UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EMPIRICAL DETERMINATION OF THE INDIVIDUAL PERMEABILITIES OF THOUSANDS OF GEOMETRICALLY DIVERSE CYCLIC HEXA- AND HEPTAPEPTIDES VIA MULTIPLEX PAMPA AND MSMS SEQUENCING

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Chad Edward Townsend

December 2020

The Dissertation of Chad Edward Townsend
is approved:

_____
Professor R. Scott Lokey, chair

_____
Professor Glenn Millhauser

_____
Professor Mark Akeson

_____
Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry.

Chapter 2: The passive permeability landscape around geometrically diverse hexa- and heptapeptide macrocycles.

LIST OF TABLES

Chapter 1: CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry.

Chapter 2: The passive permeability landscape around geometrically diverse hexa- and heptapeptide macrocycles.

ABSTRACT

Empirical determination of the individual permeabilities of thousands of geometrically diverse cyclic hexa- and heptapeptides via multiplex PAMPA and MSMS sequencing

By Chad Townsend

Much of modern pharmaceutical development has occurred within or nearby the chemical space delineated by Lipinski's rules of five (< 500 molecular weight, < 5 hydrogen bond donors, < 5 octanol/water partition coefficient, < 10 hydrogen bond acceptors) due to the practical advantages in pharmacokinetic properties evinced by such small molecules, especially cell permeation. Their small size has made them ideal for occupying small binding pockets instead of a protein's intended substrate and achieving sufficient oral bioavailability for oral delivery is generally facile. More recently, therapeutic enzymes and antibodies have joined Insulin in a class of injectable macromolecule therapeutics to great success – multiple new therapeutic antibodies are approved each year. Their large size gives them superior selectivity, specificity, and potency compared to small molecules, but also precludes them from entering cells to modify intracellular interactions (and from oral delivery). These two therapeutic modalities leave a great many intracellular interactions which occur over a larger surface area (i.e., not involving native small molecules) "undruggable", and the continual discovery of actionable disease-relevant interactions in this category has prompted a search for new therapeutic modalities.

Cyclic peptides of up to 1200 molecular weight have demonstrated the ability to inhibit a wide variety of such "undruggable" intracellular interactions. Their ease of synthesis combined with advances in DNA/mRNA encoded library construction and screening technologies have ensured that obtaining potent cyclic peptide leads against virtually any protein target is possible. However, these powerful encoding technologies cannot screen for permeability and engineering cell permeability into cyclic peptides remains a major barrier to their therapeutic utility. The conformational nature of cyclic peptide passive cell permeability has thus far defied computational prediction over a broad set of compounds and empirical

evaluations of cyclic peptide passive permeability have been of limited size (tens to low hundreds). The overall goals of this dissertation are to empirically evaluate the prevalence of passively cell permeable backbone geometries across thousands of geometrically diverse cyclic hexa- and heptapeptides, to derive general insights into the design of passively cell permeable cyclic peptides, and to publish a database of known permeable backbone geometries with which screening libraries can be biased towards passive cell permeability.

Chapter one covers the details of CycLS, a non-de novo cyclic peptide sequencing program using a database-matching approach to allow sequencing of entire libraries of cyclic peptides in a timely manner. I validated CycLS against a unique-mass library of 400 cyclic hexapeptomers, achieving 95% sequencing accuracy despite a ratio greater than 500:1 of decoy sequences to true sequences, and against a mass-redundant 1800-member library of cyclic hexapeptides and hexapeptomers (also found in chapter two) by resynthesis of twenty-two individual compounds over a broad range of sequencing scores. I then devised a normalized sequencing confidence metric that was able to divide the seventeen successfully sequenced resynthesized compounds from the five unsuccessfully sequenced resynthesized compounds. Direct sequencing of cyclic peptides rather than by linearization or encoding is critical for passive cell permeability assays, which are sensitive to any encoding tag or the structural elements pre-requisite to many linearization strategies. CycLS is freely available online and improves on previous work in this area by inclusion of extensive spectral pre-processing to remove noise and boost signal, which is critical to ensure sequencing quality.

CycLS is the first step of a computational workflow to associate individual compound identities with corresponding mass spectrometry quantified assay results by matching individual peak retention times given an identical LCMS method. The remainder of that workflow can be found in appendix A and is composed of CycLS and two additional steps. Processing of the assay data is done via AutoPAMPA, which performs automated peak-finding, integration, and calculation of permeation rates for the PAMPA artificial membrane permeability assay. AutoPAMPA removes data analysis as the limiting factor to throughput

for mass spectrometry quantified assays by batch-processing of hundreds to thousands of peaks per job. Finally, RTMerge performs peak alignment and matching by retention time between CycLS and AutoPAMPA outputs and generates statistics describing the physicochemical properties and structure of each compound. In addition to the work included here, this computational workflow has already been successfully used to explore the permeabilities of hundreds of lariat peptides and enables further such projects in the future.

Chapter two investigates the passive permeability of 1800 cyclic hexamers and 3600 structurally related cyclic heptamers with highly variant backbone geometries by PAMPA. Of especial interest were the effects of N-methyl residues, peptoid residues, and beta residues on passive permeability. I identified 823 hexamers and 1330 heptamers with permeation rates greater than $1 * 10^{-6}$ cm/s, a threshold at which compounds are considered passively cell permeable. I confirmed the utility of these library-derived permeabilities by correlation with the pure permeabilities of 9 resynthesized hexamers and 10 resynthesized heptamers. A matched-pair analysis revealed that peptoid and beta residues have a negative structural contribution to passive permeability that I hypothesize originates from their increased flexibility. As expected, a matched-pair analysis of stereochemistry showed little effect averaged over such diverse backbone geometries.

Library generation technologies with complete synthetic control stand to benefit greatly from the permeable hexamer and heptamer backbone geometries discovered in chapter two by selecting only the most permeable of them, but combinatorically generated libraries doing the same would be limited to small library sizes or few backbone geometries. To enable combinatorically generated libraries to better utilize known permeable backbone geometries I defined and investigated passive permeability “motifs” of length three, holding three residues of the library design static while allowing the rest to vary over a number of known permeable backbone geometries. The best motifs had median permeabilities four-fold greater than the median permeability of all other compounds with the same number of hydrogen bond donors. Bundled into motifs, these sets of permeable backbone geometries

allow combinatorically generated libraries increased size and some degree of geometric diversity.

This dissertation has thoroughly explored the impact of stereochemistry, N-methylation, and peptoid residues on the passive permeability of cyclic hexa- and heptamers and gained some insights into the effect of beta residues. In addition to these insights into the average effects on permeability, this dissertation emphasizes that backbone geometries with high intrinsic permeability may be designed into DNA/mRNA encoded screening libraries to improve the likelihood of hits with favorable pharmacokinetic properties. Lastly, the computational workflow necessary to gain these insights can be used to obtain a similar register of “privileged” backbone geometries for larger ring sizes.

ACKNOWLEDGEMENTS

The text of chapter 1 of this dissertation includes reprint of the following previously published material:

Townsend, C.; Furukawa, A.; Schwochert, J.; Pye, C. R.; Edmondson, Q.; Lokey, R. S., CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry. *Bioorg Med Chem* **2018,** *26* (6), 1232-1238.

Synthesis and tandem MS data acquisition for the first validation test library was performed by Akihiro Furukawa. Synthesis of the second validation library was performed by Chad Townsend and Quinn Edmondson. Chad Townsend wrote CycLS with guidance from Joshua Schwochert. Chad Townsend and Cameron Pye optimized the chromatography and mass spectrometry data acquisition parameters. Chad Townsend and R. Scott Lokey wrote the article. The co-author R. Scott Lokey directed and supervised the research which forms the basis for chapter 1 of the dissertation.

Chapter One

CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry.

This chapter contains text and figures from the following manuscript: Townsend, C.; Furukawa, A.; Schwochert, J.; Pye, C. R.; Edmondson, Q.; Lokey, R. S., CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry. *Bioorg Med Chem* **2018,** *26* (6), 1232-1238.

Abstract

Cyclic peptides are of great interest as therapeutic compounds due to their potential for specificity and intracellular activity, but specific compounds can be difficult to identify from large libraries without resorting to molecular encoding techniques. Large libraries of cyclic peptides are often DNA-encoded or linearized before sequencing, but both of those deconvolution strategies constrain the chemistry, assays, and quantification methods which can be used. We developed an automated sequencing program, CycLS, to identify cyclic peptides contained within large synthetic libraries. CycLS facilitates quick and easy identification of all library-members via tandem mass spectrometry data without requiring any specific chemical moieties or modifications within the library. Validation of CycLS against a library of 400 cyclic hexa-peptide peptoid hybrids (peptomers) of unique mass yielded a result of 95% accuracy when compared against a simulated library size of 234,256 compounds. CycLS was also evaluated by resynthesizing pure compounds from a separate 1800-member library of cyclic hexapeptides and hexapeptomers with high mass redundancy. Of 22 peptides resynthesized, 17 recapitulated the retention times assigned to them from the whole-library bulk assay results. Implementing a database-matching approach, CycLS is fast and provides a robust method for sequencing cyclic peptides that is particularly applicable to the deconvolution of synthetic libraries.

1.1 Introduction

Cyclic peptides have shown remarkable versatility as ligands against challenging therapeutic targets such as protein-protein interactions[1]. Cyclization can improve both potency[2, 3] and proteolytic stability[4] in peptides, and pharmacokinetic properties such as cell permeability. From a design perspective, the synthesis of cyclic peptides is highly modular, with ready access to functionality and structural diversity at the sequence and building block levels.

One-bead one-compound (OBOC) libraries provide a powerful platform for generating and screening highly diverse collections of molecules[5-8]. These libraries range can reach up to millions of members, with various hit deconvolution schemes available according to their size, composition, and usage. Examples include the labeling of multifunctional beads with a linear tag for mass spectrometry based sequencing[9-11] and chemical linearization of cyclic peptides at a labeled residue post-screening[12-15]. The above examples simplify the deconvolution of cyclic peptide libraries by allowing the use of well-developed linear peptide sequencing techniques and software packages developed for sequencing proteins. Sequencing linear tags or linearized peptides raises a few problems, however. Linearization techniques effectively add a synthetic step which can reduce the yield and purity of the library, and possibly affect signal strength. Additionally, many linearization techniques require a specific chemical moiety designed into the library, inherently limiting library composition. While linearization is sufficient for decoding individual hits from on-bead binding assays, solution-phase screening of complex library mixtures would benefit from a strategy that allows direct sequencing of individual members. Therefore, we set out to develop a direct tandem mass spectrometry-based sequencing method for cyclic peptide libraries which we could apply to the deconvolution of complex mixtures in solution.

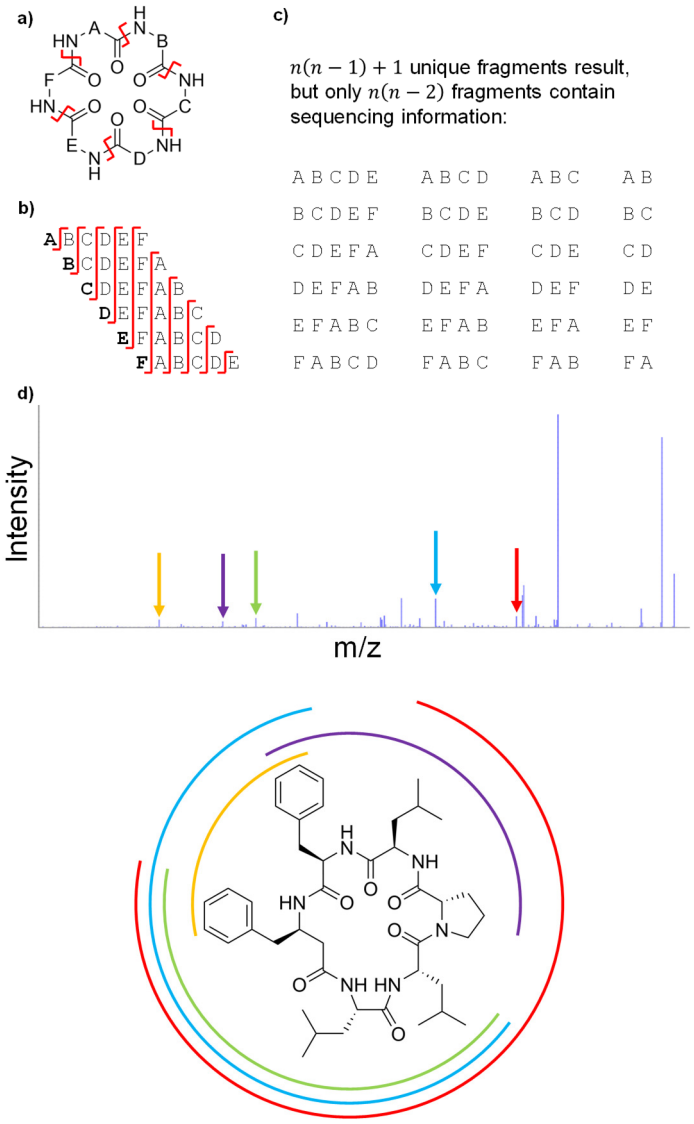


Figure 1.1 Cyclic peptide fragmentation. a) An example cyclic hexapeptide with ring-opening cleavage sites at amide bonds in red. b) All linear fragments of the example cyclic peptide with cleavage sites at amide bonds in red. c) The full-length linear fragments share a mass, so they can be counted as one unique fragment, but they contain no sequencing information. Neither do the n immonium ions generated, as they contain only a single amino acid. d) A typical MS^2 spectrum of a cyclic peptide with marked ions mapped to their position on the cyclic peptide. Even without a clear ion series, combinations of fragments can still be used to derive sequence information.

The primary difference between fragmentation of linear and cyclic peptides is the character of the ion series produced. Linear peptides have an ion series for each terminus, and both can result from an amide bond cleavage at any site within the chain. When a cleavage event occurs, one side or the other retains the charge, producing a b-ion (N-terminally), or a y-ion (C-terminally). The characteristic mass of the termini allow the peptide to be sequenced by interpreting the mass differences in sequential members of an ion series[16]. Because cyclic peptides have no termini, two amide bond cleavage events must occur for a cyclic peptide to fragment (Figure 1.1), resulting in a b-like ion, so-called because its mass is identical to a b-ion of the same composition. Thus, for a peptide of length n, the first amide bond cleavage will generate n full-length sequences which are not differentiable from each other by mass. The second amide bond cleavage will generate n fragments of each length 1 through n from each of the full-length sequences. This leads to a total of n^2 unique fragments, of which $n(n-1)$ are identifiable by a unique mass, with one additional mass characteristic of fragments derived from only a single amide bond cleavage event. Of the fragments identifiable by a unique mass, the fragments of length 1 contain no sequencing information, resulting in $n(n-2)$ fragments useful for sequencing. Due to the increased number of fragments and lack of termini with characteristic masses, these b-like ions do not form an identifiable series from which sequence information can be easily extracted. Furthermore, additional high-energy collisions may cause a cyclization and subsequent re-opening of a b-like ion, scrambling the residue ordering from its initial sequence[17, 18].

Several *de novo* cyclic peptide sequencing programs have been developed, primarily to facilitate the structure elucidation of natural products via tandem mass spectrometry[17, 19-27]. Due to the extensive chemical space of a *de novo* search, however, they take minutes per compound, which limits their application to large OBOC libraries. An automated sequencing program designed specifically to decode cyclic peptide libraries was described by Redman, et al., [28]. Their program treats the sequencing problem as a database search by using the library design as the database for a sequencing run. Each candidate compound for a specific

MS^2 spectrum is virtually fragmented and the virtual fragments are compared against the peaks present in the spectrum, after which the highest scoring peptide from the database is put forward as the correct sequence. Here we describe a conceptually similar workflow called CycLS (for “cyclic peptide whole-library sequencing”) that differs from the previous study in the introduction of extensive preprocessing of the MS^2 spectra before sequencing and in the scoring of candidate molecules. CycLS was evaluated using two cyclic hexapeptide and hexapeptomer (peptide-peptoid hybrid) libraries. The first library consisted of 400 compounds of unique mass, and therefore no resynthesis was necessary. The second library consisted of 1800 compounds with high mass redundancy, and twenty-two compounds were resynthesized to validate their sequencing results. Seventeen of the twenty-two resynthesized compounds were correctly identified, a result that is comparable to the 77% accuracy reported by Redman, et al.

1.2 Results

1.2.1 Spectral Pre-Processing to Ensure High-Quality MS² Spectra

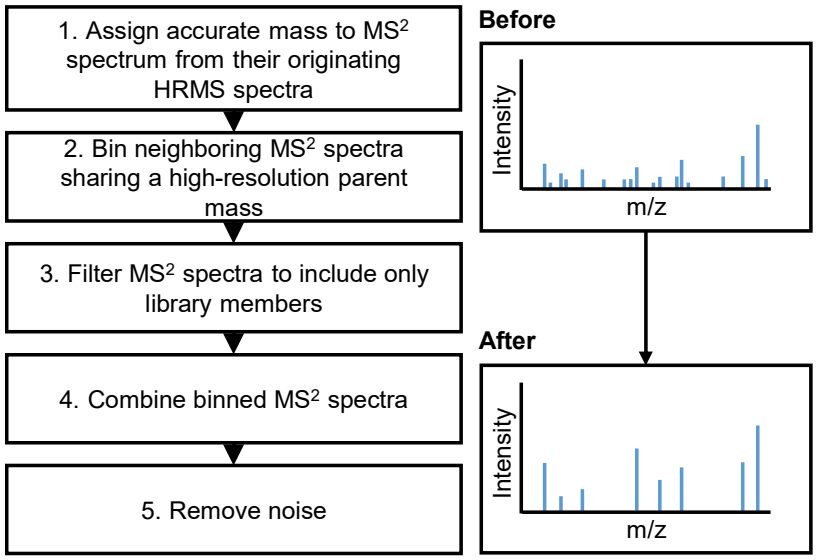


Figure 1.2 Spectral pre-processing. A summary of spectral pre-processing, highlighting the major steps. The example MS² spectrum represents an ideal result, wherein all noise peaks have been removed and all signal peaks have been significantly amplified.

To improve data quality and reduce the database of candidates to be searched, the MS² spectra were pre-processed to remove noise and amplify the signal of meaningful peaks (Figure 1.2). To maximize speed and sensitivity in the MS² mode, we obtained MS² spectra at a relatively low resolution (±0.2 m/z), and then matched each MS² spectrum to the closest parent ion from the originating MS¹ spectrum, which was obtained at higher resolution (±0.015 m/z). Next, consecutive MS² spectra sharing the same parent ion mass were grouped and assumed to be of the same chromatographic peak unless changes in signal intensity indicated closely eluting peaks of the same mass, in which case the group was split. Using the library design, which was input at the beginning of the sequencing run, each MS² spectrum was assigned a list of candidates with matching high-resolution masses. Spectra without any candidate library-members were discarded. Spectra within the same time neighborhood and sharing the same parent ion mass were then combined to improve signal to noise. The most intense MS² spectrum of each group was used as the base spectrum, to

which the other spectra were added, on the assumption that the most intense spectrum would contain the most complete set of peptide fragment peaks. Peaks were matched from the other spectra to the base spectrum on a one-to-one basis, taking the nearest m/z peak within a maximum distance to be the best match[29]. Matched peaks had their intensity added to the base spectrum. Finally, the combined MS^2 spectra were filtered to remove noise, using a threshold determined by creating an approximation of the probability density function of noise peak intensity.

1.2.2 Peak-Fragment Matching and Scoring

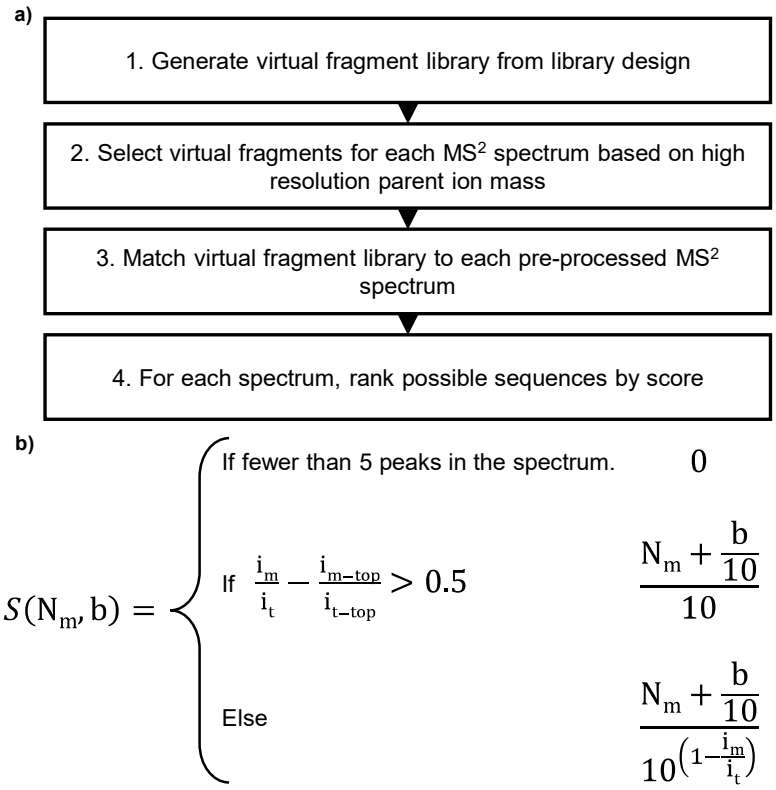


Figure 1.3 Overview of sequencing. a) Major steps of the sequencing process. b) The scoring equation for a single candidate against a single spectrum, where N_m is the number of main fragments matched, b is the number of extra times a main fragment was matched, i_m is the total intensity of the peaks matched, i_t is the total intensity of the spectrum, and i_{m-top} and i_{t-top} are missing the most intense peak in the spectrum. The first and second conditional cases penalize MS^2 spectra originating from non-peptidic contaminants, which often fragment poorly under the conditions used for peptide tandem mass spectroscopy.

The input library design was used to generate the sequences of all library-members, and each library-member was virtually fragmented at every amide bond to generate a virtual fragment library (Figure 1.3a). From the resulting “main fragments”, all possible neutral losses were generated from known fragmentation schemes[16]. This information was arranged such that each mass corresponding to any virtual fragment was traceable to the main fragment(s) from which it originated, and thus to each library member containing those main fragments. Each combined MS^2 spectrum was then compared to a subset of the virtual fragment library containing only the fragments corresponding to its candidate library-members as determined

by high-resolution mass matches. Finally, peaks and fragment masses were compared at the precision of the MS^2 data, resulting in a count of the number of matches to each main fragment of each candidate compound. Along with counting the matches to each main fragment of each candidate compound, the total intensity of the peaks matched to a candidate was also tracked. This process was repeated until each MS^2 spectrum was processed.

Scores were generated for each candidate to each spectrum using the total intensity and fragment match counts (Figure 1.3b). The first match to each main fragment was worth one point and all additional matches to that fragment were worth one tenth of a point. This weighting system was chosen because a single match to any neutral loss of a main fragment confirms the presence of that main fragment, and any additional matches only provide increased confidence in its presence. The matching score was totaled over all main fragments before being divided by a number between one and ten. That divisor was determined by comparing the spectral intensity to the total intensity of all peaks with matches to the candidate being scored, with a higher matched intensity resulting in a smaller divisor. The divisor's maximal value of ten was chosen such that the scores of candidates with high match-count and low intensity-matching were close in value to candidates with low match-count and high intensity-matching. Peak intensity was not used to weight individual matches because it was assumed that nearly all noise had been removed from the spectrum; in that case, the remaining peaks could be considered of equal reliability.

Finally, non-peptidic spectra originating from chromatographic contaminants (such as polymers), were penalized. Polymers occupy a large range of masses and are likely to overlap with a library mass in a highly multiplex mixture but tend to fragment poorly under collision induced disassociation conditions optimized for peptide fragmentation. At low intensity, the preprocessing of non-peptidic spectra results in very few peaks after noise filtration; therefore, spectra with fewer than five peaks were heavily penalized. More commonly, non-peptidic spectra have one extremely intense peak at their parent mass,

causing the noise level of the spectrum to rise above the filtration threshold. To prevent matching to that noise, spectra in which the most intense peak represents 0.5 or more of the fractional intensity matched by a candidate, the score divisor is set to ten. In addition to non-peptidic spectra, these scoring modifications also penalize peptides which fragment poorly-enough to have the same effect on the noise threshold's effectiveness.

1.2.3 Validation 1: Unique Mass Library

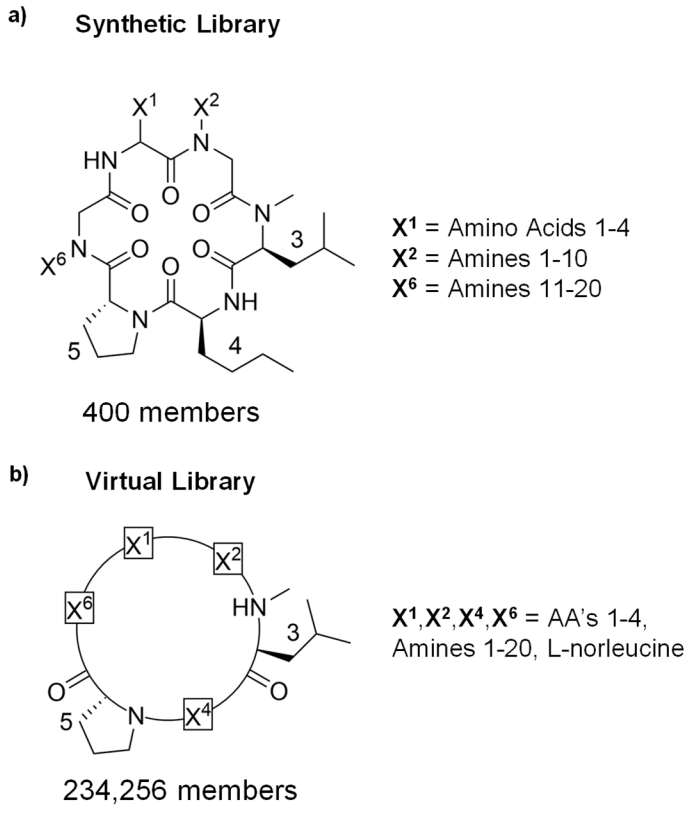


Figure 1.4 Unique mass library schematics. a) The design of the first library of cyclic hexapeptomers used to validate CycLS. Variation was focused at the side-chain level, and the library was designed such that after the final split at X^6, the ten sub-libraries contained no compounds which shared a mass. Residue identities for this design can be found in the initial study using this library. b) A large, virtual library was generated by combining the residue lists from several positions of the synthetic library. This second design was used to test CycLS, since the synthetic design contains no competing candidates for each spectrum.

To confirm the discriminatory ability of CycLS, MS^2 data were acquired for a library of 400 cyclic hexapeptomers that had been synthesized previously in a study focusing on

membrane permeability (Figure 1.4a)[30]. The library was comprised of 10 sub-libraries (defined by the residue at position X^6), each of which contained 40 compounds with unique masses. We had previously determined that all 40 members of each library were accounted for by the presence of a peak corresponding to their accurate masses in the MS^1 spectrum. However, only 345 out of 400 compounds were represented in the MS^2 data used for sequencing. The remaining compounds had signals that fell below the threshold for MS^2 spectral acquisition and were therefore not sequenced. A virtual library of 234,256 cyclic peptomer sequences was generated in which positions X^1, X^2, X^4, and X^6 were permuted among all building blocks, using common virtual residues to represent the groups of isobaric residues used in the initial library. We ran CycLS on the MS^2 spectra from the 10 sub-libraries, using all 234,256 permutations as the source of virtual fragments, and found that 328 of the 345 compounds represented in the MS^2 data were sequenced correctly.

The scoring function was examined via receiver operating characteristic (ROC) curve analysis[31]. Ninety-nine spectra matched virtual library members but were not associated with a member of the actual library, and thus were designated as false positives in the ROC analysis. Some of these spectra could have arisen from M+Na peaks from real library members, or from matching impurity peaks. It was observed that the raw score assigned to the top sequence for a spectrum had significant explanatory power for its categorization as a true or false positive, with an area under the curve of 0.96. The true and false positive rates (TPR and FPR) were found to be insensitive to the score threshold except between 1.0 (TPR 100%, FPR 36%) and 2.5 (TPR 86%, FPR 0%). A score threshold of 2 (TPR 92%, FPR 4%) was found to be a useful middle ground. Because this metric varies greatly between library designs, however, a normalized metric was sought. ROC analysis of the normalized difference between the top two scores for each spectrum resulted in a lower area under the curve, at 0.88. This normalized metric did not yield a convenient threshold for reduction of false positives, but it can be used effectively to compare sequencing confidence among

different library designs. This result prompted us to test the generality of CycLS on a larger, more diverse cyclic peptide library.

1.2.4 Validation 2: Mass-Redundant Library

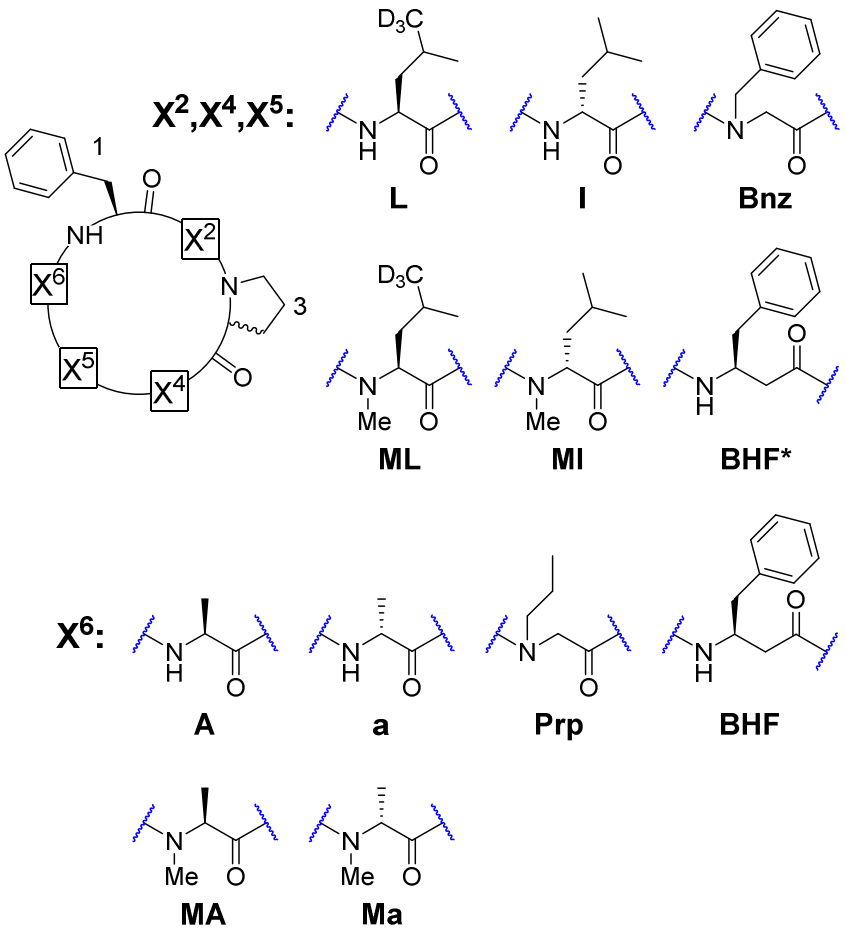


Figure 1.5 Mass-redundant library schematic. The design of the second library of cyclic hexapeptomers used to validate CycLS. Variation was focused at the backbone level, and the library was split into twelve sub-libraries of 150 compounds at position X^6. Due to the high mass redundancy, the synthetic library design was used to test CycLS. The residues follow the single letter amino acid code where possible, with uppercase letters representing L-amino acids and lowercase letters representing D-amino acids. Bnz abbreviates benzyl peptoid and BHF abbreviates L-β-homoPhe. Isotopic labeling was used to identify stereochemistry of leucine and n-methyl leucine at positions X^2, X^4, and X^5. * This residue is only present at position X^4.

To test the generality of CycLS toward other library designs, we synthesized a library of 1800 cyclic hexapeptides and hexapeptomers in which backbone elements, including

stereochemistry, N-methylation, and the presence of peptoids and beta-residues, were permuted. Stereochemistry was encoded using deuterium-labeled side chains to avoid isobaric residue masses except at the position six (Figure 1.5).

We used split-pool solid phase methods to synthesize the library, dividing it into twelve sub-libraries of 150 compounds each. Some truncations in the peptide sequences were observed, presumably due to incomplete couplings, but these compounds occupied a lower non-redundant mass range and eluted earlier, and therefore did not interfere with analysis of the library compounds.

Approximately 1200 of the 1800 library-members were sequenced, with the rest lost due to poor signal or chromatographic overlap. More chromatographic peaks were observed than expected, but this was determined to be a consequence of epimerization in a few cases and mass overlap of sodium ions in other cases. MS^2 data was collected separately for each sub-library to reduce chromatographic overlap of peaks with the same mass. Separating the analysis into sub-libraries allowed the database of potential compounds to be specified based upon the sub-library, reducing the sequencing search space of each individual CycLS run.

Due to the mass-redundancy of this library design, the identity of individual compounds was unknown. Therefore, twenty-two compounds that spanned a range of sequencing scores and lipophilicities were resynthesized to verify the sequencing results. Aside from those criteria, compounds were chosen based upon potential interest as novel, membrane permeable scaffolds. MS^2 data was acquired on the resynthesized compounds using the same method and MS^2 acquisition scheme used for the library (Table 1.1).

Seventeen of the twenty-two resynthesized compounds were judged to have been sequenced correctly, as determined by retention-time proximity, along with visual inspection and comparison of MS^2 data. For the five compounds that were not sequenced correctly, the normalized difference in score between the top two candidates for each sequence was less than 0.042 (with an average difference of 0.022), indicating that the fragment ion(s) unique to the correct sequence were not present and/or the incorrect sequence matched to noise

peaks that were above the noise removal threshold. For the seventeen compounds that were sequenced correctly, the normalized difference in scores between the top two candidates for each compound was greater than 0.08 (with an average difference of 0.30). The ten-fold difference in their averages supports the discriminatory power of the normalized score difference across library designs.

#	Sequence	Correct?	Top Score	Next Best Score	Normalized Difference	Number of Candidates
1.1	A,l,BHF,p,L,F	✓	3.9	2.7	0.32	14
1.2	BHF,ML,Bnz,P,l,F		3.6	3.6	0.006	14
1.3	Ma,L,Bnz,P,ML,F	✓	10.2	9.2	0.098	7
1.4	Ma,l,ML,p,ML,F	✓	15.9	9.3	0.42	9
1.5	a,Ml,BHF,p,Ml,F	✓	2.0			1
1.6	MA,l,l,P,l,F	✓	6.5			1
1.7	A,L,l,p,l,F		6.4	6.2	0.042	3
1.8	a,L,L,p,l,F	✓	5.7	4.8	0.16	3
1.9	A,l,L,P,L,F		6.8	6.5	0.042	3
1.10	A,l,L,p,L,F		6.5	6.3	0.039	3
1.11	BHF,ML,L,p,Ml,F	✓	4.7	3.2	0.34	9
1.12	BHF,l,BHF,p,l,F	✓	2.8	2.4	0.12	7
1.13	Ma,l,L,p,l,F	✓	5.0	4.6	0.077	3
1.14	A,L,L,P,Ml,F	✓	5.6	4.2	0.25	9
1.15	a,l,l,p,ML,F	✓	6.1	4.6	0.26	9
1.16	Ma,L,BHF,p,l,F	✓	4.9	3.9	0.20	14
1.17	MA,L,BHF,p,L,F		8.0	8.0	0.006	7
1.18	A,L,BHF,P,ML,F	✓	5.7	4.1	0.28	5
1.19	BHF,L,l,P,L,F	✓	4.3	3.7	0.13	3
1.20	BHF,l,L,p,l,F	✓	4.3	4.0	0.061	3
1.21	BHF,L,l,p,l,F	✓	5.0	4.5	0.094	3
1.22	BHF,L,L,P,l,F	✓	5.1	4.0	0.21	3

Table 1.1 Sequencing results. Twenty-two compounds with a variety of scores were resynthesized individually and matched to a library-member through retention time and MS^2 data. The Sequence column contains residue names in a comma-separated format from N-terminus to C-terminus with the C-terminus on-resin for library synthesis. Individual residue names are defined in Figure 1.5. The top score represents the score for the sequence given, while the next best score is the score of the candidate ranked second. The normalized difference is the difference between the two scores divided by the top score. The number of candidates is the number of potential sequences that the MS^2 spectra scored against could represent.

1.3 Discussion

We chose to implement CycLS as a database search using the library design to constrain the search space as compared to *de novo* sequencing. The reduced database size was critical in enabling whole-library sequencing runs to be completed quickly. Additionally, this approach is well-suited to sequencing cyclic peptides: A fragment mass-matching approach via virtual fragmentation of candidate molecules allows identification of the best matched peptide without assuming any ion series are present. Although this method ignores intuitive sequencing behaviors such as identifying amino acids by the mass differences between sequential ions in a series, those intuitions are less useful in the case of cyclic peptides where ion series may be short and overlapped.

However, this simple method becomes more error prone as data quality declines, especially when there are many candidate peptides with single differences in their sequence order. There are often only a small number of fragments capable of discrimination between candidates, and their ionization tendencies are often weak, necessitating high-quality MS^2 spectra to differentiate between candidates with confidence. We therefore believe that extensive preprocessing of the MS^2 data to increase data quality, as we have done, is critical to ensure sequencing quality.

Sodiated ion peaks (M+Na) that occupy the same mass as the protonated ion of another compound currently result in false positives when CycLS attempts to assign them a sequence. In future versions of CycLS, MS^2 spectra originating from sodium ions of library-members will be combined with their M+H spectra, boosting signal and removing them from false positive status. Incorporation of sodium ion data into sequencing will not, however, remedy the chromatographic overlap issues often caused by sodium ion peaks which appear at the same mass as another set of compounds' protonated ions. Obtaining high-resolution mass spectrometry data would enable M+Na and M+H peaks to be interpreted separately, and, on a system capable of a sufficiently narrow MS^2 isolation window, would also allow

M+Na and M+H ions of the same mass and retention time to be sequenced separately as well.

The five incorrectly sequenced cyclic peptides from the second validation study had scores that matched very closely to at least two candidates. In such cases limited resynthesis of high scoring candidates would provide the correct compound. If a score gap were observed within the candidate list, then its normalized difference could be used to gauge the confidence of that segregation of scores. For example, if the first three candidate sequences for a spectrum had nearly identical scores, but the fourth candidate sequence had a much lower score, only the first three sequences would merit resynthesis. Indeed, for 13 out of the 17 incorrectly identified sequences from the 400-member library, the correct sequence was among the top three candidates scored with CycLS.

The application of CycLS to the deconvolution of complex mixtures is primarily limited by experimental constraints. Chromatographic peak overlap resulted in many library members not having representation in the MS^2 data: 86% recovery for the unique mass library (40 compounds per mixture), and approximately 60% recovery in the mass-redundant library (150 compounds per mixture). Mixtures of greater size would increase the severity of data loss, while greater chromatographic separation and higher resolution mass spectrometry data would decrease data loss and/or allow for sequencing of more complex mixtures. These issues would be much less problematic in cases for which mixtures of only a few peptides are sequenced, such as in the deconvolution of one-bead-one-compound libraries selected from on-bead biological screening assays. For more complex mixtures, successful deconvolution will depend on the separation capabilities of the instrumentation at both the chromatographic and mass spectrometric levels.

1.4 Conclusions

In summary, we developed an automated sequencing program, CycLS, which facilitates the identification of cyclic peptides contained within large synthetic libraries. CycLS achieved an accuracy of 95% upon analysis of LCMS data from a 400-member cyclic peptomer library,

selecting from a virtual library design containing 234,256 compounds. A separate, 1800-member library of cyclic hexapeptides and hexapeptomers with high mass redundancy due to the presence of isobaric sequence variants, an accuracy of 77% was achieved as verified using 22 independently synthesized compounds.

We have accomplished our goal of rapid and accurate whole-library sequencing, but many opportunities for improvement remain. In the future, the sequence-scoring scheme will be revised to take the information content of each fragment into account. Because it can be assumed that there will be at least one static position in an OBOC library – at the last split step – that static position can be used as an anchor to start building the sequence from. By aligning fragments of varying lengths containing the anchor position, sequencing information can be obtained. This will allow fragments that contain little or no useful sequencing information to be weighted less and fragments which give solid sequencing information to be weighted more. Additionally, the initial scoring can be further improved by automated detection and increased weighting of fragments that discriminate between candidates, similarly to the post-sequencing re-examination performed by Redman, et al.[28] CycLS enables quick and easy acquisition of large datasets of synthetic cyclic peptides, especially for assay systems which make linearization less attractive.

1.5 Supplementary Figures and Discussion

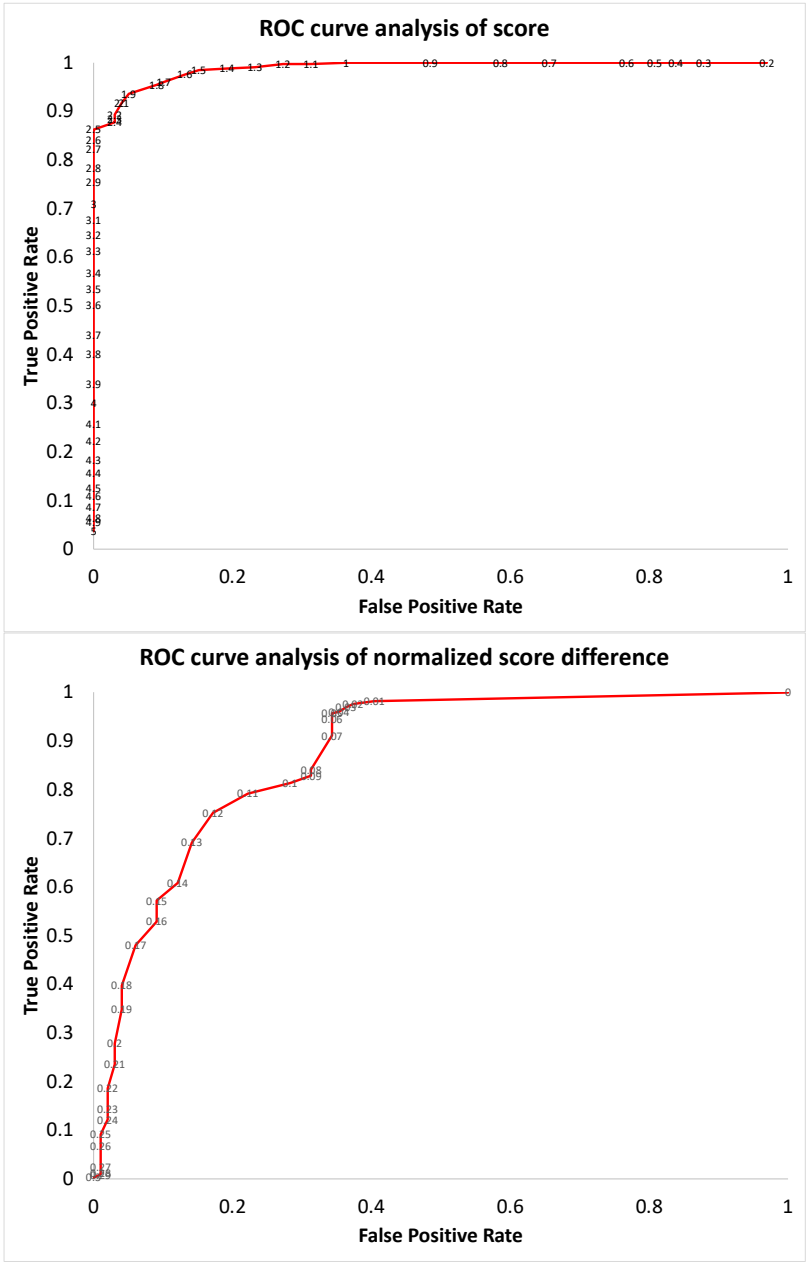


Figure S1.1 Receiver operating characteristic (ROC) analysis of CycLS score and normalized score difference. Various thresholds for score or normalized difference yielding varying amounts of true and false positives lie along each curve.

1.5.1 Additional Discussion of Scoring Metric Characteristics

Each sub-library of the unique mass validation library was first analyzed separately, comparing the known sequence to the highest-scoring sequence given by CycLS. Spectra were sorted into correctly sequenced, incorrectly sequenced, and false positives. The false

positives were spectra which were analyzed due to matching a virtual library member, but not to a member of the library as synthesized. In addition to calculating the fraction of MS^2 spectra matching to experimental library members which were correctly sequenced, the scores assigned by CycLS and the normalized difference between the two scores were compiled for receiver operating characteristic (ROC) analysis.

The raw scores assigned by CycLS were assigned to the true positive and false positive categories, with incorrectly sequenced spectra which represented members of the library as synthesized left out. The resulting curve had an area of 0.96, meaning that the scores' magnitude account for nearly the entirety of their classification as true or false positives. A score of 2 was deemed useful as a threshold under which spectra are likely sequenced incorrectly, as it reduced the false positives to 4% while preserving 92% of the true positives. This score threshold may remain useful in other data sets if the character of false positives proves to be somewhat universal between library designs.

The normalized difference between the scores assigned to the top two candidates by CycLS underwent a similar analysis. The area under the curve was 0.89, meaning that this metric is useful in classifying the true and false positives, but not as useful as the raw scores in this data set. A normalized score difference of 0.01 removes 60% of the false positives, and many true positives are excluded as the threshold is raised. Therefore, any threshold between 0.01 and 0.1 may be appropriate depending upon how many incorrect sequences are acceptable. All incorrectly sequenced peptides from the mass redundant library had a normalized difference smaller than 0.05.

1.5.2 Additional CycLS Functionality

The spectral processing used in CycLS has an automated noise threshold determination component not described fully in the discussion section for concision. While reading in the MS^2 spectra, all observed peak intensities are saved to a list to obtain the largest sample of the population of MS^2 peak intensities possible. This includes peaks from spectra which are removed from consideration in future pre-processing steps. Due to the

abundance of low-intensity peaks, nearly all of which arise from electronic noise, we made the initial assumption that all peaks arising from peptide fragmentation are likely to be within the upper 10^{th} percentile of peak signal intensities. A kernel density estimate was generated using the lower 90^{th} percentile, which resembled a normal distribution with a fat right tail. The resulting curve was treated as an approximate probability density function. The noise threshold can then be adjusted to any desired probability of a peak being noise. The default behavior is to set the noise threshold such that any peak above the threshold has $p < 0.05$ of being the result of noise, which we have observed to be only slightly lower than we would set the noise threshold by visual inspection.

1.6 Abbreviations

ACN, acetonitrile; COMU, (1-Cyano-2-ethoxy-2-oxoethylidenaminooxy)dimethylamino-morpholino-carbenium hexafluorophosphate; DBU, 1,8- Diazabicyclo[5.4.0]undec-7-ene; DCM, dichloromethane; DMF, N,N-dimethylformamide; DIPEA, diisopropylethylamine, DMSO, dimethylsulfoxide; Fmoc, 9-fluorenylmethoxycarbonyl; HATU, 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxide hexafluorophosphate; LCMS, liquid chromatography mass spectrometry; MeOH, methanol; N2, nitrogen; PTFE, Polytetrafluoroethylene; RP, reverse phase; SPPS, solid phase peptide synthesis; TFA, trifluoroacetic acid; OBOC, one-bead one-compound; ROC, receiver operating characteristic; FPR, false positive rate; TPR, true positive rate; BIC, base peak ion chromatogram.

1.7 Methods

All chemicals were commercially available and used without further purification except where mentioned.

1.7.1 Equipment and Analytical Methods

Composition and purity 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). A mixture of water (0.1% formic acid) and ACN (0.1% formic acid) was used as an eluent, with results being analyzed by MassLynx4.1. Runs were 12 minutes long with 1.2 mL/min flow rate; ACN concentration was increased stepwise from its starting concentration (30% 0-2 min, linear increase to 100% 2-10 min, 100% 10-12 min). Tandem MS runs were performed on an 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) or a 2.2 µm C18 column (Acclaim 120 250×2.1 mm, Thermo Scientific). A mixture of water (0.1% formic acid) and ACN (0.1% formic acid) was used as an eluent, with results being analyzed by XCALIBUR version 2.2 SP1.48. Runs on the 30 mm column were 6.5 min long with a 0.6 mL/min flow rate; ACN concentration was modified stepwise from its starting concentration of 10% (10% 0-0.5 min, linear increase to 100% 0.5-3.5 min, 100% 3.5-5.0 min, 10% 5.0-6.5 min). Runs on the 250mm column were 32.7 minutes long with a 0.5 mL/min flow rate; ACN concentration was increased stepwise from its starting concentration (40% 0-2 min, linear increase to 85% 2-22 min, 100% 22-27 min, 40% 27-32.7 min).

Tandem MS data of the unique mass validation library was acquired on the Orbitrap Velos Pro using an “Nth Order Double Play” scan event with the LC method described for the 30mm column, and for the cyclic hexapeptomer library and resynthesized compounds using the method described for the 250 mm column. The five most intense non-isotopic peaks (Xcalibur setting) were selected for MS^2 acquisition at every MS^1 acquisition. The MS^n activation was by collision induced disassociation with a normalized collision energy of 35.0, isolation width of 2.0 m/z, Activation Q. of 0.250, and activation time of 10.0 ms, and analyzed in the ion trap.

Raw Thermo MS^2 data from the Orbitrap Velos Pro was converted to mzML format using the msconvert program included in the Proteowizard software suite[32]. Version 3.0.10577 was used with the following command, and the resulting mzML files were used in all further analysis.

*msconvert --simAsSpectra *.raw --32 --zlib --filter "peakPicking true 1-" --filter "zeroSamples removeExtra"*

MzMine 2.23[33] was used to examine and generate chromatograms and spectra for characterization of the cyclic hexapeptomer sub-libraries and resynthesized compounds.

1.7.2 Reagent Synthesis

1.7.2.1 Fmoc addition to isotopically labeled leucine

Isotopically labeled leucine was purchased unprotected from Cambridge Isotope Laboratories, Inc. and Fmoc was added. The procedure used herein was adapted from Malkov et al.[34] and Marecek et al.[35] Isotopically labeled leucine (1 g) was added to a separate flask containing 30 mL of distilled water 80% saturated with sodium bicarbonate, then agitated by sonication until dissolved. Fmoc succinimide (Fmoc-OSu) (2.82 g, 1 molar equivalent) was added to a separate round-bottom flask in an ice bath containing 30 mL dioxane. The dissolved leucine was added dropwise to the Fmoc solution over 10 minutes with rapid stirring. The mixture was stirred 2 h on ice, then allowed to come to room temperature overnight with continued stirring. The mixture was then evaporated at 25˚ C and brought up in a minimal volume of DMF before reverse-phase purification over a SNAP Ultra C18 30g cartridge (Biotage), eluting with water (0.1% TFA)/ACN (0.1% TFA). The resulting fractions were analyzed on the Waters LC-MS system, pooled, and evaporated to a white solid.

1.7.2.2 N-methylation of Fmoc amino acids

All N-methylated amino acids used in the library synthesis were synthesized via the following procedure: Fmoc amino acid and paraformaldehyde (1:1 by mass) were added to a round-bottom flask containing 75 mL toluene per 5 g of amino acid. The flask was heated to 90˚ C before 0.2 molar equivalents of camphorsulfonic acid was added. The flask was stirred for at least 2 h, then the solution was evaporated to 3mL volume per gram of amino acid. Five mL of ethyl acetate per gram of amino acid was added. The solution was then transferred to

a separatory funnel and washed twice with an equal volume of water saturated with sodium bicarbonate and once with an equal volume of brine. The organic layer was dried with magnesium sulfate, filtered into a round-bottom flask, then evaporated to a minimal volume. Two mL of DCM per gram of amino acid was added to the flask and the oil or solid thoroughly dissolved. Once dissolved, an equal volume of trifluoroacetic acid was added, then stirred for 5 min. Finally, 3 molar equivalents of triethylsilane was added and let stir overnight. The solution was then evaporated to minimal volume before addition of 5 mL ethyl acetate per gram of amino acid, transferred to a separatory funnel and washed twice with an equal volume of distilled water and once with an equal volume of brine. The organic layer was dried, filtered into a round-bottom flask, and evaporated to a white solid, then analyzed for purity.

1.7.2.3 Resin loading procedure

Fmoc-Phe-OH and 2-chlorotrityl resin were placed in a vacuum desiccator with phosphorous pentoxide overnight. The phenylalanine was added to a flame-dried round-bottom flask. Dry DCM of sufficient volume for the resin to float freely and 4 molar equivalents of dry DIPEA was added to the flask, which was then sonicated until the phenylalanine dissolved completely. The resin was added to the flask and the flask purged of air with a flow of argon, after which the flask was agitated for 4 h. The resin was transferred to a solid-phase synthesis tube/manifold and washed with 2 resin-volumes of DMF (3x), then 2 resin-volumes DCM (3x), always keeping the solvent level above the resin. Finally, the resin was capped with a solution of 17:2:1 DCM:MeOH:DIPEA (2 resin-volumes, 3x, 15 min incubation each). Loading value was calculated via resin cleavage (1% TFA) followed by quantification by UV absorbance at 280 nm. Resin was loaded sparsely for library synthesis to avoid peptoid dimerization during the submonomer synthesis method[36].

1.7.3 Cyclic Hexapeptomer Library Synthesis

Linear peptomers were synthesized on L-phenylalanine 2-chlorotrityl resin (0.14 mmol/g) using extended Fmoc coupling (Fmoc amino acid/HATU/DIPEA in DMF, overnight)

and submonomer peptoid synthesis conditions (bromoacetic acid/DIC in DMF, 1 h, then amine in DMF, overnight). The linear peptomers were cleaved from resin with 30% HFIP in DCM. Cyclization was performed in dilute conditions (<3 mM peptomer in a solution of 1:1 ACN:THF) with COMU (2 molar equivalents) and DIPEA (10 molar equivalents) and stirred overnight at room temperature. Each sub-library was briefly purified using Isolute 103 SPE cartridge (200 mg/6 mL, Biotage).

1.7.3.1 Fmoc deprotection

Two resin-volumes of 2% DBU 2% piperidine in DMF were added to the resin and the tube was capped and agitated for 20 min. The resin was then drained and washed with 2 resin-volumes of DMF (x3) and 2 resin-volumes of DCM (x3).

1.7.3.2 Amino acid couplings

Due to low resin loading value, 12 molar equivalents of Fmoc amino acid, 11.4 molar equivalents of HATU, and 15 molar equivalents of DIPEA were used to maintain high concentration in a DMF volume large enough to cover the resin. Fmoc amino acid, HATU, and DIPEA were added to a vial, then solubilized in the minimal amount of DMF which covers the resin volume. The vial was set aside to react for 15 min before its contents were added to the resin. The SPPS tube was then capped and agitated overnight. The resin was then drained and washed with 2 resin-volumes of DMF (x3) and 2 resin-volumes of DCM (x3).

1.7.3.3 Peptoid synthesis

For the same reasons as above, the 30 molar equivalents of bromoacetic acid was added to a vial and solubilized in a minimal volume of DMF which still covers the resin volume. Fifteen molar equivalents of DIC was then added, and the vial capped, mixed, and reacted for 15 min. The vial's contents were then added to the resin. The SPPS tube was then capped and agitated for 1 h, after which the resin was drained and washed as above. Thirty molar equivalents of the amine of choice was then added directly to the resin and the volume was increased with a minimal amount of DMF before capping the tube and agitating

overnight. The reaction time was lengthened from 1 h to match the amino acid couplings for convenience. Finally, the resin was drained and washed as typical after an amino acid coupling (above).

1.7.3.4 Cleavage from resin

Two resin-volumes of 30% HFIP in DCM was added to the resin, the tube capped, then agitated for 30 min. The tube was then drained into a tared vial and washed into the tared vial with an equal volume of DCM three times. The above steps were performed a second time with an additional final wash of acetone. The contents of the vial were then evaporated.

1.7.3.5 Peptide cyclization

Cyclization was performed dilute in 10 mL of dry 1:1 ACN:THF. Five mL dry ACN and 10 molar equivalents of DIPEA were added to each vial containing a cleaved sub-library. The vials were then sonicated until all peptides were totally dissolved before addition of 5 mL dry THF, diluting the peptides to <3 mM. Two molar equivalents of COMU were added to each vial dropwise before the vial was stirred overnight. Finally, each vial was evaporated to a solid or oil and the mass obtained.

1.7.3.6 Purification of sub-libraries

Purification of each sub-library was performed on individual Isolute 103 ENV+ flash chromatography columns (200 mg/6 mL). Each sub-library was dissolved in a minimal volume of DMF (75 µL or less) and loaded onto the dry column. The column was washed with 10ml of water, followed by elution with 10 mL of methanol into tared vials. The methanol was then evaporated, and the vials weighed for final yield.

1.7.4 Characterization of the Cyclic Hexapeptomer Library

The composition and purity of each sub-library were tested by LC-MS (Waters system). Expected masses for the library were observed by mass spectroscopy (Oribtrap Velos Pro). The strong peak at 20.7 minutes in the BIC of each sub-library is a standard

cyclic peptide spiked into the analysis. Extra peaks observed at expected masses were due to sodium ions of other library masses or due to epimerization.

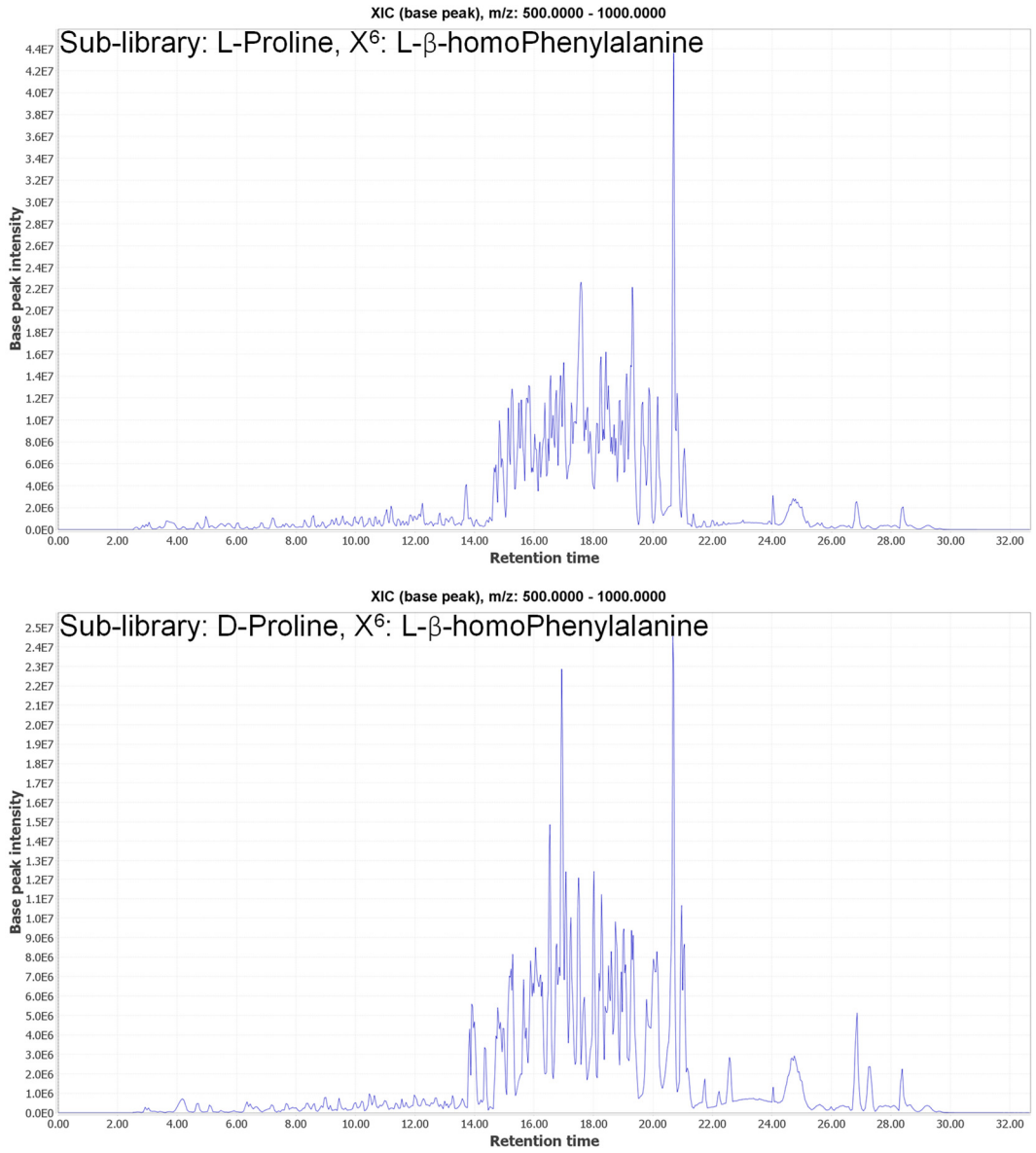


Figure S1.2.1 BIC of mass-redundant sub-libraries containing L-β-homoPhenylalanine at position six.

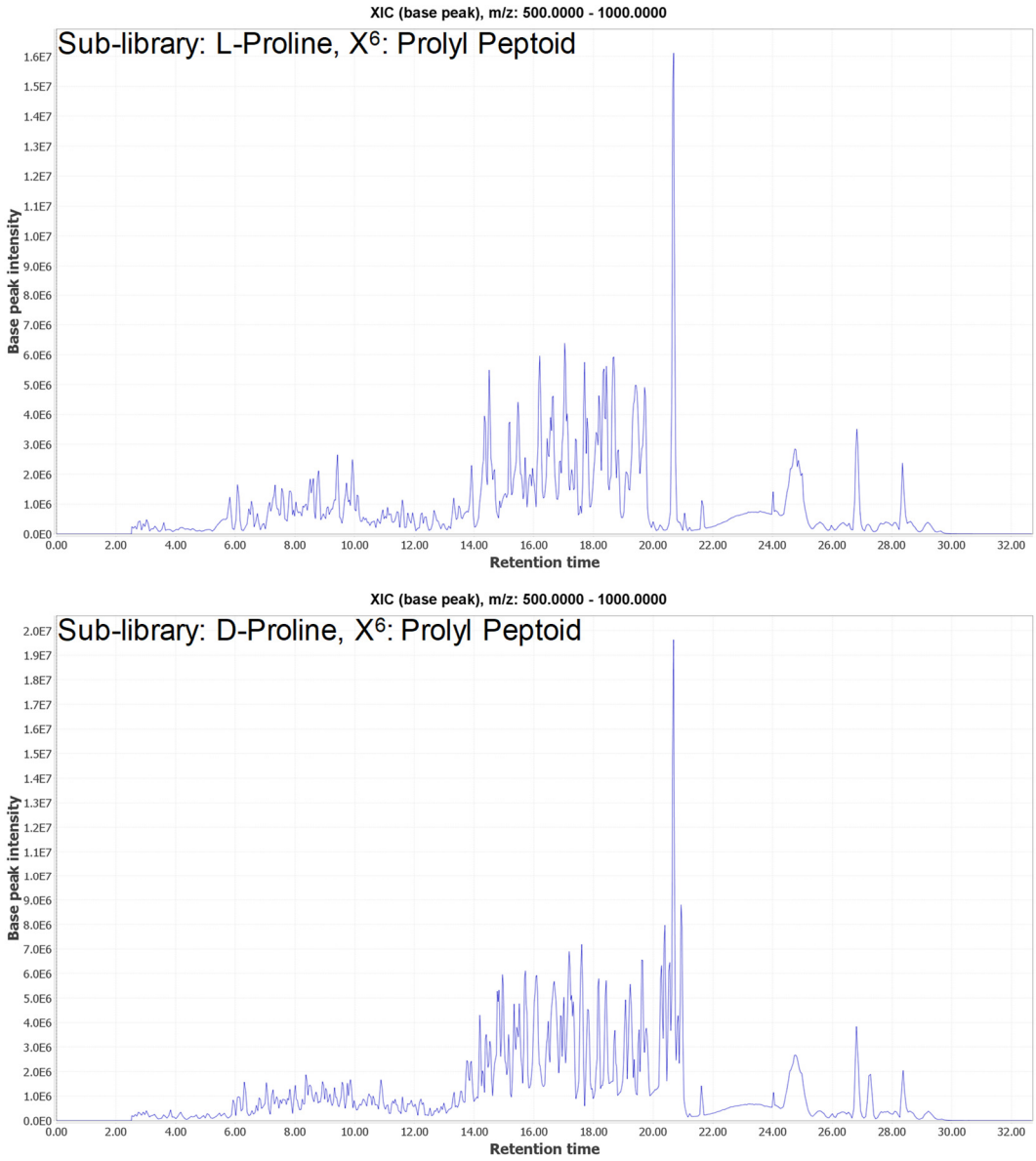


Figure S1.2.2 BIC of mass-redundant sub-libraries containing propyl peptoid at position six.

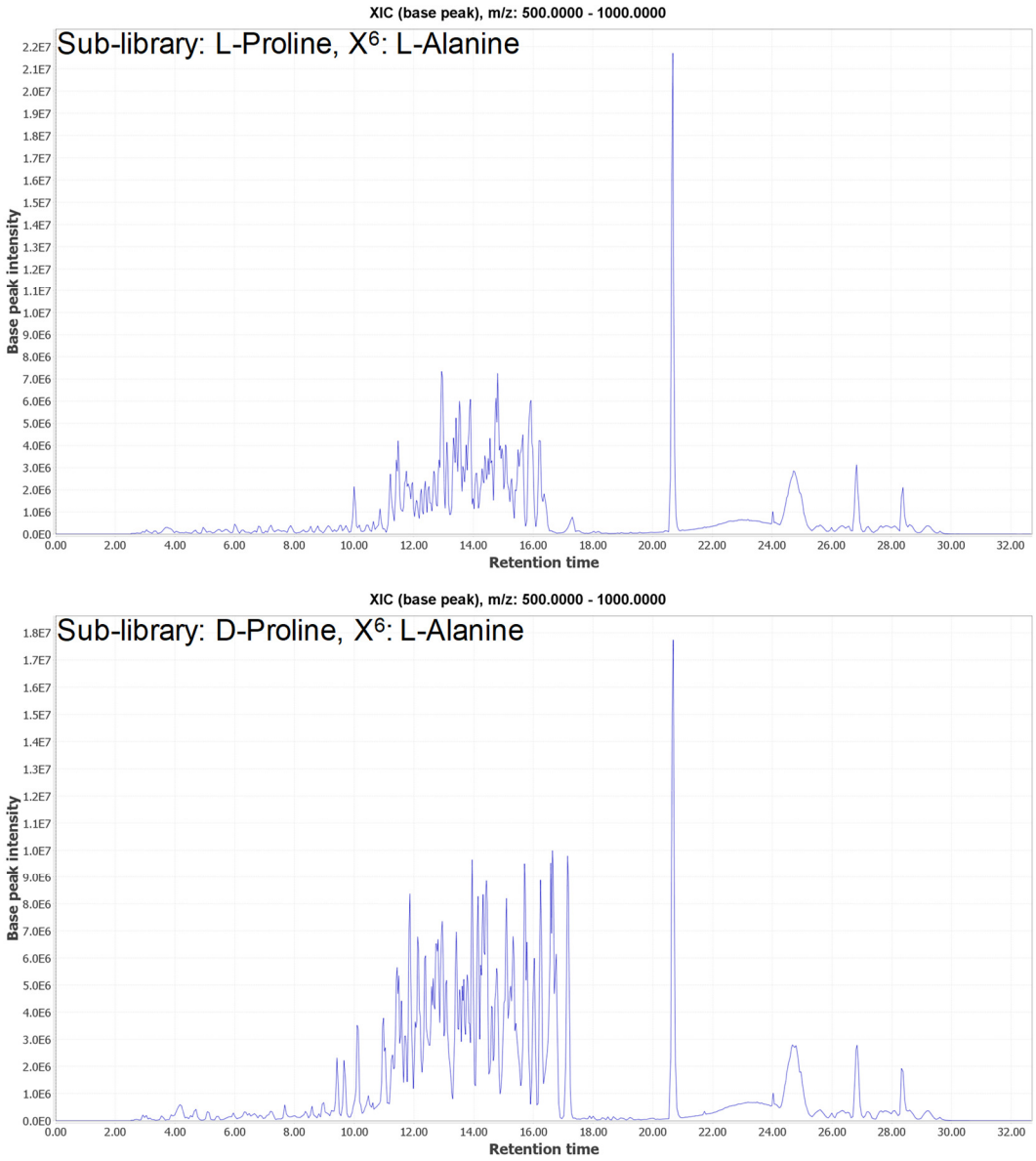


Figure S1.2.3 BIC of mass-redundant sub-libraries containing L-Alanine at position six.

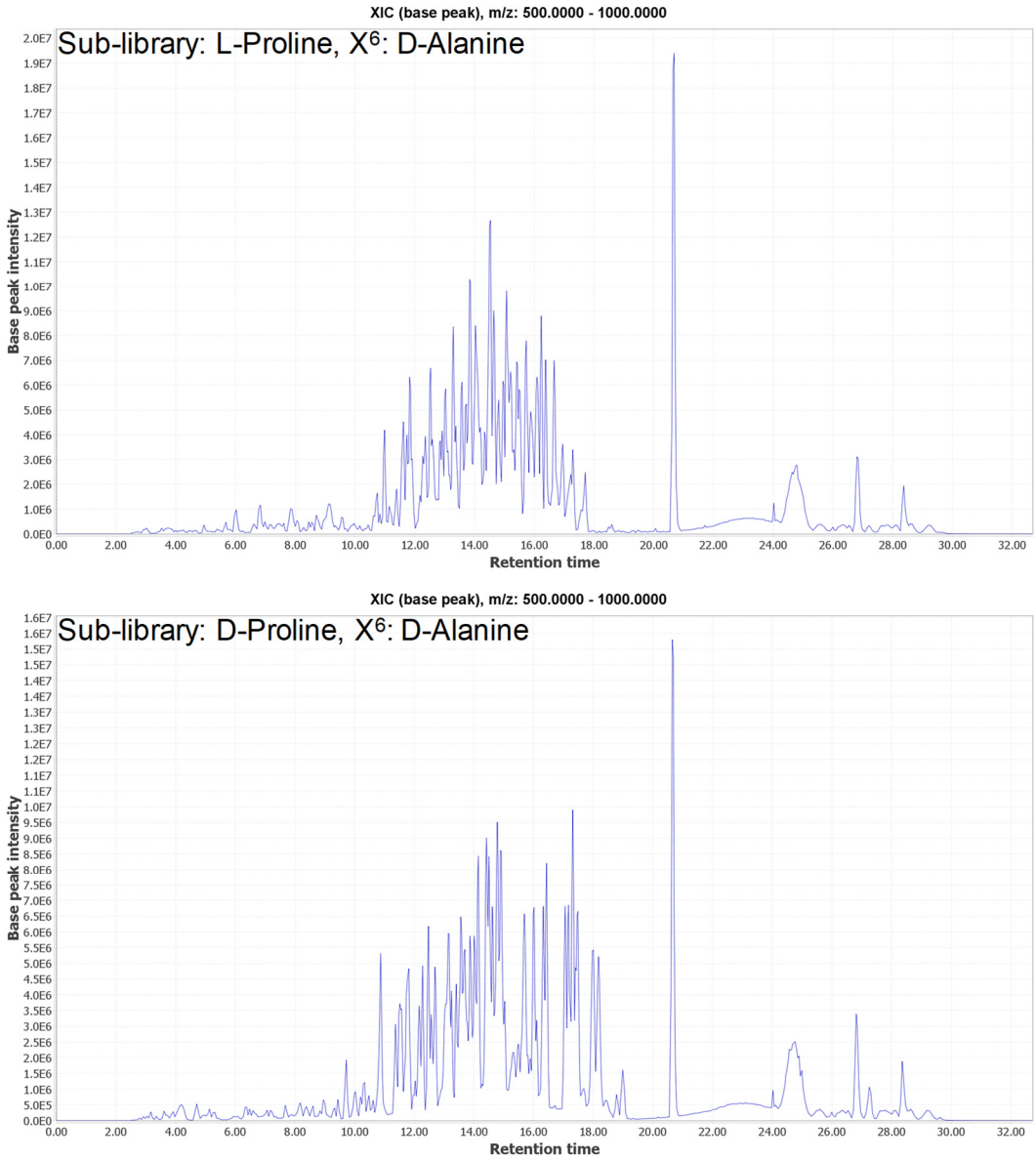


Figure S1.2.4 BIC of mass-redundant sub-libraries containing D-Alanine at position six.

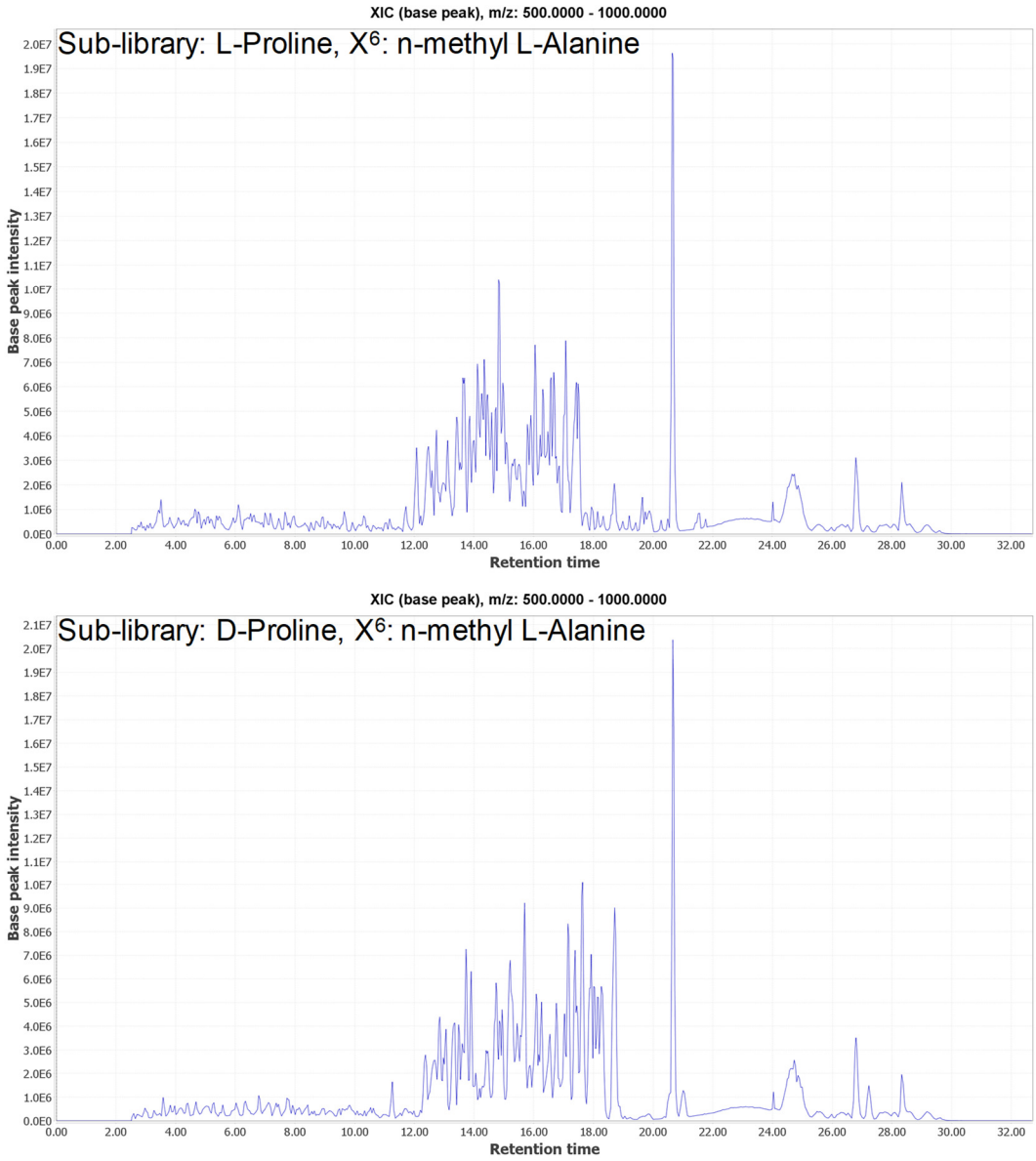


Figure S1.2.5 BIC of mass-redundant sub-libraries containing n-methyl L-Alanine at position six.

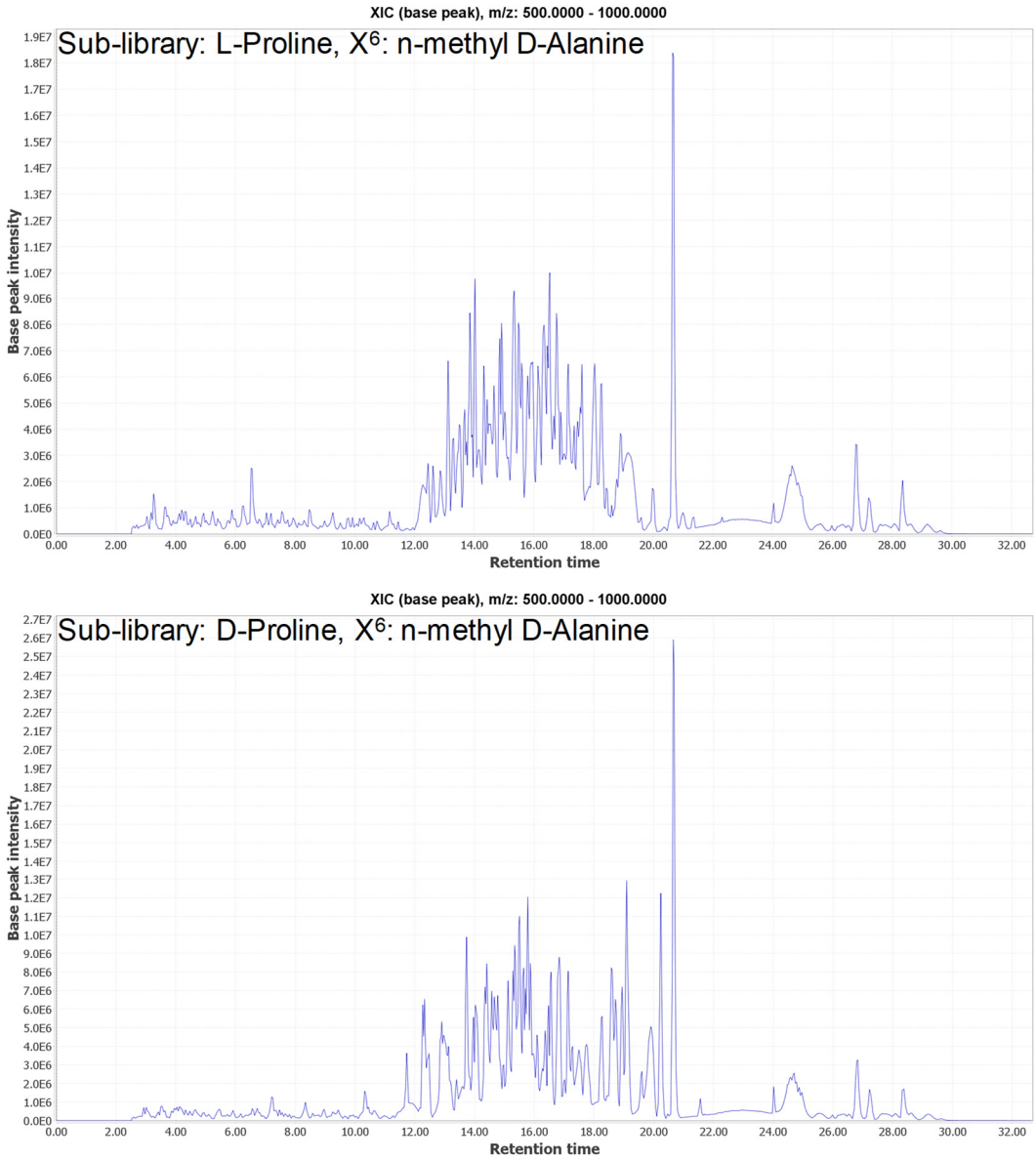


Figure S1.2.6 BIC of mass-redundant sub-libraries containing n-methyl D-Alanine at position six.

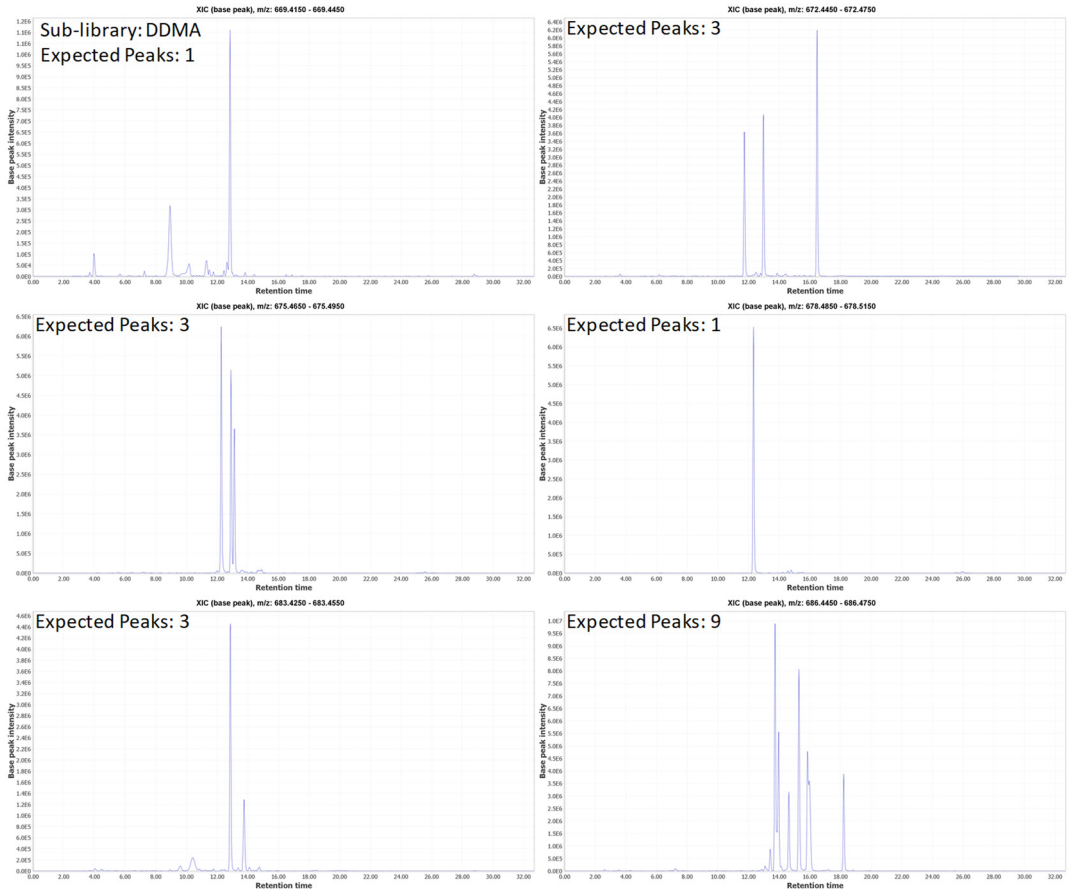


Figure S1.2.7 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

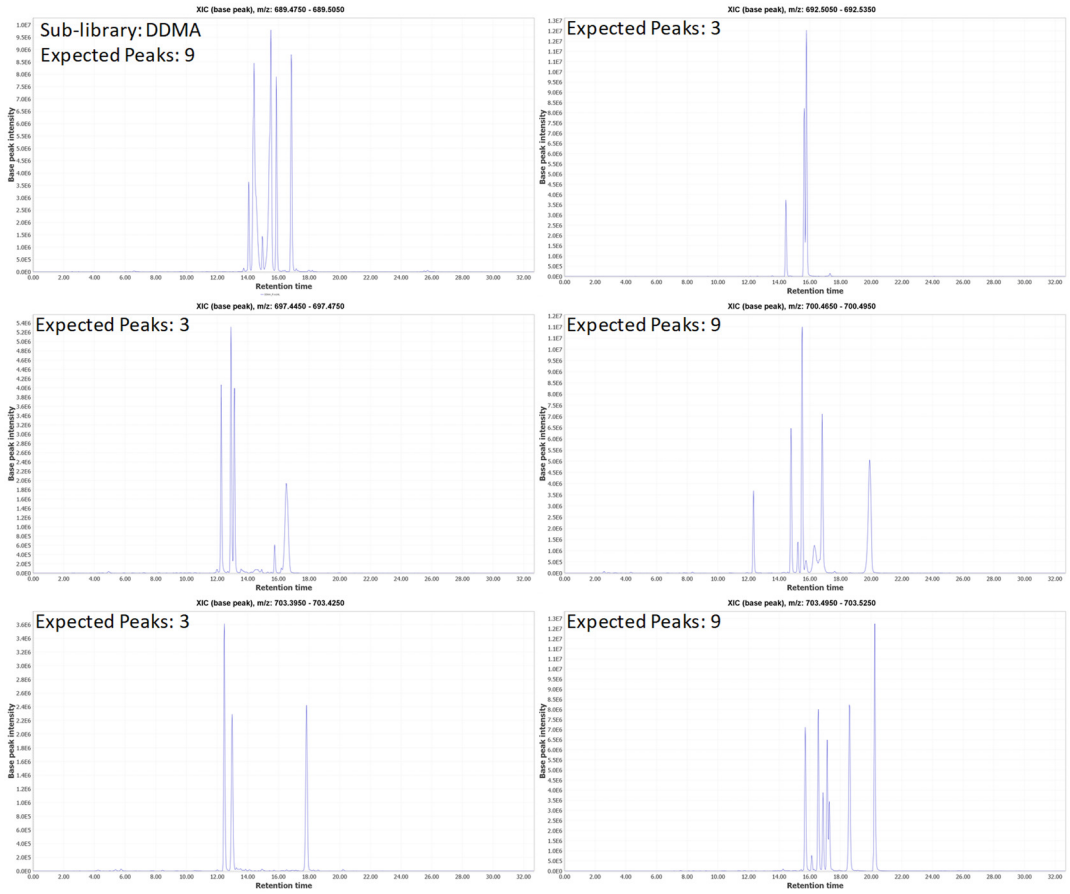


Figure S1.2.8 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

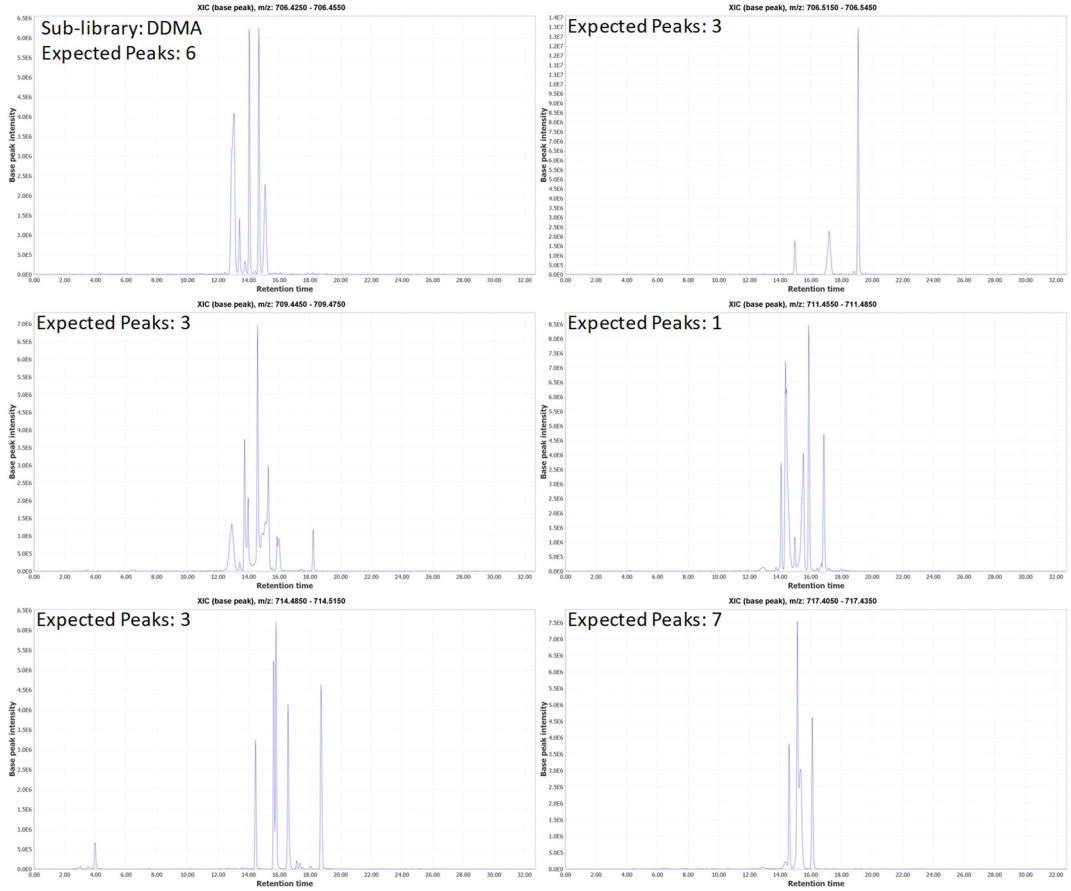


Figure S1.2.9 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

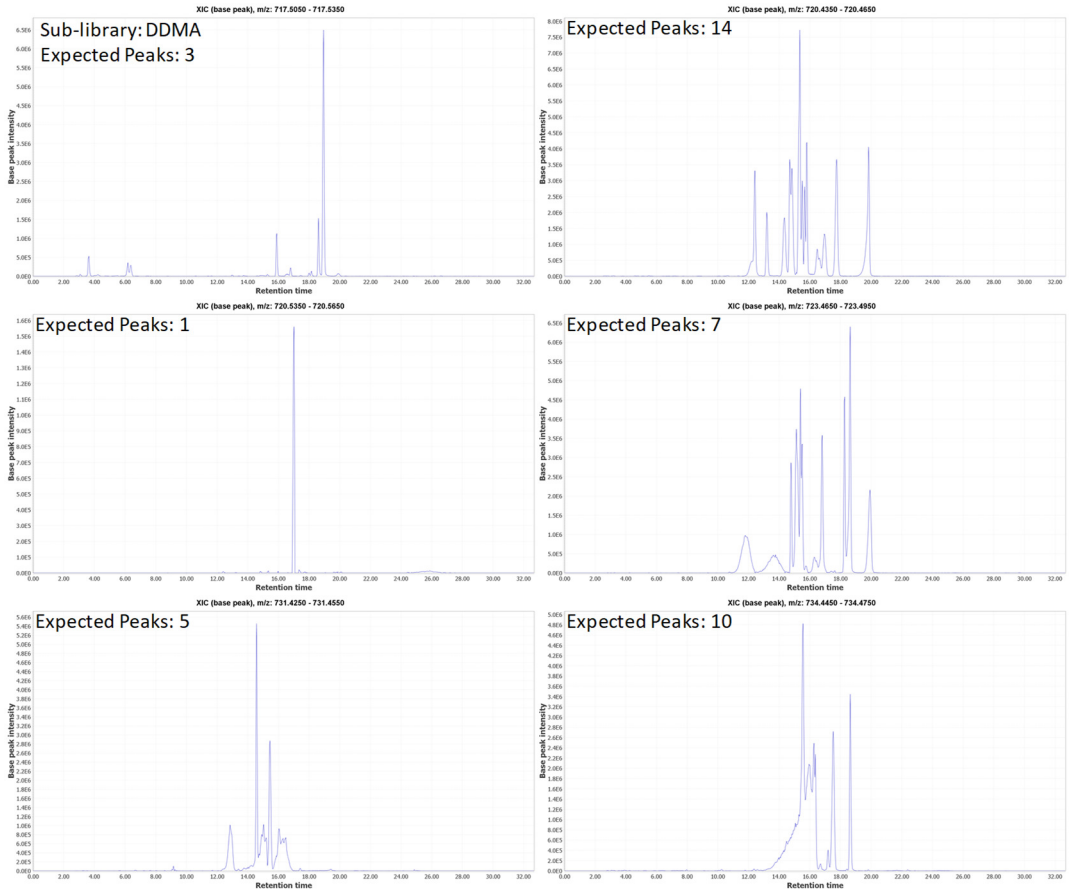


Figure S1.2.10 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

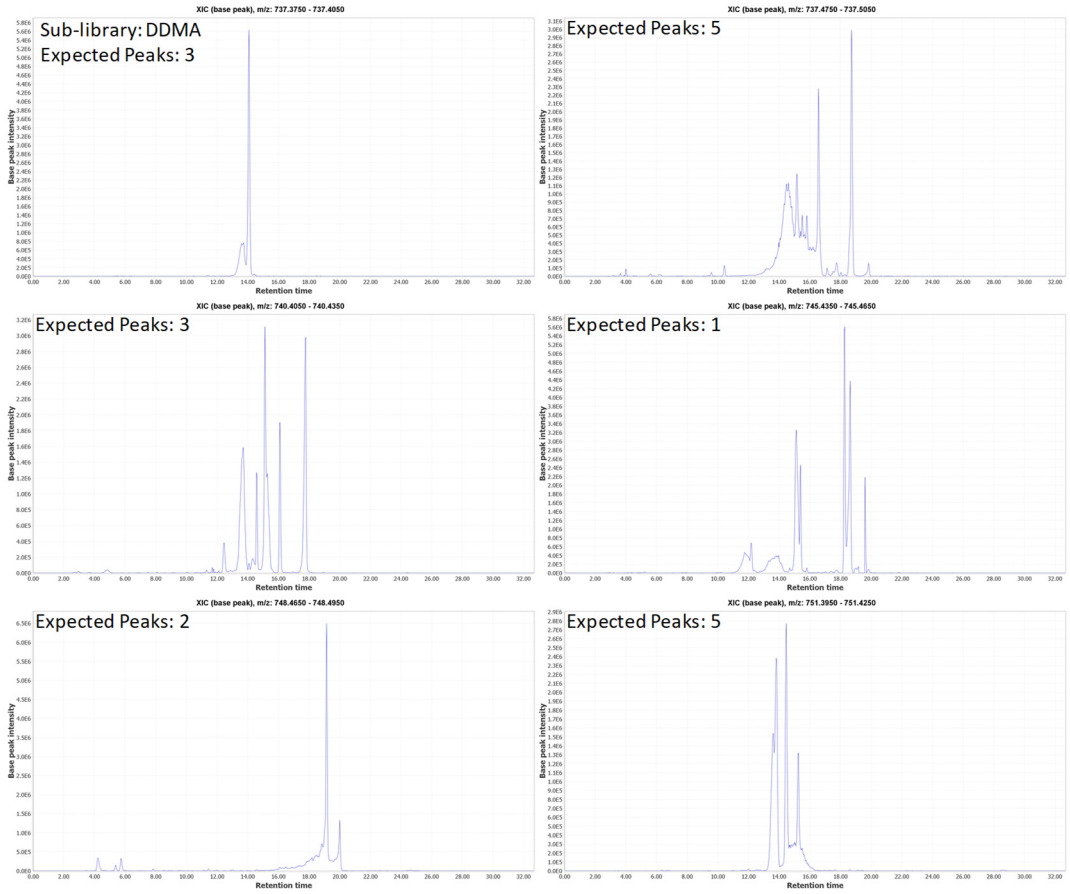


Figure S1.2.11 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

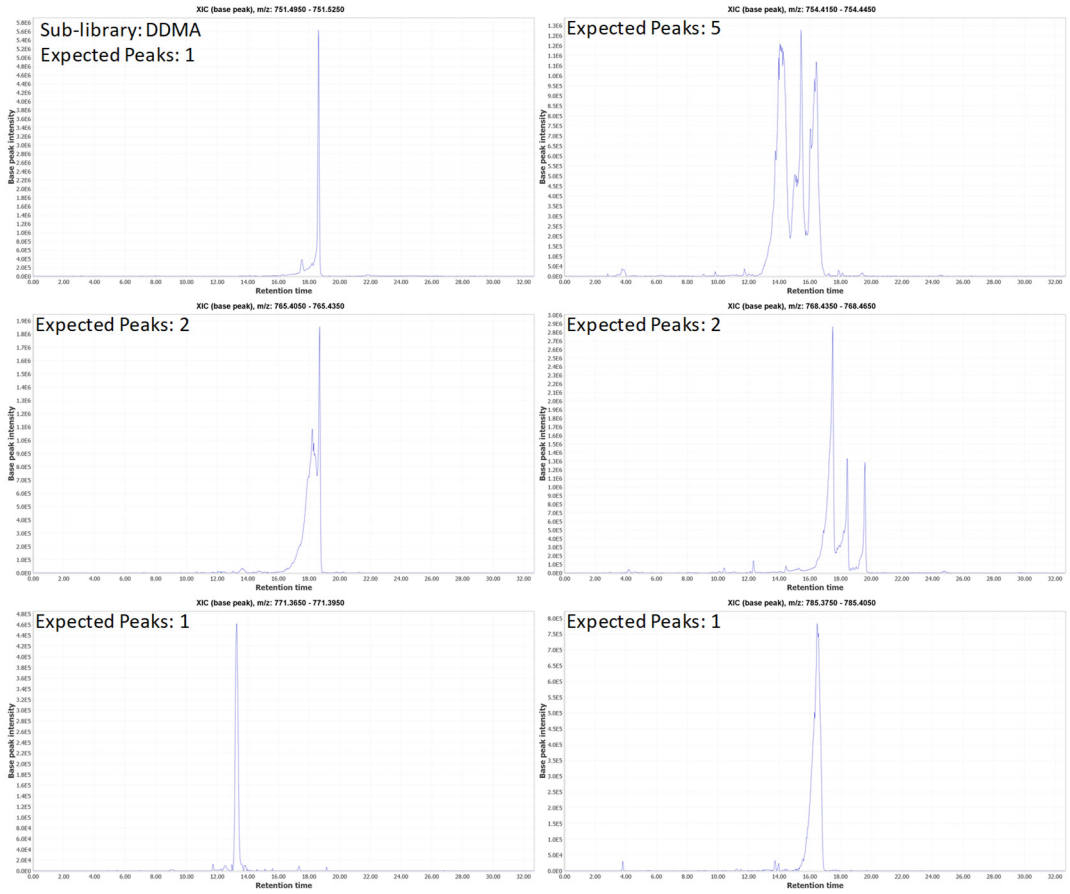


Figure S1.2.12 Examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library.

Table S1.1 Sub-library DDMA summary

Exact Mass	Expected major peaks	Observed major peaks	M+Na ion source mass	Peaks with Rt matches to M+Na source mass
668.4261	1	2		
671.4521	3	3		
674.4781	3	3		
677.5041	1	1		
682.4418	3	2		
685.4678	9	8		
688.4938	9	8		
691.5198	3	3		
696.4574	3	5		
699.4834	9	8		
702.4105	3	3		
702.5094	9	8		
705.4365	6	6		
705.5354	3	4		
708.4625	3	11	685	8
710.4731	1	7	688	7
713.4991	3	5	691	3
716.4261	7	4		
716.5251	3	3		
719.4521	14	14		
719.5511	1	1		
722.4781	7	>8		
730.4418	5	>5	708	Many
733.4678	10	>7		
736.3948	3	>2		
736.4938	5	>7		
739.4208	3	7		
744.4574	1	6	722	At least 3
747.4834	2	2		
750.4105	5	4		
750.5094	1	1		
753.4365	5	6		
764.4261	2	2		
767.4521	2	3		
770.3792	1	1		
784.3948	1	1		

Table S1.1 Summary of examination of single-ion chromatograms of expected library masses from the D-proline, n-methyl D-Alanine sub-library. Evidence for significant M+Na interference from other library compounds is noted at some library masses. At the noted library masses, M+Na peaks were identified by their retention time and intensity profile as a set. Major peaks were hand-curated to provide a general overview.

1.7.5 Resynthesis of Individual Compounds from the Mass-redundant Library

Resynthesized compounds were synthesized either manually or automatically without the use of isotopic labelling.

1.7.5.1 Manual Resynthesis

Peptides 1.1 through 1.6 were synthesized manually at 0.1 mmol scale on Fmoc-L-phenylalanine-loaded 2-chlorotrityl resin (0.8 mmol/g, Rapp Polymere) using accelerated amino acid coupling conditions (Fmoc amino acid/HATU/DIPEA in DMF, agitated at 50° C 1 h) with reduced equivalents compared to the couplings described above (4 aa/3.8 HATU/5 DIPEA). Couplings involving L-β-homophenylalanine used half the number of equivalents of amino acid. Coupling times and equivalents for peptoid couplings were also modified (8 bromoacetic acid/4 DIC, 1 h then 10 equivalents of the amine for 2 h). Couplings were monitored on the Waters LCMS system at each step and repeated if necessary. The linear peptomers were cleaved and cyclized by the method described above with the exception that the cyclization was performed entirely in ACN instead of 1:1 ACN:THF. Assessing correct or incorrect sequencing was possible of the crude cyclic peptomers, so they were not purified for this study.

1.7.5.2 Automated Resynthesis

Peptides 1.7 through 1.22 were synthesized on an automated peptide synthesizer (Prelude X, Gyros Protein Technologies). The synthesis was at a 0.05 mmol scale on L-phenylalanine load 2-chlorotrityl resin (0.8 mmol/g, Rapp Polymere). The coupling protocol began with a deprotection step using 2% DBU and 2% Piperidine in DMF. Three mL of deprotection solution was added to each reaction vessel, then the vessels were heated to 90° C and shaken for 1 min. This was repeated twice before four DMF washes (2 mL ea.), two DCM washes (2 mL ea.), and a final two DMF washes (2 mL ea.) to remove the DCM. Coupling reagents were then added: 3.8 equivalents COMU, 4 equivalents amino acid, and 4 equivalents DIPEA in 4 mL DMF. The coupling was shaken and heated to 90° C for 10 min before the wash series from above was repeated. Finally, 2 mL of 20% acetic anhydride and

1 mL of DIPEA was added to each reaction vessel and shaken for 5 min without heating to cap the remaining uncoupled peptide before another set of washes as above. This procedure was repeated for each coupling.

Cleavage occurred on the Prelude X as well, with two 45-min incubations with 4 mL 30% HFIP in DCM. Following each incubation with HFIP, the reaction vessels were collected into a waiting collection vial, after which two washes with 4 mL DCM were also collected. Cyclization was performed as with the manually resynthesized compounds.

1.7.6 Characterization of Resynthesized Compounds

As the resynthesized compounds were not purified, single ion chromatograms of their masses from analysis of crude peptide on the Velos Pro Orbitrap mass spectrometer are included here. Figures S1.3.1 through S1.3.11 abbreviate the full compound numbers by leaving out the preceding chapter designation, with the full compound number present in each figure's caption.

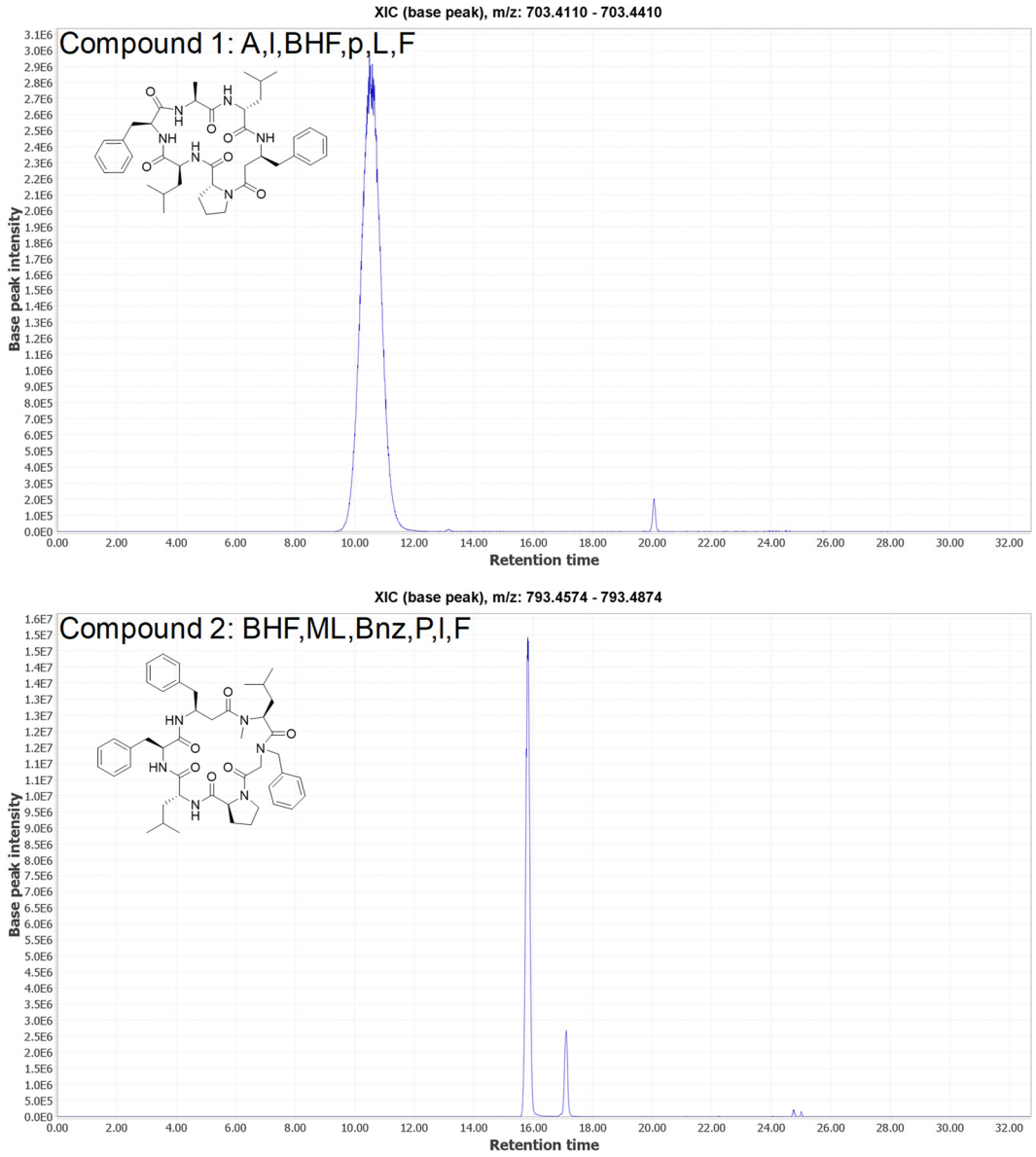


Figure S1.3.1 Crude single ion chromatograms of resynthesized compounds 1.1 and 1.2. The minor peak visible for compound 1.1 was a contaminant by MS^2. The minor peak visible for compound 1.2 was confirmed to be an epimer by MS^2.

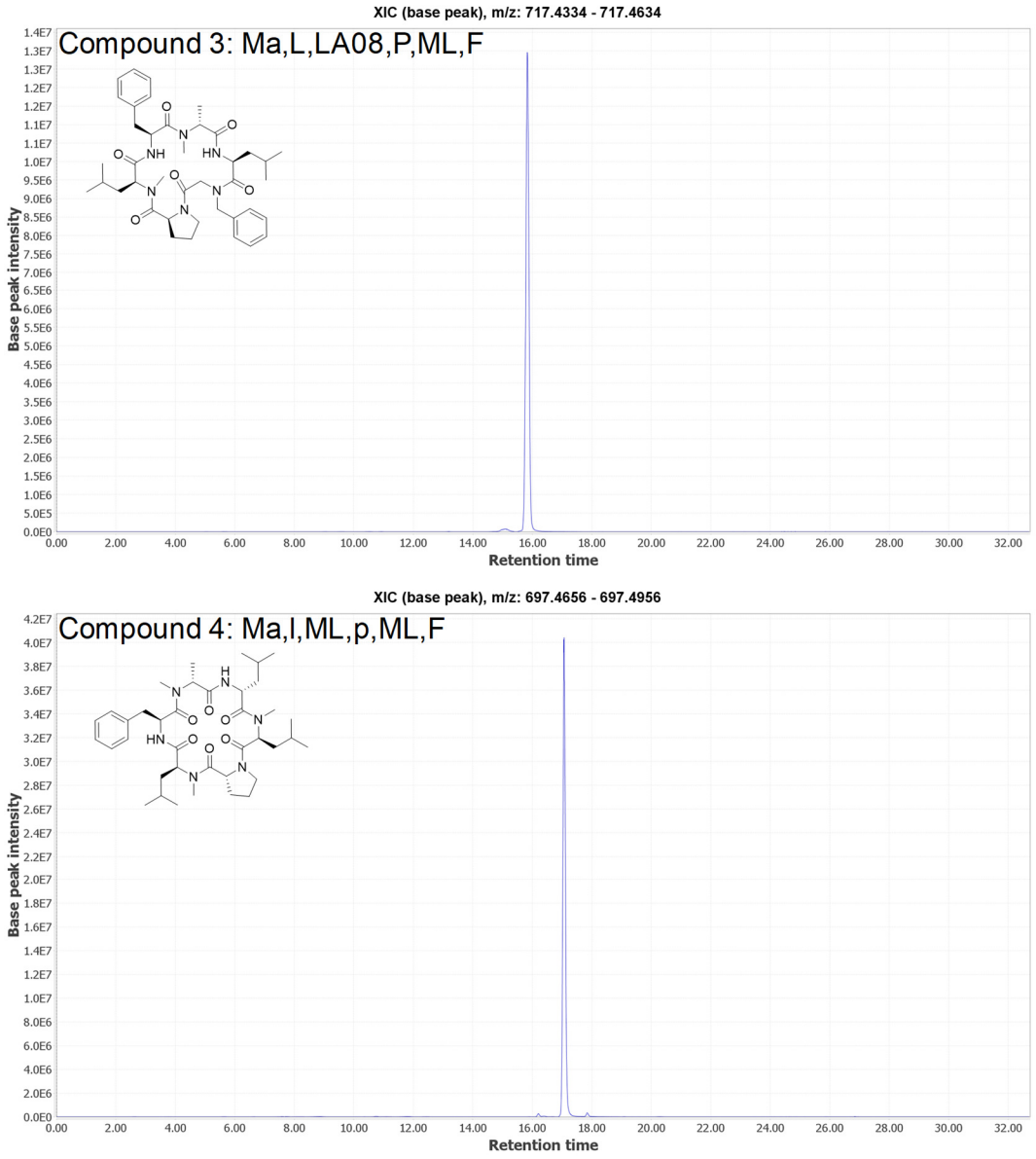


Figure S1.3.2 Crude single ion chromatograms of resynthesized compounds 1.3 and 1.4.

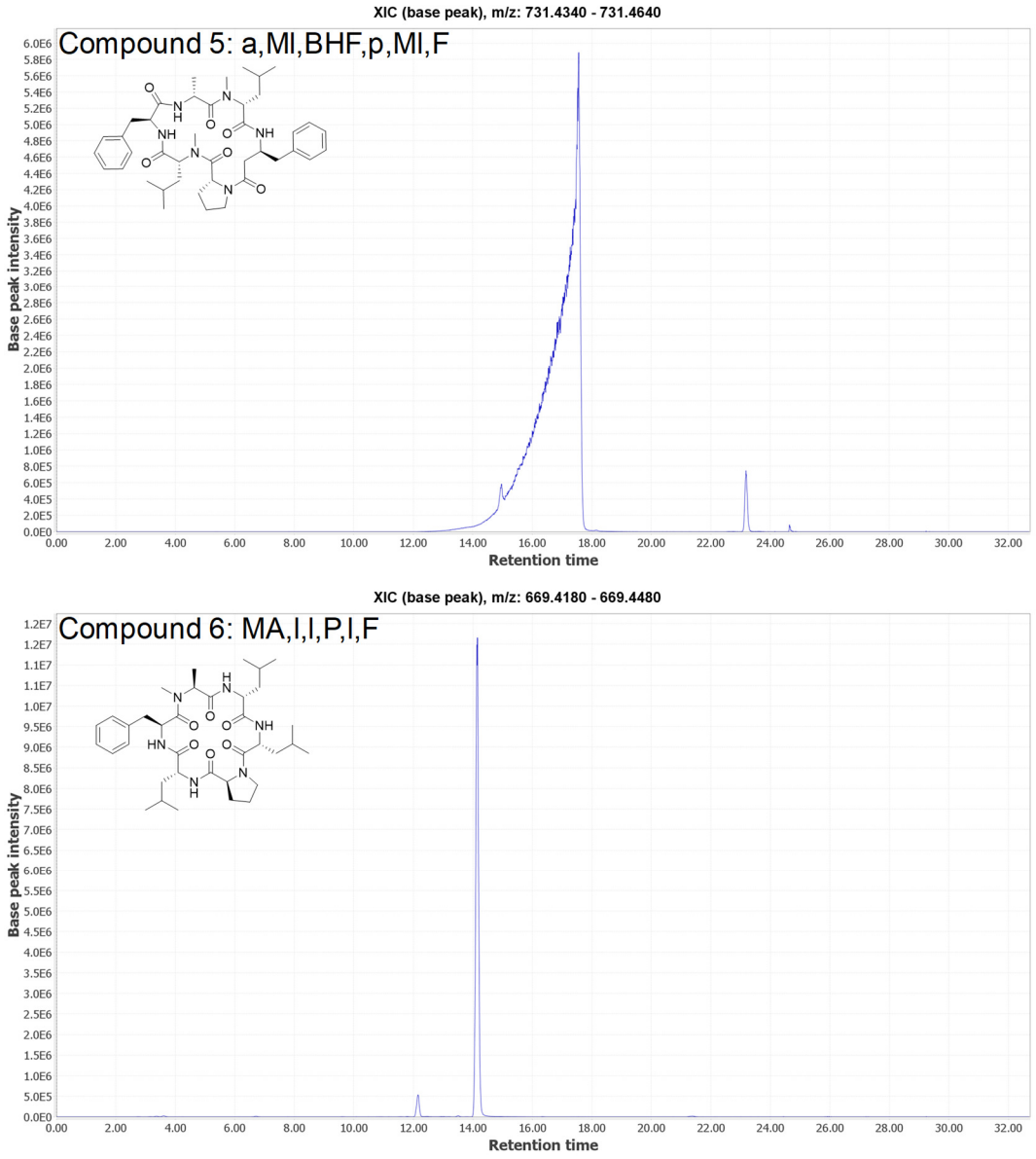


Figure S1.3.3 Crude single ion chromatograms of resynthesized compounds 1.5 and 1.6. The minor peak observed near 23 minutes for compound 1.5 was a contaminant by MS^2. The minor peak present for compound 1.6 was confirmed to be an epimer by MS^2.

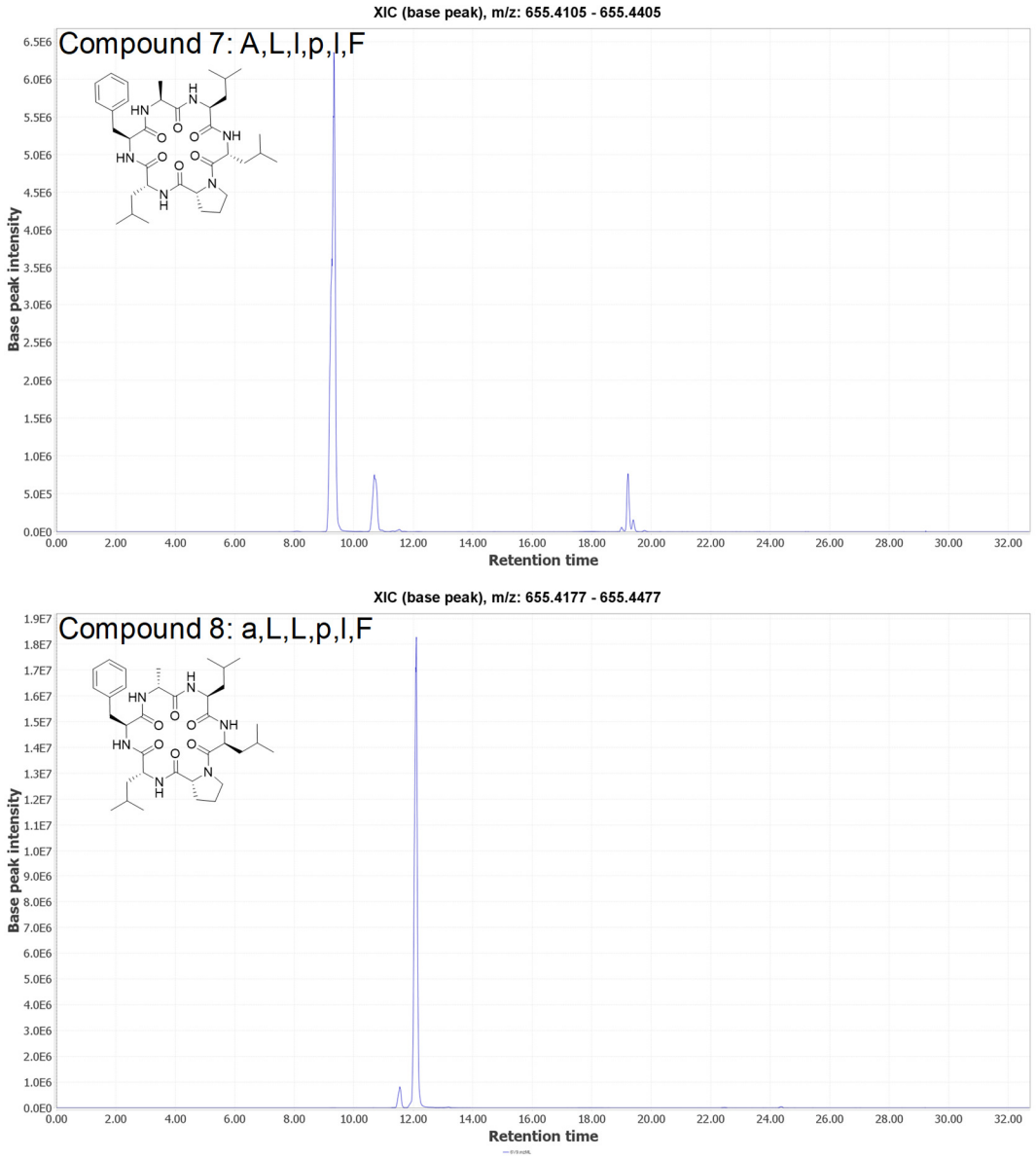


Figure S1.3.4 Crude single ion chromatograms of resynthesized compounds 1.7 and 1.8. The minor peak observed near 11 minutes for compound 1.7 was due to epimerization; the later minor peak was a contaminant or doubly-charge dimer by MS^2. The minor peak observed for compound 1.8 was confirmed to be an epimer by MS^2.

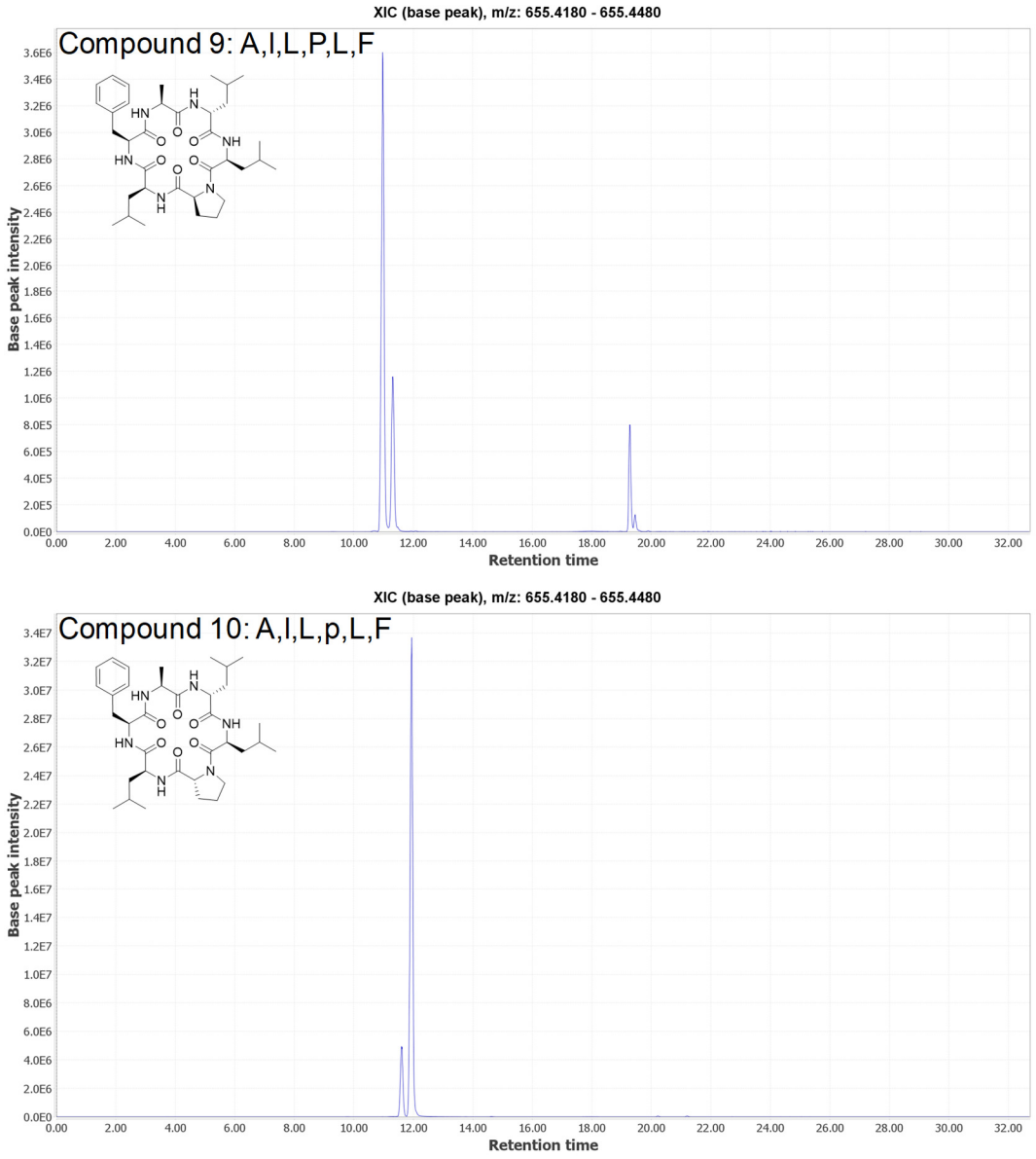


Figure S1.3.5 Crude single ion chromatograms of resynthesized compounds 1.9 and 1.10. The minor peaks distal to compound 1.9 were either contaminants or doubly charged dimers as confirmed by MS^2. The minor peaks nearby the major peaks of both compounds were confirmed as minor epimers by MS^2.

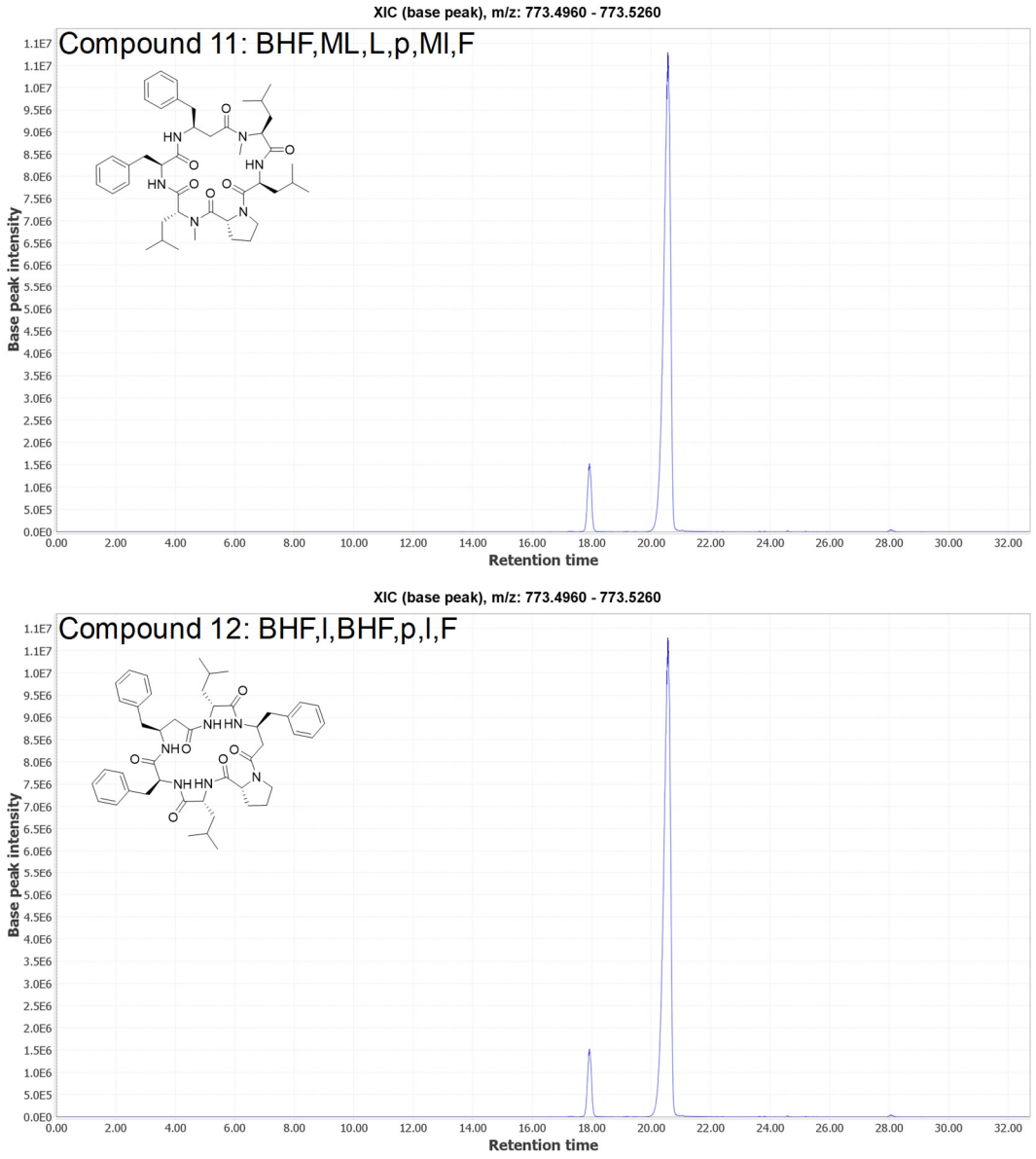


Figure S1.3.6 Crude single ion chromatograms of resynthesized compounds 1.11 and 1.12. The minor peaks observed for both compounds were epimers by MS^2.

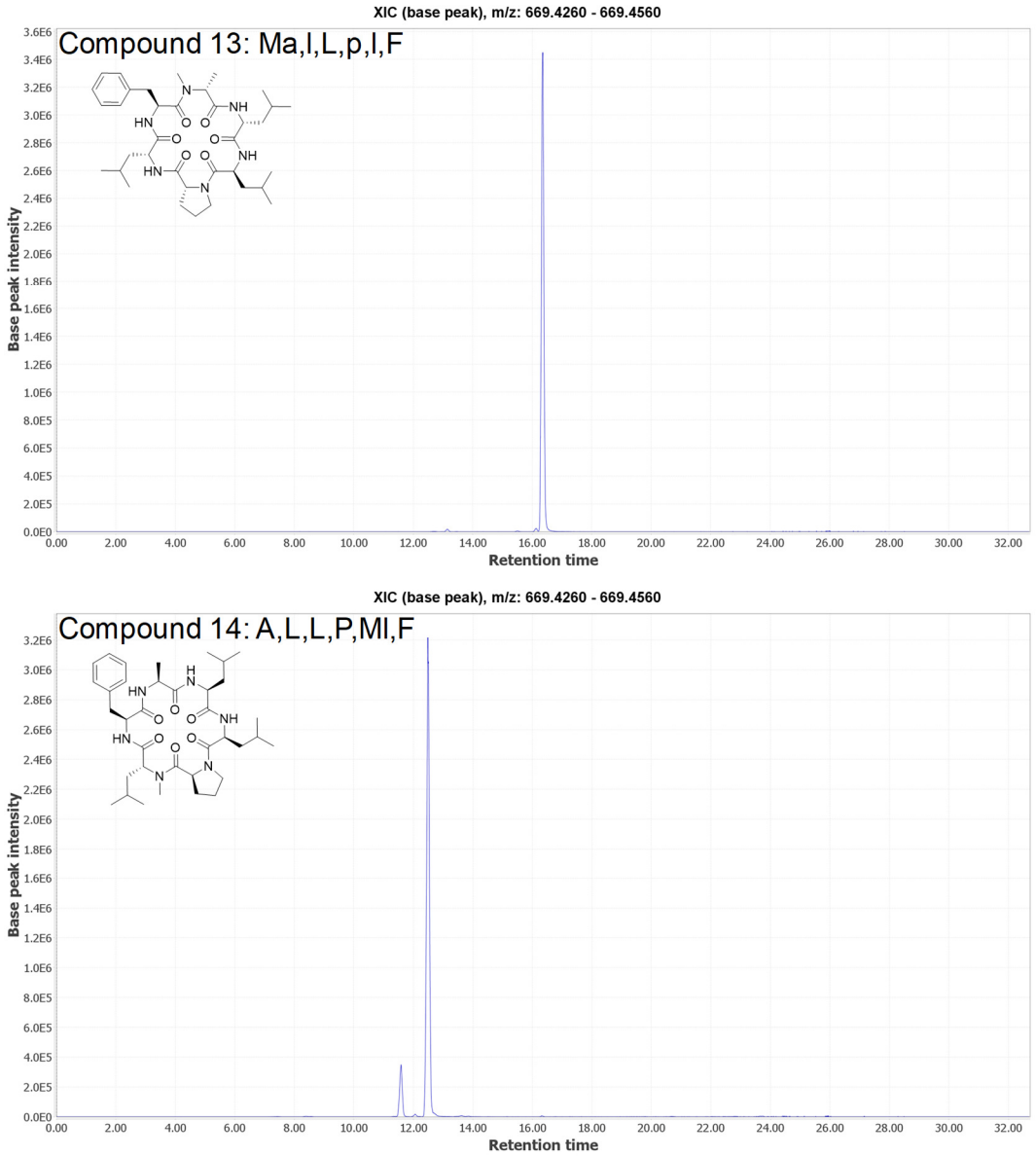


Figure S1.3.7 Crude single ion chromatograms of resynthesized compounds 1.13 and 1.14. The minor peak observed near the major peak of compound 1.14 was confirmed to be an epimer by MS^2.

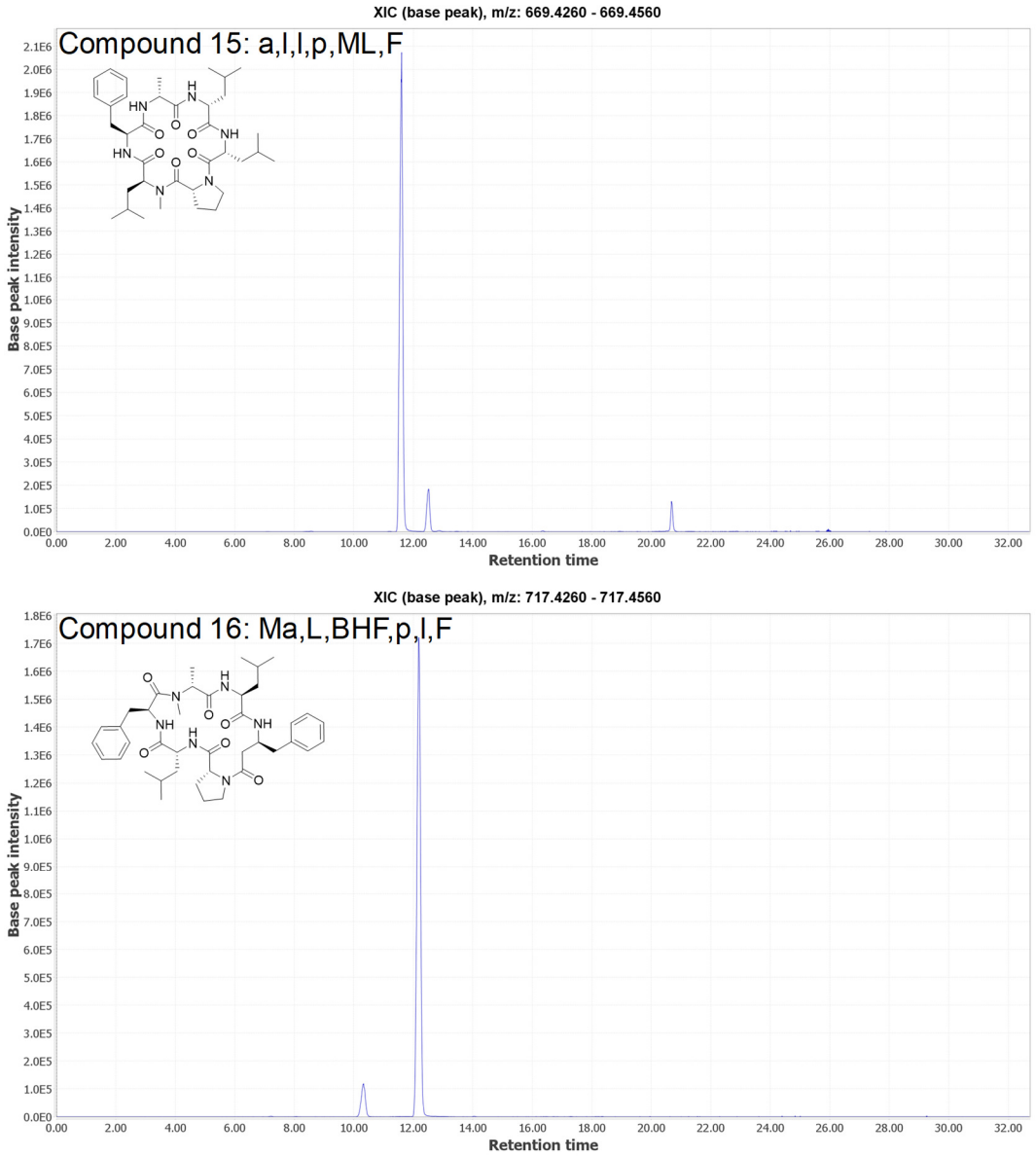


Figure S1.3.8 Crude single ion chromatograms of resynthesized compounds 1.15 and 1.16. For compound 1.15, the nearby minor peak was an epimer by MS^2, while the distal minor peak was either a contaminant or doubly charged dimer by MS^1. The minor peak in compound 1.16's chromatogram was also an epimer by MS^2.

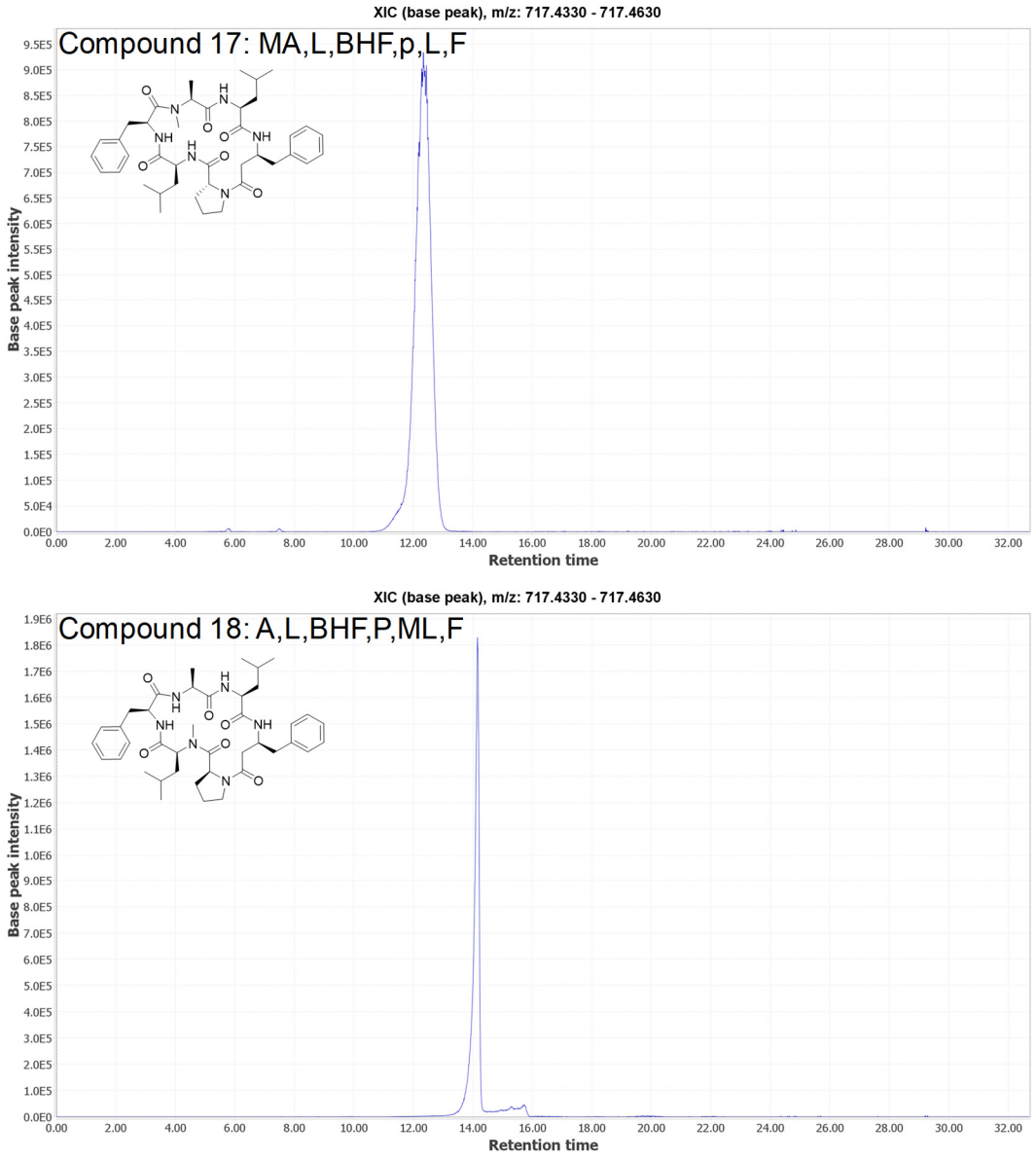


Figure S1.3.9 Crude single ion chromatograms of resynthesized compounds 1.17 and 1.18.

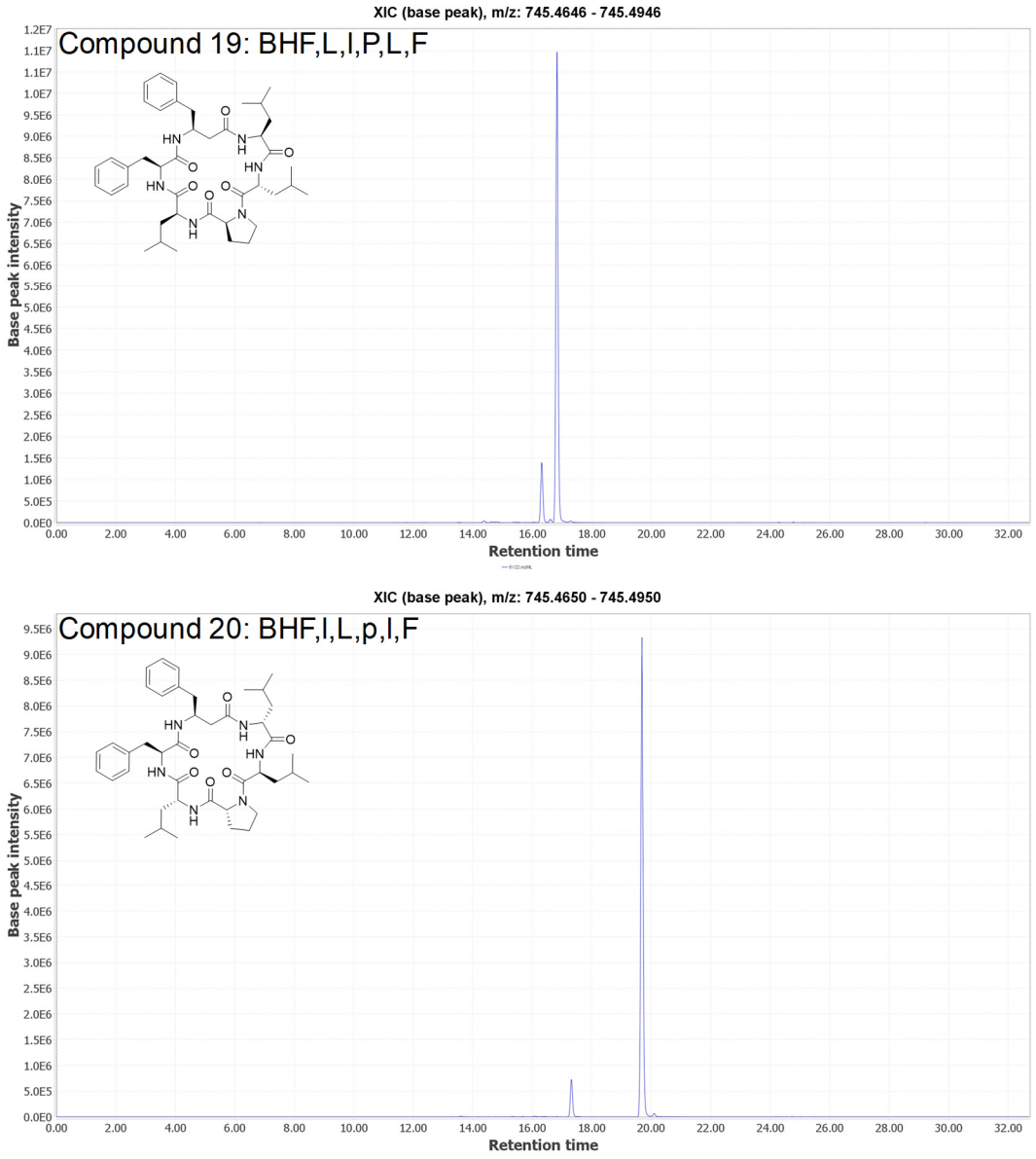


Figure S1.3.10 Crude single ion chromatograms of resynthesized compounds 1.19 and 1.20. Both minor peaks observed were minor epimers by MS^2.

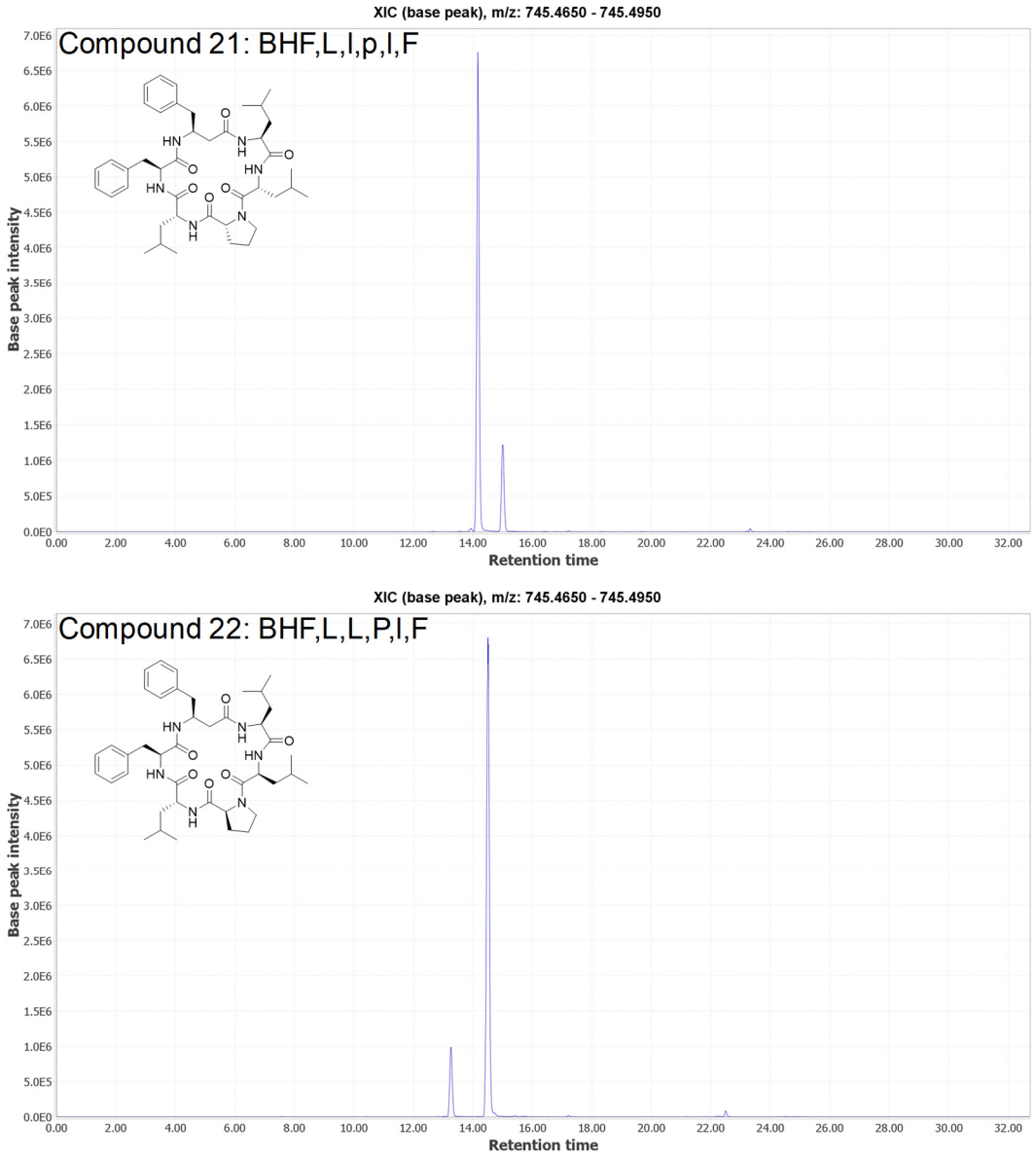


Figure S1.3.11 Crude single ion chromatograms of resynthesized compounds 1.21 and 1.22. Both minor peaks observed were minor epimers by MS^2.

1.7.7 Sequencing Validation of Resynthesized Compounds

To ascertain whether the sequence given by CycLS for a certain peak was correct, 22 of those predicted sequences were resynthesized (without isotopic labeling). The first test of whether the two represent the same compound was to compare single ion chromatograms of the resynthesized peptide and the sub-library from which it originated. In many cases this comparison alone was enough to determine the correct/incorrect status of the sequencing. In other cases, especially in the presence of closely eluting peaks, the MS^2 spectra of the library peak and the resynthesized peptide were visually compared. MS2 spectra of sodiated ions or contaminants were easily distinguishable from the library peak in question, while correctly sequenced resynthesized compounds matched their library peak of origin completely aside from mass shifts due to the absence isotopic labelling in the resynthesized compounds. In rare cases, inspection of an ion list for both spectra was necessary to determine whether the two MS2 spectra matched. Figures S1.4.1 through S1.4.11 abbreviate the full compound numbers by leaving out the preceding chapter designation, with the full compound number present in each figure's caption.

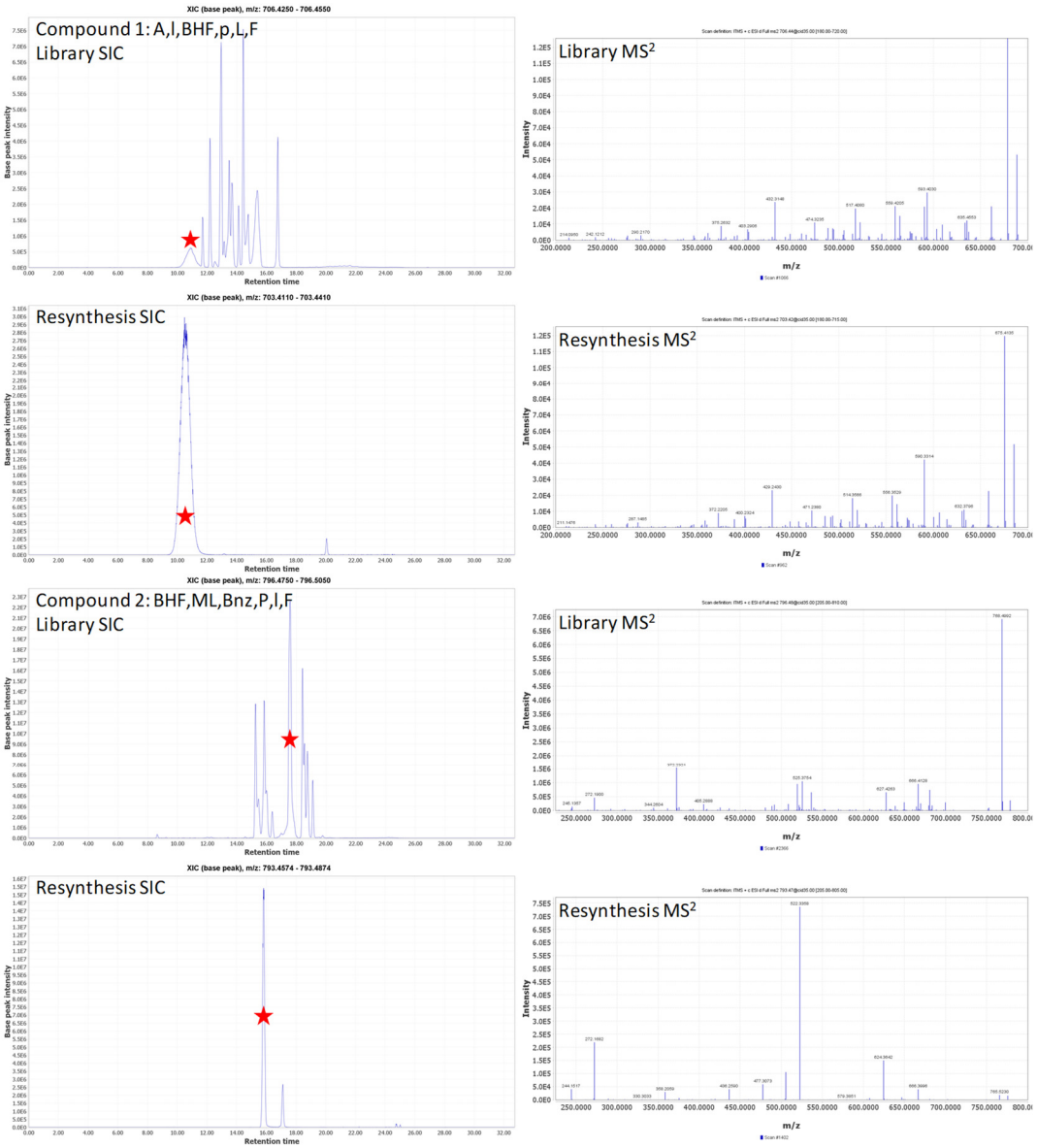


Figure S1.4.1 MS^2 analysis of compounds 1.1 and 1.2. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compound 1.1 was sequenced correctly, while compound 1.2 was sequenced incorrectly, matching to a different library member instead.

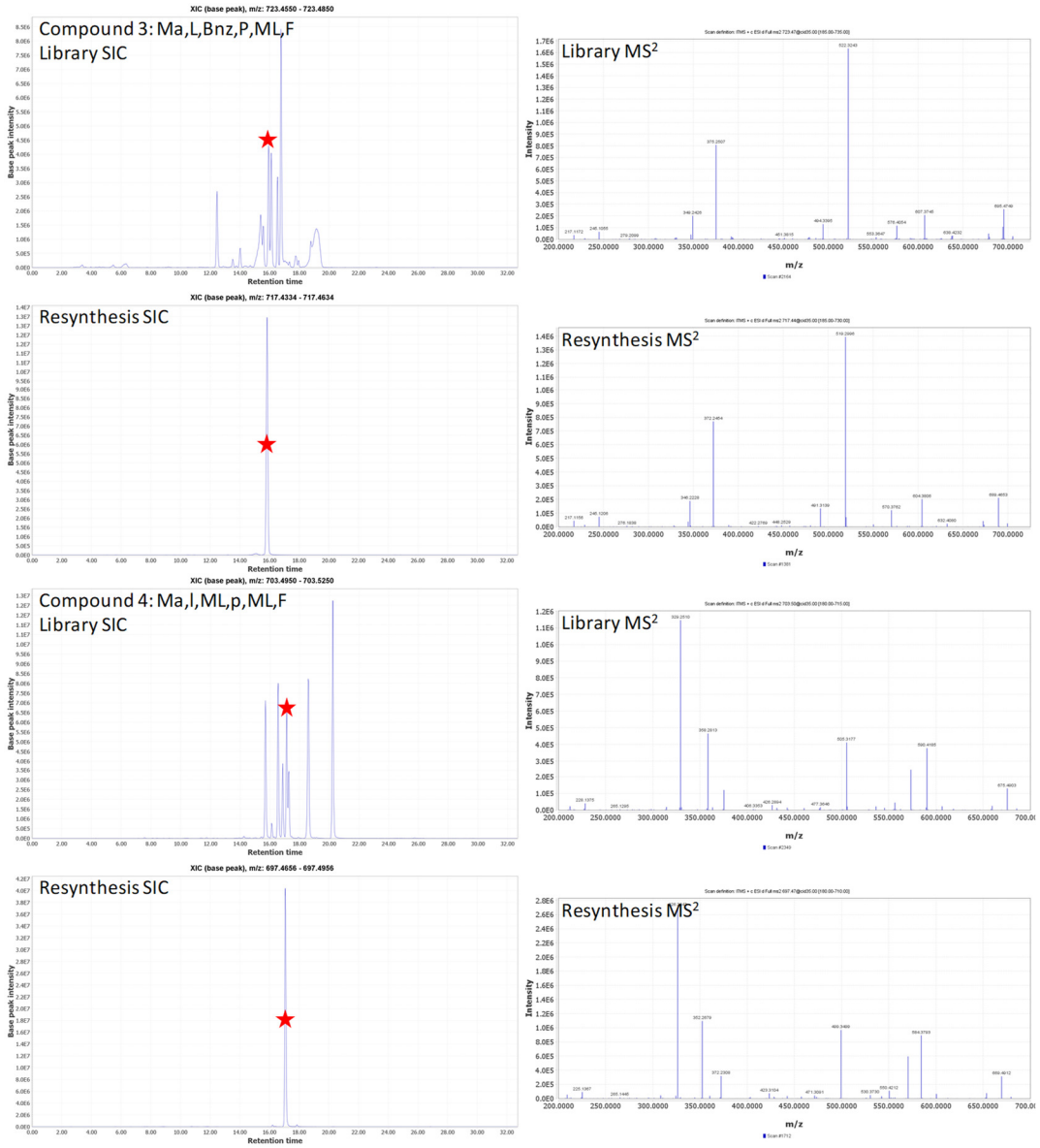


Figure S1.4.2 MS^2 analysis of compounds 1.3 and 1.4. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compounds 1.3 and 1.4 were sequenced correctly.

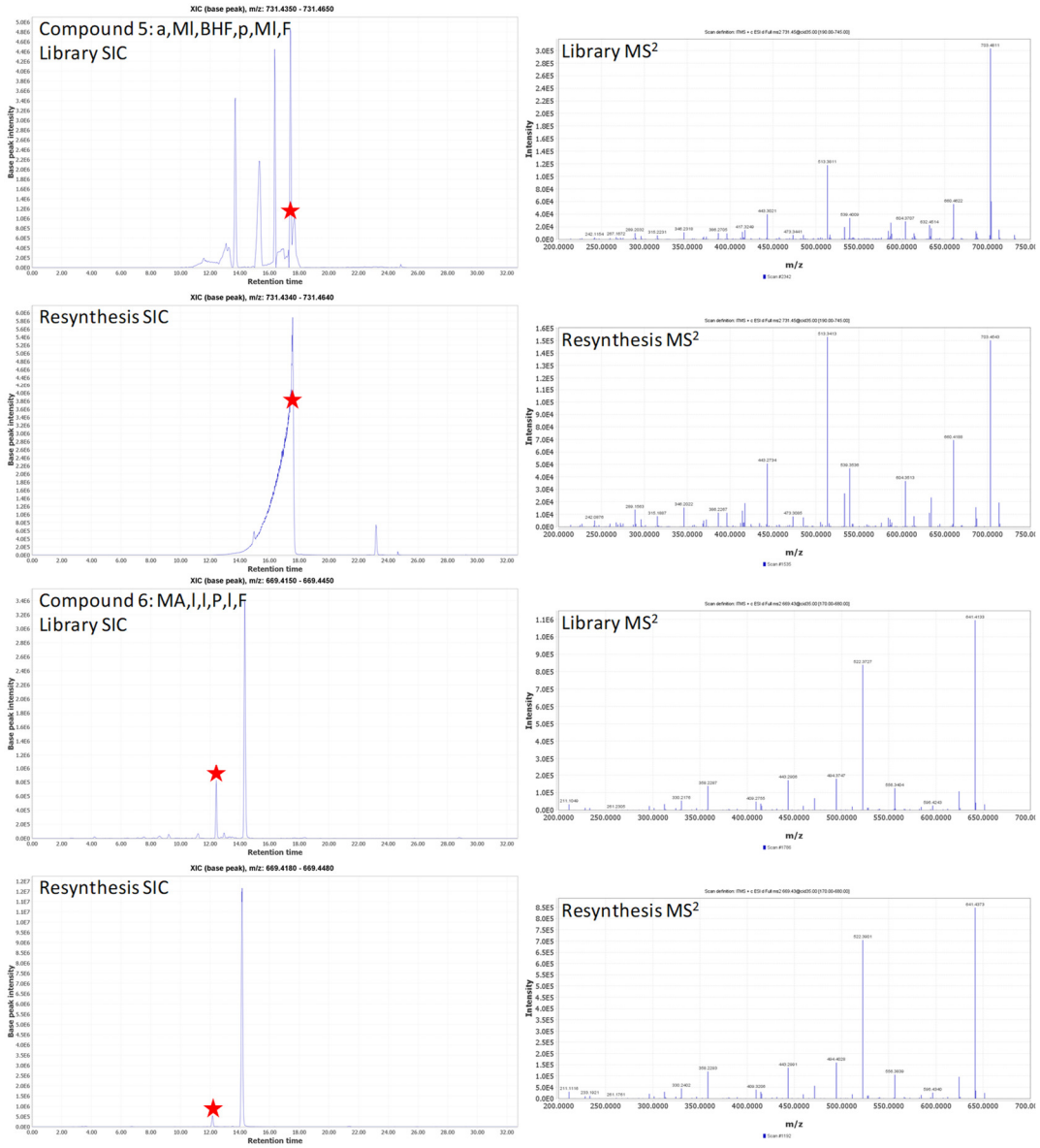


Figure S1.4.3 MS^2 analysis of compounds 1.5 and 1.6. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. The library peak corresponding to compound 1.5 was an interfering sodium ion, but CycLS successfully identified the left-tailed peak to its right from the MS^2 data at that time point despite the interfering sodium ion (and user error) and we therefore called this a correct sequencing. Compound 1.6 epimerized both in the library and its resynthesis and only its minor epimer was sequenced in the library, so we compared the MS^2 data of the minor epimers here to confirm that the sequencing was correct by proxy to the major peak.

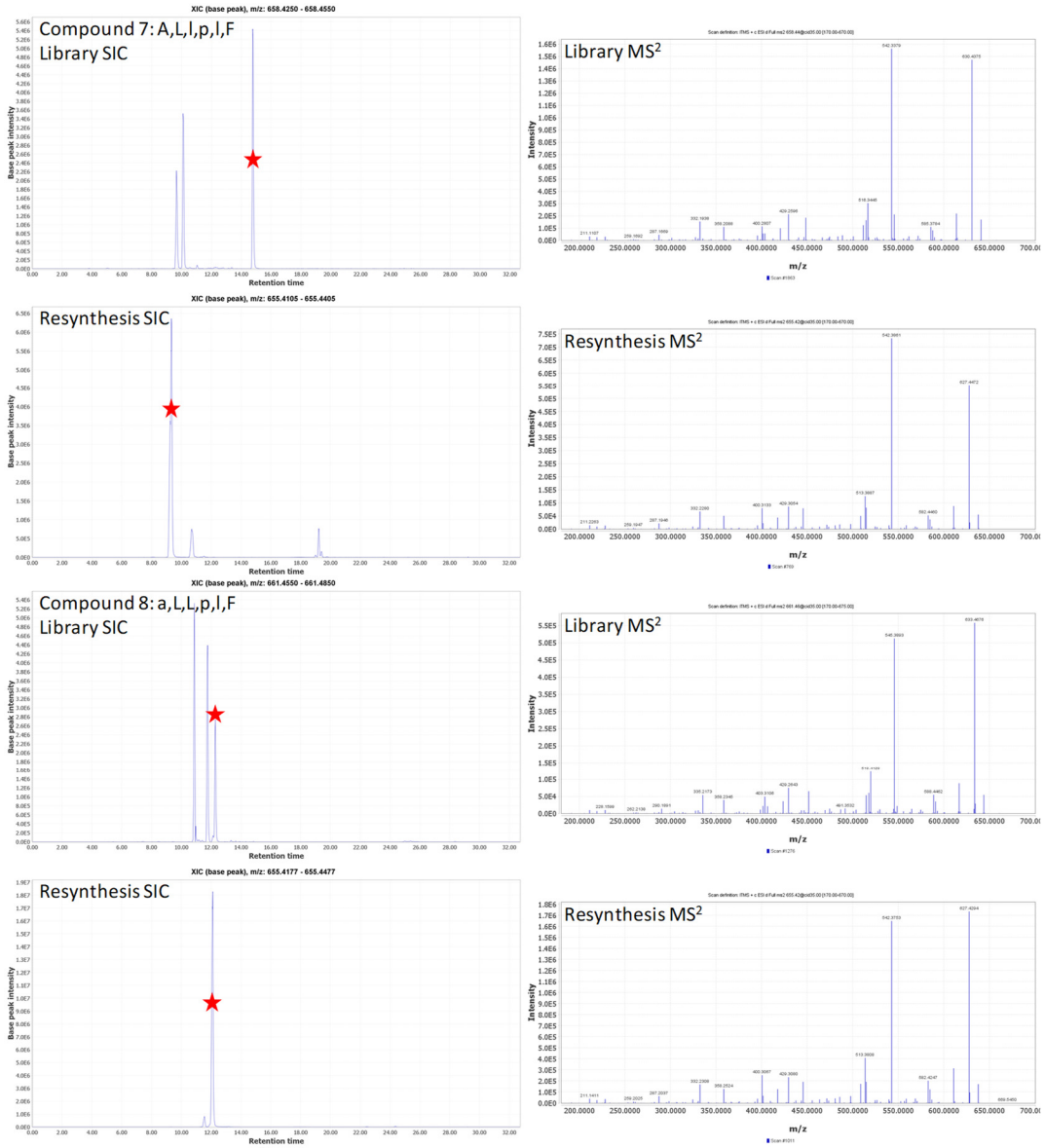


Figure S1.4.4 MS^2 analysis of compounds 1.7 and 1.8. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compounds 1.7 was sequenced incorrectly, matching to a different library member instead, and compound 1.8 was sequenced correctly.

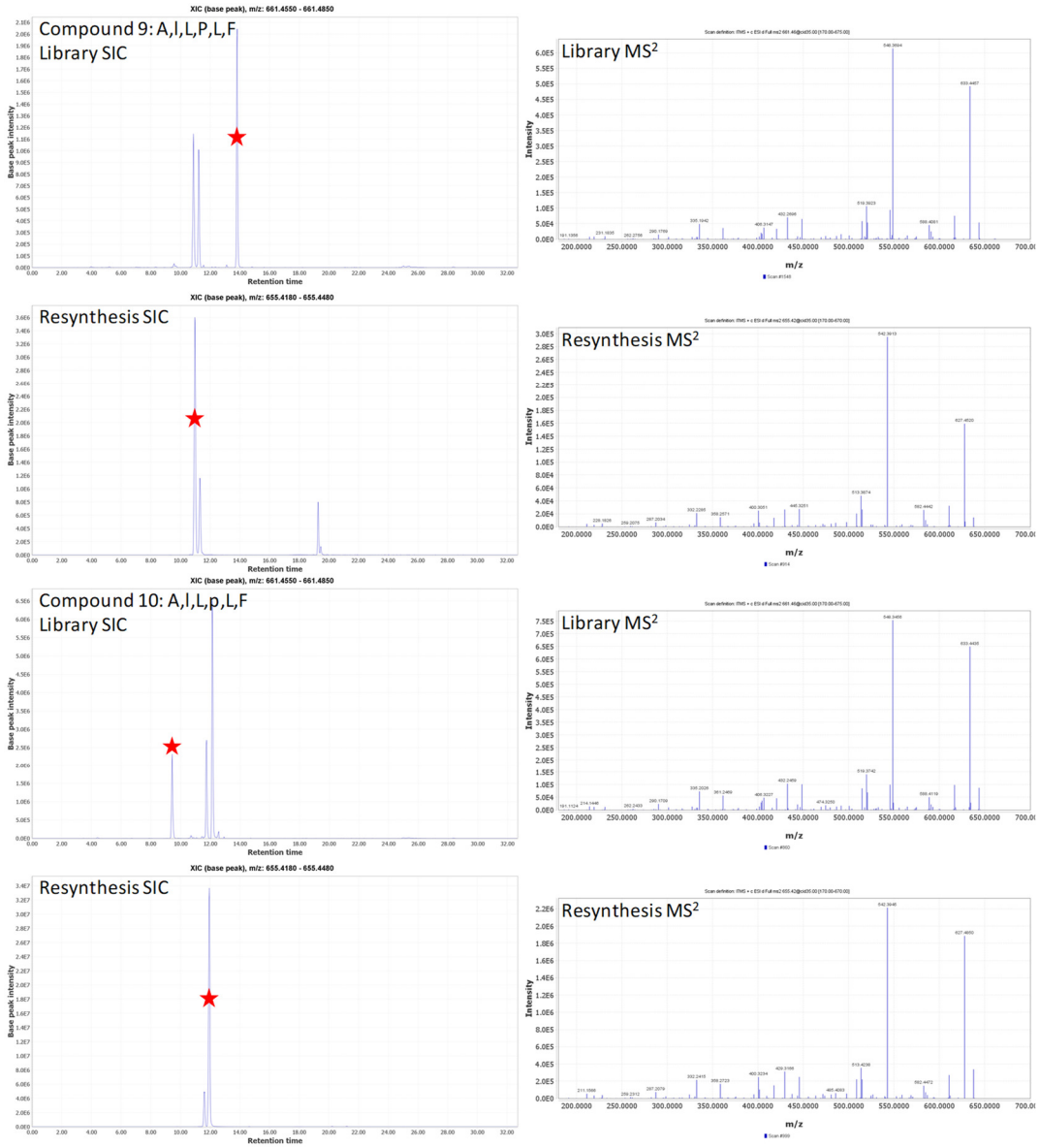


Figure S1.4.5 MS^2 analysis of compounds 1.9 and 1.10. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compounds 1.9 and 1.10 were sequenced incorrectly, matching to different library peaks instead.

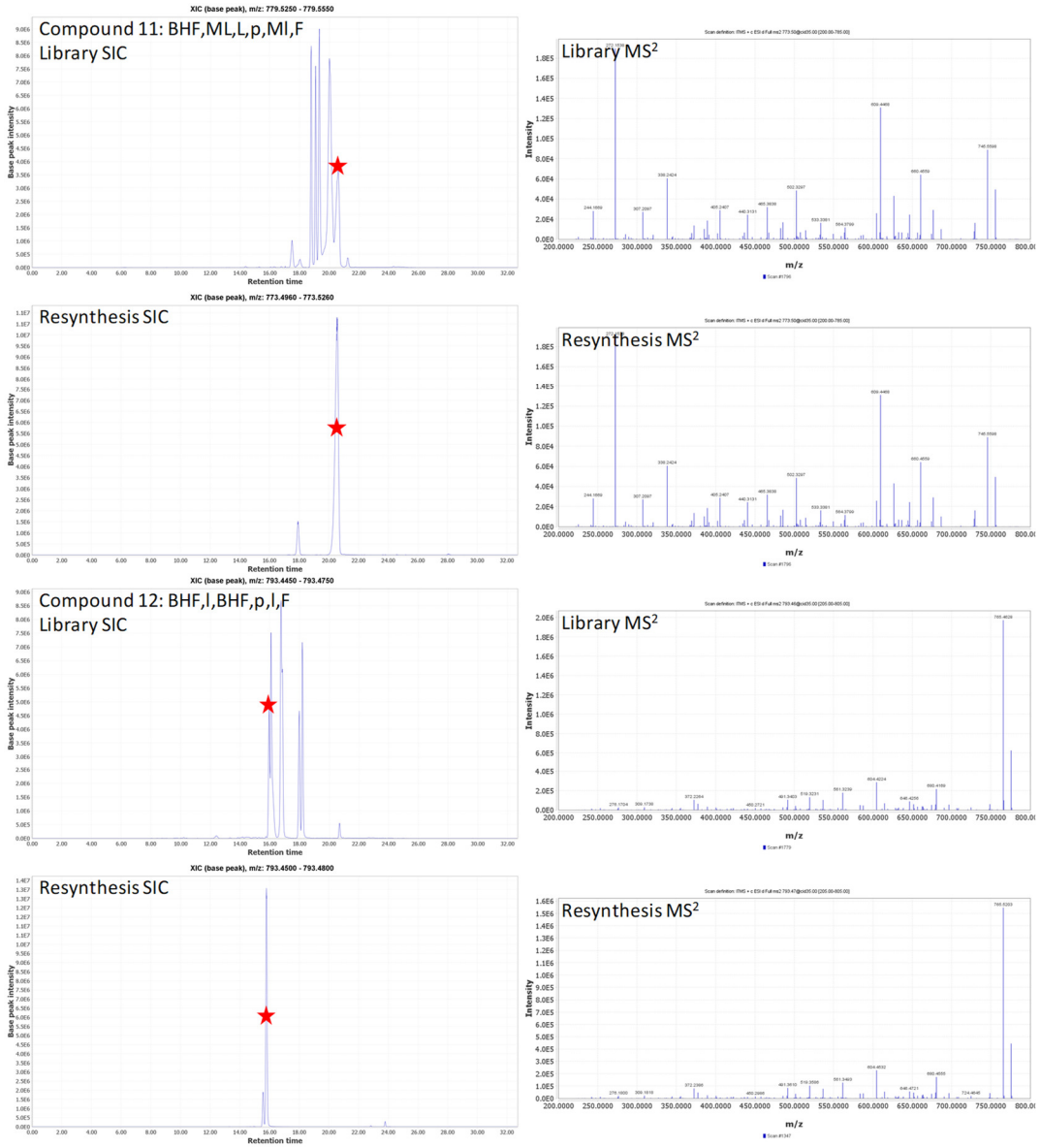


Figure S1.4.6 MS^2 analysis of compounds 1.11 and 1.12. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compounds 1.11 and 1.12 were sequenced correctly. Note that for compound 1.11, the library peak was almost exactly overlapped with a sodium ion peak of an identical mass but was correctly identified despite this. The MS2 shown for the library is the spectrum with the least influence from the sodium ion overlap.

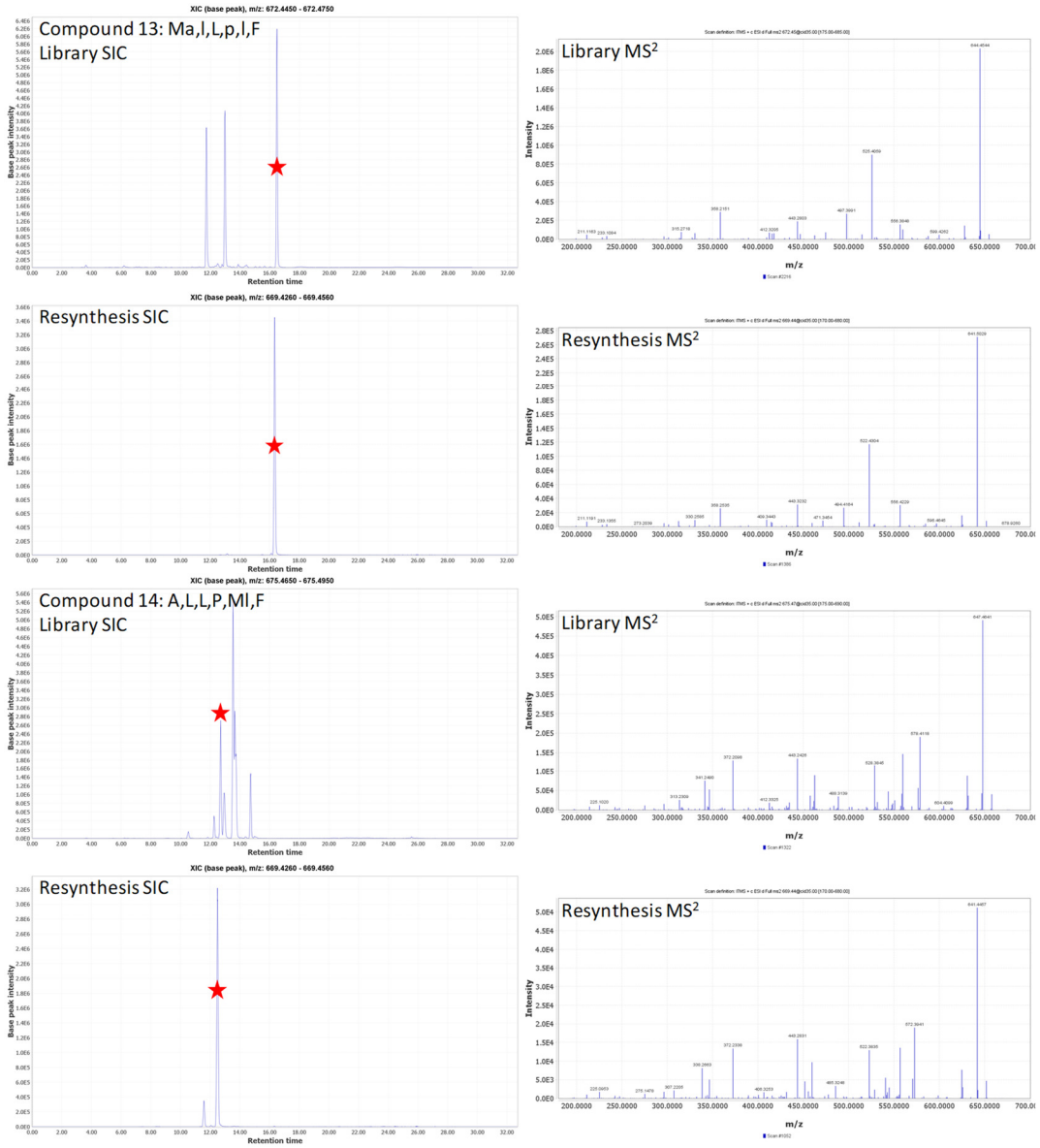


Figure S1.4.7 MS^2 analysis of compounds 1.13 and 1.14. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compound 1.13 and 1.14 were sequenced correctly.

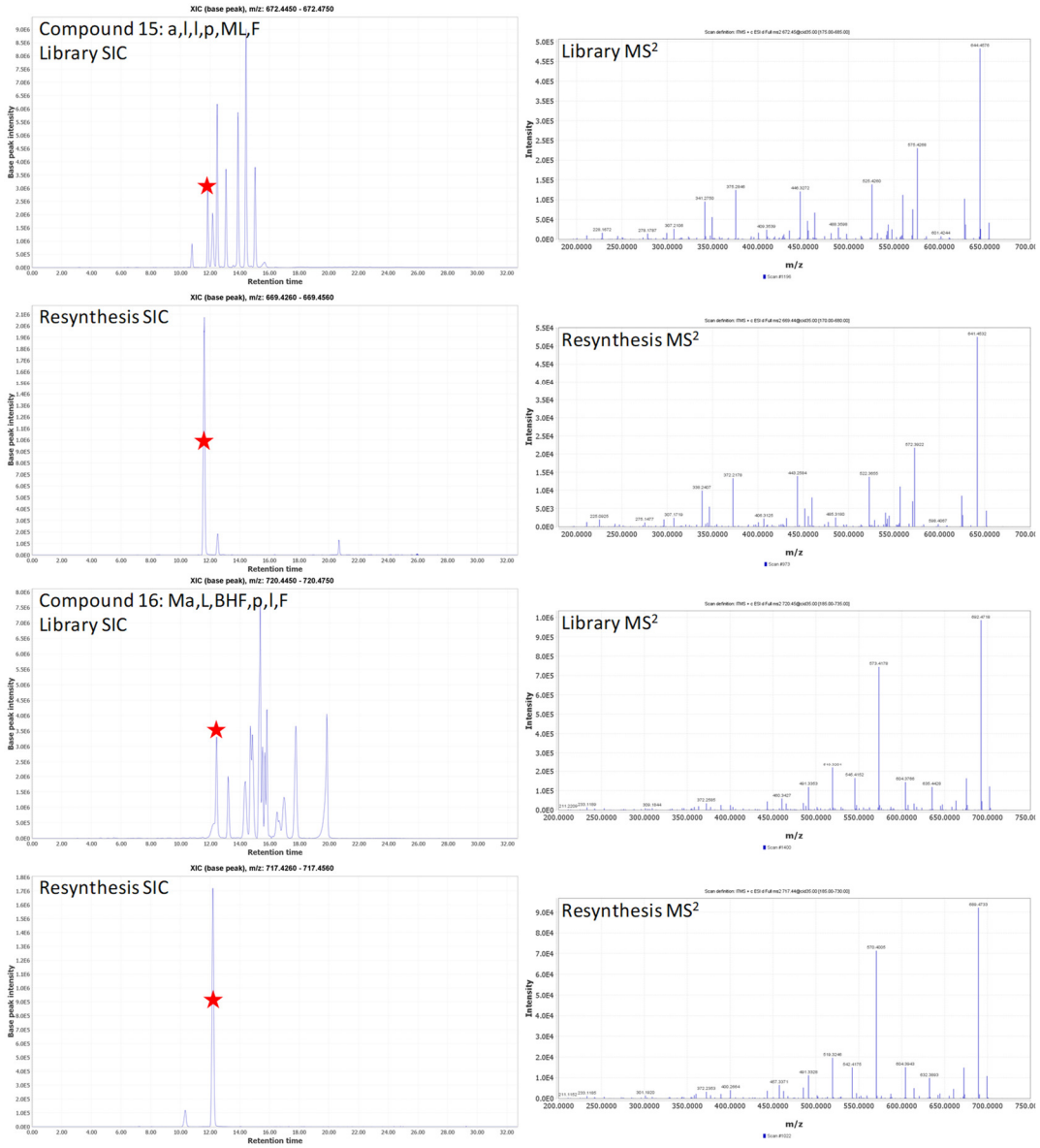


Figure S1.4.8 MS^2 analysis of compounds 1.15 and 1.16. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compounds 1.15 and 1.16 were sequenced correctly. In compound 1.16's library chromatogram, the broad peak to the left of the sequenced library peak and overlapping it is another library-member with a different sequence from the resynthesized compound.

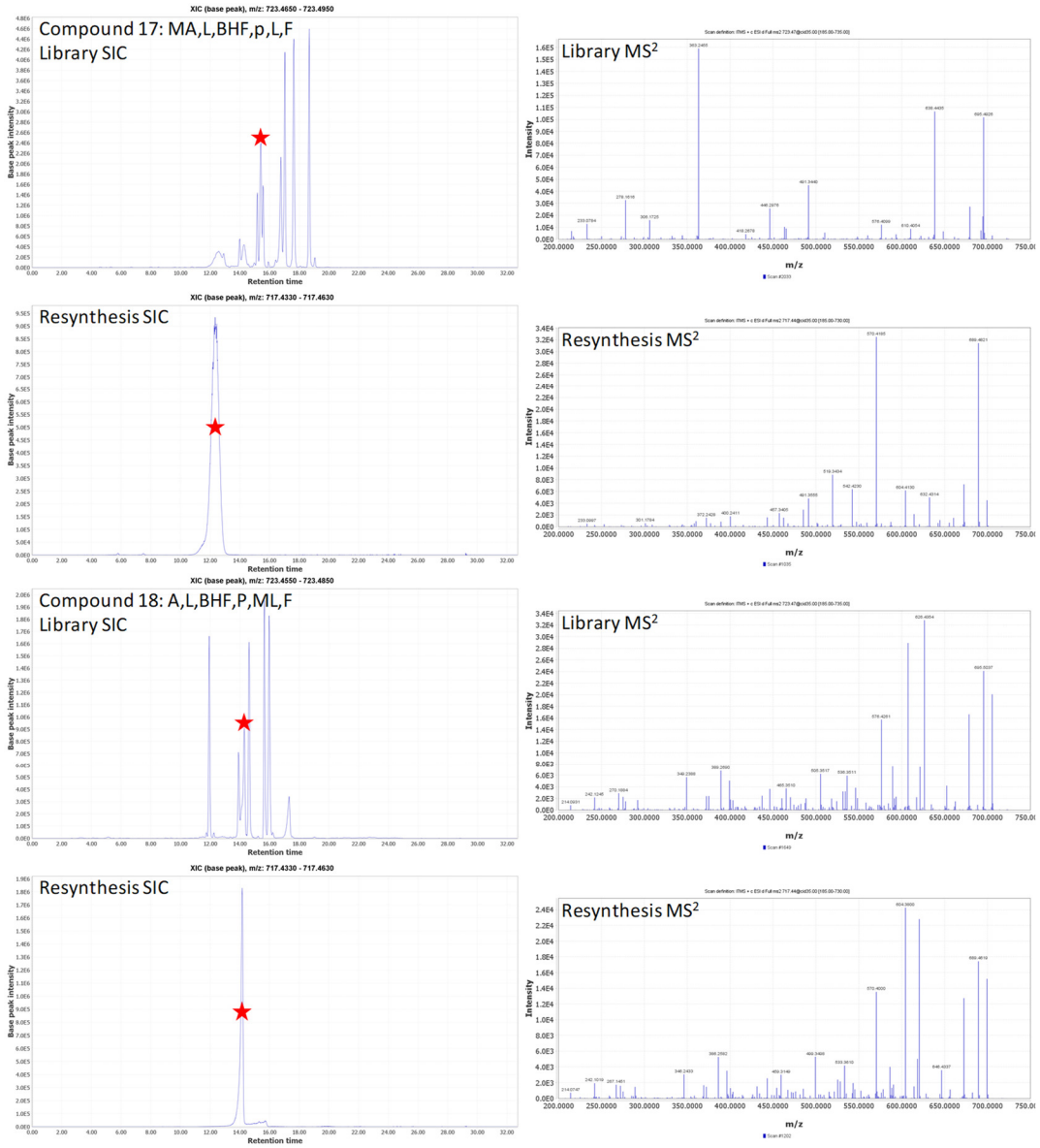


Figure S1.4.9 MS^2 analysis of compounds 1.17 and 1.18. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compound 1.17 was sequenced incorrectly, matching to a different library member instead, and compound 1.18 correctly.

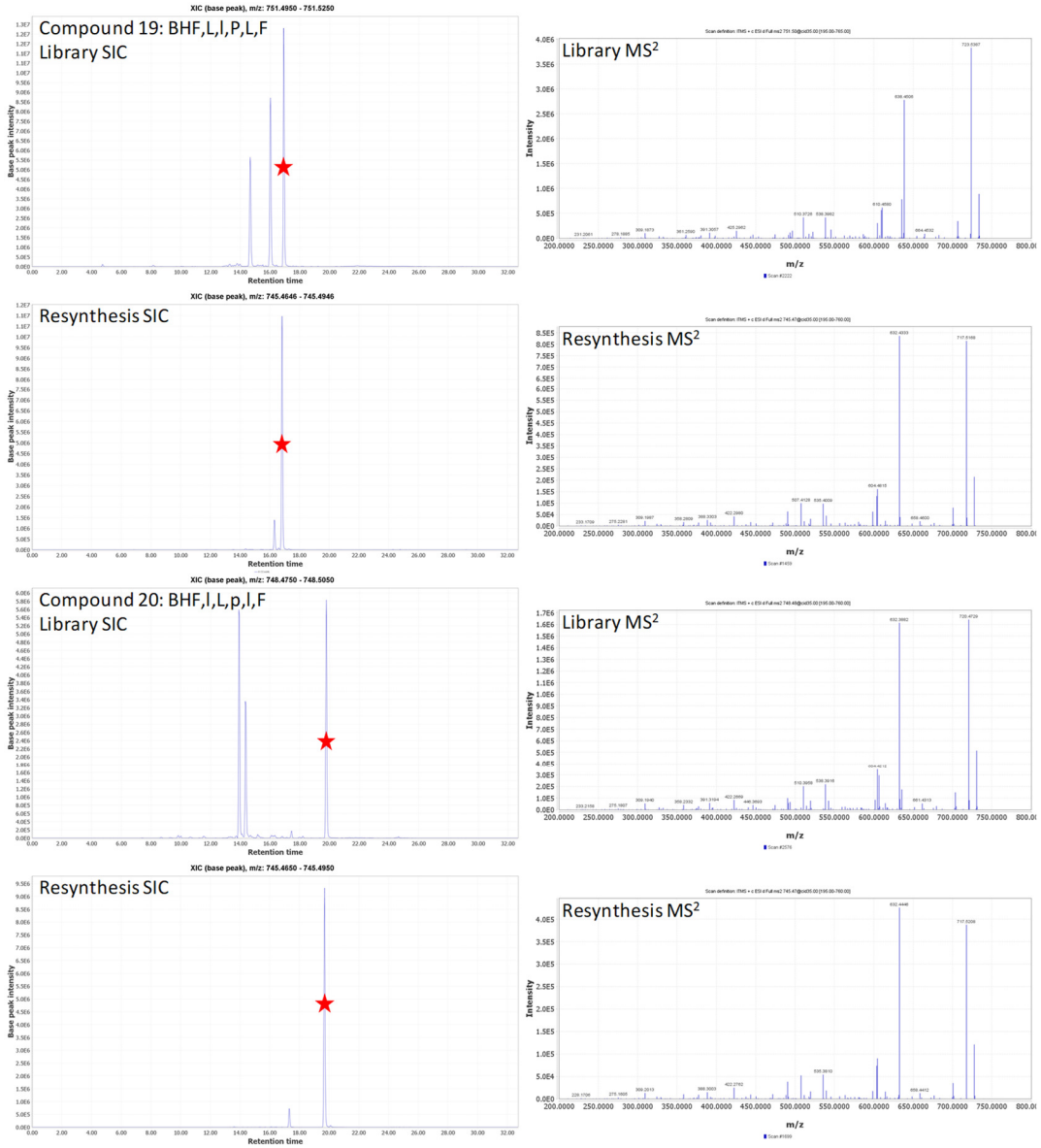


Figure S1.4.10 MS² analysis of compounds 1.19 and 1.20. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS² spectra of the two selected peaks are compared. Compounds 19 and 20 were sequenced correctly.

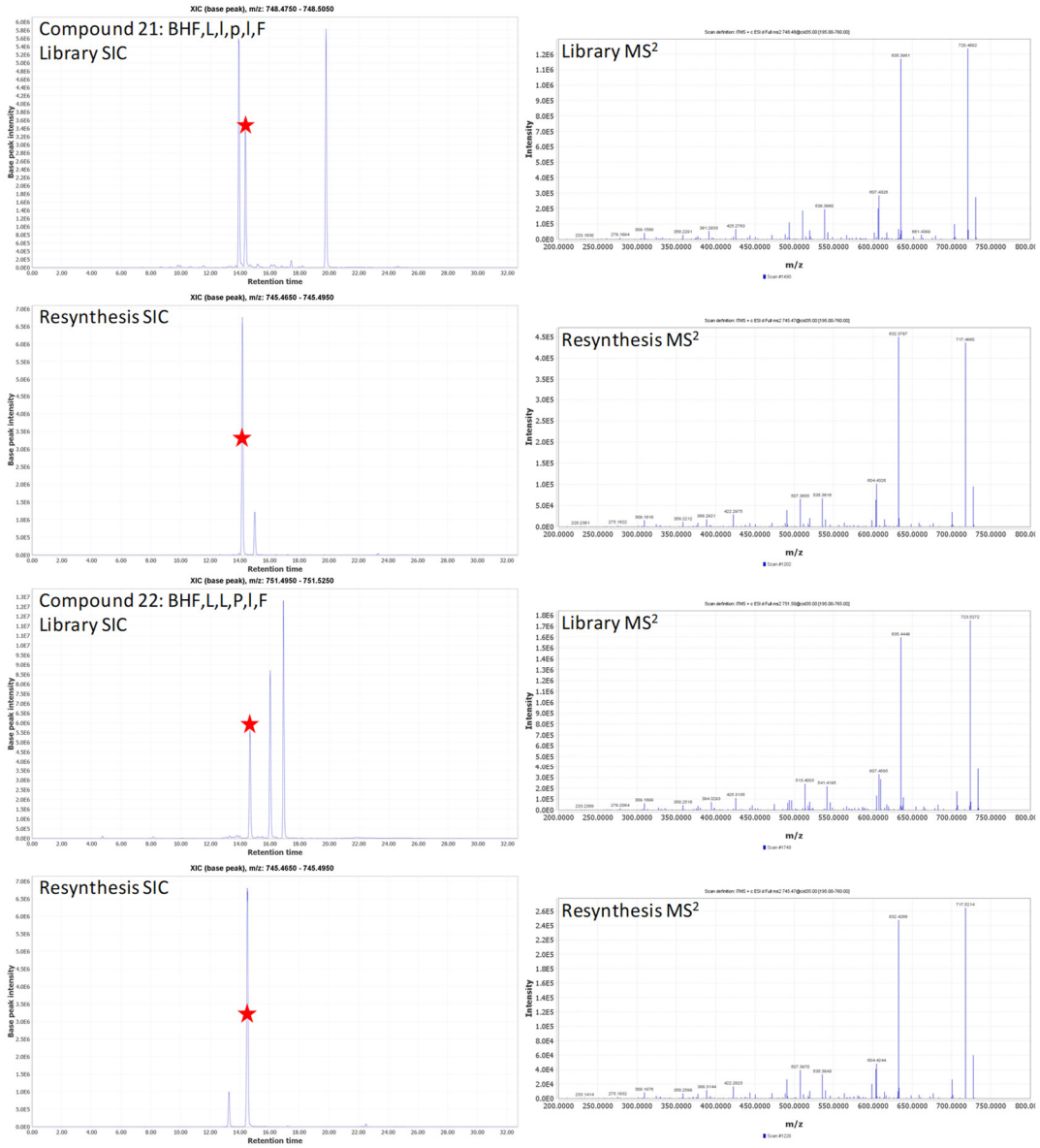


Figure S1.4.11 MS^2 analysis of compounds 1.21 and 1.22. Left: Single ion chromatograms to compare retention time are on the left, with the red stars marking the peaks compared (upper: library peak sequenced, lower: resynthesized compound). Right: MS^2 spectra of the two selected peaks are compared. Compound 1.21 and 1.22 were sequenced correctly.

1.8 Acknowledgements

We would like to thank Matthew Naylor and Justin Faris for proofreading the manuscript that became this chapter.

1.9 Author Contributions

The manuscript that became this chapter was written by Chad Townsend and R. Scott Lokey. Synthesis and data acquisition for the first validation library was performed by Akihiro Furukawa. Synthesis of the second validation library was performed by Chad Townsend and Quinn Edmondson. Chad Townsend wrote CycLS with the help of Joshua Schwochert. Cameron Pye and Chad Townsend optimized the chromatography and mass spectrometry data acquisition parameters. All authors gave approval to the final version of the manuscript that became this chapter.

1.10 Funding Sources

We wish to acknowledge Roche-Nimblegen for their financial support of this project.

1.11 Associated Content

The associated excel spreadsheet Validation_1_unique_mass.xlsx contains the sequencing results of the unique mass validation library. CycLS capabilities, installation and usage instructions, and GitHub page address can be found in Appendix A. Complete code can be found in the associated file “CycLS.py”.

Chapter Two

The passive permeability landscape around geometrically diverse hexa- and heptapeptide macrocycles.

This chapter contains text and figures from the following manuscript: Townsend, C. E.; Naylor, M.R.; Jason, E.; Pye, C.R.; Furukawa, A.; Schwochert, J. A.; Edmondson, Q.; Lokey, R. S., The passive permeability landscape around geometrically diverse hexa- and heptapeptide macrocycles. *(manuscript in preparation)*

Abstract

Recent advances in DNA and mRNA encoding technologies have enabled the discovery of high-affinity macrocyclic peptides and peptide-like ligands against virtually any protein target of interest. Unfortunately, even the most potent biochemical leads from these screening technologies often have weak cellular activity due to poor absorption. Biasing such libraries towards passive cell permeability in the design phase would facilitate development of leads against intracellular targets. We set out to empirically evaluate the intrinsic permeability of thousands of geometrically diverse hexa- and heptapeptide scaffolds by permuting backbone stereochemistry and N-methylation, and by including peptoid and β-amino acid residues at select positions, with the goals of providing a resource for biasing library-based screening efforts toward passive membrane permeability and studying the effects of the backbone elements introduced on a large number of compounds. Libraries were synthesized via standard split-pool solid phase peptide synthesis, and passive permeability was measured in pools of 150 compounds using a highly multiplexed version of the parallel artificial membrane permeability assay (PAMPA) under sink conditions. Compounds were identified using CycLS, a high-resolution mass spectrometry-based method that uses stable isotopes to encode stereochemistry and matches MSMS data to virtual fragment libraries based on the expected macrocyclic products. From the compounds that were identified with high confidence, 823 hexameric and 1330 heptameric scaffolds had PAMPA permeability coefficients greater than 1x10-6 cm/s. The prevalence of high permeability compounds in these two libraries suggests that passive permeability is achievable for hexa- and heptapeptides with highly diverse backbone geometries.

2.1 Introduction

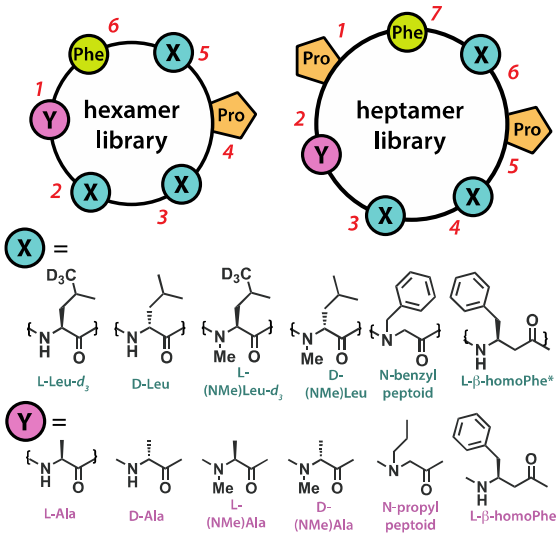


Figure 2.1. Hexamer and heptamer library design schematics. In each schematic, residues are labeled by residue number from the N-terminus to the C-terminus of the linear peptide pre-cyclization. The heptamer library was generated by addition of a second Proline position to the hexamer library. At X-labeled positions, the resin was split for separate additions of each residue and then pooled. The resin was kept separate after additions of L/D-Pro and after addition of the Y-labeled residue. The L-β-hPhe residue was not allowed at all X-labeled positions; *only position 3 on the hexamer schematic and position 4 on the heptamer schematic received L-β-hPhe additions.

Cyclic peptides have demonstrated the ability to inhibit a wide range of protein-protein interactions[37] and reach intracellular targets[38], with new cyclic peptide therapeutics reaching the market each year[39]. Their ease of synthesis and diversification have made cyclic peptides an attractive platform for high throughput screening against protein-protein interactions and other targets that are difficult to inhibit with small molecule therapeutics. However, realizing that potential by achieving membrane permeability, which is a prerequisite for oral bioavailability and the ability to engage intracellular targets, remains a major barrier to the development of therapeutic cyclic peptides[40, 41].

Powerful library construction and screening technologies such as mRNA display have made discovering potent cyclic peptide binders easier than ever[42, 43]. But as powerful as DNA/mRNA encoding technology is, the ligands produced are selected only for potency, which is often driven by charged and/or highly polar residues that are generally incompatible with passive cell permeability. Optimizing a lead compound for passive permeability without negating target binding is often difficult even without such incompatibilities. Biasing such

libraries toward cell permeability in the design phase has the potential to significantly ease optimization of hits towards favorable ADME properties.

Conformation has been shown to be highly impactful on passive permeability in cyclic peptides[44-46], with modifications to backbone geometry producing dramatic effects. In one study a single stereochemical inversion resulted in epimers with permeation rates that varied by orders of magnitude[47]. Oppositely, sidechain modifications often affect passive permeation mainly through their lipophilic contribution[30, 48, 49]. We therefore hypothesized that a large set of cyclic peptide backbone geometries of known permeability would both aid our understanding of passive permeation and be of potential use in biasing library screens towards passively permeable hits through sidechain modification of permeable backbones.

We set out to evaluate the permeability landscape in geometrically diverse cyclic hexa- and heptapeptides using a one-bead one-compound (OBOC) based synthesis and highly parallel analytical approach that we have previously reported[47]. Whereas previously we relied on resynthesis to decode stereochemical and sequence ambiguities when analyzing whole libraries, here we use stable isotopes to encode stereochemistry and an MSMS-based approach that we developed previously called CycLS[50] to distinguish among isomers with different sequences. To cover the broadest possible conformational landscape in this size range, we varied stereochemistry and N-methylation, factors that have been shown to have a significant impact on passive permeability[44, 45, 51, 52]. We also permuted other backbone features by introducing β-amino acid residues and peptoids. In addition to increasing backbone flexibility, peptoids impart distinct conformational characteristics to a macrocycle, for example, by allowing both cis- and trans-rotamers at the N-substituted amide. They also provide potential sites for side chain diversification in future compounds or libraries based on these scaffolds, and we have previously shown that substitution of peptoids at specific positions in a cyclic peptide scaffold can preserve or even enhance membrane permeability[30, 53].

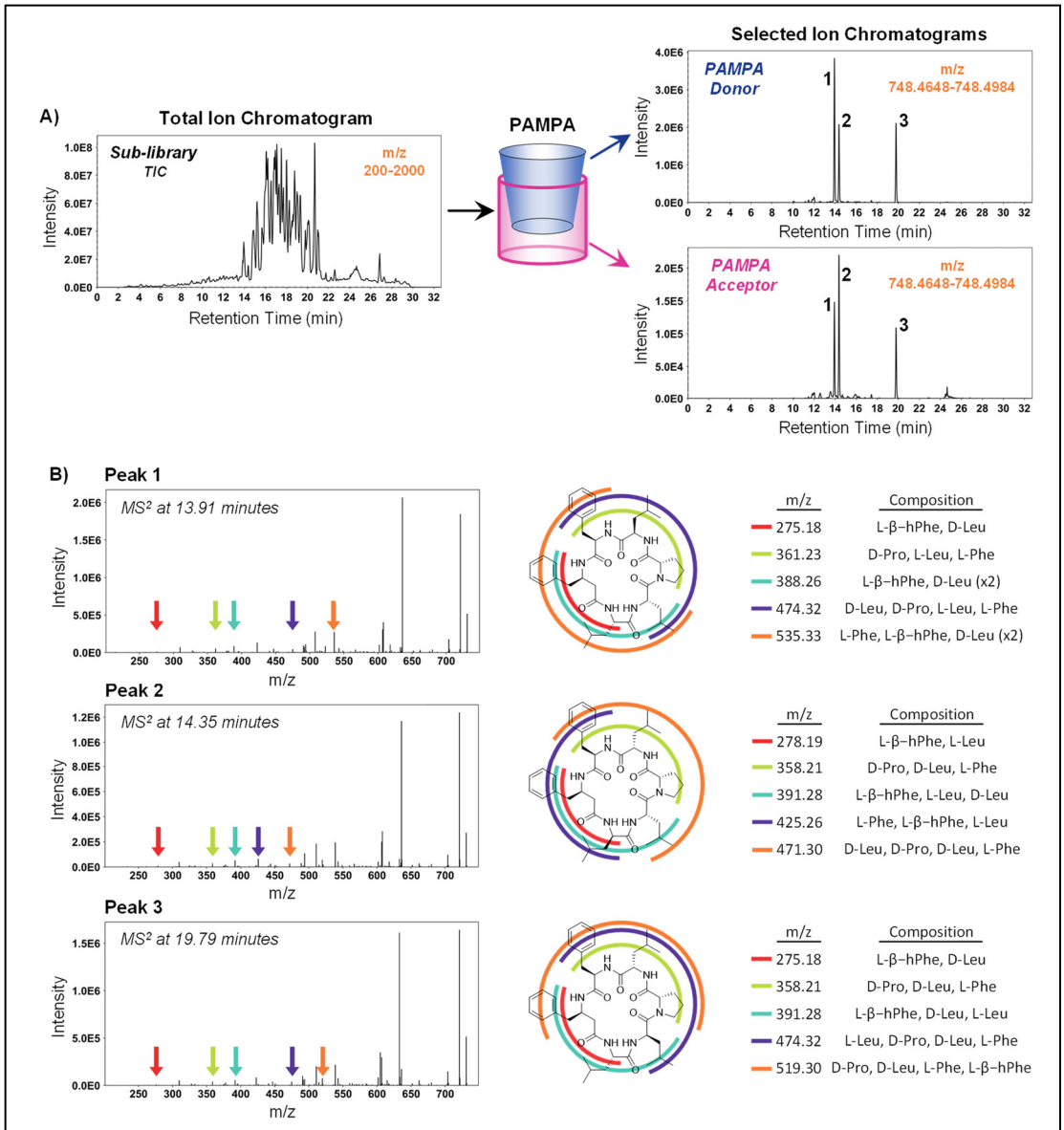


Figure 2.2 Analysis of mixtures workflow. A) A mixture of 150 cyclic hexa-peptides and peptide-peptoid hybrids had its passive permeation quantified by PAMPA. A total ion chromatogram of the mixture and extracted ion chromatograms of the donor and acceptor wells for a library mass with three expected isobaric compounds is shown. The MS^2 data of the three peaks were compared to virtual fragment libraries of each potential sequence to determine the position of an L-Leucine-d_3 residue. B) Each MS^2 spectrum corresponding to a chromatographic peak in the extracted ion chromatograms is accompanied by the chemical structure of the library member that had the strongest match. Select fragments containing sequencing information key to distinguishing among the three candidates have been marked by colored arrows on the MS^2 spectrum and arcs denoting their position on the corresponding chemical structure. The table entry corresponding to each color notes the mass/charge ratio of each fragment and its composition. Individual cyclic peptide fragments do not generally provide sequence information directly, so information from multiple fragments must be combined to determine a sequence.

Although hexapeptide and peptide-peptoid hybrids have been extensively studied using similar approaches[44, 45, 54-56], few studies have been done on the permeability landscape in heptapeptides. Here we describe an extensive and unbiased survey of the permeability landscape in cyclic hexa- and hepta- peptides and peptide-peptoid hybrids. The results not only reveal the expected dependence of passive permeability on backbone geometry, but also show a remarkably high degree of passive permeability among all library members. Additionally, we observed strong correlations between measured mixture and pure compound PAMPA permeabilities, enabling future usage of specific permeability-biased backbone geometries discovered herein to bias encoded libraries toward passive cell permeability.

2.2 Results

Using an approach that we reported previously[50], we synthesized a library of 1800 cyclic hexapeptide and 3600 cyclic heptapeptide scaffolds using split-pool solid phase peptide synthesis (SPPS), starting with L-Phe attached to the solid support through a 2-chlorotrityl linker (in describing library synthesis residues will be addressed by their number in the hexamer library design schematic in Figure 2.1 because the heptamer library was generated from a portion of the linear hexamer library). The resin was split into five pools for attachment of residue 5, which included L-Leu-d_3, D-Leu, L-(NMe)Leu-d_3, D-(NMe)Leu, and benzyl peptoid (peptoids were incorporated using the submonomer method[36]; all other residues were incorporated using standard FMOC SPPS conditions). After addition of residue 5, the resin was pooled and split once again for incorporation of either L-Pro or D-Pro. To encode the Pro stereochemistry at this position, the L-Pro and D-Pro pools were kept separate for the remainder of the synthesis. Each of the Pro pools was split into six sub-pools for the addition of residue 3, which included the same building blocks used at residue 5 with the addition of L-β-homoPhe (L-β-hPhe). Splitting and pooling was continued for incorporation of position 2, which was identical to position 5, and position 1, which included L-Ala, D-Ala, L-(NMe)Ala, D-(NMe)Ala, L-β-homo-Phe, and propyl peptoid, resulting in twelve mixtures

defined by the known identities of residues 4 and 1. Each of the 12 mixtures were split into two lots, with one lot cyclized (to generate the 1800 cyclic hexapeptides) and the other lot set aside for the addition of another L- or D-Pro to generate the 3600 cyclic heptapeptides. Cyclization was performed in dilute 1:1 ACN:THF with COMU, and each pool was purified by solid phase extraction on C18 media.

The resulting 12 hexamer and 24 heptamer sub-libraries were inspected by LCMS to assess synthetic success. Each sub-library was confirmed by LCMS to have approximately the expected number of major peaks at each relevant accurate mass through inspection of individual extracted ion chromatograms. Truncations were detected post-purification but most eluted before the start of the LCMS gradient and none shared a mass with an expected library compound. An analysis of all major peaks at each expected library mass performed on a single exemplar heptamer sub-library confirmed 146 of the 150 expected peaks. An analysis of an exemplar hexamer sub-library from our previous work on CycLS confirmed 139 of 150 expected peaks (section 2.7.6).

Each sub-library of 150 compounds was tested for passive permeability using a version of the Parallel Artificial Membrane Permeability Assay (PAMPA) as previously reported[48] with the addition of 0.2% polysorbate 80 in the donor well and 0.2% D-α-Tocopherol polyethylene glycol 1000 succinate (TPGS) to the acceptor well as “double-sink” conditions to mitigate against compound aggregation and adsorption to the apparatus[57]. Tandem MS data were acquired for each sub-library, and the resulting spectra were analyzed using CycLS, with the peak identification process proceeding as in Figure 2.2. PAMPA and sequencing data were acquired with an identical LC method and were aligned by retention time to associate permeabilities and sequence identities. Various filters were applied to the merged data to ensure sequencing and PAMPA data quality, then peaks identified as duplicate sequences were resolved. Finally, peaks with outlier permeabilities for their sub-library and retention time were manually curated by visual inspection of the appropriate extracted ion chromatogram to ensure correct automated integration and peak alignment

between the donor and acceptor wells. Impermeable compounds, heptamers with low acceptor well intensity, and a limited set of random peaks were inspected similarly (section 2.7.5).

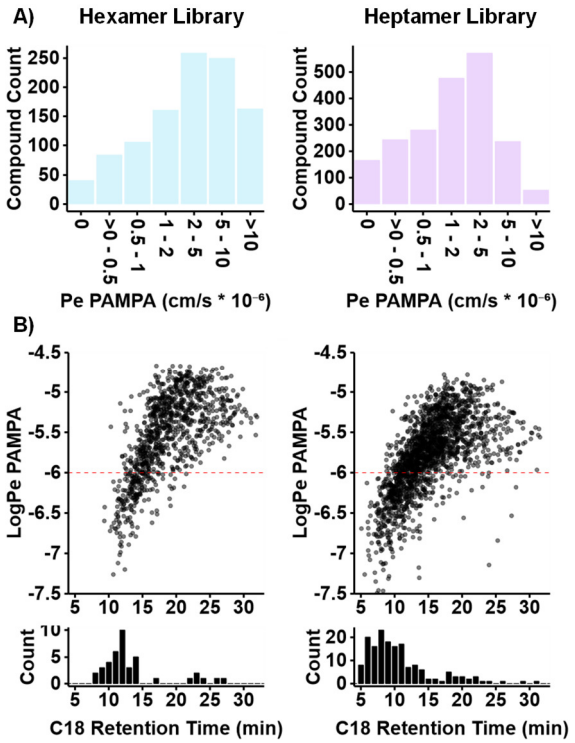


Figure 2.3. Library PAMPA permeability summary. A) Categorical histograms summarizing the PAMPA per-meabilities within each library. Bins are left-inclusive un-less noted. B) Scatterplots describing the relationship be-tween retention time on a C18 column and log PAMPA permeability for each library with a dashed line at 1 * 10-6 cm/s. Impermeable compounds could not be plotted on the log axis and were included as accompanying histograms with bin widths of one minute. The entirety of the plotted regions are within the gradient sections of the LC methods used.

Unique sequences were obtained for 1063 of 1800 hexamers (59%) and 2023 of 3600 heptamers (56%), of which 823 hexamers and 1330 heptamers permeation rates greater than 1.0 x 10^{-6} cm/s, while 153 hexamers and 47 heptamers had permeation rates of 10 x 10^{-6} cm/s or greater (Fig. 2.3A). PAMPA permeability coefficients and HPLC retention times on a reverse-phase (C18) column (Fig. 2.3B) were correlated for both libraries as has been observed previously[30, 45, 47, 48]. The majority of the 40 impermeable hexamers and 165 impermeable heptamers eluted early, in agreement with that correlation. Measured permeabilities ranged between 0.05 x 10^{-6} cm/s to 21.5 x 10^{-6} cm/s in the hexamer library and 0.01 x 10^{-6} cm/s to 16.4 x 10^{-6} cm/s in the heptamer library.

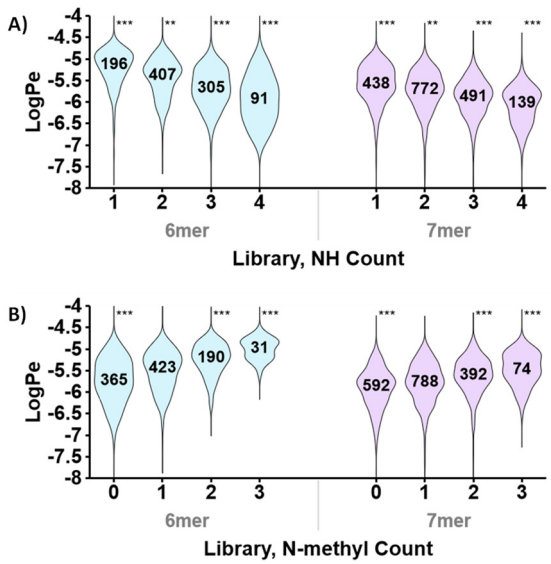


Figure 2.4 Permeability analysis by NH and N-methyl count. Violin plots of A) NH count and B) N-methylation count versus the log of PAMPA permeability for each library, with hexamers in blue and heptamers in purple. Each violin is labeled by the number of sequences used to generate that violin – positioned at each violin's median value – and by significance (* = P<0.05, ** = P<0.005, *** = P<0.0005). Significance was calculated using the Mann-Whitney U-test with probabilities being derived from a two-tailed normal distribution. The p-values gen-erated represent the probability of the data within a violin having originated from the combined distribution of the remainder of the library. Compounds with a measured PAMPA permeability of zero were not represented due to the log space Y-axis.

A simple compositional analysis was performed to demonstrate agreement with well-known trends in passive permeability. In general, removal of exposed hydrogen bond donors generally increases passive permeability in cyclic peptides[44, 45, 52, 58] and other macrocycles[49]. We binned each library by amide NH count (Fig. 2.4A) and by N-methylated residue count (Fig. 2.4B) to obtain two different views of each library differing only by their treatment of peptoid residues. As expected, median log permeabilities decreased as amide NH count increased and increased as amide N-methyl count increased for both libraries. From the similar effect magnitudes of the two counts (NH and N-methyl), we can determine that the inclusion of peptoid residues in these libraries was overall neutral to permeability. Additionally, the varying ability of individual compounds to form intramolecular hydrogen bonds can be observed through the range of permeabilities at each NH amide count, with a broad range of permeabilities observed for compounds with many NH amides and a more narrow range for those with few NH amides – though this effect is subtle in the heptamer library. This confirmation of previous work over thousands of structurally diverse cyclic peptides combined with reliable sequencing from CycLS supports the validity of the remaining analyses performed herein. However, we also synthesized these libraries with the

hope that the permeability estimates of individual compounds could be useful outside of the library setting.

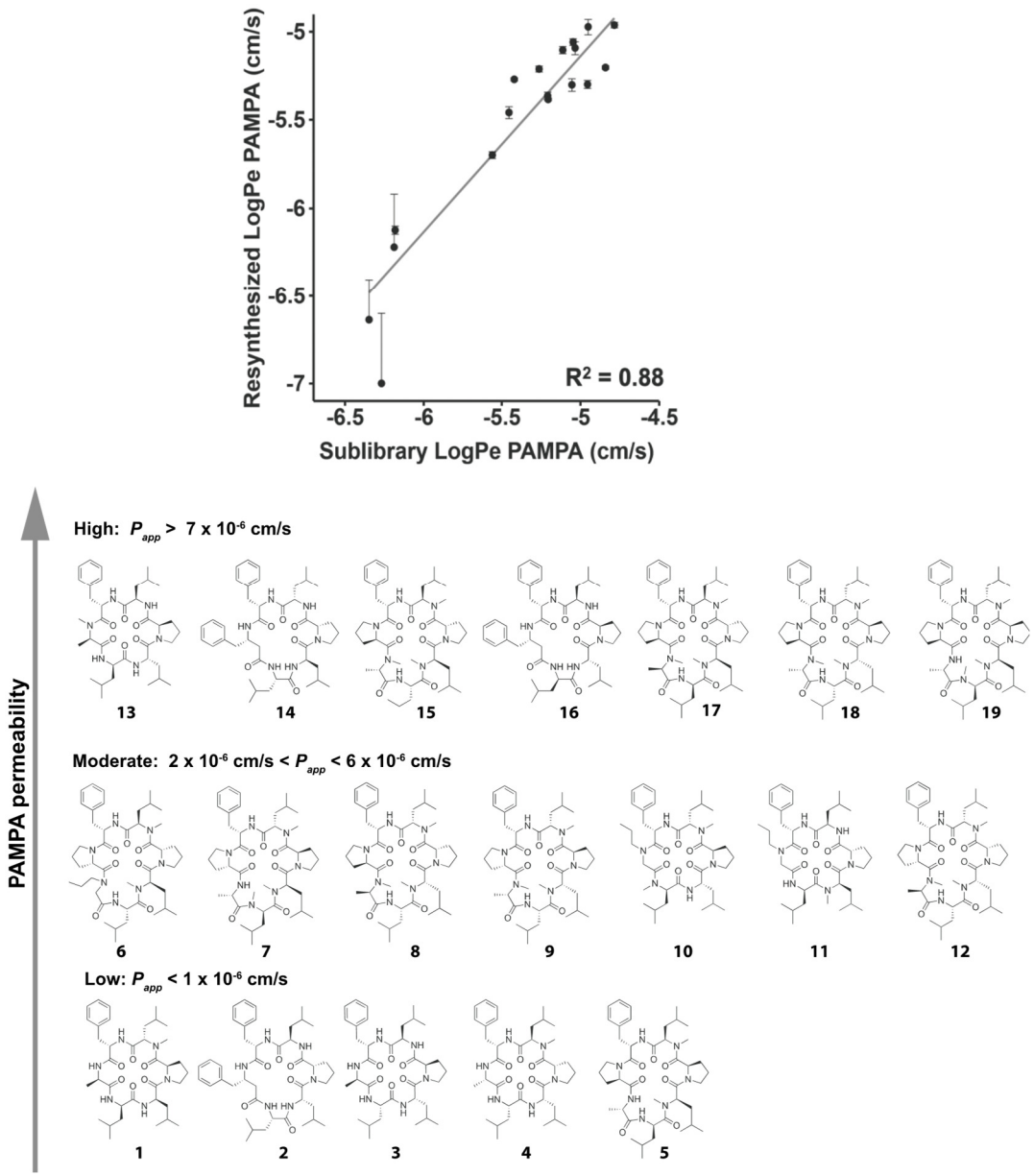


Figure 2.5 Comparison of resynthesized compound LogPe PAMPA with other permeability data. Above: Scatter plot comparing the PAMPA permeabilities for 18 resynthesized compounds with their corresponding mixture-based library permeabilities. Lower confidence intervals reaching zero could not be represented on the log scale. Fit line was arrived at by linear regression with an enforced slope of one. Below: Compound structures by their PAMPA permeability, with their numbering abbreviated from 2.# to # for simplicity. Compounds are referred to by their full names in all text and captions.

PAMPA on complex mixtures has been noted to affect the permeability of the mixture components[57, 59, 60]. To evaluate the severity of these differences and asses the utility of our data outside of bulk library analyses, we resynthesized and purified nine hexamers and ten heptamers and compared their pure PAMPA permeabilities with their mixture PAMPA permeabilities. Hexamer compounds 2.1, 2.2, 2.3, 2.4, 2.13, 2.14, and 2.16 (Fig. 2.5) were selected from the CycLS study for having sequenced correctly (compounds 1.15, 1.22, 1.8, 1.14, 1.13, 1.19, and 1.20 respectively). Compounds 2.6, 2.10, and 2.11 were selected randomly among all compounds with high sequencing confidence and non-zero permeability from sub-libraries with a propyl-peptoid residue. The remaining compounds were chosen arbitrarily among all heptamer sub-libraries for high sequencing confidence with a bias toward high permeability. Synthesis was performed similarly to the library synthesis but without isotopic labeling of L-Leu. Resynthesized compound identities were confirmed through manual inspection of tandem MS data for a similar fragmentation profile and matching retention times when run on identical chromatographic methods (section 2.7.5).

PAMPA was performed separately on each pure, resynthesized compound in quadruplicate and the results compared to their library permeabilities in Figure 2.5. Compound 2.1 could not be displayed on the log scale because it was impermeable both in the library and as a pure compound, resulting in 18 contributing data points from 19 compounds. Library and resynthesized permeability values correlated well (R-squared of 0.88), but differed by up to 0.36 log units or 2.29-fold (average 0.14 or 1.39-fold) in either direction for data above $0.5 * 10^{-6}$ cm/s, below which the measurement deviation of the pure compounds increased. This mixture effect was not found to be concentration-dependent under non-sink conditions, which, lacking sinks in both wells, should be more vulnerable to such effects (section 2.5.1).

2.3 Discussion

2.3.1 Matched-Pair Analysis

To best take advantage of the individual permeabilities of thousands of structurally diverse cyclic peptides, we performed an analysis of compound pairs differing only at a single residue. In most cases such comparisons can have their differences in permeability attributed partly to lipophilicity and partly to their impact on a peptide's conformational dynamics. While the intrinsic permeability of these scaffolds is not directly assessed by PAMPA, the "double-sink" conditions reduce the impact of water solubility on permeability, producing a closer estimate of intrinsic permeability than non-sink PAMPA. The impact of lipophilicity has been minimized by the library designs except at position Y (Fig. 2.1) and Figures 2.6 and 2.7 are colored by the NH amide count of the X-axis to clarify any differences in behavior originating from lipophilicity. In comparisons where the position of the matched pair of residues could vary, such as Figure 2.7, alternate scatterplots colored by pairing position were generated (section 2.5.2). Of particular interest were the effect of stereochemical inversions, peptoid residues, and beta residues on permeability.

2.3.1.1 Stereochemical Inversion

Single stereochemical inversions have been observed to have dramatic effects on the passive permeability of individual compounds, and we observe differences in permeability of over 70-fold in both libraries. However, we observed no systematic preference for either D- or L-stereochemistry greater than an average of 0.14 log units at any position, highlighting the context dependence of stereochemistry on conformation (and therefore, permeability). Stereoinversion had the greatest impact on the variable position *i – 1* to the static L-Phe residue, demonstrating preferences for D-Leu over L-Leu (0.14 log units hexa- and 0.12 log units heptamer) and L-(NMe)Leu over D-(NMe)Leu (0.11 log units hexa- and 0.09 log units heptamer). The adjacent proline (*i – 2* to the L-Phe) demonstrated no stereopreference in the hexamer library and only a small (0.08 log units) preference for D-Pro in the heptamer library, leading us to conclude that the static L-Phe templates the stereopreference of its *i – 1*

residue. No similar influence was observed *i + 1* to the static L-Phe residue and all other positional trends averaged lower than 0.1 log units (section 2.5.2). This matched pair analysis revealed many "permeability cliffs" (represented by points that lie off the diagonal) in which a single stereochemical inversion results in a large permeability increase or decrease. In addition, there are many points close to the diagonal that represent positions which are tolerant to stereoinversion.

2.3.1.2 Beta Residues

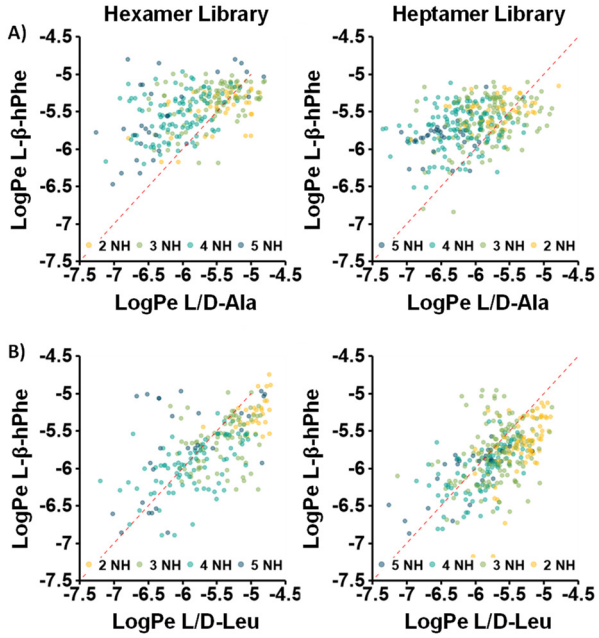


Figure 2.6 Matched-pair analysis of L-β-hPhe substitutions. Log PAMPA permeability scatters between compound pairs differing at a single position. A) L/D alanine to L-β-hPhe or B) L/D leucine to L-β-hPhe. Each point is colored by the NH amide count of the residues on the X-axis, though here there is no difference between sequences. A dashed red line of identity has been added to aid in discerning the overall impact of each mutation.

Having established that there is no systematic relationship between stereochemistry and permeability, we sought to understand the impact of β-residues on passive permeability by examining compound pairs in which L-β-hPhe replaces non-N-alkyl residues of either stereochemistry. L-β-hPhe appeared at only two positions per library and was therefore examined per position as a replacement for L/D-Ala at the Y position (Fig. 2.6A) and L/D-Leu at position X^3 in the hexamer library and X^4 in the heptamer library (Fig. 2.6B). One might expect a higher permeability from compounds including L-β-hPhe residues based on their greatly increased lipophilicity compared to Ala and Leu, especially under sink conditions where water solubility is less relevant. The large increase in permeability observed from the Ala to L-β-hPhe substitution accorded with that expectation, however the Leu to L-β-

hPhe substitution resulted in a decrease in permeability instead of the smaller, positive impact expected from a solely lipophilic perspective. Therefore, the L-β-hPhe residue must have a structural impact on passive permeability in one or both positions.

The major structural difference between Leu and L-β-hPhe is an increase in backbone flexibility. Given that the Leu to L-β-hPhe substitution is *i -1* to a proline residue in all cases, the observed decrease in permeability may be a result of increased flexibility disrupting a proline-templated β-turn conformation hosting intramolecular hydrogen bonds. This is surely not true of all affected compounds, however. One hypothesis is that the flexibility of L-β-hPhe may also be detrimental to permeability at the Y position as well, but that the greatly increased lipophilicity of L-β-hPhe compared to Ala and, potentially, the increased steric occlusion of the hPhe sidechain result in a net positive impact on permeability. Though our sampling of beta residues is limited, the increased flexibility of L-β-hPhe is detrimental in at least the Leu to L-β-hPhe substitution.

2.3.1.3 Peptoid Residues

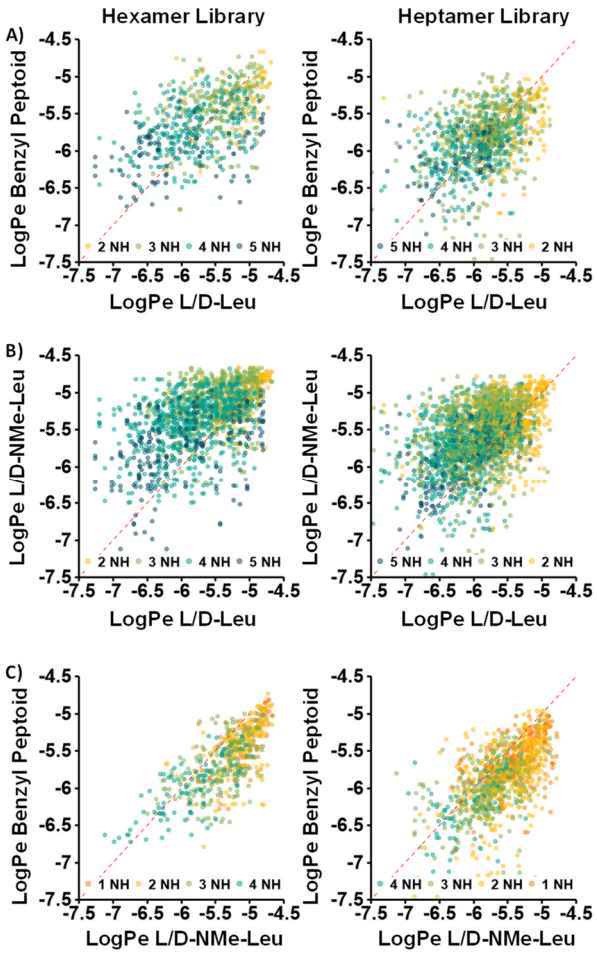


Figure 2.7 Matched-pair analysis of benzyl peptoid substitutions. Log PAMPA permeability scatters between compound pairs differing at a single position. A) L/D leucine to benzyl peptoid, B) L/D leucine to L/D N-methyl leucine, or C) L/D N-methyl leucine to benzyl peptoid. Each point is colored by the NH amide count of the residues on the X-axis. A dashed red line of identity has been added to aid in discerning the overall impact of each mutation.

Peptoid residues have great utility for introducing chemical diversity not easily available in amino acid form into combinatorial libraries, and we therefore sought to explore their impact on passive permeability in detail. Here we focus on Leu or (NMe)Leu to benzyl peptoid substitutions, which were thoroughly sampled by these libraries. Benzyl peptoid was chosen for the X-residue position because of its lipophilic similarity to (NMe)Leu, simplifying this analysis. Removal of a hydrogen bond donor by substituting benzyl peptoid for L- or D-Leu (Fig. 2.7A) resulted in slightly increased permeability in both libraries. This substitution is overall positive only because the *i - 1* position to the static L-Phe has a permeability preference for N-alkylation; the other positions are neutral (section 2.5.2). By comparison, the Leu to (NMe)Leu substitution (Fig. 2.7B) increased permeability greatly at all positions in both libraries. Consistent with the previous comparisons, the (NMe)Leu to benzyl peptoid substitution (Fig. 2.7C) decreased permeability equally at all positions in both libraries.

Neither the difference in lipophilicity between the two residues nor the position of the substitution accounts for this decrease, leading us to believe their difference in permeability originates from the structural impact of peptoid substitutions, perhaps arising from the comparative increase in backbone flexibility. Although previous work from our group has shown that strategic N-Me-to-peptoid substitutions can preserve and even enhance the permeability of individual cyclic peptide scaffolds[30, 53, 59], the present study provides a more comprehensive picture of the effect of peptoid substitutions across a wide variety of different backbone geometries. Thus, although, on average, substitution of N-Me amino acids for peptoids tends to decrease permeability, peptoids remain an attractive option for easy functional diversification for scaffolds that tolerate them.

The negative effect on permeability of peptoid and beta residues suggests that additional flexibility is unfavorable in these libraries. The proline residue(s) in each library may contribute to this behavior by templating a double β-turn conformation for maximal intramolecular hydrogen bonding that increased flexibility disrupts. While rigidity can be a successful strategy for permeability in hexa- and heptamer cyclic peptides (e.g. 1NMe3[46]), large passively permeable scaffolds such as cyclosporin A[61] are required to leverage flexibility in a chameleonic behavior[62] to maintain both water solubility and passive permeability simultaneously. The lower permeabilities observed for scaffolds containing

multiple peptoid or beta residues in these libraries may give some insight into the rarity of large permeable scaffolds.

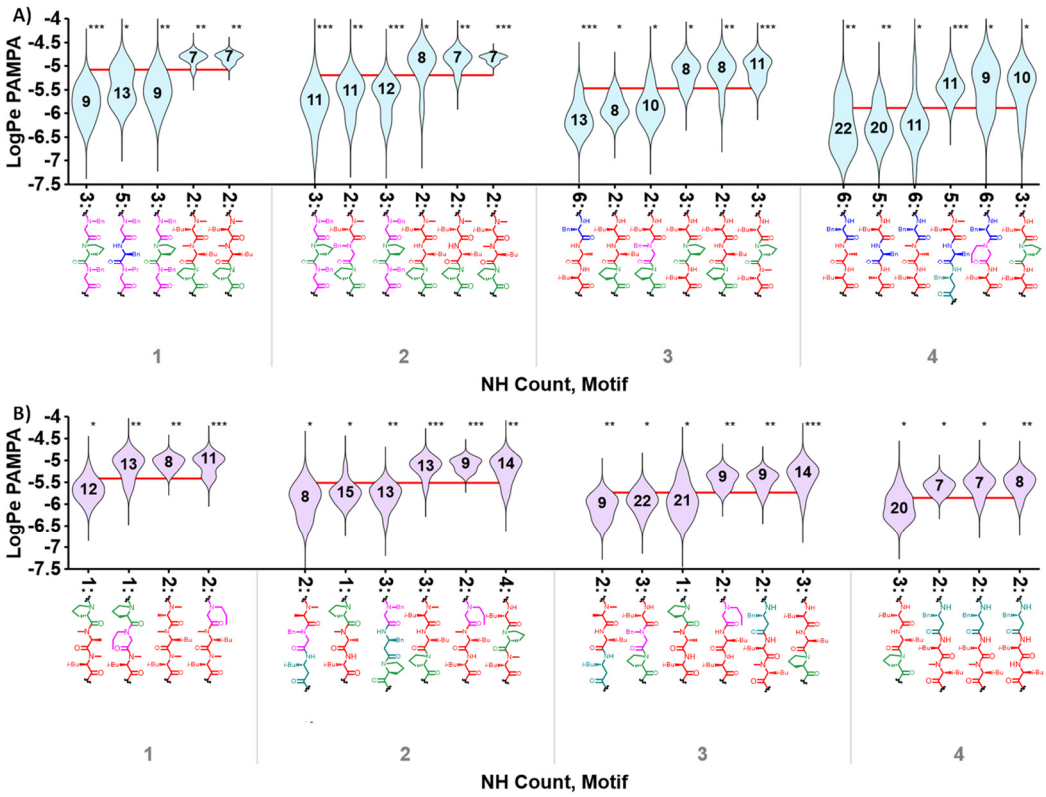


Figure 2.8 Maximal impact of 3mer motifs on hexamer and heptamer library permeabilities. Violin plots of the least and most permeable three residue long contiguous motifs were generated similarly to Figure 2.4 for each amide NH count for both libraries. Up to three permeability-enhancing motifs and three permeability-reducing motifs qualified as well-sampled (see text) are displayed within each bin. The median of all motifs within a bin is denoted by a horizontal red line. Very few motifs in the 5 amide NH bin are well-sampled and thus that bin is not displayed. Motifs denoted by their position of origin as in the library schematics (Fig. 2.1) and a structural representation colored red for aliphatic residues, magenta for peptoid residues, cyan for beta residues, green for proline residues, and blue for the static L-Phe residue.

2.3.2 Passive Permeability Motifs

Technologies such as in-vitro translation[42] and light-directed peptide array synthesis[63] can generate cyclic peptide libraries biased towards passive permeability by including only permeable backbone geometries, but other methods of library synthesis lack total synthetic control and cannot take the same approach without severely limiting their size. We therefore sought an alternative method of leveraging our permeability data for the benefit of such

combinatorial library generation techniques by investigating the efficacy of permeability-enhancing substructures. These permeability-enhancing motifs would be held static while allowing the rest of the cycle to vary over a collection of backbone geometries biased towards permeability, increasing the potential library size without abandoning a known set of permeable backbone geometries. Thus, we sought to address the prevalence of motifs with significant effects on permeability, their maximal impact on permeability, and their scope of applicability.

The permeability motifs examined herein are contiguous 3-residue sequences generated from all possible such sequences at each position of a library's schematic in a position-specific manner. Each library compound is therefore associated to a number of motifs equal to its length and each motif has many associated compound permeabilities. A motif length of three was chosen as a useful length for both hexamers and heptamers while allowing significant variation in the remainder of the peptide. For all inter-motif comparisons, we found it necessary to segregate the compounds associated with each motif by NH amide count to ensure that simple lipophilicity did not drive comparisons of motif impact on permeability. This binning scheme introduces its own bias on compound membership in a given motif by restricting the N-alkylation of the remainder of the cycle, but nonetheless results in an approximation of intrinsic permeability (section 2.5.3).

2.3.2.1 Motif Prevalence

To determine the prevalence of motifs with a significant effect on permeability we used a Mann-Whitney U-test[64] between the log permeabilities of one motif vs. the population of all log permeabilities for all motifs within the same amide NH count. We found that 146 of 366 hexamer motifs detected and 197 of 391 heptamer motifs detected significantly affected permeability for one or more NH amide counts. As expected for conformational motifs derived from such a structurally diverse compound set, a roughly equal number of significant permeability-enhancing motifs were discovered (69 hexamers, 87 heptamers) as permeability-reducing motifs (75 hexamers, 106 heptamers). Statistical significance in a U-

test alone, however, is not sufficient for a motif to be useful in practice, and we therefore implemented additional requirements for a motif to be considered well-characterized.

The ideal permeability-enhancing motif for biasing a combinatorically generated screening library contains many backbone geometries for better odds of target engagement and a greater potential library size. These factors in mind, we defined well-sampled motifs to be associated with a minimum of six compounds with known, non-zero permeabilities – which must be at least two-thirds of all theoretical library members containing that motif. These requirements resulted in a final total of 81 well-sampled permeability-enhancing motifs (25 hexamer, 56 heptamer) across all amide NH counts and 39 well-sampled permeability-reducing motifs (20 hexamer, 19 heptamer). Few permeability-reducing motifs were considered well-sampled due to the higher proportion of impermeable compounds composing them compared to permeability-enhancing motifs. Each of the 81 well-sampled permeability-enhancing motifs identified represents a set of backbone geometries potentially useful in biasing a combinatorically generated screening library towards passive permeability.

2.3.2.2 Motif Impact

To investigate the maximal effect of a well-sampled motif on passive permeability, up to three of the least and most permeable well-sampled motifs for each NH amide count were plotted in Figure 2.8. The same trends observed in the matched pair analysis were visible at some amide NH counts (benzyl peptoid in hexamer bins 1 and 2, L-β-hPhe in heptamer bin 4) but did not dominate the motifs discovered. The most permeable motifs improved upon the population median permeability at that NH amide count by between 0.28 and 0.63 log units (1.9- and 4.3-fold) in the hexamer library and between 0.36 and 0.46 log units in the heptamer library (2.3- and 2.9-fold). While these benefits are smaller than picking the only the most permeable backbone geometries, incorporating permeable backbone geometries into a library design in either fashion also serves to prevent inclusion of backbone geometries with especially poor permeability. The least permeable motifs generally had a similar level of impact, ranging between 0.43 and 0.68 log units (2.7- and 4.8-fold) lower permeability than

the population median in the hexamer library and 0.24 to 0.32 log units (1.7- and 2.1-fold) lower in the heptamer library. Reducing the requirements on well-sampled motifs had little effect on the maximal impact of permeability-enhancing motifs but increased the maximal impact of permeability-reducing motifs greatly (section 2.5.3). From this data we conclude that permeability motifs can have a sufficiently large impact to be of utility in biasing combinatorically generated libraries toward passive permeability even among this structurally diverse set of compounds.

2.3.2.3 Motif Scope

We investigated the degree of correlation between the impact on permeability of similar motifs between the hexamer and heptamer libraries to determine their structural scope. The previous positional definition for motifs made a direct comparison between the hexamer and heptamer libraries impossible so we created a position-independent motif naming scheme. This new motif naming scheme categorized residues as L-NH, L-N-Me, D-NH, D-N-Me, L-proline, D-proline, peptoid residue, or beta residue rather than use exact residue identities. While this resulted in some cases of composite motifs with less extreme impacts on permeability, few motifs were impacted (see additional discussion).

A linear regression of the permeability comparison between hexamer and heptamer libraries of all identically named motifs without any filtering resulted in an R-squared of 0.52, which was nearly the strongest correlation (0.55) found in all attempted comparisons (section 2.5.3). Although there was not a strong overall correlation between the permeabilities of individual motifs in the hexamer and heptameter libraries, some permeability-determining motifs were common to both ring sizes. For example, the D-Leu – L-Leu – D-Pro motif was highly favorable among compounds containing 3 backbone NH amides across both the hexamer and heptamer libraries while all stereochemical variants of the Leu – Bnz – Pro motif were among the least permeable. Similarly, the D-NMe-Leu – D-Leu – D-Pro motif was highly favorable among compounds containing 2 NH amides in both libraries. While instances of agreement do exist between libraries, most permeability-effecting motifs did not correlate,

leading us to conclude that their utility is limited to libraries of the same ring size and possibly the same proline placement. Despite these restrictions of scope, the conformational nature of a permeability motif should allow these motifs to maintain utility in the presence of a variety of alternative sidechains.

We have demonstrated the presence and dramatic potential impact of permeability-enhancing motifs on passive permeability and gained insights into the structural scope of their applicability. The permeability-enhancing motifs discovered expand the utility of our library data to combinatorically generated screening libraries in addition to those technologies with full synthetic control by offering sets of interrelated backbone geometries with a net positive impact on passive permeability.

2.4 Conclusions

To increase our understanding of passive permeation in cyclic peptides and to aid in biasing screening libraries composed of cyclic peptides towards passively permeable hits, we synthesized hexamer and heptamer libraries with diverse backbone geometries and quantified their PAMPA permeabilities as mixtures by MSMS. We validated that individual compounds can be extracted from our library analyses by the agreement between the pure PAMPA permeabilities of 19 resynthesized compounds with the corresponding mixture-based library peaks. Our results confirm our previous understanding of the impact of free amide NHs and N-methylation across thousands of compounds, elucidate the effects of single backbone modifications, and introduce the concept of motifs for passive permeability.

Our matched-pair analysis of backbone modifications showed that single stereochemical inversions cannot in general control the passive permeability of a set of geometrically diverse cyclic peptides but also confirmed that dramatic “permeability cliffs” were not uncommon in individual cases. Although the matched pair analysis of L-β-hPhe was complicated by differences in lipophilicity, sidechain volume, and proximity to proline resides, we hypothesize that its increased flexibility had a negative impact on permeability. The results of our matched pair analysis of benzyl peptoid led us to a similar conclusion, with the

negative effect of replacing an N-methyl residue especially telling. Our data suggests that for hexa- and heptapeptides, in most cases, the entropic cost of assuming a non-polar conformer for membrane transit is higher than the enthalpic gains compared to a more rigid scaffold. Although beta and peptoid residues are attractive for the structural and chemical diversity they provide, these findings suggest that their inclusion in libraries biased towards passive permeability should be limited where possible.

Our investigation of passive permeation in relation to 3-mer motifs revealed that “permeability motifs” and “anti-permeability motifs” exist even among compound sets with diverse backbone geometries and few shared structural elements. We found dozens of permeability-enhancing motifs sampled well enough to be of practical use in biasing combinatorically generated libraries towards passive permeability, the best of which increased permeability 2-fold to 4-fold compared to all other compounds with the same number of hydrogen bond donors. We find this effect size remarkable for libraries with such a high structural diversity. Although few motifs were shared between the hexa- and heptamer libraries, permeability-enhancing motifs likely exist for larger ring sizes as well.

We have demonstrated that individual library members can be extracted from these libraries with a good correlation to their mixture-based permeability measurement. Compounds with particularly high permeabilities may be useful to bias screening technologies with a high degree of control over library design by including many “privileged” backbones and iterating their sidechains to search for activity. The permeability motifs discovered herein extend the utility of our permeability data to library generation technologies without complete synthetic control by allowing the remainder of the cycle to vary. Any hits from such permeability-biased libraries would require fewer optimizations to achieve passive permeability and move towards oral bioavailability. Though we present only hexa- and heptamer scaffolds herein, this approach can be applied to identify such privileged backbone geometries in other systems, and we are currently using it to explore larger cycles.

2.5 Supplementary Figures and Discussion

2.5.1 Effect of Concentration on Sub-library Permeability

A dilution series of hexamer sub-library DLMA was run at 500 μM total concentration (as the final PAMPA data, with 1NMe3 standard), 100 μM total concentration, and 20 μM total concentration under non-sink PAMPA conditions to investigate the impact of sub-library concentration on PAMPA results. Possible causes of such variation include multi-component aggregation, adsorption to the plastic of the acceptor well or subsequent vessels before analysis. These experiments were not performed in sink conditions both for better signal-to-noise and because the sources of concentration dependent variation mentioned above are mitigated under sink conditions by the presence of an excess of Polysorbate-80 and TGPS. 1NMe3 was included in the 500 μM stock pre-dilution and was within the expected intra-plate percent coefficient of variation (%CV) of around 10% for compounds in the range of 1 to 10 x 10^{-6} cm/s permeation rate (as set out by Avdeef et al.[65]) for all three well-pairs. Fifty-nine manually curated peak triplets were obtained, and their standard deviations were used to calculate an average %CV of 9%, corresponding to expected intra-plate variation. We therefore find no evidence of concentration-dependence for the mixture permeabilities in non-sink conditions – where they should be susceptible to such effects – and suggest that this is also true under sink conditions – where they should not. We also propose that the differences in permeability between resynthesized pure compounds and library data originate from mixture-based effects and not concentration-based effects.

Peaks were required to have correct automated integration and have been detected in the acceptor well (thus having a known permeability) to be included in this analysis. The number of triply detected peaks was low due to low signal in the 20 μM acceptor well. For reference, 112 curated peak pairs were identified between the 500 μM and 100 μM concentrations.

2.5.2 Additional Discussion of Matched-Pair Analysis

2.5.2.1 Additional Discussion of Stereoinversion Matched Pairs

Matched pair analysis scatterplots for stereochemical inversions were not included in the main discussion due to the lack of strong trends but are included here for completeness (Figure S2.1). We did not discuss all trends below 0.1 log units because such low effect sizes may result from the 40% of theoretical library sequences with no associated permeability data. Tables S2.1 and S2.2 summarize the averaged effect of all stereoinversions in log units, including a breakdown by substitution position. Tables S2.3 and S2.4 summarize all other matched pair substitutions similarly.

Table S2.1 Averaged effect of stereoinversion matched pairs on hexamer permeability (log units)

Substitution (L->D)	Combined	Y^1	X^2	X^3	Pro^4	X^5	L-Phe^6
Ala	-0.019	-0.019					
NMe-Ala	-0.023	-0.023					
Leu	-0.002		0.024	-0.140		0.140	
NMe-Leu	-0.032		0.001	0.047		-0.107	
Pro	-0.003				-0.003		

Table S2.2 Averaged effect of stereoinversion matched pairs on heptamer permeability (log units)

Substitution (L->D)	Combined	Pro^1	Y^2	X^3	X^4	Pro^5	X^6	L-Phe^7
Ala	-0.053		-0.053					
NMe-Ala	-0.092		-0.092					
Leu	-0.034			-0.088	-0.083		0.124	
NMe-Leu	-0.029			0.026	0.001		-0.087	
Pro	0.014	-0.050				0.080		

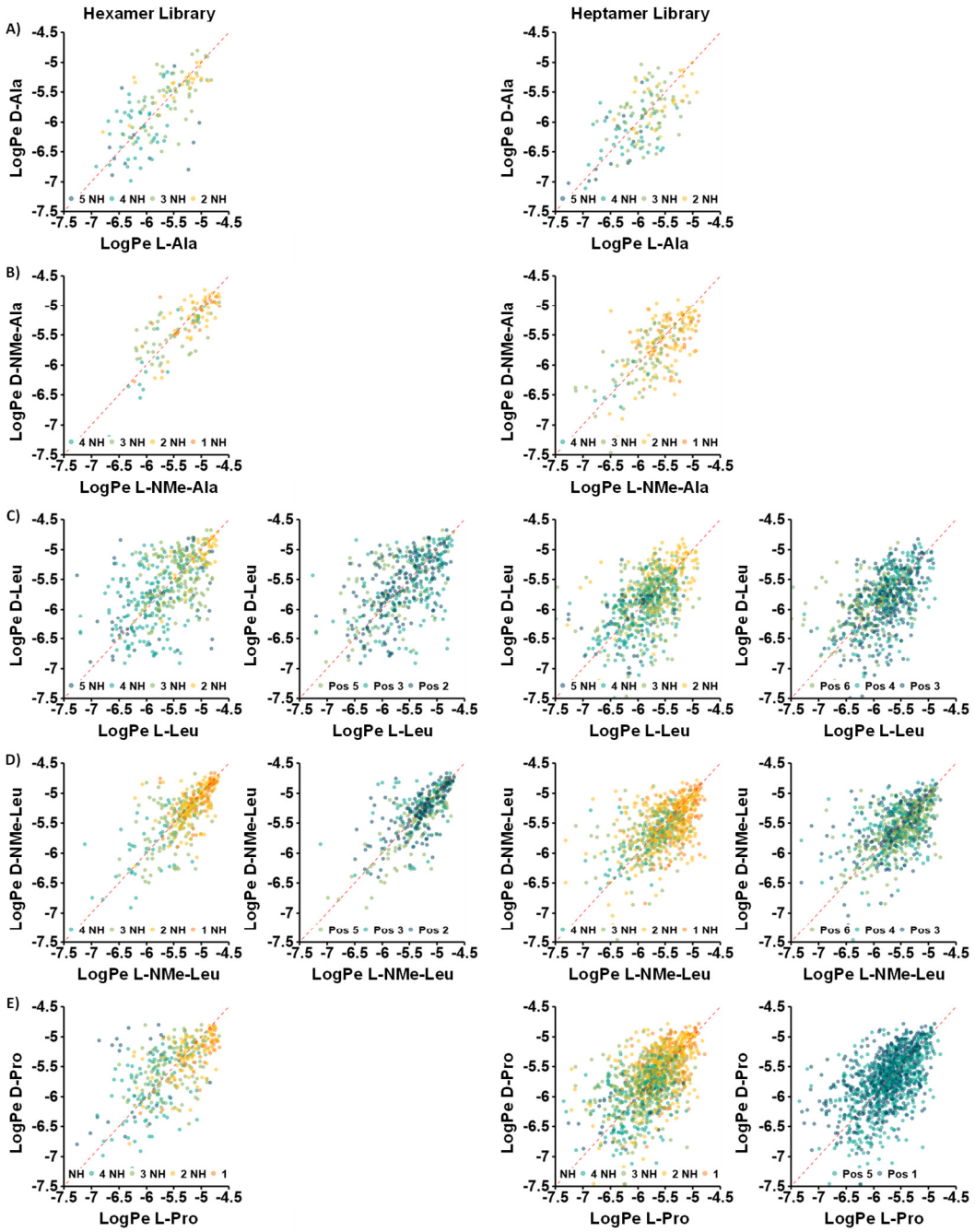


Figure S2.1 Matched-pair analysis of stereochemical inversions.
Scatterplots of the permeabilities of compound pairs containing the X-axis residue and the Y-axis residue at the same position (in this case, L-stereo on X-axis, D-stereo on Y-axis). Each library has two plots for each matched pair; the left is colored by NH amide count of the X-axis sequence and the right by substitution position. Not all such matched pairs could occur at multiple positions. A) Alanine. B) N-methyl alanine. C) Leucine. D) N-methyl leucine. E) Proline.

Table S2.3 Averaged effect of other matched pairs on hexamer permeability (log units)

Substitution	Combined	Y^1	X^2	X^3	Pro^4	X^5	L-Phe^6
L/D-Ala to L-β-hPhe	0.344	0.344					
L/D-Leu to L-β-hPhe	-0.124			-0.124			
L/D-Leu to benzyl peptoid	0.061		0.025	-0.050		0.191	
L/D-Leu to L/D-NMe-Leu	0.330		0.337	0.252		0.388	
L/D-NMe-Leu to benzyl peptoid	-0.275		-0.288	-0.311		-0.245	

Table S2.4 Averaged effect of other matched pairs on heptamer permeability (log units)

Substitution	Combined	Pro^1	Y^2	X^3	X^4	Pro^5	X^6	L-Phe^7
L/D-Ala to L-β-hPhe	0.355		0.355					
L/D-Leu to L-β-hPhe	-0.156				-0.156			
L/D-Leu to benzyl peptoid	0.025			-0.052	-0.087		0.200	
L/D-Leu to L/D-NMe-Leu	0.294			0.226	0.193		0.460	
L/D-NMe-Leu to benzyl peptoid	-0.269			-0.255	-0.263		-0.282	

2.5.2.2 Impact of Substitution Position on Benzyl Peptoid Matched-Pair Analysis

The substitutions explored in the matched-pair analysis of benzyl peptoid each contained pairs from three separate positions on the cycle (all variable X positions). The average impact of these (and other) matched pairs are summarized in Tables S2.3 and S2.4 and visualized in Figure S2.2. Figure S2.3 focuses on the individual positional trends for the Leu to Bnz and Leu to (NMe)Leu substitutions to highlight the different behavior of the variable position *i – 1* from the static L-Phe residue.

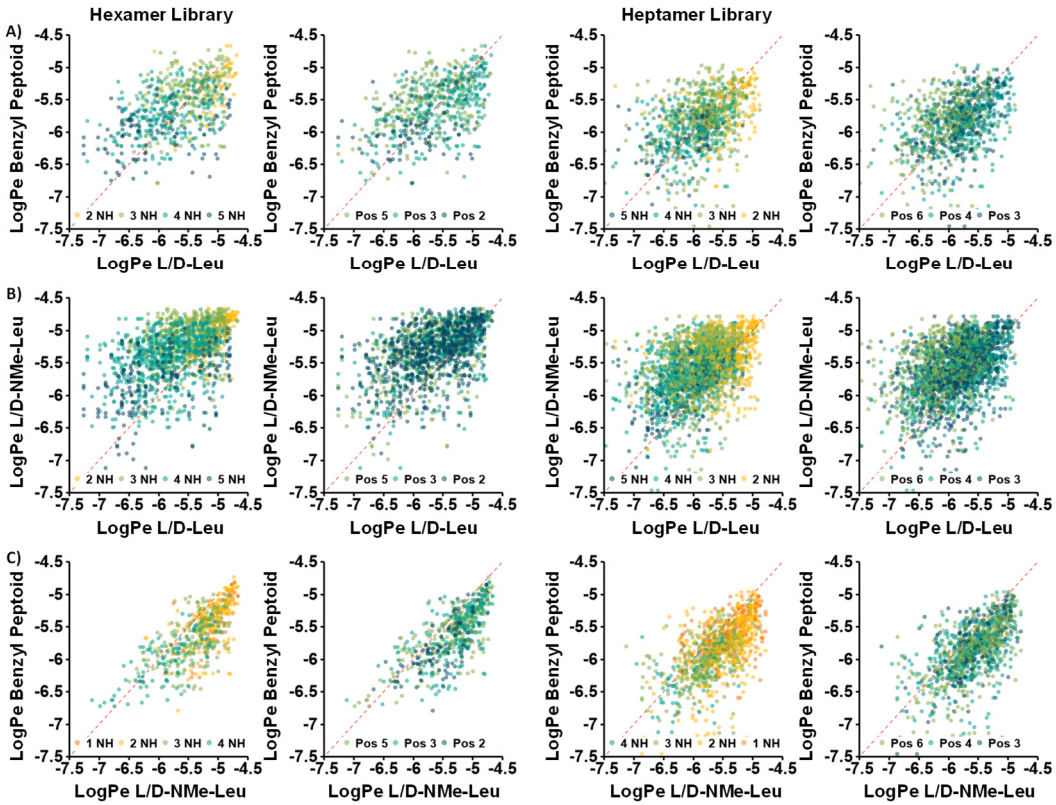


Figure S2.2 Matched-pair analysis of benzyl peptoid.
As Figure SI2.1. A) L/D leucine to benzyl peptoid. B) L/D leucine to L/D N-methyl leucine. C) L/D N-methyl leucine to benzyl peptoid.

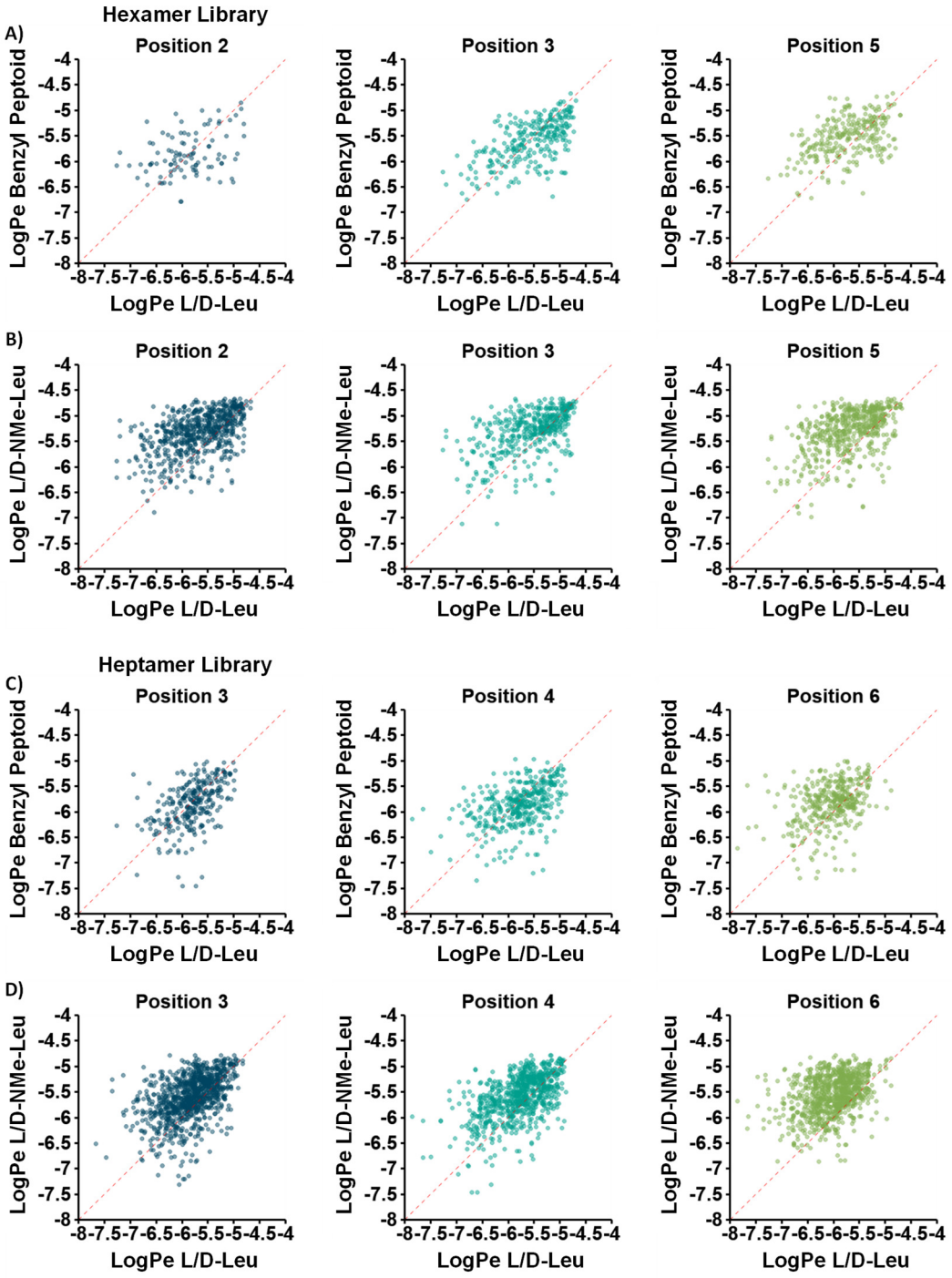


Figure S2.3 Positional breakout of matched-pair analysis of benzyl peptoid. One scatterplot per substitution position has been generated for the L/D leucine to benzyl peptoid substitution **(**A) hexamer and C) heptamer**)** and the L/D leucine to L/D N-methyl leucine substitution **(**B) hexamer, D) heptamer**)** similarly to Figure S2.1. The preference of hexamer position 5 and heptamer position 6 for N-alkylation is visible (except in part B, where the difference is less obvious).

2.5.3 Additional Discussion of Motif Analyses

2.5.3.1 Bias Induced by Binning by NH Amide Count

Although binning motifs by amide NH count mitigates the impact of lipophilicity, it also results in different biases at each count by limiting the remainder of the cycle. Motifs of NH amide counts one and five are the least affected by the binning due to the requirement that all variable residues be N-alkyl or N-H respectively but are the least populated for the same reason. NH amide counts two and four are contain only one variable N-alkyl or N-H residue respectively and therefore we expect the most impactful motifs to control that residue's position, thus narrowing the range of lipophilicity for the remaining members and yielding a more consistent permeability – and indeed this is true. At NH amide count two, only 2 (of 24) well-characterized permeability-enhancing motifs do not control the single variable-position hydrogen bond donor across both libraries. The NH amide count of four is less affected by this bias, with only 9 (of 26) motifs controlling the sole variable-position N-alkyl residue. One possible driver for this discrepancy is the difference in the importance of control over one of two hydrogen bond donors at NH amide count two versus the control of an N-alkyl blocker. Although these biases influence the significant motifs available at each NH amide count, the motifs are nonetheless based on real structural trends. We judged a structural basis for motif bias as preferable to simple lipophilicity.

Because of the limitations outline above, binning by NH count has exaggerated the trends observed in the matched pair analysis within select bins. One example visible in Figure 8A is the prevalence of benzyl peptoid in the least permeable motifs at NH amide counts one and two. At these NH amide counts the peptoid occupies a position which would otherwise contain an N-methylated residue – a substitution that generally results in lower permeability. Another example is the prevalence of L-β-hPhe at NH amide count four in Figure 8B, where it occupies position two of all three of the most permeable motifs (and all nine permeability-enhancing motifs not displayed). Given that only one N-alkyl residue is allowed at NH amide count four, we can consider the L-β-hPhe residue as a replacement for

an N-H alanine residue – a substitution known to generally increase permeability. However, such instances did not dominate the discovered motifs.

2.5.3.2 Additional Biases

We also observed that a greater number of motifs were significant for positions controlling more variable residues (residue X or Y as Figure 2.1) in both the full motif population and the well-characterized motif population across both libraries. We believe this to be a result of the structural diversity in these libraries, with a large effect on passive permeability being more likely for motifs restricting the backbone geometry as completely as possible. We observed no particular bias toward control of the proline residues either because this bias was stronger or there was no such bias.

As mentioned in the main text, impermeable compounds were uncommon in well-sampled motifs due to the nature of their filtering (two-thirds of permeabilities must be known). This was especially true of well-sampled permeability-enhancing motifs, of which only 2 of 27 hexamer motifs and 11 of 65 heptamer motifs across all NH amide counts contained between one and two impermeable compounds. None contained more than two impermeable compounds.

2.5.3.3 Effects of Various Motif Sampling Thresholds on Motif Impact

In the main text we briefly mentioned that changing the threshold filters for motifs to be considered well-sampled did not produce a dramatic effect on our results for permeability-enhancing motifs. The impact of full thresholds, half thresholds, and no thresholds – for both minimum number of compounds with associated, non-zero permeabilities and the minimum fraction of such compounds over the count of compounds theoretically containing the motif in question – on the hexamer library is shown in Table S2.5 below. As mentioned in the main text, these thresholds remove many permeability-reducing motifs from consideration due to their higher likelihood of containing sequences corresponding to impermeable compounds.

Table S2.5 Effects of varying thresholds for "well-sampled" motifs on hexamer motif prevalence and impact

NH Count: Metric	Min Samples > 6 Frac Known = 0.66	Min Samples > 3 Frac Known = 0.33	Min Samples = 0 Frac Known = 0
1: Count Positive	2	7	11
1: Count Negative	4	8	14
1: Max Positive Log Impact	0.28	0.31	0.31
1: Max Negative Log Impact	0.68	0.68	1.14
2: Count Positive	5	28	33
2: Count Negative	4	21	32
2: Max Positive Log Impact	0.38	0.44	0.48
2: Max Negative Log Impact	0.52	0.84	1.35
3: Count Positive	9	23	24
3: Count Negative	11	31	36
3: Max Positive Log Impact	0.51	0.53	0.59
3: Max Negative Log Impact	0.68	0.84	0.84
4: Count Positive	7	17	18
4: Count Negative	4	12	13
4: Max Positive Log Impact	0.63	0.72	0.72
4: Max Negative Log Impact	0.43	0.87	0.87

2.5.3.4 Consequences of a Positionally-Variant Motif Definition

The non-positional motif naming scheme proposed in the main text (categorizing residues as L-NH, D-NH, L-N-methyl, D-N-methyl, peptoid, L-proline, D-proline, or beta) resulted in multiple positional motifs being grouped under the same non-positional name. The positional motifs most often combined under the non-positional naming scheme contain no proline or beta residues because both are limited to a small number of positions. Non-positional motif names such as *L-NH, L-NH, L-NH* have many potential positions of origin in the hexamer library (XXYFX region as Figure 2.1) but fewer such ambiguities in the heptamer library due to the second proline residue. Despite this, most well-characterized non-positional motifs contain only one positional motif.

To investigate cases where unlike positional motifs were grouped under the same non-positional name, we first looked for cases in which the positional motifs differed from each other by U-test. We found 16 such hexamer motifs among the 767 motifs spread across all NH amide counts and 31 such heptamer motifs among 915 motifs similarly. We next searched for non-positional motifs in which one or more positional motifs was more significant by U-test against the population than the non-positional (combined) motif. This proved

somewhat more common, with 33 hexamers and 43 heptamers that fit these criteria. Although few, these situations each misrepresent the permeability of multiple motifs. However, the remainder of non-positional motifs not mishandled should have been sufficient to reveal any correlations between the hexamer and heptamer libraries had they existed.

2.5.4 Correlation of Permeability Motifs Between Libraries

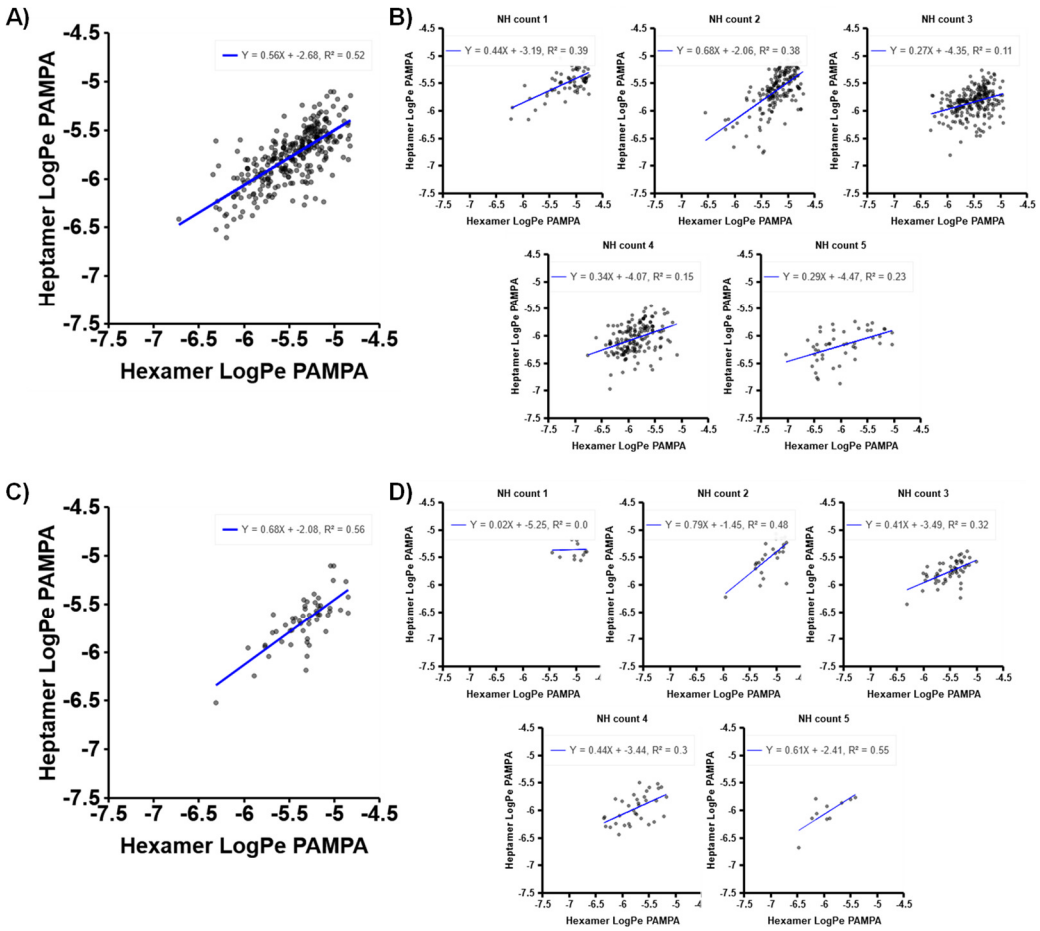


Figure S2.4 Non-positional motif comparison between libraries.
A) Scatter between median permeabilities of motifs found in both libraries. B) Scatters of median motifs found in both libraries binned by NH amide count to mitigate the effect of lipophilicity on any correlation. C) Scatter between median permeabilities of motifs with at least two-thirds of their member permeabilities known and non-zero in both libraries. D) As B) with the restrictions of C). In addition to those plots shown here, correlations were also attempted by rank-order of motifs within each library without useful results. No threshold for motif minimum sample size or statistical significance was enforced for these plots.

2.6 Abbreviations

ACN, acetonitrile; COMU, (1-Cyano-2-ethoxy-2-oxoethylidenaminooxy)dimethylamino-morpholino-carbenium hexafluorophosphate; DBU, 1,8- Diazabicyclo[5.4.0]undec-7-ene; DCM, dichloromethane; DMF, N,N-dimethylformamide; DIPEA, diisopropylethylamine, DMSO, dimethylsulfoxide; Fmoc, 9-fluorenylmethoxycarbonyl; HATU, 1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxide hexafluorophosphate; HPLC, high-pressure liquid chromatography; LCMS, liquid chromatography mass spectrometry; MeOH, methanol; N_2, nitrogen; PTFE, Polytetrafluoroethylene; RP, reverse phase; SPPS, solid phase peptide synthesis; TFA, trifluoroacetic acid; TGPS, D-α-Tocopherol polyethylene glycol 1000 succinate; UV, ultra-violet; NMR, nuclear magnetic resonance; OBOC, one-bead one-compound; PAMPA, parallel artificial membrane permeability assay; ADME, absorption, distribution, metabolism, excretion.

2.7 Methods

All chemicals were commercially available and used without further purification except where mentioned. Frequent reference to chapter 1 methods will be made for concision, though some information will be repeated as part of communicating alterations to previous methods or for convenience.

Solvents used in synthesis and Optima grade Acetonitrile and Formic Acid for HPLC use were purchased from Fisher Scientific. Fmoc-protected amino acids and coupling agents were purchased from Combi-Blocks, Oakwood, or Chem-Impex. Tri-deuterated L-Leucine and NMR solvents were purchased from Cambridge Isotope Laboratories. 2-chlorotrityl chloride polystyrene resin and L-Phe-2-chlorotrityl polystyrene resin were purchased from Rapp-Polymere. Dodecane and soy lecithin used in PAMPA were purchased from Alfa Aesar.

2.7.1 Equipment and Analytical Methods

Purification was performed on a Biotage Isolera Prime HPLC using a SNAP Ultra C18 12 g cartridge and on a Waters HPLC (Waters 1525) with an attached mass spectrometer (Micromass ZQ, waters), PDA detector (Waters 2998), sample manager (Waters 2767), and supplementary pump (Waters 515) using a 5 µm C18 column (XBridge BEH130 Prep C18 OBD 19x150 mm). All purification was reverse phase in a mixture of water and ACN, containing either 0.1% TFA or 0.1% formic acid for the Biotage and Waters systems, respectively. Collection was UV-triggered for both purification apparatus. All purification on the Biotage system followed the same LC method, starting with 2 column volumes at 30% ACN, increasing to 100% ACN over the next 12 column volumes, and remaining at 100% ACN for a further 4 column volumes. Purification on the Waters system was either a 20 min isocratic or a 20 min gradient method optimized for each compound (X% ACN 0-3 min, X%-100% 3-17 min, 100% 17-20 min).

Fraction composition and purity was tested on the same Waters system through a 3.5 µm C18 column (XBridge BEH C18 4.6×50 mm) or by direct inject to an Advion ExpressIon mass spectrometer. Analytical runs on the Waters system were 12 minutes long with 1.2 mL/min flow rate; ACN concentration was increased stepwise from its starting concentration (30% 0-2 min, linear increase to 100% 2-10 min, 100% 10-12 min). Results were analyzed by MassLynx4.1 or MassExpress, respectively.

Tandem MS runs and LC-MS analyses for PAMPA of libraries was performed on an UHPLC (UltiMate 3000, Dionex) with attached mass spectrometer (Orbitrap Velos Pro, Thermo Scientific) using a 2.2 µm C18 column (Acclaim 120 2.1×250 mm, Thermo Scientific) with a column heater set to 50° C. A mixture of water (0.1% formic acid) and ACN (0.1% formic acid) was used as an eluent, with data processed by XCALIBUR version 2.2 SP1.48. Two LCMS methods were used for library mixture PAMPA and sequencing, one for all hexamer-related runs and one for all heptamer-related runs. Both runs were 53 min long with 0.5 mL/min flow rate with the concentration of ACN varied over time. Hexamer-related runs

started at 45% ACN for 2 min followed by a 30 min gradient from 45% to 75%, then a 15 min interval at 100% ACN followed by 6 min at 45% ACN (45% 0-2 min, linear increase to 75% 2-32 min, 100% 32-47 min, 45% 47-53 min). Heptamer-related runs were identical save for a 50% ACN concentration at the start and end of each run and a maximum of 70% ACN reached in the gradient section (50% 0-2 min, linear increase to 70% 2-32 min, 100% 32-47 min, 50% 47-53 min).

Tandem MS methods were created from the “Nth Order Double Play” template directed to collect the five most intense non-isotopic peaks for MS^2 acquisition in the ion trap after each MS^1 acquisition by the FTMS. The MS^2 activation was by collision-induced disassociation with a normalized collision energy of 35.0, isolation width of 1.0 m/z, activation Q. of 0.250, and activation time of 10.0 ms. The method was set to automatically detect and avoid analysis of sodium adducts.

Final purity checks and PAMPA quantitation for pure compounds was performed on the Orbitrap using a short 1.9 µm C18 column (Hypersil Gold 2.1x30 mm, Thermo Scientific) with a 7 min method for purity checks (with UV monitoring) and a 5 min method for PAMPA quantitation (without UV monitoring) (5% ACN 0-1 min, linear increase to 95% ACN 1-6 min or 1-4 min, 95% ACN for 1 min).

Proton spectra were acquired on a Bruker 500 MHz NMR with an Avance III HD console and a 5 mm BBO smart probe. All spectra were acquired at 20° C.

2.7.2 Reagent Synthesis

All reagents synthesis is described in chapter one.

2.7.3 Cyclic Hexa- and Heptapeptomer Library Synthesis

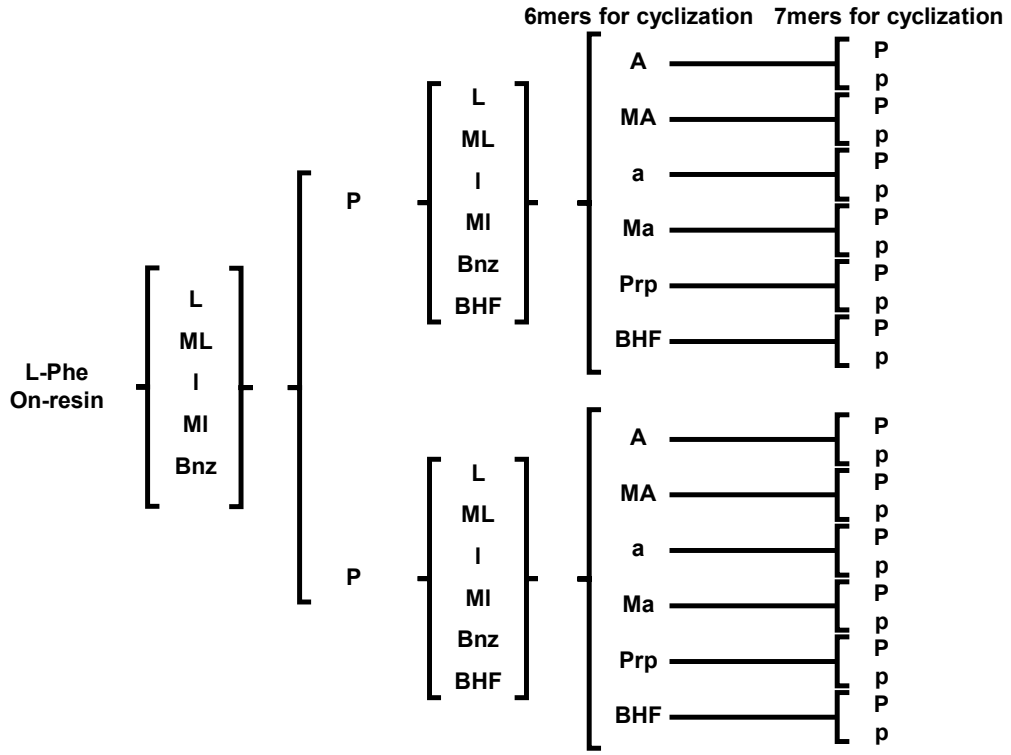


Figure S2.5 OBOC library synthetic scheme. Residues represented by their one-letter codes where possible, with capitalized letters representing L stereochemistry and lowercase representing D stereochemistry. A preceding M indicates an N-methyl residue. Propyl peptoid, benzyl peptoid, and L-β-hPhe have been abbreviated as Prp, Bnz, and BHF, respectively.

Linear peptomers were synthesized on L-phenylalanine 2-chlorotrityl resin (0.14 mmol/g) using extended Fmoc coupling (Fmoc amino acid/HATU/DIPEA in DMF, overnight) and peptoid synthesis conditions (bromoacetic acid/DIC in DMF, 1 h, then amine in DMF, overnight). Synthesis proceeded in a OBOC fashion, with the resin split and pooled at the second, fourth, and fifth residues added. Resin was split for addition of the third and sixth residues and kept separate for the remained of the synthesis in each case, resulting in 12 mixtures at that point. The resin was then split again, with half the resin destined to become the hexamer library (retained for cyclization) and half the heptamer library (proceeding to addition of a seventh residue). Each of the 12 heptamer-destined mixtures was again split for addition of the seventh residue, resulting in a final count of 12 hexamer sub-libraries and 24 heptamer sub-libraries of 150 theoretical compounds each. The linear peptomers were

cleaved from resin with 30% HFIP in DCM. Cyclization was performed in dilute conditions (<3 mM peptomer in a solution of 1:1 ACN:THF) with COMU (2 molar equivalents) and DIPEA (10 molar equivalents) and stirred overnight at room temperature. Each sub-library was briefly purified using Isolute 103 SPE cartridge (200mg/6ml, Biotage).

All individual synthetic steps were identical to those performed in chapter one and will not be repeated here.

2.7.4 Resynthesis of Individual Compounds and their Purification

Synthesis and characterization of the hexamer library and resynthesized compounds 2.1, 2.2, 2.3, 2.4, 2.13, 2.14, and 2.16 were previously reported as compound numbers 1.15, 1.22, 1.8, 1.14, 1.13, 1.19, and 1.20 respectively in chapter one. Compound numbering without a “1.#” or “2.#” prefix refers to the compound numbering system originating from the publication this chapter was based on. Some information from chapter one is repeated here, with alterations as needed, for convenience and completeness.

The sequence names of individual compounds have been abbreviated as much as possible in Table S2.6 below and in all other references by sequence to the resynthesized compounds. One-letter residue names are used where possible, with an uppercase letter representing L stereochemistry and a lowercase letter representing D stereochemistry. There are no methionine residues in these libraries, so a preceding capital M represents an N-methylated residue unambiguously. Propyl peptoid is abbreviated as Prp and benzyl peptoid as Bnz. L-β-hPhe is abbreviated as BHF. Each residue name is separated by a comma.

Table S2.6. Resynthesized compound synthetic details

ID #	Sequence	Synthetic Scale	Purification LC Method (see 2.7.1)
2.1	a,l,l,p,ML,F	0.05 mmol	Biotage
2.2	BHF,L,L,P,l,F	0.05 mmol	Biotage
2.3	a,L,L,p,l,F	0.05 mmol	Biotage
2.4	A,L,L,P,Ml,F	0.05 mmol	Biotage
2.5	p,A,l,Ml,p,Ml,F	0.2 mmol	Waters (40%-100% ACN gradient)
2.6	P,Prp,L,Ml,P,Ml,F	0.2 mmol	Waters (52% ACN isocratic)
2.7	P,A,Ml,Ml,p,ML,F	0.2 mmol	Waters (53% ACN isocratic)
2.8	p,Ma,L,ML,P,ML,F	0.2 mmol	Waters (40%-100% ACN gradient)
2.9	P,MA,L,ML,p,ML,F	0.2 mmol	Waters (48% ACN isocratic)
2.10	Prp,Ml,L,p,ML,F	0.2 mmol	Waters (50%-100% ACN gradient)
2.11	Prp,l,Ml,P,l,F	0.2 mmol	Waters (50%-100% ACN gradient)
2.12	P,Ma,L,ML,P,ML,F	0.2 mmol	Waters (50% ACN isocratic)
2.13	Ma,l,L,p,l,F	0.05 mmol	Biotage
2.14	BHF,L,l,P,L,F	0.05 mmol	Biotage
2.15	p,MA,L,Ml,P,Ml,F	0.2 mmol	Waters (53% ACN isocratic)
2.16	BHF,l,L,p,l,F	0.05 mmol	Biotage
2.17	p,Ma,l,Ml,P,Ml,F	0.2 mmol	Waters (50% ACN isocratic)
2.18	p,MA,L,ML,p,ML,F	0.2 mmol	Waters (45%-100% ACN gradient)
2.19	p,A,Ml,Ml,p,ML,F	0.2 mmol	Waters (45%-100% ACN gradient)

2.7.4.1 Synthesis Overview

All peptides were resynthesized on an automated peptide synthesizer (Prelude X, Gyros Protein Technologies) without isotopic labelling. The synthesis occurred on L-phenylalanine loaded 2-chlorotrityl resin (0.8 mmol/g, Rapp Polymere) at varying scales (Table S2.6.). Each coupling proceeded as in chapter one (1.7.5.2), with cleavage occurring on the Prelude X also as described in chapter one. Cyclization proceeded, as in chapter one, entirely in ACN rather than 1:1 ACN:THF as with the library synthesis.

The three peptoid-containing compounds were removed from the peptide synthesizer for manual peptoid addition as the library synthesis procedure but with 10 molar equivalents bromoacetic acid, 5 molar equivalents DIC, and 10 molar equivalents of amine to account for the comparatively higher loading value of the resin. The single peptoid-containing heptamer had its final proline residue added manually as the library amino acid coupling procedure but with 4 molar equivalents of Fmoc amino acid, 3.8 molar equivalents of HATU, and 6 molar equivalents of DIPEA. They were then cleaved on the Prelude X as with the other peptides.

All cyclized peptides/peptomers were then purified and characterized by LCMS and proton NMR.

2.7.4.2 Biotage Purification

Dry cyclized peptide was dissolved in a minimal volume of DMF (1 mL or less) and loaded onto the SNAP cartridge after equilibration. Collection was controlled by a UV threshold or manually triggered. The method was run as described in section 2.7.1 and the collected fractions were analyzed on the Waters LC-MS system (also as described in 2.7.1), pure fractions combined, and evaporated to a solid or oil before transfer to a tared vial.

2.7.4.3 Waters Purification

Up to 60 mg of crude cyclized peptide was dissolved in DMSO (300 µL or less) and placed into the Waters LC-MS autosampler, then injected onto the prep column with an individualized method as Table S2.6. Fraction collection was triggered by UV or manually and fractions were analyzed by direct inject to an Advion ExpressIon MS. Pure fractions were combined, evaporated to a solid or oil, and transferred to a tared vial.

2.7.5 Assay Details and Data Analysis

2.7.5.1 PAMPA Assay

PAMPA was performed and analyzed as in Naylor et al. from 2017[48] with a few differences introduced by sink conditions[57] and in the concentration of analytes. All other details were consistent. The donor well solution was prepared with the addition of 0.2% polysorbate-80 and the acceptor solution had 0.2% TGPS added. Sample preparation was complicated by the need to equalize the polymer presence in the final donor and acceptor solutions. After plating the 100 µL each from the donor and acceptor wells of the experiment, 100 µL of ACN and 100 µL of PBS buffer with 5% DMSO and either 0.2% polysorbate-80 or 0.2% TPGS was added to each well such that both polymers were present in each sample before quantitation on the Velos Pro Orbitrap LC-MS system.

PAMPA on mixtures was performed at 500μM total concentration, resulting in a maximum theoretical 3.3 μM concentration for each putative peptide. 1NMe3[46] was added as an internal standard at 1 μM (exact mass 754.50, 1.98 ± 0.14 * 10^{-6} cm/s under these conditions). Where there was insufficient signal (heptamer sub-libraries DDAL, DLMAL), injection volumes were increased from 5 μL to 20 μL. Our initial PAMPA data set showed a strong correlation between the pure resynthesized compound permeabilities and the library permeabilities. Combined with the low deviation observed in section 2.5.1 for PAMPA on the same mixture at varying concentrations, we found it unnecessary to perform further replicates of the library data.

Pure compounds were assayed at a concentration of 10 μM in the donor well with four replicates to establish a standard deviation after any bad wells were removed. Carbamazepine was used as an internal standard at 10 μM.

2.7.5.2 Automated Data Analysis Script Usage

All raw data was converted to mzML format using Proteowizard 3.0.10577 as in chapter one using the command:

```
msconvert -v log.txt --simAsSpectra *.RAW --32 --zlib --filter
"peakPicking true 1-" --filter "zeroSamples removeExtra"
```

PAMPA data was processed automatically to ease the burden of processing PAMPA of large mixtures. Automated PAMPA data processing was performed using the command:

```
python AutoPAMPA.py -g -u jobfile.xlsx
```

This command was used once for each full PAMPA experiment, with the specific masses to monitor in each well and the PAMPA equation parameters listed in the job spreadsheet. Additional parameters modifying expected MZ precision, peak and peak bound detection, smoothing behavior, and expected retention time window were also optimized to improve peak detection, fitting, and integration. Further details are can be found in Appendix A.

All sub-libraries and resynthesized compounds were sequenced using an updated version of CycLS (only data output format and exploratory analysis features differ) with a command like:

```
python CycLS.py -o mixture_name * mixture_name.mzML
p;BHF;l,L***,Ml,ML***,LA08;l,L***,Ml,ML***,LA08,BHF;P;l,L***,Ml,ML***,LA08;F
```

Where the long constraint string of residue names allowed at each position is specific to the potential sequences in each mixture (or each mixture of origin for resynthesized peptides).

As noted in chapter one, CycLS generates a score for each set of MS^2 spectra grouped together and a sequencing confidence metric (referred to as the “normalized score difference” in chapter one). The score has been observed to trend with peak intensity within a set of compositionally identical compounds differing only by sequence, likely due to an increased count for rare fragments. When those fragments are critical in differentiating between sequences, the sequencing confidence is also increased for such peaks. Thus, an intensity-based tandem MS collection method is sensible. Only the top-scoring sequence and accompanying statistics were used for downstream analysis.

Given that all data was acquired in identical LC conditions, PAMPA and sequencing data could be merged by retention time. An in-house python script (RTMerge.py) was used to align individual peaks a command like this:

```
python RTmerge.py -o output_prefix config.xlsx
```

The resulting spreadsheet containing the merged permeability and sequencing data was then filtered and curated to arrive at the final set of sequence-permeability pairs used for the analyses presented earlier in the chapter.

2.7.5.3 Data Filtration and Curation to Ensure Quality

The merged data was subjected to sequential filtering steps to ensure high data quality and sequencing accuracy before further analysis. The first two filters were on sequencing statistics generated by CycLS. Any sequence with a raw score of less than 2.0

was discarded; low raw scores can indicate a non-peptidic spectrum, that all matches were to noise, or that key fragments were not captured. The next filter was based on the normalized score difference between the top scoring sequence for that spectrum and the second-best score (referred to as sequencing confidence above). The threshold was set low, at 0.005, because true sequences begin to be discarded at any higher threshold. All pure resynthesized compounds had a normalized difference of at least 0.1, however. These filters discarded low-quality or non-peptidic MS^2 spectra that happened to align with a relevant peak and the most ambiguous sequencing results.

The added polymers from the “double-sink” condition PAMPA shared a mass range with the heptamer library. This interference was generally low in intensity but still necessitated increased caution in accepting PAMPA data for peaks with low signal intensity. We therefore implemented a minimum integrated intensity requirement of eight million in the donor well and marked all peaks with an integrated intensity of less than one million in the acceptor well for later manual curation.

We encountered many instances of peaks with duplicate top-scoring sequences. We resolved each instance of duplicate top sequences by preserving only the peak with the best sequencing quality statistics. Both raw sequencing score and sequencing confidence were considered in choosing the best peak available for each sequence, with priority given to sequencing confidence only when raw scores were similar between candidates. The duplicate with the highest raw score was chosen to be preserved unless the duplicate with the second highest raw score was within two raw score units and had a higher sequencing confidence metric. In that case, the improvement in sequencing confidence was required to be distance*10% higher for the duplicate with the lower raw score to be preserved. For example, candidate A has a raw score of 12 and candidate B has a raw score of 10.5, a difference of 1.5 units. Candidate B has a sequencing confidence metric of .25 to candidate A’s 0.2. The difference between confidence metrics is 0.05, which is greater than the required 15% improvement (distance*10%), so candidate B is chosen to represent that sequence.

As a final quality assurance, a limited manual curation of the data was performed. First, all impermeable peaks and all heptamer peaks with a low acceptor well intensity were inspected. A second holistic inspection was performed on each sub-library focused on peaks with outlier permeabilities for their retention time. Over 800 peaks were inspected during this process, resulting in the removal of 88 hexamers and 119 heptamers. Peaks were mainly removed for incorrect integration in either the donor or acceptor well due to peak overlap, overlap with sink PAMPA polymer signal, or other events causing poor signal. A small number of peaks were falsely measured as impermeable due to a failure in peak alignment due to dramatically differing peak shapes between wells or shifts in retention time outside of the allowed tolerances. Data with greater than 95% T was also removed due to the high sensitivity of the permeation rate equation to error in that regime. The overall data attrition can be seen in figure S2.6.

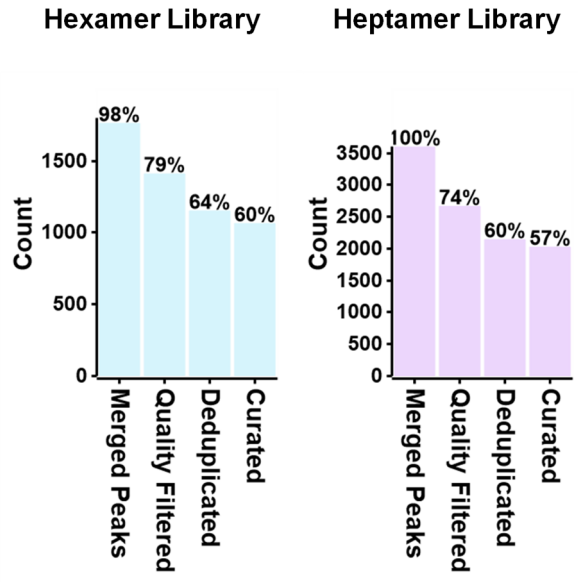


Figure S2.6 Data attrition through various quality filters.
Bar plots (and percentages) normalized to the theoretical library size for each of the hexamer and heptamer libraries describe the initial retention time matches between the sequencing and permeability data, and the count remaining after filtering for sequencing quality and peak integrity, deduplication of peaks with identical top sequences, and manual curation.

2.7.6 Characterization of Library Mixtures

Each sub-library (12 hexamer and 24 heptamer) was assessed by mass spectroscopy on the Velos Pro Orbitrap by examination of the TIC and extracted ion chromatograms for expected masses. While the expected masses for truncations were observed, they universally eluted much earlier and shared no masses with the libraries. The base peak extracted ion chromatograms [500-1000 m/z range] of each heptamer sub-library's tandem MS acquisition run can be found below (Figures S2.7.1 through S2.7.12). These runs did not contain sink polymer and therefore were not extended to 53 min to allow for elution of polymer as with the PAMPA analysis runs. Retention time matching was not affected by this difference because the runs were identical up to the extension. Libraries began eluting between 5 and 10 min into the method and ceased eluting between 20 and 32 min. Hexamer sub-libraries were characterized in the CycLS paper and can be found in that SI document.

One sub-library of each library was analyzed peak-by-peak as a representative for that library. The hexamer sub-library labeled DDMA was analyzed in brief in the CycLS paper and a more detailed analysis of the heptamer sub-library labeled LLAL can be found below. Table S2.7 lists each expected mass for sub-library LLAL and the number of compounds expected, and major peaks observed at each mass. Extra peaks, where observed, are categorized by their identity as interfering peaks from another expected mass (sodium adducts and C^{13} envelope interference) or unknown. In some cases, expected peaks were not observed, either because they did not synthesize, did not ionize, or are overlapped by another peak. Some individual cases require more explanation, as indicated by a superscripted letter.

Table S2.7. Synthetic validation of heptamer sub-library LLAL

Library Mass	Expected Peaks	Observed Peaks[a]	Peaks Absent[b]	Sodium Adducts (source)	C^{13} Envelope Interference	Unknown
751.463	1	2				1
754.482	3	5				2
757.501	3	3				
760.520	1	1				
765.479	3	3				
768.498	9	12				3
771.517	9	11				2
774.535	3	4				1
779.495	3	8		3 (757.501)		2
782.513	9	15				6
785.448	3	4				1
785.532	9	10				1
788.466	6	9		3 (765.479)		
788.551	3	4			1	
791.485	3	13	2	12 (768.498)		
793.510[c]	1	10		9 (771.517)		
796.529[d]	3	7		1 (774.535)		3
799.463	7	9				2
799.548	3	3				
802.482[e]	14	20		6 (779.495)		
802.567	1	2			1	
805.501[f]	7	18	1	12 (782.513)	1	
813.479	5	6		1 (791.485)		
816.498	10	11		1 (793.510)		
819.432	3	4				1
819.517	5	14		6 (796.529)	2	1
822.451[g]	3	10	1	8 (799.463)		
827.495	1	8		6 (805.501)		1
830.513	2	2				
833.448	5	5				
833.532	1	1				
836.466	5	11		5 (813.479)		1
847.463	2	2				
850.482	2	4		2 (827.495)		
853.416	1	2				1
867.432	1	1				

[a]Major peaks only, though sometimes this designation was used generously when there was no obvious threshold. Contaminant peaks from plasticizers and similar are not included.

[b]Number of peaks below the expected count after [M+Na] ions and envelope interference peaks (and overlaps) are accounted for. Missing peaks may not have synthesized or may not be visible due to overlapping peaks of higher intensity originating from other masses. 793.501[c]: The expected number of [M+Na] peaks originating from mass 771.517 are all present, but some highly overlapped peaks were more difficult to differentiate. 796.529[d]: The expected number of [M+Na] peaks originating from mass 774.535 were observed but were low enough in intensity for only one to be categorized a major peak. 802.482[e]: A peak considered minor in mass 779.495 was judged significant as an [M+Na] peak in mass 802.482 due to a peak overlap. 805.501[f]: Two [M+H] peaks were rendered unusable by an overlapping [M+Na] peak, two [M+Na] peaks were overpowered by [M+H] signal, and two expected [M+Na] peaks at nearly identical retention times could not be differentiated at this mass (too overlapped). This accounts for the 3 missing [M+Na] peaks from mass 782.513. 822.451[g]: One [M+Na] peak originating from mass 799.463 was overpowered by an [M+H] peak from mass 822.451.

Of 150 expected peaks, 146 were observed with an additional 108 peaks present from various sources. The most common cause of extra peaks was interference of sodium adducts from other expected masses (74/108). An internal calibrant was not used, so mass resolution was not high enough to distinguish the sodium adducts by their mass difference alone. The tandem MS method was set to avoid collection of MS^2 data from sodium adducts, but the minimum isolation window width could not distinguish between an expected mass and a nearby sodium adduct where there was chromatographic overlap between the two. Such cases universally led to poor sequencing results compared to other peaks of the same mass unless the expected mass was far more intense than the overlapped sodium adduct. We believe this is true both because sodium adducts lack fragments from the expected sequences for that mass and because sodium adducts fragment less efficiently than protonated species under the ionization conditions used.

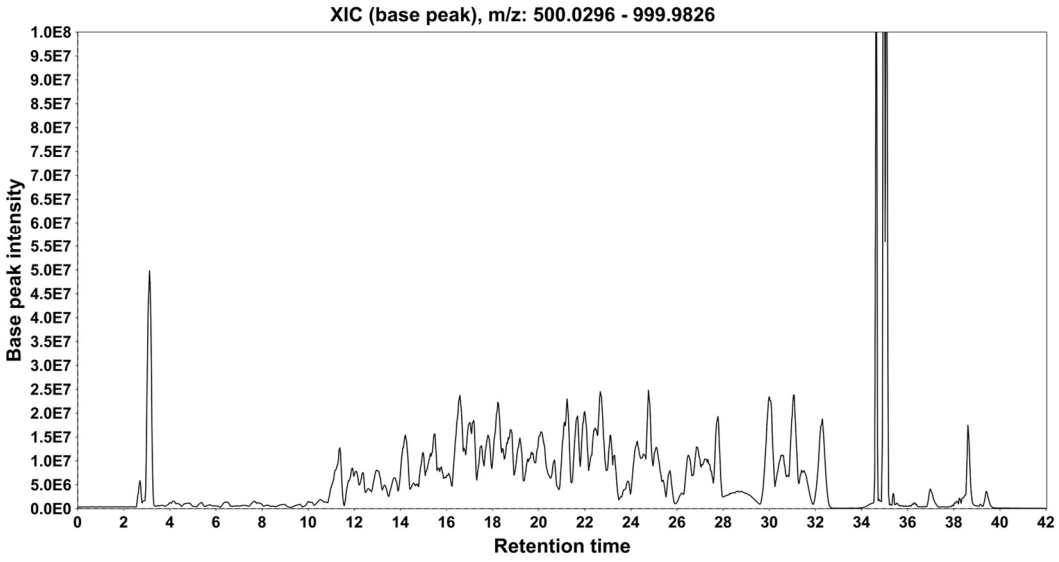
In rare cases, extra peaks originated from the carbon-13 envelope of nearby masses which were especially intense in comparison (5/108). Other cases (29/108) are hypothesized to be slow-exchanging conformers or minor epimers. Minor epimerization was observed in some resynthesized compounds, supporting an expectation of epimers in the libraries. Such peaks, where sequenced, either exhibited obvious differences in their tandem MS spectra from all other peaks – marking them as epimers – or had spectra nearly or completely identical to other peaks. Most such peaks had low intensity and were consequently not sequenced.

Many instances of peak overlap between isobaric peaks occurred. However, in most cases the overlap did not prevent correct assignment of sequencing data or integration. Loss of a potential library signal (either sequencing or PAMPA) to chromatographic overlap of multiple major peaks was observed 17 times in this sub-library (6 caused by sodium adduct peaks) and inaccurate automated integration of especially broad peaks caused unusable PAMPA data in 4 further instances.

Peaks were only classified as missing if they were impossible to locate both as [M+H] and [M+Na] even if their overlap with other peaks would make them effectively unusable. Only four expected peaks were missing in this analysis and 11 missing in the hexamer sub-library DDMA analysis. The rarity of missing peaks suggests that synthesis of nearly all library compounds occurred in sufficient quantity for detection. However, many peaks had low enough intensity that signal in the PAMPA acceptor well was insufficient for a meaningful integration. This issue was exacerbated by the baseline signal caused by the addition of sink polymers, particularly in the heptamer library's expected mass range as noted in section 3.3. Of 149 peaks with associated PAMPA and sequencing data, 11 peaks did not meet the minimum donor well intensity threshold and 16 did not meet minimum sequencing quality criteria, resulting in 124 peaks recovered. Sequence deduplication and manual curation resulted in a final total of 87 data points of a theoretical 150.

Heptamer Sublibrary DBD

D-Proline, L-β-hPhe, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DBL

L-Proline, L-β-hPhe, X_3, X_4, D-Proline, X_6, L-Phe

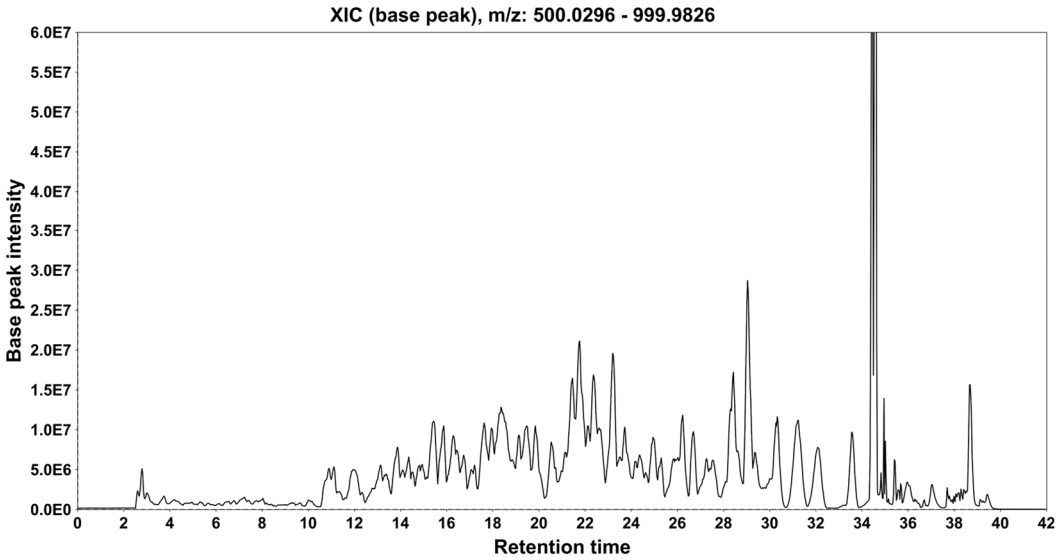
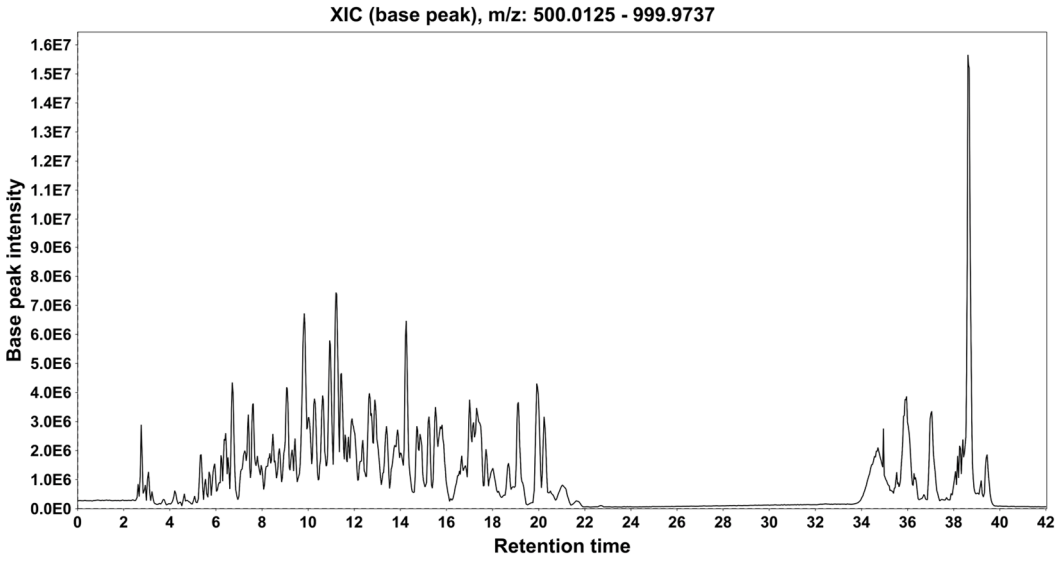


Figure S2.7.1 Extracted ion chromatograms of sub-libraries DBD and DBL.

Heptamer Sublibrary DDAD

D-Proline, D-Ala, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DDAL

L-Proline, D-Ala, X_3, X_4, D-Proline, X_6, L-Phe

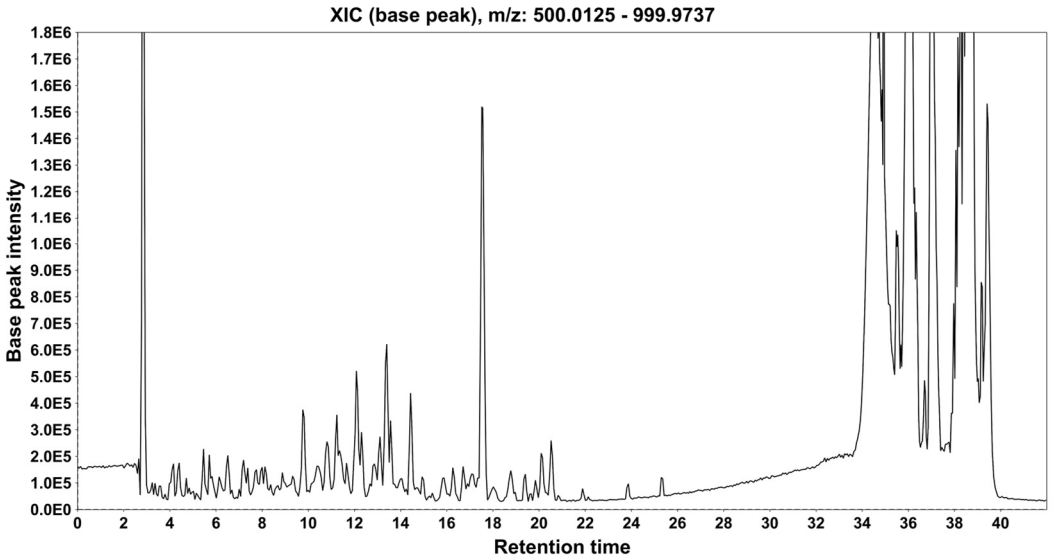
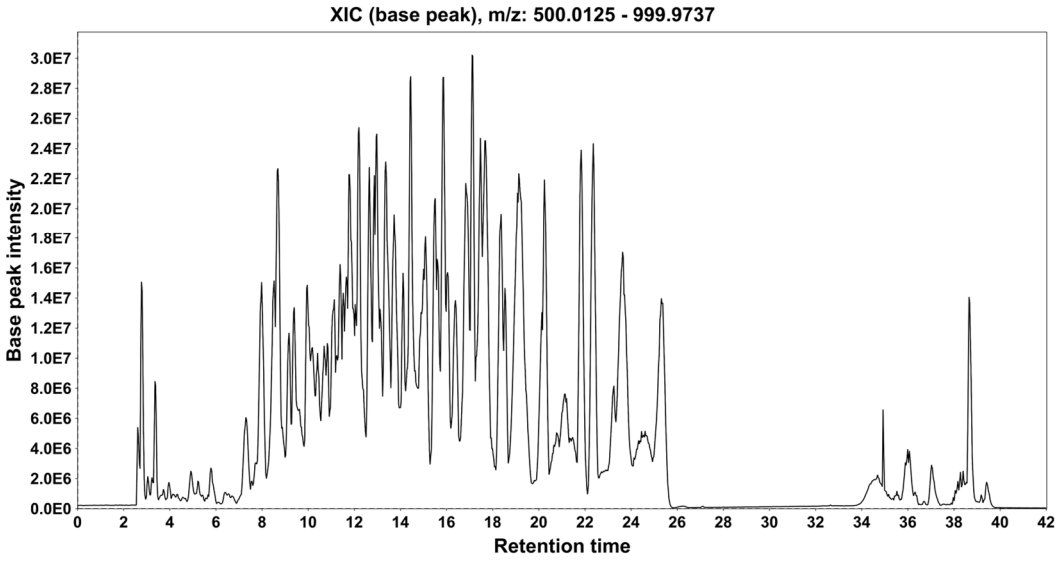


Figure S2.7.2 Extracted ion chromatograms of sub-libraries DDAD and DDAL.

Heptamer Sublibrary DDMAD

D-Proline, D-NMe-Ala, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DDMAL

L-Proline, D-NMe-Ala, X_3, X_4, D-Proline, X_6, L-Phe

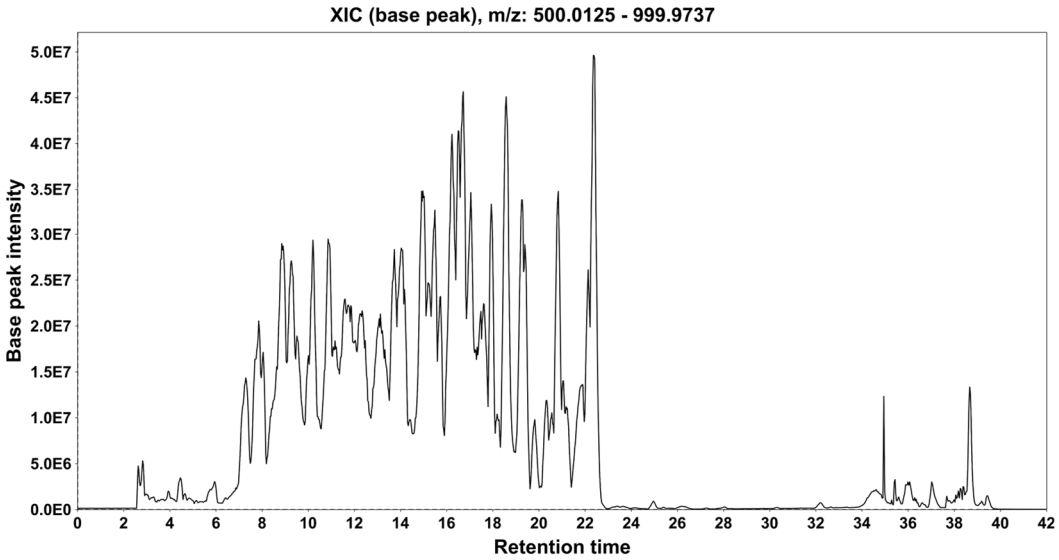
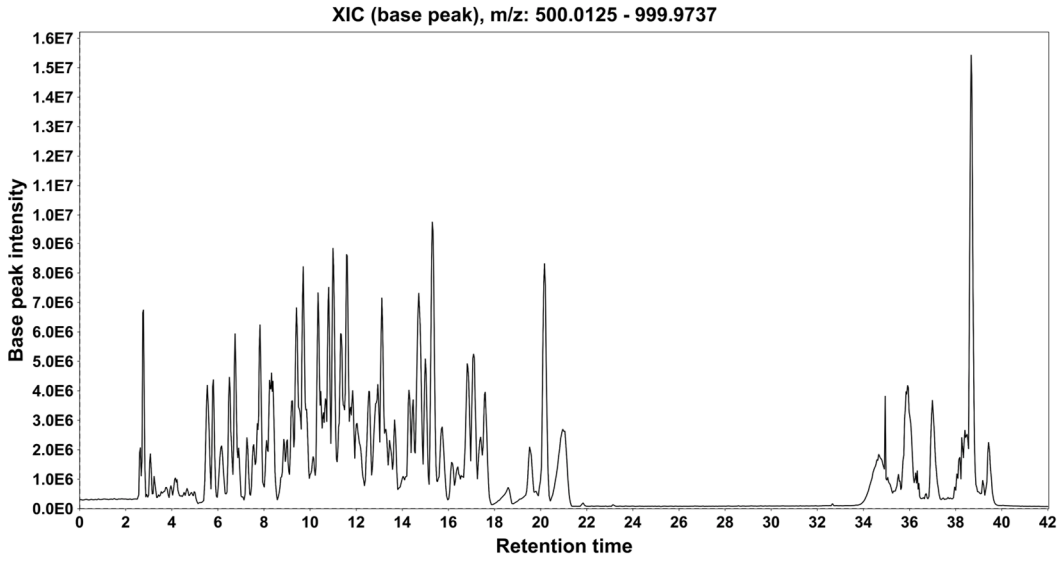


Figure S2.7.3 Extracted ion chromatograms of sub-libraries DDMAD and DDMAL.

Heptamer Sublibrary DLAD

D-Proline, L-Ala, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DLAL

L-Proline, L-Ala, X_3, X_4, D-Proline, X_6, L-Phe

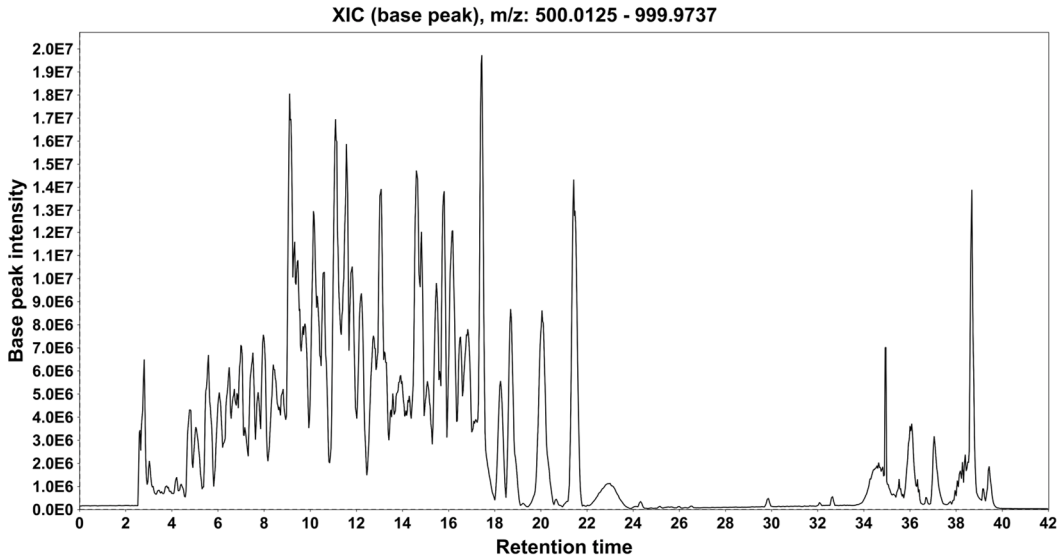
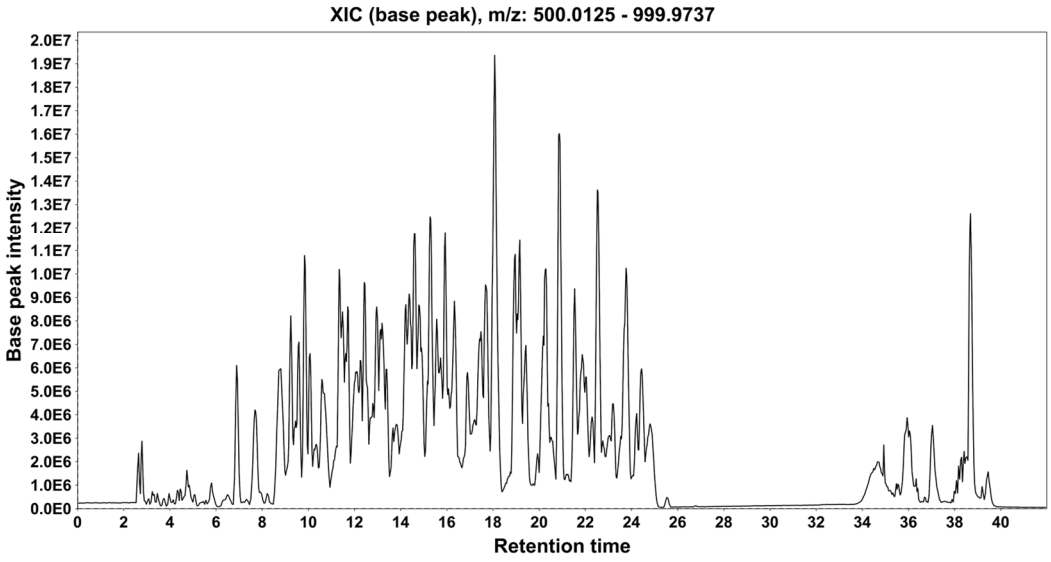


Figure S2.7.4 Extracted ion chromatograms of sub-libraries DLAD and DLAL.

Heptamer Sublibrary DLMAD

D-Proline, L-NMe-Ala, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DLMAL

L-Proline, L-NMe-Ala, X_3, X_4, D-Proline, X_6, L-Phe

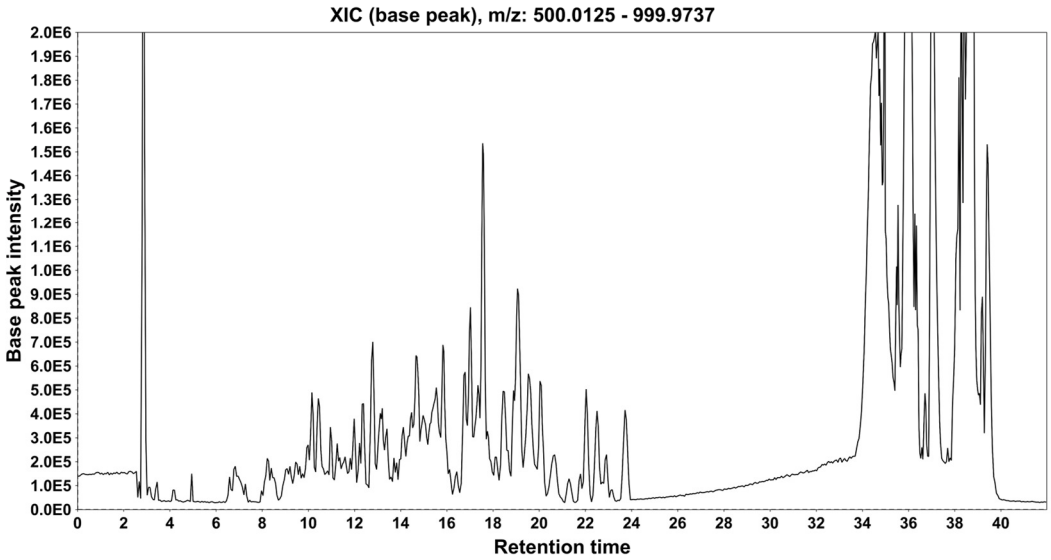
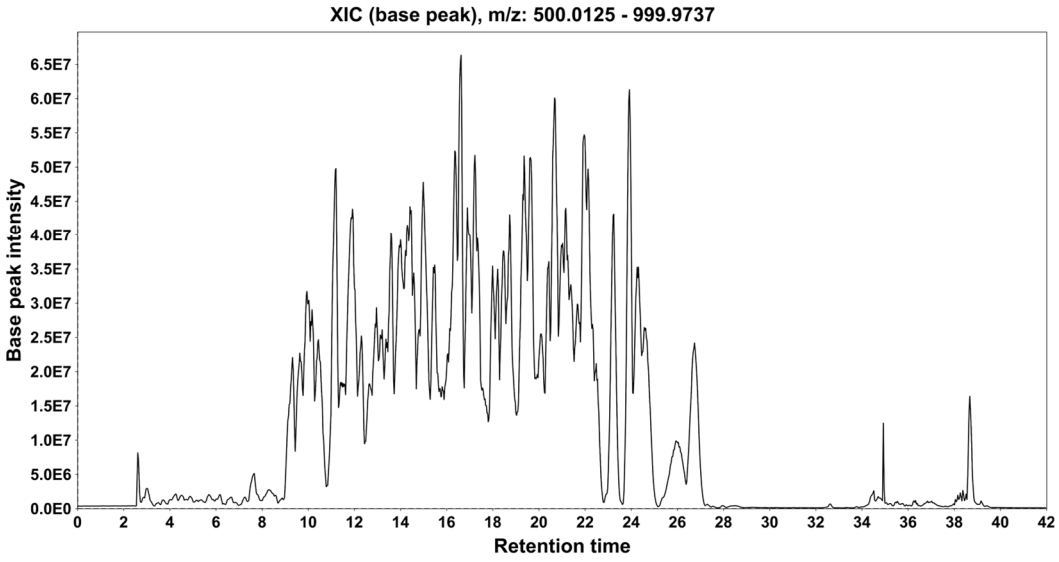


Figure S2.7.5 Extracted ion chromatograms of sub-libraries DLMAD and DLMAL.

Heptamer Sublibrary DPD

D-Proline, Propyl peptoid, X_3, X_4, D-Proline, X_6, L-Phe



Heptamer Sublibrary DPL

L-Proline, Propyl peptoid, X_3, X_4, D-Proline, X_6, L-Phe

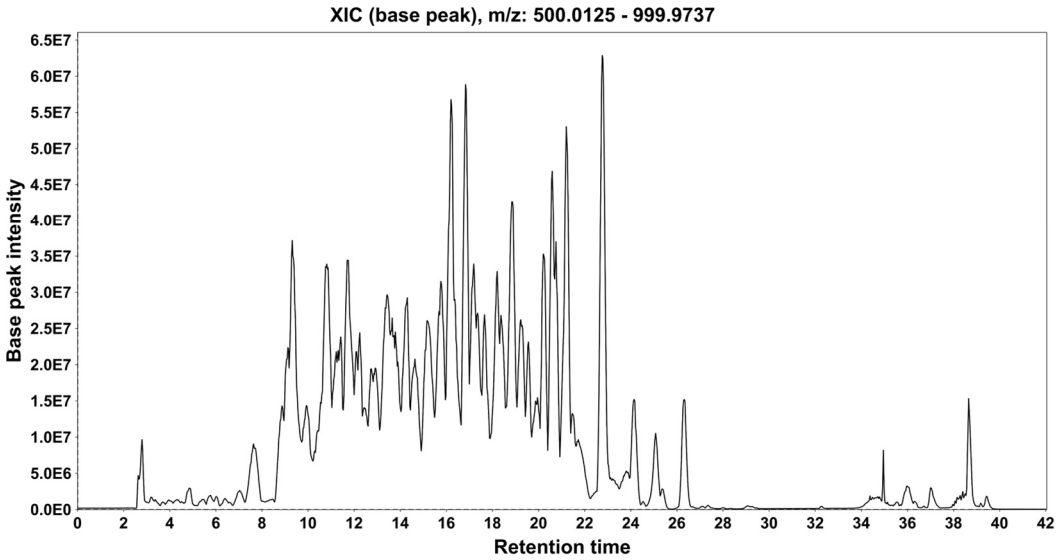
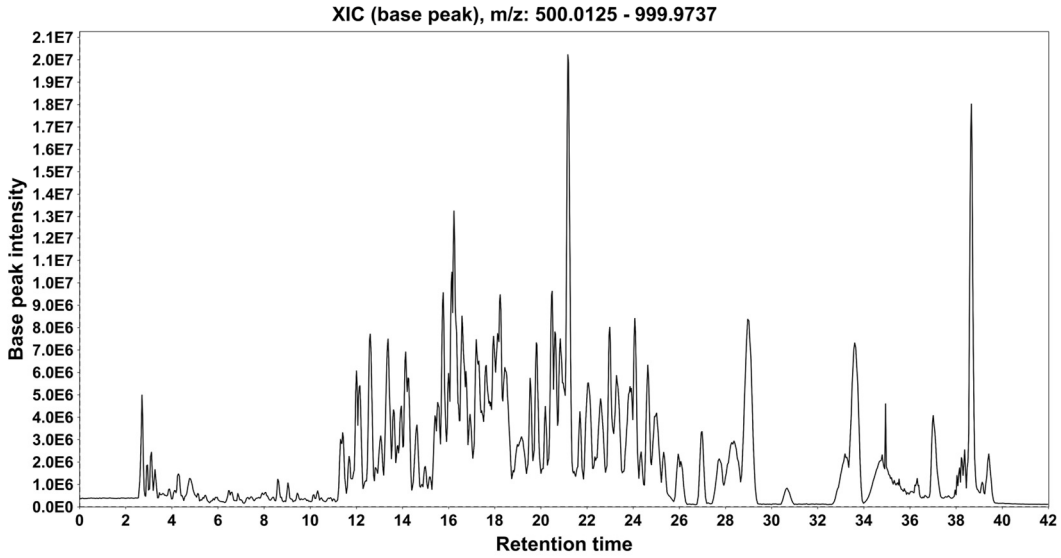


Figure S2.7.6 Extracted ion chromatograms of sub-libraries DPD and DPL.

Heptamer Sublibrary LBD

D-Proline, L-β-hPhe, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LBL

L-Proline, L-β-hPhe, X_3, X_4, L-Proline, X_6, L-Phe

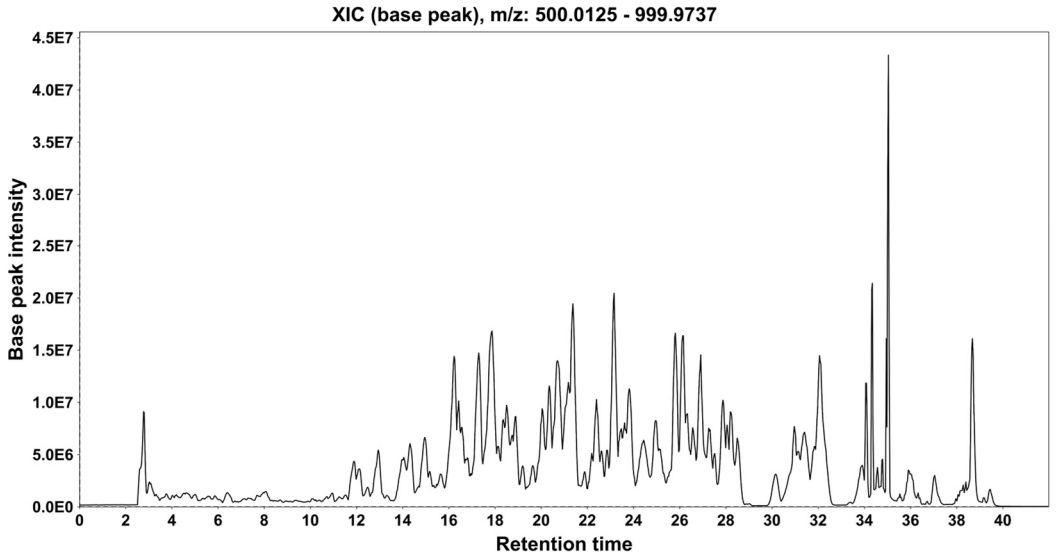
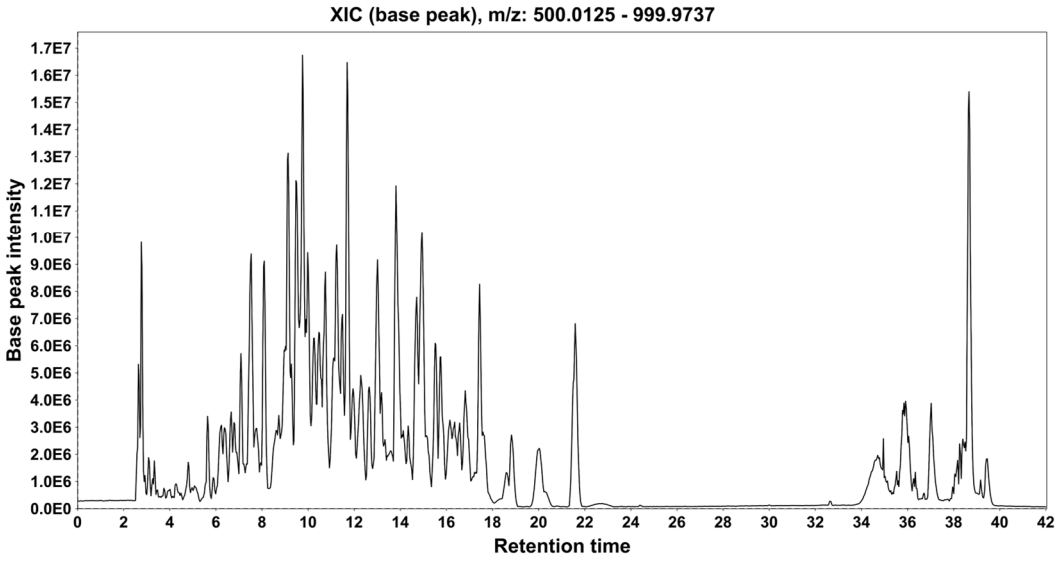


Figure S2.7.7 Extracted ion chromatograms of sub-libraries LBD and LBL.

Heptamer Sublibrary LDAD

D-Proline, D-Ala, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LDAL

L-Proline, D-Ala, X_3, X_4, L-Proline, X_6, L-Phe

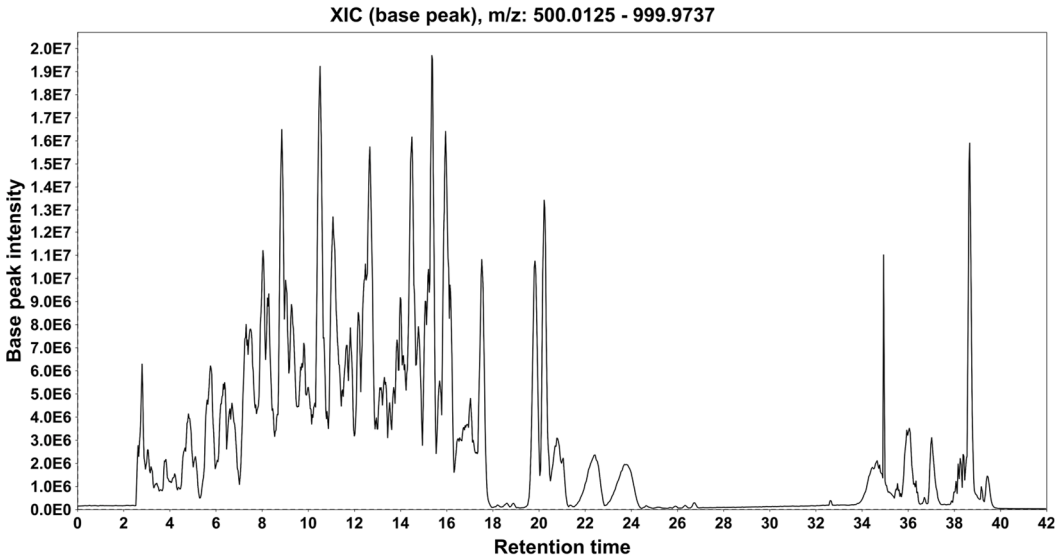
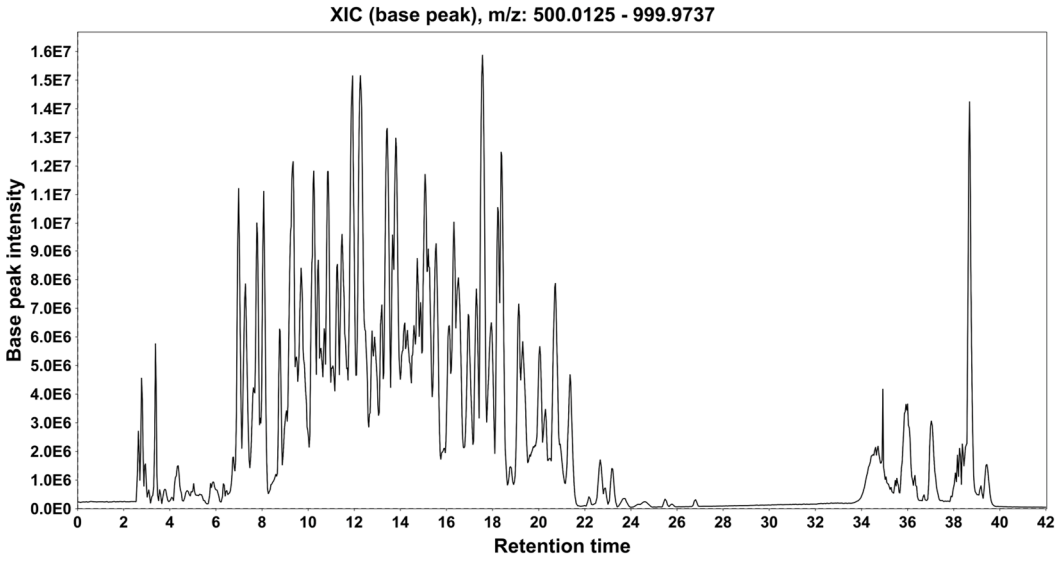


Figure S2.7.8 Extracted ion chromatograms of sub-libraries LDAD and LDAL.

Heptamer Sublibrary LDMAD

D-Proline, D-NMe-Ala, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LDMAL

L-Proline, D-NMe-Ala, X_3, X_4, L-Proline, X_6, L-Phe

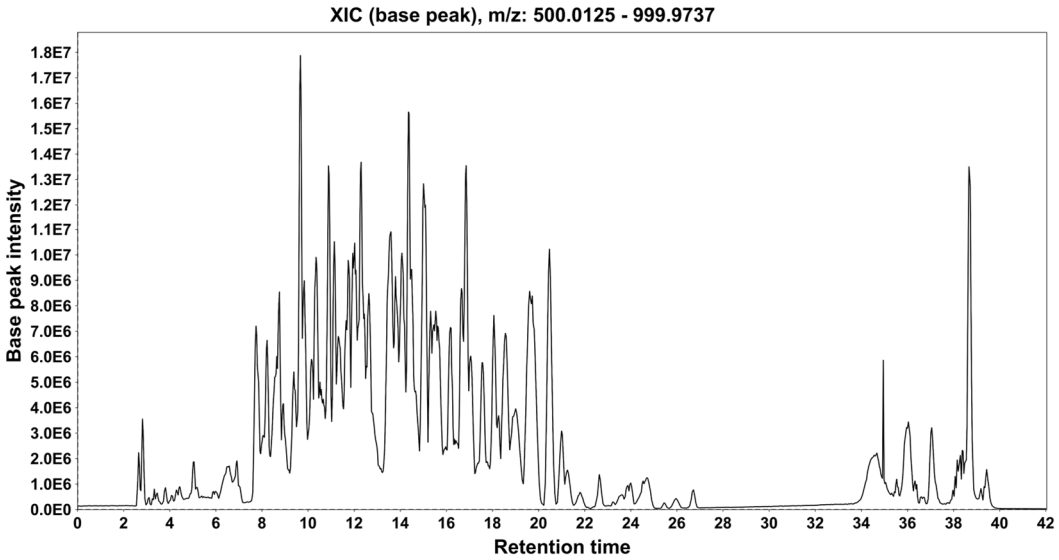
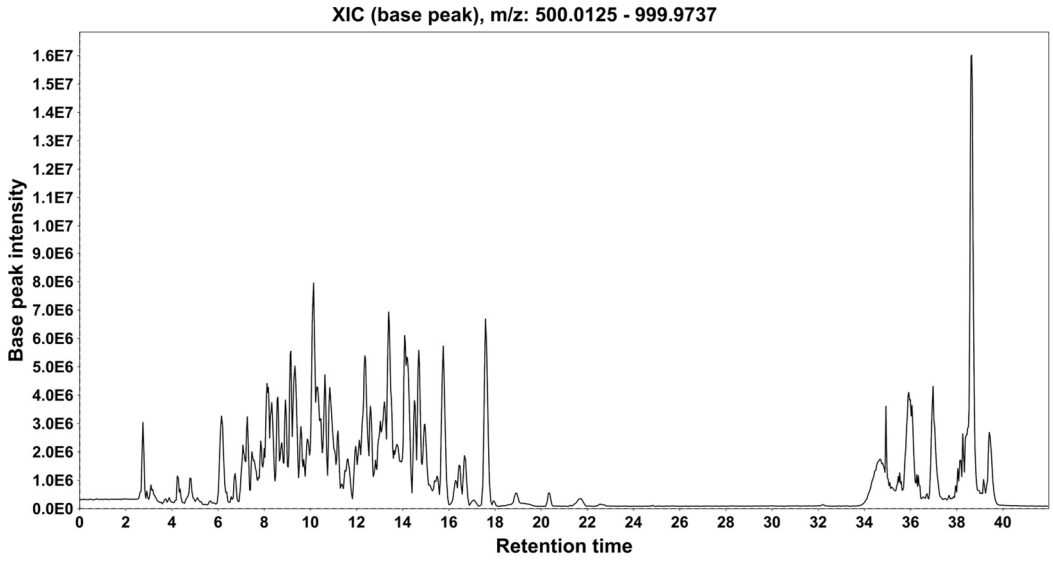


Figure S2.7.9 Extracted ion chromatograms of sub-libraries LDMAD and LDMAL.

Heptamer Sublibrary LLAD

D-Proline, L-Ala, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LLAL

L-Proline, L-Ala, X_3, X_4, L-Proline, X_6, L-Phe

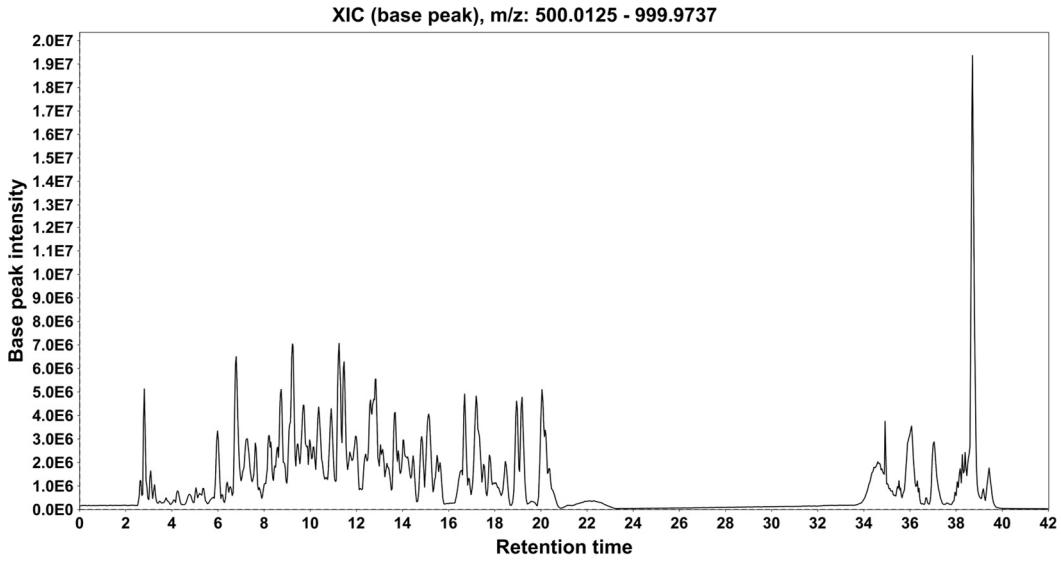
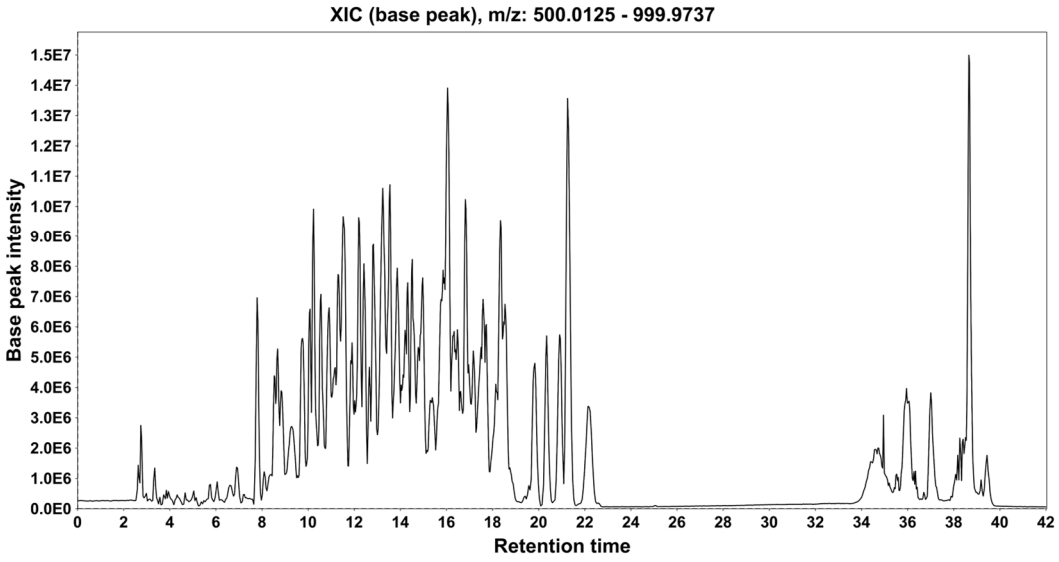


Figure S2.7.10 Extracted ion chromatograms of sub-libraries LLAD and LLAL.

Heptamer Sublibrary LLMAD

D-Proline, L-NMe-Ala, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LLMAL

L-Proline, L-NMe-Ala, X_3, X_4, L-Proline, X_6, L-Phe

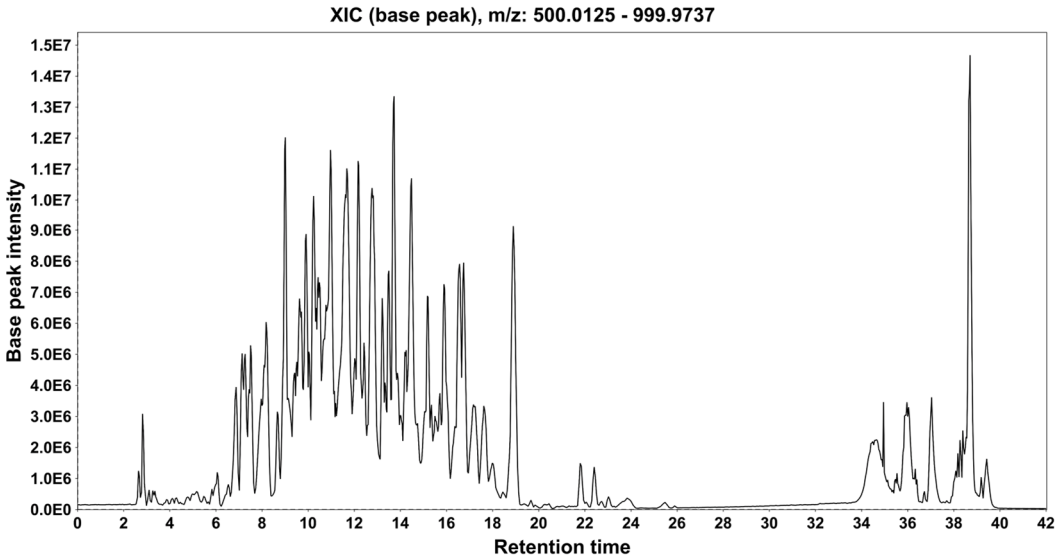
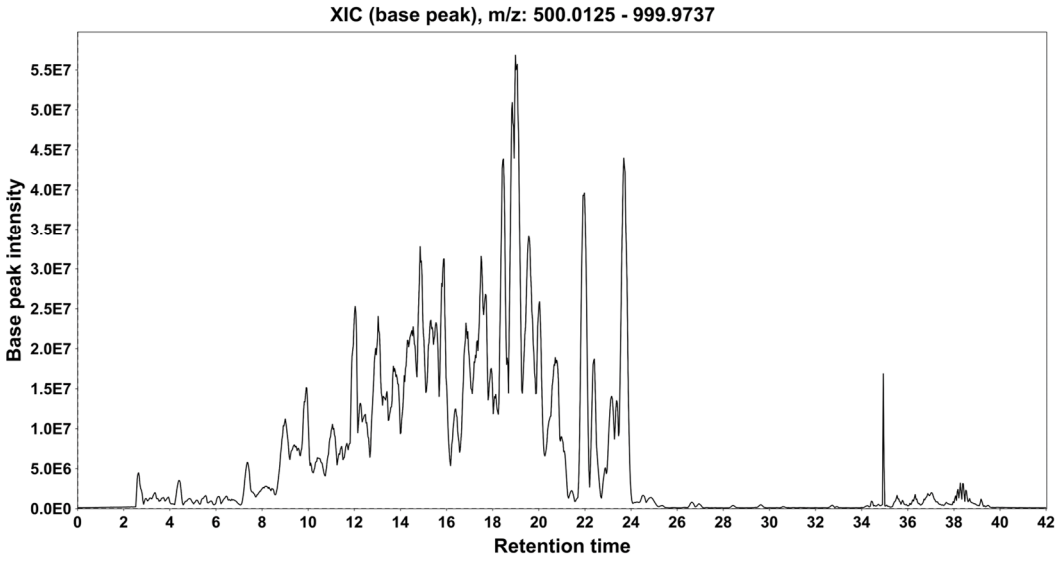


Figure S2.7.11 Extracted ion chromatograms of sub-libraries LLMAD and LLMAL.

Heptamer Sublibrary LPD

D-Proline, Prolyl peptoid, X_3, X_4, L-Proline, X_6, L-Phe



Heptamer Sublibrary LPL

L-Proline, Prolyl peptoid, X_3, X_4, L-Proline, X_6, L-Phe

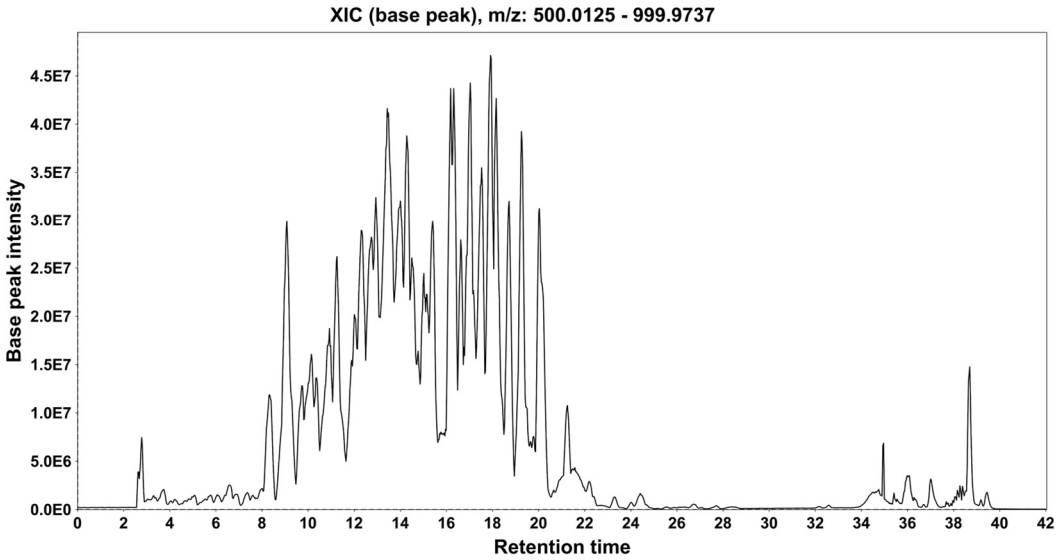


Figure S2.7.12 Extracted ion chromatograms of sub-libraries LPD and LPL.

2.7.7 Characterization and Sequencing Validation of Resynthesized Compounds

Table S2.8. Resynthesized compound PAMPA permeability values and corresponding library permeabilities

ID #	Sequence	Pure PeE^{-6} cm/s	Library PeE^{-6} cm/s
2.1	a,l,l,p,ML,F	0.0 ± 0.0	0.1
2.2	BHF,L,L,P,l,F	0.1 ± 0.2	0.5
2.3	a,L,L,p,l,F	0.2 ± 0.2	0.5
2.4	A,L,L,P,Ml,F	0.6 ± 0.6	0.7
2.5	p,A,l,Ml,p,Ml,F	0.8 ± 0.0	0.8
2.6	P,Prp,L,Ml,P,Ml,F	2.0 ± 0.1	2.7
2.7	P,A,Ml,Ml,p,ML,F	3.5 ± 0.3	2.7
2.8	p,Ma,L,ML,P,ML,F	4.1 ± 0.1	4.8
2.9	P,MA,L,ML,p,ML,F	4.3 ± 0.2	6.3
2.10	Prp,Ml,L,p,ML,F	5.0 ± 0.4	8.8
2.11	Prp,l,Ml,P,l,F	5.0 ± 0.3	11.1
2.12	P,Ma,L,ML,P,ML,F	5.4 ± 0.1	2.7
2.13	Ma,l,L,p,l,F	6.2 ± 0.2	5.4
2.14	BHF,L,l,P,L,F	6.3 ± 0.1	14.4
2.15	p,MA,L,Ml,P,Ml,F	7.9 ± 0.4	3.5
2.16	BHF,l,L,p,l,F	8.1 ± 0.7	9.2
2.17	p,Ma,l,Ml,P,Ml,F	8.8 ± 0.4	7.1
2.18	p,MA,L,ML,p,ML,F	10.7 ± 1.1	7.8
2.19	p,A,Ml,Ml,p,ML,F	10.9 ± 0.4	9.6

Pure compounds were characterized by mass and UV absorbance on the Velos Pro Orbitrap and by proton NMR spectra. Sequencing was validated by retention time matching and comparison of MS^2 spectra by eye and by ion listing. Tandem MS data were not expected to be completely identical due to the absence of isotopic labeling in the pure compounds, but correct sequencing results in obvious matching of ionization profiles. Crudes were used for sequencing validation to confirm the major peak as the desired product (true in all cases) and any minor peaks, when present, as epimers. Compounds 2.1, 2.2, 2.3, 2.4, 2.13, 2.14, and 2.16 had their sequencing confirmed in chapter one (compound numbers 1.15, 1.22, 1.8, 1.14, 1.13, 1.19, and 1.20, respectively) and only their characterization by MS and NMR are presented here. All compounds were validated to have sequenced correctly and all compounds were >95% pure by UV absorbance (except 2.2 and 2.14, >80% pure). Figures S2.8.1 through S2.8.19 abbreviate the full compound numbers by leaving out the preceding chapter designation, with the full compound number present in each figure's caption.

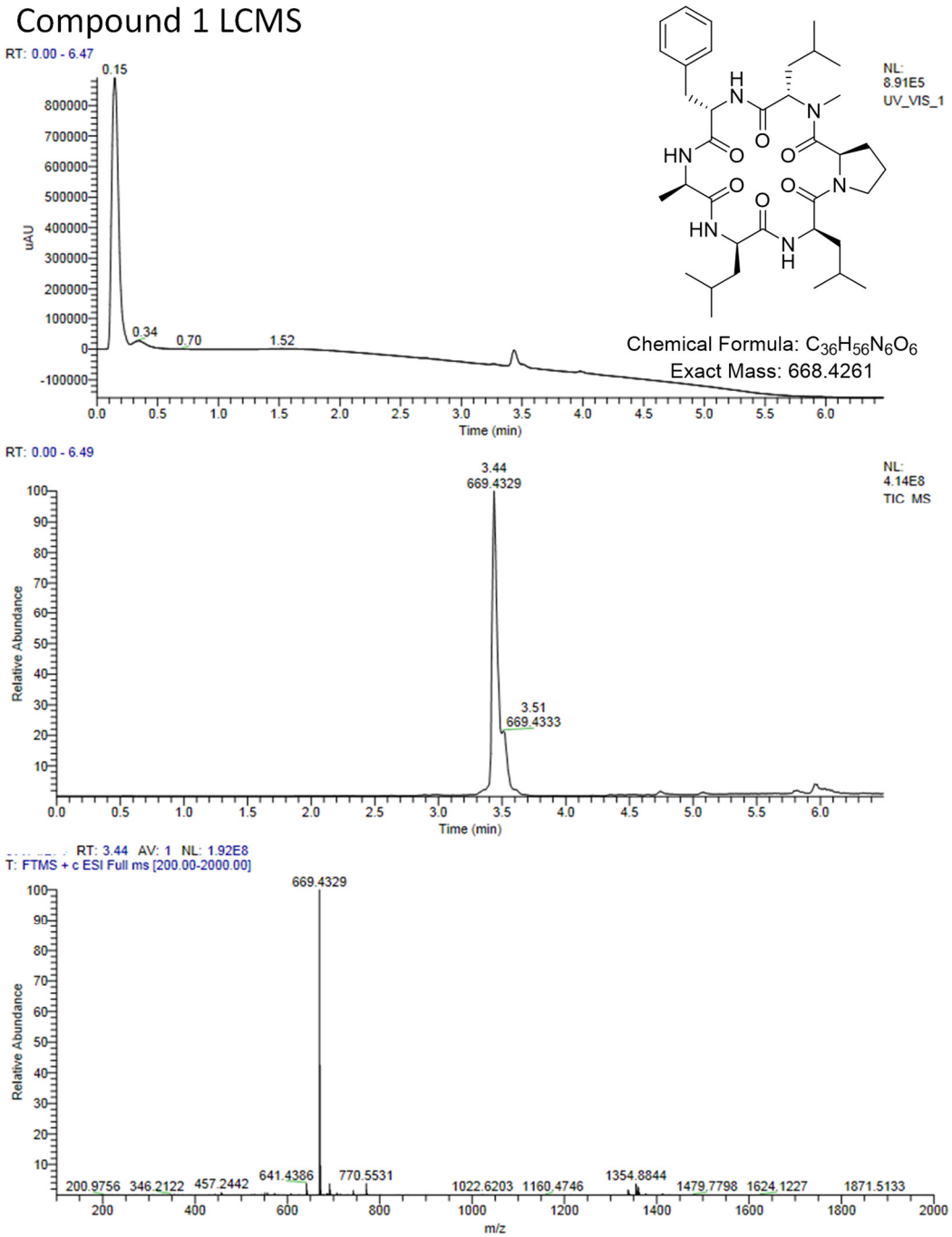
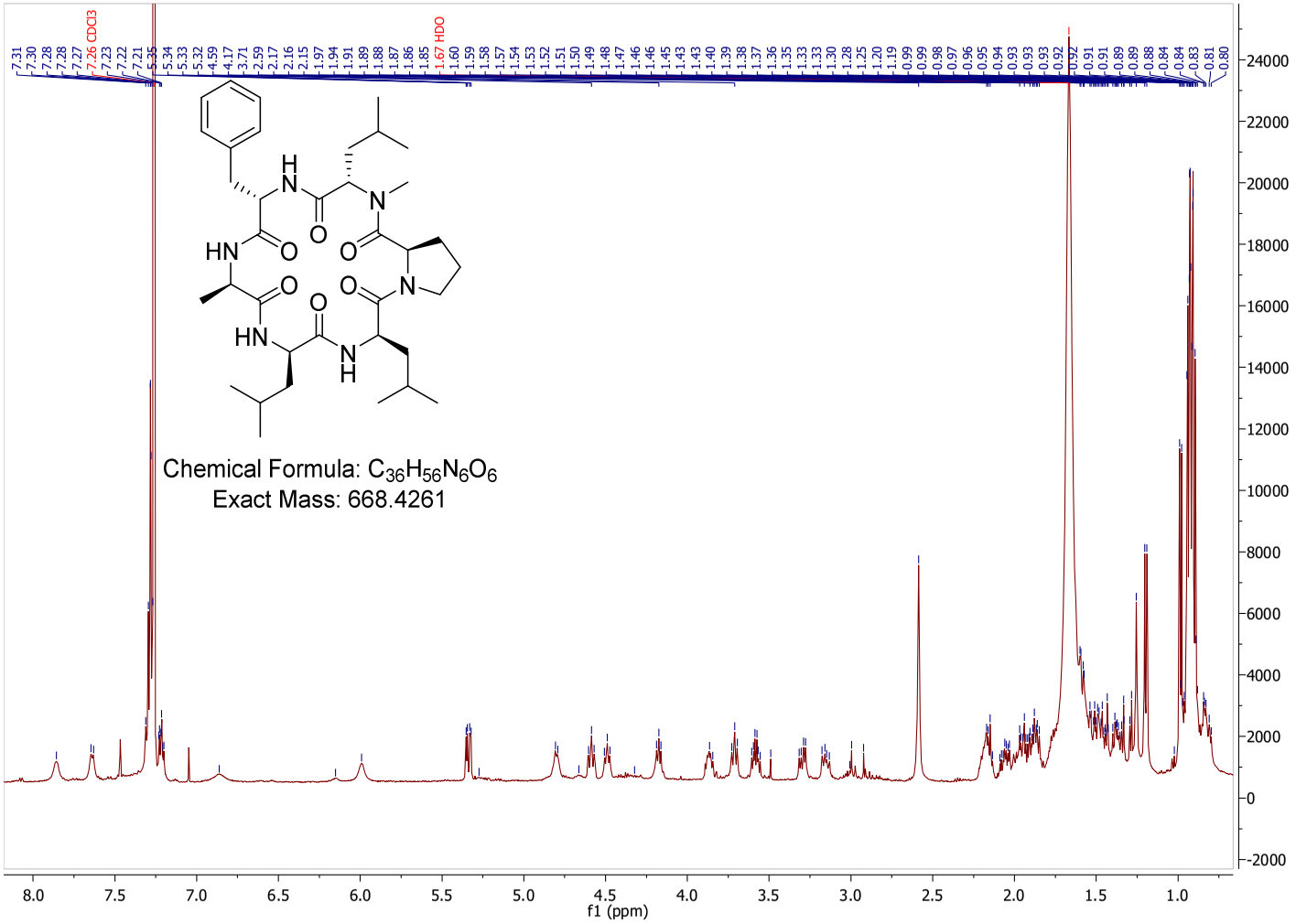


Figure S2.8.1 Compound 2.1 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.15**.**

Compound 1 H1 NMR ($CDCl_3$)



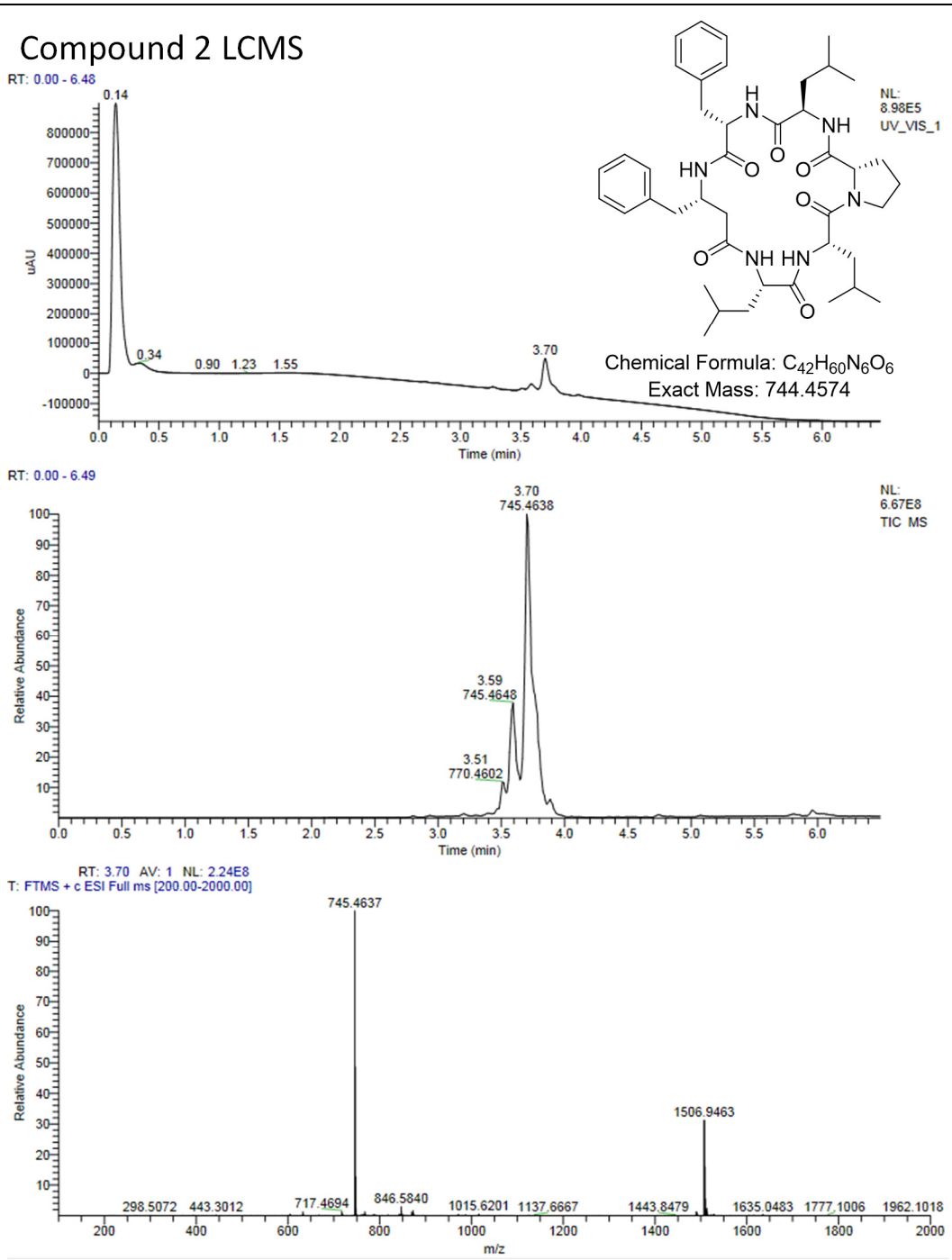
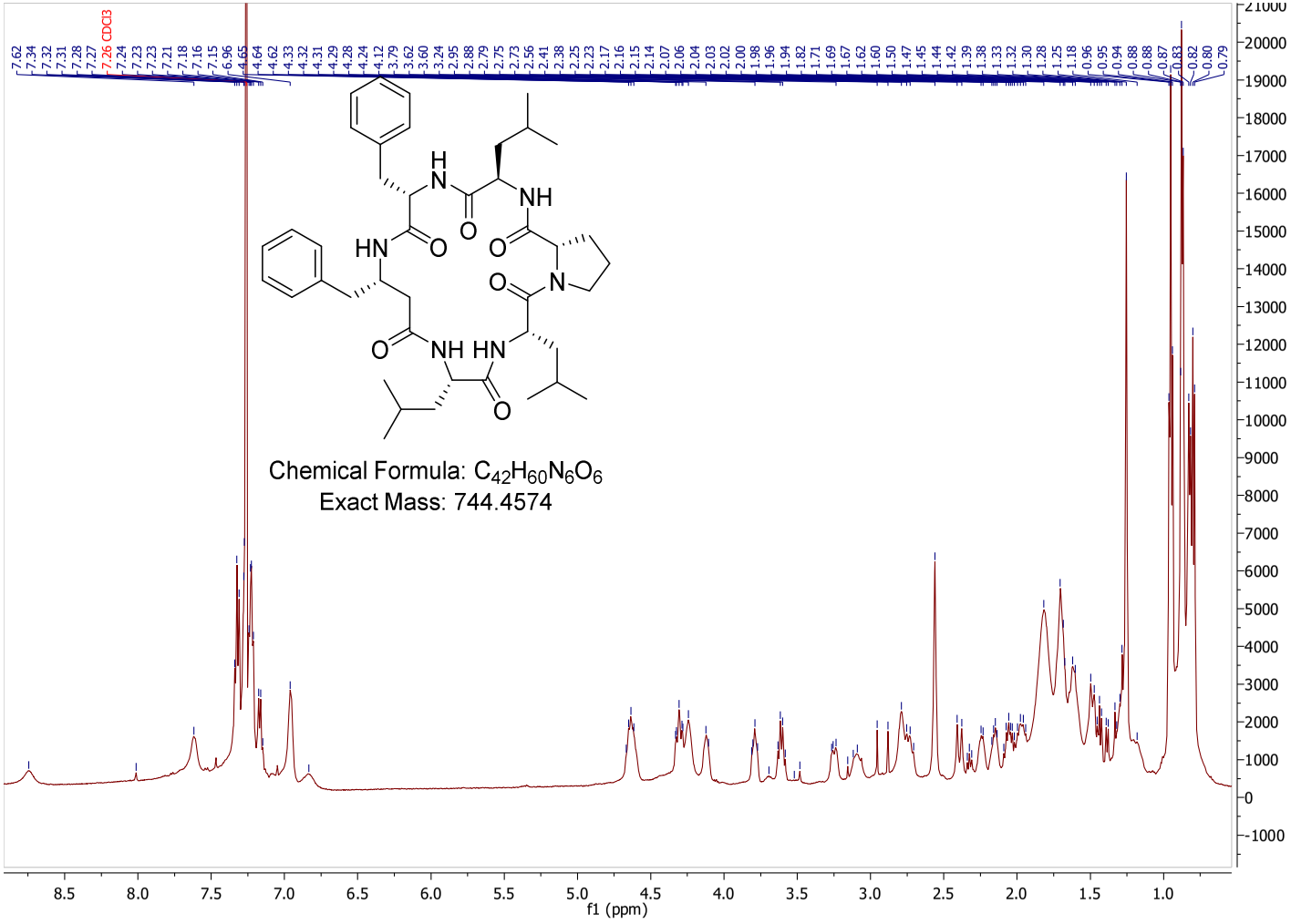


Figure S2.8.2 Compound 2.2 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.22**.**

Compound 2 H1 NMR ($CDCl_3$)



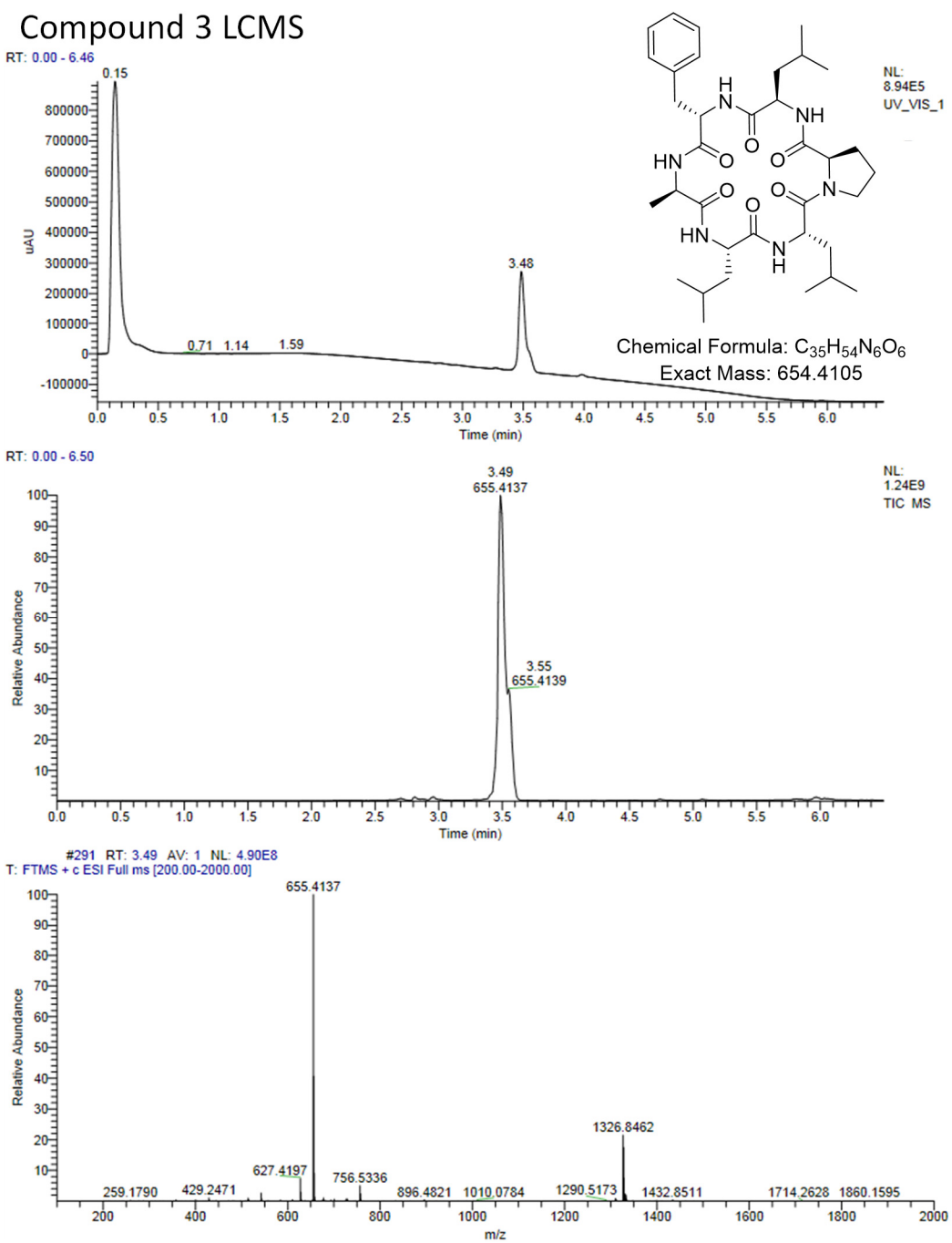
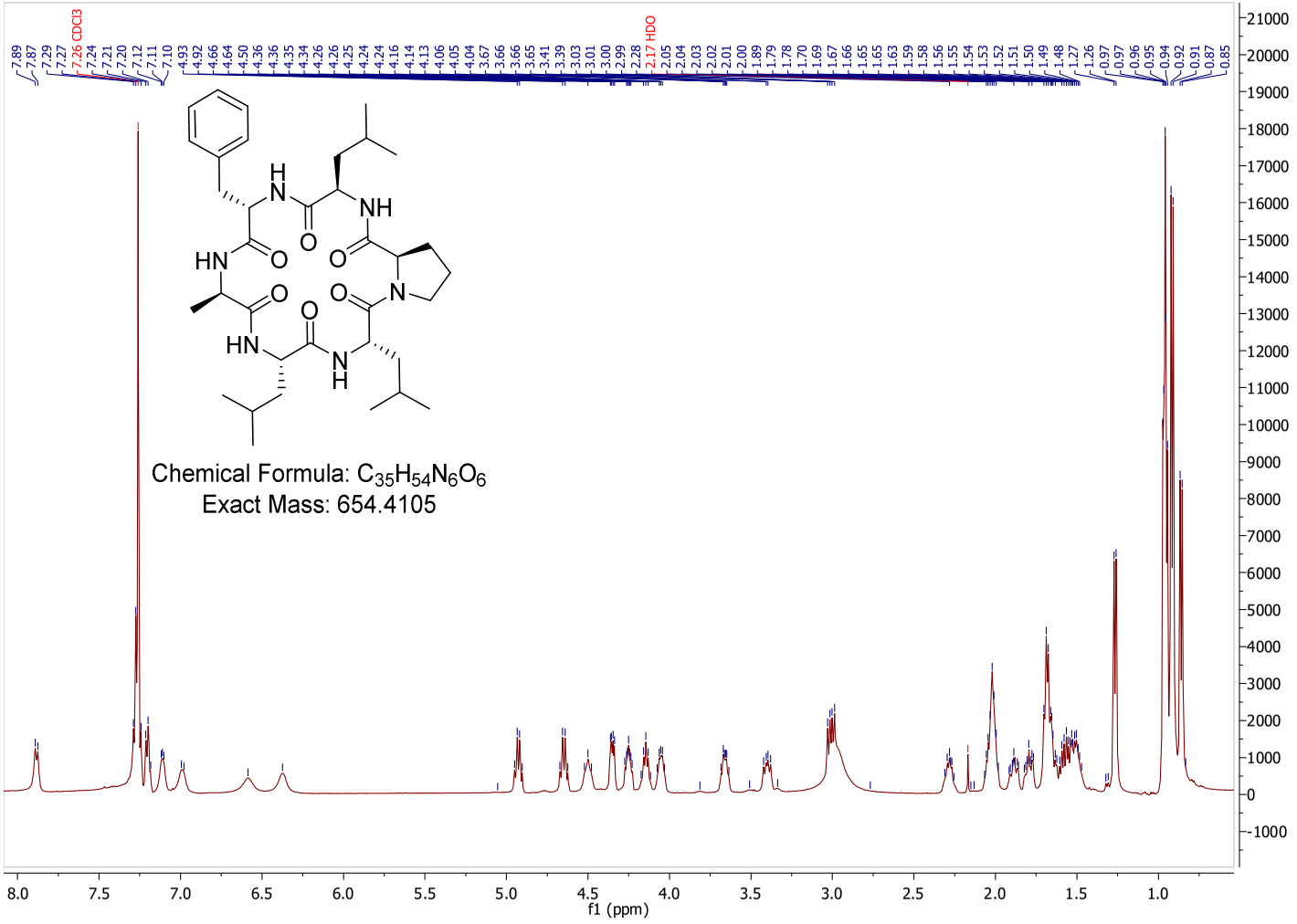


Figure S2.8.3 Compound 2.3 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.8.

Compound 3 H1 NMR ($CDCl_3$)



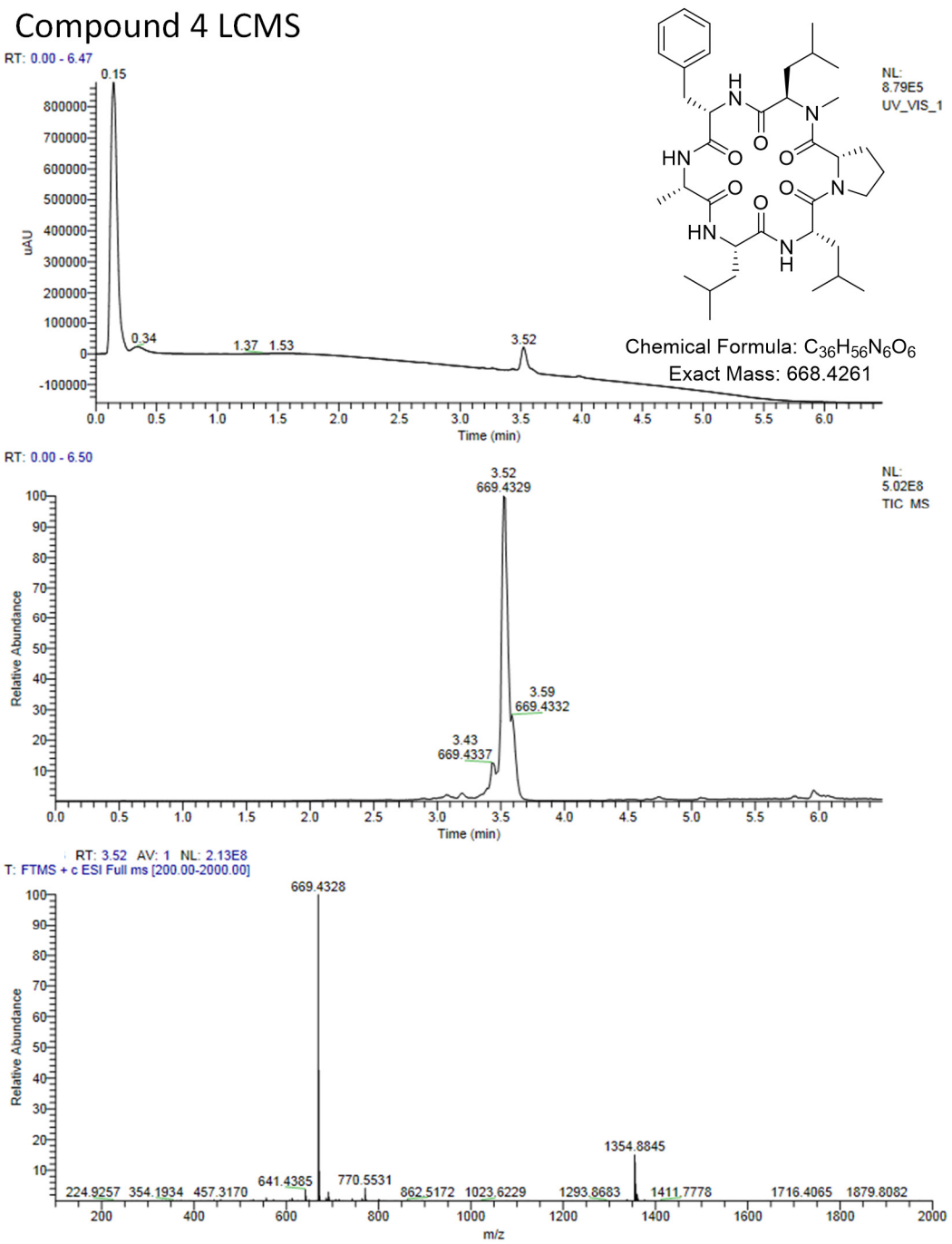
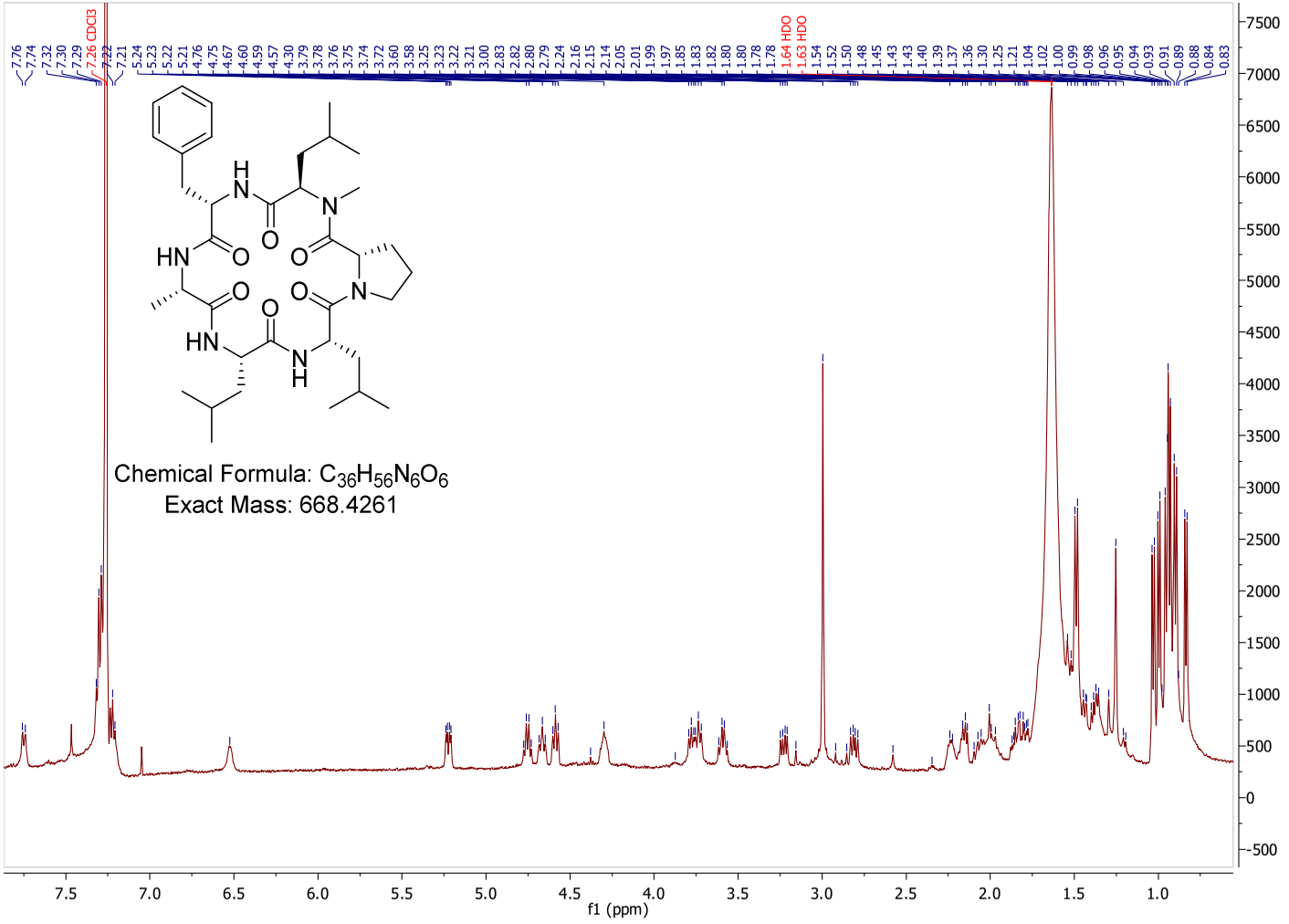


Figure S2.8.4 Compound 2.4 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.14**.**

Compound 4 H1 NMR ($CDCl_3$)



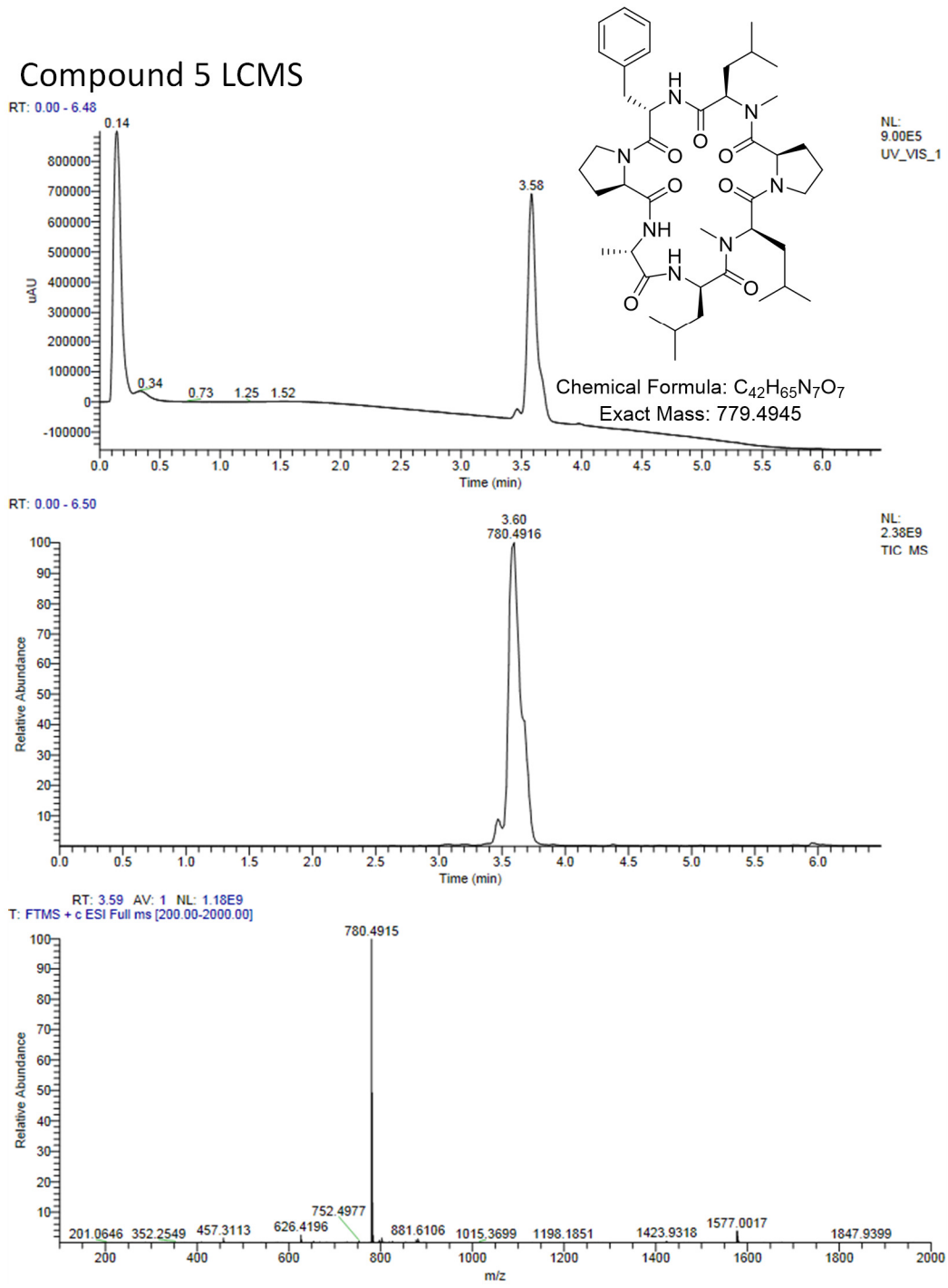
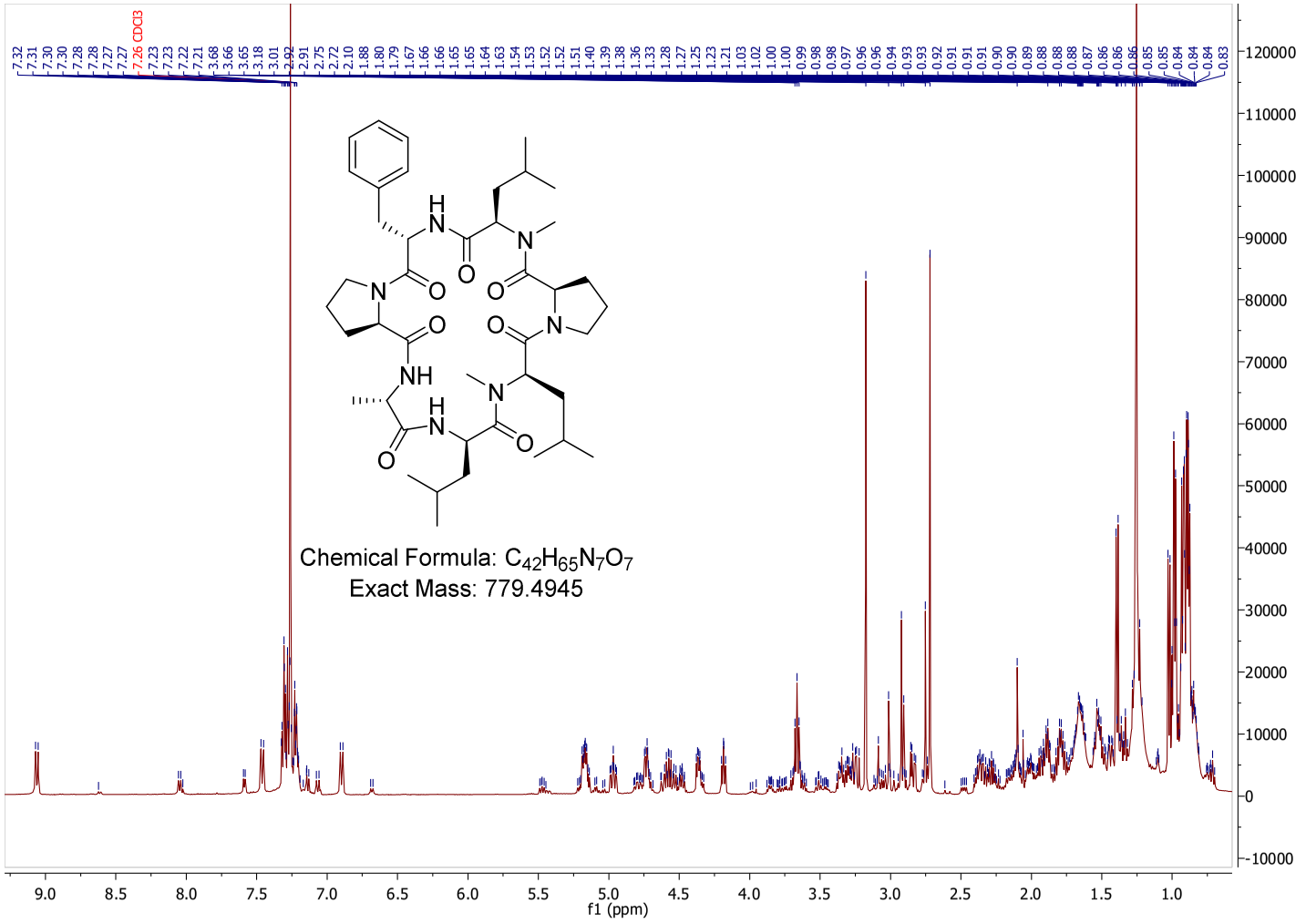


Figure S2.8.5 Compound 2.5 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 5 H1 NMR ($CDCl_3$)



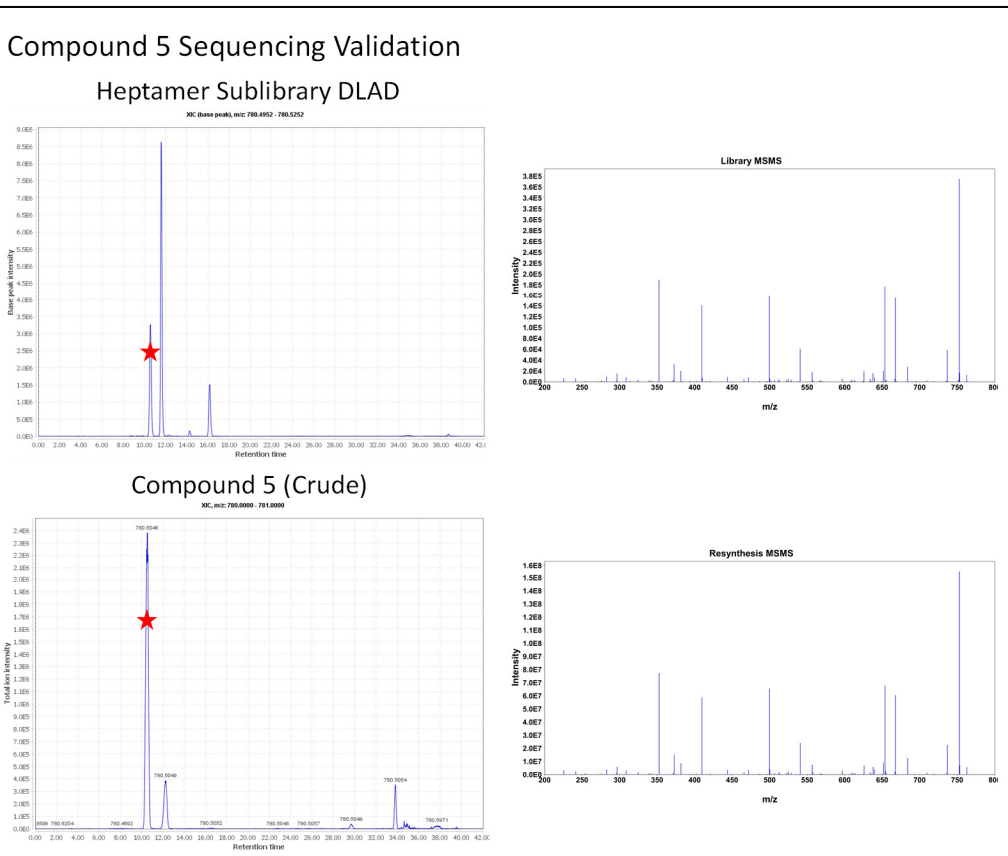


Figure S2.8.5 Compound 2.5 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

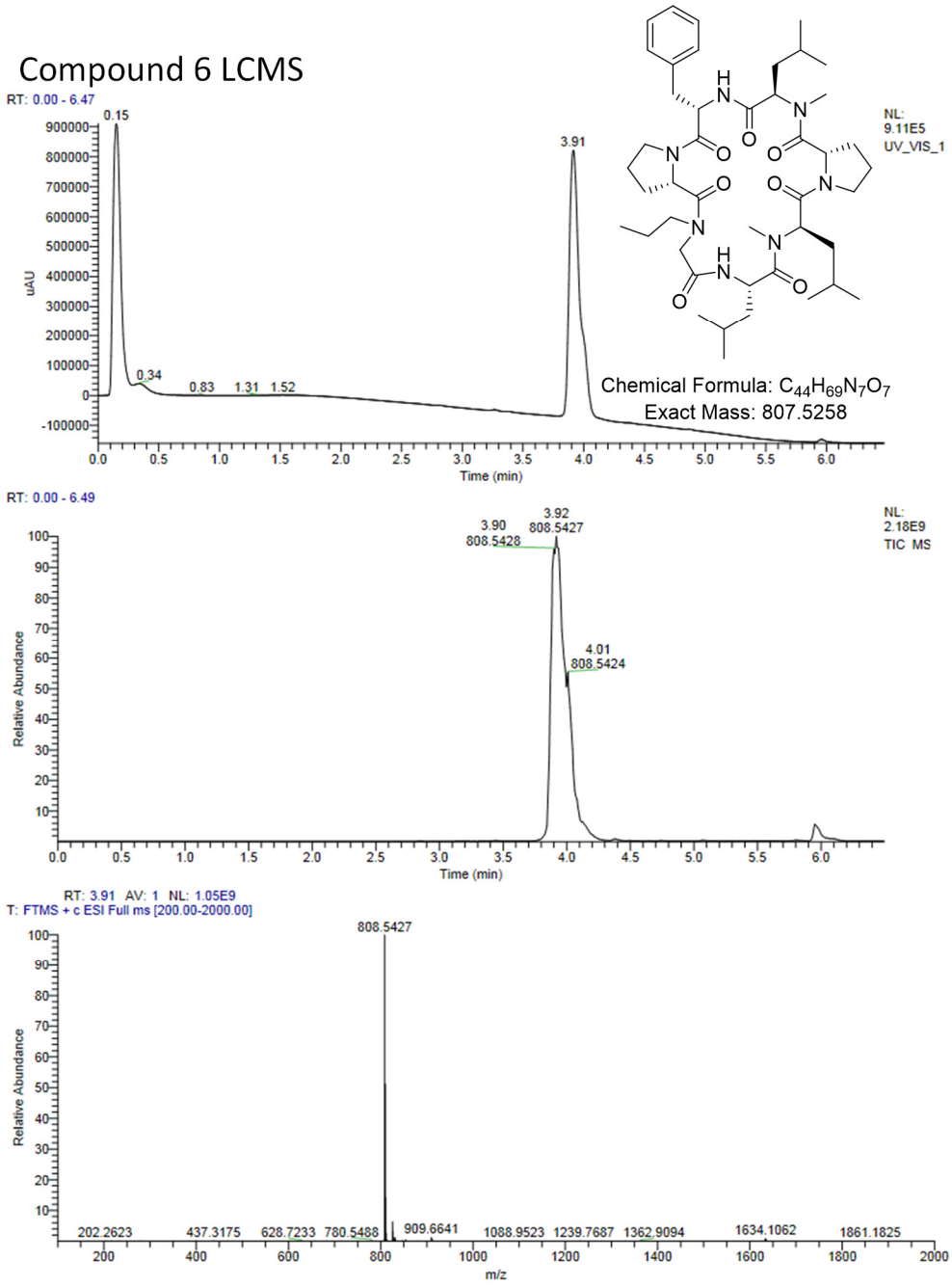
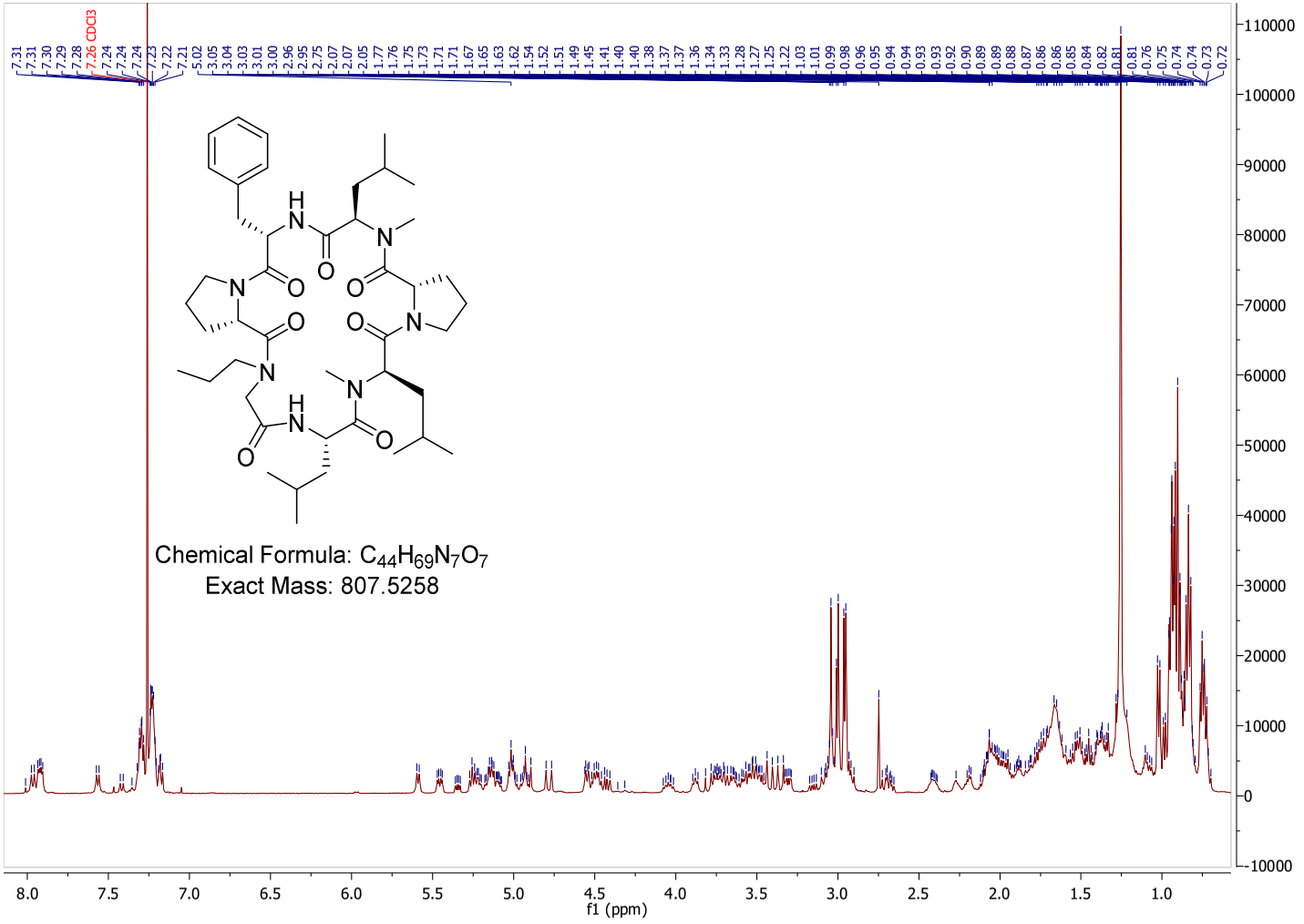


Figure S2.8.6 Compound 2.6 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 6 H1 NMR ($CDCl_3$)



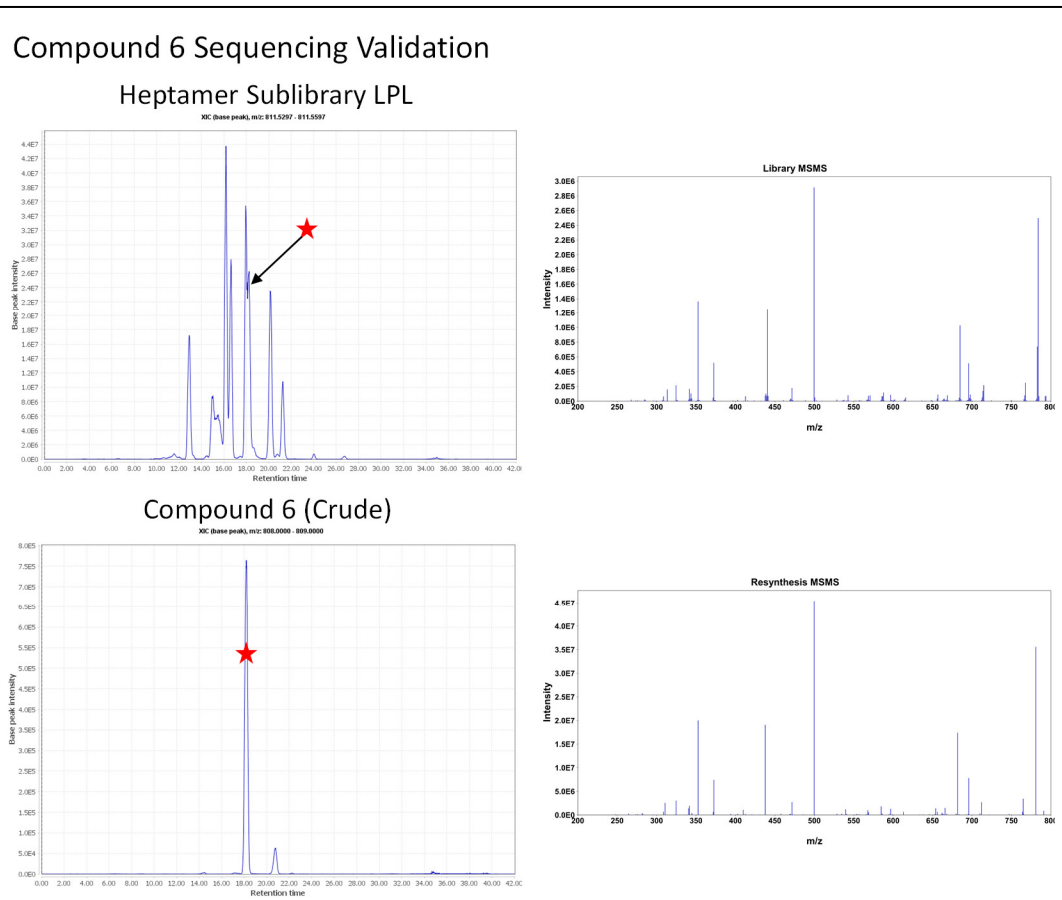


Figure S2.8.6 Compound 2.6 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

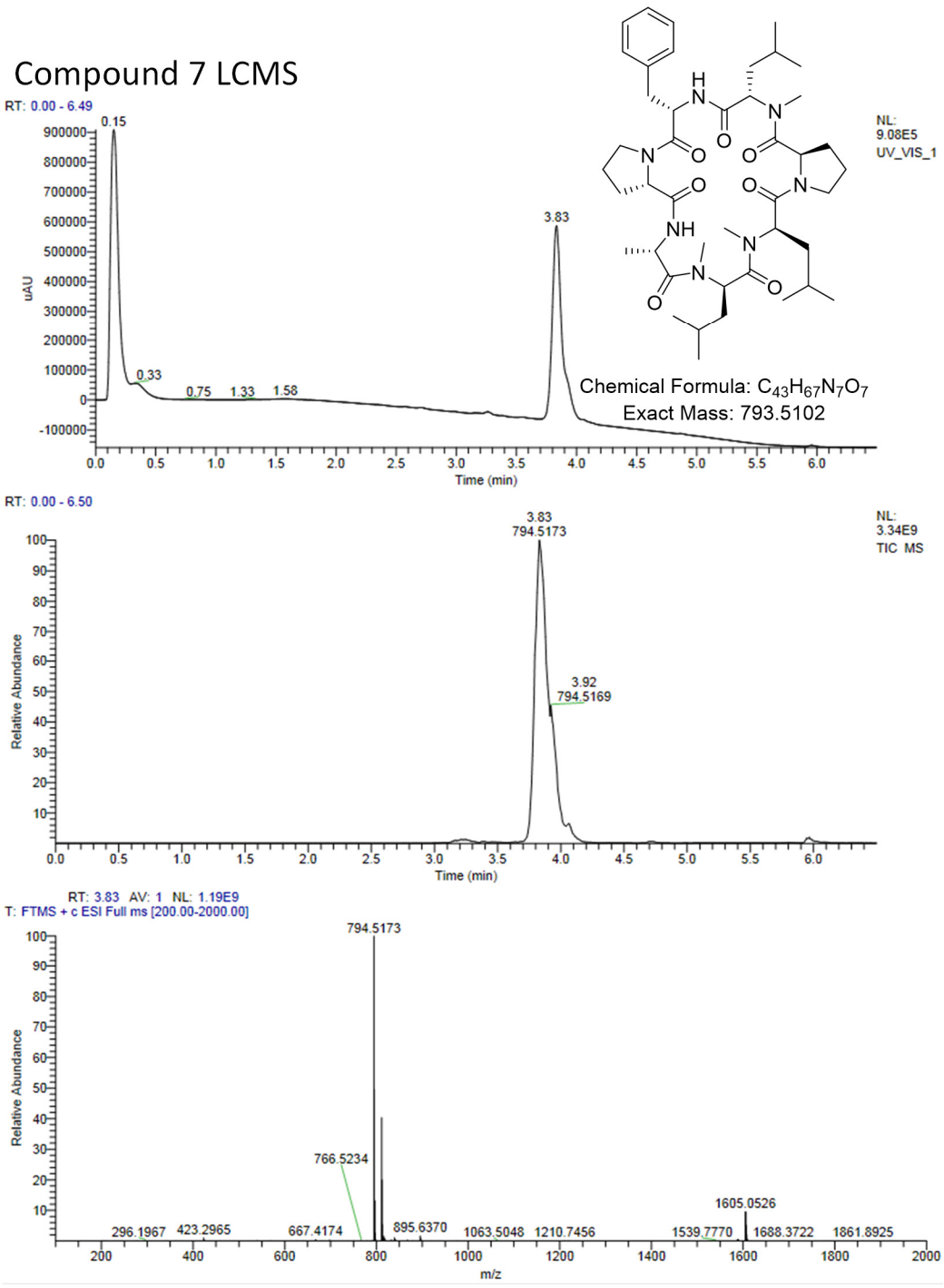
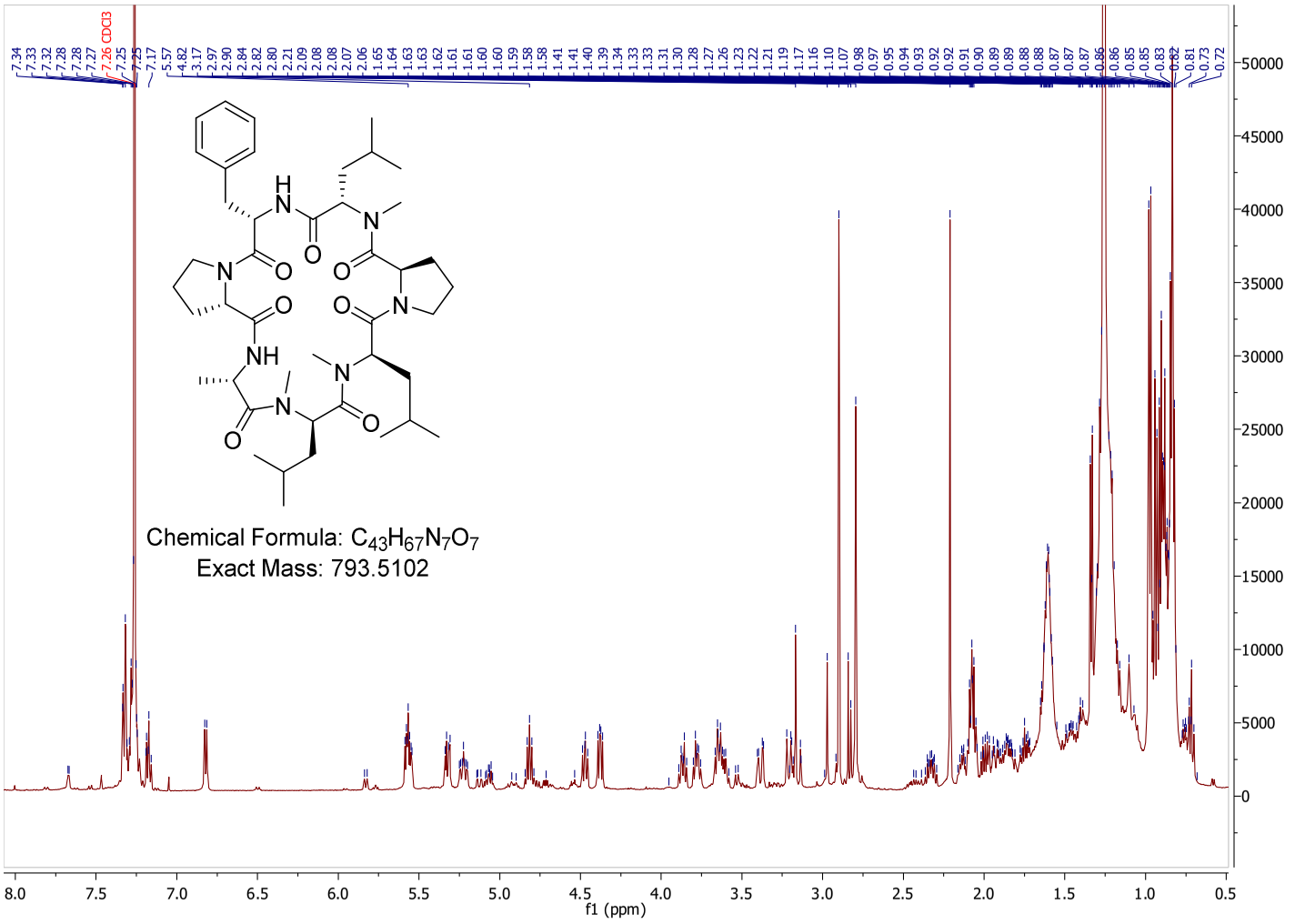


Figure S2.8.7 Compound 2.7 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 7 H1 NMR ($CDCl_3$)



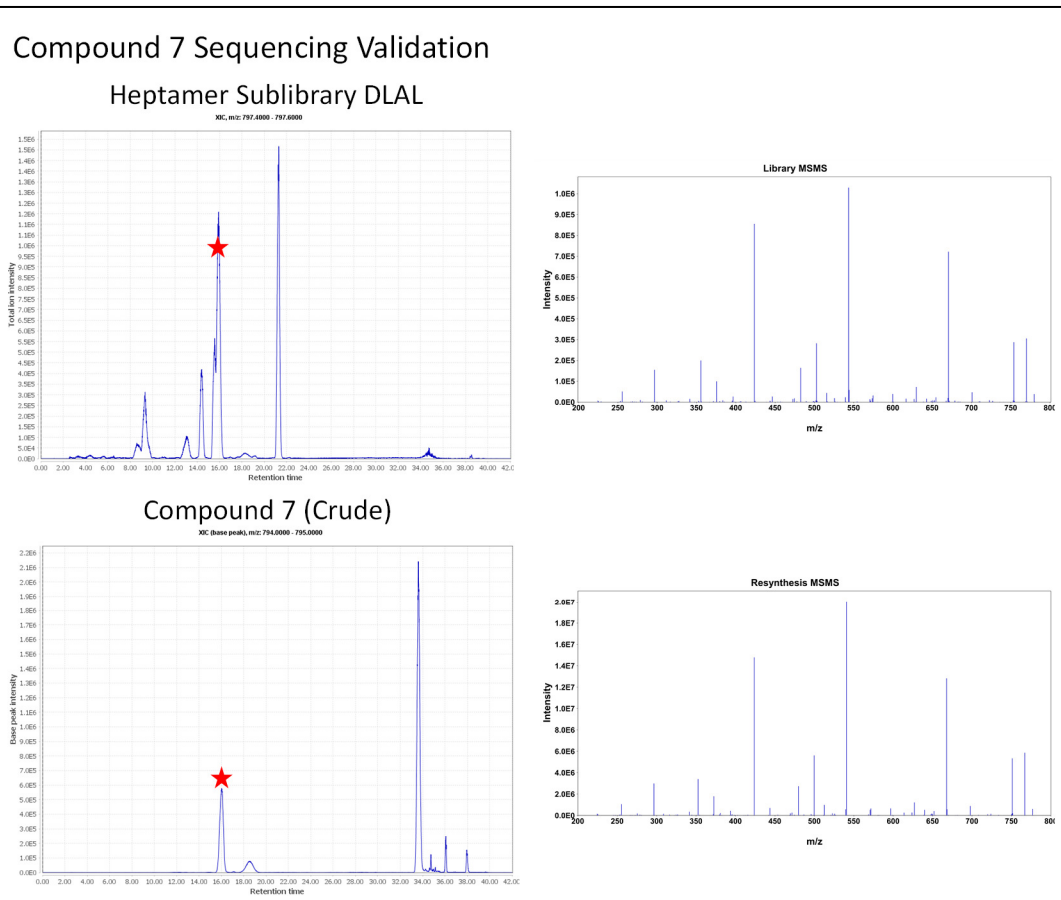


Figure S2.8.7 Compound 2.7 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

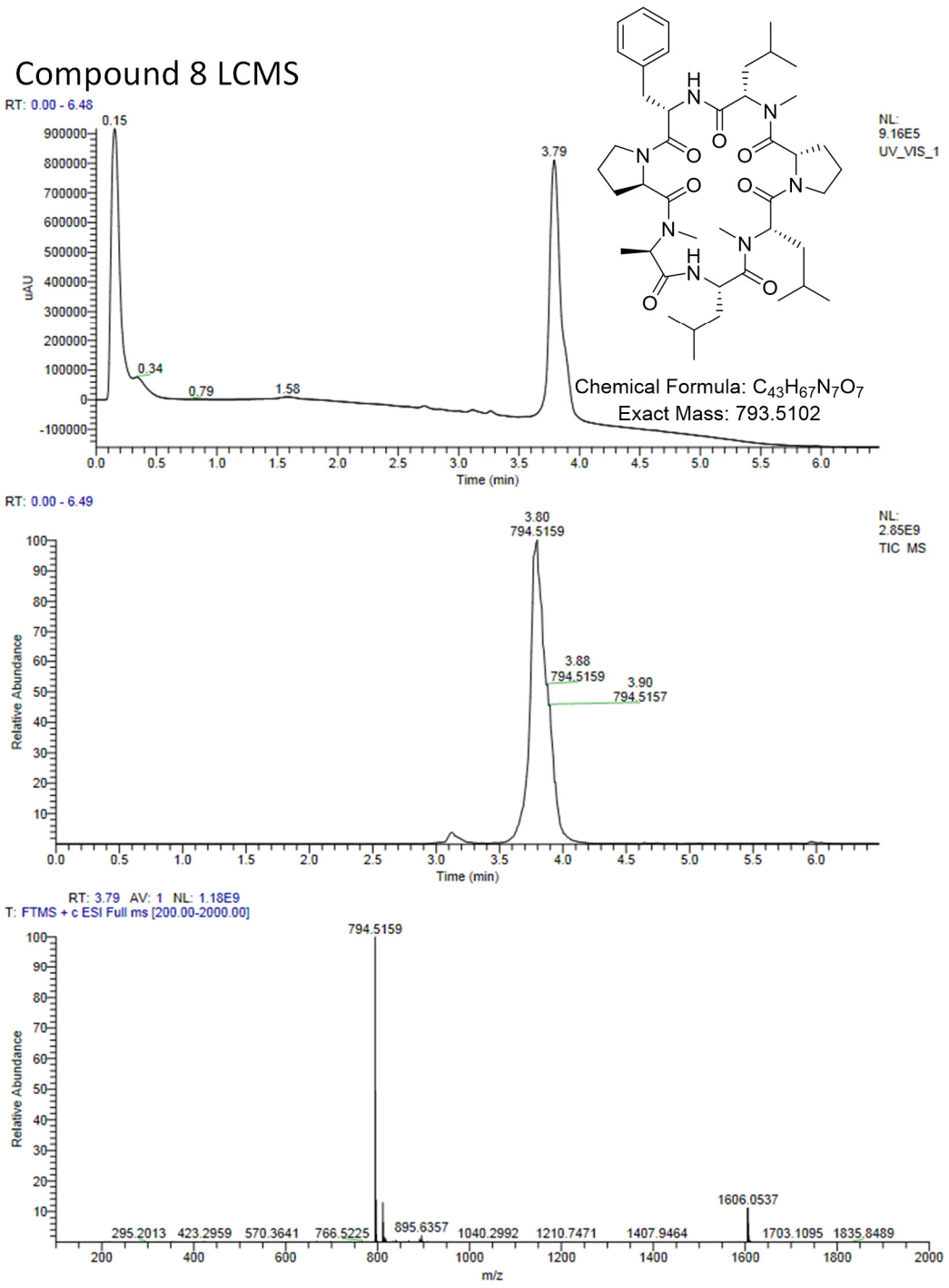
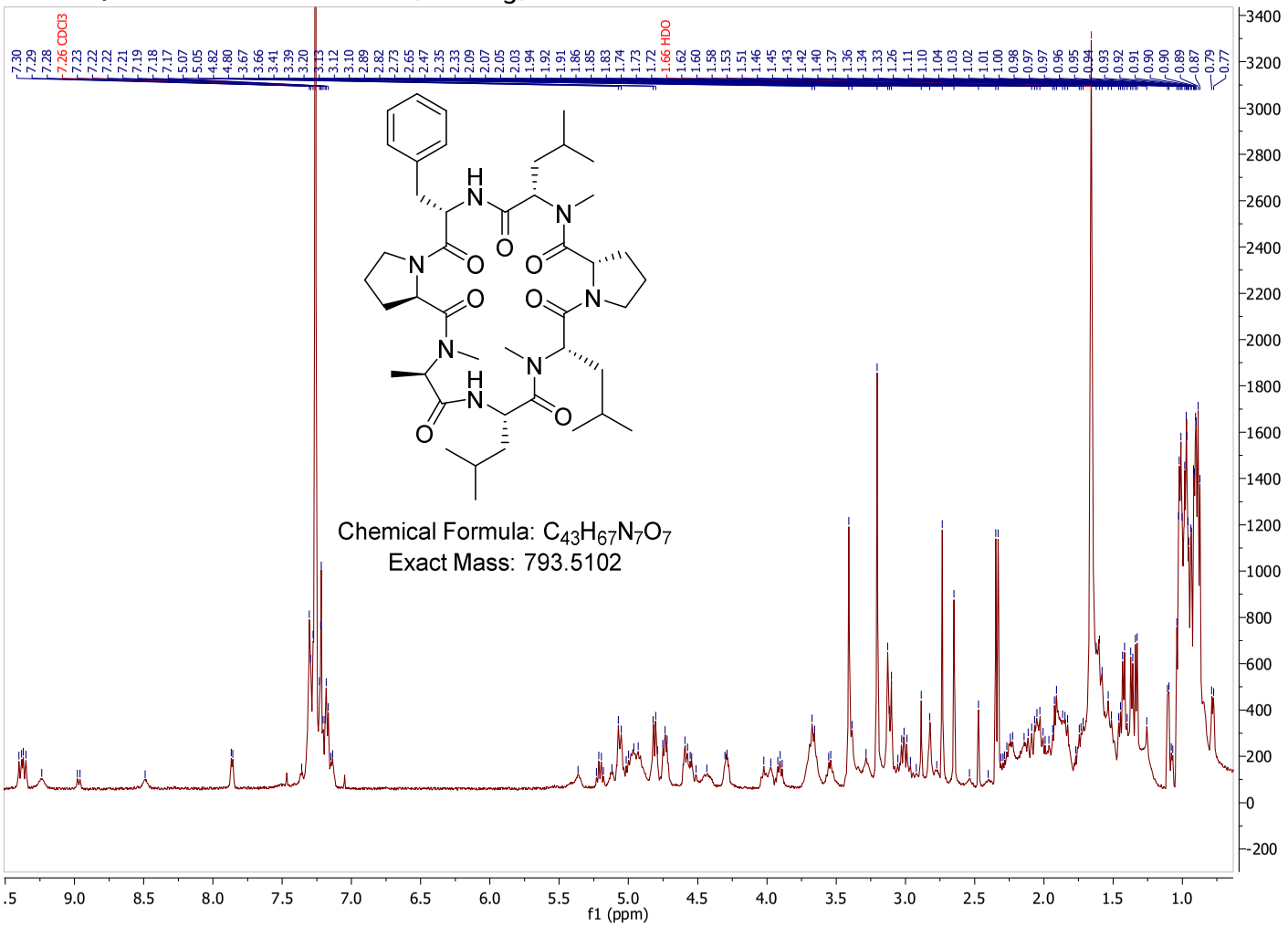


Figure S2.8.8 Compound 2.8 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 8 H1 NMR ($CDCl_3$)



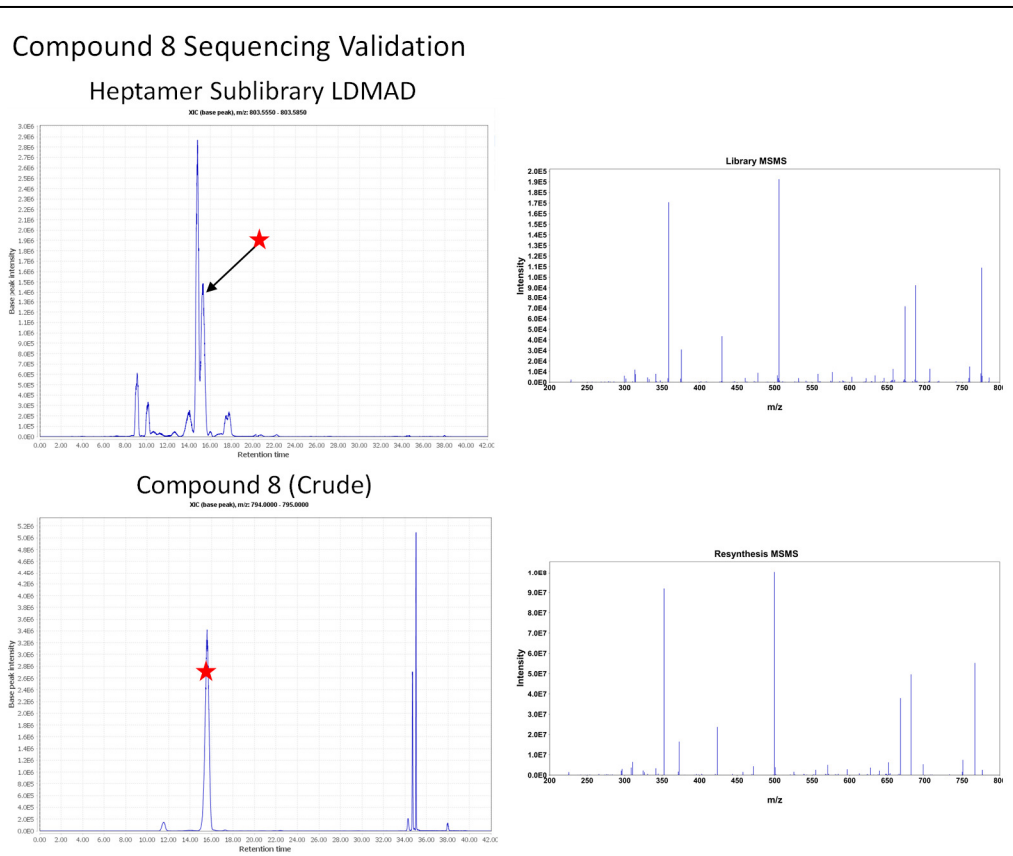


Figure S2.8.8 Compound 2.8 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

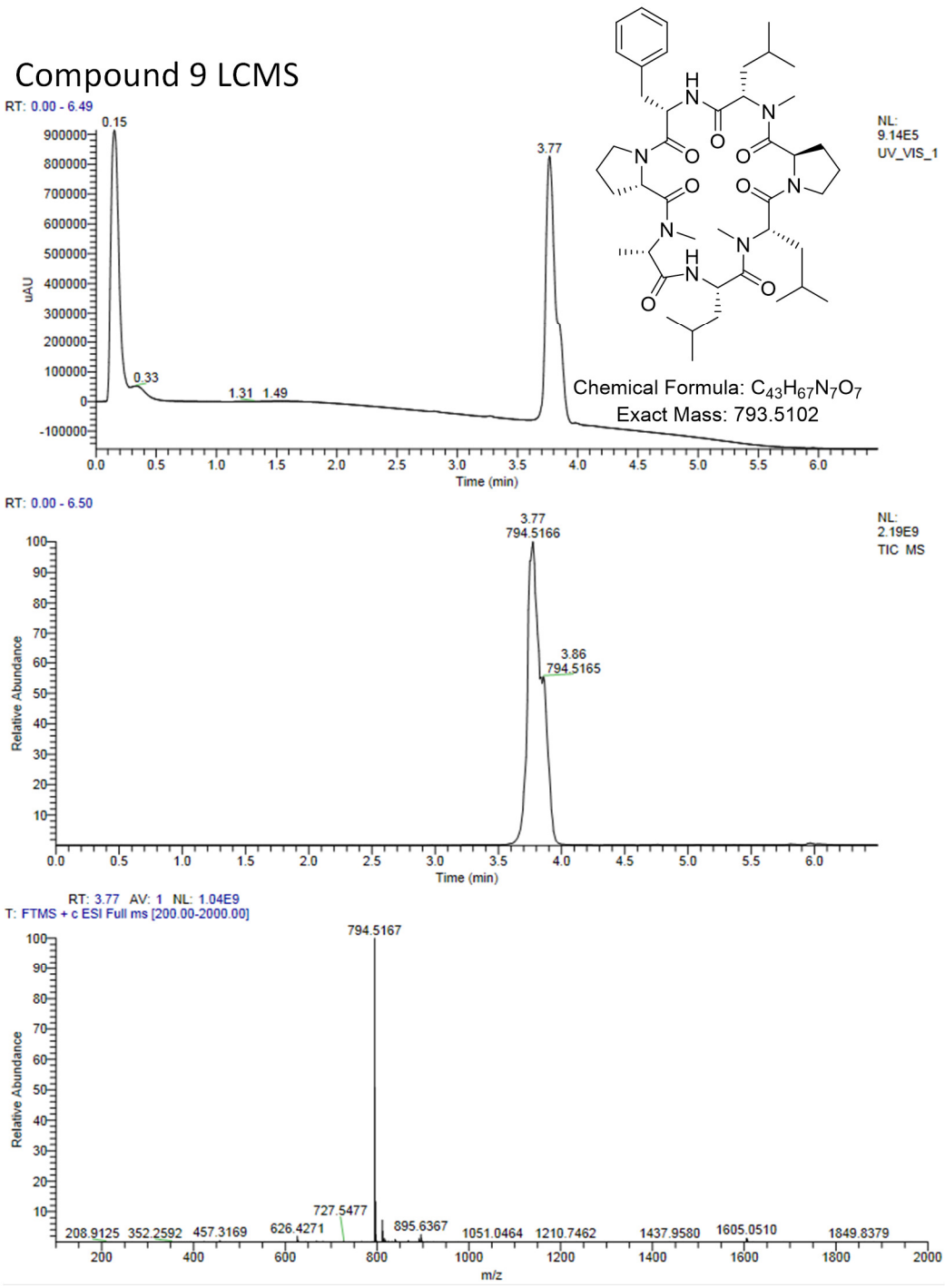
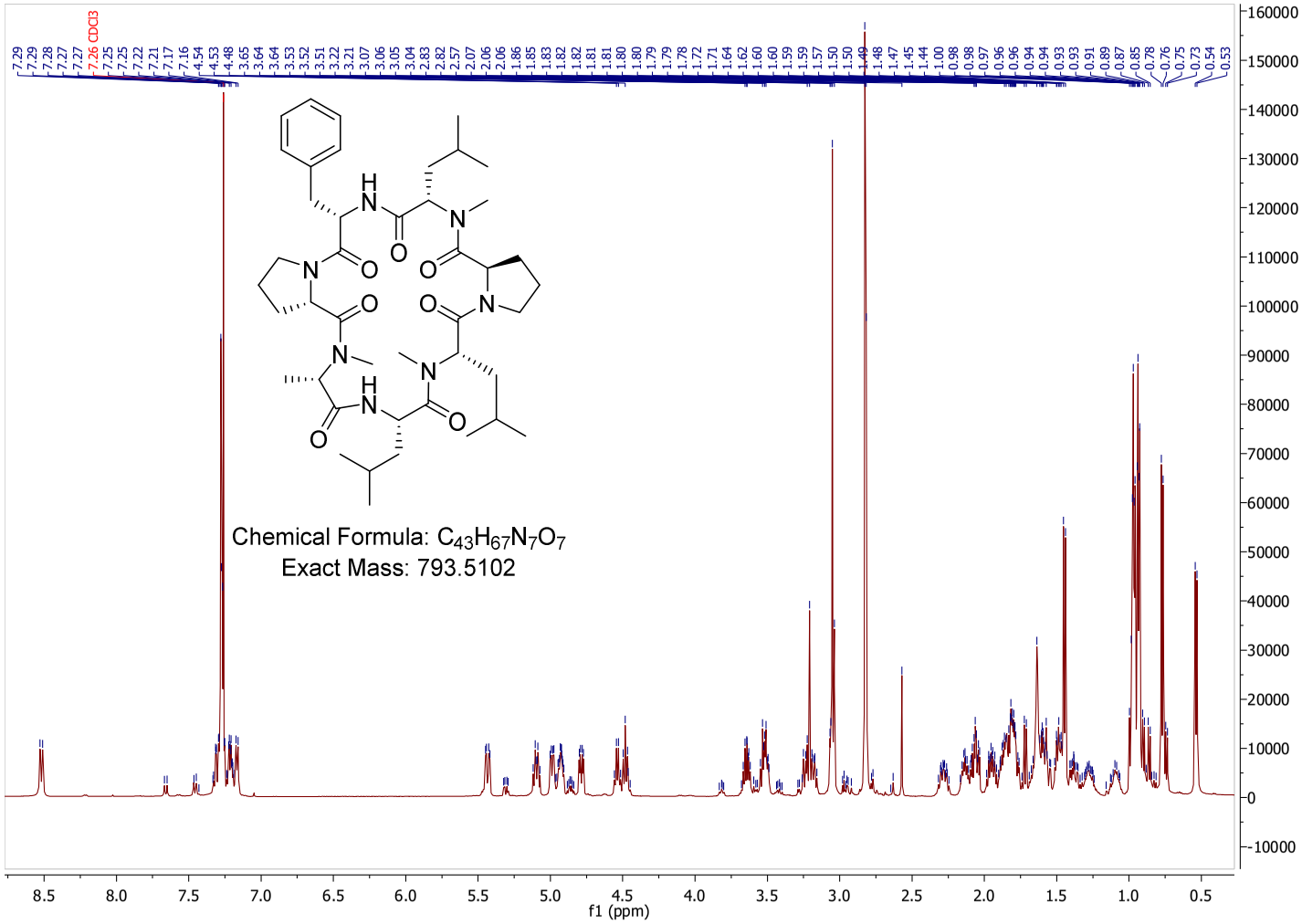


Figure S2.8.9 Compound 2.9 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 9 H1 NMR ($CDCl_3$)



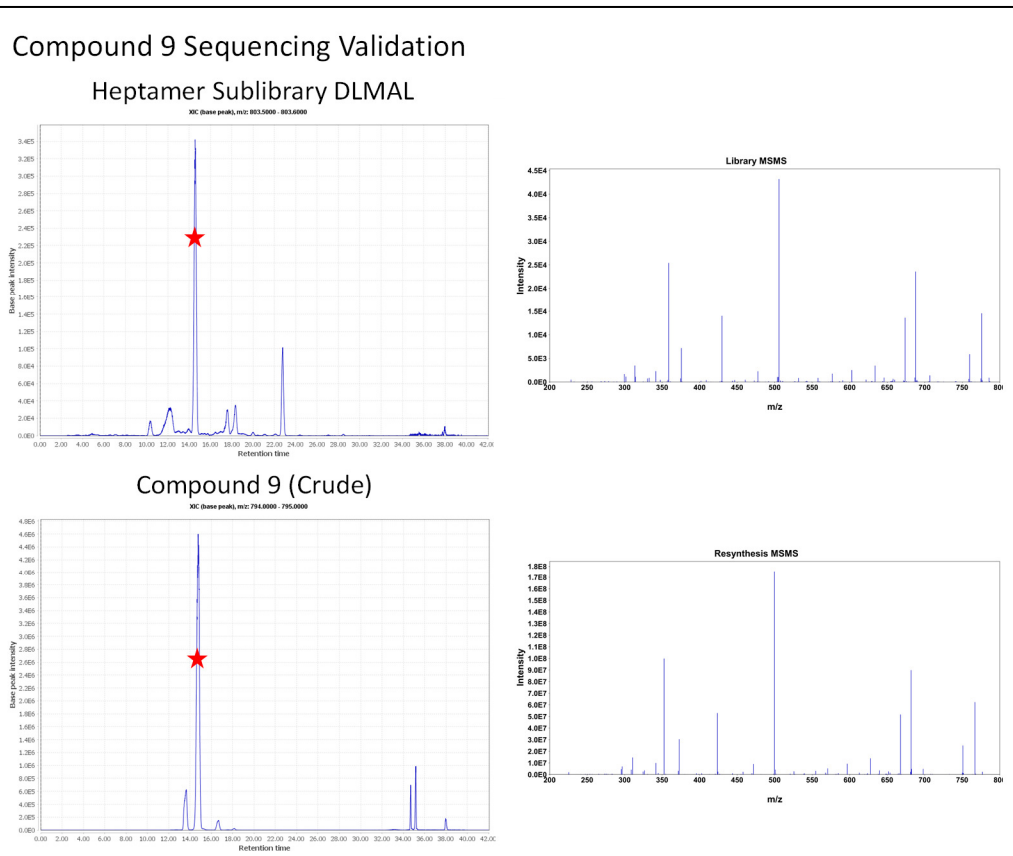


Figure S2.8.9 Compound 2.9 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

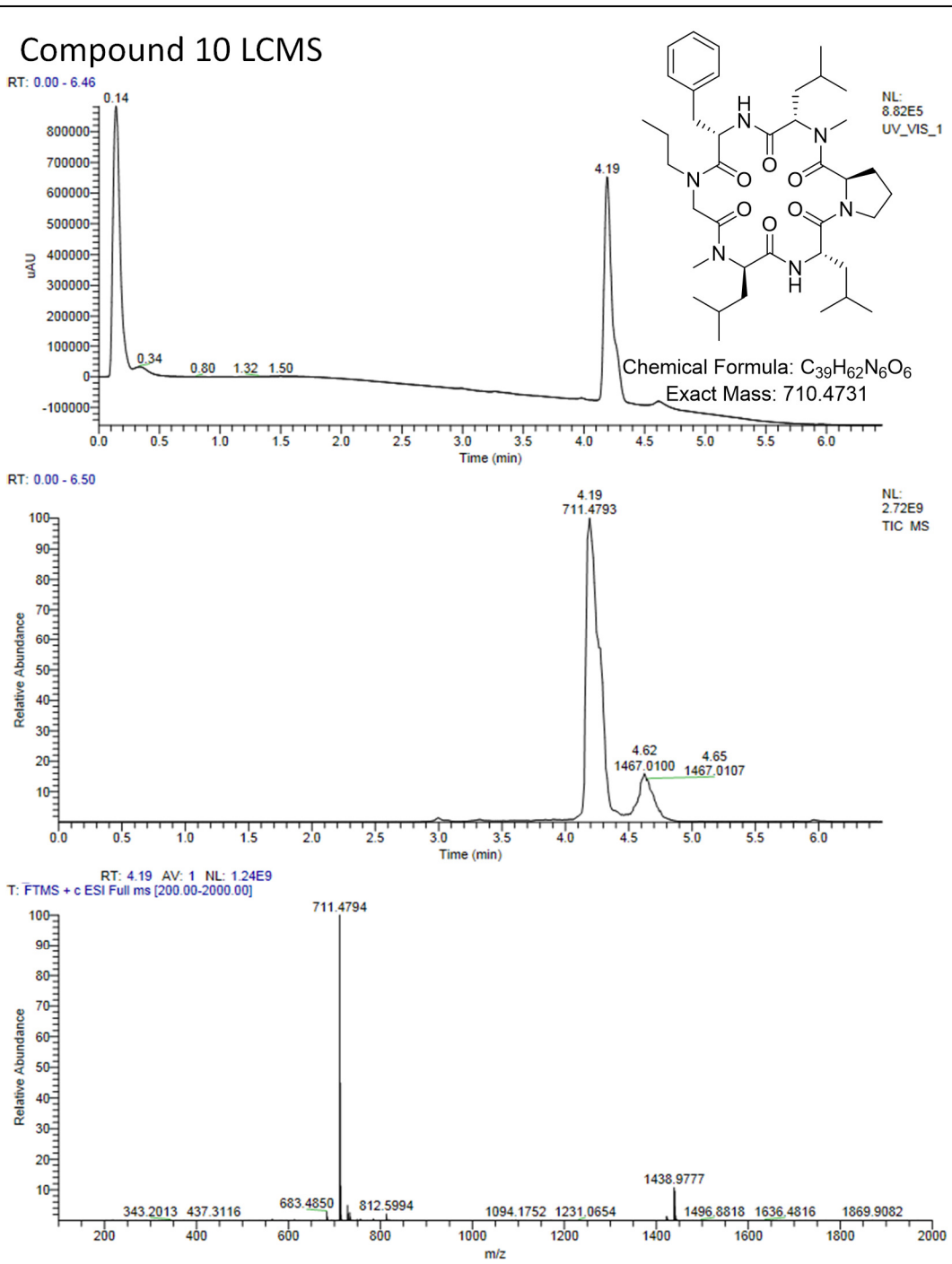
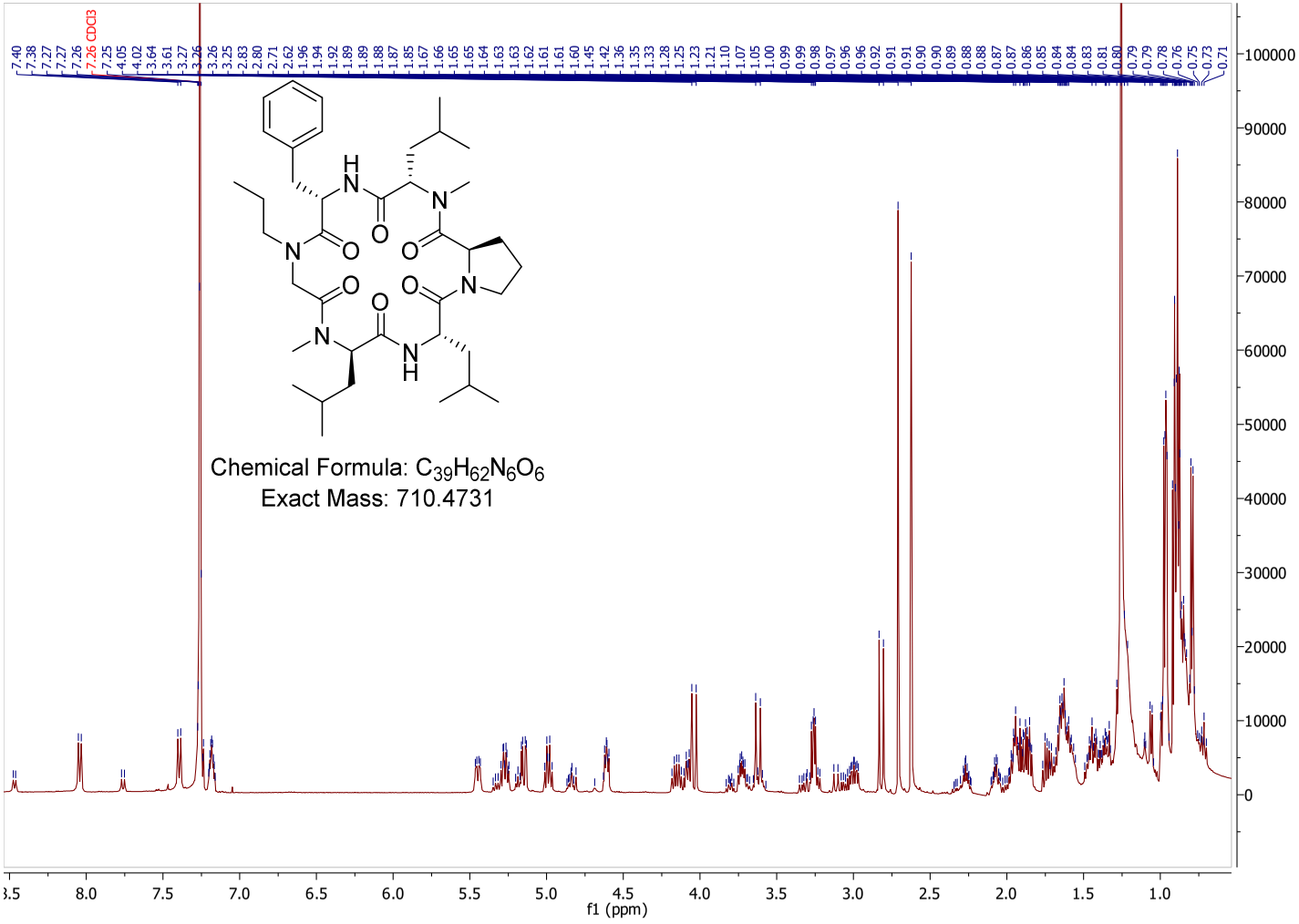


Figure S2.8.10 Compound 2.10 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 10 H1 NMR ($CDCl_3$)



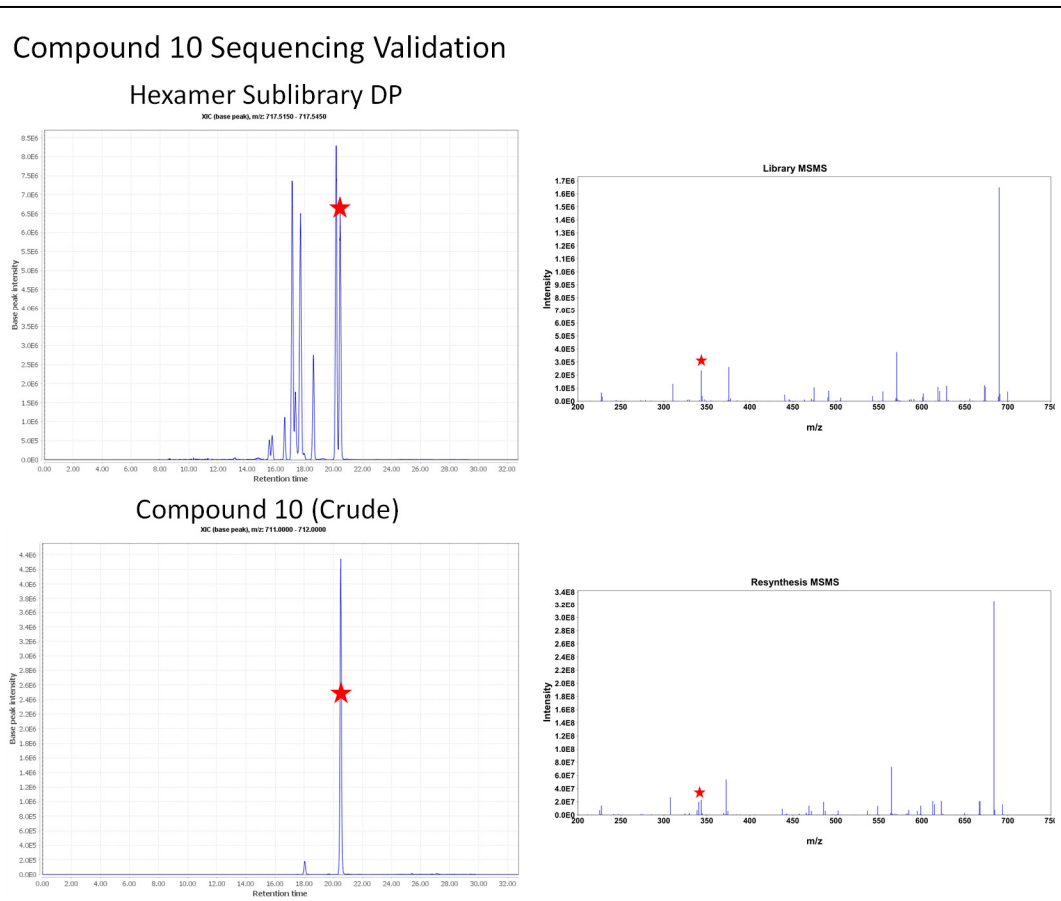


Figure S2.8.10 Compound 2.10 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound. Generally, isotopic labelling causes only small shifts in peak position, but in this case causes two peaks to stack atop one another in the library spectrum but not in the MS^2 spectrum of the resynthesized compound. The apparent difference between the two MS^2 spectra is marked by the small red stars.

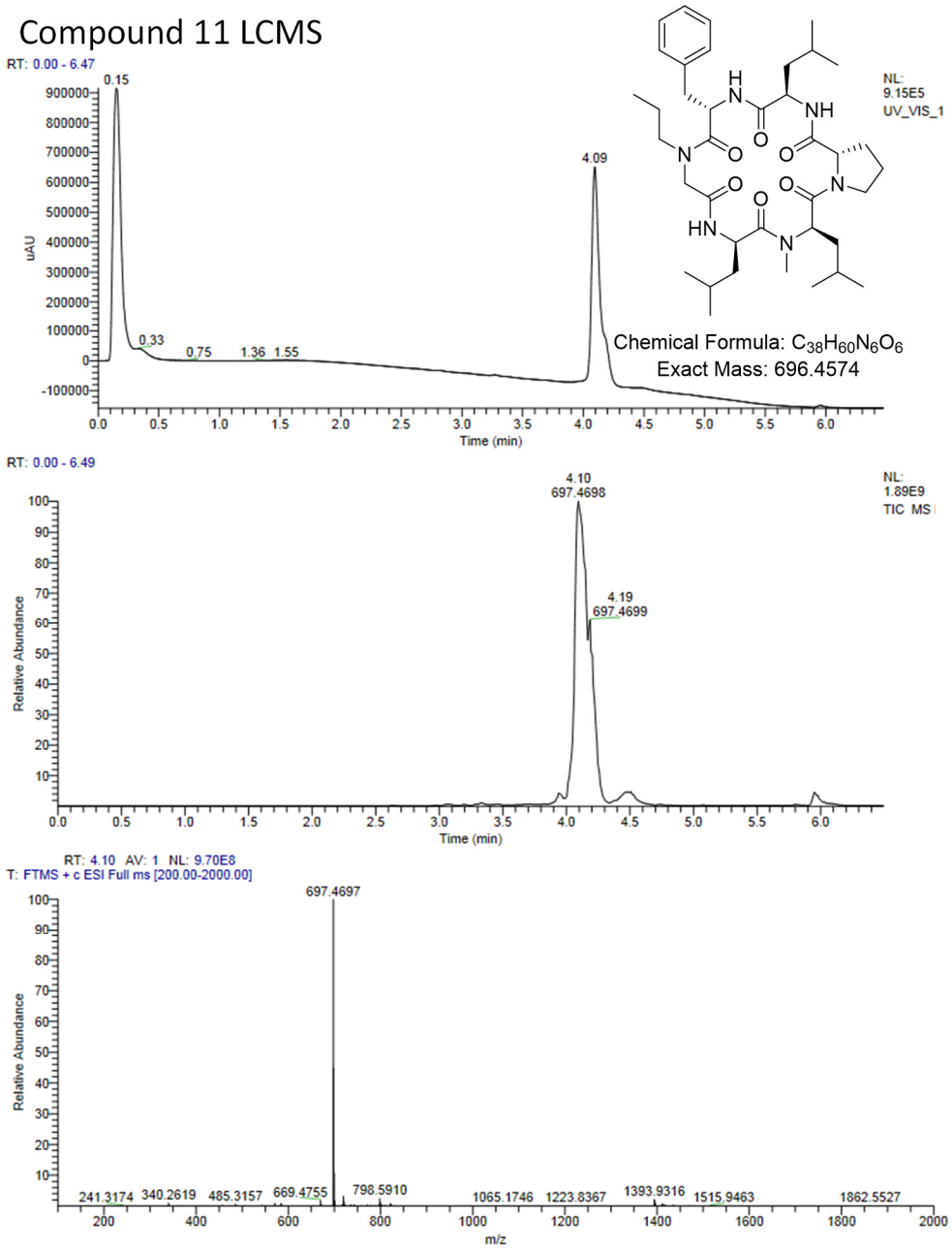
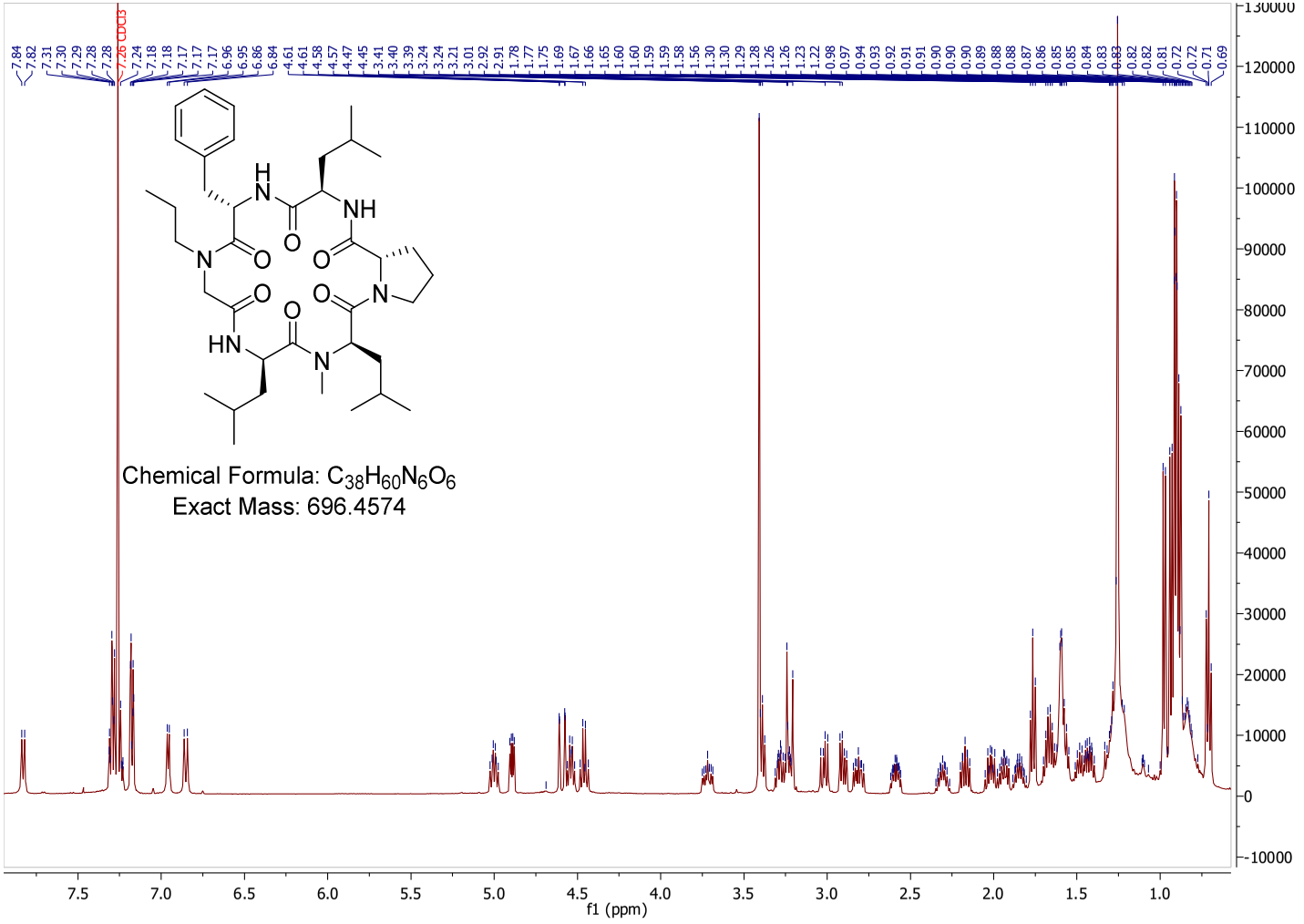


Figure S2.8.11 Compound 2.11 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 11 H1 NMR ($CDCl_3$)



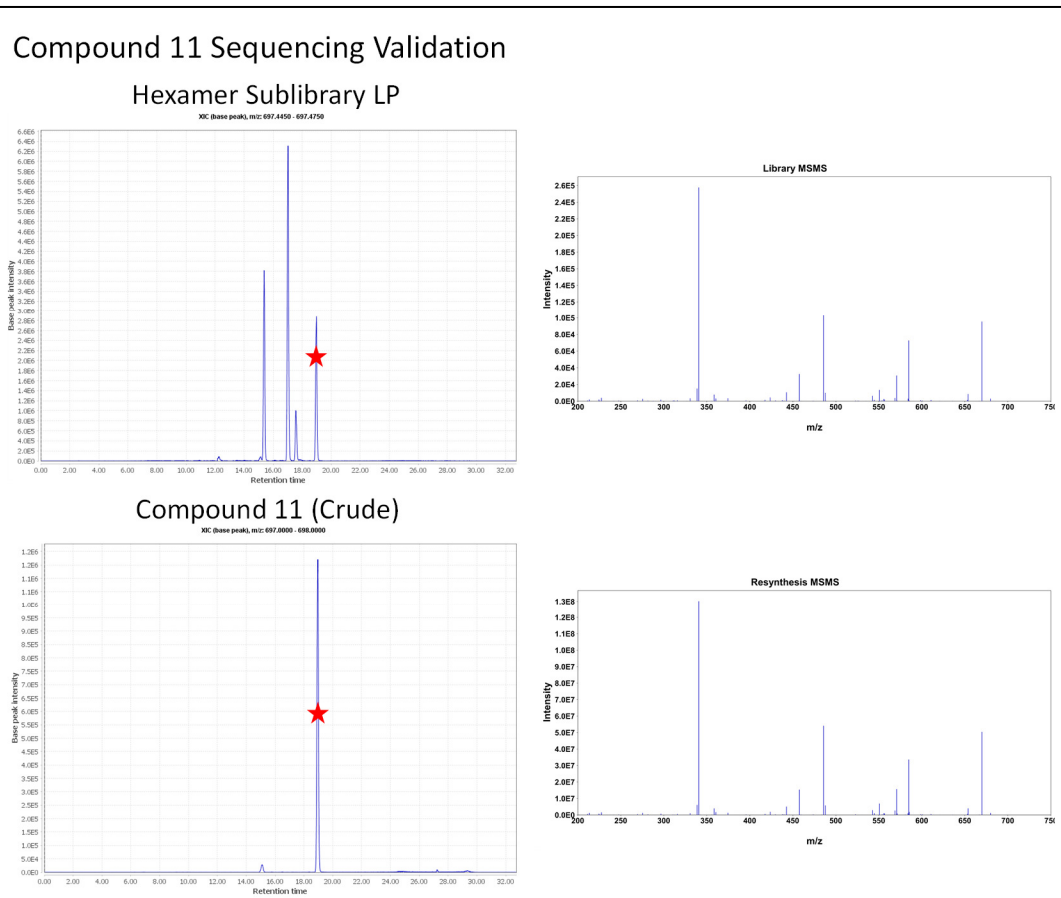


Figure S2.8.11 Compound 2.11 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

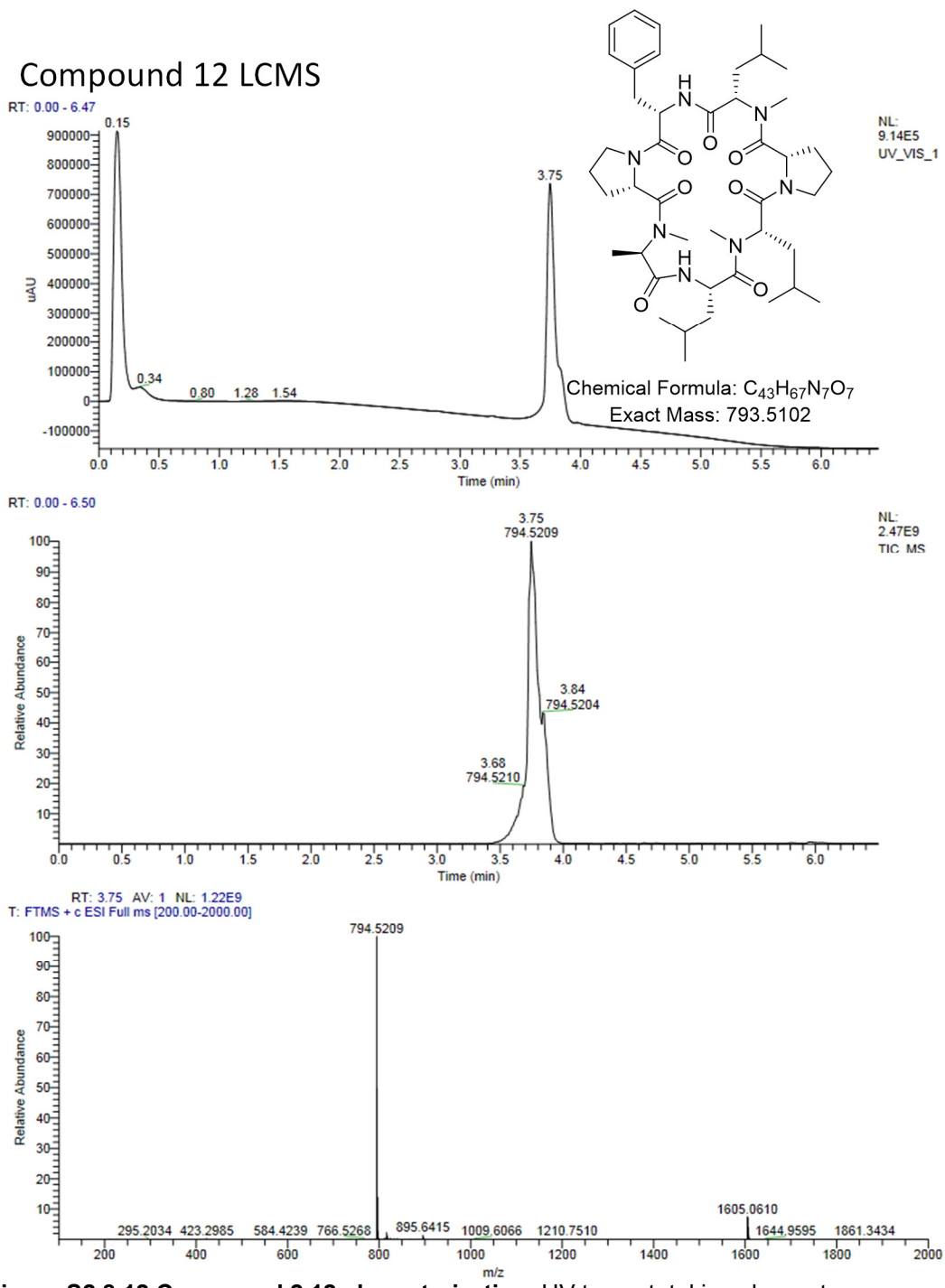
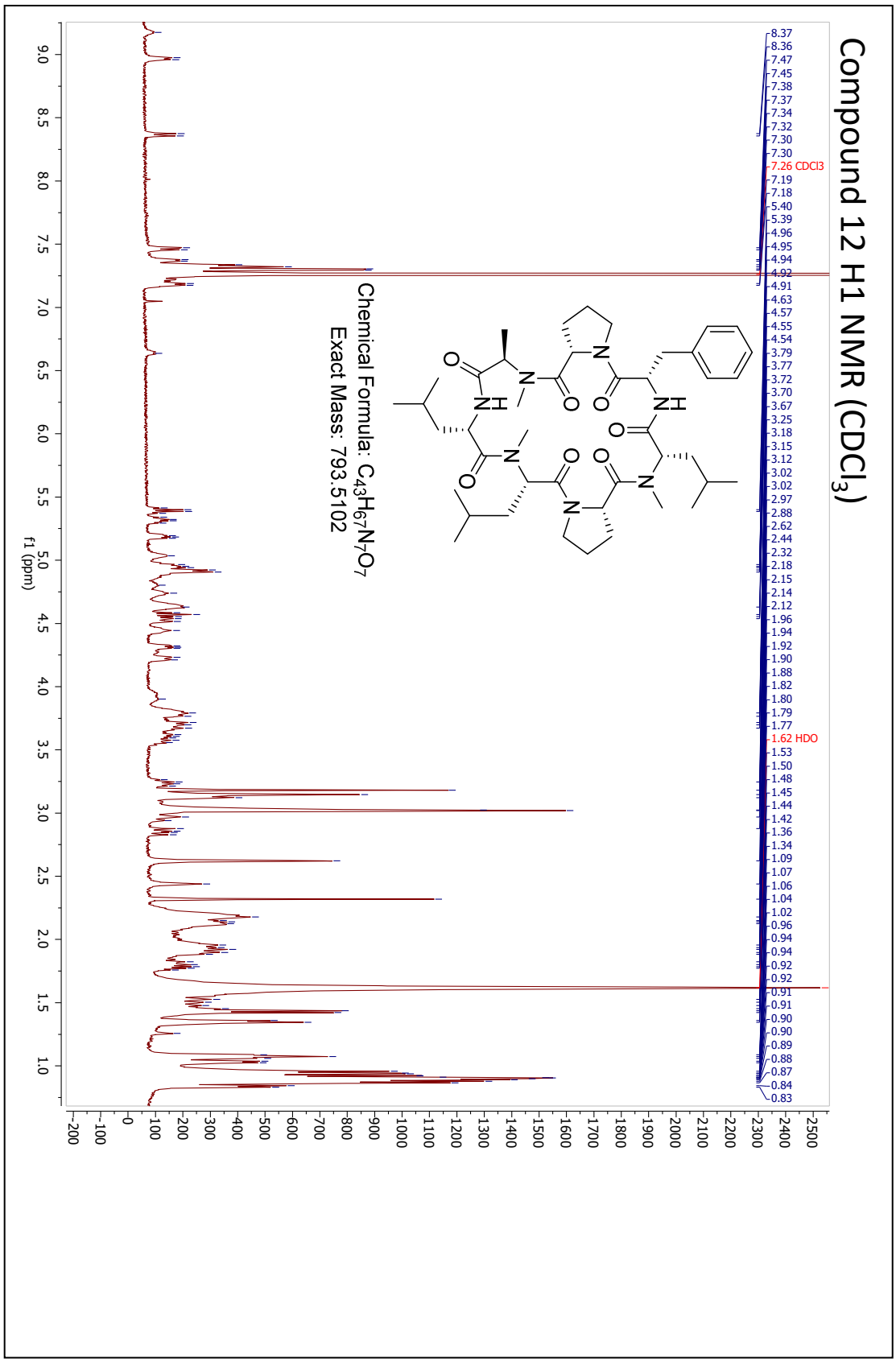


Figure S2.8.12 Compound 2.12 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.


Compound 12 H1 NMR ($CDCl_3$)
Chemical Formula: $C_{43}H_{67}N_7O_7$
Exact Mass: 793.5102
f1 (ppm)
7.26 CDCl3
1.62 HDO

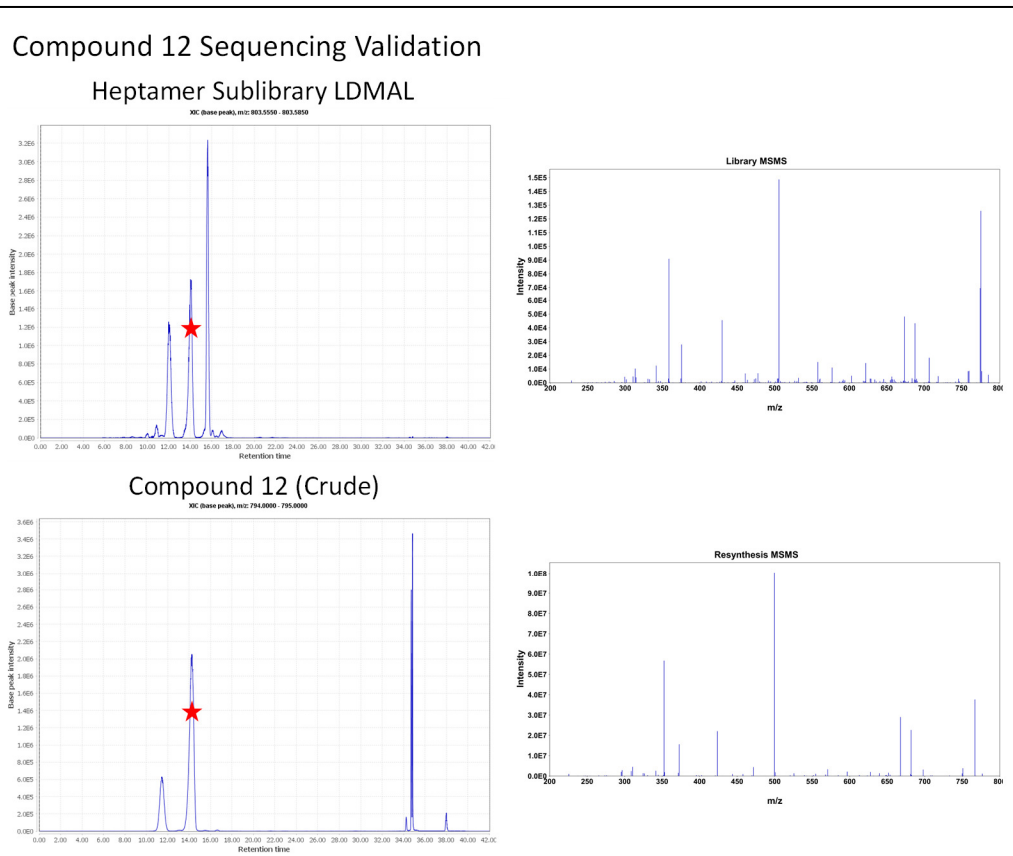


Figure S2.8.12 Compound 2.12 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound. Many peaks in the library spectrum are much weaker in the resynthesis spectrum, but still present.

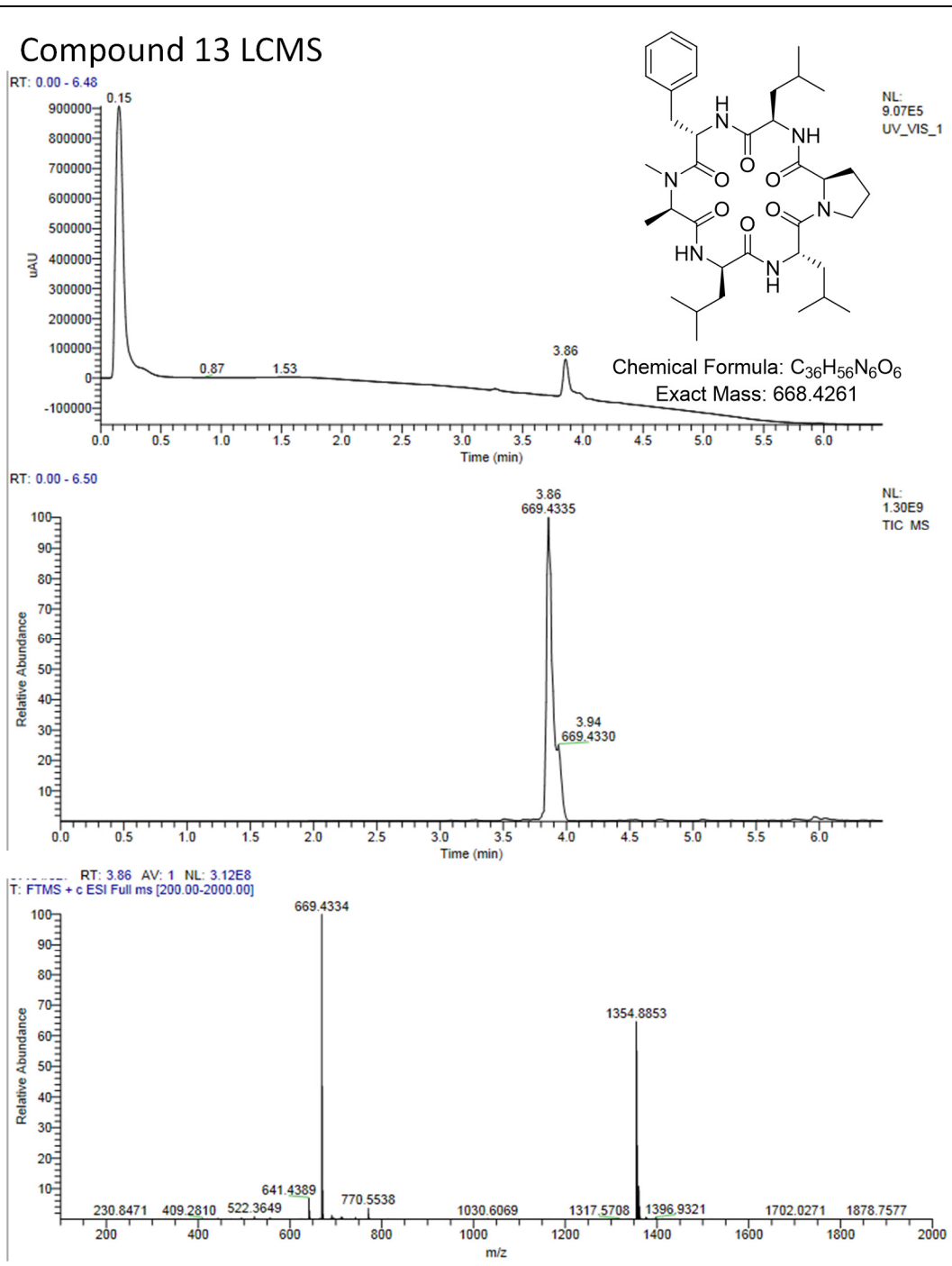
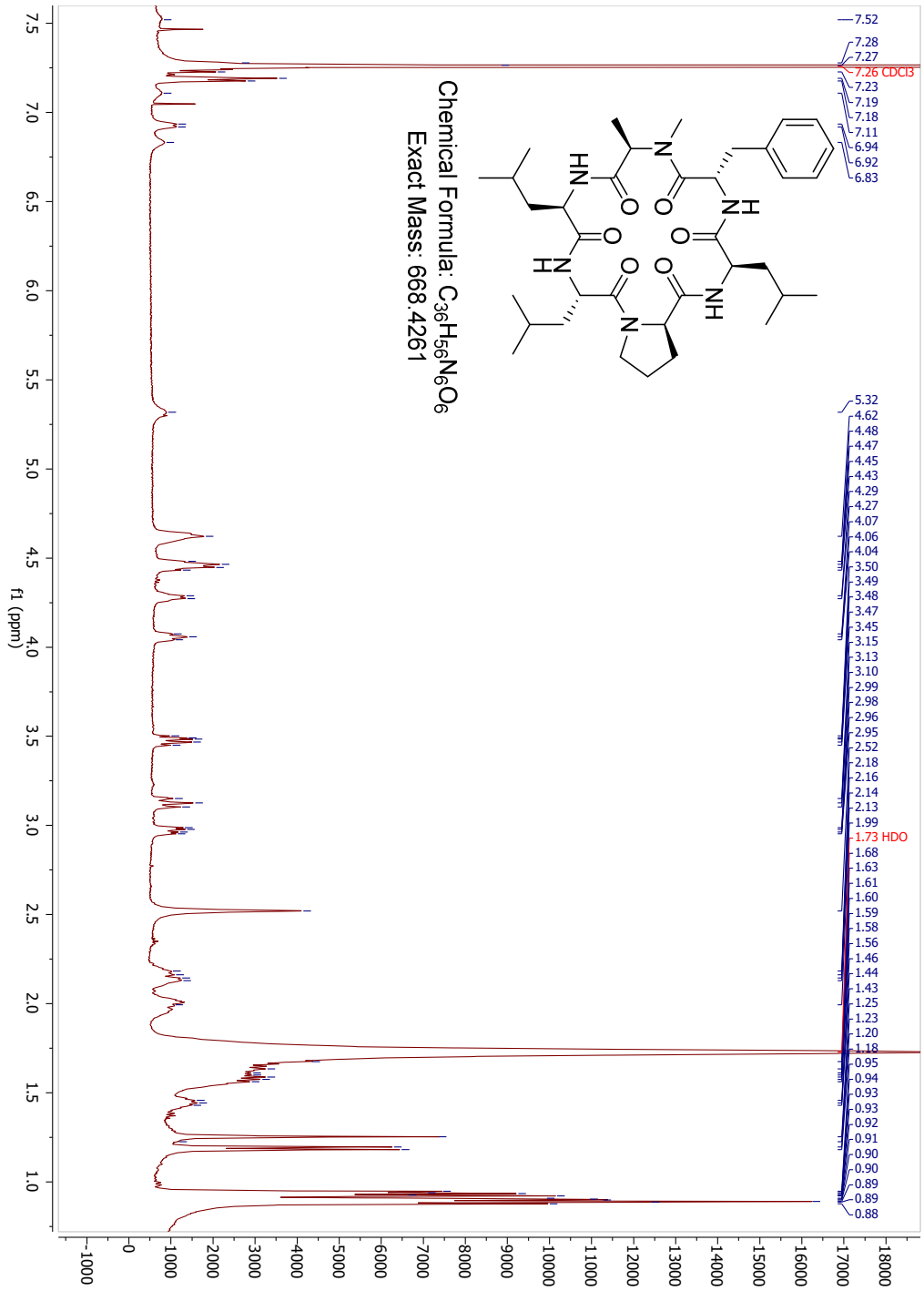


Figure S2.8.13 Compound 2.13 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.13.

Compound 13 H1 NMR ($CDCl_3$)



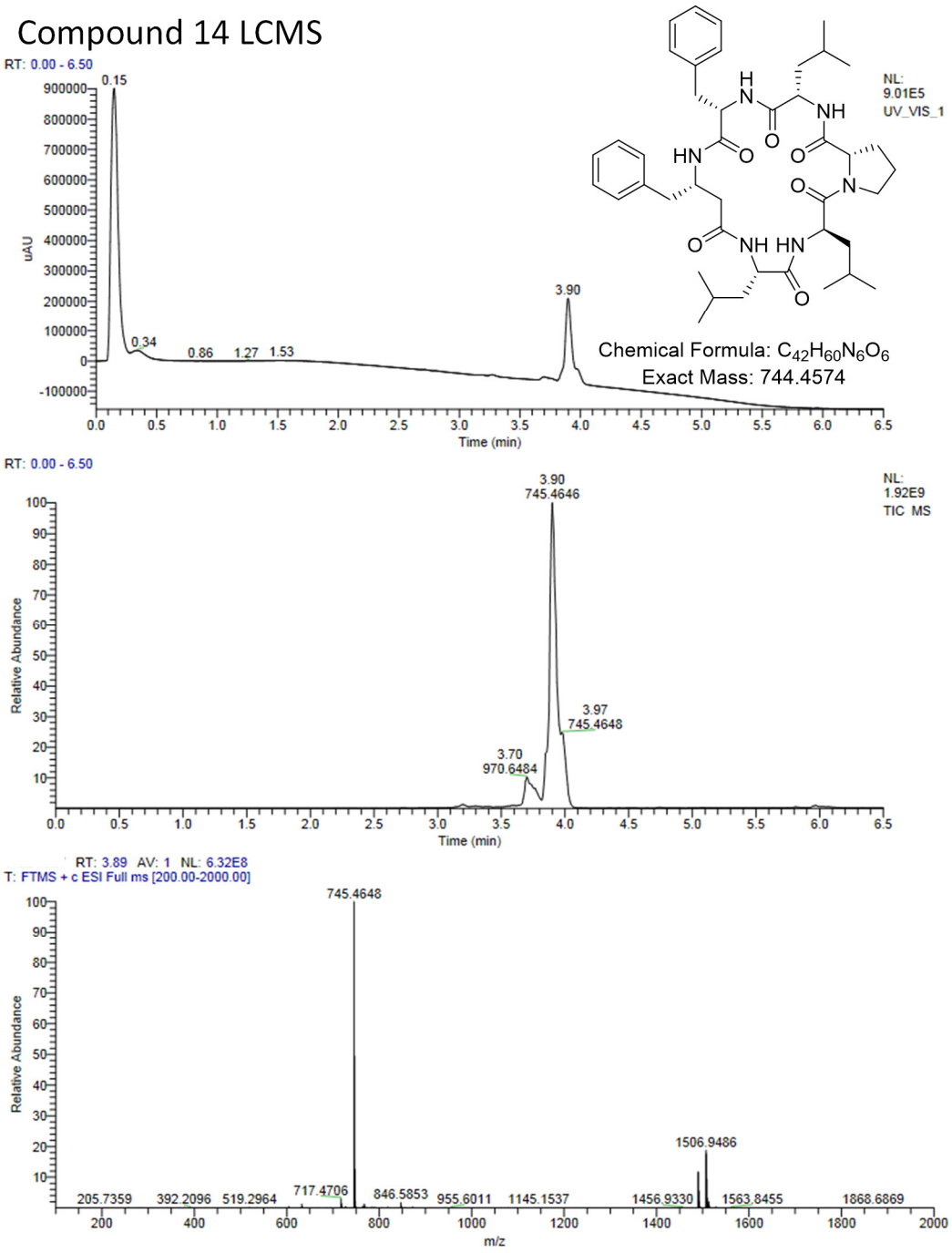
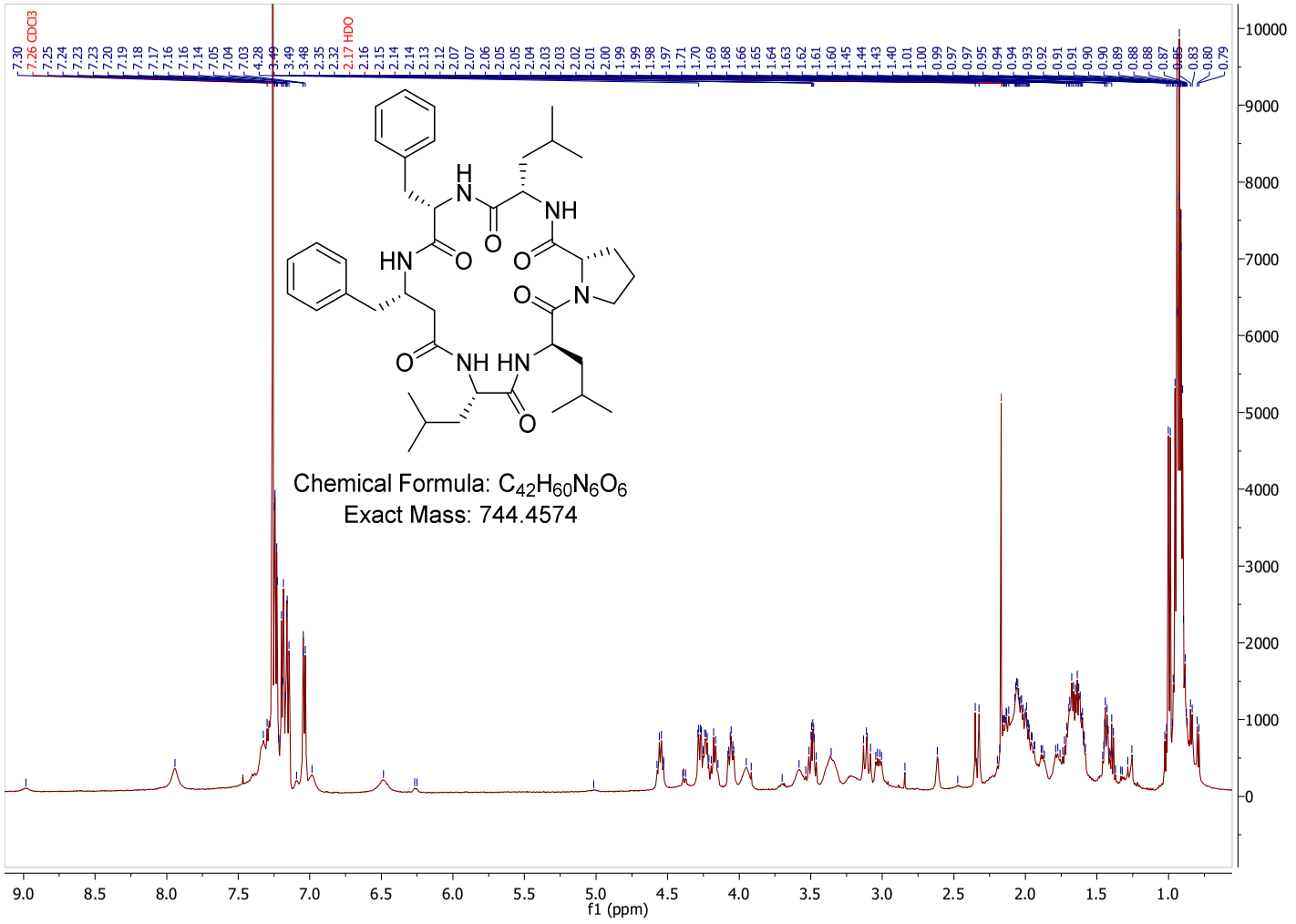


Figure S2.8.14 Compound 2.14 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.19.

Compound 14 H1 NMR ($CDCl_3$)



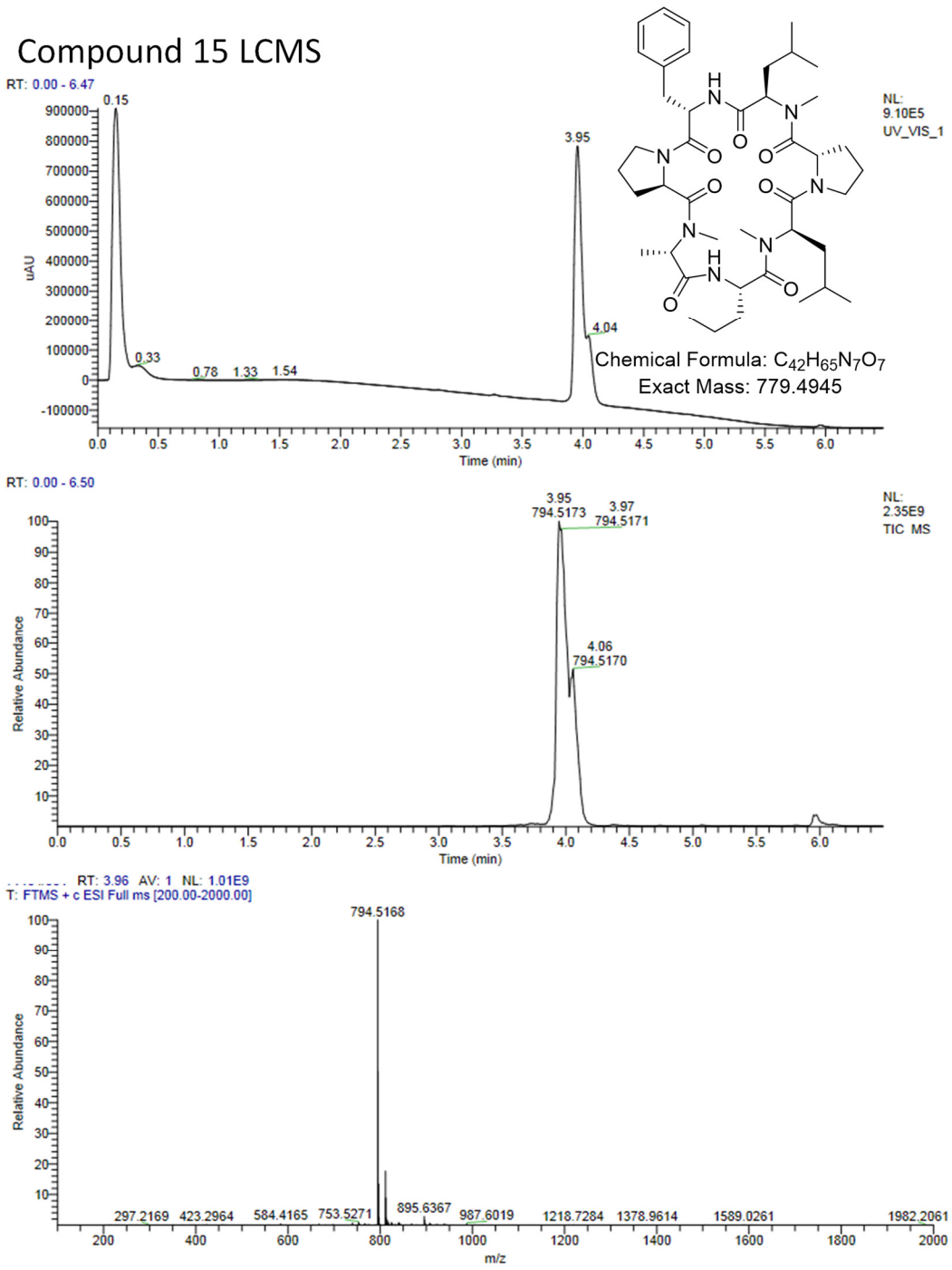
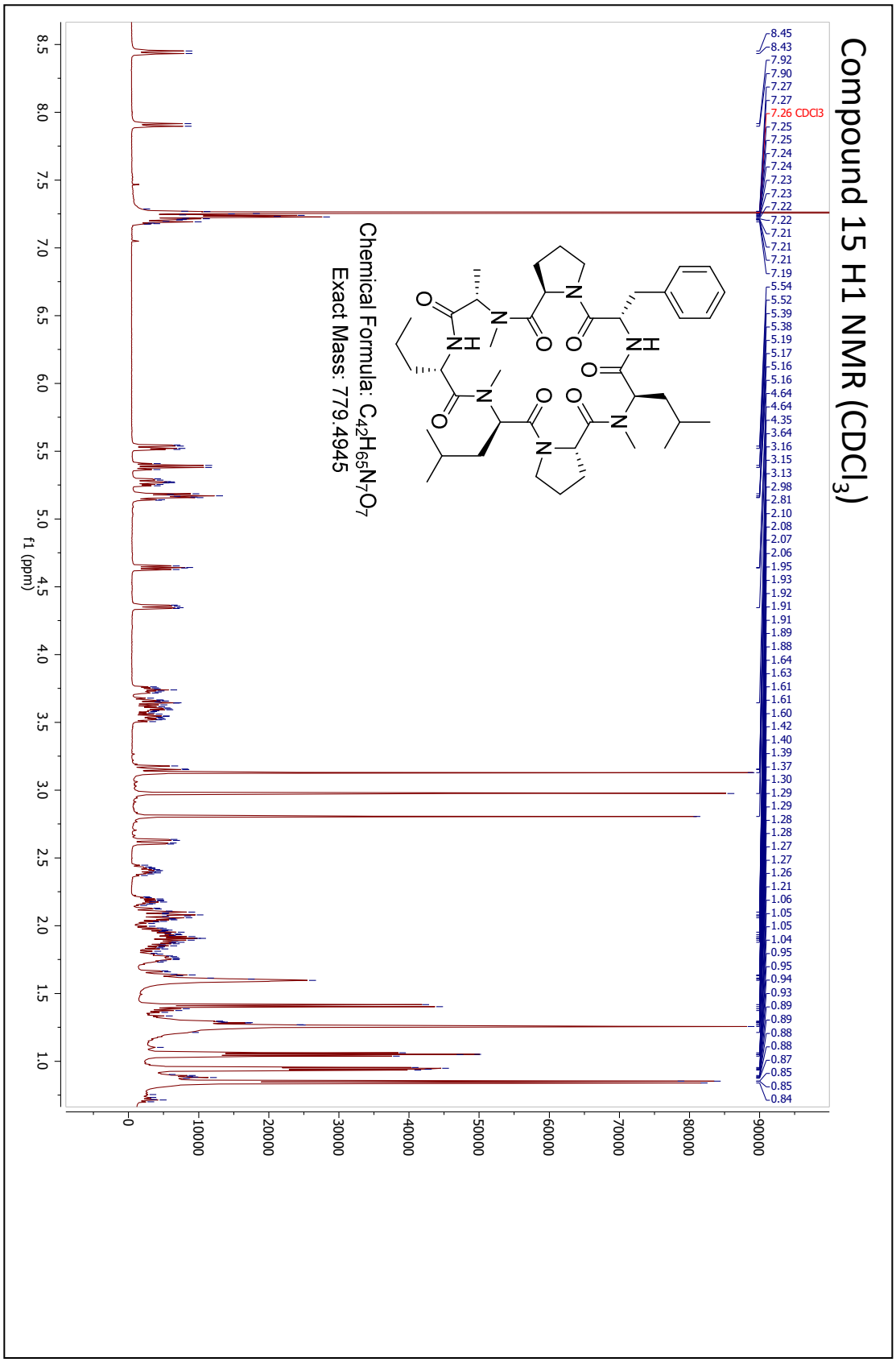


Figure S2.8.15 Compound 2.15 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.


Compound 15 H1 NMR ($CDCl_3$)
Chemical Formula: $C_{42}H_{65}N_7O_7$
Exact Mass: 779.4945
f1 (ppm)
7.26 CDCl3

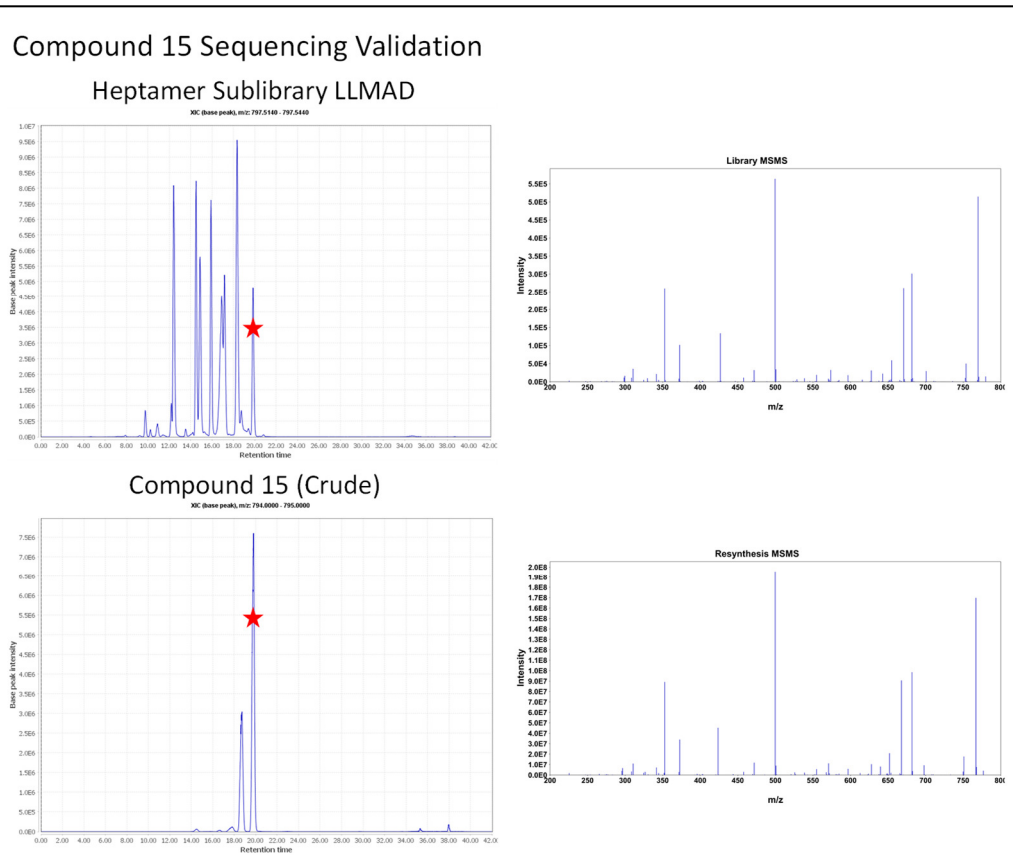


Figure S2.8.15 Compound 2.15 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

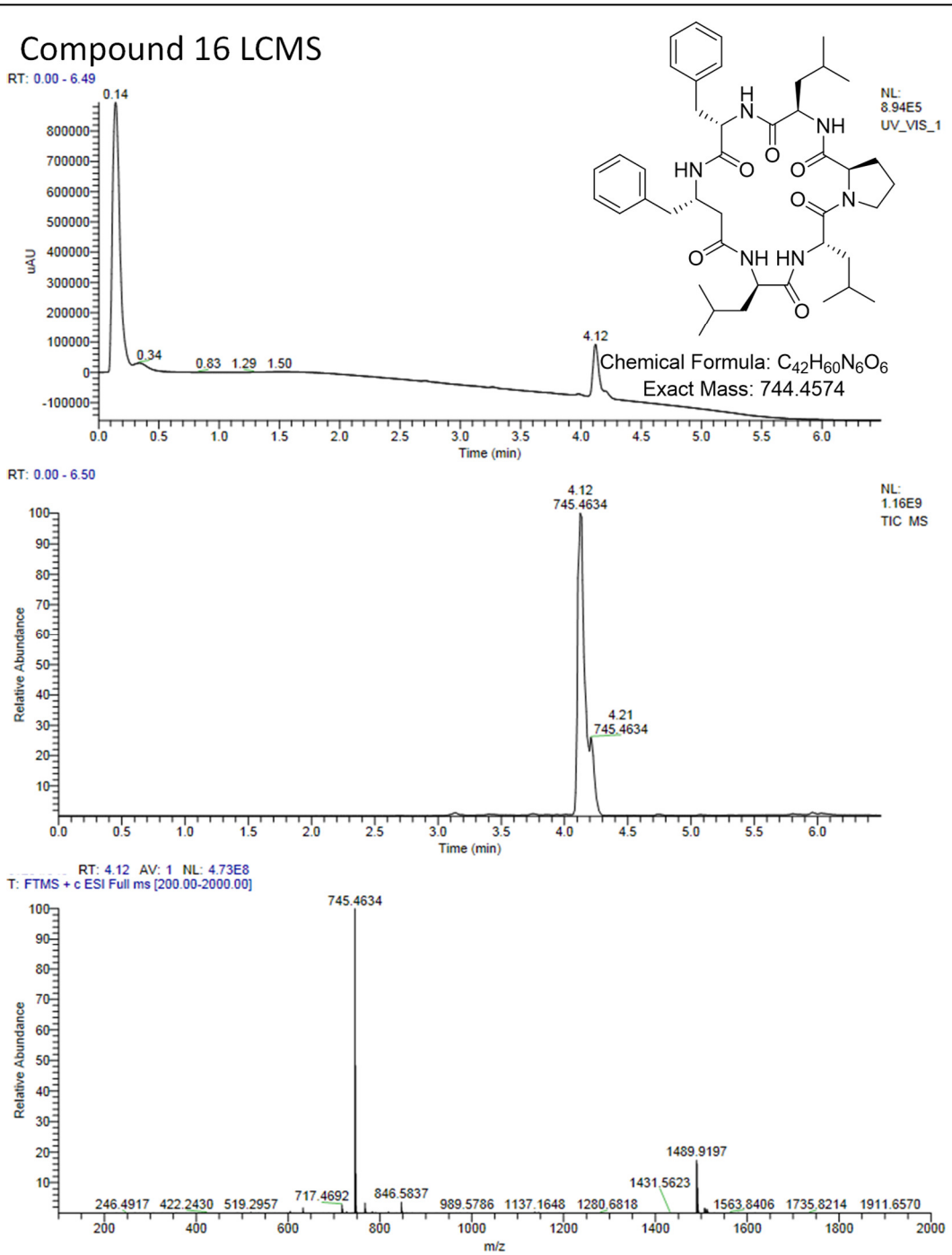
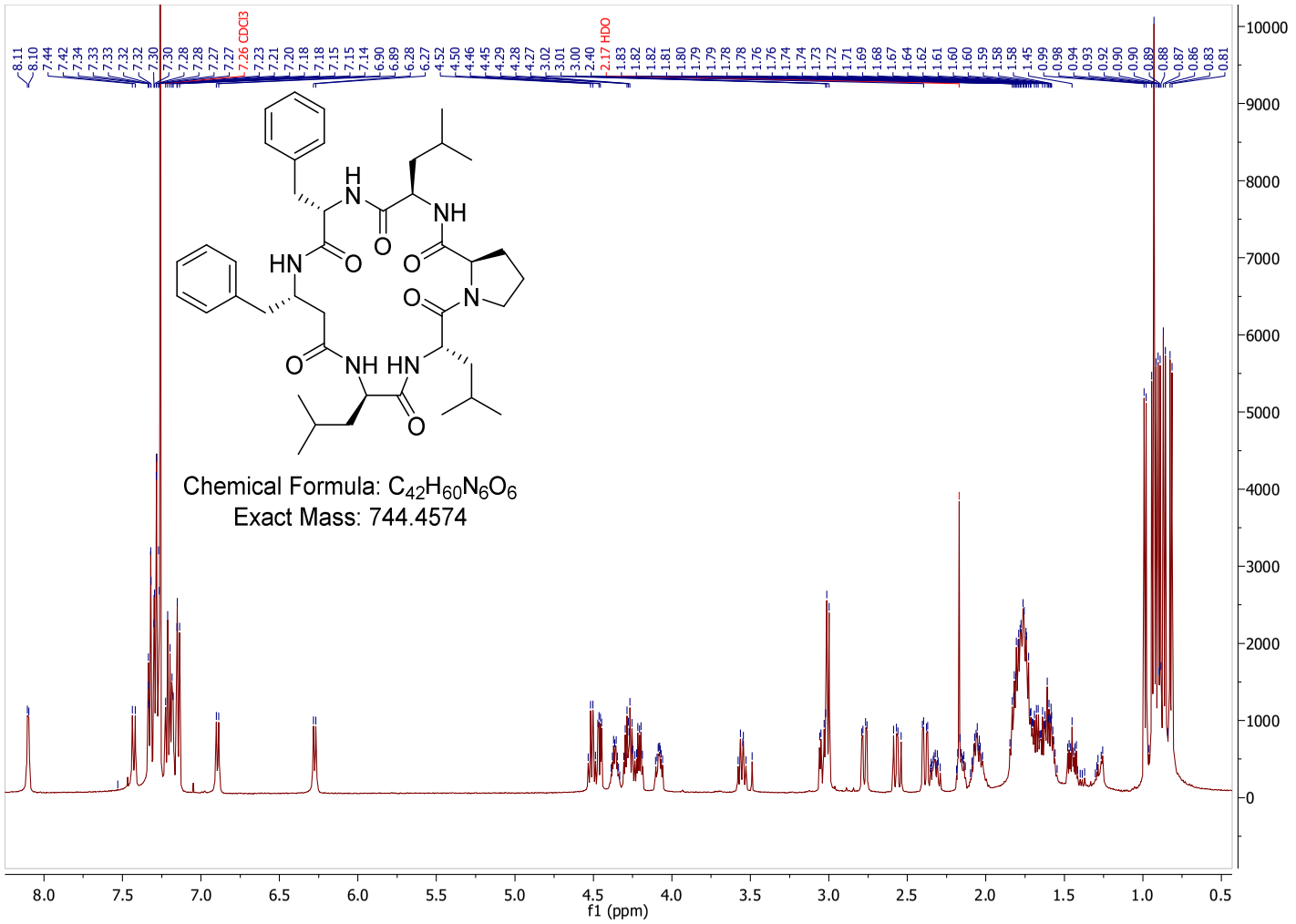


Figure S2.8.16 Compound 2.16 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ follows. Sequencing validation can be found in chapter one as compound 1.20.

Compound 16 H1 NMR ($CDCl_3$)



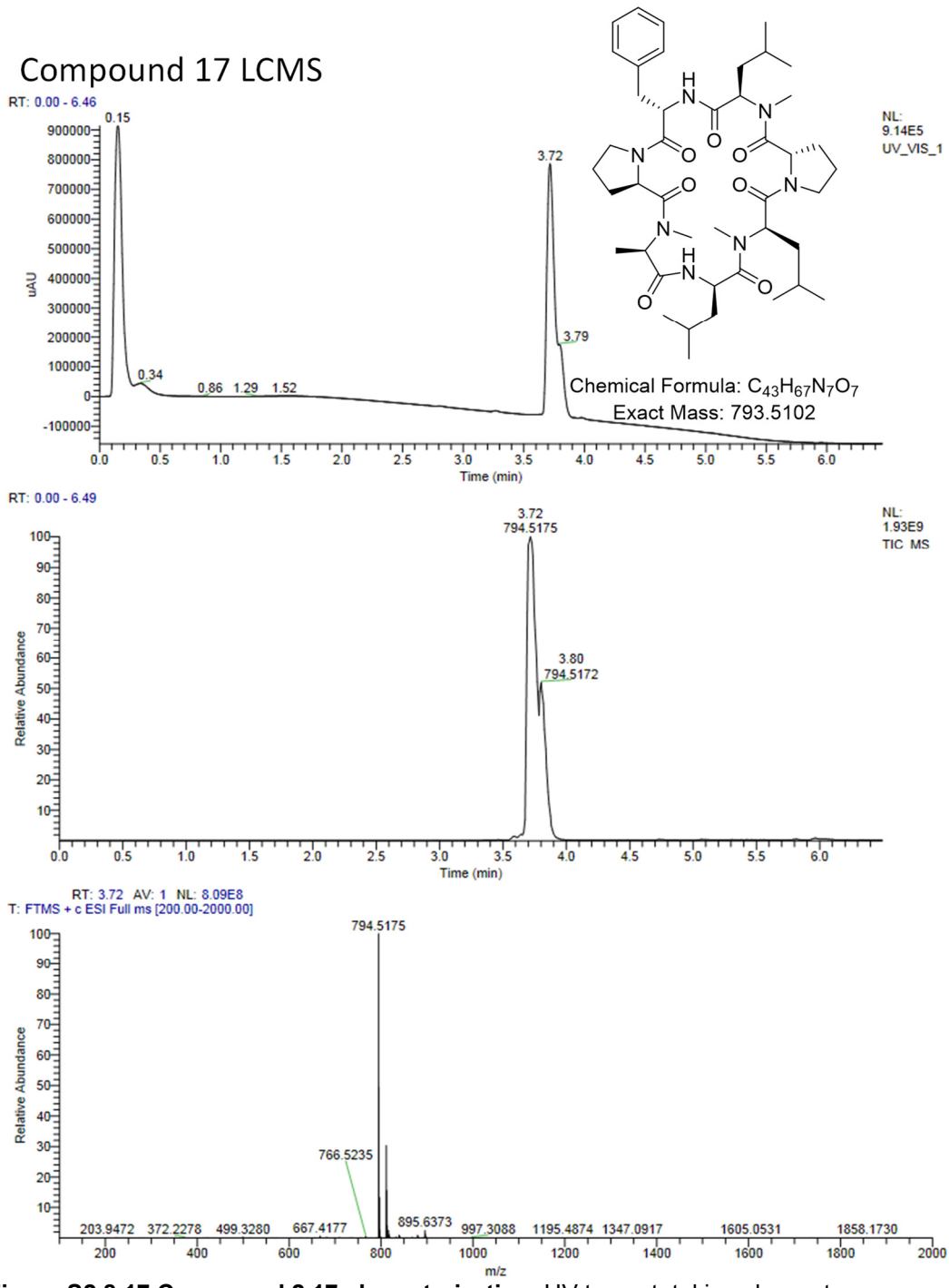
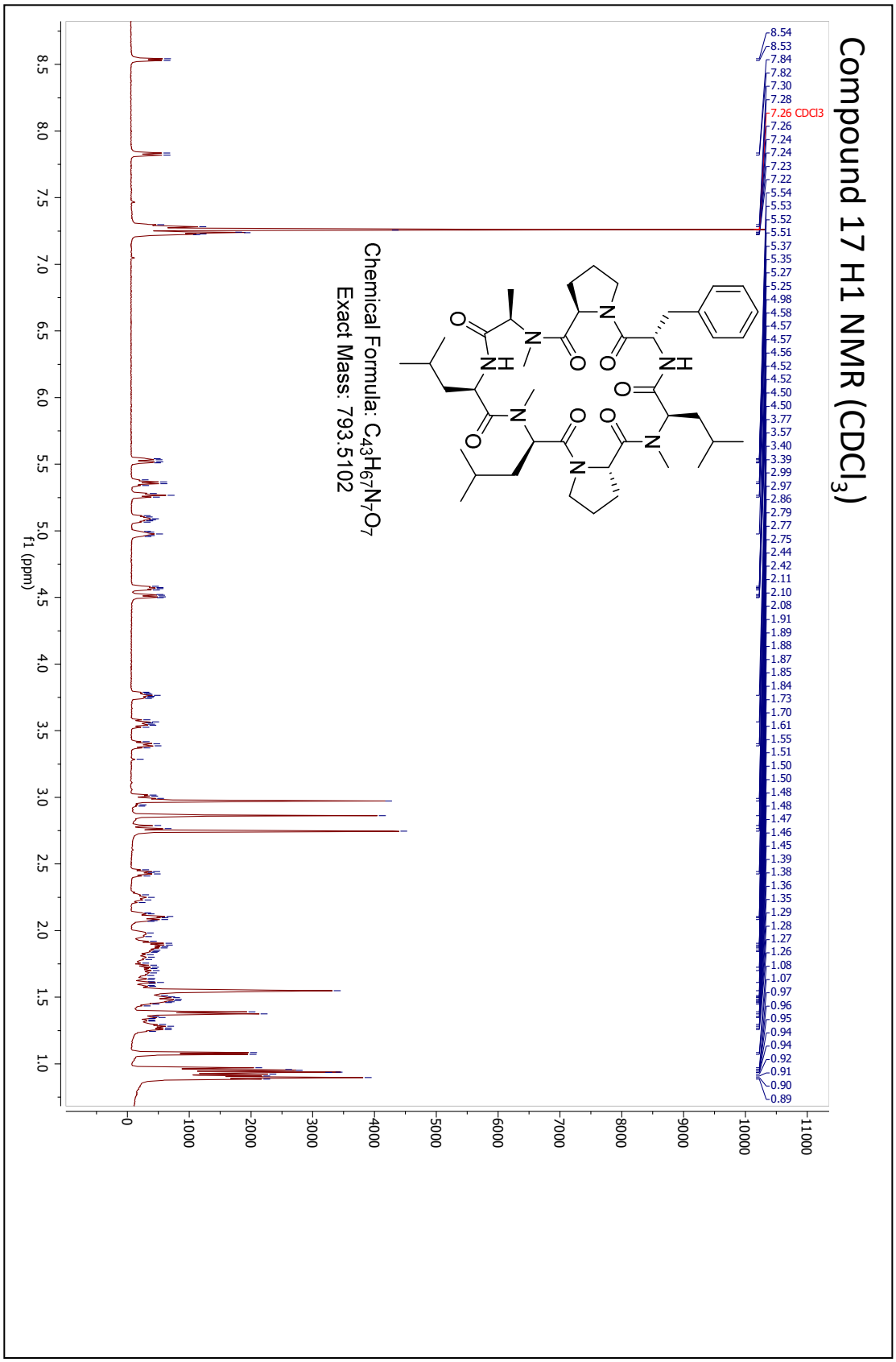


Figure S2.8.17 Compound 2.17 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.


Compound 17 H1 NMR ($CDCl_3$)
Chemical Formula: $C_{43}H_{67}N_7O_7$
Exact Mass: 793.5102
f1 (ppm)
7.26 CDCl3

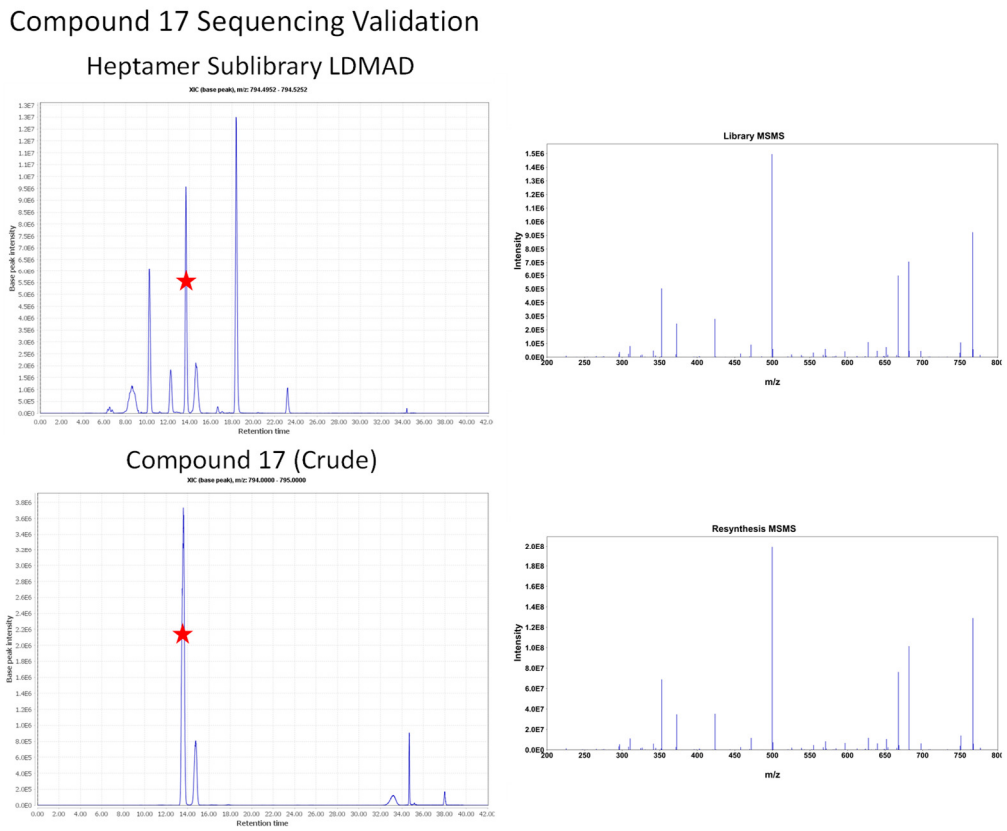


Figure S2.8.17 Compound 2.17 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

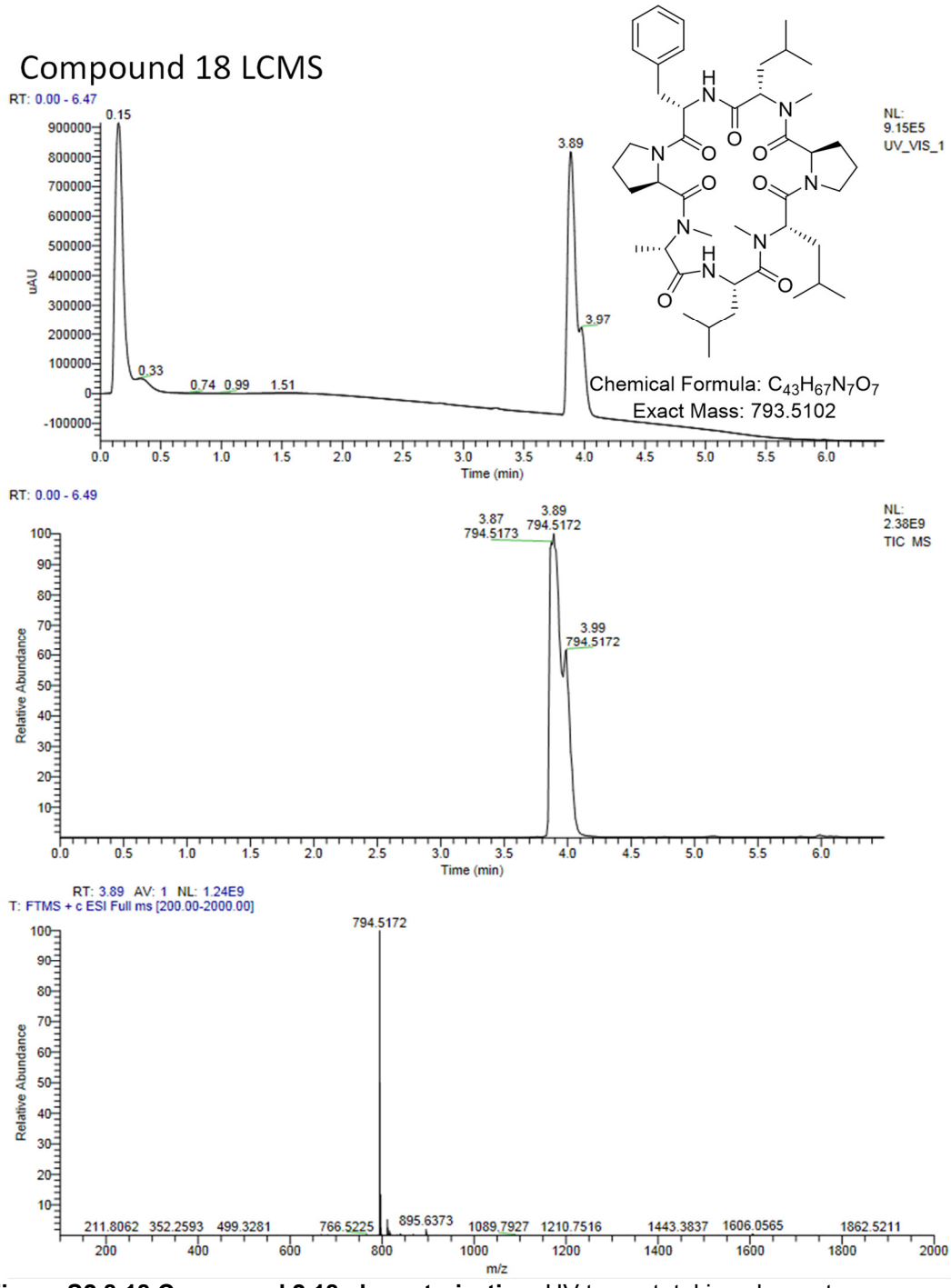
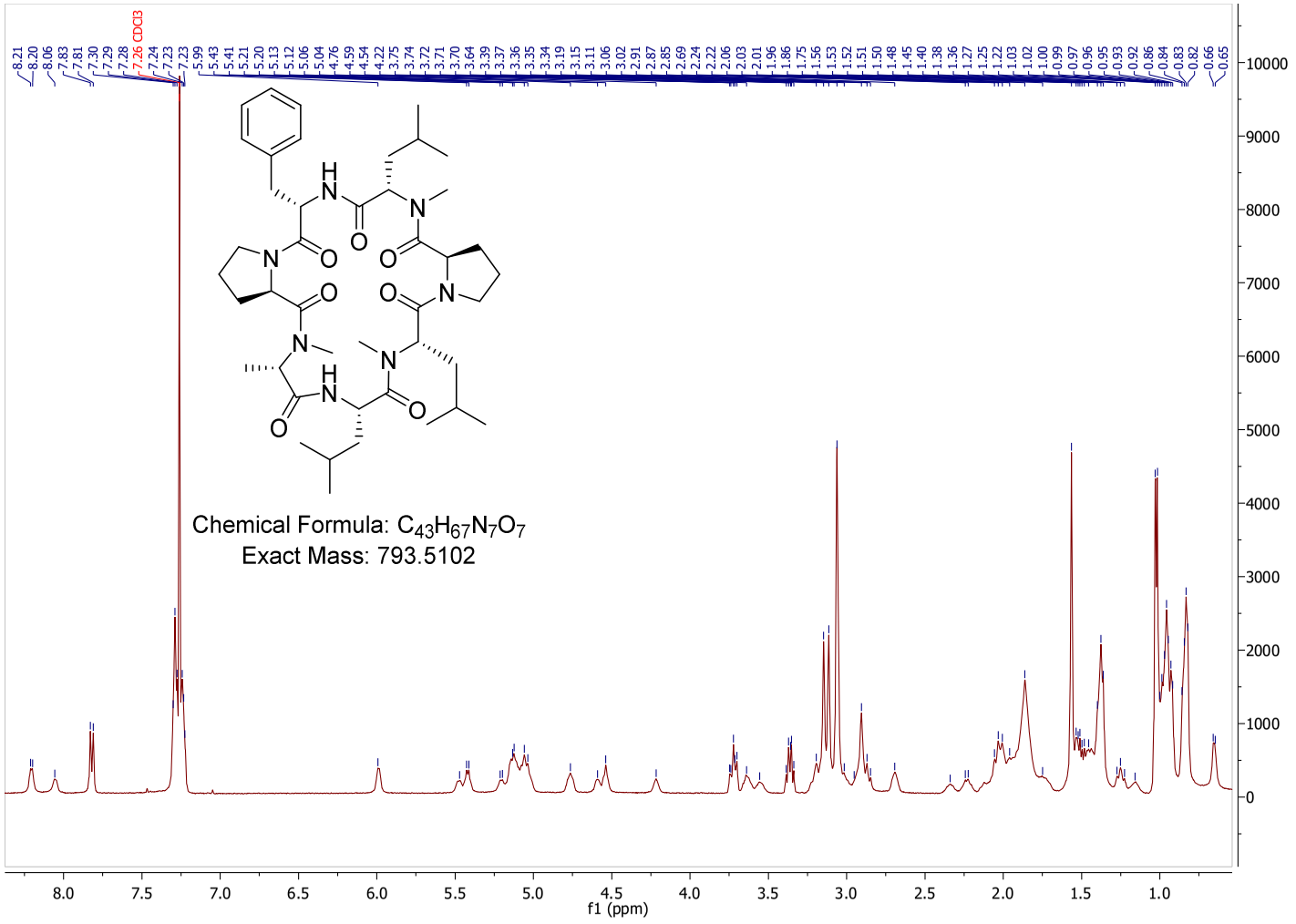


Figure S2.8.18 Compound 2.18 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 18 H1 NMR ($CDCl_3$)



Chemical Formula: $C_{43}H_{67}N_7O_7$
Exact Mass: 793.5102

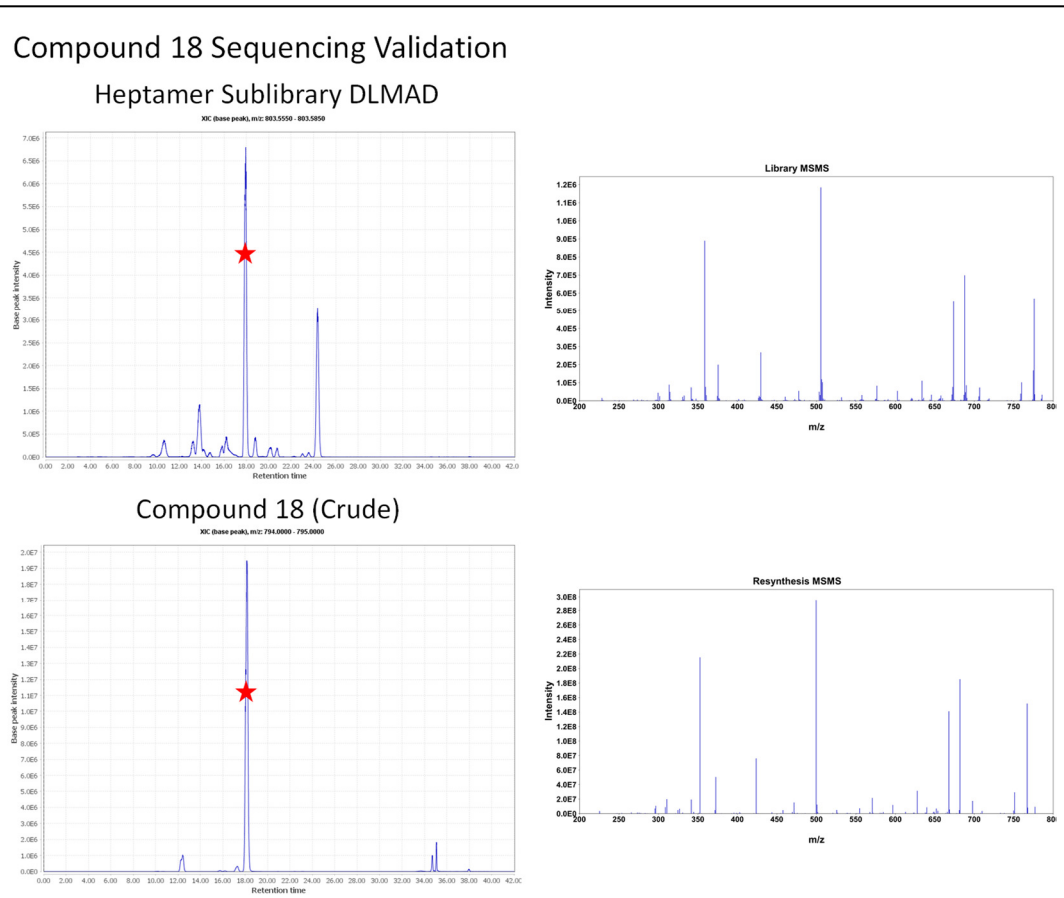


Figure S2.8.18 Compound 2.18 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

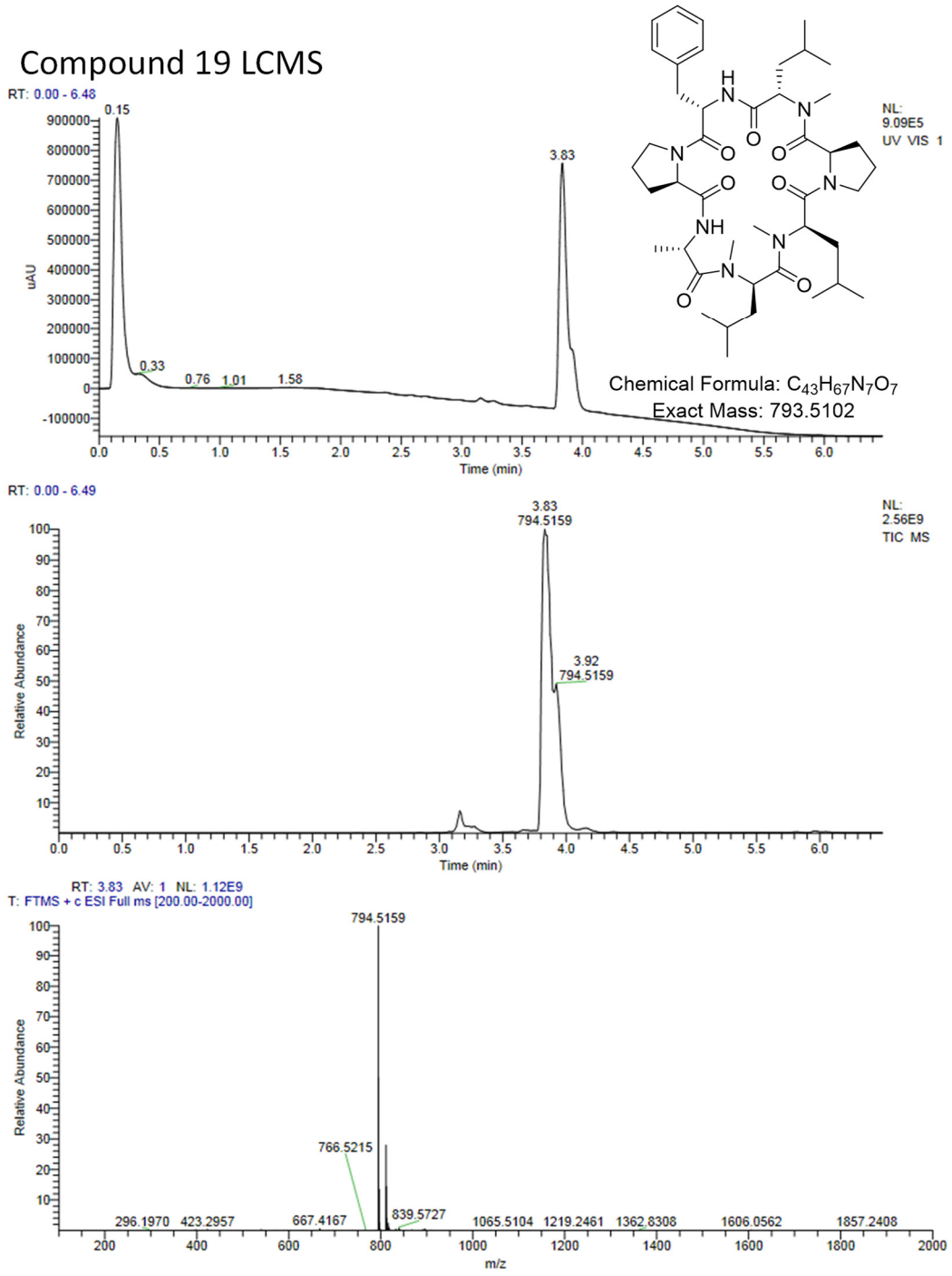
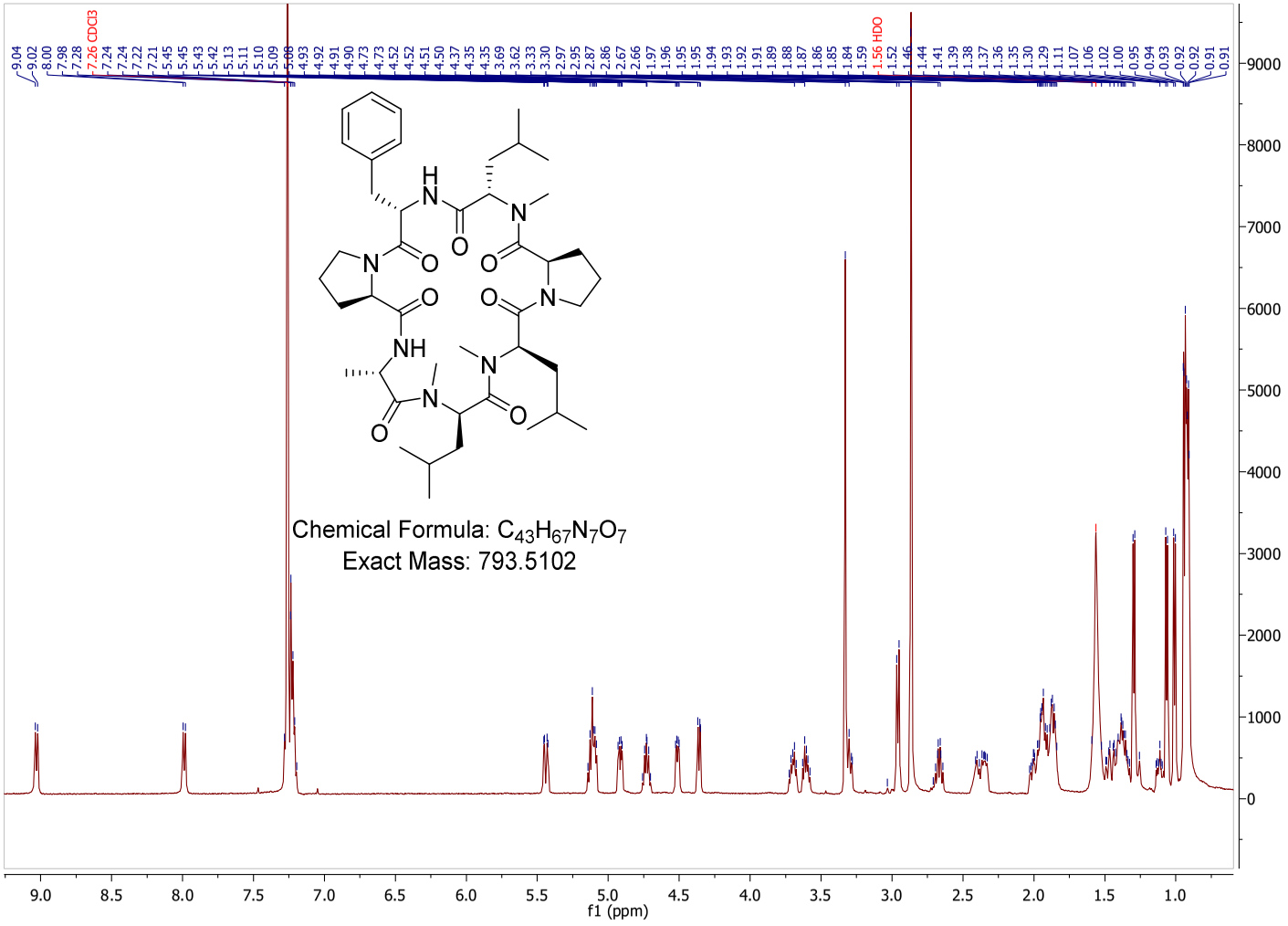


Figure S2.8.19 Compound 2.19 characterization. UV trace, total ion chromatogram, and spectrum of the product peak. A proton NMR spectrum in $CDCl_3$ and a comparison of retention time and MS^2 spectra between the library and resynthesized compound data follow.

Compound 19 H1 NMR ($CDCl_3$)



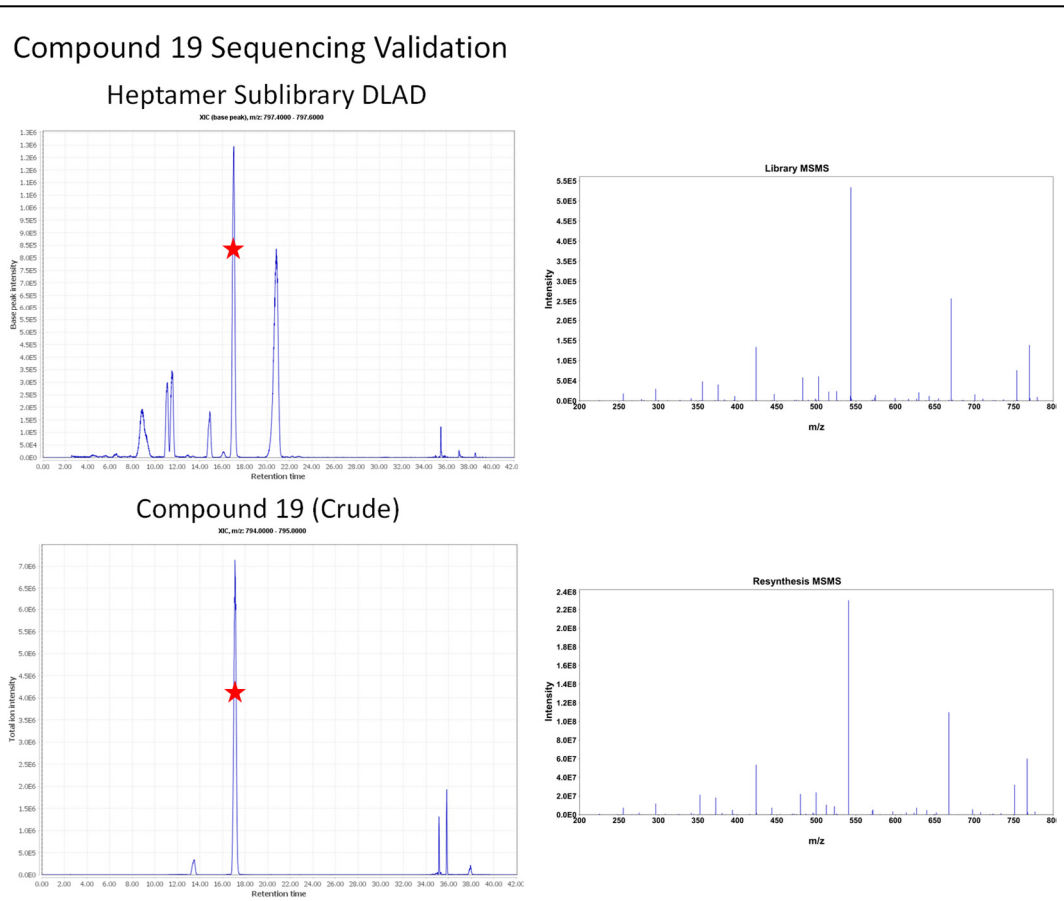


Figure S2.8.19 Compound 2.19 sequencing validation. A comparison between the retention times of the library peaks and resynthesized compounds, with the peaks of interest marked by a read star. The MS^2 spectra of the marked chromatographic peaks are displayed to their right. A high degree of similarity is expected between peaks if the resynthesized sequence is correct, though they will not be identical due to the absence of tri-deuterated L-leucine in the resynthesized compound.

2.8 Acknowledgements

We would like to acknowledge Colin Kelly and Victoria G. Klein for providing valuable feedback throughout this project.

2.9 Author Contributions

Chad Townsend, Eva Jason, and Quinn Edmondson synthesized all libraries. Chad Townsend and Eva Jason synthesized all individual compounds. Chad Townsend, Matthew Naylor, Cameron Pye, Joshua Schwochert, and Akihiro Furukawa developed PAMPA and tandem mass spectrometry LCMS methods. Chad Townsend wrote in-house python scripts assisted by Eva Jason and with advice from Cameron Pye and Joshua Schwochert. Chad Townsend performed all data analysis. Chad Townsend and Scott Lokey wrote the manuscript that became this chapter.

2.10 Funding Sources

We gratefully acknowledge NIH grant GM131135 for funding.

2.11 Associated Content

The associated spreadsheets “Curated Hexamer Permeability Data.xlsx” and “Curated Heptamer Permeability Data.xlsx” contain all curated library data used for this analysis. Associated spreadsheets “Hexamer Motifs by NH Count.xlsx” and “Heptamer Motifs by NH Count.xlsx” contain a summary of all passive permeability motifs identified. Further information on the software pipeline used for this analysis including capabilities, installation and usage instructions, and GitHub address can be found in Appendix A. Complete code is contained in the associated files “CycLS.py”, “AutoPAMPA.py”, and “RTMerge.py”.

Appendix A

Automated multiplex PAMPA data processing and peak alignment for association of permeability and sequencing data originating from CycLS

This chapter contains text and figures from the following manuscript: Townsend, C. E.; Naylor, M.R.; Jason, E.; Pye, C.R.; Furukawa, A.; Schwochert, J. A.; Edmondson, Q.; Lokey, R. S., The passive permeability landscape around geometrically diverse hexa- and heptapeptide macrocycles. *(manuscript in preparation)*

A.1 Introduction

Completion of my efforts to empirically explore the PAMPA permeability of thousands of cyclic peptides required the development of new analytical capacities, the first of which was sequencing of cyclic peptides by tandem mass spectrometry via CycLS. However, the analytical bottleneck of PAMPA data processing also required a solution. Completion of AutoPAMPA has reduced processing time from hours per thousand peaks to minutes per thousand peaks in both PAMPA and shake-flask partition coefficient experiments. Merging the results of CycLS and AutoPAMPA via a retention time matching script (RTMerge) allowed me to complete a data processing pipeline which has since been used to explore the permeabilities of lariat peptides and cyclic peptides of larger ring sizes.

In addition to use in our own lab, publishing my work on GitHub has allowed these programs to benefit others as well. CycLS is currently being used by the Knight group at University of North Carolina, Chapel Hill to sequence mixtures of linear peptomers. AutoPAMPA has seen interest from startups seeking to increase the throughput of their own PAMPA data analysis and its general capacity for batch-processed peak integration. This appendix contains installation and usage instructions for CycLS, AutoPAMPA, and RTmerge along with additional information on AutoPAMPA and RTMerge and a summary of the revisions to CycLS since its publication.

A.2 CycLS

CycLS is a program designed to identify individual cyclic peptides from a library of known design by interpreting tandem mass spectrometry data. This allows for assays and analysis to be performed on complex mixtures containing some degree of mass redundancy without adding library design constraints. CycLS can also be used to sequence linear peptides or peptomers, taking advantage of the b- and y-ion series for increased accuracy. Additionally, CycLS can be set to sequence truncations or generate E-values for scores for amenable library designs. Finally, when attempting to adapt CycLS to your use case, the query mode allows exploration of individual MS^2 spectra alone or matched against any viable sequence both textually and visually to quickly diagnose problems with peak matching.

A.2.1 Installation

The easiest way to get the packages required to run CycLS is to install the Anaconda Python distribution (version 4.3.1 tested) from Continuum Analytics, then install RDkit (version 2016.09.4 tested), peakutils (version 1.0.3 or higher), openpyxl (version 2.4.1 tested), statsmodels (version 0.6.1 tested), seaborn (version 0.7.1 tested), and pymzml.(version 0.7.7 only). Use the "conda install package-name" command to install all those packages except RDkit, peakutils, and pymzml. See the RDkit GitHub page for installation instructions. Using “pip install package-name” for peakutils and pymzml is the easiest way to add those packages to an Anaconda installation if they are not present in conda channels, and it may be necessary to install pymzml 0.7.7 as a clone of the archive pymzml GitHub branch. Aside from pymzml, it is likely that more current versions of these packages are compatible.

A.2.2 Usage

Run CycLS.py via the command line using your python interpreter.

A.2.2.1 Required arguments:

Targets: The first positional argument may be either an asterisk signifying that CycLS is to search all spectra for library members or a comma-separated list of protonated ion exact masses (without whitespace).

Example: '647.523,831.978,745.069' would search only for MS^2 spectra originating from those three protonated ions.

mzML: The second positional argument must contain the path to the mzML format data file to be analyzed.

Constraint: The third positional argument must be a string communicating the library composition. Building blocks within a position are comma-separated, with positions separated by semi-colons (potentially necessitating surrounding the constraint string in double-quotes to prevent their interpretation by the shell). As with the *Targets* field, no whitespace is permitted. Building blocks are defined in the amino acid database file, with single-letter amino acid codes implemented in the example database provided.

Example: 'L,A,D;E,Q,K;P,G,R' is a tripeptide with three possibilities for building blocks at each position.

A.2.2.2 Optional arguments:

-d, --database: Sets the name of the residue name and SMILES string database to be used for library generation, defaulting to 'aadatabase.txt'. An example database has been included in the repository. Though it is not currently necessary for CycLS, we have rearranged the SMILES strings for each residue such that they begin with the N-terminus of the residue and end with the C-terminus of the residue. Note that any SMILES string can be defined as a 'residue', and some non-amino acid examples are present in the provided residue database.

Example: L-alanine has been defined as the line: 'A N[C@@H](C)C(=O)O'.

-e: Activates expected value calculation for the scores of each candidate molecule by generating a decoy database and scoring it. This option significantly increases run time due

to the increased number of candidate evaluations necessary per spectrum. Thus far, decoy database sizes have been under 1000, leading to unreliable expected values and making testing of this functionality difficult.

-l, --linear: Use this flag if the library to be sequenced is linear. CycLS was first tested against linear peptides in its development and is capable of sequencing linear libraries at least as well as it sequences cyclic peptides.

-n: Sets an intensity threshold below which all peaks in the processed MS^2 spectra are assumed to be noise and thrown out. If above 1.0, the number provided is the intensity cutoff used. If below 1.0, the number provided is treated as the maximum probability allowable probability of a peak being due to noise. In this case, a noise threshold is automatically generated to fulfill that condition. Defaults to 100.0, which may not be a useful threshold for your mass spectroscopy system.

-o, --out: Sets the prefix of the output file and defaulting to 'Sequencing'. The default value results in the two output files 'Sequencing_Out.xlsx' and 'Sequencing_Results.xlsx'. Output is given in full in the 'Out' file and summarized in 'Results' file.

-p: Sets the precision of MS^2 and MS^1 spectra m/z values in that order, defaulting to 0.3 and 0.02, respectively. These values may not be appropriate for HRMS set-ups and may need adjustment for your mass spectrometry system. The two values are entered comma-separated, and without spaces (Ex: '0.3,0.02').

-q, --query: Activates an interactive results inspection mode after completing normal operations. Allows inspection of combined MS^2 spectra alone or against a compound to inspect virtual fragment matches through text or graphically. Further usage instructions are supplied interactively.

-r, --rules: Allows filtering of the generated library of compounds based upon position-independent constraints using a residue name, a comparison operator, and an integer value. Alternatively, 'AlogP' (via RDKit) and 'MolWeight' may be used to compare to those properties

instead. Comparing to AlogP gives erroneous results for libraries containing residues that are not amino acids. Incompatible with the -t argument.

Example: '-r "G < 3;R > 2;AlogP > 2.27"' would allow only compounds with less than 3 glycine residues, greater than 2 arginine residues, and AlogP greater than 2.27 to continue forward. The three components of a rule are separated by a space, with rules separated by semicolons, and the entire field surrounded by double-quotes to prevent the shell from interpreting the spaces and semicolons.

-s, --scanranges: Sets the minimum and maximum scan numbers to be considered, with one scan range expected per target given (and assigned to targets in order), or a single scan range if the target was '*'.

Example: If the targets 654 and 708 had been entered and '-s 175-354,378-412' would yield assign scan range 175-354 to target 654 and scan range 378-412 to target 708. As observed here, there should be no spaces and multiple scan ranges are comma-separated.

-t, --truncate: Enables the generation of all possible synthetic truncations of the library resultant from incomplete peptide couplings in addition to full length library members. May significantly increase run time for large libraries. Incompatible with rule usage (*-r*) in its current implementation.

-u: Sets the number of worker processes allowed to CycLS during multi-processing operations, defaulting to the number of CPU cores minus one.

-v, --verbose: Sets verbosity level, defaulting to zero. Prints general status announcements to the terminal at level 1 or higher.

A.2.2.3 Input File Preparation:

CycLS uses the pymzml package to read in spectra data from mzML format files. We suggest using Proteowizard's msconvert program to convert mass spectrometry data to the mzML format. We used the command line version of msconvert from Proteowizard version 3.0.10577 with some modifiers to strip unnecessary data:

```
msconvert --simAsSpectra *.raw --32 --zlib --filter "peakPicking
true 1-" --filter "zeroSamples removeExtra"
```

Including UV data causes pymzml version 0.7.7 to crash and should therefore be avoided.

A.2.2.4 Interpreting Output:

CycLS outputs a Results file and an Out file, with varying metrics included. Both files include all information necessary to locate the MS^2 spectra used, including mass, retention time, and the scan number (or scan numbers for spectra which were combined).

Out

The Out file includes the score and its components (Unique Matches, Redundant Matches, and Percent Intensity Matched) for each candidate to each (combined) MS^2 spectrum. Unique Matches represents the count of initial fragments (those present pre-neutral loss) to which any match was found, including all neutral losses which originated from an initial fragment. 0.1 is added to the Redundant Matches field for each match beyond the first traced back to a given initial fragment. The Percent Intensity Matched is the sum of the intensity of the peaks for which there was at least one fragment match divided by the total intensity of the spectrum. In the simple case, the equation below is followed, where M_u is the count of unique matches, M_r is the count of redundant matches, and I_m is the fraction of intensity matched.

$$S(M_u, M_r, I_m) = \frac{M_u + \frac{M_r}{10}}{10^{(1-I_m)}}$$

Results

The Results file includes the sequence of the top candidate to each spectrum, the top score, the next best score (if there were multiple candidates), the average score of all candidates, the number of candidates, a sequencing confidence metric (discussed below), and any expected value-related statistics.

The magnitude of a score can only be directly compared to the score of other candidates against the same spectrum with confidence; despite this, it has been observed that higher top scores relative to the rest of the same library can be used as an indicator of sequencing confidence. The normalized difference between the top and next best scores is a superior indicator of sequencing confidence because it expresses the ambiguity of the MS^2 spectrum and because it is comparable between libraries. Top sequences with higher normalized score differences are more likely to be correct, though some correctly sequenced spectra are discarded at any threshold. Low normalized score differences between similar sequences often signifies that the correct composition of the compound in question has been determined, but crucial evidence on the sequence at one or more sites is missing. In such cases, the correct sequence is usually among the top three candidates for the spectrum.

A.2.2.5 Known Bugs and Issues:

Isotopes are handled via asterisks in the constraint string and represent deuterium only: L*** represents triple-deuterated L-Leucine using the SMILES representation of L-Leucine and interpreting the asterisks on the fly. It is likely that not all neutral losses are accounted for, though the most common are.

A.2.3 CycLS Revisions

Although the core functionality of CycLS has not changed since publishing chapter one, some minor additions have been made. In addition to fixing bugs (that did not affect my results), sequencing confidence has been added to the output Results spreadsheet among other output formatting changes. Additionally, the data exploration mode (-q) has been improved to aid new users in troubleshooting poor results.

A.2.4 Associated Content

A copy of CycLS.py and an amino acid database text file accompany this dissertation and can also be found at https://GitHub.com/LokeyLab/CycLS.

A.3 AutoPAMPA

AutoPAMPA is a python script designed to process data from PAMPA, partition coefficient, or general chromatographic data acquired from complex mixtures to ease data analysis and increase practical throughput. It accepts mzML format raw data and performs peak-finding, peak bounding, and integration, and peak alignment across paired wells by retention time. Parameters are mainly controlled by a configuration excel file and data is output in excel format. Visual inspection of the automated integration is accomplished by optional generation of vector graphics files showing stacked chromatograms of each well for each target mass.

AutoPAMPA has been a great aid in minimizing the hassle of data analysis for assays quantified by mass spectrometry, even those not performed on mixtures. Several iterations can be necessary to tune the bounding parameters to a dataset, but this can be completed via several brief visual inspections over as little as five minutes. The script fails when peaks are highly overlapped and when the baseline is a significant contributor to peak height, but peak overlap can be troublesome even for manual integration.

A.3.1 Installation

AutoPAMPA has identical package requirements to CycLS (less RDKit) and was tested on the same package version numbers.

A.3.2 Usage

Run AutoPAMPA.py via the command line using your python interpreter.

A.3.2.1 Required arguments:

config: The only positional argument must contain the file path to a specially formatted excel file containing the input parameters for an AutoPAMPA job. An example configuration file is included in this repository. A thorough explanation of the parameters contained can be found below.

A.3.2.2 Optional arguments:

-o, --out: Sets the prefix of the output files (defaults to 'Expt'). This prefix will be applied to both output files as well as any directories created by the -g argument.

-g, --graph: Generates stacked extracted ion chromatograms for each target mass for each experiment (well group) and saves them as vector graphics files (SVG). Generated plots have peaks, bounds, and fit curves labeled. Individual peaks are assigned numbers based on their ordering in the reference well and these labels are added to other wells if successfully aligned and above a peak height threshold (to avoid the worst clutter).

-u, --gauss: Activates gaussian integration mode, in which a gaussian is fit to each peak and its parameters used for an exact integration. Default is to simply sum the intensities within the bounds of a peak on the assumption of approximately consistent scan durations (reasonable for simple MS methods in our experience). A code block for taking time differences into account is commented out nearby if this does not hold true for your data.

-m, --msevents: Used to unweave alternating event spectra in the case of a multi-event MS method, resolving what would otherwise be jagged peaks. Defaults to 1.

-i, --ion: Sets the type of ion to search for. Defaults to proton, but accepts "proton", "sodium", "ammonia" or "H+", "Na+", "NH4+". Can also be set to any floating-point value (including negative ones).

-v, --verbose: Sets verbosity level, defaulting to zero. Values 1 or higher causes more messages to be printed and 2 or higher causes the graphic generation to use the smoothed chromatogram (instead of the unsmoothed) to visualize what the gaussian fits are being generated from.

A.3.2.3 Input File Preparation:

AutoPAMPA expects its main input from an excel configuration file. An example file is included as associated content to demonstrate the proper format and as a reference for the

sections below. The excel workbook is composed of four sheets: Parameters, Experiment Type, Wells, and Targets. Input mzML files should be prepared as described for CycLS.

Parameters

Parameters is composed of settings which are "global" in the sense of affecting the entire job. Formatted as a set of parameter names and values, one set per row, and read in by exact match to their name. All values are expected as floating-point convertible unless otherwise mentioned.

MZ Precision should be adjusted to the practical precision of your MS system. If this parameter is set too low, expect to see gaps in extracted ion chromatograms. Set too high, it results in a higher noise level or even phantom peaks.

Minor Peak Detection Threshold governs peak detection in the reference well, with all peaks not above the maximum peak height detected multiplied by *this* discarded. Usually set to 0.001 and only rarely adjusted.

Peak Bound Detection Sensitivity governs the threshold for rate of change of slope for a peak bound to be called. Usually set at 50 and only rarely adjusted.

Begin Bound Detection Below Fractional Height governs the fraction of a peak's maximum height above which bounds cannot be called above. Set to 0.9 by default assuming neat peak shapes, but flat or jagged peak tops may require lowering this significantly at an increasing risk of integrating multiple overlapping peaks as one.

Maximum Number of Peaks to Report Per Target expects an integer value. If more peaks than this number are detected, the peak with the lowest intensity is discarded iteratively until there are no longer too many peaks. Used to reduce output of unwanted data.

Maximum Expected Peak Width (as fraction of run time) specifies a maximum reasonable peak width beyond which peaks are discarded. Useful to prevent situations where broad, shallow peaks may not have bounds properly detected.

Savitzky-Golay Smoothing Window expects an odd integer value, governing the window length of the SG filter. A longer window results in more smoothing and should be increased for data with a high scan rate or long run time.

Savitzky-Golay Smoothing Order expects an integer value and governs the order of the polynomial fit within the window by the SG filter. A lower value results in more smoothing.

Retention Time Window expects two float values separated by a whitespace, representing when to start processing peaks in a run and when to stop (in seconds). Alternatively, an asterisk can be used to represent that the entire run should be processed. Used to reduce output of unwanted data.

Volume of Donor Well (ml), *Volume of Acceptor Well (ml)*, *Active Surface Area of Membrane (cm^2)*, and *Assay Run Time (s)* are experimental parameters of the PAMPA apparatus necessary to the calculation of permeation rate. They are a required part of the configuration file even if solely processing data of other experiment types but are only accessed for PAMPA experiments.

Experiment Type

AutoPAMPA can, in addition to processing PAMPA data, also process partition coefficient experimental data, providing automated calculation of ratio and log ratio, or simply be set to integrate peaks and provide the raw integration values. This worksheet is organized in two columns, one for experiment names (must be unique strings), and one for experiment type. Valid experiment types are "PAMPA", "Ratio", and "Integrate".

Wells

This worksheet, organized in six columns, is used to associate well-pairs and mzML files to experiment names. The first column contains the experiment name to associate the rest of the row with and must match a name declared in the Experiment Type worksheet. The second through fourth columns should be filled with file paths (or file names if in the same directory) to appropriate mzML files.

The second column should contain the file path to a reference well or nothing. The first reference well give for each experiment name is accepted and the rest are ignored. The third and fourth columns are the donor well column and the acceptor well column (or hydrocarbon and water columns for partition coefficient experiments). Well-pairs must be declared in the same row to correctly associate them. If there are multiple well-pairs sharing the same target mass list, they can be associated to the same experiment name (one well-pair per row).

AutoPAMPA requires a reference well to be the anchor for peak alignment between associated wells, and, if none is provided, will use the donor well (or well A, column 3) on the assumption that the donor well has greater signal than the acceptor well (or well B, column 4). For a pure integration job, the first well encountered will be assigned as the reference well.

The fifth and sixth columns allow compensation for retention time drift between acquisition of the reference well and the donor and acceptor wells (assuming linear or nearly linear drift). Values should be in seconds.

Targets

This worksheet contains 4 columns used to specify the masses of interest for the various experiment names defined in the Experiment Type worksheet. Each row specifies a single exact mass for investigation. As usual, the first column contains the experiment name for the row to be associated with. The second column allows optional association of a name with a target mass (useful for keeping track of internal standards). The third column describes contains the exact masses, which should have enough digits to accommodate the precision defined in *Parameters*.

The final column is an optional location for taking manual control of peak calling and integration bounds of a target mass for an experiment name. Multiple peak locations and

bounds can be overwritten per row in the format: [Peak, left bound, right bound] in seconds without spaces. Multiple sets of these brackets should be separated by whitespace if needed.

Example: [110,97,123] [543,510,567]

A.3.2.4 Interpreting Output:

AutoPAMPA outputs two different excel files after a successful run and (optionally) many vector graphics files that visually represent all integrated peaks. Example output files are also available in this repository. One of the output files is a summary, ending in *_Results.xlsx* and contains averaged results in the case of multiple well-pairs or wells associated with a single experiment name. The other output file ending in *_Out.xlsx* contains the raw integration and calculated statistics for each individual well-pair or well.

Both are output with three worksheets, one for each job type (PAMPA, Ratio, Integrate), and each has its own output format after the first 5 columns, which are universal. The universal columns are: Experiment name, target mass name, exact mass, peak number, and retention time. Peak number is assigned based on peak alignments to the reference well for a well-set. PAMPA statistics are calculated as in the supplemental information of Naylor et al.[48] Non-detection of a peak in the acceptor well sets permeation rate to zero. A %T of over 100% sets permeation rate to "Invalid".

_Results.xlsx

This output file is useful for visual inspection of integration through the hyperlinks and an overview of the data.

PAMPA Reported statistics include (in order): The averaged percent transmission, the averaged percent recovery, the averaged permeation rate (10E-6 cm/s), the standard deviation of the permeation rate (if possible), the integrated recovery well intensity, the averaged integrated donor well intensity, the averaged integrated acceptor well intensity, the peak bound retention times, and (if -g) a hyperlink to the relevant stacked trace.

Ratio Reported statistics include (in order): The averaged integrated Well A intensity, the Well A standard deviation (if possible), the averaged integrated Well B intensity, the Well B standard deviation (if possible), the ratio of A/B, the log ratio of A/B, the peak bound retention times, and (if -g) a hyperlink to the relevant stacked trace.

Integrate Reported statistics include (in order): The averaged integrated intensity, the standard deviation of that intensity, the peak bound retention times, and (if -g) a hyperlink to the relevant stacked trace.

_Out.xlsx

This output file is useful to examine replicates for bad wells or for analyses that treat well-pairs separately.

PAMPA Reported statistics include (in order): The reference well file path, the reference well integrated intensity, the peak bound retention times, and a set of columns for each well pair. The repeated set contains the donor well file path, the donor well integrated intensity, the acceptor well file path, the acceptor well integrated intensity, the percent transmission, the percent recovery, and the permeation rate (10E-6 cm/s) all for that well-pair.

Ratio Reported statistics include (in order): The reference well file path, the reference well integrated intensity, the peak bound retention times, and a set of columns for each well pair. The repeated set contains the Well A file path, the Well A integrated intensity, the Well B file path, the Well B integrated intensity, and the A/B ratio for that well-pair.

Integrate Reported statistics include (in order): The reference well file path, the reference well integrated intensity, the peak bound retention times, and a set of columns for each well. The repeated set contains each a well's file path and its integrated intensity.

A.3.2.5 Known Bugs and Issues:

Non-scalar retention time drift between associated wells causes incorrect peak alignments in parts of the run that align poorly. Current corrective measures only account for scalar drift.

A.3.3 Data Processing Summary

First, spectral data files are parsed and organized into well-sets (e.g. sets of acceptor-donor pairs containing the same compounds) termed experiments. For each input spectral data file (each representing a single well), single-ion chromatograms are built for all masses present in a target mass list using the level of precision specified. Multi-event MS methods cause periodic zeroes in mzML format intensity data, so the next step (if necessary) is to remove them from intra-peak spaces to ensure continuous peaks. A Savitsky-Golay filter is then applied to reduce high-frequency noise and ease peak and peak bound detection. Peak detection is performed using the first and second derivatives of each extracted chromatogram, with peaks a minimum of 7 scans apart. Peaks are called with an emphasis on catching all possible peaks, then pruned of peaks of a height less than one thousandth of the most intense peak. A separate minimum peak intensity threshold is then applied to disqualify obvious noise peaks, reducing needless peak bound searches.

Peak bounds are identified by searching outward left and right from identified peaks for changes in slope above a specified threshold, with possible bounds ignored when near the top of a peak. Often integration will place bounds near a jagged peak-top without such a measure, so typically bounds are not called until the lower 80% of the peak's maximal height is reached. This parameter is adjusted individually to each data set and has the potential to cause multiple peaks to be integrated under a single call if set aggressively. A maximum peak bound width is also specified based on typical peak widths to discard unreasonable sets of bounds.

Despite discarding peaks of insignificant magnitude, peak calling still results in many more peaks called than desired. To remove spurious peaks, peak bounds are checked for significant overlap or subset status and removed if overlapped by more than 10% of their area. When multiple peaks share an area, only the most intense peak is retained. This strategy has been highly successful at retaining an appropriate number of called peaks except in cases where high-frequency noise is not sufficiently smoothed. The maximum

number of peaks specified is then used for a final reduction of peaks, if necessary, by dropping the least intense peaks first. Automatically detected peaks and peak bounds can be overridden to increase integration accuracy by specifying retention time coordinates in the configuration file.

If gaussian integration is enabled, each peak has a gaussian fit to the coordinates within its assigned bounds and the parameters of the fit are used to perform an exact integration. We find that peaks are generally close-enough to gaussian for a fit to be successful and useful. If peak shapes are not gaussian, the basic integration mode sums the intensities within the peak bounds on the assumption that the scan interval is consistent over that duration (approximately true for simple MS methods). All data analysis published herein was using the gaussian integration mode.

Once all peaks are assigned integrated intensities, all associated wells have their peaks aligned to a reference well. If there is no true reference well, a PAMPA experiment will automatically choose the donor well to take this role. In general, the well with the most intense signal should be used as the reference well because it makes the best template for all other peaks to be aligned with. Peaks are allowed a maximum of 3 s or (1 second per five minutes LC method length rounded up) leeway when matching to the reference peaks. Peak-width is also considered to avoid false matches; aligned peaks which differ in width by more than 5-fold are discarded. Any scalar retention time drift can be corrected for by offset fields in the configuration file. Monotonically increasing drift with increasing retention time can be compensated for, but generally results in failure to align peaks at one edge or the other of a run. Other types of retention time drift may result in misalignment of closely spaced peaks.

Data is output in the form of two excel files, each with one sheet for each supported experiment type (to create a rigid data format). The excel file ending in "_Results.xlsx" serves as a summary, containing averages of relevant statistics (%T, %R, PeE^{-6} cm/s for PAMPA), standard deviations, and, optionally, hyperlinks to generated graphical outputs. The excel file ending in "_Out.xlsx" contains the data generated for each well in total and reports more

statistics. When the command contains *-g*, stacked traces of all associated wells are generated for each mass are output as vector graphics files with peaks called, peak bounds, gaussian fit, and peak labels (if successfully aligned) displayed. The stacked plots aid in checking the fitness of automated integration, but their generation greatly increases the script's run time.

A.3.4 Associated Content

A copy of AutoPAMPA.py and example configuration and output/results spreadsheets accompany this dissertation and can also be found at https://GitHub.com/LokeyLab/AutoPAMPA.

A.4 RTMerge

RTMerge is a python script designed to take output from CycLS and AutoPAMPA (whether PAMPA data or other supported assay types) and associate sequences with chromatographic peaks by retention time. Using these three tools together allows for a full pipeline to convert sequencing and assay data acquired from complex mixtures into assay data on individual compounds. Like AutoPAMPA, RTMerge is controlled mainly through a configuration excel file and outputs an excel spreadsheet.

Merging uses the target masses and retention times of the assay data as the anchor for all matching due to the lower time resolution of MS^2 sampling. An interval-matching algorithm is used to accommodate inexact matches to the mass and time precisions given in the configuration file. Multiple matches to a single peak may occur if all matches are within the mass and time precision windows, in which case the closest retention time is retained, and the other matches discarded. Fields in the configuration file allow this process to account for any scalar chromatographic drift between the sequencing and assay data. Because only chromatographic peaks to which a sequence was assigned are reported in the output file, peaks with no sequencing match in any assay are dropped, as are all MS^2 clusters not corresponding to a chromatographic peak. This process removes irrelevant MS^2 clusters and low-intensity assay peaks with no recorded sequencing data, easing further curation.

In addition to merging one or more assay data sets with sequencing data, RTMerge generates SMILES strings, categorizes residues by stereochemistry and N-alkylation, and calculates RDKit's molALogP atomistic lipophilicity metric[32] (an implementation of the 1999 Wildman and Crippen AlogP) for each reported sequence.

A.4.1 Installation

As RTMerge requires successful installation of the required packages for CycLS and AutoPAMPA, no additional packages need be installed. However, only openpyxl and RDKit are required beyond those packages associated with a standard Anaconda installation.

A.4.2 Usage

A.4.2.1 Required arguments:

config: The only positional argument must contain the file path to a specially formatted excel file containing most of the input parameters for RTMerge. An example configuration file is included in this repository. A thorough explanation of the parameters contained can be found below.

A.4.2.2 Optional arguments:

-o, --out: Sets the prefix of the output files, ending in "_Merged.xlsx". If not set, the output file is simply named "Merged.xlsx".

A.4.2.3 Input File Preparation:

RTMerge expects a configuration file, which directs further input of assay and sequencing data in the form of the excel file output of AutoPAMPA and CycLS. An example file is included to demonstrate the proper format. Using it as a reference while reading the below is recommended.

Global Parameters

This worksheet is composed of settings which are "global" in the sense of affecting the entire job. Parameters are read in from columns one and two, with parameters recognized by name in column one.

Library Constraint A string representing the composition of the library using residue names from the amino acid database file. This is the same format as the CycLS constraint string, with amino acids present at the same position separated by commas and positions separated by semicolons. This string must cover all possible sequences for RTMerge to generate full SMILES strings for all reported sequences.

Cyclic Library? A boolean value, with True representing a cyclic library and False representing a linear library. Required for accurate SMILES string generation.

Amino Acid Database File Expects a string representing the file path to an amino acid database text file in the same format as used by CycLS (each row containing an amino acid name and SMILES string separated by a tab). The SMILES strings must be of N-to-C format for the SMILES string generation and other statistics to be accurate.

Example: L N[C@@H](CC(C)C)C(=O)O

Mass Precision (m/z) A float value representing the maximum difference between assay and sequencing exact masses before a match is refused. As CycLS attempts to get a high-resolution mass for each cluster of MS^2 data, the acquisition precision should be the same for both assay and sequencing.

Time Precision (s) A float value representing the maximum difference between assay and sequencing retention times in seconds before a match is refused.

Assay Data

This worksheet is used to specify the file paths to the assay data files and their assay types (columns one and two). The same file can be specified multiple times for different assay types as needed (types: PAMPA, Ratio, Integrate). If multiple assays types are merged

simultaneously, columns for all assay types will be present for each row. More details on that in the next section.

Experiments

This worksheet is used to specify each separate experiment (defined here as a set of wells containing the same compounds) by name and the file path of the sequencing data corresponding to each experiment (columns one and two). Additionally, columns three and four contain offsets to the assay data retention times and masses to allow them to match to sequencing data in the presence of scalar (or close to scalar) chromatographic drift or poor MS calibration. Despite these offsets, the retention times from the assay data are output rather than the sequencing retention times because the multi-event structure of the sequencing results in a low time-resolution.

If multiple experiments of the same name are given, the later mention will clobber the earlier one. If an experiment is present in multiple assay types, peaks will be labeled in the initial four columns (including retention time) by the first assay listed there is a match to. A sequence-peak match to any assay is sufficient and does not need to occur across all assays.

A.4.3.4 Interpreting Output:

RTMerge outputs a single excel sheet with two levels of column headers. The top level indicates the source of that column (sequencing, newly generated by RTMerge, or an assay data file) while the second level indicates the column contents. Only sequence-peak matches are output. Most sequencing statistics from CycLS and nearly all assay data columns are incorporated and can be interpreted as suggested in their respective readme.

RTMerge generates additional columns including SMILES strings, RDKit's molAlogP statistic, a breakdown of residue identity, stereochemistry, and N-alkylation by position, and one column each for short-form stereochemical and N-alkylation patterns.

A.4.3 Associated Content

A copy of RTMerge.py and an example configuration spreadsheet accompanies this dissertation and can also be found at https://GitHub.com/LokeyLab/RTMerge. The library data spreadsheets accompanying chapter two have a similar format to the output of RTMerge.

SUPPLEMENTAL FILES

Files associated with Chapter 1 include: A letter of permission to reprint the contents of chapter 1 signed by all co-authors and a data spreadsheet "Validation_1_unique_mass.xlsx" containing the analysis of the unique mass validation sequencing results.

Files associated with Chapter 2 include: Spreadsheets "Curated Hexamer Permeability Data.xlsx" and "Curated Heptamer Permeability Data.xlsx" containing all curated permeability data analyzed. Spreadsheets "Hexamer Motifs by NH Count.xlsx" and "Heptamer Motifs by NH Count.xlsx" containing all motifs analyzed.

Files associated with Appendix A include: "CycLS.py" containing the full python code for CycLS. "CycLSExampleAminoAcidDatabse.txt", an associated data file necessary to run CycLS.py. "AutoPAMPA.py" containing the full python code for AutoPAMPA. "AutoPAMPAExampledConfig.xlsx", an example configuration file appropriate for input to AutoPAMPA. "AutoPAMPAExample_Out.xlsx" and "AutoPAMPAExample_Results.xlsx", example output files from AutoPAMPA. "RTMerge.py" containing the full python code for RTMerge. "RTMergeExampleConfig.xlsx", an example configuration file appropriate for input to RTMerge.

REFERENCES

1. Qian, Z.; Dougherty, P. G.; Pei, D., Targeting intracellular protein-protein interactions with cell-permeable cyclic peptides. *Curr Opin Chem Biol* **2017,** *38*, 80-86.

2. Gao, Y.; Kodadek, T., Direct comparison of linear and macrocyclic compound libraries as a source of protein ligands. *ACS Comb. Sci.* **2015,** *17* (3), 190-5.

3. Sia, S. K.; Carr, P. A.; Cochran, A. G.; Malashkevich, V. N.; Kim, P. S., Short constrained peptides that inhibit HIV-1 entry. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** *99* (23), 14664-9.

4. Di, L., Strategic approaches to optimizing peptide ADME properties. *AAPS J* **2015,** *17* (1), 134-43.

5. Salmon, S. E.; Liu-Stevens, R. H.; Zhao, Y.; Lebl, M.; Krchnak, V.; Wertman, K.; Sepetov, N.; Lam, K. S., High-volume cellular screening for anticancer agents with combinatorial chemical libraries: a new methodology. *Mol. Divers.* **1996,** *2* (1-2), 57-63.

6. Wang, Y.; Xiao, W.; Zhang, Y.; Meza, L.; Tseng, H.; Takada, Y.; Ames, J. B.; Lam, K. S., Optimization of RGD-Containing Cyclic Peptides against alphavbeta3 Integrin. *Mol. Cancer Ther.* **2016,** *15* (2), 232-40.

7. Qian, Z.; Upadhyaya, P.; Pei, D., Synthesis and screening of one-bead-one-compound cyclic peptide libraries. *Methods Mol. Biol.* **2015,** *1248*, 39-53.

8. Gao, Y.; Amar, S.; Pahwa, S.; Fields, G.; Kodadek, T., Rapid lead discovery through iterative screening of one bead one compound libraries. *ACS Comb. Sci.* **2015,** *17* (1), 49-59.

9. Liu, R.; Marik, J.; Lam, K. S., A novel peptide-based encoding system for "one-bead one-compound" peptidomimetic and small molecule combinatorial libraries. *J. Am. Chem. Soc.* **2002,** *124* (26), 7678-80.

10. Wang, X.; Zhang, J.; Song, A.; Lebrilla, C. B.; Lam, K. S., Encoding method for OBOC small molecule libraries using a biphasic approach for ladder-synthesis of coding tags. *J. Am. Chem. Soc.* **2004,** *126* (18), 5740-9.

11. Joo, S. H.; Xiao, Q.; Ling, Y.; Gopishetty, B.; Pei, D., High-throughput sequence determination of cyclic peptide library members by partial Edman degradation/mass spectrometry. *J. Am. Chem. Soc.* **2006,** *128* (39), 13000-9.

12. Lee, J. H.; Meyer, A. M.; Lim, H. S., A simple strategy for the construction of combinatorial cyclic peptoid libraries. *Chem. Commun. (Camb.)* **2010,** *46* (45), 8615-7.

13. Liang, X.; Vezina-Dawod, S.; Bedard, F.; Porte, K.; Biron, E., One-Pot Photochemical Ring-Opening/Cleavage Approach for the Synthesis and Decoding of Cyclic Peptide Libraries. *Org. Lett.* **2016,** *18* (5), 1174-7.

14. Simpson, L. S.; Kodadek, T., A Cleavable Scaffold Strategy for the Synthesis of One-Bead One-Compound Cyclic Peptoid Libraries That Can Be Sequenced By Tandem Mass Spectrometry. *Tetrahedron Lett.* **2012,** *53* (18), 2341-2344.

15. Gurevich-Messina, J. M.; Giudicessi, S. L.; Martinez-Ceron, M. C.; Acosta, G.; Erra-Balsells, R.; Cascone, O.; Albericio, F.; Camperi, S. A., A simple protocol for combinatorial cyclic depsipeptide libraries sequencing by matrix-assisted laser desorption/ionisation mass spectrometry. *J Pept Sci* **2015,** *21* (1), 40-5.

16. Paizs, B.; Suhai, S., Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005,** *24* (4), 508-548.

17. Liu, W. T.; Ng, J.; Meluzzi, D.; Bandeira, N.; Gutierrez, M.; Simmons, T. L.; Schultz, A. W.; Linington, R. G.; Moore, B. S.; Gerwick, W. H.; Pevzner, P. A.; Dorrestein, P. C., Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Anal. Chem.* **2009,** *81* (11), 4200-9.

18. Bleiholder, C.; Osburn, S.; Williams, T. D.; Suhai, S.; Van Stipdonk, M.; Harrison, A. G.; Paizs, B., Sequence-scrambling fragmentation pathways of protonated peptides. *J. Am. Chem. Soc.* **2008,** *130* (52), 17774-89.

19. Novak, J.; Lemr, K.; Schug, K. A.; Havlicek, V., CycloBranch: De Novo Sequencing of Nonribosomal Peptides from Accurate Product Ion Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2015,** *26* (10), 1780-6.

20. Mohimani, H.; Liu, W. T.; Kersten, R. D.; Moore, B. S.; Dorrestein, P. C.; Pevzner, P. A., NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. *J. Nat. Prod.* **2014,** *77* (8), 1902-9.

21. Kavan, D.; Kuzma, M.; Lemr, K.; Schug, K. A.; Havlicek, V., CYCLONE--a utility for de novo sequencing of microbial cyclic peptides. *J. Am. Soc. Mass Spectrom.* **2013,** *24* (8), 1177-84.

22. Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. A., Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2012,** *109* (47), 19196-201.

23. Mohimani, H.; Liu, W. T.; Yang, Y. L.; Gaudencio, S. P.; Fenical, W.; Dorrestein, P. C.; Pevzner, P. A., Multiplex de novo sequencing of peptide antibiotics. *J. Comput. Biol.* **2011,** *18* (11), 1371-81.

24. Mohimani, H.; Liu, W. T.; Mylne, J. S.; Poth, A. G.; Colgrave, M. L.; Tran, D.; Selsted, M. E.; Dorrestein, P. C.; Pevzner, P. A., Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. *J. Proteome Res.* **2011,** *10* (10), 4505-12.

25. Mohimani, H.; Yang, Y. L.; Liu, W. T.; Hsieh, P. W.; Dorrestein, P. C.; Pevzner, P. A., Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **2011,** *11* (18), 3642-50.

26. Ng, J.; Bandeira, N.; Liu, W. T.; Ghassemian, M.; Simmons, T. L.; Gerwick, W. H.; Linington, R.; Dorrestein, P. C.; Pevzner, P. A., Dereplication and de novo sequencing of nonribosomal peptides. *Nat Methods* **2009,** *6* (8), 596-9.

27. Ngoka, L. C.; Gross, M. L., Multistep tandem mass spectrometry for sequencing cyclic peptides in an ion-trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **1999,** *10* (8), 732-46.

28. Redman, J. E.; Wilcoxen, K. M.; Ghadiri, M. R., Automated mass spectrometric sequence determination of cyclic peptide library members. *J. Comb. Chem.* **2003,** *5* (1), 33-40.

29. Hansen, M. E.; Smedsgaard, J., A new matching algorithm for high resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **2004,** *15* (8), 1173-80.

30. Furukawa, A.; Townsend, C. E.; Schwochert, J.; Pye, C. R.; Bednarek, M. A.; Lokey, R. S., Passive Membrane Permeability in Cyclic Peptomer Scaffolds Is Robust to Extensive Variation in Side Chain Functionality and Backbone Geometry. *J Med Chem* **2016,** *59* (20), 9503-9512.

31. Fawcett, T., An introduction to ROC analysis. *Pattern Recognition Letters* **2006,** *27* (8), 861-874.

32. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** *30* (10), 918-20.

33. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M., MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010,** *11*, 395.

34. Malkov, A. V.; Vrankova, K.; Cerny, M.; Kocovsky, P., On the selective N-methylation of BOC-protected amino acids. *J Org Chem* **2009,** *74* (21), 8425-7.

35. Marecek, J.; Song, B.; Brewer, S.; Belyea, J.; Dyer, R. B.; Raleigh, D. P., A simple and economical method for the production of 13C,18O-labeled Fmoc-amino acids with high levels of enrichment: applications to isotope-edited IR studies of proteins. *Org Lett* **2007,** *9* (24), 4935-7.

36. Zuckermann, R. N.; Kerr, J. M.; Kent, S. B. H.; Moos, W. H., Efficient method for the preparation of peptoids [oligo(N-substituted glycines)] by submonomer solid-phase synthesis. *J Am Chem Soc* **1992,** *114* (26), 10646-10647.

37. Villar, E. A.; Beglov, D.; Chennamadhavuni, S.; Porco, J. A., Jr.; Kozakov, D.; Vajda, S.; Whitty, A., How proteins bind macrocycles. *Nat Chem Biol* **2014,** *10* (9), 723-31.

38. Nielsen, D. S.; Shepherd, N. E.; Xu, W.; Lucke, A. J.; Stoermer, M. J.; Fairlie, D. P., Orally Absorbed Cyclic Peptides. *Chem Rev* **2017,** *117* (12), 8094-8128.

39. Zorzi, A.; Deyle, K.; Heinis, C., Cyclic peptide therapeutics: past, present and future. *Curr Opin Chem Biol* **2017,** *38*, 24-29.

40. Liras, S.; McClure, K. F., Permeability of Cyclic Peptide Macrocycles and Cyclotides and Their Potential as Therapeutics. *ACS Med Chem Lett* **2019,** *10* (7), 1026-1032.

41. White, A. M.; Craik, D. J., Discovery and optimization of peptide macrocycles. *Expert Opin Drug Discov* **2016,** *11* (12), 1151-1163.

42. Passioura, T.; Suga, H., A RaPID way to discover nonstandard macrocyclic peptide modulators of drug targets. *Chem Commun (Camb)* **2017,** *53* (12), 1931-1940.

43. Franzini, R. M.; Neri, D.; Scheuermann, J., DNA-Encoded Chemical Libraries: Advancing beyond Conventional Small-Molecule Libraries. *Acc Chem Res* **2014,** *47* (4), 1247-1255.

44. Wang, C. K.; Northfield, S. E.; Colless, B.; Chaousis, S.; Hamernig, I.; Lohman, R. J.; Nielsen, D. S.; Schroeder, C. I.; Liras, S.; Price, D. A.; Fairlie, D. P.; Craik, D. J., Rational design and synthesis of an orally bioavailable peptide guided by NMR amide temperature coefficients. *Proc Natl Acad Sci U S A* **2014,** *111* (49), 17504-9.

45. Wang, C. K.; Northfield, S. E.; Swedberg, J. E.; Colless, B.; Chaousis, S.; Price, D. A.; Liras, S.; Craik, D. J., Exploring experimental and computational markers of cyclic peptides: Charting islands of permeability. *Eur J Med Chem* **2015,** *97*, 202-13.

46. Rezai, T.; Yu, B.; Millhauser, G. L.; Jacobson, M. P.; Lokey, R. S., Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. *J Am Chem Soc* **2006,** *128* (8), 2510-1.

47. Hewitt, W. M.; Leung, S. S.; Pye, C. R.; Ponkey, A. R.; Bednarek, M.; Jacobson, M. P.; Lokey, R. S., Cell-permeable cyclic peptides from synthetic libraries inspired by natural products. *J Am Chem Soc* **2015,** *137* (2), 715-21.

48. Naylor, M. R.; Ly, A. M.; Handford, M. J.; Ramos, D. P.; Pye, C. R.; Furukawa, A.; Klein, V. G.; Noland, R. P.; Edmondson, Q.; Turmon, A. C.; Hewitt, W. M.; Schwochert, J.; Townsend, C. E.; Kelly, C. N.; Blanco, M. J.; Lokey, R. S., Lipophilic Permeability Efficiency Reconciles the Opposing Roles of Lipophilicity in Membrane Permeability and Aqueous Solubility. *J Med Chem* **2018,** *61* (24), 11169-11182.

49. Over, B.; Matsson, P.; Tyrchan, C.; Artursson, P.; Doak, B. C.; Foley, M. A.; Hilgendorf, C.; Johnston, S. E.; Lee, M. D. t.; Lewis, R. J.; McCarren, P.; Muncipinto, G.; Norinder, U.; Perry, M. W.; Duvall, J. R.; Kihlberg, J., Structural and conformational determinants of macrocycle cell permeability. *Nat Chem Biol* **2016,** *12* (12), 1065-1074.

50. Townsend, C.; Furukawa, A.; Schwochert, J.; Pye, C. R.; Edmondson, Q.; Lokey, R. S., CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry. *Bioorg Med Chem* **2018,** *26* (6), 1232-1238.

51. Rader, A. F. B.; Reichart, F.; Weinmuller, M.; Kessler, H., Improving oral bioavailability of cyclic peptides by N-methylation. *Bioorg Med Chem* **2018,** *26* (10), 2766-2773.

52. 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.; Kalgutkar, A. S.; Bauman, J. N.; Zhang, Y.; Liras, S.; Price, D. A.; Mathiowetz, A. M.; Jacobson, M. P.; Lokey, R. S., On-resin N-methylation of cyclic peptides for discovery of orally bioavailable scaffolds. *Nat Chem Biol* **2011,** *7* (11), 810-7.

53. Schwochert, J.; Turner, R.; Thang, M.; Berkeley, R. F.; Ponkey, A. R.; Rodriguez, K. M.; Leung, S. S.; Khunte, B.; Goetz, G.; Limberakis, C.; Kalgutkar, A. S.; Eng, H.; Shapiro, M. J.; Mathiowetz, A. M.; Price, D. A.; Liras, S.; Jacobson, M. P.; Lokey, R. S., Peptide to Peptoid Substitutions Increase Cell Permeability in Cyclic Hexapeptides. *Org Lett* **2015,** *17* (12), 2928-31.

54. Goetz, G. H.; Philippe, L.; Shapiro, M. J., EPSA: A Novel Supercritical Fluid Chromatography Technique Enabling the Design of Permeable Cyclic Peptides. *ACS Med Chem Lett* **2014,** *5* (10), 1167-72.

55. Guixer, B.; Arroyo, X.; Belda, I.; Sabido, E.; Teixido, M.; Giralt, E., Chemically synthesized peptide libraries as a new source of BBB shuttles. Use of mass spectrometry for peptide identification. *J Pept Sci* **2016,** *22* (9), 577-91.

56. Ciudad, S.; Bayó-Puxán, N.; Varese, M.; Seco, J.; Teixidó, M.; García, J.; Giralt, E., 'À La Carte' Cyclic Hexapeptides: Fine Tuning Conformational Diversity while Preserving the Peptide Scaffold. *ChemistrySelect* **2018,** *3* (8), 2343-2351.

57. Oh, M. H.; Lee, H. J.; Jo, S. H.; Park, B. B.; Park, S. B.; Kim, E. Y.; Zhou, Y.; Jeon, Y. H.; Lee, K., Development of Cassette PAMPA for Permeability Screening. *Biol Pharm Bull* **2017,** *40* (4), 419-424.

58. Fouche, M.; Schafer, M.; Berghausen, J.; Desrayaud, S.; Blatter, M.; Piechon, P.; Dix, I.; Martin Garcia, A.; Roth, H. J., Design and Development of a Cyclic Decapeptide Scaffold with Suitable Properties for Bioavailability and Oral Exposure. *ChemMedChem* **2016,** *11* (10), 1048-59.

59. Furukawa, A.; Schwochert, J.; Pye, C. R.; Asano, D.; Edmondson, Q. D.; Turmon, A.; Klein, V.; Ono, S.; Okada, O.; Lokey, R. S., Drug-like properties in macrocycles above MW 1000: Backbone rigidity vs. side-chain lipophilicity. *Angew Chem Int Ed Engl* **2020**.

60. Zhang, J.; Maloney, J.; Drexler, D. M.; Cai, X.; Stewart, J.; Mayer, C.; Herbst, J.; Weller, H.; Shou, W. Z., Cassette incubation followed by bioanalysis using high-resolution MS for in vitro ADME screening assays. *Bioanalysis* **2012,** *4* (5), 581-93.

61. el Tayar, N.; Mark, A. E.; Vallat, P.; Brunne, R. M.; Testa, B.; van Gunsteren, W. F., Solvent-dependent conformation and hydrogen-bonding capacity of cyclosporin A: evidence from partition coefficients and molecular dynamics simulations. *J Med Chem* **1993,** *36* (24), 3757-64.

62. Whitty, A.; Zhong, M.; Viarengo, L.; Beglov, D.; Hall, D. R.; Vajda, S., Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. *Drug Discov Today* **2016,** *21* (5), 712-7.

63. Bhushan, K. R., Light-directed maskless synthesis of peptide arrays using photolabile amino acid monomers. *Org Biomol Chem* **2006,** *4* (10), 1857-9.

64. Mann, H. B.; Whitney, D. R., On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.* **1947,** *18* (1), 50-60.

65. Avdeef, A., *Absorption and drug development: solubility, permeability, and charge state*. John Wiley & Sons: 2012.