UNIVERSITY OF CALIFORNIA

SANTA CRUZ

COMPOUND ACTIVITY MAPPING: INTEGRATING CHEMICAL AND BIOLOGICAL PROFILING FOR THE FUNCTIONAL ANNOTATION OF NATURAL PRODUCT LIBRARIES

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMISTRY AND BIOCHEMISTRY

by

Kenji Long Kurita

September 2015

The Dissertation of Kenji L. Kurita
is approved:

_____
Professor Roger Linington, Advisor

_____
Professor Joseph Konopelski, Chair

_____
Professor R. Scott Lokey

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

Table of Figures

Chapter 1

Chapter 2

Chapter 4

Abstract

Compound Activity Mapping: Integrating Chemical and Biological Profiling for the Functional Annotation of Natural Product Libraries

By Kenji Kurita

Natural products research has had a significant impact on human-health and our understanding of the natural world as a pillar of pharmacognosy, organic chemistry, ecology, and chemical biology. But while this science has yielded countless discoveries such as penicillin, taxol, and artimesinin and will continue to improve quality of life around the world, the idea that natural products is a panacea of chemical diversity has been challenged by problems including the endless rediscovery of known compounds, the immense time required to isolate and elucidate structures, and the need for large amounts of scarce compounds to exceed the ever raising bar of biological annotation for drug approval. This thesis will provide examples of the use of integrated biological and chemical annotation of natural product libraries for the comprehensive functional annotation of natural product libraries, a new platform to expedite the dereplication and structural assignment of natural products libraries, and a study using genome annotation tools to look at the diversity of secondary metabolite biosynthetic gene clusters across a large set of cultured bacterial clades. Each chapter will discuss how using these modern techniques enabled the discovery of the quinocinnolinomycins (**3.1-3.4**), the elucidation of the aryl-polyenes (APE_{EC} and APE_{VF}), and the deeper understanding of the biological effects and constitution of natural products libraries through the dereplication of phencomycin. The net result of

all these technologies is that they change natural products research from an intensely focused effort to discover the most potent compounds for a particular disease, to a hypothesis and data driven exploration of the subtle interactions of secondary metabolites within biological settings.

Acknowledgements

The text of this dissertation includes a reprint of the following previously published material:

Kurita, K. L., & Linington, R. G. (2015). Connecting Phenotype and Chemotype: High-Content Discovery Strategies for Natural Products Research. *Journal of Natural Products*, *78*(3), 587–596.

The co-author listed in this publication directed and supervised the research which forms the basis for the dissertation

1. CONNECTING PHENOTYPE AND CHEMOTYPE: HIGH-CONTENT DISCOVERY STRATEGIES FOR NATURAL PRODUCTS RESEARCH

1.1. Introduction

Natural products have historically played a major role in the discovery and development of a diverse array of therapeutics including antibiotics, anticancer agents, antifungal drugs and analgesics. The modern era of natural products discovery has been driven in large part by continued innovation in both bioassay screening systems and analytical methods for the discovery of secondary metabolites with unique structures and biological properties. These efforts have led to an impressive diversity of new drugs, and the discovery of countless bioactive small molecules with value as chemical probes and sources of inspiration for medicinal chemistry campaigns. However, despite significant developments in these areas, natural products discovery is still challenged by a number of issues that have hampered the field for over 50 years.

In 1981, Drs. Matthew Suffness and John Douros from the U.S. National Cancer Institute published an opinion piece in *Trends in Pharmacological Sciences*,[1] in which they presented some of the problems and solutions associated with the then "new" field of anticancer drug discovery from natural sources, and discussed their outlook for the future. Reading their paper, it is remarkable how many of the challenges they identified remain substantial barriers to efficient discovery of bioactive natural products today. In this review of strategies for high-content

biological and chemical characterization in natural product discovery, we will begin by revisiting some of the issues raised by Drs. Suffness and Douros in 1981, and briefly discuss our interpretations of these issues for the field as we see them in 2015.

"…most active materials are undetectable, and those that are tend to be discovered repeatedly."

The issue of re-isolation was a problem then, and remains a significant challenge today. Despite dramatic advances in analytical hardware (high-field cryoprobe NMR and benchtop accurate mass LC-MS systems) it is a rare student that has not isolated a known compound at some time during their Ph.D. studies. Owing to of the large number of compounds now isolated from natural sources, rediscovery is becoming the norm rather than the exception in many instances. A number of metabolomics approaches have been developed to circumvent this issue, as will be discussed in more detail below, but new methods are still required to integrate these approaches with biological data in order to identify compounds with the highest value as novel bioactive lead compounds.

"…cytotoxicity tests are sensitive to any cell killing substance and give many false leads."

Traditional colorimetric live/dead assays say nothing about target, with the result that active extracts from these assays must be selected based on raw potency, rather than mechanistic behavior. Given that even some new compounds will likely hit targets for which there are already drugs on the market, it is important that modern natural products discovery programs take advantage of multi-parametric profiling tools for screening where possible, and use these methods for the targeted discovery of compounds with novel biological functions. A number of unbiased biological profiling platforms are discussed below, including examples of their use for the discovery or characterization of natural products with unique biological properties.

"The design and development of in vitro screens which are specific for detection of key mechanisms of drug action is a challenging task."

This issue has largely been resolved, thanks to the development of a vast array of target- and pathway-based high-throughput screening platforms. However, because many of these assay platforms are relatively complex or time-consuming to run, it is still true that mechanistic assays are hard to implement broadly. There is therefore still a need for the creation of new unbiased screening tools that characterize bioactive extracts in terms of broad mode of action (MOA) classifications, as a complement to the two extremes of live/dead cytotoxicity, and target-based screening methods.

"The isolation and purification of active compounds present in minute quantities in a crude extract is a time consuming and difficult task…"

Just as was true in 1981, natural products discovery remains difficult! Despite the advances in hardware mentioned above and the development of numerous derivatization, labeling, and analytical methods for compound identification, detailed and unequivocal determination of the constitution and configuration of complex natural products is a time-consuming task that typically requires a significant investment of resources and material. The development of integrated tools that consider both biological MOA predictions and chemical constitution of natural products extracts is beginning to provide solutions to this issue by ensuring that compounds selected for full structural characterization are of the highest priority in terms of both structural and/or biological novelty. The third section of this review will discuss this integrated approach, including both the advantages and current limitations of these strategies.

If the 20th century was the age of structure-driven natural products discovery, then the 21st century promises to be the age of function-driven natural products research. There remains a high degree of value in "old" natural products for which the biological attributes remain poorly characterized, but deriving accurate functional information for natural products libraries on a global scale remains a major challenge. This review will cover methods for untargeted chemical and biological characterization, and will present a perspective on future directions for the integration

of these analytical platforms for the de novo prediction of natural product structures and MOAs from complex screening libraries.

1.2. Chemical Characterization Strategies

1.2.1. Preamble.

Chemical characterization of natural products has progressed dramatically from early studies, which relied heavily on degradation, derivatization and the synthesis of structural subunits to solve chemical structures[2] to the modern scenario where even the largest and most complex structures can be determined using microscale analytical techniques.[3,4] Although many of these methods have seen incredible development since the creation of the earliest instruments[5-7] this review will focus on the broad characterization of natural product libraries, rather than the development of techniques to aid in the structure determination for individual compounds. For recent reviews of the development of MS technologies and the use of NMR-based metabolomics in natural products, see Carter,[8] Jarmusch and Cooks,[7] and Robinette et al.[9]

Thin-layer chromatography (TLC) emerged as the first method for parallelized characterization of natural product extracts, and is still widely used as a rapid, low-resolution method for profiling chemical constitution of natural product extracts; however, high-performance liquid chromatography (HPLC) and hyphenated techniques have all but completely replaced TLC for most natural products discovery applications, because of their increased resolution and greater information content

(Figure 1-1).[10-12] The use of HPLC retention time in combination with ultraviolet and visible absorbance spectra allows the profiling and comparison of extracts within any screening library and has been used widely by industry and academia. In an early example Miller et al. used stream splitting and automated fraction collection in a compound-by-compound bioactivity and dereplication process for the discovery of clavulanic acid,[13,14] paving the way for adoption of this approach by many other research groups. The primary disadvantages of these techniques are that HPLC protocols are time-consuming, analysis and dereplication are performed on a compound-by-compound basis, and saving fractions is not practical for large libraries.[10]

The rapid improvement in resolution and throughput introduced by ultra-performance liquid chromatography (UPLC), bench-top HRMS, and advances in NMR experimentation and technologies like 1.7 mm cryogenic NMR probes have recently changed the chemical characterization landscape of natural products libraries from a compound-by-compound dereplication process to a situation where analysis can reveal an unbiased global view of all metabolites in a given library, as will be described below.

1.2.2. Mass Spectrometric Profiling Methods.

Owing to its sensitivity and relatively high throughput, MS-based techniques have come to the forefront of rapid chemical characterization. Studies have demonstrated the coverage and accuracy of such techniques for representative fungal compound libraries (Figure 1-1).[15] The use of multivariate statistical methods such as

principal component analysis has also been used to discover unique compounds from MS-based untargeted analysis of libraries of *Myxococcus xanthus* strains and Ascidian-associated Actinomycetales.[16,17] Similarly, traditional metabolomics platforms including versions of XCMS have been used to discover novel compounds from organisms as well studied as *Streptomyces coelicolor.*[18] In this last study structural characterization was assisted by the use of tandem MS, which allowed structural information to be incorporated into MS-based dereplication and discovery. More recently, MS^2 fragmentation pattern matching has been used to develop Molecular Networking as a dereplication strategy for identifying known compounds and ascribing structural classes to unknown metabolites.[19,20] The use of MS fragmentation patterns for compound identification is a standard tool in traditional metabolomics analysis (e.g., electron impact fragmentation in most GC-MS systems). However, the use of relative mass differences in fragmentation spectra to connect compounds from a given structural family, coupled with network analysis to visualize the relatedness of analytes in a given sample set, provide new opportunities for the rapid characterization and visualization of the metabolic capacity of sets of samples regardless of source origin or the availability of pure compound standards for every analyte. Finally, Müller and co-workers have developed a new approach to the acquisition of MS^2 data for complex natural product samples, which generates a "scheduled precursor list" of features present in extracts of microbial cultures but not the corresponding medium blanks, and uses this list to direct subsequent MS^2 data acquisition.[21] Advances such as this improve the coverage of relevant molecules over

traditional MS^2 selection methods that rely on signal intensity for fragmentation selection, and are indicative of the new approaches to data acquisition being developed. These techniques are moving the field towards the comprehensive untargeted metabolomics profiling of complex natural products mixtures.

1.2.3. Nuclear Magnetic Resonance Profiling Methods.

While less common than MS-based techniques, developments in NMR experimentation and instrumentation have led to a significant rise in the use of NMR-based metabolomics for the profiling of crude extracts in recent years. The discovery of iotrochotrazine by 1H NMR comparisons of extracts enriched for compounds obeying Lipinski's "rule of five" exemplifies the utility of this strategy.[22] Similar to MS approaches, standardized acquisition and databases can be used to identify chemical constituents from crude mixtures.[23] The primary advantages of NMR-based chemical profiling over MS-based strategies are that (1) the analysis is quantitative, unlike MS-based approaches where poor ionization or ion suppression by other metabolites can preclude the observation of all constituents in an extract, and that (2) structural information is more readily derived from the data, particularly if 1H spectra are augmented with TOCSY or phase-sensitive HSQC experiments (Figure 1-1). The structural information inherent in two-dimensional (2D) experiments has been used extensively for the characterization of chemical components of insect and spider venom, fireflies, and ladybugs.[24-27] Integration of NMR spectroscopy with biological data has been used to identify pheromones in Caenorphabditis elegans through differential analysis by 2D NMR spectroscopy (DANS).[28] Similar to MS-based

metabolomics strategies, this study was able to identify specific signals corresponding to the ascarosides that have synergistic effects with other pheromones and were therefore unidentifiable by activity-guided fractionation. This elegant approach lays the foundation for integrating biological and chemical profiling for the discovery of molecules correlated with a specific phenotype in a given biological assay.

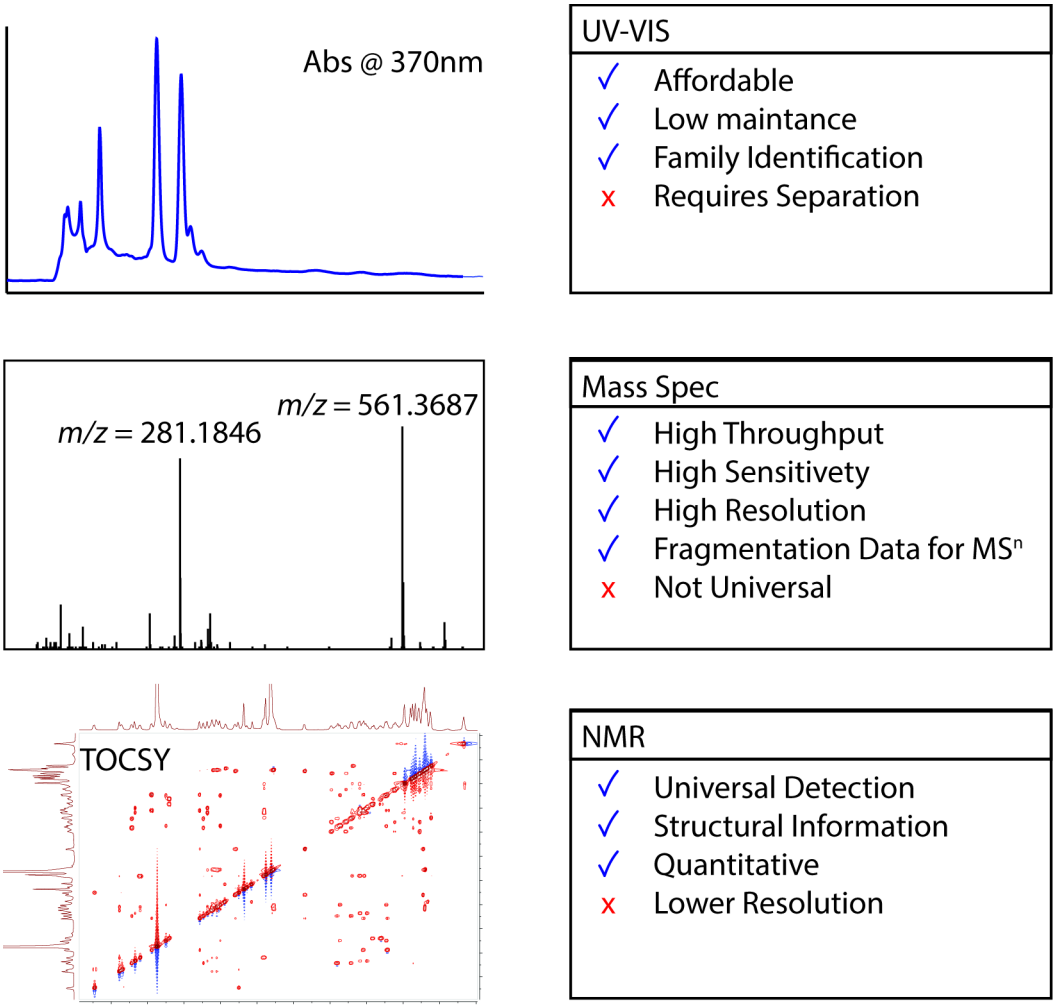


Figure 1-1: A summary of advantages and limitations of common chemical profiling strategies for natural products libraries.

1.3. Profiling Strategies for Biological Characterization.

1.3.1. Preamble.

Natural product screening has made significant progress since the early development of disk diffusion assays for microbial pathogens and colorimetric live/dead screens for mammalian cell lines. Recent developments in screening hardware and informatics now offer a wealth of readily accessible tools for the detailed biological characterization of compound libraries against almost any target system. These advances are providing opportunities for the early mode of action (MOA) prediction for bioactive compounds, which in turn is driving a "function-first" selection process for lead discovery and development (Figure 2).[29]

Although there are many examples of innovative screening systems for specific molecular targets and processes, we will restrict our focus in this review to unbiased assay systems that offer tools for the broad classification of bioactive compounds independent of specific MOAs, because of the inherent value that these tools offer the natural products chemist in terms of early global characterization of complex natural product libraries. Within this general area, the majority of development has been focused on four main target systems: mammalian cancer cell lines, yeast, bacteria and early vertebrate models. Each of these will be discussed in turn, highlighting recent advances and the advantages and limitations of each system for natural products research.

1.3.2. Mammalian Cell Screening.

Multiparametric screening in mammalian cells was first pioneered as a systematic strategy for the evaluation of compound mode of action by the development of the NCI 60-cell-line screen from 1985-1990.[30] This platform is the original "high-content" screening platform for natural products research, and has been used successfully to determine the MOAs of numerous natural products. For example, extracts containing salicylihalamides, potent vacuolar ATP-ase inhibitors, were first identified based on their particular NCI 60-cell-line profile.[31] This platform is still in regular use and is very information rich, but is logistically impractical for widespread library screening, given the quantities of material required to screen against the entire 60-cell-line panel, and the inherently low throughput of such a system.

In recent years, cytological profiling, broadly defined as multiparametric evaluation of cellular response to compound treatment, has gained increasing attention as a complement to target-based and colorimetric live/dead screening assays. Cytological profiling is most commonly performed on mammalian cell lines, and can incorporate a variety of analytical techniques, including microarrays, MS-based metabolomics, gene signatures, and high-content automated microscopy.[32-36] Several of these approaches have been employed for the investigation of natural product libraries, as outlined below.

1.3.2.1. Image-Based Screening.

Image-based screening was first widely adopted in industry because early systems were expensive, and required substantial informatics support to analyze the

resulting image files. More recently, the hardware cost has dropped and the analytical software has improved, making this a routine tool in academic screening centers. Image-based screening has been used to develop a number of unbiased whole cell phenotypic screening platforms.[32,37] In our own laboratory we have developed a modified version of the platform initially reported by Altschuler and co-workers[32] in order to create a tool suitable for the examination of complex natural product libraries.[38] This tool characterizes cell morphology using a set of structural and cell cycle fluorescent stains to extract hundreds of size and shape metrics for cells under drug pressure at sub-lethal concentrations. Subsequent informatics analysis compares these size and shape metrics to those for untreated control cells, and uses the differences in these values to create a numerical fingerprint that provides a graphical representation of the phenotypic differences between treated and control cells.

We have demonstrated that this tool can be used to classify the MOA of active constituents from complex natural product mixtures. Subsequent image-guided peak library fractionation can be used to pinpoint active compounds, and directly verify the cytological profiling signatures of these individual constituents, making the platform a powerful one for the discovery of natural products with unique phenotypic profiles (Figure 1-2).

A complementary approach that uses a combination of fluorescence and brightfield imaging for the characterization of cellular phenotypes was recently reported by Osada and co-workers.[39] This platform, termed MorphoBase, uses imaging data for two cell lines (HeLa and *src*^{ts}-NRK) to characterize the phenotypic

effect of compounds on cell development, and compares these phenotypic profiles to those of over 200 reference compounds of known mode of action to make direct predictions about the pathways or processes being disrupted by test compounds/extracts. MorphoBase has been used in conjunction with a proteomic profiling platform termed ChemProteoBase[40] for the de novo prediction of the mode of action of a new fungal metabolite, pyrrolizilactone.[41] In this work, both MorphoBase and ChemProteoBase identified strong clustering between pyrrolizilactone and test compounds known to inhibit proteasome function. Subsequent in vitro evaluation of 20S proteasome function confirmed this prediction, with the strongest inhibition of trypsin-like activity, providing an elegant demonstration of the use of unbiased profiling platforms for the direct prediction of bioactive natural products of unknown MOA.

Overall, image-based screens offer a large amount of biological annotation for natural product screening libraries in a format and timeframe that is appropriate for medium-throughput primary screens that number in the thousands of wells. We expect that the continuing improvements in screening hardware and software tools (e.g., the ability to perform high-throughput live cell imaging) will further lower the barrier to entry for these screening platforms, and that image-based profiling is likely to become a mainstay of future natural product discovery programs.

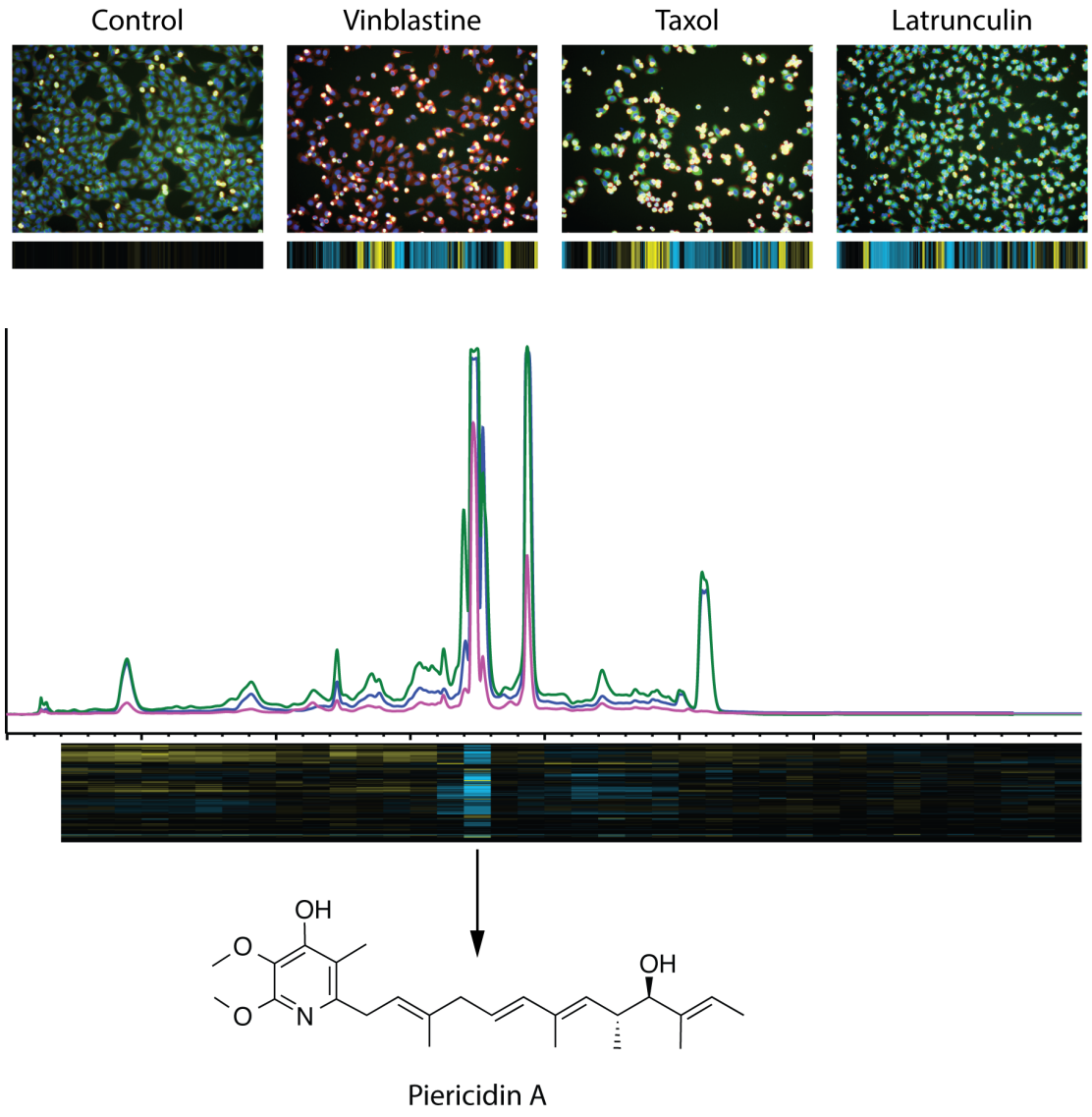


Figure 1-2: (Above) Images of control cells, test cells treated with purified natural products, and their corresponding cytological profiles. (Below) Example of the use of cytological profiling-driven peak library screening and bioactive compound discovery for piericidin A, an inhibitor of the mitochondrial electron transport chain.

1.3.2.2. Gene Expression Profiling Platforms

In addition to image-based approaches, a number of powerful gene profiling methods have been developed that are of relevance to the natural products community. The

"Connectivity Map", developed by researchers at the Broad Institute, was the first MOA profiling tool to compare the gene expression profiles of test compounds to a set of known bioactive molecules.[42] This platform is finding widespread use in the biomedical community beyond the prediction of compound MOAs, and has already been cited over 1000 times since its publication in 2006. In the natural products area, this system has been used to profile compounds from a range of sources, including a recent study that used Connectivity Map profiles to compare the bioactivity of intact Gila monster venom to the drug Byetta®, which is a synthetic derivative of a lead compound derived from this venom mixture currently in clinical use to treat diabetes.[43]

Another gene profiling method recently applied to the characterization of natural product modes of action is the Functional Signal Ontology (FUSION) system developed by researchers at the University of Texas Southwestern Medical Center.[44] This powerful platform uses the gene expression signatures of six key genes in HCT116 cells, as well as two genes with low variance as internal controls, to map the effect of treatment with either miRNAs, siRNAs, or natural product extracts. The team was able to demonstrate that these selected genes displayed non-colinearity of response under different treatment conditions, but that treatments of siRNAs or miRNAs from related pathways gave related FUSION signatures, and that FUSION signature matching can be used to accurately characterize the pathways targeted by specific bioactive natural products. More recently, this platform has been used to identify DDR2 as the molecular target of a new family of alkaloid natural products,

discoipyrroles A – D, demonstrating the power of this untargeted approach for molecular target determination for natural products.[45]

1.3.3. Yeast Profiling.

The baker's yeast, *Saccharomyces cerevisiae,* is a popular model system for studying mammalian cell biology thanks to the conservation of many of the genes implicated in human disease.[46] *Saccharomyces cerevisiae* has therefore become a powerful model organism for studying the mode of action of bioactive small molecules.[47] This has been aided by the creation of an ordered 5100-member gene deletion mutant library for all non-essential genes,[48-50] that permits the systematic evaluation of the effect of test compounds on gene deletion mutants for the prediction of compound MOAs. Coupled with the systematic evaluation of synthetic interactions between 5.4 million gene-gene pairs that has created a comprehensive gene interaction network map for *S. cerevisiae,*[50-52] this platform now represents a mature and powerful strategy for exploring chemical genetic properties of small molecules, including natural products.

Synthetic lethality screening uses the hypersensitivity of single gene deletion mutants to treatment with test compounds to indirectly report on compound molecular targets. If a single gene is non-essential, but treatment of that deletion strain with a bioactive small molecule causes lethality, then the small molecule must disrupt a compensatory pathway that is complementary to the function of the deleted gene product. By using the susceptibility of gene deletion mutants to test compounds in conjunction with the global genetic interaction network map, it is therefore in theory

possible to determine the specific target of any individual compound, provided that this target has a homologue in *S. cerevisiae*, and that the compound is active against this yeast protein.

There have been several recent examples of the use of this technology for the determination of natural product MOAs, including the discovery that the macrocyclic lipopeptide papuamide B targets phosphatidylserine in yeast,[53] and the determination that the marine sponge metabolite girolline targets Elongation Factor 2, and therefore exerts its anti-inflammatory activity through inhibition of protein synthesis at the elongation step.[54]

1.3.4. Antibiotic Profiling.

Antimicrobial assays were some of the earliest assays used in natural products discovery, including the original serendipitous discovery of penicillin, and are still in widespread use around the world for the early characterization of natural product extract libraries. Although simple assays such as disk diffusion, cross streak, and well-plate liquid culture growth assays against individual pathogens are rapid and cheap, the number of published natural products with antibiotic activities now means that rates of rediscovery using these methods are extremely high. To overcome this limitation, a number of unbiased antibiotic screening platforms have been developed that provide multi-parametric characterization of the effects of natural product extracts on bacterial cell development. These tools provide direct information about compound class and/or MOA for active constituents, and can be used to rapidly triage

large natural product libraries so that development effort is focused on those few extracts with highest potential for the discovery of new classes of antibiotics.

1.3.4.1. BioMAP Screening

The BioMAP screening platform, developed in our laboratory in 2012, uses a panel of Gram-positive and Gram-negative bacterial pathogens to create activity profiles across the panel, in analogous fashion to the NCI 60-cell-line screen described above.[55] By comparing these BioMAP profiles to profiles for a suite of commercially available antibiotics, it is possible to identify extracts that contain members of known classes of antibiotics, and to prioritize extracts with unique BioMAP signatures for further development. We have used this platform to discover new classes of antibiotics,[55] and to profile large numbers of pure compounds and extracts from collaborative partners from academia and industry. This technology is readily transferable to any research laboratory with access to basic microbiology facilities and a standard plate reader, and has successfully been implemented by other research groups, including institutions in developing nations such as Indonesia.

1.3.4.2. Bacterial Cytological Profiling.

Although BioMAP profiling is very efficient at identifying extracts with unique antibiotic profiles, it does not provide information about the molecular targets or MOAs of these active constituents. To address this issue, a number of research groups have turned to image-based screening to explore antibiotic MOA profiles. MOA determination using cell imaging is challenging for bacterial targets, because

bacterial cells are typically 100 times smaller than mammalian targets such as HeLa cells, making it technically difficult to acquire images of high enough resolution for cytological profiling in a high-throughput manner. In addition, most automated microscopy systems do not have pre-programmed modules to directly score images of bacterial cells, complicating the analytical component of this approach. Notwithstanding these challenges, two bacterial cytological profiling strategies have recently been reported.

The first, developed in our laboratory, uses high-throughput imaging of a chromosomally GFP-tagged strain of *V. cholerae* at 40 x magnification and a bespoke image analysis software platform to quantify cell size and shape features.[56] These size and shape features are used to provide a numerical description of the phenotypes of individual cells under varying concentrations of either test extracts or training set antibiotics of known MOA. The progression of phenotypes is then compared to those for the training set antibiotics and these phenotypic "trajectories" used to predict compound MOAs. In the initial study 58 antibiotics were profiled to generate the training set phenotypic trajectories. Comparing these trajectories to those of a set of natural product extracts identified four bioactive compounds with predicted MOAs. Of these, three (novobiocin, cosmomycin D, cycloprodigiosin) had predicted MOAs that concurred with previous literature, while the fourth (pentachloropseudilin) had its MOA predicted for the first time.

In a second study, cells were examined at higher magnification, using FM4-64 to stain cell membranes, DAPI to stain the nucleus, and SYTOX green to stain cells

with permeabilized membranes.[57] The platform was used to examine the effects on cell morphology of 41 antibiotics from 26 separate structural classes, and was able to demonstrate a strong clustering of compounds by phenotype that closely paralleled the known MOAs for these compounds. In addition, the authors examined a novel antibiotic natural product, spirohexenolide A, and proved that it rapidly collapses the proton motor force using a combination of bacterial cytological profiling and complementary secondary assays. This approach provides more detailed information about cell shape and the fate of specific cellular components, but at a lower throughput than the previous study. The development of motorized SCLM stages and automated 100 x water immersion objectives offer new opportunities for further method development in this area, though this has yet to be applied to natural product MOA determination.

1.3.5. Zebrafish Imaging.

In vivo imaging represents another substantial advance for natural product screening. Just as the early conotoxin screening in whole animals revealed a wealth of neurological activities for individual components of these complex mixtures,[58] so in vivo screening in zebrafish (*Danio rerio*) is providing a new strategy for the broad evaluation of natural products libraries. Advantages of this strategy include: a whole animal response; the ability to simultaneously measure both efficacy and off-target toxicity; the identification of developmental defects; the measurement of neurological and behavioral factors; and the ability to perform live animal time-resolved assays that look at temporal effects of compounds on animal health and survival.[59]

Although zebrafish have now been used for a wide array of targeted assays,[60-62] and as a tool for downstream target identification or validation,[63] there are still few examples of untargeted phenotypic screening in zebrafish, particularly for natural products.

One innovative system that has recently been developed incorporates both in vivo zebrafish screening and micro-scale fractionation for the simultaneous bioassay and physical characterization of plant extracts.[64,65] This system has been used to identify both angiogenesis inhibitors from African plant extracts,[66] and anticonvulsant compounds from Philippine medicinal plant *Solanum torvum.*[67]

Zebrafish screening has also been developed in industry, with Novartis reporting the results from profiling their in-house collection of 12,000 purified natural products.[68] This impressive study, likely the largest of its kind, identified 114 phenotypic hits from this primary screen, including 50 compounds that caused developmental arrest without necrosis. This set of compounds contained molecules known to disrupt the mitochondrial electron transport chain, leading the authors to hypothesize a similar mechanism for other compounds displaying this phenotype. Subsequent transcriptional profiling of these compounds revealed that many of them did indeed target specific complexes of the mitochondrial electron transport chain, but also revealed instances where these two profiling systems did not agree, highlighting the importance of careful secondary screening for MOA predictions derived from high-throughput multiparametric profiling primary screens.

Certainly, the development of new screening systems in live animal models offers the potential for the rapid and detailed profiling of complex libraries, with the capacity to examine broader physiological characteristics of extracts and lead compounds than is possible using simple cell-based or enzyme assays. It will be interesting to see how these tools continue to evolve in the coming years as assay platforms develop in terms of liquid handling and image/ phenotype analysis.

1.3.6. Multiparametric Screens and the Future of Natural Products Screening.

Overall, multi-parametric screening tools are offering new opportunities to the natural products community for the rapid and efficient classification of complex natural product libraries. These tools provide new methods for the early prioritization of extracts and compounds with unique biological properties, and are a valuable complement to traditional live/dead screening systems for the discovery of next-generation therapeutic lead compounds. With the widespread availability of screening centers in academic institutions, development and implementation of these screening tools is well within the reach of most natural products research groups. Given the obvious benefit that such screening methods offer for natural products discovery, we expect that these approaches will enjoy increasing prominence within the natural products community in the coming years.

1.4. Integrating Chemical Characterization and Biological Profiling Datasets

1.4.1. Preamble.

Bioinformatics tools are becoming essential in natural products research, as advances in experimental throughput and the complexity of data obtained from genomic, chemical, and biological profiling make manual interpretation difficult or impossible. As previously mentioned, many laboratories have now developed sophisticated platforms to discover and classify biosynthetic gene clusters, to connect biosynthetic gene clusters to their gene products, and to classify complex small molecule libraries based on their chemical signatures.[20,28,69-71] Recently, the integration of proteomics, metabolomics, and genomics has allowed genes, enzymes, and their small molecule products to be connected informatically for the discovery of bioactive compounds,[72] providing examples of how integrated multiparametric profiling can be used to solve complex analytical problems, such as the connection of genes to molecules. While these techniques are powerful and have significantly advanced our understanding of natural products genomics and biosynthesis, there are a number of difficulties that preclude the facile integration of multiparametric chemical and biological screening information for natural products discovery.

1.4.2. Challenges with Multiparametric Data Integration.

The requirement for the integration of chemical constitution and biological screening techniques favors MS based chemical profiling strategies because of their throughput, resolution, and sensitivity; however, most developed metabolomics

techniques require binary control and experimental groups looking for the correlation of genes with a defined outcome. Therefore, these analyses require the library to be manually curated. Instead, integrated profiling strategies require the use of untargeted metabolomics approaches that report on the presence of all constituents, whether or not the structures of all of these components are known. These tools can be developed with relative ease to create lists of individual components (defined by retention time and HRMS properties) and their distribution throughout the natural products library, but, connecting these components to specific structures is a much more challenging task which currently hampers the use of this approach for broad scale library characterization.

1.4.2.1. Concentrations, Timescales and the Analysis of Mixtures.

Since natural products libraries are extremely complicated mixtures, often with large variations in the concentration of different analytes, dynamic range is an issue for both screening and metabolomics platforms. This large variation in concentrations requires both the chemical profiling strategy and the biological screen to be sensitive, but to have the ability to characterize compounds at a range of concentrations. Typically, this is done by selecting a concentration for profiling that gives useful data for the majority of extracts, and performing a second profiling experiment at higher dilution factor for extracts that give either a strongly cidal readout in the profiling assay, or a saturated signal in the chemical analysis (typically a problem for accurate mass analyses such as ESI-qTOF). Furthermore, it is important that the analyses are configured such that the lower limits of detection are similar for

the two systems. This is important because without this bioactive compounds can be ignored, either because they were observed in the metabolomics system but not classified as active, or because the extract was classified as active, but the compound concentration was below the detection limit in the chemical analysis.

1.4.2.2. Technical Requirements for the Integration of High-Content Datasets.

Some of the major challenges in integrating high-content datasets involve how the data are processed and integrated. Generalizable strategies for either chemical or biological annotation such as those described above are useful; however, directly integrating data from these analytical platforms is often difficult or impossible using existing tools. For example, while multivariate statistical methods such as principal component analysis are effective for discovering unique compounds from MS-based metabolomics libraries, it is difficult to confidently assign biological information to the resulting components when these statistical methods are extended to include high-content screening.

It is our opinion that an integration strategy should aim to correlate every detectable chemical feature with undefined phenotypes or screening profiles. In this way, the data should draw hypotheses about the biological activity of each detectable compound in the library for a global view of the chemical and chemical-genetic potential in the library. This resource would be invaluable for dereplicating known compounds, identifying modes of action, finding new biological activities using orthogonal screens, and discovering new compounds. For example, when newly

developed biological screens are relatively low throughput, we can avoid re-screening samples containing frequent nuisance molecules like hydroxamic acid-containing metal chelators, pan-specific kinase inhibitors like staurosporine, or grossly cytotoxic anthracyclines by cherry picking the natural products library to avoid extracts previously annotated by multiparametric screening systems as containing these compound classes. The prediction of the broad MOAs of bioactive molecules can also be useful to avoid inclusion of extracts containing compounds with potential negative host interactions such as those associated with DNA damage, highlighting just a couple of situations where the target-independent characterization of biological and chemical properties of natural products libraries can be used to improve the discovery workflow for next-generation natural products-based therapeutics.

1.5. Introduction to CAM and Chemical and Biological Data Integration.

We have recently developed a new integrated profiling platform, termed Compound Activity Mapping (CAM), which profiles natural products libraries using a combination of image-based cytological profiling and untargeted UPLC-TOF metabolomics to directly identify and characterize all bioactive constituents of any natural products screening library against HeLa cells (Figure 4). This tool is capable of generating networks that cluster extracts and their bioactive constituents based on biological and chemical similarities, such that each cluster contains a list of related compounds predicted to cause a specific phenotypic effect on HeLa cell development, and the extracts that contain these bioactive constituents. Using this tool we are discovering a wealth of new bioactive constituents from our microbially-derived

natural products library, as well as providing phenotypic annotations for a large number of known compounds, some of which have not previously been characterized in terms of mammalian cell MOA.

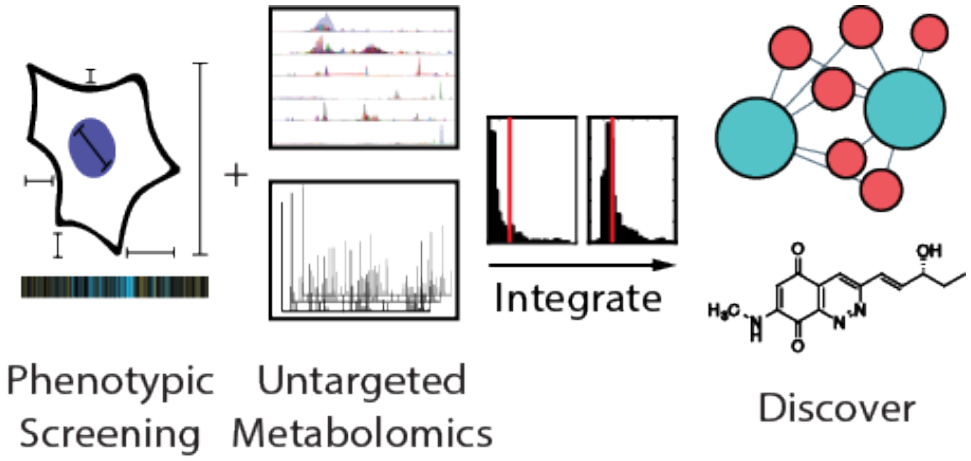


Figure 1-3: A cartoon representation of the integration of biological and chemical profiling demonstrating how the combined data may be used to find bioactive constituents from complex natural product extracts.

The rest of this thesis will discuss the development and application of CAM to the Linington Lab extract libraries as well as touch upon collaborative work in which bioinformatic analysis of bacterial genomic data was used for compound discovery. First, the hypothesis, preliminary results, and the proof of concept experiments will be described (Chapter 2). The true power of CAM will be illustrated by its application for the discovery of quinocinnolinomycin from an extract library derived from marine sediment samples collected as part of the Panama International Cooperative Biodiversity Group (Chapter 3). The discussion of CAM will conclude with expansion of CAM to include more of the Linington Lab extract library, its use in

collaboration with BioSortia, and the future third generation CAM platform which will incorporate MS^2 and MS^e data (Chapter 4). The final chapter will address genes to molecules approaches to natural product discovery and structure the elucidation of the aryl-polyenes, efforts by our lab to sequence 50 bacterial genomes, the specific sequencing and analysis of the abyssomycin gene cluster.

1.6. Future Perspective on "Big Data" in Natural Product Discovery

A rapid expansion in the resolution and throughput of academic screening data is currently taking place as high-throughput screening centers become more prevalent in universities and research institutes. Coupled with increasingly affordable and reliable MS tools and advances in the use of NMR methods for direct analysis of complex mixtures, we are poised to "open the box" on natural product discovery and transition from the traditional "grind and find" model, to a scenario in which we possess a priori knowledge about the constitution and MOA of all bioactive constituents of any screening library in advance of the isolation and detailed biological evaluation of individual compounds. Expansion of this approach to include whole genome sequence data for producing organisms is an obvious next step for improving the accuracy and coverage of molecular identification, and is close to becoming a reality as robust and affordable sequencing and genome assembly methods come of age. By extending this strategy from single profiling approaches to the integration of multiple profiling methods, each of which provides complementary but orthogonal information about the constitution and function of secondary metabolites from natural products libraries, we can now consider the possibility of

developing universal characterization methods that describe the precise constitutions and biological activities of all members of any complex natural product library. The implications of developing such tools are widespread, with many fields set to benefit. Areas of future application of these technologies include chemotaxonomy, chemical ecology and interspecies interactions, botanicals research, natural product drug discovery, and human microbiome research, to name a few. The era of "Big Data" is here for natural products; it is already changing the field, and we are only beginning to see the impact that multiparametric biological and chemical evaluation of will have on natural products discovery. It is an exciting time to be involved in natural products research, and we are fascinated to see what new discoveries this next generation of sophisticated tools will bring.

1.7. References:

(1) Suffness, M.; Douros, J. D. *Trends Pharmacol. Sci.* **1981**, *2*, 307.

(2) Woodward, R. B.; Brehm, W. J.; Nelson, A. L. *J. Am. Chem. Soc.* **1947**, *69*, 2250.

(3) Hamada, T.; Matsunaga, S.; Fujiwara, M.; Fujita, K.; Hirota, H.; Schmucki, R.; Güntert, P.; Fusetani, N. *J. Am. Chem. Soc.* **2010**, *132*, 12941.

(4) Molinski, T. F. *Nat. Prod. Rep.* **2010**, *27*, 321.

(5) Purcell, E. M. *Science* **1953**, *118*, 431.

(6) Bloch, F. *Science* **1953**, *118*, 425.

(7) Jarmusch, A. K.; Cooks, R. G. *Nat. Prod. Rep.* **2014**, *31*, 730.

(8) Carter, G. T. *Nat. Prod. Rep.* **2014**, *31*, 711.

(9) Robinette, S. L.; Brüschweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2011**, *45*, 288.

(10) Hook, D. J.; Pack, E. J.; Yacobucci, J. J.; Guss, J. *J. of Biomol. Screening* **1997**, *2*, 145.

(11) Wong, M. Y.; Steck, P. A.; Gray, G. R. *J. Biol. Chem.* **1979**, *254*, 5734.

(12) Miller, R. D.; Huckstep, L. L.; McDermott, J. P.; Queener, S. W.; Kukolja, S.; Spry, D. O.; Elzey, T. K.; Lawrence, S. M.; Neuss, N. *J. Antibiot.* **1981**, *34*, 984.

(13) Miller, R. D.; Neuss, N. *J. Antibiot.* **1978**, *31*, 1132.

(14) Hook, D. J.; More, C. F.; Yacobucci, J. J.; Dubay, G.; O'Connor, S. *J. Chromatogr. A* **1987**, *385*, 99.

(15) Nielsen, K. F.; Månsson, M.; Rank, C.; Frisvad, J. C.; Larsen, T. O. *J. Nat. Prod.* **2011**, *74*, 2338.

(16) Cortina, N. S.; Krug, D.; Plaza, A.; Revermann, O.; Müller, R. *Angew. Chem. Int. Ed.* **2011**, *51*, 811.

(17) Hou, Y.; Braun, D. R.; Michel, C. R.; Klassen, J. L.; Adnani, N.; Wyche, T. P.; Bugni, T. S. *Anal. Chem. 84*, 4277.

(18) Sidebottom, A. M.; Johnson, A. R.; Karty, J. A.; Trader, D. J.; Carlson, E. E. *ACS Chem. Biol.* **2013**, *8*, 2009.

(19) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1743.

(20) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686.

(21) Hoffmann, T.; Krug, D.; Hüttel, S.; Müller, R. *Anal. Chem.* **2014**, *86*, 10780.

(22) Grkovic, T.; Pouwer, R. H.; Vial, M. L.; Gambini, L.; Noël, A.; Hooper, J. N. A.; Wood, S. A.; Mellick, G. D.; Quinn, R. J. *Angew. Chem. Int. Ed.* **2014**, *53*, 6070.

(23) Bingol, K.; Bruschweiler Li, L.; Li, D.-W.; Brüschweiler, R. *Anal. Chem.* **2014**, *86*, 5494.

(24) Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Brüschweiler, R. *Anal. Chem.* **2007**, *79*, 7748.

(25) Taggi, A. E.; Meinwald, J.; Schroeder, F. C. *J. Am. Chem. Soc.* **2004**, *126*, 10364.

(26) Gronquist, M.; Meinwald, J.; Eisner, T.; Schroeder, F. C. *J. Am. Chem. Soc.* **2005**, *127*, 10810.

(27) Deyrup, S. T.; Eckman, L. E.; McCarthy, P. H.; Smedley, S. R.; Meinwald, J.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 9753.

(28) Pungaliya, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708.

(29) Schulze, C. J.; Linington, R. G. In *Natural Products: Discourse, Diversity, and Design*; Osbourn, A., Goss, R. J., Carter, G. T., Eds.; wiley.com: Oxford, 2014; pp 373–396.

(30) Alley, M. C.; Scudiero, D. A.; Monks, A.; Hursey, M. L.; Czerwinski, M. J.; Fine, D. L.; Abbott, B. J.; Mayo, J. G.; Shoemaker, R. H.; Boyd, M. R. *Cancer Res* **1988**, *48*, 589.

(31) Erickson, K. L.; Beutler, J. A.; Cardellina, J. H.; Boyd, M. R. *J. Org. Chem.* **1997**, *62*, 8188.

(32) Perlman, Z. E.; Slack, M. D.; Feng, Y.; Mitchison, T. J.; Wu, L. F.; Altschuler, S. J. *Science* **2004**, *306*, 1194.

(33) Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G.-W.; Tao, C.

Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. *Nat. Chem. Biol.* **2008**, *4*, 59.

(34) Feng, Y.; Mitchison, T. J.; Bender, A.; Young, D. W.; Tallarico, J. A. *Nat. Rev. Drug. Discov.* **2009**, *8*, 567.

(35) Mitchison, T. J. *ChemBioChem* **2005**, *6*, 33.

(36) Lorang, J.; King, R. W. *Genome Biol.* **2005**, *6*, 228.

(37) Tanaka, M.; Bateman, R.; Rauh, D.; Vaisberg, E.; Ramachandani, S.; Zhang, C.; Hansen, K. C.; Burlingame, A. L.; Trautman, J. K.; Shokat, K. M.; Adams, C. L. *PLoS Biol.* **2005**, *3*, e128.

(38) Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. *Chem. Biol.* **2013**, *20*, 285.

(39) Futamura, Y.; Kawatani, M.; Kazami, S.; Tanaka, K.; Muroi, M.; Shimizu, T.; Tomita, K.; Watanabe, N.; Osada, H. *Chem. Biol.* **2012**, *19*, 1620.

(40) Muroi, M.; Kazami, S.; Noda, K.; Kondo, H.; Takayama, H.; Kawatani, M.; Usui, T.; Osada, H. *Chem. Biol.* **2010**, *17*, 460.

(41) Futamura, Y.; Kawatani, M.; Muroi, M.; Aono, H.; Nogawa, T.; Osada, H. *ChemBioChem* **2013**, *14*, 2456.

(42) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. *Science* **2006**, *313*, 1929.

(43) Aramadhaka, L. R.; Prorock, A.; Dragulev, B.; Bao, Y.; Fox, J. W. *Toxicon*

2013, *69*, 160.

(44) Potts, M. B.; Kim, H. S.; Fisher, K. W.; Hu, Y.; Carrasco, Y. P.; Bulut, G. B.; Ou, Y.-H.; Herrera-Herrera, M. L.; Cubillos, F.; Mendiratta, S.; Xiao, G.; Hofree, M.; Ideker, T.; Xie, Y.; Huang, L. J.-S.; Lewis, R. E.; MacMillan, J. B.; White, M. A. *Sci. Signal.* **2013**, *6*, ra90.

(45) Hu, Y.; Potts, M. B.; Colosimo, D.; Herrera-Herrera, M. L.; Legako, A. G.; Yousufuddin, M.; White, M. A.; MacMillan, J. B. *J. Am. Chem. Soc.* **2013**, *135*, 13387.

(46) Karathia, H.; Vilaprinyo, E.; Sorribas, A.; Alves, R. *PLoS ONE* **2011**, *6*, e16015.

(47) Luesch, H.; Wu, T. Y. H.; Ren, P.; Gray, N. S.; Schultz, P. G.; Supek, F. *Chem. Biol.* **2005**, *12*, 55.

(48) Winzeler, E. A.; Shoemaker, D. D.; Astromoff, A.; Liang, H.; Anderson, K.; Andre, B.; Bangham, R.; Benito, R.; Boeke, J. D.; Bussey, H.; Chu, A. M.; Connelly, C.; Davis, K.; Dietrich, F.; Dow, S. W.; Bakkoury, El, M.; Foury, F.; Friend, S. H.; Gentalen, E.; Giaever, G.; Hegemann, J. H.; Jones, T.; Laub, M.; Liao, H.; Liebundguth, N.; Lockhart, D. J.; Lucau-Danila, A.; Lussier, M.; M'Rabet, N.; Menard, P.; Mittmann, M.; Pai, C.; Rebischung, C.; Revuelta, J. L.; Riles, L.; Roberts, C. J.; Ross-MacDonald, P.; Scherens, B.; Snyder, M.; Sookhai-Mahadeo, S.; Storms, R. K.; Véronneau, S.; Voet, M.; Volckaert, G.; Ward, T. R.; Wysocki, R.; Yen, G. S.; Yu, K.; Zimmermann, K.; Philippsen, P.; Johnston, M.; Davis, R. W. *Science* **1999**,

285, 901.

(49) Giaever, G.; Chu, A. M.; Ni, L.; Connelly, C.; Riles, L.; Véronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; André, B.; Arkin, A. P.; Astromoff, A.; Bakkoury, El, M.; Bangham, R.; Benito, R.; Brachat, S.; Campanaro, S.; Curtiss, M.; Davis, K.; Deutschbauer, A.; Entian, K.-D.; Flaherty, P.; Foury, F.; Garfinkel, D. J.; Gerstein, M.; Gotte, D.; Güldener, U.; Hegemann, J. H.; Hempel, S.; Herman, Z.; Jaramillo, D. F.; Kelly, D. E.; Kelly, S. L.; Kötter, P.; LaBonte, D.; Lamb, D. C.; Lan, N.; Liang, H.; Liao, H.; Liu, L.; Luo, C.; Lussier, M.; Mao, R.; Menard, P.; Ooi, S. L.; Revuelta, J. L.; Roberts, C. J.; Rose, M.; Ross-Macdonald, P.; Scherens, B.; Schimmack, G.; Shafer, B.; Shoemaker, D. D.; Sookhai-Mahadeo, S.; Storms, R. K.; Strathern, J. N.; Valle, G.; Voet, M.; Volckaert, G.; Wang, C.-Y.; Ward, T. R.; Wilhelmy, J.; Winzeler, E. A.; Yang, Y.; Yen, G.; Youngman, E.; Yu, K.; Bussey, H.; Boeke, J. D.; Snyder, M.; Philippsen, P.; Davis, R. W.; Johnston, M. *Nature* **2002**, *418*, 387.

(50) Ooi, S. L.; Pan, X.; Peyser, B. D.; Ye, P.; Meluh, P. B.; Yuan, D. S.; Irizarry, R. A.; Bader, J. S.; Spencer, F. A.; Boeke, J. D. *Trends Genet.* **2006**, *22*, 56.

(51) Tong, A. H.; Evangelista, M.; Parsons, A. B.; Xu, H.; Bader, G. D.; Pagé, N.; Robinson, M.; Raghibizadeh, S.; Hogue, C. W.; Bussey, H.; Andrews, B.; Tyers, M.; Boone, C. *Science* **2001**, *294*, 2364.

(52) Costanzo, M.; Baryshnikova, A.; Bellay, J.; Kim, Y.; Spear, E. D.; Sevier, C. S.; Ding, H.; Koh, J. L. Y.; Toufighi, K.; Mostafavi, S.; Prinz, J.; St Onge, R.

P.; VanderSluis, B.; Makhnevych, T.; Vizeacoumar, F. J.; Alizadeh, S.; Bahr, S.; Brost, R. L.; Chen, Y.; Cokol, M.; Deshpande, R.; Li, Z.; Lin, Z.-Y.; Liang, W.; Marback, M.; Paw, J.; San Luis, B.-J.; Shuteriqi, E.; Tong, A. H. Y.; van Dyk, N.; Wallace, I. M.; Whitney, J. A.; Weirauch, M. T.; Zhong, G.; Zhu, H.; Houry, W. A.; Brudno, M.; Ragibizadeh, S.; Papp, B.; Pál, C.; Roth, F. P.; Giaever, G.; Nislow, C.; Troyanskaya, O. G.; Bussey, H.; Bader, G. D.; Gingras, A.-C.; Morris, Q. D.; Kim, P. M.; Kaiser, C. A.; Myers, C. L.; Andrews, B. J.; Boone, C. *Science* **2010**, *327*, 425.

(53) Parsons, A. B.; Lopez, A.; Givoni, I. E.; Williams, D. E.; Gray, C. A.; Porter, J.; Chua, G.; Sopko, R.; Brost, R. L.; Ho, C. H.; Wang, J.; Ketela, T.; Brenner, C.; Brill, J. A.; Fernandez, G. E.; Lorenz, T. C.; Payne, G. S.; Ishihara, S.; Ohya, Y.; Andrews, B.; Hughes, T. R.; Frey, B. J.; Graham, T. R.; Andersen, R. J.; Boone, C. *Cell* **2006**, *126*, 611.

(54) Fung, S.-Y.; Sofiyev, V.; Schneiderman, J.; Hirschfeld, A. F.; Victor, R. E.; Woods, K.; Piotrowski, J. S.; Deshpande, R.; Li, S. C.; de Voogd, N. J.; Myers, C. L.; Boone, C.; Andersen, R. J.; Turvey, S. E. *ACS Chem. Biol.* **2014**, *9*, 247.

(55) Wong, W. R.; Oliver, A. G.; Linington, R. G. *Chem. Biol.* **2012**, *19*, 1483.

(56) Peach, K. C.; Bray, W. M.; Winslow, D.; Linington, P. F.; Linington, R. G. *Mol. BioSyst.* **2013**, *9*, 1837.

(57) Nonejuie, P.; Burkart, M.; Pogliano, K.; Pogliano, J. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 16169.

(58) Olivera, B. M.; Rivier, J.; Clark, C.; Ramilo, C. A.; Corpuz, G. P.; Abogadie, F. C.; Mena, E. E.; SR, W.; Hillyard, D. R.; Cruz, L. J. *Science* **1990**, *249*, 257.

(59) Zon, L. I.; Peterson, R. T. *Nat. Rev. Drug. Discov.* **2005**, *4*, 35.

(60) Basu, S.; Sachidanandan, C. *Chem. Rev.* **2013**, *113*, 7952.

(61) Rihel, J.; Prober, D. A.; Arvanites, A.; Lam, K.; Zimmerman, S.; Jang, S.; Haggarty, S. J.; Kokel, D.; Rubin, L. L.; Peterson, R. T.; Schier, A. F. *Science* **2010**, *327*, 348.

(62) Raldúa, D.; Piña, B. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 685.

(63) Crawford, A.; Esguerra, C.; de Witte, P. *Planta Med* **2008**, *74*, 624.

(64) Challal, S.; Bohni, N.; Buenafe, O. E.; Esguerra, C. V.; de Witte, P. A. M.; Wolfender, J.-L.; Crawford, A. D. *Chimia* **2014**, 1.

(65) Bohni, N.; Cordero-Maldonado, M. L.; Maes, J.; Siverio-Mota, D.; Marcourt, L.; Munck, S.; Kamuhabwa, A. R.; Moshi, M. J.; Esguerra, C. V.; de Witte, P. A. M.; Crawford, A. D.; Wolfender, J.-L. *PLoS ONE* **2013**, *8*, e64006.

(66) Crawford, A. D.; Liekens, S.; Kamuhabwa, A. R.; Maes, J.; Munck, S.; Busson, R.; Rozenski, J.; Esguerra, C. V.; de Witte, P. A. M. *PLoS ONE* **2011**, *6*, e14694.

(67) Challal, S.; Buenafe, O. E. M.; Queiroz, E. F.; Maljevic, S.; Marcourt, L.; Bock, M.; Kloeti, W.; Dayrit, F. M.; Harvey, A. L.; Lerche, H.; Esguerra, C. V.; de Witte, P. A. M.; Wolfender, J.-L.; Crawford, A. D. *ACS Chem. Neurosci.* **2014**, *5*, 993.

(68) Lai, K.; Selinger, D. W.; Solomon, J. M.; Wu, H.; Schmitt, E.; Serluca, F. C.; Curtis, D.; Benson, J. D. *ACS Chem. Biol.* **2013**, *8*, 257.

(69) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; BREITLING, R. *Nucleic Acids Res.* **2011**, *39*, W339.

(70) Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Brown, L. C. W.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 412.

(71) Bumpus, S. B.; Evans, B. S.; Thomas, P. M.; Ntai, I.; Kelleher, N. L. *Nat. Biotechnol.* **2009**, *27*, 951.

(72) Gubbens, J.; Zhu, H.; Girard, G.; Song, L.; Florea, B. I.; Aston, P.; Ichinose, K.; Filippov, D. V.; Choi, Y. H.; Overkleeft, H. S.; Challis, G. L.; van Wezel, G. P. *Chem. Biol.* **2014**, 1.

2. INTEGRATING SECONDARY METABOLOMICS AND PHENOTYPIC SCREENING

2.1. Introduction:

Natural products are a valuable source of compounds for drug discovery because of their structural complexity, chemical diversity, and selectivity for specific biological targets. For these reasons natural products and their derivatives represent almost three quarters of FDA approved cancer treatments; however, traditional natural products drug discovery suffers from high rates of rediscovery and challenges with identification of molecular targets.[1,2] In addition, the identification of individual bioactive constituents from complex mixtures often requires multiple rounds of purification, drastically slowing the discovery process and making natural product libraries poorly compatible with modern high-throughput screening programs. Despite the difficulties of natural products drug discovery, natural product libraries continue to be integrated with modern screening techniques because of natural products' inherent wealth of chemical diversity. If natural product libraries are to remain an integral part of modern drug discovery there is therefore a pressing need to develop new methods for the prioritization of leads from high-throughput screens.

Big data "Omics" systems have revolutionized data acquisition and analysis in almost all fields of the physical sciences and have provided a foundation to answer broader questions about biological and ecological systems. While earlier studies typically focused on evaluating small numbers of genes, proteins, or metabolites, recent approaches are becoming more complex. The field of natural products is no

exception. Recent studies have covered a range of approaches including the use of meta-omic characterization of marine microbial communities to identify biosynthetic gene clusters responsible for the production of the clinical anticancer drug ET-743,[3] the use of an informatic search algorithm "iSNAP" for Non-ribosomal Peptide (NRPS) de-replication,[4] and the creation of platforms like AntiSMASH[5] and NaPDoS[6] for the rapid identification, annotation, and phylogenetic analysis of biosynthetic gene clusters from genomic and meta-genomic sequence data. However, to date, these systems have not integrated biological and chemical datasets into a single platform.

Recently we developed a high-content image-based phenomics screen for the discovery of bioactive natural products from complex mixtures.[7] Using this system we clustered marine microbially derived prefractions based on mechanistic class, and demonstrated that the phenotypic profiles of purified individual constituents are consistent with the phenotypes induced by the crude prefractions. This "function-first" approach to natural products discovery helps eliminate the high rates of rediscovery but is slow because chemical annotation of each prefraction still requires extensive follow-up, making the examination of large libraries labor intensive. This new study extends the system by integrating data from ultraperformance liquid chromatography coupled time-of-flight mass spectrometry (UPLC TOF-MS) in order to correlate biological phenotypes with the chemical constitution of test prefractions.

UPLC-MS has become a mainstay for investigating chemical constituents in biological and environmental samples because of its broad applicability to a range of

analytes, and the value of the information it provides in terms of resolution, sensitivity, and mass accuracy. The subsequent generation and use of database systems for searching and comparing compounds from mass spectrometric experiments is essential for natural products drug discovery and has been applied for discovery,[8] compound identification[9], bacterial strain de-replication,[10] and high-throughput profiling.[11,12] This analysis has reduced the high rates of rediscovery and aided in compound identification; however, the data interpretation is not often automated and the analysis of large libraries of samples is difficult because of the enormous numbers of MS features detected in each run. We are able to distill the pertinent bioactive *m/z* features from a large MS dataset by integrating biological screening data. Integrating the biological data allows us to create filters by associating mass to charge *(m/z)* features with specific phenotypes and eliminating features with no observed activity in the phenotypic screen.

Herein we show how Compound Activity Mapping has fundamentally altered our approach to the investigation of natural products libraries by overcoming many of the problems of natural products drug discovery including secondary chemical analysis throughput, high rates of rediscovery, and sample prioritization. The strategic combination of UPLC TOF-MS secondary metabolomics and high-content screening provides the first global evaluation of all bioactive compounds from a large library of complex mixtures for a specific biological system. While we have applied this approach to Cytological Profiling (CP), Compound Activity Mapping is designed to integrate UPLC-MS data with any orthogonal high-content assay, and as such we

imagine its application with whole cell/organism genomics,[10] transcriptomics,[13] and antibiotic profiling.[14]

2.2. Hypothesis and Design:

Current advances in high-content screening and MS based profiling strategies have drastically improved traditional live or dead screens and bioassay guided fractionation methods for natural products discovery, but there are still many drawbacks to these techniques.[1,2,15] Because natural product libraries are complex mixtures of unknown constitution and titer, it is often unclear whether samples with different bioactivities are distinct because they contain different secondary metabolites with differing modes of action, are mixtures of compounds, or contain the same compound at varying concentrations. Similarly, these same properties of natural product libraries make MS data acquisition and analysis difficult because there can be hundreds of peaks of different magnitudes with no clear indication of which compounds are bioactive.

The central assumption of modern natural products discovery is that individual molecules, or families of molecules, from natural products libraries are responsible for both the measured biological effects in assays and the chemical signatures from spectrophotometer experiments, typically coupled to ultraperformance liquid chromatography (UPLC). It follows that the biological and MS data from the same extract should be reciprocally related by chemical constitution; therefore, we hypothesize that by using informatics techniques, it should be possible to correlate mass spectral features with specific, but not predefined,

biological phenotypes, and to use these data to predict the molecular constitution and phenotypic activity of specific molecules directly from complex mixtures. To test this hypothesis we developed Compound Activity Mapping, an interactive software tool that integrates data from high content screens with a database of validated mass to charge (m/z) features from a set of standard UPLC-coupled TOF-MS experiments (Figure 2-1).

By estimating the biological activity for each of the validated *m/z* features, Compound Activity Mapping is able to filter prefraction MS data so that hundreds of observed features are reduced to a small number of pertinent features that are predicted to induce the phenotype observed in cells treated with that prefraction. To gain a global view of the bioactivity and chemical constitution of complex natural product libraries, Compound Activity Mapping uses MS data to augment the prefraction's CP fingerprint so that prefractions that induce similar phenotypes and have the same validated *m/z* features appear as a cluster in a network diagram. The combination of these two techniques allows us to quickly triage commonly encountered bioactive molecules and generate hypotheses about which *m/z* features drive the activity in prefractions with unique bioactivities. In order to confirm the fidelity of our system we examined a 312-member subset from our prefractionated marine natural products library that had been previously screened in our in house CP assay and partially chemically annotated.[3,7]

In order to overcome some of the technical difficulties associated with the MS analysis of natural products libraries such as concentration differences between

samples and varying numbers of unknown analytes in each sample, we created an alternate approach to technical replicates and peak alignment that we refer to as feature validation (Chapter 2.5.4.1). A single MS experiment can yield hundreds of signals after molecular feature extraction (MFE). Comparing multiple runs in different ionization modes and detector settings adds even more complexity. The validation method we developed allows us to store only the highest quality data from two technical replicates and combine positive and negative ionization mode data and eliminates extraneous peaks due to detector ringing from exceptional concentrations of certain compounds.

Targeted and nontargeted MS based metabolomics have been used to identify molecular markers for disease states and have been used to identify secondary metabolites from producing organisms; however, in most cases, the number of experimental test groups was relatively low and static. Some recent examples in the field of natural products include the use of nontargeted metabolomics to discover new secondary metabolites from *Streptomyces coelicolor* and *Myxococcus xanthus* after media source perturbation and targeted mutagenesis, respectively.[8,16,17] These studies successfully used integrated MS-based metabolomics approaches, but with static independent variables: samples from one organism were either perturbed, or they were not. We have expanded on this approach using high-content screening to allow us to examine a natural products library by neither limiting nor defining the phenotypic profiles. Instead of comparing secondary metabolites between just two

groups, we correlate thousands of compounds with any number of known and unknown potential modes of action, as defined by the phenotypic profiles.

To integrate the validated MS data for each prefraction with the corresponding biological data we start by estimating the biological activity of individual mass features by creating a synthetic fingerprint that is the average of the biological fingerprints of prefractions that contain that feature (Chapter 2.5.4.3). We use two metrics to quantify the overall activity of an *m/z* feature and its biological specificity that we have termed "activity score" and "cluster score." The activity score is calculated from the synthetic fingerprint and is the sum of the square of the individual biological parameters and represents the severity of the phenotype of a particular *m/z* feature. The cluster score describes the consistency of the biological phenotypes induced by prefractions that contain the *m/z* feature (Chapter 2.5.4.4). If an *m/z* feature has a high cluster and activity score, it indicates that the prefractions containing that *m/z* feature induce a similar, strong phenotype, and therefore, that the *m/z* feature likely drives the activity of those prefractions in which it is present. The corollary is that if a feature has a low cluster and activity score, it is present in extracts with different phenotypes and on average has little to no activity (e.g. primary metabolites that are present in many prefractions).

One of the modules of Compound Activity Mapping plots *m/z* features in Cartesian coordinates with activity score on the vertical axis, the cluster score on the horizontal axis with the color of the dot corresponding to the retention time. These activity plots allow the user to quickly analyze prefractions because *m*/z features

associated with strong and consistent phenotypes stand out from those associated with many different or weak phenotypes. By filtering *m/z* features based on activity score and cluster score following blank subtraction, Compound Activity Mapping considers and outputs only those features that are candidates for the observed bioactivity.

Compound Activity Mapping filters and plots are useful tools for carefully examining individual prefractions and prefractions with similar bioactivities; however, this piecewise method does not provide a broad view of the entire dataset. In contrast, network analysis is a powerful tool for analyzing correlated data that can display both global trends and fine-scale information about the behaviors of individual constituents. There are many examples in which combinations of proteomics, transcriptomics, and metabolomics networks have been used to discover relationships within datasets including the identification of shifts in *Bacillus subtilis* regulation between two carbon sources,[5,18] and the use of "Molecular Networks" to identify classes of compounds from tandem mass spectrometry experiments based on fragmentation pattern similarities.[6,19] We have modified these approaches significantly to incorporate two independent datasets by using chemical similarity observed as shared *m/z* features to link samples displaying similar phenotypes.

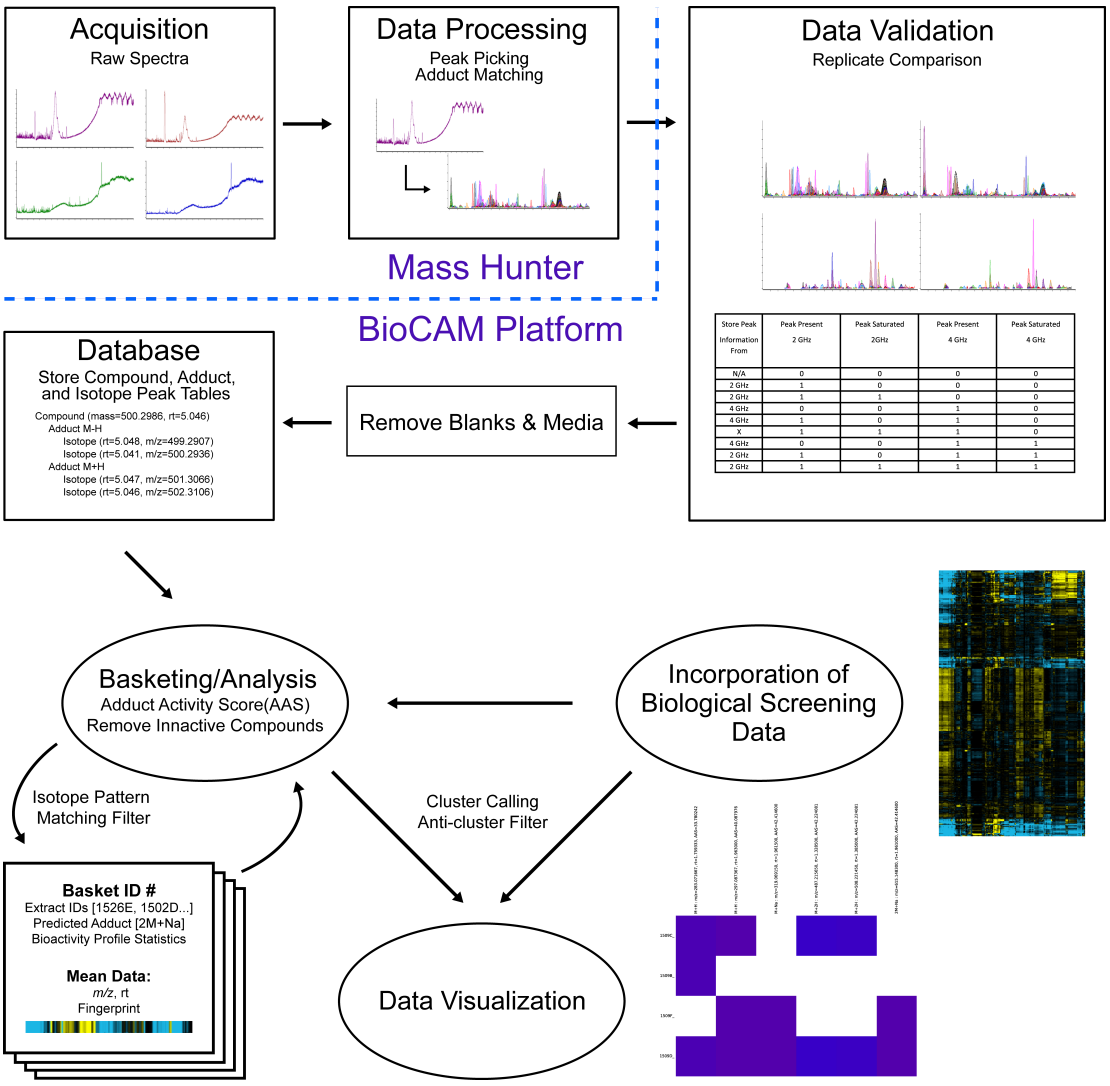


Figure 2-1: A flowchart depicting the workflow of Compound Activity Mapping from the Mass Hunter software to customoized graphical outputs.

For our study system, biologically related samples can be linked together by similarity scores and chemically related samples can be linked by shared *m/z* features. However it is not obvious whether samples far from each other in the biological fingerprint network are distinct because they have different chemistries, or if they

share the same compounds but at different titers. Also, in networks that only include mass feature relatedness, there is no delineation between clusters defined by biologically active or inactive compounds. To simultaneously use information from both datasets, we created an algorithm to reposition the n-order biological fingerprints by shared chemistry from the MS data that we have termed vector compression. First, a table of *m/z* feature connections between prefractions is created from the validated and filtered MS data to make a connectivity table. For each prefraction, we calculate a vector to every other prefraction with which it shares chemistry in the connectivity table. Each vector is weighted based on the likelihood that the *m/z* feature shared by the two prefractions are responsible for the observed phenotype of both samples based on the phenotypes of the two prefractions and the synthetic fingerprint of the analyte. These vectors are then normalized by the number of shared features to limit the movement of prefractions based on families of compounds or compounds with many adducts. In order to limit the change of the prefraction's fingerprint based on any one vector, we scaled each vector twice: once by the largest likelihood observed for that prefraction and again by the largest likelihood observed for all prefractions. Finally, each prefraction's connection vectors are summed, normalized by the number of vectors, and added to the prefraction fingerprint to move the fingerprint in n-dimensional space. Iterations of this algorithm augment each sample's fingerprint/position in the network, gradually drawing together samples with related bioactivity and shared *m/z* features whose synthetic fingerprints closely resemble those of the prefractions (Figure 2-2).

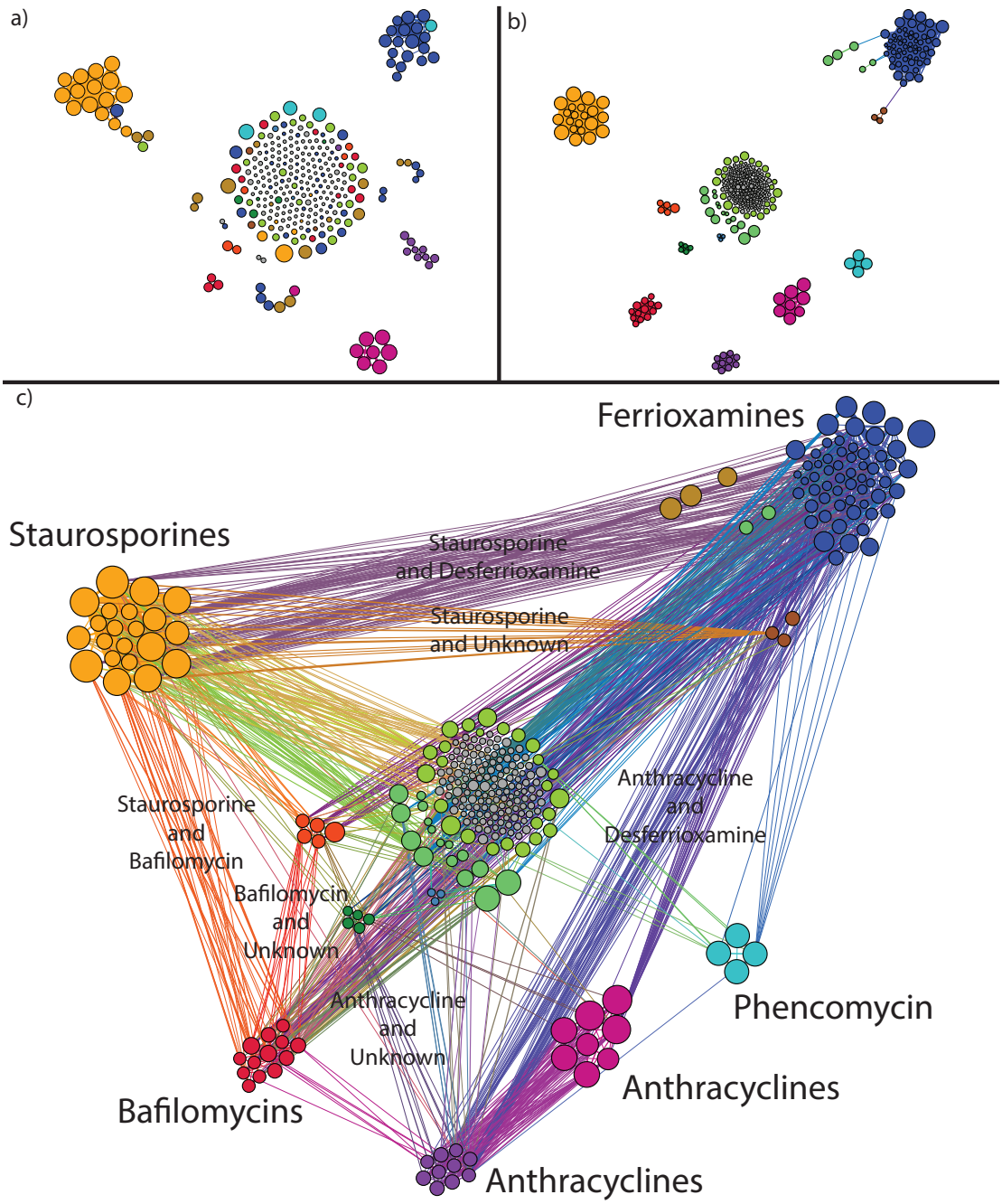


Figure 2-2: Network representations of the extracts clustered by the similarity of their cytological profile. a) before vector compression b) after vector compression and c) after vector compression with the known active metabolites from each cluster.

Following vector compression, repositioning based on related phenotypes and MS features, relationships between prefractions can be visualized using a network diagram. The prefractions, represented by nodes, are connected by edges that can be defined by the similarity of two nodes' augmented biological fingerprint or by shared *m/z* features. In this way it is easier to develop hypotheses about related samples and the molecular features responsible for inducing observed phenotypes. This visualization allows the user to condense biological clusters of prefractions that share the same chemistry at varying concentrations and to separate biological clusters of varying chemistry with similar modes of action.

2.3. Results and Discussion:

In order to test Compound Activity Mapping we applied it to a 312-member microbial extract library that had been previously partially chemically annotated.[7] Compound Activity Mapping can accurately predict which chemical constituents drive the observed CP phenotypes in samples containing either a single bioactive compound, mixtures of bioactive compounds, or previously unknown bioactive compounds. We are able to overcome many of the pitfalls of traditional natural products discovery by using networks to analyze the entire library and annotate clusters possessing distinct biological and chemical properties. Using these integrated biological and chemical fingerprints we can distinguish large clusters defined by the presence of known bioactive natural products, identify previously unknown relationships between samples, and prioritize chemically and biologically distinct prefractions for future lead discovery.

2.3.1. Network Analysis of Chemical and Biological Datasets:

Using the network analysis modules of Compound Activity Mapping we can visualize the original CP data as groups of prefractions (nodes) that are connected by edges if their fingerprints are closely related - have a Pearson correlation greater than 0.9 (Figure 2-2). The network approach improves upon hierarchical clustering approaches because it does not constrain the high-content data to a single dimension; however, distributing the dataset based solely on biological phenotype is not always accurate because there are still prefractions that do not group correctly according to previous annotation and the MS data. After each cycle of vector compression, we can track the changes in biological fingerprints as a function of iteration by observing the creation and destruction of edges. After 749 iterations, equilibrium is reached: from this point on very few new edges are created or destroyed (Figure 2-2). The resulting network contains eight large (greater than five nodes) clusters that are biologically and chemically distinguished by the presence of frequently encountered microbial natural products (anthracyclines, bafilomycins, ferrioxamines, *etc*.). The prefractions present in these clusters are consistent with previous annotation and many discrepancies between the presence of known compounds and misleading arrangement in hierarchical clustering are removed. Chemical and biological network analysis also accounts for differences in phenotypes that could arise from concentration variation and mixtures of bioactive compounds while leaving the phenotypes of prefractions with no biologically significant shared chemistry unaltered.

2.3.2. Cluster Annotation:

The most prominent cluster that is defined by the presence of a single compound is the cluster containing the pan-specific kinase inhibitor staurosporine (yellow circles, Figure 2-2). Based on the biological and chemical datasets, Compound Activity Mapping predicts a single *m/z* feature shared amongst prefractions 1502D, 1502E, 1502F, 1504E, 1505D, 1505E, 1513D, 1513E, 1526D, 1526F, 1530C, 1530D, and 1530E corresponding to the $[M + H]^+$ adduct of staurosporine (Figure 2-3). This result is consistent with the annotation for several of these prefractions previously reported by Schulze *et al*.[7,8] To experimentally validate the predicted activity of staurosporine, we examined a dilution series of a commercial sample of staurosporine by CP to determine the phenotypic effects of the pure compound across a broad concentration range. Encouragingly, the synthetic fingerprints for the $[M + H]^+$ and the $[M + Na]^+$ adducts of staurosporine from the Compound Activity Mapping predictions both closely resemble the fingerprint of the pure compound at 0.43 μM (Figure 2-4). This result shows that Compound Activity Mapping is able to accurately predict the phenotypic effects of specific bioactive secondary metabolites directly from natural products libraries, and can connect chemical and biological attributes of specific metabolites from the primary screening data of complex extracts.

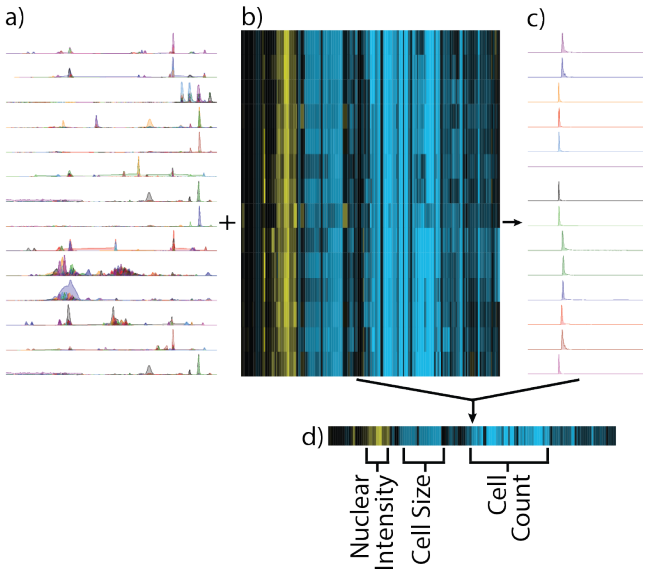


Figure 2-3: (a) Total Compound Chromatograms (TCC) for 14 prefractions in the staurosporine cluster. Visually, no single *m/z* feature can be identified as the common bioactive constituent from the aligned TCCs. (b) CP fingerprints for each of the prefractions. (c) Extracted Ion Chromatograms (EIC) for the $[M + H]^+$ adduct of staurosporine. (d) The synthetic fingerprint of the $[M+H]^+$ adduct of staurosporine and the fingerprint of commercial staurosporine (0.43 μM).

The presence of multiple bioactive compounds in some prefractions is to be expected for natural products libraries, where initial extracts are of unknown constitution and titer. Whole genome sequencing has shown that many Actinobacteria contain upwards of 35 biosynthetic gene clusters,[9,20] any of which may be producing secondary metabolites under a given set of fermentation conditions. To demonstrate that Compound Activity Mapping can accurately characterize prefractions containing multiple bioactive metabolites, we examined a case that exhibited 'mixed-mode' phenotype in the original CP screen.

Prefraction 1504E initially displayed a distinct phenotype to the other prefractions in the test set. Examination of the UPLC-MS data for this prefraction

revealed the presence of two frequently occurring features. These features were identified as the $[M + H]^+$ adduct of staurosporine and the $[M + Na]^+$ adduct of bafilomycin A1, based on molecular formulae and retention time matches with previously reported compounds.[7,10] Although both compounds are potently bioactive against mammalian cells (pan-specific kinase inhibitor and vacuolar ATPase inhibitor, respectively), the CP fingerprint for the original prefraction was not representative of the biological activities of either metabolite. However, the predicted synthetic fingerprints for each of these metabolites share significant similarities with the CP profiles of the respective pure compounds (0.83 and 0.64 Pearson similarity scores for staurosporine and bafilomycin), demonstrating that this method is effective at directly predicting the biological behaviors of individual constituents in complex mixtures of metabolites with different modes of action, even if the CP fingerprint of the prefraction does not accurately represent the fingerprints of either of the individual constituents (Figure 2-4).

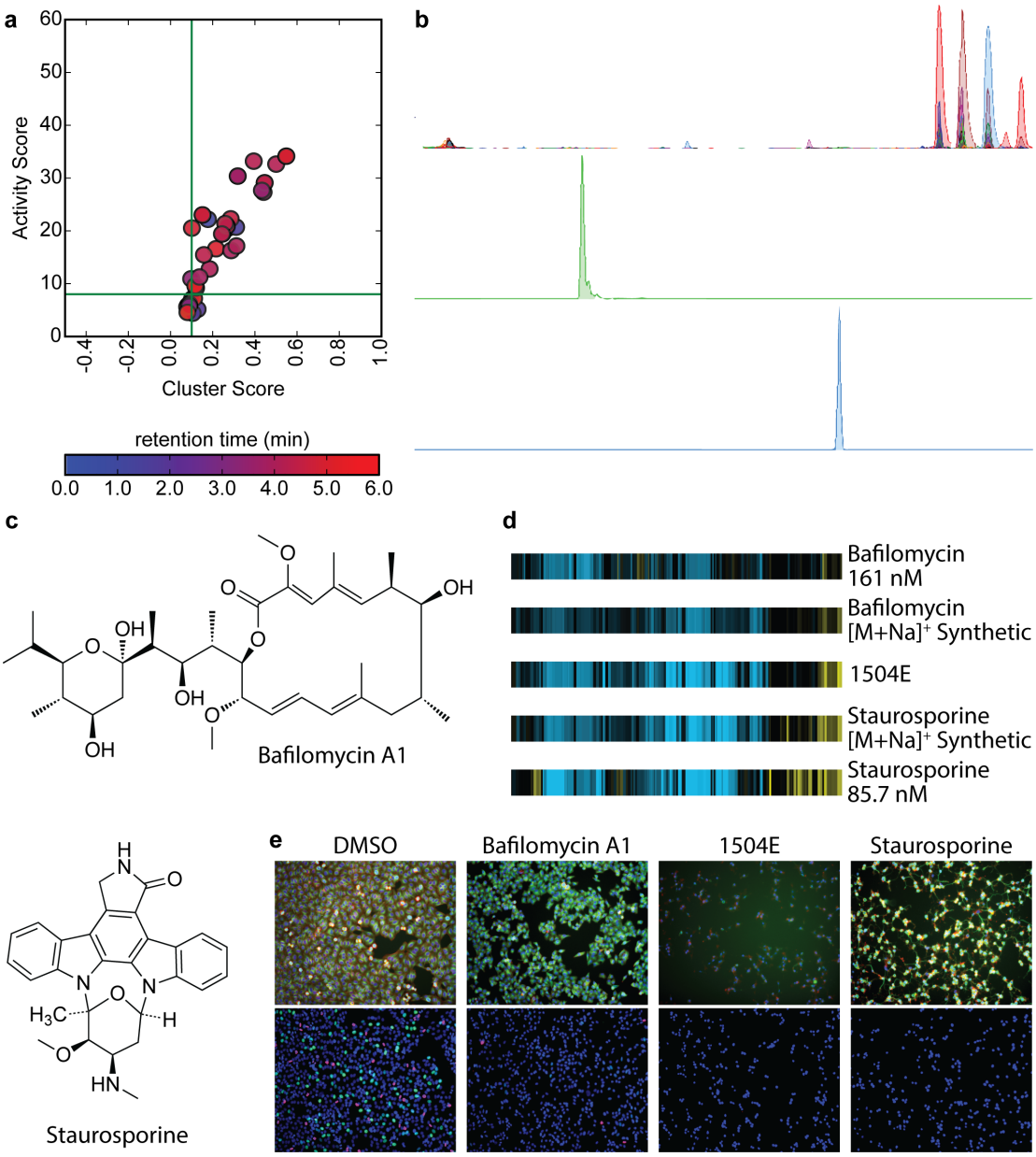


Figure 2-4: (a) The activity plot visually shows active *m/z* features, with highlighted $[M+H]^+$ of staurosporine (green) and the $[M+H]^+$ of bafilomycin A1 (blue). (b) From top to bottom: TCC of 1504E, EIC of staurosporine $[M+H]^+$, and EIC of bafilomycin A1 $[M+H]^+$. (c) Structures of bafilomycin A1 (top) and staurosporine. (bottom) (d) Observed and synthetic CP fingerprints of both bafilomycin A1 and staurosporine, as well as the mixed-mode phenotype observed with 1504E. (E) Cell images used to generate the original CP fingerprints.

2.3.3. Annotation of compounds with multiple phenotypes:

Anthracyclines represent a large family of compounds that are frequently encountered in Actinobacterial natural products discovery. This class of compounds is well represented among the prefractions in our screening library, but is split into two distinct clusters in the network diagram (Fig. 3), despite containing significant shared chemistry. We hypothesized that this difference in phenotype was due to concentration differences between the prefractions, leading to intermediate vs. end-point phenotypes for the two cluster classes. We envisioned that at a specific concentration the effect of treatment with anthracyclines would switch from a classic G1-S stall to an apoptotic phenotype. To test this hypothesis we performed a dilution series of prefraction 1498D and observed a rapid transition in phenotype within a 4-fold dilution window (Figure 2-5). This observation indicates that, while in many instances cytological profiles remain constant over wide concentration gradients, these clusters can become segregated where defined phenotypic shifts occur with changing concentration. As the annotation of the platform with known compounds matures, we predict that these clusters will afford additional value, in that they will report on relative concentrations of specific metabolites, in addition to providing structural and biological information about active constituents. Thus the anthracycline example illustrates how Compound Activity Mapping is able to retain and report on phenotypic information, even in situations where compound concentrations are significantly different.

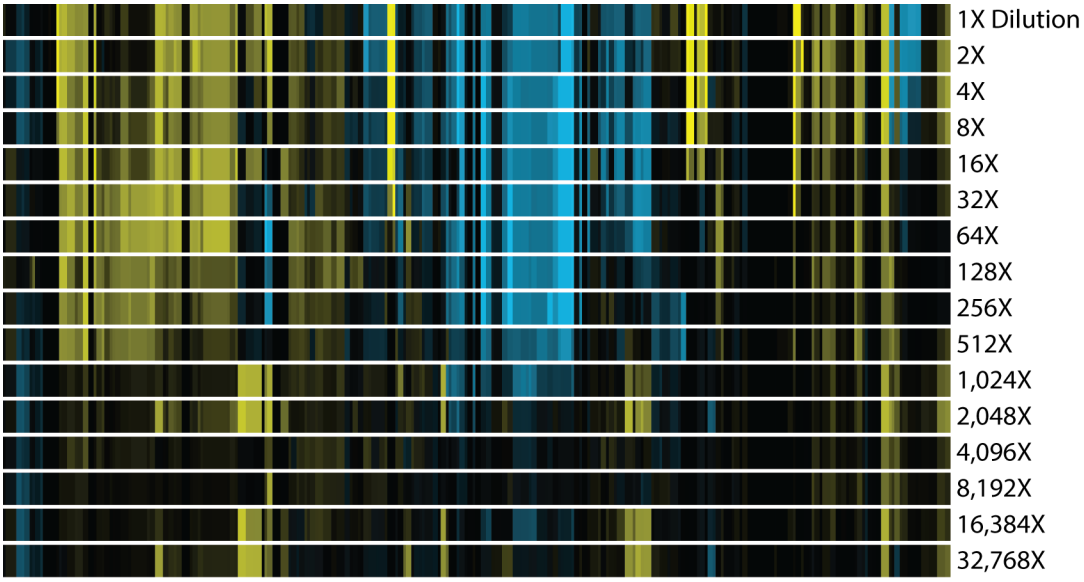


Figure 2-5: Cytological profiles of a 2-fold dilution series of 1498F clearly illustrating how anthracycline compounds may have multiple distinct phenotypes at different concentrations

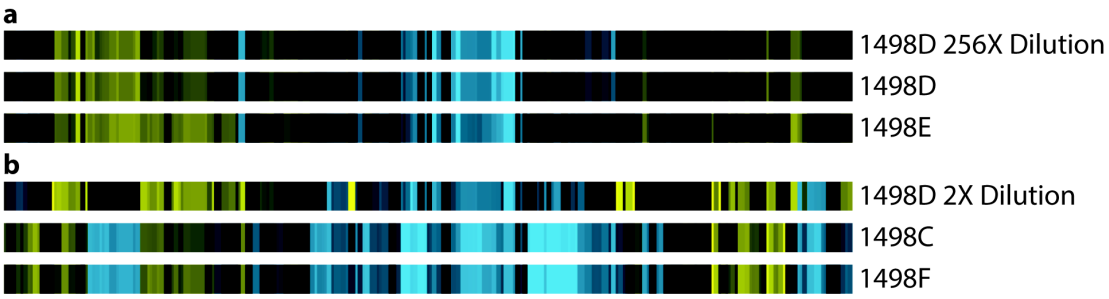


Figure 2-6: Cytological profiles from two anthracycline clusters (a) fractions exhibiting G1/S stall (b) fractions causing apoptosis.

2.3.4. Examining Unknown Clusters:

The main advantage of Compound Activity Mapping is that it allows the correlation of biological data based on shared chemistry in order to identify prefractions with novel bioactivity and chemistry. For example, the cluster containing prefractions B, C, D, and F from extract 1509 did not contain compounds previously identified in our library (Teal circles, Figure 2-7). Comparing the activity plots of the prefractions, one compound is present in all four prefractions **.** Isolation of the compound defined by the *m/z* feature ($[M+H]^+$ = 283.0714, rt = 1.75 min) revealed the known bioactive compound phencomycin by HR-TOFMS and ^{1}H NMR.[11,12,21] The related metabolite 1,6-dicarbomethoxy phenazine ($[M+H]^+$ = 297.0873, rt = 1.96 min), which shares the same UV profile and a similar synthetic fingerprint, was also isolated and confirmed by ^{1}H NMR. Several other *m/z* features appear in the output for 1509D in addition to phencomycin and 1,6-dicarbomethoxy phenazine, some with even higher activity and cluster scores; however, by tracking the values of the weighting function through iterations of our chemical and biological redistribution algorithm, we were able to demonstrate that the phenazine compounds were responsible for the observed bioactivity due to the large log order differences in the feature likelihood scores between the phenazine features and all others (red and blue lines for phenazines, versus green, navy and yellow lines, Figure 2-7). Following mass-guided purification of both metabolites their activities were examined by CP. Both phencomycin and 1,6-dicarbomethoxy phenazine $[M+H]^+$ possessed CP fingerprints with high similarities to both the predicted synthetic fingerprints, and the

fingerprint of the original prefraction (Figure 2-7), demonstrating that compound activity mapping can accurately predict the identities of unknown bioactive constituents, and that the predicted biological properties correlate well with the actual bioactivity fingerprints for the pure materials.

It is important to note that prior to chemical and biological integration the phenazine-containing prefractions were dispersed throughout the DNA cluster, suggesting that their mechanism of action involved DNA damage or the obstruction of core replication or transcription processes. Previous work has indicated that phencomycin is capable of binding DNA, leading to a G1/S mitotic stall. Examination of the images of cells treated with phencomycin and 1,6-dicarbomethoxy phenazine show a decrease in mitotic cells (reduced pHH3 staining) and a concomitant decrease in cells actively undergoing DNA synthesis (reduced EdU staining) indicating a G1/S stall phenotype that is common to modulators of DNA synthesis (Figure 2-7).

The elucidation of the phenazine structures demonstrates Compound Activity Mapping's ability to synthesize meaningful correlations between multi-parametric chemical and biological datasets in order to discover relationships between prefractions that were not obvious in either dataset alone. Had phencomycin not been previously discovered by traditional screening methods, this result would illustrate Compound Activity Mapping's ability to identify and biologically characterize novel bioactive secondary metabolites in an automated high throughput manner.[10,13]

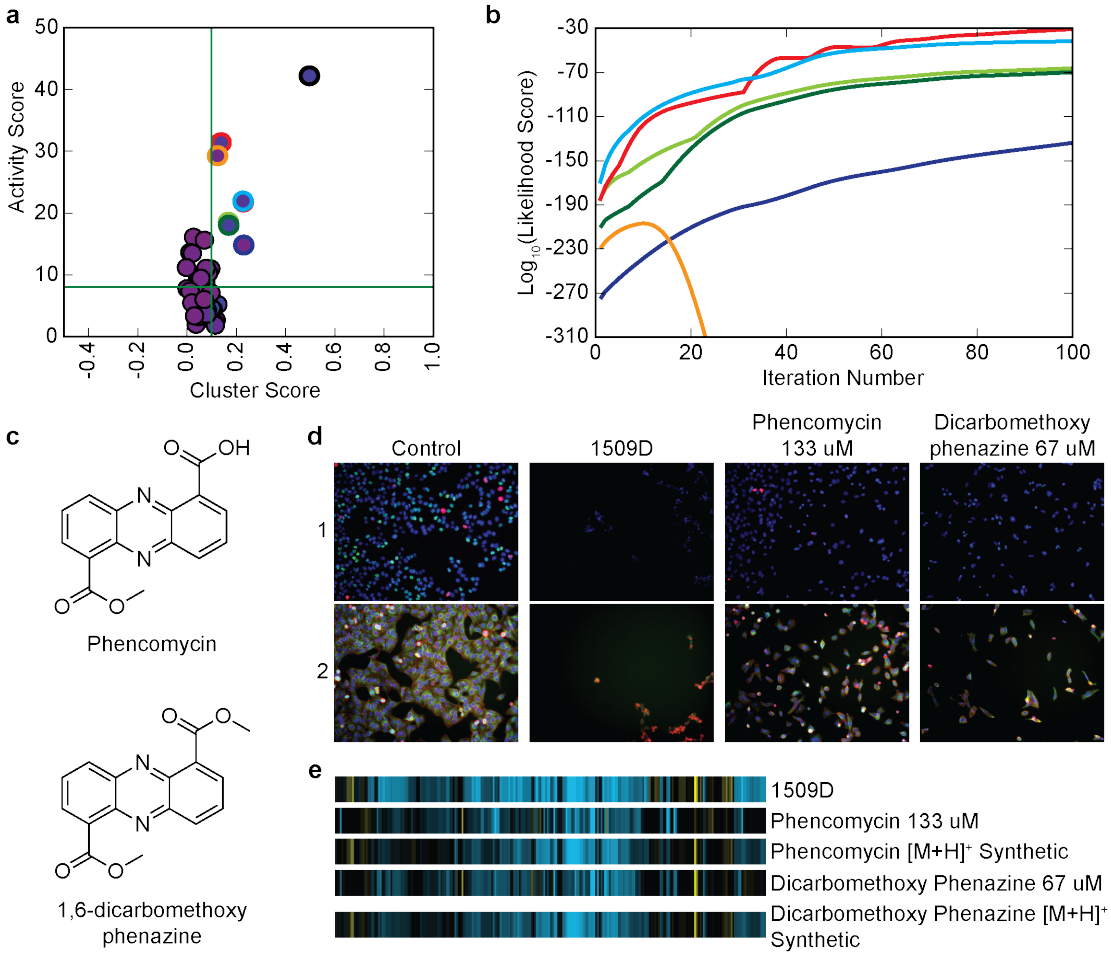


Figure 2-7(a) The activity plot of 1509D contains several active *m/z* features. (b) Likelihood scores for *m/z* features plotted as a function of iteration (two of the most likely features are overlapped, with cyan being visible over red). Color-coding between plots was done by compound when it could be assigned or by *m/z* feature otherwise. Note: the most active compound shown in the activity plot is not shown in the likelihood plot because it falls to zero (undefined in log-scale) within several iterations. (c) Structures of phencomycin and 1,6-dicarbomethoxy phenazine. (d) Cell images from DMSO control, prefraction 1509D, pure phencomycin 133 μM, and pure dicarbomethoxy phenazine 67 μM. (e) Fingerprints of pure phencomycin and dicarbomethoxy phenazine compared with the original 1509D prefraction and the calculated synthetic fingerprints of the $[M+H]^+$ adducts of phencomycin and dicarbomethoxy phenazine.

2.3.5. Limitations and Considerations:

While this new strategy greatly aids in the analysis of high-throughput high-content biological screening results in conjunction with MS data, there are three limitations worth addressing. First, if a *m/z* feature is detected in only one prefraction, then its synthetic fingerprint will be identical to that of the prefraction and no correlation analysis with other prefractions can be performed. We anticipate that as Compound Activity Mapping matures with the inclusion of additional library members, the number of unique *m/z* features will decrease and truly unique features will therefore represent important lead molecules that can be directly selected for further analysis. Second, if a compound *y* is only expressed in conjunction with a more broadly expressed molecule *x* that has a strong phenotype, the true phenotype of *y* may be masked by the dominate activity of *x*. Compound *y* will still be listed as active and presented in the activity plots, but the synthetic fingerprint will likely not represent the true activity of the molecule. This is an inherent limitation of the technology but, again, expansion of the datasets by increasing the number of prefractions analyzed should reduce the probability of encountering metabolites only as co-expressed pairs. Third, at this time Compound Activity Mapping cannot identify in source fragmentation and is only based on MS^1 data. Expansion of this technology to include tandem MS could aid in feature alignment and the identification of compound families. Informatics strategies for such an approach are now well established, making this a realistic and accessible improvement to the platform for research teams with access to the required mass spectrometry hardware.

2.4. CONCLUSIONS:

Herein we have demonstrated that Compound Activity Mapping is able to overcome several of the traditional limitations of natural product drug discovery through the correlation and integration of large orthogonal chemical and biological datasets. Through the application of this automated high throughput tool to a 312-member prefraction plate from our library of marine-derived microbial extracts, we have demonstrated that Compound Activity Mapping is able to predict the active constituent(s) of bioactive mixtures directly from primary screening data, and to enable the grouping of prefractions into both known and unknown chemical and biological clusters. In this way, both the constitution and mode of action of extracts are elucidated at the primary screening stage. This reverses the traditional process of natural product drug discovery by providing researchers with hypotheses about all compounds in all extracts prior to lead selection and prioritization. We have applied Compound Activity Mapping to natural products and CP; however, it is designed to integrate any MS based dataset with any high-content dataset and we imagine its use in a variety of fields including environmental studies, chemical ecology, and genome-guided natural products discovery.

2.5. METHODS:

2.5.1. Library Preparation*:*

Cell culture, extraction, and prefractionation is described in detail by Schulze, *et al*.[7,13] Microbes were isolated from marine sediment off the west coast of the United States, American Samoa, and Hawaii, grown under standard fermentation conditions with XAD-16 resin, extracted with 1:1 methanol/dichloromethane, and fractionated on a reverse phase C_{18} column with an eleutropic series of water and methanol. These fractionated extracts or prefractions were dried and resuspended in 1 mL of dimethylsulfoxide and diluted 1:5 in DMSO for CP screening and 1:50,000 in 50% v/v methanol/water into Corning V bottom 96 well plates for MS analysis.

2.5.2. UPLC TOF-MS:

UPLC TOF-MS experiments were performed using an Agilent 1260 binary pump in low dwell volume mode, an Agilent column oven heated to 45° C, and an Agilent 6230 Time-of-flight Mass Spectrometer with an electrospray ionization (ESI) source. One μL of sample, dissolved in 50% v/v methanol/water, was injected onto a 1.8 μm particle size, 50 x 2.3 mm I.D., ZORBAX RRHT column. Each sample was subjected to a water-acetonitrile gradient from 10% to 90% organic over 4 min with a 1.5 min hold at 90% organic before a 3 min re-equilibration. The flow rate was maintained at 0.8 mL/min. Formic acid, 200 μL/L, was added to both the water and the acetonitrile. Water, 1 mL/L, was added to the acetonitrile.

The mass spectrometer was run with a detector range from 100 to 1700 *m*/z. The ESI source was operated with a desolvation temperature of 350° C and a drying gas flow rate of 11 L/min. The fragmentor voltage was held at 135 V. In positive ESI mode, the capillary voltage was ramped from 2500 V at 0 min to 2750 V at 1 min, and to 3000 V at 3 min. In negative ESI mode, the capillary voltage was held at 2750 V. Each sample was run in both positive and negative ESI source modes and in both high resolution (4GHz) and extended dynamic range (2GHz) detector modes. We selected peaks using the Find By Molecular Feature algorithm in Agilent Mass Hunter Software with raw ion cutoff of 300 counts and a compound cutoff of 1000 counts. Selected peaks were exported in a CEF file containing isotopic abundance, rt, and metadata.

2.5.3. Cytological Profiling:

Methods for cell culture and staining are identical to those presented by Schulze, *et al*.[7,14] HeLa cells were plated in two 384-well plates and incubated for 24 hours at 37°C. The plates were then stained with either nuclear or cytoskeletal stain sets. Finally, plates were imaged with a 10x objective lens with four images per well for each stain wavelength in a plate. Prefraction-treated wells were compared with DMSO-treated wells affording a 248-parameter fingerprint for each prefraction-treated well indicating the positive or negative perturbations in each parameter.

2.5.4. Compound Activity Mapping:

Unless stated otherwise, *m/z* feature alignment between samples uses high-resolution mass (ppm), retention time, and an isotope pattern matching method adapted from Pluskal *et al*. [22]

2.5.4.1. MS Data Validation:

We developed a decision tree to align *m*/z retention time (rt) pairs between extended dynamic range (2 GHz) and high-resolution (4 GHz) detector modes in order to store the most reliable data from both positive and negative electrospray ionization (ESI) experiments. After initial data acquisition, processing, and CEF file (peak list) output, we align and subtract MeOH blank and SYP media run peaks using 20 ppm, 0.4 minutes, and 0.5 isotopic score difference windows from the sample prefraction run peak lists. We also remove detector ringing from peaks above the detector saturation limit by removing all peaks within 0.4 minutes and 1 mass unit of the most abundant peak in saturated data. Then we align peaks between 4 GHz and 2 GHz modes with 10 or 20 ppm, 0.4 minutes, 0.5 isotopic score difference windows. In order to store only the most accurate data, each aligned peak is assigned a tag that indicates whether or not the *m/z* feature is present or present and saturated in both 4 GHz and 2 GHz modes. We use a decision tree to select the best value for aligned features, with priority being given to *m*/z values that are not saturated from the 4 GHz data. We store post-validated *m/z* features in a SQLite database for rapid indexing during incorporation with biological data.

2.5.4.2. Integrating TOF-MS Data and Cytological Profiling Data:

In order to integrate the CP and MS datasets, each *m/z* feature stored in the database is ascribed a synthetic fingerprint, an activity score, and a cluster score.

2.5.4.3. Synthetic Fingerprints:

The synthetic fingerprint of an MS feature is a set of pairs, one for each attribute measured in cytological profiling, where each pair is the average and standard deviation of the attribute value for the set prefractions that contain that MS feature.

$$F = \{f_1, f_2, \dots, f_{j-1}, f_j\}$$

$$f_k = \{a_1, a_2, \dots, a_{n-1}, a_n\}$$

$$SyntheticFingerprint(f) = \{(\bar{a}_1, \sigma_{a_1}), (\bar{a}_2, \sigma_{a_2}), \dots, (\bar{a}_{n-1}, \sigma_{a_{n-1}}), (\bar{a}_n, \sigma_{a_n})\}$$

Equation 2-1: calculation of the synthetic fingerprint where F is a set of fractions f_1-f_j each with attributes a_1-a_n. The average and standard deviation for each attribute across all the fractions is stored in a set of two value lists.

2.5.4.4. Activity Score:

From the synthetic fingerprint we calculate an activity score that is defined as the sum of the square of the means of each attribute.

$$\boldsymbol{ActivityScore(f_k \text{ or } feature) = \sum_{i=1}^{n} a_i^2}$$

Equation 2-2: The activity score is calculated as the magnitude of the cytological profile vector.

2.5.4.5. Cluster Score:

We also calculate a cluster score for each MS feature that is the average of the Pearson correlations between all combinations of two biological fingerprints of the prefractions that contains that MS feature.

$$\boldsymbol{ClusterScore(feature)} = \frac{\sum_{q=1}^{j}\sum_{q=1}^{j}\boldsymbol{Pearson}\left(f_p, f_q\right)^3 - j}{j^2 - j}$$

Equation 2-3: The pairwise Pearson correlation between each fraction containing a particular *m/z* features is cubed, summed, and divided by the number of relationships in order to calculate the cluster score or the consistency of the *m/z* features biological phenotype.

2.5.5. Biological and Chemical Fingerprints:

We identify correlations between the chemical and biological datasets and visualize the entire dataset by incorporating chemical relatedness into the biological fingerprints. To use information from both datasets we created an algorithm to reposition the n-order biological fingerprints by shared chemistry from the MS dataset. A table of *m/z* feature connections between samples is created from the validated and filtered MS data to make a connectivity table. The algorithm then repositions each of the prefractions in n-dimensional space using weighted difference vectors between two prefractions containing the same *m/z* feature. The weighting is calculated based on the likelihood that the *m/z* feature connecting the two prefractions is responsible for the observed phenotype of both samples. In order to limit the change of the prefraction's fingerprint based on any one vector, we scaled the vectors

twice: once by the largest weighting observed for that prefraction and again by the largest weighting observed for all the prefractions. Each prefraction is then moved by the sum of the *m/z* connection vectors. Performing multiple iterations of this algorithm augments each sample's fingerprint/position in the network such that samples with related bioactivity and shared *m/z* features are drawn together (Figure 2-2).

2.5.6. Network Visualization:

We use NetworkX in python to create and edit networks and Gephi to visualize and analyze networked data. In all networks shown, nodes represent prefractions and edges connect nodes if the prefractions have Pearson correlations greater than 0.9. Using Gephi, we visually rank nodes by size proportional to prefraction activity score, and identify distinct clusters (represented by different colors) using network modularity with weighted edges and a resolution of one. We use Gephi's built in Force Atlas algorithm to distribute nodes 2-dimensionally with default parameters except: repulsion of 1,000, attraction of 15, gravity of 60, and adjust by sizes.

2.6. References:

(1) Newman, D. J.; Cragg, G. M. *Journal of Natural Products*. March 23, 2012, pp 311–335.

(2) Gerwick, W. H.; Moore, B. S. *Chemistry & Biology*. Elsevier Ltd January 27, 2012, pp 85–98.

(3) Rath, C. M.; Janto, B.; Earl, J.; Ahmed, A.; Hu, F. Z.; Hiller, L.; Dahlgren, M.; Kreft, R.; Yu, F.; Wolff, J. J.; Kweon, H. K.; Christiansen, M. A.; Håkansson, K.; Williams, R. M.; Ehrlich, G. D.; Sherman, D. H. *ACS Chemical Biology*. November 18, 2011, pp 1244–1256.

(4) Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. A. *Proceedings of the National Academy of Sciences of the United States of America*. November 20, 2012, pp 19196–19201.

(5) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; BREITLING, R. *Nucleic Acids Research*. Oxford University Press July 2011, pp W339–W346.

(6) Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. *PLoS ONE*. Public Library of Science March 29, 2012, p e34064.

(7) Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. *Chemistry & Biology*. Elsevier Ltd February 21, 2013, pp 285–295.

(8) Sidebottom, A. M.; Johnson, A. R.; Karty, J. A.; Trader, D. J.; Carlson, E. E. *ACS Chemical Biology*. September 20, 2013, pp 2009–2016.

(9) Nielsen, K. F.; Månsson, M.; Rank, C.; Frisvad, J. C.; Larsen, T. O. *Journal of Natural Products*. November 28, 2011, pp 2338–2348.

(10) Hou, Y.; Braun, D. R.; Michel, C. R.; Klassen, J. L.; Adnani, N.; Wyche, T. P.; Bugni, T. S. *Analytical Chemistry*. April 23, 2012, pp 4277–4283.

(11) Ito, T.; Odake, T.; Katoh, H.; Yamaguchi, Y.; Aoki, M. *Journal of Natural Products*. May 27, 2011, pp 983–988.

(12) Cuthbertson, D. J.; Johnson, S. R.; Piljac-Žegarac, J.; Kappel, J.; Schäfer, S.; Wüst, M.; Ketchum, R. E. B.; Croteau, R. B.; Marques, J. V.; Davin, L. B.; Lewis, N. G.; Rolf, M.; Kutchan, T. M.; Soejarto, D. D.; Lange, B. M. *Phytochemistry*. Elsevier Ltd July 1, 2013, pp 187–197.

(13) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. *Science*. American Association for the Advancement of Science September 29, 2006, pp 1929–1935.

(14) Wong, W. R.; Oliver, A. G.; Linington, R. G. *Chemistry & Biology*. Elsevier Ltd November 21, 2012, pp 1483–1495.

(15) December 27, 2011, pp 2545–2555.

(16) National Acad Sciences November 20, 2012, pp 19196–19201.

(17) Cortina, N. S.; Krug, D.; Plaza, A.; Revermann, O.; Müller, R. *Angewandte Chemie International Edition*. WILEY-VCH Verlag December 7, 2011, pp

811–816.

(18) Buescher, J. M.; Liebermeister, W.; Jules, M.; Uhr, M.; Muntel, J.; Botella, E.; Hessling, B.; Kleijn, R. J.; Le Chat, L.; Lecointe, F.; Mäder, U.; Nicolas, P.; Piersma, S.; Rügheimer, F.; Becher, D.; Bessieres, P.; Bidnenko, E.; Denham, E. L.; Dervyn, E.; Devine, K. M.; Doherty, G.; Drulhe, S.; Felicori, L.; Fogg, M. J.; Goelzer, A.; Hansen, A.; Harwood, C. R.; Hecker, M.; Hubner, S.; Hultschig, C.; Jarmer, H.; Klipp, E.; Leduc, A.; Lewis, P.; Molina, F.; Noirot, P.; Peres, S.; Pigeonneau, N.; Pohl, S.; Rasmussen, S.; Rinn, B.; Schaffer, M.; Schnidder, J.; Schwikowski, B.; Van Dijl, J. M.; Veiga, P.; Walsh, S.; Wilkinson, A. J.; Stelling, J.; Aymerich, S.; Sauer, U. *Science*. American Association for the Advancement of Science March 2, 2012, pp 1099–1103.

(19) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proceedings of the National Academy of Sciences of the United States of America*. National Acad Sciences June 26, 2012, pp E1743–E1752.

(20) Nett, M.; Ikeda, H.; Moore, B. S. *Natural Product Reports*. The Royal Society of Chemistry 2009, pp 1362–1384.

(21) Chatterjee, S.; Vijayakumar, E. K. S.; Franco, C. M. M.; Maurya, R.; Blumbach, J.; Ganguli, B. N. *The Journal of Antibiotics*. 1995, pp 1353–1354.

(22) Pluskal, T.; Uehara, T.; Yanagida, M. *Analytical Chemistry*. American Chemical Society April 26, 2012, pp 4396–4403.

3. COMPOUND ACTIVITY MAPPING: SECOND GENERATION PLATFORM FOR THE INTEGRATION OF SECONDARY METABOLOMICS AND HIGH-CONTENT SCREENING

3.1. Introduction:

Notwithstanding the historical importance of natural products in drug discovery[1] the field is facing a number of challenges that impact the relevance of natural products research in modern biomedical science. Among these are the increasing rates of rediscovery of known classes of natural products, and the high rates of attrition of bioactive natural products in secondary assays due to limited information about compound modes of action in primary whole cell assays.[2] Although many pharmaceutical companies recognize that natural products are an important component of drug discovery programs because of the different pharmacologies of natural products and synthetic compounds,[3] there is a reluctance to return to "grind and find" discovery methods. Therefore there is a strong need for technologies that address these issues and provide new strategies for the discovery of lead compounds with unique structural and/or biological properties.[2]

While natural product libraries are generally considered to be better suited to whole cell screens than target-based screening methods, traditional approaches based on live/dead assays for bioactivity guided fractionation are slow and have low probabilities of finding compounds with unique biological properties because the modes of action of lead compounds are not investigated until late in the discovery process. The challenge with whole cell screening is that target identification is often a

complex and time-consuming process, even with the latest advances in proteomic, metabolomic, and affinity capture methods. More generally, natural product drug discovery is difficult in any assay system because extract libraries are typically complex mixtures of small molecules in varying titers, making it difficult to distinguish biological outcomes.[4] This is compounded by issues of additive effects of multiple bioactive compounds, and the presence of nuisance compounds that cause false positives in assay systems. In order to address these issues our laboratory has recently developed several image-based screening platforms that are optimized for natural product discovery.[5-8] The cytological profiling platform optimized by Schulze *et al*. characterizes the biological activities of extract library members using untargeted phenotypic profiling, and uses the phenotypic profiles of natural products extracts and training set compounds of known mode of action to characterize this bioactivity landscape.[9,10] This cytological profiling tool forms the basis of the biological characterization component of the Compound Activity Mapping (CAM) platform, as described below.

In the area of chemical characterization of natural product libraries, untargeted metabolomics is gaining attention as a method for evaluating the chemical constitution of natural products libraries.[11-14] This approach has been used successfully for strain prioritization and preliminary extract dereplication; however, this approach typically requires in-house compound databases that take years to create and are often not transferable. Modern "genes-to-molecules" and untargeted metabolomics approaches taking advantage of PCA and MS^2 spectral comparisons

have also been developed in order to quickly dereplicate complex extracts and distinguish noise and nuisance compounds from new molecules.[15-18] Finally, targeted approaches have expanded the coverage and information content of tandem mass spectrometry-based approaches to secondary metabolomics.[12] Unfortunately, while these techniques are well suited to the discovery of new chemical scaffolds, they are unable to describe the function or biological activities of the compounds they identify. Therefore, there is still a need for new approaches to systematically identify novel bioactive scaffolds from complex mixtures.

In order to overcome some of these outstanding challenges we developed a platform called CAM that integrates phenotypic screening information from our cytological profiling assay with untargeted metabolomics data on the extract library to directly predict the identities and biological activities of all bioactive compounds in a given extract library directly from complex mixtures. By correlating individual mass signals with specific phenotypes from the high-content cell-based screen the CAM algorithm allows the prediction of the identities and modes of action of these biologically active molecules, which in turn provides a mechanism for the rational selection of lead compounds for further development based on biological and/or chemical properties. Chapter 2 of this thesis described in detail the first generation of this platform and the pilot study on a well-characterized subset of the Linington Lab library. To evaluate the utility of this platform for natural products discovery we examined a 234-member NP extract library, from which we derived 58,032 biological measurements and 10,977 mass spectral features (Figure 3-1). Using an updated

version of the Platform for the integration of these data led to the creation of a Compound Activity Map for this library comprised of 13 clusters containing 16 compounds from 11 compound classes, as well as the discovery of four new compounds, quinocinnolinomycins A - D, which are first example of microbial natural products containing the unusual cinnoline core.

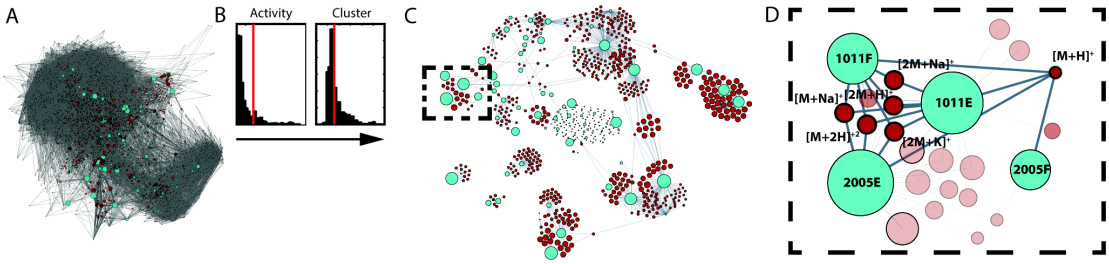


Figure 3-1: Illustration of how CAM enables discovery. (A) Network of extracts (light blue) connected by edges to *m/z* features (red) detected in the extract. (B) Scaled down histograms of the Activity and Cluster Scores for all *m/z* features with cutoffs indicated as red lines. (C) Network displaying *m/z* features associated with consistent bioactivity. (D) Zoom in of the staurosporine cluster with extracts and relevant adducts labeled.

This chapter will describe the impetus for Panama compound library, provide a brief description of the rational for colony selection, and discuss the improvements to CAM which simplify the integration and analysis, remove any data augmentation that are introduced by "vector compression", and change the network visualization strategy. Importantly, several molecules like staurosporine, rifamycin, echinomycin, and bafilomycin A1 that were used for the validation of the original platform are present in this second library and are easily identified illustrating that the general strategy for sample processing, data acquisition, analysis, and integration yield predictable and reproducible results.

3.2. Generating the Linington Panama Compound Library:

3.2.1. Background:

Over four years other lab members went to Panama in order to collect sediment samples from marine environments around the coasts of Panama and share the results of screening campaigns with collaborators in the Panama International Cooperative Biodiversity Group. This collaborative effort between many United States (U.S.) universities, U.S. companies, and Panamanian institution works to discover new anticancer, antiparasite, immunomodulatory, neuromodulatory and agrochemical lead compounds as well as provide technology transfer to Panamanian institutions. The U.S. National Institutes of Health and the U.S. National Science Foundation fund ICBG programs in many scientifically developing countries around the world. The Linington Lab contributed to this work by producing compound libraries of Actinomycetales bacteria extracts from Panamanian Marine sediment samples.

At the onset, Bailey J. Dickey worked as an undergraduate volunteer to develop the first 96-well plate containing 52 prefractions of 13 bacterial extracts. DOW Agro and EISAI Pharmaceuticals evaluated these prefractions for biological activity. The extracts were also screened in the relatively green cytological profiling assay. From this Bailey identified several anthracycline producing strains as potent cytotoxins. RLPA-1002D had interesting activity in CP as well as in EISAI's undisclosed assay and was therefore prioritized for analysis. A peak library[7] was

performed and the lead molecule was identified as lobophorin B identified by Fenical *et al.*[19]

3.2.2. Isolation and Growth:

It is important to discuss the rational behind the selection of colonies because the strains chosen to make the library determine the diversity and number of secondary metabolites and strain libraries picked by different people have distinct chemical profiles. In order to produce an extract library with high numbers of secondary metabolites, strains were isolated from media specific for Actinomycetales. Actinomycetales is an order of bacteria associated with soil and marine sediment that contain the biomedically import genera *Salinospora, Micromonospora, Nicardia, Streptomyces, Saccharopolyspora,* and *Frankia* responsible for producing many commercial antibiotics, as well as several pathogenic strains like *Mycobacterium tuberculosis, Corynebacterium diptheriae, and Propionibacterium acnes.* The G + C content of Actinomycetales genomes is often as high as 70% while the size and shape of the genomes can range from small circular genomes of 2 Mb to 10 Mb genomes with multiple linear chromosomes. Natural products chemists have studied this order extensively because many of the members dedicated up to 5 % of their genomes to secondary metabolite production with greater than 20 distinct biosynthetic gene clusters.

The different genera have distinct morphologies; however, the Actinomycetales order can be generally be identified by several traits that were used to select colonies: first, the presence of aerial hyphae; second, spore formation; third,

colonies that dig into the plate; fourth, zones of inhibition. One of the most distinct morphological features of Actinomycetales is the presence of small aerial hyphae that are visible under higher (40-100 X) magnification and give the colonies a fuzzy appearance at lower magnifications. A useful trick for determining if the colony is just out of focus is to slowly pan with the fine focus of the microscope. Fungi also have similar structures that are easily distinguished by the presence of distinct nodes connected by the hyphae. Besides hyphae, Actinomycetales often appear to dig into the plate rather than grow on the surface as *Escherichia coli*. Also, to the naked eye, many strains appear dry and textured on the surface rather than slimy. These topographical features are often accompanied by spore-formation that gives the colonies a chalky appearance after they have grown for several weeks.

Finally, my selections were heavily biased towards colonies that exhibited zones of inhibition. These colonies appeared as if they had halos around them on the isolation plate. This, hypothetically, indicates that the colony is producing some compound to prevent the growth of other bacteria and fungus that is diffusing through the agar. It is important to note that some colonies were selected based on the presence of a zone of inhibition alone.

3.3. Significantly Altered Methods:

This methods section will focus on experimental conditions that have changed significantly from the previous version of the platform because general methods for CAM were outlined thoroughly in Chapter 2.

3.3.1. Bioactivity Profiling:

3.3.1.1. Cytological Profile Screening:

Methods for cell culture and staining were used as previously reported.[7,10] HeLa cells were plated in two 384-well plates, compound plates were pinned into those cultures, and incubated for 24 hours at 37 °C. The plates were then fixed and stained with either nuclear or cytoskeletal stain sets. Finally, plates were imaged with a 10x objective lens with four images per well for each stain. Extract-treated wells were compared with DMSO-treated wells affording a 248-parameter fingerprint for each prefraction-treated well indicating the positive or negative perturbations in each parameter.

3.3.1.2. Death Dilutions:

Before submitting each screening plate for journaling, the raw imaging data were used to count the number of cells in each well. In some instances treatment of cells with prefractions resulted in significant cell death precluding the determination of accurate cytological profiles. The prefractions that caused a reduction in cell count outside of three standard deviations of control wells were submitted for serial dilution and rescreening. For prefractions that elicited a response with acceptable cell counts, the journaled cytological profiles were used for data integration and clustering. For the prefractions that caused a three standard deviation reduction in the number of cells, the cytological profile of the first dilution with a cell count within three

standard deviations of the mean control cell count was used for clustering and integration (Figure 3-2).

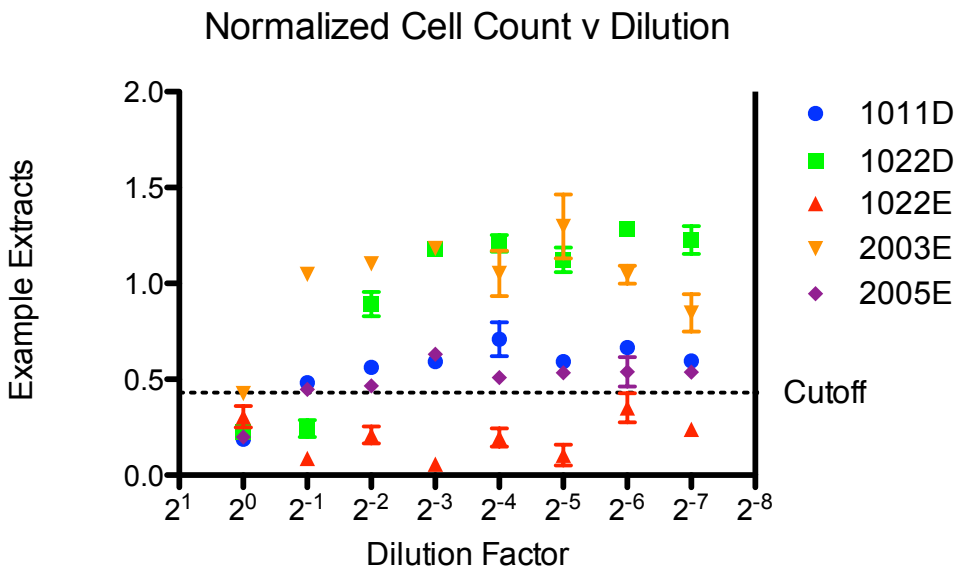


Figure 3-2: Graph of the plate normalized cell count plotted as a function of dilution factor with the absolute cutoff 0.43 plotted for several example extracts. Extract 1022E was diluted again such that the cell counts reached the acceptable levels before adding the cytological profiles to the plate data.

3.3.2. Chemical Profiling:

3.3.2.1. MS Data Alignment:

The *m/z* feature comparisons between samples were performed using high-resolution mass (ppm), retention time, and an isotope pattern matching method adapted from Pluskal *et al*.[20] We developed a decision tree to align *m/z* retention time (rt) pairs between extended dynamic range (2 GHz) and high-resolution (4 GHz) detector modes to select the most accurate data between 2 and 4 GHz modes from both positive and negative electrospray ionization (ESI) experiments (Figure 3-3).

After initial data acquisition, processing, and CEF file (peak list) output, MeOH blank and SYP media run peaks were aligned and removed using 20 ppm, 0.4 minutes, and 0.5 isotopic score difference windows from the sample prefraction run peak lists. Detector ringing from peaks above the detector saturation limit was removed by eliminating all peaks within 0.4 minutes and 1 mass unit of the most abundant peak in saturated data. Then peaks were aligned between 4 GHz and 2 GHz modes with 7 or 20 ppm, 0.4 minutes, 0.5 isotopic score difference windows. In order to store only the most accurate data, each aligned peak is assigned a tag that indicates whether or not the *m*/*z* feature is present and not saturated or present and saturated in both 4 GHz and 2 GHz modes (Figure 3-3). Priority is given to *m*/*z* values that are not saturated from the 4 GHz data. We store post-validated *m*/*z* features in a SQLite database for rapid indexing during incorporation with biological data.

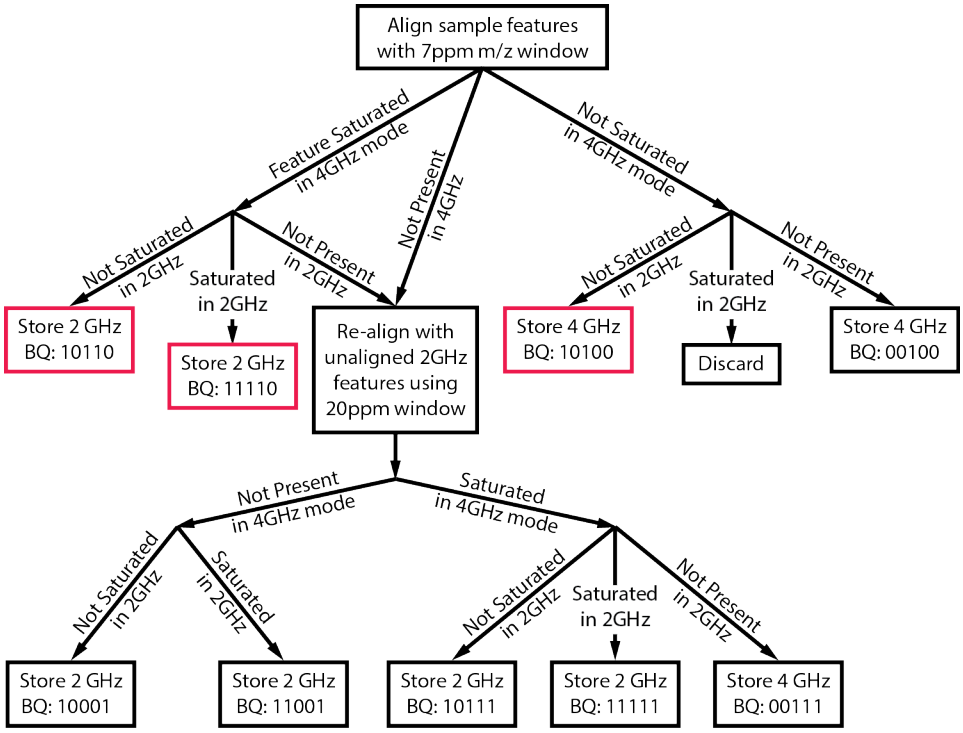


Figure 3-3: The decision tree for *m/z* feature alignment and scoring displaying how peaks are compared across mass spectrometry experiments of the same prefraction in the same ionization mode.

3.3.3. Data Integration:

In order to integrate the cytological profiling and metabolomics datasets, each *m/z* feature stored in the database is ascribed a synthetic fingerprint, an activity score, and a cluster score, which together predict the biological activity of each feature as described in Chapter 2.

3.3.3.1. Network Visualization:

We use NetworkX in python to create and edit networks and Gephi to visualize and analyze networked data. In general, blue nodes represent prefractions and are connected to red nodes representing the *m/z* features detected in those prefractions. Using Gephi, we visually rank nodes by size proportional to prefraction activity score or *m/z* feature activity score, and identify distinct clusters (represented by different colors) using network modularity with weighted edges and a resolution of one. We use Gephi's built-in Force Atlas 2 algorithm to distribute nodes two-dimensionally with default parameters except: approximate repulsion of 0.2, scaling of 10, gravity of 2, and prevent overlap (Figure 3-1 and Figure 3-7). This unbiased method was used rather than the vector compression display of just the prefractions because it is information rich and avoids altering the data in any way besides filtering biologically uninteresting molecules.

3.3.4. Fermentation and Isolation of Quinocinnolinomycins:

Bacteria frozen stock of strain RL11-047-HVF-B was struck out on solid media (DIFCO TM Marine Broth 37.4 g and 18.0 g of agar). Colonies were

inoculated into a capped 40 ml culture tube with 7.0 ml of liquid media containing 31.2 g of instant ocean, 10.0 g of Soluble Starch, 4.0 g of yeast extract, and 2.0 g of peptone per liter of water. All liquid media cultures were maintained at r.t. and shaken at 210 r.p.m. After 4 days 6 ml from the small-scale culture were used to inoculate 60.0 ml of the same media in a 250 ml wide neck Erlenmeyer flask with a 1 cm diameter metal spring coil and milk filter top. After 4 days 45 ml of this medium-scale culture were inoculated into 1 L of media in a 2.8 L wide mouth Fernbach flask containing a large spring coil and then topped with a milk filter. This culture was grown for 5 days.

The cells were then filtered using a glass filter, washed with sterile water, transferred to a 1.0 L Erlenmeyer flask, and extracted with 250 ml of 1 to 1 dichloromethane in methanol. The cell debris was filtered off and the extract solution was evaporated under vacuum. This dried extract was prefractionated using the eluotropic series of methanol water described in the methods section. The 80% methanol fraction was dried under vacuum and resuspended in minimal methanol and centrifuged. The supernatant was purified by HPLC on a (Phenomonex Kinetix 2.6 μm XB-C18 100 x 4.6 mm) using a gradient of MeCN:H_2O + 0.02% formic acid (50% MeCN for 2 min, 50%-65% MeCN over 20 min) at a flow rate of 2 ml min^{-1}. The peaks at minutes 8, 14, 15, 17 and 18 (diazaquinomycin C, and quinocinnolinomycin A-D respectively) were collected separately and dried under vacuum (**Figure 3-4**).

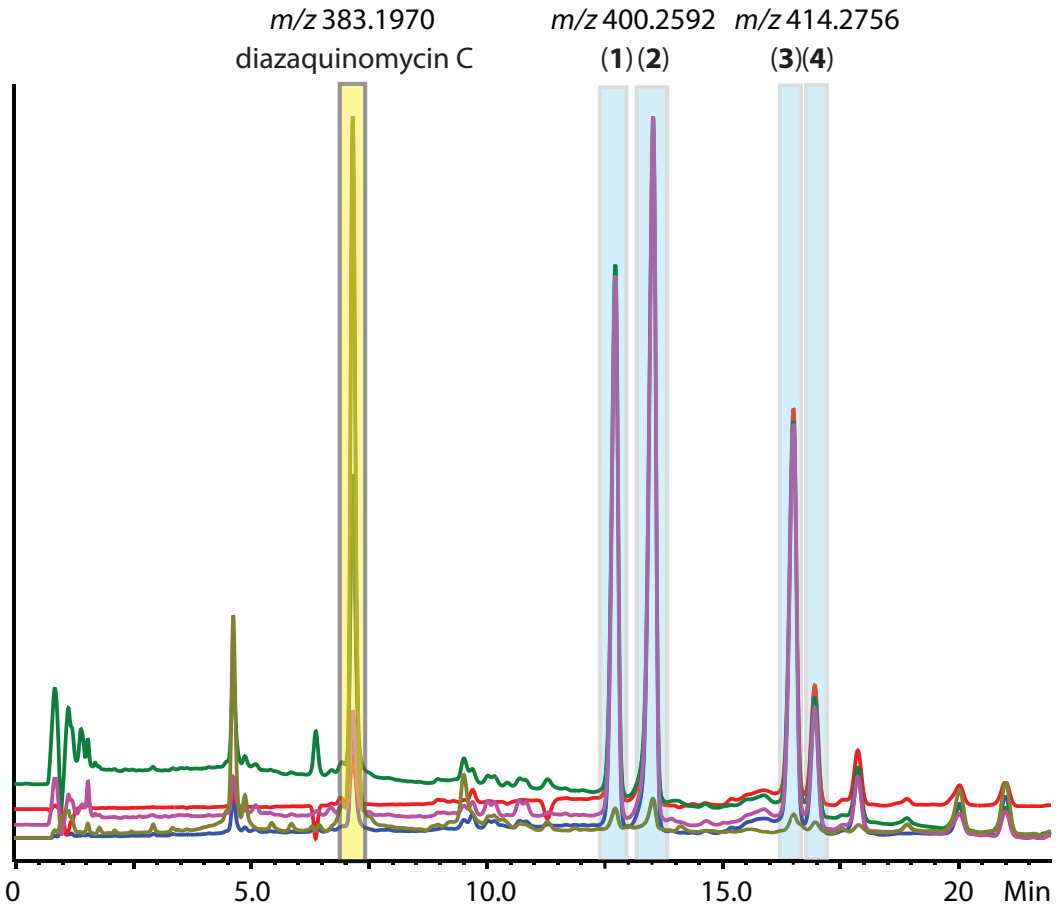


Figure 3-4: HPLC-DAD trace of RLPA-2003D with peaks labeled. Compounds **3.1-3.4** are labeled 1-4. The first two peaks are isomers that have the same *m/z* = 400.2592 and the second two peaks are isomers with the *m/z* = 414.2756.

3.3.5. Synthesis of (*S*) and (*R*)-MTPA Esters (5, 6) of Quinocinnolinomycin A:

In two vials with teflon septa pierced by a blunt needle containing a small stir bar, 1.0 mg of compound **3.1** was dried under vacuum. The vessel was flushed with argon and 0.200 mL of dry pyridine was added to each. The *R* and *S* MTPA-Cl were added in 6 fold excess, each to one of the vials, and the reaction was run for 1 hour at RT. The reaction was quenched by addition of a drop of methanol. The products were run through silica plugs, eluted with dichlormethane, and dried under vacuum. The subsequent products were purified by HPLC by (Eclipse XDB-C18 5μm 4.6 x 150 mm) reverse phase column on a gradient of using a gradient of MeOH:H_2O + 0.02% formic acid (60%-100% MeOH over 16 min) at a flow rate of 2 mL min^{-1}. The *S*-ester eluted at 11 minutes while the *R* ester eluted at 15 minutes.

3.4. Results:

3.4.1. Cytological Profiling:

Preliminary screening generated 234 prefraction cytological profiles of which 50 were serially diluted and rescreened based on low cell count (see methods). After these samples were diluted 57 of the 234 profiles had activity scores greater than 10, with 13 discrete clusters with Pearson correlations < 0.875 (**Figure 3-5**).

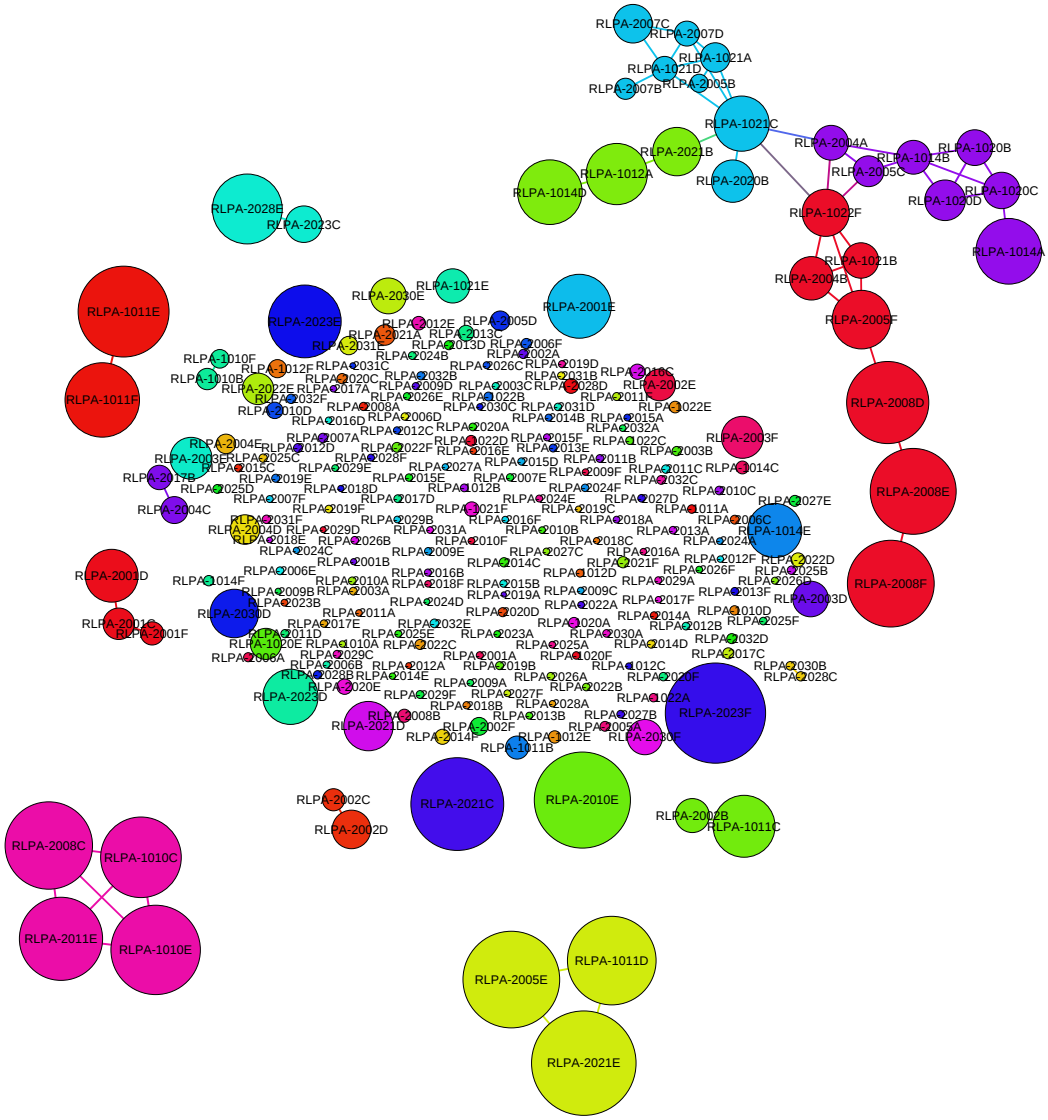


Figure 3-5: A network representation of the cytological profiling data. Prefractions that induce phenotypes with a Pearson correlation greater than 0.875 are connected and colored using Gephi's modularity package.

3.4.2. Metabolomics:

After media and blank subtraction, 10,977 features were stored into the mass feature sequel database. Of these, 346 were eliminated because they appeared in greater than ten percent of prefractions and 5310 singletons were removed, affording 5321 filtered features for network analysis.

3.4.3. Integration:

To integrate the biological and chemical datasets synthetic fingerprints, cluster scores, and activity scores were generated for each *m/z* feature in the database. An illustration of how these metrics are calculated for each *m/z*, retention time (rt) pair is shown in **Figure 3-6**. In this example, the phenotypes induced by extracts containing the $[M + Na]^+$ adduct of staurosporine are shown in an all-on-all matrix. Some of the phenotypes differ significantly. This is either due to differences in the concentration of staurosporine in the extracts, or the presence of other bioactives in the extracts. The activity score remains high because this metric only takes into account the magnitude of the fluctuations from the controls, while the cluster score is somewhat lower than expected. The average activity score is 4.66 with a standard deviation of 5.53 and the average cluster score is 0.13 with a standard deviation of 0.14. The $[M + H]^+$ adduct of staurosporine was used to define the cutoff for acceptable value of the activity score because it serves as a positive control and its activity score of 10.00 is approximately one standardization above the mean. The cluster score cutoff of 0.10 was defined by choosing a value above the median cluster score of 0.087. This value was chosen to account for differences in concentration and mixtures of metabolites

while still filtering a large percentage of the total features. Histograms and plots of all *m/z* feature activity scores and cluster scores are shown in **Figure 3-8**.

These results were used to generate activity plots for each prefraction, displaying *m/z* features as points on the graph, with the activity score on the y-axis and cluster score on the x-axis as seen in Figure 3-12. The color of each point corresponds to the retention time of that *m/z* feature. The activity and cluster score metrics were used to filter the *m/z* feature database to select for features that were correlated with strong and consistent phenotypes with activity scores greater than 10 and cluster scores greater than 0.10 respectively. After applying these filters 634 features remained that represented the *m/z* features predicted to be responsible for the observed bioactivites. A network was then generated from these 634 features in which extract nodes are connected to their corresponding *m/z* feature nodes by edges (Figure 3-7). The size of the node is defined by the activity score of the extract or half the activity score of the *m/z* feature for easy visualization. Clusters could then be assigned using the modularity feature of Gephi based on connectivities. From this we are able to observe 13 unique clusters, each of which contained the mass spectral features for the natural products predicted to be responsible for the bioactivity of the extracts (Figure 3-7).

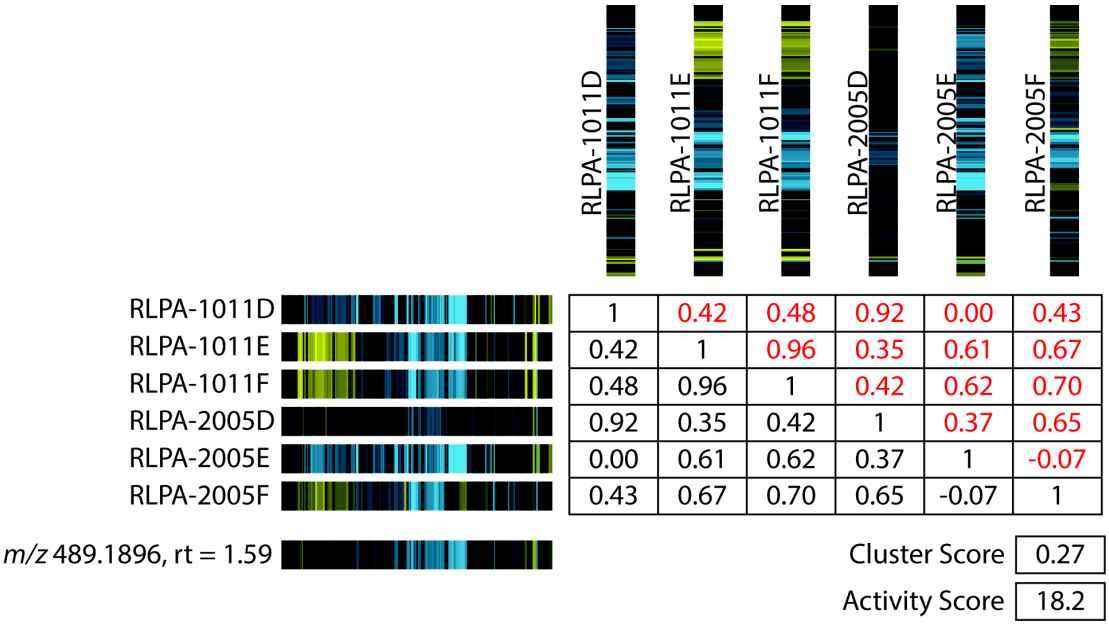


Figure 3-6: (above) Table of Pearson correlations of the cytological profiles for extracts in which the *m/z* feature (*m/z* = 489.1896, rt = 1.59) was detected. In each cytological profile, yellow stripes correspond to positive perturbations in the observed cytological attribute and blue stripes correspond to negatively perturbed attributes. (**below**) Calculated synthetic fingerprint, activity score, and cluster score of *m/z* = 489.1896.

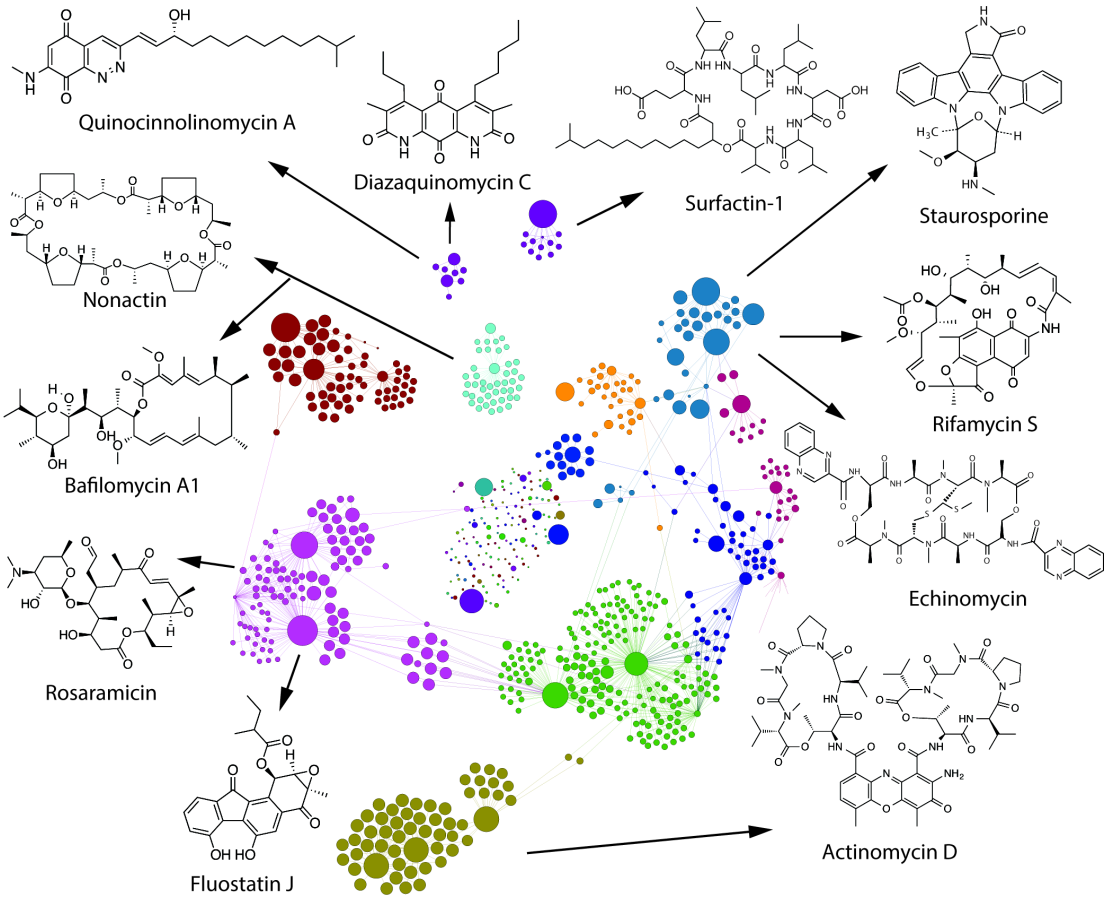


Figure 3-7: The same network from Figure 3-1 with the extracts and *m/z* features colored assigned by Gephi modularity function. Each cluster is annotated with a representative molecule from each of the confirmed families of compounds. *m/z* features with activity scores less than 10 and clusters scores less than 0.10 were removed from the network.

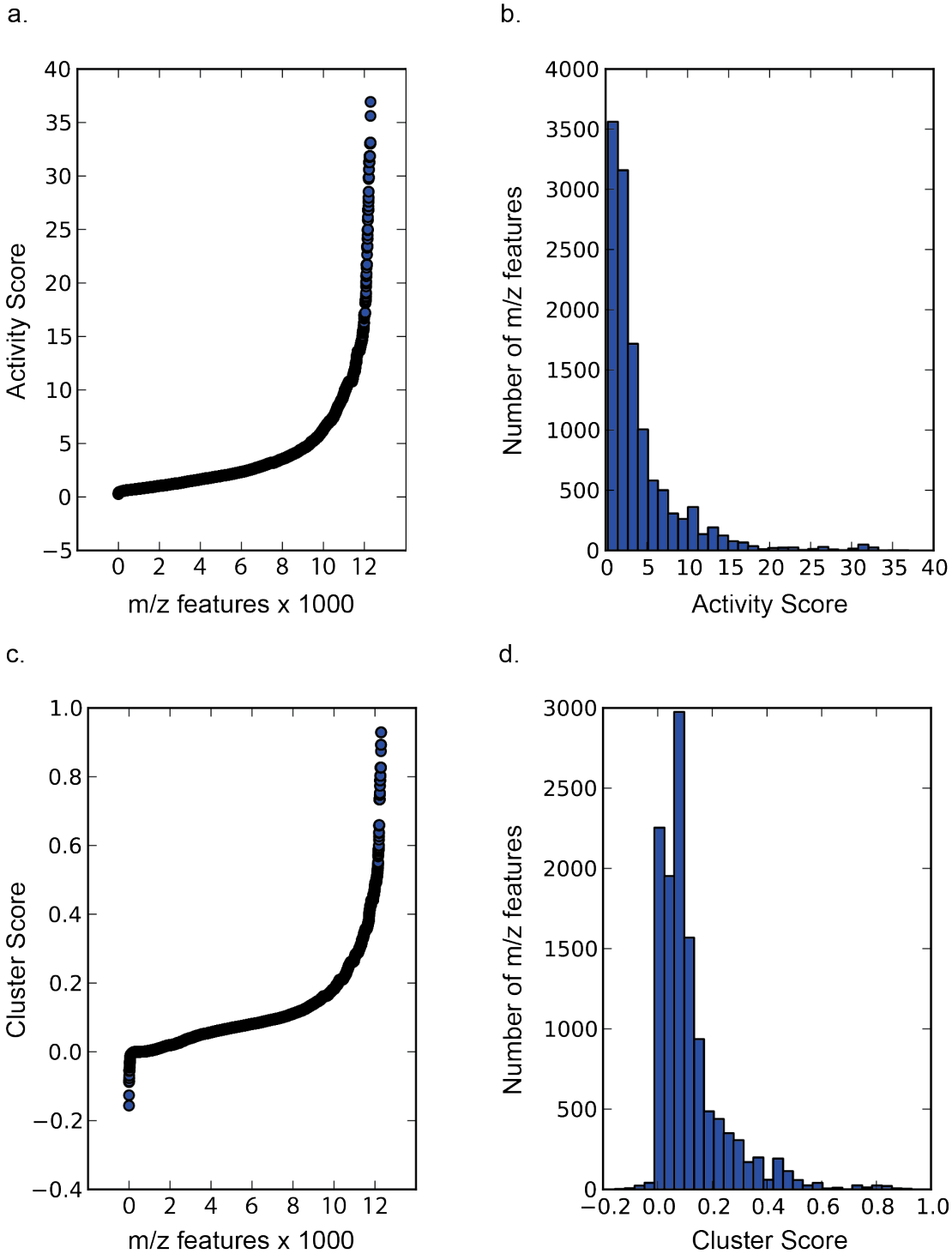


Figure 3-8: Graphs displaying the activity and cluster score values of each *m/z* feature. (a and c) depict activity and cluster score respectively versus feature count. (b and d) histograms of # of *m/z* features versus activity and cluster scores respectively.

3.5. Results and Discussion:

CAM provides a powerful new technology for the characterization of natural product libraries for bioactive compound discovery. Within the 234 extracts examined in this study, 57 have cytological profiles with activity scores above 10. All of these extracts possess associated *m/z* features from the metabolomic analysis that are predicted to be responsible for the observed activities, indicating that CAM can be successfully applied to the broad systematic characterization of complex screening libraries.

In general, these active clusters fall into one of three classes: clusters where the activity is caused by a single known natural product class; clusters where the activity is caused by the presence of multiple classes of known bioactives, and clusters where the activity is caused by bioactives that have no matches to available databases of microbially-derived natural products.

3.5.1. Clusters Containing Single Bioactives:

One example of a cluster driven by the presence of a single known bioactive class is the cluster containing extracts RLPA2008C, E and F (Figure 3-7). It is clear from the network that the chemical constitutions of RLPA2008C, E and F are distinct from the rest of the library. Three of the *m/z* features in this cluster were consistent with the $[M + H]^+$, $[M-H]^-$, and $[M+Na]^+$ adducts of a compound with the molecular formula $C_{62}H_{86}N_{12}O_{16}$. Searching the AntiMarin database (a comprehensive database of microbial and marine-derived natural products) identified actinomycin D as a match with the molecular formula. Clustering the synthetic fingerprints of these

features with the cytolofical profiles of the ENZO compound library strongly supported this result (**Figure 3-9**), with extracts RLPA2008C, E and F also clustering closely with the pure actinomycin D standard. The identification was confirmed by coinjection with a commercial standard of actinomycin D, which possessed the same *m/z* features and retention time as the predicted hits from the extract.

A second example of clustering driven by the presence of a single bioactive compound is the cluster containing extracts RLPA1011E, RLPA1011F, RLPA2005E and RLPA2005F (Figure 3-7). In this case the activity plot for RLPA1011F reveals seven *m/z* features consistent with the single molecular formula $C_{28}H_{26}N_4O_3$. Comparison of this formula to the AntiMarin database reveals a match to the pan-specific kinase inhibitor staurosporine. This assignment was confirmed by co-injection with an authentic standard of staurosporine, which had a retention time and HRMS signals that matched those for the bioactive components in these extracts.

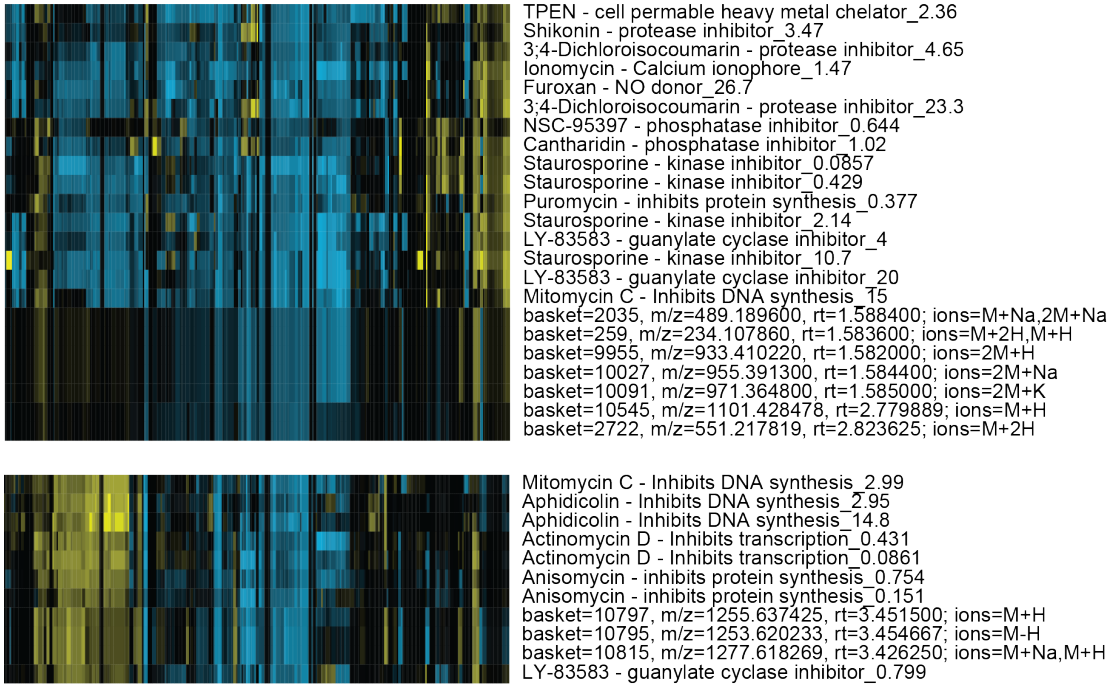


Figure 3-9: The cytological profiles of the ENZO compound library with synthetic fingerprints (the predicted cytological profiles of *m/z* features). The first cluster contains *m/z* features corresponding to the compounds staurosporine and echinomycin while the second contains *m/z* features corresponding to actinomycin D.

3.5.2. Clusters Containing Multiple Bioactives.

Although extracts 1011E and F were correctly predicted to contain staurosporine, examination of the Compound Activity Map and activity plots for extracts RLPA2005E and F revealed a second set of two *m/z* features predicted to contribute strongly to the observed biological activities of these extracts. These new *m/z* features were consistent with a compound with the molecular formula $C_{51}H_{64}N_{12}O_{12}S_2$, which corresponded to the DNA-intercalator echinomycin. Presence of this second bioactive metabolite was also confirmed by co-injection with a standard.

Importantly, although these three situations (staurosporine only, staurosporine and echinomycin) are all in one connected "super-cluster" because they are related by the extracts in which they are found, they resolve into individual sub-clusters based on the interconnectivities of the extract nodes and *m/z* features. This demonstrates that CAM can resolve even convoluted situations involving mixtures of compounds with fundamentally different biological mechanisms of action and provide useful characterization of bioactive metabolites even in situation where mixtures of bioactives cause phenotypic responses that are not closely related to either compound individually. The synthetic fingerprints of the *m/z* features corresponding to staurosporine cluster closely with the pure compound from the ENZO library and are readily distinguishable from those of the echinomycin (**Figure 3-9**).

A second example of clusters containing multiple bioactive metabolites is provided by the cluster containing extracts RLPA2021C, E and F. In this instance the cluster contains a large number of candidate *m/z* features, many of which are consistent with different members of two separate classes of natural products: the fluorenone-containing fluostatins, and the macrolide antibiotic rosaramicins (Figure 3-7). This situation is significantly more complex than the previous example, with multiple members of two separate bioactive compound classes contributing to the overall phenotypes observed for these extracts. Isolation and NMR evaluation of representative members of these two compound classes (fluostatin C,D, and J and Rosaramicin) confirmed their initial assignments, and permitted the evaluation of each compound class as pure compounds in the cytological profiling assay. The fluostatins all clustered closely with kinase inhibitors,[21,22] while rosaramicins induced only a very weak phenotype that is consistent with their previous annotation as antibiotics and not cytotoxic agents (**Figure 3-10** and **Figure 3-11**).[23] CAM was able to identify the fluostatins as the correct bioactive constituent, but because the fluostatins and the rosaramicins always appeared together, the macrolides were called as a false positive. This limitation of the platform can be resolved by analyzing larger libraries of extracts from similar organisms

because this will reduce the probability that two compounds will always be co-expressed. Once each constituent appears individually in the dataset, inactive compounds will display lower activity and cluster scores, eventually excluding them from the network.

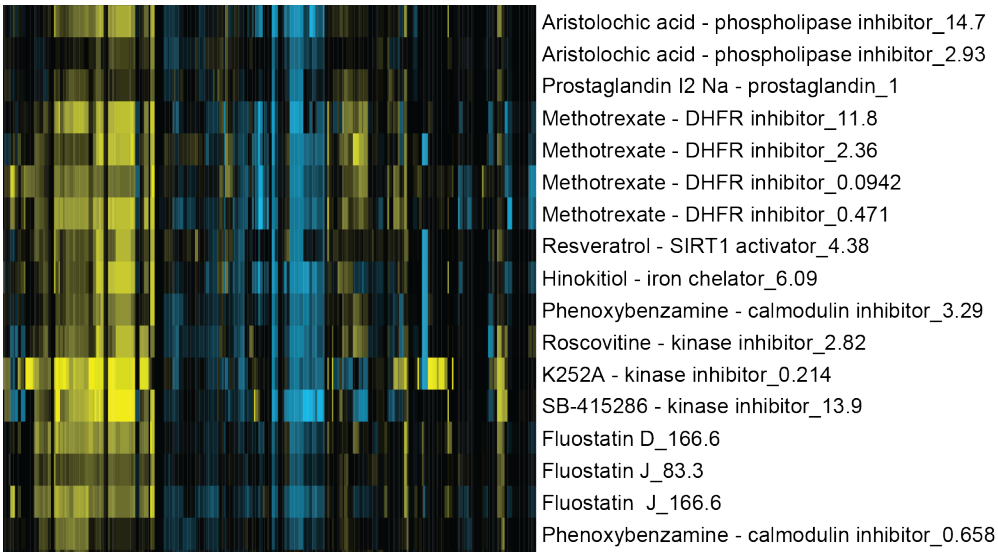


Figure 3-10: The cytological profiles of the ENZO compound library clustered with the purified fluostatins C, D, and J. The compound name is followed by the in well μM concentration.

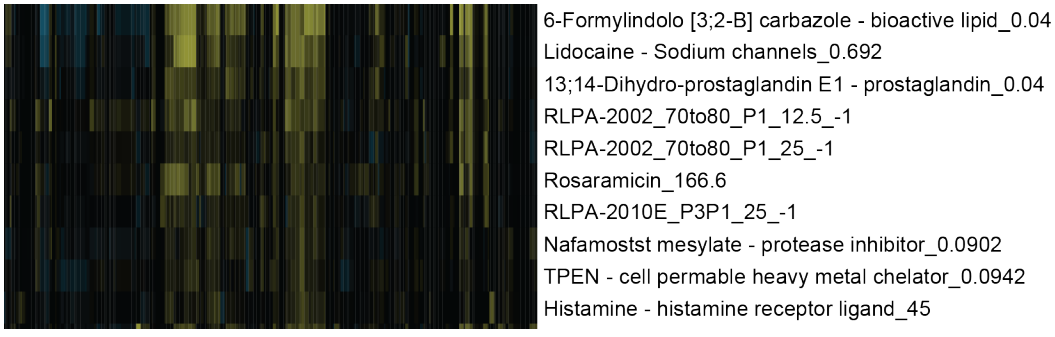


Figure 3-11: The cytological profiles of the ENZO compound library with the purified rosaramicin. The compound name is followed by the in well μM concentration.

3.5.3. Discovery of and Structure Determination of Quinocinnolinomycin:

In addition to the annotation of known bioactive metabolites, CAM is well suited to the discovery of novel compounds and the characterization of their biological attributes. Within this set of 234 extracts a number of unique clusters with high activity scores and no matches in the AntiMarin database were identified. We elected to explore one such cluster containing extracts RLPA2003E and F, because it contained just five *m/z* features that were common to both extracts. Examination of the activity plot highlighted one *m/z* feature with high activity and cluster scores (*m/z* = 400.2590, r.t. = 3.50 min, activity score 13.12, cluster score 0.57) that was prioritized for chemical analysis (Figure 3-12). LCMS analysis of this extract revealed the presence of two peaks with *m/z* features at 400.2590 amu and similar UV profiles, as well as two additional peaks that possessed the same UV profiles but had *m/z* values of 414.2756, suggestive of the presence of a family of related compounds.

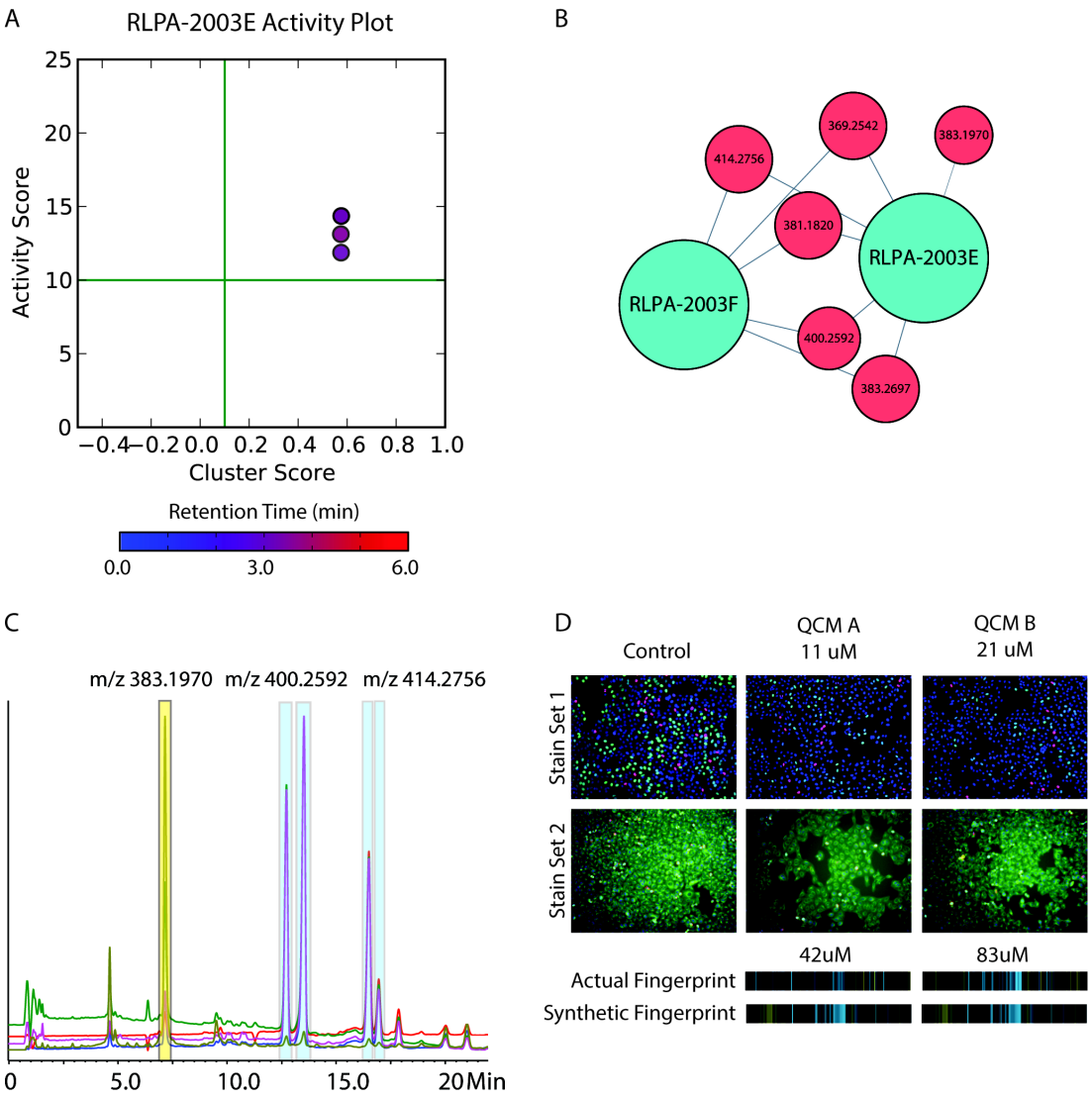


Figure 3-12: The prioritization, isolation, and confirmation of the quinocinnolinomycins A-D (3.1-3.4). (A) *m/z* features plotted on a graph of Activity Score and Cluster Score. The color of the dot corresponds to the retention time of the *m/z* feature with the color bar and scale below in minutes. (B) Isolated cluster from Figure 3-1 and Figure 3-7 with the extract labels and *m/z* for the *m/z* features. (C) HPLC trace of the RLPA-2003E and the isolation of 3.1-3.4 (Figure 3-13). (D) Cell images of pure compounds screened as a 2-fold dilution series for quinocinnolinomycins A and B in both stain sets compared to images of vehicle (DMSO) wells. (E) Comparison of the synthetic and actual cytological fingerprints of the pure compounds.

The molecular formulae $C_{23}H_{33}N_3O_3$ and $C_{24}H_{35}N_3O_3$ were determined based on the strong consensus between the $[M + H]^+$ and $[M + Na]^+$ *m/z* features for each set of two constitutional isomers. The earliest eluting compound with the molecular formula $C_{23}H_{33}N_3O_3$ was solved by NMR analysis, using a combination of ^1H, ^{13}C, gCOSY, gHSQC, gHMBC, and 1D-TOCSY spectra. Consideration of the ^1H-NMR spectrum indicated the presence of two vinylic and two aromatic signals, an aryl amine, one *N*-methyl doublet, two methyl doublets, and multiple overlapping resonances in a methylene envelope at 1.32 – 1.20 ppm. Interpretation of the ^{13}C and gHSQC spectra confirmed the presence of two ketones and four aromatic/quaternary carbons. The planar structure of the tail of the molecule was assigned from either side of the methylene region using gCOSY and HMBC correlations. 1D-TOCSY from the H11 proton to the H21 was used to confirm that the allyl oxymethine was connected to the tail through the methylene region.

The remaining $C_9N_3O_2$ belong to the headgroup. The attachment to the tail was assigned by gHMBC correlations from the vinylic protons at δ 7.01 and 7.22 to the carbons at positions 3 and 4 and gHMBC correlations from the aromatic proton on position 4 at δ 8.11 to the carbon at position 9. One of the ketones could be placed at position 5 based on gHMBC correlations from the aromatic resonance at δ 8.11. The resonance at δ 5.80 could be assigned to the vinylic proton at position 6 between the two ketones that form the quinone based on gHMBC correlations to the carbons at positions 4a, 5, 7, and 8. The gHMBC correlation from the *N*-methyl at δ

2.84 to the carbons at positions 6, 7, and 8 placed the relative orientation of the vinylic amine. Finally, the incorporation of the remaining atoms (CN_2) could only satisfy the requirement for three additional degrees of unsaturation by inclusion of the quaternary carbon at position 8a and two heteroaromatic nitrogens in positions 1 and 2, thus completing the structural assignment. The '*R*' stereochemistry of quinocinnolinomycin A was solved using Mosher's ester method (**Figure 3-15** and Table 3-1), and this assignment extended to quinocinnolinomycins B-D based on their common biosynthetic origin and the consistency in the sign of the circular dichroism spectra (Figure 3-16).

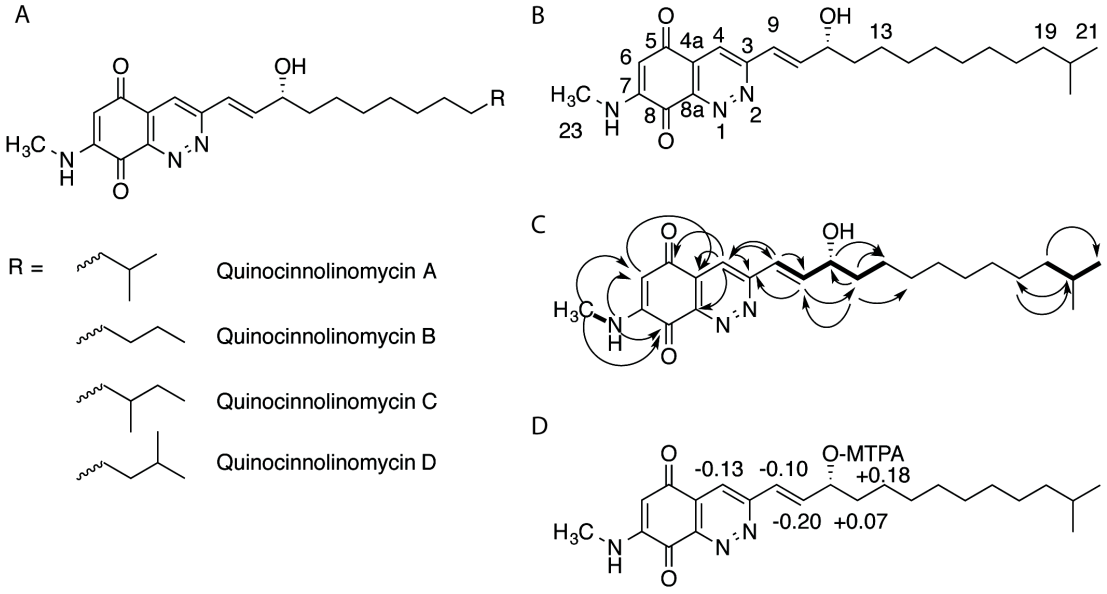


Figure 3-13: Structure elucidation of quinocinnolinomycins A-D (3.1-3.4). (A) The core and tails of quinocinnolinomycins A-D are displayed in order. (B) The structure of quinocinnolinomyin A (3.1) is displayed. The positions are numbered based on the cinnoline core. (D) $\Delta\delta^{SR}$ values for the MTPA ester analysis of the secondary alcohol to assign the absolute configuration.

Position	(1) δ_C	(1) δ_H(J in Hz)	(2) δ_C	(2) δ_H(J in Hz)	(3) δ_C	(3) δ_H(J in Hz)	(4) δ_C	(4) δ_H(J in Hz)
3	161.0		161.0		161.0		161.0	
4	118.2	8.11, s	118.4	8.11, s	118.4	8.11, s	118.4	8.12, s
4a	131.3		131.3		131.3		131.3	
5	178.1		178.1		178.0		178.1	
6	100.0	5.8, s	100.0	5.78, s	100.0	5.80, s	100.0	5.80, s
7	150.8		150.8		150.8		150.8	
8	178.0		178.0		178.1		178.1	
8a	147.0		147.0		147.0			
9	124.5	7.01, dd (16.0, 1.4)	124.7	7.01, dd (16.0, 1.5)	124.7	7.01, dd (16.0, 1.6)	124.7	7.01, dd (16.0, 1.6)
10	145.7	7.22, dd (16.0, 4.9)	145.8	7.22, dd (16.0, 4.9)	145.8	7.22, dd (16.0, 5.0)	145.8	7.22, dd (16.0, 5.0)
11	69.8	4.27, m	70.2	4.271, m	70.2	4.27, m	70.2	4.27, m
12	36.7	1.57, m	36.7	1.57, m	36.7	1.57, m	36.7	1.57, m
		1.51, m		1.51, m		1.51, m		1.51, m
13	25.0	1.37, m	25.0	1.38, m	24.9	1.4, m	24.9	1.38, m
14	29.3-29.1	1.30-1.22, m	29.2-29.0	1.32-1.20, m	29.4-29.0	1.33-1.19, m	29.4-29.0	1.37-1.22, m
15	29.3-29.1	1.30-1.22, m	29.2-29.0	1.32-1.20, m	29.4-29.0	1.33-1.19, m	29.4-29.0	1.37-1.22, m
16	29.3-29.1	1.30-1.22, m	29.2-29.0	1.32-1.20, m	29.4-29.0	1.33-1.19, m	29.4-29.0	1.37-1.22, m
17	29.3-29.1	1.30-1.22, m	29.2-29.0	1.32-1.20, m	29.4-29.0	1.33-1.19, m	29.4-29.0	1.37-1.22, m
18	26.8	1.22, m	29.2-29.0	1.32-1.20, m	29.4-29.0	1.33-1.19, m	29.4-29.0	1.37-1.22, m
19	38.5	1.12, m	29.2-29.0	1.32-1.20, m	36.0	1.25, m	26.8	1.22, m
						1.05, m		
20	27.4	1.48, m	29.2-29.0	1.32-1.20, m	33.7	1.27, m	38.5	1.12, dt (13.7, 6.4)
21	22.1	0.83, d (6.6)	31.3	1.32-1.20, m	28.9	1.33-1.19, m	27.4	1.47, m
						1.08, m		
22			14.0	0.84, t (7.0, 7.0)	11.2	0.82, t (7.4, 7.4)	22.5	0.83, m
23	29.1	2.84, d (4.9)	29.2	2.84, s	19.1	0.81, d (6.2)		
24					29.2	2.84, d (5.1)	29.2	2.84, d (5.0)
NH		8.17, q (4.7)		8.16, s		8.17, q (5.2)		8.16, q (4.8)
OH				5.12, d (4.9)		5.13, s		5.12, d (4.7)

Table 3-1/Figure 3-14: Tabulated NMR data from 3.1-3.4. All spectra were acquired in DMSO-d6 at 600 MHz and 150 MHz for ^{1}H and ^{13}C respectively. The structure of the core of the quinocinnolinomycins and each of the different tails displayed and numbered for clarity.

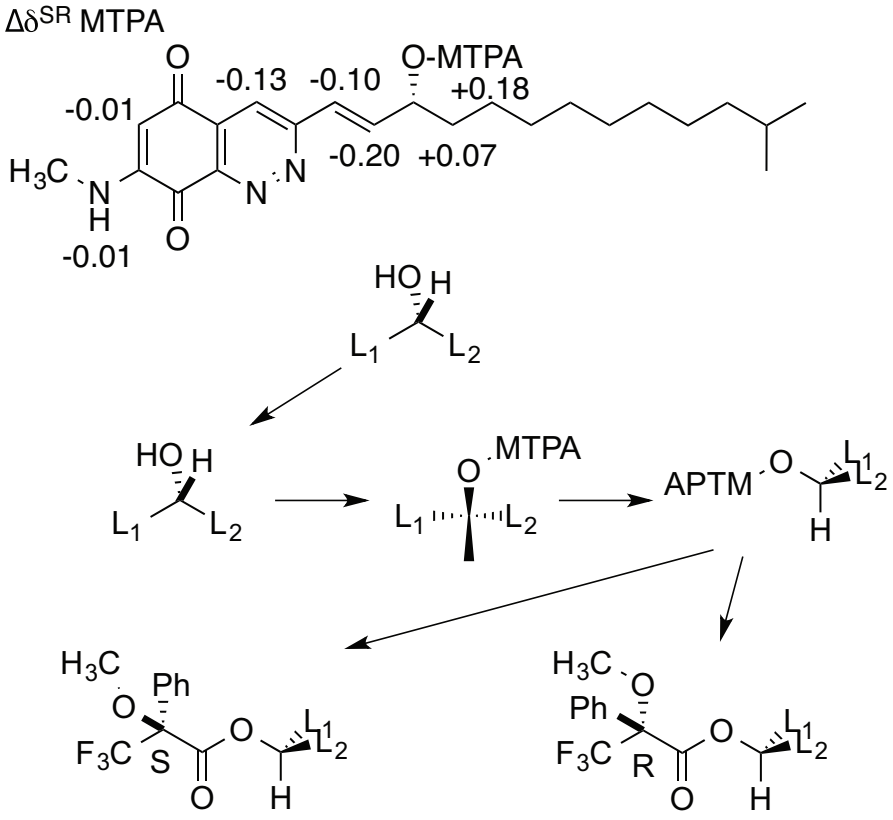


Figure 3-15: Quinocinnolinomycin A (**3.1**) is displayed with the $\Delta\delta^{SR}$ values for the modified Mosher's ester method. Shielding from in the phenyl ring in the suggested major comformer displayed below causes the affected protons to be shifted upfield for that particular diastereomer.

Position	(**3.1**) δ_H	S-Ester (**3.5**) δ_H	R-Ester (**3.6**) δ_H	$\Delta\delta^{SR} = \delta^S - \delta^R$
4	8.11	8.11	8.24	-0.13
4a				
5				
6	5.8	5.81	5.82	-0.01
7				
8				
8a				
9	7.01	6.91	7.12	-0.2
10	7.22	7.17	7.27	-0.1
11	4.27	5.77	5.77	0
12	1.54	1.84	1.77	0.07
13	1.37	1.37	1.19	0.18

Table 3-2: Tabulated 1H NMR data for (3.1, 3.5, and 3.6). All spectra were acquired in DMSO-d6 at 600 MHz.

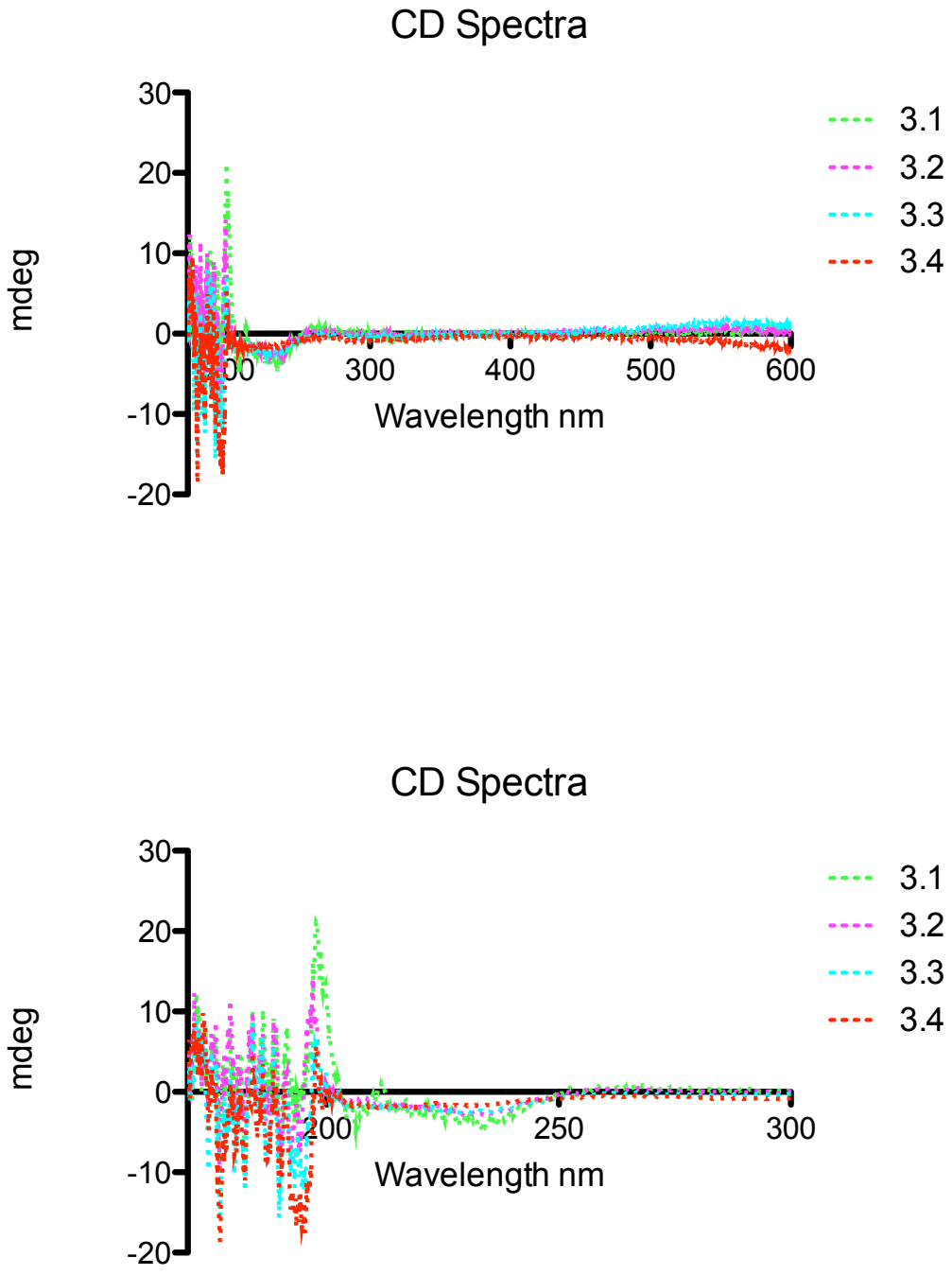


Figure 3-16: CD spectra for the four quinocinnolinomycins 3.1-3.4. All four analogues have the same sign in the range of light absorption indicating that the absolute configurations are the same.

3.5.4. Mechanism of Action of the Quinocinnolinomycins:

Purified quinocinnolinomycins A-D where rescreened as two-fold dilution series (166.7 μ M – 2.5 nM) in the cytological profiling assay (**Figure 3-17**). Clustering these cytological profiles with those of the ENZO compound library training set revealed a distinct cluster containing all four analogues over a range of concentrations between 0.3 and 83.3 μM along with the known compounds thapsigargin (calcium ATPase inhibitor),[24] tunicamycin (N-linked protein glycosylation inhibitor),[25] lycorine (ribosome inhibitor),[26] and brefeldin A (ARF GTPase inhibitor). [27] While the precise molecular targets of these compounds differ, they are all mechanistically related because they all affect the function of different components of the endoplasmic reticulum (ER) and result in ER stress and the induction of the protein unfolding response.[28-30] Active concentrations of quinocinnolinomycins A – D are present within this cluster with Pearson correlations to the other training set compounds on the order of 0.6 – 0.7, indicating very close matches between these cytological profiling fingerprints. These data strongly suggest that the quinocinnolinomycins have a mode of action that causes ER stress. Moderate ER stress may be mitigated by macroautophagy (autophagy) in mammalian cells and can lead to cell death or survival depending on the context, and is an active area of research for future cancer therapies.[29-32] Further studies to elucidate the precise molecular target of the quinocinnolinomycins will expand our understanding of the cellular processes involved with ER stress, the unfolded protein response, and autophagy with direct implications for human disease.

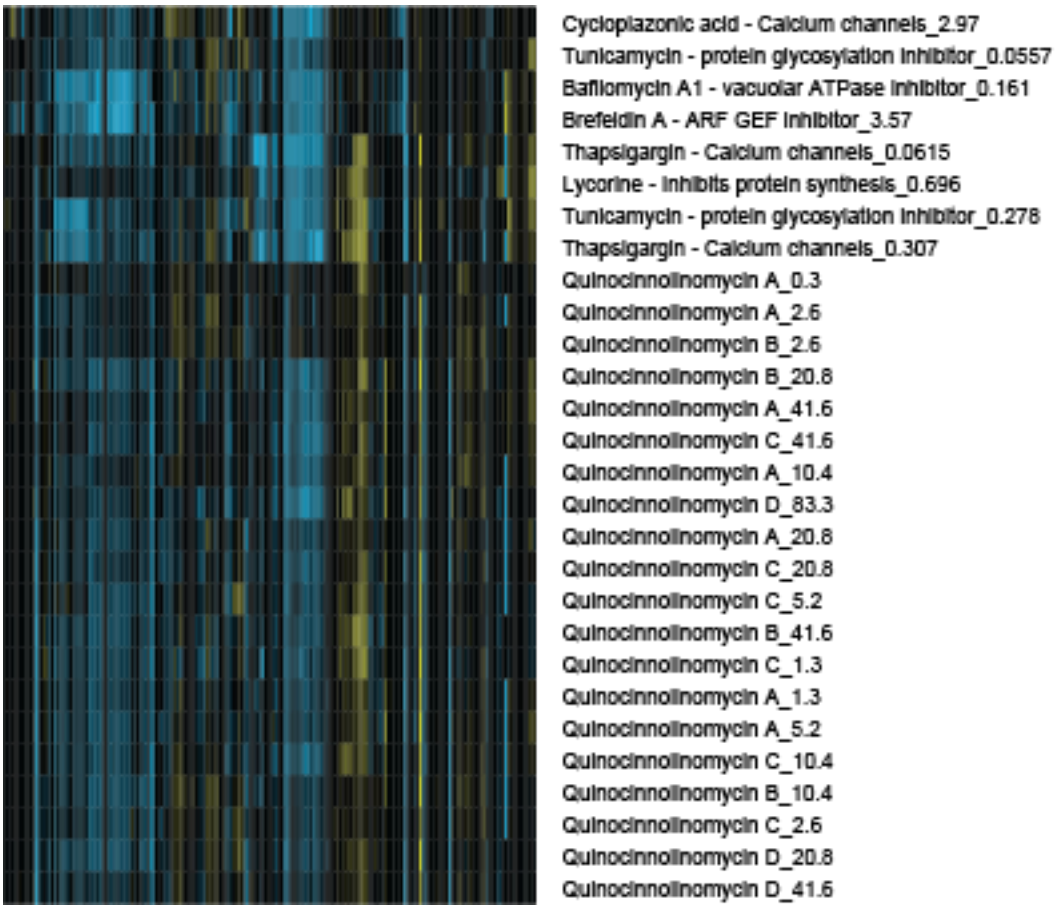


Figure 3-17: The cytological profiles of the ENZO compound library clustered with the purified quinocinnolinomycins (**3.1-3.4**) in a dilution series. The compound name is followed by the in well μM concentration. The strong similarity of the cytological fingerprints of the quinocinnolinomycins (**3.1-3.4**) with compounds known to cause endoplasmic reticulum stress (thapsigargin, tunicamycin, lycorine, and brefeldin A) suggest that (**3.1-3.4**) have a similar mechanism of action.

3.6. Limitations and Caveats of CAM:

3.6.1. The Acquisition of the MS Data is Too Slow:

Due to the length of the runs and the need to acquire data in both 2 GHz (high dynamic range) and 4 GHz (high resolution) modes in order to get accurate masses

over a large concentration range, each 96-well plate requires four cycles for a total of 56 hours of continuous data acquisition and for the instrument to be cleaned and calibrated four times. Running the instrument continuously, this timeline would be reasonable to generate spectra for the whole 6,000 member library; however, the constant use puts significant strain on the instrumentation, resulting in frequent interruptions and long delays. In the next chapter I will discuss improvements to the UPLC methods as well as the use of the waters Aquity UPLC SYNAPT MS to improve acquisition time, the need for replicates, frequent and high strain on the system.

3.6.2. Mass Spectrometry Is Not A Universal Detection Technique:

While ESI-TOF MS is able to detect the molecular ion for most organic compounds, there are some molecules that either do not ionize under the soft ionization conditions, or undergo in source transformations that are not easily predicted and complicate the assignment of a molecular formula. One known species missing from the raw data is the molecular ion for cycloprodigiosin. This compound was previously isolated from RLPA-2001C,D, and E as a hit from a peak library in the cytological profiling assay. The compound was later shown to have antibiotic activity in the vibrio cholera biofilm assay.[6] Also, the fractions containing cycloprodigiosin exhibit striking phenotypes in CP. The absence of this particular compound could be due to concentration, but at the dilution used in the assay, the pigment is visible to the naked eye.

3.6.3. Singletons:

One limitation of the platform, as we have implemented it, is that compounds that are only observed in one extract are discarded during the data integration. This is done because no significant correlations can be made between the activity and the presence of the compound if the compound only appears once in the library. Unfortunately this results in the elimination of many of the features that could be driving activity. In the future, investigating singletons from each active extract that does not share any other *m/z* features with any other extract should be a top priority.

3.6.4. Death Dilutions Lead to False Negative Loss of Biologically Active Compounds:

The death dilution series are absolutely necessary for the accurate prediction of the mode of action of compounds directly from the crude extracts; however, there are examples in which this process led to the elimination of biologically active compounds. Extracts RLPA-1012E and F and RLPA-1022E and F contain analogues of the rakacidin family of compounds. Rakacidin A and B were isolated and characterized by NMR from 1012E.[33,34] The pure compounds were screened in a 2-fold dilution series and exhibited activity, but the cell count curve showed a precipitous decline over a 4-fold dilution. Correspondingly the extract activity score dropped to 6.20 at a dilution with an acceptable total cell count. This activity is consistent with the previously reported activity of these compounds because they exhibit an 1,000-fold selectivity for hypoxic cancer cells and off target effects at high concentrations.[21,35]Performing the dilutions to reach a cell count greater than three

standard deviations from the mean control cell count caused the activity score for rakicidin A and B to drop to 2.36, below the filter threshold, and for the *m/z* feature to be removed from the network analysis.

This result both demonstrates that the death dilutions are effective for the elimination of cytological profiles of empty wells and observing the proper mechanism of action of compounds. In this case, the compounds should exhibit little to no effect at reasonable concentrations and the dilution process was effective at correcting for the overtly cytotoxic off target effects. Extracts exhibiting this dilution sensitive activity should be prioritized for screening in other assays that may reveal their true mechanism of action.

3.7. Conclusion:

By predicting the identity and mode of action of all detectable metabolites from complex extracts CAM aims to expedite the discovery process changing the traditional "blind" discovery model to a hypothesis-driven approach to novel bioactive compound discovery. CAM drastically reduces the time required to go from a hit in an assay to a lead molecule by minimizing iterative bioassay guided fractionation and screening steps, and allows hypothesis-driven exploration of NP libraries by providing a global view of compound diversity and activity across any library. In this study, analysis of the 234-member library revealed 13 unique clusters based on chemical and biological similarities. We were able to confirm the identities of 16 compounds from these clusters using a combination of analytical approaches,

providing a detailed molecular picture of the bioactivity landscape for this extract library in this biological assay.

The discovery of quinocinnolinomycins A - D highlights the utility of this platform for novel compound discovery and mode of action characterization. The cluster containing extracts RLPA2003E and F is distinct in the Compound Activity Map and contained *m/z* features suggesting the presence of unique compounds correlated with a strong and distinct phenotype. These data strongly suggested that these mass features should be prioritized for structure elucidation, leading to the discovery of this new structural class of natural products with accurately predefined biological activities.

3.8. NMR Data:

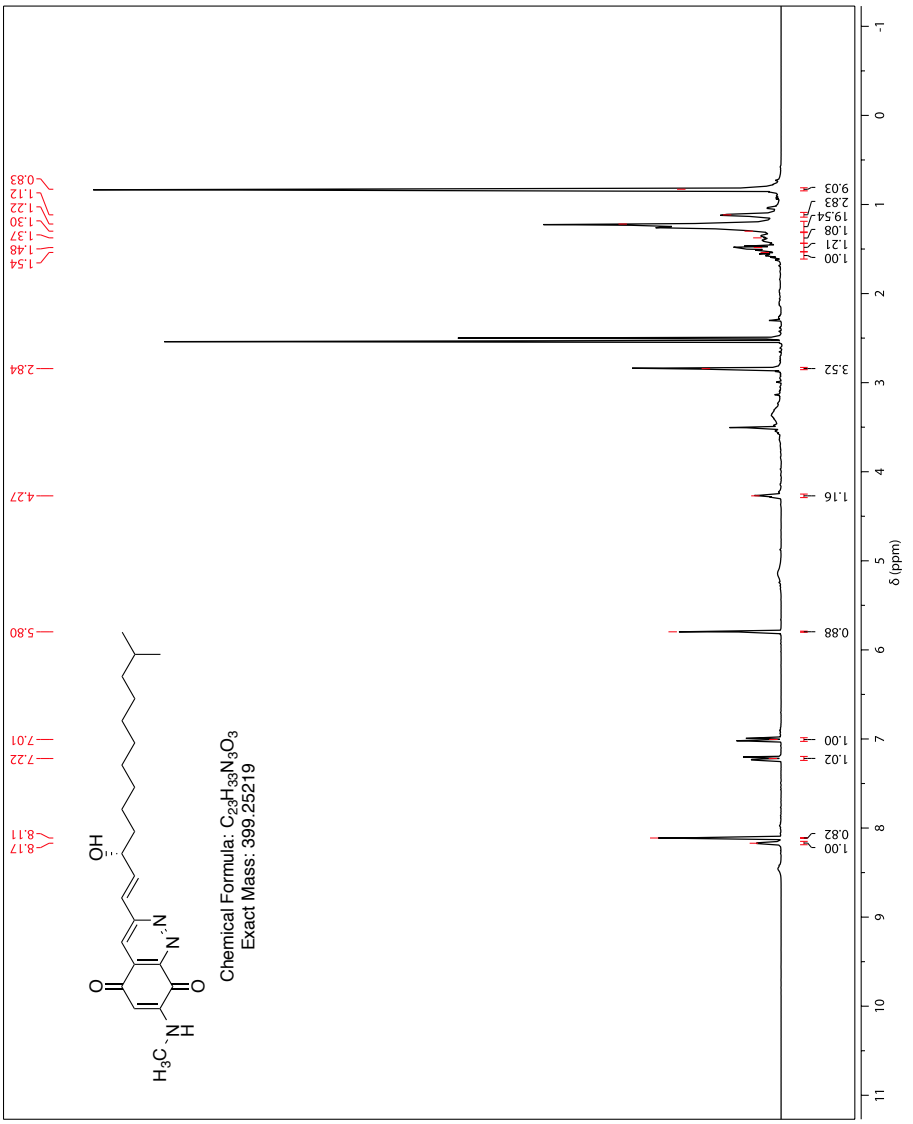


Figure 3-18: ^{1}H NMR of 3.1 in DMSO-d6.

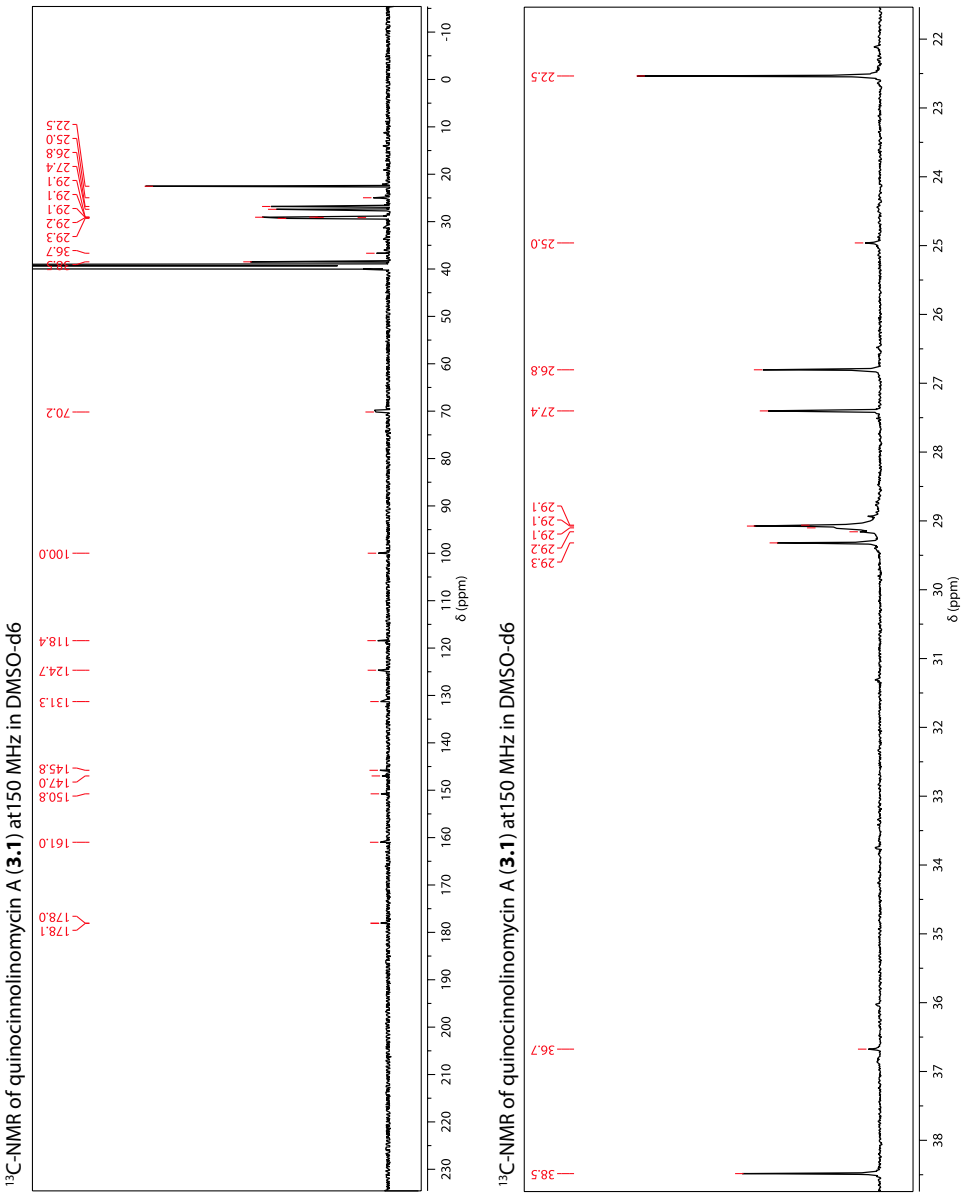


Figure 3-19: ^{13}C NMR of 3.1 in DMSO-d6.

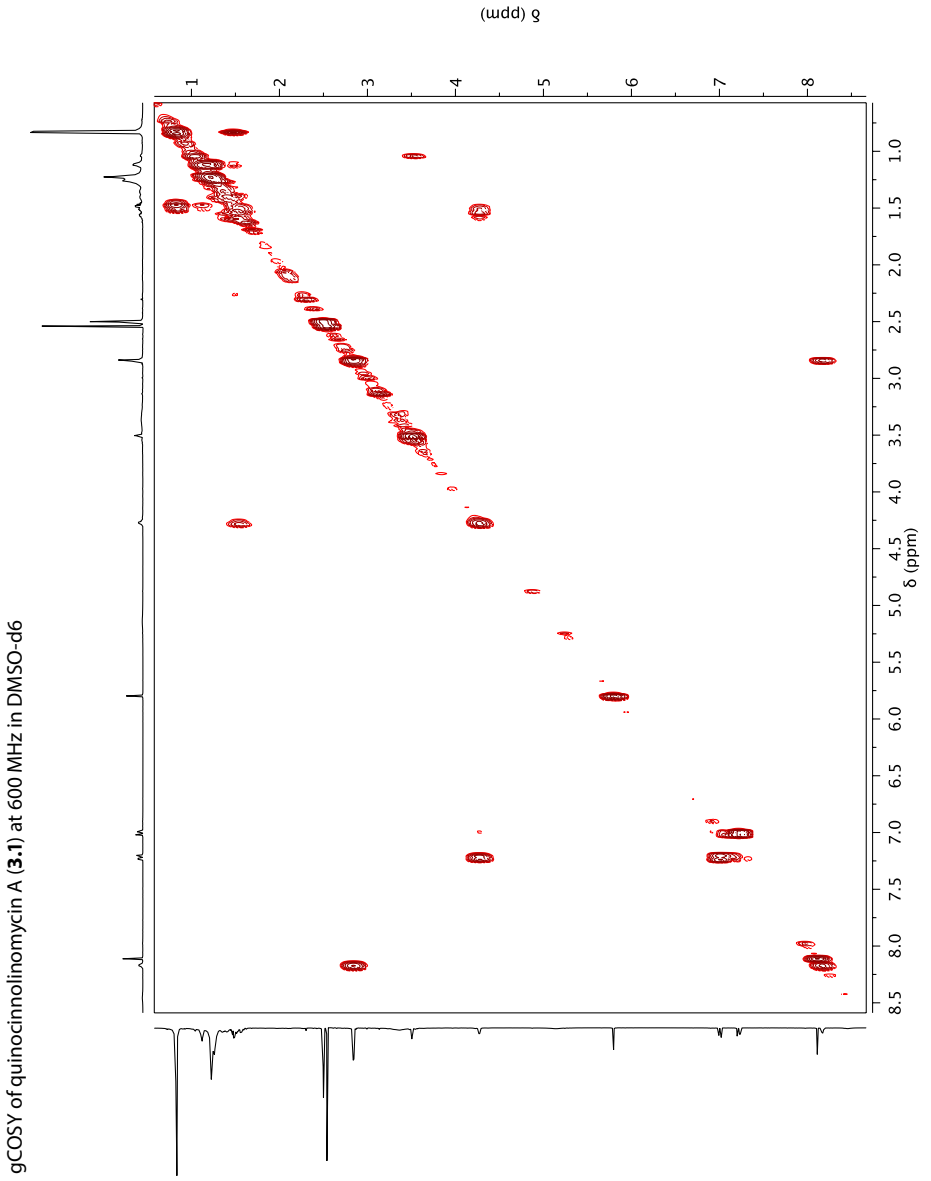


Figure 3-20: COSY of 3.1 in DMSO-d6.

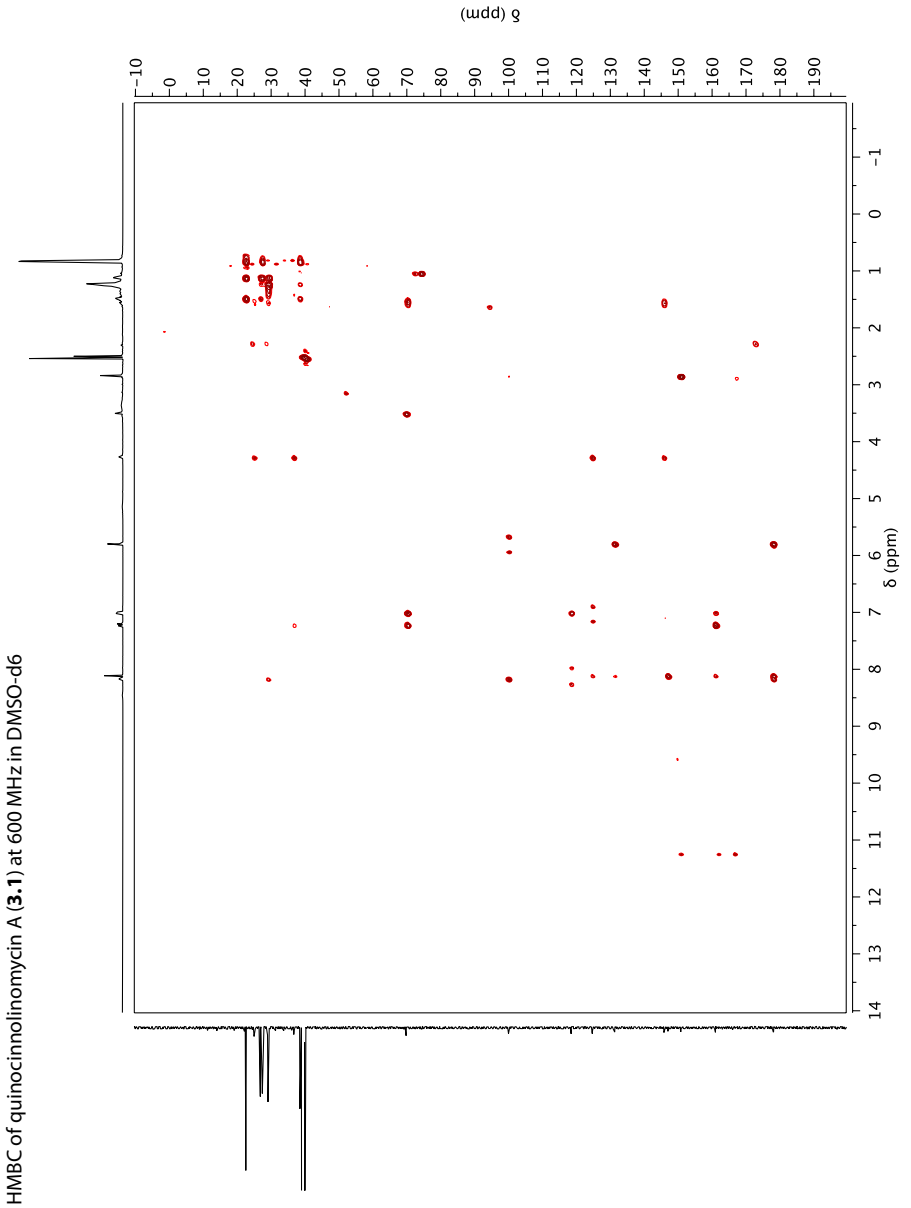


Figure 3-21: HMBC of 3.1 in DMSO-d6.

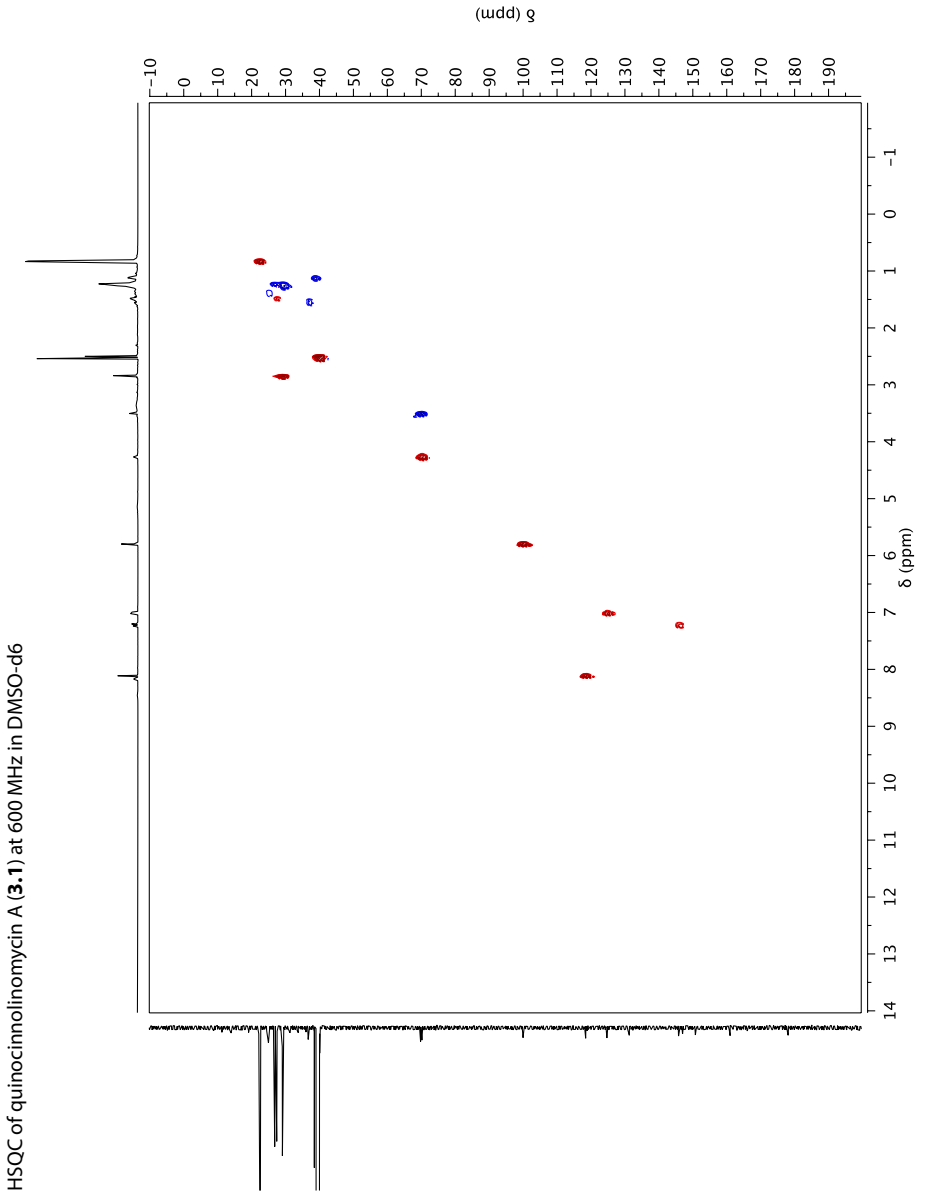


Figure 3-22: HSQC of 3.1 in DMSO-d6.

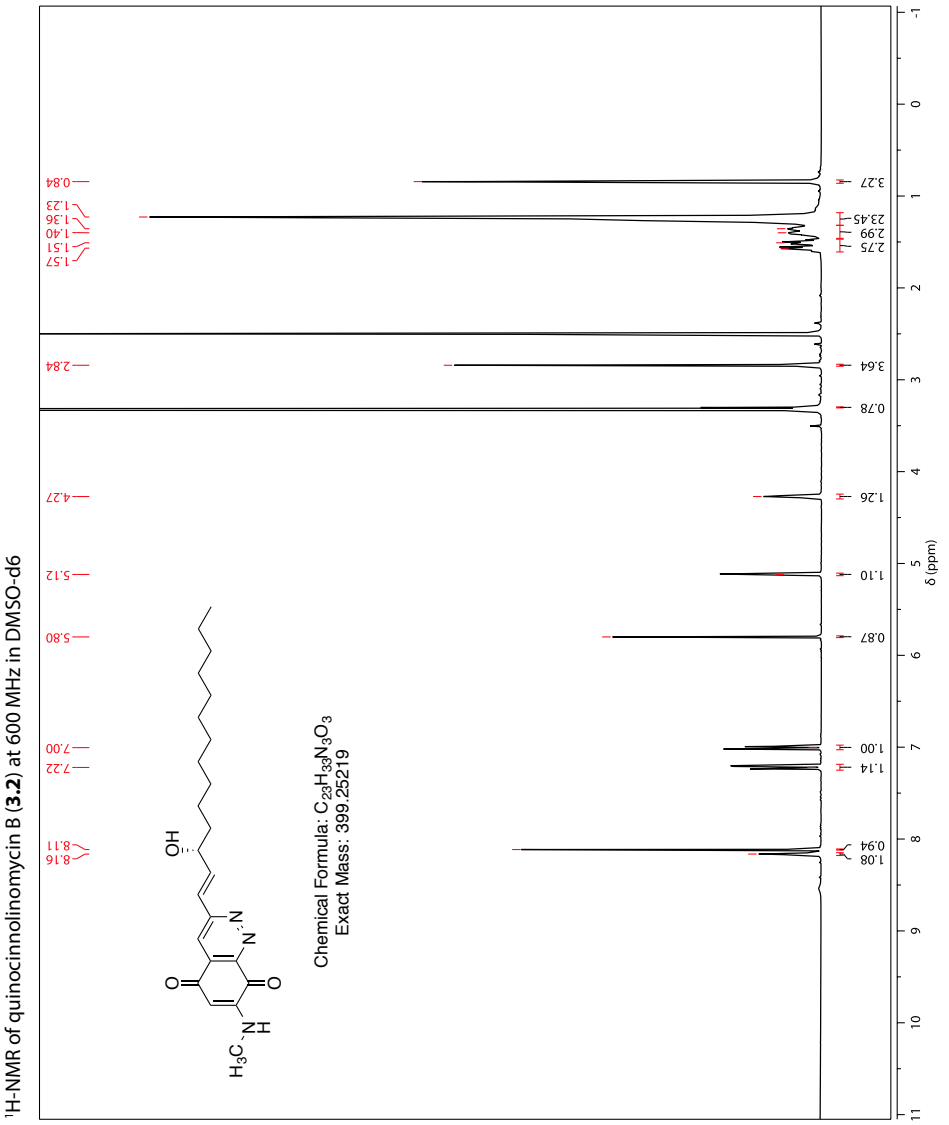


Figure 3-23: ^{1}H NMR of 3.2 in DMSO-d6.

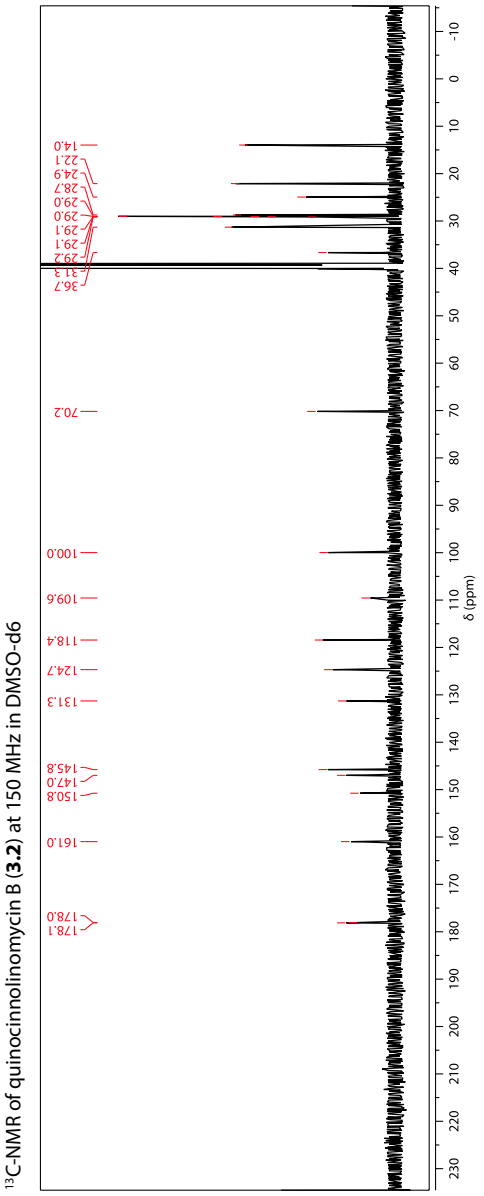


Figure 3-24: ^{13}C NMR of 3.2 in DMSO-d6.

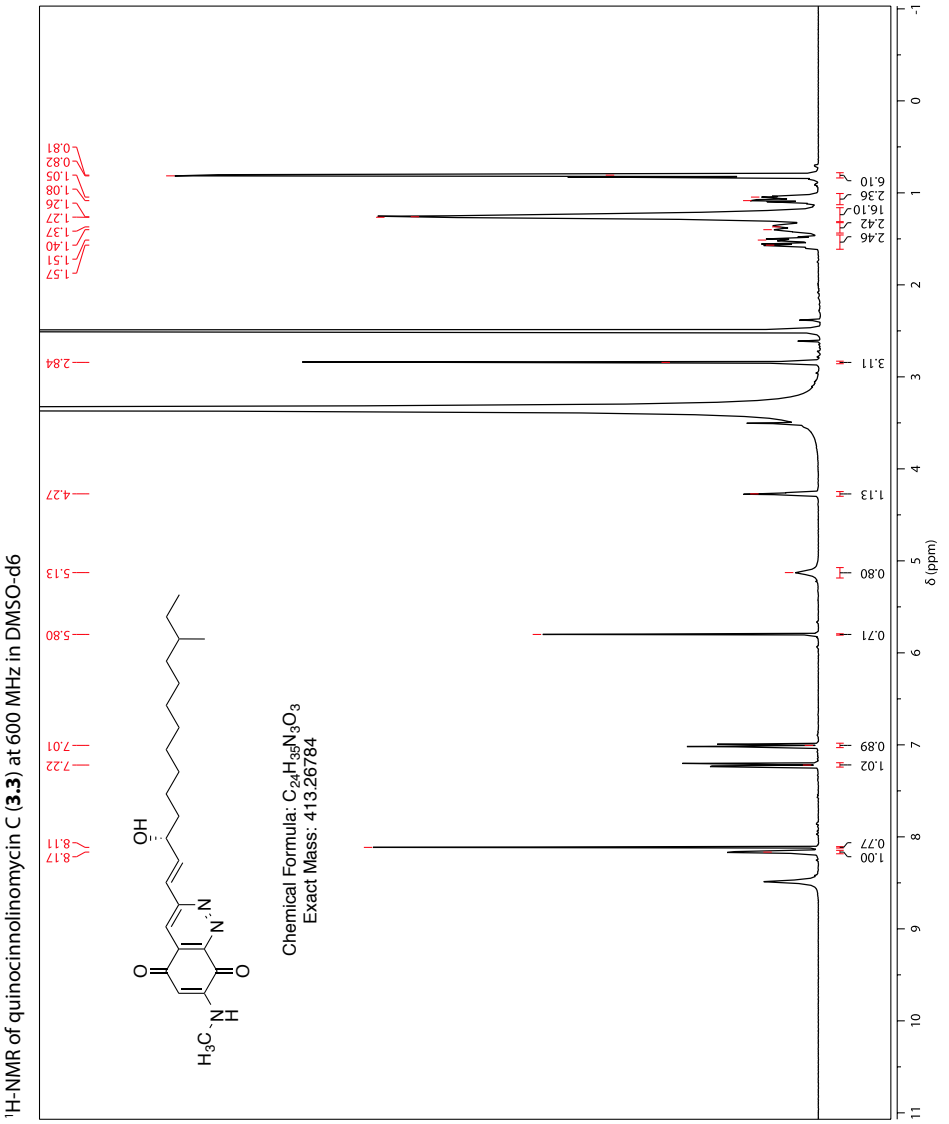


Figure 3-25: ^{1}H NMR of 3.3 in DMSO-d6.

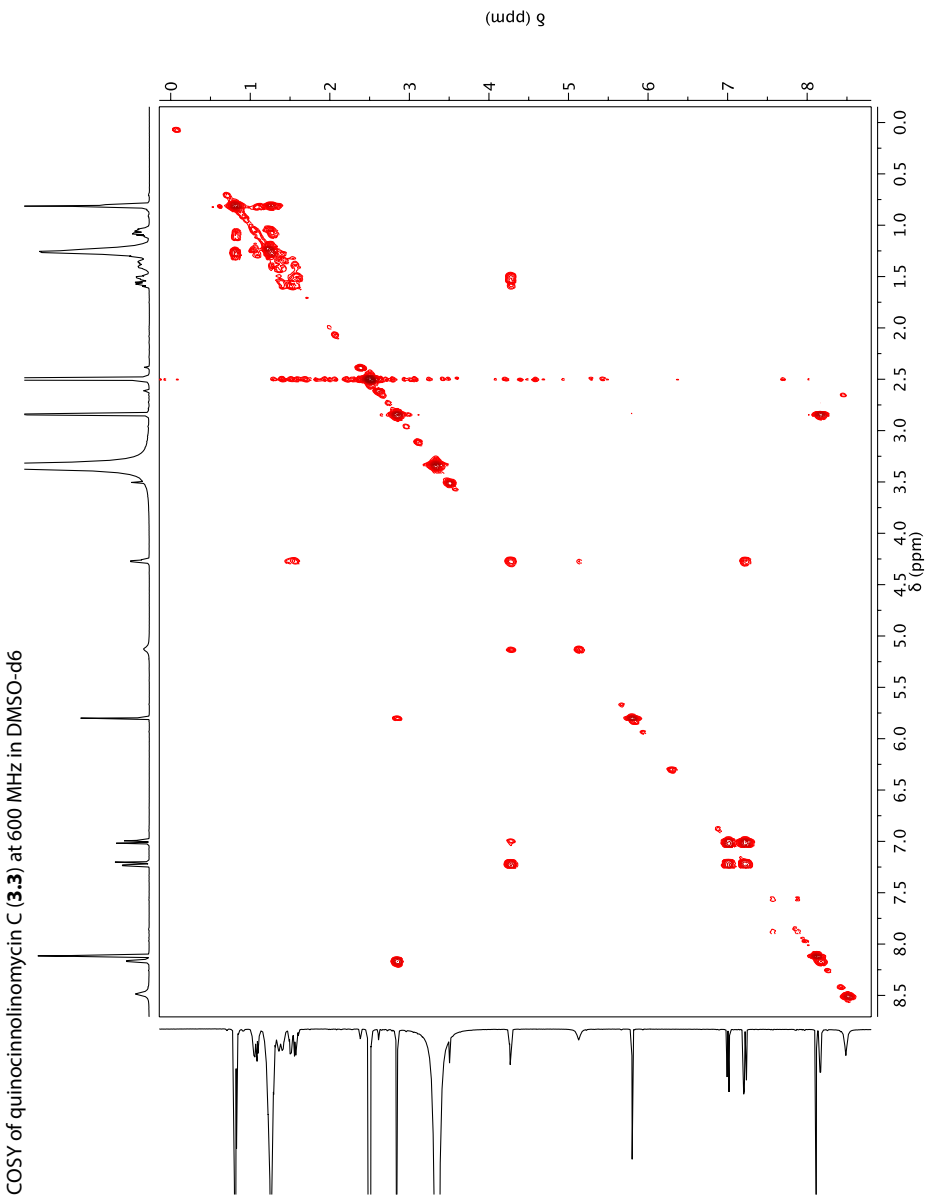


Figure 3-26: COSY of 3.3 in DMSO-d6.

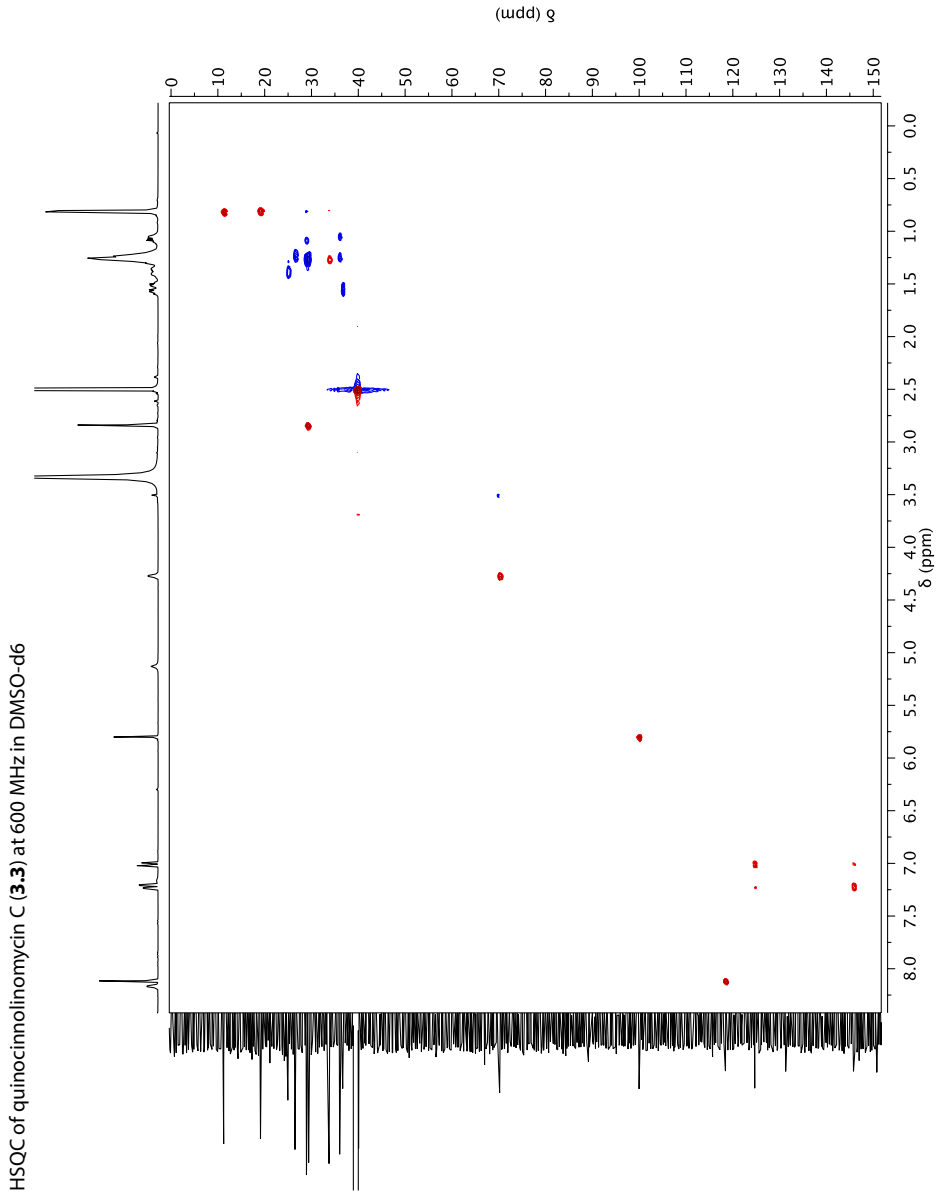


Figure 3-27: HSQC of 3.3 in DMSO-d6.

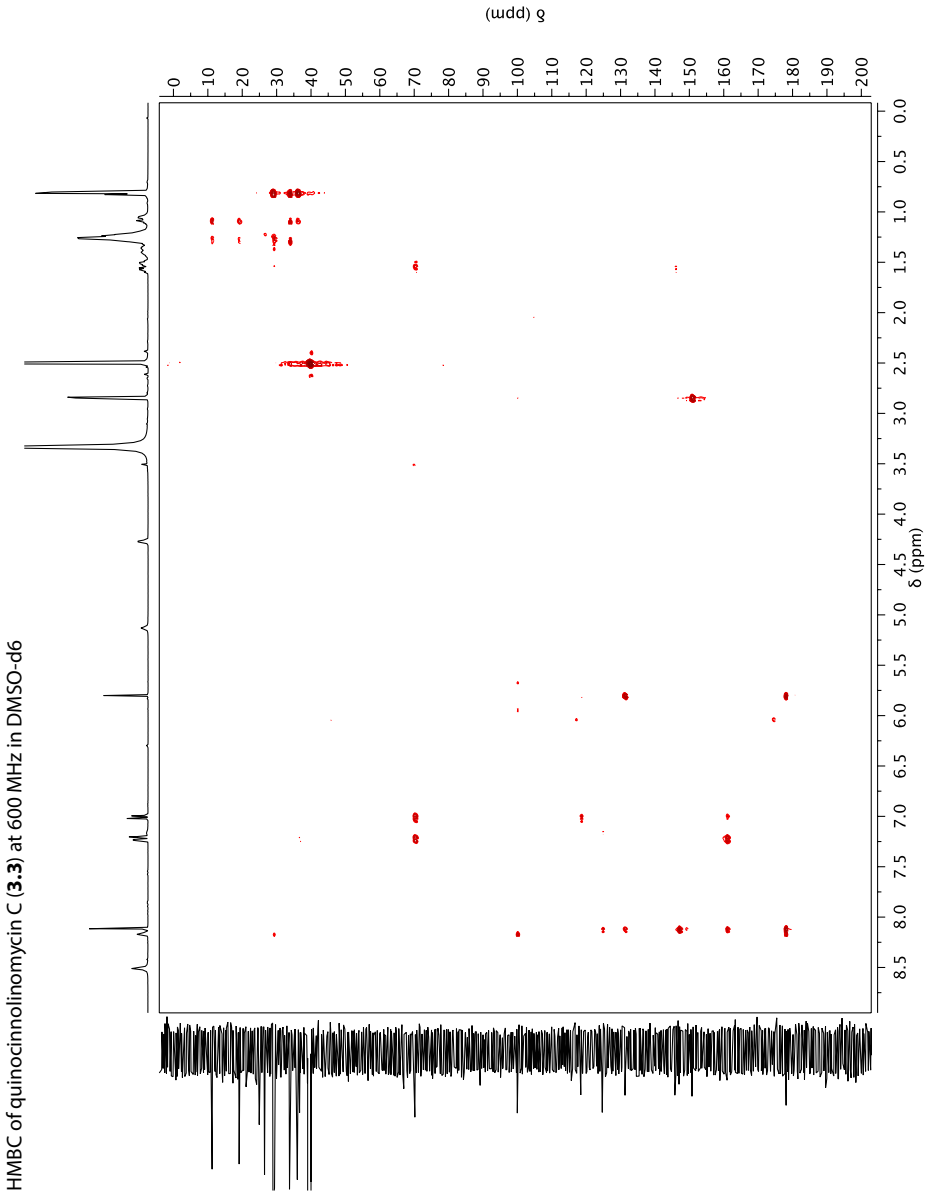


Figure 3-28: HMBC of 3.3 in DMSO-d6.

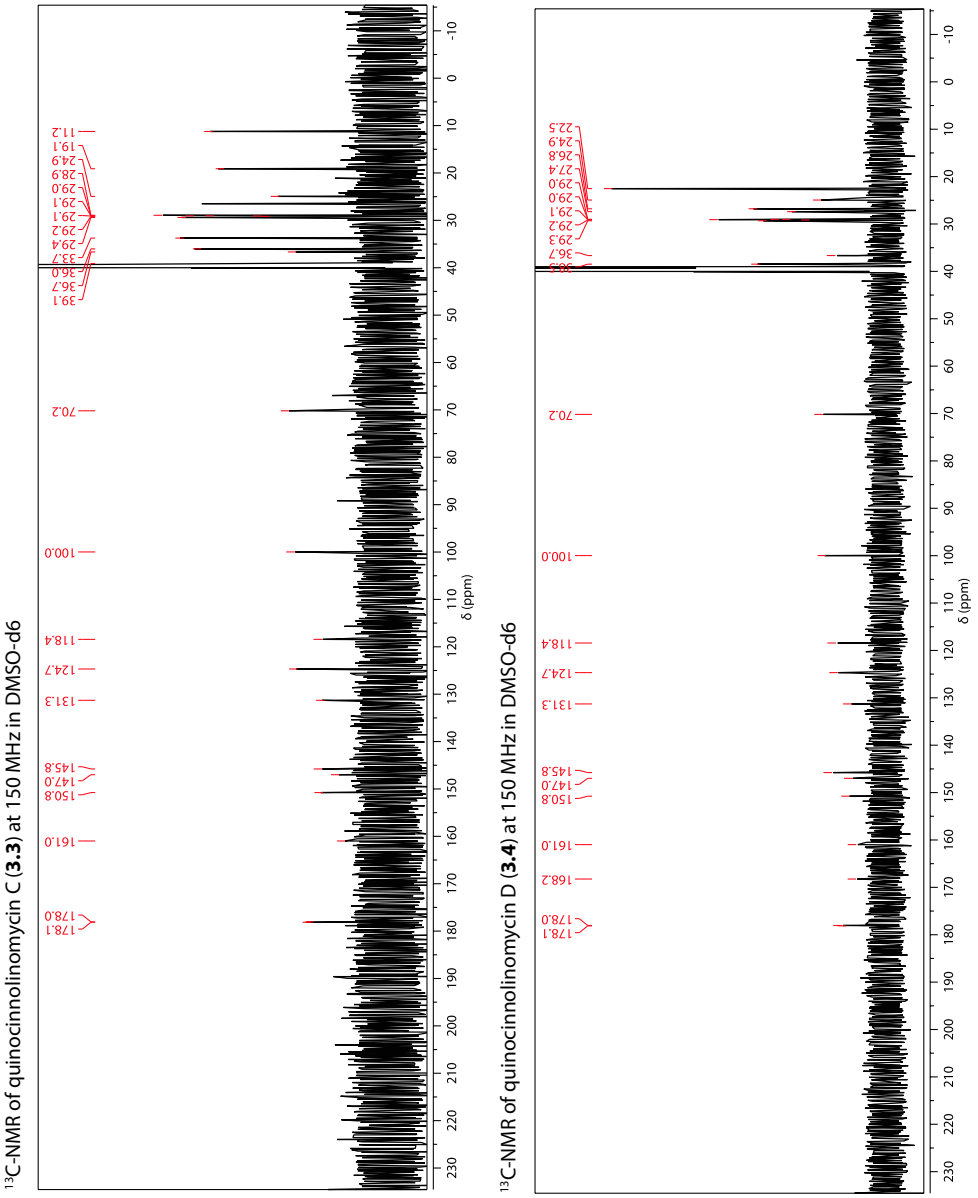


Figure 3-29: ^{13}C NMR of 3.3 and 3.4 in DMSO-d6.

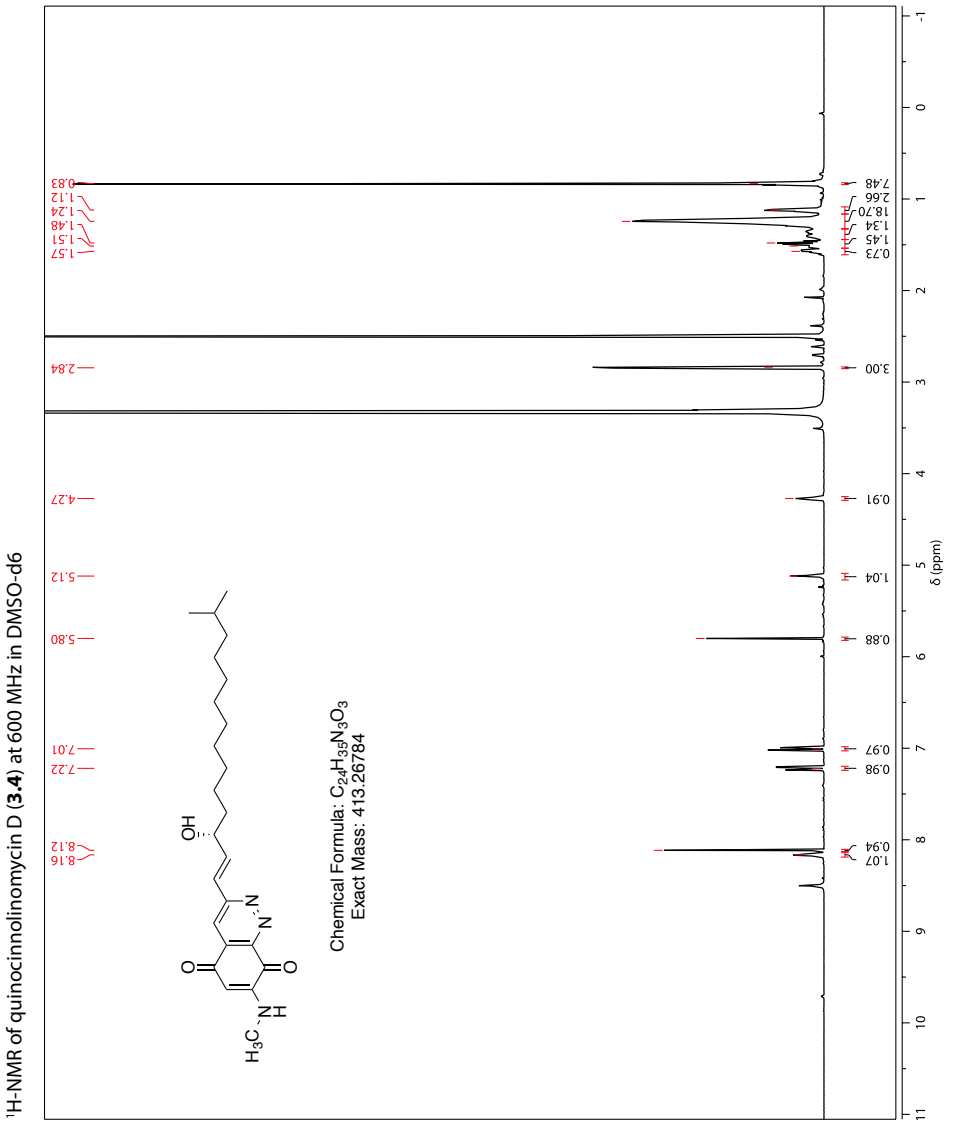


Figure 3-30: ^{1}H NMR of 3.4 in DMSO-d6.

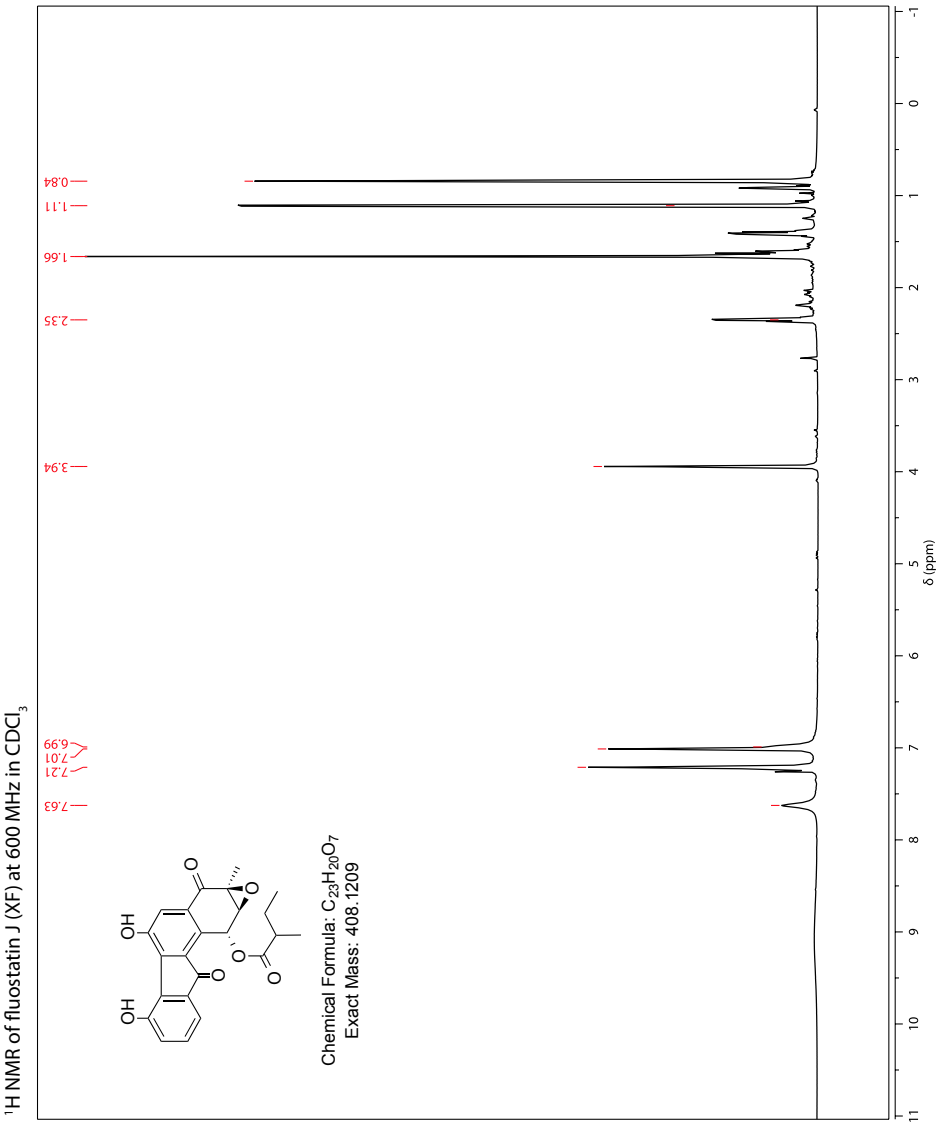


Figure 3-31: ^{1}H NMR of fluostatin J in $CDCl_3$.

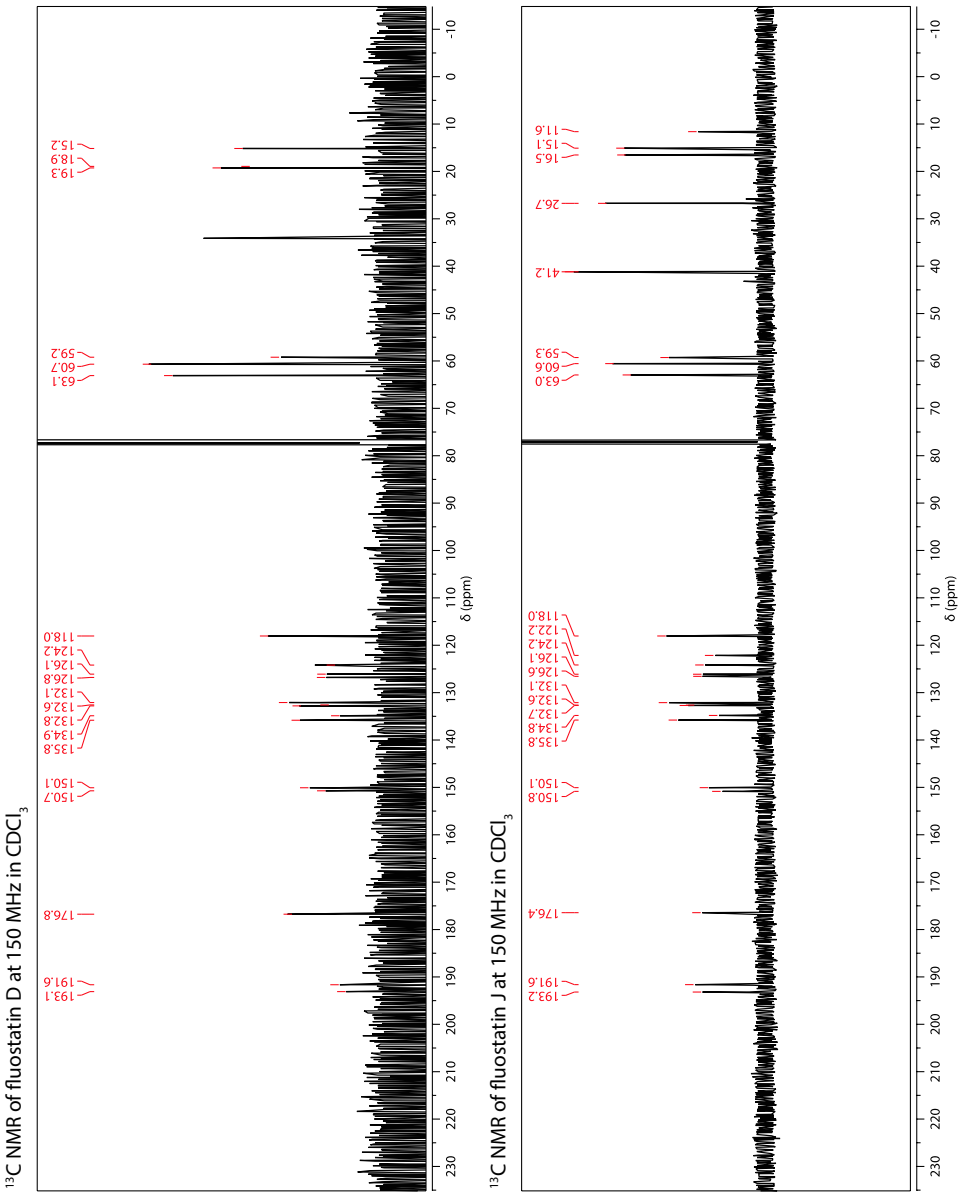


Figure 3-32: ^{13}C NMR of fluostatin D and J in $CDCl_3$.

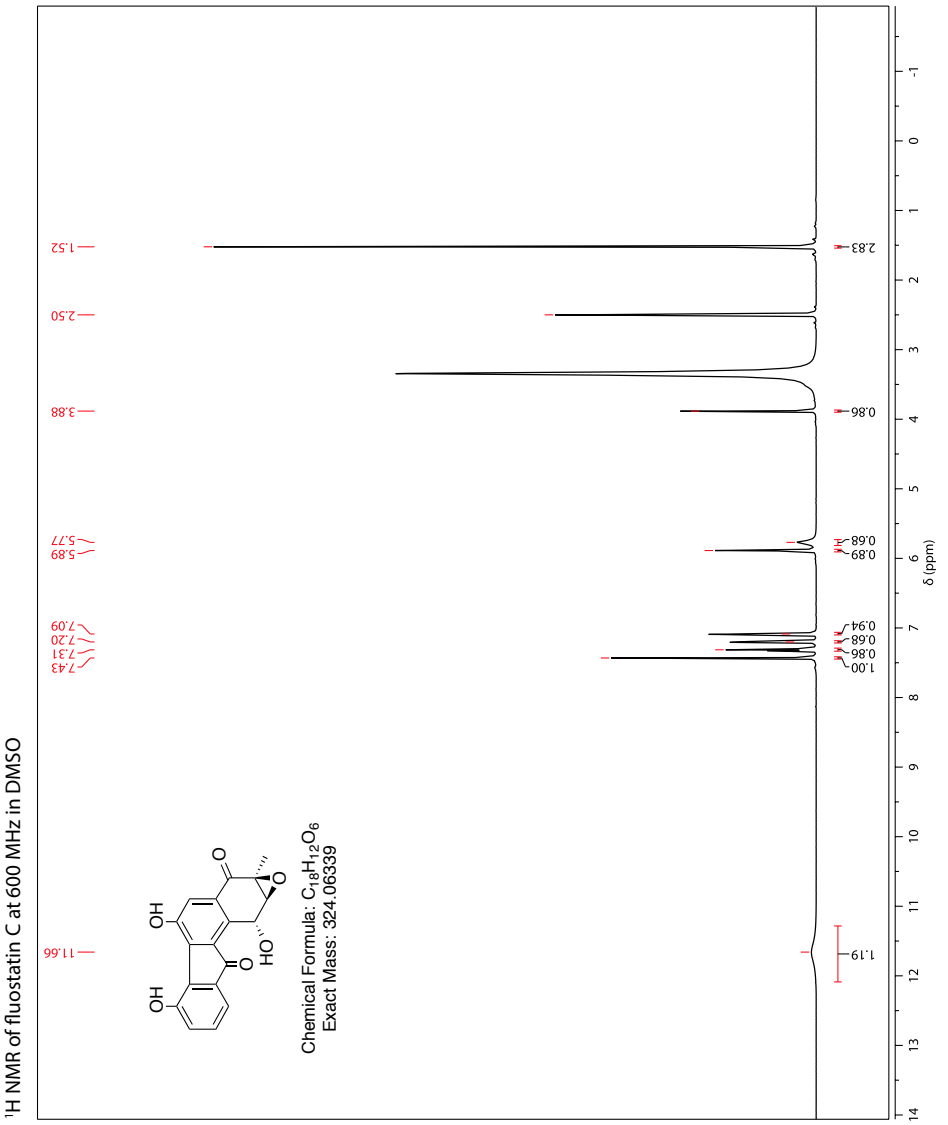


Figure 3-33: ^{1}H NMR of fluostatin C in DMSO-d6.

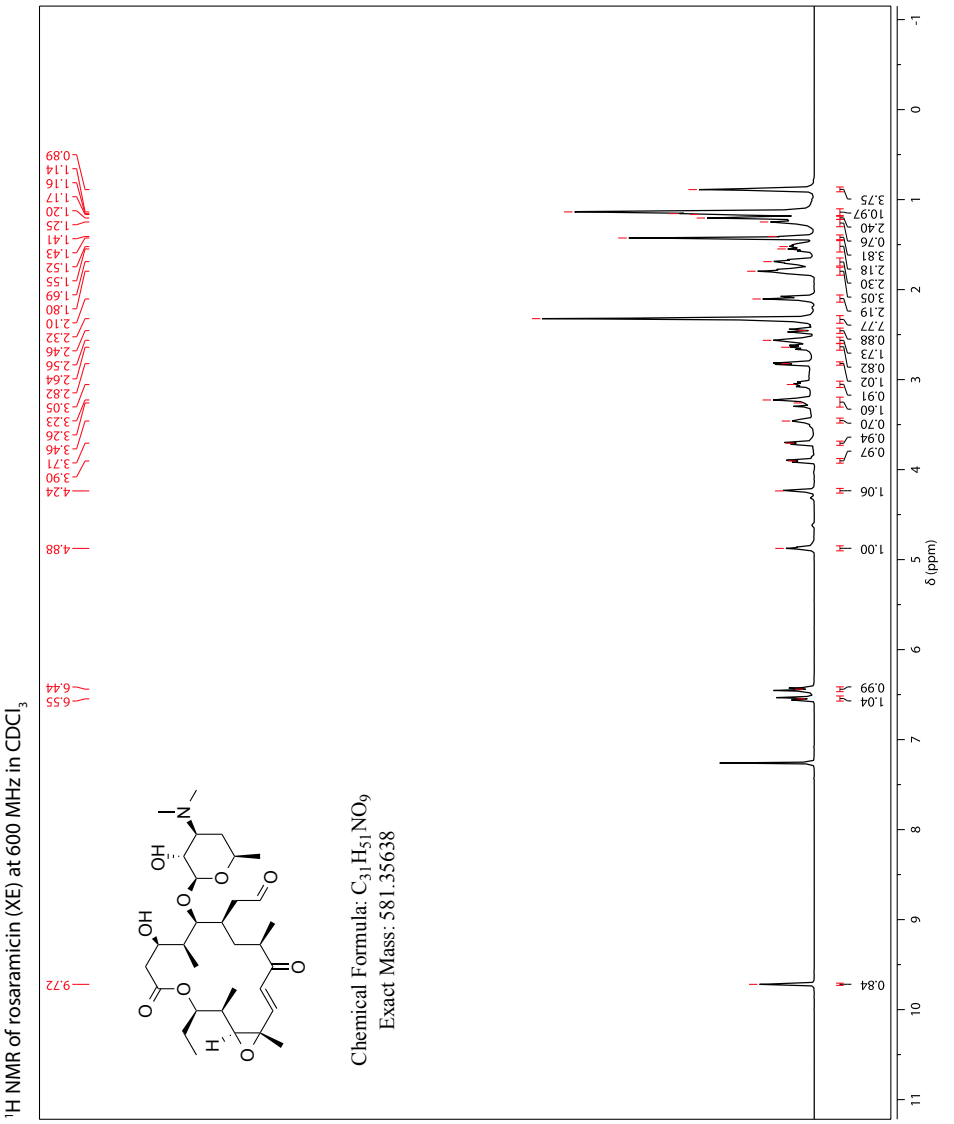


Figure 3-34: ^{1}H NMR of rosaramicin in $CDCl_3$.

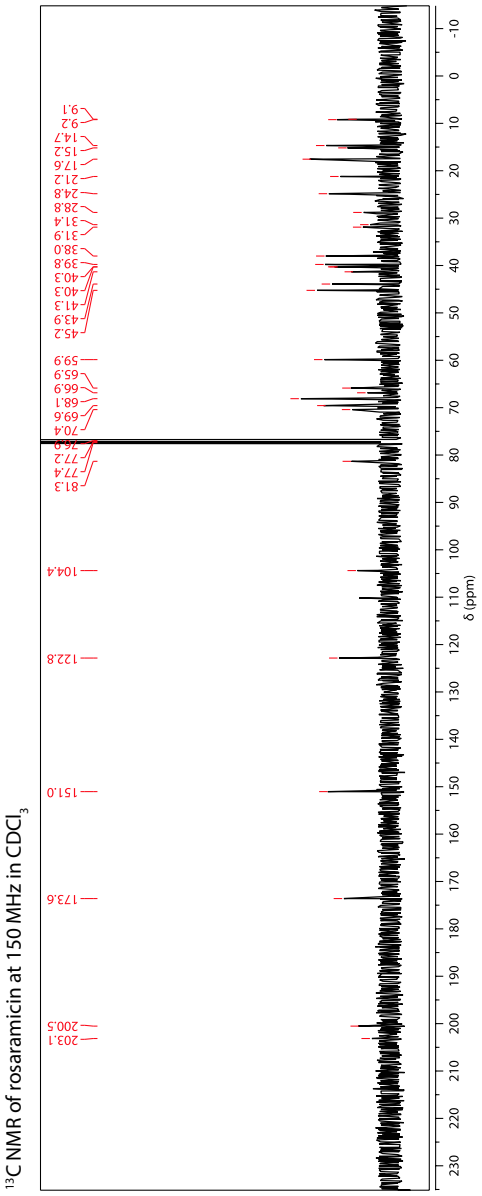


Figure 3-35: ^{13}C NMR of rosaramicin in $CDCl_3$.

3.9. References:

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2012**, *75*, 311–335.

(2) Gerwick, W. H.; Moore, B. S. *Chem. Biol.* **2012**, *19*, 85–98.

(3) Wassermann, A. M.; Lounkine, E.; Urban, L.; Whitebread, S.; Chen, S.; Hughes, K.; Guo, H.; Kutlina, E.; Fekete, A.; Klumpp, M.; Glick, M. *ACS Chem. Biol.* **2014**, *9*, 1622–1631.

(4) Suffness, M.; Douros, J. D. *Trends Pharmacol. Sci.* **1981**, *2*, 307–310.

(5) Navarro, G.; Cheng, A. T.; Peach, K. C.; Bray, W. M.; Bernan, V. S.; Yildiz, F. H.; Linington, R. G. *Antimicrobial Agents and Chemotherapy* **2014**, *58*, 1092–1099.

(6) Peach, K. C.; Bray, W. M.; Winslow, D.; Linington, P. F.; Linington, R. G. *Mol. BioSyst.* **2013**, *9*, 1837–1848.

(7) Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. *Chem. Biol.* **2013**, *20*, 285–295.

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

(9) Perlman, Z. E.; Slack, M. D.; Feng, Y.; Mitchison, T. J.; Wu, L. F.; Altschuler, S. J. *Science* **2004**, *306*, 1194–1198.

(10) Woehrmann, M. H.; Bray, W. M.; Durbin, J. K.; Nisam, S. C.; Michael, A. K.; Glassey, E.; Stuart, J. M.; Lokey, R. S. *Mol. BioSyst.* **2013**, *9*, 2604–2617.

(11) Duncan, K. R.; Crüsemann, M.; Lechner, A.; Sarkar, A.; Li, J.; Ziemert, N.; Wang, M.; Bandeira, N.; Moore, B. S.; Dorrestein, P. C.; Jensen, P. R. *Chem.*

Biol. **2015**, *22*, 460–471.

(12) Hoffmann, T.; Krug, D.; Hüttel, S.; Müller, R. *Anal. Chem.* **2014**, *86*, 10780–10788.

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

(14) Nielsen, K. F.; Månsson, M.; Rank, C.; Frisvad, J. C.; Larsen, T. O. *J. Nat. Prod.* **2011**, *74*, 2338–2348.

(15) Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Brown, L. C. W.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 412–421.

(16) Krug, D.; Zurek, G.; Revermann, O.; Vos, M.; Velicer, G. J.; Müller, R. *Applied and Environmental Microbiology* **2008**, *74*, 3058–3068.

(17) Hou, Y.; Braun, D. R.; Michel, C. R.; Klassen, J. L.; Adnani, N.; Wyche, T. P.; Bugni, T. S. *Anal. Chem.* **2012**, *84*, 4277–4283.

(18) Doroghazi, J. R.; Albright, J. C.; Goering, A. W.; Ju, K.-S.; Haines, R. R.; Tchalukov, K. A.; Labeda, D. P.; Kelleher, N. L.; Metcalf, W. W. *Nat. Chem. Biol.* **2014**, 1–8.

(19) Jiang, Z.-D.; Jensen, P. R.; Fenical, W. *Bioorganic & Medicinal Chemistry Letters* **1999**, *9*, 2003–2006.

(20) Pluskal, T.; Uehara, T.; Yanagida, M. *Anal. Chem.* **2012**, *84*, 4396–4403.

(21) Baur, S.; Niehaus, J.; Karagouni, A. D.; Katsifas, E. A. *Journal of …* **2006**.

(22) Zhang, W.; Liu, Z.; Li, S.; Lu, Y.; Chen, Y.; Zhang, H.; Zhang, G.; Zhu, Y.; Zhang, G.; Zhang, W.; Liu, J.; Zhang, C. *J. Nat. Prod.* **2012**, *75*, 1937–1943.

(23) Lin, C. C.; Chung, M.; Gural, R.; Schuessler, D.; Kim, H. K.; Radwanski, E.; Marco, A.; DiGiore, C.; Symchowicz, S. *Antimicrobial Agents and Chemotherapy* **1984**, *26*, 522–526.

(24) Treiman, M.; Caspersen, C.; Christensen, S. B. *Trends in Pharmacological Sciences* **1998**, *19*, 131–135.

(25) Heifetz, A.; Keenan, R. W.; Elbein, A. D. *Biochemistry* **1979**, *18*, 2186–2192.

(26) de Loubresse, N. G.; Prokhorova, I.; Holtkamp, W.; Rodnina, M. V.; Yusupova, G.; Yusupov, M. *Nature* **2014**, *513*, 517–522.

(27) Donaldson, J. G.; Finazzi, D.; Klausner, R. D. *Nature* **1992**, *360*, 350–352.

(28) Samali, A.; FitzGerald, U.; Deegan, S.; Gupta, S. *International Journal of Cell Biology* **2010**, *2010*, 1–11.

(29) Ding, W.-X.; Ni, H.-M.; Gao, W.; Hou, Y.-F.; Melan, M. A.; Chen, X.; Stolz, D. B.; Shao, Z.-M.; Yin, X.-M. *J. Biol. Chem.* **2007**, *282*, 4702–4710.

(30) Verfaillie, T.; Salazar, M.; Velasco, G.; Agostinis, P. *International Journal of Cell Biology* **2010**, *2010*, 1–19.

(31) Hau, A. M.; Greenwood, J. A.; Löhr, C. V.; Serrill, J. D.; Proteau, P. J.; Ganley, I. G.; McPhail, K. L.; Ishmael, J. E. *PLoS ONE* **2013**, *8*, e65250.

(32) Høyer-Hansen, M.; Jäättelä, M. *Cell Death Differ* **2007**, *14*, 1576–1582.

(33) McBrien, K. D.; Berry, R. L.; Lowe, S. E.; Neddermann, K. M.; Bursuker, I.;

Huang, S.; Klohr, S. E.; Leet, J. E. *J. Antibiot.* **1995**, *48*, 1446–1452.

(34) Sang, F.; Li, D.; Sun, X.; Cao, X.; Wang, L.; Sun, J.; Sun, B.; Wu, L.; Yang, G.; Chu, X.; Wang, J.; Dong, C.; Geng, Y.; Jiang, H.; Long, H.; Chen, S.; Wang, G.; Zhang, S.; Zhang, Q.; Chen, Y. *J. Am. Chem. Soc.* **2014**, *136*, 15787–15791.

(35) Takeuchi, M.; Ashihara, E.; Yamazaki, Y.; Kimura, S.; Nakagawa, Y.; Tanaka, R.; Yao, H.; Nagao, R.; Hayashi, Y.; Hirai, H.; Maekawa, T. *Cancer Science* **2010**, *102*, 591–596.

4. ONE STEP NATURAL PRODUCTS DISCOVERY: INTEGRATED STRUCTURE MATCHING AND MODE OF ACTION CHARACTERIZATION OF INDIVIDUAL METABOLITES FROM NATURAL PRODUCT LIBRARIES

4.1. Introduction:

The preceding chapters have addressed the application of Compound Activity Mapping (CAM) to individual 384-well plates and have shown that this method enables rapid and reliable dereplication and discovery, but has several limitations. The use of other mass spectrometers with tandem mass spectrometry capabilities and expanded dynamic range can address several of the problems with the current platform by increasing the reliability of peak binning and eliminating the need for multiple acquisitions at different detector settings. Not only can MS/MS spectra provide solutions to some of the shortcomings of CAM, the added structural information can confidently identify related compounds based on the similarity of their MS^2 fingerprints and connect these related species across samples. Combining these benefits with CAM's biological annotation capabilities promises to drastically improve the discovery workflow by providing comprehensive biological activity as well as structural information for every detectable compound in every extract for any natural product library.

This chapter will begin by looking at the statistics of the data already acquired on the Linington Extract Library, discuss the Global Natural Products Social Molecular Networking platform for MS^2 spectral comparisons and compound

dereplication with examples of the direct integration of this system with CAM, and briefly discuss future generations of the platform using Waters mass spectrometers and the UNIFI data analysis software.

4.2. Expanding the Library Coverage:

4.2.1. Preamble:

The end goal for CAM is to analyze the entire Linington sample collection in order to know the identity and mode of action of every molecule from every sample. While the analysis of the entire library is beyond the scope of this thesis, performing the data integration on the samples for which MS data has already been acquired is straightforward thanks to the design of the platform. Combining the 96 well plates containing 624 extracts from 104 organisms revealed that there is significant secondary metabolite overlap between the microbes collected in the United States (including American Samoa and Hawaii) and the microbes isolated from marine sediment from Panama. There are also small and large clusters that are specific for multiple organisms from each plate indicating that while the library is likely to contain redundancy, increasing the number of samples should increase the diversity of observed secondary metabolites. For example the phenazines from RLUS_384_1487_1538-CR04 (Plate 4) discussed in Chapter 2 and the quinocinnolinomycins and actinomycins from Panama extract library RLPA_384_1001_2032-CR01 (RLPA) discussed Chapter 3 are present as distinct

clusters while the compounds present in both plates form large clusters of extracts with overlapping chemistry (Figure 4-1).

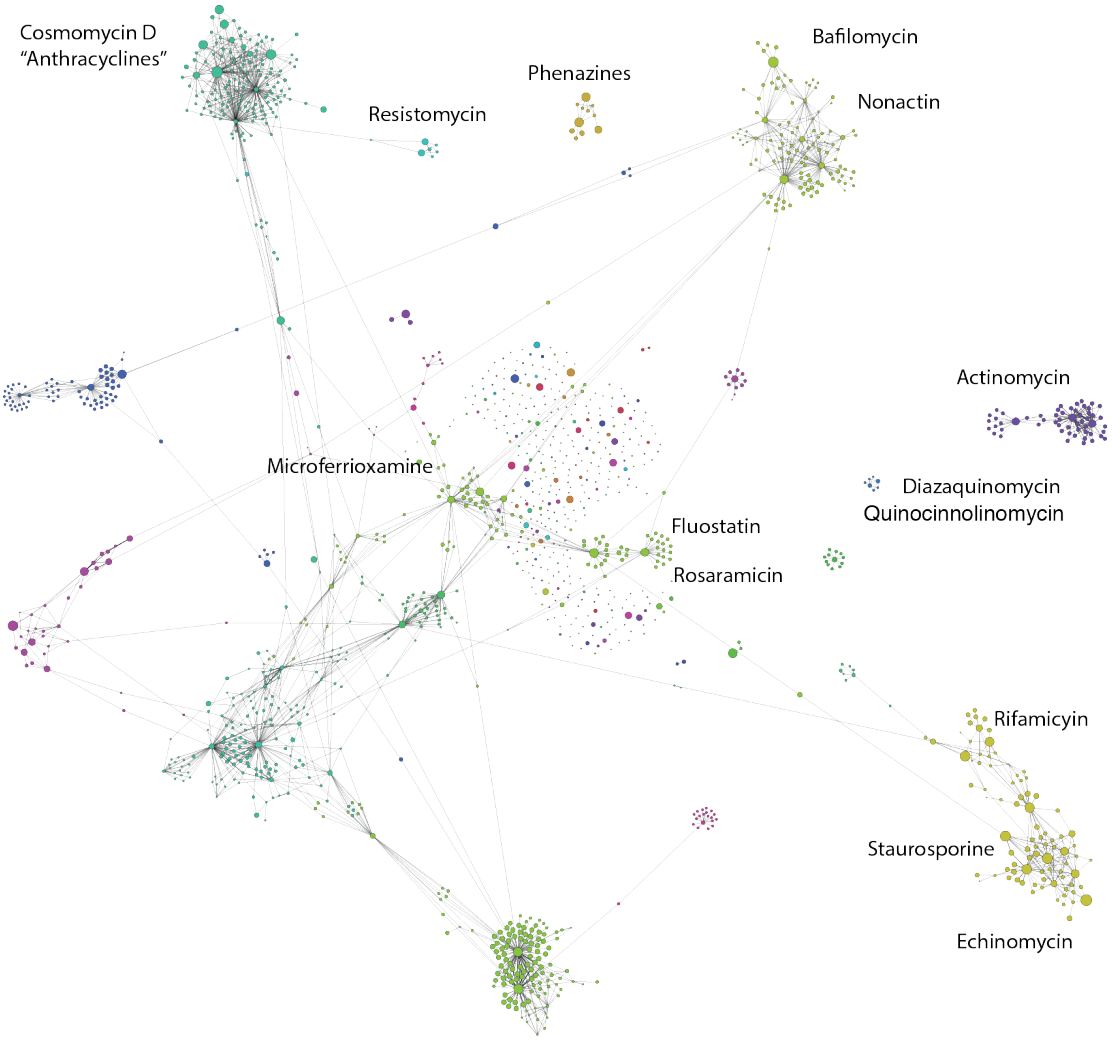


Figure 4-1: A network of all analyzed extracts from RLPA and RLUS libraries connected to *m/z* features contained within those extracts and colored by the modularity algorithm in Gephi. All previously identified compounds are still present in the network and are labeled. There is a significant amount of overlap for some common metabolites such as staurosporine, microferrioxamines, bafilomycin, and nonactin.

4.2.2. Discussion of Data Statistics:

4.2.2.1. Library Diversity and Singletons:

Central to the hypothesis on which CAM is based is the idea that as the number of samples increases, there is likely to be a significant amount of metabolite overlap between extracts. This should reduce the number of false positives because metabolites that appear to be active just because they are biosynthesized once alongside an active compound will, on average, have no bioactivity and be filtered out during the analysis process. Also, the singletons problem should become less frequent. This will result in very few *m/z* features belonging to just one extract, and therefore, any singleton should be prioritized as chemically distinct. In order to test this hypothesis the metadata for the three plates were analyzed by creating individual and combined databases and performing the biological integration.

Analysis of these data supports the general hypothesis that as the library expands the number of unique *m/*z features will decrease. When all the extract data are combined, there are 23,228 *m/z* features, including singletons, with only 150 appearing in greater than 10% of the extracts. Individually there are 10,977, 8,544, and 10,276 features for Panama, Plate 4, and RLUS_384_2163_2214-CR17 (Plate 17) respectively for a total of 29,797. This means that 6,569 features were common between all three plates. While it is unnecessary to define how many of these 6,569 *m/z*, retention time (rt) pairs were present in all three plates or present in just two plates, this strongly indicates that there is a large amount of redundancy between plates with up to 28 % of the *m/z* features occurring in more than one plate. Similarly,

from all the plates there are 11,684 *m/z* features that were present in 2 extracts from that plate and were present in less than 10% of the total extracts from all three plates. These represent all the non-singleton features that are not from either systematic contamination or primary metabolites. There were 5,321, 4,507, and 3,734 features from Panama, Plate 4, and Plate 17 respectively for a total of 13,562 features indicating that there are approximately 1,878 or 16 % of the features that are common to samples within plates and across plates. These data suggest that a comprehensive dataset including all 6,000 Linington Lab extracts is likely to contain several instances of every *m/z* feature in the database.

Less common metabolites are also likely to be redundant in the library. There were 5,310, 3,995, and 5,625 singletons from Panama, Plate 4, and Plate 17 respectively for a total of 14,930 total; however, upon combination of the plates there were only 11,394 total singletons. Suggesting that there are 3,536 or 15 % of the features that are present only once per plate but are present in more than one plate. All these data support the redundancy hypothesis and indicate that the predictive power and accuracy of CAM will increase as the size of the library increases.

4.2.2.2. Activity and Cluster Score Comparison:

Adding more extracts also affected the frequency distributions of activity and cluster scores (Figure 4-2). From the Panama plate RLPA, the mean activity score was 4.66 with a standard deviation of 5.53 and the mean cluster score was 0.13 with a standard deviation of 0.14. For all the extracts the mean activity score was 5.95 with a standard deviation of 6.76 and the mean cluster score was 0.12 with a standard

deviation of 0.15. The higher activity score and lower cluster score is from the inconsistency in the death dilutions. The extracts from Plate 4 were diluted using Nuclear Count Tranfluor EdU to report on cell count. The fingerprint of the dilution closest to the LD_{50} was used for biological profile integration. Plate 17 data was not dilute at all. As discussed in Chapter 2 and Chapter 3, selecting the proper dilution is critical for the accurate prediction of a mechanism of action; however, the combined data provides a useful insight into the flexibility of CAM.

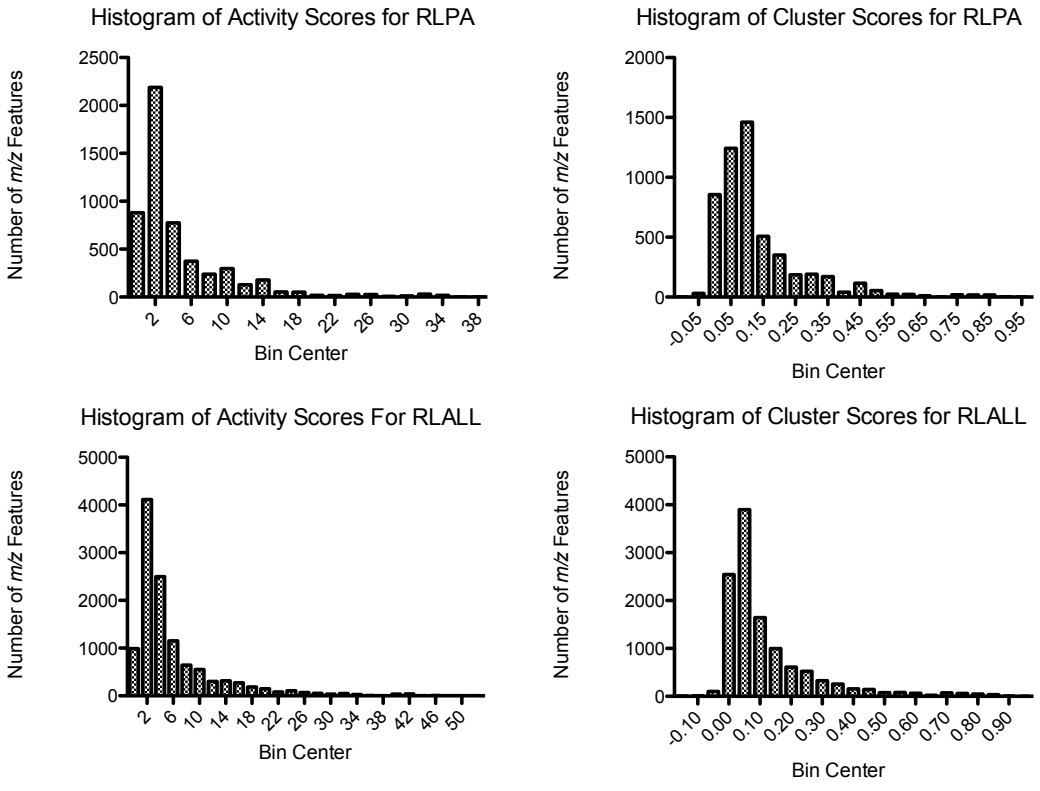


Figure 4-2: Histograms of the biological integration metrics cluster score and activity score for the Panama Plate (RLPA) and all the data (RLALL).

The inconsistency in dilutions selection can be resolved in the future, but the resulting network indicates that even with partial dilution data, CAM can provide accurate predictions of the bioactive constituents in an extract (Figure 4-1). This is important because it means that the system is robust to batch effects and partial data. Therefore, discovery and dereplication can be performed in real time as the library is being processed through the platform.

4.3. GNPS – Global Natural Products Social Molecular Networking:

4.3.1. Preamble:

The major advances in natural products over the next 20 years will come from the ability to leverage the swaths of data collected by modern instruments and biological screens to ask questions that are fundamental to our understanding of the function of secondary metabolites in ecology and biomedical research. Generally, these approaches will allow researchers to more quickly arrive structural and functional annotation in order to avoid the expensive process of compound structure elucidation and target identification. While the overarching focus of our lab is to identify potential uses for new and old compounds as molecular probes or therapeutics, the Dorrestein Lab at the University of California in San Diego focuses primarily on the use of tandem mass spectrometry data to rapidly profile classes of metabolites from live organisms, libraries, and even human skin.[1-4] We were invited by Professor Dorrestein to participate in a flagship study in which they began creating a centralized platform for the storage and analysis of natural products MS^2 data from

labs all around the world. The platform they call Global Natural Product Social Molecular Networking (GNPS) aims to take the information that remains hidden in laboratory notebooks, in-house databases, and papers and make it all generally available in the same way that the National Center for Biotechnology Information (NCBI) has curated databases of nucleotide and protein sequence information.

The two key components of the system are the GNPS database and Molecular Networking for data visualization and exploration. The GNPS database is structured so that users can add raw data to the database in the form of MZXML formatted files and may choose to annotate compounds by the scan number in which the parent mass appears. In this way, all the MS^2 data from every known compound from all the labs can be collected, stored in one central location, and explored. The compounds are assigned classifications (Gold, Silver, and Bronze) based on whether there is extra structural data to support and corresponding publications to support the assignment. Also, because ionization, adduct formation, fragmentation, and therefore the MS^2 fingerprints of compounds can change dramatically on different systems under different conditions, the database stores a large amount of metadata about the acquisitions including the instrument, detected adduct, ionization source, and strain information for the producing organism. As part of the project, I traveled to San Diego for three weeks and added 14 Gold level annotations to the GNPS library from the Panamanian extracts. In addition, I acquired the raw spectra from all 234 extracts and these were added to the GNPS library database.

One of the advantages of GNPS and all open, crowd-sourced information (even in the social media forms of Facebook, Twitter, etc.) is that the addition of raw spectra from extracts allows for information about rarefaction, coexpression of metabolites, and overall metabolite diversity to be explored in an open and accessible format. Any user can query the database to find out if their spectra match other spectra, regardless of whether they are annotated or not. While the coexpression of two metabolites may not be of interest to the data collector, this information that would otherwise be inaccessible in a hard drive is available for others for their scientific pursuits. What's more, as the dataset is continually updated and users are notified monthly if others have identified peaks from their library or if others have proposed alternate assignments, creating a living set of data that becomes richer and more informative as more users add their knowledge.

So far the discussion has focused on a database format that is searchable through online tools, but much of the power of GNPS relies on the visualization of datasets using molecular networks in which parent *m/z* peaks can be connected by the similarity of their MS^2 profiles.[1] This technique has been rapidly adopted by in the field of natural products and has been used in everything from discovery of new molecules from well studied organism like *Streptomyces coelicolor,*[5] to the detection of siderophores in white nose syndrome that is killing millions of bats from the United States,[6] to integrated genomics approaches connecting molecules to the gene clusters that encode them.[3] Originally, the algorithms were used for the clustering of similar peptides from proteomics data in order to identify proteins more quickly using

database searches.[7,8] Following this, these same principles were applied to small molecule peptides[9] and later to all types of small molecules[1] produced by living colonies. Overall the system works well for the broad characterization of large, ionizable molecules from natural product libraries and produces appealing and easily interpretable data outputs.

This platform is effective, widely adopted, and will be instrumental in modernizing the field of natural products; however, it is not without its limitations. As mentioned before, different instruments, run by different people, with different conditions will always yield slightly different results. The metadata that is stored will help to understand these differences in parent mass ionization and MS^2 spectra, but adjusting the parameters within GNPS and Molecular Networks to give the most informative data is far from an exact science. For example, the parameters cosine score of the MS^2 fingerprints and the number of peaks required for a positive identification change the connectivity of the molecular networks outputs and the number of identifications in GNPS significantly. Choosing too strict values misses identifications, but overly lenient settings cause the system to take far too long to analyze the data and will produce spurious results. More work is needed to create ways of analyzing the data so that these parameters can be set more intelligently.

4.3.2. Panama Library of Compounds:

From the 234 extracts analyzed 200,370 parent *m/z* peaks were detected with acceptable MS^2 fingerprints. The blanks were removed by assigning those spectra to group 1 in GNPS and then using a filter in cytoscape to remove any parent *m/z* node

that belongs to the blank group. After blank removal 187,448 nodes remain in the network table and are only assigned to group 2. This amount of data cannot be easily displayed and crashes the Cytoscape Viewer. Using filters for the data allows pertinent information to be viewed from the whole library. For example, of the 187,448, only 66 had MS^2 spectrum cosine scores >= 0.7 and 4 matching peaks with known compounds in the GNPS library. These nodes are displayed with their Library ID name in **Figure 4-3**.

The GNPS website allows you to change many of the parameters used for linking parent *m/z* nodes together as well as whether or not to display nodes with acceptable MS^2 spectra but no connections in the dataset. Unless otherwise stated the following parameters were used to create molecular networks from GNPS: cosine score >= 0.6; min peaks = 3; parent *m/z* range of 0.2 Daltons; Ion Tolerance = 0.5 Daltons Minimum cluster size 1. Adjusting these parameters can significantly change the data output and is adjusted on a dataset by dataset basis. These are the default settings recommended for large datasets.

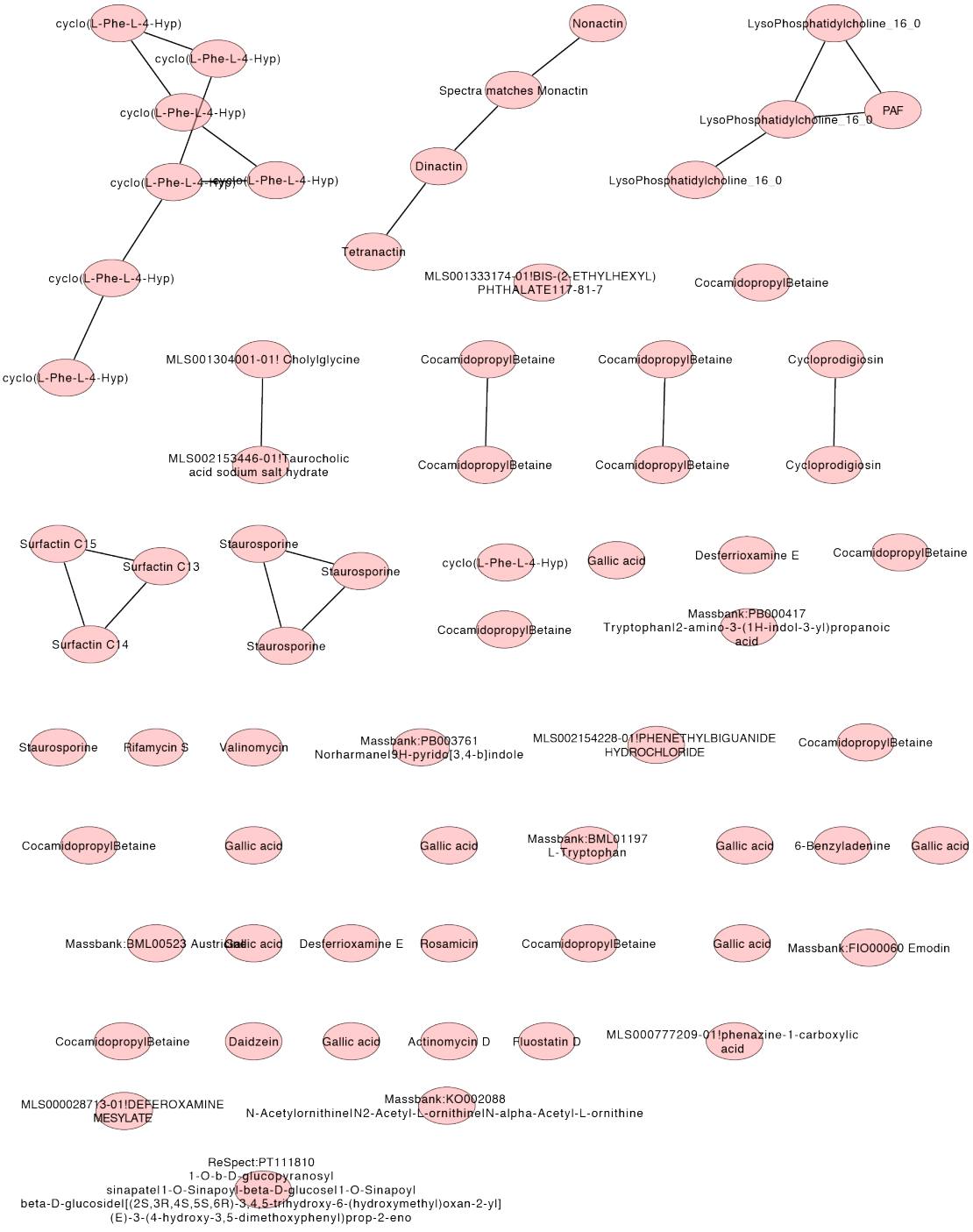


Figure 4-3: A network of all nodes from the Panama library identified by the MS^2 similarity searches from the GNPS libraries with the standard settings for analyzing large datasets.

4.3.3. Compound Activity Mapping, GNPS, and Molecular Networks:

To pare down the number of nodes in the network so that the dataset may be visually represented and enriched only for bioactive components, the CAM data was directly integrated with the molecular network. First a table of all the *m/z* features from CAM with activity scores greater than 10 and cluster scores greater 0.15 was exported from CAM. This table was then compared to the node table exported from GNPS. Parent *m/z* nodes present in the table from CAM were assigned to group 3 (as previously discussed group 1 contained nodes present in the blank runs and group 2 contained all the extract nodes). This new group contained only parent *m/z* nodes that were predicted to be responsible for the observed bioactivity. The twice-filtered molecular networking data for the Panama extract library is displayed in Figure 4-4.

Using the bioactivity filter enabled the exploration of the data in cytoscape and the analysis of the hybrid system. Comparing the manual identification to the MS^2 GNPS database identifications showed the utility and limitations of GNPS dereplication. GNPS was able to correctly match 9 out of 10 previously known compounds and identify phenazine-1-carboxylic acid that was previously not dereplicated **Figure 4-3**. The only compound that is known to be in the active extracts that was not identified was the vacuolar-ATPase bafilomycin A1. Molecular networks displayed the *m/z* and two parent *m/z* features differing by 14 amu; however, the database did not identify the $[M+Na]^+$ as bafilomycin A1 (Figure 4-5). As the database grows and more compounds are entered the identification of common metabolites like bafilomycin should be routine. Differential ionization and

fragmentation between instruments and collision cell energies will always result in different fragmentation patterns and the intensity of fragment ions, but as more users enter known compounds into library from different instruments, this problem should also be at least partially resolved.

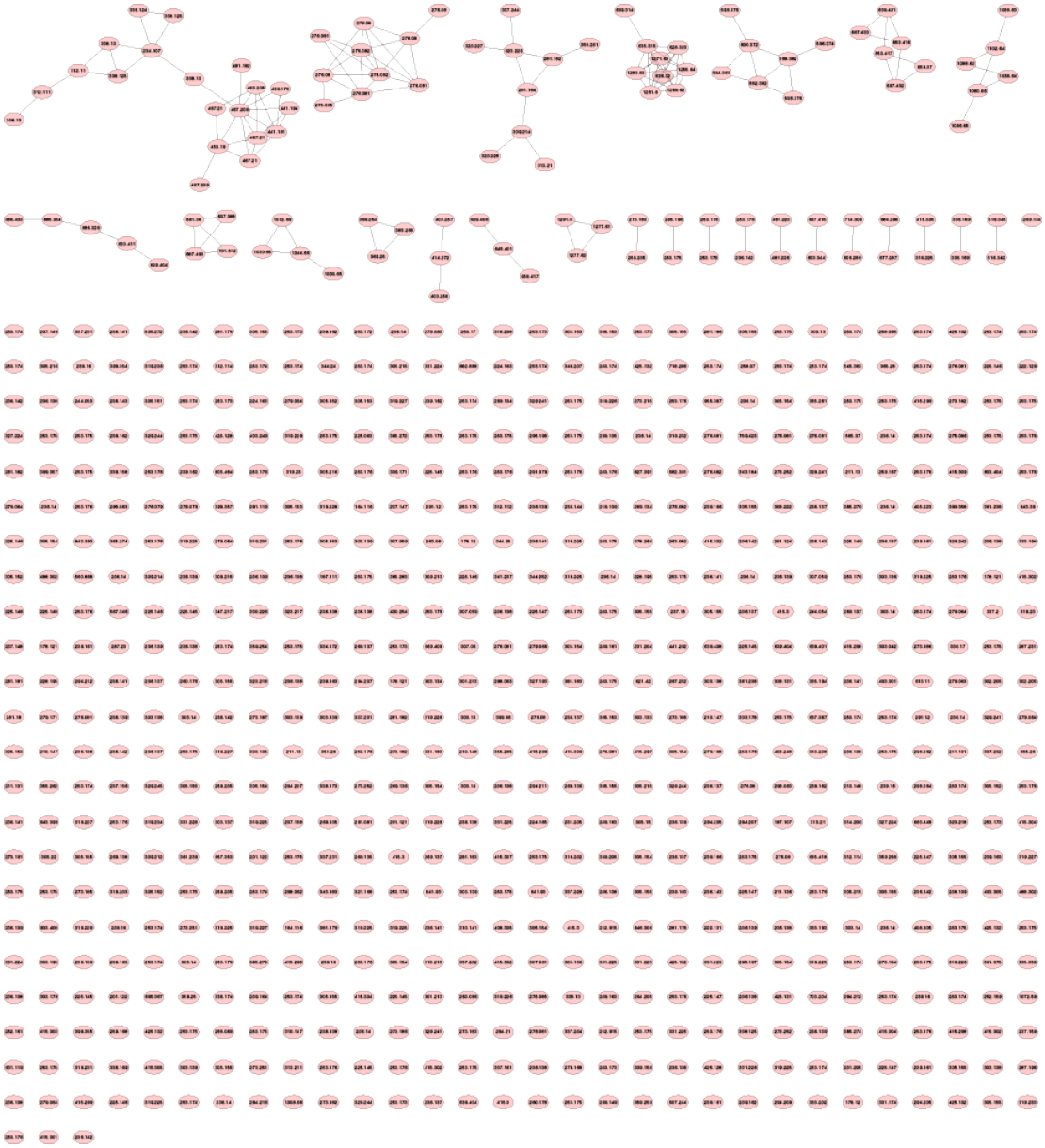


Figure 4-4: Molecular network of all the parent mass nodes with acceptable MS^2 fingerprints that are also predicted to be active in compound activity mapping. The large proportion of unconnected nodes in the network indicates that many of the masses in this compound library are structurally distinct, or either the displayed metabolite or related family members were not at concentrations sufficient for the consistent detection of fragment ions.

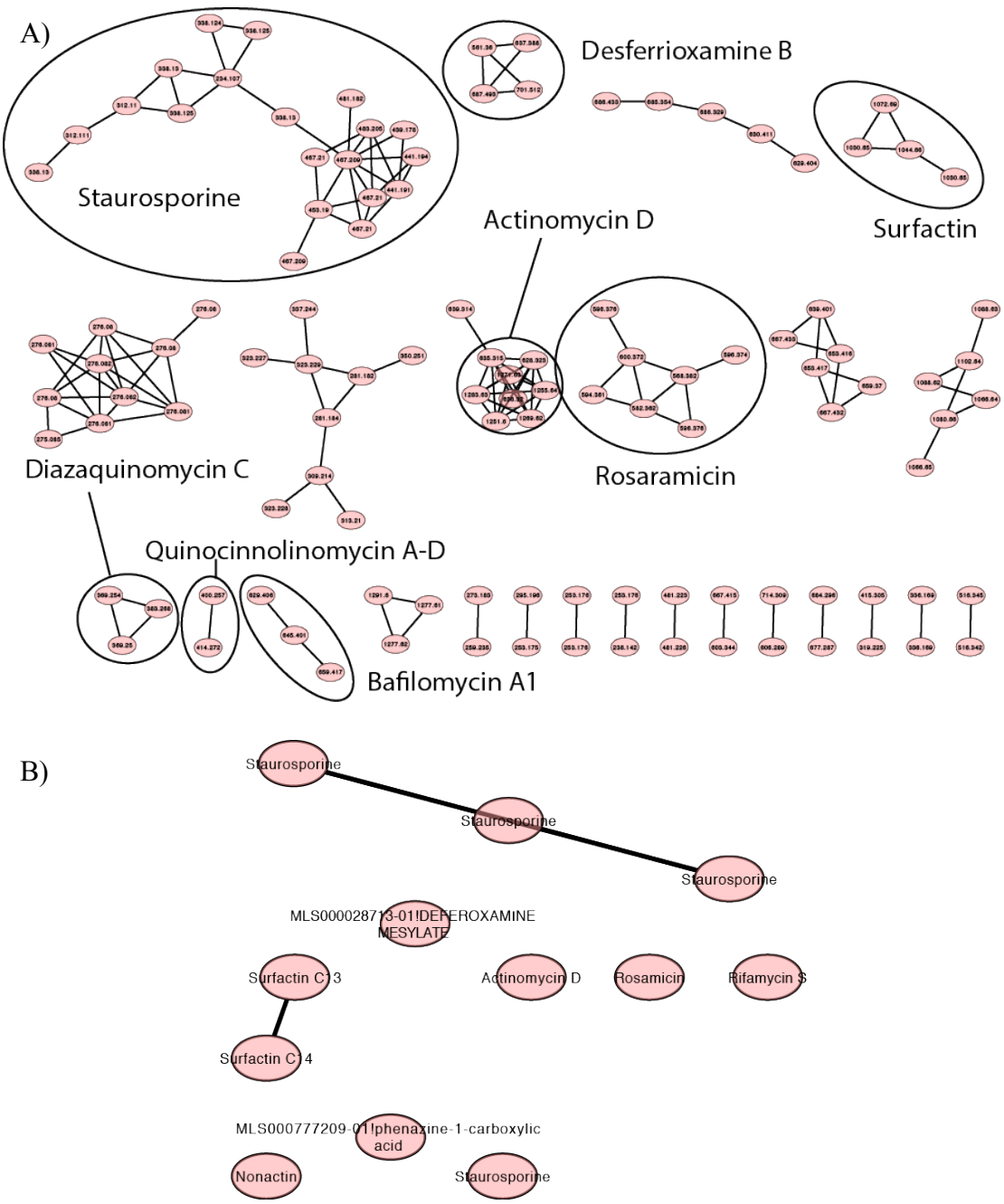


Figure 4-5: Identified structures from manually assigned from CAM and identified by GNPS. (A) Expansion of the network showing only active parent *m/z* nodes with MS^2 cosine similarity scores greater than 0.6. The compounds that were previously identified by CAM are labeled. (B) Active parent *m/z* nodes identified by comparison of MS^2 spectra from the available GNPS libraries.

4.3.4. Investigation of the Extracts of Individual Organisms:

Molecular networking is also useful for displaying compound families from one or several organisms Figure 4-6 and **Figure 4-7**. For example, examining the extract RLPA-2003, the parent *m/z* features for diazaquinomycin C and quinocinnolinomycin A-D belong to two distinct clusters. Other analogues of diazaquinomycin C are clearly present in the extract of RLPA-2003, but the MS^2-based clustering provides a way to dereplicate these compounds without further purification or the need to do coinjections. A second example of the utility of the MS^2 based clustering in combination with bioactivity profiling is the extract RLP-2021 (**Figure 4-7**). The compound rosaramicin and its analogues have been heavily studied[10-12] and derivatized because of their antibiotic activity and are now easily distinguished from the parent *m/z* nodes corresponding to the fluostatins.[13] CAM did not previously have the ability to distinguish compounds based on structural similarity and displayed these compounds in a large single cluster. Now these compounds as well as desferrioxamine B appear in the Cytoscape viewer, are positively identified by GNPS, and are connected to related compounds.

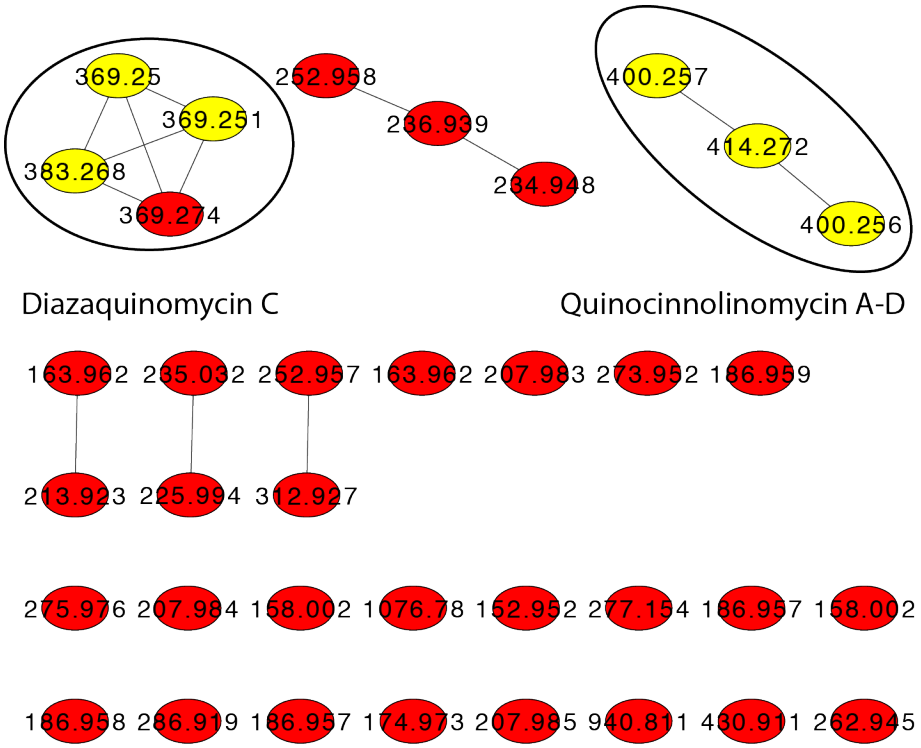


Figure 4-6: Molecular network of extract RLPA-2003. The parent m/z nodes predicted to be active by Compound Activity Mapping are highlighted in yellow. Previously derreplicated compound clusters are labeled with the positively identified natural product.

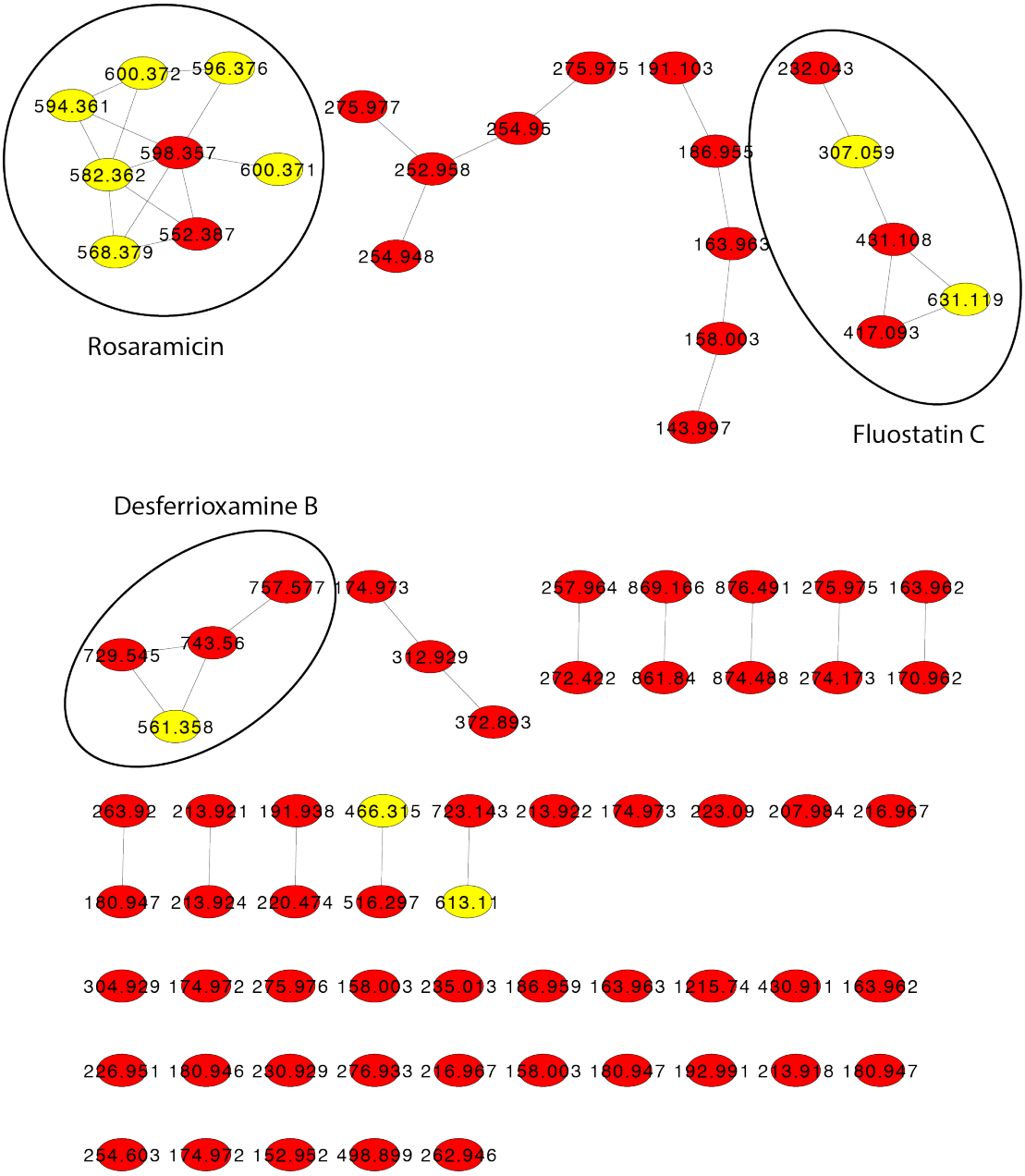


Figure 4-7: Molecular network of extract RLPA-2021. The parent m/z nodes predicted to be active by Compound Activity Mapping are highlighted in yellow. Previously derreplicated compound clusters are labeled with the positively identified natural product.

4.4. Waters Collaboration:

While it is beyond the scope of this thesis to discuss the future integration of MS^e data into CAM, a brief discussion of the capabilities of Waters Unifi system and its application to the Panama library will provide a glimpse into the future of the platform. Over the last year we have worked closely with the technology development department at Waters to use their new MS technologies to explore our natural product library. Through many conference calls and a trip to Milford to visit their department and perform method optimization, we have developed methods for the next generation of CAM including the use of MS^e. Briefly, MS^e is a technology that is similar to other tandem MS experiments that use collision assisted dissociation to fragment parent ions; however, instead of selecting for one *m/z* at a time to be fragmented, all ions are fragmented. The MS switches rapidly between two modes: low energy mode in which parent *m/z* features are detected, and high energy mode in which all observable fragments are detected. The fragment ions are aligned as best as possible to parent ions using the chromatogram and the data can be analyzed similarly to MS^2 fragmentation data.

There are several distinct advantaged to MS^e over MS^2. The biggest advantage is that fragmentation data is collected for every detectable parent ion. Data dependent MS^2 acquisitions can only fragment the ten most concentrated ions at any one time. This limits the number of metabolites that can be detected severely because co-eluting peaks will be lost if the titres or relative ionization are dramatically different. Also, focusing one ion at a time limits the number of scans that can be acquired for each

molecule, reducing the quality of data. The primary disadvantage to this strategy is that, at the moment, assigning fragment ions to precursors remains a challenge and the data output is not straightforward making this technique difficult to directly integrate with CAM.

It has already been shown that CAM is able to predict the parent masses responsible for biological activity and that these masses can be grouped by structural similarity based on MS^2 fingerprints; however, this analysis is complicated and requires a lot of steps. In order to re-integrate the information obtained in compound activity mapping into a user friendly walk up instrument, we will integrate a custom compound structure database with the fragment identification capabilities included in Unifi. By building an atlas of all bacterial natural products and using structural similarity to cluster families of molecules in a network, we will use related parent masses to map *m/z* features from extracts to predicted families.

These families can be imported into the Unifi scientific libraries and the MS^e fingerprints of these natural products can be automatically analyzed by Unifi (Figure 4-8). For example, when a small library of compounds identified in the Panama Plate were imported as a scientific library, surfactin was readily identified and the structure was confirmed accurately by Unifi without user input. This example shows how, combined with CAM and a comprehensive natural products atlas, Unifi can enable comprehensive dereplication, lead confirmation and biological activity prediction.

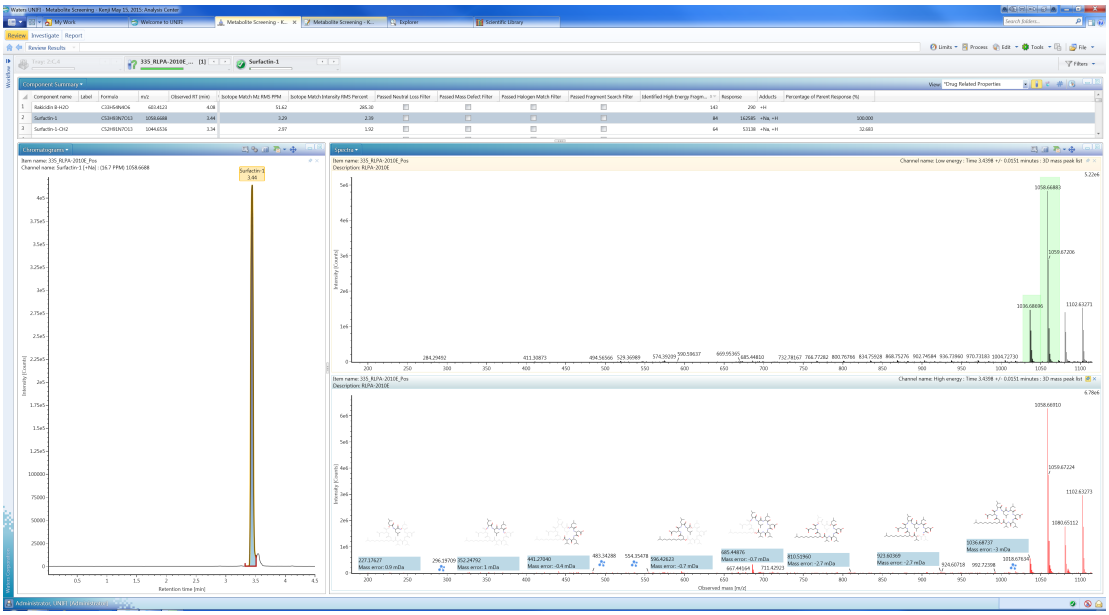


Figure 4-8: A screen capture from the Waters Unifi Natural Products Solutions Program. Top window contains the list of identified masses with surfactin highlighted. The left window shows the extracted ion chromatogram (EIC) for surfactin identified from RLPA-2010E. The upper right window is average spectrum of the low-energy (MS^1) scans over the time window from the EIC. The identified peaks highlighted in green are related $[M+H]^+$ and $[M+Na]^+$ adducts of the same parent mass. The lower right window shows the average of the high-energy (MS^e) spectra with the signals identified with their corresponding structure fragments.

4.5. References:

(1) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1743.

(2) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686.

(3) Duncan, K. R.; Crüsemann, M.; Lechner, A.; Sarkar, A.; Li, J.; Ziemert, N.; Wang, M.; Bandeira, N.; Moore, B. S.; Dorrestein, P. C.; Jensen, P. R. *Chem. Biol.* **2015**, *22*, 460.

(4) Bouslimani, A.; Porto, C.; Rath, C. M.; Wang, M.; Guo, Y.; Gonzalez, A.; Berg-Lyon, D.; Ackermann, G.; Moeller Christensen, G. J.; Nakatsuji, T.; Zhang, L.; Borkowski, A. W.; Meehan, M. J.; Dorrestein, K.; Gallo, R. L.; Bandeira, N.; Knight, R.; Alexandrov, T.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E2120.

(5) Sidebottom, A. M.; Johnson, A. R.; Karty, J. A.; Trader, D. J.; Carlson, E. E. *ACS Chem. Biol.* **2013**, *8*, 2009.

(6) Mascuch, S. J.; Moree, W. J.; Hsu, C.-C.; Turner, G. G.; Cheng, T. L.; Blehert, D. S.; Kilpatrick, A. M.; Frick, W. F.; Meehan, M. J.; Dorrestein, P. C.; Gerwick, L. *PLoS ONE* **2015**, *10*, e0119668.

(7) Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. *PNAS* **2007**, *104*, 6140.

(8) Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2007**, *7*, 113.

(9) Kersten, R. D.; Yang, Y.-L.; Xu, Y.; Cimermancic, P.; Nam, S.-J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat. Chem. Biol.* **2011**, *7*, 794.

(10) Lin, C. C.; Chung, M.; Gural, R.; Schuessler, D.; Kim, H. K.; Radwanski, E.; Marco, A.; DiGiore, C.; Symchowicz, S. *Antimicrobial Agents and Chemotherapy* **1984**, *26*, 522.

(11) Rfimann, H.; Jaret, R. S. *J. Chem. Soc., Chem. Commun.* **1972**, 1270a.

(12) Anzai, Y.; Sakai, A.; Li, W.; Iizaka, Y.; Koike, K.; Kinoshita, K.; Kato, F. *J. Antibiot.* **2010**, *63*, 325.

(13) Zhang, W.; Liu, Z.; Li, S.; Lu, Y.; Chen, Y.; Zhang, H.; Zhang, G.; Zhu, Y.; Zhang, G.; Zhang, W.; Liu, J.; Zhang, C. *J. Nat. Prod.* **2012**, *75*, 1937.

5. GENES TO MOLECULES AND BACK AGAIN: GENOME MINING APPROACHES TO NATURAL PRODUCTS DISCOVERY

5.1. Introduction:

The biosynthesis of natural products continues to be extensively studied due to the importance of secondary metabolites for biomedical applications, as well as for the fundamental understanding of the role these molecules play in the natural environment. Understanding the significant role of the human microbiome in human health through microbe-microbe and microbe-host interactions has also bolstered the field of bacterial biosynthesis.[1] The use of genomic information combined with the improvements in analytical technologies has ushered in a new era of natural products discovery and has provided a glimpse into the outstanding potential for accessing new chemistry through the isolation and heterologous expression of "cryptic" or unexpressed metabolites in standard microbiology hosts.[2] The investigation of individual and related gene clusters have been extensively studied as a method for production optimization,[3] derivatization,[4] bioengineering,[5] and the metagenomic annotation of producing strains.[6] Overviews of the synthetic machinery used in these applied studies has been highlighted in many reviews and is beyond the scope of this discussion,[7,8] but as our understanding of how the modules involved in biosynthesis may be connected increase,[9] the potential for plug and play heterologous expression of novel natural products will become a reality.[10]

While this expanding subdiscipline of natural products is increasingly accessible due to the continuously declining cost of genomic sequencing and

assembly, the identification of biosynthetic gene clusters has been limited to a subset of the classes of natural products.[11-15] In order to expand the breadth of detectable gene clusters, the Fischbach group at the University of California at San Francisco, in collaboration with Professor Marnix H. Medema of the University of Groningen, The Netherlands, developed a Hidden Markov Model based method to expand upon antiSMASH, a similar online tool developed by these labs, to detect more classes of gene clusters in order to create a global census of all bacterial biosynthetic gene clusters from 1,154 sequenced bacterium for which the sequence data is publicly available.[15-17]

Briefly, the new ClusterFinder algorithm takes nucleotide sequences and converts them into Pfam domains or protein family domains. Pfam domains are assigned by comparison to a database of assigned protein families using the HMMER package.[18] Each domain is assigned a probability of being part of a gene cluster based on the frequency at which it appears in a database of 732 manually curated known biosynthetic gene clusters in reference to the preceding and subsequent domain. While it may seem that this system would bias for gene clusters similar to those in the training set, ClusterFinder has low training set bias because each gene is broken up into small segments. These segments (domains) are continuously reused to create different types of secondary metabolites, and therefore the presence of homologous domains in tandem may be used to find new classes of biosynthetic gene clusters.

The ClusterFinder algorithm was used to identify biosynthetic gene clusters from 1,154 sequenced genomes representing all publically available sequenced

bacterial genomes. From these genomes 33,351 putative biosynthetic gene clusters were identified, 10,742 with high confidence and 22,627 with low confidence. An all-by-all matrix of the evolutionary distance between clusters was then constructed of the high confidence identifications. This matrix was used to create a network of the gene clusters.[19] Several well-studied families of gene clusters are taxonomically widely distributed across bacteria taxonomy. These families include NRPS-independent siderophores, O-Antigens, capsular polysaccharides, and carotenoids, which are all distinguished in the network (Figure 5-1).[20-23] While it was expected that there would be several small families of gene clusters specific for closely related strains, there was large, unidentified family containing 1,021 gene clusters that was widely distributed across many bacterial clades. This chapter will describe the characterization of this family of gene clusters that produce aryl-polyenes (APEs). The APE family is the largest family of gene clusters even exceeding the carotenoids with 870 gene clusters (Figure 5-1, Figure 5-2, and Figure 5-3).[24]

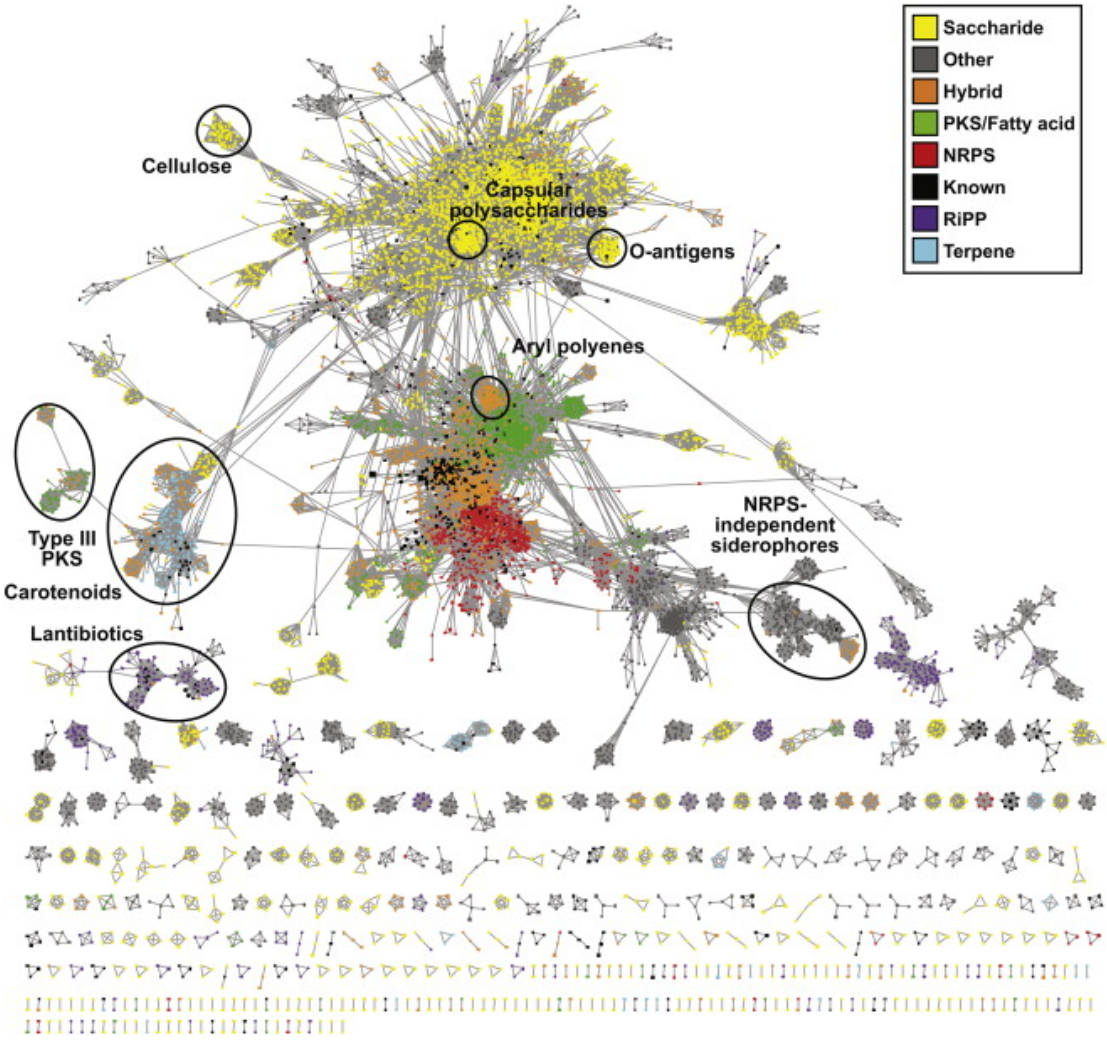


Figure 5-1: A similarity network of high confidence biosynthetic gene clusters (nodes) connected if their similarity score[19] is greater than 0.5. The edges are weighted by this similarity score. The largest connected component contains 72% of the gene clusters and contains oligosaccharides, nonribosomal peptides (NRPs), polyketides/lipids indicating that these types of biosynthetic gene clusters are common to many different families of gene clusters. The genes are colored based on gene cluster type and taxonomically widely distributed gene clusters such as NRPS-independent siderophores, O-Antigens, capsular polysaccharides, and carotenoids are circled on the graph. The APEs are also circled. Adapted from Cimermancic *et al.*[24]

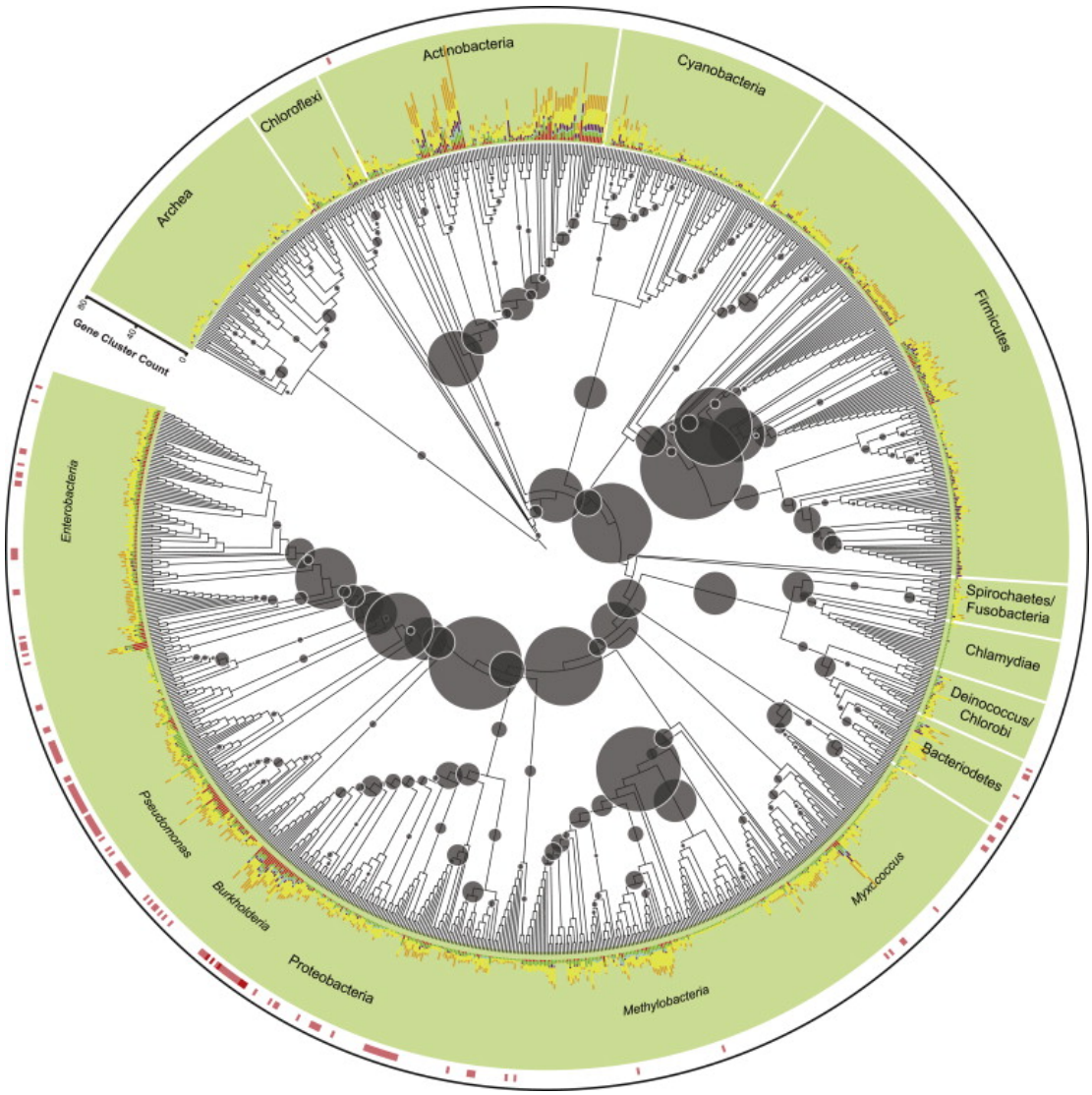


Figure 5-2: A phylogenetic tree of the 1,154 organisms used in the biosynthetic gene cluster analysis with the high confidence gene clusters arrayed as colored bars around the outside of the cladogram and the circles within the tree indicating the amount of diversity at each node. The Gene clusters are colored based on the class with the same color key as in Figure 5-1. Organisms in which an APE related biosynthetic gene cluster was predicted are labeled with a red bar on the outside of the circle. APE genes are distributed widely across clades of gram-negative bacteria. Adapted from Cimermancic *et al.*[24]

This gene cluster family could be split into three distantly related subfamilies. Only subfamily three contained gene clusters for which the product had been identified. These products, flexirubin[25,26] and xanthomonoadin,[27-29] are aryl polyenes (APE)s and have been extensively studied, but the relationship between these gene clusters had not been previously identified. The family and three subfamilies are extremely widely dispersed across gram-negative bacteria with 36.4% of the strains from a typical genera containing a related gene cluster; however, subfamilies 1 and 2 had no known small molecule products. One example for each subfamily was chosen for characterization in order to understand the relationship between these extremely common related gene clusters. The isolation and structure elucidation of the APE_{EC} and APE_{VF} molecules is presented in this chapter.

5.2. Structure Assignment:

In order to characterize the compounds produced by the flexirubin like families 1 and 2 Dr. Jan Claesen of the Fischbach Laboratory at the University of California at San Francisco, Mission Bay amplified the 15.5 kb gene cluster containing 18 genes from *Escherichia coli* strain CFT073 into a plasmid (Figure 5-3).[30,31] Colonies of *E. coli* Top10 transformed with this plasmid exhibited a strong yellow pigment. The plasmid was maintained using antibiotic resistance cassettes and the transformed strain was used to isolate the gene product from family 1, APE_{EC}. The product from family 2, APE_{VF}, was isolated from *Vibrio fischeri* strain ES114 WT. The gene cluster was confirmed by knocking out the gene cluster from the WT

strain and heterologous expression. The strains described above were shipped to UCSC for growth and chemical characterization.

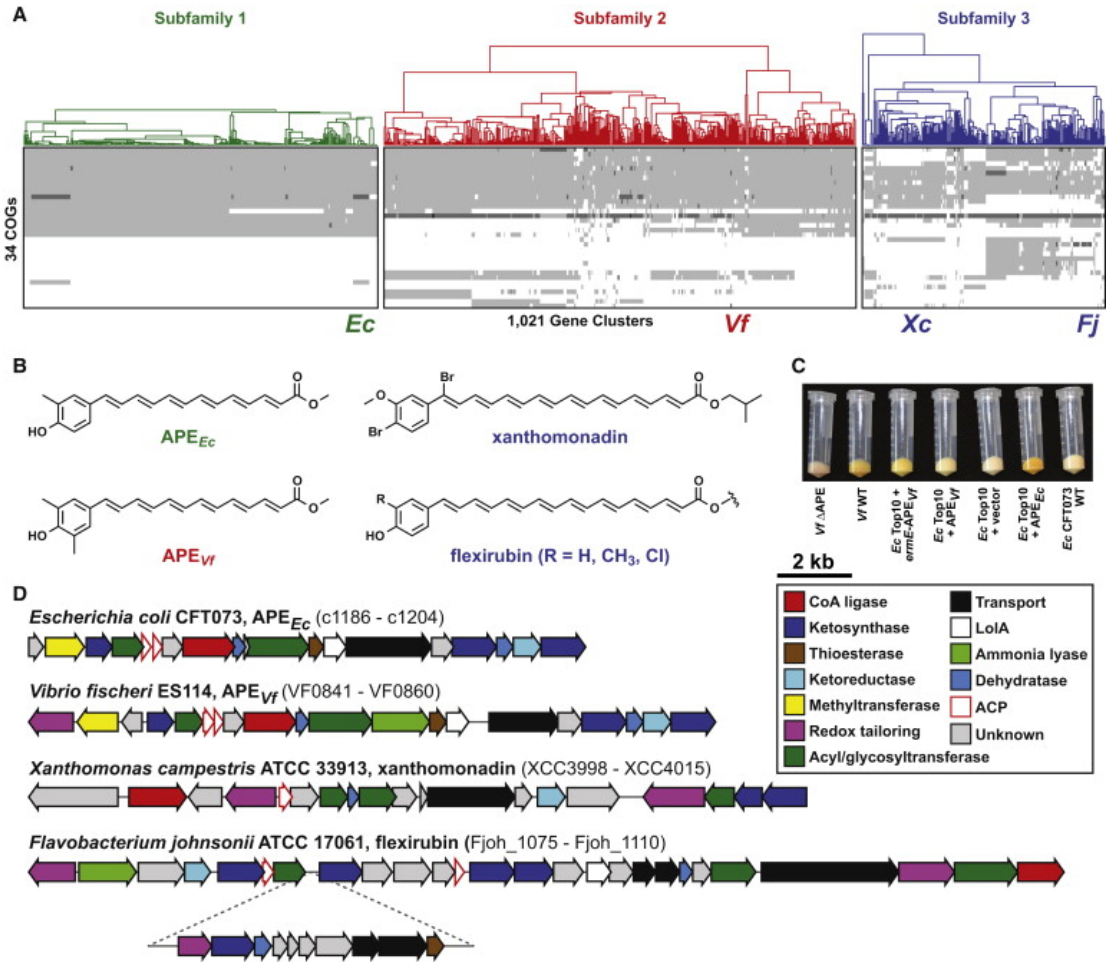


Figure 5-3: The subfamily identification, gene cluster analysis, pigment confirmation, and small molecule product structures of the APEs. (A) The three subfamilies of 1,021 gene clusters in the APE family divided into the three subfamilies. The heatmap represents the presence of Clusters of Orthologous Groups (COGs) generated by OrthoMCL[32] using the adapted distance metric[19] where grey represents one COG and dark grey represents the presence of two or three COGs. The locations of the clusters from *E. coli* CFT073, *V. fischeri* ES114, *Xanthomonas campestri*, and *Flavoacterium johnsonii*, are indicated. (B) Structures of the new APE_{EC}, APE_{VF}, xanthomonadin, and flexirubin. (C) Cell pellets from the strains used for the isolation and confirmation of the APE gene clusters. *V. fischeri* WT, *E. coli* Top10 expressing APE_{VF}, and the *E. coli* Top10 expressing multiple copies of the *E. coli* CFT073 APE gene cluster all appear yellow, while vector controls and knockouts of the same strains do not show significant yellow pigmentation. (D) The gene cluster blueprints with protein segments labeled for the four organisms highlighted in part A of this figure. The collapsed region in the flexirubin gene cluster represents the alkyl tail of the molecule not shown in part B. Adapted from Cimermancic *et al.*[24]

5.2.1. HPLC Based Polyene-Production Control Experiment:

In order to determine if the compounds from subfamilies 1 and 2 were being produced in culture and to optimize the extraction conditions, medium scale 50 mL cultures were grown and extracted. Consistent with the results displayed in Figure 5-3-C the *V. fischeri* ES114 and *E. coli* Top10 expressing the CFT073 APE gene cluster clearly produce pigment based on HPLC-DAD monitoring 441 nm Figure 5-4. With the small scale extraction protocol in hand, large scale cultures could be started and prepared for APE isolation.

5.2.1.1. HPLC Based Polyene-Production Experimental:

Escherichia coli strains were grown in LB media buffered with 50 mM TRIS at pH 7.5. Kanamycin and ampicillin were added at a final concentration of 50 μg/mL to maintain plasmids, and when necessary 1.5% agar was added to make solid media. *V. fischeri* ES114 and *Vibrio fischeri* ES114 ΔAPE were grown in LBS media made with 50 mM TRIS buffer pH 7.5, 20 g of sodium chloride, 3 mL of glycerol, 10 g of tryptone, and 5 g of yeast extract per liter of media. *Xanthomonas campestri* was cultured in NYGB media made from 5 g of peptone, 3 g of yeast extract, and 20 g of glycerol. *E. coli* strains were grown at 37 °C and shook at 250 rpm. *V. fischeri* strains and *X. campestri* were grown at 30 °C and shook at 120 rpm. Each organism was inoculated into 10 ml of the appropriate media in a 50 mL culture tube and incubated for 8h. Then a 10% inoculum was used to inoculate 50 mL of media. Cultures were grown to saturation at 24 h, transferred to 50 mL falcon tubes, and spun down at 7830

rpm for 20 min. The supernatant was carefully decanted, the cell pellets were frozen in liquid nitrogen, and lyophilized.

The dried cells were extracted by transferring them to a 20 mL amber scintillation vial containing 5.0 mL of 2:1 methylene chloride/methanol and stirring them vigorously. After 1 h, 2.0 ml 0.5 M potassium hydroxide was added and stirred for 1 h to hydrolyze the pigments. The solutions were neutralized with 2.0 M sulfuric acid, the organic layer was collected by liquid/liquid separation in a 50 mL separatory funnel and then dried over sodium sulfate. The solution was filtered and dried under a stream of nitrogen. The dried extracts were suspended in 0.2 mL of acetone (HPLC grade), transferred to 0.5 mL eppendorf tubes, spun down for 5 min at 3,000 rpm, and transferred to an insert in an amber LCMS vial.

Samples were subjected to a gradient of 30% acetonitrile in water with 0.02% formic acid to 90% acetonitrile over 20 min, after a 2 min initial hold, and then a 3 min 100% acetonitrile wash. Samples were injected on a C_{18} RP 2.6 micron particle size 100 mm x 4.6 mm Kinetix column (Phenomenex, USA) using an Agilent 1200 HPLC with diode array detector.

5.2.2. Special Considerations for Structure Elucidation:

Difficulties in the isolation of related aryl-polyenes from *Lysobacter enzymogenes* are well known.[33] To date structure elucidation efforts for this class of compounds have relied primarily on infrared spectroscopy (IR), ultraviolet spectroscopy (UV), mass spectrometry (MS), and some chemical manipulations, but due to the light sensitivity and limited material, no NMR spectra have previously

been reported.[27] By developing isolation conditions that rigorously exclude exposure to light, we isolated sufficient material to complete the first solution NMR characterization of a molecule of this type, and have confirmed all elements of the structure elucidation through careful and exhaustive examination of 1D and 2D NMR spectra.

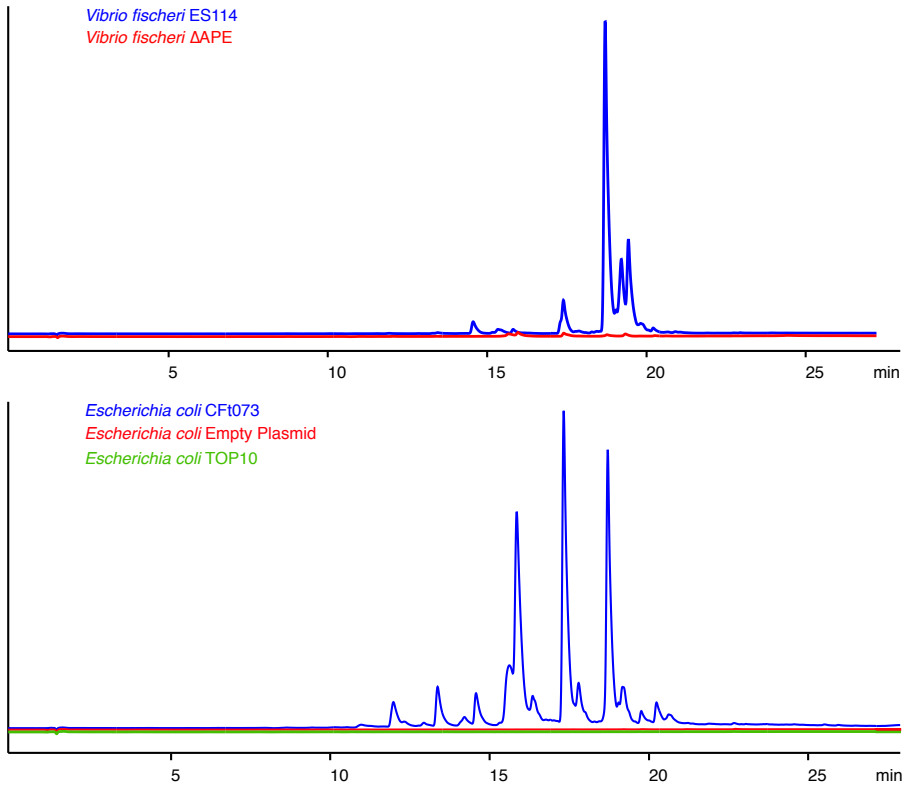


Figure 5-4: HPLC injections monitoring 441 nm for the presence of APE production. (Above) *V. fischeri* ES114 WT extract and *V. fischeri* ES114 ΔAPE extract overlaid. (Below) Extracts of *E. coli* TOP10 with heterologous expression of the *E. coli* APE gene cluster, an empty plasmid, and without plasmid. The chromatograms and the color of the cells in Figure 5-3 show that the expression of APE is dependent on the presence of the identified gene clusters.

5.2.3. The Aryl Polyene from Family 1, *E. coli*:

The pure APE_{EC}, a red amorphous powder, was determined to have a molecular formula of $C_{21}H_{22}O_3$ based on the observation of the $[M-H]^-$ adduct at 321.1496 *m/z* (Δppm = -0.310) and analysis of one and two-dimensional NMR experiments. The chemical formula was confirmed based on ^1H NMR (Figure 5-10) and HSQC (Figure 5-13) assignment of 15 aromatic and vinylic protons, one aromatic methyl singlet, one methoxy singlet, and one potential broad singlet phenolic proton at 8.43 ppm (**Figure 5-5**). From the TOCSY (Figure 5-16) spectrum it was clear that the molecule contained two independent spin systems. One spin system was defined as a phenyl ring with a 1,2,4 substitution pattern based on classical H18-H19 ortho-coupling constants ($^3J_{HH}$ = 7.2 Hz), meta-coupling between H15 and H19 ($^4J_{HH}$ = 2.1 Hz), and HMBC (Figure 5-14) correlations from H19 and H20 to C17, H19 and H20 to C15, the aromatic methyl singlet to C15 and C16, and the phenolic proton to C17 and C16 (**Figure 5-6**). The second spin system was defined as a long conjugated polyene terminating at a methyl ester and the 1,2,4-phenyl ring. The terminus of the polyene chain at the phenyl ring was identified based on HMBC signals from H15 and H19 to C13 as well as ROESY (Figure 5-15) signals between H15 and H13, and H19 and H13.

The methyl ester was identified via an HMBC correlation from the singlet methoxy proton signal at 3.70 ppm to the quaternary carbon C1 at 167.7 ppm. Protons H2 (doublet, ^1H 5.93 ppm $^3J_{HH}$ = 15.2 Hz; ^{13}C 120.5 ppm) and H3 (doublet of doublets, ^1H 7.33 ppm $^3J_{HH}$ = 15.1, 11.4; ^{13}C 145.1 ppm) displayed strong COSY

(Figure 5-12) correlations to one another, and both possessed HMBC correlations to the ester carbonyl at C1. These chemical shifts and coupling constants are indicative of the presence of an alpha-beta unsaturated ester. The assignment of the polyene chain continued through H4 based on HMBC and COSY correlations. Of the remaining C_8H_7 one quaternary carbon is contained in the phenyl ring connecting the aromatic functionality to the polyene, leaving the remaining constituents (C_7H_7; all between ^{1}H 6.85 – 6.40 ppm and ^{13}C 126 – 138 ppm) as a contiguous all-*trans* polyene chain connecting the aromatic head group with the methyl ester tail. The all-*trans* configuration is suggested by the absence of the '*cis* peak' centered around 340 nm in the UV spectrum that is a diagnostic marker for alkene chains that possess at least one region of non-linear (angulated) region of lesser symmetry, caused by the presence of *cis*-olefin(s).[34]

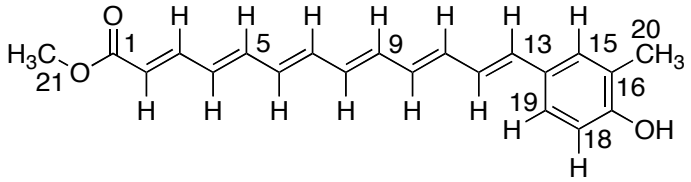


Figure 5-5: The structure of the APE_{EC} polyene.

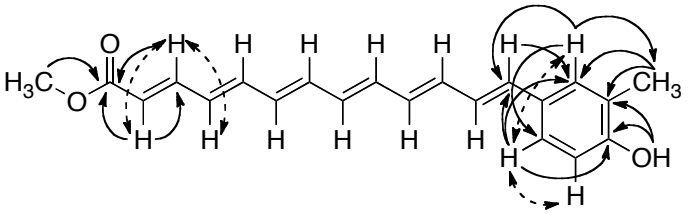


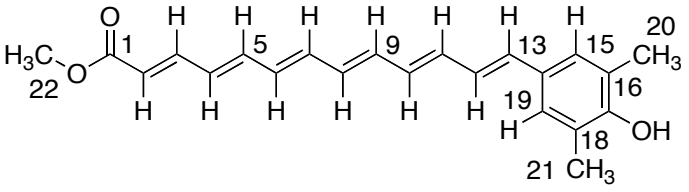
Figure 5-6: The structure of APE_{EC} with COSY (dashed lines) and HMBC (solid lines) correlations.

5.2.4. Aryl Polyene from Family 2, *V. fischeri*:

APE_{VF}, a red amorphous powder, was determined to have a molecular formula of $C_{22}H_{24}O_3$ based on the observation of the $[M-H]^-$ adduct at 335.1652 *m/z* (Δppm = 0.0) and analysis of one and two-dimensional NMR experiments (Figure 5-7). Comparison of the NMR spectra in acetone-d6 to that of APE_{EC} in acetone-d6 indicated that the polyene segments of the two molecules were very similar based on related chemical shifts (Figure 5-10 and Figure 5-17). To alleviate solubility issues, one and two-dimensional experiments were repeated in DMSO-d6. The alpha-beta unsaturated methyl ester motif was assigned based on both COSY (Figure 5-19) correlations between H2 (doublet, ^{1}H 5.97 ppm $^3J_{HH}$ = 15.2 Hz) and H3 (doublet of doublets, ^{1}H 7.31 ppm $^3J_{HH}$ = 15.2, 11.5) and HMBC (Figure 5-21) signals from both H2 and H3 to C1 at 166.7 ppm, as well as HMBC correlation from the methoxy proton singlet at 3.66 ppm to ester carbonyl C1 (Figure 5-8). As with the previous structure assignment, H4 was assigned based on COSY (Figure 5-19) correlation to H3 and HMBC correlations to C2 and C3. While signal overlap complicated interpretation of the COSY spectrum, H5 could be assigned based on HMBC

correlations to C4 and C3 as well as an HMBC correlation to C5 from H3 (assigned in conjunction with HSQC data, Figure 5-7 and Figure 5-8).

The one aromatic singlet in the downfield region of the spectrum (H15, H19; ^{1}H 7.06 ppm; ^{13}C 126.7 ppm) integrated for two protons, suggesting a 1,2,4,6-tetra-substituted symmetric aromatic group. The aromatic methyl singlet (^{1}H 2.15 ppm; ^{13}C 16.3 ppm) integrating for six protons and the phenol signal at 8.47 ppm suggested para substitution of the polyene and phenolic OH moieties, with the methyl groups either ortho or meta to the phenolic OH. An HMBC correlation from singlet aromatic protons H15 and H19 to C13, coupled with through space ROESY (Figure 5-22) correlations between H13 and H15/H19 proved that the substitution pattern of the phenol was 1,2,4,6 substituted. As with the previous structure assignment, completion of the structure elucidation was accomplished by consideration of the remaining double bond equivalents and the chemical shifts for the ^{1}H and ^{13}C resonances for the remaining atoms, which unequivocally determined that the aromatic head group and the methyl ester tail be connected via a linear polyene chain.



Chemical Formula: $C_{22}H_{24}O_3$
Exact Mass: 336.1725

Figure 5-7: The structure of the APE_{VF}.

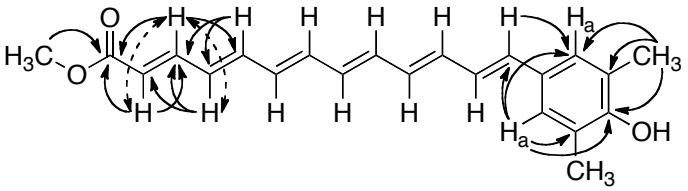


Figure 5-8: The structure of the APE_{VF} derived polyene with COSY (dashed lines) and HMBC (solid lines) correlations.

5.2.5. All-trans Conjugated Polyenes:

Consideration of the UV-profiles of the isolated peaks APE_{EC} and APE_{VF} with previously reported data on alpha and beta carotenoids suggests and all-*trans* structure for both molecules.[34-37] A *cis*-double bond within extended polyene chains breaks the linearity of the molecule, resulting in a shorter chain and new absorption axis. The result is what is known as a *cis*-peak in the UV spectra between 310 and 370 nm. In both the *V. fischeri* and *E. coli* UV-profiles there is little or no absorbance between 310-370 nm, indicating all-*trans* configurations for both structures (Figure 5-9).

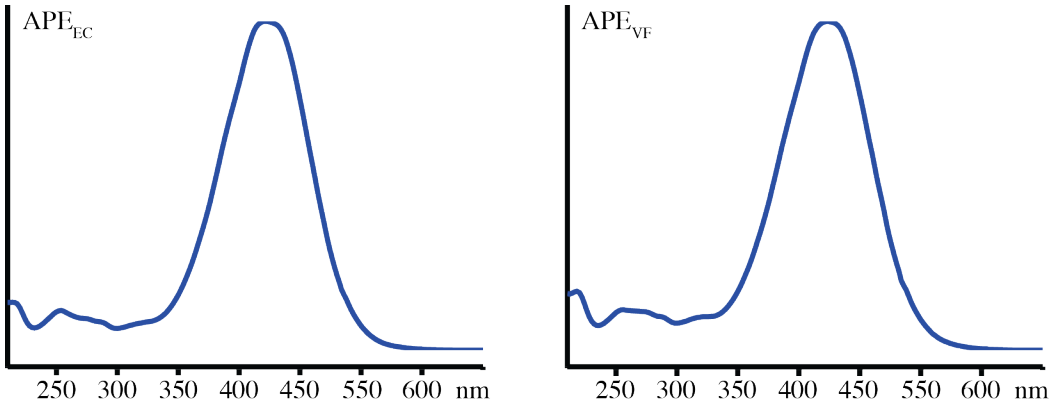


Figure 5-9: The UV-Vis absorbance spectrum for APE_{EC} and APE_{VF} without the presence of the *cis*-peak from 310 to 370 nm.

5.2.6. Mass Spectrometry:

Compounds were analyzed on an Agilent UPLC-ESI-TOF mass spectrometer, comprising a 1260 binary pump in low dwell volume mode, an Agilent column oven heated to 45°C, and an Agilent 6230 Time-of-flight Mass Spectrometer with an electrospray ionization (ESI) source. A sample of 1 μL, dissolved in 50% v/v methanol/water, was injected onto a 1.8 μm particle size, 50 x 2.3 mm I.D. Zorbax RRHT column. Each sample was subjected to a MeCN:H_2O gradient from 10% to 90% MeCN over 4 min followed by 1.5 min at 90% MeCN at a flow rate of 0.8 mL min^{-1}. Formic acid, 200 μl/L, was added to both the water and the acetonitrile. Water, 1 mL min^{-1}, was added to the acetonitrile. The mass spectrometer was run with a detector mass range of 100 to 1700 *m/z*. The ESI source was operated with a desolvation temperature of 350 C and a drying gas flow rate of 11 L min^{-1}. The fragmentor voltage was held at 135 V. In positive ESI mode, the capillary voltage was ramped from 2500 V at 0 min to 2750 V at 1 min, and to 3000 V at 3 min. In

negative ESI mode, the capillary voltage was held at 2750 V. Each sample was run in high-resolution (4GHz) detector mode.

5.3. Growth and Purification:

5.3.1. Fermentation of *E. coli* Strains:

Cultures were grown in LB Broth Miller from Fisher containing tryptone 10 g, yeast extract 5 g, sodium chloride 10 g, buffered with 50 mM TRIS at pH 7.5 per L of media. After autoclaving the media and letting it cool to 60 °C, kanamycin and ampicillin were added via sterile filtration at final concentrations of 50 μg/mL to maintain plasmids. When necessary, 1.5% agar was added to prepare solid media. For large-scale preparation the following growth process was repeated eight times, 4 L per iteration, to produce a total of 32 L of culture. Bacteria were grown on solid media at 37°C overnight after streaking them on solid media. Colonies were used to inoculate 10 mL of media in a 50 mL culture tube. Cultures were grown in the dark at 37°C and shaken at 250 rpm. After 8 h the small-scale culture was used to inoculate 100 mL of antibiotic-containing media in a 250 mL wide neck Erlenmeyer flask and grown under the same conditions overnight. Finally 50 mL of this medium-scale culture was used to inoculate 1 L that was subsequently grown for 3 days, spun down at 4000 rpm at 4 °C for 20 min, transferred to a 50 mL falcon tube, and lyophilized. After the cells were spun down, all the subsequent steps were conducted in the dark with the use of red LED headlamp.

5.3.2. Fermentation of *V. fischeri* Strains:

Vibrio fischeri strain ES114 was grown in LBS media: 10 g tryptone, 5 g of yeast extract, 20 g of NaCl, 3 mL of glycerol, buffered in 50 mM Tris-HCl at pH 7.5 per L of media. When necessary, 1.5% agar was added to prepare solid media. Colonies were struck out on solid plates and left overnight at RT. Single colonies were inoculated into 7 mL of media in a 40 mL culture tube and shaken at RT overnight at 100 rpm. This small scale was inoculated into 50 mL of media in a wide neck Erlenmeyer flask for 18 h at which time 35 mL was transferred into 1.0 L of media in a 2.8 L wide neck flask and shaken for 3 days. The cells were pelleted and lyophilized before extraction.

5.3.3. Extraction:

The same process was used to extract both *E. coli* and *V. fischeri* separately. The dried cell pellets were split into two 1 L Erlenmeyer flasks containing 500 mL of 1:2 methanol/dichloromethane, shaken for 1 h at 180 rpm, stirred vigorously with a magnetic stir bar for 1 hour, then vacuum filtered, and the solution concentrated to dryness under vacuum. The cell debris was re-extracted three times in this fashion and all extracts for each strain were combined into a 1 L round bottom flask. The dried extract was suspended in 400 mL of 1:2 methanol/dichloromethane at room temperature. A saponification reaction was performed on each extract by stirring the solution rapidly with a magnetic stir bar and adding 200 mL of 0.5 M potassium hydroxide. The reaction was carried out for 1 h at which time the mixture was neutralized with 2.0 M sulfuric acid to pH 7.0 and transferred to a 2 L separatory

funnel. The organic layer was collected, washed three times with brine, once with deionized water, dried over sodium sulfate, transferred through a paper filter into a 500 mL round bottom flask, and concentrated to dryness under vacuum. The dried extracts were suspended in 10 mL of acetone and carried forward to purification.

5.3.4. Purification:

E.coli materials were purified on RP-HPLC using a two step purification protocol. Firstly, crude material was purified on a semi-prep RP column (Phenomenex Synergi Fusion-RP, 250 x 10 mm, 10 μm) using a gradient of acetonitrile MeCN:H_2O + 0.02% formic acid (32% MeCN for 26 min, 100% MeCN for 9 min, 20% MeCN for 2 min, and a 9 minute re-equilibration) at a flow rate of 4 mL min^{-1}. The peak eluting at 16 min displaying a strong UV absorbance at 441 nm was collected and re-purified using an analytical column (Phenomonex Kinetix 2.6 μm XB-C18 100 x 4.6 mm) using a gradient of MeCN:H_2O + 0.02% formic acid (50% MeCN for 2 min, 50%-65% MeCN over 20 min) at a flow rate of 2 mL min^{-1}. APE_{EC}, the peak eluting at 16 min that displayed the correct UV spectra, was collected, dried under vacuum, and stored at -20°C in a 5 mL amber vial. Standard one and two-dimensional NMR experiments were performed on a Varian 600 MHz cryoprobe NMR in acetone-d6.

The *V. fischeri* extract was first purified by RP-HPLC analytical column (Phenomonex Kinetex 5μm XB-C18 250 x 4.6 mm) using a gradient of MeCN:H_2O + 0.02% formic acid (50%-60% MeCN 2 min, 60%-73.8% MeCN over 11 min, 73.8%-95% over 1 min, 95%-100% over 3 min, 100% for 1 min) at a flow rate of 2 mL min-

[1]. The peak at 9.5 min with absorbance at 441 nm was collected and re-purified on an analytical column (Phenomonex Synergi 10 μm Fusion-RP 250 x 4.6 mm) using a gradient of methanol (MeOH):H_2O + 0.02% formic acid (50% MeOH for 2 min, 50%-90% MeOH over 15 min, 100% MeOH for 2 min) at a flow rate of 2 mL min^{-1}. APE_{VF}, the peak eluting at 18 min that displayed the correct UV spectra, was collected, dried under vacuum, and stored at -20°C in a 5 ml amber vial. Standard one and two-dimensional NMR experiments were performed on a Varian 600 MHz cryoprobe NMR in both acetone-d6 and DMSO-d6.

5.4. NMR Spectra for APE_{EC} and APE_{VC}:

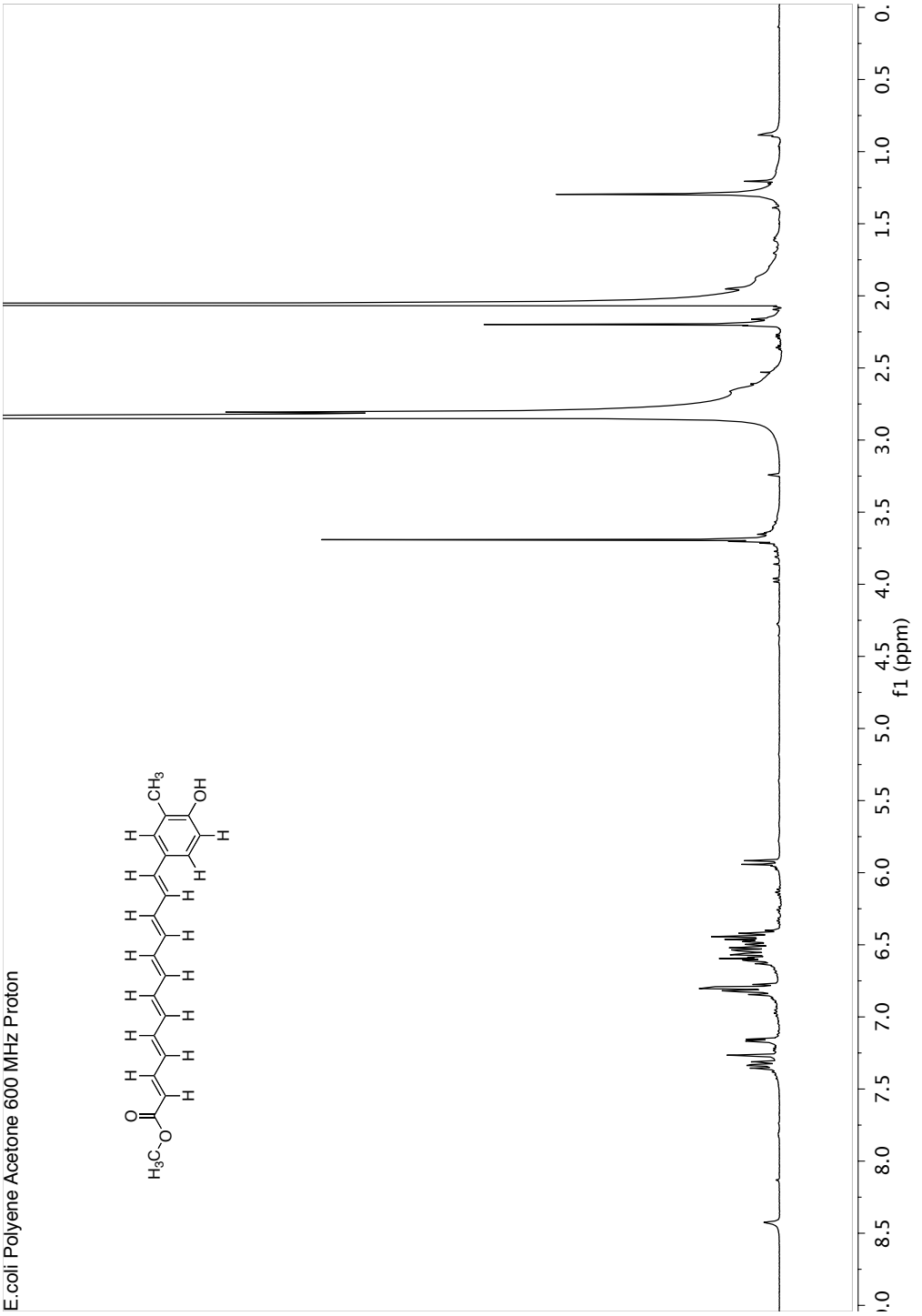


Figure 5-10: ^{1}H-NMR of APE_{EC} in acetone-d6.

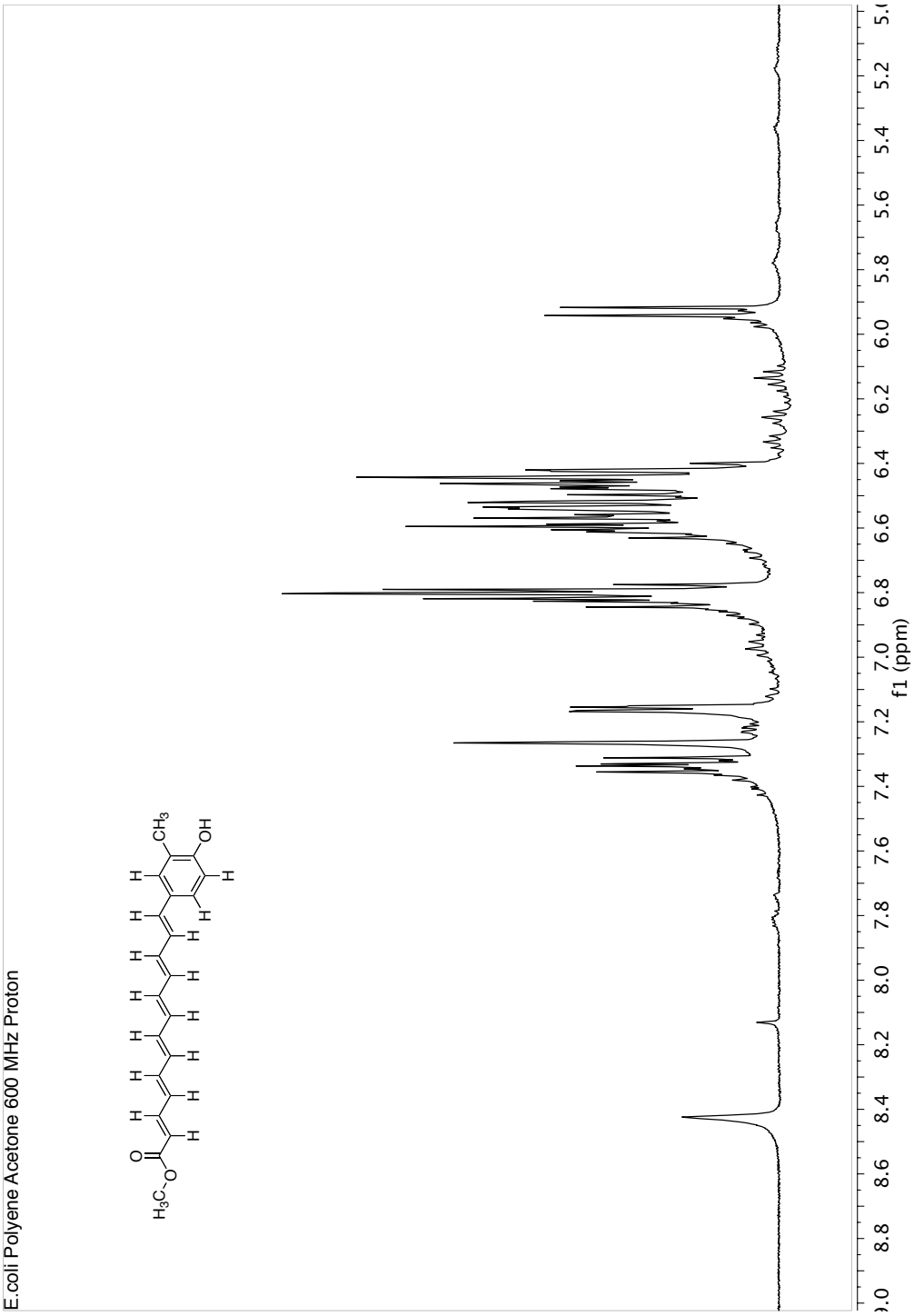


Figure 5-11: Expansion of the ^{1}H-NMR of APE_{EC} in acetone-d6.

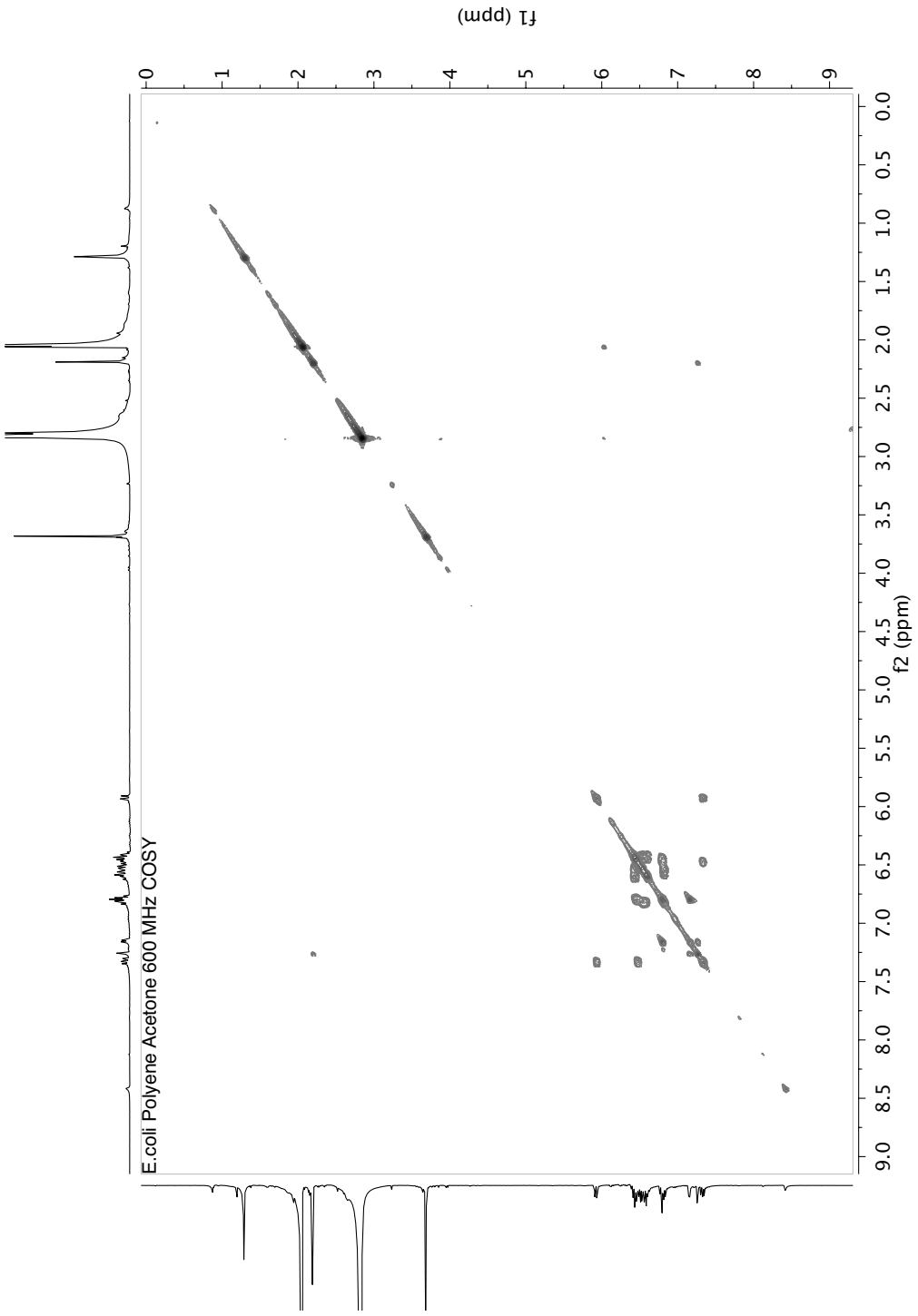


Figure 5-12: COSY of APE_{EC} in acetone-d6.

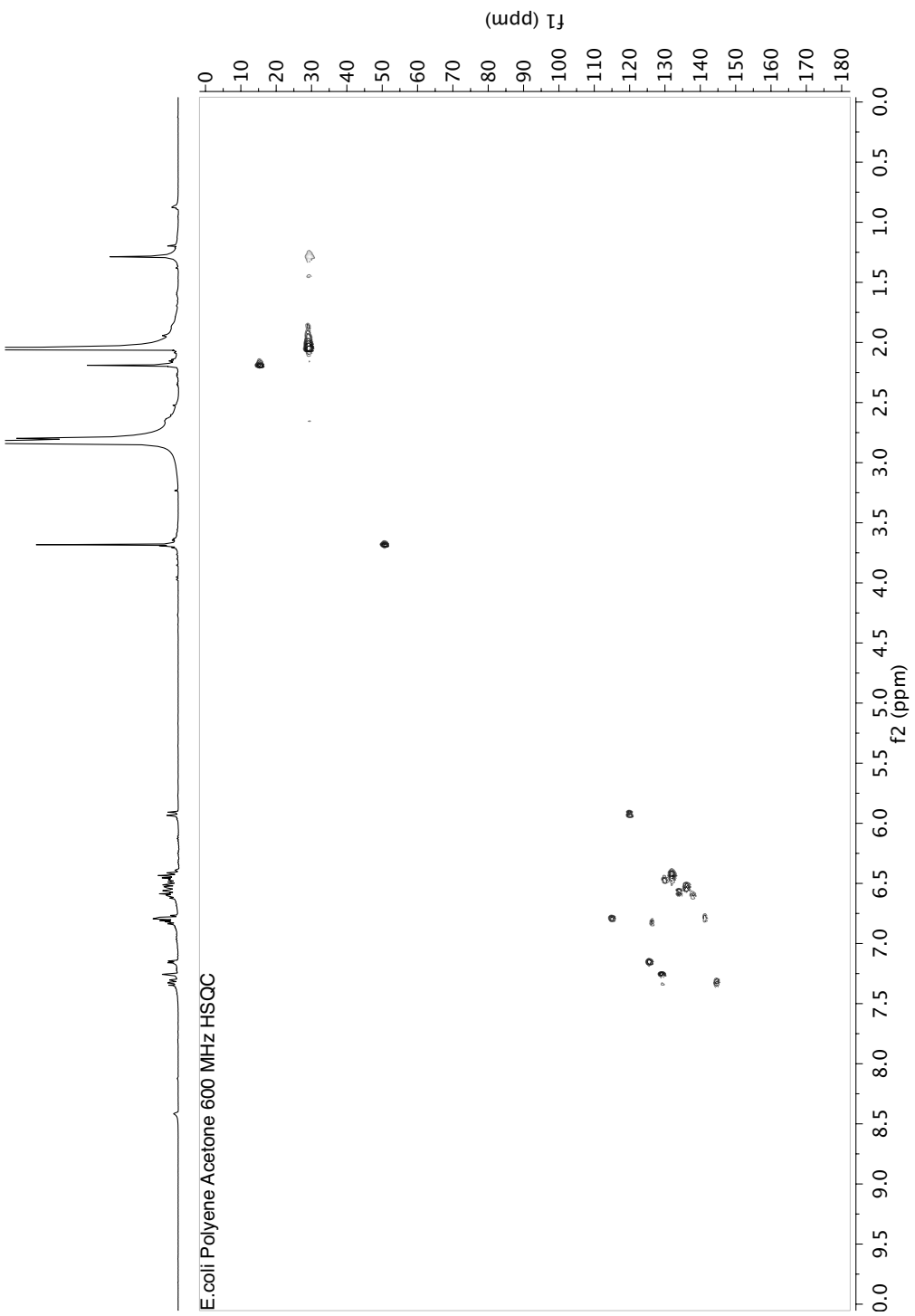


Figure 5-13: HSQC of APE_{EC} in acetone-d6.

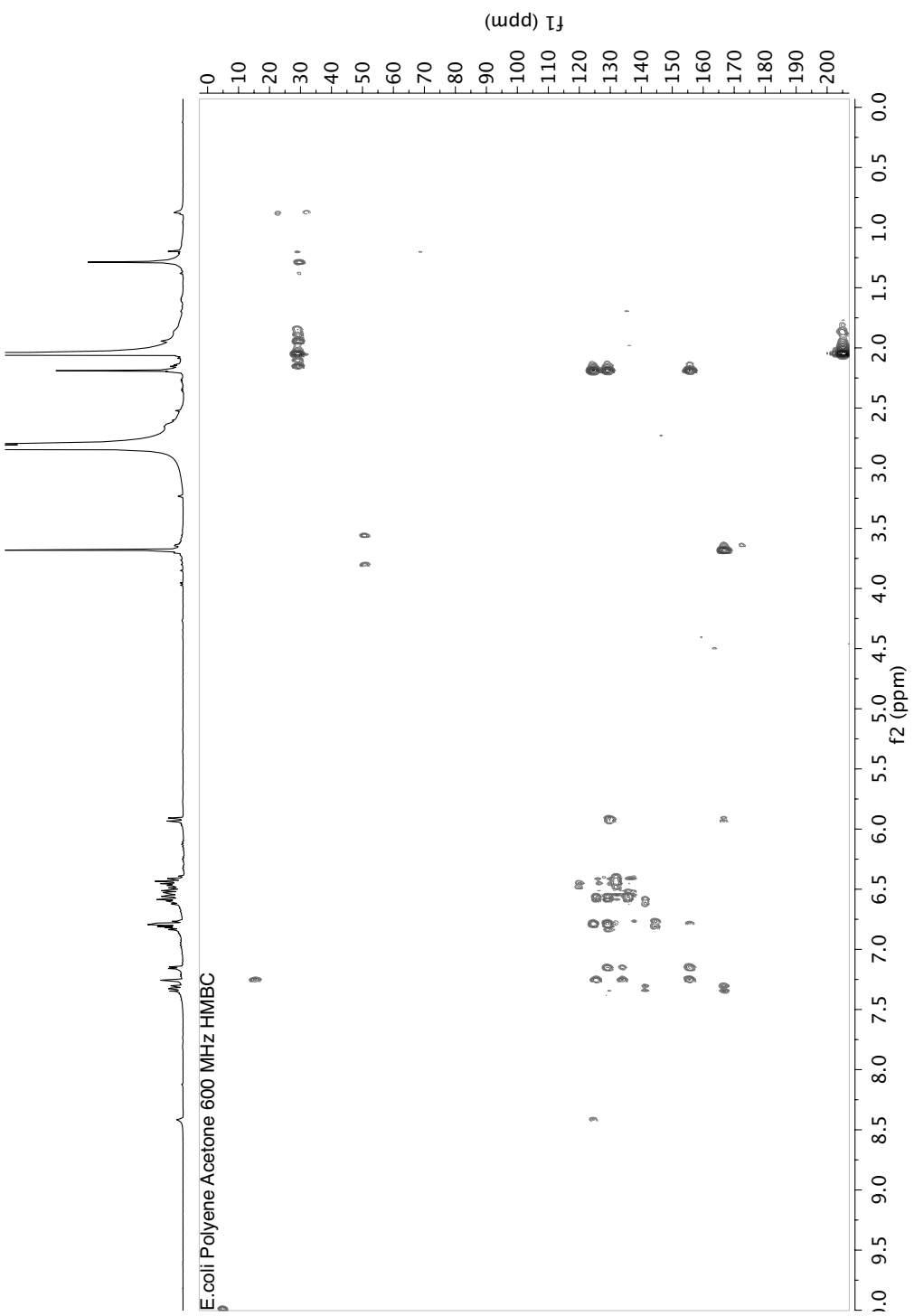


Figure 5-14: HMBC of APE_{EC} in acetone-d6.

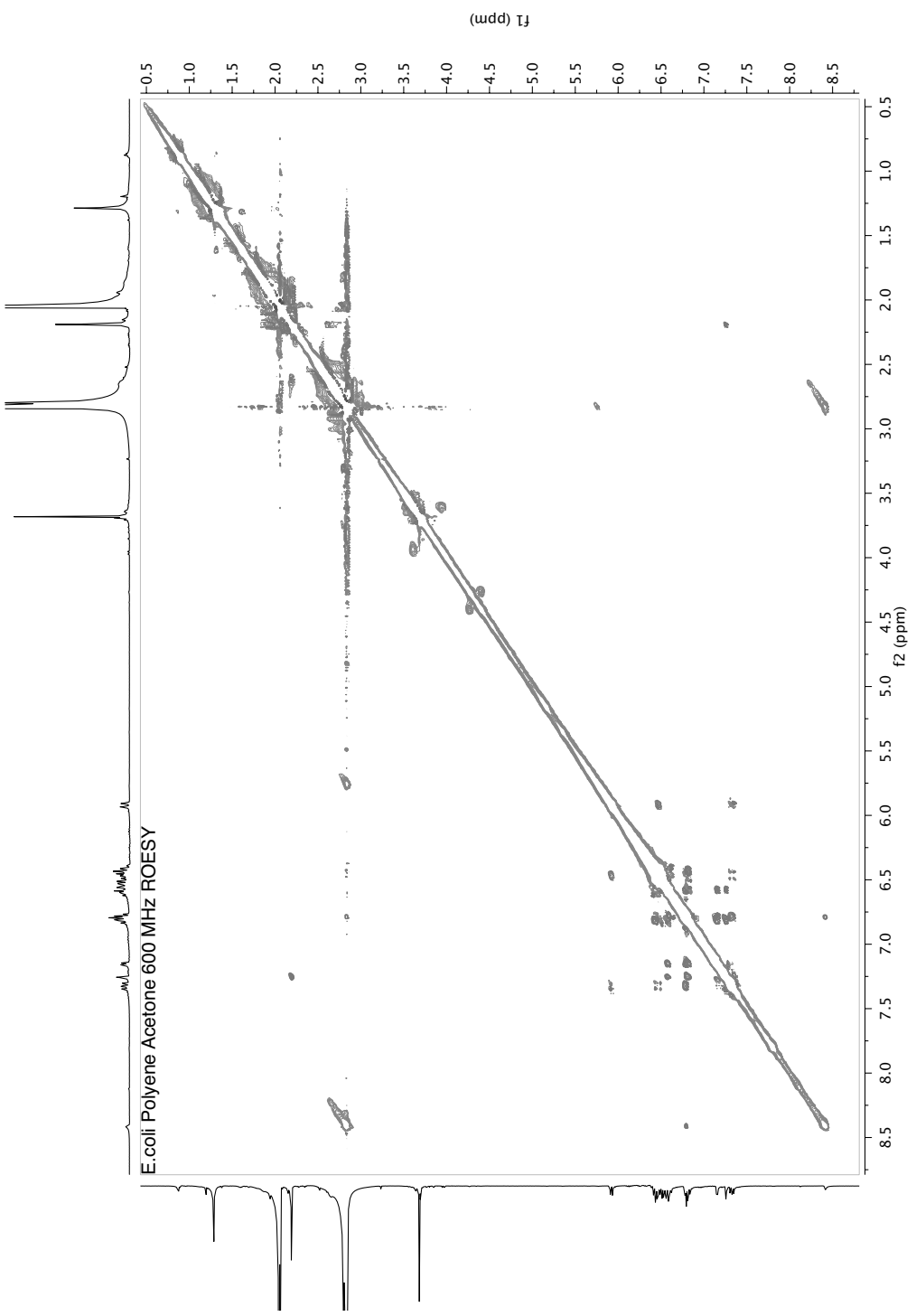


Figure 5-15: ROESY of APE_{EC} in acetone-d6.

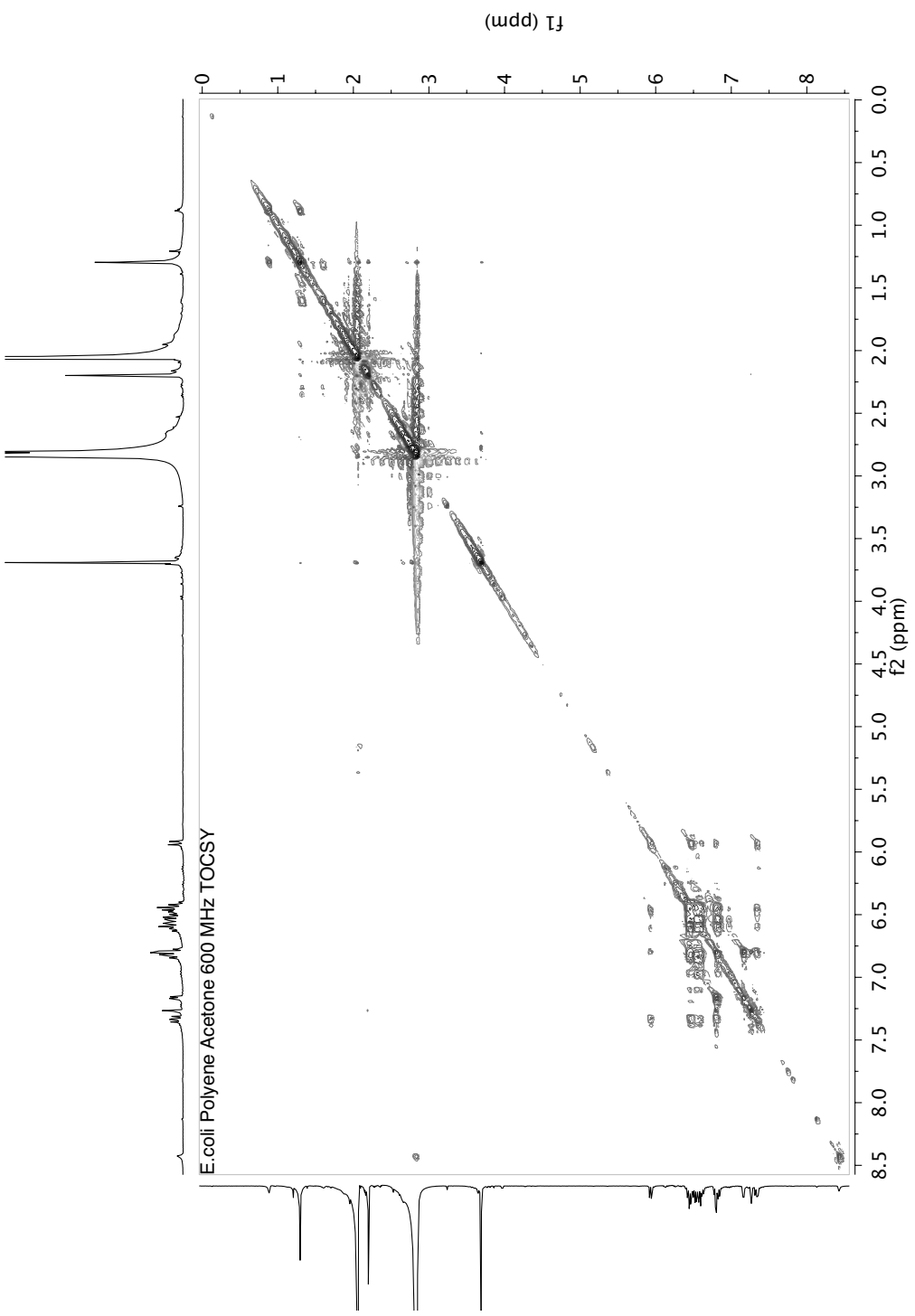


Figure 5-16: TOCSY of APE_{EC} in acetone-d6.

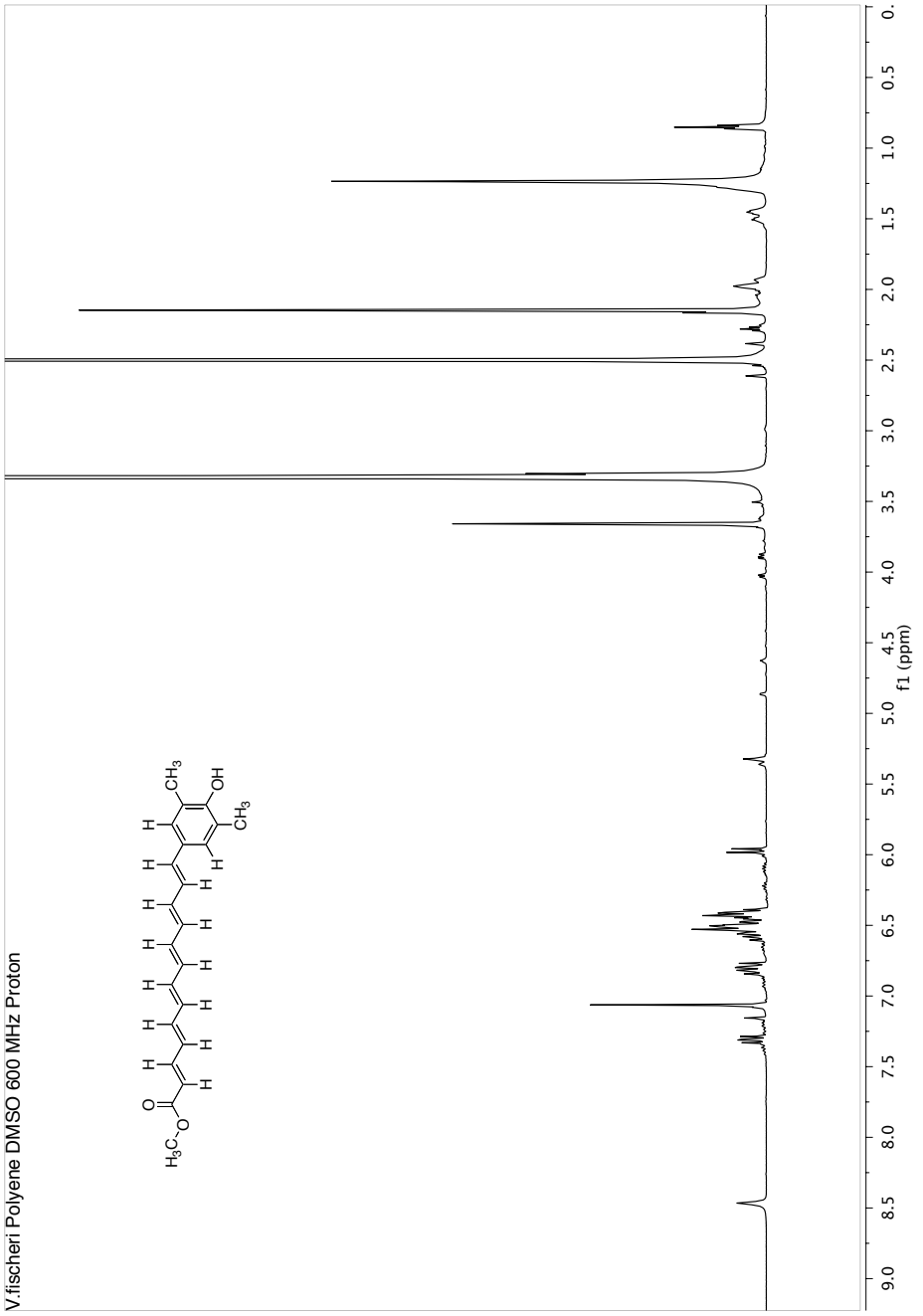


Figure 5-17: ^{1}H NMR of APE_{VF} in DMSO-d6.

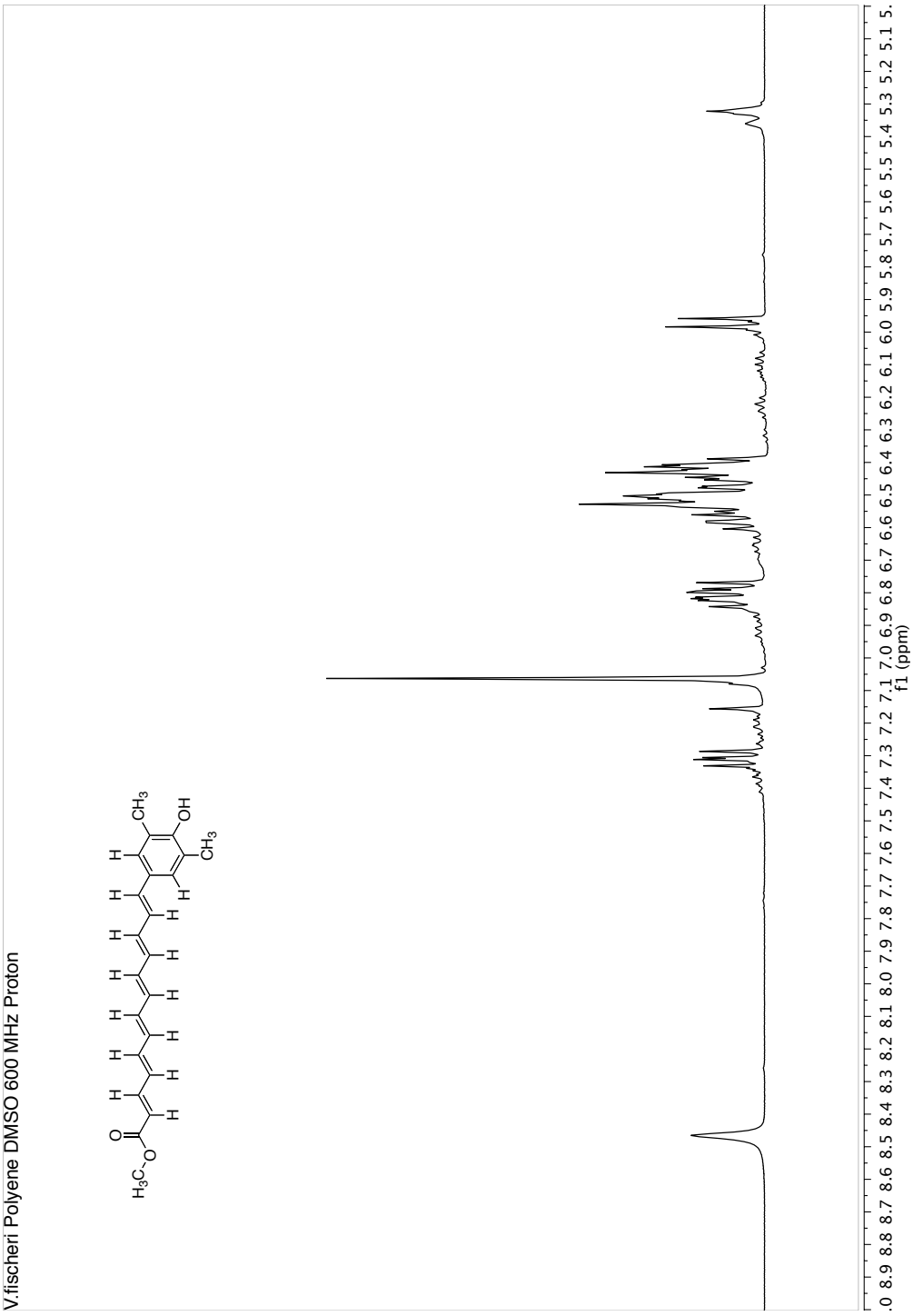


Figure 5-18: Expansion of the ^{1}H NMR of APE_{VF} in DMSO-d6.

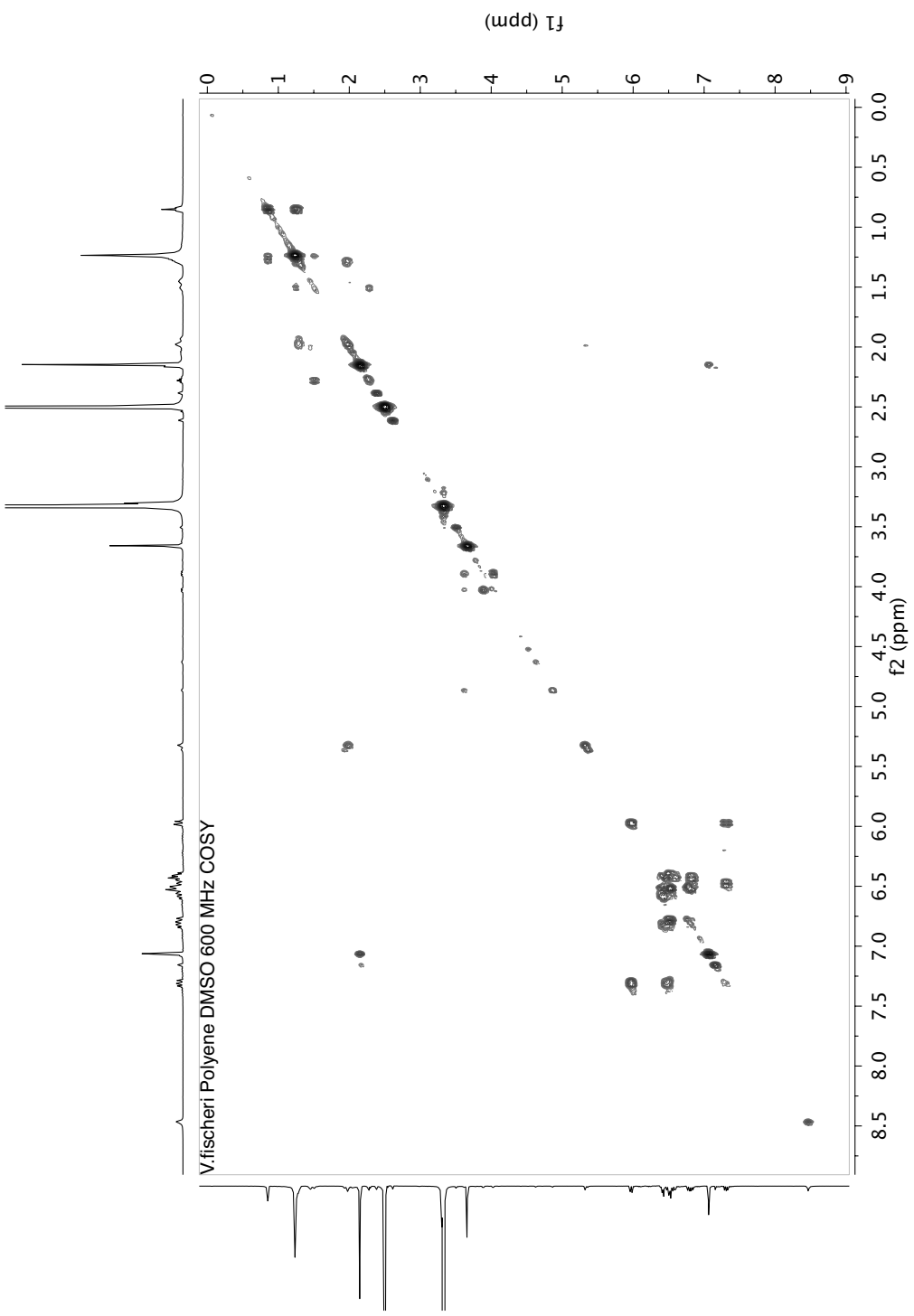


Figure 5-19: COSY of APE_{VF} in DMSO-d6.

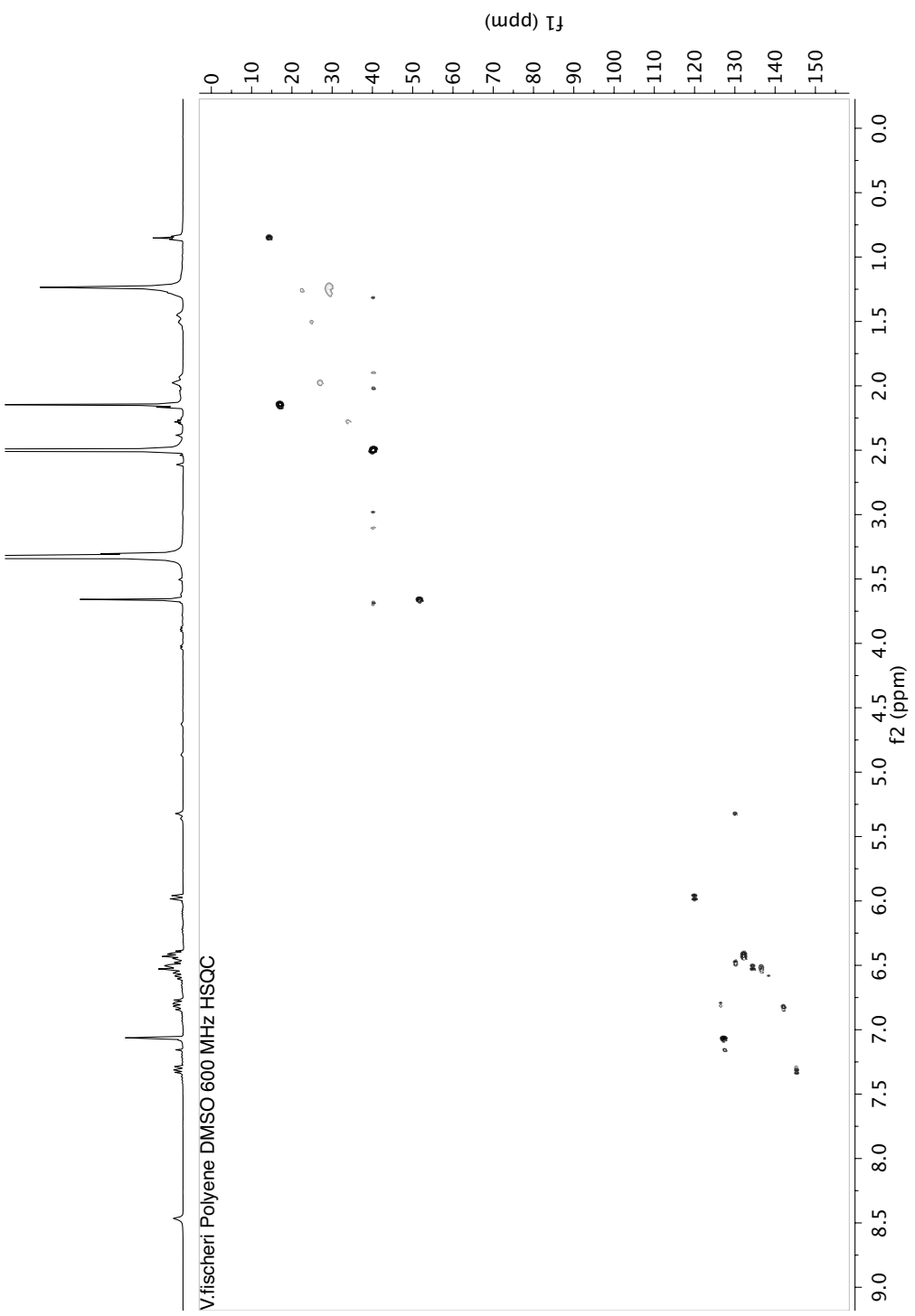


Figure 5-20: HSQC of APE_{VF} in DMSO-d6.

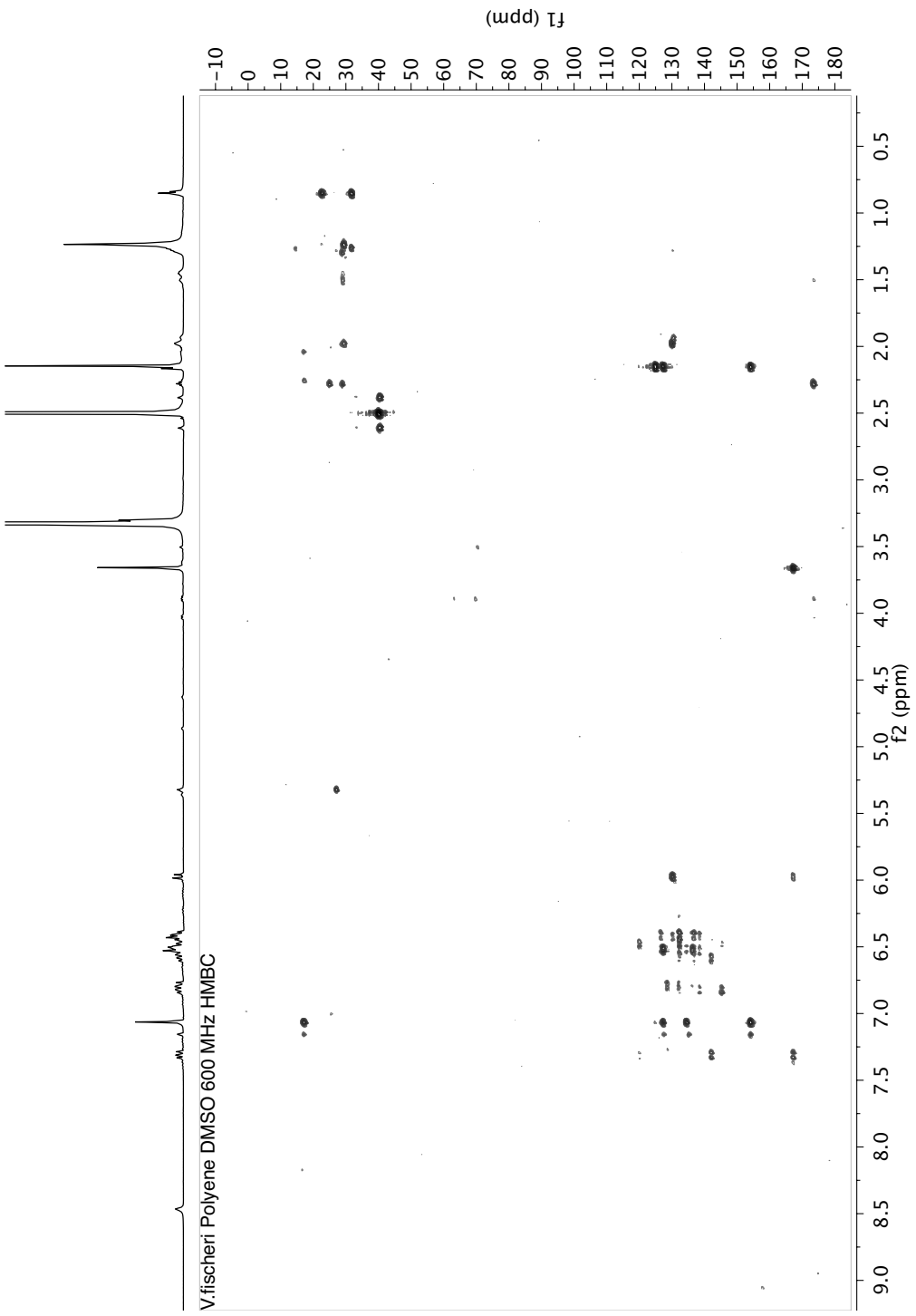


Figure 5-21: HMBC of APE_{VF} in DMSO-d6.

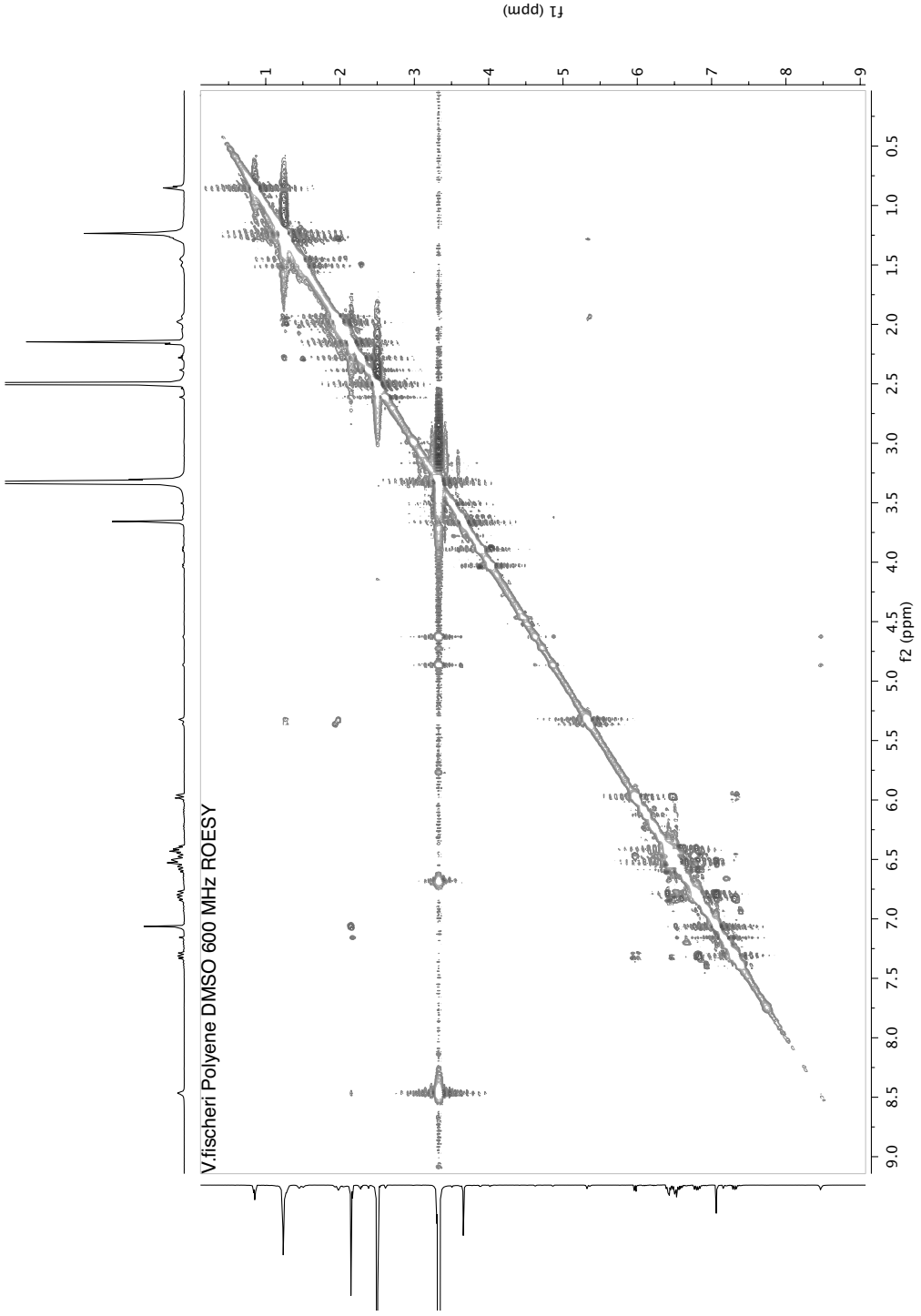


Figure 5-22: ROESY of APE_{VF} in DMSO-d6.

5.5. References:

(1) Donia, M. S.; Cimermancic, P.; Schulze, C. J.; Wieland Brown, L. C.; Martin, J.; Mitreva, M.; Clardy, J.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 1402–1414.

(2) Molinski, T. F. *Org. Lett.* **2014**, *16*, 3849–3855.

(3) Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Jiang, H.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. *Nature* **2013**, *496*, 528–532.

(4) Anzai, Y.; Iizaka, Y.; Li, W.; Idemoto, N.; Tsukada, S.-I.; Koike, K.; Kinoshita, K.; Kato, F. *J IND MICROBIOL BIOTECHNOL* **2009**, *36*, 1013–1021.

(5) Donia, M. S.; Ravel, J.; Schmidt, E. W. *Nat. Chem. Biol.* **2008**, *4*, 341–343.

(6) Rath, C. M.; Janto, B.; Earl, J.; Ahmed, A.; Hu, F. Z.; Hiller, L.; Dahlgren, M.; Kreft, R.; Yu, F.; Wolff, J. J.; Kweon, H. K.; Christiansen, M. A.; Håkansson, K.; Williams, R. M.; Ehrlich, G. D.; Sherman, D. H. *ACS Chem.*

Biol. **2011**, *6*, 1244–1256.

(7) Walsh, C. T.; Fischbach, M. A. *J. Am. Chem. Soc.* **2010**, *132*, 2469–2493.

(8) Baunach, M.; Franke, J.; Hertweck, C. *Angew. Chem. Int. Ed.* **2014**, n/a–n/a.

(9) Dutta, S.; Whicher, J. R.; Hansen, D. A.; Hale, W. A.; Chemler, J. A.; Congdon, G. R.; Narayan, A. R.; Håkansson, K.; Sherman, D. H.; Smith, J. L.; Skiniotis, G. *Nature* **2014**, *510*, 512–517.

(10) Bian, X.; Plaza, A.; Zhang, Y.; Müller, R. *J. Nat. Prod.* **2012**, *75*, 1652–1655.

(11) Weitnauer, G.; Mühlenweg, A.; Trefzer, A.; Hoffmeister, D.; Süssmuth, R. D.; Jung, G.; Welzel, K.; Vente, A.; Girreser, U.; Bechthold, A. *Chem. Biol.* **2001**, *8*, 569–581.

(12) Starcevic, A.; Zucko, J.; Simunkovic, J.; Long, P. F.; Cullum, J.; Hranueli, D. *Nucleic Acids Res.* **2008**, *36*, 6882–6892.

(13) Li, M. H.; Ung, P. M.; Zajkowski, J.; Garneau-Tsodikova, S.; Sherman, D. H. *BMC Bioinformatics* **2009**, *10*, 185.

(14) Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D. *Fungal Genetics and Biology* **2010**, *47*, 736–741.

(15) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; BREITLING, R. *Nucleic Acids Res.* **2011**, *39*, W339–W346.

(16) Blin, K.; Medema, M. H.; Kazempour, D.; Fischbach, M. A.; BREITLING, R.; Takano, E.; Weber, T. *Nucleic Acids Res.* **2013**, *41*, gkt449–W212.

(17) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee,

S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W.; BREITLING, R.; Takano, E.; Medema, M. H. *Nucleic Acids Res.* **2015**, gkv437.

(18) Finn, R. D.; Clements, J.; Eddy, S. R. *Nucleic Acids Res.* **2011**, *39*, W29–W37.

(19) Lin, K.; Zhu, L.; Zhang, D.-Y. *Bioinformatics* **2006**, *22*, 2081–2086.

(20) Challis, G. L. *ChemBioChem* **2005**, *6*, 601–611.

(21) Samuel, G.; Reeves, P. *Carbohydr. Res.* **2003**, *338*, 2503–2519.

(22) Walter, M. H.; Strack, D. *Nat. Prod. Rep.* **2011**, *28*, 663–692.

(23) Rehm, B. H. A. *Nature Publishing Group* **2010**, *8*, 578–592.

(24) Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Brown, L. C. W.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 412–421.

(25) Fuchs, S. W.; Bozhüyük, K. A. J.; Kresovic, D.; Grundmann, F.; Dill, V.; Brachmann, A. O.; Waterfield, N. R.; Bode, H. B. *Angew. Chem. Int. Ed.* **2013**, *52*, 4108–4112.

(26) McBride, M. J.; Xie, G.; Martens, E. C.; Lapidus, A.; Henrissat, B.; Rhodes, R. G.; Goltsman, E.; Wang, W.; Xu, J.; Hunnicutt, D. W.; Staroscik, A. M.; Hoover, T. R.; Cheng, Y.-Q.; Stein, J. L. *Applied and Environmental Microbiology* **2009**, *75*, 6864–6875.

(27) Andrewes, A. G.; Hertzberg, S.; Liaaen-Jensen, S.; Starr, M. P. *Acta Chem Scand* **1973**, *27*, 2383–2395.

(28) Andrewes, A. G.; Jenkins, C. L.; Starr, M. P.; Shepherd, J. *Structure of xanthomonadin I, a novel dibrominated aryl-polyene pigment produced by the bacterium Xanthomonasjuglandis*; Tetrahedron …, 1976.

(29) Goel, A. K.; Rajagopal, L.; Nagesh, N.; Sonti, R. V. *Journal of Bacteriology* **2002**, *184*, 3539–3548.

(30) Quan, J.; Tian, J. *PLoS ONE* **2009**, *4*, e6441.

(31) Quan, J.; Tian, J. *Nat Protoc* **2011**.

(32) Li, L.; Stoeckert, C. J.; Roos, D. S. *Genome Research* **2003**, *13*, 2178–2189.

(33) Wang, Y.; Qian, G.; Li, Y.; Wang, Y.; Wang, Y.; Wright, S.; Li, Y.; Shen, Y.; Liu, F.; Du, L. *PLoS ONE* **2013**, *8*, e66633.

(34) Baraldi, I.; Benassi, E.; Spalletti, A. *Spectrochim Acta A Mol Biomol Spectrosc* **2008**, *71*, 543–549.

(35) Jurkowitz, L.; Loeb, J. N.; Brown, P.; Wald, G. *Nature* **1959**, *184*, 614–624.

(36) Tsukida, K.; Saiki, K.; Takii, T.; Koyama, Y. *J. Chromatogr. A* **1982**, *245*, 359–364.

(37) Zechmeister, L.; Lemmon, R. M. *J. Am. Chem. Soc.* **2001**, *66*, 317–322.

Bibliography

Alley, M. C.; Scudiero, D. A.; Monks, A.; Hursey, M. L.; Czerwinski, M. J.; Fine, D. L.; Abbott, B. J.; Mayo, J. G.; Shoemaker, R. H.; Boyd, M. R. *Cancer Res* **1988**, *48*, 589.

Andrewes, A. G.; Hertzberg, S.; Liaaen-Jensen, S.; Starr, M. P. *Acta Chem Scand* **1973**, *27*, 2383–2395.

Andrewes, A. G.; Jenkins, C. L.; Starr, M. P.; Shepherd, J. *Structure of xanthomonadin I, a novel dibrominated aryl-polyene pigment produced by the bacterium Xanthomonasjuglandis*; Tetrahedron …, 1976.

Anzai, Y.; Iizaka, Y.; Li, W.; Idemoto, N.; Tsukada, S.-I.; Koike, K.; Kinoshita, K.; Kato, F. *J Ind Microbiol Biotechnol* **2009**, *36*, 1013–1021.

Anzai, Y.; Sakai, A.; Li, W.; Iizaka, Y.; Koike, K.; Kinoshita, K.; Kato, F. *J. Antibiot.* **2010**, *63*, 325.

Aramadhaka, L. R.; Prorock, A.; Dragulev, B.; Bao, Y.; Fox, J. W. *Toxicon* **2013**, *69*, 160.

Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A. *PNAS* **2007**, *104*, 6140.

Baraldi, I.; Benassi, E.; Spalletti, A. *Spectrochim Acta A Mol Biomol Spectrosc* **2008**, *71*, 543–549.

Basu, S.; Sachidanandan, C. *Chem. Rev.* **2013**, *113*, 7952.

Baunach, M.; Franke, J.; Hertweck, C. *Angew. Chem. Int. Ed.* **2014**, n/a–n/a.

Baur, S.; Niehaus, J.; Karagouni, A. D.; Katsifas, E. A. *Journal of …* **2006**.

Bian, X.; Plaza, A.; Zhang, Y.; Müller, R. *J. Nat. Prod.* **2012**, *75*, 1652–1655.

Bingol, K.; Bruschweiler Li, L.; Li, D.-W.; Brüschweiler, R. *Anal. Chem.* **2014**, *86*, 5494.

Blin, K.; Medema, M. H.; Kazempour, D.; Fischbach, M. A.; Breitling, R.; Takano, E.; Weber, T. *Nucleic Acids Res.* **2013**, *41*, gkt449–W212.

Bloch, F. *Science* **1953**, *118*, 425.

Bohni, N.; Cordero-Maldonado, M. L.; Maes, J.; Siverio-Mota, D.; Marcourt, L.; Munck, S.; Kamuhabwa, A. R.; Moshi, M. J.; Esguerra, C. V.; de Witte, P. A.

M.; Crawford, A. D.; Wolfender, J.-L. *PLoS ONE* **2013**, *8*, e64006.

Bouslimani, A.; Porto, C.; Rath, C. M.; Wang, M.; Guo, Y.; Gonzalez, A.; Berg-Lyon, D.; Ackermann, G.; Moeller Christensen, G. J.; Nakatsuji, T.; Zhang, L.; Borkowski, A. W.; Meehan, M. J.; Dorrestein, K.; Gallo, R. L.; Bandeira, N.; Knight, R.; Alexandrov, T.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E2120.

Buescher, J. M.; Liebermeister, W.; Jules, M.; Uhr, M.; Muntel, J.; Botella, E.; Hessling, B.; Kleijn, R. J.; Le Chat, L.; Lecointe, F.; Mäder, U.; Nicolas, P.; Piersma, S.; Rügheimer, F.; Becher, D.; Bessieres, P.; Bidnenko, E.; Denham, E. L.; Dervyn, E.; Devine, K. M.; Doherty, G.; Drulhe, S.; Felicori, L.; Fogg, M. J.; Goelzer, A.; Hansen, A.; Harwood, C. R.; Hecker, M.; Hubner, S.; Hultschig, C.; Jarmer, H.; Klipp, E.; Leduc, A.; Lewis, P.; Molina, F.; Noirot, P.; Peres, S.; Pigeonneau, N.; Pohl, S.; Rasmussen, S.; Rinn, B.; Schaffer, M.; Schnidder, J.; Schwikowski, B.; Van Dijl, J. M.; Veiga, P.; Walsh, S.; Wilkinson, A. J.; Stelling, J.; Aymerich, S.; Sauer, U. *Science* **2012**, *335*, 1099.

Bumpus, S. B.; Evans, B. S.; Thomas, P. M.; Ntai, I.; Kelleher, N. L. *Nat. Biotechnol.* **2009**, *27*, 951.

Carter, G. T. *Nat. Prod. Rep.* **2014**, *31*, 711.

Challal, S.; Bohni, N.; Buenafe, O. E.; Esguerra, C. V.; de Witte, P. A. M.; Wolfender, J.-L.; Crawford, A. D. *Chimia* **2014**, 1.

Challal, S.; Buenafe, O. E. M.; Queiroz, E. F.; Maljevic, S.; Marcourt, L.; Bock, M.; Kloeti, W.; Dayrit, F. M.; Harvey, A. L.; Lerche, H.; Esguerra, C. V.; de Witte, P. A. M.; Wolfender, J.-L.; Crawford, A. D. *ACS Chem. Neurosci.* **2014**, *5*, 993.

Challis, G. L. *ChemBioChem* **2005**, *6*, 601–611.

Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Brown, L. C. W.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 412–421.

Cortina, N. S.; Krug, D.; Plaza, A.; Revermann, O.; Müller, R. *Angew. Chem. Int. Ed.* **2011**, *51*, 811.

Costanzo, M.; Baryshnikova, A.; Bellay, J.; Kim, Y.; Spear, E. D.; Sevier, C. S.; Ding, H.; Koh, J. L. Y.; Toufighi, K.; Mostafavi, S.; Prinz, J.; St Onge, R. P.;

VanderSluis, B.; Makhnevych, T.; Vizeacoumar, F. J.; Alizadeh, S.; Bahr, S.; Brost, R. L.; Chen, Y.; Cokol, M.; Deshpande, R.; Li, Z.; Lin, Z.-Y.; Liang, W.; Marback, M.; Paw, J.; San Luis, B.-J.; Shuteriqi, E.; Tong, A. H. Y.; van Dyk, N.; Wallace, I. M.; Whitney, J. A.; Weirauch, M. T.; Zhong, G.; Zhu, H.; Houry, W. A.; Brudno, M.; Ragibizadeh, S.; Papp, B.; Pál, C.; Roth, F. P.; Giaever, G.; Nislow, C.; Troyanskaya, O. G.; Bussey, H.; Bader, G. D.; Gingras, A.-C.; Morris, Q. D.; Kim, P. M.; Kaiser, C. A.; Myers, C. L.; Andrews, B. J.; Boone, C. *Science* **2010**, *327*, 425.

Crawford, A. D.; Liekens, S.; Kamuhabwa, A. R.; Maes, J.; Munck, S.; Busson, R.; Rozenski, J.; Esguerra, C. V.; de Witte, P. A. M. *PLoS ONE* **2011**, *6*, e14694.

Crawford, A.; Esguerra, C.; de Witte, P. *Planta Med* **2008**, *74*, 624.

Cuthbertson, D. J.; Johnson, S. R.; Piljac-Žegarac, J.; Kappel, J.; Schäfer, S.; Wüst, M.; Ketchum, R. E. B.; Croteau, R. B.; Marques, J. V.; Davin, L. B.; Lewis, N. G.; Rolf, M.; Kutchan, T. M.; Soejarto, D. D.; Lange, B. M. *Phytochemistry* **2013**, *91*, 187.

de Loubresse, N. G.; Prokhorova, I.; Holtkamp, W.; Rodnina, M. V.; Yusupova, G.; Yusupov, M. *Nature* **2014**, *513*, 517–522.

Deyrup, S. T.; Eckman, L. E.; McCarthy, P. H.; Smedley, S. R.; Meinwald, J.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 9753.

Ding, W.-X.; Ni, H.-M.; Gao, W.; Hou, Y.-F.; Melan, M. A.; Chen, X.; Stolz, D. B.; Shao, Z.-M.; Yin, X.-M. *J. Biol. Chem.* **2007**, *282*, 4702–4710.

Donaldson, J. G.; Finazzi, D.; Klausner, R. D. *Nature* **1992**, *360*, 350–352.

Donia, M. S.; Cimermancic, P.; Schulze, C. J.; Wieland Brown, L. C.; Martin, J.; Mitreva, M.; Clardy, J.; Linington, R. G.; Fischbach, M. A. *Cell* **2014**, *158*, 1402–1414.

Donia, M. S.; Ravel, J.; Schmidt, E. W. *Nat. Chem. Biol.* **2008**, *4*, 341–343.

Duncan, K. R.; Crüsemann, M.; Lechner, A.; Sarkar, A.; Li, J.; Ziemert, N.; Wang, M.; Bandeira, N.; Moore, B. S.; Dorrestein, P. C.; Jensen, P. R. *Chem. Biol.* **2015**, *22*, 460.

Dutta, S.; Whicher, J. R.; Hansen, D. A.; Hale, W. A.; Chemler, J. A.; Congdon, G. R.; Narayan, A. R.; Håkansson, K.; Sherman, D. H.; Smith, J. L.; Skiniotis, G. *Nature* **2014**, *510*, 512–517.

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

Erickson, K. L.; Beutler, J. A.; Cardellina, J. H.; Boyd, M. R. *J. Org. Chem.* **1997**, *62*, 8188.

Feng, Y.; Mitchison, T. J.; Bender, A.; Young, D. W.; Tallarico, J. A. *Nat. Rev. Drug. Discov.* **2009**, *8*, 567.

Finn, R. D.; Clements, J.; Eddy, S. R. *Nucleic Acids Res.* **2011**, *39*, W29–W37.

Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A. *J. Proteome Res.* **2007**, *7*, 113.

Fuchs, S. W.; Bozhüyük, K. A. J.; Kresovic, D.; Grundmann, F.; Dill, V.; Brachmann, A. O.; Waterfield, N. R.; Bode, H. B. *Angew. Chem. Int. Ed.* **2013**, *52*, 4108–4112.

Fung, S.-Y.; Sofiyev, V.; Schneiderman, J.; Hirschfeld, A. F.; Victor, R. E.; Woods, K.; Piotrowski, J. S.; Deshpande, R.; Li, S. C.; de Voogd, N. J.; Myers, C. L.; Boone, C.; Andersen, R. J.; Turvey, S. E. *ACS Chem. Biol.* **2014**, *9*, 247.

Futamura, Y.; Kawatani, M.; Kazami, S.; Tanaka, K.; Muroi, M.; Shimizu, T.; Tomita, K.; Watanabe, N.; Osada, H. *Chem. Biol.* **2012**, *19*, 1620.

Futamura, Y.; Kawatani, M.; Muroi, M.; Aono, H.; Nogawa, T.; Osada, H. *ChemBioChem* **2013**, *14*, 2456.

Gerwick, W. H.; Moore, B. S. *Chem. Biol.* **2012**, *19*, 85–98.

Giaever, G.; Chu, A. M.; Ni, L.; Connelly, C.; Riles, L.; Véronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; André, B.; Arkin, A. P.; Astromoff, A.; Bakkoury, El, M.; Bangham, R.; Benito, R.; Brachat, S.; Campanaro, S.; Curtiss, M.; Davis, K.; Deutschbauer, A.; Entian, K.-D.; Flaherty, P.; Foury, F.; Garfinkel, D. J.; Gerstein, M.; Gotte, D.; Güldener, U.; Hegemann, J. H.; Hempel, S.; Herman, Z.; Jaramillo, D. F.; Kelly, D. E.; Kelly, S. L.; Kötter, P.; LaBonte, D.; Lamb, D. C.; Lan, N.; Liang, H.; Liao, H.; Liu, L.; Luo, C.; Lussier, M.; Mao, R.; Menard, P.; Ooi, S. L.; Revuelta, J. L.; Roberts, C. J.; Rose, M.; Ross-Macdonald, P.; Scherens, B.; Schimmack, G.; Shafer, B.; Shoemaker, D. D.; Sookhai-Mahadeo, S.; Storms, R. K.; Strathern, J. N.; Valle, G.; Voet, M.; Volckaert, G.; Wang, C.-Y.; Ward, T. R.; Wilhelmy, J.; Winzeler, E. A.; Yang, Y.; Yen, G.; Youngman, E.; Yu, K.; Bussey, H.; Boeke, J. D.; Snyder, M.; Philippsen, P.; Davis, R. W.; Johnston, M. *Nature* **2002**, *418*, 387.

Goel, A. K.; Rajagopal, L.; Nagesh, N.; Sonti, R. V. *Journal of Bacteriology* **2002**, *184*, 3539–3548.

Grkovic, T.; Pouwer, R. H.; Vial, M. L.; Gambini, L.; Noël, A.; Hooper, J. N. A.; Wood, S. A.; Mellick, G. D.; Quinn, R. J. *Angew. Chem. Int. Ed.* **2014**, *53*, 6070.

Gronquist, M.; Meinwald, J.; Eisner, T.; Schroeder, F. C. *J. Am. Chem. Soc.* **2005**, *127*, 10810.

Gubbens, J.; Zhu, H.; Girard, G.; Song, L.; Florea, B. I.; Aston, P.; Ichinose, K.; Filippov, D. V.; Choi, Y. H.; Overkleeft, H. S.; Challis, G. L.; van Wezel, G. P. *Chem. Biol.* **2014**, 1.

Høyer-Hansen, M.; Jäättelä, M. *Cell Death Differ* **2007**, *14*, 1576–1582.

Hamada, T.; Matsunaga, S.; Fujiwara, M.; Fujita, K.; Hirota, H.; Schmucki, R.; Güntert, P.; Fusetani, N. *J. Am. Chem. Soc.* **2010**, *132*, 12941.

Hau, A. M.; Greenwood, J. A.; Löhr, C. V.; Serrill, J. D.; Proteau, P. J.; Ganley, I. G.; McPhail, K. L.; Ishmael, J. E. *PLoS ONE* **2013**, *8*, e65250.

Heifetz, A.; Keenan, R. W.; Elbein, A. D. *Biochemistry* **1979**, *18*, 2186–2192.

Hoffmann, T.; Krug, D.; Hüttel, S.; Müller, R. *Anal. Chem.* **2014**, *86*, 10780.

Hook, D. J.; More, C. F.; Yacobucci, J. J.; Dubay, G.; O'Connor, S. *J. Chromatogr. A* **1987**, *385*, 99.

Hook, D. J.; Pack, E. J.; Yacobucci, J. J.; Guss, J. *J. of Biomol. Screening* **1997**, *2*, 145.

Hou, Y.; Braun, D. R.; Michel, C. R.; Klassen, J. L.; Adnani, N.; Wyche, T. P.; Bugni, T. S. *Anal. Chem. 84*, 4277.

Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. A. *PNAS* **2012**, *109*, 19196.

Ito, T.; Odake, T.; Katoh, H.; Yamaguchi, Y.; Aoki, M. *J. Nat. Prod.* **2011**, *74*, 983.

Jarmusch, A. K.; Cooks, R. G. *Nat. Prod. Rep.* **2014**, *31*, 730.

Jiang, Z.-D.; Jensen, P. R.; Fenical, W. *Bioorganic & Medicinal Chemistry Letters*

1999, *9*, 2003–2006.

Jurkowitz, L.; Loeb, J. N.; Brown, P.; Wald, G. *Nature* **1959**, *184*, 614–624.

Karathia, H.; Vilaprinyo, E.; Sorribas, A.; Alves, R. *PLoS ONE* **2011**, *6*, e16015.

Kersten, R. D.; Yang, Y.-L.; Xu, Y.; Cimermancic, P.; Nam, S.-J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat. Chem. Biol.* **2011**, *7*, 794.

Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D. *Fungal Genetics and Biology* **2010**, *47*, 736–741.

Krug, D.; Zurek, G.; Revermann, O.; Vos, M.; Velicer, G. J.; Müller, R. *Applied and Environmental Microbiology* **2008**, *74*, 3058–3068.

Lai, K.; Selinger, D. W.; Solomon, J. M.; Wu, H.; Schmitt, E.; Serluca, F. C.; Curtis, D.; Benson, J. D. *ACS Chem. Biol.* **2013**, *8*, 257.

Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. *Science* **2006**, *313*, 1929.

Li, L.; Stoeckert, C. J.; Roos, D. S. *Genome Research* **2003**, *13*, 2178–2189.

Li, M. H.; Ung, P. M.; Zajkowski, J.; Garneau-Tsodikova, S.; Sherman, D. H. *BMC Bioinformatics* **2009**, *10*, 185.

Lin, C. C.; Chung, M.; Gural, R.; Schuessler, D.; Kim, H. K.; Radwanski, E.; Marco, A.; DiGiore, C.; Symchowicz, S. *Antimicrobial Agents and Chemotherapy* **1984**, *26*, 522.

Lin, K.; Zhu, L.; Zhang, D.-Y. *Bioinformatics* **2006**, *22*, 2081–2086.

Lorang, J.; King, R. W. *Genome Biol.* **2005**, *6*, 228.

Luesch, H.; Wu, T. Y. H.; Ren, P.; Gray, N. S.; Schultz, P. G.; Supek, F. *Chem. Biol.* **2005**, *12*, 55.

Mascuch, S. J.; Moree, W. J.; Hsu, C.-C.; Turner, G. G.; Cheng, T. L.; Blehert, D. S.; Kilpatrick, A. M.; Frick, W. F.; Meehan, M. J.; Dorrestein, P. C.; Gerwick, L. *PLoS ONE* **2015**, *10*, e0119668.

McBride, M. J.; Xie, G.; Martens, E. C.; Lapidus, A.; Henrissat, B.; Rhodes, R. G.; Goltsman, E.; Wang, W.; Xu, J.; Hunnicutt, D. W.; Staroscik, A. M.; Hoover, T. R.; Cheng, Y.-Q.; Stein, J. L. *Applied and Environmental Microbiology* **2009**, *75*, 6864–6875.

McBrien, K. D.; Berry, R. L.; Lowe, S. E.; Neddermann, K. M.; Bursuker, I.; Huang, S.; Klohr, S. E.; Leet, J. E. *J. Antibiot.* **1995**, *48*, 1446–1452.

Molinski, T. F. *Org. Lett.* **2014**, *16*, 3849–3855.

Molinski, T. F. *Nat. Prod. Rep.* **2010**, *27*, 321.

Muroi, M.; Kazami, S.; Noda, K.; Kondo, H.; Takayama, H.; Kawatani, M.; Usui, T.; Osada, H. *Chem. Biol.* **2010**, *17*, 460.

Navarro, G.; Cheng, A. T.; Peach, K. C.; Bray, W. M.; Bernan, V. S.; Yildiz, F. H.; Linington, R. G. *Antimicrobial Agents and Chemotherapy* **2014**, *58*, 1092–1099.

Nett, M.; Ikeda, H.; Moore, B. S. *Nat. Prod. Rep.* **2009**, *26*, 1362.

Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2012**, *75*, 311–335.

Nielsen, K. F.; Månsson, M.; Rank, C.; Frisvad, J. C.; Larsen, T. O. *J. Nat. Prod.* **2011**, *74*, 2338–2348.

Nonejuie, P.; Burkart, M.; Pogliano, K.; Pogliano, J. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 16169.

Olivera, B. M.; Rivier, J.; Clark, C.; Ramilo, C. A.; Corpuz, G. P.; Abogadie, F. C.; Mena, E. E.; SR, W.; Hillyard, D. R.; Cruz, L. J. *Science* **1990**, *249*, 257.

Ooi, S. L.; Pan, X.; Peyser, B. D.; Ye, P.; Meluh, P. B.; Yuan, D. S.; Irizarry, R. A.; Bader, J. S.; Spencer, F. A.; Boeke, J. D. *Trends Genet.* **2006**, *22*, 56.

Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Jiang, H.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. *Nature* **2013**, *496*, 528–532.

Parsons, A. B.; Lopez, A.; Givoni, I. E.; Williams, D. E.; Gray, C. A.; Porter, J.; Chua, G.; Sopko, R.; Brost, R. L.; Ho, C. H.; Wang, J.; Ketela, T.; Brenner, C.; Brill, J. A.; Fernandez, G. E.; Lorenz, T. C.; Payne, G. S.; Ishihara, S.; Ohya, Y.; Andrews, B.; Hughes, T. R.; Frey, B. J.; Graham, T. R.; Andersen, R. J.; Boone, C. *Cell* **2006**, *126*, 611.

Peach, K. C.; Bray, W. M.; Winslow, D.; Linington, P. F.; Linington, R. G. *Mol. BioSyst.* **2013**, *9*, 1837.

Perlman, Z. E.; Slack, M. D.; Feng, Y.; Mitchison, T. J.; Wu, L. F.; Altschuler, S. J. *Science* **2004**, *306*, 1194–1198.

Pluskal, T.; Uehara, T.; Yanagida, M. *Anal. Chem.* **2012**, *84*, 4396–4403.

Potts, M. B.; Kim, H. S.; Fisher, K. W.; Hu, Y.; Carrasco, Y. P.; Bulut, G. B.; Ou, Y.-H.; Herrera-Herrera, M. L.; Cubillos, F.; Mendiratta, S.; Xiao, G.; Hofree, M.; Ideker, T.; Xie, Y.; Huang, L. J.-S.; Lewis, R. E.; MacMillan, J. B.; White, M. A. *Sci. Signal.* **2013**, *6*, ra90.

Pungaliya, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708.

Purcell, E. M. *Science* **1953**, *118*, 431.

Quan, J.; Tian, J. *PLoS ONE* **2009**, *4*, e6441.

Quan, J.; Tian, J. *Nat Protoc* **2011**.

Raldúa, D.; Piña, B. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 685.

Rath, C. M.; Janto, B.; Earl, J.; Ahmed, A.; Hu, F. Z.; Hiller, L.; Dahlgren, M.; Kreft, R.; Yu, F.; Wolff, J. J.; Kweon, H. K.; Christiansen, M. A.; Håkansson, K.; Williams, R. M.; Ehrlich, G. D.; Sherman, D. H. *ACS Chem. Biol.* **2011**, *6*, 1244–1256.

Rehm, B. H. A. *Nature Publishing Group* **2010**, *8*, 578–592.

Rfimann, H.; Jaret, R. S. *J. Chem. Soc., Chem. Commun.* **1972**, 1270a.

Rihel, J.; Prober, D. A.; Arvanites, A.; Lam, K.; Zimmerman, S.; Jang, S.; Haggarty, S. J.; Kokel, D.; Rubin, L. L.; Peterson, R. T.; Schier, A. F. *Science* **2010**, *327*, 348.

Robinette, S. L.; Brüschweiler, R.; Schroeder, F. C.; Edison, A. S. *Acc. Chem. Res.* **2011**, *45*, 288.

Samali, A.; FitzGerald, U.; Deegan, S.; Gupta, S. *International Journal of Cell Biology* **2010**, *2010*, 1–11.

Samuel, G.; Reeves, P. *Carbohydr. Res.* **2003**, *338*, 2503–2519.

Sang, F.; Li, D.; Sun, X.; Cao, X.; Wang, L.; Sun, J.; Sun, B.; Wu, L.; Yang, G.; Chu, X.; Wang, J.; Dong, C.; Geng, Y.; Jiang, H.; Long, H.; Chen, S.; Wang, G.; Zhang, S.; Zhang, Q.; Chen, Y. *J. Am. Chem. Soc.* **2014**, *136*, 15787–15791.

Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. *Chem. Biol.* **2013**, *20*, 285.

Schulze, C. J.; Linington, R. G. In *Natural Products: Discourse, Diversity, and Design*; Osbourn, A., Goss, R. J., Carter, G. T., Eds.; wiley.com: Oxford, 2014; pp 373–396.

Sidebottom, A. M.; Johnson, A. R.; Karty, J. A.; Trader, D. J.; Carlson, E. E. *ACS Chem. Biol.* **2013**, *8*, 2009.

Starcevic, A.; Zucko, J.; Simunkovic, J.; Long, P. F.; Cullum, J.; Hranueli, D. *Nucleic Acids Res.* **2008**, *36*, 6882–6892.

Suffness, M.; Douros, J. D. *Trends Pharmacol. Sci.* **1981**, *2*, 307–310.

Taggi, A. E.; Meinwald, J.; Schroeder, F. C. *J. Am. Chem. Soc.* **2004**, *126*, 10364.

Takeuchi, M.; Ashihara, E.; Yamazaki, Y.; Kimura, S.; Nakagawa, Y.; Tanaka, R.; Yao, H.; Nagao, R.; Hayashi, Y.; Hirai, H.; Maekawa, T. *Cancer Science* **2010**, *102*, 591–596.

Tanaka, M.; Bateman, R.; Rauh, D.; Vaisberg, E.; Ramachandani, S.; Zhang, C.; Hansen, K. C.; Burlingame, A. L.; Trautman, J. K.; Shokat, K. M.; Adams, C. L. *PLoS Biol.* **2005**, *3*, e128.

Tong, A. H.; Evangelista, M.; Parsons, A. B.; Xu, H.; Bader, G. D.; Pagé, N.; Robinson, M.; Raghibizadeh, S.; Hogue, C. W.; Bussey, H.; Andrews, B.; Tyers, M.; Boone, C. *Science* **2001**, *294*, 2364.

Treiman, M.; Caspersen, C.; Christensen, S. B. *Trends in Pharmacological Sciences* **1998**, *19*, 131–135.

Tsukida, K.; Saiki, K.; Takii, T.; Koyama, Y. *J. Chromatogr. A* **1982**, *245*, 359–364.

Verfaillie, T.; Salazar, M.; Velasco, G.; Agostinis, P. *International Journal of Cell Biology* **2010**, *2010*, 1–19.

Walsh, C. T.; Fischbach, M. A. *J. Am. Chem. Soc.* **2010**, *132*, 2469–2493.

Walter, M. H.; Strack, D. *Nat. Prod. Rep.* **2011**, *28*, 663–692.

Wang, Y.; Qian, G.; Li, Y.; Wang, Y.; Wang, Y.; Wright, S.; Li, Y.; Shen, Y.; Liu, F.; Du, L. *PLoS ONE* **2013**, *8*, e66633.

Wassermann, A. M.; Lounkine, E.; Urban, L.; Whitebread, S.; Chen, S.; Hughes, K.; Guo, H.; Kutlina, E.; Fekete, A.; Klumpp, M.; Glick, M. *ACS Chem. Biol.* **2014**, *9*, 1622–1631.

Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E1743.

Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W.; BREITLING, R.; Takano, E.; Medema, M. H. *Nucleic Acids Res.* **2015**, gkv437.

Weitnauer, G.; Mühlenweg, A.; Trefzer, A.; Hoffmeister, D.; Süssmuth, R. D.; Jung, G.; Welzel, K.; Vente, A.; Girreser, U.; Bechthold, A. *Chem. Biol.* **2001**, *8*, 569–581.

Winzeler, E. A.; Shoemaker, D. D.; Astromoff, A.; Liang, H.; Anderson, K.; Andre, B.; Bangham, R.; Benito, R.; Boeke, J. D.; Bussey, H.; Chu, A. M.; Connelly, C.; Davis, K.; Dietrich, F.; Dow, S. W.; Bakkoury, El, M.; Foury, F.; Friend, S. H.; Gentalen, E.; Giaever, G.; Hegemann, J. H.; Jones, T.; Laub, M.; Liao, H.; Liebundguth, N.; Lockhart, D. J.; Lucau-Danila, A.; Lussier, M.; M'Rabet, N.; Menard, P.; Mittmann, M.; Pai, C.; Rebischung, C.; Revuelta, J. L.; Riles, L.; Roberts, C. J.; Ross-MacDonald, P.; Scherens, B.; Snyder, M.; Sookhai-Mahadeo, S.; Storms, R. K.; Véronneau, S.; Voet, M.; Volckaert, G.; Ward, T. R.; Wysocki, R.; Yen, G. S.; Yu, K.; Zimmermann, K.; Philippsen, P.; Johnston, M.; Davis, R. W. *Science* **1999**, *285*, 901.

Woehrmann, M. H.; Bray, W. M.; Durbin, J. K.; Nisam, S. C.; Michael, A. K.; Glassey, E.; Stuart, J. M.; Lokey, R. S. *Mol. BioSyst.* **2013**, *9*, 2604–2617.

Wong, W. R.; Oliver, A. G.; Linington, R. G. *Chem. Biol.* **2012**, *19*, 1483.

Woodward, R. B.; Brehm, W. J.; Nelson, A. L. *J. Am. Chem. Soc.* **1947**, *69*, 2250.
Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686.

Young, D. W.; Bender, A.; Hoyt, J.; McWhinnie, E.; Chirn, G.-W.; Tao, C. Y.; Tallarico, J. A.; Labow, M.; Jenkins, J. L.; Mitchison, T. J.; Feng, Y. *Nat. Chem. Biol.* **2008**, *4*, 59.

Zechmeister, L.; Lemmon, R. M. *J. Am. Chem. Soc.* **2001**, *66*, 317–322.

Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Brüschweiler, R. *Anal. Chem.* **2007**, *79*, 7748.

Zhang, W.; Liu, Z.; Li, S.; Lu, Y.; Chen, Y.; Zhang, H.; Zhang, G.; Zhu, Y.; Zhang, G.; Zhang, W.; Liu, J.; Zhang, C. *J. Nat. Prod.* **2012**, *75*, 1937.

Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. *PLoS ONE* **2012**, *7*, e34064.

Zon, L. I.; Peterson, R. T. *Nat. Rev. Drug. Discov.* **2005**, *4*, 35.