

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Cheminformatic Approaches to Decipher Natural Product –Target – Disease Associations

Permalink

<https://escholarship.org/uc/item/37x7j671>

Author

Delgadillo, David Alexander

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

SANTA CRUZ

**CHEMINFORMATIC APPROACHES TO DECIPHER NATURAL PRODUCT –
TARGET – DISEASE ASSOCIATIONS**

A dissertation submitted in partial satisfaction

of the requirement for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMISTRY AND BIOCHEMISTRY

by

David Alexander Delgadillo

March 2022

The Dissertation of David Alexander
Delgadillo is approved:

Professor John B. MacMillan, Advisor

Professor R. Scott Lokey, Chair

Professor Laura M. Sanchez

Peter Biehl
Vice Provost and Dean of Graduate Studies

Copyright © by
David A. Delgadillo
2022

TABLE OF CONTENTS

List of Figures	iv
List of Tables	vi
List of Schemes	vii
List of Definitions	viii
Abstract	x
Acknowledgements	xii
Dedication	xvi
1 SCREENING STRATEGIES FOR NATURAL PRODUCTS – TARGET – DISEASE ASSOCIATION DISCOVERY	1
1.1 Brief History of Natural Products	1
1.2 Screening Strategies for Natural Products	2
Drug Discovery: Target-Based vs Phenotypic Screens	
1.3 Natural Products and Cancer	5
1.4 Natural Product Fraction Library Generation (crude vs fractionation)	7
1.5 Non-Small Cell Lung Cancer Natural Products Screen	14
1.6 Conclusion	17
1.7 Materials & Methods	19
2 Isolation and Structural and Biological Characterization of Ikarugamycin	24
2.1 Analysis of NSCLC Screening Data	25
2.2 Isolation of the Bioactive Metabolites from SNB-040 and SNE-002 for NSCLC Toxicity	28
2.3 Biological Activity of Ikarugamycin	42
2.4 Semi-Synthetic Derivatization	46
2.9 Conclusion and Discussion	50
2.10 Materials and Methods	52
3 Directed-Message Passing Neural Network Application to the Natural Products Atlas	57

3. 1 Challenges in Physical High Throughput Screening Strategies	58
3.2 Virtual Natural Products Libraries	61
3.3 <i>In silico</i> Screening of Natural Products	63
3.4 Proof of Principle Application to the Natural Products Atlas	68
3.5 Conclusion	73
3.6 Materials and Methods	75
4 Application of Directed Message Passing Neural Network to Rapidly Identify Anti-Viral Natural Products	81
4.1 Natural Products as a trove of bioactive compounds	82
4.2 Onset of the SARS-CoV-2 Virus	83
4.3 Target Selection for SARS-CoV-2	84
4.4 <i>In-silico</i> Approaches to SARS-CoV-2 Drug Discovery	87
4.5 Application of DMPNN to Identify NPs Active Against SARS-CoV-2 Proteases	88
4.6 Conclusion	90
4.7 Materials and Methods	94
Epilogue	107
APPENDICES	109
Bibliography	182

LIST OF FIGURES

Figure 1.1 Generic representation of target-based screening strategies and phenotypic screening strategies.

Figure 1.2 Crystal structure of PT2399 bound to HIF-2 α and the structural similarity to the FDA-approved drug Welireg.

Figure 1.3 Schematic overview of cheminformatic approaches to natural product drug discovery.

Figure 1.4 NSCLC by histology and adenocarcinoma mutations.

Figure 1.5 Selected structures that are representative of first line treatments for NSCLC.

Figure 1.6 The figure depicts the main druggable genetic targets and their involvement in the main signaling pathways in NSCLC.

Figure 1.7 Representation of NSCLC cell line clustering via different characterizations.

Figure 1.8 General bioactive natural product discovery workflow in the MacMillan Lab.

Figure 1.9 Genomic Characterization and Chemical Sensitivities of NSCLC Cell Line Panel.

Figure 2.1 Natural product fraction prioritization workflow.

Figure 2.2 Phenotypic readout of preliminary NSCLC screen.

Figure 2.3 Representation of different cytotoxicity profiles observed in preliminary NSCLC screen.

Figure 2.4 Purification schematic for the isolation of bioactive metabolites identified in SNB-040.

Figure 2.5 Structure of ikarugamycin, a microbial natural product.

Figure 2.6 Structure of ikarugamycin analog, capsimycin D.

Figure 2.7 Structure of ikarugamycin analog, capsimycin B.

Figure 2.8 Structure of ikarugamycin analog, capsimycin F.

Figure 2.9 Key COSY and HMBC correlations of **4**.

Figure 2.10 Key NOESY correlations of **4**.

Figure 2.11 Structure of ikarugamycin analog, capsimycin C.

Figure 2.12 Structure of ikarugamycin analog, xlamenemycin C.

Figure 2.13 Structure of ikarugamycin analog, SS8201 D.

Figure 2.14 Structure of Ikarugamycin analog, capsimycin E.

Figure 2.15 A. Volcano plot of ikarugamycin (**2.5**) against 40 NSCLC cell lines and 2 HBEC cell lines. **B.** Overlaid cell viability plots of ikarugamycin against a representative set of NSCLC cell lines.

Figure 2.16 IC₅₀ curve of representative NSCLC cell lines against ikarugamycin (**1**).

Figure 2.17 Collection of PTMs screened against NSCLC.

Figure 2.18 Cytotoxicity assay of ikarugamycin (**1**) and propargyl-IKA against representative NSCLC cell lines.

Figure 3.1 A Sequential steps applied in virtual screening workflows to identify bioactive natural products. **B** Ligand- and structure-based virtual screening approaches and some of their associated computational methods

Figure 3.2 Representation of a message passing neural network (MPNN) that iteratively aggregates local chemical features for molecular property prediction.

Figure 3.3 Representative classes of antibiotics of the modern era, excluding the arsenic-containing antibiotics of the early twentieth century.

Figure 3.4 Schematic representation of machine learning in antibiotic discovery that demonstrates how the combination of *in silico* predictions and empirical investigations can lead to the discovery of new antibiotics.

Figure 3.5 Cross validation of predicted natural products with antimicrobial characteristics.

Figure 4.1 Antivirals drugs by source from 2015 to 09/2019, $n = 185$.

Figure 4.2 Schematic of SARS-CoV-2 replication mechanism within host.

Figure 4.3 IC₅₀ curves generated via fluorescent protein engagement assay.

Figure 4.4 t-Distributed stochastic neighbor embedding (t-SNE) of all molecules from the training datasets and the NP Atlas, with Closthioamide.

LIST OF TABLES

Table 1.1 Examples of natural products isolated from phenotypic screens.

Table 2.1 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **1**.

Table 2.2 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **2**.

Table 2.3 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **3**.

Table 2.4 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **4**.

Table 2.5 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **5**.

Table 2.6 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **6**.

Table 2.7 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **7**.

Table 2.8 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **8**.

Table 2.9 Anticancer activity of PTMs against representative NSCLC cells.

Table 3.1 Commercially and publicly available large natural product libraries.

Table 3.3 List of hyperparameters for each respective dataset.

Table 4.1 Examples of active molecules derived from drug repurposing for SARS-CoV-2.

Table 4.2 List of publicly available drug repurposing campaigns for compounds against Betacoronavirus genus.

Table 4.3 List of hyperparameters for each respective dataset

LIST OF SCHEMES

Scheme 2.1 Reaction conditions for the selective reduction of α,β -unsaturated carbonyls in **1**.

Scheme 2.2 Altered reaction conditions of **Scheme 2.1**.

Scheme 2.3 Reaction conditions for the Birch reduction employed on **1**.

Scheme 2.4 Epoxidation of **1** utilizing *m*CPBA.

Scheme 2.5 General reaction conditions for the Juliá-Colonna epoxidation of **1**.

Scheme 2.6 Reaction scheme of the epoxide ring opening of compound **3**.

Scheme 4.1 NPs identified with high bioactivity against key SARS-CoV-2 proteins via in silico screening.

Scheme 4.2 Several NPs with a high propensity to be active against SAR-CoV-2 Mpro via DMPNN.

Scheme 4.3 Schematic presentation of the total synthesis of Closthioamide.

LIST OF DEFINITIONS

AC50	½ maximal effect measured
AKT	protein kinase B
ALK	anaplastic lymphoma receptor tyrosine kinase
ATP	adenosine triphosphate
BCL11A	B-cell lymphoma/leukemia 11A
COSY	correlation spectroscopy
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EC50	½ maximal effect
EGFR	epidermal growth factor receptor
EML4	echinoderm microtubule-associated protein-like 4
FDA	US Food and Drug Administration
GI50	½ growth inhibition
HMBC	heteronuclear multiple bond correlation
HNRNPM	heterogeneous nuclear ribonucleoprotein M
HPLC	high performance liquid chromatography
HSQC	heteronuclear single-quantum correlation
IC50	½ maximal inhibition
LC-MS	liquid chromatography-mass spectrometry
LDL	low-density lipoprotein
LD50	½ maximal lethal dose
miRNA	micro ribonucleic acid
mRNA	messenger ribonucleic acid
mTOR	mammalian target of rapamycin
NFκB	nuclear factor κB
NMR	nuclear magnetic resonance spectroscopy

NOESY nuclear Overhauser effect spectroscopy
PI3K phosphoinositide 3-kinase
RNA ribonucleic acid
siRNA small interfering ribonucleic acid
TOCSY total correlation spectroscopy
DMPNN directed-message passing neural network

ABSTRACT

CHEMINFORMATIC APPROACHES TO DECIPHER NATURAL PRODUCT – TARGET – DISEASE ASSOCIATIONS

By David Alexander Delgadillo

Cancer is the second leading cause of death in the United States, with many of those deaths attributed to lung cancer. NSCLC accounts for nearly 85% of all lung cancer cases, making NSCLC a leading cause of cancer-related death in the United States. A chemistry-driven de novo discovery strategy utilizing cheminformatics recently identified ikarugamycin (IKA) as a potent and selective inhibitor of cellular proliferation amongst NSCLC cell lines. However, a detailed characterization of IKA and several analogs has yet to be performed within the context of NSCLC. This work aimed to further investigate the relationship between IKA analogs and the effect that structural diversity may have on the potency and selectivity of its' antiproliferative properties concerning NSCLC. The chemical characterization and biological cytotoxicity profiling of IKA and its' analogs against several NSCLC cell lines will drive the development of IKA towards clinical relevance. Biological evaluation of several IKA analogs revealed that the double bond within the 5-6-5 ring moiety is crucial to selective antiproliferative activity against NSCLC HCC44, H23, and Calu-1. All selective analogs exhibited an IC₅₀ value within the 0.09-1.00 μ M range. Along with the strong toxicity trends we have already observed, IKA also appears to be interacting with a novel biological target outside of the commonly acted on genes (i.e., EGFR, ALK, BRAF, ROS1). This work also explored strategies to embed synthetic handles for future click pull-down experiments for target identification.

Concurrently, this thesis contains work on the utilization of the NP Atlas as a compound repository for virtual drug screening – a first of its kind. To address the emerging viral epidemic of SARS-CoV-2, I employed the open-source Chemprop algorithm to train a Directed-Message Passing Neural Network (DMPNN) to identify key chemical descriptors that can be attributed to disrupting the viral replication mechanism. I utilized open-source screening data provided by the National Center for Biotechnology Information (NCBI) to train the DMPNN and

subsequently utilized the trained neural network to score compounds found within the NP Atlas. Through this process, I was able to identify and validate the targeted interaction between closthioamide and the main protease of SARS-CoV-2. This work serves as a proof of principle for adjacent computational drug discovery strategies that may help scientist prioritize their efforts and lower the cost of resources necessary to screen libraries that are greater than 24,000 molecules.

ACKNOWLEDGEMENTS

Those whom I have the pleasure to call my mentors over the years have always imparted wonderful words of wisdom unto me. “The graduate school experience is less about what you learn through your experiments and more of what you learn about yourself.” “Grit is the ultimate tool for success – your experiments will fail, your journey will be tough, but developing grit will allow you to overcome those trialing moments.” “Graduate school is a marathon and not a race.” While these wise words express the exhaustive nature of the effort that necessitates grit and pacing oneself, they also suggest growth and perseverance upon completion of the great ‘race’. Through and through, my graduate experience has been immensely rewarding, and one of the greatest aspects of that is finally getting to reflect on the last four and a half years through writing this volume. Gratefully, the journey that encompassed my pursuit for the doctorate was not traveled alone and I have many to thank for encouraging, supporting, and inspiring me to attain by doctorate.

First, I would like to thank John MacMillan for being my advisor. Since the very first meeting I had with John, he encouraged me to pursue my curiosity by engaging in one of my many hypothetical rants about machine learning and its’ potential role in the field of natural products. Albeit a conversation had in the spur of the moment, that conversation solidified my desire to work under an advisor that was as supportive and engaging as John. Over the course of my degree, I have learned to be grateful for our early-morning discussions and debates which allowed me space to grow both as a scientist and a person. Thank you for always entertaining my ideas and encouraging me to find a way to make them work, rather than pointing out their flaws (which were plenty). I sincerely appreciate the guidance and support you provided me throughout my journey.

Secondly, I would like to thank the members of my committee. Professor Scott Lokey, thank you for always being flexible with your schedule and for challenging me to dig deep into the chemical rationale of my proposals. Professor Laura Sanchez, thank you for being so supportive throughout this process and for constantly sharing opportunities with the students

at UC Santa Cruz. Dr. Joshua Schwochert, thank you for the invaluable insight regarding the process of taking an idea and seeing it through at all costs. And lastly, I would like to thank Professor Nikolas Sgourakis for the incredible support as a scientist. Dr. Sgourakis provided me with a logical, yet critical, sound board for my ideas and always assured me that I had place in science whenever I would spiral into a pit of self-doubt.

To my previous mentor, Dr. Ryan Baxter, thank you. You gave me a chance. You allowed me to join your honors organic chemistry course even though I wasn't a chemistry major, you gave me a chance to learn. You asked me if I wanted to conduct research in your lab over the summer, you gave me a chance to conduct scientific experiments. You encouraged me to apply to the CAMP and UC LEADS programs, you wrote countless letters of recommendation over the years – you gave me chance. Thank you for introducing me to the beautiful world of organic chemistry and for giving me a chance to become *someone*.

To my life coach, Dr. Peter Mai, thank you. You taught me how to run my first column purification, but most importantly you taught me how to fight for what I want out of life. Thank you for always checking in on me and for snapping me back into reality when I would start talking about life at a biotech company.

Over the years, I have had the pleasure to work alongside some incredible staff members whose expertise and willingness to help ensured that the completion of my doctorate was as free from pain as possible. I would like to take this time to thank Karen Meece for her incredible patience and relentless support throughout my time at UCSC, as well as for all that she does for all the graduate students she oversees. I would also like to thank Jack Lee for showing me how to master the instruments in the NMR facility and for all our insightful conversations, you were an absolute pleasure to work with. Yuliana Ortega and Xingci Situ both played integral roles in providing me with reassurance and validation as I navigated my journey through higher education – thank you for your endless support for students and our communities. Lastly, I would like to thank DeSean Greene for always being so genuine and for making the late nights in the lab a lot less lonely.

To my amazing cohort: graduate school would have not been nearly as enjoyable as it was without each and every one of you. I would like to thank the following students for their unconditional love and bottomless kindness: Jonathan Philpott, Vivien Cherrete, Jeremy Barnett, Kaitlyn Vian, Anna Johnston, and Megan Freyman. I am extremely grateful to have gotten to know you all so well and it has been an absolute honor to have traveled this journey alongside you.

It is only fair that I apologize to my lab mates for having to tolerate me before I go through and thank each of you – sorry for having my experiments in every nook of the lab and for not cleaning it up before I left. Aswad, thank you for being such a great friend and an amazing role model. You have taught me what it means to come into the lab each day with a plan in mind and how to think critically about science. Sahar, thank you for always being so kind and for running the cytotoxicity assays that I never dared to do myself. Rahul, thank you for always being a sound board for my synthetic troubles and for having such a contagious smile. Scott, you escaped to Texas before I could finish this degree, but you left me with culinary experiences unlike any other; never forget – have the courage to follow your convictions. Riley, you are an extremely inquisitive person and I have no doubt that you will have a successful scientific career. Thank you for early morning coffees and for allowing me to mentor you during your first year, I hope you learned the good scientific habits and forgot the bad ones. Lastly, I would like to thank my rocks in the lab: Anam, Duy, and Victor. Anam, thank you for always helping me work through the tough moments in graduate school and for being such a great friend – you remain one of the most knowledgeable people I know, keep pushing forward! Duy, thank you for always feeding me and for always talking science with me. The opportunity I have at Caltech would not have been a reality had you not rushed to show me the post by Derek Lowe. You have always been like a brother to me, thank you for always having my back and for rescuing me whenever I was in need. Victor, from the moment we met I knew we were going to be friends – for better or for worse. Thank you for allowing me to feel seen, I'm not sure where I would be without our late-night conversations regarding the sacrifices we had to make

to pursue our graduate degrees. The three of you have helped me grow not only as a scientist, but as human being – thank you for making the journey so enjoyable.

Thank you to my wonderful partner, Sierra, for helping me get through the most tumultuous times in graduate school – lightning complexes, wildfires, deadly diseases. Through it all, you have continued to cheer me on and reassure me that things will work out as they were meant to. Thank you for pushing me to pursue my dreams and for inspiring me to be a better version of myself every day. You and your family

Last, but not least, I would like to thank my family. They are the foundation upon which this achievement stands. Each one of you has made a sacrifice so that I could have the opportunity to pursue this degree, and for that I am eternally grateful. Mom, thank you for taking us to the library every Tuesday and for always giving into my tantrums when I wanted to go to the Natural History Museum. Dad, thank you for always teaching me your handy ways for helping me build my weird science projects for the science olympiads. Together, you fostered an insatiable curiosity and provided unconditional support at every level of my education. I cannot put into words how grateful I am to have you two in my life and how proud I am to be your son.

This thesis is dedicated to my mom and dad, Martha Leticia Valencia-Delgadillo and Jose Alfredo Delgadillo, and to both of the Delgadillo and Valencia families. Me puse las pilas y si se pudo.

CHAPTER ONE

**SCREENING STRATEGIES FOR NATURAL PRODUCTS – TARGET – DISEASE
ASSOCIATION DISCOVERY**

1.1 Brief History of Natural Products

The relationship between natural products and human health has long existed and was born of survival. Humans have turned to nature to find the earliest forms of medicine and have continued to explore the expansive chemical space that nature has provided. Natural products, in the context of this dissertation, are described as bioactive chemical compounds sourced from microorganisms, plants, and animals. These compounds are secondary metabolites that are not directly involved in development or reproduction but are understood to have co-evolved with their source organism to help navigate biotic and abiotic stresses or changes.¹ Through this evolution, nature has managed to construct diverse sets of complex chemical entities with potent and selective bioactivity. It was only within the last 250 years that we began to understand how these natural products exerted their medicinal effects.¹ Technological advancements in analytical techniques, such as chromatography and spectroscopy allow us to narrow down the medicinal effects of extracts to a singular chemical entity. Developments of atomic theory have allowed scientists to better understand how the structure of these chemical entities relates to their biological activity. Natural products continue to be a rich source of bioactive molecules that have tremendous importance to human health. Many of the therapeutics that form the cornerstones of modern medicine have been directly influenced by the study and characterization of natural products.

The importance of natural products is best captured by the Nobel Prize in Physiology or Medicine awarded to Dr. Selman Waksman in 1952 for the discovery of streptomycin and its' bioactivity against *Mycobacterium tuberculosis*.² This discovery illuminated the powerful therapeutic potential of compounds isolated from natural sources and solidified their importance in combating human diseases. The strong hold of natural products research was fortified by the 2015 Nobel Prize in Physiology or Medicine awarded to Drs. Youyou Tu, William Campbell, and Satoshi Ōmura for the discovery of the anti-malarial artemisinin and the anti-parasitic avermectin.^{3,4} Although great strides have been made in the fight against human

disease, the emergence of new diseases and increasing antibiotic drug resistance serve as a reminder that natural products research must continue and novel screening strategies must emerge.

1.2 Screening Strategies for Natural Product Drug Discovery: Target-Based Screens vs Phenotypic Screens

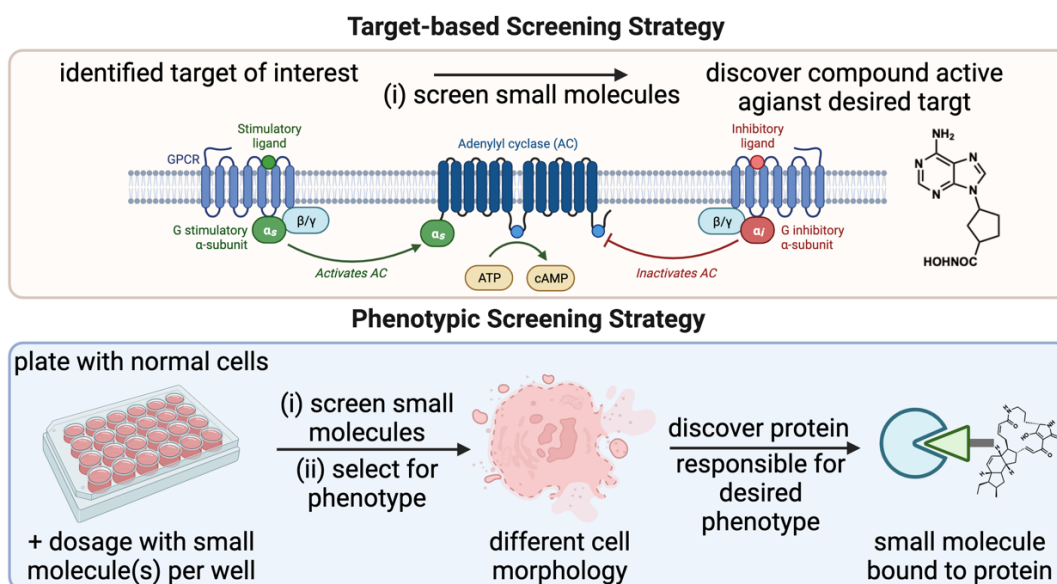


Figure 1.1 Generic representation of target-based screening strategies and phenotypic screening strategies.

Decades of screening methods including natural product samples have been employed in industry and academia alike. The usefulness of crude natural product extracts in primary screening has decreased as assay techniques became more complex, more target-based, and more high-throughput. This has led to the increased use of prefractionated natural product libraries in high throughput screening efforts.^{5,6} High throughput screening is described as the use of automated equipment to rapidly test large compound libraries for biological activity at the model organism, cellular, pathway, or molecular level. High throughput screening can be divided into two main paradigms: target-based screening and phenotypic screening (**Figure 1.1**). Target-based screening takes significant work at the front end to validate the physiological

benefit in the disease context. However, a major advantage is that knowledge of the molecular target and/or mechanisms allows drug discovery tools, such as mutational analysis, crystallography, computational modeling, and a variety of other techniques to deduce how a drug interacts with the target. This enables the efficient creation of structure-activity relationships, biomarkers, and subsequent generations of the medication acting on the target.

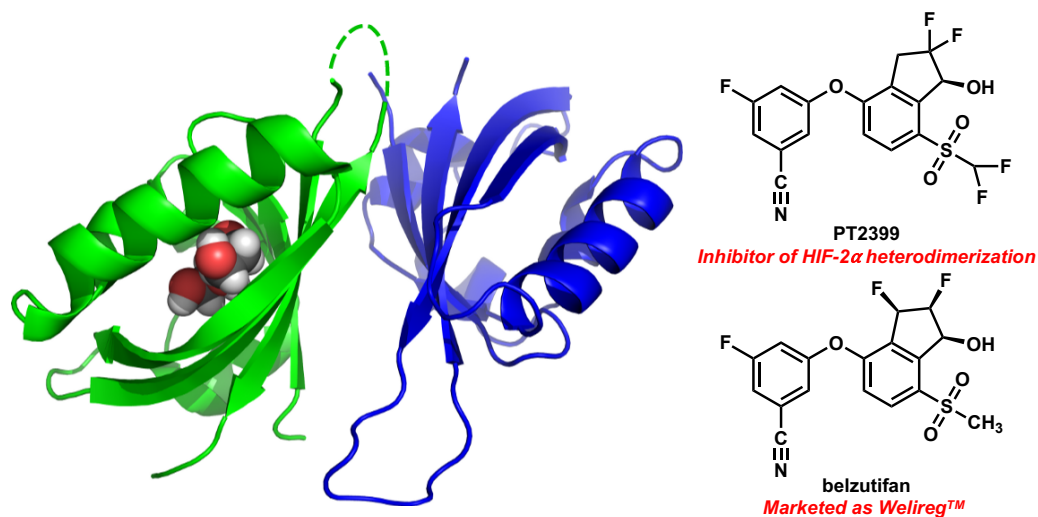


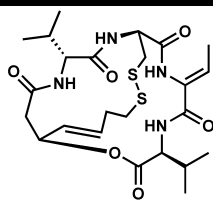
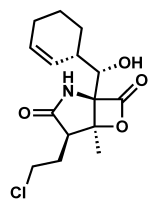
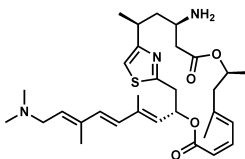
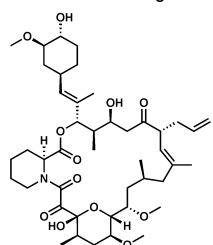
Figure 1.2 Crystal structure of PT2399 bound to HIF-2 α and the structural similarity to the FDA-approved drug Welireg.

A notable example of successful target-based HTS is the discovery of the HIF-2 antagonist PT2399 (**Figure 1.2**). Hypoxia inducible factors (HIF-1, HIF-2, and HIF-3) accumulate in the cell nucleus under low oxygen conditions (hypoxia), such as those found in solid tumors. This accumulation upregulates the transcription of several genes that allow for the cell to adapt to the low oxygen conditions, thereby further supporting the growth and metastasis solid tumors. NMR-based ligand binding assays of drug fragments and HTS screens looking for disruption of protein/protein interactions paved the way for scientists to identify compounds that could disrupt the endogenous HIF-2 α -HIF-1 β heterodimer. The identified HIF-2 α ligand was then subject to medical chemistry optimization to yield PT2399,

which has been further optimized and is now being marketed as Welireg (belzutifan) for adult patients with von Hippel-Lindau disease.⁷⁻⁹

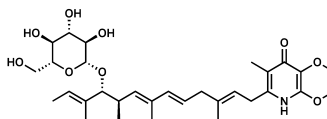
Although target-based screening has allowed for many advances in modern drug discovery, high attrition rates in Phase II and III clinical trials attributed to poor target selection, poor drug efficacy, and narrow scope have limited the approach. Unbiased phenotypic screening allows for the bioactivities of natural products to be evaluated at the cellular, tissue, or whole organism level. In addition, phenotypic screening has the capability of identifying various compounds with diverse structural characteristics that interact with potentially unknown biological targets or that exhibit novel mechanisms of action. On account of this, several phenotypic events associated with a given disease have been exploited in the phenotypic screening of natural products and are summarized in **Table 1.1**.

Table 1.1 Examples of natural products isolated from phenotypic screens.

Phenotypic Screen	Isolated Natural Products	Representative Structure	Biological Activity
	Romidepsin		HDAC Inhibition ^{10,11}
Cytotoxicity	Salinosporamide A		Proteasome Inhibition ¹²
	Pateamine		Translation Inhibition Via eIF4A ¹³
Immuno-suppression	FK506		Immuno-suppressive activity by targeting FABP/CaN ^{14,15}

Filopodia
protrusion
inhibition

Glucopiericidin A



Cancer metastasis
inhibition by
functional targeting
of GLUT-1¹⁶

1.3 Natural Products and Cancer

Since the passing of the National Cancer Act of 1971, cancer remains a leading cause of morbidity and mortality on a global scale. To date, natural products or their derivatives account for more than 57% of small molecule cancer therapeutics¹⁷. Paclitaxel (Taxol), for example, was developed from the bark of the Pacific yew tree and is one of the most regularly used chemotherapy medications¹⁸. Although many cancer drug discovery efforts have moved towards synthetic compound libraries or bioconjugate strategies, the screening of natural products against cancer remains crucial to discovering bioactive pharmacophores and identifying novel oncogenic drug targets.

Given the complexity of cancer genomics, natural product drug discovery efforts have often elected to conduct phenotypic screens in order to identify potential drug leads. Since the identification of taxol, several cell-based phenotypic assays have emerged in an effort to isolate novel anticancer therapeutics. The ability for scientists to develop novel anti-cancer screens would not be as accessible as it is today without the US National Cancer Institute (NCI). The NCI compiled a panel of 60 cell lines each possessing a diverse histology and representing seven types of human cancers (brain, colon, leukemia, lung, melanoma, ovarian, and renal).³³⁷ The establishment and characterization of this panel has enabled the developments of assays measuring cell proliferation and/or death that can be interrogated through various readouts, such as ³H-thymidine uptake, metabolic activity indicators, and trypan blue exclusion.³³⁸ For example, the natural product glucopiericidin A (GPA) was isolated while screening microbial samples in a filopodia protrusion inhibition assay. Filopodia are cell membrane projections that contribute to tumor metastasis. The phenotypic assay was designed to identify novel inhibitors of the protrusions and to characterize the molecular mechanism that implicates filopodia

protrusion in tumor metastasis. Kitagawa et al. utilized human epidermal A431 cells that are over express epidermal growth factor (EGF) receptors and exhibit filopodial protrusion within 30 minutes.¹⁶ The bioassay guided isolation led the group to identify GPA and piericidin A (PA) as strong inhibitors of filopodial protrusion that acted synergistically.

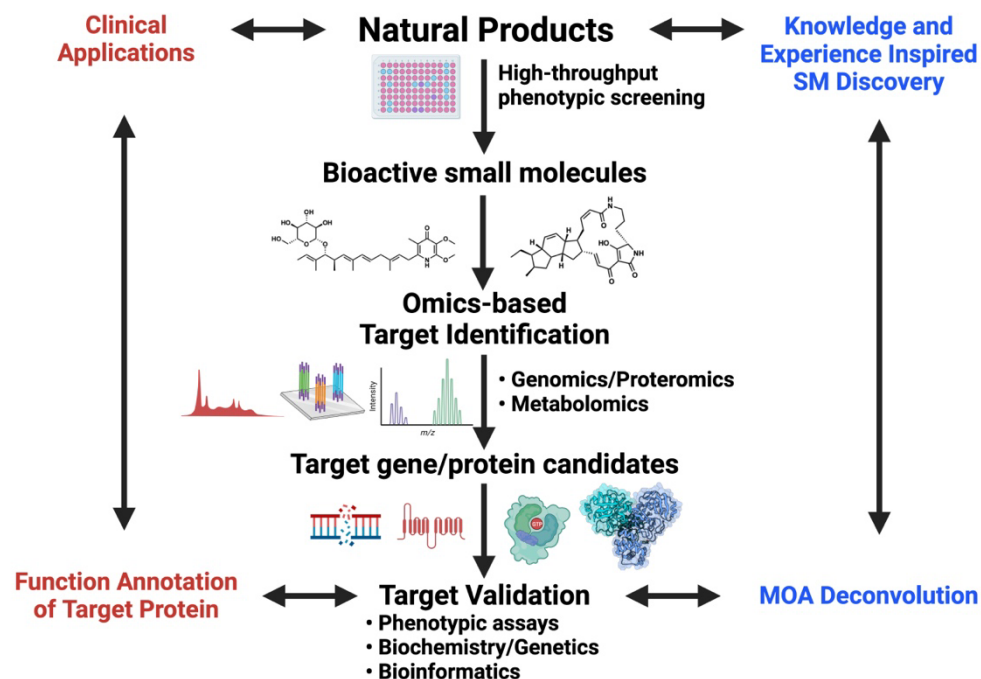


Figure 1.3 Schematic overview of cheminformatic approaches to natural product drug discovery.

Although PA was previously identified as a mitochondrial complex I inhibitor, GPA exhibited considerably weaker inhibition against mitochondrial respiration. This led the team to utilize chemical genomic screening to elucidate the mode of action of GPA and metabolomics (CE-TOFMS) as a method of target identification – a first of its kind. Their analysis identified glucose transporter 1 (GLUT1) as the functional target of GPA. The story of GPA highlights the utility of unbiased phenotypic screens in cancer therapeutic research and showcases how omics-based methods, including chemical genomics³³⁹, proteomics³⁴⁰, and metabolomics³⁴¹, can be used to uncover new therapeutic targets (**Figure 1.3**). The unbiased phenotypic screening of natural products can provide a number of unique bioactive small molecules. Target

identification of these bioactive natural products with omics-based methods allows us to annotate potentially new therapeutic targets and protein candidates. This enables the deconvolution of the mode of action of the natural product and functional annotation of the target proteins in specific biological systems. Based on the newly identified structural and biological characterization of bioactive molecules, new synthetic small molecules can be discovered. Collectively, this approach raises the importance of unbiased phenotypic screening of natural products for therapeutic applications.

1.4 Non-Small Cell Lung Cancer Overview

Lung cancer accounts for the highest mortality rate amongst both men and women in the USA. Non-small cell lung cancer (NSCLC) is group of genetic diseases with three major subgroups¹⁹ (**Figure 1.4**). Based on advances in genomic and mutational analysis, it is now recognized that up to 60% of adenocarcinomas and 50% to 80% of squamous cell carcinomas (SCC) have a known oncogenic driver mutation^{20,21} (**Figure 1.4**). These driver mutations eventually cause unregulated growth, proliferation, and rapidly generate resistance to therapy. The genetic and epigenetic mutations that give rise to cancer also present vulnerabilities in cancer cells that serve as potential targets for therapeutic intervention.

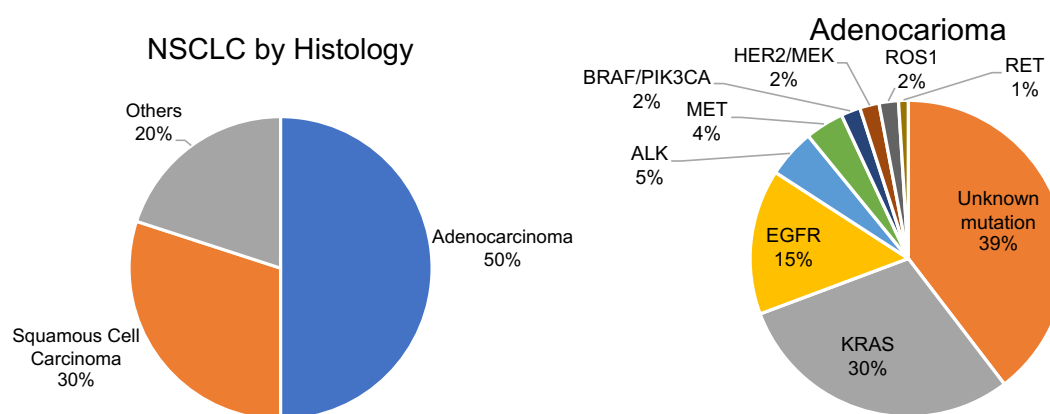


Figure 1.4 NSCLC by histology and adenocarcinoma mutations.

Historically, NSCLC has been treated with platinum-based doublets. The typical survival rate of patients that only receive chemotherapy is approximately 6%, whereas those who are eligible to receive immune-based therapies ranges from 15% to 50%. Clinical trials of second-line treatments, such as pemetrexed and docetaxel, have generated overall response rates of 9.1% and 8.8%, respectively²²⁻²⁴. There are now a number of therapeutic pathways with specific oncoprotein inhibitors that target lung cancers addicted to these oncogenic driver mutations. The main therapeutic pathways and their associated drugs (**Figure 1.5**) are described below.

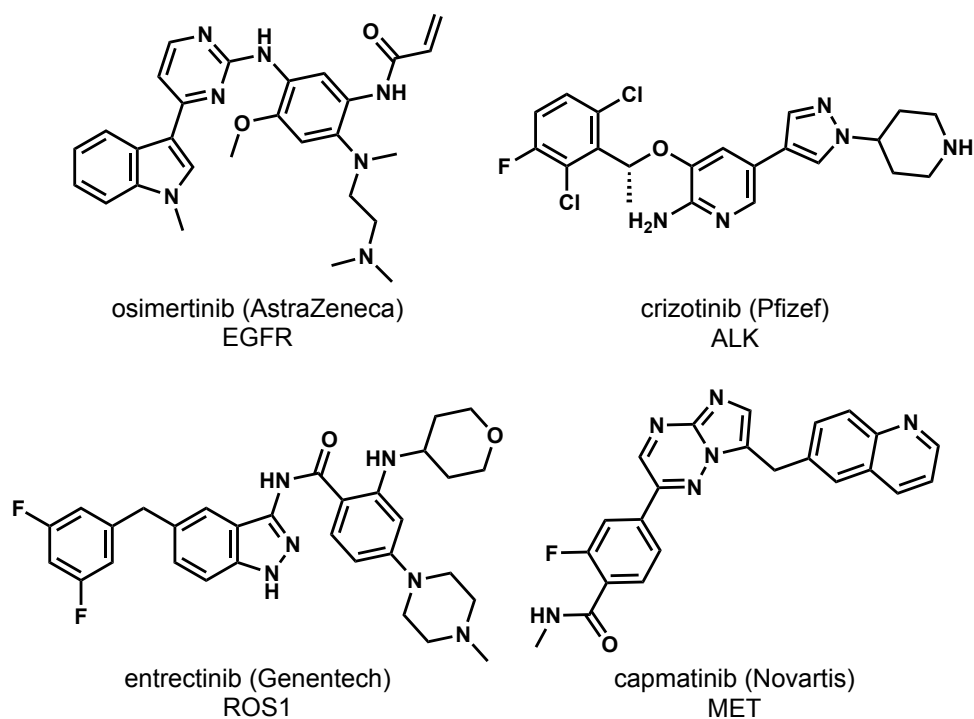


Figure 1.5 Selected structures that are representative of first-line treatments for NSCLC.

1.4.1 Sensitizing EGFR Mutations

EGFR mutations are the most prevalent among nonsmokers and former light smokers. In 45 percent of patients, deletions in exon 19 and point mutations in exon 21 (L858R) are the most prevalent EGFR gene alterations. Osemertinib is a medication that is currently in use.

(Tagrisso; AstraZeneca) (**Figure 1.5**), which is the current standard of care for first-line treatment; dacomitinib (Vizimpro; Pfizer); afatinib (Gilotrif; Boehringer Ingelheim); erlotinib (Tarceva; Genentech); and gefitinib (Iressa; AstraZeneca).²⁵

Additionally, the use of osimertinib has recently moved to the adjuvant setting for stage IB-IIIa NSCLCs that have a sensitizing EGFR mutation.²⁶ Tumors with EGFR mutations do not respond to immune checkpoint inhibitors (ICIs) except for the atezolizumab (Tecentriq; Genentech) quadruplet regimen.²⁵

1.4.2 EGFR Exon 20 Insertion Mutations

EGFR exon 20 insertion mutations account for 4% to 10% of all EGFR mutations. Patients with EGFR exon 20 mutations have previously had poor results when treated with existing EGFR tyrosine kinase inhibitors (TKIs). Amivantamab, a bispecific antibody directed against MET receptors and EGFR, was examined in patients who had progressed on or after platinum-based treatment. In the phase 1 CHRYSALIS study (NCT02609776), amivantamab elicited an ORR of 40% (3.7% were complete responses and 36.3% were partial responses). Immature data for median OS and PFS have been shown to be 22.8 months and 8.3 months, respectively.

1.4.3 ALK Rearrangements

ALK fusions are caused by a rearrangement in the ALK gene, which codes for a tyrosine kinase, and another gene product, most often EML4. The fusion product is a constitutively active kinase that promotes cellular proliferation and survival.²⁰

Tumors with ALK mutations do not respond to ICIs.²⁰ Current drugs include alectinib (Alecensa; Genentech), the standard of care for first-line treatment; brigatinib (Alunbrig; Takeda Oncology); ceritinib (Zykadia; Novartis); crizotinib (Xalkori; Pfizer) (**Figure 1.5**); and lorlatinib (Lorbrena; Pfizer).

1.4.4 ROS1 Rearrangements

In terms of molecular function, the ROS1 tyrosine kinase is quite similar to ALK. ROS1 mutations are more common in people who do not have EGFR mutations, KRAS mutations, or ALK gene fusions. The response to ICIs is shortened, with ORRs of just 17%.²⁰ Current drugs include crizotinib, a preferred first-line treatment; ceritinib; entrectinib (Rozlytrek; Genentech) (**Figure 1.5**), a preferred first-line treatment; and lorlatinib, which is reserved for second-line treatment.²⁰

1.4.5 BRAF V600E Mutations

BRAF is a serine/threonine kinase in the MAP/ERK kinase pathway. BRAF gene mutations are linked to more aggressive tumor histology and a worse prognosis. Patients who have BRAF mutations react to ICIs at a rate of 24%.²⁰ Current targeted agents for these mutations include dabrafenib (Tafinlar; Novartis) plus trametinib (Mekinist; Novartis), which is preferred; or vemurafenib (Zelboraf; Genentech) monotherapy.²⁰

1.4.6 NTRK1/2/3 Gene Fusions

TRK proteins (TRKA, TRKB, and TRKC) encoded by NTRK genes play a crucial role in the cellular development, differentiation, and death of peripheral and central nervous system neurons. NTRK fusions occur at a rate ranging from 0.2 to 4% in NSCLC..^{27,28} It is unknown whether there are ethnic-related or social behavior-related predilections for NTRK mutations.²⁹ Current therapies include larotrectinib (Vitrakvi; Bayer) and entrectinib.

1.4.7 MET exon 14 (METex14) Skipping Mutations

METex14 skipping mutations are detected in around 3% of NSCLC cases and are more common in females, adults 70 years or older, nonsmokers, and patients with pulmonary sarcomatoid cancer. METex14 skipping mutations are linked with a poor prognosis, and unlike KRAS and BRAF mutations, the response to immunotherapy is shortened to ORRs of 16% to

17%. Current guideline-recommended agents include capmatinib (Tabrecta; Novartis) (**Figure 1.5**), tepotinib (Tepmetko; EMD Serono), and crizotinib. Additionally, the investigational drug savolitinib (AZD6094; AstraZeneca) is a selective MET inhibitor that is being studied.²⁰

1.4.8 RET Rearrangements

RET rearrangements occur when the RET gene combines with another gene, resulting in a fusion RET protein that is overexpressed and promotes cellular proliferation. RET fusions are oncogenic drivers in 1% to 2% of NSCLC diagnosis.

Immunotherapy response is minimal with responses of 6%. The current agents that can be used include selpercatinib (Retevmo; Eli Lilly and Company), a preferred treatment; pralsetinib (Gavreto; Blueprint Medicines and Genentech), a preferred treatment; and cabozantinib and vandetanib (Caprelsa; Sanofi Genzyme).³⁰⁻³³

1.4.9 PD-1/PD-L1 Axis

ICIs that target PD-1/PD-L1 axis work by reversing tumor-mediated inactivation of T cells and improving immune antitumor response. Classwise, PD-1 receptor inhibitors include nivolumab (Opdivo; Bristol Myers Squibb), pembrolizumab (Keytruda; Merck), and more recently cemiplimab (Libtayo; Regeneron Pharmaceuticals and Sanofi Genzyme), whereas atezolizumab and durvalumab (Imfinzi; AstraZeneca) inhibit PD-L1.³⁴

ICIs are primarily utilized in patients who do not have driver mutations and have almost removed the need for chemotherapy alone in the first-line context, except in situations where immunotherapy is contraindicated. ICIs are now used in all first-line NSCLC regimens in this context. Pembrolizumab, atezolizumab, or cemiplimab can also be treated as monotherapy when PD-L1 expression is 50% or higher.³⁴⁻³⁶

1.4.10 HER2 Mutations

HER2 (or ERBB2) differs from EGFR (ERBB1) in that it does not have an endogenous ligand. It promotes oncogenesis through heterodimerization with other members of the ERBB family that then activate various kinase pathways. Despite a rough start with other anti-HER2 agents, TDM-1 or ado-trastuzumab emtansine and trastuzumab deruxtecan have shown much higher ORRs.³⁷⁻³⁹

1.4.11 KRAS

KRAS is a G protein with GTPase activity that is involved in the MAP/ERK pathway; point mutations in the KRAS gene are prevalent at codon 12. KRAS mutations indicate poor survival and nonresponsiveness to EGFR TKIs. Furthermore, KRAS mutations do not appear to impact chemotherapeutic effectiveness and, unlike many other driver mutations, appear to react to immunotherapy.

Despite years of research on the subject, attempts at inhibiting KRAS met with failure. However, more recently, hope has been restored due to presented data from a phase 2 trial (NCT03600883) of sotorasib (Lumakras; Amgen), a TKI that inhibits the KRAS G12C mutation by binding to KRAS in its inactive GDP state. The KRAS G12C mutation occurs in approximately 13% of patients with NSCLC, and therefore accounts for roughly half of all KRAS mutations.⁴⁰

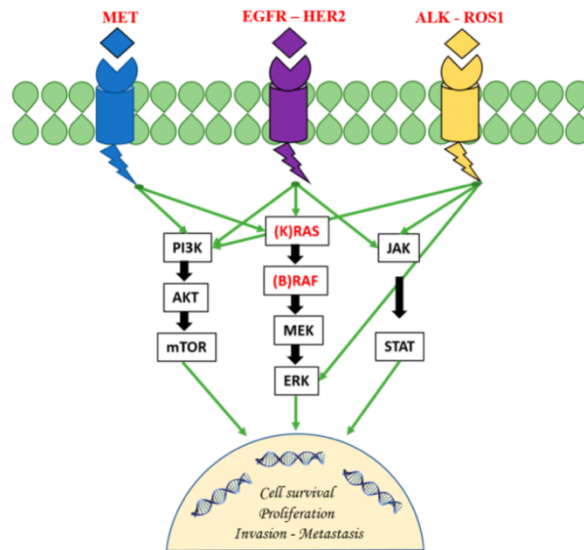


Figure 1.6 The figure depicts the main druggable genetic targets and their involvement in the main signaling pathways in NSCLC.

To summarize, targeted treatments and immunotherapies have transformed the treatment of NSCLC. The use of molecular and immunological tools and ideas has resulted in significant breakthroughs in lung cancer diagnosis. In addition to the epidermal growth factor receptor (EGFR), other molecular targets, such as microRNAs, HER3, and immune checkpoint inhibitors, are being discovered on a regular basis, spurring the development of new therapeutics. Many clinical trials for targeted therapy and immunotherapy drugs are now underway, with promising and exciting findings to date. These trials will aid in the definition of the role of targeted therapy in the treatment of lung cancer, including the role of immune monotherapies, combination immunotherapies, and combinations of targeted therapies with immunotherapies, as well as the optimal timing of these therapies and whether they should be used in early-stage versus late-stage disease. Targeted therapy may one day shift the treatment paradigm for lung cancer, giving patients with few treatment options a hopeful outcome. The hunt for predictors of response to targeted medications continues to be an important subject in clinical research. The ultimate curative option for NSCLC may lie in our ability to couple therapies (either targeted therapies or immunotherapies) with well annotated molecular biomarkers.

1.5 Non-Small Cell Lung Cancer Screen

Although several oncogenic mutational pathways have been characterized (**Figure 1.6**), the current number of actionable mutations only account for roughly 20% of the known oncogenic genome. NSCLC is a highly heterogeneous disease with a mean non-synonymous mutation burden of ~250 mutations/tumor. The heavy mutational burden presents a challenge to understanding the molecular drivers that give rise to this disease, thus posing a complex challenge for novel therapeutic discovery efforts.

Non-small cell lung cancer (NSCLC) cell lines

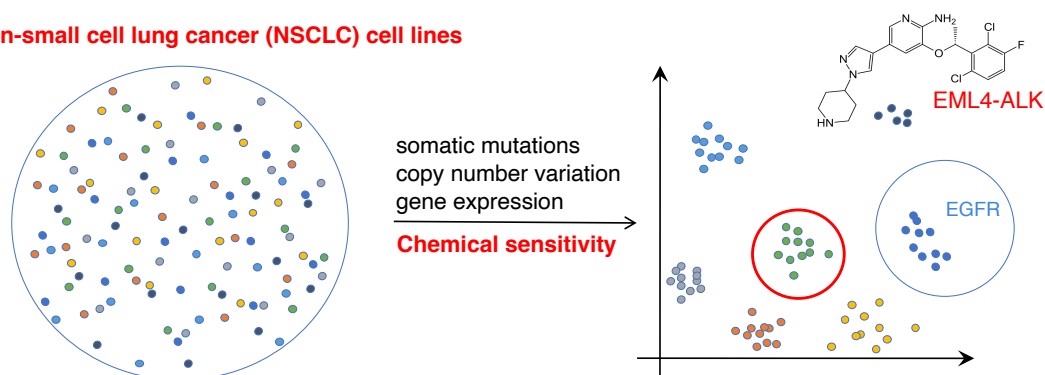


Figure 1.7 Representation of NSCLC cell line clustering via different characterizations. Cell lines may be clustered according to similar genomic signatures, such as somatic mutations (EML4-ALK fusion), copy number variations, or gene expression (EGFR; treated with crizotinib). NSCLC cell lines may also be clustered based on their chemical sensitivities.

In order to address the complex issue of NSCLC drug discovery, our lab participated in a large collaborative effort to design a chemistry-driven screen for the nomination of patient-matched therapeutic interventions. To discover and exploit new vulnerabilities in cancer, (including those which may not be classical “driver mutations”)⁴¹, it is necessary to: A) identify them among the many random changes that occur in a cancer cell; B) have a clinically useful biomarker signifying their presence; and C) an agent to specifically target them. Our screen took an unbiased approach to this problem (**Figure 1.7**): we first utilized our own natural product fraction library to identify a natural product agent that is toxic to a subset of well annotated

NSCLC cell lines, but not to immortalized normal human lung epithelial (and other non-cancer) cells; secondly, we identify tumor molecular correlates that successfully predict sensitivity to the natural product in our preclinical NSCLC models. This creates a novel observation from which we can use bioinformatics to study the cellular responses to the toxic agent to generate hypotheses for the mechanism of action of the agent.⁴²

1.5.1 Natural Product Fraction Library Generation

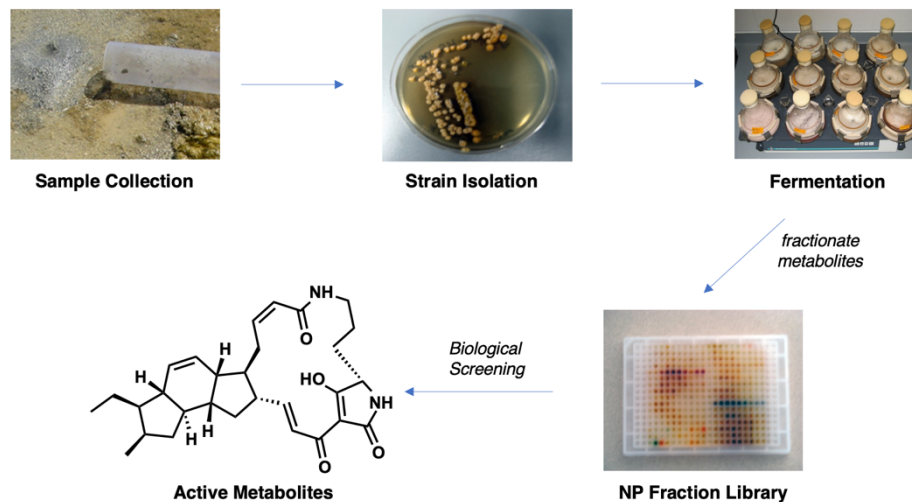


Figure 1.8 General bioactive natural product discovery workflow in the MacMillan Lab.

The MacMillan lab natural product fraction library has been developed through various sample collections of unique marine derived Actinomycetes, Firmicutes, and other non-traditional microbial sources as well as hydrothermal vent associated anaerobic bacteria and Actinomycetes. The collected samples then undergo strain isolation that are then cultured in-house by mimicking the native conditions of the organism. The isolated strains are then fermented in large scale liquid fermentation cultures. Following the ideal growth curve for the given strain, XAD-7 resin is introduced to collect the excreted bacterial metabolites. The resin harboring the bacterial metabolites is then filtered from the fermentation broth and washed with

acetone, which releases the metabolites from the resin to produce a liquid suspension of natural products. The acetone is then dried down to yield a crude extract that is then subject to fractionation via liquid chromatography (**Figure 1.8**). Each fraction has been characterized by LC-UV-MS and has been logged into our Anti-Base repository. The library currently consists of greater than 6,500 natural product fractions from over 3000 different organisms representing 10 different orders and over 200 genera. The chemical diversity of this library contains polyketides, alkaloids, terpenes and hybrid structures of the like.

1.5.2 Panel of Non-Small Cell Lung Cancer Cell Lines

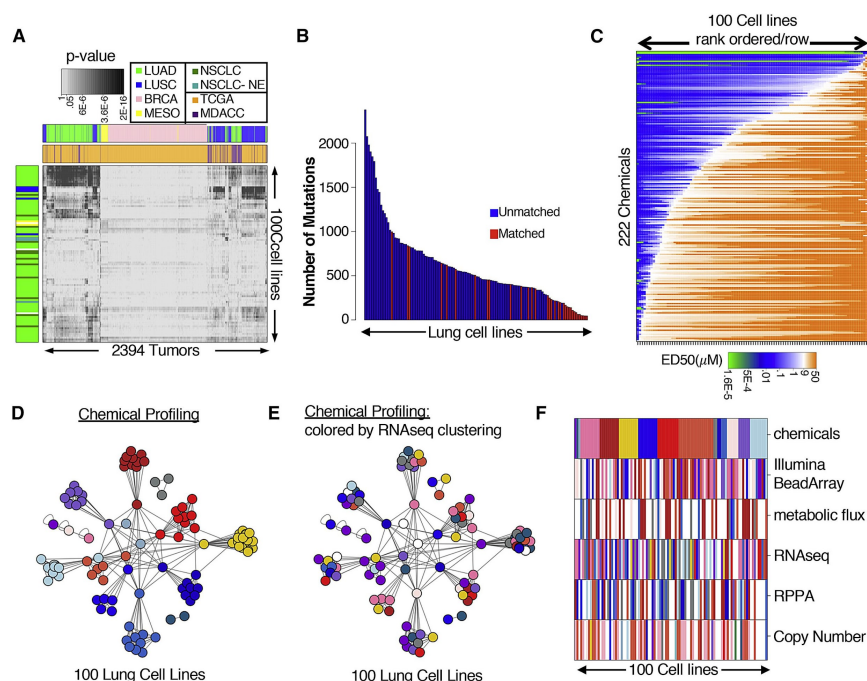


Figure 1.9 Genomic Characterization and Chemical Sensitivities of NSCLC Cell Line Panel. **A** p-values (Pearson) comparing tumors (MDACC, orange; TCGA, purple) and cell lines colored by source. **B** Number of mutations called in the matched (red) and unmatched (blue) subsets. **C** NSCLC sensitivity (ED_{50}) to POPS rank ordered by row. **D** APC clustering by similarity of POPS ED_{50} responses. Nodes are colored according to cluster membership. **E** APC clustered by similarity of POPS ED_{50} responses (as in **D**). Nodes are colored according to cluster membership defined by RNA-seq-based APC. **F** APC clustering across all datasets. Cell lines are ordered according to cluster membership in chemical APC. Each cell line is colored according to cluster membership in the indicated datasets. Reprinted with permission from Elsevier.

Next, in our approach to identifying selective natural product toxins, we employed the world's largest and best-characterized panel of lung cancer cell lines, xenografts, and immortalized normal lung epithelial cells established by Dr. John Minna and Dr. Adi Gazdar. The phenotypic variation of the NSCLC and immortalized bronchial epithelial cell lines was evaluated using legacy whole-genome transcription array data (**Figure 1.9 A**). Whole-exome sequencing (WES), RNA sequencing (RNA-Seq), tiled SNP arrays, reverse phase protein array (RPPA), and heavy carbon tracing was employed to provide a high-resolution molecular characterization of the cell lines. The panel mutational variation of the cell panel was also compared to the B cells of patients and categorized based on whether or not they matched (**Figure 1.9 B**). The chemical sensitivities of the cells, in respect to the screened compounds, were ranked by potency and activity (**Figure 1.8 C**). Cell lines that exhibited similar chemical sensitivity profiles were clustered together (**Figure 1.8 D**) and the clusters were then overlaid with their respective gene expression profiles (**Figure 1.8 E and F**). In summary, our group was able to devise a tiered high-throughput screening strategy to screen large chemical libraries across several highly annotated NSCLC cell lines.⁴³ The isolation, chemical characterization, and biological evaluation of a hit produced by this screen will be discussed in Chapter 2.

1.6 Conclusion

The establishment and utilization of natural product libraries remain a critical source for drug discovery efforts. Coupled with focused high throughput assays and bioinformatics, natural products provide an unrivaled path toward generating novel drug-target-disease associations. Although these strategies may produce natural products that have been previously isolated, they can provide scientists with novel chemical probes to better understand the molecular mechanisms that underpin these diseases. The remaining challenges for natural product drug discovery lie within our ability to quickly isolate, characterize, and re-supply active compounds to researchers investigating the utility of compounds from nature. Developments

in analytical techniques, such as microcrystalline electron diffraction, and biosynthetic gene cluster (BGC) engineering are poised to help relieve some of the final challenges faced by the natural products chemistry field. Nevertheless, natural products continue to push our understanding of chemistry and the roles in which it facilitates our everyday lives.

1.7 Materials and Methods

1.7.1 General Procedures

Low-resolution/ESI-MS data were measured using an Agilent 1200 series LC/MS system with a reversed phase C18 column (Phenomenex Kinetex C18 Evo, 30 mm X 4.6 mm, 2.6 μ m) at a flow rate of 0.3 mL/min.

1.7.2 Cell Lines

Most NSCLC lines used in this study were part of the NCI and HCC (Hamon Cancer Center at UT Southwestern) series of cell lines, with the exception of THLE-2, THLE-3, A427, A549, Calu.1, Calu.6 (American Type Culture Collection; ATCC), Cal.12T (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH; DSMZ), DFCI.024, DFCI.032 (Dana Farber Cancer Institute, courtesy of Pasi Jänne), EK VX, Hop62 (NCI-60 panel), PC9 (Johns Hopkins University School of Medicine, courtesy of Bert Vogelstein). Cell lines from these collections were cultured in RPMI 1640 (GIBCO, 2.05mM L-glutamine) supplemented with 5% FBS (GIBCO) and 1% penicillin/streptomycin (GIBCO). Normal bronchiole epithelia-derived cell lines (Ramirez et al., 2004) were grown in ACL4 (RPMI 1640 supplemented with 0.02 mg/ml insulin, 0.01 mg/ml transferrin, 25 nM sodium selenite, 50 nM hydrocortisone, 10 mM HEPES, 1 ng/ml EGF, 0.01 mM ethanolamine, 0.01 mM O-phosphorylethanolamine, 0.1 nM triiodothyronine, 2 mg/ml BSA, 0.5 mM sodium pyruvate) with 2% FBS and 1% penicillin/streptomycin. Normal liver lines, THLE-2 and THLE-3, were grown in the Bronchial Epithelial Cell Growth Medium (Lonza, CC-3170) supplemented with 10% FBS and 1% penicillin/streptomycin. All cell lines were maintained in a humidified environment in the presence of 5% CO₂ at 37°C. All cell lines were DNA fingerprinted (Powerplex 1.2 Kit, Promega) and mycoplasma free (myco kit, Boca Scientific). All chemicals beginning with the prefix SW are from the UT Southwestern Chemical Library. THZ1 was obtained from

Calbiochem, ciliobrevin from Tocris, GSK923295 from SellekChem, HET-0016 from Santa Cruz Biotechnology, nocodazole from Sigma-Aldrich.

1.7.3 Genomic Characterization

SNP Arrays

Whole-genome single nucleotide polymorphism (SNP) array profiling was done using the Illumina Human1M-Duo DNA Analysis BeadChip (Illumina). Cell line DNA was hybridized according to manufacturer instructions. Processing was first performed using Illumina BeadStudio to generate the 'Log R Ratio' which measures the relative probe intensity compared with normal diploid controls. The package DNACopy in the R statistical software environment was then used to segment the data. Final copy number variation was interpreted as the log₂ segmented copy number values.

RNAseq and Whole Exome Sequencing

FastQC (Babraham Bioinformatics Institute) was used to check the sequencing quality, and high-quality reads were mapped to the human reference genome (hg19) along with the gene annotation data (genecode v19) from Genecode database using STAR (v2.4.2). RSeQC was applied for assessing RNA sample quality Gene-level expression was reported in fragments per kilobase per million reads (FPKM) by Cufflinks.

Illumina BeadChip Microarray

Raw Illumina HumanWG-6 v3.0 BeadChip files were obtained from the Gene Expression Omnibus using accession number GEO: GSE32026 and normalized as described previously (Kim et al., 2016). Briefly, Data were background-corrected using the 'MBCB' package in R, which provides a model-based background correction method similar to an RMA correction with affymetrix arrays. Data were then quantile-normalized to produce equivalent expression distributions among cell lines.

Germline Variant Filtering

The UTSW-92 panel of the cell lines corresponded to those in which we have tumor DNA but corresponding matched non-tumorigenic DNA is not available. These correspond to 68 lines from the ‘training set’ of cell lines and 24 lines from the ‘testing set’ of cell lines. For these, we developed a pipeline to filter out the most probable germline mutations and enrich for somatically acquired mutations. Reads were aligned as described to the hg19 reference and filtered for non-synonymous lesions (missense, non-sense, splice site mutations) (mean of 5,049 mutations/cell). We next removed any site that was annotated as corresponding to a germline mutation in the matched dataset (mean of 1,248 mutations/cell). Using publicly available datasets such as the thousand genome project (TGP) as an exclusion criterion or the catalog of somatic mutations in cancer (COSMIC) as an inclusion criterion may aid in enriching for somatic mutations. We removed variants (defined based on genomic position) that were found in > 12% of the TGP (TGP filter) and where the difference in the UTSW panel frequency and the TGP frequency was < 1.8% (allele difference filter). We also removed, on a gene-level basis, genes that were highly mutated (mutated at any site in > 40% of cell lines) in the UTSW panel (mutation any site filter), but present at a low frequency (< 13%) in COSMIC (Cosmic filter) and in the UTSW-34 matched panel (< 20%) (UTSW-34 filter). This resulted in a final mean mutation count of 718 mutations/cell. We developed a strategy to find a data driven way select optimal filter cutoffs from these datasets. We selected 12 evenly distributed values for the TGP filter between 0.02% and 20%, for the allele difference filter between -10% and 10%, for the mutated any-site filter between 1.8% and 80%, for the Cosmic filter between 0.13% and 20% (\log_{10} scale), and for the UTSW-34 filter between 2.9% and 50%. Selecting all possible combinations of these filters resulted in 248,832 possible combinations. For each filter combination, we can plot the number of mutations that pass the filters, with the strictest filter combination resulting in the fewest variant being annotated as ‘somatic’ and the most lenient resulting in the most variants being included. To select the optimal filter combination in a data-

driven way, we fit a cubic function to the plot of filter index (x values) versus a number of mutations included at each filter index (y-axis) and selected the value on the plot which results in the minimized second derivative for each cell line.

1.7.4 Small Molecule Cytotoxicity Assays

Our chemical library, consisting of ~200,000 chemicals (Figures S1B and S1G), was initially screened at a single dose (2.5 μM) in a single well for each compound against a panel of 12 NSCLC cell lines. Toxicity data were converted to an activity score according to the following equation:

$$AC = -1 \times \left(100 - \frac{x}{\text{median}(x_{\text{control}})} \times 100\right)$$

so that an activity score indicates percent kill relative to on-board DMSO controls. We subsequently converted activity scores to z-scores for each chemical across the 12 cell line panel and selected chemicals with $z \leq -3$ in at least one cell line, resulting in 15,483 chemicals (single dose cohort). These chemicals were then re-screened in triplicate against the same 12 NSCLC cell lines along as well as an immortalized human bronchial epithelial cell line (HBEC30KT) at the screening dose of 2.5 μM (confirmation dataset). From this dataset, we used two criteria to select chemicals for further follow-up. We first filtered for chemicals with a bimodal pattern of response from our panel of cell lines. Specifically, we selected chemicals with > 40% toxicity to a subset of cell lines and < 20% toxicity to the remaining NSCLC's and HBEC30KT. As determined in downstream dose-response studies, compounds that met this criterion typically displayed IC50's in the range of our screening dose or lower for a subset of the NSCLC lines and IC50 values > 10 μM in the remaining cell lines in the panel and the HBEC30KT cell lines. In terms of chemical selectivity, we expect this selection to result in compounds with at least a $\frac{1}{2}$ log difference in response between sensitive and resistant cell lines. We also used a selection method to capture potent chemicals with more of a continuous distribution of cytotoxicity in our 12 cell line panel. For each compound, the responses of the

cell lines were ranked from most sensitive to least. The difference (Δn) in response between each pair of ranked cell line activities for each compound was calculated. The S-score is the maximum difference (Δn_{\max}) between two cell lines' responses in the ranked list of responses to the compound. The two cell line responses that define the S-score, therefore, demarcate a boundary between sensitive and resistant response groups in the ranked list of responses for each compound. We selected chemicals for follow-up to be those with the S-score > 40%, while enforcing the criteria that the chemical not be toxic to HBEC30KT (< 20% observed toxicity). These chemicals were subjected to chemistry review that removed compounds with known or suspected promiscuous (off-target) behavior based on historical screening data, structural alerts, and PAINS substructures. Following resupply (1 – 5 mg of powder per compound) and analytical quality control for identity and purity (LC/MS), 447 compounds were assayed in a multi-dose format (12 point dose-response curves in $\frac{1}{2}$ log dilutions with the doses ranging from 50 pM to 50 μ M) against the same panel of 12 NSCLC cell lines plus the HBEC30KT cell line. Each compound was assayed twice in this format and the dose-response curves were compared. In cases where experimental replicates differed by more than 3-fold, we performed a third dose-response experiment and averaged the two experimental replicates that were in closest agreement. We used the same unimodal (S-score) method to select a total of 208 chemicals to be screened across the entire panel of 100 cell lines. In this case, we rank-ordered average $\log_{10}(\text{IC}_{50})$ values for each compound and applied a threshold of 0.5 log units for the S-score.

CHAPTER TWO
ISOLATION AND STRUCTURAL AND BIOLOGICAL CHARACTERIZATION OF
IKARUGAMYCIN

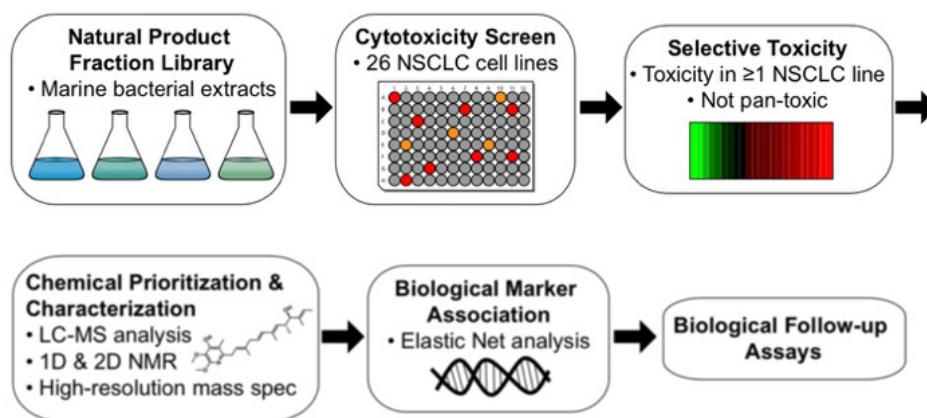


Figure 2.1 Natural product fraction prioritization workflow.

2.1 Analysis of NSCLC Screening Data

In Chapter 1, I described a nonbiased screening platform designed to identify selective natural product toxins for non-small cell lung cancer (NSCLC).⁴³ As the assay was shown to be functional, it was critical to conduct follow-up studies on the identified hits. ‘Hits’ can be described to have produced the desired phenotypic expression, cytotoxicity. In this chapter, I will describe how we took the phenotypic readout of the assay and prioritized a subset of fractions for further investigation (**Figure 2.1**).

2.1.1 Identification of Subtype-selective Inventions of Non-Small Cell Lung Cancer

As part of our efforts in our screening platform designed to identify select toxins for non-small cell lung cancer (NSCLC), we carried out a primary screen of synthetic chemicals, siRNAs, and natural product fractions against representative members of the NSCLC cell lines. The phenotypic readout of our natural product fractions reflected various levels of cytotoxicity (**Figure 2.2**). The representative NSCLC cell lines are displayed along the y-axis and the natural product fractions are displayed along the x-axis. Fractions that exhibited toxicity to the NSCLC panel are represented by green bars, while the fractions that exhibited no toxicity are represented by red bars.

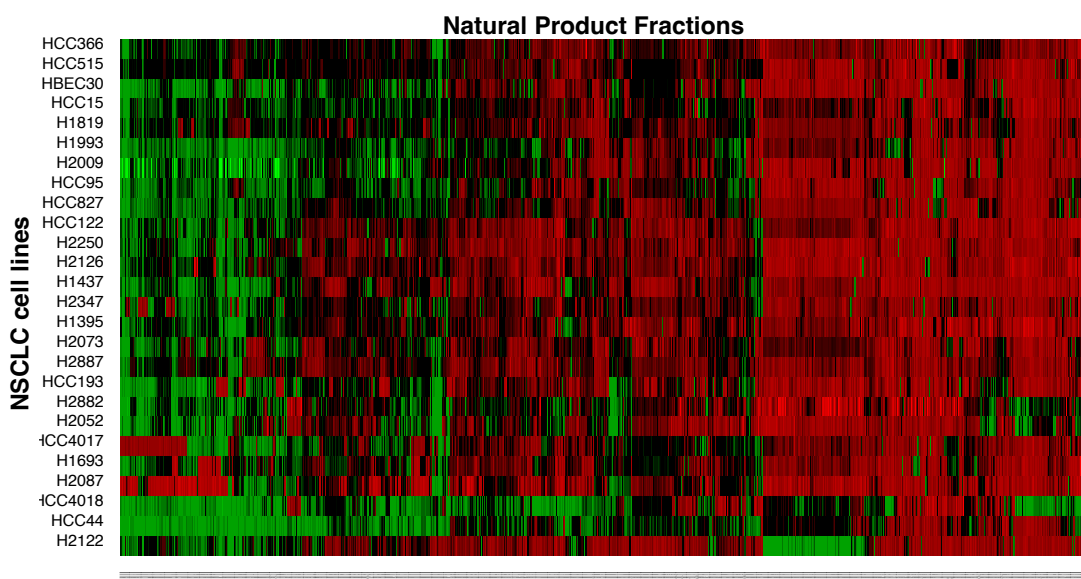


Figure 2.2 Phenotypic readout of preliminary NSCLC screen. Natural product fractions are shown across the x-axis, while NSCLC cell lines are shown across the y-axis. Green represents cytotoxic readout; red represents non-toxic read out.

Our traditional screening methods for the identification of chemical toxins utilized standard concentrations of 5 or 10 μM . As we set out to find selective toxins, screening the natural product fraction library at our standard concentrations presented a unique challenge due to the potency of many natural products. As such, we elected to carry out a limited dose-response for all fractions at 5, 1.8, 0.5, and 0.18 μM . Carrying out the screen in a limited dose-response allowed us to quickly identify those fractions that were both potent and selective over a larger concentration range. In doing so, we were able to prioritize for fractions that exhibited the desired cytotoxicity profile (**Figure 2.3**). For example, we identified fractions 15 and 16 from our strain SNB-039 as eliciting a pan-toxic response from our representative NSCLC cell lines. Fraction 16 from our strain SNA-096 and fraction 10 from our strain SNA-097 elicited a cytotoxicity profile that was selective for one cell line. Although pan-toxic and single-cell selective cytotoxicity profiles can be used to identify novel NSCLC toxins, our initial intention was to identify cell subtype-selective toxins. Our hypothesis was that by prioritizing for subtype-

selective toxicity profiles, we would be able to cluster the cell lines that were sensitive to our fractions for further genomic interrogation.⁴⁴ Once the implicated natural product was isolated, we would then use the compound as a probe to better understand the genetic make-up that underpinned the sensitivity to the probe. As such, we decided to move forward with the identification of the bioactive natural product fraction in fractions 13 and 14 of our strain SNB-040.

SNB-040 fractions 13 and 14: selective for a few cell lines

[C]	HBEK-30KT	HCC366	H1993	H2009	H2122	HCC15	HCC827	H2073	HCC44	H2887	HCC193	H1819	HCC515	H2347	HCC95	H2250	H1437	
SNB-040-13	0.18	1.82858	-0.10205	0.197159	0.529948	-3.90215	7.50741	-1.49939	-9.97888	-3.53422	-2.99257	-8.88815	2.57643	-0.38493	5.876281	-1.81186	3.43944	0.817387
SNB-040-13	0.55	-0.3497617	51.8351	0.5274042	2.75573	-2.26959	-29.01583	-31.9513	2.19436	-11.04878	-19.0884	-82.09248	-17.94579	-6.09361	-2.12374	-1.57376	3.93353	2.28416
SNB-040-13	1.05	-76.8418	96.75926	-75.8899	-41.84926	6.955743	-79.14972	46.79073	86.2858	94.83162	-82.43996	97.96471	84.54939	-88.61437	-71.42303	-88.02987	-13.91115	1.834824
SNB-040-13	0.05	96.3362	-97.62511	97.84466	-106.5332	94.69795	-95.19986	84.37975	86.3865	-97.44654	-97.60161	-98.01331	-96.29628	-97.2891	-97.46573	-86.69166	-98.09728	-25.73482
SNB-040-14	0.18	0.7782176	2.542116	0.2007621	-1.536368	-5.723119	-12.07146	-10.68327	-7.381722	-2.965849	-3.651489	-15.16561	-6.804036	-3.200526	2.081979	-0.0499978	-3.186828	0.702942
SNB-040-14	0.55	-2.471447	-71.17397	-5.060446	1.302935	-1.028975	-58.9795	-48.06662	-33.10688	-17.48317	-29.83836	-80.64539	-29.08953	-0.226133	-6.803807	-7.784667	0.8307836	1.508217
SNB-040-14	1.05	-32.21320	95.56478	-86.39421	-101.0587	4.460775	-67.76071	57.72481	74.22344	-55.86082	65.37313	-88.06143	55.38459	-24.49978	32.49469	-20.9235	-17.2926	2.804304
SNB-040-14	0.05	98.16468	-97.74488	-97.99749	-103.2189	92.68705	-95.82001	90.54113	-97.96647	-97.17852	-96.80316	-98.09853	-95.13991	-97.43974	-97.00789	-97.006	91.91898	-46.92305

SNB-039 fractions 15 and 16: pan-toxic

[C]	HBEK-30KT	HCC366	H1993	H2009	H2122	HCC15	HCC827	H2073	HCC44	H2887	HCC193	H1819	HCC515	H2347	HCC95	H2250	H1437	
SNB-039-15	0.18	85.9986	49.4087	89.88187	89.8965	72.39323	84.9029	73.3794	89.8394	97.1394	89.7999	85.6802	95.2678	96.3854	84.8449	78.2411	54.65091	39.24426
SNB-039-15	0.55	88.83823	77.5349	92.05017	-92.84122	-76.82318	-95.33662	-78.05078	-94.80574	-95.48262	85.91998	-86.86884	86.47294	-85.56644	-87.2376	-85.43026	-64.99881	-74.94312
SNB-039-15	1.05	96.87995	-78.19072	95.00456	-96.16575	-89.28362	-95.74217	-90.73588	-95.50244	-94.20938	-97.79108	-99.85819	-99.72211	-95.8019	-91.42394	-85.34238	-81.45839	-81.45839
SNB-039-15	0.05	96.12466	85.57506	97.71939	-104.8424	96.0078	96.30871	95.05445	96.07836	96.66256	-96.71881	-95.50187	-96.6011	96.81725	-97.06197	-93.88467	61.77869	61.77869
SNB-039-16	0.18	30.77433	-43.72076	-62.55092	-80.62083	-22.81242	-88.2181	-49.62059	-71.67211	-92.18115	-46.38343	-81.87833	-33.67881	-55.106	-71.84641	-46.96029	-4.984473	64.23478
SNB-039-16	0.55	-58.95263	-64.79375	-79.7488	-83.7604	-52.82349	-92.083	-67.78725	-85.55317	-94.79708	-93.94096	-93.86651	-56.17953	-68.18752	-77.04944	-23.80483	-18.85196	67.74474
SNB-039-16	1.05	-88.73979	-74.7467	-92.16068	-95.48039	-79.75093	-94.88702	-87.03013	-95.74347	-95.79269	-86.39171	-96.51183	-88.19659	-78.92873	-89.15173	-89.52836	-61.20802	-73.30516
SNB-039-16	0.05	97.89317	84.96722	95.75822	-104.2517	88.21377	96.37728	92.9747	-96.0268	-96.24109	-94.93289	-97.69123	-98.7871	-98.13214	-93.18451	-86.44077	-78.91882	-77.74925

SNA-096 fraction 16 and SNA-097 fraction 10: selective to one cell line

[C]	HBEK-30KT	HCC366	H1993	H2009	H2122	HCC15	HCC827	H2073	HCC44	H2887	HCC193	H1819	HCC515	H2347	HCC95	H2250	H1437	
SNA-096-16	0.18	-0.847999	-0.858857	-0.4727144	-0.8169523	-0.80925	-4.34401	-5.731465	-16.91568	-3.765476	-0.3777134	-0.888657	-0.7044109	-2.454453	-7.86282	2.335191	1.189212	-10.35459
SNA-096-16	0.55	-1.91297	-10.85544	-10.16426	1.189758	96.39547	-17.50731	-4.607521	-30.50032	-11.05592	8.408069	-8.69508	-7.176849	-0.6952676	-6.340884	0.9174312	6.123744	-44.36871
SNA-096-16	1.05	-4.312368	-22.35546	-13.85791	2.102176	96.8978	-34.90662	-8.084054	-35.96952	-38.75949	-3.669897	-4.47855	-3.703985	-4.052131	-11.34181	-0.0716823	-4.60995	-41.89317
SNA-096-16	0.05	1.275381	-24.59094	-9.94721	-0.902602	-86.6532	-7.670688	-8.586505	-23.20969	-5.23028	1.532366	0.779546	-1.97864	-7.011843	-12.75116	-5.766222	-7.015836	-11.06682
SNA-097-10	0.18	3.122767	-2.33873	-29.0934	-4.614753	4.612888	2.619039	-17.9647	-99.8488	-88.80522	-41.46385	-39.61133	-34.98916	-20.10906	-41.63276	-45.77955	37.96331	-5.10971
SNA-097-10	0.55	1.820944	-7.81702	1.465887	-1.801508	-10.82439	-78.93838	1.629679	-40.0044	90.44668	-12.89617	37.24809	-47.33828	18.36116	4.716991	65.70946	15.54996	-13.3365
SNA-097-10	1.05	0.6125037	-11.59151	-43.94889	-9.434425	-3.971173	-8.234017	-22.86773	-57.26706	91.36131	-46.66887	-44.2906	-22.23842	16.38152	-36.8093	-34.69677	-19.82454	-0.397012
SNA-097-10	0.05	-0.1692397	-16.4908	-14.84994	-8.24278	-6.277823	-18.36203	-90.57713	-42.17713	-98.81287	-25.17342	-28.07586	-10.04555	-23.71765	-33.1923	12.21891	5.785701	-11.99505

Figure 2.3 Representation of different cytotoxicity profiles observed in preliminary NSCLC screen.

Having set out to find novel NSCLC toxins with the best chance of becoming clinically relevant, our preliminary datasets utilized HBEK30KT and HCC366 cells alongside the representative NSCLC cell lines. HBEK30KT cells are immortalized lung epithelial cells – in other words, normal healthy human lung epithelial cells that are non-cancerous. HCC366 cells are NSCLC cells that have expressed a large mutational load and have been identified to develop resistance to treatment, such as the acquired resistance to the DDR2 inhibitor dasatinib. Including the HBEK30KT cell line in our preliminary screen allowed us to prioritize natural product fractions that exhibited discriminative activity NSCLC cell lines and not against the HBEK30KT, thus pushing the identified bioactive natural product fraction closer to clinical relevancy. Screening against the HCC366 cell lines, and those of its kind, allowed us to further prioritize natural product fractions that exhibited selective activity against treatment-resistant

NSCLC cell lines with the idea that a hit would be able to provide insight into a novel druggable target.

We identified the natural product fractions SNB-040-13 through SNB-040-17 and SNE-002-15 through SNE-002-19 as exhibiting selective toxicity profiles against NSCLC cell lines and discriminatory non-toxic profiles against HBEC30KT cell lines. Amongst the NSCLC cell lines, we were able to identify cell lines that were either sensitive (NPF caused cytotoxicity readout) or resistant (NPF did not cause cytotoxicity readout). We utilized representative cell lines, HCC366 and Calu-1 (sensitive) and HCC44 and H650 (resistant), to direct our bioassay-guided isolation of the active metabolite found in the fractions of interest.

2.2 Isolation of the Bioactive Metabolites from SNB-040 and SNE-002 for NSCLC Toxicity

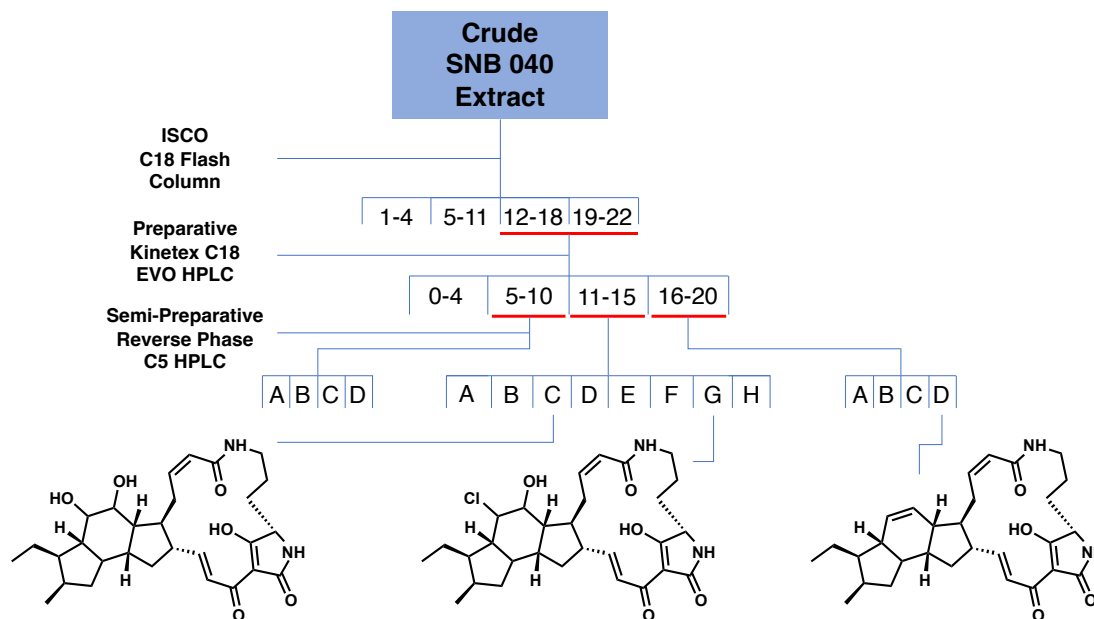


Figure 2.4 Purification schematic for the isolation of bioactive metabolites identified in SNB-040.

Fractions SNB-040-13 through SNB-040-17 are NPFs derived from the marine bacterial strain *Streptomyces carpaticus*. Analysis of the active fractions by LC-UV-MS showed a number of metabolites with λ_{\max} absorptions at 225nm and 325nm correlating to a mass to charge range between 450 and 650 m/z $[M+H]^+$. A large-scale regrow (20L) by shake

fermentation was carried out to obtain sufficient material for the purification, isolation, and full chemical and biological characterization of the active metabolites (**Figure 2.4**). The crude extract was initially purified using reversed-phase (C18) flash columns with step gradients (30%-100% MeOH:H₂O) to yield 22 fractions. Fractions 12 through 22 were found to possess modest to strong selectivity and potency for the representative NSCLC cell lines. Therefore, fractions 12 through 22 were pooled together and purified using preparative-scale reversed-phase HPLC (Phenomenex Kinetex C18 Evo) with a linear gradient (10%-100% ACN:H₂O) to yield 21 fractions. The resulting fractions were then separated according to their elution times and re-collected according to their observed mass to charge ratio via LC-UV-MS analysis. The fraction groups (F5-F10, F11-F15, and F16-F20) were then subject to further purification utilizing semi-preparative reversed-phase HPLC (Phenomenex Luna C5) with a linear gradient (40%-90% ACN:H₂O) until pure peaks were isolated.

Fractions SNE-002-15 through SNE-002-19 are NPFs derived from the marine bacterium SNE-002 that was isolated from a sediment sample collected from a hypersaline lake at East Plana Cay, Bahamas. Bacterial spores were collected via stepwise centrifugation and isolated on a humic acid media. Analysis of 16S rRNA revealed 99% identity to *Streptomyces xlamenemycin*. Analysis of the active fractions by LC-UV-MS showed a number of metabolites with λ_{\max} absorptions at 225nm and 325nm correlating to a mass to charge range between 450 and 650 m/z [M+H]. A large-scale regrow (20L) by shake fermentation was carried out to obtain sufficient material for the purification, isolation, and full chemical and biological characterization of the active metabolites. The excreted metabolites were collected using XAD-7 resin and the resulting crude extract was initially purified using reversed-phase (C18) flash columns with step gradients (30%-100% MeOH:H₂O) to yield 22 fractions. Fractions found to possess modest to strong selectivity and potency for the representative NSCLC cell lines were pooled together and purified using preparative-scale reversed-phase HPLC (Phenomenex Kinetex C18 Evo) with a linear gradient (10%-100% ACN:H₂O) to yield 21 fractions. The resulting fractions were then separated according to their elution times and re-collected

according to their observed mass to charge ratio via LC-UV-MS analysis. The fraction groups were then subject to further purification utilizing semi-preparative reversed-phase HPLC (Phenomenex Luna C5) with a linear gradient (40%-90% ACN:H₂O) until pure peaks were isolated, following the same purification schematic laid out in **Figure 2.4**.

2.2.1 Ikarugamycin

High-resolution mass spectrometry analysis determined an m/z [M+H] of 479.2908 corresponding to a molecular formula of C₂₉H₃₈N₂O₄ (calculated m/z 479.2910) and accounting for 12 degrees of unsaturation. Using 1- and 2-D ¹H NMR techniques the natural product was successfully characterized as ikarugamycin, a polycyclic tetramate macrolactam (PTM) (**Figure 2.5**, **Table 2.1**), and confirmed by comparing data to that found in the literature. Originally isolated as an antibiotic⁴⁵, ikarugamycin was recently identified as a promising lead for anticancer therapeutics but has not been characterized for cytotoxicity against NSCLC cell lines. Other PTMs such as discodermide,^{46,47} ikarugamycin oxide,^{48,49} cylindramide⁵⁰ have also been reported to have anticancer activities against various cancer cells.

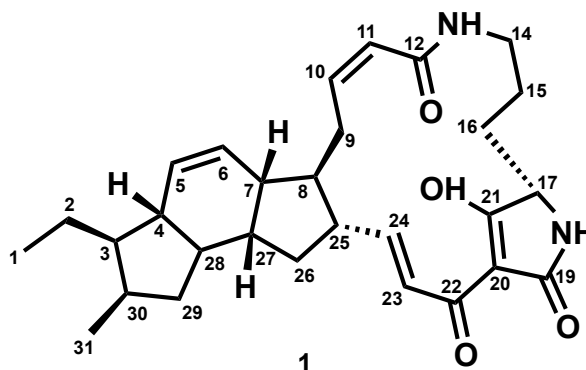


Figure 2.5 Structure of ikarugamycin, a microbial natural product.

Table 2.1 ¹H (800 MHz) and ¹³C (800 MHz) spectroscopic data of **1**.

Position	δ H, Mult (J in Hz)	δ C, Mult	Position	δ H, Mult (J in Hz)	δ C, Mult
1	0.98, t (7.0)	13.28	17	3.95, br s	61.32
2	1.41, m; 1.51, m	21.63	NH-18	6.15, br s	-
3	1.62, m	47.72	19	-	173.98

4	1.42, m	46.97	20	-	100.38
5	5.99, d (10.0)	131.59	21	-	195.81
6	5.74, dt	128.07	22	-	175.51
7	2.56, m	42.92	23	7.19, d (15.4)	122.18
8	1.20, m	48.3	24	6.83, dd (15.4, 10.6)	152.86
9	2.44, d (10.6); 3.52, m	25.33	25	2.57, m	49.51
10	6.11, td	141.13	26	1.29, m; 2.18, m	36.71
11	5.87, d (10.6)	123.98	27	2.12, m	41.76
12	-	166.31	28	1.62, m	48.6
NH-13	5.92, br s	-	29	0.74, ddd (12.0, 12.0, 6.8); 2.14, m	38.46
14	2.68, s; 3.74, br s	38.87	30	2.31, ddd (7.6, 7.6, 7.6)	33.05
15	1.29, m; 1.63, m	21.1	31	0.92, d (7.2)	17.71
16	1.87, m; 2.07, m	27.67			

Measured in CDCl₃. δ values given in ppm.

2.2.2 Capsimycin D

A chloride-containing ikarugamycin (**1**) analog was isolated from *S. carpaticus* and became a high priority target due to the relatively low number of published halogenated PTMs. With an ionized molecular peak at m/z 531.26, compound **2** showed an isotopic peak at m/z 533.26 with a high relative intensity of 3:1 in the LC-MS spectrum; indicating a halogenated substituent compound. The molecular formula for **2** was determined as C₂₉H₃₉ClN₂O₅ and confirmed by comparing the HR-MS data found in the literature. The structural characterization of **2** was determined by extensive NMR data analysis and confirmed with NMR data found in the literature (**Figure 2.6, Table 2.2**). Although capsimycin D is a recently known compound⁵¹, it has not been characterized for cytotoxicity against a panel of NSCLC cell lines.

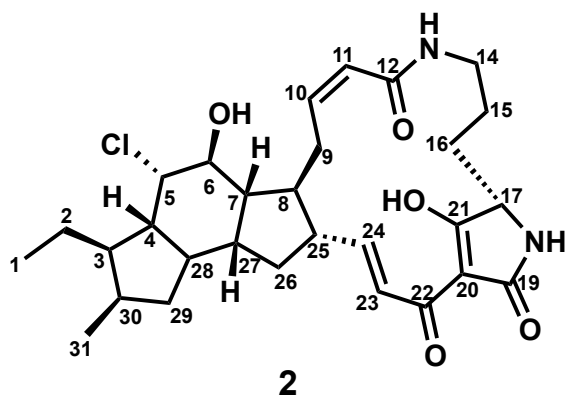


Figure 2.6 Structure of ikarugamycin analog, capsimycin D.

Table 2.2 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **2**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	0.94, t (7.3)	12.8	17	3.88, dd (5.5, 2.1)	61.6
2	1.35, m	21.2	NH-18	-	
3	1.76, d (3.3)	44.7	19		173.65
4	1.77, m	47	20		100.8
5	3.13, dd (3.8, 2.0)	57.7	21		197.1
6	2.89, d (3.8)	53.5	22		175.6
7	2.07, m	47.3	23	7.13, d (15.4)	122.18
8	2.14, m	45.6	24	6.83, dd (15.4, 10.6)	153
9	2.53, dd (17.3, 3.0); 3.38, m	26.4	25	2.57, m	49.51
10	6.06, ddd (11.5, 11.5, 3.4)	141.6	26	1.29, m; 2.13, m	35.6
11	5.84, dd (11.5, 1.3)	123.7	27	2.07, m	42.6
12		167.2	28	1.61, m	41.1
NH-13	-		29	0.75, ddd (12.0, 12.0, 6.8); 2.19, d (7.6)	38.6
14	2.65, br t (11.2); 3.55, ddd (11.2, 4.9, 3.0)	39	30	2.21, m	32.6
15	1.29, m; 1.63, m	21.1	31	0.92, d (6.8)	17.71
16	1.18m, 2.05 m	27.5			

Measured in DMSO- d_6 . δ values given in ppm.

2.2.3 Capsimycin B

An ikarugamycin (**1**) analog was isolated from *S. carpaticus* and became a priority for isolation due to its' unique mass, an ionized molecular peak at m/z 495.28 (calculated 495.2859). The molecular formula for **3** was determined as $C_{29}H_{39}N_2O_5$ and confirmed by comparing the HR-MS data found in the literature. The structural characterization of **3** was determined by extensive NMR data analysis and confirmed with NMR data found in the literature (**Figure 2.7**, **Table 2.3**). Although capsimycin B is a recently known compound⁵¹, it has not been characterized for cytotoxicity against a panel of NSCLC cell lines.

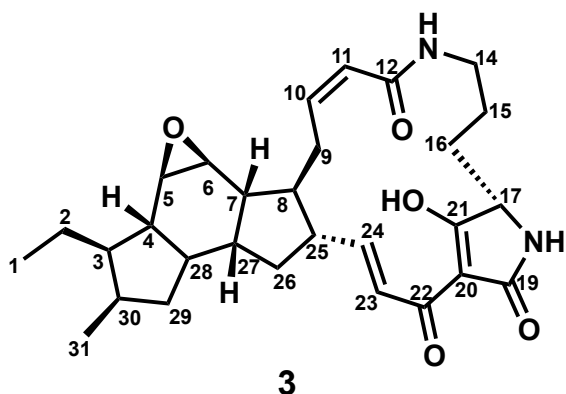


Figure 2.7 Structure of ikarugamycin analog, capsimycin B.

Table 2.3 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **3**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	0.98, t (7.0)	13.28	17	3.95, br s	61.32
2	1.41, m; 1.51, m	21.63	NH-18	6.15, br s	-
3	1.62, m	47.72	19	-	173.98
4	1.42, m	46.97	20	-	100.38
5	5.99, d (10.0)	131.59	21	-	195.86
6	5.74, dt	128.07	22	-	175.51
7	2.56, m	42.92	23	7.11, d (15.4)	122.68
8	1.20, m	48.3	24	6.73, dd (15.4, 10.6)	151.86
9	2.44, d (10.6); 3.52, m	25.33	25	2.57, m	49.51
10	6.06, td	140.33	26	1.29, m; 2.18, m	36.71
11	5.79, d (10.6)	124.18	27	2.12, m	41.76
12	-	166.31	28	1.62, m	48.6

NH-13	5.92, br s	-	29	0.74, ddd (12.0, 12.0, 6.8); 2.14, m	38.46
14	2.68, s; 3.74, br s	38.87	30	2.31, ddd (7.6, 7.6, 7.6)	33.05
15	1.29, m; 1.63, m	21.1	31	0.92, d (7.2)	17.71
16	1.87, m; 2.07, m	27.67			

Measured in CDCl₃/CD₃OD. δ values given in ppm.

2.2.4 Capsimycin F

Capsimycin F (**4**) was obtained as white amorphous solid with UV-Vis absorptions λ_{\max} 327 and 223 nm (**Figure 2.8**). Its positive ion HRESIMS revealed a pseudomolecular ion peak at m/z 527.3108 $[M+H]^+$, corresponding to a molecular formula of C₃₀H₄₂N₂O₆ (calcd for C₃₀H₄₃N₂O₆ 527.3121). Its ¹H NMR signals in CD₃OD exhibited four downfield signals at δ_H 6.07 (td, 11.3, 3.8 Hz), 5.85 (d, 11.5 Hz), and 7.40 (d, 15.4Hz), 6.76 (dd, 15.4, 10.3Hz), suggesting the existence of one Z- and one E- double bonds (**Table 2.4**). The analysis of ¹³C and HSQC NMR spectra concluded 30 carbon signals, including one methoxyl, two methyls, seven methylenes, 15 methines (including three hetero substituted carbons and four olefinic carbons), and five downfield shifted quaternary carbons including one amide carbon at δ_C 167.9 and characteristic tetramic acid signals at δ_C 175.9, 101.1, 197.2 and 173.0 (**Table 2.4**).

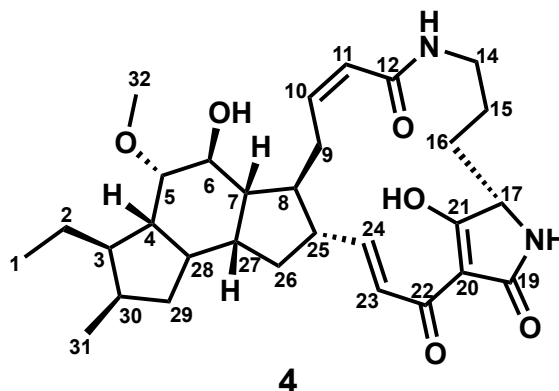


Figure 2.8 Structure of ikarugamycin analog, capsimycin F.

Position	δ_H , Mult (J in Hz)	δ_C , Mult	Position	δ_H , Mult (J in Hz)	δ_C , Mult
1	0.96 (t, J = 7.3)	12.6	17	3.86, s	61.9
2	1.39 m	21.7	NH-18	-	-

3	1.75 m	42.9	19	-	175.9
4	1.48 m	46.7	20	-	101.1
5	3.40 m	81.5	21	-	197.2
6	4.09 m	67.8	22	-	173.0
7	2.03 m	47.9	23	7.40 (d, J = 15.4)	122.4
8	2.03 m	44.8	24	6.76 (dd, J = 15.4, 10.3)	152.3
9	3.49 m, 2.47 m	26.7	25	2.38 m	50.3
10	6.07 (td, J = 11.3, 3.8)	141.7	26	2.08 m, 1.25 m	35.3
11	5.85 (d, J = 11.5)	123.7	27	1.55 m	42.0
12	-	167.9	28	2.02 m	42.0
NH-13	-	-	29	2.16 m, 0.67 m	39.3
14	3.45 m, 2.67 (t, J = 10.9)	39.0	30	2.17 m	32.8
15	2.01 m, 1.97 m	21.0	31	0.90 (d, J = 6.4)	17.0
16	2.01 m, 1.87 m	27.4	32		58.0

Measured in CD₃OD. δ values given in ppm.

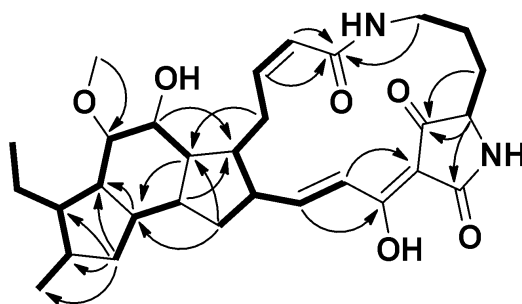
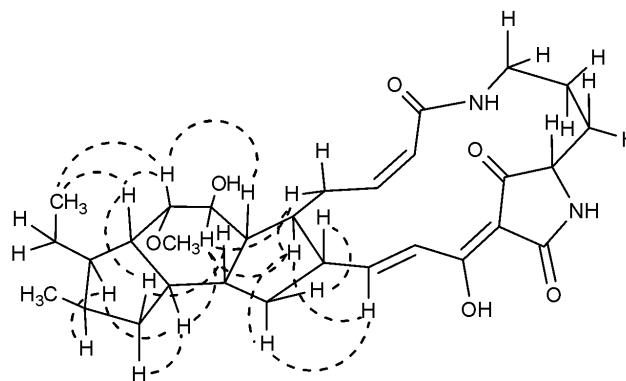


Figure 2.9 Key COSY and HMBC correlations of **4**.

In the COSY spectrum, the signals for spin systems CH₃-CH₂-CH(CH₃)-CH-CH-CH-CH (CH₃-1 through H₂-2, H-3, H-4, H-5, H-6 to H-7 and from H-3 through H-4 to CH₃-5), CH₂-CH-CH (from H₂-29 through H-28 to H-27), and CH-CH-CH₂-CH-CH-(CH₂)-CH-CH (from H-11 through H-10, H₂-9, H-8, H-25 and H-24 to H-23, and from H-25 to H₂-26) were clearly observed. Due to the close overlapping of some proton signals, the connectivity of the above three spin systems were built up based on the interpretation of its HMBC spectrum. The long-range correlations from H-29 to C-30, C-31, C-3 and C-4 as well as from H-28 to C-4 indicated rings A and B were fused at C-4 and C-28. Meanwhile, the relationship of ring B and C was

suggested by the observation of correlations from H-6 and H-27 to C-8, from H-7 and H₂-26 to C-28, from H-9 and H₂-26 to C-7. One methoxyl group was assigned at C-5 (δ_c 81.5) due to the observation of HMBC correlation from OCH₃ (δ_H 3.43) to C-5. Therefore, the molecule was determined to have a 5,6,5 tricyclic ring segment. The COSY correlations established the spin system CH₂-CH₂-CH₂-CH (from H₂-14 through H₂-15 and H₂-16 to H-17), which was proved to be connected with α , β unsaturated amide by the HMBC signals from H-10, H-11 and H₂-14 to C-12 (δ_c 167.8). The carbon chemical shifts of tetramic acid (δ_c 176.4, 100.4, 194.0, 174.4) were ascribed to C-19, C-20, C-21 and C-22 based on the HMBC correlations from H-23, H-24 to C-22, from H-23 to C-20 as well as from H-17 to carbonyl C-19 and C-21. Therefore, the planar structure of **4** was established to be a 5,6,5 tricyclic tetramic acid amide (**Figure 2.9**), an analog of ikarugamycin (**1**).



The stereochemistry of **4** was established on the NOESY correlations. The correlations from H-4 to CH₃-1, H-5 and H-29 β , from H-5 to H-7, and from H-29 β to H-27 suggested H-4, H-5, H-7, H-27 in ring A and B adapted β orientation. Meanwhile, the NOESY correlations from H-25 to H-26 β revealed the β orientation of H-25. (**Figure 2.10**). The observed NOESY signals from H-28 to H-29 α , H-6, and H-3, from H-3 to H-30, from H-8 to H-6, H-24 and 26 α , and from H-26 α to H-24 demonstrated H-3, H-6, H-8, H-28, H-30 were α orientation as depicted in **Figure 2.10**. Since all natural occurring PTMs are having L-ornithine incorporated in the molecule,

given the biogenetic consideration, H-17 was considered to being β orientation. Therefore, the molecular structure of **4** was as shown in **Figure 2.8**.

2.2.5 Capsimycin C

Capsimycin C (**5**) was isolated as a white amorphous solid. Its positive ion HRESIMS revealed a pseudomolecular ion peak at m/z 513.2960 $[M+H]^+$, corresponding to a molecular formula of $C_{29}H_{40}N_2O_6$ (calc'd for $C_{29}H_{41}N_2O_6$, 513.2965) with UV-Vis absorptions λ_{max} 326 and 222 nm. The similarity of its ^{13}C NMR spectrum and **4** suggested it an analog of **4**. The only difference lies in the change of the chemical shifts at C-5 and C-6 in addition to the absence of the methoxyl carbon signal. The above information and its molecular formula hinted a hydroxyl instead of a methoxyl group was substituted at C-5. Its planar structure was further confirmed by the COSY and HMBC spectra. The NOESY spectrum of **5** displayed similar correlations as **4**, establishing the α orientation of H-3, H-8, H-28, H-30 and the β orientation of H-4, H-7, H-25 and H-27. H-5 was concluded to be α orientation due to its correlations with H-4 and CH_3 -1, while the α orientation of H-6 was deduced from its correlation with H-8, H-28 and H-9 α . Derived from same precursor as **4**, H-17 of capsimycin F was also believed to adapt β orientation. Therefore, the structure of **5** was determined as depicted in **Figure 2.11** (**Table 2.5**).

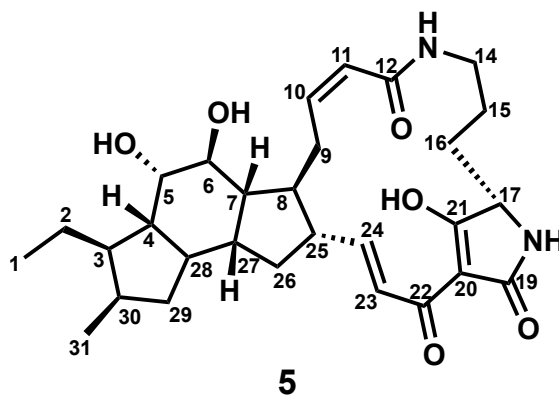


Figure 2.11 Structure of ikarugamycin analog, capsimycin C.

Table 2.5 ¹H (600 MHz) and ¹³C (100 MHz) spectroscopic data of **5**.

Position	δ H, Mult (J in Hz)	δ C, Mult	Position	δ H, Mult (J in Hz)	δ C, Mult
1	0.96 (t, J = 7.3)	12.4	17	3.85, m	61.7
2	1.36 m	21.7	NH-18	-	-
3	1.78 m	42.9	19	-	175.9
4	1.44 m	45.5	20	-	101.1
5	3.85 m	71.6	21	-	197.3
6	3.87 m	73.3	22	-	173.0
7	1.99 m	47.9	23	7.13 (d, J = 15.4)	122.4
8	2.08 m	44.5	24	6.79 (dd, J = 15.4, 10.3)	152.3
9	3.47 m, 2.53 m	26.8	25	2.39 m	50.2
10	6.09 (td, J = 1 1.3, 3.9)	142.0	26	2.16 m, 1.23 m	35.8
11	5.84 (d, J = 11.3)	123.5	27	2.02 m	42.7
12	-	168.0	28	1.60 m	41.2
NH-13	-	-	29	2.18 m, 0.70 m	39.3
14	3.41 m, 2.67 m	39.0	30	2.21 m	33.1
15	1.48 m, 1.19 m	21.1	31	0.91 (d, J=6.6)	17.0
16	1.99 m, 1.86 m	27.4			

Measured in CD₃OD. δ values given in ppm.

2.2.6 Xlamenemycin C & SS8201 D

Xlamenemycin C (**6**) was obtained as a white amorphous solid with UV-Vis absorptions λ_{max} 251 and 324 nm. Its positive HRESIMS revealed a pseudomolecular ion peak at m/z : 575.2532 [M-H]⁻, corresponding to a molecular formula of C₃₀H₄₀ClN₂O₇, (calc'd for C₃₀H₃₉ClN₂O₇ 575.2524). The UV spectrum and the similarity of ¹³C NMR data of **6** with **4** and **5** indicated it was another ikarugamycin analog.

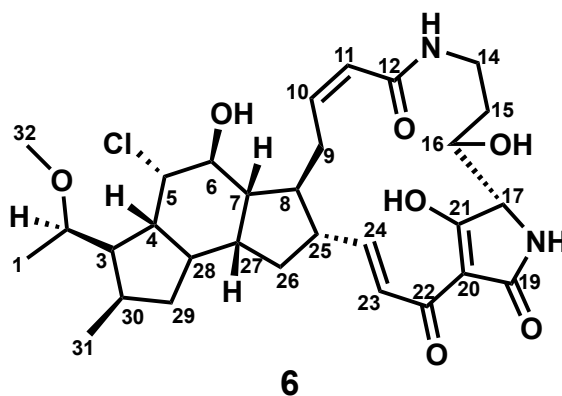


Figure 2.12 Structure of ikarugamycin analog, xlamenemycin C.

The 5,6,5 tricyclic ring skeleton of **6** was able to be identified and connected with a spin system (from H₂-14, H₂-15 and H-16 to H-17) through an amide functionality by the analysis of COSY and HMBC signals. However, chemical shift of H-16 at 3.97 ppm ascribed it an oxygenated methine. Furthermore, we saw one methoxyl group showing strong HMBC correlation to C-2 (δ_{C} 77.5). The above functionalities of **6** resembled with another PTM SS8201D (**7**) (**Figure 2.13**, **Table 2.7**).¹⁰ The carbon chemical shift difference at C-5 and C-6 (shift from 58.7 and 54.6 to 65.0 and 73.4) and the molecular formula C₃₀H₃₉ClN₂O₇ ascribed Cl and OH at C-5 and C-6, respectively. The planar structure of **6** was further confirmed by the analysis of 2D spectra.

The stereochemistry of **6** was established on the analysis of NOESY spectra. The correlations from H-28 to H-6, H-26 β , H-30 and H-29 β set up their same orientation as corresponding protons in **7**. Meanwhile, clear signals arisen from the correlations between CH₃-1, H-4 and H-5 assigned α orientation of H-5, H-4 and CH₃-1. β -hydroxyl L-ornithine has been

reported to be incorporated in several PTMs such as alteramide A, cylindramide and clifednamide B. Using the same strategy for determining the stereochemistry at H-17 of clifednamide B,¹¹ we calculated the dihedral angles of H-16-C16-C17-H17 as 87° ($J < 0.5$ Hz) of β orientation versus 20° ($J > 7$ Hz) of α orientation of H-17 (Chem3D-Ultra 12.0, MM2). The fact of the broad singlet at H-17 indicated the β orientation of the OH group. Therefore, the structure of xlamenemycin C is as shown in **Figure 2.12** (Table 2.6).

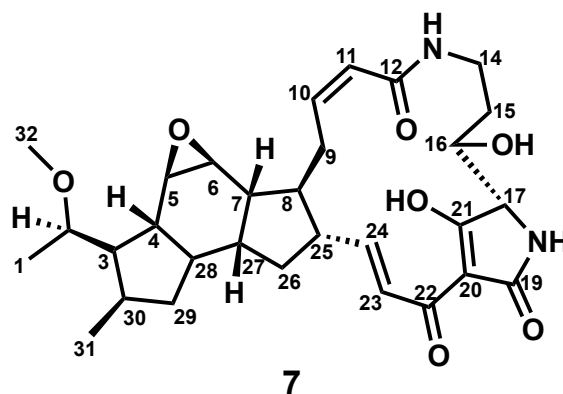


Figure 2.13 Structure of ikarugamycin analog, SS8201 D.

Table 2.6 ¹H (600 MHz) and ¹³C (100 MHz) spectroscopic data of **6**.

Position	δ H, Mult (J in Hz)	δ C, Mult	Position	δ H, Mult (J in Hz)	δ C, Mult
1	1.21 (d, J=6.2)	17.4	17	3.99, m	68.6
2	3.49 m	77.5	NH-18	-	-
3	2.14 m	47.0	19	-	176.4
4	2.05 m	44.2	20	-	100.4
5	4.32 (t, J=2.9)	65.0	21	-	194.0
6	4.11 s	73.4	22	-	174.4
7	2.10 m	47.3	23	7.09 (d, J=15.4)	122.4
8	2.15 m	45.9	24	6.79 (dd, J=15.4, 10.3)	151.9
9	3.63 m, 2.41 m	26.0	25	2.36 m	49.9
10	6.09 (td, J = 11.2, 2.9)	140.8	26	2.10 m, 1.31 m	35.4
11	5.84 (d, J = 11.7)	123.2	27	2.05 m	41.3
12	-	167.8	28	2.09 m	42.2
NH-13	-	-	29	2.10 m; 0.79 (dd, J=20.6, 11.3)	39.5
14	3.44 m, 2.83 (t, J = 11.8)	37.2	30	2.24 m	34.1

15	1.53 m, 1.40 m	31.4	31	1.04 (d, J=7.1)	16.5
16	3.97 m	71.5	32	3.31 s (overlap)	54.5

Measured in CD₃OD. δ values given in ppm.

Table 2.7 ¹H (600 MHz) and ¹³C (100 MHz) spectroscopic data of 7.

Position	δ H, Mult (J in Hz)	δ C, Mult	Position	δ H, Mult (J in Hz)	δ C, Mult
1	1.30 (d, J = 6.2)	17.2	17	3.74, m	66.4
2	3.43 m	78.1	NH-18	-	-
3	1.96 (dd, J = 20.7, 9.9)	50.4	19	-	178.3
4	0.96 m	47.5	20	-	102.5
5	3.31 m (overlap)	58.7	21	-	194.7
6	2.98 (d, J = 3.9)	54.6	22	-	184.5
7	2.38 m	40.3	23	7.40 (d, J = 15.4)	131.6
8	1.64 m	46.6	24	6.32 (dd, J = 15.3, 9.6)	143.0
9	3.57 m, 2,36 m	25.8	25	2.34 m	47.4
10	6.14 (td, J = 11.3, 2.4)	140.1	26	2.08 m, 1.18 m	37.6
11	5.90 (d, J = 11.6)	124.1	27	1.72 m	41.6
12	-	167.9	28	1.22 m,	47.5
NH-13	-	-	29	2.06 m, 0.68 (dd, J = 20.8, 11.9)	39.5
14	3.55 m, 2.84 (t, J = 12.3)	36.7	30	2.36 m	34.2
15	1.64 m, 1.36 m	31.3	31	1.01 (d, J = 7.1)	16.8
16	3.94 m	71.7	32	3.35 s	54.6

Measured in CD₃OD. δ values given in ppm.

2.2.7 Capsimycin E

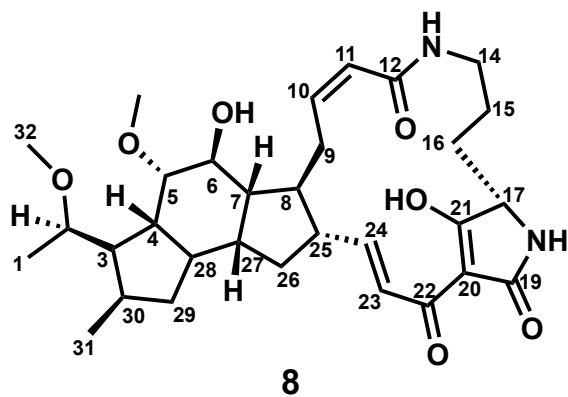


Figure 2.14 Structure of Ikarugamycin analog, capsimycin E.

Table 2.8 ¹H (800 MHz) and ¹³C (800 MHz) spectroscopic data of **8**.

Position	δ H, Mult (J in Hz)	δ C, Mult	Position	δ H, Mult (J in Hz)	δ C, Mult
1	1.22 (d, J = 6.2)	17.8	17	3.82, br d (4.0)	62.8
2	3.40 m	79.1	NH-18	-	-
3	2.05 m	46.1	19	-	176.3
4	1.68 dd (11.2, 2.5)	44.5	20	-	102.3
5	3.44 m (overlap)	83.1	21	-	197.8
6	4.1 br	68.7	22	-	174.5
7	2.02 m	49.3	23	7.23 (d, J = 15.4)	124.6
8	2.01 m	45.9	24	6.71 (dd, J = 15.3, 9.6)	152.9
9	2.46d (15.4), 3.48 m	27.4	25	2.34 m	49.4
10	6.1 (td, J = 11.3, 3.5)	142.6	26	2.08 m, 1.25 m	36.6
11	5.85 (d, J = 11.4)	124.5	27	1.92 m	42.6
12	-	168.9	28	1.54 m	43.1
NH-13	-	-	29	2.04 m, 0.68 m	40.5
14	2.64 br t (11.1), 3.40 m	38.9	30	2.16 m	34.8
15	1.16 m, 1.52 m	22.3	31	1.01 (d, J = 7.1)	17.8
16	1.82 m, 1.98 m	28.3	32/33	3.29 s, 3.41 s	55.6, 58.4

2.3 Biological Characterization of Ikarugamycin

The original isolation of ikarugamycin (**1**) in 1972 keenly identified it as a toxin against the protozoa *tertrahymena pyriformis*, exhibiting an excellent potency of 1.0 μ g/mL.⁴⁵ However, ikarugamycin quickly lost clinical relevance due to the group reporting intraperitoneal and intravenous LD₅₀ values of 6 mg/kg and 2 mg/kg, respectively, in mice. Ikarugamycin has also been reported to be cytotoxic (IC₅₀ 21.3 nM) in HL-60 leukemia cells by inducing apoptosis.⁵² Posecu et al. confirmed apoptotic cell death by immunoblotting to observe the cleavage of caspase-9, -8, and -3. Although the group did not directly identify a molecular target of **1**, they presented evidence that ikarugamycin induces DNA damage, increased intracellular, and activates p38 MAP kinase⁵² – providing key insights to the mechanism of action. The initial biological characterizations of ikarugamycin, in the context of protozoa and HL-60 cells, coupled with our NSCLC screening results allowed us to formulate a sound hypothesis for the

thorough characterization of **1** and how its structure gives rise to the selective toxicity observed in the context of NSCLC.

2.3.1 Cytotoxicity of Ikarugamycin

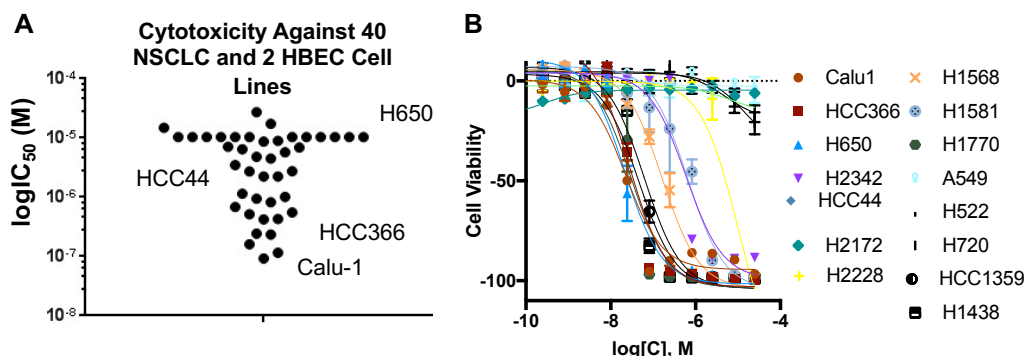


Figure 2.15 A. Volcano plot of ikarugamycin (**2.5**) against 40 NSCLC cell lines and 2 HBEC cell lines. **B.** Overlaid cell viability plots of ikarugamycin against a representative set of NSCLC cell lines. Both plots illuminate the selective toxicity trend of **2.5** and its ability to discriminate against sensitive and resistant cell lines. Determined by Celltiter Glo™ in triplicate.

Screening **2.5** against a panel of 40 NSCLC and 2 HBEC cell lines allowed us to observe a stark cellular subtype-selective cytotoxicity pattern. This is best visualized by the volcano plot (**Figure 2.15 A**) where the experimental IC_{50} values are observed to have a clear cut-off for certain NSCLC and HBEC cell lines at the 10 μ M region. We characterized those cells with IC_{50} values at, or above, 10 μ M as being resistant to toxin **1**. The overlaid cell viability plots (**Figure 2.15 B**) confirmed the trends observed in the volcano plot and allowed us to categorize the NSCLC cell lines that exhibited an IC_{50} value below 2 μ M as sensitive cell lines; those with an IC_{50} value between 2 μ M and 10 μ M were categorized with ‘intermediate’ sensitivity. Parsing out the cells that are sensitive to **1** allows for us to interrogate what biological mechanisms, or lack thereof, illicit the sensitivity. Identifying the biological make-up of ikarugamycin sensitivity could prove to be critical for achieving patient-matched NSCLC interventions down the road.

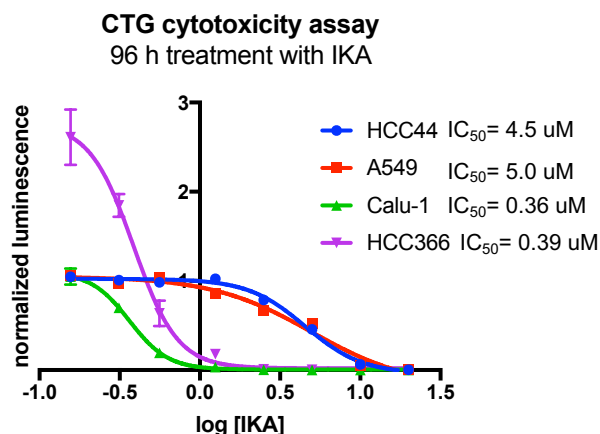


Figure 2.16 IC₅₀ curve of representative NSCLC cell lines against ikarugamycin (1).

We decided to develop a focused group of NSCLC cell lines that would be representative of those that we identified as resistant, intermediate, and sensitive to ikarugamycin treatment. The more focused screen resulted in a shift in our categorial parameters. We moved towards the use of resistant cell lines HCC44 and A549 which revealed ikarugamycin to possess a half maximal inhibitory concentration (IC₅₀) of 4.5 μM and 5.0 μM, respectively. We utilized the sensitive cell lines Calu-1 and HCC366 which identified ikarugamycin to illicit an IC₅₀ of 0.36 μM and 0.39 μM, respectively (**Figure 2.16**).

2.3.2 Cytotoxicity of Ikarugamycin Analogs

In order to probe the structure-activity-relationship (SAR) of ikarugamycin, we sought out to isolate any naturally occurring analogs. We hypothesized that the isolated analogs would provide enough structural diversity to direct a more focused semi-synthetic derivatization in our quest to characterize the minimum pharmacophore of ikarugamycin. The culmination of our ikarugamycin analog isolation efforts is listed in **Table 2.9** and the associated compounds are structurally characterized in chapter 2.2. Compounds **8** through **11** were not structurally characterized as part of this thesis, but their structures can be found in **Figure 2.17**.

Table 2.9 Anticancer activity of PTMs against representative NSCLC cells.

Compound	HCC366 (μM)	HCC44 (μM)	H650 (μM)	Calu-1 (μM)	HCC4017 (μM)
----------	----------------	---------------	--------------	----------------	-----------------

1	0.12	8.25	>24	0.09	0.47
2	>24	>24	>24	>24	-
3	>24	>24	>24	>24	3.8
4	>24	>24	>24	>24	>24
5	>24	>24	>24	>24	>24
6	-	>24	-	-	14.0
7	-	10.1	-	-	6.8
8	>24	>24	>24	>24	-
9	>24	>24	>24	>24	-
10	10.2	>24	>24	>24	-
11	0.17	3.2	>24	0.007	-

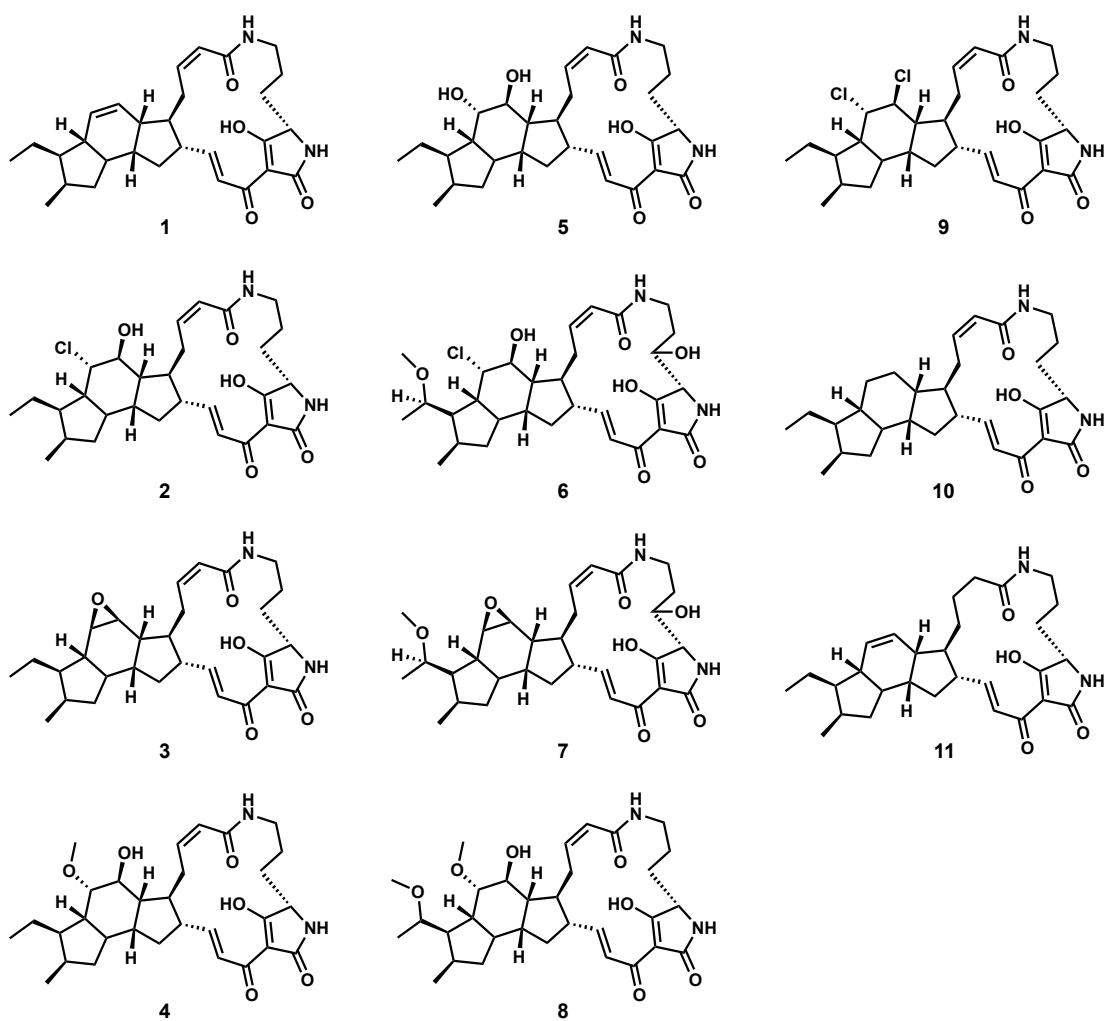


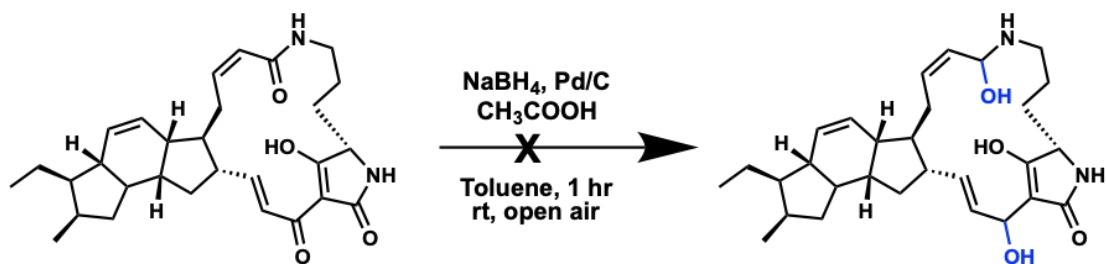
Figure 2.17 Collection of PTMs screened against NSCLC.

2.4 Semi-Synthetic Derivatization

Upon analysis of the PTMs isolated from SNB-040 and SNE-002, we concluded that the preservation of the alkene embedded within the 5-6-5 tricyclic system was critical to the biological activity of Ikarugamycin (**1**). Bearing that in mind, we sought to selectively derivatize **1** near commonly known reactive functionality such as the tetramic acid and the conjugated (α,β -Unsaturated) carbonyls while preserving the alkene embedded within the 5-6-5 tricyclic system. Those efforts are described below.

2.4.1 Selective Reduction of Conjugated (α,β -Unsaturated) Carbonyls

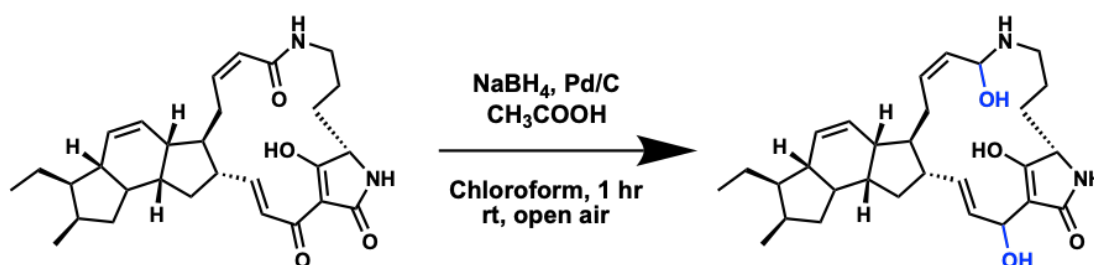
We prioritized the reduction of the α,β -unsaturated carbonyls in an effort to explicitly rule out any potential for the non-conjugated alkene to be implicated as the main pharmacophore. The methodology that we initially implemented utilized sodium borohydride in acetic acid for the deoxygenation of our α,β -unsaturated carbonyls.⁵³ The reaction was carried out in open air for 1 hour with 4 eq of NaBH₄, 2 eq of Acetic Acid, and 2.5 mol% of Palladium on carbon (**Scheme 2.1**).



Scheme 2.1 Reaction conditions for the selective reduction of α,β -unsaturated carbonyls in **1**.

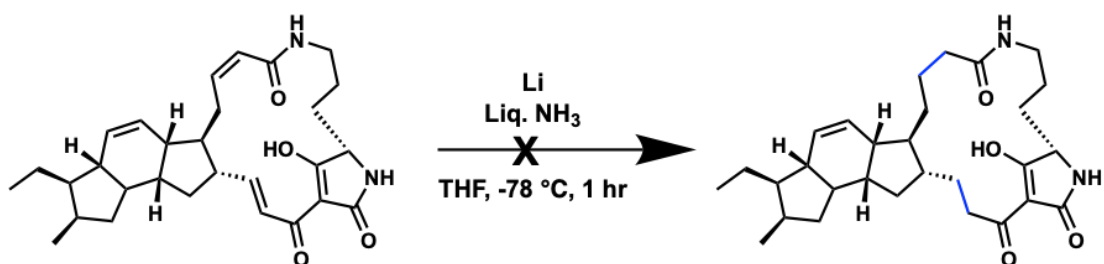
The reaction did not yield the expected results, but I observed a small amount of starting material (**1**) remaining in solution. Given that I did not allow the reaction to resume for very long, I questioned whether **1** was going into solution. The original isolation paper for **1** highlights the lack of solubility across traditional organic solvents, thus presenting a challenge

for the given methodology. Having run several ^1H NMR experiments on **1** utilizing CDCl_3 , I decided to substitute toluene with chloroform and implement a fresh batch of Pd/C (**Scheme 2.2**). According to LCMS signatures, I observed several reduction events along with a major amount of remaining starting material. I decided to set **Scheme 2.2** up again, but with a longer reaction time of 3 hours. This resulted in a cleaner conversion of **1** to a reaction product correlated with pseudo-mass signatures of 489 $[\text{M}+\text{H}]^+$ and 487 $[\text{M}-\text{H}]^-$. I was unable to isolate enough material from this small-scale reaction to characterize via ^1H NMR, therefore the proposed structure found in **Scheme 2.2** was from the observed mass signatures and the reactivity of NaBH_4 described in the literature.⁵³



Scheme 2.2 Altered reaction conditions of **Scheme 2.1**.

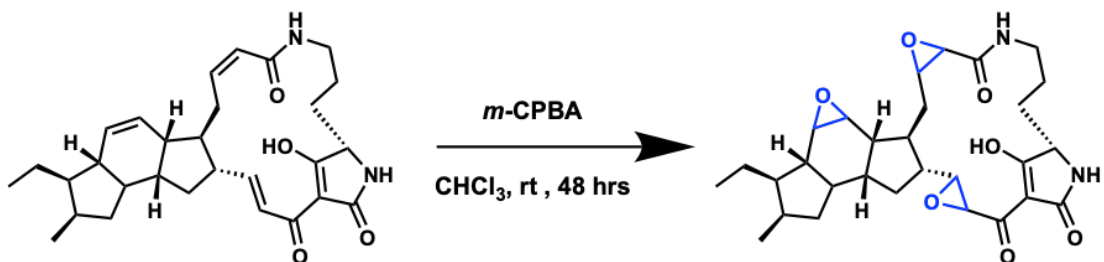
I later attempted to utilize the conjugated nature of the α,β -unsaturated carbonyls for selective reduction via a dissolving metal reduction. I decided to utilize the Birch reduction methodology⁵⁴ (**Scheme 2.3**) with the understanding that the delocalized π -system would selectively engage with the desired α,β -unsaturated carbonyls, as well as providing low-temperature reaction conditions that would help with the degradation of **1**. After allowing the reaction to run for one hour, I only observed starting material. This would indicate that I was unable to appropriately react lithium with liquid ammonia, therefore stunting the production of the unpaired electron required to initiate the reduction.



Scheme 2.3 Reaction conditions for the Birch reduction employed on **1**.

2.4.2 Epoxidation of Ikarugamycin (**1**)

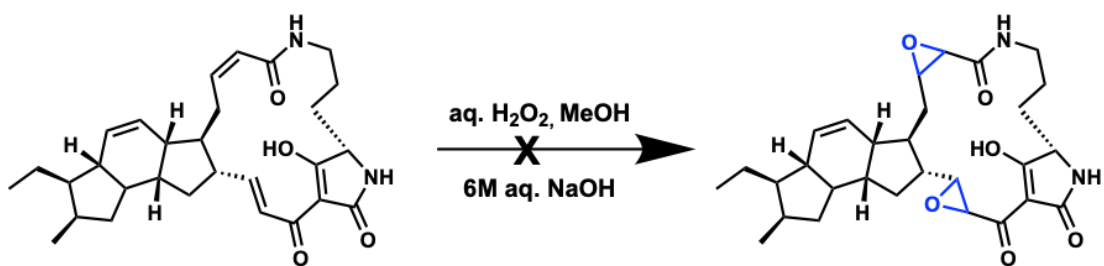
The cytotoxicity screening of **3** resulted in decreased activity against the representative NSCLC cell panel, but exhibited more bioactivity than the less rigid analogs of ikarugamycin. With this in mind, we decided to explore several epoxidation methodologies to gain a better insight of the tolerable functionality. Addition of 2 equivalents of *m*CPBA to **1** suspended in CHCl_3 over 48 hours (**Scheme 2.4**) yielded predominant pseudo-mass signatures of 527 $[\text{M}+\text{H}]^+$ and 525 $[\text{M}-\text{H}]^-$ that correspond to triple epoxidation event. An aliquot taken after 24 hours revealed a predominant pseudo-mass signature of 504 $[\text{M}-\text{H}]^-$ which corresponds to a unknown reaction intermediate. Isolation and purification of the 526 m/z species proved to be difficult due to its' lack of UV absorption and therefore was never fully characterized via ^1H NMR.



Scheme 2.4 Epoxidation of **1** utilizing *m*CPBA.

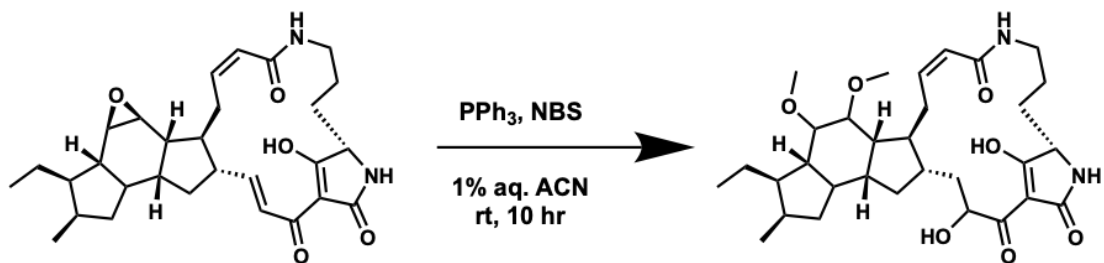
We next attempted to perform an epoxidation on the α,β -unsaturated carbonyls by implementing Juliá-Colonna (**Scheme 2.5**) epoxidation conditions.^{55,56} The idea was to preserve the non-conjugated alkene embedded within the 5-6-5 tricyclic ring system while

epoxidating the electron deficient olefins found in the α,β -unsaturated carbonyls. However, the reaction did not yield the desired products and the starting material was observed to have been consumed by the reactants.



Scheme 2.5 General reaction conditions for the Juliá-Colonna epoxidation of **1**.

Lastly, we decided to attempt a synthetic derivitization of one of the isolated analogs, compound **3**. The addition of 1.2 equivalents of triphenylphosphine and 1.2 equivalents of NBS to compound **3** suspended in 1% aqueous ACN yielded interesting results (**Scheme 2.6**). The expected outcome of a bromo-hydroxyl functionalization following the opening of the epoxide ring was not achieved based on the lack of brominated mass signatures. I did, however, observe a pseudo-mass signature of 559 $[M+H]^+$ and no remaining mass signatures of the starting material. The 559 m/z corresponds to an isomer of the compound **8**, which has the addition of two methoxy groups and one hydroxyl group.



Scheme 2.6 Reaction scheme of the epoxide ring opening of compound **3**.

2.5 Conclusions and Discussion

Due to their interesting biological properties,² many efforts have been done in the synthesis³ and biosynthesis⁴ of natural occurring tetramic acids. Among all tetramic acids from natural resources, polycyclic tetramate macrolactams (PTMs) are structurally complex compounds with a 5,5-dicyclic ring fusing with a 20-member tetramic acid lactam ring or a 5,5,6/5,5,6 tricyclic ring fusing with a 16-member tetramic acid lactam ring. Originally isolated as an antibiotic PTM, ikarugamycin was recently identified as a promising lead for anticancer therapeutics.¹² Other PTMs such as discoderamide,⁶ ikarugamycin oxide,⁹ cylindramide¹³ have also been reported to have anticancer activities against various cancer cells. Our biological evaluation of ikarugamycin, and its' corresponding analogs, provided great insight into the relationship between the structure of the natural product toxin and its ability to exhibit cytotoxicity across several representative NSCLC cell lines (**Table 2.8**). Although the structural diversity found in naturally occurring PTMs was limited to a few regions within the molecule, we were able to extrapolate two conclusions.

Compounds that contained functionality across the alkene found in the 5,6,5 ring system of ikarugamycin(**1**) exhibited a dramatic decrease in cytotoxicity. Given that 5,6,5 tricyclic system does not provide a reactive warhead, we presume that the alkene provides structural rigidity to ikarugamycin. The structural rigidity may play a role in allowing the compound to properly fit into a minor groove of the target protein and allowing the conjugated enone to react in a Michael Addition fashion. Some of the analogs containing epoxides (**3 & 7**) showed modest activity against sensitive cell lines – the incorporation of the epoxide adds some rigidity to the tricyclic system and could explain the slight improvement in cytotoxicity compared to the other analogs. To further support this conclusion, Yu et al. observed a similar cytotoxicity pattern where compounds **1** and **3** induced cell death in the pancreatic carcinoma cell line PANC-1 with IC₅₀ values of 1.30 μ M and 3.33 μ M, respectively. These results highlight the importance of maintaining structural rigidity within the tricycle region of the molecule during future semi-synthetic derivatization efforts.

Secondly, analogs containing an additional hydroxyl group adjacent to the embedded tetramic acid observed in compounds **6** and **7** exhibited negligible cytotoxicity decreases. These results provide insight to tolerable functionality and point derivatization efforts to focus on the exploration of structural diversity within the macrolactam region. Having observed this trend, I set out to install a propargyl group with the goal of conducting a 'click' pull-down assay to help us identify the target of ikarugamycin. The mass of the propargyl-IKA compound was confirmed via HRMS, but the position of the installation proved to be difficult due to the overlapping signals in the up field region of the NMR spectra. We believe that the propargyl resides near the tetramic acid. Unfortunately, the installation of the propargyl group dramatically decreased the cytotoxicity of ikarugamycin against the NSCLC cell lines H2122 and Calu-1 (**Figure 2.18**).

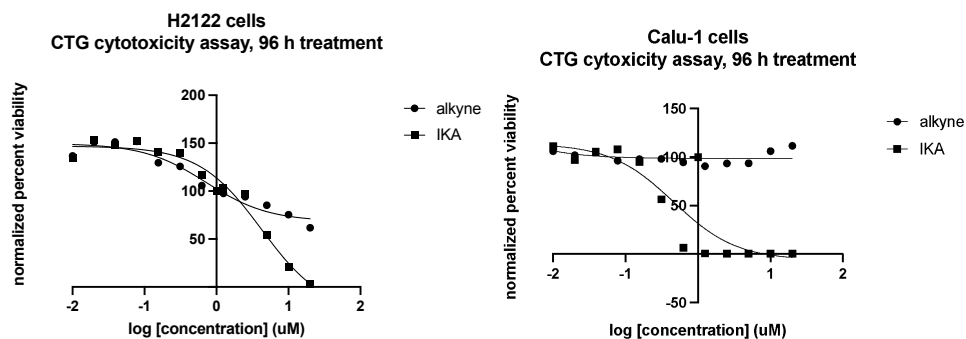


Figure 2.18 Cytotoxicity assay of ikarugamycin (**1**) and propargyl-IKA against representative NSCLC cell lines.

In conclusion, our preliminary screen for selective toxins against NSCLC allowed us to identify ikarugamycin (**1**) as a potent and selective agent for NSCLC. The chemical and biological characterization of ikarugamycin and its analogs has provided powerful insights on key the functionalities that give rise to the unique selective toxicity profile exhibited across a panel of NSCLC cell lines. Our studies have laid the foundation for medical chemists to begin semi-synthetic derivatization in order to tune the potency and potentially tune the selectivity of **1**. Utilizing **1** as a probe, we were also able to identify sensitive subtypes of NSCLC which are currently being investigated on the basis of target gene expression and TFEB localization.

2.6 Materials and Methods

2.6.1 General experimental procedures

UV spectra were recorded on a Shimadzu UV-1601 UV–Vis spectrophotometer. ^1H and ^2D NMR spectral data were recorded at 800 MHz in CD_3OD , CDCl_3 , $\text{CD}_3\text{OD}/\text{CDCl}_3$, or $\text{DMSO-}d_6$ on a Varian System or Bruker spectrometer, and chemical shifts were referenced to the corresponding solvent residual signal. ^{13}C NMR spectra were acquired at 100 MHz on a Varian System spectrometer and at 800MHz on a Bruker spectrometer. High resolution ESI-TOF mass spectra were collected on a ThermoFisher Orbitrap or provided by The Scripps Research Institute, La Jolla, CA. Low-resolution LC/ESI-MS data were measured using an Agilent 1200 series LC/MS system with a reversed-phase C18 column (Phenomenex Luna, 150 mm x 4.6 mm, 5 μm) at a flow rate of 0.7 mL/min. Preparative HPLC was performed on an Agilent 1200 series instrument with a DAD detector, using a C18 column (Phenomenex Luna, 250 x 10.0 mm, 5 μm). ODS (50 μm , Merck) were used for column chromatography.

2.6.2 Collection and Phylogenetic Analysis of Strain SNB-040

The marine-derived actinomycete, strain SNB-040, was isolated from a sediment sample collected from Sweetings Cay, Bahamas. Bacterial spores were collected via stepwise centrifugation as follows: 2 g of sediment was dried over 24 h in an incubator at 35 °C and the resulting sediment added to 10 mL sH_2O containing 0.05% Tween 20. After a vigorous vortex for 10 min, the sediment was centrifuged at 2500 rpm for 5 min (4 °C). The supernatant was removed and transferred into a new tube and centrifuged at 18,000 rpm for 25 min (4 °C) and the resulting spore pellet collected. The resuspended spore pellet (4 mL sH_2O) was plated on a humic acid media, giving rise to individual colonies of SNB-040 after two weeks. Analysis of the 16S rRNA sequence of SNB-040 revealed 99% identity to *Streptomyces carpaticus*.

2.6.3 Collection and Phylogenetic Analysis of Strain SNE-002

The marine-derived actinomycete, strain SNE-002, was isolated from a sediment sample collected from hypersaline lake at East Plana Cay, Bahamas (22°36'33"N, 73°33'24"W). Bacterial spores were collected via stepwise centrifugation as follows: 2 g of sediment was dried over 24 h in an incubator at 35 °C and the resulting sediment added to 10 mL sH₂O containing 0.05% Tween 20. After a vigorous vortex for 10 min, the sediment was centrifuged at 2500 rpm for 5 min (4 °C). The supernatant was removed and transferred into a new tube and centrifuged at 18,000 rpm for 25 min (4 °C) and the resulting spore pellet collected. The resuspended spore pellet (4 mL sH₂O) was plated on a humic acid media, giving rise to individual colonies of SNE-002 after two weeks. Analysis of the 16S rRNA sequence of SNB-040 revealed 99% identity to *Streptomyces xlamensis*.

2.6.4 Cultivation and Extraction of SNB-040 and SNE-002

Bacterium SNB-040 was cultured in 20 2.8 L Fernbach flasks each containing 1 L of a seawater based medium (10 g starch, 4 g yeast extract, 2 g peptone, 1 g CaCO₃, 40 mg Fe₂(SO₄)₃·4H₂O, 100 mg KBr) and shaken at 200 rpm at 27 °C. After seven days of cultivation, sterilized XAD-7-HP resin (20 g/L) was added to adsorb organic products, and the culture and resin were shaken at 200 rpm at 27 °C for 2 h. The resin was filtered through cheesecloth, washed with deionized water, and eluted with acetone. The acetone soluble fraction was dried *in vacuo* to yield 3.8 g of extract.

Bacterium SNE-002 was cultured in 20 2.8 L Fernbach flasks each containing 1 L of a seawater based medium (10 g starch, 4 g yeast extract, 2 g peptone, 1 g CaCO₃, 40 mg Fe₂(SO₄)₃·4H₂O, 100 mg KBr) and shaken at 200 rpm at 27 °C. After seven days of cultivation, sterilized XAD-7-HP resin (20 g/L) was added to adsorb organic products, and the culture and resin were shaken at 200 rpm at 27 °C for 2 h. The resin was filtered through cheesecloth,

washed with deionized water, and eluted with acetone. The acetone soluble fraction was dried *in vacuo* to yield 6.1 g of extract.

2.6.5 Extraction and Isolation.

The extract of either SNB-040 or SNE-002 was suspended in aqueous MeOH (MeOH-H₂O, 9:1, 100 mL) and extracted with hexanes (3 x 100 mL portions). The hexanes extract was evaporated in vacuo to leave 241 mg. The aqueous MeOH extract then subjected to reverse phase flash column chromatograph (C18) to yield 22 fractions with a step gradient of MeOH and H₂O (30:80-100:0). Fractions 12 through 22 were pooled together and purified using preparative-scale reversed-phase HPLC (Phenomenex Kinetex C18 Evo, 250 x 21.2 mm, 10 mL/min, 5 µm) with a linear gradient (10:90-100:0 ACN:H₂O) to yield 21 fractions. The later fractions were grouped (F5-F10, F11-F15, and F16-F20) and were then subject to further purification utilizing semi-preparative reversed-phase HPLC (Phenomenex Luna C5, 250 x 10.0 m, 2.5 mL/min, 5 µm) with a linear gradient (40:60-90:10 ACN:H₂O) until pure peaks were isolated.

2.6.6 Antiproliferative Bioassays.

Cytotoxicity assays Cell lines were cultured in 10 cm dishes (Corning, Inc.) in NSCLC cell-culture medium: RPMI/L-glutamine medium (Invitrogen, Inc.), 1000 U/ml penicillin (Invitrogen, Inc.), 1 mg/ml streptomycin (Invitrogen, Inc.), and 5% fetal bovine serum (Atlanta Biologicals, Inc.). Cell lines were grown in a humidified environment in the presence of 5% CO₂ at 37 °C. For cell viability assays, HCC366, A549, Calu-1, H650, HCC4017 and HCC44 cells (60 µL) were plated individually at a density of 750 and 500 cells/well, respectively, in 384 well microtiter assay plates (Bio-one; Greiner, Inc.). After incubating the assay plates overnight under the growth conditions described above, purified compounds were dissolved and diluted in DMSO and subsequently added to each plate with final compound concentrations ranging from 1 µM to 2 pM and a final DMSO concentration of 0.5%. After an incubation of 96 h under

growth conditions, Cell Titer Glo™ reagent (Promega, Inc.) was added to each well (10 µL of a 1:2 dilution in NSCLC culture medium) and mixed. Plates were incubated for 10 min at room temperature and luminescence was determined for each well using an EnVision multi-modal plate reader (Perkin-Elmer, Inc.). Relative luminescence units were normalized to the untreated control wells (cells plus DMSO only).

2.6.7 Natural Product Fraction Cytotoxicity Screening

Screening Protocol

The NSCLC cell line was cultured in 10 cm dishes (Corning, Inc.) in NSCLC culture medium (RPMI/L-glutamine medium (Invitrogen, Inc.), 1000 U/ml penicillin (Invitrogen, Inc.), 1 mg/ml streptomycin (Invitrogen, Inc.), and 5% fetal bovine serum (Atlanta Biologicals, Inc.) for primary screening. All cell lines were maintained in a humidified environment in the presence of 5% CO₂ at 37 °C. For cell viability assays, NSCLC cell lines (60 µL) were plated in 384-well microtiter assay plates (Bio-one; Greiner, Inc.) at a cell density that would allow for 70 – 80% confluency by the end of the incubation period (96 h). After incubating the assay plates overnight under the growth conditions described above, NPFs or pure compounds were added to each plate for confirmation studies (three replicates per compound per cell line) and at twelve half-log doses ranging from 50 µg/mL to 50 pg/mL for dose-response studies (3 replicates per dose per cell line). In all experiments, we maintained a final DMSO concentration of 0.5%. After an incubation of 96 hours under growth conditions, Cell Titer Glo™ reagent (Promega, Inc.) was added to each well (10 µL of a 1:2 dilution in NSCLC culture medium) and mixed. Plates were incubated for 10 minutes at room temperature and luminescence was determined for each well using an EnVision multi-label plate reader (Perkin-Elmer, Inc.). Assay with the HBEC30KT was assayed using a published protocol (Kim, Hyun S., et al., *Cell*, 2013. 155(3): 552-566). All NSCLC and HBEC30KT viability assays typically displayed Z' values greater than 0.5.

The 26 NSCLC Cell Lines

HBEC30KT, HCC366, H1993, H2009, H2122, HCC15, HCC827, H2073, HCC44, H2887, HCC193, H1819, HCC515, HCC95, H2250, H1437, HCC122, H2126, H2347, H1693, HCC4017, H2087, H2052, H1395, HCC4018 & H2882.

CHAPTER THREE
DIRECTED-MESSAGE PASSING NEURAL NETWORK APPLICATION TO THE NATURAL
PRODUCTS ATLAS

3.1 Challenges in Physical High Throughput Screening Strategies

3.1.1 Creating and Screening Natural Product Libraries

As mentioned in Chapter 1, natural product libraries typically come in two forms – crude extracts or semi-purified fractions. Conducting high throughput screens using these two mediums has provided many therapeutic leads over the years, but the method is not perfect. On one hand, screening crude extracts allows for drug discovery groups to quickly discern which organisms are producing compounds with the desired phenotypic read out and inform them which extract to push forward into purification. On the other hand, crude extracts contain complex mixtures of numerous chemical entities that may interfere with the assay and result in a false positive hit.⁵⁷ Pre-fractionated high throughput screens were designed with this issue in mind and provide a solution by iteratively reducing the number of compounds that are dosed into an assay. Furthermore, groups have coupled pre-fractionated screening with concentration dose gradients to aid in the identification of true hits.^{58,59} While these efforts have greatly assisted in our ability to parse out false positives and true positives, current high throughput natural products screening strategies leave out great detail regarding the relationship between the structures in question and their associated bioactivity.⁶⁰

Are natural products drug discovery efforts looking for a ‘hit’ or are they looking for a potential drug? The latter begs for thorough characterization between a given natural product and its’ bioactivity within the context of the screen, that of which is lost in current physical screening strategies. Ideally, the development of a pure natural product library would allow for accurate screening that produces rich chemoinformatic data that is critical for true drug discovery. However, the task of isolating and structurally characterizing each natural product sourced from an organism *prior* to conducting a high throughput screen would require extensive labor and countless resources.

3.1.2 High-throughput screening with Fraction Libraries and Pure Compound Libraries

The pure natural product libraries that do exist present accessibility challenges and are not conducive to collaborative drug discovery efforts. Although combinatorial chemistry enabled vast libraries of synthesized compounds for HTS, early libraries had low chemical diversity and so provided few authorized medications in the previous 25 years.⁶¹ Pure natural product libraries might considerably boost structural variety in chemical libraries, but the expenses involved with building them can be prohibitive (**Table 3.1**). This is owing to the resource-intensive stages needed with the purification and characterization of individual molecules in sufficient amount.⁶² So, to construct more chemically varied screening libraries, techniques like fragment-based drug discovery (FBDD) and diversity-oriented synthesis (DOS) were developed.⁶³ Despite this, only around 20% of the core ring scaffolds observed in natural products are represented in most commercially accessible synthetic collections or compound libraries. Ultimately, a complementary collection of varied source species should provide the scientific community access to structurally diverse pure chemicals to be used in HTS efforts for further examination as possible therapeutic leads

Table 3.1 Commercially and publicly available large natural product libraries.⁶⁰

Company/institute	Sample type (number)	Number of screening samples			Ref.
		Extracts (sample source #)	Fractions (extract source #)	Compounds (type)	
<i>Albany Molecular Research, Inc. (AMRI)</i>	B/F and P (>190 000)	102 000 (23 375)	209 000 (12 349)	—	144
<i>AnalytiCon Discovery</i>	B/F (na); P (na); SS (>25 000)	—	—	>25 000 (SS); >5000 (NP)	145
<i>Bioinformatics Institute Singapore (BII)—A*STAR Natural Product Library</i>	B/F (>120 000); P (>37 000)	~270 000 (>157 000)	~70 000 (na)	2600 (NP)	146
<i>Developmental Therapeutics Program—The National Cancer Institute</i>	MI (>20 000); B/F (>25 000); P (>80 000)	>230 000 (>108 000)	326 000 (46 570)	419 (NP set IV) ^d	147
<i>Fondazione Ricerca per la Vita (FIIRV)</i>	B/F (>15 000)	166 000 (15 000)	—	—	148
<i>Fundación MEDINA</i>	B/F (190 000)	>130 000 (na)	—	—	149
<i>Griffith Institute for Drug Discovery (GRIDD)—Nature Bank</i>	MI and P (30 000)	10 000 (10 000)	50 000 (10 000)	—	150
<i>InterBioScreen (IBS)</i>	MI (na); B/F (na) and P (na)	—	—	>67 000 (NP)	151
<i>Magellan BioScience Group, Inc.</i>	MB (10 000); F (55 000)	>15 000 (na)	—	—	152
<i>Mycosynthetix</i>	F (>55 000)	55 000 (na)	—	—	153
<i>Natural Products Discovery Institute (NPDI)</i>	B/F (>30 000); P (>20 000)	80 000 (na)	—	—	154
<i>PharmaMar</i>	MI (>118 000); MB (>100 000)	100 000 (na)	—	—	155
<i>PhytoPharmacon</i>	P (4000)	4000 (4000)	25 000 (4000)	500 (NP)	156
<i>RIKEN Natural Products Repository (NPDepo)</i>	B/F (na) and P (na)	—	—	8000 (NP)	157
<i>The Institut de Chimie des Substances Naturelles (ICSN)</i>	MI, B/F, and P (>7000)	14 000 (7000)	—	—	158
<i>The Natural Products Library Initiative at the Scripps Research Institute (Florida)</i>	B (>5500)	8500 (na)	3400 (na)	450 (NP)	159
<i>The University of Mississippi—National Center for Natural Products Research</i>	MI and F (>2000); P (>18 000)	>20 000 (>20 000)	>43 000 (>3400)	~700 (NP)	160
<i>Unigen (PhytoLogix Library)</i>	P (8000)	9000 (na)	200 000 (na)	—	161

B= bacteria; F= fungi; MB= marine bacteria; MI= marine invertebrates; P= plant; NP= pure natural products; SS= semi-synthetic (NP-based); na= data not available.

3.1.3 Role of Virtual Compound Libraries for Large Scale Screens

Virtual compound libraries provide a rich alternative to physical compound libraries by increasing accessibility and allowing for a broad scope of high throughput *in silico* screening. Virtual screening is a computer approach that is used to scan through libraries of small molecules in order to determine the structures that are most likely to bind to the target of a pharmaceutically relevant target.⁶⁴ Because of its distinct benefits over experimental high-throughput screening (HTS), it has emerged as a critical step in early-stage drug development. These advantages include being drug target-relevant, being competitively priced, and being efficient.

In this context, the virtual screening of virtual compound libraries is a promising *in silico* method using in the drug discovery process. Traditionally, the availability of 3D structures of target proteins directed our ability perform *in silico* screening of virtual compound libraries. Within the context of screening natural products, an indispensable condition in performing virtual screening is the availability of appropriate virtual natural product compound libraries.⁶⁵ The ideal virtual natural product library to accelerate broad *in silico* high throughput screening would provide comprehensive structural characterization for each entry in an open-access format that allows for easily downloadable content.

3.2 Virtual Natural Products Libraries

3.2.1 Overview of Virtual Natural Product Databases

Virtual natural products databases are as close as we can come to a large (>20,000) SS library of pure natural product compounds. The assembly of, and access to, these databases are critical for virtual screening to identify potential NP-based drugs or key functionality that elicits a desired bioactivity. A major issue in the assembly of virtual NP databases stems from the publication of structures exclusively in pictorial format, such as in the annual reviews of Marine Natural Products. This makes it difficult to retrieve the compounds to be computationally analyzed and subsequently integrated into a central molecular

database.⁶⁶ The first stage in any exploratory molecular analyses and, to some extent, in the discovery of NP-based drugs or other active components, is virtual screening. For example, virtual screening of known NPs can save time on sample extraction and purification, delaying the wet lab phase until the theoretical selection of optimal candidates. Using contemporary cheminformatics technologies thereby streamlines research, saves time and money, and improves findings.

The current landscape of virtual natural product databases is plenty, each of which offering different content and providing various analysis tools. In a broad review of 123 resources listing NP structures, Sorokina et. al. found that 92 were open-source and only 50 of those contained molecular structures that could be retrieved for further analysis. Sorokina et. al. compared all of the selected NP databases on the following: NP source organism type, estimated size (number of NP molecules with correct structures), number of unique molecules in COCONUT, percentage of molecules with stereochemistry, can the data be freely browsed (open-source), registration requirements, and if they were actively maintained or updated (**Table 3.2**, which can be found in the Appendix).^{67,68} Their review keenly identified that 40 of the 50 open-source databases also shared a significant overlap of at least 50% of compounds with at least one other dataset. This analysis proved to be crucial to choosing the appropriate virtual NP database to be used as a compound library for virtual screening. I elected to use the Natural Products Atlas because of its' strong structural foundation and selection for microbial natural products, that of which have become increasingly important to modern natural product drug discovery efforts.

3.2.2 The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery

Although microbial natural products only make up 11% of FDA approved natural product drugs, the field of natural products has shifted its' focus towards microbial sources due to their high propensity for bioactivity and rising interests in biosynthetic gene cluster

manipulation. Given the substantial fraction of antibiotics sourced from microbial organisms, it seems clear that utilizing a virtual natural product database would yield rich results for virtual screening efforts. The Natural Products Atlas provides a concise repository of microbial natural products with easily accessible structures that can be applied to various *in silico* screening methodologies.⁶⁹

The Natural Products Atlas addresses the existing, yet fragmented, database landscape for microbial natural products. Out of the currently available databases, some are only accessible through commercial means and do not provide comprehensive characterization for each entry (i.e., AntiBase, Dictionary of Natural Products, MarinLit). The databases that are open access are often difficult to download (i.e., NPEdia) or cater to a niche within the field (i.e., AfroDB, NuBBEDB, StreptomeDB).^{70–75} More than 29,000 microbially-derived natural products are included in the NP Atlas, which is designed to cover all microbially-derived natural products published in peer-reviewed primary scientific literature. The database contains taxonomic information to inquire about the distribution of compounds originating from different species. Researchers can use the search, explore, and discover aspects of the NP Atlas to find new information. Compounds can be found by utilizing a structure drawing tool or by searching for their structure, name, molecular weight, chemical formula, InChiKey, SMILES, etc. Compounds with structural similarities can be clustered and nodes analyzed using NP Atlas's explore feature, while those with different structures can be interrogated using the find feature.

3.3 *In silico* Screening of Natural Products

3.3.1 Overview

Unprecedented opportunities for rationalizing drug discovery are opening up because of advances in our knowledge of protein–ligand interactions, as well as the rapidly expanding number of 3D-structures of potentially useful and empirically proven ligands. In order to benefit from previously released knowledge, the human brain is being asked to do actions that are just

unrealistic. Thus, today's understanding of NPs may be used to develop more efficient and effective methods. Computer simulations like virtual screening studies have previously satisfied these requirements in the field of pharmaceutical chemistry.⁷⁶ They are required to make use of the available structural information, to grasp specific molecular recognition events, and to elucidate the function of the target macromolecule. The 'needles in a haystack' of bioactive natural compounds can be found using rationalized approaches, but computational approaches have only recently made their way into the field of natural product drug discovery research as auxiliary screening methods.⁷⁷ Virtual screening methods have innovated the discovery of new compounds with specific bioactivity, assessing *in silico* large structural libraries against a bioreceptor or biological in parallel to physical screening efforts favor the reduction of financial efforts, streamline discovery infrastructure, and reduce the time required to develop drug-target-disease associations.

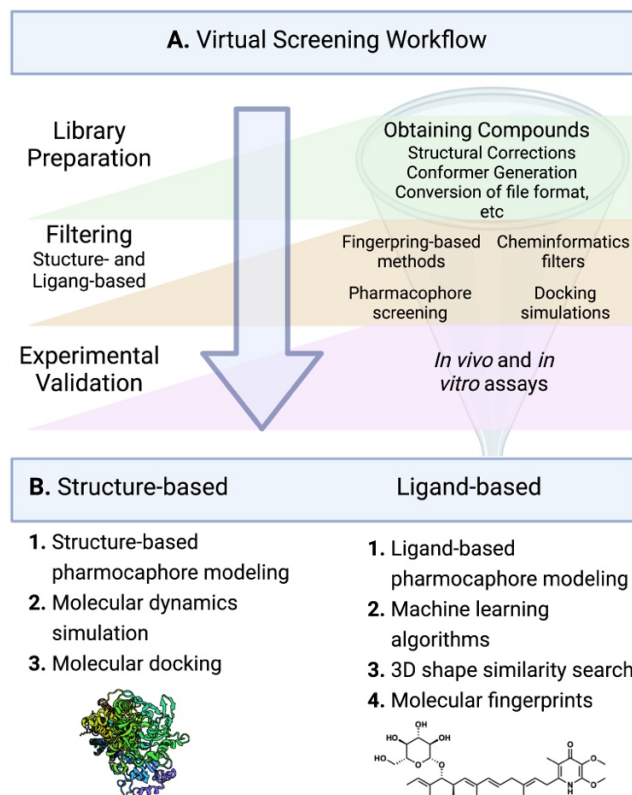


Figure 3.1 **A** Sequential steps applied in virtual screening workflows to identify bioactive natural products. **B** Ligand- and structure-based virtual screening approaches and some of their associated computational methods.⁷⁸

In silico methods involve sequential and hierarchical steps that aim at filtering and selecting compounds with desirable physiochemical, pharmacokinetic, and pharmacodynamic properties. All virtual screening workflows begin with library preparation, which involves obtaining structures of the compounds (developing a novel library or utilizing an available database such as the ones in **Table 3.2**), converting them into a readable format (SMILES, SDF, MOL₂, etc.), and scrubbing the structures for stereochemical correction or valence errors. Following library preparation, virtual screening workflows involve the selection and application of computational methods (docking simulations, chemoinformatics, QSAR, etc.) that are appropriate for the established hypothesis. The final step corresponds to experimental validation using *in vitro* and *in vivo* assays, such as enzymatic inhibition and cell line cytotoxicity assays.

3.3.2 Chemoinformatic Models for Molecular Property Prediction

Predicting molecular properties is critical for rapidly developing new drugs. Chemists can quickly filter through vast libraries of molecules using accurate property prediction models. Some techniques, like chemoinformatics, quantitative structure activity relationship (QSAR), docking, molecular similarity, network pharmacology, and pharmacogenomic computational *de novo* design, to name a few, have been found to cut the cost of drug development by as much as 50%.⁷⁹ In recent years, machine learning has come to play an increasingly important role in predicting molecular properties.⁸⁰⁻⁸² Machine learning offers a quick, low-cost, and accurate framework for developing a property prediction model.

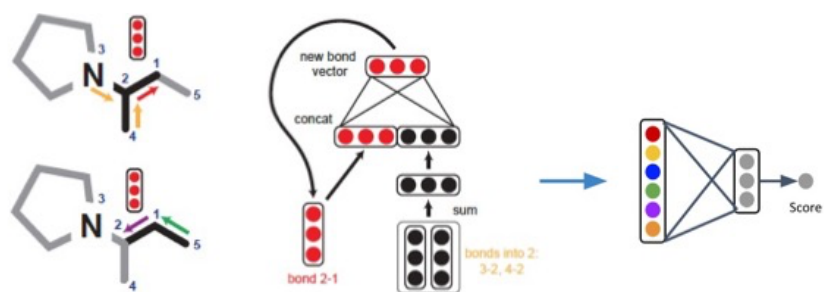


Figure 3.2 Representation of a message passing neural network (MPNN) that iteratively aggregates local chemical features for molecular property prediction.

Many current approaches to molecular property prediction, such as Dragon descriptors or Morgan (ECFP) fingerprints,⁸³ rely on standard machine learning models. Improvements in property prediction accuracy using these models has primarily come through the development of better molecular descriptors. Other approaches have explored the use of 3D atomic coordinates to augment the information provided by these models. Instead, a different line of study has focused on developing more robust models. Descriptors and fingerprints have been the subject of some research, but SMILES strings and the molecular graph have also been employed as input.^{84–87} Graph convolutional neural networks and message passing neural networks (**Figure 3.2**) are examples of the latter method.⁸⁸ These models work directly with the molecular graph's atoms and bonds, allowing them to generate a molecular representation that is more relevant to the feature or qualities that are being studied.

3.3.4 Directed-Message Passing Neural Network Development

An MPNN is a form of neural network model that is especially built to function on graphs, as opposed to other types of networks. When an MPNN is run, it receives as input an undirected graph G that has node characteristics x_v and bond features e_{vw} . Atoms are the nodes of the network, while bonds are the edges. In chemistry, the graph represents a molecule. Following the incorporation of this featurization as an input, the MPNN functions in two stages.⁸⁸ In the beginning, there is a message passing phase, in which information is sent throughout the graph in order to construct a neural representation of the entire graph. The second step is

the readout phase, during which the neural representation of the graph is utilized to create predictions based on the data. There are T stages in the information propagation process during the message passing phase. Vertex v is connected with a set of hidden h_v^t states and messages m_v^t , which are updated on each time step t by calling a message function M_t and a vertex updated function U_t according to:

$$\begin{aligned} m_v^{t+1} &= \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \\ h_v^{t+1} &= U_t(h_v^t, m_v^{t+1}) \end{aligned}$$

where $N(v)$ is the set of neighbors of v in graph G , and h_v^0 is some function of the initial atom features x_v .

During the readout phase, a readout function R is applied to the collection of final concealed states h_v^T in order to produce the following prediction:

$$\hat{y} = R(\{h_v^T | v \in G\})$$

The readout function works by first building a single representation h of the whole graph by summing the final hidden states:

$$h = \sum_{v \in G} h_v^T$$

Then, a feed-forward neural network f is applied to h to produce:

$$\hat{y} = f(h)$$

It is necessary to train the MPNN from beginning to finish, training each feature vector and ending with the final representation graph, with backpropagation of the gradient of loss occurring throughout both the reading and message passing phases. By adjusting the loss function, the MPNN may be trained in either a regression or a classification context, depending on the situation.^{88,89}

The fundamental distinction between D-MPNNs and normal MPNNs is the type of the messages that are conveyed across the molecule during the message passing phase. Unlike the conventional MPNN framework, which assumes that messages are focused on atoms, the

D-MPNN framework centers messages on bonds. The D-MPNN, in particular, keeps two representations of the message focused on the connection between atoms v and w : one from atom v to atom w (m_{vw}^t) and one from atom w to atom v (m_{wv}^t). As a result, rather than aggregating data from nearby atoms, the D-MPNN gathers data from neighboring bonds. The message of each bond is modified depending on all incoming bond messages m_{kv}^t where $k \in \{N(v) \setminus w\}$. The D-MPNN has better control over the flow of information throughout the molecule as a result of this structure, with messages concentrated on bonds and a differentiation between the two orientations of bond messages and can thus generate more informative molecular representations.^{88–90}

3.4 Proof of Principle Application to the Natural Products Atlas

3.4.1 Rise in Antibiotic Resistance

Natural products have provided a significant foundation for the development of antibiotic drugs in the 90 years since the discovery of penicillin (**1**).⁹¹ The use of natural products to create new molecular entities for almost every disease is also well established.⁹² Six of the nine antibiotic classes depicted in **Figure 3.3** are naturally occurring compounds, with the remaining three (sulfonamides, fluoroquinolones, and oxazolidinones) created entirely through synthetic chemistry.^{93–101} The structural diversity and complexity of natural product antibiotics tend to offer unique mechanisms of action with selective target interaction, especially when compared to synthetic classes.

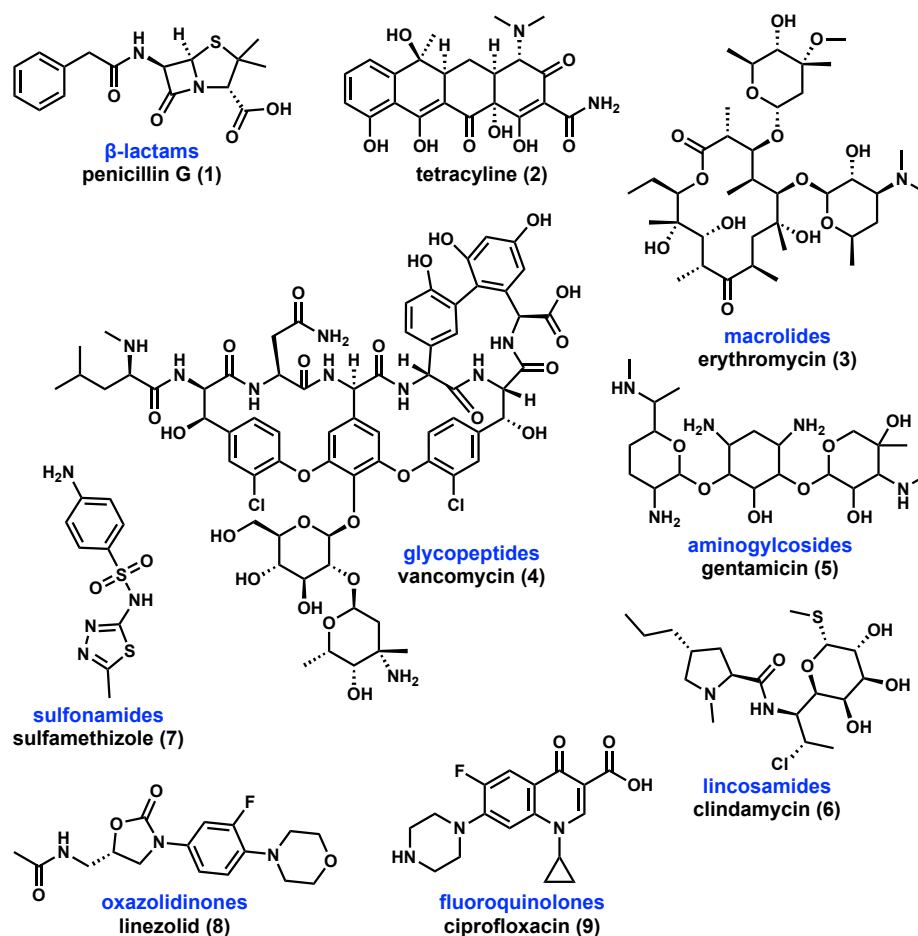


Figure 3.3 Representative classes of antibiotics of the modern era, excluding the arsenic-containing antibiotics of the early twentieth century.

Antibiotic resistance has spread as a result of the widespread use of antibiotics to treat nonbacterial infections.¹⁰² The United States is likely to use newly developed antibiotics quickly, potentially shortening their efficacy lifetime. In the United States and the United Kingdom, nearly 40–60% of hospital-acquired *S. aureus* strains are methicillin-resistant.³ It is estimated that the livestock industry in the United States consumes an astounding 80 percent of all antibiotics produced.¹⁰⁴ Antibiotic resistance was responsible for at least two million illnesses and 23,000 deaths in the United States as of 2013.¹⁰⁵ With nearly 450,000 cases of drug-resistant tuberculosis, antibiotic resistance is also a global issue.¹⁰⁶ Since the implementation of HTS in the 1980s, no new chemical antibiotics have been discovered using this method.¹⁰⁷

The rise of global antibiotic resistance and subsequent shortcomings of HTS efforts illuminate the necessity for novel approaches to antibiotic discovery to increase the rate at which new antibiotics are identified.

3.4.2 Training DMPNN to Predict for Antibiotic Bioactivity

I first utilized an open source message passing neural network, Chemprop (<https://github.com/chemprop/chemprop>), for molecular property prediction to train a D-MPNN that would predict the likelihood that a molecule would inhibit the growth of *E. coli*. The entire training process was done in accordance with the procedure found in Stokes et. al (**Figure 3.4**).¹⁰⁸ The authors were able to provide me with their primary training dataset composed of 2,335 compounds (deduplicated library containing 1,760 FDA-approved drugs and 800 natural products) that were binarized as hit or non-hit. Following binarization, I utilized these data to train a binary classification model that predicts the likelihood of a novel chemical inhibiting *E. coli* growth based on its structure. To do this, I used a directed-message passing deep neural network model⁸⁹, which converts a molecule's graph representation to a continuous vector using a directed bond-based message passing strategy. This method constructs a molecular representation iteratively by aggregating the properties of individual atoms and bonds. The model works by encoding information about surrounding atoms and bonds in "messages" that are sent along bonds. The model produces higher-level bond messages containing information about bigger chemical substructures by repeatedly performing this message passing procedure. The bond signals at the highest level are then concatenated into a single continuous vector describing the complete molecule.

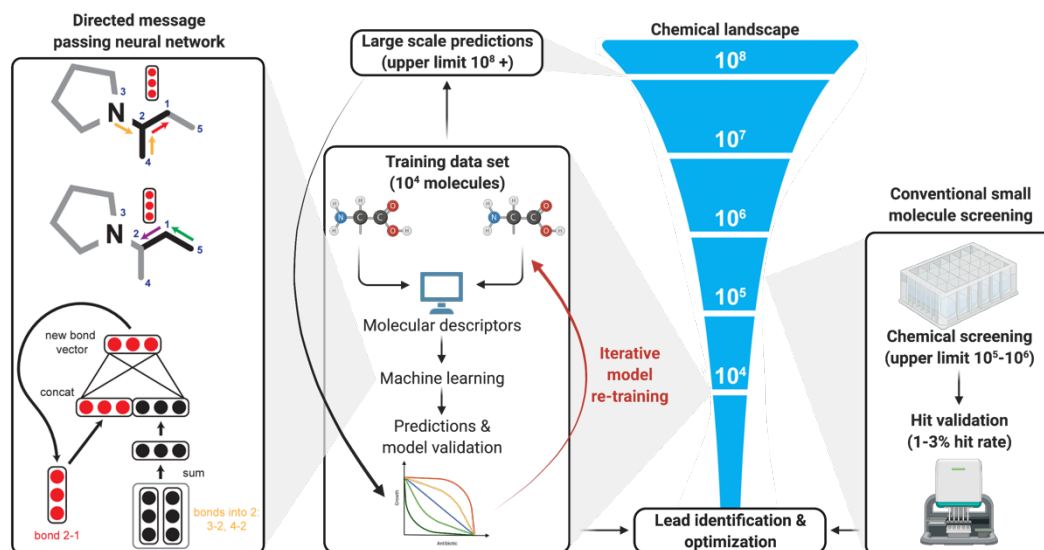


Figure 3.4 Schematic representation of machine learning in antibiotic discovery that demonstrates how the combination of *in silico* predictions and empirical investigations can lead to the discovery of new antibiotics.

The D-MPNN was then further optimized by adding computed molecular features provided by the cheminformatic package RDKit so that the D-MPNN could effectively learn to associate key features with the molecule's respective classification. I then improved the resilience of the approach even further by employing an ensemble of classifiers and predicting hyperparameters using Bayesian optimization to give the D-MPNN the best shot at accurately learning key molecular features.⁸⁸ The resulting model achieved a receiver operating characteristic curve-area under the curve (ROC-AUC) score of 0.982 and a precision recall curve-area under the curve (PRC-AUC) score of 0.734 on the test data.

3.4.3 Application of Trained Model to NP Atlas as Drug Hub

Once the D-MPNN model was optimized, I then applied an ensemble of models trained on twenty folds to identify potential antibacterial molecules from the Natural Products Atlas.⁶⁹ This database, at the time of the study, consisted of 20,035 microbial natural products with the associated cross references. Although microbial natural products have a high propensity for antibacterial properties, this database would allow for cross validation of the predictive results

in order to assess the ability for the D-MPNN to accurately interrogate structures as complex as natural products.

3.4.4 Results and Cross Validation via Literature

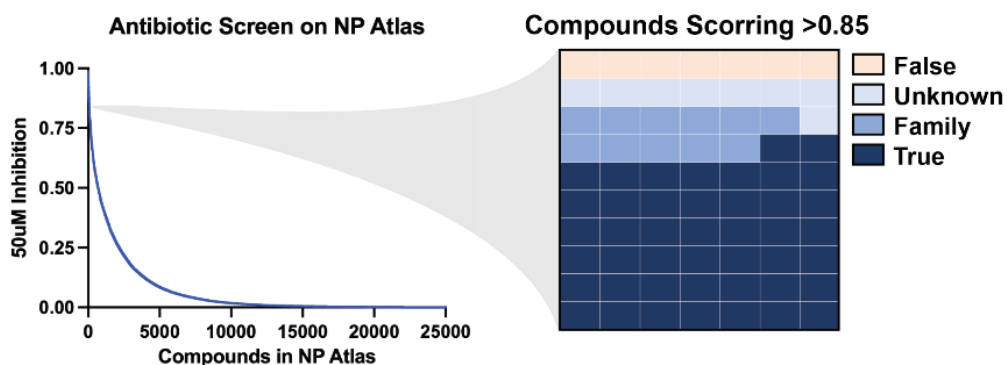


Figure 3.5 Cross validation of predicted natural products with antimicrobial characteristics.

Given that many natural products discovery efforts include antimicrobial screens for novel compounds, we were able to compare predicted scores for the natural products produced by the D-MPNN to the annotated antimicrobial characteristics found in the given literature (**Figure 3.5**). Our comparison of compounds identified by the D-MPNN to the characteristics of those compounds annotated in the literature encouraged us to believe that the algorithm was capable of attributing learned chemical characteristics to molecules as complex as natural products. Some of the microbial natural products identified by the D-MPNN include: Tetracycline (*Streptomyces* sp. CB01913), Actinomycin Z3 (*Streptomyces fradiae*), and Mitomycin C (*Streptomyces vertuillatus*) (**Table 3.3**).

On the opposite end of the scoring, the algorithm seemed to have more false negatives, but the low scoring compounds seemed to share lower molecular weights. Since the scoring metric is relative to the rest of the dataset, we could argue that the lower scores are a function of the compounds having fewer chemical descriptors correlated with bioactivity per structure. This could be addressed by splitting the datasets according to relative molecular weights to achieve appropriate scoring. The large mass variability within the dataset could also

explain the false positives observed in our top scoring region. For example, our DMPNN scored the lipodepsipeptide (LDP) Tolaasin C within the 90th percentile but was the only analog within the family that was reported to not possess antimicrobial activity. The lack of activity was correlated to the hydrolysis of its' lactone ring that was preserved amongst every bioactive LDP in their study. This acute change in functionality may be saturated by the large number of chemical descriptors accumulating due the high mass of the compound.

In the cases where the antimicrobial properties were unknown, it was simply due to the design of the bioassay utilized to isolate the NPs. For example, Leporzine C was isolated through a collaborative screening effort between Albany Molecular Research Incorporated (AMRI) and the Cystic Fibrosis Foundation Therapeutics (CFFT) to identify natural products that may act as correctors or potentiators of F508del-CFTR function. Collaborative efforts such as these present an impressive opportunity to discover targeted therapeutics for a given human disease, but such focus can inherently leave out antimicrobial/antibiotic characterizations of the identified compounds. The lack of antimicrobial/antibiotic characterizations of novel natural products identified through adjacent phenotypic screens can be addressed by repurposing the identified compounds through the means discussed in this thesis. Once these compounds have been identified by the DMPNN, it would be possible to screen them against the antimicrobial assay used to develop the training set for experimental validation.

3.5 Conclusion

The initial antibacterial prediction screen of the natural products found within the NP Atlas served to be a genuine proof of principle. I think it is most important to keep in mind what the network is actually achieving throughout the process. Each message passing step aggregates information from neighboring atoms, meaning each step of message passing moves information one step through the graph. If a molecule has two atoms that are, for instance, more than 3 bonds away from each other then those two atoms will never be able to relay information during the message passing phase given that the Chemprop algorithm utilizes

3 passing step by default. Nevertheless, both representations will be included in the final molecular representation since the embeddings of all atoms and bonds are summed throughout the process.

A simple solution would be to increase the number of message passing steps so that most atoms can communicate during message passing, but that can potentially introduce more concerning issues. Increasing the number of message passing steps will result in a much slower network and can potentially saturate the atom representations to cause all atoms to appear similar to the network, thereby losing important local chemical information. In conclusion, message passing neural networks are incredibly useful for characterizing localized graphs embedded within complex natural products. In the context of medicinal chemistry efforts, this can be interpreted as a way to identify bioactive functionality embedded within the structure of natural products. Such a tool can then be used to train message passing neural networks to select for optimized pharmacophores or bioactive functional groups that may have been missed by traditional natural product screening efforts.

3.9 Materials and Methods

Model training and predictions

A directed-message passing neural network (Chemprop), like other message passing neural networks, learns to predict molecular properties directly from the graph structure of the molecule, where atoms are represented as nodes and bonds are represented as edges. For every molecule, we reconstructed the molecular graph corresponding to each compound's SMILES string and determined the set of atoms and bonds using the open-source package RDKit (Landrum, 2006). Next, we initialized a feature vector, as described in Yang et al. (2019b), for each atom and bond based on computable features:

1. Atom features: atomic number, number of bonds for each atom, formal charge, chirality, number of bonded hydrogens, hybridization, aromaticity, atomic mass.
2. Bond features: bond type (single/double/triple/aromatic), conjugation, ring membership, stereochemistry.

The model applies a series of message passing steps where it aggregates information from neighboring atoms and bonds to build an understanding of local chemistry. In Chemprop, on each step of message passing, each bond's featurization is updated by summing the featurization of neighboring bonds, concatenating the current bond's featurization with the sum, and then applying a single neural network layer with non-linear activation. After a fixed number of message-passing steps, the learned featurizations across the molecule are summed to produce a single featurization for the whole molecule. Finally, this featurization is fed through a feed-forward neural network that outputs a prediction of the property of interest. Since the property of interest in our application was the binary classification of whether a molecule inhibits the growth of *E. coli*, the model is trained to output a number between 0 and 1, which represents its prediction about whether the input molecule is growth inhibitory.

In addition to the basic D-MPNN architecture described above, we employed three model optimizations⁸⁹:

Additional molecule-level features

While the message passing paradigm is excellent for extracting features that depend on local chemistry, it can struggle to extract global molecular features. This is especially true for large molecules, where the longest path through the molecule may be longer than the number of message-passing iterations performed, meaning information from one side of the molecule does not inform the features on the other side of the molecule. For this reason, we chose to concatenate the molecular representation that is learned via message passing with 200 additional molecule-level features computed with RDKit.⁸⁵

Additional molecule-level features

While the message passing paradigm is excellent for extracting features that depend on local chemistry, it can struggle to extract global molecular features. This is especially true for large molecules, where the longest path through the molecule may be longer than the number of message-passing iterations performed, meaning information from one side of the molecule does not inform the features on the other side of the molecule. For this reason, we chose to concatenate the molecular representation that is learned via message passing with 200 additional molecule-level features computed with RDKit.⁸⁵

Hyperparameter optimization

The performance of machine learning models is known to depend critically on the choice of hyperparameters, such as the size of the neural network layers, which control how and what the model is able to learn. We used the Bayesian hyperparameter optimization scheme, with 20 iterations of optimization to improve the hyperparameters of our model (see **Table 3.3**). Bayesian hyperparameter optimization learns to select optimal hyperparameters

based on performance using prior hyperparameter settings, allowing for rapid identification of the best set of hyperparameters for any model.¹⁰⁸

Table 3.3 List of hyperparameters for each respective dataset.

Dataset	Hyperparameter	Range	Value
Antibiotic	Number of message-passing steps	[2, 6]	5
	Neural network hidden size	[300, 2400]	1600
	Number of feed-forward layers	[1, 3]	1
	Dropout probability	[0, 0.4]	0.35
	Neural network hidden size	[300, 2400]	1100
	Number of feed-forward layers	[1, 3]	1
	Dropout probability	[0, 0.4]	0.35

Ensembling

Another standard machine learning technique used to improve performance is ensembling, where several copies of the same model architecture with different random initial weights are trained and their predictions are averaged. We used an ensemble of 20 models, with each model trained on a different random split of the data (Dietterich, 2000). Our initial training dataset consisted of 2,335 molecules, with 120 compounds (5.14%) showing growth inhibitory activity against *E. coli*, as defined by endpoint $OD_{600} < 0.2$.¹⁰⁸

Our experimental procedure consisted of four phases: (1a) a training phase to evaluate the optimized but non-ensembled model and (1b) training the ensemble of optimized models; (2) a prediction phase; (3) a retraining phase; and (4) a final prediction phase. We began by evaluating our model on the training set of 2,335 molecules using all optimizations except for ensembling, in order to determine the best performance of a single model. Here, we randomly split the dataset into 80% training data, 10% validation data, and 10% test data. We trained our model on the training data for 30 epochs, where an epoch is defined as a single pass through all of the training data, and we evaluated it on the validation data at the end of each epoch. After training was complete, we used the model parameters that performed best on the validation data and tested the model with those parameters on the test data. We repeated this

procedure with 20 different random splits of the data and averaged the results. After we were satisfied with model performance, we conducted predictions on new datasets. Since we wanted to maximize the amount of training data and were no longer interested in measuring performance on the test set, we trained new models on the training data from each of 20 random splits, each with 90% training data, 10% validation data, and no test data.

We lastly compared the prediction outputs of our augmented D-MPNN with a D-MPNN without RDKit features; a feedforward DNN model with the same depth as our D-MPNN model with hyperparameter optimization using RDKit features only; the same DNN instead using Morgan fingerprints (radius 2) as the molecular representation; and RF and SVM models using the same Morgan fingerprint representations. We used the scikit-learn implementation of a random forest classifier with all of the default parameters except for the number of trees, where we used 500 instead of 10.¹⁰⁹ When making predictions, we output the growth inhibition probability for each molecule according to the random forest, which is the proportion of trees in the model that predict a 1 for that molecule. Similarly, we used the scikit-learn implementation of a support vector machine with all of the default parameters. When making predictions, we output the signed distance between the Morgan fingerprint of the molecule and the separating hyperplane that is learned by the SVM. This number represents how much the model predicts a molecule is antibacterial, with large positive distances meaning most likely antibacterial and large negative distances meaning most likely not. Although the signed distance is not a probability, it can still be used to rank the molecules according to how likely they are to be antibacterial.

To predict the toxicity of molecules for possible in vivo applications, we trained a Chemprop model on the ClinTox dataset. This dataset consisted of 1,478 molecules, each with two binary properties: (a) clinical trial toxicity and (b) FDA-approval status. Of these 1,478 molecules, 94 (6.36%) had clinical toxicity and 1,366 (92.42%) were FDA approved. Using the same methodology as described in phase (1) of our experimental procedure, the Chemprop model was trained on both properties simultaneously and learned a single molecular

representation that was used by the feed-forward neural network layers to predict toxicity. We utilized the same RDKit features as in our other models, except for that the ClinTox model was an ensemble of five models and used the following optimal hyperparameters: message-passing steps = 6; neural network hidden size = 2200; number of feed-forward layers = 3; and dropout probability = 0.15.

Chemical analyses

We utilized Tanimoto similarity to quantify the chemical relationship between molecules predicted in our study. The Tanimoto similarity of two molecules is a measure of the proportion of shared chemical substructures in the molecules.¹¹⁰ To compute Tanimoto similarity, we first determined Morgan fingerprints (computed using RDKit) for each molecule using a radius of 2 and 2048-bit fingerprint vectors. Tanimoto similarity was then computed as the number of chemical substructures contained in both molecules divided by the total number of unique chemical substructures in either molecule. The Tanimoto similarity is thus a number between 0 and 1, with 0 indicating least similar (no substructures are shared) and 1 indicating most similar (all substructures are shared). Morgan fingerprints with radius R and B bits are generated by looking at each atom and determining all of the substructures centered at that atom that include atoms up to R bonds away from the central atom. The presence or absence of these substructures is encoded as 1 and 0 in a vector of length B, which represents the fingerprint. For t-SNE analyses, plots were created using scikit-learn's implementation of t-Distributed Stochastic Neighbor Embedding. Here, we first used RDKit to compute Morgan fingerprints for each molecule using a radius of 2 and 2048-bit fingerprint vectors. We then used t-SNE with the Jaccard distance metric to reduce the data points from 2048 dimensions to the two dimensions that are plotted. Note that Jaccard distance is another name for Tanimoto distance, and Tanimoto distance is defined as: Tanimoto distance = 1 - Tanimoto similarity. Thus, the distance between points in the t-SNE plots is an indication of the Tanimoto similarity of the corresponding molecules, with greater distance between molecules indicating lower

Tanimoto similarity. We used scikit-learn's default values for all t-SNE parameters besides the distance metric.

Software and Algorithms

Chemprop	Yang, et al., 2019	https://github.com/swansonk14/chemprop
RDKit	Landrum, 2016	https://github.com/rdkit

CHAPTER FOUR
APPLICATION OF DIRECTED MESSAGE PASSING NEURAL NETWORK TO RAPIDLY
IDENTIFY ANTI-VIRAL NATURAL PRODUCTS

4.1 Natural Products as a trove of bioactive compounds

Natural Products (NPs) remain a rich source of bioactive compounds that pave the road for the development of novel therapeutics for hard-to-treat pathogens. It is postulated that the rapid evolution of treatment-resistant pathogens could be rivaled by the diverse microbial NP chemical space.¹¹¹ NPs offer structural diversity and bioactivity that many synthetic therapeutics aim to mimic and incorporate into their scaffolds. As of 2019 there have been 441 therapeutics that are NPs or NP-derived compounds that have been FDA approved, and 424 synthetic therapeutics that mimic NPs or contain a NP pharmacophore.¹⁷ Of the FDA approved drugs that were explicitly categorized as antivirals, 19.5 percent of them were NP-derived or were inspired by the structural features found in NPs (**Figure 4.1**).¹⁷

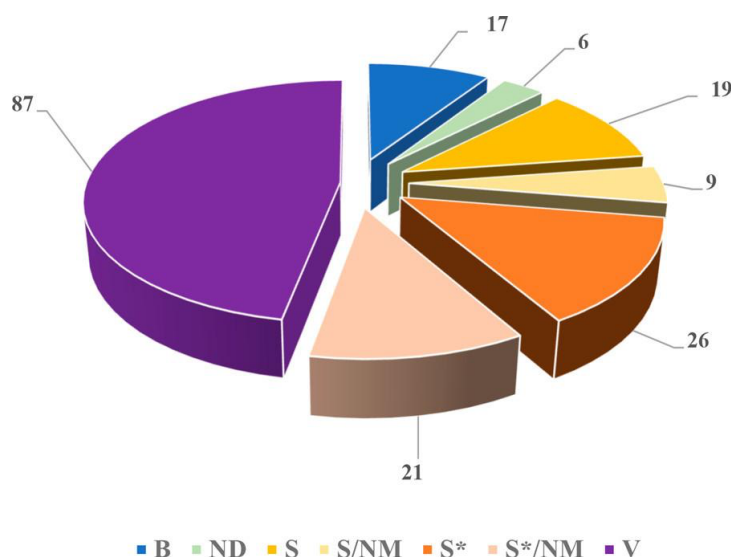


Figure 4.1 Antivirals drugs by source from 2015 to 09/2019, $n = 185$.¹⁷

Behind each great natural product, there lies a rigorous screening effort followed by thorough structure-activity-relationship (SAR) studies. The field of NP drug discovery, at the industry level and at the academic level, has greatly benefited from technological screening advancements such as genome mining and engineering; induced and heterologous expression of NP biosynthetic gene clusters;¹¹² high-content phenotypic bioactivity profiling;¹¹³ HRMS metabolomics;¹¹⁴ new dereplication methods and new molecular mode of action elucidation

platforms (Reviewed in Atanasov 2021, nature reviews drug discovery).¹¹⁵ Advancements in screening platforms that are more target-oriented and higher throughput have led drug discovery efforts to shift away from the crude extract samples and towards partially purified samples. The partially purified extracts allow for the parsing of complex mixtures that are inherent to crude NP extracts which has translated to better performance amongst molecularly targeted assay designs. Screening such an expansive set of complex fractions through these screening methods would still take a considerable amount of time – time we cannot afford amidst a pandemic. Ideally, the construction of pure NP libraries would yield the greatest drug-target-disease association, but the cost associated with the assembly of such a library is inhibitory. This is largely due to the generation of pure compounds through intramural isolations or collaborations, which can be limited by resource-intensive steps associated with the purification and characterization of individual compounds of sufficient quantity.⁶⁰ However, over 120 computational libraries exist that contain thousands of pure NPs from a variety of sources (plants, bacteria, fungus, metazoa, insects, and food) in the form of SMILES (Simplified Molecular Input Entry System) chains or other chemical identification notations. Although 98 of the NP databases require clearance, 50 of them remain open access such as the NP Atlas, NPASS, and CMAUP.^{116, 117} These open-access databases, coupled with in silico screening, are crucial for drug discovery when a global health crisis limits our ability to physically screen for a potential therapeutic.

4.2 Onset of the SARS-CoV-2 Virus

The COVID-19 pandemic, which caused about 180 million illnesses and 4 million fatalities globally by June 2021, necessitated a swift, vast, and effective therapeutic response.¹¹⁸ Since the emergence of the pandemic in late December 2019, researchers have been studying both the causative agent, SARS-CoV-2 virus, and the host response to the virus in order to better understand the disease pathogenesis, the structure of the constituent viral proteins, and identify key actionable proteins in order to guide the rapid development of both

antiviral therapeutics and host-directed agents.¹¹⁹ Clearly, developing new, effective medications for COVID-19 has not been possible because drug development and clinical testing often take 10–15 years. As a result, many scientists and clinicians have pursued the repurposing of existing drugs, clinical trial candidates, and approved natural products that have already been in clinical use and whose toxicity and preliminary pharmacokinetics have been deciphered in successful efforts to identify compounds with bioactivity against COVID-19.¹²⁰

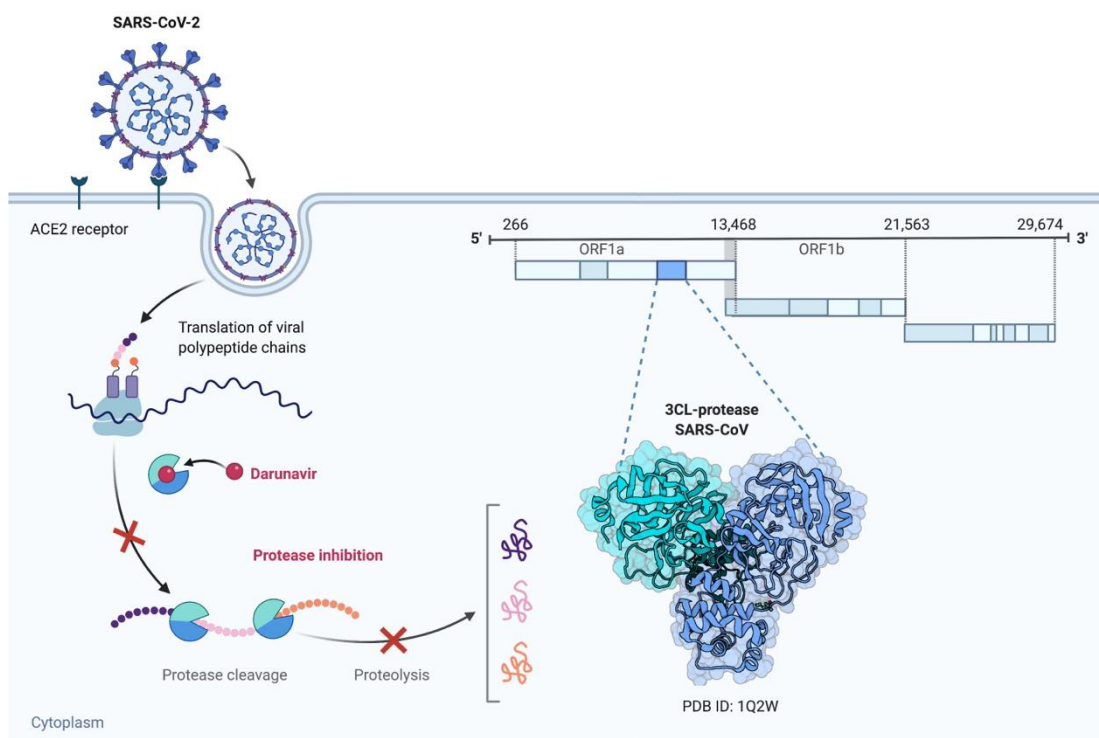


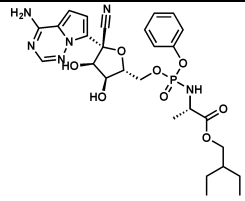
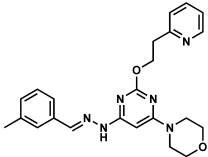
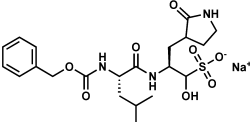
Figure 4.2 Schematic of SARS-CoV-2 replication mechanism within host.

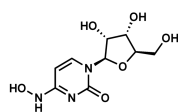
4.3 Target Selection for SARS-CoV-2

The Betacoronavirus genus includes SARS-CoV, MERS-CoV, and SARS-CoV-2, all which contain genomes with 5'-methylated caps at the N terminus and a 3'-poly-A tail at the C terminus, as well as a highly conserved order of genes linked to replication/transcription and structural components.¹²¹ In the first step of the SARS-CoV-2 life cycle, the spike protein on the outer surface of the virion is responsible for binding to the host receptor (ACE2) for attachment to the cell membrane, which is followed by viral and host cellular membrane fusion

and the release of viral genomic RNA into the cells. Subsequently, host ribosomes are hijacked to produce the two viral replicase polyproteins, which can further be processed into 16 mature nonstructural proteins (NSPs) through two virus-encoding proteases: main protease a 3C-like protease (3CLpro) and papain-like protease (PLpro).¹²² The first three peptide cleavages are performed by PLpro. The remainders are cleaved by 3CLpro (also called Mpro). These NSPs can assemble into the replication and transcription complex (RTC) to initiate viral RNA replication and transcription (**Figure 4.2**).¹²³ Several drug repurposing efforts have been carried out to identify compounds that selectively target these mechanisms of replication within the context of SARS-CoV-2. These efforts have successfully identified remdesivir as a potent inhibitor of the SARS-CoV-2 RNA polymerase with an EC₅₀ value of 0.77 μM.¹²⁴ The concept of drug repurposing can be thought of as the utilization of a predetermined compound library, such as the ZINC15 database or the NP Atlas.

Table 4.1 Examples of active molecules derived from drug repurposing for SARS-CoV-2.

Molecule	Name	Target	SARS-CoV-2 activity in Vero cells	SARS-CoV-2 activity on other cell types
	Remdesivir	RNA-dependent RNA polymerase	EC ₅₀ 0.77 μM ¹²⁴	Human epithelial cell culture (EC ₅₀ 0.01 μM); Calu-3 (EC ₅₀ 0.28mM) ¹²⁵
	Apilimod	PIKfyve	EC ₅₀ 0.023 μM ¹²⁶	293T cells (EC ₅₀ 0.012mM) ¹²⁷
	GC376	Mpro (Ki 12 nM)	EC ₅₀ 0.91 μM ¹²⁸	Not tested

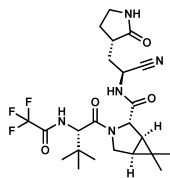


EIDD-1931

RNA-dependent
RNA polymerase

IC₅₀ 0.3
μM¹²⁹

Calu-3
(IC₅₀ 0.08mM)¹³⁰



PF-
07321332

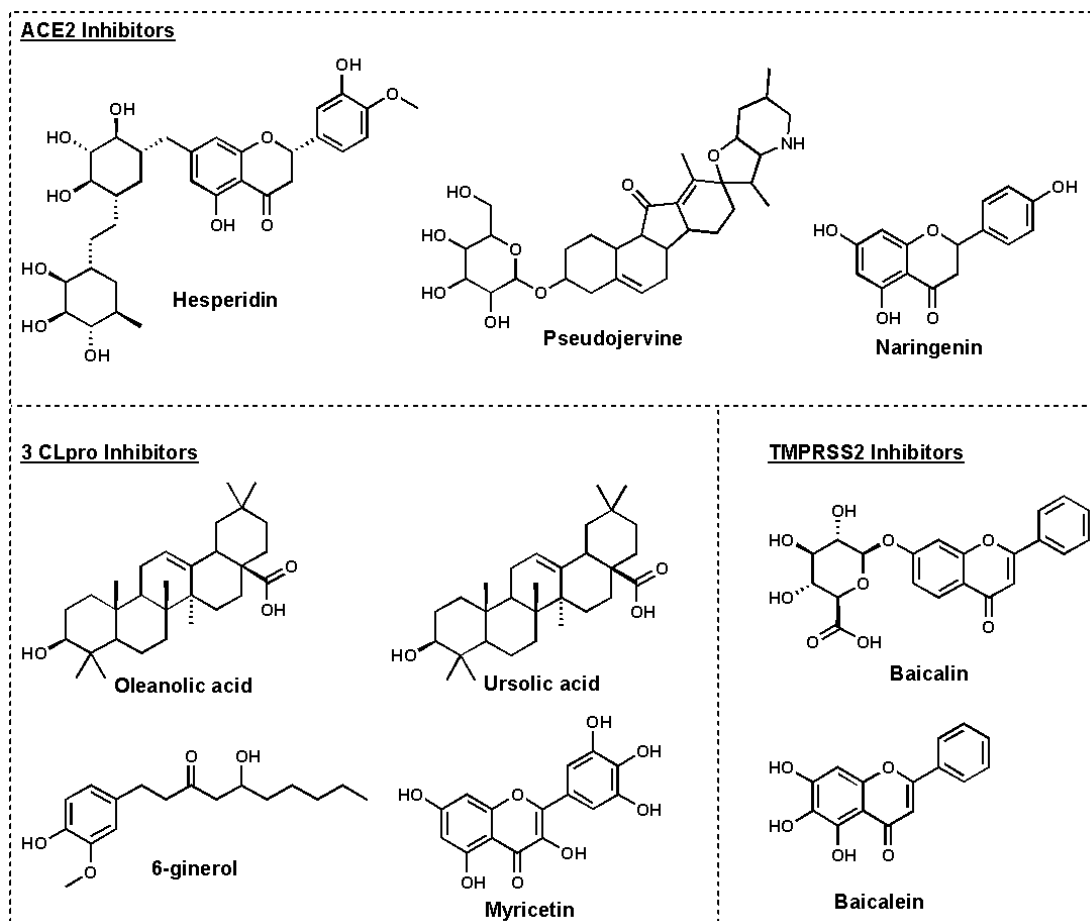
Mpro (Ki 3.11
nM)

EC₅₀ 74.5
nM¹³¹

See Ref [132]

The 3CL protease, also known as Main Protease (Mpro), plays a vital role in processing the polyproteins that are translated from the viral RNA. 3CL Protease inhibitors that can block viral replication are promising potential drug candidates that could be used to treat patients suffering with the COVID19 coronavirus infection. Thus, blocking viral replication by inhibiting 3CLpro is one of the key strategies in the drug development process (**Figure 4.2**). Some drugs in this aspect include GC376 or PF-07321332 (PAXLOVID™) which inhibit the 3C-like protease (3CLpro). In the CAS registry, the highest number of patents and potential drug candidates have been registered against 3CLpro among all the target proteins of SARS-CoV-2, which directly reflect its potentiality as a drug target.

4.4 *In-silico* Approaches to SARS-CoV-2 Drug Discovery



Scheme 4.1 NPs identified with high bioactivity against key SARS-CoV-2 proteins via *in silico* screening.

Several efforts have utilized virtual libraries of NPs to identify a potential therapeutic to combat the emerging SARS-CoV-2 virus. *In silico* screening trials are a powerful step to identify NPs with a high propensity for bioactivity when resources and time are limited. Computational simulations and neural networks decrease our reliance on time-consuming physical screening procedures and allow for a more streamlined discovery of drug-target-disease associations.

Some of the most promising *in silico* efforts for anti-SARS-CoV-2 NPs focus on glycosylated flavonols, flavanones, terpenoids, and alkaloids that are predicted to interact with key proteins (**Scheme 4.1**).^{133,134,135} Many of the *in-silico* screening efforts of large NP databases often come in the form of docking simulation that bears a heavy computational load

and specialized software. The recent advancements in machine learning technology presents a path for in silico screening with minimal computational loads that is readily accessible.

4.5 Application of DMPNN to Identify NPs Active Against SARS-CoV-2 Proteases

Given recent advancements in machine learning, the field is adequately prepared for the application of algorithmic solutions for molecular property prediction to identify novel therapeutic NP leads. The implementation of methodologies that allow early drug discovery to be performed largely in silico enables the exploration of vast bioactivity-rich chemical space that can streamline the formation of drug-target-disease associations. Recent advancements in modeling neural network-based molecular representations have allowed for learning automation in mapping molecules into continuous vectors that can subsequently be used to predict their properties. An important application of this innovation is best illustrated by three key steps taken to successfully identify a novel antibiotic through a combination of in silico predictions and empirical investigations. The first step required is to train a deep neural network model to predict a desired property using a dataset of molecules. Secondly, the resulting model is applied to several chemical libraries to identify potential lead compounds with the desired property. Lastly, ranking the compounds based on the model's predicted score and selecting leads based on chemical structure and availability. Utilizing this approach, coupled with the vast bioactivity-rich chemical space of natural product libraries represents an important opportunity to streamline drug discovery efforts.

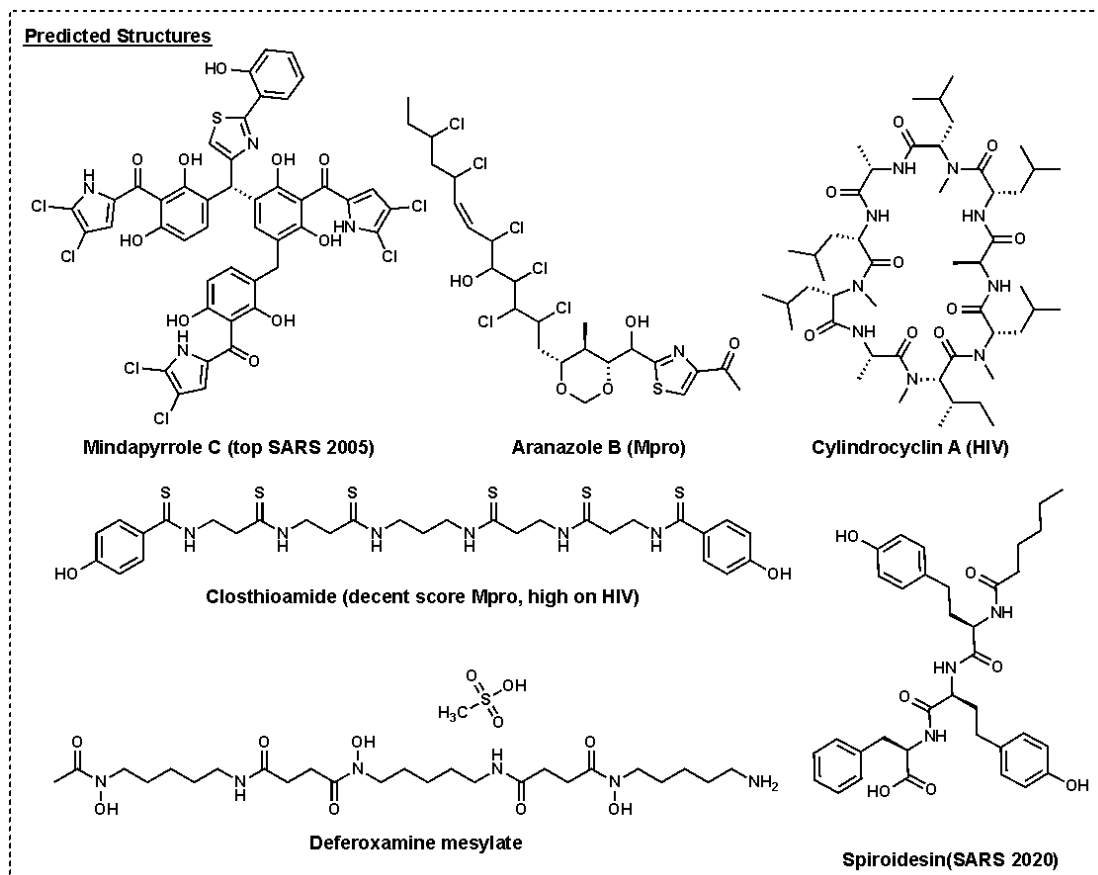
Table 4.2 List of publicly available drug repurposing campaigns for compounds against Betacoronavirus genus.

Main Target	Compounds	Hits	Source
<i>3CL Protease of SARS-CoV-2</i>	880 fragments	78	Diamond Light Source Group
<i>SARS-CoV-2 Replication</i>	1,484	88	NCBI

<i>SARS-CoV-2 Cytotoxicity</i>	5,632	67	NCBI
<i>3CL Protease of SARS-CoV</i>	290,726	405	NCBI
<i>PL Protease of SARS-CoV</i>	233,891	697	NCBI
<i>HIV Replication</i>	41,127	1,443	MoleculeNet

Herein, we report the application of deep neural network models to predict the bioactivity of NPs against the SARS-CoV-2 M-protease. Our approach consisted of three stages. First, we chose publicly available datasets that screened small molecules against the main proteases of both the SARS-CoV and SARS-CoV-2 viral strains (**Table 4.2**). We also used untargeted screens that provided information on small molecules that inhibited SARS-CoV-2 replication. Lastly, we chose to include a large dataset composed of compounds that have been screened against HIV replication so that we would select for antivirals from a broad sense. We used the Chemprop algorithm in combination with open-source datasets to train six independent deep neural network models to predict activity against key proteins that allow for viral entry. We then applied the resulting models to the >20,000 microbial-derived NPs listed within the NP Atlas library to identify potential lead compounds with activity against key viral proteins. Lastly, the NPs were ranked according to the model's predicted score and the compounds with a relative pre-specified score were cross-examined against the other model predictions. We selected a list of NPs based on the relative prediction score across the panel of models, availability, and synthetic feasibility. Through this approach, we successfully identified Closthioamide and Deferroxamine mesylate from the NP Atlas to possess activity against the SARS-CoV-2 MProtease.

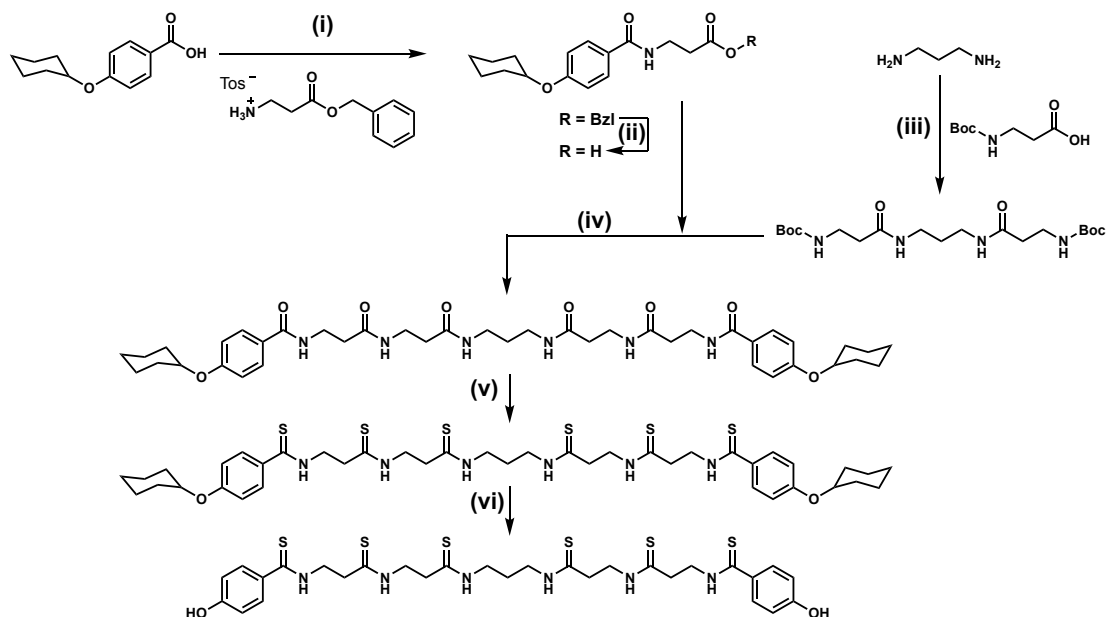
4.6 Conclusion



Scheme 4.2 Several NPs with a high propensity to be active against SAR-CoV-2 Mpro via DMPNN.

Once we possessed the read-out scores from our antiviral screen, we began to prioritize the top scoring compounds based on several conditions. Initially, we selected for the top scoring compounds within each respective screen that were readily available to purchase or synthesize. We also provided a list of high scoring compounds to colleagues within the department, and our in-house chemical screening center, in the case that they had some of the compounds in storage. The top scoring compounds varied in chemical complexity spanning from cyclic peptides to siderophores (**Scheme 4.2**).

Many of the compounds identified by our screens have been characterized for adjacent bioactivity. For example, Cyclosporin A was the highest scoring compound within our HIV screen and has been shown to exhibit modest cytotoxicity ($IC_{50} \sim 11 \mu M$) against COLO-320, HL-60, L1210 and human T lymphocyte Jurkat cells. The compound we ultimately selected to synthesize in house, Closthioamide, was also shown to be a potent inhibitor of bacterial DNA gyrase; however, its molecular mechanism differs from that of the quinolones and aminocoumarins. We decided to test Closthioamide, Cyclosporin A, Cyclosporin A, Ribocyclophane C, and Desferroxamine mesylate using a SARS-CoV-2 specific 3CLpro assay kit (Catalog #79955-1) purchased from BPS Biosciences (CA).



Scheme 4.3 Schematic presentation of the total synthesis of Closthioamide.

Given the structural simplicity of Closthioamide, I decided to evoke a modified synthetic route analogous to that provided by Hertweck et. al (**Scheme 4.3**). The synthesis began with the generation of 4-(cyclohexyloxy)benzoic acid (**13**) from the addition of cyclohexane to methyl 4-hydroxybenzoate in the presence of boron trifluoride diethyl etherate under reflux. The protected benzoic acid was then utilized to form the amide (**14**) via coupling with a benzyl ester in the presence of DCC and HOBt in triethylamine (i). The terminal amide (**14**) was then

subjected to a lithium hydroxide reduction (ii) to remove the benzyl group to afford the propanoic acid (15). The amide backbone (9) was synthesized utilizing 1,3-diaminopropane and boc-beta-alanine in the presence of DDC and HOBT in DMF (iii). Removal of the boc group proved to be challenging, but was eventually achieved by subjecting compound 9 to acid dioxane under constant sonication followed by a methanolic quench with continuous sonication to afford bis(hydrochloride). The protected hexamide backbone (16) was then synthesized by converging our propanoic acid and bis(hydrochloride) in the presence of Hünig's Base, HOBT, and DCC (iv). Thionation via Lawesson's reagent in pyridine (v) worked remarkably well to afford the protected hexathioamide 17, which was subsequently deprotected utilizing trifluoromethanesulfonic acid in TFA (vi) to yield closthioamide (18).



Figure 4.3 IC₅₀ curves generated via fluorescent protein engagement assay.

The natural products in **Scheme 4.2** were selected based on their high scores across all six DPMNN models and their accessibility (immediately available in-house or synthetically feasible). Cyclosporin A and Cyllindrocyclin A were provided by the lab of Prof. Scott Lokey. Ribocyclophane C was sourced from the Chemical Screening Center at UCSC. We also chose to include a known protease inhibitor, GC376, and a reported protease inhibitor, MG-132, found in the literature at the time. Our assay resulted in hits for both Closthioamide (IC₅₀ 8.602 μ M) and Desferroxamine mesylate (IC₅₀ 3.303 μ M) – a 40 percent positive hit rate for our predicted protease inhibitors.

The results of our studies indicate that the Chemprop algorithm is capable of training DMPNNs for the prediction of antiviral activity of the natural products found within the NP Atlas.

The rapid development of machine learning algorithms is quickly become a tool for drug discovery and is quickly approaching a realm of accuracy that will allow for its application to complex structures such as natural products. Graphical neural networks allow for a rapid

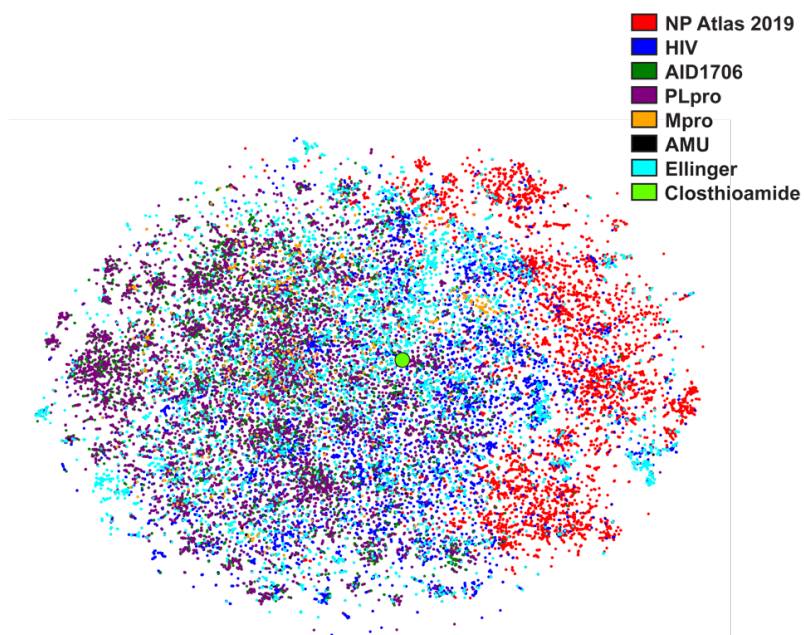


Figure 4.4 t-Distributed stochastic neighbor embedding (t-SNE) of all molecules from the training datasets (blue, forest green, purple, orange, black, and neon blue) and the NP Atlas (red), with Closthioamide (neon green).

characterization of vast chemical libraries without inheriting the large computational burden observed by other *in silico* methods. This study encompassed expansive chemical diversity by natural products isolated from various microbial organisms and thousands of FDA-approved drugs (**Figure 4.3**). The t-SNE plot above reveals the chemical relationships between the libraries used to train the DMPNNs and the NP Atlas, with Closthioamide highlighted in neon green. Carrying out a physical screen of this magnitude would have countless hours and numerous resources that proved to be scarce in the face of a global pandemic.

4.7 Materials and Methods

SARS-CoV-2 genome encodes for more than 20 proteins, including the main protease (Mpro) which is a 3C-like protease (3CLP) that happens to share 96.1% similarity with the 3CLP of the SARS-CoV. Utilizing a series of open-source datasets from the National Center for Biotechnology Information (NCBI) we began to train several D-MPNN's to recognize chemical features that resulted in bioactivity against replication in both SARS-CoV and SARS-CoV-2. The datasets were chosen to span from in vitro assays of FDA-approved compounds to fragment screens for targeted protease binding. The high level of similarity between the SARS-CoV and SARS-CoV-2 encouraged us to utilize previously assembled compound libraries that characterize the activity of known therapeutics against Mpro/3CLP in the context of SARS-CoV. The referenced targeted screens focused on either of the two polyproteins encoded by the coronavirus, the 3CL protease (M-pro) and a papain-like protease (Plpro). The D-MPNN's were trained under varying conditions; some with added RDKit features, some with random training splits, and some with scaffolded training splits. The goal was to identify which conditions would accurately interpret the given data sets so that we could have a well-trained network to utilize on the NP Atlas data set.

SARS-CoV Data

A comprehensive bioassay conducted by The Scripps Research Institute Molecular Screening Center titled "Summary of probe development efforts to identify inhibitors of the SARS- coronavirus 3C-like Protease (3CLP)" provided a data set of 290,726 molecules with 405 of those exhibiting activity against 3CLP via fluorescence. Broad repurposing hub evaluation set from SARS-CoV 3CL protease containing 41 experimentally validated hits along with 5630 other molecules. The 41 validated hits were combined with the Scripps screen to yield a dataset containing 290,767 compounds and 446 hits. This dataset was used to train a D-MPNN with the parameters founds in **Table 4.3**.

We also utilized a dataset from a bioassay that detects activity against SARS-CoV in yeast models via PL protease inhibition titled “qHTS o Yeast-based Assay for SARS-CoV PLP: Summary.” The dataset is a combination of a broad screen and their follow up validation screen that contains 233,891 compounds and 697 hits. This dataset was used to train a D-MPNN with the parameters founds in **Table 4.3**.

SARS-CoV-2 Data

Fragments screened for 3CL protease binding using crystallography techniques. Data was sourced from the Diamond Light Source group. The dataset contained 880 fragments with 78 hits. This dataset was used to train a D-MPNN with the parameters founds in **Table 4.3**.

FDA-approved compounds screened against SARS-CoV-2 *in vitro*. Data was sourced from “In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication.” The dataset contained 1,484 FDA-approved compounds with 88 identified hits. This dataset was used to train a D-MPNN with the parameters founds in **Table 4.3**.

Table 4.3 List of hyperparameters for each respective dataset.¹³⁶⁻¹⁴²

Dataset	Hyperparameter	Range	Value
Antibiotic	Number of message-passing steps	[2, 6]	5
	Neural network hidden size	[300, 2400]	1600
	Number of feed-forward layers	[1, 3]	1
	Dropout probability	[0, 0.4]	0.35
AID1706	Number of message-passing steps	[2, 6]	3
	Neural network hidden size	[300, 2400]	1300
	Number of feed-forward layers	[1, 3]	2
	Dropout probability	[0, 0.4]	0.4
PLpro	Number of message-passing steps	[2, 6]	3
	Neural network hidden size	[300, 2400]	1300
	Number of feed-forward layers	[1, 3]	2
	Dropout probability	[0, 0.4]	0.1
Mpro	Number of message-passing steps	[2, 6]	4
	Neural network hidden size	[300, 2400]	800

amu_sars_cov_2	Number of feed-forward layers	[1, 3]	3
	Dropout probability	[0, 0.4]	0.35
	Number of message-passing steps	[2, 6]	2
	Neural network hidden size	[300, 2400]	1400
Ellinger	Number of feed-forward layers	[1, 3]	2
	Dropout probability	[0, 0.4]	0.3
	Number of message-passing steps	[2, 6]	6
	Neural network hidden size	[300, 2400]	1100
HIV	Number of feed-forward layers	[1, 3]	1
	Dropout probability	[0, 0.4]	0.35
	Number of message-passing steps	[2, 6]	3
	Neural network hidden size	[300, 2400]	1000
	Number of feed-forward layers	[1, 3]	2
	Dropout probability	[0, 0.4]	0.35

Compounds screened against SARS-CoV-2 *in vitro*. Data was sourced from “Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large-scale drug repurposing collection.” The dataset contained 5,632 compounds with 67 identified hits. This dataset was used to train a D-MPNN with the parameters found in **Table 4.3**.

HIV Data

Dataset	Metric	Score
Antibiotic	ROC-AUC	0.951744 ± 0.025224
AID1706	ROC-AUC	0.794167 ± 0.031480
PLpro	ROC-AUC	0.763580 ± 0.031110
Mpro	ROC-AUC	0.850082 ± 0.056475
amu_sars_cov_2	ROC-AUC	0.676327 ± 0.077956
Ellinger	ROC-AUC	0.760277 ± 0.074501
HIV	ROC-AUC	0.778134 ± 0.000013

To ensure that our results also included previously identified antivirals we utilized a publicly available dataset from the MoleculeNet benchmark repository that contained

experimentally measured abilities to inhibit HIV replication. The dataset contained 41,127 compounds with 1,443 identified as confirmed active or confirmed moderately active. This dataset was used to train a D-MPNN with the parameters found in **Table 4.3**.

Application of Models

After D-MPNN model development and optimization using the respective training datasets, we subsequently applied an ensemble of models trained on twenty folds to identify potential antivirals from the Natural Products Atlas. This library contains 20,035 natural products isolated from bacterial, fungal, and cyanobacterial source organisms, with 18,959 of the compounds listed as being unique by the MongoDB COllection of Open Natural prodUcTs (COCONUT). Here, prediction scores for each compound were determined and molecules were ranked based on their probability of displaying antiviral activity. There were no identified overlaps between the compounds in the training datasets and the NP Atlas dataset.

Hit Prioritization

The molecules predicted to display antiviral characteristics were ranked relative to the model that produced the scores. We used a prediction cutoff of (>0.9 Antibiotic, >0.6 HIV, etc.) The prediction scores produced by the appropriate models were then cross examined to see if we could identify any natural products that scored well across several models. The compounds that scored exceptionally on respective models and those that appeared to score well across multiple models were then evaluated based on structural characteristics, commercial availability, and synthetic feasibility.

Additional molecule-level features

While the message passing paradigm is excellent for extracting features that depend on local chemistry, it can struggle to extract global molecular features. This is especially true for large molecules, where the longest path through the molecule may be longer than the

number of message-passing iterations performed, meaning information from one side of the molecule does not inform the features on the other side of the molecule. For this reason, we chose to concatenate the molecular representation that is learned via message passing with 200 additional molecule-level features computed with RDKit.

Hyperparameter optimization

The performance of machine learning models is known to depend critically on the choice of hyperparameters, such as the size of the neural network layers, which control how and what the model is able to learn. We used the Bayesian hyperparameter optimization scheme, with 20 iterations of optimization to improve the hyperparameters of our model (see **Table 4.3**). Bayesian hyperparameter optimization learns to select optimal hyperparameters based on performance using prior hyperparameter settings, allowing for rapid identification of the best set of hyperparameters for any model.

Ensembling

Another standard machine learning technique used to improve performance is ensembling, where several copies of the same model architecture with different random initial weights are trained and their predictions are averaged. We used an ensemble of 20 models, with each model trained on a different random split of the data (Dietterich, 2000).

Chemical analyses

We utilized Tanimoto similarity to quantify the chemical relationship between molecules predicted in our study. The Tanimoto similarity of two molecules is a measure of the proportion of shared chemical substructures in the molecules. To compute Tanimoto similarity, we first determined Morgan fingerprints (computed using RDKit) for each molecule using a radius of 2 and 2048-bit fingerprint vectors. Tanimoto similarity was then computed as the number of chemical substructures contained in both molecules divided by the total number of

unique chemical substructures in either molecule. The Tanimoto similarity is thus a number between 0 and 1, with 0 indicating least similar (no substructures are shared) and 1 indicating most similar (all substructures are shared). Morgan fingerprints with radius R and B bits are generated by looking at each atom and determining all of the substructures centered at that atom that include atoms up to R bonds away from the central atom. The presence or absence of these substructures is encoded as 1 and 0 in a vector of length B, which represents the fingerprint. For t-SNE analyses, plots were created using scikit-learn's implementation of t-Distributed Stochastic Neighbor Embedding. Here, we first used RDKit to compute Morgan fingerprints for each molecule using a radius of 2 and 2048-bit fingerprint vectors. We then used t-SNE with the Jaccard distance metric to reduce the data points from 2048 dimensions to the two dimensions that are plotted. Note that Jaccard distance is another name for Tanimoto distance, and Tanimoto distance is defined as: $\text{Tanimoto distance} = 1 - \text{Tanimoto similarity}$. Thus, the distance between points in the t-SNE plots is an indication of the Tanimoto similarity of the corresponding molecules, with greater distance between molecules indicating lower Tanimoto similarity. We used scikit-learn's default values for all t-SNE parameters besides the distance metric.



Figure 4.4 Graphical representation of validation bioassay.

Bioassay Validation of predicted hits

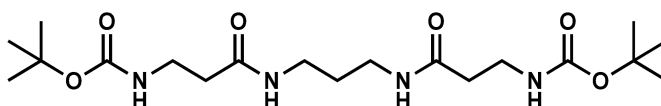
SARS-CoV-2 specific 3CLpro assay kits (Catalog #79955-1) were purchased from BPS Biosciences (CA) and assay was carried out as per the manufacturer's recommendations. In Brief, in a 96-well plate 4 ng/ μ l 3CLpro-MBP tagged enzyme (120 ng per reaction) in 30 μ l of assay buffer was pre-incubated with varying concentrations of selected natural products for 40 min. The enzymatic reaction was initiated by adding 10 μ l (50 μ M final) fluorescent substrate. Fluorescence kinetic measurements were taken in Envision 2105 multimode plate reader for 16 h with 360/40 excitation and 460/40 nm emission. For IC50 calculation, drugs were screened at 0 to 100 μ M dose range. Positive control for no enzyme inhibition was 1% DMSO with 4 ng of enzyme and 50 μ M of substrate. 100 μ M of GC-376 served as inhibitor control and blanks were wells with 1% DMSO with 50 μ M of substrate without enzyme. All the values were subtracted from blank values and percent inhibition calculations compared to DMSO controls.

Software and Algorithms

Chemprop	Yang, et al., 2019	https://github.com/swansonk14/chemprop
RDKit	Landrum, 2016	https://github.com/rdkit

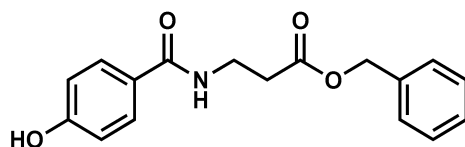
Total Synthesis of Closthioamide

All reagents were obtained from commercial suppliers (Sigma Aldrich, TCI, etc.) and used without further purification unless otherwise explained. Reactions were carried out under inert gas (N₂) by using the Schlenk technique in dried solvents. Dry methanol was obtained by desiccation with magnesium and subsequent distillation. Dimethylformamide was dried over molecular sieve (4 Å), distilled and stored over molecular sieve (4 Å) under N₂. Dry dichloromethane (DCM) was generated by distillation from calciumhydride suspension. Triethylamine and diisopropylethylamine (Hünig's Base) were dried over potassium hydroxide and distilled. Dioxane was dried over potassium hydroxide. Methanol, chloroform, dichloromethane, ethyl acetate and benzyl alcohol were distilled prior to use. Open column chromatographic separations were executed on silica gel (Kieselgel 60, 15-40 µm, Merck KGaA). Reaction progresses were monitored by thin layer chromatography (TLC) (silica gel on aluminium sheets 20 x 20 cm with fluorescent dye 254 nm, Merck KGaA), GC-MS or HPLC-MS. β-Alanine benzylester p-toluenesulfonate (4) was purchased from Bachem AG, while larger amounts were easily obtained according to a standard procedure. 4-(Cyclohexyloxy)benzoic acid (3b) could be provided by various suppliers (Life Chemicals, Aronis, etc.)

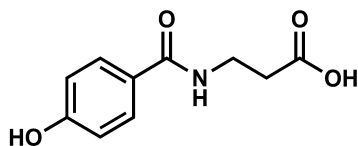


Compound **9**. *tert*-Butyl 3,3'-(propane-1,3-diylbis(azanediyl))bis(3-oxopropane-3,1-diyl)dicarbamate (**9**) was synthesized using a simplified literature procedure.¹⁴³ To a solution of Boc-β-Ala-OH (7.2 g, 42 mmol) and 1-hydroxy-benzotriazole (HOBt) (6.7 g, 48 mmol) in dry dimethylformamide (DMF) (80 mL) was added *N,N'*-Diisopropylcarbodiimide (DCC) (9.20 g, 48 mmol). The mixture was stirred for one hour at room temperature (rt), subsequently treated with 1,3-diaminopropane (1.48 g, 1.66 mL, 20 mmol), and stirred for an additional 96 h at rt. The solvent was evaporated under reduced pressure and the residue was and the residue was

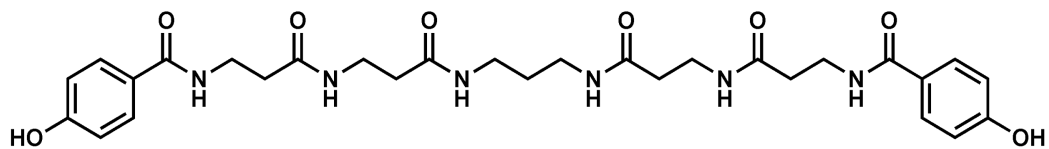
mixed with aqueous sodium hydroxide solution (xx mL, 0.2M) and filtered. The white solid product was washed with aqueous sodium hydroxide solution (50 mL, 0.2M), with water (2 x 50 mL), with hydrochloric acid (2 x 25 mL, 1M) and dried in high vacuum to yield the desired product (7.2 g, 86%).



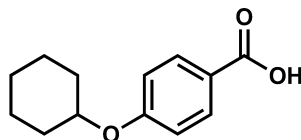
Compound **10**. Benzyl 3-(4-hydroxybenzamido)propanoate (**10**) was synthesized in analogy to standard procedures.¹⁴³ β -Alanine benzylester p-toluenesulfonate (7.03 g, 20 mmol), EDC (4.60 g, 24 mmol) and HOBT (3.35 g, 24.8 mmol) were dissolved in dry DCM (50 mL). Subsequently, triethylamine (3.64 g, 5.0 mL, 36 mmol) was added and the mixture was stirred at rt for 42 h. The solution was diluted with 150 mL ethyl acetate and washed twice with 50 mL saturated aqueous ammonium chloride solution in 150 mL water. The organic extract was dried with sodium sulfate and the solvent was evaporated. The solid crude product was mixed with hot chloroform (15 mL) and diethyl ether was added dropwise until cloudiness. Crystallization was induced in an ultrasonic bath. Diethyl ether (30 mL) was added and the solid was removed by filtration and dried in fine vacuum (6.9 g, 88%).



Compound **11**. 3-(4-Hydroxybenzamido)propanoic acid (**11**). 4 Benzyl 3-(4-hydroxybenzamido) propanoate (**10**) (1.8 g, 4.7 mmol) and Lithium Hydroxide (10% loading, 281 mg, 11.75 mmol) were suspended in THF with H₂O (3:2, 50 mL). Reaction was left to stiff for 4 hours at rt. Subsequently, the suspension was filtered, and the solvent of the filtrate was evaporated under reduced pressure yielding the pure desired product (1.40 g, ~97%).

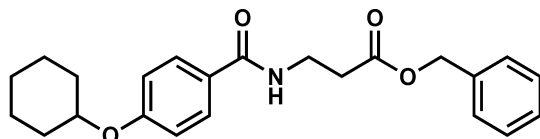


Compound **12**. N,N'-(3,7,13,17-Tetraoxo-4,8,12,16-tetraazanonadecane-1,19-diyl)bis(4-hydroxy benzamide), "closamide" (**12**). tert-Butyl 3,3'-(propane-1,3-diyl)bis(azanediyl))bis(3-oxopropane-3,1-diyl)dicarbamate (**9**) (104 mg, 0.25 mmol) was dissolved in a solution of hydrogen chloride in dried methanol (1 mL, 5.2 M). The solution was treated in an ultrasonic bath at 50 °C for 2.25 h. The solvent was removed under reduced pressure, and the residue was mixed with dry DMF (2.5 mL). Triethylamine (61 mg, 84 μ L, 0.6 mmol), 3-(4-hydroxybenzamido)propanoic acid (**6a**) (110 mg, 0.53 mmol), HOBt (81 mg, 0.62 mmol) and DCC (124 mg, 0.6 mmol) were added, and the mixture was stirred at rt for 21 h before the solvent was removed in vacuo. The residue was treated with chloroform (5 mL) in an ultrasonic bath and filtered. The white solid was reprecipitated from methanol yielding the desired product (91.5 mg, 61%).

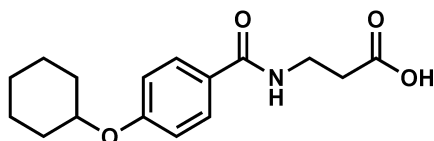


Compound **13**. 4-(Cyclohexyloxy)benzoic acid (**13**) was prepared in accordance with a reported procedure.¹⁴³ Methyl 4-hydroxybenzoate (15.2 g, 0.1 mol) and cyclohexene (80 mL) were mixed and boron trifluoride diethyl etherate (7.1 g, 6.3 mL, 0.05 mol) was added. The mixture was heated under reflux for 2 h and cooled down to rt. Ethyl acetate (50 mL) was added and the solution was washed with aqueous sodium hydroxide solution (3 x 100 mL, 5%) and once with water (100 mL). The organic layer was dried with sodium sulfate and the solvent was evaporated. The oily residue was mixed with water (50 mL), methanol (100 mL) and acetone (30 mL), sodium hydroxide was added (20 g, 0.5 mol) and the emulsion was heated under reflux for 3 h. The solvent was largely evaporated under reduced pressure and the residue was

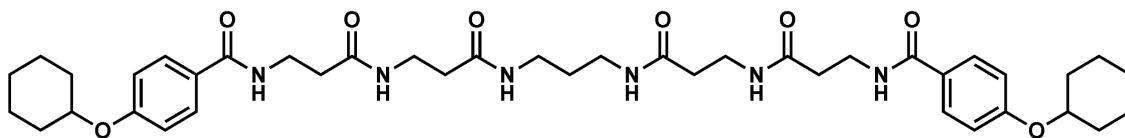
treated with hydrochloric acid (120 mL, 18%). The precipitate formed was removed by filtration and recrystallized from glacial acetic acid yielding the desired product (14.77 g, 67%).



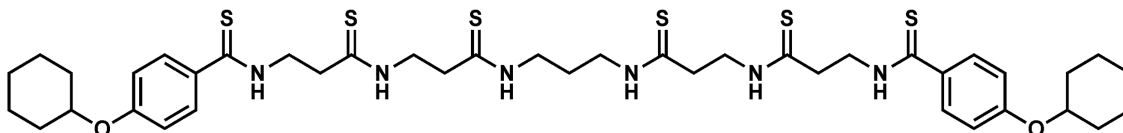
Compound **14**. Benzyl 3-(4-(cyclohexyloxy)benzamido)propanoate (**14**). 4-(Cyclohexyloxy)benzoic acid (**13**) (4.63 g, 21 mmol), β -alanine benzylester p-toluenesulfonate (7.39 g, 20 mmol), HOBt (3.38 g, 25 mmol) and dicyclohexylcarbodiimide (DCC) (4.95 g, 24 mmol) were suspended in dry DCM, and Hünig's Base (3.87 g, 5.16 mL, 30 mmol) was slowly added. The mixture was stirred at rt for 18 h, and the solvent was removed under reduced pressure. The residue was mixed with ethyl acetate, washed twice with hydrochloric acid (50 mL, 1M) and once with aqueous sodium hydroxide solution (50 mL, 0.2 M), dried with sodium sulfate, and the solvent was evaporated. The oily residue was purified by flash chromatography (cyclohexane : ethyl acetate = 4 : 1) yielding the desired product as a colorless oil, which crystallized within several hours (7.3 g, 96%).



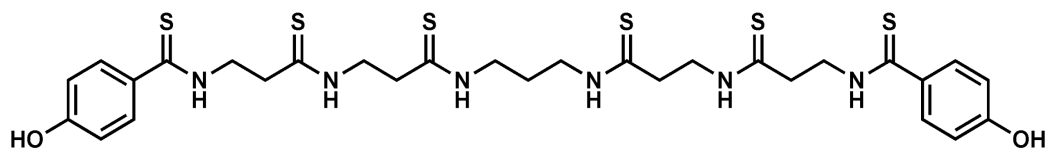
Compound **15**. 3-(4-(Cyclohexyloxy)benzamido)propanoic acid (**15**). Benzyl 3-(4-(cyclohexyloxy)-benzamido)propanoate (**14**) (2.67 g, 7 mmol) and Lithium Hydroxide (10% loading, 281 mg, 11.75 mmol) were suspended in THF with H₂O (3:2, 50 mL). Reaction was left to stiff for 4 hours at rt. The mixture was filtered and the solvent of the filtrate was evaporated (2.06 g, >99%).



Compound **16**. N,N'-(3,7,13,17-Tetraoxo-4,8,12,16-tetraazanonadecane-1,19-diyl)bis(4-(cyclohexyl- oxy)benzamide), "Chx2closamide" (**16**). tert-Butyl 3,3'-(propane-1,3-diyl)bis(azanediyl))bis- (3-oxopropane-3,1-diyl)dicarbamate (**9**) (208 mg, 0.5 mmol) was dissolved in dry dioxane (5 mL), and a solution of hydrogen chloride in dioxane (5 mL, 3.4M) was added. The mixture was sonicated for 1 h at 50 °C. Subsequently, the suspension was diluted with dry methanol (5 mL) and sonicated for further 2 h at 50 °C. The solvent was removed at low pressure and the residue was mixed with dry DMF (20 mL). Hünig's Base (161 mg, 215 μ l, 1.25 mmol), 3-(4-(cyclohexyloxy)benzamido)-propanoic acid (**15**) (306 mg, 1.05 mmol), HOBt (162 mg, 1.24 mmol) and DCC (230 mg, 1.20 mmol) were added. After stirring at 30 °C for 68 h, the solvent was evaporated in fine vacuum and the residue was treated with ethyl acetate (10 mL) in an ultrasonic bath and centrifugated. The precipitate was washed once with aqueous hydrochloric acid (25 mL, 1M) and twice with ethyl acetate (10 mL) by sonication and subsequent centrifugation. The white solid was dried in fine vacuum to yield the desired product (338 mg, 89%).



Compound **17**. N,N'-(3,7,13,17-Tetrathioxo-4,8,12,16-tetraazanonadecane-1,19-diyl)bis(4-(cyclohexyl- oxy)benzothioamide), "Chx2closthioamide" (**17**). Chx2closamide (**16**) (38.1 mg, 50 μ mol) and Lawesson's Reagent (121.4 mg, 300 μ mol) were suspended in dry pyridine (1 mL) and stirred in a closed vial for 23 h at 100 °C. The solvent was removed under reduced pressure and the residue was mixed with ethyl acetate (10 mL), washed twice with aqueous sodium hydroxide solution (10 mL, 1M), and twice with hydrochloric acid (10 mL, 1M). The organic layer was dried with sodium sulfate and the solvent was evaporated yielding the desired product in good purity (41.3 mg, 96%). (The substance could further be purified by open column chromatography (chloroform : methanol = 9 : 1, R_f = 0.5)).



Compound **18**. N,N'-(3,7,13,17-Tetrathioxo-4,8,12,16-tetraazanonadecane-1,19-diyl)bis(4-hydroxy- benzothioamide), “closthioamide” (**18**).

Chx2closthioamide (**17**) (50.0 mg, 58.2 μmol) was dissolved in trifluoroacetic acid (TFA) (500 μL) at 0 $^{\circ}\text{C}$, and a mixture of trifluoromethanesulfonic acid in TFA (500 μL , 2M) was added dropwise. The mixture was stirred for 12 minutes at 0 $^{\circ}\text{C}$ under continuous TLC monitoring. The reaction was rapidly quenched by adding diluted aqueous sodium bicarbonate solution (8 mL) and the product was extracted with ethyl acetate. The organic layer was washed with diluted brine (11 mL) and aqueous sodium bicarbonate solution (3 x 10 mL), dried with sodium sulfate and the solvent was removed under reduced pressure to yield the desired product in good purity (37.5 mg, 93%). The crude product could further be purified by column chromatography (chloroform : methanol = 9 :1, R_f = 0.4) or by preparative HPLC obtaining highest purity.

Epilogue

Having experimented in with both physical and virtual screening efforts for the discovery of natural product - target - disease associations, I feel encouraged to speculate on the future direction of natural products drug discovery. Large-scale collaborative efforts have connected world class experts in medicine, pharmacognosy, and numerous other fields to develop new-age screening methodologies to address human health concerns. These collaborations have brought together the advancements of cell-free and cell-based screening technologies, next-generative compound libraries, and cutting-edge analytical techniques to provide rich information on bioactive natural products. Furthermore, the assembly of structural-based virtual libraries of natural products coupled with rapidly developing computational screening technologies have provided drug discovery efforts with highly detailed structure-activity relationship. At the current moment, the use of both of these approaches in parallel could yield powerful improvements on our abilities to rapidly interrogate thousands of natural products for their bioactivity while simultaneously aggregating chemical information provided by their structure.

Our ability to characterize the bioactivity of natural products, with clinical relevance, is dependent on the structural elucidation – that of which presents a fairly time intensive bottleneck to drug discovery efforts. The current workflow for the structural elucidation of novel natural products incorporates the concurrent application of spectroscopic techniques that rely on inference of connectivity (NMR, HRMS, etc.). Although these techniques have produced a wealth of structural information for chemists in all fields, they often require high sample purity and do not produce unambiguous structural determination. Single crystal X-ray diffraction has monopolized chemists abilities to provide unequivocal structural information about position, orientation, connectivity, and placement of individual atoms and bonds within a given molecule. However, this technique has proven to be challenging for natural product application due to the amount of material necessary to form uniform crystals at the scale required for diffraction. A development in the utilization of the electron cry-microscopy (cryoEM) method microcrystal

electron diffraction (MicroED) presents a promising opportunity for natural product chemists to produce unambiguous structural determination – especially for the metabolites isolated in low yields, possessing multiple quaternary centers, or that are simply too difficult to purify.

With the recent development in our ability to produce unambiguous structural characterizations of natural products, I can envision the creation of structural database of NPs that is akin to the PDB. Employing such a database with the ever-evolving capabilities of machine learning algorithms and cheminformatic techniques may lay the foundation for the next generation of natural product screening. Lastly, associating the learned chemical descriptors of a given natural product in three-dimensional space with its relative bioactivity could potentially provide unrivaled virtual screening capabilities for the field of pharmacognosy. Nevertheless, the future of natural products chemistry continues to look bright, and I am excited to see what the next couple of decades may provide.

Appendix

NMR Tables and Spectra by Molecule

Ikarugamycin (1)

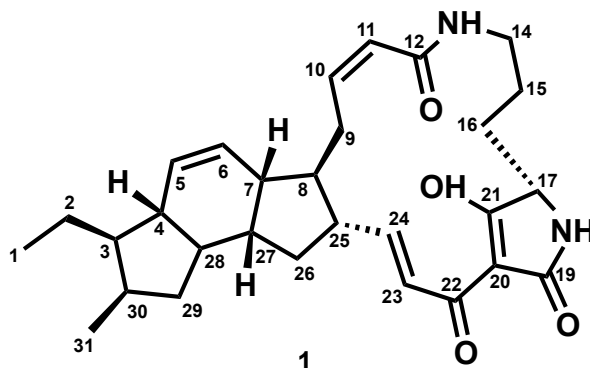


Table 2.1 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of 1.

Position	δH , Mult (<i>J</i> in Hz)	δC , Mult	Position	δH , Mult (<i>J</i> in Hz)	δC , Mult
1	0.98, t (7.0)	13.28	17	3.95, br s	61.32
2	1.41, m; 1.51, m	21.63	NH-18	6.15, br s	-
3	1.62, m	47.72	19	-	173.98
4	1.42, m	46.97	20	-	100.38
5	5.99, d (10.0)	131.59	21	-	195.81
6	5.74, dt	128.07	22	-	175.51
7	2.56, m	42.92	23	7.19, d (15.4)	122.18
8	1.20, m	48.3	24	6.83, dd (15.4, 10.6)	152.86
9	2.44, d (10.6); 3.52, m	25.33	25	2.57, m	49.51
10	6.11, td	141.13	26	1.29, m; 2.18, m	36.71
11	5.87, d (10.6)	123.98	27	2.12, m	41.76
12	-	166.31	28	1.62, m	48.6
NH-13	5.92, br s	-	29	0.74, ddd (12.0, 12.0, 6.8); 2.14, m	38.46
14	2.68, s; 3.74, br s	38.87	30	2.31, ddd (7.6, 7.6, 7.6)	33.05
15	1.29, m; 1.63, m	21.1	31	0.92, d (7.2)	17.71
16	1.87, m; 2.07, m	27.67			

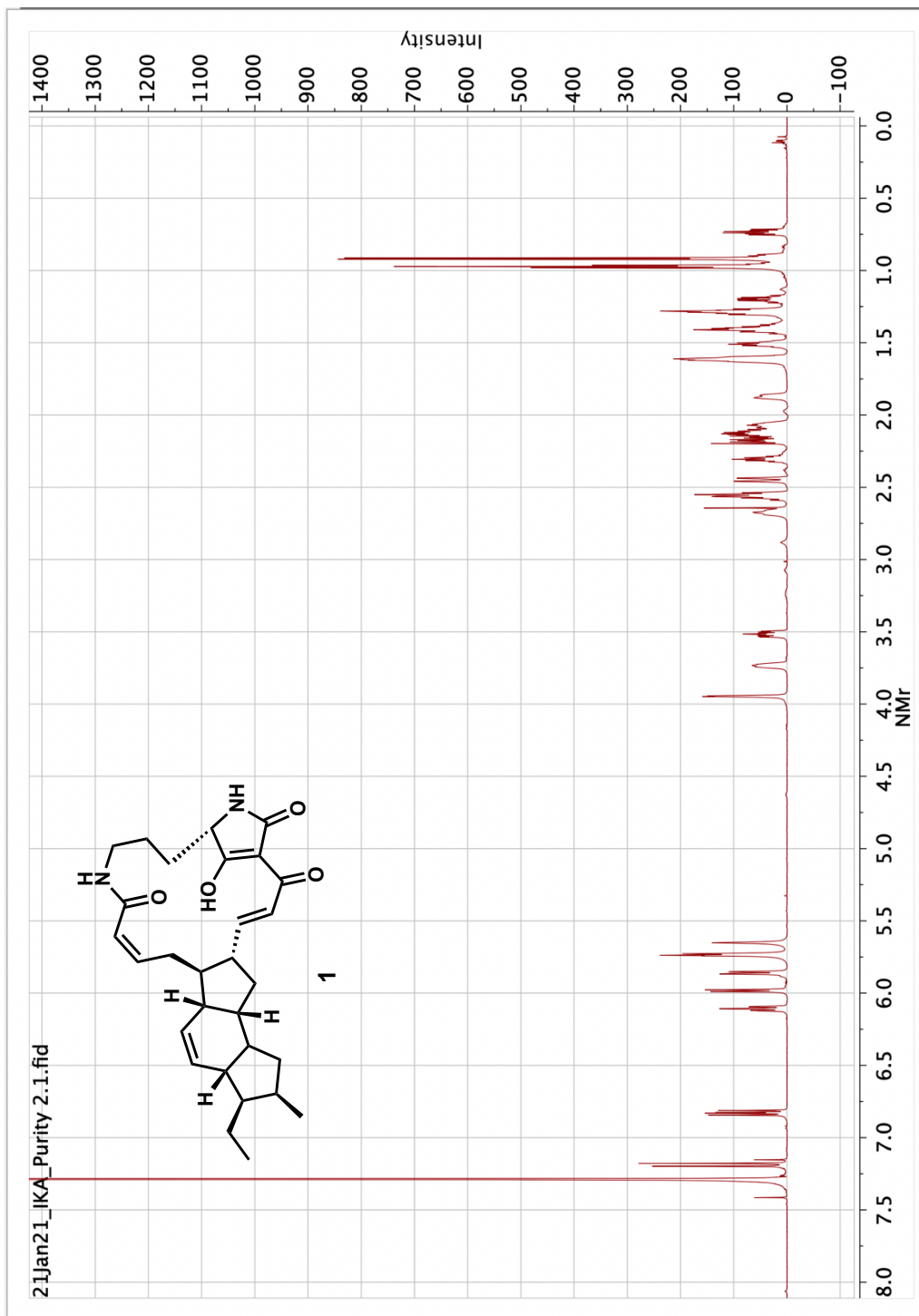


Figure 2.19 Ikarugamycin (1).

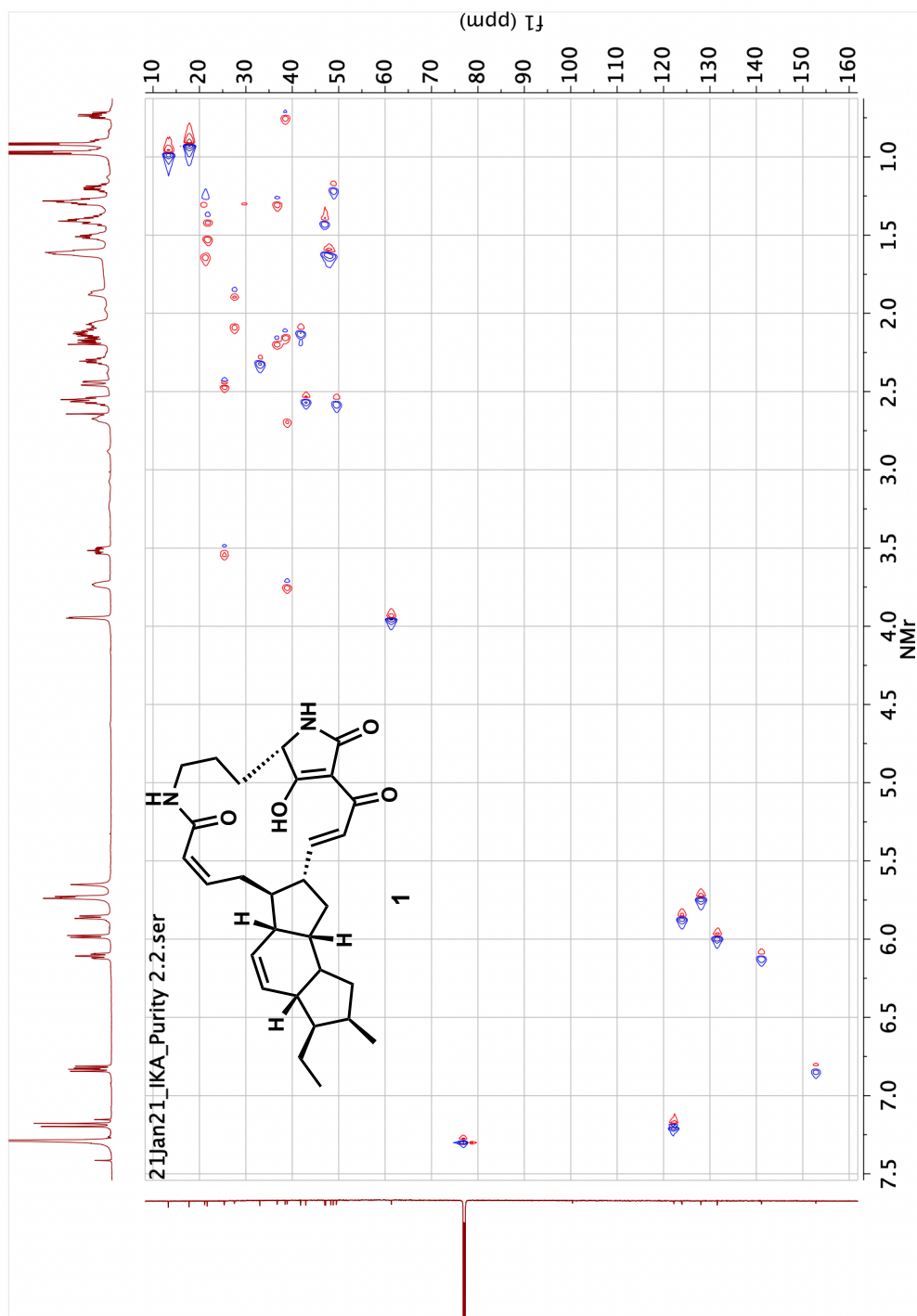


Figure 2.20 Ikarugamycin (1).

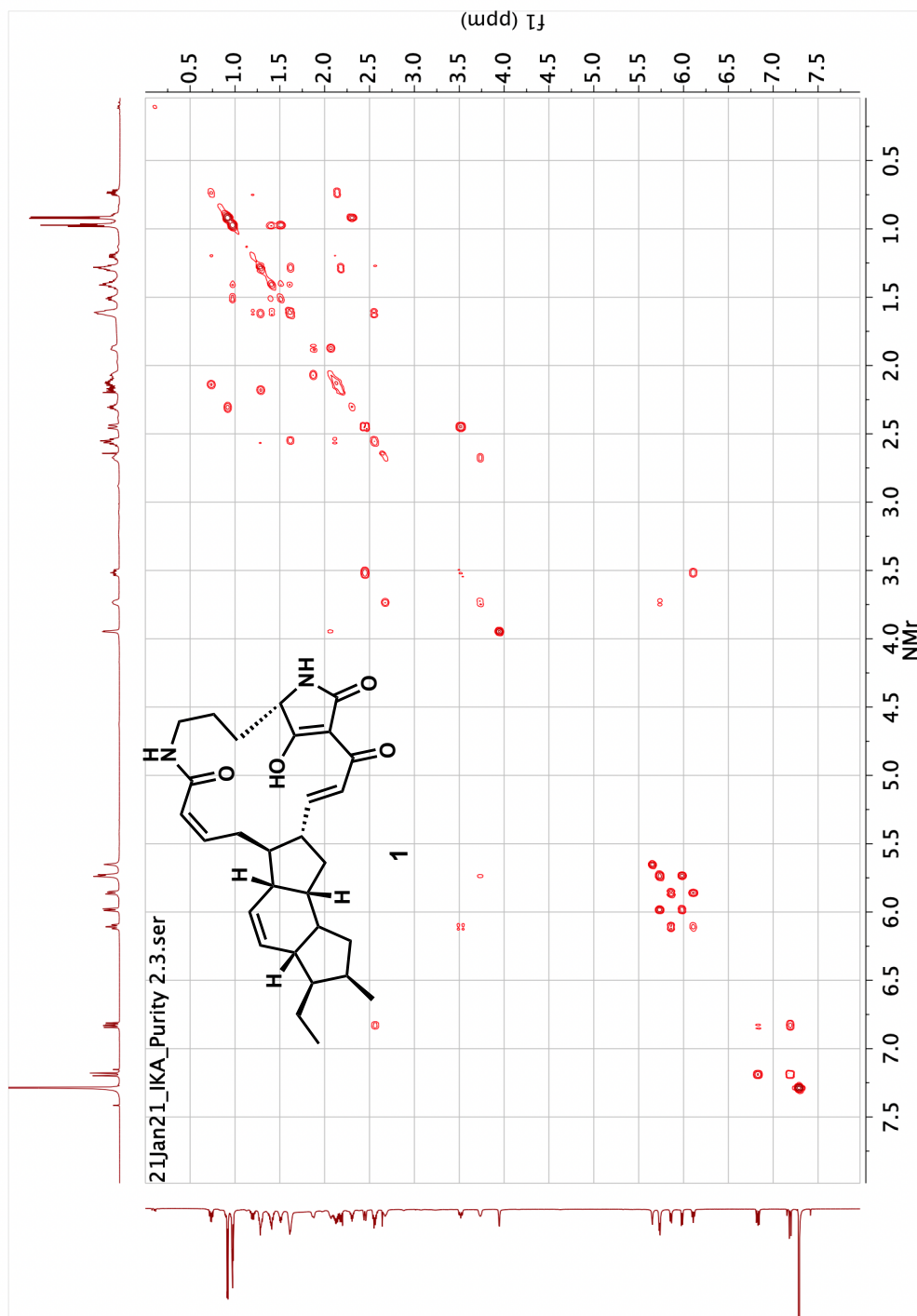


Figure 2.21 Ikarugamycin (1).

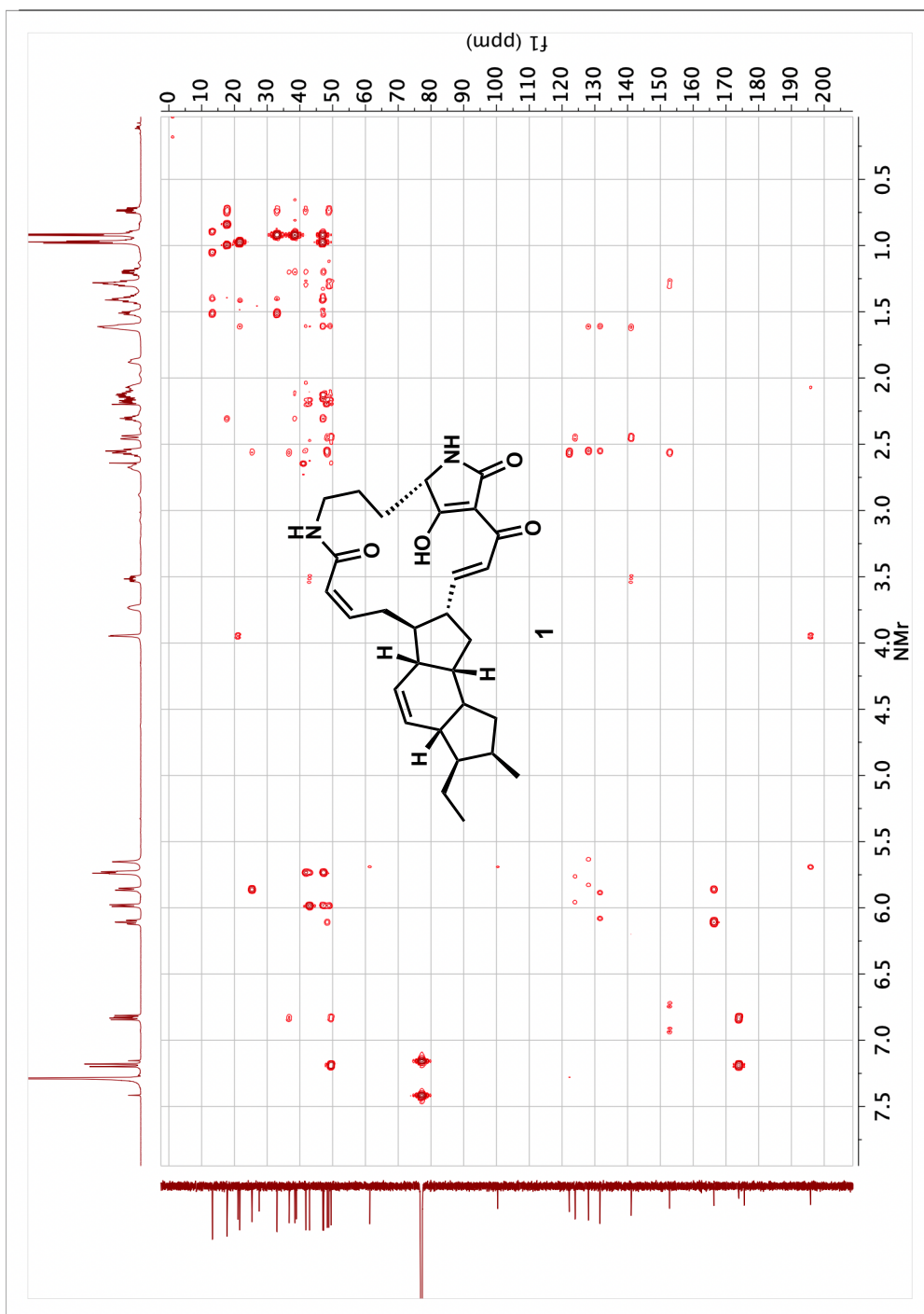


Figure 2.22 Ikarugamycin (1).

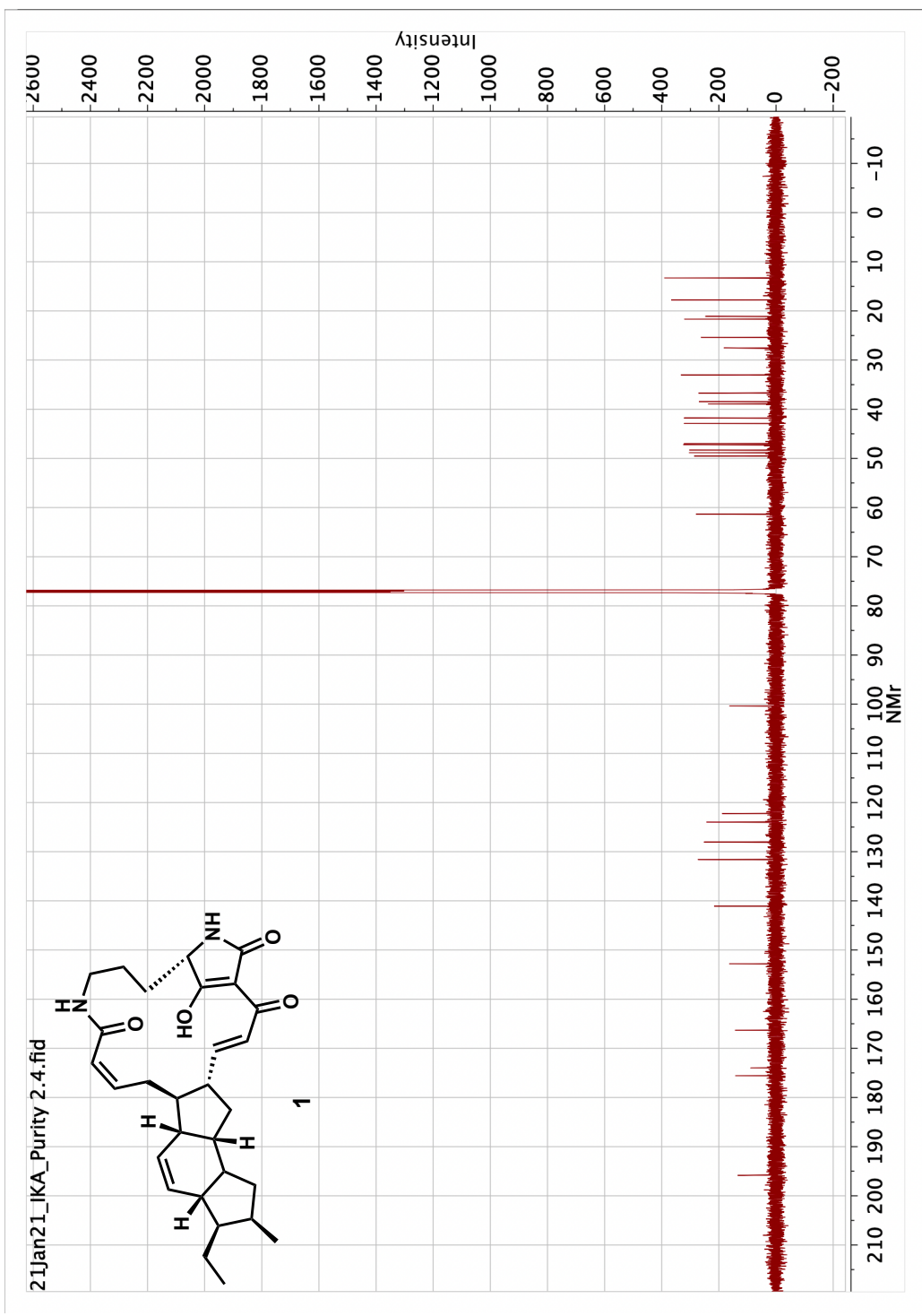


Figure 2.23 Ikarugamycin (1).

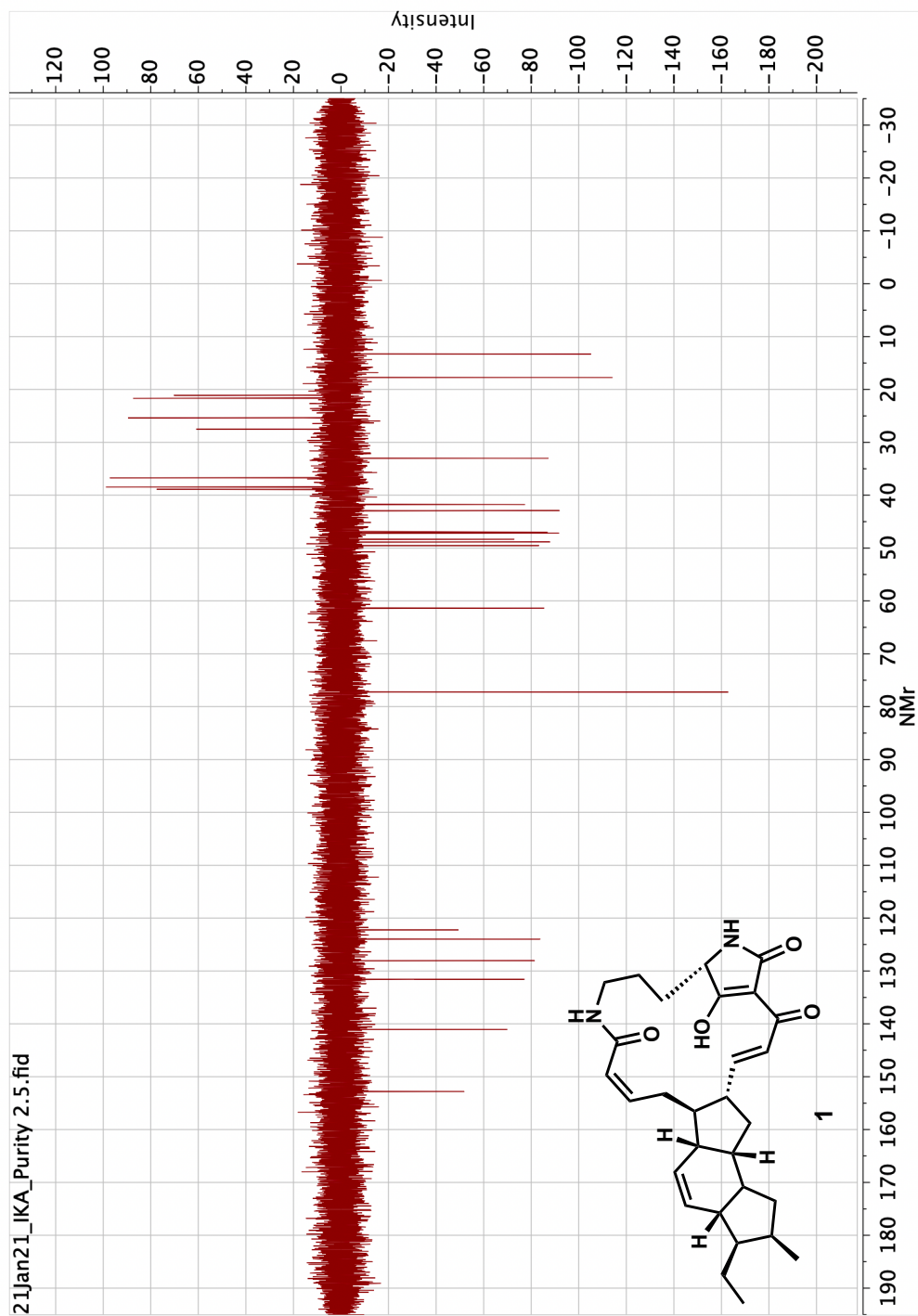


Figure 2.24 Ikarugamycin (1).

Capsimycin D (2)

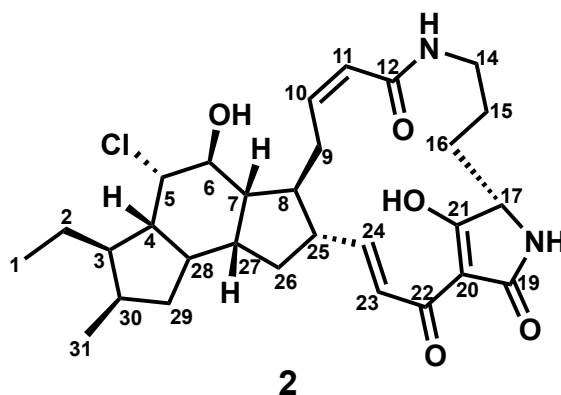


Table 2.2 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **2**.

Position	δH , Mult (<i>J</i> in Hz)	δC , Mult	Position	δH , Mult (<i>J</i> in Hz)	δC , Mult
1	0.94, t (7.3)	12.8	17	3.88, dd (5.5, 2.1)	61.6
2	1.35, m	21.2	NH-18	-	
3	1.76, d (3.3)	44.7	19		173.65
4	1.77, m	47	20		100.8
5	3.13, dd (3.8, 2.0)	57.7	21		197.1
6	2.89, d (3.8)	53.5	22		175.6
7	2.07, m	47.3	23	7.13, d (15.4)	122.18
8	2.14, m	45.6	24	6.83, dd (15.4, 10.6)	153
9	2.53, dd (17.3, 3.0); 3.38, m	26.4	25	2.57, m	49.51
10	6.06, ddd (11.5, 11.5, 3.4)	141.6	26	1.29, m; 2.13, m	35.6
11	5.84, dd (11.5, 1.3)	123.7	27	2.07, m	42.6
12		167.2	28	1.61, m	41.1
NH-13	-		29	0.75, ddd (12.0, 12.0, 6.8); 2.19, d (7.6)	38.6
14	2.65, br t (11.2); 3.55, ddd (11.2, 4.9, 3.0)	39	30	2.21, m	32.6
15	1.29, m; 1.63, m	21.1	31	0.92, d (6.8)	17.71
16	1.18m, 2.05 m	27.5			

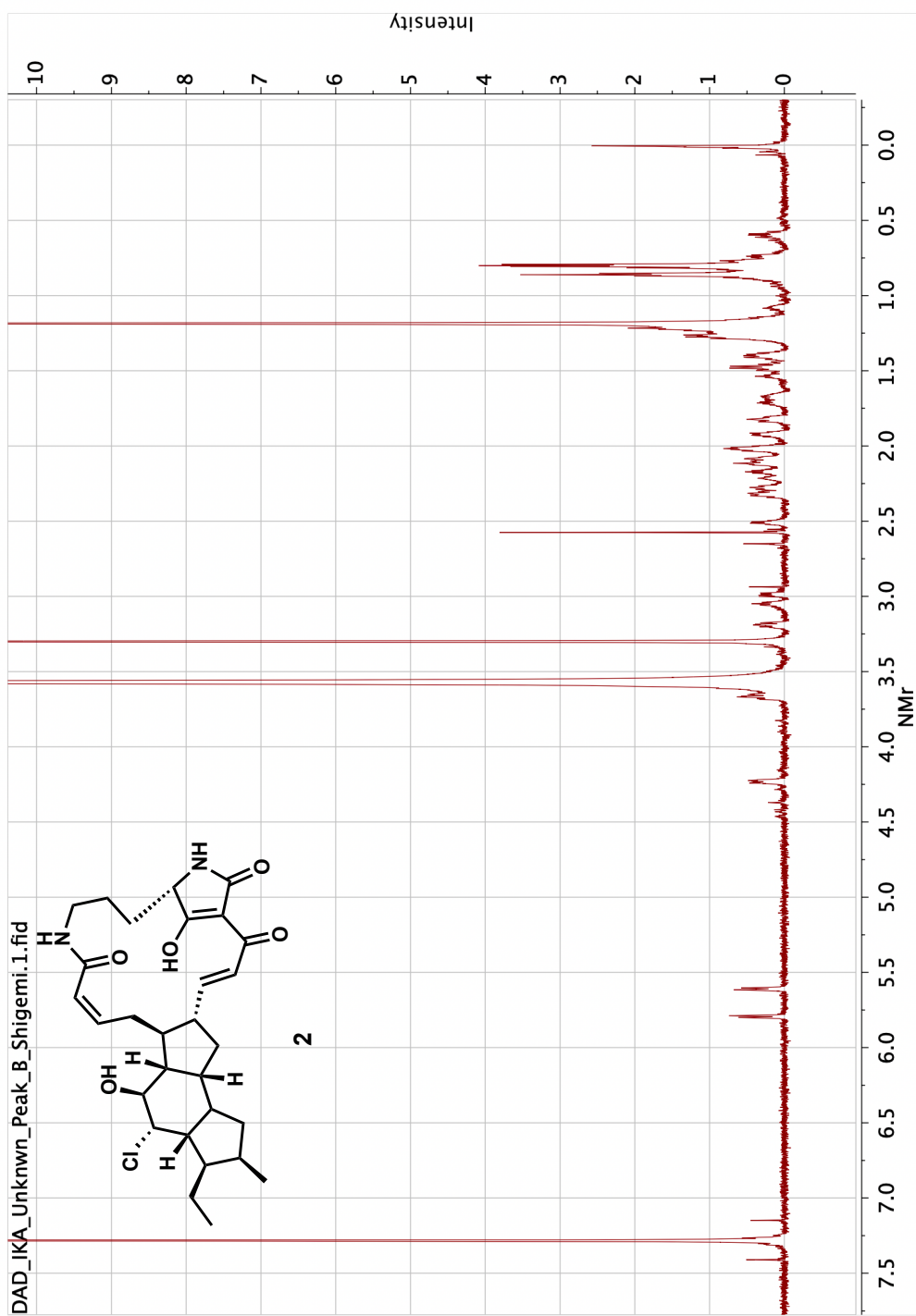
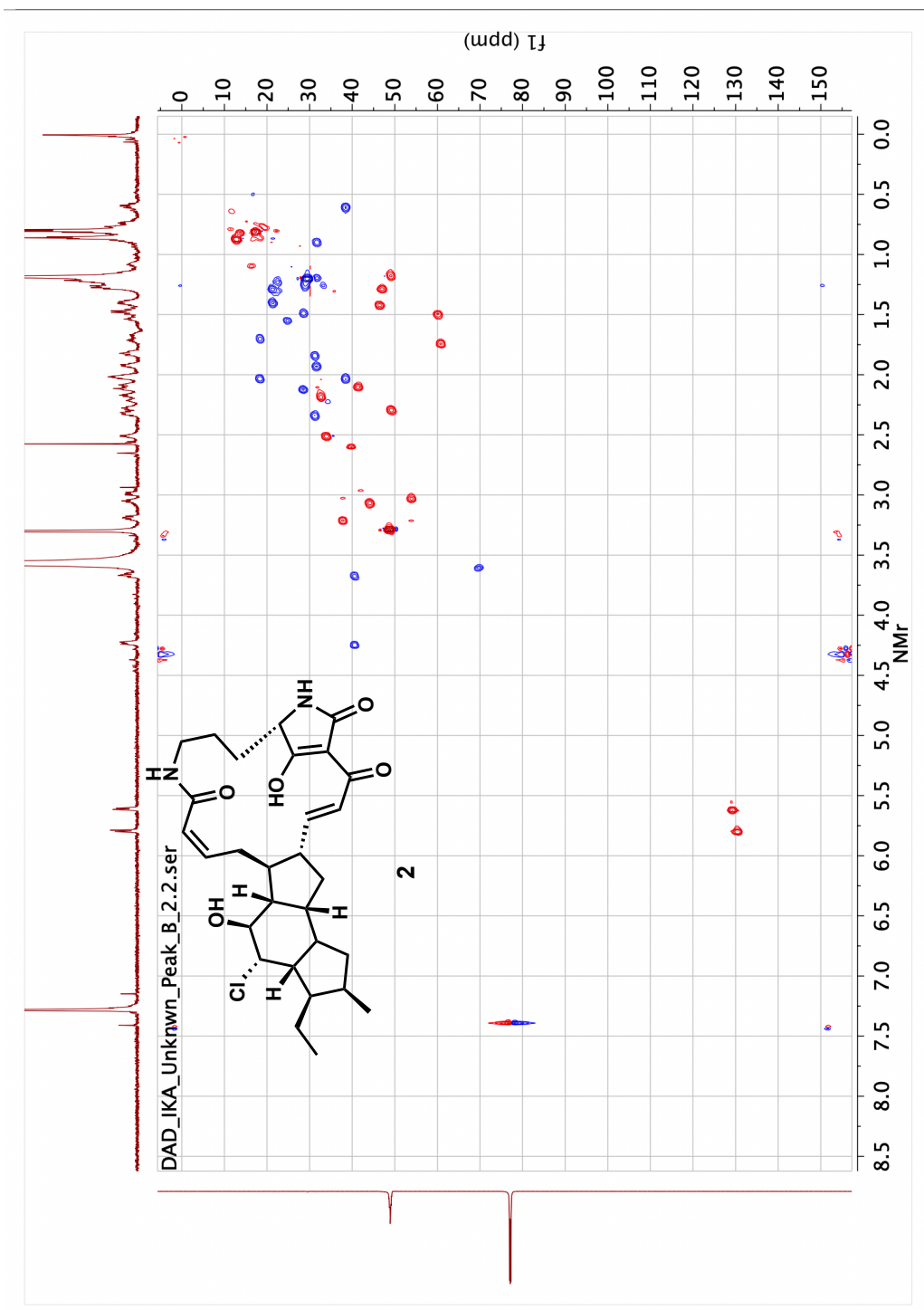


Figure 2.25 Capsimycin D (2).



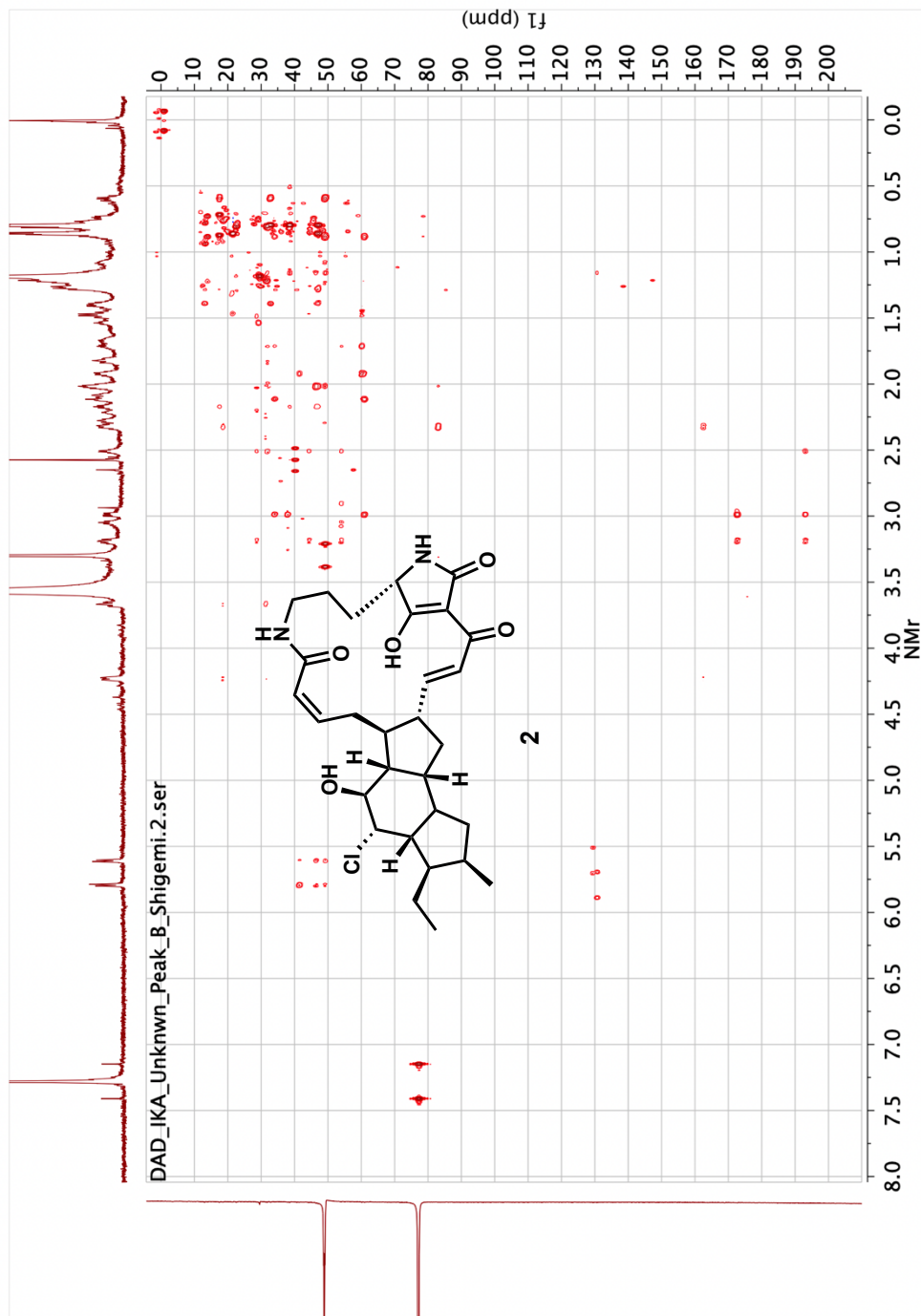


Figure 2.27 Capsimycin D (2).

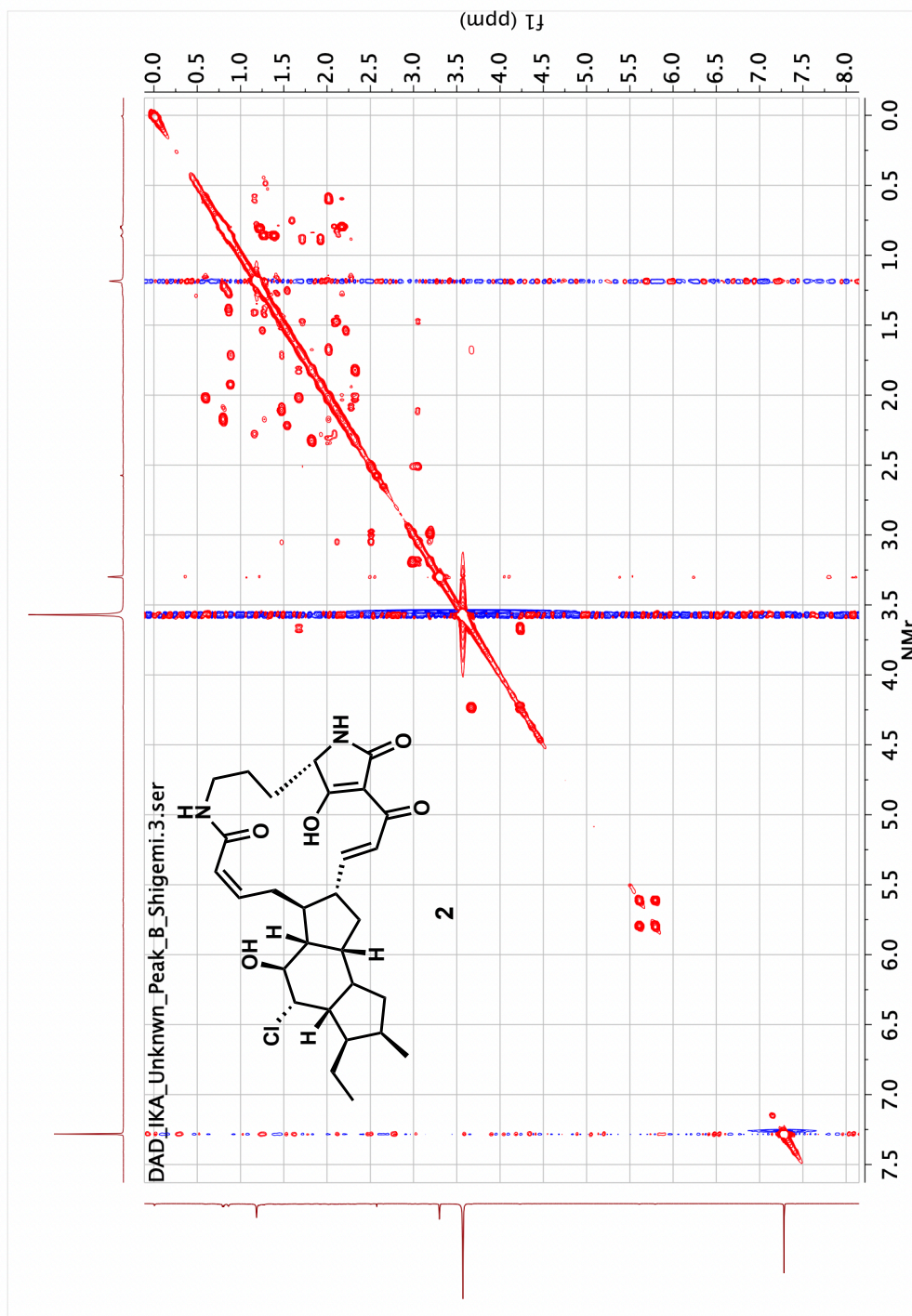


Figure 2.28 Capsimycin D (2).

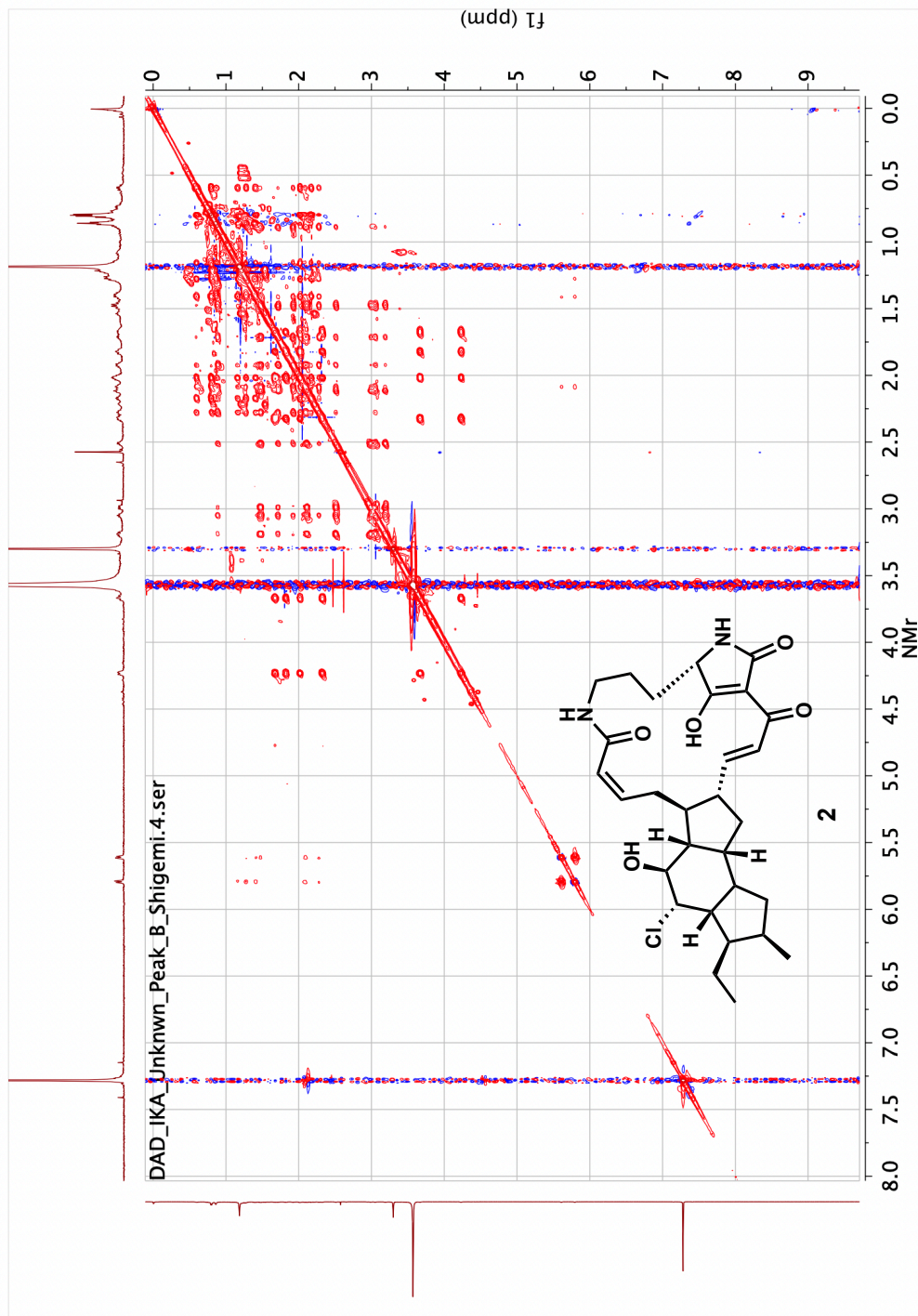
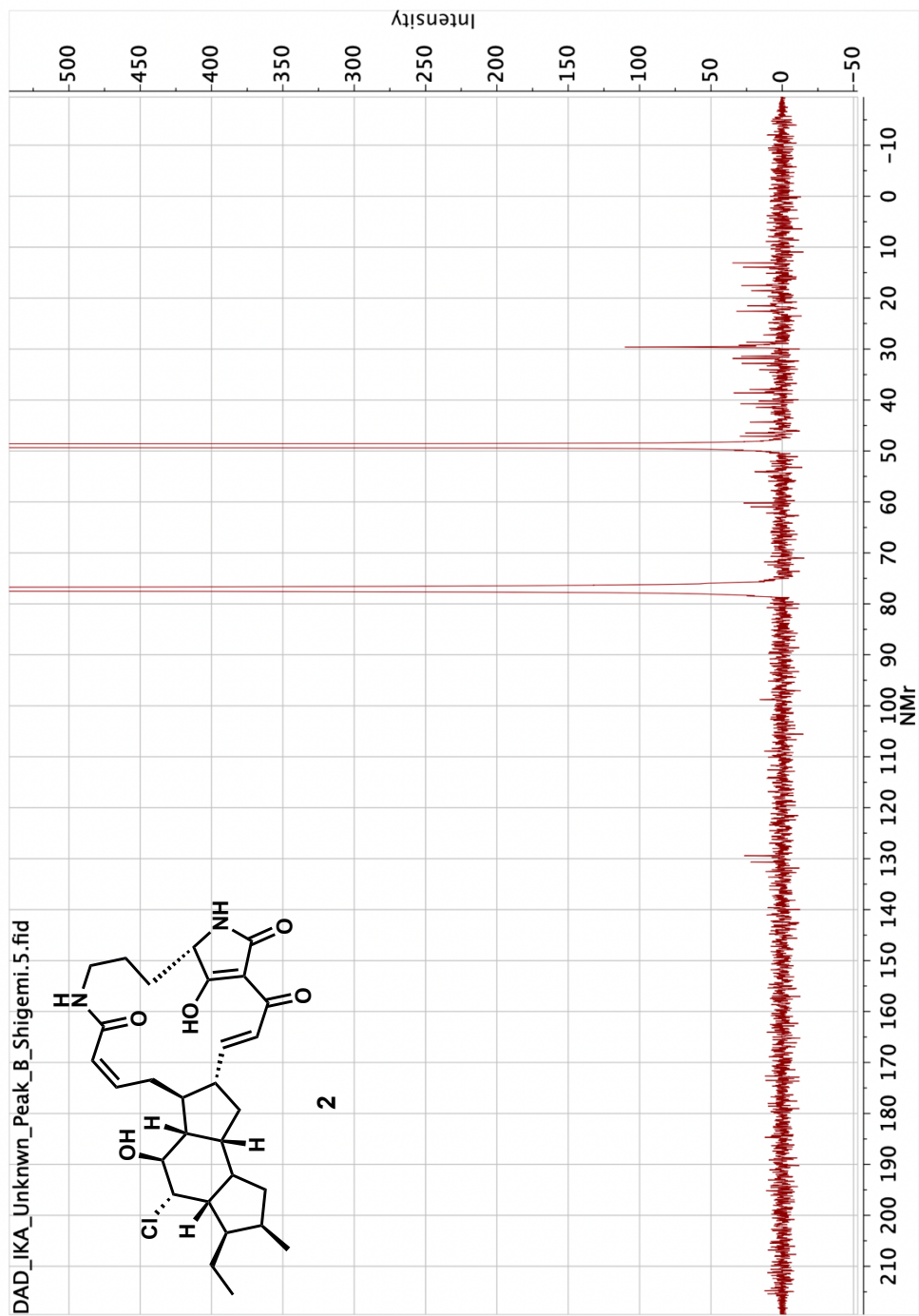


Figure 2.29 Capsimycin D (2).



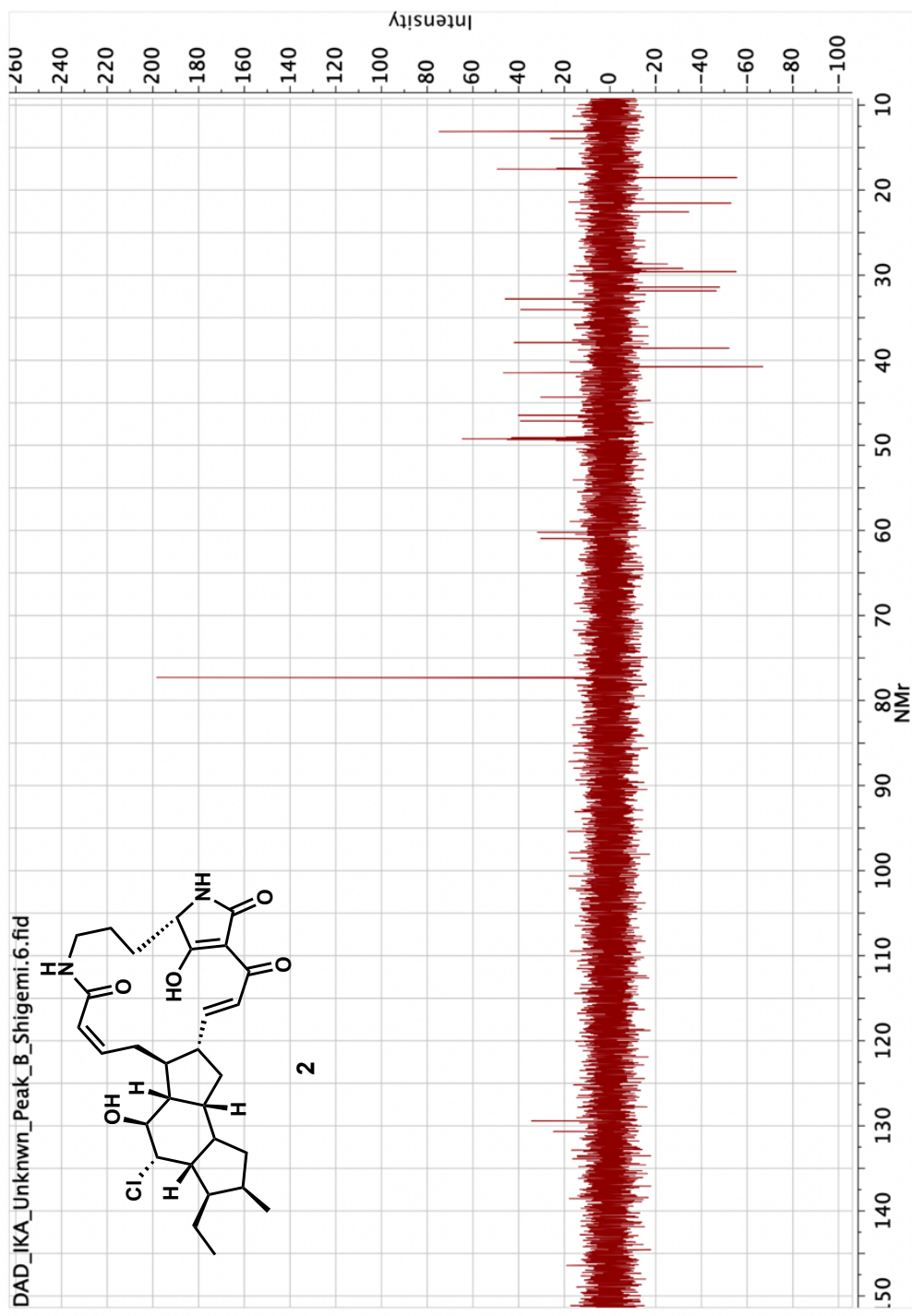


Figure 2.31 Capsimycin D (2).

Capsimycin B (3)

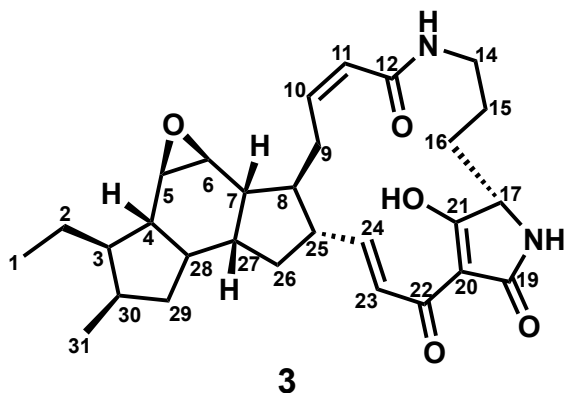


Table 2.3 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **3**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	0.98, t (7.0)	13.28	17	3.95, br s	61.32
2	1.41, m; 1.51, m	21.63	NH-18	6.15, br s	-
3	1.62, m	47.72	19	-	173.98
4	1.42, m	46.97	20	-	100.38
5	5.99, d (10.0)	131.59	21	-	195.86
6	5.74, dt	128.07	22	-	175.51
7	2.56, m	42.92	23	7.11, d (15.4)	122.68
8	1.20, m	48.3	24	6.73, dd (15.4, 10.6)	151.86
9	2.44, d (10.6); 3.52, m	25.33	25	2.57, m	49.51
10	6.06, td	140.33	26	1.29, m; 2.18, m	36.71
11	5.79, d (10.6)	124.18	27	2.12, m	41.76
12	-	166.31	28	1.62, m	48.6
NH-13	5.92, br s	-	29	0.74, ddd (12.0, 12.0, 6.8); 2.14, m	38.46
14	2.68, s; 3.74, br s	38.87	30	2.31, ddd (7.6, 7.6, 7.6)	33.05
15	1.29, m; 1.63, m	21.1	31	0.92, d (7.2)	17.71
16	1.87, m; 2.07, m	27.67			

Measured in $\text{CDCl}_3/\text{CD}_3\text{OD}$. δ values given in ppm.

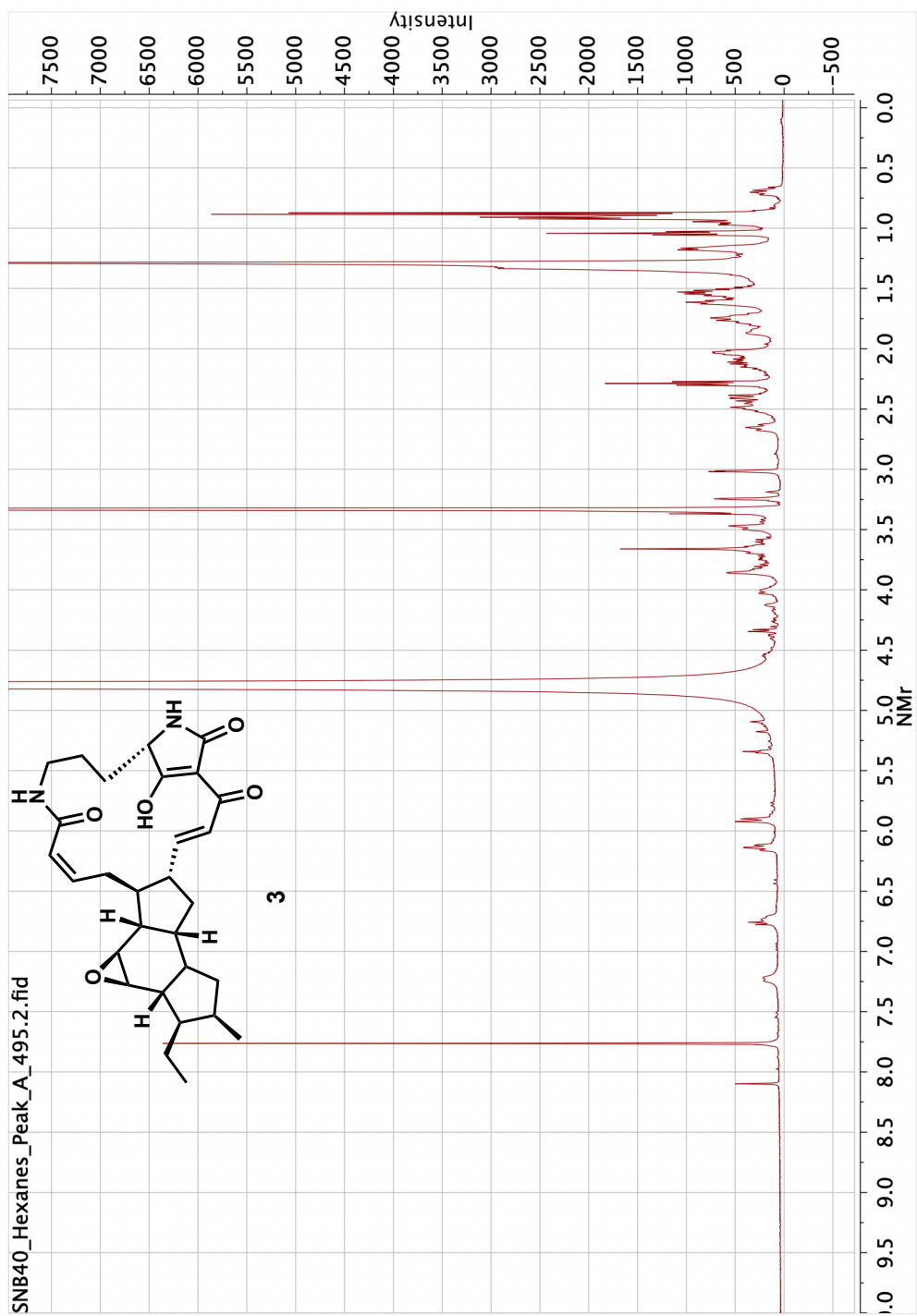


Figure 2.32 Capsimycin B (3).

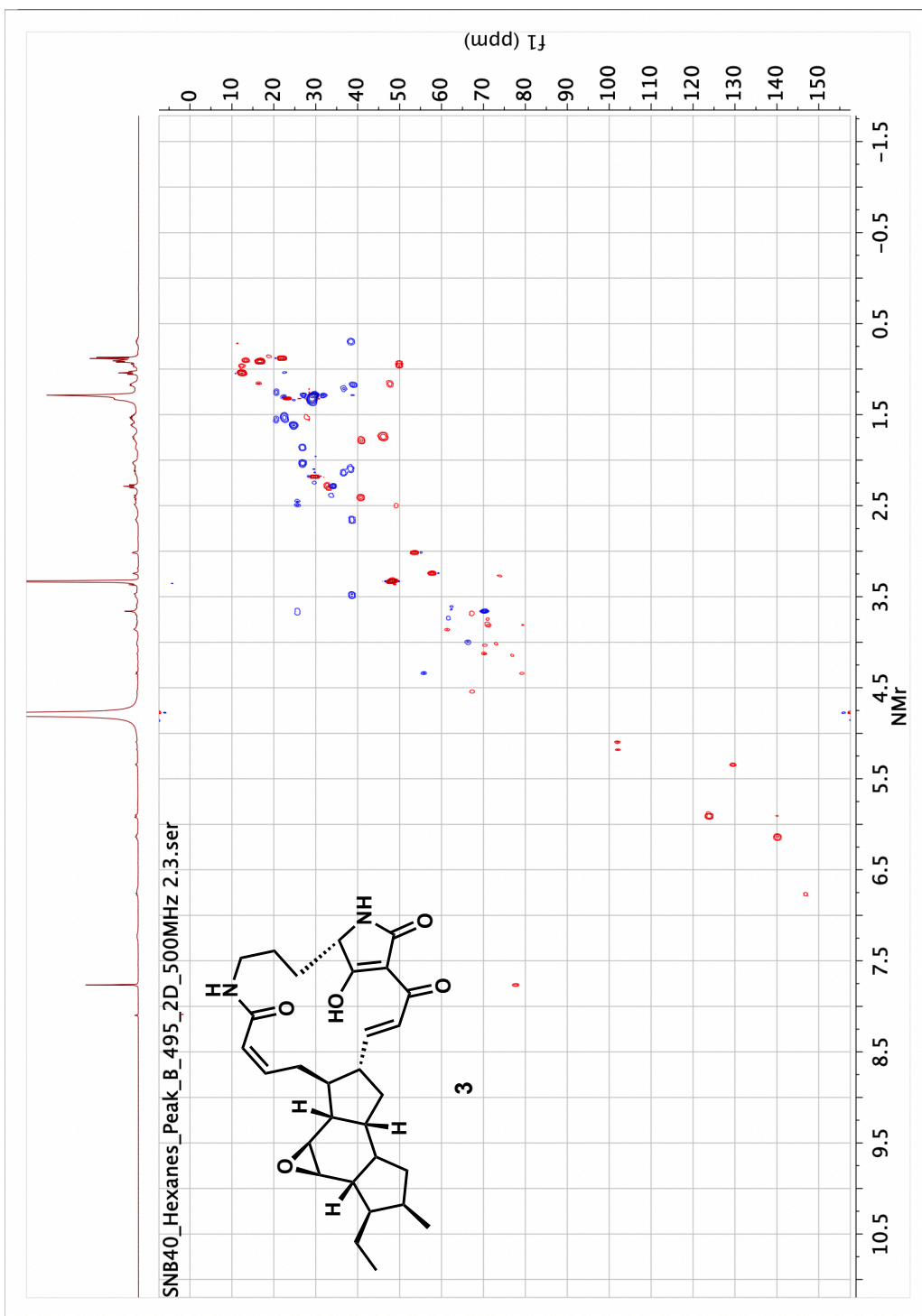


Figure 2.33 Capsimycin B (3).

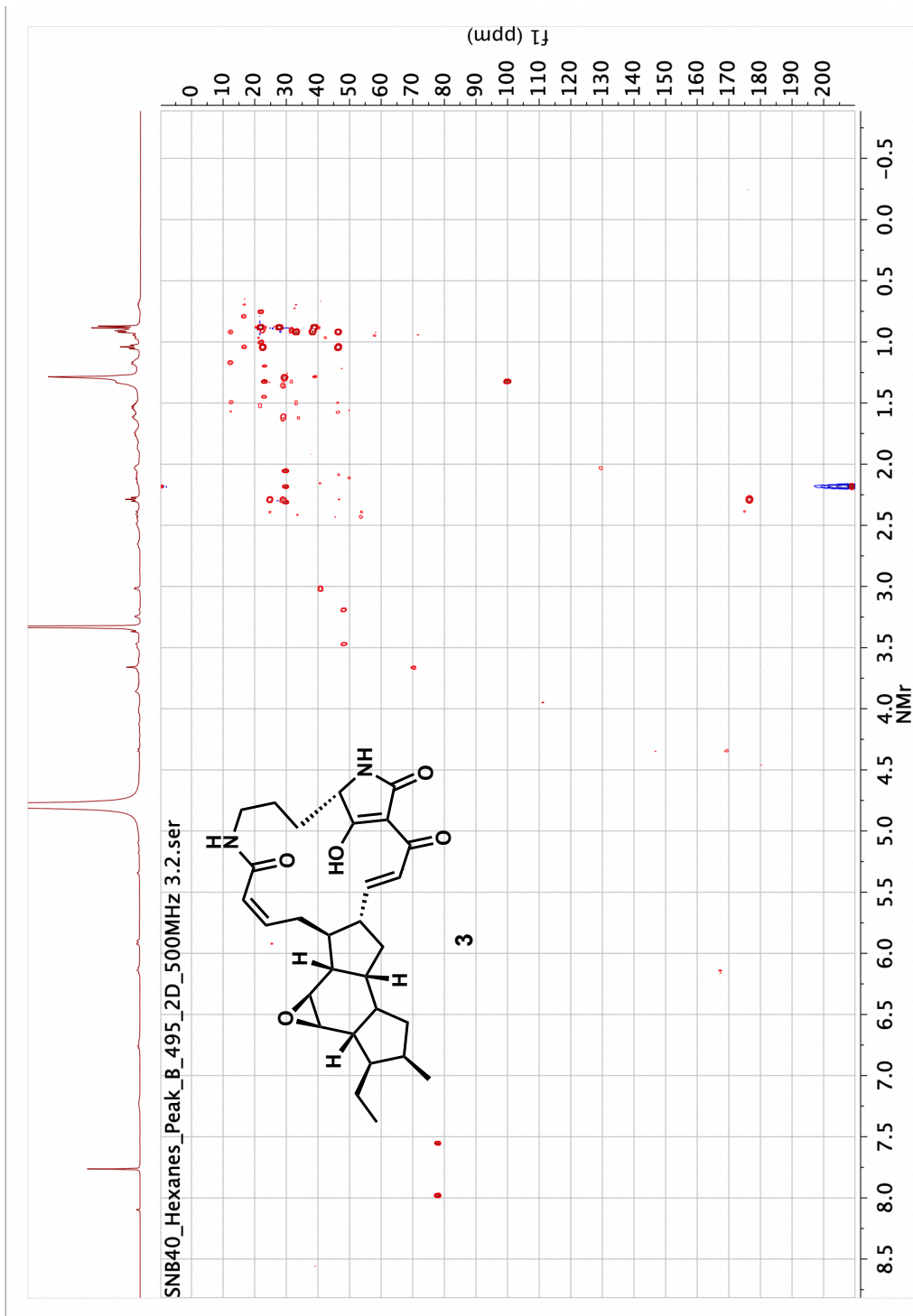


Figure 2.34 Capsimycin B (3).

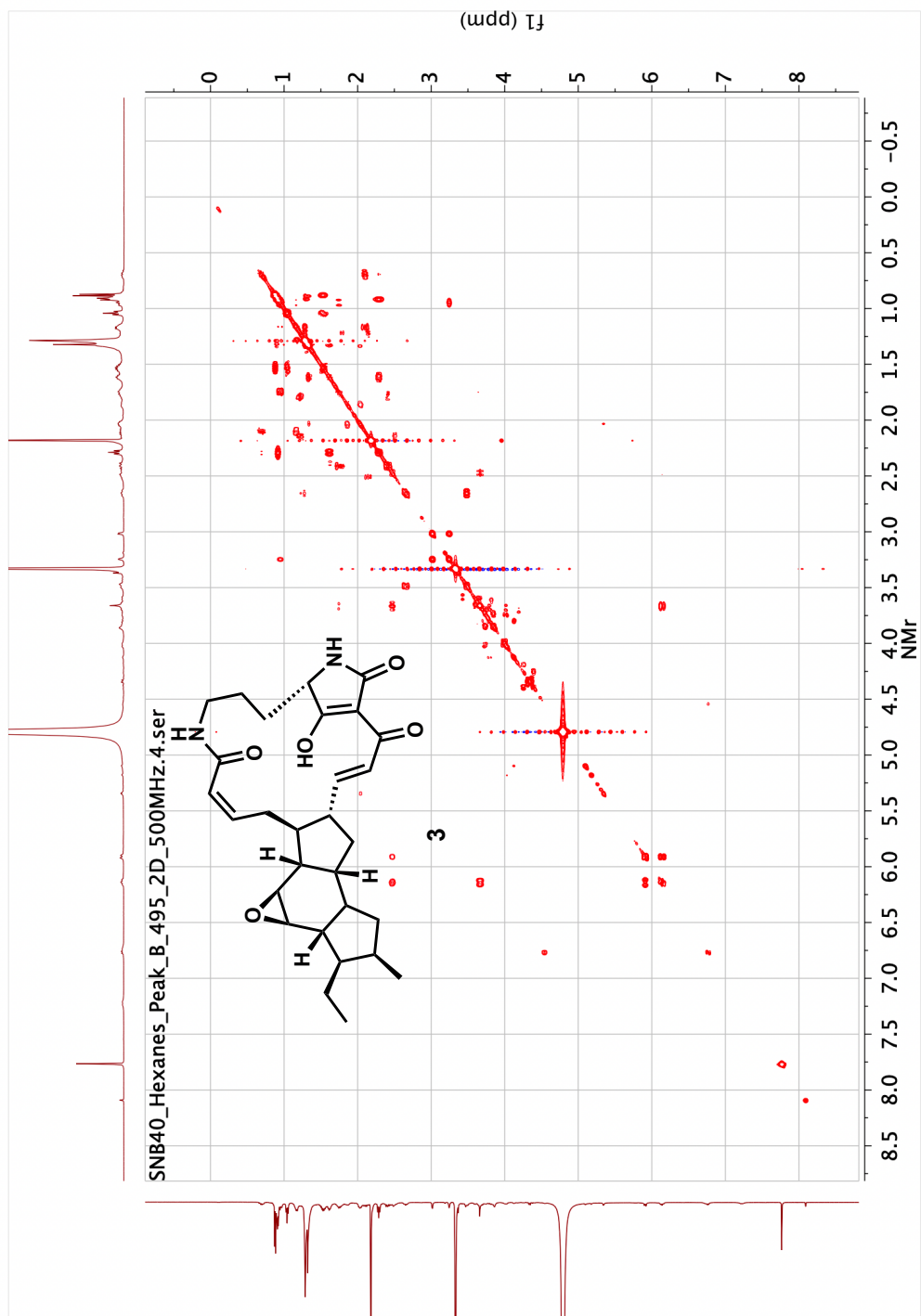


Figure 2.35 Capsimycin B (3).

Capsimycin F (4)

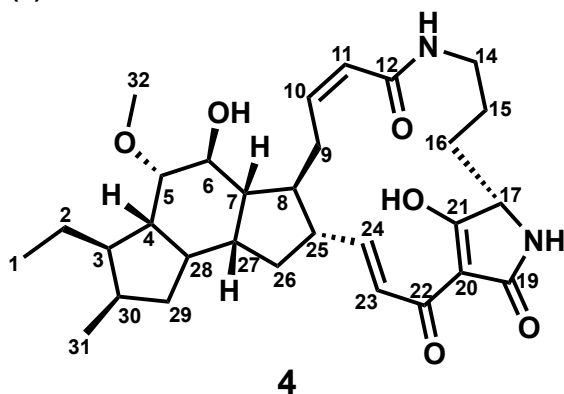


Table 2.4 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **4**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	0.96 (t, $J = 7.3$)	12.6	17	3.86, s	61.9
2	1.39 m	21.7	NH-18	-	-
3	1.75 m	42.9	19	-	175.9
4	1.48 m	46.7	20	-	101.1
5	3.40 m	81.5	21	-	197.2
6	4.09 m	67.8	22	-	173.0
7	2.03 m	47.9	23	7.40 (d, $J = 15.4$)	122.4
8	2.03 m	44.8	24	6.76 (dd, $J = 15.4, 10.3$)	152.3
9	3.49 m, 2.47 m	26.7	25	2.38 m	50.3
10	6.07 (td, $J = 11.3, 3.8$)	141.7	26	2.08 m, 1.25 m	35.3
11	5.85 (d, $J = 11.5$)	123.7	27	1.55 m	42.0
12	-	167.9	28	2.02 m	42.0
NH-13	-	-	29	2.16 m, 0.67 m	39.3
14	3.45 m, 2.67 (t, $J = 10.9$)	39.0	30	2.17 m	32.8
15	2.01 m, 1.97 m	21.0	31	0.90 (d, $J = 6.4$)	17.0
16	2.01 m, 1.87 m	27.4	32		58.0

Measured in CD_3OD . δ values given in ppm.

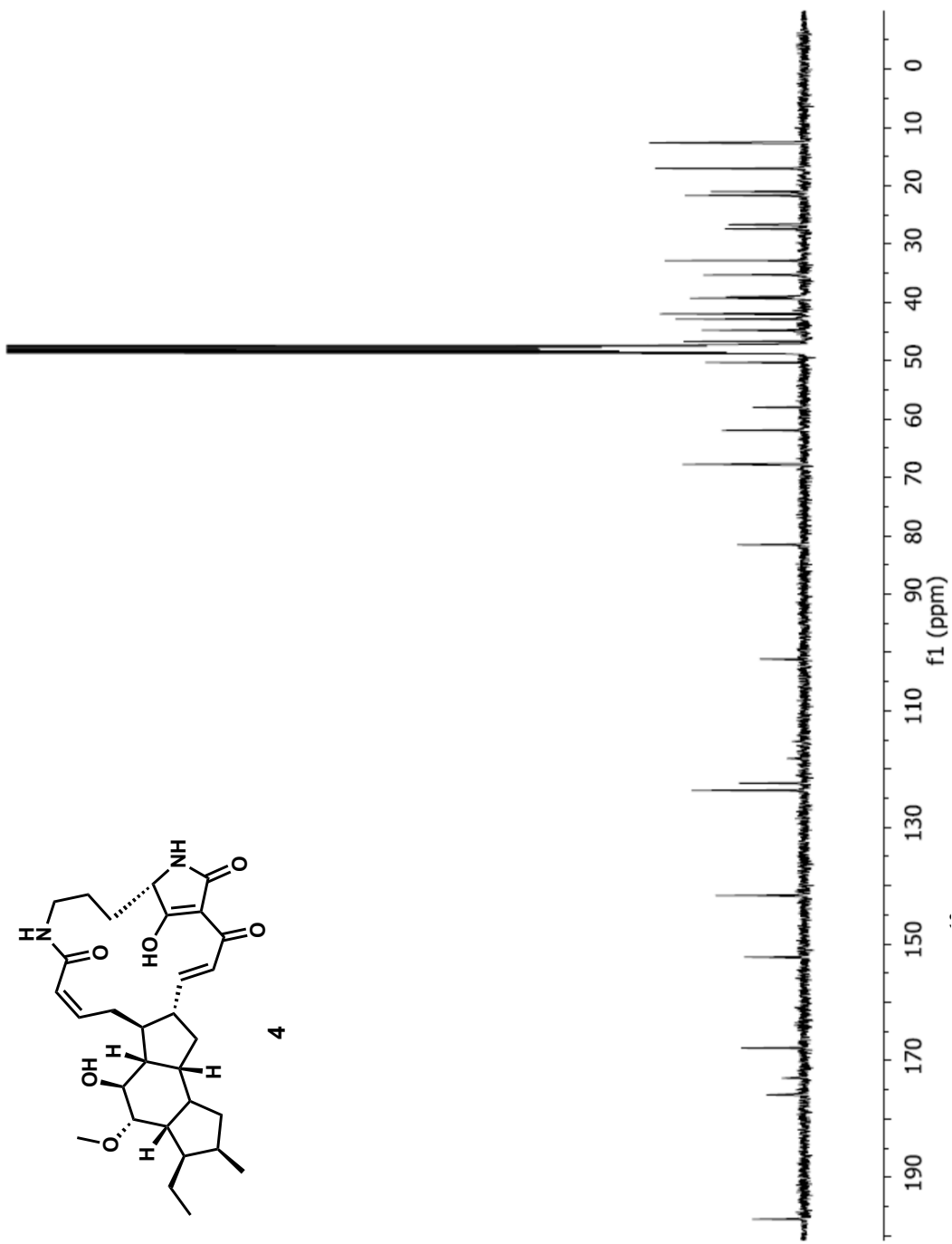


Figure 2.37 Capsimycin F (4).

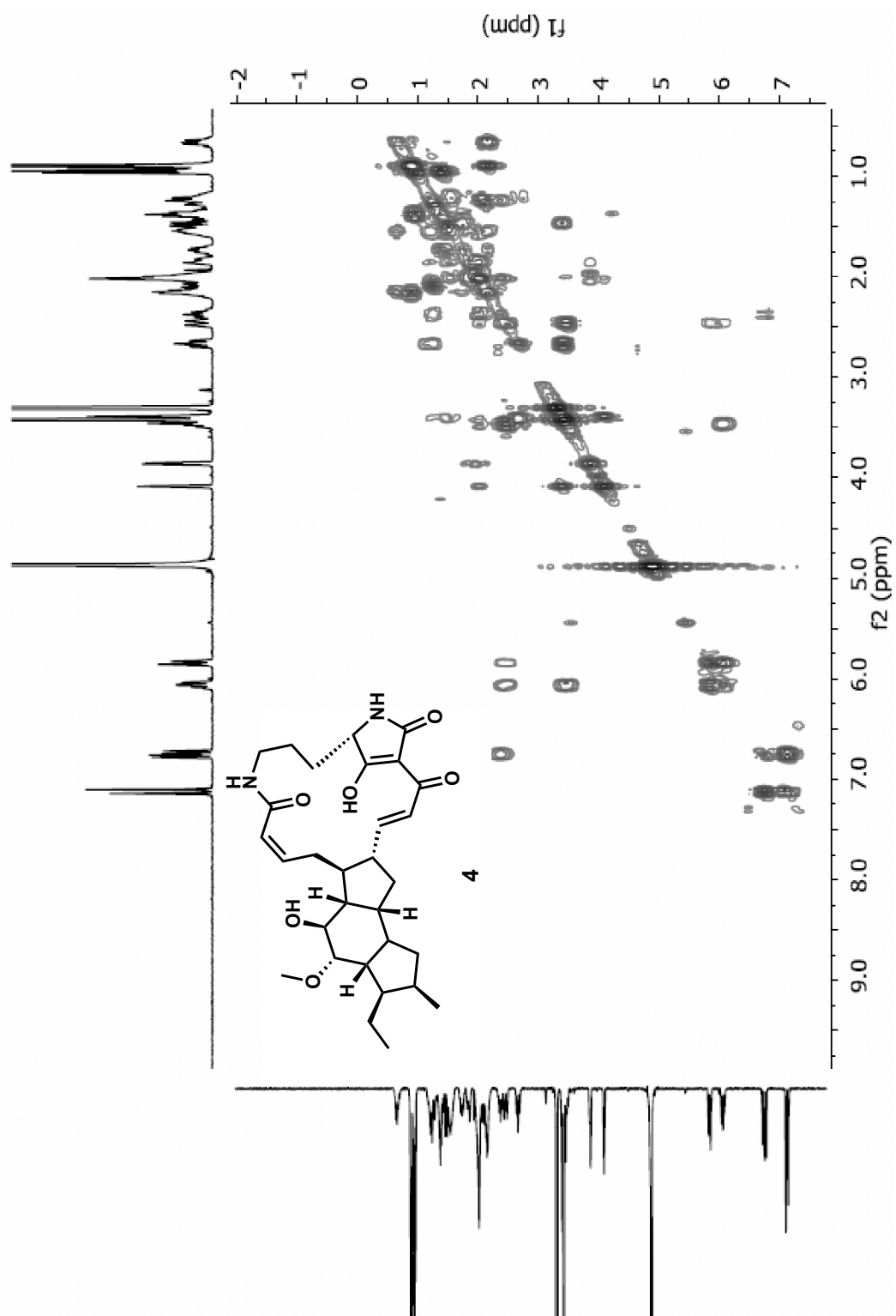


Figure 2.38 Capsimycin F (4).

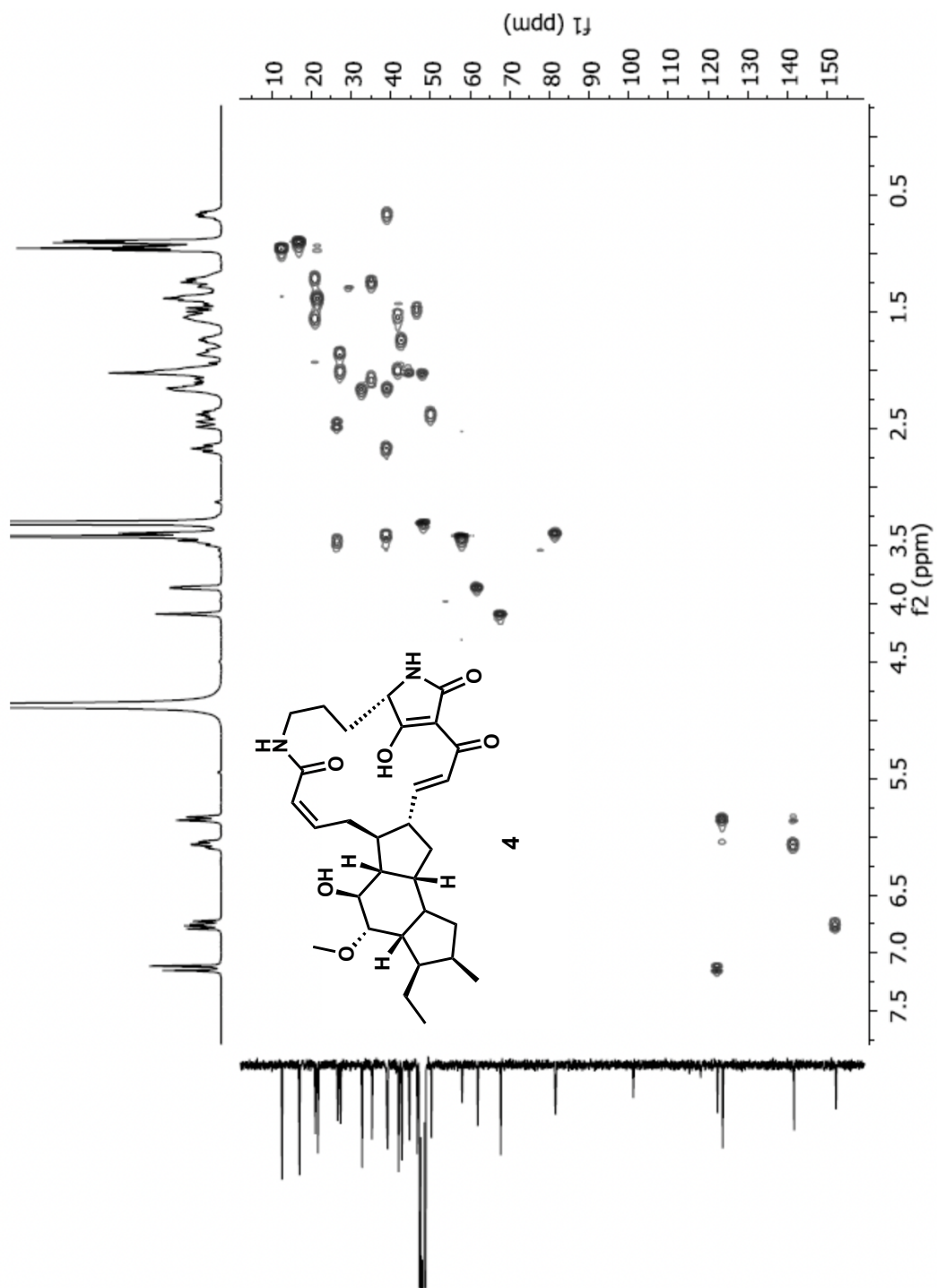


Figure 2.39 Capsimycin F (4).

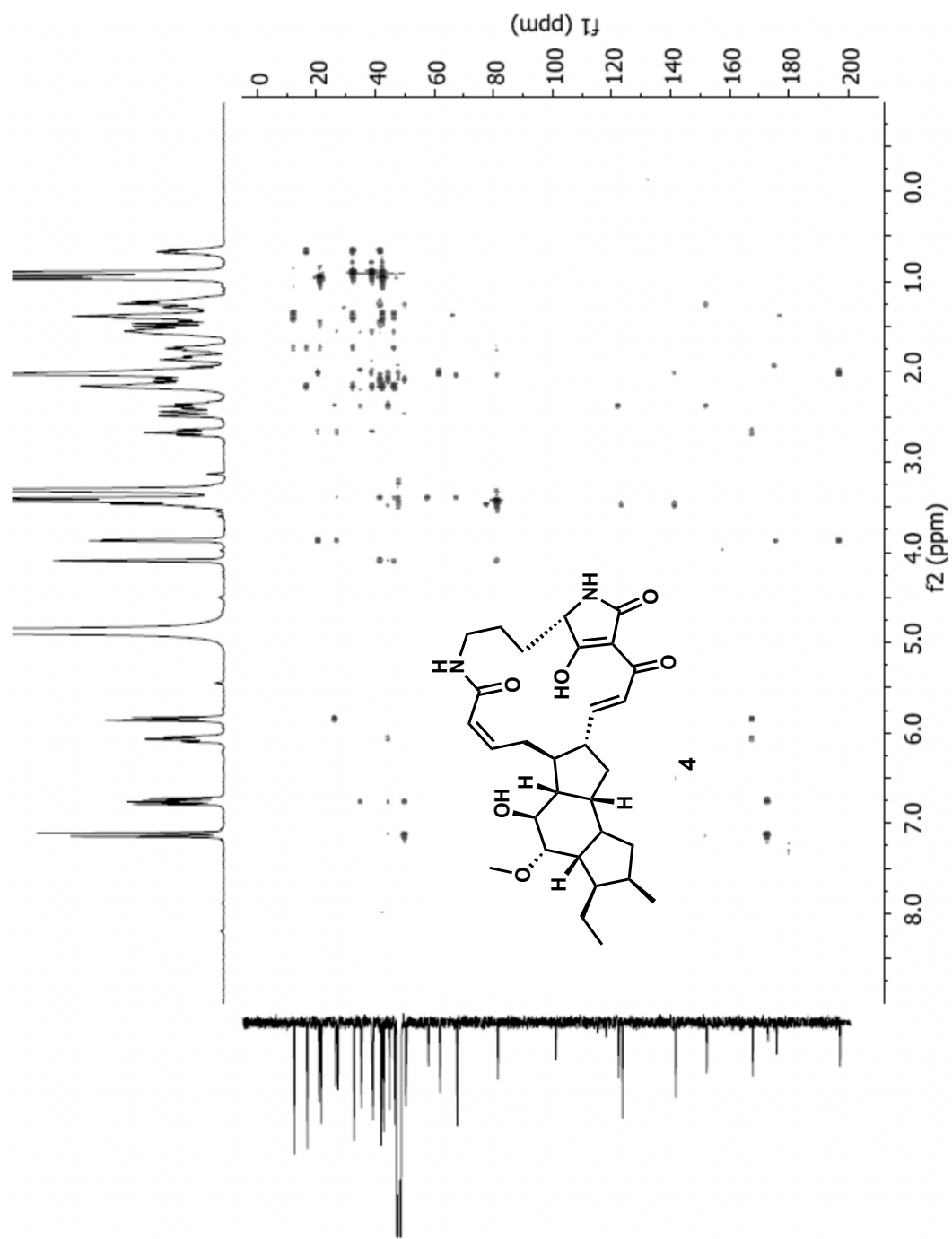


Figure 2.40 Capsimycin F (4).

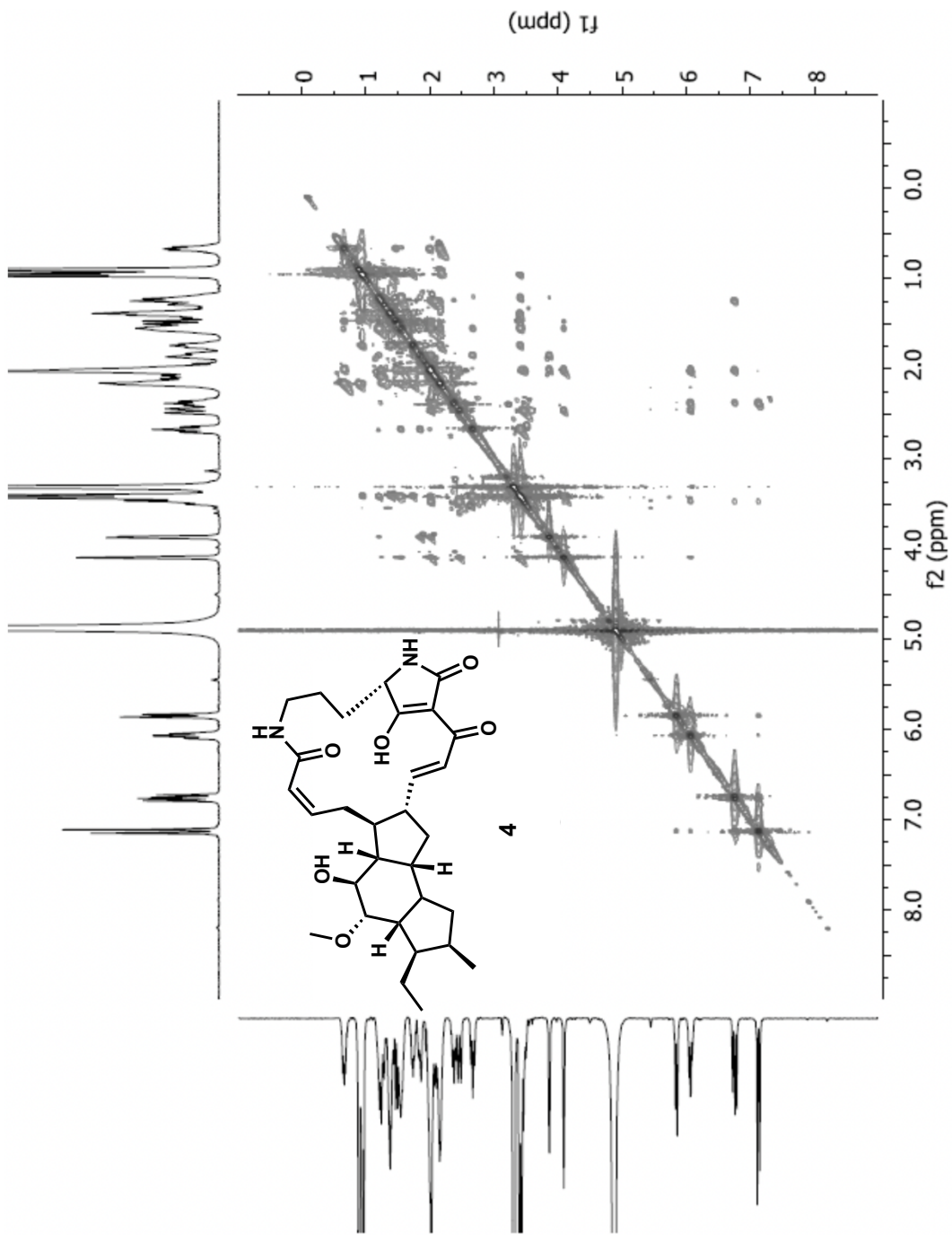


Figure 2.41 Capsimycin F (4).

Capsimycin C

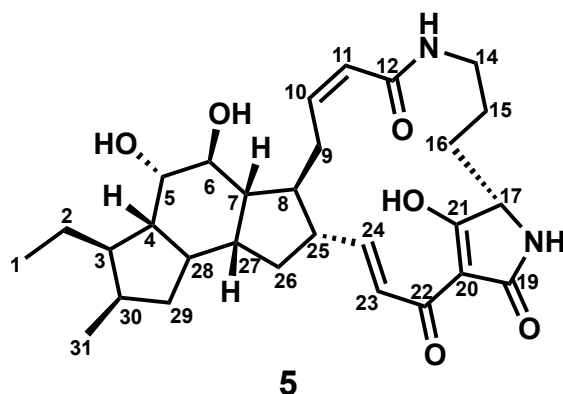


Table 2.5 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **5**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	0.96 (t, $J = 7.3$)	12.4	17	3.85, m	61.7
2	1.36 m	21.7	NH-18	-	-
3	1.78 m	42.9	19	-	175.9
4	1.44 m	45.5	20	-	101.1
5	3.85 m	71.6	21	-	197.3
6	3.87 m	73.3	22	-	173.0
7	1.99 m	47.9	23	7.13 (d, $J = 15.4$)	122.4
8	2.08 m	44.5	24	6.79 (dd, $J = 15.4, 10.3$)	152.3
9	3.47 m, 2.53 m	26.8	25	2.39 m	50.2
10	6.09 (td, $J = 1.3, 3.9$)	142.0	26	2.16 m, 1.23 m	35.8
11	5.84 (d, $J = 11.3$)	123.5	27	2.02 m	42.7
12	-	168.0	28	1.60 m	41.2
NH-13	-	-	29	2.18 m, 0.70 m	39.3
14	3.41 m, 2.67 m	39.0	30	2.21 m	33.1
15	1.48 m, 1.19 m	21.1	31	0.91 (d, $J = 6.6$)	17.0
16	1.99 m, 1.86 m	27.4			

Measured in CD_3OD . δ values given in ppm.

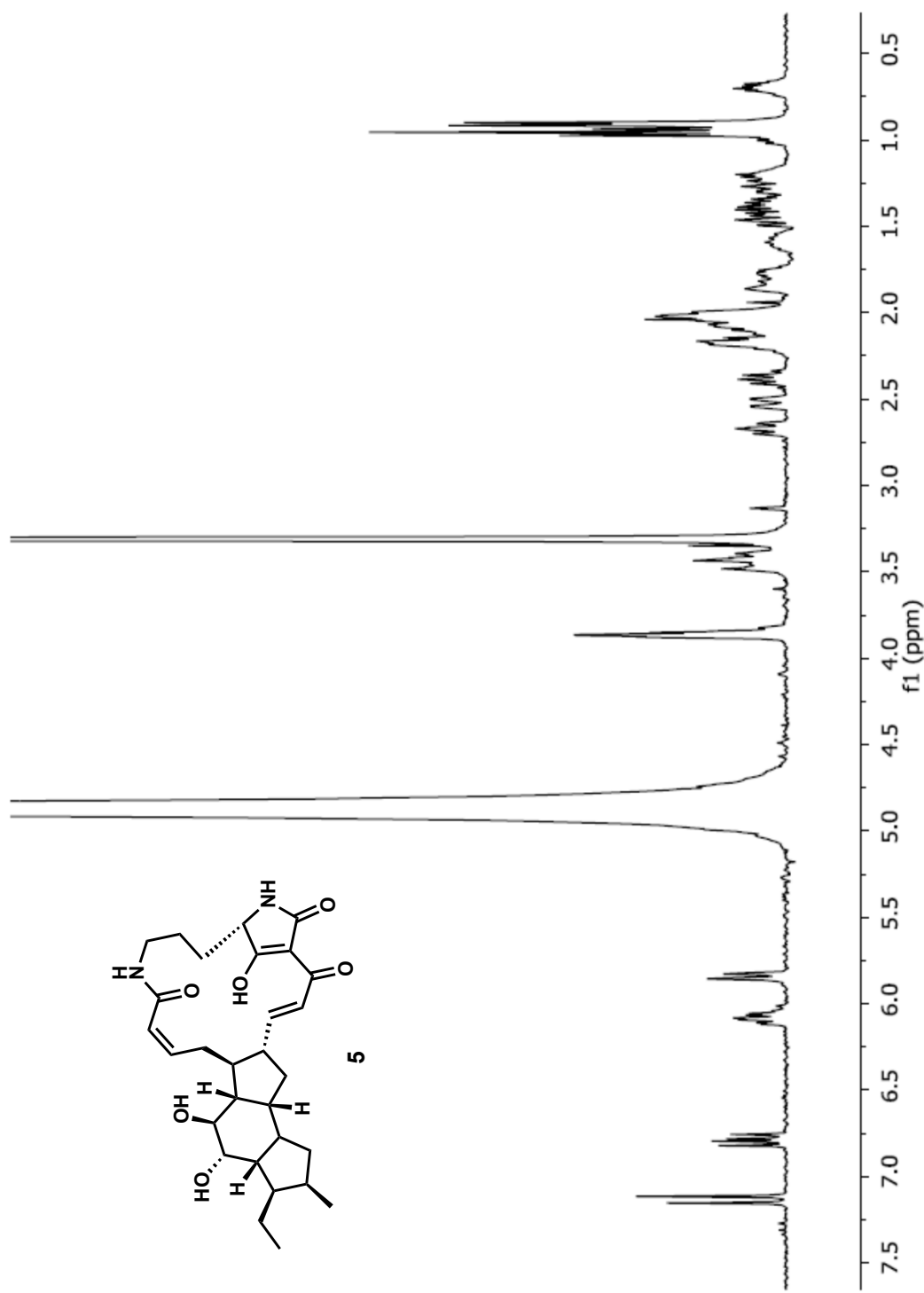


Figure 2.41 Capsimycin C (5).

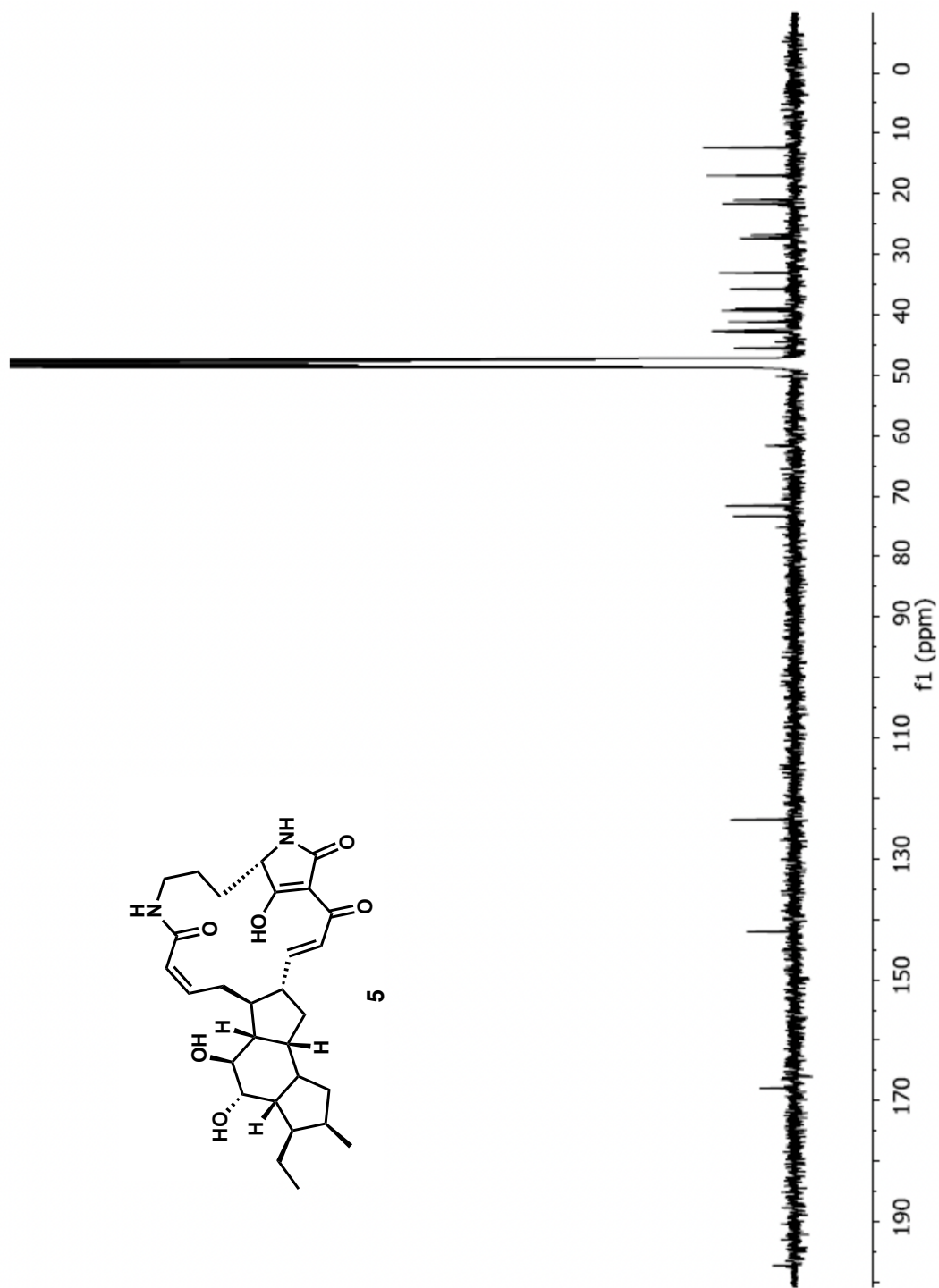
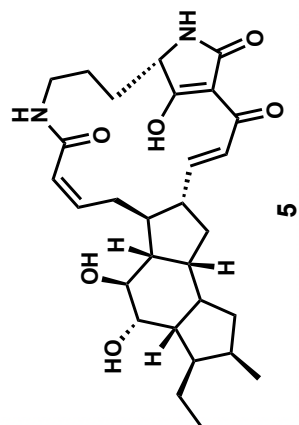


Figure 2.42 Capsimycin C (5).

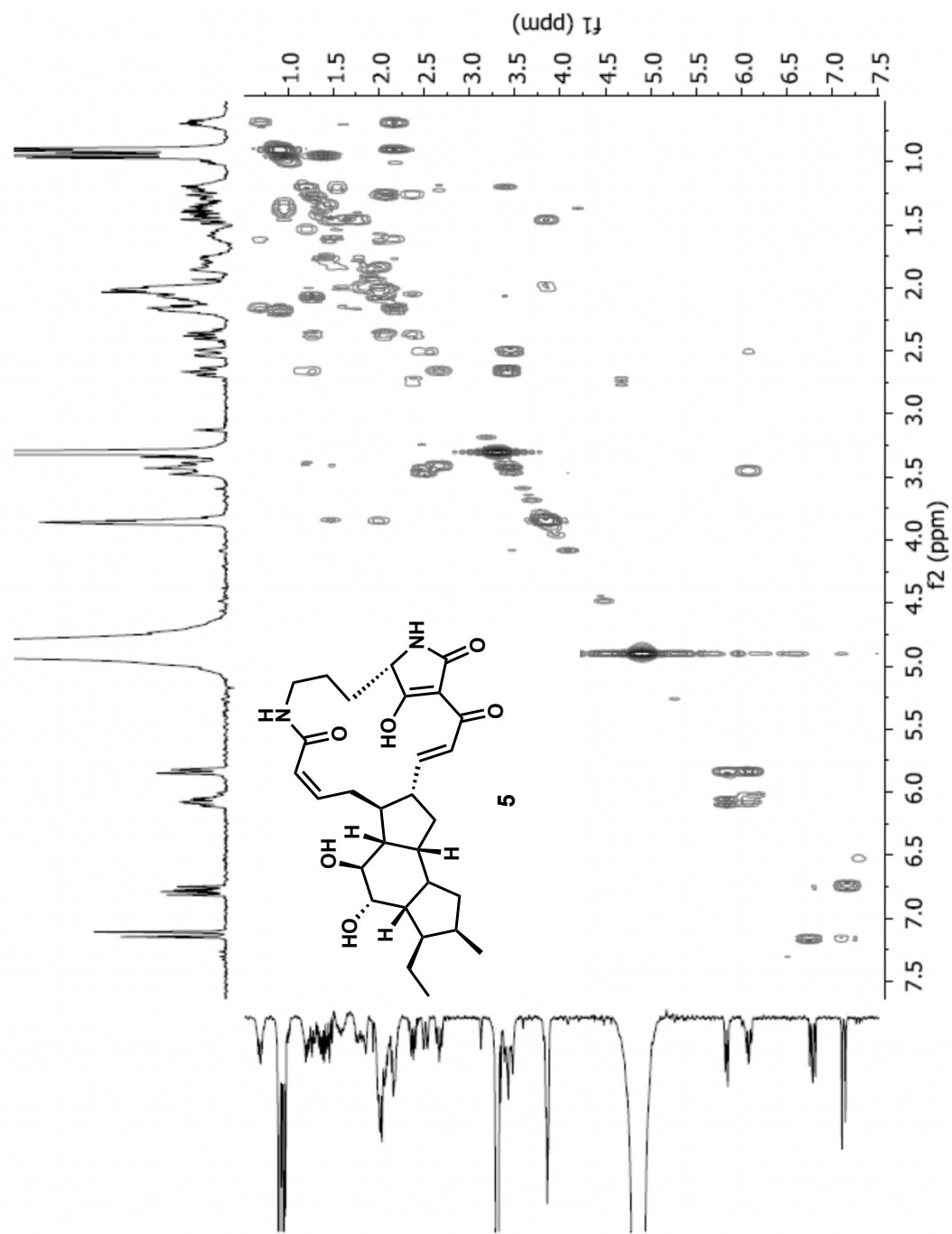


Figure 2.43 Capsimycin C (5).

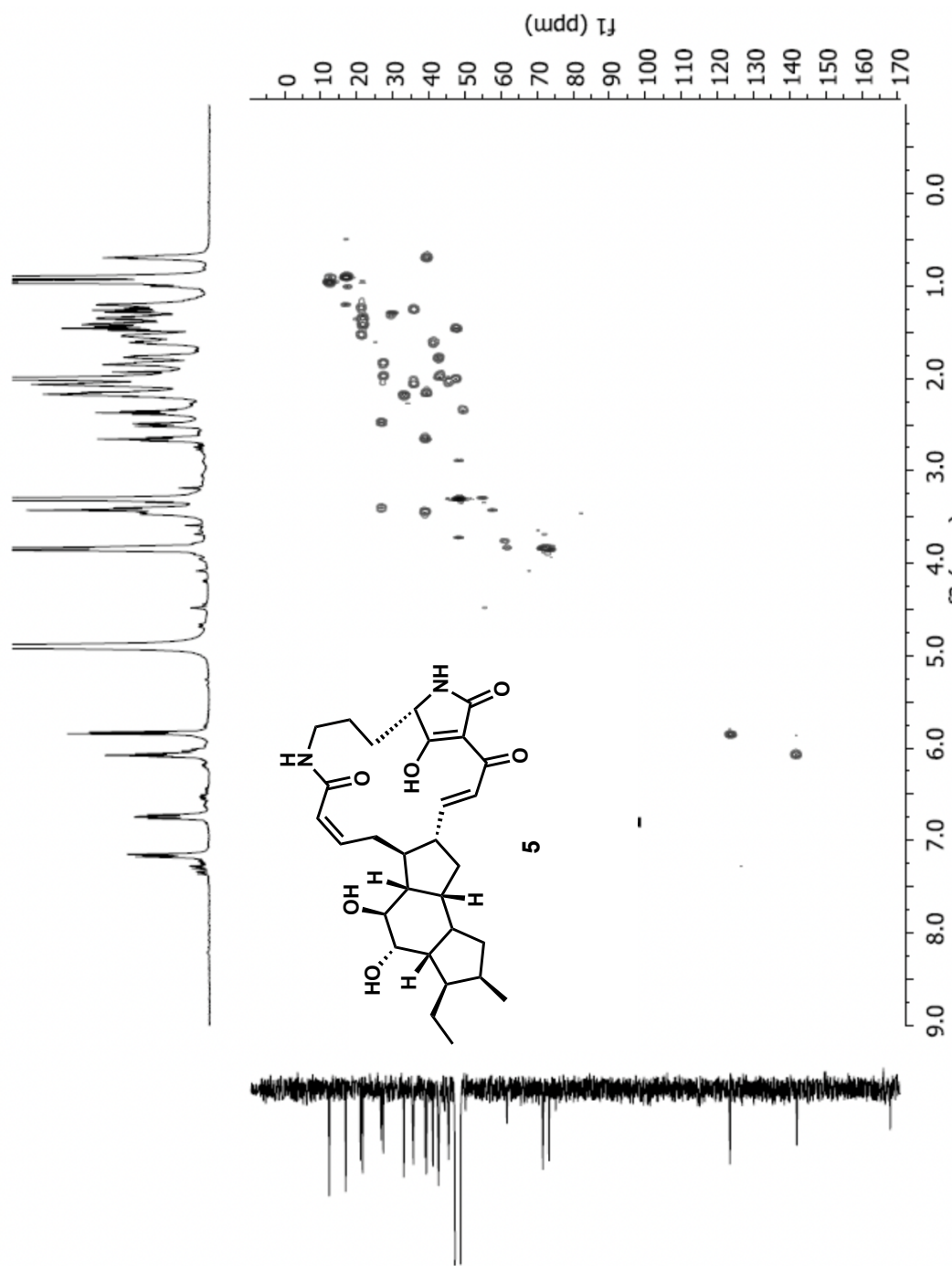


Figure 2.44 Capsimycin C (5).

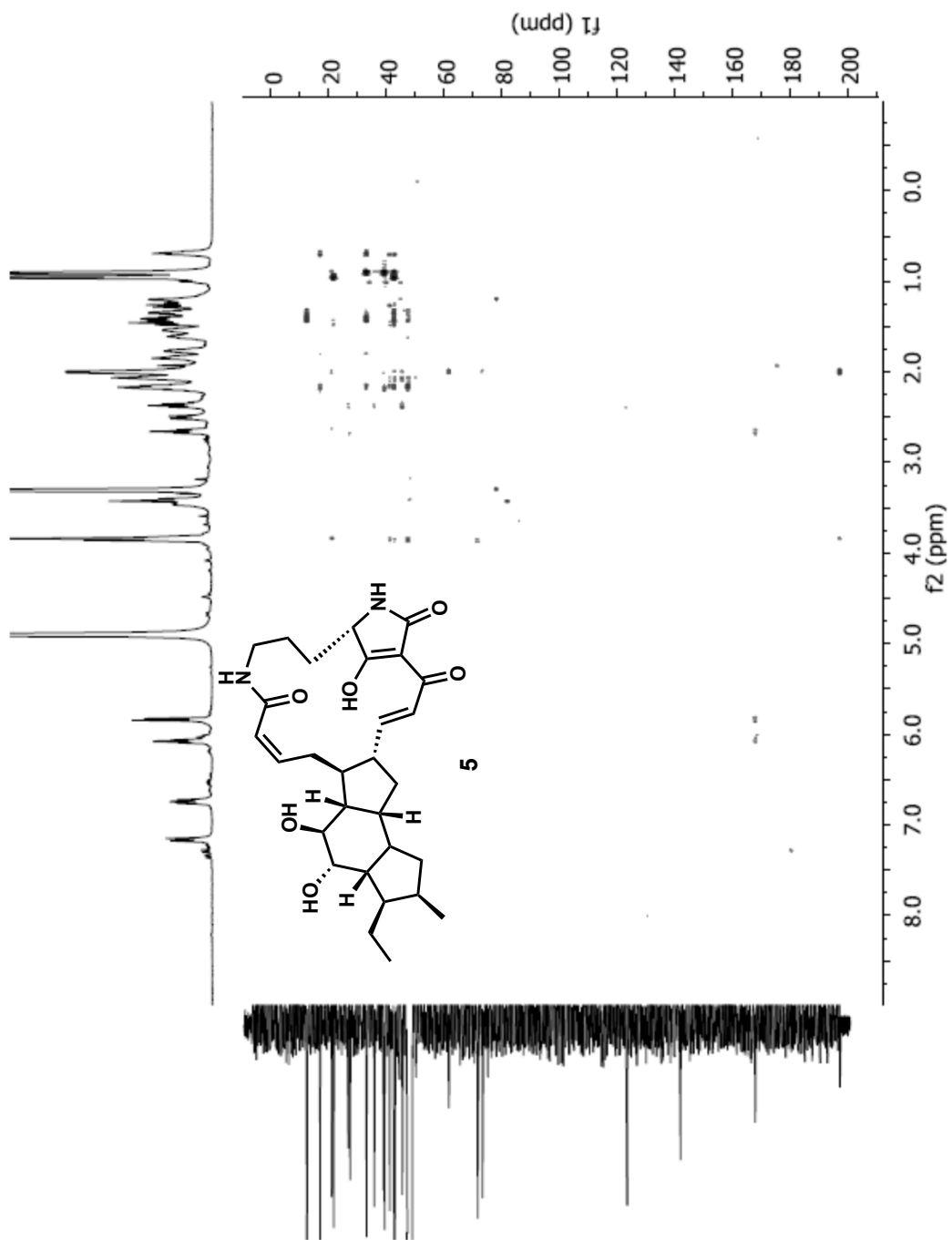


Figure 2.45 Capsimycin C (5).

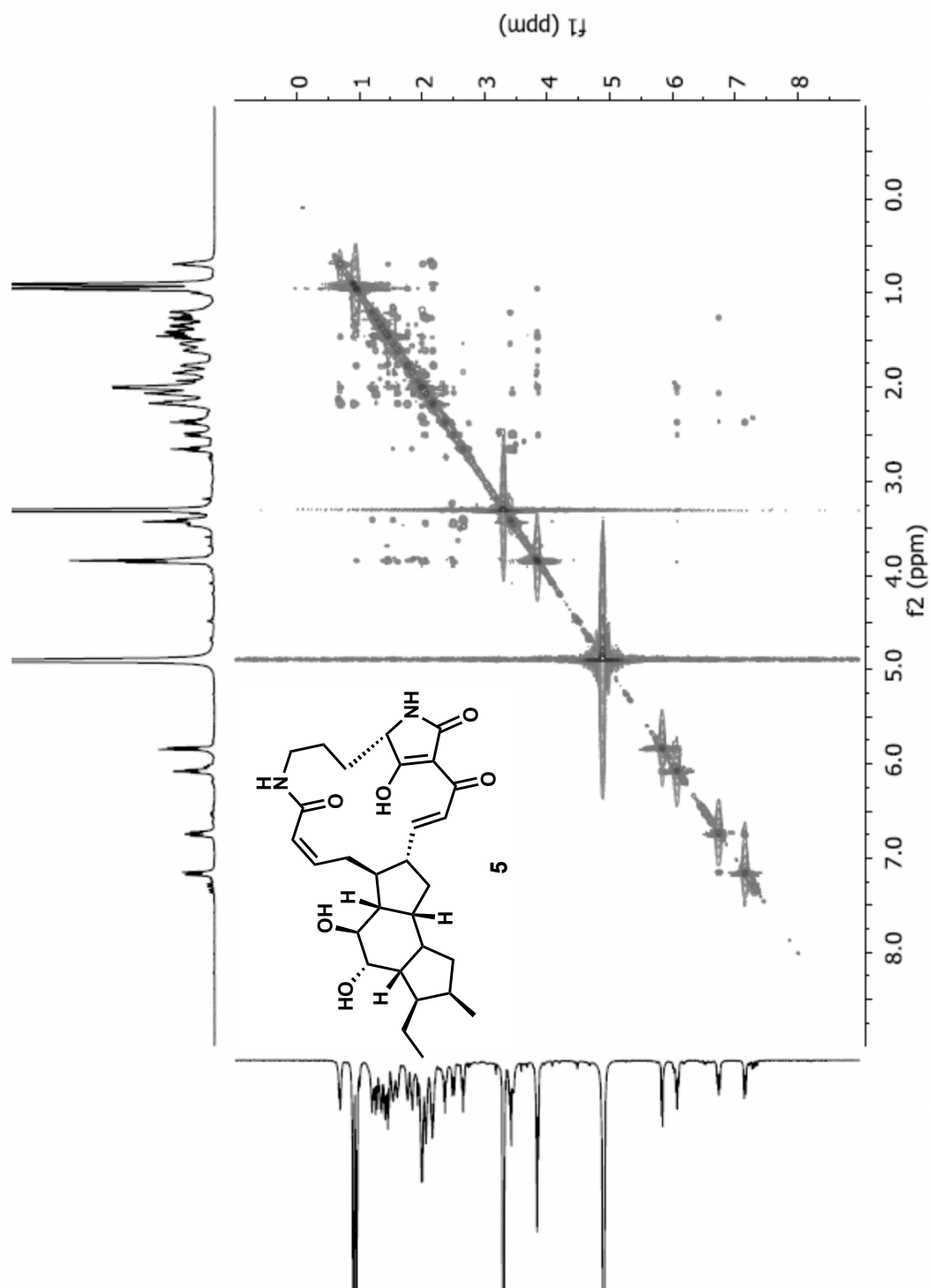


Figure 2.46 Capsimycin C (5).

Xlamenemycin C (6)

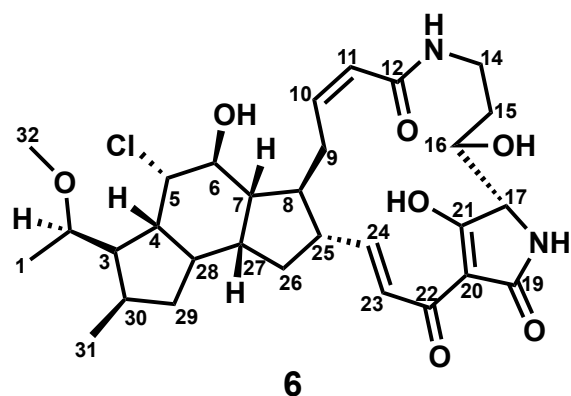
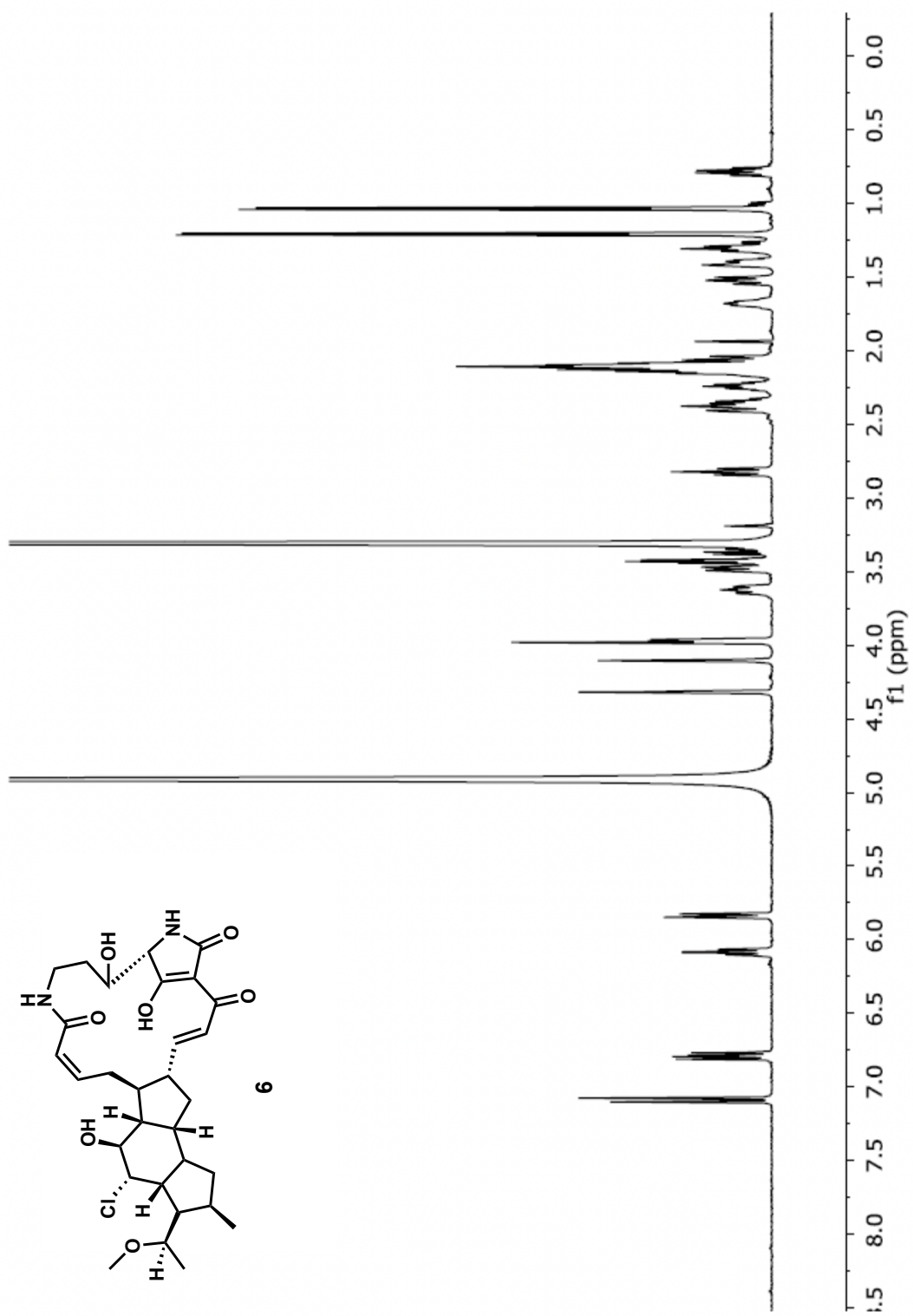


Table 2.6 ^1H (600 MHz) and ^{13}C (100 MHz) spectroscopic data of **6**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	1.21 (d, $J=6.2$)	17.4	17	3.99, m	68.6
2	3.49 m	77.5	NH-18	-	-
3	2.14 m	47.0	19	-	176.4
4	2.05 m	44.2	20	-	100.4
5	4.32 (t, $J=2.9$)	65.0	21	-	194.0
6	4.11 s	73.4	22	-	174.4
7	2.10 m	47.3	23	7.09 (d, $J=15.4$)	122.4
8	2.15 m	45.9	24	6.79 (dd, $J=15.4, 10.3$)	151.9
9	3.63 m, 2.41 m	26.0	25	2.36 m	49.9
10	6.09 (td, $J = 11.2, 2.9$)	140.8	26	2.10 m, 1.31 m	35.4
11	5.84 (d, $J = 11.7$)	123.2	27	2.05 m	41.3
12	-	167.8	28	2.09 m	42.2
NH-13	-	-	29	2.10 m; 0.79 (dd, $J=20.6, 11.3$)	39.5
14	3.44 m, 2.83 (t, $J = 11.8$)	37.2	30	2.24 m	34.1
15	1.53 m, 1.40 m	31.4	31	1.04 (d, $J=7.1$)	16.5
16	3.97 m	71.5	32	3.31 s (overlap)	54.5



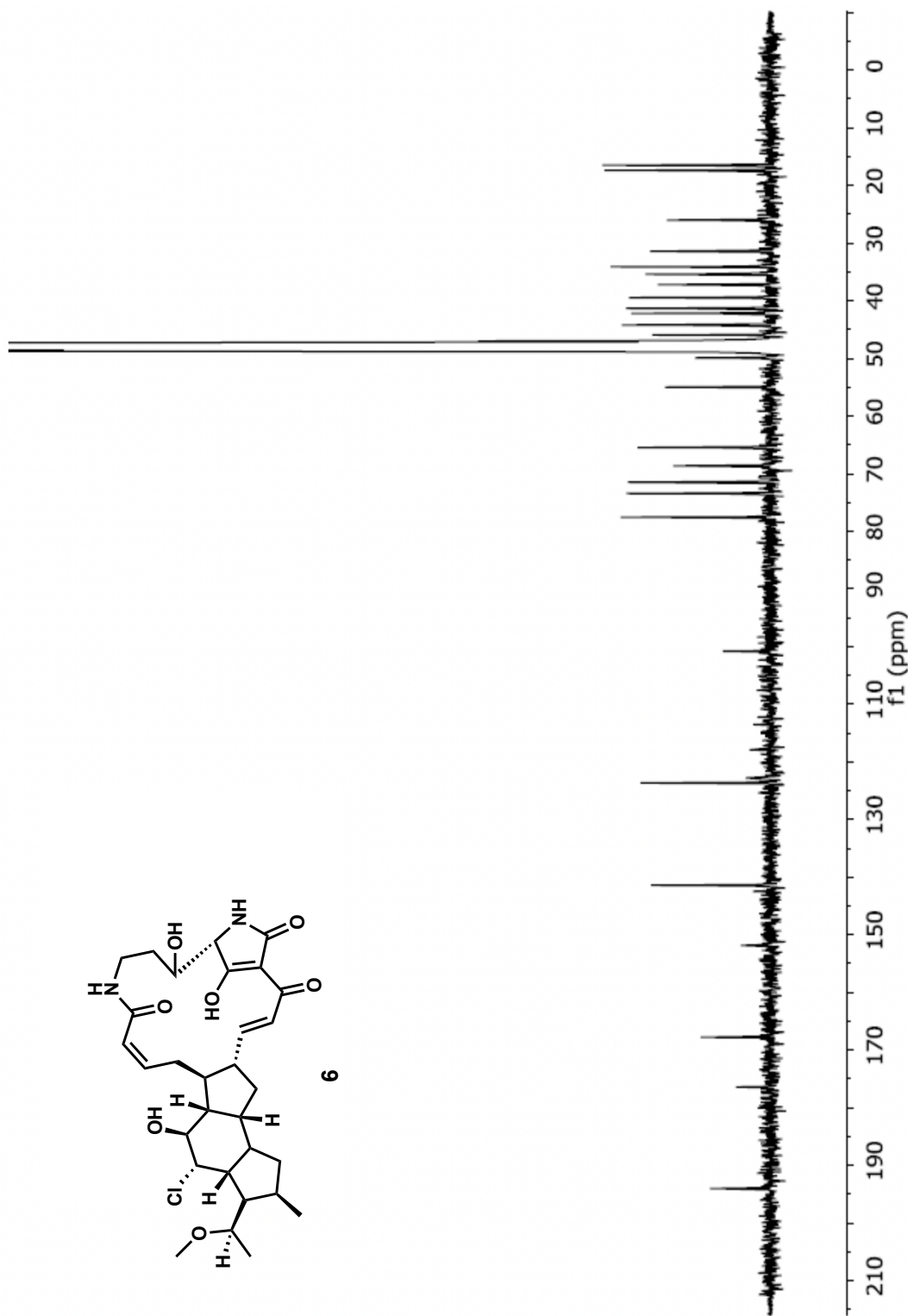
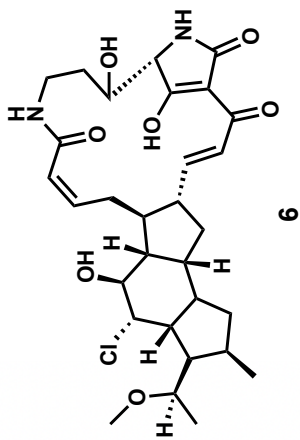


Figure 2.48 Xlmaenemycin C (6).

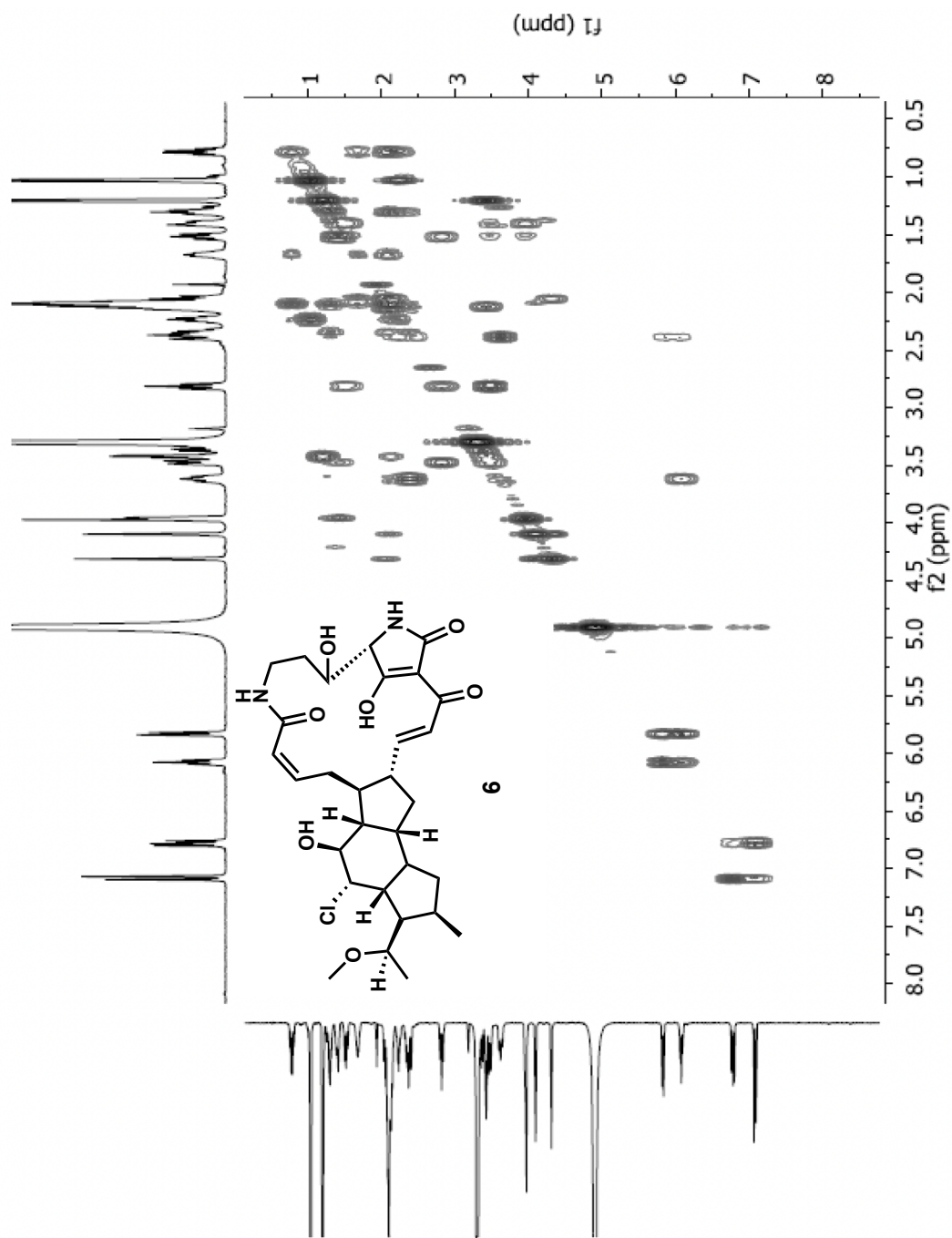


Figure 2.49 Xlmaenemycin C (6).

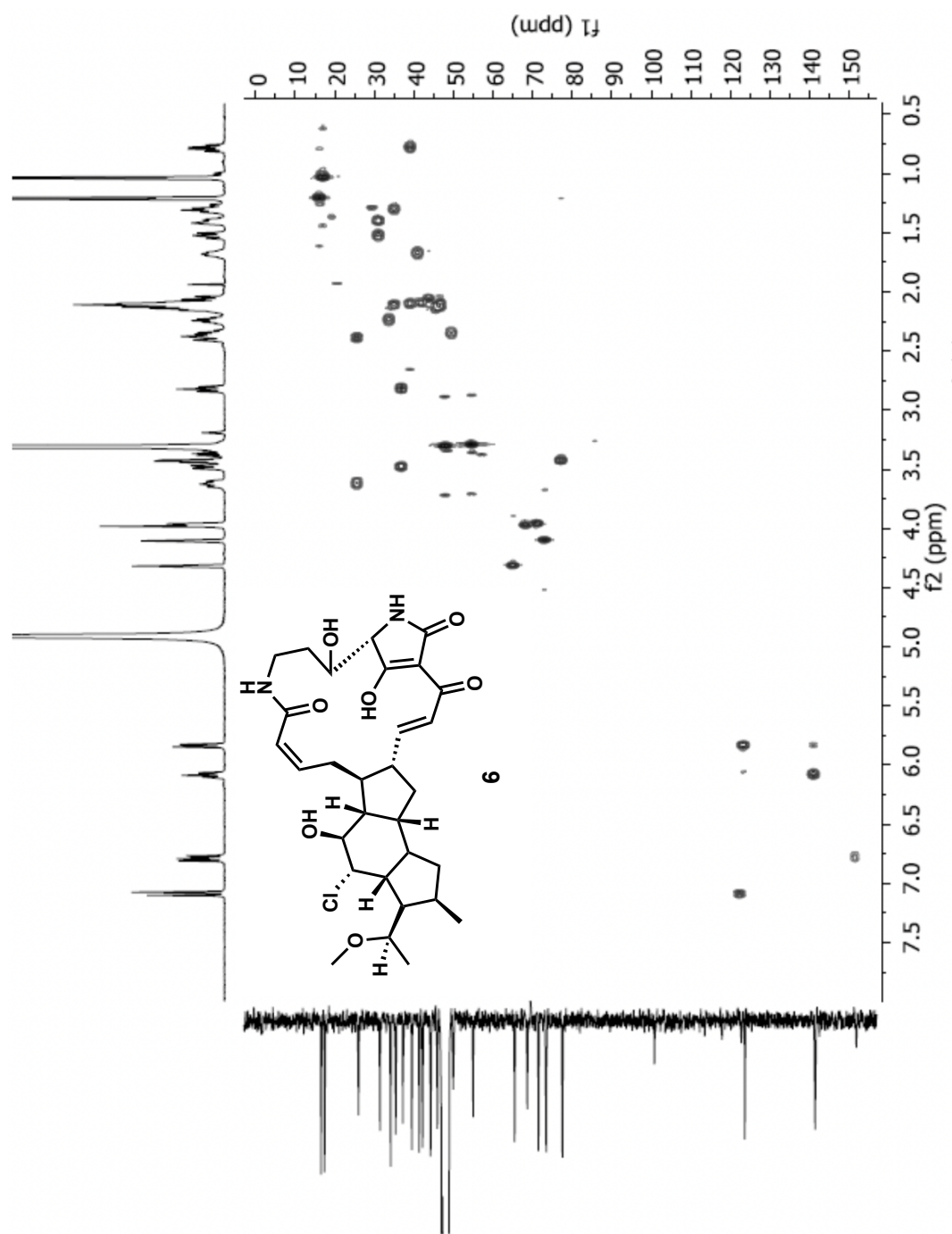


Figure 2.50 Xlmaenemycin C (6).

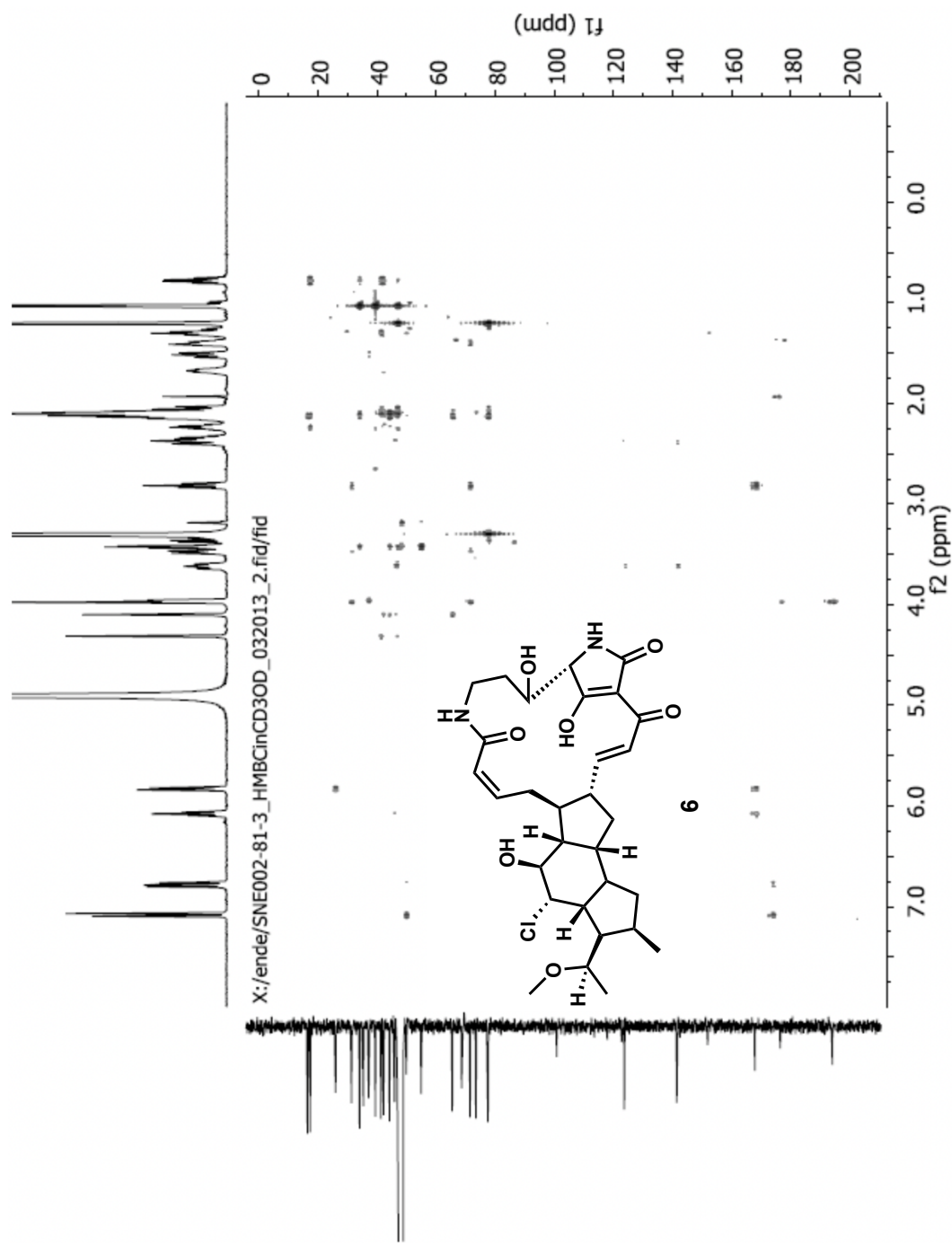


Figure 2.51 Xlmaenenmycin C (6).

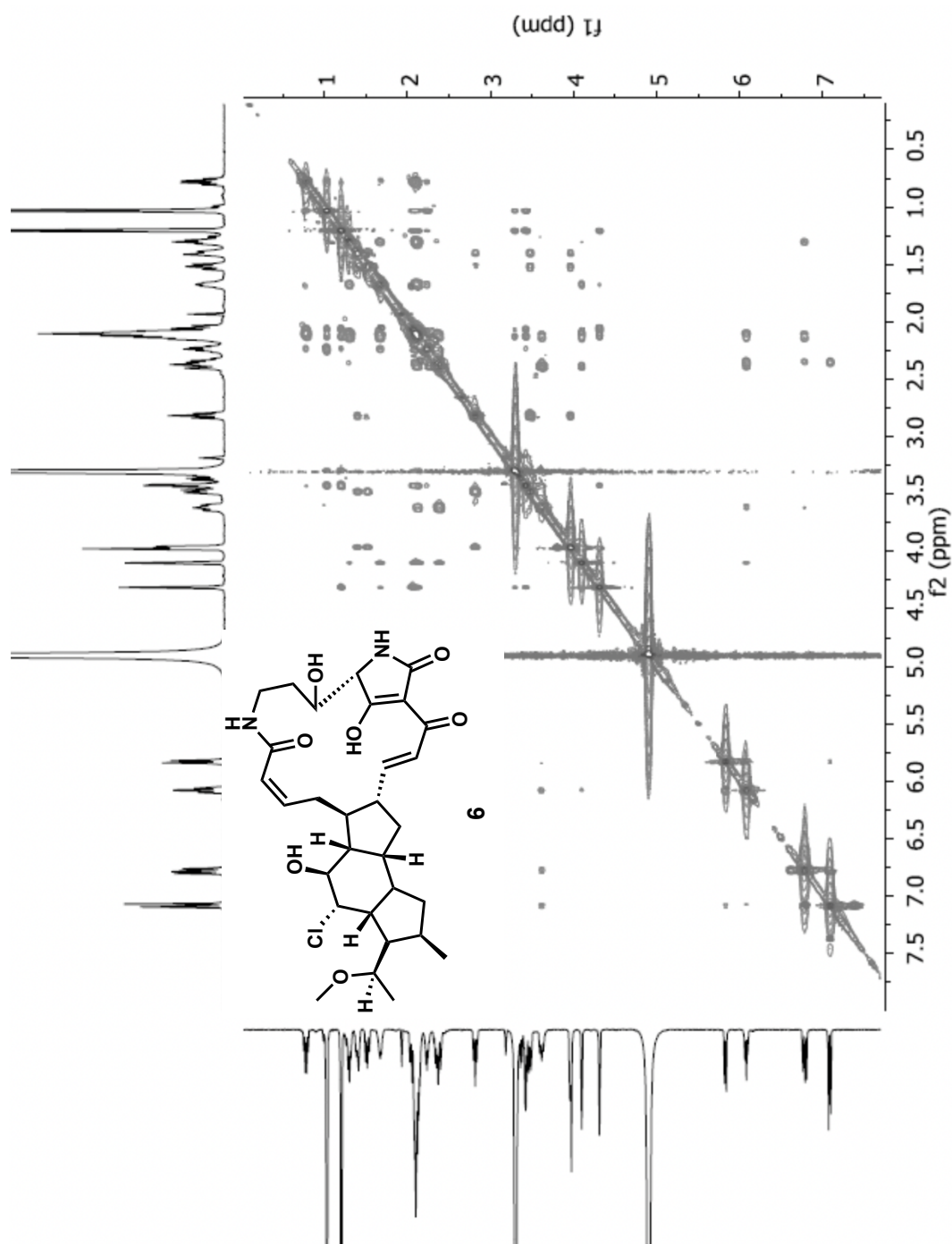


Figure 2.52 Xlmaenemycin C (6).

Capsimycin E (8)

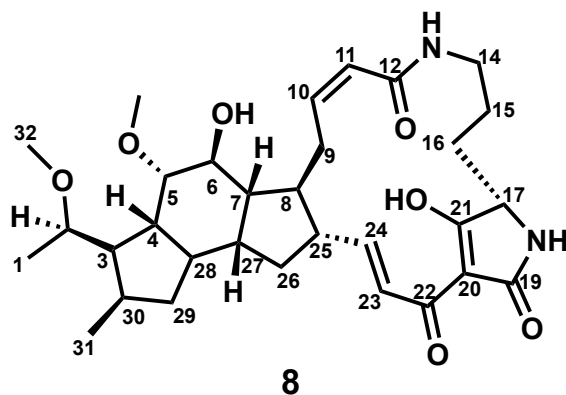


Table 2.8 ^1H (800 MHz) and ^{13}C (800 MHz) spectroscopic data of **8**.

Position	δH , Mult (J in Hz)	δC , Mult	Position	δH , Mult (J in Hz)	δC , Mult
1	1.22 (d, $J = 6.2$)	17.8	17	3.82, br d (4.0)	62.8
2	3.40 m	79.1	NH-18	-	-
3	2.05 m	46.1	19	-	176.3
4	1.68 dd (11.2, 2.5)	44.5	20	-	102.3
5	3.44 m (overlap)	83.1	21	-	197.8
6	4.1 br	68.7	22	-	174.5
7	2.02 m	49.3	23	7.23 (d, $J = 15.4$)	124.6
8	2.01 m	45.9	24	6.71 (dd, $J = 15.3, 9.6$)	152.9
9	2.46d (15.4), 3.48 m	27.4	25	2.34 m	49.4
10	6.1 (td, $J = 11.3, 3.5$)	142.6	26	2.08 m, 1.25 m	36.6
11	5.85 (d, $J = 11.4$)	124.5	27	1.92 m	42.6
12	-	168.9	28	1.54 m	43.1
NH-13	-	-	29	2.04 m, 0.68 m	40.5
14	2.64 br t (11.1), 3.40 m	38.9	30	2.16 m	34.8
15	1.16 m, 1.52 m	22.3	31	1.01 (d, $J = 7.1$)	17.8
16	1.82 m, 1.98 m	28.3	32/33	3.29 s, 3.41 s	55.6, 58.4

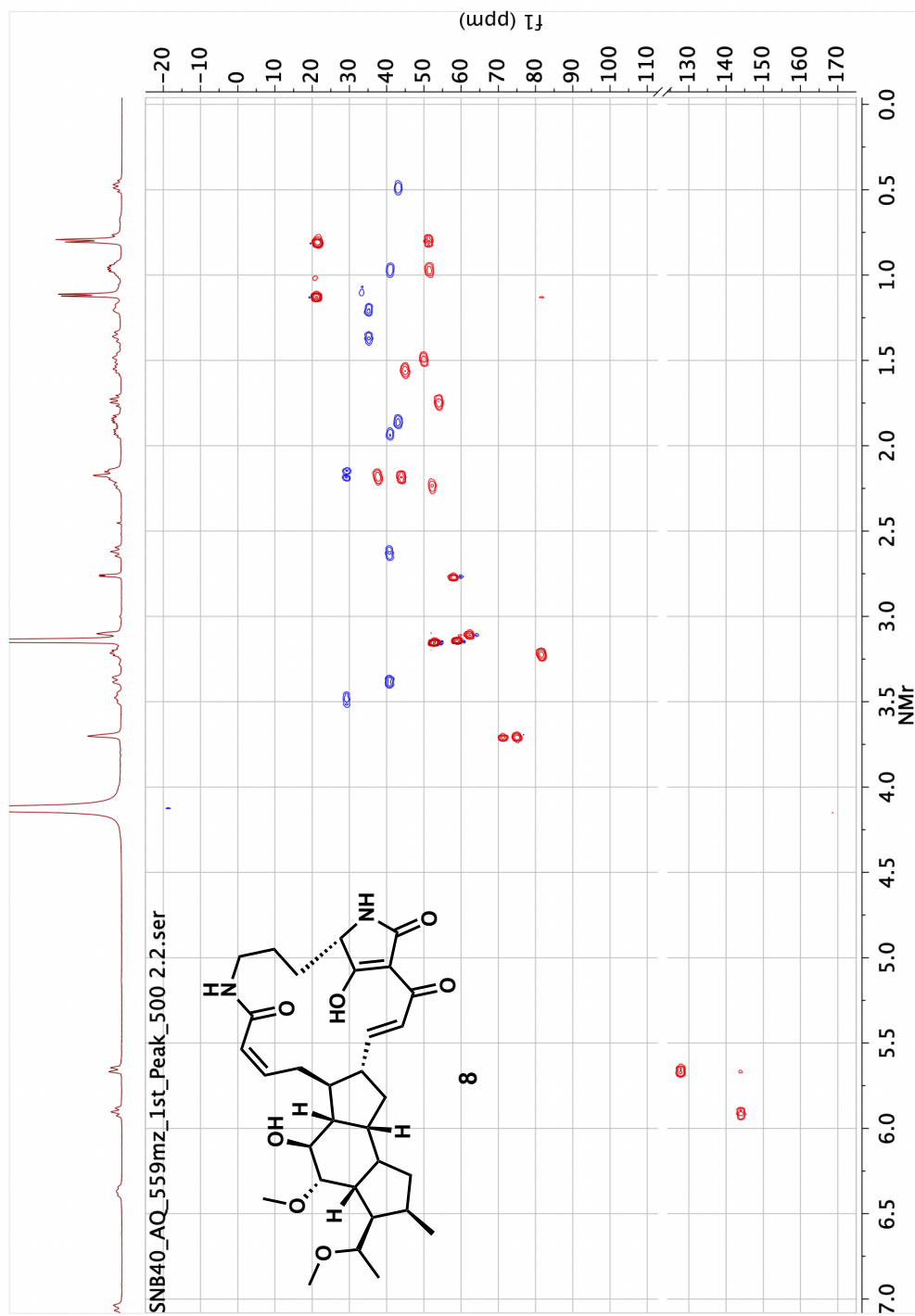


Figure 2.53 Capsimycin E (8).

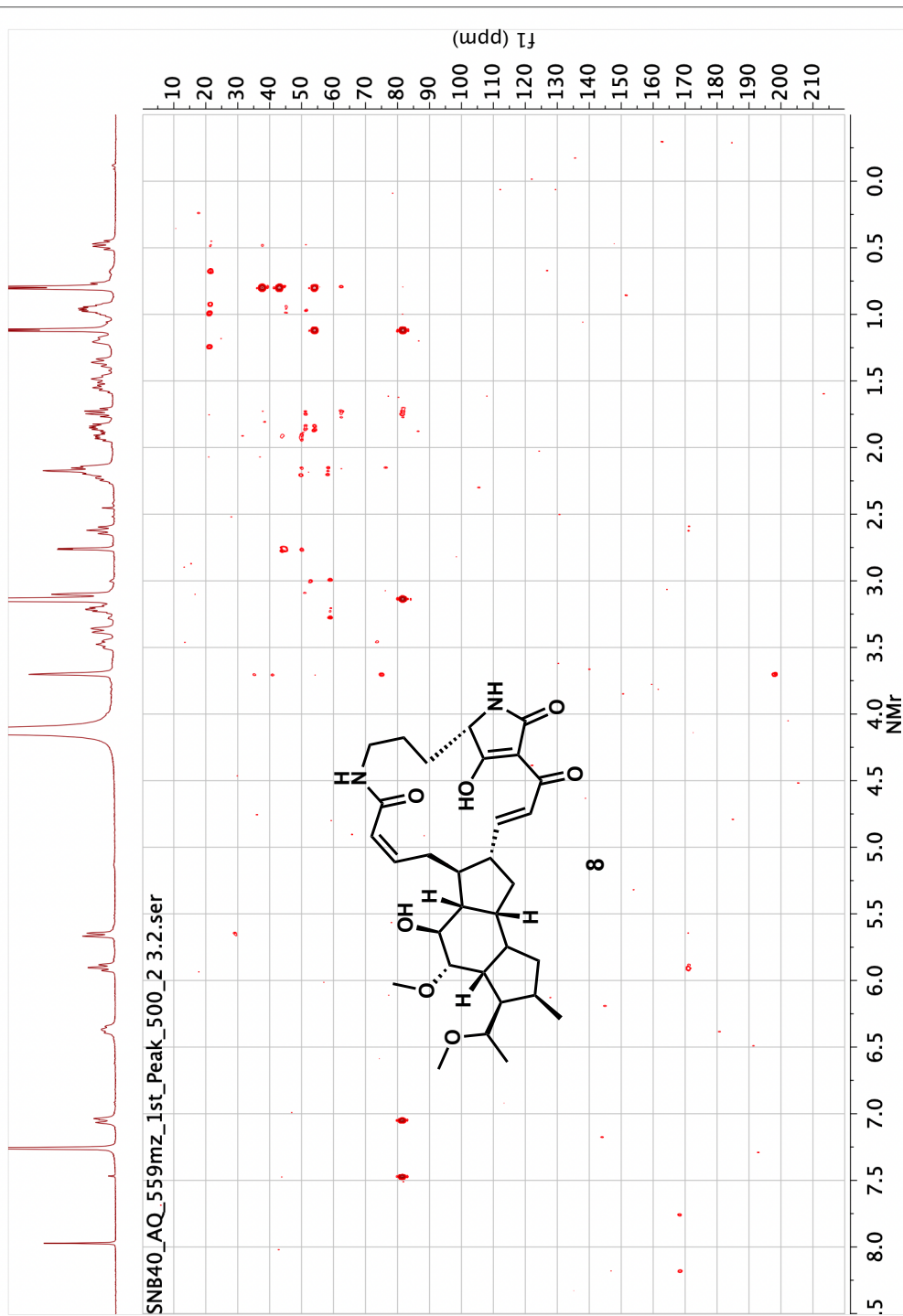


Figure 2.54 Capsimycin E (8).

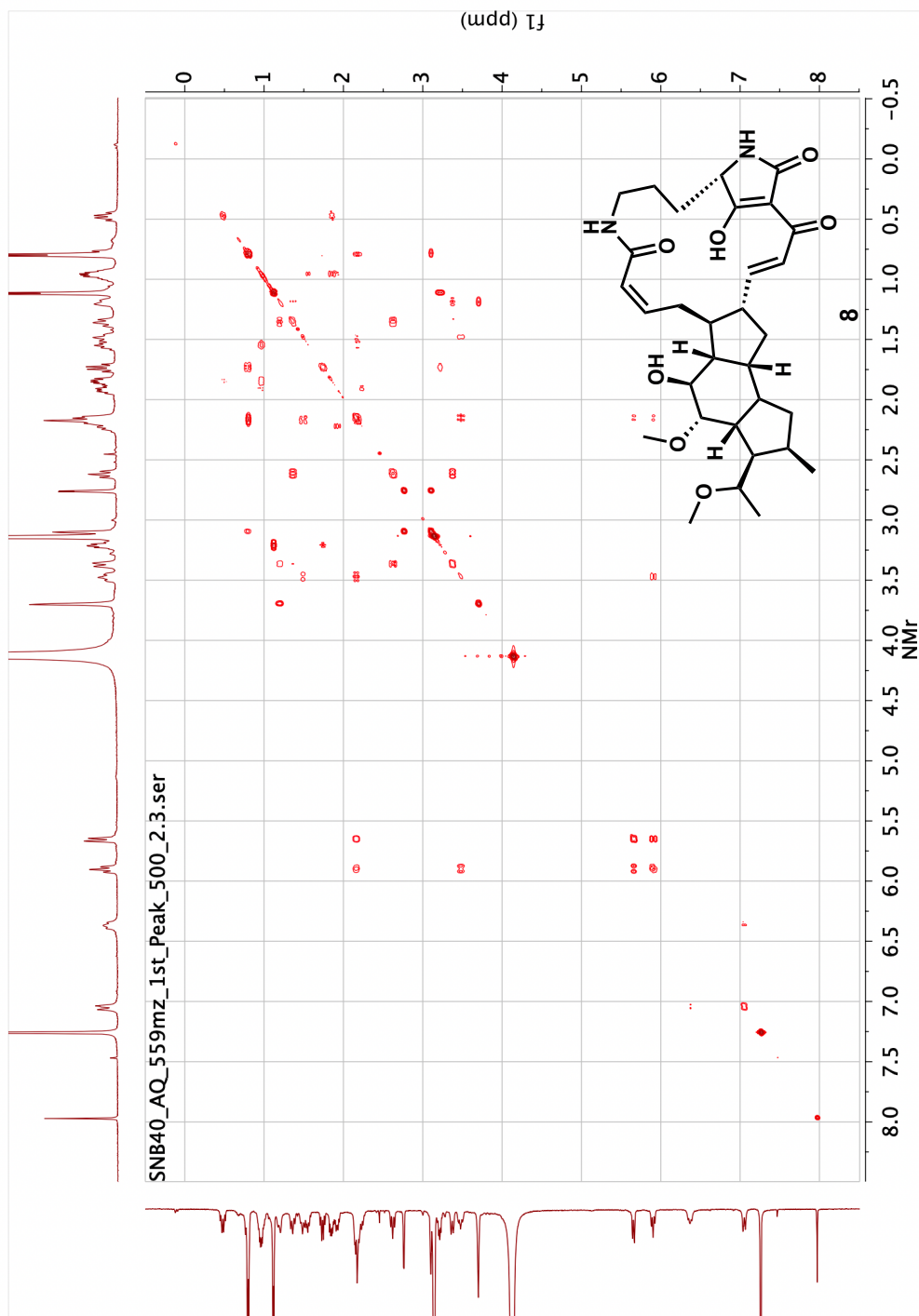


Figure 2.55 Capsimycin E (8).

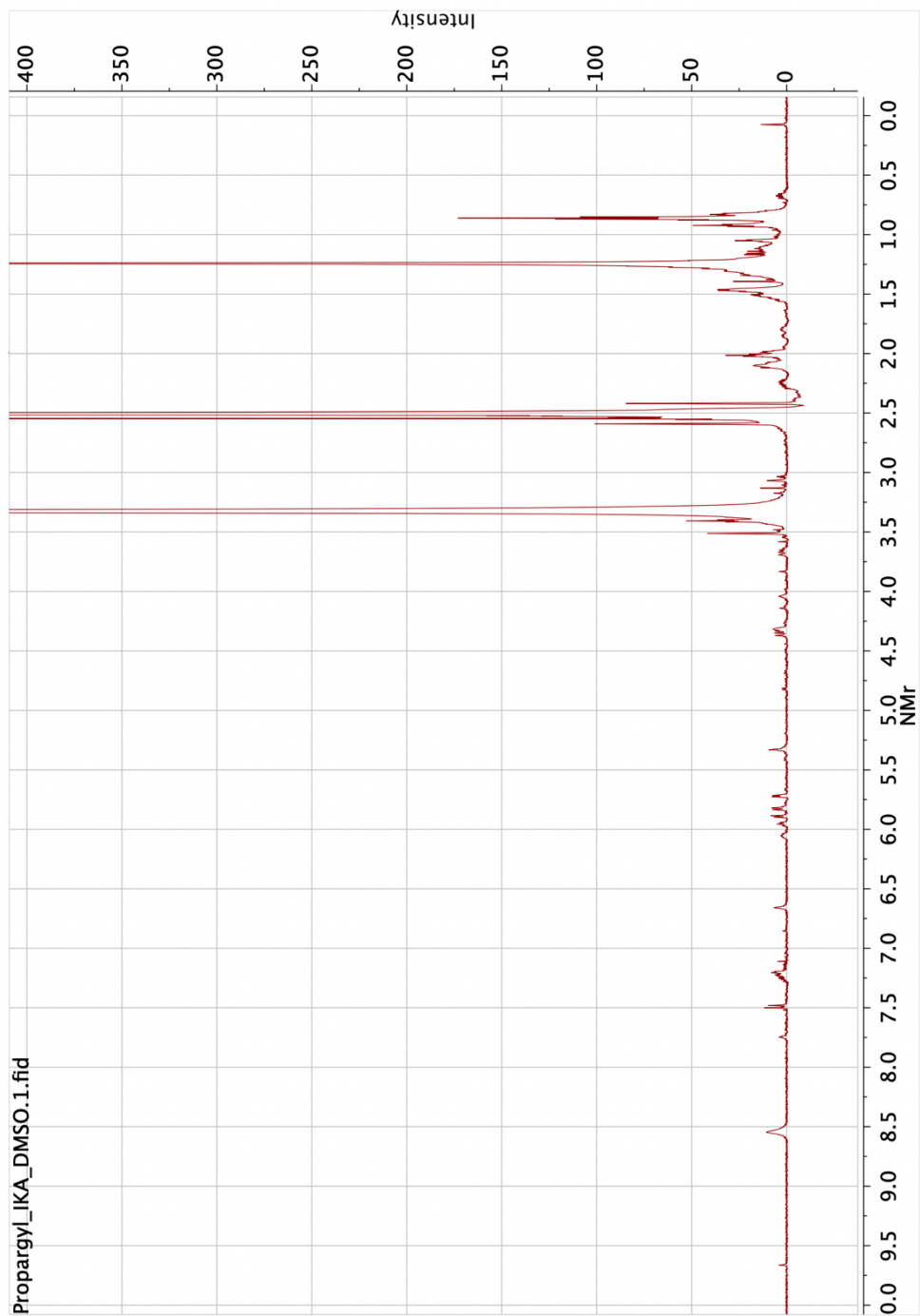


Figure 2.56 Initial Propargyl-IKA.

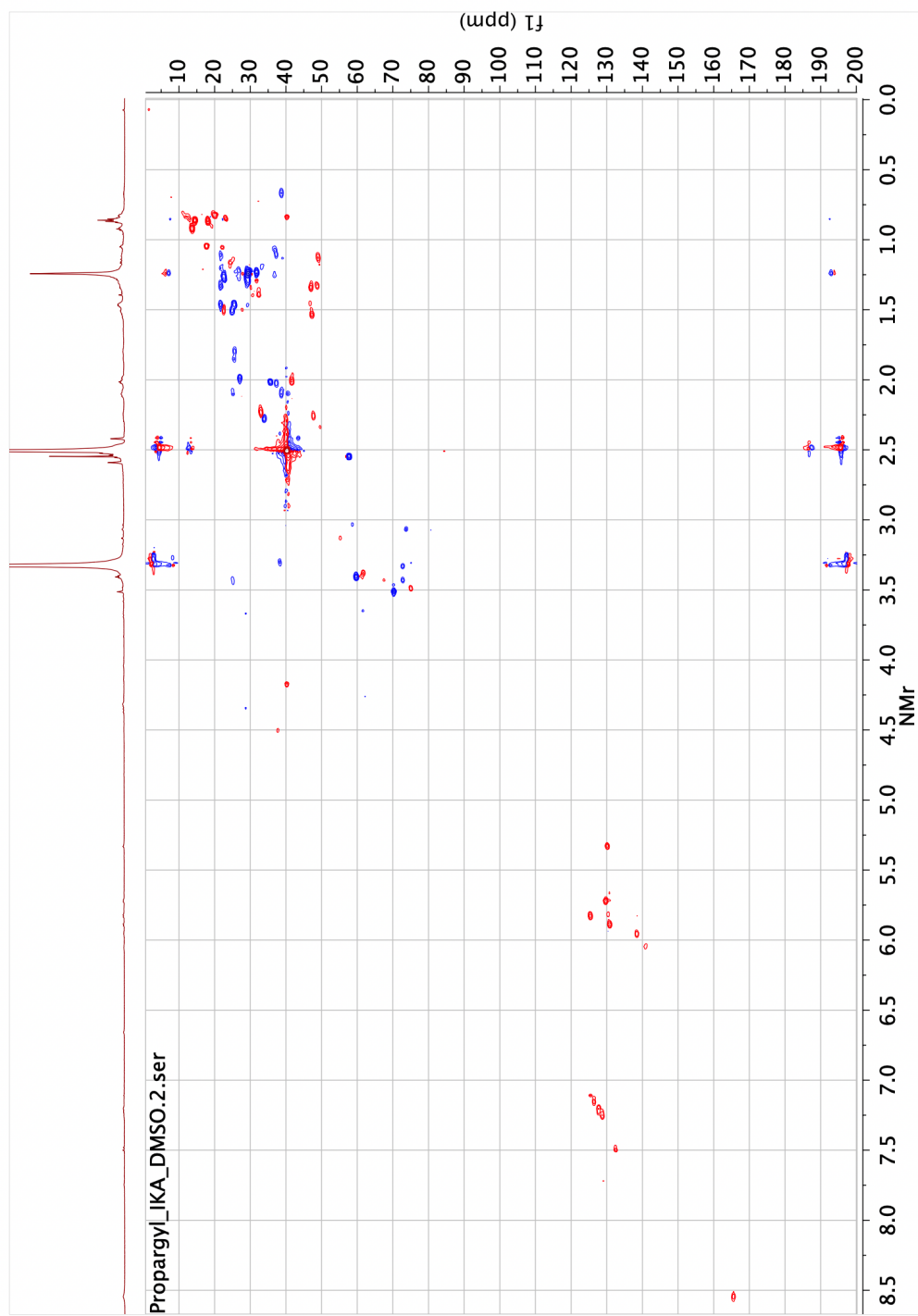


Figure 2.57 Initial Propargyl-IKA.

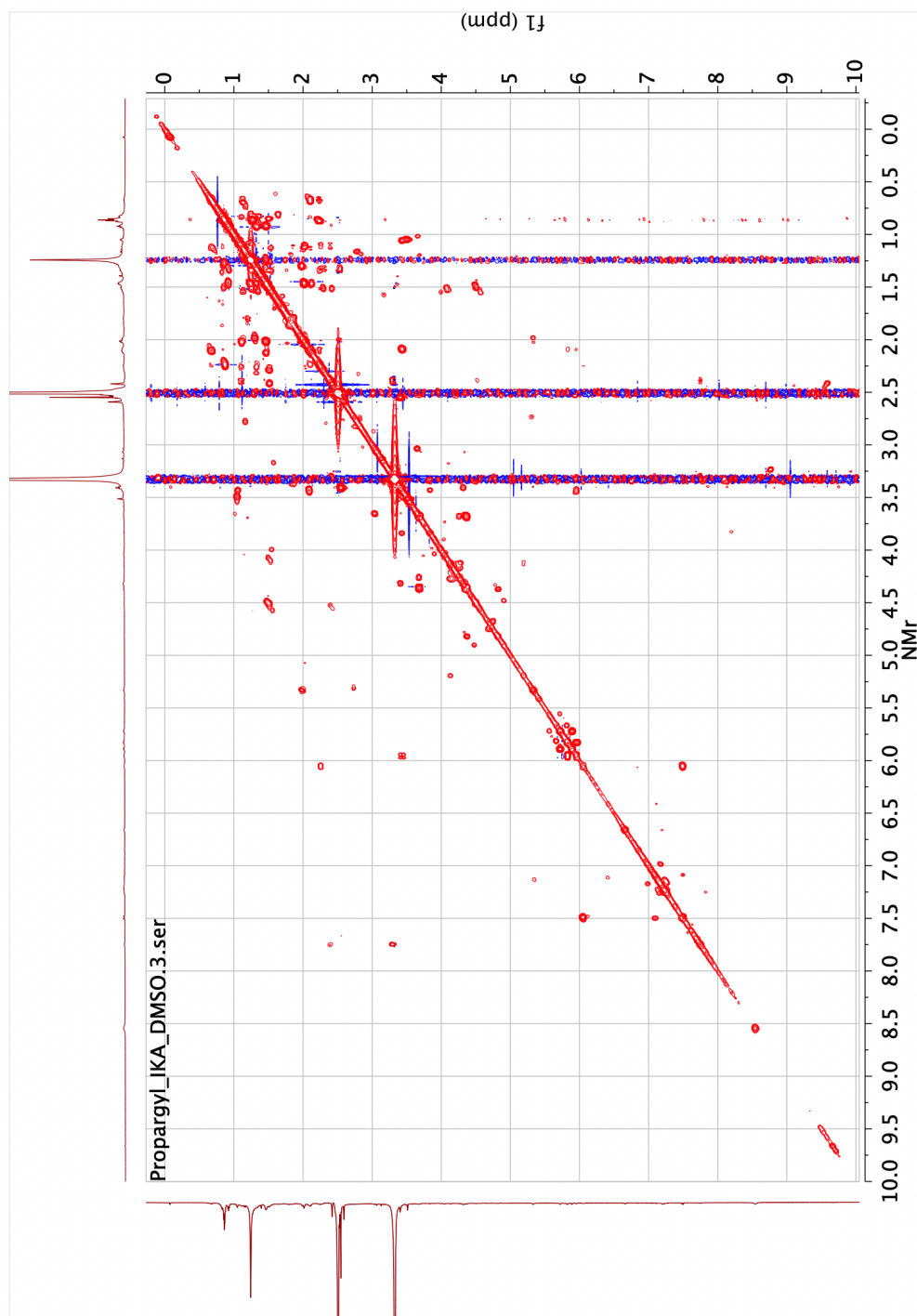


Figure 2.58 Initial Propargyl-IKA.

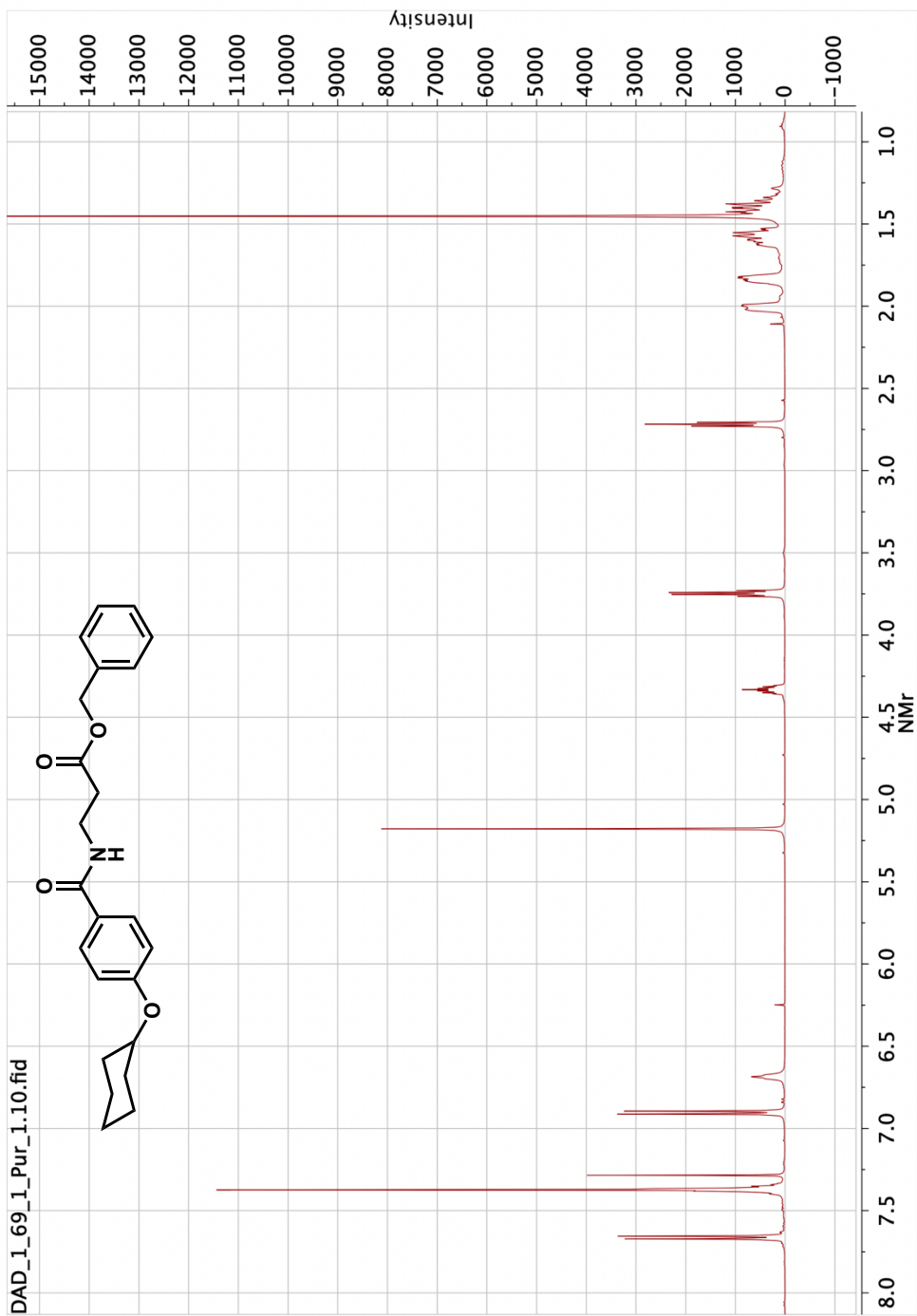


Figure 4.5 Reaction 1-69-1 product.

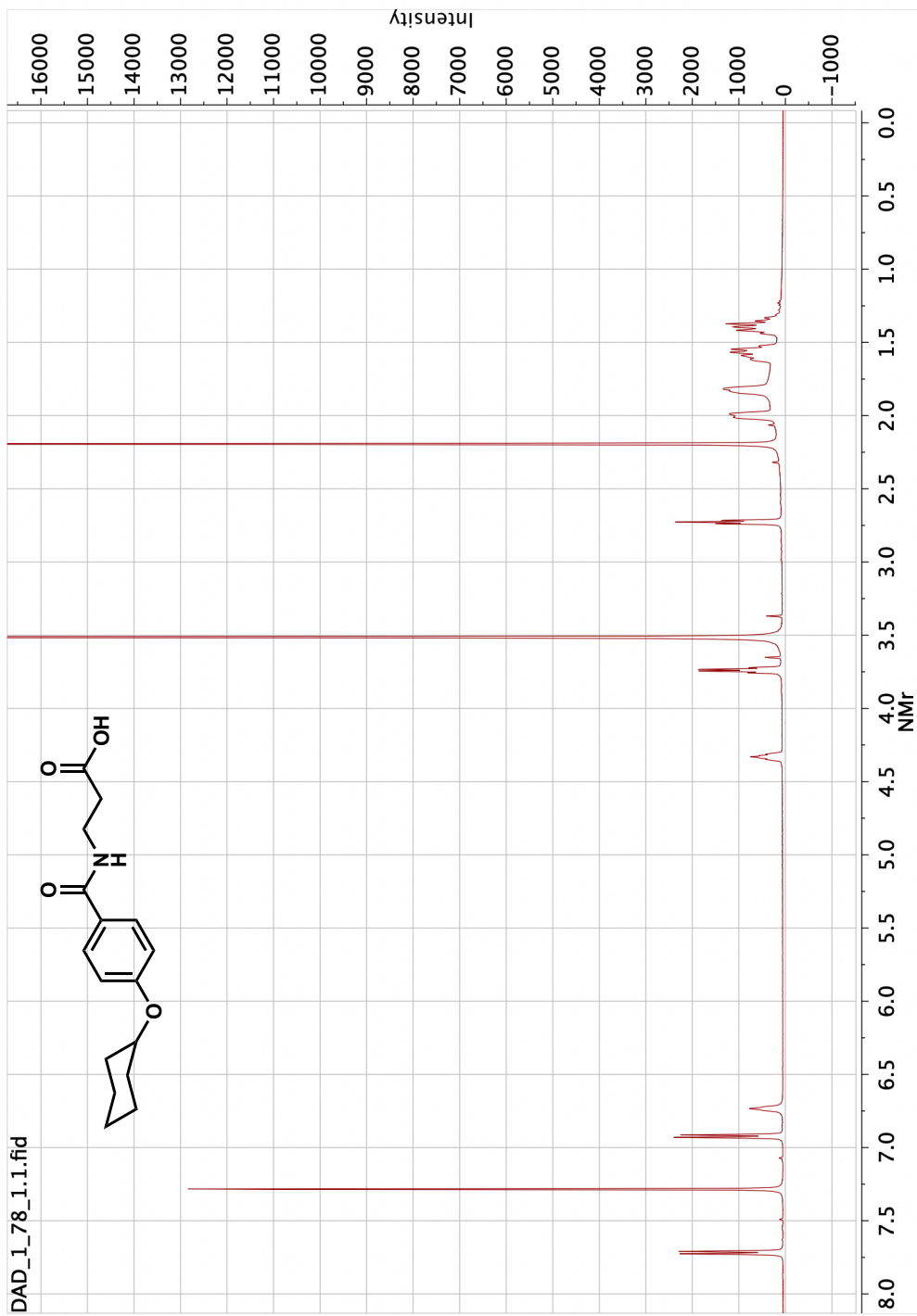


Figure 4.6 Reaction 1-78-1 product.

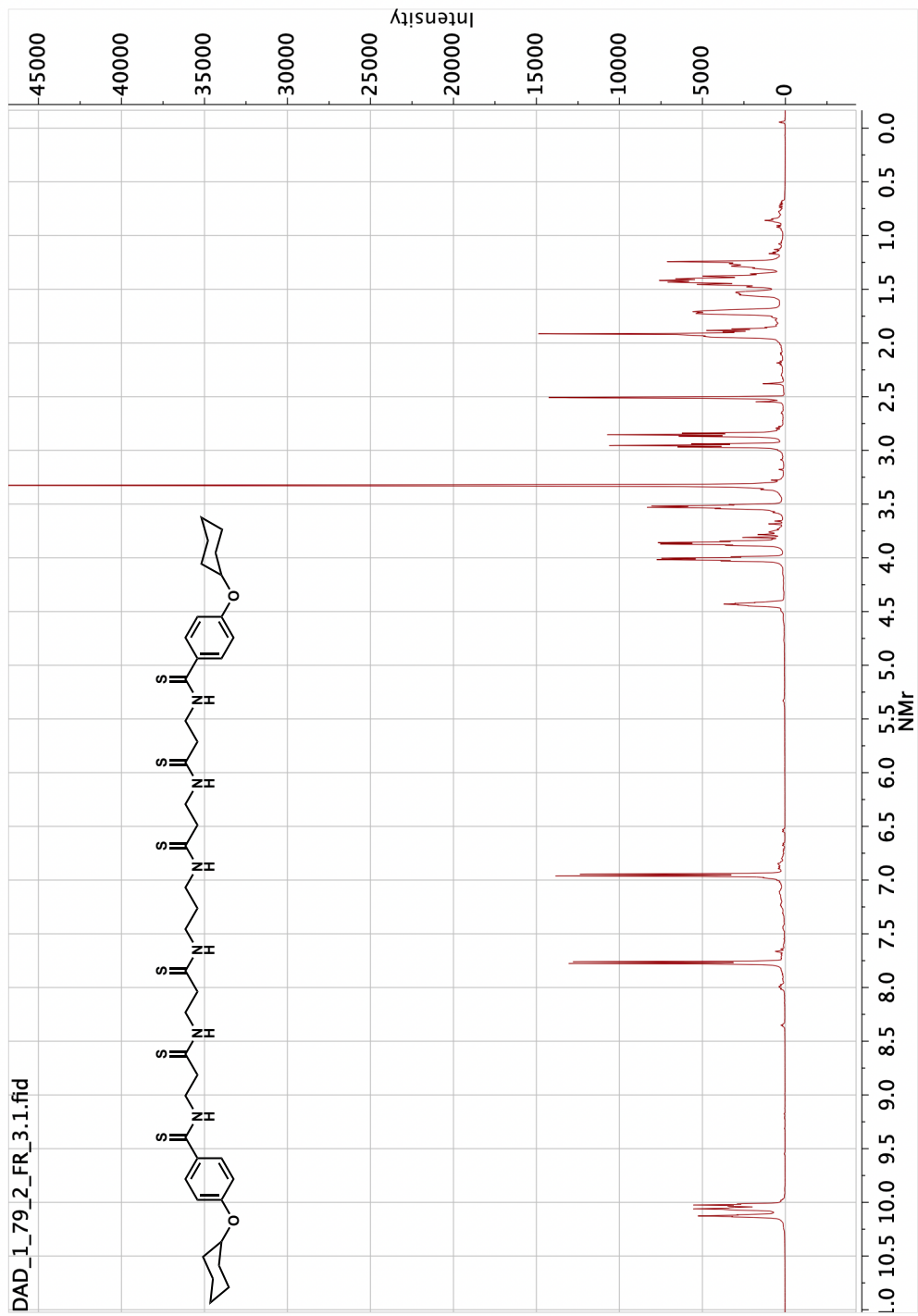


Figure 4.9 Reaction 1-79-2 product.

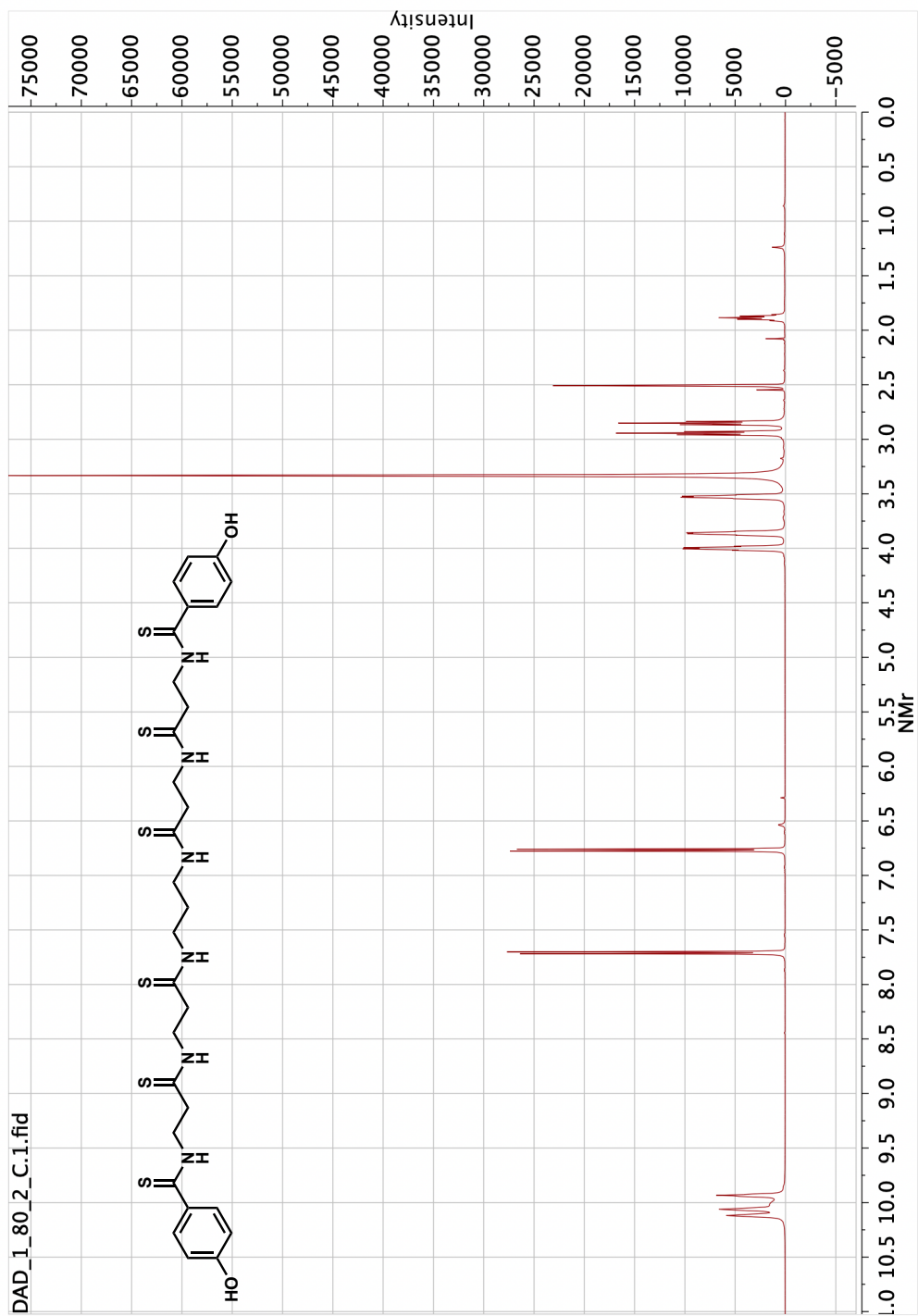


Figure 4.10 Closthioamide.

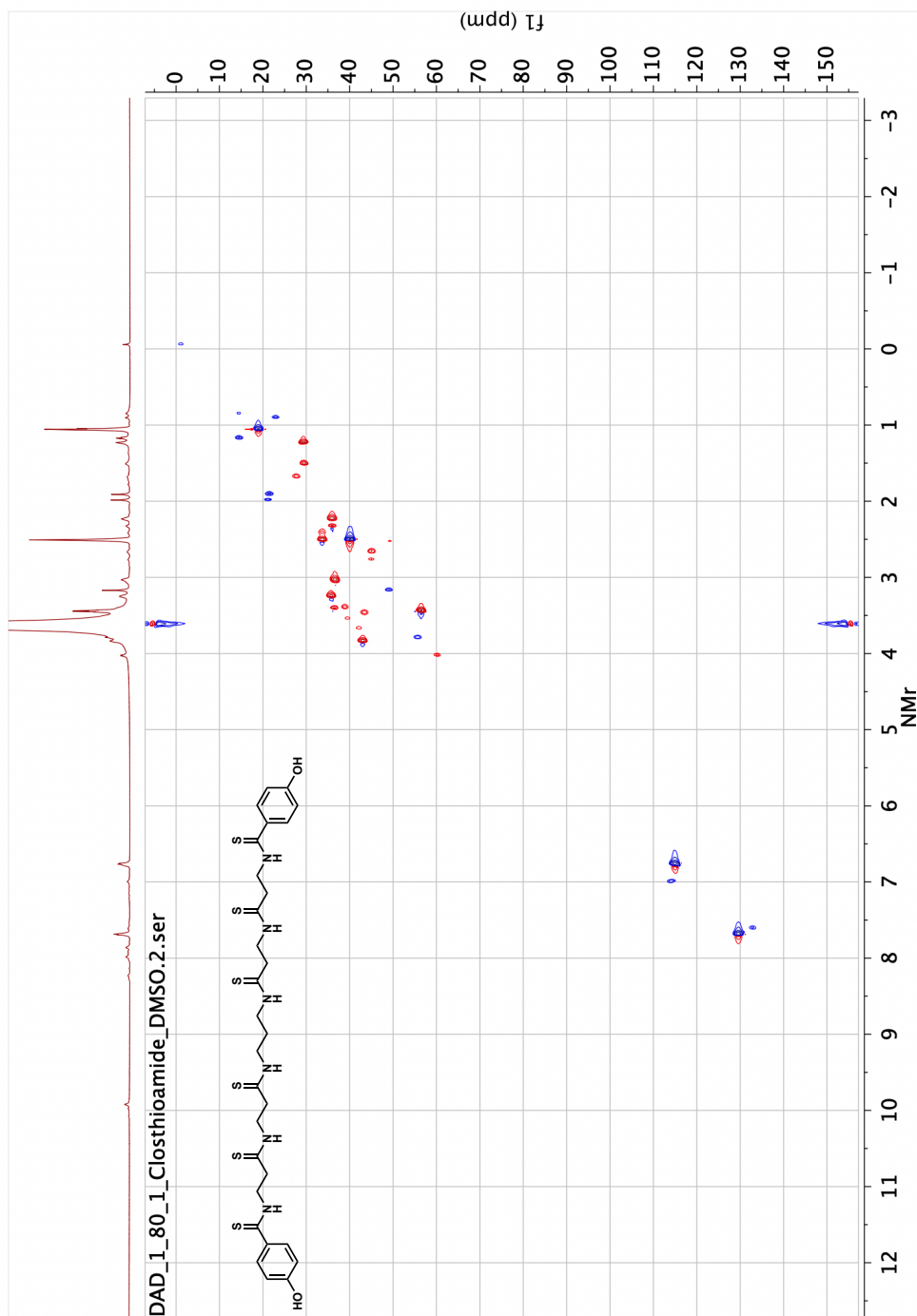


Figure 4.11 Closthioamide.

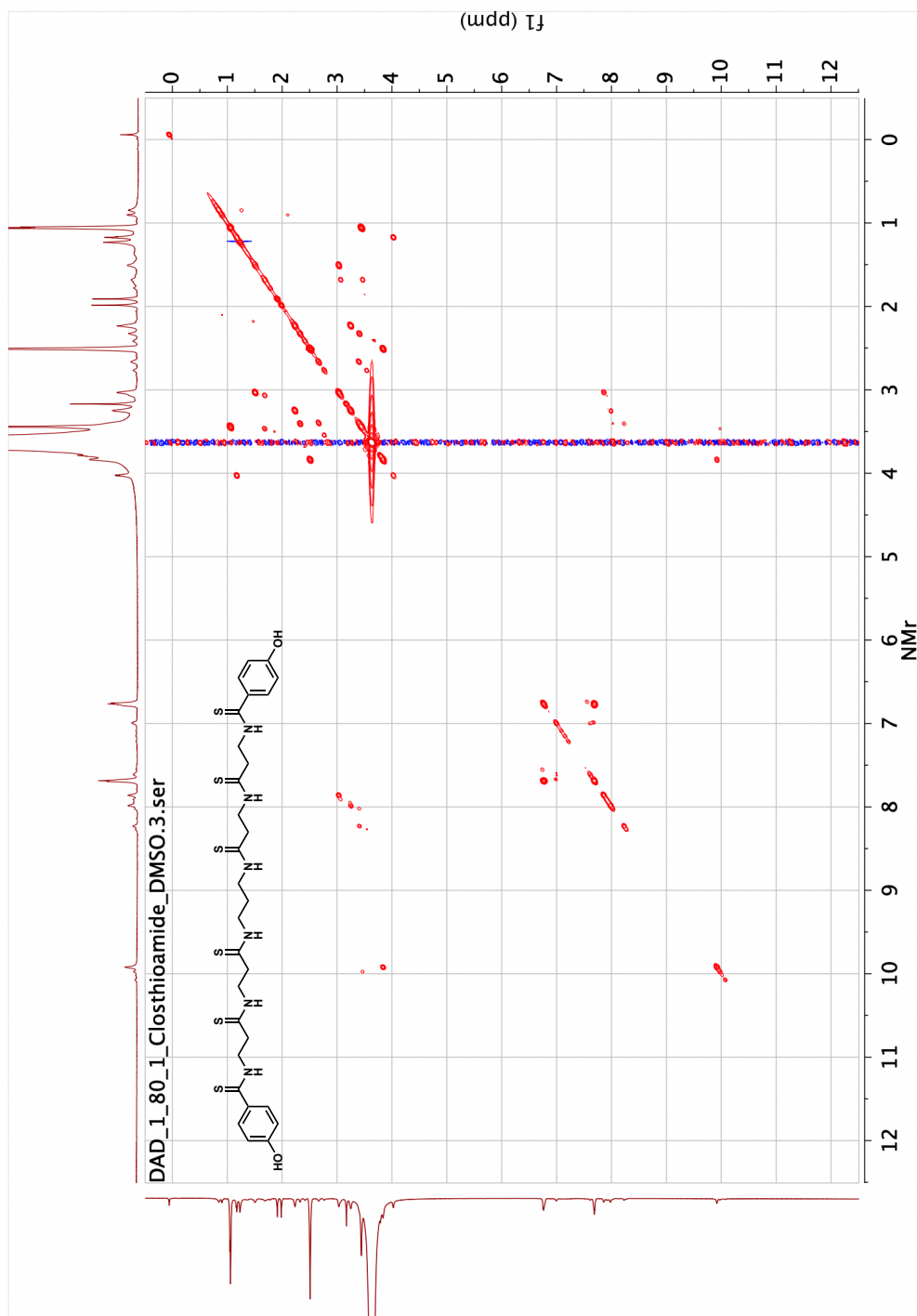


Figure 4.12 Closthioamide.

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name. ¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>3DMET</i>	generalistic	18248	x	yes	no	yes	no	no	no
<i>AfroCancer</i>	tm, plants, africa	390	365	yes	NA	no	no	NA	no
<i>AfroDB</i>	tm, plants, africa	954	874	yes	no	no	no	yes	no
<i>AfroMalariaDB</i>	tm, plants, africa	265	252	yes	NA	no	no	NA	no
<i>Afrotryp</i>	tm, plants, drug-like, africa	321	x	unknown	NA	no	no	NA	unknown
<i>Alkamid database</i>	plants, structure	300	x	yes	no	no	no	no	yes
<i>Ambinter-Greenpharma natural compound library (GPNCL)</i>	generalistic, industrial	>150000	x	no	yes	yes	yes	no	unknown
<i>AnalytiCon Discovery MEGx</i>	bacteria, plants, industrial	5147	4908	yes	yes	yes	no	yes	no
<i>AntiBase</i>	drug-like	>40000	x	no	no	no	yes	yes	unknown
<i>AntiMarin</i>	marine, drug-like	>60000	x	no	unknown	no	yes	unknown	unknown
<i>ATBD (Animal Toxin Database)</i>	toxins	1000	x	unknown	unknown	unknown	no	unknown	unknown
<i>Ayurveda</i>	tm, asia	950	x	no	yes	unknown	unknown	unknown	unknown
<i>Berdy's Bioactive Natural Products Database</i>	generalistic	x	x	no	unknown	no	yes	unknown	unknown
<i>BiGG</i>	metabolites	7339	x	yes	no	yes	no	NA	yes
<i>Binding DB</i>	drug-like	x	x	yes	no	no	no	yes	yes

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>BIOFAQUIM</i>	plant, fungi, america	420	400	yes	no	yes	no	yes	yes
<i>BioPhytMol</i>	drug-like, plants, asia	633	x	yes	no	yes	no	no	yes
<i>BitterDB</i>	food	654	631	yes	no	yes	no	no	yes
<i>BRENDA</i>	metabolites	x	x	yes	no	yes	no	yes	yes
<i>CamMedNP</i>	tm, plants, africa	>2500	x	yes, but proprietary format	no	no	no	yes, but proprietary format (MDB readable by MOE)	unknown
<i>Carotenoids Database</i>	structure	1174	991	yes	no	yes	no	yes	yes
<i>CAS registry/SciFinder</i>	chemicals	>300000	x	no	yes	yes	yes	unknown	unknown
<i>CEMTDD - Chinese Ethnic Minority Traditional Drug Database</i>	tm, plants, asia	4060	x	yes	no	no	no	no	yes
<i>CHDD (Chinese Traditional Medicinal Herbs database)</i>	tm, plants, asia	>30000	x	unknown	unknown	no	unknown	unknown	unknown
<i>ChEBI</i>	chemicals	15736	14621	yes	no	yes	no	yes	yes
<i>Chem-TCM</i>	plants, tm, asia	>12000	x	no	yes	no	yes	unknown	unknown
<i>ChemBank</i>	chemicals	x	x	yes	no	no	no	unknown	unknown
<i>ChEMBL</i>	chemicals	1899	1581	yes	no	yes	no	yes	no
<i>ChemBridge diversity datasets</i>	generalistic, industrial	x	x	no	yes	no	no	unknown	unknown
<i>ChemDB</i>	plants, asia	>1000	x	unknown	unknown	no	no	unknown	unknown

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>ChemIDplus</i>	drug-like, toxins	9042	x	yes	no	yes	no	no	no
<i>ChemSpider</i>	chemicals	9732	9029	yes	no	yes	no	yes	no
<i>CHMIS-C</i>	plants, tm, asia	>8000	x	yes	no	no	no	unknown	unknown
<i>CMAUP</i>	plants	47645	20873	yes	no	no	no	yes	yes
<i>CNPD (Chinese Natural Products Database)</i>	generalistic	>57000	x	unknown	unknown	no	unknown	unknown	unknown
<i>ConMedNP</i>	plants, tm, africa	3118	2504	yes	NA	NA	NA	NA	no
<i>CSLS/NCI (Chemical Structure Lookup Service)</i>	metabolites	x	x	yes	no	no	no	yes	no
<i>Database of Indonesian Medicinal Plants</i>	plants, tm, asia	6776	x	yes	no	no	no	no	no
<i>DESMSCI (Dragon Exploration System on Marine Sponge Compounds Interactions)</i>	marine	x	x	yes	no	no	no	no	unknown
<i>DFC (Dictionary of Food COmpounds)</i>	food	>41000	x	no	yes	yes	yes	yes	unknown
<i>DMNP (Dictionary of Marine Natural Products)</i>	marine	>30000	x	no	yes	yes	yes	yes	unknown
<i>DNP (Dictionary of Natural Products) by</i>	generalistic	> 230000	x	no	yes	yes	yes	yes	unknown

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>Chapman and Hall (also known as CHEMnetBase)</i>									
<i>Drugbank NPs</i>	drug-like	2617	2617	yes	no	yes	no	yes	yes
<i>eBasis</i>	food		x	no	yes	yes	yes	unknown	unknown
<i>ETCM (Encyclopedia of Traditional Chinese Medicine)</i>	tm, asia	7274	x	yes	no	yes	no	no	yes
<i>ETM-DB</i>	tm, plants, africa	1795	1653	yes	no	yes	no	no	yes
<i>FooDB</i>	food	24215	22223	yes	no	yes	no	yes	yes
<i>GNPS</i>	dereplication	7619	6708	yes	no	yes	no	yes	no
<i>HIM (Herbal Ingredients in-vivo Metabolism database)</i>	drug-like, tm, plants	1261	962	yes	no	no	no	unknown	unknown
<i>HIT (Herbal Ingredients Targets)</i>	drug-like, tm, plants	524	472	yes	no	no	no	unknown	unknown
<i>HMDB</i>	dereplication	x	x	yes	no	yes	no	yes	yes
<i>IMPPAT</i>	tm, plants, asia	9596	x	yes	no	yes	no	no	yes
<i>InflamNat</i>	drug-like	552	536	yes	NA	NA	NA	yes	yes
<i>Indofine Chemical Company Inc. natural products</i>	generalistic, industrial	56	46	yes	no		no	yes	no
<i>InPACdb</i>	drug-like, plants, asia	124	121	yes	no	no	no	yes	unknown

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>InterBioScreen Ltd (IBS)</i>	generalistic, industrial	68350	67292	yes	yes	yes	no	yes	no
<i>iSMART</i>	tm, plants, asia	x	x	yes	no	yes	no	no	no
<i>KEGG</i>	metabolites	x	x	yes	no	yes	no	no	no
<i>KNAPSaCK</i>	plants	10265	8887	yes	no	yes	no	no	yes
<i>Lichen Database</i>	fungi	249	156	yes	no	no	no	yes	yes
<i>LOPAC1280 by Merck</i>	drug-like	1280	x	no	yes	yes	unknown	unknown	unknown
<i>MAPS database</i>	plants, asia	x	x	unknown	unknown	no	no	unknown	unknown
<i>Marine Compound Database (MCDB)</i>	marine	182	x	yes	no	no	no	unknown	unknown
<i>Marine Natural Product Database (MNPD)</i>	marine	6000	x	yes	no	no	unknown	unknown	unknown
<i>MarineLit</i>	marine	>29000	x	no	yes	yes	yes	yes	unknown
<i>Massbank</i>	dereplication	x	x	yes	no	yes	no	no	no
<i>MedPServer</i>	plants, tm, asia, drug-like	1124	x	yes	no	yes	no	no	yes
<i>MetaCyc</i>	metabolites	x	x	yes	no	yes	no	yes	yes
<i>METLIN</i>	dereplication	x	x	yes	yes	yes	no	no	no
<i>Mitishamba database</i>	plants, africa	1102	1010	yes	no	no	no	no	yes
<i>NADI</i>	tm, plants	3000	x	no	yes	unknown	yes	unknown	unknown
<i>NANPDB</i>	plants, africa	6832	3913	yes	no	yes	no	yes	no

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>NaprAlert</i>	generalistic	>155000	x	no	yes	yes	yes	unknown	yes
<i>NAPROC-13</i>	dereplication	>18000	x	yes	no	yes	no	no	no
<i>NCI DTP data</i>	drug-like	418	404	yes	no	no	no	yes	no
<i>NeMedPlant</i>	tm, plants, asia	100	x	yes	no	no	no	no	yes
<i>NIST</i>	chemicals	x	x	no	no	yes	yes	yes	unknown
<i>NMRDATA</i>	dereplication	x	x	unknown	yes	yes	unknown	unknown	unknown
<i>NMRShiftDB</i>	dereplication	1875	x	yes	no	yes	no	yes	no
<i>Novel Antibiotics database</i>	drug-like	5430	x	yes	no	no	yes	no	yes
<i>NPACT</i>	plants, drug-like	1573	1453	yes	no	yes	no	yes	no
<i>NPASS</i>	plants, bacteria, metazoa, fungi	30858	27479	yes	no	yes	no	yes	yes
<i>NPAtlas</i>	bacteria, fungi	20035	18959	yes	no	yes	no	yes	yes
<i>NPCARE</i>	plants, marine, bacteria, drug-like	1370	1364	yes	no	yes	no	yes	no but contains impact of NPs on different cancer tissues and associated genes
<i>NPEdia</i>	generalistic	18016	16190	yes	no	no	no	no	yes
<i>NPL (library)</i>	plants, drug-like	814	x	no	NA	NA	NA	NA	unknown
<i>NuBBEDB</i>	plants, insects, america	2215	2022	yes	no		no	yes	no
<i>Open Source Malaria</i>	drug-like	842	x	yes	no	yes	no	yes	no

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>p</i> -ANAPL (Pan-African Natural Product Library)	plants, africa	538	467	yes	no		no	NA	no
PAMDB	metabolites, bacteria	x	x	yes	no	yes	no	yes	yes
Phenol-explorer	food	862	681	yes	no	NA	no	yes	yes
Phytochemica	plants, tm, asia	571	x	yes	no	no	no	no	yes
PhytoHub	food, plants	1200	x	yes	no	yes	no	no	yes
Pi Chemicals System Natural Products	generalistic, industrial	405	x	yes	no	yes	no	no	no
Prestwick	plants, industrial	320	x	no	yes	yes	yes	unknwn	unknown
ProCarDB	structure, bacteria	304	x	yes	no	no	no	no	yes
PubChem	chemicals	3529	2835	yes	no	yes	no	yes	no
REAXYS	chemicals	>220000	x	no	yes		yes	unknown	unknown
ReSpect	dereplication	4767	711	yes	no	no	no	yes	yes
SANCDDB	plants, africa	623	592	yes	no	yes	no	no	yes
Seaweed Metabolite Database (SWMD)	marine	1110	423	yes	no	no	no	yes	yes
Specs Natural Products	generalistic, industrial	745	745	yes	yes	no	no	unknwon	no
Spektraris NMR	dereplication	248	242	yes	no	no	no	yes	no
StreptomeDB	bacteria	6415	3610	yes	no	no	no	no	yes
Super Natural II	generalistic	320670	235436	yes	no	no	no	no	no

Table 3.2 List of Natural Products databases cited in scientific literature since 2000. The list is ordered alphabetical order of the database name.¹⁶³⁻³³⁶

Database name	NP type	Estimated size (number of NP molecules with correct structures)	Number of unique molecules in COCONUT	is open (data can be freely browsed)	requires a registration	is updated	is commercial	Molecule structures easily retrievable (download link, data packed in one file, bulk download option)	has extensive metadata (organism, tissue, geo info, ...)
<i>Super Scent</i>	other	2100	x	yes	no	no	no	no	no
<i>Super Sweet</i>	food, metabolites	15000	x	yes	no	no	no	no	no
<i>TargetMol Natural Compound Library</i>	generalistic, industrial	1680	x	no	yes	yes	yes	no	unkown
<i>TC-MC</i>	tm, asia, plants	>20000	x	yes	no	yes	no	no	yes
<i>TCMDB@Taiwan</i>	tm, asia, plants	58351	50891	yes	no	yes	no	yes	no
<i>TCMID</i>	tm, asia, plants	12549	10572	yes	no	no	no	no	yes (but difficult to extract)
<i>TCMSP</i>	tm, asia, plants	29384	x	yes	no	no	no	unknown	unknown
<i>TIM</i>	tm, asia, plants	1829	x	no	unknown	no	no	no	unknown
<i>TIPdb</i>	asia, plants, drug-like	8656	7752	yes	no	no	no	yes	no
<i>TMDB</i>	plants, metabolites	1393	x	yes	no	no	no	unknown	yes
<i>TPPT</i>	plants, toxins, europe	1583	1486	yes	no	no	no	yes	yes
<i>TriForC</i>	plants	266	x	yes	no	no	no	no	yes
<i>UEFS</i>	plants, america	503	481	yes	no	no	no	yes	no
<i>UNPD (Universal Natural Products Database)</i>	generalistic	213100	156984	yes	no	no	no	yes	no
<i>VIETHERB</i>	plants, asia	10887	x	yes	unknown	no	no	unknown	unknown
<i>YaTCM</i>	tm, asia, plants	47696	x	yes	no	yes	no	no	no
<i>ZINC natural products catalogue</i>	generalistic	85198	67336	yes	no	yes	no	yes	no

Table 3.3 Top scoring compounds in the NP Atlas with predicted antimicrobial activity.

Natural Products	50uM Inhibition	NPAID	Molecular Weight	Literature Bioactivity
<i>Thiazomycin</i>	0.88276043	NPA000855	1435.55	TRUE
<i>Enniatin F</i>	0.90372952	NPA000946	681.912	FAMILY
<i>Tolaasin C</i>	0.91668656	NPA002448	2005.475	FALSE
<i>Sch 54445</i>	0.94944679	NPA002511	597.02	TRUE
<i>Nocathiacin III</i>	0.8570237	NPA004612	1268.298	TRUE
<i>Tolaasin A</i>	0.85528483	NPA004692	1959.362	TRUE
<i>Scyptolin A</i>	0.85230808	NPA004834	981.542	TRUE
<i>MJ347-81F4-B</i>	0.89153915	NPA004941	1423.539	TRUE
<i>Isopyoverdin</i>	0.90438383	NPA005819	1195.184	UNKNOWN
<i>Not named</i>	0.95638429	NPA005907	1141.428	UNKNOWN
<i>Polymyxin E4</i>	0.95669381	NPA006001	1141.428	FAMILY
<i>Precolibactin C</i>	0.94074521	NPA006156	796.029	TRUE
<i>POH 2</i>	0.93932198	NPA006221	1030.192	UNKNOWN
<i>Actinomycin Y1</i>	0.85299864	NPA007759	1305.838	TRUE
<i>Antagonistic factor</i>	0.92958705	NPA008055	1204.483	TRUE
<i>Actinomycin Z3</i>	0.92574602	NPA008209	1319.865	FAMILY
<i>(+)-rugulosin A</i>	0.89788688	NPA008436	542.496	TRUE
<i>Thienamycin</i>	0.87884712	NPA008843	272.326	TRUE
<i>Precolibactin B</i>	0.92184092	NPA008996	712.914	TRUE
<i>6-demethyltetracycline</i>	0.91438579	NPA009270	430.413	FAMILY
<i>Polymyxin B5</i>	0.94657051	NPA009889	1203.499	TRUE
<i>Polymyxin-P1</i>	0.95743854	NPA010156	1191.444	TRUE
<i>Dactylocycline B</i>	0.94511582	NPA010526	712.105	TRUE
<i>5-Hydroxy-7-chlortetracycline</i>	0.98724271	NPA010649	494.884	FAMILY
<i>Tetracycline</i>	0.91914074	NPA010761	444.44	TRUE
<i>Enterolysin A</i>	0.87823657	NPA010873	1613.088	TRUE
<i>E. coli ferritin</i>	0.88475053	NPA010924	1152.384	UNKNOWN
<i>Enterocin EJ97</i>	0.91271102	NPA010953	1834.376	TRUE
<i>Plectosphaeroic acid B</i>	0.91889298	NPA010992	792.852	UNKNOWN
<i>Cypemycin</i>	0.8765226	NPA012286	2096.532	TRUE
<i>Plw-alpha</i>	0.88969634	NPA012477	1798.19	FALSE
<i>Plantaricin W</i>	0.94437841	NPA012512	1751.094	FALSE
<i>Gassericin B2</i>	0.93412083	NPA012635	1032.233	FAMILY
<i>Not named</i>	0.95998715	NPA012930	1155.455	UNKNOWN
<i>Polymyxin E7</i>	0.95851634	NPA013003	1169.482	FAMILY

<i>Terramycin</i>	0.97405366	NPA014607	460.439	TRUE
<i>MJ347-81F4-A</i>	0.90775723	NPA014699	1437.566	TRUE
<i>Plectosphaeroic acid C</i>	0.87970339	NPA014844	810.848	UNKNOWN
<i>NC0604</i>	0.89455987	NPA014943	1511.664	TRUE
<i>Sporidesmin E</i>	0.85439744	NPA015639	506.027	FAMILY
<i>Polymyxin Ile-E8</i>	0.96373628	NPA015649	1183.509	TRUE
<i>Nocathiacin I</i>	0.9194022	NPA016398	1437.566	TRUE
<i>Dactylocycline A</i>	0.92815099	NPA017120	698.122	TRUE
<i>Polymyxin Nva-E2</i>	0.96279448	NPA017349	1127.401	FAMILY
<i>Polymyxin Nva-E1</i>	0.96265357	NPA017927	1141.428	FAMILY
<i>Porfiromycin</i>	0.8962043	NPA018217	348.359	TRUE
<i>Polymyxin B6</i>	0.93444857	NPA018392	1219.498	TRUE
<i>Leporzine C</i>	0.90074278	NPA018659	589.692	UNKNOWN
<i>Thiazomycin E1</i>	0.87389428	NPA019285	1254.4	FALSE
<i>Thiazomycin A</i>	0.90039885	NPA019485	1449.577	TRUE
<i>Bleomycin A2</i>	0.9244478	NPA020036	1415.576	TRUE
<i>Bleomycin B2</i>	0.8900986	NPA020040	1425.53	TRUE
<i>Bleomycin A2'-b</i>	0.92294955	NPA020041	1369.462	TRUE
<i>Putisolvin I</i>	0.87089291	NPA020351	1380.691	FALSE
<i>Putisolvin II</i>	0.88751653	NPA020352	1380.691	FALSE
<i>Mitomycin A</i>	0.85054562	NPA020530	349.343	TRUE
<i>Mitomycin C</i>	0.89814951	NPA020531	334.332	TRUE
<i>Actinoplanone A</i>	0.8782621	NPA020793	584.965	TRUE
<i>Cleomycin B2</i>	0.86025122	NPA021050	1437.541	TRUE
<i>Tallysomycin B</i>	0.94876293	NPA021115	1586.683	TRUE
<i>Tallysomycin A</i>	0.93939183	NPA021116	1714.858	TRUE
<i>Polymyxin S1</i>	0.86964102	NPA021130	1178.401	FAMILY
<i>Bleomycin B1'</i>	0.93940063	NPA021447	1312.366	TRUE
<i>Bleomycin A2'-C</i>	0.85810379	NPA021448	1406.483	TRUE
<i>Bleomycin A2'-A</i>	0.93672009	NPA021449	1383.489	TRUE
<i>Bleomycin demethyl A2</i>	0.94856917	NPA021450	1400.541	TRUE
<i>Paenialvin A</i>	0.8737903	NPA021780	1892.451	TRUE
<i>Paenialvin B</i>	0.86614138	NPA021781	1876.452	TRUE
<i>Paenialvin C</i>	0.86440064	NPA021782	1878.424	TRUE
<i>Paenialvin D</i>	0.91274652	NPA021783	1924.493	TRUE
<i>Cerecidin A1</i>	0.87626275	NPA024682	1988.542	TRUE
<i>Grisemycin</i>	0.86647432	NPA024684	1833.235	FALSE

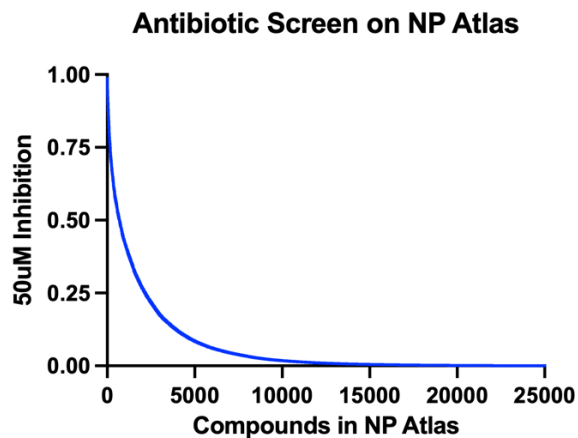


Figure 3.6 Results of virtual screen of NP Atlas for the identification of NPs with antimicrobial activity.

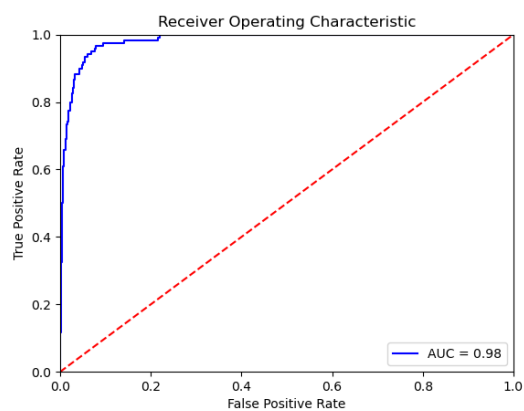


Figure 3.7 Area Under the Curve Receiver Operating Characteristics Curve for DMPNN trained on Ecoli data set.

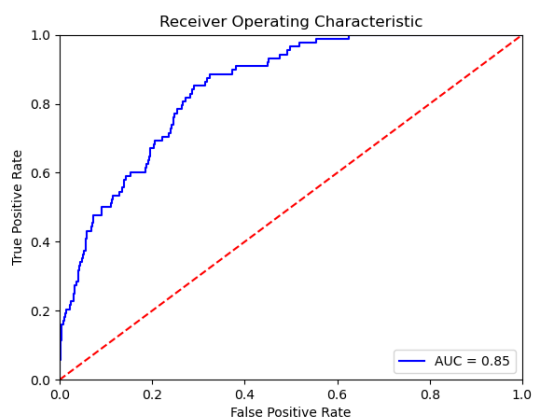


Figure 4.12 Area Under the Curve Receiver Operating Characteristics Curve for DMPNN trained on AMU_sars_cov_2 data set.

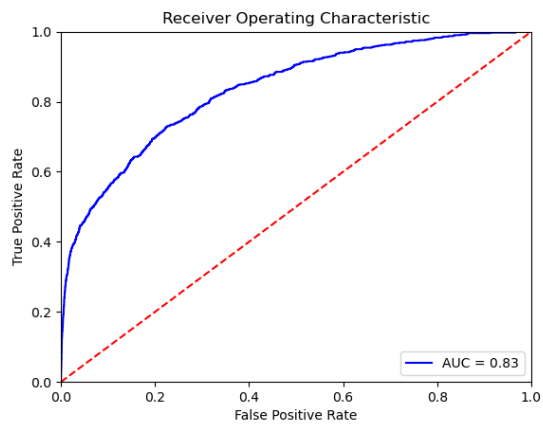


Figure 4.13 Area Under the Curve Receiver Operating Characteristics Curve for DMPNN trained on PLpro data set.

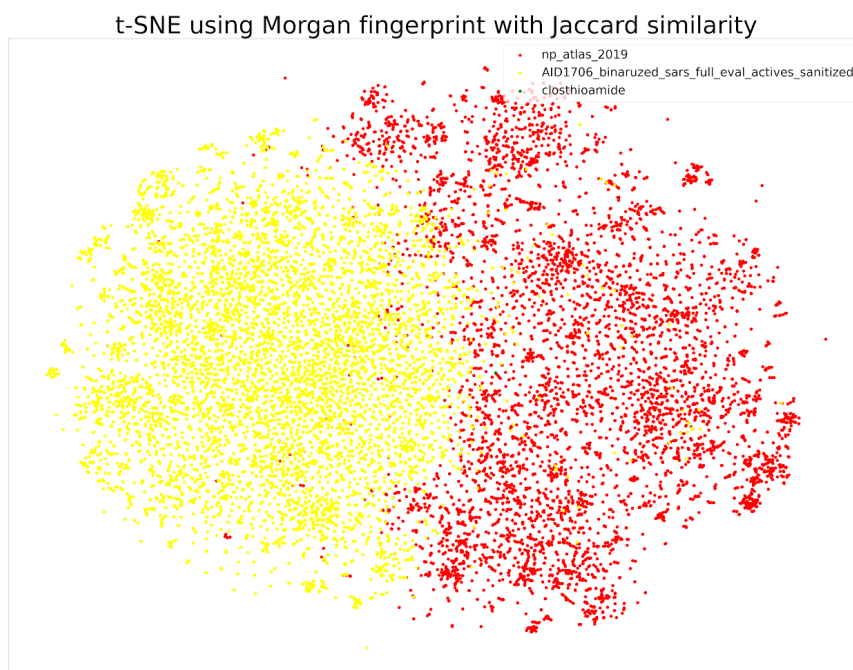


Figure 4.14 t-SNE of the training set (yellow), the natural products atlas (red), and the identified bioactive compound closthioamide (green).

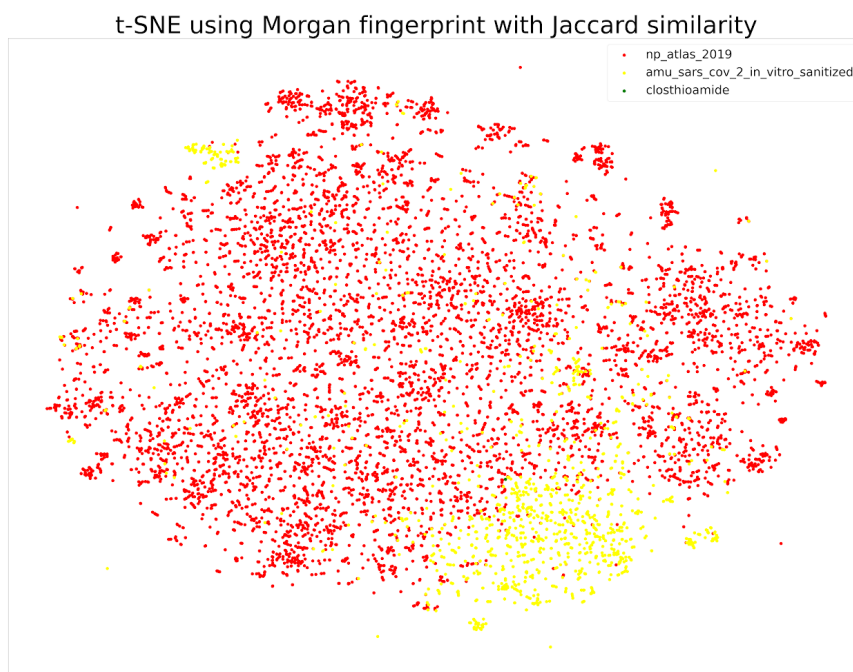


Figure 4.15 t-SNE of the training set (yellow), the natural products atlas (red), and the identified bioactive compound closthioamide (green).

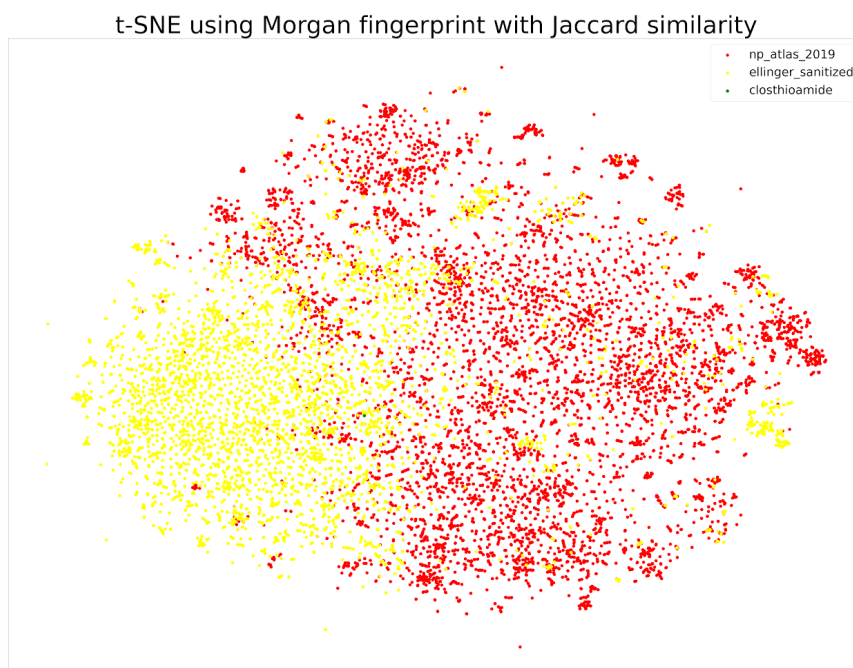


Figure 4.16 t-SNE of the training set (yellow), the natural products atlas (red), and the identified bioactive compound closthioamide (green).

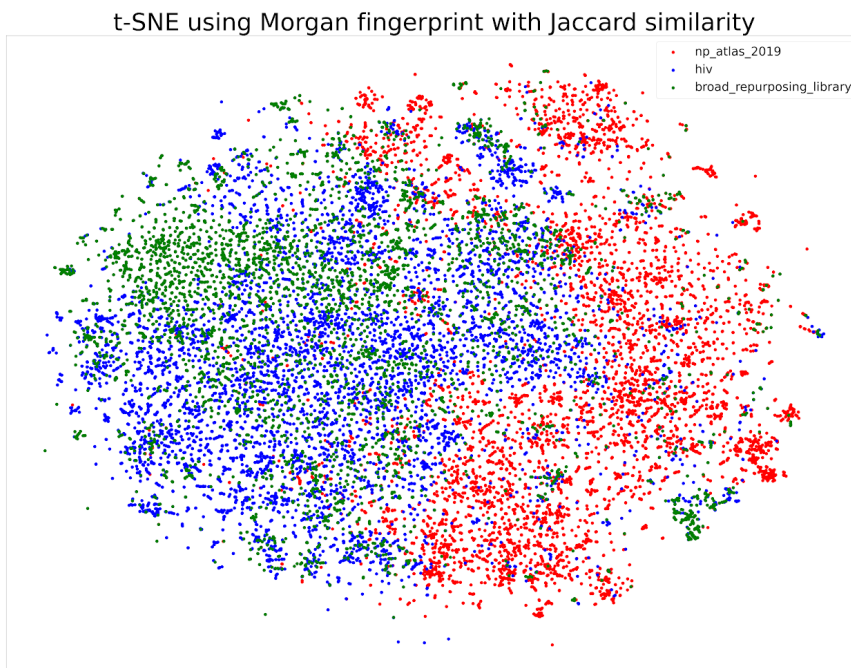


Figure 4.17 t-SNE of the natural products atlas (red), hiv dataset (blue), and the broad repurposing library (green).

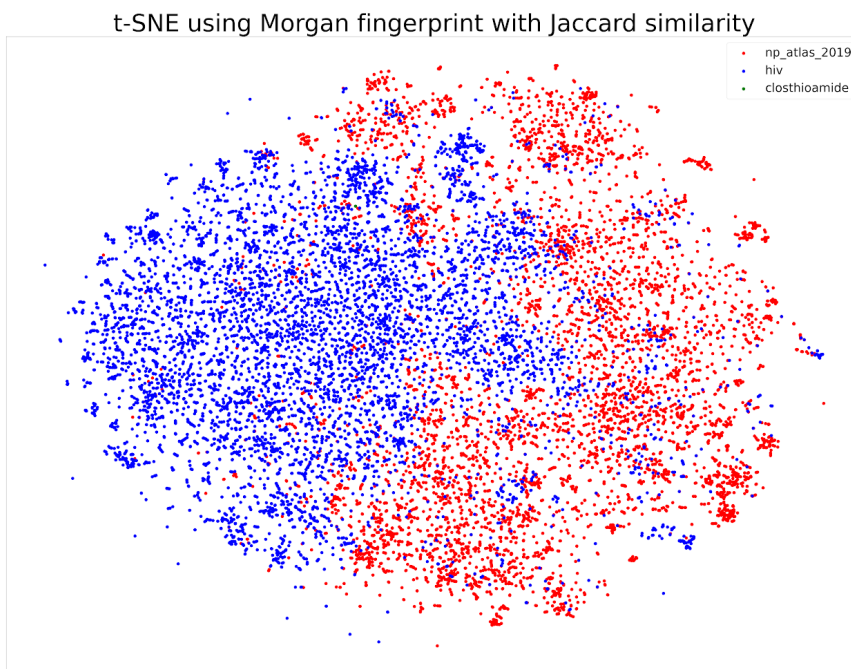


Figure 4.18 t-SNE of the natural products atlas (red), hiv dataset (blue), and closthioamide(green).

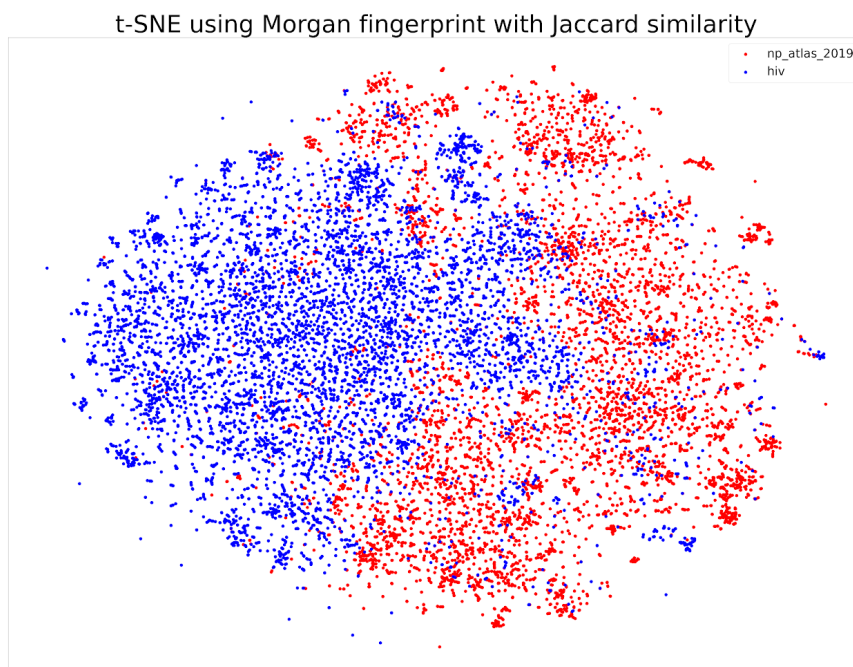


Figure 4.19 t-SNE of the natural products atlas (red) and the hiv drug repository (blue).

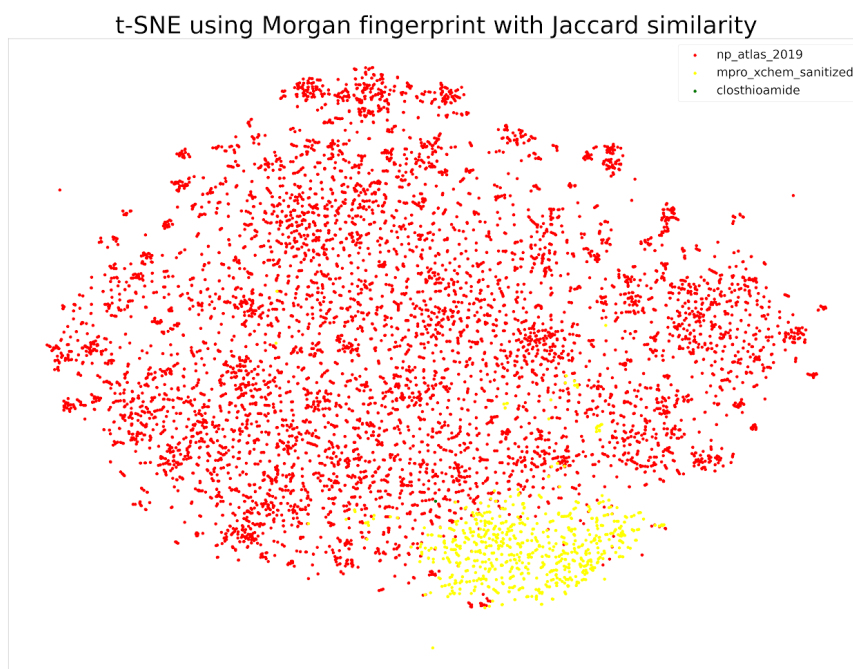


Figure 4.20 t-SNE of the training set (yellow), the natural products atlas (red), and the identified bioactive compound closthioamide (green).

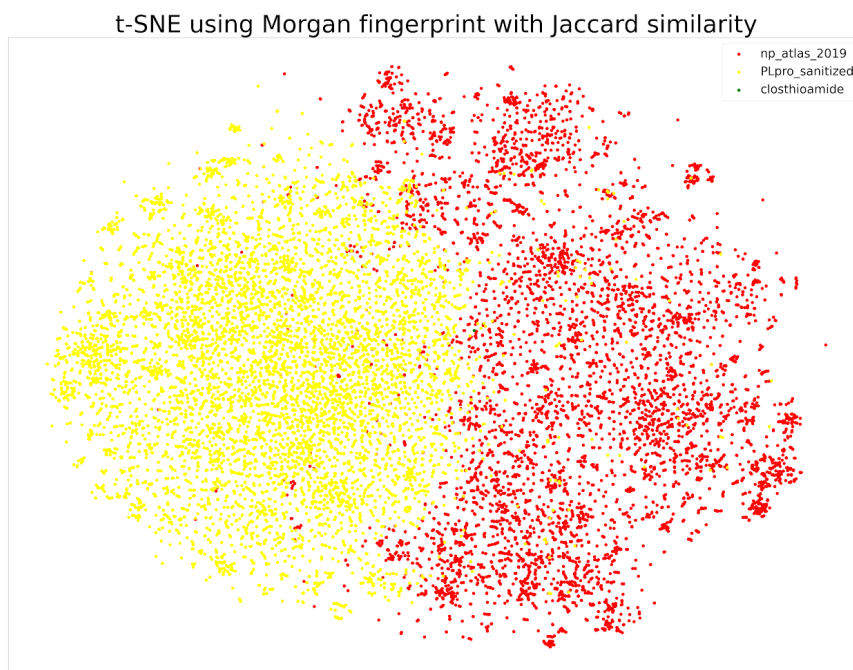


Figure 4.21 t-SNE of the training set (yellow), the natural products atlas (red), and the identified bioactive compound closthioamide (green).

Bibliography

1. Namdeo, A. G. & Others. Plant cell elicitation for production of secondary metabolites: a review. *Pharmacogn. Rev.* **1**, 69–79 (2007).
2. Schatz, A. & Waksman, S. A. Effect of Streptomycin and Other Antibiotic Substances upon Mycobacterium tuberculosis and Related Organisms.,. *Proc. Soc. Exp. Biol. Med.* **57**, 244–248 (1944).
3. Burg, R. W. *et al.* Avermectins, new family of potent anthelmintic agents: producing organism and fermentation. *Antimicrob. Agents Chemother.* **15**, 361–367 (1979).
4. Egerton, J. R. *et al.* Avermectins, new family of potent anthelmintic agents: efficacy of the B1a component. *Antimicrob. Agents Chemother.* **15**, 372–378 (1979).
5. Månsson, M. *et al.* Explorative solid-phase extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification. *J. Nat. Prod.* **73**, 1126–1132 (2010).
6. Henrich, C. J. *et al.* Deguelins, Natural Product Modulators of NF1-Defective Astrocytoma Cell Growth Identified by High-Throughput Screening of Partially Purified Natural Product Extracts. *J. Nat. Prod.* **78**, 2776–2781 (2015).
7. Scheuermann, T. H. *et al.* Allosteric inhibition of hypoxia inducible factor-2 with small molecules. *Nat. Chem. Biol.* **9**, 271–276 (2013).
8. Scheuermann, T. H. *et al.* Isoform-Selective and Stereoselective Inhibition of Hypoxia Inducible Factor-2. *J. Med. Chem.* **58**, 5930–5941 (2015).
9. Chen, W. *et al.* Targeting renal cell carcinoma with a HIF-2 antagonist. *Nature* **539**, 112–117 (2016).
10. Nakajima, H., Kim, Y. B., Terano, H., Yoshida, M. & Horinouchi, S. FR901228, a Potent Antitumor Antibiotic, Is a Novel Histone Deacetylase Inhibitor. *Experimental Cell Research* vol. 241 126–133 (1998).
11. Ueda, H. *et al.* FR901228, a novel antitumor bicyclic depsipeptide produced by chromobacterium violaceum No. 968. I. Taxonomy, fermentation, isolation, physico-

- chemical and biological properties, and antitumor activity. *The Journal of Antibiotics* vol. 47 301–310 (1994).
12. Feling, R. H. *et al.* Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus *Salinospora*. *ChemInform* vol. 34 (2003).
 13. Low, W.-K. *et al.* Inhibition of eukaryotic translation initiation by the marine natural product pateamine A. *Mol. Cell* **20**, 709–722 (2005).
 14. Harding, M. W. Galat, a, Uehling, DE & Schreiber, SL A receptor for the immunosuppressant FK506 is a cis-trans peptidyl-prolyl isomerase. *Nature* **341**, 758–760 (1989).
 15. Kino, T. *et al.* FK-506, a novel immunosuppressant isolated from a *Streptomyces*. I. Fermentation, isolation, and physico-chemical and biological characteristics. *The Journal of Antibiotics* vol. 40 1249–1255 (1987).
 16. Kitagawa, M., Ikeda, S., Tashiro, E., Soga, T. & Imoto, M. Metabolomic identification of the target of the filopodia protrusion inhibitor glucopiericidin A. *Chem. Biol.* **17**, 989–998 (2010).
 17. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
 18. Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P. & McPhail, A. T. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.* **93**, 2325–2327 (1971).
 19. Chan, B. A. & Hughes, B. G. M. Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. *Transl. Lung Cancer Res.* **4**, 36–54 (2015).
 20. National Comprehensive Cancer Network® (NCCN®). *NCCN Guidelines for Patients® Lung Cancer Screening*. (National Comprehensive Cancer Network® (NCCN®), 2019).
 21. Lim, J. U. Overcoming Osimertinib Resistance in Advanced Non-small Cell Lung Cancer. *Clinical Oncology* vol. 33 619–626 (2021).

22. Scagliotti, G. V. *et al.* Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J. Clin. Oncol.* **26**, 3543–3551 (2008).
23. MRSNY, R. J. & DAUGHERTY, A. *Proteins and Peptides: Pharmacokinetic, Pharmacodynamic, and Metabolic Outcomes.* (CRC Press, 2009).
24. Reck, M. *et al.* Overall survival with cisplatin-gemcitabine and bevacizumab or placebo as first-line therapy for nonsquamous non-small-cell lung cancer: results from a randomised phase III trial (AVAL). *Ann. Oncol.* **21**, 1804–1809 (2010).
25. Ramalingam, S. S. *et al.* Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. *N. Engl. J. Med.* **382**, 41–50 (2020).
26. Wu, Y.-L. *et al.* Osimertinib in Resected EGFR-Mutated Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **383**, 1711–1723 (2020).
27. Vaishnavi, A. *et al.* Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer. *Nat. Med.* **19**, 1469–1472 (2013).
28. Wang, W. *et al.* P2.03-09 The Real World of NTRK Fusion Data in the Chinese Lung Cancer Populations: A Multicenter Study. *J. Thorac. Oncol.* **13**, S719 (2018).
29. Guo, Y. *et al.* Recent Progress in Rare Oncogenic Drivers and Targeted Therapy For Non-Small Cell Lung Cancer. *Oncotargets. Ther.* **12**, 10343–10360 (2019).
30. Drilon, A. *et al.* Efficacy of Selpercatinib in RET Fusion-Positive Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **383**, 813–824 (2020).
31. Gainor, J. F. *et al.* Registrational dataset from the phase I/II ARROW trial of pralsetinib (BLU-667) in patients (pts) with advanced RET fusion non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* vol. 38 9515–9515 (2020).
32. Drilon, A. *et al.* Cabozantinib in patients with advanced RET-rearranged non-small-cell lung cancer: an open-label, single-centre, phase 2, single-arm trial. *Lancet Oncol.* **17**, 1653–1660 (2016).
33. Yoh, K. *et al.* Vandetanib in patients with previously treated RET-rearranged advanced

- non-small-cell lung cancer (LURET): an open-label, multicentre phase 2 trial. *Lancet Respir Med* **5**, 42–50 (2017).
34. Sezer, A. *et al.* Cemiplimab monotherapy for first-line treatment of advanced non-small-cell lung cancer with PD-L1 of at least 50%: a multicentre, open-label, global, phase 3, randomised, controlled trial. *Lancet* **397**, 592–604 (2021).
 35. Mok, T. S. K. *et al.* Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. *Lancet* **393**, 1819–1830 (2019).
 36. Herbst, R. S. *et al.* Atezolizumab for First-Line Treatment of PD-L1-Selected Patients with NSCLC. *N. Engl. J. Med.* **383**, 1328–1339 (2020).
 37. Minchom, A. R. *et al.* Amivantamab compared with real-world therapies in patients with NSCLC with EGFR Exon 20 insertion mutations who have progressed after platinum doublet chemotherapy. *J. Clin. Orthod.* **39**, 9052–9052 (2021).
 38. Li, B. T. *et al.* Ado-Trastuzumab Emtansine for Patients With HER2-Mutant Lung Cancers: Results From a Phase II Basket Trial. *J. Clin. Oncol.* **36**, 2532–2537 (2018).
 39. Smit, E. F. *et al.* Trastuzumab deruxtecan (T-DXd; DS-8201) in patients with HER2-mutated metastatic non-small cell lung cancer (NSCLC): Interim results of DESTINY-Lung01. *J. Clin. Orthod.* **38**, 9504–9504 (2020).
 40. Tanaka, N. *et al.* Clinical Acquired Resistance to KRASG12C Inhibition through a Novel KRAS Switch-II Pocket Mutation and Polyclonal Alterations Converging on RAS-MAPK Reactivation. *Cancer discovery* vol. 11 1913–1922 (2021).
 41. Cancer Target Discovery and Development Network *et al.* Towards patient-based cancer therapeutics. *Nat. Biotechnol.* **28**, 904–906 (2010).
 42. Cantrell, M. A. & Kuo, C. J. Organoid modeling for cancer precision medicine. *Genome Med.* **7**, 32 (2015).
 43. McMillan, E. A. *et al.* Chemistry-First Approach for Nomination of Personalized

- Treatment in Lung Cancer. *Cell* **173**, 864–878.e29 (2018).
44. Fojo, T. *et al.* Identification of non-cross-resistant platinum compounds with novel cytotoxicity profiles using the NCI anticancer drug screen and clustered image map visualizations. *Crit. Rev. Oncol. Hematol.* **53**, 25–34 (2005).
 45. Jomon, K., Kuroda, Y., Ajisaka, M. & Sakai, H. A new antibiotic, ikarugamycin. *J. Antibiot.* **25**, 271–280 (1972).
 46. Gunasekera, S. P., Gunasekera, M. & McCarthy, P. Discodermide: a new bioactive macrocyclic lactam from the marine sponge *Discodermia dissoluta*. *J. Org. Chem.* **56**, 4830–4833 (1991).
 47. Kobayashi, J. & Ishibashi, M. Bioactive metabolites of symbiotic marine microorganisms. *Chem. Rev.* **93**, 1753–1769 (1993).
 48. Ito, S. & Hirata, Y. The Structure of Ikarugamycin, an Acyltetramic Acid Antibiotic Possessing a Unique as-Hydrindacene Skeleton. *BCSJ* **50**, 1813–1820 (1977).
 49. Lacret, R. *et al.* New ikarugamycin derivatives with antifungal and antibacterial properties from *Streptomyces zhaozhouensis*. *Mar. Drugs* **13**, 128–140 (2014).
 50. Kanazawa, S., Fusetani, N. & Matsunaga, S. Cylindramide: Cytotoxic tetramic acid lactam from the marine sponge *Halichondria cylindrata* Tanita & Hoshino. *Tetrahedron Lett.* **34**, 1065–1068 (1993).
 51. Yu, H.-L. *et al.* Structural diversity of anti-pancreatic cancer capsimycins identified in mangrove-derived *Streptomyces xiamenensis* 318 and post-modification via a novel cytochrome P450 monooxygenase. *Scientific Reports* vol. 7 (2017).
 52. Popescu, R. *et al.* Ikarugamycin induces DNA damage, intracellular calcium increase, p38 MAP kinase activation and apoptosis in HL-60 human promyelocytic leukemia cells. *Mutat. Res.* **709-710**, 60–66 (2011).
 53. Hutchins, R. O. & Natale, N. R. Sodium borohydride in acetic acid. A convenient system for the reductive deoxygenation of carbonyl tosylhydrazones. *J. Org. Chem.* **43**, 2299–2301 (1978).

54. Hook, J. M. & Mander, L. N. Recent developments in the Birch reduction of aromatic compounds: applications to the synthesis of natural products. *Nat. Prod. Rep.* **3**, 35–85 (1986).
55. Akagawa, K., Hirata, T. & Kudo, K. ChemInform Abstract: Asymmetric Epoxidation of Enones by Peptide-Based Catalyst: A Strategy Inverting Julia-Colonna Stereoselectivity. *ChemInform* vol. 47 (2016).
56. Chen, W.-P. & Roberts, S. M. Julia–Colonna asymmetric epoxidation of furyl styryl ketone as a route to intermediates to naturally-occurring styryl lactones. *J. Chem. Soc. Perkin 1* **0**, 103–106 (1999).
57. Muchiri, R. N. & van Breemen, R. B. Drug discovery from natural products using affinity selection-mass spectrometry. *Drug Discov. Today Technol.* **40**, 59–63 (2021).
58. Suenaga, H. *et al.* Phenotypic screening system using three-dimensional (3D) culture models for natural product screening. *J. Antibiot.* **74**, 660–666 (2021).
59. Buss, A. D. & Butler, M. S. *Natural Product Chemistry for Drug Discovery*. (Royal Society of Chemistry, 2010).
60. Wilson, B. A. P., Thornburg, C. C., Henrich, C. J., Grkovic, T. & O’Keefe, B. R. Creating and screening natural product libraries. *Nat. Prod. Rep.* **37**, 893–918 (2020).
61. Newman, D. Faculty Opinions recommendation of Natural Products in High Throughput Screening: Automated High-Quality Sample Preparation. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2017)
doi:10.3410/f.727591292.793531724.
62. Butler, M. S. The role of natural product chemistry in drug discovery. *J. Nat. Prod.* **67**, 2141–2153 (2004).
63. Chávez-Hernández, A. L., Sánchez-Cruz, N. & Medina-Franco, J. L. Fragment Library of Natural Products and Compound Databases for Drug Discovery. *Biomolecules* **10**, (2020).
64. Tio, G. A. STRUCTURE-BASED VIRTUAL SCREENING OF INDONESIAN NATURAL

PRODUCT COMPOUNDS AS EBOLA VIRUS VP30 PROTEIN INHIBITORS.

International Journal of GEOMATE vol. 17 (2019).

65. Rollinger, J. M., Stuppner, H. & Langer, T. Virtual screening for the discovery of bioactive natural products. *Prog. Drug Res.* **65**, 211, 213–49 (2008).
66. Lung, J. *et al.* Virtual Screening and In Vitro Evaluation of PD-1 Dimer Stabilizers for Uncoupling PD-1/PD-L1 Interaction from Natural Products. *Molecules* **25**, (2020).
67. Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **12**, 20 (2020).
68. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminform.* **13**, 2 (2021).
69. van Santen, J. A. *et al.* The Natural Products Atlas: An open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
70. Laatsch, H. *AntiBase: the natural compound identifier.* (Wiley-Vch Weinheim, 2017).
71. Chassagne, F., Cabanac, G., Hubert, G., David, B. & Marti, G. The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products®. *Phytochem. Rev.* **18**, 601–622 (2019).
72. Blunt, J. W. & Munro, M. H. G. MarinLit database. *University of Canterbury* (2012).
73. Ntie-Kang, F. *et al.* AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* **8**, e78085 (2013).
74. Valli, M. *et al.* Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* **76**, 439–444 (2013).
75. Lucas, X. *et al.* StreptomeDB: a resource for natural compounds isolated from *Streptomyces* species. *Nucleic Acids Res.* **41**, D1130–6 (2013).
76. Luo, L., Zhong, A., Wang, Q. & Zheng, T. Structure-Based Pharmacophore Modeling, Virtual Screening, Molecular Docking, ADMET, and Molecular Dynamics (MD) Simulation of Potential Inhibitors of PD-L1 from the Library of Marine Natural Products. *Mar. Drugs* **20**, (2021).

77. Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M. & Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front Chem* **8**, 343 (2020).
78. Santana, K. *et al.* Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. *Front Chem* **9**, 662688 (2021).
79. Basak, S. C. Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. *Curr. Comput. Aided Drug Des.* **8**, 1–2 (2012).
80. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **18**, 435–441 (2019).
81. Klambauer, G., Hochreiter, S. & Rarey, M. Machine Learning in Drug Discovery. *J. Chem. Inf. Model.* **59**, 945–946 (2019).
82. Brown, N. *Artificial Intelligence in Drug Discovery*. (Royal Society of Chemistry, 2020).
83. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **56**, 237–248 (2006).
84. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
85. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. http://www.rdkit.org/RDKit_Overview.pdf.
86. Hall, L. H., Mohney, B. & Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **31**, 76–82 (1991).
87. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
88. Yang, K. *et al.* Are Learned Molecular Representations Ready for Prime Time? *ChemRxiv* (2019) doi:10.26434/chemrxiv.7940594.v2.
89.) K. Y. (M. *Are Learned Molecular Representations Ready for Prime Time?* (Massachusetts Institute of Technology, Department of Electrical Engineering and

- Computer Science, 2019).
90. Swanson, K. W.). Message passing neural networks for molecular property prediction. (Massachusetts Institute of Technology, 2019).
 91. Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzae. *Br. J. Exp. Pathol.* **10**, 226 (1929).
 92. Cordell, G. A. & Colvard, M. D. Natural products and traditional medicine: turning on a paradigm. *J. Nat. Prod.* **75**, 514–525 (2012).
 93. Cole, M. A. & Elkan, G. H. Transmissible resistance to penicillin G, neomycin, and chloramphenicol in *Rhizobium japonicum*. *Antimicrob. Agents Chemother.* **4**, 248–253 (1973).
 94. Nguyen, F. *et al.* Tetracycline antibiotics and resistance mechanisms. *Biol. Chem.* **395**, 559–575 (2014).
 95. Heilman, F. R., Herrell, W. E., Wellman, W. E. & Geraci, J. E. Some laboratory and clinical observations on a new antibiotic, erythromycin (ilotycin). *Proc. Staff Meet. Mayo Clin.* **27**, 285–304 (1952).
 96. Geraci, J. E., Heilman, F. R., Nichols, D. R., Ross, G. T. & Wellman, W. E. Some laboratory and clinical experiences with a new antibiotic, vancomycin. *Proc. Staff Meet. Mayo Clin.* **31**, 564–582 (1956).
 97. Yoshizawa, S., Fourmy, D. & Puglisi, J. D. Structural origins of gentamicin antibiotic action. *EMBO J.* **17**, 6437–6448 (1998).
 98. Zimbelman, J., Palmer, A. & Todd, J. Improved outcome of clindamycin compared with beta-lactam antibiotic treatment for invasive *Streptococcus pyogenes* infection. *Pediatr. Infect. Dis. J.* **18**, 1096–1100 (1999).
 99. Lipton, J. H. Incompatibility between Sulfamethizole and Methenamine Mandelate. *N. Engl. J. Med.* **268**, 92–93 (1963).
 100. Moellering, R. C. Linezolid: the first oxazolidinone antimicrobial. *Ann. Intern. Med.* **138**,

- 135–142 (2003).
101. Crump, B., Wise, R. & Dent, J. Pharmacokinetics and tissue penetration of ciprofloxacin. *Antimicrob. Agents Chemother.* **24**, 784–786 (1983).
102. Davies Julian & Davies Dorothy. Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
103. Lee, A. S. *et al.* Methicillin-resistant *Staphylococcus aureus*. *Nat Rev Dis Primers* **4**, 18033 (2018).
104. Kennedy, D. Time to deal with antibiotics. *Science* **342**, 777 (2013).
105. Martens, E. & Demain, A. L. The antibiotic resistance crisis, with a focus on the United States. *J. Antibiot.* **70**, 520–526 (2017).
106. World Health Organization. *WHO consolidated guidelines on tuberculosis. Module 4: treatment - drug-resistant tuberculosis treatment.* (World Health Organization, 2020).
107. Brown, D. G., May-Dracka, T. L., Gagnon, M. M. & Tommasi, R. Trends and exceptions of physical properties on antibacterial activity for Gram-positive and Gram-negative pathogens. *J. Med. Chem.* **57**, 10144–10161 (2014).
108. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **181**, 475–483 (2020).
109. Kramer, O. Scikit-Learn. in *Machine Learning for Evolution Strategies* (ed. Kramer, O.) 45–53 (Springer International Publishing, 2016).
110. Dunn, T. B. *et al.* Diversity and Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c01013.
111. Álvarez-Bardón, M. *et al.* Screening Marine Natural Products for New Drug Leads against Trypanosomatids and Malaria. *Mar. Drugs* **18**, (2020).
112. Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13**, 509–523 (2015).
113. Kremb, S. & Voolstra, C. R. High-resolution phenotypic profiling of natural products-

- induced effects on the single-cell level. *Sci. Rep.* **7**, 44472 (2017).
114. Wu, C., Kim, H. K., Van Wezel, G. P. & Choi, Y. H. Metabolomics in the natural products field--a gateway to novel antibiotics. *Drug Discov. Today Technol.* **13**, 11–17 (2015).
115. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., International Natural Product Sciences Taskforce & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
116. Zeng, X. *et al.* NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
117. Zeng, X. *et al.* CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res.* **47**, D1118–D1127 (2019).
118. National Academies of Sciences, Engineering, and Medicine *et al.* *Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies*. (National Academies Press, 2020).
119. Gao, J. & Sun, F. Drug discovery to treat COVID-19 two years after its outbreak. *Drug Discov. Ther.* **15**, 281–288 (2021).
120. Singh, T. U. *et al.* Drug repurposing approach to fight COVID-19. *Pharmacol. Rep.* **72**, 1479–1508 (2020).
121. King, A. M. *et al.* Selection for constrained peptides that bind to a single target protein. *Nat. Commun.* **12**, 6343 (2021).
122. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
123. Allam, A. E. *et al.* In silico study of natural compounds from sesame against COVID-19 by targeting M pro , PL pro and RdRp. *RSC Adv.* **11**, 22398–22408 (2021).
124. Pruijssers, A. J. *et al.* Remdesivir Inhibits SARS-CoV-2 in Human Lung Cells and Chimeric SARS-CoV Expressing the SARS-CoV-2 RNA Polymerase in Mice. *Cell Rep.* **32**, 107940 (2020).

125. Sheahan, T. P. *et al.* An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Sci. Transl. Med.* **12**, (2020).
126. Riva, L. *et al.* Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113–119 (2020).
127. Gayle, S. *et al.* Identification of apilimod as a first-in-class PIKfyve kinase inhibitor for treatment of B-cell non-Hodgkin lymphoma. *Blood* **129**, 1768–1778 (2017).
128. Fu, L. *et al.* Both Boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease. *Nat. Commun.* **11**, 4417 (2020).
129. Agostini, M. L. *et al.* Small-Molecule Antiviral β -d-N 4-Hydroxycytidine Inhibits a Proofreading-Intact Coronavirus with a High Genetic Barrier to Resistance. *J. Virol.* **93**, (2019).
130. Huang, J., Song, W., Huang, H. & Sun, Q. Pharmacological Therapeutics Targeting RNA-Dependent RNA Polymerase, Proteinase and Spike Protein: From Mechanistic Studies to Clinical Trials for COVID-19. *J. Clin. Med. Res.* **9**, (2020).
131. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J. Enzyme Inhib. Med. Chem.* **36**, 1646–1650 (2021).
132. Owen, D. R. *et al.* An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science* **374**, 1586–1593 (2021).
133. Meneguzzo, F., Ciriminna, R., Zabini, F. & Pagliaro, M. Review of Evidence Available on Hesperidin-Rich Products as Potential Tools against COVID-19 and Hydrodynamic Cavitation-Based Extraction as a Method of Increasing Their Production. *Processes* **8**, 549 (2020).
134. Matondo, A. *et al.* Oleanolic Acid, Ursolic Acid and Apigenin from *Ocimum basilicum* as Potential Inhibitors of the SARS-CoV-2 Main Protease: A Molecular Docking Study. *International Journal of Pathogen Research* 1–16 (2021).

135. Lin, C. *et al.* Study of Baicalin toward COVID-19 Treatment: In silico Target Analysis and in vitro Inhibitory Effects on SARS-CoV-2 Proteases. *Biomed Hub* **6**, 122–137 (2021).
136. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **181**, 475–483 (2020).
137. National Center for Biotechnology Information. "PubChem Bioassay Record for AID 1706, Source: The Scripps Research Institute Molecular Screening Center" PubChem, <https://pubchem.ncbi.nlm.nih.gov/bioassay/1706>. Accessed 25 January, 2022.
138. National Center for Biotechnology Information. "PubChem Bioassay Record for AID 652038, Source: National Center for Advancing Translational Sciences (NCATS)" PubChem, <https://pubchem.ncbi.nlm.nih.gov/bioassay/652038>. Accessed 25 January, 2022.
139. National Center for Biotechnology Information. "PubChem Bioassay Record for AID 485353, qHTS of Yeast-based Assay for SARS-CoV PLP, Source: National Center for Advancing Translational Sciences (NCATS)" PubChem, <https://pubchem.ncbi.nlm.nih.gov/bioassay/485353>. Accessed 25 January, 2022.
140. Touret, F. *et al.* In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. *bioRxiv* 2020.04.03.023846 (2020).
141. Ellinger, B. *et al.* Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large-scale drug repurposing collection. *Research Square* (2020).
142. Fragments screened for 3CL protease binding using crystallography techniques. Data is sourced from the Diamond Light Source group.
143. Kloss, F., Lincke, T. & Hertweck, C.. Highly Efficient Total Synthesis of the Clostridium-Derived anti-MRSA Antibiotic Closthioamide. *Eur. J. of Org. Chem.* 1429–1431 (2011).
144. Albany Molecular Research Inc: Natural Product Libraries, <https://www.amriglobal.com>, accessed 2022.
145. AnalytiCon Discovery: Libraries and Collections, <https://ac-discovery.com>, accessed

- 2022.
146. Bioinformatics Institute (BII): A*STAR Natural Product Library, <http://www.bii.a-star.edu.sg>, accessed 2022.
 147. Fondazione Istituto Insubrico Ricerca per la Vita (FIIRV): Scientific Asset, <http://www.ricercaperlavita.it/en>, accessed 2022.
 148. Fundación MEDINA: Natural Product Libraries, <http://www.medinadiscovery.com>, accessed 2022.
 149. Griffith Institute for Drug Discovery (GRIDD): Nature Bank, <https://www.griffith.edu.au/institute-drug-discovery>, accessed 2022.
 150. InterBioScreen: Natural Compound (NC) Collection, <https://www.ibscreen.com>, accessed 2022.
 151. Magellan BioScience Group Inc: Oceans of Possibilities, <http://www.magellanbioscience.com>, accessed 2022.
 152. Mycosynthetix Inc: The Healing Power of Nature, <http://www.mycosynthetix.com>, accessed 2022.
 153. Natural Products Discovery Institute (NPDI): A Division of The Baruch S. Blumberg Institute, <http://www.npdi-us.org>, accessed 2022.
 154. PharmaMar: Marine Compound Library. <https://www.pharmamar.com>, accessed 2022.
 155. PhytoPharmacon: Natural Product Library, <http://www.phytopharmacon.com>, accessed 2022.
 156. RIKEN: Natural Products Depository (NPDepo), http://www.riken.jp/dmp/english/index_en.html, accessed 2022.
 157. A. Marinetti, The Institut de Chimie des Substances Naturelles (ICSN): Past and Present, *Eur. J. Org. Chem.*, 2018,(42), 5774–5776 CrossRef CAS.
 158. The Scripps Research Institute (TSRI): The Natural Products Library (NPL) at TSRI, <https://www.scripps.edu>, accessed 2018.
 159. The Univeristiy of Mississippi National Center for Natural Products Research,

- <https://pharmacy.olemiss.edu/ncnpr/>, accessed 2019.
160. Unigen: PhytoLogix, <https://unigen.net>, accessed 2018.
161. InterLink Biotechnologies: Natural Products, <http://www.interlinkbiotech.com>, accessed 2018.
162. Bano Mirza S, Bokhari H, Qaiser Fatmi M (2015) Exploring natural products from the biodiversity of Pakistan for computational drug discovery studies: collection, optimization, design and development of a chemical database (ChemDP). <https://www.ingentaconnect.com/content/ben/cad/2015/00000011/00000002/art00003>. Accessed 9 Sept 2019.
163. FooDB. <http://foodb.ca/>. Accessed 3 Oct 2019.
164. omicX. In: omicX [Internet]. <https://omictools.com/>. Accessed 9 Oct 2019.
165. Sorokina M (2020) List of natural products databases. Figshare. <https://doi.org/10.6084/m9.figshare.11926047.v1>.
166. Dictionary of Natural Products 28.1. <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=DB01289ACAA79C222859E1CD8A98A894>. Accessed 9 Oct 2019.
167. Reaxys. <https://www.reaxys.com/#/search/quick>. Accessed 9 Oct 2019.
168. Dictionary of Marine Natural Products 2018. <http://dmnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=824F3121F9A123D4684A7A8289F618E2>. Accessed 9 Oct 2019.
169. Dictionary of Food Compounds 2018. <http://dfc.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml;jsessionid=60BDE6E1AE536A1C52AFB65A680DC289>. Accessed 18 Oct 2019.
170. Johnson SG (2014) NIST Standard Reference Database 1A v17. In: NIST [Internet]. <https://www.nist.gov/srd/nist-standard-reference-database-1a-v17>. Accessed 9 Oct 2019.
171. MarinLit. <http://pubs.rsc.org/marinlit/>. Accessed 9 Oct 2019.

172. AntiBase. <https://application.wiley-vch.de/stmdata/antibase.php>. Accessed 9 Oct 2019.
173. Wiley-VCH—AntiBase. <https://application.wiley-vch.de/stmdata/antibase.php>. Accessed 21 Oct 2019.
174. MassBank of North America (MoNa). <http://mona.fiehnlab.ucdavis.edu/>. Accessed 16 Oct 2019.
175. MassBank | European MassBank (NORMAN MassBank) mass spectral database. <http://massbank.normandata.eu/MassBank/>. Accessed 16 Oct 2019.
176. MassBank | MSSJ MassBank Mass Spectral DataBase. <http://www.massbank.jp/>. Accessed 16 Oct 2019.
177. NMRdata. <http://www.nmrdata.com/>. Accessed 15 Oct 2019.
178. Molecular Diversity Preservation International (MDPI). <https://www.mdpi.org/>. Accessed 15 Oct 2019.
179. ISDB by oolonek. <http://oolonek.github.io/ISDB/>. Accessed 15 Oct 2019.
180. Shen J, Xu X, Cheng F, Liu H, Luo X, Shen J, et al (2003) Virtual screening on natural products for discovering active compounds and target information. <https://doi.org/10.2174/0929867033456729>. Accessed 20 May 2019.
181. Natural Products Atlas. <https://www.npatlas.org/joomla/>. Accessed 16 Oct 2019.
182. Lichen Database. In: MTBLS999: A database of high-resolution MS/MS spectra for lichen metabolites [Internet]. <https://www.ebi.ac.uk/metabolights/MTBLS999>. Accessed 16 Oct 2019.
183. TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531123/>. Accessed 29 Apr 2019.
184. Yanuar A, Mun'im A, Lagho ABA, Syahdi RR, Rahmat M, Suhartanto H (2011) Medicinal plants database and three dimensional structure of the chemical compounds from medicinal plants in Indonesia. ArXiv11117183 Q-Bio. <http://arxiv.org/abs/1111.7183>. Accessed 22 Oct 2019.

185. DrugBank | nutraceutical search.
<https://www.drugbank.ca/drugs?utf8=%E2%9C%93&nutraceutical=1&filter=true>.
Accessed 17 Oct 2019.
186. Novel Antibiotics Database. <http://www.antibiotics.or.jp/journal/database/database-top.htm>. Accessed 18 Oct 2019.
187. Dr.V,Umashankar (2018) InPACdb | Indian-Plant-Anticancer-Compound-DB.
<https://github.com/inpacdb/Indian-Plant-Anticancer-Compound-DB-inpacdb>. Accessed 17 Oct 2019.
188. Compound Sets—NCI DTP Data—National Cancer Institute—Confluence Wiki.
<https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets>. Accessed 18 Oct 2019.
189. OSM—Open Source Malaria. <http://opensourcemalaria.org/>. Accessed 18 Oct 2019.
190. PhytoHub. <http://phytohub.eu/>. Accessed 16 Oct 2019.
191. International Venom and Toxin Database. <http://www.kingsnake.com/toxinology>.
192. Snake Neurotoxin Database. http://sdmc.i2r.a-star.edu.sg/Templar/DB/snake_neurotoxin.
193. MOLLUSK toxin database. <http://research.i2r.a-star.edu.sg/MOLLUSK>.
194. UEFS Natural Products. <http://zinc12.docking.org/catalogs/uefsnp>. Accessed 6 Nov 2019.
195. Journal of Natural Products. <https://pubs.acs.org/journal/jnprdf>.
196. Marine Drugs. <https://www.mdpi.com/journal/marinedrugs>.
197. A database of natural products and chemical entities from marine habitat.
<http://www.bioinformation.net/003/003000032008.htm>. Accessed 6 Nov 2019.
198. Ambinter-Greenpharma natural compound library (GPNCL). In: Greenpharma [Internet].
<https://www.greenpharma.com/products/compound-librairies/>. Accessed 9 Oct 2019.
199. ChemBridge | Screening Library | Diversity Libraries.
https://www.chembridge.com/screening_libraries/diversity_libraries/. Accessed 16 Oct 2019.

200. LOPAC1280. Library of pharmacologically active compounds. In: Sigma-Aldrich [Internet]. <https://www.sigmaaldrich.com/life-science/cell-biology/bioactive-small-molecules/lopac1280-navigator.html>. Accessed 16 Oct 2019.
201. Prestwick Chemical. The Prestwick Phytochemical Library, a collection of natural products. <http://www.prestwickchemical.com/libraries-screening-lib-phyto.html>. Accessed 16 Oct 2019.
202. Targetmol | Natural Compound Library. <https://www.targetmol.com/compound-library/Natural-Compounds-Library>. Accessed 16 Oct 2019.
203. AnalytiCon Discovery, Screening Libraries. In: AnalytiCon Discovery [Internet]. <https://ac-discovery.com/screening-libraries/>. Accessed 16 Oct 2019.
204. InterBioScreen | Natural Compounds. <https://www.ibscreen.com/natural-compounds>. Accessed 9 Oct 2019.
205. INDOFINE Chemical Company. http://www.indofinechemical.com/Media/sdf/sdf_files.aspx. Accessed 16 Oct 2019.
206. Pi Chemicals System. http://www.pipharm.com/catalog_products/list?category=28. Accessed 16 Oct 2019.
207. Specs. Compound management services and research compounds for the life science industry. <https://www.specs.net/index.php>. Accessed 16 Oct 2019.
208. ZINC Specs Natural Products. <http://zinc.docking.org/catalogs/specsnp/>. Accessed 16 Oct 2019.
209. WHO monographs on selected medicinal plants. (World Health Organization, 1999).
210. WHO monographs on selected medicinal plants. (World Health Organization, 2009).
211. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 46, doi:10.1093/nar/gky092 (2018).
212. Ahmed, J. et al. SuperSweet—a resource on natural and artificial sweetening agents. *Nucleic Acids Res* 39, doi:10.1093/nar/gkq917 (2011).
213. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. A systematic comparison of

- the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* 14, doi:10.1186/1471-2105-14-112 (2013).
214. Ashfaq, U. A., Mumtaz, A., ul-Qamar, T. & Fatima, T. MAPS database: medicinal plant activities, phytochemical and structural database. *Bioinformation* 9, doi:10.6026/97320630009993 (2013).
215. Banerjee, P. et al. Super Natural II—a database of natural products. *Nucleic Acids Res* 43, doi:10.1093/nar/gku886 (2015).
216. Berdy, J. & Kertesz, M. in *Chemical information* (ed H. R. Collier) (Berlin Heidelberg, 1989).
217. Blin, K. et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* 45, doi:10.1093/nar/gkx319 (2017).
218. Blunt, J., Munro, M. & Upjohn, M. The role of databases in marine natural products research. *Handb Mar Nat Prod.* 1, doi:10.1007/978-90-481-3834-0_6 (2012).
219. Blunt, J. W. et al. Marine natural products. *Nat Prod Rep* 35, doi:10.1039/C7NP00052A (2018).
220. Boonen, J. et al. Alkamid database: chemistry, occurrence and functionality of plant N-alkylamides. *J Ethnopharmacol* 142, doi:10.1016/j.jep.2012.05.038 (2012).
221. Bultum, L. E., Woyessa, A. M. & Lee, D. ETM-DB: integrated Ethiopian traditional herbal medicine and phytochemicals database. *BMC Complement Altern Med.* 19, doi:10.1186/s12906-019-2634-1 (2019).
222. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 46, doi:10.1093/nar/gkx935 (2018).
223. Chang, K. W. et al. iSMART: an integrated cloud computing web server for traditional Chinese medicine for online virtual screening, de novo evolution and drug design. *J Biomol Struct Dyn* 29, doi:10.1080/073911011010524988 (2011).
224. Chen, C. Y. C. TCM Database@Taiwan: the World's Largest Traditional Chinese Medicine database for drug screening in silico. *PLOS ONE.* 6,

- doi:10.1371/journal.pone.0015939 (2011).
225. Chen, Y., Bruyn, K. o. p. s. C. & Kirchmair, J. Data Resources for the computer-guided discovery of bioactive natural products. *J Chem Inf Model* 57, doi:10.1021/acs.jcim.7b00341 (2017).
226. Choi, H. et al. NPCARE: database of natural products and fractional extracts for cancer regulation. *J Cheminformatics*. 9, doi:10.1186/s13321-016-0188-5 (2017).
227. Crawford, M. J. & Clardy, J. Bacterial symbionts and natural products. *Chem Commun*. 47, doi:10.1039/c1cc11574j (2011).
228. Dagan-Wiener, A. et al. BitterDB: taste ligands and receptors database in 2019. *Nucleic Acids Res* 47, doi:10.1093/nar/gky974 (2019).
229. Davis, G. D. J. & Vasanthi, A. H. R. Seaweed metabolite database (SWMD): a database of natural compounds from marine algae. *Bioinformatics* 5, doi:10.6026/97320630005361 (2011).
230. Derese, S., Oyim, J., Rogo, M. & Ndakala, A. Mitishamba database: a web based in silico database of natural products from Kenya plants. (University of Nairobi, 2015).
231. Dunkel, M. et al. SuperScent—a database of flavors and scents. *Nucleic Acids Res* 37, doi:10.1093/nar/gkn695 (2009).
232. Ehrman, T. M., Barlow, D. J. & Hylands, P. J. In silico search for multi-target anti-inflammatories in Chinese herbs and formulas. *Bioorg Med Chem* 18, doi:10.1016/j.bmc.2010.01.070 (2010).
233. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* 48, doi:10.1021/ci700286x (2008).
234. Fang, X., Shao, L., Zhang, H. & Wang, S. CHMIS-C: a comprehensive herbal medicine information system for cancer. *J Med Chem* 48, doi:10.1021/jm049838d (2005).
235. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* 40, doi:10.1093/nar/gkr1178 (2012).

236. Fishedick, J. T., Johnson, S. R., Ketchum, R. E. B., Croteau, R. B. & Lange, B. M. NMR spectroscopic search module for Spektraris, an online resource for plant natural product identification—Taxane diterpenoids from *Taxus* × media cell suspension cultures as a case study. *Phytochemistry* 113, doi:10.1016/j.phytochem.2014.11.020 (2015).
237. Füllbeck, M., Michalsky, E., Dunkel, M. & Preissner, R. Natural products: sources and databases. *Nat Prod Rep* 23, doi:10.1039/B513504B (2006).
238. Gabrielson, S. W. SciFinder. *J Med Libr Assoc.* 106, doi:10.5195/jmla.2018.515 (2018).
239. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res* 45, doi:10.1093/nar/gkw1074 (2017).
240. Gilson, M. K. et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44, doi:10.1093/nar/gkv1072 (2016).
241. Gu, J. et al. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE* 8, doi:10.1371/journal.pone.0062839 (2013).
242. Guijas, C. et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal Chem* 90, doi:10.1021/acs.analchem.7b04424 (2018).
243. Günthardt, B. F., Hollender, J., Hungerbühler, K., Scherlinger, M. & Bucheli, T. D. Comprehensive toxic plants-phytotoxins database and its application in assessing aquatic micropollution potential. *J Agric Food Chem* 66, doi:10.1021/acs.jafc.8b01639 (2018).
244. Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J. Cheminformatics.* 10, doi:10.1186/s13321-018-0293-8 (2018).
245. 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.* 14, doi:10.1038/nrd4510 (2015).
246. Hastings, J. et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 41, doi:10.1093/nar/gks1146

- (2013).
247. Hatherley, R. et al. SANCDB: a South African natural compound database. *J Cheminformatics* 7, doi:10.1186/s13321-015-0080-8 (2015).
248. Haug, K. et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 41, doi:10.1093/nar/gks1004 (2013).
249. He, Q. Y. et al. ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res* 36, doi:10.1093/nar/gkm832 (2008).
250. Huang, J. et al. CEMTDD: the database for elucidating the relationships among herbs, compounds, targets and related diseases for Chinese ethnic minority traditional drugs. *Oncotarget*. 6, doi:10.18632/oncotarget.3789 (2015).
251. Huang, W. et al. PAMDB: a comprehensive *Pseudomonas aeruginosa* metabolome database. *Nucleic Acids Res* 46, doi:10.1093/nar/gkx1061 (2018).
252. Ibezim, A., Debnath, B., Ntie-Kang, F., Mbah, C. J. & Nwodo, N. J. Binding of anti-Trypanosoma natural products from African flora against selected drug targets: a docking study. *Med Chem Res* 26, doi:10.1007/s00044-016-1764-y (2017).
253. Ikram, N. K. K. et al. A virtual screening approach for identifying plants with anti H5N1 neuraminidase activity. *J Chem Inf Model* 55, doi:10.1021/ci500405g (2015).
254. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res* 47, doi:10.1093/nar/gky1048 (2019).
255. Johnson, S. R. & Lange, B. M. Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front Bioeng Biotechnol.*, doi:10.3389/fbioe.2015.00022 (2015).
256. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, doi:10.1093/nar/gkw1092 (2016).

257. Kang, H. et al. HIM-herbal ingredients in vivo metabolism database. *J Cheminformatics*. 5, doi:10.1186/1758-2946-5-28 (2013).
258. Khalifa, S. A. et al. Marine natural products: a source of novel anticancer drugs. *Mar Drugs* 17, doi:10.3390/md17090491 (2019).
259. Kim, S. K., Nam, S., Jang, H., Kim, A. & Lee, J. J. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complement Altern Med*. 15, doi:10.1186/s12906-015-0758-5 (2015).
260. King, Z. A. et al. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 44, doi:10.1093/nar/gkv1049 (2016).
261. Klementz, D. et al. StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res*. 44, doi:10.1093/nar/gkv1319 (2016).
262. Kuhn, S. & Schlörer, N. E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2— a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn Reson Chem* 53, doi:10.1002/mrc.4263 (2015).
263. Lagunin, A. A. et al. Computer evaluation of hidden potential of phytochemicals of medicinal plants of the traditional Indian ayurvedic medicine. *Biomeditsinskaia Khimiia*. 61, doi:10.18097/PBMC20156102286 (2015).
264. Lang, G. et al. Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. *J Nat Prod* 71, doi:10.1021/np8002222 (2008).
265. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*. 42, doi:10.1093/nar/gkt1068 (2014).
266. Lei, J. & Zhou, J. A marine natural product database. *J Chem Inf Comput Sci* 42, doi:10.1021/ci010111x (2002).
267. Li, B. et al. YaTCM: yet another traditional Chinese medicine database for drug discovery. *Comput Struct Biotechnol J*. 16, doi:10.1016/j.csbj.2018.11.002 (2018).
268. Lim, E. et al. T3DB: a comprehensively annotated database of common toxins and their

- targets. *Nucleic Acids Res* 38, doi:10.1093/nar/gkp934 (2010).
269. López-Pérez, J. L., Therón, R., Olmo, E. & Díaz, D. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics* 23, doi:10.1093/bioinformatics/btm516 (2007).
270. Loub, W. D., Farnsworth, N. R., Soejarto, D. D. & Quinn, M. L. NAPRALERT: computer handling of natural product research data. *J Chem Inf Model* 25, doi:10.1021/ci00046a009 (1985).
271. Maeda, M. H. & Kondo, K. Three-Dimensional Structure Database of Natural Metabolites (3DMET): a novel database of curated 3D structures. *J Chem Inf Model* 53, doi:10.1021/ci300309k (2013).
272. Mahesh, S. K., Fathima, J. & Veena, V. G. in *Natural Bio-active compounds: volume 2: chemistry, pharmacology and health care practices* (eds M. K. Swamy & M. S. Akhtar) (Springer Singapore, 2019).
273. Mangal, M., Sagar, P., Singh, H., Raghava, G. P. S. & Agarwal, S. M. NPACT: naturally occurring plant-based anti-cancer compound-activity-target database. *Nucleic Acids Res* 41, doi:10.1093/nar/gks1047 (2013).
274. Meetei, P. A. et al. NeMedPlant: a database of therapeutic applications and chemical constituents of medicinal plants from north-east region of India. *Bioinformation*. 8, doi:10.6026/97320630008209 (2012).
275. Miettinen, K. et al. The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis. *Nucleic Acids Res* 46, doi:10.1093/nar/gkx925 (2018).
276. Mohanraj, K. et al. IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics. *Sci Rep*, doi:10.1038/s41598-018-22631-z (2018).
277. Nakamura, K. et al. KNAPSAck-3D: a three-dimensional structure database of plant metabolites. *Plant Cell Physiol* 54, doi:10.1093/pcp/pcs186 (2013).
278. Naveja, J. J., Rico-Hidalgo, M. P. & Medina-Franco, J. L. Analysis of a large food chemical database: chemical space, diversity, and complexity. *F1000Research*,

- doi:10.12688/f1000research.15440.2 (2018).
279. Neveu, V. et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res* 45, doi:10.1093/nar/gkw980 (2017).
280. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79, doi:10.1021/acs.jnatprod.5b01055 (2016).
281. Nguyen-Vo, T. H. et al. VIETHERB: a database for Vietnamese herbal species. *J Chem Inf Model* 59, doi:10.1021/acs.jcim.8b00399 (2019).
282. Ntie-Kang, F. et al. ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Adv.* 4, doi:10.1039/c3ra43754j (2014).
283. Ntie-Kang, F. et al. CamMedNP: building the Cameroonian 3D structural natural products database for virtual screening. *BMC Complement Altern Med.* 13, doi:10.1186/1472-6882-13-88 (2013).
284. Ntie-Kang, F. et al. Molecular modeling of potential anticancer agents from African medicinal plants. *J Chem Inf Model* 54, doi:10.1021/ci5003697 (2014).
285. Ntie-Kang, F. et al. Virtualizing the p-ANAPL library: a step towards drug discovery from African medicinal plants. *PLoS ONE* 9, doi:10.1371/journal.pone.0090655 (2014).
286. Ntie-Kang, F. et al. NANPDB: a resource for natural products from Northern African sources. *J Nat Prod.* 80, doi:10.1021/acs.jnatprod.7b00283 (2017).
287. Ntie-Kang, F. et al. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS ONE* 8, doi:10.1371/journal.pone.0078085 (2013).
288. Nupur, L. N. U. et al. ProCarDB: a database of bacterial carotenoids. *BMC Microbiol* 16, doi:10.1186/s12866-016-0715-6 (2016).
289. Onguéné, P. A. et al. The potential of anti-malarial compounds derived from African medicinal plants, part III: an in silico evaluation of drug metabolism and pharmacokinetics profiling. *Org Med Chem Lett.* 4, doi:10.1186/s13588-014-0006-x (2014).

290. Otasek, D., Morris, J. H., Bouças, J., Pico, A. R. & Demchak, B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol* 20, doi:10.1186/s13059-019-1758-4 (2019).
291. Palhares, R. M. et al. Medicinal plants recommended by the World Health Organization: DNA barcode identification associated with chemical analyses guarantees their quality. *PLoS ONE*, doi:10.1371/journal.pone.0127866 (2015).
292. Pathania, S., Ramakrishnan, S. M. & Bagler, G. Phytochemica: a platform to explore phytochemicals of medicinal plants. *Database*, doi:10.1093/database/bav075 (2015).
293. Pence, H. E. & Williams, A. ChemSpider: an online chemical information resource. *J Chem Educ* 87, doi:10.1021/ed100697w (2010).
294. Pereira, F. & Aires-de-Sousa, J. Computational methodologies in the exploration of marine natural product leads. *Mar Drugs* 16, doi:10.3390/md16070236 (2018).
295. Pilon, A. C. et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep.* 7, doi:10.1038/s41598-017-07451-x (2017).
296. Pilon-Jiménez, B. A., Saldívar-González, F. I., Díaz-Eufracio, B. I. & Medina-Franco, J. L. BIOFACQUIM: a Mexican compound database of natural products. *Biomolecules*. 9, doi:10.3390/biom9010031 (2019).
297. Polur, H., Joshi, T., Workman, C. T., Lavekar, G. & Kouskoumvekaki, I. Back to the roots: prediction of biologically active natural products from ayurveda traditional medicine. *Mol Inform.* 30, doi:10.1002/minf.201000163 (2011).
298. Potshangbam, A. M. et al. MedPServer: a database for identification of therapeutic targets and novel leads pertaining to natural products. *Chem Biol Drug Des* 93, doi:10.1111/cbdd.13430 (2019).
299. Qiao, X., Hou, T., Zhang, W., Guo, S. & Xu, X. A 3D structure database of components from Chinese traditional medicinal herbs. *J Chem Inf Comput Sci* 42, doi:10.1021/ci010113h (2002).

300. Quinn, R. J. et al. Developing a drug-like natural product library. *J Nat Prod* 71, doi:10.1021/np070526y (2008).
301. Ramirez-Gaona, M. et al. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.* 45, doi:10.1093/nar/gkw1058 (2017).
302. Rodriguez-Amaya, D. B., Kimura, M., Godoy, H. T. & Amaya-Farfan, J. Updated Brazilian database on food carotenoids: factors affecting carotenoid composition. *J Food Compos Anal* 21, doi:10.1016/j.jfca.2008.04.001 (2008).
303. Rothwell, J. A. et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*, doi:10.1093/database/bat070 (2013).
304. Ru, J. et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J Cheminformatics.* 6, doi:10.1186/1758-2946-6-13 (2014).
305. Sagar, S., Kaur, M., Radovanovic, A. & Bajic, V. B. Dragon exploration system on marine sponge compounds interactions. *J Cheminformatics* 5, doi:10.1186/1758-2946-5-11 (2013).
306. Sarethy, I. P., Srivastava, N. & Pan, S. in *Natural Bio-active compounds: volume 1: production and applications* (eds M. S. Akhtar, M. K. Swamy, & U. R. Sinniah) (Springer, 2019).
307. Seiler, K. P., Kuehn, H., Happ, M. P., DeCaprio, D. & Clemons, P. A. Using ChemBank to probe chemical biology. *Curr Protoc Bioinforma.* 22, doi:10.1002/0471250953.bi1405s22 (2008).
308. Sharma, A. et al. BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. *J Cheminformatics.* 6, doi:10.1186/s13321-014-0046-2 (2014).
309. Sitzmann, M., Filippov, I. V. & Nicklaus, M. C. Internet resources integrating many small-molecule databases1. *SAR QSAR Environ Res* 19, doi:10.1080/10629360701843540 (2008).

310. Skinnider, M. A. et al. Genomes to natural products Prediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res* 43, doi:10.1093/nar/gkv1012 (2015).
311. Sorokina, M. & Steinbeck, C. NaPLoS: a natural products likeness scorer—web application and database. *J Cheminformatics*. 11, doi:10.1186/s13321-019-0378-z (2019).
312. Sparks, T. C., Wessels, F. J., Lorsbach, B. A., Nugent, B. M. & Watson, G. B. The new age of insecticide discovery—the crop protection industry and the impact of natural products. *Pestic Biochem Physiol.*, doi:10.1016/j.pestbp.2019.09.002 (2019).
313. Srinivasan, K. N. et al. SCORPION, a molecular database of scorpion toxins. *Toxicon* 40, doi:10.1016/S0041-0101(01)00182-9 (2002).
314. Sterling, T. & Irwin, J. J. ZINC 15—ligand discovery for everyone. *J Chem Inf Model* 55, doi:10.1021/acs.jcim.5b00559 (2015).
315. Tawfike, A. F., Viegelmann, C. & Edrada-Ebel, R. in *Metabolomics tools for natural product discovery: methods and protocols* (eds U. Roessner & D. A. Dias) (Humana Press, 2013).
316. Tomasulo, P. ChemIDplus—super source for chemical and drug information. *Med Ref Serv Q*. 21, doi:10.1300/J115v21n01_04 (2002).
317. Tomiki, T. et al. RIKEN natural products encyclopedia (RIKEN NPEDIA), a chemical database of RIKEN natural products depository (RIKEN NPDEPO). *J Comput Aid Chem*. 7, doi:10.2751/jcac.7.157 (2006).
318. Tung, C. W. et al. TIPdb-3D: the three-dimensional structure database of phytochemicals from Taiwan indigenous plants. *Database.*, doi:10.1093/database/bau055 (2014).
319. Vetrivel, U., Subramanian, N. & Pilla, K. InPACdb—Indian plant anticancer compounds database. *Bioinformatics* 4, doi:10.6026/97320630004071 (2009).
320. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global

- Natural Products Social Molecular Networking. *Nat Biotechnol.* 34, doi:10.1038/nbt.3597 (2016).
321. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 3, doi:10.1038/sdata.2016.18 (2016).
322. Williams, A. J., Martin, G. E. & Rovnyak, D. Modern NMR approaches to the structure elucidation of natural products: volume 1: instrumentation and software. (Royal Society of Chemistry, 2016).
323. Williamson, A. E. et al. Open Source Drug Discovery: highly potent antimalarial compounds derived from the Tres Cantos Arylpyrroles. *ACS Cent Sci.* 2, doi:10.1021/acscentsci.6b00086 (2016).
324. Willighagen, E. L. et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminformatics* 9, doi:10.1186/s13321-017-0220-4 (2017).
325. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, doi:10.1093/nar/gkx1037 (2018).
326. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 46, doi:10.1093/nar/gkx1089 (2018).
327. Xu, H. Y. et al. ETCM: an encyclopaedia of traditional Chinese medicine. *Nucleic Acids Res* 47, doi:10.1093/nar/gky987 (2019).
328. Xu, J. & Yang, Y. Traditional Chinese medicine in the Chinese health care system. *Health Policy* 90, doi:10.1016/j.healthpol.2008.09.003 (2009).
329. Yabuzaki, J. Carotenoids Database: structures, chemical fingerprints and distribution among organisms. *Database J Biol Databases Curation.*, doi:10.1093/database/bax004 (2017).
330. Ye, H. et al. HIT: linking herbal active ingredients to targets. *Nucleic Acids Res* 39, doi:10.1093/nar/gkq1165 (2011).
331. Yongye, A. B., Waddell, J. & Medina-Franco, J. L. Molecular scaffold analysis of natural

- products databases in the public domain. *Chem Biol Drug Des* 80, doi:10.1111/cbdd.12011 (2012).
332. Yuan, H., Ma, Q., Ye, L. & Piao, G. The traditional medicine and modern medicine from natural products. *Molecules* 21, doi:10.3390/molecules21050559 (2016).
333. Yue, Y. et al. TMDB: a literature-curated database for small molecular compounds found from tea. *BMC Plant Biol* 14, doi:10.1186/s12870-014-0243-1 (2014).
334. Zeng, X. et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res* 46, doi:10.1093/nar/gkx1026 (2018).
335. Zeng, X. et al. CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res* 47, doi:10.1093/nar/gky965 (2019).
336. Zhang, R., Lin, J., Zou, Y., Zhang, X. J. & Xiao, W. L. Chemical space and biological target network of anti-inflammatory natural products. *J Chem Inf Model* 59, doi:10.1021/acs.jcim.8b00560 (2019).
337. Shoemaker, R. H., The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer* 2006, 6 (10), 813-823.
338. Harvey, A. & Cree, I.. High-Throughput Screening of Natural Products for Cancer Therapy. *Planta Medica* 76, 1080–1086 (2010).
339. Chemical Genomics: A Systematic Approach in Biological Research and Drug Discovery. *Current Issues in Molecular Biology* (2002). doi:10.21775/cimb.004.033
340. Schirle, M., Bantscheff, M. & Kuster, B.. Mass Spectrometry-Based Proteomics in Preclinical Drug Discovery. *Chemistry & Biology* 19, 72–84 (2012).
341. Wishart, D. S.. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery* 15, 473–484 (2016).