

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Student Research Projects, Dissertations, and
Theses - Chemistry Department

Chemistry, Department of

Winter 12-10-2021

Developing Techniques for the Identification of Non-Canonical RNA Pairing and Analysis of LC-MS Datasets

Christopher Jurich

University of Nebraska-Lincoln, cjurich2@huskers.unl.edu

Follow this and additional works at: <https://digitalcommons.unl.edu/chemistrydiss>

 Part of the [Chemistry Commons](#)

Jurich, Christopher, "Developing Techniques for the Identification of Non-Canonical RNA Pairing and Analysis of LC-MS Datasets" (2021). *Student Research Projects, Dissertations, and Theses - Chemistry Department*. 111.

<https://digitalcommons.unl.edu/chemistrydiss/111>

This Article is brought to you for free and open access by the Chemistry, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Student Research Projects, Dissertations, and Theses - Chemistry Department by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

DEVELOPING TECHNIQUES FOR THE IDENTIFICATION OF
NON-CANONICAL RNA PAIRING AND ANALYSIS OF LC-MS DATASETS

By

Christopher P. Jurich

A THESIS

Presented to the Faculty of

The Graduate College of the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Chemistry

Under the Supervision of Professor Joseph D. Yesselman

Lincoln, Nebraska

December, 2021

DEVELOPING TECHNIQUES FOR THE IDENTIFICATION OF NON-CANONICAL RNA PAIRING AND ANALYSIS OF LCMS DATASETS

Christopher P. Jurich, M.S.

University of Nebraska, 2021

Advisor: Joseph D. Yesselman

Non-canonical pairing dynamics in ribonucleic acid (RNA) structure and statistical analysis of metabolomics liquid chromatography mass spectrometry (LC-MS) datasets are two difficult problems that stand as open challenges.

RNA folding algorithms are used across various disciplines to predict structures when experimental elucidation techniques are inconvenient or impractical. Though successful and widely adopted, folding algorithms make simplifying assumptions for loop regions due to their complex interactions and associated difficulty with generating energy parameters for relevant non-canonical pairs. Modeling assumptions and a lack of energy parameters for loops limit accuracy in these functionally critical regions of RNA. This work describes a new technique for probing non-canonical loop interactions through the combined analysis of dimethyl sulfate (DMS) and three-dimensional crystallographic data. We demonstrate that DMS data encodes information about non-canonical pairing, which can describe these interactions in an efficient, high throughput manner.

Metabolomics aims to understand biological processes through the analysis of small molecule metabolites. The field primarily uses ^1H nuclear magnetic resonance (NMR) spectroscopy as well as LC-MS to identify and quantitate metabolites. With even simple samples having hundreds or thousands of metabolites, researchers in the field have developed software pipelines to make metabolomics studies a tractable task. Numerous packages exist for the analysis of either

^1H NMR or LC-MS data, but current offerings force researchers to use multiple packages to analyze both spectral data types. To address the need for a metabolomics package capable of analyzing both spectral types, we have developed new LC-MS functionality for the NMR metabolomics package MVAPACK.

Preface

Understanding non-canonical pairing in ribonucleic acid (RNA) structure¹⁻² and statistical analysis of metabolomics liquid chromatography mass spectrometry (LC-MS) datasets are two challenging objectives that rely on computational modeling and tool development.³⁻⁴ This thesis summarizes my work in (1) identifying experimentally derived markers to characterize non-canonical RNA pairing and (2) adding implementations of field-standard LC-MS processing techniques to the MVAPACK metabolomics software package.⁵

RNA plays critical roles throughout the cell, including splicing, translation, transcription regulation, and gene silencing⁶⁻⁸. To perform these functions RNA fold into complex structures that can respond to cellular stimuli. A number of experimental techniques have been developed to elucidate these structures, including X-Ray Crystallography (XRC), cryoelectron microscopy (Cryo-EM), optical melting and cross-linking studies. While these techniques have effectively characterized RNA structures for decades, they are slow to perform with single structures often taking weeks or months to solve. As a result, full elucidation techniques are not viable for many projects.

Researchers have overcome experimental limitations by developing RNA folding algorithms that predict structure from sequence alone. While an XRC or Cryo-EM study could take years to perform, an ensemble of potential structures can be predicted in minutes or seconds. RNA folding algorithms rely on a simplified thermodynamic model of nucleotide pairing that generates secondary structures containing helices and loops. Helices are composed of canonical AU/UA and CG/GC as well as wobble GU/UG pairs whereas loops contain nucleotides participating in other non-canonical pairing modes. RNA folding packages represent the most success-

ful class of algorithms in all of bioinformatics. They see global accuracies of 60-70% with improved performance for smaller RNAs on the order of 200 nucleotides or less.²⁸ As a result, RNA folding algorithms are widely used in primer assembly, mRNA vaccine design, and bioengineering at large. While RNA folding algorithms are largely able to predict which nucleotides are contained in helices and loops, they cannot predict non-canonical interactions due to the lack of thermodynamic parameters on these pairs set has necessitated modelling assumptions that limit accuracy in loop regions.

Reduced predictive accuracy for loops poses a problem in the field as loops are often function critical regions of RNA.¹⁶ For example, Sarcin-Ricin (SR) loops anchor elongation factor G (EF-G) during mRNA-TRNA translocation, enabling the elongation phase of protein synthesis.¹⁷ SR loops rely on loop region stability and could not carry out their biological function in their absence. Inaccurate prediction of loops limits identification of structure-function relationships, potentially leading to erroneous understanding of RNAs. These limitations can only be overcome through mass collection of data for non-canonical pairing modes. Collecting data on non-canonical pairing modes would then enable increased predictive accuracy for loop regions and advance the state-of-the-art RNA folding algorithms.

Metabolomics encompasses the comprehensive characterization of small molecule metabolites from a variety of biological samples that includes tissues, cell lysates, and biofluids.¹⁸⁻¹⁹ The composite of these small molecule markers is termed a metabolome and provides insight into both regular biological processes and disease states because small molecules are the end products of enzymatic reactions and are involved in most cellular processes.²⁰ Unlike other “omics” disciplines including proteomics, genomics and transcriptomics, metabolomics allows high resolution measurement of metabolite abundance, giving the field unprecedented quantitative

precision.²¹ Metabolomics relies on both ^1H NMR and LC-MS spectroscopy to measure and identify the presence of metabolites. With simple metabolomes having hundreds of metabolites, researchers in the field have responded by developing software tools for automated metabolomics analysis.²² Despite dozens of software packages being available from both commercial and academic developers, most limit their input data to either NMR or LC-MS, forcing a combined analysis to incorporate at least two packages. The XCMS, OpenMS and Maven packages offer LC-MS functionality but not NMR and the Metabolab, NMRPipe and MVAPACK packages offer NMR functionality but not LC-MS.^{5,23-27} The inability of the field to offer a package which analyzes both NMR and LC-MS data presents an issue as metabolomics researchers must perform multiple analyses to gain maximum coverage of the metabolome.

References

1. Lemieux, S. RNA Canonical and non-canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research*, 2002, 30, 4250–4263. <https://doi.org/10.1093/nar/gkf540>.
2. Das, J.; Mukherjee, S.; Mitra, A.; Bhattacharyya, D. non-canonical Base Pairs and Higher Order Structures in Nucleic Acids: Crystal Structure Database Analysis. *Journal of Biomolecular Structure and Dynamics*, 2006, 24, 149–161. <https://doi.org/10.1080/07391102.2006.10507108>.
3. Zhou, B.; Xiao, J. F.; Tuli, L.; Resson, H. W. LC-MS-Based Metabolomics. *Mol. Biosyst.*, 2012, 8, 470–481. <https://doi.org/10.1039/c1mb05350g>.
4. Xiao, J. F.; Zhou, B.; Resson, H. W. Metabolite Identification and Quantitation in LC-MS/MS-Based Metabolomics. *TrAC Trends in Analytical Chemistry*, 2012, 32, 1–14. <https://doi.org/10.1016/j.trac.2011.08.009>.
5. Worley, B.; Powers, R. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 2014, 9, 1138–1144. <https://doi.org/10.1021/cb4008937>.
6. van den Hoogenhof, M. M. G.; Pinto, Y. M.; Creemers, E. E. RNA Splicing. *Circulation Research*, 2016, 118, 454–468. <https://doi.org/10.1161/circresaha.115.307872>.
7. Cramer, P. Organization and Regulation of Gene Transcription. *Nature*, 2019, 573, 45–54. <https://doi.org/10.1038/s41586-019-1517-4>.
8. Meister, G.; Tuschl, T. Mechanisms of Gene Silencing by Double-Stranded RNA. *Nature*, 2004, 431, 343–349. <https://doi.org/10.1038/nature02873>.
9. Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Research*, 2003, 31, 3406–3415. <https://doi.org/10.1093/nar/gkg595>.
10. Kings Oluoch, I.; Akalin, A.; Vural, Y.; Canbay, Y. A Review on RNA Secondary Structure Prediction Algorithms. 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), 2018. <https://doi.org/10.1109/ibigdelft.2018.8625347>.
11. Scott, L. G.; Hennig, M. RNA Structure Determination by NMR. *Bioinformatics*, 2008, 29–61. https://doi.org/10.1007/978-1-60327-159-2_2.
12. Serra, M. J.; Axenson, T. J.; Turner, D. H. A Model for the Stabilities of RNA Hairpins Based on a Study of the Sequence Dependence of Stability for Hairpins of Six Nucleotides. *Biochemistry*, 1994, 33, 14289–14296. <https://doi.org/10.1021/bi00251a042>.
13. Rybarczyk, A.; Szostak, N.; Antczak, M.; Zok, T.; Popena, M.; Adamiak, R.; Blazewicz, J.; Szachniuk, M. New in Silico Approach to Assessing RNA Secondary Structures with non-canonical Base Pairs. *BMC Bioinformatics*, 2015, 16. <https://doi.org/10.1186/s12859-015-0718-6>.
14. White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B. Recognition of the Four Watson–Crick Base Pairs in the DNA Minor Groove by Synthetic Ligands. *Nature*, 1998, 391, 468–471. <https://doi.org/10.1038/35106>.
15. Olson, W. K.; Colasanti, A. V.; Lu, X.-J.; Zhurkin, V. B. Watson-Crick Base Pairs: Character and Recognition. *Wiley Encyclopedia of Chemical Biology*, 2008. <https://doi.org/10.1002/9780470048672.webc452>.

16. Schudoma, C.; May, P.; Nikiforova, V.; Walther, D. Sequence–Structure Relationships in RNA Loops: Establishing the Basis for Loop Homology Modeling. *Nucleic Acids Research*, 2009, 38, 970–980. <https://doi.org/10.1093/nar/gkp1010>.
17. Shi, X.; Khade, P. K.; Sanbonmatsu, K. Y.; Joseph, S. Functional Role of the Sarcin–Ricin Loop of the 23S RRNA in the Elongation Cycle of Protein Synthesis. *Journal of Molecular Biology*, 2012, 419, 125–138. <https://doi.org/10.1016/j.jmb.2012.03.016>.
18. Clish, C. B. Metabolomics: An Emerging but Powerful Tool for Precision Medicine. *Molecular Case Studies*, 2015, 1, a000588. <https://doi.org/10.1101/mcs.a000588>.
19. Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: Beyond Biomarkers and towards Mechanisms. *Nature Reviews Molecular Cell Biology*, 2016, 17, 451–459. <https://doi.org/10.1038/nrm.2016.25>.
20. Wishart, D. S. Metabolomics for Investigating Physiological and Pathophysiological Processes. *Physiological Reviews*, 2019, 99, 1819–1875. <https://doi.org/10.1152/physrev.00035.2018>.
21. Liu, X.; Ser, Z.; Locasale, J. W. Development and Quantitative Evaluation of a High-Resolution Metabolomics Technology. *Analytical Chemistry*, 2014, 86, 2175–2184. <https://doi.org/10.1021/ac403845u>.
22. Spicer, R.; Salek, R. M.; Moreno, P.; Cañueto, D.; Steinbeck, C. Navigating Freely-Available Software Tools for Metabolomics Analysis. *Metabolomics*, 2017, 13. <https://doi.org/10.1007/s11306-017-1242-7>.
23. Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 2006, 78, 779–787. <https://doi.org/10.1021/ac051437y>.
24. Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nature Methods*, 2016, 13, 741–748. <https://doi.org/10.1038/nmeth.3959>.
25. Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine. *Current Protocols in Bioinformatics*, 2012. <https://doi.org/10.1002/0471250953.bi1411s37>.
26. Ludwig, C.; Günther, U. L. MetaboLab - Advanced NMR Data Processing and Analysis for Metabolomics. *BMC Bioinformatics*, 2011, 12. <https://doi.org/10.1186/1471-2105-12-366>.
27. Delaglio, F.; Grzesiek, S.; Vuister, Geerten W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *Journal of Biomolecular NMR*, 1995, 6. <https://doi.org/10.1007/bf00197809>.
28. Doshi, K. J.; Cannone, J. J.; Cobaugh, C. W.; Gutell, R. R. *BMC Bioinformatics*, 2004, 5, 105. <https://doi.org/10.1186/1471-2105-5-105>.

TABLE OF CONTENTS

LIST OF FIGURES	10
CHAPTER 1: Probing non-canonical RNA Pairing Through DMS	11
CHAPTER 2: Developing LC-MS Functionality for MVAPACK	38
CHAPTER 3: Creating Synthetic Data to Validate MVAPACK's New Functionality	69
CHAPTER 4: Summary of Work	81

LIST OF FIGURES

Figure 1.1: DMS Modification of Adenosine and Cytosine	15
Figure 1.2: Junction Motifs Provide Ideal Opportunities to Create non-canonical Pairs	17
Figure 1.3: Synthetic RNA Libraries Provide Useful Data For Pairing Analysis	20
Figure 1.4: C-C DMS Values Correspond to Weak or No Pairing	29
Figure 1.5: GA Pairing Modes Exhibit Differing DMS Value Ratios	31
Figure 1.6: DMS Reactivity Values Identify Diverse Junction Topologies	33
Figure 2.1: LC-MS Data Allows High Precision Metabolite Identification	41
Figure 2.2: PCAs Allow Visualization of High Dimensional Spaces	42
Figure 2.3: MVAPACK LC-MS Pipeline Overview	44
Figure 2.4: EICs Isolate Single Peaks	47
Figure 2.5: Gaussian Derivative Transforms Identify Peak Regions	50
Figure 2.6: Matrix Generation	53
Figure 2.7: Matrix Normalization	56
Figure 2.8: Peak Imputation Accounts For Missing LC-MS Peaks	59
Figure 2.9: Peak Filtration Selects Significant Features	62
Figure 3.1: Idealized EICs Are Generated For Each Metabolite	74
Figure 3.2: Noise Is Added at Three Levels to Vary Spectra Quality	75
Figure 3.3: MVAPACK's PCA Model Closely Matches the Idealized Version	77
Figure 3.4: MVAPACK's Performance Across Spectral Quality Level	78

CHAPTER 1: PROBING NON-CANONICAL RNA PAIRING THROUGH DMS

1.1 The Role of non-canonical Pairing in RNA

Previously thought to only be an intermediate between DNA and proteins, RNAs are now known to be responsible for a growing list of critical biological functions including translation of proteins, gene regulation, and mRNA maturation.¹⁻³ RNA's roles within the cell require folding from primary sequence to secondary structure governed by strong base-base interactions.⁵ Base interactions occur between any two of the four bases adenine (A), cytosine (C), guanine (G) and uracil (U), and specific pairings are what define and drive the folded secondary structure of an RNA. Of all possible pair combinations, the AU/UA, GC/CG, or canonical pairs, are the most stable, with GU/UG wobble pairs having comparable stability depending on context.⁶ Any other pair is non-canonical and less stable than a canonical counterpart. The canonical, wobble and non-canonical pairs of an RNA fully describe its secondary structure. With RNA structure driving functional roles, the field has developed both high- and low-resolution structure elucidation techniques to understand structure-function relationships. Though effective, these techniques have accompanying constraints that has limited their use.

X-Ray Crystallography (XRC) and Cryo-EM are high resolution elucidation techniques which provide full three-dimensional characterization of a given RNA's structure.¹²⁻¹³ These techniques generate ample structural information but come at the cost of long acquisition times. Generating full 3D models can take months to years for a single RNA and is increasingly difficult for longer strands. Cryo-EM may see decreased runtimes as technology improves, but current timelines for high resolution structure elucidation make these techniques impractical for high volume usage. Crosslinking is a lower resolution structural elucidation technique which instead reports on spatial distances between nucleotides.¹⁵ This technique uses various affinity

binding agents to induce the formation of cross-links between proximal nucleotides. Nucleotide distance constraints can be derived from these constraints which inform global structure dynamics. Cross-linking is more rapid than XRC or Cryo-EM but produces much lower resolution data. Deriving a full structure is still a complicated process and further assumes that cross linking agents have minimal impact on global structure conformation. The long timelines of high resolution techniques and insufficient data quality of low resolution techniques limits usage of these protocols for many applications. Researchers in the field have realized these limitations and consequently developed algorithms to predict the structure of RNA.

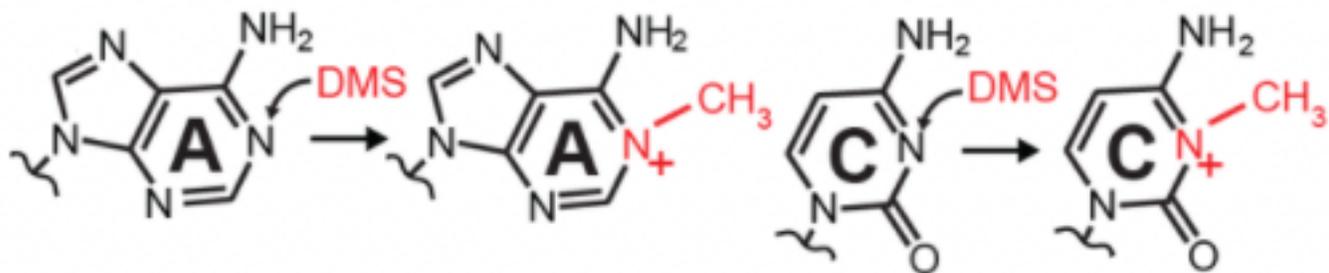
Thermodynamic RNA folding algorithms predict secondary structure by finding the minimum free energy (MFE) for a sequence. Estimated MFE values are generated using the nearest neighbor model which assumes that the energy contribution from a basepair step (or two consecutive base pairs) is always the same.⁷ A total of 36 canonical basepair steps exist and corresponding parameters were determined via high resolution optical melting experiments and NMR experiments.²⁵⁻²⁶ When combined, the nearest neighbor model and optical melting parameters enable energy estimation for arbitrary structures. Energy minimization as a technique is the dominant paradigm for RNA folding algorithms and nearest neighbor physics models are utilized in Mfold, RNAfold, RNAstructure and NUPACK, among others.⁸⁻¹¹ Existing RNA folding algorithms have been a massive success for the field, finding widespread adoption for all RNA researchers across disciplines. Despite this success, the nearest neighbor model makes estimations and has an incomplete view of basepair energetics. If addressed, RNA folding algorithms could see further accuracy improvements.

The limited number of canonical pairs allowed for practical generation of energy parameters, but 220 basesteps with a single non-canonical pair exist, making an analogous experiment impractical from combinatorial explosion alone. Nearest neighbor models address this lack of explicit parameters for con-canonical interactions via simplifying assumptions. For example, the energy penalty associated with junction formation is a heuristic rule, penalizing larger junctions regardless of potential sequence identity. These simplifications, combined with a lack of parameters, limit the accuracy of the nearest neighbor model for RNA loop prediction. Reduced accuracy for loop prediction poses a problem as RNA loops are often functional critical regions. Improved prediction of RNA loops requires acquisition of new datasets describing these structural features and their corresponding non-canonical pairs. The vast number of basepair steps to be probed makes optical melting an impractical choice for energy parameter expansion. Other popular techniques in the field provide limitations with XRC and Cryo-EM being too time consuming and cross-linking providing insufficient resolution. An alternative method for structure determination is dimethyl sulfate (DMS).

For over 40 years, DMS has been used to describe the structure of RNA without the use of next-generation sequencing. DMS has seen increased usage in recent years with next-generation sequencing has enabling thousands of RNAs to be analyzed in a one pot reaction. Dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq) chemical mapping offers high throughput RNA structure elucidation via exposure to the small molecule DMS.¹⁶ DMS selectively methylates the N1 of adenine and N3 of cytosine (Figure 1.1) via electrophilic substitution when these nitrogens are solvent exposed. Methylated nucleotides undergo mutations when the RNA are reverse transcribed by group II intron reverse transcriptase (TGIRT), resulting in a pool of mutated complementary deoxyribose nucleic acids (cDNAs). The mutational frequency at

each given nucleotide can be related to the rate of methylation on the RNA.²⁷ Sequencing and analysis of cDNA pools yields reactivity profiles which provide nucleotide level information for each analyzed RNA. Each nucleotide receives a value that correlates with its mutational rate, with lower values being widely accepted as markers of pairing since the N1's and N3's of paired nucleotides are unlikely to react with DMS. DMS-MaPseq data will likely report on non-canonical pairings as the N1 of A and N3 of C are central to the hydrogen bonds that drive pairing for these nucleotides.¹⁸ The high throughput capability of DMS-MaPseq is suited to developing a fundamental model of RNA non-canonical pairing. Thousands of RNA can be processed in parallel, enabling wide coverage of possible non-canonical pairing modes in a single experiment.

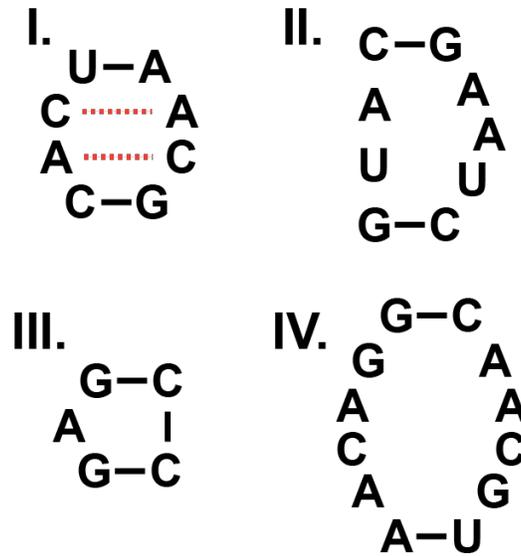
Figure 1.1: DMS Modification of Adenosine and Cytosine



DMS selectively methylates the N1 of A (left) and N3 of C (right) when these atoms are solvent accessible.

Secondary structure junction motifs are an ideal system to gather data on non-canonical RNA pairing. A junction is composed of two or more loops flanked by two Watson-Crick pairs (Figure 1.2). Secondary structure is typically described via dot-bracket notation which represents nucleotides in helices with parentheses, “(“ or “)”, and nucleotides in loops with dots, “.”. Junctions additionally use ampersands to show that looped regions share flanking pairs. A junction with two loops of size 3 and 2 has the following dot-bracket notation “(...(&)..)”. Despite not participating in canonical pairing, loop nucleotides often contain stabilize a junction via non-canonical interactions with other loop nucleotides. The exact non-canonical interactions seen vary by the size and number of constituent loops, but small, symmetrical motifs are conducive to symmetrical pairing.¹⁰ For example, a junction with two size 3 loops will typically have 2 closing canonical pairs and 3 non-canonical loop pairs. Junctions also play critical biological roles and as a result have been widely catalogued in databases like the protein data bank (PDB). The PDB provides 3D atomic data for RNAs which allows for easy classification of non-canonical pairs types through the X3DNA software tool.¹¹ To combine the abilities of DMS chemical mapping and X3DNA classifications, we have developed RNA libraries which provide multiple instances of junctions with known 3D structures and pair types. By performing a deep dive on this combined dataset, we have demonstrated that DMS chemical mapping encodes data that identifies the non-canonical pairing mode a nucleotide is participating in.

Figure 1.2: Junction Motifs Provide Ideal Opportunities to Create non-canonical Pairs



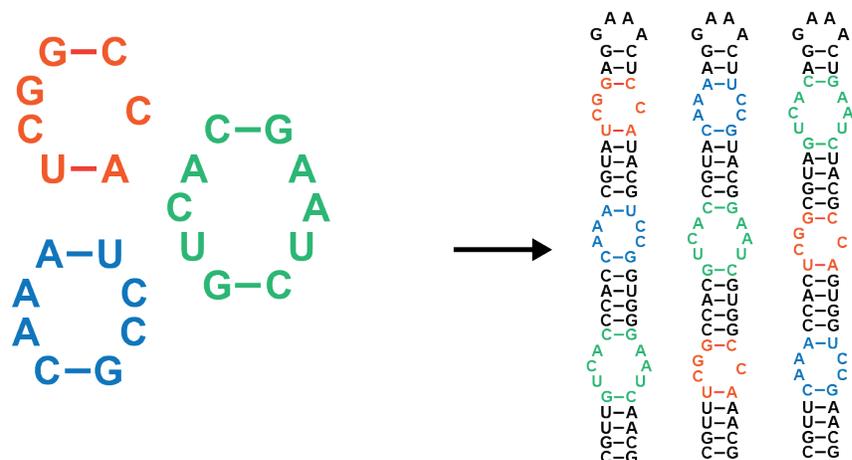
Junctions are structural motifs wherein two or more loop regions are enclosed by the same flanking pairs. These motifs vary in size and are characterized by the size and number of loops present with I being a 2x2, II being a 2x3, III being a 1x0 and IV being a 4x4. As seen in III, junctions may have loops of size zero, also known as bulges. The proximity of loop nucleotides promotes non-canonical interactions which have been highlighted with orange dotted lines in I.

Transcriptomics experiments probe entire transcriptomes from cells and resulting sequencing data has been made public for a variety of organisms. Despite public availability, these datapoints are not ideal for a deep dive into junction pairing dynamics. The RNA for these published datasets are typically from biological sources and have relatively few junctions with known non-canonical pairing.¹⁹ In addition, these structures have many complex tertiary interactions that will complicate analysis as long-range interactions are always a potential explanation for deviations in expected DMS signal. Another key issue with natural RNAs is the lack of repeated junction motifs. Whether a pattern of DMS reactivity is an anomaly or a consistent pattern can only be determined through the comparison of multiple instances of the same junction. The lack of repeats for each junction sequence limits the confidence of patterning associated with each junction. Larger sequences are also more likely to adopt multiple conformations either locally or globally, adding further challenge to potential analyses.²⁰ As a result, designing a synthetic library is an attractive option to enable a deep dive into non-canonical RNA pairing.

Synthetic RNA libraries are the optimal choice for analyzing junction dynamics as we can design stable sequences which provide multiple examples of junctions of interest. This approach addresses other limitations of analyzing existing data as smaller sequence size limits the potential for tertiary interactions. Smaller sequence size is also advantageous for predictive purposes as thermodynamic folding algorithms are known to be much more accurate for smaller RNA strands than larger ones.²¹ Mass pools of sequences can then be generated with higher confidence that the strands and motifs will form into the predicted structures. Engineering a sequence pool also provides the benefit of sequence repeats as each junction of interest can be repeated and implemented across different sequence contexts. Observing the same junction across different sequences provides further confidence that a pattern is meaningful and repeatable.

To effectively probe non-canonical RNA pairing, we have developed and performed DMS chemical mapping on a synthetic RNA library with 7,807 different junctions. This library was designed specifically to have as many instances of junctions as possible with a total of 7,807 unique sequence motifs being present 6.72 times on average (Figure 5). The resulting DMS data has been normalized, aggregated and analyzed to identify unique and unexpected patterns of pairing. By combining these patterns with pairing classifications of solved PDB structures we have identified repeatable patterns that can be confidently tied to specific pairing modes. In general, we demonstrate that DMS reactivities provide information about non-canonical pairing modes in 3D RNA structure.

Figure 1.3: Synthetic RNA Libraries Provide Useful Data For Pairing Analysis



Synthetic RNA libraries are generated from pools of junction sequences as seen on the left. Each motif is repeated multiple times across different RNA constructs to mass generate data for each junction analyzed.

1.2 Materials and Methods

1.2.1 Library Design

The objective for our RNA library was to sample a wide set of non-canonical pairings and create multiple instances of each pair type. We designed our library to span a wide range of RNA junctions. Each junction was inserted into an average of 10 distinct RNA constructs, to enable identification of trends. We started with all combinations of junctions that had AU or CG closing pairs and loops of up to size three which corresponds to a total of 115,584 junctions.. Limiting the size of loops to three is conducive to stability as loops become less stable as they grow in size and foster RNA-RNA tertiary contacts with other strands in the pool.¹⁹ These constraints left 112,896 possible junctions and we next removed sequence motifs that did not have an A or C in their loops as these are the only nucleotides sensitive to DMS modification. This results in 109,760 junctions. We then filtered out junctions with low predicted stability by placing each candidate junction into three sets of hypothetical helices and checking if their predicted structures were consistent with the desired secondary structure. RNA folding algorithms are accurate for small strands and a junction that misfolds suggests it may not fold properly in our final RNA constructs.²⁰ To further reduce the number of junction possibilities, we only kept junctions with at least one CG/GC closing pair. The resulting junction counts are seen in Table 1.1

Junction Size Summary

Left Loop Size	Right Loop Size	Number of Sequences
0	1	18
1	0	18
2	0	78
0	2	80
1	1	93
2	1	288
1	2	288
0	3	331
3	0	352
1	3	1002
3	1	1016
2	2	1172
3	2	3350
2	3	3355
3	3	13750

Summary of junction counts after closing pair, loop identity and stability requirements were enforced.

After establishing a refined pool of junctions of varying size up to a left and right loop both with size 3 (denoted 3x3), a subset were chosen for use in designing RNA constructs. All junctions with three or fewer unpaired nucleotides were chosen which include the following sizes: 0x1, 1x0, 1x1, 0x2, 2x0, 1x2, 2x2, 3x0 and 0x3. Larger junctions entries were prioritized by the number of A's or C's in the junctions to give us the most DMS active nucleotides.

Each junction in the selected pool was then repeated 10 times and randomly selected into groups of six which were separated by seven helices. Each construct has a common start 5' sequence, common 3' sequence and a GAAA normalization hairpin. Three canonical pairs were used to connect the 6 junctions, normalization hairpin and common start and end sequences for each construct. Connecting helices were randomly generated and contained only AU and CG pairs. A candidate RNA was only kept if it was predicted secondary structure utilizing Vienna's folding algorithm matched the hypothetical target structure. Constructs with four or more consecutive GC pairs were also discarded even if predicted to fold properly with candidate motifs being returned to the original pool. Having more than three consecutive GC pairs is known to cause issues with premature stops during reverse transcription in the DMS workup of the RNA.^{20,24}

To ensure both broad coverage of non-canonical pairings and ease of analysis our goal was to generate a 7,500 sequence library. We found 7,500 to be the ideal size due to price and difficulty of designing larger libraries. With our design protocol we generated over 8,000 sequences. The final subset of 7,500 were selected using requirements designed to ensure successful sequencing and analysis. Sequences were selected such that the overall pool length variance is less than 10% to reduce PCR bias. Additionally we ensured that the Levenshtein edit distance between each construct is at least 10.²¹ Levenshtein distance describes the minimum edit distance

between two RNA strands and is important that individual sequence reads must be differentiated between possible RNA constructs. When a pool satisfying these requirements was created, there were approximately 8,000 sequences and the final pool was selected giving preference to those designs with lower ensemble defect calculated by Vienna's RNAfold. Selected sequences were then converted into DNA sequences by replacing uracil (U) with thymine (T) and adding the T7 promoter to the front of each sequence. The sequence pool was ordered from Agilent (product number G7220A).

1.2.2 Probing Libraries with DMS

The 7,500 sequence library was ordered through Agilent as a dry oligo pool. We resuspended the oligo pool with 50 μ L of 1X Tris-EDTA (TE) buffer from Fischer Bioreagents, part number BP2473-1. We PCR amplified the library with primers TTCTAATACGACTCAC-TATAGG, GTTGTTGTTGTTGTTTCTTT. We used q5 master mix from New England Biolabs, product number M0492L.

Step	Temperature	Time	Cycles
Initial Denaturation	98 C	2 mins	1
Denaturation	98 C	2 mins	18
Annealing	57 C	30 secs	18
Extension	72 C	30 secs	18
Hold	4 C	INF	

We purified the double stranded product by 2% agarose gel. What kit did we use to purify the double stranded product.

Resulting double stranded DNA (dsDNA) is then purified and mixed with 0.4 M Sodium Cacodylate (from Electron Microscopy Sciences, part number 11655) and 250 mM $MgCl_2$ (from

Alfa Aesar J61014) to re-fold RNA. Folded RNA is mixed with 2.5 uL of DMS (15 %/%) and left to react for 6 minutes. The reaction was quenched by BME from Agros Organics, part number 125470010. Reaction is purified by spin column and the concentration is measured by qubit (what kit was the procedure). The purified RNA is reverse transcribed by TGIRT-III for 2 hours. Product cDNA is again purified by spin column. Downstream PCR is next performed before an egel. A final spin column purification is performed before concentration is measured by qubit and the product is diluted down to 1 nM.

The prepared library was sent to the University of Kansas Medical Center for sequencing on an Illumina NovaSeq 6000 system. Library was part of a 1.13 billion read chip and received approximately 50 million of these reads directly. Resulting data was deposited in gunzipped fastq files.

1.2.3 Data Processing and Analysis

The zipped fastq files were decompressed and partitioned equally into 500 groups to make analysis practical on available hardware. Partitioning was performed by placing the n^{th} read into the $n \bmod 500^{\text{th}}$ file. For example, the 1st, 501st, 1001st, etc. read were placed into the first partition, the 2nd, 502nd, 1002nd were placed into the second partition, and so on. Bar-coding was performed on each of the 500 partitions using the novobarcode de-multiplexing application from novocraft. Once de-multiplexed, each of the 500 partitions was analyzed with the DREEM software package to align the reads in each partition to one of the 7,500 sequences in the pool.¹⁷ The DREEM pipeline further built mutational histograms for each of the 7,500 sequences in each of the 500 partitions. A mutational histogram is a representation of the mutation rates observed for each nucleotide in a sequence. These mutation rates are later used to generate

the reactivity profiles for each construct. Mutational histograms were finally combined across all partitions to yield 7,500 unique mutational histograms for the entire pool.

Normalization was performed using a GAAA tetraloop present in each of the 7,500 constructs. This normalization scheme takes the average reactivity of the three A's in the tetraloop and uses this value as the normalization factor the entire construct. This average is set to one and the other values in the construct are divided by this value. Once normalized, the A and C reactivity values for each of the motif sequences are aggregated and grouped across all constructs. This process was carried out with the assistance of a python script which decomposes each designed RNA into a motif graph which keeps track of the reactivity values across different constructs. The result is a number of measurements for each of the A's and C's in a given motif sequence.

1.2.4 Combining DMS Data with Solved Junction Structures

We curated a set junction motif structures from XRC PDB entries. Interactions with proteins, ligands, and other RNAs are known to impact three-dimensional conformation and as a result were removed from the set of PDB entries. A final quality cutoff was to remove entries where the resolution was greater than 4.0 Å. RNA pairing is driven by hydrogen bonds which occur on the range of 3.0 Å or less, meaning resolution worse than 4.0 Å could provide misleading results.²² This sampling of the PDB resulted in 659 unique sequence motifs with a total of 1,342 unique PDB entries.

We applied the X3DNA analysis tool to each of the 1,342 junctions to generate pair classifications the non-canonical base pairs present in each junction.¹⁴ Only pairs with at least one A or C were kept and their corresponding DMS values were extracted from previous analysis such that each pair, its classification and its values are all combined. For each pair, a number of values

were computed including pairwise minimum value, maximum value and the ratio of the higher and lower value when both nucleotides in the pair are DMS active. The ratio of a non-canonical pair where both nucleotides are DMS active (i.e. AA, CA, AC, CC) is calculated as follows with reactivities r_1 and r_2 :

$$ratio = \frac{\max(r_1, r_2)}{\min(r_1, r_2)}$$

Note that for this equation the minimum value is 1, representing a pair with equal DMS reactivities and a higher value corresponds to a pair where the two DMS values are very different.

1.3 Results and Discussion

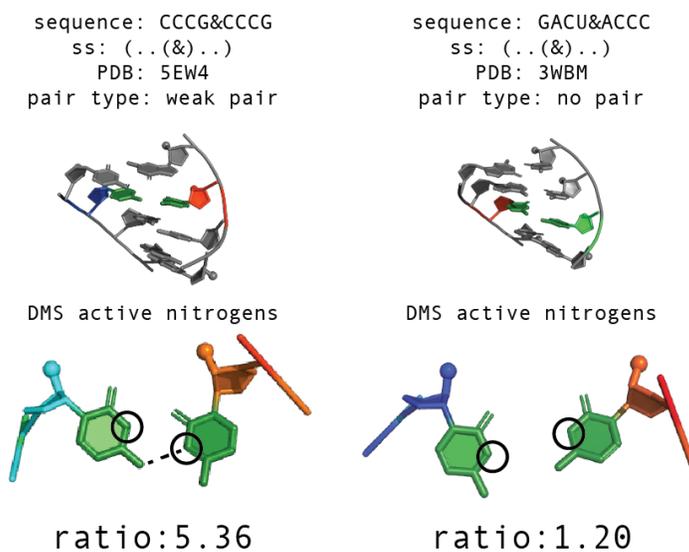
1.3.1 CC Pairs

DMS values for CC pairs signal if the pair is in one of two pairing conformations. Conventional analysis of DMS suggests that since CC pairs do not form canonical pairings, both nucleotides should have high reactivities as both N3's would not participate in pairing and thus be solvent exposed. During our analysis of CC pairs and their reactivity ratios, we saw a bi-modal set of values with ratios either being close to one or much higher than one. Two indicative examples of these values are shown in Figure 1.4.

Analysis with X3DNA showed that CC pairs with high ratios participated in a form of weak pairing where a single hydrogen bond is formed. This hydrogen bond interacts with the N3 of one cytosine but does not see participation from the other cytosine's N3. The specific orientation leaves one N3 shielded from solvent and the other N3 solvent exposed. As a result, the DMS reactivities for each C vary and result in a high ratio for the pair. This pairing mode is seen in the CCCG&CCCG motif on the left side of Figure 1.4. One of the weak pairing C-C pairs is shown in Figure 1.4 with the dotted line marking the hydrogen bond.

CC pairs with ratios close to 1 tended to not participate in pairing at all. This dynamic is seen on the right side of Figure 1.4 with the GACU&ACCC junction motif. The relevant CC pair from this motif features two C's that are too far apart to form a single hydrogen bond as in the CCCG&CCCG motif. Because both of the N3's are solvent exposed, the resulting in a ratio is closer to 1.

Figure 1.4: C-C DMS Values Correspond to Weak or No Pairing



Two examples of CC pairing. The junction motif on the left features a CC pair with weak pairing and a corresponding high DMS reactivity ratio. The relevant CC pair is shown in the bottom left with the DMS active N3's circled in black and the sole hydrogen bond represented by a dashed black line. The junction motif on the right features a CC pair with no pair and a DMS reactivity ratio near 1. Bottom right shows an enhanced view of the relevant pair with DMS active N3's circled in black.

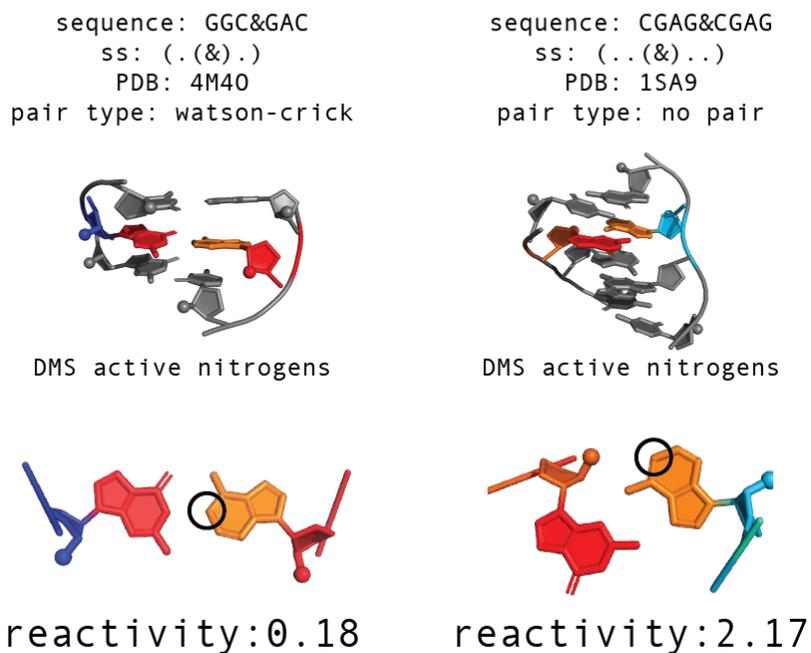
1.3.2 GA Pairs

DMS values for GA pairs provide binary information as to whether the nucleotides are participating in pairing or not. Analysis of GA pairs is limited to the magnitude from the DMS active adenine. In our analysis, we identified two modes of reactivity being either near zero or very high which correspond to being involved in a pair or sheared, respectively.

Inspection of 3D structures showed that GA pairs whose adenine had reactivity near zero tended to form a non-canonical GA pair. An example structure is shown in the left pane of Figure 1.5 with the GGC&GAC motif. The relevant GA pair is in close proximity and forms stabilizing hydrogen bonds, limiting solvent access to the involved adenine leading to a reactivity of 0.18. This pair's spatial arrangement is highlighted in the bottom left of Figure 1.5 with the DMS active N1 of adenine highlighted with a black circle.

GA pairs whose adenine had much higher reactivity adopted an alternate non-pairing configuration that exposes the N1 of adenine. An example structure is shown in the right pane of Figure 1.5 with the CGAG&CGAG motif. In this instance, the sheared conformation of the pair leaves the N1 of adenine exposed to solvent leading to a high reactivity value of 2.17. The arrangement of these nucleotides is shown in the bottom right of the figure where the exposed N1 of adenine is highlighted with a black circle.

Figure 1.5: GA Pairing Modes Exhibit Differing DMS Value Ratios



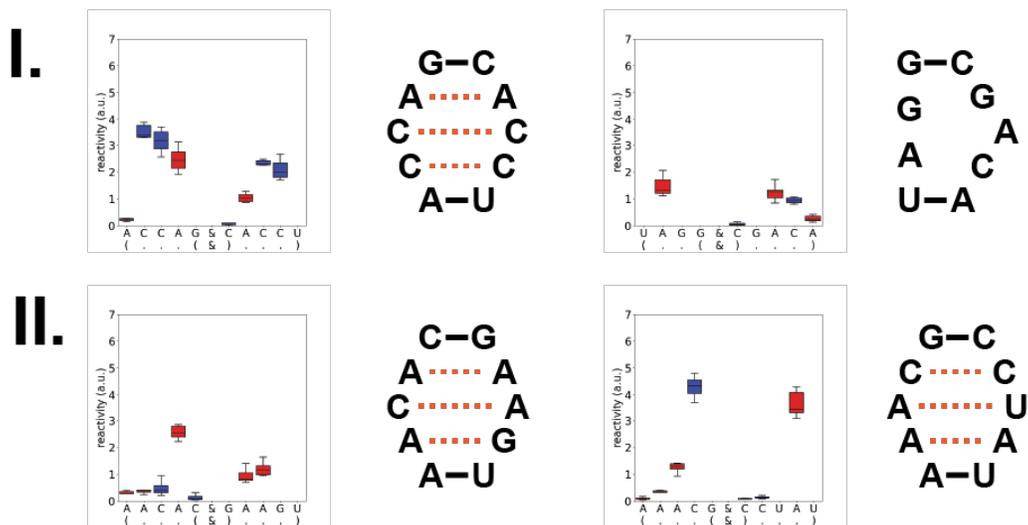
Two examples of GA pairing demonstrating the two modes observed in our curated PDB dataset. The junction on the right features a pairing mode which keeps the N1 of adenine hidden from solvent. The relevant pair is shown in plane in the bottom left. A motif with a sheared GA pair is shown on the left side of the figure. There a sheared conformation leaves the N1 of adenine exposed to solvent and able to react with DMS. An enhanced view of the relevant pair in the bottom right shows the N1 of adenine with a black circle.

DMS reactivity also identify a variety of junction topologies for junctions that do not have 3D crystallography data. Examples of different topologies are shown in Figure 1.6 and the selected junctions are indicative examples of patterns repeated observed in the full dataset.

Many observed junctions did conform to established intuition of DMS reactivity with canonically paired nucleotides having low reactivity values and non-canonically paired nucleotides having higher values as seen in row I of Figure 1.6. This highlights a key point that our analysis does not suggest that the current interpretation of DMS reactivity values is inaccurate. Instead, the established wisdom is largely correct but needs further refinement for some non-canonical interactions.

Row II of Figure 1.6 shows junctions which demonstrate the ability of non-canonically paired nucleotides to assume lower DMS reactivity values. As demonstrated by the analysis of GA pairs, lower DMS values are still consistent with non-canonical pairing. Given this, junctions with DMS profiles similar to those in row II correspond to well-formed structures with non-canonical interactions that shield the N1's and N3's of their respective adenine and cytosines. This finding represents a significant advance in our understanding of DMS reactivity as the C in the left loop of the bottom left structure in Figure 1.6 would previously be assumed to participate in a canonical pair elsewhere in its structure.

Figure 1.6: DMS Reactivity Values Identify Diverse Junction Topologies



The above junctions show the superposition of DMS reactivity profiles for each of the included junction motifs across a number of instances in the designed RNA library. Reactivity values for each nucleotide are shown as boxplots and are color coded with blue and red corresponding to adenine and cytosine, respectively. Each boxplot diagram is shown with its corresponding secondary structure and likely non-canonical basepairs are shown in dotted orange lines for the three symmetrical junctions presented.

Combining DMS data and X3DNA pair classifications enabled direct comparisons between 3D non-canonical pairs and reactivity values. In contrast to conventional analysis of DMS reactivity, values provide more information than a bi-modal classification of participation in canonical pairing or a lack thereof. Our deep dive into the CC and GA pairs demonstrates that low reactivity DMS values can be associated with non-canonical pairings and that DMS values additionally encode information about 3D pairing modes.

Beyond specific pairing examples, analysis of our designed RNA library shows that the current analysis of DMS reactivities is not wholly lacking. Many observed junctions conformed to conventional wisdom with canonically paired and loop nucleotides adopting low and high DMS reactivity values, respectively. A number of loop nucleotides did record low reactivity values, however, and it is clear the bi-modal classification does not always hold. The junction profiles presented in Figure 1.6 also demonstrate that DMS reactivity profiles are consistent across different sequence contexts. This finding also suggests that low reactivity values for loop nucleotides are not an anomaly as each motif compiled into the boxplot diagrams would need to misfold similarly across all instances for this to occur.

1.4 Summary and Future Directions

Our deep dive into DMS patterns of non-canonically paired nucleotides demonstrates that there is a wealth of structural information demanding further analysis. Importantly, the existence of patterns for both CC and GA pairs shows that DMS can provide information on non-canonical pairs with one or two DMS active nucleotides. Of the ten possible non-canonical pairs, only GG and UU cannot be analyzed via DMS reactivity studies. As a result, further studies provide an opportunity to significantly improve experimental coverage of non-canonical pairs. Our work has demonstrated that DMS studies can efficiently canvas the vast space of possible junctions. The creation of only a few more junction libraries could provide enough values to build robust datasets describing non-canonical pairing.

Improved datasets then provide opportunity to enhance the efficacy computational modeling. DMS reactivities are frequently combined with *in silico* prediction algorithms to guide higher precision folding.²³ Existing approaches use DMS values as an additional constraint, assuming that reactivities correlate with the “pairedness” of a given nucleotide. The understanding that lower DMS reactivity values are consistent with non-canonical pairing will improve algorithm accuracy as these values will no longer fuel the prediction of erroneous pairs.

Beyond incorporation into thermodynamic folding models, non-canonical DMS datasets provide an opportunity to develop algorithms for the prediction of reactivity profiles. The ability to predict a DMS profile for an RNA provides an alternate method for structure elucidation and validation. Direction comparison of predicted and actual DMS reactivity profiles is rapid and would reduce reliance on the imperfect thermodynamic models common in the field.

References

1. Yao, R.-W.; Wang, Y.; Chen, L.-L. Cellular Functions of Long Noncoding RNAs. *Nature Cell Biology*, 2019, 21, 542–551. <https://doi.org/10.1038/s41556-019-0311-8>.
2. Payne, J. L.; Khalid, F.; Wagner, A. RNA-Mediated Gene Regulation Is Less Evolvable than Transcriptional Regulation. *Proceedings of the National Academy of Sciences*, 2018, 115, E3481–E3490. <https://doi.org/10.1073/pnas.1719138115>.
3. Gry, M.; Rimini, R.; Strömberg, S.; Asplund, A.; Pontén, F.; Uhlén, M.; Nilsson, P. Correlations between RNA and Protein Expression Profiles in 23 Human Cell Lines. *BMC Genomics*, 2009, 10. <https://doi.org/10.1186/1471-2164-10-365>.
4. Pardi, N.; Hogan, M. J.; Porter, F. W.; Weissman, D. mRNA Vaccines — a New Era in Vaccinology. *Nature Reviews Drug Discovery*, 2018, 17, 261–279. <https://doi.org/10.1038/nrd.2017.243>.
5. Vandivier, L. E.; Anderson, S. J.; Foley, S. W.; Gregory, B. D. The Conservation and Function of RNA Secondary Structure in Plants. *Annual Review of Plant Biology*, 2016, 67, 463–488. <https://doi.org/10.1146/annurev-arplant-043015-111754>.
6. Lemieux, S. RNA Canonical and non-canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research*, 2002, 30, 4250–4263. <https://doi.org/10.1093/nar/gkf540>.
7. Turner, D. H.; Mathews, D. H. NNDB: The Nearest Neighbor Parameter Database for Predicting Stability of Nucleic Acid Secondary Structure. *Nucleic Acids Research*, 2009, 38, D280–D282. <https://doi.org/10.1093/nar/gkp892>.
8. Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Research*, 2003, 31, 3406–3415. <https://doi.org/10.1093/nar/gkg595>.
9. Lorenz, R.; Bernhart, S. H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 2011, 6. <https://doi.org/10.1186/1748-7188-6-26>.
10. Zadeh, J. N.; Steenberg, C. D.; Bois, J. S.; Wolfe, B. R.; Pierce, M. B.; Khan, A. R.; Dirks, R. M.; Pierce, N. A. NUPACK: Analysis and Design of Nucleic Acid Systems. *Journal of Computational Chemistry*, 2010, 32, 170–173. <https://doi.org/10.1002/jcc.21596>.
11. Reuter, J. S.; Mathews, D. H. RNAstructure: Software for RNA Secondary Structure Prediction and Analysis. *BMC Bioinformatics*, 2010, 11. <https://doi.org/10.1186/1471-2105-11-129>.
12. Holbrook, S. R.; Kim, S.-H. RNA Crystallography. *Biopolymers*, 1997, 44, 3–21. [https://doi.org/10.1002/\(sici\)1097-0282\(1997\)44:1<3::aid-bip2>3.0.co;2-z](https://doi.org/10.1002/(sici)1097-0282(1997)44:1<3::aid-bip2>3.0.co;2-z).
13. Zhang, K.; Li, S.; Kappel, K.; Pintilie, G.; Su, Z.; Mou, T.-C.; Schmid, M. F.; Das, R.; Chiu, W. Cryo-EM Structure of a 40 KDa SAM-IV Riboswitch RNA at 3.7 Å Resolution. *Nature Communications*, 2019, 10. <https://doi.org/10.1038/s41467-019-13494-7>.
14. Lu, X.-J.; Bussemaker, H. J.; Olson, W. K. DSSR: An Integrated Software Tool for Dissecting the Spatial Structure of RNA. *Nucleic Acids Research*, 2015, gkv716. <https://doi.org/10.1093/nar/gkv716>.
15. Harris, M. E.; Christian, E. L. RNA Crosslinking Methods. *Methods in Enzymology*, 2009, 127–146. [https://doi.org/10.1016/s0076-6879\(09\)68007-1](https://doi.org/10.1016/s0076-6879(09)68007-1).

16. Zubradt, M.; Gupta, P.; Persad, S.; Lambowitz, A. M.; Weissman, J. S.; Rouskin, S. DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing in Vivo. *Nature Methods*, 2016, 14, 75–82. <https://doi.org/10.1038/nmeth.4057>.
17. Aviran, S.; Lucks, J. B.; Pachter, L. RNA Structure Characterization from Chemical Mapping Experiments. 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2011. <https://doi.org/10.1109/allerton.2011.6120379>.
18. Tijerina, P.; Mohr, S.; Russell, R. DMS Footprinting of Structured RNAs and RNA–Protein Complexes. *Nature Protocols*, 2007, 2, 2608–2623. <https://doi.org/10.1038/nprot.2007.380>.
19. Butcher, S. E.; Pyle, A. M. The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Accounts of Chemical Research*, 2011, 44, 1302–1311. <https://doi.org/10.1021/ar200098t>.
20. Eddy, S. R. How Do RNA Folding Algorithms Work? *Nature Biotechnology*, 2004, 22, 1457–1458. <https://doi.org/10.1038/nbt1104-1457>.
21. Yujian, L.; Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29, 1091–1095. <https://doi.org/10.1109/tpami.2007.1078>.
22. Halder, A.; Data, D.; Seelam, P. P.; Bhattacharyya, D.; Mitra, A. Estimating Strengths of Individual Hydrogen Bonds in RNA Base Pairs: Toward a Consensus between Different Computational Approaches. *ACS Omega*, 2019, 4, 7354–7368. <https://doi.org/10.1021/acsomega.8b03689>.
23. Lorenz, R.; Hofacker, I. L.; Stadler, P. F. RNA Folding with Hard and Soft Constraints. *Algorithms for Molecular Biology*, 2016, 11. <https://doi.org/10.1186/s13015-016-0070-z>.
24. Cordero, P.; Kladwang, W.; VanLang, C. C.; Das, R. Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference. *Biochemistry*, 2012, 51, 7037–7039. <https://doi.org/10.1021/bi3008802>.
25. Schroeder, S. J.; Turner, D. H. Optical Melting Measurements of Nucleic Acid Thermodynamics. *Methods in Enzymology*, 2009, 371–387. [https://doi.org/10.1016/s0076-6879\(09\)68017-4](https://doi.org/10.1016/s0076-6879(09)68017-4).
26. Andronescu, M.; Condon, A.; Hoos, H. H.; Mathews, D. H.; Murphy, K. P. Computational Approaches for RNA Energy Parameter Estimation. *RNA*, 2010, 16, 2304–2318. <https://doi.org/10.1261/rna.1950510>.
27. Tijerina, P.; Mohr, S.; Russell, R. DMS Footprinting of Structured RNAs and RNA–Protein Complexes. *Nature Protocols*, 2007, 2, 2608–2623. <https://doi.org/10.1038/nprot.2007.380>.

CHAPTER 2: Developing LC-MS Functionality for MVAPACK

2.1 The Need For Complete Metabolomics Software Suites

Metabolomics aims to understand cellular processes through the comprehensive characterization of small molecule metabolites.¹ Small molecules are the final product of many cell processes and can be found in a variety of tissues, cell lysates and biofluids.²⁻⁴ An organism's metabolome is the composite of these small molecules and is typically composed of thousands of metabolites.⁵ The vast number of known metabolites and their corresponding high abundances in the cell make the metabolome a robust proxy for understanding fundamental processes and disease states alike. Accurate identification and measurement of small molecules is critical to characterizing an organism's metabolome. Need for high fidelity metabolite measurement has made ¹H nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC-MS) standard techniques for any metabolomics studies.⁶⁻⁷

¹H NMR and LC-MS are complementary techniques that make coverage of the metabolome a tractable task.⁸ Although identification and measurement of isolated small molecules by either technique is highly accurate today, biological mixtures feature increased noise and interactions that complicate analysis. The rapid runtimes, non-destructive nature, and high reproducibility of NMR make it ideal for identifying new biomarkers in a metabolomics study.⁹ In contrast, the greater volume and labelling capabilities of LC-MS make the technique effective for further analysis of the metabolome once biologically relevant metabolites have been identified (Figure 2.1).¹⁰ For both NMR and LC-MS, the volume of data produced is impractical for manual analysis. Researchers in the field have addressed the abundance of generated data through the development of software pipelines that automate analysis and aggregation of metabolite measurements.

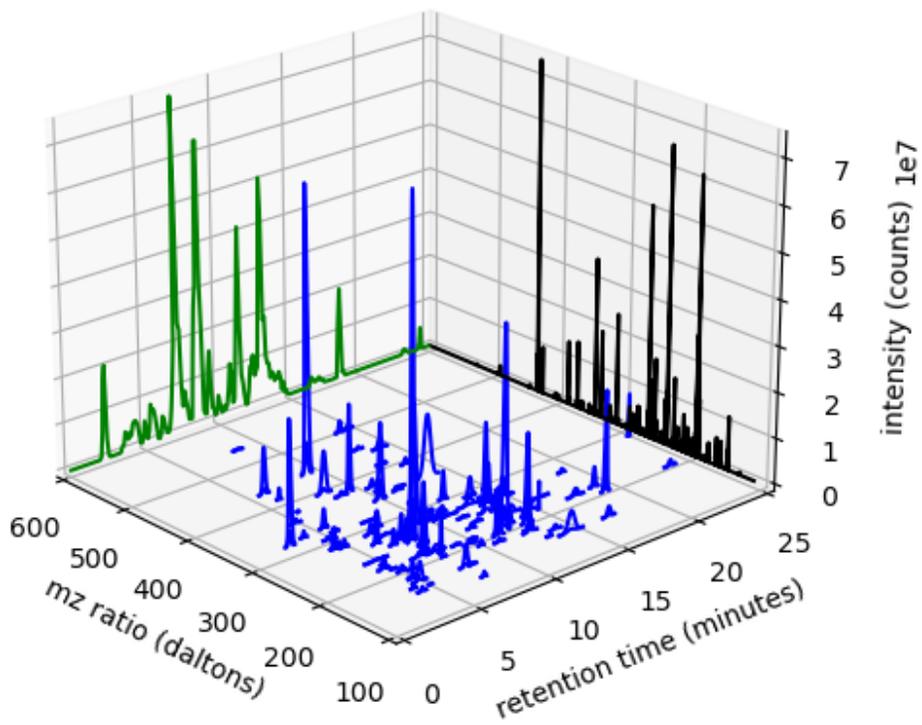
Metabolomics software packages make metabolomics analyses a tractable task by automating metabolite identification, quantitation, and statistical analysis. Software pipelines begin with automated peak picking that enables rapid identification of metabolites across dozens of biological replicates.¹¹ With modern computational power, selection of metabolite peaks is performed in minutes and results in a high dimensional description of the metabolome. Making sense of the resulting peak features is then achieved via the preparation of a feature matrix which aggregates metabolite abundances across all replicates in a study. Ensuring accurate grouping of features is another non-trivial task that relies on statistical models to cluster analogous metabolite peaks across different samples. Feature matrices are further refined by selecting only metabolite features which appear consistently within a biological replicate group and vary versus other groups. The resulting feature matrix contains metabolites which describe the underlying biological differences between groups. Feature matrices contain hundreds of metabolites and the field has turned to statistical models like principal component analysis (PCA), partial least squares (PLS) and orthogonal projection onto latent structures (OPLS) to enable rapid visualization of these high dimensional spaces (Figure 2.2).¹² Distilling metabolomic spectral data into statistical models for convenient visualization and analysis is universal in the field, and this functionality is provided by dozens of software packages.

Despite the abundance of metabolomics packages, most limit users to analyzing either NMR or LC-MS data. OpenMS, XCMS, and Maven are popular programs for analyzing LC-MS data but have no facilities for raw NMR data.¹³⁻¹⁵ NMRProcFlow, NMRPipe and MVAPACK all provide solutions for NMR metabolomics analyses but leave users unable to work with LC-MS data.¹⁶⁻¹⁸ The lack of a metabolomics software package offering both LC-MS and NMR function-

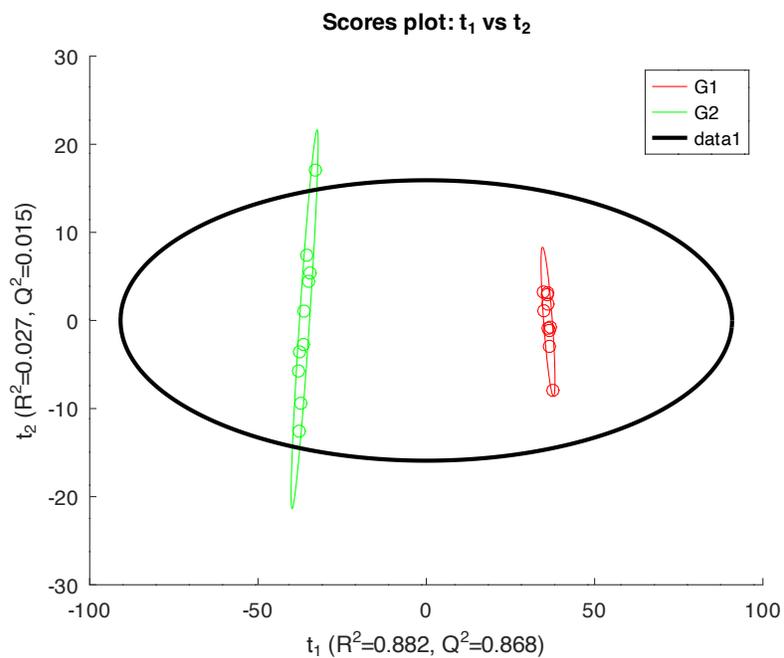
ality forces researchers to either limit their analyses to one spectral type or invest time into learning multiple software packages. Creating a package capable of analyzing both ^1H NMR and LC-MS data would alleviate these problems, allowing users to leverage the information provided by both spectral types. We believe MVAPACK is an ideal software package to address these needs. MVAPACK features a modular approach with existing PCA, OPLS and PLS modeling capabilities. As a result, end-to-end analysis of LC-MS data would not require re-implementation of these techniques.

Adding field standard LC-MS processing techniques to MVAPACK will transform the package into a unique one stop shop for metabolomics analysis. The wealth of available LC-MS processing techniques provides ample algorithmic options for each step of analysis. As a result, adding LC-MS processing to MVAPACK has been an exercise in implementation rather than novel algorithm development. The following chapter outlines my work in adding field-standard LC-MS processing techniques to MVAPACK.

Figure 2.1: LC-MS Data Allows High Precision Metabolite Identification



LC-MS data is a three dimensional description of both the retention time and mass-to-charge ratio of sampled compounds. Access to both dimensions enables high precision identification of metabolites.

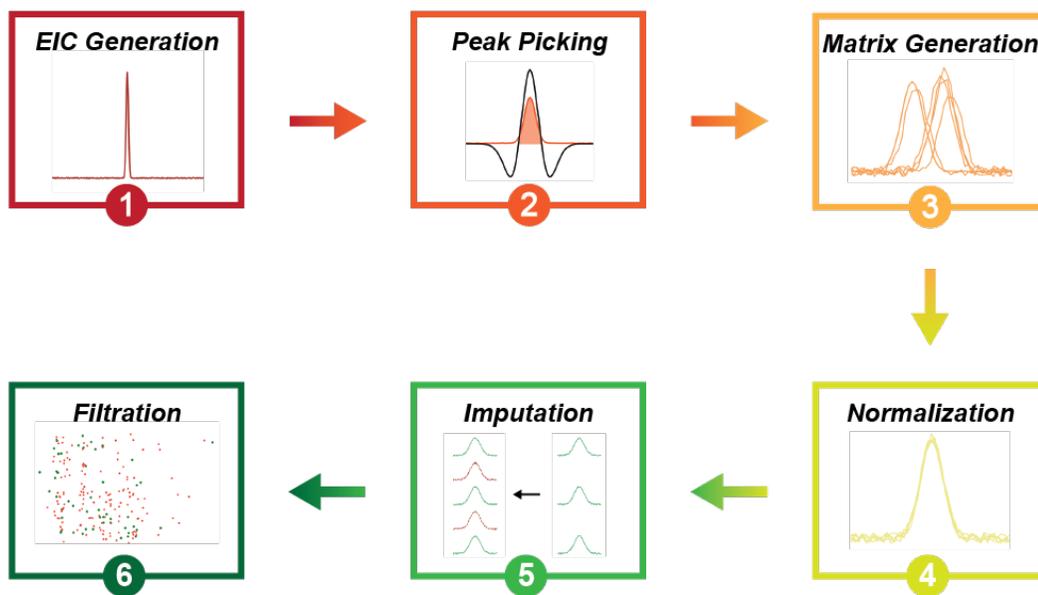
Figure 2.2: PCAs Allow Visualization of High Dimensional Spaces

PCA plots cluster samples using their similarities in variance. Here each dot represents all of the peaks in a given experimental replicate. Replicates are color coded and the green and red ellipses correspond to the 95% confidence levels for each respective group. The black ellipse is the 95% confidence interval for all data in the PCA analysis.

2.2 Materials and Methods

MVAPACK's new LC-MS functionality utilizes previously published algorithms accepted as standard in the metabolomics community. This functionality is incorporated consistently with the rest of MVAPACK, utilizing a modular approach where each step of the LC-MS data pipeline is performed by a single function call with multiple choices for analytical method. The steps are as follows: (1) Extracted Ion Chromatogram (EIC) generation, (2) peak picking, (3) peak matrix generation, (4) peak matrix normalization, (5) peak matrix imputation and (6) peak matrix filtration. An overview of the new LC-MS functionality added to MVAPACK is presented in Figure 2.3.

Figure 2.3: MVAPACK LC-MS Pipeline Overview



A summary of MVAPACK's LC-MS processing functionality. Six total steps are performed to go from raw LC-MS data to a filtered feature matrix.

2.2.1 EIC Generation

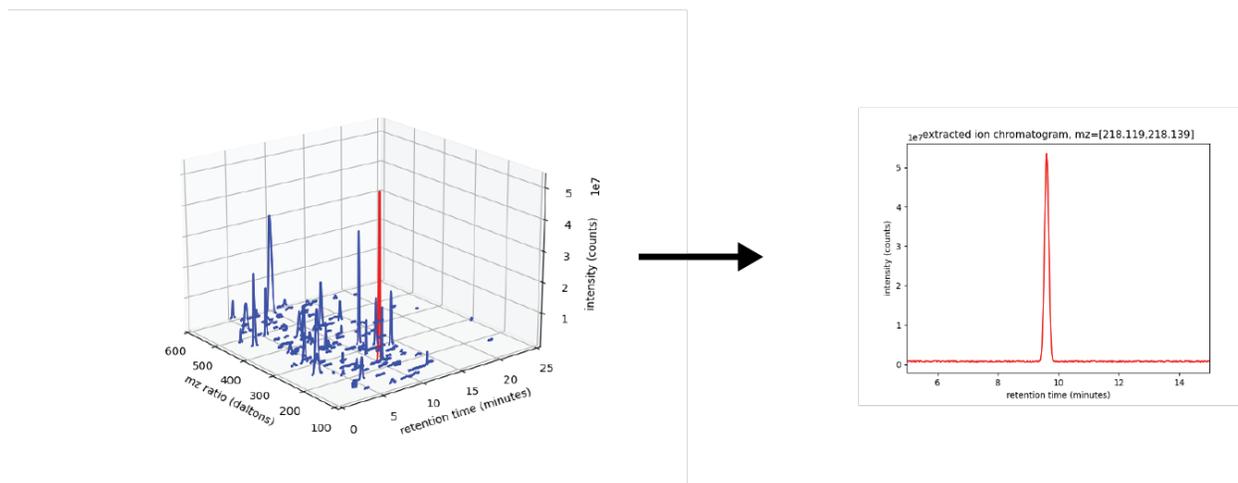
An EIC is a group of LC-MS intensities (cts) with very similar mass-to-charge ratios (mz) ordered by retention time (rt) in ascending order. EIC generation is essential for peak picking as the initially 3D LC-MS data is simplified to a collection of 2D spectra (Figure 2.4). In our implementation, we utilize a strategy similar to that described in XCMS.²⁴ A single function call to `create_eics()` converts a replicate file containing raw scans into an EIC file with uniform binning. The width of mz binning is a user specified value with a default of 0.05 daltons. Acceptable input file formats include the partially binary encoded .mzML and .mzXML as well as the plaintext proteoWizard .txt format.¹⁹

Replicate files often exceed 10 gigabytes (GB) and a backend C++ parser was written to ensure reasonable EIC generation times. Octave is a high level language whose scripts are interpreted by a Java Virtual Machine (JVM). Using a JVM is convenient for language implementation purposes but comes at the cost of execution speed. Additionally, decompressing the binary encoded portions of .mzML and .mzXML files is a difficult task in Octave. The C++ language is known for high performance and ability to work with binary data, making it a clear choice for a parser backend. Octave additionally provides an application program interface (API) to communicate with compiled C++ code. Developing a C++ backend allowed MVAPACK's LC-MS functionality to overcome Octave's performance bottlenecks and integration with the Octave C++ API hides implementation details from end-users.

Our implementation identifies the supplied file format and performs validation checks prior to EIC generation. Checks are specific to each of the three possible file formats and ensure the data is not mal-formed. Once input data is validated, EICs are generated by sorting each data

point into its appropriate m/z bin. Multiple points often exist within the same m/z bin and the single value with the highest intensity is kept for each retention time point within a given EIC. This step is designed to remove baseline noise points. After generation, EICs are saved to a plaintext file so that replicates files do not have to be parsed again for future analysis.

Figure 2.4: EICs Isolate Single Peaks



EIC generation reduces 3D LC-MS data to a collection of 2D spectra. The above example shows how a sample peak shown in red is extracted from the full spectra on the left to make a 2D spectra on the right.

2.2.2 Peak Picking

The next step in MVAPACK's LC-MS pipeline is peak picking via the `pick_peaks()` function. Selected peaks correspond to individual metabolites and are used in the final feature matrix, making accurate peak picking a priority. Additionally, ordinary replicates have in excess of 20,000 EICs, which adds performance considerations to our implementation. To provide users with EIC peak picking that is both accurate and efficient, we have developed two wave-form based options for selecting LC-MS peaks in MVAPACK.

The first option for peak picking is a Gaussian second derivative wavelet similar to that described in XCMS.¹⁴ Applying a wavelet to a raw signal series creates a waveform describing the curvature of the original signal at each point. As seen in Figure 2.5, a Gaussian second derivative transform will cross zero near peak boundary regions. Locations of zero crossings are impacted by the size of the wavelet being used so the full width at half maximum (FWHM) of the raw signal is taken and used to inform the Gaussian second derivative wavelet size. Using this description of peak curvature, peaks are defined as regions between zero crossings where the maximum peak value of the raw series is above some signal-to-noise or intensity threshold. The default cutoff in MVAPACK is being greater than 10 times the average intensity in the current EIC. To address performance concerns, we built a C++ function to apply the Gaussian second derivative wavelet to the raw signal. Like the C++ backend parser used for EIC generation, this code utilizes Octave's C++ API and is an implementation detail that users do not need to worry about.

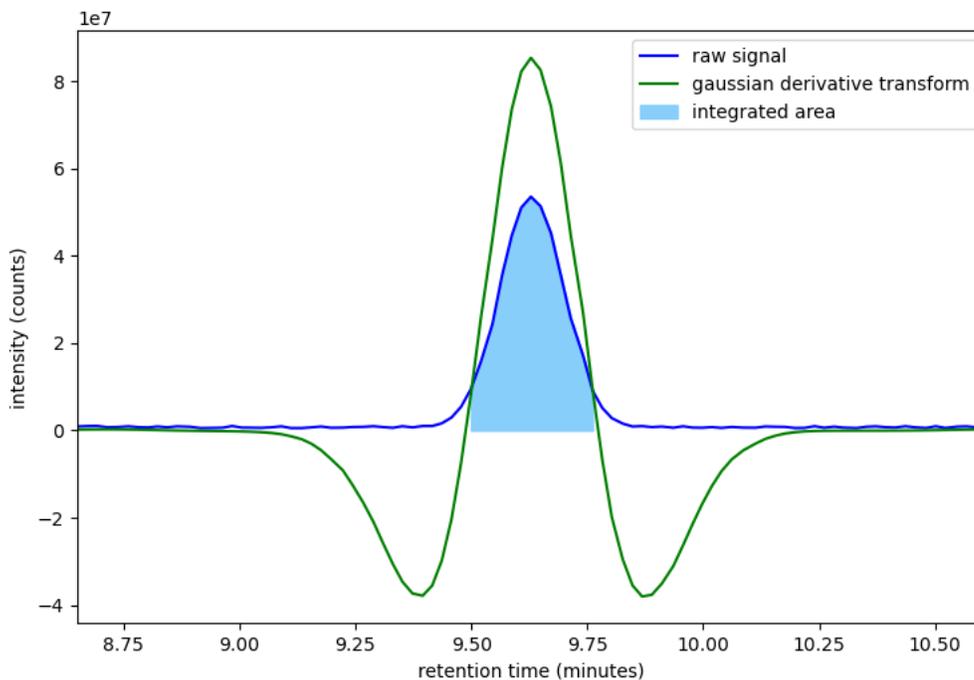
The second option for peak picking is a Savitzky-Golay filter. Savitzky-Golay filters perform a smoothing of the raw signal by making each point in a waveform the weighted average of nearby points.²⁰ Like the Gaussian second derivative transform, Savitzky-Golay filtration creates

a waveform that measures the slope of the original signal. It also sees zero crossing similar to the Gaussian second derivative transform in peak regions. As a result, Savitzky-Golay peak picking uses an identical protocol for actual peak picking wherein the area between zero crossings is integrated to arrive at the total intensity for the peak. Octave provides a performant Savitzky-Golay filter implementation which we use in MVAPACK.

Beyond wavelet transformation strategy, both the Gaussian second derivative transform and Savitzky-Golay filter have the same behavior. Each method is capable of selecting multiple peaks per EIC up to a specified number. Likewise, each approach finds the rt, mz, max intensity, integration and width for each peak. Integrations are performed via trapezoidal integration in the regions between zero crossings and widths are found by measuring the FWHM for each peak.

The result of the `pick_peaks ()` function is a matrix containing all peaks for a replicate. Each row of the matrix corresponds to a single peak with columns holding values for the mz, rt, intensity, integration and width. MVAPACK offers a utility function named `save_peaks ()` that exports a peak matrix to a comma separated values (CSV) format.

Figure 2.5: Gaussian Derivative Transforms Identify Peak Regions



The application of a Gaussian second derivative wavelet (in green) to a raw EIC (in blue). Integration of the peak is performed on the area between zero crossings where the maximum intensity satisfies a specified cutoff.

2.2.3 Matrix Generation

After selecting peaks from each replicate, MVAPACK provides the `generate_matrix()` function to aggregate metabolite features globally. Matrix generation is critical for downstream statistical analysis via PCA, OPLS or PLS and requires clustering in two dimensions across all available replicates. Peak clustering is a difficult but necessary task as both biological and instrument-based noise lead to drift even in high quality datasets.²¹ Similar to EIC generation and peak picking, this step requires both performance and accuracy. For MVAPACK, we have provided users with two implementation options: an OpenMS style root mean square difference (RMSD) approach as well as an ObiWarp correlated time warping method.^{13,22}

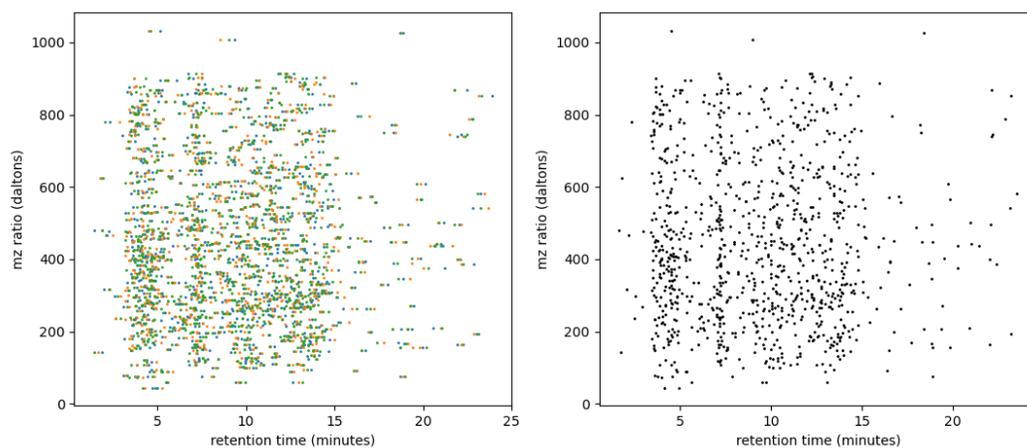
The OpenMS RMSD approach views peaks as points in an (rt,mz) coordinate space and groups replicate peaks with low corresponding distances. The distance between two peaks from different replicates is defined below:

$$Distance = \sqrt{\frac{(mz_1 - mz_2)^2}{mz_scale_factor} + \frac{|rt_1 - rt_2|}{rt_scale_factor}}$$

Where mz_1 and rt_1 corresponds to p_1 and mz_2 and rt_2 correspond to p_2 , respectively. The `mz_scale_factor` and `rt_scale_factor` are constants set by the user with default values of 0.05 dalton and 90 seconds, respectively. This RMSD metric addresses the anisotropic nature of LC-MS data. Drifts in retention time are common and expected to be larger whereas drifts in mz ratio are rare and expected to be very low.²¹ Taking the square of the mz difference and the absolute value of the retention time difference allows larger changes in retention time and penalizes changes in mz ratio. Using a smaller `mz_scale_factor` and a larger `rt_scale_factor` further promotes alignments that avoid changes in mz but allow movement in retention time space.

Each candidate peak is merged into the existing cluster with the lowest RMSD distance. Distances are calculated by looping through candidate peaks and making every possible peak vs cluster comparison. The result is a large number of comparisons which would be slow to perform if implemented in Octave. We have implemented this part of MVAPACK's LC-MS pipeline in C++ to make rapid peak alignment possible. This routine is called through the Octave C++ API and this implementation detail is hidden from the user.

The other option for peak clustering is the ObiWarp time warping algorithm. ObiWarp takes a top down approach to feature grouping, viewing each replicate as a matrix of raw intensities to globally align via warping of retention times.²² Time warping is a popular technique used by dozens of alignment algorithms in the field.²³ ObiWarp's implementation was made fully open source by its creators and as a result has been included in a number of packages including BioConductor, XCMS and Maven.^{14,15} This algorithm aggregates a replicate's intensities into a matrix and minimizes the correlated difference between a given replicate and a reference replicate via retention time adjustments. Correlated distances can be measured with Pearson's correlation coefficient, covariance, dot product or Euclidean distance. ObiWarp already being written in C++ allowed for convenient inclusion into MVAPACK as an efficient method for peak clustering. Beyond implementation details, this algorithm's top-down approach makes it ideal for datasets where large retention time drifts occur. Unlike the RMSD approach, it is effective at performing accurate multi-minute alignments.

Figure 2.6: Matrix Generation

The feature clustering step of MVAPACK's LC-MS data pipeline aligns peaks across different replicates. Seen on the left are collections of peaks color coded to represent different replicates. The right shows the aligned peaks which have been clustered to eliminate differences in retention times. Each black dot corresponds to a cluster of peaks across different replicates.

2.2.4 Feature Normalization

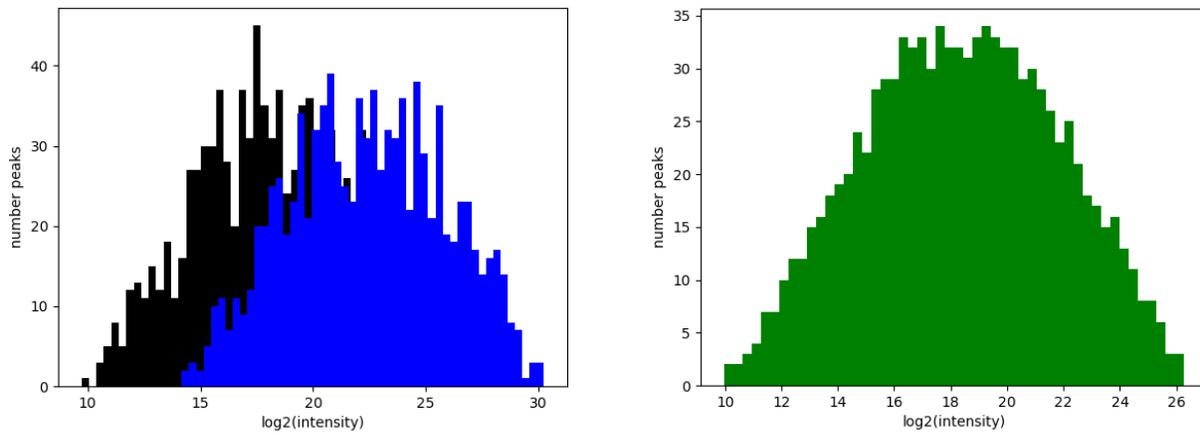
LC-MS metabolomics data experiences significant variance from both biological and instrumental sources. Despite its prevalence, signal fluctuations must be addressed to enable accurate inter-replicate quantitation and downstream statistical analysis.²⁴ Without normalization, statistical models have difficulty discerning biologically significant signal variation from random noise. Metabolomics researchers have developed a consensus approach of first taking the base two logarithm of peak intensities before applying normalization schemes to minimize the variation present in datasets (Figure 2.2.7). We have developed the `normalize_matrix()` function to combat this variation. Our implementation in MVAPACK provides three protocols for normalization: maximum value normalization, p-norm normalization and quantile normalization.

Maximum value normalization adjusts for peak variation by setting the most intense peak in each replicate to one and scaling the rest of the peaks to that value. It is the simplest approach available in MVAPACK and is useful for datasets with lower variation. This normalization scheme will not be effective for many datasets but tends not to distort the underlying data significantly.

P-norm normalization is a normalization scheme frequently used to adjust high dimensional machine-learning regression models.²⁶ Each replicate's normalization factor is generated by summing the cumulative intensity of peaks. This method makes use of Octave's built in norm function and creates a normalization factor that includes information about all peaks in the replicate.

Quantile normalization is a higher order technique standard in many LC-MS metabolomics packages. This normalization scheme assumes that each replicate has comparable distribu-

tions of intensities and makes their respective composites identical.²⁷ MVAPACK's quantile normalization implementation first sorts all intensities in each replicate within a matrix where each column represents a different replicate. Each intensity value is then set to the average intensity found in each row. Although this approach sees more change to the underlying values, the lack of parameterization provides a strong check against overfitting.

Figure 2.7: Matrix Normalization

Two distributions of un-normalized peak intensities are seen on the left with different replicates being colored in black and blue, respectively. After normalization, differences between replicate peak intensities are eliminated as seen by the green distributions on the right.

2.2.5 Peak Imputation

In addition to inter-replicate variation, LC-MS metabolomics experiments experience random missing peaks for both biological and instrumental reasons. Peaks are known to be missing completely at random (MCAR) in up to 20% of replicates even in high quality LC-MS datasets.²⁸ Filling in missing values is critical for final statistical models which see performance decreases when feature values are missing.²⁹ We have provided the `impute()` function in MVAPACK to give users techniques for imputing their LC-MS peak matrices. The `impute()` function has three available protocols: mean imputation, mean distribution imputation and kNN imputation.

Mean imputation sets missing peak to the average of other intensities for the same feature within a given replicate group. The mean approach is a simple first order technique for addressing missing values. An example is shown in Figure 2.2.8. A major limitation to mean imputation is the tendency for this technique to create imputed features with artificially low variation. In particular, the more peaks missing from a given feature, the greater the reduction in variation. The side effect of reduced variation is less pronounced when more replicates are present in an experimental group.

Mean distribution imputation behaves similarly to mean imputation and but also adds stochastic noise to the imputed peaks to reduce the variation reduction effect. This protocol measures both the mean and standard deviation of the existing features for a given experimental group. The measured mean and standard deviation are used to establish a hypothetical value distribution for each feature which is sampled to impute missing peaks. Adding random noise addresses the tendency of mean imputation to artificially lower feature variation, preserving existing trends in the feature matrix.

kNN imputation is a machine learning method for imputation that predicts missing values using observed relationships between peak intensities.³⁰ Unlike both mean and mean distribution, kNN considers more than the non-zero peak values for a given feature. kNN instead identifies patterns between peaks within each replicate and uses these patterns to inform imputation. Similar to mean distribution, kNN imputation incorporates variation into its estimates as neighbor peaks used will vary in relative abundances between replicates.

Figure 2.8: Peak Imputation Accounts For Missing LC-MS Peaks

		feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_N
group_1	rep_1	17.28	14.60	15.11	14.08	15.08	14.02	14.20
	rep_2	17.02	14.97	15.42		14.94	14.22	14.50
	rep_3		15.03	15.09	14.38	15.20	14.10	14.14
	rep_4	17.28	14.74	15.43	14.05	14.86	13.93	14.60
	rep_5	17.22	15.03	15.40	14.17	15.14		14.21
group_2	rep_6	17.85	16.50	14.47	15.07	13.77	11.76	16.22
	rep_7	17.82		14.49	14.80	13.64	11.68	16.04
	rep_8	17.69	16.58		14.81	14.00	11.84	15.87
	rep_9	17.58	16.47	14.50	14.96	13.66	11.85	16.33
	rep_10	17.87	16.48	14.69	14.84	13.74		15.86

		feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_N
group_1	rep_1	17.28	14.60	15.11	14.08	15.08	14.02	14.20
	rep_2	17.02	14.97	15.42	14.17	14.94	14.22	14.50
	rep_3	17.20	15.03	15.09	14.38	15.20	14.10	14.14
	rep_4	17.28	14.74	15.43	14.05	14.86	13.93	14.60
	rep_5	17.22	15.03	15.40	14.17	15.14	14.06	14.21
group_2	rep_6	17.85	16.50	14.47	15.07	13.77	11.76	16.22
	rep_7	17.82	16.51	14.49	14.80	13.64	11.68	16.04
	rep_8	17.69	16.58	14.54	14.81	14.00	11.84	15.87
	rep_9	17.58	16.47	14.50	14.96	13.66	11.85	16.33
	rep_10	17.87	16.48	14.69	14.84	13.74	11.78	15.86

The top matrix describes shows a feature matrix with a number of missing peaks highlighted in red. The bottom matrix has imputed these missing values with the mean values for each of the existing peaks in each replicate group, highlighted in green.

2.2.6 Peak Filtration

Peak filtration is performed as a final processing step to select metabolites that best describe variation in the biological system at hand. Desirable metabolite features have low variation within their experimental groups and large variation across experimental groups. Features with these characteristics are useful for building statistical models and understanding biological processes alike (Figure 2.2.9). We have developed the `filter_features()` function for MVAPACK users to select significant peaks. This function has three filtering modes: maximum variance, max fold change, and analysis of variance (ANOVA).

Maximum variance, or coefficient of variation (CV), describes the variation present within an experimental group for a given feature. Maximum variance is measured for each experimental group by dividing the standard deviation of the group's intensities by the mean of the group's intensities. A feature is only kept when CV's for each experimental group are below a given cutoff, typically 0.20 or 0.15. The maximum variation filter directly targets features with low inter-group variation.

Max fold change describes the variation present between experimental groups for each feature. First, the ratio of average intensity differences between each of the experimental groups is found. The max fold change is the greatest of these values and a feature is only kept when it is above a specified cutoff, usually 2. The max fold change filter directly targets features with high variation between experimental groups.

ANOVA measures variance both within and between experimental groups. Unlike maximum variance and max fold change, ANOVA relies on the F statistical test to classify grouped variation.³¹ As a result, the ANOVA test generates a p-value which indicates if the variation is

statistically significant. A feature is only kept when this p-value is below a specified cutoff. The default value is 0.05, corresponding to a 95% probability that the variation is significant.

Figure 2.9: Peak Filtration Selects Significant Features

		feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_N
group_1	rep_1	10.09	15.08	20.25	11.97	11.67	20.30	8.19
	rep_2	7.07	18.21	15.84	11.47	13.63	23.65	8.94
	rep_3	8.64	20.34	11.85	8.45	14.91	15.02	8.24
	rep_4	7.04	18.55	21.80	15.21	15.24	25.10	5.87
	rep_5	7.33	20.35	16.47	8.39	18.19	18.02	6.37
group_2	rep_6	10.22	18.04	11.73	13.87	21.48	25.55	10.31
	rep_7	12.31	20.89	15.48	9.86	19.15	23.61	7.99
	rep_8	12.54	22.59	12.17	15.31	23.78	22.35	10.79
	rep_9	13.48	20.61	20.38	12.38	21.64	25.71	10.06
	rep_10	13.03	25.87	15.52	15.20	15.48	28.14	10.85
	group_1_cv	0.16	0.12	0.23	0.26	0.16	0.20	0.18
	group_2_cv	0.10	0.13	0.23	0.17	0.16	0.09	0.12
	fold change	4.28	3.09	2.19	2.22	5.58	4.65	2.48
	Keep?	keep	keep	drop	drop	keep	drop	keep

This example feature matrix is composed of two experimental groups with five replicates each. Max variance (CV) and fold change has been calculated for each of the features and only those with CV below 0.20 and fold change greater than 2 are kept.

2.3 Results and Discussion

This work represents an expansion of MVAPACK's functionality to handle both ^1H NMR and LC-MS datasets. LC-MS processing has been implemented as a modular process with six sequential steps: (1) EIC generation, (2) peak picking, (3) peak matrix generation, (4) peak matrix normalization, (5) peak matrix imputation and (6) peak matrix filtration. Each step has been implemented with user-friendly wrapper functions that make use of different protocols convenient. Different analytical techniques for each step can be specified with the same function calls, limiting the learning curve for these new functions. Octave compatible C++ code has also been developed to improve performance for EIC generation, peak picking and matrix generation. Using C++ for performance critical aspects of the pipeline makes MVAPACK's performance competitive with existing commercial and open-source packages. This performance is helpful for widespread adoption as users will not have to sacrifice analytical run time to analyze both ^1H NMR and LC-MS data. The included techniques also represent standard approaches in the field for LC-MS metabolomics analysis. Protocols have been borrowed from established packages like XCMS, OpenMS, Maven and ObiWarp, allowing users to directly translate existing processing scripts from other packages for use in MVAPACK while still using existing analytical techniques.^{13-15,21} This potential for easy conversion is another strength of MVAPACK's new functionality as users will not have to limit their analytical options to process both spectral data types.

Adding an LC-MS pipeline has made MVAPACK a unique one-stop shop for metabolomics analysis. MVAPACK users can now extract metabolite features from both spectral data types, providing a unique analytical offering not seen elsewhere in the metabolomics community. This work represents a unique effort in the field that enables users greater coverage of the metab-

olome without having to make large sacrifices in terms of execution speed or processing abilities. New functionality has also been implemented in manner consistent with existing NMR functionality, providing a convenient experience for end-users. As a result, LC-MS data can now be incorporated into a study with only minor alterations to existing MVAPACK scripts. Using a wrapper function approach is also conducive to the addition of new functionality over time. The presented work represents the development of an LC-MS infrastructure for MVAPACK, opening the door for future developers to new analytical protocols with ease.

2.4 Summary and Future Directions

MVAPACK's new LC-MS functionality has been implemented in a consistent, modular format that opens the door for the addition of new protocols. MVAPACK's LC-MS functionality represents common techniques in the field, but a number of additional algorithms and processing steps can still be included. While every LC-MS processing package provides facilities for peak picking, alignment, and so forth, few provide multiple options for each step. MVAPACK's current offering is comparable to existing package in terms of analytical techniques, but further addition of published protocols would elevate the package. BioConductoR is the poster child for this concept, providing dozens of techniques for each step of LC-MS metabolomics processing. The clear next step for this project is the addition of more protocols for each processing step. Providing users with more options will lead to increase usage in the community as researchers see the value in using MVAPACK. As a result, combined ^1H NMR and LC-MS studies will become more prevalent and enhanced coverage of the metabolome will be seen throughout the field.

An additional next step is to validate MVAPACK's LC-MS functionality through data benchmarking. Comparing the package's results on the same datasets versus existing analytical techniques will first aid future development. Documenting current performance is important for the addition of new functionality so that improvements and pessimization can be directly identified. The analysis of benchmark data provides the opportunity to profile performance and identify bottlenecks and potential bugs in the package. Benchmarking also allows future developers to understand the strengths and weaknesses of newly added functionality. Validation is also important for user confidence and widespread adoption. Users will be more likely to adopt a new package if they are confident in its ability to accurately characterize their datasets.

References

1. Idle, J. R.; Gonzalez, F. J. *Metabolomics*. *Cell Metabolism*, 2007, 6, 348–351. <https://doi.org/10.1016/j.cmet.2007.10.005>.
2. Luo, X.; Li, L. *Metabolomics of Small Numbers of Cells: Metabolomic Profiling of 100, 1000, and 10000 Human Breast Cancer Cells*. *Analytical Chemistry*, 2017, 89, 11664–11671. <https://doi.org/10.1021/acs.analchem.7b03100>.
3. Ramautar, R.; Mayboroda, O. A.; Somsen, G. W.; de Jong, G. J. *CE-MS for Metabolomics: Developments and Applications in the Period 2008-2010*. *ELECTROPHORESIS*, 2010, 32, 52–65. <https://doi.org/10.1002/elps.201000378>.
4. Huang, Q.; Tan, Y.; Yin, P.; Ye, G.; Gao, P.; Lu, X.; Wang, H.; Xu, G. *Metabolic Characterization of Hepatocellular Carcinoma Using Nontargeted Tissue Metabolomics*. *Cancer Research*, 2013, 73, 4992–5002. <https://doi.org/10.1158/0008-5472.can-13-0308>.
5. Nobeli, I.; Thornton, J. M. *A Bioinformatician's View of the Metabolome*. *BioEssays*, 2006, 28, 534–545. <https://doi.org/10.1002/bies.20414>.
6. Wishart, D. S. *Quantitative Metabolomics Using NMR*. *TrAC Trends in Analytical Chemistry*, 2008, 27, 228–237. <https://doi.org/10.1016/j.trac.2007.12.001>.
7. Zhou, B.; Xiao, J. F.; Tuli, L.; Ransom, H. W. *LC-MS-Based Metabolomics*. *Mol. BioSyst.*, 2012, 8, 470–481. <https://doi.org/10.1039/c1mb05350g>.
8. Pan, Z.; Raftery, D. *Comparing and Combining NMR Spectroscopy and Mass Spectrometry in Metabolomics*. *Analytical and Bioanalytical Chemistry*, 2006, 387, 525–527. <https://doi.org/10.1007/s00216-006-0687-8>.
9. Emwas, A.-H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Gowda, G. A. N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. *NMR Spectroscopy for Metabolomics Research*. *Metabolites*, 2019, 9, 123. <https://doi.org/10.3390/metabo9070123>.
10. Gika, H. G.; Wilson, I. D.; Theodoridis, G. A. *LC-MS-Based Holistic Metabolic Profiling. Problems, Limitations, Advantages, and Future Perspectives*. *Journal of Chromatography B*, 2014, 966, 1–6. <https://doi.org/10.1016/j.jchromb.2014.01.054>.
11. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. *Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics*. *Metabolites*, 2018, 8, 31. <https://doi.org/10.3390/metabo8020031>.
12. Worley, B.; Powers, R. *Multivariate Analysis in Metabolomics*. *Current Metabolomics*, 2012, 1, 92–107. <https://doi.org/10.2174/2213235x11301010092>.
13. Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *OpenMS – An Open-Source Software Framework for Mass Spectrometry*. *BMC Bioinformatics*, 2008, 9. <https://doi.org/10.1186/1471-2105-9-163>.
14. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification*. *Analytical Chemistry*, 2006, 78, 779–787. <https://doi.org/10.1021/ac051437y>.
15. Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. *LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine*. *Current Protocols in Bioinformatics*, 2012. <https://doi.org/10.1002/0471250953.bi1411s37>.

16. Jacob, D.; Deborde, C.; Lefebvre, M.; Maucourt, M.; Moing, A. NMRProcFlow: A Graphical and Interactive Tool Dedicated to 1D Spectra Processing for NMR-Based Metabolomics. *Metabolomics*, 2017, 13. <https://doi.org/10.1007/s11306-017-1178-y>.
17. Delaglio, F.; Grzesiek, S.; Vuister, Geerten W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *Journal of Biomolecular NMR*, 1995, 6. <https://doi.org/10.1007/bf00197809>.
18. Worley, B.; Powers, R. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, 2014, 9, 1138–1144. <https://doi.org/10.1021/cb4008937>.
19. Adusumilli, R.; Mallick, P. Data Conversion with ProteoWizard MsConvert. *Methods in Molecular Biology*, 2017, 339–368. https://doi.org/10.1007/978-1-4939-6747-6_23.
20. Press, W. H.; Teukolsky, S. A. Savitzky-Golay Smoothing Filters. *Computers in Physics*, 1990, 4, 669. <https://doi.org/10.1063/1.4822961>.
21. Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C. Critical Assessment of Alignment Procedures for LC-MS Proteomics and Metabolomics Measurements. *BMC Bioinformatics*, 2008, 9. <https://doi.org/10.1186/1471-2105-9-375>.
22. Prince, J. T.; Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Analytical Chemistry*, 2006, 78, 6140–6152. <https://doi.org/10.1021/ac0605344>.
23. Smith, R.; Ventura, D.; Prince, J. T. LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review. *Briefings in Bioinformatics*, 2013, 16, 104–117. <https://doi.org/10.1093/bib/bbt080>.
24. Mizuno, H.; Ueda, K.; Kobayashi, Y.; Tsuyama, N.; Todoroki, K.; Min, J. Z.; Toyo'oka, T. The Great Importance of Normalization of LC-MS Data for Highly-Accurate Non-Targeted Metabolomics. *Biomedical Chromatography*, 2016, 31, e3864. <https://doi.org/10.1002/bmc.3864>.
25. Webb-Robertson, B.-J. M.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; Waters, K. M. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *Journal of Proteome Research*, 2015, 14, 1993–2001. <https://doi.org/10.1021/pr501138h>.
26. Gentile, C. *Machine Learning*, 2003, 53, 265–299. <https://doi.org/10.1023/a:1026319107706>.
27. Hicks, S. C.; Okrah, K.; Paulson, J. N.; Quackenbush, J.; Irizarry, R. A.; Bravo, H. C. Smooth Quantile Normalization. *Biostatistics*, 2017, 19, 185–198. <https://doi.org/10.1093/biostatistics/kxx028>.
28. Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D. Normalization and Missing Value Imputation for Label-Free LC-MS Analysis. *BMC Bioinformatics*, 2012, 13. <https://doi.org/10.1186/1471-2105-13-s16-s5>.
29. Josse, J.; Pagès, J.; Husson, F. Multiple Imputation in Principal Component Analysis. *Advances in Data Analysis and Classification*, 2011, 5, 231–246. <https://doi.org/10.1007/s11634-011-0086-7>.
30. Lee, J. Y.; Styczynski, M. P. NS-KNN: A Modified k-Nearest Neighbors Approach for Imputing Metabolomics Data. *Metabolomics*, 2018, 14. <https://doi.org/10.1007/s11306-018-1451-8>.

31. Kim, T. K. Understanding One-Way ANOVA Using Conceptual Figures. *Korean Journal of Anesthesiology*, 2017, 70, 22. <https://doi.org/10.4097/kjae.2017.70.1.22>.

CHAPTER 3: Creating Synthetic Data to Validate MVAPACK's New Functionality

3.1 The Need for Synthetic LC-MS Datasets

Benchmarking LC-MS metabolomics software is a critical final step to method development.¹ Using benchmarks allows developers to understand algorithm behavior and provides users assurance that a package can be used reliably in their own work. Critical analysis of an LC-MS package also informs potential users about the strengths and weaknesses of a given implementation, guiding its usage in a metabolomics study. Despite the importance of benchmarking, there is no clear consensus on how algorithm performance should be measured.²

A common benchmarking technique is to run a previously analyzed dataset through new software and compare results. This approach is suitable if existing software provides a robust description of the benchmark dataset, but it is difficult to know if the original analysis obtained full and accurate coverage of the data. Biological noise, quantitation limits, and complex inter-metabolite interactions result in unidentified and missing peaks in even the simplest of LC-MS datasets.³ The lack of an established ground truth in experimental datasets poses a direct problem to comparative validation. While various software packages will correctly analyze large, easily identifiable peaks, the propensity for mischaracterization of edge cases to occur reduces the efficacy of current benchmarking approaches. Given the inherent nature of noise and missing peaks in experimental LC-MS datasets, an alternative approach could improve validation and benchmarking of new software.

Using simulated LC-MS data directly addresses issues with conventional benchmarking of metabolomics software. Simulated data has ground truth, with each peak's waveform and metabolite identity being fully characterized.⁴ Full understanding of simulated spectra enables straightforward assessment of algorithmic performance. Features selected by an algorithm can be

compared directly to the known peak population in the spectra enabling calculations of false positive and false negative rates. Ground truth understanding becomes especially useful when synthetic datasets incorporate noise and missing values to mimic the qualities of real spectra. Reducing spectral quality enables stress testing of peak picking, imputation, and normalization methods, providing feedback on their sensitivity to these factors.⁶ These goals cannot be achieved with experimental data as missing values and noise are not tunable factors. Despite the clear advantages of using simulated LC-MS data for benchmarking and validating metabolomics software for MVAPACK's validation and benchmarking phase.⁸ To achieve our goal of validating MVAPACK, we developed our own synthetic LC-MS data using existing simulation software as a starting point.

To effectively validate MVAPACK, we have created a synthetic dataset that contains wide metabolite variation and spectral quality. Using the existing ViMMS software package as a source of peaks, we have built an idealized metabolite peak library.⁵ To emulate the size and experimental grouping characteristics seen in real LC-MS metabolomics experiments, two groups of ten replicates were created from the same base set of metabolite features. Each replicate's features were given multipliers to ensure measurable and significant variation in the overall system. Lastly, we probe MVAPACK's ability to deal lower spectral quality data through the creation of 8 additional dataset versions with varying levels of missing peaks and added noise.

3.2 Materials and Methods

3.2.1 Data Generation

Simulated metabolite peaks from the ViMMS software package were used as the base feature set for our LC-MS metabolomics dataset. ViMMS generates data by first sampling the human metabolome databank (HMDB) and then using kernel density machine learning algorithms to predict the shapes of metabolite peaks. The package generates isotoped features for over 19,000 known metabolites and we initially created all possible metabolites. To ensure well-defined peak waveforms in the final data, the profiles of each metabolite were manually inspected and 7,000 peaks with multiple isotopic peaks were selected. Common reasons for peak exclusion were large baseline noise, baseline drift, the existence of too many small peaks in an extracted ion chromatogram (EIC), or a lack of data points within a single EIC.⁷ Selected peaks were idealized through Gaussian fitting. Fitted models were used to generate the ideal EICs for each of the molecules (Figure 3.1). At this point, we had created a base library of metabolites for building simulated spectra.

To ensure our synthetic dataset has overall behavior similar to LC-MS metabolomics datasets, we designed our dataset to have two groups of ten replicates each. After the master set of all metabolite EICs was created, each replicate was generated by systematically varying the intensities of each EIC. We randomly selected 935 of the 4,680 metabolites to vary at statistically significant levels across replicate groups. Significance of metabolite variation is tied to coefficient of variation (CV), defined as the standard deviation of multipliers divided by the average of these multipliers as well as fold change, the ratio of average metabolite intensity between groups. In our dataset, significant features have $CV < 20\%$ for both groups and fold change greater than or equal to two. Multipliers with a desired variance level can be generated trivially by creating a set of random multipliers and keeping them if their measured CV is on a desired range. The 935 metabolites with significant variation were each given multipliers with $CV < 20\%$ and all others

were given CV's > 30%. Significant metabolites were given inter-group fold changes between 2.2 and 4 whereas non-significant were given values ranging from 0 to 1.7. Combining fold changes and CV's, we arrived at a full multiplier set for each of the 4,680 metabolites in each of the twenty simulated replicates. The full multiplier set was validated for significance through the generation of principal component analysis (PCA) plots.

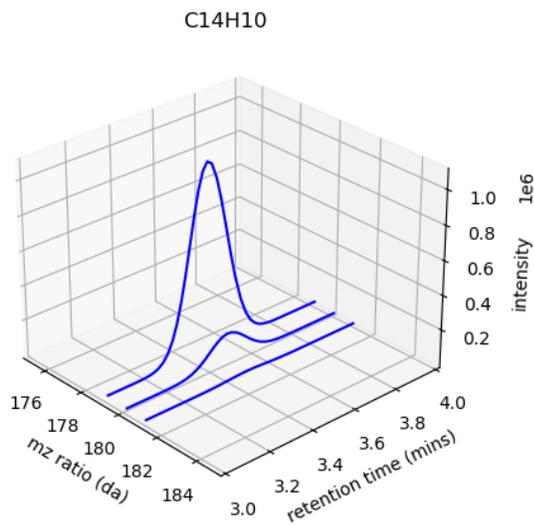
After creating the base set of replicates, we created eight additional sets with identical raw peaks but varying levels of noise and missing features. We created versions of the simulated data with varying quality to stress test MVAPACK's processing abilities. The addition of noise was implemented at the metabolite EIC level by randomly adding signal to the baseline of each waveform. We added baseline noise as a percentage of the maximum intensity in each EIC and used levels of 0%, 5% and 10% (Figure 3.2). Missing peaks were removed at random from the subset of significant peaks at rates of 0%, 10% and 20%. Removal of significant peaks was performed such that at most 20% of the peaks were missing from each feature across all replicates in an experimental group. This choice was made to ensure that significant peaks are kept in the analysis as features with more than 20% of peaks missing are discarded in MVAPACK. Between noise and missing value possibilities, we generated a total of nine distinct datasets with varying levels of data quality.

3.2.2 Analysis with MVAPACK

Data was analyzed with MVAPACK using a standard processing script. After data was converted from .mzML format to EICs, peak picking was performed using a Gaussian Second Derivative wavelet transform and a signal-to-noise cutoff of 10 was used.⁹ Peak alignment was performed using an OpenMS-style root mean square deviation (RMSD) approach and default mass-to-charge (mz) and retention time (rt) scaling factors of 0.05 daltons and 120 seconds were

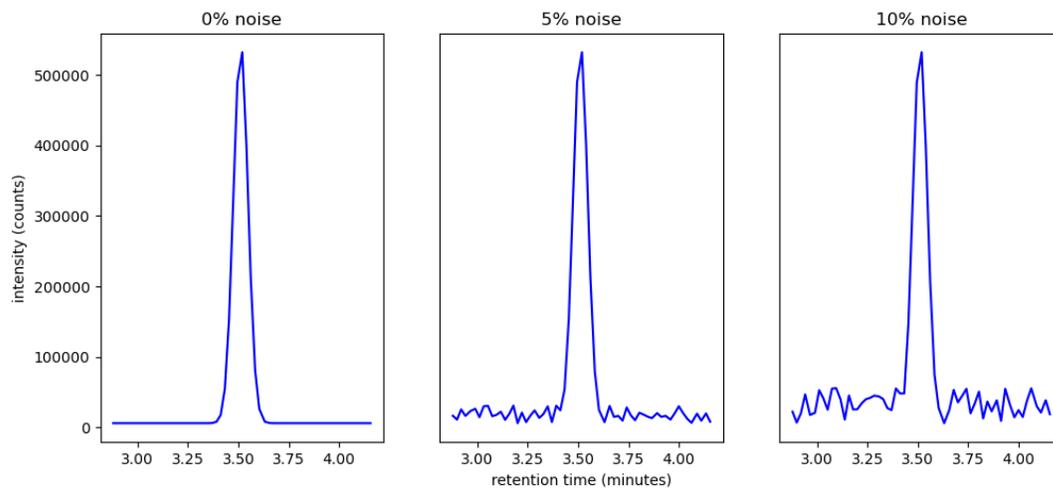
used, respectively.¹⁰ A non-parametric quantile normalization scheme was used to normalize the resulting data matrix and missing peaks were imputed using basic mean imputation.¹¹ The resulting data matrices for each of the nine conditions were modelled using principal components analysis (PCA).⁸

Figure 3.1: Idealized EICs Are Generated For Each Metabolite



Idealized metabolite EICs for the molecule anthracene. Three different peaks are seen which have each been idealized to a Gaussian waveform.

Figure 3.2: Noise Is Added at Three Levels to Vary Spectra Quality

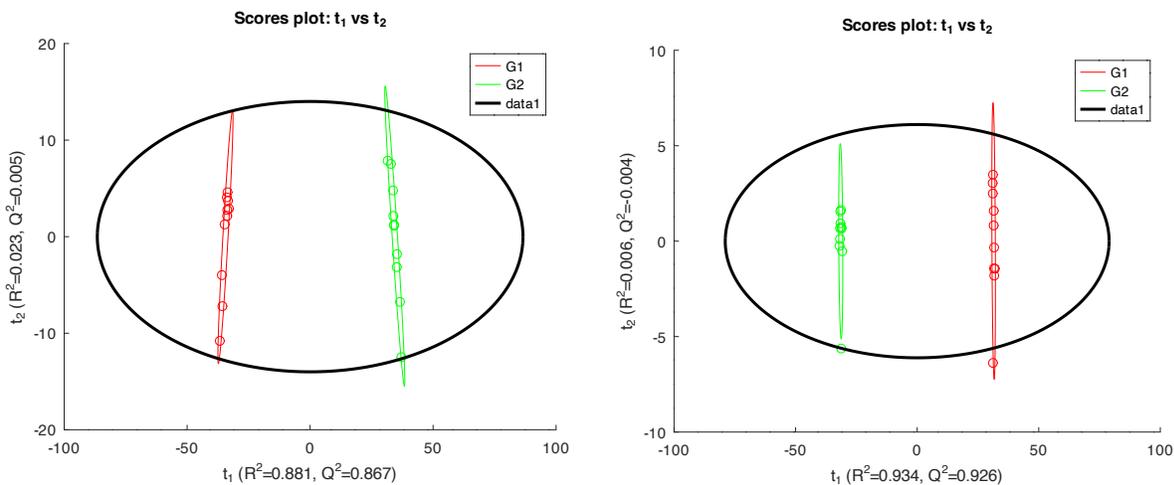


Three levels of noise that have been added to the EICs during replicate generation. Noise is generated by creating an array of random values on the range of $[0, \%max_peak]$ which is then added to the baseline.

3.3 Results and Discussion

Our protocol led to the successful creation of a simulated LC-MS dataset that contains both significant variation amongst metabolite peaks as well as varying levels of quality. When analyzed with MVAPACK, we saw strong agreement between expected results and the package's characterization of the data. PCA plots describe the degree to which variation is explainable via high level components. Pearson R^2 values are generated for each component with better data having the sum of the first two components as close to one as possible. Figure 3.3 shows that for the base case with no additional noise or peaks removed, MVAPACK performs well, showing a cumulative R^2 of 0.904 for a dataset expected to have a cumulative R^2 of 0.940. This close agreement in terms of variation explained suggests that MVAPACK is accurately characterizing the data it analyzes when there is high spectral quality. Analysis of spectra with varying quality suggests that MVAPACK still processes the data well regardless of noise levels but struggles as peaks are removed at random (Figure 3.4). Specifically we saw that cumulative R^2 pessimization was driven only by the number of missing peaks. For a given missing peak level, pessimization was nearly constant with respect to changes in noise level. Missing peaks are a problem for MVAPACK however, seeing pessimization of 13% and 25% for data with 10% and 20% of peaks missing, respectively. This finding suggests that the imputation models in MVAPACK needs improvement.

Figure 3.3: MVAPACK's PCA Model Closely Matches the Idealized Version



The above data is for the simulated data case with no noise and no missing peaks. The left shows the PCA plot generated by MVAPACK and the right shows the idealized version.

Figure 3.4: MVAPACK's Performance Across Spectral Quality Level

Pessimization of First Two Principal Components

		percent missing		
		0%	10%	20%
noise level	0%	0.36%	13.00%	22.00%
	5%	0.36%	13.40%	25.50%
	10%	0.36%	13.10%	26.90%

Summary of MVAPACK's pessimization across varying spectral quality. Pessimization is defined as the difference of the ideal model's first two component R^2 and the same first two component R^2 from MVAPACK's model.

3.4 Summary and Future Directions

Here we describe the successful generation of a synthetic LC-MS dataset for the purpose of validating a metabolomics software package. By using ViMMS as a source for peak information, we have created data that closely mirrors feature sets seen in real LC-MS datasets. Idealizing these peaks has also enabled high precision tuning of spectral quality. Creating nine different sets of data with spectral quality has been critical for understanding MVAPACK's performance sensitivity. While it performs well when all peaks are present, performance decreases are seen as peaks are randomly removed. In addition, we saw that MVAPACK is robust to the addition of baseline noise. This study suggests that further refinement is needed for MVAPACK's imputation algorithms.

Further improvement can be made to our simulated data through the addition of baseline drift as well as retention time drift. Both are common features of LC-MS data and their inclusion would enable further stress testing of MVAPACK and other algorithms. Retention time drift in particular is worthwhile to investigate as feature alignment is a major component of most metabolomics packages and is challenging to benchmark especially with real experimentally acquired datasets.

References

1. Eghbalnia, H. R.; Romero, P. R.; Westler, W. M.; Baskaran, K.; Ulrich, E. L.; Markley, J. L. Increasing Rigor in NMR-Based Metabolomics through Validated and Open Source Tools. *Current Opinion in Biotechnology*, 2017, 43, 56–61. <https://doi.org/10.1016/j.copbio.2016.08.005>.
2. Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Analytical Chemistry*, 2005, 78, 567–574. <https://doi.org/10.1021/ac051495j>.
3. Trötz Müller, M.; Guo, X.; Fauland, A.; Köfeler, H.; Lankmayr, E. Characteristics and Origins of Common Chemical Noise Ions in Negative ESI LC-MS. *Journal of Mass Spectrometry*, 2011, 46, 553–560. <https://doi.org/10.1002/jms.1924>.
4. Schulz-Trieglaff, O.; Pfeifer, N.; Gröpl, C.; Kohlbacher, O.; Reinert, K. LC-MSsim – a Simulation Software for Liquid Chromatography Mass Spectrometry Data. *BMC Bioinformatics*, 2008, 9. <https://doi.org/10.1186/1471-2105-9-423>.
5. Wandy, J.; Davies, V.; J. J. van der Hooft, J.; Weidt, S.; Daly, R.; Rogers, S. In Silico Optimization of Mass Spectrometry Fragmentation Strategies in Metabolomics. *Metabolites*, 2019, 9, 219. <https://doi.org/10.3390/metabo9100219>.
6. Awan, M. G.; Awan, A. G.; Saeed, F. Benchmarking Mass Spectrometry Based Proteomics Algorithms Using a Simulated Database. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2021, 10. <https://doi.org/10.1007/s13721-021-00298-3>.
7. Zhang, W.; Zhao, P. X. Quality Evaluation of Extracted Ion Chromatograms and Chromatographic Peaks in Liquid Chromatography/Mass Spectrometry-Based Metabolomics Data. *BMC Bioinformatics*, 2014, 15. <https://doi.org/10.1186/1471-2105-15-s11-s5>.
8. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 2012, 1, 92–107. <https://doi.org/10.2174/2213235x11301010092>.
9. Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 2006, 78, 779–787. <https://doi.org/10.1021/ac051437y>.
10. Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nature Methods*, 2016, 13, 741–748. <https://doi.org/10.1038/nmeth.3959>.
11. Hicks, S. C.; Okrah, K.; Paulson, J. N.; Quackenbush, J.; Irizarry, R. A.; Bravo, H. C. Smooth Quantile Normalization. *Biostatistics*, 2017, 19, 185–198. <https://doi.org/10.1093/biostatistics/kxx028>.

CHAPTER 4: Summary of Work

This thesis described my work in identifying markers of non-canonical pairing in DMS data as well as developing LC-MS functionality for the MVAPACK metabolomics package and its subsequent validation through simulated LC-MS data.

We demonstrated that DMS experiments do encode information about 3D structure and specifically non-canonical pairings. Presented examples for CC and GA pairs show that low values which previously would have been previously interpreted as canonical pairing can be consistent with non-canonical pairing modes. This work opens the door for further studies of relationships between DMS reactivity modes and non-canonical pairing as well as the establishment of quantitative models.

We also successfully added functionality for LC-MS metabolomics analysis to MVAPACK, making it a one-stop shop for metabolomics. New functionality represents popular techniques in the field, providing users with re-implementations of previously described algorithms. Implementation was performed in a granular manner such that future developers can incorporate additional analytical techniques. Adding other algorithms is the clear next step for this project as providing more choices for users will lead to more widespread usage in the metabolomics community.

We generated a simulated LC-MS dataset for validating MVAPACK. This data is designed to test the package's ability to generate full PCA models from metabolomics LC-MS data. Our approach features both empirically informed EICs as well as finely tuned intensities and spectral quality levels. Tuning these parameters is critical for the dataset as it allows for stress testing and robust assessment of MVAPACK or any other metabolomics package. Further work for this project would include the addition of baseline drift as well as retention time drift.