UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Reaching for the High Hanging Fruit: Working Towards Better Drug Discovery

Permalink

https://escholarship.org/uc/item/85d755jc

Author

Pye, Cameron

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Cruz

Reaching for the High Hanging Fruit: Working Towards Better Drug Discovery

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

DOCTOR OF PHILOSOPHY

in

Chemistry & Biochemistry

by

Cameron R. Pye

March 2017

The Dissertation of Cameron R. Pye
is approved:

Professor Seth Rubin, chair

Professor R. Scott Lokey

Assistant Professor Jevgenij Raskatov

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

Introduction 1

Bibliography 7

1 A Strategy for Direct Chemical Activation of the Retinoblastoma Protein 8

1.1 Abstract 8

1.2 Introduction 8

1.3 Results 9

1.3.1 FP Rb/E2F^{TMR} High Throughput Screen 11

1.3.2 Sanford Burnham Institute Screening Campaign 16

1.4 Discussion 23

1.5 Methods 25

1.5.1 Protein and peptide reagents 25

1.5.2 Isothermal Titration Calorimetry 27

1.5.3 Fluorescence Polarization Assay and Screen 27

1.6 Bibliography 28

2 Non-Classical Size Dependence of Permeation Defines Bounds for Passive Adsorption of Large Drug Molecules. 31

2.1 Foreword 31

2.2 Abstract 32

2.3 Introduction 32

2.4 Results 35

2.4.1 Aggregation Imposes a Limit on Apparent Permeability 35

2.4.2 Lipophilicity Scanning Library 38

2.4.3 Solubility Adjustment via Filtration 42

2.4.4 Non-Classical Diffusion of bRo5 Compounds 45

2.4.5 Extension to other scaffolds 46

2.5 Discussion and conclusions 47

2.6 Experimental 51

2.6.1 General 51

2.6.2 Synthesis of Pure Cyclic Permethylated Peptides 52

2.6.3 Synthesis of Cyclic Peptide Libraries 54

2.6.4 Synthesis of natural products and unrelated pure compounds 56

2.6.5 General Analytical Procedures 64

2.7 Characterization of Compounds 70

2.8 LC/MS Characterization of Selected Pure Compounds 70

2.8.2 Extracted Ion Chromatograms for Lipophillicity Scanning Mixtures 92

2.8.3 Characterization of Natural Products and Analogs 99

2.9 Bibliography 109

3 Natural Products: Are We Close to the End? 112

3.1 Foreword 112

3.2 Abstract 112

3.3 Introduction 113

3.4 Results and Discussion 114

3.4.1 Exploring Trends in Chemical Diversity 114

3.4.2 Source Diversity vs. Structural Diversity 120

3.4.3 Evaluating the Chemical Space Occupied by Natural Products 124

3.4.4 Limitations of this Analysis/ Points to Consider 131

3.5 Future Perspective 132

3.6 Experimental Methods 133

3.6.1 Dataset Creation and Curation 133

3.6.2 Tanimoto and Tversky Scoring 134

3.7 Bibliography 142

4 Future Direction 146

4.1 Chapter 1 146

4.2 Chapter 2 146

4.3 Chapter 3 148

5 Concluding Remarks 150

Figure 1-1 Fluorescence polarization screen for enhancers of Rb-E2F binding. (a) Sample data from the primary screen. FP ratio is plotted for each well in a 384-well plate. The wells contain phosphorylated RbNP and compounds (grey diamonds) or DMSO (blue squares, “negative control”), unphosphorylated RbNP (green triangles, “positive control”), or free E2FTMR alone (purple crosses). The boxed red diamonds are hits that increase the FP ratio of E2FTMR in the presence of phosphorylated RbNP. (b) Follow-up assay in which the effect of the compounds on E2FTMR FP ratio was determined in the absence (purple bars) and presence (grey bars) of phosphorylated RbNP. The phosphorylated RbNP negative (P, blue bar) and unphosphorylated RbNP (U, green bar) positive controls are shown at left. Hits were validated (red asterisks) if they yielded low FP ratios similar to controls in the RbNP target-minus (dashed purple line) assay and high FP ratios in the RbNP target-positive assay (dashed green line). 10

Figure 1-2 FP ratio for each well in the high throughput screen of the ChemDiv library. The wells contain phosphorylated Rb^{NP} and compounds (grey circles) or DMSO (blue circles, “negative control”), unphosphorylated Rb^{NP} (green circles, “positive control”. The red circles are hits that increase the FP ratio of E2F^{TMR} in the presence of phosphorylated Rb^{NP} and satisfy our total fluorescence criterion. 12

Figure 1-3 Effect of validated primary screen hits in a complete Rb^{NP} protein titration. FP ratios were measured and K_d values for E2F^{TMR}-Rb^{NP}binding were determined as in Figure 3a. 16

Figure 1-4 3D scatter plot of the Sanford-Burnham screening results. Red points represent the positive controls (un-phosphorylated Rb), green points represent negative controls (E2F^{TMR} alone), and blue points represent the assay compounds. 17

Figure 1-5 Structure-based strategy for activation of Rb. (a) Structure of phosphorylated Rb (from PDB code: 4ELJ). Docking between the Rb N-terminal domain (RbN, yellow) and the pocket domain (brown) occurs across two interfaces. Interface 1 is mediated by a pocket helix that is nucleated by T373 phosphorylation. Interface 2 is near the LxCxE-binding cleft in the pocket. The E7 peptide (cyan backbone), which is shown bound at its site in the unphosphorylated pocket (from PDB code: 4GUX), clashes with RbN residues at the interface. (b) Phosphorylation of sites in the Rb interdomain linker induces a conformational change that allosterically inhibits E2FTD binding. We find that an LxCxE peptide acts as an activator by binding Rb and inhibiting the RbN-pocket interdomain association. 18

Figure 1-6 . LxCxE peptide acts as an Rb activator. (a) Titration of unphosphorylated and phosphorylated RbNP into E2FTMR. In the presence of the E7 LxCxE peptide and full-length E7 protein (pink), the affinity is increased. (b) EC50 measurement of LxCxE peptide and E7 protein activity. Compound #478337 from Fera et al.9 (grey) and LxCxE variant peptides from Cyclin D and LIN52 do not show activity... 19

Figure 1-7 The LxCxE peptide and E7 protein activators do not affect E2F binding to an Rb mutant in which interdomain docking is disrupted. Binding measurements are shown of E2F^{TMR} to an Rb^{NP} construct containing Q736A and K740A mutations (Rb$^{\otimes I2}$). As described previously[16], E2F binding to the unphosphorylated (K$_d$=7.5 ± 0.3 nM) and phosphorylated (K$_d$=12 ± 2 nM) mutant are similar. Addition of 2 ∝M LxCxE peptide (K$_d$=10 ± 1 nM) or E7 protein (K$_d$=10.5 ± 0.3 nM) does not increase E2F affinity, indicating that the activating effect is not independent of interdomain docking. We confirmed using ITC that the LxCxE peptide still binds Rb$^{\otimes I2}$ with K$_d$=180 ± 20 nM. 21

Figure 1-8 The E7 LxCxE peptide increases affinity of E2FTD for phosphorylated Rb. Representative ITC curves and average K$_d$ measurements are shown for E2F1TD titration into unphosphorylated (a) and phosphorylated (b) RbNP. 22

Figure 1-9 Electrospray ionization mass spectra of intact samples of unphosphorylated and phosphorylated Rb^{NP}. Samples were prepared by desalting. Each spectrum shows a single, dominant peak at the indicated molecular weight. The ~160 Da increase in the phosphorylated sample corresponds to addition of two phosphates. 26

Figure 2-1 . Octapeptide Split Pool Library (A) Split-pool synthetic scheme for per-methylated Ala scan library. All library members contained a Tyr and Pro residue and the remaining 6 positions were either Ala, Leu, or Ile. (B) PAMPA permeability of Ala scan library for each unique composition. Di-Ala substitution was the dominantly permeable species in Leu and Ile containing library sub-pools. 36

Figure 2-2 Lipophilicity Scanning Library 38

Figure 2-3 Variable Ring Size Lipophilicity Scanning Library. All plots include octa- (blue), nona- (orange), and decapeptides (grey). (A) logPapp vs SFlog$K_{hc/w}$ should be a linear relationship (eq 1), we observe deviation from linearity in each system at SFlog$K_{hc/w}$ > 1. (B) Experimental 1,9-decadiene partition coefficients (SFlog$K_{hc/w}$) vs. AlogP. The Linear regression (broken red line) was used to extrapolate values outside the detection limits. (C) logP_{app} of PAMPA vs MDCK-LE for individual compounds. Linear shape of the data suggests that the PAMPA assay is correlated with the more biologically relevant cell based assay. The trends of individual compounds were in agreement with those observed in the mixtures (Table 2-1). 40

Figure 2-4 Solubility Adjusted Intrinsic Permeabilities Filtration solubility (unbroken red lines) and logP_{app} for the (a) octa-, (b) nona- , (c) decapeptides (Figure 2-2) (blue markers). Solubility-adjusted Po values were obtained by dividing the P_{app} by the relative solubility in MeOH for each point (orange markers). The linearity of Po and $K_{hc/w}$ is indicated by a linear fit of the solubility-adjusted values (red broken lines). 44

Figure 2-5 Membrane diffusion vs molecular Volume Diffusion was determined by eq 3 using experimental values of P_o and $K_{hc/w}$. The volume calculated from a 2D structure using a parameterized method. Small, Ro5 molecules (green points) roughly follow an exponential trend while the bRo5, lipophilicity scanning library members (Figure 2-2) (blue points) are subject to a much steeper size penalty. A selection of unrelated synthetic and natural products were subjected to the same analysis to ensure the trends are general (orange points). 47

Figure 2-6 Permeable Space A crude approximation of the minimum log$K_{hc/w}$ required to maintain a permbeability of 10^{-6}cm/s (black line), using D_{mem} values from the Einstein-Stokes for Ro5 compounds (Vol < 850 Å^3) and a 3rd order polynomial fit of the relationship between Log$K_{hc/w}$ and molecular volume for bRo5 compounds (Vol > 850 Å^3). High logK$_{hc/w}$ will likely result in poor aqueous solubility limiting the Papp. This in turn applies an upper limit to the size of a molecule for passive membrane permeability (green shaded region) since the necessary logK$_{hc/w}$ required to overcome the size penalty would render the compound insoluble. 50

Figure 2-7 logSol vs Log K$_{hc/w}$ 67

Figure 2-8 log(f$_{sol}$)Figure 2-8 vs SF Log K$_{hc/w}$ 68

Figure 2-9 P$_{o\ ABL}$ vs elogK$_{hc/w}$ for the permethylated lipophillicity scanning library. 69

Figure 3-1 Examining Structural Diversity. A) Number of compounds published per year and rate of novel compound isolation as a percentage of total natural product isolation. B) Median maximum Tanimoto scores as a function of time. Median average deviation shown as shaded blue region. C) Absolute number of low similarity compounds (T < 0.4) per year. 117

Figure 3-2 Comparison of the time dependent tanimoto scoring vs a randomly shuffled version of dataset (broken red line). 119

Figure 3-3. Examining Source Diversity. A) Plot of median maximum Tanimoto score by year for the full dataset (blue) and the intra-sub-group values for the cyanobacterial sub-group (red) B) Plot of intra-sub-group median maximum Tanimoto scores by year for bacterial sub-groups C) Plot of intra-sub-group median maximum Tanimoto scores by year for marine sub-groups D) Plot of extra-sub-group median maximum Tanimoto scores by year for marine sub-groups E) Violin plots for intra-sub-group median maximum Tanimoto scores for bacterial and marine sub-groups F) Violin plots for extra-sub-group. 122

Figure 3-4 Theoretical vs. Actual Structural Diversity. A) Examples of the four major classes of cyclic tetrapeptides found in nature B) Violin plots indicating the distribution of Tanimoto scores between all members of 65 randomly selected cyclic peptides (10 trials, lanes 1 - 10) and between all 65 cyclic tetrapeptides from our natural product dataset (lane REAL). 126

Figure 3-5 Cluster Analysis for Natural Product Diversity. A) Network diagram displaying all molecules as clusters based on Tversky structural similarities. Compounds with no structural similarity partners appear as singletons in the bottom region of the figure. B) Expansion of region of network diagram indicating erythromycin compound class. C) Example structure from erythromycin cluster. 128

Figure 3-6 Examples of natural products with low (<0.4) Tanimoto scores, indicating compound name, source, year of discovery and isolation method. 130

Figure 3-7 137

Figure 3-8 Cyclic Tetrapeptide Structures from Natural Product Dataset 137

Figure 4-1 Similarity network of all cyclic peptide natural products in the Anti-Marin database. Each edge represents a Dice similarity score >0.8 and each node is a molecule. Nodes are colored based on their calculated octanol/water partition coefficients (AlogP). 148

Table 1-1 Structures of validated hits in the primary screen. 14

Table 2-1 Pure compound composition and assay data. Red residue names indicate D stereochemistry; black residue names indicate L stereochemistry. Methods for PAMPA and MDCKII-LE are described below. AlogP is the octanol/water partition coefficient as calculated by the Ghose and Crippen atomic contribution model. [17,25] 42

For Trillium:

Thanks for putting up with my erratic schedule, our cabin existence,

and even pretending to enjoy our lab karaoke nights.

I couldn’t have done it without you.

On to our next chapter!

And for Dougie:

Whose bright light still shines in my darkest of times.

Miss you bud.

Introduction

When I first began my graduate career at UC Santa Cruz, I assumed that I would be continuing along the path of synthetic organic research that I began in my undergraduate research. During my coursework and cumulative exams encountered in my first year, I found I had a passion for chemical biology and ultimately joined Scott Lokey's lab. Over the course of my PhD research, I have had the opportunity to work with collaborators in both academia and the pharmaceutical industry. The following thesis has three distinct chapters, which were projects that I worked on with three different PI's: Seth Rubin, Scott Lokey (my PI), and Roger Linnington. I have had the great fortune of being able to work with these three brilliant faculty members and a PI that was willing to tolerate my time being shared amongst the other two. Each chapter represent the summary of separate research enterprises, though all under the umbrella of modern drug discovery.

Traditional drug discovery has been successful at targeting a variety of molecular targets that mostly encompass g-protein coupled receptors (GPCRs) and a variety of enzymes.[1] However, it has been hypothesized that the majority of the "druggable genome" exists outside of these two broad classes and are often regulated via protein-protein interactions (PPIs).[1] These PPIs are large interfaces that are more diffuse than the binding pockets of receptors or active sites of enzymes and require large ligands to modulate the interactions.[2] However, an empirical survey of drugs and found that the majority of successful, orally bioavailable molecules shared a set of simple properties.[3] This property set has become cannon for drug-likeness known as Lipinski's Rule of 5. Drugging the portion of the genome that demands larger ligands has been the focus of my research; venturing beyond the Rule of 5 (bRo5). The following chapters include attempts to discover a compound for a non-cannnonical target, an effort to expand the rule set beyond the Rule of 5, and a survey of nature's solution to drug discovery.

Chapter 1 encompasses my work with Seth Rubin and focuses on a high throughput screening campaign to identify molecules that stabilize the Rb-E2F interaction. The Retinoblastoma protein (Rb) regulates the cell cycle by binding the transcription factor E2F, effectively sequestering it in the nucleus and keeping the cell in the G1 phase. This repressive function in part explains why Rb is a potent tumor suppressor protein. Rb is mutated or constitutively inactivated in almost all human carcinomas.[4,5] This make Rb an attractive drug target for a wide variety of cancers and would require a molecule that increases or up regulates Rb's function in the cell.

Rb is phosphorylated by the Cyclin-dependent kinases (Cdks) during cell cycle progression. Rb phosphorylation results in its release of E2F, which in turn activates gene expression for the transition from G1 to S phase. Jason Burke from the Rubin lab had previously shown the mechanism by which phosphorylation induces Rb-E2F dissocistion. Threonine 373 in Rb is phosphorylated and nucleates a conformational change, decreasing Rb's affinity for E2F.[6] The therapeutic strategy for directly targeting Rb is to stabilize the phosphorylated Rb-E2F (pRb-E2F) construct thus preventing cell cycle progression. The feasibility of this proposed mechanism of activation is bolstered by a viral peptide fragment, which is shown to bind to the pRb-E2F complex and stabilize the interaction. This innovative therapeutic strategy is orthogonal to direct inhibition of the cyclin-CDK's responsible for the phosphorylation of Rb, which has proved unsuccessful due to off target effects and potential for rescue by other highly homologous kinases. [7]

In order to find a molecule that accomplished the desired stabilization for the Rb-E2F complex, we designed an assay that would be amenable to high throughput screening. We chose the fluorescence polarization (FP) assay because of its simplicity, rapid read out, and minimal use of reagents, which allows for easy scaling. After developing the assay and establishing the viral peptide fragment acted on the system in the expected manner, we screened an in-house collection of ~25k diverse druglike molecules. Hit compounds discovered in the primary screen were cherry picked and selected for confirmation in dilution series and secondary assays. Unfortunately, the most promising group of molecules shared a common toxoflavin core which is known to undergo redox cycling with common reducing agents such as DTT employed in our assay.[8] Despite these negative results, our assay proved robust. We screened a larger collection of compounds (~250k) at the Sanford-Berham institute. This screening campaign gave rise to a handful of hit molecules but those that were commercially available did not repeat their activity in our hands.

The failure to find useful compounds that act on pRB in the desired manner suggest that these challenging drug targets require a new modality of compounds and conventional screening collections don’t offer the necessary chemistries to access these molecular targets. The viral peptide that elicits the desired effect has poor drug like properties as a linear peptide with charged residues and as such makes for a poor drug candidate. Peptidic macrocyclic compounds, such as those found in natural products are one alternative modality that is being investigated for targeting Rb. These cyclic systems can have drug like properties despite their larger size and is the focus of research of the Lokey Lab; the ongoing collaboration between the Lokey and Rubin labs aims to leverage this therapeutic class towards finding a successful pRb-E2F stabilizer.

Chapter 2 makes up the bulk of my thesis research and represents my research with the one and only Scott Lokey. The Lokey lab’s research aims to understand how peptidic macrocycles can maintain permeability despite their size. Macrocycles, as alluded to earlier, are a class of compounds that are frequently found in natural products.[9] Some of these compounds, such as the FDA approved drug cyclosporine A (CSA) maintains oral bioavailability and cell permeability despite it being over twice the molecular weight (1.2 kDa) of what is normally considered the cutoff (500 Da) for passive permeability. The Lokey lab has been working on understanding how these molecules break the rules and maintain the drug like pharmacokinetic (PK) properties in an effort to apply them to synthetic systems for drug design. CSA contains a variety of non-natural amino acids. It has been shown that it is able to adopt a highly lipophilic low dielectric conformation as well as several hydrophilic conformations including one that binds to its molecular targets.[10] It has been postulated that this chameleonic behavior is what allows CSA to passively permeate cells despite its size and that the non-natural residues facilitate this dynamism by allowing an intramolecular network of hydrogen bonds to effectively hide the polar atoms in the backbone amides. Previous work in the Lokey lab has demonstrated that entirely synthetic hexapeptide scaffolds containing a mixture of d & l as well as N-methylated hydrophobic residues can maintain cell permeabilities and even oral bioavailability on par with CSA suggesting that there is a path to understanding and employing the chemical properties that give rise to CSA’s exceptional behavior.[11]

Current theory of passive cell permeability predicts a positive correlation between lipophilicity and a negative correlation between permeability and molecular weight.[12] Though these models do a good job of explaining the permeabilities of small molecules (MW<500 Da), they fail to recapitulate the size penalty observed as molecular weight approaches 1 kDa.[13] To better understand this relationship, a test system was designed to probe the relationship of lipophilicity, size, and passive permeability. A previous member of our lab, William Hewitt, demonstrated that a split pool library of compounds can be assayed for permeability in complex mixtures and can simultaneously be quantified and deconvoluted by liquid chromatography - mass spectroscopy (LC-MS).[14] By creating a library of compounds whose conformation could not contribute to their lipophilicity in three different ring sizes we could study the size dependence of permeability in the space well above the 500 Da range.

By assessing these compounds' experimental lipophilicities and solubilities to compare the intrinsic permeabilities vs size, we observed a steep size dependence that is not explained by the current passive permeability models. To confirm that the size dependence was not an artifact of the system studied, a series of natural product macrocycles as well as a hand full of commercially available drug compounds were assayed in an identical manner and they displayed a similar trend suggesting the results are general beyond our model compounds. Though the mechanism by which this size dependence is not yet understood, helping to define design rules in this larger chemical space is will allow for the more rational synthesis effort and potentially more useful drug compounds.

Chapter 3 encompasses my work with Roger Linnington. In an effort to cheminformatically assess the chemical space of macrocylces as synthesized by nature I liberated the chemical structures of the compounds in a chemical database that Roger had purchased that contained all known microbial and marine natural products from 1941 -2012. As payment for use of the database Roger asked me for "a few hours of my time" to investigate the novelty of natural products as a function of time. What started as a simple survey blossomed into an involved and illuminating analysis of natural product discovery over time. To add to the thoroughness of the study, the authors of the resulting manuscript manually curated the literature of 90+ journals from 2012 to present day and extracted the microbial and marine natural products as well as any source organism information about each molecule annotated.

The principal methodology of the study made use of molecular similarity scoring by which one is able to computationally quantitate the similarity between any given two molecules. We asked what the highest similarity of a given molecule to any from the years prior as a way to assess the rate of novel chemical discovery in the natural products field. Rediscovery of known compounds or very subtle derivatives thereof is one of the chief problems in modern natural products discovery efforts. By looking at the database of compounds through the lens of this time dependent novelty metric, we could ask a myriad of other questions that asses the given novelty or uniqueness of various source organisms, coverage of compound classes, or discovery strategies that are enriched for unique compounds.

Though unique in their focus, each chapter explores an aspect modern drug discovery. High throughput screening is a staple of modern drug discovery, and though its application in the Rb screening campaign bore no fruit, it validates the chemical strategy for direct pRb activation and paves the way for future work with better collections of molecules. The work towards understanding the role of size and lipophilicity in passive membrane diffusion helps to better define what the boundaries are for passive, oral drugs. The retrospective analysis of natural products allows us to understand what chemistries have proven successful for nature's drug discovery efforts and where new avenues for diverse chemical matter may come from in the future. All together I hope this work will allow for continued innovation and understanding in the chemical biology of drug discovery.

Bibliography

(1) Hopkins, A. L.; Groom, C. R. *Nat. Rev. Drug Discov.* **2002**, *1* (9), 727–730.

(2) Villar, E. A.; Beglov, D.; Chennamadhavuni, S.; Porco, J. A.; Kozakov, D.; Vajda, S.; Whitty, A. *Nat. Chem. Biol.* **2014**, *10* (9), 723–731.

(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23* (1), 3–25.

(4) Classon, M.; Harlow, E. *Nat Rev Cancer* **2002**, *2* (12), 910–917.

(5) Knudsen, E. S.; Wang, J. Y. *Mol Cell Biol* **1997**, *17* (10), 5771–5783.

(6) Burke, J. R.; Deshong, A. J.; Pelton, J. G.; Rubin, S. M. *J Biol Chem* **2010**, *285* (21), 16286–16293.

(7) Asghar, U.; Witkiewicz, A. K.; Turner, N. C.; Knudsen, E. S. *Nat. Rev. Drug Discov.* **2015**, *14* (2), 130–146.

(8) Johnston, P. A. *Curr Opin Chem Biol* **2011**, *15* (1), 174–182.

(9) Bockus, A. T.; McEwen, C. M.; Lokey, R. S. *Curr. Top. Med. Chem.* **2013**, *13* (7), 821–836.

(10) El Tayar, N.; Mark, A. E.; Vallat, P.; Brunne, R. M.; Testa, B.; van Gunsteren, W. F. *J. Med. Chem.* **1993**, *36* (24), 3757–3764.

(11) White, T. R.; Renzelman, C. M.; Rand, A. C.; Rezai, T.; Mcewen, C. M.; Gelev, V. M.; Turner, R. A.; Linington, R. G.; Leung, S. S. F.; Kalgutkar, A. S.; Bauman, J. N.; Zhang, Y.; Liras, S.; Price, D. A.; Mathiowetz, A. M.; Jacobson, M. P.; Lokey, R. S. *Nat. Chem. Biol.* **2011**, *7* (11), 810–817.

(12) Xiang, T.-X. X.; Anderson, B. D. *J. Membr. Biol.* **1994**, *140* (2), 111–122.

(13) Leung, S. S. F.; Sindhikara, D.; Jacobson, M. P. *J. Chem. Inf. Model.* **2016**, *56* (5), 924–929.

(14) Hewitt, W. M.; Leung, S. S. F.; Pye, C. R.; Ponkey, A. R.; Bednarek, M.; Jacobson, M. P.; Lokey, R. S. *J. Am. Chem. Soc.* **2015**, *137* (2), 715–721.

1 A Strategy for Direct Chemical Activation of the Retinoblastoma Protein

1.1 Abstract

The retinoblastoma (Rb) tumor suppressor protein negatively regulates cell proliferation by binding and inhibiting E2F transcription factors. Rb inactivation occurs in cancer cells upon Cyclin-dependent kinase (Cdk) phosphorylation, which induces E2F release and activation of cell cycle genes. We present a strategy for activating phosphorylated Rb with molecules that bind Rb directly and enhance affinity for E2F. We developed a fluorescence polarization assay that can detect the effect of exogenous compounds on modulating affinity of Rb for the E2F transactivation domain. We found that a peptide capable of disrupting the compact inactive Rb conformation increases affinity of the repressive Rb-E2F complex. Our results demonstrate the feasibility of discovering novel molecules that target the cell cycle and proliferation through directly targeting Rb rather than upstream kinase activity.

1.2 Introduction

Rb regulates proliferation through controlling the cell cycle, differentiation, senescence, and cell survival.[1–4] Whereas Rb orchestrates proper cellular signals with the mechanics of cell cycle progression, cancer cells almost invariably have alterations in Rb pathway components, which enables uncontrolled cell proliferation.[1,2,4–7] Importantly, while deletion of the gene and full loss of the Rb protein is observed in some cancers, in the vast majority of cases, Rb pathway inactivation in cancer cells is achieved through activation of Cyclin/Cdk complexes or inactivation of proteins that inhibit Cdk activity.[2,5,7] Thus, chemotherapeutic strategies that directly promote Rb activity would be relevant to most tumors.

We describe here a novel approach to reversing Rb inactivation with molecules that directly bind Rb itself. There are potential therapeutic advantages to such compounds over current Cdk inhibitors, including potency and specificity.[8] In addition, the specificity of such molecules would give them unprecedented advantages for studying the Rb pathway and its role in tumor suppression. For example, currently the only chemical approach to preventing Rb inactivation is through Cdk inhibition, which has off-target effects from preventing phosphorylation of other substrates.[8] Despite these motivations, beyond molecules that specifically inhibit Rb association with viral oncoproteins[9], no direct chemical probes of Rb exist.

1.3 Results

Rb arrests cells largely due to its ability to repress E2F-mediated gene expression[3]. Rb binds E2F primarily through an association of its so-called “pocket” domain with the E2F transactivation domain (E2FTD). E2FTD binding by the pocket is necessary for Rb activity in growth suppression, cell cycle control, and E2F inhibition[3]. The Rb pocket domain has an additional protein interaction cleft known as the “LxCxE” site, which binds oncogenic viral proteins and cellular proteins containing the LxCxExΦ sequence motif (Φ is a hydrophobic)[3,10]. Several specific Cdk phosphorylation events inhibit E2F binding upon S phase entry[11–14], however T373 phosphorylation has the most pronounced effect in quantitative in vitro assays[15,16]. Evidence also suggests that T373 phosphorylation is the most critical

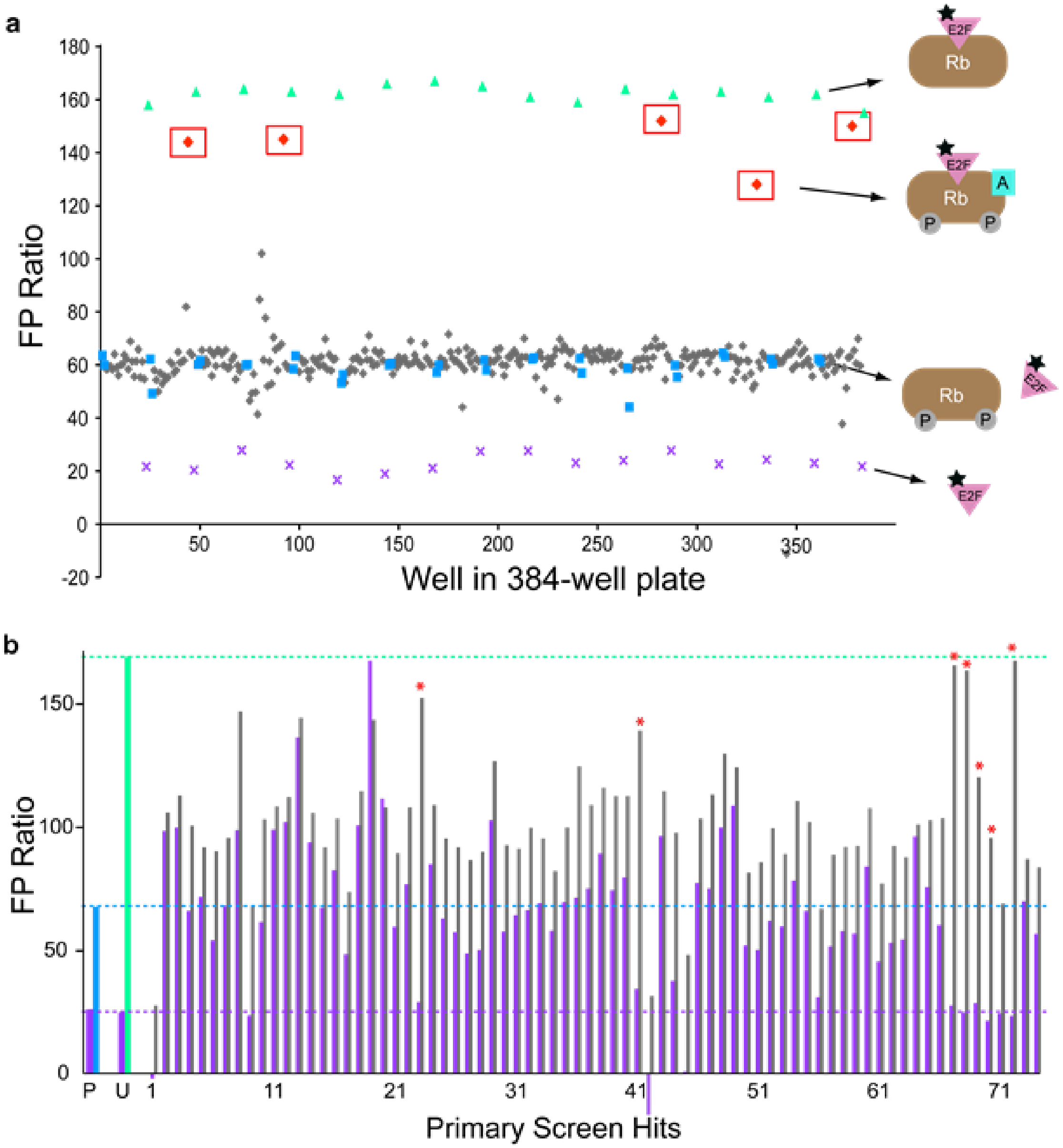


Figure 1-1 Fluorescence polarization screen for enhancers of Rb-E2F binding. (a) Sample data from the primary screen. FP ratio is plotted for each well in a 384-well plate. The wells contain phosphorylated RbNP and compounds (grey diamonds) or DMSO (blue squares, “negative control”), unphosphorylated RbNP (green triangles, “positive control”), or free E2FTMR alone (purple crosses). The boxed red diamonds are hits that increase the FP ratio of E2FTMR in the presence of phosphorylated RbNP. (b) Follow-up assay in which the effect of the compounds on E2FTMR FP ratio was determined in the absence (purple bars) and presence (grey bars) of phosphorylated RbNP. The phosphorylated RbNP negative (P, blue bar) and unphosphorylated RbNP (U, green bar) positive controls are shown at left. Hits were validated (red asterisks) if they yielded low FP ratios similar to controls in the RbNP target-minus (dashed purple line) assay and high FP ratios in the RbNP target-positive assay (dashed green line).

event for Rb inactivation in vivo. T373 is the only phosphorylation site sufficient for Cdk-induced inactivation of Rb in cells, and mutation of T373 significantly inhibits Cdk-induced Rb-E2F dissociation and E2F activation[17,18,19].

We set out to identify molecules that directly activate Rb by stabilizing the association of E2FTD with Cdk-phosphorylated Rb. To search for such compounds, we crafted a fluorescence polarization assay that is amenable to high throughput screening. An E2^{FTD} peptide (human E2F2 amino acids 409-428) was synthesized with a tetramethylrhodamine dye (TMR) at its N-terminus (E2F^{TMR}). We assayed binding of E2FTMR to an Rb protein construct (Rb^{NP}) that contains the Rb N-terminal domain (RbN) and pocket domain but lacks internal loops in each domain (residues 55–787, Δ245–267,Δ582–642)[16]. This minimized Rb construct contains two phosphorylation sites (T356 and T373) and the structural elements necessary and sufficient for recapitulating the inhibitory effect of T373 phosphorylation on E2^{FTD} binding.[16]

1.3.1 FP Rb/E2F^{TMR} High Throughput Screen

We assayed fluorescence polarization (FP) ratios in 384-well format using 10 nM E2F^{TMR} (**Error! Reference source not found.**a). The FP ratio for free E2F^{TMR} is ~20 in our assay conditions, and the FP ratio increases upon addition of 10 nM phosphorylated Rb^{NP} and 10 nM unphosphorylated Rb^{NP} to ~60 and ~165 respectively. The more modest FP ratio increase upon addition of phosphorylated Rb^{NP} reflects its approximately 10-fold weaker affinity for E2F^{TMR} compared to the affinity of unphosphorylated Rb^{NP} for E2F^{TMR}.[16] We conducted a pilot screen using a 21,120 small molecule library from ChemDiv that contains groups of analogs based on ~1200 structurally diverse “drug-like” scaffolds. 50 μM of each compound was incubated with 10 nM phosphorylated Rb^{NP}, and then 10 nM E2F^{TMR} was added. Control wells were used that had no protein (E2FTMR-only) or had 0.5% (by volume) DMSO added to either unphosphorylated or phosphorylated Rb^{NP}.

Sample data from one 384-well plate are shown in Figure 1a, and results for the entire screen can be seen in Figure 1-2. We looked for hit compounds that raised the FP ratio towards that of unphosphorylated, tighter-binding Rb^{NP}. In this primary screen, the assay had an average Z' = 0.82 ± 0.02.[20] 74 compounds were selected as hits (0.35% hit rate) that had FP ratios higher than the average FP of phosphorylated RbNP with B-scores[21] of 10 (Z-score of 7.7) or higher. These hits also satisfied the criterion that the measured overall fluorescence intensity was less than three standard deviations above average control fluorescence intensity.

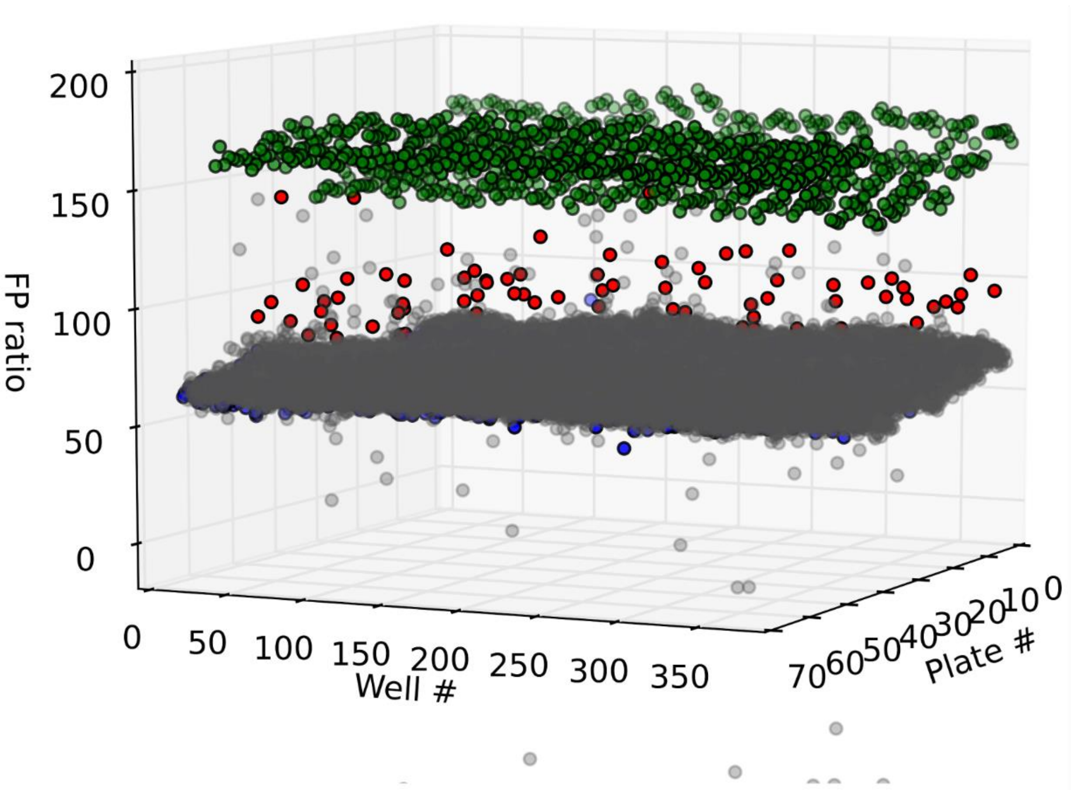


Figure 1-2 FP ratio for each well in the high throughput screen of the ChemDiv library. The wells contain phosphorylated Rb^{NP} and compounds (grey circles) or DMSO (blue circles, "negative control"), unphosphorylated Rb^{NP} (green circles, "positive control". The red circles are hits that increase the FP ratio of $E2F^{TMR}$ in the presence of phosphorylated Rb^{NP} and satisfy our total fluorescence criterion.

We recognized that because we are seeking compounds that increase the FP ratio, which reflects enhanced Rb binding, a molecule that induces E2F^{TMR} aggregation would be detected in the screen as a hit.[22] To rule out these false-positives, we tested screen hits in a follow-up “target-minus” assay in which phosphorylated Rb^{NP} was left out (**Error! Reference source not found.**b). Most of the initial hits resulted in a perturbed FP ratio in the absence of target, likely either the result of intrinsic fluorescence or compound induced aggregation of the TMR-peptide. We did find 7 hits in the library that had no effect in the absence of Rb^{NP} and enhanced the FP ratio in the presence of phosphorylated Rb^{NP} (red asterisks in **Error! Reference source not found.**b and **Error! Reference source not found.**). 4 of the 7 compounds increased the affinity of E2F^{TMR} for phosphorylated Rb^{NP} in a complete protein titration (Figure 1-3). Those four validated hits contain a common core scaffold based on 1,6-dimethylpyrimido[5,4-e][1,2,4]triazene-5,7(1H,6H)-dione. Such triazene compounds are known to generate reactive oxygen species (ROS) that may induce off-target effects in cells.[23] We found that triazene compound activity in the FP assay was lost in the presence of catalase (not shown), so we did not pursue them further.

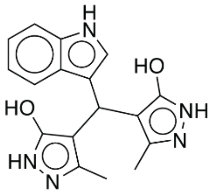
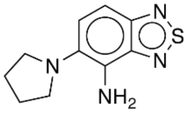
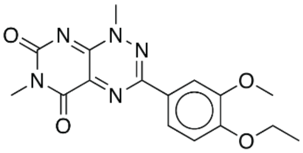
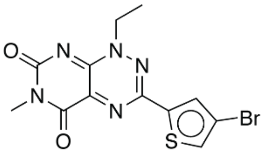
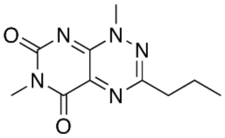
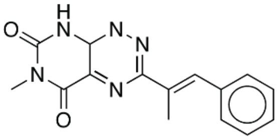
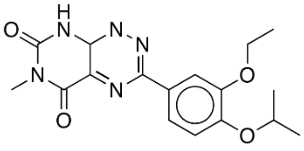
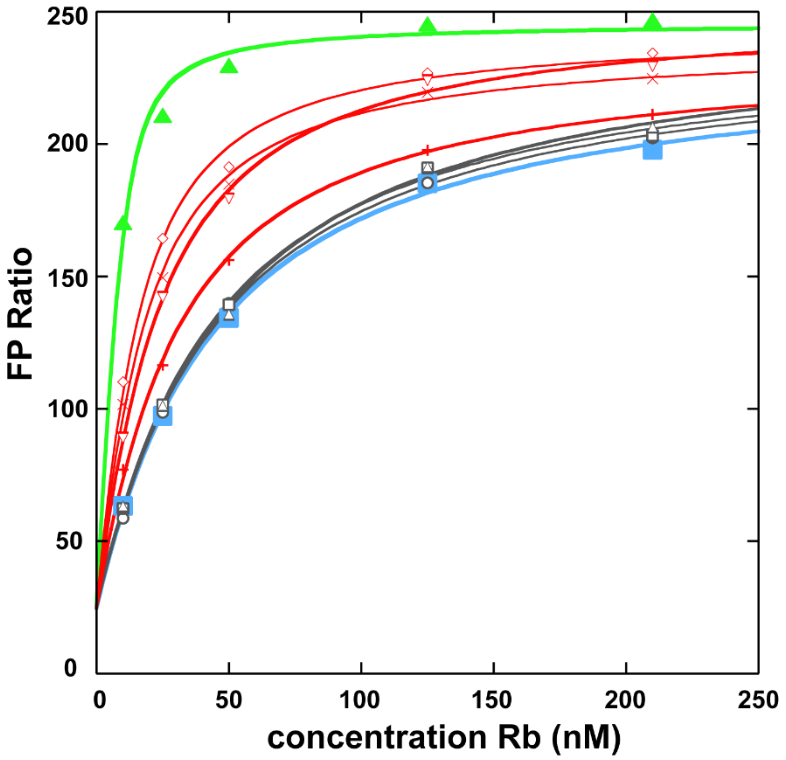
Structure	Hit Compound No.
	**23**
	**41**
	**67**
	**68**
	**69**
	**70**
	**72**

Table 1-1 Structures of validated hits in the primary screen.



		E2F^{TMR} K$_d$ (nM)
▲	unphosphorylated Rb^{NP}	2.2 ± 0.6
■	phosphorylated Rb^{NP}	42 ± 3
phosphorylated Rb^{NP} + hit compound #		
○	23	40 ± 2
□	41	39 ± 3
▽	67	19 ± 1
◇	68	11 ± 1
△	69	39 ± 3
+	70	26 ± 2
×	72	13 ± 1

Figure 1-3 Effect of validated primary screen hits in a complete Rb^{NP} protein titration. FP ratios were measured and K_d values for $E2F^{TMR}$-Rb^{NP}binding were determined as in Figure 3a.

1.3.2 Sanford Burnham Institute Screening Campaign

Encouraged by the robust and scalable nature of the FP assay we engaged the Sanford Burnham institute to screen their much larger library of ~320,000 compounds against Rb. Presumably with this 10 fold larger library the would drastically increase the likelihood of finding a genuine compound that elicited the desired stabilization of the Rb/E2F. To screen that number of compounds in a reasonable amount of time, the screening center employs 1536 format (4x denser than the 384 format employed at the UCSC Chemical Screening Center at the time of the screen). The assay passed all initial pilot screen controls run in house to verify that the reduced assay volume and different dispensing methods would not affect the z-score of the assay. They then proceeded to run all ~320k compounds against our assay and the results mirrored that of our in house screening. After eliminating compounds with intrinsic fluorescence in the TMR wavelength and those compounds that did not repeat in a dilution series resulting in 108 potential hit compounds (Figure 1-4). Of those hits, only 20 compounds had IC_{50} <10µM as estimated by a 5-point titration as performed by the Sanford-Burnham collaborators.

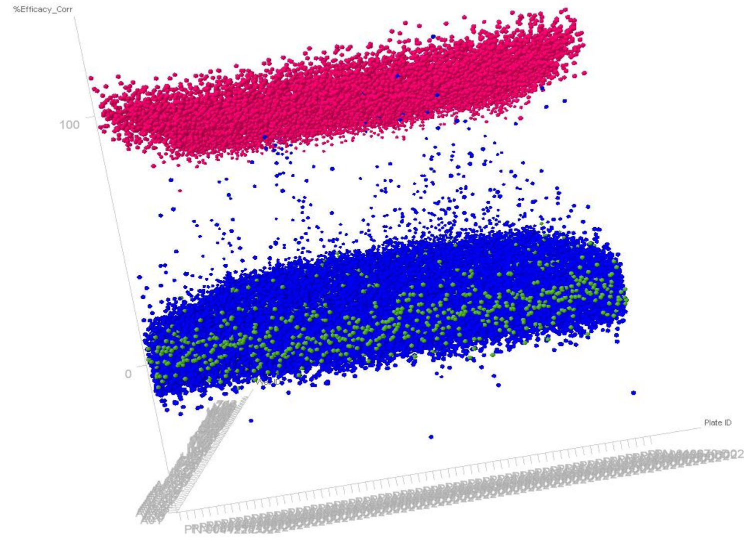


Figure 1-4 3D scatter plot of the Sanford-Burnham screening results. Red points represent the positive controls (un-phosphorylated Rb), green points represent negative controls (E2F^{TMR} alone), and blue points represent the assay compounds.

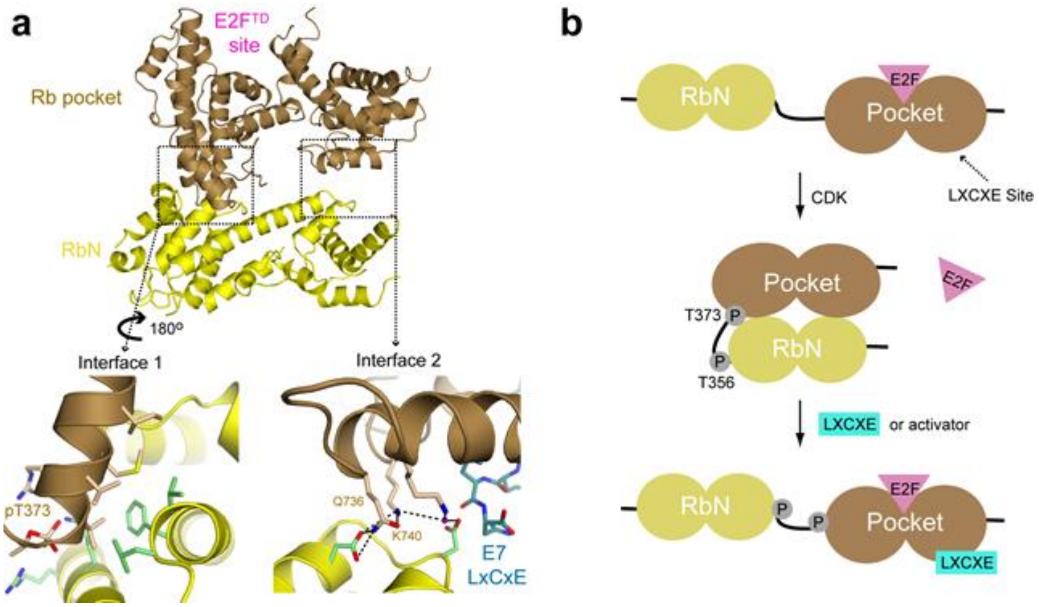


Figure 1-5 Structure-based strategy for activation of Rb. (a) Structure of phosphorylated Rb (from PDB code: 4ELJ). Docking between the Rb N-terminal domain (RbN, yellow) and the pocket domain (brown) occurs across two interfaces. Interface 1 is mediated by a pocket helix that is nucleated by T373 phosphorylation. Interface 2 is near the LxCxE-binding cleft in the pocket. The E7 peptide (cyan backbone), which is shown bound at its site in the unphosphorylated pocket (from PDB code: 4GUX), clashes with RbN residues at the interface. (b) Phosphorylation of sites in the Rb interdomain linker induces a conformational change that allosterically inhibits E2FTD binding. We find that an LxCxE peptide acts as an activator by binding Rb and inhibiting the RbN-pocket interdomain association.

We next explored a strategy for developing Rb activators that is motivated by the structural mechanism underlying how the Rb-E2F^{TD} complex is inhibited by Rb phosphorylation (Figure 1-5).[16] T373 phosphorylation induces an interdomain association between RbN and the pocket, which allosterically opens the E2F^{TD} binding cleft to weaken affinity.[16] RbN-pocket association occurs across two interfaces (Figure 1-5a), both of which must be formed to open up the E2FTD-binding site and disrupt interactions between the pocket and E2F^{TD}. One interface is anchored by the first helix of the pocket domain, which is nucleated by T373 phosphorylation and docks into a hydrophobic groove in RbN. The RbN position at the second interface is close to the LxCxE binding site in the pocket domain (see for example the structure of the human papilloma virus (HPV) E7 LxCxE peptide-pocket domain complex).[10,16] From a structural alignment (Figure 1-5a), the binding of the LxCxE peptide and interdomain docking appear incompatible.

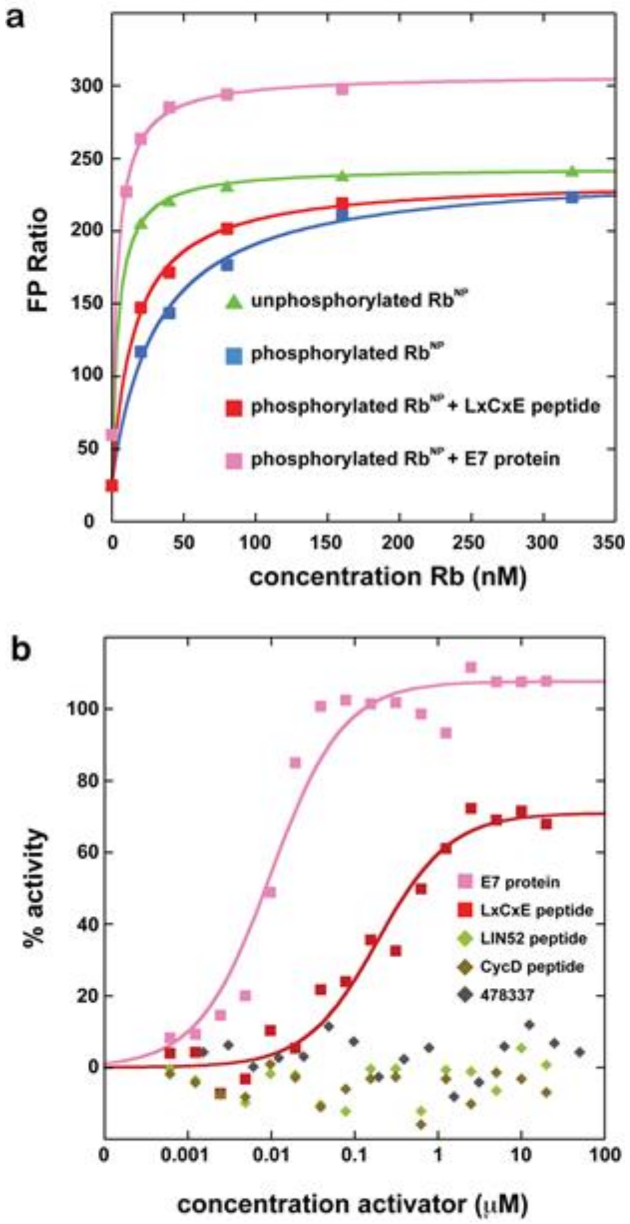


Figure 1-6 . LxCxE peptide acts as an Rb activator. (a) Titration of unphosphorylated and phosphorylated RbNP into E2FTMR. In the presence of the E7 LxCxE peptide and full-length E7 protein (pink), the affinity is increased. (b) EC50 measurement of LxCxE peptide and E7 protein activity. Compound #478337 from Fera et al.9 (grey) and LxCxE variant peptides from Cyclin D and LIN52 do not show activity.

We reasoned that molecules that inhibit interdomain docking would stabilize E2F^{TD} binding to phosphorylated Rb by preventing the allosteric opening of the E2F binding site (Figure 1-5b).

Moreover, because the LxCxE peptide binds near the second RbN-pocket interface (Interface 2 in Figure 1-5a), we hypothesized that it would disrupt docking and act as an activator. We tested the effects of the LxCxE peptide from the HPV E7 protein on $E2F^{TMR}$ binding to Rb^{NP} using the FP assay as in the pilot screen but using a full protein titration (Figure 1-6a). The affinity of $E2F^{TMR}$ for unphosphorylated RbNP (K_d = 4.3 ± 0.2 nM) is 8-fold tighter than its affinity for phosphorylated RbNP in this assay (K_d = 31 ± 4 nM). In the presence of 10 µM E7 LxCxE peptide, phosphorylated RbNP binds $E2F^{TMR}$ with 2-fold higher affinity (K_d = 16 ± 2 nM), which implicates the LxCxE peptide as an example of a desired Rb activator molecule. In the presence of 2 µM full length E7 protein, $E2F^{TMR}$ binds phosphorylated Rb^{NP} with similar affinity (K_d = 4.6 ± 0.3 nM) as unphosphorylated Rb^{NP}.

We found that the LxCxE peptide and full-length E7 increased RbNP-$E2F^{TMR}$ affinity with EC50 = 190 ± 40 nM and EC50 = 10 ± 2 nM respectively (Figure 1-6b). The greater potency of the full-length protein correlates with its known 20-fold greater affinity for the Rb pocket domain.[24] We suggest that the greater highest activity of the protein (~100%) compared to the peptide (~70%) may result from the fact that its larger size is better suited for occluding the RbN-pocket interface. We found that the LxCxE-peptide and E7 protein do not affect E2FTD binding to Rb if the docking interface is mutated (Q736A/K740A) (Figure 1-7), which supports further our proposal that these activators function by disrupting interdomain docking. We also tested two LxCxE-like peptides from Cyclin D (no hydrophobic in +2 position) and LIN52 (LxSxE), which have weak affinity for Rb.[25] These variant peptides and a compound previously reported to bind the LxCxE cleft (compound #478337 from Fera et al. 9) show no effect in the Rb activation assay (Figure 1-6b).

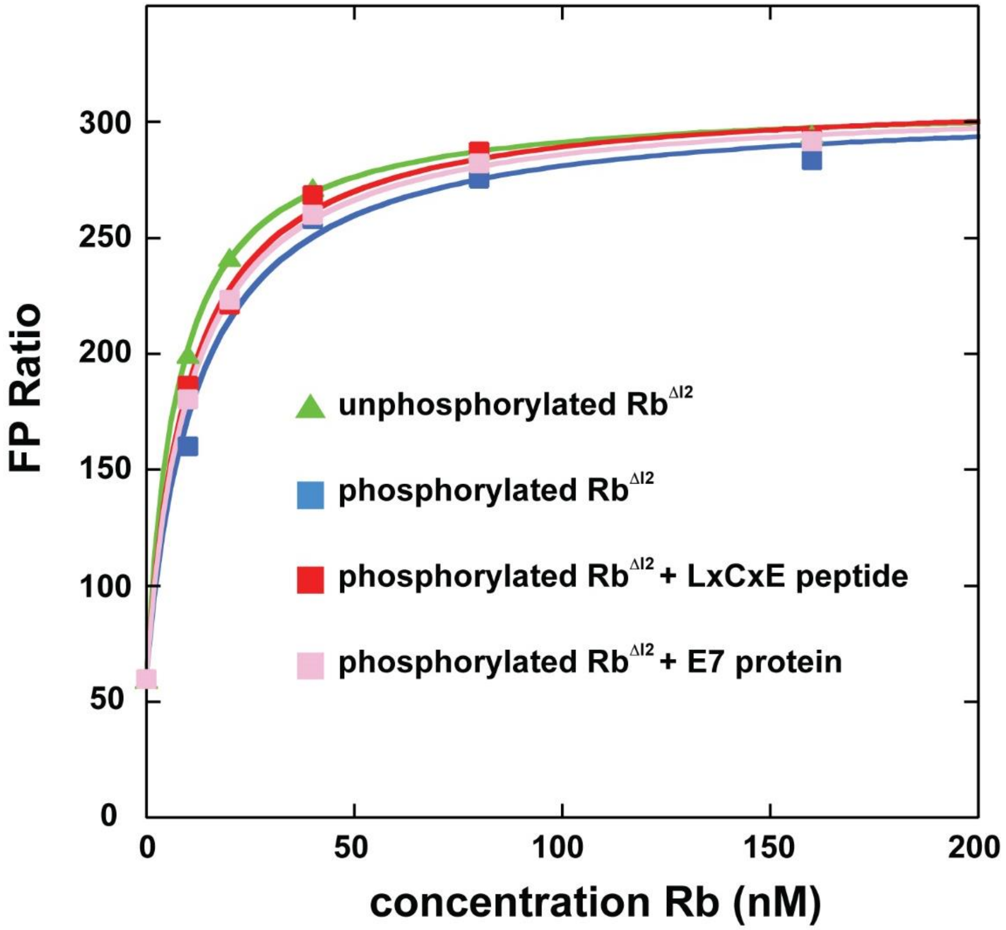


Figure 1-7 The LxCxE peptide and E7 protein activators do not affect E2F binding to an Rb mutant in which interdomain docking is disrupted. Binding measurements are shown of E2F$^{\text{TMR}}$ to an Rb$^{\text{NP}}$ construct containing Q736A and K740A mutations (Rb$^{\otimes}_{\text{I2}}$). As described previously[16], E2F binding to the unphosphorylated (K$_d$=7.5 ± 0.3 nM) and phosphorylated (K$_d$=12 ± 2 nM) mutant are similar. Addition of 2 ∝M LxCxE peptide (K$_d$=10 ± 1 nM) or E7 protein (K$_d$=10.5 ± 0.3 nM) does not increase E2F affinity, indicating that the activating effect is not independent of interdomain docking. We confirmed using ITC that the LxCxE peptide still binds Rb$^{\otimes}_{\text{I2}}$ with K$_d$=180 ± 20 nM.

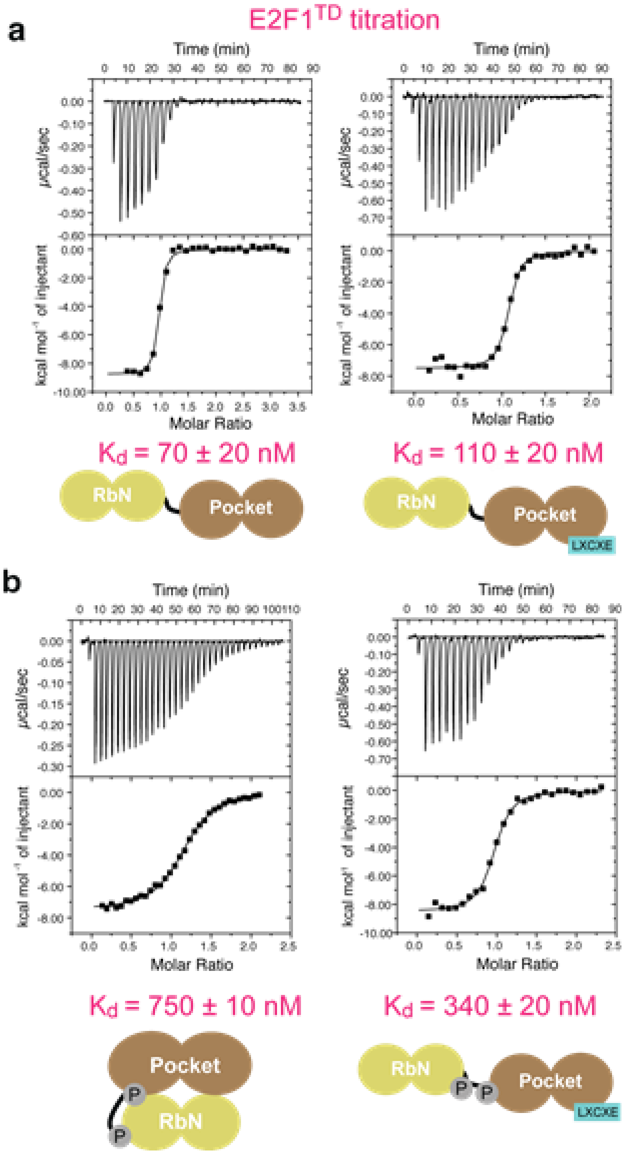


Figure 1-8 The E7 LxCxE peptide increases affinity of E2FTD for phosphorylated Rb. Representative ITC curves and average K_d measurements are shown for E2F1TD titration into unphosphorylated (a) and phosphorylated (b) RbNP.

To confirm the stabilizing effect of the LxCxE peptide in an orthogonal assay, we measured affinities using isothermal titration calorimetry (Figure 1-8). We found that E2F1TD binds unphosphorylated RbNP with similar affinity in the absence (K_d = 70 ± 20 nM) and presence (K_d =

110 ± 20 nM) of excess E7 LxCxE peptide (Figure 4a). In contrast, the affinity of E2FTD for phosphorylated RbNP is enhanced in the presence of LxCxE peptide (K_d = 340 ± 20 nM) compared to in its absence (K_d = 750 ± 10 nM) (Figure 1-8b). The observed increase in affinity that is specific for phosphorylated Rb demonstrates that molecules that interfere with the structural changes induced by Rb phosphorylation can act as Rb activators. We note that while the fold-change due to the LxCxE peptide is similar in the ITC assay as in the FP assay, measured RbNP-E2F affinities are greater using FP, perhaps due to the hydrophobic TMR dye.

1.4 Discussion

In conclusion, we have developed a robust fluorescence polarization assay for screening molecules that modulate the binding between Rb and E2F, and we found that an LxCxE peptide from the HPV E7 protein, which is known to bind at the RbN-pocket interface[10], increases affinity of the complex. Several observations support the idea that isolated LxCxE peptides or derivatives could be used as Rb-E2F stabilizers in cells. While the E7 and related viral oncoproteins disrupt Rb-E2F complexes in cells to stimulate proliferation, in each case the LxCxE-containing domain is insufficient. Additional domains that directly inhibit Rb-E2F association, are required, and in the case of E7, the additional domain targets binding of the Rb C-terminal domain to E2F.[26–28] Notably, expression of an SV40 virus T-antigen protein mutant, which contains the LxCxE motif but lacks the Rb-E2F dissociating domain, enhances the population of Rb-E2F complexes relative to free E2F in fibroblast cells[27]. Moreover, the fact that T373 mutation inhibits Rb-E2F dissociation and E2F activation in cells[17–19] suggests that the affinity increase achieved here, which negates the effect of T373 phosphorylation, may effectively modulate Rb activity in vivo.

While the E7 LxCxE peptide is effective in vitro, its likely poor pharmacokinetic properties and extended binding structure make it a suboptimal lead for a therapeutic or chemical probe. However, we envision developing peptide mimics that circumvent these shortcomings such as stapled or cyclic peptides. A group of thiadiazolidinedione compounds has been reported to competitively inhibit Rb-LxCxE association.[9] However, in an experiment with one such compound reported to bind Rb with 200 nM affinity (#478337 in Fera et al.[9]), we did not observe any effect on E2F^{TD} affinity for phosphorylated Rb^{NP} (Figure 1-6b). It may be that the specific molecular requirements of inhibiting viral protein LxCxE binding to the pocket cleft and disrupting the RbN-pocket interdomain docking are distinct. Indeed, the location of the LxCxE peptide binding site is adjacent to but not directly overlapping the interdomain interface (Figure 1-5a). The full-length E7 protein may be a more effective activator than the peptide (Figure 1-6) because additional interactions occlude this interface. We propose that LxCxE peptide derivatives may be more active if they are extended at their N-terminus to overlap more extensively with the RbN docking surface in the pocket. With respect to further screening, our results suggest that libraries containing compounds with larger scaffolds, such as natural products derived libraries, may be more suitable for producing lead activators. We anticipate that ideal compounds would access the hydrophobic pocket bound by the leucine in the 'LxCxE' peptide and also contact the Q736/K740 surface that forms the docking interface.

In addition to the molecules discussed here, we propose also developing Rb activators that bind the RbN groove forming the primary RbN-pocket interface (Interface 1 in Figure 1-5a). By inhibiting docking of the phosphorylated pocket helix, these molecules may stabilize E2FTD binding to phosphorylated Rb as desired. The pocket helix contacts the groove using hydrophobic residues along one face of an amphipathic helix. This type of interaction has been successfully targeted by small molecules.[29–34] A classic example is the p53-MDM2 interface, formed by residues in the i, i+4, and i+7 positions of a p53 helix, for which cis-imadazoline (Nutlin) and spirooxindole (MI-219) inhibitors have been found.[31,33,34] Therefore, while our screen results suggest that disrupting the RbN-pocket association is a challenge for small molecules, recent successes in protein-protein interaction inhibition suggests the promise of finding a lead compound for direct Rb activation.

1.5 Methods

1.5.1 Protein and peptide reagents

Rb^{NP} and E2F1TD (E2F1 residues 409-426) were expressed in E. coli as GST-fusion proteins and purified with GS4B sepharose as previously described16. Following elution from the affinity resin, the fusion protein was diluted three-fold into a buffer containing 25 mM Tris and 1 mM DTT (pH 8.0). Protein was then loaded onto a Source Q ion exchange column equilibrated in the same low salt buffer and eluted from the column in a gradient of 0-1 M NaCl. The fusion protein eluted in a single peak and was digested overnight at 4 °C in the presence of 1% (by mass) TEV protease. The samples were loaded again onto GS4B to remove the free GST, and the proteins were collected and concentrated to ~5 mg/mL for future assays. Phosphorylation of Rb^{NP} was achieved as previously described using 10% (by mass) purified Cdk2-CycA in a reaction containing 5 mM ATP and 20 mM MgCl215. After an overnight reaction at 4°C, quantitative phosphorylation on two sites was validated by observation of an increase in molecular mass of ~160 Da using electrospray ionization mass spectrometry (Figure 1-9). Synthetic E7 LxCxE (DLYCYEQLN), LIN52 (TDLEASLLSFEKLDRAphosSPDLWPE), Cyclin D (MEHQLLCCEVETIRRAY), and TMR-E2F2TD (QDDYLWGLEAGEGISDLFD) peptides were ordered from Genscript, LLC.

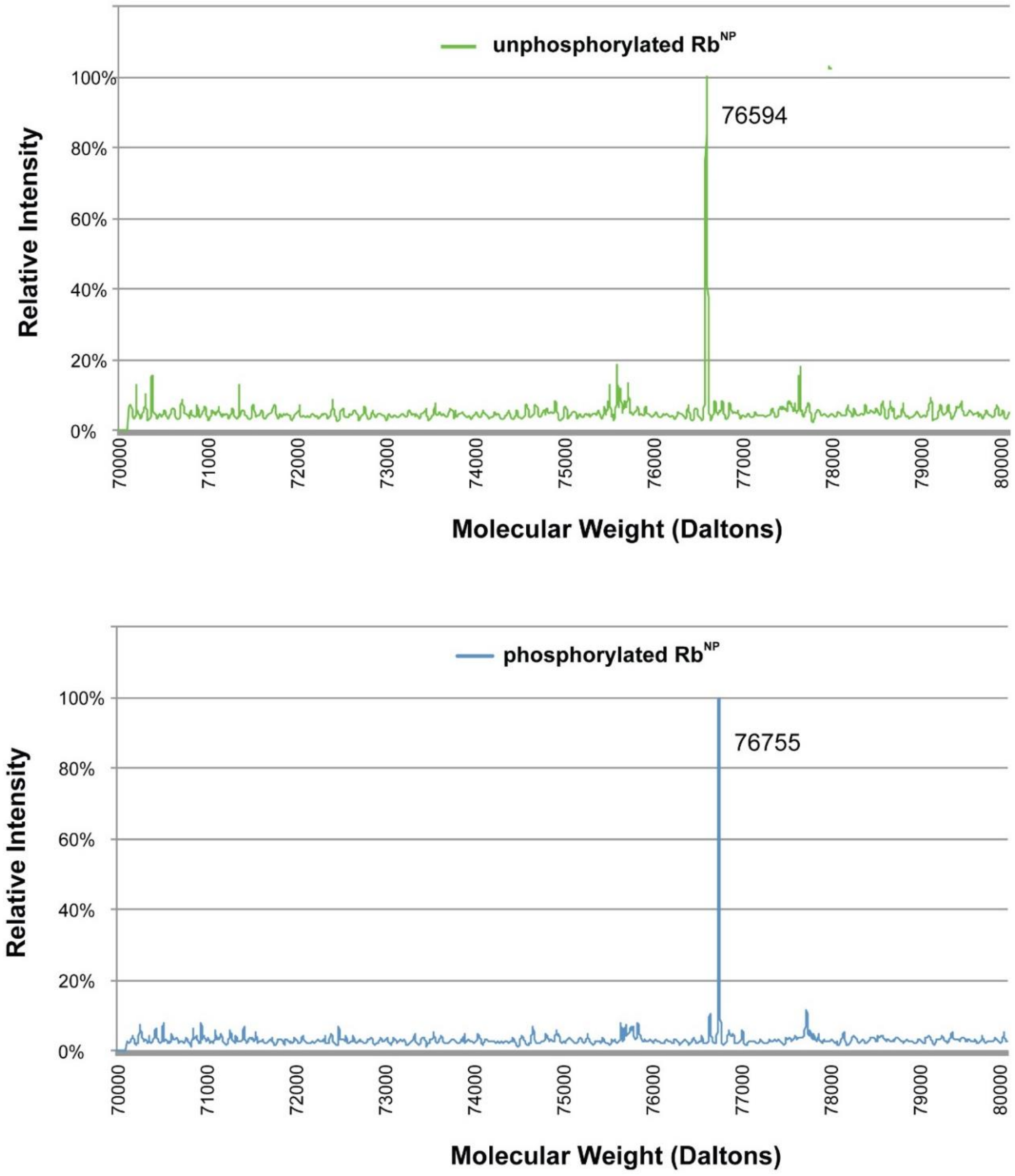


Figure 1-9 Electrospray ionization mass spectra of intact samples of unphosphorylated and phosphorylated $\mathrm{Rb}^{\mathrm{NP}}$. Samples were prepared by desalting. Each spectrum shows a single, dominant peak at the indicated molecular weight. The ~160 Da increase in the phosphorylated sample corresponds to addition of two phosphates.

1.5.2 Isothermal Titration Calorimetry

ITC experiments were performed with a VP-ITC calorimeter from Microcal, LLC. (now supported by Malvern Instruments). Prior to the measurements, unphosphorylated and phosphorylated Rb^{NP} and $E2F1^{TD}$ were dialyzed overnight in the same beaker against a buffer containing 40 mM Tris, 100 mM NaCl, and 1 mM 2-mercaptoethanol (pH 8.0). Measurements were made with an $E2F1^{TD}$ concentration of 1 mM, an Rb^{NP} concentration of 15-25 µM, and an LxCxE peptide concentration of 100 µM. The reported K_d values are the average of 2-3 measurements, and the standard deviation is reported as the error.

1.5.3 Fluorescence Polarization Assay and Screen

Fluorescence polarization measurements were made in black, untreated 384-well plates (Corning). For the screen, 20 µL of a 40 nM solution of Rb^{NP} was dispensed using a Matrix Wellmate peristaltic pump. Compounds or DMSO were pin-transferred using a 200 nL pin tool (Perkin Elmer) to a final concentration of 50 µM. Then 20 µL of a 20 nM $E2F^{TMR}$ solution was added for a resulting final concentration of 20 nM Rb^{NP} and 20 nM $E2F^{TMR}$. All solutions were prepared using a buffer containing 40 mM Tris, 100 mM NaCl, 1 mM DTT, and 0.1% Tween-20 (pH 8.0). Total fluorescence and fluorescence polarization were measured using a Perkin-Elmer Envision plate reader. An excitation filter centered around a wavelength of 531 nM and with a bandwidth of 20 nM was used along with a emission filter centered around 595 nM and with a bandwidth of 60 nM. Fluorescence polarization ratios were calculated as FP = 1000*(S-G*P)/(S+G*P), where S is intensity of fluorescence parallel to excitation plane, P is perpendicular fluorescence intensity, and G is a correction factor to ensure positive ratio values. A positive control of unphosphorylated Rb^{NP} was used in each screen plate along with a negative control of DMSO added to phosphorylated Rb^{NP}. Additionally, 10 nM free $E2F^{TMR}$ was used as a positive inhibitory control and represents the lowest potential FP value. This control was also used in the absence of Rb to test whether compounds induced nonspecific effects. For the protein titration experiments, 10 µM of E7 LxCxE peptide, 2 µM of E7 protein, or 50 µM compound was added to prepared solutions of Rb^{NP} at the different indicated concentrations. Binding constants were determined from fits of the protein titration data using a two-site binding, and the y-intercept was fixed to the FP value of the $E2F^{TMR}$ peptide. Reported errors in the K_d from FP measurements are curve-fitting errors. The EC50 measurement was performed using conditions similar to the screen except 20 nM Rb^{NP} was used. We report % activity = (FPphosRb+activator – FPphosRb) / (FPunphosRb – FPphosRb).

1.6 Bibliography

(1) Burkhart, D. L.; Sage, J. *Nat Rev Cancer* **2008**, *8* (9), 671–682.

(2) Classon, M.; Harlow, E. *Nat Rev Cancer* **2002**, *2* (12), 910–917.

(3) Dick, F. A.; Rubin, S. M. *Nat Rev Mol Cell Biol* **2013**, *14* (5), 297–306.

(4) Lipinski, M. M.; Jacks, T. *Oncogene* **1999**, *18* (55), 7873–7882.

(5) Knudsen, E. S.; Knudsen, K. E. *Nat Rev Cancer* **2008**, *8* (9), 714–724.

(6) Malumbres, M.; Barbacid, M. *Nat Rev Cancer* **2001**, *1* (3), 222–231.

(7) Sherr, C. J. *Science (80-.).* **1996**, *274* (5293), 1672–1677.

(8) Stone, A.; Sutherland, R. L.; Musgrove, E. A. *Crit Rev Oncog* **2012**, *17* (2), 175–198.

(9) Fera, D.; Schultz, D. C.; Hodawadekar, S.; Reichman, M.; Donover, P. S.; Melvin, J.; Troutman, S.; Kissil, J. L.; Huryn, D. M.; Marmorstein, R. *Chem Biol* **2012**, *19* (4), 518–528.

(10) Lee, J. O.; Russo, A. A.; Pavletich, N. P. *Nature* **1998**, *391* (6670), 859–865.

(11) Rubin, S. M. *Trends Biochem Sci* **2013**, *38* (1), 12–19.

(12) Buchkovich, K.; Duffy, L. A.; Harlow, E. *Cell* **1989**, *58* (6), 1097–1105.

(13) DeCaprio, J. A.; Ludlow, J. W.; Lynch, D.; Furukawa, Y.; Griffin, J.; Piwnica-Worms, H.; Huang, C. M.; Livingston, D. M. *Cell* **1989**, *58* (6), 1085–1095.

(14) Lees, J. A.; Buchkovich, K. J.; Marshak, D. R.; Anderson, C. W.; Harlow, E. *Embo J* **1991**, *10* (13), 4279–4290.

(15) Burke, J. R.; Deshong, A. J.; Pelton, J. G.; Rubin, S. M. *J Biol Chem* **2010**, *285* (21), 16286–16293.

(16) Burke, J. R.; Hura, G. L.; Rubin, S. M. *Genes Dev* **2012**, *26* (11), 1156–1166.

(17) Brown, V. D.; Phillips, R. A.; Gallie, B. L. *Mol Cell Biol* **1999**, *19* (5), 3246–3256.

(18) Knudsen, E. S.; Wang, J. Y. *Mol Cell Biol* **1997**, *17* (10), 5771–5783.

(19) Lents, N. H.; Gorges, L. L.; Baldassare, J. J. *Cell Cycle* **2006**, *5* (15), 1699–1707.

(20) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. *J Biomol Screen* **1999**, *4* (2), 67–73.

(21) Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A. *J Biomol Screen* **2003**, *8* (6), 634–647.

(22) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. *J Med Chem* **2015**, *58* (17), 7076–7087.

(23) Johnston, P. A. *Curr Opin Chem Biol* **2011**, *15* (1), 174–182.

(24) Jones, R. E.; Wegrzyn, R. J.; Patrick, D. R.; Balishin, N. L.; Vuocolo, G. A.; Riemen, M. W.; Defeo-Jones, D.; Garsky, V. M.; Heimbrook, D. C.; Oliff, A. *J. Biol. Chem.* **1990**, *265* (22), 12782–12785.

(25) Guiley, K. Z.; Liban, T. J.; Felthousen, J. G.; Ramanan, P.; Litovchick, L.; Rubin, S. M. *Genes Dev.* **2015**, *29* (9), 961–974.

(26) Fattaey, A. R.; Harlow, E.; Helin, K. *Mol Cell Biol* **1993**, *13* (12), 7267–7277.

(27) Patrick, D. R.; Oliff, A.; Heimbrook, D. C. *J Biol Chem* **1994**, *269* (9), 6842–6850.

(28) Zalvide, J.; Stubdal, H.; DeCaprio, J. A. *Mol Cell Biol* **1998**, *18* (3), 1408–1415.

(29) Bullock, B. N.; Jochim, A. L.; Arora, P. S. *J Am Chem Soc* **2011**, *133* (36), 14220–14223.

(30) Cummings, C. G.; Hamilton, A. D. *Curr Opin Chem Biol* **2010**, *14* (3), 341–346.

(31) Grasberger, B. L.; Lu, T.; Schubert, C.; Parks, D. J.; Carver, T. E.; Koblish, H. K.; Cummings, M. D.; LaFrance, L. V; Milkiewicz, K. L.; Calvo, R. R.; Maguire, D.; Lattanze, J.; Franks, C. F.; Zhao, S.; Ramachandren, K.; Bylebyl, G. R.; Zhang, M.; Manthey, C. L.; Petrella, E. C.; Pantoliano, M. W.; Deckman, I. C.; Spurlino, J. C.; Maroney, A. C.; Tomczuk, B. E.; Molloy, C. J.; Bone, R. F. *J Med Chem* **2005**, *48* (4), 909–912.

(32) Gunther, J. R.; Parent, A. A.; Katzenellenbogen, J. A. *ACS Chem Biol* **2009**, *4* (6), 435–440.

(33) Shangary, S.; Qin, D.; McEachern, D.; Liu, M.; Miller, R. S.; Qiu, S.; Nikolovska-Coleska, Z.; Ding, K.; Wang, G.; Chen, J.; Bernard, D.; Zhang, J.; Lu, Y.; Gu, Q.; Shah, R. B.; Pienta, K. J.; Ling, X.; Kang, S.; Guo, M.; Sun, Y.; Yang, D.; Wang, S. *Proc Natl Acad Sci U S A* **2008**, *105* (10), 3933–3938.

(34) Vassilev, L. T.; Vu, B. T.; Graves, B.; Carvajal, D.; Podlaski, F.; Filipovic, Z.; Kong, N.; Kammlott, U.; Lukacs, C.; Klein, C.; Fotouhi, N.; Liu, E. a. *Science* **2004**, *303* (5659), 844–848.

2 Non-Classical Size Dependence of Permeation Defines Bounds for Passive Adsorption of Large Drug Molecules.

2.1 Foreword

The following work was completed as collaboration between all the following authors:

Cameron R. Pye†, William M. Hewitt†, Joshua Schwochert†, Terra D. Haddad†, Chad E. Townsend†, Lyns Etienne†, Yongtong Lao†, Chris Limberakis‖, Akihiro FurukawaΔ, Alan M. Mathowetz, David A. Price○, Spiros Liras○, and R. Scott Lokey*†

† Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064, United States

‖ World Wide Medicinal Chemistry, Groton Laboratories, Pfizer Inc. Groton, Connecticut 06340, United States

○ World Wide Medicinal Chemistry, Cambridge Laboratories, Pfizer Inc. Cambridge, Massachusetts 02139, United States

Δ Modality Research Laboratories, Daiichi Sankyo Co., Ltd, 1-2-58 Hiromachi, Shinagawa-ku, Tokyo 140-8710, Japan

The Following was published in ACS Journal of Medicinal Chemistry:

Pye, C. R.; Hewitt, W. M.; Schwochert, J.; Haddad, T. D.; Townsend, C. E.; Etienne, L.; Lao, Y.; Limberakis, C.; Furukawa, A.; Mathiowetz, A. M.; Price, D. A.; Liras, S.; Lokey, R. S. Nonclassical Size Dependence of Permeation Defines Bounds for Passive Adsorption of Large Drug Molecules. *J. Med. Chem.* **2017**, *60*, 1665–1672.

2.2 Abstract

Macrocyclic peptides are considered large enough to inhibit “undruggable” targets, but the design of passively cell-permeable molecules in this space remains a challenge due to the poorly understood role of molecular size on passive membrane permeability. Using split-pool combinatorial synthesis, we constructed a library of cyclic, per-N-methlyated peptides spanning a wide range of calculated lipohilicities (0 < AlogP < 8) and molecular weights (~800Da < MW < ~1200Da). Analysis by the Parallel Artificial Membrane Permeability Assay (PAMPA) revealed a steep drop-off in apparent passive permeability with increasing size, in stark disagreement with current permeation models. This observation, corroborated by a set of natural products, helps define criteria for achieving permeability in larger molecular size regimes and suggests an operational cut-off beyond which passive permeability is constrained by a sharply increasing penalty on membrane permeation.

2.3 Introduction

Predicting membrane permeability is a key challenge in medicinal chemistry, as diffusion across membranes is required for orally delivered drugs and for drugs with intracellular targets. In modern medicinal chemistry, the “drug-likeness” of a compound is often evaluated by comparing its properties to those of a historical training set of drugs, giving rise to knowledge-based structure-property relationships to help guide the design of drugs with favorable properties. These quantitative (and qualitative) models include Lipinski’s Rule of 5 (Ro5), which provides a simple metric for predicting ADME properties based on molecular weight (MW), number of hydrogen bond donors and acceptors, and octanol-water partition coefficient (LogP).[1] Compounds that do not follow Lipinski’s rules are more likely to be rejected from screening collections or eliminated from consideration in medicinal chemistry campaigns, thus reinforcing the criteria and potentially hindering exploration outside this well-defined chemical space. Indeed, there are a number of notable outliers that challenge the wisdom of a strict adherence to these ADME rules. Particularly noteworthy are compounds with molecular weights above 1000 Da that show significant passive membrane permeability and, in some cases, oral bioavailability.[2] Of these rule-breaking molecules, some are macrocyclic natural products such as the cyclic peptide cyclosporine A (CSA), a 1202-Da compound that is cell permeable by passive diffusion and has an oral bioavailability of 28%. Outliers such as CSA suggest that compounds outside the Rule of 5 are able to exhibit favorable ADME properties and that potentially vast regions of chemical space remain underutilized. Recent work has focused on the classification of these "beyond Rule of 5" (bRo5) compounds using QSPR techniques based on calculable molecular properties that often belie their physical underpinnings.[2–4] The question remains to what extent CSA and similar compounds are true outliers, or whether they point toward untapped chemical potential and the prospect of designing permeable molecules with molecular weights well outside conventional boundaries.

There is an extensive body of literature on the physics of passive membrane permeability, dating back to the solubility-diffusion theory (1) proposed by Overton in the 19th century.[5] This model treats the membrane as a thin slab of hydrocarbon solvent in which the rate of membrane permeation (the permeability coefficient Pm) is equal to the hydrocarbon/water partition coefficient $K_{hc/w}$ times the membrane diffusion coefficient D, divided by the membrane thickness (∂):

(1) $P\_m=(K_{(hc/w)}\ D)/\partial$

(2) $D=\ (k_B T)/6\pi\eta r$

This model's only form of size dependence is the diffusion term, which has been approximated by the Stokes-Einstein equation (2) and is dependent on the radius of the solute (r), the viscosity of the medium (η), and the temperature (T). However, this diffusion term alone does not accurately capture the observed size dependence on permeability, even within Ro5 space. More recent models have attempted to address this discrepancy by asserting that diffusion through the ordered membrane more closely resembles the diffusion of solutes in polymers.[6] In polymer-type diffusion, the solute migrates through the medium by traversing a series of stochastically appearing holes which are of greater volume than that of the compound.[7] Extending this principle, Xiang and Anderson proposed that there was an additional penalty associated with a lateral pressure that must be overcome during partitioning into the membrane in order to create a cavity of sufficient volume to accommodate the solute.[8,9] Recently these models have been successfully applied to the prediction of permeabilities for compounds in the “extended Ro5” size range (500~850 Da) but have yet to successfully predict the permeability of compounds approaching the weight of CSA.[10,11]

Only a handful of natural products and, more recently, synthetic cyclic peptides, have been described with molecular weights above ~1000 that show bone fide passive membrane permeability.[12,13] This raises the question whether or to what extent the design rules developed for extended Ro5 compounds will apply to these significantly larger systems. Herein, we report the application of a synthetic library approach towards understanding the impact of lipophilicity and molecular size on the intrinsic permeability of peptidic macrocycles in the size range of CSA and similar natural products. We designed a library of cyclic peptides with MW > ~800, and assayed their permeability in order to elucidate the size dependence of permeability in this space. All compounds were permethylated at the backbone to eliminate intramolecular hydrogen bonding (IMHB) and the resulting conformational effects as potentially complicating factors. After accounting for aqueous solubility and using experimentally determined hydrocarbon-water partition coefficients, we observed a steep, unprecedented size penalty for increasing molecular size in the bRo5 space that cannot be accounted for using current physics-based or QSPR models of permeability.

2.4 Results

2.4.1 Aggregation Imposes a Limit on Apparent Permeability

In order to quantify permeability over a wide range of sizes and lipophilicities, we synthesized a split pool library of hydrophobic octapeptides that spanned a MW range from 784 to 938. All library members contained a Tyr residue for linkage to the resin via the side chain phenol, as well as a D-Pro residue to facilitate cyclization. The additional residues were comprised of combinations of Ala, Leu, and Ile. All of the library members had the same backbone stereochemistry arising from a previous, unpublished scaffold whose high predicted permeability was compromised by poor solubility. Additionally, in the final synthetic step before cleavage from solid support all the scaffolds were permethylated. Complete N-methylation was designed to eliminate the possibility of intramolecular hydrogen bonding in the backbone, thus providing a controlled model system for evaluating the impact of size and lipophilicity independent of backbone conformation. A separate library was synthesized in tandem using Ile in place of Leu to test whether the shape of the alkyl sidechains impact any observed permeability trends.

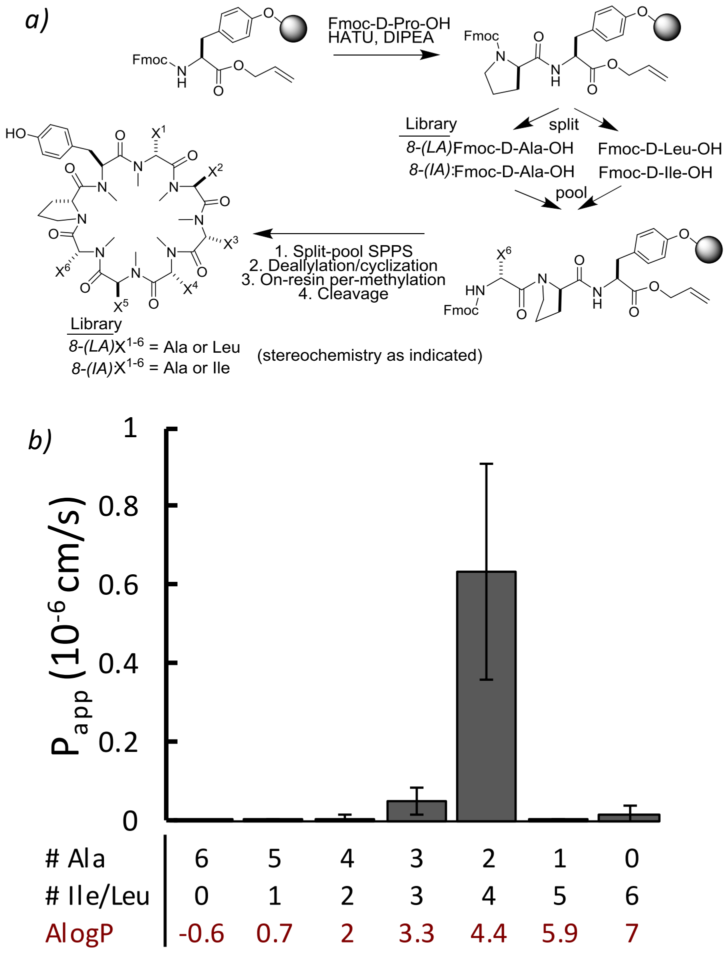
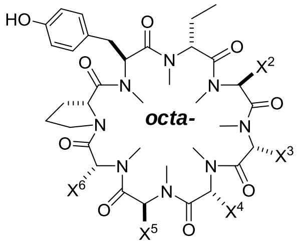
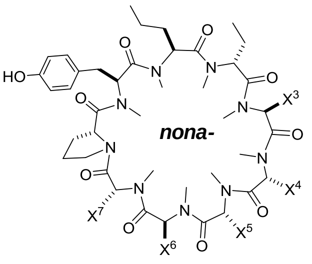

a)
Fmoc
N
H
O
O
Fmoc-D-Pro-OH
HATU, DIPEA
Fmoc
N
O
N
H
O
O
split
Library
8-(LA)Fmoc-D-Ala-OH
8-(IA):Fmoc-D-Ala-OH
Fmoc-D-Leu-OH
Fmoc-D-Ile-OH
pool
HO
X1
X2
X3
X4
X5
X6
1. Split-pool SPPS
2. Deallylation/cyclization
3. On-resin per-methylation
4. Cleavage
HN
Fmoc
Library
8-(LA)X1-6 = Ala or Leu
8-(IA)X1-6 = Ala or Ile
(stereochemistry as indicated)
b)
Papp (10-6 cm/s)
1
0.8
0.6
0.4
0.2
0
# Ala 6 5 4 3 2 1 0
# Ile/Leu 0 1 2 3 4 5 6
AlogP -0.6 0.7 2 3.3 4.4 5.9 7

Figure 2-1 . **Octapeptide Split Pool Library** (A) Split-pool synthetic scheme for per-methylated Ala scan library. All library members contained a Tyr and Pro residue and the remaining 6 positions were either Ala, Leu, or Ile. (B) PAMPA permeability of Ala scan library for each unique composition. Di-Ala substitution was the dominantly permeable species in Leu and Ile containing library sub-pools.

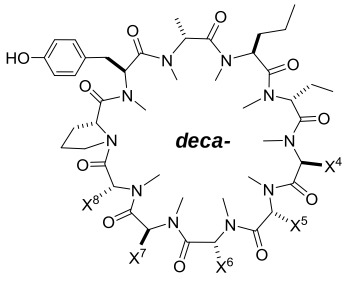
The parallel artificial membrane permeability assay (PAMPA) is a model of membrane permeability that is widely used as a cell-free system for measuring passive permeability and has been shown to correlate with in vivo oral bioavailability.[14,15] PAMPA allows for the assessment of a compound’s passive membrane permeability without the convoluting factors of paracellular diffusion and active efflux and transport mechanisms present in cell-based models. Previously we have shown that the assay can be multiplexed, using LC/MS to enable simultaneous quantification and deconvolution of complex mixtures.[16] Using this methodology, we discovered several novel cyclic hexapeptides with good to excellent membrane permeability. Both the Leu/Ala and Ile/Ala libraries were assayed using this multiplexed PAMPA methodology in order to evaluate the permeability of all library species. Only compounds comprised of two Ala and four Xaa residues (Xaa = Leu or Ile) showed appreciable permeability (Figure 2-1b). This result suggested that there is an optimal lipophilicity window in which significant permeability can be achieved within this system. We selected individual compounds from the library for synthesis and assayed their permeabilities in PAMPA as well as a cell-based permeability assay using a low-efflux MDCK cell line in which P-glycoprotein expression has been reduced (MDCKII-LE).[17] While the magnitudes of the permeabilities for the individual compounds were higher, the trends observed in the libraries were in good agreement with the trends observed in the assay of mixtures, demonstrating that the trend was not an artifact of the multiplexed in vitro assay and that the observed permeabilities correlated with composition irrespective of the specific sequences. (Figure 2-7).



MW 799-1009



MW 912-1123



MW 997-1207

X = Ala $-CH_3$

Abu $-CH_2CH_3$

Nva $-CH_2CH_2CH_3$

Nle $-CH_2CH_2CH_2CH_3$

Figure 2-2 **Lipophilicity Scanning Library**

2.4.2 Lipophilicity Scanning Library

In order to more thoroughly investigate the combined effect of lipophilicity and MW on permeability, we designed a combinatorial library of permethylated octa-, nona-, and decapeptides. As in the previous library, each member contained a Tyr and Pro residue, and the variable positions were populated with a combination of simple, straight chain aliphatic residues: Ala, aminobutyric acid (Abu), norvaline (Nva), and norleucine (Nle) (Figure 2-2). The library was designed to span a wide range of molecular weights (797-1237) and lipophilicity ($0 < \text{AlogP} < 8$), with compositional isomers that sampled lipophilicities in increments of one –CH2– group, corresponding to 0.4 log units. The octa-,nona-, and decapeptide libraries were analyzed separately. Permeabilities were determined for all –CH2– (i.e., AlogP) increments, which eluted separately on a reverse phase (C18) HPLC gradient (H2O→ACN), and were thus evaluated as groups of constitutional isomers (Figure 2-2a). As ring size increased, the overall permeabilities decreased, even among compounds in the same AlogP range. This general inverse correlation between MW and permeability has been observed previously for small sets of low-MW compounds.[18] However, we could not determine whether the magnitude of the MW penalty observed in our system was consistent with current models. In addition, we observed a bilinear relationship between permeability and lipophilicity. This decrease in apparent permeability beyond AlogP ~ 4.5 has been reported for other model systems and is consistent with knowledge-based models such as the Rule of 5 and others which place an upper limit on lipophilicity.[1,19,20] However, there is a lack of consensus as to whether this decrease in permeability at higher AlogP values is due to poor solubility or sequestration into the PAMPA membrane, and the phenomenon may depend on the particular system involved.[20,21]

In contrast to AlogP, which is a calculated descriptor based on octanol/water partition coefficients, experimental partition coefficients between water and hydrocarbon solvents such as 1,9-decadiene have been shown to provide a good correlation with experimental permeabilities in physics-based permeability models.[22] Octanol can form hydrogen bonds with solute and therefore provides a poor model for the desolvation of highly polar groups such as hydrogen bond donors.[23] However, in a series of homologous compounds that differ only by hydrocarbon chain lengths, a high correlation is observed between octanol/water and hydrocarbon/water partition coefficients.[23] In order to determine whether the apparent permeabilities observed in our system are concordant with current physics-based models, we set out to 1) determine their hydrocarbon/water partition coefficients experimentally and 2) account for the decrease in solubility at higher lipophilicities in order to determine their intrinsic permeabilities.

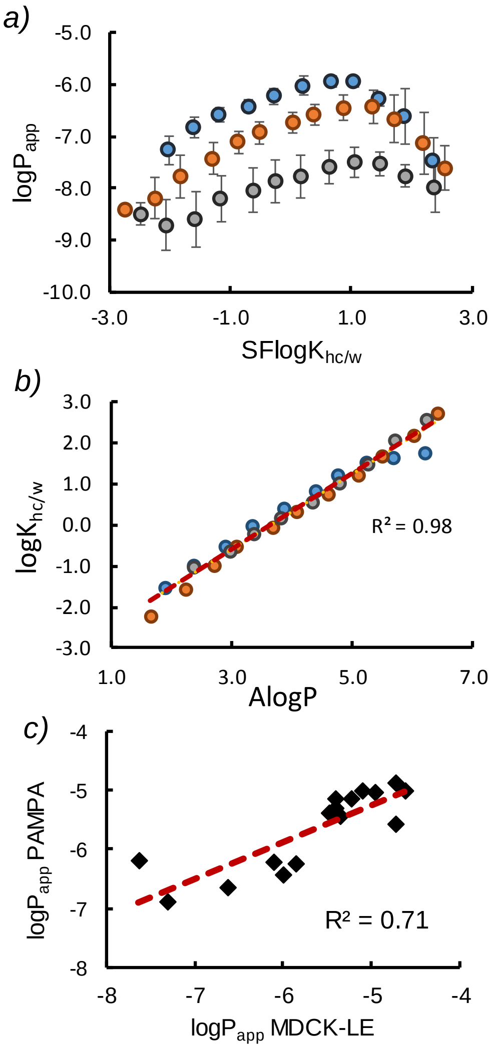


Figure 2-3 **Variable Ring Size Lipophilicity Scanning Library**. All plots include octa- (blue), nona- (orange), and decapeptides (grey). (A) logPapp vs SFlogK$_{hc/w}$ should be a linear relationship (eq 1), we observe deviation from linearity in each system at SFlogK$_{hc/w}$ > 1. (B) Experimental 1,9-decadiene partition coefficients (SFlogK$_{hc/w}$) vs. AlogP. The Linear regression (broken red line) was used to extrapolate values outside the detection limits. (C) logP$_{app}$ of PAMPA vs MDCK-LE for individual compounds. Linear shape of the data suggests that the PAMPA assay is correlated with the more biologically relevant cell based assay. The trends of individual compounds were in agreement with those observed in the mixtures (Table 2-1).

We selected 1,9-decadiene as the hydrocarbon solvent for shake-flask partition coefficient measurements (SFlog$K_{hc/w}$) since it has been reported to give the best correlation to experimental permeabilities.[8] We found a very good linear fit over ~5 log units between experimental 1,9-decadiene/buffer partition coefficients and AlogP (Figure 2-3b), suggesting that partitioning into hydrocarbon in this system is dominated by the aliphatic side chains and that conformational effects (e.g., by way of steric shielding) are either small or similar for all library members. A linear regression was used to extrapolate log$K_{hc/w}$ values for compounds that were below the detection limits of the instrument in either of the two phases. The MDCK-LE permeabilities for individual compounds were measured and compared to their permeabilities in the PAMPA system. The linearity of the data suggests that the PAMPA assay is correlated with the more biologically relevant cell based assay. The lower values in PAMPA vs. MDCK may be related to differences in membrane retention between the two systems. [24]

	Position												PAMPA	LE-MDCK	
cpd.	1	2	3	4	5	6	7	8	9	10	MW	AlogP	P_e (10^{-6} cm/s)	P_e (10^{-6} cm/s)	% recovery
8.1	Leu	Leu	Leu	Leu	Leu	Leu	Pro	Tyr			1037.4	6.8	0.033	<0.1	54.81
8.2	Leu	Ala	Leu	Leu	Leu	Ala	Pro	Tyr			953.3	4.3	19	2.76	26.95
8.3	Ala	Leu	Ala	Leu	Ala	Ala	Pro	Tyr			869.1	1.9	1	0.38	46.00
8.4	Ala	Ala	Ala	Ala	Ala	Ala	Pro	Tyr			785	-0.6	0.023	0.66	78.87
8.5	Ala	Ala	Leu	Leu	Leu	Leu	Pro	Tyr			953.3	4.3	8	10.31	77.84
8.6	Leu	Leu	Ala	Ala	Leu	Leu	Pro	Tyr			953.3	4.3	4	4.47	73.69
8.7	Leu	Leu	Leu	Leu	Ala	Ala	Pro	Tyr			953.3	4.3	6	7.30	73.46
8.8	Ala	Leu	Leu	Leu	Leu	Ala	Pro	Tyr			953.3	4.3	11	8.96	76.68
8.9	Leu	Ala	Leu	Leu	Ala	Leu	Pro	Tyr			953.3	4.3	19	12.72	92.13
8.10	Leu	Ala	Leu	Ala	Leu	Leu	Pro	Tyr			953.3	4.3	24	9.68	82.50
8.11	Abu	Ala	Ala	Nle	Ala	Ala	Pro	Tyr			841	1.5	0.24	0.22	74.01
8.12	Abu	Nle	Abu	Nle	Nle	Ala	Pro	Tyr			911	3.9	3.3	4.02	70.72
8.13	Abu	Nva	Nva	Nva	Nva	Nva	Pro	Tyr			939	4.8	4	7.29	73.92
8.14	Abu	Nle	Nle	Nle	Nle	Nle	Pro	Tyr			1009	7.1	<0.01	<0.1	89.71
9.1	Nva	Abu	Ala	Abu	Ala	Ala	Abu	Pro	Tyr		940	1.7	0.05	0.13	63.23
9.2	Nva	Abu	Ala	Nle	Ala	Nle	Ala	Pro	Tyr		996	3.7	0.8	0.61	77.88
9.3	Nva	Abu	Nva	Nva	Nva	Abu	Nva	Pro	Tyr		1038	5.1	4.5	3.56	67.95
9.4	Nva	Abu	Nle	Nle	Nle	Nle	Nle	Pro	Tyr		1121	7.9	<0.01	<0.1	88.10
10.1	Ala	Nva	Abu	Ala	Abu	Abu	Ala	Abu	Pro	Tyr	1039	1.9	0.12	<0.1	62.86
10.2	Ala	Nva	Abu	Nle	Abu	Ala	Nva	Abu	Pro	Tyr	1095	3.8	1.4	0.58	80.93
10.3	Ala	Nva	Abu	Nle	Abu	Nle	Nle	Abu	Pro	Tyr	1151	5.8	0.8	below LOD	below LOD

Table 2-1 Pure compound composition and assay data. Red residue names indicate D stereochemistry; black residue names indicate L stereochemistry. Methods for PAMPA and MDCKII-LE are described below. AlogP is the octanol/water partition coefficient as calculated by the Ghose and Crippen atomic contribution model. [17,25]

2.4.3 Solubility Adjustment via Filtration

The solvent-diffusion model predicts a linear relationship between log of intrinsic permeability ($\log P_o$) and membrane partition coefficient ($\log K_{mem}$) with unit slope, and this behavior has been observed for a variety of compounds, including small linear peptides.[26,27] At higher lipophilicities our cyclic peptides deviate substantially from this expected linear relationship, prompting us to test whether this decrease in apparent permeability could be attributed to a decrease in water solubility at the higher $\log K_{hc/w}$ values (Figure 2-3A).

Apparent permeability coefficients (P_{app}) are typically calculated based on the assumption that the concentration of compound added to the donor compartment (in practice usually ~10 µM) represents the solution concentration of the permeable species in the assay. In order to determine intrinsic permeabilities (P0), we used a filtration solubility assay to determine the concentration of soluble, monomeric (or small < 200-nm oligomeric) compound for each AlogP (or

SFlog$K_{hc/w}$) increment. Solutions were prepared using the same concentration and buffer conditions as those used in the permeability assay. The solutions were then filtered through 0.2-µ nylon filters so that only very small aggregates or monomeric species could pass through, and then quantitated by LC/MS using samples dissolved in MeOH as standards. The buffer-to-MeOH intensity ratios for each group of isomers thus provided a "solubility decrement" factor for adjusting Papp values based on true solubilities. Above log$K_{hc/w}$ ~ -1, solubilities fell off precipitously as a function of log$K_{hc/w}$ (Figure 2-4a-c).

Analogous to the adjustment commonly applied to permeants with ionizable groups, we adjusted the Papp value by the soluble fraction available (f_{sol}) using eq 3.[28] When we applied this solubility adjustment to Papp, we observed a linear relationship as predicted by models relating intrinsic permeability Po to $K_{hc/w}$ as indicated in Figure 2-4 by the regression lines (orange markers). While linearity is observed (consistent with a linear free energy relationship between partition coefficients and membrane transport rates), the slope is notably less than 1 (0.47,0.48,0.53 for the octa-,nona-,and deca- libraries respectively). This suggests that the change in P_o is not fully explained by the change in $K_{hc/w}$.

1/P_app =1/(P_o*f_sol)

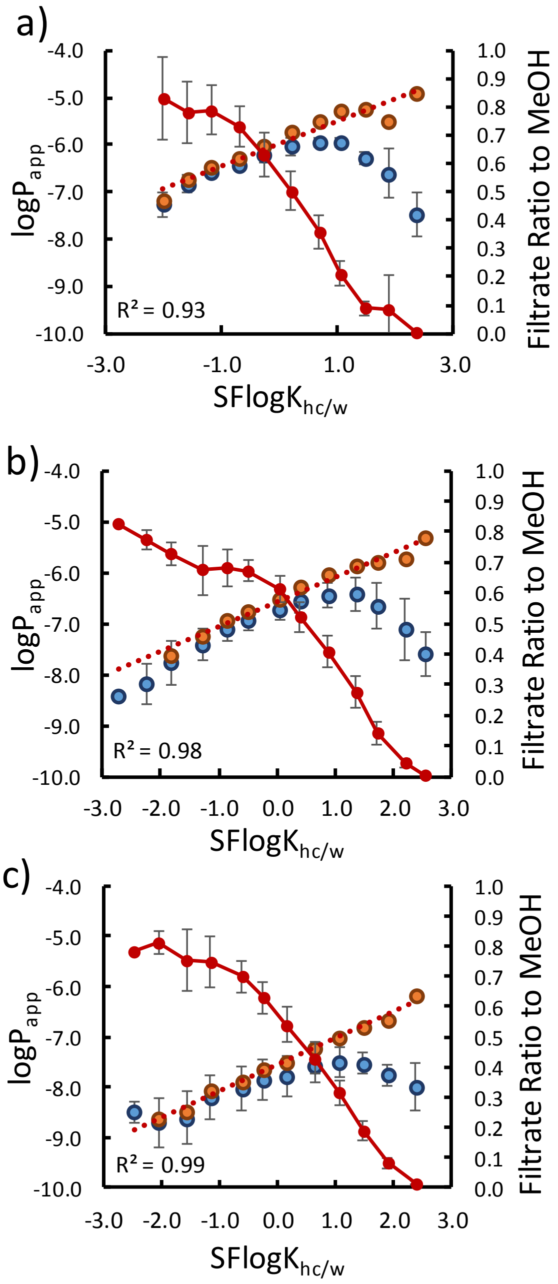


Figure 2-4 **Solubility Adjusted Intrinsic Permeabilities** Filtration solubility (unbroken red lines) and $\log P_{app}$ for the (a) octa-, (b) nona- , (c) decapeptides (Figure 2-2) (blue markers). Solubility-adjusted Po values were obtained by dividing the P_{app} by the relative solubility in MeOH for each point (orange markers). The linearity of Po and $K_{hc/w}$ is indicated by a linear fit of the solubility-adjusted values (red broken lines).

2.4.4 Non-Classical Diffusion of bRo5 Compounds

Despite the linear relationship between $\log K_{hc/w}$ and $\mathrm{Log}P_0$, the slope was significantly less than 1 and the intercepts decreased with increasing ring size, suggesting that there is an additional permeability penalty not accounted for with the solubility adjustment. The solubility-diffusion model has served as the basis for more modern treatments of predicting the role of size in membrane diffusion (e.g., the barrier-domain model).[7,8,11] In order to compare scaffolds of different lipophilicities and sizes, one can solve eq 1 for the apparent membrane diffusion term D_{mem} to get the following:

$$D_{mem}=(P_0*\partial)/K_{mem}$$

where Po is our solubility-adjusted P_{app}, K_{mem} is the $K_{hc/w}$ value from the 1,9-decadiene partition experiments, and ∂ is the thickness of the membrane in cm.[29] Using a reported thickness 0.125 cm, we plotted diffusions of compounds across the PAMPA membrane. It should be noted that solving for D_{mem} in the solubility diffusion equation is not necessarily evidence for a diffusion-mediated mechanism, since the size dependence may arise from a combinations of factors including partitioning penalties.[8] Plotting the apparent membrane diffusion of small molecules (using published experimental $\log K_{hc/w}$ and $\log P_o$ values[30]) versus their McGowan volume[31] reveals a roughly exponential relationship that follows the same form as Einstien-Stokes diffusion (Figure 2-5, black broken line). However, with our library compounds we find a steep drop-off beginning at a molecular volume of ~1000 Å3 (corresponding to a MW of ~1000Da) that deviates from the exponential approximation (Figure 2-5 blue points).

2.4.5 Extension to other scaffolds

To ensure that this trend was not particular to the macrocyclic scaffolds used in our study, we selected a variety of commercially available or synthetically tractable cyclic peptide natural products in this size range that we anticipated could be passively permeable. This diverse set of compounds allowed us to extend our observations beyond the permethylated model system to scaffolds that are potentially more relevant in a drug discovery context. Using the panel of assays described above, we determined their apparent D_{mem} values (Figure 2-5, orange markers). These data followed the same sharp downward trend above a volume of ~1000 Å3 (MW ~ 975) that was seen for the octa-, nona-, and decapeptides described above. The sharp decrease in $\log D_{mem}$ observed in this MW range is neither predicted by models with molecular volume dependences in the diffusion term (V^{-n}) nor in the partition term (e^{-Vn}) (nor a combination of both).[7,8,18]

Generalizing these results, the $\log K_{hc/w}$ required to achieve the generally accepted minimum threshold for permeability of Papp = 1×10^{-6} cm/s can be approximated at a given molecular volume (Figure 2-6). As volume increases, the $\log K_{hc/w}$ required to overcome the size penalty is so great that aqueous solubility begins to rapidly decline. In practice, this applies an upper limit on the size of compounds that are likely to be passively permeable and confines large molecular weight scaffolds to existing in a narrow lipophilicity window (green shaded area). These large compounds must have lipophilicities that are high enough to overcome the steep size dependence observed but low enough to maintain reasonable aqueous solubility. It should be noted that the relationship observed in our system between molecular size and intrinsic permeability is in agreement with the observation by Doak, et al. that very few orally delivered drugs exist beyond MW ~ 1000.[2]

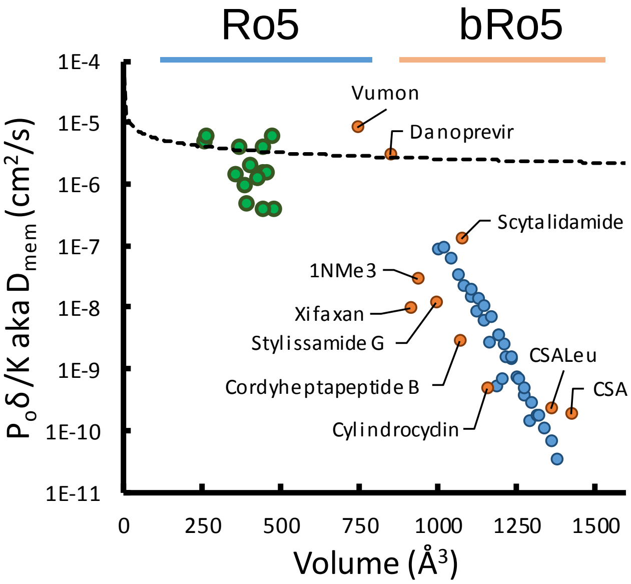


Figure 2-5 **Membrane diffusion vs molecular Volume** Diffusion was determined by eq 3 using experimental values of P_o and $K_{hc/w}$. The volume calculated from a 2D structure using a parameterized method. Small, Ro5 molecules (green points) roughly follow an exponential trend while the bRo5, lipophilicity scanning library members (Figure 2-2) (blue points) are subject to a much steeper size penalty. A selection of unrelated synthetic and natural products were subjected to the same analysis to ensure the trends are general (orange points).

2.5 Discussion and conclusions

In this work we set out to investigate the relationship between lipophilicity, size, and membrane permeability in a molecular weight range yet to be thoroughly interrogated. By using split-pool combinatorial syntheses and multiplexed assays we were able to evaluate thousands of compounds displaying a wide range of sizes and lipophilicities. The use of permethylated amide backbones and straight chain alkyl sidechains allowed us to probe compositional trends without convoluting factors arising from intramolecular hydrogen bonding and associated conformational effects. Physical models that describe permeability refer to the intrinsic permeability (P_o) as opposed to the experimentally observed apparent permeability (P_{app}), which can make comparison to experimental values difficult. Notably, the slopes of the relationships between $\log P_o$ and $\log K_{hc/w}$ were less than unity, indicating that the incremental increase in $\log K_{hc/w}$ for each additional –CH2- group corresponds to a smaller relative increase in $\log P_0$. Given that molecular size also increases with lipophilicity in these plots, the deviation from unity is consistent with a strong size dependence on permeability. By adjusting our experimental P_{app} for solubility we were able to compare P_0 values with other published datasets. Solving the general permeability equation (eq 1) for the size dependent diffusion term we discovered a steep size dependence for compounds with volumes in the bRo5 molecular weight range. We can only speculate about the mechanistic origin of this size dependence. It appears that there could be more than one rate limiting step for the permeation of these large hydrophobic macrocycles. It is clear from Figure 2-5 that Stokes-Einstein diffusion cannot account for the size dependence observed, but other more restrictive diffusion schemes akin to the diffusion of solutes through polymeric matrices might be at play as has been suggested by others.[6] Additionally, increased re-solvation penalties associated with departitioning from the membrane may cause membrane exit to become an additional rate limiting step. [32]

The model compounds presented here – permethylated species with only one hydrogen bond donor, the phenolic hydroxyl – are not necessarily representative of many bRo5 compounds that contain multiple hydrogen bond donors, where conformational effects may be much more important. Polar groups in the backbone of therapeutically relevant macrocycles have been shown to be important for protein target binding4 and the ability to adopt lipophilic conformations is required to mitigate the deleterious effect of these polar groups on permeability. The low dielectric conformation that CSA adopts in the membrane substantially decreases the desolvation cost of partitioning, and might minimize the effective radius, allowing CSA not only to “smuggle” polar backbone groups into the cell, but also to do so in a conformation that minimizes the cavity size needed to accommodate it.[8] These 3D effects of conformation are not captured in the simple 2D molecular volume used in plotting Dmem, which may be why CSA lies to the right of the regression line in Figure 2-5. Previous studies into the effect of size on permeability have used the 3D, radius of gyration descriptor; this descriptor requires the knowledge of the conformational landscape which is difficult to determine for large ring systems described in this paper and as such is not used.[3] However, the separate set of diverse compounds (Figure 2-5 orange points) containing hydrogen bond donors (and conformation-dependent properties) followed roughly the same sharp decline in membrane diffusion as our model compounds when using the thermodynamic solubility adjustments and equilibrium $\log K_{hc/w}$ values, suggesting that the observed size dependence is not particular to scaffolds that lack conformational effects. It should be noted that these compounds are also cyclic peptide species, and though their chemical nature is substantially more diverse than the library described herein, the generality of these trends has yet to be thoroughly investigated with non-peptidic compounds in this space.

Predicting precise values of P_o ab initio for a diverse set of molecules is a difficult task that requires knowledge of the molecules' conformational states as well as the composition of the membrane in question. However, the bulk physical properties that contribute to the intrinsic permeability are relatively simple: K_{mem} (approximated as $K_{hc/w}$) and size (approximated as ~MW). In contrast to intrinsic permeability (P_0), apparent permeability (P_{app}) depends on additional factors which must be considered, including the aqueous boundary layer, pKa, and aqueous solubility. Even the most permeable species in this range are well below the aqueous boundary threshold (around 10^{-4} cm/s), leaving solubility and pK_a as the main factors to consider aside from size and lipophilicity when assessing the potential apparent permeability of bRo5 molecules. To achieve K_{mem} values that are high enough to offset the intrinsic size penalty in bRo5 space, any ionizable groups are likely to be prohibitively deleterious. The fraction of available compound depends on both solubility and pK_a, and if a large compound is not fully soluble at the required lipophilicity to permeate, the addition of an ionizable group, while it may substantially increase the soluble fraction, will likely decrease K_{mem} prohibitively. However, in principle a balance could be achieved with the incorporation of solubilizing groups (e.g., morpholine or other tertiary amines) whose pK_a values are close to neutral. It has been shown in previous studies that conformational flexibility can help improve solubility of bRo5 macrocyclic compounds in some cases, though the exact mechanism by which the “chameleonicity” gives rise to these desirable properties is still poorly understood.[33–35] Poor aqueous solubility can also be alleviated (to a point) by modern pharmaceutical formulation chemistry. The use of nm-scale dispersions and other solubility-enhancing delivery systems such as emulsifiers and liposomes can greatly improve the oral bioavailability of highly lipophilic drugs.[36]

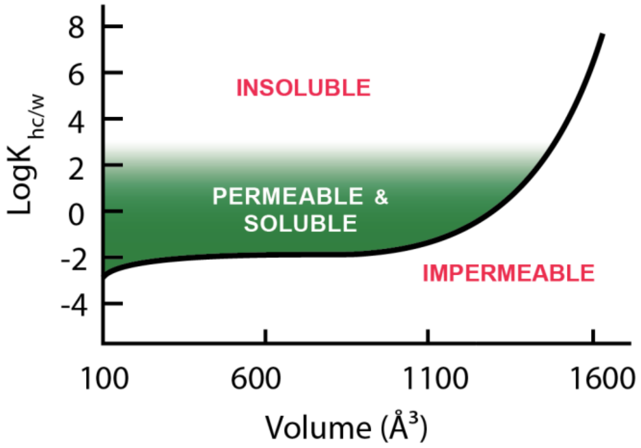


Figure 2-6 **Permeable Space** A crude approximation of the minimum logK$_{hc/w}$ required to maintain a permbeability of 10^{-6}cm/s (black line), using D$_{mem}$ values from the Einstein-Stokes for Ro5 compounds (Vol < 850 Å^3) and a 3rd order polynomial fit of the relationship between LogK$_{hc/w}$ and molecular volume for bRo5 compounds (Vol > 850 Å^3). High logK$_{hc/w}$ will likely result in poor aqueous solubility limiting the Papp. This in turn applies an upper limit to the size of a molecule for passive membrane permeability (green shaded region) since the necessary logK$_{hc/w}$ required to overcome the size penalty would render the compound insoluble.

Cyclic peptides and peptide-containing macrocycles have received a lot of attention due to their synthetic tractability and a renewed enthusiasm for their potential to exhibit drug-like pharmacokinetic properties.[37] While a variety of combinatorial strategies (DNA-encoded, mRNA display, and phage display libraries etc.) have been developed that can yield macrocycles with potencies that rival those of antibodies, they tend to produce molecules with multiple polar side chains. While these residues can be critical for high affinity, they can virtually eliminate the possibility for passive cell permeability.[37,38] There are other modalities that can be employed to effect cell permeability which rely on various non-passive or pseudo-passive mechanisms (e.g., micropintocytosis, membrane disruption, possibly other uncharacterized transport pathways).[39–41] For example, charge-mediated transport mechanisms employed by incorporating a series of cationic residues have been demonstrated to overcome the passive permeability problems by bypassing the mechanism altogether.[42] Yet, although these strategies have been effective in cell culture and even in animals, there are still no compounds in this class that have progressed through clinical trials, perhaps due to the complexities associated with cation-mediated transport in the varied cell types found in biological tissues. Despite the severe constraints placed on molecular volume by conventional, passive membrane diffusion, the permeable chemical space between MW 500 and 1000 remains virtually untapped. Nature has provided us with a diverse set of examples that manage to walk the line between permeability and solubility. In addition to CSA, scaffolds with very different structures such as cylindrocyclin and patellamide C have drug-like passive permeabilities.[7] Therefore, while the challenge of achieving both permeability and solubility increases sharply beyond the Rule of 5, so does the number of synthetically accessible scaffolds. Mining this chemical space for compounds of therapeutic value will be the challenge of this field for decades to come.

2.6 Experimental

2.6.1 General

All chemicals were commercially available and used without further purification. For reverse-phase column chromatography, a SNAP C18 cartridge (30 g, Biotage) was employed. NMR spectra were recorded in CDCl3 , Pyridien d5, or DMSO d6 (depending on solubility) on a Varian

500 MHz NMR instrument with Unity Plus console and 5 mm broad-band probe at 25 °C. Chemical shifts were referenced to residual solvent proton signals. The purities of individual compounds were tested by HPLC (Waters 1525) with an attached mass spectrometer (Micromass ZQ, waters) and PDA detector (Waters 2998) through a 3.5 μm C18 column (XBridge BEH C18 4.6 × 50 mm). The mixture of water (0.1% formic acid) and ACN (0.1% formic acid) was used as an eluent, of which ACN percentage was increased stepwise (0–2 min, 20%; 2–10 min, linear gradient from 20 to 100%; 10–12 min, 100%) with a flow rate of 1.2 mL/min. Pure, diverse compounds were analyzed on the same instrument and solvent system using a Thermo Accucore C18 column (50x4.6mm) with 2.6μm packing material. The ACN was again increased stepwise (0–1 min, 20%; 2–10 min, linear gradient from 20 to 100%; 2–4 min, 100%) with a flow rate of 1.9 mL/min. The absorbance at 220 nm wavelength was used to calculate the purity and all individual compounds were at least 95% pure. LC-MS analyses for PAMPA were performed on a UHPLC (UltiMate 3000, Dionex) with attached mass spectrometer (Orbitrap Velos Pro, Thermo Scientific) through a 1.9 μm C18 column (Hypersil GOLD 30 × 2.1 mm, Thermo Scientific). The mixture of water (0.1% formic acid) and ACN (0.1% formic acid) was used as an eluent, using a short trap and elute method ACN was held at an initial value of 10% for 1min followed by a rapid linear gradient (30sec) to 95% which was held for 1.75min before returning to starting column conditions with a flow rate of 1 mL/min.

2.6.2 Synthesis of Pure Cyclic Permethylated Peptides

2.6.2.1 *Loading of 2-Chlorotrityl resin*

For resin bound for on-resin cyclization, Fmoc-*O*-allyl-Tyr (0.954 g, 0.43 mmol) was added to a flame dried round bottom flask and dried in a vacuum desiccator with phosphorous pentoxide overnight. 2 g of cesium carbonate was added to the flask and it was purged with N_2. 50mL of dry DCM was cannula transferred into the flask followed by 2.5 mL of DIPEA transferred via syringe. After sonication for 10 min, 5 g of 2-chlorotrityl resin was added under a stream of N_2 and allowed to shake for 4 h. The resin was capped with a 15 mL solution of 1:2:17 MeOH:DIPEA:DMF (3 x 15 min). The resin was washed with DMF (3 x 15 mL) followed by DCM (3 x 15 mL). The loading value was calculated by mass increase of dried, loaded resin.

2.6.2.2 Automated Peptide Synthesis

The linear peptide sequences were synthesized and cyclized using an automated peptide synthesizer (Prelude, Protein Technologies). Fmoc deprotections were carried out with 2% 1,8-diazabicycloundec-7-ene (DBU) and 2% piperidine in DMF for 15 min. Couplings were performed using 4 eq Fmoc-protected amino acid, 3.8 eq O-(azabenzotriazol-1-yl)- N,N,N',N'-tetramethyluronium hexafluorophosphate (HATU) and 6 eq N,N-diisopropylethylamine (DIPEA) in N,N-dimethylformamide (DMF, 0.1 M with respect to amino acid) for 1 h. After each coupling and deprotection step, the resin was washed with DMF (3×), dichloromethane (DCM) (3×) and DMF (3×).

2.6.2.3 On-Resin O-allyl and Fmoc deprotection

After the addition of the final residue, deallylation and final Fmoc removal were performed simultaneously with a solution of 1 eq $Pd(Ph_3P)_4$ in THF containing 10% (v/v) piperidine for 3 h. A chelating wash was performed to remove traces of palladium using 5% (w/v) sodium diethyldithiocarbamate andin 5% (v/v) DIPEA in DMF, followed by the previously described DMF-DCM-DMF resin wash sequence.

2.6.2.4 On-Resin Cyclization

Cyclization was performed with 3 eq (Benzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (PyBOP), 3 eq 1-hydroxy-7-azabenzotriazole (HOAt), and 6 eq DIPEA in DMF for 3 h, followed by resin washing with five final DCM washes to remove residual DMF.

2.6.2.5 On-Resin Permethylation

Following cyclization, cyclic peptides were subjected to an on-resin permethylation. To the resin-bound peptides was added lithium *tert*-butoxide (3 equiv / per theoretical eq NH) in DMSO / THF (1:1) using enough volume to fully cover all resin and dissolve the Li *t*-BuOH resulting in an approximate concentration of ~1M . The resin was agitated for 15 min, after which methyl iodide was added. After an additional 15 min agitation, the resin was washed exhaustively with water, followed by washings with DMF (3x), MeOH (3x), and DCM (3x).

2.6.2.6 Peptide Cleavage

Peptides were cleaved with a 5% (v/v) trifluoroacetic acid (TFA) in DCM. The filtrate was concentrated and the crude residue was then purified by reverse phase (Biotage SNAP C18) automated flash chromatography (ACN:H20) and lyophilized to afford white amorphous solids.

2.6.3 Synthesis of Cyclic Peptide Libraries

The permethylated libraries were synthesized using the “split-pool” strategy starting with the allyl ester of N-fluorenylmethyloxycarbonyl (Fmoc)-protected tyrosine linked via the phenolic hydroxyl group to 2-chlorotrityl polystyrene resin (0.4 mmol/g loading value) as described above.

2.6.3.1 Manual Fmoc Deprotection

After washing the resin with the previously described DMF-DCM-DMF sequence, Fmoc deprotections were carried out with 2% 1,8-diazabicycloundec-7-ene (DBU) and 2% piperidine in DMF. A volume of this deprotection solution approximately twice that of the resin was added to the SPPS tube and agitated on a rotary shaker for 15 min. The solution was then drained and without washing the resin another round of deprotection was added and shaken for an additional 15 min. The resin was then washed using the standard sequence.

2.6.3.2 Manual Coupling Procedure:

For all couplings, a coupling solution was prepared by dissolving 4 eq of the amino acid and 3.8 eq of HATU in DMF (0.2 M with respect to amino acid). DIPEA (6 eq) was added and the solution was allowed to pre-activate for 30 min, then added to the deprotected, washed resin. The solution was agitated for 45 min on a rotary shaker.

2.6.3.3 Split-Pool Synthesis:

1 g of resin loaded with Fmoc-L-Tyr-OAllyl was coupled with Fmoc-D-proline as outlined above, then segregated into four separate polypropylene SPPS vessels and deprotected for 30 min. The resin was washed with DMF (3x), MeOH (3x), DCM (3x), and DMF (3x), then coupled with either N-Fmoc-D-norleucine, N-Fmoc-D-norvaline, N-Fmoc-D-aminobutyric acid, or N-Fmoc-D-alanine as outlined above. The resin was washed, pooled into a single vessel, thoroughly mixed, and the split-pool process repeated 4 more times (following the same stereochemistry as compound 8.1-8.4) to generate four sub-libraries of 256 linear heptapeptides. To keep the library size relatively small, all sub-libraries were kept separate and then coupled with N-Fmoc-D-aminobutyric acid.

After the final coupling, each portion of resin was washed and subjected to the simultaneous N/C-terminus deprotection. After deprotection, the resin was washed, cyclized as previously described, washed again, and subjected to the described on-resin permethylation conditions. The resin-bound sub-libraries were cleaved with a 5% solution of TFA in DCM. The collected filtrates were concentrated and the residue redissolved in DMSO at a concentration of 100 mg/mL to give stock solutions of four 256-member sublibraries of permethylated octapeptides.

Libraries of permethylated nonapeptides were generated in a similar manner except after coupling of the final Abu residue, an additional coupling of N-Fmoc-L-norvaline was performed to give four sub-libraries of 256 linear nonapeptides, which were cyclized, permethylated, cleaved, and reconstituted in DMSO as described above using an aliquot of the previously loaded resin.

Libraries of permethylated decapeptides were generated in a similar manner except after the coupling of the final Nva residue (in the preparation of nonapeptides), an additional coupling of N-Fmoc-D-alanine was performed to give four sub-libraries of 256 linear decapeptides, which were cyclized, permethylated, cleaved, and reconstituted in DMSO as described above.

2.6.4 Synthesis of natural products and unrelated pure compounds

Peptides were prepared with the following procedures; Stylissamide G, Cordyheptapeptide B, Scytallidamide B, and Cylindrocyclin were prepared using microwave assisted peptide couplings with installation of N-methyl amino acids followed by solution phase cyclization with COMU. 1NMe3 was prepared as previously described.[43] Cyclosporine A (BMT-to- Leu) was suynthesized using microwave assisted peptide couplings with installation of the appropriate Fmoc-protected N-Me amino acids followed by solution phase cyclization with COMU, with a modified permethylation for the first 4 amino acids. In the case of CSA starting with the Leucine that replaced the BMT residue as the first amino acid on resin the first 3 couplings were done with unmethylated amino acids. The Fmoc was removed, replaced with a Nosyl protecting group and the tetrapeptide was subjected to on resin permethylation. The remaining couplings and amino acids were added as described below.

2.6.4.1 Loading of 2-Chlorotrityl resin

Fmoc-Xaa (10 mmol) was added to a flame dried round bottom flask and dried in a vacuum dessicator with phosphorous pentoxide overnight. 50mL of dry DCM was cannula transferred into the flask followed by 2.5 mL of DIPEA transferred via syringe. After sonication for 10 min, 5 g of 2-chlorotrityl resin was added under a stream of N_2 and allowed to shake for 4 h. Resin was capped with a 15 mL solution of 1:2:17 MeOH:DIPEA:DMF (3 x 15 min). Resin was washed with DMF (3 x 15 mL) followed by DCM (3 x 15 mL). Loading value was calculated by mass increase of dried, loaded resin.

2.6.4.2 Microwave Reaction Conditions

Microwave assisted reactions were done in a CEM Discover microwave reactor with an open reaction vessel. Reactions were heated to 50°C with 12 watts of power while temperature was monitored with an IR probe from within the reaction vessel.

2.6.4.3 Amino Acid Coupling under Microwave Condtions

A solution of 4 eq of Fmoc-Xaa, 4 eq of HBTU, and 4 eq of HOAT, and 6 eq of DIPEA in DMF was allowed to pre-react for 5 min. This solution was added to the deprotected peptide on-resin and allowed to react for 10 min at 50°C with microwave heating. The solution was drained and resin was washed 3 times with DMF (3 x 3 mL) and DCM (3 x 3 mL). The reaction was monitored by LCMS and repeated until starting material was no longer observed.

2.6.4.4 Amino Acid Coupling onto N-Methylated Amino Acids under Microwave Conditions

A solution of 4 eq of Fmoc-Xaa, 4 eq of HATU, and 6 eq of DIPEA in DMF was allowed to pre-react for 5 min. This solution was added to the deprotected peptide on-resin and allowed to react for 30 min at 50°C under microwave heating. The solution was drained and resin was washed

DMF (3 x 3 mL) and DCM (3 x 3 mL). The reaction was monitored by LCMS and repeated until starting material was no longer observed.

2.6.4.5 Removal of the N-Fmoc Protecting Group Under Microwave Conditions

A solution of 2% piperidine and 2% DBU in DMF was added to the resin. The reaction was allowed to react for 5 min at 50°C under microwave heating then drained. The resin was washed with DMF (3 x 3 mL) and DCM (3 x 3 mL).

2.6.4.6 Peptide Cleavage

Complete linear peptides were cleaved off resin in 5 resin volumes of 2.5% TFA in DCM for 4 min three times with a DCM wash equivalent to 5 resin volumes in between each step. Solvent was removed under N_2 followed by dissolution in acetone or DCM and evaporation under reduced pressure. Residual TFA was removed *in vacuo* overnight.

2.6.4.7 Cyclization with COMU

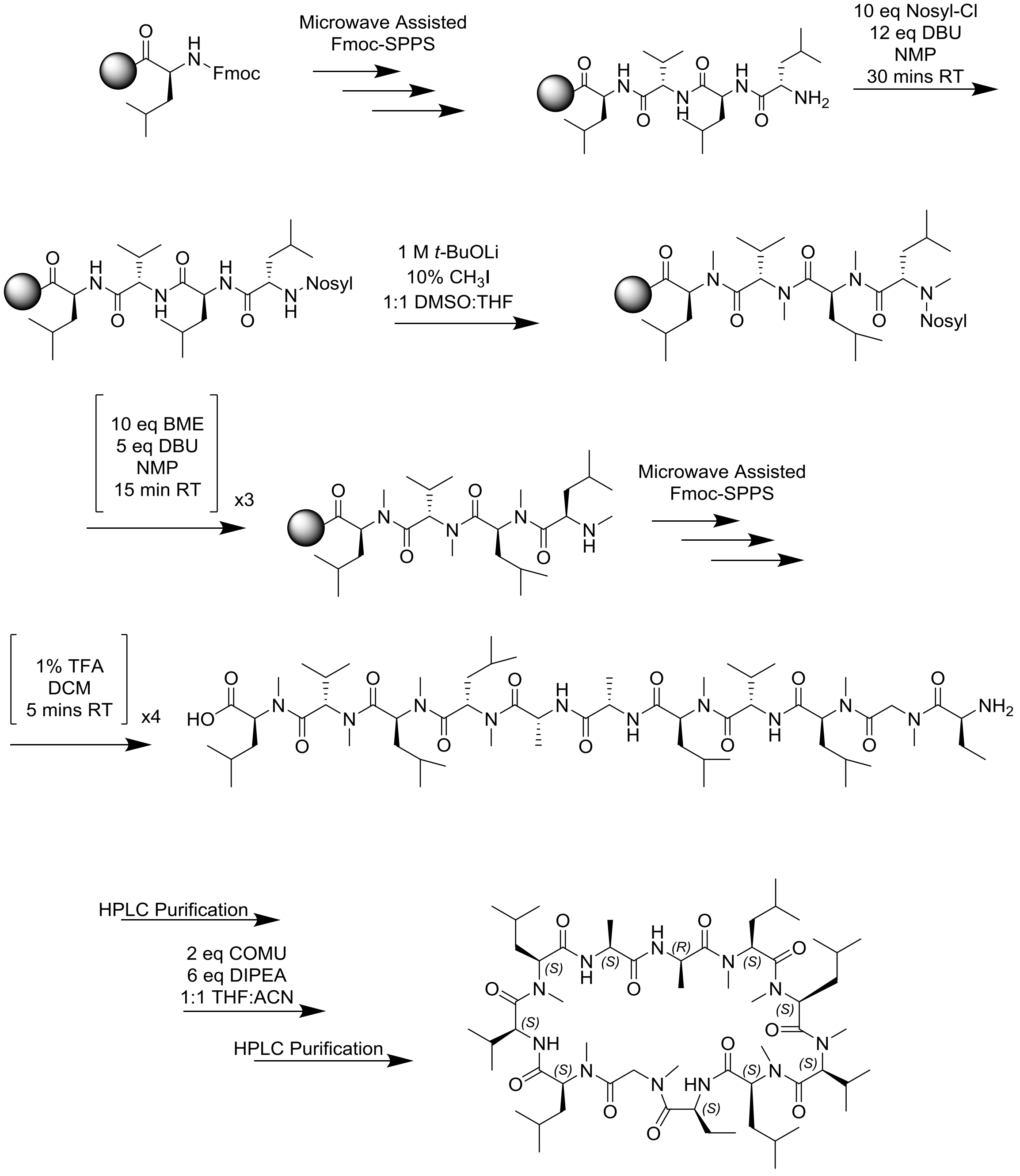
Linear peptides were dissolved in 20 mL of dry acetonitrile with 4 eq of DIPEA and added dropwise to a solution of (final concentration 1 mg crude peptide per mL) 1:1 THF/ACN containing 2 eq of COMU. Reactions were stirred from 0.5 to 24 h until complete cyclization was achieved as monitored by LCMS. The reaction was reduced *in vacuo* for purification via HPLC.

2.6.4.8 Purification of Peptides

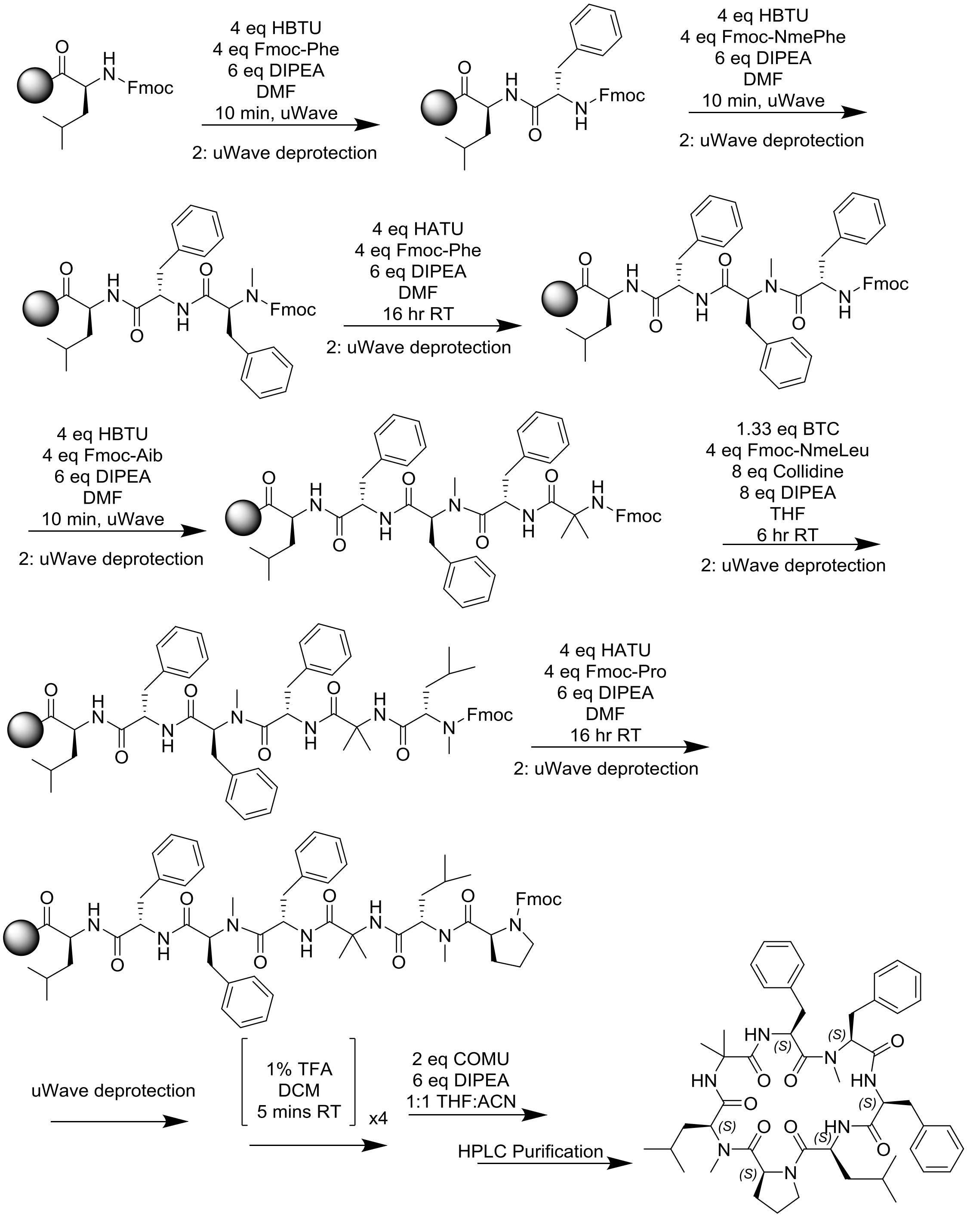
COMU by-products were removed after solution phase cyclization on a Biotage Isolera Prime system equipped with a KP-C18-HS 12g column eluting with H2O/Acetonitrile modified with 0.1% TFA. Peptides were further purified when necessary on a Waters mass-directed prep system equipped with an XBridge BEH130 5µm 19x150 C18 column eluting with H_2O/Acetonitrile modified with 0.1% formic acid.

2.6.4.9 Synthesis of Cyclosporine A BMT to Leucine

Sequence: Leu - Val - Leu - Leu - DAla - LAla - LNmeLeu - LVal - LNmeLeu - Sarc - LAbu

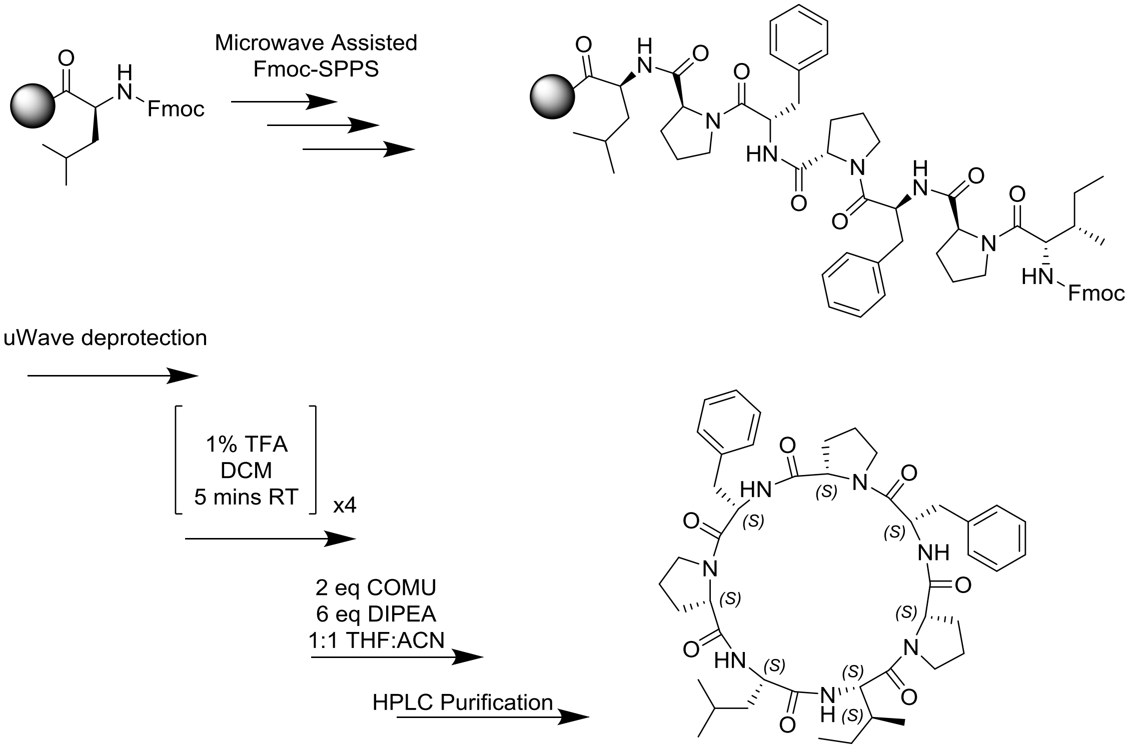

Fmoc
Microwave Assisted
Fmoc-SPPS
NH_2
10 eq Nosyl-Cl
12 eq DBU
NMP
30 mins RT
Nosyl
1 M *t*-BuOLi
10% CH_3I
1:1 DMSO:THF
Nosyl
10 eq BME
5 eq DBU
NMP
15 min RT
x3
Microwave Assisted
Fmoc-SPPS
1% TFA
DCM
5 mins RT
x4
HO
NH_2
HPLC Purification
2 eq COMU
6 eq DIPEA
1:1 THF:ACN
HPLC Purification
(S)
(R)

2.6.4.10 Synthesis of Scytallidamide B

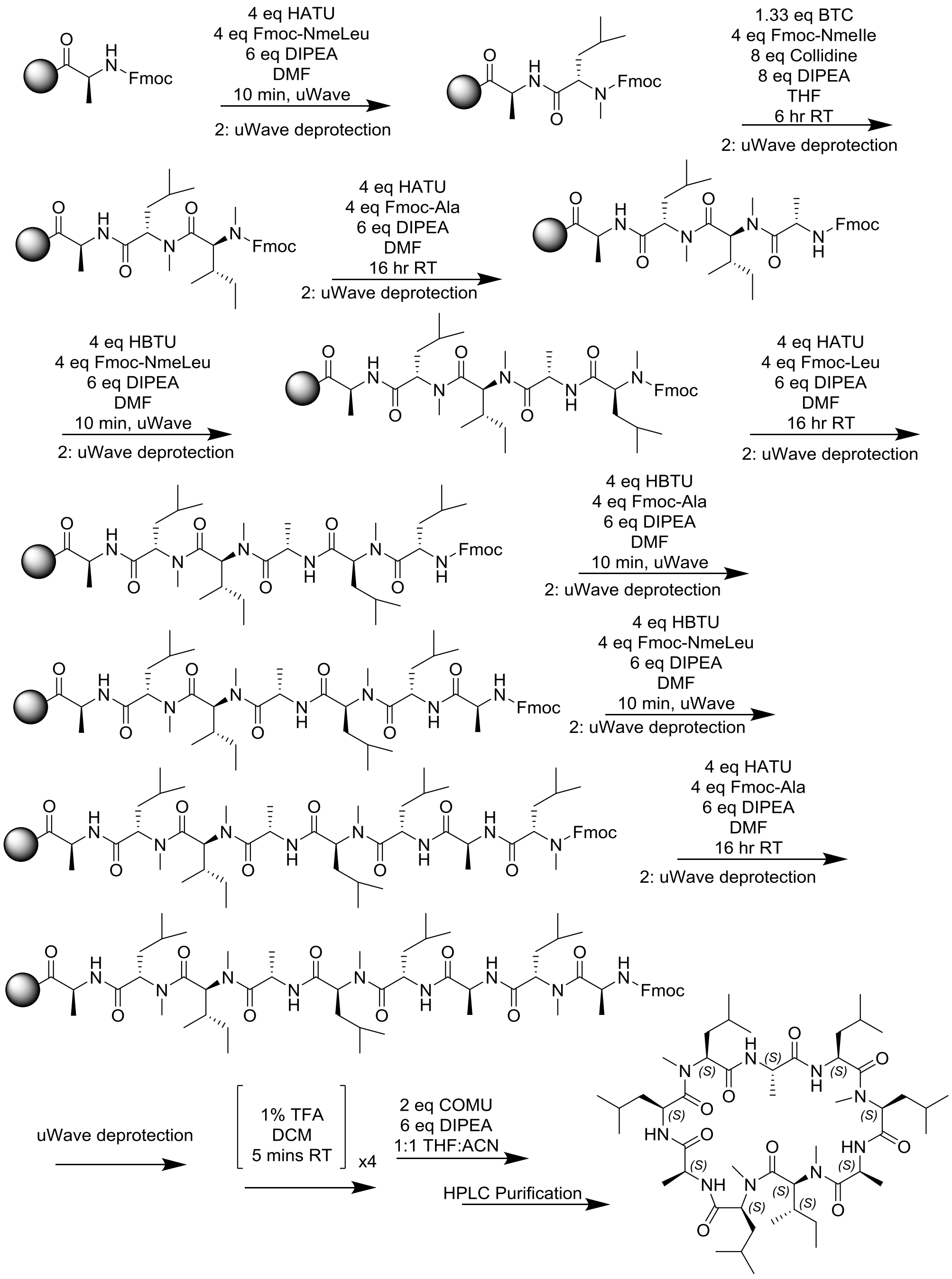


4 eq HBTU
4 eq Fmoc-Phe
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
4 eq HBTU
4 eq Fmoc-NmePhe
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
4 eq HATU
4 eq Fmoc-Phe
6 eq DIPEA
DMF
16 hr RT
2: uWave deprotection
4 eq HBTU
4 eq Fmoc-Aib
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
1.33 eq BTC
4 eq Fmoc-NmeLeu
8 eq Collidine
8 eq DIPEA
THF
6 hr RT
2: uWave deprotection
4 eq HATU
4 eq Fmoc-Pro
6 eq DIPEA
DMF
16 hr RT
2: uWave deprotection
uWave deprotection
1% TFA
DCM
5 mins RT
x4
2 eq COMU
6 eq DIPEA
1:1 THF:ACN
HPLC Purification
Fmoc
(S)

2.6.4.11 *Synthesis of Styllisamide G*


Fmoc
Microwave Assisted
Fmoc-SPPS
Fmoc
uWave deprotection
1% TFA
DCM
5 mins RT
x4
2 eq COMU
6 eq DIPEA
1:1 THF:ACN
HPLC Purification
(S)

2.6.4.12 Synthesis of Cylindrocyclin A


Fmoc
4 eq HATU
4 eq Fmoc-NmeLeu
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
1.33 eq BTC
4 eq Fmoc-NmeIle
8 eq Collidine
8 eq DIPEA
THF
6 hr RT
2: uWave deprotection
4 eq HATU
4 eq Fmoc-Ala
6 eq DIPEA
DMF
16 hr RT
2: uWave deprotection
4 eq HBTU
4 eq Fmoc-NmeLeu
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
4 eq HATU
4 eq Fmoc-Leu
6 eq DIPEA
DMF
16 hr RT
2: uWave deprotection
4 eq HBTU
4 eq Fmoc-Ala
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
4 eq HBTU
4 eq Fmoc-NmeLeu
6 eq DIPEA
DMF
10 min, uWave
2: uWave deprotection
4 eq HATU
4 eq Fmoc-Ala
6 eq DIPEA
DMF
16 hr RT
2: uWave deprotection
uWave deprotection
1% TFA
DCM
5 mins RT
x4
2 eq COMU
6 eq DIPEA
1:1 THF:ACN
HPLC Purification
(S)

2.6.5 General Analytical Procedures

2.6.5.1 *HPLC-MS Analysis for reaction and purity assessment:*

Purity of individual peptides was performed using HPLC-MS(Waters AutoPure system equipped with a Micromass ZQ200 mass spectrometer). ~1mg/mL samples were prepared in MeOH and 10µL was injected and run using a gradient of H20 → ACN on a Thermo AccuCore C18 column (4.6x50mm). Starting gradient conditions were varied depending on the lipophilicity of the analyte to elute the compound in the middle of the run. Since the compounds ware being assayed for permeability, absolute purity was not crucial as long as the presence of isobaric contaminants were not present that would convolute the assay analysis.

2.6.5.2 *Permeability Analysis of Cyclic Peptide Mixtures by PAMPA:*

A 96-well donor plate with 0.45 µm hydrophobic Immobilon-P membrane supports (Millipore) and a 96-well Teflon acceptor plate were used in the PAMPA permeability test. The acceptor plate was prepared by adding 300 µL of 5% DMSO in pH = 7.4 phosphate-buffered saline (PBS) to each well. Donor well solutions of the cyclic peptide libraries were prepared by diluting 10 µL of the DMSO stock solutions prepared above to a final volume of 200 µL with PBS (pH 7.4). The suspensions were centrifuged to remove any insoluble material. A 1% (w/v) solution of lecithin in dodecane was prepared and sonicated before use. 5 µL of the dodecane / lecithin solution was carefully applied to the membrane supports in the wells of the donor plate, with care being taken to not touch the pipet tip to the membrane. Without allowing this solution to evaporate, 150 µL of the peptide solutions were added to the donor wells. The donor plate was then placed on top of the acceptor plate so that the artificial membrane was in contact with the buffer solution below. A lid was placed on the donor well, and the system was covered with a glass evaporating dish and left overnight (~18 h) at room temperature. A wet paper towel was placed on the inside of the chamber to prevent evaporation.

After ~18hr (exact time recorded and used for subsequent calculations) the donor and acceptor plate were separated and 100µL of each well (donor and acceptor) were transferred to a 96 well plate for quantification. The plate was immediately sealed with a pierceable plate cover to prevent any sample evaporation. These solutions were analyzed by UPLC on a C18 Thermo HyperSil column (2.1x30mm, 3µm) with an accurate mass MS detector (Thermo Scientific Orbitrap VelosPro) using +/- 0.02 AMU windows for integration. Permeability (%T) was quantified as the ratio of analyte areas in the acceptor well divided by a theoretical equilibrium ratio based on amounts of combined analyte found in the donor and acceptor wells as follows:

$$\%T = \frac{R_A}{\left(\frac{R_A V_A + R_D V_D}{V_A + V_D}\right)}$$

Where RA and RD are the integrations analyte in the acceptor and donor wells, respectively, and VA and VD are the volumes of the acceptor (300 µL) and donor (150 µL) wells, respectively.

Permeation rates (P_{app}) were calculated from %T by the following equations:

$$P_{app} = -\frac{V_A V_D}{V_A + V_D} \times \frac{\ln(1 - \%T)}{At}$$

Where *A* is the active surface area of the filter support (0.24 cm^2), and *t* is time of the incubation period in seconds.

2.6.5.3 Permeability Analysis of Pure Cyclic Peptides by PAMPA

Pure compounds were analyzed via PAMPA as outlined above for mixtures. Donor-well solutions were prepared at 10µM.

2.6.5.4 Permeability Analysis of Pure Cyclic Peptides by RRCK:

Cell permeability was determined using RRCK cells (Pfizer, Inc. Groton, CT)[17]. RRCK cells were generated in house as a subclone of Madin-Darby Canine Kidney wild-type (MDCK-WT) cells that displayed low expression of endogenous P-glycoprotein (~ 1–2% of MDCK-WT cells, based on mRNA level). Cells were cultured in minimal essential medium α with supplements and passaged when 70–80% confluent. Cell monolayer flux studies were conducted five days after seeding in 24-well transwell inserts, 1.0 μm pore size, (Becton Dickinson, Cowley, UK) at 4.2×10^4 cells/cm^2. Donor and acceptor solutions were prepared from HBSS containing HEPES at 20 mM, pH 7.4. Stock solutions of test compounds, prepared at 10 mM in DMSO, were used to prepare donor solutions of 2 μM compound in 0.05% (v/v) DMSO. Apparent permeability (P_{app}) was determined in apical to basolateral (AB) direction in triplicate by incubating with compound for 2 h at 37 °C. Samples of the medium were analyzed by liquid-chromatography tandem mass spectrometry (LC-MS/MS). P_{app} values were calculated according to the equation $P_{app} = (Q/t) \times 1/C_0 \times 1/A$, where Q is the sampled concentration in the acceptor compartment, t is the incubation time, C_0 is the initial concentration in the donor compartment and A is the area of the filter of the transwell plate.

2.6.5.5 Differences in Pure vs Mixture Values:

It should be noted that the masses (m/z) used for analysis of the mixtures represent the average for all possible isobaric combinations. The pure compounds however, represent one possible combination. As such they are not directly comparable to the averaged mixtures but the overall trends still hold as demonstrated by Figure 2-7.

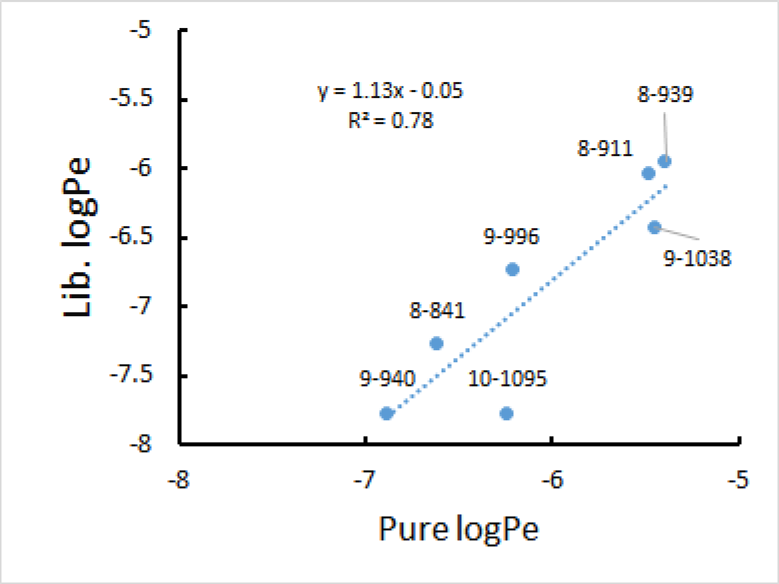


Figure 2-7 **logSol vs Log $K_{hc/w}$**

2.6.5.6 Filtration Solubility Assay Procedure:

Relative thermodynamic solubilities were assayed by comparing a filtered aqueous solution to a solution prepared from the same stock solution dissolved in MeOH. Aqueous solutions of each sublibrary were prepared the night before by mixing 10 µL of DMSO stock into 1 mL of PBS (pH 7.4) and allowing to incubate for 16 h. Solutions were then centrifuged at 16,000 x g for 5 min. The supernatant was then taken up in a 1 mL syringe and filtered using a 7 mm^2 nylon filter with a pore size of 0.2 µ (Nalgene cat. no. 179-0020). The first half of the filtrate was discarded so as to passivate the filter with analyte. The remainder of the solution was dispensed in to HPLC vial for analysis. MeOH solutions of each sublibrary were prepared immediately before analysis by mixing 10 µL of DMSO stock into 1 mL of HPLC grade MeOH. These solutions were analyzed by UPLC on a C18 Thermo HyperSil column (2.1x30mm, 3µM) with an accurate mass MS detector (Thermo Scientific Orbitrap VelosPro) using +/- 0.02 AMU windows for integration.

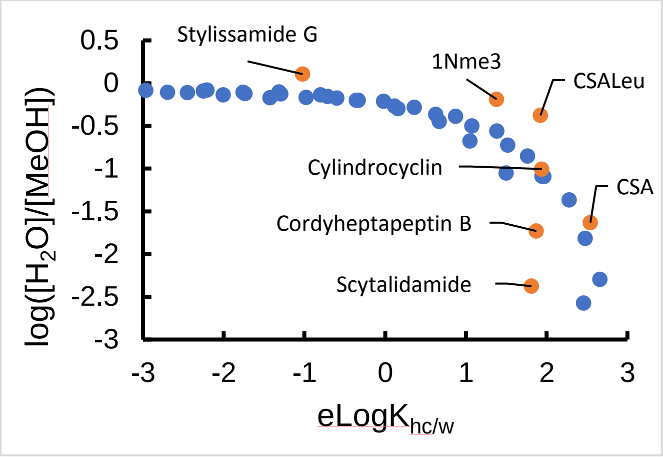


Figure 2-8 **log(f_{sol})**Figure 2-8 **vs SF Log $K_{hc/w}$**

2.6.5.7 Shake Flask Log$K_{hc/w}$ Analysis:

Reagent grade 1,9-decadiene was purchased from Sigma Aldrich and used as is. Prior to use for the assay, 10 mL 1,9-decadiene was saturated with H_2O by shaking with an equal volume of PBS (pH 7.4) and allowing the emulsion to completely separate. 1 µL of a 100mg/mL DMSO stock in the case of the library, or 1 µL of a 10 mM DMSO stock of the pure compounds was added to a 1.5 mL centrifuge tube and then 500 µL of each solvent (1,9-decadiene and H_2O) was added and agitated by vortexing (1 min) and sonicating (30 min). The emulsions were then spun down in a centrifuge for 5 min at 16,000 x g and allowed to sit overnight. 100 µL of each phase was carefully removed with special care taken not to allow contamination from the pipette tip. These solutions were then evaporated overnight under a gentle stream of N_2 and re-suspended in 100 µL of MeOH for quantification via UPLC/MS. These solutions were analyzed by UPLC on a C18 Thermo HyperSil column (2.1 x 30 mm, 3 µ) with an accurate mass MS detector (Thermo

Scientific Orbitrap VelosPro) using +/- 0.02 AMU windows for integration. The intensities were then divided (1,9-decadiene/H_2O) to afford the $K_{hc/w}$ value.

2.6.5.8 Aqueous Boundary Layer PAMPA

In order to determine the effect of the aqueous boundary layer (ABL) in PAMPA with our compounds the assay was run without the dodecane/lipid mixture. Instead, 5 µL of DMSO was used to wet the filter of the 96-well filter plate and the apparatus was assembled as previously described. Since compounds will permeate at their rate of diffusion the assay time was shortened to 2 h to avoid reaching equilibrium. The permeabilities were calculated using the standard P_{app} equation described above and then adjusted using the filtration solubility ratios as previously described to get $P_{o\ ABL}$ values. The values obtained were assumed to be the permeabilities of the compounds either of monomeric species or small (<200-nm) aggregates through the aqueous boundary layer and filter. The downturn at higher lipophilicities is likely due to the presence of small aggregate species which diffuse slower; the P_o's are much greater than those observed in conventional PAMPA suggesting that the effect of the ABL is minimal on our large compounds.

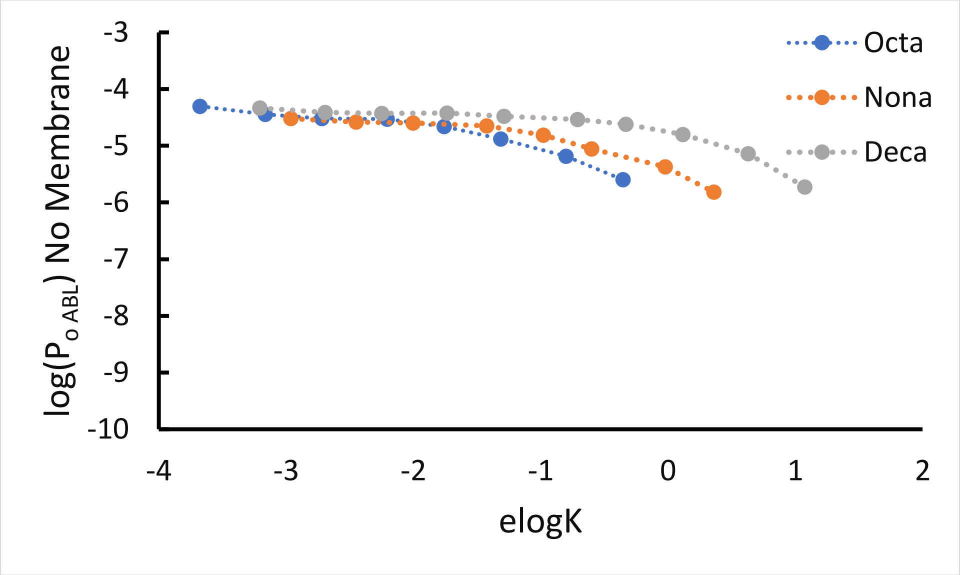


Figure 2-9 $P_{o\ ABL}$ vs elog$K_{hc/w}$ for the permethylated lipophillicity scanning library.

2.6.5.9 *Membrane Diffusion Calculation:*

For the library members (permethyl octa- ,nona- , and deca-peptides) diffusions were calculated accounting for the small effect of the aqueous boundary layer as measured. The equation used was as follows:

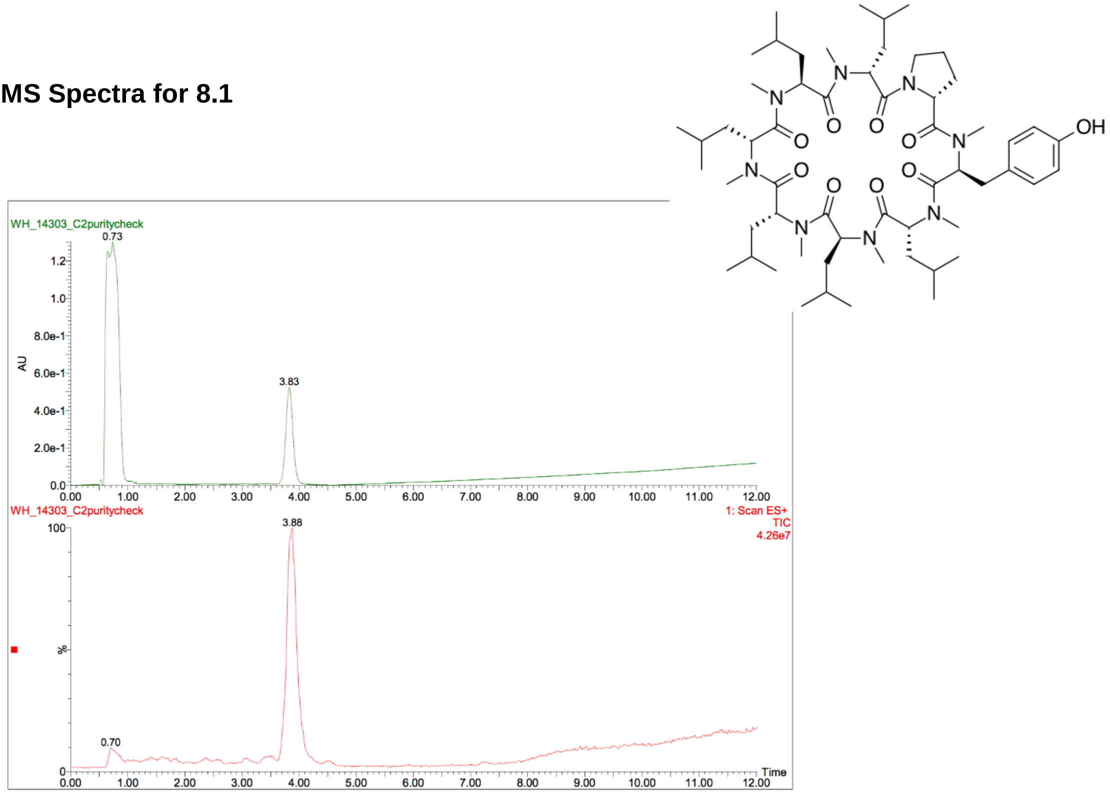
$$D_{mem} = \frac{P_o * \partial_{mem}}{K_{hc/w}}$$

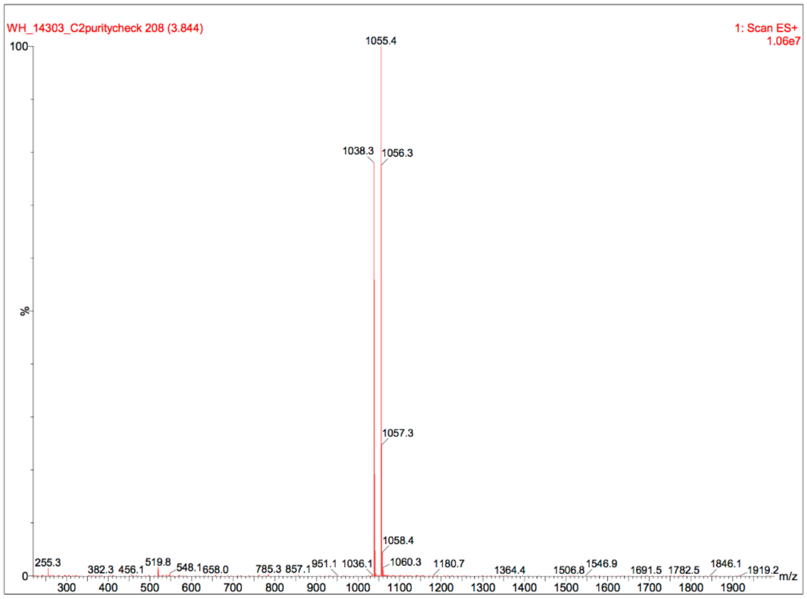
The thickness of the membrane, ∂_{mem} , was 0.0125 cm as previously reported

2.7 Characterization of Compounds

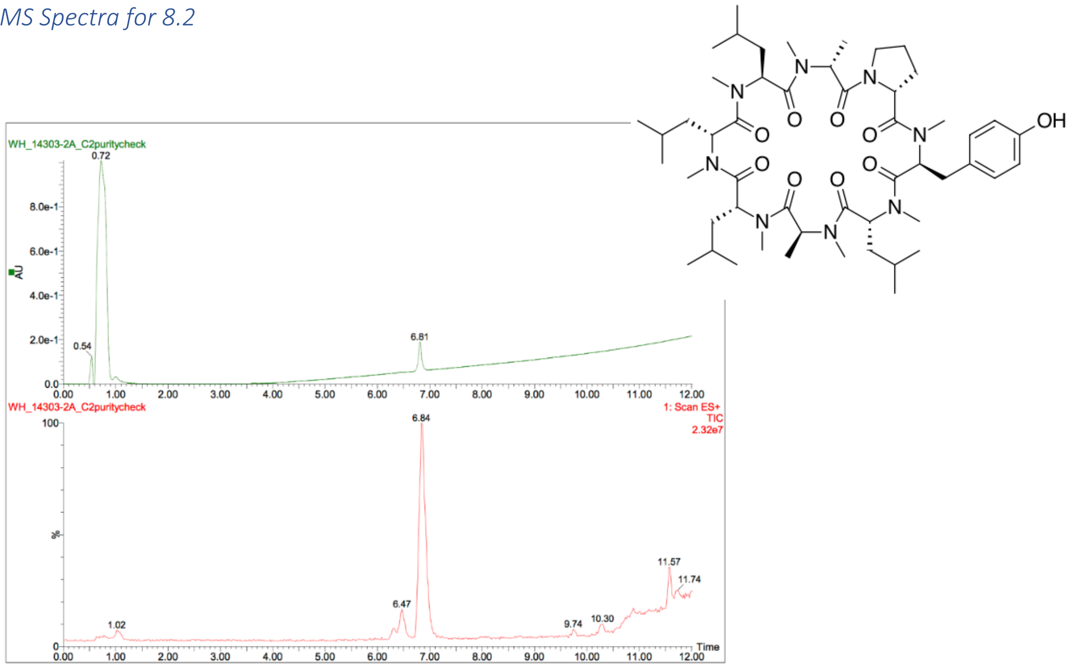
2.8 LC/MS Characterization of Selected Pure Compounds

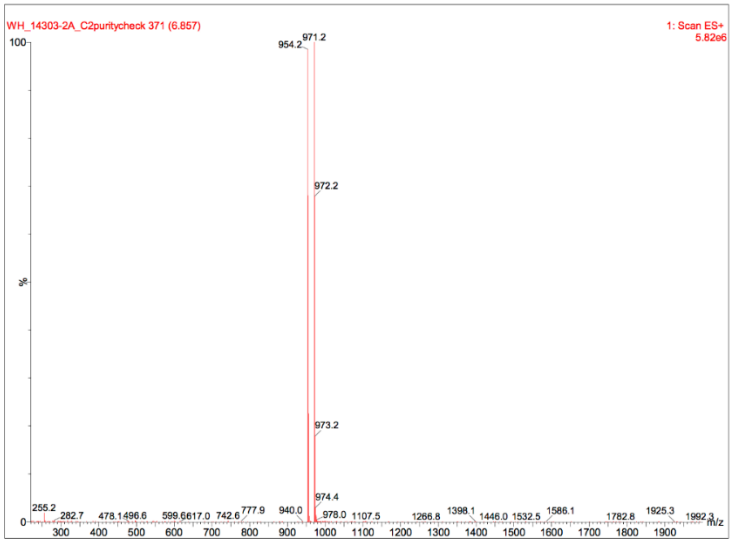
LCMS Spectra for 8.1



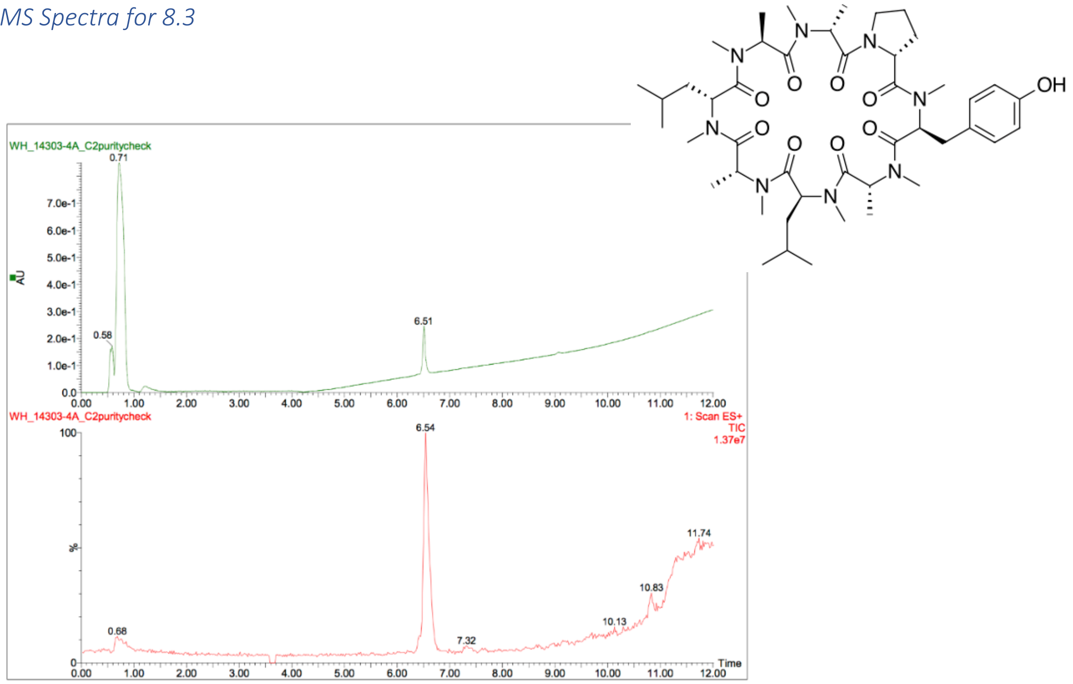


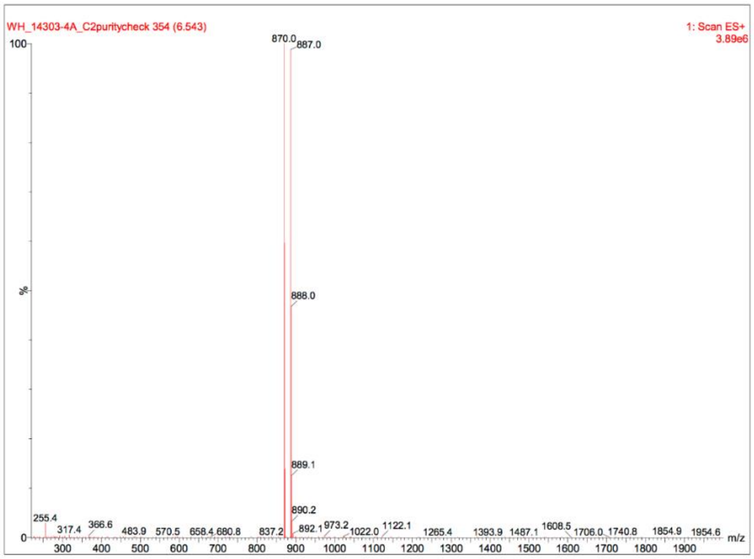
LCMS Spectra for 8.2



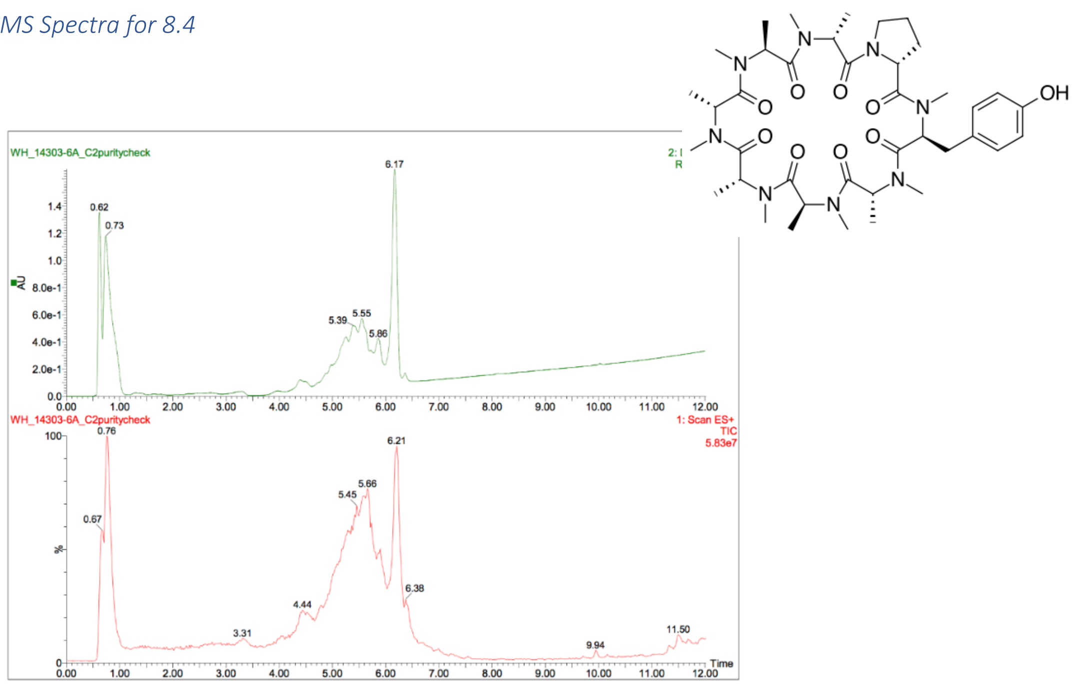


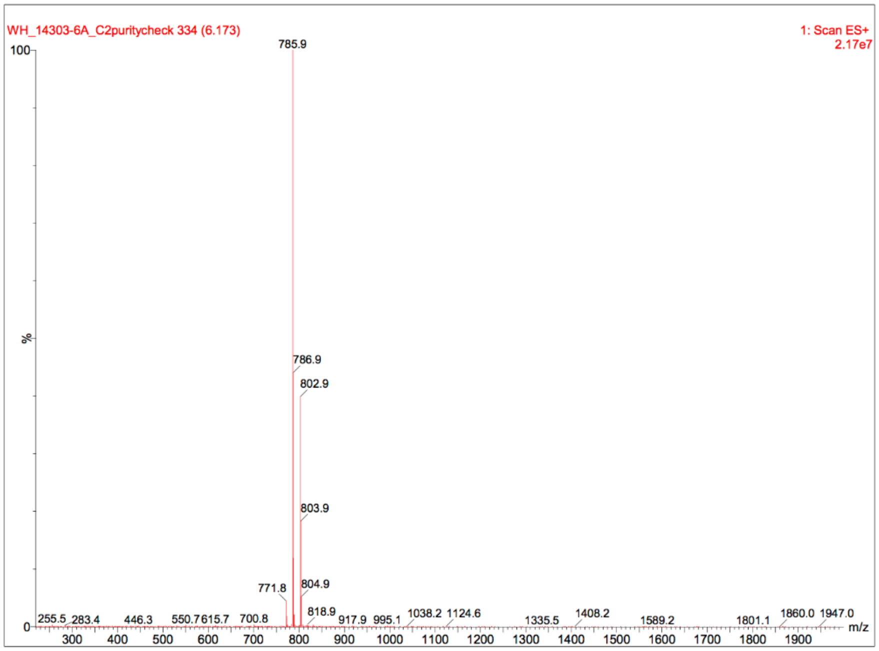
LCMS Spectra for 8.3



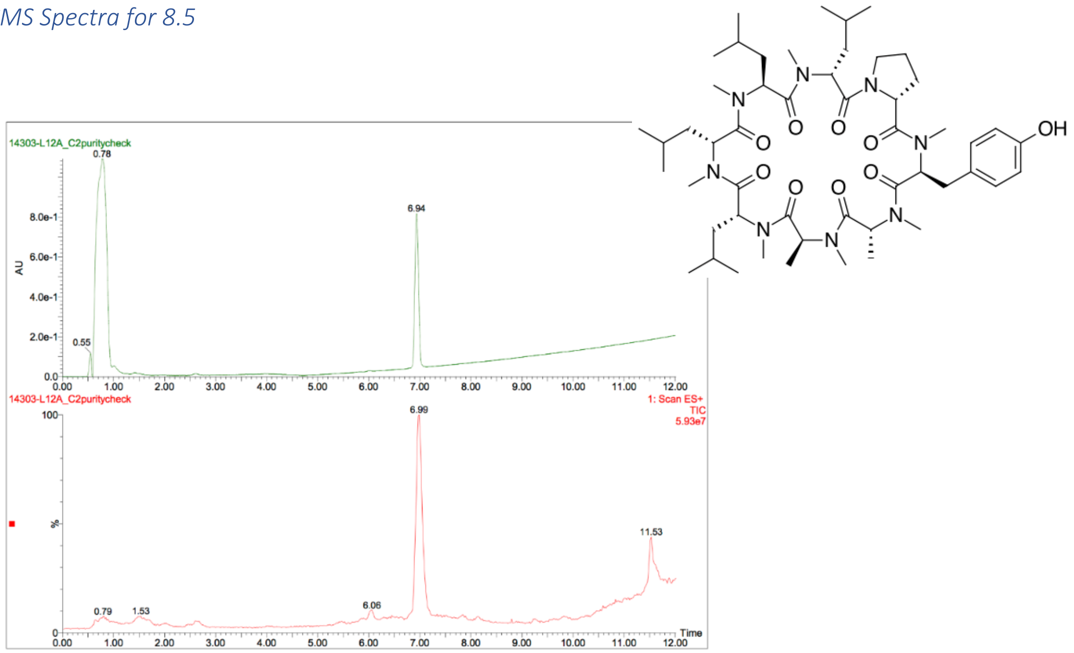


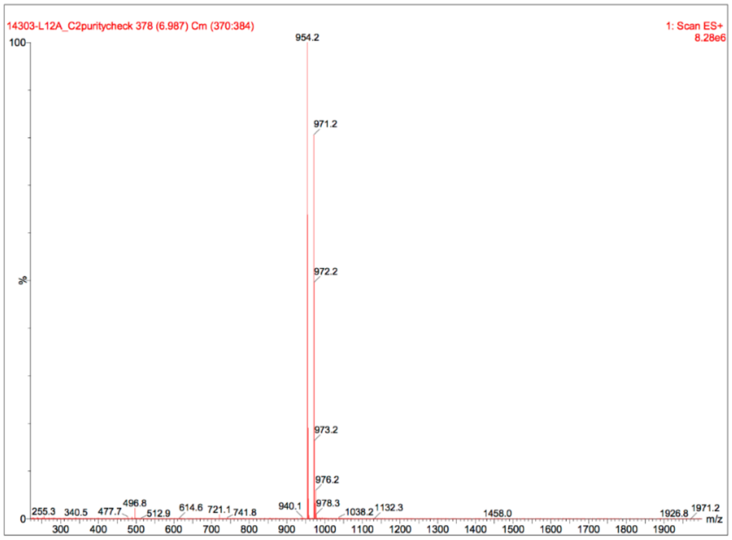
LCMS Spectra for 8.4



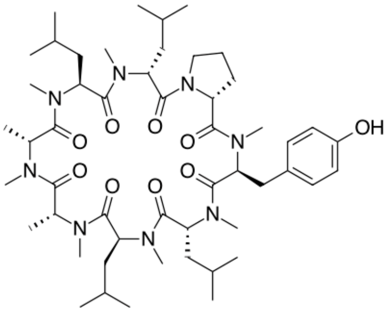


LCMS Spectra for 8.5

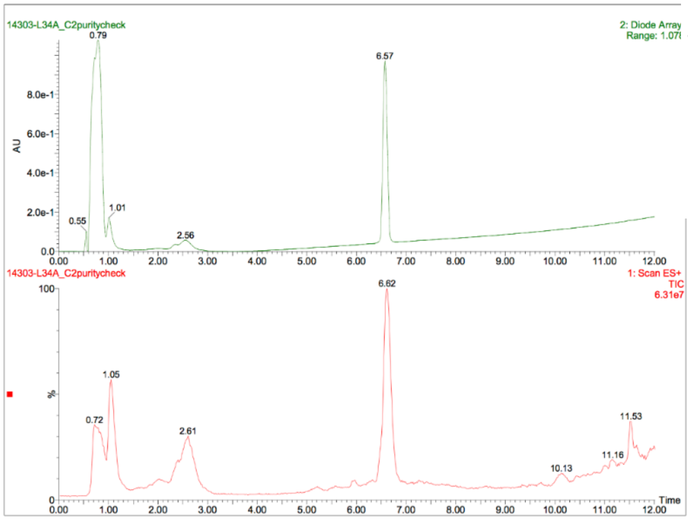


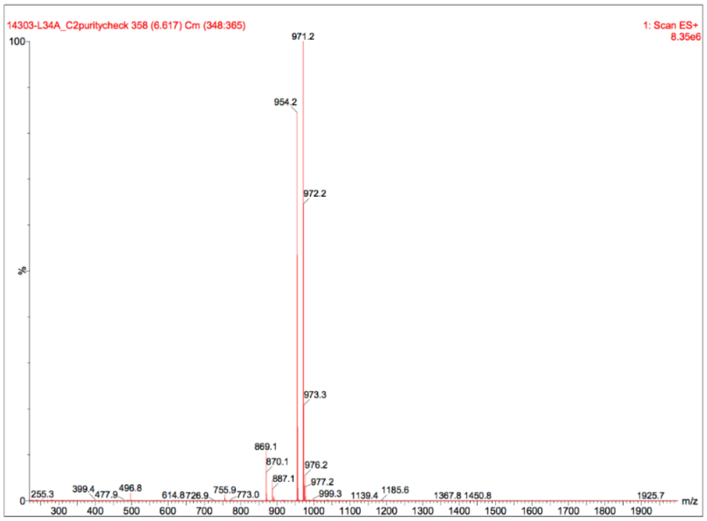


2.8.1.1 LCMS Spectra for 8.6

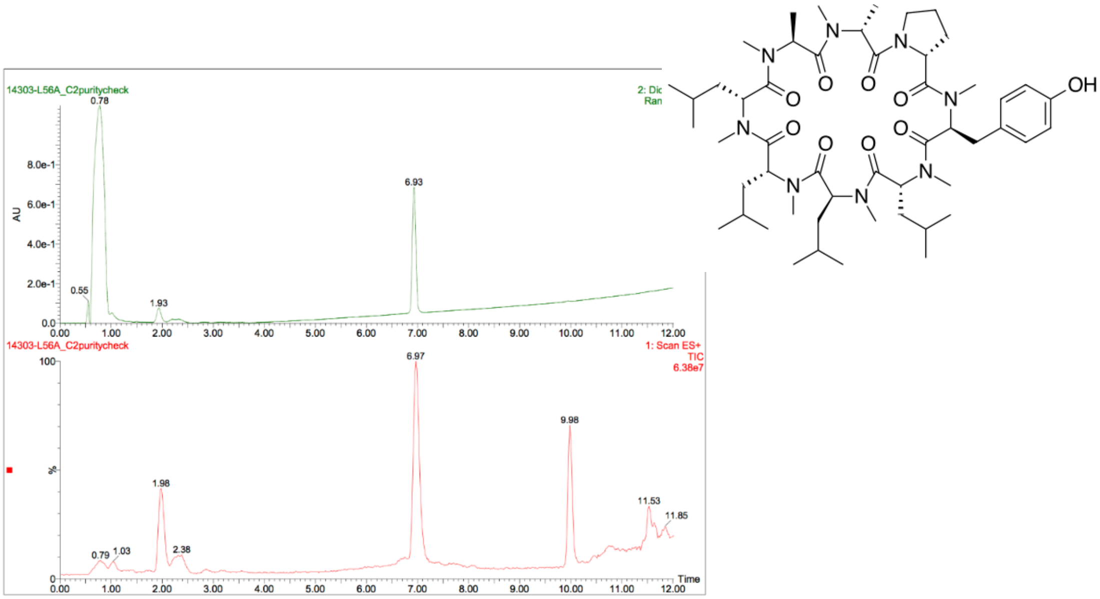


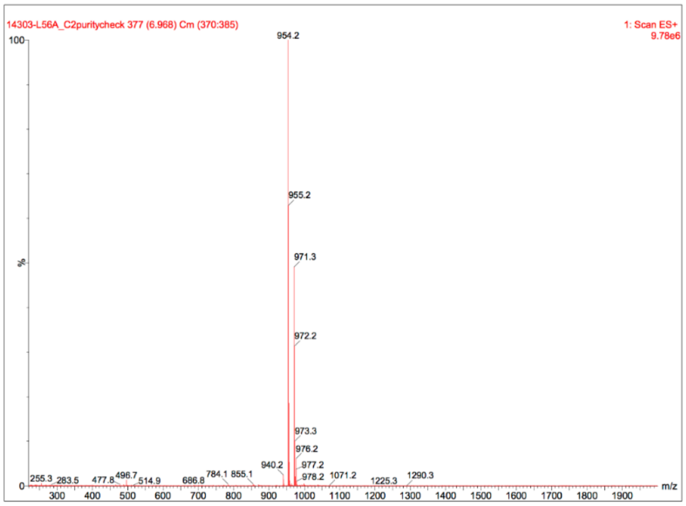
N
N
N
N
N
N
N
N
O
O
O
O
O
O
O
O
OH



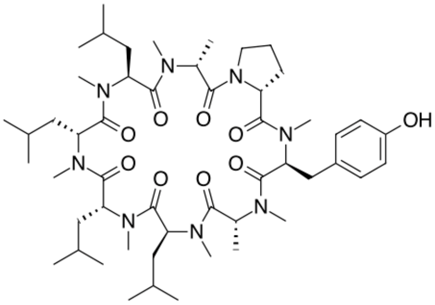


2.8.1.2 LCMS Spectra for 8.7

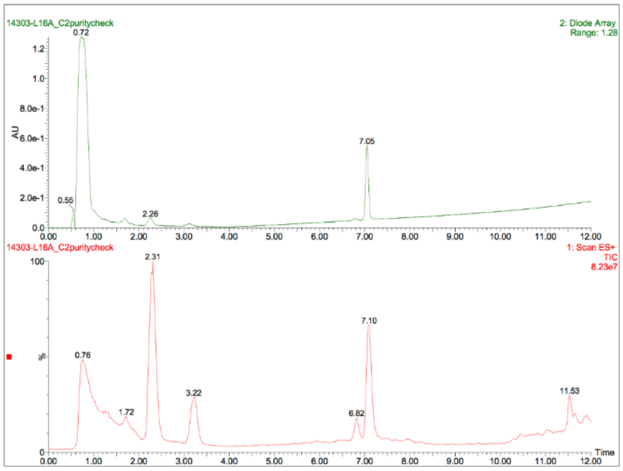


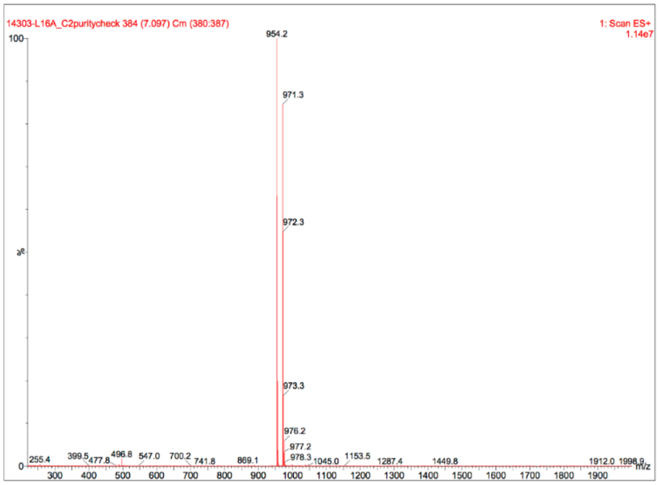


2.8.1.3 LCMS Spectra for 8.8

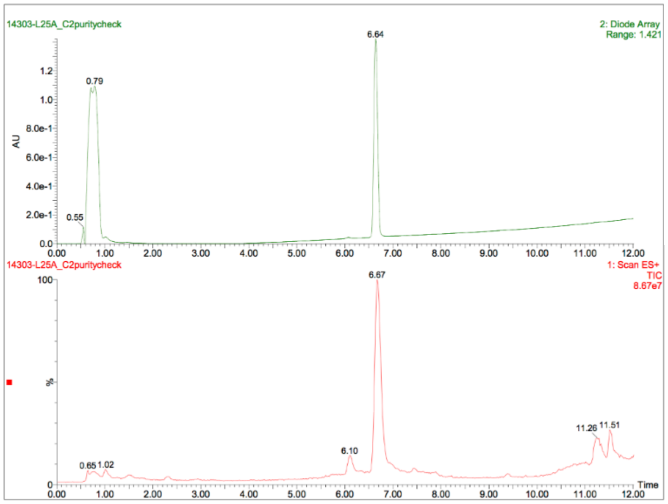


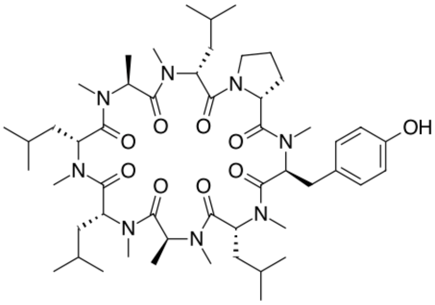
N
N
N
O
O
O
N
OH
N
O
O
N
N
N
O
O
O
N

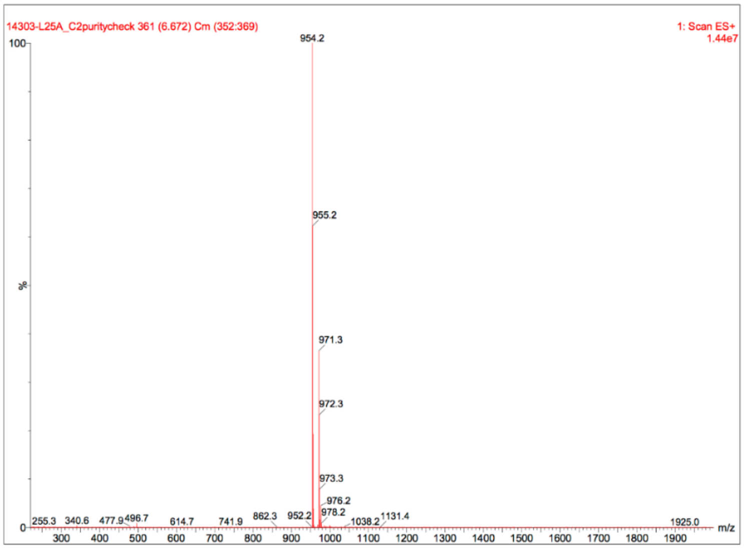




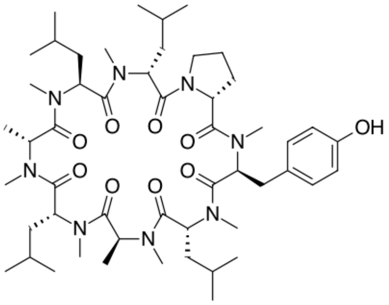
2.8.1.4 LCMS Spectra for 8.9

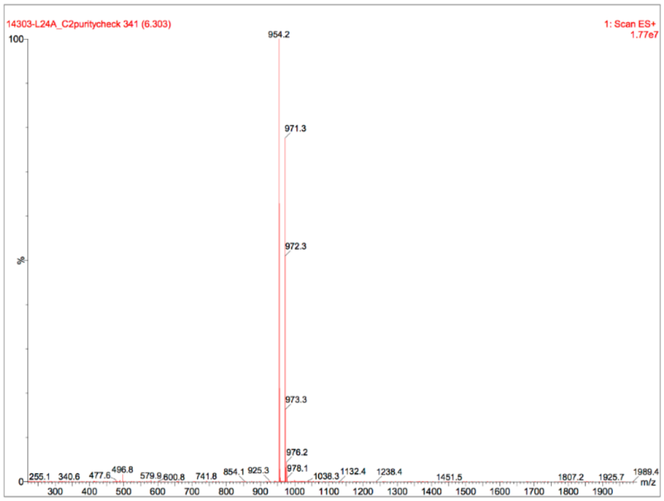


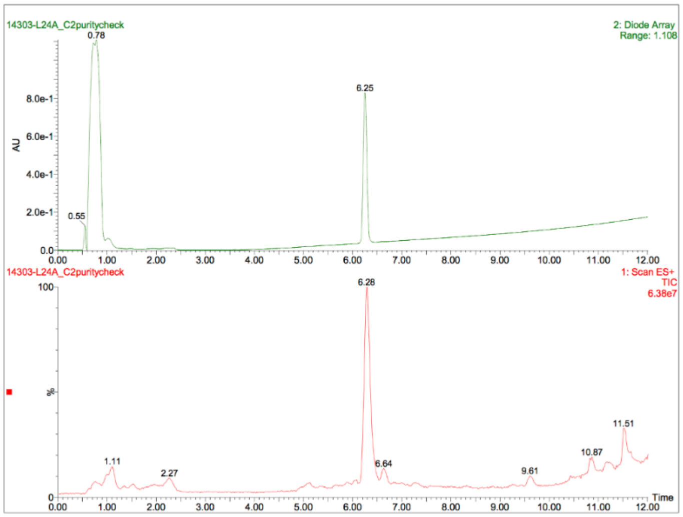




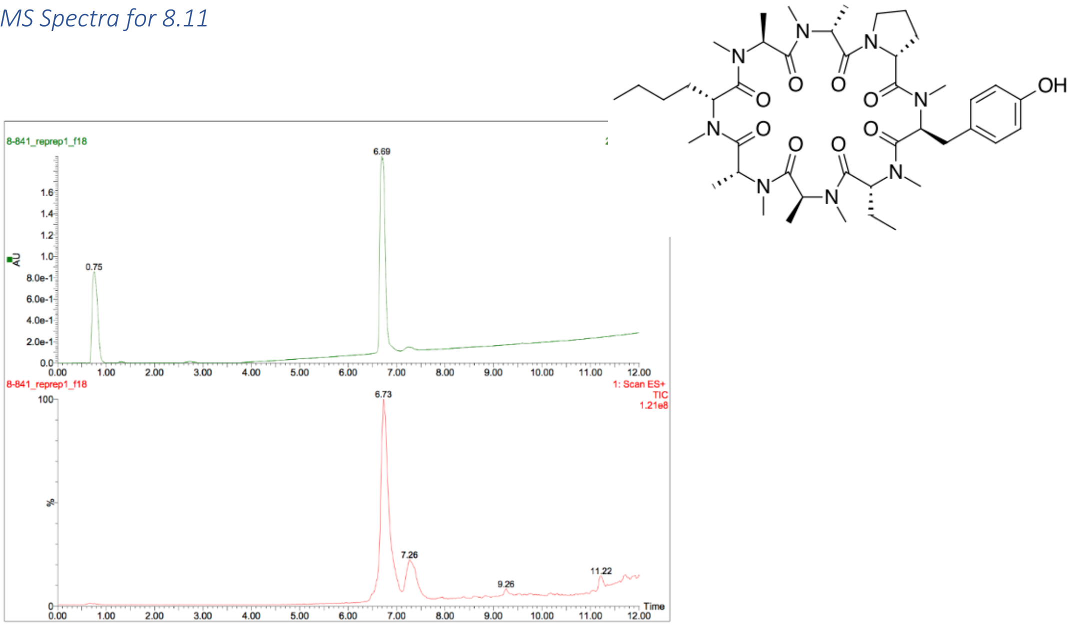
2.8.1.5 LCMS Spectra for 8.10

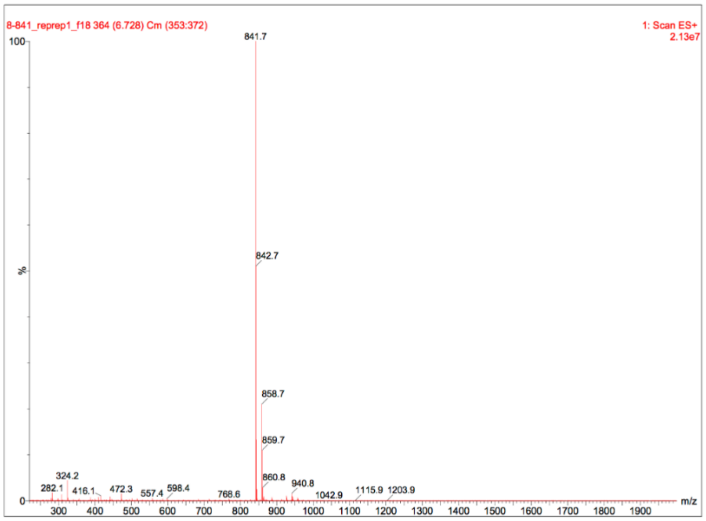




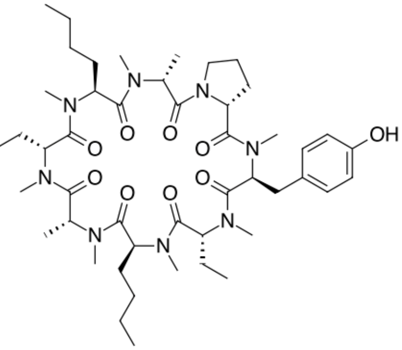


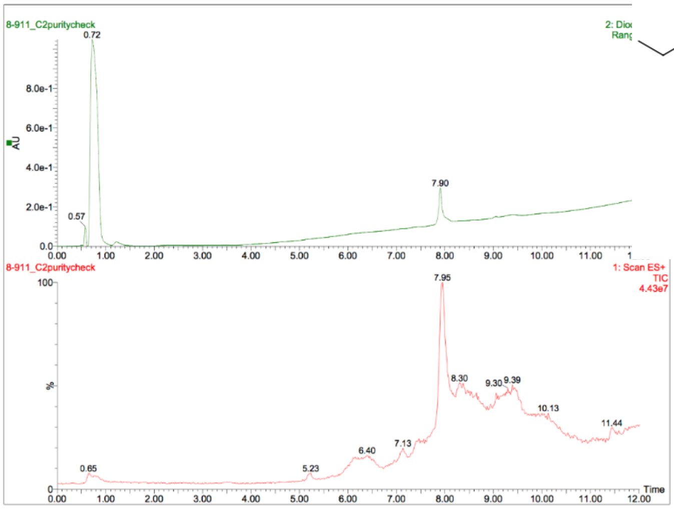
LCMS Spectra for 8.11

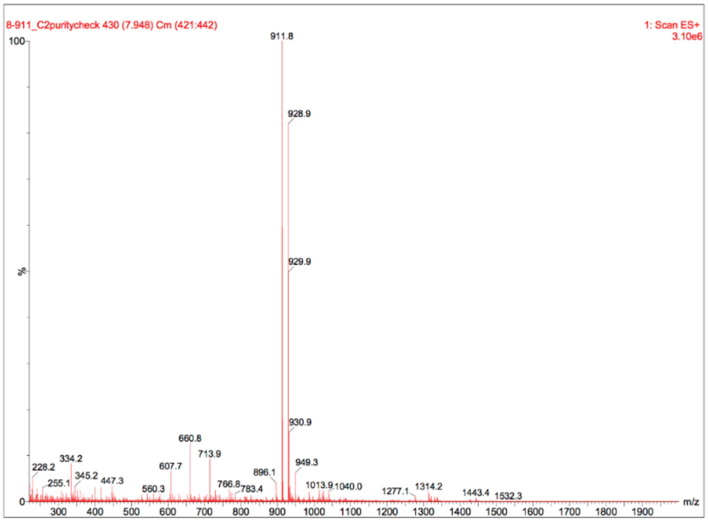




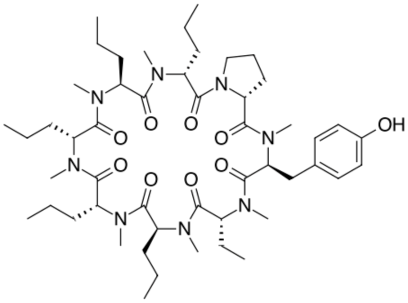
2.8.1.6 LCMS Spectra for 8.12

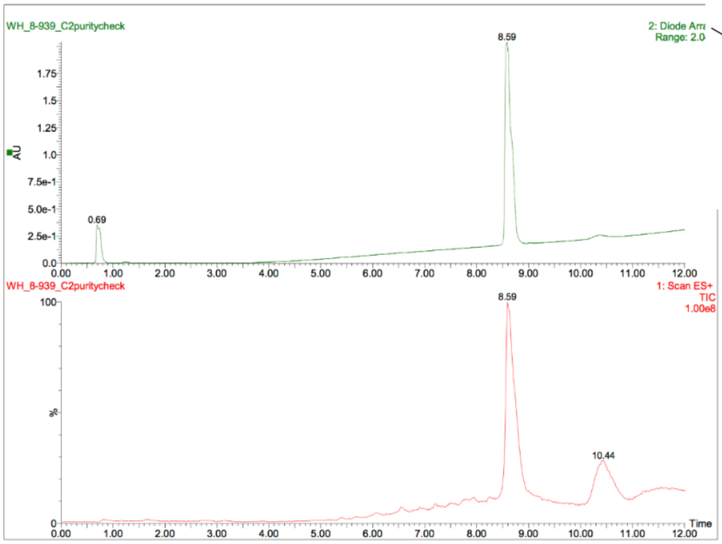


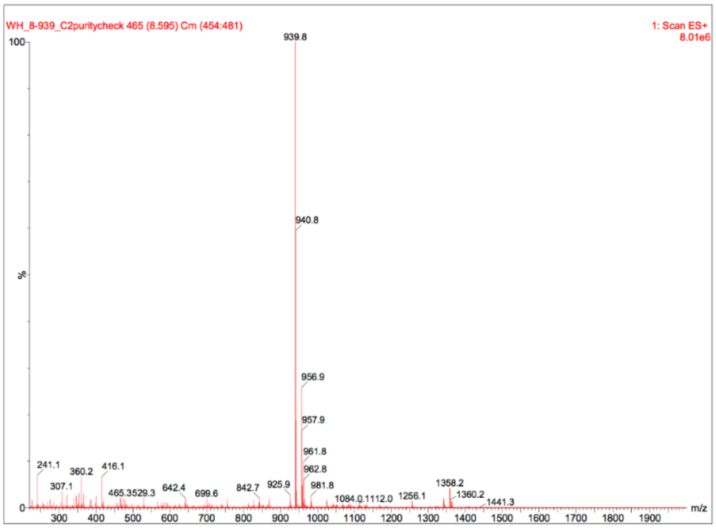




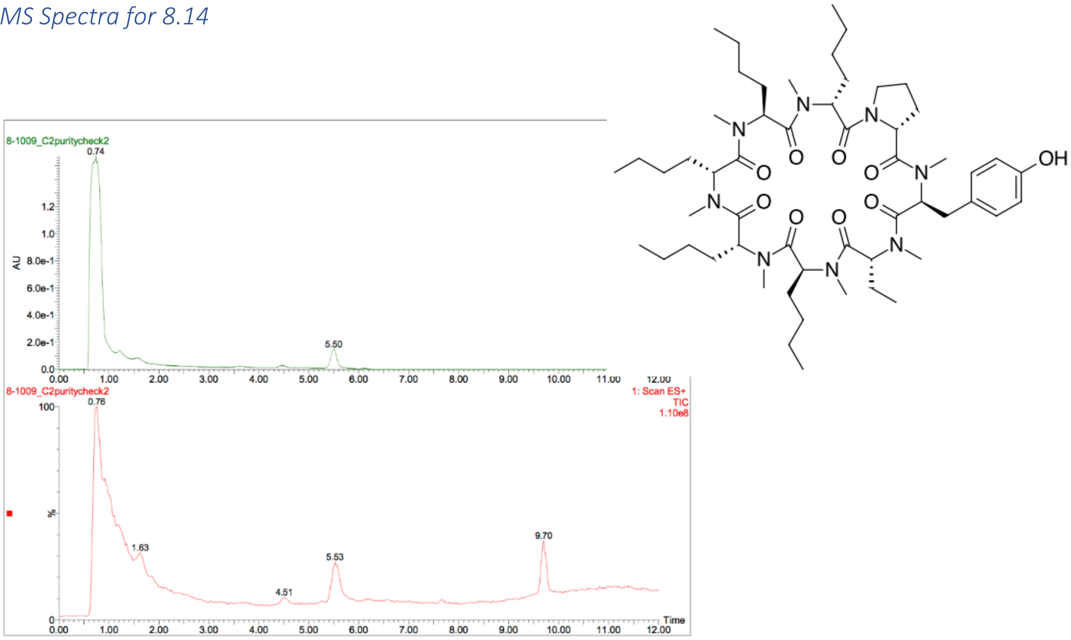
2.8.1.7 LCMS Spectra for 8.13

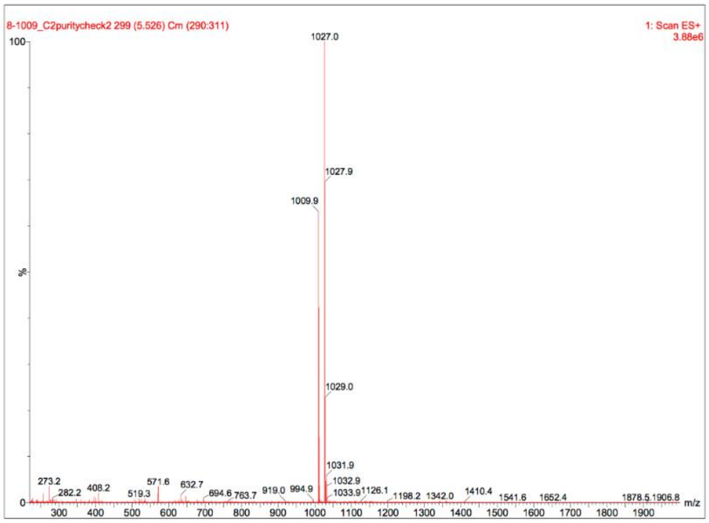




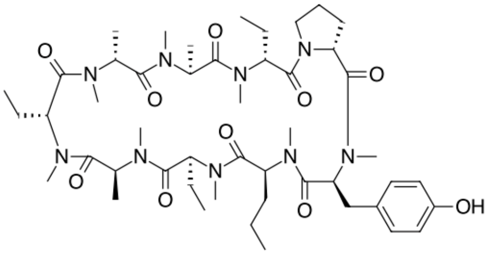


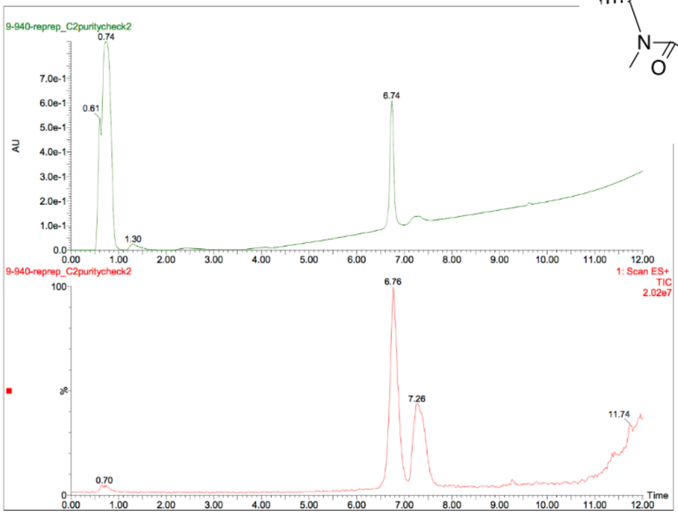
LCMS Spectra for 8.14

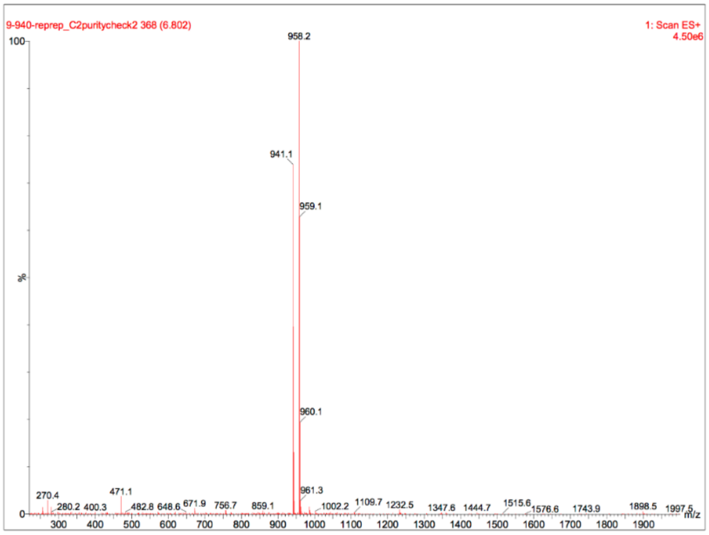




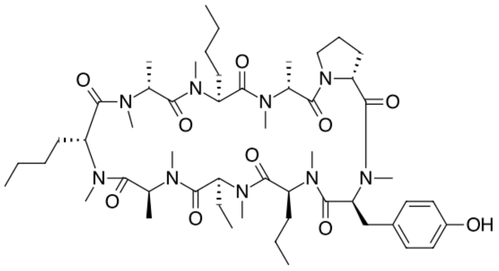
LCMS Spectra for 9.1

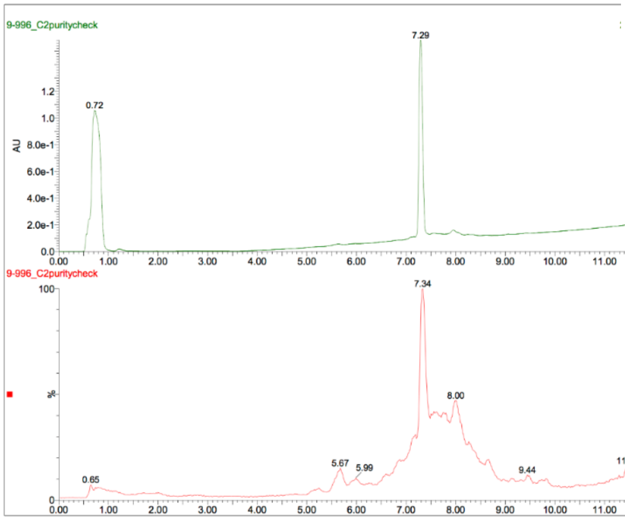

O
N
O
N
O
N
O
N
O
N
N
O
N
O
N
O
N
O
N
O
OH

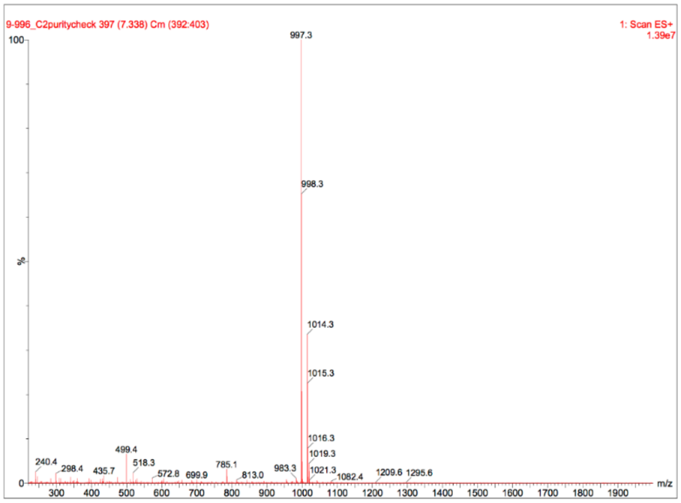




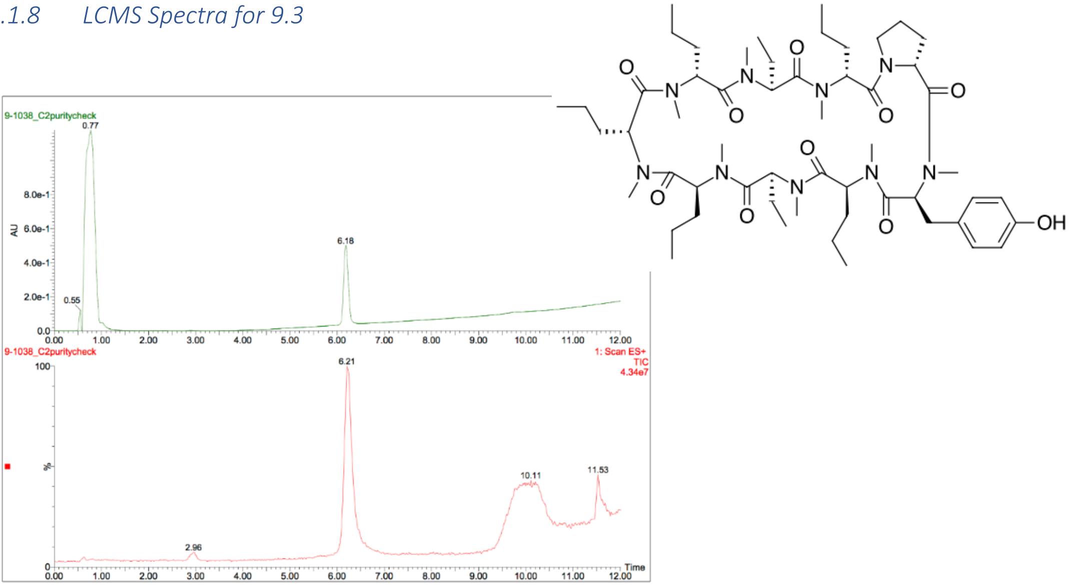
LCMS Spectra for 9.2

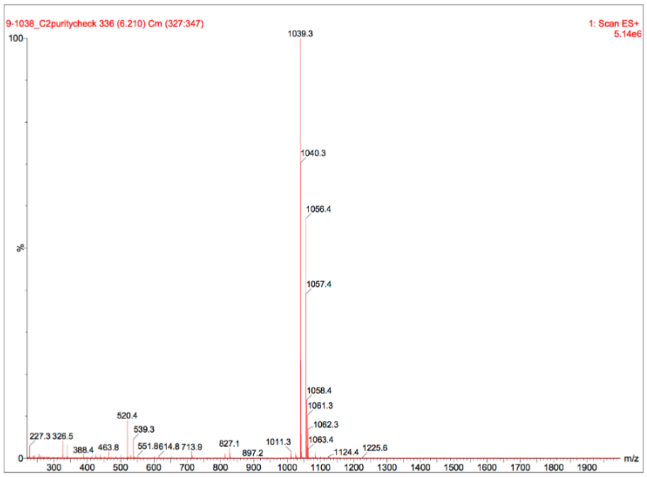

O
N
O
N
O
N
O
N
O
O
N
N
O
N
O
N
O
N
O
N
O
OH



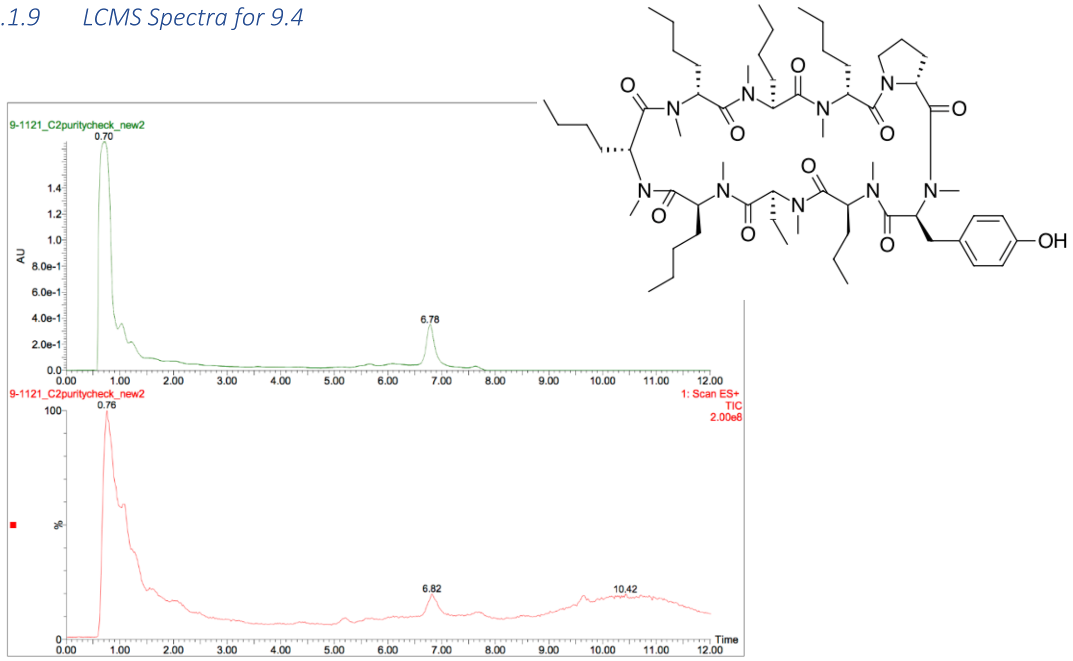


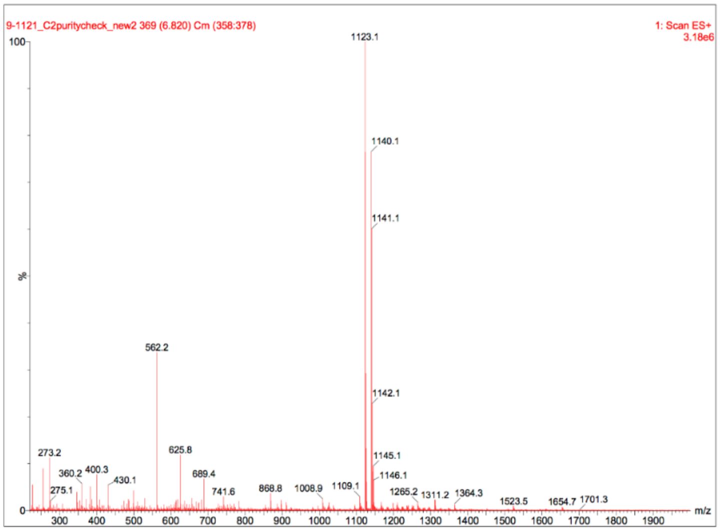
2.8.1.8 LCMS Spectra for 9.3



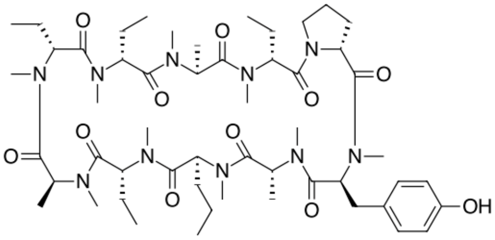


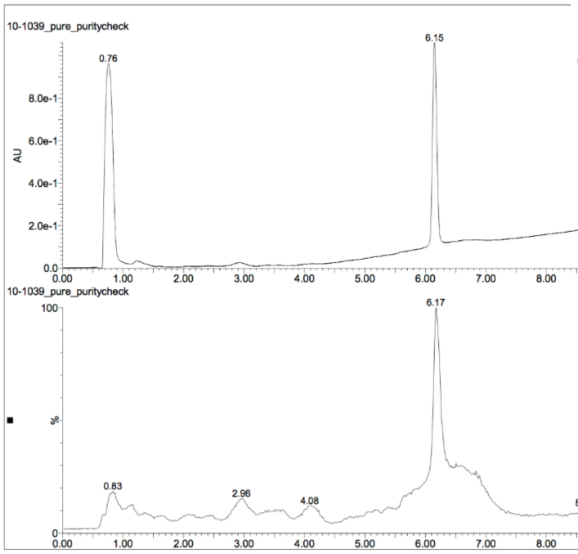
2.8.1.9 LCMS Spectra for 9.4

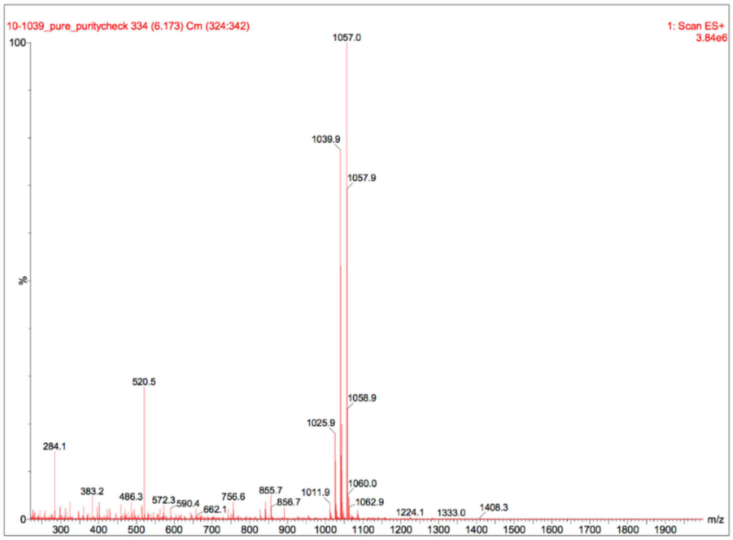




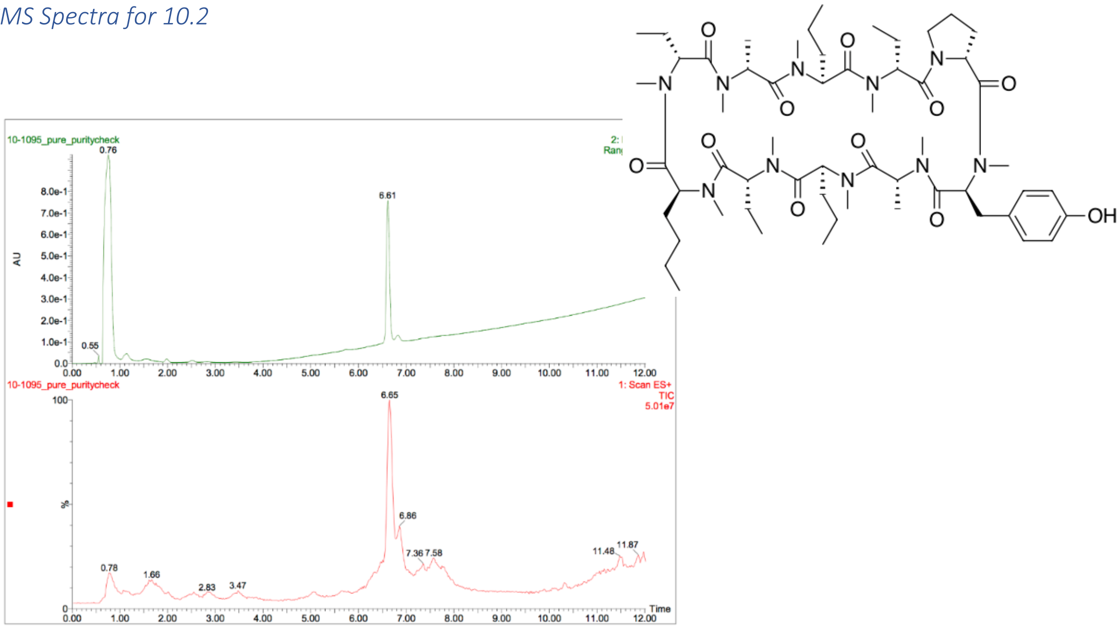
2.8.1.10 LCMS Spectra for 10.1

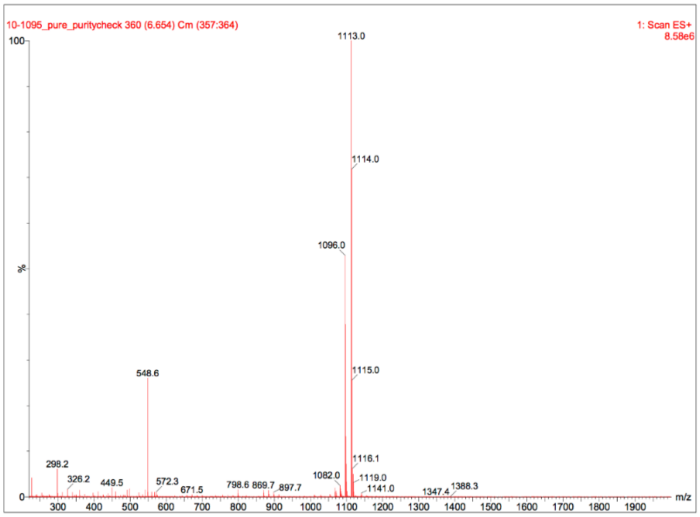




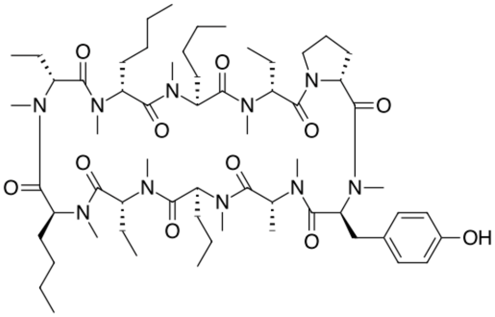


LCMS Spectra for 10.2

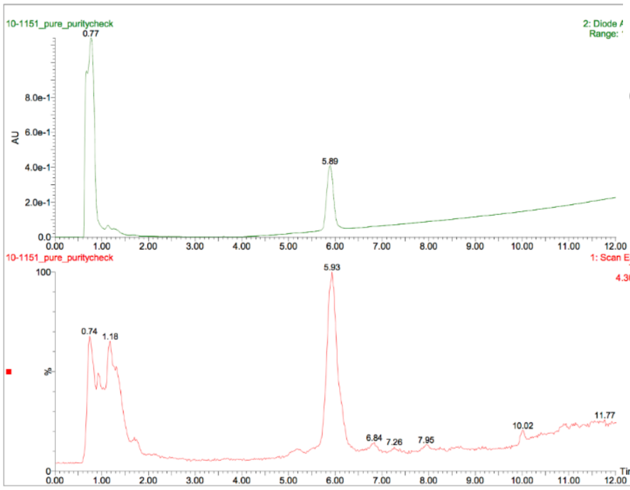


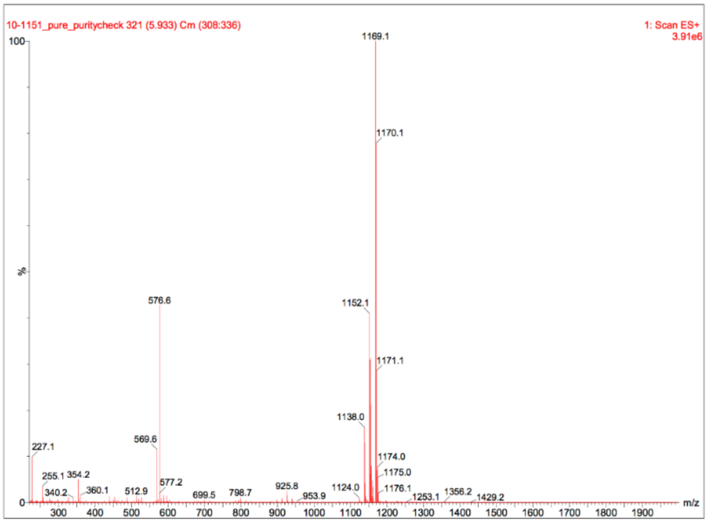


2.8.1.11 LCMS Spectra for 10.3



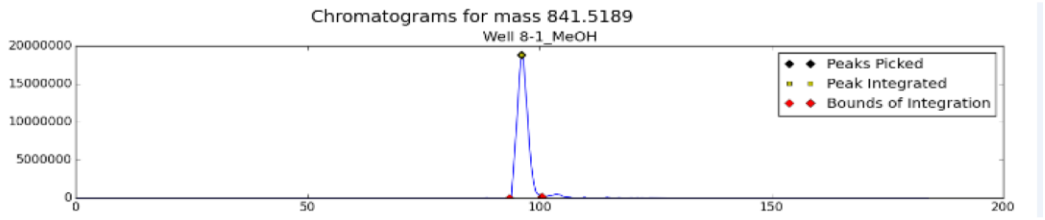
O
N
N
N
N
N
O
O
O
O
O
O
N
N
N
N
N
O
O
O
O
OH

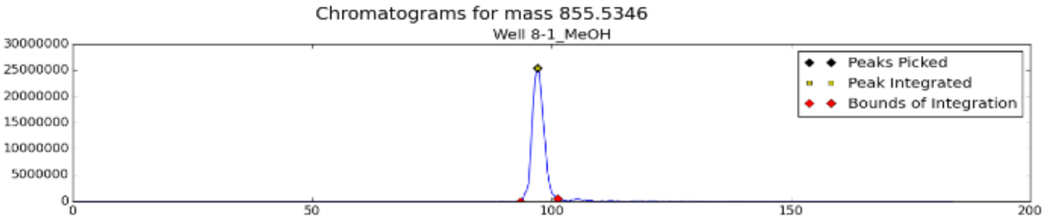


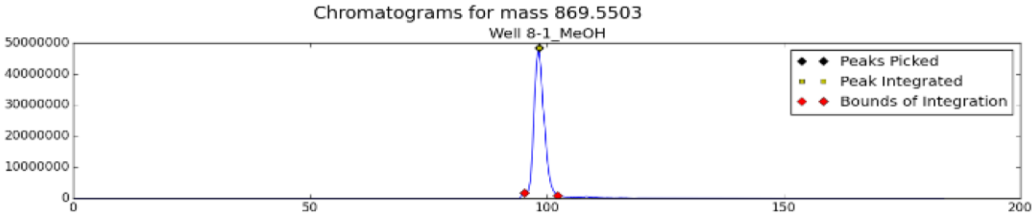


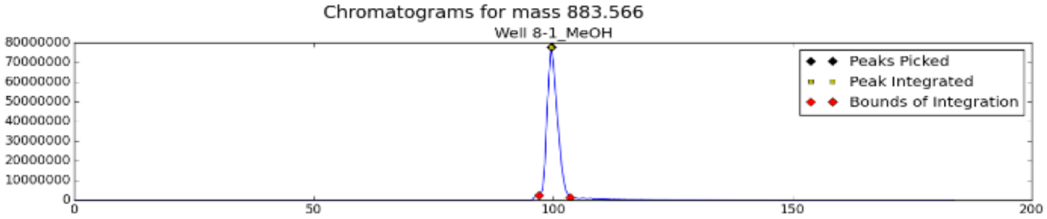
2.8.2 Extracted Ion Chromatograms for Lipophillicity Scanning Mixtures

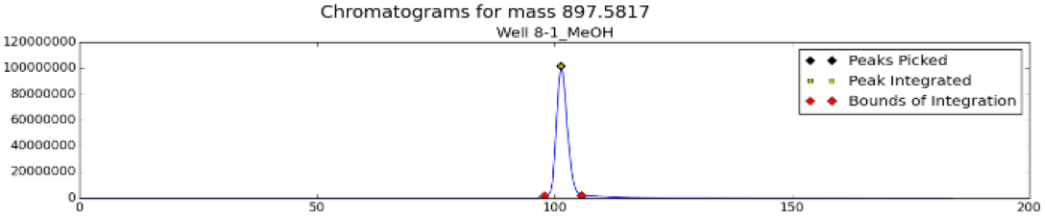
Octapeptide Extracted Ion Chromatograms

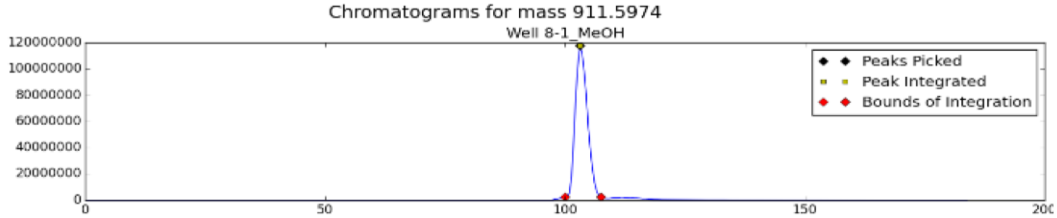


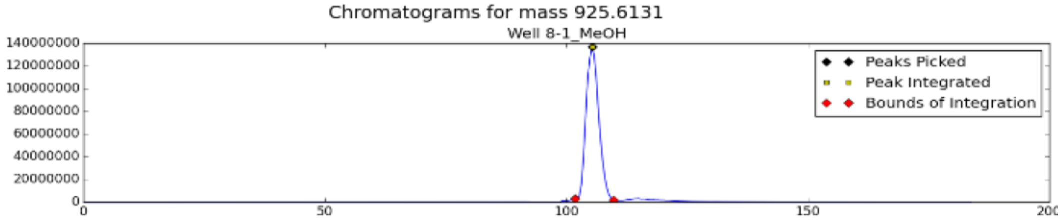


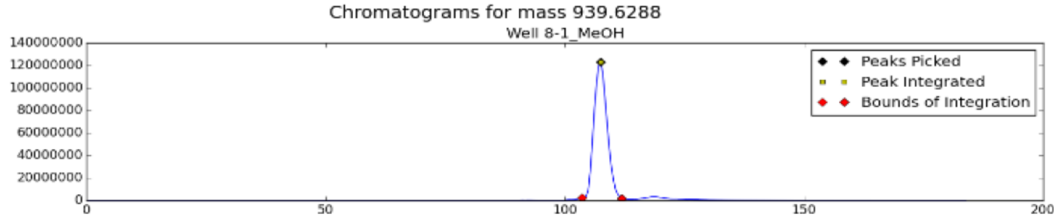


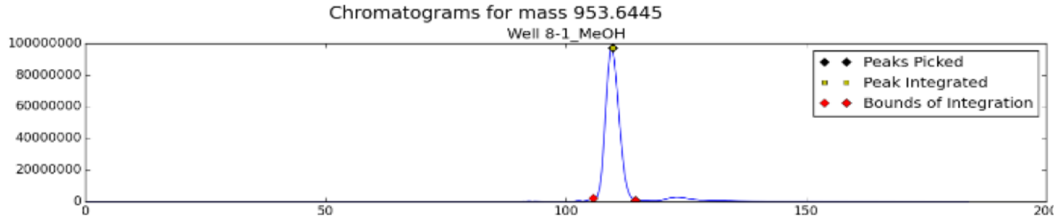


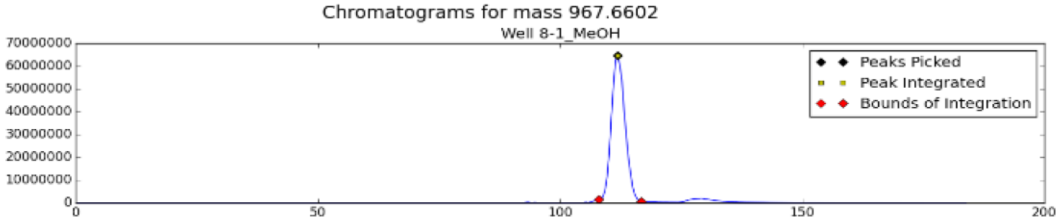


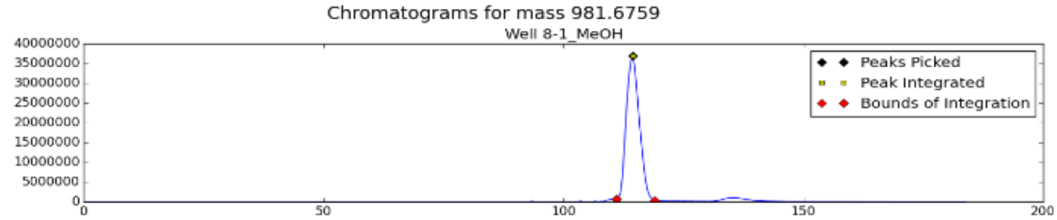

Chromatograms for mass 911.5974
Well 8-1_MeOH
120000000
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration

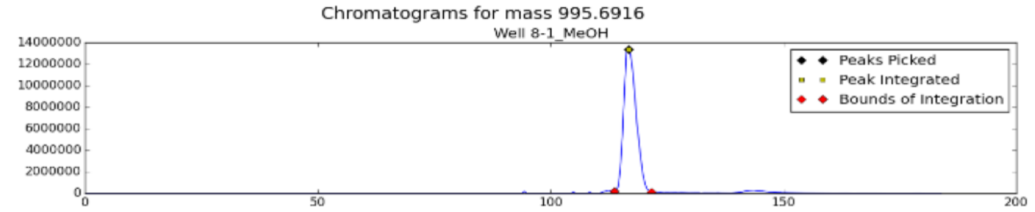

Chromatograms for mass 925.6131
Well 8-1_MeOH
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration


Chromatograms for mass 939.6288
Well 8-1_MeOH
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration

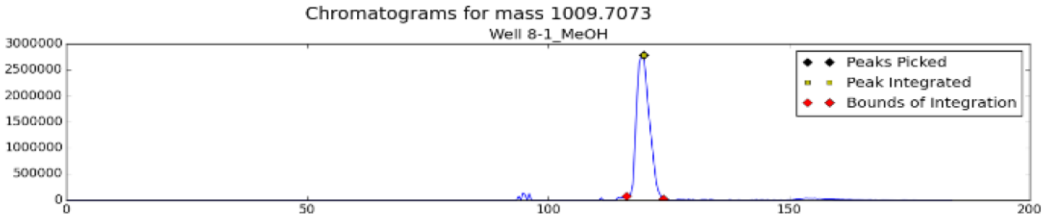

Chromatograms for mass 953.6445
Well 8-1_MeOH
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration


Chromatograms for mass 967.6602
Well 8-1_MeOH
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration


Chromatograms for mass 981.6759
Well 8-1_MeOH
40000000
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
0
50
100
150
200
◆ ◆ Peaks Picked
■ ■ Peak Integrated
◆ ◆ Bounds of Integration

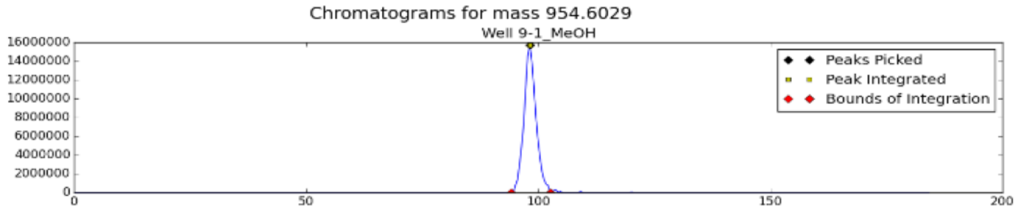

Chromatograms for mass 995.6916

Well 8-1_MeOH

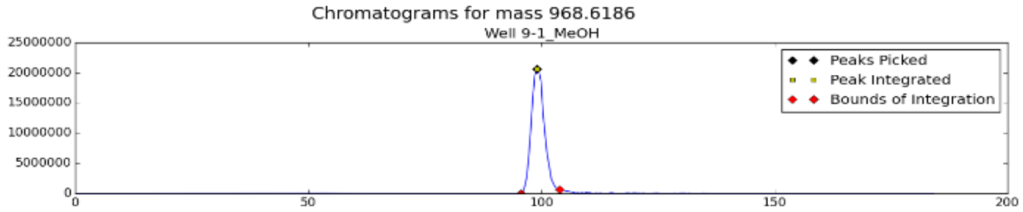

Chromatograms for mass 1009.7073

Well 8-1_MeOH

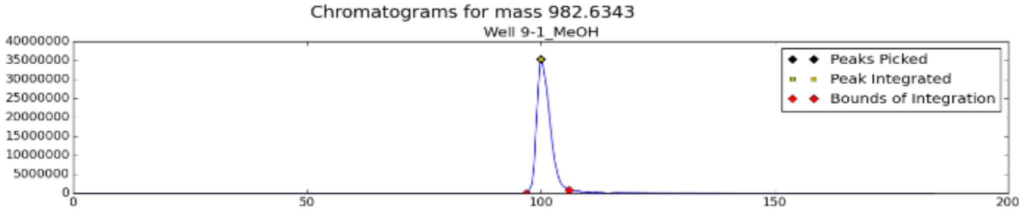
Nonapeptide Extracted Ion Chromatograms


Chromatograms for mass 954.6029

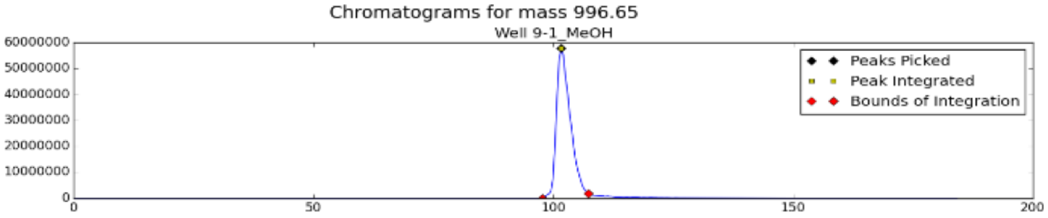
Well 9-1_MeOH


Chromatograms for mass 968.6186

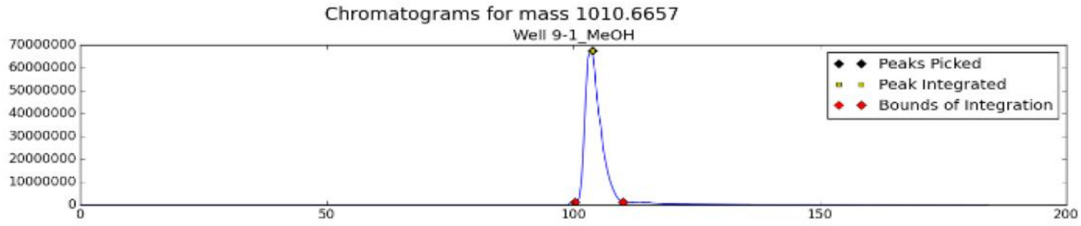
Well 9-1_MeOH

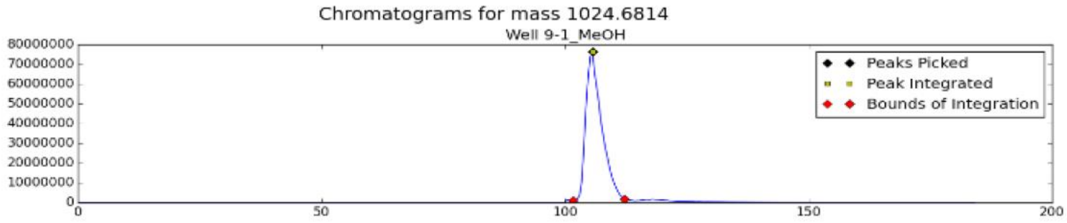

Chromatograms for mass 982.6343

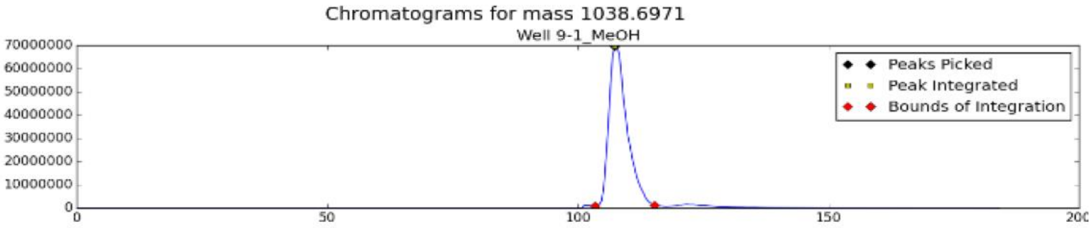
Well 9-1_MeOH

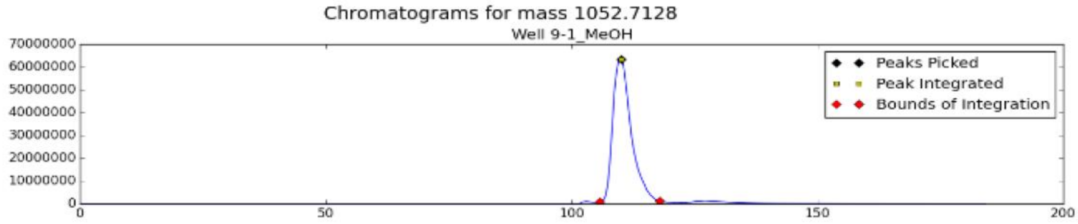

Chromatograms for mass 996.65

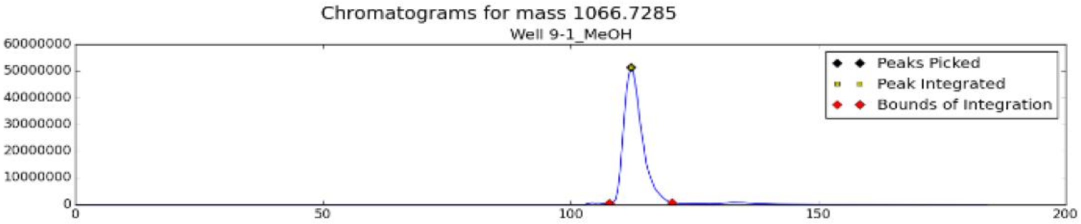
Well 9-1_MeOH

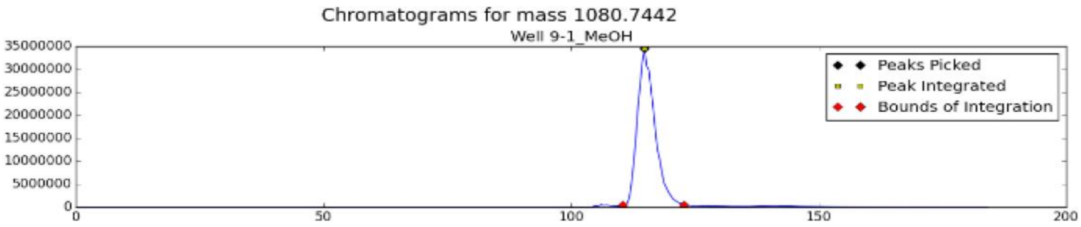

Chromatograms for mass 1010.6657
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200

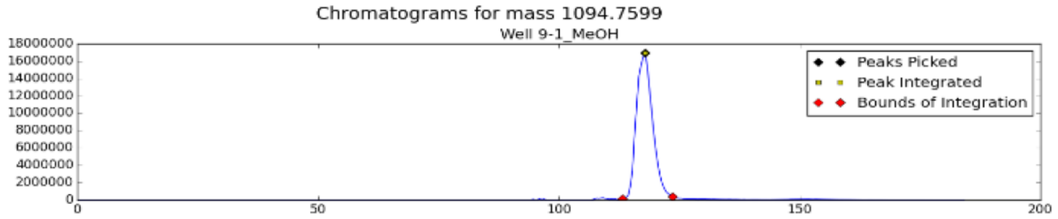

Chromatograms for mass 1024.6814
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200

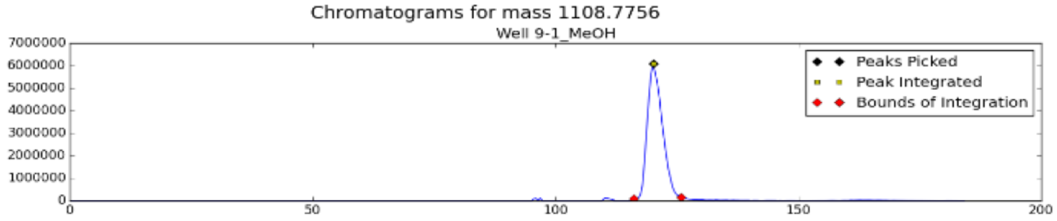

Chromatograms for mass 1038.6971
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200

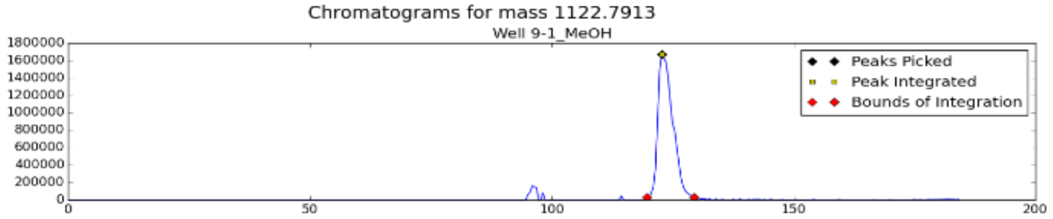

Chromatograms for mass 1052.7128
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200


Chromatograms for mass 1066.7285
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200

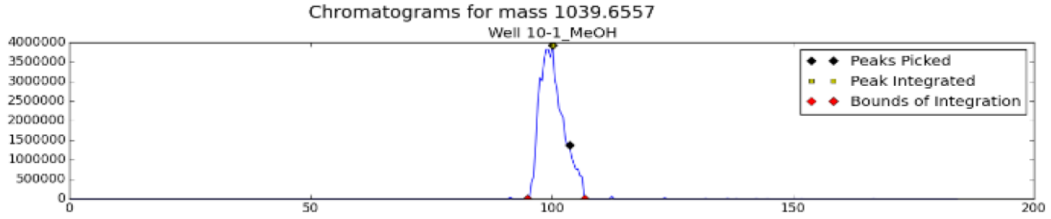

Chromatograms for mass 1080.7442
Well 9-1_MeOH
Peaks Picked
Peak Integrated
Bounds of Integration
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
0
50
100
150
200

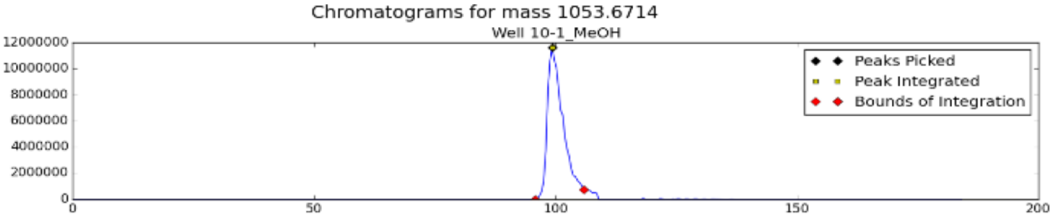


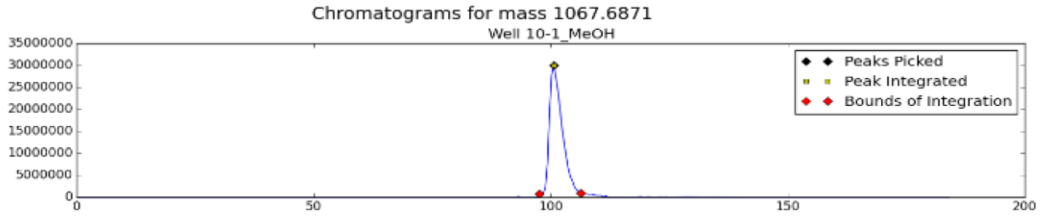


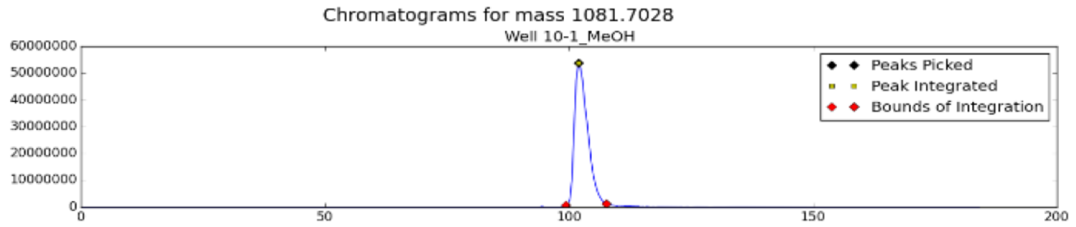


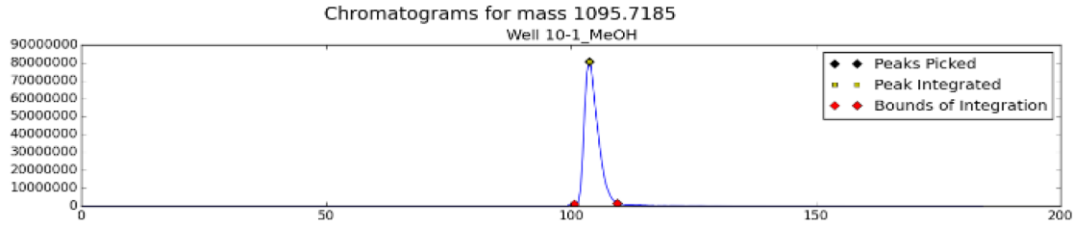
Decapeptide Extracted Ion Chromatograms

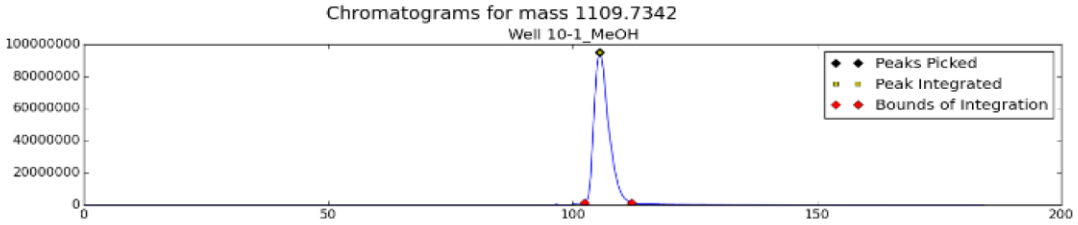


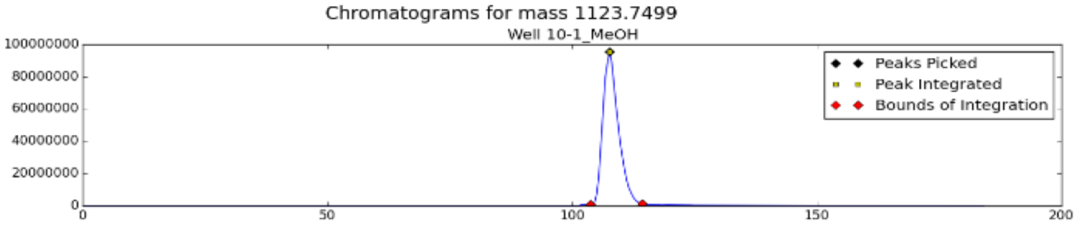


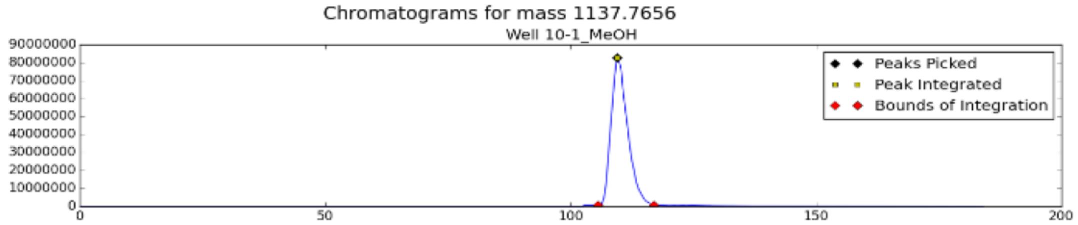


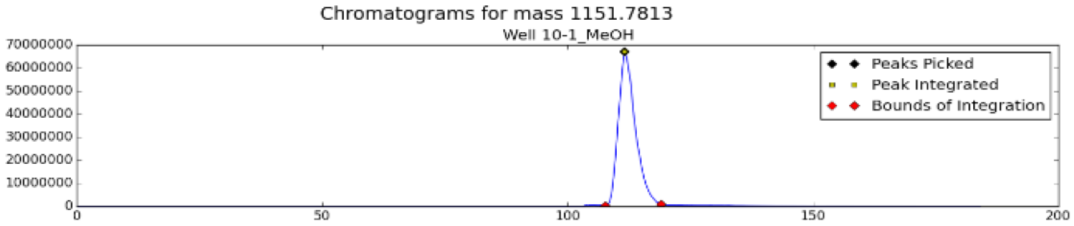

Chromatograms for mass 1081.7028
Well 10-1_MeOH
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration

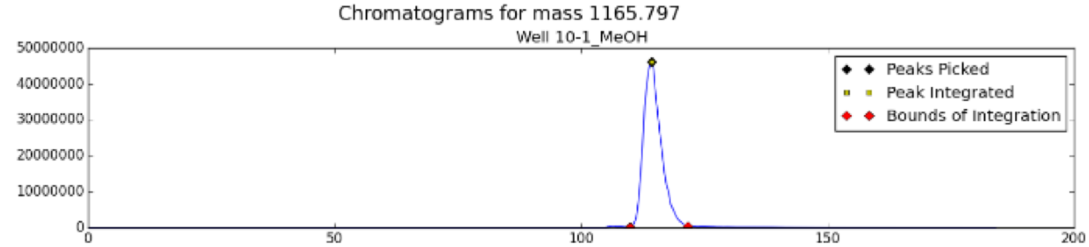

Chromatograms for mass 1095.7185
Well 10-1_MeOH
90000000
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration


Chromatograms for mass 1109.7342
Well 10-1_MeOH
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration

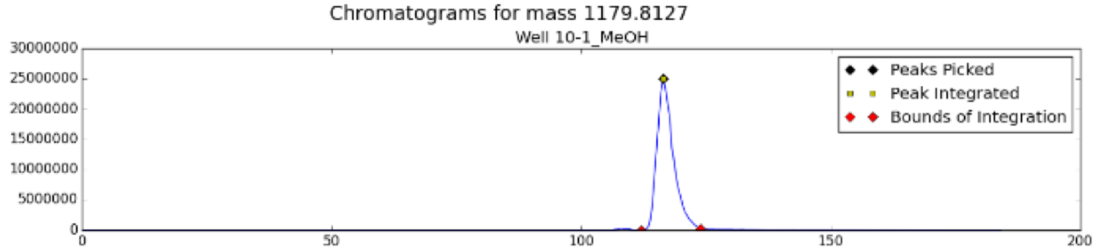

Chromatograms for mass 1123.7499
Well 10-1_MeOH
100000000
80000000
60000000
40000000
20000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration


Chromatograms for mass 1137.7656
Well 10-1_MeOH
90000000
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration

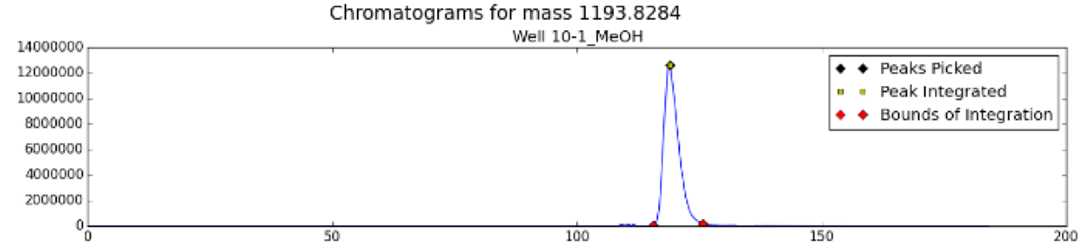

Chromatograms for mass 1151.7813
Well 10-1_MeOH
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
50
100
150
200
Peaks Picked
Peak Integrated
Bounds of Integration


Chromatograms for mass 1165.797

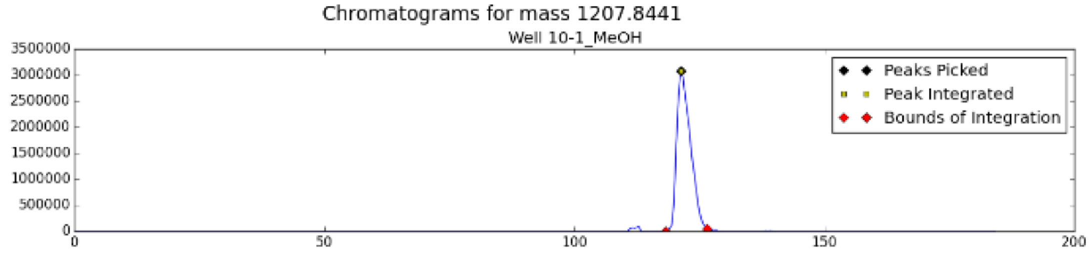
Well 10-1_MeOH


Chromatograms for mass 1179.8127

Well 10-1_MeOH


Chromatograms for mass 1193.8284

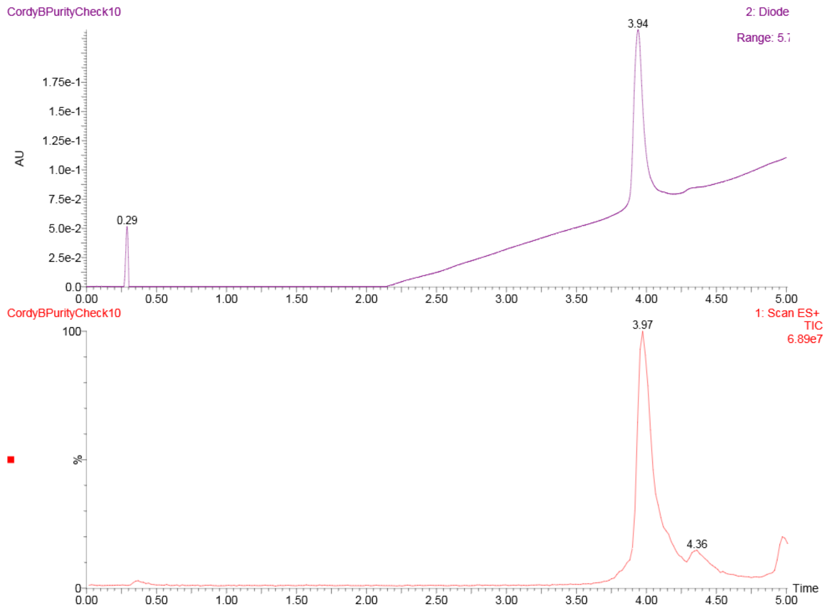
Well 10-1_MeOH

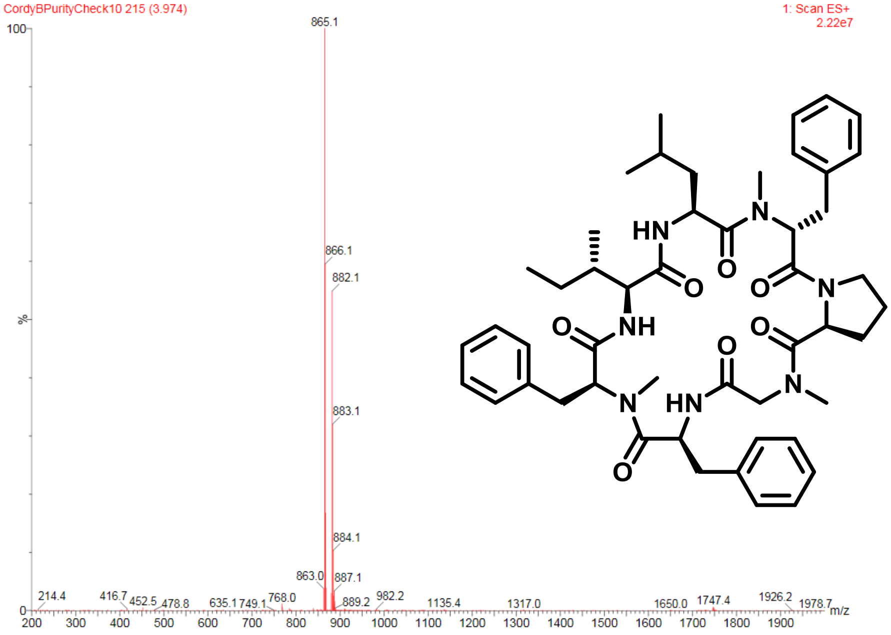

Chromatograms for mass 1207.8441

Well 10-1_MeOH

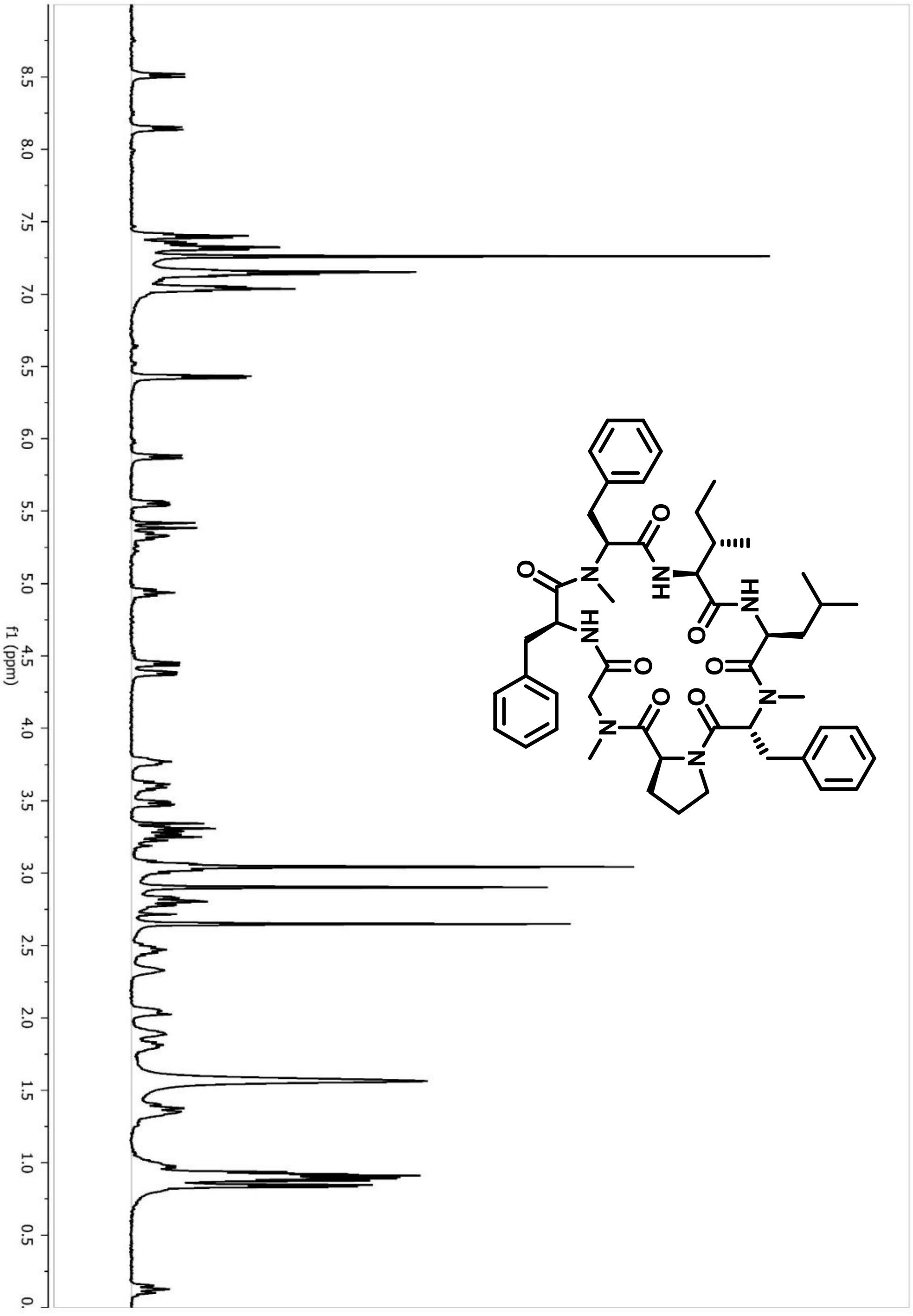
2.8.3 Characterization of Natural Products and Analogs

2.8.3.1 LCMS Spectra for Cordyheptapeptide B

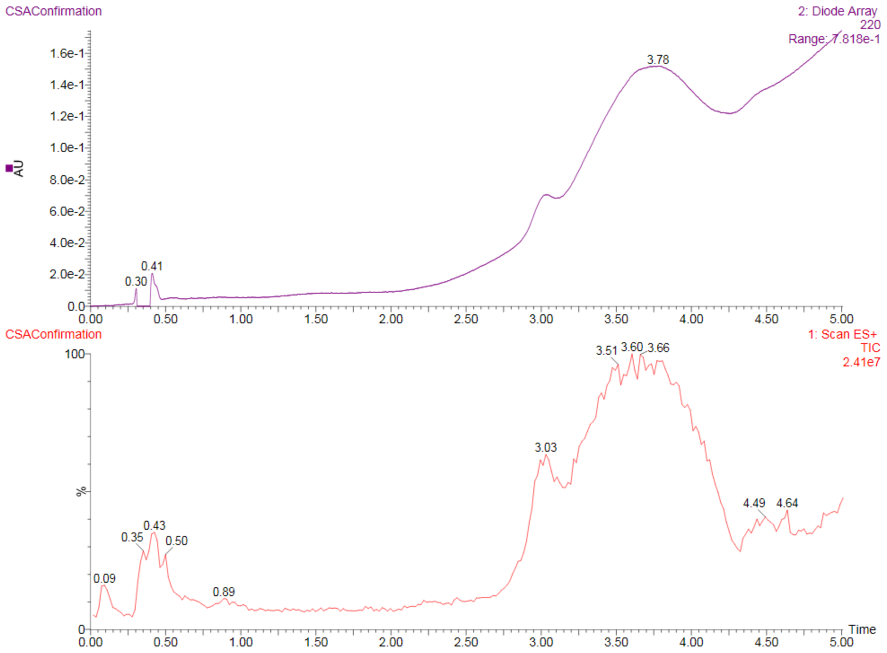


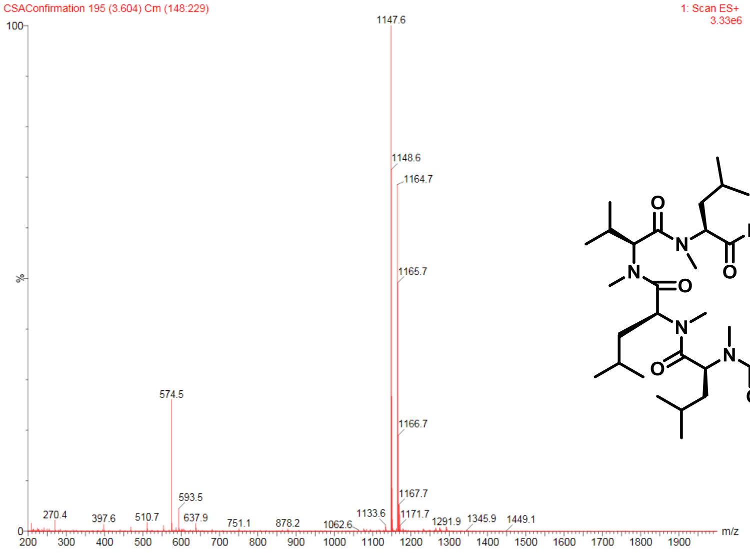


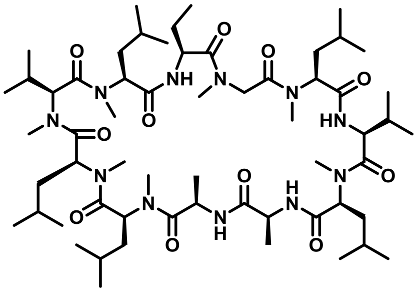
2.8.3.2 NMR Spectra for Cordyheptapeptide B in $CDCl_3$



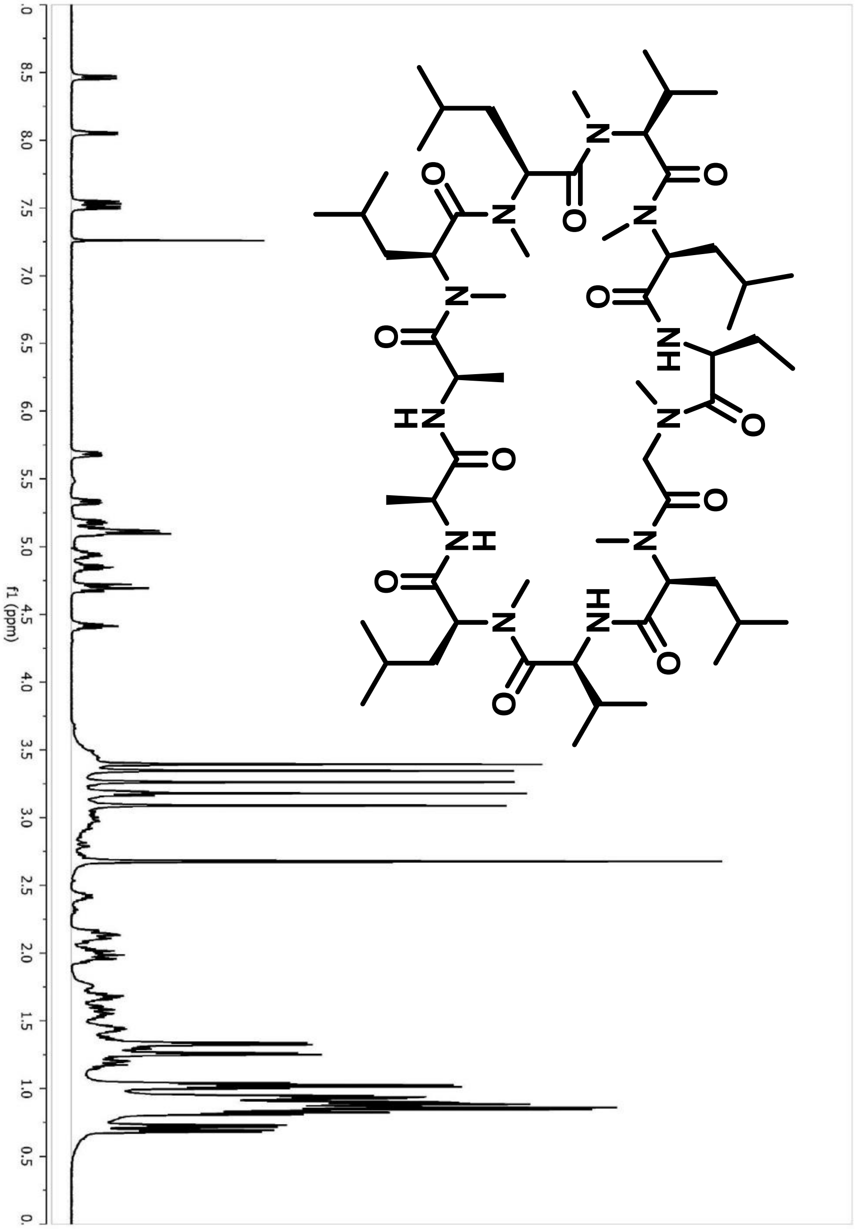
2.8.3.3 LCMS Spectra for Cyclosporine A BMT to Leucine



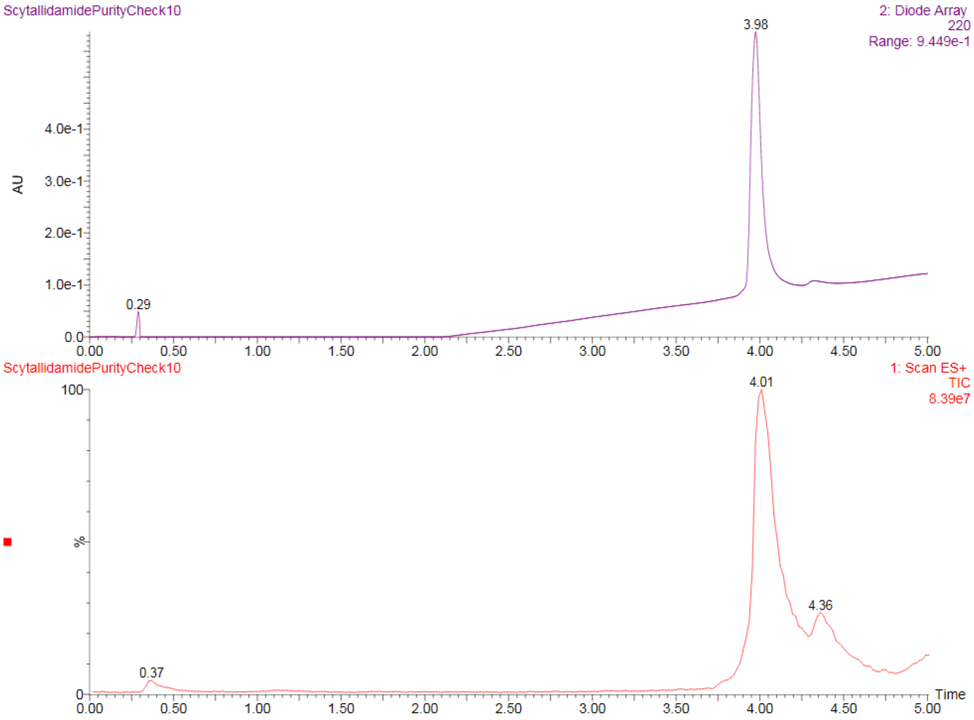


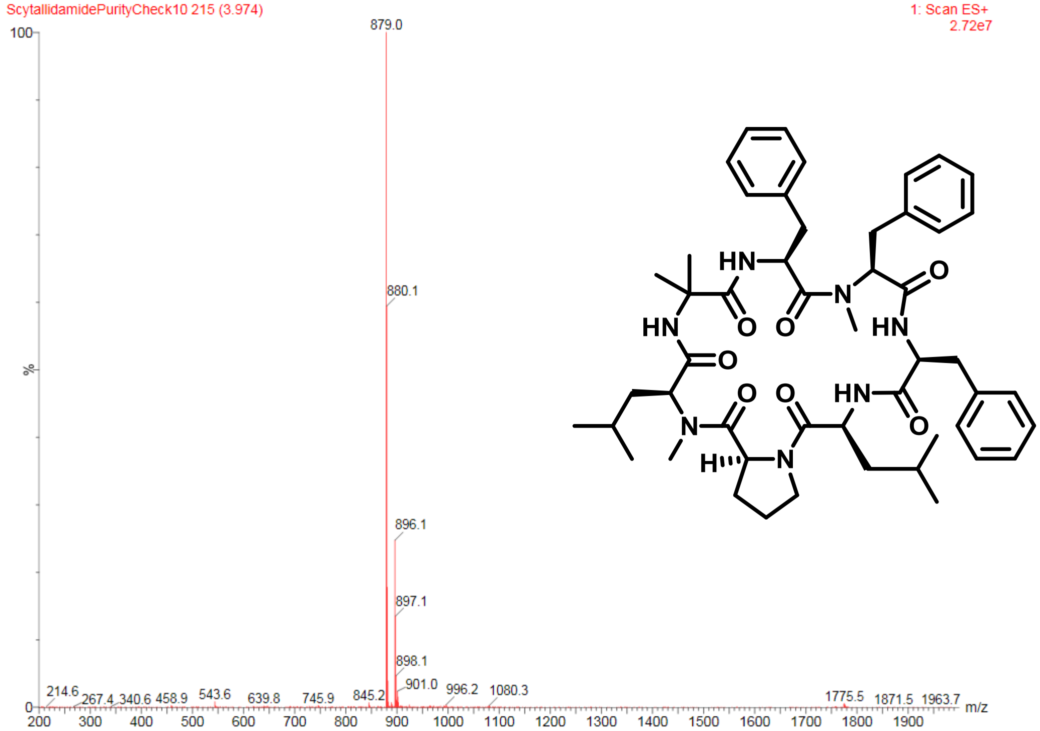


2.8.3.4 NMR Spectra for Cyclosporine A BMT to Leucine in $CDCl_3$

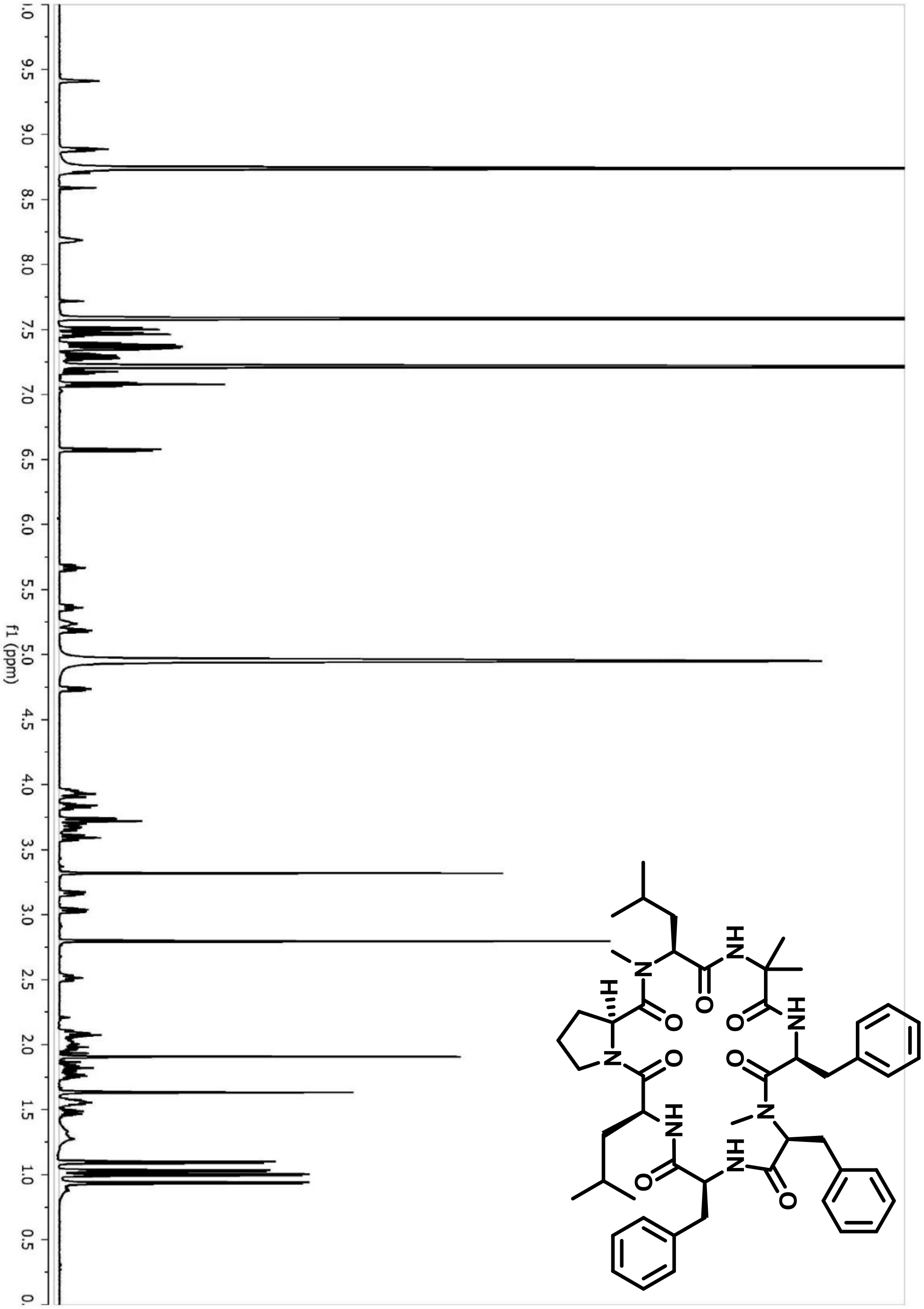


2.8.3.5 LCMS Spectra for Scytallidamide B

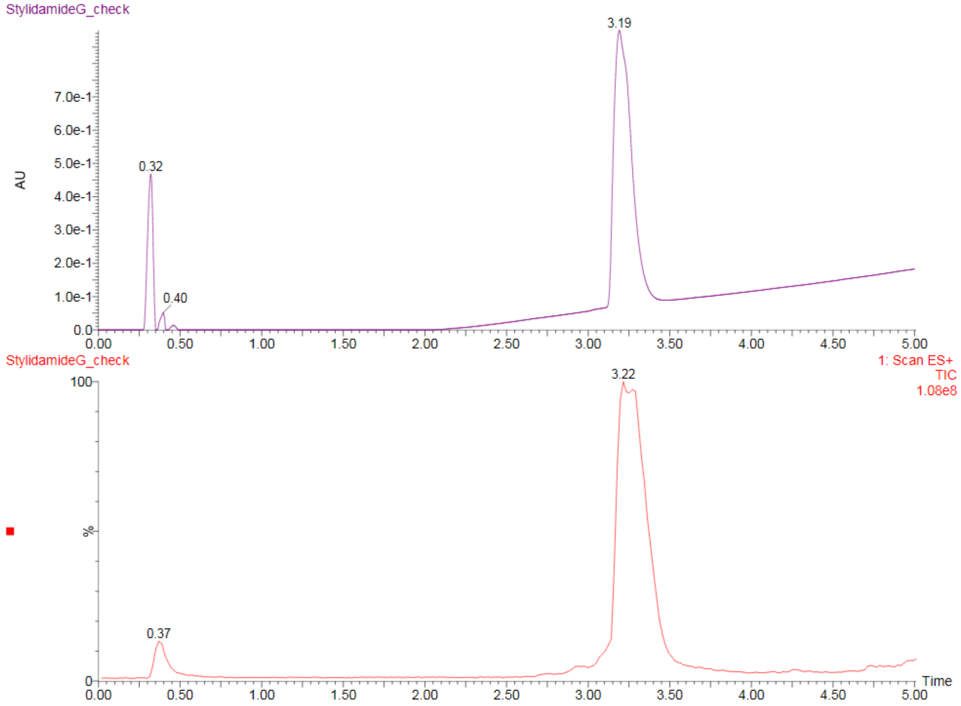

ScytallidamidePurityCheck10
2: Diode Array
220
Range: 9.449e-1
AU
3.98
0.29
4.0e-1
3.0e-1
2.0e-1
1.0e-1
0.0
0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
5.00
ScytallidamidePurityCheck10
1: Scan ES+
TIC
8.39e7
100
%
0
4.01
4.36
0.37
Time

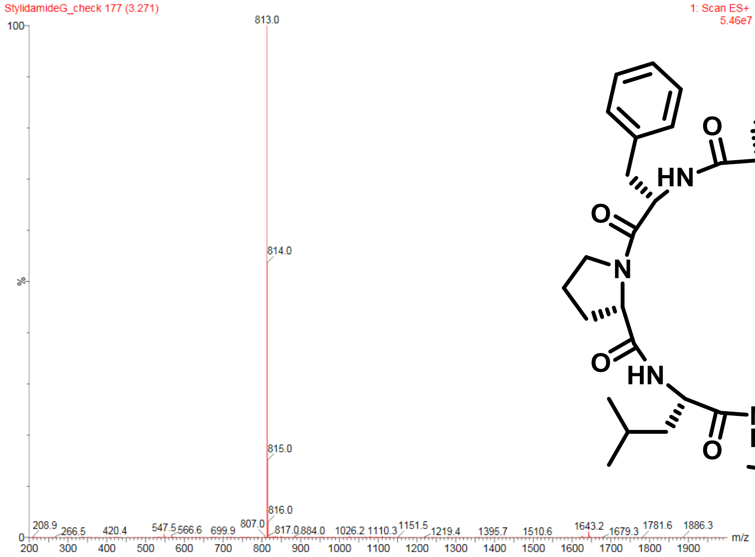

ScytallidamidePurityCheck10 215 (3.974)
1: Scan ES+
2.72e7
100
%
0
879.0
880.1
896.1
897.1
898.1
901.0
214.6
267.4
340.6
458.9
543.6
639.8
745.9
845.2
996.2
1080.3
1775.5
1871.5
1963.7
m/z
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
HN
HN
HN
HN
N
N
N
O
O
O
O
O
O
O
H

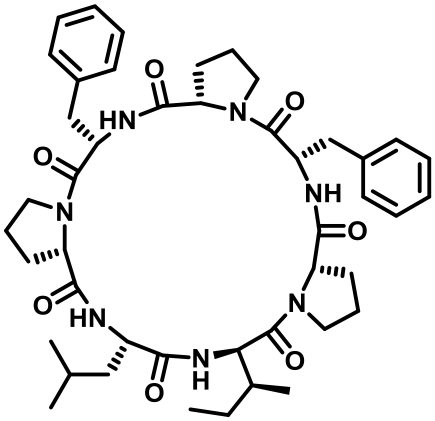
2.8.3.6 NMR Spectra for Scytallidamide B in Pyridine d5



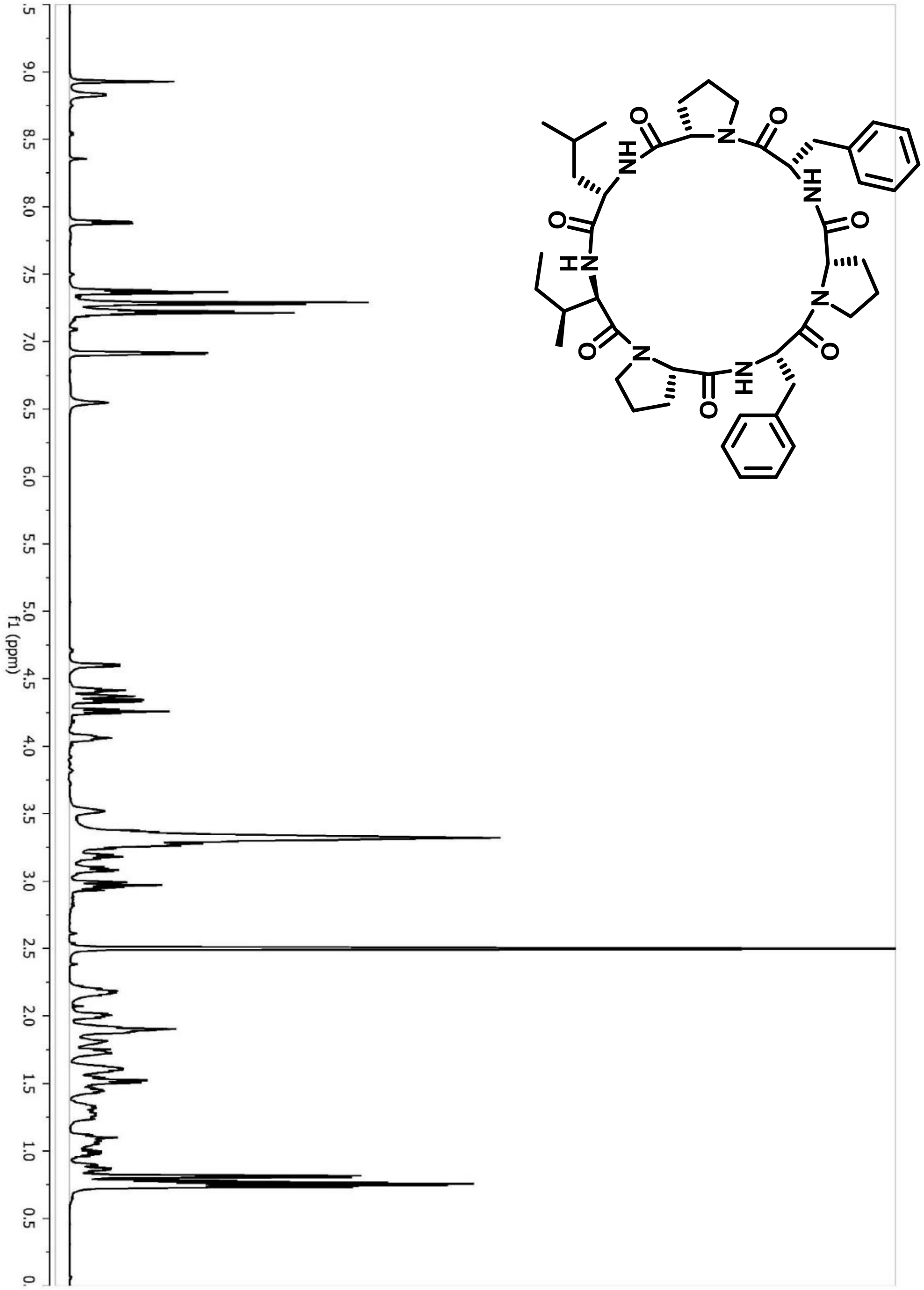
2.8.3.7 LCMS Spectra for Styllisamide G



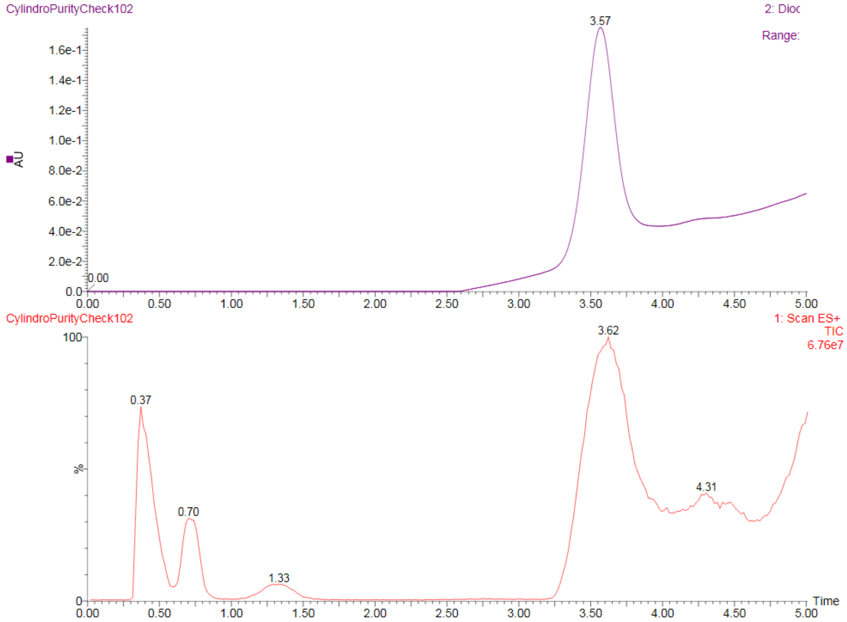


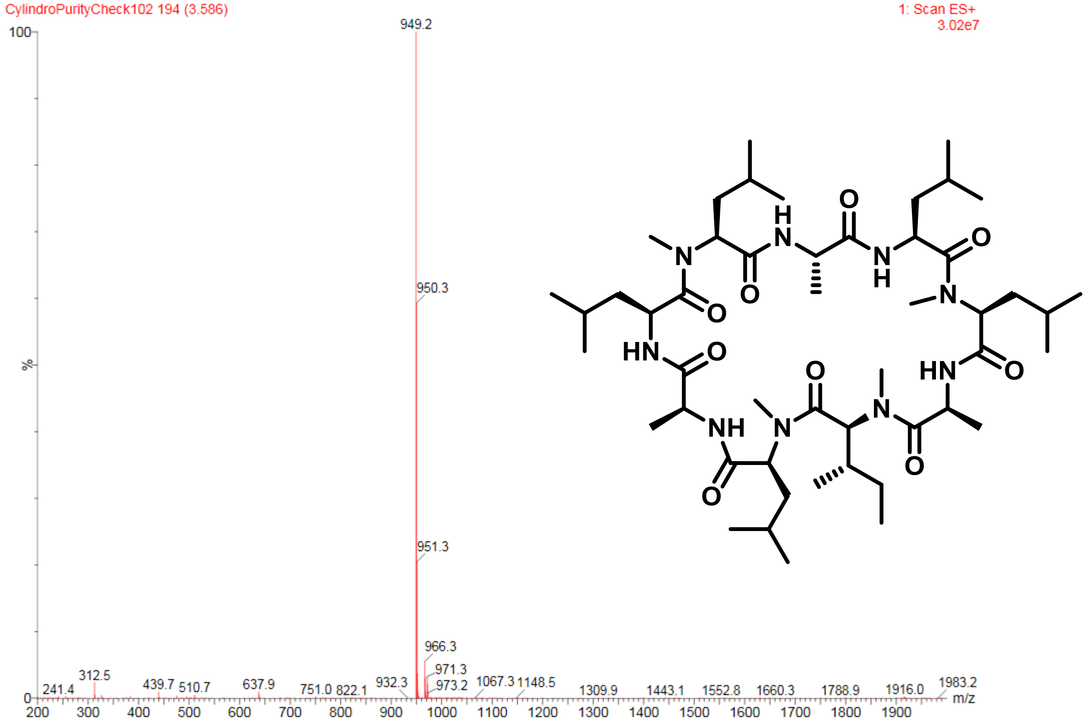

HN
O
N
O
NH
O
N
O
N
H
O
HN
O
N
O

2.8.3.8 NMR Spectra for Styllisamide G in DMSO d6

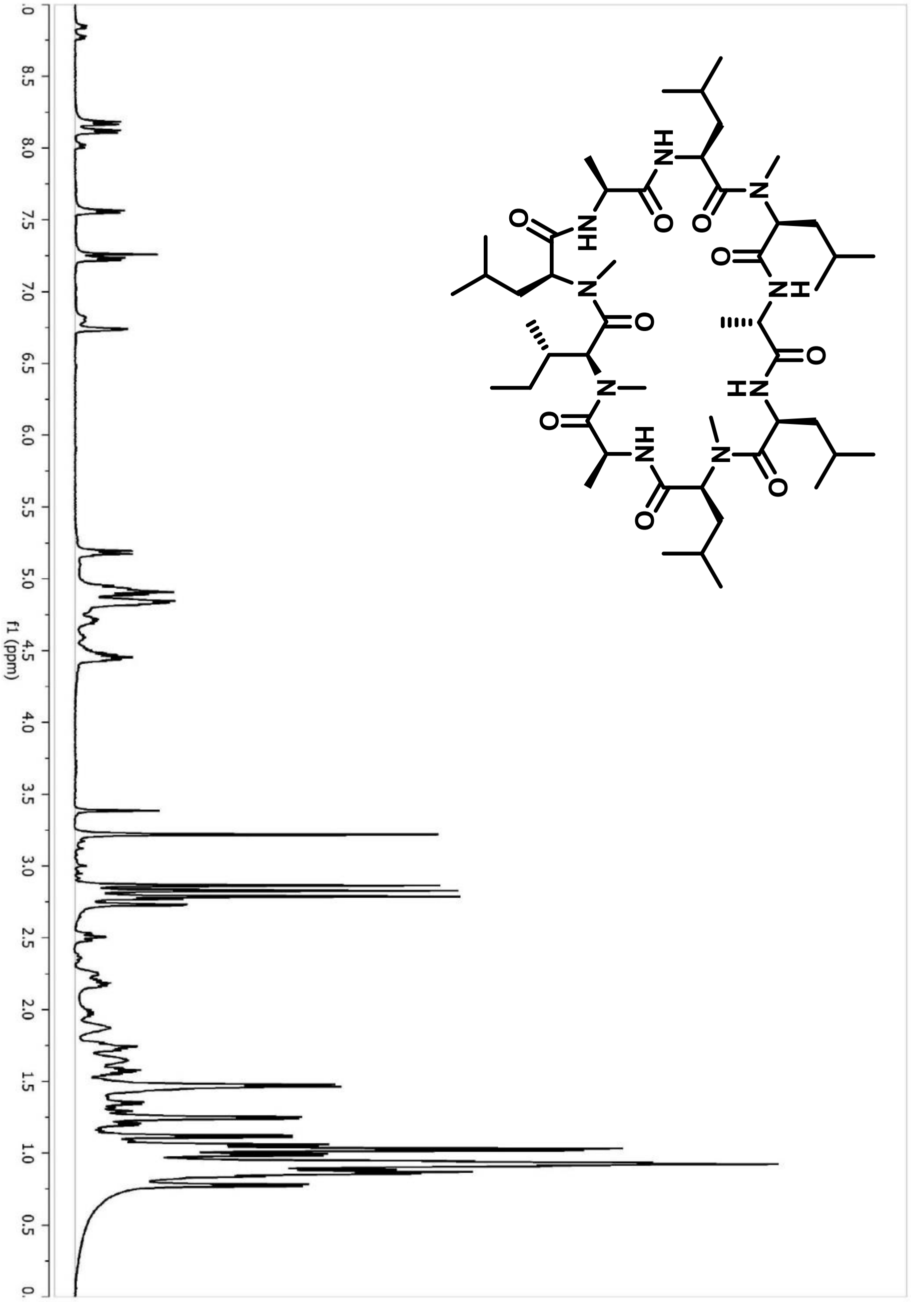


2.8.3.9 LCMS Spectra for Cylindrocyclin A





2.8.3.10 NMR Spectra for Cylindrocyclin A in $CDCl_3$



2.9 Bibliography

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **2001**, *46* (1), 3–26.

(2) Doak, B. C. C.; Over, B. B.; Giordanetto, F.; Kihlberg, J. *Chem. Biol.* **2014**, *21* (9), 1115–1142.

(3) Guimarães, C. R. W.; Mathiowetz, A. M.; Shalaeva, M.; Goetz, G.; Liras, S. *J. Chem. Inf. Model.* **2012**, *52* (4), 882–890.

(4) Villar, E. A.; Beglov, D.; Chennamadhavuni, S.; Porco, J. A.; Kozakov, D.; Vajda, S.; Whitty, A. *Nat. Chem. Biol.* **2014**, *10* (9), 723–731.

(5) Overton, E. *Vierteljahrsschr. Naturforsch. Ges. Zurich* **1899**, *44*, 88–135.

(6) Lieb, W. R.; Stein, W. D. *Nature* **1971**, *234* (50), 220–222.

(7) Lieb, W. R.; Stein, W. D. *J. Membr. Biol.* **1986**, *92* (2), 111–119.

(8) Xiang, T.-X. X.; Anderson, B. D. *J. Membr. Biol.* **1994**, *140* (2), 111–122.

(9) Xiang, T. X.; Anderson, B. D. *Biophys. J.* **1994**, *66* (March), 561–573.

(10) Leung, S. S. F.; Mijalkovic, J.; Borrelli, K.; Jacobson, M. P. *J. Chem. Inf. Model.* **2012**, *52* (6), 1621–1636.

(11) Leung, S. S. F.; Sindhikara, D.; Jacobson, M. P. *J. Chem. Inf. Model.* **2016**, *56* (5), 924–929.

(12) Bockus, A. T.; McEwen, C. M.; Lokey, R. S. *Curr. Top. Med. Chem.* **2013**, *13* (7), 821–836.

(13) Fouché, M.; Schäfer, M.; Berghausen, J.; Desrayaud, S.; Blatter, M.; Piéchon, P.; Dix, I.; Martin Garcia, A.; Roth, H.-J. *ChemMedChem* **2016**, *11* (10), 1048–1059.

(14) Kansy, M.; Senner, F.; Gubernator, K. *J. Med. Chem.* **1998**, *41* (7), 1007–1010.

(15) *Pharmacokinetic Optimization in Drug Research*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta: Zürich, 2001.

(16) Hewitt, W. M.; Leung, S. S. F.; Pye, C. R.; Ponkey, A. R.; Bednarek, M.; Jacobson, M. P.; Lokey, R. S. *J. Am. Chem. Soc.* **2015**, *137* (2), 715–721.

(17) Di, L.; Whitney-Pickett, C.; Umland, J. P.; Zhang, H.; Zhang, X.; Gebhard, D. F.; Lai, Y.; Federico, J. J.; Davidson, R. E.; Smith, R.; Reyner, E. L.; Lee, C.; Feng, B.; Rotter, C.; Varma, M. V; Kempshall, S.; Fenner, K.; El-Kattan, A. F.; Liston, T. E.; Troutman, M. D. *J. Pharm. Sci.* **2011**, *100* (11), 4974–4985.

(18) Nitsche, J. M.; Kasting, G. B. *J. Pharm. Sci.* **2013**, *102* (6), 2005–2032.

(19) Furukawa, A.; Townsend, C. E.; Schwochert, J.; Pye, C. R.; Bednarek, M. A.; Lokey, R. S. *J. Med. Chem.* **2016**, *59* (20), 9503–9512.

(20) Fujikawa, M.; Nakao, K.; Shimizu, R.; Akamatsu, M. *Bioorg. Med. Chem.* **2007**, *15* (11), 3756–3767.

(21) Wang, C. K.; Northfield, S. E.; Swedberg, J. E.; Colless, B.; Chaousis, S.; Price, D. A.; Liras, S.; Craik, D. J. *Eur. J. Med. Chem.* **2015**, *97*, 202–213.

(22) Mayer, P. T.; Anderson, B. D. *J. Pharm. Sci.* **2002**, *91* (3), 640–646.

(23) Goodwin, J. T.; Conradi, R. A.; Ho, N. F. H.; Burton, P. S. *J. Med. Chem.* **2001**, *44* (22), 3721–3729.

(24) Avdeef, A. *Absorption and Drug Development*; Wiley: Hoboken, New Jersey, 2003.

(25) Ghose, A.; Ghose, A.; Viswanadhan, V.; Viswanadhan, V.; Wendoloski, J.; Wendoloski, J. *J. Phys. Chem. A* **1998**, *102* (21), 3762–3772.

(26) Terry Mayer, P.; Xiang, T. X.; Niemi, R.; Anderson, B. D. *Biochemistry* **2003**, *42* (6), 1624–1636.

(27) Xiang, T. X.; Chen, J.; Anderson, B. D. *J. Membr. Biol.* **2000**, *177* (2), 137–148.

(28) Gutknecht, J.; Tosteson, D. C. *Science* **1973**, *182* (4118), 1258–1261.

(29) Nielsen, P. E.; Avdeef, A. *Eur. J. Pharm. Sci.* **2004**, *22* (1), 33–41.

(30) Wohnsland, F.; Faller, B. *J. Med. Chem.* **2001**, *44* (6), 923–930.

(31) Abraham, M. H.; McGowan, J. C. *Chromatographia* **1987**, *23* (4), 243–246.

(32) Filipe, H. A. L.; Salvador, A.; Silvestre, J. M.; Vaz, W. L. C.; Moreno, M. J. *Mol. Pharm.* **2014**, *11* (10), 3696–3706.

(33) Whitty, A.; Zhong, M.; Viarengo, L.; Beglov, D.; Hall, D. R.; Vajda, S. *Drug Discov. Today* **2016**, *21* (5), 712–717.

(34) Bockus, A. T.; Schwochert, J. A.; Pye, C. R.; Townsend, C. E.; Sok, V.; Bednarek, M. A.; Lokey, R. S. *J. Med. Chem.* **2015**, *58* (18), 7409–7418.

(35) Over, B.; Matsson, P.; Tyrchan, C.; Artursson, P.; Doak, B. C.; Foley, M. A.; Hilgendorf, C.; Johnston, S. E.; Lee, M. D.; Lewis, R. J.; McCarren, P.; Muncipinto, G.; Norinder, U.; Perry, M. W. D.; Duvall, J. R.; Kihlberg, J. *Nat. Chem. Biol.* **2016**, *12* (12), 1065–1074.

(36) Porter, C. J. H.; Trevaskis, N. L.; Charman, W. N. *Nat Rev Drug Discov* **2007**, *6* (3), 231–248.

(37) Georghiou, G.; Kleiner, R. E.; Pulkoski-Gross, M.; Liu, D. R.; Seeliger, M. A. *Nat. Chem. Biol.* **2012**, *8* (4), 366–374.

(38) Josephson, K.; Ricardo, A.; Szostak, J. W. *Drug Discov. Today* **2014**, *19* (4), 388–399.

(39) Kaplan, I. M.; Wadia, J. S.; Dowdy, S. F. *J. Control. Release* **2005**, *102* (1), 247–253.

(40) Henriques, S. T.; Huang, Y.-H.; Chaousis, S.; Sani, M.-A.; Poth, A. G.; Separovic, F.; Craik, D. J. *Chem. Biol.* **2015**, *22* (8), 1087–1097.

(41) Chu, Q.; Moellering, R. E.; Hilinski, G. J.; Kim, Y.; Grossmann, T. N.; Yeh, J. T.-H.; Verdine, G. L.; Yeh, T.; Verdine, G. L. *Medchemcomm* **2014**, *6* (c), 111–119.

(42) Trinh, T. B.; Upadhyaya, P.; Qian, Z.; Pei, D. *ACS Comb. Sci.* **2016**, *18* (1), 75–85.

(43) White, T. R.; Renzelman, C. M.; Rand, A. C.; Rezai, T.; Mcewen, C. M.; Gelev, V. M.; Turner, R. A.; Linington, R. G.; Leung, S. S. F.; Kalgutkar, A. S.; Bauman, J. N.; Zhang, Y.; Liras, S.; Price, D. A.; Mathiowetz, A. M.; Jacobson, M. P.; Lokey, R. S. *Nat. Chem. Biol.* **2011**, *7* (11), 810–817.

3 Natural Products: Are We Close to the End?

3.1 Foreword

The following work was completed as collaboration between all the following authors:

Cameron Pye[1], Matthew Bertin[2], R Lokey[3], William Gerwick[2], Roger G Linington[4]

[1] University of California Santa Cruz, [2] University of California San Diego, [3] University of California, Simon Fraser University

Submitted to Proceedings of the National Academy of Sciences of the United States of America

3.2 Abstract

Understanding of the capacity of the natural world to produce secondary metabolites is important to a broad range of fields, including drug discovery, ecology, biosynthesis and chemical biology among others. Both the absolute number and the rate of discovery of natural products have increased significantly in recent years. However, there is a perception and concern that the fundamental novelty of these discoveries is decreasing relative to previously known natural products. This study presents a quantitative examination of the field from the perspective of both number of compounds and compound novelty using a dataset of all published microbial and marine-derived natural products. This analysis aimed to explore a number of key questions, such as how the rate of discovery of new natural products has changed over the past decades, how the average natural product structural novelty has changed as a function of time, whether exploring novel taxonomic space affords an advantage in terms of novel compound discovery and if it is possible to estimate how close we are to having described all of the chemical space

covered by natural products. Our analyses demonstrate that most natural products being published today bear structural similarity to previously published compounds, and that the range of scaffolds readily accessible from nature is limited. However, the analysis also shows that the field continues to discover appreciable numbers of natural products with no structural precedent. Together these results suggest that the development of innovative discovery methods will continue to yield compounds with unique structural and biological properties.

3.3 Introduction

Many of today's small molecule therapeutics trace their origins to natural products, estimated variably as providing or inspiring the development of between 50-70% of all agents in clinical use today.[1] And whereas the natural environment is frequently identified as a rich source of unique chemical diversity for pharmaceutical lead compound discovery,[2,3] the rediscovery of known natural product structures is an increasing challenge for the field.[4,5,6] The concern in some quarters is that natural product diversity accessible by "top-down" approaches (e.g. bioassay- or chemical signature-guided isolation) has been largely exhausted, and that existing discovery models are no longer capable of delivering novel lead compo7unds.[7] In this regard, it has been proposed that 'bottom-up' approaches (e.g. genetic information-driven natural product isolation) have the capacity to access the unexpressed genetic potential of microorganisms, and can thus lead to a "renaissance" in the field of natural products.[8] Ultimately, the veracity of these propositions will be revealed in their relative success records, and perhaps the more reasoned view is that the success of the discipline is best achieved by a diversity of approaches.

Antibiotic discovery is an example where there is particular concern over the ability of top-down natural products investigations to yield fundamentally new classes of agents.[9,10] Almost all of the early antibiotic scaffolds were derived from natural sources, and there have been no new clinically-approved natural product-based antibiotics discovered for over 30 years.[1,11] Even those that have entered the market more recently, such as daptomycin15 and tiacumicin B1,16,17 have their discovery origins back in the 1980's. This lack of discovery of new natural product-based antibiotics, in spite of substantial effort in this area by both academia and industry, raises the question of whether all of the clinically relevant natural product-based antibiotics have already been discovered. This would of course present a quite terrifying prospect for patients and the biomedical community alike.

In order to provide a perspective on the issue of natural product structural diversity, we have performed a series of analyses on the structures of all microbial and marine-derived natural products published during the period 1941 – 2015. These analyses were designed to examine the rates of natural product discovery over time as well as the relationship between year of discovery and structural novelty. In this regard, such an analysis provides a description of the current state of natural products research and its ability to potentially yield new classes of therapeutic agents in the future.

To accomplish this objective, we assembled a dataset comprising all published microbial and marine-derived natural products from the period 1941- 2015. The data for the period 1941 - 2011 are contained in the commercial database AntiMarin. The data for the period 2012 - 2015 were assembled for this study through manual curation of all published articles from a large panel of journals in the chemistry and chemical biology arena (for details of data set construction see Supporting Information)

3.4 Results and Discussion

3.4.1 Exploring Trends in Chemical Diversity

How has the rate of discovery of new natural products changed as a function of time?

Initially we examined the rate of natural product discovery as a function of time. Figure 3-1A shows that the number of compounds being published from these sources increased dramatically from relatively few compounds per year in the 1940's to an average of ~1600 per year over the last two decades. The rate of increase in numbers of newly reported compounds was greatest from the 1970's through the mid-1990's, and has remained relatively constant since then. There are likely a number of factors that contributed to the dramatic rise in natural product discoveries in the last part of the previous century. In the 1940's this was still a new field, with relatively few practitioners. Moreover, the analytical tools available were very limited, meaning that structure determination was extremely challenging and time consuming. However, after the success of early therapeutic leads from nature, particularly the early antibiotics, the number of research groups in this area increased significantly. With the advent of better instrumentation (high performance liquid chromatography, nuclear magnetic resonance spectroscopy, mass spectrometry) and the invention of two-dimensional NMR methods in the mid-1980's, the process of compound isolation and structure determination improved greatly, leading to a steady increase in the annual number of published structures from nature.

It is less clear why the number of published compounds has continued to increase or remain steady over the past two decades, even as the pharmaceutical industry has largely exited the natural products arena.[12,13] This continued productivity might reflect the increasing globalization of natural products research, with most countries that possess significant academic research infrastructure also supporting vibrant natural products research programs. For example, there has been a considerable increase in productivity in natural products research from China, Korea, Brazil and India during this period.20 From a global perspective, the numbers indicate that there is still a healthy focus on natural products research. Moreover, the robust productivity of these efforts demonstrates that the natural world continues to provide large numbers of new and bioactive molecules year on year.

Has the rate of novel natural product discovery changed in recent years?

Although the raw number of new molecules published per year provides some information about the productivity of the natural products research community, it provides no information about the structural novelty of the compounds being reported. In order to explore the structural relationships within this set of natural products and thus gain insight into this question, we calculated Tanimoto similarity scores between all molecule pairs (for a discussion of chemical similarity scoring methods, see Supporting Information).[14] We then separated the compounds into bins based on year of discovery, and determined the highest Tanimoto score between each molecule in a given bin and all of the molecules published in previous years. This analysis provides a measure of the structural novelty of a given compound at the time at which it was discovered. Taken together, these results provide one metric for evaluating both trends in structural novelty (based on median values) and diversity (based on median average deviation).

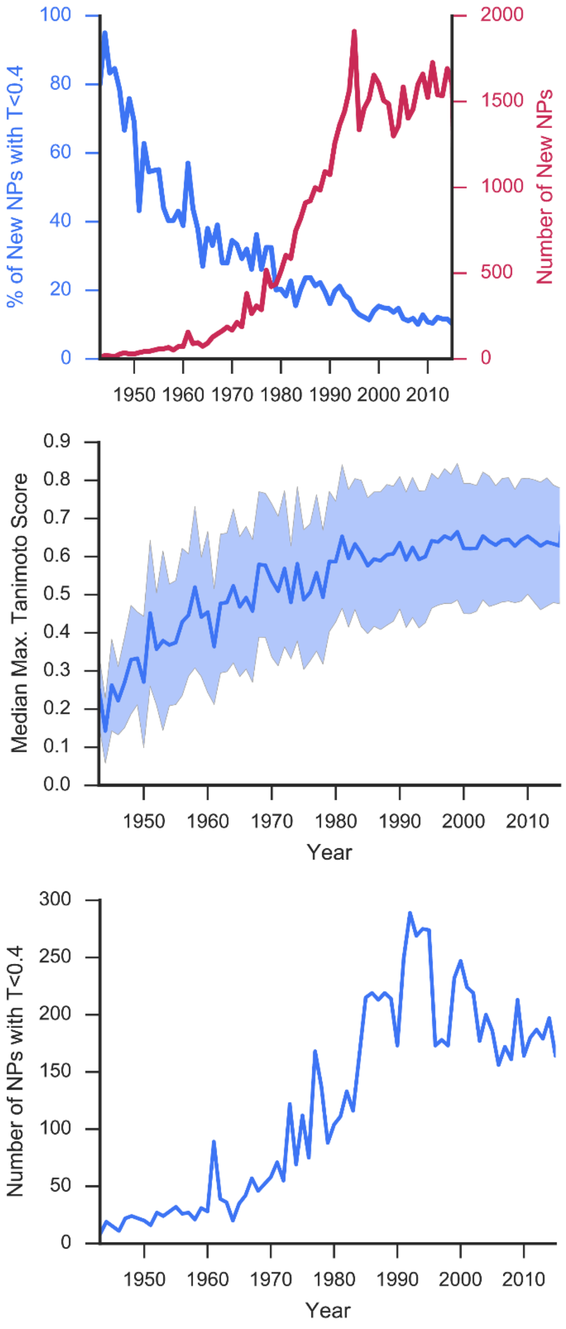


Figure 3-1 Examining Structural Diversity. A) Number of compounds published per year and rate of novel compound isolation as a percentage of total natural product isolation. B) Median maximum Tanimoto scores as a function of time. Median average deviation shown as shaded blue region. C) Absolute number of low similarity compounds (T < 0.4) per year.

As shown in Figure 3-1B, it is clear that in parallel with the steady increase in number of compounds reported per year, median maximum Tanimoto scores also increased rapidly from the 1950’s to 1970’s (Figure 3-1B, blue line). The rate of increase tapered during the 1980’s and 1990’s to reach a plateau at ~0.65 by the mid-1990’s, a value where it remains today. A cursory review of these data might suggest that the field of natural products is no longer discovering novel chemical entities, and that natural products chemical space has been largely described. However, contained within this huge compendium of structures are many examples of fundamentally unique molecules, often with unprecedented structural and/or functional attributes. Therefore, in addition to considering the median Tanimoto score distribution, it is also important to evaluate the distribution of molecules with low similarity scores ($T < 0.4$) (Figure 3-1C). This analysis is interesting as it reveals that the number of novel compounds increased in parallel to the increase in number of compounds reported through the mid-1990's, followed by a steady or slightly decreasing rate of novel compound discovery in recent years. It is impressive and significant that the absolute number of molecules with low similarities remains high over this most recent period, despite the ever-increasing bar set by the addition of thousands of new structures to the dataset each year.

Overall, this analysis indicates that the discovery rate of new molecular architectures among natural products has increased since the origins of this field, and has remained at a significant rate despite the ever-increasing number of published natural products (Figure 3-1C). However, it should also be noted that an increasing number of the total reported natural products do have structural precedent in the literature, and thus constitute derivative structures.

To ensure that the shape of the curve in Figure 3-1B was not simply an artifact of comparing a given compounds against an increasing corpus of other molecules a null/random dataset was generated. To do this the data set was shuffled (ie date and repopulated into bins the same size as the original dataset year bins. This effectively randomized dataset uses the same compounds as the original and as such should serve as a an effective null. The same retroactive maximum tanimoto analysis was performed on the shuffled dataset and the shape of the curve is quite different from the real dataset (Figure 3-2 red line). The curve starts lower but and increases more rapidly eventually plateauing at 1.0. This is likely in part because compounds that are effectively identical to the fingerprinting scheme are not compared in the original dataset because these analogues are frequently reported in the same publication and are thus not compared in the original analysis since they do not precede each other. The plateau of the original dataset ~0.6 is likely due to the pressures for novelty that are exerted by the field. In either case the randomized / null dataset is substantially different than our results implying that the observation is not just an artifact of the analysis.

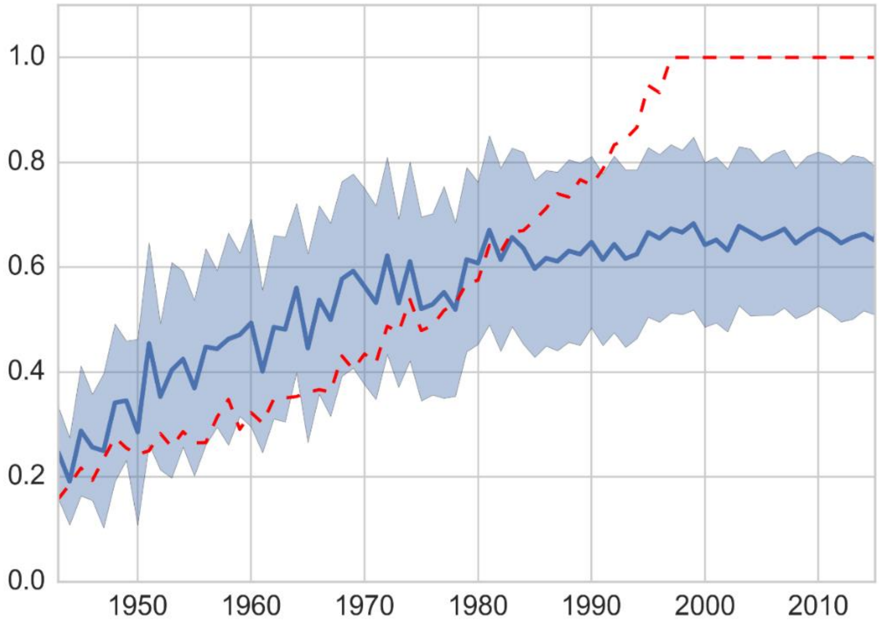


Figure 3-2 Comparison of the time dependent tanimoto scoring vs a randomly shuffled version of dataset (broken red line).

Overall, structurally unique compounds represent a decreasing percentage of the total number of compounds isolated from natural sources (Figure 3-1A). Therefore, if structural novelty is an important and valued component of natural products research, a central question for the field becomes how do we prioritize the discovery of these unique molecules from within this large pool of natural products with known structural scaffolds?

3.4.2 Source Diversity vs. Structural Diversity

Does exploring novel taxonomic space afford an advantage in terms of novel compound discovery?

It has long been a tenet of natural products discovery research that examination of unexplored and unusual source organisms, or those from unique environments, provides opportunities for finding novel natural products. Recent examples of such habitats include caves,[15] hydrothermal vents,[16] Arctic[17] and Antarctic waters,[18] plant endophytes,[19,20] and vertebrate[21–23] and invertebrate[24] microbiota. To examine the relationship between organism type and chemical diversity we sub-divided the dataset into sub-groups within two major designations (bacterial and marine).

As an example of the impact of studying a unique type of source organism on structural novelty, we first examined the compounds in the cyanobacterial sub-group. Cyanobacteria were selected for this analysis because i) they are morphologically distinct from other organisms and not easy to mis-assign in terms of fundamental taxonomic classification ii) they have been show to possess the genes required to produce many of the classes of compounds associated with this phylum,[25,26] limiting the risk that these molecules are actually produced by endosymbionts from other phyla iii) there are a sufficient number of research groups studying cyanobacteria that the results are not likely to be biased by research strategies of an individual research group[27,28] iv) the investigation of cyanobacterial metabolites significantly post-dates the exploration of other sources of natural products, providing an ideal model for examining the impact of exploring new biological space on chemical novelty.

In the first plot (Figure 3-3A), temporal variation in median Tanimoto scores for all compounds (blue line) were compared to the median Tanimoto scores between only the cyanobacterial metabolites (red line). As expected, the trend of the cyanobacterial compound data shows that study of a new source organism initially yielded compounds with little similarity to one another. Over time these median Tanimoto values gradually increase, suggesting that the easily accessible chemical diversity from cyanobacteria was described during this period (~1980 - 2000). Ultimately these values have come to match those observed for the broader set of natural products (blue line), indicating that the study of cyanobacterial chemistry has now reached maturity.

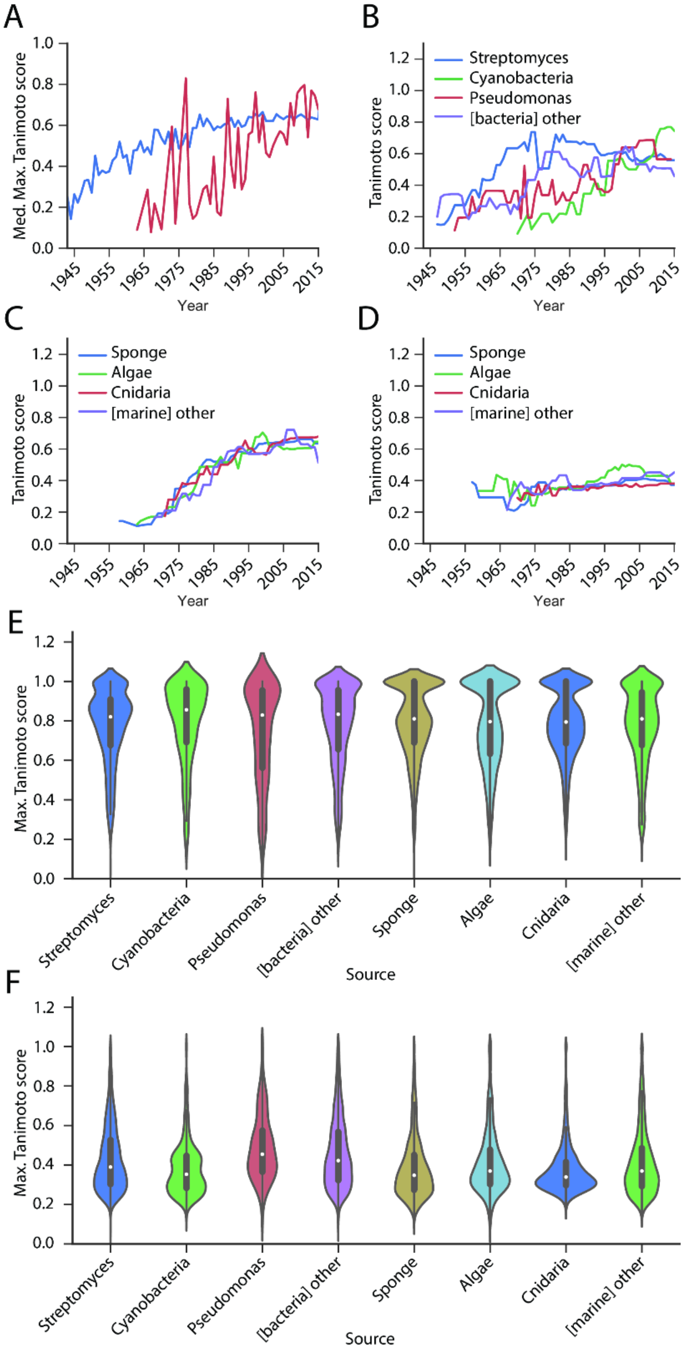


Figure 3-3. Examining Source Diversity. A) Plot of median maximum Tanimoto score by year for the full dataset (blue) and the intra-sub-group values for the cyanobacterial sub-group (red) B) Plot of intra-sub-group median maximum Tanimoto scores by year for bacterial sub-groups C) Plot of intra-sub-group median maximum Tanimoto scores by year for marine sub-groups D) Plot of extra-sub-group median maximum Tanimoto scores by year for marine sub-groups E) Violin plots for intra-sub-group median maximum Tanimoto scores for bacterial and marine sub-groups F) Violin plots for extra-sub-group.

median maximum Tanimoto scores for bacterial and marine sub-groups.

To examine the prevalence of this phenomenon in natural products discovery we plotted the temporal progression of in-class similarity scores for a selection of sub-groups from the dataset. Looking at the sub-groups for the bacterial designation (Streptomyces, Pseudomonas, Cyanobacteria, other; Figure 3-3B) a number of interesting observations emerge. Firstly, as expected, all plots have upward trends with values in later years in the region 0.6, suggesting that there is a limited period in which unique chemistry can be easily found within a given source type. This phenomenon is even more pronounced among the sub-groups from the marine environment. Figure 3-3C shows that compound similarities follow very similar trends, regardless of source sub-group, demonstrating that we have a mature understanding of the predominant chemistries likely to be encountered from these sources. By contrast, the data for bacteriala sources (Figure 3-3B) do not follow such strong trends. In the case of Streptomyces the number of compounds is similar to those from the Porifera sub-group (6547 vs 7263) and the taxonomic classification is narrower (genus vs. phylum). However, the median maximum Tanimoto scores are more variable, with values in the past 30 yeaars decreasing consistently. There are many factors that could contribute to this trend including application of new discovery strategies, higher chemical diversity, differences in program objectives or discovery models among researchers in the two sub-disciplines. However, what is apparent from these results is that at present the average molecule reported from bacteria of the genus Streptomyces is less similar to other Streptomyces compounds than is the case with compounds within the sub-group Porifera.

Interestingly, when compounds in a given marine sub-group were compared against compounds from all other sub-groups in the marine set the median Tanimoto scores remained low, regardless of year (Figure 3-3D). This supports the long-held understanding in the marine natural products community that the chemistries derived from a given organism type are often fundamentally different to the chemistries encountered from all other organisms in that environment. These results suggest that novel sources of natural products have been, and remain, an important and productive source of novel chemical diversity, albeit with a limited period of expected novel compound discovery. The counterpoint to this conclusion is that, in the absence of significant innovation in discovery approaches, there are diminishing returns in terms of the discovery of fundamentally new chemical diversity from continued investigations of the same classes of organisms.

Finally, to examine the absolute chemical diversity within each sub-group we calculated the maximum Tanimoto score for each compound compared to either other compounds within that sub-group (Figure 3-3E), or to other members of the source type (bacterial or marine; Figure 3-3F), irrespective of year of discovery. Figure 3-3E illustrates that most sub-groups contain moderate to large numbers of compounds with high similarities to one or more other compounds in that class. This is particularly apparent for Porifera, algae and Cnidaria, all of which have large 'hammerhead' distributions (Figure 3-3E). However, when compared to the other sub-groups within that source type (e.g. Cnidaria vs all other marine compounds; Figure 3-3F) the distribution of maximum Tanimoto scores is centered around very low values, confirming the previous observation that the chemistries from these source organisms are typically not found elsewhere in the marine world.

3.4.3 Evaluating the Chemical Space Occupied by Natural Products

How much of the ‘natural product-like’ chemical space is actually occupied by natural products?

Recently, there have been a number of efforts to describe natural product chemical space, and to use this space in various ways including as a boundary for designing ‘natural product like’ synthetic screening libraries.[29,30] For a given biosynthetic class of natural products, a very large number of theoretical molecules can be created from primary building blocks, such as amino acids, sugars, acetate and propionate, mevalonate etc. Using these diverse and often chiral components, natural product libraries should therefore exceed synthetic libraries in terms of structural diversity of chemical scaffolds.

One way to explore this hypothesis is to examine the chemical diversity within classes of compounds that are easily and accurately characterized by boundary conditions. To explore this idea we examined the chemical diversity of all currently published cyclic tetrapeptides from our dataset. Cyclic tetrapeptides were chosen because these are relatively easy to identify from the dataset, and the building blocks are well defined. There are numerous ways to estimate the theoretical number of molecules that are possible within this class; thus, to simplify the analysis, we considered only the 20 proteinogenic amino acids as possible building blocks. Although this oversimplifies the analysis (e.g. non-proteinogenic and D-amino acids are excluded), it provides at least a minimum theoretical limit for possible structural diversity. With four positions in the peptide and 20 amino acid building blocks there are 40,110 possible molecules that can be produced. This is less than the 204 possible compounds one might expect, because of the rotational symmetry of some conceivable products. However, to our surprise, examination of the natural product database developed in this study revealed just 65 cyclic tetrapeptides, the majority of which fall into just four structural classes (Figure 3-4A and Figure 3-8). The largest class, containing 27 members, incorporates a residue rarely encountered elsewhere in nature containing an alkyl chain terminating in either an epoxy ketone or an ethyl ketone; trapoxin A[31] and apicidin A[32] are well known members of this family. Compounds in this class have been isolated by 20 different research groups over a 23-year span, suggesting that that this cluster is large because this structural motif is relatively widespread in the environment, rather than having been the subject of intense investigation by a small number of specialized research teams.

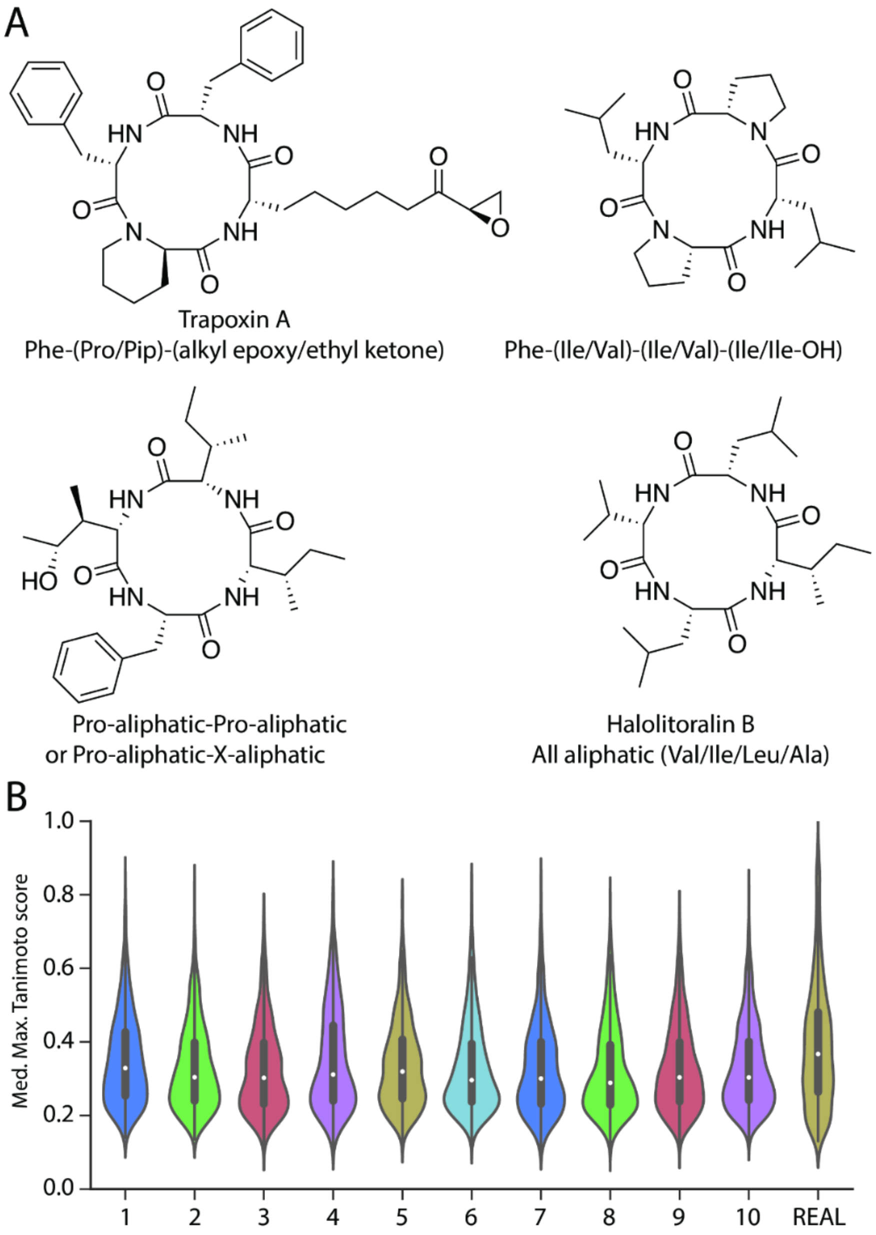


Figure 3-4 Theoretical vs. Actual Structural Diversity. A) Examples of the four major classes of cyclic tetrapeptides found in nature B) Violin plots indicating the distribution of Tanimoto scores between all members of 65 randomly selected cyclic peptides (10 trials, lanes 1 - 10) and between all 65 cyclic tetrapeptides from our natural product dataset (lane REAL).

The precise reasons for this disconnection between theoretical and observed chemical diversities are unknown, but are likely due to the combined factors of physical organic chemistry (i.e. some of the conceivable molecules are not structurally plausible or are difficult to form) and chemical ecology (i.e. molecules produced in nature must confer a competitive advantage to the producing organism). Although not explored further in this current report, this trend that only a few representatives in a given structure class are produced in nature is a general phenomenon that holds true for many of the compound classes published to date. Therefore, we conclude that although the theoretical chemical space offered by natural products is very high, the number of actual compounds discovered within a given chemical class is quite low, despite all of the different screening and isolation approaches that have been used by the field in the preceding 70 years.

Furthermore, it is a general observation that compounds that are produced within a class often share many structural features, indicating that the compounds produced are not a random sub-set of the possible molecules, but rather possess moderate diversification within a large number of highly refined structural constraints. To explore this phenomenon within the cyclic tetrapeptide example we calculated the Tanimoto scores between all members of a set of 65 randomly selected cyclic tetrapeptides from among the 40,110 possible unique structures. Comparing the results from ten independent trials of this experiment with the distribution from the naturally occurring molecules (Figure 3-4B) we observed much higher frequencies of structural relatedness within the naturally occurring compounds than within any of the ten randomly selected compound sets. This result supports the idea that, at least in the case of cyclic tetrapeptides, structural diversity in nature is centered around a select set of key scaffolds, rather than being randomly distributed throughout the available chemical space. An inescapable conclusion of this observation is that selection pressures at work in the natural world significantly limit the scope of structural diversity created in most compound classes.

We next turned our attention to the question of how the 52,395 unique molecules in our dataset compared to one another in terms of structural relatedness. In order to allow molecules that are substructures of larger scaffolds to achieve high similarity scores we employed a different distance metric for clustering molecular structures. Tversky scores (α=0.1,β=0.9) were calculated for each pair of molecules in each direction (i,j & j,i) and averaged to afford a single similarity score.[33] Edges between nodes in the network represent averaged Tversky scores greater than or equal to 0.8. Finally, structure graphics were overlaid using the ChemViz Cytoscape plugin. In this way, we created a network diagram that describes the structural relatedness of all 52,395 compounds within the database (Figure 3-5).

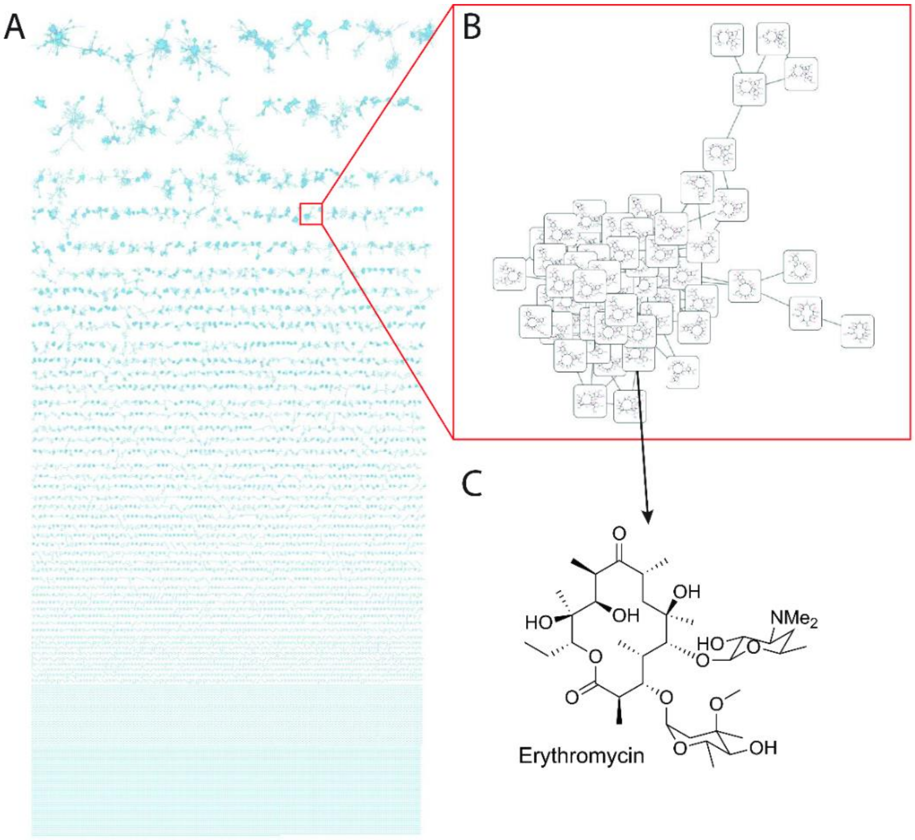


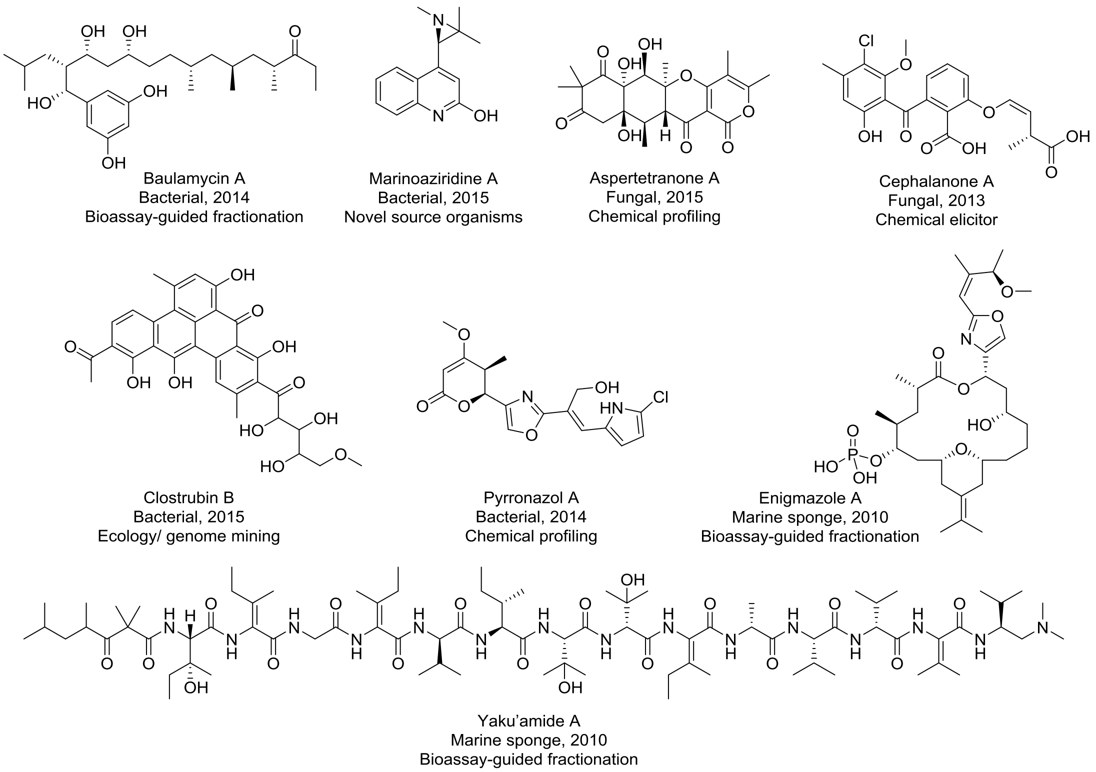
Figure 3-5 **Cluster Analysis for Natural Product Diversity. A)** Network diagram displaying all molecules as clusters based on Tversky structural similarities. Compounds with no structural similarity partners appear as singletons in the bottom region of the figure. **B)** Expansion of region of network diagram indicating erythromycin compound class. **C)** Example structure from erythromycin cluster.

This analysis revealed that the dataset contained 6,414 clusters comprised of two or more compounds, which together accounted for 40,229 compounds. In addition, the network contained a further 12,166 compounds that did not belong to any cluster. Therefore, 76.8% of the chemical space occupied by published natural products from these sources is described by less than 6,500 scaffolds, indicating that these known NP scaffolds occupy a relatively narrow percentage of the total available ‘natural product like’ chemical space. This observation suggests that either the selective pressures on the molecular evolution of NPs from nature are convergent for particular compound classes, or that we have not yet developed the technologies required to access natural compounds from the vast majority of the chemical space they are predicted to occupy.

How close are we to having described all of the chemical space covered by natural products?

This key question is extremely difficult to answer definitively without further improvements in a number of areas, such as: a) enhancing the ability to accurately predict structural classes from genome sequence data, b) refining the capacity to find biosynthetic genes from genomic DNA, and c) developing the facility to derive complete whole genome sequencing data from complex environmental samples. Nevertheless, it is clear from the temporal trends in similarity values discussed above that a significant number of molecules published in recent years bear close structural similarity to established scaffolds.

On the other hand, over 23% of the NPs in the database have low structural similarity to all other compounds, and therefore appear as singletons in the network diagram. Examples of compounds with low similarity scores are shown in Figure 3-6. These compounds, all of which possessed maximum Tanimoto scores of 0.4 or less at the time of discovery, encompass a broad array of biosynthetic origins, structural complexities, source organisms and discovery methods, suggesting that there remains significant opportunity for novel compound discovery across a broad range of sources and scientific approaches. There are many unique and structurally novel compounds in the natural products world, and the rate of discovery of these unusual structures is not changing significantly, despite the presence of a large and growing canon of known NP structures (Figure 3-1C). While there appears to be much left to discover that is truly novel, significant innovation will be required to access these in an efficient manner and retain the impressive historical rate of novel compound discovery from the natural world.



Baulamycin A
Bacterial, 2014
Bioassay-guided fractionation

Marinoaziridine A
Bacterial, 2015
Novel source organisms

Aspertetranone A
Fungal, 2015
Chemical profiling

Cephalanone A
Fungal, 2013
Chemical elicitor

Clostrubin B
Bacterial, 2015
Ecology/ genome mining

Pyrronazol A
Bacterial, 2014
Chemical profiling

Enigmazole A
Marine sponge, 2010
Bioassay-guided fractionation

Yaku'amide A
Marine sponge, 2010
Bioassay-guided fractionation

Figure 3-6 Examples of natural products with low (<0.4) Tanimoto scores, indicating compound name, source, year of discovery and isolation method.

3.4.4 Limitations of this Analysis/ Points to Consider

It is important to recognize that novel structure, while scientifically intriguing, is not necessarily the primary driver of interest in natural products. That position arguably rests in the realm of biology. It is the biological roles of natural products that are the ultimate source of their value to human society and the environment. Moreover, it is widely known that small and subtle changes in a molecule's structure have the capacity to alter it from being "inactive" to being "exquisitely potent". Therefore it must be recognized that there remains substantial value in known and derivative natural product structures, provided that these structures possess unique attributes from a biological perspective.

Natural products occupy chemical space that is not well represented by synthetic libraries, and thus there remains high value in natural product scaffolds for biomedical applications.[34,35] Activity of a natural product against a biomolecular target that has no other small molecule modulator is of enormous value, regardless of its relative structural novelty. In this regard, comparatively little is known about the biological activities of most natural products in either ecological or biomedical contexts because in many cases evaluation of natural product bioactivities has been limited to basic cytotoxicity, antibacterial and antifungal whole cell assays. A vast opportunity to discover and characterize the biological function of natural products in complex biological systems therefore remains largely unexplored. Such investigation need not be limited to new or novel compounds but rather should seek to understand their endogenous roles in nature and/or human health applications, regardless of structural novelty.

A further limitation of the current analysis is that it only incorporates published structures, and does not provide any estimates as to the global capacity for natural product production. Additionally, it only includes organisms that have been studied to date. The majority of taxonomic space has had no systematic examination of its capacity for natural products production.[36] Nevertheless, the result of this current analysis suggests that unstudied source organisms with unique metabolic and/or environmental constraints should represent excellent sources of structurally novel natural products. Finally, this analysis does not take into consideration the products of ‘cryptic’ biosynthetic gene clusters, many of which are predicted to produce novel natural products based on bioinformatic analyses.[37,38]

3.5 Future Perspective

The results of our analysis indicate that the future for natural products is very bright indeed. From a variety of lines of evidence, including genetic analysis of the sequenced genomes of microorganisms and the trends documented herein, a large reservoir of chemical space exists in NPs. This has yet to be fully explored via traditional approaches, although accessing novel genetic resources as well as new biological prioritization methods are assisting these endeavors.[39,40] Nevertheless, it is imperative that the field aggressively innovates if we are to avoid increasing redundancy of effort and marginalization of natural products research in the areas of chemical biology and biotechnology. Given the trends observed in these data, it is reasonable to suggest that in most cases traditional NP discovery platforms implemented on traditional source organisms will lead predominantly to the isolation of traditional, well known chemical entities.

The goal of discovering a new lead compound to treat human disease is exceptionally complex. Not only does the agent need to be efficacious, but also have appropriate bioavailability, ADME and PK properties, and lack of toxicity. As a result, it is not only attractive, but absolutely necessary, to employ a diversity of approaches to the discovery of such a lead structure, be it “top-down” or “bottom-up” natural products discovery, screening of synthetic libraries, medicinal chemistry, fragment-based, structure based design, or some combination of the above. We should not become overly wedded to one approach, as robustness in the field of drug discovery derives in large part from this diversity of approaches, and is therefore to be embraced.

Natural products have important roles to play beyond simply the reporting of novel chemical structures. Rather, contributions to ecology, biotechnology and biomedicine will continue to form the backbone for natural products research programs in the coming decades. Development of new strategies and methods to better integrate natural product libraries into the modern biotechnology arena should therefore be considered a critical focus for academia, funding agencies and industry alike. With the many innovations that are being developed in this highly multidisciplinary field, it will be exciting to see where the next 70 years of natural products research will lead!

3.6 Experimental Methods

3.6.1 Dataset Creation and Curation

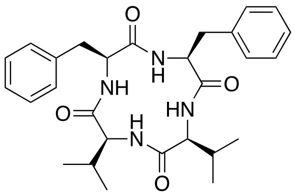
The dataset used for these analyses was created in two parts. For the years 1941 - 2011 we used the commercial database AntiMarin, removing all entries without discrete chemical structures, and ascribing year of discovery to the earliest citation available for each compound. In addition, all compounds annotated as synthetic, semi-synthetic or contaminant/ adduct were removed. Because AntiMarin is not available after mid-2012, we then created a new dataset from the primary literature that aimed to replicate the selection criteria for compound inclusion. To accomplish this, we searched the abstracts and titles of every article published in the period 2012 - 2015 from a panel of 48 journals. These journals were selected because they encompassed the vast majority of compounds published in the AntiMarin dataset. Each abstract and title was searched for keywords, including the genus names of all bacterial and fungal species for which natural products had been previously reported, as well as a set of keywords appropriate for marine-derived compounds (e.g. cyanobacteria, Porifera, Cnidaria etc). Metadata for each matching article was downloaded, provided that at least one mol file was associated with the citation. Using a custom software tool created in-house we manually curated the resulting 29,062 articles to identify all of the novel natural products and their associated common names and biological origins. These data were collated into a single data file containing 52,395 compounds, and all structures examined manually to eliminate any non-natural products included through indexing errors.

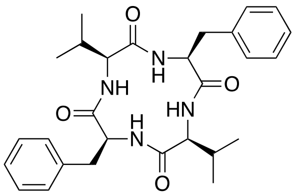
In order to perform the analyses on chemical diversity as a function of source organism (Figure 3-3), sources for each molecule were defined as follows. In most entries, the source data is reported as a string of text (e.g. "Porifera Dysidea herbacea"). Each text string was searched for presence of source organism keywords (bacterial genus names, fungal genus names, marine Phylum names) and all cases where a positive match was obtained were ascribed to that source. Compounds with multiple matching keywords, or compounds where no match was obtained were removed, providing a sub-set of the dataset comprising 50,093 compounds.

3.6.2 Tanimoto and Tversky Scoring

Structural similarity scoring is used extensively in medicinal chemistry and virtual screening to develop drug leads. In brief, the workflow for such analyses involves two steps. Firstly, each molecule must be described as a binary string of molecular features in a process known as 'fingerprinting'. Secondly, these binary strings must be compared against one another and scored using an appropriate similarity metric (Tanimoto, Tversky, cosine, etc).

Numerous variations exist for both the fingerprinting and scoring steps. For fingerprinting, compounds are most commonly described as a binary bitstrings that represent the presence of absence of a set of predefined structural features. In this study Morgan fingerprints were used as implemented in the RDKit software library[14]. These fingerprints describe molecules in terms of the neighborhood of each atom in the molecule (element, charge, presence in ring, number of adjacent heavy atoms, number of hydrogens attached) within a given bond radius. Bond radius is important, as it strongly impacts the values of the similarity scores obtained by comparing bitstrings. Too high a bond radius and most compounds score poorly in terms of similarity. Too low, and all compounds are scored as highly similar. For libraries containing a high diversity of compound structures (e.g. natural products or large untargeted virtual screening libraries) a bond radius of 2 is commonly used, as was the case for most analyses in this study. For compounds with high similarity in their core structures (e.g. cyclic peptides) a higher bond radius is preferable in order to improve the resolving power of the scoring system. For example, two cyclic peptides cyclo-Val-Val-Phe-Phe and cyclo-Val-Phe-Val-Phe receive a Tanimoto similarity score of 1 (identical) when fingerprinted with a bond radius of 2 because this radius is too short to relate the beta positions on the side chains to one another. By contrast, a bond radius of 4 affords a Tanimoto score of 0.77, driven by the Val-Val and Phe-Phe relationships in one structure but not the other.



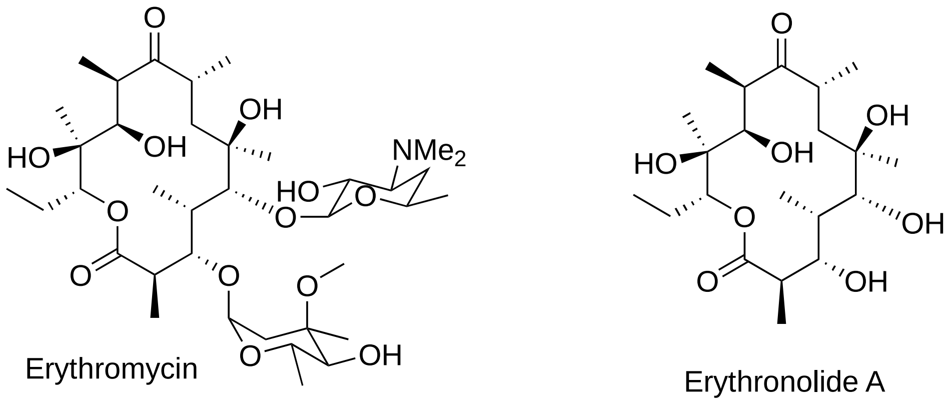


Path radius	Tanimoto scores
2	1.0
3	0.91
4	0.77

For similarity scoring, most methods relate the number of features common to both molecules to the total number of measured features. In the case of Tanimoto scoring for molecules A and B, the score (T_S) is defined as the number of features common to both molecules, divided by the total number of unique features:

$$T_S = \frac{A \cap B}{A \cup B}$$

The implication of using this scoring method is that feature absence impacts the Tanimoto score with the same weight as feature presence. For example, the Tanimoto score between the glycosylated polyketide erythromycin and its aglycone erythronolide A is just 0.46, despite one being a sub-structure of the other. Nevertheless, Tanimoto scoring has been proven to provide a good general description of chemical similarities across varied compound sets.[41]


Erythromycin

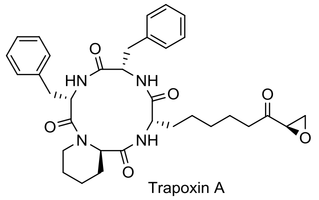
Erythronolide A

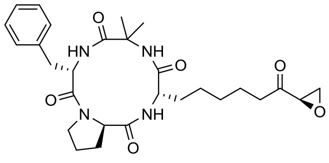
Figure 3-7

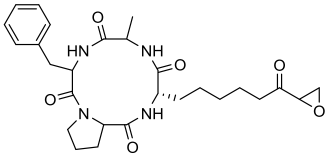
To address the issue of substructure relatedness, a number of alternative scoring methods exist that perform bi-directional scoring of each compound as a substructure of the other. Of these, Tversky scoring remains a popular and valuable method in the computational chemistry community.[33] In brief, this approach works by considering to features that are unique to A and B, and weighting these using a weighting factor in order to provide a measure of how well each compound is a sub-unit of the other. In this example, weighting factors were set at 0.1 and 0.9, giving Tversky scores of: erythronolide A → erythromycin A = 0.82; erythromycin A → erythronolide A = 0.52. Taking an average of these two values gives an overall score of 0.67 for these two compounds. Tversky scoring was using for Figure 3-5 in this study because it more accurately relates natural products that are derivatives or shunt products of the same biosynthetic pathway than is possible with Tanimoto scoring, which is important when defining compound classes.

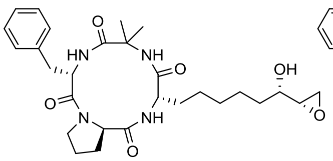
Figure 3-8 Cyclic Tetrapeptide Structures from Natural Product Dataset

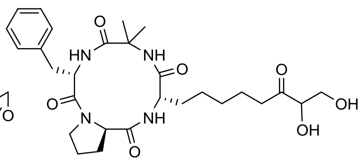
3.6.3 Cyclic Tetra peptide dataset

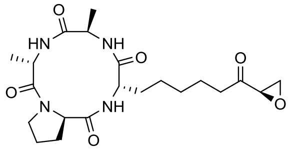

Trapoxin A

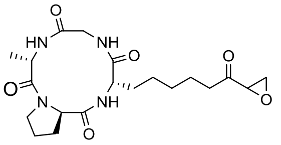

Chlamydocin; SL-3440

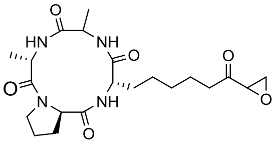

Ala-chlamydocin

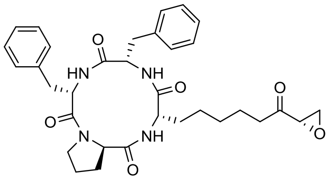

Dihydrochlamydocin

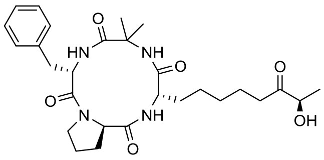

Deoxy-dihydroxychlamydocin

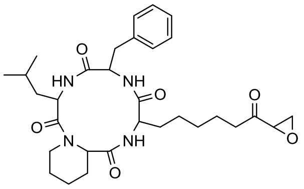

HC Toxin I

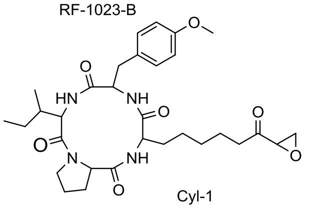

HC Toxin II

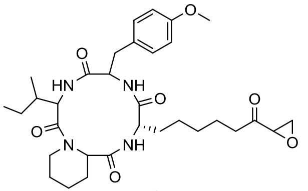

HC Toxin III

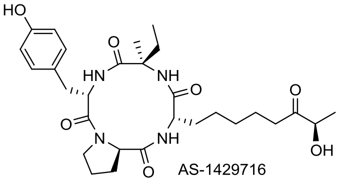

RF-1023-B

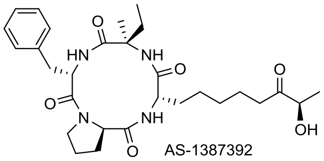

TAN-1746

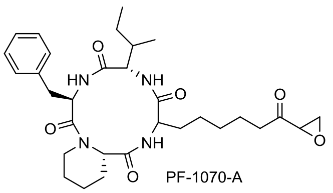

WF 3161; FR 900261

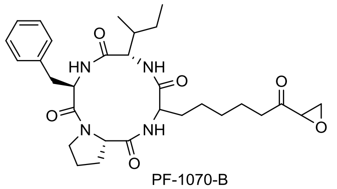

Cyl-1

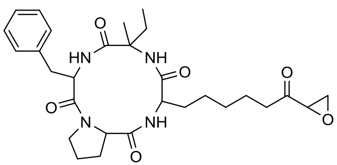

Cyl-2

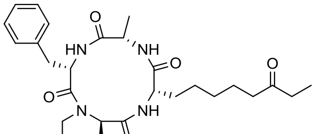

AS-1429716

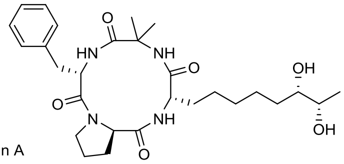

AS-1387392

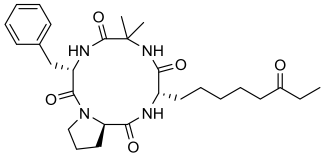

PF-1070-A

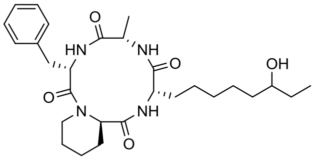

PF-1070-B

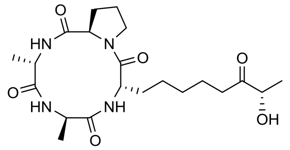

Phoenistatin

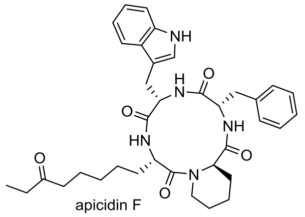

CNL629.512; Microsporin A

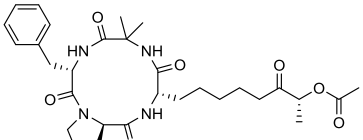

Diheteropeptin

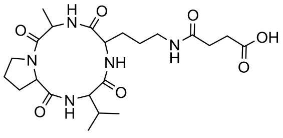

Chlamydocin analogue 2

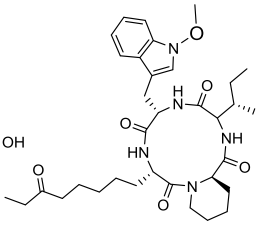

CNL629.514; Microsporin B

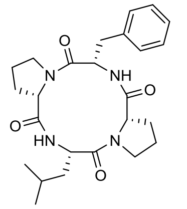

JM-47

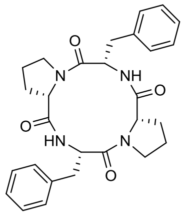

apicidin F

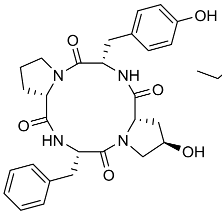

Chlamydocin analogue 3

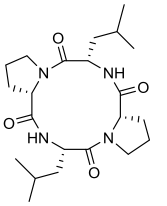

trichoderide

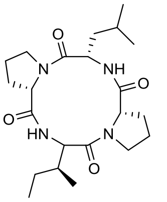

Apicidin Ia; OSI 2040

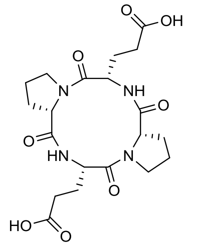

cyclo-[Phenylalanyl-prolyl-leucyl-prolyl]

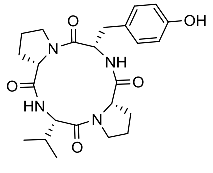

cyclo-[Phenylalanyl-prolyl]2

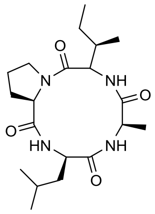

cyclo[L-Tyr-L-Pro-L-Phe-trans-4-hydroxy-L-Pro]

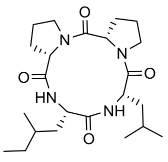

cyclo-[Leucyl-prolyl]2

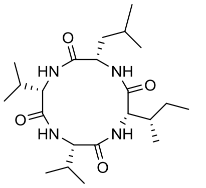

cyclo-[L-isoleucyl-L-prolyl-L-leucyl-L-prolyl]

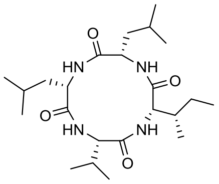

cyclo[L-alpha-Glutamyl-L-prolyl-L-alpha-Glutamyl-L-prolyl]

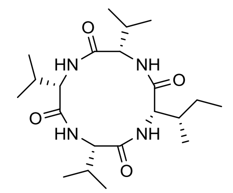

cyclo[L-Pro-L-Tyr-L-Pro-L-Val]


cyclo-[Isoleucyl-prolyl-leucyl-alanyl]

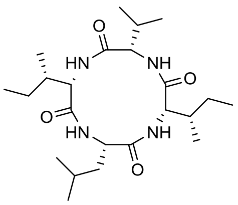

Fenestin A


cyclo[L-Ile-L-Leu-L-Val-L-Val]

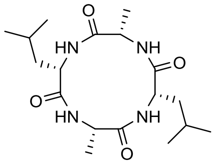

cyclo[L-Ile-L-Leu-L-Leu-L-Val]



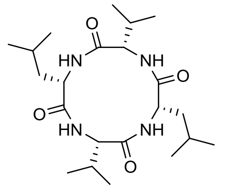
cyclo[Ile-Val-Val-Val]



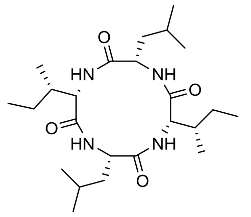
Halolitoralin C



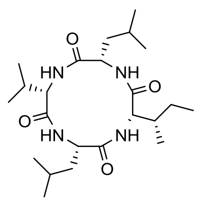
cyclo[L-Leu-L-Ala-L-Leu-L-Ala]



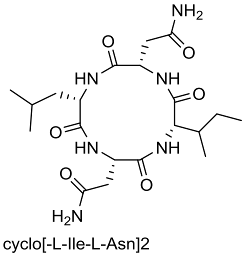
cyclo[L-Val-L-Leu-L-Val-L-Leu]



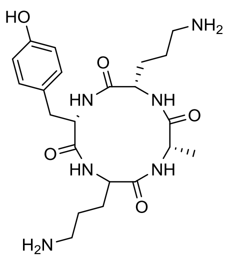
cyclo-[Leucyl-isoleucyl]2

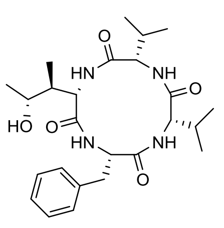


Halolitoralin B

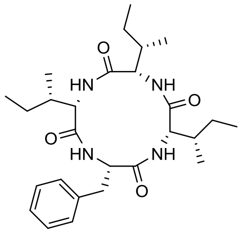


cyclo[-L-Ile-L-Asn]2

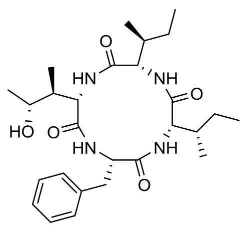




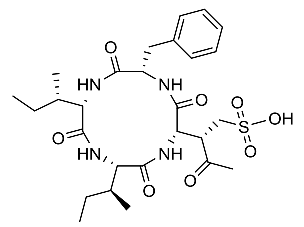
WSS 2217



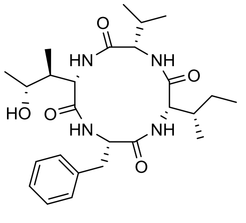
WSS 2218



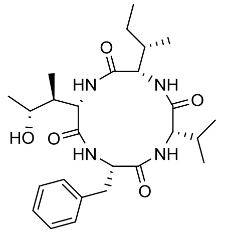
WSS 2219



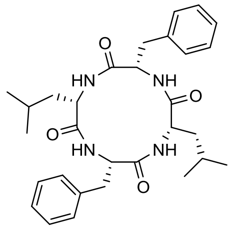
WSS 2220



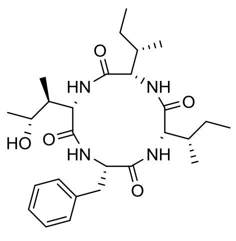
WSS 2221

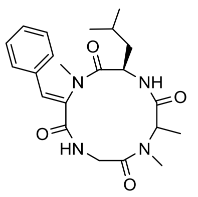


WSS 2222

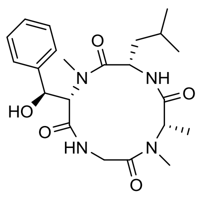


cyclo-[Phenylalanyl-leucyl]2

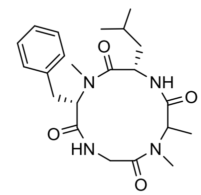




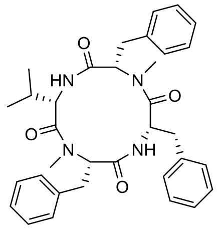
Tentoxin



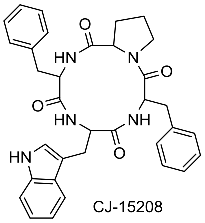
tentoxin B



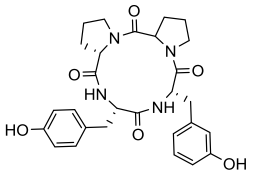
Dihydrotentoxin



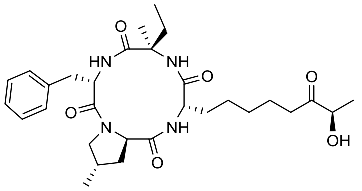
Hirsutide



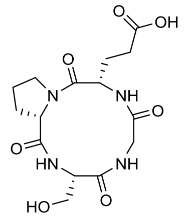
CJ-15208



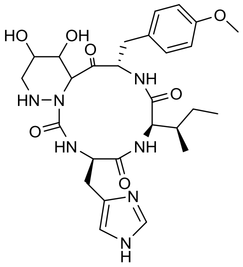
cyclo-[proply-proplyl-tyrosyl-tyrosine]



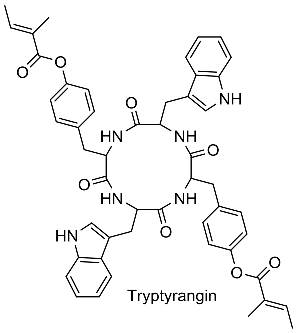
FR-235222



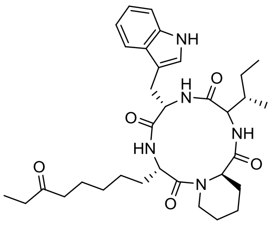
cyclo[L-alpha-Glutamyl-glycyl-L-seryl-L-prolyl]



Glomecidin



Tryptyrangin



Apicidin A

3.7 Bibliography

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2016**, *79*, 629–661.

(2) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129.

(3) Gwynn, M. N.; Portnoy, A.; Rittenhouse, S. F.; Payne, D. J. *Ann. N. Y. Acad. Sci.* **2010**, *1213*, 5–19.

(4) Kong, D.-X.; Guo, M.-Y.; Xiao, Z.-H.; Chen, L.-L.; Zhang, H.-Y. *Chem. Biodivers.* **2011**, *8*, 1968–1977.

(5) Walsh, C. T. *Nat. Chem. Biol.* **2015**, *11*, 620–624.

(6) Ju, K.-S.; Gao, J.; Doroghazi, J. R.; Wang, K.-K. A.; Thibodeaux, C. J.; Li, S.; Metzger, E.; Fudala, J.; Su, J.; Zhang, J. K.; Lee, J.; Cioni, J. P.; Evans, B. S.; Hirota, R.; Labeda, D. P.; van der Donk, W. A.; Metcalf, W. W. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12175–12180.

(7) Jensen, P. R.; Chavarria, K. L.; Fenical, W.; Moore, B. S.; Ziemert, N. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 203–209.

(8) Bachmann, B. O.; Van Lanen, S. G.; Baltz, R. H. *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 175–184.

(9) Wright, G. D. *Can. J. Microbiol.* **2014**, *60*, 147–154.

(10) Walsh, C. T.; Wencewicz, T. A. *J. Antibiot. (Tokyo).* **2014**, *67*, 7–22.

(11) Patridge, E.; Gareiss, P.; Kinch, M. S.; Hoyer, D. *Drug Discov. Today* **2016**, *21*, 204–207.

(12) Koehn, F. E.; Carter, G. T. *Nat. Rev. Drug Discov.* **2005**, *4*, 206–220.

(13) David, B.; Wolfender, J.-L.; Dias, D. A. *Phytochem. Rev.* **2015**, *14*, 299–315.

(14) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(15) Derewacz, D. K.; McNees, C. R.; Scalmani, G.; Covington, C. L.; Shanmugam, G.; Marnett, L. J.; Polavarapu, P. L.; Bachmann, B. O. *J. Nat. Prod.* **2014**, *77*, 1759–1763.

(16) Andrianasolo, E. H.; Haramaty, L.; McPhail, K. L.; White, E.; Vetriani, C.; Falkowski, P.; Lutz, R. *J. Nat. Prod.* **2011**, *74*, 842–846.

(17) Abbas, S.; Kelly, M.; Bowling, J.; Sims, J.; Waters, A.; Hamann, M. *Mar. Drugs* **2011**, *9*, 2423–2437.

(18) Thushara Diyabalanage, †; Charles D. Amsler, ‡; James B. McClintock, ‡ and; Bill J. Baker*, †. **2006**.

(19) Kusari, S.; Hertweck, C.; Spiteller, M. *Chem. Biol.* **2012**, *19*, 792–798.

(20) Schueffler, A.; Anke, T.; Kaiser, M.; Brun, R.; Hamer, M.; Parr-Dobrzanski, B.; Che, Y.; Ortiz-López, F. J.; Bills, G. F.; Liu, X.; An, Z.; Omura, S.; Giacobbe, R.; Abruzzo, G.; Hickey, E.; Liberator, P.; Xu, D.; Roemer, T.; Singh, S. B. *Nat. Prod. Rep.* **2014**, *31*, 1425–1448.

(21) Motley, J. L.; Stamps, B. W.; Mitchell, C. A.; Thompson, A. T.; Cross, J.; You, J.; Powell, D. R.; Stevenson, B. S.; Cichewicz, R. H. *J. Nat. Prod.* **2016**, acs.jnatprod.6b00772.

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

(23) Theodore, C. M.; Stamps, B. W.; King, J. B.; Price, L. S. L.; Powell, D. R.; Stevenson, B. S.; Cichewicz, R. H. *PLoS One* **2014**, *9*, e90124.

(24) Lin, Z.; Torres, J. P.; Ammon, M. A.; Marett, L.; Teichert, R. W.; Reilly, C. A.; Kwan, J. C.;

Hughen, R. W.; Flores, M.; Tianero, M. D.; Peraud, O.; Cox, J. E.; Light, A. R.; Villaraza, A. J. L.; Haygood, M. G.; Concepcion, G. P.; Olivera, B. M.; Schmidt, E. W. *Chem. Biol.* **2013**, *20*, 73–81.

(25) Calteau, A.; Fewer, D. P.; Latifi, A.; Coursin, T.; Laurent, T.; Jokela, J.; Kerfeld, C. A.; Sivonen, K.; Piel, J.; Gugger, M. *BMC Genomics* **2014**, *15*, 977.

(26) Kehr, J.-C.; Gatte Picchi, D.; Dittmann, E. *Beilstein J. Org. Chem.* **2011**, *7*, 1622–1635.

(27) Salvador, L. A.; Taori, K.; Biggs, J. S.; Jakoncic, J.; Ostrov, D. A.; Paul, V. J.; Luesch, H. *J. Med. Chem.* **2013**, *56*, 1276–1290.

(28) Kleigrewe, K.; Almaliti, J.; Tian, I. Y.; Kinnel, R. B.; Korobeynikov, A.; Monroe, E. A.; Duggan, B. M.; Di Marzo, V.; Sherman, D. H.; Dorrestein, P. C.; Gerwick, L.; Gerwick, W. H. *J. Nat. Prod.* **2015**, *78*, 1671–1682.

(29) Rosén, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. *J. Med. Chem.* **2009**, *52*, 1953–1962.

(30) Pascolutti, M.; Campitelli, M.; Nguyen, B.; Pham, N.; Gorse, A.-D.; Quinn, R. J. *PLoS One* **2015**, *10*, e0120942.

(31) Itazaki, H.; Nagashima, K.; Sugita, K.; Yoshida, H.; Kawamura, Y.; Yasuda, Y.; Matsumoto, K.; Ishii, K.; Uotani, N.; Nakai, H. *J. Antibiot. (Tokyo).* **1990**, *43*, 1524–1532.

(32) Darkin-Rattray, S. J.; Gurnett, A. M.; Myers, R. W.; Dulski, P. M.; Crumley, T. M.; Allocco, J. J.; Cannova, C.; Meinke, P. T.; Colletti, S. L.; Bednarek, M. A.; Singh, S. B.; Goetz, M. A.; Dombrowski, A. W.; Polishook, J. D.; Schmatz, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 13143–13147.

(33) Kunimoto, R.; Vogt, M.; Bajorath, J. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 523–531.

(34) López-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. *Drug Discov. Today* **2012**, *17*, 718–726.

(35) and, M. F.; Schmidt, J. M. **2002**.

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

(37) Yamanaka, K.; Reynolds, K. A.; Kersten, R. D.; Ryan, K. S.; Gonzalez, D. J.; Nizet, V.; Dorrestein, P. C.; Moore, B. S. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 1957–1962.

(38) Rutledge, P. J.; Challis, G. L. *Nat. Rev. Microbiol.* **2015**, *13*, 509–523.

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

(40) Park, S. R.; Tripathi, A.; Wu, J.; Schultz, P. J.; Yim, I.; McQuade, T. J.; Yu, F.; Arevang, C.-J.; Mensah, A. Y.; Tamayo-Castillo, G.; Xi, C.; Sherman, D. H. *Nat. Commun.* **2016**, *7*, 10710.

(41) Willett, P. *Drug Discov. Today* **2006**, *11*, 1046–1053.

4 Future Direction

4.1 Chapter 1

Though the outcome of the screening campaign was less than desirable, the goal of targeting pRb has not died with the false positives. The FP assay developed provides a robust and quick assay for subsequent screening or follow up for any other discovery techniques. One such alternate screening technique is ongoing with our collaborators at Roche Nimblegen. Nimblegen is a synthesis platform that allows for the synthesis of millions of spatially encoded, unique peptides on a chemically functionalized slide using DLP mirror arrays and photolabile protecting groups. These resulting arrays of peptides can then be assayed for protein binding by incubating the a fluorescently tagged protein of interest. A specialized fluorescence scanner can effectively read the intensity of each unique peptide.

The Lokey Lab collaboration with Roche has been focused on enabling their technology to include the synthesis of macrocyclic compounds. By using a sidechain attachment point through a Glu sidechain carboxyl, the cyclization can be conducted on-array similar to the on-resin cyclization approach we have employed in our lab. Using our understanding of the pharmacokinetic properties of macrocycles, our collaboration is premised on enabling the Roche Nimblegen group to synthesize cell permeable macrocycles and screen these libraries on the array. The screening of pRb scaffolds is currently underway and the best hits from the assay on the array will be synthesized in the Lokey lab and assayed in the FP pRb-E2F assay for efficacy.

4.2 Chapter 2

The work performed in Chapter 2 was one of the first attempts to quantify the effect of size on passive membrane permeability of compounds with MW > 500. We found a trend that was in stark disagreement with the current models and this work helps define what the likely permeable space is in the beyond Rule of 5 size regime. However, the mechanism by which this size dependence is manifested is still unknown and as such there is still work to be done to fully elucidate the theory and create more accurate, predictive models.

It is often speculated that the conformational dynamism or “chameleonicity” that is observed in scaffolds like cyclosporine A must serve to increase their solubility while maintaining lipophilicity. This hypothesis remains poorly supported. Devising a system that would shed direct light on the role of conformational flexibility would be of great interest to the field. To accomplish this goal, a direct read out of flexibility would need to be used in order to understand the relationship between flexibility, solubility and lipophilicity. This could be accomplished using NMR T2 relaxation experiments of scaffold that incorporated a variety of rotatable bonds and N-alkylation in their backbone. Additionally, the dynamics between amorphous solid (aggregate) and the fully solubilized monomer in solution are not well understood for these systems. Working with experts in dynamic light scattering may allow one to observe the thermodynamic distributions of the monomer / aggregate / large aggregate system. Depending on the kinetics and practical aspects of the experiment, filtration or dilution experiments may allow one to gain kinetic information about this system.

The influence of shape on the size dependence observed is another variable which has not been studied. Experimentally determining the shape of numerous scaffolds is difficult, particularly if the conformations are solvent dependent like CSA and other macrocyclic compounds. However, Ion Mobility Separation (IMS) may serve a surrogate for a direct measurement of cross sectional area in a low dielectric environment (vacuum). By taking a variety of scaffolds with known permeabilities, solubilities, and logK’s one could investigate whether scaffolds which flew faster in the IMS stage of an appropriately equipped mass spectrometer might be correlated with increased apparent diffusion in the permeability experiment.

4.3 Chapter 3

The cheminformatic, natural products analysis in chapter 3, though very interesting, was not the initial impetus to investigate the natural products structural database. I originally was interested in exploring the structural diversity of macrocycles in nature. The similarity comparisons that inspired the work conducted in chapter 3 was originally inspired my attempts at clustering and visualizing the macrocyclic natural products chemical space. One on the more successful techniques I employed for visualization was the production of a network from pairwise similarities. (Figure 4-1)

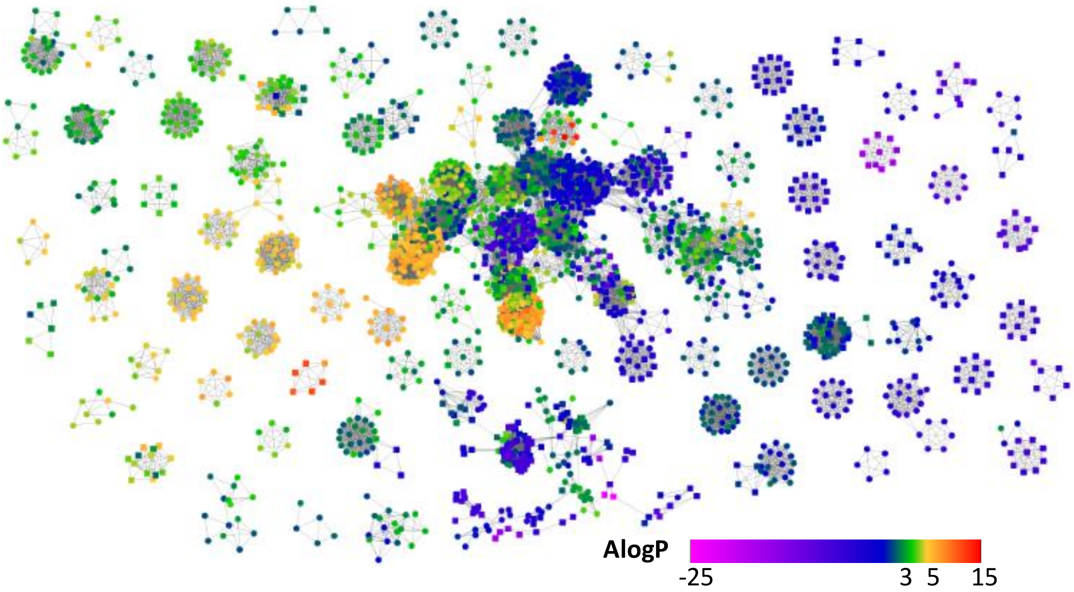


Figure 4-1 Similarity network of all cyclic peptide natural products in the Anti-Marin database. Each edge represents a Dice similarity score >0.8 and each node is a molecule. Nodes are colored based on their calculated octanol/water partition coefficients (AlogP).

To create this visualization, all pairwise Dice similarity scores were calculated for each molecule in the dataset which had been filtered for compounds containing at least 3 cyclic amide bonds. The Dice scores that were greater than an arbitrary threshold (0.8) were considered “edges” in the resulting network. This leads to the effective visual clustering of groups of similar compounds when a force directed layout algorithm is employed. Nodes, which represent molecules can be colored to convey additional information such as a property of each node. In Figure 4-1 the nodes are colored by the calculated lipophilicity (AlogP) of each compound. We can easily observe from this that compounds that are chemically similar also tend to share property similarity such as calculated lipophilicity. This may seem obvious, but even subtle changes in structure can bring large changes in lipohpillicity such as the addition of a charged group. As we discuss at length in Chapter 2, lipophilicity is a central property governing cell permeability, we can boldly infer that there are groups of structures that might be likely to be cell permeant and those that are not compounds that have intracellular targets and perhaps those with extracellular targets respectively.

There remains an interesting series of cheminformatic analyses to be performed on this rich dataset of nature's macrocycles. I've been developing algorithms to extract backbones from the structures to investigate properties of the scaffolds vs the sidechain appendages and how these differences might be conserved between molecules whose calculable properties suggest potential passive permeability. Interrogating the chemical composition of the monomeric species that make up these molecules may also serve an enriched set of matter to draw from for synthetic systems that are created for biological activity. Even a cursory look at the chemical space shows that there are wide variety of interesting chemistries that bear little resemblance to CSA, often regarded as the posterchild of beyond Rule of 5 drug-likeness. I think that exploring the space synthetically around the islands sampled by nature will give rise to a myriad of useful molecules to be employed in the pursuit of better chemical entities for drug discovery. I plan on completing these proposed analyses and publishing this work shortly after the conclusion of my dissertation.

5 Concluding Remarks

My time at UC Santa Cruz will be a cherished chapter in my life. I’ve been fortunate enough to make an exceptional set of friendships with some incredible people. The Chemistry program at Santa Cruz offers a since of collegiality and collaborative atmosphere I hope I can find and create throughout the rest of my life. Though my research path was a circuitous one I wouldn’t change a thing, the opportunity to work with myriad of wonderful colleagues and explore a variety of research interests has been a privilege. See you on the flipside.