UNIVERSITY OF CALIFORNIA, SAN DIEGO

Enhancing Natural Products Structural Dereplication and Elucidation with Deep Learning Based Nuclear Magnetic Resonance Techniques

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

in

NanoEngineering

by

Chen Zhang

Committee in charge:

William H. Gerwick, Chair
Garrison W. Cottrell, Co-Chair
Gaurav Arya
Seth M. Cohen
Chambers C. Hughes
Preston B. Landon
Liangfang Zhang

2017

The Dissertation of Chen Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my beloved parents, Jianping Zhao and Xiaojing Zhang whose words of encouragement and push for tenacity ring in my ears. My cousin Min Zhang has never left my side and is very special.

I also dedicate this dissertation to my mentors who have shown me fascinating views of the world throughout the process, and those who have walked me through the valley of the shadow of frustration. I will always appreciate all they have done, especially Bill Gerwick, Gary Cottrell and Preston Landon for showing me the gate to new frontiers, Sylvia Evans, Pieter Dorrestein, Shu Chien, Liangfang Zhang, and Gaurav Arya for their great encouragement, and Wood Lee and Yezifeng for initially showing me the value of freedom, and continuously answering my questions regarding social sciences, humanities, literature, and arts.

I dedicate this work and give special thanks to my wonderful friends, especially Niu Du, Sima Yazdani, Yerlan Idelbayev, Jie Min, Zheng Long, Mike Shao, Bettina Lehman, Sheri Harris and Han Huang for not giving me up throughout the entire doctorate program. All of you have been my best cheerleaders.

EPIGRAPH

"So doth the woodbine, the sweet honeysuckle
Gently entwist; the female ivy so
Enrings the barky fingers of the elm.
O, how I love thee! how I dote on thee!"

--*"A Midsummer Night's Dream" ACT IV, sc. i., By W. Shakespeare*

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI – Artificial Intelligence

CD – Circular Dichroism

$CHCl_3$ – Chloroform

CH_2Cl_2 – Dichloromethane

CH_2N_2 – Diazomethane

CH_3CN – Acetonitrile

CNN – Convolutional Neural Networks

COSY – Correlation Spectroscopy

DMSO – Dimethyl sulfoxide

ECCD – Exciton-Coupled Circular Dichroism

ESI – Electrospray Ionization

EtOAc – Ethyl Acetate

EtOH – Ethanol

FDVA – 1-fluoro-2,4-dinitrophenyl-5-valine amide

FT – Fourier Transform

GCMS – Gas Chromatography Mass Spectrometry

GNPS – Global Natural Products Social Molecular Networking

H2BC – Heteronuclear 2 Bond Correlation

HMBC – Heteronuclear Multiple Bond Correlation

HPLC – High Pressure Liquid Chromatography

HR – High Resolution

HSQC – Heteronuclear Single Quantum Correlation

IR – Infrared Spectroscopy

LCMS – Liquid Chromatography Mass Spectrometry

LR – Low Resolution

MeOH – Methanol

MNP – Marine Natural Products

MS – Mass spectrometry

NCE – New Chemical Entity

NMR – Nuclear Magnetic Resonances

NOE – Nuclear Overhauser Effect

NP – Natural Product

NRPS – Non-Ribosomal Peptide Synthetase

NTD – Neglected Tropical Diseases

PKS – Polyketide Synthase

ROESY – Rotating-Frame Nuclear Overhauser Effect Correlation Spectroscopy

RP – Reverse Phase

SCUBA – Self Contained Underwater Breathing Apparatus

TOCSY – Total Correlation Spectroscopy

TOF – Time of Flight

UV – Ultraviolet

VLC – Vacuum Liquid Chromatography

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my entire committee for their time, feedback and support of my research but especially my advisor Professor William H. Gerwick and co-advisor Professor Garrison W. Cottrell. Professor William H. Gerwick and Professor Garrison W. Cottrell have always been enthusiastic and strongest supporters of my research and my future as an inventor as they played significant roles in many of my realized dreams, including showing the gate to a "digital frontier", being awarded the Frontiers of Innovation Scholarship, acquiring additional funds to advance the research on SMART and most importantly finding and allowing a brilliant graduate student and later my friend, Yerlan Idelbayev, to work with. I greatly appreciate their mentorship, support, and guidance over the past years.

I would also like to acknowledge Dr. Preston B. Landon at the University of California, San Diego, Department of Bioengineering for igniting my passion of creative engineering. He offered me in his laboratory at night times working with a brilliant graduate student, Alexander H. Mo, on designing and characterizing nanocarriers, and I appreciate Preston's continued mentorship, support, and guidance in my future.

I would also like to acknowledge my entire family, especially my parents Jianping Zhao and Xiaojing Zhang and my cousin Min Zhang for their love, support, and encouragement throughout my educational years, as I would never had made it this far without any of them. I especially appreciate all of those video calls with my parents

as well as their financial support, and food and shelters provided by my cousin and her family during my summer and winter schools in Shanghai before 2011.

I would also like to acknowledge all of my good friends, especially Han Huang, Mike Shao, Niu Du, Jie Min, Lawrence Januar, Xiao Wang, Sima Yazdani, Zheng Long, Bettina Lehman and Sheri Harris and her family. I would like to thank them for constantly visiting me or caring me. They have all been reasons to have hope in my life.

I would like to also acknowledge my collaborators, Changlun Shao, Yiwen Tao, Emily Mevers, Evgenia Glukhov and her son David, Enora Briand, Benjamin Naman, Matthew Bertin, Eugene Lin, Lena Gerwick and her son Erik, Bailey Miller, Cameron Coates, Karin Kleigrewe, John Lee, Nathan Moss, Tiago Leao, Valentina Steverlynck, Virginia Xu, Susan Golden, Niclas Engene, Thomas Williamson, Brian Marquez, Pieter Dorrestein, Sanjeev Rao, Yufei Wang, Yashwanth Nannapaneni, Nicolas Roberts, Virginia Xin Xu, Ratnesh Lal, Alice Yepremeyan and her mom Margo, Mike Hwang, Joon Lee, Woraphong Janetanakit, Bettina Lehman and all of the past and present Gerwick lab, Cottrell lab and Lal lab members. It has been a true pleasure to be able to work alongside each of them and I appreciate all of their advice and intriguing discussions.

I would like to also acknowledge the UCSD Nanoengineering administration, especially Dana Jimenez, for her help finding me financial supports. The SIO administration, for providing me the piano to play during lunch times. The UCSD NMR and Mass Spectrometry facilities, especially Anthony Mrse and Brendan Duggan, for their immense knowledge on acquiring 2D NMR data, and Yongxuan Su for his effort

on acquiring high resolution mass spectra. The Seth Cohen lab for their generosity of providing FT IR instrument usage.

Chapter 2, in part, includes a reprint as it appears in the Marine Drugs. 2017, 15(4), 121, with the following authors Chen Zhang, C. Benjamin Naman, Niclas Engene, and William H. Gerwick. The dissertation author was a primary investigator and first author of this paper.

Chapter 3, is a reprint as it appears in the Scientific Reports. 2017, 7(1), 14243, with the following authors, Chen Zhang, Yerlan Idelbayev, Nicholas Roberts, Yiwen Tao, Yashwanth Nannapaneni, Brendan M. Duggan, Jie Min, Eugene C. Lin, Erik C. Gerwick, Garrison W. Cottrell, and William H. Gerwick. The dissertation author was the primary investigator and is the co-first author of this paper.

Chapter 4, in essence, is currently being prepared for submission in 2017, with the following authors Yiwen Tao, Chen Zhang, Yerlan Idelbayev, Svetlana Nikoulina, Evgenia Glukhov, C. Benjamin Naman, Garrison W. Cottrell and William H. Gerwick. The dissertation author was the primary investigator and will be the co-first author of this material.

VITA

EDUCATION AND FIELD OF STUDY

University of California, San Diego, La Jolla, CA, USA 2013-2017
Doctor of Philosophy in Nanoengineering
Advisors: William H. Gerwick and Garrison W. Cottrell

University of California, San Diego, La Jolla, CA, USA 2011-2013
Master of Science in Nanoengineering

Ocean University of China, Qingdao, Shandong, China 2007-2011
Bachelor of Science in Marine Chemistry

AWARDS

Frontier of Innovation Scholarship, University of California, San Diego 2014-2015

PUBLICATIONS

Zhang, C., Idelbayev, Y., Roberts, N., Tao, Y., Nannapaneni, Y., Duggan, B. M., Min, J., Lin, E. C., Gerwick, E. C., Cottrell, G. W., Gerwick, W. H. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research *Scientific Reports* **2017**, 7(1), 14243.

Zhang, C., Naman, C. B., Engene, N., Gerwick, W. H. Laucysteinamide A, a Hybrid PKS/NRPS Metabolite from a Saipan Cyanobacterium, cf. Caldora penicillata *Marine Drugs* **2017**, 15(4), 121.

Mo, A. H., Zhang, C., Landon, P. B., Janetanakit, W., Hwang, M. T., Santacruz-Gomez, K., Colburn, D. A., Dossou, S. M., Lu, T., Cao, Y., Sant, V., Sud, P. L., Akkiraju, S., Shubayev, V. I., Glinsky, G., Lal, R. Dual-Functionalized Theranostic Nanocarriers *ACS Applied Materials & Interfaces* **2016**, *8*, 14740-14746.

Mo, A. H., Landon, P. B., Gomez, K. S., Kang, H., Lee, J., Zhang, C., Janetanakit, W., Sant, V., Lu, T., Colburn, D. A., Akkiraju, S., Dossou, S., Cao, Y., Lee, K. F., Varghese, S., Glinsky, G., Lal, R. Magnetically-Responsive Silica-Gold Nanobowls for Targeted Delivery and SERS-Based Sensing *Nanoscale* **2016**, *8*, 11840-11850.

Santacruz-Gomez, K., Silva-Campa, E., Melendrez-Amavizca, R., Teran-Arce, F., Mata-Haro, V., Landon, P.B., Zhang, C., Pedroza-Montero, M., Lal, R. Carboxylated

Nanodiamonds Inhibit γ-Irradiation Damage of Human red Blood Cells *Nanoscale* **2016**, *8*, 7189-7196.

Landon, P. B., Mo, A. H., Printz, A. D., Emerson, C., Zhang, C., Janetanakit, W., Colburn, D. A., Akkiraju, S., Dossou, S., Chong, B., Glinsky, G., Lal, R. Asymmetric Colloidal Janus Particle Formation Is Core-Size-Dependent *Langmuir* **2015**, *31*, 9148-9154.

Shao, C. L., Linington, R. G., Balunas, M. J., Centeno, A., Boudreau, P., Zhang, C.; Engene, N., Spadafora, C., Mutka, T. S., Kyle, D. E., Gerwick, L., Wang, C. Y., Gerwick, W. H. Bastimolide A, a Potent Antimalarial Polyhydroxy Macrolide from the Marine Cyanobacterium *Okeania hirsute Journal of Organic Chemistry* **2015**, *80*, 7849-7855.

Mo, A. H., Landon, P. B., Emerson, C. D., Zhang, C., Anzenberg, P., Akkiraju, S., Lal R. Synthesis of Nanobowls with a Janus Template *Nanoscale* **2015**, *7*, 771-775.

Landon, P. B., Lee, J., Hwang, M. T., Mo, A. H., Zhang, C., Neuberger, A., Meckes, B., Gutierrez, J. J., Glinsky, G., Lal, R. Energetically Biased DNA Motor Containing a Thermodynamically Stable Partial Strand Displacement state *Langmuir* **2014**, *30*, 14073-14078.

Landon, P. B., Mo, A. H., Zhang, C., Emerson, C. D., Printz, A. D., Gomez, A. F., DeLaTorre, C. J., Colburn, D. A., Anzenberg, P., Eliceiri, M., O'Connell, C., Lal, R. Designing Hollow Nano Gold Golf Balls *ACS Applied Materials & Interfaces* **2014**, *6*, 9937-9941.

Wang, M., et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34*, 828-837.

PUBLICATION IN PREPARATION

Tao, Y., Zhang, C., Idelbayev, Y., Nikoulina, S., Glukhov, E., Naman, C. B., Cottrell, G. W., Gerwick, W. H. Viequeamides and Aurilide D: Exceptionally Anti-cancer Cytotoxic Cyclic Depsipeptides from Marine Cyanobacteria Rivularia sp. and *Moorea producens*, **2017**, in prep.

ABSTRACT OF THE DISSERTATION

Enhancing Natural Products Structural Dereplication and Elucidation with Deep Learning Based Nuclear Magnetic Resonance Techniques

by

Chen Zhang

Doctor of Philosophy in Nanoengineering

University of California, San Diego, 2017

Professor William Gerwick, Chair
Professor Garrison Cottrell, Co-Chair

Nature Products Research (NPR) has a long history of revealing bioactive constituents of natural origin, both as single drug leads within modern western medicine and as mixtures of bioactive constituents enriching traditional medicines. Identifying

bioactive constituents in complex mixtures such as those obtained from extracting marine algae has been relying on multidisciplinary techniques, such as bioactivity-guided or spectroscopic-guided fractionation and purification. In this regard, milestones of scientific achievements of NPR have been hailed by applying novel technologies, such as improved separation or purification, spectroscopic hardware with detection limits of natural abundance, software algorithms for accelerating data collecting and processing, and high-throughput screening.

In most NPR, the characterization of novel compounds as well as the dereplication of known compounds entails the collection and analysis of NMR spectra. This involves the running of 1D and 2D NMR spectroscopic experiments for the purpose of partial structure construction, assemblage and relative stereochemistry determination. As exciting advancements in the rapid genetic and proteomic approaches have made their way into NPR, conventional NMR practices have become one of several bottlenecks in the characterization and dereplication of new compounds. In regard to this challenge, we leveraged the advantages of Non Uniform Sampling Nuclear Magnetic Resonance (NUS NMR) and Artificial Intelligence (AI) to create Small Molecule Accurate Recognition Technology (SMART) as a tool to speed up marine natural products discovery. Fast NMR techniques like NUS NMR have the potential to further reduce detection limits while maintaining the same sampling time and quality. Next, we applied over 4000 experimental Heteronuclear Single Quantum Correlation (HSQC) spectra for the AI training. The outcome is that the AI algorithm provided us with structurally insightful AI embedding maps with nodes and clusters representing correlations of related families of natural products. By testing different

HSQC spectra using this algorithm, we can greatly accelerate the rate of known compound identification as well as rapidly generating hypotheses about the relationship of new molecules to those used for the training - based entirely on their NMR properties. Specifically, the 2D NMR spectra of a series of unknown compounds isolated from two different marine cyanobacteria were recognized by the SMART belonging to a specific class of marine depsipeptides.

CHAPTER 1

INTRODUCTION

One of many goals of the drug discovery process is to provide an understanding of the pharmacological mechanisms by which small molecules can impact biological systems, and thereby in some cases provide new treatments for disease in humans. A key aspect of this research therefore depends on the efficient and accurate determination of the chemical structure of natural products. This thesis is therefore focused on enhancing the available techniques by which natural product structures can be determined.

1.1 Uses of Natural Products in the History

Natural Products (NPs) are characteristic secondary metabolites, a particular type of chemicals produced by living organisms (i.e. plants or microbes). Living organisms (especially plants and microbes) and human beings coexist in an intricate relationship – people benefit from living organisms as a source of commodity products, or people appreciate some organisms as cultural symbols or as charms.

Before chemical compounds could be structurally defined, the demand for plants with attractive biological properties, some of which were limited in supply, led to the practice of herbalism. The methodologies of herbalism are loosely based on scientific methods. Here, those attractive NPs include 1) the additives that give desired or pleasant odor, color or flavor to food and drinks, 2) scents or perfumes that stimulate human

olfaction, 3) pills, powders or potions that cure human diseases, promote human desires or serve recreational purposes, 4) raw materials from commodity cash crops. Common examples of species or natural products in these categories are tea, coffee, tobacco, alcohol, narcotics, fragrances, spices, rubber, quinine, etc.. The roles of those plant derivatives were used as poisons for hunting or murdering, hallucinogens, stimulants or narcotics, and medicine to save lives. From an ethno-pharmacognosy perspective, living organisms continuously become a prolific source of clues and medical materials, which supply health systems of human ethnic cultures to promote health, prevent sickness, restore health, and rehabilitate[1].

Reviewing the history of natural remedies in this dissertation is no more than a student excursion into some aspects of the historic association between humans and their healing plants. In prehistoric societies, tribal medicine man, shamans, or healers emerged to promote remedies that are mostly of vegetable origin. There is increasing archaeological evidence that indicates that medicinal plants were regularly applied by human beings in prehistoric times. In multiple prehistoric cultures, natural products were orally administrated for disease treatment or psychotherapies.[2, 3] It should be stressed that mixtures were formed based on common sense along with with intuition or magical exorcism, thus characterizing such botanical products as 'leech-craft'. Historical records from between the 4th millennium BC and 1st millennium BC showed that these previously inextricably combined magical, religious and empirico-rational elements began to separate. At least three types of healers were recorded in Kahoun Papyrus and Ramesseum Papyrus by the ancient Egyptians, namely, the physician, the priest of the Sekhmet and the sorcerer.[4] According to the papyrus, opium, cannabis,

myrrh, frankincense, fennel, cassia, senna, thyme, henna, juniper, linseed, castor oil, coriander, cumin, willow, sycamore, acacia, pomegranate, garlic, onions, aloes, figs, and gum produced by *Acacia arabica* are among the herbal medicines used by the ancient Egyptians.[5] Strikingly, Egyptian prescriptions contain indications about weights and measures of those ingredients. This phenomenon was not found in clay cuneiform records of the contemporary but later Mesopotamian civilizations (3th millennium BC to 539 BC), such as the Assyrians and Akkadians.

The materia medica of Mesopotamia inscribed in 660 clay tablets by cuneiform as examined and revealed by Thompson et al.[6] includes approximately 250 medicinal plants, as well as natural commodities of medical use, such as alcohols, oils, honey, and wax. Similar to their Egyptian counterparts, Mesopotamian medical texts are organized by a series of treatments followed by related symptoms, or even of a single symptom as the headings of a tablet. Mesopotamian herbal medicine consisted of fruits, common vegetables, cereals, spices and condiments, flowers, resins, gums, and other native flora. The genera identified by Thompson et al. are *Calendula*, *Styrax*, *Sinapis*, *Hyoscymus*, *Myrrha*, *Ambar*, *Asafoetida*, *Mentha*, *Lupinus*, *Solanum*, *Ricinus*, *Ferula*, *Foeniculum*, *Anthemis*, *Papaver*, *Glycyrrhiza*, *Rhus*, *Mandragora*, *Nerium*, *Opopanax*, *Rhamnus*, *Artemesia*, *Cannabis*, *Conium*, *Funiperus*, *Crocus*, *Thymus*, *Urtica*, *Oenanthe*, *Ruta*, *Citrus*, *Allium*, *Anacyclus*, *Cumin*, *Anemone*, *Origanum*, etc. Although Mesopotamian medicine is collectively characterized by religious and magic, for example, incantations are applied during the treatment, there are still rational and effective herbal usages as deciphered by Thompson et al., such as hellebore, hyoscyamus, mandrake, opium, hemp, etc.

India is mainly in the tropical region and has a great reservoir of biodiversity, as well as most diseases peculiar to the tropics. Under such conditions, remarkable developments of herbal medicine practices are recorded in the form of writings left by Aryan migrants or conquers (Vedic Aryans).[7] The earliest record of the Indian herbal medicine is found in the Rig-Veda, one of the oldest repositories of human learning, compiled between 4500 BC and 1600 BC, in which the medical use of the Sowa plant was described. The Vedic Aryans had knowledge of at least 100 medicinal plants. The later Charak Samhita (100 BC to 800 BC) contains accumulation of over 700 medicines.[7] After the flourishing of Buddhism, numerous classics on Ayurveda (500 BC to 100 BC) that are the outgrowth of the great Vedas were compiled by medical specialists like Bhikshu Atreya, Chakradatta, Sarangadhar, Bangasen, etc. Ayurveda medical books like Kalpastanum or Vrikshayurveda present detailed information on herbal drugs, including their collection information, culture methods, storage methods, timing and dose in regard of methods of drug administration.[7] Despite the existence of magical and religious elements, many of the old Hindu remedies were justified by modern pharmaceutical sciences. For example, concentrated aqueous extracts of *Acacia catechu* was used as an astringent, and it is one of the most powerful astringents known even at the present time.[8]

Ancient Persians quickly inherited herbal medicine knowledge from the Assyrians to the west and Indians to the east with the expansion of the Persian Empire. The thriving of herbal medicine practices was not-surprisingly driven by the religious view of Zoroastrianism. The Vendidad (1st millennium BC to 1st millennium AD?), one of the six Zoroastrian Avesta, categorizes the medical practices of Ancient Persia,

including surgery, herbal medicine, and Zoroastrian incantations, of which the later was the most praised.[9] Herbal drugs mentioned in the Avesta are hempseed, shaeta (a kind of gold or yellow plant), ghana, fraspata (a substance that causes mature and rotting of fruits).[9] The Achaemenid Empire (550 BC to 330 BC) made great strides in medicine as they established schools in medicine and hospitals with Pahlavi literature and a variety of medical equipment.[9] Unfortunately, due to wars and conflicts in Persia, detailed description of herbal remedies are not known in the present time. However, by absorbing and expanding the knowledge of their predecessors, namely, the Mesopotamians and Indians, Zoroastrian medical knowledge based on monotheism could contribute to the later emerging Greek practice of medicine. This occurred despite the cardinal feature of the Greek system, namely polytheism.

Unlike other early cultures who graphically carved their animals while ignoring plants, early Greek cultures, like Minoan civilization in Crete and Dorian Greeks, depicted medically significant plants in their frescos. For example, those from around 1450 BC in Knossos, Crete contains pictures of gourds, irises, sage, etc.[10] Records of ancient Greek medicine from the Pre-Hippocratic era are mainly limited to Homer's Iliad and Odyssey. In those epics, various potions, powders, and poultices were described and administered to different areas of the bodies of the Homeric heroes. [11] Homer's works were a symbol of the dawn of humanism. They were also a sign that ancient Greeks started to rely more on observations and to develop practical treatments using scientific approaches. Still, pre-Hippocratic Greek medicine can be divided into three categories, namely, Asclepius medicine, physical training, and the medical schools. Asclepius medicine is a belief that placing patients in sanatoriums like temples

of the God of Medicine, Asclepius, can lead to the recovery of their health. Both physical training and medical school requires understanding of herbal medicine. The famous medical schools, like Cnidus and Cos, finally gave birth to the Hippocratic medical science. Meanwhile, Pythagoras schools introduced mathematics into medical science. Around 300 drug plants were used by Hippocrates (459 BC to 370 BC), including *Absinthe, Anise, Anthemis, Aristolochia*, *Asphodel*, *Atriplex*, *Polygonum*, *Bryonia*, *Lappa*, *Carduus benedictus*, *Dancus*, *Centaurium*, *Chenopodium*, *Cinnamomom*, *Cinquefoil*, *Coriandrum*, *Cyclamen*, *Teucrium*, *Hyoscyamus*, *Conium*, *Isatis*, *Malva*, *Melilotus*, *Mentha artemesia*, *Olea*, *Parthenium*, *Phaseolus*, *Potentilla*, *Ricinus*, *Solanum*, *Thymus*, *Viola*, etc..[8] Hippocrates classified herbal medicine by physiological activity. Specifically, *Centaurium umbellatum* Gilib were applied against fever; garlic against intestine parasites; opium, henbane, deadly nightshade, and mandrake were used as narcotics; fragrant hellebore and haselwort as emetics; sea onion, celery, parsley, asparagus, and garlic as diuretics; oak and pomegranate as astringents.[12]

Following the Pythagoras school, the ancient Greeks developed medical theories to explain physiological mechanisms of herbal remedies. The philosopher, Empedocles, founder of the Sicilian school of medicine is known for his physical theory of the four elements, fire, air, water, and earth, which were the constant realities behind the physical world. In this regard, Empedocles considered plants pivotal during the interchange of the elements, and thus advocated vegetarianism.[13] Inspired by Empedocles, Hippocrates applied the idea of humorism to medicine and believed that an excess or deficiency of

any of four distinct bodily fluids in a person directly influences their temperament and health.[14]

Aristotle's botanical writings placed great emphasis on empiricism, the biodiversity, and most importantly, the inquiry of biological or physiological causation. Greatly influenced by Aristotle's works, Theophrast (371 BC to 287 BC) further classified more than 500 medicinal plants, with descriptions of their toxic activity and doses, and thus, founded the 'botanical sciences'. Later in the Greco-Roman era, Celsus (25 BC to 50 AD?) quoted around 250 medicinal plants. Dioscorides (40 AD to 90 AD?), "the father of pharmacognosy", compiled the work "De Materia Medica", offering data on 657 medicinal plants. Pliny the Elder (23 AD to 79 AD) recorded about 1,000 medicinal plants in his work "Historia naturalis". Together, Dioscorides' and Pliny's encyclopedic works influenced researchers in Medieval and Renaissance Europe. The Roman pharmacist, Galen (131 AD to 200 AD), compiled the first list of drugs "De succedanus", with notes that dereplicate similar or identical physiological activities of the covered drugs. Galen's work extended Dioscorides' drug list by including new drugs or new therapies. For example, *Uvae-ursi folium*, was and still is applied as an uroantiseptic and mild diuretic today. [12]

After the fall of the West Rome Empire in 476 AD, Greco-Roman medicine and its theories were imported, inherited, and further developed by scholars in the Persian Sassanid Empire and later Saracen Empires. This movement began with digestion of Greek medicinal works by translation of works of Galen and other scholars into Farsi, Syriac, or Arabic. Yuhanna ibn Masawaih (John Mesue the Elder) (777 AD to 857 AD) composed medical texts on a number of topics, including ophthalmology, fevers,

leprosy, headache, melancholia, etc.[15] Ibn Masawaih pointed out general directions to choose, prepare, and administor single and compound remedies, and he evaluated Galenic medicine based on their practical medical effects. In order to judge and classify medicine, Ibn Masawaih took eight factors[15] into consideration when examining the materials of the medicine, namely, complexion, texture, flavor, odor, color, the timing to harvest and the duration of storage, and the environment of the collection. According to those eight factors, Ibn Masawaih thought heavier, denser, thicker, tenacious characters made desired medical materials; however, when it comes to purgative drugs, the lightest were always preferred. In regard to the complexion of a drug, warming, humidifying materials were generally preferable. For the texture of medicine, softer or smother textures are better. For odor and flavor, pleasant, sweet taste and sweet scent are more beneficial to health. Ibn Masawaih advised the drug manufacturers to know the timing of plant harvest and drug expiration date. According to him, sweet or salty materials should be harvested at medium maturity, and thin or bitter materials should be collected when fully ripe. Furthermore, the environment of the harvest site was an important factor, because humid materials grow better in dry conditions and vice versa. Herbals of strong drug effect are dispersed, while plants with weaker virtues should be cultured in a concentrated area.[15] We now know that the pigments that plants produce determine their colors. Also, we now know that the physical environment does provide hints for potential bioactivity of the target species during sample collection. For instance, if a marine cyanobacterium lives as a macroscopic tuft in a competitive environment such as a coral reef, it most likely produces noxious natural products that keeps it from being eaten.

Ibn Masawaih outlined the basis of pharmaceutical technology by setting up four operations of post-harvest processing of herbal materials to enhance or change the properties of collected samples. By correctly applying those new technologies, he thought the drugs could be rendered safer and more effective. The four operations are decoction (which is the act of heating), infusion (which is the action of soaking in boiling liquid), lavation (which is the action of rinsing or washing at room temperature), trituration (which is the action of grinding, crushing or squeezing). The extent to which these operations were performed depended on the evaluation results of the medical materials concerning the eight factors. Generally, thick, heavy materials required extended and intensified heating and could withstand forceful grinding. In contrast, delicate, brittle, or thin materials must be heated at lower temperature for shorter time and should avoid washing. In short, post-harvest processing of plant samples should be precise and carefully preserve the essences of the plants.[15] Naturally, in this regard, when heavy and light materials are mixed, there is a demand that they should be separated.

Another major contemporary medical figure of Ibn Masawaih was Hunain Ibn Ishaq (809 AD to 873 AD), a student of the former. Hunain ibn Ishaq and his students travelled to the Greek Byzantine Empire to learn rare Galenic works. Totally, Ibn Ishaq and his students accurately translated and extensively commented on 129 Galenic works into Syriac or Arabic.[16, 17] Their labors provided the Arabic world with more Galenic classics than survive today in their original Greek. By digesting and absorbing Galenic medical classics, Ibn Ishaq and his students created a consistent medical system, which

benefited the pharmaceutical science of their own time, as well as the bloom of medical writing in 10th to 13th century AD in the middle east.

Under the influence of Galenic medical theories, as early as in the 8th century, Jabier ibn Hayyan (721 AD to 815 AD?) extracted anaesthetic materials from Persian herbs for local or general anesthetization purpose.[18] Ibn Hayyan also explored methods to extract and purify herbal essences. For example, he applied beaker, cucurbit and retort to distill and crystalize citric acid, acetic acid and tartaric acid.[19] In the 10th century, with the development of chemical tools, Muhammad ibn Zakariya al-Razi (865 AD to 925 AD?) and Abu Ali ibn Sina (980 AD to 1037 AD) extracted and purified more essences of medical materials. Razi was known to discover and purify alcohol and further use alcohol as a solvent for drug extraction from herbal plants.[20] Ibn Sina was known by his Canon of Medicine with encyclopedic content, systematic arrangement, and combination of Galenic medicine with Aristotle's logic and philosophy, which was widely accepted in Europe by the end of Medieval times.[21] Later in the 13th century, Ibn al-Baitar (1197 AD to 1248 AD) compiled his book "Compendium on Simple Medicaments and Foods", describing over 1,000 medicinal plants, of which around 300 medicines were novel drugs at that time. Persian and Saracen medical science and technology laid the foundation for European medicine. Techniques they developed, such as distillation, crystallization, and the use of alcohol as an antiseptic and a solvent, are widely applied now.

From what has been discussed above, we can conclude that the development of pre-Renaissance herbal medical science is boosted by the following factors. 1) A Humanism belief (See Pericles' Funeral Oration by Thucydides[22]) makes people realize

happiness. Body health can be defined, and thus, looking for drugs and applying correct medical treatments can achieve the cure of diseases. 2) A database of herbal medicine is essential. Unfortunately, the accumulation of medical data has been slow but is highly desired. Expansion of the world's knowledge of natural product medicines has been a long and continuous effort, and has required a peaceful and free environment, and thus should not be interrupted by wars or government regulations. Furthermore, classification of medical materials within this database according to their properties and origins is essential. 3) New technologies, such as distillation and crystallization, as well as the hardware for these processes, have continuously spurred the data accumulation process. 4) New concepts of medicine such as the "four humors" by Galen, the "eight factors" to categorize medical materials and the "four operations" to separate or purify those materials by Ibn Masawaih has contributed to the upgrading of pharmaceutical technologies as well as the expansion of the medical database. 5) Observation, experience and experimentation leads to more in depth and accurate scientific knowledge of medicinal materials. Therefore, it is within my PhD study that applications of new modern technologies to accelerate drug discovery and to develop modern pharmaceutical science are explored.

1.2 Acquiring Natural Products in the History (Isolation Methods)

As discussed in the previous section, natural products (NPs) provided the only source of pharmaceuticals for thousands of years. One thing we know about innovation in natural products research is that it typically occurs when natural product researchers

who have mastered two or more quite different fields use the framework in one to think freshly about natural products research itself.

The accumulation of knowledge of medicinal plants was further promoted by the invention of the Gutenberg Printing Press in the middle 15th century. The availability and popularity of medicinal books lowered the threshold to access medical knowledge. These printing techniques also paved the way for preservation and further accumulation of natural products databases. Another benefit of better printing techniques was that they promoted the dissemination of ideas of observation, experience and experimentation. An example of this influence can be found in ACT IV, sc. i. of William Shakespeares' "A Middle Summer Night's Dream" (1595).

"So doth the woodbine, the sweet honeysuckle

Gently entwist; the female ivy so

Enrings the barky fingers of the elm.

O, how I love thee! how I dote on thee!"

In this quotation, woodbine, also known as the great convolvulus, is a kind of climbing plant that grows into a left- handed helix, while honeysuckle is another kind of climbing plant that grows into right- handed helices.[23] Shakespeare here uses the entwisting of the two plants as a metaphor of the passion of love between Bottom and Titania. Another example of natural history observation recorded in literature can be found in Samuel Taylor Coleridge's "Work without Hope" (Feb 21st,1825).

"All Nature seems at work. Slugs leave their lair —

The bees are stirring — birds are on the wing —

And Winter slumbering in the open air,

Wears on his smiling face a dream of Spring!

And I the while, the sole unbusy thing,

Nor honey make, nor pair, nor build, nor sing."

In this quotation, slugs are observed to look for (building) lairs in spring of the northern hemisphere but not mate. Slugs are nocturnal and hermaphroditic, which usually results in breedin in the nights during the fall and laying eggs before winter comes.[24]

Following the scientific method advocated and practiced by Francis Bacon, progress in chemistry and chemical engineering in the 18th century led to a deeper understanding of the basic chemical elements and chemical reactions that give rise to natural products. Soon after, these advances resulted in practical uses in the pharmaceutical industry as well as natural products research. In the 19th century, developments in physiology, biochemistry and microbiology led to the field of pharmacology.[25] Rational medical treatments began to be widely accepted in clinical and scientific fields. Ever since, the extraction and purification of natural products has contributed to laboratory and industrial drug production by inspiring efforts in chemical synthesis and biotechnology. In this regard, new extraction, isolation, and purification technologies have revolutionized natural products discovery, yielding numerous bioactive compounds. For example, chromatography techniques[26] described by Mikhail Tsvet have facilitated the profiling and purifying of compounds from complex mixtures containing substances that are very close in polarity or other properties.

The desire to isolate bioactive molecules from plants or microbials gave birth to modern organic chemistry and the pharmaceutical industry. As more natural products

of commercial or therapeutic values were isolated and purified, researchers like Liebig and Woehler[27] began to study natural products that possess the same atomic composition but very different properties, e.g. isocyanic acid/cyanic acid (stable) and fulminic acid (explosive). We now know these types of compounds by the term 'isomers' as described by Berzelius[28]. Discovery of isomers led chemists to realize that natural products were defined not only by the atomic compositions but also by the arrangement and connectivity of the atoms. Atomic compositions and their arrangement together decide the properties and bioactivities of natural products. Following this hypothesis, Woehler synthesized urea from the starting material ammonium nitrate, which laid the foundation for natural products structural determination.

Structural determination of natural products is a scholarly pursuit that has been benefited by multidisciplinary efforts, integrating knowledge from such diverse fields as geometry, total synthesis, crystallography, spectroscopy, microbiology, structural biology, chemical biology, etc. Rewardingly, the unequivocal characterization of natural product structures fuels the development of other disciplines, such as chemical ecology, biosynthesis, and pharmaceutical sciences. A famous example in this regard is that Louis Pasteur, who had a background in crystallography, studied the bacterial impact on stereochemistry of tartaric acid[29].

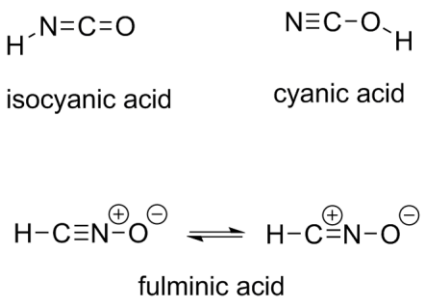


Figure 1.1. The structures of isocyanic acid, cyanic acid and fulminic acid.

In early 19th century, with more chemical structures being available, de Gassicourt et al. proposed that pure compounds are favored over mixtures for pharmaceutical purpose[30]. It should be stressed that the positive bioactivity of chemical mixtures should not be undermined, however, the constituents of the mixtures and their quantity should be made clear. Following de Gassicourt's principals, many isolation methods were developed to obtain pure compounds. An early example in this regard is that in 1817 Friedrich Sertuerner isolated pure morphine by extracting opium with hot water and precipitating colorless morphine crystals using ammonia. Other bioactive compounds extracted from plants in the 19th century included salicyclic acid, digitoxin, ouabain, pilocarpine, and cocaine[31-35].

Common plant extraction techniques still in use today include liquid-liquid extraction, solid-liquid extraction, maceration, expression, digestion, percolation and so on. In natural products research, extraction and initial fractionation has been standardized to assure an extract of consistent quality. In modern natural products research, new chemical discovery as well as dereplication of known compounds is achieved using proper extraction procedures and solvents of different polarity, and

multiple analytical tools, such as high-performance liquid chromatography, mass spectrometry, nuclear magnetic resonances spectroscopy, ultraviolet-visible, infrared, X-ray crystallography and others. Additionally, biosynthetic techniques such as DNA sequencing and bioinformatics, is providing complementary pieces of information in regard to natural products structural determination. However, the modern techniques of natural products research gives rise to a large amount of analytical. In order to turn these data into useful information regarding the structure of the compound, if it is known or closely related to a known compound, a thorough and extensive literature search is required, which is time consuming.

From what has been discussed in this section, we can see that historically, ground-breaking technical innovations in the natural products research are creative modifications and/or combinations of existing technologies of the time. The reason for new and better technologies to replace existing technologies is not that outdated technologies are useless but that new technologies are more efficient and become widely adopted by users. Outdated technologies can still be creatively put into new use, but embracing new technologies generally benefits the discipline itself.

1.3 Potentials of Natural Products Industrialization

In the 20th century, organic chemistry delivered better pharmaceuticals that were created by modifying natural products (semi-synthesis) or through total synthesis of natural product-like molecules. While an in-depth understanding has been achieved for natural products as small molecules acting on druggable pharmacological targets, many new biological features of the natural products are emerging. For the modern

pharmaceutical industry, large quantities of new drugs with multiple modes of bioactivity are required to effectively cure diseases that require new therapeutic approaches. In this regard, decades of studies have shown that a reliable basis for the discovery of a new drug is to pattern it after drugs of known chemical structure, as chemical structures determine how the small molecules act on target proteins and thus, the biological activity of the small molecules.

Over 70% of all medicines are either natural products or their derivatives[36]. Natural products continue to be a prolific source of clinical candidates and drug leads. These compounds remain an important inspiration for novel medicines. Many have subsequently advanced into clinical trials or onto the market. By 2014, the US Food and Drugs Administration (FDA) had approved 1453 new chemical entities (NCEs), of which 547 were natural products and/or derivatives[37-39], with an average of 7.7 natural products and/or derivatives approved annually. Among those approved natural products, most of them have an origin from microorganisms. Famous examples of clinically approved natural product medicines are the antibiotic penicillin (**1**)[40] and the anti-cancer agent paclitaxel (**2**)[41].

Other remarkable natural products include anti-microbial agents (retapamulin (**3**) and fidaxomicin (**4**)), anti-malarial agents (e.g., artemisinin (**5**)) immunosuppressants (rapamycin (**6**) and cyclosporin (**7**)), anti-inflammatory agents (colchicine (**8**) and cannabidol (**9**)), anti-diabetes (e.g., resveratrol (**10**)), blood cholesterol lowering agents (e.g., lovastatin (**11**)), and anti-cancer agents (capsaicin (**12**), and arglabin (**13**))[42, 43].

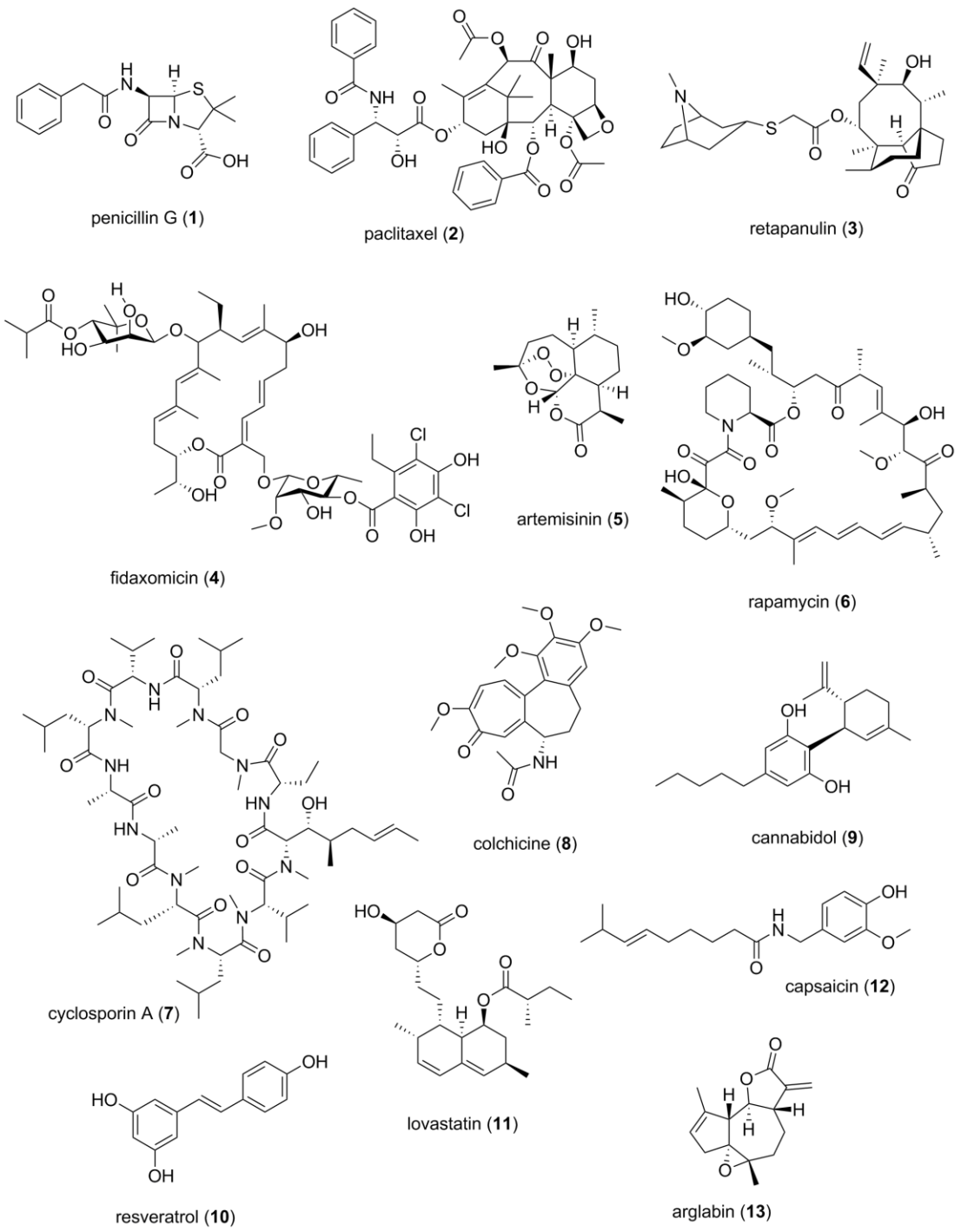


Figure 1.2. A selection of bioactive natural products.

Natural products discovery continues to be prioritized by pharmaceutical industries, either directly or through the acquisition of small companies that engage in natural products research (see discussion below). This is due to the failure of other drug discovery methods, such as combinatorial chemistry, which failed to yield anticipated drug leads in key therapeutic areas, such as antibiotic drug resistance, metabolic diseases and the neurosciences. Currently, it is widely agreed that natural products continue to be major sources of novel drug leads.

Although natural products research has been considered indispensable in drug discovery, in recent years pharmaceutical industries either terminated or substantially cut their investment in natural product research. This is due to the structural complexity of natural products and the time consuming and expensive nature of their structure determination. In addition, total synthesis of complex natural products is still a challenge, resulting in problems of supply. Fortunately, novel drug discovery technologies and increased knowledge of natural products have once again put natural products discovery into the spotlight of the pharmaceutical and biotech industries. The pursuit of novel natural product-derived drug leads is driven by profitable therapeutic demands to treat diseases or clinical conditions, such as cancer, neuronal system diseases, metabolic diseases and multi-drug resistance. Plants, microorganisms, fungi and bacteria have been invaluable for leads in drug design. Because bioactive natural products have been obtained from terrestrial and marine plants, microorganisms, fungi, and bacteria, the pharmaceutical industries are integrating and supporting efforts aimed at the conservation of biodiversity. Indeed, natural products research provides a very strong rationale for the importance of these conservation efforts and the sustainable use

of biodiversity. Meanwhile, natural products discovery provides benefits to the source countries and local communities who are the stewards of those natural product resources[44].

However, in the developed world, there is a chronic shortage of investment in research and development in poverty-related diseases, which occur almost exclusively in poor, developing countries. This is because they offer negligible marketable or profitable returns to research and industry. It seems too costly and too risky to invest in drugs for neglected diseases that occur almost exclusively in developing countries, particularly those with the lowest per capita income and least infrastructural development[45].

Neglected Tropical Diseases (NTDs) are an example. NTDs create one of the heaviest social burdens on the society of tropical countries, including those in East and Southeast Asia, Oceania, Central and South America. With the potential for economic stagnation, this burden has the potential to become catastrophic. NTDs, such as leishmaniasis, Chagas' disease, malaria and schistosomiasis, are among the diseases causing the greatest social burden. Schistosomiasis alone affects an estimated 32 million people in the Pacific Region[46], threatens 240 million people worldwide and contributes to approximately 500,000 deaths per year (https://tinyurl.com/yaq8peuu).

Besides death, NTDs commonly result in long-term illness, impaired childhood development, and/or severe and permanent disabilities and deformities. Long-term illness and impaired childhood development lead to further decreases in productive capabilities in affected regions like China, Laos, Burma, East Timor, Papua New Guinea, the Solomon Islands, Panama, and others, which are developing countries with a large

lower-income population (https://tinyurl.com/d3non2). As a consequence, the health care systems in endemic areas of these countries are underdeveloped, and people generally live with a lack of education, poor sanitation, and substandard housing conditions. Worse still, because of the high price of pharmaceuticals in those endemic areas, and high drug development cost, reduction in the burden of NTDs and possible elimination of the infection are hindered. Thus, non-profit organizations and charity groups are expected to step in to provide financial and technological support in this regard.

1.4 Historical Investigations into Marine Natural Products

The ocean is the largest biome on earth, covering 71% of the surface of the planet. The sea is the origin of the initial diversification of life on earth. The earliest known fossils are marine stromatolites, laminar structures produced by ancient cyanobacteria, discovered in Greenland[47], between 3.9 to 3.7 billion years old. In comparison, the oldest terrestrial fossil record corresponds to Palaeozoic fungi, Ornatifilum lornensis[48], dating back to about 440 million years ago. Marine organisms have thus had more time for genetic diversification than their terrestrial counterparts. Furthermore, marine organisms have to adapt to different stressors, like high salinity, high pressure near the deep sea floor, high temperatures near thermal vents, etc., which has resulted in the further diversification of marine life. Therefore, the marine environment is an extraordinary reservoir of biodiversity.

There is a strong biological and ecological rational for marine organisms to produce bioactive secondary metabolites. The fact that marine organisms such as

tunicates, sponges, cyanobacteria, macroalgae, and gastropods, are oftentimes sessile indicates their need for defensive, offensive or digestive weapons to ward off or kill their predators or grazers. In this regard, marine organisms have evolved complex chemical warfare mechanisms[49] that involve the production of a variety of chemically diverse secondary metabolites, which bind to specific receptors associated with their competitors or predators. Besides, relationships between eukaryotic hosts and environmentally acquired microbial symbionts have been observed and investigated as long-term interactions between marine organisms, and result in a variety of relationships, such as mutualism, parasitism and commensalism. The sessile nature of certain marine invertebrates, such as marine sponges, has caused them to acquire defensive, offensive or digestive toxic natural products from their symbionts. Many of those toxins[50] were shown to be drug leads for human diseases as well as to have ecological functions such as predator deterrence. Commonly, two types of toxic mechanisms for those marine substances have been observed: active mechanisms where toxins are actively administrated into the foreign organisms; passive mechanism, whereby an organism repels competing organisms from intruding its territory (e.g. antibiotic-type toxins are probably examples of passive defense). Furthermore, natural products based chemical communication[51] strongly impacts marine populations and community processes, such as foraging, dominance, mating and breeding.

One of the earliest scientific collection trips was the Challenger expedition[52] of 1872 to 1876. During this expedition, marine collections were made mainly by “blind” trawling and dredging, which meant that the collector did not have a view of the collection site, nor the environment of the collected samples. The invention of the self-

contained underwater breathing apparatus (SCUBA)[53] made accessible collection sites to a maximum water depth of about 50 m below sea level. The shallow water marine habitats to these depths harbor a great diversity of marine organisms that are rich producers of natural products.

Thanks to technologies like the SCUBA, chemical investigations into the collected marine species quickly became a rewarding discipline. Marine natural products (MNPs) discovery efforts have also been driven by the need for new chemical structures, because the combination of a new chemical structure and unique bioactivity improves the chances of a compound becoming a useful pharmaceutical. Furthermore, the occurrence of multi-drug resistance bacteria is increasing faster than the approval rate of new drugs from any source. In this regard, the rich biodiversity of marine organisms is offering a vast resource of potentially useful therapeutics.

Over the years, scientists have been intrigued by the effects of thousands of marine toxins from diverse organisms, such as tunicates, mollusks, macroalgae, and sponges. These have been observed during toxicology or ecological studies of marine invertebrates, beginning with initial investigations in the late 1960s. The first MNP discovery was reported in 1940 by Lederer et al., a red-violet pigment obtained from the sea anemone Calliactis parasitica[54]. Since the 1950s, the number of MNPs isolated from marine organisms has exponentially increased. In 1951, Bergmann et al. reported the isolation of unprecedented arabino- and ribo-pentosyl nucleosides[55-57] obtained from the marine sponge *Cryptotethya crypta* collected off shore of the Florida coast. Those nucleosides eventually give rise to the clinical anti-cancer derivatives vidarabine and cytarabine. Later, in the early 1970s, a total of 191 MNPs were documented[58] by

Scheuer. Since then, continued basic scientific research on the chemistry and pharmacology of MNPs gave rise to the first marine drug, ziconotide[59] (alias, ω-conotoxin MVIIA), a peptide isolated from a tropical marine cone snail *Conus magus*. In 2004, the ziconotide was approved by the US Food and Drug Administration (FDA) under the trade name Prialt for the treatment of chronic pain such as from cancer. The anti-cancer drug trabectedin[60] (alias, yondelis or ecteinascidin-743), originally obtained from the tropical marine squirt *Ecteinascidia turbinate*, was approved in 2015 by US FDA for the treatment of liposarcoma and leiomyosarcoma.

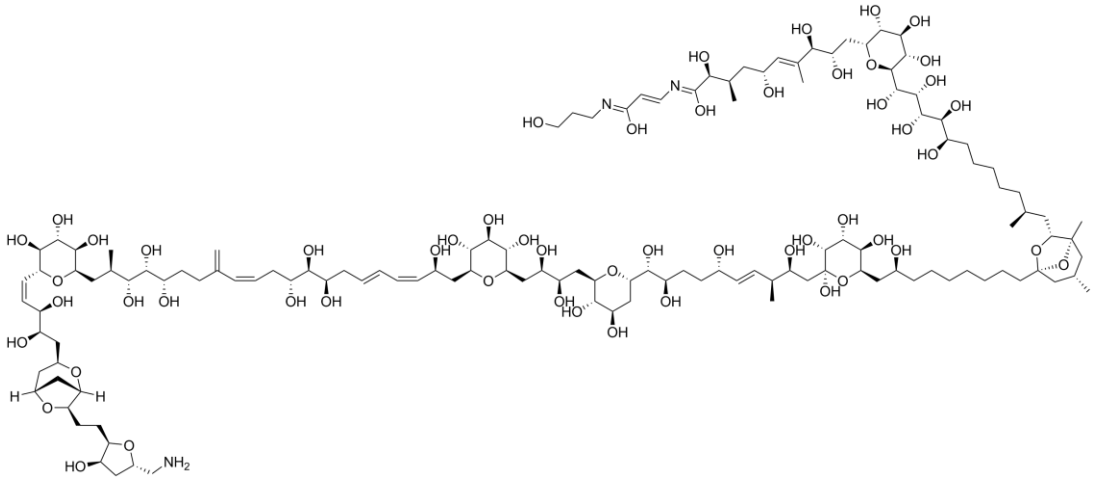


Figure 1.3. The structure of palytoxin (**14**).

Palytoxin (**14**)[61], a polyketide produced by several anthozoans, is an intense vasoconstrictor. (2-Hydroxyethyl) dimethylsulfonium (**15**)[62] isolated from bryozoan sea chervils, *A. gelatinosum*, *A. hirsutum* or *A. diaphanum*, is the cause of Dogger Bank Itch Syndrome among North Sea fishery workers. The macrolide lactones, bryostatins (**16**),[63] also found in bryozoans, are currently under clinical trial for anti-cancer, anti-HIV, and Alzheimer's disease. A series of saponins, asterosaponins (**17**)[64-66] isolated from starfish

and sea cucumbers that belong to the phylum Echinodermata, showed multi-modal bioactivities such as hemolysis, anti-cancer, anti-viral, mammalian neuromuscular transmission blockage, anti-inflammatory, and analgesic properties. Tetrodotoxin (**18**)[67], a potent sodium channel blocker, was found in some platyhelminthes (e.g. marine worms) as well as tetraodontiformes (e.g. pufferfish). Conotoxins[68], polypeptides found in Conidae in the phylum of Mollusca, were determined to be potent sodium channel inhibitors or acetylcholine receptor inhibitors. The depsipeptide kahalalide F (**19**)[69] isolated from both the Sacoglossan, *Elysia rufescens,* also in the phylum of Mollusca, and the green alga *Bryopsis* sp., is currently in Phase II clinical trial against cancer such as melanoma, and hepatocellular carcinoma. Tetramethylammonium containing natural products[70], such as murexine (**20**) and its derivitives were utilized by species of Tonnidae and Muricidae as active cholines. The sponge *Hymeniacidon* sp. produces a feeding deterrent sesquiterpene isocyanide[71] (9-isocyanopupukeanane (**21**)) that possesses a very unique structure. Halichondrin B (**22**)[72], a potent mitotic inhibitor isolated from a sponge collected in the Japan Sea, *Halichondria okadai,* is toxic against a range of tumors, such as breast carcinoma[73] and liposarcoma[74], and has been exploited as a template for the development of the clinically approved drug, Eribulin (**23**). Ciguatoxin (**24**)[75], brevetoxin (**25**)[76, 77], saxitoxin (**26**)[78], and maitotoxin (**27**)[79] now known to be produced by dinoflagellates, are powerful neurotoxins that bind to the mammalian sodium channel. Also isolated from dinoflagellates, okadaic acid (**28**)[80], is a primary cause of diarrhetic shellfish poisoning. It was later was shown to be a potent inhibitor of protein phosphatases. Domoic acid (**29**),[81] originally isolated from two red algae in Japan, caused human sickness upon ingestion of mussels (*Mytilus edulis*). In

the latter case, the toxin was shown to derive from diatoms. Toxins produced by marine algae can be bio-magnified in shellfish or other marine organism, which ultimately causes food poisoning for mammals as well as human beings.

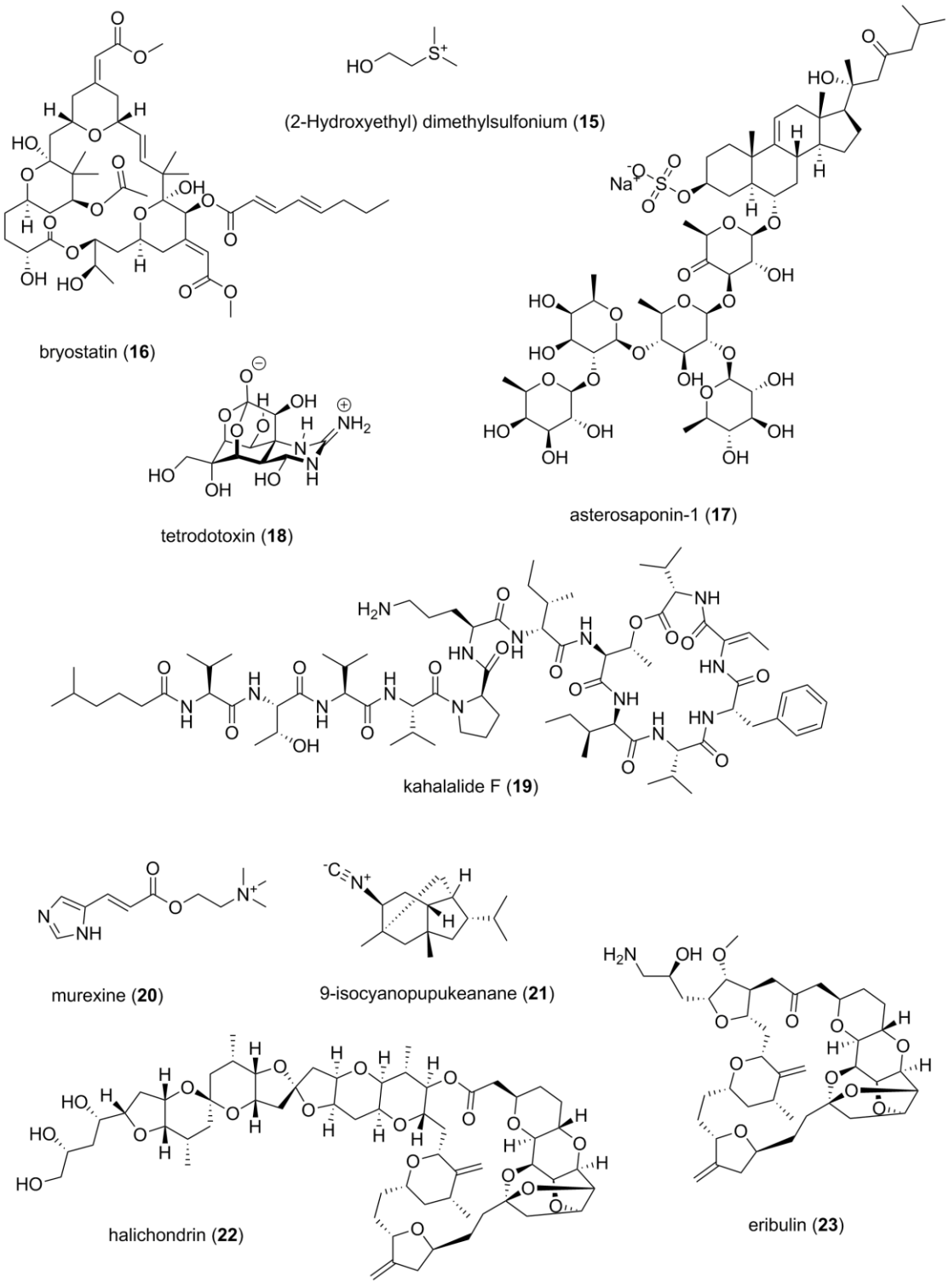


Figure 1.4. The structures of marine natural products (**15**-**23**).

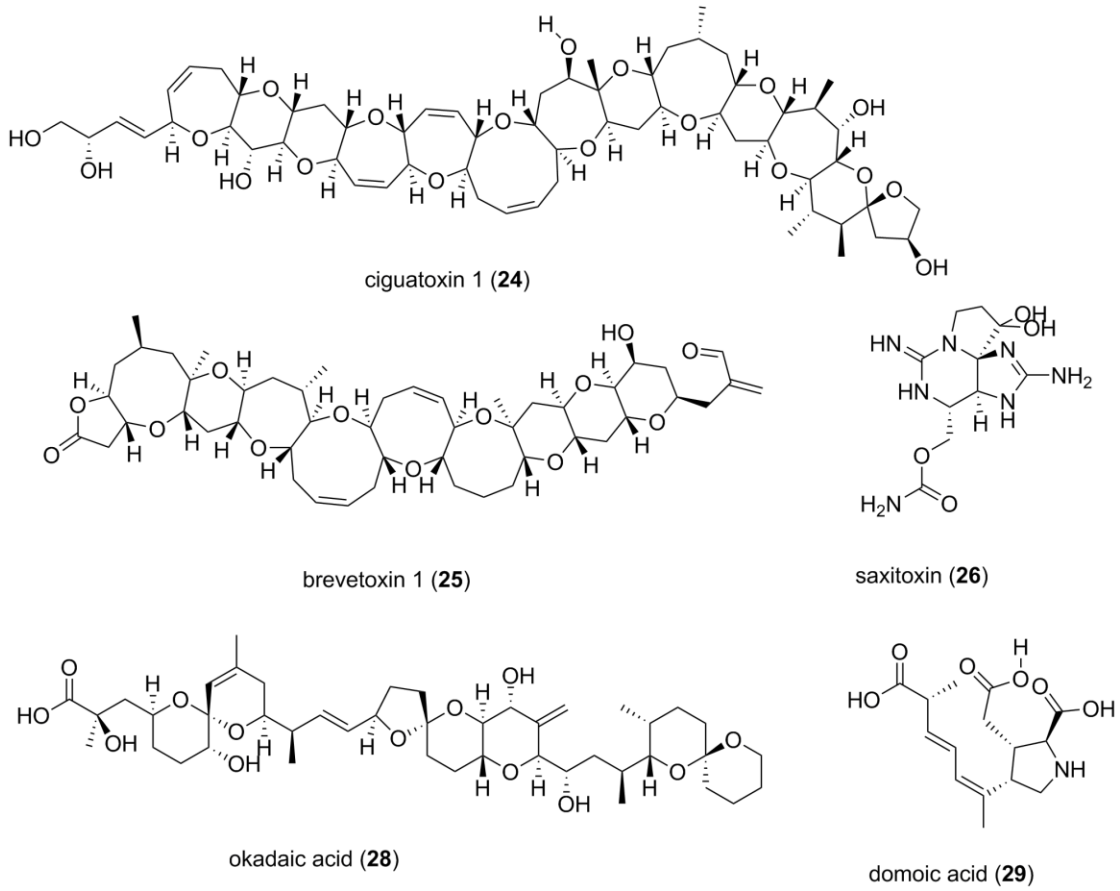


Figure 1.5. The structures of marine natural products (**24**-**26**, **28**-**29**).

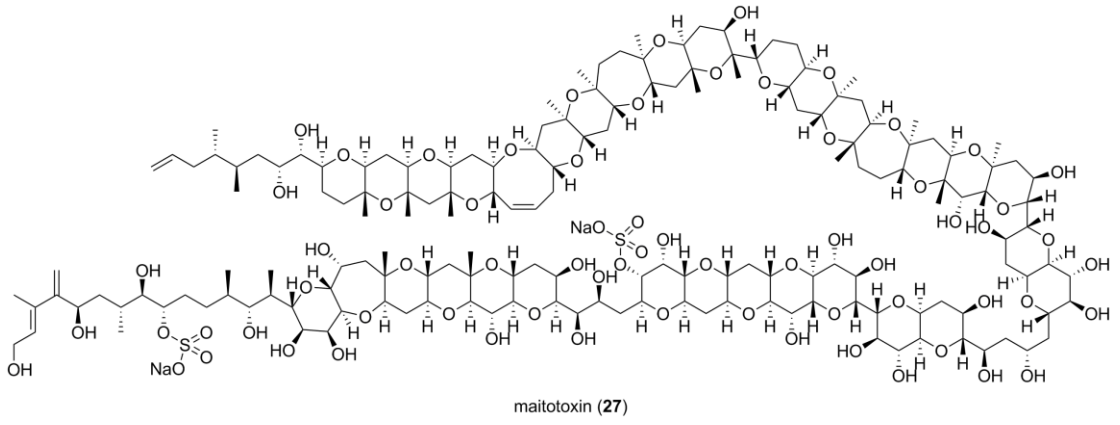


Figure 1.6. The structure of maitotoxin (**27**).

In many cases, it seems likely that marine invertebrates are the producers of their diverse MNPs. However, some filter feeders are known to contain large amounts of symbiotic bacteria. For example, marine bacteria can comprise as high as 40% of the biomass of many sponges[82]. Besides, analysis of the chemical structures and geographic distribution of the particular MNPs suggests that symbiotic microorganisms are oftentimes the true biosynthetic producers of marine invertebrate natural products. Both compound localizations studies and genetic studies have indicated the true producers of the MNPs to be marine microorganisms in a number of cases[83-85].

The highly cytotoxic peptide, dolastatin 10 (**30**)[86], initially isolated from the sea hare of the phylum Mollusca, *Dolabella auricularia*, showed extraordinary anti-cancer activity via an anti-tubulin mechanism of action. Follow-up studies[87] revealed that its close analogues, the symplostatins, as well as dolastatin 10 itself, actually derive from the marine cyanobacterium *Symploca* sp. Therefore, it is believed that the sea hares acquired dolastatin 10 by feeding on marine cyanobacteria that were the true producers of the product. Later, the antineoplastic drug dolastatin 10 was modified, attached to an antibody, and developed into brentuximab vendotin[88], which has been approved for treating anaplastic large cell lymphoma and Hodgkin's lymphoma.

The aforementioned trabectedin (ET-743) (**31**), a clinically approved anti-tumor agent originally found in the Caribbean mangrove tunicate *Ecteinascidia turbinate*, has also been ltraced to be the product of an associated marine microorganism, *Endoecteinascidia frumentensis*. The results of the full genome sequencing of the microbe indicated that the compound was an outcome of a symbiotic interaction between the microbe and its tunicate host[89].

Didemnin B (**32**) and dehydrodidemnin B (**33**)[90] were among the first marine derived agents that went into clinical trials as antitumor agents. Didemnin B was first isolated from the Caribbean encrusting ascidian *Tridemnum solidum*. However, later both compounds were determined to be produced by the potentially symbiotic marine alpha-proteobacteria *Tistrella mobilis*. The didemnin B gene cluster was identified in the genome of this bacterium[91, 92]. Furthermore, imaging mass spectrometry of the microbe revealed chemical precursors to the didemnins.

The structural similarity of the toxic pederins (**34**),[93] produced by terrestrial beetles of the genus *Paederus*, and the marine sponge-derived pederins, such as mycalamides A (**35**) and B (**36**)[94], onnamide F (**37**)[95], theopederins K (**38**) and L (**39**)[96], led researchers to beleive that micro-symbionts could be responsible for the production of these compounds. Using single cell separation technologies, follow-up studies showed that two distinct marine microorganisms from the phyla Tectomicrobia and Entotheonella, were isolated from the Japanese sponge *Theonella swinhoei*. Both of the microorganisms contain gene clusters that produce the onnamides as well as other pederins[97].

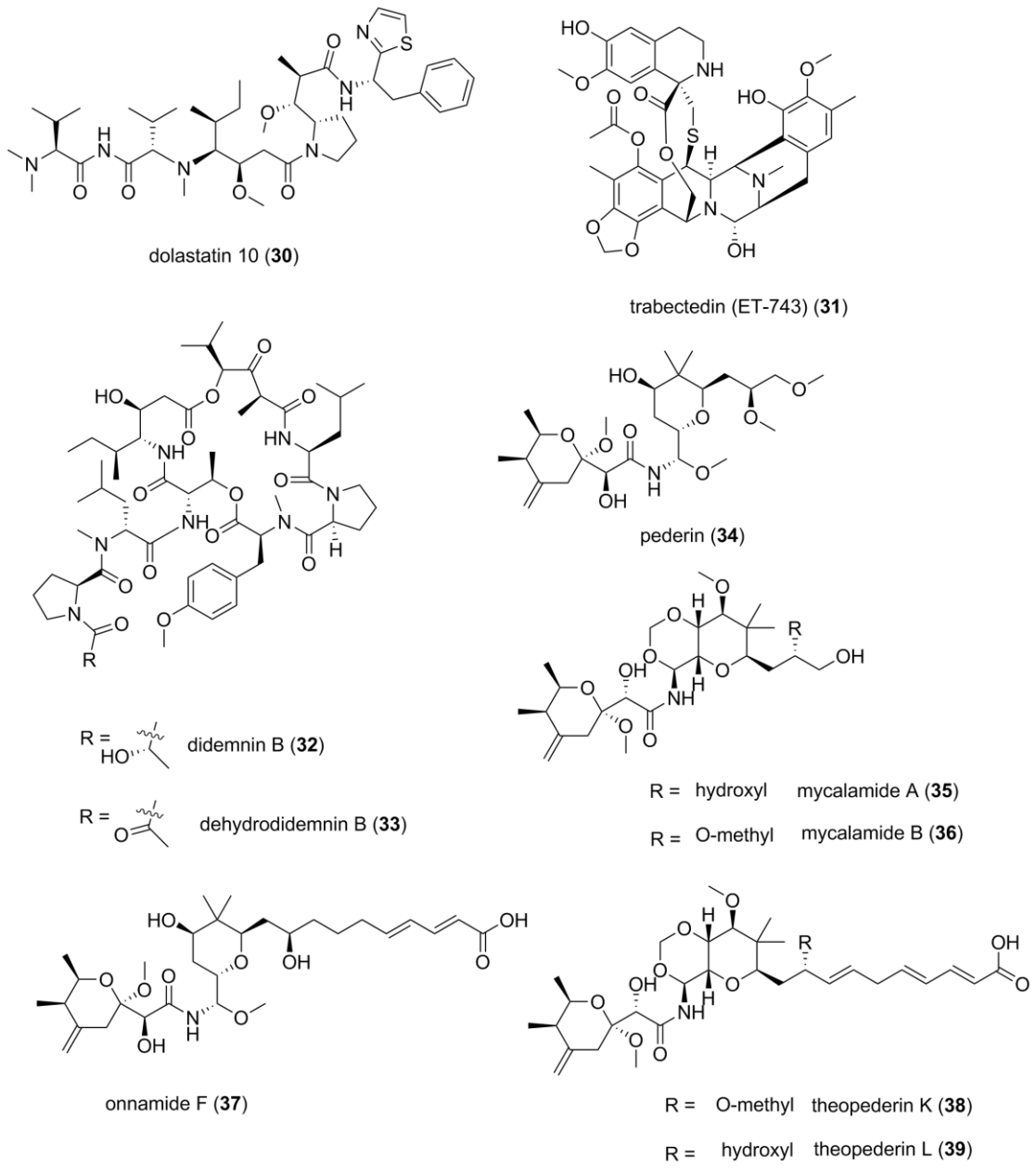


Figure 1.7. The structures of compounds (**30**-**39**).

With a growing number of such examples reported in which microorganisms are the real producers of these bioactive moecules, recent efforts have focused increasingly on marine microorganisms.

By the middle of 2017, a total of 29,605 MNPs have been reported (Marinlit Database, Accessed September 15, 2017). Amongst this dizzying array of compounds, nearly 800 metabolites from cyanobacteria have been identified.[98] Cyanobacteria are a trustworthy source of potential drug leads for several reasons. First, marine cyanobacteria may have evolved this extensive capacity to produce such bioactive NPs because they require defensive weapons that are poisonous to cyanobacteria grazers[99]. Some of their weapons are also poisonous to parasites, cancer cells and treating other human diseases. Further, because cyanobacteria live symbiotically with sponges or fungi, they provide a range of services including photosynthesis, nitrogen fixation, UV protection, and defensive toxins. Cyanobacteria are able to serve these functions using specific chemical molecules that they produce via elaborate biosynthetic processes. In many cases, these cyanobacterial-derived molecules have been found biologically active against cancer cells or human pathogens. Furthermore, cyanobacteria have been a source of quorum-sensing molecules, chemical signals that they use to monitor their own population density and accordingly modulate gene expression[100]; these also have potential drug applications.

Nearly 800 compounds have been obtained from marine cyanobacteria, with a large proportion of those produced by filamentous forms, i.e., the genera *Moorea/Lyngbya*, *Oscillatoria*, and *Symploca*. The presence of non-ribosomal peptide synthetase and/or polyketide synthetase gene clusters in cyanobacteria gives rise to a wide range of structurally diversified bioactive compounds. Several potent anti-cancer type agents have been isolated from cyanobacteria, such as the cryptophycins (**40**-**41**)[101], curacins (**42**-**43**)[102, 103], apratoxins (**44**-**45**)[104, 105], somocystinamide A (**46**)[106, 107], and

others. Some mixtures or proteins produced by cyanobacteria are capable of eradicating virus particles, such as spirulan[108], cyanovirin-N[109], and scytovirin[110]. Efficient anti-protozoa compounds are also reported from tropical cyanobacteria, including, nostocarboline (**47**)[111], bastimolide A (**48**)[112], aerucyclamides (**50**-**51**)[113, 114], and tumonoic acids (**52**-**54**)[115]. Several protease inhibitors are of cyanobacterial origin, including the microginins (**55**)[116], aeruginosins (**56**)[117], and gallinamide A (**57**)[118, 119]. Immunosuppressants like microcolins (**58**-**59**)[120] are found in cyanobacteria. Sodium channel blockers were obtained from cyanobacteria, such as kalkitoxin (**60**)[121, 122].

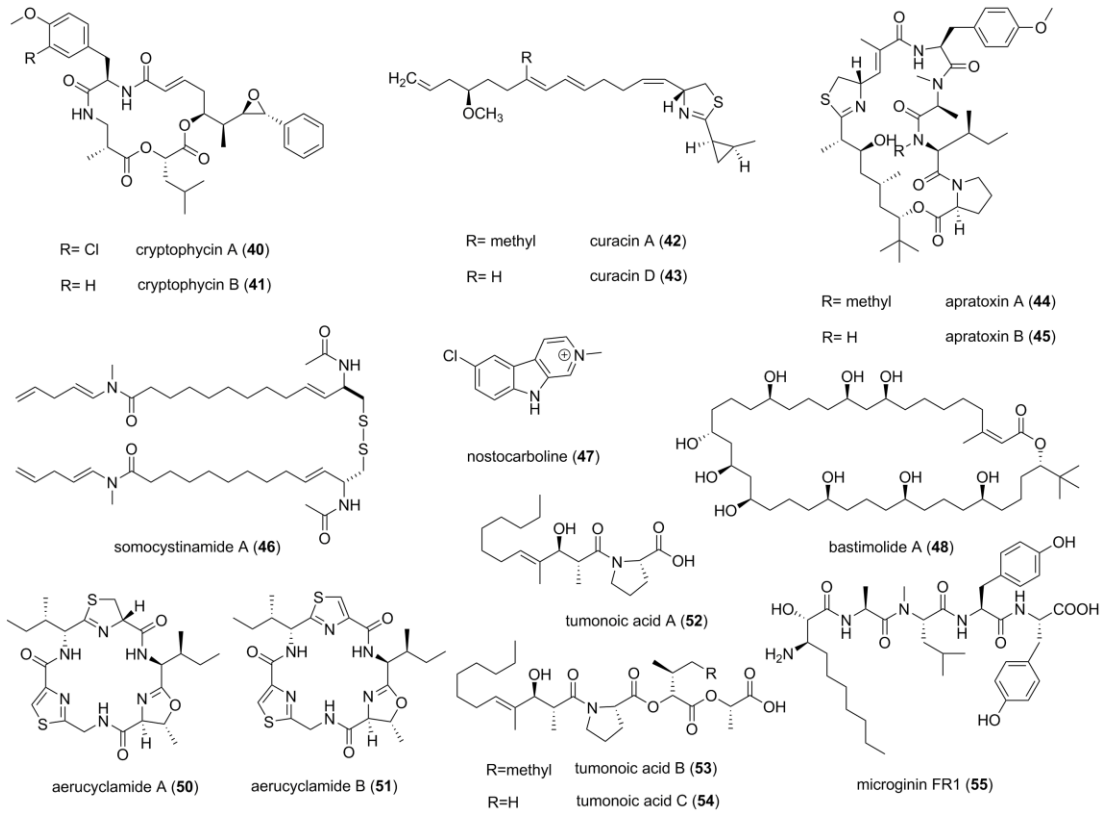


Figure 1.8. The structures of some cyanobacteria derived compounds (**40**-**55**).

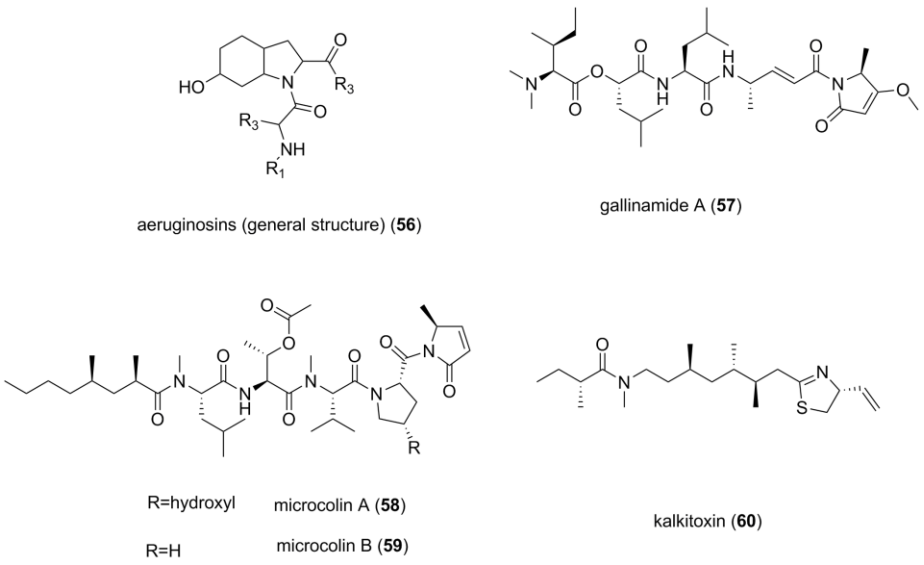


Figure 1.9. The structures of some cyanobacteria derived compounds (**56**-**60**).

In addition to the general bottlenecks in the drug discovery process, the industrial development of many promising MNP is hindered by unique drawbacks, such as having a sustainable source, structural complexity, and difficulty in chemical synthesis. Still, the marine drug discovery and development pipeline remains rewarding and, until now, eight US FDA or European Medicines Agency (EMEA) marine drugs (https://tinyurl.com/glck7xf) are approved and thirteen MNPs or their derivatives are in different phases of the clinical trials.

1.5 Techniques that Enhance Marine Natural Products Discovery

In most MNPs research, novel compound characterization as well as known compound dereplication practices entail collection and analysis of NMR spectra, including running 1D and 2D NMR spectroscopic experiments for the purposes of molecular moiety construction, assemblage and relative stereochemistry determination.

As inspiring advancement of rapid genetic and metabolomics approaches, such as mass spectral guided genome mining[123] and MS-based metabolomics networking[124, 125], have made their way into MNPs research. Nevertheless, conventional NMR practices have been a bottleneck and yet are an indispensable step in the characterization and dereplication of MNPs. In the meantime, conventional NMR practices are considered impracticable for studying secondary metabolites of unculturable or slow-growing microbes. Besides, these microbes are oftentimes living symbiotically with their hosts. In this case, their secondary metabolites are only produced under difficult to reproduce ecological conditions. Furthermore, conventional 2D NMR techniques are notw well suited to studying chemical signaling within microbial communities or between hosts and parasites. The state-of-the-art 1.7 mm cryoprobe NMR instrument in the SSPPS at UCSD has pushed the limits for natural products discovery to just several nanomoles.[126] Combined with MS, ECCD and IR spectroscopy techniques in the past years, we have very powerful techniques by which to study the structures of MNPs. Nevertheless, the conventional step-wise type unselective 2D NMR pulse sequences usually require large numbers of scan which increases the sampling time.

Therefore, it is a question whether using the current cutting-edge NMR hardware and adopting novel pulse sequences or sampling methods, will it be possible to further push the detection limits down to the picomolar scale while maintaining the same sampling time and quality? Or can we work at the picomolar scale and achieve these detection results in even less time or or with fewer spectral artifacts?

Novel pulse sequence programs, such as prompted by Hadamard gate NMR quantum computing techniques alongside with sparse sampling methods,[127] has thus

provided multiple choices to answer the question. However, this depends on the result of experimentation using highly specialized techniques and methods along with the 1.7 mm cryoprobe NMR instrument, and forms the basis of one aspect of my thesis.

1.5.1. Single-Scan 2D NMR Spectroscopy

The single-scan 2D NMR scheme is derived from the concept of liquid NMR quantum computer. An NMR quantum computer is defined by using the spin-up and spin-down states of a spin -1/2 nucleus to process qubits of information simultaneously. A molecule in the solution with multiple nuclear spins may function as a liquid state NMR quantum computer in which each spin constitutes a qubit. A qubit represents a superposition of two quantum states at the same time, while a bit only stands for one state. The Hadamard matrix is used to encode and decode qubits. With the idea of qubits -the superposition of states- in mind, the single-scan 2D NMR technique[128] is crystallized. The single-scan 2D NMR shortens the NMR data recording time to milliseconds by reducing acquisition time and correlated evolution time with a continuous single-scan of the 2D **time-domain** waves. This is achieved through spatially encoding a phase incremented RF irradiation along the *z*-axis of the sample. In this case, the strong magnetic field in the *z*-axis is a gradient instead of being a static field. This spatial phase encoding is kept during the mixing process and later decoded by a second acquisition gradient. The second acquisition gradient simultaneously picks all the site-specific RF signals being generated by the sample. The single-scan multidimensional NMR is compatible with COSY, TOCSY and HSQC experiments.

1.5.2. 2D Hadamard Spectroscopy

2D Hadamard spectroscopy[129] requires encoding and decoding in the **frequency-domain** of the pulses and signals. Frequencies picked from a quick 1D NMR experiment are exploited for frequency-domain excitation. Those frequencies are encoded as 90 degrees or -90 degrees selective pulses. The free induction decays (FID) detection is completed and useful FID signals are extracted followed by Fourier transformation to give the 2D spectra. The Hadamard spectroscopy works for ^{1}H-^{1}H COSY, TOCSY, NOESY experiments. The Hadamard spectroscopy also works for ^{1}H-^{13}C HETCOR and HSQC[130] experiments.

1.5.3. Sparse Sampling NMR Spectroscopy

Conventional 2D NMR spectroscopy applies discrete Fourier transform so that the experiments are very time consuming when generating high frequency resolution in the indirect dimension of the spectra. Non-Fourier transform methods in combination with non-uniform sampling (NUS)[131, 132] allow high resolution along the indirect dimensions in 2D spectra while reducing the sampling time. During the data collection time, some of the increments in the indirect dimension can be be skipped so and later reconstructed using algorithems like Iterated Soft Threshold (IST) or Maximum Entropy.[133] The sparse sampling method is designed for fewer sample collections and delivering an estimation of the fully sampled spectra.

The novel sparse sampling methods, specifically, NUS and compressed sensing are compatible with HSQC.

1.5.4. 2D NMR Spectra Analysis via Machine Learning

Some researchers have applied different algorithms in 2D NMR spectra comparison, like the 2D NMR peak alignment algorithm[134]. But those techniques are not powerful enough to accurately file 2D NMR spectra into an order of NP families. This results from several reasons. Compound concentration, solvent effect, and the gestalt effect of single functional group alternation on environmental ^{1}H and ^{13}C NMR chemical shifts all raise the difficulty for computer assisted 2D NMR data analysis. The second general area is that conventional NMR experiments tend to introduce artifacts into the spectra, for which the existing overlap methods cannot distinguish from genuine peaks. New facial recognition technology, neural networks, derived from cognitive science, has thus generated new thoughts on computer assisted 2D data analysis. Compared with conventional machine learning methods, which require a large basis of known training samples for each category, the machine learning technique is designed for small training samples for each category. The category numbers for metric learning can be very large and unknown during training. Thus the algorithm is a fit for 2D NMR spectra of NPs. There are unlimited categories for compound families and a lot of the families are unknown to the researchers. Plus, each category contains less than 30 different compounds as an estimated average, such as the curacins, apratoxins and lyngbyabellins. It should be mentioned that the high resolution spectra obtained through

the new 2D NMR techniques discussed in the previous section can potentially raise the successful rate of metric learning assisted spectra profiling due to few artifacts in the spectra.

In short, a desired 2D NMR spectra comparison system will perform two tasks: detection and ranking. *Detection:* "Is the value of curacins compatible with spectra A?" This detection job is performed by comparing the scalar "energy" of a compound family label with a threshold value. Here, the term, energy, is an imaginary concept that measures the compatibility of two configurations of the variables. The system must be trained to generate the scalar energy values for 2D spectra, in such a way that the value is large when the input spectra are less like curacins. *Ranking:* "spectra B or spectra C, which is more compatible with curacins?" This is one stage further than detection because the system is to be trained to produce a complete ranking of all outputs, rather than only present the best one.

1.5.5. Absolute Structural Assignment via ECCD

Circular dichroism (CD) is caused by the distinctive absorption rate of left- and right-handed circularly polarized UV light when the light is passing through a chiral media (see Figure 1.10 for CD visualization). Thus, CD is measured as a difference ($\Delta\varepsilon$) in molar absorptivity of chiral compounds in solution for left- and right- handed circularly polarized UV light.

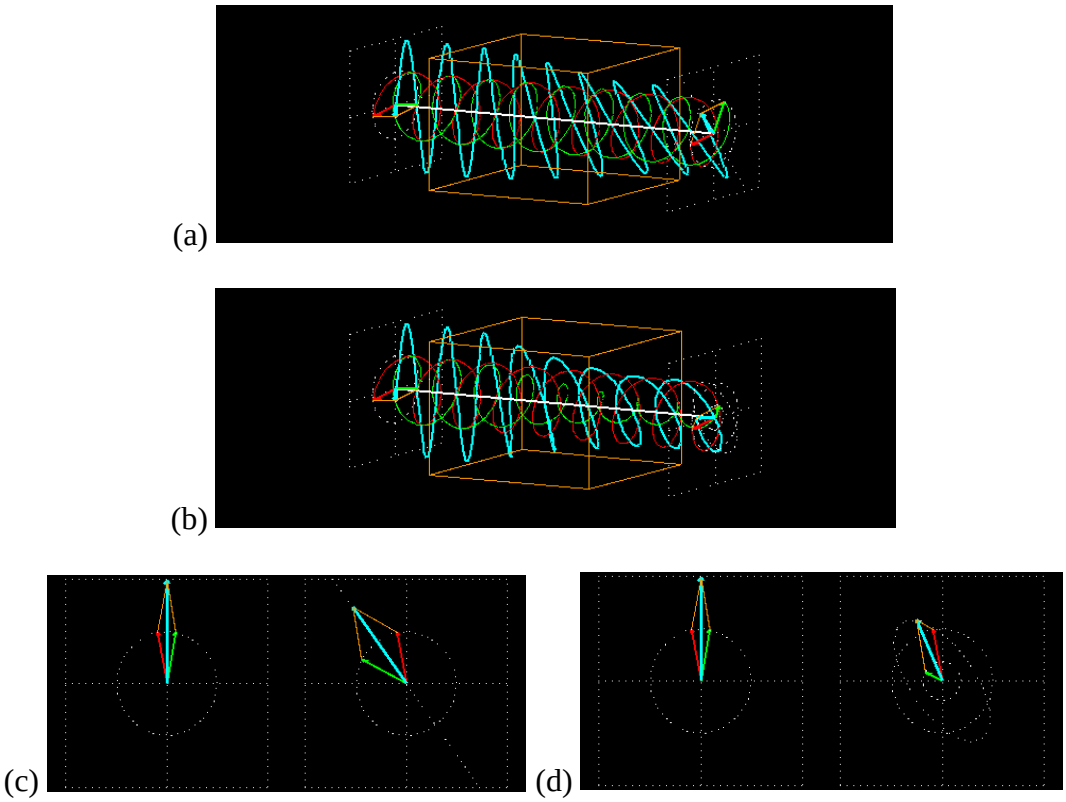


Figure 1.10. The visualization of CD. The green helix in (a) and (b) stands for the right- handed circularly polarized UV light. In (c) and (d), they are presented as the green vector rotating CW around the original point in an x-y plane. In either (a) or (b), the light propagates from left to right. The red helix in (a) and (b) stands for the left-handed circularly polarized UV light. In (c) and (d), they are presented as the red vector rotating CCW around the original point in the x-y plane. The blue helix in (a) and (b) stands for the superposition of the right- handed and the left- handed circularly polarized UV light. In (c) and (d), they are presented as the blue vector in the x-y plane. The orange box in (a) and (b) stands for media. (a) The absorption of left- and right-handed circularly polarized UV light by the media is equal, while the reflection of the two light in the media is different. That's why in (c) the linear combination of the two vectors tilts. (b) The absorption of left- and right-handed circularly polarized UV light by the media is not equal, and the reflection of the two light in the media is different. The amplitude of the right- handed light is damped by the media. (d) CD effect is explained by the blue resultant vector cycling in the elliptic orbital in the same direction as the left- handed light with larger amplitude, after the two light passes through the media.(Pictures Source: http://www.enzim.hu/~szia/cddemo/edemo16.htm)

Exciton coupling between two or more chromophores gives rise to very intense "split" Cotton effects ($\Delta\varepsilon > 20$ $mol^{-1}dm^{-3}cm^{-1}$) that can be useful in assigning the

absolute configuration of a molecule. A pair of enantiomers gives rise to mirror images across the $\Delta\varepsilon = 0$ line in CD spectra. (see Figure 1.11) The signs of the split Cotton effects reflect the absolute stereochemistry of the electric transition moments of interacting chromophores. Previous study[135] has shown that ECCD is suitable for pure compounds at nanomole concentrations. More experiments are needed to test ECCD's picomolar detection ability.

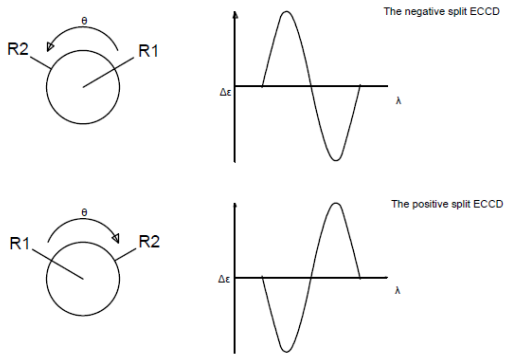


Figure 1.11. ECCD illustration. The through-space coupling of two interacting chromophores in a chiral molecule solution gives rise to ECCD. The ECCD curves are dual signaled CD curves. The left section of the picture contains the Newman projection of two enantiomers. If the dihedral angle θ is measured CCW, the ECCD shows negative first Cotton effect against the positive direction of the wavenumber λ axis (upper right chart). If the dihedral angle theta is measured CW, the ECCD shows positive first Cotton effect against the positive direction of the wavenumber λ axis (lower right chart). R_1 and R_2 are the two interacting chromophores within the two enantiomers. (Pictures drawn with AutoCAD 2015)

1.6 Dissertation Contents

The primary focus of the research described in following chapters is the creation of the Small Molecule Accurate Recognition Technology (SMART) to accelerate

natural products discovery, with a secondary focus on the characterization of anti-cancer secondary metabolites from marine cyanobacteria. Chapter 2 describes the isolation and structure determination of a new secondary metabolite, laucysteinamide A, from a collection of cyanobacterium cf. *Caldora penicillata* from Lau Lau Bay, off the central east coast of the Saipan island, a US territory. The planar structure of laucysteinamide A was resolved by an interplay of UV, IR, and mass spectrometry and 1D or 2D NMR spectroscopy. The new compound is a structural analogue of a previously isolated disulfide dimer, somocystinamide A, obtained from a *Lyngbya majuscula/Schizothrix* sp. assemblage, containing the same terminal olefin and amide moieties. In contrast, laucysteinamide A features a thiazoline ring, a functional group that has been previously observed in other marine natural products, such as curacins A-E[102, 103, 136, 137] and apratoxins A-D, F-H and A sulfoxide[104, 105, 138-140]. Laucysteinamide A was mildly cytotoxic to H-460 human non-small cell lung cancer cells (IC_{50} = 11 μM).

Chapter 3 presents the creation, characterization and application of a digital frontier, a Fast NMR and deep-learning based 2D NMR analysis tool, the Small Molecule Accurate Recognition Technology (SMART) to streamline marine natural products discovery process. The prototype of SMART leveraged the benefits of a Fast NMR technique, the Non-Uniform Sampling (NUS) 2D NMR and a face recognition algorithm, the Convolutional Neural Networks (CNN), to facilitate NMR data collection and analysis. The NUS pulse sequences were successfully adapted to the 600 MHz Bruker 1.7 mm cryoprobe NMR instrument. Meanwhile, the modified CNN was trained with an HSQC dataset containing thousands of spectra. The prototype demonstrated its capacity to over 85% of accuracy to associate unknown HSQC spectra to their known

analogues, as well as reduce the NMR collection time to at least 25% of the original time. Furthermore, SMART possesses several interesting aspects, for example, accurate recognition of 2D NMR spectra even in the presence of noise.

Chapter 4 describes the isolation and structural determination of new marine depsipeptides, viequeamides A2, A3, C, D and aurilide D, from the fractionated extracts of two quite disparate marine cyanobacteria, a *Rivularia* sp. from Vieques, Puerto Rico, and a *Moorea producens* from American Samoa. Also from these two collections, viequeamides A and B were dereplicated. The structures of all the obtained compounds were determined using a combination of LCMS based Global Natural Products Social Molecular Networking (GNPS) and the SMART platform described in the previous chapter. The initial fractions of each sample were first analyzed by LC-MS/MS followed by GNPS. The results showed that several fractions from both species contained analogues of viequeamides. The purified compounds from those fractions were then underwent NUS HSQC experiments. Those HSQC spectra confidently associated to a depsipeptides class most likely related to the viequeamide or veraguamide classes of compounds by SMART. Subsequently, the 3D structures of those compounds were established using various NMR, MS, and X-Ray crystallography techniques.

The dissertation finishes with a conclusion and future work chapter that consists of a brief summary about the chapters 2 to 4 and some exciting potential future directions for SMART.

1.7 Chapter 1 References

1. Labadie, R. P.; Vandernat, J. M.; Simons, J. M.; Kroes, B. H.; Kosasi, S.; Vandenberg, A. J. J.; Thart, L. A.; Vandersluis, W. G.; Abeysekera, A.;

Bamunuarachchi, A.; Desilva, K. T. D., An Ethnopharmacognostic Approach to the Search for Immunomodulators of Plant-Origin. *Planta Medica* **1989**, (4), 339-348.

2. Lietava, J., Medicinal-Plants in a Middle Paleolithic Grave Shanidar-Iv. *Journal of Ethnopharmacology* **1992,** 35, (3), 263-266.

3. Merlin, M. D., Archaeological evidence for the tradition of psychoactive plant use in the old world. *Economic Botany* **2003,** 57, (3), 295-323.

4. Haas, L. F., Papyrus of Ebers and Smith. *Journal of Neurology Neurosurgery and Psychiatry* **1999,** 67, (5), 578-578.

5. Aboelsoud, N. H., Herbal medicine in ancient Egypt. *Journal of Medicinal Plants Research* **2010,** 4, (2), 82-86.

6. Thompson, R. C., *The Assyrian herbal*. Luzac and co.: London, 1924; p 2 p. L., xxvii, 294 p.

7. Chopra, R. N., *Chopra's indigenous drugs of India*. 2d ed.; U.N. Dhur: Calcutta, 1958; p xxxii, 816 p.

8. Wheelwright, E. G., *Medicinal plants and their history*. Dover Publications: New York, 1974; p 288 pages.

9. Elgood, C., *A medical history of Persia and the eastern caliphate from the earliest times until the year A.D. 1932*. University Press: Cambridge Eng., 1951; p xii, 616 p.

10. Tucker, A. O., Identification of the rose, sage, iris, and lily in the "Blue Bird Fresco" from Knossos, Crete (ca. 1450.BCE). *Economic Botany* **2004,** 58, (4), 733-736.

11. Osler, W., The Greek Genius (Reprinted from the Evolution of Modern Medicine, 1921). *Jama-Journal of the American Medical Association* **1988,** 259, (20), 3065-3065.

12. Petrovska, B. B., Historical review of medicinal plants' usage. *Pharmacogn Rev* **2012,** 6, (11), 1-5.

13. Long, A. A., Thinking and Sense-Perception in Empedocles: Mysticism or Materialism? *The Classical Quarterly* **1966,** 16, (2), 256-276.

14. Sigerist, H. E., *A history of medicine : II, early Greek, Hindu, and Persian medicine*. Oxford University Press: New York, 1987; p xvi, 352 p.

15. De Vos, P., The Prince of Medicine: Yuhanna ibn Masawayh and the Foundations of the Western Pharmaceutical Tradition. *Isis* **2013,** 104, (4), 667-712.

16. Brock, S. P., *Aspects of translation technique in antiquity*.

17. Porter, R., *The Cambridge history of medicine*. Rev. ed ed.; Cambridge University Press: Cambridge ; New York, 2006; p 408 pages.

18. Watson, R. R.; Preedy, V. R., *Botanical medicine in clinical practice*. CABI: Wallingford, UK ; Cambridge, MA, 2008; p xviii, 915 pages.

19. Holmyard, E. J., *Makers of chemistry*. 1931.

20. Modanlou, H. D., A tribute to Zakariya Razi (865 - 925 AD), an Iranian pioneer scholar. *Arch Iran Med* **2008,** 11, (6), 673-7.

21. McGinnis, J., *Avicenna*. Oxford University Press: Oxford ; New York, 2010; p xiv, 300 pages.

22. Thucydides; Hammond, M.; Rhodes, P. J., The Peloponnesian War. In *Oxford world's classics*, Oxford University Press,: Oxford ; New York, 2009; pp 1 online resource (lxiv, 708 pages).

23. Darwin, C.; Darwin, F., *The power of movement in plants*. D. Appleton: New York, 1885; p x, 592 p.

24. Bett, J. A., The Breeding Seasons of Slugs in Gardens. *Proceedings of the Zoological Society of London* **2009,** 135, (4), 559-568.

25. Drews, J., Drug discovery: A historical perspective. *Science* **2000,** 287, (5460), 1960-1964.

26. Zechmeister, L., Mikhail Tswett -- The Inventor of Chromatography. *Isis* **1946,** 36, (2), 108-109.

27. Liebig, J., Liebig and Wöhler. In *A History of Chemistry,* Springer: 1964; pp 294-336.

28. Esteban, S., Liebig–Wöhler Controversy and the Concept of Isomerism. *Journal of Chemical Education* **2008,** 85, (9), 1201.

29. Gal, J., Louis Pasteur, language, and molecular chirality. I. Background and dissymmetry. *Chirality* **2011,** 23, (1), 1-16.

30. Berman, A., The Cadet circle: representatives of an era in French pharmacy. *Bulletin of the History of Medicine* **1966,** 40, (2), 101.

31. Mahdi, J.; Mahdi, A.; Bowen, I., The historical analysis of aspirin discovery, its relation to the willow tree and antiproliferative and anticancer potential. *Cell proliferation* **2006,** 39, (2), 147-155.

32. Diefenbach, W. C.; Meneely, J. K., Digitoxin - a Critical Review. *Yale Journal of Biology and Medicine* **1949,** 21, (5), 421-431.

33. Arnaud, A., Organic chemistry. *Journal of the Chemical Society, Abstracts* **1898,** 74, (0), A545-A612.

34. Ferguson, M. M., Pilocarpine and other cholinergic drugs in the management of salivary gland dysfunction. *Oral surgery, oral medicine, oral pathology* **1993,** 75, (2), 186-191.

35. Karch, S. B., Cocaine: history, use, abuse. *Journal of the Royal Society of Medicine* **1999,** 92, (8), 393-397.

36. Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **2016,** 79, (3), 629-61.

37. Patridge, E.; Gareiss, P.; Kinch, M. S.; Hoyer, D., An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discov Today* **2016,** 21, (2), 204-7.

38. Griesenauer, R. H.; Kinch, M. S., 2016 in review: FDA approvals of new molecular entities. *Drug Discov Today* **2017**.

39. Kinch, M. S.; Haynesworth, A.; Kinch, S. L.; Hoyer, D., An overview of FDA-approved new molecular entities: 1827-2013. *Drug Discovery Today* **2014,** 19, (8), 1033-1039.

40. Fleming, A., The story of penicillin. *J Am Inst Homeopath* **1946,** 39, 154-7.

41. Wani, M. C.; Taylor, H. L.; Wall, M. E.; Coggon, P.; McPhail, A. T., Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. *J Am Chem Soc* **1971,** 93, (9), 2325-7.

42. Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J., The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* **2015,** 14, (2), 111-29.

43. Atanasov, A. G.; Waltenberger, B.; Pferschy-Wenzig, E. M.; Linder, T.; Wawrosch, C.; Uhrin, P.; Temml, V.; Wang, L.; Schwaiger, S.; Heiss, E. H.; Rollinger, J. M.; Schuster, D.; Breuss, J. M.; Bochkov, V.; Mihovilovic, M. D.; Kopp, B.; Bauer, R.; Dirsch, V. M.; Stuppner, H., Discovery and resupply of pharmacologically active plant-derived natural products: A review. *Biotechnol Adv* **2015,** 33, (8), 1582-1614.

44. Cragg, G. M.; Katz, F.; David, J. N. A.; Rosenthal, J., The impact of the United Nations Convention on Biological Diversity on natural products research. *Natural Product Reports* **2012,** 29, (12), 1407-1423.

45. Hotez, P. J.; Molyneux, D. H.; Fenwick, A.; Kumaresan, J.; Sachs, S. E.; Sachs, J. D.; Savioli, L., Current concepts - Control of neglected tropical diseases. *New England Journal of Medicine* **2007,** 357, (10), 1018-1027.

46. King, C. H.; Dangerfield-Cha, M., The unacknowledged impact of chronic schistosomiasis. *Chronic Illn* **2008,** 4, (1), 65-79.

47. Nutman, A. P.; Bennett, V. C.; Friend, C. R. L.; Van Kranendonk, M. J.; Chivas, A. R., Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* **2016,** 537, (7621), 535-+.

48. Smith, M. R., Cord-forming Palaeozoic fungi in terrestrial assemblages. *Botanical Journal of the Linnean Society* **2016,** 180, (4), 452-460.

49. Pohnert, G., Chemical defense strategies of marine organisms. *Chemistry of Pheromones and Other Semiochemicals I* **2004,** 239, 179-219.

50. Puglisi, M. P.; Sneed, J. M.; Sharp, K. H.; Ritson-Williams, R.; Paul, V. J., Marine chemical ecology in benthic environments. *Natural Product Reports* **2014,** 31, (11), 1510-1553.

51. Hay, M. E., Marine Chemical Ecology: Chemical Signals and Cues Structure Marine Populations, Communities, and Ecosystems. *Annual Review of Marine Science* **2009,** 1, 193-212.

52. Herdman, W. A., 5. Preliminary Report on the Tunicata of the "Challenger" Expedition. *Proceedings of the Royal Society of Edinburgh* **2014,** 10, 458-472.

53. Dill, R. F.; Shumway, G., Geologic Use of Self-Contained Diving Apparatus. *Aapg Bulletin-American Association of Petroleum Geologists* **1954,** 38, (1), 148-157.

54. Lederer, E.; Teissier, G.; Huttrer, C., The isolation and chemical composition of calliactine, pigment of the sea anemone "Sagartia parasitica"(Calliactis effoeta). *Bull. Soc. Chim. Fr* **1940,** 7, 608-615.

55. Bergmann, W.; Feeney, R. J., CONTRIBUTIONS TO THE STUDY OF MARINE PRODUCTS. XXXII. THE NUCLEOSIDES OF SPONGES. I. *The Journal of Organic Chemistry* **1951,** 16, (6), 981-987.

56. BERGMANN, W.; BURKE, D. C., Contributions to the Study of Marine Products. XL. The Nucleosides of Sponges. 1 IV. Spongosine2. *The Journal of organic chemistry* **1956,** 21, (2), 226-228.

57. BERGMANN, W.; STEMPIEN Jr, M. F., Contributions to the Study of Marine Products. XLIII. The Nucleosides of Sponges. V. The Synthesis of Spongosine1. *The journal of Organic Chemistry* **1957,** 22, (12), 1575-1577.

58. Scheuer, P. J., *Chemistry of marine natural products*. Academic Press: New York,, 1973; p xi, 201 p.

59. McIntosh, M.; Cruz, L. J.; Hunkapiller, M. W.; Gray, W. R.; Olivera, B. M., Isolation and structure of a peptide toxin from the marine snail Conus magus. *Arch Biochem Biophys* **1982,** 218, (1), 329-34.

60. Worthen, L. R.; Marine Technology Society.; University of Rhode Island. College of Pharmacy., *Food-drugs from the sea : proceedings, 1972*. Marine Technology Society: Washington, 1973; p xix, 396 pages.

61. Moore, R. E.; Scheuer, P. J., Palytoxin - New Marine Toxin from a Coelenterate. *Science* **1971,** 172, (3982), 495-&.

62. Carle, J. S.; Christophersen, C., Dogger Bank itch. The allergen is (2-hydroxyethyl)dimethylsulfoxonium ion. *Journal of the American Chemical Society* **1980,** 102, (15), 5107-5108.

63. Pettit, G. R.; Herald, C. L.; Doubek, D. L.; Herald, D. L.; Arnold, E.; Clardy, J., Anti-Neoplastic Agents .86. Isolation and Structure of Bryostatin-1. *Journal of the American Chemical Society* **1982,** 104, (24), 6846-6848.

64. Yasumoto, T.; Hashimot.Y, Properties and Sugar Components of Asterosaponin a Isolated from Starfish. *Agricultural and Biological Chemistry* **1965,** 29, (9), 804-&.

65. Yasumoto, T.; Hashimoto, Y., Properties of Asterosaponin B Isolated from a Starfish Asterias Amurensis. *Agricultural and Biological Chemistry* **1967,** 31, (3), 368-+.

66. Ikegami, S.; Kamiya, Y.; Tamura, S., Studies on Asterosaponins .1. Isolation and Structure of a New Steroid, Asterosapogenine-I, from Asterosaponin-a and Asterosaponin-B. *Agricultural and Biological Chemistry* **1972,** 36, (10), 1777-&.

67. Goto, T.; Kishi, Y.; Takahashi, S.; Hirata, Y., The Structure of Tetrodotoxin. *Tetrahedron Letters* **1963**, (30), 2105-2113.

68. Olivera, B. M.; Hillyard, D. R.; Rivier, J.; Woodward, S.; Gray, W. R.; Corpuz, G.; Cruz, L. J., Conotoxins - Targeted Peptide Ligands from Snail Venoms. *Acs Symposium Series* **1990,** 418, 256-278.

69. Hamann, M. T.; Scheuer, P. J., Kahalalide-F - a Bioactive Depsipeptide from the Sacoglossan Mollusk Elysia-Rufescens and the Green-Alga Bryopsis Sp. *Journal of the American Chemical Society* **1993,** 115, (13), 5825-5826.

70. Blankenship, J. E.; Langlais, P. J.; Kittredge, J. S., Identification of a cholinomimetic compound in the digestive gland of Aplysia californica. *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology* **1975,** 51, (1), 129-137.

71. Burreson, B. J.; Scheuer, P. J.; Finer, J.; Clardy, J., 9-Isocyanopupukeanane, a Marine Invertebrate Allomone with a New Sesquiterpene Skeleton. *Journal of the American Chemical Society* **1975,** 97, (16), 4763-4764.

72. Hirata, Y.; Uemura, D., Halichondrins - Antitumor Polyether Macrolides from a Marine Sponge. *Pure and Applied Chemistry* **1986,** 58, (5), 701-710.

73. Towle, M. J.; Nomoto, K.; Asano, M.; Kishi, Y.; Yu, M. J.; Littlefield, B. A., Broad Spectrum Preclinical Antitumor Activity of Eribulin (Halaven (R)): Optimal Effectiveness under Intermittent Dosing Conditions. *Anticancer Research* **2012,** 32, (5), 1611-1619.

74. Kawano, S.; Asano, M.; Adachi, Y.; Matsui, J., Antimitotic and Non-mitotic Effects of Eribulin Mesilate in Soft Tissue Sarcoma. *Anticancer Research* **2016,** 36, (4), 1553-1561.

75. Scheuer, P. J.; Takahashi, W.; Tsutsumi, J.; Yoshida, T., Ciguatoxin: isolation and chemical nature. *Science* **1967,** 155, (3767), 1267-8.

76. Shimizu, Y.; Chou, H. N.; Bando, H.; Van Duyne, G.; Clardy, J., Structure of brevetoxin A (GB-1 toxin), the most potent toxin in the Florida red tide organism Gymnodinium breve (Ptychodiscus brevis). *J Am Chem Soc* **1986,** 108, (3), 514-5.

77. Shimizu, Y.; Bando, H.; Chou, H. N.; Vanduyne, G.; Clardy, J. C., Absolute-Configuration of Brevetoxins. *Journal of the Chemical Society-Chemical Communications* **1986**, (22), 1656-1658.

78. Schantz, E. J.; Ghazarossian, V. E.; Schnoes, H. K.; Strong, F. M.; Springer, J. P.; Pezzanite, J. O.; Clardy, J., Structure of Saxitoxin. *Journal of the American Chemical Society* **1975,** 97, (5), 1238-1239.

79. Murata, M.; Naoki, H.; Iwashita, T.; Matsunaga, S.; Sasaki, M.; Yokoyama, A.; Yasumoto, T., Structure of Maitotoxin. *Journal of the American Chemical Society* **1993,** 115, (5), 2060-2062.

80. Tachibana, K.; Scheuer, P. J.; Tsukitani, Y.; Kikuchi, H.; Van Engen, D.; Clardy, J.; Gopichand, Y.; Schmitz, F. J., Okadaic acid, a cytotoxic polyether from two marine sponges of the genus Halichondria. *Journal of the American Chemical Society* **1981,** 103, (9), 2469-2471.

81. FATTORUSSO, E., LI. 1980. Amins acids from marine algae. *P. J. Sehuer led.] Marine natural products. VQ* 3, 105-107.

82. Vacelet, J.; Donadey, C., Electron microscope study of the association between some sponges and bacteria. *Journal of Experimental Marine Biology and Ecology* **1977,** 30, (3), 301-314.

83. Haygood, M. G.; Schmidt, E. W.; Davidson, S. K.; Faulkner, D. J., Microbial symbionts of marine invertebrates: opportunities for microbial biotechnology. *J Mol Microbiol Biotechnol* **1999,** 1, (1), 33-43.

84. Piel, J., Bacterial symbionts: prospects for the sustainable production of invertebrate-derived pharmaceuticals. *Curr Med Chem* **2006,** 13, (1), 39-50.

85. Simmons, T. L.; Coates, R. C.; Clark, B. R.; Engene, N.; Gonzalez, D.; Esquenazi, E.; Dorrestein, P. C.; Gerwick, W. H., Biosynthetic origin of natural

products isolated from marine microorganism-invertebrate assemblages. *Proc Natl Acad Sci U S A* **2008,** 105, (12), 4587-94.

86. Pettit, G. R.; Kamano, Y.; Herald, C. L.; Tuinman, A. A.; Boettner, F. E.; Kizu, H.; Schmidt, J. M.; Baczynskyj, L.; Tomer, K. B.; Bontems, R. J., Antineoplastic Agents .136. The Isolation and Structure of a Remarkable Marine Animal Antineoplastic Constituent - Dolastatin 10. *Journal of the American Chemical Society* **1987,** 109, (22), 6883-6885.

87. Luesch, H.; Moore, R. E.; Paul, V. J.; Mooberry, S. L.; Corbett, T. H., Isolation of dolastatin 10 from the marine cyanobacterium Symploca species VP642 and total stereochemistry and biological evaluation of its analogue symplostatin 1. *Journal of Natural Products* **2001,** 64, (7), 907-910.

88. Senter, P. D.; Sievers, E. L., The discovery and development of brentuximab vedotin for use in relapsed Hodgkin lymphoma and systemic anaplastic large cell lymphoma. *Nature Biotechnology* **2012,** 30, (7), 631-637.

89. Schofield, M. M.; Jain, S.; Porat, D.; Dick, G. J.; Sherman, D. H., Identification and analysis of the bacterial endosymbiont specialized for production of the chemotherapeutic natural product ET-743. *Environmental Microbiology* **2015,** 17, (10), 3964-3975.

90. Rinehart, K. L., Jr.; Gloer, J. B.; Hughes, R. G., Jr.; Renis, H. E.; McGovren, J. P.; Swynenberg, E. B.; Stringfellow, D. A.; Kuentzel, S. L.; Li, L. H., Didemnins: antiviral and antitumor depsipeptides from a caribbean tunicate. *Science* **1981,** 212, (4497), 933-5.

91. Tsukimoto, M.; Nagaoka, M.; Shishido, Y.; Fujimoto, J.; Nishisaka, F.; Matsumoto, S.; Harunari, E.; Imada, C.; Matsuzaki, T., Bacterial production of the tunicate-derived antitumor cyclic depsipeptide didemnin B. *J Nat Prod* **2011,** 74, (11), 2329-31.

92. Xu, Y.; Kersten, R. D.; Nam, S. J.; Lu, L.; Al-Suwailem, A. M.; Zheng, H.; Fenical, W.; Dorrestein, P. C.; Moore, B. S.; Qian, P. Y., Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *J Am Chem Soc* **2012,** 134, (20), 8625-32.

93. Pavan, M.; Bo, G., Ricerche sulla differenziabilita, natura e attivita del principio tossico di Paederus fuscipes Curt.(Col. Staph.). *Mem. Soc. Ent. It* **1952,** 31, 67-82.

94. Perry, N. B.; Blunt, J. W.; Munro, M. H. G.; Thompson, A. M., Antiviral and Antitumor Agents from a New-Zealand Sponge, Mycale Sp .2. Structures and Solution Conformations of Mycalamide-a and Mycalamide-B. *Journal of Organic Chemistry* **1990,** 55, (1), 223-227.

95. Vuong, D.; Capon, R. J.; Lacey, E.; Gill, J. H.; Heiland, K.; Friedel, T., Onnamide F: A new nematocide from a southern Australian marine sponge, Trachycladus laevispirulifer. *Journal of Natural Products* **2001,** 64, (5), 640-642.

96. Paul, G. K.; Gunasekera, S. P.; Longley, R. E.; Pomponi, S. A., Theopederins K and L. Highly potent cytotoxic metabolites from a marine sponge Discodermia species. *Journal of Natural Products* **2002,** 65, (1), 59-61.

97. Wilson, M. C.; Mori, T.; Ruckert, C.; Uria, A. R.; Helf, M. J.; Takada, K.; Gernert, C.; Steffens, U. A. E.; Heycke, N.; Schmitt, S.; Rinke, C.; Helfrich, E. J. N.; Brachmann, A. O.; Gurgui, C.; Wakimoto, T.; Kracht, M.; Crusemann, M.; Hentschel, U.; Abe, I.; Matsunaga, S.; Kalinowski, J.; Takeyama, H.; Piel, J., An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **2014,** 506, (7486), 58-+.

98. Leao, P. N.; Pereira, A. R.; Liu, W. T.; Ng, J.; Pevzner, P. A.; Dorrestein, P. C.; Konig, G. M.; Vasconcelosa, V. M.; Gerwick, W. H., Synergistic allelochemicals from a freshwater cyanobacterium. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** 107, (25), 11183-11188.

99. Paul, D. G. N.; Valerie, J., Production of secondary metabolites by filamentous tropical marine cyanobacteria: ecological functions of the compounds. *Journal of Phycology* **1999,** 35, (6), 1412-1421.

100. Usher, K. M.; Bergman, B.; Raven, J. A., Exploring cyanobacterial mutualisms. *Annual Review of Ecology Evolution and Systematics* **2007,** 38, 255-273.

101. Smith, C. D.; Zhang, X. Q.; Mooberry, S. L.; Patterson, G. M. L.; Moore, R. E., Cryptophycin - a New Antimicrotubule Agent Active against Drug-Resistant Cells. *Cancer Research* **1994,** 54, (14), 3779-3784.

102. Gerwick, W. H.; Proteau, P. J.; Nagle, D. G.; Hamel, E.; Blokhin, A.; Slate, D. L., Structure of Curacin-a, a Novel Antimitotic, Antiproliferative, and Brine Shrimp Toxic Natural Product from the Marine Cyanobacterium Lyngbya-Majuscula. *Journal of Organic Chemistry* **1994,** 59, (6), 1243-1245.

103. Marquez, B.; Verdier-Pinard, P.; Hamel, E.; Gerwick, W. H., Curacin D, an antimitotic agent from the marine cyanobacterium Lyngbya majuscula. *Phytochemistry* **1998,** 49, (8), 2387-2389.

104. Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J.; Corbett, T. H., Total structure determination of apratoxin A, a potent novel cytotoxin from the marine cyanobacterium Lyngbya majuscula. *Journal of the American Chemical Society* **2001,** 123, (23), 5418-5423.

105. Gutierrez, M.; Suyama, T. L.; Engene, N.; Wingerd, J. S.; Matainaho, T.; Gerwick, W. H., Apratoxin D, a potent cytotoxic cyclodepsipeptide from Papua New Guinea collections of the marine cyanobacteria Lyngbya majuscula and Lyngbya sordida. *Journal of Natural Products* **2008,** 71, (6), 1099-1103.

106. Nogle, L. M.; Gerwick, W. H., Somocystinamide A, a novel cytotoxic disulfide dimer from a Fijian marine cyanobacterial mixed assemblage. *Organic Letters* **2002,** 4, (7), 1095-1098.

107. Wrasidlo, W.; Mielgo, A.; Torres, V. A.; Barbero, S.; Stoletov, K.; Suyama, T. L.; Klemke, R. L.; Gerwick, W. H.; Carson, D. A.; Stupack, D. G., The marine lipopeptide somocystinamide A triggers apoptosis via caspase 8. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (7), 2313-2318.

108. Hayashi, T.; Hayashi, K.; Maeda, M.; Kojima, I., Calcium spirulan, an inhibitor of enveloped virus replication, from a blue-green alga Spirulina platensis. *Journal of Natural Products* **1996,** 59, (1), 83-87.

109. Boyd, M. R.; Gustafson, K. R.; McMahon, J. B.; Shoemaker, R. H.; OKeefe, B. R.; Mori, T.; Gulakowski, R. J.; Wu, L.; Rivera, M. I.; Laurencot, C. M.; Currens, M. J.; Cardellina, J. H.; Buckheit, R. W.; Nara, P. L.; Pannell, L. K.; Sowder, R. C.; Henderson, L. E., Discovery of cyanovirin-N, a novel human immunodeficiency virus-inactivating protein that binds viral surface envelope glycoprotein gp120: Potential applications to microbicide development. *Antimicrobial Agents and Chemotherapy* **1997,** 41, (7), 1521-1530.

110. Bokesch, H. R.; O'Keefe, B. R.; McKee, T. C.; Pannell, L. K.; Patterson, G. M. L.; Gardella, R. S.; Sowder, R. C.; Turpin, J.; Watson, K.; Buckheit, R. W.; Boyd, M. R., A potent novel anti-HIV protein from the cultured cyanobacterium Scytonema varium. *Biochemistry* **2003,** 42, (9), 2578-2584.

111. Becher, P. G.; Beuchat, J.; Gademann, K.; Juttner, F., Nostocarboline: Isolation and synthesis of a new cholinesterase inhibitor from Nostoc 78-12A. *Journal of Natural Products* **2005,** 68, (12), 1793-1795.

112. Shao, C. L.; Linington, R. G.; Balunas, M. J.; Centeno, A.; Boudreau, P.; Zhang, C.; Engene, N.; Spadafora, C.; Mutka, T. S.; Kyle, D. E.; Gerwick, L.; Wang, C. Y.; Gerwick, W. H., Bastimolide A, a Potent Antimalarial Polyhydroxy Macrolide from the Marine Cyanobacterium Okeania hirsuta. *Journal of Organic Chemistry* **2015,** 80, (16), 7849-7855.

113. Portmann, C.; Blom, J. F.; Kaiser, M.; Brun, R.; Juttner, F.; Gademann, K., Isolation of Aerucyclamides C and D and Structure Revision of Microcyclamide 7806A: Heterocyclic Ribosomal Peptides from Microcystis aeruginosa PCC 7806 and Their Antiparasite Evaluation. *Journal of Natural Products* **2008,** 71, (11), 1891-1896.

114. Portmann, C.; Blom, J. F.; Gademann, K.; Juttner, F., Aerucyclamides A and B: Isolation and synthesis of toxic ribosomal heterocyclic peptides from the cyanobacterium Microcystis aeruginosa PCC 7806. *Journal of Natural Products* **2008,** 71, (7), 1193-1196.

115. Harrigan, G. G.; Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Nagle, D. G.; Biggs, J.; Park, P. U.; Paul, V. J., Tumonoic acids, novel metabolites from a cyanobacterial assemblage of Lyngbya majuscula and Schizothrix calcicola. *Journal of Natural Products* **1999,** 62, (3), 464-467.

116. Neumann, U.; Forchert, A.; Flury, T.; Weckesser, J., Microginin FR1, a linear peptide from a water bloom of Microcystis species. *Fems Microbiology Letters* **1997,** 153, (2), 475-478.

117. Ersmark, K.; Del Valle, J. R.; Hanessian, S., Chemistry and biology of the aeruginosin family of serine protease inhibitors. *Angew Chem Int Ed Engl* **2008,** 47, (7), 1202-23.

118. Miller, B.; Friedman, A. J.; Choi, H.; Hogan, J.; McCammon, J. A.; Hook, V.; Gerwick, W. H., The marine cyanobacterial metabolite gallinamide A is a potent and selective inhibitor of human cathepsin L. *J Nat Prod* **2014,** 77, (1), 92-9.

119. Linington, R. G.; Clark, B. R.; Trimble, E. E.; Almanza, A.; Urena, L. D.; Kyle, D. E.; Gerwick, W. H., Antimalarial peptides from marine cyanobacteria: isolation and structural elucidation of gallinamide A. *J Nat Prod* **2009,** 72, (1), 14-7.

120. Koehn, F. E.; Longley, R. E.; Reed, J. K., Microcolins A and B, new immunosuppressive peptides from the blue-green alga Lyngbya majuscula. *J Nat Prod* **1992,** 55, (5), 613-9.

121. Berman, F. W.; Gerwick, W. H.; Murray, T. F., Antillatoxin and kalkitoxin, ichthyotoxins from the tropical cyanobacterium Lyngbya majuscula, induce distinct temporal patterns of NMDA receptor-mediated neurotoxicity. *Toxicon* **1999,** 37, (11), 1645-1648.

122. Wu, M.; Okino, T.; Nogle, L. M.; Marquez, B. L.; Williamson, R. T.; Sitachitta, N.; Berman, F. W.; Murray, T. F.; McGough, K.; Jacobs, R.; Colsen, K.; Asano, T.; Yokokawa, F.; Shioiri, T.; Gerwick, W. H., Structure, synthesis, and biological properties of kalkitoxin, a novel neurotoxin from the marine cyanobacterium Lyngbya majuscula. *Journal of the American Chemical Society* **2000,** 122, (48), 12041-12042.

123. Kersten, R. D.; Ziemert, N.; Gonzalez, D. J.; Duggan, B. M.; Nizet, V.; Dorrestein, P. C.; Moore, B. S., Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences of the United States of America* **2013,** 110, (47), E4407-E4416.

124. Winnikoff, J. R.; Glukhov, E.; Watrous, J.; Dorrestein, P. C.; Gerwick, W. H., Quantitative molecular networking to profile marine cyanobacterial metabolomes. *Journal of Antibiotics* **2014,** 67, (1), 105-112.

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

Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J. Q.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrovr, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W. Y.; Liu, X. T.; Zhang, L. X.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** 34, (8), 828-837.

126. Dalisay, D. S.; Molinski, T. F., Structure Elucidation at the Nanomole Scale. 2. Hemi-phorboxazole A from Phorbas sp. *Organic Letters* **2009,** 11, (9), 1967-1970.

127. Ndukwe, I. E.; Shchukina, A.; Kazimierczuk, K.; Butts, C. P., Rapid and safe ASAP acquisition with EXACT NMR. *Chemical Communications* **2016,** 52, (86), 12769-12772.

128. Frydman, L., Single-scan multidimensional NMR. *Comptes Rendus Chimie* **2006,** 9, (3), 336-345.

129. Kupce, E.; Freeman, R., Two-dimensional Hadamard spectroscopy. *Journal of Magnetic Resonance* **2003,** 162, (2), 300-310.

130. Ruan, K.; Yang, S. T.; Van Sant, K. A.; Likos, J. J., Application of Hadamard spectroscopy to automated structure verification in high-throughput NMR. *Magnetic Resonance in Chemistry* **2009,** 47, (8), 693-700.

131. Mobli, M.; Maciejewski, M. W.; Schuyler, A. D.; Stern, A. S.; Hoch, J. C., Sparse sampling methods in multidimensional NMR. *Physical Chemistry Chemical Physics* **2012,** 14, (31), 10835-10843.

132. Kazimierczuk, K.; Orekhov, V. Y., Accelerated NMR Spectroscopy by Using Compressed Sensing. *Angewandte Chemie-International Edition* **2011,** 50, (24), 5556-5559.

133. Hoch, J. C.; Stern, A. S., *NMR data processing*. Wiley-Liss: New York, 1996; p xi, 196 pages.

134. Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S., Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Analytical Chemistry* **2011,** 83, (5), 1649-1657.

135. Matile, S.; Berova, N.; Nakanishi, K.; Fleischhauer, J.; Woody, R. W., Structural studies by exciton coupled circular dichroism over a large distance: Porphyrin derivatives of steroids, dimeric steroids, and brevetoxin B. *Journal of the American Chemical Society* **1996,** 118, (22), 5198-5206.

136. Yoo, H. D.; Gerwick, W. H., Curacins B and C, new antimitotic natural products from the marine cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **1995,** 58, (12), 1961-1965.

137. Ueoka, R.; Hitora, Y.; Ito, A.; Yoshida, M.; Okada, S.; Takada, K.; Matsunaga, S., Curacin E from the Brittle Star Ophiocoma scolopendrina. *Journal of Natural Products* **2016,** 79, (10), 2754-2757.

138. Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J., New apratoxins of marine cyanobacterial origin from Guam and Palau. *Bioorganic & Medicinal Chemistry* **2002,** 10, (6), 1973-1978.

139. Tidgewell, K.; Engene, N.; Byrum, T.; Media, J.; Doi, T.; Valeriote, F. A.; Gerwick, W. H., Evolved Diversification of a Modular Natural Product Pathway: Apratoxins F and G, Two Cytotoxic Cyclic Depsipeptides from a Palmyra Collection of Lyngbya bouillonii. *Chembiochem* **2010,** 11, (10), 1458-1466.

140. Thornburg, C. C.; Cowley, E. S.; Sikorska, J.; Shaala, L. A.; Ishmael, J. E.; Youssef, D. T. A.; McPhail, K. L., Apratoxin H and Apratoxin A Sulfoxide from the Red Sea Cyanobacterium Moorea producens. *Journal of Natural Products* **2013,** 76, (9), 1781-1788.

CHAPTER 2

ISOLATION AND STRUCTURAL ELUCIDATION OF LAUCYSTEINAMIDE A, A HYBRID PKS/NRPS METABOLITE FROM A SAIPAN CYANOBACTERIUM, *CF. CALDORA PENICILLATA*

2.0. Abstract

A bioactivity guided study of a cf. *Caldora penicillata* species, collected during a 2013 expedition to the Pacific island of Saipan, Northern Mariana Islands (a commonwealth of the USA), led to the isolation of a new thiazoline-containing alkaloid, laucysteinamide A (**1**)[1]. Laucysteinamide A is a new monomeric analogue of the marine cyanobacterial metabolite, somocystinamide A (**2**), a disulfide-bonded dimeric compound that was isolated previously from a Fijian marine cyanobacterium. The structure and absolute configuration of laucysteinamide A (**1**) was determined by a detailed analysis of its NMR, MS, and CD spectra. In addition, the highly bioactive lipid, curacin D (**3**), was also found to be present in this cyanobacterial extract. The latter compound was responsible for the potent cytotoxicity of this extract to H-460 human non-small cell lung cancer cells *in vitro*.

2.1 Introduction

Field collections of Pacific tropical marine cyanobacteria have been a prolific source for a wide range of novel bioactive marine natural products[2]. Whereas the Northern Mariana Islands have not previously been identified as “hotspots” of marine

biodiversity, their coral reef habitats have been a source of chemically-prolific strains of marine cyanobacteria. For example, obyanamide, a high nM LC_{50} human KB cancer cell cytotoxic agent, was isolated from a collection of the marine cyanobacterium *Lyngbya confervoides* from Saipan, a US territory[3]. The relatively unexplored marine biodiversity of this region is thus an exciting resource for screening for new biologically active natural products.

Such screening efforts are especially relevant to the discovery of anticancer lead compounds, as many of our current the Food and Drug Administration (FDA) approved agents in this therapeutic class are derived from or have been patterned after natural products. Examples include the vinca alkaloids, taxanes, dolastatins (a lead for auristatin E), and halichondrin B (lead for eribulin)[4], as well as a number of preclinical leads such as curacin A, discodermolide and the ixabepilone[5]. Common to all of these agents is their targeting and disruption of the pivotal function of microtubules within cancer cells, leading to apoptotic cell death.

In 2013, we surveyed the natural populations of marine cyanobacteria from Saipan and made collections of those present in sufficient biomass to support ensuing chemical and pharmacological investigations. One such shallow water (1–2 m) collection, obtained from Lau Lau Bay, was comprised of several centimeter long tufts of a pinkish-purple colored filamentous cyanobacterium. These were found growing distinctively from the tops of stipes of the brown alga *Turbinaria* sp. Subsequently, the collected biomass was extracted in the laboratory and subjected to a combination of bioassay and NMR guided isolation efforts. We describe here the isolation, structure

elucidation, and biological activity of laucysteinamide A (**1**) (Figure 2.1), a novel hybrid PKS/NRPS (2-methyl-4-thiazolinyl) cytotoxic compound from this Saipan cyanobacterial collection. A second and highly bioactive metabolite of this collection was the previously described compound curacin D (**3**), a potent inhibitor of microtubule assembly[6]. Laucysteinamide A is structurally related to somocystinamide A (**2**), a neurotoxic and cytotoxic compound (IC_{50} = 3 nM) previously isolated from a mixed assemblage of Fijian marine cyanobacteria[7]. In the course of these studies, the source organism was examined in detail by light microscopy and was found to correspond to the recently described species *Caldora penicillata* (Figure 2.2)[8].

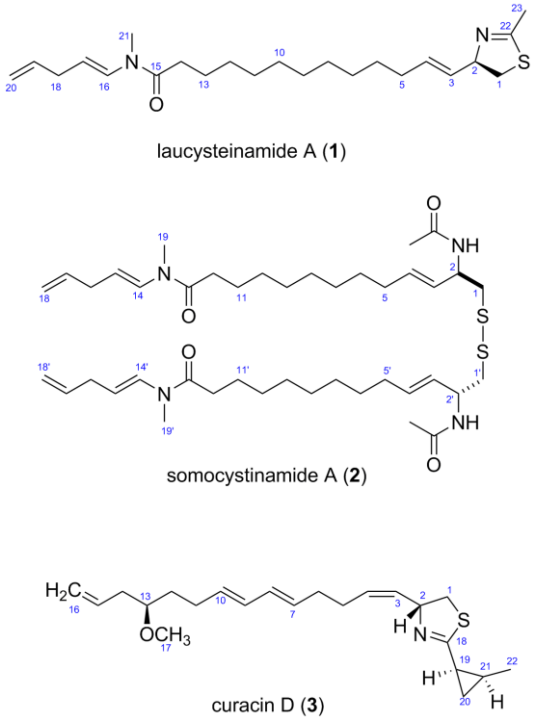


Figure 2.1. Structures of Compounds (**1**–**3**).

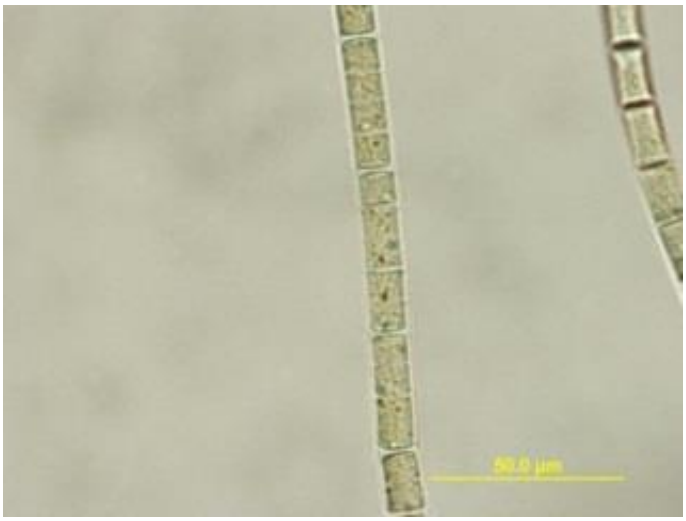


Figure 2.2. Microscopic image of filaments of the cf. *Caldora penicillata* species (100×).

2.2. Results and Discussion

SPL-3Feb13-2, a cf. *Caldora penicillata* collected at Lau Lau Bay in Saipan, was repeatedly extracted with a 2:1 mixture of dichloromethane (DCM) and methanol (MeOH) to afford a total of 6.73 g of extract. Two lipophilic fractions obtained from vacuum liquid chromatography of this extract (fraction C: 20% ethyl acetate (ETAC)/hexanes *v/v*; fraction D: 40% ETAC/hexanes *v/v*) possessed interesting ^{1}H-NMR features, such as *J*-coupled peaks in the alkenyl and typical peptidyl alpha proton regions. Moreover, they demonstrated strong toxicity in the brine shrimp model (100% toxicity at 3 μg/mL), and thus were selected for further investigation. These two combined fractions (36.5 mg) were repeatedly chromatographed by normal phase HPLC to afford two compounds; 3.6 mg (0.05% extraction yield *w/w* dry) of compound **1** as an optically active oil $[\alpha]_D^{26}$ = +17.1° (*c* = 0.86, $CHCl_3$), and 2.7 mg (0.04% extraction

yield *w/w* dry) of compound **3**. Dereplication using MS-MS based molecular networking[9] (Figure 2.3) and ^{1}H-NMR[6] indicated that compound **3** was the known cyanobacterial metabolite, curacin D. However, compound **1**, assigned here the trivial name laucysteinamide A, had MS and NMR features unlike any known compound, and hence its structural and biological properties were investigated as reported below.

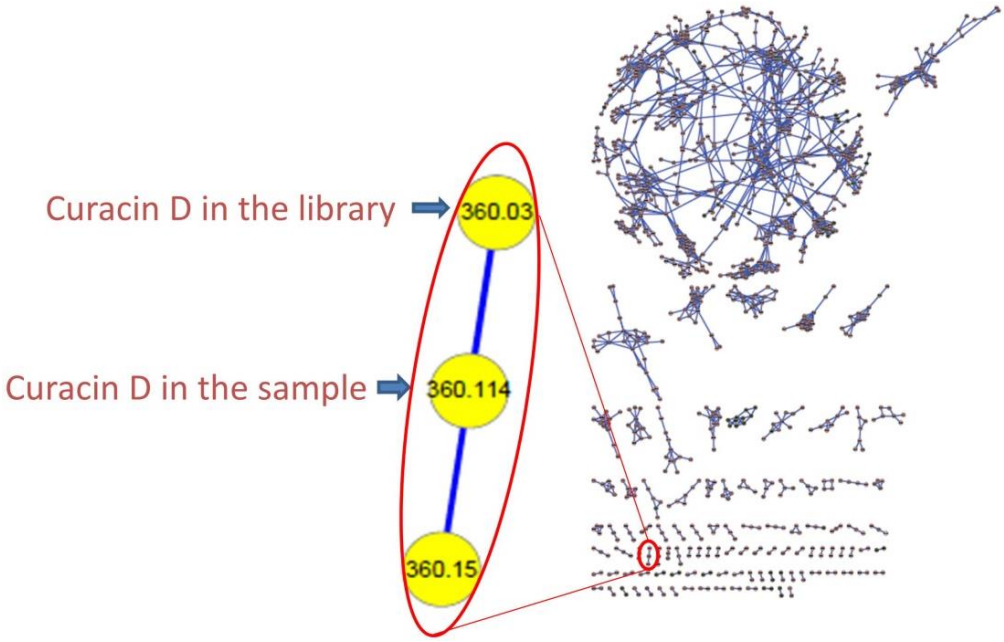


Figure 2.3. Representative Molecular Network. (**Left**) the expanded cluster of curacin D includes one peak in the LC-MS/MS of the present extract along with the node incorporated from a pure compound library; (**Right**) the entire molecular network of the crude extract, fractions C–J, and the pure compound library. The circular cluster represents structurally similar molecular families from either crude extract or vacuum liquid chromatography (VLC) fractions that share similar MS/MS fragmentation patterns[9].

The MS of **1** provided a nominal mass of *m/z* 391 $[M + H]^{+}$, initially suggesting a molecular formula typical for a diisooctyl phthalate contaminant ($C_{24}H_{38}O_4$). However, an isotope ratio fitting calculation eliminated this possibility, and the

molecular formula $C_{23}H_{38}N_2OS$ was derived (*m/z* 391.2778, calcd; six degrees of unsaturation). A preliminary analysis of ^{1}H- and ^{13}C-NMR spectra of **1** in C_6D_6 showed resonances for two deshielded methyl groups, an amide *N*-methyl group in two equilibrating tautomeric configurations [δ_H 2.90 (s) and 2.35 ppm (s) (3H together, CH_3-21)] and a methyl group attached to an sp^2 carbon [δ_H 1.98 (3H, s, CH_3-23)]. Additionally, eight sp^2 carbons were observed for **1**, forming one mono-substituted vinylidene moiety, two di-substituted *trans* alkenes, and two carbonyl or imine functional groups. At this point, the remaining degree of unsaturation could be explained by either the presence of a ring structure or a carbon-nitrogen triple bond.

The substructures of an amide (**1a**) and a 2-methyl-4-thiazolinyl ethenyl moiety (**1b**) were constructed according to interpretation of 1D and 2D NMR data obtained from ^{1}H-^{1}H Correlation spectroscopy (COSY) and ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC), Heteronuclear 2 Bond Correlation (H2BC), and Heteronuclear Multiple Bond Correlation (HMBC) experiments (Figure 2.4). The HSQC data allowed for the assignment of the combined three-proton signals at δ_H 2.35 and 2.90 to the *cis/trans* isomers of the *N*-methyl amide carbon (δ_C 31.4 and 29.6, C-21), respectively. The HMBC correlations from δ_H 2.35 to the sp^2 methine and carbonyl carbons, δ_C 129.0 (δ_H 7.80, d, J = 14.4 Hz, C-16) and 170.5 (C-15), respectively, were not observed in the ^{1}H-^{13}C H2BC spectrum, thus suggesting the occurrence of an *N*-methyl enamide moiety. For the configurational isomer with CH_3-21 at δ_H 2.90, HMBC correlations were observed with analogous carbons at δ_C 130.1 (δ_H 6.42, d, J = 13.8 Hz) and 170.2. By COSY and *J*-coupling analysis, the H-16 methine resonances were

located adjacent to another set of divided proton signals [δ_H 4.61 (dt, *J* = 13.8, 6.6 Hz) and 4.66 (dt, *J* = 14.4, 7.2 Hz) (1H together, H-17)], with corresponding carbon resonances at δ_C 107.2 and 106.9. The magnitude of the *J*-coupling between H-16 and H-17 suggested a *trans*-double bond configuration. A deshielded methylene group [δ_H 2.60 (t, *J* = 6.1 Hz) and 2.68 (t, *J* = 6.1 Hz) (2H together, H-18), δ_C 34.7] was adjacent to C-17 according to H2BC correlations. ^{1}H-^{1}H COSY data sequentially connected the protons of CH_2-18, alkenyl CH-19 [(δ_H 5.78 (m) and 5.80 ppm (m) (1H together), δ_C 137.5 and 138.0] and the alkenyl terminus CH_2-20 [δ_H 5.03 (m, 2H), 114.9 and 115.2]. Completing partial structure **1a** was a moderately deshielded methylene group at δ_H 2.09 (2H, t, *J* = 7.2 Hz), δ_C 33.6), located next to the carbonyl by HMBC and H2BC correlations. The two sets of ^{1}H-NMR and ^{13}C-NMR chemical shift data for each carbon and proton from C-14 to C-21 is explained by conformational anisotropy caused by *cis/trans* isomerism of the corresponding *N*-methyl quaternary imine (Figure 2.5), as described previously[7, 10]. The *cis*/*trans* ratio between these isomers (0.44:1) was calculated in C_6D_6 at 20 °C using peak integrals from the ^{1}H-NMR spectrum.

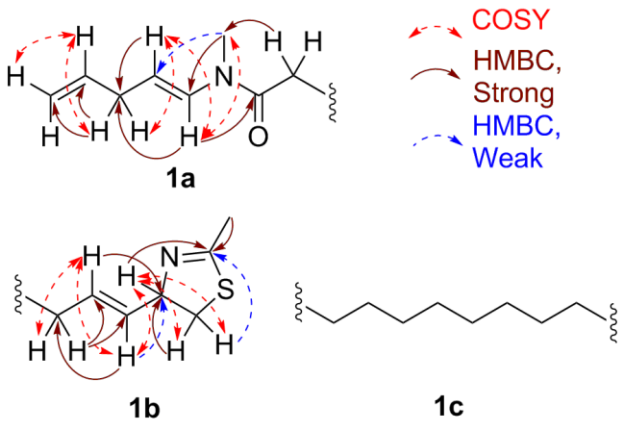


Figure 2.4. Selected COSY and HMBC correlations for **1a** and **1b**, two partial substructures of laucysteinamide A (**1**), plus the intervening substructure **1c**.

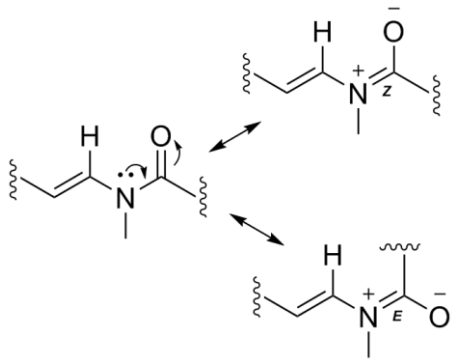


Figure 2.5. The imine-enamine tautomerism results in two sets of chemical shifts for atoms in this region of laucysteinamide A (**1**).

Table 2.1. NMR Spectroscopic Data for Laucysteinamide A (**1**) in Benzene-d_6.

Position	δ_H (*J* in Hz)	δ_C (*J* in Hz)	COSY	H2BC	HMBC
1	3.01, dd (10.8, 8.4)	40.3	2.79, 4.80	79.2	32.7, 79.2, 164.7
	2.79, dd (10.8, 8.4)		3.01, 4.80	79.2	164.7

Table 2.1. NMR Spectroscopic Data for Laucysteinamide A (**1**) in Benzene-d_6. (Cont'd)

Position	δ_H (*J* in Hz)	δ_C (*J* in Hz)	COSY	H2BC	HMBC
2	4.8, ddd (6.6)	79.2	2.79, 3.01, 1.97, 5.53	40.3	164.7, 130.1, 132.4
3	5.53, dd (15.6, 6.6)	130.1	1.97, 4.80	79.2, 132.4	32.7, 132.4
4	5.69, dt (15.6, 7.8,)	132.4	1.97	32.7, 130.1	29.9, 32.7, 130.1
5	1.97, m	32.7	4.80, 5.53/5.69	132.4	40.3
6	1.25, m	29.9			30.0
7 to 12	1.25–1.4, m	30.0			
13	1.69, m	25.3	1.28, 1.93, 2.09	33.6	30.0
14	2.09, t (7.2)	33.6	1.69	25.3	30.0
	1.93, t (7.8)	34.7	1.69	25.3	170.2
15		170.5			
		170.2			
16	6.42, d (13.8)	130.1	4.61	107.2	34.7, 29.6, 170.5
	7.80, d (14.4)	129.0	4.66	106.9	31.4, 34.7
17	4.61, dt (13.8, 6.6)	107.2	2.60, 6.42	129.0	34.7/34.8, 34.7, 129.0, 137.5
	4.66, dt (14.4, 7.2)	106.9	2.68, 7.80	137.5	34.7/34.8
18	2.60, t (6.1)	34.7	4.61, 5.78	107.2/106.9	114.9/115.2, 107.2, 130.1, 137.5
	2.68, t (6.1)	34.8	4.66, 5.80		114.9/115.2, 106.9, 138.0
19	5.78, m	137.5	2.60, 5.03	114.9	130.1
	5.80, m	138.0	2.68, 5.03	115.2	129.0
20	5.03, m	114.9	2.60/2.68	137.5/138.0	34.7/34.8

Table 2.1. NMR Spectroscopic Data for Laucysteinamide A (**1**) in Benzene-d_6. (Cont'd)

Position	δ_H (*J* in Hz)	δ_C (*J* in Hz)	COSY	H2BC	HMBC
	5.03, m	115.2	2.60/2.68	37.5/138.0	34.7/34.8
21	2.90, s	29.6	6.48		130.1, 107.2/106.9, 170.5
	2.35, s	31.4			129.0, 170.5
22		164.7			79.2
23	1.98, s	20.3			164.7

The remaining unassigned methyl signal from the ^{1}H-NMR spectrum [δ_H 1.99 ppm (3H, s, CH_3-23)] showed a strong HMBC correlation with an imine carbon [δ_C 164.7 ppm (C-22)]. A heteroatom-substituted methylene group [δ_H 3.01 (1H, dd, *J* = 10.8, 8.4 Hz, H-1a) and 2.79 (1H, dd, *J* = 10.8, 8.4 Hz, H-1b)] and a more deshielded methine moiety [δ_H 4.80 (1H, ddd, *J* = 6.6 Hz, CH-2), δ_C 79.2] also showed HMBC corrections with this imine carbon. By COSY and H2BC, the H_2-1 protons were shown to be adjacent to the C-2 methine, and this could be extended to an alkenyl methine group [δ_H 5.53 (1H, dd, *J* = 15.6, 6.6 Hz, H-3)]. Proton H-3 showed COSY and H2BC correlations with another vinyl proton [δ_H 5.69 (1H, dd, *J* = 15.6, 7.8 Hz, H-4); δ_C 132.4 (C-4)]; the large *J*-value between H-3 and H-4 indicated a *trans*-relationship. This latter CH group was adjacent to an allylic methylene group (C-5) as revealed by multiple COSY and HMBC correlations (Figure 2.4, Table 2.1). The positioning of a sulfur between C-23 and C-1 and nitrogen atom between C-23 and C-2 was supported by comparison with the ^{1}H- and ^{13}C-NMR data in benzene-d_6 for the 2-alkyl-4-ethenyl-

thiazoline moiety present in both curacin D[6] and curacin A[11]. Altogether, these data suggested the presence of a methylene-substituted 2-methyl-4-ethenyl-thiazoline subunit in laucysteinamide A (**1**) (Figure 2.1).

The two partial structures, **1a** and **1b**, were connected by a saturated linear alkyl chain (**1c**, C-6 to C-13) to form the final planar structure of compound **1**. The insertion of eight methylene groups between C-5 and C-14 satisfied the molecular formula, and was supported by COSY and HMBC correlations between the terminal atoms of partial structures **1a** and **1b** and the shielded methylene envelope of resonances of partial structure **1c** (Figure 2.4). Consequently, the planar structure of laucysteinamide (**1**) was shown to be an alternately condensed form of a monomer of the symmetrical dimeric metabolite somocystinamide A (**3**)[7].

2.2.1. Stereochemistry

The absolute configuration of laucysteinamide A was determined by comparison of energy-minimized molecular models of **1** with observed exciton coupling circular dichroism (ECCD) data. Computational molecular models of enantiomeric forms of **1** were subjected to energy minimization with MOPAC software (ChemBioDraw Ultra 13.0, PerkinElmer Inc., Waltham, MA, USA)[12], the results of which are shown in Figure 2.6. The through-space coupling of nearby interacting chromophores gives rise to diagnostic angle-dependent exciton coupling in the circular dichroism spectrum[13]. In the case of laucysteinamide A, a thiazoline chromophore is present in the vicinity of the C-3/C-4 alkenyl moiety. The coupling of these chromophores gives rise to a corresponding split Cotton effect, as shown in Figure 2.6. The ECCD spectrum of

compound **1** (Figure 2.6.11) showed a negative local maximum at 223 nm, corresponding to the thiazoline chromophore. The maximum expected from the C-3/C-4 alkene would be around 190 nm, but was not observed in the spectrum due to solvent absorptions. However, the CD spectrum clearly showed a negative first Cotton effect, and thus, the absolute configuration of compound **1** is confidently assigned as 2*R*. Compound **2**, which is an analogue of compound **1**, was assigned previously with a 2*R*,2′*R* absolute configuration, and it showed a similar optical rotation $[\alpha]_D^{22}$ = +13.5° (*c* = 0.75, $CHCl_3$) to compound **1**, thus providing additional support for t the absolute configuration of compound **1** as 2*R*[7].

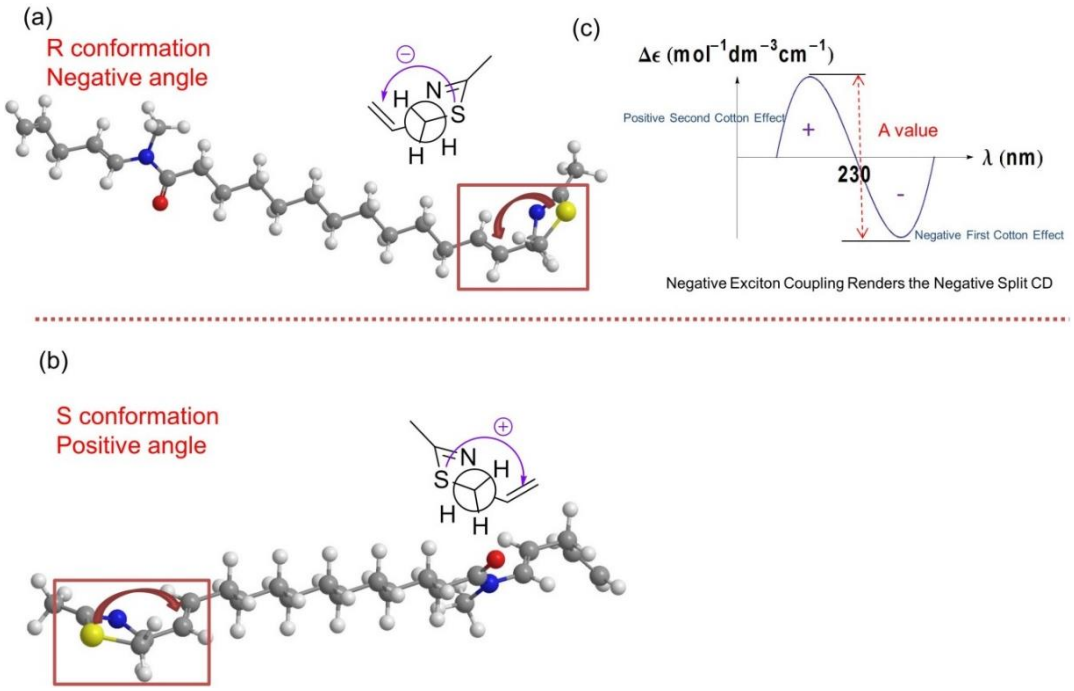


Figure 2.6. Molecular Modeling and exciton coupling circular dichroism (ECCD) Prediction. (**a**) The stick-ball model of the *R* configuration of compound **1** and its Newman projection of the relevant thiazoline moiety. The angle between the two chromophores in the model is negative (counterclockwise). Atom definitions: red, oxygen; yellow, sulfur; blue, nitrogen; gray, carbon; white, hydrogen; (**b**) The stick-ball model of the *S* configuration of compound **1** and its Newman projection of the relevant thiazoline moiety. The angle between the two chromophores in the model is positive (clockwise). Atom definitions as in panel a; (**c**) The *R* conformation of compound **1** with negative exciton coupling renders the negative split ECCD curve as indicated by the experimental result shown in Figure 2.6.11.

2.2.2. Bioactivity

Although laucysteinamide A (**1**) is structurally analogous to somocystinamide A (**2**), which was previously shown to be a potent inhibitor of angiogenesis and cancer cell proliferation (IC_{50} = 3 nM against Jurkat leukemia cells[14]), **1** was only mildly cytotoxic to H-460 human non-small cell lung cancer cells (IC_{50} = 11 μM), compared with doxorubicin (EC_{50} = 0.032 μM). Thus, the cytotoxic activity of this fraction

(fraction D) was almost entirely due to the antitubulin properties of curacin D (**3**)[6]. The brine shrimp toxicity data of all VLC fractions and crude extract are shown in Table 2.6.1; unfortunately, compound **1** decomposed before it could be evaluated for this biological activity.

2.2.3. Biosynthetic Considerations

Based on its essentially linear structure, we propose that laucysteinamide A is assembled by a hybrid PKS/NRPS pathway as it possesses sections logically deriving from amino acid residues (cysteine and *N*-methylglycine) and nine acetate units (Figure 2.6.12). The assembly of **1** could be initiated by loading an acetate unit followed by an NRPS-mediated condensation with cysteine, heterocyclization and dehydration to form the thiazoline ring, followed by six cycles of polyketide extension. While the first ketide extension is only partially reduced, the ensuing five are fully reduced. Next, another NRPS module is envisaged to incorporate a glycine residue, followed by an *S*-adenosyl methionine assisted methylation of the glycine *N*-atom (Figure 2.6.12). After two more cycles of acetate extension, the chain termination (Figure 2.6.12) likely concludes with a sulfotransferase-aided sulfonate esterification of the β-hydroxy group, thioesterase hydrolysis and then coincident decarboxylation and sulfate elimination. This termination sequence is identical to that found in the previously described curacin A pathway[15].

2.3 Conclusion

Marine cyanobacteria, such as *Caldora penicillata*, are a continuing source of new and bioactive molecules of interest for natural product drug discovery research. A novel thiazoline containing alkaloid, laucysteinamide (**1**), along with known compounds, curacin D (**3**), were isolated from the marine cyanobacterium, cf. *Caldora penicillata*, collected from island of Saipan. Laucysteinamide A (**1**) was determined to be mildly cytotoxic to H-460 human non-small lung cancer cells, and was isolated together with the known potent bioactive molecule, curacin D (**3**), which was initially identified by LC-MS/MS molecular networking dereplication efforts. Laucysteinamide A (**1**) belongs to a unique cyanobacterial metabolite class of mixed PKS/NRPS biosynthetic origin, analogous to the previously reported somocystinamide A (**2**). We propose that the acetated l-cysteine is extended with six more malonyl CoA-derived acetate units. This intermediate is then further extended with an *N*-methyl glycine unit, followed by two additional acetates. The proposed biosynthesis of **1** is terminated by a decarboxylation very similar to that described for curacin A[16].

2.4 Experimental Methods

2.4.1. General Experimental Procedures

Optical rotations were recorded with a JASCO P-2000 polarimeter; ECCD spectra were measured in CH_2Cl_2 using a JASCO J-810 spectropolarimeter (Jasco, Easton, MD, USA). UV spectra were recorded with a Beckman Coulter DU800 spectrophotometer (Indianapolis, IN, USA); IR spectra were collected on a Bruker

ALPHA-P FTIR spectrometer with a diamond ATR (Bruker Optics, Billerica, MA, USA). ^{1}H NMR and 2D NMR spectra of laucysteinamide A (**1**) were measured on the Bruker 600 MHz NMR spectrometer with a 5 mm inverse detection triple resonance (^{1}H-^{13}C/^{15}N/^{2}D) cryoprobe (Bruker Biospin, Billerica, MA, USA). The remaining ^{1}H- and ^{13}C-NMR spectra were measured on a JEOL ECA 500 MHz spectrometer (JEOL, Akishima, Tokyo, Japan) or Varian XSens 2 channel (^{1}H/^{13}C) NMR cryoprobe optimized for direct observation of ^{13}C-NMR, with samples dissolved in $CDCl_3$ or C_6D_6 (Varian, Palo Alto, CA, USA). Low-resolution MS spectra were recorded on a Thermo Finnigan LCQ, operating in positive ion ESI mode, coupled to a Thermo Finnigan Surveyor Plus liquid chromatography system (Thermo Scientific, Waltham, MA, USA). HRMS data were obtained with an Agilent 6230 TOF-MS (Agilent, Santa Clara, CA, USA) under positive ion ESI-TOF-MS conditions and provided by the University of California, San Diego (UCSD) Small Molecule MS Facility. Semipreparative HPLC was carried out using a Waters 515 pump system with a Waters 996 PDA detector (Waters Corporation, Milford, MA, USA).

2.4.2. Sample Material

A marine cyanobacterial sample, assigned the code SPL-3FEB13-2, was collected in February 2013 from shallow water in Lau Lau Bay, Saipan (GPS coordinates: 15°09’35.5’’ N and 145°45’25.5’’ E). The taxonomy of the collected sample was determined to be cf. *Caldora penicillata* by microscopic characterization. The 2.5–15 cm long cyanobacterial filaments possessed a mucilagenous base and grew mainly from the tops of *Turbinaria* sp. The tan/purple colored sample was preserved in

approximately 750 mL of isopropanol in sea water at -20 °C prior to laboratory extraction. A voucher specimen is preserved and available from the Gerwick Voucher Collection, Scripps Institution of Oceanography, University of California San Diego.

2.4.3. Extraction and Isolation

The sample biomass was defrosted and then extracted with DCM/MeOH (2:1), eight times, to yield 6.73 g of dark green crude extract. The extracted cyanobacterial biomass was 71.3 g in dry weight. This crude extract was subjected to a stepped-gradient fractionation (hexanes/EtOAc and EtOAc/MeOH) by vacuum liquid chromatography (VLC) over normal phase silica gel to give ten fractions (A–J). Fraction C (eluted with 20% EtOAc/hexanes) was again separated by normal phase chromatography on silica gel with a stepwise gradient of hexanes/EtOAc to give six sub-fractions (Ca-Cf). Fraction Cb was further purified by chromatography on a normal-phase Luna HPLC column (100 Å, 5 μm, 250 × 1000 mm, isocratic solvent system comprised of 10% EtOAc/hexanes over 45 min; flow rate 3 mL/min; PDA detection) to give 2.7 mg of compound **3** (t_R = 13.2 min)**,** which had a pale yellowish color. The ^{1}H- and ^{13}C-NMR spectra of **3** in C_6D_6 matched literature reported values for curacin D[6]. The optical rotation value of **3**, $[\alpha]_D^{25}$ = +33.3° (*c* = 0.14, $CHCl_3$), matched the reported value $[\alpha]_D^{25}$ = +33° (*c* = 0.14, $CHCl_3$)[6].

The ^{1}H-NMR spectrum of fraction D in $CDCl_3$ showed peaks with coupling patterns in the 4.0 ppm to 7.5 ppm region, and was selected for further investigation. The major peak in this fraction by LC-MS/MS analysis showed an $[M + H]^+$ at *m/z* 391.11; by MarinLit searching and Molecular Network analysis, this compound did not

correlate with any known compound. Fraction D was further separated with HPLC as above on a normal-phase Luna column (100 Å, 5 μm, 250 × 1000 mm, solvent system of a linear gradient starting with 100% 1:3 EtOAc/hexanes for 25 min before being ramped to 100% 1:1 EtOAc/hexanes in 10 min followed by maintenance in 100% 1:1 EtOAc/hexanes for another 5 min; flow rate 3 mL/min; PDA detection at 254 nm), giving rise to 3.6 mg of compound **1** (t_R = 34.5 min).

Laucysteinamide A (**1**): pale yellowish oil. $[\alpha]_D^{26}$ = +17.07° (*c* = 0.86, $CHCl_3$); UV λ_{max} (CH_2Cl_2) 223 nm; IR (neat) λ_{max} 2927, 2844, 1662, 1634, 1464, 1394, 1338, 1161, 1087, 912 cm^{-1}; ^{1}H-, ^{13}C-, and 2D-NMR see Table 2.1; ESIMS *m/z* 391 $[M + H]^+$; HRESITOFMS *m/z* 391.2777 $[M + H]^+$ (calcd for $C_{23}H_{39}N_2OS$, 391.2778).

Curacin D (**3**): pale yellow oil. $[\alpha]_D^{25}$ = +33.3° (*c* = 0.14, $CHCl_3$), Lit. $[\alpha]_D^{25}$ = +33° (*c* = 0.14, $CHCl_3$) [5]; UV λ_{max} (hexanes) 224 (ϵ 9 000) nm ^{1}H-NMR (C_6D_6, 500 MHz) δ 6.1 (1H, m, H-9), 6.0 (1H, m, H-8), 5.79 (1H, ddt, *J* = 16.2, 11.0, 7.2 Hz, H-15), 5.64 (1H, dd, *J* = 10.5, 10.4 Hz, H-3), 5.55 (1H, dt, *J* = 14.5, 7.1 Hz, H-10), 5.49 (1H, bdt, *J* = 14.5, 7.3 Hz, H-7), 5.38 (1H, m, H-4), 5.05 (2H, m, H-16), 5.03 (1H, m, H-2), 3.12 (3H, s, -OMe), 3.05 (1H, m, H-13), 3.03 (1H, dd, *J* = 10.3, 8.3 Hz, H-1b), 2.74 (1H, dd, *J* = 10.3, 10.3 Hz, H-1a), 2.2 (4H, m, H-11,14), 2.1 (2H, m, H-5), 2.0 (2H, m, H-6), 1.67 (1H, td, *J* = 8.3, 5.5 Hz, H-18), 1.58 (1H, m, H-12a), 1.54 (1H, m, H-12b), 1.17 (3H, d, *J* = 6.3 Hz, H-21), 1.15 (1H, m, H-19b), 0.95 (1H, m, H-20), 0.67 (1H, ddd, *J* = 8.1, 8.1, 4.3 Hz, H-19a); ^{13}C-NMR (C_6D_6, 100 MHz) δ 168.44 (C17), 135.09 (C15), 132.29 (C10), 131.39 (C9), 131.12 (C3), 131.05 (C7), 130.87 (C8), 130.57 (C4), 116.54 (C16), 79.56 (C13), 74.1 (C2), 56.05 (OMe), 39.69 (C1), 37.91 (C14), 33.39 (C12),

32.51 (C6), 28.51 (C11), 27.67 (C5), 19.86 (C18), 15.71 (C20), 13.96 (C19), 12.06 (C21); LC LRMS $[M + H]^+$ *m/z* 360.114.

2.4.4. Molecular Networking

The fractions and crude extract were each diluted to 1 mg/mL in MeOH for LC-MS with automated dependent MS/MS scanning. The following gradient was used: initiated in 50% CH_3CN in MilliQ H_2O with 0.1% formic acid (*v/v*) for 4 min, then CH_3CN increased to 99% in a linear gradient over 12 min and then maintained at this percentage for 5 min. Finally, the CH_3CN was reduced to 50% and stabilized for 3 min. MS data from two scan events were acquired: (1) scan positive MS, window from *m*/*z* 190–2000; (2) scan MS/MS in data-dependent mode for the most intense ions from the first scan. The raw Thermo XCalibur data files were processed with MSConvert to produce.mxz files, and these were submitted for molecular networking using the GNPS platform[16]. The resulting molecular networks were graphically represented using Cytoscape[17].

2.4.5. Biological Testing

In vitro cytotoxicity studies were performed using H-460 human non-small cell lung cancer cells as previously described[18]. Briefly, H-460 cells were added to 96-well plates at 3.33×10^4 cells/mL of Roswell Park Memorial Institute (RPMI) 1640 medium with fetal bovine serum (FBS) and 1% penicillin/streptomycin. The cells were incubated overnight (37 °C, 5% CO_2) in a volume of 180 μL per well to allow recovery before treatment with test compounds. Compounds were dissolved in Dimethyl sulfoxide

(DMSO) to a stock concentration of 1 mg/mL. Working solutions were made through serial dilution in RPMI 1640 medium without FBS, with 20 μL added to each well producing final compound concentrations of 10, 3, 1, 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001 μg/mL. An equal volume of RPMI 1640 without FBS was added to wells designated as negative controls for each plate. Plates were incubated for approximately 48 h before MTT staining. Plates were read at 570 and 630 nm using a Thermo Electron Multiskan Ascent plate reader (Thermo Scientific, Waltham, MA, USA).

Brine shrimp toxicity studies were performed as previously described[19]. Briefly, brine shrimp eggs were hatched for 24 h in brine solution. Each well of the bioassay plate was prefilled with 2 mL brine solution, 200 μL brine shrimp culture mixture (around 15 brine shrimp in brine solution), and 300 μL brine solution, sequentially. Test fractions and compounds were dissolved and added to wells in 10 μL DMSO to afford a final concentration of 3 μg/mL and 30 μg/mL in each well, in duplicate per concentration. A sample of 10 μL DMSO was added to separate wells as a negative control. After 24 h, the number of dead non-moving brine shrimp was counted with the aid of a dissecting microscope. Acetone (1.2 mL) was added to each well to sacrifice the shrimp and the total number of dead shrimp was counted. The difference in the two counts represents the number of live shrimp at the end of the test period.

2.5 Chapter 2 Acknowledgements

We gratefully acknowledge the government of the Commonwealth of the Northern Mariana Islands for permission to collect and study marine samples. This research was partially supported by National Institute of Health (NIH) grant GM107550

(William H. Gerwick) and NIH postdoctoral fellowship T32 CA009523 (C. Benjamin Naman). We also thank Bailey Miller and Paul Boudreau for help with the cyanobacterial collections, Enora Briand for the cyanobacterium photomicrographs, Yongxuan Su (UCSD) for the HRMS data acquisition, Matthew Bertin for optical rotation instrument usage, Brendan Duggan (UCSD) and Anthony Mrse (UCSD) for the NMR instrument usage, Seth Cohen for FTIR instrument usage and Emily Mevers and Changlun Shao for comments on the configuration analysis, and John Lee for running the H-460 bioassay.

Chapter 2, in part, includes a reprint as it appears in the Marine Drugs. 2017, 15(4), 121, with the following authors Chen Zhang, C. Benjamin Naman, Niclas Engene, and William H. Gerwick. The dissertation author was a primary investigator and first author of this paper.

2.6 Chapter 2 Appendix

Table 2.6.1. Brine Shrimp Assay Results of the 10 Fractions (A-J) and Crude Extract of the Sample

Conc.	Tray #		A	B	C	D	E	F	G	H	I	J	Crude
3 µg/mL	1	death rate	-0.24	0.08	1	0.18	0	0	0.18	0	0	0	0
3 µg/mL	2	death rate	-0.1	0	1	0	0.14	0.06	0.11	0	0	0	0.06
		average	-0.17	0.04	1	0.09	0.07	0.03	0.15	0	0	0	0.03
30 µg/mL	3	death rate	0.25	1	0.95	1	1	0.89	0.71	0.86	0	0.14	1
30 µg/mL	4	death rate	0.29	0.83	1	1	1	1	1	0.79	0.09	0.12	1
		average	0.27	0.92	0.98	1	1	0.94	0.86	0.82	0.05	0.13	1

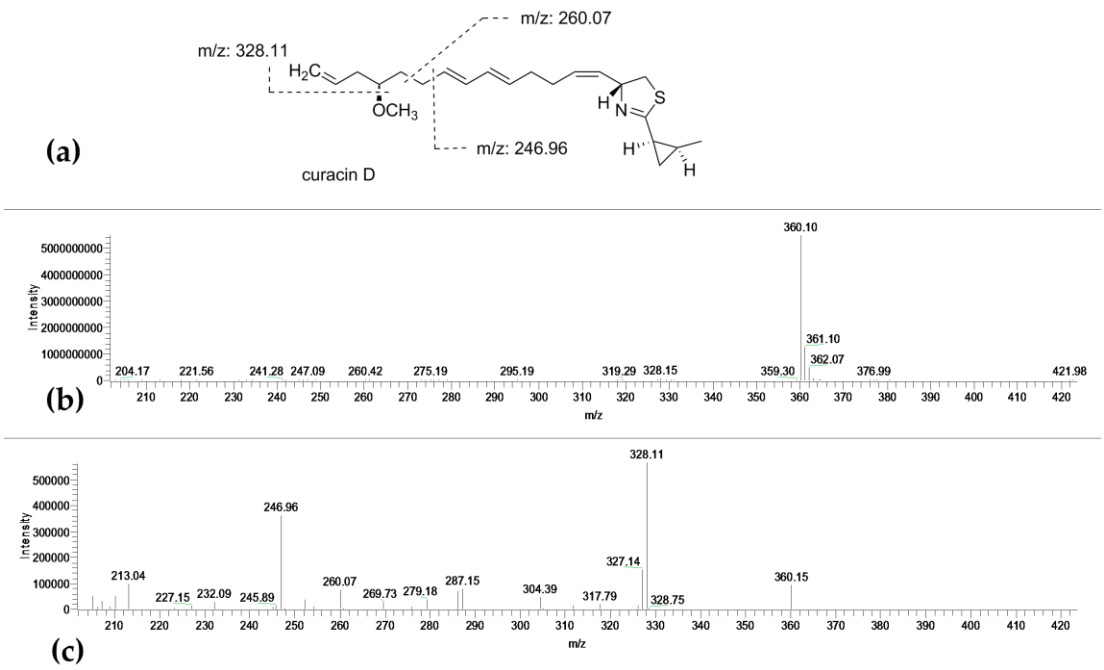


Figure 2.6.1. (**a**) Fragmentation analysis of curacin D (**3**). (**b**) MS of curacin D (**3**) (positive ion mode). (**c**) MS/MS (positive ion mode) spectra of curacin D (**3**).

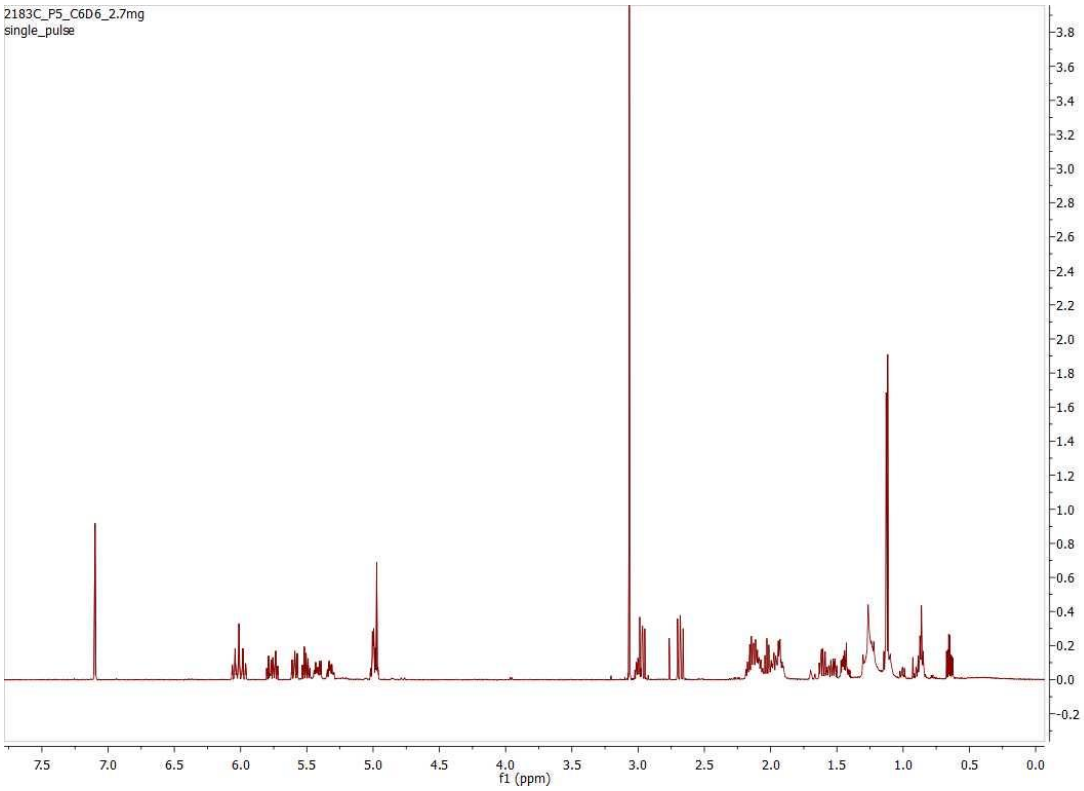


Figure 2.6.2. ^{1}H NMR spectra of curacin D (**3**) in C_6D_6.

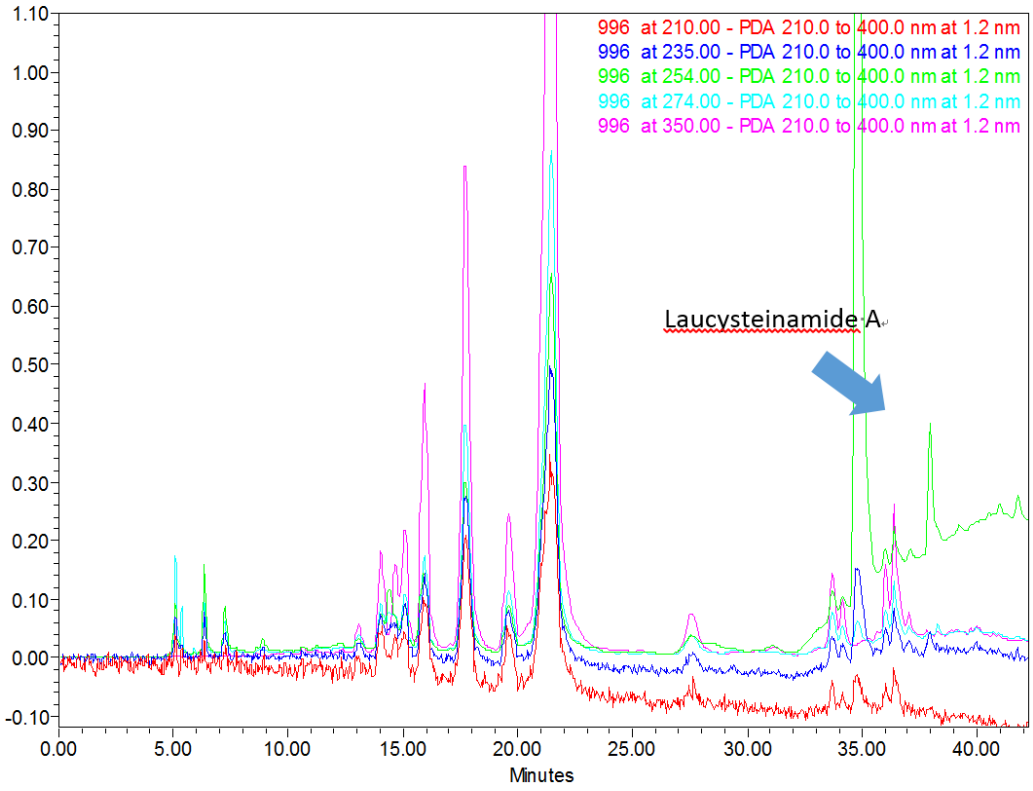


Figure 2.6.3. HPLC chromatogram of laucysteinamide A (**1**).

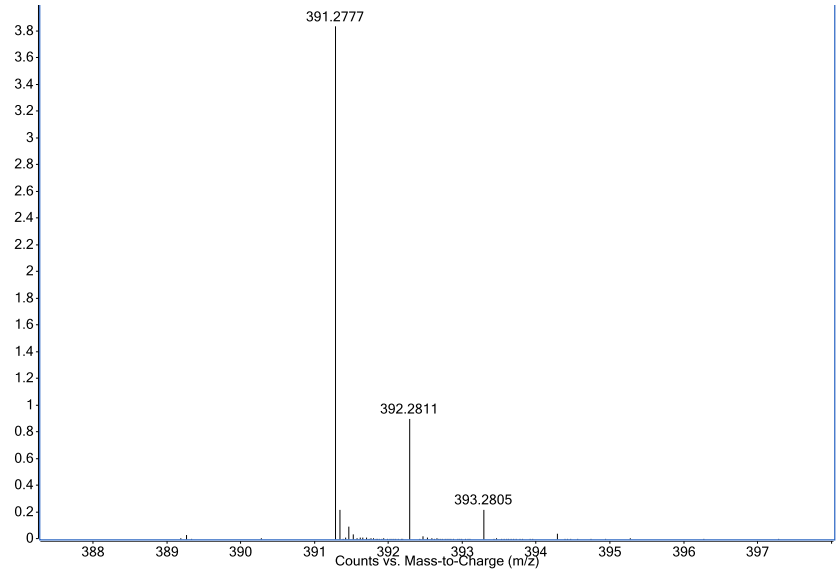


Figure 2.6.4. HRESITOFMS results of laucysteinamide A (**1**).

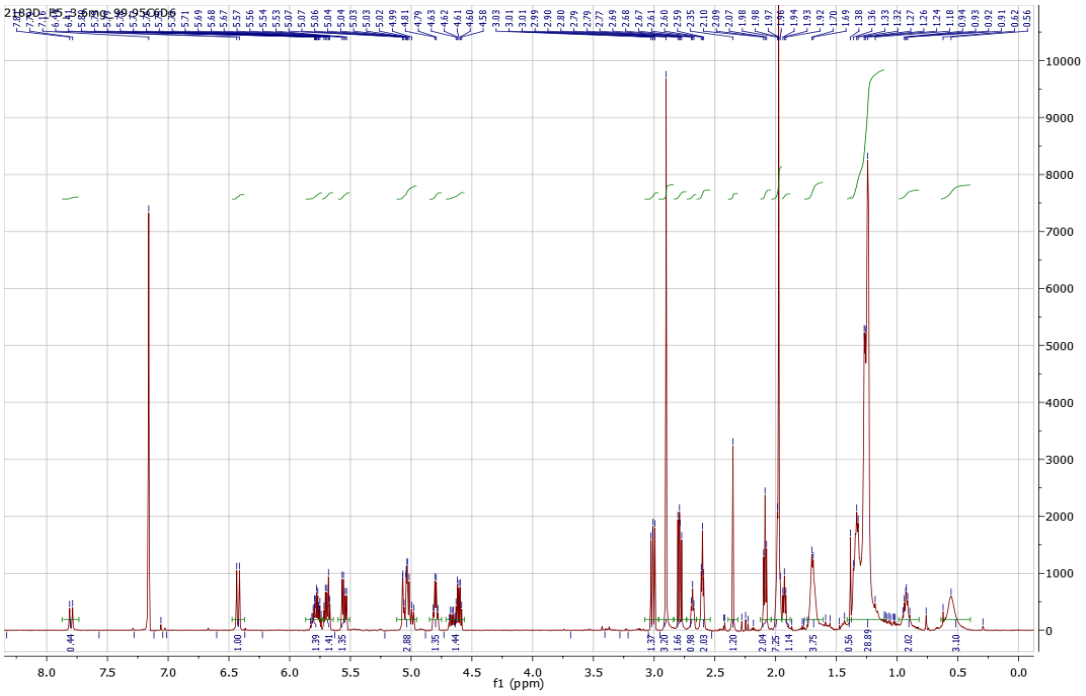


Figure 2.6.5. ^{1}H NMR spectra of Laucysteinamide A (**1**) in C_6D_6.

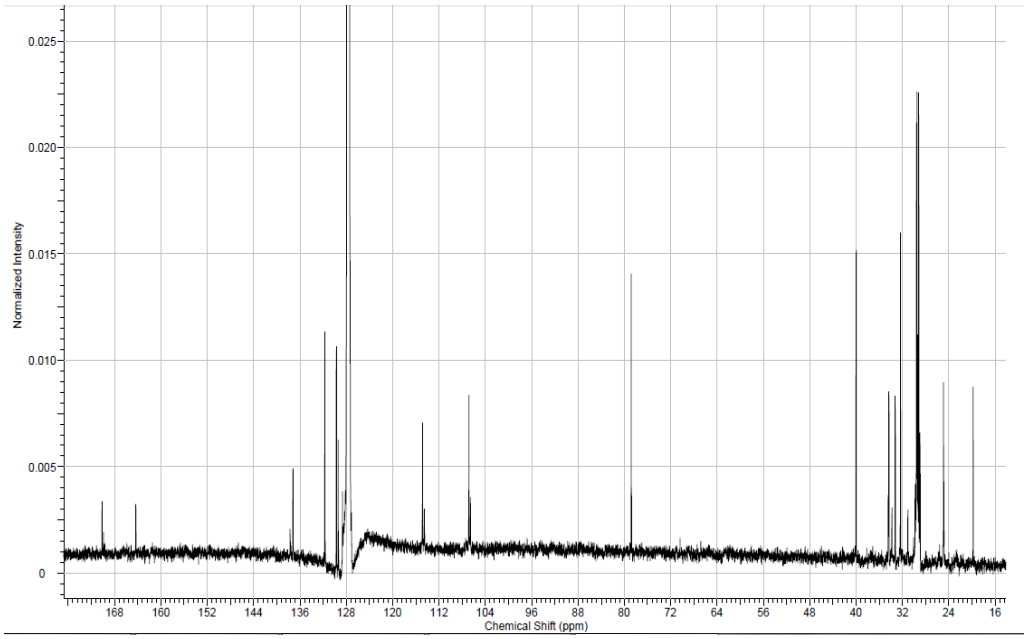


Figure 2.6.6. ^{13}C NMR spectra of Laucysteinamide A (**1**) in C_6D_6.

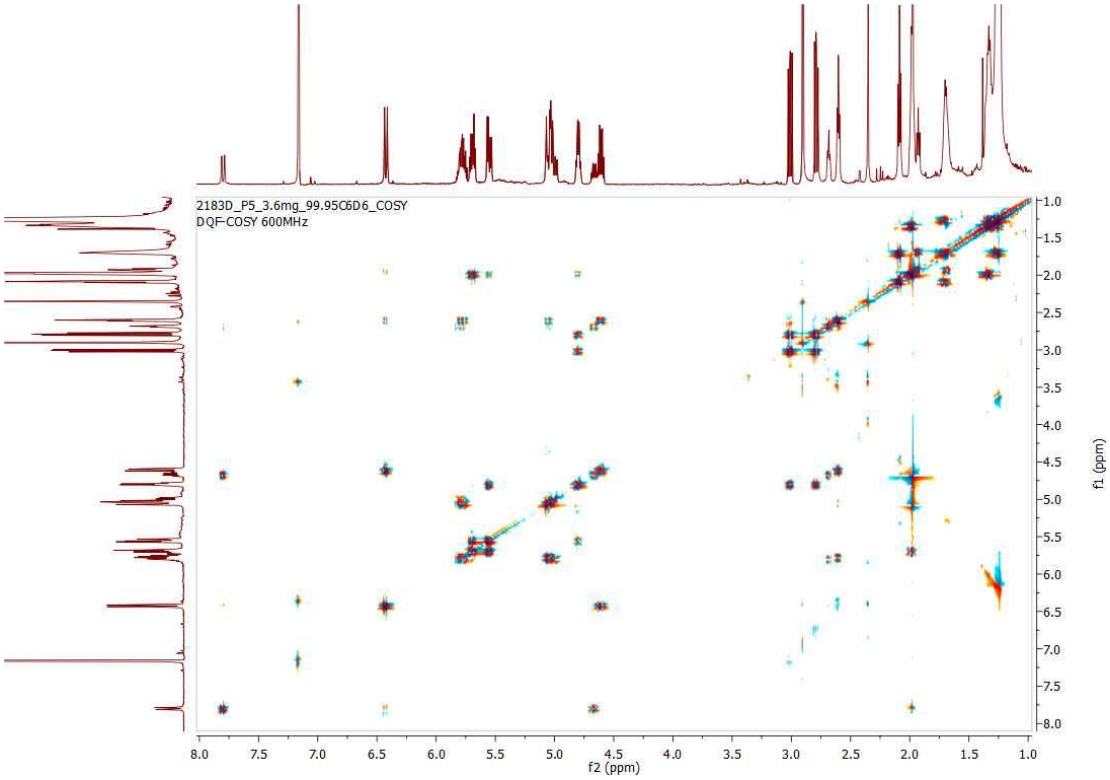


Figure 2.6.7. COSY spectra of Laucysteinamide A (**1**) in C_6D_6.

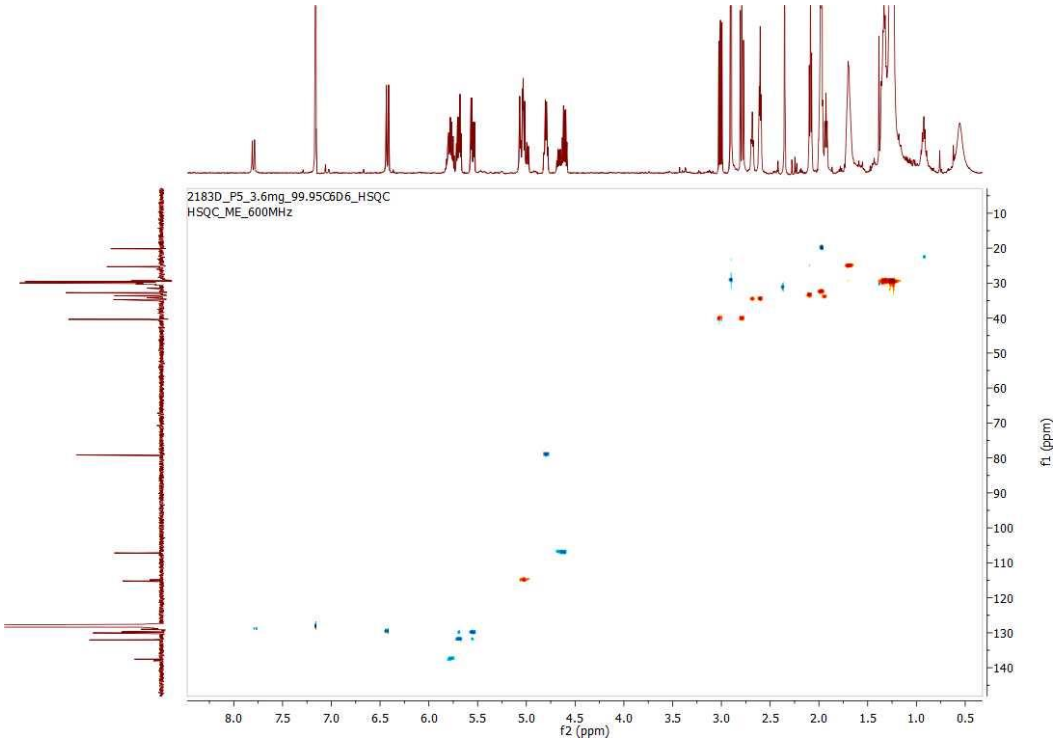


Figure 2.6.8. ^{1}H-^{13}C HSQC spectra of Laucysteinamide A (**1**) in C_6D_6.

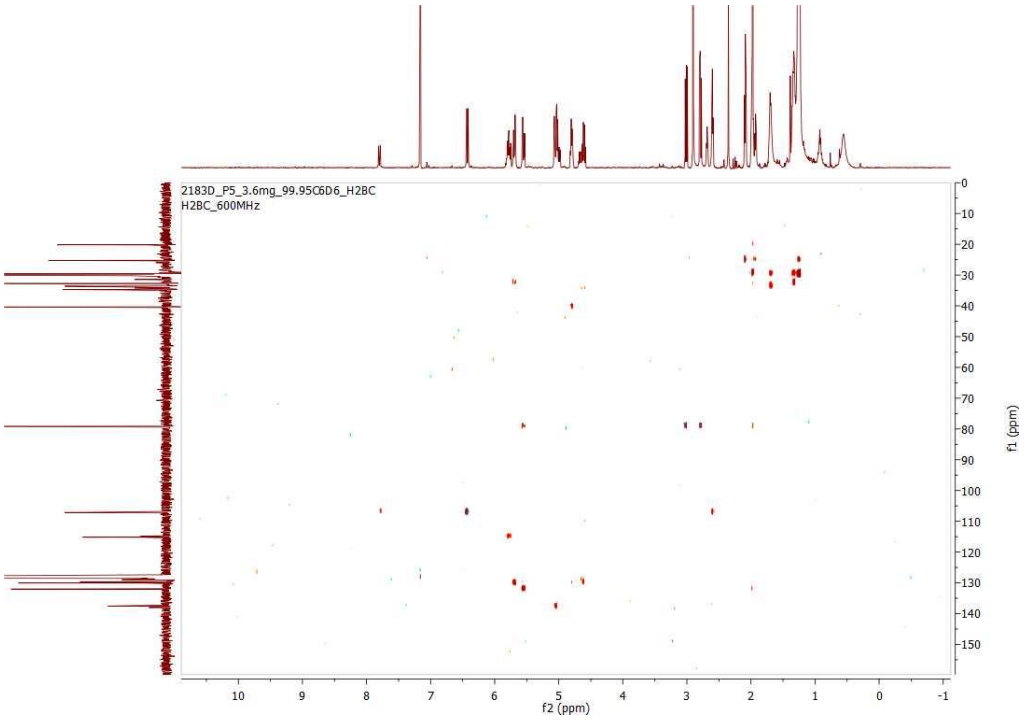


Figure 2.6.9. H2BC spectra of Laucysteinamide A (**1**) in C_6D_6.

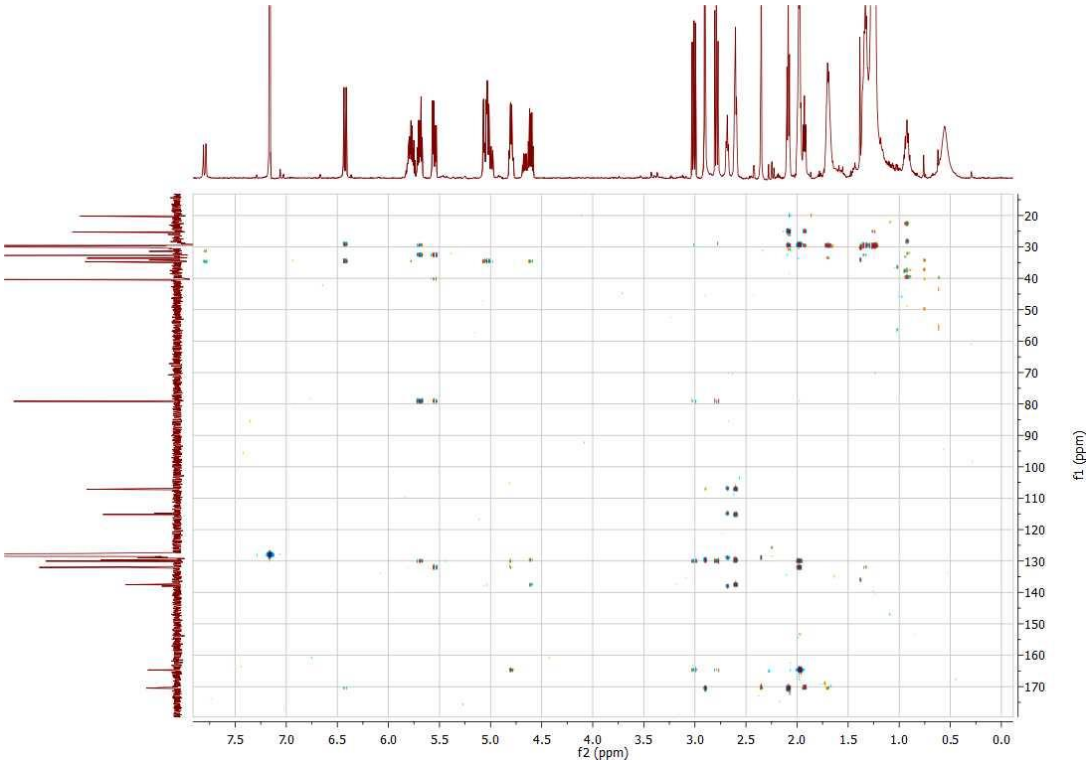


Figure 2.6.10. HMBC spectra of Laucysteinamide A (**1**) in C_6D_6.

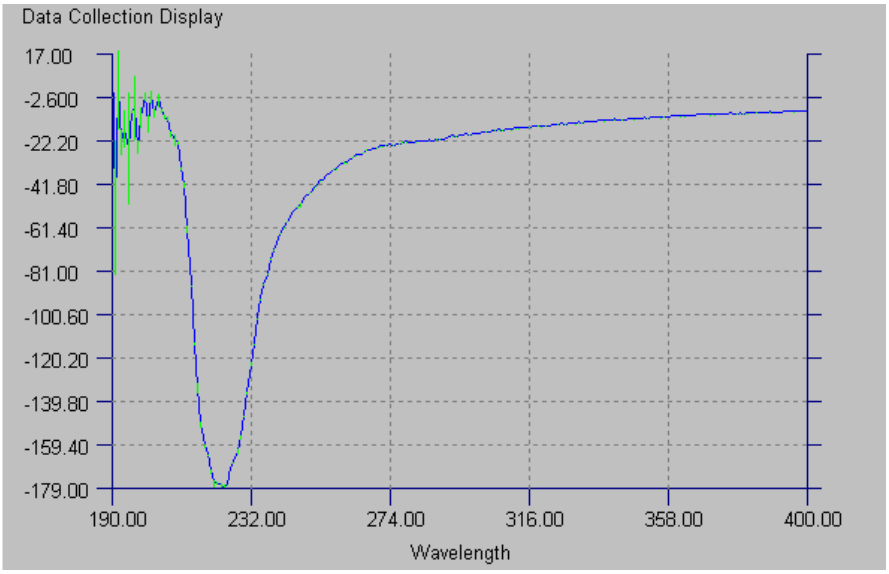


Figure 2.6.11. ECCD Spectrum of Laucysteinamide A (**1**). The compound was dissolved in dichloromethane for the experiment. The region above 200 nm is obscured by solvent absorptions.

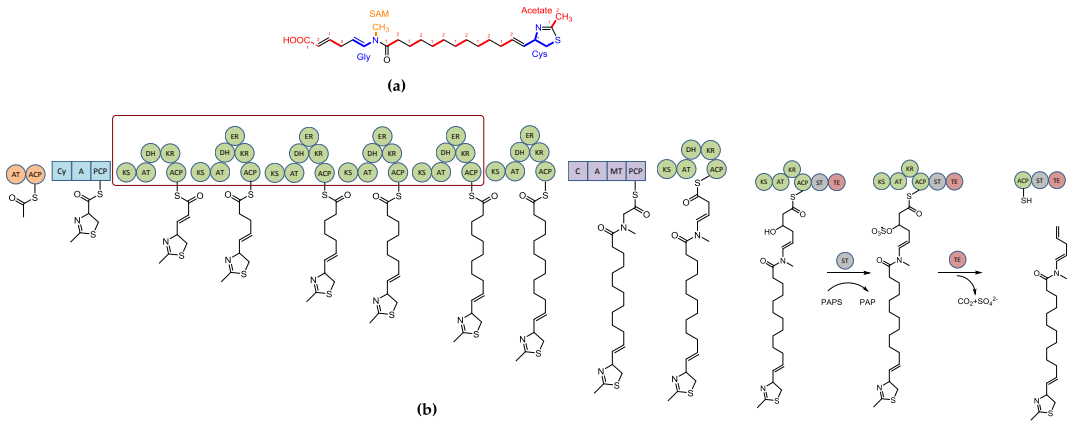


Figure 2.6.12. Biosynthetic scheme proposed for laucysteinamide A. (**a**) A hypothesized hybrid PKS/NRPS pathway of biosynthetic precursors. (**b**) The hybrid PKS/NRPS pathway prediction before chain termination with enzymatic domain. The last three steps show the predicted chain termination mechanism in laucysteinamide A (**1**) biosynthesis. This proposed biosynthetic pathway is based on that described for curacin A biosynthesis process[20, 21]. Abbreviations: ACP, acyl carrier protein; KS, β-ketoacyl-ACP synthase; KR, β-ketoacyl-ACP reductase; AT, acyl transferase; DH, β-hydroxy-acyl-ACP dehydratase; ER, enoyl reductase; MT, N-methyl transferase; PCP, peptidyl carrier protein; Cy, condensaton/cyclization domain; A, adenylation domain; ST, sulfotransferase; PAPS, adenosine 3-phosphate 5-phosphosulfate; PAP, adenosine 3-phosphate 5-phosphate; TE, thioesterase.

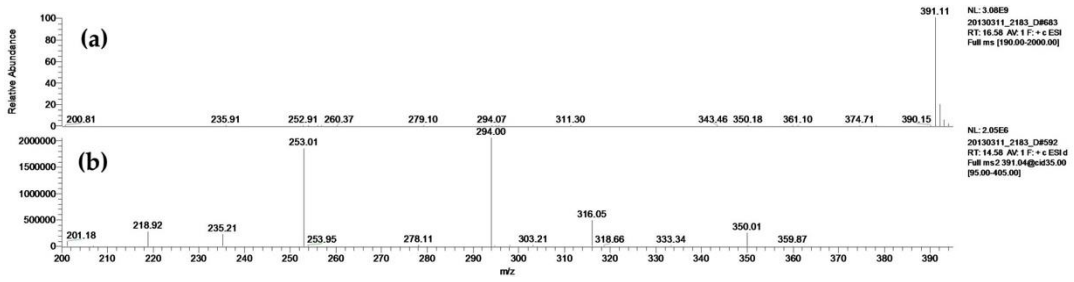


Figure 2.6.13. (**a**) Low resolution MS (positive ion mode) of laucysteinamide A (**1**). (**b**) MS/MS (positive ion mode) spectra of **1**.

2.7 Chapter 2 References

1. Zhang, C.; Naman, C. B.; Engene, N.; Gerwick, W. H., Laucysteinamide A, a Hybrid PKS/NRPS Metabolite from a Saipan Cyanobacterium, cf. Caldora penicillata. *Marine Drugs* **2017,** 15, (4).

2. Blunt, J. W.; Copp, B. R.; Keyzers, R. A.; Munro, M. H.; Prinsep, M. R., Marine natural products. *Nat Prod Rep* **2015,** 32, (2), 116-211.

3. Williams, P. G.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J., Isolation and structure determination of obyanamide, a novel cytotoxic cyclic depsipeptide from the marine cyanobacterium Lyngbya confervoides. *J Nat Prod* **2002,** 65, (1), 29-31.

4. Kingston, D. G. I., Tubulin-Interactive Natural Products as Anticancer Agents. *Journal of Natural Products* **2009,** 72, (3), 507-515.

5. Dumontet, C.; Jordan, M. A., Microtubule-binding agents: a dynamic field of cancer therapeutics. *Nature Reviews Drug Discovery* **2010,** 9, (10), 790-803.

6. Marquez, B.; Verdier-Pinard, P.; Hamel, E.; Gerwick, W. H., Curacin D, an antimitotic agent from the marine cyanobacterium Lyngbya majuscula. *Phytochemistry* **1998,** 49, (8), 2387-2389.

7. Nogle, L. M.; Gerwick, W. H., Somocystinamide A, a novel cytotoxic disulfide dimer from a Fijian marine cyanobacterial mixed assemblage. *Organic Letters* **2002,** 4, (7), 1095-1098.

8. Engene, N.; Tronholm, A.; Salvador-Reyes, L. A.; Luesch, H.; Paul, V. J., Caldora Penicillata Gen. Nov., Comb. Nov (Cyanobacteria), a Pantropical Marine Species with Biomedical Relevance. *Journal of Phycology* **2015,** 51, (4), 670-681.

9. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X. T.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L. X.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C., Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* **2013,** 76, (9), 1686-1699.

10. Katritzky, A. R.; Hall, C. D.; El-Dien, B.; El-Gendy, M.; Draghici, B., Tautomerism in drug discovery. *Journal of Computer-Aided Molecular Design* **2010,** 24, (6-7), 475-484.

11. Gerwick, W. H.; Proteau, P. J.; Nagle, D. G.; Hamel, E.; Blokhin, A.; Slate, D. L., Structure of Curacin-a, a Novel Antimitotic, Antiproliferative, and Brine Shrimp Toxic Natural Product from the Marine Cyanobacterium Lyngbya-Majuscula. *Journal of Organic Chemistry* **1994,** 59, (6), 1243-1245.

12. Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P., RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *Journal of Computational Chemistry* **2006,** 27, (10), 1101-1111.

13. Harada, N.; Chen, S. L.; Nakanishi, K., Quantitative Definition of Exciton Chirality and Distant Effect in Exciton Chirality Method. *Journal of the American Chemical Society* **1975,** 97, (19), 5345-5352.

14. Wrasidlo, W.; Mielgo, A.; Torres, V. A.; Barbero, S.; Stoletov, K.; Suyama, T. L.; Klemke, R. L.; Gerwick, W. H.; Carson, D. A.; Stupack, D. G., The marine lipopeptide somocystinamide A triggers apoptosis via caspase 8. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (7), 2313-2318.

15. Gehret, J. J.; Gu, L. C.; Gerwick, W. H.; Wipf, P.; Sherman, D. H.; Smith, J. L., Terminal Alkene Formation by the Thioesterase of Curacin A Biosynthesis STRUCTURE OF A DECARBOXYLATING THIOESTERASE. *Journal of Biological Chemistry* **2011,** 286, (16), 14445-14454.

16. Wang, M. X.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Criisemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.;

Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J. Q.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrovr, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W. Y.; Liu, X. T.; Zhang, L. X.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** 34, (8), 828-837.

17. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **2003,** 13, (11), 2498-2504.

18. 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., Feasibility of Drug Screening with Panels of Human-Tumor Cell-Lines Using a Microculture Tetrazolium Assay. *Cancer Research* **1988,** 48, (3), 589-601.

19. Meyer, B. N.; Ferrigni, N. R.; Putnam, J. E.; Jacobsen, L. B.; Nichols, D. E.; Mclaughlin, J. L., Brine Shrimp - a Convenient General Bioassay for Active-Plant Constituents. *Planta Medica* **1982,** 45, (1), 31-34.

20. Gu, L. C.; Wang, B.; Kulkarni, A.; Gehret, J. J.; Lloyd, K. R.; Gerwick, L.; Gerwick, W. H.; Wipf, P.; Hakansson, K.; Smith, J. L.; Sherman, D. H., Polyketide Decarboxylative Chain Termination Preceded by O-Sulfonation in Curacin A Biosynthesis. *Journal of the American Chemical Society* **2009,** 131, (44), 16033-+.

21. Chang, Z. X.; Sitachitta, N.; Rossi, J. V.; Roberts, M. A.; Flatt, P. M.; Jia, J. Y.; Sherman, D. H.; Gerwick, W. H., Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **2004,** 67, (8), 1356-1367.

CHAPTER 3

DEVELOPING SMALL MOLECULE ACCURATE RECOGNITION TECHNOLOGY (SMART), A DEEP LEARNING AND FAST 2D NMR BASED DIGITAL FRONTIER TO ENHANCE NATURAL PRODUCTS DISCOVERY

3.0. Abstract

Various algorithms comparing 2D NMR spectra have been explored for their ability to dereplicate natural products as well as determine molecular structures. However, spectroscopic artefacts, solvent effects, and the interactive effect of functional group(s) on chemical shifts combine to hinder their effectiveness. Here, we leveraged Non-Uniform Sampling (NUS) 2D NMR techniques and deep Convolutional Neural Networks (CNNs) to create a tool, SMART, that can assist in natural products discovery efforts. First, an NUS heteronuclear single quantum coherence (HSQC) NMR pulse sequence was adapted to a state-of-the-art nuclear magnetic resonance (NMR) instrument, and data reconstruction methods were optimized, and second, a deep CNN with contrastive loss was trained on a database containing over 2,054 HSQC spectra as the training set. To demonstrate the utility of SMART, several newly isolated compounds were automatically located with their known analogues in the embedded clustering space, thereby streamlining the discovery pipeline for new natural products.

3.1 Introduction

As a discipline, natural products research (NPR) enables and benefits numerous downstream research fields, such as chemical biology, chemical ecology, drug discovery and development, pharmacology and the total chemical synthesis of natural products (NPs). In this regard, approximately 70% of all approved drugs are NPs, their analogues, or a chemical modification of an existing NP[1]. In addition to these academic and societal benefits, NPR provides a powerful incentive for the conservation and sustainable use of biodiversity and biodiverse habitats[2].

An important step in NPR is dereplication, the process of assessing the uniqueness of a new compound in relationship to all known ones. In most NPR, traditional compound dereplication practices have entailed the collection and analysis of nuclear magnetic resonance (NMR) spectra, including running 1D and 2D NMR spectroscopic experiments for the purposes of molecular framework construction, assembly, and relative stereochemistry determination. More recently, mass spectrometric approaches and mass spectrometry (MS)-based molecular networking[3], in part stimulated by integration with DNA sequencing and genome mining[4,5] have been integrated into NPR workflows. Nevertheless, conventional NMR practices are still indispensable to the characterization and dereplication of NPs. Unfortunately, 2D NMR experiments can be time consuming, especially when the sample is relatively scarce. Furthermore, 2D NMR-based structural assignments can sometimes take protracted periods of time to accomplish due to the inherent structural complexity of some NPs.

Along with relatively recent improvements in mass spectrometry, circular dichroism and infrared spectroscopy techniques, state-of-the-art cryoprobe NMR

instruments have reduced the sample requirements for NPs discovery to just a few nanomoles.[6] Nevertheless, acquisition of NMR spectra may still require a relatively large number of time consuming scans before Fourier transformation of the data. Furthermore, conventional 2D NMR spectroscopy relies upon linear sampling of the frequency evolution in the indirect dimension (usually the ^{13}C dimension). When generating high frequency resolution in the indirect dimension, extensive sampling is required and the experiments become very time consuming. Modification of conventional uniform sampling to non-uniform sampling (NUS)[7-13] allows the number of experiments in the indirect dimension to be reduced, thereby reducing the overall time of the experiment. The NUS method is designed to reduce the number of data collection experiments while at the same time delivering an accurate estimation of the fully sampled spectrum.

To streamline compound dereplication or even structure determination, algorithms have been applied for 2D NMR spectra comparisons, such as the 2D NMR peak alignment algorithm[14,15]. However, these techniques are not powerful enough to accurately classify 2D NMR spectra into the correct NP family. This arises for several reasons, such as compound concentration, solvent effects, and the interactive effect of a single functional group alteration on ^{1}H and ^{13}C NMR chemical shifts, all of which combine to increase the difficulty for computer assisted 2D NMR data analysis. At the same time, artefacts are often introduced into NMR spectra, and this makes it difficult for existing pattern recognition or overlap methods to distinguish genuine peaks from artefacts. However, artificial intelligence technologies, such as deep learning[16,17], have generated new approaches for meeting these challenges. Compared with conventional

machine learning methods, which require the cumbersome process of selecting and creating features that might be suboptimal for a given task, deep learning allows creation of the most suitable set of features within the process of training, without any design or involvement by the investigator. Moreover, the deep learning method works well even when the number of categories is very large and unknown during the training process. Thus, deep learning is an ideal method by which to analyse and categorize 2D NMR spectra of NPs. For NPs, there are an essentially unlimited number of categories for different compound families, with many being unknown at the present time. Additionally, it is quite common for each category to contain fewer than 50 different members; in the work of our laboratory with marine cyanobacterial NPs, this is the case for all of the molecular families we have encountered over the past 40 years, including the curacins[18-20], apratoxins[21], lyngbyabellins[22] and majusculamides[23-25].

Other approaches for automatic recognition of NMR spectra have appeared in the literature or private sector. The typical approach is to create grids over the data and then compute similarities based on how many points fall into the same grid cells[26]. This approach can miss peaks that are near one another that happen to fall in different grid cells, so an enhancement of this approach is to use multiple grid resolutions and offsets before computing the similarities[27]. Our convolutional network approach automatically does this by using overlapping convolutions combined with increasing-sized receptive fields through pooling the results from previous layers. However, our method of deciding similarity is learned by the network through nonlinear dimensionality reduction via training it to map together those compounds it recognizes as being from

the same family, and to map different families to different locations in the underlying space.

Another method involves computer-aided structure elucidation (CASE, ACD/Labs) which is largely based on applying a least-squares regression (LSR) approach for comparing NMR chemical shifts; this tactic is not powerful enough to satisfactorily accommodate solvent effects, instrumental artefacts, or weak signal issues[14,15]. An early effort using machine learning applied to NMR spectra was reported in (Wolfram et al., 2006)[28]. They used Probabilistic Latent Semantic Indexing (PLSI), a method usually applied to text documents for information retrieval purposes. PLSI maps documents into a lower dimensional space using a probabilistic analogue to Singular Value Decomposition (SVD) applied to a document by word count matrix. To apply PLSI to compounds, the typical multi-scale and shifted grid cell approach was used, treating each grid cell as a "word" in the compound. This is essentially learning a linear mapping from the feature space to a reduced space, and thus is not as powerful as using a nonlinear deep network.

In our approach, heteronuclear single quantum correlation (HSQC)[29] spectra are recorded using a 2D NMR pulse sequence that utilizes the large heteronuclear coupling between directly bonded nuclei within an organic molecule to correlate directly bonded atoms (*e.g.* ^{1}H and ^{13}C, with ^{1}H being defined as the direct dimension and ^{13}C the indirect dimension). The peaks of those correlated nuclei in the 2D HSQC spectra are generated by detecting coherences that connect states whose total *z*-angular momentum quantum numbers differ by one order (*i.e.* single-quantum coherences). In this regard, an HSQC spectrum is deemed as the 'fingerprint' or 'face' for a natural product

molecule, and thus is highly discriminating. Specifically, within a 2D HSQC spectrum, signals in the direct dimension can be distinguished if they have shifts of 0.02 ppm or greater, and in the indirect dimension if they have shifts of 0.1 ppm or greater. Furthermore, most ^{1}H chemical shifts occur between 0.5 and 9.5 ppm, whereas in the ^{13}C dimension chemical shifts typically occur between 10 and 215 ppm, which gives rise to 922,500 distinguishable positions within a 2D HSQC spectrum. When summed over all protonated carbons in a molecule of 20 carbons with attached protons, the number of potential combinations is in the tens of millions, and is thus highly discriminatory. In addition, this technique avoids detection of double-quantum coherence, resulting in relatively few artefacts. In contrast, the commonly used heteronuclear multiple bond correlation (HMBC) experiment detects two and three bond correlations by selecting smaller multiple bond heteronuclear coupling constants (around 5-10 Hz for ^{1}H-^{13}C versus one bond of 125-170 Hz) for double-quantum and zero-quantum transfer. Therefore, while the HMBC experiment produces an even larger amount of theoretical information, it is prone to introducing artefacts and its complexity makes it more difficult to interpret. In addition, the HSQC when performed with NUS discussed above is a relatively quick and efficient experiment for data accumulation.

Here, we report the development of the Small Molecule Accurate Recognition Technology (SMART) prototype, a system that integrates the benefits of NUS NMR with advances in deep learning to enhance and improve the efficiency of NPs dereplication. To create SMART, a database of training examples containing 2D HSQC spectra of 2,054 compounds was compiled. These examples were used to train a deep network that learns to map the spectra into a cluster space where similar compounds are

near one another and dissimilar compounds are far apart. To perform this function, we use a deep convolutional neural network (CNN) employing a siamese architecture[30] as described in the methods section. A siamese network amplifies the number of training examples by training on pairs of spectra that are labelled "same" or "different," rather than training on individual examples. The network then learns features of the spectra that are relevant to their similarities and differences, and uses this to create the cluster space. The resulting mapping then generalizes to new compounds, placing them in the space near compounds with similar HSQC spectra. We evaluate SMART by holding back several known NPs from different families from the training set, and then show that SMART maps them into their proper location within the cluster space. We also present here the rapid identification of a newly isolated natural product compound family as a result of SMART's ability to cluster similar compounds together. HSQC spectra were collected for several nonribosomal peptide synthetase (NRPSs)-derived NPs that had been isolated from two marine cyanobacteria. These novel spectra were accurately mapped by SMART into the 'viequeamide' subfamily of NPs.

3.2 Results and Discussion

3.2.1. The SMART Prototype

SMART is a user-friendly, AI-based dereplication and analysis tool that uses 2D NMR data to rapidly associate newly isolated NPs with their known analogues. SMART has been designed to mimic the normal path of experiential learning in that additional 2D NMR spectral inputs can be used to enrich its database and improve its performance. In short, SMART aims to become an experienced associate to natural products

researchers as well as other classes of organic chemists. The SMART workflow consists of three steps, *1*) 2D NMR data acquisition by NUS HSQC pulse sequence, *2*) 2D NMR spectral analysis by deep CNN, resulting in an embedding of the spectra into a similarity space of NPs, and *3*) molecular structure dereplication or determination by the user (Figure 3.1). This process gives users rapid access to a well-organized map of structurally determined NPs, and helps ensure that SMART's insights are chemically rational. In this regard, SMART capitalizes on the wealth of molecular fingerprints, namely 2D HSQC spectra, built over the past four decades[31,32], and reciprocally, we anticipate that 2D HSQC spectral databases will experience an accelerating expansion as a result of SMART's application.

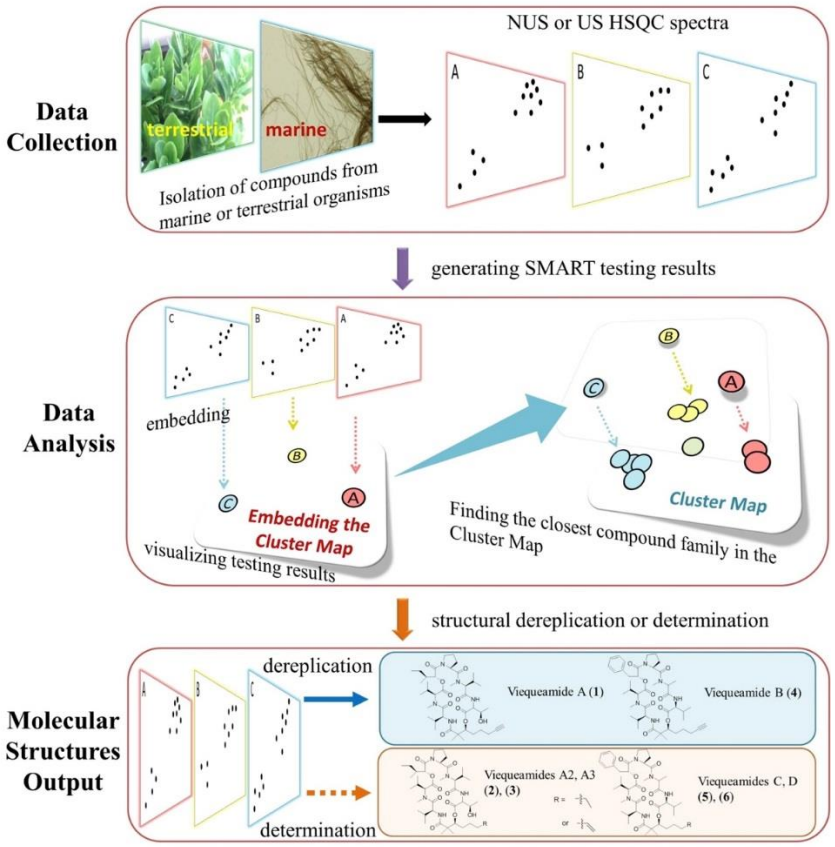


Figure 3.1. Workflow for the Small Molecule Accurate Recognition Technology (SMART). Experimental HSQC spectra of newly isolated pure natural product molecules collected using either NUS HSQC pulse sequences or conventional HSQC techniques, are automatically embedded by SMART into a cluster space near similar, previously-characterized compounds. The resultant embedding in the cluster map is visualized using the Bokeh visualization package[72], where each node represents an HSQC spectrum processed by SMART. The node colours in a local area of the clustering map designate compounds from the same journal articles and thus of the same natural product family. This facile method allowed tracking of compounds into SMART, but is not of paramount significance in that some compounds reported in different publications display closer relationships in SMART and by structural comparison than to compounds within their same article. When available, the node labels are the compound names; otherwise, the labels are for the organism from which the compound derives. Node distance is proportional to relatedness, a quantification of molecular structural similarity. The 2D cluster map is created by performing Principal Component Analysis (PCA) of the 10D space outputs to reduce to 2D. Optionally, the top 5, 10 and 20 closest nodes in the 10D space are available in text format. The proof-of-concept experiments are illustrated: Dereplication (solid blue arrow) of viequeamides A (**1**) and B (**4**), and determination (dashed orange arrow) of viequeamides A2 (**2**), A3 (**3**), C (**5**) and D (**6**), isolated from 1) *Rivularia sp.*, collected in Vieques, Puerto Rico and 2) *Moorea sp.*, collected in American Samoa.

The workflow (Figure 3.1) of SMART begins with recording the NUS HSQC spectrum for a pure small organic molecule; in the case of NPR, this is a substance extracted and purified from an organism of interest, but just as easily could be a small molecule produced from organic synthesis, biosynthesis or from a metabolomic study. A small molecule is defined here as one whose transverse relaxation time constant (T_2) is on the same order of magnitude as its longitudinal relaxation time constant (T_1) when dissolved in liquid solution. In other words, the nuclear spins of a small molecule should be synchronized between 10^7 to 10^8 Larmor precession cycles during a liquid state 2D HSQC experiment[33]. Nevertheless, the SMART concept is not inherently confined to small molecule NUS NMR spectra, considering the ability of NMR to structurally characterize molecules of many sizes and types. NUS HSQC experiments are highly advantageous for small molecule structure elucidation compared with conventional pulse sequences due to their rapid acquisition, few spectral artefacts, and intrinsic high resolution. Nevertheless, as discussed below, conventional 2D HSQC spectra can be provided to the AI system and spectral recognition achieved. In fact, the initial database of HSQC spectra that were compiled to train the SMART system was acquired in this manner.

Due to lower sampling density, NUS HSQC requires alternative approaches to convert the indirectly sampled time domain into visual spectra of the frequency domain, and thus methods other than the Discrete Fourier Transform are required. To this end, Iterated Soft Thresholding (IST)[34,35] followed by the Maximum Entropy Method (MEM)[36,37] was applied to NUS data collected for the model compound strychnine. In

order to achieve convergence in local minima, the Lagrange multiplier was applied to weight the regularization function, the L_1 norm, in the IST routine. Previous studies[12] have shown that IST is superior to Maximum Entropy Reconstruction (MaxEnt)[38] (not to be confused with MEM) in NUS NMR data reconstruction, owing to the simplicity of IST with fewer adjustable parameters and the resultant ease of application. Nevertheless, IST suffers slower convergence compared to MaxEnt for spectra with a high dynamic range. However, it has been shown that changing the step sizes in IST can achieve visualization of the final spectra indistinguishable from those reconstructed by a well-tuned MaxEnt process[39]. The MEM can then be applied after Fourier Transformation of the IST reconstructed data in the direct dimension, resulting in an improvement that derives from the fact that MEM is biased towards the enhancement of smaller line widths.[40] For the model compound, the HSQC correlation signals of the C-11 methylene protons (3.11 ppm and 2.67 ppm) to their subtending carbon were visibly strengthened after sequentially applying IST (400 iterations) and MEM (3 iterations) compared with application of IST (400 iterations) with Linear Predictions (LP) during data reconstruction of the non-uniformly sampled 2D NMR spectra (Figure 3.2).

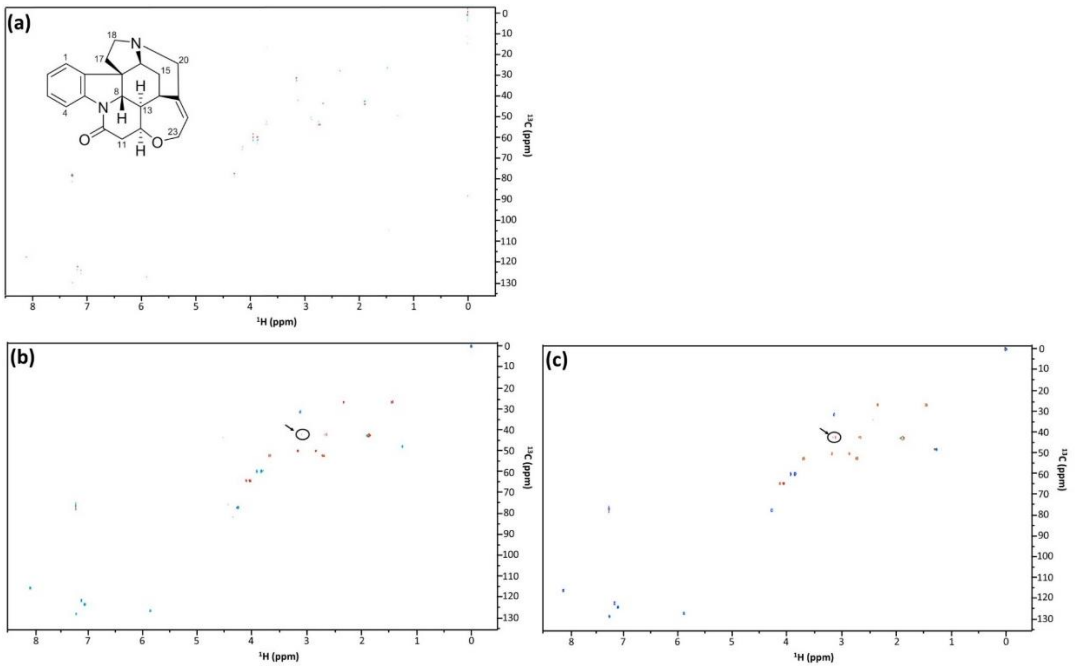


Figure 3.2. Data reconstruction results of a non-uniformly sampled (NUS) HSQC experiment. All of the three full HSQC spectra were recorded with a 50 nmole strychnine sample in $CDCl_3$ on a 600 MHz Bruker 1.7 mm cryoprobe instrument, using 32 out of a total 128 increments (25% sampling density) in the indirect dimension and 8 scans. The differences between the three spectra were that (a) was transformed with the maximum entropy method (MEM) alone, (b) was transformed with the iterative soft thresholding (IST) alone, and (c) was transformed with IST followed by MEM. The doublet (see black arrows and circles in (b) and (c)) associates with the protons on the methylene (C-11) adjacent to the ketone in strychnine.

Our deep learning method is based on a siamese neural network architecture[41]. A siamese network is comprised of a pair of identical networks that are trained with pairs of inputs. These are mapped to a representational space where similar items are near one another and different items are further apart. As a result, it produces a clustering of the input space based on a similarity signal. In our case, it first maps the input HSQC spectra into a ten dimensional space, which was then further mapped into a lower two dimensional space where HSQC spectra are clustered.

Because HSQC spectra are inherently a visual input, we used convolutional neural networks (CNNs)[42] as the components of our siamese network. CNNs are currently the best method for image processing in the computer vision community, and have revolutionized the field of computer vision.[42-44] Like standard neural networks, they are trained by backpropagation of errors.[45] CNNs are structured to learn local visual features that are replicated across the input, hence the term "convolutional". The local maximum of these features are then input to another layer that learns local features over the previous layer of features, and this process is repeated for several layers. In previous work, it has been shown that the feature maps resulting from each convolutional layer become more abstract as the layers of the network are traversed. We show the first layer features in Figure 3.3. By using the local maxima of feature responses over nearby locations in the input, the network will generalize to patterns that are shifted in the (f_1, f_2) plane of the spectra, *i.e.*, it achieves some translation invariance. Thus, the network is inherently hierarchical, like the mammalian visual system, and learns more and more abstract features in deeper layers of the network. In a siamese network, the final layer is not trained to classify the inputs; instead, a set of units are trained to give similar patterns of activation for similar inputs (as given in the teaching signal) and different patterns of activation for inputs that are labelled as different. Hence, they produce a clustering in the space of unit activations[46].

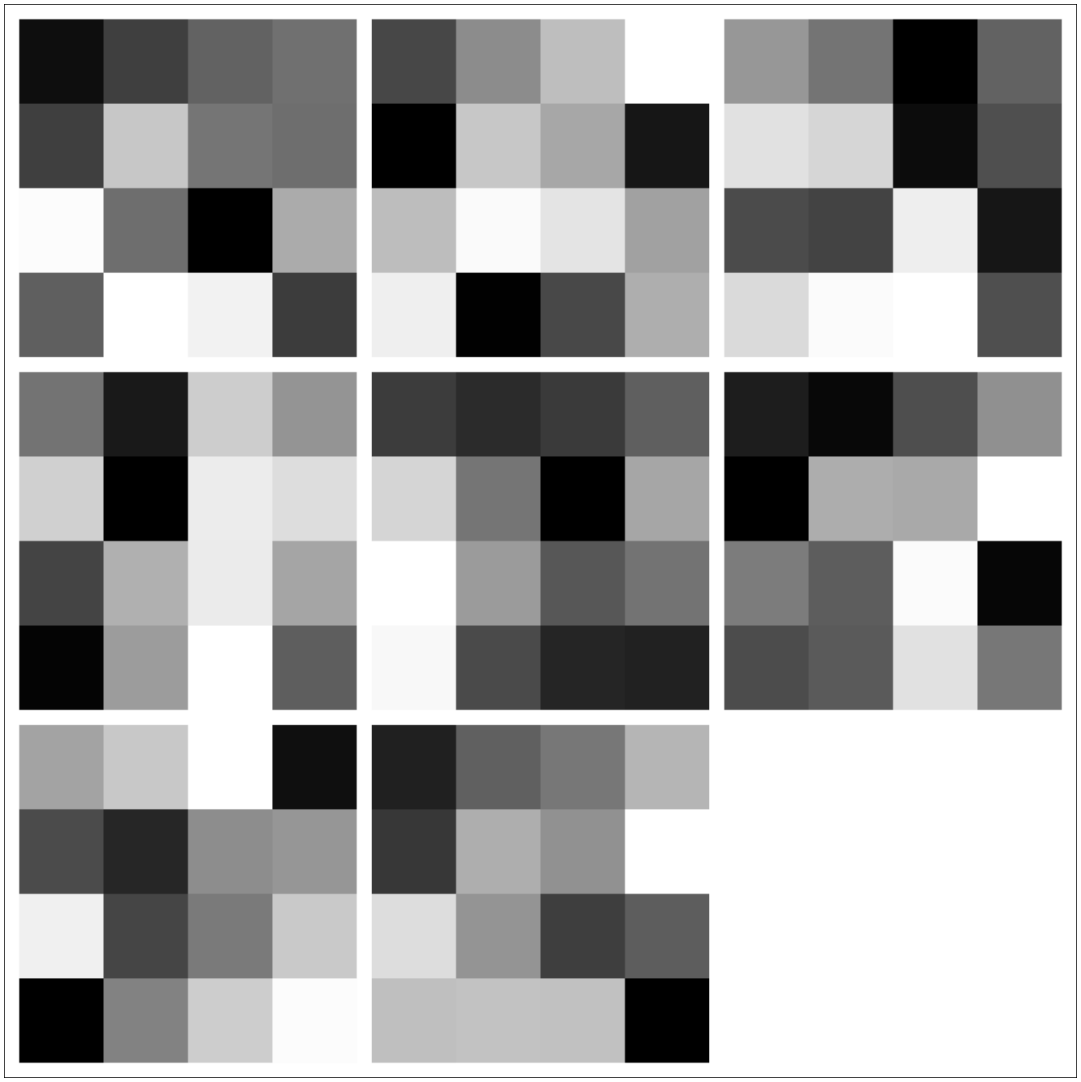


Figure 3.3. Features learnt by the first convolutional layer of the CNN. Feature maps were extracted from convolution layer 1 in Table 3.1, with the eight blocks of 4×4 pixels in this figure corresponding to the results of each of the eight filters applied to the HSQC dataset.

As a result, molecules that are similar in HSQC spectra will be mapped to nearby locations in the output space. If the network generalizes well, it will place novel molecules near known ones that have similar NMR spectra. This allows the system to rapidly identify candidate known molecules that may have similar chemical features to the novel molecule, allowing the user to search through a small subset of known

molecules for similar compounds. In our initial approach, we used ten output units (*i.e.*, a 10 dimensional space), which can be visualized by applying Principal Components Analysis (PCA) to reduce the 10 dimensions to two.

3.2.2. Network Training and Results

The neural network was trained using stochastic gradient descent[47] with the Adagrad[48] update rule. To speed the training, we employed batch normalization[49], which reduces the internal covariance shift by standardizing the distribution of features on each layer. The network was found to train 7 times faster (the wall clock time) using batch normalization.

When training the CNN, the datasets (see the Methods section for details) were divided into three subsets; the training set containing 80% of the spectra, used to adjust the parameters of the network, the validation set containing 10% of the data used for early stopping, and a test set containing the remaining 10% of the data (for details, see Methods). The test set consisted of HSQC spectra that were not used during the training process. The error from the validation set was monitored to prevent overfitting on unseen data. The test spectra were then embedded in the cluster map to locate their nearest neighbours. In this way, the test HSQC spectra were matched with other structurally similar compounds (*e.g.*, from the same compound family or by possessing a high Tanimoto similarity score).

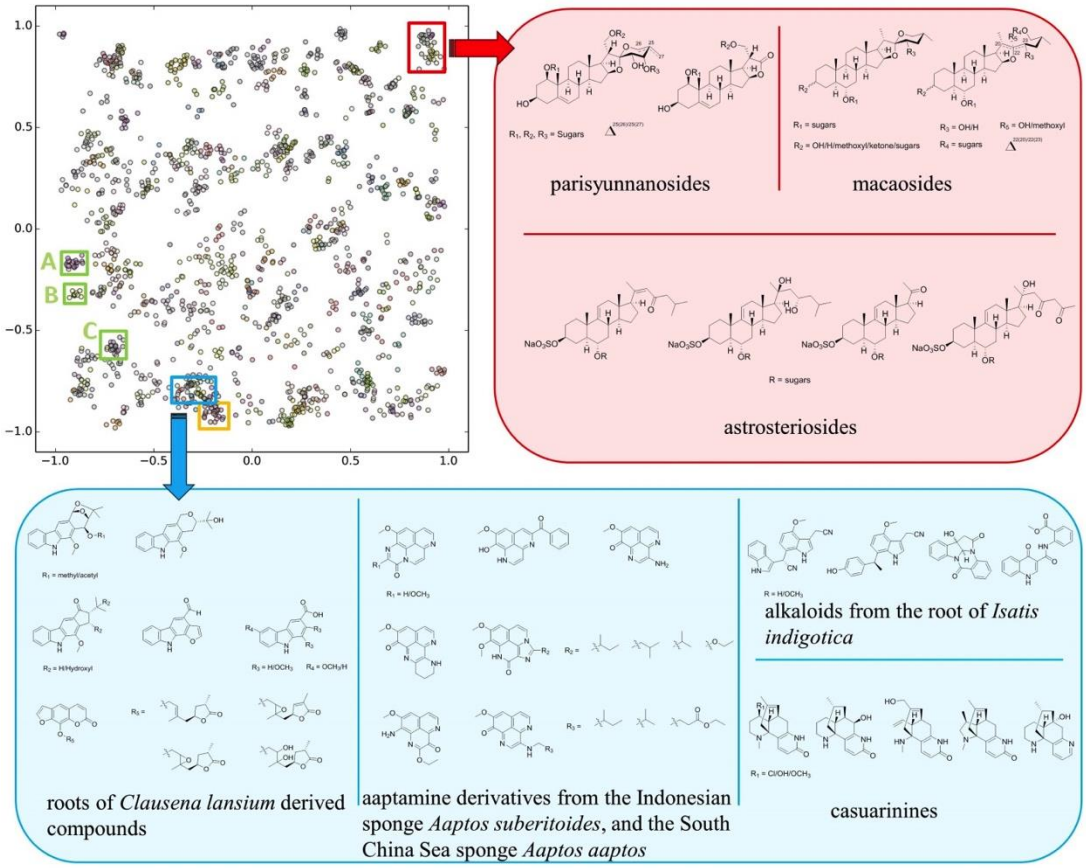


Figure 3.4. The SMART cluster map based on training result of 2,054 HSQC spectra over 83,000 iterations, with inset boxes representing different compound classes discussed in the text.

To produce visually comparable results, the outputs of both the training and the test sets in SMART were visualized in a two dimensional map (Figure 3.4). Each node represents an HSQC spectrum processed by SMART. The node colours designate compounds originating from different research articles (e.g. usually different compound families). When available, the node labels are the compound names; otherwise, the labels are for the organism from which the compound derives. Here the dimension of embedding refers to the dimensions of the cluster space into which the siamese network maps the compounds. For example, if the siamese network had two outputs, we would

be embedding the compounds into 2D. However, we have found that this is too restrictive, and does not perform well. Rather, in preliminary work we found that 10 dimensions provides optimal accuracy and precision-recall performance. Our illustrations in Figure 3.4 are constructed by taking the 10D output of the network and applying PCA to map the 10D cluster space into 2D for illustration purposes. To evaluate the training algorithm, a smaller dataset containing 400 HSQC spectra was first mapped into node clusters with 4,800 training iterations (Figure 3.6.1 and 3.6.2 for the cluster map with analysis), and subsequently, we trained on a larger dataset of 2,054 for a total of 83,000 iterations. The tight structural similarity between the compounds and their locations in the cluster map is evident (Figure 3.2).

Structurally similar NPs were found to form distinct clusters in the map. Three clusters are discussed here to highlight this distinct clustering of different molecular entities, one for a terpenoid family, and two for aromatic alkaloid groups (Figure 3.4). A cluster comprised of 40 nodes (red box, Figure 3.4) was found to contain three saponin variants together with other corresponding triterpenoids. The three saponin variants, parisyunnanosides[50], macaosides[51], and astrosteriosides[52], are of different geographic origins and are produced by organisms from different biological orders. The parisyunnanosides were isolated from the rhizomes of the terrestrial plant *Paris polyphylla Smith var. yunnanensis* originating in Lijiang, Yunnan Province, mainland China. The macaosides were obtained from the aerial parts of the terrestrial plant *Solanum macaonense* collected in Kaohsiung, Taiwan. Finally, the astrosteriosides were isolated from the marine starfish, *Astropecten monacanthus* found around Cát Bà island, Haiphong, Vietnam. The parisyunnanosides have been reported to be toxic to leukaemia

cells[50] whereas the macaosides and astrosteriosides have been found to be anti-inflammatory[51,52]. A second cluster consisting of 42 nodes (blue box, Figure 3.4) was comprised of poly-heterocyclic aromatic alkaloids. Within this cluster there are four major molecular families (Table 3.6.1) with the heterocyclic components being a pyrrole, imidazole, pyridine, or pyrazine, or a combination of these. Notably, several congeners of aaptamine, isolated from two varieties of *Aaptos* species collected in different geographic locations, are found in this cluster. A third cluster was composed of phenolic amides known as the teuvissides[53] (orange box, Figure 3.4); these latter compounds are reported to possess anti-hyperglycaemic properties and were isolated from *Teucrium viscidum* collected in Fujian Province, mainland China. The above discussion highlights the alternate basis for compound clustering by SMART as compared with geographical, pharmacological or source organism methods.

To explore the significance of cluster-to-cluster distance in the clustering map, we evaluated the types of structures present in three clusters that were well defined and in varying proximity to one another (green boxes A, B and C of Figure 3.4). Cluster A was composed of oxidized steroids of highly similar structure to one another from the plants *Aphanamixis polystachya*[54] and *A. grandifolia*[55], whereas nearby cluster B was formed from a series of triterpene glycosides[56]. The more distant cluster C contained several diterpenoids[57]. Visually, it seems generally correct that oxidized steroids are more similar to triterpenes than they are to diterpenoids. In comparison, the averaged 2D Tanimoto score[58] (a distance measure based on planar structures of compounds) between compounds in the cluster A and B, $T_{AB} = 44$, slightly exceeded the value $T_{AC} = 43$ between compounds in the cluster A and C (Figure 3.6.3 for molecular structures),

which indicates that the deep CNN method is better at quantifying and representing structural differences among compound subfamilies than the algorithm used to generate 2D Tanimoto scores. The average intra-cluster Tanimoto score of the cluster containing uralsaponins A, B, C, F, M, T, V, W, X and Y is 96.3 whereas the cluster containing aphanamixoids C, D, E, F and G is 95.7. The average intra-cluster Tanimoto score of the cluster containing ebractenoids A, B C, D, E, F, G, H, I and J is 69.4. All of these intra-cluster Tanimoto scores are higher than the inter-cluster Tanimoto score $T_{AB} = 44$ or $T_{AC} = 43$. Therefore, it is apparent that the SMART clustering map not only recognizes closely similar compounds, but also appropriately places clusters of different compounds in their proper context relative to one another.

3.2.3. Related Work

Again, the aforementioned grid-cell approaches[28] are similar to ours in that the shifted grid positions can be thought of as corresponding to the first layer of convolutions, which have small receptive fields (like grid cells), and they are shifted across the input space like shifted grids. Also, our approach uses layers of convolutions that can capture multi-scale similarities. The grid-cell approaches, however, use hand-designed features (i.e. counts of peaks within each grid cell), and the similarities are computed by simple distance measures. This is not enough data to train the deep network. In particular, PLSI and LSR are linear techniques applied to hand-designed features. Furthermore, other representations, for example the 'tree-based' method[59], also rely on data structures designed by the researcher. Our approach, using deep networks and gradient descent, allows higher-level and nonlinear features to be learned in the

service of the task. This approach is similar to modern approaches for computer vision, which since 2012 has shifted away from hand-designed features to deep networks and learned features, and has led to orders of magnitude better performance. Similarly to how deep networks applied to computer vision tasks have learned to deal with common problems, such as recognizing objects and faces in different lighting conditions and poses, our CNN pattern recognition-based method can overcome solvent effects, instrumental artefacts, and weak signal issues.

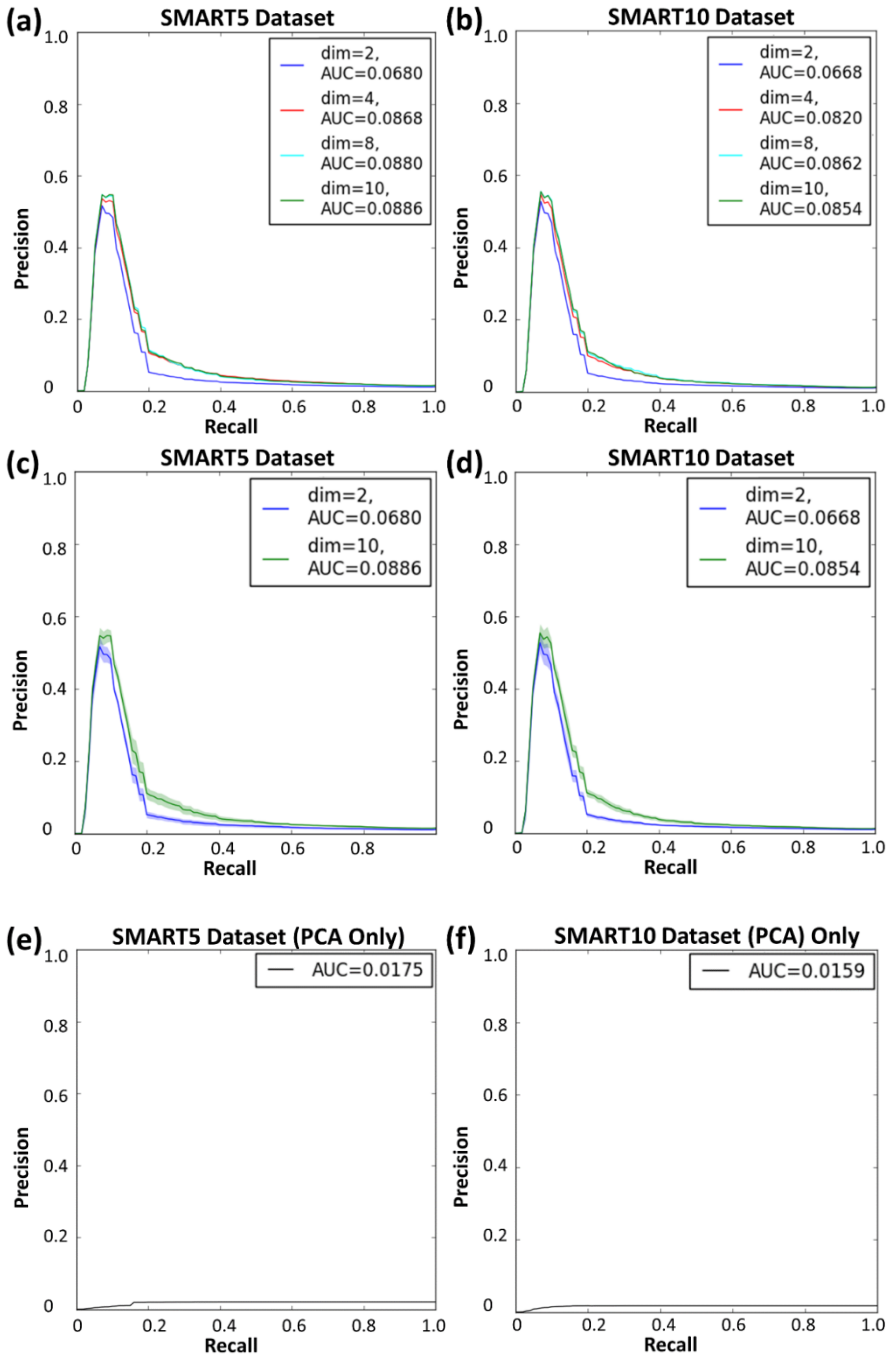


Figure 3.5. Precision-recall curves measured across 10-fold validation for different dimensions (dim) of embeddings. (a) and (b) Mean precision-recall curves on test HSQC spectra for SMART5 and SMART10 datasets, respectively. (c) and (d) Mean precision-recall with error curves (grey) for SMART5 and SMART10, respectively. (e) and (f) Mean precision-recall curves for SMART5 and SMART10 clustered by Principal Component Analysis (PCA) without use of the CNN. AUC: area under the curve

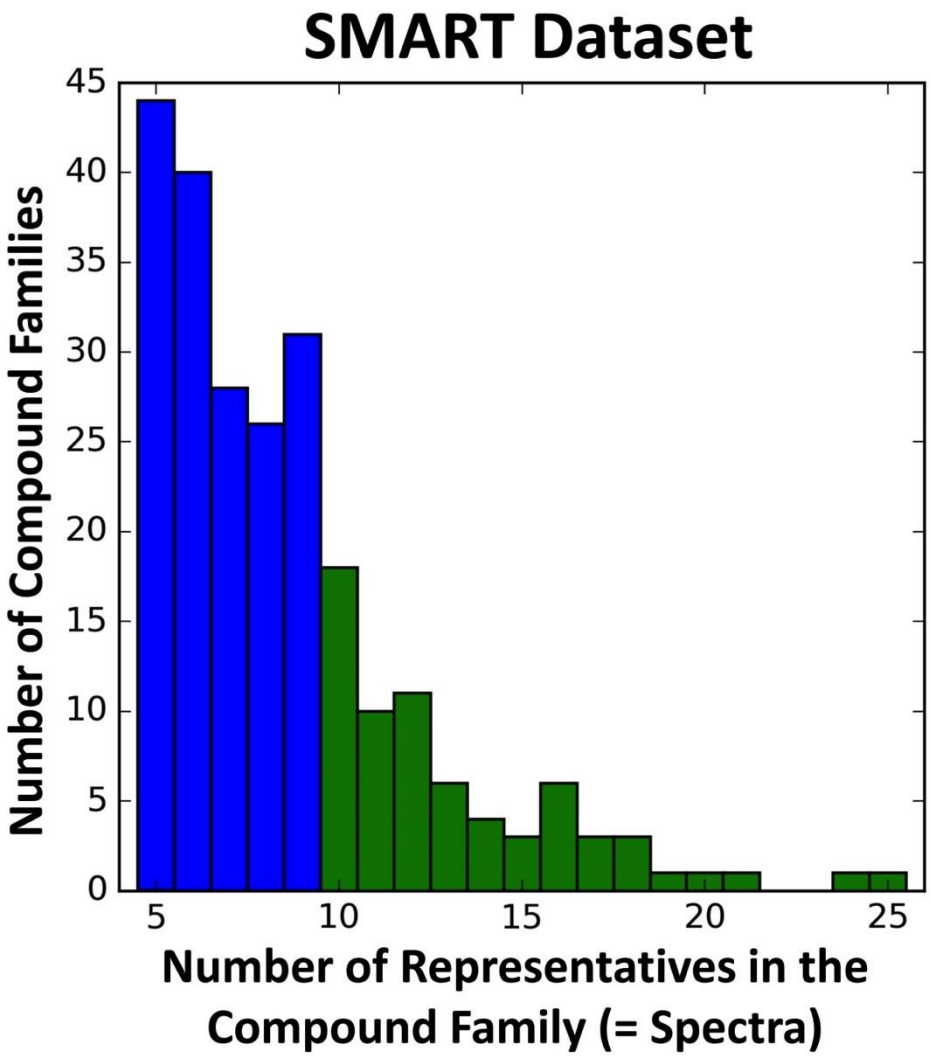


Figure 3.6. Distribution in the Training Dataset of Numbers of Families Containing Different Numbers of Individual Compounds. (a) The SMART5 training set contains 238 compound subfamilies, giving rise to 2,054 HSQC spectra in total. (Blue and Green) (b) The SMART10 training set contains 69 compound subfamilies and is composed of 911 HSQC spectra in total. (Green only)

It is difficult to directly compare Wolfram et al.'s results to ours because they used a much smaller dataset (132 compounds) from 10 well-separated families. This is

not enough data to train the deep network. To further compare our approach with other NMR pattern recognition approaches, we generated precision-recall curves (Figure 3.5) using SMART trained with the SMART5 and SMART10 databases (Figure 3.6). Considering SMART as a search engine, precision recall curves help evaluate the SMART's performance to find the most relevant chemical structures, while taking into account the non-relevant compounds that are retrieved. In our approach to HSQC spectra recognition/retrieval, precision is a measure of the percentage of correct compounds over the total number retrieved, while recall is the percentage of the total number of relevant compounds. Therefore, higher precision indicates a lower false positive rate, and higher recall indicates a lower false negative rate. The precision-recall curves of our approach show high precision peaks at low recall rates, suggesting that SMART retrieves at least some relevant structures in the first 10%-20% of compounds retrieved, and thus indicates that SMART returns accurate chemical structures. To compare this to a linear embedding, we performed PCA on the SMART5 and SMART10 databases separately. The precision recall curves of those PCA results are much worse than those processed by the CNN.

3.2.4. SMART recognition of noisy HSQC spectra

Because white Gaussian noise is often seen in experimental HSQC spectra, we investigated the robustness of the SMART to recognize HSQC spectra in the presence of significant noise. By adding noise to HSQC spectra in the SMART10 database and measuring the Euclidean distance of those noisy spectra to their original ones, we were able to observe that as noise intensity increases so does the distance increase from the

original location in the 2D cluster map. However, the noisy spectra were still effectively recognized as being closely related to their original compounds (Figure 3.7 and Appendix).

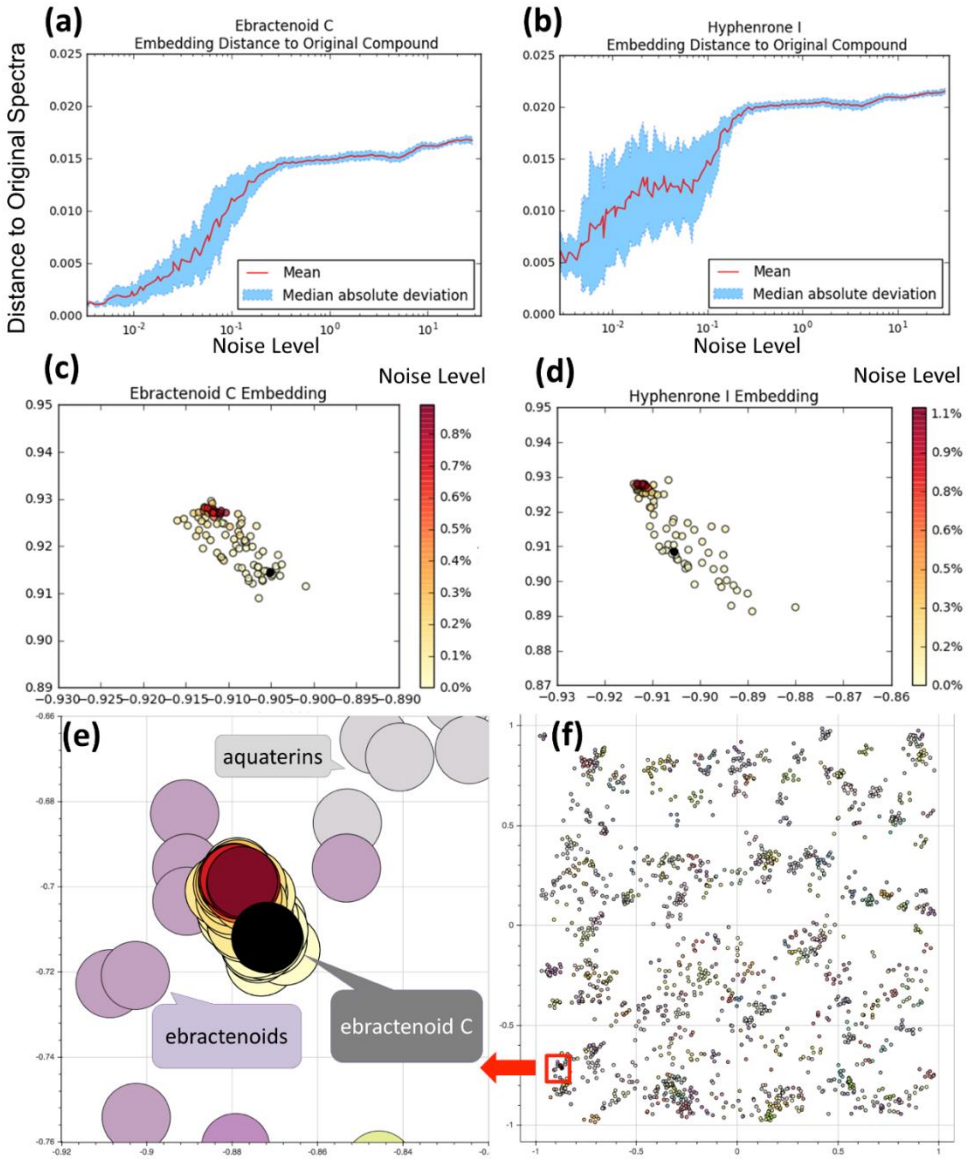


Figure 3.7. Distance of the noisy spectra measured against the original spectra of ebracenoid C and hyphenrone I. The distance measure in the y axis of the ebractenoid plot (a) and hyphenrone plot (b) is the same as the cluster map in Figures 4 and 7 (f). The noise level is defined by dividing pixels altered over the total number of pixels of the HSQC spectra. The results visualized in the 2D cluster maps with each node representing one noisy spectra, and with node color intensity as a function of the noise level, for the ebractenoids (c) and hyphenrones (d). The original image without added noise is shown as the black node in these 2D cluster maps. We then embedded the nodes for the ebractenoids in (c) to a global view of the 2D cluster map in (f), and zoom in on the red box in (f) as shown in (e). Note, larger node sizes are used to depict compounds in (e) versus (c).

3.2.5. SMART characterization of Viequeamides of NRPS origin

As a practical example of the functional use of the SMART workflow to discover new NPs, we used it to rapidly characterize a group of unknown marine cyclic depsipeptides from two marine cyanobacteria: 1) *Rivularia sp.* collected in Vieques, Puerto Rico and 2) *Moorea sp.* collected in American Samoa. These compounds were isolated in the course of our ongoing efforts to discover marine cyanobacterial NPs with anti-cancer properties[60]. Metabolites from these two collections were purified by high performance liquid chromatography (HPLC), and then ^{1}H-^{13}C HSQC data were collected with 100% sampling density, but using the NUS pulse sequence in the indirect dimension for all six purified compounds. Data reconstruction as described above for the six samples yielded HSQC spectra, and these were subjected to the SMART workflow to embed them in the cluster map. We found that the six nodes clustered with nodes for the previously characterized viequeamides A (**1**) and viequeamides B (**4**). After an analysis of various 2D NMR spectra, and MS, IR and UV data, the planar structures of the four new compounds were determined (Figure 3.1, compounds **2**, **3**, **5**, **6**). The absolute configurations of these compounds were then elucidated by Marfey's analysis and/or X-ray crystallography, completing their structure determination. Evaluation of the toxicity of the pure compounds to H-460 human lung cancer cells revealed that two possessed relatively potent cancer cell toxicity properties; viequeamide A2 (**2**) had an $IC_{50} = 0.62 \pm 0.046$ μM and viequeamide A3 (**3**) had an $IC_{50} = 1.98 \pm 0.038$ μM. Viequeamides B (**4**), C (**5**) and D (**6**) showed no appreciable H-460 cytotoxicity. (See Chapter 4 for details)

3.3 Conclusion

SMART is the first combination of NUS 2D NMR and deep CNNs. This tool streamlines dereplication and determination of natural product families from multiple organisms and facilitates their isolation and structural elucidation. While compound families represented the metadata associated with HSQC spectra in this study, it is very possible to associate and integrate biological, pharmacological and ecological data with SMART, and thereby create new tools for enhanced discovery and development of biological active NPs as well as other small molecules. Ultimately, this leads to an increased appreciation for the structural diversity and therapeutic potential of natural products.

By both quantitative and qualitative inspection of SMARTs cluster space, the following properties were suggested by the results: 1) the distance between nearby nodes of a clustering map is a measure of the structural similarity between compounds that share molecular moieties (*e.g.* functional groups, carbon skeletons, etc.), 2) chimeric compounds with structural features comprised of two independent families of compounds reside near or in between the component clusters (for example, saponins are located near and between other glycosides and terpenoids, in Figure 3.6.2), 3) this accuracy of placement of new compounds in SMART should be enhanced as the size of the training set grows, 4) as the size of the training set increases for a given compound class, the accuracy of placement of a new test compound in that family improves, 5) even in the presence of random spectral noise, spectra are strongly associated to their structural chemical analogues. Nevertheless, the accuracy of recognition correlating to the signal-to-noise ratio of HSQC spectra remains to be determined, as does the impact

of solvent effects on chemical shifts or extraneous peaks appearing in the spectrum from electronic sources or impurities. But as more compounds are added to the training set, the SMART system will naturally improve in accuracy and robustness, thereby accelerating natural product structural elucidation and thus drug discovery.

SMART has an immediate value in NP drug discovery efforts by providing rapid and automatic compound dereplication and assignment to molecular structure families. With further refinement of the SMART workflow, such as training for spectra of the same compound with different S/N ratios, deeper understanding of other parameters that impact spectral recognition, combining with other fast NMR techniques, SMART has the potential to enhance NPR and enable new directions of experimentation at the chemistry-biology interface.

3.4 Experimental Methods

3.4.1. Training Set Collection and Processing

The dataset of HSQC spectra was compiled from available online sources. We removed spectra that showed no peaks (*i.e.*, the spectra in the publication appeared blank). We collected all usable ^{1}H-^{13}C HSQC spectra (4,105 in total), including a few cases of the same compound in different deuterated solvents, from the supporting information of the *Journal of Natural Products*, years 2011, 2012, 2013, 2014 and 2015. In addition, the HSQC spectra of somocystinamide A[61] and swinholide A[62] in the supporting information of *Organic Letters* were also included in the dataset. Around 2,056 spectra were removed from this series, because their molecular class had less than 5 HSQC spectra. All spectra were collected and initially processed by the following

steps: (1) The HSQC spectra were saved as png format grayscale images at a resolution of 512 × 512 pixels (the minimum resolution in the proton dimension is 51.2 pixels per ppm and in the ^{13}C dimension it is 2.8 pixels per ppm.); (2) lines surrounding spectral edges, annotations, chemical structures, and other annotations were deleted using Photoshop such that only the HSQC signals and noise were present in the images; (3) images were rotated and/or flipped when necessary to ensure that the horizontal dimension was the direct ^{1}H dimension with chemical shifts increasing from right to left, and the vertical dimension was the indirect ^{13}C dimension with chemical shifts increasing from top to bottom; (4) images were uniformly converted into black (signal and noise) and white (spectral background); (5) images from the same publication were pooled and labelled as the same training class, as all of the publications we utilized reported compounds from a single family; (6) a cross shaped 3 × 3 median filter[63] was applied to remove unwanted salt-and-pepper noise; however, no other enhancements were applied (Figure 3.6.4 for an example of spectra input preparation). Essentially, in this study, the relevant quantity for measuring similarity was the positions and shapes of the various peaks with respect to one another, and not their overall location.

Figure 3.6 shows the distribution of spectra number within each compound family in the complete dataset. From Figure 3.6, we observe that the dataset has a skewed distribution of images per class. Hence, in order to make the training stable and comparison fair, we created two different datasets: SMART5 and SMART10, containing all spectra of compound families (*e.g.* veraguamides[64], ebractenoids[57], naphthomycins[65], viequeamides, etc.) with at least 5 and 10 HSQC spectra, respectively, per family. In total we have 238 categories (2,054 spectra) for SMART5, the largest

having 25 and the lowest having 5 spectra per compound family. Further restricting the data to contain at least 10 spectra per molecular class results in only 69 categories (911 spectra) in SMART10, which we found to be too few for effective training. Hence, all of our experiments used SMART5.

When training the neural network (see below for description), we used a 10-fold cross-validation scheme, randomly shuffling the dataset and then splitting it into training, validation, and test sets in proportions 8:1:1. We repeated the procedure 10 times such that all images became part of a test set one time. The results we report here were averaged across these ten networks

3.4.2. NUS 2D NMR Data Generation

In order to generate an independent test set, we developed an optimized NUS pulse sequence using an NMR standard (strychnine, 50 nmole TCI America, Catalog No. S0249). This optimized method was then applied to several newly isolated NPs (e.g. the viequeamides). The viequeamides were isolated from two different marine cyanobacteria; *Rivularia sp.* collected in Vieques, Puerto Rico[60] and *Moorea sp.* collected in American Samoa. Detailed isolation and structural elucidation of these compounds will be published separately. The 2D NMR spectra were recorded on a 600 MHz Bruker Avance III spectrometer with a 1.7 mm Bruker TXI MicroCryoProbe™ using TopSpin 2.1. The solvent $CDCl_3$ contained 0.03% v/v trimethylsilane (δ_H 0.0 and δ_C 77.16 as internal standards using trimethylsilane and $CDCl_3$, respectively). All spectra were recorded with the sample temperature at 298 $^{\circ}$K.

The data shown in Figure 3.2 were acquired using the NUS edited hsqcedetgpsisp2.3 HSQC pulse sequence. Data were acquired as 4096 × 32 points (32 out of 128 t_1 increments, 25% NUS) in direct and indirect dimensions, respectively, giving a total acquisition time of a quarter of its conventional counterpart. Spectral windows in direct and indirect dimensions were 7183.9 and 24118.9 Hz respectively. Data in both Figure 3.2 (b) and (c) were processed using NMRPipe[66] by applying zero filling (round final size to power of 2) in both dimensions. Spectra in Figure 3.2 (b) were processed by applying IST as implemented in hmsIST[12] with 400 iterations followed by forward-backward LP sequentially, while spectra in Figure 3.2 (c) were processed by applying IST with 400 iterations followed by MEM with the standard deviation of time-domain noise set to 200. The viequeamides spectra were acquired and processed the same way as Figure 3.2 (c), except that the indirect dimension was sampled with 100% NUS (256 out of 256 t_1 increments).

3.4.3. The Deep Siamese Network

As depicted in Table 3.1, the overall deep CNN siamese architecture used in this study is similar to AlexNet[42], and consists of 8 layers comprised of 4 convolutional layers followed by 4 fully connected layers. This network is used as the two "twins" in the siamese network. The output layer contains vectors in $\mathbb{R}^K$. Here, K is the embedding dimension. The energy loss function defined in equation 2 (below) is applied to the outputs of the embedding layer (layer 8). We ran several experiments to find the best K and measured the accuracy on the validation set. Empirically, for the given dataset, $K = 10$ gave us the best results.

Table 3.1. The Architecture of the Deep CNN Used in This Study[a]

Layer Number	Layer Type	Number of Filters (Stride 1)	Size	Additional Information
1.	convolutional	8	4 × 4	maxpool 4 × 4 stride 2
2.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
3.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
4.	convolutional	16	4 × 4	maxpool 4 × 4 stride 2
5.	fully connected	-	128	dropout 0.5
6.	fully connected	-	128	dropout 0.5
7.	fully connected	-	128	dropout 0.5
8.	fully connected	-	*K*	*K*-dimensional embedding layer

[a] the dimensionality of the input data is 512 × 512

3.4.4. Loss Function

Siamese networks are trained with an energy function that is minimized by gradient descent. The design of the energy function determines the way in which pairs of items are pushed together or pulled apart. There are at least two such functions that have been used[30] in the literature; here, we used a modified version of the spring model developed by Hadsell *et al.*[41]. The energy function is described with the following notation; for example *i*, the input vector is represented as x_i, and the output label as y_i. The output label is defined as a "one hot" vector, where if there are *k* categories, y_i is a *k*-dimensional binary vector, and if the category is *c*, y_i is 1 at the c^{th} position and 0 everywhere else. Meanwhile, x_i, the input HSQC spectra, is treated as a vector.

We treat our neural network as a function G_w, where *W* is the weights of the network. Then the output of the neural network is $G_w(x)$. $G_w(x)$ is a vector of dimension *K*, a hyperparameter of the system. We then define the distance function *d* between images x_i and x_j:

$$\boldsymbol{d(x_i , x_j) = \| G_w(x_i) - G_w(x_j) \|} \qquad \text{(Equation } \mathbf{1}\text{)}$$

where $\|\cdot\|$ is the Euclidean distance function.

Now we can define the energy function L to be minimized as[41]:

$$L(x_i, x_j) = \begin{cases} \frac{1}{2}\max(0, d(x_i, x_j) - m)^2, & \text{if } y_i = y_j \\ \frac{1}{2}\max\left(0, m - d(x_i, x_j)\right)^2, & \text{otherwise} \end{cases} \quad \text{(Equation 2)}$$

where m is a hyperparameter that defines a margin. In this case, if y_i and y_j are the same category and the squared distance between the output representations of x_i and x_j is more than a margin, then this distance is minimized, otherwise it is unchanged. If they are different, then we should increase the distance between them up to the margin m. Once they are pushed at least m distance apart, and the loss becomes 0. This loss function penalizes large distances between pairs of outputs for images in the same category (first line), but for outputs from different categories, a penalty is assigned only if they are within m units. This loss function ensures that the output space forms well-behaved clusters during training. The difference between this objective function and the one used in Hadsell *et al*[41] is that no margin was used within the same category. Empirically, we find this objective function gives superior results.

3.4.5. Training Details of the Siamese Network

We implemented our system using the Theano[67] and Lasagne (http://tinyurl.com/hl9dy9y) deep network packages. The siamese network was trained using mini-batch stochastic gradient descent with the Adagrad[45] update rule, following the protocol introduced by Hadsell *et al*[41]. We applied the packages. Specifically, 50% of each mini-batch has negative samples ((x_i,x_j,y_i,y_j) s.t. $(y_i \neq y_j)$), and 50% has positive samples ((x_i,x_j,y_i,y_j) s.t. $(y_i = y_j)$). The Adagrad update rule tunes the step size

automatically in real time, making learning stable in later iterations. We used hyperbolic tangent as the activation function for all layers including the output layer. The weights were initialized using Xavier initialization[68]. The initial learning rate was $\alpha = 0.001$, and the mini-batch size was 256. We applied dropout regularization[69] on layers 5, 6, and 7 of the network, and batch normalization[49]. We found that applying batch normalization speeds convergence by a factor of 7. The total number of parameters in the network is 399,102, considering that the number of parameters triples when batch normalization is applied. We used Amazon EC2 instances to run our experiments.

We recorded precision-recall curves (Figure 3.5) of SMART's performance by randomly selecting HSQC spectra from the test dataset and retrieving known compounds according to their distance to the test compound in the cluster map. In this regard, precision was calculated by dividing the number of true positives over the combination of the number of true positives and the number of false positives. Recall was calculated as the number of true positives over the combination of the number of true positives plus the number of false negatives. At each level of recall, there is a different level of precision. The area under the precision recall curve (AUC) is then a standard measure of performance (larger is better). In our case, for each compound in the test dataset, we measured a precision recall curve by calculating precision (the number of retrieved compounds that are relevant) and recall (the number of relevant compounds that are retrieved) of the retrieved HSQC spectra from the training dataset within an expanding hypersphere centred at the compound in the test dataset. These final precision recall curves were averaged over the test dataset. The CNNs that we used in this regard were trained for 10,000 iterations on the SMART5 and SMART10 datasets

with 10-fold cross validation for embedding dimensions k = 2, 4, 8, and 10 (Figure 3.5). To compare our results to a linear embedding, we separately performed PCA on the SMART5 and SMART10 databases. Specifically, we embedded the PCA results in high dimensional Euclidean space (k = 10, chosen to match the number of dimensions used in the CNN training). The precision recall curves of the randomized results are also shown in Figure 3.5.

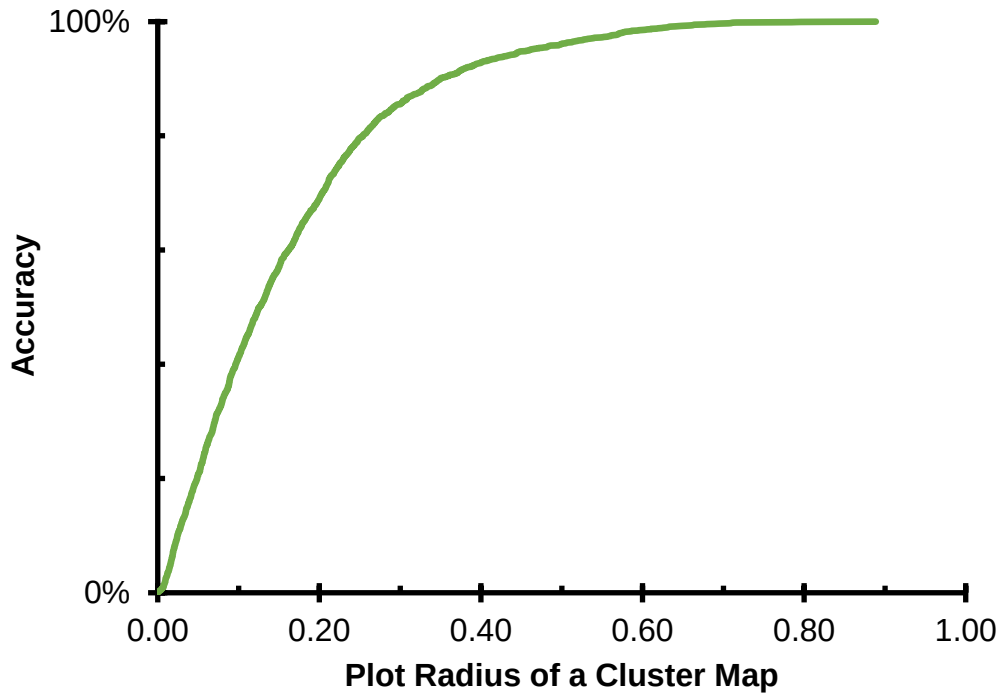


Figure 3.8. Plot of the Accuracy of SMART as the radius around a project point increases. This figure shows the fraction of correct families captured by a hypersphere of the given radius around each node in the cluster map. The distances between nodes in the cluster map has no physical meaning, but is a quantification of HSQC spectral similarity. SMART can achieve good accuracy (proper placement in the map of a new compound to its correct compound family) within 0.5 radius of a 2-dimensional cluster map, and even better for a 10-dimensional map.

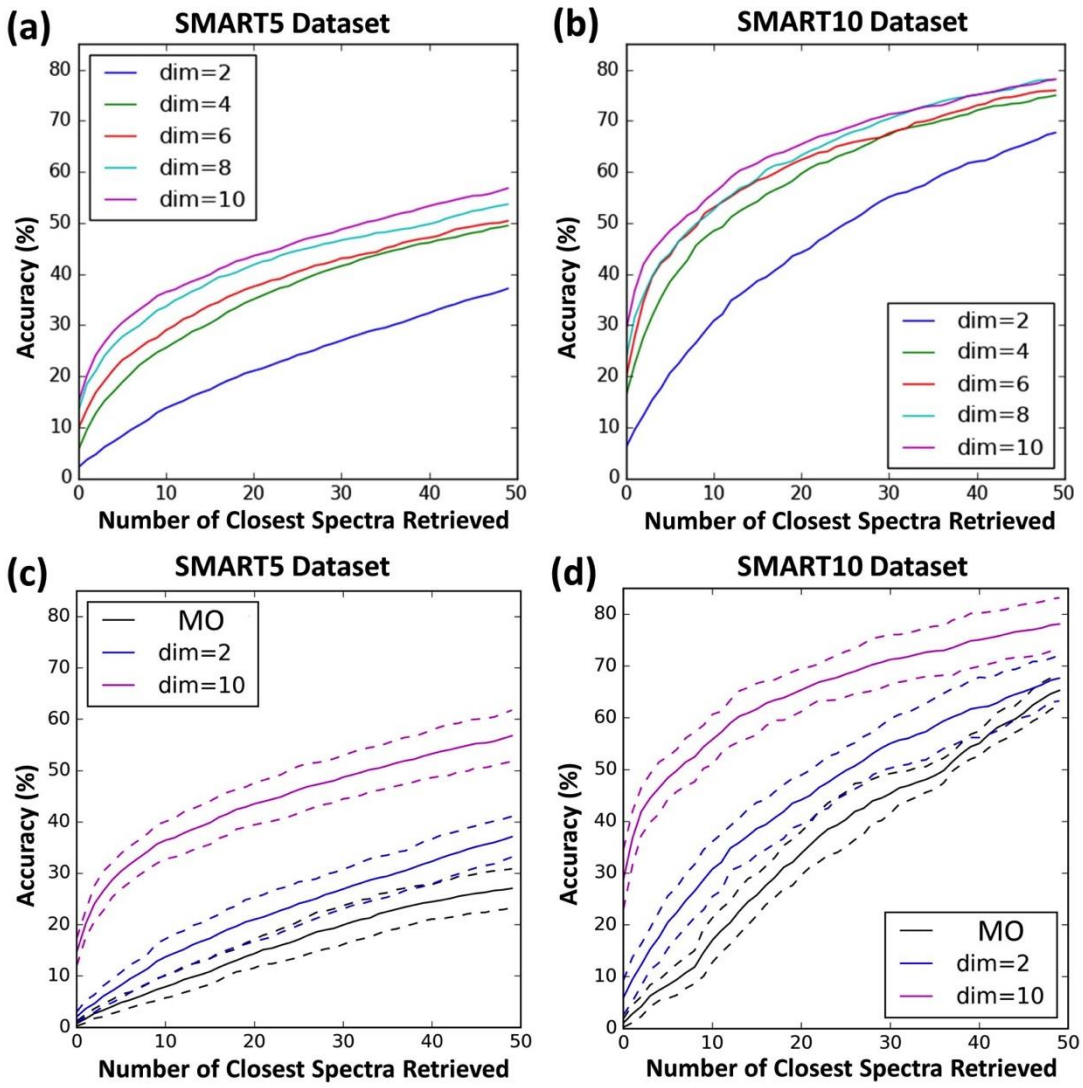


Figure 3.9. Closest retrieval curves measured across 10-fold validation for different dimensions (dim) of embeddings. For (a) and (b), mean closest retrieval curves on test sets for SMART5 and SMART10 datasets, respectively. For (c) and (d), mean closest retrieval curves with error curves ($\mu \pm \sigma$, dashed lines) for SMART5 and SMART10, respectively. In (c) and (d), the black plot (MO, most frequently occurring) is a baseline prediction of random compound associations on the basis of the number of members in a compound family. Specifically, the category with the most members is picked as the first compound association, the second most members as the second one, etc. This order is the same irrespective of the compound being considered.

We also used 10-fold cross validation to estimate performance (Figure 3.8 and 3.9). Specifically, a different 10% of the training set was held out as a test set 10 times,

and the results were averaged to report performance. For each fold of the cross validation, we held out 10% of the data for early stopping. In this way, all of our HSQC spectra were used for testing. Here, the complete split was 8:1:1, training:validation:test. The iterations stop at the point in training where the error on the hold-out set is minimized. Here, the error was a measure of average precision on the hold-out set.

We employed the Tensorflow package (https://tinyurl.com/y9lz45sa) to visualize the features that were learnt by the first layers of the CNN. The results of the first convolutional layer are shown in Figure 3.3.

3.4.6. Validation of the Model on "Novel" Categories

To evaluate whether the system performs properly with new categories of molecules, we performed the following three experiments. In SMART5, we removed the HSQC spectra of three categories of compounds (ebractenoids, naphthomycins, and veraguamides) from each of three common NP families (terpenoids, polyketides, and peptides, respectively), for each experiment, and used those removed spectra as a test set. During training, each subfamily was given a different label, however, this information was only provided to the training algorithm in the sense of "same/different category" as in Equation 2. This experiment thus tested whether a subfamily of terpenoids that was unfamiliar to the network would be mapped close to the other terpenoids. For example, there are 10 compounds in the terpenoid subfamily of ebractenoids that were not used during training. During testing, they were presented to the network, and their distance to the other terpenoids measured. This experiment was repeated for the naphthomycins, and veraguamides, and their location in the embedding

space was evaluated for whether they were properly mapped to their respective families (*e.g.* polyketides and peptides, respectively). This experiment revealed that the ebractenoids clustered with the terpenes and terpenoids in the 10-dimensional space (Table 3.6.2). Similarly, the naphthomycins and veraguamides were subjected to a similar experiment (Table 3.6.3 and Table 3.6.4) and confirmed that SMART was able to properly place compounds to which it was naive.

Finally, we trained the siamese network using all of SMART5, supplemented with HSQC spectra for viequeamide A (**1**) and B (**4**) (2 spectra), parguerene, precarriebowmide, palmyrolide and three isomers (4 spectra), somocystinamide and a derivative (2 spectra), and columbamides A, B and C (3 spectra). This was exposed to the six newly collected HSQC spectra [subsequently identified as the viequeamides, e.g. the two known viequeamides A (**1**) and B (**4**) and four new viequeamides A2 (**2**), A3 (**3**), C (**5**) and D (**6**)] using the 100% NUS sampling method. Training was stopped after a fixed number of iterations. The 10-dimensional output of this test is presented in the Chapter 3 Appendix (Table 3.6.5).

3.4.7. Tanimoto Score Calculation

Averaged Tanimoto Score for compounds between the three clusters in Figure 3.1 was calculated using the PubChem Score Matrix Service[70].

3.4.8. Recognition of noisy HSQC spectra

Using Matlab 2013, we created a 2D matrix of white Gaussian noise to simulate the noise in real-time measurements. Next, we applied 2D Fast Fourier Transform (FFT) to this 2D noise matrix. The transformed FFT results for these noisy spectra were sized

to match those of two randomly selected compounds (hyphenrone I and ebractenoid C) from the SMART10 database[57,71]. We also calculated the noise intensity in the spectra by dividing the number of noisy pixels by the total number of pixels. The noise matrix was then added to the two HSQC spectra, and the intensity of the noise was then increased consecutively in a finite arithmetic progression of 140 steps, rendering 140 noisy spectra for each compound. In addition, at each noise level, the white noise was again randomized 100 times, rendering a total of 14,000 noisy spectra. These noisy HSQC data were then processed by the convolutional neural networks pre-trained with SMART10 for over 10,000 iterations. The results are shown as two distance vs. noise plots in Figure 3.7 (a) and (b). The distance measure displayed in the vertical axis of these two plots was in the same units as the cluster map in Figure 3.4. The results were also visualized in 2D cluster maps with each node representing one noisy spectrum, with the intensity of the node color representing the noise level (Figure 3.7 (c) and (d)). The original image without added noise is shown as the black node in those 2D cluster maps. In order to further visualize the internode distance between nodes that represent noisy spectra and those that represent our training dataset, we embedded the nodes of the noisy spectra in Figure 3.7 (c) in a global view of the 2D cluster map shown in Figure 3.7 (f), and provided a zoomed-in view of the ebractenoids clusters in Figure 3.7 (e). Figure 3.7 (e) shows that noisy HSQC spectra are clustered close to their original spectrum, and thus, noise to the levels we have evaluated, has little effect on the ability of SMART to accurately place compounds into their appropriate location (ebractenoids in this case). Selected noise maps are provided in the Appendix."

3.5 Chapter 3 Acknowledgements

We thank Drs. Anthony Mrse and Xi Liu for their technical support, Dr. Preston B. Landon, Jianping Zhao, Sanjeev Rao, Yufei Wang, Zheng Long, Xin Xu, and David Glukhov for useful discussions. We also thank Ruslan Tilemisov for donating a Titan X GPU for this study. Finally, we thank Aaron Landon for donating a GEFORCE GTX 1080 GPU for this study.

This work was supported by a UC San Diego Frontiers of Innovation Scholarship (C.Z., G.W.C and W.H.G.), National Institute of Health GM107550 (W.H.G.), UC San Diego Graduate Student Growth & Excellence Initiative (C.Z.), The Bolashak International Scholarship (Y.I.), Science and Technology Project of Guangdong Province (2013B021100021) (Y.T.), and the National Science Foundation (SMA 1041755) (G.W.C.).

Chapter 3, is a reprint as it appears in the Scientific Reports. 2017, 7(1), 14243, with the following authors, Chen Zhang, Yerlan Idelbayev, Nicholas Roberts, Yiwen Tao, Yashwanth Nannapaneni, Brendan M. Duggan, Jie Min, Eugene C. Lin, Erik C. Gerwick, Garrison W. Cottrell, and William H. Gerwick. The dissertation author was a primary investigator and first author of this paper.

3.6 Chapter 3 Appendix

Table 3.6.1. Detailed information regarding the compounds in the blue box of Figure 3.4.

Molecular Labels	Collection Information	Bioactivity
roots of *Clausena lansium* derived compounds[73]	collected from Quỳ Hợp District, Nghệ An Province, Vietnam in March 2011	anti-inflammation, inhibition of superoxide anion generation or elastase release,
aaptamine derivatives from the Indonesian sponge *Aaptos suberitoides*[74]	collected by scuba diving in Ambon, Indonesia in October 1996 at the depth of 3 m	cytotoxic activity against the murine lymphoma L5178Y cell line
aaptamine derivatives from the South China Sea Sponge *Aaptos aaptos*[75]	collected off Woody (Yongxing) Island and Seven Connected Islets in the South China Sea in June 2007	cytotoxicities against HL60, K562, MCF-7, KB, HepG2, and HT-29 cells
alkaloids from the root of *Isatis indigotica*[76]	collected from Anhui Province, China in December 2009	antiviral, against influenza virus A/Hanfang/359/95 (H3N2) or inhibition of Coxsackie virus B3 replication
casuarinines[77]	collected from Zhenghe County of Fujian Province, China, in October 2010	neuroprotective effect against hydrogen peroxide (H_2O_2)-induced neuronal cell damage in human neuroblastoma SH-SY5Y cells or inhibition of acetylcholinesterase (AChE)

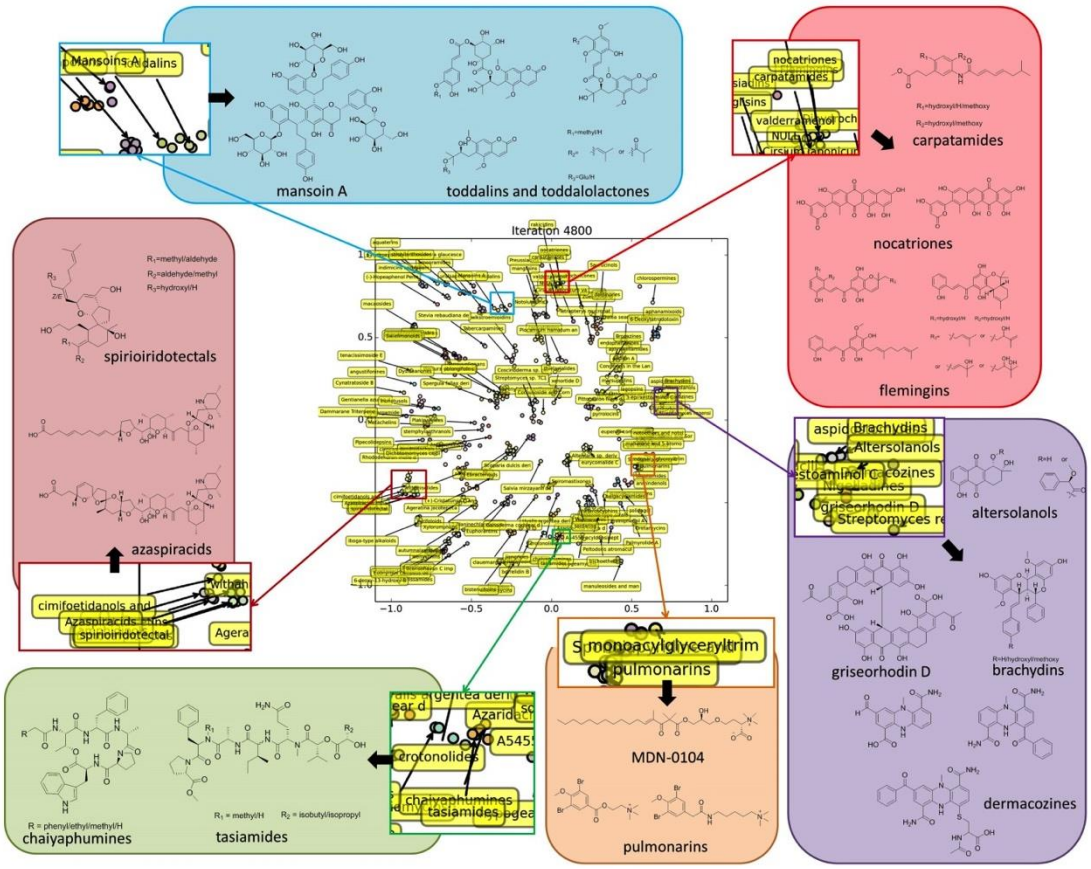


Figure 3.6.1. A cluster map containing 400 compounds after 4,800 training iterations. This smaller cluster map shows a distribution of different families of compounds on this map. The names of the compounds are shown as yellow labels.

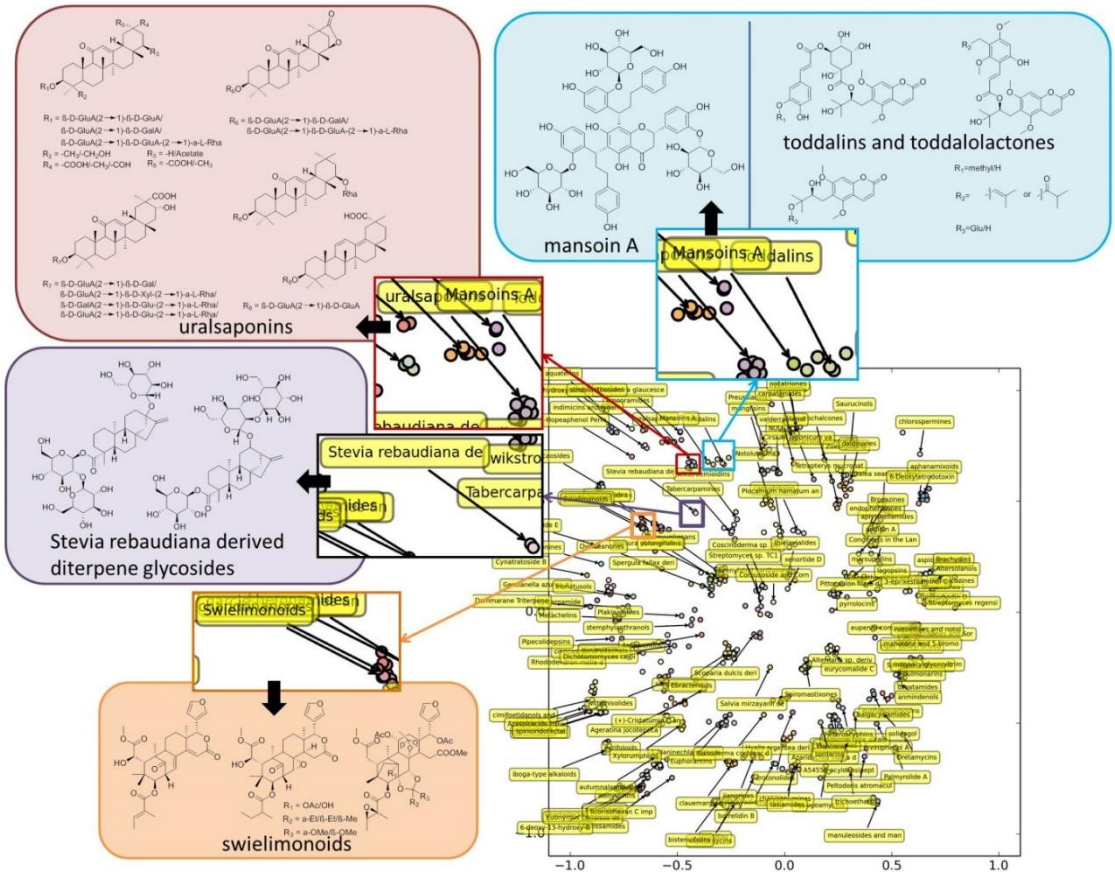


Figure 3.6.2. A cluster map containing 400 compounds after 4,800 training iterations. (Continued) The structural change from one family to another appears to be continuous and evolving. Specifically, we observed some intermediates *i.e.* uralsaponins[78] and *Stevia rebaudiana* derived diterpene glycosides[79] (saponins) between mansion A[80] (glycosides) and swielimonoids[81] (terpenoids).

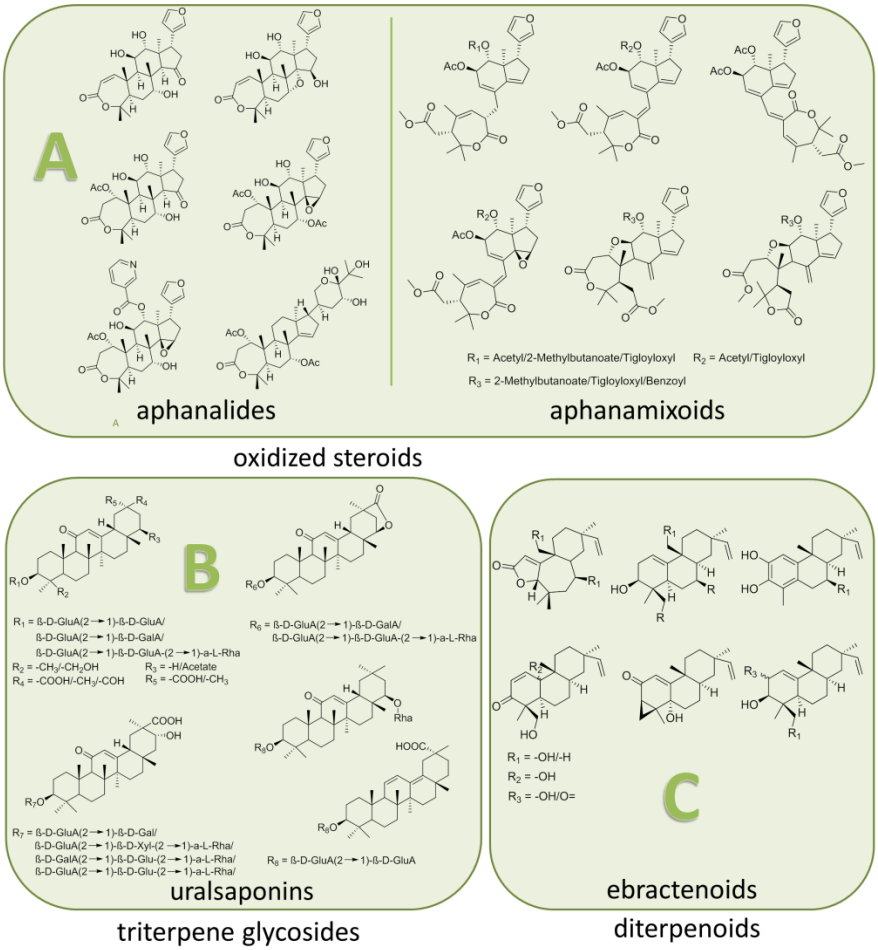


Figure 3.6.3. Molecular structures associating with the HSQC spectra within the three green boxes of Figure 3.4. Cluster A: oxidized steroids from two articles in the Journal of Natural Products, from plants *Aphanamixis polystachya* and *Aphanamixis grandifolia*, respectively. Cluster B: triterpene glycosides isolated from the roots of *Glycyrrhiza uralensis Fisch*. Cluster C: diterpenoids isolated from the roots of *Euphorbia ebracteolata*.

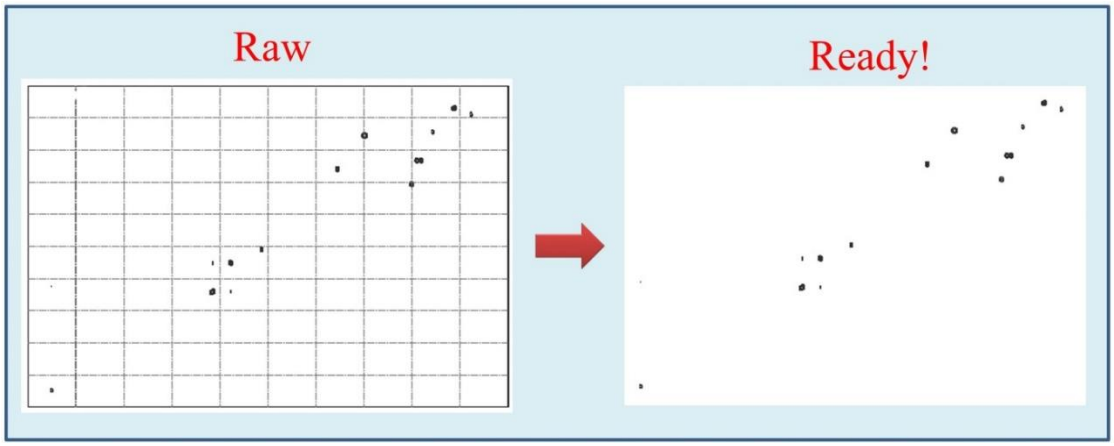


Figure 3.6.4. An example of preparing input HSQC spectra for SMART. Adobe Acrobat was used to convert .PDF spectra into .PNG files. Then Adobe Photoshop was used to remove axis labels, frames, grids and artificial marks, if there are any, so that only the HSQC signal patterns and experimental noise were left on a white background. Another important application of Adobe Photoshop in this preparation was to turn the spectra into black (signal and noise) and white (background). Spectra were flipped and rotated if there was a switch between f1 and f2 dimensions to make sure that ^{1}H NMR was on the horizontal dimension and ^{13}C NMR was on the vertical dimension. Hence, the input of the SMART is a binary image.

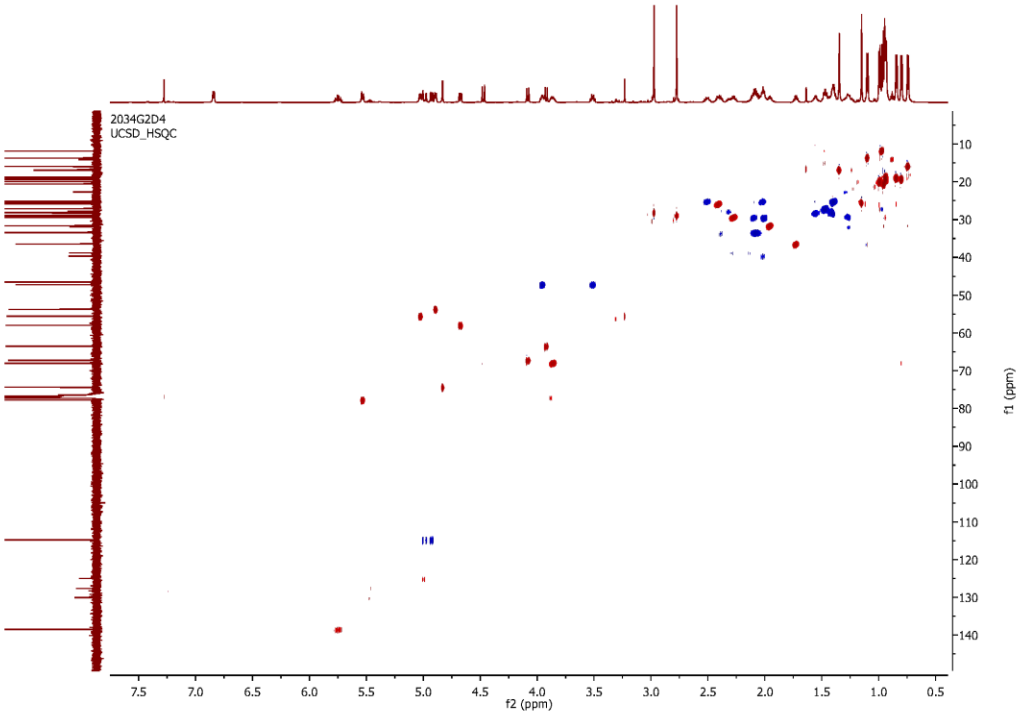


Figure 3.6.5. ^{1}H-^{13}C HSQC spectra of viequeamide A2 (2) in $CDCl_3$.

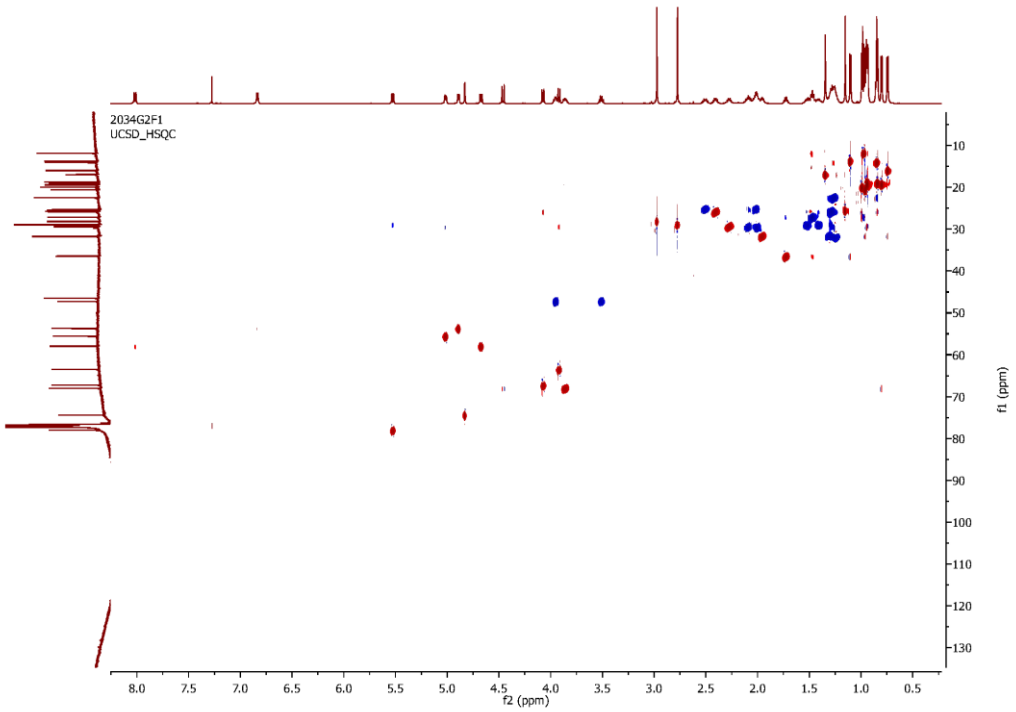


Figure 3.6.6. ^{1}H-^{13}C HSQC spectra of viequeamide A3 (3) in $CDCl_3$.

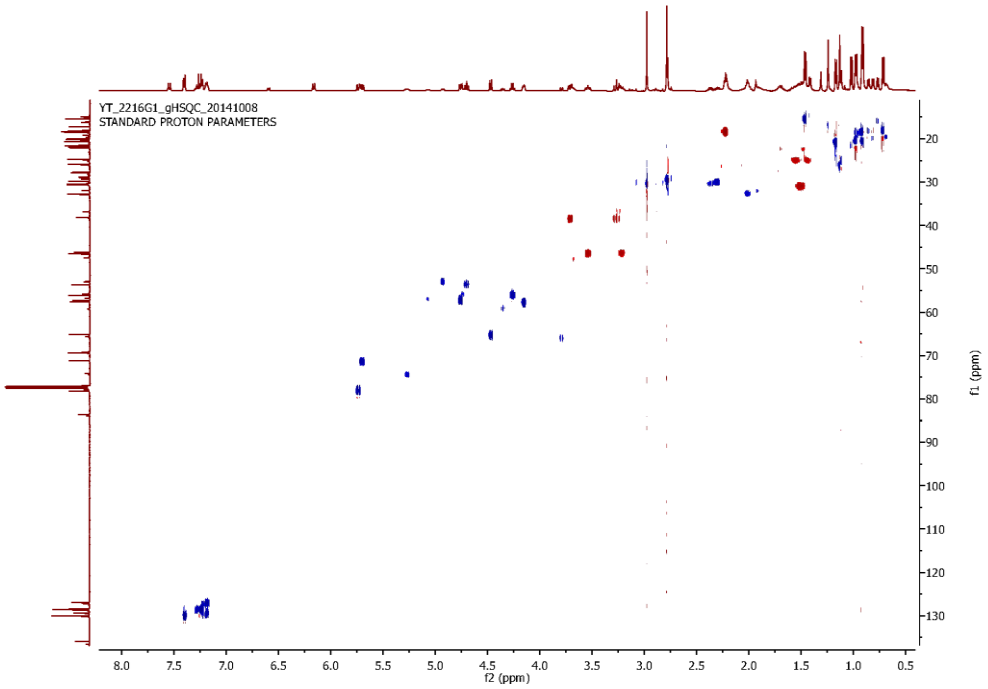


Figure 3.6.7. ^{1}H-^{13}C HSQC spectra of viequeamide B (4) in $CDCl_3$.

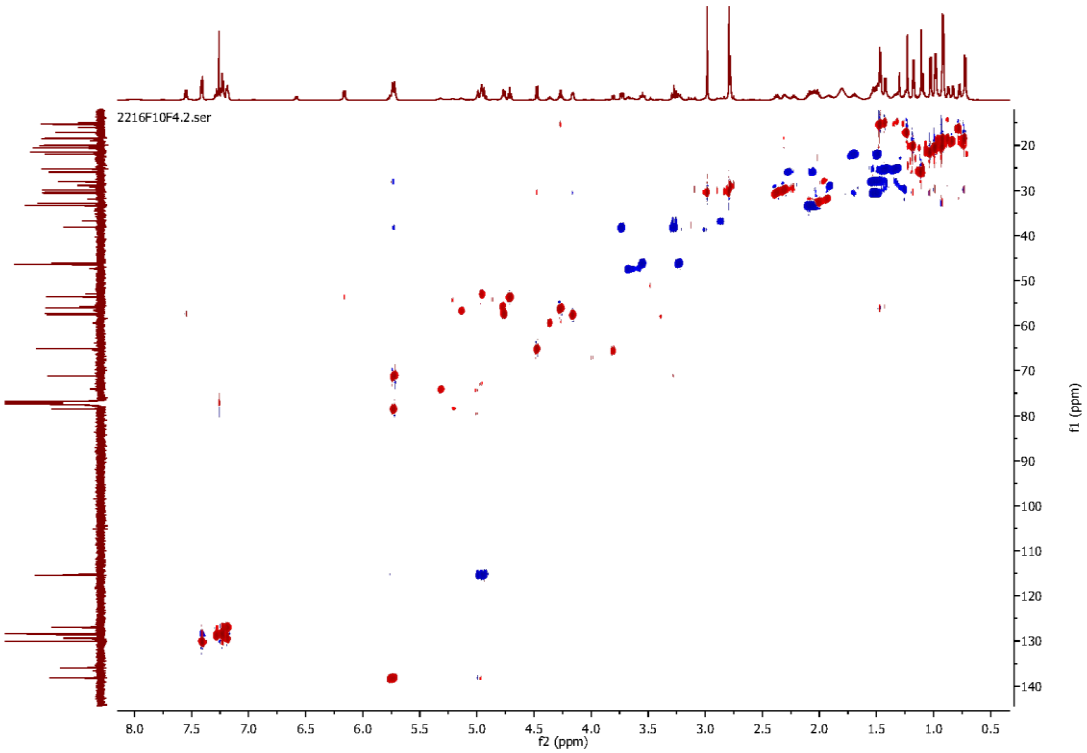


Figure 3.6.8. ^{1}H-^{13}C HSQC spectra of viequeamide C (5) in $CDCl_3$.

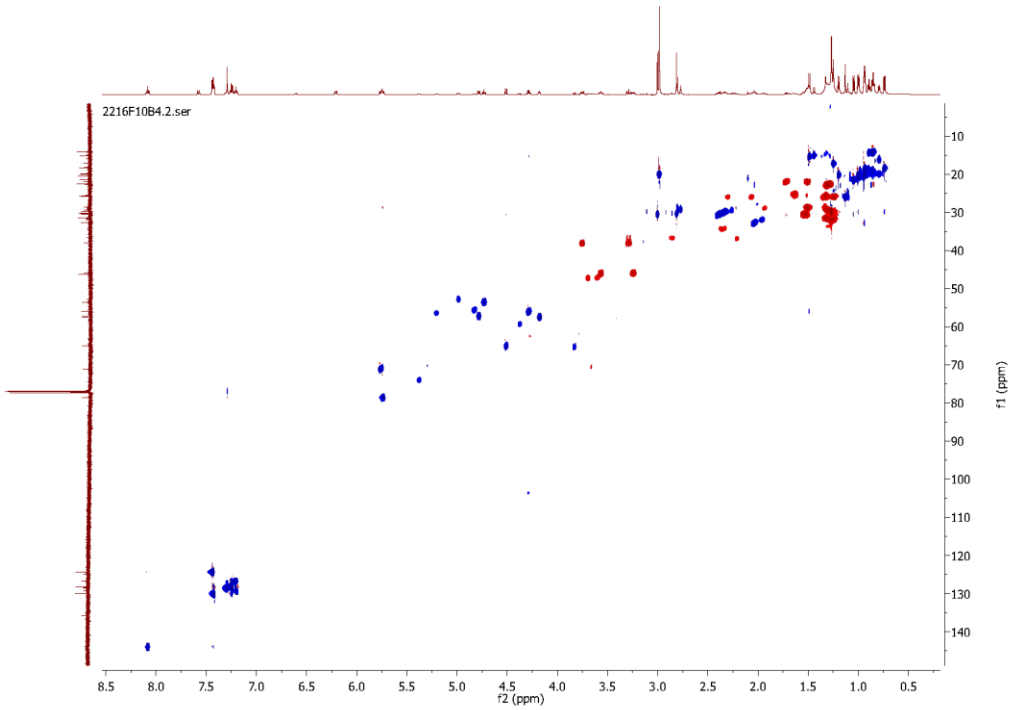


Figure 3.6.9. ^{1}H-^{13}C HSQC spectra of viequeamide D (6) in $CDCl_3$.

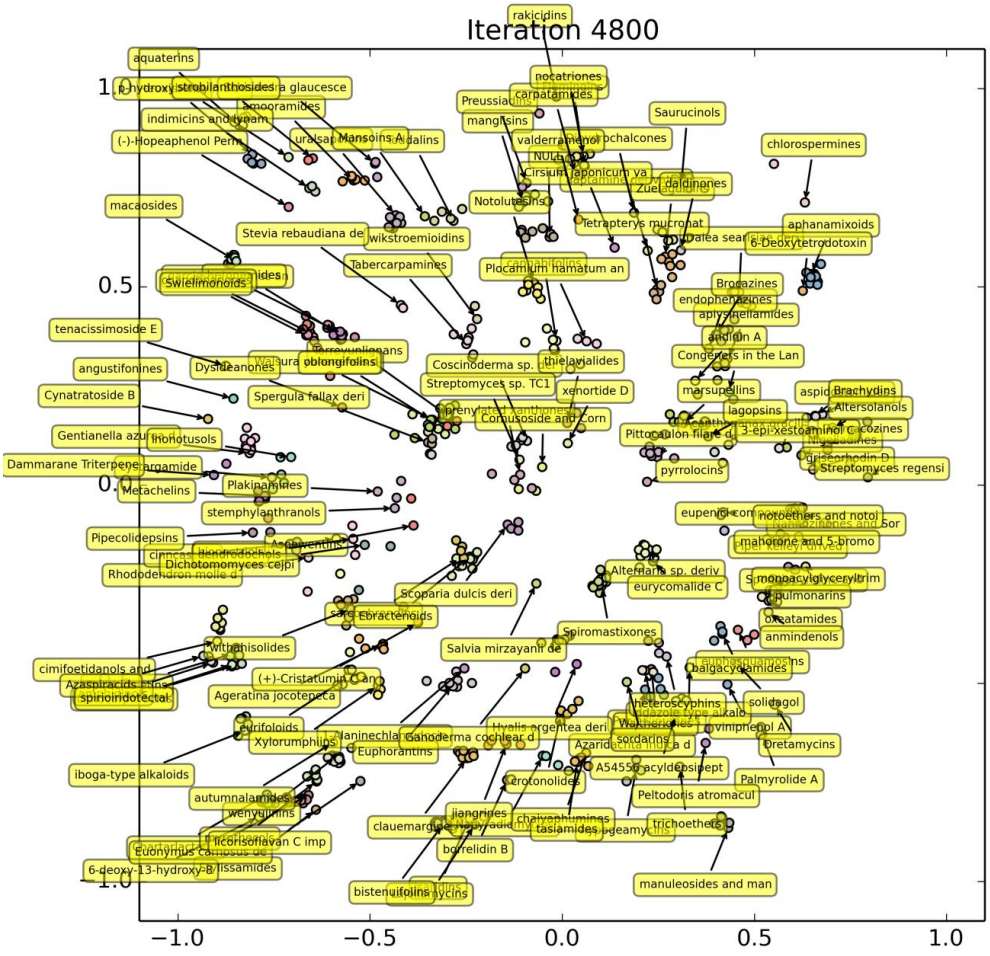


Figure 3.6.10. The cluster map containing 400 compounds after 4,800 iterations with node labels.

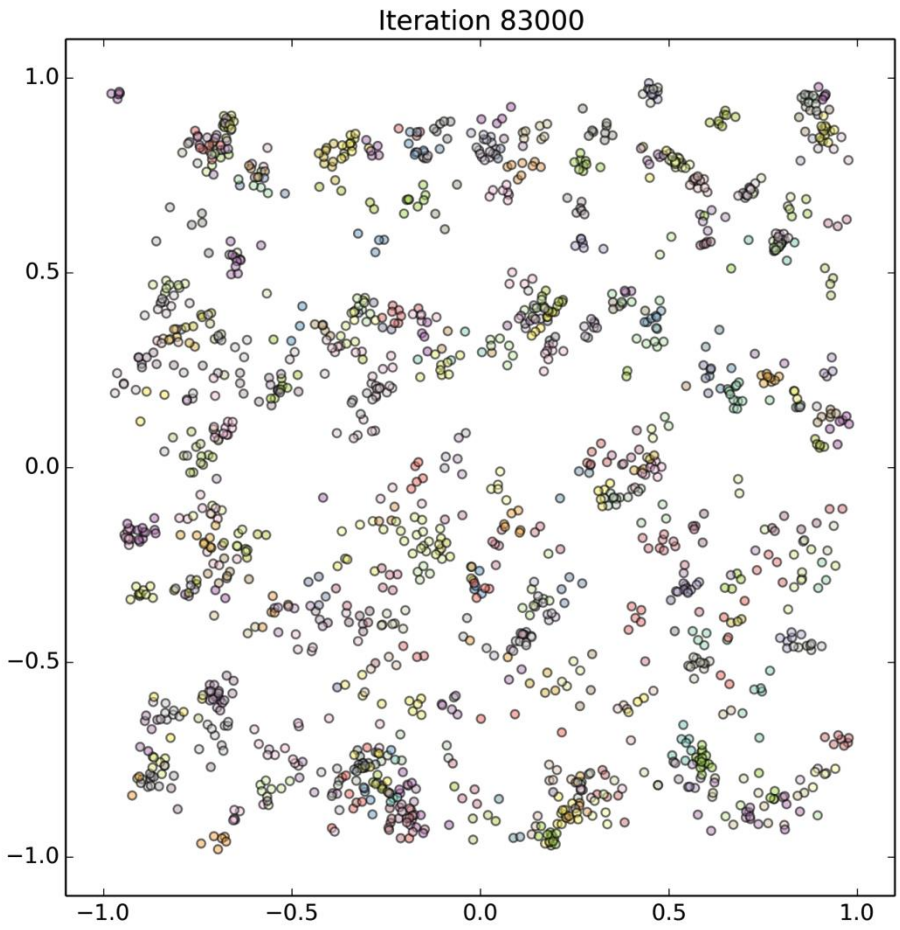


Figure 3.6.11. The cluster map containing 2,054 compounds after 83,000 iterations.

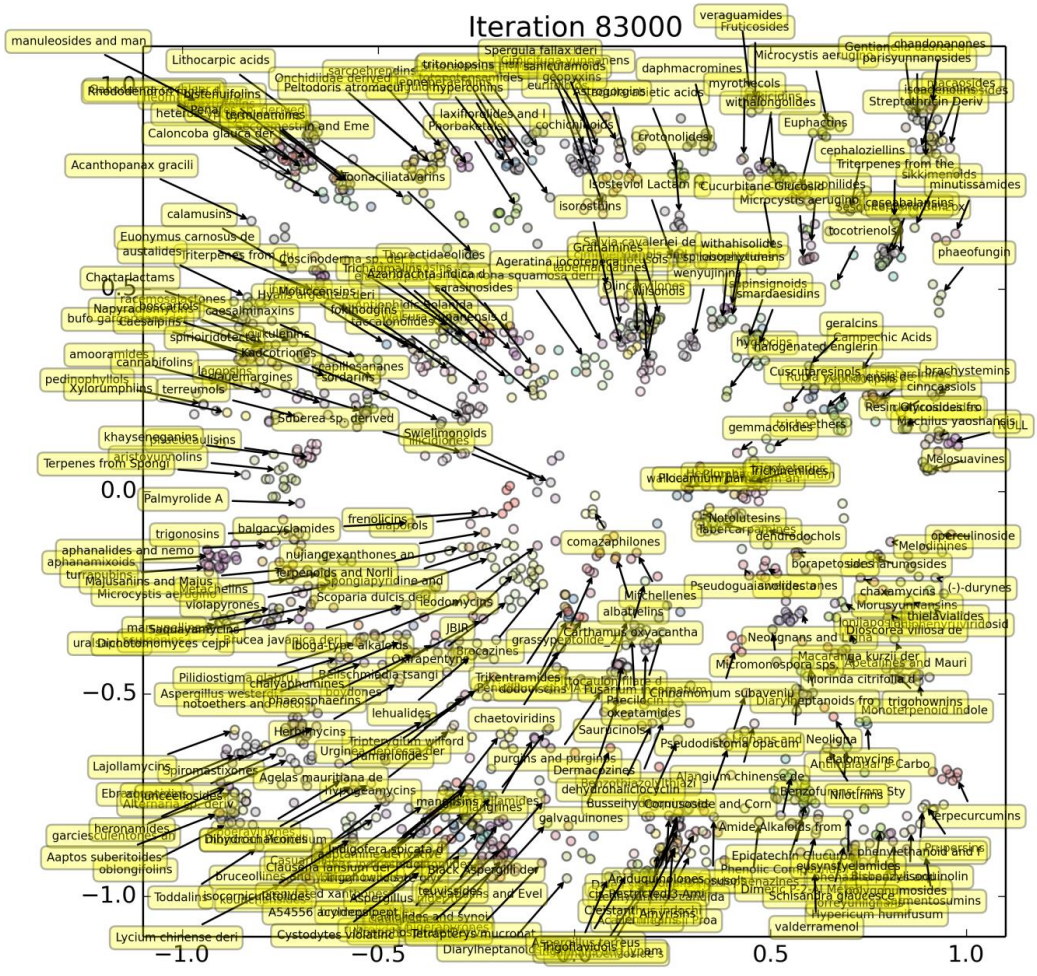


Figure 3.6.12. The cluster map containing 2,054 compounds after 83,000 iterations with labels.

Table 3.6.2. Lists of top 20 closest compound families for each spectra within the held-out set of ebractenoids in order of their distance to the test compounds in 10D space.[a]

Ebractenoid A	Ebractenoid B	Ebractenoid C	Ebractenoid D	Ebractenoid E	Ebractenoid F	Ebractenoid G	Ebractenoid H	Ebractenoid I	Ebractenoid J
bistenuifolins	lamesticumins	diaporols	oblongifolins	uralsaponins	elaeocarpucins	Lignans and Neolignans from Sinocalamus affinis	chaetoviridins	turrapubins	garciesculentones and garciesculenxanthone
withalongolides	Onchidiidae derived Bis-γ-pyrone Polypropionates	Brocazines	bistenuifolins	boscartols	khayseneganins	tritoniopsins	Flemingins	turrapubinis	amooramides
triptersinines	comazaphilones	phaeocaulisins	spirioiridotectal	aquaterins	Saquayamycins	junceellosides	triptersinines	Pittocaulon filare derived sespquiterpenes	Grahamines
clauemarginines	neomaclafungins	Neolignans and Lignans from Machilus robusta	trigohowninsins	turrapubins	Acacia mearnsii Proanthocyanidinss	trigonosins	kouitchensides	aphanamixoids	Phorbaketals
Astrogorgins	Diarylheptanoids from Alpinia katsumadai	indiosides	saniculamoids	spirioiridotectal	Saquayamycins	triptersinines	Clausenalansium derived compounds	Astrogorgins	amooramides
manglisins	gemmacolides	Saquayamycins	spirioiridotectal	Salvia cavaleriei derived ent-Kaurane	macaosides	capillosanes	withahisolides	clauemarginines	comazaphilones

Table 3.6.2. Lists of top 20 closest compound families for each spectra within the held-out set of ebractenoids in order of their distance to the test compounds in 10D space.[a] (Cont'd)

Ebractenoid A	Ebractenoid B	Ebractenoid C	Ebractenoid D	Ebractenoid E	Ebractenoid F	Ebractenoid G	Ebractenoid H	Ebractenoid I	Ebractenoid J
bistenuifolins	Bisbenzylisoquinoline Alkaloids Stephania epigaea derived	Cimicifuga yunnanensis derived Triterpenes	bruceollines and yadanzioli des	notoethers and notoincisols	trigohownins	JBIR-21	oblongifolins	Astrogorgins	aphanamixoids
Astrogorgins	caseabalansins	Diarylheptanoids and Flavonoids from Viscum album	trigonosins	uralsaponins	teuvissides	benzyl benzoate glycosides	Lignans and Neolignans from Sinocalamus affinis	triptersinines	Lithocarpic acids
Grahamines	aphanamixoids	Machilus yaoshansis derived dammarane glycosides	calamusins	notoethers and notoincisols	macaosides	macaosides	Sinocalamus affinis derived triterpenoids and steroids	Astrogorgins	Zuelaguidins
Chartarlactams	notoethers and notoincisols	Lignans and Neolignans from Sinocalamus affinis	cochichinoids	fokihodgins	Lignans and Neolignans from Sinocalamus affinis	Chartarlactams	teuvissides	Grahamines	Trichagmalins
turrapubins	Ganoderma cochlear derived Triterpenoids	clethroidosides	turrapubins	hypogeamycins	Euonymus carnosus derived Lupane Triterpenoids Euonymus carnosus	Pilidiostigma glabrum derived dibenzofurans	Pilidiostigma glabrum derived dibenzofurans	aphanamixoids	Melosuavines
rakicidins	Euphorantins	Lycium chinense derived neolignanamides and lignanamides	Isatis indigotica derived alkaloids	trigonosins	Prupersins	Metachelins	Aniduquinolones	hygrocins	Astrogorgins

Table 3.6.2. Lists of top 20 closest compound families for each spectra within the held-out set of ebractenoids in order of their distance to the test compounds in 10D space.[a] (Cont'd)

Ebractenoid A	Ebractenoid B	Ebractenoid C	Ebractenoid D	Ebractenoid E	Ebractenoid F	Ebractenoid G	Ebractenoid H	Ebractenoid I	Ebractenoid J
manglisins	khaysenegans	Rubia yunnanensis derived Arborinane-type Triterpenoids and Anthraquinones	bruceollines and yadanziolides	Neolignans and Lignans from Machilus robusta	trigohownins	Secoemestrin and Emericellenes	Metachelins	Microcystis aeruginosa derived protease inhibitors	terminamines
myrothecols	Diarylheptanoids from Alpinia katsumadai	frenolicins	turrapubins	Salvia cavaleriei derived ent-Kaurane	macaosides	khaysenegans	capillosanes	Notolutesins	Benzothiazolylthiazinodihydroisoquinolines from human red hair
benzyl benzoate glycosides	Plumbaginesand Plumbagosides	taccalonolides	Lignans and Neolignans from Sinocalamus affinis	trigoheterins	Euphactins	aquaterins	Salvia cavaleriei derived ent-Kaurane	Walsura yunanensis derived limonoids	cimifoetidanols and cimifoetidanosides
iboga-type alkaloids	Euphorantins	uralsaponins	iboga-type alkaloids	Pittocaulon filare derived sesquiterpenes	benzyl benzoate glycosides	diaporols	diaporols	Chartarlactams	Microcystis aeruginosa derived micropeptins
turrapubins	cimifoetidanols and cimifoetidanosides	Machilus yaoshansis derived dammarane glycosides	Saquayamycins	myrothecols	elaeocarpucins	peptidolipins	tritoniopsins	Herdmanines	gemmacolides
Diarylheptanoids from Alpinia katsumadai	Astrogorgins	phaeofungin	aristoyunnolins	withalongolides	uralsaponins	notoethers and notoincisols	Alternaria sp. derived metabolites	bistenuifolins	amooramides
Walsura yunanensis derived limonoids	Pseudoguaianolides and Guaianolides from Inula hupehensis	Machilus yaoshansis derived dammarane glycosides	trigohownins	teuvissides	triptersinines	Spergula fallax derived glycosides	junceellosides	Swielimonoids	aphanamixoids

Table 3.6.2. Lists of top 20 closest compound families for each spectra within the held-out set of ebractenoids in order of their distance to the test compounds in 10D space.[a] (Cont'd)

Ebractenoid A	Ebractenoid B	Ebractenoid C	Ebractenoid D	Ebractenoid E	Ebractenoid F	Ebractenoid G	Ebractenoid H	Ebractenoid I	Ebractenoid J
comazaphilones	indimicins and lynamicins	uralsaponins	Onchidiidae derived Bis-γ-pyrone Polypropionates	Euonymus carnosus derived Lupane Triterpenoids Euonymus carnosus	trigoheterins	Clausenalansium derived compounds	hypogeamycins	aquaterins	aphanalides and nemoralisins

a. The first row contains the names of each test compound within the ebractenoids family (bold). Compound families that are terpenoids/terpenes within the top 20 closest compounds are coloured red, and can be considered “hits”; non-terpenoids are coloured black. The blocks containing the same family of compounds within the top 20 hits for each ebractenoid are similarly coloured.

Table 3.6.3. Lists of top 20 closest compound families for each spectra within the held-out set of naphthomycins in order of their distance to the test compounds in 10D space.[a]

naphthomycin L	naphthomycin M	naphthomycin N
Clausena lansium derived compounds	JBIR-94-125	dysolenticins
Euphorantins	rakicidins	Trigohowilols
Clausena lansium derived compounds	Pittocaulon filare derived sespquiterpenes	Acacia mearnsii Proanthocyanidinss
Trigohowilols	Dimeric P-2-AI Metabolites	Busseihydroquinones
Penares sp. derived triterpenoids	laxiflorolides and laxiflorins	Melodinines
Triterpenes from the Leaves of Rosa laevigata	Plocamium hamatum and Plocamium costatum derived halogenated monoterpenes	Apetalines and Mauritines
Trigohowilols	Morusyunnansins	tritoniopsins
albatrelins	Grahamines	trigonosins
albatrelins	Annona squamosa derived bistetrahydrofuran annonaceous acetogenines with mosher	Clausena lansium derived compounds
triptersinines	Ganoderma cochlear derived Triterpenoids	lonijaposides
Onchidiidae derived Bis-γ-pyrone Polypropionates	Ganoderma cochlear derived Triterpenoids	Ganoderma cochlear derived Triterpenoids
Notolutesins	neomaclafungins	Triterpenes from the Leaves of Rosa laevigata
Neolignans and Lignans from Machilus robusta	diaporols	sapinsignoids
withahisolides	aquilarabietic acids	Euphorantins
Terpecurcumins	Monoterpenoid Indole Alkaloids from Gardneria ovata	Herdmanines
Oxirapentyns	aquilarabietic acids	comazaphilones
comazaphilones	Euphorantins	Cimicifuga yunnanensis derived Triterpenes
Amide Alkaloids from Piper boehmeriaefolium	Melosuavines	withahisolides
Zephyranthes candida derived alkaloids	withalongolides	cephaloziellins
Cornusoside and Cornolactones	dodoviscins	Notolutesins

a. The first row contains the names of each test compound within the naphthomycins family (bold). Compound families that are polyketides within the top 20 closest compounds are coloured red, and can be considered “hits”; non-polyketides are coloured black. The blocks containing the same family of compounds within the top 20 hits for each naphthomycin are similarly coloured.

Table 3.6.4. Lists of top 20 closest compound families for each spectra within the held-out set of veraguamides in order of their distance to the test compounds in 10D space.[a]

tetrahydro veraguamide A	veraguamide A	veraguamide B	veraguamide C	veraguamide D	veraguamide E	veraguamide F	veraguamide G
autumnalamides	noroleanane triterpenoids from Paeonia rockii	austalides	Lithocarpic acids	Lithocarpic acids	Microcystis aeruginosa derived protease inhibitors	austalides	withalongolides
terminamines	autumnalamides	Microcystis aeruginosa derived protease inhibitors	minutissamides	grassypeptolides	Tauromantellic acid	Aeruginosins	noroleanane triterpenoids from Paeonia rockii
phoslactomycins	Caesalpins	noroleanane triterpenoids from Paeonia rockii	lamesticumins	Peltodoris atromaculata derived fulvinols	austalides	Tauromantellic acid	Lithocarpic acids
triptersinines	wilsonols	Siderophore from Streptomyces sp. YM5-799	geldanamycins	geldanamycins	Melosuavines	Balticidins	cimifoetidanols and cimifoetidanosides
phaeocaulisins	Moluccensins	cimifoetidanols and cimifoetidanosides	grassypeptolides	lamesticumins	Lithocarpic acids	Lithocarpic acids	Lithocarpic acids
wilsonols	junceellosides	withalongolides	sarcoehrendins	Onchidiidae derived Bis-γ-pyrone Polypropionates	NN′-Methyleno-didemnin A from Tridemnum solidum	Melosuavines	Ganoderma cochlear derived Triterpenoids
lystabactins	lagopsins	Plocamium hamatum and Plocamium costatum derived halogenated monoterpenes	NN′-Methyleno-didemnin A from Tridemnum solidum	Lithocarpic acids	Aeruginosins	Microcystis aeruginosa derived protease inhibitors	cochichinoids
cochichinoids	racemosalactones	Walsura yunanensis derived limonoids	lamesticumins	NN′-Methyleno-didemnin A from Tridemnum solidum	Siderophore from Streptomyces sp. YM5-799	Walsura yunanensis derived limonoids	spirastrellolides
noroleanane triterpenoids from Paeonia rockii	Penicillium sp. MA-37 derived meroterpenoids and diphenyl ethers	cimifoetidanols and cimifoetidanosides	Lithocarpic acids	minutissamides	noroleanane triterpenoids from Paeonia rockii	eusynstyelamides	terminamines
caesalminaxins	dysolenticins	Lithocarpic acids	Microcystis aeruginosa derived	Phorbaketals	cimifoetidanols and	pouosides	viequeamides

Table 3.6.4. Lists of top 20 closest compound families for each spectra within the held-out set of veraguamides in order of their distance to the test compounds in 10D space.[a] (Cont'd)

tetrahydro veraguamide A	veraguamide A	veraguamide B	veraguamide C	veraguamide D	veraguamide E	veraguamide F	veraguamide G
			protease inhibitors		cimifoetidanosides		
dysolenticins	sarasinosides	NN′-Methyleno-didemnin A from Tridemnum solidum	noroleanane triterpenoids from Paeonia rockii	noroleanane triterpenoids from Paeonia rockii	Microcystis aeruginosa derived protease inhibitors	Siderophore from Streptomyces sp. YM5-799	hypercohins
austalides	Cedrus deodara derived tubulin inhibitors	Aeruginosins	cimifoetidanols and cimifoetidanosides	cimifoetidanols and cimifoetidanosides	sarcoehrendins	cimifoetidanols and cimifoetidanosides	austalides
viequeamides	austalides	taccalonolides	isorosthins	sarcoehrendins	sikkimenoids	noroleanane triterpenoids from Paeonia rockii	cimifoetidanols and cimifoetidanosides
sarcoehrendins	viequeamides	borapetosides	Pipecolidepsins	Pittocaulon filare derived sespquiterpenes	Chartarlactams	isorosthins	grassypeptolide_2
sarasinosides	dysolenticins	sarcoehrendins	austalides	Microcystis aeruginosa derived protease inhibitors	Balticidins	cimifoetidanols and cimifoetidanosides	garciesculentones and garciesculenxanthone
tasiamides	sarcoehrendins	Lithocarpic acids	Peltodoris atromaculata derived fulvinols	Onchidiidae derived Bis-γ-pyrone Polypropionates	Microcystis aeruginosa derived protease inhibitors	Microcystis aeruginosa derived protease inhibitors	Euphorantins
terminamines	gukulenins	Tabercarpamines	Onchidiidae derived Bis-γ-pyrone Polypropionates	noroleanane triterpenoids from Paeonia rockii	grassypeptolides	brachystemins	noroleanane triterpenoids from Paeonia rockii
dysolenticins	phoslactomycins	Microcystis aeruginosa derived protease inhibitors	homotemsirolimuses	minutissamides	tasiamides	Tauromantellic acid	cephaloziellins
austalides	austalides	Grahamines	Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans	Sinocalamus affinis derived triterpenoids and steroids	lystabactins	NN′-Methyleno-didemnin A from Tridemnum solidum	cimifoetidanols and cimifoetidanosides

Table 3.6.4. Lists of top 20 closest compound families for each spectra within the held-out set of veraguamides in order of their distance to the test compounds in 10D space.[a] (Cont'd)

tetrahydro veraguamide A	**veraguamide A**	**veraguamide B**	**veraguamide C**	**veraguamide D**	**veraguamide E**	**veraguamide F**	**veraguamide G**
dysolenticins	sarasinosides	hypercohins	Onchidiidae derived Bis-γ-pyrone Polypropionates	Lithocarpic acids	Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans	Microcystis aeruginosa derived protease inhibitors	Onchidiidae derived Bis-γ-pyrone Polypropionates

a. The first row contains the names of each test compound within the veraguamides family (bold). Compound families that are peptides within the top 20 closest compounds are coloured red, and can be considered “hits”; non-peptides are coloured black. The blocks containing the same family of compounds within the top 20 hits for each veraguamide are similarly coloured.

Table 3.6.5. Lists of top 50 closest compound families for each spectra within the newly isolated members of the viequeamides family in order of their distance to the test compounds in 10D space.[a]

viequeamide A	viequeamide A2	viequeamide A3[b]	viequeamide B	viequeamide C[b]	viequeamide D
cimifoetidanols and cimifoetidanosides	khayseneganins	aphanamixoids	Astrogorgins	cephaloziellins	cimifoetidanols and cimifoetidanosides
cephaloziellins	Euphorantins	Swielimonoids	gemmacolides	wenyujinins	sarasinosides
sarasinosides	Astrogorgins	tabernaricatines	gemmacolides	myrothecols	Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans
Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans	turrapubins	wenyujinins	Astrogorgins	Notolutesins	spirastrellolides
spirastrellolides	Gentianella azurea derived triterpenoids	Swielimonoids	teuvissides	vitextrifolins	jiangrines
jiangrines	Euphorantins	Plocamium hamatum and Plocamium costatum derived halogenated monoterpenes	aphanamixoids	myrothecols	cimifoetidanols and cimifoetidanosides
Gentianella azurea derived triterpenoids	aphanamixoids	aphanamixoids	Notolutesins	cimifoetidanols and cimifoetidanosides	cephaloziellins
chlorajaponilides	Triterpenes from the Leaves of Rosa laevigata	hypercohins	Grahamines	Gentianella azurea derived triterpenoids	chlorajaponilides
Cucurbitane Glucosides from Machilus yaoshansis	sedonans	Clausena lansium derived compounds	cephaloziellins	caseabalansins	Cucurbitane Glucosides from Machilus yaoshansis
cimifoetidanols and cimifoetidanosides	Swielimonoids	veraguamides	Plocamium hamatum and Plocamium costatum derived halogenated monoterpenes	Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans	Gentianella azurea derived triterpenoids
caseabalansins	gemmacolides	Torreyunlignans	morrocan Penicillium citrinum derived compounds	Astrogorgins	Phorbaketals
caseabalansins	Gentianella azurea derived triterpenoids	Aeruginosins	Dimeric P-2-AI Metabolites	aphanamixoids	caseabalansins
Phorbaketals	cannabifolins	veraguamides	Ebractenoids	Gentianella azurea derived triterpenoids	caseabalansins
Dalea searlsiae derived compounds	withalongolides	Xylorumphiins	gemmacolides	Walsura yunanensis derived limonoids	veraguamides
veraguamides	turrapubins	triptersinines	Plumbagines and Plumbagosides	gemmacolides	Dalea searlsiae derived compounds
Euphorantins	lamesticumins	spirastrellolides	khayseneganins	Dalea searlsiae derived compounds	lamesticumins
uralsaponins	cephaloziellins	cephaloziellins	aquaterins	spirastrellolides	Euphorantins

Table 3.6.5. Lists of top 50 closest compound families for each spectra within the newly isolated members of the viequeamides family in order of their distance to the test compounds in 10D space.[a] (Cont'd)

viequeamide A	viequeamide A2	viequeamide A3[b]	viequeamide B	viequeamide C[b]	viequeamide D
Euphorantins	viequeamides	aphanalides and nemoralisins	Torreyunlignans	cimifoetidanols and cimifoetidanosides	Swielimonoids
cimifoetidanols and cimifoetidanosides	Pittocaulon filare derived sespquiterpenes	triptersinines	Walsura yunanensis derived limonoids	Phorbaketals	cimifoetidanols and cimifoetidanosides
myrothecols	Microcystis aeruginosa derived protease inhibitors	Scoparia dulcis derived diterpnoids	Pittocaulon filare derived sespquiterpenes	Swielimonoids	Euphorantins
Notolutesins	aphanamixoids	Notolutesins	spirastrellolides	laxiflorolides and laxiflorins	Gentianella azurea derived triterpenoids
Swielimonoids	uralsaponins	khayseneganins	aphanamixoids	Gentianella azurea derived triterpenoids	Gentianella azurea derived triterpenoids
Gentianella azurea derived triterpenoids	Bisbenzylisoquinoline Alkaloids Stephania epigaea derived	Swielimonoids	viequeamides	Acacia mearnsii Proanthocyanidinss	myrothecols
lamesticumins	Euphorantins	Herdmanines	amooramides	aphanamixoids	uralsaponins
Euphorantins	aphanamixoids	Tripterygium wilfordii derived dihydroagarofurans	lamesticumins	Grahamines	Euphorantins
wenyujinins	tritoniopsins	khayseneganins	fokihodgins	Sesquiterpene Benzoxazoles and Sesquiterpene Quinones from Dactylospongia elegans	Notolutesins
Gentianella azurea derived triterpenoids	cimifoetidanols and cimifoetidanosides	Plocamium hamatum and Plocamium costatum derived halogenated monoterpenes	Scoparia dulcis derived diterpnoids	aphanamixoids	chandonanones
cimifoetidanols and cimifoetidanosides	Dalea searlsiae derived compounds	Trichagmalins	Dalea searlsiae derived compounds	sarasinosides	cephaloziellins
viequeamides	bistenuifolins	cimifoetidanols and cimifoetidanosides	cimifoetidanols and cimifoetidanosides	caseabalansins	Pittocaulon filare derived sespquiterpenes
Pittocaulon filare derived sespquiterpenes	Brocazines	withahisolides	Caesalpins	Onchidiidae derived Bis-γ-pyrone Polypropionates	cimifoetidanols and cimifoetidanosides
morrocan Penicillium citrinum derived compounds	caseabalansins	Grahamines	uralsaponins	Euphorantins	Gentianella azurea derived triterpenoids
chandonanones	neomaclafungins	Xylorumphiins	Fruticosides	marsupellins	viequeamides
cephaloziellins	Apetalines and Mauritines	Astrogorgins	nujiangexanthones and nujiangefolins	Euphorantins	Oxirapentyns
aphanamixoids	chandonanones	Diarylheptanoids from Alpinia katsumadai	uralsaponins	Euphorantins	wenyujinins
Gentianella azurea derived triterpenoids	veraguamides	Clausena lansium derived compounds	Swielimonoids	Notolutesins	bistenuifolins

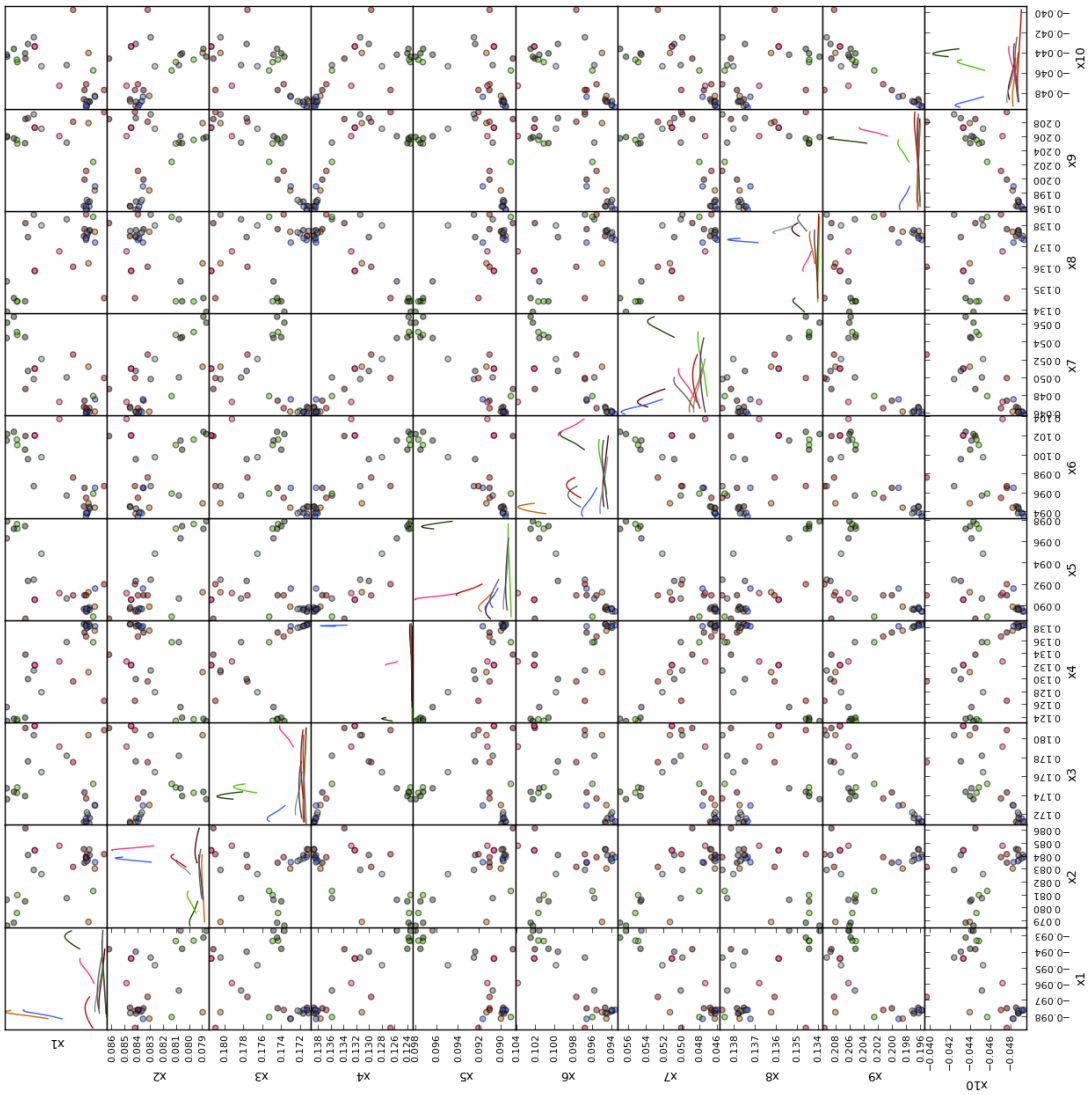
Table 3.6.5. Lists of top 50 closest compound families for each spectra within the newly isolated members of the viequeamides family in order of their distance to the test compounds in 10D space.[a] (Cont'd)

viequeamide A	viequeamide A2	viequeamide A3[b]	viequeamide B	viequeamide C[b]	viequeamide D
Swielimonoids	neomaclafungins	Diarylheptanoids from Alpinia katsumadai	Salvia cavaleriei derived ent-Kaurane	morrocan Penicillium citrinum derived compounds	morrocan Penicillium citrinum derived compounds
Oxirapentyns	chlorajaponilides	withalongolides	Xylorumphiins	cimifoetidanols and cimifoetidanosides	tritoniopsins
cephaloziellins	jiangrines	cimifoetidanols and cimifoetidanosides	bistenuifolins	Zuelaguidins	Swielimonoids
Annona squamosa derived bistetrahydrofuran annonaceous acetogenines with mosher	diaporols	endophenazines	aphanamixoids	Bisbenzylisoquinoline Alkaloids Stephania epigaea derived	cephaloziellins
tritoniopsins	veraguamides	aphanamixoids	aphanamixoids	Gentianella azurea derived triterpenoids	aphanamixoids
bistenuifolins	cephaloziellins	trigohownins	cimifoetidanols and cimifoetidanosides	Dimeric P-2-AI Metabolites	cephaloziellins
aphanamixoids	Cucurbitane Glucosides from Machilus yaoshansis	gemmacolides	Euphorantins	cephaloziellins	aphanamixoids
aphanamixoids	gemmacolides	pedinophyllols	comazaphilones	Walsura yunanensis derived limonoids	Astrogorgins
cephaloziellins	Cimicifuga yunnanensis derived Triterpenes	teuvissides	Trichagmalins	Cucurbitane Glucosides from Machilus yaoshansis	Annona squamosa derived bistetrahydrofuran annonaceous acetogenines with mosher
Astrogorgins	Oxirapentyns	withalongolides	aphanalides and nemoralisins	jiangrines	veraguamides
veraguamides	turrapubins	Euphactins	aphanamixoids	amooramides	aphanamixoids
Grahamines	caesalminaxins	Walsura yunanensis derived limonoids	endophenazines	sarcoehrendins	isorosthins
Astrogorgins	bruceollines and yadanziolides	gemmacolides	tabernaricatines	chlorajaponilides	khayseneganins
khayseneganins	neomaclafungins	Walsura yunanensis derived limonoids	Zuelaguidins	teuvissides	Astrogorgins
Dalea searlsiae derived compounds	isorosthins	cephaloziellins	hypercohins	amooramides	gemmacolides

a. Close compound families of each newly obtained viequeamide spectra are listed in each column with the compound names in bold. Compounds in the viequeamide family are highlighted in red; non-viequeamides are coloured black. The blocks containing the same peptidic family of compounds within the top 50 hits for each of the viequeamides are similarly coloured.

b. Due to the small number of viequeamides in the training set (only 2), viequeamides A3 and C were not closely associated with this family.

Visualization of 10D embeddings using 10 randomly selected compound families and 3 randomly selected examples from each family (diagonal is the projection for every compound on that axis)



Total number of pairs of inputs in the training process

A siamese network is comprised of a pair of identical networks that are trained with pairs of inputs. For SMART5: there are 5,982 positive pairs, and 2,103,476 negative pairs. For SMART10 there were 3,787 positive pairs and 410,718 negative pairs. Notice that number of pairs grows with an order of O(n^2). During training, the positive and negative pairs must be balanced. We generated minibatches of 200 with 100 pairs randomly chosen from the positive pair set, and 100 from the negative pair set. The pairs are resampled each time.

This training strategy explains one reason why a siamese network is appropriate for this task. To train a deep network, typically on the order of a million examples are used. The advantage of the siamese network is that they amplify a small data set by training on pairs as opposed to single examples. A second reason why siamese networks are appropriate is that we need a cluster space, rather than a classification of compounds into families. If the system were simply trained to take an example and classify it, this would not be appropriate for new compounds for which the category is unknown. The cluster map generated by SMART places the new compounds into a similarity space with known compounds.

Reasons for software and parameters selection

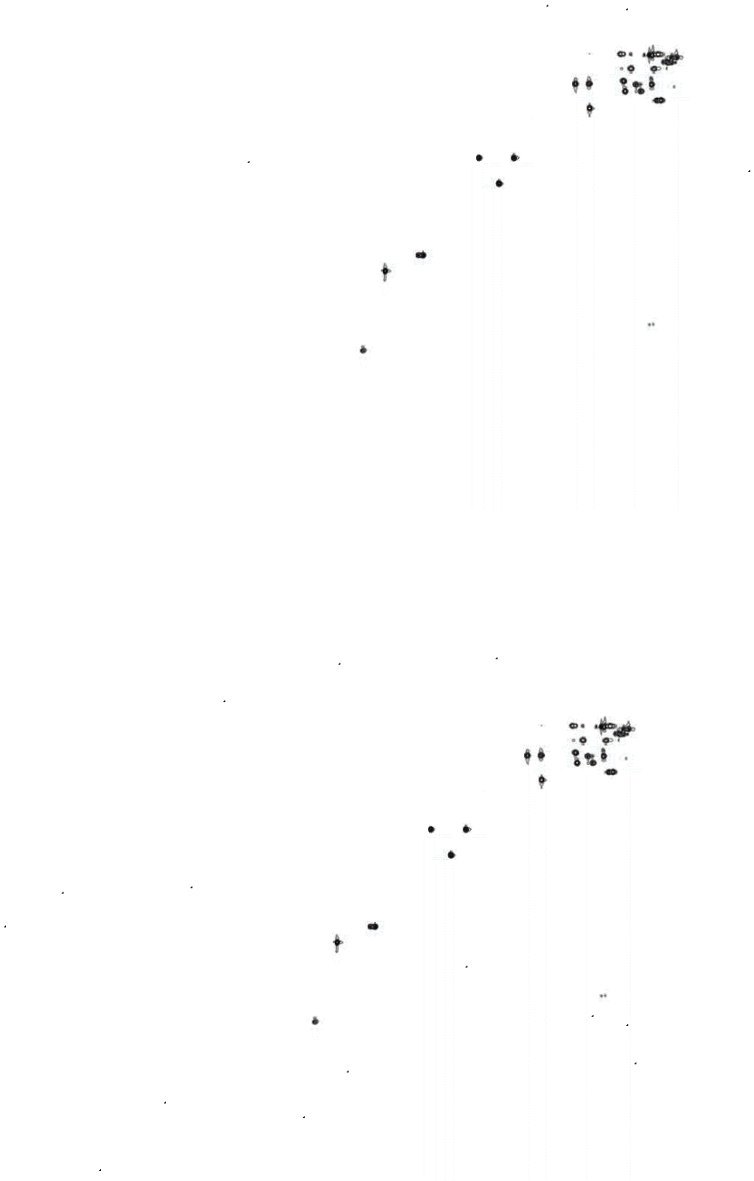
There exist a number of different frameworks to perform deep learning, including Torch, Tensorflow, Theano, Caffe, mxnet, etc. There is no particular advantage of one over another, except for the tradeoff between ease of use and flexibility. The authors of this manuscript were most familiar with Theano, and at the time of project's start, this was the version of a deep learning framework that had a good (native) python interface, as well as a stable development version.

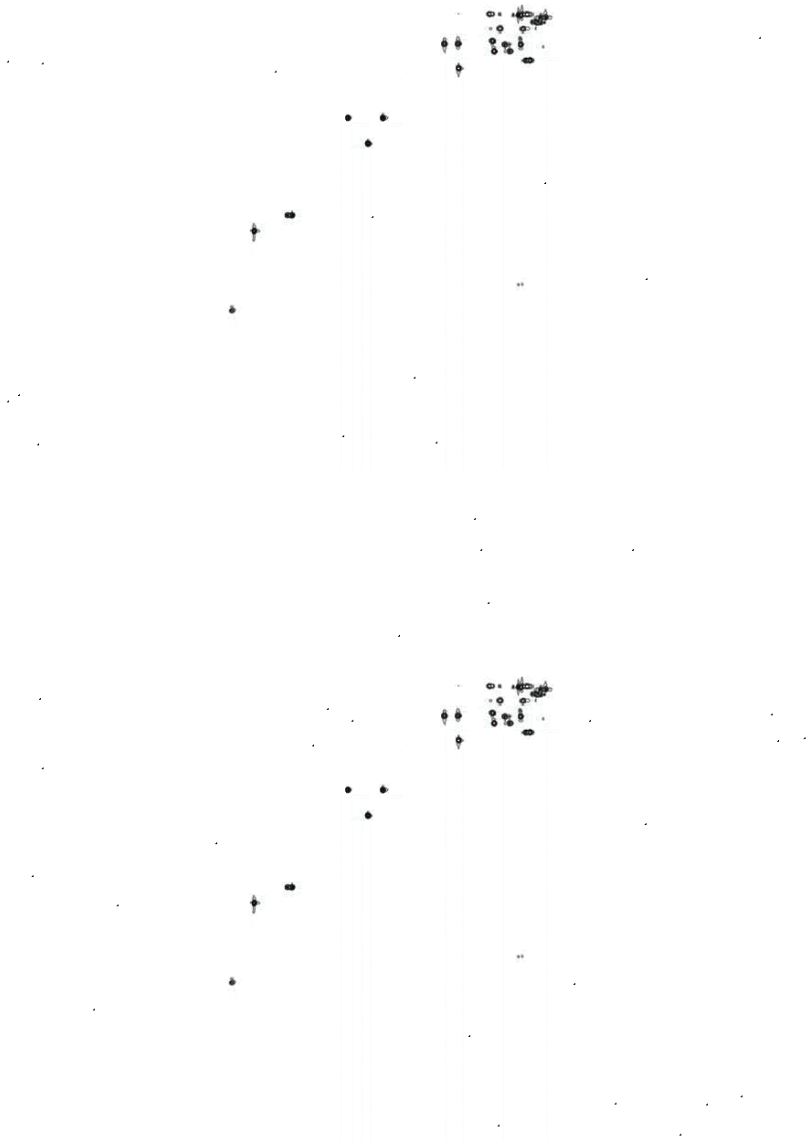
The choice of parameters and hyperparameters is completely empirical in the entire deep learning field, with no best method except trial and error. The reported parameters led to good results in all experiments. When searching for hyperparameters, we looked into GPU utilization/runtime per iteration. There is a tradeoff between batch size, number of iterations of learning, and wall-clock time. GPUs are very good at processing batches of examples for training, and the more memory a GPU has, the larger batch it can process (higher utilization), which will reduce the noise in Stochastic gradient descent (SGD). However, the larger the batch, the more time it requires, but this in turn, reduces the number of SGD iterations that are required. Another essential parameter to tune is the learning rate; if the learning rate is too high, the optimization procedure can diverge, whereas smaller rates may terminate before reaching the best minimum of the objective function. Based on preliminary experiments, we chose a batch size of 200 pairs and an initial learning rate of 0.002. Other parameters were likewise chosen based on preliminary experiments, for example: margin = 0.02, l2 regularization multiplier on dense layers: 0.0001, update rule = adagrad.

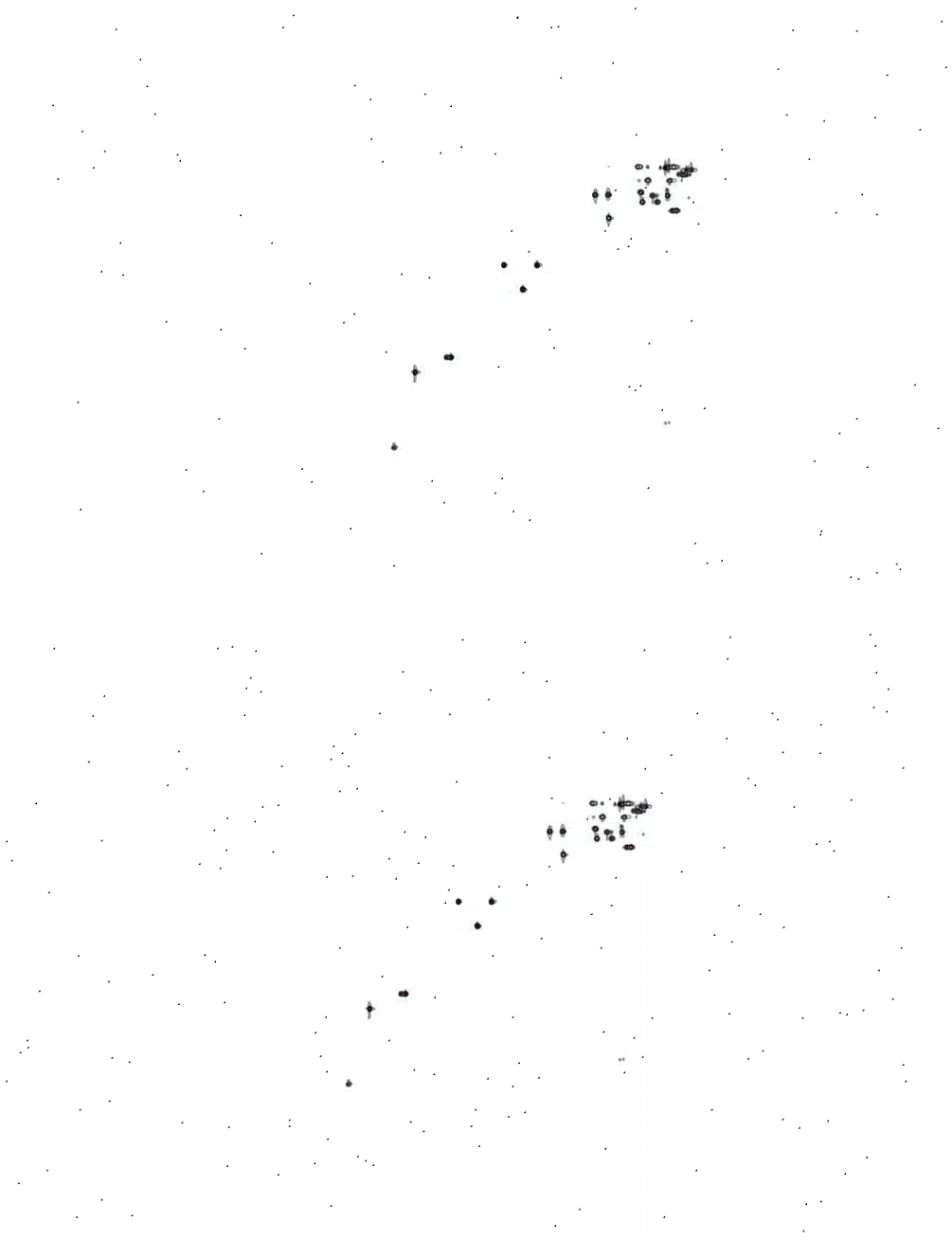
SMART training speed

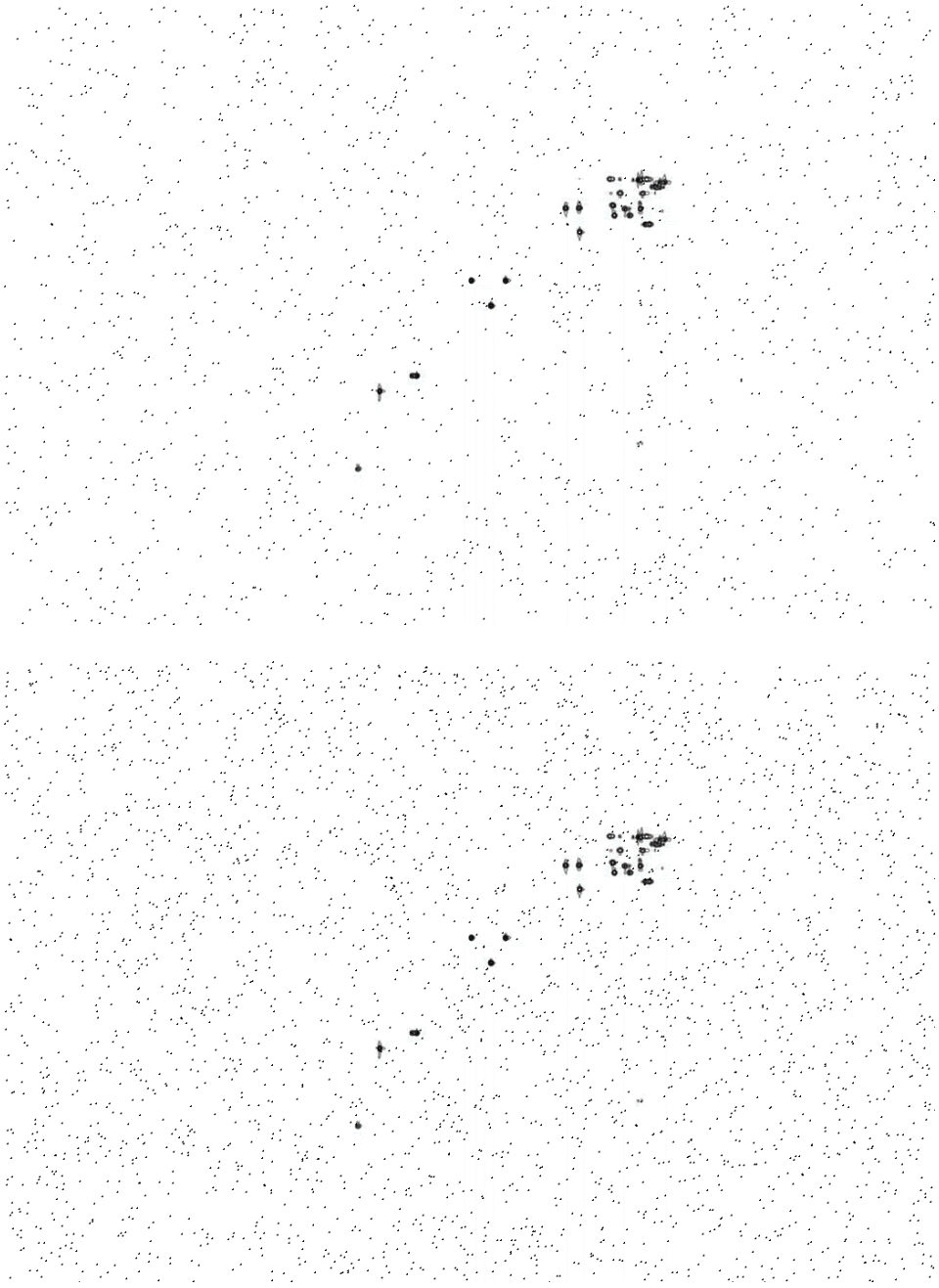
Our initial experiments on SMART5 with k=2 dimensional embedding were run for 100k iterations on an Amazon EC2 g2.2xlarge instance (using NVIDIA GRID K520 GPU) which required 8 days. Using batch normalization, the time was reduced to 28 hours. Our final experiments on SMART5 were run on an Nvidia Titan X (Maxwell), and we limited the number of iterations to 15k; this was completed in 3-4 hours.

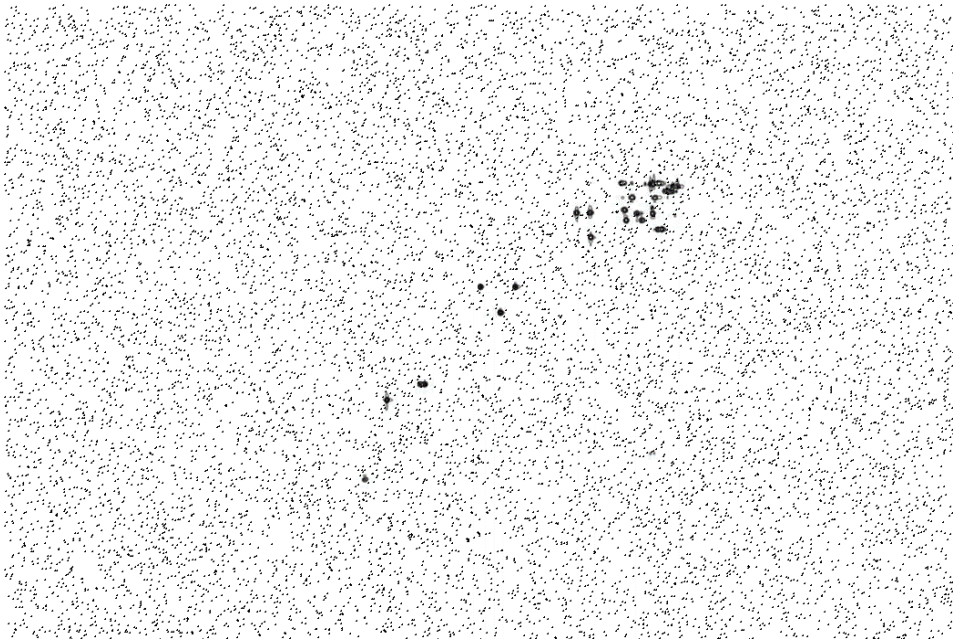
Noisy HSQC spectra of ebractenoid C

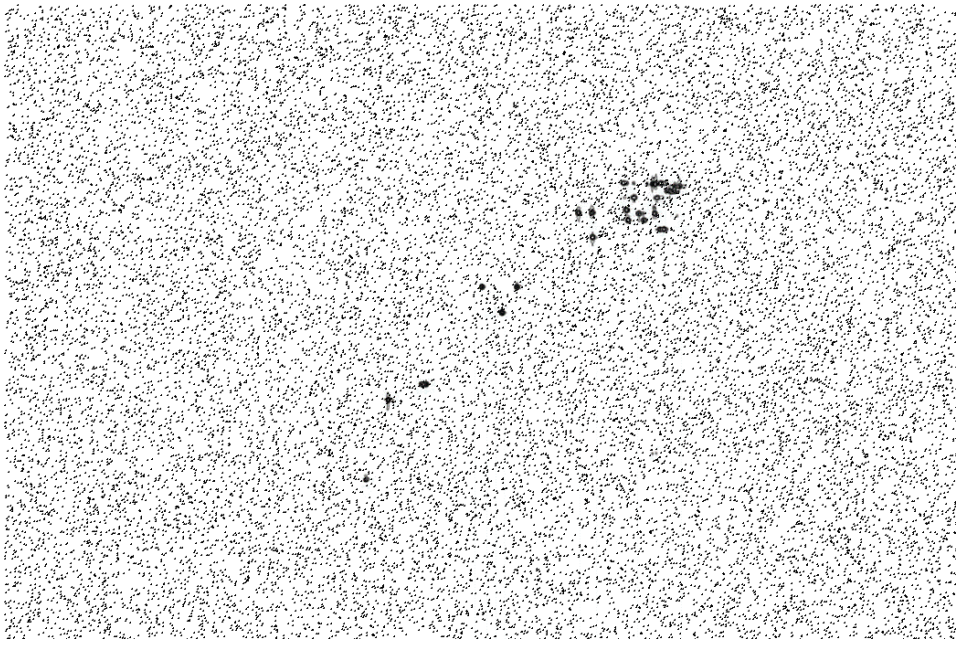




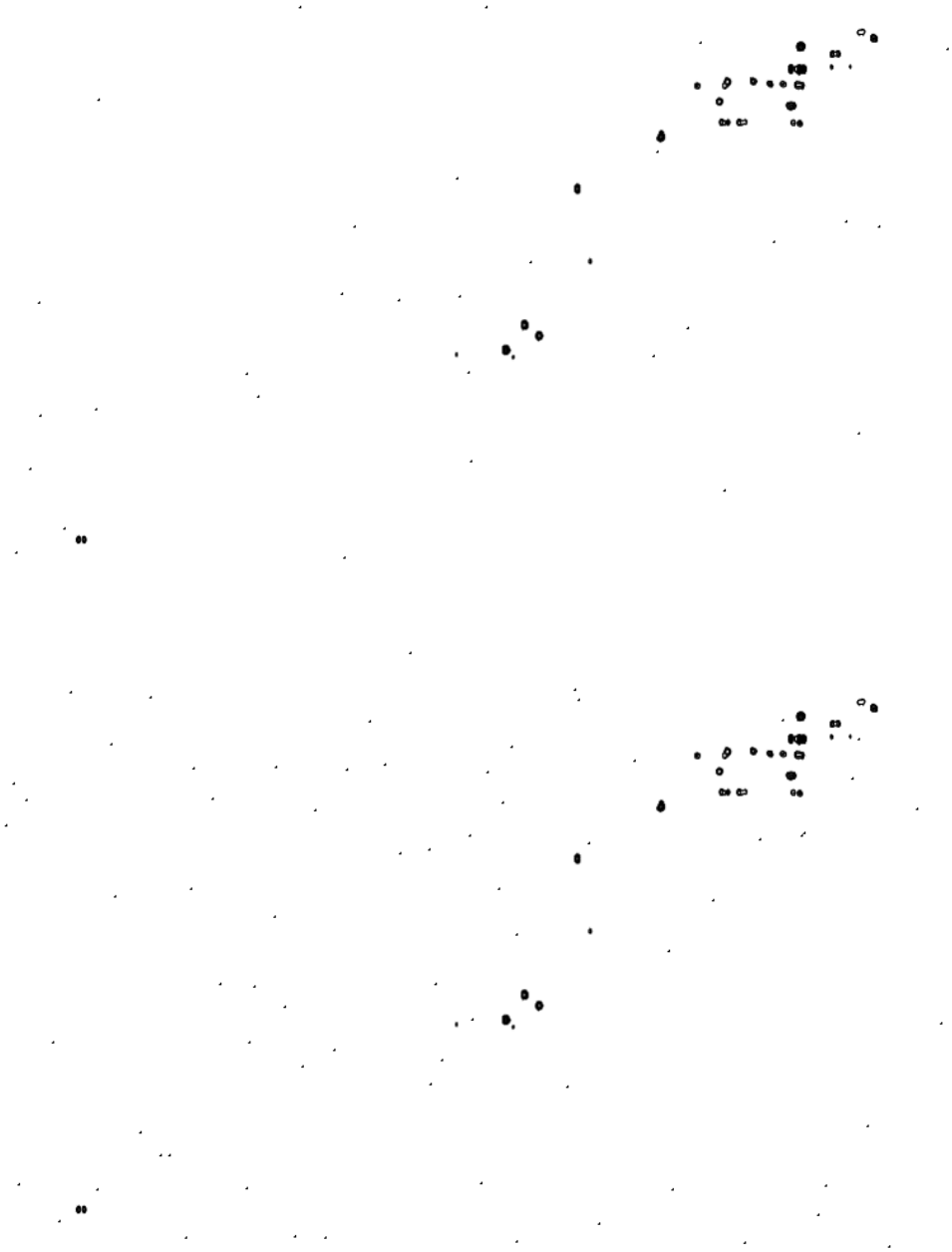


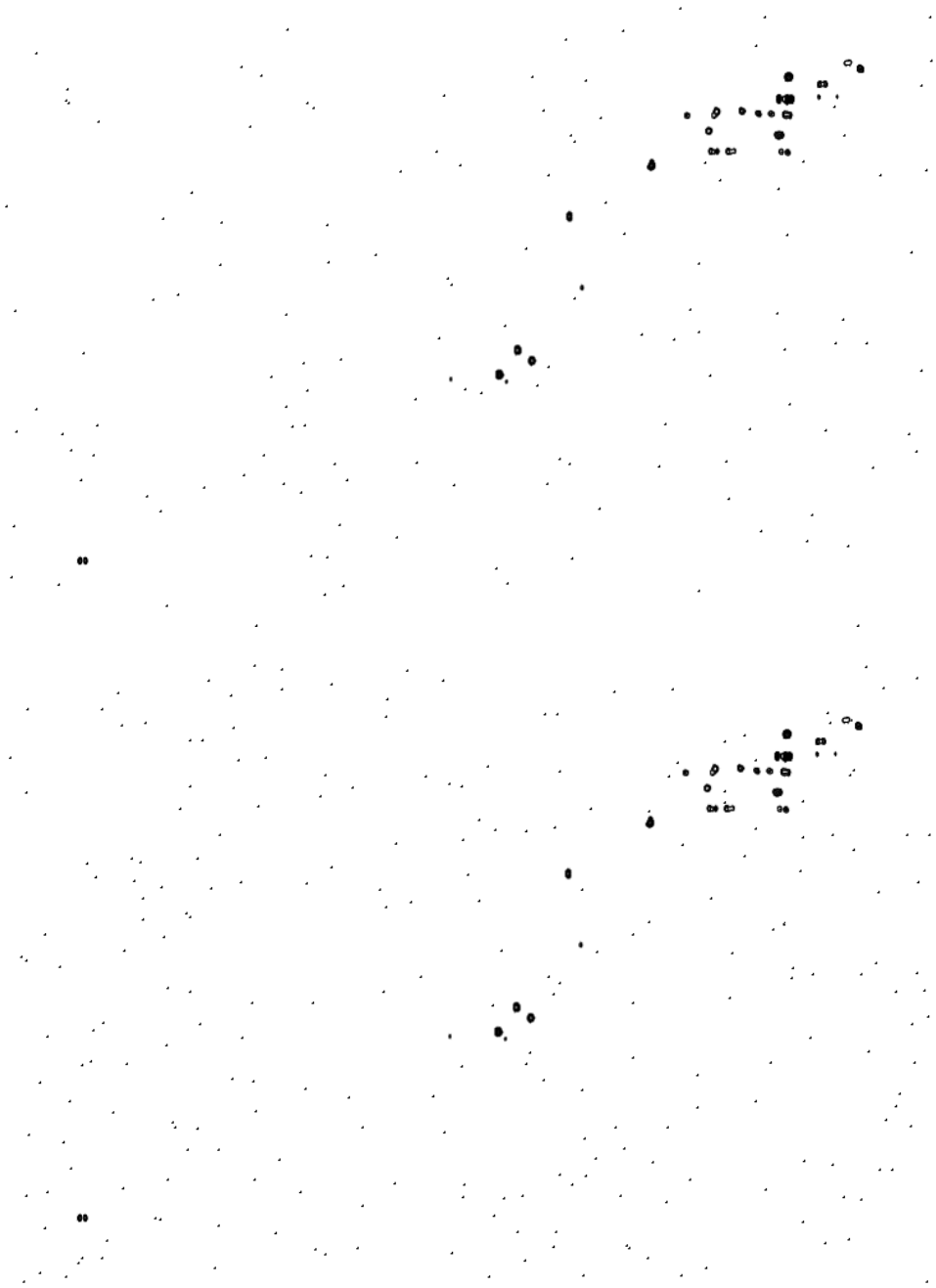


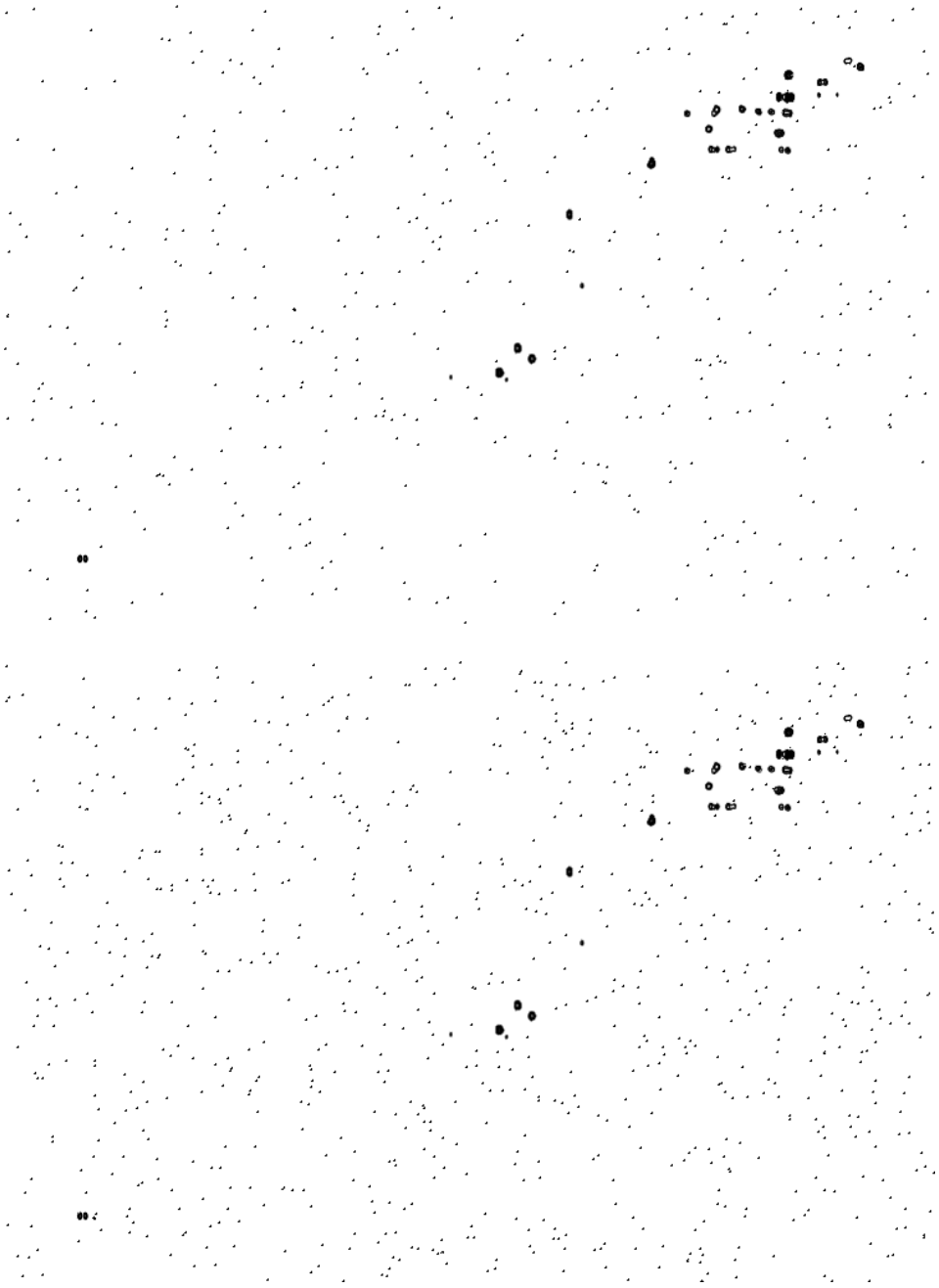


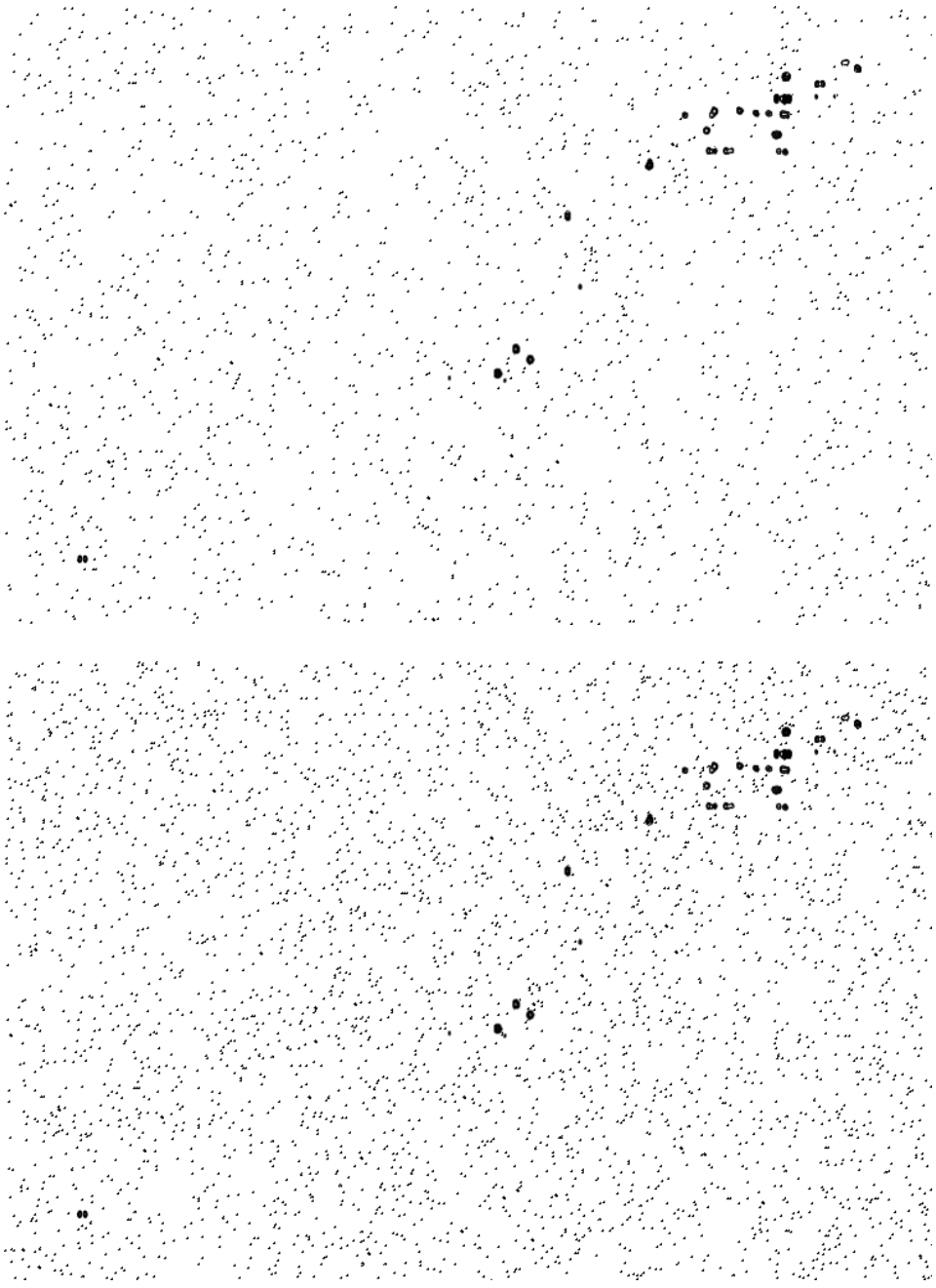


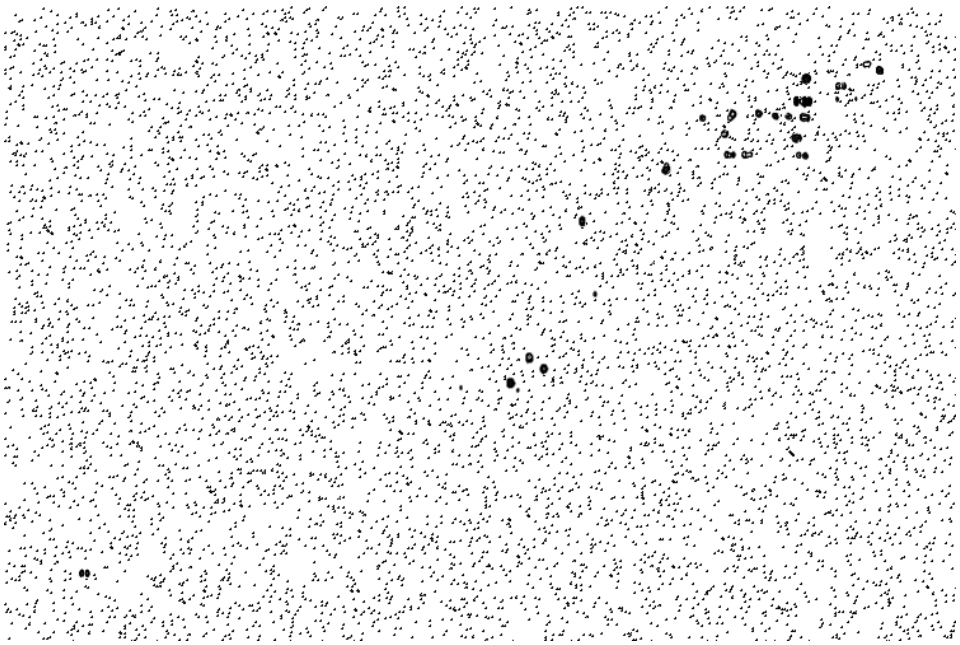
Noisy HSQC spectra of hyphenrone I

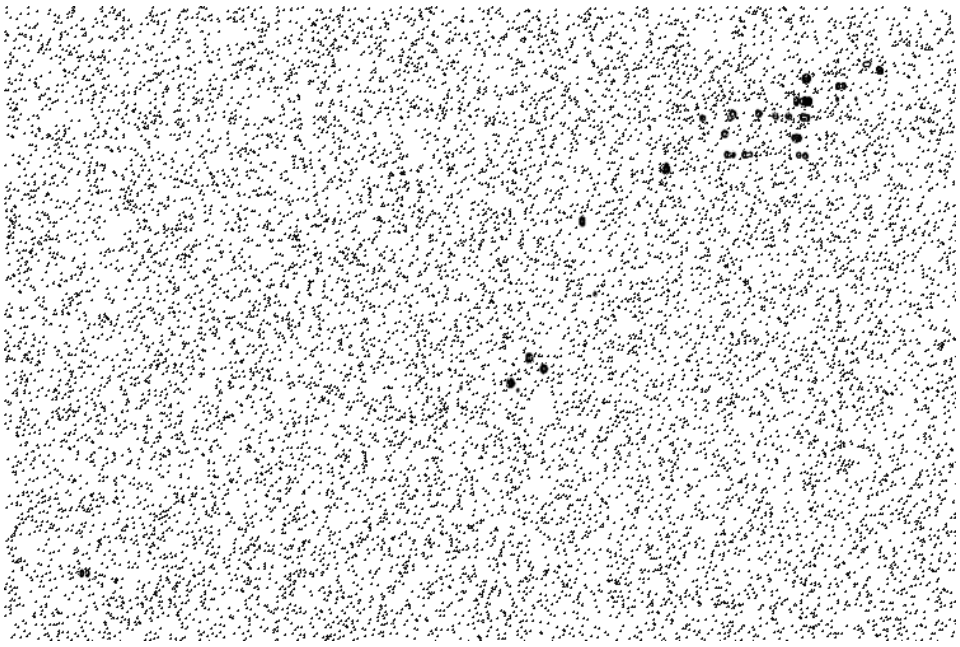












3.7 Chapter 3 References

1. Newman, D. J.; Cragg, G. M., Natural Products As Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **2016,** 79, (3), 629-661.

2. Kursar, T. A.; Caballero-George, C. C.; Capson, T. L.; Cubilla-Rios, L.; Gerwick, W. H.; Gupta, M. P.; Ibanez, A.; Linington, R. G.; McPhail, K. L.; Ortega-Barria, E.; Romero, L. I.; Solis, P. N.; Coley, P. D., Securing economic benefits and promoting conservation through bioprospecting. *Bioscience* **2006,** 56, (12), 1005-1012.

3. Liu, W. T.; Lamsa, A.; Wong, W. R.; Boudreau, P. D.; Kersten, R.; Peng, Y.; Moree, W. J.; Duggan, B. M.; Moore, B. S.; Gerwick, W. H.; Linington, R. G.; Pogliano, K.; Dorrestein, P. C., MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in Streptomyces roseosporus. *Journal of Antibiotics* **2014,** 67, (1), 99-104.

4. Medema, M. H.; Kottmann, R.; Yilmaz, P.; Cummings, M.; Biggins, J. B.; Blin, K.; de Bruijn, I.; Chooi, Y. H.; Claesen, J.; Coates, R. C.; Cruz-Morales, P.; Duddela, S.; Dusterhus, S.; Edwards, D. J.; Fewer, D. P.; Garg, N.; Geiger, C.; Gomez-Escribano, J. P.; Greule, A.; Hadjithomas, M.; Haines, A. S.; Helfrich, E. J. N.; Hillwig, M. L.; Ishida, K.; Jones, A. C.; Jones, C. S.; Jungmann, K.; Kegler, C.; Kim, H. U.; Kotter, P.; Krug, D.; Masschelein, J.; Melnik, A. V.; Mantovani, S. M.; Monroe, E. A.; Moore, M.; Moss, N.; Nutzmann, H. W.; Pan, G. H.; Pati, A.; Petras, D.; Reen, F. J.; Rosconi, F.; Rui, Z.; Tian, Z. H.; Tobias, N. J.; Tsunematsu, Y.; Wiemann, P.; Wyckoff, E.; Yan, X. H.; Yim, G.; Yu, F. G.; Xie, Y. C.; Aigle, B.; Apel, A. K.; Balibar, C. J.; Balskus, E. P.; Barona-Gomez, F.; Bechthold, A.; Bode, H. B.; Borriss, R.; Brady, S. F.; Brakhage, A. A.; Caffrey, P.; Cheng, Y. Q.; Clardy, J.; Cox, R. J.; De Mot, R.; Donadio, S.; Donia, M. S.; van der Donk, W. A.; Dorrestein, P. C.; Doyle, S.; Driessen, A. J. M.; Ehling-Schulz, M.; Entian, K. D.; Fischbach, M. A.; Gerwick, L.; Gerwick, W. H.; Gross, H.; Gust, B.; Hertweck, C.; Hofte, M.; Jensen, S. E.; Ju, J. H.; Katz, L.; Kaysser, L.; Klassen, J. L.; Keller, N. P.; Kormanec, J.; Kuipers, O. P.; Kuzuyama, T.; Kyrpides, N. C.; Kwon, H. J.; Lautru, S.; Lavigne, R.; Lee, C. Y.; Linquan, B.; Liu, X. Y.; Liu, W.; Luzhetskyy, A.; Mahmud, T.; Mast, Y.; Mendez, C.; Metsa-Ketela, M.; Micklefield, J.; Mitchell, D. A.; Moore, B. S.; Moreira, L. M.; Muller, R.; Neilan, B. A.; Nett, M.; Nielsen, J.; O'Gara, F.; Oikawa, H.; Osbourn, A.; Osburne, M. S.; Ostash, B.; Payne, S. M.; Pernodet, J. L.; Petricek, M.; Piel, J.; Ploux, O.; Raaijmakers, J. M.; Salas, J. A.; Schmitt, E. K.; Scott, B.; Seipke, R. F.; Shen, B.; Sherman, D. H.; Sivonen, K.; Smanski, M. J.; Sosio, M.; Stegmann, E.; Sussmuth, R. D.; Tahlan, K.; Thomas, C. M.; Tang, Y.; Truman, A. W.; Viaud, M.; Walton, J. D.; Walsh, C. T.; Weber, T.; van Wezel, G. P.; Wilkinson, B.; Willey, J. M.; Wohlleben, W.; Wright, G. D.; Ziemert, N.; Zhang, C. S.; Zotchev, S.

B.; Breitling, R.; Takano, E.; Glockner, F. O., Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology* **2015,** 11, (9), 625-631.

5. Walsh, C. T., A chemocentric view of the natural product inventory. *Nature Chemical Biology* **2015,** 11, (9), 620-624.

6. Molinski, T. F., NMR of natural products at the 'nanomole-scale'. *Natural Product Reports* **2010,** 27, (3), 321-329.

7. Breton, R. C.; Reynolds, W. F., Using NMR to identify and characterize natural products. *Natural Product Reports* **2013,** 30, (4), 501-524.

8. Mobli, M.; Maciejewski, M. W.; Schuyler, A. D.; Stern, A. S.; Hoch, J. C., Sparse sampling methods in multidimensional NMR. *Physical Chemistry Chemical Physics* **2012,** 14, (31), 10835-10843.

9. Kazimierczuk, K.; Orekhov, V. Y., Accelerated NMR Spectroscopy by Using Compressed Sensing. *Angewandte Chemie-International Edition* **2011,** 50, (24), 5556-5559.

10. Palmer, M. R.; Suiter, C. L.; Henry, G. E.; Rovnyak, J.; Hoch, J. C.; Polenova, T.; Rovnyak, D., Sensitivity of Nonuniform Sampling NMR. *Journal of Physical Chemistry B* **2015,** 119, (22), 6502-6515.

11. Hyberts, S. G.; Arthanari, H.; Wagner, G., Applications of Non-Uniform Sampling and Processing. *Novel Sampling Approaches in Higher Dimensional NMR* **2012,** 316, 125-148.

12. Hyberts, S. G.; Milbradt, A. G.; Wagner, A. B.; Arthanari, H.; Wagner, G., Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. *Journal of Biomolecular NMR* **2012,** 52, (4), 315-327.

13. Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S.; Hoch, J. C., Data Sampling in Multidimensional NMR: Fundamentals and Strategies. *Novel Sampling Approaches in Higher Dimensional NMR* **2012,** 316, 49-77.

14. Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S., Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Analytical Chemistry* **2011,** 83, (5), 1649-1657.

15. Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J., Toward more reliable C-13 and H-1 chemical shift prediction: A systematic comparison of neural-network and least-squares regression based approaches. *Journal of Chemical Information and Modeling* **2008,** 48, (1), 128-134.

16. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** 521, (7553), 436-444.

17. Schmidhuber, J., Deep learning in neural networks: An overview. *Neural Networks* **2015,** 61, 85-117.

18. Gerwick, W. H.; Proteau, P. J.; Nagle, D. G.; Hamel, E.; Blokhin, A.; Slate, D. L., Structure of Curacin-a, a Novel Antimitotic, Antiproliferative, and Brine Shrimp Toxic Natural Product from the Marine Cyanobacterium Lyngbya-Majuscula. *Journal of Organic Chemistry* **1994,** 59, (6), 1243-1245.

19. Yoo, H. D.; Gerwick, W. H., Curacins B and C, new antimitotic natural products from the marine cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **1995,** 58, (12), 1961-1965.

20. Marquez, B.; Verdier-Pinard, P.; Hamel, E.; Gerwick, W. H., Curacin D, an antimitotic agent from the marine cyanobacterium *Lyngbya majuscula*. *Phytochemistry* **1998,** 49, (8), 2387-2389.

21. Tarsis, E. M.; Rastelli, E. J.; Wengryniuk, S. E.; Coltart, D. M., The apratoxin marine natural products: isolation, structure determination, and asymmetric total synthesis. *Tetrahedron* **2015,** 71, (31), 5029-5044.

22. Choi, H.; Mevers, E.; Byrum, T.; Valeriote, F. A.; Gerwick, W. H., Lyngbyabellins K-N from Two Palmyra Atoll Collections of the Marine Cyanobacterium Moorea bouillonii. *European Journal of Organic Chemistry* **2012**, (27), 5141-5150.

23. Marner, F. J.; Moore, R. E.; Hirotsu, K.; Clardy, J., Majusculamides a and B, 2 Epimeric Lipodipeptides from Lyngbya-Majuscula-Gomont. *Journal of Organic Chemistry* **1977,** 42, (17), 2815-2819.

24. Carter, D. C.; Moore, R. E.; Mynderse, J. S.; Niemczura, W. P.; Todd, J. S., Structure of Majusculamide-C, a Cyclic Depsipeptide from Lyngbya-Majuscula. *Journal of Organic Chemistry* **1984,** 49, (2), 236-241.

25. Moore, R. E.; Entzeroth, M., Majusculamide-D and Deoxymajusculamide-D, 2 Cytotoxins from Lyngbya-Majuscula. *Phytochemistry* **1988,** 27, (10), 3101-3103.

26. Bodis, L.; Ross, A.; Bodis, J.; Pretsch, E., Automatic compatibility tests of HSQC NMR spectra with proposed structures of chemical compounds. *Talanta* **2009,** 79, (5), 1379-1386.

27. Hinneburg, A.; Egert, B.; Porzel, A., Duplicate detection of 2D-NMR Spectra. *Journal of Integrative Bioinformatics* **2007,** 4, (1), 64.

28. Wolfram, K.; Porzel, A.; Hinneburg, A., Similarity search for multi-dimensional NMR-spectra of natural products. *Knowledge Discovery in Databases: Pkdd 2006, Proceedings* **2006,** 4213, 650-658.

29. Levitt, M. H., *Spin dynamics : basics of nuclear magnetic resonance 2nd edn, 345*. John Wiley & Sons: Chichester, England ; Hoboken, NJ, 2008; p 345.

30. Chopra, S.; Hadsell, R.; LeCun, Y., Learning a similarity metric discriminatively, with application to face verification. *2005 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings* **2005**, 539-546.

31. Aue, W. P.; Bartholdi, E.; Ernst, R. R., 2-Dimensional Spectroscopy - Application to Nuclear Magnetic-Resonance. *Journal of Chemical Physics* **1976,** 64, (5), 2229-2246.

32. Bodenhausen, G.; Freeman, R.; Turner, D. L., 2-Dimensional J-Spectroscopy - Proton-Coupled C-13 NMR. *Journal of Chemical Physics* **1976,** 65, (2), 839-840.

33. Levitt, M. H., *Spin dynamics : basics of nuclear magnetic resonance, 2nd edn, 38*. John Wiley & Sons: Chichester, England ; Hoboken, NJ, 2008; p xxv, 714 pages.

34. Papoulis, A., New Algorithm in Spectral Analysis and Band-Limited Extrapolation. *IEEE Transactions on Circuits and Systems* **1975,** 22, (9), 735-742.

35. Lin, E. C.; Opella, S. J., Sampling scheme and compressed sensing applied to solid-state NMR spectroscopy. *Journal of Magnetic Resonance* **2013,** 237, 40-48.

36. Burg, J. P., *Maximum entropy spectral analysis, Ph.D. thesis*. Stanford University, Stanford, California: S.l., 1975; p 1 v.

37. Burg, J. P., *A New analysis technique for time series data*. NATO advanced study institute on signal processing, Enschede, Netherlands, (1968).

38. Donoho, D. L.; Johnstone, I. M.; Hoch, J. C.; Stern, A. S., Maximum-Entropy and the Nearly Black Object. *Journal of the Royal Statistical Society Series B-Methodological* **1992,** 54, (1), 41-81.

39. Hoch, J. C.; Stern, A. S., *NMR Data Processing, 140-144*. Wiley-Liss: New York, 1996; p xi, 196 pages.

40. Hoch, J. C.; Stern, A. S., *NMR data processing, 93*. Wiley-Liss: New York, 1996; p xi, 196 pages.

41. Hadsell, R.; Chopra, S.; LeCun, Y., Dimensionality Reduction by Learning an Invariant Mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* **2006,** 2, 1735-1742.

42. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25* **2012**, 1097-1105.

43. Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.

44. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* **2015**.

45. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J., Learning Representations by Back-Propagating Errors. *Nature* **1986,** 323, (6088), 533-536.

46. Russell, S. J.; Norvig, P., *Artificial intelligence : a modern approach*. 3rd ed. 728-729; Prentice Hall: Upper Saddle River, N.J., 2010; p xviii, 1132 pages.

47. Russell, S. J.; Norvig, P., *Artificial intelligence : a Modern Approach*. 3rd ed. 720; Prentice Hall: Upper Saddle River, N.J., 2010; p xviii, 1132 pages.

48. Duchi, J.; Hazan, E.; Singer, Y., Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* **2011,** 12, 2121-2159.

49. Ioffe, S.; Szegedy, C., Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* **2015,** abs/1502.03167.

50. Kang, L. P.; Liu, Y. X.; Eichhorn, T.; Dapat, E.; Yu, H. S.; Zhao, Y.; Xiong, C. Q.; Liu, C.; Efferth, T.; Ma, B. P., Polyhydroxylated Steroidal Glycosides from Paris polyphylla. *Journal of Natural Products* **2012,** 75, (6), 1201-1205.

51. Lee, C. L.; Hwang, T. L.; Yang, J. C.; Cheng, H. T.; He, W. J.; Yen, C. T.; Kuo, C. L.; Chen, C. J.; Chang, W. Y.; Wu, Y. C., Anti-Inflammatory Spirostanol and Furostanol Saponins from Solanum macaonense. *Journal of Natural Products* **2014,** 77, (8), 1770-1783.

52. Thao, N. P.; Cuong, N. X.; Luyen, B. T. T.; Van Thanh, N.; Nhiem, N. X.; Koh, Y. S.; Ly, B. M.; Nam, N. H.; Van Kiem, P.; Van Minh, C.; Kim, Y. H., Anti-inflammatory Asterosaponins from the Starfish Astropecten monacanthus. *Journal of Natural Products* **2013,** 76, (9), 1764-1770.

53. Lv, H. W.; Zhu, M. D.; Luo, J. G.; Kong, L. Y., Antihyperglycemic Glucosylated Coumaroyltyramine Derivatives from Teucrium viscidum. *Journal of Natural Products* **2014,** 77, (2), 200-205.

54. Cai, J. Y.; Chen, D. Z.; Luo, S. H.; Kong, N. C.; Zhang, Y.; Di, Y. T.; Zhang, Q.; Hua, J.; Jing, S. X.; Li, S. L.; Li, S. H.; Hao, X. J.; He, H. P., Limonoids from Aphanamixis polystachya and Their Antifeedant Activity. *Journal of Natural Products* **2014,** 77, (3), 472-482.

55. Zhang, Y.; Wang, J. S.; Wei, D. D.; Gu, Y. C.; Wang, X. B.; Kong, L. Y., Bioactive Terpenoids from the Fruits of Aphanamixis grandifolia. *Journal of Natural Products* **2013,** 76, (6), 1191-1195.

56. Song, W.; Si, L. L.; Ji, S.; Wang, H.; Fang, X. M.; Yu, L. Y.; Li, R. Y.; Liang, L. N.; Zhou, D. M.; Ye, M., Uralsaponins M-Y, Antiviral Triterpenoid Saponins from the Roots of Glycyrrhiza uralensis. *Journal of Natural Products* **2014,** 77, (7), 1632-1643.

57. Liu, Z. G.; Li, Z. L.; Bai, J.; Meng, D. L.; Li, N.; Pei, Y. H.; Zhao, F.; Hua, H. M., Anti-inflammatory Diterpenoids from the Roots of Euphorbia ebracteolata. *Journal of Natural Products* **2014,** 77, (4), 792-799.

58. Rogers, D. J.; Tanimoto, T. T., Computer Program for Classifying Plants. *Science* **1960,** 132, (3434), 1115-1118.

59. Castillo, A. M.; Uribe, L.; Patiny, L.; Wist, J., Fast and shift-insensitive similarity comparisons of NMR using a tree-representation of spectra. *Chemometrics and Intelligent Laboratory Systems* **2013,** 127, 1-6.

60. Boudreau, P. D.; Byrum, T.; Liu, W. T.; Dorrestein, P. C.; Gerwick, W. H., Viequeamide A, a Cytotoxic Member of the Kulolide Superfamily of Cyclic Depsipeptides from a Marine Button Cyanobacterium. *Journal of Natural Products* **2012,** 75, (9), 1560-1570.

61. Nogle, L. M.; Gerwick, W. H., Somocystinamide A, a novel cytotoxic disulfide dimer from a Fijian marine cyanobacterial mixed assemblage. *Organic Letters* **2002,** 4, (7), 1095-1098.

62. Andrianasolo, E. H.; Gross, H.; Goeger, D.; Musafija-Girt, M.; McPhail, K. P.; Leal, R. M.; Mooberry, S. L.; Gerwick, W. H., Isolation of swinholide A and related glycosylated derivatives from two field collections of marine cyanobacteria. *Organic Letters* **2005,** 7, (7), 1375-1378.

63. Gonzalez, R. C.; Woods, R. E., *Digital image processing*. 2nd ed. 233-237; Prentice Hall: Upper Saddle River, N.J., 2002; p xx, 793 pages.

64. Mevers, E.; Liu, W. T.; Engene, N.; Mohimani, H.; Byrum, T.; Pevzner, P. A.; Dorrestein, P. C.; Spadafora, C.; Gerwick, W. H., Cytotoxic Veraguamides, Alkynyl Bromide-Containing Cyclic Depsipeptides from the Marine Cyanobacterium cf. Oscillatoria margaritifera. *Journal of Natural Products* **2011,** 74, (5), 928-936.

65. Yang, Y. H.; Fu, X. L.; Li, L. Q.; Zeng, Y.; Li, C. Y.; He, Y. N.; Zhao, P. J., Naphthomycins L-N, Ansamycin Antibiotics from Streptomyces sp CS. *Journal of Natural Products* **2012,** 75, (7), 1409-1413.

66. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular Nmr* **1995,** 6, (3), 277-293.

67. Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; Bengio, Y.; Bergeron, A.; Bergstra, J.; Bisson, V.; Snyder, J. B.; Bouchard, N.; Boulanger-Lewandowski, N.; Bouthillier, X.; de Brébisson, A.; Breuleux, O.; Carrier, P.-L.; Cho, K.; Chorowski, J.; Christiano, P.; Cooijmans, T.; Côté, M.-A.; Côté, M.; Courville, A.; Dauphin, Y. N.; Delalleau, O.; Demouth, J.; Desjardins, G.; Dieleman, S.; Dinh, L.; Ducoffe, M.; Dumoulin, V.; Kahou, S. E.; Erhan, D.; Fan, Z.; Firat, O.; Germain, M.; Glorot, X.; Goodfellow, I.; Graham, M.; Gulcehre, C.; Hamel, P.; Harlouchet, I.; Heng, J.-P.; Hidasi, B.; Honari, S.; Jain, A.; Jean, S.; Jia, K.; Korobov, M.; Kulkarni, V.; Lamb, A.; Lamblin, P.; Larsen, E.; Laurent, C.; Lee, S.; Lefrancois, S.; Lemieux, S.; Léonard, N.; Lin, Z.; Livezey, J. A.; Lorenz, C.; Lowin, J.; Ma, Q.; Manzagol, P.-A.; Mastropietro, O.; McGibbon, R. T.; Memisevic, R.; van Merriënboer, B.; Michalski, V.; Mirza, M.; Orlandi, A.; Pal, C.; Pascanu, R.; Pezeshki, M.; Raffel, C.; Renshaw, D.; Rocklin, M.; Romero, A.; Roth, M.; Sadowski, P.; Salvatier, J.; Savard, F.; Schlüter, J.; Schulman, J.; Schwartz, G.; Serban, I. V.; Serdyuk, D.; Shabanian, S.; Simon, É.; Spieckermann, S.; Subramanyam, S. R.; Sygnowski, J.; Tanguay, J.; van Tulder, G.; Turian, J.; Urban, S.; Vincent, P.; Visin, F.; de Vries, H.; Warde-Farley, D.; Webb, D. J.; Willson, M.; Xu, K.; Xue, L.; Yao, L.; Zhang, S.; Zhang, Y., Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* **2016,** abs/1605.02688.

68. Glorot, X.; Bengio, Y., Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the International Conference on*

Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics.

69. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **2014,** 15, 1929-1958.

70. National Center for Biotechnology Information PubChem Score Matrix Service. http://tinyurl.com/hdtpe23

71. Yang, X. W.; Li, M. M.; Liu, X.; Ferreira, D.; Ding, Y.; Zhang, J. J.; Liao, Y.; Qin, H. B.; Xu, G., Polycyclic Polyprenylated Acylphloroglucinol Congeners Possessing Diverse Structures from Hypericum henryi. *J Nat Prod* **2015,** 78, (4), 885-95.

72. Bokeh Development Team Bokeh: Python library for interactive visualization. http://tinyurl.com/hzalr73

73. Shen, D. Y.; Chan, Y. Y.; Hwang, T. L.; Juang, S. H.; Huang, S. C.; Kuo, P. C.; Thang, T. D.; Lee, E. J.; Damu, A. G.; Wu, T. S., Constituents of the Roots of Clausena lansium and Their Potential Anti-inflammatory Activity. *Journal of Natural Products* **2014,** 77, (5), 1215-1223.

74. Pham, C. D.; Hartmann, R.; Muller, W. E. G.; de Voogd, N.; Lai, D. W.; Proksch, P., Aaptamine Derivatives from the Indonesian Sponge Aaptos suberitoides. *Journal of Natural Products* **2013,** 76, (1), 103-106.

75. Yu, H. B.; Yang, F.; Sun, F.; Ma, G. Y.; Gan, J. H.; Hu, W. Z.; Han, B. N.; Jiao, W. H.; Lin, H. W., Cytotoxic Aaptamine Derivatives from the South China Sea Sponge Aaptos aaptos. *Journal of Natural Products* **2014,** 77, (9), 2124-2129.

76. Chen, M. H.; Gan, L. S.; Lin, S.; Wang, X. L.; Li, L.; Li, Y. H.; Zhu, C. G.; Wang, Y. A.; Jiang, B. Y.; Jiang, J. D.; Yang, Y. C.; Shi, J. G., Alkaloids from the Root of Isatis indigotica. *Journal of Natural Products* **2012,** 75, (6), 1167-1176.

77. Tang, Y.; Fu, Y.; Xiong, J.; Li, M.; Ma, G. L.; Yang, G. X.; Wei, B. G.; Zhao, Y.; Zhang, H. Y.; Hu, J. F., Casuarinines A-J, Lycodine-Type Alkaloids from Lycopodiastrum casuarinoides. *Journal of Natural Products* **2013,** 76, (8), 1475-1484.

78. Song, W.; Si, L. L.; Ji, S.; Wang, H.; Fang, X. M.; Yu, L. Y.; Li, R. Y.; Liang, L. N.; Zhou, D. M.; Ye, M., Uralsaponins M-Y, Antiviral Triterpenoid Saponins from the Roots of Glycyrrhiza uralensis. *Journal of Natural Products* **2014,** 77, (7), 1632-1643.

79. Ibrahim, M. A.; Rodenburg, D. L.; Alves, K.; Fronczek, F. R.; McChesney, J. D.; Wu, C. M.; Nettles, B. J.; Venkataraman, S. K.; Jaksch, F., Minor Diterpene Glycosides from the Leaves of Stevia rebaudiana. *Journal of Natural Products* **2014,** 77, (5), 1231-1235.

80. Campana, P. R. V.; Coleman, C. M.; Teixeira, M. M.; Ferreira, D.; Braga, F. C., TNF-alpha Inhibition Elicited by Mansoins A and B, Heterotrimeric Flavonoids Isolated from Mansoa hirsuta. *Journal of Natural Products* **2014,** 77, (4), 824-830.

81. Cheng, Y. B.; Chien, Y. T.; Lee, J. C.; Tseng, C. K.; Wang, H. C.; Lo, I. W.; Wu, Y. H.; Wang, S. Y.; Wu, Y. C.; Chang, F. R., Limonoids from the Seeds of Swietenia macrophylla with Inhibitory Activity against Dengue Virus 2. *Journal of Natural Products* **2014,** 77, (11), 2367-2374.

CHAPTER 4

SMART ASSISTED ISOLATION AND STRUCTURAL ELUCIDATION OF VIEQUEAMIDES AND AURILIDE A, ANTI-CANCER CYCLIC DEPSIPEPTIDES FROM THE MARINE CYANOBACTERIA *RIVULARIA SP.* AND *MOOREA SP.*

4.0 Abstract

Combining molecular networking and bioassay-guided fractionation enabled the targeted isolation of four anti-cancer cyclic depsipeptides, viequeamides A, A2, and A3 (**1-3**) and aurilide D (**4**) from the extracts of a collection of a marine cyanobacterium (*Rivularia sp.*) originating from the vicinity of Vieques, Puerto Rico. Besides, three known viequeamides B, C and D (**8-10**) were obtained from a marine blackish green alga collected from American Samoa, identified as the benthic cyanobacterium (*Moorea producens*). Their planar structures and absolute configurations were dereplicated and/or elucidated by a deep learning aided comprehensive analysis of spectroscopic and chromatographic data. Metabolite **2** is distinctive from known agent **1** by hydrogenation on the terminal acetylene residue on the 2,2-dimethyl-3-hydroxy-7-octynoic acid (Dhoya) domain, whereas in metabolite **3** presents a 2,2-dimethyl-3-hydroxy-7-octanoic acid (Dhoaa) domain. Aurilide D (**4**) is an epimer of known compound aurilide A (**5**) by altering the stereocenter of the methylleucine domain. The absolute structures of **9** and **10** were determined by referring to the stereocenters within the crystal structure of their analogue **8**. Compounds **1-4** demonstrate strong cytotoxicity against the H-460 human lung cell line (IC_{50} = 4.23 $\pm$ 0.171 μM, **1**; IC_{50} = 0.62 $\pm$ 0.046 μM, **2**; IC_{50} = 1.98

± 0.038 μM, **3**; IC_{50} = 2.3 ± 0.10 nM, **4**, respectively), while compounds **8-10** were inactive to H460 cells at 1.0 μg/mL.

4.1 Introduction

Lung cancer continues to be the leading cause of cancer death in the United States and worldwide for the past ten years. Cancer statistics showed that nearly as many Americans, regardless of age, die of lung cancer every year as die of prostate, breast, and colon cancer combined[1]. Despite success of chemotherapy and targeted biological therapies, remarkable limitations in drug safety and drug efficacy still exist among current medical treatments. Therefore, additional anti-cancer agents with different modes of action are needed, and it is thus hoped that novel data technology based marine natural products (MNPs) research can aid in this search[2].

Previously, a variety of data technology based natural products discovery approaches were created, and thus swiftly employed to streamline the screening of chemotypes of natural products. One highlighted approach bridged chemotypes of peptidic and glycosylated natural products to their biosynthetic pathways using MS (Mass Spectrometry)-guided genome mining[3, 4]. A recent approach capitalized on the richness of MS data of natural products via an online digital curation platform (*i.e.* Global Natural Products Social Molecular Networking (GNPS))[5], which facilitated high-throughput small peptide dereplication[6]. Considering that NMR provides information that is non-accessible by other spectroscopic methods, efforts have also been made to automate analysis of 2D NMR spectra collected using natural product mixtures[7, 8]. Additionally, the effectiveness of neural networks and least-squares

regression analysis with regard to 1D NMR chemical shift prediction has been compared[9]. In our previous study, we have already leveraged the benefits of HSQC NMR spectroscopy with a deep learning technique, Convolutional Neural Networks (CNN), to create the Small Molecule Accurate Recognition Technology (SMART) as a tool to associate unknown HSQC spectra to their known analogues[10]; however, its practical use remains to be demonstrated in our ongoing anti-cancer drug discovery projects.

Cyclic depsipeptides of the kulolide superfamily were isolated from marine nudibranchs and marine cyanobacteria, and have demonstrated a variety of anti- cancer bioactivities *in vitro*[11]. A subfamily of the cyclic depsipeptides, the viequeamides, were first extracted from a marine button cyanobacterium (*Rivularia sp.*) in the vicinity of Vieques, Puerto Rico. It was previously subjected to toxicity evaluation against H-460 human lung cancer cell lines. Intriguingly, viequeamide A (**1**), with a (2R,3S)-2-hydroxy-3-methylpentanoic acid (2R,3S-Hmpa) moiety and a threonine (Thr) moiety, was much more cytotoxic against H460 human lung cancer cells (IC_{50} = 60 $\pm$ 10 nM) than other members of the viequeamide subfamily[12]. The outstanding nanomolar level cancer cell toxicity of viequeamide A (**1**) thus triggered endeavors for its total synthesis, and further evaluation of the bioactivity of both the natural product and its synthetic counterparts have ensued. The first total synthesis of viequeamide A (**1**) was reported by Wang et al. in 2013[13]. Unfortunately, compared to the isolated natural product, the bioactivity of the synthetic viequeamide A (**1**) against H-460 human lung cancer cells was exceptionally poor (IC_{50} up to 100 μM). In response to the two contradictory results previously reported, a detailed reinvestigation of the bioactivity of isolated viequeamide

A (**1**) was launched within our group. The newly isolated viequeamide A (**1**) from the same cyanobacterial crude extract as previously reported[12] demonstrated strong H-460 cytotoxicity (IC_{50} = 60 $\pm$ 10 nM). Furthermore, also from the same crude extract of the cyanobacterium, three novel cyclic depsipeptides were identified using the GNPS[5], followed by the SMART[10].

Prompted by the discovery of those new cyclic depsipeptides, we carried out a more in-depth investigation of structural variations within the viequeamide family from other marine cyanobacteria using state-of-the-art dereplication tools. In this regard, LC-MS/MS-based bioinformatic evaluation of a crude CH_2Cl_2/MeOH extract of an American Samoa collection of *Moorea sp.* led to the detection of a series of viequeamides. By applying the deep learning based HSQC spectra recognition tool, SMART[10], the investigation of the extract rapidly disclosed the presence of three additional viequeamides.

Here, we report the isolation, structural elucidation and bioactivity investigation of viequeamide A, A2, A3 (**1-3**), and aurilide D (**4**) that were obtained in *Rivularia sp.* collected from Vieques, Puerto Rico during a follow up study of the bioactivity of viequeamide A (**1**), as well as viequeamide B, C, and D (**8-10**) isolated from an extract of dark-green cyanobacerium, *Moorea producens*, collected in American Samoa. (See Figure 4.1 for chemical structures) Dereplication of **1** and **8**, and structural determination of **2**, **3**, **4**, **9**, and **10**, was based on SMART analysis of 2D NMR data, as well as confirmation by X-ray crystallography. Subsequent evaluation to H-460 human lung cancer cells confirmed the moderately cytotoxic properties of viequeamide A (**1**) (IC_{50} = 4.23 $\pm$ 0.171 μM), as derived from the cyanobacterium *Rivularia sp.*, and also

unveiled three moderate to highly potent anti-cancer cyclic depsipeptides, viequeamide A2 (**2**) (IC_{50} = 0.62 $\pm$ 0.046 μM), A3 (**3**) (IC_{50} = 1.98 $\pm$ 0.038 μM) and aurilide D (**4**) (IC_{50} = 2.3 $\pm$ 0.10 nM). These metabolites have the structural hallmarks of deriving from a combination of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) biosynthetic pathways[14, 15]. In contrast, viequeamide B (**8**), C (**9**) and D (**10**) originated from the cyanobacterium *Moorea sp.*, and were inactive to H-460 cells at 1.0 μg/mL.

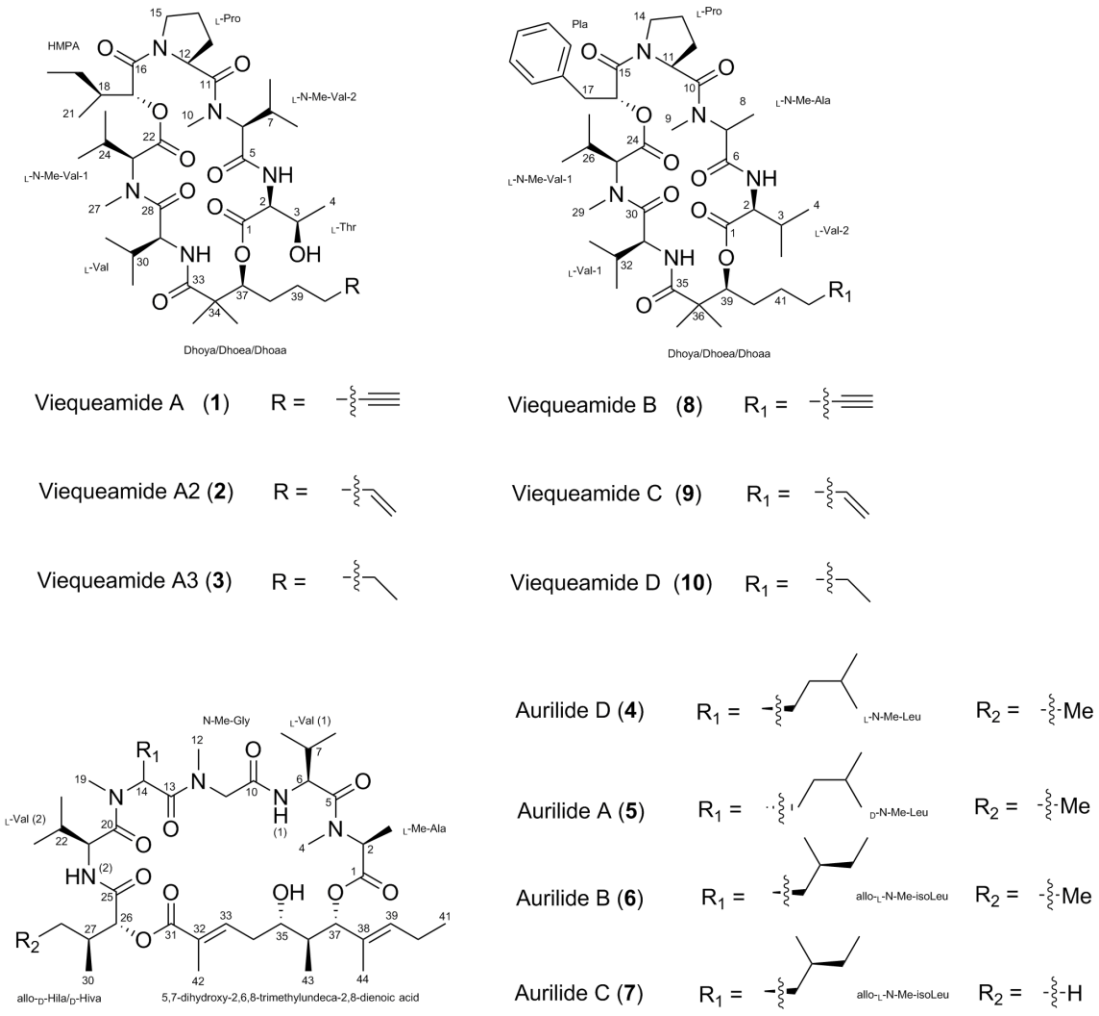


Figure 4.1. Structures of Viequeamides A-A3 (**1-3**), B-D (**8-10**), Aurilide D (**4**), and Aurilide A-C (**5-7**).

4.2 Results and Discussion

The VLC fraction eluting with ethyl acetate from chromatography of the extract of *Rivularia sp.*[12] was further purified over a self-packed C_{18} SPE column, giving rise to two SPE fractions. The relatively non-polar SPE fraction was preserved at -80 ^{o}C before being defrosted at 21 $\pm$ 1 ^{o}C.

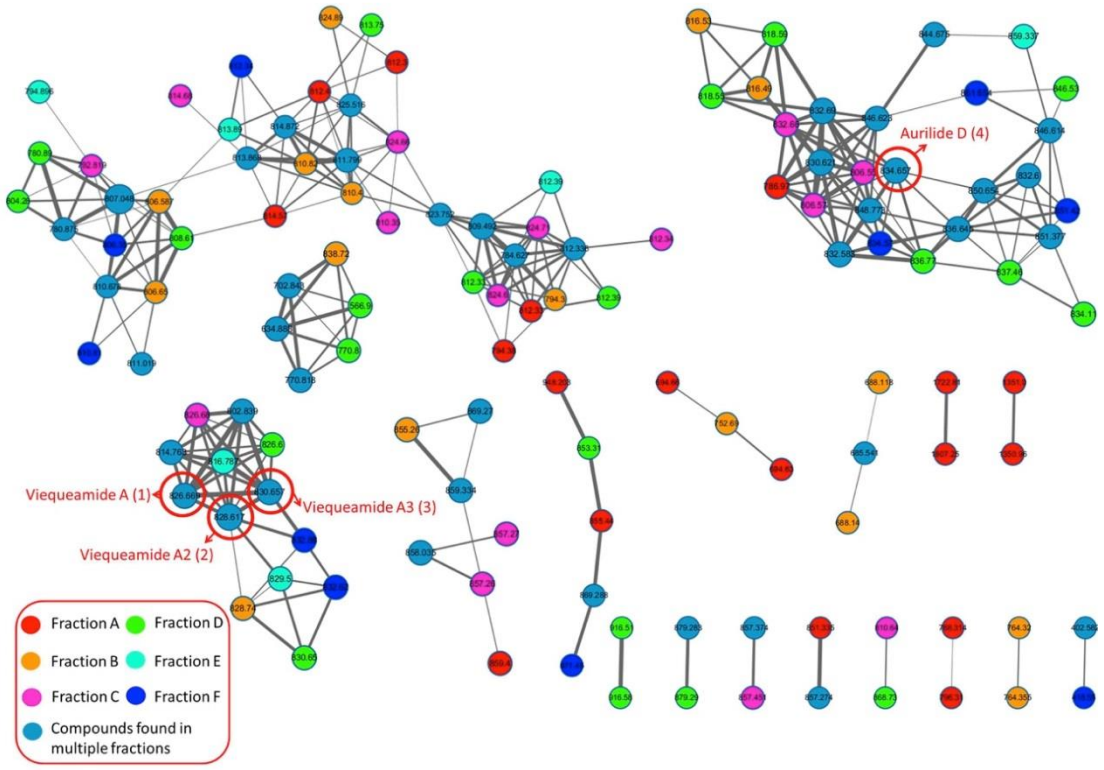


Figure 4.2. The molecular network of all prefractions of the *Rivularia sp.* sample as visualized in Cytoscape 3.1. Each node stands for a MS/MS spectrum obtained via LCMS and is labeled with the precursor parent mass. Each edge stands for quantified correlation between the two MS/MS spectra. The edge thickness indicates the cosine similarity values. Nodes are colored according to fractions collected at different retentions time during the HPLC fractionation of the sample, as noted in the lower right section of the chart. (See Chapter 4 Appendix for chromatogram and retention time for each fraction)

Molecular networking containing MS/MS data of the relatively non-polar fractions was constructed using GNPS[5, 16]. (See Figure 4.2) In this molecular network, the nodes showing precursor MS1 data from the fraction were differently colored. Overlapping experimental nodes with the same precursor MS1 and MS2 data from the GNPS database are circled in red. The strength of correlations between the two MS2 precursor data are denoted with edges, whose labels reflecting cosine values. The dereplicated nodes of the SPE fraction derived viequeamide A (**1**) (*m/z* 804.78 $[M+H]^+$)

along with viequeamide A2 (**2**) (*m/z* 806.82 $[M+H]^+$), and viequeamide A3 (**3**) (*m/z* 808.87 $[M+H]^+$) clustered with **1** (*m/z* 826.40 $[M+Na]^+$) in the GNPS database. Additionally, their related MS/MS fragmentation patterns were compared with that of **1** in the database to verify that the fragmentation patterns resembled each other, respectively. (See Figure 4.3) This SPE fractions were then subjected to further reversed-phase HPLC purification, which finally afforded viequeamide A (**1**, 5.2 mg), viequeamide A2 (**2**, 1.8 mg), viequeamide A3 (**3**, 1.3 mg), and aurilide D (**4**, 2.0 mg).

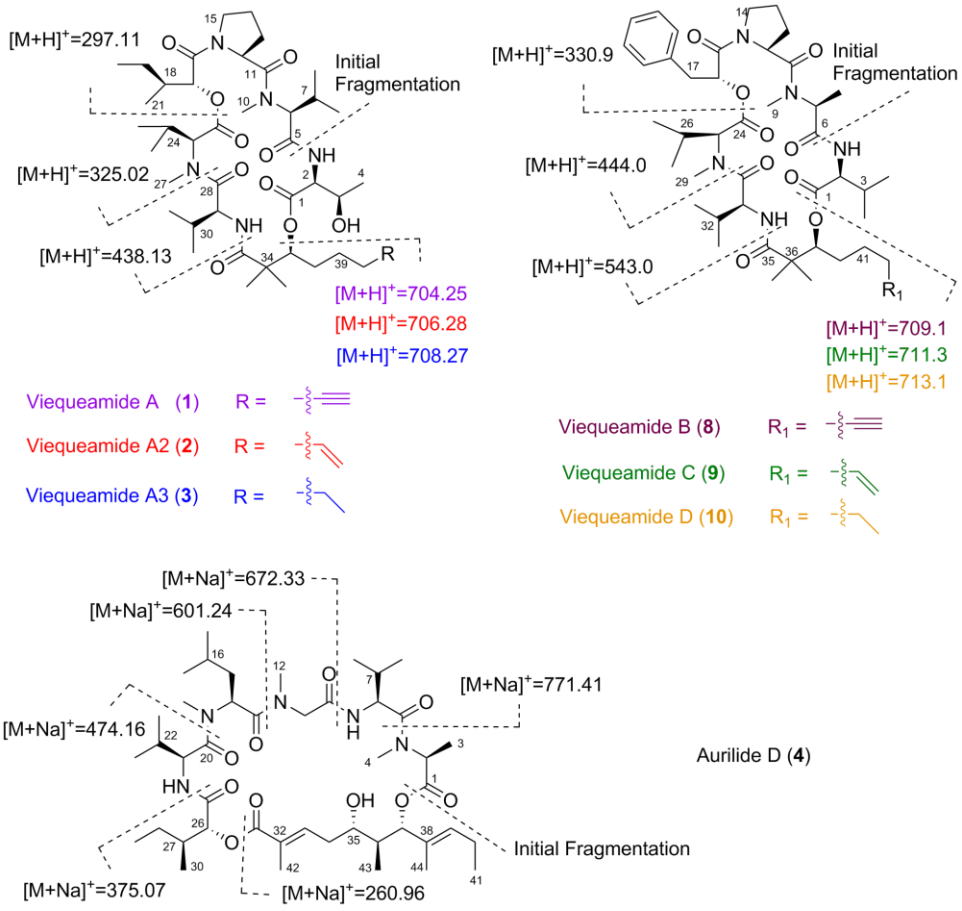


Figure 4.3. MS/MS fragmentation patterns of viequeamide A2 (**2**), viequeamide A3 (**3**), aurilide D (**4**), viequeamide C (**9**), viequeamide D (**10**).

A detailed analysis of the HRMS data of viequeamide A (**1**) revealed a chemical with *m/z* 826.4940 $[M+Na]^+$, (Δ 0.4 mmu), giving rise to the molecular formula $C_{42}H_{69}N_5O_{10}$ (Ω = 11). The ^1H NMR spectra in $CDCl_3$ demonstrated the presence of two amide NH groups at δ 8.04 (NH, threonine) and 6.83 (NH, valine) and 2 *N*-methylamide groups at δ 2.98 (H_3-27), and 2.77 (H_3-10), revealing compound **1** as containing peptide sections. The ^{13}C NMR spectrum in $CDCl_3$ contained 7 carbonyl groups with chemical shifts at δ 174.7 (C33), 172.8, (C28), 172.1 (C11), 170.3 (C22), 169.2 (C1), 168.7 (C16), 168.2 (C5), which accounted for 7 unsaturated degrees, leaving unassigned 4 degrees of unsaturateation in the compound.

The HSQC spectra of pure compounds **1** - **3** were subjected to dereplication using the Small Molecule Accurate Recognition Technology (SMART) platform[10]. The 10 dimensional test results are shown in the Appendix of Chapter 3. Three significant observations were made from this data list. First, compounds **1** and **2** were closely associated with the molecular family that were previously labeled as "viequeamides" in the training set. Second, compound **3** clustered with the "veraguamides" molecular family and thus revealed its peptidic nature. Nevertheless, follow-up analysis of the HSQC spectra of compound **3** suggested that it might still be a viequeamide analogue. Finally, multiple molecular families in the top 50 hits of these three newly isolated compounds had no peptidic feature. In the full list of the top 50 closest compound families, we highlight those blocks containing the same peptidic family of compounds. Further structural and biochemical interrogation of the three compounds is described in detail herein.

A detailed analysis of the 2D NMR data sets, including DQF-COSY, HSQC, HSQC-TOCSY and HMBC, allowed assembly of the peptide and polyketide sections of compound **1**. Generally, by inspection of HSQC-TOCSY, COSY and HMBC, **1** was confirmed to consist of five amino acids; one valine, two *N*-methylvalines, one proline and one threonine, plus one 2-hydroxy-3-methylpentanoic acid (HMPA, leucic acid) and one 2,2-dimethyl-3-hydroxy-7-octynoic acid (Dhoya) moiety. HMBC correlations between C11 (δ 172.1) and the *N*-methyl protons (δ 2.77) on the *N*-methylvaline (2, C5-10) indicated that the first *N*-methylvaline (2, C5-10) was linked to the proline residue (C11-15). HMBC correlation between the α-proton (δ 4.68) of theronine (C1) and the C5 (δ 168.2) on the carbonyl group of the *N*-methylvaline linked these two residues together via an amide bond. The second *N*-methylvaline (C22-27) was linked with the HMPA residue (C16-21) via an ester bond and the valine residue (C28-32) throught its *N*-terminus via an amide bond. In turn, the valine residue (C28-32) was connected with the Dhoya (C33-42) through an amide bond. Specifically, the α-proton (δ 4.89) of this valine was correlated with the carbonyl C33 (δ 174.7) of the Dhoya residue. The planar structure of the Dhoya domain, whose terminal alkyne accounted for 2 degrees of unsaturation, was solved by HSQC-TOCSY and HMBC correlations. Because the α-hydroxy proton H-37 (δ 5.55) correlated with the carbonyl C1 (δ 169.2) of the threonine residue, it was deduced that these residues were connected via an ester bond. The ROESY correlation between the α-proton of the HMPA and (H_2-15, δ 3.95 and 3.52) on the proline moiety thus closed the macrocyclic structure of compound **1**, and contributed to one more unsaturated degree. In addition, the three facts that, 1) in TOCSY spectra, H-12, H_2-13, H_2-14 and H_2-15 were in single spin system, 2) the *N*-

terminal of the proline was a tertiary amide, and 3) a closed ring accounted for the remaining 1 unsaturated degree were indicative of the five membered proline ring structure (C11-15). A careful comparison of the ^{1}H and ^{13}C NMR data for compound **1** with that published for viequeamide A showed that they were identical.

Although the newly isolated **1** showed similar optical rotation data ($[\alpha]_D^{23}$ -32.6) compared with previously isolated **1** ($[\alpha]_D^{23}$ -32.6), the bioassay results for the natural product and a published report for synthetic viequeamide A[13] were significantly different. Therefore, we elected to unequivocally determine the stereochemistry of **1** via a single-crystal X-ray diffraction study. Crystals were obtained by slowly diffusing hexane into an acetone solution of **1** until hexane to acetone ratio was 1:9, yielding a well formed crystal of $0.33 \times 0.30 \times 0.08$ mm^3 over 16 hours. The planar structure and all stereocenters of **1** were firmly established (Figure 4.4 Left) with the absolute configuration of 2*S*, 3*R*, 6*S*, 12*S*, 17*R*, 18*S*, 23*S*, 29*S*, 37*S*, via refinement of the absolute structure parameter [0.11(5)].

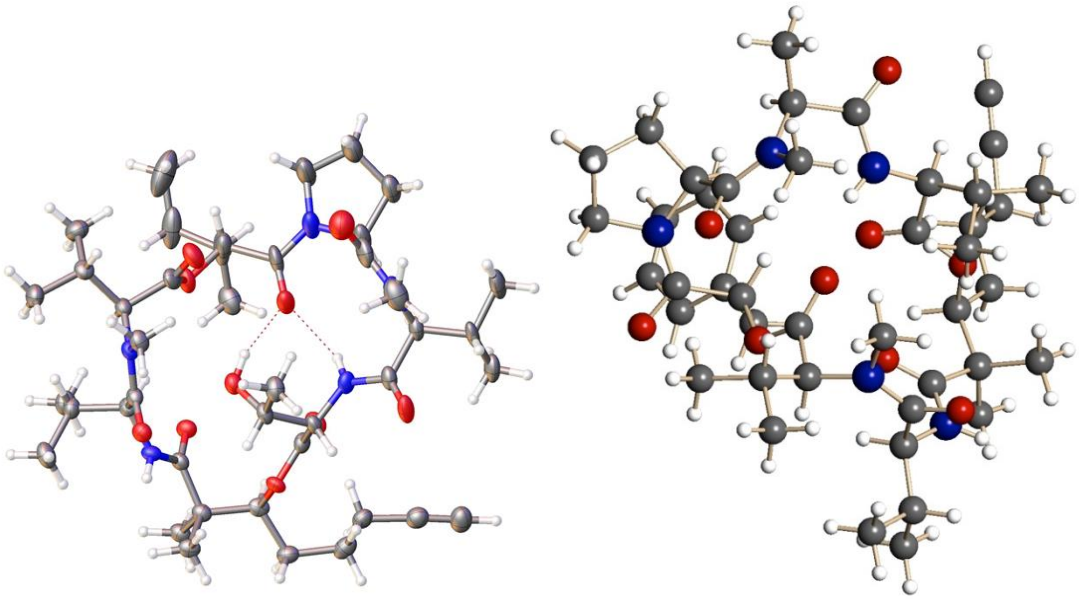


Figure 4.4. X-ray crystallographic results of viequeamide A (**1**) (left) and viequeamide B (**8**) (right).

The HR ESI-TOF-MS data of viequeamide A2 (**2**) [*m/z* 828.5087 $[M+Na]^+$, (Δ -0.7 mmu)] and A3 (**3**) [*m/z* 830.5245 $[M+H]^+$, (Δ -0.6 mmu)] gave rise to the molecular formulas of the two new compounds as $C_{42}H_{71}N_5O_{10}$ (Ω = 10) and $C_{42}H_{73}N_5O_{10}$ (Ω = 9). By analysis of the 1D and 2D NMR spectra for viequeamide A2 (**2**) and A3 (**3**) in comparison with that of viequeamide A (**1**) (See Figure 4.5), the only variation between the three compounds was the residue of Dhoya moiety in viequeamide A (**1**) was substituted with a 2,2-dimethyl-3-hydroxy-7-octenoic acid (Dhoea) in viequeamide A2 (**2**), and a 2,2-dimethyl-3-hydroxy-7-octanoic acid (Dhoaa) in viequeamide A3 (**3**).

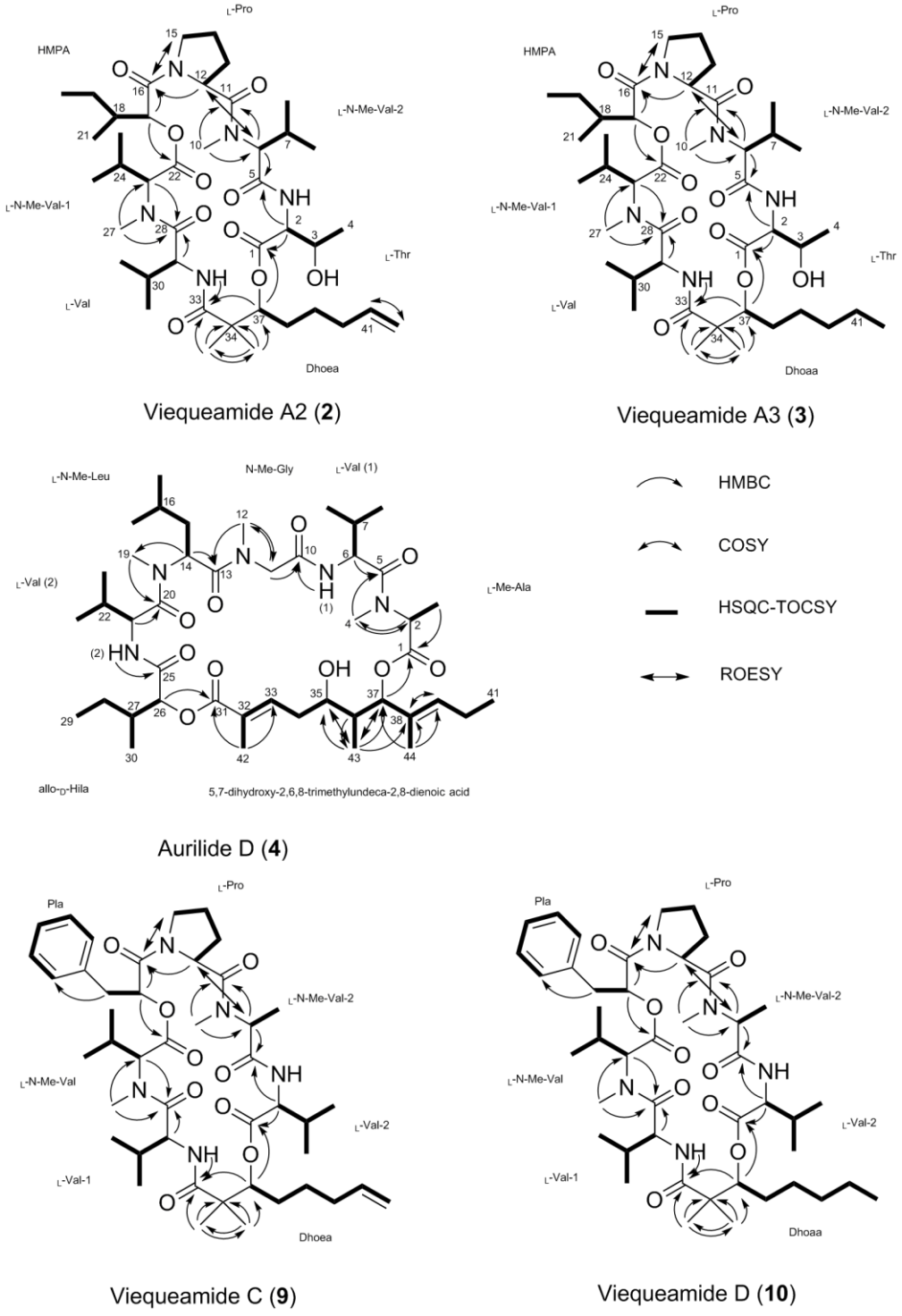


Figure 4.5. Key 2D NMR correlations of viequeamide A2 (**2**), viequeamide A3 (**3**), aurilide D (**4**), viequeamide C (**9**), and viequeamide D (**10**).

Because the Dhoya residue can be biosynthetically formed from a desaturase acting on either a Dhoea or Dhoaa, similar to the formation of the acetylene residue formed by JamB in jamaicamide B,[17, 18] and because this process does not alter the C-37 stereocenter of compounds **2** or **3**, we predict the stereochemistry of the corresponding residues to be 37*S*. This is similar to what has also been observed in other series of cyanobacterial compounds, such as the kulolides[19], pitipeptolides[20, 21], antanapeptins[22], and cocosamides[23]. Since the newly isolated **2** ($[\alpha]_D^{25}$ -31.8) and **3** ($[\alpha]_D^{23}$ -32.3) showed similar optical rotation data compared with **1** ($[\alpha]_D^{23}$ -32.6), also isolated from this same cyanobacterium, the configuration of the amino acids in **2** and **3** were also assigned to be L and the Hmpa residue to be 17*R*, 18*S*.

The molecular formula of aurilide D (**4)** was determined as $C_{44}H_{75}N_5O_{10}$ (Ω = 10) on the basis of HR ESI-TOF-MS [*m/z* 856.5402 $[M+Na]^+$, (Δ -0.5 mmu)]. The IR absorption peaks at 3352.6 (O-H st), 1739.5 (C=O st), 1677.8 (C=O st or NH δ and N–C=O st sy), 1644.0 (C=O st or NH δ and N–C=O st sy), and 1251.6 (C–N st) cm^{-1} in were assigned to hydroxy groups, esters, and/or amides, respectively. The ^{1}H NMR spectra in $CDCl_3$ showed the presence of two amide NH groups at δ 7.38 (NH(1)) and 6.67 (NH(2)) and 3 *N*-methylamide groups at δ 3.22 (H_3-4), 2.96 (H_3-19), and 2.90 (H_3-12), revealing a peptidic component to compound **4**. The ^{13}C NMR spectrum in $CDCl_3$ offered evidence for 7 carbonyl groups with chemical shifts at δ 172.0 (C20), 171.9 (C5), 170.3 (C1), 170.3 (C25), 170.0 (C13), 169.0 (C10), 168.9 (C31), leaving unassigned 3 degrees of unsaturation. A comprehensive 2D NMR study including DQF-COSY, HSQC HSQC-TOCSY and HMBC spectra acquired in $CDCl_3$ allowed assignment of all resonances from the ^{1}H NMR and ^{13}C NMR spectra and gave rise to

a planar structure with that was identical to the previously reported metabolite aurilide A (**5**). Generally, *N*-methylleucine (C13-19) was connected with the *N*-methylglycine (C13-12) at its C-terminus and a valine (2, C20-24) at its *N*-terminus via amide bonds. HMBC correlations between C10 (δ 169.0) and the proton NH(1) (δ 7.38) on the secondary amide indicated that this *N*-methylglycine (C13-12) was linked by the other valine (1, C5-9). The proton NH(2) (δ 6.67) on valine (2, C20-24) was correlated with C25 (δ 170.3) on the Hmpa in HMBC spectra, which showed the *N*-terminus of the valine (2, C20-24) was connected with the carboxyl group on the HMPA via an amide bond. The C5 (δ 170.3) on the valine (1, C5-9) was correlated with the methyl protons H_3-4 (δ 3.22), which was assigned to the *N*-methyl group of the *N*-methylalanine (C1-4). A dihydroxy acid moiety (C31-C44) was assembled with HMBC, DQF-COSY and HSQC-TOCSY correlations (See Figure 4.5), which closed the cyclic peptide ring of aurilide D (**4**) with the 3 remaining unsaturated degrees assigned. This deduction was proven by HMBC correlations between the α-proton H26 (δ 4.75) on the HMPA residue and C31 (δ 168.9) on the carbonyl group of dihydroxy acid moiety, as well as between the C1 carbonyl carbon (δ 170.3) of the *N*-methylalanine residue and the proton H37 (δ 4.98) on the oxygenated methylene C37 (δ 82.6).

The absolute stereochemistry problem of compound **4** was initially approached by comparing the chemical shifts and coupling constants of the peaks in the ^{1}H NMR and ^{13}C NMR spectra recorded in deuterated benzene so as to be fully comparable with previous reports.[24] Notably, in deuterated benzene, the coupling constant $^{3}J_{H_6,H_7}$ between the protons H-7 (δ 2.08, m) and H-6 (δ 4.92, dd) was smaller for **4** (5.6 Hz) than reported for aurilide A (**5)** (7.0 Hz) (See Figure 4.1 for the chemical structure), while the coupling

constant $^{3}J_{H_6,H_{NH_2}}$ between the protons H-6 (δ 4.92, dd) and NH(1) (δ 7.38) was larger for **4** (9.2 Hz) than **5** (7.0 Hz). This inconsistency of experimental data for compound **4** with the literature for compound **5** resulted in a detailed inspection of all stereocenters of compound **4**.

In order to determine the absolute stereochemistry of the five amino acid residues in **4**, the acid hydrolysates were analyzed by chiral LCMS[25]. The absolute stereochemistry of the four amino acids, *N*-methylalanine, valine (1), *N*-methylleucine, and valine (2) were determined to be all L. Part of the hydrolysate of **4** was also used to determine the absolute configuration of the HMPA residue by chiral GCMS, the stereochemistry of which was assigned as allo-D. Thus, aurilide D (**4**) differs from aurilide A (**5**) by a change in the stereochemistry of the *N*-methylleucine residue.

ROESY correlations were observable from H_3-43 (δ 0.74) to H-35 (δ 3.81) and H-37 (δ 4.98). Also, the ^{1}H NMR and ^{13}C NMR chemical shifts and proton coupling constants within the 5,7-dihydroxy-2,6,8-trimethylundeca-2,8-dienoic acid moiety in **4** and the known compound **5** were identical. Therefore, the absolute configuration of the 5,7-dihydroxy-2,6,8-trimethylundeca-2,8-dienoic acid moiety of **4** can safely be predicted as 32*E*, 35*S*, 36*S*, 37*S*, 38*E*.

Aurilide A (**5**) was first isolated from the internal organs of the Japanese sea hare *Dolabella auricularia*, a sea slug that feeds on marine cyanobacteria[26] and collected from the coast of Azuri, Shima peninsula in Mie Prefecture, Japan, by Suenaga et al..[27] Bioactivity evaluations showed **5** was highly toxic against HeLa S3 tumor cells (IC_{50} = 13.2 nM), with its potency against other cancer cell lines also reported.[24] Interestingly, within the same report, an epimer of **5**, 6-epi-aurilide A was reported with a much

reduced cytotoxic effect against HeLa S3 tumor cells (IC_{50} > 4.8 μM). Mechanism of action studies showed that **5** activated the proteolysis of optic atrophy 1 (OPA1) protein by selectively binding to prohibitin 1 (PHB1) in the mitochondria, which triggered further mitochondria-induced apoptosis.[28, 29] Two analogues of **5**, aurilide B (**6**) and C (**7**), were later isolated from a marine cyanobacterium, *Lyngbya majuscula*, collected in Alotau Bay, Papua New Guinea. Compound **6** was found more toxic to both H-460 human lung cancer cell lines (LC_{50} = 0.01 μM) and neuro-2a mouse neuroblastoma cells cell lines (LC_{50} = 0.04 μM) than compound **7** (LC_{50} = 0.05 μM for H-460 cells and LC_{50} = 0.13 μM against neuro-2a cells). Compound **6** was also found to highly toxic to other cancer cell lines, such as from leukemia, renal, and prostate cancers, with GI_{50} values less than 10 nM.[30] The structure of **6** differed from that of **5** by the substitution of D-methylleucine in **5** with the allo-L-methylisoleucine in **6**.

Three additional analogues of compounds **1**-**3**, namely, viequeamide B, C and D (**8**-**10**), were detected in the extract of the *Moorea producens* sample by clustering their MS spectra with existing cyanobacterial networks[16] using GNPS. MS/MS spectra of **8**-**10** were manually compared with published results[12] so as to verify their fragmentation patterns. HPLC purification of the three relatively polar VLC fractions (E-G), which were highlighted by the molecular networks, afforded three pure compounds **8**-**10**. HR ESI-TOF-MS data of viequeamide B (**8**) [*m/z* 830.4668 $[M+Na]^+$, $C_{44}H_{65}N_5O_9Na$ (Δ -0.8 mmu)], viequeamide C (**9**) [*m/z* 832.4823 $[M+Na]^+$, $C_{44}H_{67}N_5O_9Na$ (Δ -0.99 mmu)], and viequeamide D (**10**) [*m/z* 834.4980 $[M+Na]^+$, $C_{44}H_{69}N_5O_9Na$ (Δ -0.92 mmu)] confirmed the molecular formula of the three compounds. 2D HSQC spectra were then generated for the three new compounds

dissolved in $CDCl_3$. Again, after converting the collected HSQC spectra into .PNG format, we exposed these to the pre-trained SMART system containing over 2,000 HSQC training spectra in order to facilitate cdetermination of their planar structures. The clustering result showed that compounds **8** and **10** were again associated with the "viequeamides" family in the training dataset. Unfortunately, due to the small number of viequeamides in the training set (only 2), compound **9** was not closely associated with this family, at least not in the 2D visualization.

Because compound **8** was found to provide well-formed crystals from MeOH, an X-ray diffraction study was initiated. The results of this study revealed that the constitutive and stereostructure of this newly isolated compound was the same as the previously reported[12]. Since the HSQC spectra of **9** and **10** were recognized by the SMART to be closely related to that of **8**, the planar structures of **9** and **10** were deduced by comparing the ^{1}H and ^{13}C NMR data with those of **8**. Compounds **9** and **10** were thus found to be analogous to **8** by varying the levels of unsaturation of the alkyne group in the Dhoya residue, similar to the situation above with compounds **1**-**3**. Specifically, in viequeamide B (**8**), the protons H_2-42 (δ 2.24, m) on the distal methylene group correlated with acetylenic carbons C-43 (δ 83.53) and C-44 (δ 69.29) by HMBC, and H44 (δ 1.93, t) correlated with C-44 (δ 69.29) by HSQC, revealing that this residue terminated with an alkyne group. In viequeamide C (**9**), the analogous H_2-42 protons (δ 2.08, m) correlated with the olefin carbons C-43 (δ 138.13) and C-44 (δ 115.38) by HMBC, and terminal H_2-42 protons correlated with H-43 (δ 5.70-5.75, m) and H_2-44 (δ 5.01-4.91, m) in the TOCSY spectrum. Consistently, the H-43 and H_2-44 protons

correlated with C-43 (δ 138.13) and C-44 (δ 115.38) by HSQC, revealing that compound **9** terminated with an alkene group.

Similarly, analysis of the 2D NMR correlations of viequeamide D (**10**) identified that it contained the saturated Dhoaa residue. In **10**, H_2-42 (1.25, m) correlated with H_2-43 (δ 1.26, m) and H_3-44 (δ 0.82, t) by TOCSY. By HSQC the H_2-43 and H_3-44 protons correlated with C-43 (δ 22.65) and C-44 (δ 14.13), respectively. The 37*S* configurations of **9** and **10** were predicted based on the assumption that the biosynthetic processes did not alter any of the stereocenters of the three analogues, as discussed above for compounds **2** and **3**. Because the newly isolated compounds **9** ($[\alpha]_D^{25}$ -82.2) and **10** ($[\alpha]_D^{25}$ -80.1) showed similar optical rotation data compared with **8** ($[\alpha]_D^{25}$ -85.8) isolated from the same cyanobacterium sample, the configurations of all of the amino acids are predicted to be L as well.

Viequeamides A, A2, A3 (**1**-**3**), aurilide D (**4**), and viequeamides B, C, D (**8**-**10**) were evaluated for cytotoxic activity to H-460 human lung cancer cells. Intriguingly, viequeamide A2 with a Dhoea moiety was around 3-fold more toxic (IC_{50} = 0.62 $\pm$ 0.046 μM) than A3 (IC_{50} = 1.98 $\pm$ 0.038 μM) with a Dhoaa moiety, and around 7-fold more toxic than A (IC_{50} = 4.23 $\pm$ 0.171 μM) with a Dhoya moiety. The IC_{50} for aurilide D was 2.3 $\pm$ 0.10 nM to H-460 cells, approximately 6-fold more toxic than its epimer, aurilide A[27]. With their strong *in vitro* cancer cell toxicity, it is thus hoped that viequeamide A2, A3 and aurilide D strengthen the growing arsenal of potential cyanobacterial-derived preclinical anticancer drug leads.

4.3 Conclusions

A chemical investigation into cytotoxic extracts of two marine cyanobacteria, a *Rivularia* sp. from Vieques, Puerto Rico, and a *Moorea producens* from American Samoa, led to the isolation of a series of depsipeptides, namely viequeamides A-A3, B-D and aurilide D. The planar structures of the viequeamides were established using a combination of the MS/MS-based dereplication algorithm, GNPS, and the 2D NMR-based deep learning tool, SMART. The planar structure of aurilide D was solved by using conventional MS and 2D NMR data. The absolute configuration was studied for all compounds using a combination of Marfey's analysis and X-ray crystallography. Viequeamides A2 and A3 are analogous to viequeamide A, with the terminal alkyne of viequeamide A changed to either an alkene or alkane in each case. Similarly, viequeamide B, C, and D are also structural analogues with the terminal alkyne being replaced with either an alkene or alkane. Aurilide D differs from the previously reported aurilide A by a change in the stereochemistry of the *N*-methylleucine residue. As a result, aurilide D exhibited even more potent cytotoxicity against H-460 human lung cancer cells with an IC_{50} of 2.3 ± 0.10 nM.

4.4 Experimental Methods

4.4.1. General Experimental Procedures

Optical rotations were measured on a Jasco P-2000 polarimeter. IR spectra were measured on a Thermo Electron Corporation Nicolet IR 100 FT-IR. UV/visual-light spectra were recorded on a Beckman Coulter DU 880 spectrophotometer. NMR spectra were collected on the Varian Unity 500 MHz spectrometers (500 and 125 MHz for the

^{1}H and ^{13}C nuclei, respectively) with the solvent $CDCl_3$ containing 0.03% v/v trimethylsilane (δH 0.0 and δC 77.16 as internal standards using trimethylsilane and $CDCl_3$, respectively). Non Uniform Sampling (NUS) HSQC spectra was recorded on the 600 MHz Bruker Avance III DRX600 600 MHz spectrometer with a 1.7 mm Bruker TXI MicroCryoProbe™ using TopSpin 2.1. The solvent $CDCl_3$ contained 0.03% v/v trimethylsilane (δ_H 0.0 and δ_C 77.16 as internal standards using trimethylsilane and $CDCl_3$, respectively). All spectra were recorded with the sample temperature at 298 K. High-resolution mass spectra were collected on an Agilent 6230 TOF-MS under positive ion ESI-TOF-MS conditions and provided by the University of California, San Diego (UCSD) Small Molecule MS Facility. Chiral LC-MS analysis was accomplished on a Thermo Finnigan LCQ, operating in positive ion ESI (Electrospray Ionization) mode, coupled to a Thermo Finnigan Surveyor Plus liquid chromatography system with a Phenomenex Kinetex 5 μm C_{18} 100 Å column (4.60 × 100 mm). HPLC was performed using chromeleon 7 software with Thermo dionex ultimate 3000 pump and a RS diode array detector. All solvents were HPLC grade except for water, which was produced by a Millipore Milli-Q system. Chiral MS data were obtained with chemical standards obtained from Sigma-Aldrich.

4.4.2. Cyanobacterial Collections and Morphological Identification

The collection and morphological study of the cyanobacterium *Rivularia sp.* was reported previously[12]. The cyanobacterium *Moorea producens* in this study was blackish green, growing on a coral reef at 5 to 10 feet in Fagaalu Park, American Samoa. (See Figure 4.6) Collection of the *Moorea* sample was made by hand using snorkeling

gear in July of 2014. Morphological study was performed using an Olympus IX51 epifluorescent microscope (1000× optical zoom) equipped with an Olympus U-CMAD3 camera. Morphological comparison and putative taxonomic identification of the cyanobacterial specimen was performed in accordance with modern classification systems[31].

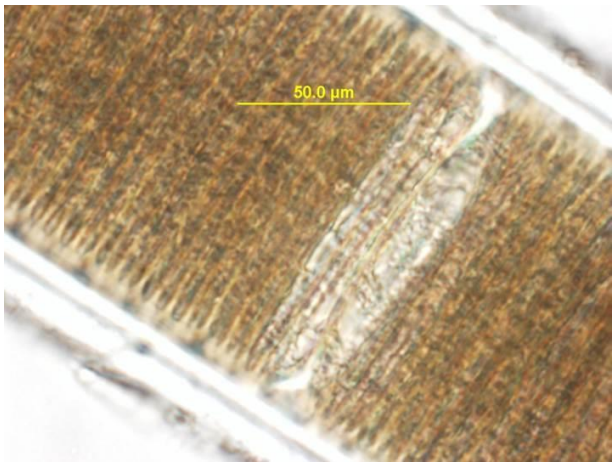


Figure 4.6. A microscopic image of the American Samoa *Moorea producens* collection.

4.4.3. Extraction and Isolation

The cyanobacterium *Rivularia sp.* was extracted and fractionated with vacuum liquid chromatography (VLC) as described previously[12]. The EtOAc eluted VLC fraction was subjected to a self-packed C_{18} solid-phase extraction (SPE) with a Strata 6 mL column and 1 g of C_{18}-E (55 µm, 70 Å). The EtOAc eluted fraction was further fractionated with 30 mL 50% acetonitrile (ACN)/50% H_2O, followed by 30 mL of 100% ACN, yielding a relatively polar fraction and a relatively non-polar one. The non-polar fraction was kept frozen at -80 ºC before being defrosted at 21 ± 1 ºC. The non-polar

fraction was eluted with a Synergi 4 μm Hydro-RP 80 Å column (10.00 × 250 mm) and an isocratic 75% ACN/25% H_2O gradient at the flowrate of 3 mL/min over 30 min, at which time the gradient was ramped to 100% ACN over 3 min, then held at 100% ACN for 5 min, before being ramped back to 75% ACN/25% H_2O within 2 min. This reversed-phase HPLC gave rise to four sub-fractions at 16.1 min, 19.5 min, 23.5 min and 33.0 min, respectively.

All four of the fractions were further purified by reversed-phase HPLC using the same column (10.00 × 250 mm Synergi 4 μm Hydro-RP 80 Å column) with different gradients applied. Specifically, the sub-fraction collected at 16.1 min was subjected to HPLC purification [75% ACN/25% H_2O, detection at 211 nm], giving compound **1** (5.2 mg). The sub-fraction obtained at 19.5 min was subjected to HPLC purification [75% ACN/25% H_2O, detection at 211 nm], giving compound **4** (2.0 mg). The sub-fraction obtained at 23.5 min was subjected to HPLC purification [85% ACN/15% H_2O, detection at 211 nm], giving compound **2** (1.8 mg). The sub-fraction obtained at 23.5 min was subjected to HPLC purification [90% ACN/10% H_2O for 5 min, ramping to 95% ACN/5% H_2O over 25 min, then ramping to 100% ACN over 5 min, finally ramping back to 90% ACN/10% H_2O over 5 min, detection at 211 nm], giving compound **3** (1.3 mg).

The cyanobacterial sample of American Samoa *Moorea producens* was extracted with 2:1 CH_2Cl_2/CH_3OH to afford 2.0 g of dried extract. A total of 1.8 g of the extract was fractionated by silica gel vacuum liquid chromatography (VLC) using a stepwise gradient solvent system of increasing polarity starting from 100% hexanes to 100% MeOH (nine fractions, A-I). The three fractions eluting with 60% EtOAc/40%

hexanes (fraction E), 80% EtOAc/20% hexanes (fraction F), and 100% EtOAc (fraction G) were separated further using the Synergi 4 μm Hydro-RP 80 Å column (10.00 × 250 mm), at a flowrate of 3 mL/min over 24 min. The gradient was initially held isocratic at 75% ACN/25% H_2O for 5 min, then ramped to 100% ACN over 25 min and held at 100% ACN for 2 min, before being ramped back to 75% ACN/25% H_2O within 3 min. This reversed-phase HPLC separation process gave rise to six sub-fractions at 9.5 min, 13.5 min, 16.0 min, 17.5 min, 21.2 min, and 24.7 min, respectively. The fractions at 13.5 min and 16.0 min were later combined for further purification using the same column and gradient. Again, this repeated separation gave rise to four additional sub-fractions at 9.5 min, 13.5 min, 16.0 min and 17.5 min. The two separation processes afforded pure viequeamides B (**8**) (30 mg), C (**9**) (1.8 mg) and D (**10**) (1.5 mg).

Viequeamide A (**1**): colorless, rhombic crystal; $[\alpha]_D^{25}$ -32.6 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 204 (4.30) nm; IR (neat) ν_{max} 3387, 2967, 2230, 1749, 1641, 1455, 1366, 1308, 1154, 1122, 733 cm^{-1}; see Table 4.1 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 826.4940 (calculated for $C_{42}H_{69}N_5O_{10}Na$ 826.4937).

Viequeamide A2 (**2**): colorless, amorphous solid; $[\alpha]_D^{25}$ -31.8 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 204 (4.30) nm; IR (neat) ν_{max} 3387, 2967, 1749, 1641, 1455, 1366, 1308, 1154, 1122, 733 cm^{-1}; see Table 4.2 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 828.5087 (calculated for $C_{42}H_{71}N_5O_{10}Na$ 828.5093).

Viequeamide A3 (**3**): colorless, amorphous solid; $[\alpha]_D^{23}$ -32.0 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 204 (4.30) nm; IR (neat) ν_{max} 3389, 2967, 1749, 1641, 1455, 1366, 1308, 1154, 1122, 733 cm^{-1}; see Table 4.3 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 830.5245 (calculated for $C_{42}H_{73}N_5O_{10}Na$ 830.5250).

Aurilide D (**4**): colorless, amorphous solid; $[\alpha]_D^{25}$ -17.2 (*c* 0.34, MeOH); UV (MeOH) λ_{max} (log ε) 222 (4.65) nm; IR (neat) ν_{max} 3481 (br), 3353 (br), 2964, 1740, 1687, 1646, 1251, 1205 cm^{-1}; see Table 4.4 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 856.5402 (calculated for $C_{44}H_{75}N_5O_{10}Na$ 856.5406).

Viequeamide B (**8**): colorless, block crystal; $[\alpha]_D^{25}$ -85.8 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 203 (4.20) nm; IR (neat) ν_{max} 3395, 2965, 2820, 2230, 1728, 1645, 1520, 1458, 1366, 1298, 1187 cm^{-1}; see Table 4.5 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 830.4668 (calculated for $C_{44}H_{65}N_5O_9Na$ 830.4680).

Viequeamide C (**9**): colorless, amorphous solid; $[\alpha]_D^{25}$ -82.2 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 203 (4.48) nm; IR (neat) ν_{max} 3397, 2967, 2830, 1720, 1641, 1520, 1458, 1366, 1298, 1187 cm^{-1}; see Table 4.6 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 832.4823 (calculated for $C_{44}H_{67}N_5O_9Na$ 832.4836).

Viequeamide D (**10**): colorless, amorphous solid; $[\alpha]_D^{25}$ -80.1 (*c* 0.10, MeOH); UV (MeOH) λ_{max} (log ε) 203 (4.35) nm; IR (neat) ν_{max} 3387, 2967, 2832, 1721, 1640, 1520, 1457, 1368, 1300, 1188 cm^{-1}; see Table 4.7 for NMR data; HR-ESI-TOF-MS $[M+Na]^+$ *m/z* 834.4980 (calculated for $C_{44}H_{69}N_5O_9Na$ 834.4993).

Table 4.1. Summary of NMR Data (in $CDCl_3$) for Viequeamide A (1).

residue	position	δ_C, type	δ_H (*J* in Hz)
Thr	1	169.2, C	
	2	58.1, CH	4.68, dd (10.3, 3.1)
	3	68.2, CH	3.87, m
	4	19, CH_3	0.80, d (6.4)
	OH		4.48, d (12.4)
	NH		8.04, d (10.4)
N-Me-Val-2	5	168.2, C	
	6	67.4, CH	4.06, d (10.7)
	7	26.0, CH	2.41, m
	8	19, CH_3	0.84, d (6.7)
	9	20.1, CH_3	0.93 – 1.00
	10	29.0, CH_3	2.77, s
Pro	11	172.1, C	
	12	55.7, CH	5.01, dd (8.4, 3.1)
	13	29.7, CH_2	2.09, m
			2.03, m
	14	25.4, CH_2	2.52, m
	15	47.4, CH_2	3.95, m
			3.52, m
Hmpa[a]	16	168.7, C	
	17	74.5, CH	4.82, d (2.0)
	18	36.6, CH	1.72, m
	19	27.3, CH_2	1.43 – 1.55
	20	12.0, CH_3	0.93 – 1.00
	21	13.8, CH_3	1.10, d (6.7)
N-Me-Val-1	22	170.3, C	
	23	63.7, CH	3.91, d (10.7)
	24	29.5, CH	2.28
	25	19, CH_3	0.93 – 1.00
	26	19, CH_3	0.93 – 1.00
	27	28.3, CH_3	2.98, s
Val	28	172.8, C	
	29	53.9, CH	4.89, dd (7.3, 2.3)
	30	31.7, CH	1.96, m

Table 4.1. Summary of NMR Data (in $CDCl_3$) for Viequeamide A (1). (Cont'd)

residue	position	δ_C, type	δ_H (*J* in Hz)
	31	20.7, CH_3	0.93 – 1.00
	32	16.1, CH_3	0.75, d (6.6)
	NH		6.83, d (7.4)
Dhoya[b]	33	174.7, C	
	34	46.7, C	
	35	17.1, CH_3	1.36, s
	36	25.6, CH_3	1.18, s
	37	77.4, CH	5.55, dd (8.9)
	38	27.9, CH_2	1.76, m
			1.43 – 1.55
	39	24.8, CH_2	1.43 – 1.55
	40	18.1, CH_2	2.19 – 2.32
	41	84.2, C	
	42	68.9, CH	1.93, t (2.6)

a Hmpa, 2-hydroxy-3-methylpentanoic acid.

b Dhoya, 2,2-dimethyl-3-hydroxy-7-octynoic acid.

Table 4.2. Summary of NMR Data (in $CDCl_3$) for Viequeamide A2 (2). (Cont'd)

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
Thr	1	169.0, C			
	2	58.0, CH	4.68, dd (10.3, 3.1)	1, 3, 5	3, 4
	3	68.0, CH	3.86, m		2, 4
	4	19, CH_3	0.80, d (6.4)	2, 3	2, 3
	OH		4.47, d (12.4)	3	1, 2, 3, 4
	NH		8.02, d (10.3)	5	2, 3, 4
N-Me-Val-2	5	168.0, C			
	6	67.3, CH	4.07, d (10.7)	5, 7, 10, 11	7, 8, 9, 10
	7	25.8, CH	2.41, m	6	6, 8, 9, 10
	8	19, CH_3	0.84, d (6.8)	6, 7, 9	6, 7, 9, 10
	9	20.0, CH_3	0.93 – 1.00		6, 7, 8, 10
	10	29.5, CH_3	2.77, s	6, 11	6, 7, 8, 9
Pro	11	172.2, C			
	12	55.6, CH	5.03, dd (8.4, 2.9)		13, 14, 15
	13	29.5, CH_2	2.09, m		12, 14, 15
			2.02, m		
	14	25.2, CH_2	2.51, m		12, 13, 15
	15	47.2, CH_2	3.96, m	13, 14	12, 13, 14
			3.51, m	13, 14	
Hmpa	16	168.5, C			
	17	74.4, CH	4.83, d (1.4)	18-21	18, 19, 20, 21
	18	36.4, CH	1.72, m	19, 20, 21	17, 19, 20, 21
	19	27.1, CH_2	1.46, m	18, 20, 21	17, 18, 20, 21
	20	12.0, CH_3	0.93 – 1.00	18, 20, 21	17, 18, 19, 21
	21	13.7, CH_3	1.10, d (6.8)		17, 18, 19, 20
N-Me-Val-1	22	170.3, C			
	23	63.5, CH	3.92, d (10.8)	22, 24, 27, 28	24, 25, 26
	24	29.3, CH	2.27, m		23, 25, 26
	25	19.4, CH_3	0.92 – 0.96		23, 24, 26
	26	19.1, CH_3	0.92 – 0.96		23, 24, 25
	27	28.1, CH_3	2.97, s	23, 28	
Val	28	172.6, C			

Table 4.2. Summary of NMR Data (in $CDCl_3$) for Viequeamide A2 (2). (Cont'd)

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
	29	53.7, CH	4.89, dd (7.3, 1.6)	28, 30, 33	30, 31, 32,
	30	31.6, CH	1.95, m		29, 31, 32
	31	20.5, CH_3	0.95, d, obscr		29, 30, 32
	32	16.0, CH_3	0.74, d (6.6)	29, 30, 31	29, 30, 31
	NH		6.84, d (7.3)	33	29, 30, 31, 32
Dhoea[a]	33	174.7, C			
	34	46.5, C			
	35	16.9, CH	1.35, s	33, 34, 36, 37	
	36	25.5, CH_3	1.15, s	33, 34, 35, 37	
	37	78.0, CH	5.53, dd (10.1)	1, 33, 34	38, 39, 40,41, 42
	38	28.3, CH_2	1.37 – 1.51		37, 39, 40, 41, 42
	39	25.1, CH_2	1.37 – 1.51	41	37, 38, 40, 41, 42
	40	33.4, CH_2	2.04 – 2.15	41, 42	37, 38, 39, 41, 42
	41	138.5, CH	5.75, ddt (5.5, 8.5, 14.5)	40, 42	37, 38, 39, 40, 42
	42	114.8, CH_2	4.99, dd (1.5, 14.5)	40, 41	37, 38, 39, 40, 41
			4.99, dd (1.5, 8.5)		

a Dhoea, 2,2-dimethyl-3-hydroxy-7-octenoic acid.

Table 4.3. Summary of NMR Data (in $CDCl_3$) for Viequeamide A3 (3).

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
Thr	1	168.9, C			
	2	58.0, CH	4.68, dd (10.3, 3.1)	1, 3, 5	3, 4
	3	68.0, CH	3.86, m		2, 4
	4	19, CH_3	0.80, d (6.4)	2, 3	2, 3
	OH		4.47, d (12.4)	3	1, 2, 3, 4
	NH		8.02, d (10.3)	5	2, 3, 4
N-Me-Val-2	5	168.0, C			
	6	67.3, CH	4.07, d (10.7)	5, 7, 10, 11	7, 8, 9, 10
	7	26.0, CH	2.41, m	6	6, 8, 9, 10
	8	19, CH_3	0.84, d (6.8)	6, 7, 9	6, 7, 9, 10
	9	20.1, CH_3	0.99, d (6.4)		6, 7, 8, 10
	10	29.0, CH_3	2.77, s	6, 11	6, 7, 8, 9
Pro	11	172.0, C			
	12	55.6, CH	5.02, dd (8.4, 2.9)		13, 14, 15
	13	29.5, CH_2	2.09, m		12, 14, 15
			2.02, m		
	14	25.2, CH_2	2.51, m		12, 13, 15
	15	47.2, CH_2	3.95, m	13, 14	12, 13, 14
			3.51, m	13, 14	
Hmpa	16	168.5, C			
	17	74.4, CH	4.83, d (1.4)	18–21 weak	18, 19, 20, 21
	18	36.4, CH	1.73 m	19, 20, 21	17, 19, 20, 21
	19	27.1, CH_2	1.46, m	18, 20, 21	17, 18, 20, 21
	20	11.9, CH_3	0.98, t (6.0)	18, 20, 21	17, 18, 19, 21
	21	13.7, CH_3	1.10, d (6.8)		17, 18, 19, 20
N-Me-Val-1	22	170.3, C			
	23	63.5, CH	3.92, d (10.8)	22, 24, 27, 28	24, 25, 26

Table 4.3. Summary of NMR Data (in $CDCl_3$) for Viequeamide A3 (3). (Cont'd)

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
	24	29.3, CH	2.27, m		23, 25, 26
	25	19, CH_3	0.93 – 0.97		23, 24, 26
	26	19, CH_3	0.93 – 0.97		23, 24, 25
	27	28.2, CH_3	2.97, s	23, 28	
Val	28	172.6, C			
	29	53.7, CH	4.89, dd (7.3, 1.6)	28, 30, 33	30, 31, 32,
	30	31.7, CH	1.95, m		29, 31, 32
	31	20.5, CH_3	0.95, d (5.6)		29, 30, 32
	32	16.0, CH_3	0.74, d (6.6)	29, 30, 31	29, 30, 31
	NH		6.84, d (7.3)	33	29, 30, 31, 32
Dhoaa[a]	33	174.7, C			
	34	46.5, C			
	35	16.9, CH_3	1.34, s	33, 34, 36, 37	
	36	25.5, CH_3	1.15, s	33, 34, 35, 37	
	37	78.1, CH	5.53, dd (1.9, 10.5)	1, 33, 34	38, 39, 40,41, 42
	38	29.0, CH_2	1.73, m		37, 39, 40, 41, 42
			1.37 – 1.51		
	39	31.9, CH_2	1.20 – 1.32	41	37, 38, 40, 41, 42
	40	22.5, CH_2	1.20 – 1.32	41, 42	37, 38, 39, 41, 42
	41	25.8, CH_2	1.20 – 1.32	40, 42	37, 38, 39, 40, 42
	42	14.1, CH_3	0.85, t (6.7)	40, 41	37, 38, 39, 40, 41

a Dhoaa, 2,2-dimethyl-3-hydroxy-7-octanoic acid.

Table 4.4. Summary of NMR Data (in $CDCl_3$) for Aurilide D (4).

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
Me-Ala	1	170.3, C			
	2	58.6, CH	3.79, m	1,3,4,5	2, 3
	3	13.7, CH_3	1.41, d (7.0)	1,2	2,3
	4	36.5, CH_3	3.22, s	2,5	4
Val-1	5	171.9, C			
	6	54.1, CH	4.92, dd (9.3, 5.8)	5,8,9,10	6, 7, 8, 9
	7	31.5, CH	2.08, m		6, 7, 8, 9
	8	19.7, CH_3	1.02, d (6.3)	6,7,9	6, 7, 8, 9
	9	17.5, CH_3	1.01, d (6.3)	6,7,8	6, 7, 8, 9
	NH (1)		7.38 br, d (9.3)	6,7,10,11	6, 7, 8, 9
Me-Gly	10	169.0, C			
	11	51.1, CH_2	4.03, d (17.6)	10,12	11
			3.48, d (17.6)	12,13	11
	12	36.3, CH_3	2.90, s	11,13	12
Me-Leu	13	170.3, C			
	14	52.0, CH	5.17, dd (7.1, 7.1)	13,15,16,19,20	14, 15, 16, 17, 18
	15	38.0, CH_2	1.77, m	13,14,16,17,18	14, 15, 16, 17, 18
			1.46, m		
	16	24.7, CH	1.46, m		14, 15, 16, 17, 18
	17	22.9, CH_3	0.87, d (6.3)	15,16	14, 15, 16, 17, 18
	18	22.8, CH_3	0.87, d (6.3)	15,16	14, 15, 16, 17, 18
	19	30.6, CH_3	2.96, s	18,20	19
Val-2	20	172.0, C			
	21	54.3, CH	4.65, t (8.6)	20,22,24,25	21, 22, 23, 24
	22	31.0, CH	2.08, m		21, 22, 23, 24
	23	19.6, CH_3	0.94, d (6.0)	21,22,24	21, 22, 23, 24
	24	18.5, CH_3	0.94, d (6.0)	21,22,23	21, 22, 23, 24
	NH (2)		6.67 br, d (8.9)	21,22,25	21, 22, 23, 24
Hmpa	25	170.0, C			
	26	77.8, CH	4.79, d (5.7)	25,27,28,30	26, 27, 28, 29, 30
	27	36.9, CH	2.03, m	26,30	26, 27, 28, 29, 30
	28	25.8, CH_2	1.46, m	26,27,29,30	26, 27, 28, 29, 30
			1.23, m	26,27,29,30	26, 27, 28, 29, 30
	29	11.7, CH_3	0.93, t (7.0)	27,28	26, 27, 28, 29, 30
	30	14.6, CH_3	0.98, d (6.6)	26,27,28	26, 27, 28, 29, 30
Dhtmda[a]	31	168.9, C			

Table 4.4. Summary of NMR Data (in $CDCl_3$) for Aurilide D (4). (Cont'd)

residue	position	δ_C, type	δ_H (*J* in Hz)	HMBC	HSQC-TOCSY
	32	128.3, C			
	33	144.2, CH	7.23, dd (9.8, 4.6)	31,32,34,35,42	33, 34, 35
	34	30.8, CH_2	2.23, m	32,33,35,36	33, 34, 35
	35	71.4, CH	3.81, m	33,44	33, 34, 35
	36	40.2, CH	2.10, m	35,37	36, 37, 43
	37	82.6, CH	4.98, d (11.2)	1,35,36,38,39,43	36, 37, 43
	38	130.4, C			
	39	134.3, CH	5.50, t (7.1)	37,40,44,41	39, 40, 41
	40	21.0, CH_2	2.01, m	38,39,41	39, 40, 41
	41	14.0, CH_3	0.93, t (7.0)	39,40	39, 40, 41
	42	12.6, CH_3	1.88, s	31,32,33	42
	43	10.4, CH_3	0.74, d (7.2)	35,36,37	35, 36, 37
	44	10.9, CH_3	1.54, s	37,38,39	44

a Dhtmda, 5,7-dihydroxy-2,6,8-trimethylundeca-2,8-dienoic acid

Table 4.5. Summary of NMR Data (in $CDCl_3$) for Viequeamide B (8).

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
Val-1	1	169.98, C	
	2	57.22, CH	4.77, dd (9.6, 4.0)
	3	29.72, CH	2.32, m
	4	20.53, CH_3	0.99, d (6.7)
	5	18.32, CH_3	0.73, d (6.9).
	NH(1)		7.56, d (9.6)
N-Me-Ala	6	170.51, C	
	7	56.03, CH	4.28, q (6.9)
	8	15.26, CH_3	1.48, d (7.0)
	9	30.29, CH_3	2.80, s
Pro	10	171.1, C	
	11	57.48, CH	4.17, m
	12	30.52, CH_2	1.52, m
	13	21.92, CH_2	1.71, m
			1.50, m
	14	46.05, CH_2	3.55, m
			3.24, m
Pla	15	167.13, C	
	16	71.08, CH	5.71, dd (11.4, 5.1)
	17	38.06, CH_2	3.73, dd (11.8, 5.0)
			3.29, t (11.4)
	18	135.8, C	
	19	129.99, CH	7.42, d (7.3)
	20	128.4, CH	7.25, m
	21	126.88, CH	7.20, m
	22	128.4,CH	7.25, m
	23	129.99, CH	7.42, d (7.3)
N-Me-Val	24	170.62, C	
	25	65.04, CH	4.47, d (9.0)
	26	30.45, CH	2.38, m
	27	19.95, CH_3	1.18, d (6.7)
	28	21.45, CH_3	1.04, d (6.7)
	29	30.39, CH_3	2.99, s
Val-2	30	172.77, C	
	31	53.59, CH	4.72, t (8.6)
	32	32.7, CH	2.02, m
	33	18.41, CH_3	0.93, d (6.7)
	34	20.03, CH_3	0.93, d (6.7)

Table 4.5. Summary of NMR Data (in $CDCl_3$) for Viequeamide B (8). (Cont'd)

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
	NH(2)		6.16, d (8.8)
Dhoya[a]	35	174.84, C	
	36	46.35, C	
	37	17.07, CH_3	1.26, s
	38	25.83, CH_3	1.15, s
	39	78.04, CH	5.75, dd (11.3, 2.0)
	40	27.65, CH_2	1.72, m
			1.59, m
	41	24.68, CH_2	1.55, m
			1.44, m
	42	18.06, CH_2	2.24, m
	43	83.53, C	
	44	69.29, CH	1.95, t (2.67)

a Dhoya, 2,2-dimethyl-3-hydroxy-7-octynoic acid.

Table 4.6. Summary of NMR Data (in $CDCl_3$) for Viequeamide C (9).

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
Val-1	1	169.96, C	
	2	57.25, CH	4.76, dd (9.5, 3.9)
	3	29.82, CH	2.30, m
	4	20.55, CH_3	0.98, d (6.8)
	5	18.32, CH_3	0.72, d (6.9)
	NH(1)		7.55, d (9.5)
N-Me-Ala	6	170.54, C	
	7	56.07, CH	4.27, q (6.9)
	8	15.32, CH_3	1.47, d (6.9)
	9	30.35, CH_3	2.79, s
Pro	10	171.13, C	
	11	57.56, CH	4.16, m
	12	30.55, CH_2	1.52, m
	13	21.97, CH_2	1.73 – 1.64, m
			1.50, m
	14	46.09, CH_2	3.55, m
			3.22, m
Pla	15	167.2, C	
	16	71.11, CH	5.75 – 5.70, m
	17	38.08, CH_2	3.73, dd (11.8, 5.0)
			3.28, t (11.6)
	18	135.86, C	
	19	130.06, CH	7.41, d (7.3)
	20	128.42, CH	7.23, t (7.2)
	21	126.9, CH	7.19, m
	22	128.42, CH	7.23, t (7.2)
	23	130.06, CH	7.41, d (7.3)
N-Me-Val	24	170.63, C	
	25	65.06, CH	4.47, d (9.0)
	26	30.5, CH	2.37, m
	27	19.99, CH_3	1.18, d (6.7)
	28	21.51, CH_3	1.03, d (6.7)
	29	30.43, CH_3	2.98, s
Val-2	30	172.8, C	
	31	53.6, CH	4.71, t (8.7)
	32	32.81, CH	2.02, m
	33	18.48, CH_3	0.92, d (6.7)

Table 4.6. Summary of NMR Data (in $CDCl_3$) for Viequeamide C (9). (Cont'd)

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
	34	20.07, CH_3	0.92, d (6.7)
	NH(2)		6.16, d (8.7)
Dhoea[a]	35	174.96, C	
	36	46.37, C	
	37	17.11, CH_3	1.23, s
	38	25.92, CH_3	1.11, s
	39	78.44, CH	5.75 – 5.70, m
	40	28.01, CH_2	1.52, m
	41	25.14, CH_2	1.42, m
			1.33, m
	42	33.32, CH_2	2.08, m
	43	138.13, CH	5.75 – 5.70, m
	44	115.38, CH_2	5.01 – 4.91, m

a Dhoea, 2,2-dimethyl-3-hydroxy-7-octenoic acid.

Table 4.7. Summary of NMR Data (in $CDCl_3$) for Viequeamide D (10).

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
Val-1	1	170, C	
	2	57.33, CH	4.75, dd (9.5, 3.9)
	3	29.86, CH	2.29, m
	4	20.58, CH_3	0.97, d (6.8)
	5	18.38 , CH_3	0.71, d (6.9)
	NH(1)		7.54, d (9.5)
N-Me-Ala	6	170.51, C	
	7	56.1, CH	4.26, q (6.9)
	8	15.32, CH_3	1.46, d (7.0)
	9	30.32, CH_3	2.78, s
Pro	10	171.18, C	
	11	57.56, CH	4.15, dd (7.6, 2.8)
	12	30.59, CH_2	1.54 – 1.47, m
	13	21.99, CH_2	1.73 – 1.64, m
			1.50, m
	14	46.12, CH_2	3.56 – 3.51, m
			3.23 – 3.18, m
Pla	15	167.24, C	
	16	71.2, CH	5.73, dd (11.4, 4.9)
	17	38.14, CH_2	3.72, dd (11.8, 5.0)
			3.26, t (11.5)
	18	135.97, C	
	19	130.12, CH	7.40, d (7.6)
	20	128.38, CH	7.22, t (7.2)
	21	126.89, CH	7.17, t (7.7)
	22	128.38, CH	7.22, t (7.2)
	23	130.12, CH	7.40, d (7.6)
N-Me-Val	24	170.66, C	
	25	65.12, CH	4.48, d (8.9)
	26	30.53, CH	2.36, m
	27	20.01, CH_3	1.16, d (6.7)
	28	21.51, CH_3	1.01, d (6.7)
	29	30.42, CH_3	2.97, s
Val-2	30	172.84, C	
	31	53.63, CH	4.70, t (8.7)
	32	32.83, CH	2.05 – 1.97, m
	33	18.46, CH_3	0.91, d (6.4)
	34	20.09, CH_3	0.91, d (6.4)

Table 4.7. Summary of NMR Data (in $CDCl_3$) for Viequeamide D (10). (Cont'd)

Residue	Position	δ_C, Type	δ_H (*J* in Hz)
	NH(2)		6.18, d (8.8)
Dhoaa[a]	35	175.06, C	
	36	46.38, C	
	37	17.15, CH_3	1.24, s
	38	25.97, CH_3	1.10, s
	39	78.8, CH	5.71, dd (10.4, 3.6)
	40	28.74, CH_2	1.48, m
	41	25.7, CH_2	1.29, m
			1.21, m
	42	31.59, CH_2	1.25, m
	43	22.65, CH_2	1.26, m
	44	14.13, CH_3	0.82, t (6.6)

a Dhoaa, 2,2-dimethyl-3-hydroxy-7-octanoic acid.

4.4.4. Molecular Networking

The defrosted relatively non-polar fraction of the *Rivularia sp.* sample and the nine fractions and the crude extract of the *Moorea producens* sample were analyzed by LC-MS/MS to generate the molecular networks. Using Thermo Finnigan Surveyor Autosampler-Plus/LC-Pump-Plus/PDA-Plus system coupled to a Thermo Finnigan LCQ Advantage Max mass spectrometer adapted with a Phenomenex Kinetex C_{18} 100 Å 100 × 4.60 mm column, with a flow rate of 700 μL/min and the elution system of 30% ACN/70% HCOOH acidified water for 5 min, next ramping to 99% ACN/1% HCOOH acidified water over 17 min, then 99% ACN/1% HCOOH acidified water for 3 min, furthermore ramping back to 30% ACN/70% HCOOH acidified water over 1

min, finally 30% ACN/70% HCOOH acidified water for 3 min. ESI conditions were set according to previously determined parameters[16]. MS spectra applying four scan events were acquired: the first scan was positive MS mode with a window from *m/z* 190-2000; the second was a selective MS/MS scan of the most intense ion from the first scan, also using ESI mode; the third scan was a selective MS/MS scan of the second most intense ion from the first scan; and the fourth scan was a selective MS/MS scan of the third most intense ion from the first scan.

The data were converted to .mzXML format, via msconvert, before being subjected to GNPS for network visualization with minimum cosine score[16] set 0.7. The network was embedded with MS/MS data of ions below *m/z* 2000 from available databases on the GNPS website. Other algorithm parameters for the networking were set as follows: include ion tolerance (0.3 Da) and parent mass tolerance (0.3 Da). Node colors were chosen based on the source files of the MS/MS. The edge label was set to display cosine scores between the two nodes at both ends of the edge representing two source files (higher cosine score meaning higher similarity).

4.4.5. Small Molecular Accurate Recognition Technology (SMART) Analysis of 2D HSQC Spectra

The training details of the SMART was published separately[10]. The HSQC spectra for compounds (**2**, **3**, **8**-**10**) shown in the Chapter 4 Appendix were acquired using the NUS edited hsqcedetgpsisp2.3 HSQC pulse sequence. Data were acquired as 4096 × 32 points (256 out of 256 t_1 increments, 100% NUS) in ^{1}H and ^{13}C dimensions, respectively. Spectral windows in direct and indirect dimensions were 7183.9 and

24118.9 Hz respectively. HSQC spectra were processed using NMRPipe[32] by applying zero filling (round final size to power of 2) in both dimensions. Collected NUS 2D data were processed by applying IST as implemented in hmsIST[33] with 400 iterations followed by MEM with the standard deviation of time-domain noise set to 200.

4.4.6. X-ray Diffraction Analysis for Viequeamide A (1) and Viequeamide B (8)

Colorless crystals of viequeamide A (**1**) were obtained from evaporation of a 90% acetone/10% hexane solution in the open air at room temperature. Similarly, colorless crystals of viequeamide B (**8**) were obtained from the evaporation of its methanol solution in the open air at room temperature. The crystal data were collected at 100(2) K on an Agilent Gemini ultra diffractometer using Cu Kα radiation (λ = 1.54718 Å). The 3D crystal structures were solved by direct methods via SHELXS-97. The refinement was applied using full-matrix least-squares difference Fourier techniques with anisotropic displacement parameters for all the non-hydrogen atoms. All hydrogen atoms were placed in calculated positions and refined using a riding model with the relative isotropic parameters. The crystallographic data for **1** and **8** have been deposited at the Cambridge Crystallographic Data Center with the deposition number: 1457231 and 1457232, respectively. Crystal data for **1**: $C_{42}H_{69}N_5O_{10}$, Mr = 833.06, monoclinic, space group P 1 21 1 with a = 14.6362(6) Å, b = 10.5766(4) Å, c = 31.0319(14) Å, β = 92.558(2)°, V = 4799.0(3) Å^3, Z = 4, Dx = 1.153 g/cm^3, μ(Cu Kα) = 0.668 mm^{-1}, and F(000) = 1808.0. Crystal dimensions: 0.33 × 0.30 × 0.08 mm^3. Independent reflections: 17186 (R_{int} = 0.0496). The final R_1 values were 0.0695, absolute structure parameter = 0.11(5), wR_2 = 0.1518 (I > 2σ(I)). Crystal data for **8**:

$C_{44}H_{65}N_5O_9$, Mr = 808.01, orthorhombic, space group P 21 21 21 with a = 9.4670(3) Å, b = 12.2088(4) Å, c = 37.4659(11) Å, β = 90°, V = 4330.3(2) Å^3, Z = 4, Dx = 1.153 Mg/m^3, μ(Cu Kα) = 0.702 mm^{-1}, and F(000) = 1744. Crystal dimensions: 0.34 × 0.30 × 0.26 mm^3. Independent reflections: 7753 (R_{int} = 0.0496). The final R_1 values were 0.0353, absolute structure parameter = 0.09 (54), wR_2 = 0.0951 (I > 2σ(I)).

4.4.7. Absolute Configuration of the Peptidic Moiety of Viequeamide A2 (2) and A3 (3)

The specific optical rotation, ^{1}H and ^{13}C NMR shifts, and $J_{H,H}$ values of **2** and **3** were all consistent with those of viequeamide A (**1**), which indicated that those two compounds likely possessed identical absolute configurations, and thus further stereochemical analyses of **2** and **3** were not performed.

4.4.8. Absolute Configuration of Aurilide D (4) by Marfey's Analysis and Chiral GCMS

Aurilide D (**4**, 0.5 mg) was hydrolyzed in 500 μL of 6 N HCl at 110 °C for 18 h, then dried under a stream of N_2 and further dried under vacuum, and the resulting solids were dissolved in 500 μL of CH_2Cl_2 and partitioned into a 300 μL and 200 μL sample.

The 300 μL sample was concentrated to dryness in a fresh vial and redissolved in 300 μL of 1 M $NaHCO_3$ (aq), and a Teflon stir bar and 1.1 mL of 1-fluoro-2,4-dinitrophenyl-5-L-valine amide (L-FDVA) in acetone (at a concentration of 1.0 mg/mL) was added. The reaction mixture was then stirred at 35 °C for 1 h, and then quenched with 2 N HCl (150 μL). The reaction mixture was then transferred to a fresh vial with 3 × 500 μL of CH_3CN, concentrated to dryness under N_2 (g) flow, and transferred to an

LCMS vial again through a syringe filter with 3 × 0.5 mL of CH_3CN. This mixture and all L-FDVA derivatized authentic amino acid standards were subjected to LC-MS [column: Phenomenex Kinetex C_{18}, 4.60 × 100 mm, 5 μm 100 Å; mobile phase, CH_3CN in 0.1% (*v/v*) aqueous HCOOH; flow rate, 0.40 mL/min] using a linear gradient (5-50% CH_3CN over 75 min) according to Marfey's method, respectively. Retention times for the authentic standards were as follows: D/L-*N*-Me-Ala (55.05 and 54.56 min), L-*N*-Me-Ala (54.46 min), D/L-Val (66.27 and 57.61 min), L-Val (57.74 min), D/L-*N*-Me-Leu (70.27 and 66.46 min), and L-*N*-Me-Leu (66.33 min). The hydrolysate peaks with the expected *m/z* ratio were found at 54.61, 57.74, and 66.54 min which correspond to L-*N*-Me-Ala, L-Val, and L-*N*-Me-Leu.

The above 200 μL sample was concentrated to dryness and redissolved in 400 μL of dry DMF while stirring. CH_3I (20 μL) was added to the solution in excess. After stirring for 16 h, the reaction was concentrated to dryness under N_2 (g) flow and the residue quickly transferred to a GCMS vial with 3 × 50 μL of CH_2Cl_2. 5 μL of the product and previously prepared authentic standards (2*S*,3*S*-Hmpa, 2*S*,3*R*-Hmpa, 2*R*,3*R*-Hmpa, and 2*R*,3*S*-Hmpa) were injected into the GCMS system, respectively. (For all GC-MS, 35 ^oC for 15 min followed by a ramp of 1.5 ^oC/min until 60 ^oC was reached at which time the GC was ramped at 50 ^oC/min until 170 ^oC where the temperature was held for 5 min. The ion source was held at 250 ^oC and the positive ion scans from 50 to 1000 *m/z* were collected from 7 min until the end). Retention times of 2*S*, 3*S*-Hmpa, 2*S*, 3*R*-Hmpa, 2*R*, 3*R*-Hmpa, and 2*R*, 3*S*-Hmpa were 45.66, 44.88, 45.23, and 45.11 min, respectively. The retention times of the sample corresponded to (2*R*, 3*S*)-Hmpa (45.08).

^{1}H and ^{13}C NMR data of 5,7-dihydroxy-2,6,8-trimethylundeca-2,8-dienoic acid residue in **4** were virtually identical with those of aurilide as well as aurilide B and C[30]. Therefore, it is likely that these closely related molecules all have identical relative and absolute configurations for this fragment. Owing to the trace amounts of isolated **4**, a stereochemical analysis of that fragment was not performed.

4.4.9. Absolute Configuration of the Peptidic Moiety of Viequeamide C (**9**) and D (**10**)

The specific optical rotation, ^{1}H and ^{13}C NMR shifts, and $J_{H,H}$ values of **9** and **10** were all consistent with those of viequeamide B (**8**), indicating that they likely had identical absolute configurations, and thus stereochemical analyses of **9** and **10** were not performed.

4.4.10. Biological Activity

Cytotoxicity was measured in NCI-H460 human lung tumor cells using previously reported methods, with cell viability being determined by MTT reduction[34]. Cells were seeded in 96-well plates at 6000 cells/well in 180 μL of Roswell Park Memorial Institute (RPMI) medium. Twenty-four hours later, the test chemicals were dissolved in DMSO and diluted into medium without fetal bovine serum; 20 μL of this resultant mixture was added to each well. DMSO was less than 0.5% of the final concentration. After 48 h, the medium was removed and cell viability determined.

4.5 Chapter 4 Acknowledgements

We thank Dr. Y. Su (UCSD Chemistry and Biochemistry Mass Spectrometry Facility), and Prof. A. L. Rheingold (UCSD) for HR-ESI-TOFMS and X-ray Diffraction Analysis, respectively. We also thank Dr. Brendan Duggan and Dr. Anthony Mrse for access to the NMR instruments and their excellent technical support. Finally, we would like to thank Dr. Seth Cohen for access to the FTIR instrument in his lab. The work was supported by National Institutes of Health grant CA100851 (WHG), Science and Technology Project of Guangdong Province (2013B021100021), Guangdong Natural Science Foundation (2016A030313588), Fund of the Education Bureau of Guangzhou City (1201610155) and construction fund of high-level university from Guangzhou Medical University.

Chapter 4, in essence, is currently being prepared for submission in 2017, with the following authors Yiwen Tao, Chen Zhang, Yerlan Idelbayev, Svetlana Nikoulina, Evgenia Glukhov, C. Benjamin Naman, Garrison W. Cottrell and William H. Gerwick. The dissertation author was a primary investigator and first author of this paper.

4.6 Chapter 4 Appendix

NMR Spectroscopic Data for viequeamide A2 (**2**)

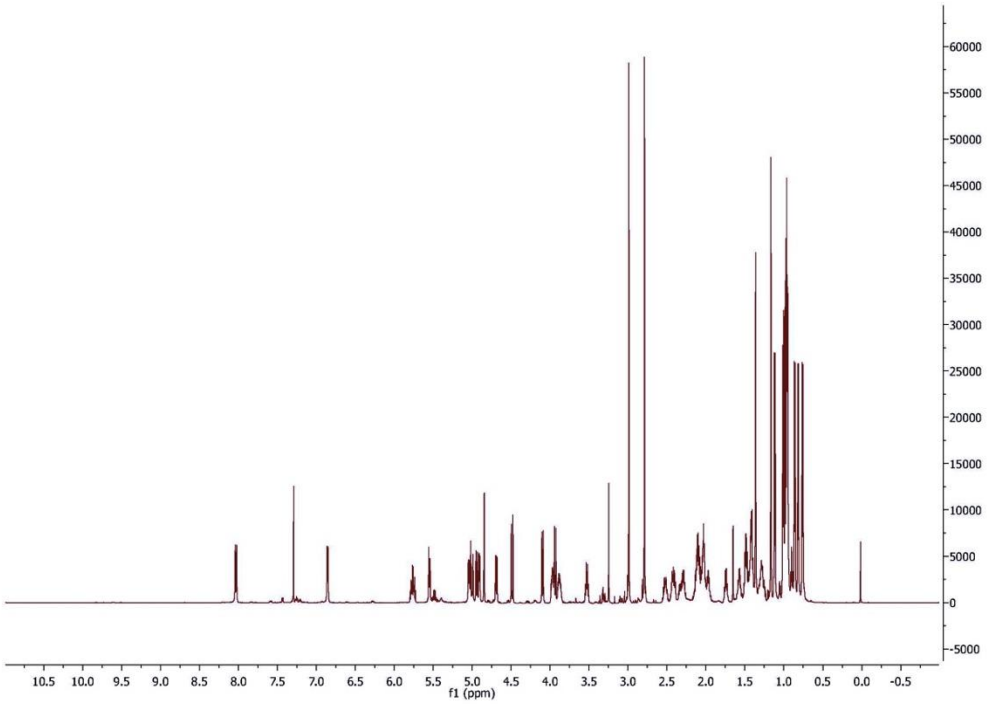


Figure 4.6.1 ^{1}H NMR (600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

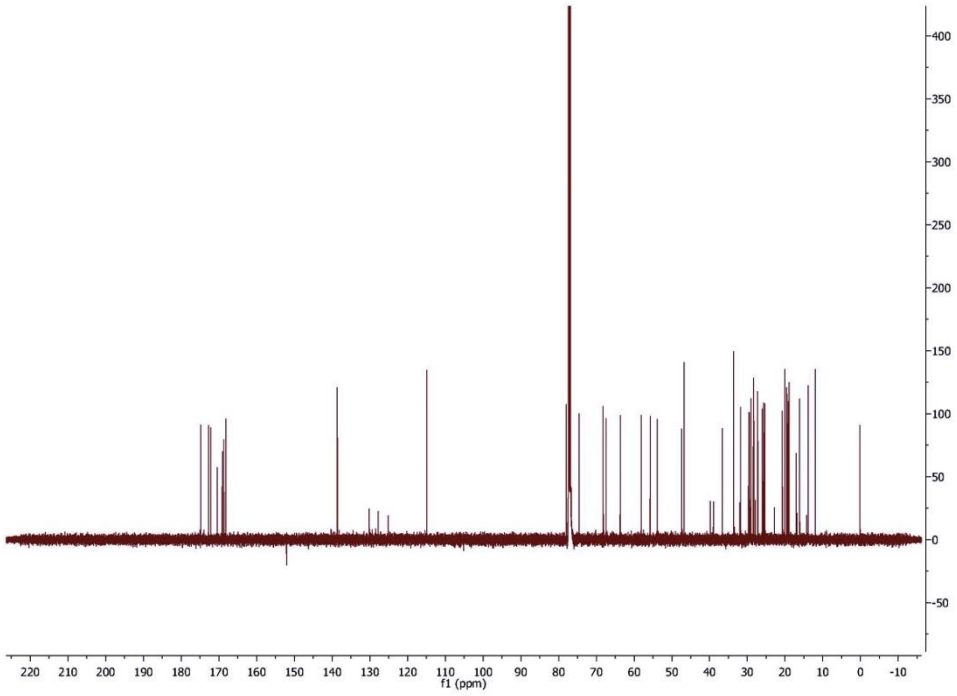


Figure 4.6.2 ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

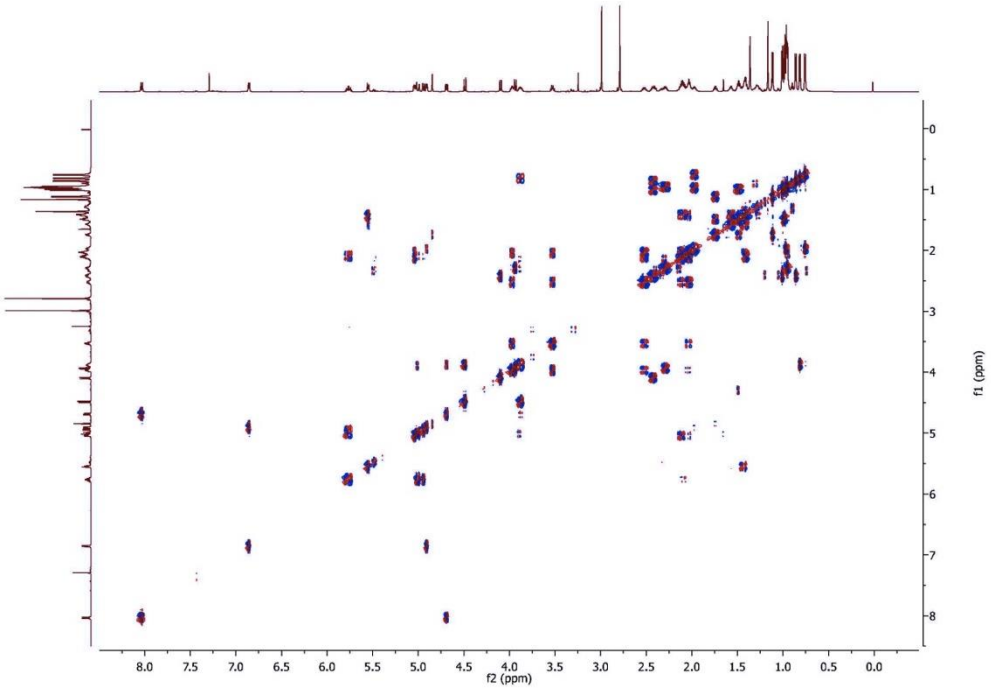


Figure 4.6.3 DQF-COSY (1H 600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

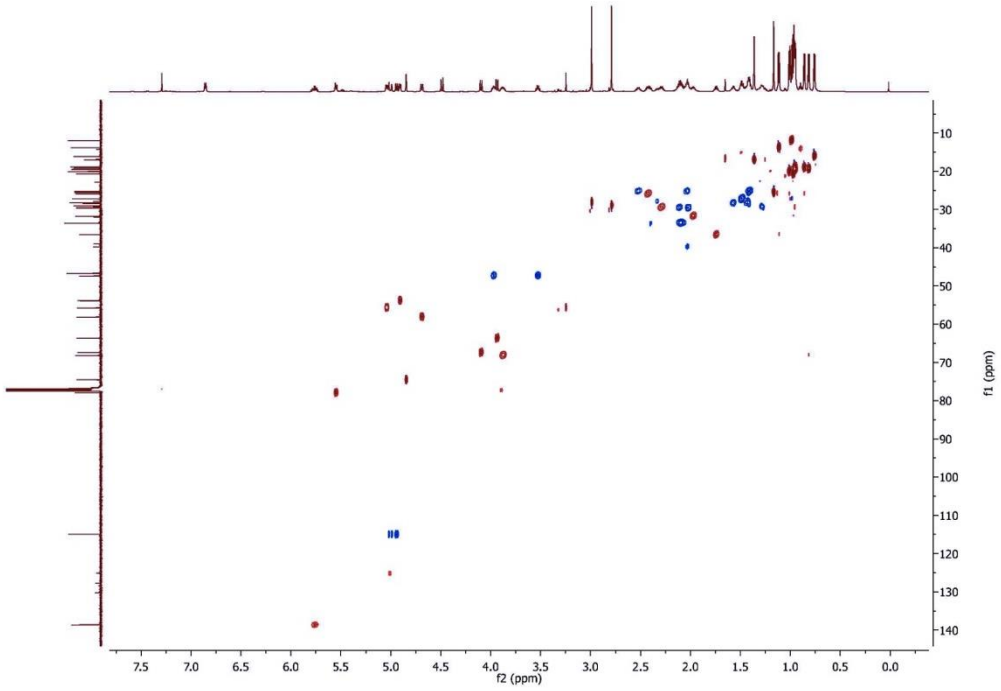


Figure 4.6.4 HSQC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

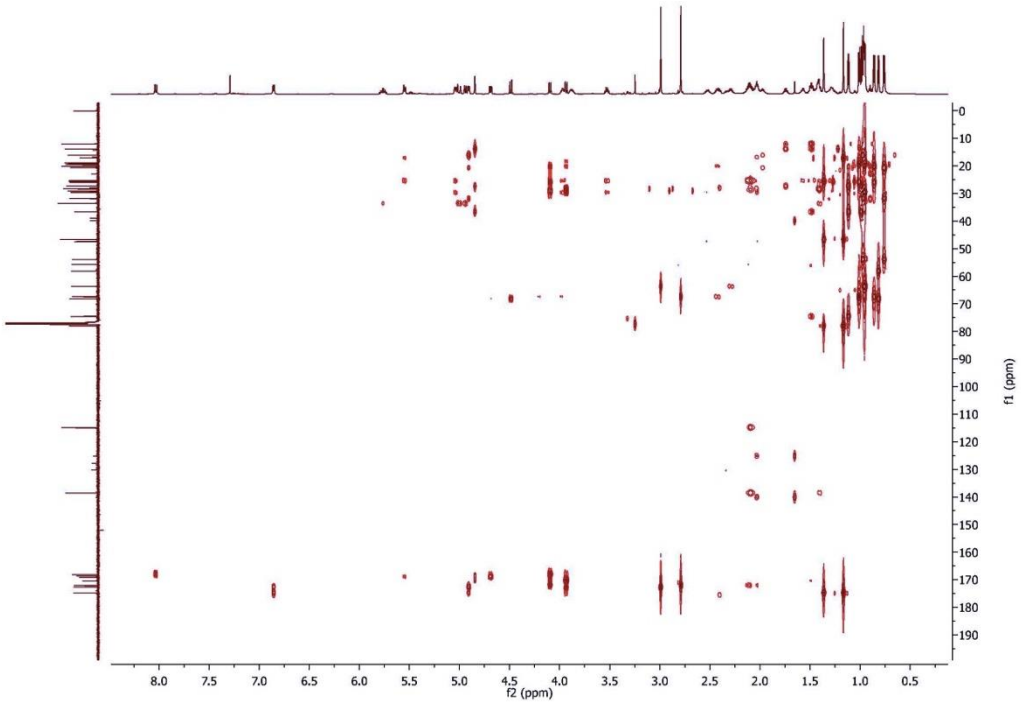


Figure 4.6.5 HMBC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

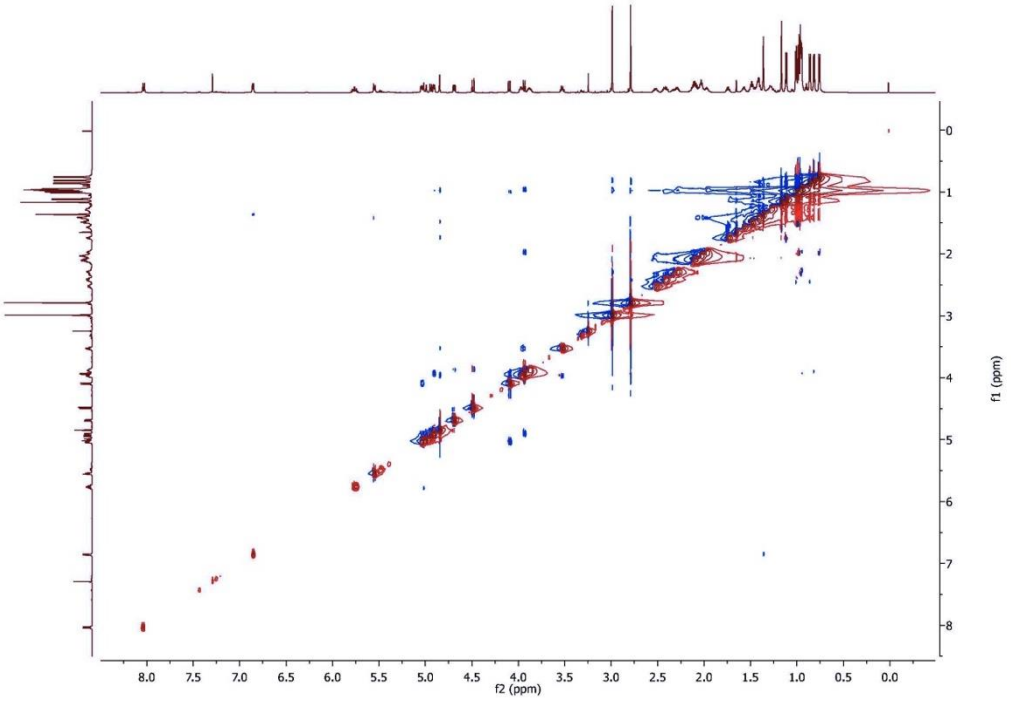


Figure 4.6.6 ROESY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

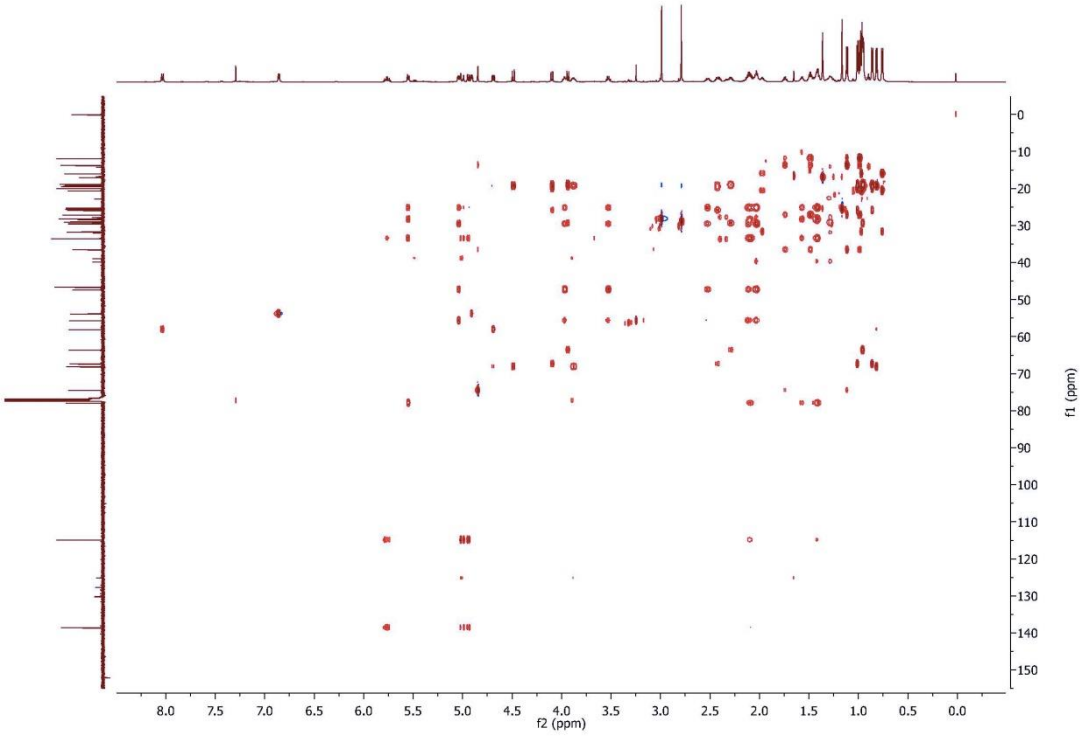


Figure 4.6.7 HSQC-TOCSY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A2 (**2**).

NMR Spectroscopic Data for Viequeamide A3 (**3**)

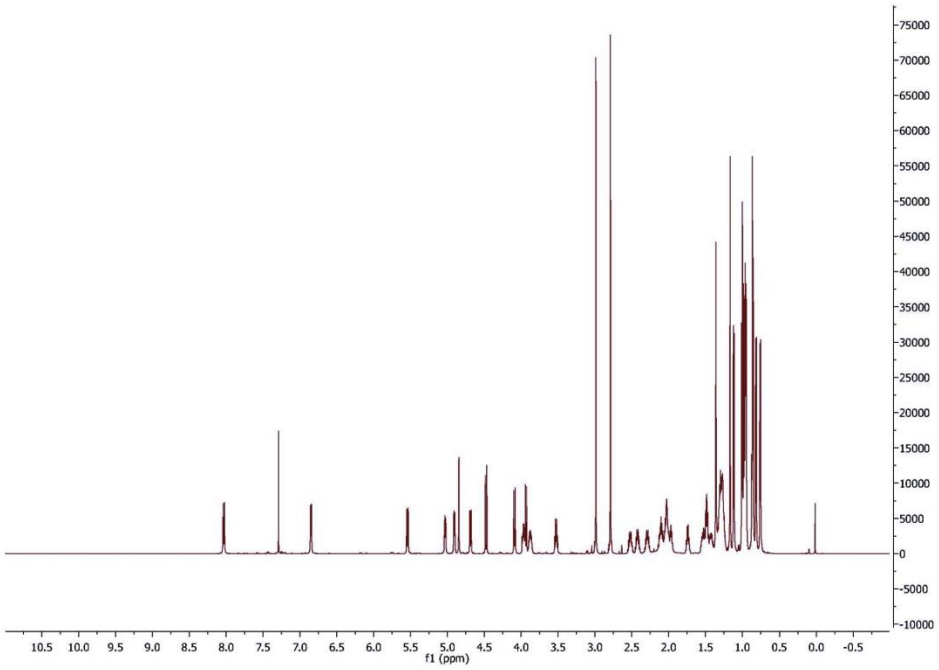


Figure 4.6.8 ^{1}H NMR (600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

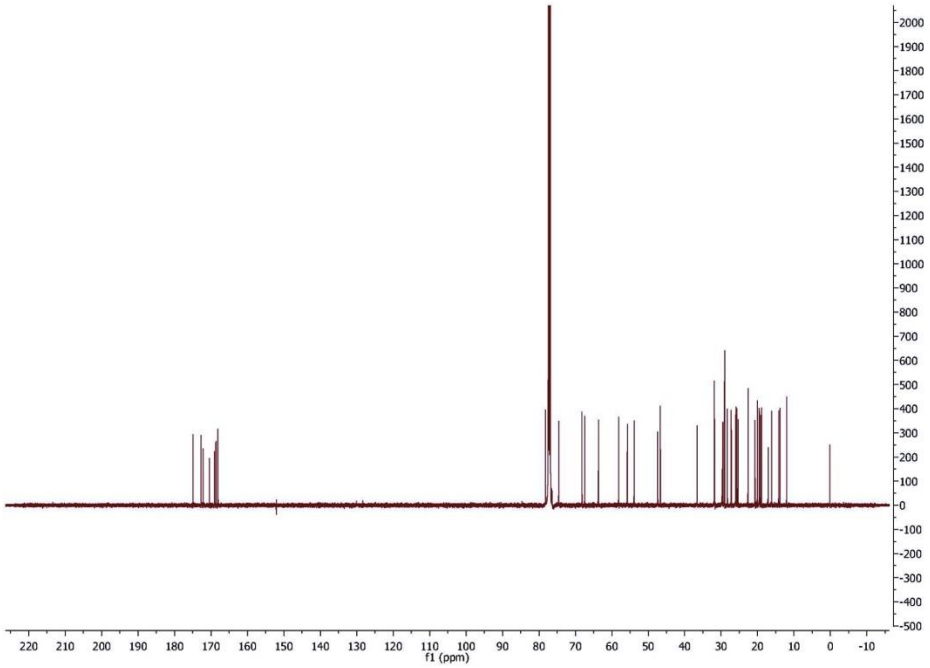


Figure 4.6.9 ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

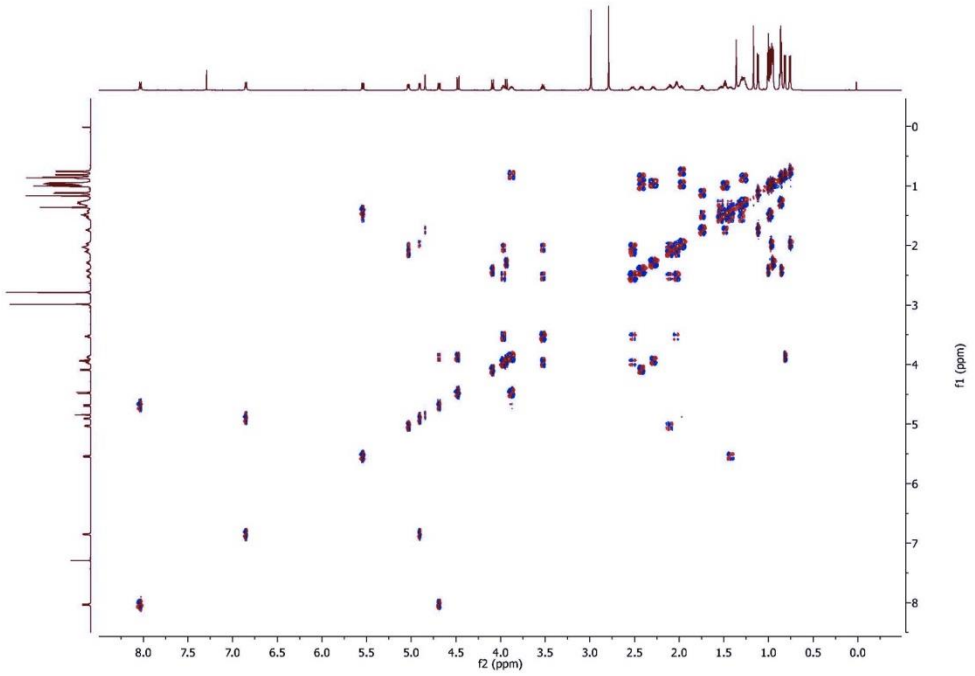


Figure 4.6.10 DQF-COSY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

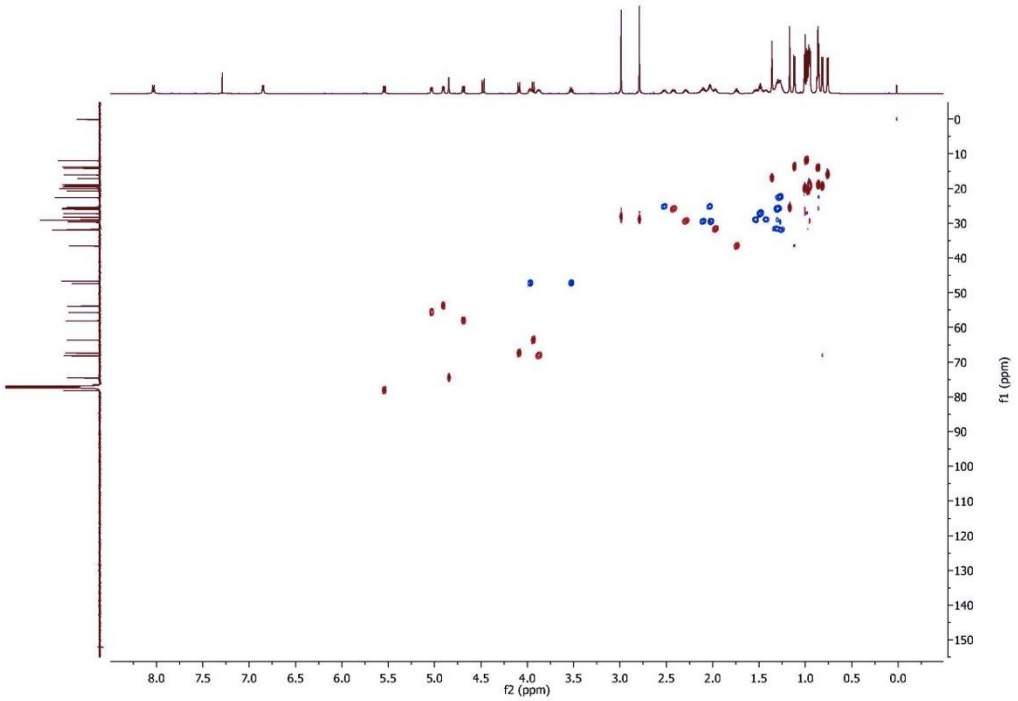


Figure 4.6.11HSQC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

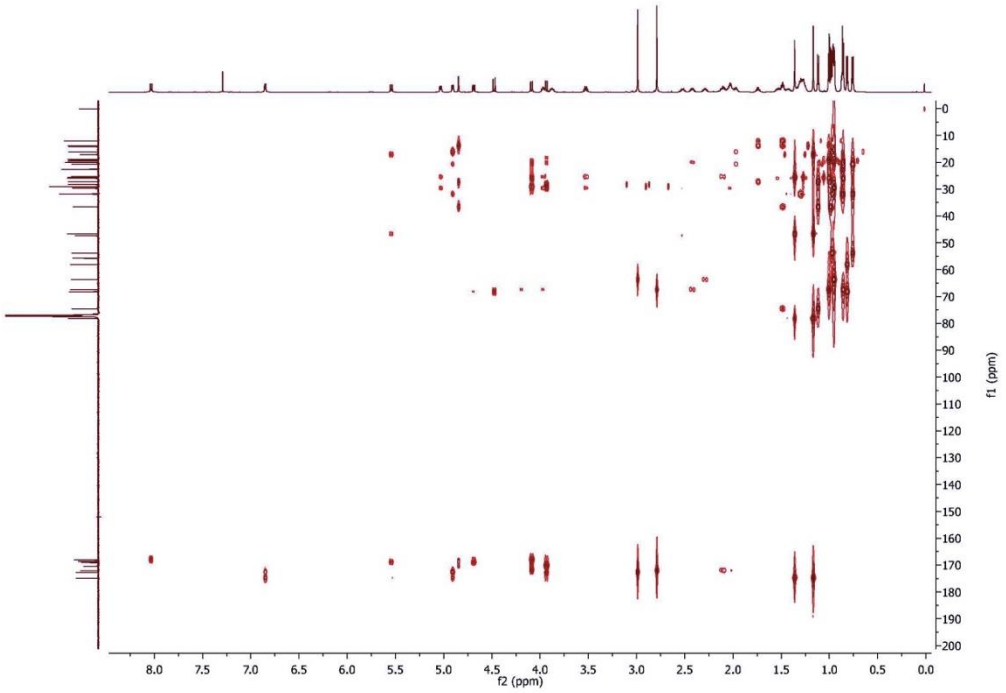


Figure 4.6.12 HMBC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

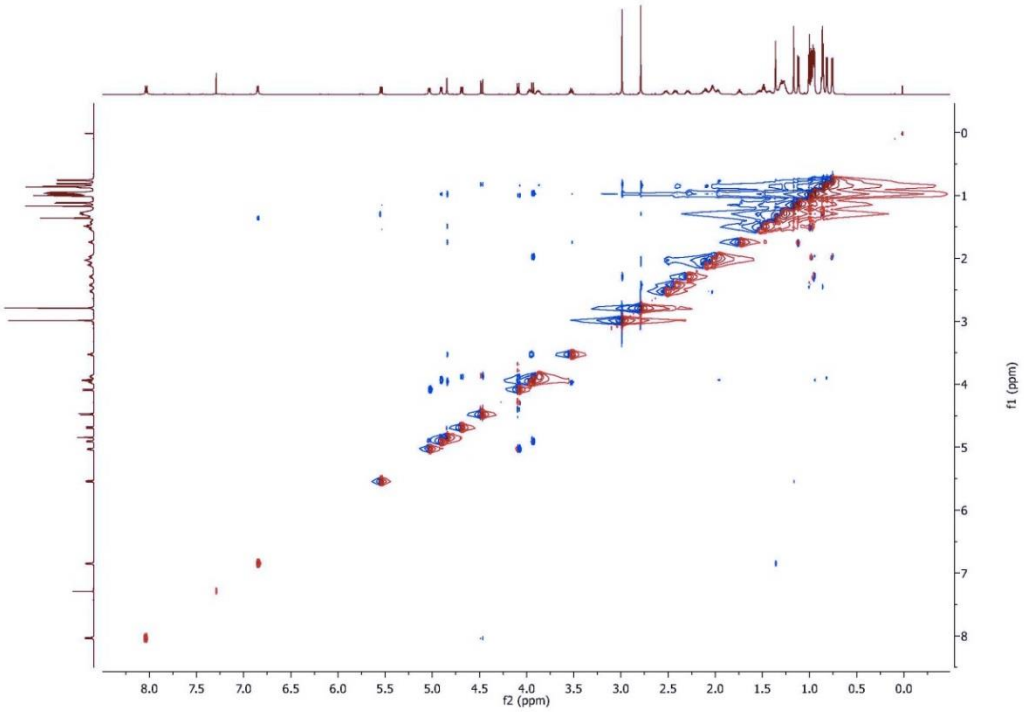


Figure 4.6.13 ROESY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

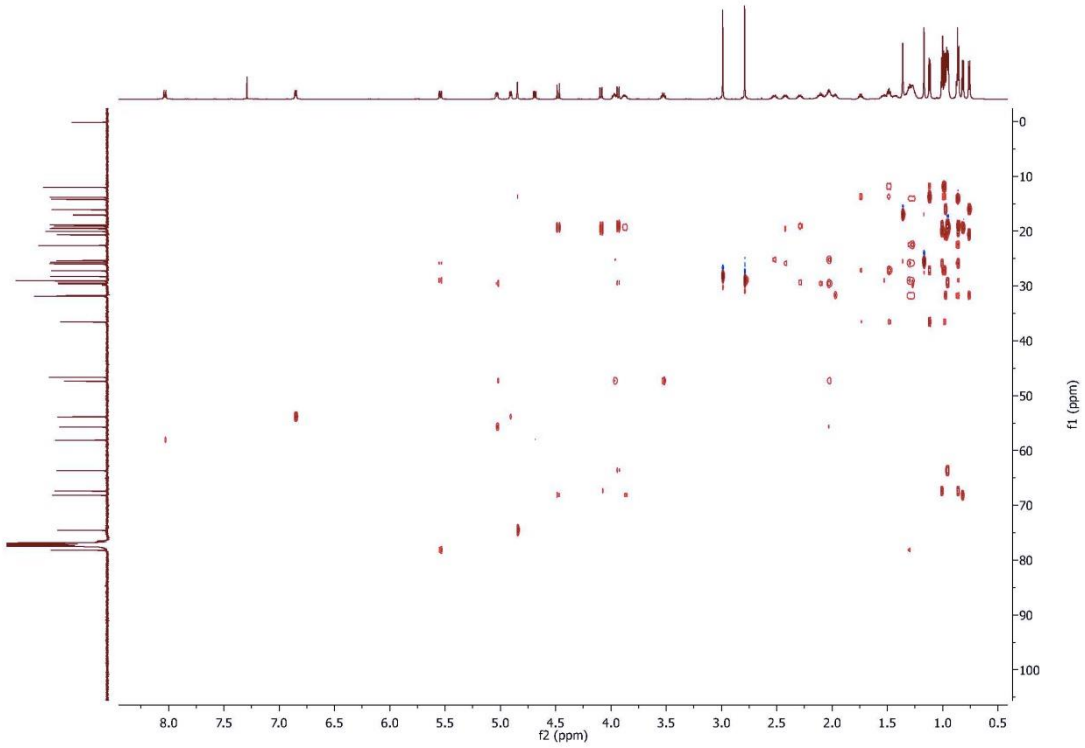


Figure 4.6.14 HSQC-TOCSY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide A3 (**3**).

NMR Spectroscopic Data for Aurilide D (**4**)

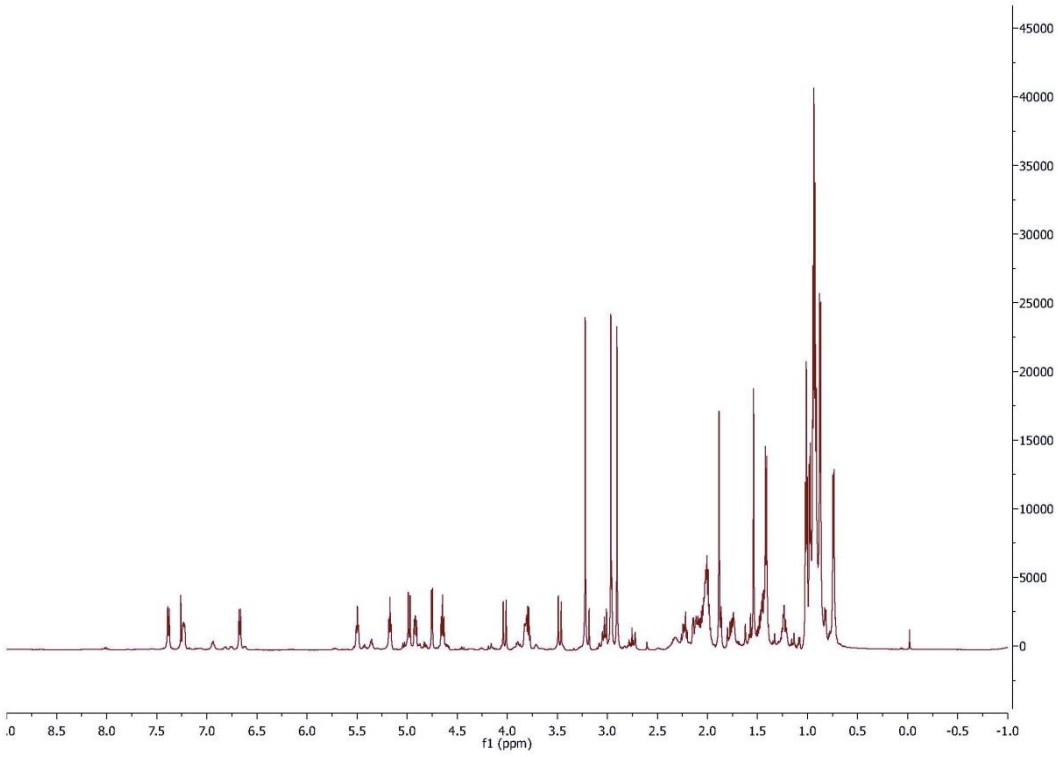


Figure 4.6.15 ^{1}H NMR (600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

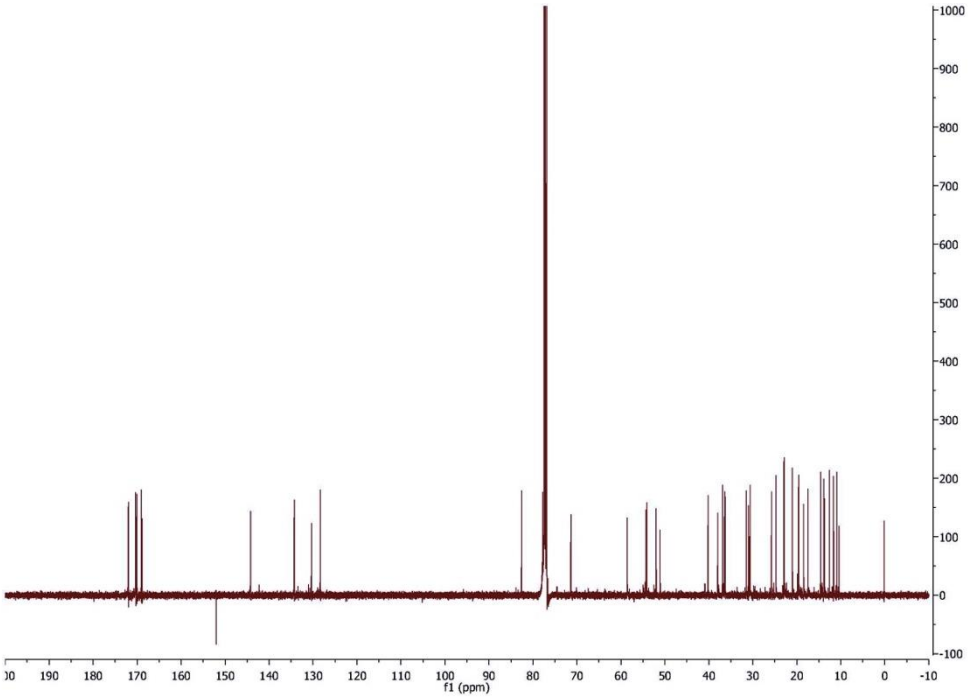


Figure 4.6.16 ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

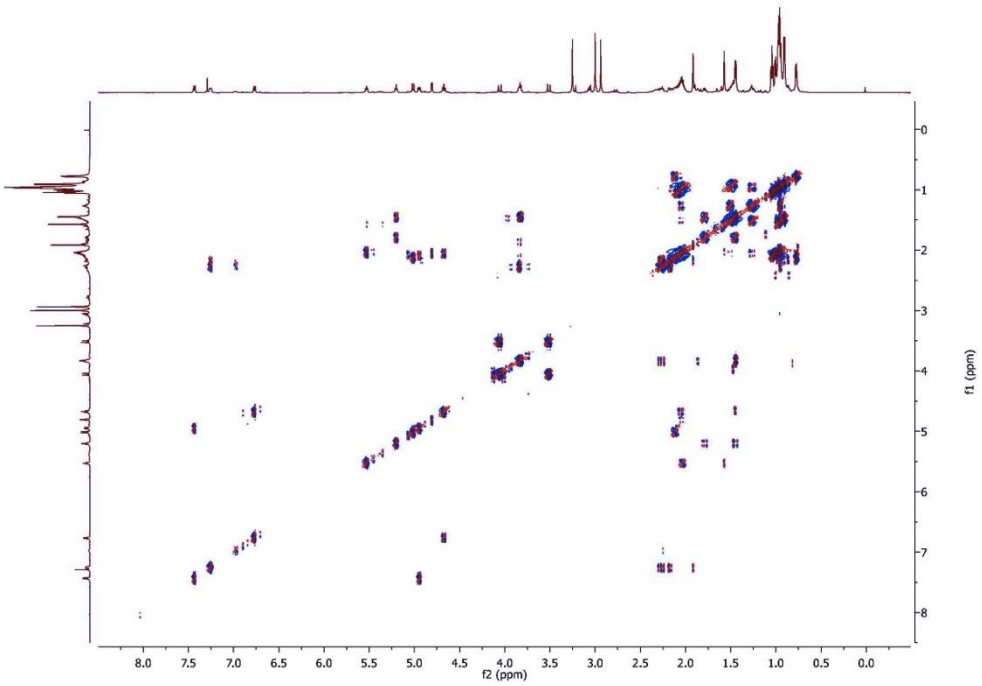


Figure 4.6.17 DQF-COSY (1H 600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

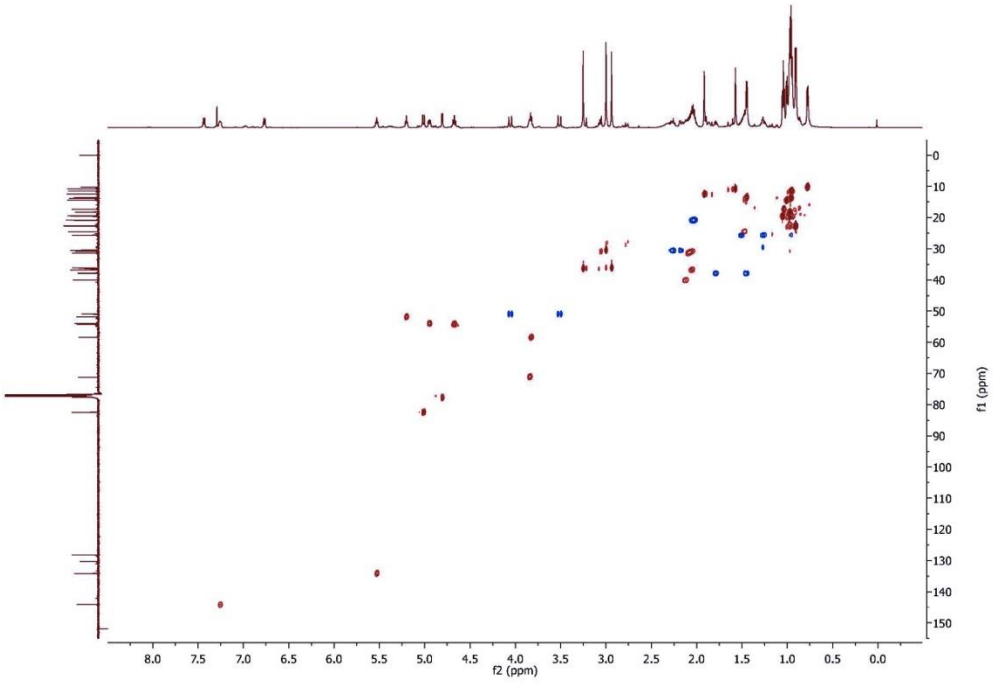


Figure 4.6.18 HSQC (^{1}H 600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

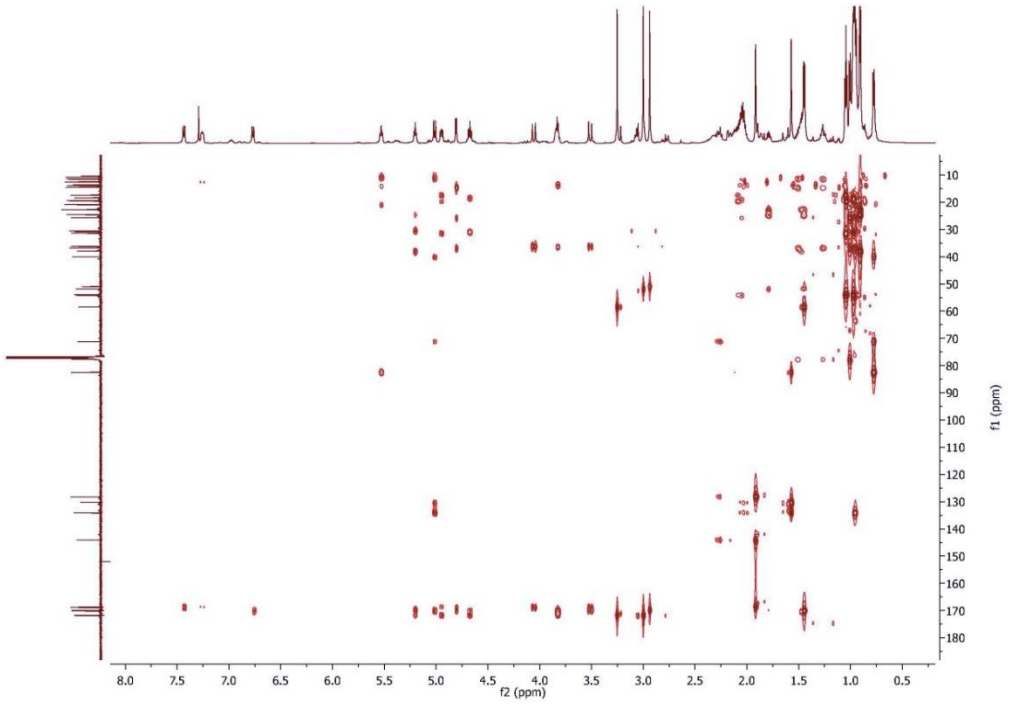


Figure 4.6.19 HMBC (^{1}H 600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

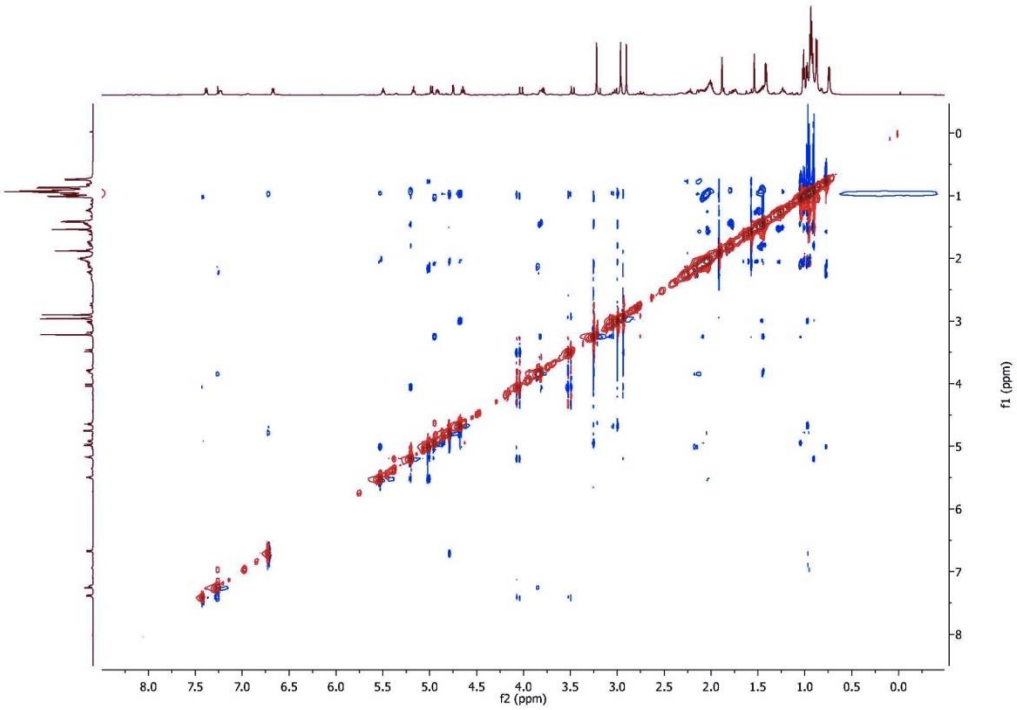


Figure 4.6.20 ROESY (^{1}H 600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

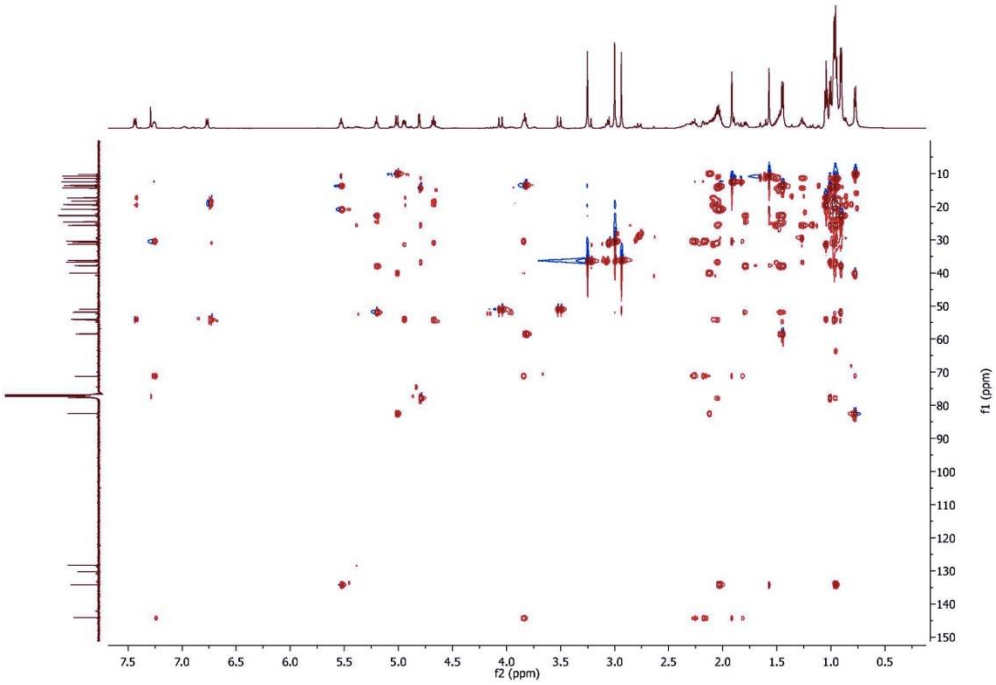


Figure 4.6.21 HSQC-TOCSY (^{1}H 600 MHz, $CDCl_3$) spectrum of Aurilide D (**4**).

NMR Spectroscopic Data for Viequeamide C (**9**)

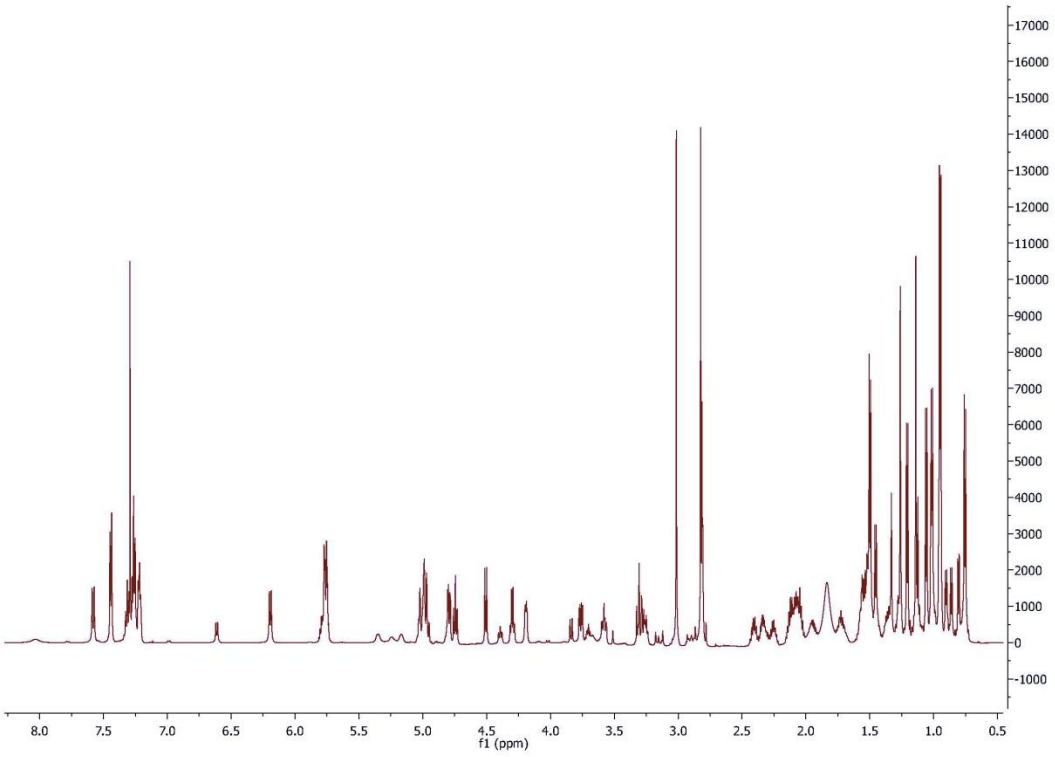


Figure 4.6.22 ^{1}H NMR (600 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

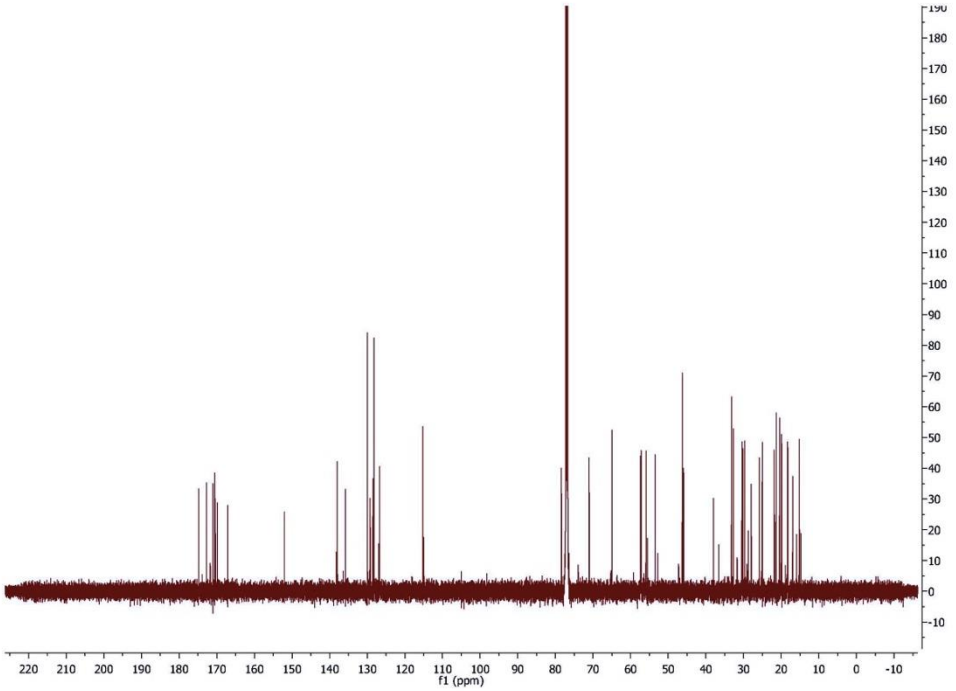


Figure 4.6.23 13C NMR (125 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

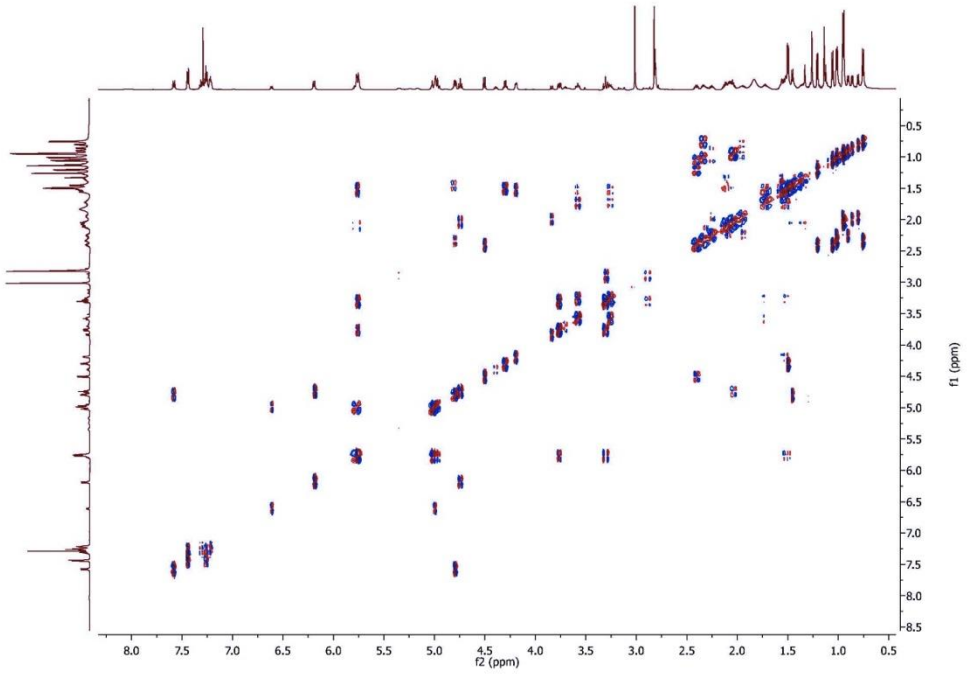


Figure 4.6.24 DQF-COSY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

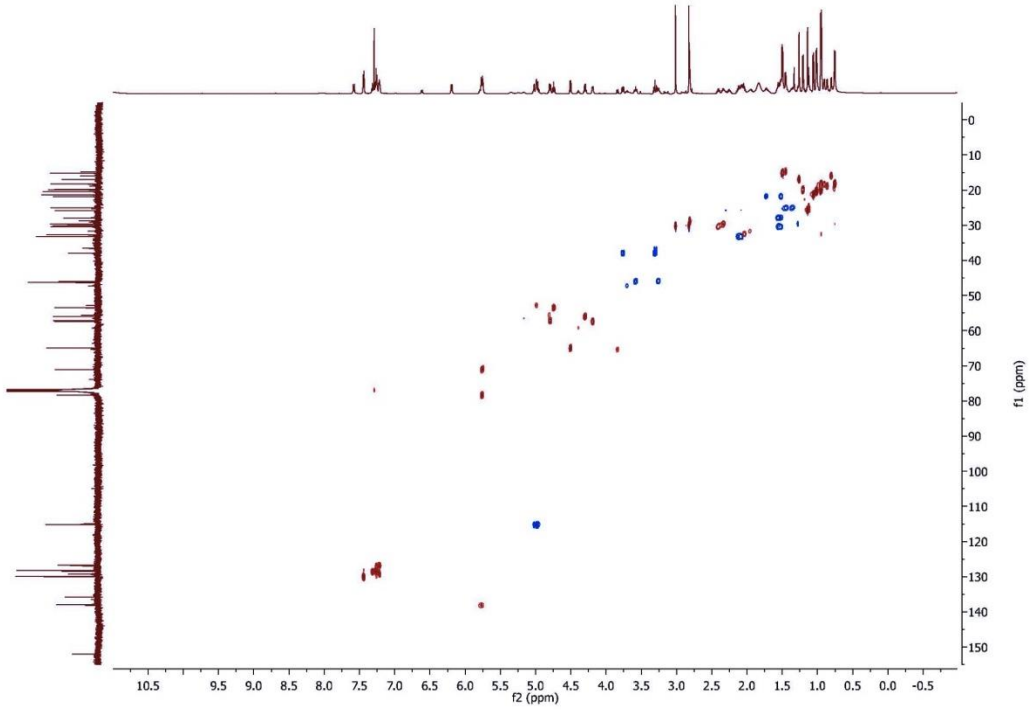


Figure 4.6.25 HSQC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

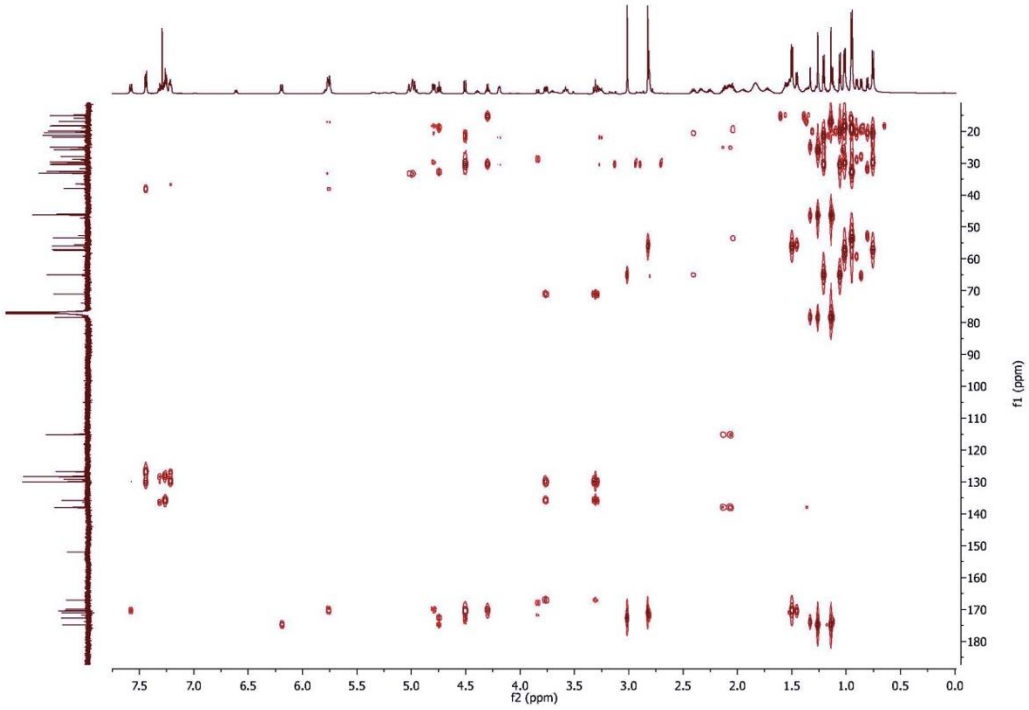


Figure 4.6.26 HMBC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

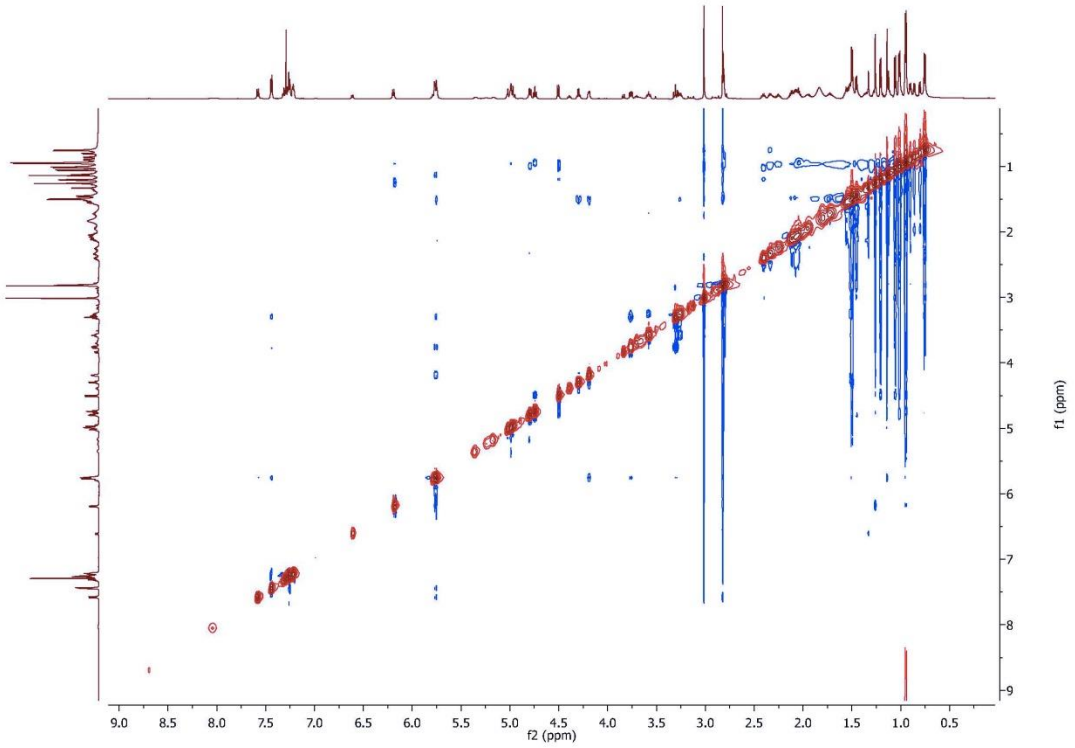


Figure 4.6.27 ROESY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide C (**9**).

NMR Spectroscopic Data for Viequeamide D (**10**)

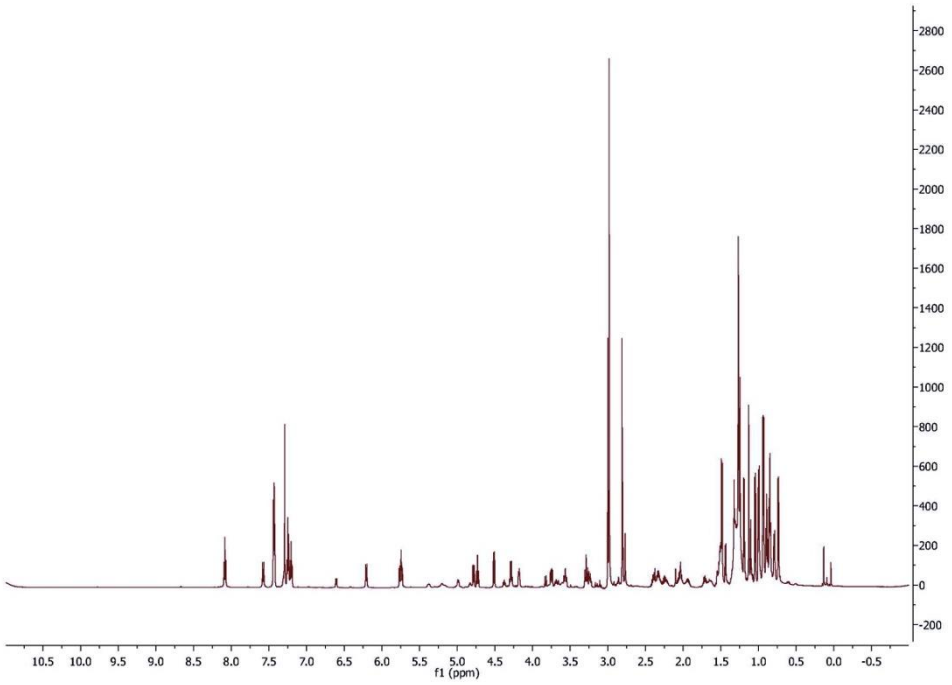


Figure 4.6.28 ^{1}H NMR (600 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

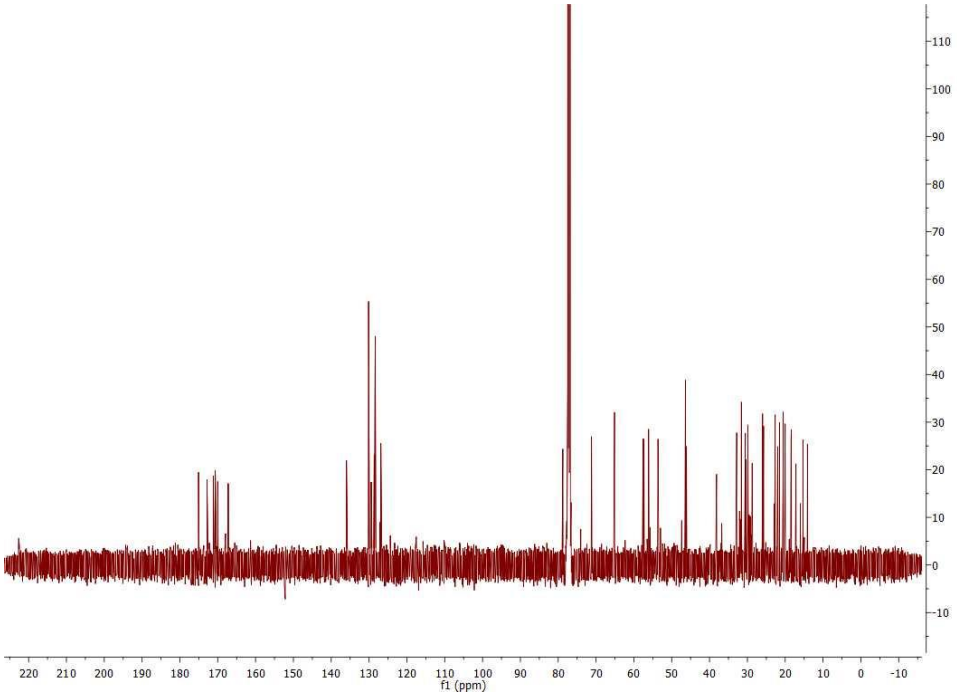


Figure 4.6.29 ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

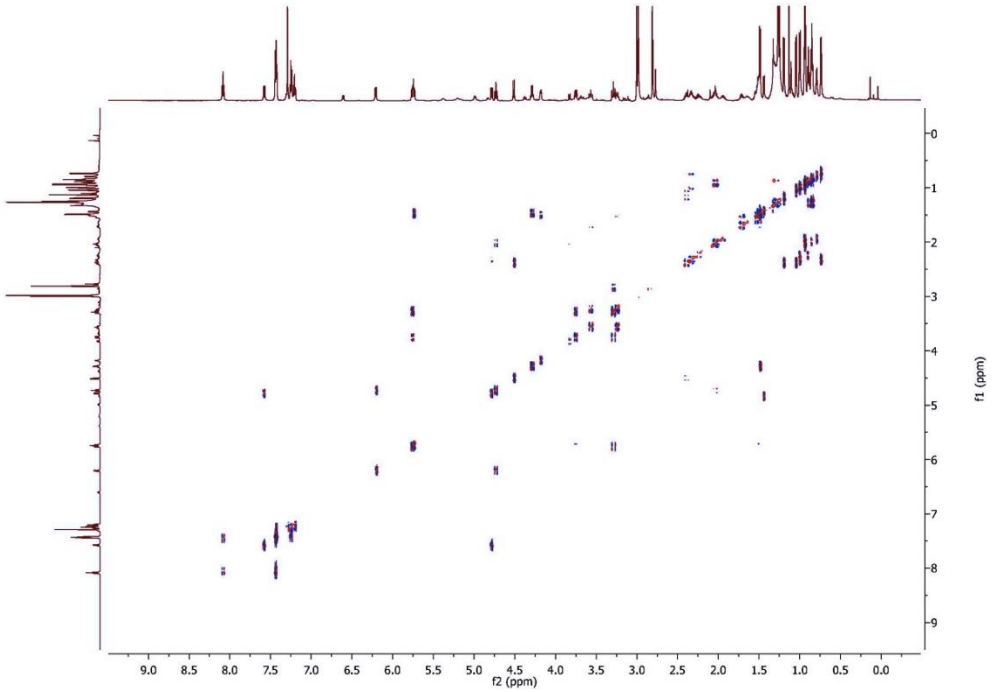


Figure 4.6.30 DQF-COSY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

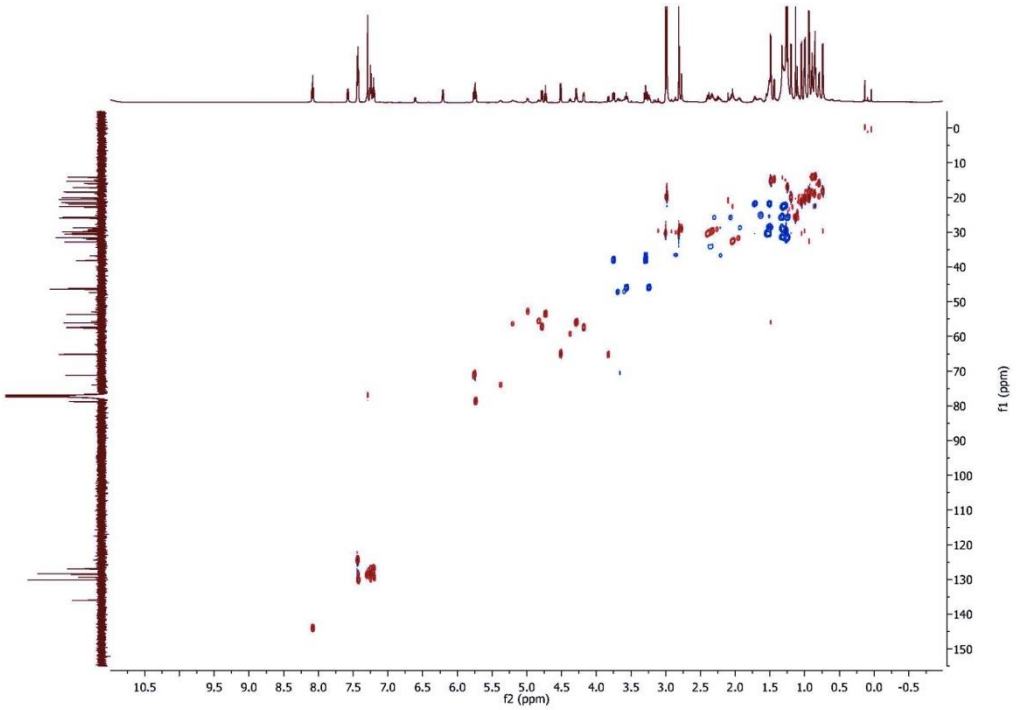


Figure 4.6.31 HSQC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

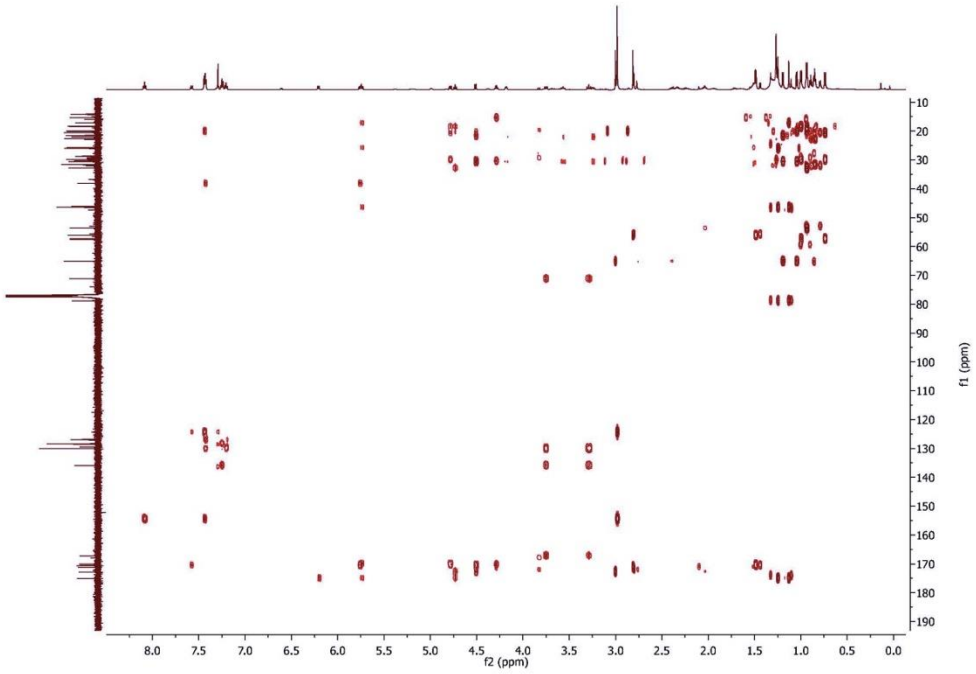


Figure 4.6.32 HMBC (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

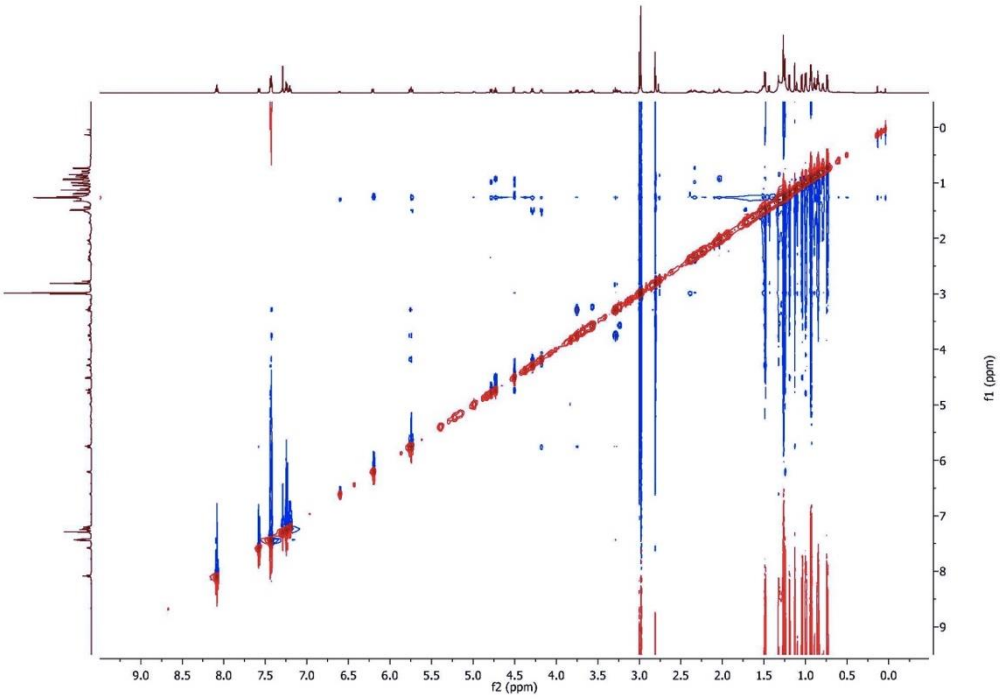


Figure 4.6.33 ROESY (^{1}H 600 MHz, $CDCl_3$) spectrum of viequeamide D (**10**).

Analysis of Absolute Configuration of Aurilide D (4)

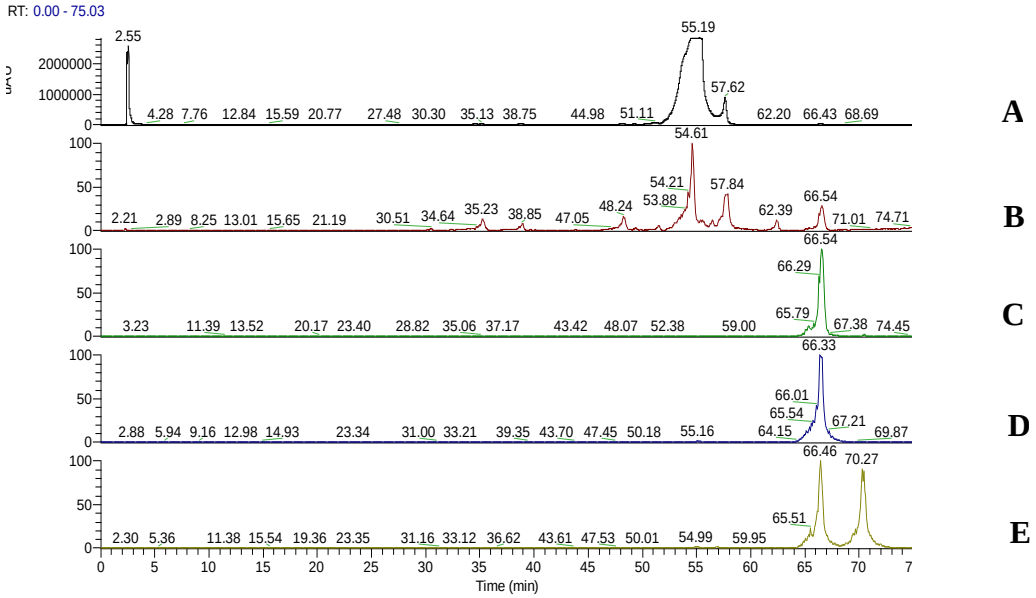


Figure 4.6.34 Marfey's L-FDAA Derivatives of aurilide D hydrolysate and standard *N*-Me-Leu via LCMS. **A**: UV of aurilide D hydrolysate; **B**: Positive ion trace of aurilide D hydrolysate; **C**: Positive ion trace *m/z* 425.46-426.46 of aurilide D hydrolysate (*N*-Me-Leu); **D**: Positive ion trace of standard L-*N*-Me-Leu; **E**: Positive ion trace of standard D/L-*N*-Me-Leu.

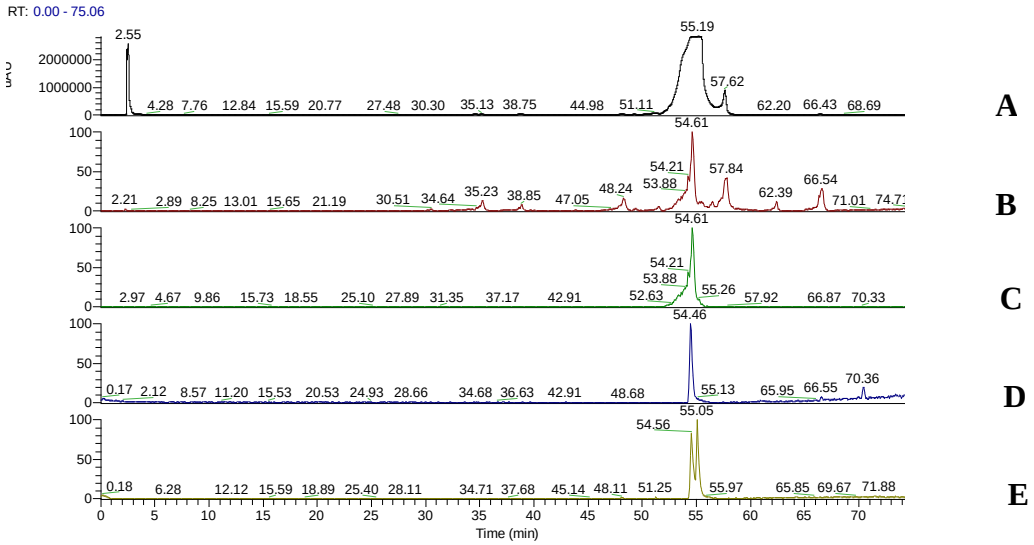


Figure 4.6.35 Marfey's L-FDAA Derivatives of aurilide D hydrolysate and standard *N*-Me-Ala via LCMS. **A**: UV of aurilide D hydrolysate; **B**: Positive ion trace of aurilide D hydrolysate; **C**: Positive ion trace *m/z* 383.45-384.45 of aurilide D hydrolysate (*N*-Me-Ala); **D**: Positive ion trace of standard L-*N*-Me-Ala; **E**: Positive ion trace of standard D/L-*N*-Me-Ala.

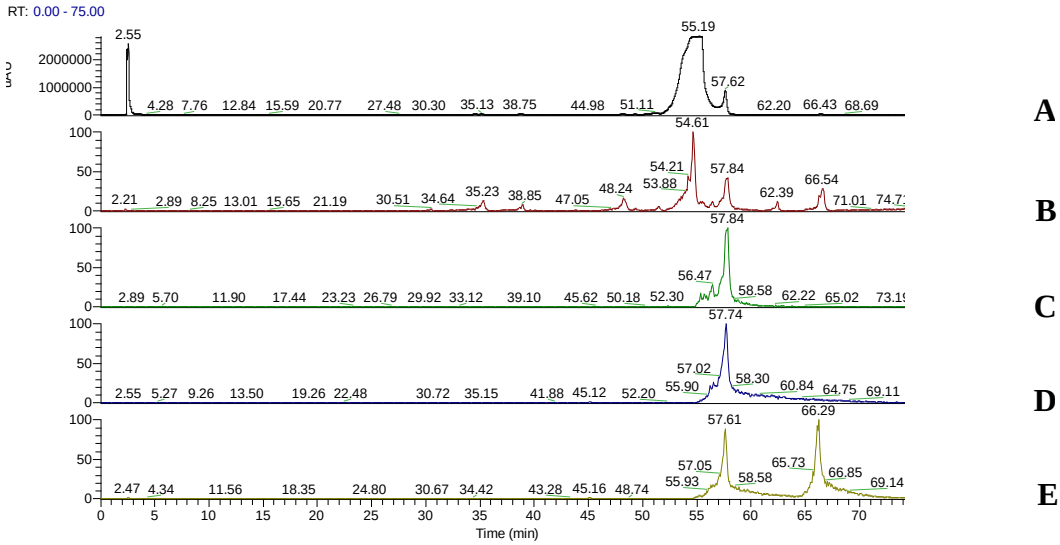


Figure 4.6.36 Marfey's L-FDAA Derivatives of aurilide D hydrolysate and standard Val via LCMS. **A**: UV of aurilide D hydrolysate; **B**: Positive ion trace of aurilide D hydrolysate; **C**: Positive ion trace *m/z* 397.30-398.30 of aurilide D hydrolysate (Val); **D**: Positive ion trace of standard L-Val; **E**: Positive ion trace of standard D/L-Val.

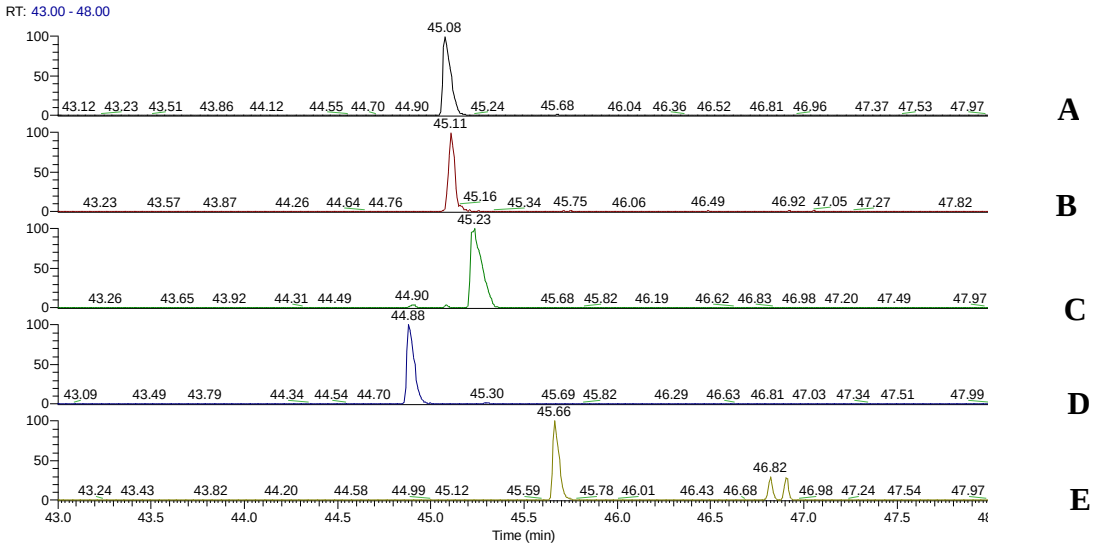


Figure 4.6.37 Steroanalysis of Hmpa residue via GCMS. **A**: Positive ion trace of *m/z* 89.5-90.5 of aurilide D hydrolysate; **B**: Positive ion trace *m/z* 89.5-90.5 of authentic standard 2*R*, 3*S*-Hmpa; **C**: Positive ion trace *m/z* 89.5-90.5 of authentic standard 2*R*, 3*R*-Hmpa; **D**: Positive ion trace *m/z* 89.5-90.5 of authentic standard 2*S*, 3*R*-Hmpa; **E**: Positive ion trace *m/z* 89.5-90.5 of authentic standard 2*S*, 3*S*-Hmpa.

X-ray crystallographic data for Viequeamide A (1) and B (8)

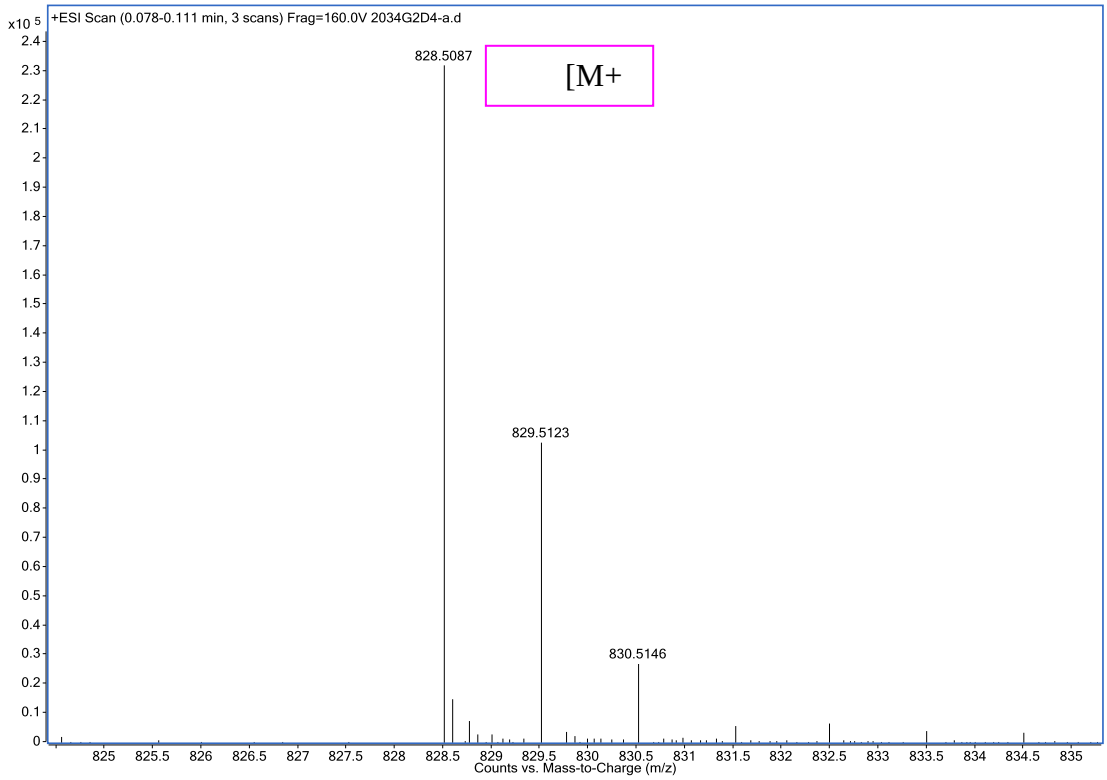
Table 4.6.1 Crystal data and structure refinement for Viequeamide A (**1**).

Identification code	gerw15_sq
Empirical formula	C43.50H72 N5O10.50 (mol of **1**and 0.5 acetone)
Formula weight	833.06
Temperature	100.0 K
Wavelength	1.54178 Å
Crystal system	Monoclinic
Space group	P 1 21 1
Unit cell dimensions	a = 14.6362(6) Å = 90°. b = 10.5766(4) Å = 92.558(2)°. c = 31.0319(14) Å = 90°.
Volume	4799.0(3) $Å^3$
Z, Z'	4, 2
Density (calculated)	1.153 Mg/m^3
Absorption coefficient	0.668 mm^{-1}
F(000)	1808
Crystal size	0.33 x 0.30 x 0.08 mm^3
Theta range for data collection	3.022 to 68.971°.
Index ranges	-17<=h<=17, -12<=k<=11, -37<=l<=36
Reflections collected	82162
Independent reflections	17186 [R(int) = 0.0496]
Completeness to theta = 67.679°	99.5 %
Absorption correction	Semi-empirical from equivalents
Max. and min. transmission	0.668 and 0.633
Refinement method	Full-matrix least-squares on F^2
Data / restraints / parameters	17186 / 1 / 1092
Goodness-of-fit on F^2	1.053
Final R indices [I>2sigma(I)]	R1 = 0.0695, wR2 = 0.1518
R indices (all data)	R1 = 0.0733, wR2 = 0.1537
Absolute structure parameter	0.11(5) [abs stereochem. confirmed]
Extinction coefficient	n/a
Largest diff. peak and hole	0.366 and -0.560 $e.Å^{-3}$

Table 4.6.2 Crystal data and structure refinement for Viequeamide B (**8**).

Identification code	gerw12	
Empirical formula	C44 H65 N5 O9	
Formula weight	808.01	
Temperature	100(2) K	
Wavelength	1.54178 Å	
Crystal system	Orthorhombic	
Space group	P 21 21 21	
Unit cell dimensions	a = 9.4670(3) Å	= 90°.
	b = 12.2088(4) Å	= 90°.
	c = 37.4659(11) Å	= 90°.
Volume	4330.3(2) $Å^3$	
Z	4	
Density (calculated)	1.239 Mg/m^3	
Absorption coefficient	0.702 mm^{-1}	
F(000)	1744	
Crystal size	0.340 x 0.300 x 0.260 mm^3	
Theta range for data collection	4.721 to 68.837°.	
Index ranges	-11<=h<=11, -14<=k<=11, -39<=l<=44	
Reflections collected	21438	
Independent reflections	7753 [R(int) = 0.0255]	
Completeness to theta = 66.500°	98.9 %	
Absorption correction	Multi-scan	
Refinement method	Full-matrix least-squares on F^2	
Data / restraints / parameters	7753 / 0 / 533	
Goodness-of-fit on F^2	1.064	
Final R indices [I>2sigma(I)]	R1 = 0.0353, wR2 = 0.0951	
R indices (all data)	R1 = 0.0357, wR2 = 0.0954	
Absolute structure parameter	0.09(4)	
Extinction coefficient	n/a	
Largest diff. peak and hole	0.354 and -0.314 $e.Å^{-3}$	

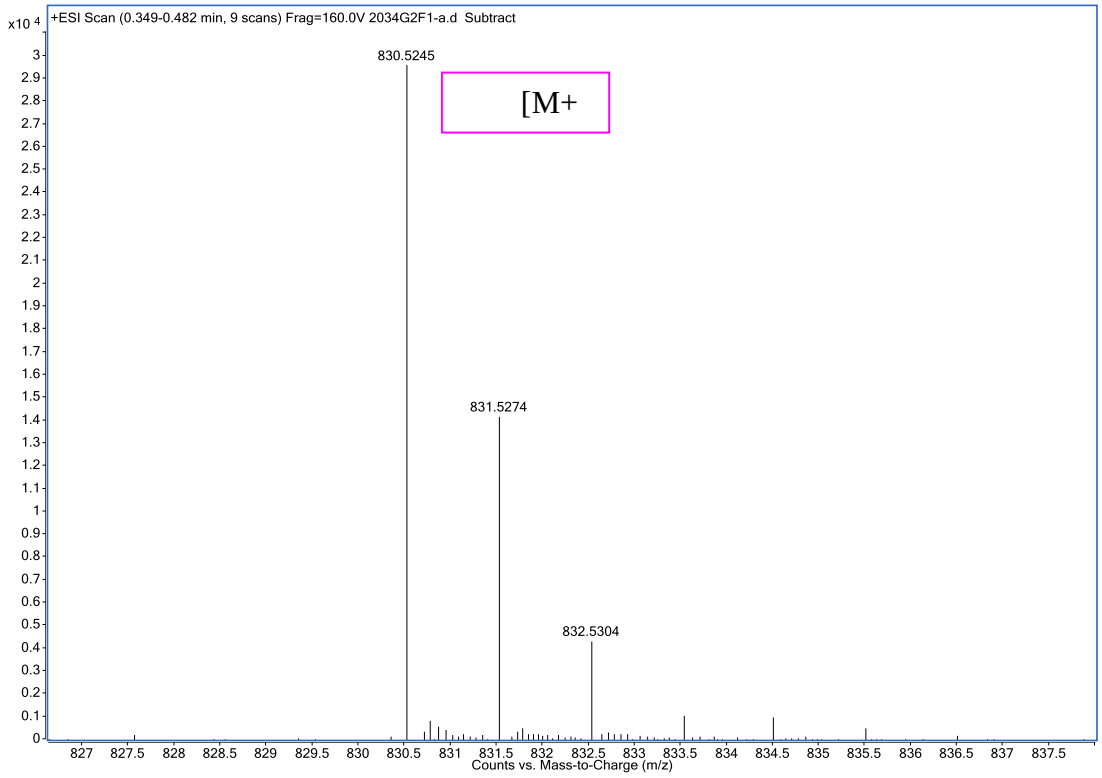
HRMS data for Viequeamide A2, A3, C, D and Aurilide D



Search Results:

Mass Measured	Theo. Mass	Delta (ppm)	Composition
828.5087	828.5093	-0.7	$[C_{42}H_{71}N_5O_{10}Na]^+$

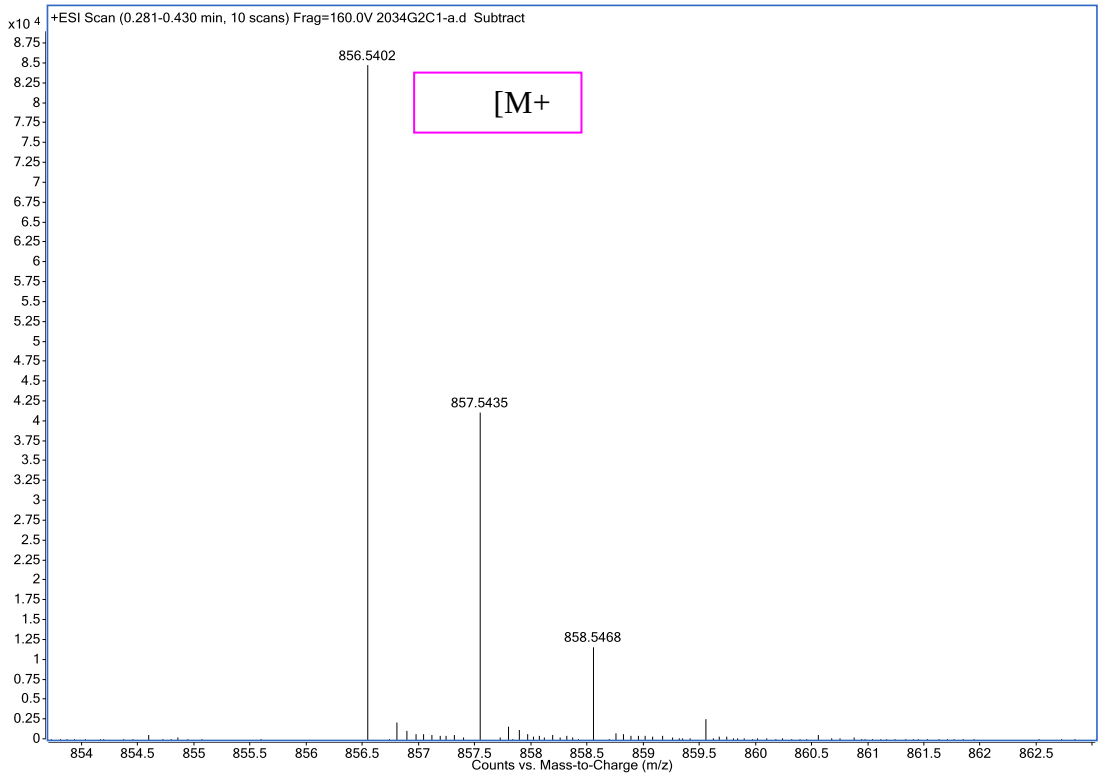
Figure 4.6.38 HR-ESI-TOFMS data of viequeamide A2.



Search Results:

Mass Measured	**Theo. Mass**	**Delta (ppm)**	**Composition**
830.5245	830.5250	-0.6	$[C_{42}H_{73}N_5O_{10}Na]^+$

Figure 4.6.39 HR-ESI-TOFMS data of viequeamide A3.



Search Results:

Mass Measured	Theo. Mass	Delta (ppm)	Composition
856.5402	856.5406	-0.5	$[C_{44}H_{75}N_5O_{10}Na]^+$

Figure 4.6.40 HR-ESI-TOFMS data of aurilide D.

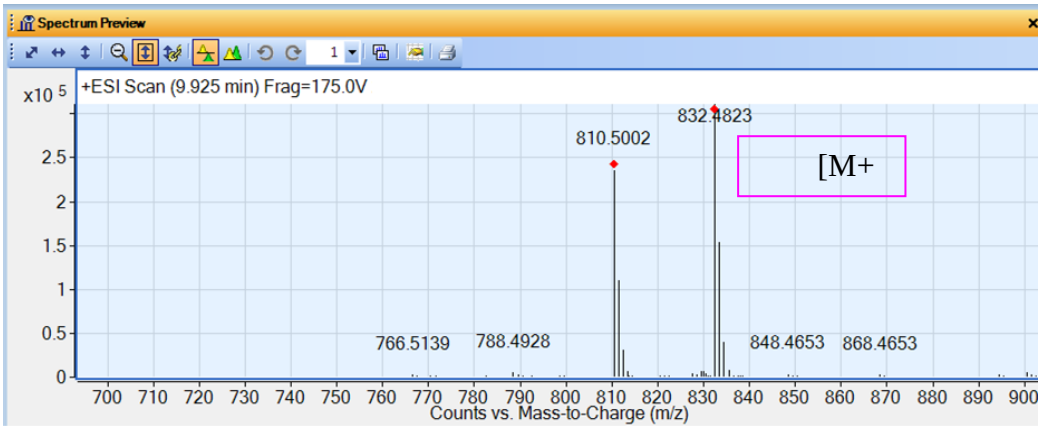


Figure 4.6.41 HR-ESI-TOFMS data of viequeamide C

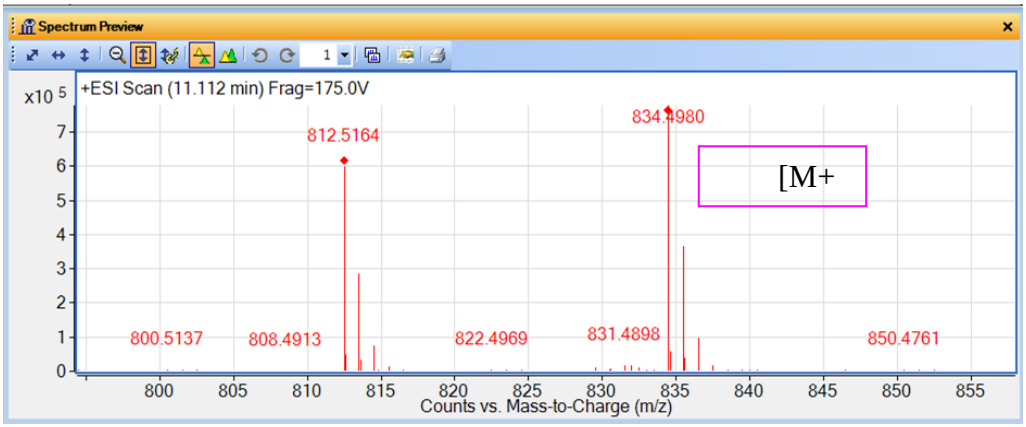


Figure 4.6.42 HR-ESI-TOFMS data of viequeamide D.

Biological Assays

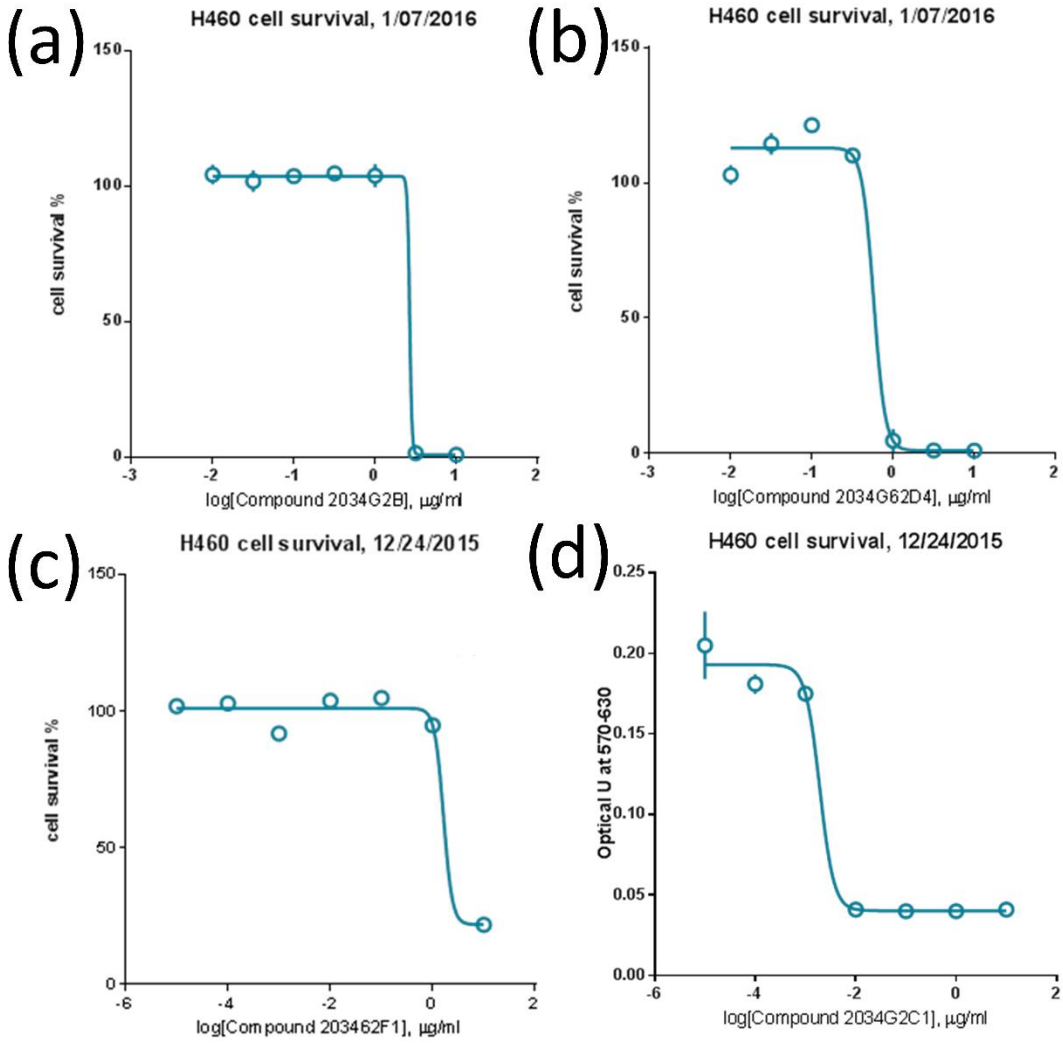


Figure 4.6.43 H460 bioassay dose response curves for (a) viequeamide A (**1**) (IC_{50} = 4.23 $\pm$ 0.171 µM), (b) viequeamide A2 (**2**) (IC_{50} = 0.62 $\pm$ 0.046 µM), (c) viequeamide A3 (**3**) (IC_{50} = 1.98 $\pm$ 0.038 µM) and (d) aurilide D (**4**) (IC_{50} = 2.3 $\pm$ 0.10 nM).

4.7 Chapter 4 References

1. Siegel, R. L.; Miller, K. D.; Jemal, A., Cancer Statistics, 2015. *Ca-a Cancer Journal for Clinicians* **2015,** 65, (1), 5-29.

2. Naman, C. B.; Rattan, R.; Nikoulina, S. E.; Lee, J.; Miller, B. W.; Moss, N. A.; Armstrong, L.; Boudreau, P. D.; Debonsi, H. M.; Valeriote, F. A.; Dorrestein, P. C.; Gerwick, W. H., Integrating Molecular Networking and Biological Assays To Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an American Samoan Marine Cyanobacterium. *Journal of Natural Products* **2017**.

3. Kersten, R. D.; Yang, Y. L.; Xu, Y. Q.; Cimermancic, P.; Nam, S. J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C., A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature Chemical Biology* **2011,** 7, (11), 794-802.

4. Kersten, R. D.; Ziemert, N.; Gonzalez, D. J.; Duggan, B. M.; Nizet, V.; Dorrestein, P. C.; Moore, B. S., Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences of the United States of America* **2013,** 110, (47), E4407-E4416.

5. Wang, M. X.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Criisemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J. Q.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrovr, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.;

Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W. Y.; Liu, X. T.; Zhang, L. X.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016,** 34, (8), 828-837.

6. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **2017,** 13, (1), 30-37.

7. Schroeder, F. C.; Taggi, A. E.; Gronquist, M.; Malik, R. U.; Grant, J. B.; Eisner, T.; Meinwald, J., NMR-spectroscopic screening of spider venom reveals sulfated nucleosides as major components for the brown recluse and related species. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (38), 14283-14287.

8. Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S., Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Analytical Chemistry* **2011,** 83, (5), 1649-1657.

9. Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J., Toward More Reliable 13C and 1H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *Journal of Chemical Information and Modeling* **2008,** 48, (1), 128-134.

10. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H., Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci Rep* **2017,** 7, (1), 14243.

11. Moore, R. E., Cyclic peptides and depsipeptides from cyanobacteria: A review. *Journal of Industrial Microbiology* **1996,** 16, (2), 134-143.

12. Boudreau, P. D.; Byrum, T.; Liu, W. T.; Dorrestein, P. C.; Gerwick, W. H., Viequeamide A, a Cytotoxic Member of the Kulolide Superfamily of Cyclic Depsipeptides from a Marine Button Cyanobacterium. *Journal of Natural Products* **2012,** 75, (9), 1560-1570.

13. Wang, D. Y.; Song, S. S.; Tian, Y.; Xu, Y. J.; Miao, Z. H.; Zhang, A., Total Synthesis of the Marine Cyclic Depsipeptide Viequeamide A. *Journal of Natural Products* **2013,** 76, (5), 974-978.

14. Welker, M.; von Dohren, H., Cyanobacterial peptides - Nature's own combinatorial biosynthesis. *Fems Microbiology Reviews* **2006,** 30, (4), 530-563.

15. Calteau, A.; Fewer, D. P.; Latifi, A.; Coursin, T.; Laurent, T.; Jokela, J.; Kerfeld, C. A.; Sivonen, K.; Piel, J.; Gugger, M., Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *Bmc Genomics* **2014,** 15.

16. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X. T.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L. X.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C., Molecular networking as a dereplication strategy. *Journal of Natural Products* **2013,** 76, (9), 1686-1699.

17. Zhu, X. J.; Liu, J.; Zhang, W. J., De novo biosynthesis of terminal alkyne-labeled natural products. *Nature Chemical Biology* **2015,** 11, (2), 115-U51.

18. Edwards, D. J.; Marquez, B. L.; Nogle, L. M.; McPhail, K.; Goeger, D. E.; Roberts, M. A.; Gerwick, W. H., Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chemistry & Biology* **2004,** 11, (6), 817-833.

19. Reese, M. T.; Gulavita, N. K.; Nakao, Y.; Hamann, M. T.; Yoshida, W. Y.; Coval, S. J.; Scheuer, P. J., Kulolide: A cytotoxic depsipeptide from a cephalaspidean mollusk, Philinopsis speciosa. *Journal of the American Chemical Society* **1996,** 118, (45), 11081-11084.

20. Luesch, H.; Pangilinan, R.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J., Pitipeptolides A and B, new cyclodepsipeptides from the marine cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **2001,** 64, (3), 304-307.

21. Montaser, R.; Paul, V. J.; Luesch, H., Pitipeptolides C-F, antimycobacterial cyclodepsipeptides from the marine cyanobacterium Lyngbya majuscula from Guam. *Phytochemistry* **2011,** 72, (16), 2068-2074.

22. Nogle, L. M.; Gerwick, W. H., Isolation of four new cyclic depsipeptides, antanapeptins A-D, and dolastatin 16 from a Madagascan collection of Lyngbya majuscula. *Journal of Natural Products* **2002,** 65, (1), 21-24.

23. Gunasekera, S. P.; Owle, C. S.; Montaser, R.; Luesch, H.; Paul, V. J., Malyngamide 3 and Cocosamides A and B from the Marine Cyanobacterium Lyngbya majuscula from Cocos Lagoon, Guam. *Journal of Natural Products* **2011,** 74, (4), 871-876.

24. Suenaga, K.; Mutou, T.; Shibata, T.; Itoh, T.; Fujita, T.; Takada, N.; Hayamizu, K.; Takagi, M.; Irifune, T.; Kigoshi, H.; Yamada, K., Aurilide, a cytotoxic depsipeptide from the sea hare Dolabella auricularia: isolation, structure determination, synthesis, and biological activity. *Tetrahedron* **2004,** 60, (38), 8509-8527.

25. Mevers, E.; Liu, W. T.; Engene, N.; Mohimani, H.; Byrum, T.; Pevzner, P. A.; Dorrestein, P. C.; Spadafora, C.; Gerwick, W. H., Cytotoxic Veraguamides, Alkynyl Bromide-Containing Cyclic Depsipeptides from the Marine Cyanobacterium cf. Oscillatoria margaritifera. *Journal of Natural Products* **2011,** 74, (5), 928-936.

26. Gosliner, T.; Behrens, D. W.; Williams, G. C., *Coral reef animals of the Indo-Pacific : animal life from Africa to Hawai**i exclusive of the vertebrates*. Sea Challengers: Monterey, Calif., 1996; p vi, 314 pages.

27. Suenaga, K.; Mutou, T.; Shibata, T.; Itoh, T.; Kigoshi, H.; Yamada, K., Isolation and stereostructure of aurilide, a novel cyclodepsipeptide from the Japanese sea hare Dolabella auricularia. *Tetrahedron Letters* **1996,** 37, (37), 6771-6774.

28. Semenzato, M.; Cogliati, S.; Scorrano, L., Prohibitin(g) Cancer: Aurilide and Killing by Opa1-Dependent Cristae Remodeling. *Chemistry & Biology* **2011,** 18, (1), 8-9.

29. Sato, S.; Murata, A.; Orihara, T.; Shirakawa, T.; Suenaga, K.; Kigoshi, H.; Uesugi, M., Marine Natural Product Aurilide Activates the OPA1-Mediated Apoptosis by Binding to Prohibitin. *Chemistry & Biology* **2011,** 18, (1), 131-139.

30. Han, B. N.; Gross, H.; Goeger, D. E.; Mooberry, S. L.; Gerwick, W. H., Aurilides B and C, cancer cell toxins from a Papua New Guinea collection of the

marine cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **2006,** 69, (4), 572-575.

31. Komárek, J., Modern classification of cyanobacteria. In *Cyanobacteria*, John Wiley & Sons, Ltd: 2014; pp 21-39.

32. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *Journal of Biomolecular Nmr* **1995,** 6, (3), 277-293.

33. Hyberts, S. G.; Milbradt, A. G.; Wagner, A. B.; Arthanari, H.; Wagner, G., Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. *Journal of Biomolecular Nmr* **2012,** 52, (4), 315-327.

34. 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., Feasibility of Drug Screening with Panels of Human-Tumor Cell-Lines Using a Microculture Tetrazolium Assay. *Cancer Research* **1988,** 48, (3), 589-601.

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Summary of the Work Presented in this Dissertation and Future Work

The major goal of the research present herein was to accelerate marine natural products discovery, leveraging the advantages of two cutting edge technologies, Fast NMR and deep learning, in the hope of streamlining the drug discovery process. Chapters 2 through 4 addressed this goal by the creation and characterization of Small Molecule Accurate Recognition Technology (SMART) to associate unknown compounds to their structural analogues both quickly and accurately, even in the presence of 2D NMR spectral noise. A secondary goal involved an application of the novel tool SMART to isolate and structurally characterize cancer cell cytotoxic marine cyanobacteria-derived lipopeptides (Figure 6.1). These compounds were biologically evaluated with several of them possessing potent cytotoxicity against H-460 (human lung cancer cells). A brief summary of each research chapter and future work for each part is described below.

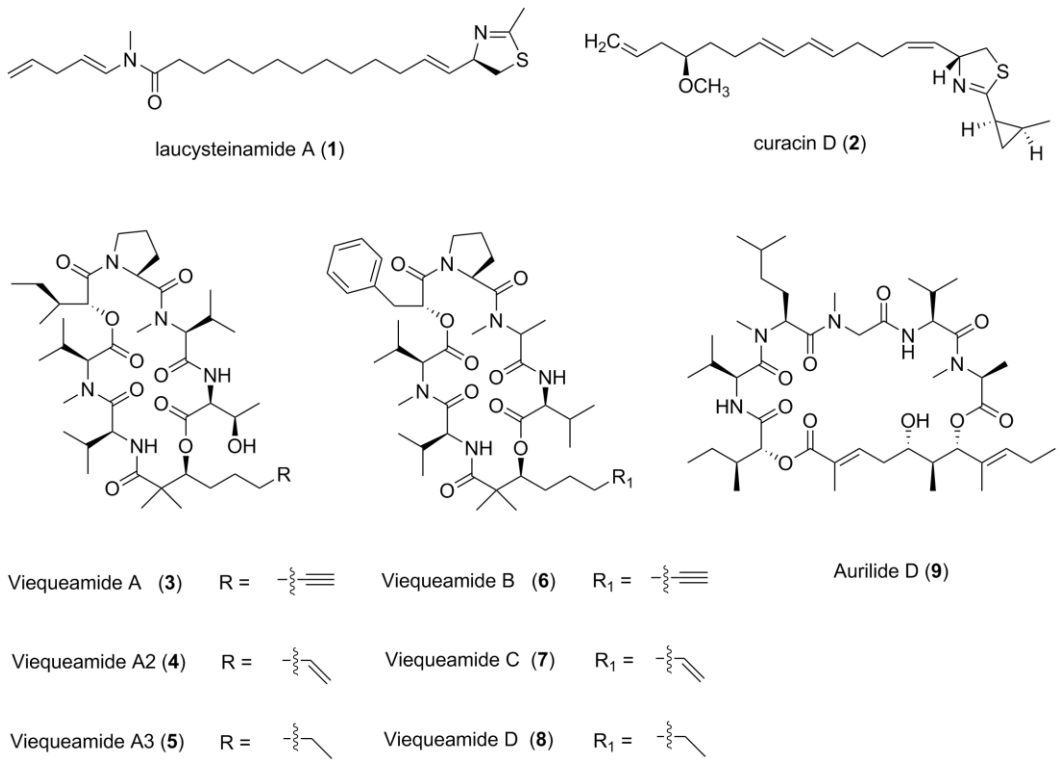


Figure 5.1. Cyanobacterial natural products that were discussed in the previous research chapters.

Chapter 2 discusses the isolation and structural determination of a novel hybrid PKS/NRPS alkaloid, laucysteinamide A, isolated from an extract of the marine cyanobacterium cf. *Caldora penicillata*. The sample was collected from Lau Lau Bay, off the central east coast of Saipan in the Western Pacific Ocean. Two pre-fractions of the crude extracts showed strong cytotoxic activity in the brine shrimp toxicity assay. Molecular Networking, a MS/MS-based dereplication tool, suggested the presence of curacin D in the two prefrctions; however, an analysis of the 1D ^{1}H NMR of one of the prefractions suggested the presence of a structural analogue(s) of the previously reported kalkitoxin, somocystinamide A or curacin D. Ultimately, purification of this prefraction led to the isolation of laucysteinamide A and a known compound, curacin

D. The planar structure of laucysteinamide A was determined using 1D and 2D NMR techniques, yielding its intriguing structural features, a thiazoline ring and a terminal olefin attached to an *N*-methylated amide. The configuration of the thiazoline ring was resolved by referring to the exciton coupling circular dichroism spectrum of the compound. Laucysteinamide A exhibits mild cytotoxic activity against the H-460 human cancer cell line (IC_{50} = 11 μM)[1].

Future directions for this research could include a chemical synthesis of laucysteinamide A in order to obtain more material for a broader biological evaluation. A dimer of laucysteinamide A, somocystinamide A, was isolated previously from a *L. majuscula/Schizothrix* sp. mixed assemblage collected from Somo Somo, Fiji and it exhibited very potent cytotoxicity. This could potentially be attributed to the existence of a disulfide functional group and the lipopeptide moiety within somocystinamide A, as suggested by these previous studies[2, 3]. However, by substituting the disulfide moiety with a thiazoline moiety, laucysteinamide A, demonstrates only mild H-460 cytotoxicity, which might mean that this novel alkaloid has other biological mechanisms of action than somocystinamide. In this regard, further investigations are required to explore this hypothesis with more materials provided by chemical synthesis. Also from the same collection that gave rise to laucysteinamide A, curacin D, a thiazoline ring containing antimitotic compound, is produced in larger quantities. Curacin D was previously isolated from a Virgin Islands collection of marine cyanobacterium, *Lyngbya majuscula* in the Atlantic Ocean. Thus, it could be of great interest to explore the genetic relationship of the gene clusters that are responsible for curacin D production in the Pacific and Atlantic Oceans. It is thus hoped that full genome sequencing of both the

species that produce curacin D can be performed and an analysis of the result might be fruitful.

Chapter 3 addresses the primary goal by presenting the conceptualization, development and characterization of a novel structural analysis tool, SMART, which is an integration of Fast 2D NMR techniques and deep convolutional learning. Specifically, we leveraged the advantages of Non Uniform Sampling Nuclear Magnetic Resonance (NUS NMR) and Convolutional Neural Networks (CNN) to develop SMART as a tool to associate unknown 2D NMR spectra to their close relatives in the NMR database. Fast NMR techniques like NUS NMR have reduced the sample requirements of NMR instruments while maintaining the same sampling time and quality. Next, over 2000 experimental Heteronuclear Single Quantum Correlation (HSQC) spectra were taken from supplementary information of the *Journal of Natural Products* for the training of the CNN. In return, the deep learning algorithm provided us structurally insightful embedding maps with nodes and clusters representing correlations of related families of natural products. The training result was compared with that of several other machine learning techniques using the same dataset, which showed that this CNN pattern recognition-based method overcomes many challenges, such as solvent effects, instrumental artifacts, and weak signal issues. The trained CNN also demonstrated its capacity to recognize noisy HSQC spectra at high signal to noise ratios. From the work in chapter 4, it is clear that SMART has the ability to accurately categorize newly isolated depsipeptides to their known relatives, namely the viequeamides.

There are numerous directions that this project could foreseeably turn towards in the future. One immediate direction is to further investigate the impact of spectral noise on SMART. Specifically, certain peaks within a given 2D NMR spectra can be deleted and the result in noisy spectra can be tested by SMART. Peaks could be completely deleted or shrunk in size. In the case where the sizes of certain peaks are reduced, the geometric centers of the reduced peaks will be kept at the same coordinates relative to their neighbor peaks. This type of test is needed in order to evaluate the tolerance of small differences among compounds within the same compound family. Secondly, it would be insightful to move peaks of a given HSQC to a new location within a radius *r* of the geometric center of the very peak. The radius *r* is determined by the distance *l* times a factor *i* ranging from 0.1, 0.2, 0.3, 0.4, 0.5. The distance *l* is defined by magnitude of the vector ***L*** directing from the geometric center of the chosen peak to the geometric center of its nearest neighbor peak. However, if any distance l_x, ($x \in [1,n]$) is larger than the mean value $\bar{l}_n$ of $l_n \ldots l_2, l_1$, $[l_n \geq l_{n-1} \ldots l_2 \geq l_1]$, then the values of *r* for the two geometric centers at each end of $\boldsymbol{L_n}$ is assigned by

$$r = \frac{(n-1)\cdot \bar{l}_{n-1} - l_1}{n-2} \cdot i, \qquad i \in [0.1, 0.2, 0.3, 0.4, 0.5]$$

This type of manipulated HSQC spectrum could give useful insights into the influence of solvent induced effects on the accuracy of SMART.

Because the ability to accurately detect organic chemicals is a multi-billion dollar industry, there is a need for rapid algorithms to identify unknown organic molecules as being structurally associated to known molecules. In order to potentially commercialize SMART for the pharmaceutical industry and other specialty chemical

industries, a web based SMART platform is required for users to easily access SMART and contribute to an ever growing 2D NMR database. Might itbe possible to create a Facebook of Molecules using SMART? In this case, the clustering map would be like a Facebook users' friendship list or connections, and the HSQC spectra would be like photos in their family album. The bioactivity or collection data are more like the "About" section or "Eduation", or "Hobbies" sections of Facebook. Unknown molecules could be "tagged" or "invited". "Recognition" of an unknown compound is similar to "finding friends". In this regard, molecules are not things any more. They are incarnated living entities. The Facebook of Molecules will be a community. Molecular structures will evolve to have better biological activities, for example, anti-cancer activity. This can be recorded in their "Timelines", which will give researchers better ideas to design drugs. Multiple constituents in the same drug product would be like marriage to cure a disease. Small molecules could get married to give birth to a product that cures a disease. Or a small molecule can marry a protein, since they bind to cause an effect. This is inter-racial marriage. But before that, small molecules will "date" a few proteins. A virtual drug screening would be like an "Event". Similar chemical structures may be interested in going to the same event. But an event is also open to molecules with novel structures.

Furthermore, the current SMART dataset of 2D NMR spectra utilizes only black and white HSQC spectra. However, as the wide application of phase edited HSQC experiments, $-CH_3$, -CH groups and $-CH_2$ groups in a molecule can be labeled in different colors, which is more informative than conventional 2D HSQC spectra. By training SMART to recognize those colored spectra, SMART could hope to be even more accurate. However, colored HSQC spectra are currently not popular in the *Journal*

of Natural Products as many labs have not had access to the phase edited HSQC techniques. Nevertheless, I envision that phase edited HSQC techniques will be the standard HSQC experiments within a few years, just as color televisions replaced monochrome TVs a few years after its entry to the market.

Chapter 5 addresses the secondary objective of this thesis with the application of SMART to recognize the structure of seven depsipeptides, viequeamides A-A3, B-D, and aurilide D from two separate collections of marine cyanobacteria, a *Rivularia* sp. from Vieques, Puerto Rico, and a *Moorea producens* from American Samoa. The *Rivularia* sp. collection has been particularly rich in secondary metabolites, such as viequeamides A and B, which were previously characterized. However, a follow-up investigation into a relatively polar prefraction of this extract yielded three new viequeamides, viequeamides A2, A3, and aurilide D. The *Moorea producens* collection was found to produce viequeamides B, C and D. Their planar structures were determined by a combination of MS/MS based Global Natural Products Social Molecular Networking (GNPS) and 2D NMR-based SMART. Their absolute configurations were determined by an interplay of LCMS-based Marfey's analysis, chiral-phase GC-MS analysis and X-ray crystallography. These new metabolites were evaluated for cytotoxicity to H-460 human lung cancer cells. Intriguingly, viequeamide A2 with a terminal alkene was around 3-fold more toxic (IC_{50} = 0.62 $\pm$ 0.046 μM) than A3 (IC_{50} = 1.98 $\pm$ 0.038 μM) with a terminal alkane, and around 7-fold more toxic than A (IC_{50} = 4.23 $\pm$ 0.171 μM) with a terminal alkyne. The IC_{50} for aurilide D was 2.3 $\pm$ 0.10 nM to H-460 cells, approximately 6-fold more toxic than its epimer, aurilide A.

Future directions for this research would be to evaluate these compounds for other biological properties, such as anti-infective, anti-inflammatory, neuromodulatory etc.[4, 5]. The isolation of the viequeamides from both the Atlantic Ocean and the Pacific Ocean indicates that the viequeamides are distributed pan-tropically, and suggests that it constitutes an ancient biosynthetic pathway that is present in many different species of cyanobacteria. It would be insightful to obtain full genome sequencing of the two collections, identify the gene clusters that are responsible for viequeamides production, and consider their evolutionary relationship.

5.2 Future Directions of Marine Natural Products Discovery

Natural products research enables and benefits numerous downstream research fields, such as chemical biology, chemical ecology, drug discovery and development, pharmacology and the total chemical synthesis of NPs. For example, approximately 70% of all approved drugs are NPs, their analogues, or a chemical modifications of an existing NP. In addition to these academic and social benefits, natural products research (NPR) provides a powerful incentive for the conservation and sustainable use of biodiversity and biodiverse habitats.

NPR fuels drug discovery and development by providing medicinal researchers with drug leads. The drug leads are then tested in animals like mice to see their desired effects as well as undesired side effects, leading to therapeutic agents for human clinical trials that will have better treatment effects and limited side effects. If several phases of clinical evaluation demonstrate the reliability of the candidate agents, then they will be reviewed by Food and Drug Administration (FDA) for commercial approval. The FDA

can both grant patients access to drugs that save their life as well as deny them the access to drugs that they badly need. At the early stages, the drug developers or inventors have to show the FDA their animal test results, clinical trial results, pharmacological mechanism studies, and most importantly, the structures of the drug molecules.

So the question is, how can we find better drug leads more quickly? One could simply create new chemical constituents and hope to find a biologist willing to test each substance with whatever pharmacological tests are available. This is not considered to be a valid approach, however. As Nature has had many millions of years during which to design molecular structures, a better approach is to 1) collect existing materials from Nature, 2) prepare extracts (a substance made by extracting a part of a raw material, often by using a solvent such as methanol or hexanes), 3) perform a first round of biological screens for all extracts, 4) isolate active and pure chemicals (natural products) from each extract showing biological activity, 5) determine planar structures of the purified compounds, 6) elucidate stereochemistry of the compounds, and 7) test each compound for more in depth types of pharmacological activity (see Figure 6.2). The collection from Nature and broad screening NPR method is a reasonable approach that has produced many useful pharmaceuticals, and should continue to do so, although that is contingent on the availability of adequate funding and appropriate research tools. It is worth pointing out that none of the steps in the workflow can be skipped nor the order of each research step can be altered, for these fit together in a logical way. For example, we cannot perform many pharmacological tests for low abundance chemicals without determining their structures first, otherwise, we may detect a compound with a valuable

property but not be able to determine what it is. This is due to the sample consuming natural of the bioactivity assays.

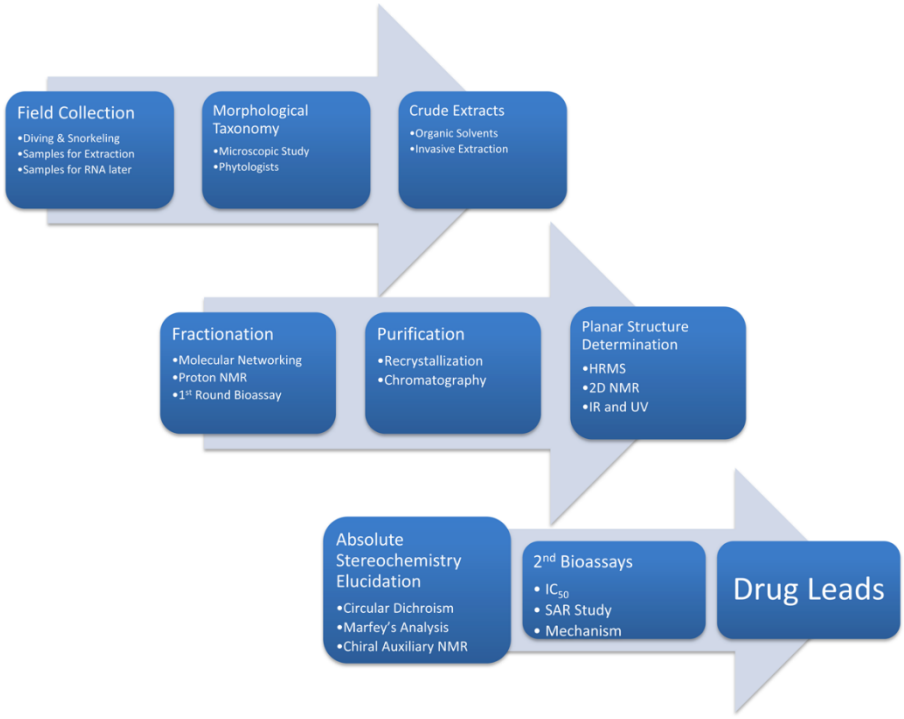


Figure 5.2. An optimized workflow for the marine natural products discovery process.

One of many bottlenecks within this NPR workflow that this thesis deals with is the planar structure determination (step 5). There are several reasons to prioritize the acceleration of this step. First, 2D NMR technique is indispensable in this step. Conventional 2D NMR experiments are time consuming, especially when the sample is scarce. Secondly, 2D NMR spectroscopic-based structural determination can take months to years considering the complexity of natural products. Thirdly, training of experts in structure elucidation of natural products is sometimes deemphasized. In this

regard, we need additional tools like GNPS and SMART for the automatic identification of chemicals or chemical families.

The SMART prototype is the first ever ensemble of 2D NMR and deep CNN. During the laucysteinamide A structural determination project, I kept dreaming of a world I thought I would never see. It took me a great deal of time to run conventional 2D NMR experiments for the structure elucidation of laucysteinamide A and dereplication of curacin D. To solve the structure of laucysteinamide A, weeks were spent carefully analyzing the 1D and 2D data sets, constructing the 3D molecular moieties, and finally piecing them together into the completed laucysteinamide A structure. It was this experience that made me want to find a method that could inform me what compound family laucysteinamide A belonged to from the available 2D spectra. Thus, I envision a future in which MNP researchers use SMART integrated Google Glass to support the heads-up structure dereplication and assignment to molecular structure families, and thus, augment their research capacity.

As a project that brings the future to the present, the SMART will give rise to many spin-off innovations. Practical applications of the metric learning in MNPs research are being realized for projects that require leaps in technology. For example, could we envision the immediate taxonomic identification of a marine cyanobacterium via an image recognition program[6] from a microscopic picture without referring to a marine biology expert or further genotyping? New species discovery would thus be boosted. What about the possibility of automatic phenotyping[7] of antiparasitic bioassays with video recognition? The drug screen efficiency could thus be improved.

Another bottleneck impeding the development of natural products discovery is the process of absolute stereochemistry determination (step 6). Approximately 50% of approved drugs are mixtures of enantiomers or single-enantiomers.[8] Nevertheless, the two enantiomers of a chiral drug may be significantly different in their biological activities, pharmacokinetics, pharmacodynamics (e.g. selectivity for target ligands), and side effects. This is due to the fact that differences in 3D structure may prevent the active enantiomer binding to the pocket due to binding of the inactive enantiomer. Currently, NMR experiments such as rotating frame nuclear Overhauser effect spectroscopy (ROESY) and nuclear Overhauser effect spectroscopy (NOESY) can contribute to the determination of relative stereochemistry of small molecules. However, in order to solve absolute stereochemistry of complex molecules, the commonly used methods include chiral NMR techniques (e.g. Mosher ester analysis), Marfey's analysis, X-ray crystallography, circular dichroism (CD) and optical rotations. Mosher ester analysis requires the presence of a hydroxy group that can be subjected to esterification. Marfey's analysis is often applied to peptides, which consumes sample and requires chiral standards. X-ray crystallography demands a relatively large amount of sample and its crystallization process is very time consuming and entirely uncertain. The circular dichroism method is sensitive and only requires nanomolar amount sample compounds; however, the success of the CD experiments hinges on the existence of UV chromophores near stereocenters in chiral compounds.

Recently, emerging techniques have been developed to quickly determine 3D structures of compounds, while saving samples. For example, the improvement in material sciences and X-ray crystallography have led to a significant increase in

efficiency of the practical use of X-ray based absolute stereochemistry determinations. For example, Inokuma et al.[9] has applied porous $Co(NCS)_2$ and ZnI_2 based crystalline host sponges to absorb nanomolar scale guest compounds in order to facilitate X-ray single-crystal diffraction analysis of those compounds. The resulting complexes allowed the successful determination of absolute 3D structures. Another example is Atomic Force Microscope (AFM) which has made its way into 3D structure determination of small molecules. In 2009, Gross et al.[10] resolved the 3D structures of several aromatic compounds utilizing atomic force tips to probe the aromatic compounds that were absorbed on top of an inorganic substrate. The limitation is that only structures of highly conjugated compounds (e.g. pentacene) can be directly imaged by this method. To extend the capacity of AFM, other efforts have been made to combine AFM with NMR and create 3D microscopy of molecules (MRFM)[11] with atomic resolution. The goal of this technology is to detect the magnetic force of atoms using atomic force probes with single nuclear spin sensitivity. The current obstacle of MRFM is that the resolution of the probe is only several nanometers, which is far from the atomic level. Thirdly, Buckingham et al. proposed a liquid state chiral NMR to study absolute stereochemistry of small molecules.[12-17] Because spin-spin coupling effects and shielding effects are identical after space inversion of a chiral center (Parity Even), conventional liquid state NMR is blind to the absolute configuration of small chiral molecules. Specifically, neither the applied static magnetic field nor the transverse magnetic pulse works in this regard. Researchers have looked for Parity Odd properties, like spin and shielding, which can distinguish two enantiomers. Because the nuclear half-integer spins are Parity Odd fermions, the rotation of light polarization by nuclear

spins is measurable. In this regard, Nuclear Spin Optical Rotation (NSOR) and Laser enhanced NMR were developed for chiral/non-chiral NMR detection.[18, 19] As perturbation theory supports the nuclear magnetic shielding polarizability is Parity Odd, two electric-field-perturbed enantiomers are not NMR identical. Thus, Buckingham et al.[17] pointed out that the timing of the electric field could be prior to the transverse magnetic pulse. The obstacle has been that the electric-field-perturbed chemical shift tensor value is beyond the current NMR detection limit. Experiments have thus been proposed to increase the nuclear magnetic shielding polarizability constant to current NMR detection limits. Professor A. David Buckingham in University of Cambridge in Britain mentioned to me in his email back in 2014 that "In the near future we shall be starting an experimental search for the predicted effects - the work will be carried out in the CNRS High Magnetic Field Laboratory in Grenoble, France in collaboration with Drs. Peer Fischer and Geert Rikken." (A. D. Buckingham, personal communication)

Furthermore, my colleague, Clinton Edwards, a PhD student at Scripps Institution of Oceanography, is currently developing a deep learning based platform for ocean imaging, using neural networks to analyze marine ecological data. His project involves 3D digital mapping of the seafloor habitats and accurately recognizing individual marine organism, e.g. coral reefs (https://tinyurl.com/yd3kkdoz). I believe his technology can be applied to purposeful field collections of target marine organism. Imagine the diver can pull out a full 3D digital scan of the sea-floor from his/her wearable devices before snorkeling or scuba diving. By analyzing the ecological conditions of the collection site, the devices can label or tag potentially bioactive marine species for the diver and save his/her collection time.

In addition, chemical biology or bionanotechnology are having inputs that could change natural products research. Coelho et al.[20] have tamed a natural monooxygenation catalytic enzyme of the cytochrome P450 family via directed evolution to perform the task of olefin cyclization. Another example is that CRISPR-CAS9 has been applied to increase biofuel productions of marine algae by editing genes that are responsible for lipid production[21]. Thus, the continued development of new methods has a positive impact on the exploration of Nature to yield products of utility to humans.

5.3 Chapter 5 References

1. Zhang, C.; Naman, C. B.; Engene, N.; Gerwick, W. H., Laucysteinamide A, a Hybrid PKS/NRPS Metabolite from a Saipan Cyanobacterium, cf. Caldora penicillata. *Marine Drugs* **2017,** 15, (4).

2. Nogle, L. M.; Gerwick, W. H., Somocystinamide A, a novel cytotoxic disulfide dimer from a Fijian marine cyanobacterial mixed assemblage. *Organic Letters* **2002,** 4, (7), 1095-1098.

3. Wrasidlo, W.; Mielgo, A.; Torres, V. A.; Barbero, S.; Stoletov, K.; Suyama, T. L.; Klemke, R. L.; Gerwick, W. H.; Carson, D. A.; Stupack, D. G., The marine lipopeptide somocystinamide A triggers apoptosis via caspase 8. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (7), 2313-2318.

4. Ji, R. R.; Xu, Z. Z.; Gao, Y. J., Emerging targets in neuroinflammation-driven chronic pain. *Nature Reviews Drug Discovery* **2014,** 13, (7), 533-548.

5. Miller, B.; Friedman, A. J.; Choi, H.; Hogan, J.; McCammon, J. A.; Hook, V.; Gerwick, W. H., The Marine Cyanobacterial Metabolite Gallinamide A Is a Potent and Selective Inhibitor of Human Cathepsin L. *Journal of Natural Products* **2014,** 77, (1), 92-99.

6. Ranzato, M.; Taylor, P. E.; House, J. M.; Flagan, R. C.; LeCun, Y.; Perona, P., Automatic recognition of biological particles in microscopic images. *Pattern Recognition Letters* **2007,** 28, (1), 31-39.

7. Ning, F.; Delhomme, D.; LeCun, Y.; Piano, F.; Bottou, L.; Barbano, P. E., Toward automatic phenotyping of developing embryos from videos. *Ieee Transactions on Image Processing* **2005,** 14, (9), 1360-1371.

8. Hutt, A. J., The Development of Single-Isomer Molecules: Why and How. *Cns Spectrums* **2002,** 7, (4), 14-22.

9. Inokuma, Y.; Yoshioka, S.; Ariyoshi, J.; Arai, T.; Hitora, Y.; Takada, K.; Matsunaga, S.; Rissanen, K.; Fujita, M., X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature* **2013,** 495, (7442), 461-+.

10. Gross, L.; Mohn, F.; Moll, N.; Liljeroth, P.; Meyer, G., The Chemical Structure of a Molecule Resolved by Atomic Force Microscopy. *Science* **2009,** 325, (5944), 1110-1114.

11. Poggio, M.; Degen, C. L., Force-detected nuclear magnetic resonance: recent advances and future challenges. *Nanotechnology* **2010,** 21, (34).

12. Buckingham, A. D., Communication: Permanent dipoles contribute to electric polarization in chiral NMR spectra. *Journal of Chemical Physics* **2014,** 140, (1).

13. Soncini, A.; Calvello, S., Room Temperature Chiral Discrimination in Paramagnetic NMR Spectroscopy. *Physical Review Letters* **2016,** 116, (16).

14. Walls, J. D.; Harris, R. A.; Jameson, C. J., Measuring chirality in NMR in the presence of a static electric field. *Journal of Chemical Physics* **2008,** 128, (15).

15. Walls, J. D.; Harris, R. A., Measuring chirality in NMR in the presence of a time-dependent electric field. *Journal of Chemical Physics* **2014,** 140, (23).

16. Garbacz, P.; Fischer, P.; Kramer, S., A loop-gap resonator for chirality-sensitive nuclear magneto-electric resonance (NMER). *Journal of Chemical Physics* **2016,** 145, (10).

17. Buckingham, A. D.; Fischer, P., Direct chiral discrimination in NMR spectroscopy. *Chemical Physics* **2006,** 324, (1), 111-116.

18. Shi, J.; Ikalainen, S.; Vaara, J.; Romalis, M. V., Observation of Optical Chemical Shift by Precision Nuclear Spin Optical Rotation Measurements and Calculations. *Journal of Physical Chemistry Letters* **2013,** 4, (3), 437-441.

19. Warren, W. S.; Mayr, S.; Goswami, D.; West, A. P., Laser-Enhanced Nmr-Spectroscopy. *Science* **1992,** 255, (5052), 1683-1685.

20. Coelho, P. S.; Brustad, E. M.; Kannan, A.; Arnold, F. H., Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* **2013,** 339, (6117), 307-310.

21. Ajjawi, I.; Verruto, J.; Aqui, M.; Soriaga, L. B.; Coppersmith, J.; Kwok, K.; Peach, L.; Orchard, E.; Kalb, R.; Xu, W. D.; Carlson, T. J.; Francis, K.; Konigsfeld, K.; Bartalis, J.; Schultz, A.; Lambert, W.; Schwartz, A. S.; Brown, R.; Moellering, E. R., Lipid production in Nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. *Nature Biotechnology* **2017,** 35, (7), 647-+.