Functional Annotation of Natural Product Extracts Through Integration of Orthogonal NMR Datasets

by

Joseph M. Egan

Master of Science, UNC Greensboro, 2016

Bachelor of Science, UNC Greensboro, 2014

Thesis to be 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:	**Joseph M. Egan**
Degree:	**Doctor of Philosophy**
Title:	**Functional Annotation of Natural Product Extracts Through Integration of Orthogonal NMR Datasets**
Committee:	**Chair:** **Paul C. H. Li** Professor, Chemistry
	Roger G. Linington Supervisor Professor, Chemistry
	Vance Williams Committee Member Professor, Chemistry
	Robert Britton Committee Member Professor, Chemistry
	Charles Walsby Examiner Associate Professor, Chemistry
	Anthony Carroll External Examiner Professor, School of Environment and Science - Ecology and Evolution Griffith University

Abstract

Natural products have provided humanity with powerful tools for human health, including combatting infections, curing diseases, and helping humans to understand the world around us. However, natural product discovery is complicated by the challenges in robust annotation and sample comparison from complex samples. New utilities which integrate orthogonal experimental data together could better describe the constitution of natural product extracts, allowing for functional annotation of individual components and streamlining the discovery process. MADByTE, an NMR processing platform designed for metabolomics and dereplication using 2D NMR data was designed to integrate data from HSQC and TOCSY experiments to create contextual networks to annotate structural characteristics of unknowns in complex samples. Addition of bioactivity profiling to the MADByTE network allows for prediction and targeted isolation of bioactive constituents, demonstrated through the isolation of collismycin A from an actinobacterial extract. When coupled to a molecular recognition platform (SMART) and an NMR prediction utility (NMRShiftDB2), substructures of molecules within complex samples can be proposed based on similarities in spectral profiles. Integration of MADByTE with DOSY experiments allows for the refinement of features based on molecular descriptors such as diffusion rates. Taken together, MADByTE represents a valuable utility for the untargeted analysis of natural products contained in complex samples and provides a new viewpoint of chemical diversity across an extract library.

Keywords: Metabolomics; 2D-NMR; Natural Products; Discovery; MADByTE

Dedication

To my amazing wife, Anna.

For reminding me how magical the world is daily and embracing every adventure.

Whose love and support I am eternally grateful for.

To my parents, Mike and Shannon.

For entertaining my independence, my inquisitiveness, and supporting me in every venture I've made known, and whose selflessness and ingenuity I admire.

To my inspiring sister, Corey.

For teaching me the world is full of creativity and adventure.

Whom I will always look up to.

To my Grandparents; Chuck, Sheila, Virginia, and Joe.

Whose stories and wisdom I carry with me every day to give me strength.

To my aunts, uncles, and cousins.

For showing me that love and support cost nothing but can mean everything.

To my amazing friends, scattered across the globe.

Words will never be enough, but I won't stop talking.

I owe you all a beer.

And to Fleetwood, Mac, Sherlock, and Bella.

None of whom can read.

Acknowledgements

This dissertation contains text and figures from:

“Development of an NMR-Based Platform for the Direct Structural Annotation of Complex Natural Products Mixtures.” Joseph M. Egan, Jeffrey A. van Santen, Dennis Y. Liu, and Roger G. Linington. *Journal of Natural Products* **2021** *84* (4), 1044-1055

Roger G. Linington directed and supervised the research which forms this dissertation.

This work was supported by NIH U41-AT008718, NSERC Discovery, and NIH F31-AT010098, and the SFU Chemistry Department.

I thank Drs. A. Lewis and E. Ye for assistance with NMR experiment selection and data acquisition. Thanks to Bruker and C. Anklin for advice, support, and training. Thanks to the Gerwick lab, the Cottrell lab, S. Kuhn, R. Moreira Borges, G. Kleks, and A. Carroll for support and helpful discussions on the development of MADByTE modules. Thanks to the MADByTE alpha testers and early adopters; Z. Al Subeh, L. Flores-Bocanegra, N. Oberlies, G. Peterson, C. Fergusson, T. Clark, T. Bergeron, R. Reher, K. B. Kang, and S. Carlson.

Table of Contents

List of Tables

List of Figures

List of Acronyms

API	Application Programming Interface
ATCC	American Type Culture Collection
BHI	Brain Hearth Infusion
BioMAP	Antibiotic Mode of Action Profile
BPP-LED	Bipolar Pulse Longitudinal Eddy Current Delay
CDCl3	Deuterated Chloroform
CFU	Colony Forming Units
CLSI	Clinical And Laboratory Standards Institute
COLMAR	Complex Mixture Analysis By NMR
COSY	Correlation Spectroscopy
COSY-IDOSY	COSY Integrated into Diffusion Ordered Spectroscopy
CRAFT NMR	Complete Reduction to Amplitude Frequency Table Analysis of Nuclear Magnetic Resonance
DEPT	Distortionless Enhancement by Polarization Transfer
DFT	Density Functional Theory
DMSO	Dimethylsulfoxide
DMSO-d_6	Deuterated Dimethylsulfoxide
DOSY	Diffusion Ordered Spectroscopy
FADES	Feature Association by Diffusion Experiments
FDA	United States Food and Drug Administration
GNPS	Global Natural Products Social Molecular Networking
GUI	Graphic User Interface
HiFSA	^{1}H Iterative Full Spin Analysis
HMBC	Heteronuclear Multiple Bond Correlation
HOSE	Hierarchical Organization of Spherical Environments
HSQC	Heteronuclear Single Quantum Coherence
HSQC-TOCSY	Heteronuclear Single Quantum Coherence - Total Correlation Spectroscopy
HSQC-IDOSY	HSQC Integrated into Diffusion Ordered Spectroscopy
HTM	Haemophilus Test Medium

ID	Identifier
INADEQUATE	Incredible Natural Abundance Double Quantum Transfer Experiment
INEPT	Insensitive Nuclei Enhanced by Polarization Transfer
LC-MS	Liquid Chromatography Coupled Mass Spectrometry
MADByTE	Metabolomics and Dereplication By Two-Dimensional Experiments
MeOD	Deuterated Methanol
Mi-BIG	Minimum Information About a Biosynthetic Gene Cluster Database
MIC_{50}	Minimum Inhibitory Concentration Required to Inhibit 50% of Growth
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOA	Mechanism of Action
MS	Mass Spectrometry
MS^2	Tandem Mass Spectrometry
MW	Molecular Weight
NB	Nutrient Broth
NMR	Nuclear Magnetic Resonance
NP	Natural Products
NPAtlas	The Natural Products Atlas
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
PLS	Partial Least Squares
qNMR	Quantitative Nuclear Magnetic Resonance
RF	Radio Frequency
SHIMS	Substructure Hypothesis by Integration of MADByTE and SMART
SMART	Small Molecule Recognition Technology
SMILES	Simplified Molecular-Input Line-Entry System
SSF	Spin System Feature
STOCSY	Statistical Total Correlation Spectroscopy
TOCSY	Total Correlation Spectroscopy

TSB	Tryptic Soy Broth
UPLC-MS	Ultrahigh Pressure Liquid Chromatography Coupled Mass Spectrometry
UV	Ultraviolet
WHO	World Health Organization

Chapter 1.

Metabolomics and Natural Products

Natural products are defined as secondary metabolites from organisms that do not play a role in basic metabolism but are often produced to help organisms compete for resources, survive in a niche environment, or directly combat other organisms that threaten their survival. Secondary metabolites are compounds produced and used by an organism that do not play an active roll in metabolism and can often be organism specific. Since the dawn of civilization, humans have used these organisms in various ways to adapt and thrive in their surroundings, but only through the last two centuries have we been able to directly isolate and observe the effects of these secondary metabolites on human pathogens in a controlled manner. Throughout the 20th century, natural product chemistry was a cornerstone in the pharmaceutical market, serving as the source or point of inspiration for over 39% of approved small molecule drugs by the FDA over the last 40 years, and expanding to ~60% if natural product inspired molecules are included.[1]

Investigation of these molecules is not without its challenges, namely complications from the source material itself and the variable abundance of potent molecules among the complex background of primary metabolites and known products that are not of interest in a particular study. These complications, and the time-consuming nature of classical natural product investigations, have led to a massive reduction in natural product drug discovery platforms adopted in the pharmaceutical industry. However, in recent years, new approaches to the investigation and characterization of these important molecules have created a new era of technological development and instrumentation capable of profiling hundreds of metabolites simultaneously that are now not only available, but are readily accessible to many institutions across the globe. This has created a massive communal interest in the investigation of natural products by combining new driven tools and advanced instrumentation.

Broadly defined, the term metabolomics refers to the detection, identification and, in some cases, quantification of metabolites produced by an organism.[2] In comparative metabolomics, each analysis compares the metabolic profile of one sample against the context of a much larger subset of samples. These investigations have been heavily

utilized in the field of primary metabolism where inherent changes in the profile of many primary metabolites can be linked to a biological perturbation or effect in an organism. Secondary metabolomics, the detection and identification of secondary metabolites, has been a slower growing field due in large part to the complexity of the metabolome of the organisms under investigations. A large bottleneck in the development and utilization of secondary metabolomics has been the accurate identification of molecules of interest, since many secondary metabolites are still unknown and therefore have no identifying characteristics which can be tied directly to phenotype or survivability of an organism.

1.1. Natural Products and Unique Challenges

1.1.1. Natural Products and Their Sources

In chemistry, the term “natural products” refers to the collection of small molecule metabolites from living organisms which are not involved in primary metabolism, but often provide an advantage for the organism to survive a niche in its larger ecosystem.[3] These molecules, despite not being directly involved in primary metabolism, can serve a variety of functions to aid the organism in survival from predation,[4] resource sequestration,[5] or to address external pressures brought on by competing organisms.

Presumably, since life began on earth simple organisms have engaged in a type of “biological warfare” through the production of natural products to address selective pressures and to increase overall survivability of the species. The expenditure of energy needed to produce these natural products is considerable, and much like other evolutionary traits they must confer some survivability advantage upon the organism to be of use. As selective pressures change on an organism, so too does its expressed metabolome – producing certain compounds only in the presence of different conditions. As early as the discovery of penicillin, manipulation of culture media or growth conditions have provided a systematic method for the discovery of new natural products and ways to increase their yield from source organisms.[6,7]

Natural products have been isolated from a variety of sources with much of the focus on plants, bacteria, and fungi as platforms for new and important molecules for human health. Botanically derived molecules, in a way, represent the transition from folk medicine to modern medicine. Ancient remedies involving botanicals have been well documented

across many cultures and provide robust hypotheses for natural product chemists; if a given botanical or preparation alleviates a symptom, there is a high likelihood that the physiologically active components are good starting points for modern medicinal approaches. Indeed, investigations of plants relevant to ancient people's lives has become its own field of study – ethnobotany – and active components are still being described from these sources today. With the discovery of the microbial world and our increased understanding of fungi and bacteria, new opportunities presented themselves for the discovery of medicines.

1.1.2. Natural Products Importance in Modern Medicine

Small molecule natural products provide a historically significant contribution to modern medicine beyond the 19th century. Perhaps one of the most famous advancements in the last century was the discovery and use of penicillin, a natural product first described by Alexander Flemming in 1929 from a culture of *Penicillium notatum* which had grown on a *Streptococcus* culture dish, yielding a zone of inhibition where the *Streptococcus* could not survive.[6] In the 1940s, cultivation and widespread production of penicillin became so important for treating infections sustained by soldiers that researchers working to optimize production and isolation at Oxford University had spores hidden in their coats for immediate transport in the event the production facilities were bombed by Axis forces.[8] By the end of the war, production of penicillin had undergone massive transformations, drastically reducing its cost as well as increasing its widespread availability and stands as a notable achievement in modern medicine development. In this way, the allied wartime effort leveraged the biological capabilities of *Penicillium* to increase survival to selective pressures beyond itself.

Natural products have become so important to the landscape of human health that several Nobel prizes have been awarded for their discovery and subsequent use to treat human conditions, including in 1945 for the discovery of penicillin to treat infections[9] and in 2015 for the discoveries of ivermectin and artemisinin, two natural products deemed by the WHO to be essential medicines.[10] Beyond the Nobel recognition, other natural products have drastically changed the treatment of human diseases, such as taxol – a secondary metabolite from the bark of the pacific yew (*Taxus brevafolia*) which displayed a novel mechanism of action which was shown to be effective in the treatment of refractory ovarian cancers.[11] Taxol was approved for use in ovarian cancers in 1992, and was shown to

double the survival rate of patients after adoption. In ovarian cancers which show resistance to platinum-based chemotherapies, combination therapies that include taxol and its derivatives, raise the survival rate and outcomes considerably.[12]

1.1.3. The Future of Natural Products

Much like human warfare, victory in the battle for antibiotics is a temporary phenomenon and repercussions of the battle can last well into the future. Even before its initial release to the public, resistance to penicillin was observed in laboratory conditions, foreshadowing the widespread resistance we see today. This creates an ever-changing landscape and the need for new tools in the treatment of infections, cancers, parasites, and human disease. As natural products are often predisposed to bioactivity, they represent a wealth of potential in the never-ending quests for small molecule therapeutics.

Starting in the 1980s, the role of natural products in pharmaceutical development has been reduced due to high up-front cost and high rates of rediscovery as the main reasons. Despite the pharmaceutical industry retraction of their natural product divisions, natural products remain important sources of both new lead molecules, and conceptual scaffolds for synthetic design, accounting for over 60% of all small-molecule drugs approved since the 1980s.[1] Therefore, their importance in the field of human health remains as strong as ever and the need for better techniques for discovery and profiling of natural products is apparent. The field of metabolomics has developed high-throughput and high-content analyses methods which could reduce the rates of rediscovery in natural product discovery pipelines and better prioritize samples with novel chemistry.

1.2. Metabolomics and Natural Products

1.2.1. Targeted Metabolomics and Dereplication

Primary and secondary metabolomics studies can be further described by whether they are targeted or untargeted in approach. A targeted metabolomics approach often uses known metabolites (primary or secondary) as markers and gauges perturbations in these molecules across many different phenotypes to find correlation of the known compound to the unknown condition of the organism.[2] Examples of this include biomarker discovery and detection,[13] urinalysis or blood-plasma investigations,[14] organism state

monitoring, cultivar and geographic origin comparisons,[15] and comparison studies of popular supplements and foodstuffs.[16]

Targeted metabolomics is an expansion of dereplication efforts devised over the last century to identify molecules based on repositories of spectral features and physical characteristics of compounds already isolated.[17] Dereplication strategies are highly amenable to automation, and save considerable amounts of time, material, and cost of investigations into natural products.

Dereplication and targeted metabolomics can be performed through a variety of methods, depending on available reference data, including mass spectrometry based investigations[18], and NMR spectroscopy based investigations.[19] However, all methods of dereplication are contingent on these data being well organized and of sufficient quality for a robust comparison to be made. Currently, there are repositories of information available for public use,[20] but few are constructed to contain raw reference data, which could be more universally leveraged than data tables from original publications.

Several databases of natural product spectra carry subscription costs and are compiled on a for-profit model, limiting their widespread access. This has provided a niche for the academic community to create, curate, and continue to develop new repositories of information that are better suited for the needs of rapid investigations into complex mixtures. Notably, the Natural Product Atlas – a platform designed in the Linington Lab – is a publicly accessible database of more than 29,000 natural products with citations to their original discovery and description.[21] In addition, the NPAtlas is cross referenced with other repositories of note, such as MIBiG (>1400 annotations) and GNPS (>1200 annotations) to facilitate robust analysis from a variety of platforms. Robust databases enable rapid comparison of molecular families, their identified origins, and provide a great starting point for de novo structure elucidation efforts.

1.2.2. Untargeted Approaches

Untargeted metabolomics, conversely, is observing spectral characteristics without the identity of many of the molecules being known.[22] This method is inherently complex, as these investigations must extensively utilize several different analytical techniques simultaneously to arrive at a convergent answer as to what metabolite may be

driving a given response. In most untargeted analyses, there exists little or no prior knowledge of the identity of a given molecule and therefore the dependence on databases of matching spectral data is inherently lower.

Untargeted metabolomics in the field of natural products has largely been driven by statistical modeling of the presence, or absence, of a particular signal or feature to compare one sample to another. In these analyses, the focus is not to detect a given compound within a complex mixture, but rather to gauge the similarities and differences of samples to each other in a batch-wise fashion. This allows for quick prioritization to be performed on many samples, which can save considerable amounts of time and resources, and allow for predictions to be made about the novelty of a given feature across many samples simultaneously. When combined with contextual metadata about the sample, strategies in untargeted metabolomics can be used to gauge the validity and legitimacy of complex samples, such as botanical supplements.[23]

When combined with the strengths of targeted metabolomics and dereplication, a vast amount of information can be generated about the plausible identity of a molecule which is identified by untargeted techniques. This approach, sometimes called targeted-untargeted metabolomics, can functionally annotate an extract by gauging what is known via dereplication or targeted analysis, and further evaluate if the known compounds explain the novelty of the extract.[22] Analyses capable of combining these schemes are well situated for the investigation of complex extracts where some metabolites may be completely new but are normally obscured by known chemistry.

1.3. Existing Methods of Metabolomic Profiling of Natural Products

1.3.1. Mass Spectrometry

Mass spectrometry has emerged as a first-choice utility for the untargeted profiling of an organism's metabolome in recent years. High resolution mass spectrometers have incredible dynamic range, are extremely sensitive, and have a resolving power which is largely unmatched in analytical instrumentation. Using MS as a driving utility in metabolomics, many tools have been designed which leverage the massive amount of

information from MS with metadata and biological evaluation to predict the bioactivity of unknown metabolites. These tools, such as the Global Natural Product Social Molecular Network (GNPS)[24], Compound Activity Mapping,[25] and Biochemometrics[23,26] have given investigators new profiling tools to investigate the complex profiles of extracts with minimal sample separation or manipulation.

Comparison strategies based in MS can leverage several layers of data for comparison simultaneously. The primary feature for comparison in most strategies is the molecular ion, generated through the analysis of a charged ion by a mass analyzer. However, with the introduction of LC-MS systems, analytes are analyzed as they elute from the LC system generating a retention time specific to a molecule. The combination of these two features form the basis for standard comparison for many MS systems and dereplication protocols.[18] In addition to this, many mass spectrometers can fragment ions through forced collisions with an inert gas, generating ionized fragments of the molecule.

The matching of fragmentation patterns in MS systems can be leveraged beyond verification of molecules against a known standard. Platforms which leverage this compare the way in which analytes fragment instead of comparing absolute values of the fragment pieces, relying on the comparison of patterns over singular features. With the introduction of molecular networking and GNPS, the patterns of fragmentation can be analyzed and compared across hundreds of samples for similar fragmentation profiles.[24] This allows associations of structurally similar molecules to be made even with changes in the overall molecular structure (such as methylation). Molecular networking enables researchers to gain a perspective on what the ions of interest may be, as comparisons are done against a growing library of reference datasets. GNPS itself is available for public use and data deposition, allowing for community curation of compound identities and fragmentation information.

Mass spectrometry-based methods of comparison are powerful but come with certain limitations. Molecules of interest must have the ability to ionize to be detected, and not all analytes ionize well in MS. Ion suppression is a well documented phenomena in MS and can be pronounced in complex sample analyses.[27,28] Additionally, the mechanisms of fragmentation in MS systems is not fully understood, and fragmentation itself provides limited information of molecular connectivity unless compared to a standard.

1.3.2. Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) spectroscopy is an analysis technique which has seen demonstrated improvement over the last 15 years in both sensitivity and resolution due to the introduction of higher field instruments and advanced processing methods.[29] Although notably less sensitive when compared mass spectrometry and other spectroscopic methods, NMR offers considerable advantages in establishing molecular connectivity. NMR spectroscopy has been used to great effect in the metabolomics community due to inherent advantages offered by this method of analysis. NMR spectroscopy is regarded as a 'universal detector' for organic molecules, is relatively quantitative under standard conditions, and is highly reproducible.[30] NMR experiments provide molecular connectivity, allowing investigators to directly annotate and solve the structure of an analyte.

NMR spectroscopy works by exploiting the relationship between nuclear spin states, excitation of nuclei by RF frequencies, and the detection of emitted signals as the nuclei relax to ground state. NMR active nuclei exhibit a property called spin, which induces a magnetic field with respect to the spin direction. Outside of a magnetic field, these magnetic moments are scattered at random in all directions. However, in the presence of a magnetic field, atomic nuclei align with respect to the field (B_0). As the atomic nuclei are still not perfectly aligned, the sum of all magnetic vectors – called the magnetic moment - determines the spin state. Nuclei can be in either a spin-up (mostly parallel with the magnetic field) or spin-down orientation (mostly anti-parallel). Energetically, in ^{1}H NMR spectroscopy the spin-up state is more favorable, and therefore more nuclei align themselves in this orientation. The energy difference between the two spin states (called the Boltzman distribution) scales with the strength of the magnetic field; the stronger the magnetic field, the more nuclei are in the lower energy state. When exposed to a RF pulse of a particular frequency, the nuclei absorb the energy, and the equilibrium is disturbed. When the RF is halted, the nuclei release the energy necessary to return to equilibrium which is measured by the detection coil. The larger the change in the bulk magnetization, the more intense the signal detected. Both ^{1}H and ^{13}C nuclei are detectable by NMR spectroscopy, and as nearly all organic molecules contain at least one 1H, NMR spectroscopy is regarded as a universal detector in organic chemistry.

^{1}H NMR analysis of pure compounds is a relatively straightforward analysis technique and is a routine technique in many fields of chemistry. Each ^{1}H nucleus emits a signal proportional to the number of nuclei in each environment, allowing quantitative relationships to be made for each signal at a given chemical shift. The chemical shifts at which these signals are detected, describe the amount of chemical shielding a given nucleus is exposed to, allowing for functional groups and environments to be deduced. NMR signals at a given chemical shift can exhibit splitting as nuclei couple to each other due to a phenomenon known as spin-spin, or scalar, coupling. This relationship splits a given NMR signal proportional to the number of neighbors which are coupled, allowing for connectivity relationships to be established.

Taken together, the quantitative and qualitative information from ^{1}H and ^{13}C NMR spectroscopy allows chemists to deduce the structure of an unknown molecule. However, when multiple compounds are present in the sample, these relationships become harder to resolve through signal overlap and the difficulty of associating of signals arising from a single molecule scales with the number of chemical species present.

NMR Analysis of Complex Samples

NMR based metabolomics using 1D NMR experiments such as ^{1}H and ^{13}C have increased drastically in popularity and utility. In many applications of 1D targeted metabolomics, diagnostic signals from metabolites are used to predict the presence of a metabolite, rather than detection of all signals from the compound. These studies allow for comparison of highly overlapped and complex spectra but rely on some prior knowledge of the metabolites expected. One such example is the development of a cranberry model system,[22] which allows for analysis of the origin of cranberry samples with direct applications for quality control in industry. In this case study, reference spectra of major metabolites were used to identify signals which could be extracted from complex spectra to confirm the presence and abundance of metabolites through 1D quantitative NMR methods and ascribe differences in the profiles to geographic origin of the botanical. Expansion of these methods have enabled the creation of workflows to authenticate botanical supplements by profiling expected metabolites from samples to reference data.[31]

Combinations of signal patterns, often referred to as barcoding or fingerprinting, have been used with great success when comparing and identifying metabolites from 1D ^{1}H NMR profiling. Using fingerprints of isolated compounds for future dereplication, analysis

pipelines specific to species or organism source can be leveraged.[32,33] HiFSA (^{1}H iterative Full Spin Analysis) has been used in the analysis of isomers which have near identical NMR profiles, and for the generation of high quality fingerprints.[32] However, as the number of components in a sample increase, the ability of these platforms to deduce sample constitution is reduced, as overlap in signals can create problems in the extraction of both chemical shift and ^{1}H quantitation.

1.3.3. 2-Dimensional NMR Spectroscopy

Although a considerable amount of information can be afforded by 1D NMR experiments, complications arise when attempting to detect and verify the presence of compounds in a crude mixture. These complications arise from overlap of resonances in the ^{1}H NMR spectrum, concentration differences in the metabolic profile of the extract, or the rise of complex couplings which crowd the spectra. Attempts have been made to deconvolute the information in 1D NMR experiments in recent years, such as CRAFT NMR[34] and various deconvolution algorithms which are made available by both vendor supplied[35] and third party software packages.[36,37] However, the challenge of deducing chemical constitution in severely overlapped spectra remains a principal challenge in NMR.

Two-dimensional NMR experiments come in a wide variety of applications and each exploit different phenomena to associate individual resonances together to better understand molecular structure. Each 2D NMR experiment can be broken down into a collection of 1D experiments which adjust predetermined variables in the pulse sequence to create an effect on the system that can be compared across multiple acquisitions. These independent variables can be adjustments in the delays between pulses, adjustments in pulse lengths, the power of each pulse, or even the shape of the pulse applied. Comparing how these variables affect the system over several acquisitions affords new knowledge pertaining to coupling, connectivity, or spatial interactions. 2D NMR is less amenable to the sample complexity effects when compared to ^{1}H 1D NMR. One reason is that it provides a secondary axis to separate resonances through an additional correlation axis that is less likely to overlap and is therefore a plausible solution to spectral complexity.

There are many 2D NMR experiment types, each of which afford new information in the molecular connectivity or physical characteristics of the analyte, such as its diffusion rate through a given mixture. These experiments can be separated into three categories –

heteronuclear correlation, homonuclear correlation, and physical/spatial correlation experiments, each demonstrating orthogonal characteristics of the molecules they analyze.

Homonuclear Correlation Spectroscopy

In NMR spectroscopy, homonuclear correlation spectroscopy is driven by atomic nuclei from the same element, provided they are the same isotope. These experiments are used to determine molecular connectivity by demonstrating through bond correlations, from coupling of the nuclei to one another. There are several routine experiments which utilize homonuclear correlations, such as **CO**rrelation **S**pectroscop**Y** (COSY), **TO**tal **C**orrelation **S**pectroscop**Y** (TOCSY) and the 2D *j*-resolved ^{1}H experiment. These ^{1}H-detected experiments are widely used, but other homonuclear experiments exist in which other nuclei are used, such as the ^{13}C detected INADEQUATE experiment, but these suffer from decreased sensitivity when compared to ^{1}H detected experiments. The decrease in sensitivity is a major drawback, and as such, these experiments are somewhat rare in practice despite their obvious utility.

Homonuclear NMR experiments are of extreme value for structure elucidation, serving as a method to establish scaffold connectivity through coupled nuclei. The COSY experiment establishes connectivity typically through 3 bonds (Figure 1.1- Top), allowing for detection of adjacent protons along a given carbon backbone, while the TOCSY allows for many more connections to be established (Figure 1.1- Bottom).

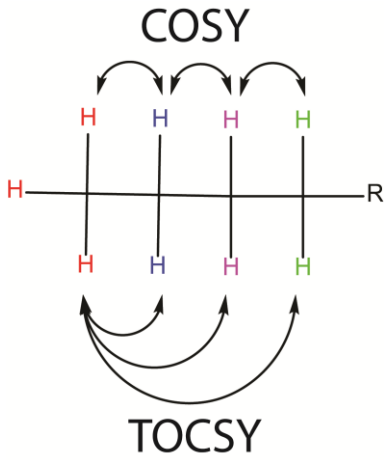


Figure 1.1. Homonuclear Correlation Spectroscopy. Top: COSY allows for adjacent neighbors in a spin system to be determined independently. Bottom: TOCSY allows for all neighbors in a spin system to be determined simultaneously.

Heteronuclear Correlation Spectroscopy

Heteronuclear experiments, such as the **H**eteronuclear **S**ingle **Q**uantum **C**oherence (HSQC) and the **H**eteronuclear **M**ultiple **B**ond **C**orrelation (HMBC) experiment are similar to homonuclear experiments in that they transfer the magnetization energy between nuclei which are connected through bonds but differ in that the nuclei observed are not of the same element. Both the HSQC and HMBC experiments are able to detect correlations of ^{1}H-^{13}C and ^{1}H-^{15}N, but differ on the number of bonds the energy must travel through.

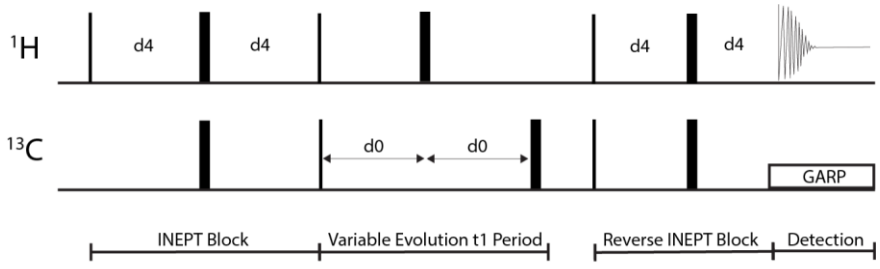


Figure 1.2. The HSQC Pulse Sequence. Combinations of pulses and timing allow for the transfer of energy through bonds of ^{1}H nuclei to ^{13}C nuclei (INEPT Block) which is then allowed to evolve over a given timeframe (d0) before being transferred back to ^{1}H for detection (Reverse INEPT Block).

The HSQC pulse sequence has four main components, the INEPT block, the variable evolution period, the reverse INEPT block, and detection (Figure 1.2). The INEPT/reverse INEPT block transfers energy between ^{1}H nuclei to ^{13}C nuclei before and after the variable

evolution period where differences in chemical shift of the ^{13}C nuclei can evolve. The detection phase includes simultaneous decoupling of the ^{13}C nuclei, allowing for simplified resonances with a reduction in coupling based splitting. When combined with a spin-lock step during the variable evolution period, it is also possible to derive multiplicity information from the phase of the resonances, affording information previously obtainable through many ^{13}C detected DEPT experiments with no additional time costs and increased sensitivity.[38]

The information afforded by the HSQC experiment is extremely valuable as it allows for the detection of directly bonded ^{13}C nuclei through the detection of ^{1}H resonances. This is a notable advantage to ^{13}C detected methods, as ^{13}C has only 1% natural abundance, and a magnetic moment which is ¼ that of ^{1}H (found at 99.98% abundance compared to other species of H). This method directly translates to higher sensitivity than ^{13}C detected pulse sequences and allows for minor compounds to be observed alongside more abundant analytes.

Physical/Spatial Correlation Experiments

In addition to molecular connectivity, NMR spectroscopy can apply experiments that are dependent on the physical characteristics of an entire molecule. DOSY – **D**iffusion **O**rdered **S**pectroscop**Y** is a method of NMR analysis which focuses not on connectivity of atoms, but rather on the physical properties of nuclei diffusing through a medium over a given time interval. DOSY itself is not a pulse sequence, such as an HSQC or TOCSY, but instead relies on the relative intensity of analytes across several identical experiments with an applied dephasing gradient and delay to allow for the molecules to diffuse at random within the solvent. After the diffusion delay, selective refocusing of the magnetization allows for resonances which are in-phase to be detected. Over the course of several acquisitions, induced changes in intensity of analytes can be used to back calculate the decay rate of the signals as a function of gradient strength. This decay rate can then be used to find the diffusion rate of the resonance which is molecule dependent. This allows for the association of many resonances originating from single analytes to be made based on their unique diffusion rates. Molecules which are small and diffuse more easily though the medium have a faster diffusion rate, and molecules which are larger in structure that are not as mobile diffuse at a slower rate.[39] Because DOSY is a processing method utilizing slightly modified pulse sequences, it is extremely versatile in the types of

experiments which it can be leveraged for, including implementation of diffusion elements into multidimensional NMR experiments.

1.4. Application of NMR Metabolomics for Functional Annotation of Natural Products

The importance of natural products in modern medicine is well established, but the challenges in the study and isolation of these metabolites have slowed the rate of discovery for truly novel compounds. However, the challenge of annotating the components from natural sources has produced remarkably innovative utilities to continuously improve our ability to understand the molecular world around us. Many of these utilities have been designed for MS, but carry with them the inherent drawbacks of MS based analyses. In recent decades, new approaches in NMR have allowed for increased sensitivity, new experiments, and processing techniques which are well suited to address the needs of the NP discovery community.

The focus of this dissertation is the development, implementation, and orthogonal applications of a metabolomics utility deemed MADByTE (**M**etabolomics **A**nd **D**ereplication **By T**wo-Dimensional **E**xperiments). MADByTE itself is not a singular experiment, rather a coupling of information from several techniques which were selected to give a new perspective on the types of metabolites present in complex mixtures. In Chapter 2, the overall rationale and design of the platform will be discussed alongside practical challenges in the development of an NMR based method for feature comparison. The framework of MADByTE networks will be discussed, and their application towards dereplication and structural annotation of similar compounds.

To prioritize bioactive molecules, an expansion of MADByTE was developed to overlay bioactivity data directly on top of the MADByTE association networks. Chapter 3 will focus on the design, rationale, and application of this extension towards the isolation of an active component from an actinobacterial extract. This application highlights the ability of the platform to be further refined by the addition of new experimental data and builds on the concept of associating data beyond NMR into the platform.

The need for robust databases for the construction and use of new workflows, as described previously, cannot be understated. New technologies which attempt to utilize

data from unknowns, such as with MADByTE, could be further leveraged if sufficient resources existed to encourage new *in silico* developments. In Chapter 4, the combination of MADByTE sample comparison to a machine learning platform (SMART),[40] originally designed for pure compounds, is explored with a new module to generate substructure predictions. This application is made possible through community contributions to a publicly accessible database, NMRShiftDB2,[41] and the *in silico* utilities designed for prediction of NMR chemical shifts.[42] SMART and NMRShiftDB2 are well curated databases, and Chapter 4 explores their integration into a new hypothesis generating utility. SHIMS (**S**ubstructure **H**ypothesis by **I**ntegration of **M**ADByTE and **S**MART).

Chapter 5 explores the combination of additional NMR experiments into MADByTE to generate robust features for comparison through the introduction of DOSY techniques. Originally designed for simplified mixtures, DOSY shows great promise in the ability to refine spin systems derived from MADByTE, especially in areas where a crowded 2D NMR spectrum can complicate analysis. Using diffusion in conjunction with MADByTE networking through a workflow called FADES (**F**eature **A**ssociation by **D**iffusion **E**xperiment**S**) allows for more robust spin system formation and association of spin system features through similar diffusion rates.

All applications of MADByTE described within this dissertation were constructed from an NMR first perspective, focusing primarily on how relatively simple analyses can give rise to information rich datasets. MADByTE was developed to put this information into real context, allowing new perspectives to be considered and offering new vantage points to the landscape of analytical technologies available to NP discovery. The landscape of utilities developed for annotation of natural product extracts is ever changing and new strategies will be needed as new challenges arise. These tools do not exist in isolation, and to be of practical use to the community at large, must be able to pivot to a variety of sample types and experimental objectives. Collectively, the development of MADByTE and the introduction of the SHIMS and FADES modules highlight the ability to integrate orthogonal datasets into an open source, context-driven tool for the functional annotation of natural products mixtures.

Chapter 2.

Design of the MADByTE Platform

2.1. Introduction

Natural products have traditionally played a central role in drug discovery, but novel bioactive compound discovery is becoming increasingly difficult as the field matures and the number of known scaffolds increases.[43] Standard approaches rely heavily on bioassay-guided fractionation, which often results in the re-isolation of known compounds and carries an inescapable material cost. To reduce the chances of re-isolation, numerous dereplication methods have been developed, including UV, mass spectrometry and NMR-based platforms.[17,44] However, many of these tools rely on in-house databases that are slow and expensive to generate, and require high coverage for exact database matching.

Methods which compare spectral features from samples rather than attempting to match single components explicitly have been developed to remove the need for large reference databases. In the field of MS-based natural products, a workflow named Global Natural Products Social Molecular Networking (GNPS) has revolutionized front-end investigations of complex samples.[24] GNPS uses molecular networking to associate MS features based on the similarity of their MS^2 profiles by analyzing fragmentation patterns for similarities. These similarities are then visualized through the creation of networks between MS features to generate clusters of features which may be structurally related. In absence of reference libraries, this profiling technique can associate molecular ions within sample pools to describe shared chemistry in a sample set. However, when coupled to external reference libraries, structural annotations can be made which provide robust starting points for structure elucidation or sample triage. These “feature first” tools improve prioritization efforts by enabling investigators to group metabolites into compound families, and to determine the distribution of these families across sample sets.

Despite the numerous advantages of this methodology, mass spectrometric analysis suffers from several inherent limitations, including variable ionization efficiency between analytes, and ion suppression.[27,28,45] In addition, mass spectrometry yields limited structural information compared to other analytical methods. Therefore, complementary methods are needed that can address the existing limitations of these approaches.

In contrast to mass spectrometry, NMR spectroscopy provides direct structural information, is a universal method of detection, and is semi-quantitative under standard conditions. NMR-based metabolomics approaches have increasingly focused on the development of platforms capable of highly accurate annotations of known primary metabolites, especially in biofluids.[13,14,46–48] These approaches have been successfully used to highlight high priority regions of the spectra, using spectral variability as a proxy to gauge potential novelty or detection of biomarkers.[49,50] However, ^{1}H NMR-based metabolomics methods cannot relate signals from the same molecule together, limiting identification options for unknowns.

As described in Chapter 1, an inherent strength of NMR-based platforms is their ability to resolve complex mixtures in two or more dimensions and are a rapid method to establish molecular connectivity. 2D NMR data offer a robust method of annotation when compared against a database of known compounds,[19,20] but are often limited by the availability of reference data.[51] Several platforms exist for annotating metabolites utilizing 2D NMR data, including dereplication utilities and targeted metabolomics in biofluids (Table 2.1).[52]

Table 2.1. NMR Utilities for Natural Product Investigations

	MetaboMiner[53]	**COLMAR**[20]	**SMART 2.0**[54]	**DEREP-NP**[55]	**MADByTE**[56]
Analysis Type	Targeted	Targeted	Targeted	Targeted	Untargeted
Designed for Mixtures	✓	✓	×	×	✓
Reference Database	Required	Required	Required	Required	Optional
Metabolite Type	1°	1°	2°	2°	2°
Sample Type	Biofluids	Biofluids	Single Compound	Single Compound	Extracts
Batch Comparison	×	×	×	×	✓
Solvent System	Buffered D_2O	Buffered D_2O/$CDCl_3$	Independent	Independent	Independent

Like MS^2 based methods, comparison of NMR spectroscopy derived fingerprints from complex samples has shown great promise in the prioritization of desirable compounds. In one study, HSQC-TOCSY spectra were used to fingerprint crude extracts from a library of bacterial isolates, and these fingerprints were used to prioritize strains enriched in NMR

derived motifs for polyketide and peptidic natural products; scaffold types with significant precedent in bioactive natural products discovery.[55]

Many existing NMR-based tools require pure, or simplified samples for accurate structure prediction or priortization.[57] To address this issue, MADByTE (Metabolomics And Dereplication By Two-dimensional Experiments) was developed to deconvolute NMR data from complex natural products mixtures into features for comparison. MADByTE annotates extract spectra, verifies components through feature matching to pure compounds, and identifies features associated with highly bioactive samples for the isolation of bioactive constituents. MADByTE works by integrating ^{1}H-^{13}C connectivity data from HSQC spectra with ^{1}H-^{1}H scalar couplings from TOCSY spectra to define scaffold substructures from multiple components simultaneously. These spin system features (SSFs) can be related between samples, or compared against reference datasets for compound dereplication, accelerating the discovery process.

Unlike many of the existing NMR-based profiling tools, MADByTE does not require a bespoke spectral reference library against which to compare NMR data (Table 2.1). This is an important distinction, as it offers a new mechanism to evaluate the chemical similarities and differences between samples, regardless of whether these constituents are known or novel natural product classes.

2.1.1. Theory

MADByTE's analysis strategy builds on the idea that if a metabolite is shared between samples, there will be shared spectral features that accompany it. This approach is common among dereplication platforms which compare MS profiles of complex samples.[18,58] The rapid adoption of these strategies have enabled discovery efforts to develop new methodologies for chemical annotation by using repeated observations of fingerprints to describe chemical constitution.[59,60] These new targeting strategies have improved up-front profiling strategies by enabling many compounds to be simultaneously annotated.

The feature comparison strategy MADByTE employs is to extract spectroscopic information from individual 2D NMR spectra (as points), generate associations between points from HSQC and TOCSY spectra (as features) and, finally, to compare features

between samples across the entire sample set. The result of this comparison is an association matrix which displays the homology of points between any two features as a ratio of points found compared to the size of the features compared. This comparison strategy allows for imperfect matches between overall features. This is advantageous, as it allows for association of similar chemical species which may have slight changes in the core scaffold. Even when comparing the same compound between samples, differences in sample constitution or temperature would hinder annotation efforts if MADByTE required each resonance to match perfectly.

Once the association matrix is constructed, MADByTE generates network visualizations which place the information into context and highlight the chemical similarities between compounds in different chemical extracts. To achieve this, a strategy was devised to create SSF nodes for each feature from a sample and to connect these nodes to a central extract ID node. On a per sample basis, this represents all features from an extract directly associated to an anchor point (Figure 2.1 – Panel A). As features are found to be similar across samples, connections are made which allow for a network to grow describing chemical similarity across the set (Figure 2.1- Panel B). Networks can be rendered to display all SSFs or display only SSFs which have similarity in the sample pool.

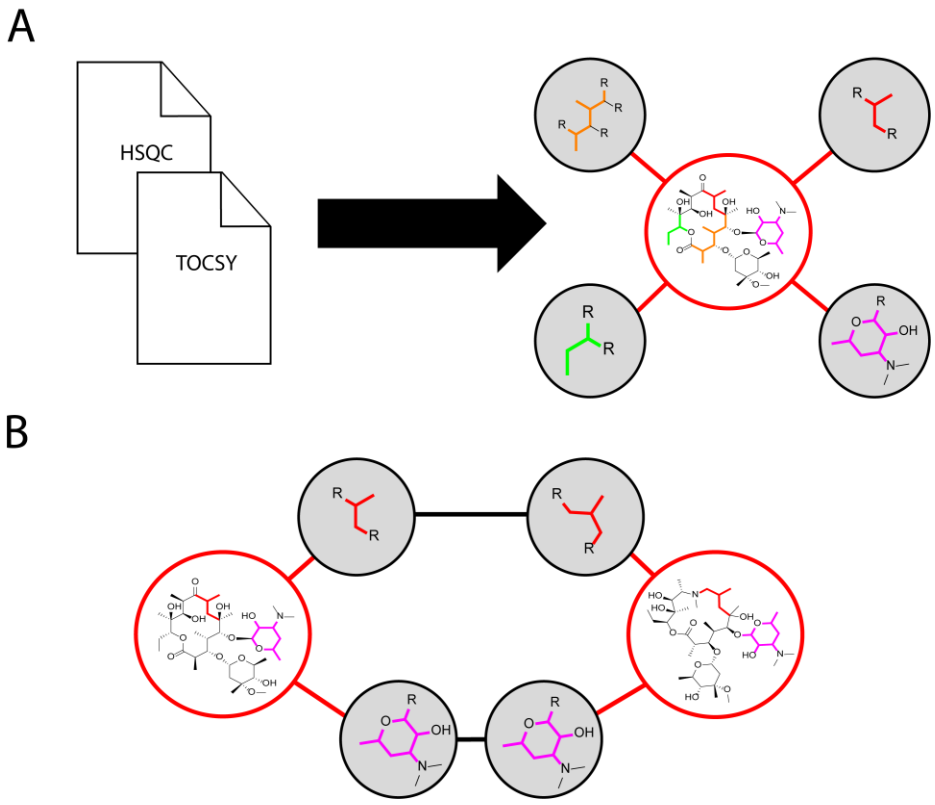


Figure 2.1. **Conceptual Overview of MADByTE Comparison. A) Peak lists from HSQC and TOCSY are deconstructed into spin system nodes (grey) which represent scaffold pieces of molecules in the sample. B) As similar molecular pieces are found in other samples, connections between nodes allow for similar samples to be associated.**

2.2. Practical Considerations

2.2.1. Experiment Selection and Setup

Pulse Sequence Selection

MADByTE is structured to use two separate 2D experiments, the HSQC and TOCSY. The HSQC pulse sequence selected for use in the MADByTE system is the hsqcedetgpsisp2.3 pulse sequence, which is optimized for sensitivity when compared to a standard HSQC pulse sequence. This is accomplished by replacing the rectangular pulses typically used in the INEPT portion of the pulse with adiabatic pulses, which have increased response from ^{13}C nuclei over a wide range of resonance frequencies.[61] In addition to this, the phase

sensitive nature of the experiment provides multiplicity information that is used in spin system construction.

The TOCSY pulse sequence used for MADByTE was dipsi2gpphzs, a robust pulse sequence which optimizes the response of in-phase signals.[62] TOCSY mixing time is an important factor, as it determines both the signal intensity and length of the detected spin system. Typical TOCSY mixing times are between 30 ms and 120 ms. Longer mixing times can reveal longer spin system associations, but if set to long probe damage can occur. A sample of 3% lactose by mass was prepared and run with mixing times of 60, 80 and 100 ms. The length of the spin system was similar between 80 and 100 ms, and reduced for the shorter mixing time of 60ms. A mixing time of 100 ms was chosen and used for the collection of all MADByTE data. To balance signal response and spectrum resolution, TOCSY spectra were processed with a sine squared function with a sine bell shift of 2.

HSQC variants include the HSQC-TOCSY experiment, where TOCSY cross peaks are arrayed on the ^{13}C axis in the F1 dimension. In principle, a combination HSQC-TOCSY offers time savings, as only one experiment must be acquired. To evaluate the usefulness of the combined HSQC-TOCSY, 8 mg of lincomycin was dissolved in 500 µL of DMSO-d_6 and subjected to HSQC-TOCSY analysis. Using 16 scans in the HSQC-TOCSY, the sensitivity loss was noticeable when compared to a TOCSY using only 2 scans (Figure 2.2). S/n was calculated for these slices with s/n of 676.3 for TOCSY at 2 scans, compared to a s/n of 160.9 for the HSQC-TOCSY. As MADByTE was designed for the investigation of crude mixtures where some metabolites are present in low abundance, individual HSQC and TOCSY spectra were collected for this study.

By requiring the HSQC as a separate experiment, explicit ^{1}H-^{13}C relationships can be assigned. This is not possible in the HSQC-TOCSY. As an added benefit of using orthogonal experiments rather than a singular experiment, MADByTE requires resonances in both the HSQC and TOCSY to be present in order to be counted as a valid resonance, ensuring artifacts from noise in one spectrum does not confound the results of the combination processing. Although HSQC experiments are notably less sensitive than TOCSY experiments, HSQC-TOCSY experiments are less sensitive than the selected HSQC experiment. This is because the HSQC-TOCSY experiment (hsqcdietgpsi) is an expansion of an already less sensitive HSQC experiment (hsqcetgpsi)

with an added spin lock in the pulse sequence which may contribute to additional loss of signal.

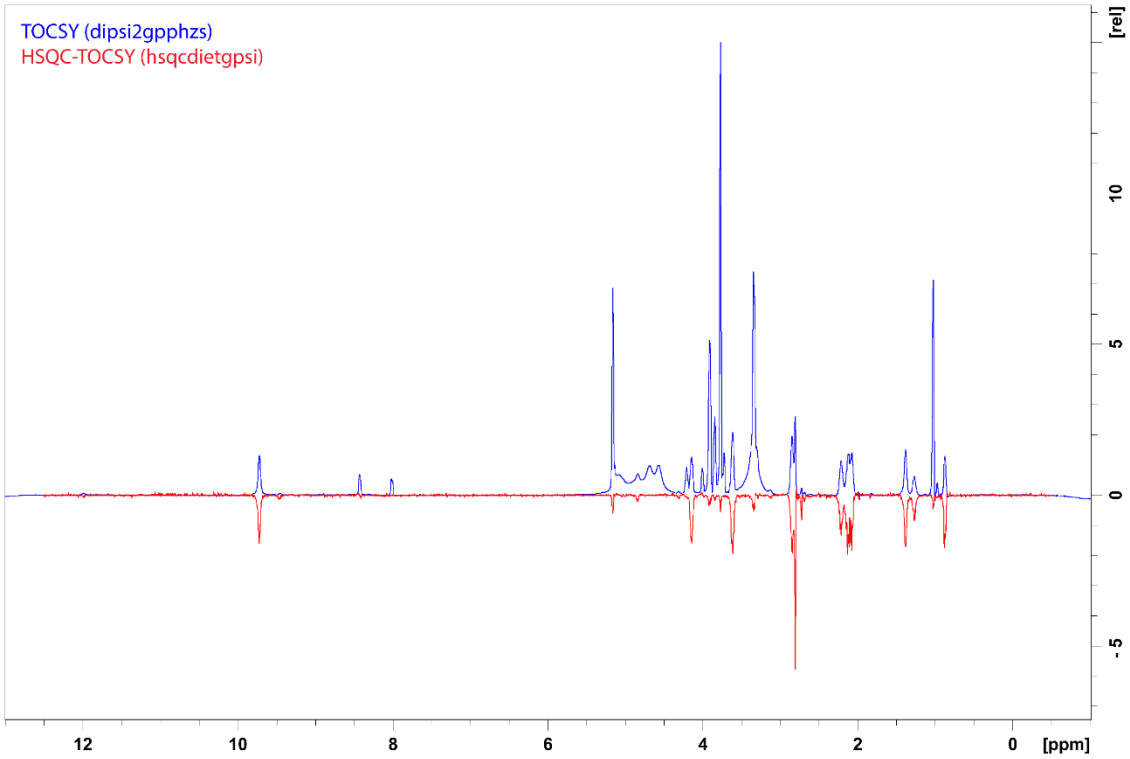


Figure 2.2. Sensitivity of TOCSY vs HSQC-TOCSY for a Spin System of Lincomycin at 3.16 ppm. The overall sensitivity improvement of the TOCSY vs HSQC-TOCSY can be seen through additional correlations and higher s/n.

2.2.2. Sample Considerations

Sample preparation for NMR metabolomics studies is extremely important to ensure that the information provided by the analysis is both meaningful and truly representative of the chemistry in an extract. In addition, external factors such as sample scarcity and

conservation place a specific burden on sensitivity considerations. Sample preparation must also be done as consistently as possible, as downstream comparisons require minimal variation between samples to ensure reliable and robust analyses.

Natural product libraries containing microbial extracts often contain samples of varying complexity and concentrations. Any profiling technique should consider the range of polarities and solubility profiles of the metabolites being analyzed. In primary metabolomics, pH is often standardized by buffering the solvent to achieve comparable metabolite profiles between samples. In natural product metabolomics analyses, buffering was not done as the samples were not obtained from aqueous prefractions and instead consistent profiles were obtained by acquisition in the same solvent (DMSO-d_6).

Often, samples from natural product libraries are stored in DMSO for long term storage, as it is generally regarded as a very effective solvent and can help to solubilize a wide range of compounds.[63] Use of DMSO carries an additional practical consideration, as this solvent is hydroscopic and absorbs water from the atmosphere readily. Over a period of time in storage, this water content can increase drastically.[64] Additionally, DMSO can be a difficult solvent to fully remove through evaporation. This complicates preparation for NMR analysis as the concentration of native DMSO and/or water in the sample can quickly overwhelm signals originating from metabolites in NMR analysis. To remove native DMSO from the extracts, samples were lyophilized for a minimum of 12 hours, resuspended in DMSO-d_6 and lyophilized for an additional 12 hours to remove as much native DMSO as possible. Dried extracts were resuspended in 280 µL DMSO-d_6 for NMR analysis to provide a consistent volume in the detection coil, providing a consistent shim profile while retaining a representative concentration profile for relative quantitation.

An additional consideration in the design of MADByTE was sample scarcity. Isolates of an organism may produce different chemical profiles over time, which complicates sample generation and replication. The Linington lab chemical library is created using a standardized approach in which 1 L scale cultures of actinobacteria are condensed into six 1 mL DMSO aliquots representing each prefraction (prefractionation procedure outlined in section 2.7.1), providing the sole supply of a given extract and a snapshot of the chemical profile of the organism at a singular point of time under these conditions. As replication of this chemical extract is not guaranteed, practical sample allocations allow for analyses by multiple platforms. For the development of MADByTE, the practical

working limitation was an aliquot of 25 µL from each 1 mL stock, accounting for 2.5% of the total volume of the chemical extract supply.

NMR Tube Selection

Concentration of the sample can have drastic effects on the profiles obtained during NMR metabolomics investigations. In some cases, some compounds are present at levels too low for detection. To increase the signal intensity in NMR spectroscopy, investigators often attempt to maximize the amount of sample in solution. In practice, NMR sensitivity is directly proportional to the amount of sample in the detection coil, shown by Equation 2.1 where S/N is the signal to noise ratio, n is the number of spins in the detection coil, γ_e is the gyromagnetic ratio of the nucleus being excited, γ_d is the gyromagnetic ratio of the nucleus being detected, B_0 is the magnetic field strength and t is the acquisition time.

$$S/N \propto n\gamma_e\sqrt{\gamma_d{}^3 B_0{}^3 t} \qquad \textbf{(2.1)}$$

In the case of sample limitation with all other factors kept consistent, an increase of sample concentration (n) by a factor of 2 directly translates to a 2-fold increase in the signal to noise ratio of the overall experiment. However, investigations with limited source material provide a challenge as access to more sample is not always feasible. There are several options which provide a way around this, including reducing the concentration of solvent in the coil while maximizing the concentration of analyte. This increases the number of spins from the analyte in the detection coil while minimizing to the number of spins contributed by the solvent.

Shigemi Tubes are unique NMR sample tubes which are ‘matched’ to solvents providing a way to achieve higher concentrations in the detection coil while retaining a favorable shim profile by mimicking solvent both above and below the sample. This is accomplished by positioning the solvent between a plug and plunger assembly that is matched to the magnetic susceptibility of the solvent. This creates a shim profile like that of a standard NMR tube filled to the recommended volume, but with a sample/volume ratio up to 2/3 higher. As an added advantage, this configuration allows for all sample signal to be directly in the detection coil, which maximizes the signal when compared to a reduced sample volume in a standard NMR tube. Small volume tubes, such as 3 mm and 1.7 mm tubes

are often used in metabolomics studies as they allow for a considerably reduced volume requirement to fill the tube, allowing for more sample to be dissolved in minimal solvent.

To select the optimal tube type for obtaining spectra with good peak shape and sensitivity in a 5 mm probe, a sample of 0.84 mg of mupirocin was suspended at 1 mg/mL in methanol, sonicated, and 450 µL was transferred to each of two separate vials. Each vial was concentrated to dryness under a stream of nitrogen gas, resuspended in either 150 µL or 250 µL of DMSO-d_6, and placed into a 3 mm NMR tube and a matched Shigemi tube, respectively. After automatic shimming using 'topshim', the 5 mm Shigemi tube showed an improvement in shimming time, compared to the 3 mm tube, which required manual shim correction to achieve decent peak shapes. The 5 mm tube also afforded higher peak area and improved signal to noise ratio compared to the 3 mm tube. The signal to noise ratio was calculated using the Bruker command SINO with the signal region between 3.98 - 4.02 ppm and the noise region defined as 7 - 8 ppm with a s/n of 5954 for the 3 mm tube and 14180 for the 5 mm Shigemi tube. All experiment settings were kept consistent for both experiments.

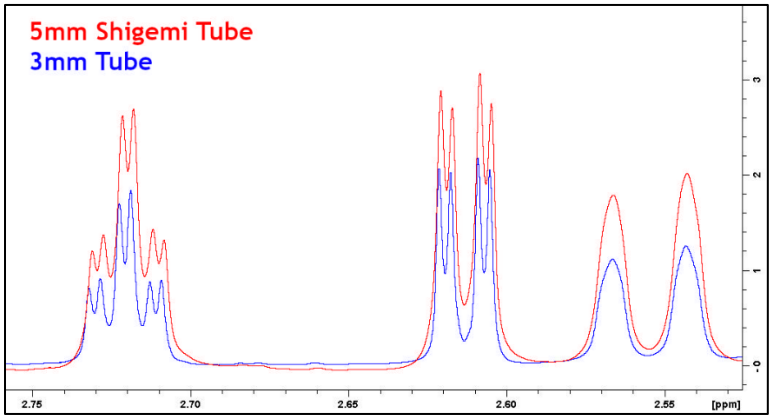


Figure 2.3. Comparison of Peak Area of 0.42 mg of Mupirocin in a 3 mm Tube vs 5 mm Shigemi Tube in a 5 mm TCI Probe

Optimizing conditions and increasing the volume of solvent to 280 µL in the 5 mm Shigemi tube allowed for easy automatic shimming. An additional benefit was the ability to use Shigemi tubes in an autosampler to increase sample throughput, as opposed to manual loading and shimming required by the 3 mm tube setup.

2.3. Data Processing

2.3.1. MADByTE Architecture

MADByTE was deigned and built in the Python programming language and comprises over 2000 lines of code and more than 90 functions. These functions encompass the full data processing pipeline spanning data import from vendor specific files to the generation of various network plots from post processed data. A simplified overview can be seen in Figure 2.4.

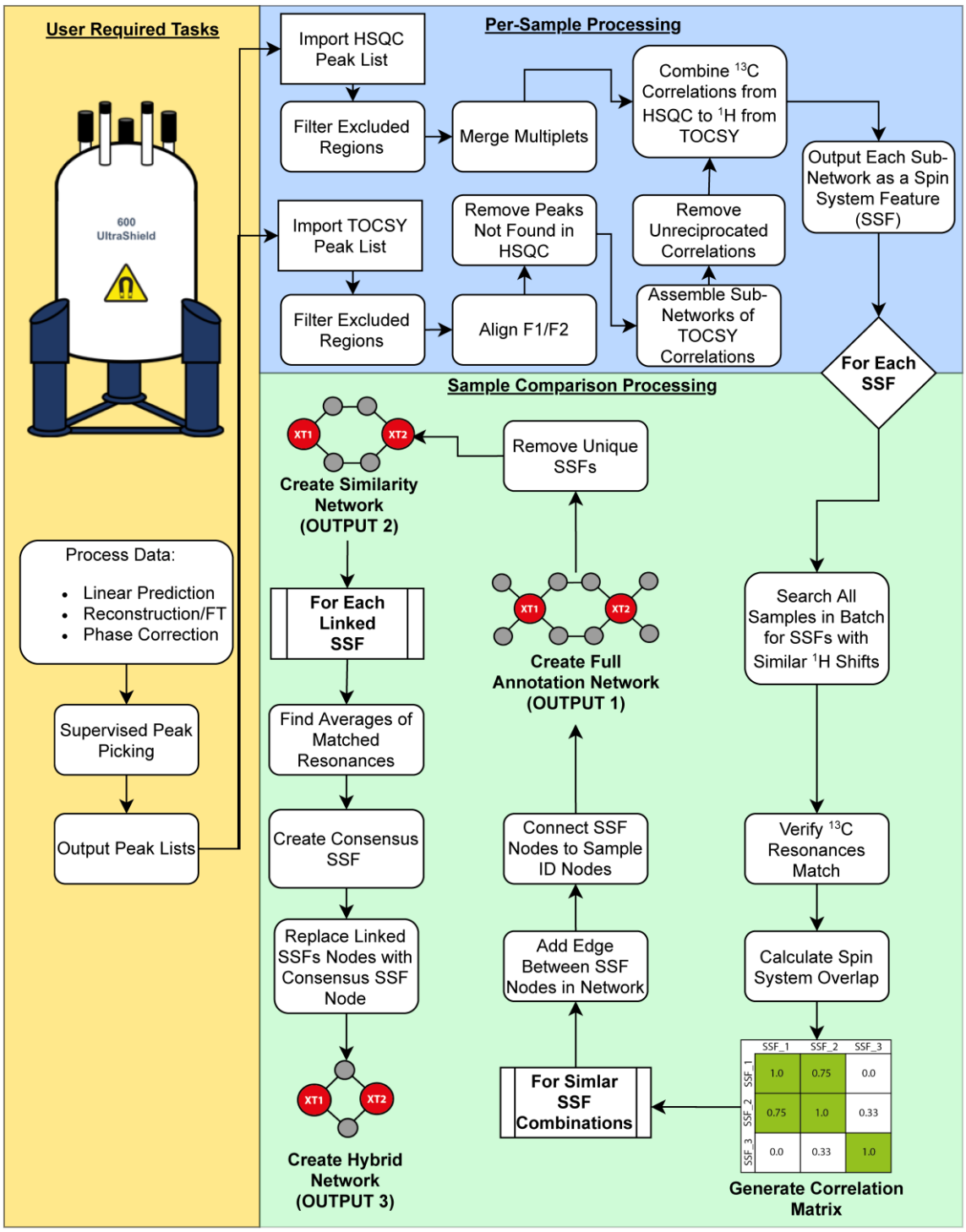


Figure 2.4. Overview of Processing Steps and Outputs from MADByTE Analysis

At several stages in the processing pipeline, output files are created for quality control and manual inspection. An example of this is the spin system master file (spin_system_master.json) which is created to summarize spin systems found in the

sample batch which displays the derived spin systems from all samples, the resonances assigned to each spin system, and the originating sample ID. The flat files produced contain essential data for MADByTE networking and are used by the downstream processing steps. Additionally, a log of every processing step and interaction is created as MADByTE generates features and networks, allowing for the manual follow-up of a particular resonance of interest by logging filtration and alignment steps explicitly.

2.3.2. Supervised Peak Picking

Investigations into automated peak picking were conducted to find a plausible method for data processing automation. Although peak picking is a relatively straightforward aspect of NMR spectroscopy, the nuances of how this is done across various platforms and in different sample conditions/types differ considerably with each algorithm. In general, most peak picking algorithms determine the validity of a peak by evaluating peak shape, peak maximum intensity, the standard deviation of the noise level of the spectrum, resolution between peaks, or a combination of several of these metrics. In pure compounds, these metrics are straightforward, and the majority of data will pass each of these checkpoints without issue or the need for major revision from an investigator.

In multidimensional NMR of complex samples, peak picking in an automated fashion is a known issue, as peak overlap is a more significant problem in crude or simplified extracts than it is with pure compounds.[65] Vendor supplied peak picking algorithms (MNova and Bruker) were trialled for their ability to pick peaks of sufficient intensity, plausible peak shapes, and their capacity to deal with overlapped regions of the spectra. In each case, exceptions to one or more of these criteria were found, often creating issues in the downstream processing steps.

Spectra containing considerable T_1 noise, which is common in samples with a pronounced solvent peak and/or water peak, demonstrate the complications of peak picking algorithms. The regions of noise caused by these signals are often considerably higher than the standard noise level of the rest of the spectra. If a generic noise “floor” is selected, these extensively noisy regions can quickly confound the peak picking algorithm and generate peak lists which overwhelmingly represent the noise region (Figure 2.5).

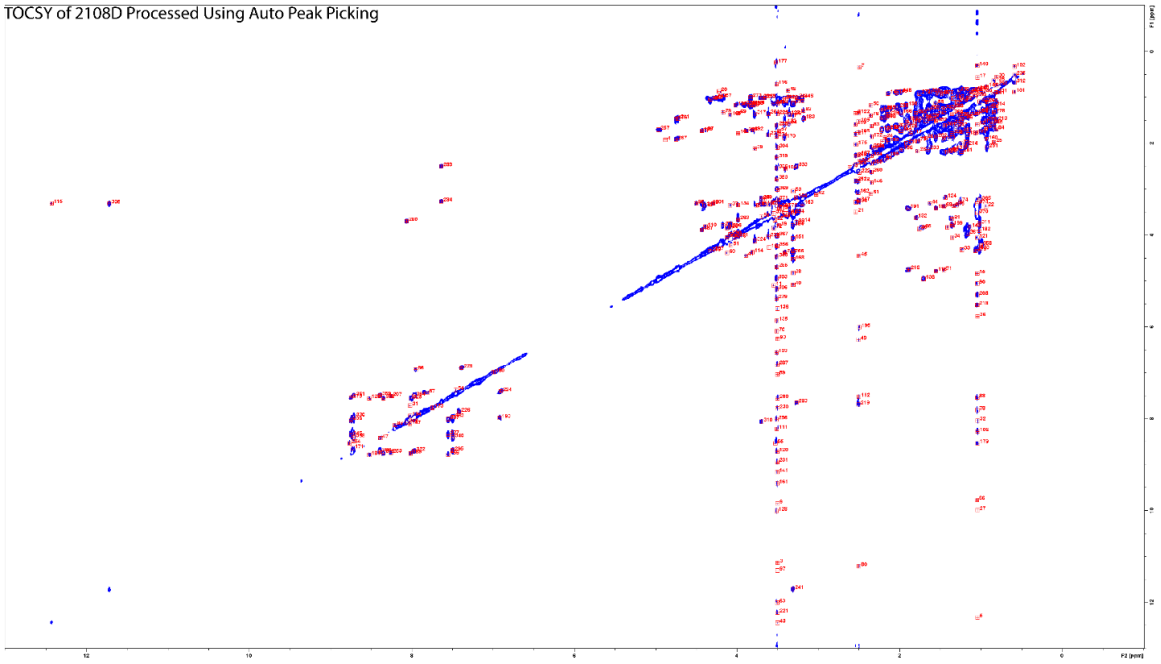


Figure 2.5. TOCSY of Extract Prefraction RLUS 2108D Processed Using Automated Peak Picking.

One method of reducing the number of erroneous peaks selected by peak picking algorithms is to set an upper threshold for the number of peaks to be picked. However, an additional complication can arise from the dynamic range of the data each spectrum contains. Major metabolites are often present at considerably higher levels than minor metabolites, which may be of interest, and regions of the spectrum which are not as susceptible to noise or peak overlaps can contain legitimate resonances at a threshold lower than that of noise in another part of the spectrum (such as the T_1 noise regions). This causes the algorithms to become overwhelmed with peaks which may be erroneous in one part of the spectrum, while ignoring real data in others.

Due to these complications, it was determined that the development of a robust processing scheme should predicate its use on data which has undergone supervised peak picking (Figure 2.6). This ensures that the data used to construct, and network features is built on the confidence of the researcher that a given signal is real and reduces the chances for false representation of the data. However, as more robust automated peak picking algorithms are under constant development and refinement, future applications of MADByTE analysis could see this recommendation relaxed.[66,67]

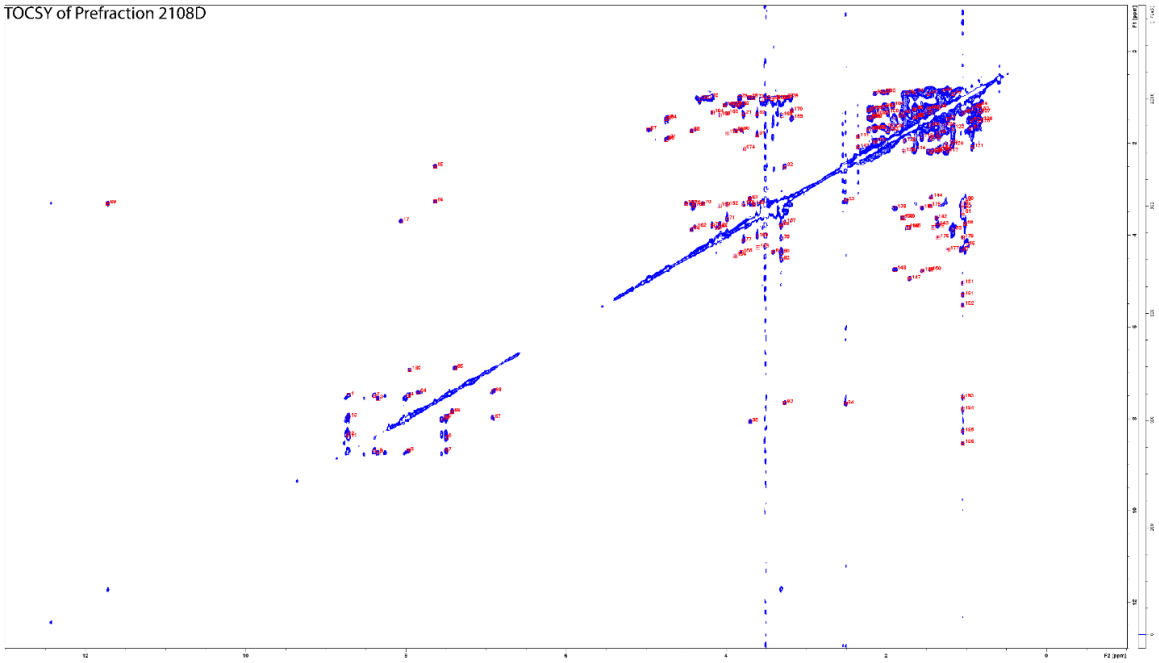


Figure 2.6. TOCSY of Prefraction RLUS 2108D Processed Using Supervised Peak Picking.

2.3.3. HSQC Data Filtration and Preprocessing

HSQC data are first filtered to remove peaks in the solvent and water signal regions, as some datapoints may be carried forward despite the requirement for supervised peak picking. In addition, regions with very high disparities between the relative ^{1}H and ^{13}C values are excluded (Figure 2.7 regions 1, 2, and 3 and tabulated in Table 2.2).

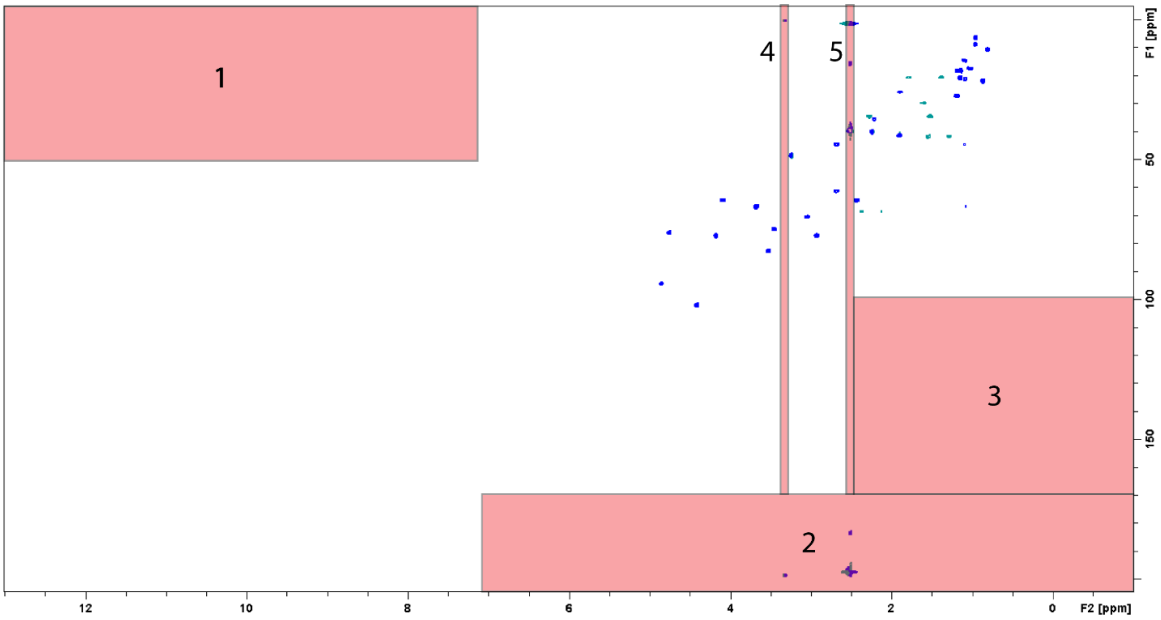


Figure 2.7. Graphic Representation of Data Filtration Regions for HSQC Data

Table 2.2. Tabulated Data Filtration Regions for HSQC Data

Proton Range (ppm)	Carbon Range (ppm)	Reason	Zone
2.48-2.52	All	DMSO T_1 Noise	5
3.28-3.32	All	H_2O T_1 Noise	4
0.0-2.4	100.0-210.0	Extraordinary Shift Disparity	3
0.0-7.0	170.0-210.0	Extraordinary Shift Disparity	2
7.0-13.0	0.0-50.0	Extraordinary Shift Disparity	1

Addressing the Retention of Multiplet Structure in Metabolomics Data

An additional consideration in the filtration of HSQC data is the efficiency of the decoupling between ^{1}H and ^{13}C. Although HSQC spectra are optimally single resonance datapoints, some multiplet structure is retained without implementation of pure shift experiments.[68] Although pure shift variants of the HSQC exist and are widely used, their overall sensitivity loss reduces their effective use in metabolomics studies.[69] To address this, after initial data filtration HSQC data points are queried for residual multiplet structure by finding resonances which could plausibly be originating from single resonances split by coupling. This is done by finding resonances within defined ^{1}H and ^{13}C ppm error tolerances. ^{1}H tolerance recommendations can be made through relation of the Karplus plot and equation with respect to a 3 bond HCCH coupling through Equation 2.2.

$$^{3}J_{ax}(\theta) = A\cos^2\theta + B\cos\theta + C \qquad \textbf{(2.2)}$$

Equation 2.2 describes the relationship of the coupling constant in Hz ($^{3}J_{ax}$ (θ)) obtained from the relationship of the dihedral angle (θ) between any two protons which display three bond coupling. Plotting the relationship for ethane (Figure 2.8), the maximum expected coupling due to bond angles of ethane would be around 15 Hz.[70,71] In a 600 MHz magnetic field, this translates to a ^{1}H tolerance of 0.025 ppm; since coupling can be either a positive or negative relationship, points within ^{1}H 0.05 ppm may be due to imperfect decoupling.

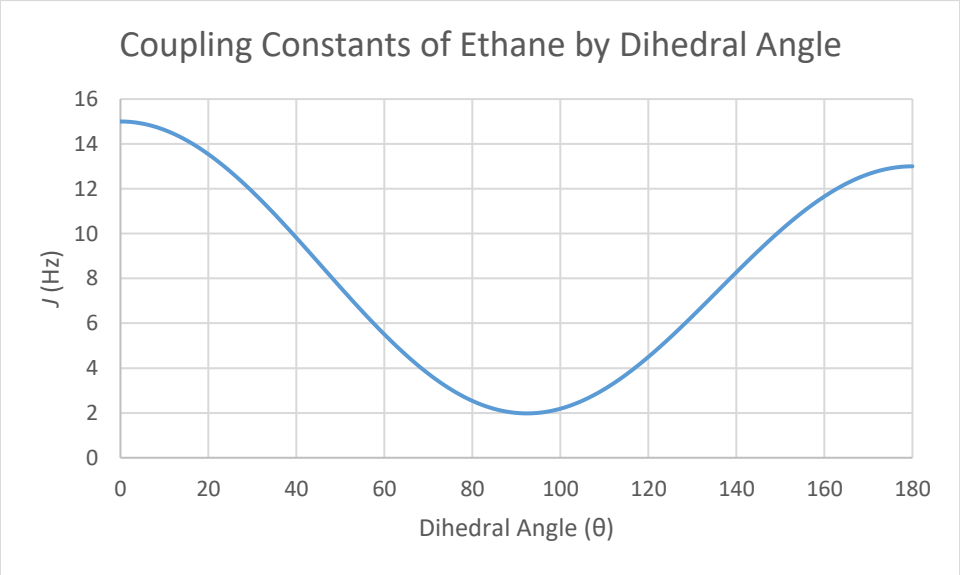


Figure 2.8. Coupling Constant Relationship as a Function of Dihedral Angle Between Two Nuclei Coupled Through 3 Bonds in Ethane.

In addition to the ^{1}H and ^{13}C tolerances, HSQC data provided in phase-sensitive mode contain important information as to whether signals could plausibly be originating from coupling relationships. For instance, if two points are within these tolerances but display differing phase (I.E. + and -), they cannot be originating from the same resonance. Points satisfying these requirements are merged, centered, and treated as single resonances for downstream comparisons.

2.3.4. TOCSY Data Filtration and Preprocessing

The homonuclear data acquired from TOCSY experiments contains important information about chemical environments and coupling information between resonances. However, resolving peaks with similar chemical shifts and extended spin systems can be challenging due to peak overlap. Left unaddressed, areas of high overlap can create false associations, which extend the TOCSY spin system feature with erroneous correlations.

TOCSY data is initially filtered by removal of the shielded region in both dimensions (<2.50 ppm), as the overlap in this region is too great to permit accurate peak picking in most cases. Fortunately, data points in this region are often incorporated through their association with downfield correlations in F2, improving spectral coverage. Points identified as self-correlations are also excluded from analysis to reduce inflation of the data. Each TOCSY proton signal is then queried to see if a matching proton signal exists within the ^{1}H ppm tolerance in the HSQC spectrum. If no HSQC data point is found, the correlation is removed. By requiring the data points to be present in both the processed TOCSY and HSQC data, false data points arising from reconstruction of non-uniform sample data as well as random noise are eliminated from further analysis.

Homonuclear experiments such as TOCSY and COSY can be artificially symmetrized through vendor supplied algorithms. While this may seem advantageous, the most symmetrisation functions introduce the risk of artefact creation or signal splitting when comparing F1 and F2– both of which can complicate the formation of trustworthy spin system features. This is especially pronounced in complex data, such as with extracts and prefractions for which the MADByTE platform was designed. To avoid these complications, symmetrisation is not recommended.

TOCSY spectra collected in DMSO-d_6 may contain valuable information about extended chemical motifs through the detection of exchangeable protons found on OH or NH groups. However, as these signals will not have associated HSQC coordinates, they are dropped from consideration in the assembly of spin system features. With the current framework of the MADByTE processing scheme, inclusion of these signals poses a difficult data filtration challenge, as they may be indistinguishable from random noise. However, as their extended spin system will likely contain other ^{1}H coordinates which are

detectable in the HSQC, the remainder of these spin systems will remain intact as long as they satisfy the remaining filtration requirements.

After data filtration, remaining TOCSY data points are aligned by merging all reported F2 peaks occurring within 0.01 ppm, allowing for small variances and peak picking errors. Points in F1 are then aligned to points in F2 within 0.02 ppm and data points outside of this margin are re-queried and merged within a margin of 0.01 ppm. This process is designed to take advantage of the increased resolution in the F2 dimension for alignment, eliminating cases of resonance duplication. An overview of the process can be seen in Figure 2.9.

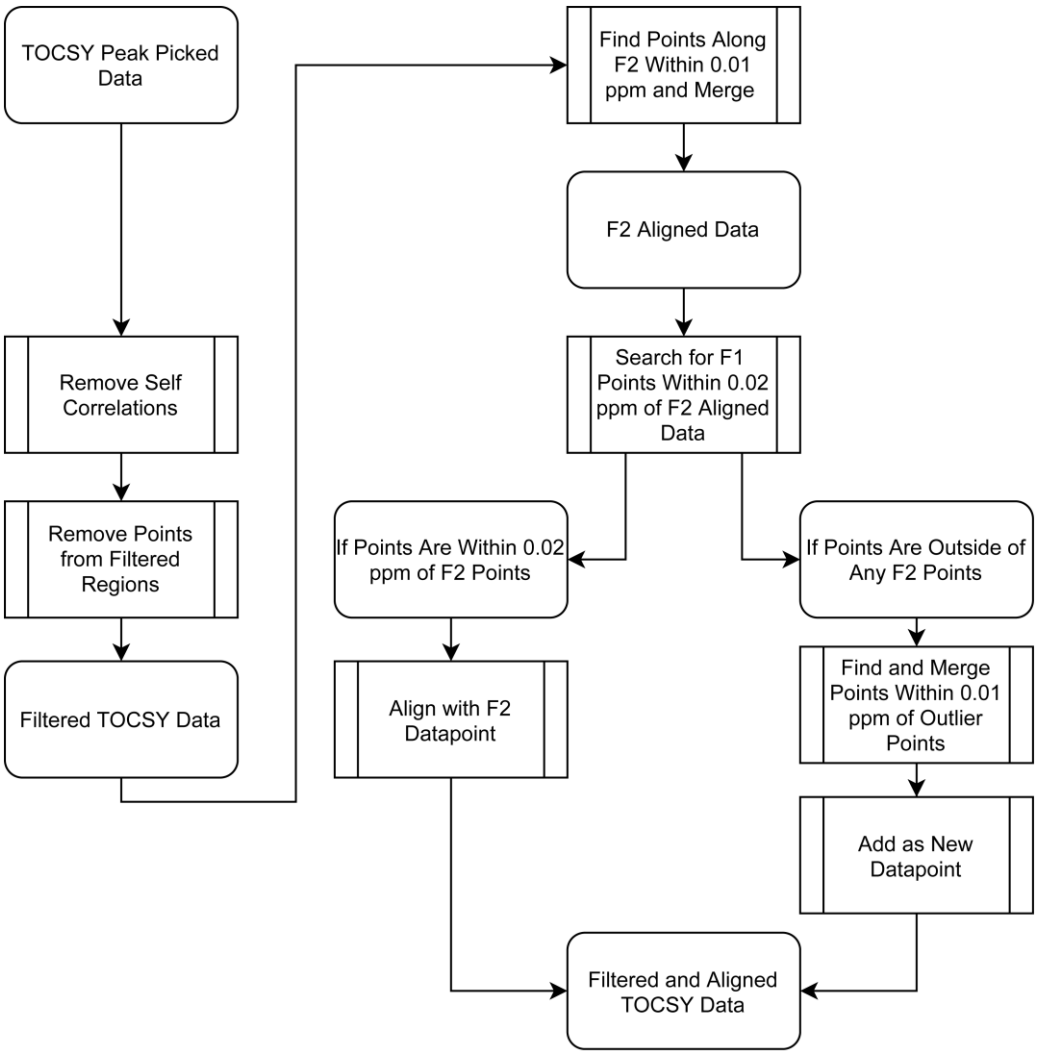


Figure 2.9. TOCSY Data Filtration and Alignment Steps

2.3.5. Feature Construction and Comparison

Following the preprocessing of both the HSQC and TOCSY data, datapoints are associated between the experiments. Initially, the TOCSY signals are compared against the HSQC peak list, and TOCSY signals without corresponding HSQC cross peaks in the ^{1}H dimension are excluded. This is done as an additional confidence step, as points arising in the TOCSY which are not found in the HSQC are either noise, generated artefacts from spectrum reconstruction, or not directly attached to ^{13}C – and therefore would not be useful in the cross comparison of these data. Therefore, any further comparison is done only between datapoints which have been identified in both spectra, have not fallen into typically noisy regions, and are of a high degree of confidence.

Spin system features for individual samples are created by generating a directed graph from each TOCSY peak table, where nodes represent ^{1}H signals in the TOCSY spectrum, and edges represent TOCSY cross peaks between ^{1}H signals. Because multiple members of each spin system should generate cross peaks to any given spin system member, nodes containing only a single connection to a spin system are removed. This requires valid resonances to be reciprocal (i.e. observed on both sides of the diagonal of the TOCSY spectrum).

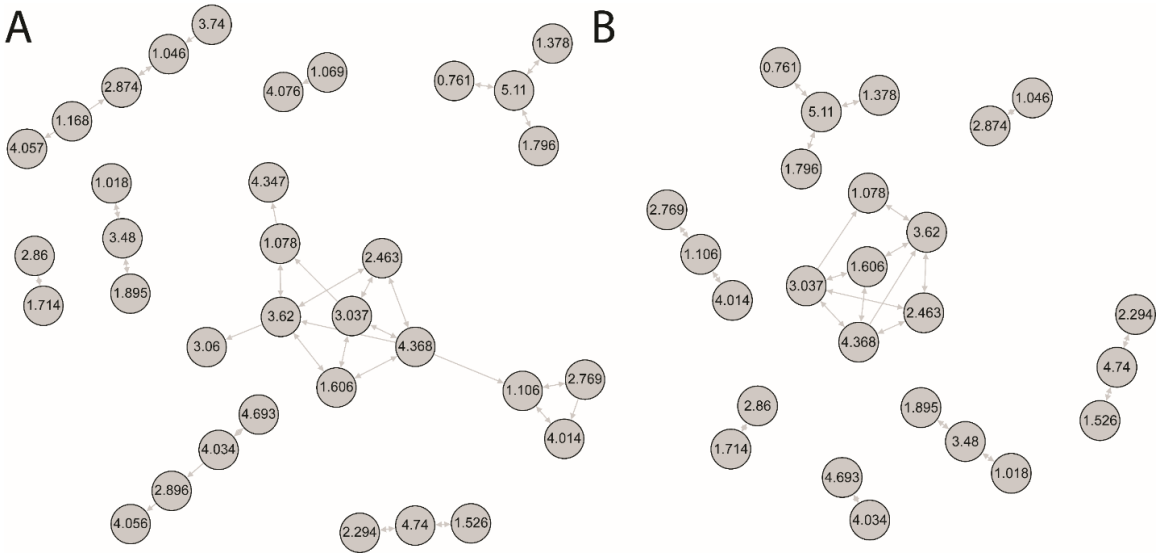


Figure 2.10. Spin System Construction from TOCSY Data of Erythromycin

The resulting graph includes sub-graphs for every unique spin system in the sample. Nodes in these sub-graphs are annotated with ^{13}C chemical shifts to form (^{1}H, ^{13}C) pairs

by integration of the HSQC peak table data. In instances where spectral overlap yields multiple candidate ^{13}C chemical shifts, values of the closest HSQC resonance are included in the node annotation. In the event that two resonances from the HSQC are equidistant to the TOCSY resonance, both carbon assignments are added as points in the spin system as they cannot be resolved. Resulting spin system features are stored and used in the following step to determine the similarity of spin systems from different samples.

Once the spin system construction step is complete, the spin systems must be compared to each other to find the amount of potential overlap between any two systems. This was accomplished by establishing a similarity ratio to determine the overlap. Each pairwise combination is scored for (^{1}H, ^{13}C) pair overlap by dividing the number of overlapping (^{1}H, ^{13}C) pairs by the total number of (^{1}H, ^{13}C) pairs in the spin system.

To calculate the similarity between two spin systems s and s', the analysis considers each spin feature independently. A spin feature consists of a proton resonance signal with a set of one or more carbon signals. The similarity ratio is then given by Equation 2.3 where the intersection of two spin systems is determined by finding the number of spin features in system s' which overlap with the features in system s.

$$Similarity(s, s') = \frac{length(s' \cap s)}{length(s)} \qquad \textbf{(2.3)}$$

Two user defined parameters are employed here; the ^{1}H error tolerance and the ^{13}C error tolerance, which default to 0.05 and 0.4 ppm respectively. Two spin features are said to overlap if a proton is found within the given tolerance, provided that proton has at least one carbon which also matches within the given tolerance. As an example, if we have two spin systems [where each spin feature is noted as $(^{1}\mathrm{H}, (^{13}\mathrm{C}_1, ^{13}\mathrm{C}_2, \ldots))$]

$$s = \{ (1.50, (13.0, 30.2)), (3.01, (20.1)), (4.44, (55.5)) \}$$

$$s' = \{(1.51, (25.9)), (1.56, (28.9))\}$$

Then the length of the intersect of s' in s is 1, because the first spin features of each system match within tolerance. This would give a similarity ratio of $1/3$, or 0.33. The similarity ratio takes values between 0 (when there is no overlap of two systems) to 1

(when systems are identical or overlap perfectly). For a MADByTE experiment, the similarity ratio between each pair of spin systems is computed and stored in a square, non-symmetrical matrix which records the overlap of spin system A with B in one dimension, and B with A in the other. The higher of the two similarity scores (A vs B or B vs A) is used to define edges between spin system nodes. This approach is appropriate because variation in compound concentrations or resolution between samples can lead to the creation of larger or smaller spin system features for the same molecule in different samples.

2.3.6. Network Visualizations

MADByTE illustrates the chemical interrelatedness of sample sets by creating network graphs that include all spin system features and include connections (edges) between features with similarity scores above a minimum threshold. These networks are used to identify interrelated spin system features between samples, which can be used to either gauge the amount of structural similarity between samples, or as a feature for downstream dereplication. In all views, extracts are represented by large nodes (red nodes in Figure 2.11), but each representation allows for different aspects of the overall network to be highlighted.

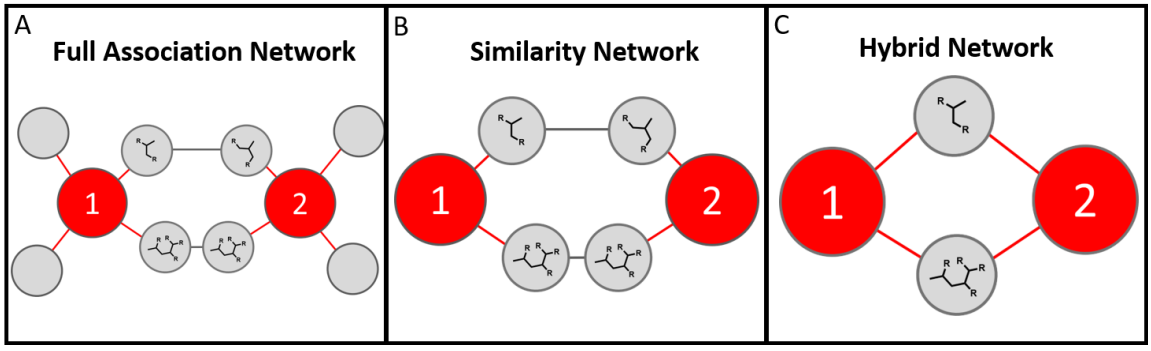


Figure 2.11. Conceptual Representations of Node Relationships from Each of MADByTE's Network Outputs

Full Association Network

The full association network (Figure 2.11 - Panel A) is a representation of every SSF node derived from every sample. This allows for nodes that share similarity to be displayed alongside nodes which house unique SSFs which are not found elsewhere in a sample set. When small sample sets are queried, this provides a comprehensive view of the

chemical environments found in each sample. The full annotation network can be used to highlight the uniqueness of each spin system, providing a strategic vantage point for novelty driven explorations of the network. However, due to the inherent complexity of large networks, the visual representation of the FAN can quickly become problematic.

Similarity Network

The similarity network (Figure 2.11 - Panel B) is a method of reducing the complexity resulting from MADByTE analysis by highlighting only SSFs which are found elsewhere in the sample set. This view reduces the number of datapoints visible by dropping the unique SSFs which can overwhelm visual analysis of large sample set comparisons. Nodes which share no connections at all are dropped from visualization, so as to reduce the overall complexity. The similarity network provides a good vantage point for studies where shared chemistry is the focus, such as compound dereplication or metabolomics co-occurrence studies.

Hybrid Network

The hybrid network (Figure 2.11- Panel C) was constructed to display the relatedness of each SSF node relationship in context of the samples with homology identified. The construction of the hybrid network involves merging independent SSFs through their shared chemical profiles. To achieve this, each set of SSF resonances from associated nodes are compared and only points found in each are retained, dropping unique resonances from consideration. Points found to be within the ^{1}H ppm and ^{13}C ppm error tolerances are averaged as new points. As the new node does not represent the totality of the SSFs, new nodes are created to represent these relationships which replace the presence of discreet SSF nodes in the network – providing a condensed node network by comparison to the similarity network.

2.3.7. Graphic User Interface

Perhaps one of the biggest hurdles in adoption of a new utility or analysis pipeline is the ease of installation and use for end users.[72] Many metabolomics software packages have been developed for community adoption, but few contain intuitive user interfaces or ample documentation for users to tailor the processing steps or resulting outputs to the context of their individual studies. In many cases, installation of the software itself can present

major challenges, often placing a burden of technical expertise on the user before they can begin using the tool. To address this, MADByTE was designed with user definable parameters – clearly documented in the user manual, and facile installation instructions for both Windows and Mac. These features were specifically designed to facilitate ease of use and rapid adoption of the utility. Consistent with this philosophy, MADByTE is provided under the MIT Software license and is available for free as a code repository from https://github.com/liningtonlab/MADByTE.

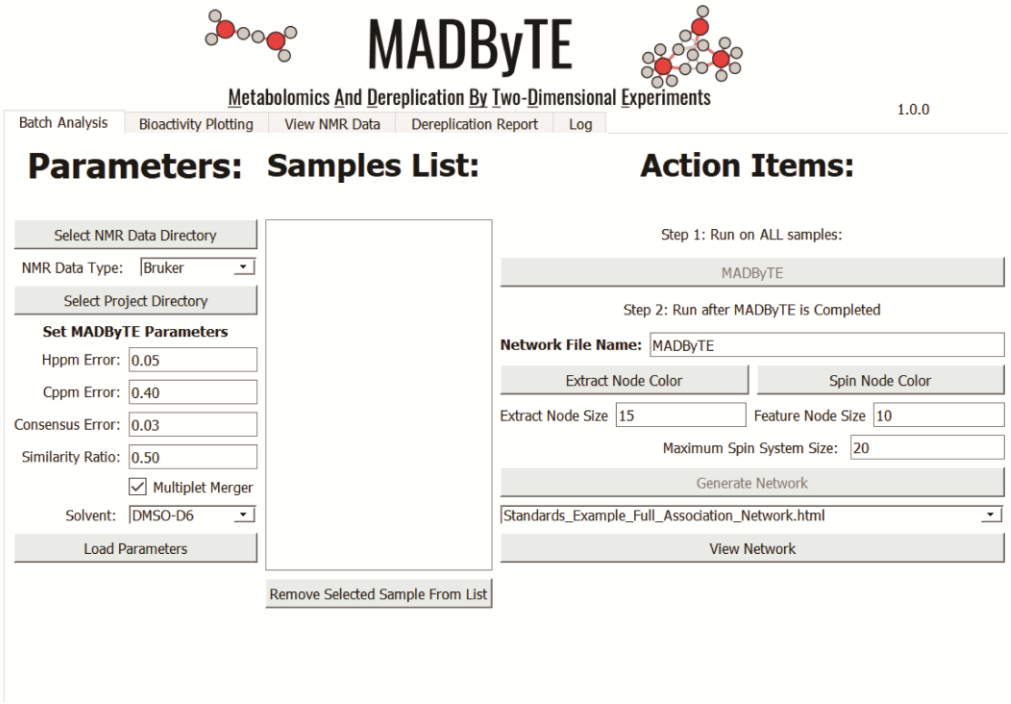


Figure 2.12. Screenshots of the MADByTE Graphic User Interface (GUI). The batch analysis menu contains all the user adjustable parameters such as the ^{1}H ppm and ^{13}C ppm cutoffs and allows for solvent and data type selections.

Flexibility of Input Data

Because MADByTE accepts post-processed peak lists rather than raw fids, users can customize processing parameters to fit their experimental design. This allows control over peak picking thresholds, linear prediction, zero filling, apodization function selection, and even application of advanced processing methods such as covariance NMR.[73] Additionally, as users may use different peak picking strategies available to them, using peak picked lists as the input data allows for users to be in full control of the input data, and affords downstream compatibility for peak picking or processing utilities yet to be

developed. Importantly, this also removes dependencies associated with vendor-specific software, affording access to a wider user base.

Nearly all aspects of data filtration in the HSQC and TOCSY preprocessing steps are directly accessible to end users through the GUI, with default values provided as starting points. This enables users with limited access to high resolution systems to utilize the utility by establishing cut-off values appropriate for the resolution of their spectra.

Interactive Displays

The MADByTE GUI provides utilities which enable interaction with the results of the processing in an intuitive manner. Because MADByTE is constructed on top of a Python-based framework, data can be quickly parsed in customized functions and displayed to the user in real time. To achieve this, MADByTE constructs node networks using networkX[74] and passes the resulting network and associated metadata to Bokeh[75] to generate an interactive data display (Figure 2.21). As a user hovers over a given spin system feature node in the network viewer, the membership of the feature is displayed in real time, providing a snapshot of real datapoints associated with abstract features used for comparison. Combined with filtration of the spin system network based on the number of points in each feature, these interactive networks allow for users to highlight and investigate regions of interest and derive features for prioritization within seconds of processing.

In addition to network views, MADByTE also contains native NMR plotting for spectral review, including options for viewing both 1D spectra and points derived from HSQC and TOCSY processing (Figure 2.22). ^{1}H NMR plotting is done using NMRglue,[76] a powerful Python-based utility for processing of NMR data, which retains important plotting information such as peak shapes and relative intensities in high fidelity.

2.3.8. Dereplication Module

Although the MADByTE analysis pipeline was constructed to address the shortcomings of standard dereplication approaches to annotation in complex mixtures, dereplication as a concept is a robust and repeatedly validated methodology in natural product research and discovery. To facilitate dereplication within the GUI environment of MADByTE, a module was constructed that allows for HSQC spectra to be queried against an in-house library

which users can establish, depending on their individual requirements. In many cases, dereplication of known chemistry can extend beyond the scope of publicly available databases.

NMR based dereplication is often performed by deriving the peak list from a sample and comparing against a database of known metabolite profiles, often specialized for a given sample type or experimental setup. Although this methodology has been shown to work in complex samples, assignments are often made on only a few key correlations rather than the entire molecular profile of the metabolite. Multiple resonances of interest found within a sample can increase confidence of assignment, but regions of high complexity caused by peak overlap may confound these approaches to annotation and are common in metabolomics studies.

Dereplication in MADByTE is done via point-to-point HSQC comparison of the datapoints from a given sample against the database references for entire molecules. The 2D nature of the HSQC spectra is advantageous as it resolves the resonance information against an additional axis, determined by the chemical shift of the directly attached ^{13}C. In addition to the increased resolution, the chemical shift of the ^{13}C nuclei provides an additional layer of confidence in assignment, as ^{1}H nuclei alone may share the same chemical shifts, despite being attached to ^{13}C nuclei which are exposed to different chemical environments. These resonances, when taken together represent a robust feature for molecular identification and provide considerably more context than either a 1D ^{1}H NMR or ^{13}C NMR can provide for a given molecule.

As MADByTE derives the resonance information from peak picked lists in the first processing steps, this information is readily available to the platform to perform dereplication. Scoring of resonance matches are performed by finding (^{1}H,^{13}C) coordinate pairs in the sample matrix which are within the defined ^{1}H ppm and ^{13}C ppm error tolerances defined. Each sample is then scored by the number of peaks found in the spectrum vs the number of peaks expected from the reference compound in the dereplication library and displayed as a table in the GUI for review.

2.4. Applications of MADByTE

2.4.1. Proof of Principal: Application to Standard Compounds

To test the effectiveness of MADByTE for grouping compound classes, a training dataset comprising ^{1}H, TOCSY and HSQC spectra for 17 commercially available natural products and natural product analogues was acquired (Figure 2.13). Following supervised peak picking, peak lists were imported into MADByTE and processed as described above. An excerpt from the resulting similarity network is presented in Figure 2.14, which demonstrates the ability of the platform to network similar compounds together through spin system features. The full network can be found in Section 2.8.2.

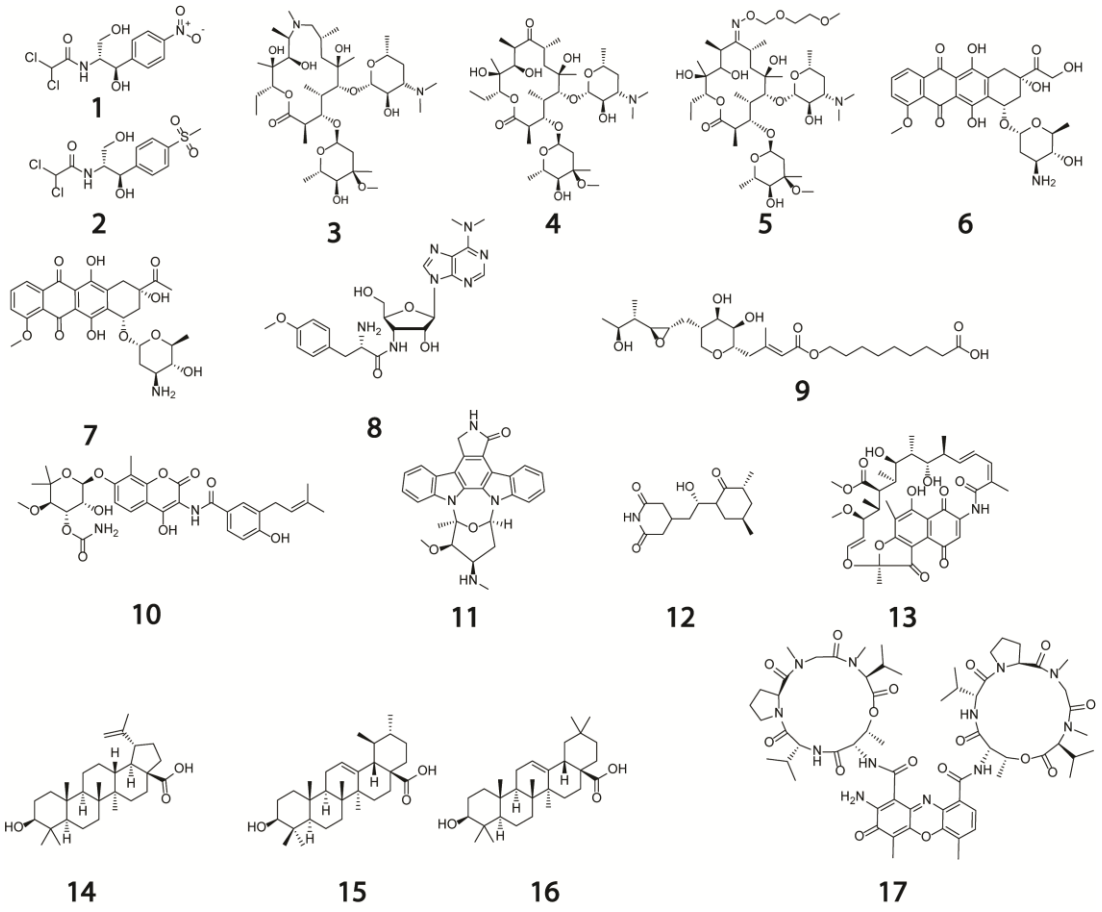


Figure 2.13. Standard Compounds Used for MADByTE Development. Compounds were chosen to represent natural products from several structural classes common to natural product investigations.

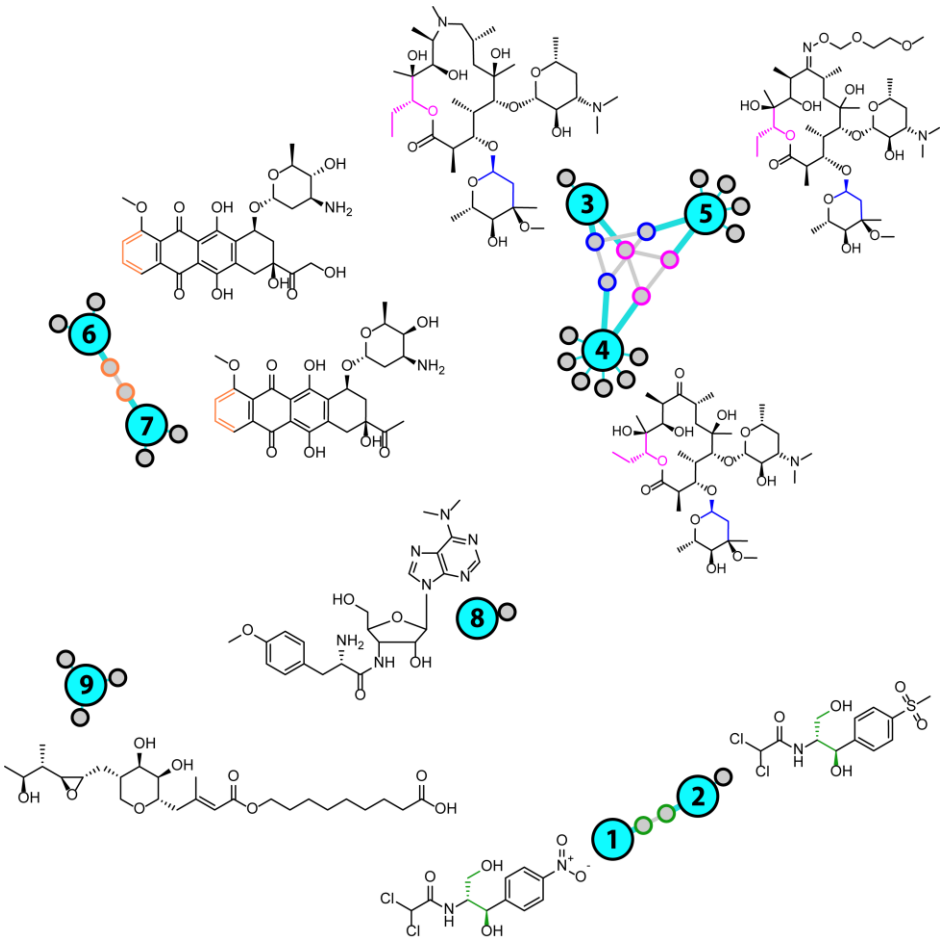


Figure 2.14. Condensed Full Annotation Network of 9 Commercially Available Compounds Used for MADByTE Development. Colored nodes can be mapped to their structural components (matching colors) through manual inspection and comparison against reference data.

The global network contained one sub-cluster (central cluster) containing three reference compounds. Closer examination of this sub-network revealed the presence of the related polyketide macrocycles azithromycin (**3**), erythromycin (**4**) and roxithromycin (**5**). These compounds were related by the presence of two major spin system features shared between all three compounds, analyzed in detail in Figure 2.15 and Table 2.3.

Review of the network (Figure 2.14) revealed other matching compound sets, including chloramphenicol (**1**)/ thiamphenicol (**2**), and epirubicin (**6**)/ daunomycin (**7**). Encouragingly, compounds with low structural similarity to other members of the training set did not form connections to these clusters. Instead these compounds (e.g. puromycin (**8**) and mupirocin (**9**)) remained as single sub-networks containing only the spin system

features identified from their own NMR spectra, without false-positive connections to other members of the test set.

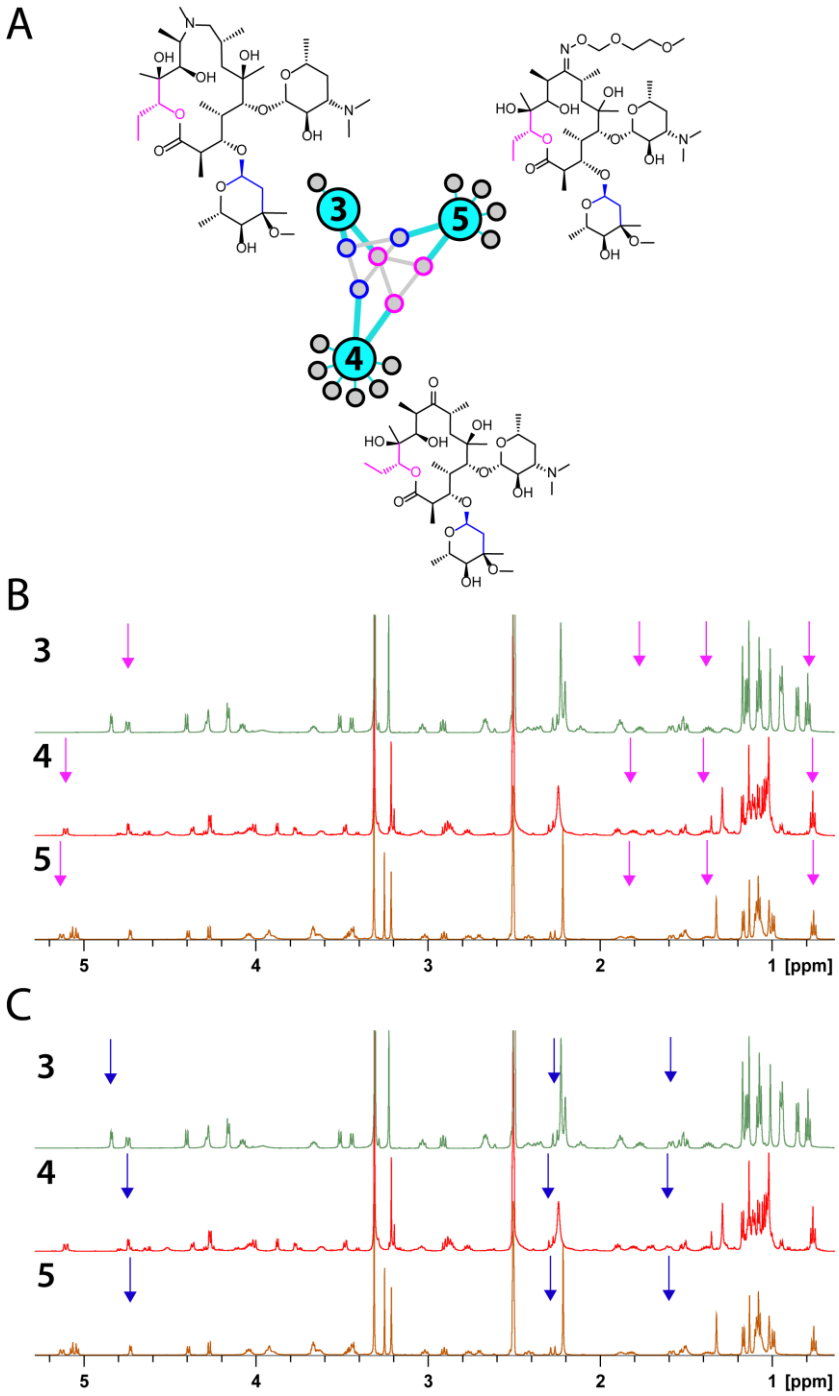


Figure 2.15. **Analysis of Macrocycle Cluster. Shared spin systems (colored node borders) were mapped back to common structural elements (corresponding color) by comparison to published assignments. For example, the central cluster of macrocyclic compounds azithromycin (3), erythromycin (4), and roxithromycin (5) contains spin systems from the cladinose sugar (blue border) and a portion of the macrocyclic core (pink border). Analysis of the proton assignments for these motifs show that ^{1}H directed methods may miss common elements due to bin sizes (Panels 3 and 4).**

Table 2.3. Spin System Features From Macrocyclic Compounds

Spin System	Members	Compound	Node Border Color (Figure 2.15)
Azithromycin_0	(0.80, 10.9),(1.37,20.9),(1.77,20.9),(4.75,76.4)	3	Pink
Erythromycin_0	(0.76, 10.8),(1.38,21.0),(1.80,21.2),(5.11,75.8)	4	Pink
Roxithromycin_0	(0.76, 10.6),(1.37,20.9),(1.80,21.1),(5.13,76.0)	5	Pink
Azithromycin_2	(1.52, 34.7),(2.28, 34.6),(4.84, 94.5)	3	Blue
Erythromycin_6	(1.53, 34.9),(2.29, 35.0),(4.74, 95.9)	4	Blue
Roxithromycin_4	(1.53, 34.9),(2.28, 34.9),(4.73, 95.6)	5	Blue

The first set of features (pink nodes) included signals from the lactone junction of the macrocyclic core, while the second set (blue nodes) contained signals from the pendant cladinose sugar. In the former case the spin-system feature contains four resonances (0.76-0.80, 1.37-1.38, 1.77-1.80, 4.75-5.13) found in each dataset (Table 2.3). These four signals are sufficient to identify this motif as a commonly shared sub-structure. The cladinose substructure includes six (^{1}H, ^{13}C) correlations, present in two discrete spin systems. One of these spin systems, containing three features (1.52-1.53, 2.28-2.29, 4.73-4.84), was detected as a core motif in all three datasets. These results demonstrate the ability of the MADByTE algorithm to connect substructures even in the absence of all possible correlations, and to use these connections to group structurally related molecules from different samples.

2.4.2. Detection of Non-Native Compounds in Complex Matrices

To evaluate the ability of MADByTE to identify metabolites in complex mixtures a set of 9 spiked prefraction samples were prepared (Table 2.4). 25 µL aliquots of each prefraction were taken from the Linington lab extract library, dried, and spiked with 0.5 mg of one of three reference compounds as shown in Table 2.4. These three compounds (novobiocin, mupirocin, and erythromycin) were selected because they are structurally dissimilar to one another, and were known not to occur in these extracts. HSQC and TOCSY spectra were collected for each sample, processed, peak picked, and subjected to MADByTE analysis, including the reference spectra for the pure reference compounds (Figure 2.16). The resulting network shows clear separation of the extract prefractions based on the presence of spiked reference compounds.

Table 2.4. Prefractions Spiked with Reference Compounds.

Sample Name	Extract Prefraction	Spiked Compound	Node ID (Figure 2.16)
1526_A_SPK	1526 A	Erythromycin	A
1526_C_SPK	1526 C	Mupirocin	B
1526_E_SPK	1526 E	Novobiocin	C
1726_A_SPK	1726 A	Novobiocin	D
1726_C_SPK	1726 C	Erythromycin	E
1726_E_SPK	1726 E	Mupirocin	F
1814_A_SPK	1814 A	Novobiocin	G
1814_C_SPK	1814 C	Mupirocin	H
1814_E_SPK	1814 E	Erythromycin	I
Mupirocin	-	-	9
Erythromycin	-	-	4
Novobiocin	-	-	10

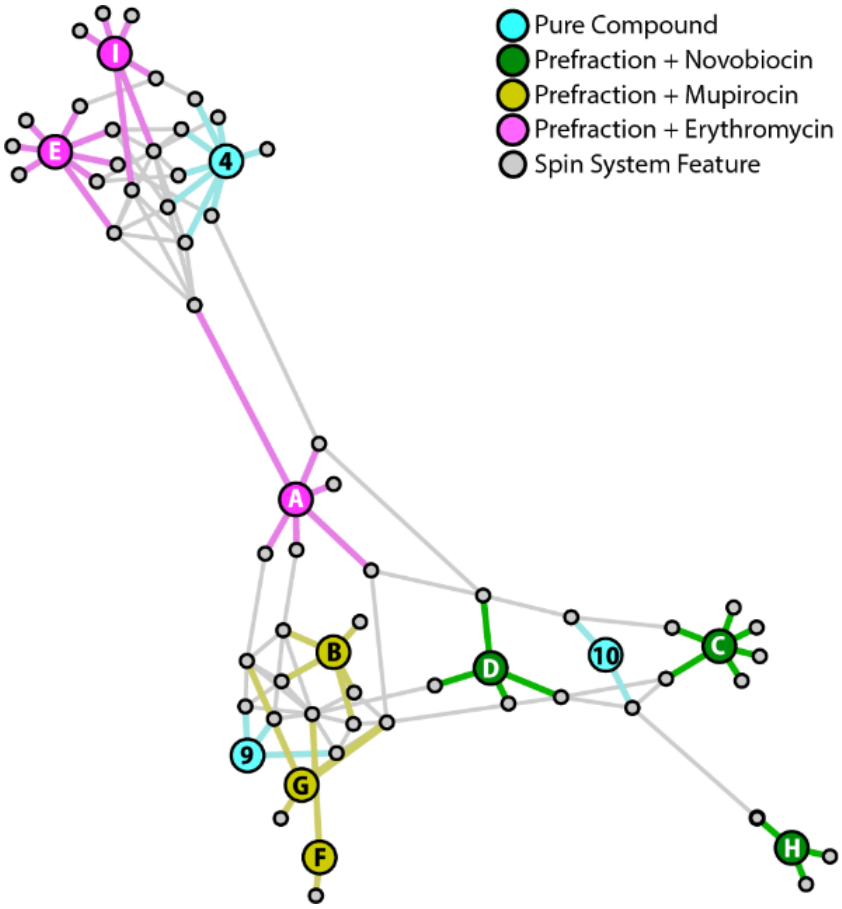


Figure 2.16. Full Annotation Network Illustrating Extract Prefractions Containing Spiked Reference Compounds. Spiked extracts (green, gold, and pink nodes) cluster through spin system features (grey nodes) to pure compound reference data (blue nodes; erythromycin (4), mupirocin (9), and novobiocin (10))

As with all MADByTE networks, nodes in this graph are grouped based on the presence of shared spin system features. Grouping by reference compound indicates that these three extracts have low chemical similarities to one another, with the exception of the reference compounds added in each case. In two of the three cases (mupirocin and erythromycin), clearly resolvable features gave network connections between all three prefractions containing the same compound. In the third case (novobiocin), extensive signal overlap in the HSQC spectra reduced the number of resolvable spin-system features, decreasing but not eliminating the interconnectedness of these three prefractions. Importantly, this study indicated that even in complex matrices, features can be constructed and used to identify scaffold motifs present.

2.4.3. Structural Dereplication in A Natural Product Extract Library

To extend MADByTE to real-world datasets, data for 85 samples from the prefractionated microbial natural products library was acquired. Following data processing to generate spin system features, these samples were combined with the pure compound dataset and the spin system similarity matrix generated using standard parameters.

This yielded a complex network containing a large number of spin system features of varying complexity with respect to membership. To highlight the shared chemistry between these extracts, the resulting hybrid network was analyzed, as it collapses linked features into their common resonances, simplifying the overall layout (Figure 2.18: panel A). Interestingly, several natural products prefractions contained spin system features that linked to reference compounds, suggesting the presence of known compound families. Two extracts in particular, RLUS 1565C and RLUS 1565D showed connection to the reference compound novobiocin, a compound known to be produced by marine actinobacteria, and therefore a plausible metabolite in this sample set. Analysis of the shared features did not account for the entire molecule, but rather, provided a structural hypothesis. Importantly, comparison of the ^{1}H NMR profiles of these extracts was not sufficient to provide definitive confirmation of this molecule within the sample due to peak overlap (Figure 2.17).

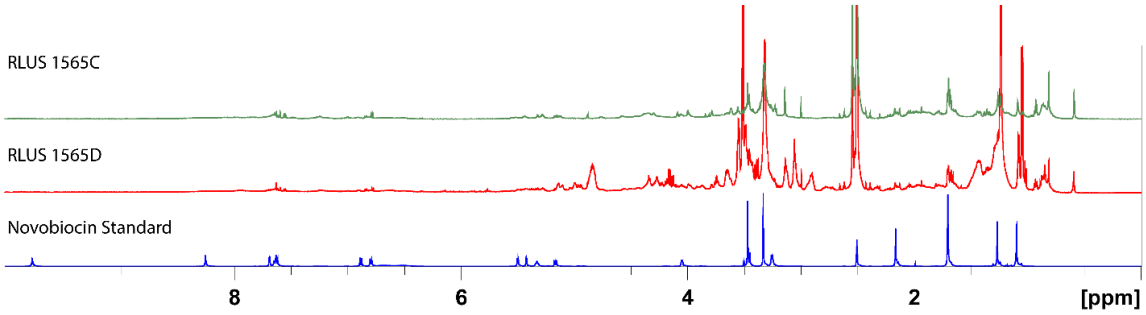


Figure 2.17. Stacked ^{1}H Profiles of Extracts RLUS 1565C and RLUS 1565D Compared to a Novobiocin Reference Standard.

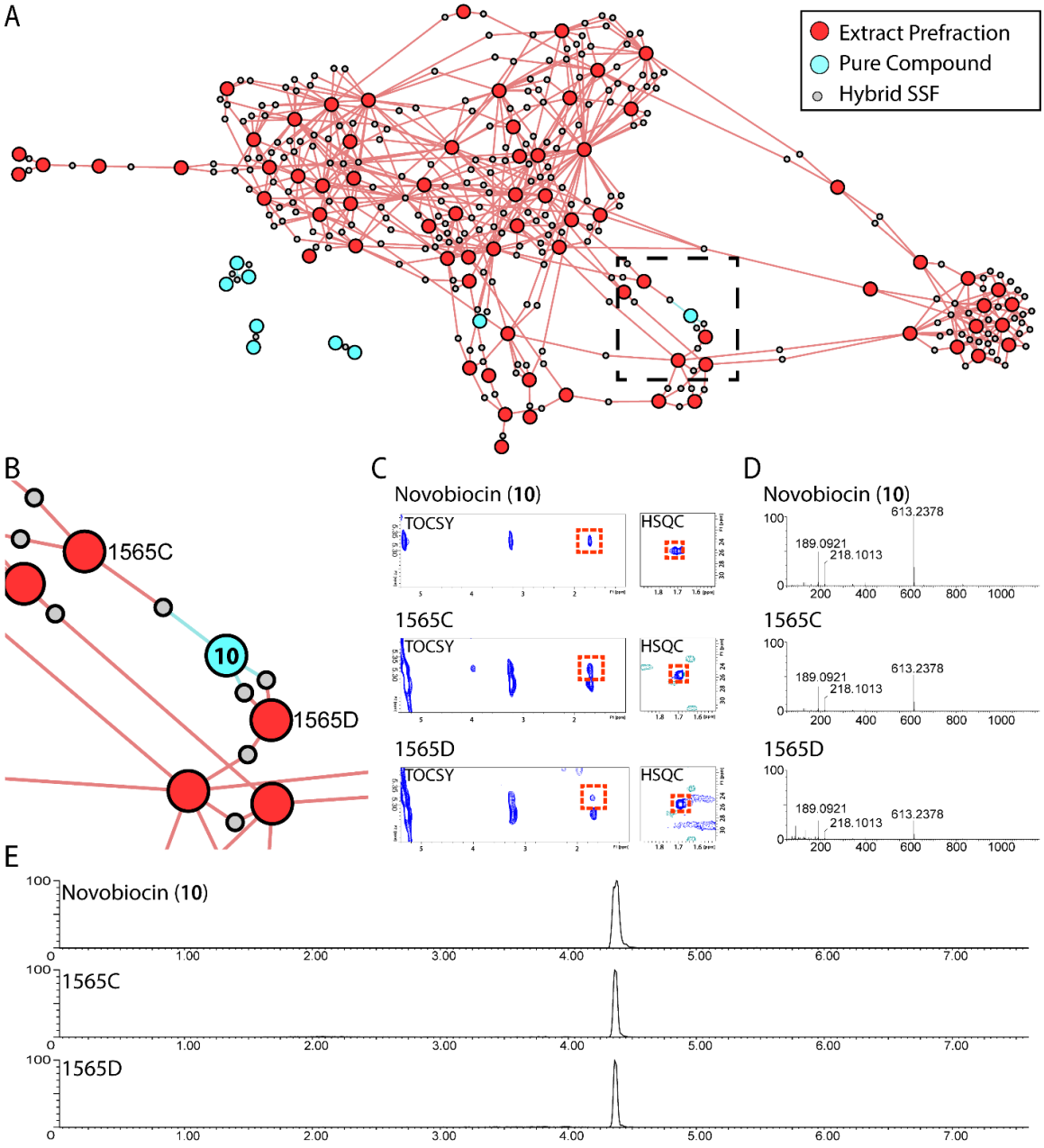


Figure 2.18. **Identification of Novobiocin in Natural Products Library Prefractions. A) Network of 85 extract prefractions and reference compounds. B) Expanded region from panel A showing node connections between novobiocin (10) and extract prefractions RLUS 1565C and RLUS 1565D. C) Expansions of TOCSY and HSQC spectra showing resonances responsible for node connections in panel B. D) HRMS spectra of novobiocin peak at 4.36 min. E) Extracted ion chromatograms for novobiocin (m/z 613.2378) in prefractions 1565C and 1565D, compared to novobiocin standard.**

UPLC-MS analysis of pre-fractions RLUS 1565C and RLUS 1565D (Figure 2.18 panels D and E) confirmed the presence of novobiocin unequivocally through confirmation of retention time and MS profiles, demonstrating the value of MADByTE for compound dereplication through partial structure homologies, even in absence of all resonances from this molecule in the complex extract.

2.5. Limitations of MADByTE Analysis

Although the findings from these experiments demonstrate a variety of utilities and applications of the MADByTE platform, several limitations became apparent which should be addressed.

Firstly, spectral overlap in the TOCSY spectrum proved to be a considerable challenge, especially in regions of high complexity such as correlations from 0-2.5 ppm. As addressed in section 2.3.4, the current strategy to mitigate this is to not consider resonance associations which are only shown in this region. However, problems with overlap and closely associated chemical shifts are not unique to this region, and as such, imperfect spin systems are created. In some cases, overlap can cause two independent spin systems to become associated through an erroneous resonance creating a larger spin system feature. This error is caused in the spin system construction step and is currently expected behaviour due to the restrictions in resolution caused by instrument and processing limitations. These large spin system features are still valid features and contain important information which could still be valuable in associations of spin systems across the sample set. Therefore, the asymmetric correlation matrix is important. In cases where a large spin system contains several smaller spin systems combined (A+B), other extracts may create features which represent only one of the smaller spin systems (A or B) and would still allow for connections to be made showing the shared chemical overlap. Innovations in pure-shift NMR, which increases the resolution between points when compared to standard NMR experiments or an increase in the magnetic field strength could also be leveraged to alleviate these problems of spectral overlap in future applications.

The loss of connection information from 0-2.5 ppm is an additional limitation when considering all possible cases of spectral overlap. This region contains information which may be fundamental in the assignment of structural motifs from relatively shielded

molecules. This limitation was observed in the standard compounds network in section 2.4.1 with compounds **14-16** (betulinic acid, ursolic acid, and oleanolic acid – respectively). Although these compounds have very similar scaffolds, the number of positions which contained valid resonances beyond 2.5 ppm were few. In these cases, the loss of connection information in the 0-2.5 ppm region proved to be the majority of the molecular scaffold, and the number of resonances which were above this cut-off threshold differed too greatly to generate linkages in the network. Therefore, MADByTE analysis of molecules containing highly shielded spin systems may yield limited structural information. An increase in the resolution of the overall spectrum – obtainable through a higher field NMR, alternative pulse sequences, or an increase in experiment acquisition time may provide a viable alternative to this filtration mechanism but could prove impractical for widespread use.

An additional limitation is that the data viewpoints afforded by the MADByTE processing pipeline represent shared resonances from different samples but cannot natively attribute those collections of resonances into direct structural motifs. In the case of the macrocyclic compounds described in Figure 2.15, the features derived by MADByTE could be mapped back due to the availability of reference data in the appropriate solvent. This, however, is not a realistic expectation when applied to unknown compounds in complex matrices. A proposed solution to this limitation would be to generate a reference library of compounds with their spin systems fully annotated. Associations to a given annotated motif could then be used to propose plausible structural components, but this is currently impractical due to the limited availability of reference compound data.

2.6. Conclusions

MADByTE has been shown to derive spin system information from TOCSY, create associations of these spin systems to their connected carbon backbones through HSQC, and construct and compare these spin system features across large sample sets. These comparisons can be leveraged for metabolomics, discovery, and dereplication studies and provides a framework for customization through its open-source nature.

The development and design of the MADByTE analysis pipeline reflects the changing face of natural products research. Utilizing the methods for structure elucidation common to natural products research, in conjunction with new comparison methods and visualization

schemes, MADByTE provides an orthogonal viewpoint of complex data which can be leveraged for hypothesis generation, sample comparison, and functional annotation of natural products in complex samples. Technological platforms such as MADByTE pivot the strategies used for de-novo structure elucidation to more complex systems than would be manageable through human analysis, allowing access to otherwise overwhelming data.

The processing steps involved in the MADByTE analysis reflect the challenges for a robust platform to be developed. Data filtering and feature comparison require a careful balance of flexibility to handle real-world data and rigidity to provide a reproducible and useful result. As complex analysis platforms are created to investigate complex data, they must also be designed with consideration towards future implementations and real-world complications.

Metabolomics and dereplication represent two extremely important steps in natural product research by providing new leads and triage methods to streamline downstream studies. Chemical extracts can take weeks to generate, but with new analysis tools focused around describing their constituents in just a few hours, researchers can focus their efforts on samples which provide the best context to test their hypotheses and make new discoveries.

2.7. Experimental Methods

2.7.1. Extract Prefraction Preparation

25 µL aliquots of extract prefractions were retrieved from our previously described actinobacterial library[77], dried via lyophilization, resuspended in 300 µL of DMSO-d_6, and lyophilized again to remove non-deuterated DMSO from the sample. Sample aliquots translated to a variable mass between 4-15 mg of dry material. Shigemi tubes were placed under high vacuum for 30 minutes prior to use to remove water vapor, and back filled with argon. Samples were dissolved in 320 µL of DMSO-d_6, sonicated, and 280 µL was placed into a matched Shigemi tube for acquisition, with care taken to ensure no solid particulate was transferred. Glass pipettes were pulled to greater length and attached to a 1000 µL micropipette for accurate solvent dispensing and transfer.

2.7.2. NMR Acquisition

All NMR spectra were recorded on Avance™ III TCI (600 MHz) or Avance™ III QCI (600 MHz) spectrometers in DMSO-d_6 (CortecNet lot Q0611) at 300K. HSQC spectra were recorded as 32 scans (TD: 4096 x 256), collected by non-uniform sampling at 50% followed by linear prediction and zero filling. TOCSY spectra were recorded as 16 scans (TD: 1024 x 128), collected by non-uniform sampling at 50% followed by linear prediction and zero filling. NUS point spreads were kept consistent between samples to ensure consistency. Proton spectra were recorded as 64 scans (TD: 131 k). All spectra were manually referenced and phased, followed by supervised peak picking.

2.7.3. Standard Compound Network

The reference compound set selected for the standard compound network were chosen to represent diverse scaffolds from natural products or natural product derivatives. Daunomycin (**7**), roxithromycin (**5**), erythromycin (**4**), puromycin (**8**), novobiocin (**10**), and cycloheximide (**12**) were obtained from Sigma-Aldrich (St. Louis, MO, USA). Ursolic acid (**15**), betulinic acid (**14**), and oleanolic acid (**16**) were purchased from Extrasynthese SA (Genay, France). Chloramphenicol (**1**) was obtained from Calbiochem (La Jolla, CA, USA). Azithromycin (**3**) and rifamycin S (**13**) were purchased from TCI (Tokyo, Japan). Thiamphenicol (**2**) was acquired from Spectrum Chemicals (Cardena, CA, USA), and actinomycin D (**17**) was purchased from RPI (Mount Prospect, IL, USA). Mupirocin (**9**) was purchased from AppliChem (Darmstadt, Germany). Epirubicin (**6**) was purchased from MP Biomedicals LLC (Solon, OH, USA), and staurosporine (**11**) was purchased from LC Laboratories (Woburn, MA, USA).

Parameters used for this study are displayed in Table 2.5. The resulting networks were exported in graphML format and processed in Gephi for visualization using the Force Atlas 2 algorithm with default parameters except; spacing = 10, dissuade hubs = True, prevent overlap = True.

Table 2.5. MADByTE Parameters Used for Standard Compound Networking

Parameter	Value
Hppm Error	0.05
Cppm Error	0.40
Consensus Error	0.03
Similarity Ratio	0.51
Merge Multiplets	True
Maximum Spin System Size	40

2.7.4. Non-Native Compound Network

Preparation of Chemical Extracts

Isolates of RL10-484-HV5-A were retrieved from cryogenic storage, streaked on SYP agar plates, and allowed to grow at room temperature for 9 days until robust colonies were observed. Colony isolates from each plate were then transferred to small-scale liquid culture conditions containing 7 mL of SYP media and 3 glass beads to increase agitation. Each small-scale culture was incubated while shaken at room temperature for 3 days. From this, 4 mL of each small-scale culture was transferred to medium-scale conditions containing 250 mL of SYP media and shaken for an additional 7 days. 40 mL of the medium-scale cultures were transferred to large-scale conditions containing 1 L of SYP media, 20 g of washed XAD-16 resin and incubated under agitation for 7 days.

Each 1 L culture was filtered to remove the culture media and the resulting cell mass and resin was stirred in 250 mL of 50:50 CH_2Cl_2:CH_3OH for 1 hour and filtered. The resulting filtrate was dried under vacuum and fractionated on a reverse phase sep-pack C-18 column yielding 6 final fractions – RLUS 1814 X,A,B,C,D,E with each fraction representing an increase in methanol concentration in the eluent, and a 7th fraction, F, representing an ethyl acetate flush. Each fraction was dried under vacuum and resuspended in 1 mL of DMSO for storage. This process was done in parallel for cultures RL10-247-HVF-C and RL10-348-HVF-A, yielding extracts RLUS 1526 X-F and RLUS 1728 X-F respectively.

Each extract was retrieved from storage, thawed, sonicated, aliquoted, and dried as described as above to prepare for NMR analysis. Samples of each compound to be added into the extract were prepared by dissolution of 0.5 mg of each in 300 µL of solvent, which

was then added to each dried extract sample, sonicated, and transferred to a Shigemi tube for data collection.

Parameters used for the detection of non-native compounds in sample extracts can be found in Table 2.6. The resulting networks were exported in graphML format and processed in Gephi for visualization using the Force Atlas 2 algorithm with default parameters except; spacing = 10, dissuade hubs = True, prevent overlap = True. For Figure 2.16, color coding was applied manually.

Table 2.6. MADByTE Parameters Used for Extract Prefractions Containing Non-native Compounds

Parameter	Value
Hppm Error	0.05
Cppm Error	0.40
Consensus Error	0.03
Similarity Ratio	0.30
Merge Multiplets	True
Maximum Spin System Size	40

2.7.5. Natural Product Extract Library Network

Samples chosen for the NP extract library network were prepared as described in section 2.7.1 and data acquisition followed the outlined protocol in section 2.7.2. Following analysis, samples were dried down under vacuum and placed in separate storage to facilitate follow up analysis. Parameters used for the detection of non-native compounds in sample extracts can be found in Table 2.7. The resulting networks were exported in graphML format and processed in Gephi for visualization using the Force Atlas 2 algorithm with default parameters except; spacing = 10, dissuade hubs = True, prevent overlap = True.

Table 2.7. MADByTE Parameters Used for Natural Product Extract Library Networking

Parameter	Value
Hppm Error	0.05
Cppm Error	0.40
Consensus Error	0.03
Similarity Ratio	0.30
Merge Multiplets	True
Maximum Spin System Size	40

2.8. Supplemental Data

2.8.1. Plausible Alternatives to Fourier Transformation in Metabolomics Data

Advances in alternative processing, such as covariance processing, have aimed to increase the practical resolution of homonuclear experiments (such as TOCSY) and was investigated for utility. Covariance processing offers a promising alternative to standard processing, as it provides an increase in the resolution in F1 when compared to the typical fast Fourier transform.[78] To evaluate covariance processing, the TOCSY spectrum from azithromycin (**3**) and an extract prefraction 2108D were processed using the covariance module included in MNova (direct covariance, square root, no filter).[79] Although this worked well for pure compounds, extract data performed poorly (Figure 2.19), introducing 'streaking' patterns and signal artefacts that made it harder to accurately pick relevant peaks from the TOCSY spectrum. For this reason, the design and applied demonstration of MADByTE did not utilize this processing, although this is still a valid avenue of preprocessing available should future experiments or utilization require it.

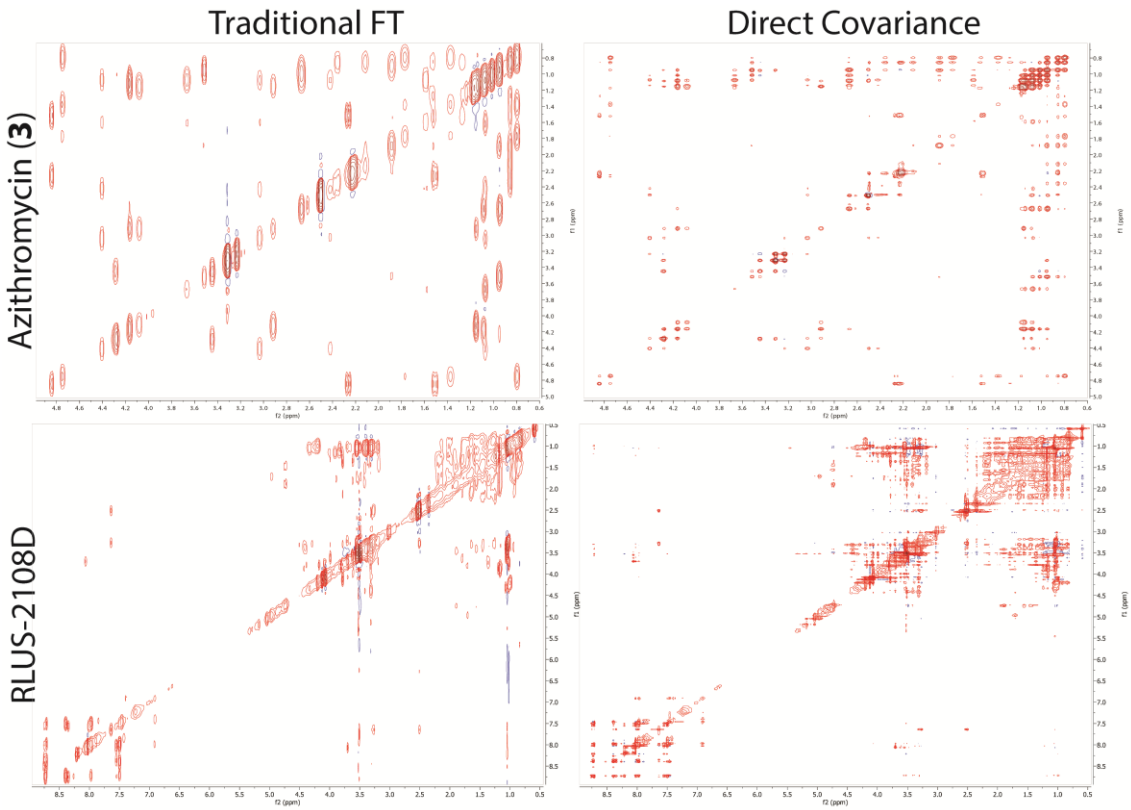


Figure 2.19. Covariance Processing on a Standard Compound and an Extract From the Linington Lab Library.

2.8.2. Full Annotation Network of 17 Standard Compounds

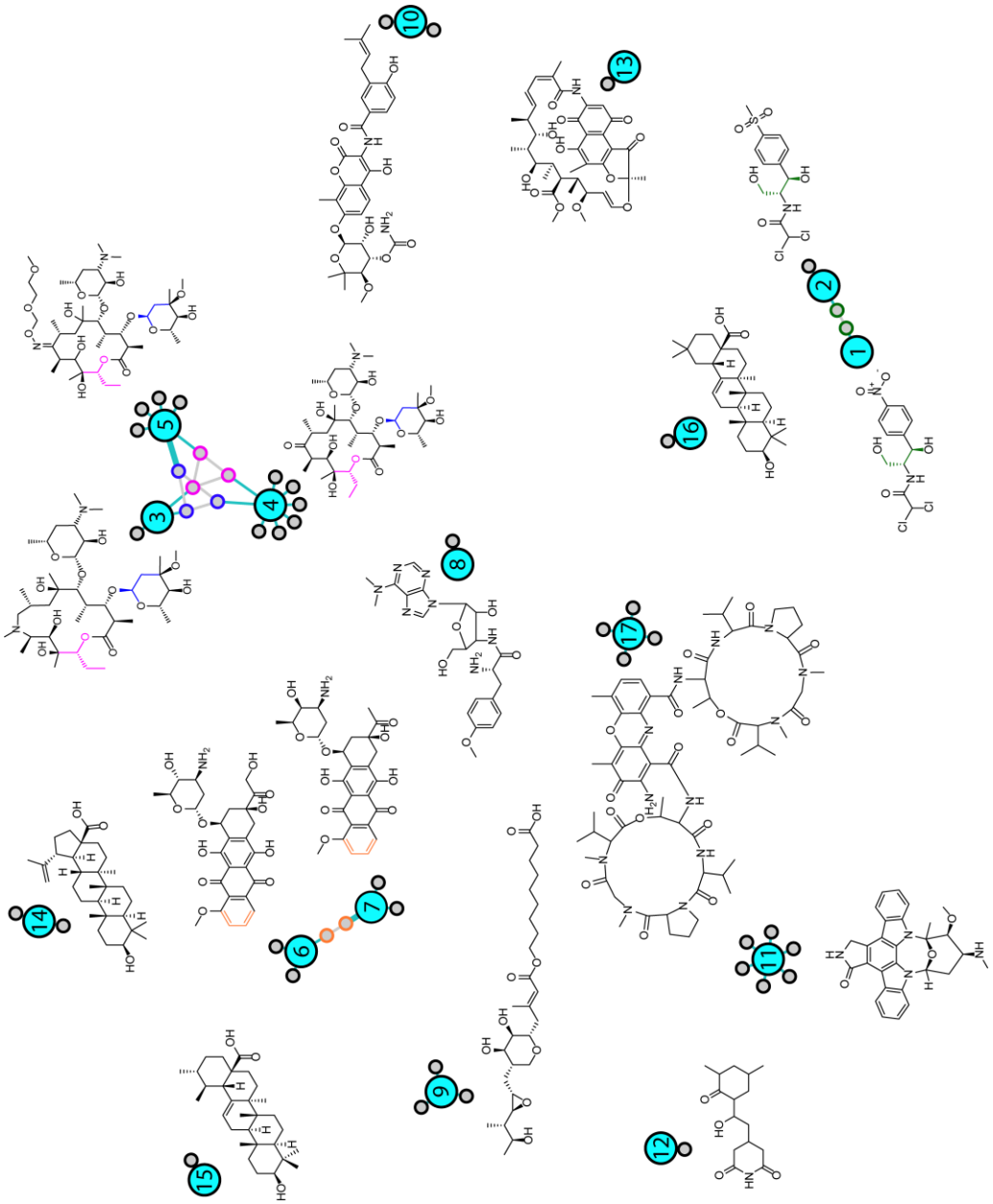


Figure 2.20. Full Annotation Network of Standard Compounds Involved in MADByTE Development Including Structure Annotations.

2.8.3. Additional MADByTE GUI Features

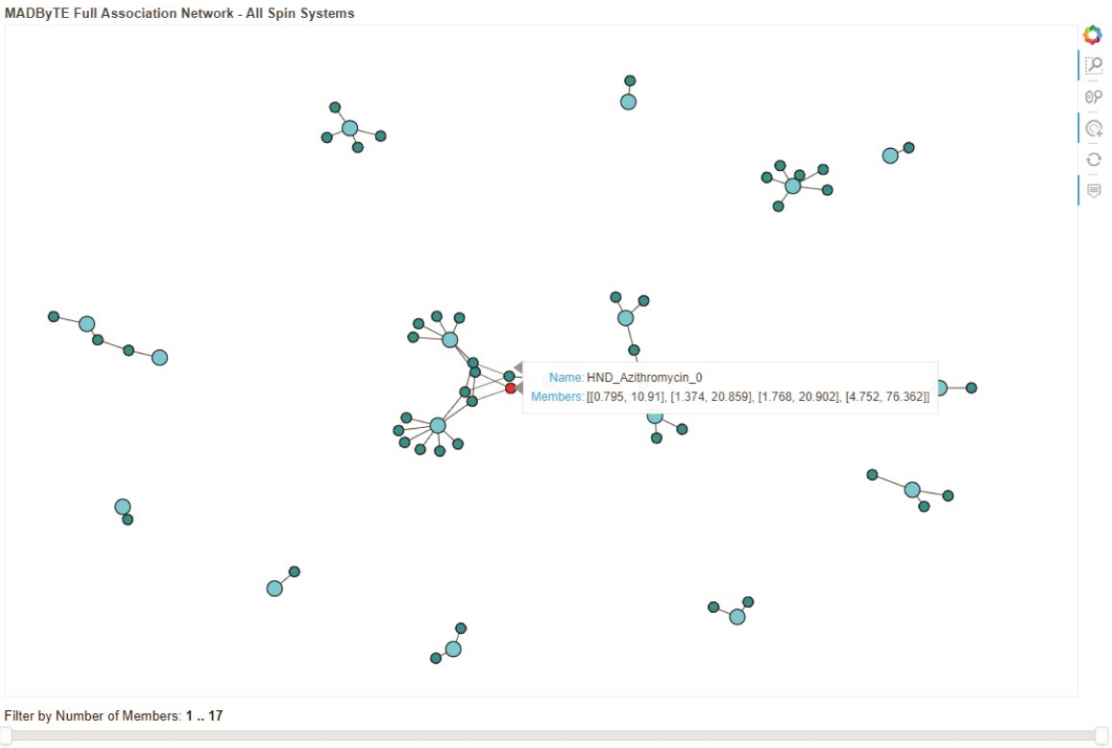


Figure 2.21. The MADByTE Network Module. The networks constructed by MADByTE analysis can be viewed using the interactive module – hovering over nodes displays SSF membership.

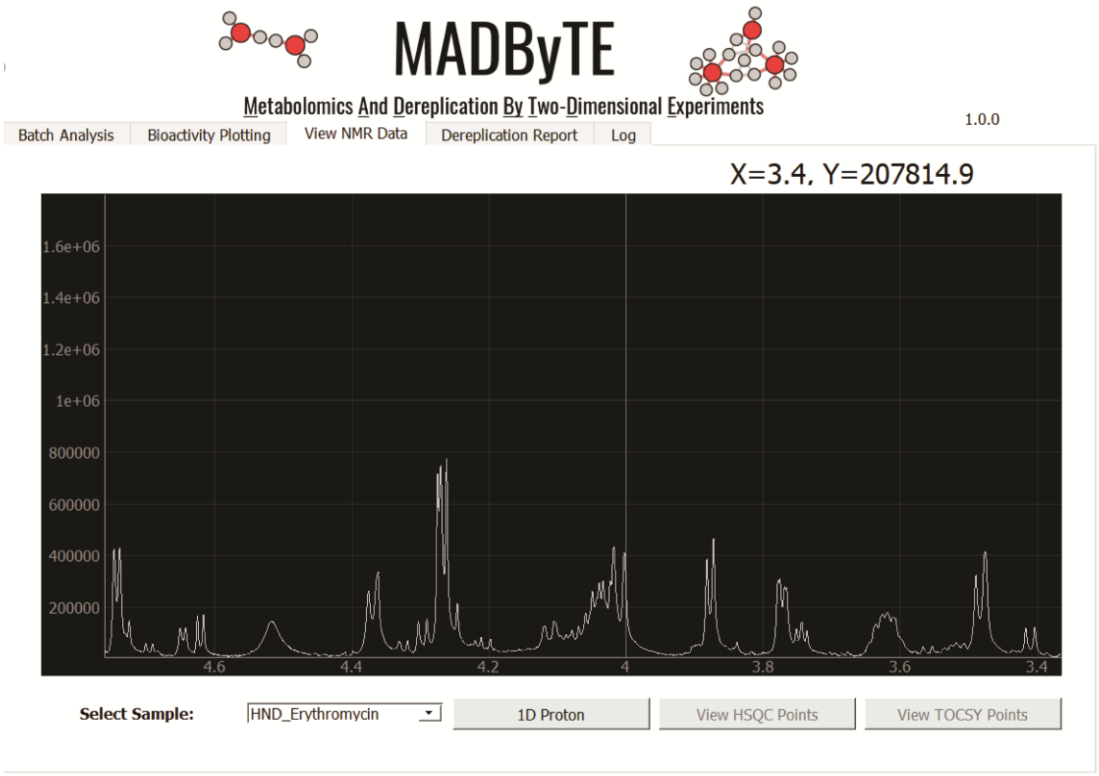


Figure 2.22. Screenshot of the MADByTE GUI Plotting Function. The plotting function built into MADByTE can display ^{1}H NMR spectra from MADByTE processed samples natively including ^{1}H spectra and points derived from TOCSY and HSQC Processing.

Chapter 3.

Integration of Bioactivity Profiling and MADByTE

3.1. Introduction

Drugs and therapies discovered from natural products are often considered as some of the most important drug discovery successes in modern medicine. Since the discovery of penicillin as an isolable fungal metabolite, the relationship between small molecules produced by living organisms and their role in the treatment of diseases have inspired the study of and search for new and novel producers of bioactive metabolites. A review of the last 40 years of all FDA approved drugs shows that natural products or natural product inspired molecules account for ~60% of small molecule drugs, demonstrating their broad usage and modern relevance as lead molecules for medicinal use.[1]

Historically, most discovery platforms aimed at identifying small molecules from natural sources relied on bioactivity guided fractionation to isolate bioactive constituents. This practice utilizes iterative rounds of separation by chromatography or partitioning with bioactivity evaluation at each step. Complex samples, through multiple rounds of this process, can be purified into their constituents using the bioactivity results as a prioritization function. This method has led to the discovery and development of many bioactive natural products,[11] but includes an inherent cost in both the time and sample amounts needed to perform biological evaluation at every step. In many cases, extensive time and effort can be spent on the characterization of a single component from a sample only to find that the molecule is already known and well characterized in the literature. In others, extensive rounds of bioassay guided fractionation can quickly reduce the amount of working material available until the supply is exhausted, and recollection of the original fraction cannot be guaranteed. Therefore, while bioactivity guided fractionation allows for the triage of non-bioactive extracts and the prioritization of others, the risk of rediscovery has led to the reduction of this method as an industry standard.

3.1.1. Bioactivity Profiling

To overcome the limitations of bioactivity guided fractionation, many discovery workflows have shifted towards a more informatics-based approach by infusing analytical methods of analysis such as MS and NMR spectroscopy into various steps in the process to characterize sample constituents. Using these data, samples can be checked for known entities through a process called dereplication. Although a powerful utility for prevention of reisolation, dereplication methodologies require bespoke libraries of reference data to be compiled before their full potential can be realized and are often tailored to single organisms, or frequently studied model organisms. In practice, when dereplication suggests a known bioactive molecule exists in a mixture, investigations are often abandoned before the full evaluation of a particular extract which may still yet yield important bioactive molecules.

Mass Spectrometry Based Methods of Biological Profiling

Because dereplication can only compare what is currently known to an investigation, new methods for predicting bioactivity from complex analytical data on these extracts have been developed in recent years which focus on the use of statistical modeling to predict features which may be giving rise to bioactivity. Biochemometrics, for example, derives PLS components for comparison and calculates the variance in a bioactivity dataset as a function of the presence or absence of a particular feature within a more complex dataset – called the selectivity ratio.[23] This allows for targeted isolation of features which may explain the bioactivity of the extract. Biochemometrics, although able to predict whether a molecule may be bioactive, does not provide context as to the MOA of a given molecule, and therefore could lead to the isolation of components which contain off target effects.

Beyond simply predicting the activity of a natural product within a complex extract, advancements in metabolomics comparison have allowed for the prediction of specific mechanisms of these analytes through comparisons of high content profiling assays. Compound Activity Mapping is one such utility which utilizes high content imaging of HeLa cells to generate fingerprints which describe the phenotypic effects of extracts in combination with HRMS data to generate hypotheses about the MOA of molecules from actinobacteria extract prefractions.[25] Importantly, Compound Activity Mapping provides these predictions by the generation of high content networks which allow for these data to be visualized and grouped based on shared features. The resulting networks provide

general descriptions about the predicted bioactivities of extracts, as well as specific predictions of molecular identities driving these bioactivity profiles. Although a very powerful utility, Compound Activity Mapping requires the generation of many biological responses through high content screening which may not be accessible to many research programs.

Mass spectrometry provides a powerful analytical method for generating features to compare across samples due to its ability to: resolve a great number of analytes through coupling to chromatography systems; the ability to selectively fragment ions of interest; and measure additional traits such as a collisional cross section. The wide variety of information which can be derived for constituents within a complex sample provide a robust method for verifying the identity of a molecule and even predicting its molecular class. However, mass spectrometry carries some inherent complications that can create problems with downstream isolation. Often the limit of detection for MS is orders of magnitude lower than that of other techniques, such as NMR spectroscopy, but as MS is not inherently quantitative, this can create instances in which features are predicted to be bioactive but are not present in high enough abundance for practical isolation. In addition, MS based methods are dependent on the analytes ability to ionize and hold a charge – a requirement that many molecules can fail to satisfy depending on their ionization mechanism or molecular structure.

3.1.2. NMR Applications

In comparison to MS, NMR spectroscopy is regarded as a “universal detector” for organic molecules, as it relies on the detection of magnetically active nuclei - typically ^{1}H - which can be found in almost all organic molecules. Additionally, NMR is semi-quantitative under standard conditions and although the limit of detection is considerably higher than MS, features that are abundant enough for detection by NMR are present in practically isolable amounts. Therefore, methods which are able to profile the biological activities of complex samples using NMR as the basis for analysis would be at a distinct advantage over MS based methods for practical use. However, without the resolving power of MS based analyses, deriving features for comparison and prediction remains a challenge.

NMR based metabolomics have recently attempted to solve this issue by using the comparison of data between extracts or samples to find common features which exhibit

similar bioactivities to explain the contributions from metabolites. STOCSY (**S**tatistical **TO**tal **C**orrelation **S**pectroscop**Y**) is a method of comparison for ^{1}H 1D spectra which utilizes the collinearity of signals in sets of ^{1}H spectra to generate a correlation matrix predicting the associations of signals which may arise from single metabolites shared between the samples.[50] When combined with statistical modeling such as OPLS-DA, predictions can be made about the metabolites' contribution to a given biological state, observed biological effect, or connect NMR fingerprints to MS features.[80] However, STOCSY requires extremely large sample sets with carefully monitored biological data outputs and has been designed for use in primary metabolomics investigations, which often can be validated against reference libraries. Despite this, STOCSY has been shown to be compatible with natural product discovery pipelines, but still requires extensive separation to predict bioactive motifs.[81]

MADByTE, like STOCSY, is a method for comparison of common features between complex samples through relationships in NMR spectra. Unlike STOCSY, however, MADByTE does not rely on complex statistical relationships of compared features or combinations with MS profiling data. MADByTE derives shared spin system features for comparison, and like MS based utilities such as Compound Activity Mapping, provides contextualized mapping of the features occurrence through the generation of a network. Using the network resulting from a MADByTE analysis of a sample set, predictions of bioactivity of linked extract nodes could be made based on their shared spin system features and bioactivity profiles.

3.2. Integration of Biological Profiling and MADByTE

3.2.1. BioMAP

Biological evaluation of extracts is an important step in the profiling of natural product libraries, allowing for the prioritization of extracts which display elevated activities against known pathogens. In most applications, a targeted organism is selected for evaluation and extracts are evaluated against a particular organism of interest, such as a drug resistant or hospital derived strain of a pathogenic bacteria. Typically, biological assays provide a metric of growth inhibition, phenotypic response, or cytotoxicity. While this strategy provides valuable information about the extract potential, it carries inherent limitations if only a single organism is profiled.

In recent years, new biological evaluation methodologies have allowed for wider screening panels to be tested with minimal sample requirements. The inherent advantages of these methods include a more comprehensive bioactivity profile and comparison of biological responses from the same extract across expanded panels. BioMAP, a platform developed to take advantage of an expanded panel of 15 bacterial strains, allows for comparisons of biological response to construct fingerprints which can be compared to known antimicrobial compounds. These comparisons of complex extracts were shown to be sufficient to predict the identity of known antibiotics contained within them, and even provide predictions of bioactive compound class. Importantly, the screening strategy developed for BioMAP provides access to high content data in a high throughput format, allowing for the rapid evaluation of large natural product libraries.

The 15 bacterial strains used in the development of BioMAP (Table 3.1) represent an important cross section of clinically relevant bacteria, known to have different responses to antibiotics from differing classes. The diversity of this panel is important, as it addresses the need for antibiotics capable of treating both Gram-positive and Gram-negative bacterial infections and provides datapoints for activity profiles against a wide variety of commonly targeted organisms. In the BioMAP platform, the differential between these responses allows for the class level prediction of active components. Establishing a link between these biological data to spectroscopic information could potentially provide a mechanism for bioactive component prediction without the need for large reference datasets.

Table 3.1. Organisms Used In BioMAP Screening

Strain	Gram (+/-)	Biosafety Level
Bacillus subtilis 168	+	1
Listeria ivanovii (BAA-139)	+	1
Enterococcus faecium (ATCC 6569)	+	1
Staphylococcus epidermis (ATCC 14990)	+	1
Staphylococcus aureus (ATCC 29213)	+	2
Methicillin-resistant *S. aureus* (MRSA) (BAA-44)	+	2
Escherichia coli K12 (BW 25113)	-	1
Providencia alcalifaciens (ATCC 9886)	-	1
Ochrobactrum anthropi (ATCC 49687)	-	1
Enterobacter aerogenes (ATCC 35029)	-	1
Acinetobacter baumanii (NCIMB 12457)	-	1
Vibrio cholerae O1 (biotype El Tor A1552)	-	2
Salmonella typhimurium LT2	-	2
Pseudomonas aeruginosa (ATCC 27835)	-	2
Yersinia pseudotuberculosis (IP2666 pIBI)	-	2

3.2.2. MADByTE Comparison of BioMAP Evaluated Extract Prefractions

The development of MADByTE (Chapter 2) involved the profiling of 85 extract prefractions from the Linington Lab actinobacteria extract library. HSQC and TOCSY experiments were used to generate features which allow for comparison of scaffold motifs present in complex mixtures, and provided a viable method for compound dereplication from a structure driven perspective. A subset of these data (34 samples) were previously screened in the BioMAP assay, providing the bioassay data necessary for the evaluation of this approach.

The MADByTE bioactivity layering function was constructed to be a generalizable utility which would not require extended bioactivity panels and could easily be adapted. To this end, construction of categorical bins is performed, and extract nodes are re-colored to provide a topographical context of activity on top of existing MADByTE networks. The initial screening of these extracts provided three metrics as outputs; active, mildly active, and inactive; activities were summed across all 15 organisms. Three bins were established to highlight broad spectrum antimicrobial activity of the Linington Lab extracts

against the BioMAP panel (mildly active– 1-4 organisms, moderately active – 5-9 organisms, and highly active 10+). Samples which hit at least one organism in BioMAP provided the MADByTE network shown in Figure 3.1 - panel A, and layering the broad-spectrum activity profiles resulted in the network displayed in Figure 3.1- panel B.

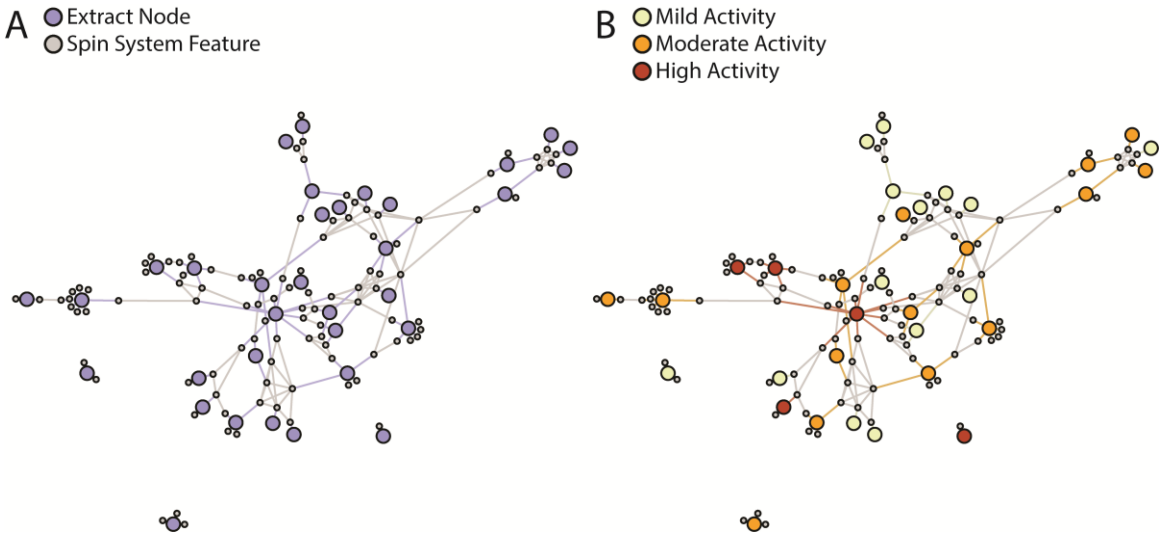


Figure 3.1. Layering of BioMAP Activity Onto a MADByTE Network. Activity profiles were established as mildly active(1-4 organisms hit), moderately active (5-9 organisms hit) and highly active (10+ organisms hit). Clusters of high bioactivity can serve as a method for prioritization and can provide structural relationships of potentially bioactive motifs present in the extract prefraction.

3.3. Isolation of an Active Component from MADByTE Networking

3.3.1. Determination of a Shared Feature for Isolation Prioritization

The bioactivity layering function provides a contextual mapping of extract activities from spectral information. Regions of low or moderate activity may possess shared chemistry, but the bioactivity layering suggests that if isolation of a bioactive component is of high importance, these structural features can be deprioritized while areas of high activity which show similar SSFs can be prioritized. Investigation of common linkages between three prefractions displaying high biological activity displayed this favorable relationship. These prefractions belonged to the same isolate, RL12_176_HVF_A, originally isolated from marine sediment in Bell Point, WA in 2012. The overlap of the SSFs from these samples

provided a plausible hypothesis for a structural feature belonging to a potentially active shared component, and therefore represented a high priority target for isolation. Comparison of the spin system features is displayed in Table 3.2. Although there were differences in the assigned ^{13}C resonances of 2108_C_8 compared to 2108_D_4 and 2108_E_5, the ^{1}H resonances are highly conserved and served as the principal feature for isolation.

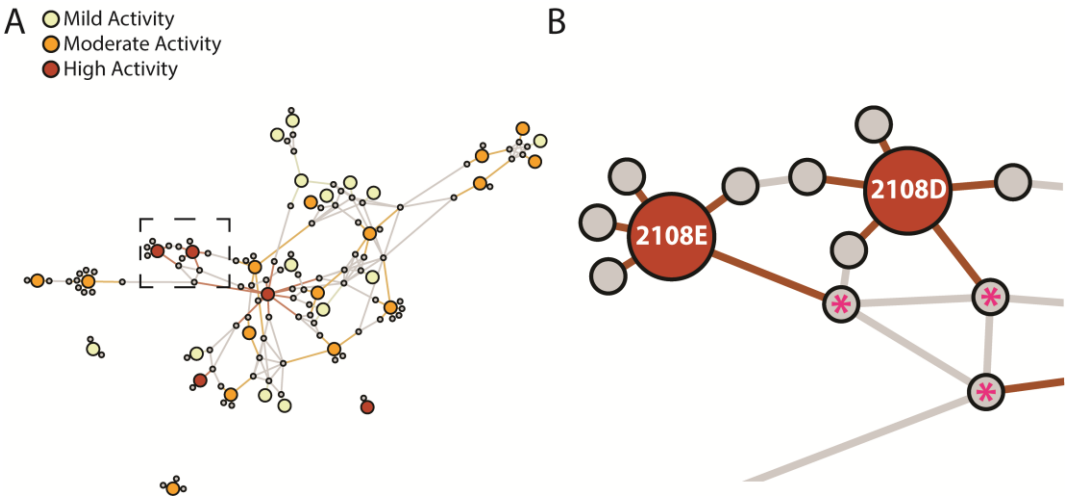


Figure 3.2. Overlap of Extract Prefractions 2108E and 2108D Provides a Plausible Target for Isolation of a Predicted Bioactive Component.

Table 3.2. Spin System Features Derived from Three Highly Bioactive Extract Prefractions

2108_E_5		2108_D_4		2108_C_8	
^{1}H ppm	^{13}C ppm	^{1}H ppm	^{13}C ppm	^{1}H ppm	^{13}C ppm
7.50	124.5	6.92	117.1	7.50	124.7
7.97	119.7	7.50	124.6	7.97	124.5
8.34	120.7	7.96	119.4	8.39	137.1
8.68	149.1	7.97	124.6	8.70	124.5
8.71	149.1	8.39	120.7	8.70	149.2
		8.72	149.1		

3.3.2. Regrowth and Extraction of the Producing Organism

To provide ample material for isolation of the pure component, a stock culture of RL12_176_HVF_A was streaked out on SYP media and allowed to grow until distinct

colonies formed. Four isolated colonies were each transferred into 7 mL of marine broth and allowed to grow for 5 days. From each culture, 3 mL was transferred into medium scale conditions (60 mL of media) and allowed to incubate for 7 days. Large scale growths were prepared by inoculation of 40 mL of the medium scale to 1L of SYP medium containing 20 g of washed XAD-16 resin and incubated for 7 days. All cultures were incubated at 27 °C and shaken at 200 rpm.

Each large-scale culture was vacuum filtered to remove the supernatant from the cellular material and resin. The resulting cell material, resin, and filter were then placed into a 1 L Erlenmeyer flask containing 250 mL of 50:50 CH_2Cl_2:CH_3OH mixture, shaken for 1 hr, and filtered to remove the solid material from the extracted metabolites. Each extract was then evaporated under vacuum until minimal solvent remained and loaded onto celite loading material for Combiflash separation. Each 1 L large scale culture was kept separate, and no samples were pooled at this point.

3.3.3. Flash Chromatography and Isolation of Collismycin A

Flash chromatography was completed on each large-scale culture using a 4 minute 10% loading followed by a 10-100% (methanol: water) gradient over 34 minutes with a 3 minute 100% methanol wash and 3 minute ethyl acetate wash at a flow rate of 20 mL/min. Aliquots of the higher methanol fractions (fractions 7-18) from cultures 1-3 were then analyzed via LCMS to assess complexity. One sample, fraction 10 from culture 3 seemingly contained only one metabolite when analyzed via LCMS (Figure 3.3) (m/*z* of 276.1) and yielded 5.83 mg of material after drying. NMR analysis of fraction 2108_S3_F10 yielded confirmation that the product was pure (> 95%) and contained the proton resonances shared between the targeted spin system features. Comparison to literature values verified this component as collismycin A (**18**).[82,83]

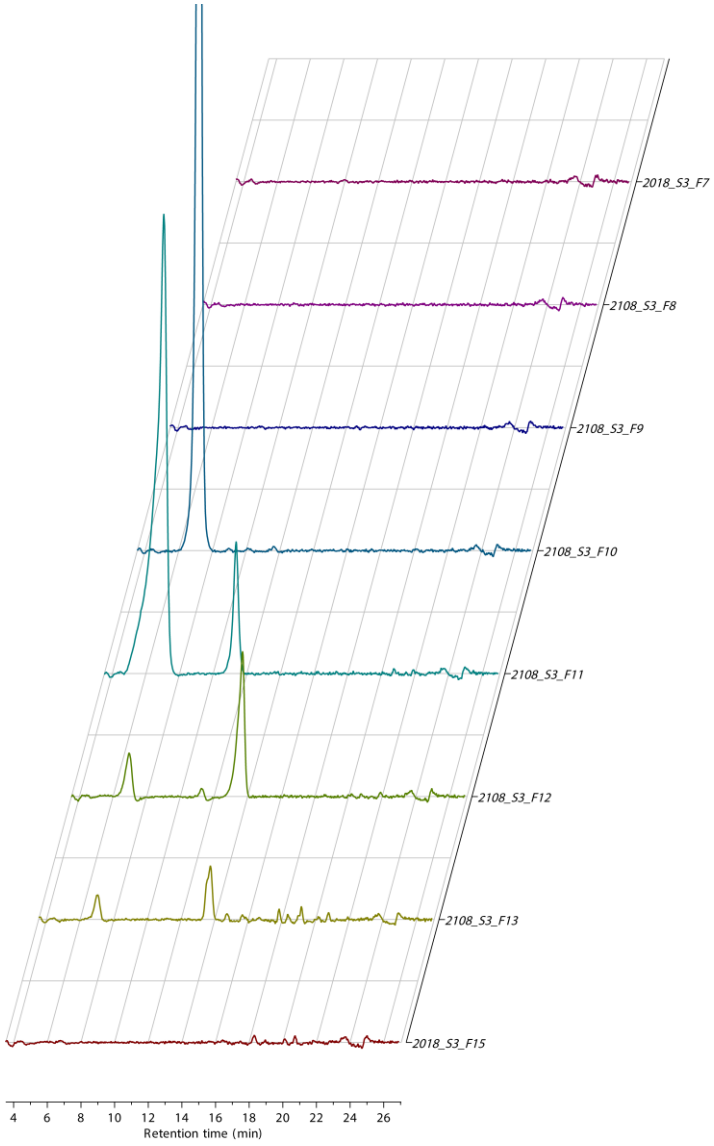


Figure 3.3. PDA Response Profiles of Combiflash Fractions from 2108 Culture 3 Analyzed Via LCMS for Complexity.

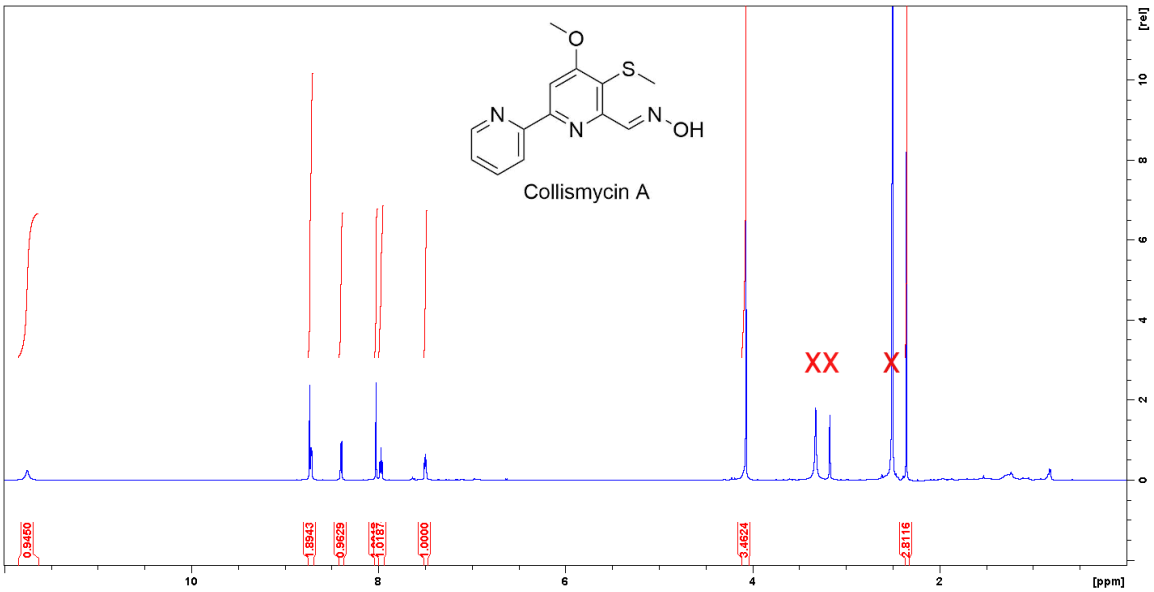


Figure 3.4. **^{1}H NMR of 2108_S3_F10 in DMSO-d_6, Determined to be Collismycin A.**

The MADByTE network of these extracts was reconstructed including the spin system information derived from the purified collismycin A and the resulting spin system feature was shown to be directly associated with the spin system features of interest.

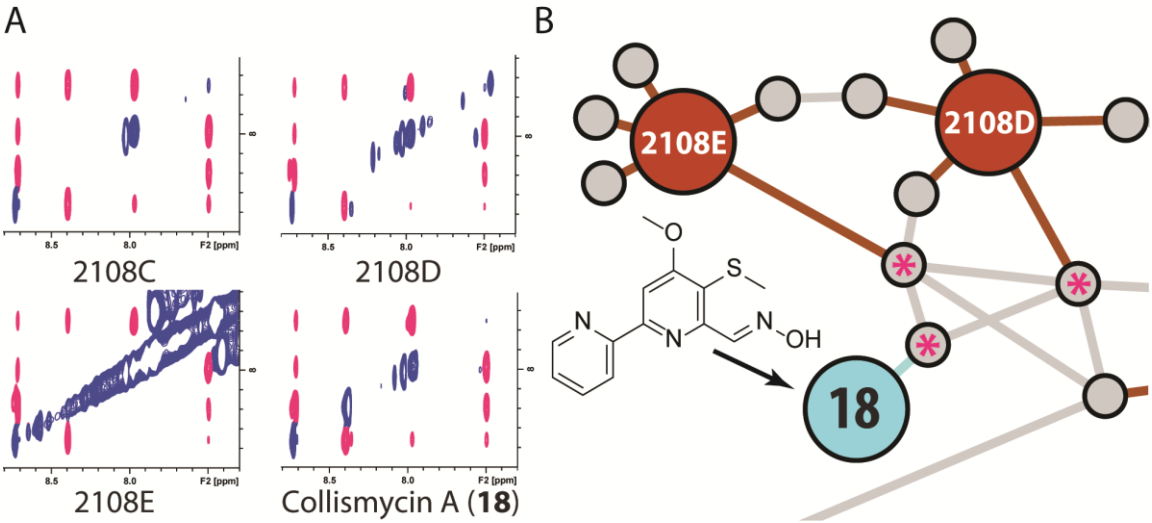


Figure 3.5. **Collismycin A (18) Showed High Overlap in the 2D TOCSY When Compared to the Prioritized AExtracts (A) and Networking (B) Revealed Direct Connections to Prioritized Spin System Features.**

3.3.4. Biological Evaluation of Collismycin A.

Antimicrobial susceptibility tests for collismycin A were performed against a select panel of bacteria (Table 3.3) using a miniaturized high throughput assay adapted from the broth microdilution method outlined by the Clinical and Laboratory Standards Institute (CLSI).[84] Bacterial test strains were individually grown on fresh Nutrient Broth (NB, ATCC Medium 3) agar, Tryptic Soy Broth (TSB, ATCC Medium 18) agar or Brain Heart Infusion (BHI, ATCC Medium 44) agar, as recommended by the American Type Culture Collection (ATCC) cultivation protocol (Table 3.3). Individual colonies were used to inoculate 3 mL of sterile NB, TSB or BHI media and grown overnight with shaking (200 rpm; 37 °C). *Listeria ivanovii* (ATCC BAA-139) and *Streptococcus pneumoniae* (ATCC 49619) were incubated overnight but not shaken (37 °C; 5% CO_2). Saturated overnight cultures were diluted in their respective media according to turbidity to achieve approximately 5 x 10^5 CFU of final inoculum density and dispensed into sterile clear polystyrene 384-well microplates (Thermo Scientific 265202) with a final screening volume of 30 µL. *L. ivanovii* was diluted with and grown in Haemophilus Test Medium (HTM; ATCC Medium 2167).

Table 3.3. Bacterial Strains and Growth Conditions for Biological Evaluation

Strain Name	Strain Number	Biosafety Level	Growth Medium	Growth Condition
Gram-Positive				
Bacillus subtilis	ATCC 6051	1	NB	37°C
Enterococcus faecalis	ATCC 29212	2	BHI	37°C
Enterococcus faecium	ATCC 6569	2	BHI	37°C
Listeria ivanovii	BAA-139	1	BHI-A; HTM	37°C; 5% CO_2
Staphylococcus aureus (Methicillin-Resistant)	BAA-44	2	TSB	37°C
Staphylococcus aureus (Methicillin-Sensitive)	ATCC 29213	2	TSB	37°C
Staphylococcus epidermidis	ATCC 14990	1	TSB	37°C
Streptococcus pneumoniae	ATCC 49619	2	BHI	37°C; 5% CO_2
Gram-Negative				
Acinetobacter baumanii	ATCC 19606	2	TSB	37°C
Escherichia coli	K-12 MG1655	1	NB	37°C
Klebsiella aerogenes	ATCC 35029	1	NB	37°C
Klebsiella pneumoniae	ATCC 700603	2	NB	37°C
Ochrobactrum anthropi	ATCC 49687	1	TSB	37°C
Providencia alcalifaciens	ATCC 9886	1	TSB	37°C
Pseudomonas aeruginosa	ATCC 27853	2	TSB	37°C
Salmonella enterica	ATCC 13311	2	NB	37°C
Shigella sonnei	ATCC 25931	2	NB	37°C
Vibrio cholerae	A1552 El Tor	2	TSB	37°C
Yersinia pseudotuberculosis	ATCC 6904	2	BHI	37°C

DMSO solutions of test collismycin A and antibiotic controls were prepared as a 1:1 dilution series and pinned into each assay plate (200 nL) using a high throughput pinning robot (Tecan Freedom EVO 100) to achieve final screening concentrations ranging from 128 µM to 3.91 nM. In each 384-well plate; lane 1 was reserved for DMSO vehicle and culture medium; lane 2 reserved for DMSO vehicle, culture medium and target bacteria; lanes 23 and 24 reserved for antibiotic controls, DMSO vehicle, culture medium and target bacteria. After compound pinning, assay plates were read as T_0 at OD_{600} using an automated plate

reader (Molecular Devices SpectraMax i3x), sealed with a lid and placed in a humidity-controlled incubator at 37 °C for 18-20 h. Final OD_{600} values were obtained on the same plate reader for T_{20} values. *L. ivanovii* and *S. pneumoniae* were incubated in a separate incubator (37 °C; 5% CO_2). Resulting growth curves for each dilution series were used to determine the MIC_{50} values for all test compounds following standard procedures.

Table 3.4. Biological Evaluation of Collismycin A

Organism	MIC_{50} (μM)
Bacillus subtilis	32
Enterococcus faecalis	>128
Enterococcus faecium	>128
Listeria ivanovii	>128
*Staphylococcus aureus (*Methicillin-Sensitive*)*	>128
*Staphylococcus aureus (*Methicillin-Resistant*)*	>128
Staphylococcus epidermis	>128
Streptococcus pneumoniae	64
Ochrobactrum anthropi	>128
Escherichia coli	32
Klebsiella aerogenes	64
Klebsiella pneumoniae	16
Providencia alcalifaciens	128
Salmonella enterica	32
Shigella sonnei	64
Yersinia pseudotuberculosis	>128
Acinetobacter baumanii	128
Pseudomonas aeruginosa	32
Vibrio cholerae	128

Encouragingly, the results from the antimicrobial susceptibility tests showed that collismycin A was active in 11 out of the 19 organisms tested, reflecting a hit rate comparable to the initial screening results and confirming that incorporation of bioactivity information into MADByTE networking allows for associations between structural motifs and bioassay activity to be made.

3.4. Variations of Bioactivity Prediction Using MADByTE

The prioritization and isolation of collismycin A by using bioactivity layered networks reinforces that the layering of orthogonal data generates strong hypotheses for natural product isolation efforts. The amount of bioassay data available aided in generating a robust hypothesis and was strengthened through the wide panel of bacteria screened. The summation of these results allowed for the isolation of a compound displaying broad spectrum activity, which can be a favorable lead in drug discovery efforts. However, the ability to filter data such as the ability to prioritize activity against Gram positive or Gram negative in a selective fashion could be of great interest for finding selective agents over broad-spectrum activities. To demonstrate this approach, the MADByTE network from Figure 3.1 was further processed (Figure 3.6) to provide viewpoints offering summed response profiles (Figure 3.6-Panel B) Gram negative (Figure 3.6-Panel C), and Gram positive (Figure 3.6- Panel D) in direct comparison. This new viewpoint shows that although the nodes which led to the isolation of collismycin A were highly bioactive from a broad-spectrum point of view, other bioactive nodes were selectively active against one or another organism type.

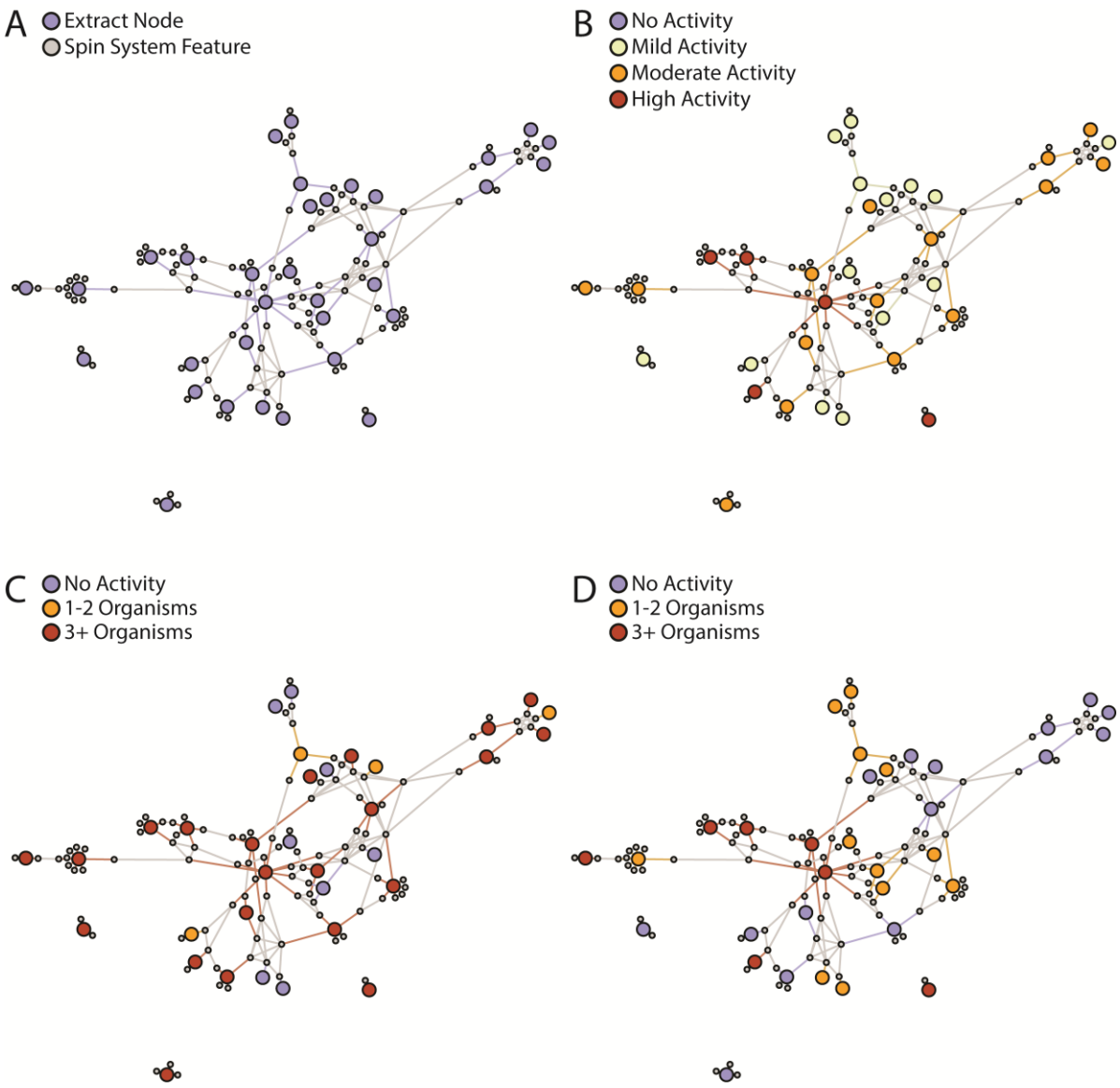


Figure 3.6. Differential Layering of Bioactivity Information on a MADByTE Network. A) MADByTE base network with no bioactivity based color coding. B) Summation of all bioactivity profiles from BioMAP screening. C) Overlay of Gram positive activity profiles. D) Overlay of Gram negative activity profiles.

High throughput applications can generate vast amounts of biological information simultaneously, but these applications are often out of reach for many smaller natural product laboratories or screening campaigns. To allow for generalizable use of the MADByTE bioactivity layering, the bioactivity module performs no processing on provided biological data and instead implements color coding based on user provided bins. As

users establish the bins they wish to use, responses can be generalized to a wide variety of scalable metrics such as MIC values, live/dead binary responses, or inhibition of a target expression. By allowing users to determine what constitutes a hit in the bioactivity response, investigators can optimize the utility for their experimental design and constraints. Extensions of this principal could yield networks which display counter-screening results which could further refine the bioactivity prediction and subsequent prioritization.

3.5. Limitations and Future Directions

The generation of a synthetic spin system feature through the combination of the shared resonances of 2108_C_8, 2108_D_4, and 2108_E_5 produced ^{1}H resonances which could be used for isolation. However, as noted in section 3.3.1, the ^{13}C resonances determined for 2108_C_8 were dissimilar to those compared between 2108_D_4 and 2108_E_5. Investigation of these spectra showed that 2108_C_8 contained several resonances in this region belonging to a minor constituent which confounded the spin system assignment function. This suggests that this organism produces several compounds which may be structurally related but are different enough in their ^{13}C shifts to cause complications when attempting to assign values between the HSQC and TOCSY. As noted in Chapter 2, further refinement, and adoption of pure shift HSQC and TOCSY experiments may be sufficient to mitigate this issue but must overcome sensitivity losses when working with complex extracts.

3.6. Conclusions

MADByTE networks have shown the ability to derive structural information from complex samples and apply new visualization techniques to generate context. As shown through the proportion of approved drugs in the last four decades, natural products represent important molecules for human health and many NP discovery pipelines remain focused on the rapid prioritization of bioactive molecules. As demonstrated through the prioritization and isolation of collismycin A, combining structural context from MADByTE with generalizable layering of bioactivity data, targeted isolation can greatly streamline the discovery pipeline. Importantly, as MADByTE utilizes structural features derived from

experimental spectra and is not contingent on existing databases for comparison, these untargeted pipelines hold great value in discovery.

Chapter 4.

Substructure Hypothesis by Integration of SMART and MADByTE

4.1. Introduction

Characterization of natural products in extracts is a challenging task, even in cases where a molecular identity is known. A robust method of dereplication can inform investigators about the known chemistry within a given sample and provide a hypothesis of novelty if no matches are found. However, as most dereplication libraries are lab specific, there currently are limited options for thorough dereplication from an NMR perspective. Centralized databases of pure compound spectra would greatly aid future targeted metabolomics efforts by enabling direct comparison to reference data, but few have been constructed which are specific to natural product investigations and house the types of molecular entities often encountered.[51] The databases of spectra from each laboratory can contain massive gaps in coverage, which can limit the databases utility when applied to new projects or organisms, and are rarely shared between groups. In cases where databases containing chemical shift information exist, reports are often provided in a single-solvent – resulting in less overall coverage and limiting widespread adoption of these procedures. In primary metabolomics databases, for which there are a number of targeted molecules submitted and referenced, most are provided for molecules in buffered aqueous solutions, which is not optimal for NP investigations.

Specialized databases with a natural products focus have been constructed, such as the MetIDB,[85] or the JEOL NP NMR database which allow for specialized applications. However, MetIDB is no longer available to the public, and the JEOL NP database was constructed from a ^{13}C perspective and lacks total ^{1}H coverage for all molecules, representing a challenge for comprehensive representation. Currently, amalgamation of these databases is slow moving, and solutions beyond standard databasing are needed to allow for more rapid discovery and annotation of natural products in mixtures.

Since the early 2000s, calculation of predicted NMR shifts for organic molecules has been an area of rapid development and expansion, with DFT methods revolutionizing the

accuracy of predicted spectra but at the cost of computational power and time.[86] Many applications of these calculation methods have focused on structure verification and resonance assignments and have been applied extensively to complex natural product molecules.[87] More recently, hybrid methods which use availability of reference compound information to inform calculation efforts have also shown great promise, a high degree of accuracy, and have cut down drastically on the computational time costs associated with generating high accuracy predictions on large numbers of molecules.[42,88]

Mixture analysis pipelines have increasingly embraced calculation methods to access chemical shift information from proposed structures. If analysts have knowledge of candidate structure or metabolites which are plausible, modelling and prediction allows complex NMR data to be readily queried allowing highly targeted analysis without the need of a bespoke database.[89] However, application of these calculation based approaches requires prior knowledge of the molecules that may be in the mixture which is not a feasible requirement when applied to discovery pipelines where the target objective is the discovery of new molecules.

Inspired by the improvement of metabolomics pipelines for comparison of sample constitution, natural product investigations have increasingly relied on the comparison of features across samples rather than direct verification of an analyte. Increasingly, these workflows have been paired with machine learning algorithms and platforms which leverage a considerable throughput advantage afforded by automated processing. These methods, although powerful, carry significant limitations when addressing mixtures. Most ML platforms are trained on model data, which originate from single molecule samples in optimal conditions, whereas more complex samples may confound the analysis pipeline when more signals are present to consider.

4.1.1. Challenges with MADByTE Analysis

At its core, MADByTE is a sample comparison utility aimed at deriving features for comparison. As described in Chapter 2, it utilizes an orthogonal data comparison strategy to generate spin system features from complex samples. The strength of using MADByTE is that it directly compares sample to sample without requiring an external database and spin system features can be mapped across a sample subset to determine spin systems shared by molecules in extracts. From a discovery point of view, this information can be

leveraged as a metric of novelty in cases where few or no similar spin systems are found, but these associations are sample and experiment dependent. As a metric of similarity, associated features across samples can suggest shared chemical motifs, but it is a significantly more difficult task to assign motif elements to the SSFs directly without a database of reference compounds to provide context.

4.1.2. Small Molecule Accurate Recognition Technology (SMART)

Recently, a new machine learning platform, SMART (Small Molecule Accurate Recognition Technology) has been proposed as a new structural dereplication utility for pure compounds.[40] SMART uses a comparison strategy similar to image recognition technologies by comparing HSQC spectra to user supplied peak lists, attempting to recognize molecules that may exist in the spectra, analogous to an image recognition algorithm identifying an object in a picture. At its core, SMART is constructed with over 50,000 natural products HSQC spectra as its training set, providing an unmatched scale of dereplication data to users through this molecular recognition method.

Comparison by SMART is an all-on-all query against the points in the database, attempting to fit as much of the provided data as possible. The practical restriction on this strategy is that the algorithm is attempting to map all provided HSQC datapoints to a single entity. Comparing this problem to the image recognition analogy, the system may be able to identify an image of a lamp against a blank background but identifying the object in a natural setting of a living room may fail. Due to the all-on-all comparison strategy, complex mixtures where several compounds contribute to data complexity remain a challenge.

4.1.3. Combinations of Context

MADByTE spin system features, although small with respect to the entire molecule, represent resonances from HSQC spectra which have been associated as originating from the same molecule. If SMART analysis can find molecules which contain the MADByTE SSFs, it should be possible to use the context of SMART results to provide structural information. Importantly, SMART is a dereplication utility at its core, but uses a method of comparison which does not make a definitive claim of identity. Rather, it suggests the types of molecules a match may fit the available data. The classes of molecules returned can be of substantial use to inform structure elucidation, even if the verification of a single

molecule cannot be made. A new MADByTE module, SHIMS (**S**ubstructure **H**ypothesis by **I**ntegration of **M**ADByTE and **S**MART) was designed to provide plausible substructures for MADByTE features by providing a frame of context as to what kind of molecules and chemical environments the chemical shift patterns in SSFs may be describing.

4.2. SMART Usage and Data Structure

4.2.1. SMART Training Dataset

As with all machine learning algorithms, the strength of the platform is proportional to the size and quality of the data training set used for its creation. SMART originally was trained with 2,054 HSQC datasets, including a sizable portion of natural products classes. The variety of information was important to the performance, as many natural product classes have less than 50 described members.[40] Unlike conventional ML approaches which require active selection of features for comparison, SMART utilizes a process called deep learning, which derives features without interaction with an investigator. As HSQC spectra form patterns which can be displayed as images, a convolutional neural network was chosen as the inputs of choice, as they have greatly increased the effectiveness of image recognition ML approaches. With the release of SMART 2.0, the number of HSQC spectra used for the training set was increased to 53,076 created from primary data sources (25,454 HSQC spectra from the JEOL NP database) and simulated compounds (27,642 HSQC spectra generated from ACD Labs predictor).[54] Training on these new data provided SMART 2.0 with a considerable boost in coverage.

4.2.2. SMART Searching

SMART, as a utility, is a web-based application hosted by the University of California San Diego with an intuitive graphic user interface. Users can submit peak picked lists directly for query as flat files, or by providing tabulated HSQC resonances. The algorithm is relatively quick, with most queries taking less than 30 seconds to complete. The top 100 matches made by the algorithm are then displayed in a results table which contains the structure of the compound, the compound name, a SMILES string representing the chemical structure, and a cosine score describing the accuracy of the prediction (Figure 4.1). Results can be downloaded as a csv flat file containing all information except the rendered structure.

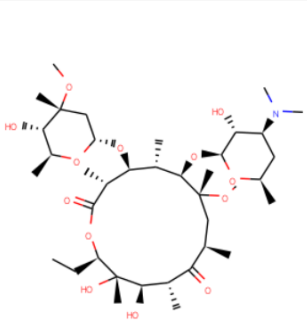
Name	Structure	Cosine score	MW
CLARITHROMYCIN		0.9637729367981128	747.5

Figure 4.1. Example of SMART Results Returned for an HSQC Spectra. Results are provided for the top 100 compounds which are plausible scaffold matches to the provided HSQC peak lists and contain the molecule name, structure, cosine score, and molecular weight.

Submission of peak lists to SMART 2.0 was a manual process, accepting csv files with certain format requirements which were not directly compatible with vendor peak list output files. As MADByTE was constructed to extract the resonance information from these vendor output files already, simple adjustments to the output formatting allowed for SMART to use these scripts to parse output lists from Topspin and MNova, removing the transcription requirement from the workflow with the release of SMART 2.1. To expand the dereplication capabilities of MADByTE, this new formatting script was added into the dereplication module, allowing users to directly format their MADByTE ready data for SMART searching. As SMART is a web-based application, interaction through an API is possible and was configured to allow for automatic submission of queries, facilitating high-throughput operation.

4.3. Design of SHIMS Processing

SHIMS processing attempts to describe the origin of SSFs contained in samples through comparison of the common elements of their HSQC spectra enabling prediction of molecular class and providing context as to where SSFs may arise from. This is accomplished through the integration of several open platforms and providing an overview of likely candidates. An overview of the process involved can be seen in Figure 4.2.

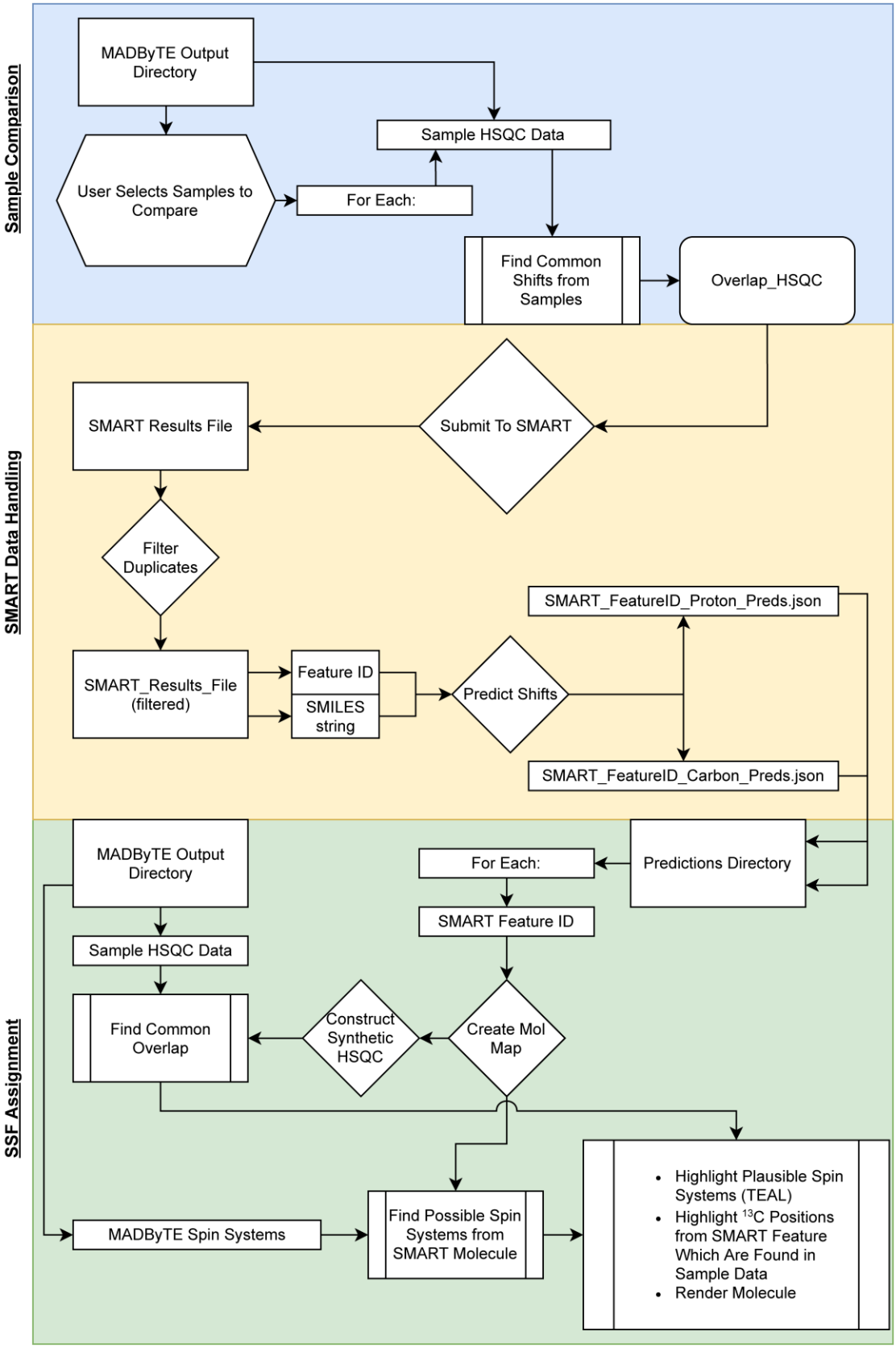


Figure 4.2. Overview of SHIMS Processes, Filtration Steps, and Data Handling.

4.3.1. Obtaining SMART Results

The initial strategy of comparison of MADByTE features using SMART aimed to find all molecules where the resonances from a SSF are found, and then to compare their structures to find common elements. It was proposed that the spin systems derived from MADByTE would be sufficient to return results from SMART, as the peak lists could be treated as truncated HSQC datasets. Unfortunately, this proved to be impractical, as the comparison strategy in SMART returned limited results with small datasets, as the model was not designed for fragment-based searches. Using the spin system erythromycin_0 from the standard compound experiment in Chapter 2, over 100 results were returned as plausible matches with cosine similarity scores above 0.5. This demonstrated that the algorithm was finding many compounds whose spectra were predicted to be closely related to the spin system. Notably, however, none of the returned structures were macrocyclic compounds, showing a failure to identify the class of compound from which this spin system was derived.

Spin systems are representative of small portions of these secondary metabolites they are derived from, and shared spin system features between samples highlights the overlap between the spectra. Using a single SSF for comparison showed the platform's limitations when searching directly in SMART, as the data contained in a SSF are too sparse for robust comparison using this platform.

4.3.2. Filtration of SMART Results

Occasionally, several instances of compound recognition returned from SMART show duplicate scaffolds or names, despite having different cosine scores associated with the match. This is due to the way in which the HSQC spectra were originally obtained for the training data set, and the metadata associated with each. In some cases, stereochemistry was neglected when the simulated HSQC were created, leading to some minor discrepancies when the same compound was simulated more than once. In others, the solvent used for simulation was different than that of the experimentally acquired spectra.

Filtering of these data to remove duplicates was done through comparison of the supplied SMILES strings from the results file. SMILES strings are a representation of the molecule noting positions of substitution and stereochemistry. Although stereochemistry can affect

chemical shift, it was determined that some compounds provided in the SMART database do not attempt to represent stereochemistry, instead predicting their NMR shifts from flat structures. To standardize the SMART results, it was determined to not factor in stereochemistry when removing duplicates.

To compare these molecules without stereochemistry, SMILES strings provided in the results are converted to an InChi-Key, a truncated version of the InChi which describes the flat structure of the compound. This method allowed for considerably different SMILES strings to be converted and compared, revealing that the underlying flat structures were the same. As duplicates are found, the result with the highest cosine score is kept and all others are dropped. In the case of erythromycin, filtration of the query reduced the top 20 results to 10 to consider (Table 4.1 and Table 4.2).

Table 4.1. Top 20 SMART Results from Erythromycin HSQC Data

Cosine Score	DBID	From	MW	Name From SMART
0.972	v2.1_92052	ACD_Labs	747.5	CLARITHROMYCIN
0.970	v2.1_86233	ACD_Labs	717.5	Erythromycin B
0.962	v2.1_91700	ACD_Labs	747.5	"Clarithromycin (Biaxin, Klacid)"
0.957	v2.1_91474	ACD_Labs	733.5	Erythromycin (E-Mycin)
0.953	v2.1_95226	ACD_Labs	748.5	Massbank:EA019001 Azithromycin
0.953	v2.1_71379	ACD_Labs	733.5	Erythromycin
0.953	v2.1_91247	ACD_Labs	733.5	Erythromycin A
0.953	v2.1_96810	ACD_Labs	733.5	Massbank:UF407401 Erythromycin
0.950	v2.1_5797	Jeol	748.5	erythromycin A oxime
0.950	v2.1_96020	ACD_Labs	687.4	Massbank:KO003661 Oleandomycin
0.949	v2.1_97044	ACD_Labs	748.5	HMDB:HMDB01916-1833 Azythromycin
0.947	v2.1_21508	Jeol	832.5	lankamycin
0.947	v2.1_91544	ACD_Labs	748.5	Azithromycin (Zithromax)
0.943	v2.1_96812	ACD_Labs	747.5	Massbank:UF408501 Clarithromycin
0.942	v2.1_94284	ACD_Labs	834.5	MLS001074061-01!DIRITHROMYCIN
0.942	v2.1_86234	ACD_Labs	719.4	Erythromycin C
0.941	v2.1_94617	ACD_Labs	861.5	MLS001074901-01!erythromycin ethylsuccinate
0.939	v2.1_95227	ACD_Labs	747.5	Massbank:EA019101 Clarithromycin
0.938	v2.1_87632	ACD_Labs	876.6	Megalomicin A

Table 4.2. Top Results from SMART After Duplicate Filtration

Cosine Score	DBID	From	MW	Name From SMART	Inchi_Key
0.972	v2.1_92052	ACD_Labs	747.5	CLARITHROMYCIN	AGOYDEPGAOXOCK
0.970	v2.1_86233	ACD_Labs	717.5	Erythromycin B	IDRYSCOQVVUBIJ
0.957	v2.1_91474	ACD_Labs	733.5	Erythromycin (E-Mycin)	ULGZDMOVFRHVEP
0.953	v2.1_95226	ACD_Labs	748.5	Massbank:EA019001 Azithromycin	MQTOSJVFKKJCRP
0.950	v2.1_5797	Jeol	748.5	erythromycin A oxime	KYTWXIARANQMCA
0.950	v2.1_96020	ACD_Labs	687.4	Massbank:KO003661 Oleandomycin	RZPAKFUAFGMUPI
0.947	v2.1_21508	Jeol	832.5	lankamycin	JQMACDQCTNFQMM
0.942	v2.1_94284	ACD_Labs	834.5	MLS001074061-01!DIRITHROMYCIN	WLOHNSSYAXHWNR
0.942	v2.1_86234	ACD_Labs	719.4	Erythromycin C	MWFRKHPRXPSWNT
0.941	v2.1_94617	ACD_Labs	861.5	MLS001074901-01!erythromycin ethylsuccinate	NSYZCCDSJNWWJL
0.938	v2.1_87632	ACD_Labs	876.6	Megalomicin A	LRWRQTMTYVZKQW

4.3.3. Resonance Prediction and Assignment from SMART Results

Although the SMART platform can attempt to assign structures to their HSQC profiles, it does not perform assignments of the data to the suggested scaffolds. Without access to the training data and molecular assignments for each molecule, this data must be generated for comparison to assign a positional identity to the SSFs derived from MADByTE. NMR shift prediction from the proposed compounds was seen as a viable method to access these assignment data.

Despite the availability of NMR simulation software, few utilities which can generate high quality prediction data are amenable to high throughput automation. ACD Labs Predictor, which was used to generate the simulated data used for SMART training, is a manual use tool requiring several steps and does not provide any utilities for automated processing. Other prediction utilities provide extremely accurate values through molecular modeling but come at the cost of computational power and time which cannot be afforded for high

throughput automated analyses and are more suited to the generation of chemical shift libraries for reference.[90]

NMRshiftDB2, a database containing depositions of NMR data freely available to the community, also hosts several simulation utilities which offer a variety of methods for NMR prediction such as HOSE code and neural network based predictions for ^{1}H and ^{13}C shift prediction.[41,42,91] These utilities are open to public use and are provided independently of the database through self contained Java archives, allowing for use on a local machine. HOSE code prediction methods predict chemical shift values based on surrounding chemical environments, and work by relating similar chemical environments across many molecules. Predictions are generated rapidly compared to molecular modeling approaches and are generated by supplying a SMILES string as structural input, which is provided within the SMART results.

Because the HOSE code prediction utility relies on user submitted assignment data as a basis for shift prediction, the prediction quality is dependent on the number of times a particular chemical environment has been previously assigned. In some cases, only a few reference points can be collected which have considerably different shift values, causing the difference between the minimum and maximum values to be substantial (Table 4.3). Predictions were scored as good(0-5 ppm), fair (5-8 ppm), poor (8-15 ppm), and bad (15+ ppm) depending on the minimum and maximum values obtained through HOSE code predictions. Often, the mean values provided by the prediction are of sufficient quality for comparison, however, there remain cases where these shifts differ from published assignment data.

Table 4.3. ^{13}C Predictions for Erythromycin Via HOSE Code Shift Prediction

Atom ID	Min	Mean	Max	Prediction Range	Prediction Range Quality
0	9.8	9.8	9.8	0.0	ND
1	27.7	27.7	27.7	0.0	ND
2	62.0	83.8	94.1	32.1	Bad
4	172.4	172.9	173.5	1.1	Good
6	37.2	43.1	45.9	8.7	Poor
7	11.1	12.6	15.8	4.7	Good
8	69.5	79.5	92.7	23.2	Bad
10	95.7	98.5	101.8	6.1	Fair
11	43.3	45.2	47.1	3.8	Good
12	74.0	80.5	89.7	15.7	Bad
13	19.1	21.8	24.4	5.3	Fair
15	48.6	51.9	57.2	8.6	Poor
16	63.8	74.3	84.9	21.1	Bad
18	70.6	73.4	76.0	5.4	Fair
19	13.3	17.7	19.0	5.7	Fair
21	33.0	37.7	44.2	11.2	Poor
22	13.0	13.7	14.2	1.2	Good
23	53.4	77.8	92.7	39.3	Bad
25	88.9	101.5	107.8	18.9	Bad
27	62.9	73.1	77.1	14.2	Poor
28	9.8	20.8	24.9	15.1	Bad
29	26.3	36.1	44.6	18.3	Bad
30	67.6	67.6	67.6	0.0	ND
32	38.0	42.4	45.3	7.3	Fair
33	38.0	42.4	45.3	7.3	Fair
34	48.6	73.1	92.8	44.2	Bad
36	73.7	73.9	74.2	0.5	Good
37	20.4	23.8	24.1	3.7	Good
39	24.7	37.7	53.3	28.6	Bad
40	39.1	13.6	47.6	8.5	Poor
41	14.0	15.6	18.9	4.9	Good
42	214.3	214.3	214.3	0.0	ND
44	42.2	48.3	54.4	12.2	Poor
45	9.6	10.7	11.5	1.9	Good
46	70.3	74.9	81.2	10.9	Poor
48	73.7	74.3	75.4	1.7	Good
49	11.4	23.8	33.8	22.4	Bad

For every molecule with predicted shifts, a key is constructed which outlines the connectivity information for all ^{13}C atoms. These positions are then assigned the predicted shift information, yielding an atom connectivity map and their associated resonance IDs. This molecular map allows for molecular characteristics to be queried per carbon such as the expected number of connected ^{1}Hs (providing multiplicity), and the identities of adjacent carbons.

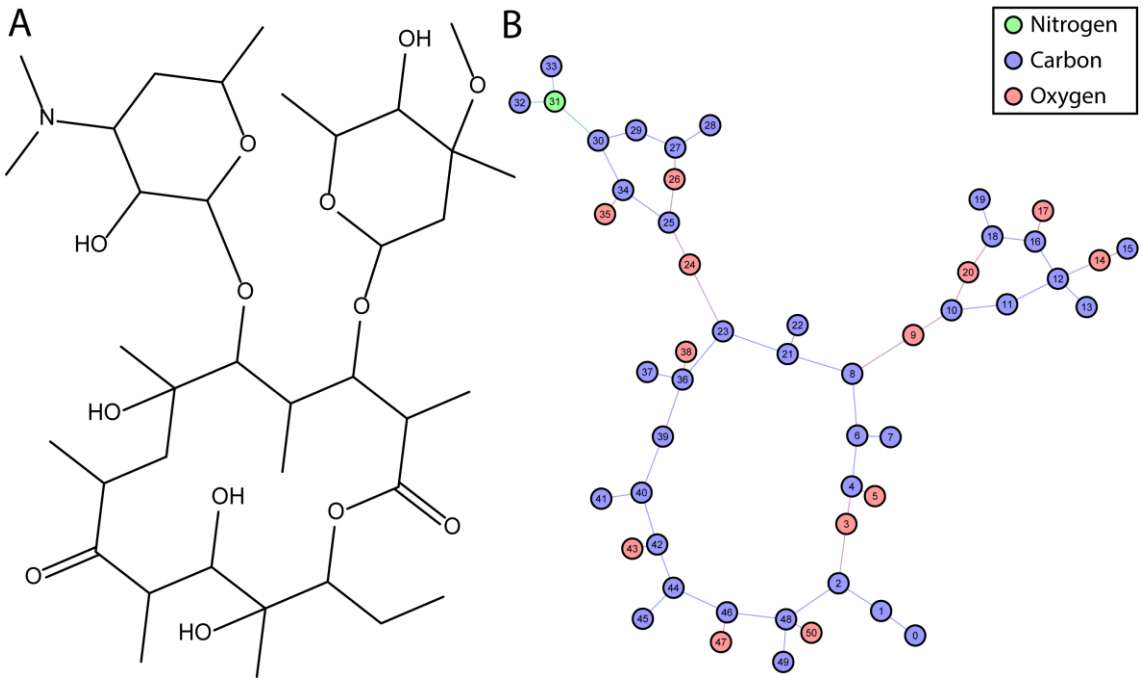


Figure 4.3. Erythromycin (A) and the Resulting Carbon Connectivity Map (B). The molecular map allows for visual reference for predicted chemical shifts provided in SHIMS, access to expected multiplicity for each carbon position, and context for adjacent carbon positions.

4.3.4. Assignment of MADByTE Spin System Features

Assignment of SSFs from MADByTE is done from a "feature first" perspective, attempting to map a given spin system feature to tentative molecules returned from SMART. Following the construction of the reference map of the tentative molecule, resonances are retrieved for every position in the SSF. The tentative molecule is then searched for ^{13}C resonances which have predicted mean shifts close to a ^{13}C resonance in the SSF. If a match is made, the adjacent carbon is compared to other ^{13}C resonances in the SSF; this is repeated until there are no more ^{13}C positions remaining in the SSF, and the plausible SSF is returned.

As many portions of a scaffold may contain areas where ^{13}C shifts are similar, the returned results are further analyzed utilizing multiplicity information. As MADByTE retains the phase information provided by the HSQC, each position which matches ^{13}C shift information is compared to the expected multiplicity. For example, a spin system whose phase is entirely positive could contain no CH_2 moieties, therefore any plausible assignments which rely on CH_2 positions in the tentative molecule are dropped from consideration.

Even with these filtration steps, it is plausible that several portions of a molecule are returned as valid if they have similar ^{13}C resonances and phase patterns. To provide the best prediction of identity, the remaining plausible spin systems are compared to the SSF shift values directly, and the assignment in best overall agreement is returned as the best match for a given spin system in each molecule. An overview of the assignment process can be seen in Figure 4.4, detailing filtration requirements.

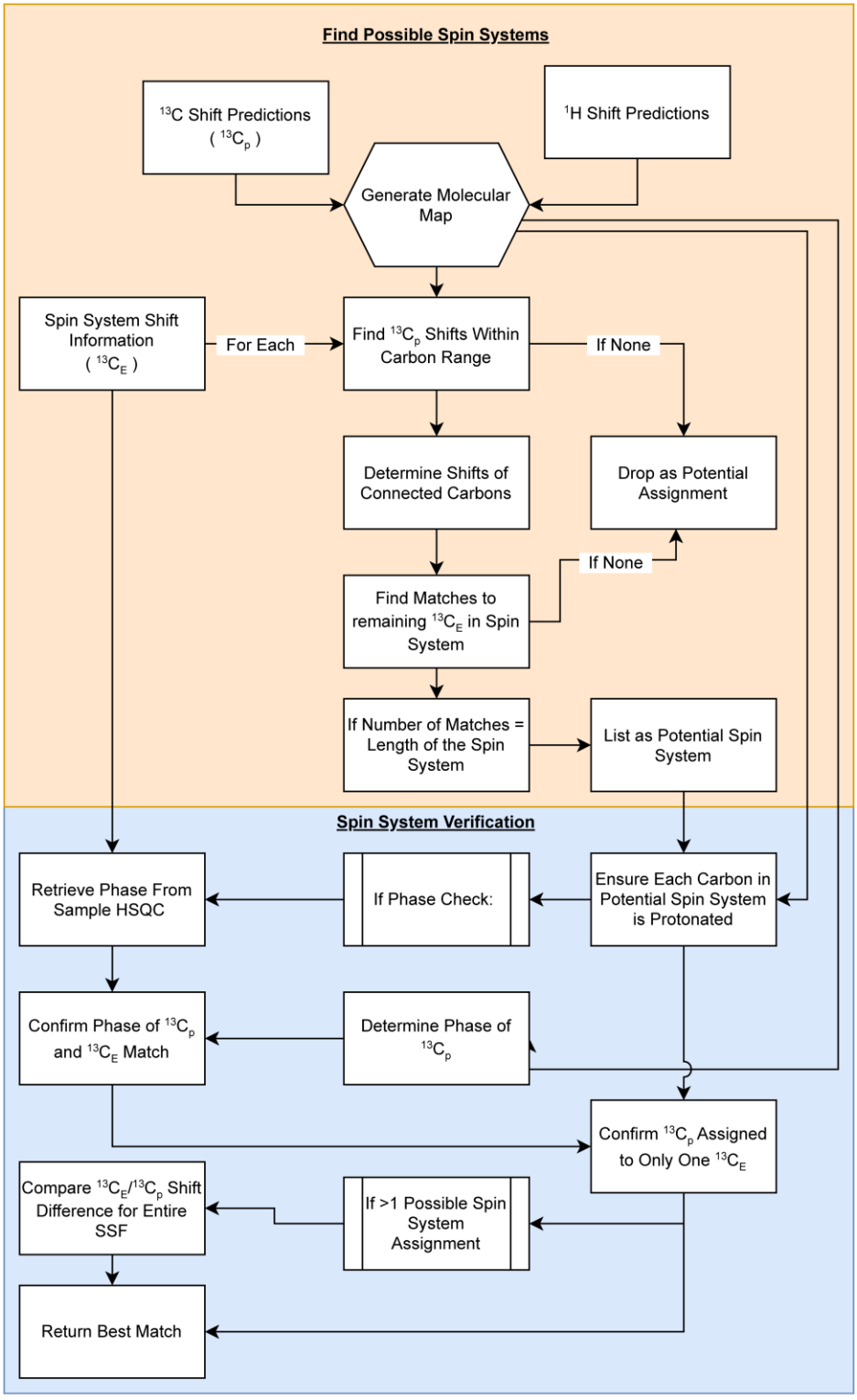


Figure 4.4. Overview of Spin System Feature Assignment. Spin system features (SSFs) are compared to tentative matches from each scaffold through ^{13}C resonance matching, phase agreement, and quality of shift agreement.

4.3.5. HSQC Scoring

Although SMART compares experimental HSQC spectra (from user input) to reference HSQC data, the result returned does not contain information on what resonances are present and which are missing. To provide users with context as to which HSQC resonances are present, the synthetic HSQC constructed for each molecule is compared against the sample HSQC data and visualized on the molecule structure in the SHIMS results.

As noted in Section 4.3.3, there are often cases where the prediction quality can vary greatly within a molecule and comparison of the full synthetic HSQC spectrum would yield many false positive associations. To ensure only predictions which are the most robust are compared against the experimental data, resonances which have predicted ^{13}C value ranges greater than 8 ppm, or ^{1}H predictions with prediction ranges greater than 1.0 ppm are removed prior to HSQC mapping and scoring.

4.3.6. Molecular Class Prediction

SMART prediction of compounds based on HSQC pattern matching provides a method for prediction of molecular class from these data, as the results returned from SMART often contain repeated classes. Although SMART does not provide a breakdown of the represented molecular classes, a neural network tool called NPClassifier is able to predict molecular superclass from SMILES representations.[92] By categorizing the top 100 results from SMART into their molecular superclass, it is possible to gauge which molecular families are heavily represented and therefore are plausible descriptors.

The prediction of molecular class by this method may provide important information in the investigation of unknown compounds. Although a definitive identity may be beyond the scope of SMART, prediction of probable molecular class can enable access to expected UV-Vis spectroscopy profiles and a plausible starting point for structure elucidation efforts. Molecular class predictions and their representation in SMART results may be useful but should also be used with caution as overrepresentation in SMART by a molecular superclass may bias results.

4.4. Proof of Principal: Pure Compounds

The MADByTE data from the pure compound network leveraged for MADByTE development in Chapter 2 was used for the development and evaluation of the SHIMS module. To evaluate the performance of SHIMS on pure compound assignments, spin system assignments made by comparison to reference data were evaluated independently by the SHIMS platform.

4.4.1. Erythromycin, Roxithromycin, and Azithromycin

SMART results from the erythromycin data returned the top 100 candidates that the system gauged as possible identifications. The top 10 results from SMART all possessed cosine scores of >0.95 (excellent agreement), but the top result was clarithromycin – not the expected erythromycin A. As all results in the top 10 were matches to predicted HSQC spectra, it is reasonable to conclude that the top candidates proposed by SMART still require a reasonable level of scrutiny and that accuracy will improve as more experimental data is provided. Of the top 100 compounds, 89 were viable candidates for categorization by NPClassifier, allowing for prediction of the molecular class of the compound (Table 4.4). Overwhelmingly, SMART recognizes the data as originating from a macrolide superclass, which is the correct designation for erythromycin A.

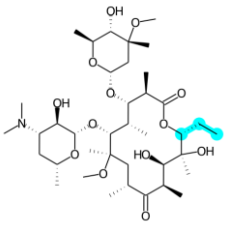
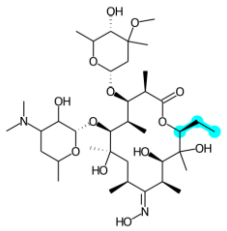
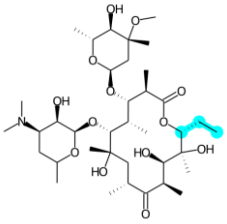
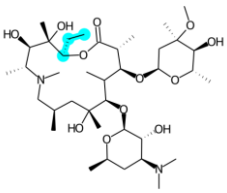
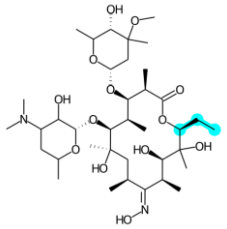
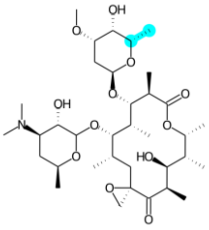
Table 4.4. NPClassifier Classifications of SMART Candidate Molecules from Erythromycin A HSQC Data

Molecular Superclass	Number of Candidates in SMART Results
Macrolide	63
Polyether	7
Linear polyketide	5
Steroid	7
Cyclic polyketide	4
Oligopeptide	1
Fatty acyl	1
Diterpenoid	1

Manual investigation of the spin system features from erythromycin attributed Erythromycin_0 to the ethyl appendage at the lactone junction through comparison to reference data. SHIMS processing conducted on this sample found 6 compounds with

cosine similarity scores above 0.95 that contained at least one substructure hypothesis for this spin system (Table 4.5), and all but one were ethyl appendages at the lactone linkages in macrocyclic compounds. SHIMS was unable to correctly find assignments for Erythromycin_6, the cladinose substructure. Review of these data showed that the predictions for ^{13}C for this molecule were further from the real values than allowed by the comparison in positions 10 and 11 (Figure 4.3, panel B). Position 10 had a reasonably close chemical shift prediction (Min: 95.7 ppm, Mean: 98.5 ppm, Max: 101.8 ppm) to the reported value of 95.8 ppm while position 11 was predicted as having a ^{13}C shift of (Min: 43.3 ppm, Mean: 45.2 ppm, Max: 47.1 ppm), but reference data places this ^{13}C shift as 34.8 ppm in DMSO-d_6.[93]

Table 4.5. Proposed Substructures from SHIMS for the Spin System Erythromycin_0

Compound	Cosine Score	HSQC Match Score	Proposed Substructure
Clarithromycin	0.972	0.47	
Erythromycin B	0.970	0.55	
Erythromycin A	0.957	0.5	
Azithromycin	0.953	0.40	
Erythromycin A oxime	0.950	0.44	
Oleandomycin	0.950	0.63	

4.4.2. Obtaining SMART Results from Complex Samples

If samples processed by MADByTE share spin system features, then there are verified HSQC and TOCSY features that are common between them that may extend beyond the SSF. Therefore, comparison of common points in the complex HSQC profiles of these samples may be sufficient to achieve enough scaffold coverage to highlight potential structures in SMART.

Comparison of the full HSQC spectra of the three macrocyclic compounds from chapter 2, azithromycin, erythromycin, and roxithromycin was done to evaluate this strategy. These compounds were chosen as some resonances associated with the core scaffold are shared between spectra and others differ by considerable amounts (Figure 4.5).

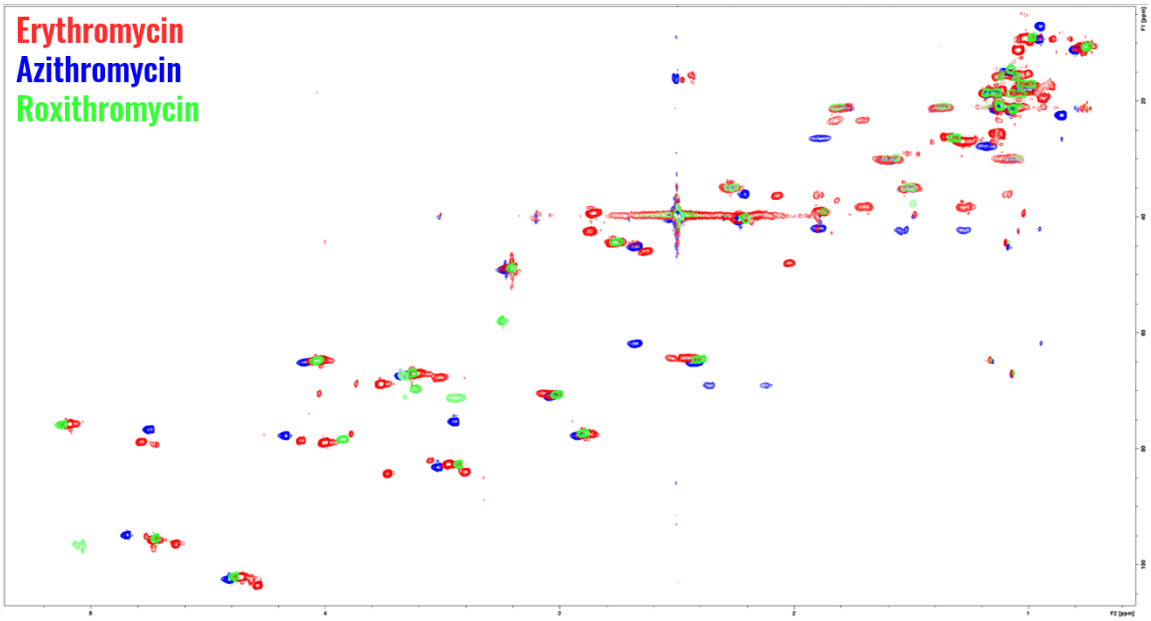


Figure 4.5. HSQC Stack Plot of 3 Macrocyclic Antibiotics, Erythromycin (Red), Azithromycin (Blue), and Roxithromycin (Green).

To generate the profile of common points between them, unique points in each spectrum are subtracted, thereby creating a new consensus HSQC displaying the homology (Figure 4.6). As with MADByTE comparison, the similarity requirements between any two points are adjustable with ^{1}H and ^{13}C ppm values. To maximize the amount of plausible overlap

due to signal shifts, these error values are wider than the suggested cut-off values for MADByTE analysis (H_ppm: 0.05, C_ppm: 3.0 ppm).

Using the resultant consensus HSQC (Figure 4.6 – Panel D) as the input for SMART provided results which were considerably more informative. The homology of points between the three different macrocycles provided enough datapoints for SMART results with high cosine scores. Encouragingly, 17 out of the top twenty results were macrocyclic compounds (Table 4.6), demonstrating that the shared features between these compounds were enough to correctly predict the compound class they represent.

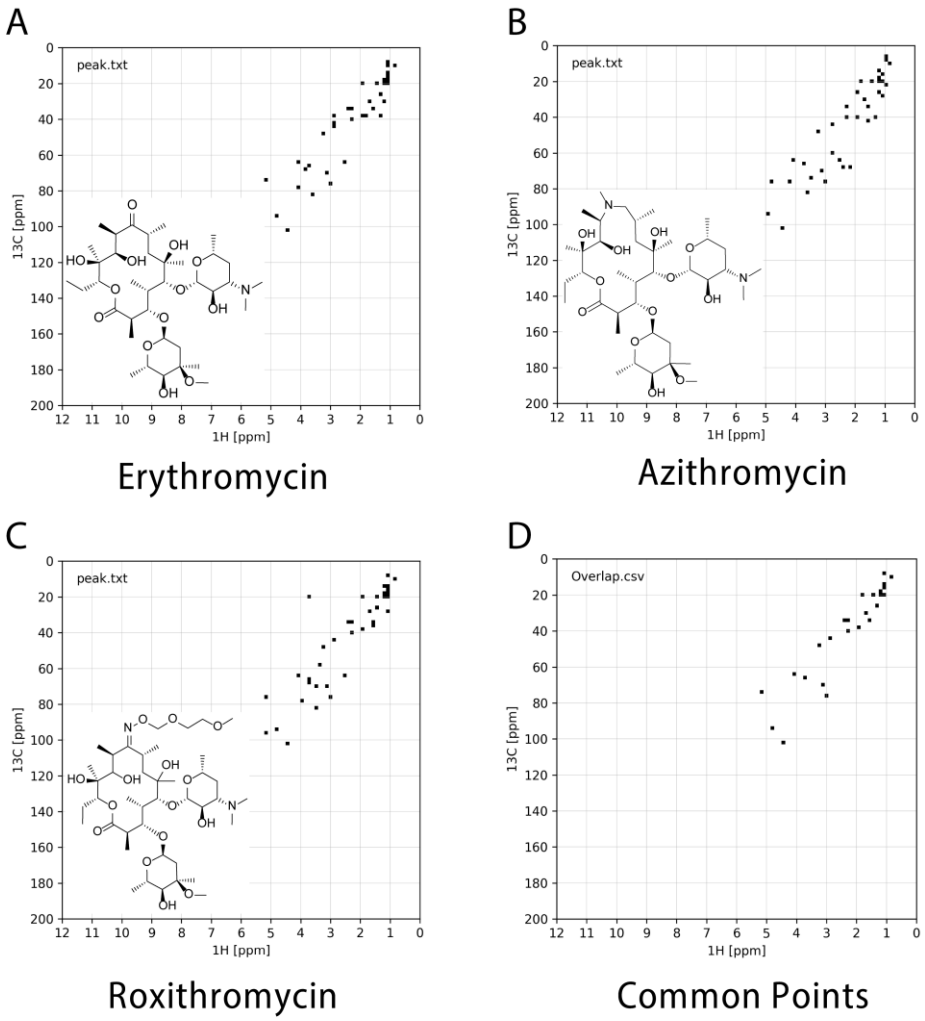


Figure 4.6. Comparison of 3 Macrocyclic Compounds (A: Erythromycin,B: Azithromycin,C: Roxithromycin) HSQC Spectra Yield a Consensus HSQC Representing Common Elements (D)

Table 4.6. Top 20 Results from SMART Obtained by Querey of the Synthetic HSQC Generated from Macrocyclic Compounds After Duplicate Filtration

Cosine Score	DBID	MW	Compound Name	Molecular Superclass
0.950	v2.1_92052	747.5	Clarithromycin	Macrolides
0.947	v2.1_91474	733.5	Erythromycin	Macrolides
0.938	v2.1_94617	861.5	Erythromycin ethylsuccinate	Macrolides
0.923	v2.1_86233	717.5	Erythromycin B	Macrolides
0.920	v2.1_5797	748.5	Erythromycin A oxime	Macrolides
0.913	v2.1_21508	832.5	Iankamycin	Macrolides
0.913	v2.1_95226	748.5	Azithromycin	Macrolides
0.899	v2.1_87634	960.6	Megalomicin C1	Macrolides
0.896	v2.1_94287	836.5	Roxithromycin	Macrolides
0.894	v2.1_96020	687.4	Oleandomycin	Macrolides
0.894	v2.1_83505	546.3	Antibiotic A-31438	Macrolides
0.893	v2.1_87632	876.6	Megalomicin A	Macrolides
0.892	v2.1_85972	527.3	Dihydropicromycin	Macrolides
0.886	v2.1_23179	500.3	Dolabriferol C	Linear polyketides
0.885	v2.1_9063	715.5	Anhydroerythromycin A	Macrolides
0.882	v2.1_87633	918.6	Megalomicin B	Macrolides
0.879	v2.1_94284	834.5	Dirithromycin	Macrolides
0.876	v2.1_86234	719.4	Erythromycin C	Macrolides
0.865	v2.1_102388	715.5	Erythromycin A enol ether	Macrolides
0.855	v2.1_9112	752.4	Marsdenoside D	Steroids
0.846	v2.1_3794	689.4	7-hydroxy-6-demethyl-6-deoxy-erythromycin D	Diterpenoids

All three of the compounds used for comparison were represented in Table 4.6, demonstrating the ability of spectra comparison to retrieve plausible results for compound overlap. Further, SHIMS processing of the these results was able to retrieve plausible substructures for a shared spin system by these molecules from MADByTE networking (Erythromcyin_0: Figure 4.7, Azithromycin_0: Supplemental Data: Figure 4.11, Roxithromycin_0: Supplemental Data, Figure 4.12). Substructure prediction was limited to potential compounds with a SMART cosine score of greater than 0.85. Notably, all predicted substructure matches were to macrolide compounds.

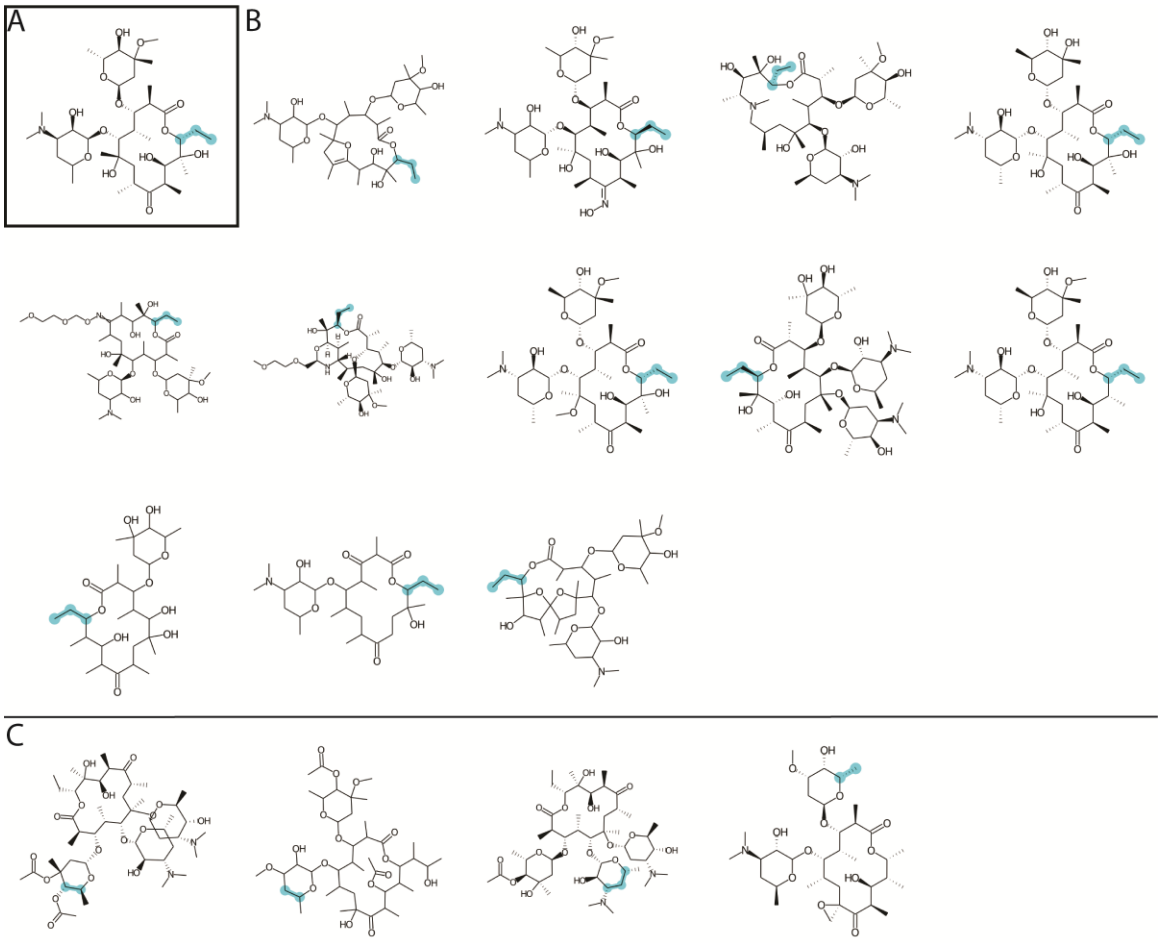


Figure 4.7. Substructure Hypotheses for Spin System Feature Erythromcyin_0 From Comparison of Three Macrocyclic Compounds. A) The correct substructure predicted on the correct molecule from which this SSF was derived. B) Correct substructure hypotheses on incorrect molecular entities. C) Incorrectly predicted substructures.

The results of this comparison demonstrate the ability of the SHIMS processing to find plausible substructures for spin system features when comparing compounds from a pure compound perspective. The HSQC comparison approach was able to retrieve plausible compounds from the SMART platform, suggesting this approach may be applicable to more complex cases where SMART analysis between more complex samples may fail.

4.5. Application to Complex Mixtures

To evaluate the ability of SHIMS to provide substructure hypotheses for spin system features derived from complex samples, the samples from Section 4.4.2 which were spiked with erythromycin were compared to gauge the ability of the platform to correctly identify erythromycin as a tentatively shared metabolite, as well as provide an accurate hypothesis for the spin system features from these complex samples. HSQC spectra from RLUS 1814E_SPK, 1526A_SPK, and 1726C_SPK were compared for common points (Figure 4.8). The SMART results from this consensus HSQC were promising, with erythromycin returned as the top result with a cosine score of 0.83.

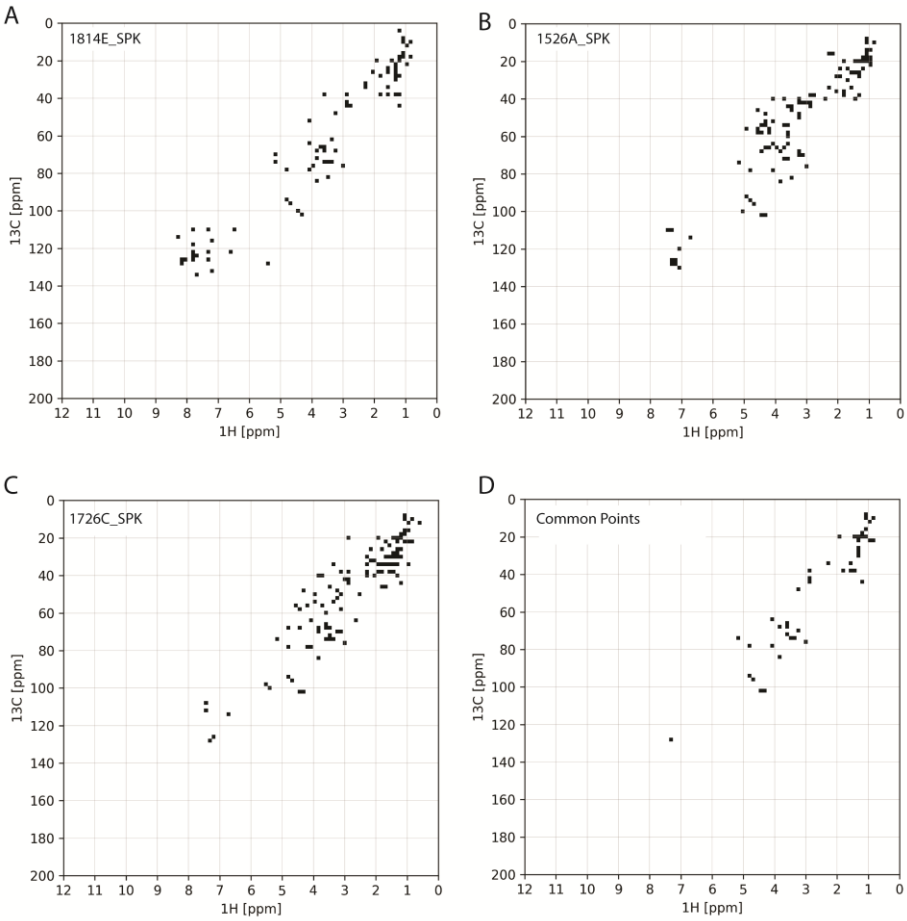


Figure 4.8. HSQC Peaks and Resulting Overlap of Extract Prefractions Spiked With Erythromycin. Peak lists from 1814E_SPK (A),1526A_SPK (B), and 1726C_SPK (C) were compared for common points (^{1}H Tolerance: 0.05 ppm, ^{13}C Tolerance: 1.0 ppm) yielding the consensus HSQC for SHIMS analysis (D).

From the spiked compounds MADByTE network, one node from these samples, 1814_E_SPK_0, connected to the known spin system erythromycin_0, which had previously been correctly predicted by SHIMS. Substructure predictions were limited to the top 5 returned results from SMART, and only three of these contained substructures where the SHIMS system predicted could be plausible for assignment. All three predictions showed the correctly predicted motif as the ethyl appendage in a macrocyclic compound (Figure 4.9). Similar, but truncated substructure predictions were found for the spin system features 1526_A_SPK_0 and 1726_C_SPK_2 although these SSFs are not connected to the erythromycin_0 SSF in the MADByTE network. Manual investigations of the resonances in these SSFs indicated that the shifts attributed to this motif are present in these SSFs, but are combined with other resonances due to spectral overlap.

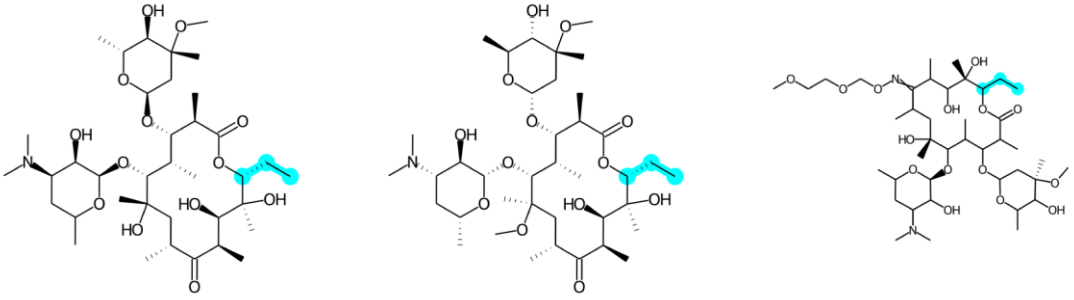


Figure 4.9. SHIMS Predicted Substructures for SSF 1814_E_SPK_0. All three proposed substructures were predicted to be ethyl appendages at the lactone junction in macrolide compounds.

To test the effectiveness of SHIMS on prefractions that have not been spiked with known compounds, 1565C and 1565D were compared as they were previously shown to contain novobiocin by MADByTE networking. The results of SHIMS showed that although novobiocin itself was not in the SMART results from HSQC comparison, very similar compounds were returned which contained several major substructures also present in novobiocin. The MADByTE network (Figure 4.10) showed that novobiocin was linked to these prefractions through Novobiocin_0 linkages to the extract spin systems 1565_D_3 and 1565_C_0.

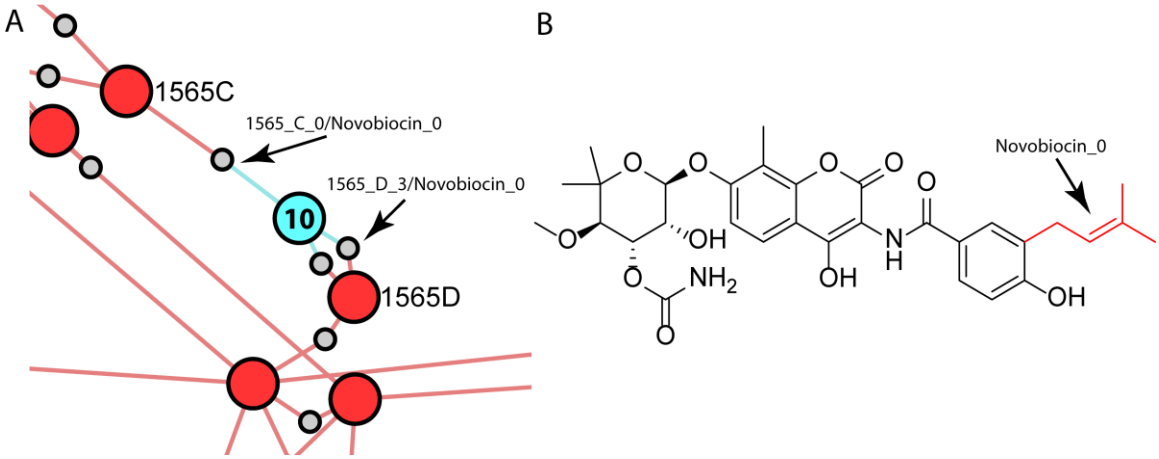


Figure 4.10. Hybrid Node Network Linking Novobiocin to Prefractions RLUS 1565C and RLUS 1565D. The spin system linked to both extract prefractions, Novobiocin_0 (A), pertains to the isoprene subunit of novobiocin (B).

The spin system Novobiocin_0 pertains to an isoprene subunit on novobiocin, and this subunit was represented in many of the top SMART results. However, neither of the spin systems derived from these prefractions were able to be mapped onto this motif. This represents a challenging case, as the assembly of the spin system Novobiocin_0 was done based on the coupling seen in the TOCSY spectra. In this case, the allylic coupling across the isoprene unit was strong enough to generate valid TOCSY resonances, generating a spin system across a quaternary carbon position. This spin system therefore cannot be mapped accurately by SHIMS, as the substructure prediction does not make assumptions as to whether allylic coupling can be allowed.

4.6. Limitations of SHIMS

The SHIMS module presents a promising case for the description of individual metabolites in complex mixtures but carries significant limitations which must be overcome for future development and adoption to be practical. Perhaps the most notable limitation of SHIMS is in the prediction of NMR shifts from the provided SMART data. Although amenable to high throughput automation, the predictors used in the SHIMS pipeline do not provide high accuracy predictions for all chemical environments. This limitation is driven by the lack of deposited reference data into NMRShiftDB2, as the HOSE code based approach uses these references to derive plausible shift patterns for each derived HOSE environment. An additional complication to this problem is the limited availability of these data for any

molecule in several solvent conditions, which can have major consequences on the shielding or deshielding of certain nuclei.

SHIMS relies directly on the results returned from SMART analysis, which can be restricting due to the availability of training data for the neural network. Much of the data used in the training of SMART 2.0 was generated using spectral simulation and chemical shift prediction which provides access to compounds for which there is no primary reference data. Although this contributes to the success of SMART for many cases, the simulated spectra are only made using two primary solvents, MeOD and $CDCl_3$ which can affect the ability of the platform to return results for queries made from other solvents, such as our chosen DMSO-d_6.

These limitations, both in the simulation of chemical shifts and in the availability of training data for SMART will be overcome in the near future, as the NMR community makes a concerted push towards databasing reference data in centralized repositories. As these databases are constructed and curated, the availability of high-quality reference data for utilities such as these will contribute greatly towards the accuracy of these systems and allow for the future development of computational utilities for NMR mixture analysis.

These limitations have a considerable effect on the ability of SHIMS to map MADByTE features into plausible structures, but an existing limitation is the ability of MADByTE to construct robust features in severely overlapped spectra. As demonstrated in Section 4.5, complex spectra have the potential to generate large spin system features which still allow useful networking but are not usable candidates for predictions of chemical motifs. New strategies in achieving higher resolution HSQC and TOCSY experiments, such as improved covariance processing and increased sensitivity of pure shift TOCSY experiments would bring about a considerable leap in the ability of MADByTE to construct these spin systems.

4.7. Conclusions and Future Directions

MADByTE in combination with SHIMS represents the first known strategy to derive plausible substructures from complex mixture data using 2D NMR experiments. The HSQC comparison strategy described allows for complex spectra to be consolidated into shared resonances, providing a mechanism for obtaining results from SMART through

comparison of similar spectra. The use of MADByTE networking to suggest samples to be compared allows for prioritization for this comparison strategy, removing the need to compare every sample. The integration with SMART allows for a predictive model to be generated and analysis by NPClassifier provides robust predictions of molecular class from complex sample data. Chemical shift prediction of SMART results through NMRshiftDB2 and the creation of the molecular map predicts plausible scaffold motifs which can describe the experimental shift data derived from samples through MADByTE.

SHIMS was able to correctly predict substructures in a variety of cases involving pure compounds and in mixtures, showing a great deal of promise in targeted applications. However, the inability to match potential substructures that can arise due to complex coupling, highlighted by the failure to identify the shared motif between novobiocin and 1565C and 1565D show that some limitations exist. Although SHIMS can provide these hypotheses, the number of plausible results returned by the module can be overwhelming and filtration in cases where the sample constitution is completely unknown remains a challenge. As this comparison approach provides candidates that may be shared between extracts linked through MADByTE analysis, SMART results could be filtered and compared to MS data which would drastically reduce the number of false positives.

SHIMS represents the fusion of separate utilities developed for orthogonal purposes to achieve plausible prediction of molecular scaffolds present in mixtures. The integration of these utilities to describe sample constitution demonstrates the complexity of cheminformatics when applied to real world data and highlight their importance in the changing landscape of discovery-based metabolomics. These utilities provide an NMR focused approach towards the functional annotation of complex mixtures, and integration with additional data such as proposed substructures obtained via an MS based approach,[94] could provide future opportunities for increased confidence in substructure prediction. As future utilities are developed and combined, the ability to provide more robust annotations through the combination of orthogonal data, machine learning, and molecular modeling will drastically increase the speed and ease of investigations of complex samples and their potentially important components.

4.8. Supplemental Data:

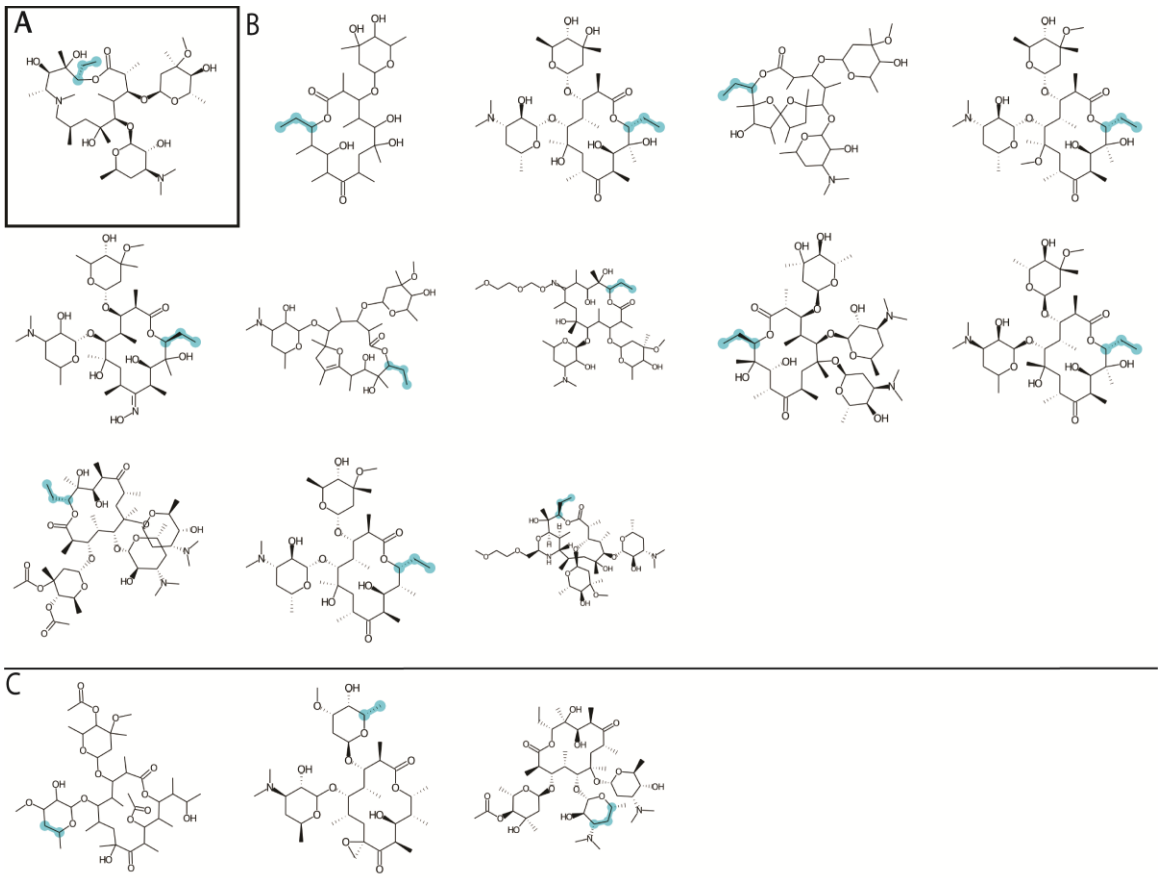


Figure 4.11. Substructure Hypotheses for Spin System Feature Azithromycin_0 From Comparison of Three Macrocyclic Compounds. A) The correct substructure predicted on the correct molecule from which this SSF was derived. B) Correct substructure hypotheses on incorrect molecular entities. C) Incorrectly predicted substructures.

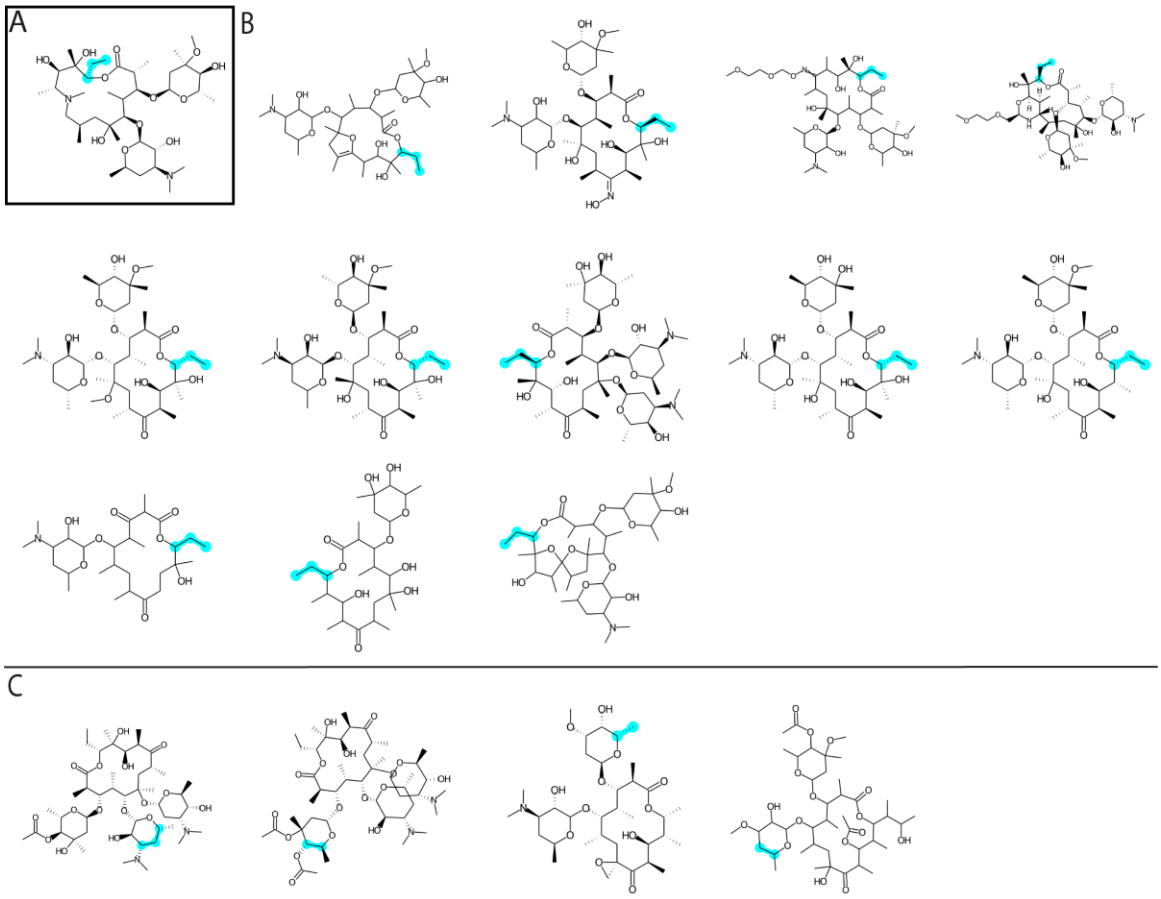


Figure 4.12. Substructure Hypotheses for Spin System Feature Roxithromycin_0 From Comparison of Three Macrocyclic Compounds. A) The correct substructure predicted on the correct molecule from which this SSF was derived. B) Correct substructure hypotheses on incorrect molecular entities. C) Incorrectly predicted substructures.

Chapter 5.

MADByTE Feature Association by Diffusion Experiments

5.1. Introduction

5.1.1. Diffusion NMR in Mixture Analysis

Natural products represent some of the most challenging cases of mixture analysis for annotation and characterization for most analytical platforms. As complex extracts, metabolites are present in variable amounts, often unknown, and represent difficult molecular scaffolds. NMR investigations into the constitution of mixtures are often aided by the addition of new methodologies such as orthogonal experiments which allow for new information to confirm or suggest the identity of molecules present in the mixture. One method of investigation which has been applied to simplified mixtures for decades is DOSY, or **D**iffusion **O**rdered **S**pectroscop**Y** which utilizes the ability of molecules to diffuse through a liquid to derive new information about molecular structure.

Diffusion Ordered Spectroscopy

The basic DOSY method is a pseudo 2D NMR experiment which allows for the determination of physical characteristics of the molecules analyzed, rather than directly establishing connectivity through bonds.[95] In solution NMR spectroscopy, molecules are suspended in a solvent and naturally diffuse within the solution during the timeframe of an NMR experiment. In most NMR experiments, this diffusion is of little consequence as the pulses are applied to provide consistent energy transfer along the z-axis and molecules experience the effect of pulses uniformly. This allows for the other variables, such as the T_1 delay in the HSQC, to be the independent variable in the experiment. In DOSY, pulses are shaped by a gradient to purposely cause a differential effect along the z-axis of the NMR tube which affects the amount of focusing/refocusing energy transferred to a given nucleus.

As molecules diffuse over the timeframe of the experiment, changes in the observed energy upon refocusing can be associated to the movement of the molecule. By altering

the power associated with this gradient pulse, nuclei which are in focus are attenuated differentially. This change in signal intensity can be compared as a function of the gradient strength, which allows for the derivation of a coefficient for a particular resonance that describes the rate of diffusion. In a single molecule, resonances would be affected uniformly by this gradient pulse, and therefore have the same diffusion coefficient. When represented as a 2D plot, the ^{1}H resonances can be associated against their diffusion coefficients, allowing for resonances of the same molecule to be determined (Figure 5.1).

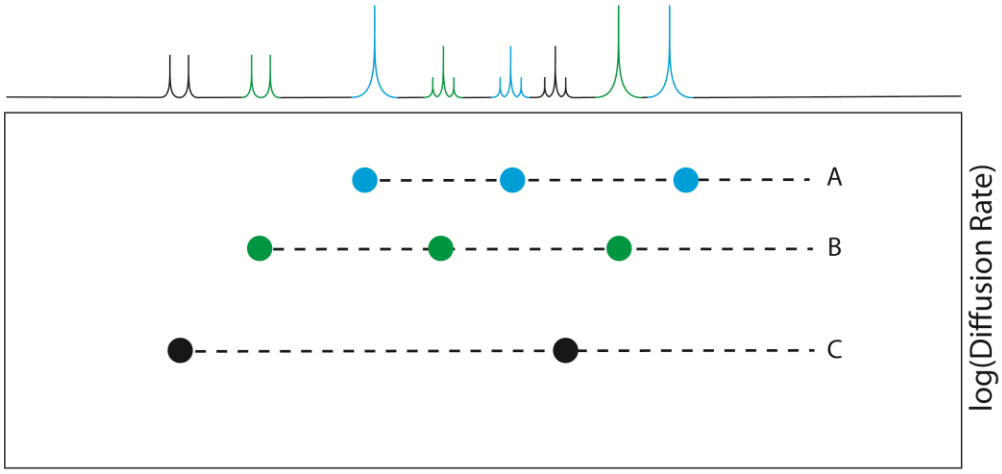


Figure 5.1. DOSY Conceptual Plot. DOSY yields a pseudo 2D NMR spectrum by plotting resonances against their derived diffusion rates. Compounds with different diffusion rates (A vs B vs C) resolve along the Y-axis due to their physical characteristics.

Several factors can determine the amount of diffusion a molecule undergoes during a DOSY experiment including the physical size of the molecule, the viscosity of the solvent, and changes in conformation. In general, larger molecules will diffuse slower due to their inability to move easily through solvent. This relationship is one of the primary uses of DOSY and is often used to estimate the size of large molecules, such as mixtures of polymers.[96,97]

Advanced DOSY

Although DOSY typically shows success in the analysis of simple mixtures, complexity of the sample is still a defining limitation. Baseline resolution of the signal is often required in order to pick and fit peaks in each 1D plane of the experiment and overlap in signals can

distort the calculations used to determine the diffusion rate of a signal.[98] Because of this, multidimensional experiments which incorporate DOSY elements have been developed.[99] Conceptually, these experiments are like standard DOSY processing in that they calculate the changes in signal intensity for a given pulse gradient strength, which can then be used to calculate a diffusion rate for the signal. However, because each signal is now separated along a secondary axis, the degree of overlap for a resonance is reduced. The result is a pseudo 3D NMR spectrum in which standard 2D NMR spectra are displayed along a third axis of diffusion rates.

This strategy of incorporating DOSY elements into standard 2D experiments is attractive for mixture analysis, but requires substantial increases in experiment collection time as each plane of the pseudo 3D spectrum contains the information from a conventionally sampled 2D dataset. As the accuracy of diffusion rate calculations depends greatly on the number of gradient pulses applied, the need for more planes can quickly outpace reasonable experimental timeframes and is still considered a major limitation to the application of DOSY for complex mixture analysis.

To address this limitation, the development of the COSY-IDOSY experiment was undertaken by Nilsson et al. in which the underlying DOSY element (BPP-LED)[100] was removed and replaced by a modified gradient enhanced COSY sequence, reducing the number of required transients by 32 fold.[101] This drastically reduced the experimental time required for collection, allowing for more practical applications towards mixture analysis.

5.1.2. Limitations of MADByTE Analysis

MADByTE analysis derives spin system features from complex samples by associating HSQC resonances through TOCSY correlations. When considering a pure compound, these spin system features represent different scaffold motifs in a relatively straightforward manner – a molecule is broken up into its smaller spin systems and each spin system would be represented as a spin system feature node in the overall network. However, considering that multiple spin systems should be present for each detected molecule within a mixture, these small motifs may not be sufficient to describe the overall character of a component within a mixture. If multiple spin system features are derived for a given molecule in a mixture, then association of these spin system features together before comparison could allow for a more detailed comparison of conserved motifs. However,

neither TOCSY nor HSQC experiments allow for further connectivity or associations to be established and therefore other experimental information would need to be incorporated to provide a way to link two otherwise independent features.

Another limitation of MADByTE processing is that ^{1}H overlap can cause features to become fused through false connections brought on by resonances within the defined ^{1}H ppm error. These spin systems are often considerably larger than would be practical and are falsely comprised of otherwise independent spin systems with a small number of resonances in common. If these resonances could be split and pooled when compared to the TOCSY spin systems, spin system features from complex samples would more accurately reflect the overall makeup of the sample by teasing out scaffold information from otherwise confounding features.

5.1.3. Feature Association by Diffusion Experiments

To address both limitations in the formation and association of spin system features from complex samples, it was proposed that combination of DOSY experiments to derive diffusion rates for signals in these data could be leveraged. With diffusion rates associated for each resonance, SSFs with the same diffusion rates can be merged (feature fusion) and confounded SSFs can be split into more refined features (feature fission). The combination of both feature fusion and feature fission forms the basis for the expansion of MADByTE known as FADES (**F**eature **A**ssociation by **D**iffusion **E**xperiment**S**).

Feature Fusion

Spin systems that would otherwise be independent entities (Figure 5.2 - Panel A) would contain resonances which match diffusion rates in other spin systems, allowing for these features to be associated together (Figure 5.2 - Panel B). Taken as larger collections (Figure 5.2 - Panel C), associations of these spin systems could better describe the molecules within the mixture than any one spin system feature and may better differentiate shared compounds.

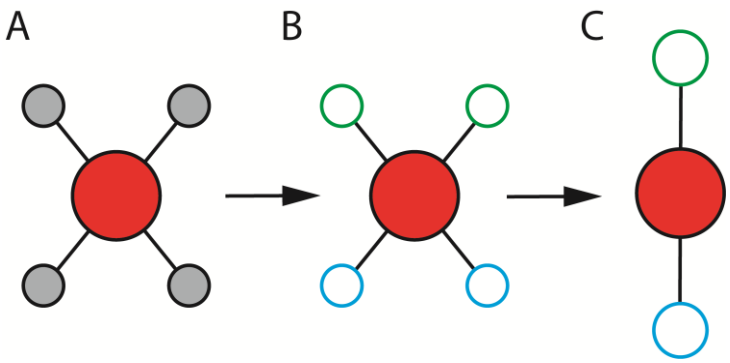


Figure 5.2. **Feature Association by Diffusion Experiments. A) Independent spin systems from MADByTE analysis could originate from one or many compounds. B) By comparing the diffusion rates of resonances within the spin system features, those arising from the same molecule could be linked and C) fused into composite features.**

Feature Fission

Feature fission would work in the opposite direction of feature fusion, by splitting complex spin systems based on diffusion rates prior to the spin system formation step (Figure 5.3). Peak lists from TOCSY would be first sorted by their corresponding ^{1}H diffusion rates, and these sub-lists would then be subjected to standard MADByTE analysis. The reduced complexity of the TOCSY would allow for otherwise overlapping systems to be split into smaller and more robust SSFs.

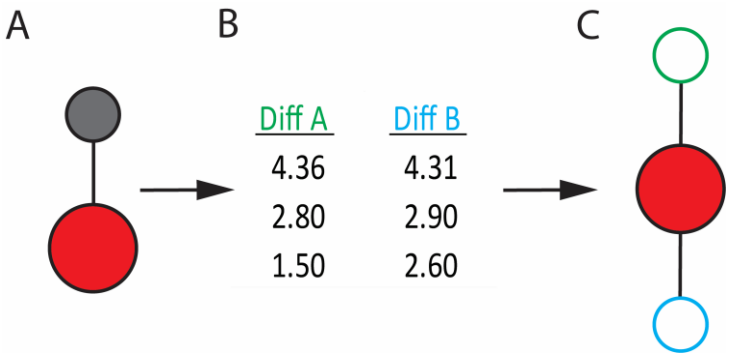


Figure 5.3. **Feature Fission by Diffusion Experiments. A) Spin system features derived from overlapped resonances contain spin systems from several compounds fused together. B) By comparing the diffusion rates of ^{1}H signals before feature creation, these complex SSFs can be split C) into smaller SSFs.**

5.2. Experimental Considerations

5.2.1. Important Parameters in DOSY Experiments

The diffusion coefficient (D) for a given molecule can be determined through the relationship of change in signal intensity as a function of the applied gradient strength (g) and length of the gradient pulse (δ), as shown in Equation 5.1 where I_0 is the initial intensity with no gradient applied, I is the observed intensity, Δ is the diffusion time allowed, and γ is the gyromagnetic ratio of the observed nuclei.[102]

$$I = I_0 e^{-D\gamma^2 g^2 \delta^2 (\Delta - \frac{\delta}{3})} \qquad (\mathbf{5.1})$$

From these relationships, molecule specific factors (D and γ) determine signal attenuation but cannot be changed experimentally. As long as the remainder of the NMR experiment remains the same, the factors affecting signal intensity that can be manipulated on a per transient basis include g, δ, and Δ. The values of Δ and δ are set for all transients but must be determined experimentally to ensure proper signal attenuation. The gradient strength (g) is calculated depending on the number of planes in the overall experiment.

5.2.2. Creation of a Model System

To evaluate the DOSY experiments, a 1:1:2 mixture of erythromycin (**4**), cycloheximide (**12**), and nystatin (**19**) was suspended in DMSO-d_6 and HSQC and TOCSY spectra were obtained. These molecules were selected to ensure the components in the system were representative of natural product scaffolds, were different in molecular size and mass, and contained multiple spin systems expected from the MADByTE processing.

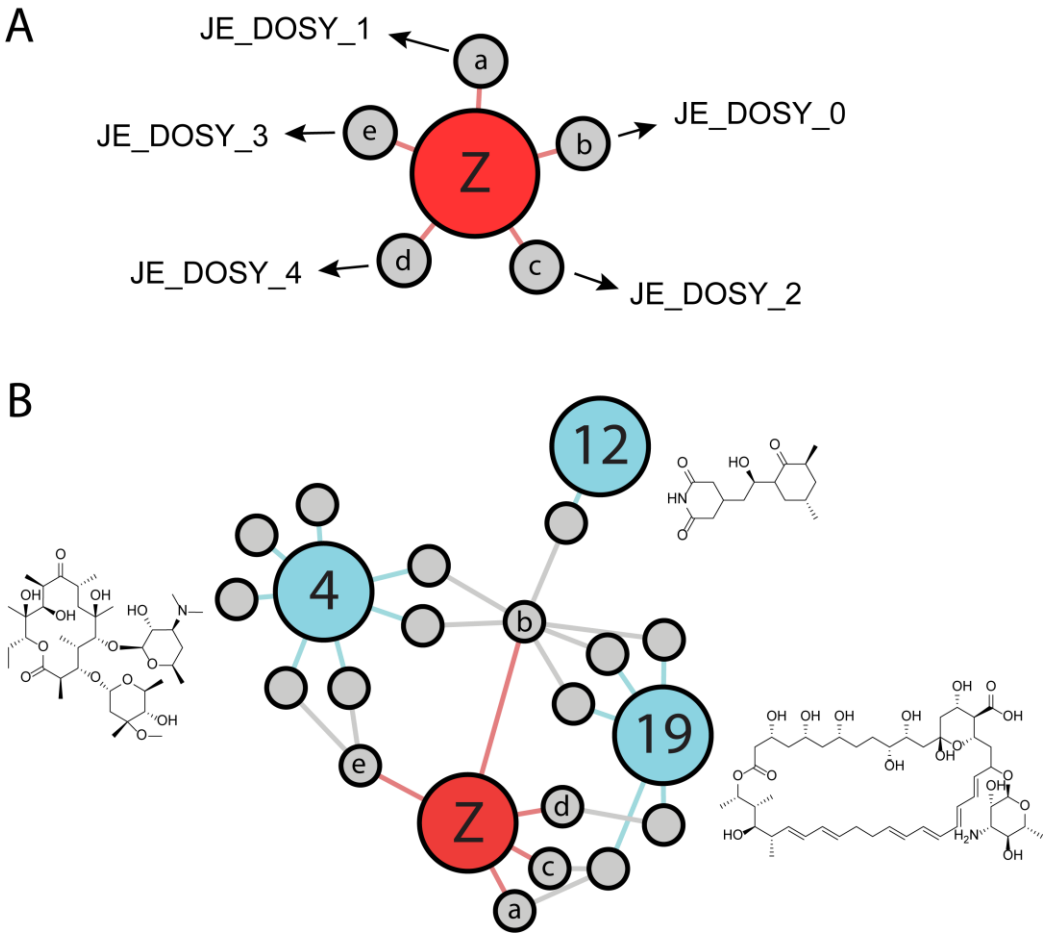


Figure 5.4. **MADByTE Network of DOSY Sample. A) The mixture of three components yielded five SSFs, including one (node b) that represents extensive overlap which may be refined through feature fission. B) MADByTE networking with standards showed five nodes originating from node Z, where nodes a,c, and d belong to nystatin (19), and should group together through feature fusion.**

Processing the model mixture in MADByTE led to the generation of five SSFs of varying confidence (Figure 5.4 – Panel A). Three spin systems were of reasonable length, nodes a (6 members long), c and d (each two members long), and the remaining spin systems were of suspect length, nodes e (11 members) and b (33 members). A summary of these memberships can be found in Table 5.4. Although these spin systems contain useful resonances, they are suspected to be the product of resonance overlap and each may hold potential for refinement. When processed alongside the reference compounds, cases for implementing both feature fission and feature fusion are made (Figure 5.4 - Panel B).

5.3. DOSY Experiments

Pseudo 2D DOSY

To evaluate the performance of a pseudo 2D DOSY, a 1:1:2 mixture of erythromycin, cycloheximide, and nystatin was prepared and analyzed with ledbpgp2s. The resulting DOSY plot showed that the separation of resonances by their diffusion rates (y-axis – logarithmic scale) performed well in areas where peaks are well resolved, such as in areas of higher deshielding. However, in areas where peaks overlap, such as 1-3 ppm, resonances were unable to be resolved into their respective diffusion rates.

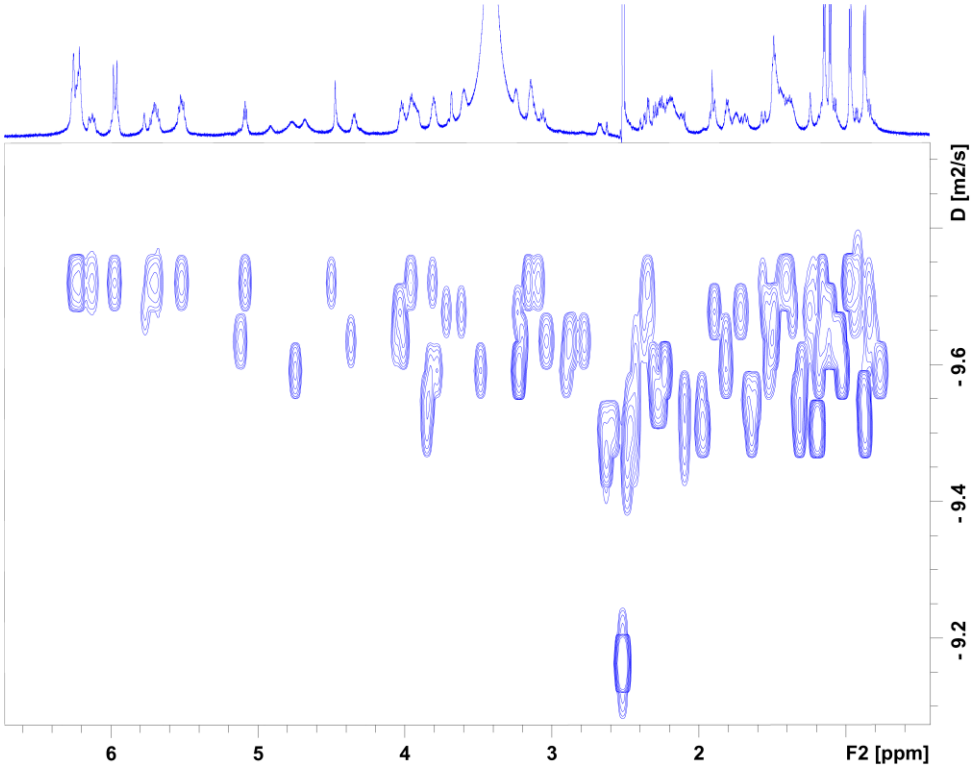


Figure 5.5. Pseudo 2D DOSY Plot of Erythromycin, Nystatin, and Cycloheximide. Well resolved resonances, such as those which are deshielded and displayed good baseline separation showed good agreement in the determination of a diffusion rate. However, areas of high complexity, 1-3 ppm, showed a reduced ability to determine diffusion rates associated with the rest of the molecule.

As this represented a simplified mixture of only three components, it was determined that more complex cases, such as the extract prefractions analyzed by MADByTE would not be able to determine consistent diffusion rates for whole molecules. As the number of

components within a DOSY mixture increases, the loss in resolution and accuracy of diffusion rate determination become difficult to manage.

Pseudo 3D DOSY

To address issues of sample complexity, advanced DOSY variants of 2D NMR experiments have been put forward, as they offer a secondary axis of resolution to otherwise overlapped signals. Positions which would experience overlap in a 1D projection are often resolved in a 2D NMR experiment, and comparison of signal decay for these positions can be better managed. The COSY-IDOSY was selected for evaluation as it offers this increased resolution advantage, is proton detected, and contains ^{1}H-^{1}H coupling information which could aid in the assignment of spin system features derived from MADByTE. A sample of 1:1:2 of erythromycin, cycloheximide, and nystatin was prepared and analyzed by COSY-IDOSY (Figure 5.6).[101] Analogous to 1D DOSY experiments, a set of COSY experiments were collected with an increase in the applied gradient pulse strength to attenuate resonances. The rate of attenuation of each signal can then be used to calculate a diffusion rate for each peak in the spectrum, and signals with similar diffusion rates can be represented as individual planes in a 3D plot. Three planes of interest arise from this processing, with their mean diffusion rates allowing for filtration of signals which are likely to arise from the same molecule (Figure 5.7).

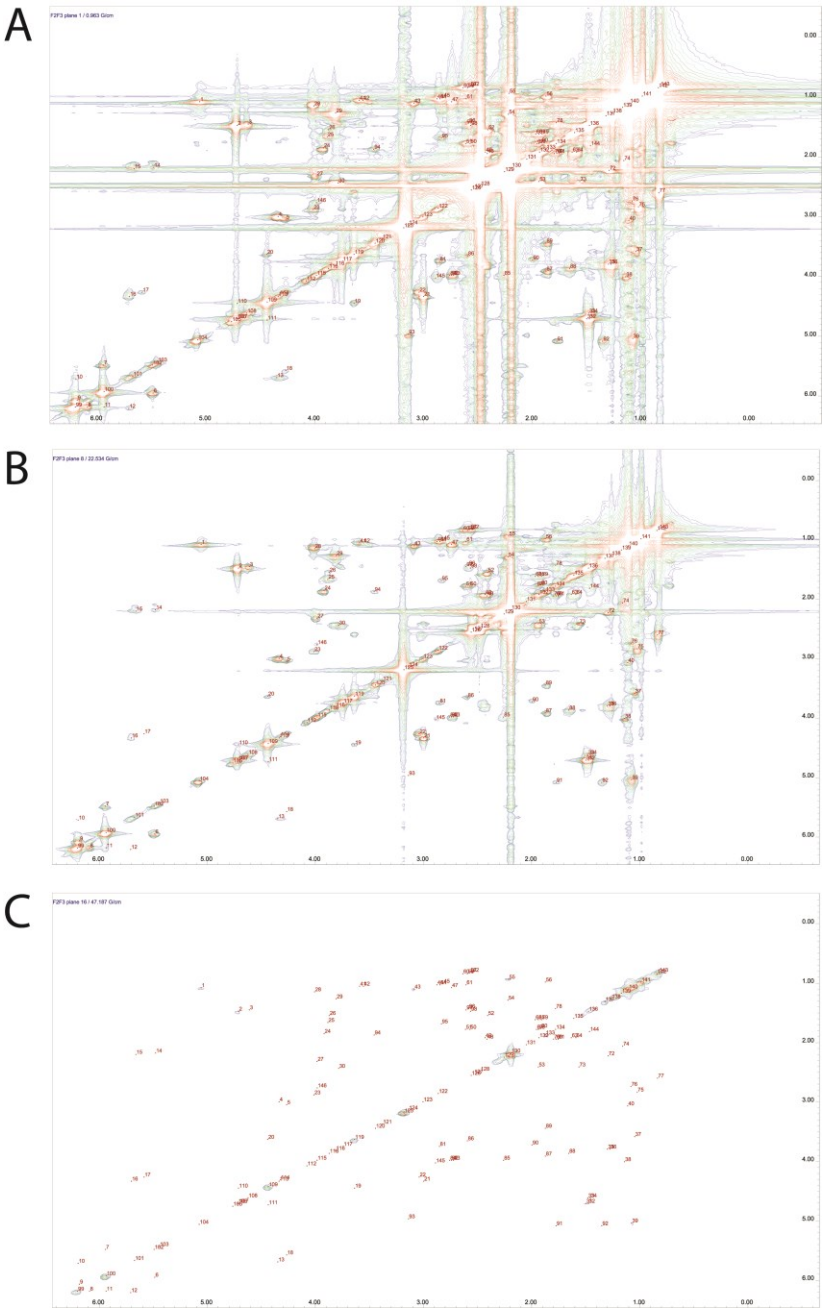


Figure 5.6. **Planes of COSY Spectra of 1:1:2 Erythromycin:Cycloheximide:Nystatin Mixture In DMSO-d_6 for COSY-IDOSY. A) With minimal applied gradient strength, the attenuation of most signals in the mixture is negligible, and peaks are at their full intensity. B) At the middle gradient strength, signals have begun to attenuate and are no longer visible at the same intensity. C) At the highest gradient strength applied, all but the most intense signals have fully attenuated and are no longer visible.**

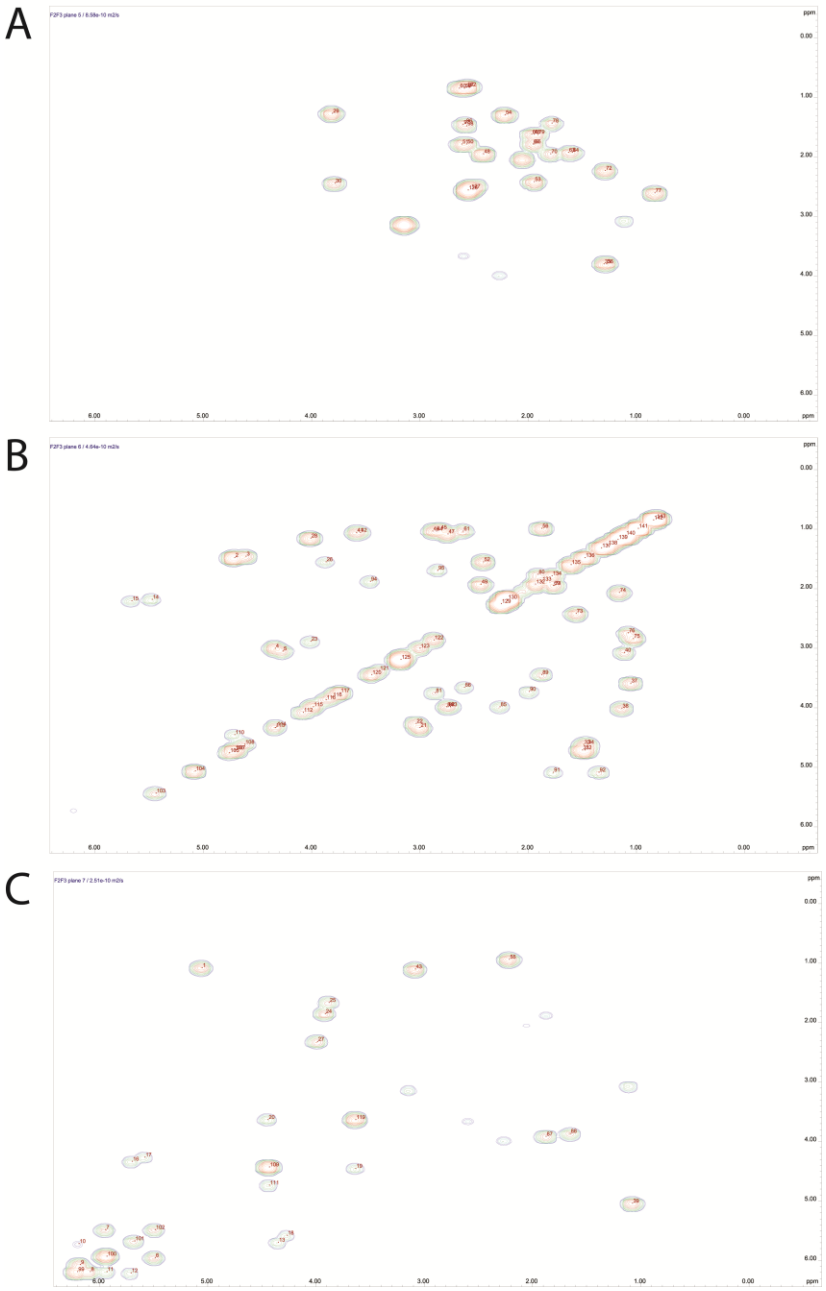


Figure 5.7. **Three Planes of the COSY-IDOSY Experiment. The 3 planes derived from the COSY-IDOSY experiment which display peaks that could be fit. The derived diffusion planes A) $8.58e^{-10}$ B) $4.64e^{-10}$ and C) $2.51e^{-10}$ m^2/s represent clustering of individual resonances which decay with similar rates.**

5.4. Association of Diffusion Data to Spin System Features

With a promising result showing three main planes of DOSY separation, the next step would be to compare these resonances to the resonances found in MADByTE spin systems to see whether distinct groupings can be validated.

5.4.1. Processing of Diffusion Data

Analysis of COSY-IDOSY Data

There are several methods of processing in Dynamics Center for COSY-IDOSY data which allow for near-automated processing to be conducted on Topspin processed data. One option available is automated peak picking which attempts to peak pick above a given threshold in every plane of the experiment. However, this process proved to be impractical, as only a few peaks from each plane were picked and fittings did not provide any tangible results. As peak picking is a notable problem in complex data, a manual solution was needed to address this data processing shortcoming. The COSY-IDOSY data were processed in Topspin to produce the 3D representation and the first plane was selected, representing the lowest applied gradient strength where no peak attenuation is expected. Peaks were picked manually following the same convention as MADByTE peak picking and exported as a peak list. This peak list was applied to all planes of the DOSY data in Dynamics Center, allowing for uniform positioning across all planes.

As can be seen in Figure 5.7-B, most cross peaks in the diagonal lined up with the middle plane, although this is likely a byproduct of the method of analysis which attempts to derive diffusion constants from these peaks. This method of analysis used the intensity of each peak at the point provided and did not attempt to integrate signals or deconvolute signals that may be overlapping. In cases where the points show good decay (Figure 5.8) diffusion rates can be calculated. If two signals are overlapped at the point a peak is picked, then the dissimilar attenuation rates may introduce error into the quality of the fit of the peak. Uncertainty in pseudo 3D DOSY spectra are represented as stretching in the F1 plane (Figure 5.9). The mean of peaks affected by this may be biased towards the middle plane of the spectrum, resulting in a falsely associated signal due to overfitting.

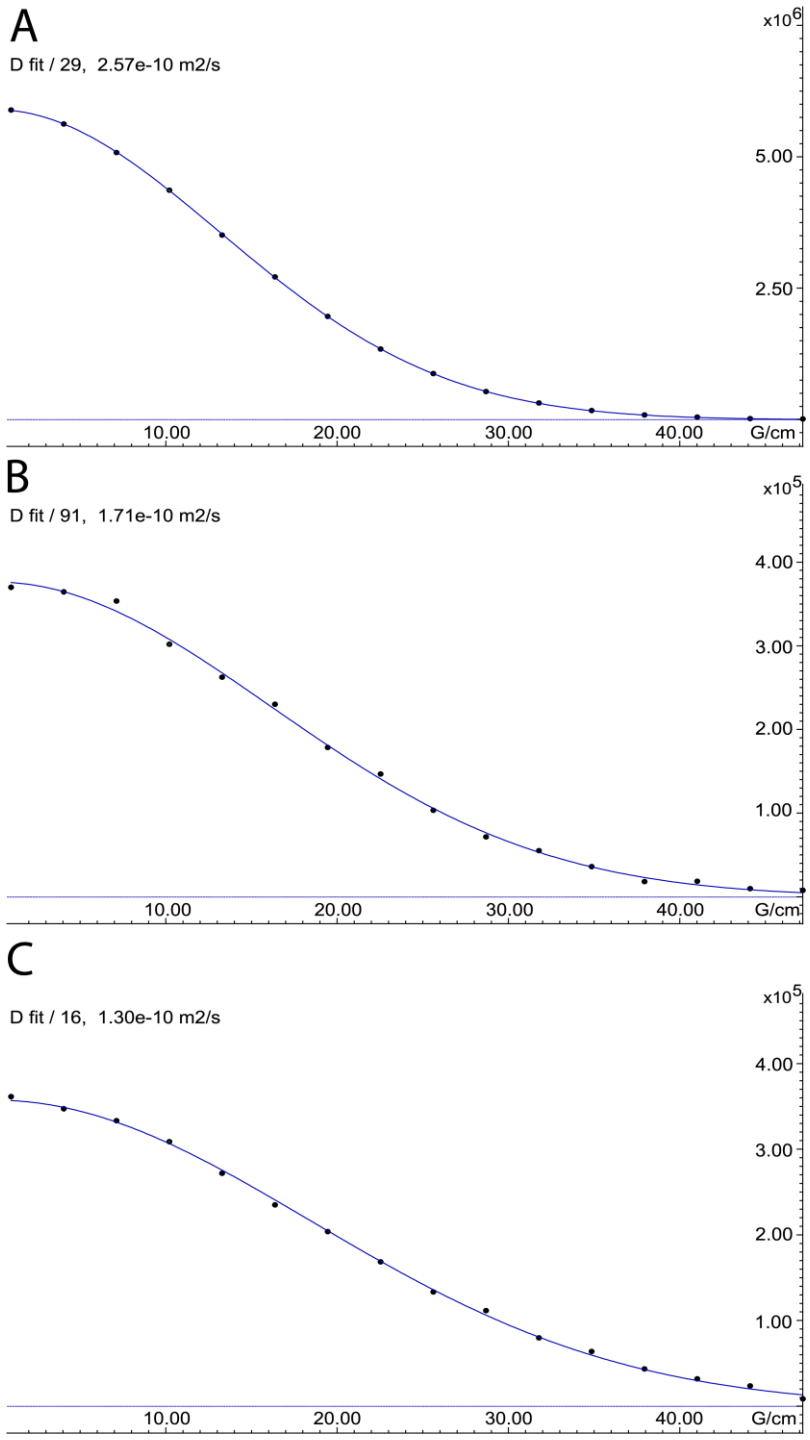


Figure 5.8. **Decay Curves for Peaks From Different DOSY Planes. A) A peak from DOSY Plane 5 shows full attenuation around 40 G/cm of applied gradient strength, A signal from Plane 6 B) shows around 90% attenuation around the final gradient strength of 47 G/cm, and a signal from Plane 7 C) shows a slower decay rate than either.**

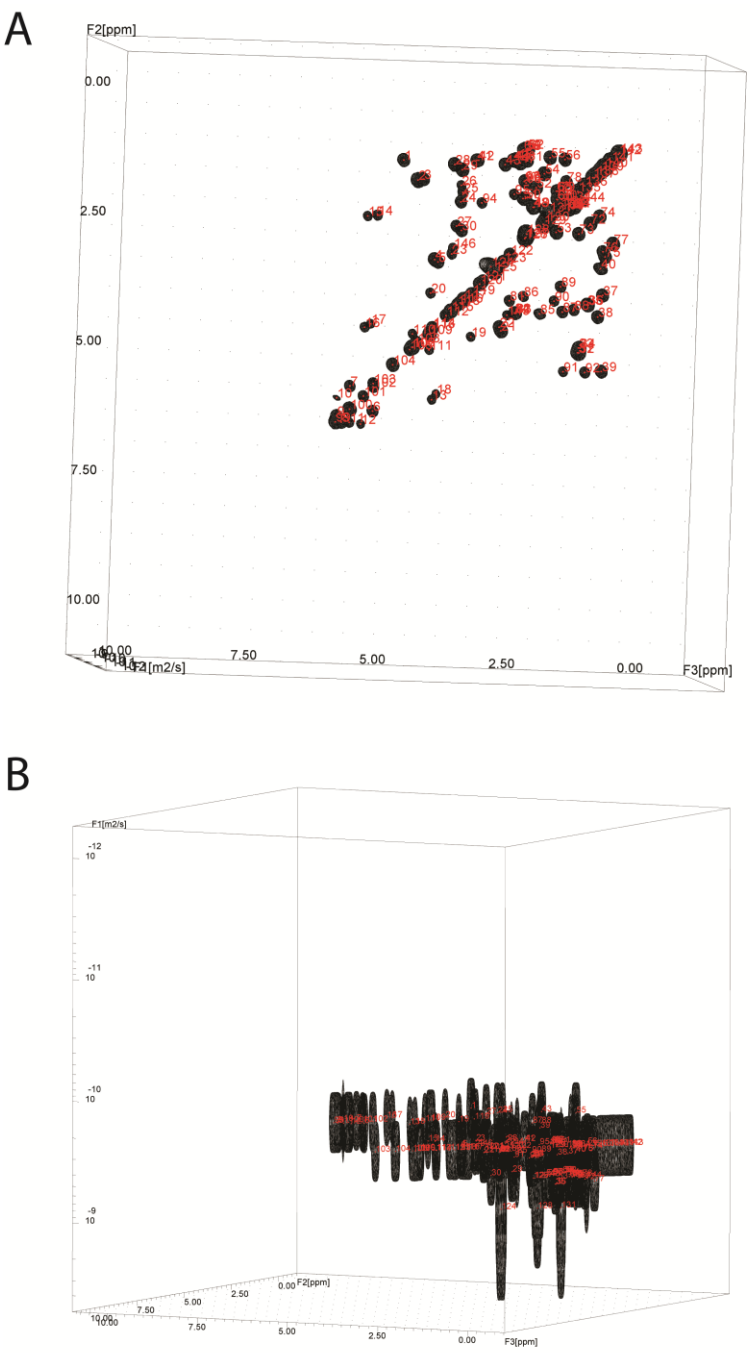


Figure 5.9. **3D Representation of COSY-IDOSY Results for Erythromycin, Cycloheximide, and Nystatin Mixture. A) Datapoints representing the picked peaks can be seen from an F2/F3 perspective and resemble a typical COSY experiment. B) Viewed from F2/F3/F1 perspective, points can be seen to align along one of 4 planes identified depending on the fit of the decay curve. The line widths along the F1 dimension represent the quality of fit, and points in the 4th plane show the lowest quality of fit, as they extend through the other planes as well.**

Table 5.1. Summary of Calculated Diffusion Rates from COSY-IDOSY Planes

Plane	Mean (m^2/s)	Min (m^2/s)	Max (m^2/s)	Standard Deviation (m^2/s)
DOSY Plane 5	2.60E-10	2.52E-10	2.83E-10	6.10E-12
DOSY Plane 6	1.87E-10	1.37E-10	2.64E-10	2.87E-11
DOSY Plane 7	1.26E-10	1.17E-10	1.33E-10	3.41E-12

Peaks assigned to each plane were recorded and the diagonal relationships were removed to calculate the mean and standard deviation of diffusion rate for each grouping (Table 5.1). As can be seen, the ranges of DOSY planes 6 and 5 overlap, which would complicate assignment of spin system points to a diffusion plane. Points were removed from further consideration if they were above or below two standard deviations of the mean in each plane. This resulted in distinct bins for each plane which did not allow for overlap (Table 5.2) which can then be independently compared against MADByTE spin system features.

Table 5.2. Refined Bins of Diffusion Planes from COSY-IDOSY

Plane	Min (m^2/s)	Max (m^2/s)
DOSY Plane 5	2.52E-10	2.68E-10
DOSY Plane 6	1.37E-10	2.43E-10
DOSY Plane 7	1.22E-10	1.32E-10

5.4.2. Spin System Association

New spin system features are created from the combination of diffusion data and MADByTE SSFs. Each point in a SSF is compared against peak lists from a DOSY plane for points which are within the defined ^1H error. As each diffusion plane represents COSY correlations, when a match to ^1H_a is made the associated ^1H_b resonance is cross checked

against the SSF to ensure it is present. If both are confirmed, the resonance is added to a new DOSY associated SSF and the process is repeated for remaining ^{1}H resonances (Figure 5.10). Once comparison of all SSFs from a sample is completed, the process is repeated for the remaining DOSY planes.

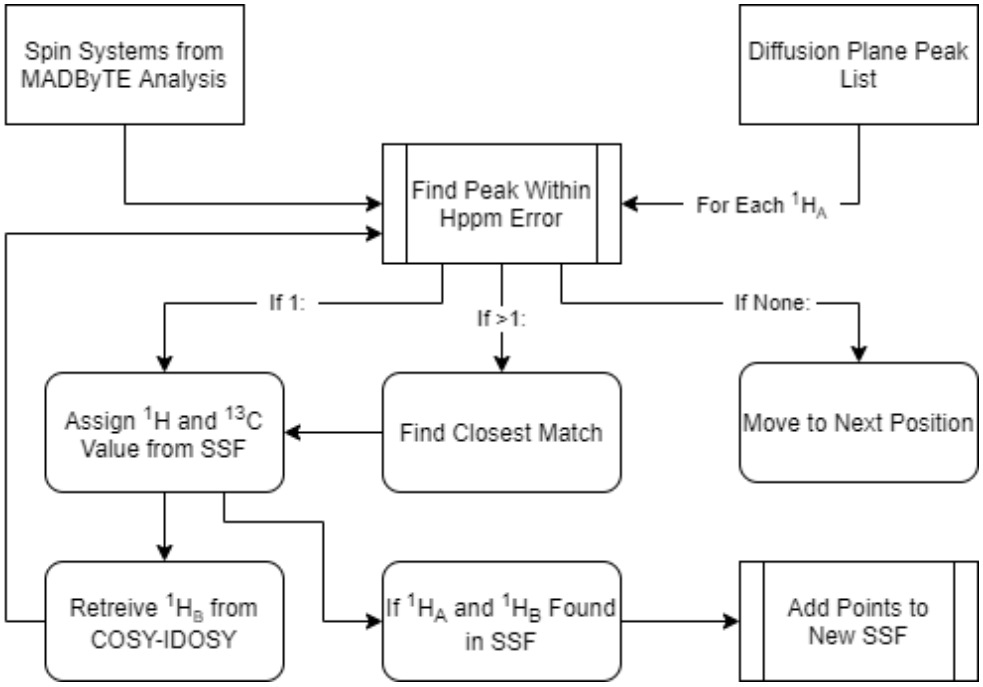


Figure 5.10. Overview of Spin System Refinement Using DOSY Planes. Each SSF from MADByTE analysis is compared to a DOSY plane peak list for matches. If a match is found, the reciprocal point is queried. If a match is made, the points are added to a new SSF annotated with the plane association.

Each DOSY plane can therefore split the original SSFs into smaller systems, directly allowing for feature fission. A new spin system master file is created indicating these new spin system features and their membership to allow for MADByTE processing and network generation.

5.5. Results and Discussion

5.5.1. Feature Fission

The 1:1:2 mixture of erythromycin, cycloheximide, and nystatin presented a valid case for feature fission through the association of DOSY data where a single spin system feature was constructed from several independent spin systems from a complex sample (Figure 5.11 – Panel A, starred node). This was due to the inherent complexity of the TOCSY data, in which proton resonances were too close for the system to resolve as independent. However, through the introduction of the diffusion data, this spin system was split into three new SSFs, each originating from a different plane of the DOSY plot (Figure 5.11 – Panel C, starred nodes). In the case of the original spin system network, cycloheximide (**12**) shared only associations to this feature. However, after feature fission, the new spin system feature associated with DOSY plane 5 (Figure 5.11 – Panel C, green bordered node) showed improved connectivity to cycloheximide and represented a more reasonable spin system size (9 members) when compared to the original feature (33 members).

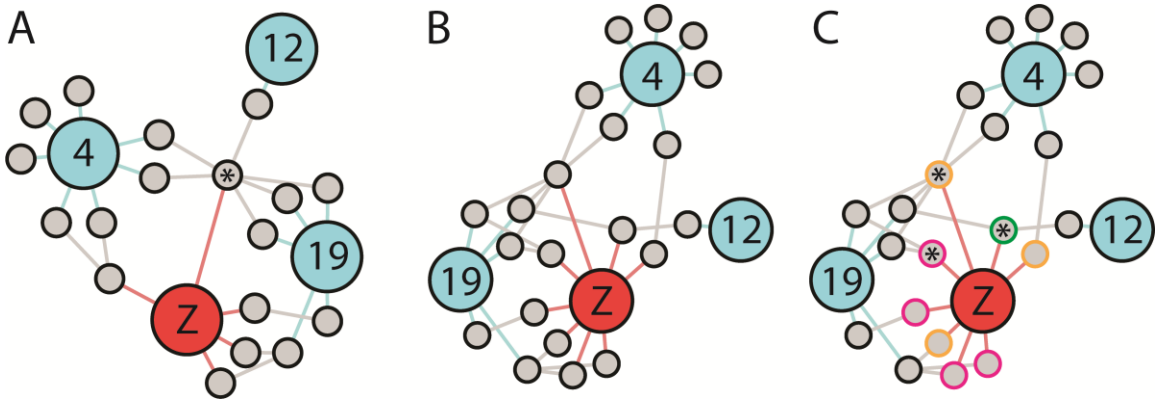


Figure 5.11. Feature Association by Diffusion Experiments – Feature Fission of a Mixture of 3 Components. A) The original MADByTE Network of the mixture produced a spin system through clustering of many signals together (starred node). B) Feature fission filters the SSFs into new diffusion associated SSFs. C) Color coding of SSFs by their associated DOSY plane (Plane. 5: green, Plane 6: yellow, Plane 7: pink). The complex node from panel A is split into three new nodes (starred) each associated with a different diffusion rate.

5.5.2. Feature Fusion

Feature fusion of SSFs from a sample allows for the clustering of SSFs from each DOSY plane, which allows for a more compact version of the network to be constructed. In the case of this mixture, Erythromycin (**4**) clustered to only nodes which were associated with an average diffusion rate of 1.87E-10 m^2/s, and cycloheximide (**12**) clustered only to the fusion node associated with an average of 2.60E-10 m^2/s. This suggests that the resonances within these features belong to these components and that connections from nystatin (**19**) to these nodes may be due to chance overlap of resonances.

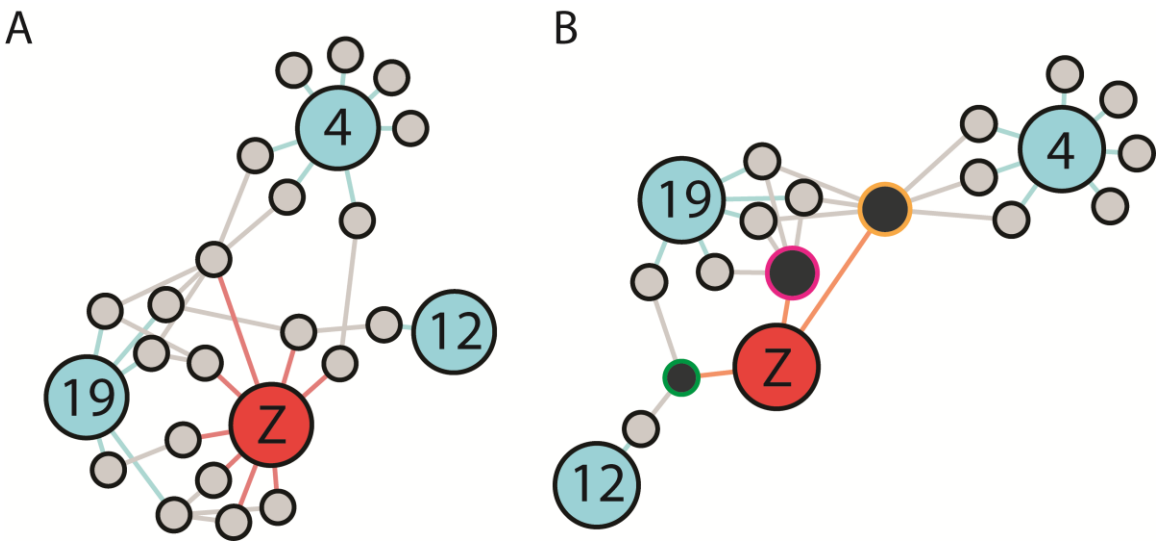


Figure 5.12. Feature Association by Diffusion Experiments - Feature Fusion of a Mixture of 3 Components. A) MADByTE Network from feature fission separating spin system features by diffusion rate. B) Combination of spin systems with the same diffusion plane (Plane 5: green border, Plane 6: yellow border, Plane 7: pink border)

In general diffusion NMR spectroscopy, molecular size greatly influences the ability of a molecule to diffuse such that the larger the molecule, the slower the diffusion rate.[39] The calculated diffusion rates of each plane suggest that the smallest molecule, cycloheximide has the fastest diffusion rate, and nystatin would conversely diffuse slowest.

5.5.3. Feature Association by Diffusion Experiments

A practical application of feature fusion can be seen from Equation 5.1, which states that the diffusion constant of a molecule should remain stable in independent experiments if all other experimental values remain constant. Therefore, fusion features which exhibit similar diffusion constants and share structural SSF connections could greatly aid in the annotation of unknown molecules when applied to extracts. However, this would require an improvement in DOSY methods to derive diffusion rates for a mixture of very different analytes, which may interact and complex in unpredictable ways in complex samples. This remains a challenge, as most applications of DOSY require substantial optimization and manual processing to achieve accurate diffusion rates. The introduction of an internal standard holds promise for this comparison strategy, providing a method for compensation of matrix effects which may affect diffusion rates in mixtures.[103]

As MADByTE processing already requires two 2D experiments (HSQC and TOCSY), any addition of experiments in the sample analysis pipeline would need to afford a considerable advantage and carefully balance the need for additional experiment time. 2D DOSY offers a considerable time advantage when compared to the COSY-IDOSY, but did not afford enough information to be of practical use. Other 3D DOSY experiments could be incorporated into a FADES approach, such as the HSQC-IDOSY[104], but require considerable amounts of time when compared to COSY-IDOSY; a 16 point gradient of the HSQC-IDOSY (1280 x 128 x 16, ns = 16) would require 15 hrs of experiment time compared to the COSY-IDOSY (2048 x 512 x 16, ns = 2) experiment time of 5 hrs. Therefore, application of the HSQC-IDOSY experiment to complex mixtures may not be practical on scale. Another DOSY method, the Hadamard encoded 3D DOSY-TOCSY, shows promise for future integration into MADByTE analysis as it reports an increase in both resolution and a 10-fold reduction of experiment time when compared to standard DOSY-TOCSY methods.[105] If optimized, this method could replace the standard TOCSY acquired for MADByTE, allowing for a standard processing pipeline to be developed which natively incorporated diffusion separation into MADByTE.

5.6. Conclusions

The implementation of 2D DOSY experiments was shown to be insufficient to achieve separation in the diffusion dimension and therefore was not promising for development.

However, through the incorporation of 3D DOSY experiments, it was shown that two limitations of MADByTE analysis could be independently addressed through feature fusion and feature fission. Feature fission allows for the refinement of MADByTE features suffering from spectral overlap in standard 2D experiments and was able to separate the spin system originating from cycloheximide from a complex of overlapped resonances. Similarly, feature fusion allows for the association of otherwise independent spin systems which arise from several molecules in a mixture to be associated based on diffusion rate, increasing the amount of structural coverage which can be attributed to a single molecule in a mixture. Application of this strategy to complex extracts could yield great improvements to MADByTE network analysis, especially when combined with the introduction of an internal standard for determination of comparable diffusion constants.

5.7. Future Perspectives of MADByTE

These pilot studies demonstrate an additional utility of the MADByTE system to associate complex NMR data into contextualized features which attempt to describe the chemical constitution of a mixture. In the perspective of the entire MADByTE platform, these additional data show that the advancements of NMR in the last few decades have provided advanced strategies for mixture analysis and metabolomics. Improvements in pulse sequences,[106,107] non-uniform sampling,[108] automation of NMR processing using machine learning,[109] and the introduction of ultra-high field NMR instrumentation[29] all have direct impacts on the future perspectives of MADByTE. As noted in Chapter 4, combinations of predictive utilities with these advancements may position NMR spectroscopy as the analytical platform of choice for functional annotation of natural product mixtures in the coming years.

As these methods improve, expandable utilities which allow for comparison of these data will become increasingly important. Currently, GNPS (Global Natural Product Social Molecular Networking) represents the state of the art in community developed tools which allow for the comparison of data to generate hypotheses of chemical constitution from spectral data.[24] The introduction of GNPS to the natural product community has revolutionized workflows and tool development in the field, demonstrating the value of these context creating utilities.[94,110,111] However, hypothesis generating utilities such as GNPS and MADByTE still require investigator intuition and manual analysis of the results and cannot be relied upon to provide a definitive answer in every use case. The roll of

these utilities is to aid the investigator - not replace them – and will require steady development and community feedback to be of widespread use.[72]

MADByTE is constructed from an NMR first perspective but does not exist in isolation. Advancements in MS and NMR annotation platforms are often pitted against one another in general discussion, but instead represent two powerful complimentary analytical utilities often used in combination. MS based platforms demonstrate increased sensitivity, dynamic range, and throughput when compared to 2D NMR studies but suffer from a loss of structural information which is accessible in NMR spectroscopy. Future development of utilities which leverage both platforms together offer considerable promise for discovery pipelines, especially when combined with robust databases to supplement in-silico predictions.[112,113]

5.8. Experimental

HSQC and TOCSY

All HSQC and TOCSY NMR spectra were recorded on a Bruker Avance™ III QCI (600 MHz) spectrometers in DMSO-d_6 (CortecNet lot Q0611) at 300°K. HSQC spectra were recorded as 32 scans (TD: 4096 x 256), collected by non-uniform sampling at 50% followed by linear prediction and zero filling. TOCSY spectra were recorded as 16 scans (TD: 1024 x 128), collected by non-uniform sampling at 50% followed by linear prediction and zero filling. NUS point spreads were kept consistent between samples to ensure consistency. HSQC and TOCSY spectra were manually peak picked for MADByTE analysis.

Pseudo 2D DOSY

The sample prepared for pseudo 2D DOSY experiments was prepared as 1.0 mg/mL erythromycin, 1.0 mg/mL of cycloheximide, and 2.0 mg/mL of nystatin prepared in a Shigemi tube with a filling of 280 µL. DOSY spectra were collected using the ledbpgp2s pulse sequence with a pulse length of 1.7 ms and a diffusion delay of 0.1 s with 16 gradient increments from 2% - 98%. Spectra were baseline and phase corrected in Topspin 3.6.2 and processed in Dynamics Center 2.6.3. Peak picking was done using automated peak picking and values were compared using intensities at peak positions.

COSY-IDOSY

The sample prepared for COSY-IDOSY experiments was prepared as 1.0 mg erythromycin, 1.0 mg of cycloheximide, and 2.0 mg of nystatin prepared in a Shigemi tube with a filing of 400 µL. Nystatin, being nearly three times the molecular mass of cycloheximide was provided at double the concentration to avoid dynamic range issues. DOSY spectra were recorded using the dosyicosy3d pulse sequence with a pulse length of 4.0 ms and a diffusion delay of 0.1 s with 16 gradient increments from 2%-98% with 2 scans (1024x512) per increment. Spectra were Fourier transformed in F1 and F2 in Topspin 3.6.2 and DOSY processed in Dynamics Center 2.6.3. Peak picking was done manually using the first plane (2% gradient strength) and applied to all remaining planes. Peaks were selected if they were above the noise threshold, exhibited good peak shape in both dimensions, and were not suspected as being noise or artefacts from spectral processing. The DOSY view was constructed using a logarithmic display with a minimum threshold of $1.0e^{-12}$ and a maximum of $1.0e^{-8}$.

MADByTE Processing

Initial MADByTE networks were constructed using MADByTE v1.3.0 using the parameters in Table 5.3. Initial processing of the mixture sample was done without reference compounds, yielding the single sample network displayed in Figure 5.4 – panel A. Introduction of the standard compounds and reprocessing using the same parameters yielded the network in Figure 5.4 – Panel B.

Table 5.3. MADByTE Parameters for 1:1:2 Mixture of Erythromycin, Cycloheximide, and Nystatin

Parameter	Value
Hppm Error	0.05
Cppm Error	0.40
Consensus Error	0.03
Similarity Ratio	0.50
Merge Multiplets	True
Maximum Spin System Size	40

Creation of the feature fission network was done using a modified version of MADByTE v1.3.0 using manual construction of the spin system master file incorporating the output

of DOSY plane association. Network construction parameters were identical to parameters in Table 5.3, producing the network framework in Figure 5.11 – panel B. The feature fusion network was created through manual combination of diffusion associated nodes from the feature fission network, retaining connections from the component nodes from feature fission MADByTE processing. All networks were visualized using Gephi 0.9.2 using the Force Atlas 2 algorithm with default settings except spacing = 10, dissuade hubs = True, prevent overlap = True.

5.9. Supplemental Data

5.9.1. Spin System Assignment from Initial MADByTE Analysis

Table 5.4. Spin System Assignment from MADByTE Analysis of DOSY Sample.

Resonance	Spin System
[0.73, 10.90]	JE_DOSY_0
[0.86, 12.25]	JE_DOSY_0
[1.07, 21.35]	JE_DOSY_0
[1.10, 17.82]	JE_DOSY_0
[1.17, 72.87]	JE_DOSY_0
[1.18, 26.31]	JE_DOSY_0
[1.31, 40.03]	JE_DOSY_0
[1.38, 21.10]	JE_DOSY_0
[1.49, 42.30]	JE_DOSY_0
[1.63, 34.88]	JE_DOSY_0
[1.80, 39.85]	JE_DOSY_0
[1.80, 42.35]	JE_DOSY_0
[1.81, 39.85]	JE_DOSY_0
[1.81, 42.35]	JE_DOSY_0
[2.24, 26.86]	JE_DOSY_0
[2.28, 38.34]	JE_DOSY_0
[2.37, 42.55]	JE_DOSY_0
[2.56, 36.8]	JE_DOSY_0
[2.59, 36.8]	JE_DOSY_0
[2.62, 56.01]	JE_DOSY_0
[2.77, 44.51]	JE_DOSY_0
[3.02, 70.86]	JE_DOSY_0
[3.03, 70.86]	JE_DOSY_0
[3.06, 70.86]	JE_DOSY_0

Resonance	Spin System
[3.62, 67.09]	JE_DOSY_0
[3.66, 69.06]	JE_DOSY_0
[3.83, 64.91]	JE_DOSY_0
[4.01, 65.60]	JE_DOSY_0
[4.34, 75.24]	JE_DOSY_0
[4.36, 102.15]	JE_DOSY_0
[5.09, 75.73]	JE_DOSY_0
[5.11, 75.73]	JE_DOSY_0
[6.24, 132.37]	JE_DOSY_0
[0.96, 16.92]	JE_DOSY_1
[5.51, 135.50]	JE_DOSY_1
[5.51, 131.21]	JE_DOSY_1
[2.18, 31.71]	JE_DOSY_1
[5.97, 129.45]	JE_DOSY_1
[5.95, 131.08]	JE_DOSY_1
[5.70, 134.24]	JE_DOSY_2
[6.20, 132.81]	JE_DOSY_2
[1.04, 9.10]	JE_DOSY_3
[1.04, 17.90]	JE_DOSY_3
[1.04, 67.00]	JE_DOSY_3
[2.86, 39.43]	JE_DOSY_3
[1.04, 16.37]	JE_DOSY_3
[1.04, 11.16]	JE_DOSY_3
[1.04, 21.35]	JE_DOSY_3
[1.04, 17.97]	JE_DOSY_3
[4.04, 64.99]	JE_DOSY_3
[1.16, 64.76]	JE_DOSY_3
[1.04, 70.23]	JE_DOSY_3
[1.89, 57.47]	JE_DOSY_4
[3.95, 65.49]	JE_DOSY_4

5.9.2. COSY-IDOSY Analysis Data Table

Table 5.5. Calculated Diffusion Rates from COSY-IDOSY Analysis Per Peak Position

Peak number	F1 [ppm]	F2 [ppm]	D [m^2/s]	error
1	1.08	5.06	1.25E-10	1.06E-12
2	1.47	4.71	1.81E-10	7.56E-13
3	1.45	4.61	1.70E-10	2.79E-12
4	3.00	4.34	1.82E-10	7.42E-13
5	3.04	4.26	1.72E-10	1.52E-12
6	5.95	5.48	1.26E-10	1.14E-12
7	5.48	5.94	1.27E-10	1.32E-12
8	6.18	6.09	1.28E-10	1.62E-12
9	6.07	6.18	1.26E-10	1.39E-12
10	5.72	6.19	1.17E-10	2.21E-11
11	6.18	5.94	1.22E-10	4.39E-12
12	6.21	5.71	1.22E-10	4.31E-12
13	5.69	4.35	1.30E-10	4.72E-12
14	2.18	5.48	1.59E-10	5.17E-12
15	2.20	5.66	1.67E-10	5.65E-12
16	4.33	5.70	1.30E-10	3.36E-12
17	4.26	5.58	1.33E-10	5.50E-12
18	5.57	4.26	1.32E-10	7.64E-12
19	4.45	3.64	1.25E-10	5.43E-12
20	3.63	4.44	1.26E-10	5.50E-12
21	4.33	3.00	1.81E-10	1.00E-12
22	4.26	3.04	1.71E-10	2.01E-12
23	2.88	4.01	1.78E-10	8.33E-12
24	1.85	3.92	1.24E-10	8.18E-13
25	1.66	3.88	1.28E-10	1.56E-12
26	1.54	3.87	1.37E-10	3.27E-12
27	2.32	3.98	1.25E-10	1.82E-12
28	1.15	4.01	1.80E-10	6.34E-13
29	1.26	3.81	2.57E-10	9.43E-13
30	2.43	3.78	2.55E-10	8.35E-12
31	4.71	1.50	1.81E-10	6.87E-13
32	4.71	1.47	1.81E-10	7.07E-13
33	4.61	1.48	1.68E-10	3.26E-12
34	4.61	1.46	1.69E-10	1.98E-12

Peak number	F1 [ppm]	F2 [ppm]	D [m²/s]	error
35	3.79	1.30	2.57E-10	2.22E-12
36	3.79	1.27	2.58E-10	2.87E-12
37	3.58	1.05	1.82E-10	3.05E-12
38	4.00	1.14	1.75E-10	1.50E-12
39	5.04	1.07	1.25E-10	2.07E-12
40	3.07	1.11	1.49E-10	1.42E-11
41	1.05	3.59	1.83E-10	1.31E-12
42	1.05	3.56	1.82E-10	2.39E-12
43	1.10	3.09	1.26E-10	7.07E-13
44	1.03	2.85	1.81E-10	6.35E-13
45	1.01	2.82	1.81E-10	8.32E-13
46	1.03	2.88	1.81E-10	2.22E-12
47	1.08	2.74	1.78E-10	5.21E-13
48	1.94	2.41	2.63E-10	1.67E-12
49	1.92	2.43	2.43E-10	3.39E-12
50	1.78	2.57	2.56E-10	3.33E-12
51	1.78	2.61	2.65E-10	4.45E-12
52	1.54	2.41	1.86E-10	1.78E-12
53	2.41	1.94	2.54E-10	1.06E-12
54	1.29	2.22	2.62E-10	1.74E-12
55	0.94	2.21	1.23E-10	2.74E-12
56	0.98	1.87	1.81E-10	1.08E-12
57	0.82	2.57	2.60E-10	1.95E-12
58	0.84	2.59	2.64E-10	5.21E-12
59	0.84	2.59	2.64E-10	5.21E-12
60	0.84	2.63	2.60E-10	2.25E-12
61	1.03	2.61	1.76E-10	1.67E-12
62	0.82	2.55	2.63E-10	3.13E-12
63	1.92	1.63	2.55E-10	3.09E-12
64	1.92	1.59	2.55E-10	3.10E-12
65	1.61	1.94	2.57E-10	3.27E-12
66	1.78	1.94	2.52E-10	2.53E-12
67	1.78	1.95	2.56E-10	3.47E-12
68	1.61	1.97	2.57E-10	3.78E-12
69	1.94	1.77	2.51E-10	2.94E-12
70	1.94	1.80	2.55E-10	2.79E-12
71	1.94	1.76	2.51E-10	1.67E-12
72	2.22	1.29	2.60E-10	1.98E-12
73	2.41	1.56	1.93E-10	2.07E-12
74	2.06	1.16	2.49E-10	2.96E-12
75	2.83	1.02	1.83E-10	2.08E-12

Peak number	F1 [ppm]	F2 [ppm]	D [m²/s]	error
76	2.74	1.08	1.83E-10	4.80E-12
77	2.60	0.83	2.60E-10	2.39E-12
78	1.43	1.78	2.64E-10	2.83E-12
79	1.61	1.91	2.54E-10	2.37E-12
80	1.76	1.91	2.50E-10	2.86E-12
81	3.75	2.85	1.71E-10	3.85E-12
82	3.98	2.74	1.76E-10	2.85E-12
83	3.98	2.72	1.78E-10	2.45E-12
84	3.98	2.76	1.72E-10	3.81E-12
85	3.98	2.26	1.58E-10	1.47E-11
86	3.65	2.59	1.57E-10	1.31E-11
87	3.91	1.87	1.24E-10	1.10E-12
88	3.86	1.66	1.29E-10	1.78E-12
89	3.44	1.87	1.79E-10	3.20E-12
90	3.72	2.00	1.75E-10	4.36E-12
91	5.08	1.77	1.71E-10	6.86E-12
92	5.08	1.35	1.74E-10	3.59E-12
93	4.97	3.14	7.91E-09	1.87E-09
94	1.87	3.46	1.95E-10	5.64E-12
95	1.69	2.83	1.85E-10	5.51E-12
96	1.43	2.58	2.67E-10	7.47E-12
97	1.45	2.61	2.62E-10	1.14E-11
98	1.47	2.56	2.62E-10	1.36E-11
99	6.18	6.20	1.26E-10	1.02E-12
100	5.93	5.94	1.21E-10	1.12E-12
101	5.67	5.67	1.21E-10	1.08E-12
102	5.48	5.49	1.26E-10	1.03E-12
103	5.43	5.44	1.60E-10	2.20E-12
104	5.06	5.07	1.75E-10	2.79E-12
105	4.75	4.76	1.70E-10	4.15E-12
106	4.71	4.72	1.81E-10	9.56E-13
107	4.71	4.70	1.78E-10	1.05E-12
108	4.61	4.62	1.67E-10	1.78E-12
109	4.43	4.43	1.25E-10	5.18E-13
110	4.45	4.71	1.68E-10	1.01E-11
111	4.73	4.43	1.24E-10	4.80E-12
112	4.08	4.08	1.69E-10	2.92E-12
113	4.33	4.34	1.58E-10	2.28E-12
114	4.31	4.32	1.62E-10	2.15E-12
115	3.98	3.98	1.75E-10	1.19E-12
116	3.82	3.81	2.26E-10	4.18E-12

Peak number	F1 [ppm]	F2 [ppm]	D [m²/s]	error
117	3.75	3.74	1.78E-10	1.00E-12
118	3.86	3.87	1.51E-10	3.07E-12
119	3.63	3.64	1.29E-10	9.90E-13
120	3.44	3.44	1.73E-10	1.26E-12
121	3.37	3.38	1.72E-10	3.08E-12
122	2.86	2.86	1.83E-10	2.07E-12
123	3.00	3.00	1.62E-10	4.92E-12
124	3.14	3.14	6.75E-10	1.01E-10
125	3.18	3.18	1.75E-10	5.53E-12
126	2.55	2.56	2.68E-10	3.77E-12
127	2.53	2.54	2.83E-10	7.11E-12
128	2.48	2.48	6.59E-10	3.38E-11
129	2.25	2.25	2.28E-10	3.08E-12
130	2.18	2.19	1.79E-10	1.45E-12
131	2.04	2.05	7.24E-10	1.06E-10
132	1.92	1.93	2.51E-10	3.59E-12
133	1.87	1.87	1.76E-10	1.02E-11
134	1.78	1.78	2.40E-10	3.70E-12
135	1.59	1.61	2.49E-10	2.16E-12
136	1.47	1.47	2.16E-10	3.53E-12
137	1.31	1.32	1.77E-10	1.42E-12
138	1.26	1.26	1.89E-10	1.73E-12
139	1.17	1.17	2.40E-10	3.22E-12
140	1.10	1.10	1.73E-10	9.42E-13
141	0.98	0.99	1.79E-10	8.88E-13
142	0.84	0.84	2.22E-10	5.22E-12
143	0.82	0.82	2.22E-10	5.00E-12
144	1.81	1.46	3.99E-10	4.82E-11
145	4.02	2.89	1.79E-10	4.99E-12
146	2.76	3.98	1.94E-10	1.04E-11

References

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2020**, *83* (3), 770–803.

(2) Roberts, L. D.; Souza, A. L.; Gerszten, R. E.; Clish, C. B. *Curr. Protoc. Mol. Biol.* **2012**, *98* (1), 1–34.

(3) Editorial. *Nat. Chem. Biol.* **2007**, *3* (7), 351–351.

(4) Gershenzon, J.; Dudareva, N. *Nat. Chem. Biol.* **2007**, *3* (7), 408–414.

(5) Hider, R. C.; Kong, X. *Nat. Prod. Rep.* **2010**, *27* (5), 637–657.

(6) American Chemical Society International Historic Chemical Landmarks. The discovery and development of penicillin 1928-1949 http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html (accessed Jan 25, 2021).

(7) Liu, M.; Grkovic, T.; Liu, X.; Han, J.; Zhang, L.; Quinn, R. J. *Synth. Syst. Biotechnol.* **2017**, *2* (4), 276–286.

(8) Lax, E. *The Mould in Dr. Florey's Coat*; Abacus, 2005.

(9) Tan, S. Y.; Tatsumura, Y. *Singapore Med. J.* **2015**, *56* (7), 366–367.

(10) WHO. *World Health Organization Model List of Essential Medicines*, 21st ed.; WHO: Geneva, 2019.

(11) Wall, M. E.; Wani, M. C. *J. Ethnopharmacol.* **1996**, *51* (1–3), 239–254.

(12) Oberlies, N. H.; Kroll, D. J. *J. Nat. Prod.* **2004**, *67* (2), 129–135.

(13) Giskeødegård, G. F.; Madssen, T. S.; Euceda, L. R.; Tessem, M.; Moestue, S. A.; Bathen, T. F. *NMR Biomed.* **2019**, *32* (10), e3927.

(14) Verhoeven, A.; Slagboom, E.; Wuhrer, M.; Giera, M.; Mayboroda, O. A. *Anal. Chim. Acta* **2017**, *976*, 52–62.

(15) Hasanpour, M.; Saberi, S.; Iranshahi, M. *Planta Med.* **2020**, *86* (3), 212–219.

(16) Lee, J. E.; Lee, B. J.; Chung, J. O.; Kim, H. N.; Kim, E. H.; Jung, S.; Lee, H.; Lee, S. J.; Hong, Y. S. *Food Chem.* **2015**, *174* (C), 452–459.

(17) Hubert, J.; Nuzillard, J. M.; Renault, J. H. *Phytochem. Rev.* **2017**, *16* (1), 55–95.

(18) El-Elimat, T.; Figueroa, M.; Ehrmann, B. M.; Cech, N. B.; Pearce, C. J.; Oberlies, N. H. *J. Nat. Prod.* **2013**, *76* (9), 1709–1716.

(19) Zani, C. L.; Carroll, A. R. *J. Nat. Prod.* **2017**, *80* (6), 1758–1766.

(20) Robinette, S. L.; Zhang, F.; Brüschweiler-Li, L.; Brüschweiler, R. *Anal. Chem.* **2008**, *80* (10), 3606–3611.

(21) van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G. *ACS Cent. Sci.* **2019**, *5* (11), 1824–1833.

(22) Turi, C. E.; Finley, J.; Shipley, P. R.; Murch, S. J.; Brown, P. N. *J. Nat. Prod.* **2015**, *78* (4), 953–966.

(23) Kellogg, J. J.; Graf, T. N.; Paine, M. F.; McCune, J. S.; Kvalheim, O. M.; Oberlies, N. H.; Cech, N. B. *J. Nat. Prod.* **2017**, *80* (5), 1457–1466.

(24) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.;

Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A. P.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrov, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34* (8), 828–837.

(25) Kurita, K. L.; Glassey, E.; Linington, R. G. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (39), 11999–12004.

(26) Britton, E. R.; Kellogg, J. J.; Kvalheim, O. M.; Cech, N. B. *J. Nat. Prod.* **2018**, *81* (3), 484–493.

(27) Jessome, L. L.; Volmer, D. A. *LCGC North Am.* **2006**, *24* (5), 498–510.

(28) Annesley, T. M. *Clin. Chem.* **2003**, *49* (7), 1041–1044.

(29) Moser, E.; Laistler, E.; Schmitt, F.; Kontaxis, G. *Front. Phys.* **2017**, *5*, 1–15.

(30) Simmler, C.; Napolitano, J. G.; McAlpine, J. B.; Chen, S. N.; Pauli, G. F. *Curr. Opin. Biotechnol.* **2014**, *25* (1), 51–59.

(31) Turbitt, J. R.; Colson, K. L.; Killday, K. B.; Milstead, A.; Neto, C. C. *Phytochem. Anal.* **2020**, *31* (1), 68–80.

(32) Napolitano, J. G.; Lankin, D. C.; Graf, T. N.; Brent Friesen, J.; Chen, S. N.;

McAlpine, J. B.; Oberlies, N. H.; Pauli, G. F. *J. Org. Chem.* **2013**, *78* (7), 2827–2839.

(33) Qiu, F.; Imai, A.; McAlpine, J. B.; Lankin, D. C.; Burton, I.; Karakach, T.; Farnsworth, N. R.; Chen, S. N.; Pauli, G. F. *J. Nat. Prod.* **2012**, *75* (3), 432–443.

(34) Krishnamurthy, K. *Magn. Reson. Chem.* **2013**, *51* (12), 821–829.

(35) Cobas, C.; Seoane, F.; Domínguez, S.; Sykora, S. *Spectrosc. Eur.* **2011**, *23* (1), 26–30.

(36) Chylla, R. A.; Hu, K.; Ellinger, J. J.; Markley, J. L. *Anal. Chem.* **2011**, *83* (12), 4871–4880.

(37) Hughes, T. S.; Wilson, H. D.; De Vera, I. M. S.; Kojetin, D. J. *PLoS One* **2015**, *10* (8), 1–16.

(38) Mandal, P. K.; Majumdar, A. *Concepts Magn. Reson. Part A Bridg. Educ. Res.* **2004**, *20* (1), 1–23.

(39) Edward, J. T. *J. Chem. Educ.* **1970**, *47* (4), 261–270.

(40) Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H. *Sci. Rep.* **2017**, *7* (1), 14243.

(41) Kuhn, S.; Schlörer, N. E. *Magn. Reson. Chem.* **2015**, *53* (8), 582–589.

(42) Kuhn, S.; Johnson, S. R. *ACS Omega* **2019**, *4* (4), 7323–7329.

(43) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (22), 5601–5606.

(44) Dumolin, C.; Aerts, M.; Verheyde, B.; Schellaert, S.; Vandamme, T.; Van der Jeugt, F.; De Canck, E.; Cnockaert, M.; Wieme, A. D.; Cleenwerck, I.; Peiren, J.; Dawyndt, P.; Vandamme, P.; Carlier, A. *mSystems* **2019**, *4* (5), e00437-19.

(45) Matuszewski, B. K.; Constanzer, M. L.; Chavez-Eng, C. M. *Anal. Chem.* **1998**, *70* (5), 882–889.

(46) Bingol, K.; Brüschweiler, R. *J. Proteome Res.* **2015**, *14* (6), 2642–2648.

(47) Bingol, K.; Li, D. W.; Brüschweiler-Li, L.; Cabrera, O. A.; Megraw, T.; Zhang, F.; Brüschweiler, R. *ACS Chem. Biol.* **2015**, *10* (2), 452–459.

(48) Walker, L. R.; Hoyt, D. W.; Walker, S. M.; Ward, J. K.; Nicora, C. D.; Bingol, K. *Magn. Reson. Chem.* **2016**, *54* (12), 998–1003.

(49) Gu, H.; Pan, Z.; Xi, B.; Asiago, V.; Musselman, B.; Raftery, D. *Anal. Chim. Acta* **2011**, *686* (1–2), 57–63.

(50) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77* (5), 1282–1289.

(51) McAlpine, J. B.; Chen, S.-N.; Kutateladze, A.; MacMillan, J. B.; Appendino, G.; Barison, A.; Beniddir, M. A.; Biavatti, M. W.; Bluml, S.; Boufridi, A.; Butler, M. S.; Capon, R. J.; Choi, Y. H.; Coppage, D.; Crews, P.; Crimmins, M. T.; Csete, M.; Dewapriya, P.; Egan, J. M.; Garson, M. J.; Genta-Jouve, G.; Gerwick, W. H.; Gross, H.; Harper, M. K.; Hermanto, P.; Hook, J. M.; Hunter, L.; Jeannerat, D.; Ji, N.-Y.; Johnson, T. A.; Kingston, D. G. I. I.; Koshino, H.; Lee, H.-W.; Lewin, G.; Li, J.; Linington, R. G.; Liu, M.; McPhail, K. L.; Molinski, T. F.; Moore, B. S.; Nam, J.-W.; Neupane, R. P.; Niemitz, M.; Nuzillard, J.-M.; Oberlies, N. H.; Ocampos, F. M. M. M.; Pan, G.; Quinn, R. J.; Reddy, D. S. S.; Renault, J.-H.; Rivera-Chávez, J.; Robien, W.; Saunders, C. M.; Schmidt, T. J.; Seger, C.; Shen, B.; Steinbeck, C.; Stuppner, H.; Sturm, S.; Taglialatela-Scafati, O.; Tantillo, D. J.; Verpoorte, R.; Wang, B.-G.; Williams, C. M.; Williams, P. G.; Wist, J.; Yue, J.-M.; Zhang, C.; Xu, Z.; Simmler, C.; Lankin, D. C.; Bisson, J.; Pauli, G. F. *Nat. Prod. Rep.* **2019**, *36* (1), 35–107.

(52) Robinette, S. L.; Brüschweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2012**, *45* (2), 288–297.

(53) Xia, J.; Bjorndahl, T. C.; Tang, P.; Wishart, D. S. *BMC Bioinformatics* **2008**, *9*, 1–16.

(54) Reher, R.; Kim, H. W.; Zhang, C.; Mao, H. H.; Wang, M.; Nothias, L. F.;

Caraballo-Rodriguez, A. M.; Glukhov, E.; Teke, B.; Leao, T.; Alexander, K. L.; Duggan, B. M.; Van Everbroeck, E. L.; Dorrestein, P. C.; Cottrell, G. W.; Gerwick, W. H. *J. Am. Chem. Soc.* **2020**, *142* (9), 4114–4120.

(55) Buedenbender, L.; Habener, L. J.; Grkovic, T.; Kurtböke, D. I.; Duffy, S.; Avery, V. M.; Carroll, A. R. *J. Nat. Prod.* **2018**, *81* (4), 957–965.

(56) Egan, J. M.; van Santen, J. A.; Liu, D. Y.; Linington, R. G. *J. Nat. Prod.* **2021**, *84* (4), 1044–1055.

(57) Howarth, A.; Ermanis, K.; Goodman, J. M. *Chem. Sci.* **2020**, *11* (17), 4351–4359.

(58) Paguigan, N. D.; El-Elimat, T.; Kao, D.; Raja, H. A.; Pearce, C. J.; Oberlies, N. H. *J. Antibiot.* **2017**, *70* (5), 553–561.

(59) Jonsson, P.; Gullberg, J.; Nordström, A.; Kusano, M.; Kowalczyk, M.; Sjöström, M.; Moritz, T. *Anal. Chem.* **2004**, *76* (6), 1738–1745.

(60) Posma, J. M.; Garcia-Perez, I.; Heaton, J. C.; Burdisso, P.; Mathers, J. C.; Draper, J.; Lewis, M.; Lindon, J. C.; Frost, G.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2017**, *89* (6), 3300–3309.

(61) Boyer, R. D.; Johnson, R.; Krishnamurthy, K. *J. Magn. Reson.* **2003**, *165* (2), 253–259.

(62) Thrippleton, M. J.; Keeler, J. *Angew. Chemie - Int. Ed.* **2003**, *42* (33), 3938–3941.

(63) Cheng, X.; Hochlowski, J.; Tang, H.; Hepp, D.; Beckner, C.; Kantor, S.; Schmitt, R. *J. Biomol. Screen.* **2003**, *8* (3), 292–304.

(64) Waybright, T. J.; Britt, J. R.; McCloud, T. G. *J. Biomol. Screen.* **2009**, *14* (6), 708–715.

(65) Williamson, M. P.; Craven, C. J. *J. Biomol. NMR* **2009**, *43* (3), 131–143.

(66) Klukowski, P.; Augoff, M.; ZieRba, M.; Drwal, M.; Gonczarek, A.; Walczak, M. J. *Bioinformatics* **2018**, *34* (15), 2590–2597.

(67) Tikole, S.; Jaravine, V.; Rogov, V.; Dötsch, V.; Güntert, P. *BMC Bioinformatics*

2014, *15* (1), 1–7.

(68) Kaltschnee, L.; Kolmer, A.; Timári, I.; Schmidts, V.; Adams, R. W.; Nilsson, M.; Kövér, K. E.; Morris, G. A.; Thiele, C. M. *Chem. Commun.* **2014**, *50* (99), 15702–15705.

(69) Emwas, A. H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Nagana Gowda, G. A.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. *Metabolites* **2019**, *9* (123), 1–39.

(70) Pachler, K. G. R.; Wessels, P. L. *J. Mol. Struct.* **1969**, *3* (3), 207–218.

(71) Minch, M. J. *Concepts Magn. Reson.* **1994**, *6* (1), 41–56.

(72) Chang, H. Y.; Colby, S. M.; Du, X.; Gomez, J. D.; Helf, M. J.; Kechris, K.; Kirkpatrick, C. R.; Li, S.; Patti, G. J.; Renslow, R. S.; Subramaniam, S.; Verma, M.; Xia, J.; Young, J. D. *Anal. Chem.* **2021**, *93* (4), 1912–1923.

(73) Short, T.; Alzapiedi, L.; Brüschweiler, R.; Snyder, D. *J. Magn. Reson.* **2011**, *209* (1), 75–78.

(74) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring network structure, dynamics, and function using NetworkX https://networkx.github.io/.

(75) Bokeh Development Team. Bokeh: Python library for interactive visualization http://www.bokeh.pydata.org.

(76) Helmus, J. J.; Jaroniec, C. P. *J. Biomol. NMR* **2013**, *55* (4), 355–367.

(77) Wong, W. R.; Oliver, A. G.; Linington, R. G. *Chem. Biol.* **2012**, *19* (11), 1483–1495.

(78) Brüschweiler, R.; Zhang, F. *J. Chem. Phys.* **2004**, *120* (11), 5253–5260.

(79) Cobas, C. *Stan's Libr.* **2014**, *V* (February), 1–8.

(80) Robinette, S. L.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2013**, *85* (11), 5297–5303.

(81) Grienke, U.; Foster, P. A.; Zwirchmayr, J.; Tahir, A.; Rollinger, J. M.; Mikros, E. *Sci. Rep.* **2019**, *9* (1), 11113.

(82) Shindo, K.; Yamagishi, Y.; Okada, Y.; Kawai, H. *J. Antibiot.* **1994**, *47* (9), 1072–1074.

(83) Garcia, I.; Vior, N. M.; González-Sabín, J.; Braña, A. F.; Rohr, J.; Moris, F.; Méndez, C.; Salas, J. A. *Chem. Biol.* **2013**, *20* (8), 1022–1032.

(84) Melven P. Weinstien. *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically*, 11th ed.; Clinical and Laboratory Standards Institute: Wayne, PA, 2018.

(85) Mihaleva, V. V.; Te Beek, T. A. H.; Van Zimmeren, F.; Moco, S.; Laatikainen, R.; Niemitz, M.; Korhonen, S. P.; Van Driel, M. A.; Vervoort, J. *Anal. Chem.* **2013**, *85* (18), 8700–8707.

(86) Bagno, A. *Chem. - A Eur. J.* **2001**, *7* (8), 1652–1661.

(87) Rychnovsky, S. D. *Org. Lett.* **2006**, *8* (13), 2895–2898.

(88) Das, S.; Edison, A. S.; Merz, K. M. *Anal. Chem.* **2020**, *92* (15), 10412–10419.

(89) Bakiri, A.; Hubert, J.; Reynaud, R.; Lanthony, S.; Harakat, D.; Renault, J. H.; Nuzillard, J. M. *J. Nat. Prod.* **2017**, *80* (5), 1387–1396.

(90) Yesiltepe, Y.; Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; Renslow, R. S. *J. Cheminform.* **2018**, *10* (1), 1–16.

(91) Jonas, E.; Kuhn, S. *J. Cheminform.* **2019**, *11* (1), 1–7.

(92) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L. F.; Reher, R.; Kang, K. Bin; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. *ChemRxiv*. 2020.

(93) Barber, J.; Gyi, J. I.; Lian, L.; Morris, G. A.; Pye, D. A.; Sutherland, J. K. *J. Chem. Soc. Perkin Trans. 2* **1991**, No. 10, 1489–1494.

(94) Van Der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (48), 13738–13743.

(95) Morris, K. F.; Johnson, C. S. *J. Am. Chem. Soc.* **1992**, *114* (8), 3139–3141.

(96) Li, W.; Chung, H.; Daeffler, C.; Johnson, J. A.; Grubbs, R. H. *Macromolecules* **2012**, *45* (24), 9595–9603.

(97) Monakhova, Y. B.; Diehl, B. W. K.; Do, T. X.; Schulze, M.; Witzleben, S. *J. Pharm. Biomed. Anal.* **2018**, *149*, 128–132.

(98) Nilsson, M.; Morris, G. A. *Chem. Commun.* **2007**, No. 9, 933–935.

(99) Margueritte, L.; Markov, P.; Chiron, L.; Starck, J. P.; Vonthron-Sénécheau, C.; Bourjot, M.; Delsuc, M. A. *Magn. Reson. Chem.* **2018**, *56* (6), 469–479.

(100) Wu, D. H.; Chen, A. D.; Johnson, C. S. *J. Magn. Reson. Ser. A* **1995**, *115* (2), 260–264.

(101) Nilsson, M.; Gil, A. M.; Delgadillo, I.; Morris, G. A. *Chem. Commun.* **2005**, No. 13, 1737–1739.

(102) Johnson, C. S. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34* (3–4), 203–256.

(103) Macchioni, A.; Ciancaleoni, G.; Zuccaccia, C.; Zuccaccia, D. *Chem. Soc. Rev.* **2008**, *37* (3), 479–489.

(104) Mclachlan, A. S.; Richards, J. J.; Bilia, A. R.; Morris, G. A.; Wiley, J. **2009**, *2009* (July), 1081–1085.

(105) Viel, S.; Caldarelli, S. *Chem. Commun.* **2008**, No. 17, 2013–2015.

(106) Hansen, A. L.; Kupče, E.; Li, D.-W.; Bruschweiler-Li, L.; Wang, C.; Brüschweiler, R. *Anal. Chem.* **2021**, *93* (15), 6112–6119.

(107) Paudel, L.; Adams, R. W.; Király, P.; Aguilar, J. A.; Foroozandeh, M.; Cliff, M. J.; Nilsson, M.; Sándor, P.; Waltho, J. P.; Morris, G. A. *Angew. Chemie - Int. Ed.* **2013**, *52* (44), 11616–11619.

(108) Hoch, J. C.; Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S. *Acc. Chem. Res.* **2014**, *47* (2), 708–717.

(109) Klukowski, P.; Augoff, M.; ZieRba, M.; Drwal, M.; Gonczarek, A.; Walczak, M. J. *Bioinformatics* **2018**, *34* (15), 2590–2597.

(110) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L. F.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Commun.* **2018**, *9* (1), 1–12.

(111) Cao, L.; Gurevich, A.; Alexander, K. L.; Naman, C. B.; Leão, T.; Glukhov, E.; Luzzatto-Knaan, T.; Vargas, F.; Quinn, R.; Bouslimani, A.; Nothias, L. F.; Singh, N. K.; Sanders, J. G.; Benitez, R. A. S.; Thompson, L. R.; Hamid, M. N.; Morton, J. T.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Friedberg, I.; Knight, R.; Venkateswaran, K.; Gerwick, W. H.; Gerwick, L.; Dorrestein, P. C.; Pevzner, P. A.; Mohimani, H. *Cell Syst.* **2019**, *9* (6), 600-608.e4.

(112) Kuhn, S.; Colreavy-Donnelly, S.; de Andrade Silva Quaresma, L. E.; de Andrade Silva Quaresma, E.; Borges, R. M. *Metabolomics* **2020**, *16* (12), 1–5.

(113) Kuhn, S.; Colreavy-Donnelly, S.; Santana De Souza, J.; Borges, R. M. *Faraday Discuss.* **2019**, *218*, 339–353.