this file to filter features based on chromatographic peak shape and signal intensity. Features were eliminated if either of these factors interfered with the accuracy of the SIL detection. This manual filtering step resulted in 100 features that could be grouped into two major compound classes based on their isotope labeling patterns (Figure 5.1). Forty-nine features corresponded to the desferrioxamine family of hydroxamate siderophores and 51 features corresponded to the lobosamide macrolactam polyketides previously discovered in our lab. Both classes could be confidently linked to their BGCs on the basis of their SIL incorporation patterns and showed remarkable diversity in labeled products despite originating from only two distinct BGCs.

5.3. Desferrioxamines

The first group of similarly labeled features possessed significant labeling by [1-^{13}C]acetate (2-6 positions), [1-^{13}C]propionate (1-2 positions), and [1-^{15}N]glutamate (5-6 positions), but not labeling by [methyl-13C]methionine (Figures 5.1, 5.2). The only BGC in Table 5.4 containing two or more [1-^{13}C]propionate labels and six or more [1-^{15}N]glutamate labels is BGC 6. This group of labeled MS features was consistently associated with iron-adducts and related fragment ions, suggesting a family of related siderophores. BGC 6 was predicted to produce a siderophore by antiSMASH, and had 100% match with the BGC responsible for producing desferrioxamine B (**5.1**) and E (**5.2**).

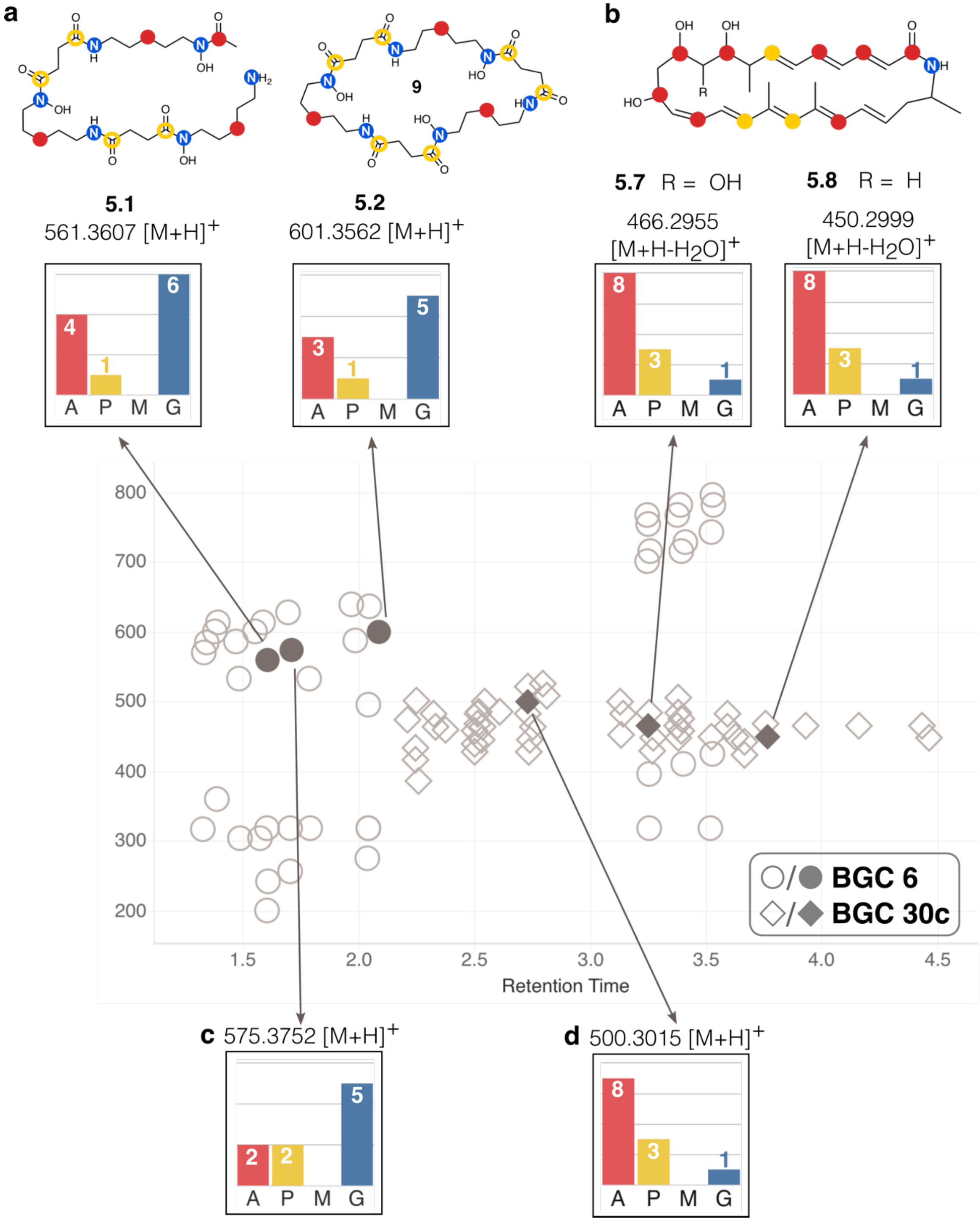


Figure 5.1 Overview of IsoAnalyst Results from *Micromonospora* sp.
Diagram showing all MS features with detected SIL incorporation in two or more conditions, following manual filtering and grouping according to SIL incorporation profiles. (a) known desferrioxamines, (b) known lobosamides, (c) new desferrioxamine, (d) new lobosamide. Diamonds represent features with SIL incorporation patterns related to the lobosamides, and circles represent features with SIL incorporation patterns related to the desferrioxamines.

5.3.1. Known Desferrioxamine Siderophores

Desferrioxamines are hydroxymate siderophores that are made up of three subunits of N-hydroxycadaverine, which are acylated by the acyl-dependent acyl transferase, DesC, with either acetyl-CoA or succinyl-CoA.[8] Des D catalyzes the oligomerization of either two units of N-hydroxy-N-succinylcadaverine (HSC) and one N-hydroxy-N-acetylcadaverine (HAC) to yield the linear **5.1**, or three units of HSC to yield the cyclized compound **5.2**.[9] N-Hydroxycadaverine is derived from lysine, which has two nitrogen atoms than can be labeled by [1-^{15}N]glutamate and one carbon atom that can be labeled by [1-^{13}C]acetate (Table 5.3). Succinyl-CoA can be labeled in one position by [1-^{13}C]acetate through the TCA cycle, and one position by [1-^{13}C]propionate through the transformation of methylmalonyl-CoA to succinyl-CoA (Table 5.3). In both cases, the position of the SIL incorporation is ambiguous due to the symmetry of succinate (Figure 5.1). These positions are represented by open circles in Figure 5.1, but only one of the two indicated positions may be labeled in a given molecule. In all, **5.1** is expected to be labeled in six positions by [1-^{13}C]acetate, two postions by [1-^{13}C]propionate, and six positions by [1-^{15}N]glutamate (Figure 5.1). Enrichment of SIL in compound **5.1** was detected for all six nitrogen atoms derived from lysine in the [1-^{15}N]glutamate condition, four of the six expected subunits in the [1-^{13}C]acetate condition, and one of the two expected subunits in the [1-^{13}C]propionate condition (Figure 5.1). The cyclized **5.2** has one additional position expected to be labeled by [1-^{13}C]propionate due to an additional incorporation of succinyl-CoA, however, **5.2** demonstrated slightly less SIL incorporation in both [1-^{13}C]acetate and [1-^{15}N]glutamate compared to **5.1** (Figure 5.1). These differences are likely due to a lower signal intensity for the isotopologue peaks of **5.2** rather than biosynthetic differences between **5.1** and **5.2**.

In addition to **5.1** and **5.2**, four desferrioxamine derivatives that have been reported in the literature were identified based on MS spectra, database matching, and the manual interpretation of SIL patterns. The identities of all compounds were confirmed by co-injection with authentic standards (Figure 5.2, Appendix A Figures A9-14). Desferrioxamine D2 (**5.3**) contains one ornithine-derived subunit instead of the lysine derived N-hydroxycadaverine, due to the substrate promiscuity of the lysine decrarboxylase DesA.[10] Although ornithine is expected to have one more position labeled by [1-^{13}C]acetate than lysine (Table 5.3), we detected less [1-^{13}C]incorporation in **5.3** compared to **5.2** (Figure 5.2). This discrepancy can again be attributed to a difference in signal intensity, indicating that the variable intensities between biosynthetically related compounds is often a limiting factor in the exact interpretation of SIL incorporation in this study.

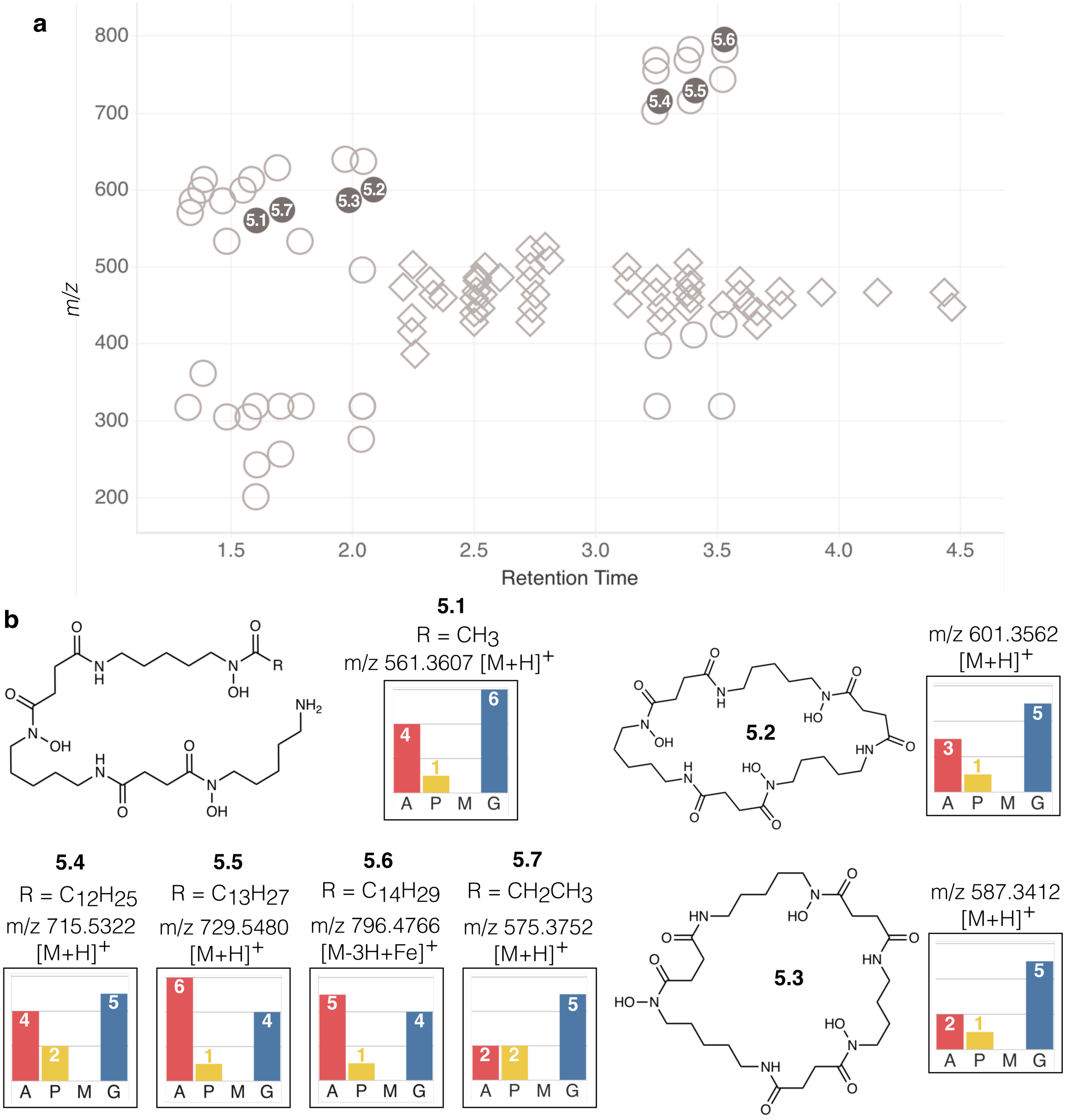
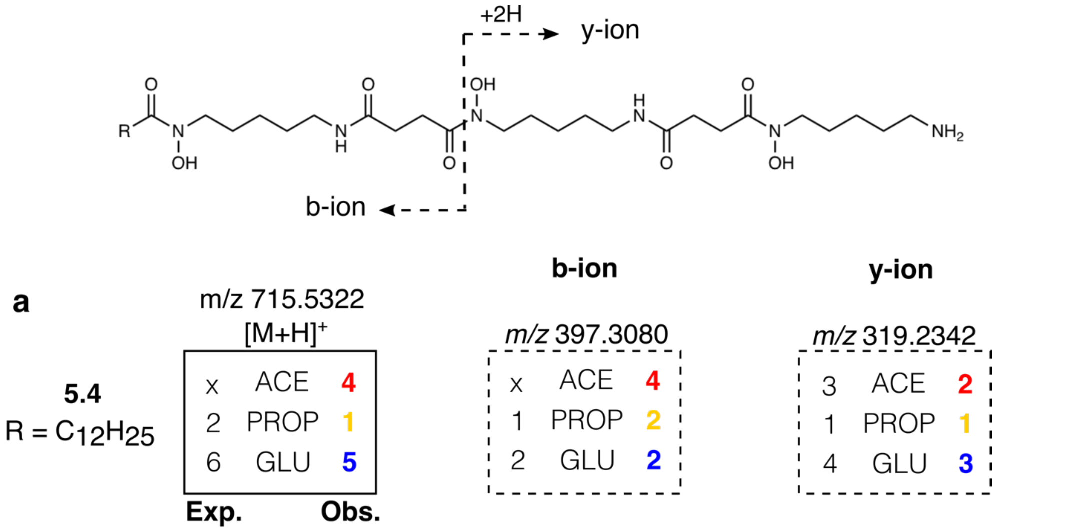
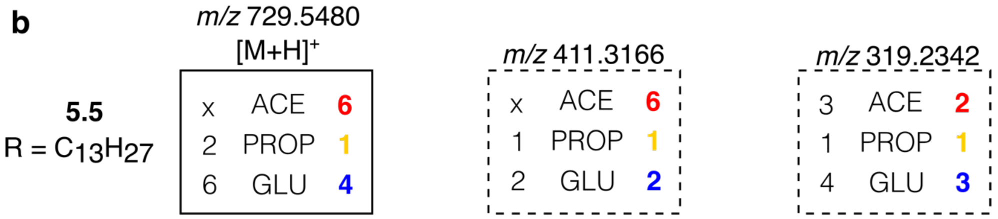


Figure 5.2 SIL Incorporation in Desferrioxamines
(a) Diagram showing all MS features with detected SIL incorporation in two or more conditions, following manual filtering and grouping according to SIL incorporation profiles. (b) SIL incorporation detected by IsoAnalyst in desferrioxamine B and related analogs. SIL incorporation shown for MS features indicated as filled circles.

Most of the structural variability in the desferrioxamine family is derived from the substrate flexibility of DesC which has been shown to also accept larger acyl substrates besides acetyl-CoA and succinyl-CoA.[8] N-Acylated desferrioxamine derivatives have been described in a few studies.[11–14] A paper previously published in our lab describes microferrioxamines A (**5.4**), B (**5.5**), and C (**5.6**), which are linear aliphatic siderophores containing acyl chains of varying lengths[15] (Figures 5.2, 5.3). Due to the varying lengths of the acyl tail, the number of maximum theoretical positions labeled by [1-^{13}C]acetate cannot be accurately predicted (Figure 5.3). Despite this, the microferrioxamines have





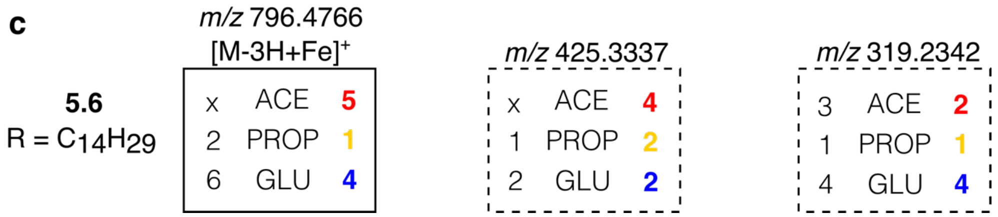


Figure 5.3 SIL Incorporation in Microferrioxamines and Fragment Ions
Expected (left) and observed (right) labeling for microferrioxamines A (a), B (b), C (c), and their corresponding b-ion and y-ion fragments. The b-ion contains the acyl tail group and retains more [1-^{13}C]acetate incorporation which varies across the different analogues. The SIL incorporation across the y-ion was consistent in all SIL tracers, except [1-^{15}N]glutamate (c), which is likely due to differences in ion intensity. The [1-^{13}C]acetate incorporation cannot be fully predicted for these molecules, as the length of the acyl chain varies and cannot be directly predicted from the BGC, and these positions are represented by ‘x’ in the expected labeling by [1-^{13}C]acetate for each compound.

clearly related SIL incorporation patterns in the [M+H]+ and $[M\text{-}3H+Fe]^+$ ions (Figure 5.2) as well as the detected fragment ions (Figure 5.3). The same y-ion (*m/z* 319.2342) was detected for all three microferrioxamines, and the corresponding b-ions have higher [1-^{13}C]acetate incorporation, indicating the relative position of the acyl tail. While the substrate variation among these derivatives does produce minor differences in detected SIL patterns, additional data such as mass fragmentation patterns and the presence of iron adducts assist in categorizing these compounds as a biosynthetically and structurally related family.

5.3.2. New Desferrioxamine Siderophore

Review of the remaining members of this compound group identified one molecule with a diagnostic ferrioxamine labeling pattern and iron-adduct that had no match in existing natural products databases (Figure 5.1c). This new molecule possessed an $[M+H]^+$ adduct at *m/z* 575.3752 corresponding to the molecular formula $C_{26}H_{50}N_6O_8$ suggesting an analogue of **5.1** containing an additional CH_2 subunit. This putative desferrioxamine derivative (**5.7**) showed related labeling to **5.1**, with a nearly identical labeling pattern in the [1-^{15}N]glutamate condition, but decreased labeling by [1-^{13}C]acetate, and increased labeling by [1-^{13}C]propionate (Figure 5.4a). This suggested that the additional CH_2 subunit present in **5.7** derived from substitution of one [1-^{13}C]acetate subunit with an additional [1-^{13}C]propionate moiety. To further understand the structural differences between these compounds, we examined the MS/MS fragmentation data. The y-ion *m/z* 319.2343 was detected as an MS feature for both **5.1** and **5.7** and possessed identical SIL patterns in both cases (Figure 5.4b). The corresponding b-ion fragments *m/z* 243.1342 and *m/z* 257.1505 were detected for **5.1** and **5.7** respectively. These b-ions have identical labeling in the [1-^{15}N]glutamate condition, but relative to each other, *m/z* 243.1342 is enriched in [1-^{13}C]acetate, while *m/z* 257.1505 is enriched in [1-^{13}C]propionate (Figure 5.4b) The combination of the mass difference of 14.0 Da, corresponding to CH_2, and the enriched [1-^{13}C] propionate labeling in comparison to the analogous fragment in **5.1**, indicates that propionyl-CoA replaces acetyl-CoA in the acylation of the terminal N-hydroxycadaverine residue, confirming the identification of this molecule as a new member of the hydroxamate siderophore family of natural products. Together, these results show that this strain can produce a large suite of desferrioxamine derivatives, and that these ions can be easily grouped together

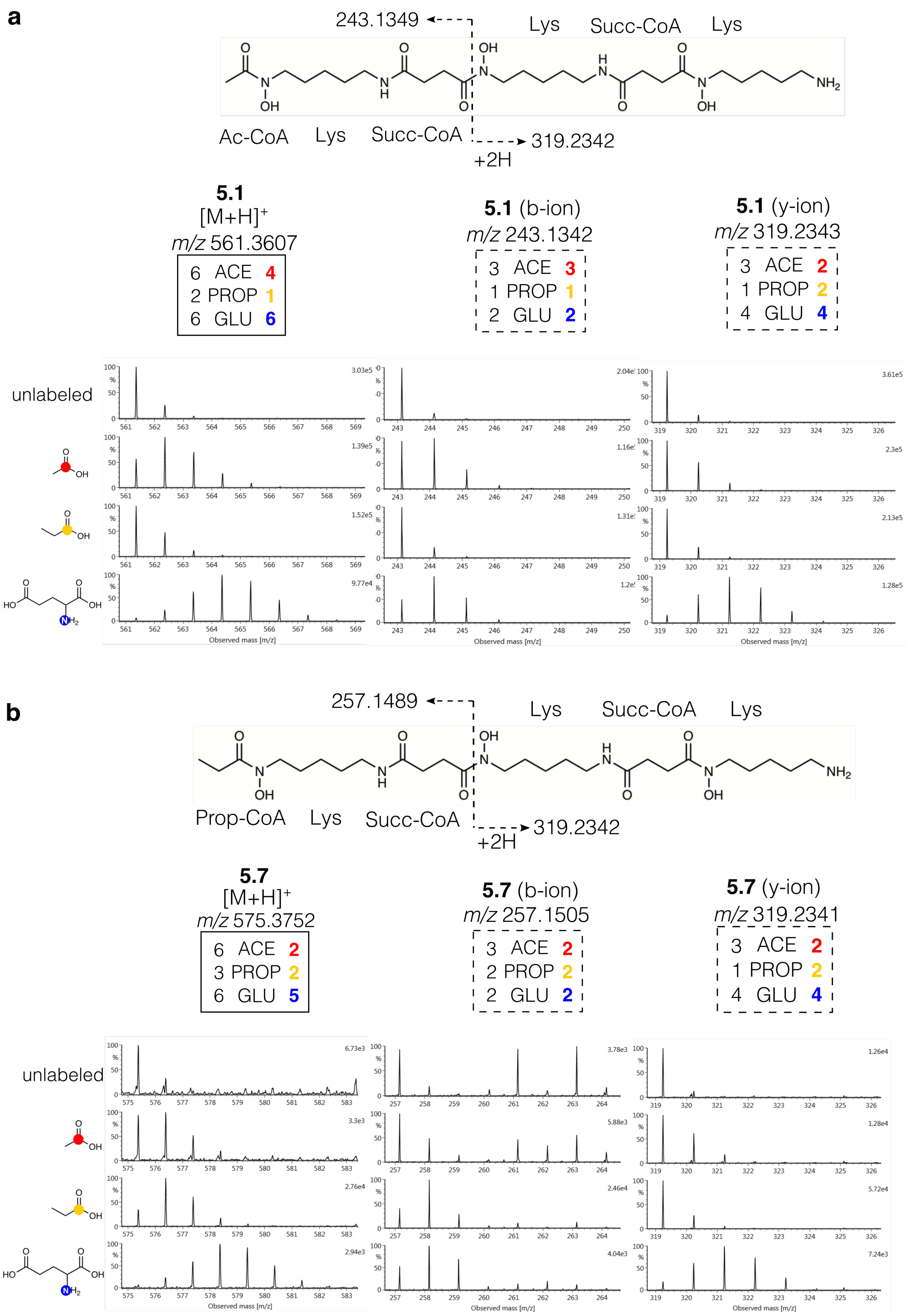


Figure 5.4 SIL Incorporation in Desferroxamine B and Unknown Desferrioxamine

(a) Structure of desferrioxamine B (**5.1**), and the mass spectra of the [M+H]$^+$ ion, the b-ion and y-ion fragments of **5.1**. (b) Putative structure of the new desferrioxamine (**5.7**) and the mass spectra of the [M+H]$^+$ ion, the b-ion and y-ion fragments of **5.7**. Expected (left) and observed (right) SIL incorporation for each ion are shown in boxes.

by relating SIL patterns detected using the IsoAnalyst platform. In total, I identified seven compounds produced by the desferrioxamine BGC, including the novel desferrioxamine analogue **5.7**.

5.4. Lobosamides

5.4.1. Lobosamide Biosynthetic Gene Cluster

The second major class of labeled molecules in this dataset also had closely related labeling patterns in the [1-^{13}C]acetate, [1-^{13}C]propionate, and [1-^{15}N]glutamate conditions (Figure 5.1). Comparison of these labeling patterns (8 x [1-^{13}C]acetate, 3 x [1-^{13}C]propionate, 1 x [1-^{15}N]glutamate) to the annotated BGC list identified two BGCs with an appropriate combination of biosynthetic modules; clusters 30a and 30c (Table 5.4). Both BGCs include a single [1-^{15}N]glutamate incorporation, with BGCs 30a and 30c incorporating five and three [1-^{13}C]propionate units respectively. Based on these data alone, the molecules could derive from either BGC 30a or 30c because incomplete SIL incorporation or incomplete detection of isotope labeling could limit our ability to accurately define the number of SIL incorporations in each molecule. From the [1-^{13}C]acetate predictions in Table 5.4 we see that both BGCs have a higher theoretical maximum for acetate incorporation than the eight [1-^{13}C]acetate units detected in this compound class. BGC 30a includes a maximum of ten [1-^{13}C]acetate incorporations while BGC 30c includes a maximum of eleven (Table 5.4). However, these predictions include both direct and indirect incorporation pathways. Specifically, the methylmalonyl-CoA building blocks that are predicted to be labeled by [1-^{13}C]propionate in these BGCs are also expected to be labeled indirectly by [1-^{13}C]acetate. Considering only the direct incorporation of [1-^{13}C]acetate through malonyl-CoA and not methylmalonyl-CoA, BGCs 30a and 30c incorporate five and eight labeled precursors respectively (Table 5.4). The strong and consistent incorporation of a minimum of eight acetate units in the molecules of this class thus prioritizes BGC 30c as the cluster responsible for the production of this compound family. This prediction is further strengthened by the observation that the predicted direct incorporation for BGC 30c (8 x [1-^{13}C]acetate, 3 x [1-^{13}C]propionate, 1 x [1-^{15}N]glutamate) exactly matches the observed SIL patterns in the compound family (Figure 5.1).

5.4.2. Diversity of Detected Lobosamides

BGC 30c has previously been shown to produce the lobosamide family of natural products.[3] All the major compounds in this group displayed diagnostic *m/z* features and distinctive $[M+H-H_2O]^+$ in-source fragments consistent with the lobosamide family. The presence of lobosamide C (**5.8**) was confirmed by HPLC purification and NMR comparison, while lobosamide A (**5.9**) was identified in the extracts by UPLC-MS comparison to an authentic standard (Appendix A Figures A15-A18). Known lobosamide structures only accounted for 6 of the 51 MS features with related SIL patterns. Most of the remaining features were produced in low titer and had mass differences related to the degree of unsaturation and varying oxidations of the known lobosamide scaffold, suggestive of a suite of lobosamide analogues. These features were detected across more than half of the chromatographic separation time, indicating a wide range of polarities despite having the same structural units (Figure 5.1). I also noticed that despite the large number of features with lobosamide-like SIL incorporation, many of these features shared the same *m/z* value but were eluted at different retention times, further suggesting that the lobosamide BGC produces a variety of isomers. Aside from substrate flexibility, these configurational isomers add to the complexity of the chemical phenotype of a BGC. Although not all isomers are produced in isolable quantities, the IsoAnalyst approach allows for the easy detection and association of these related features for further investigation and optimization for compound isolation.

5.4.3. Isolation and Characterization of Lobosamide D

I isolated a representative molecule from this class with an $[M+H]^+$ *m/z* feature at 500.3012 and a calculated molecular formula of $C_{29}H_{41}NO_6$. This molecule did have a $[M+H-H_2O]^+$ adduct, however it demonstrated a slightly different distribution of ion adducts than the known lobosamide compounds (Figure 5.5). These distinctive in-source dehydration fragments also appeared in the MS/MS spectra, however, no structurally informative MS/MS fragments were detected for any of the lobosamides. The ^1H NMR spectrum of the isolated compound bore low similarity to the ^1H spectra for the known lobosamides (Figure 5.6a). While the initial chemical characterization data provided ambiguous evidence for the structural relatedness of my isolated compound with other lobosamides, it was clear from the SIL incorporation data that the new compound was derived from the same biosynthetic pathway. Using a combination of 1D and 2D NMR experiments (Appendix B) I determined the planar structure of this new metabolite as a 5-5-6 fused ring system variant of **5.9** which we named lobosamide D (**5.10**) (Figure 5.7a). The full absolute configuration of this new molecule was determined using a combination of extensive 1D-selective and 2D ROESY experiments, coupled with configurational assignments based on sequence data for selected keto-reductase (KR) domains[3] (Figure 5.7b). A detailed description of the NMR experiments used to solve this structure is provided in the methods section at the end of this chapter.

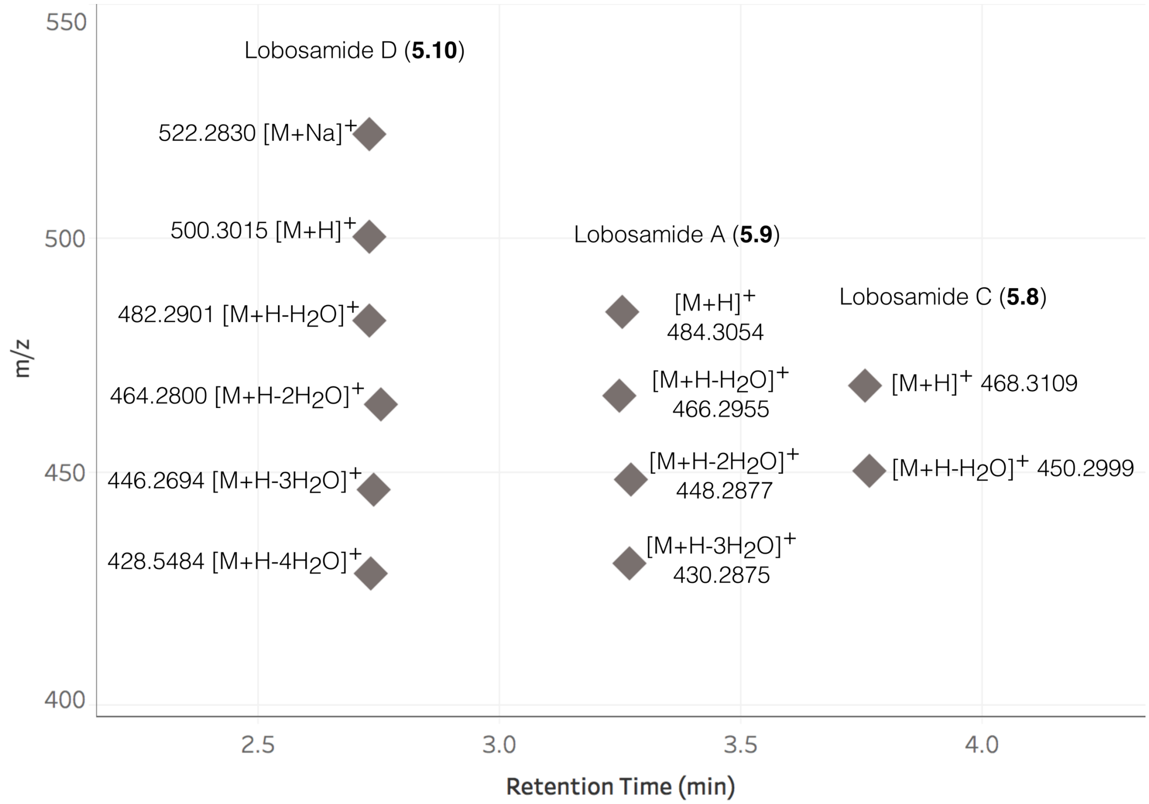


Figure 5.5 MS Adduct and In-Source Fragment Ions of Lobosamides
Adducts and in-source fragment ions corresponding to **5.8**, **5.9**, and **5.10**, which had SIL incorporation detected by IsoAnalyst. The dehydration and multi-dehydration ions are present for nearly all of the detected lobosamides, however the increased fragments and adduct ions of **5.10** could not be accounted for by increased ion intensity.

The proposed biosynthesis of this new member of the lobosamide family via epoxidation followed by intramolecular cyclization (Figure 5.6b) is analogous to the production of dracolactams A and B from a common macrolactam precursor[16] and the production of mirilactams C-E from the macrocyclic precursor mirilactam A.[17] It has been proposed that the key epoxidation step in this proposed biosynthetic pathway is catalyzed by a cytochrome P450 epoxidase that is encoded outside of the macrolactam

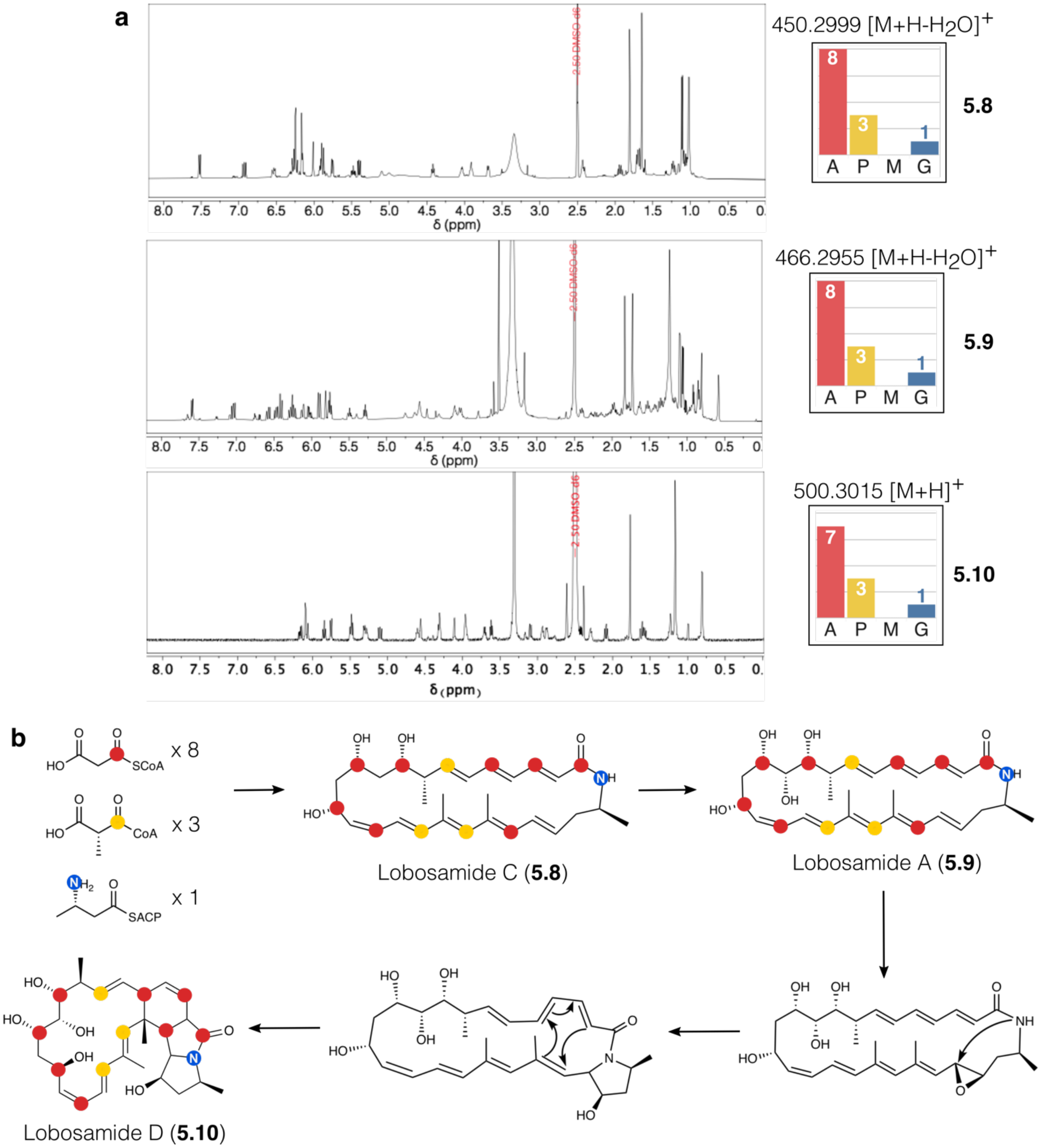


Figure 5.6 Comparison of NMR and SIL Incorporation in Lobosamides
(a) Comparison of the ^{1}H NMR spectra and IsoAnalyst profiles of lobosamides **5.8**, **5.9**, and **5.10**. (b) Biosynthetic scheme for post-PKS tailoring reactions which transform **5.8** to **5.10.** This biosynthetic scheme is a hypothesis put forth by the Abe group, where a variety of similar cyclized macrolactams have been discovered.[18]

core BGC.[18] This transformation has been observed in other macrolactam BGCs, but no epoxidase is found in any of the BGCs that are known to produce these cyclized polyene macrolactam compounds, supporting this hypothesis.[16–21] This complicates the process of predicting intramolecular cyclization events and other complex late-stage structural rearrangements for structures such as **5.10** based solely on the BGC sequence. Importantly, IsoAnalyst is able to identify biogenetic relationships between molecules, even for those which have significant structural rearrangements, highlighting a key advantage of this new platform for linking molecules to their cognate BGCs.

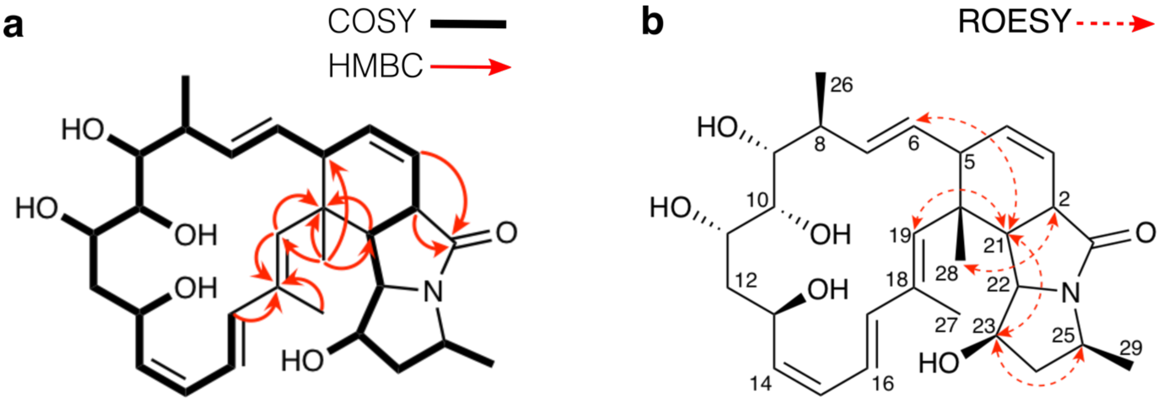


Figure 5.7 Key NMR Correletions for Lobosamide D (5.10)
(a) COSY correlations indicated by bold bonds. Key HMBC correlations indicated by red solid arrows. (b) Key ROESY correlations indicated by dashed red arrows.

5.5. Conclusion

In this chapter I have demonstrated how a complete genome sequence can be analyzed for BGCs and manually curated based on expected substrates using freely available computational tools. This BGC analysis workflow was developed in collaboration with our colleagues Dr. Marnix Medema and Dr. Justin van der Hooft, and represents the utilization of the most recent tools and databases available. The complete BGC analysis described in this chapter would allow for the careful selection of SIL tracers to optimize discovery of compounds from a given organism. This flexibility sets IsoAnalyst apart from many currently available tools that require multiply labeled or specialized precursors to identify SIL incorporation, and allows for a wide range of both targeted and untargeted experimental designs. Using four general SIL tracers that are widely applicable to multiple biosynthetic classes, I successfully identified two compound families present by their SIL incorporation patterns and further characterized ten specific compounds from within these groups. This analysis significantly reduced the number of interesting *m/z* features to manually investigate, and easily associated *m/z* features across the entire run on the basis of SIL incorporation.

Although I show in Table 5.4 that the general SIL tracers used in this study do have good coverage of the BGCs present in the *Micromonospora* sp. genome, I only succeeded in identifying novel compounds within known classes. The application of IsoAnalyst to different growth conditions, perturbations, or genetically modified organisms offers future opportunities for novel compound class discovery through the characterization of novel or unknown BGCs. Despite this, I have shown that novel chemistry can be discovered from known BGCs and many unknown *m/z* features with identifiable SIL incorporation detected by IsoAnalyst can also be associated with a known biosynthetic class.

5.6. Methods

5.6.1. Parallel SIL Fermentation with *Micromonospora* sp*.*

The parallel SIL fermentation protocol, sample extraction, and UPLC-MS method used to analyze *Micomonospora* sp. in this chapter were performed exactly as described in Chapter 3.

5.6.2. BGC Analysis of Complete Micromonospora sp. Genome

The genome of the *Micromonospora sp*. isolate was first mined for BGCs using antiSMASH version 5.2.0-8ecc354, resulting in 26 BGCs. AntiSMASH can predict numerous natural product compound classes including non-ribosomal peptides, type I, II, or III polyketides, lanthipeptides, and terpenes. These predicted BGCs were used to list the type and number of substrates needed to biosynthesize the product. In some cases, such as for NRPS, our analysis provided input for the predicted substrate specificity of each BGC domain detected, which helped to list the various expected amino acids used as a substrate. Likewise, the PKS domain architecture information was used to assess the number of expected acetate, malonate, methoxymalonate, and methylmalonate substrate building blocks incorporated into the polyketide. In some cases, the output from this analysis was further investigated by looking at the closest associated MIBiG entry for each BGC to find more information on the expected substrates and other chemical moieties likely to be incorporated based on similarity of enzyme-coding genes to those found in experimentally characterized pathways. Literature information and uniqueness of enzymatic domains present were taken into consideration in generating the substrate predictions shown in Table 5.2 ultimately a result of computationally predicted substrates where possible and expert knowledge on biochemical pathways where needed and possible based on known enzyme functions, for example in the case of methylations by SAMs.

5.6.3. Large-Scale Fermentation and Extraction of *Micromonospora* sp.

Micromonospora sp. was grown in 2.8 L Fernbach flasks containing 20.0 g of Amberlite XAD-16 adsorbent resin, a stainless steel spring, and 1 L of the minimal media as described above for use in the SIL experiments. Four large scale fermentations were performed without the supplementation of SIL precursors. Cultures were shaken at 200 rpm for 6 days, at which time the cultures were filtered by vacuum filtration on Whatman glass microfiber filters. The cells and resin were collected and extracted with 250 mL of 1:1 methanol/dicholormethane. The organic extract was collected by vacuum filtration and dried by rotary evaporation. The crude organic extract was initially separated into seven fractions by a stepwise methanol/water elution (10, 20, 40, 60, 80, 100 vol/vol) and an additional ethyl acetate wash step on a RediSep Rf C18 cartridge (Teledyne Isco) using a Teledyne Isco CombiFlash Rf flash chromatography system.

5.6.4. Isolation and Characterization of Lobosamide C and D

Purification of **5.8** and **5.10** was carried out using an Agilent 1200 series HPLC system. Lobosamide C (**5.8**) (0.7 mg) was isolated from the 100% methanol pre-fraction by HPLC (Phenomenex Kinetix XB-C18 5μm, 250 x 4.6 mm) using an isocratic separation (45% methanol + 0.02% formic acid, 45% H_2O + 0.02% formic acid, and 10% isopropyl alcohol with 0.02% formic acid) with a flow rate of 1.2 mL/min for 20 minutes. Lobosamide C (**5.8**) was eluted from the column at 34.4 min and was collected by UV detection at 300 nm. Lobosamide D (**5.10**) (0.4 mg) was isolated from the 40% methanol pre-fraction by HPLC (Phenomenex Kinetix XB-C18 5μm, 250 x 4.6 mm) using an isocratic separation (23% acetonitrile + 0.02% formic acid and 77% H_2O + 0.02% formic acid) with a flow rate of 1.2 mL/min for 20 minutes. Lobosamide D (**5.10**) was eluted from the column at 15.0 min and was collected by UV detection at 280 nm.
Lobosamide D: $[\alpha]_D$ = -166.7 (c 0.042, MeOH); UV (MeOH), λ_{max} 275 nm, log $\mathcal{E}$ = 3.45; HRMS (*m/z*): $[M+H]^+$ calcd. for $C_{29}H_{41}NO_6$, 500.3007; found, 500.3012. (see Appendix B

Table B1 for NMR shifts), SMILES: C[C@H]1C[C@@H](O)[C@]2([H])[C@@]3([H])[C@]4(C)/C=C(C)/C=C/C=C\[C@H](O)C[C@H](O)[C@@H](O)[C@H](O)[C@@H](C)/C=C/[C@@H]4C=C[C@]3([H])C(N12)=O

5.6.5. Full NMR Characterization of Lobosamide D

The molecular formula $C_{29}H_{41}NO_6$ was calculated from the $[M+H]^+$ ion detected at *m/z* 500.3012 (calcd. 500.3007, Δppm = 1.0), indicating 10 degrees of unsaturation. Review of the ^{1}H and phase-sensitive HSQC spectra revealed 4 x CH_3, 2 x CH_2, and 20 x CH including 9 olefinic signals. Further evaluation of the ^{13}C and HMBC spectra identified a further 3 x qC signals, including one olefinic carbon, and one amide carbonyl. These components accounted for all carbons, 36 of 41 protons, one oxygen and one nitrogen from the molecular formula. The remaining five oxygens and five protons were assigned as hydroxy groups, based on the presence of five broad exchangeable signals in the proton spectrum and five signals in the HSQC (^{1}H 3.35 - 4.61 ppm, ^{13}C 65.5 - 75.7 ppm) consistent with oxygenated methines. This completed the detection of all the elements in the molecular formula in the spectral data.

To solve the planar structure the COSY spectrum was used to identify one large spin system that incorporated all but two of the protonated carbons in the molecule. Starting from an olefinic doublet at 6.07 ppm (H17) sequential COSY correlations to 6.16 (H16), 5.84 (H15) and 5.48 (H14) ppm identified a diene motif. COSY signals from 5.48 (H14) sequentially to 4.61 (H13), diastereotopic methylene protons at 1.23 and 1.62 (H12), and oxygenated methine protons at 3.97 (H11), 3.35 (H10), and 2.92 (H9) indicated a polyhydroxylated region which in turn was connected to another olefin through COSY correlations sequentially from 2.29 (H8) to 5.47 (H7) and from 5.47 (H7) to 5.10 (H6). This completed the linear portion of the spin system highlighted in red in Figure 5.8a.

The next section of the structure elucidation was complicated by the presence of a large number of aliphatic methine signals, suggestive of a complex fused ring system. From the last olefinic proton (H6) a series of sequential COSY signals connected protons at 2.88 (H5), 5.32 (H4), 5.75 (H3), 3.10 (H2), 2.09 (H21), 3.62 (H22), 3.96 (H23), a diastereotopic methylene at 1.58 and 2.42 (H24), 3.71 (H25) and a terminal methyl

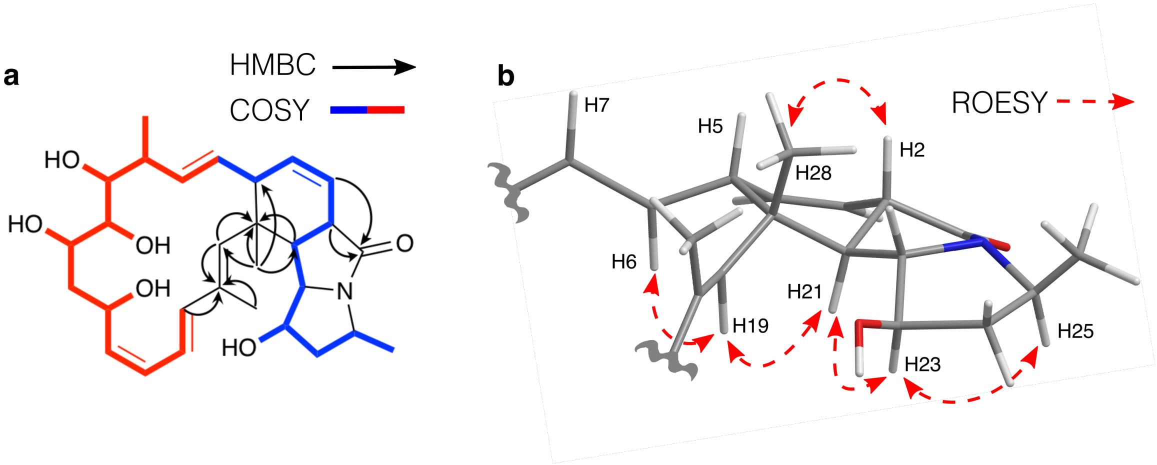


Figure 5.8 Key NMR Correlations for Configurational Analysis
(a) Key COSY and HMBC correlations used to solve the planar structure of lobosamide D (**5.10**). (b) Diagram of key ROESY correlations used in the configurational analysis of lobosamide D (**5.10**).

signal at 1.16 ppm. This completed the second component of the spin system illustrated in blue in Figure 5.8a.

Assembly of the tetracyclic ring system in **5.10** required extensive use of HMBC data. This was complicated by severe signal overlap for two of the methyl signals (H28 and H29). Key HMBC signals from H2 and H3 to the amide carbonyl carbon (C1) at 172.2 placed the carbonyl at C2. This left a total of five carbons; 1 x aliphatic CH_3, 1 x vinylic CH_3, 1 x olefinic CH, 2 x qC. This was suggestive of a trisubstituted olefin, which was supported by the presence of a methyl singlet at ^{1}H 1.76 ^{13}C 16.1 ppm (H27) which showed HMBC correlations to the olefinic methine (C19) and a quaternary carbon at 134.9 (C18). In addition, this vinyl methyl signal (H27) showed a strong HMBC correlation to C17, placing it at the terminus of the major spin system. The olefinic methine signal (H19) showed reciprocal HMBC correlations to C17, C18 and C27, confirming this assignment. H19 possessed two additional HMBC correlations, one to the remaining quaternary carbon at 38.9 ppm (C20) and the other to the remaining methyl carbon at 19.4 ppm (C28). Closer examination of the quaternary carbon at 38.9 (C20) revealed an HMBC correlation from H21, closing the macrocyclic ring. Finally, a strong HMBC correlation from methyl H28 to C20 placed the remaining methyl group on carbon 20. Additional HMBC correlations from this methyl group to C5 and C21 confirmed the presence of the 6, 5 fused ring system, and completed all of the carbon-carbon bond connections in the molecule.

The five hydroxyl groups were located on carbons 9, 10, 11, 13 and 23 based on COSY correlations from each broad OH proton to its associated methine proton (Appendix B Figure B4). Finally, the remaining nitrogen atom on the amide functional group was connected to the two remaining open positions in the molecule (C22 and C25). This assignment was supported by the chemical shifts for these two carbons (65.5 and 47.1 respectively) and completed the planar structure assignment, accounting for all of the degrees of unsaturation.

Lobosamide D (**5.10**) was isolated from the same strain, *Micromonospora sp*. RL09-050-HVF-A, which was previously published by our laboratory as the producing organism of lobosamides A-C. To determine the complete absolute configurations of lobosamides A-C we previously obtained a full genome sequence for this strain and identified the *lob* PKS biosynthetic gene cluster. This sequence data, in conjunction with extensive dipolar coupling NMR experiments, defined the absolute configuration at every position in the molecule.[3]

Given the common polyketide core precursor between lobosamides A-C and lobosamide D (**5.10**) and the absence of any other relevant polyketide BGCs in the genome of the producing organism we hypothesize that this same BGC also produces lobosamide D. The absolute configurations of positions 9, 11, and 13 of lobosamides A-C were previously determined by genetic analyses of three ketoreductase (KR) domains responsible for these hydroxyls. The relative configurations of the methyl at position 8 and the hydroxyl at position 10 were determined experimentally and the configurations were fully assigned through comparison to the hydroxyl stereocenters (9,11, and 13) produced by KR domains. The stereocenter at position 25 in all of the lobosamides and related compounds including salinilactam and micromonolactam, is derived from a 3-aminobutyrate starter unit. The enzymatic mechanism that produces this starter unit was characterized for incednine and demonstrated stereospecificity in producing (*S*)-3-aminobutyrate. This stereospecificity helped assign positions 25*S* for lobosamides A-C, and this conserved stereospecificity for compounds containing the (S)-3-aminobutyrate starter unit has been shown experimentally in other related natural products (mirilactam, micromonolactam, etc). The same absolute configuration is assigned to position 25 in lobosamide D. Based on these previously established data, we determined the absolute configuration of the following positions of lobosamide D to be 8*S*, 9*R*, 10*R*, 11*S*, 13*R*, 25*S*.

The double bond configurations in lobosamide D (**5.10**) are also expected to match the configurations in the lobosamides. Double bond configurations were consistent in lobosamides A-C except for the olefin at C14-C15, which is 14E in lobosamide B and 14Z in lobosamides A and C. Evaluation of key coupling constants between olefinic protons in lobosamide D confirmed that it matched the configuration of lobosamide A in all positions including C14 ($^3J_{H14\text{-}H15}$ = 10.1 Hz). The full double bond configurations for lobosamide D are 3Z, 6E, 14Z, 16E, and 18E ($^3J_{H3\text{-}H4}$ = 9.8 Hz, $^3J_{H6\text{-}H7}$ = 14.6 Hz, $^3J_{H14\text{-}H15}$ = 10.1 Hz, and $^3J_{H16\text{-}H17}$ = 15.7 Hz) The double bond at position 18E was corroborated by ROESY correlations between H19/H21 and H5/H27, indicating that H19 and H27 are on opposite sides of the 6-membered ring.

The full absolute configuration of the 5-5-6 fused ring system was determined using ROESY correlations. When possible, ROESY correlations from 2D ROESY experiments were confirmed by selective 1D ROESY experiments. ROESY correlations between H21/H6, H21/H19, and H21/H23 indicated that protons H6, H19, H21, and H23 were all located on the same face of the ring system (Appendix B Figures B7-8, Figure 5.8b) A ROESY correlation between H2/H28 and the trans relationship between H2 and H21 ($^3J_{H2\text{-}H21}$ = 13.0 Hz) indicated that H2 and H28 are located on the same face of the six membered ring (Appendix B Figure B7, Figure 5.8b).

The absolute configuration at position 25*S* was previously determined by BGC analysis which indicated the starter unit (*S*)-3-aminobutyrate. However, no correlations could be observed in the 2D ROESY spectra between H23 and H25 or H29 in order to establish the relationship between the fused ring system and the known configuration at position 25. To address this, we performed a 1D ROESY experiment selectively irradiating H25 and observed a correlation to H23 (Appendix B Figure B10). This connected the known absolute configuration at position 25 to the relative configurations in the 6,5,5 ring system determined by ROESY experiments and allowed us to assign the full absolute configuration of lobosamide D as 2*R*, 5*S*, 8*S*, 9*R*, 10*R*, 11*S*, 13*R*, 20*R*, 21*S*, 22*S*, 23*R*, 25*S*.

References

1. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

2. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

3. Schulze, C. J. *et al.* Genome-Directed Lead Discovery: Biosynthesis, Structure Elucidation, and Biological Evaluation of Two Families of Polyene Macrolactams against Trypanosoma brucei. *ACS Chem. Biol.* **10**, 2373–2381 (2015).

4. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).

5. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362--W367 (2011).

6. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).

7. Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).

8. Ronan, J. L. *et al.* Desferrioxamine biosynthesis: Diverse hydroxamate assembly by substrate-tolerant acyl transferase DesC. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170068 (2018).

9. Telfer, T. J., Gotsbacher, M. P., Soe, C. Z. & Codd, R. Mixing Up the Pieces of the Desferrioxamine B Jigsaw Defines the Biosynthetic Sequence Catalyzed by DesD. *ACS Chem. Biol.* **11**, 1452–1462 (2016).

10. Burrell, M., Hanfrey, C. C., Kinch, L. N., Elliott, K. A. & Michael, A. J. Evolution of a novel lysine decarboxylase in siderophore biosynthesis. *Mol. Microbiol.* **86**, 485–499 (2012).

11. Yang, Y. L. *et al.* Connecting chemotypes and phenotypes of cultured marine microbial assemblages by imaging mass spectrometry. *Angew. Chemie - Int. Ed.* **50**, 5839–5842 (2011).

12. Sidebottom, A. M., Johnson, A. R., Karty, J. A., Trader, D. J. & Carlson, E. E. Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from *Streptomyces coelicolor* M145. *ACS Chem. Biol.* **8**, 2009–2016 (2013).

13. D’Onofrio, A. *et al.* Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria. *Chem. Biol.* **17**, 254–264 (2010).

14. Traxler, M. F., Watrous, J. D., Alexandrov, T., Dorrestein, P. C. & Kolter, R. Interspecies interactions stimulate diversification of the *Streptomyces coelicolor* secreted metabolome. *MBio* **4**, e00459-13 (2013).

15. Schulze, C. J. *et al.* ‘function-first’ lead discovery: Mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20**, 285–295 (2013).

16. Hoshino, S. *et al.* Mycolic Acid Containing Bacterium Stimulates Tandem Cyclization of Polyene Macrolactam in a Lake Sediment Derived Rare Actinomycete. *Org. Lett.* **19**, 4992–4995 (2017).

17. Hoshino, S. *et al.* Mirilactams C–E, novel polycyclic macrolactams isolated from combined-culture of *Actinosynnema mirum* NBRC 14064 and mycolic acid-containing bacterium. *Chem. Pharm. Bull.* **66**, 660–667 (2018).

18. Hoshino, S., Onaka, H. & Abe, I. Activation of silent biosynthetic pathways and discovery of novel secondary metabolites in actinomycetes by co-culture with mycolic acid-containing bacteria. *J. Ind. Microbiol. Biotechnol.* **46**, 363–374 (2019).

19. Hoshino, S. *et al.* Niizalactams A-C, Multicyclic Macrolactams Isolated from Combined Culture of Streptomyces with Mycolic Acid-Containing Bacterium. *J. Nat. Prod.* **78**, 3011–3017 (2015).

20. Oh, D. C., Poulsen, M., Currie, C. R. & Clardy, J. Sceliphrolactam, a polyene macrocyclic lactam from a wasp-associated *Streptomyces* sp. *Org. Lett.* **13**, 752–755 (2011).

21. Derewacz, D. K., Covington, B. C., McLean, J. A. & Bachmann, B. O. Mapping Microbial Response Metabolomes for Induced Natural Product Discovery. *ACS Chem. Biol.* **10**, 1998–2006 (2015).

Chapter 6.

Conclusions

In this thesis I have described the development, optimization, and validation of an MS-based workflow for the untargeted characterization of complex chemical phenotypes produced by BGCs in microorganisms. The ultimate goal of this work is to provide a hypothesis-generating workflow for connecting compounds to BGCs using parallel SIL experiments. I have demonstrated various limitations to the IsoAnalyst approach, and highlighted important factors to consider when applying this approach to different microorganisms. I have shown the IsoAnalyst can accurately detect SIL incorporation into compounds with well-established biosynthetic pathways, and identify novel variants of compounds from known classes. The flexibility of this tool will offer opportunities for the creative application towards understanding natural product biosynthesis and novel compound discovery.

6.1. Limitations

The metabolic limitations of SIL incorporation need to be considered and optimized prior to SIL feeding and I have shown that considerations of both core minimal media components and SIL tracer selection are essential to experimental optimization. In Chapter 2 I discussed the steps that went into developing the minimal medium used throughout this thesis. However, this optimization overall resulted in a lack of diversity in the compounds produced by *Micromonospora* sp. in the full application of the workflow in Chapter 5. The application of this medium with the four SIL tracers I selected was sufficient for the initial development of this method, but a further expansion of both growth media and SIL tracers is necessary to expand upon the types of compounds identified. Although the use of minimal media is often limited in eliciting natural products, I demonstrated in Chapter 2 that a rich media can also be used and still result in sufficient SIL incorporation for some SIL tracers. The IsoAnalyst workflow lends itself to the parallelization of many conditions and would be especially complementary to one-strain many compounds (OSMAC) approaches which test different types of media or elicitors of biosynthesis.

The MS signal intensity is the main technical limiting factor in the IsoAnalyst platform. I showed in Chapter 3 that signal intensity was a significant factor in the reliability of statistical measurements of heavy isotopologue peaks. This was further addressed in Chapter 5, as both the desferrioxamine and lobosamide families had more variability in SIL incorporation among different analogues due to signal intensity than enzymatic substrate flexibility. This limitation underpins the fact that IsoAnalyst does not provide an exact interpretation of the biosynthesis of every detected compound. Rather, it is a general way of associating MS features with related biosynthetic origins and generating hypotheses about the BGC responsible for compound production. Factors such as MS signal intensity, chromatographic peak shape, and overlap of related compounds greatly influence the data interpretation and currently still require manual interrogation to interpret.

While signal intensity is the biggest limitation in determining SIL incorporation, BGC analysis and substrate prediction also restrict our ability to connect the compounds identified with the correct BGC. In Chapter 5, I demonstrated how a few simple criteria can be used to qualitatively assess the likelihood of identifying a compound from a given BGC using a given set of SIL precursors. This is a generalizable approach that can be used to optimize SIL selection on the basis of the BGCs present in the genome. This process also entails many manual data curation steps, however the current rate of growth of BGC databases will likely increase the automation and confidence of these predictions as more BGC sequences are discovered and their cognate product structures confirmed.[1] Still, there are many BGCs that are detectable in the genome but have unreliable substrate predictions, and little resemblance to other BGCs which are not likely to be discovered using IsoAnalyst alone. IsoAnalyst is a complementary technique that can be used alongside more established approaches such a molecular networking, heterologous expression, and correlation-based BGC discovery to draw connections between compounds and BGCs.

6.2. Future Perspectives

There are many intriguing ideas about how this platform can be applied to natural product discovery and MS metabolomics in general. Using IsoAnalyst, one can focus on the entire chemical phenotype produced by a BGC, and assess how this phenotype changes under different environmental or genetic conditions. Rather than looking at a single product of a BGC, IsoAnalyst can assess the differences in the full profile of compounds produced by a BGC in a sample. This can be used to detect optimal conditions for production of specific analogues, or to better understand how substrate flexibility influences the ecological function of specialized metabolism. Evolutionary pressures may select for more highly specific biosynthetic enzymes to produce the most potent compound structure, much like structure-activity relationships are established in a laboratory setting during drug development. Yet we find more often than not that biosynthetic enzymes employ this substrate flexibility to produce a variety of related structures. IsoAnalyst represents a step towards understanding natural product discovery from a more holistic biosynthetic and ecological perspective. By catering the experimental design to a particular organism and compound class, IsoAnalyst can be used to observe whole phenotypic changes from different conditions, including co-culture. IsoAnalyst takes a systematic approach to understanding natural product biosynthesis with the assumption that BGCs are not hard-coded to produce bioactive compounds, but rather interact adaptably with the environment to produce different chemical compositions. In heterologous expression, and other genetic techniques used to induce compound production in native hosts, identifying the target compound structure can still be difficult while optimizing the system.[2] IsoAnalyst can identify related compounds that are significantly structurally modified but derived from the same biosynthetic precursors, such as in the case of lobosamide D, and could assist in the identification of unexpected biosynthetic products in heterologous host systems.

Advances in genomic sequencing, publicly available computational tools, and BGC databases have motivated large scale studies of BGC sequences and products.[3] Both correlation and feature based approaches are used separately and together to make connections between specific BGCs and compounds, and to draw big-picture conclusions about the distribution and variation of BGCs and the natural products they produce.[1,4,5] IsoAnalyst can be used in conjunction with correlation based approaches by comparing SIL incorporation in organisms with related BGCs to identify products. This allows for the comparison of not just single products, but overall phenotypic variation between organisms with similar biosynthetic potential. These insights will further our understanding of BGC convergent and divergent evolution by offering a chemistry-centric perspective of BGC diversity orthogonally to well-established genetic comparisons. IsoAnalyst is also complementary to feature-based approaches which focus on metabolite detection in MS1 but often in MS/MS features as well. Molecular networking has become one of the most widely used MS metabolomics tools for identifying related compounds and associating structures on the basis of related fragmentation.[6] IsoAnalyst can support the conclusions drawn from molecular networking analysis and contribute to structural hypotheses on the basis of which fragment ions retain or lose isotopic enrichment from specific SIL tracers. Overall, IsoAnalyst fills a gap in both correlation and feature based approaches by offering complementary biosynthetic information about *m/z* features that can be used to generate hypotheses and guide follow-up studies on specific compound classes.

IsoAnalyst can also be applied to microbial culture in other creative ways. Detection of SIL tracer incorporation into natural products over time can be used to observe compound production throughout the growth phase, or as a result of introducing a metabolic perturbation. By applying IsoAnalyst as a targeted workflow to observe specific groups of compounds, the timing and influence of environmental factors on chemical phenotype can be more carefully interrogated. Addition of SIL tracers at different time points can also be used to optimize the efficiency of SIL incorporation and directly observe the timing of biosynthetic activity throughout the growth phase. These applications are possible because of the flexibility of the IsoAnalyst experimental approach, and the statistical application used to detect SIL incorporation. By determining the number of detectable SIL tracers incorporated into m/z features, IsoAnalyst simplifies the data interpretation compared to metabolic flux measurements and other untargeted methods that compare overall SIL incorporation between conditions. The IsoAnalyst approach focuses on natural product biosynthesis in terms of the substrates used rather than the metabolic flux through the pathway to identify the entire metabolic phenotype in a single MS metabolomics experiment. The novelty and flexibility of this approach will offer many new opportunities for applications in natural product discovery, biosynthesis, and ecological function.

References

1. Van Der Hooft, J. J. J. *et al.* Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).

2. Zhang, M. M. *et al.* CRISPR-Cas9 strategy for activation of silent Streptomyces biosynthetic gene clusters. *Nat. Chem. Biol.* **13**, 607–609 (2017).

3. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).

4. Kenshole, E., Herisse, M., Michael, M. & Pidot, S. J. Natural product discovery through microbial genome mining. *Curr. Opin. Chem. Biol.* **60**, 47–54 (2021).

5. Eldjárn, G. H. *et al.* Ranking microbial metabolomic and genomic links using correlation-based and feature-based scoring functions. *bioRxiv* 2020.06.12.148205 (2020) doi:10.1101/2020.06.12.148205.

6. Zhang, M. M., Wang, Y., Ang, E. L. & Zhao, H. Engineering microbial hosts for production of bacterial natural products. *Nat. Prod. Rep.* **33**, 963–987 (2016).

Appendix A.

MS and NMR Data for Known Compounds and Commercial Standards

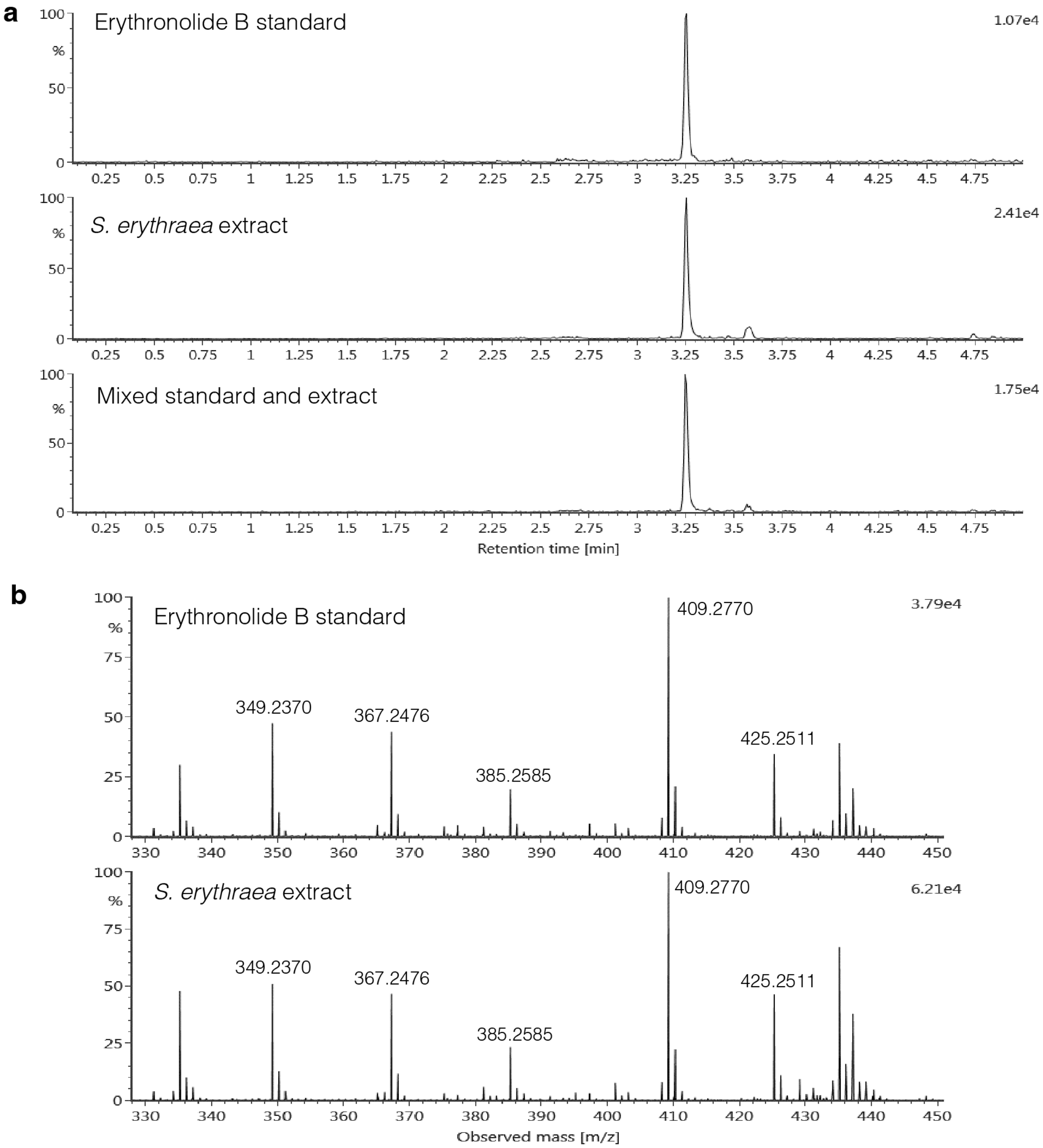


Figure A1. Erythronolide B Standard Co-Injection and MS

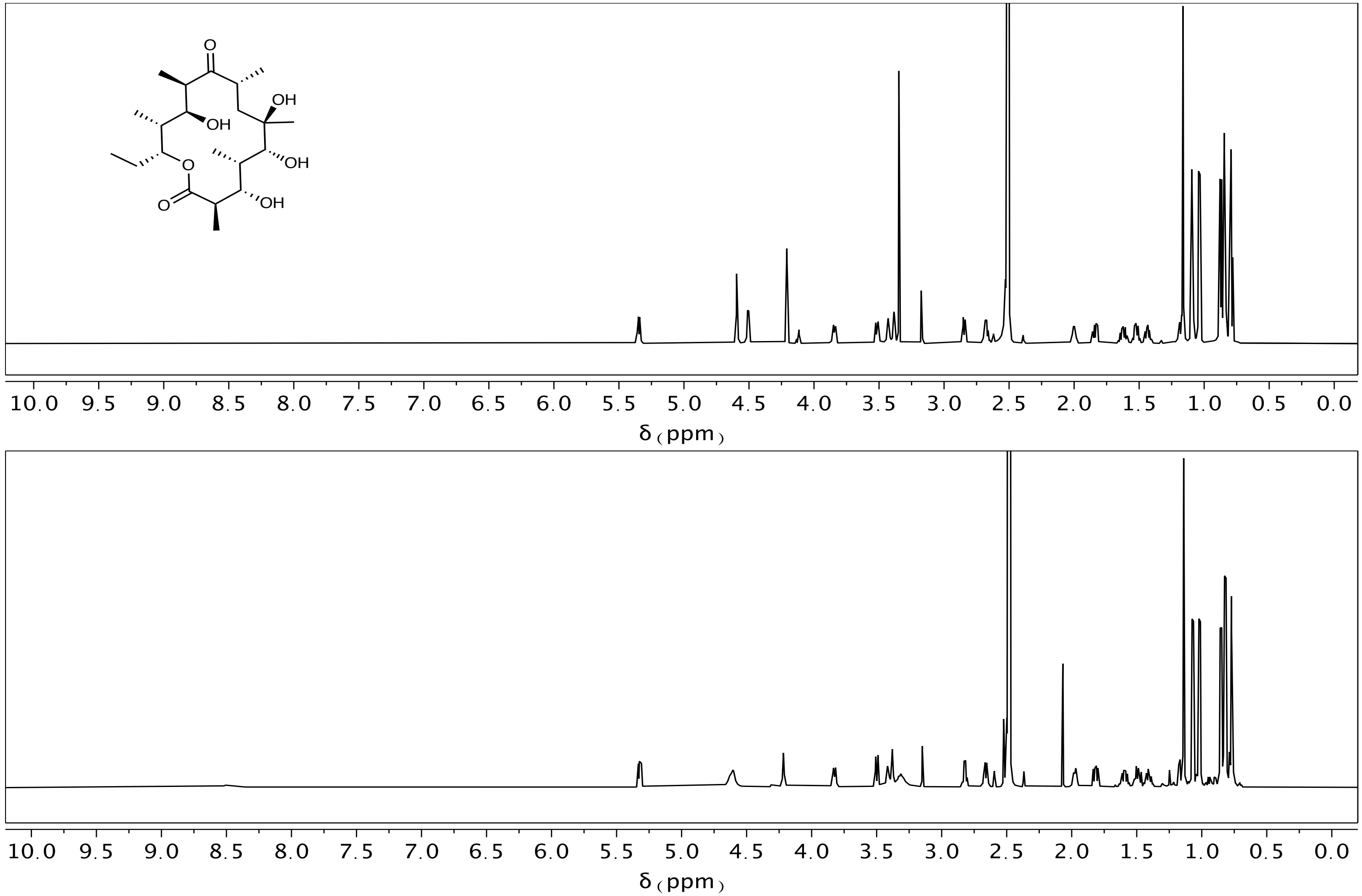
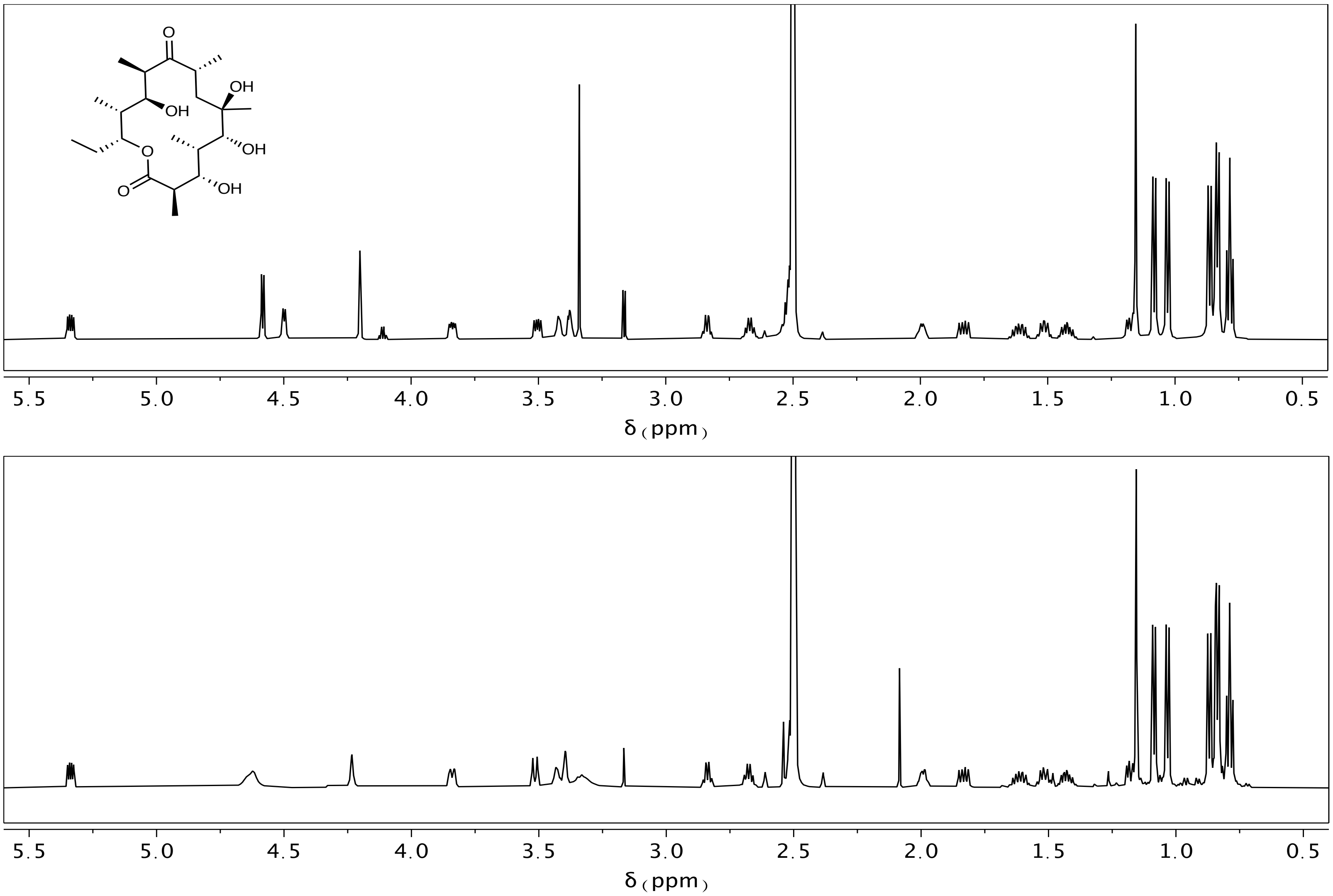


Figure A2. **^{1}H NMR Spectra of Erythronolide B acquired in DMSO-d_6 at 600MHz**
Commercial standard (top) and isolated compound (bottom)



Expansion of Figure A2.

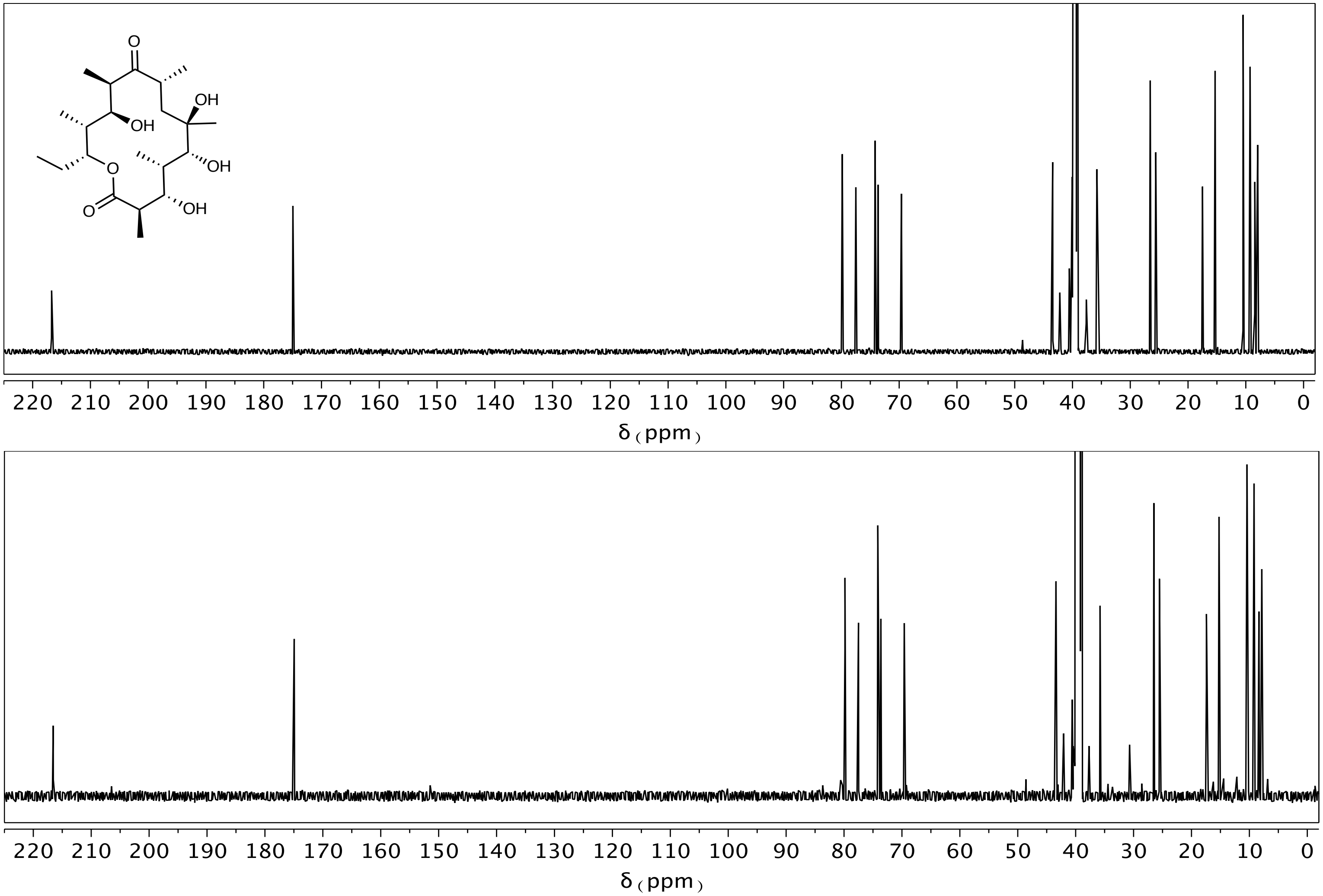
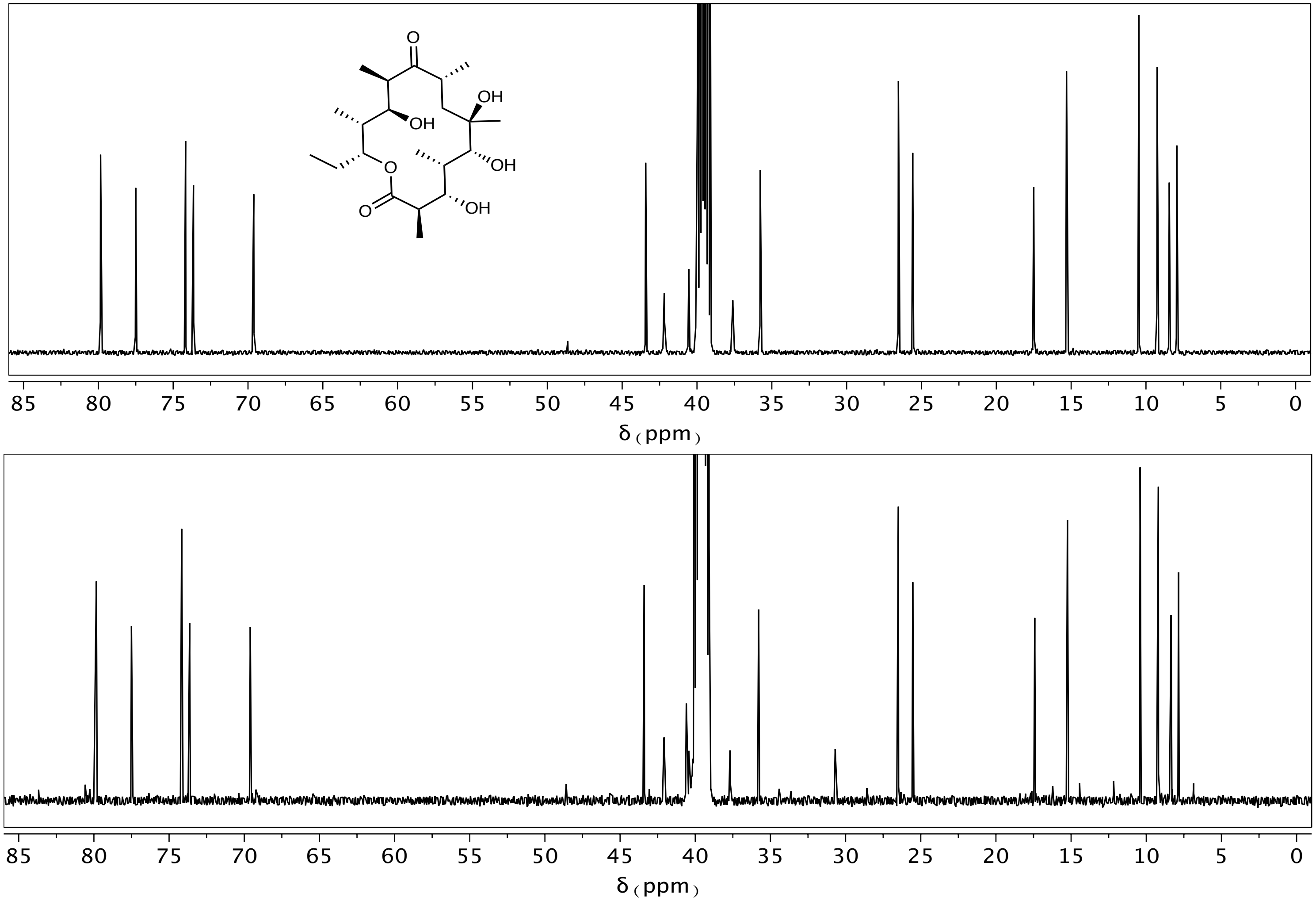


Figure A3. **^{13}C NMR Spectra of Erythronolide B acquired in DMSO-d_6 at 150 MHz**
Commercial standard (top) and isolated compound (bottom)



Expansion of Figure A3

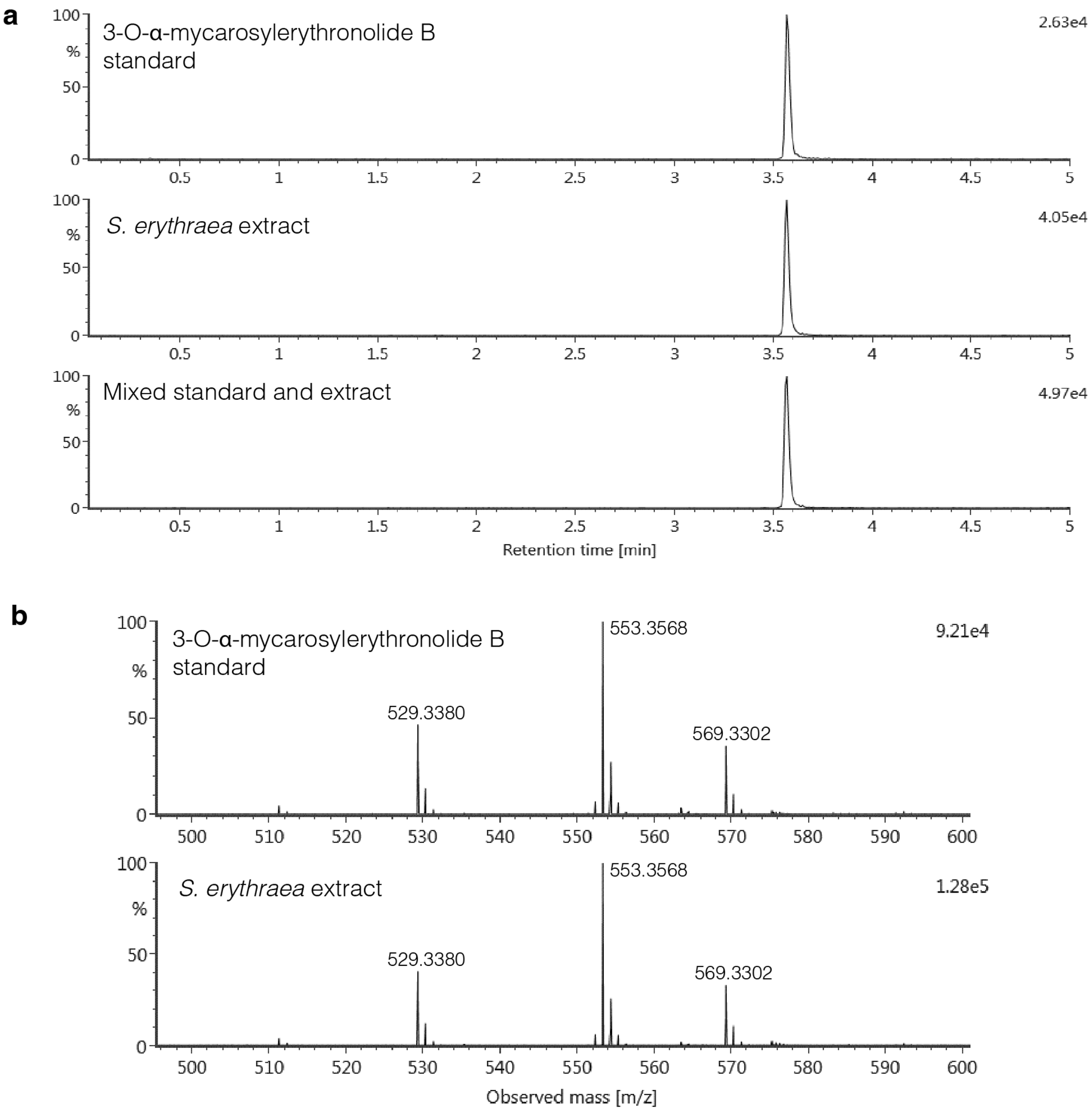


Figure A4. **3-O-α-mycarosylerythronolide B Standard Co-Injection and MS**

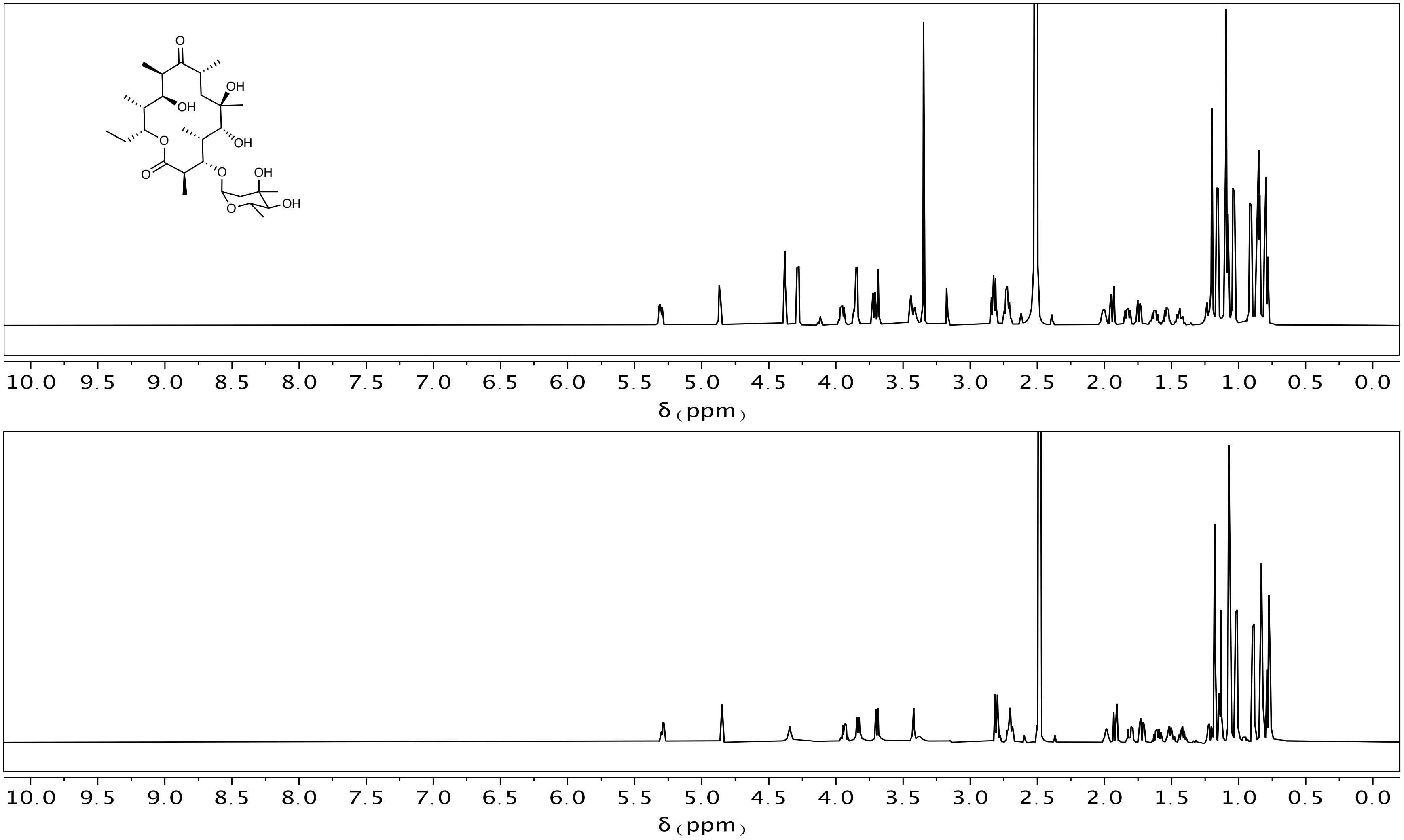
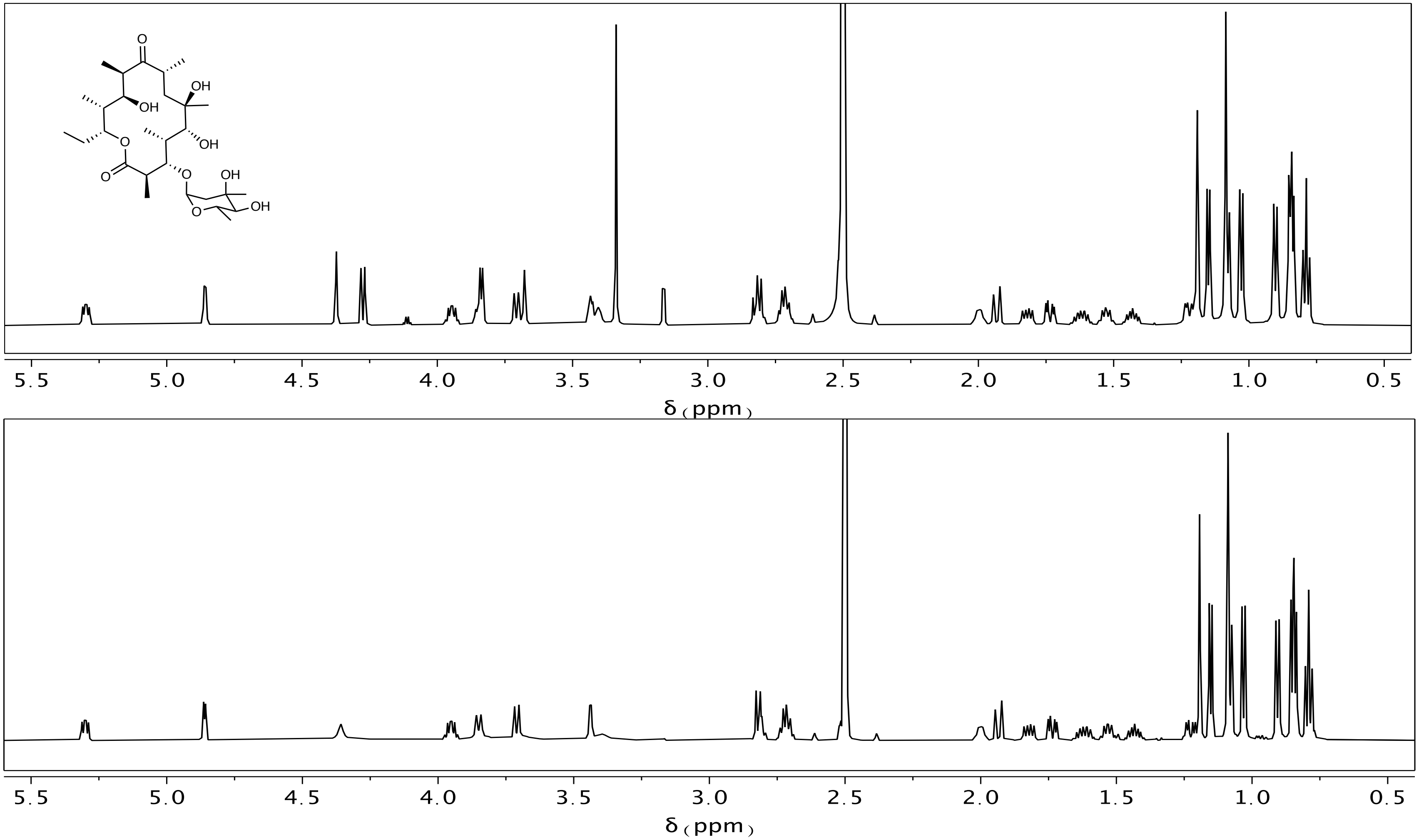


Figure A5. **^{1}H NMR Spectra of 3-O-α-mycarosylerythronolide B acquired in DMSO-d_6 at 600MHz**
Commercial standard (top) and isolated compound (bottom)



Expansion of Figure A5

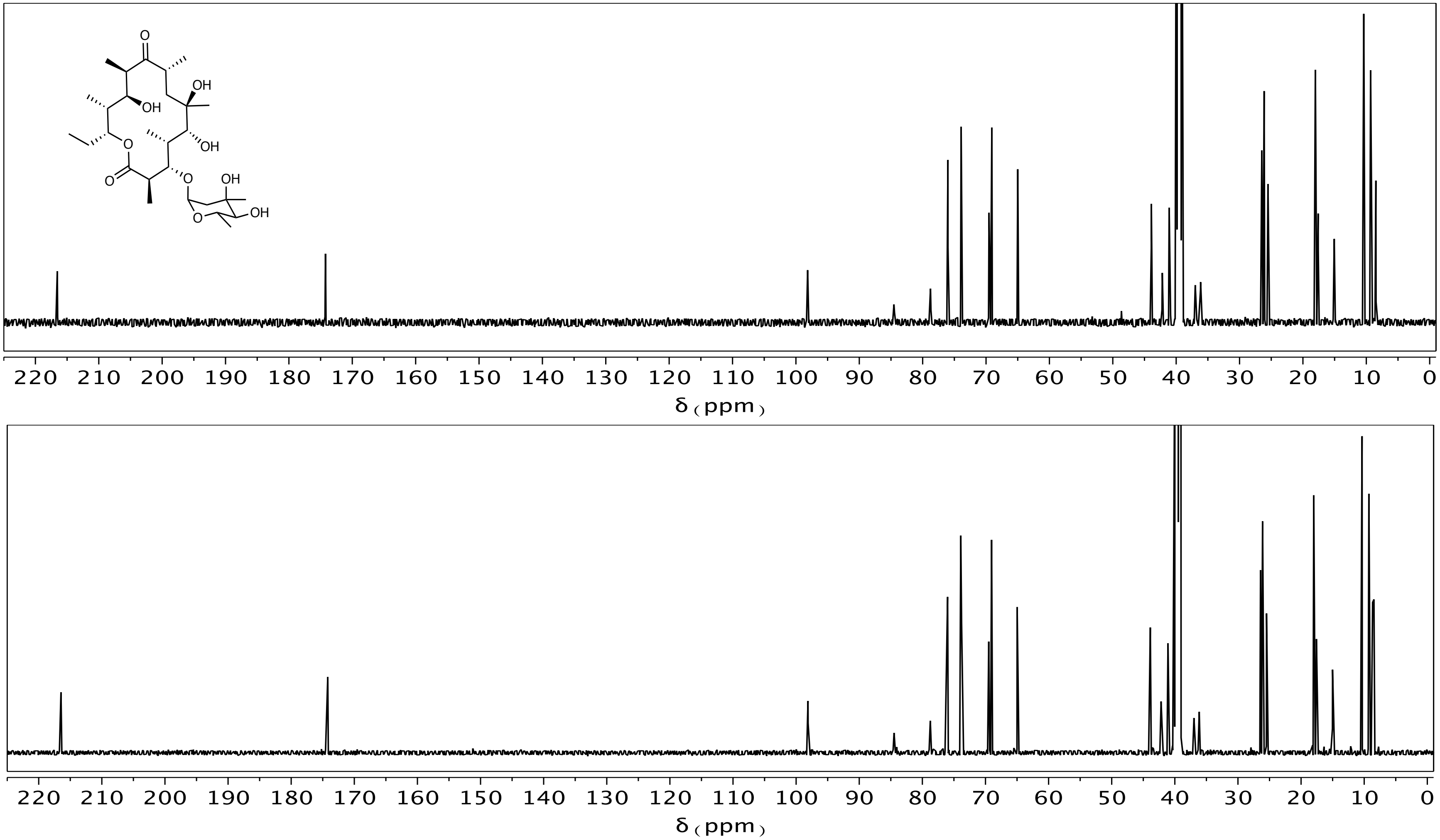
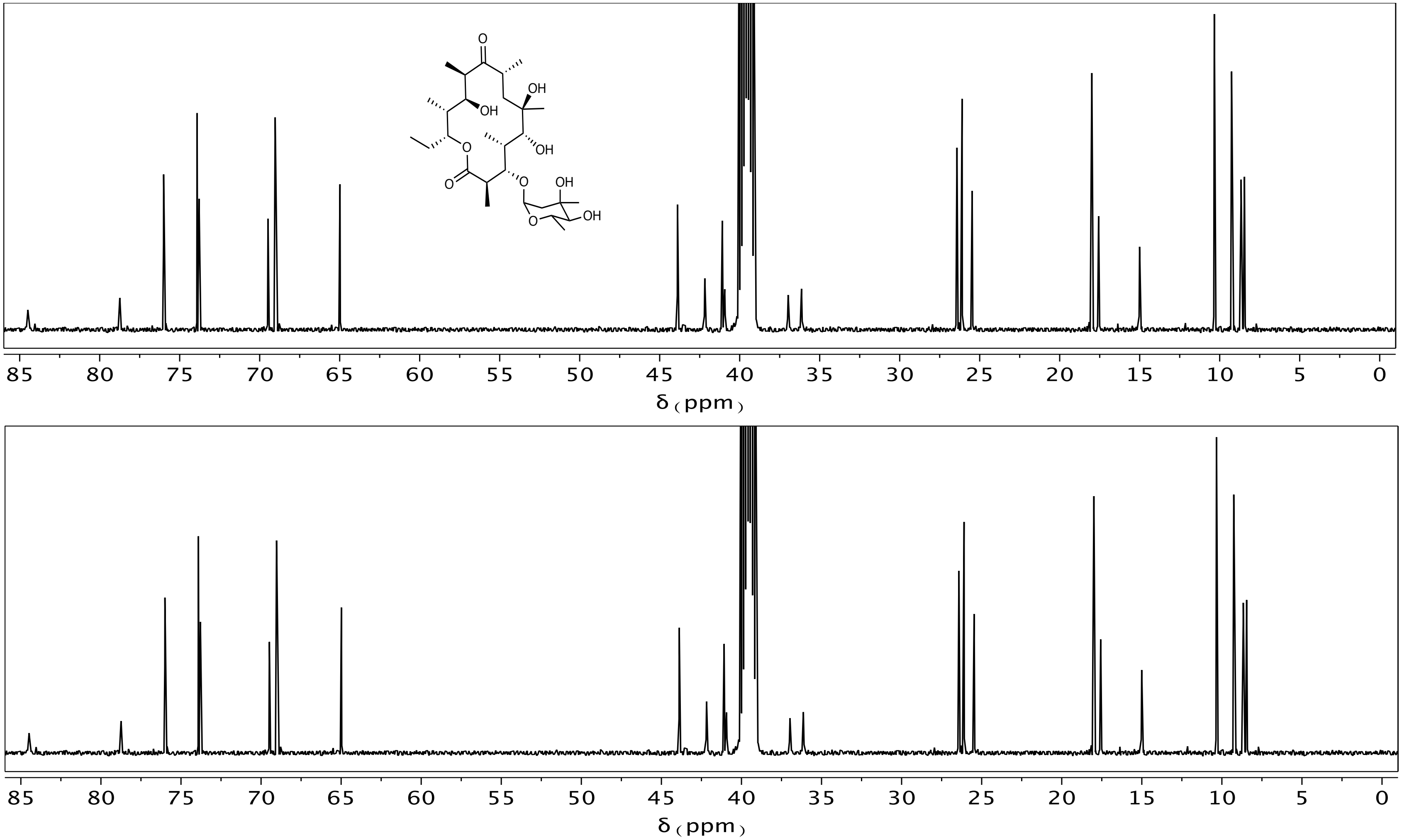


Figure A6. ^{13}C NMR Spectra of 3-O-α-mycarosylerythronolide B acquired in DMSO-d_6 at 150 MHz
Commercial standard (top) and isolated compound (bottom)



Expansion of Figure A6

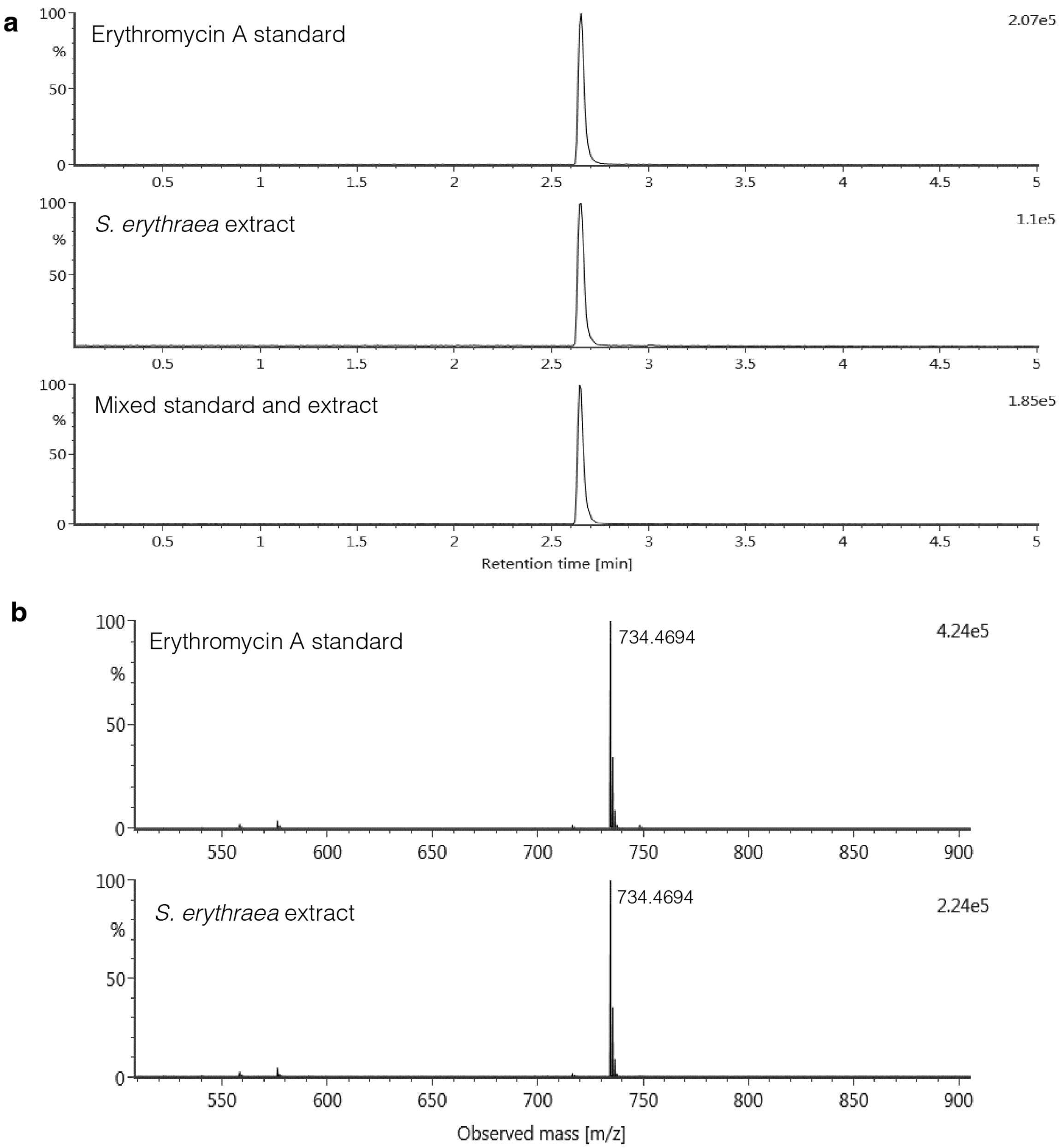


Figure A7. Erythromycin A Standard Co-Injection and MS

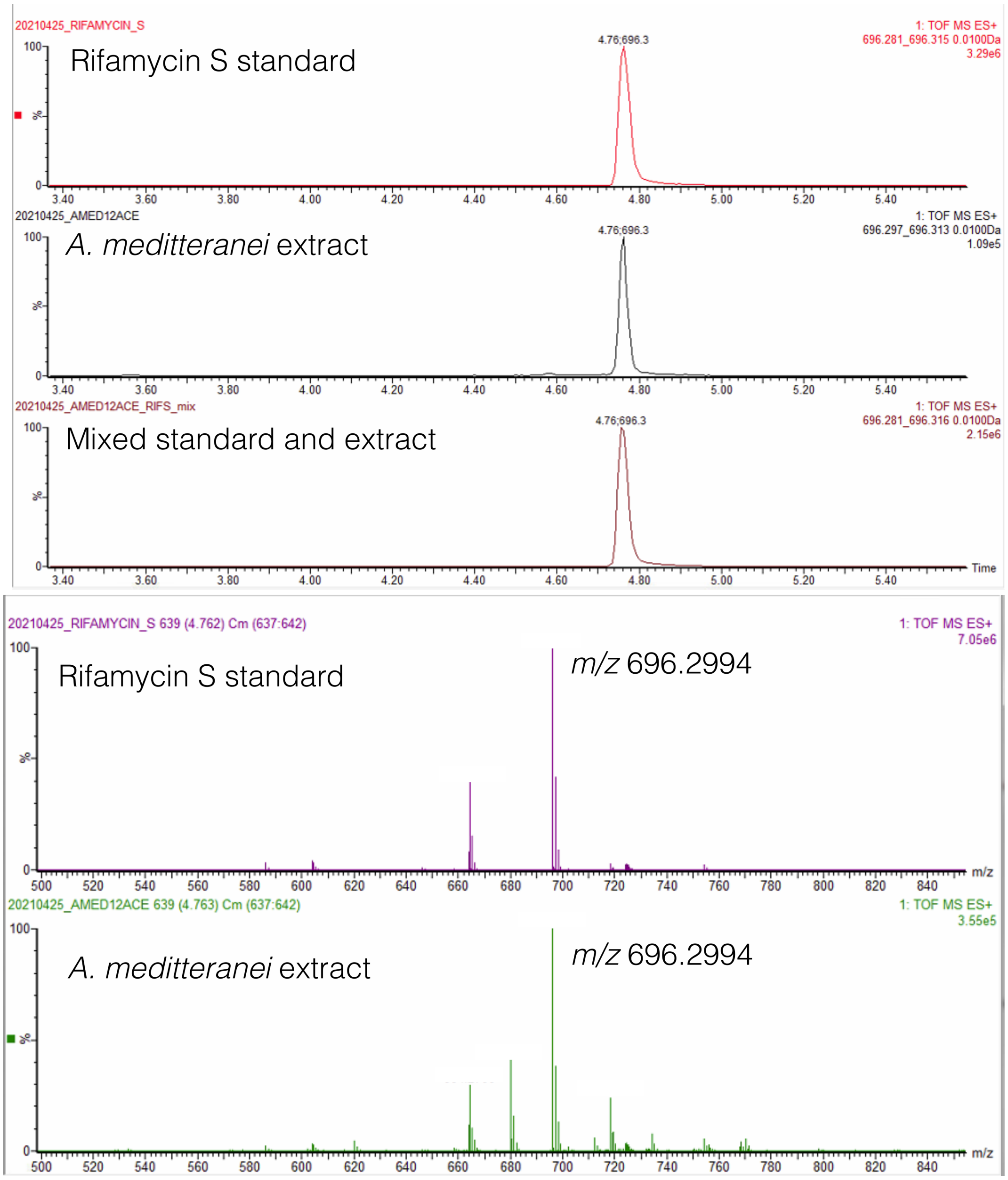


Figure A8. **Rifamycin S Standard Co-Injection and MS**

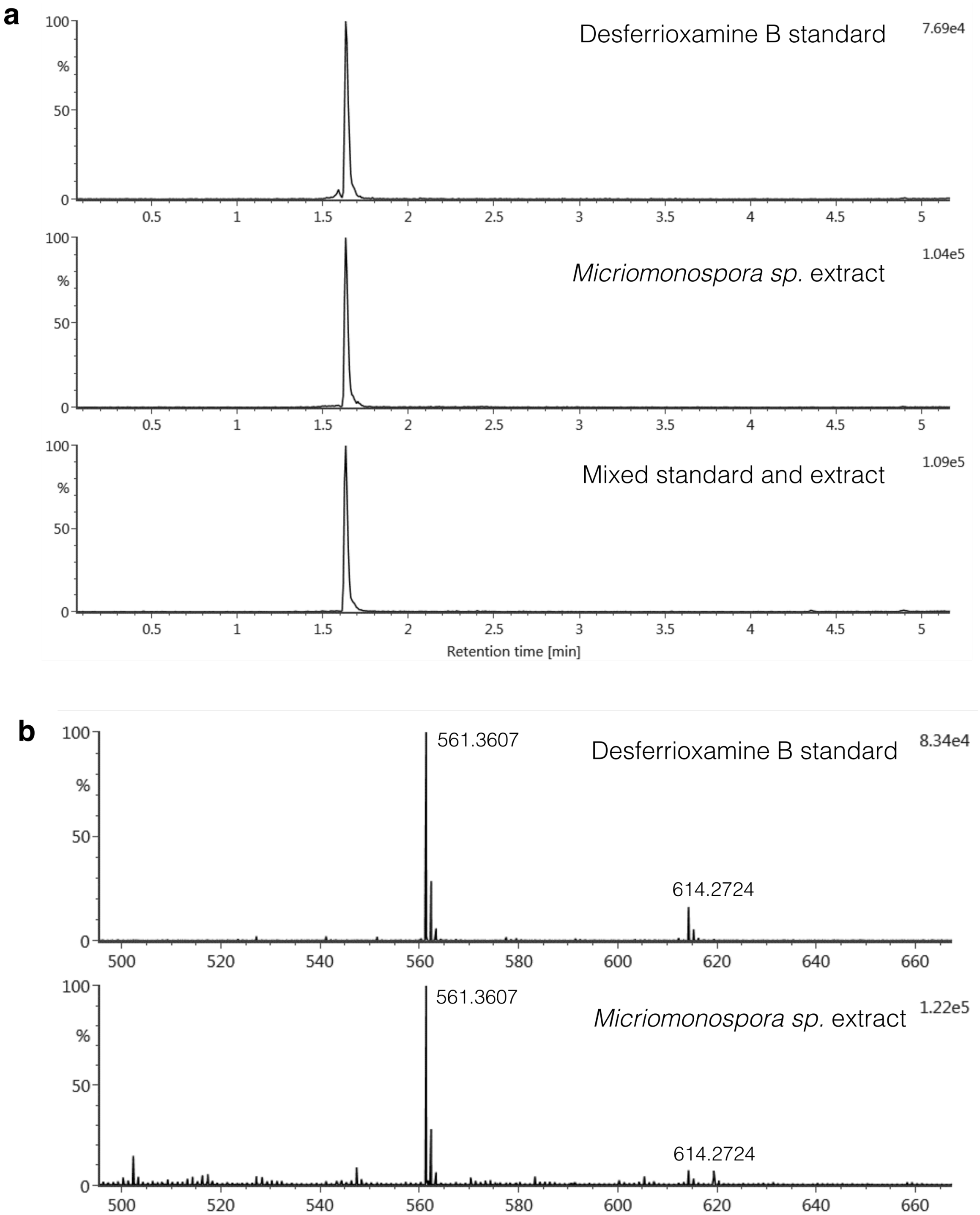


Figure A9. Desferrioxamine B Standard Co-Injection and MS

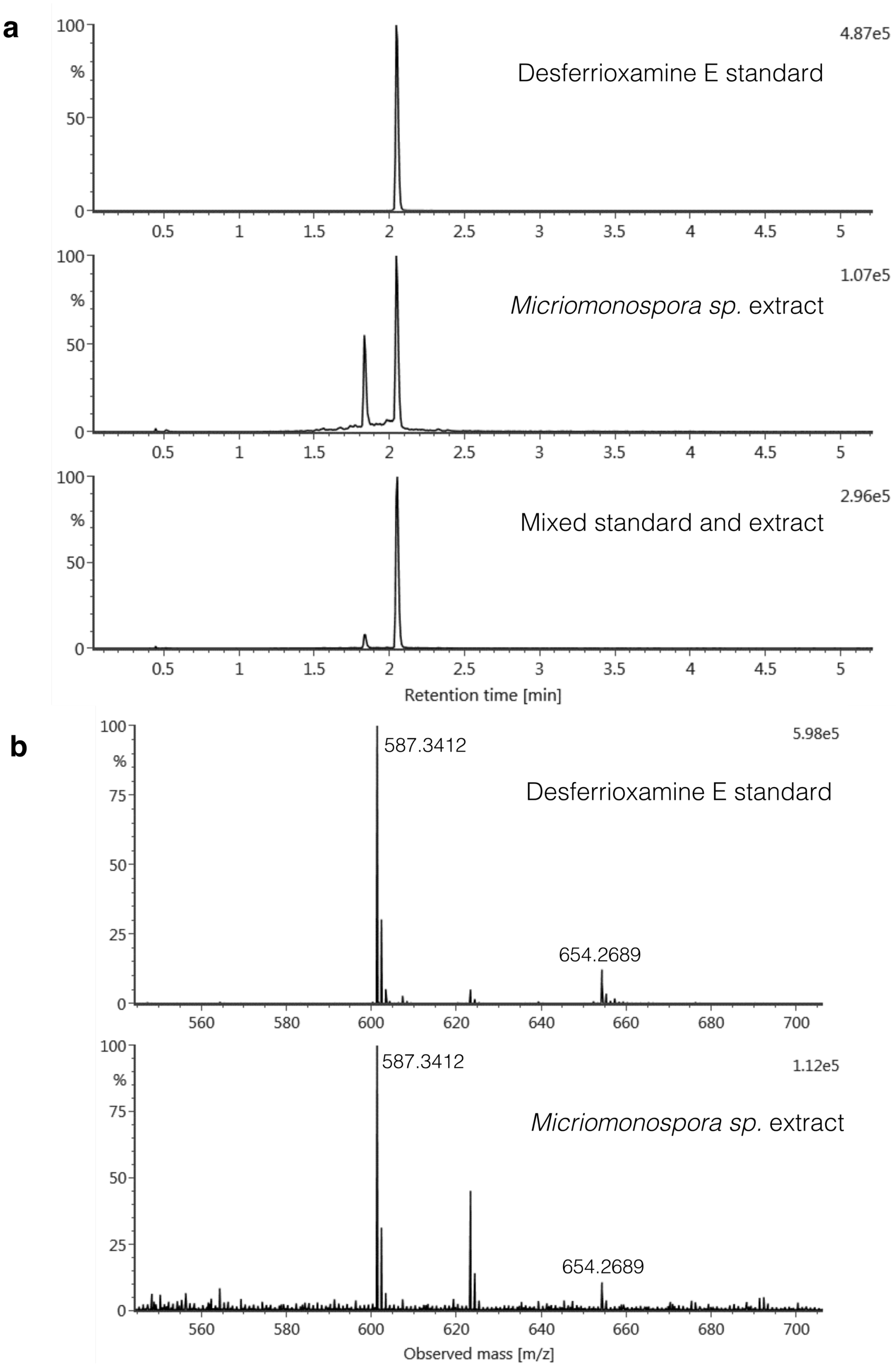


Figure A10. Desferrioxamine E Standard Co-Injection and MS

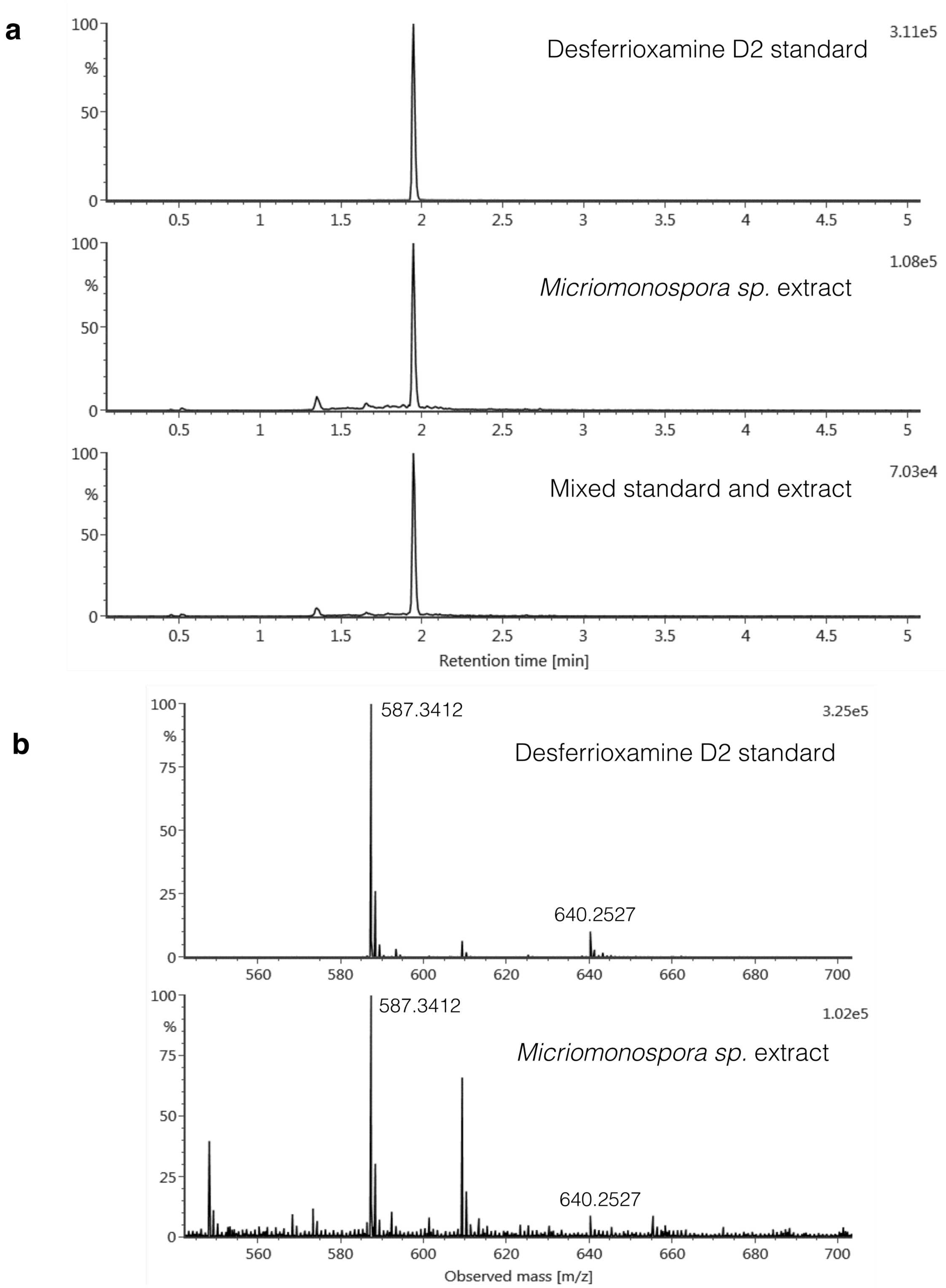


Figure A11. Desferrioxamine D2 Standard Co-Injection and MS

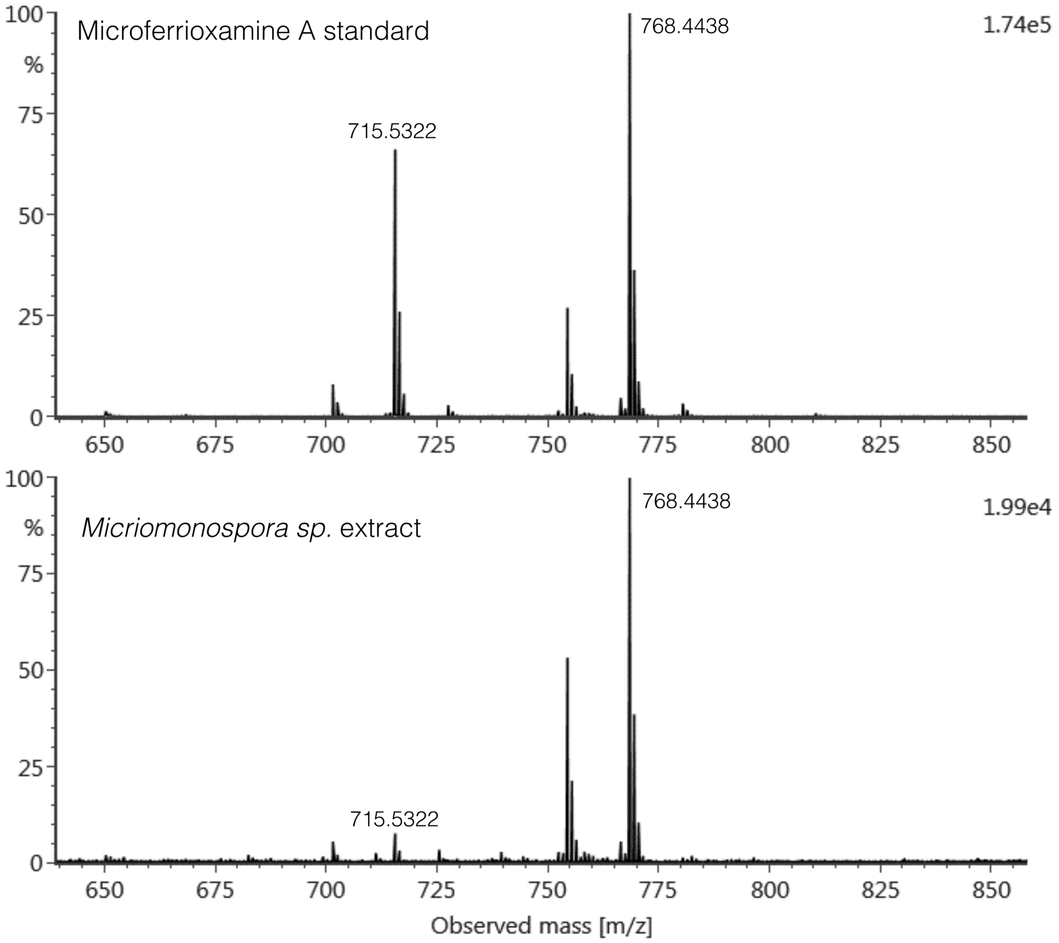


Figure A12. Microferrioxamine A Standard and extract MS

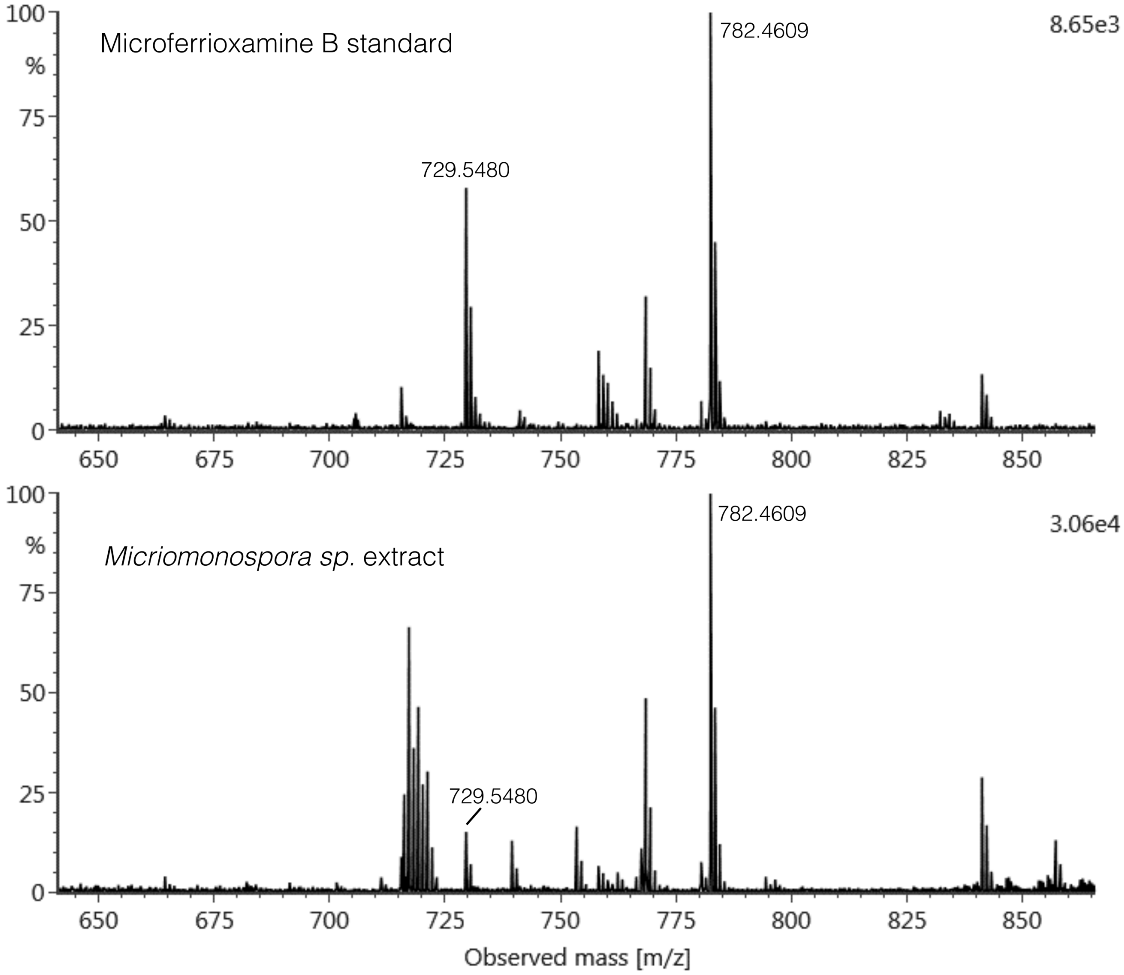


Figure A13. Microferrioxamine B Standard and extract MS

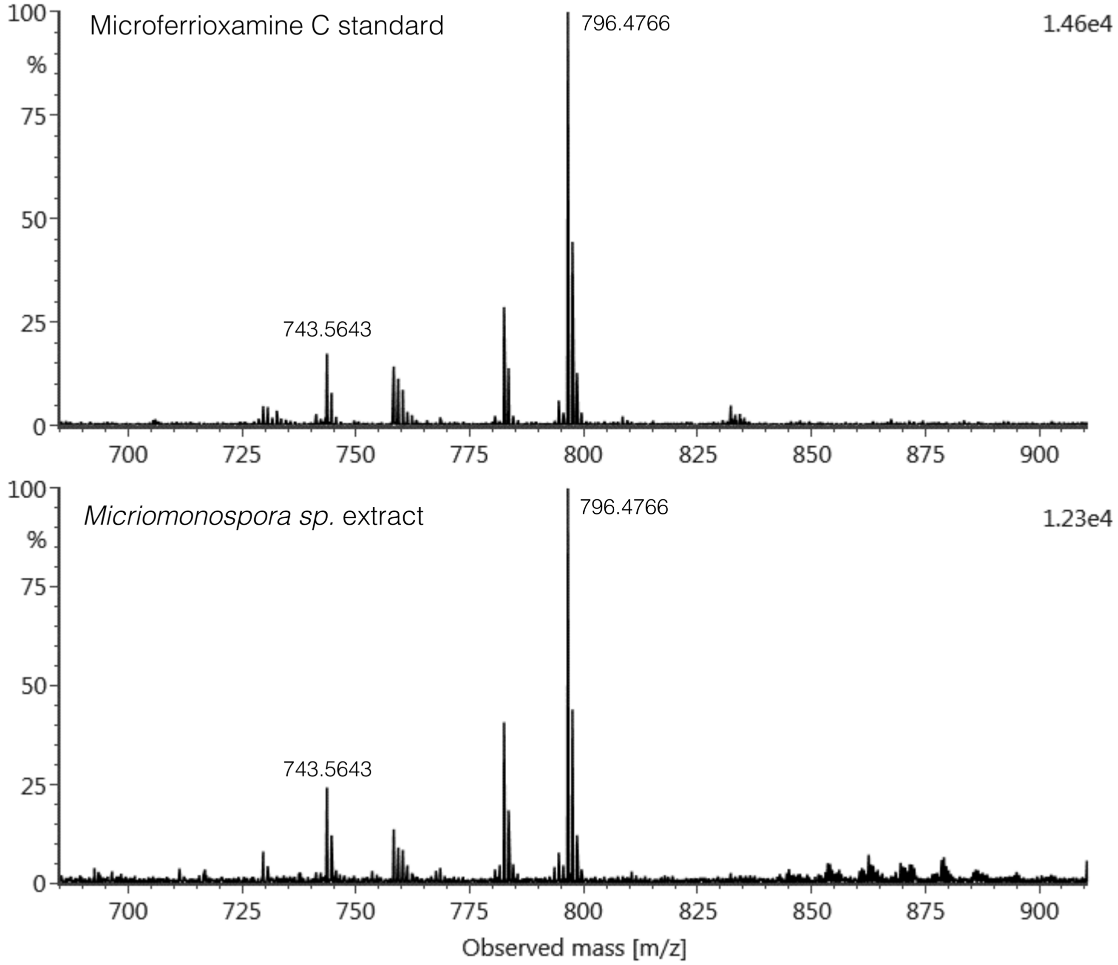


Figure A14. Microferrioxamine C Standard and extract MS

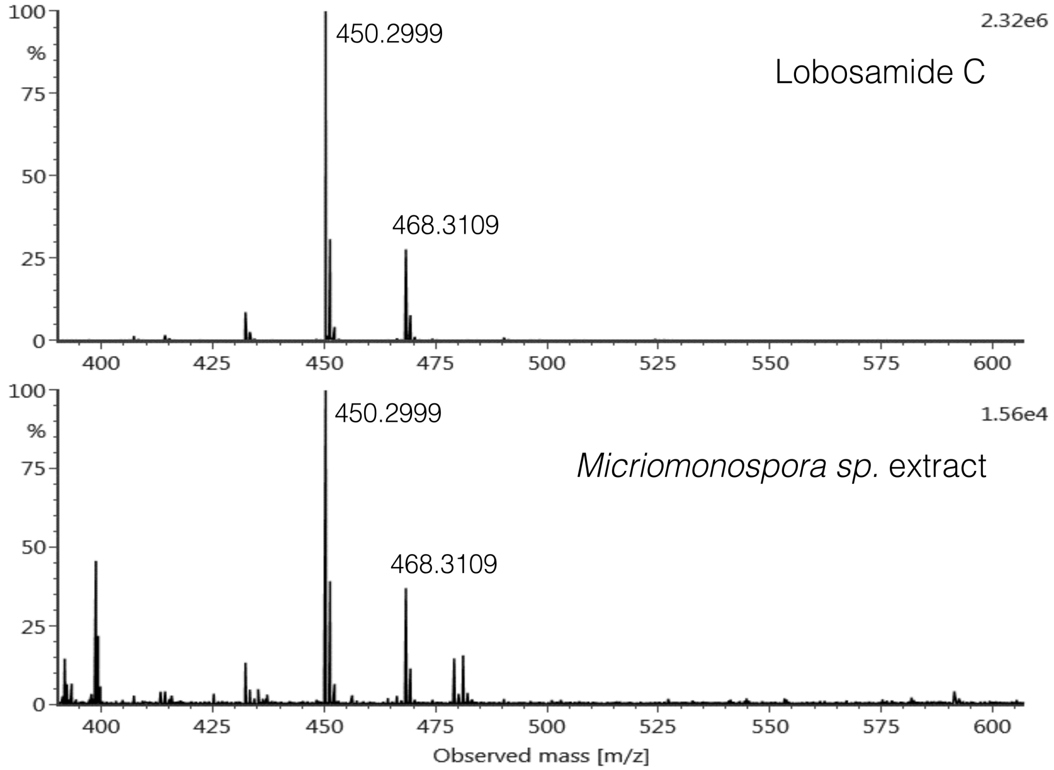


Figure A15. Lobosamide C Standard and extract MS

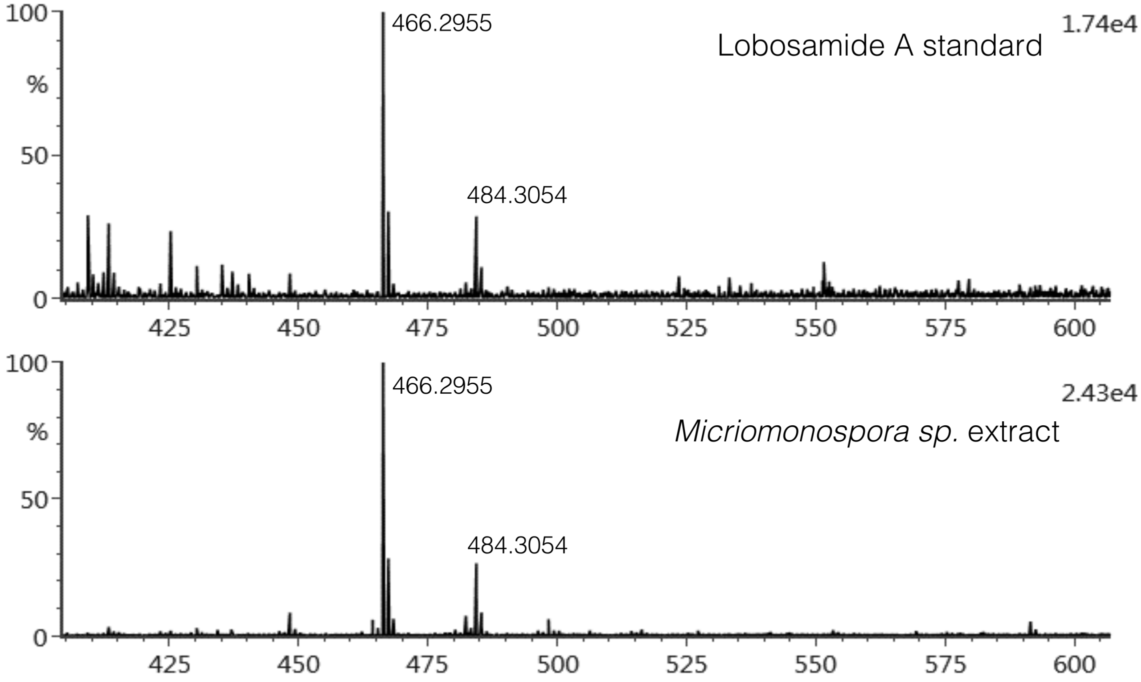


Figure A16. Lobosamide A Standard and extract MS

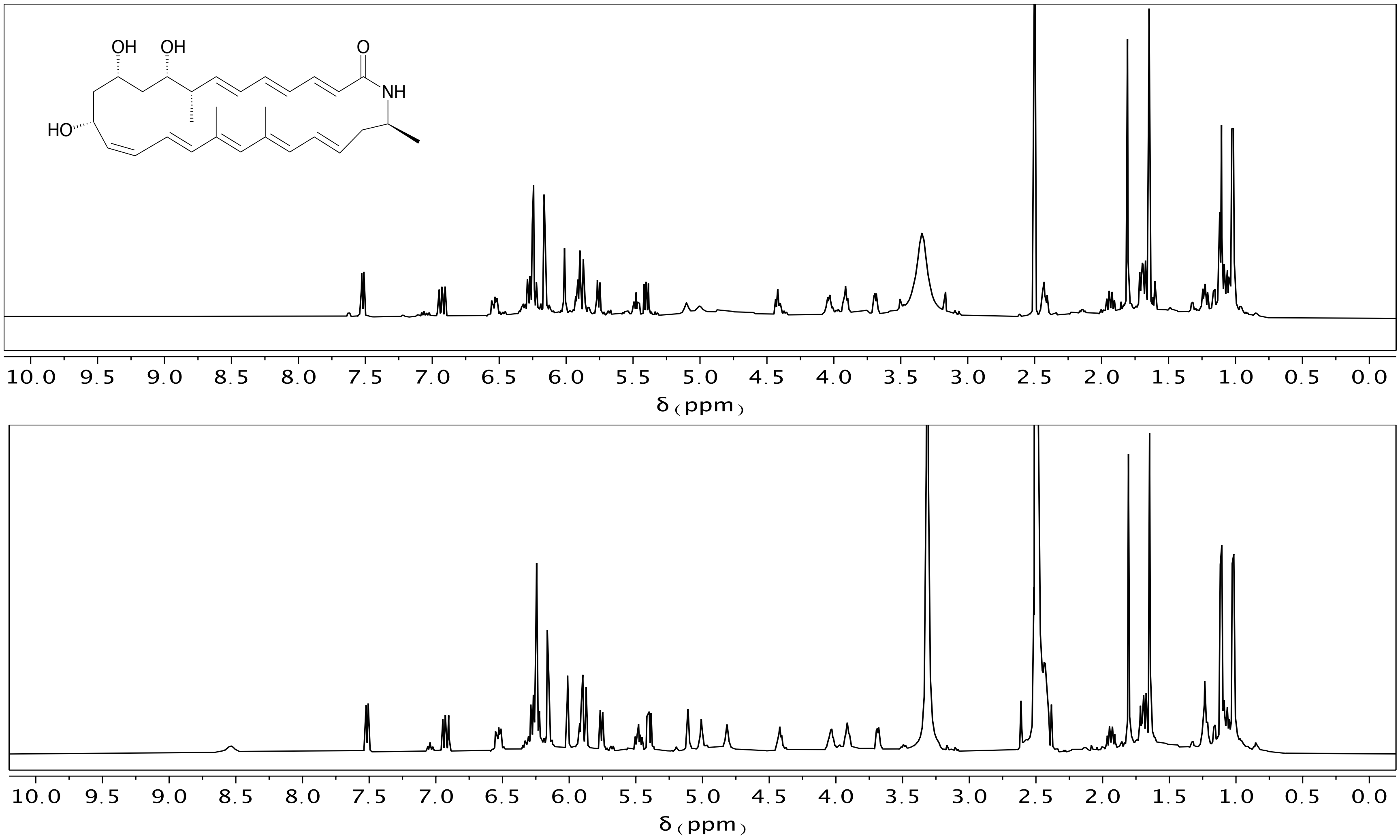


Figure A17. **^{1}H NMR Spectra of Lobosamide C acquired in DMSO-*d_6* at 600MHz**
Standard (top) and isolated compound (bottom)

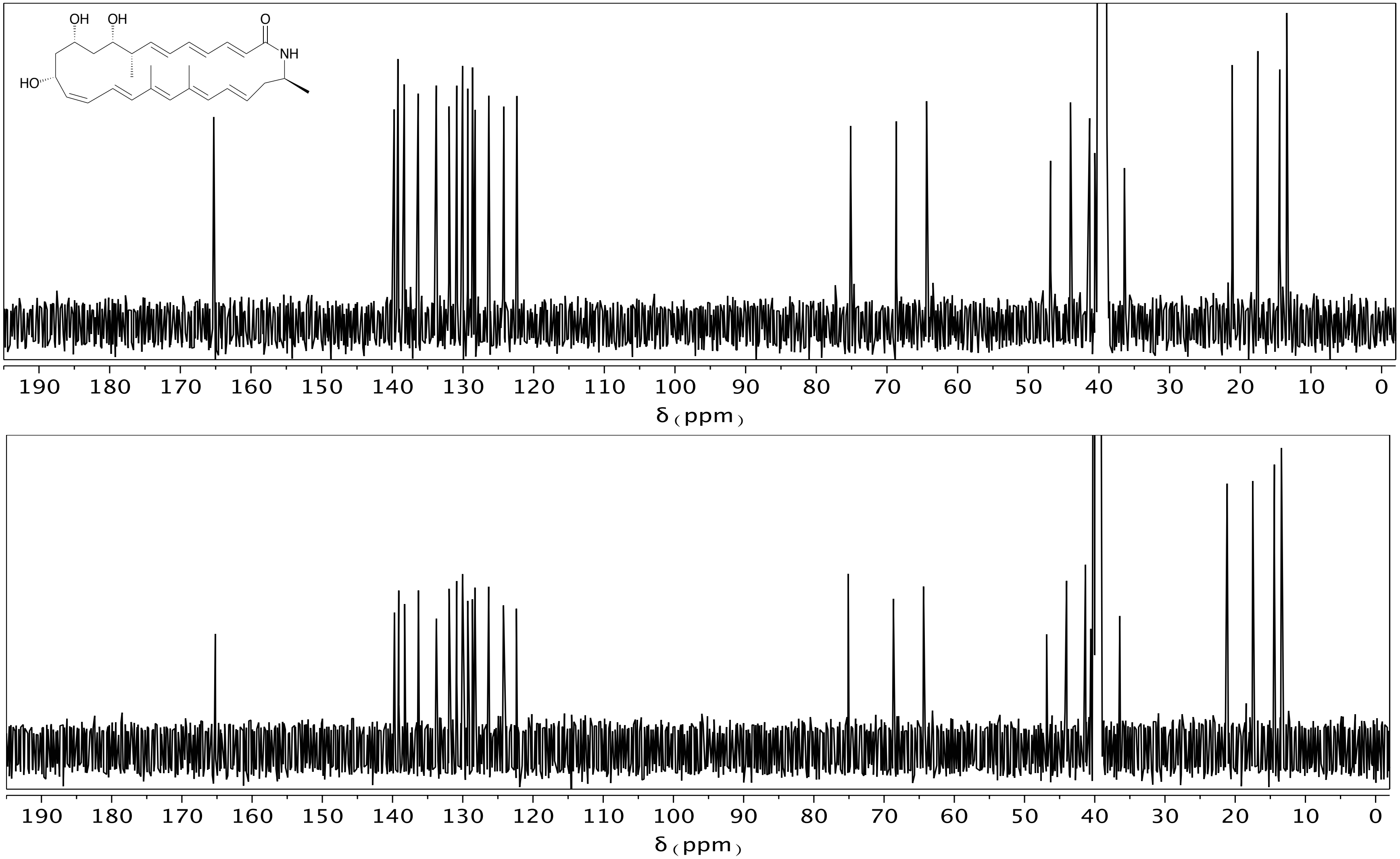


Figure A18. **^{13}C NMR Spectra of Lobosamide C acquired in DMSO-*d_6* at 150MHz**
Standard (top) and isolated compound (bottom)

Appendix B.

MS and NMR Data for Lobosamide D

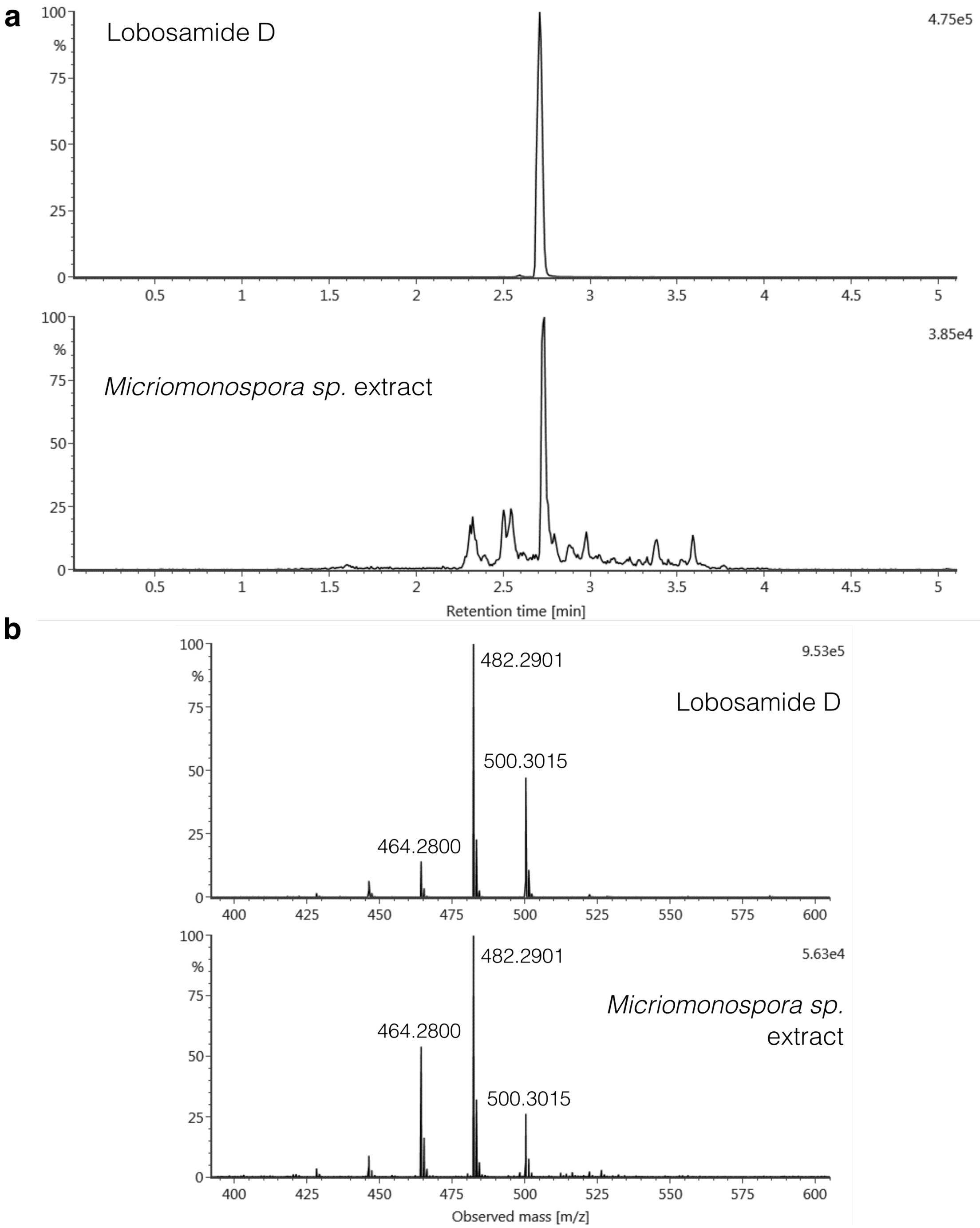


Figure B1. Lobosamide D Chromatogram and MS

Table B1. NMR Signals for Lobosamide D

Position	δ H (ppm)	m	*J*(Hz)	δ C
1				172.2
2	3.10	d	13.0	44.1
3	5.75	d	9.8	121.1
4	5.32	m		132.3
5	2.88	d	9.8	52.6
6	5.10	dd	9.8,14.6	127.5
7	5.47	dd		134.4
8	2.29	m		37.2
9	2.92	dd	8.0,8.0	74.4
10	3.35			75.7
11	3.97	m		70.9
12a	1.23	m		40.4
12b	1.62	m		
13	4.61	m		68.8
14	5.48	dd		138.1
15	5.84	dd	10.1,10.1	125.2
16	6.16	dd	10.1,15.7	122.3
17	6.07	d	15.7	137.9
18				134.9
19	6.10	s		136.9
20				38.9
21	2.09	dd	9.3, 13.0	53.2
22	3.62	dd	6.6, 9.3	65.5
23	3.96	m		73.5
24a	1.58	m		45.0
24b	2.42	m		
25	3.71	m		47.1
26	0.80	d	6.7	12.2
27	1.76	s		16.1
28	1.17	s		19.4
29	1.16	d		21.0
OH(9)	4.35			
OH(10)	4.34			
OH(11)	4.16			
OH(13)	4.60			
OH(23)	5.30			

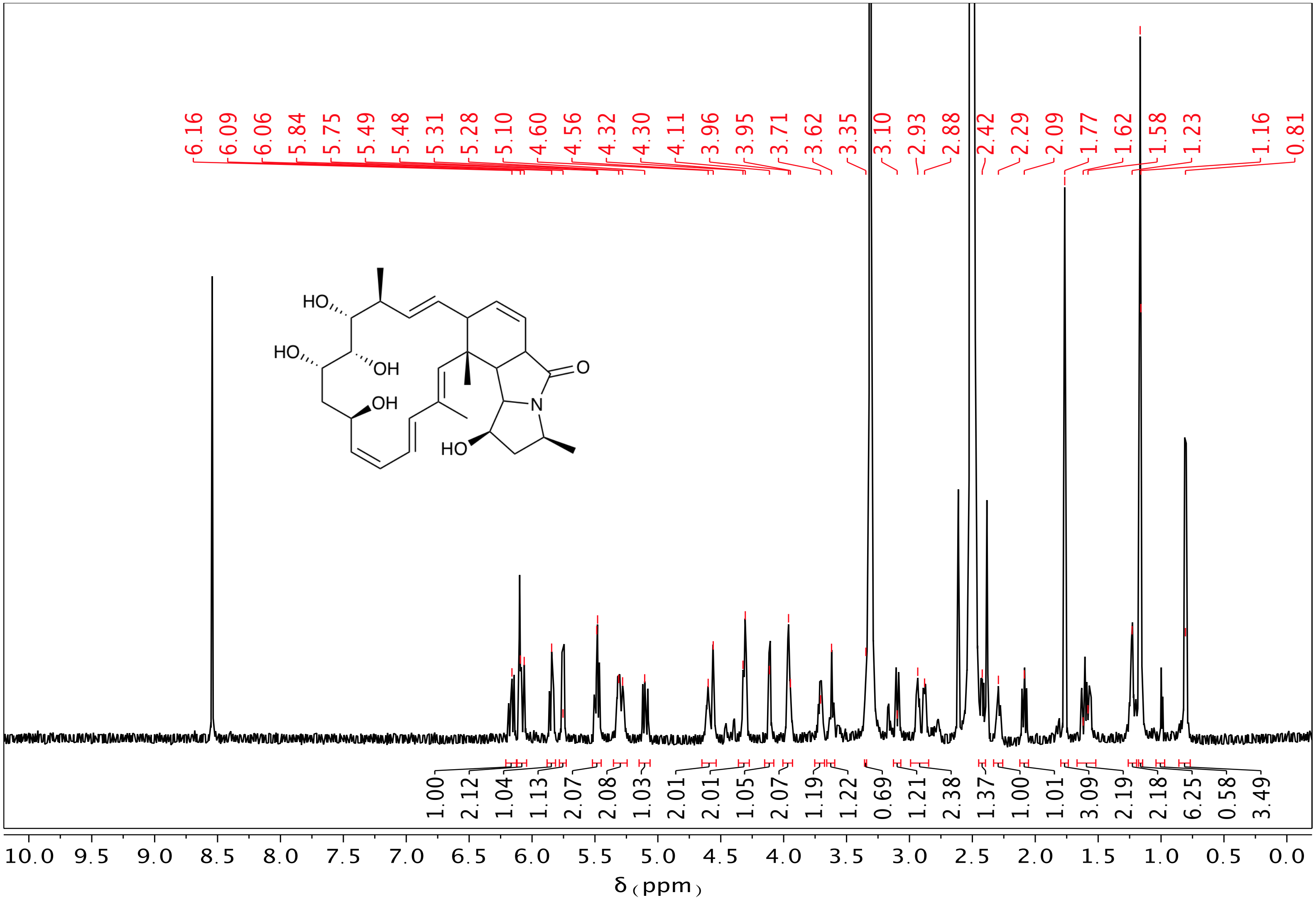
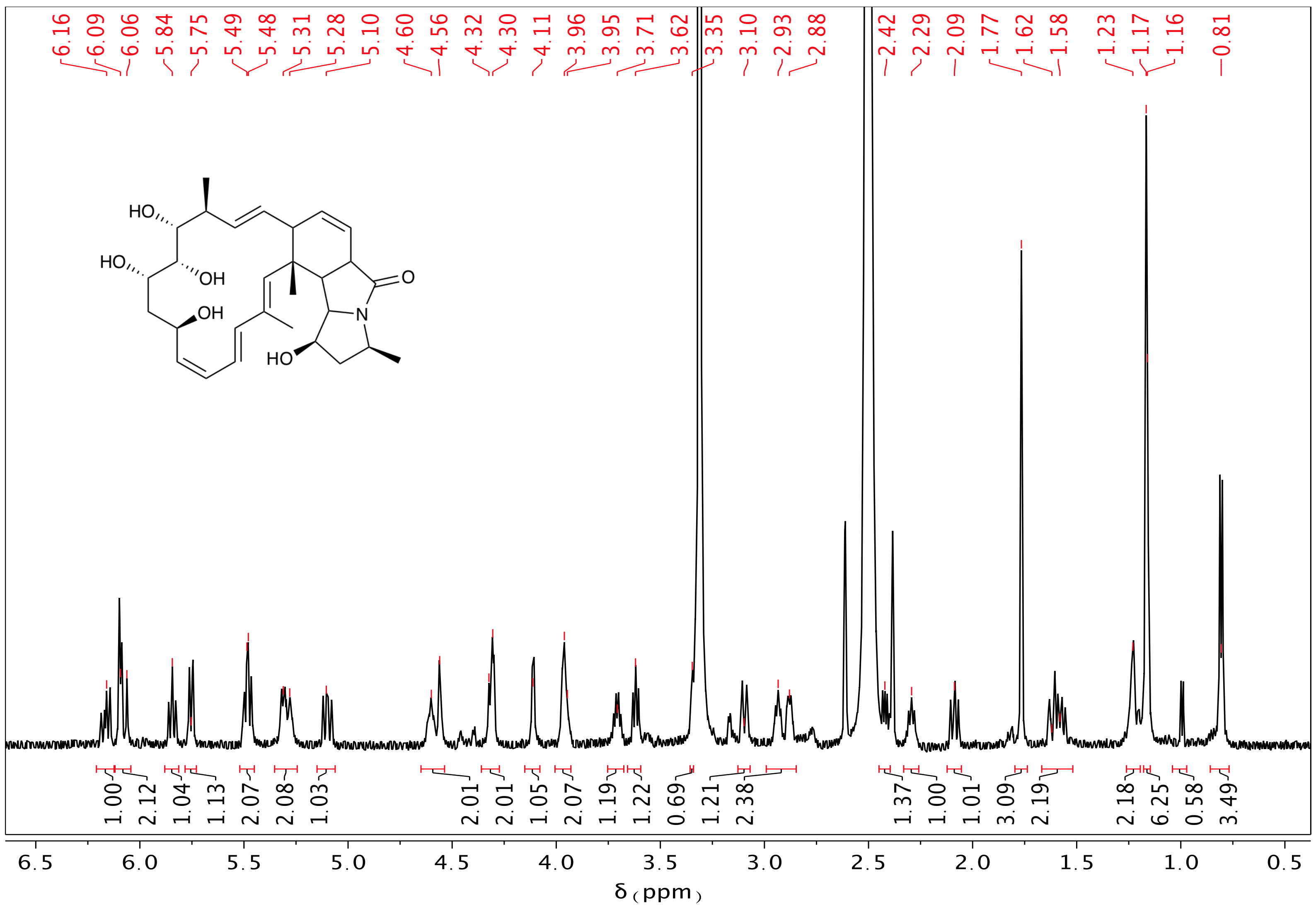


Figure B2. ^{1}H NMR Spectrum of Lobosamide D acquired in DMSO-d_6 at 600MHz



Expansion of Figure B2

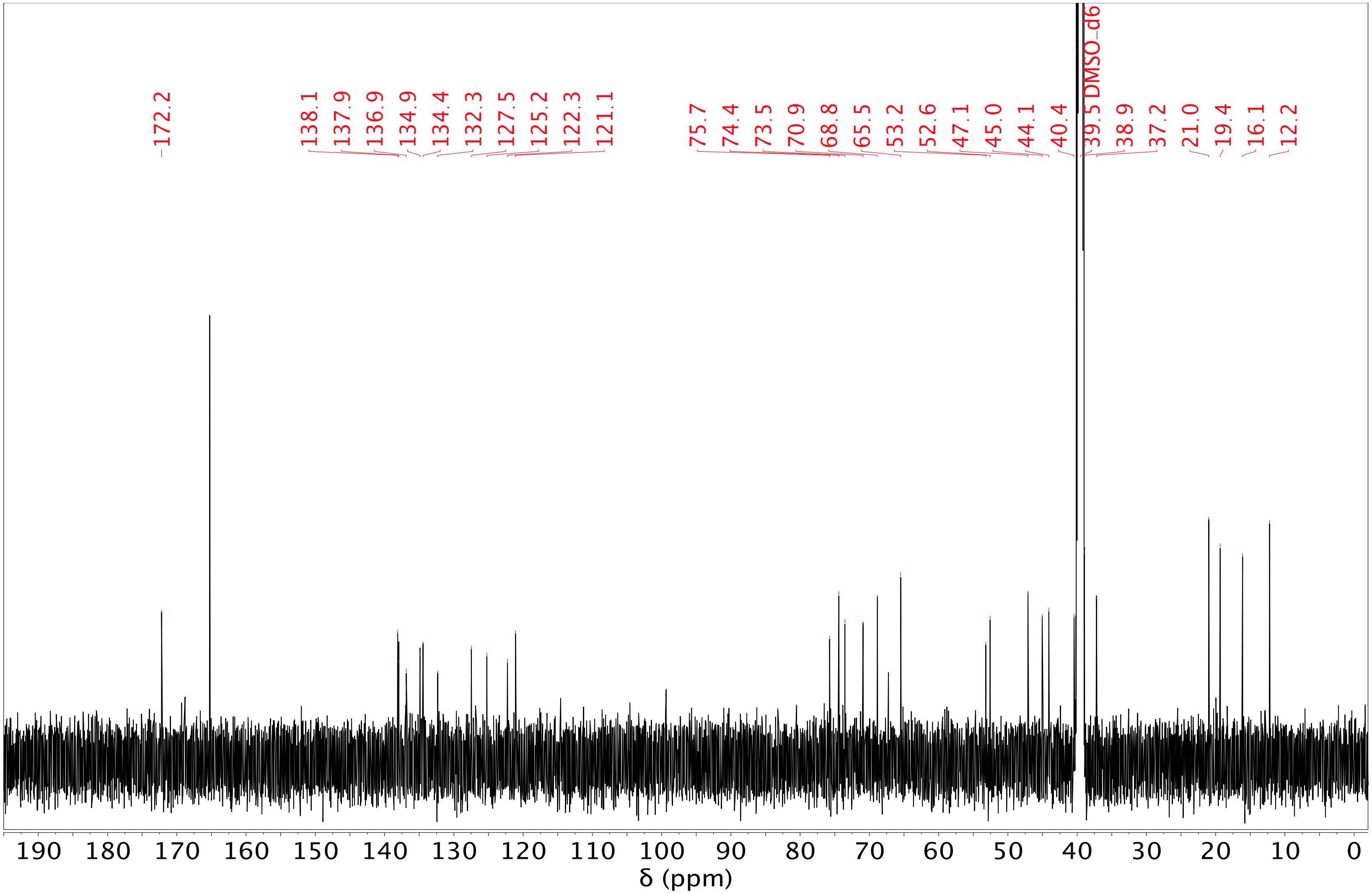


Figure B3. 13C NMR Spectrum of Lobosamide D acquired in DMSO-d_6 at 150MHz

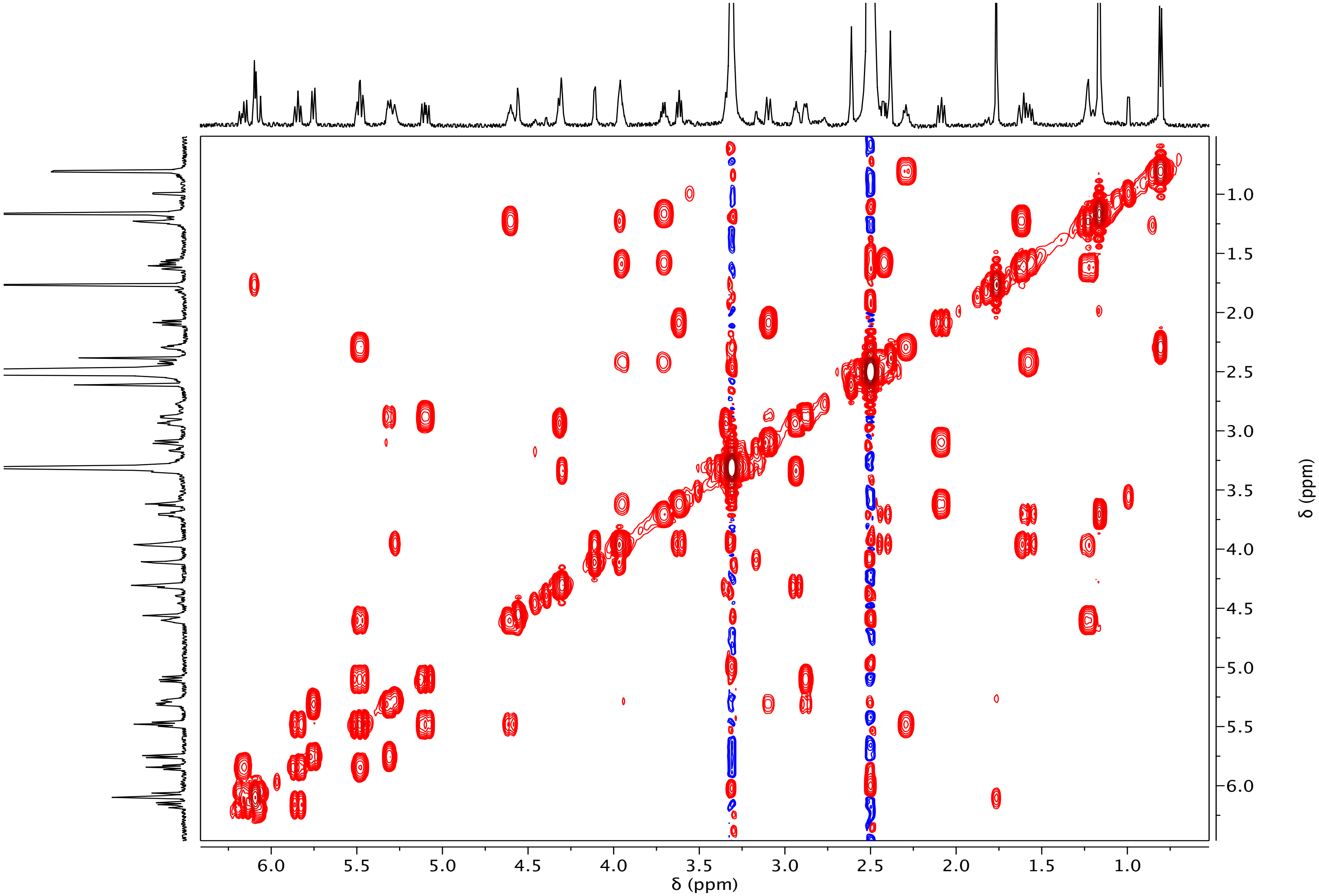


Figure B4. COSY of Lobosamide D in DMSO-d_6 at 600MHz

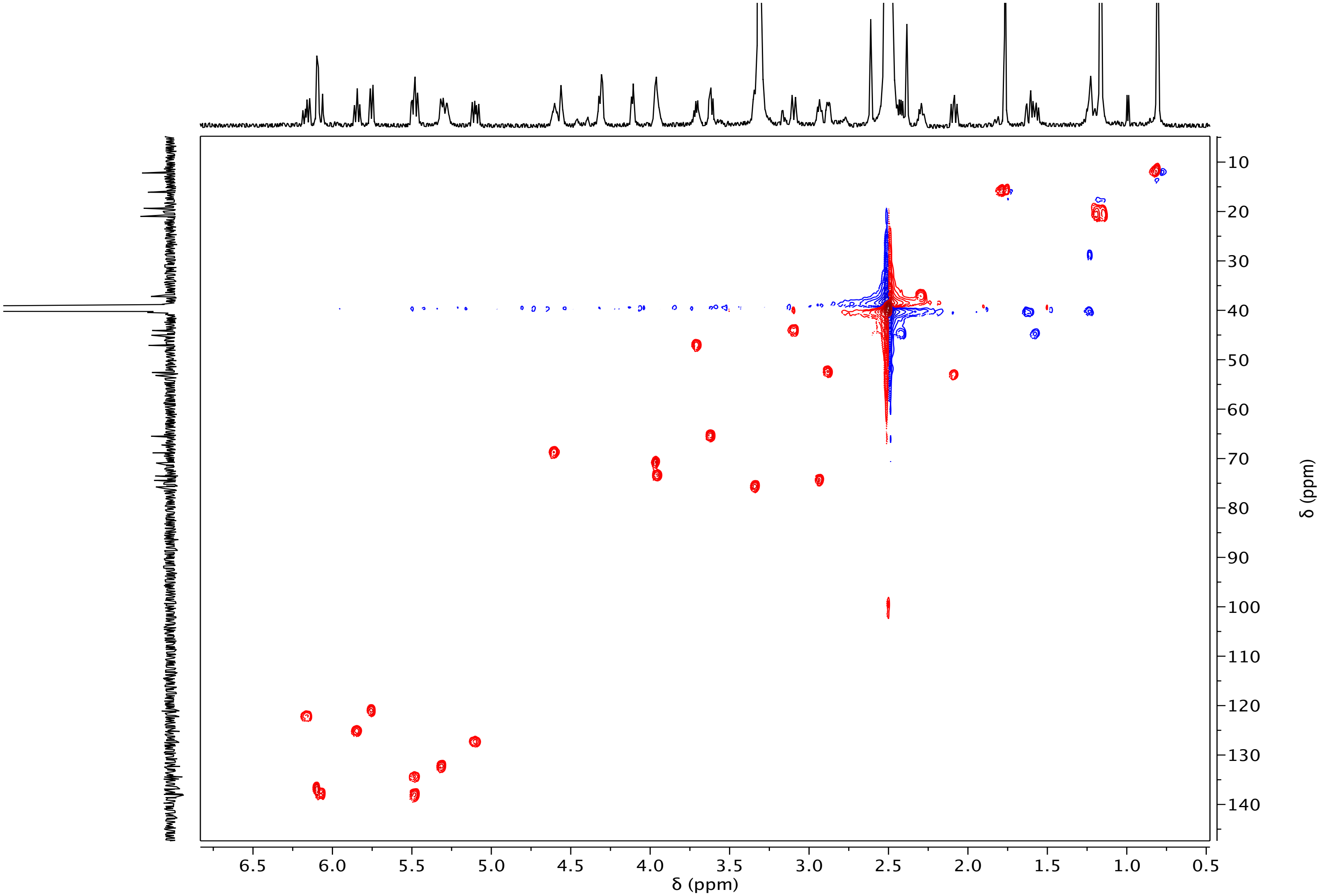


Figure B5. HSQC of Lobosamide D in DMSO-d_6 at 600MHz

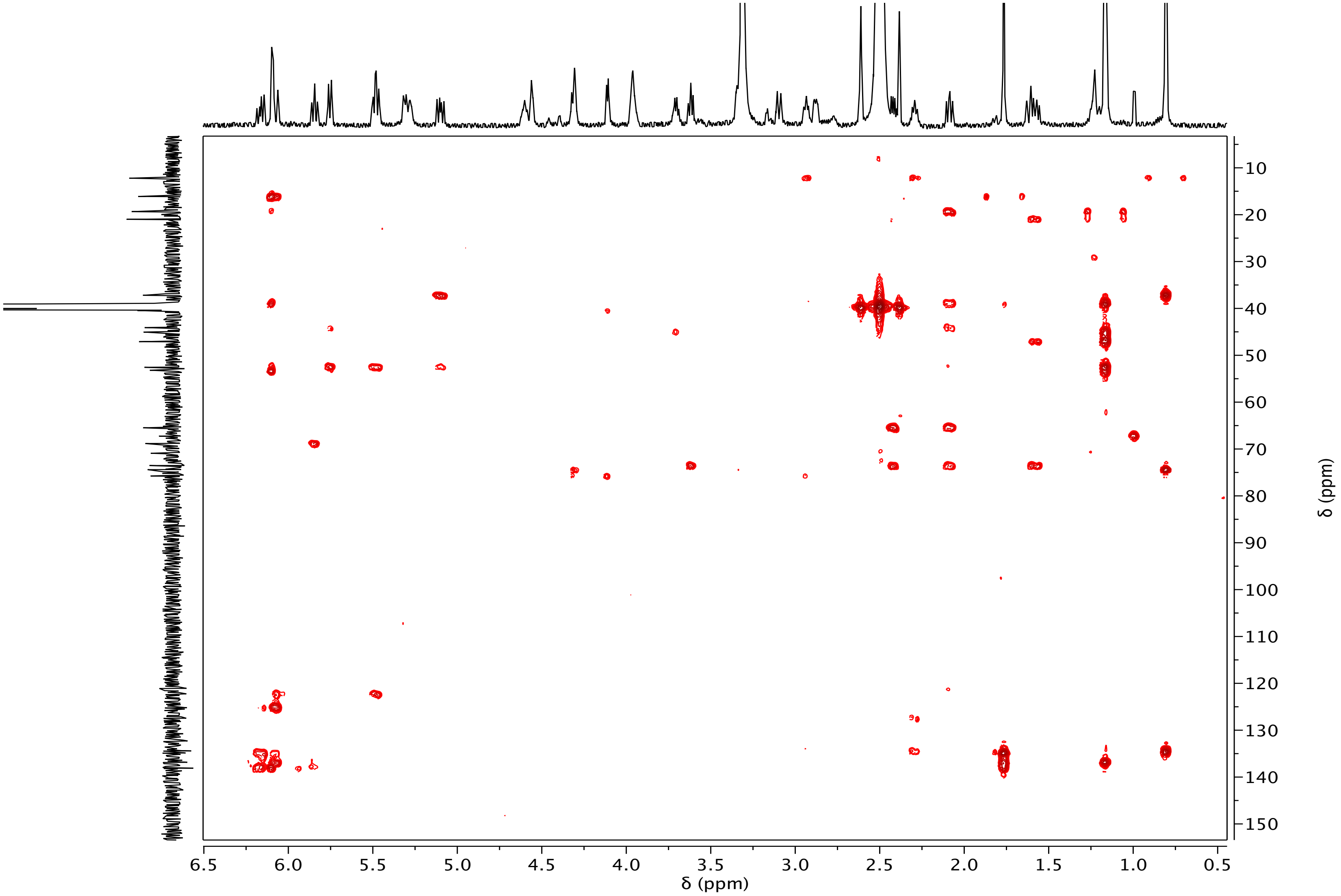


Figure B6. HMBC of Lobosamide D in DMSO-d_6 at 600MHz

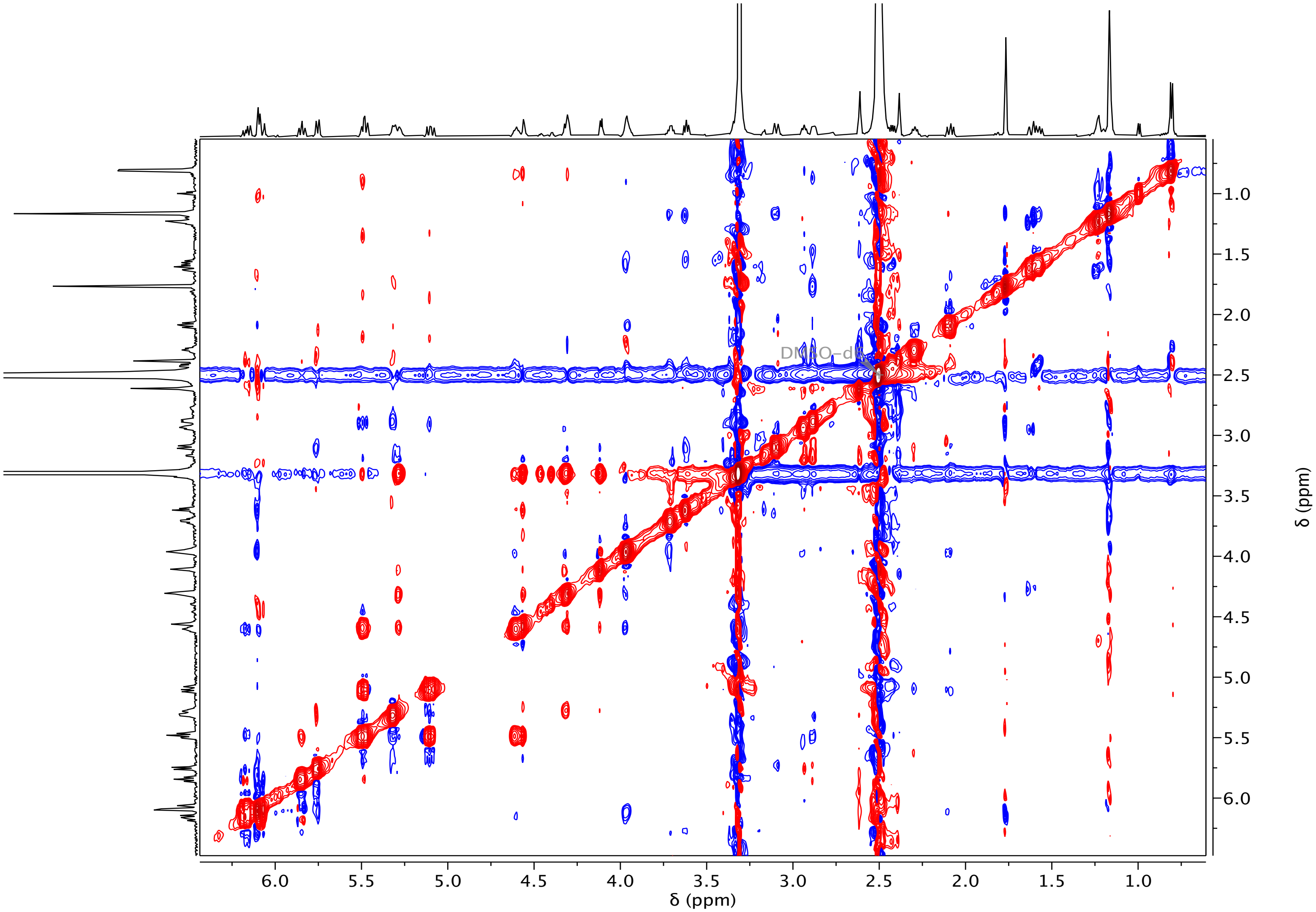


Figure B7. ROESY of Lobosamide D in DMSO-d_6 at 600MHz

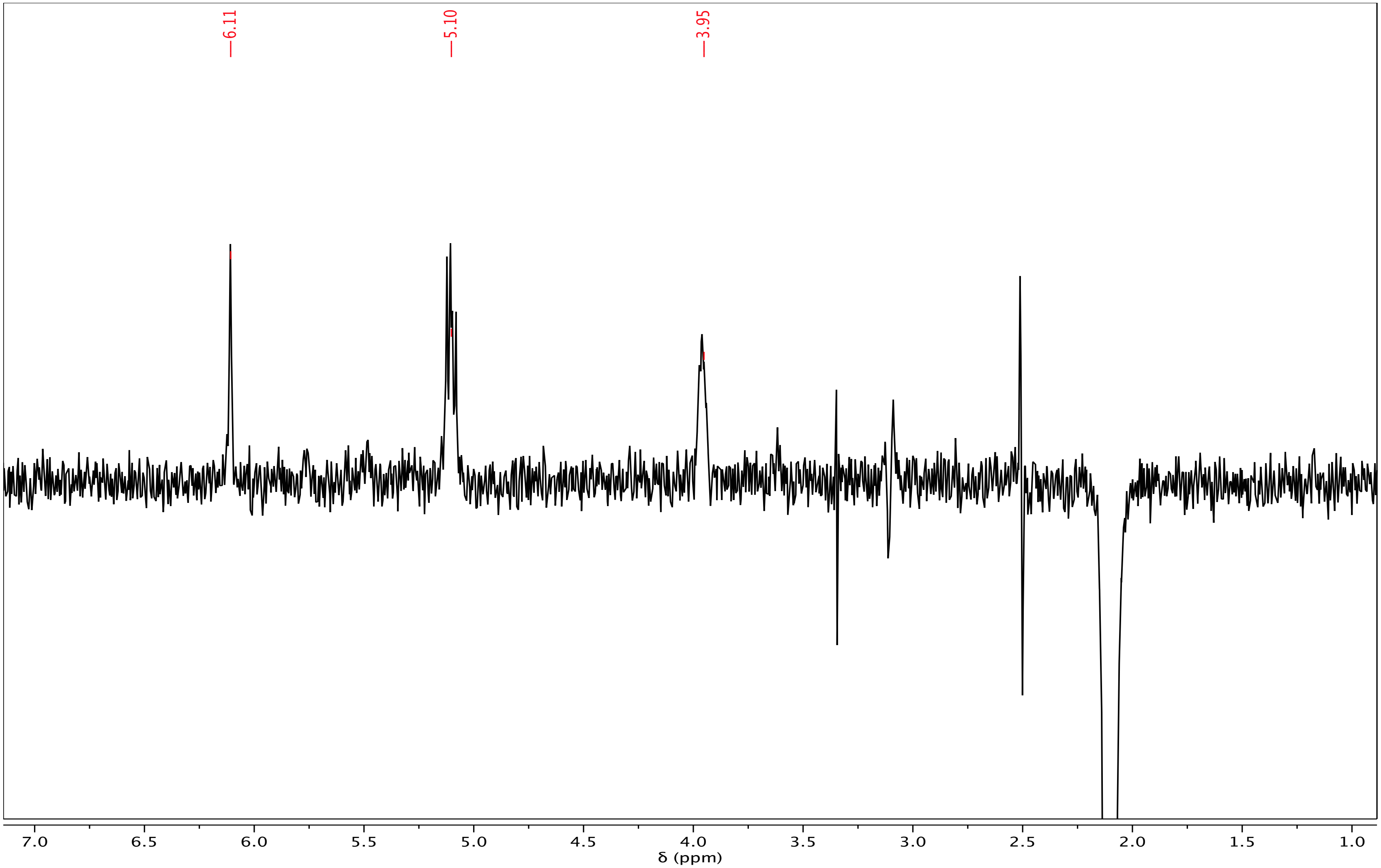


Figure B8. Selective 1D ROESY for H21 of (2.09 ppm) in DMSO-d_6 at 600MHz

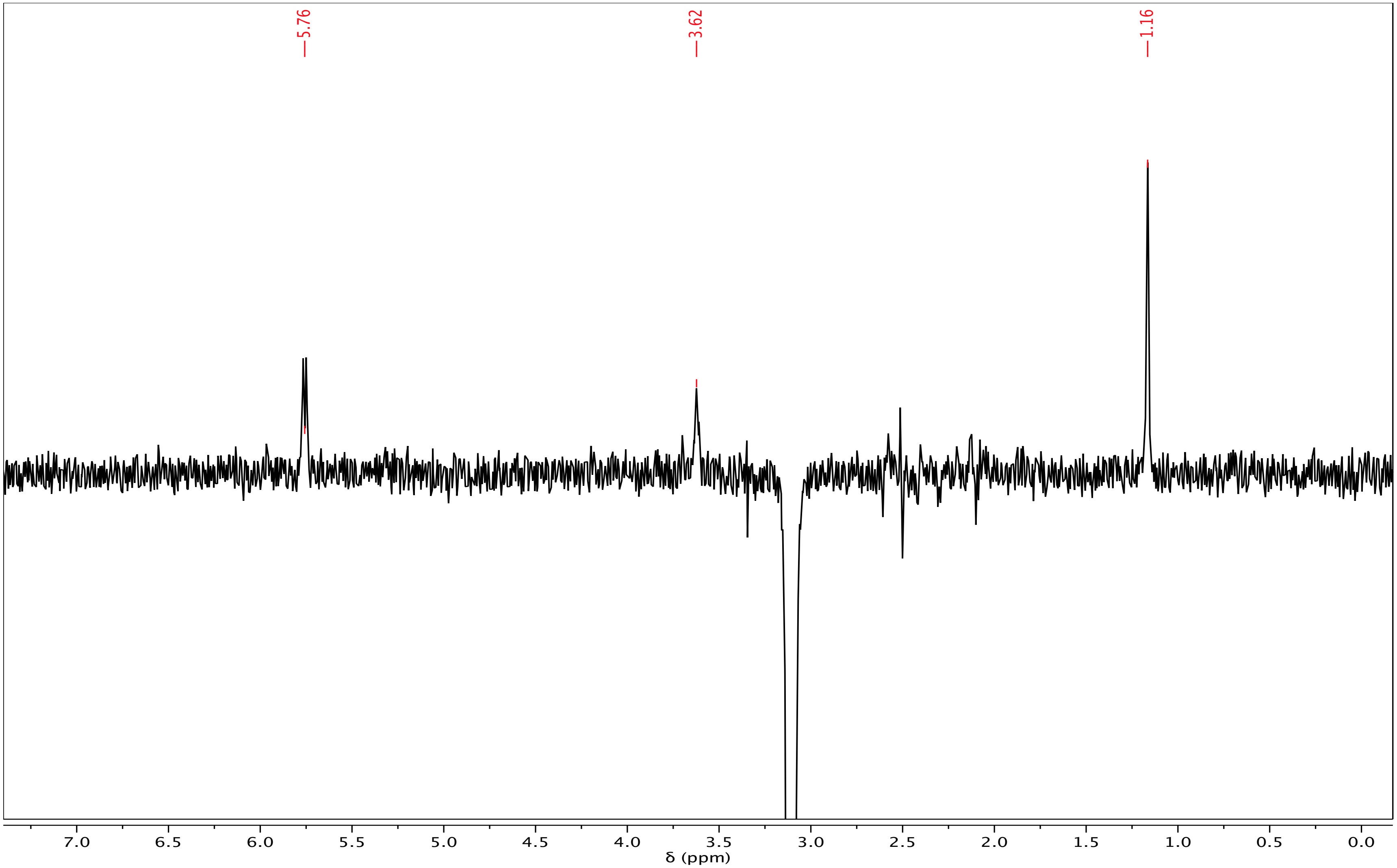


Figure B9. Selective 1D ROESY for H2 of (3.10 ppm) in DMSO-d_6 at 600MHz

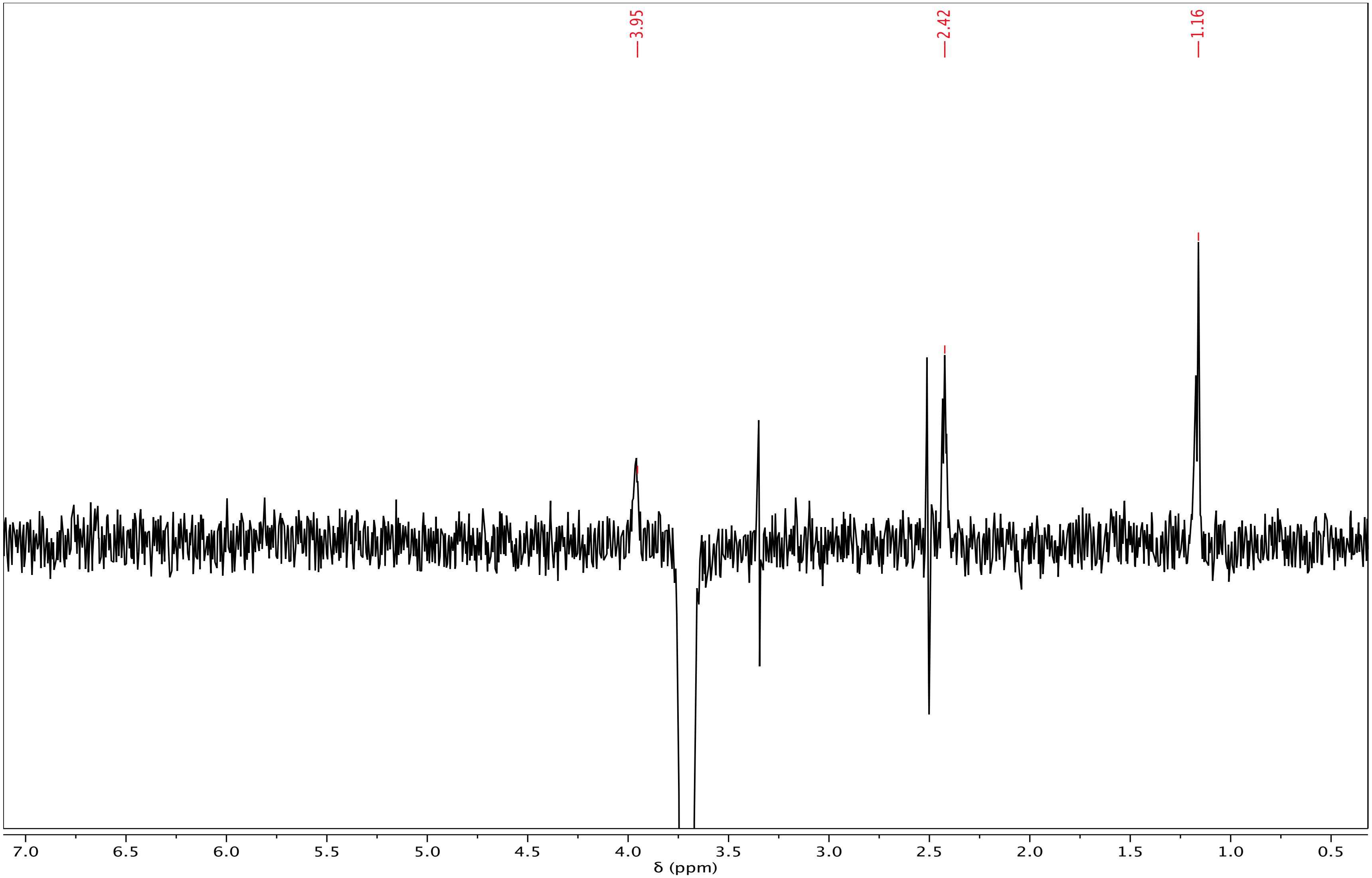


Figure B10. Selective 1D ROESY for H25 of (3.71 ppm) in DMSO-*d*$_6$ at 600MHz