Strategies for Improving Multiomic Metabolite Identifications Using Compound Libraries, Machine Learning, and Structural Mass Spectrometry

By

Jaqueline-Mae Arenas Picache

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Chemistry
September 30, 2020
Nashville, Tennessee

Approved:

John A. McLean, Ph. D.

Renã A. Robinson, Ph. D.

Lauren E. Buchanan, Ph. D.

D. Borden Lacy, Ph. D.

This work is dedicated to the rare disease community.

Thank you for your inspiration and strength.

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Dr. John A. McLean, for his guidance and support. I greatly appreciate the opportunity to work in the McLean research group alongside some inspirational individuals. My time in the McLean group was filled with innovation, exploration, and fulfillment. I was able to satisfy my scientific curiosities and have multiple opportunities to share my work with others in various settings for which I am ever grateful for. It has been my sincerest honor to have completed my dissertation research in the McLean research group.

I would also like to thank my dissertation committee members, Dr. Lauren E. Buchanan, Dr. D. Bordon Lacy, and Dr. Reña A. Robinson, for all of their insight and direction. They've helped me be the best scientist I can be and remained steadfast in their support. Their advice has pushed me to think critically and broaden the scope of my work.

This research would not be possible without my colleagues in the McLean research group. Their years of continued support, innovated discussions, side-by-side teaching and learning, as well as motivation and laughter have meant the world to me. Dr. Jody C. May, Dr. Stacy D. Sherrod, Dr. Alexandra Shrimpe-Rutledge, and Dr. Simona Codreanu were vital in the development of my technical skills and scientific discourse. I would also like to thank Andrzej Balinski, Berkley M. Ellis, and James C. Poland for their support, invigorating discussions, and comradery.

I would like to thank my family and friends for their continued support and encouragement. I could not have done this without them. To my father, Joris, you've been a constant pillar of strength and reason. When I felt tired or weak, you bolstered my confidence and

ensured to me that things would work out. To my mother, Jocelyn, - thank you for your constant love and care You've always believed in me especially when I doubted myself and that has been invaluable on this journey. To my siblings, Jennifer and Justin, thank you for being two of my best friends. You've helped me in countless ways through this journey and I couldn't have done it without you. To Titan, thank you for your constant love and protection. You give me hope like no one else. To Zachary, thank you for your support, patience, and love. You've helped me find strength, hope, and joy. To Katelyn, thank you for being a light and constant source of affirmation when I struggled. You've been a rock for me and I would not be the person I am without you.

I would be remiss if I didn't acknowledge the community I found here in Nashville through the climbing community. Stephanie, Jessie, and Casey – thank you for all of your comradery. Climbing with you has taught me perseverance and self-love. I could not have completed my graduate school journey without your support system.

Lastly, I would like to acknowledge the funding sources for this research: the National Institutes of Health, the Warren Fellowship, the College of Arts and Science, the Center for Innovative Technologies, and the Vanderbilt Chemical Biology Interface training program.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

Mass Spectrometry-based Metabolomics

1.1 Introduction and Instrumentation

The ability to globally characterize the molecular profile of a system has implications in biomedical research fields such as systems biology, drug and natural product discovery, precision medicine, and diagnostics among others.[1–4] Mass spectrometry (MS) has become the central technique in metabolomic endeavors.[5,6] However, MS alone is often not enough to capture the complexity of biological samples. In order to maximize analytical peak capacity and therefore, analyte coverage, orthogonal separations techniques must be employed. A common technique used in conjunction with MS is liquid chromatography (LC–MS) which greatly improves analyte coverage but can be challenging because of retention time variability and co-elution issues.[7] In molecular metabolomics, analyte identifications are assigned confidence levels (5 – low confidence to 1 – high confidence) as depicted in **Figure 1.1**.[8] LC-MS will often result in Level 3 IDs. Ideally, enough information should be collected to reach a Level 2 ID which is often good enough to enable biological inferences from the data. In order to move from a Level 3 ID to a Level 2 ID, an additional analytical dimension is required. Two techniques used for Level 2 IDs are fragmentation (MS/MS) and ion mobility (IM). Given that most analytes in metabolomic studies are small molecule metabolites and exogenous exposures, IM is preferable because small molecules often have similar fragmentation patterns.[9] *I propose that IM-MS, in conjunction with other technologies described herein, can improve the confidence of small molecule identification assignments when compared to traditional LC-MS experiments.*

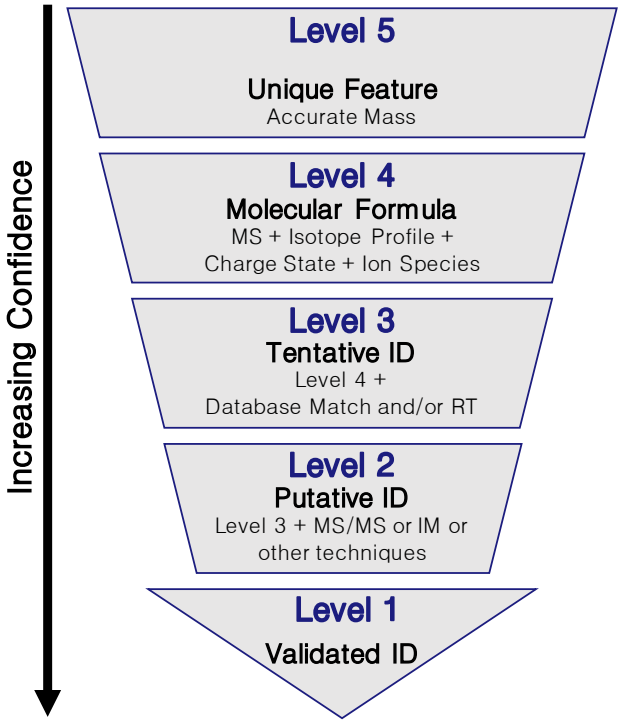


Figure 1.5. Metabolomic analyte identification confidence levels and associated analytical techniques. (Figure adapted from Schrimpe-Rutledge, et al.[8])

Briefly, IM, as depicted in **Figure 1.2**, is a separations technique that is selective to an analyte's gas phase structure.[1,10] More compact, higher charged ions traverse the drift tube faster than their bulkier, lower charged counterparts. This separation increases analytical peak capacity over LC-MS alone as well as measures an analyte's mobility through means of an experimental drift time.[11] This drift time can be used to calculate a collision cross section (CCS), a unique molecular identifier that describes an analyte's averaged rotational surface area.[12] Previous work has indicated that CCS values obtained using a standardized method can be measured within a 0.3% relative standard deviation (RSD).[13] All of the studies within this dissertation were completed using the Agilent 6560, a commercially available drift tube ion mobility–mass spectrometer depicted in **Figure 1.2**. Key challenges in the IM-MS field are a lack of consensus reference standard values as well as complicated informatics. These challenges need to be addressed if IM is going to be used in metabolomic analyte identification.

1.2 Chapter Overviews

In the following chapters, we address these challenges by developing tools that standardize CCS values. Furthermore, we describe strategies in which CCS can be used to aide in metabolomic experiments and data analysis.

1.2.1 Mass Spectrometry Metabolite Library of Standards Specifically, Chapter 2 details development of a collision cross section library of primary metabolites based on analytical standards in the Mass Spectrometry Metabolite Library of Standards (MSMLS) using a commercially available ion mobility-mass spectrometer (IM-MS). From the 554 unique compounds in the MSMLS plate library, we obtained a total of 1246 CCS measurements over a

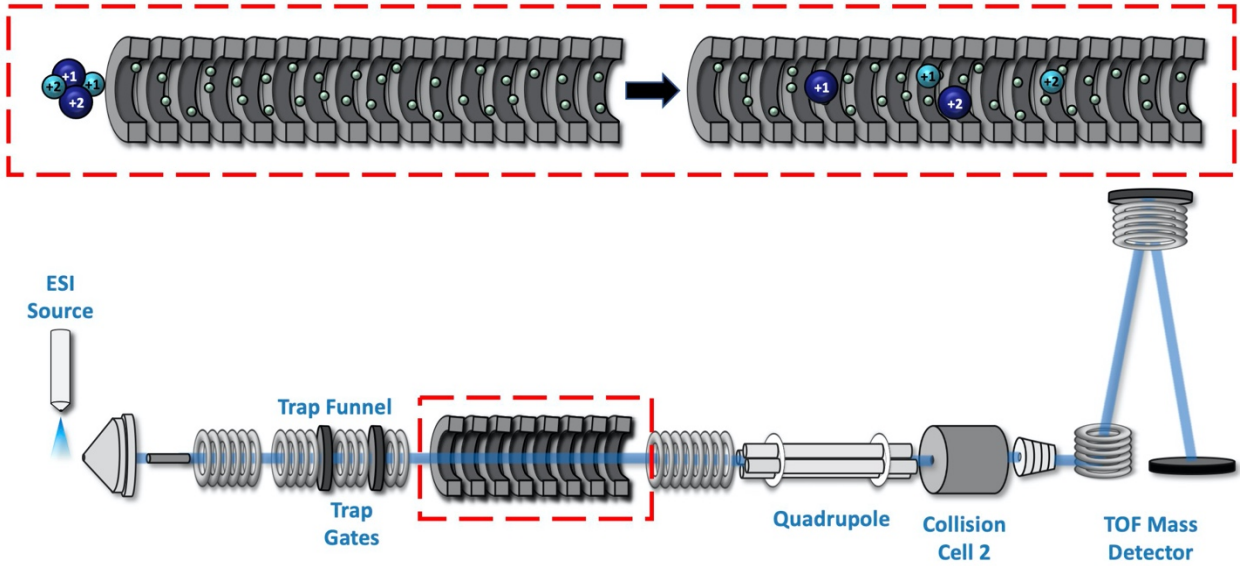


Figure 1.6. Schematic of the Agilent 6560 The Agilent 6560 is a commercially available drift tube ion mobility – mass spectrometer. The callout box indicates how more compact, higher charge ions traverse the drift tube region faster than less compact, lower charged ions.

wide range of biochemical classes and adduct types. Resulting data analysis demonstrated that the curated CCS library provides broad molecular coverage of metabolic pathways and highlights intrinsic mass−mobility relationships for specific metabolite super classes. The separation and characterization of isomeric metabolites were assessed, and all molecular species contained within the plate library, including isomers, were critically evaluated to determine the analytical separation efficiency in both the mass (*m/z*) and mobility (CCS/ΔCCS) dimension required for untargeted metabolomic analyses. To further demonstrate the analytical utility of CCS as an additional molecular descriptor, a well-characterized biological sample of human plasma serum (NIST SRM 1950) was examined by LC-IM-MS and used to provide a detailed isomeric analysis of carbohydrate constituents by ion mobility.

1.2.2 The Unified Collision Cross Section Compendium Chapter 3 details a large, Unified CCS compendium of >3800 experimentally acquired CCS values obtained from traceable molecular standards and measured with drift tube ion mobility-mass spectrometers. An interactive visualization of this compendium along with data analytic tools have been made openly accessible. Represented in the compendium are 14 structurally-based chemical super classes, consisting of a total of 80 classes and 157 subclasses. Using this large data set, regression fitting and predictive statistics have been performed to describe mass-CCS correlations specific to each chemical ontology. These structural trends provide a rapid and effective filtering method in the traditional untargeted workflow for identification of unknown biochemical species. The utility of the approach is illustrated by an application to metabolites in human serum, quantified trends of which were used to assess the probability of an unknown compound belonging to a given class. CCS-based filtering narrowed the chemical search space by 60% while increasing the confidence in the

remaining isomeric identifications from a single class, thus demonstrating the value of integrating predictive analyses into untargeted experiments to assist in identification workflows. The predictive abilities of this compendium will improve in specificity and expand to more chemical classes as additional data from the IM-MS community is contributed. Instructions for data submission to the compendium and criteria for inclusion are provided.

1.2.3 Supervised Inference for Feature Taxonomy from Ensemble Randomizations Chapter 4 details a machine learning algorithm referred to as the supervised inference of feature taxonomy from ensemble randomization (SIFTER), which supports the identification of features derived from untargeted ion mobility-mass spectrometry experiments. SIFTER utilizes random forest machine learning on three analytical measurements derived from IM-MS (collision cross section), mass-to-charge, and mass defect (Δm) to classify unknown features into a taxonomy of chemical kingdom, super class, class, and subclass. Each of these classifications is assigned a calculated probability as well as alternate classifications with associated probabilities. After optimization, SIFTER was tested against a set of molecules not used in the training set. The average success rate in classifying all four taxonomy categories correctly was found to be >99%. Analysis of molecular features detected from a complex biological matrix and not used in the training set yielded a lower success rate where all four categories were correctly predicted for ~80% of the compounds. This decline in performance is in part due to incompleteness of the training set across all potential taxonomic categories, but also resulting from a nearest-neighbor bias in the random forest algorithm. Ongoing efforts are focused on improving the class prediction accuracy of SIFTER through expansion of empirical data sets used for training as well as improvements to the core algorithm.

1.2.4 Crowd-sourced Chemistry Chapter five discusses the power of crowd-sourced chemistry. Mass spectrometry is used in multiple omics disciplines to generate large collections of data. This data enables advancements in biomedical research by providing global profiles of a given system. One of the main barriers to generating these profiles is the inability to accurately annotate omics data, especially small molecules. To complement pre-existing large databases that are not quite complete, research groups devote efforts to generating personal libraries to annotate their data. Scientific progress is impeded during the generation of these personal libraries because the data contained within them is often redundant and/or incompatible with other databases. To overcome these redundancies and incompatibilities, we proposed that communal, crowd-sourced databases be curated in a standardized fashion. A small number of groups have shown this model is feasible and successful. While the needs of a specific field will dictate the functionality of a communal database, we discuss some features to consider during database development. Special emphasis is made on standardization of terminology, documentation, format, reference materials, and quality assurance practices. These standardization procedures enable a field to have higher confidence in the quality of the data within a given database. We also discuss the three conceptual pillars of database design as well as how crowd-sourcing is practiced. Generating open-source databases requires front-end effort, but the result is a well curated, high quality data set that all can use. Having a resource such as this fosters collaboration and scientific advancement.

1.3 Acknowledgments

This dissertation chapter was adapted from the abstracts of several published manuscripts including “Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry” by Charles M. Nichols, James N

Dodds, Bailey S. Rose, Jaqueline A. Picache, Caleb B. Morris, Simona G. Codreanu, Jody C. May, Stacy D. Sherrod, and John A. McLean published in *Analytical Chemistry* **2018**, *90* (24), 14484-14492; "Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities" by Jaqueline A. Picache, Bailey S. Rose, Andrzej Balinski, Katrina L. Leaptrot, Stacy D. Sherrod, Jody C. May, and John A. McLean published in *Chemical Science* **2019**, *10* (4), 983–993; "Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization" by Jaqueline A. Picache, Jody C. May, and John A. McLean published in *Analytical Chemistry* **2020**, Just Accepted, DOI: https://dx.doi.org/10.1021/acs.analchem.0c02137.; and "Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analysis" by Jaqueline A. Picache, Jody C. May, and John A. McLean published in *ACS Omega* **2020**, *5* (2), 980-985.

I would like to acknowledge the McLean research group for their support. Financial support for this research was provided by the National Institutes of Health (NIH NIGMS R01GM092218 and NIH NCI 1R03CA222452-01) and the NIH supported Vanderbilt Chemical Biology Interface training program (5T32GM065086-16). This work was supported in part using the resources of the Center for Innovative Technology (CIT) at Vanderbilt University.

1.4 References

(1) J. C. May and J. A. McLean, *Annu. Rev. Anal. Chem*, **2016**, 9, 387–409.

(2) S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nat. Rev. Drug Discov.*, **2010**, 9, 203–14.

(3) R. A. Quinn, J. A. Navas-molina, E. R. Hyde, J. Song, Y. Vázquez-baeza, G. Humphrey, J. Gaffney, J. J. Minich, A. V Melnik, J. Herschend, J. Dereus, A. Durant, R. J. Dutton, M. Khosroheidari and C. Green, *mSystemsmSystems*, **2016**, 1, e00038-16.

(4) R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. J. Butte, E. A. Ashley, M. Gerstein, K. C. Nadeau, H. Tang and M. Snyder, *Cell*, **2012**, 148, 1293–1307.

(5) D. Houle, D. R. Govindaraju and S. Omholt, *Nat. Rev. Genet.*, **2010**, 11, 855–66.

(6) J. C. May, R. L. Gant-Branum and J. A. McLean, *Curr. Opin. Biotechnol.*, **2016**, 39**,** 192–197.

(7) J. S. D. Zimmer, M. E. Monroe, W. J. Qian and R. D. Smith, *Mass Spectrom. Rev.*, **2006**, 25, 450–482.

(8) A. C. Schrimpe-Rutledge, S. G. Codreanu, S. D. Sherrod and J. A. Mclean, *J. Am. Soc. Mass Spectrom*, , DOI:10.1007/s13361-016-1469-y.

(9) T. Kind and O. Fiehn, *Bioanal. Rev.*, **2010**, 2**,** 23–60.

(10) X. Zheng, R. Wojcik, X. Zhang, Y. M. Ibrahim, K. E. Burnum-Johnson, D. J. Orton, M. E. Monroe, R. J. Moore, R. D. Smith and E. S. Baker, *Annu. Rev. Anal. Chem.*, **2017**, 10, 71–92.

(11) J. A. McLean, B. T. Ruotolo, K. J. Gillig and D. H. Russell, *Int. J. Mass Spectrom.*, **2005**, 240, 301–315.

(12) E. A. Mason and E. W. McDaniel, *Transport Properties of Ions in Gases*, John Wiley & Sons, Ltd., New York City, NY, **1988**.

(13) S. M. Stow, T. J. Causon, X. Zheng, R. T. Kurulugama, T. Mairinger, J. C. May, E. E. Rennie, E. S. Baker, R. D. Smith, J. A. McLean, S. Hann and J. C. Fjeldsted, *Anal. Chem.*, **2017**, 89, 9048–9055.

CHAPTER II

Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry

2.1 Introduction

From the central dogma of molecular biology, studies of genomics, transcriptomics, and proteomics provide higher order information about gene and protein expression to better understand implicated phenotypes.[1,2] However, these approaches provide limited information about real-time production of chemical species related to cellular metabolism as a function of external stimuli or phenotype of interest. To address the need for rapid characterization of cellular metabolism, metabolomics seeks to uncover molecular information on a per-molecule basis by examining expressed cellular products that can be correlated with a specific phenotype, stimuli, or other experimental conditions.[3]

While several analytical approaches have been utilized to study metabolism and related cellular processes (*e.g.* NMR, electrochemistry, etc.),[4,5] mass spectrometry (MS) is gaining wide-spread adoption as a result of its high throughput, low limits of detection, and molecular specificity. Mass spectrometers can collect chemical information on the microsecond (μs) time scale,[6] and with the rise of high-resolution, accurate mass techniques such as time of flight (TOF), Orbitrap, and ion cyclotron instruments, a unique chemical formula can often be generated based solely on mass measurement for a specific analyte signal.[7,8] While identifying a specific chemical formula is advantageous, many metabolic pathways include isomeric molecules covering a range of biological classes, such as carbohydrates (*e.g.* glucose/galactose),[9] nucleosides (*e.g.*

adenosine/deoxyguanosine), and lipids (7-dehydrocholesterol/desmosterol).[10] As biological function follows molecular structure, characterization of isomeric species is imperative for complete molecular identification and accurate pathway analysis. In many MS experiments, fragmentation techniques such as collision induced dissociation (CID) or electron transfer dissociation (ETD) are utilized to provide structural information about a specific analyte measured in the study.[11,12] However, as many metabolite isomers are less than 300 Dalton, these compounds often possess identical fragmentation spectra at similar energy thresholds and hence molecular fingerprinting by MS/MS and high resolution precursor mass is often not specific enough to identify a unique molecular structure.[13] Furthermore, as quadrupoles isolate on nominal mass, molecules with different molecular formulas but similar exact mass (*i.e.* nominal mass isobars) cannot be isolated, thereby complicating MS/MS analysis.[14] To address these challenges, pre-separation techniques such as gas and liquid chromatography,[15,16] and more recently ion mobility spectrometry,[17] have been interfaced prior to mass analysis to provide enhanced structural recognition and increased analyte coverage. For untargeted analysis, metabolomic databases (*e.g.* METLIN, HMDB, etc)[18] include multiple descriptors of analyte information (*e.g.* accurate mass, ion adduct form, fragmentation pattern, and retention time) to increase confidence in molecular identification.[19] With the advent of commercially-available ion mobility-mass spectrometers in 2006,[20] collision cross section (CCS) has become an additional molecular descriptor for untargeted experiments. CCS measurements are being standardized across instrumental platforms using rigid experimental protocols, and as such provide a molecular descriptor independent of system settings which are transferable between laboratories.[17,21–23] These collected CCS measurements provide the capability to distinguish isomeric species in complex mixtures, provided enough resolution is accessible in the IM dimension.[24] In order to provide additional confidence in molecular

identification for untargeted metabolomic analysis, significant efforts are being made in the IM community to establish reliable CCS databases for analyzing unknown features across a range of biochemical classes, including lipids, metabolites, and xenobitics.[21,22,25,26] In this work, we use uniform field IM-MS to develop a library of CCS values focused on primary metabolites established with analytical standards to facilitate chemical identification in untargeted metabolomic workflows. Furthermore, we demonstrate the utility of these measurements by analyzing a commercially available extract of human serum (NIST 1950) which has been characterized previously in traditional GC and LC-MS experiments.[27,28]

2.2 Experimental Methods

2.2.1 MSMLS Sample Preparation. The Mass Spectrometry Metabolite Library of Standards (MSMLS, IROA technologies) is supplied as dried standards distributed across seven 96-well plates (Sigma-Aldrich; St. Louis, MO) and each well contains 5 µg of analytical standard. All solvents used to reconstitute the analytes prior to analysis, including water (H_2O), methanol (MeOH), acetonitrile (ACN), isopropanol (IPA), and chloroform (CHCl3) were Optima LC-MS grade purchased from Fisher Scientific (Fair Lawn, NJ). Stock solutions of the hydrophilic standards were prepared by adding 100 µL 1:9 (MeOH: H_2O) to each well prior to mixing on a waving rotator for 5 minutes. The stocks were then distributed in 20 µL aliquots throughout five 96-well plates (Waters part no. 186005837). Stock plates that were not immediately analyzed were capped and transferred to -80 °C for storage. Working solutions of the hydrophilic standards were prepared by adding 80 µL of water with 0.1% formic acid to the 20 µL stock solutions, sealed with plate covers (Waters part no. 186006332), and subsequently mixed on a waving rotator for 5 minutes. The hydrophobic analyte set was prepared similarly, where stock solutions were prepared

with 100 µL 2:1:1:0.3 (MeOH: CHCl3: IPA: H_2O), and distributed in 20 µL aliquots throughout five 96-well plates. Working solutions were prepared by adding 80 µL of 1:1 (MeOH: IPA). The concentration of the working solutions used for IM-MS analysis was 10 µg/mL.

2.2.2 Collision Cross Section Measurements. CCS measurements for the MSMLS were obtained on a commercially available drift tube ion mobility-mass spectrometer (DTIMS, Agilent 6560) operated with nitrogen gas (3.95 Torr) at room temperature (~25 °C) and using both single-field and stepped-field approaches previously established in an international inter-laboratory study.[21] The single-field CCS values reported here were measured in triplicate, while the stepped-field values were collected in a single acquisition. Stepped-field measurements were acquired using an automated flow injection analysis (FIA) stepped-field approach described previously.[29] Briefly, the FIA method was performed with a liquid chromatography system (Agilent 1290) modified with a 100 µL sample loop (Agilent part no. G4226-87303) coupled to an IM-MS (6560, Agilent). 20 µL of the working solution was injected from the 96-well plate with 1:1 (water: isopropanol) as the carrier solvent. For traditional stepped-field CCS determination by FIA, following a 0.5 s delay, an entrance potential was stepped every 0.5 min. in increments of 100 V from 1074 V to 1674 V; the first step from 1074 to 1174 occurred at 1.0 minute rather than 0.5 min. For single-field CCS determination using FIA, 4 µL of sample was injected into the carrier solvent at a flow of 800 µL/min. Data was collected for 0.5 min, followed by a 0.4 min postrun flushing cycle. A drift tube entrance voltage of 1574 V was used. DTIMS exhibits a linear relationship between drift time and CCS,[6] and single-field CCS values are determined by first measuring the drift time of ions (ESI Low Concentration Tuning Mix, Agilent) with a known CCS. The calibrant ions were infused for 0.5 minutes while IM-MS spectra are collected; calibration experiments were preformed

intermittently to ensure instrument stability. IM-MS Browser (Agilent, B.08) was used to plot the linear regression of the calibration ions for single field experiments, and the instrumental coefficients β and T_{fix}, were extracted and used to convert raw ion drift times to CCS.[21] The resulting single- and stepped-field CCS library can be found in the Supporting Information.

2.2.3 IM-MS Source and Drift Cell Conditions. To obtain high coverage of analytes within the MSMLS, both electrospray (Agilent Jet Stream, AJS) and chemical ionization (APCI) sources were used. The majority of the samples collected with the AJS in both ion modes were measured using the following conditions: gas temperature, 250 °C; drying gas, 8 L/min; nebulizer, 60 psig; sheath gas temperature, 300 °C; sheath gas flow, 11 L/min; capillary voltage (VCap), 3500 V; nozzle voltage, 800 V; fragmentor, 340 V; octopole 1 RF Vpp, 750 V. All metabolites were first investigated using the AJS source; those which were not observed in either ion polarity were subsequently investigated using the APCI source under the following conditions: gas temperature, 250 °C; vaporizer, 200 °C; drying gas, 7 L/min; nebulizer, 30 psig; VCap, 3800 V; corona, 5 μA; fragmentor, 350 V; octopole 1 RF Vpp 750 V. Some of the low m/z ions (typically ≤ 200 Da) exhibited metastable ion dissociation in the DTIMS which resulted in uncorrelated mobilities (**Supplemental Figure S2.1**). In these cases, we increased the fragmentor potential to > 350 V and decreased the Trap Funnel RF to ≤ 80 V_{pp} to culminate the ion signal into a single IM distribution. The IM-MS settings for the CCS values reported herein are as follows: 0.9 frames/s; 18 IM transients/frame; 60 ms max drift time; 600 TOF transients/IM transient; 20000 μs trap fill time; 180 μs trap release time; drift tube exit voltage, 224 V; rear funnel entrance voltage, 217.5 V; rear funnel exit voltage, 45 V.

2.2.4 Nonlinear Regression Analysis. Iterative nonlinear regression modeling for the super classes was performed using GraphPad Prism 7, and 99% confidence intervals were generated for each biomolecular super class. Three fits were tested for each super class: power fit (PF), 4-parameter sigmoidal (4P), and 5-parameter sigmoidal (5P). The most parsimonious fit was chosen by a probabilistic comparison of the corrected Akaike information criterion (AICc) values.

2.2.5 Human Serum Preparation. Protein precipitation was performed by adding 800 µL of ice cold MeOH to 100µL NIST 1950 serum and stored at -80 ˚C for one hour. The sample was centrifuged at 14,000 rpm for 5 minutes before collecting the supernatant. Next, 2.4 mL ice cold methyl tert-butyl ether and 800 µL ice cold water were added. The sample was vortexed then centrifuged at 10,000 rpm at 4˚C for 10 minutes. The polar and nonpolar fractions were separated and dried separately *in vacuo.* Samples were stored at -20 ˚C until analysis. Dried fractions were resuspended in 200 µL of the initial mobile phase solvent and analyzed via LC-IM-MS.

2.2.6 Liquid Chromatography. LC-IM-MS was performed on the prepared NIST 1950 serum using HILIC chromatography for the hydrophilic layer of the liquid-liquid extraction. For this method, 4 µL of sample was injected onto a column heated to 40 °C. The Millipore SeQuant Zic-HILIC (2.1 x 100 mm, 3.5 µm) column was used with mobile phase A and B being 9:1 and 1:9 (water: acetonitrile, buffered with 5 mM ammonium formate), respectively. The mobile phase flow rate was 200 µL/min. The gradient was initially held at 98 %B from 0 to 1 minutes, decreased to 45 %B from 1 to 20 minutes, held at 45 % B from 20 to 22 minutes, increased to 98 %B from 22 to 40 minutes, and subsequently held at 98 %B from 40 to 45 minutes before the next injection.

2.3 Results and Discussion

2.3.1 MSMLS Plate Coverage In total, the MSMLS plates analyzed in this work contained 554 unique compounds across a large breadth of biological classes found in canonical metabolite pathways (See **Figure 2.1A**). Of these 554 analytes, one or more CCS values were measured for 417, resulting in *ca.* 75% coverage. Of the remaining *ca.* 25% that did not result in an acceptable CCS measurement, approximately half of these did not yield appreciable signals, presumably because of difficulties in ionization under the conditions used here, or poor ion transmission efficiency due to the lower mass of many of these analytes and the limits of RF frequency used in ion transfer optics. For the remaining species, there is in some cases, evidence for metastable dissociation in the drift tube, resulting in uncorrelated mobility and poor peak shape (See **Appendix B, Figure B2.1**). For purposes of this manuscript, we have chosen to report only values for signals demonstrating high signal intensity and reproducibility. Collectively, these 417 analytes produced 1246 CCS measurements using both positive (701 measurements) and negative ion polarities (545 measurements) across several adduct types (*e.g.* $[M+Na]^+$, $[M-H]^-$, *etc.*, see **Appendix B, Figure B2.2**). Analyte identification and relevant descriptors (chemical name, formula, KEGG ID, Metlin ID, adduct type, measured mass, CCS, and other information) have been uploaded to Metabolomics Workbench,[30] and are provided in the Supporting information as two Personal Compound Database Libraries (PCDL Manager, Agilent), one corresponding to single-field measurements and the other to stepped-field CCS measurements.

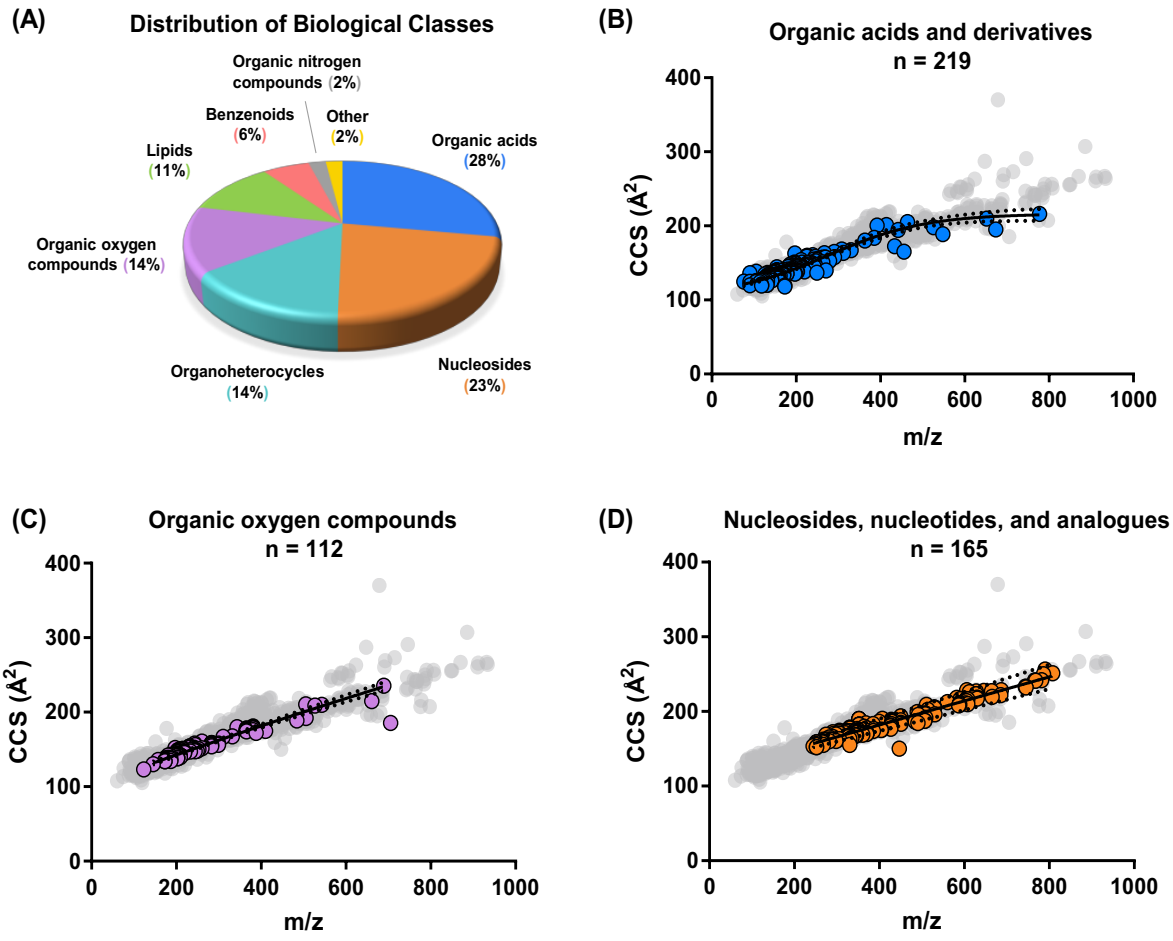


Figure 2.1. MSMLS Distribution and Conformation Plots (A) Distribution of biological categories associated with the primary metabolites examined in the MSMLS plate library. **(B, C,** and **D)** Conformational space plots of three singly charged molecular super classes contained in the MSMLS library. Representative nonlinear regression fits (solid black lines) along with 99 % confidence intervals (black dotted lines) are shown for each. Gray dots denote all molecules CCS values obtained in the library. All CCS error bars are smaller than their respective symbols. **(B)** "Organic acids and derivatives" with a 4-parameter sigmoidal fit. **(C)** "Organic oxygen compounds" with a power fit and **(D)** "Nucleosides, nucleotides, and analogues" with a power fit. (Figure adapted from Nichols, et al.[51])

2.3.2 Mass–Mobility Correlation Analysis In these data, we observed several distinct relationships between m/z and CCS for individual structural super classes represented in the MSMLS library similar to previous IM-MS literature.[31–35] Mass/mobility relationships have been shown to have utility as an additional rapid identifier of biomolecular class for uncharacterized biological samples,[36] making the mathematical description of these relationships by nonlinear regression modeling particularly useful. Unlike the canonical biochemical classes (nucleotides, proteins, carbohydrates, and lipids), metabolites exhibit less distinguished structural differences between chemical classes, and so several mathematical fits were investigated in order to find mass/mobility correlations which exhibit high class specificity. Fits and confidence intervals for representative super classes are shown in **Figure 2.1 (B, C,** and **D)**, and detailed mathematical expressions are provided in the Supporting information (**Appendix B, Equations B2.1-6**).

2.3.3 Metabolic Pathway Coverage As the MSMLS was designed to provide analytical standards of primary metabolism, we also evaluated metabolite coverage using pathway analysis by inputting KEGG IDs for all of the analytes measured in our CCS database in MetaboAnalyst 4.0.[37] In total, 64 pathways were covered with a wide range of biological activity including key metabolic processes such as the citric acid cycle, amino acid metabolism, and glycolysis (**Figure 2.2A** and **Appendix B, Table B2.1**). Pathway coverage presented in this work is solely based on analyte coverage from the standards, and therefore provides qualitative reflection on the number of analytes in each specific pathway which are accounted for in the CCS library. MetaboAnalyst 4.0 also provides detailed information for specific pathways of interest, wherein molecular coverage can be evaluated on a per-analyte basis. For example, 10 pivotal metabolites in the citric acid cycle

(see **Figure 2.2B**) are represented within the standards, and out of 20 total, 8 of these molecules exhibited a measurable CCS (green), while only 2 (orange) were observable in the mass spectrum, but did not result in a collected CCS, due to low ion intensity. Of note, many other compounds described in the KEGG pathways which are not components in the standard set (10 compounds, light blue) are protein enzymes or oxidized derivatives, and only 3 of these 10 are available for purchase as analytical standards. Hence it is unlikely that 100 % coverage of canonical pathways is obtainable with chemical standards. As an analogy, it is not necessary to have 100 % peptide coverage for a specific protein in proteomic analysis for confident identification.

2.3.4 Isomers in Metabolomics Of the more than 500 compounds in the MSMLS library, almost one-third (31%) have a chemical formula in common with another compound, forming an isomeric pair, **Appendix B, Figure B2.3**. Isomeric compounds are ubiquitous in metabolomic processes across a wide range of biological classes, for example the carbohydrate rearrangements for glucose 6-phosphate isomerization to fructose 6-phosphate in glycolysis. **Figure 2.2B** highlights two key metabolic intermediates of the citric acid cycle, citrate and isocitrate, which are constitutional rearrangements of a single hydroxyl group along the central carbon chain. As these compounds have the same chemical formula, they will also possess identical masses, requiring additional separation in the chromatographic dimension for increased identification confidence in pathway analysis.[17,38] In the example depicted in Fig. 2.2B, ion mobility allows for differentiation of these two isomeric metabolites (CCS = 143.1 $Å^2$ vs. CCS = 142.7 $Å^2$, for citrate and isocitrate, respectively), which are indistinguishable by mass alone. Adding the ion mobility dimension to analysis.[17,38] In the example depicted in Fig. 2.2B, ion mobility allows for differentiation of these two isomeric metabolites (CCS = 143.1 $Å^2$ vs. CCS = 142.7 $Å^2$, for citrate and isocitrate,

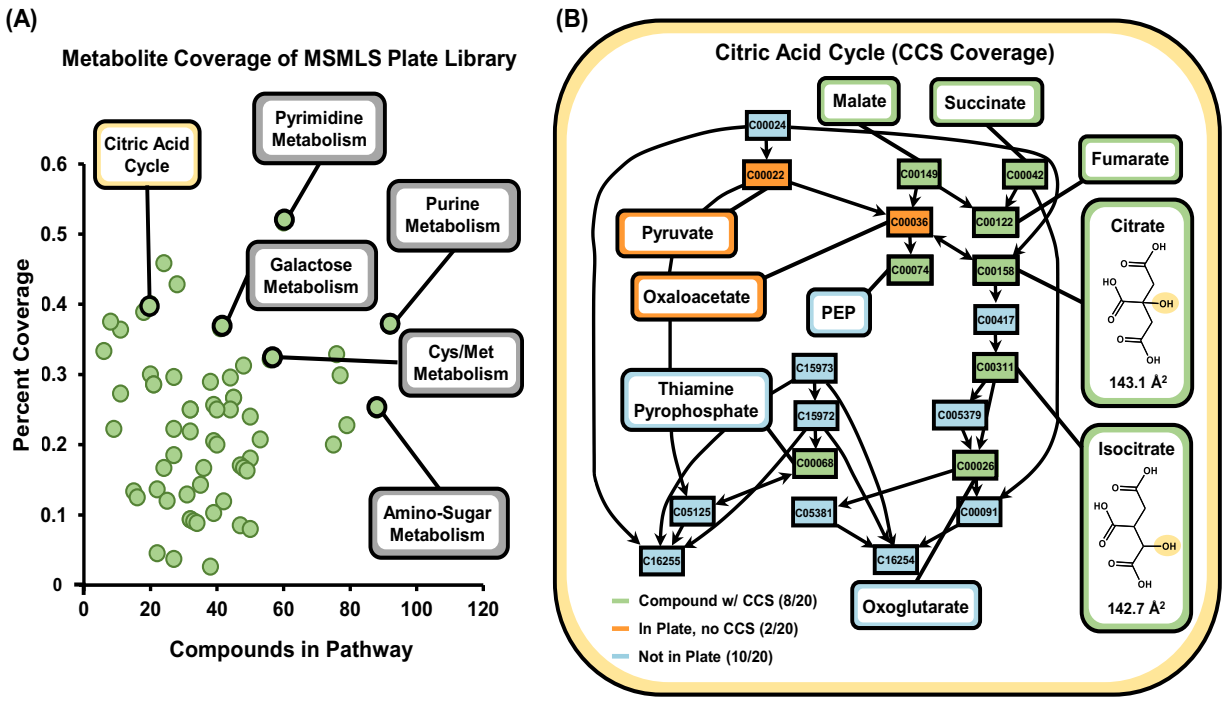


Figure 2.2. Metabolite Pathway Analysis (A) KEGG pathway coverage of metabolites with measured CCS evaluated in this study. A total of 64 pathways are covered by metabolites in our CCS library based on the MSMLS. After a specific pathway is selected **(B)**, metabolite-specific coverage can be evaluated. In many pathways, isomerization is a key intermediate in primary metabolism, noted by the callouts for citrate and isocitrate in the citric acid cycle. (Figure adapted from Nichols, et al.[51])

respectively), which are indistinguishable by mass alone. Adding the ion mobility dimension to existing untargeted workflows allows for additional separation and characterization of isomeric metabolites that interfaces within the timescale of traditional chromatographic techniques.[6]

As a specific example, adenosine 5-diphosphate, adenosine 3,5-diphosphate, and 2'-deoxyguanosine-5'-diphosphate are nucleoside isomers which are key metabolites in purine metabolism and are depicted in **Figure. 2.3A**. Note that the only structural difference between A-5-DP (blue) and A-3,5-DP (green) is the location of a phosphate group from the central ribose unit. These two isomers are in turn differentiated structurally from 2'-deoxyguanosine-5'-diphosphate (DGDP, orange) by molecular substitutions on the purine ring, where a hydroxyl group has been relocated from the ribose sugar to the guanine ring, as well as an amine rearrangement in the same region. Structurally, these three nucleoside compounds are also constitutional isomers, a subcategory of isomeric compounds which have been heavily characterized in previous ion mobility literature.[39–41] Also noted in **Figure 2.3A**, adduct formation has a substantial effect on the overall selectivity of the IM separation. Specifically, each nucleoside isomer has a distinct cross sectional distribution which are distinguishable in the protonated $[M+H]^+$ and sodium adducted $[M+Na]^+$ species, however the rearrangement of the phosphate group between A-5-DP and A-3,5-DP provides no resolution for the deprotonated form observed in negative ion mode $[M-H]^-$. Other metabolite separations in this study were more readily separated in negative ion mode such as the isomers L-glutamic acid and N-methyl-D-aspartic acid (see **Appendix B, Figure B2.4)**.

The broad range of chemical diversity present in small molecules presents unique advantages in the range of ion types that can be utilized. Collectively, these results demonstrate the advantage of utilizing both ion polarities in untargeted analysis wherein various charge adducts formed during

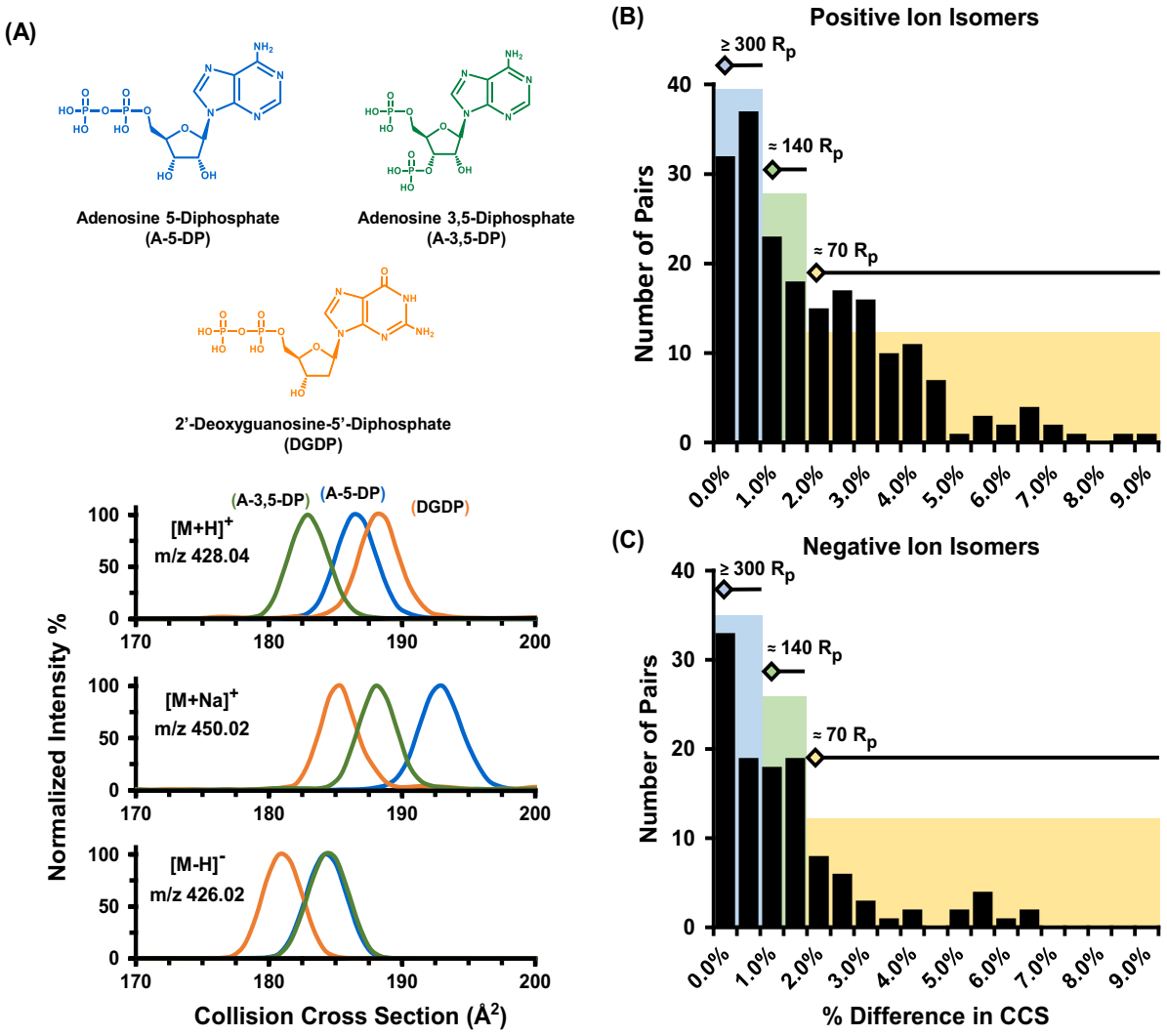


Figure 2.3. Isomer Analysis (A) IM separation of nucleoside isomers (chemical structures illustrated at the top of the panel) for $[M+H]^+$, $[M+Na]^+$, and $[M-H]^-$ ion forms, respectively. For these particular isomers, enhanced separation is noted for the sodium adducts, while the negative mode A-5-DP $[M-H]^-$ and A-3,5-DP CCS distributions are indistinguishable. After sorting all observed isomer sets in the MSMLS dataset, pairwise matches were created and evaluated based on their percent difference in CCS. The resulting difficulty in separations is noted in panels **(B)** for positive and **(C)** negative ion forms. (Figure adapted from Nichols, et al.[51])

the ionization process can be exploited in order to substantially enhance the selectivity in IM-MS analysis, by increasing the absolute CCS difference between isomers. This allows a significant improvement in separation without instrumental upgrades that would otherwise be necessary to achieve improved separation *via* increased resolving power. This enhanced separation, in turn, provides additional confidence in identification through CCS library matches and enhanced ion mobility resolution. A potential future direction in the field will utilize molecular modeling and machine learning approaches for prediction of adduct specific CCS values and optimal separation conditions.

2.3.5 IM-MS Separation in Primary Metabolites In addition to enhanced separation through charged, adduct formation, recent advances in ion mobility resolving power have provided increased separation coverage of isomeric species.[42,43] In order to determine the resolving power in the IM dimension needed for untargeted metabolomic experiments, we analyzed pairwise matches of all isomers which provided a usable CCS and binned the resulting pairs by percent difference in cross section (%ΔCCS). In brief, analytes with identical chemical formulas were grouped into isomeric sets and were subsequently matched in a pairwise comparison. Each pairwise match was generated using an enumeration strategy wherein a percent difference in CCS was calculated for each possible combination of isomers. Most isomeric sets consist of 2-3 compounds, whereas the largest isomeric set was comprised of 9 unique analytes (see **Appendix B, Figure B2.3**). In one example, there are 5 sugar compounds which share the same chemical formula ($C_6H_{13}O_9PNa^+$, exact *m/z* 283.0195), which results in a total of 10 pairwise isomer matches in this analysis. The percent difference in CCS for all isomer matches were calculated, and the compiled results for all isomer pairings are displayed in **Figure 2.3B** (positive ion mode)

and **Figure 2.3C** (negative ion mode). Approximately half of the isomer pairs generated are ≥ 2.0% different in CCS and require *ca.* 70 resolving power (CCS/ΔCCS) to separate at half height.[43,44] In order to separate additional isomers, more resolving power would be required (*ca.* 140 for ~1.0-2.0% difference in CCS, and *ca.* >300 for ~0.5% CCS difference). Currently, only two commercially available IM-MS platforms provide this level of resolving power (*i.e.* atmospheric pressure DTIMS and trapped IMS),[39,44] although several research instrument prototypes have been developed which are capable of accessing resolving powers in excess of 300 (CCS/ΔCCS).[6,45]

While IM instruments are continually increasing in resolving power capabilities, current untargeted metabolomic workflows identification is based first on primary mass measurement and subsequently supported with retention time, isotope ratios, and fragmentation matching. From this viewpoint, it is also imperative to describe how much resolving power in the mass dimension is necessary for metabolomic studies. By sorting the entire MSMLS library based on primary mass alone, our analysis shows that most analytes (64%) are resolvable based only on the mass dimension utilizing 40,000 mass resolving power (*e.g.* high resolution TOF, see **Figure 2.4**). Increasing levels of mass resolving power (300,000 for Orbitrap and up to 40,000,000 for FT-ICR, respectively)[46,47] provides minimal increases in resolution of these metabolites (*ca.* 3% more). As *ca.* 30% of the compounds in the MSMLS library are isomers, essentially no level of increased mass spectrometry efficiency (short of exited state isomer resolution with MS resolving power of *ca.* 10 billion as theorized by Marshall and coworkers.[48]) will be able to resolve these compounds, and hence orthogonal separation techniques are still required (*i.e.* GC, LC, or IM). Modest resolving power for commercially available IM instrumentation (*ca.* 70 CCS/ΔCCS) resolves an additional 10% of compounds in the library, which outweighs the benefits of additional mass

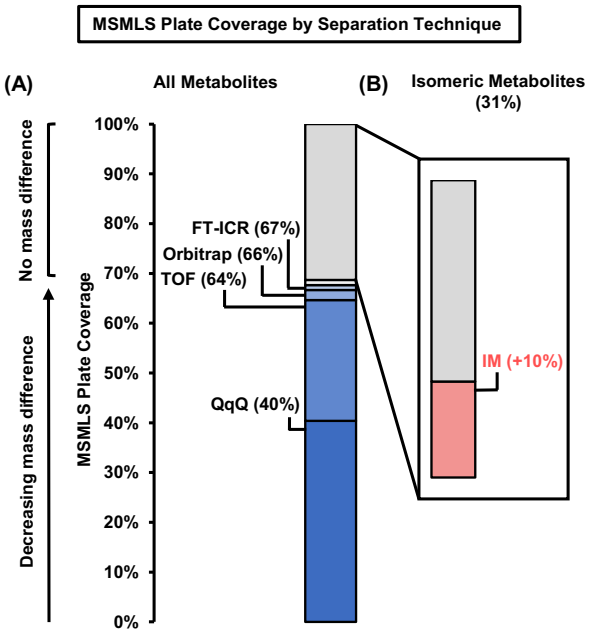


Figure 2.4. MSMLS plate coverage using different separation strategies. (A) Many analytes contained in the library can be resolved in the mass dimension at modest resolving power (TOF Rp = 40,000), with only incremental increases in coverage resulting from the use of an instrument with significantly higher resolving power (FT-ICR Rp = 40,000,000). **(B)** The addition of IM prior to mass analysis allows for isomeric separation and thus increases plate coverage by 10%. (Figure adapted from Nichols, et al.[51])

resolving power beyond 40,000 (*e.g.* TOF MS). We note, however, that this analysis does not consider mass measurement accuracy, which is typically higher for FTMS instruments (Orbitrap and FT-ICR). Nevertheless, in order to obtain the widest scope of molecular coverage in untargeted workflows, possessing sequential separation dimensions based on chemical affinity, gas-phase area, and m/z (LC-IM-MS) would strengthen analyte identification strategies.

2.3.6 LC-IM-MS Characterization of NIST 1950 Serum The NIST 1950 human serum standard has been previously characterized in the literature,[27,28,49] and is analyzed in this work to underscore the importance of isomeric characterization in untargeted experiments. Separation and characterization of isomeric species in biological extracts often requires multiple steps of chemical separation in order to gain increased confidence in assigning molecular structure. For example, the base peak chromatogram in **Figure 2.5A** shows the molecular complexity of the NIST 1950 human serum and the extracted ion chromatogram (lower trace) details a specific molecular feature at *m/z* 203.0528 that elutes into an unresolved broad peak over a *ca.* 2 minute chromatographic window. This broad distribution in the elution profile indicates the potential presence of multiple isomeric forms with similar, yet not identical, retention times. Although TOF MS has high resolving power (*ca.* 40,000), potentially two chemical formula are within 10 ppm of the measured m/z ($C_6H_{12}O_6Na$ and $C_7H_8N_4O_2Na$, at 1.7 ppm and 8.3 ppm respectively; see **Figure 2.5C**). While assignment of this feature to chemical formula $C_6H_{12}O_6Na$ is more probable due to lower observed mass error, isotope ratios were used to confirm this molecular formula assignment, wherein the relative abundance of the M+1 peak in the serum more closely aligns with the isotope model for $C_6H_{12}O_6Na$ as opposed to $C_7H_8N_4O_2Na$ (**Figure 2.5B**). However, even after a specific molecular

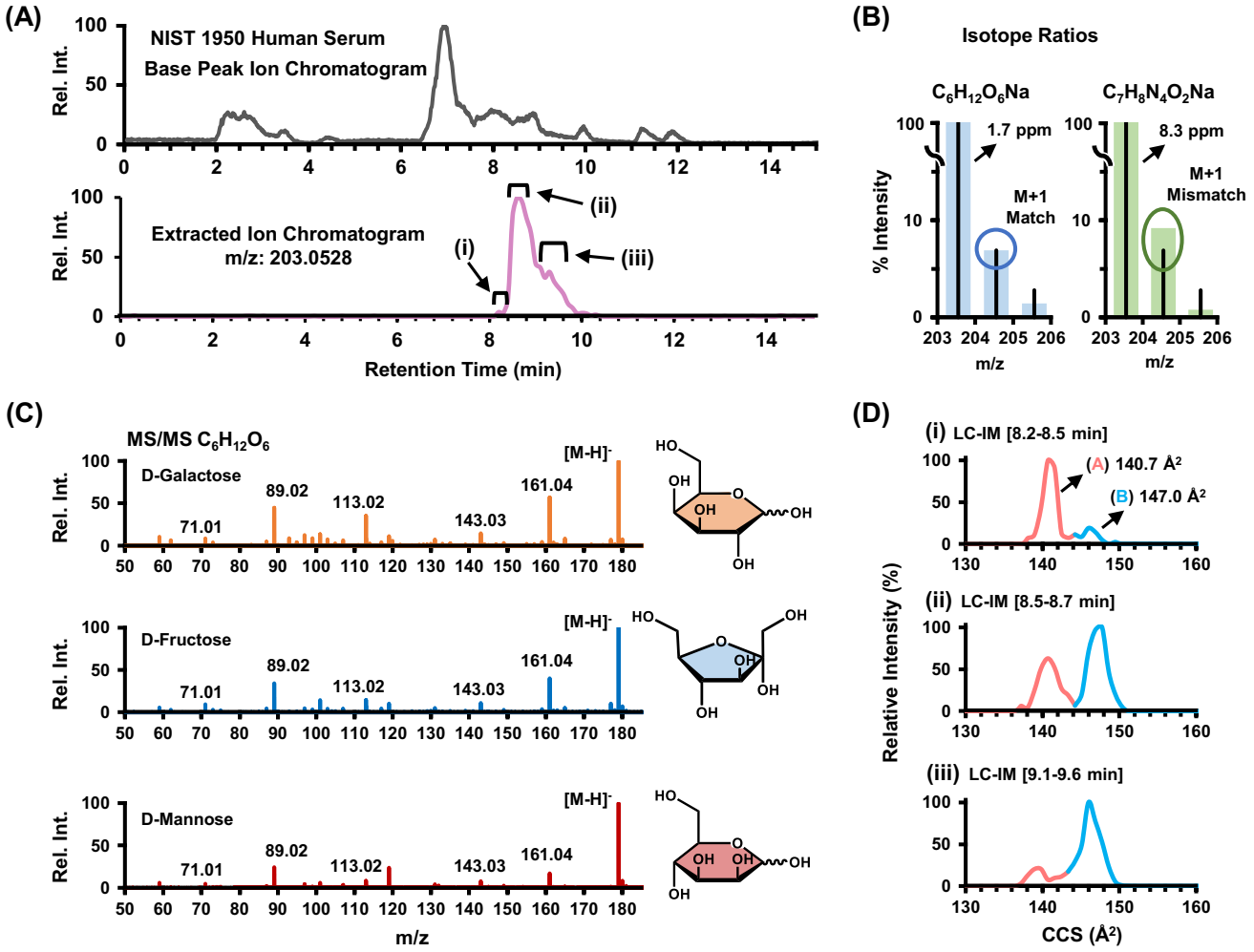


Figure 2.5. NIST 1950 Human Serum Analysis (A) HILIC base peak chromatogram for NIST 1950 human serum sample and (lower trace) the extracted ion chromatogram of *m/z* 203.0528, which consists of two distributions of interest that were further examined by isotope ratio pattern and ion mobility for structural characterization. **(B)** Expected and measured isotope ratio abundances for two possible chemical formulas corresponding to *m/z* 203 within 10 ppm. The chemical formula $C_6H_{12}O_6$ $[M+Na]^+$ more closely aligns with experimental measurements from the NIST serum both on basis of mass accuracy (2 ppm) and isotope ratio pattern (M+1). **(C)** Fragmentation spectra for isomers of with shared chemical formula $C_6H_{12}O_6$ $[M-H]^-$. **(D)** Selected ion mobility distributions for m/z 203 extracted over three time points in the chromatographic dimension. (Figure adapted from Nichols, et al.[51])

formula is determined, 9 potential isomers (including both constitutional rearrangements and stereochemistry for this chemical formula) exist within the MSMLS standards, all carbohydrates. These isomers possess almost identical fragmentation profiles (see $[M-H]^-$ ion, **Figure 2.5C)**, and sophisticated algorithms for identification by MS/MS are needed, an observation which has been previously noted in other carbohydrate studies.[50] Note that **Figure 2.5C** utilizes the deprotonated ion of $C_6H_{12}O_6$, as the $[M+Na]^+$ species noted in the other panels provides no fragmentation spectra due to ejection of the sodium charge carrier during collisional activation. Although previous studies utilized relative abundance ratios of fragment ions to determine molecular structure, this technique is time intensive and currently is not readily amended to rapid structural determination in untargeted workflows.[50] Similar to the chromatographic profile, ion mobility distributions obtained at three separate time points in the chromatogram (roman numerals) also indicate two separate chemical species present in the serum (**Figure 2.5D**). The collision cross sections measured for these two distributions helps narrow potential chemical structures from 9 potential isomeric forms down to 4 tentative identifications based on a CCS match within 1%. The smaller distribution at 140.7 $Å^2$ (light red, A) closely aligns with 3 isomers of $C_6H_{12}O_6$ in the standards (fructose, galactose and mannose at *ca.* 141.5 $Å^2$. The larger distribution at 147.0 $Å^2$ (light blue, B) closely aligns with α-D-glucose, which is noted at 146.3 $Å^2$ in the database. Although in this example ion mobility does not provide definitive identification of the compounds observed in the NIST serum, it does significantly reduce the possible candidates from the 9 potential structures noted in the database. By using collision cross section as additional metric for tentative identification, additional confidence in identifying molecular signatures can be gained in untargeted metabolomics.

2.4 Conclusions

In this work we have developed a CCS library based on primary metabolites obtained from the MSMLS library of analyte standards. As many key intermediates across metabolic pathways are formed through isomerization processes, utilizing orthogonal dimensions of separation in addition to mass analysis is imperative to fully characterize metabolic pathways. The intrinsic mass/mobility relationships for metabolites noted in this work, and others, illustrates a reproducible method of characterization for biochemical classes which interfaces seamlessly into the timescale of traditional LC/GC-MS workflows. Furthermore, we demonstrated that while additional resolving power in the m/z dimension is always advantageous, the diminishing returns of these efforts may not offset the additional analysis time required for ultra-high resolution mass acquisition (*i.e.* FT processes). However, orthogonal separation techniques such as LC and IM can often resolve many isomeric forms, facilitating their identification for a more comprehensive understanding of the biochemical implications of experimental samples. Finally, we have demonstrated the advantages of adding CCS as a molecular descriptor in untargeted metabolomic analyses through characterization of a well-studied human serum extract (NIST 1950) by LC-IM-MS.

2.5 Acknowledgements

This dissertation chapter was adapted from a published manuscript titled “Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry” by Charles M. Nichols, James N Dodds, Bailey S. Rose, Jaqueline A. Picache, Caleb B. Morris, Simona G. Codreanu, Jody C. May, Stacy D. Sherrod, and John A. McLean published in *Analytical Chemistry* **2018**, *90* (24), 14484-14492.

Financial support for aspects of this research was provided by The National Institutes of Health (NIH NIGMS R01GM092218 and NCI R03CA222-452-01) and under Assistance Agreement No. 83573601 awarded by the U. S. Environmental Protection Agency. This work has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the funding agencies and organizations. EPA does not endorse any products or commercial services mentioned in this publication.

2.6 References

(1) Crick, F. *Nature* **1970**, *227* (5258), 561–563.

(2) Joyce, A. R.; Palsson, B. Ø. *Nat. Rev. Mol. Cell Biol.* **2006**, *7* (3), 198–210.

(3) Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; Ebbels, T.; Kell, D. B.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjöström, M.; Trygg, J.; Wulfert, F. *Metabolomics* **2007**, *3* (3), 231–241.

(4) Sumner, S.; Snyder, R.; Burgess, J.; Myers, C.; Tyl, R.; Sloan, C.; Fennell, T. *J. Appl. Toxicol.* **2009**, *29* (8), 703–714.

(5) Kimmel, D. W.; Dole, W. P.; Cliffel, D. E. *J. Lipids* **2017**, *2017*, 1–9.

(6) May, J. C.; McLean, J. A. *Analytical Chemistry*. 2015, pp 1422–1436.

(7) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Kumar, D.; Showalter, M. R.; Arita, M.; Fiehn, O. **2018**, No. March 2017, 513–532.

(8) Kim, S.; Rodgers, R. P.; Marshall, A. G. *Int. J. Mass Spectrom.* **2006**, *251* (2–3), 260–265.

(9) Hofmann, J.; Hahm, H. S.; Seeberger, P. H.; Pagel, K. *Nature* **2015**, *526* (7572), 241–244.

(10) Kyle, J. E.; Zhang, X.; Weitz, K. K.; Monroe, M. E.; Ibrahim, Y. M.; Moore, R. J.; Cha, J.; Sun, X.; Lovelace, E. S.; Wagoner, J.; Polyak, S. J.; Metz, T. O.; Dey, S. K.; Smith, R. D.; Burnum-Johnson, K. E.; Baker, E. S. *Analyst* **2016**, *141* (5), 1649–1659.

(11) Yost, R. A.; Enke, C. G.; McGilvery, D. C.; Smith, D.; Morrison, J. D. *Int. J. Mass Spectrom. Ion Phys.* **1979**, *30* (2), 127–136.

(12) Lareau, N. M.; May, J. C.; McLean, J. A. *Analyst* **2015**, *140* (10), 3335–3338.

(13) Dodds, J. N.; May, J. C.; McLean, J. A. *Anal. Chem.* **2017**, *89* (1), 952–959.

(14) Ekroos, K.; Chernushevich, I. V.; Simons, K.; Shevchenko, A. *Anal. Chem.* **2002**, *74* (5), 941–949.

(15) Kanani, H.; Chrysanthopoulos, P. K.; Klapa, M. I. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2008**, *871* (2), 191–201.

(16) Lu, W.; Bennett, B. D.; Rabinowitz, J. D. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2008**, *871* (2), 236–242.

(17) Mairinger, T.; Causon, T. J.; Hann, S. *Curr. Opin. Chem. Biol.* **2018**, *42*, 9–15.

(18) Smith, C. a; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27* (6), 747–751.

(19) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (12), 1897–1905.

(20) Pringle, S. D.; Giles, K.; Wildgoose, J. L.; Williams, J. P.; Slade, S. E.; Thalassinos, K.; Bateman, R. H.; Bowers, M. T.; Scrivens, J. H. *Int. J. Mass Spectrom.* **2007**, *261* (1), 1–12.

(21) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; Hann, S.; Fjeldsted, J. C. *Anal. Chem.* **2017**, *89* (17), 9048–9055.

(22) Paglia, G.; Williams, J. P.; Menikarachchi, L.; Thompson, J. W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J.; Palsson, B. O.; Astarita, G. *Anal. Chem.* **2014**, *86* (8), 3985–3993.

(23) Paglia, G.; Astarita, G. *Nat. Protoc.* **2017**, *12* (4), 797–813.

(24) Pu, Y.; Ridgeway, M. E.; Glaskin, R. S.; Park, M. A.; Costello, C. E.; Lin, C. *Anal. Chem.* **2016**, *88* (7), 3440–3443.

(25) Paglia, G.; Angel, P.; Williams, J. P.; Richardson, K.; Olivos, H. J.; Thompson, J. W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A.; Plumb, R. S.; Grant, D. F.; Palsson, B. O.; Langridge, J.; Geromanos, S.; Astarita, G. *Anal. Chem.* **2015**, *87* (2), 1137–1144.

(26) Zheng, X.; Aly, N. A.; Zhou, Y.; Dupuis, K. T.; Bilbao, A.; Paurus, V. L.; Orton, D. J.; Wilson, R.; Payne, S. H.; Smith, R. D.; Baker, E. S. *Chem. Sci.* **2017**, *8*, 7724–7736.

(27) Simón-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.;

Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E. *Anal. Chem.* **2013**, *85* (24), 11725–11731.

(28) Telu, K. H.; Yan, X.; Wallace, W. E.; Stein, S. E.; Simón-Manso, Y. *Rapid Commun. Mass Spectrom.* **2016**, *30* (5), 581–593.

(29) Nichols, C. M.; May, J. C.; Sherrod, S. D.; McLean, J. A. *Analyst* **2018**, *143* (7), 1556–1559.

(30) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44* (D1), D463–D470.

(31) May, J. C.; Goodwin, C. R.; Lareau, N. M.; Leaptrot, K. L.; Morris, C. B.; Kurulugama, R. T.; Mordehai, A.; Klein, C.; Barry, W.; Darland, E.; Overney, G.; Imatani, K.; Stafford, G. C.; Fjeldsted, J. C.; McLean, J. A. *Anal. Chem.* **2014**, *86* (4), 2107–2116.

(32) Hines, K. M.; Ashfaq, S.; Davidson, J. M.; Opalenik, S. R.; Wikswo, J. P.; McLean, J. A. *Anal. Chem.* **2013**, *85* (7), 3651–3659.

(33) Hines, K. M.; Ross, D. H.; Davidson, K. L.; Bush, M. F.; Xu, L. *Anal. Chem.* **2017**, *89* (17), 9023–9030.

(34) Dwivedi, P.; Wu, P.; Klopsch, S. J.; Puzon, G. J.; Xun, L.; Hill, H. H. *Metabolomics* **2008**, *4* (1), 63–80.

(35) McLean, J. A. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (10), 1775–1781.

(36) Goodwin, C. R.; Fenn, L. S.; Derewacz, D. K.; Bachmann, B. O.; McLean, J. A. *J. Nat. Prod.* **2012**, *75* (1), 48–53.

(37) Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. *Nucleic Acids Res.* **2018**, *46* (W1), W486–W494.

(38) May, J. C.; McLean, J. A. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 387–409.

(39) Groessl, M.; Graf, S.; Knochenmuss, R. *Analyst* **2015**, *140* (20), 6904–6911.

(40) Li, H.; Giles, K.; Bendiak, B.; Kaplan, K.; Siems, W. F.; Hill, H. H. *Anal. Chem.* **2012**, *84* (7), 3231–3239.

(41) Giles, K.; Williams, J. P.; Campuzano, I. *Rapid Commun. Mass Spectrom.* **2011**, *25* (11), 1559–1566.

(42) D'Atri, V.; Causon, T.; Hernandez-Alba, O.; Mutabazi, A.; Veuthey, J. L.; Cianferani, S.; Guillarme, D. *J. Sep. Sci.* **2018**, *41* (1), 20–67.

(43) Dodds, J. N.; May, J. C.; McLean, J. A. *Anal. Chem.* **2017**, *89* (22), 12176–12184.

(44) Fernandez-Lima, F. A.; Kaplan, D. A.; Park, M. A. *Rev. Sci. Instrum.* **2011**, *82* (12).

(45) Deng, L.; Ibrahim, Y. M.; Baker, E. S.; Aly, N. A.; Hamid, A. M.; Zhang, X.; Zheng, X.; Garimella, S. V. B.; Webb, I. K.; Prost, S. A.; Sandoval, J. A.; Norheim, R. V.; Anderson, G. A.; Tolmachev, A. V.; Smith, R. D. *ChemistrySelect* **2016**, *1* (10), 2396–2399.

(46) N. Nikolaev, E.; N. Vladimirov, G.; Jertz, R.; Baykut, G. *Mass Spectrom.* **2013**, *2*, S0010–S0010.

(47) Zubarev, R. A.; Makarov, A. *Anal. Chem.* **2013**, *85* (11), 5288–5296.

(48) Marshall, A. G.; Hendrickson, C. L.; Shi, S. D.-H. *Anal. Chem.* **2002**, *74* (9), 252 A-259 A.

(49) Bowden, J. A.; Heckert, A.; Ulmer, C. Z.; Jones, C. M.; Koelmel, J. P.; Abdullah, L.; Ahonen, L.; Alnouti, Y.; Armando, A. M.; Asara, J. M.; Bamba, T.; Barr, J. R.; Bergquist, J.; Borchers, C. H.; Brandsma, J.; Breitkopf, S. B.; Cajka, T.; Cazenave-Gassiot, A.; Checa, A.; Cinel, M. A.; Colas, R. A.; Cremers, S.; Dennis, E. A.; Evans, J. E.; Fauland, A.; Fiehn, O.; Gardner, M. S.; Garrett, T. J.; Gotlinger, K. H.; Han, J.; Huang, Y.; Neo, A. H.; Hyötyläinen, T.; Izumi, Y.; Jiang, H.; Jiang, H.; Jiang, J.; Kachman, M.; Kiyonami, R.; Klavins, K.; Klose, C.; Köfeler, H. C.; Kolmert, J.; Koal, T.; Koster, G.; Kuklenyik, Z.; Kurland, I. J.; Leadley, M.; Lin, K.; Maddipati, K. R.; McDougall, D.; Meikle, P. J.; Mellett, N. A.; Monnin, C.; Moseley, M. A.; Nandakumar, R.; Oresic, M.; Patterson, R.; Peake, D.; Pierce, J. S.; Post, M.; Postle, A. D.; Pugh, R.; Qiu, Y.; Quehenberger, O.; Ramrup, P.; Rees, J.; Rembiesa, B.; Reynaud, D.; Roth, M. R.; Sales, S.; Schuhmann, K.; Schwartzman, M. L.; Serhan, C. N.; Shevchenko, A.; Somerville, S. E.; St. John-Williams, L.; Surma, M. A.; Takeda, H.; Thakare, R.; Thompson, J. W.; Torta, F.; Triebl, A.; Trötzmüller, M.; Ubhayasekera, S. J. K.; Vuckovic, D.; Weir, J. M.; Welti, R.; Wenk, M. R.; Wheelock, C. E.; Yao, L.; Yuan, M.; Zhao, X. H.; Zhou, S. *J. Lipid Res.* **2017**, *58* (12), 2275–2288.

(50) Fang, T. T.; Bendiak, B. *J. Am. Chem. Soc.* **2007**, *129* (31), 9721–9736.

(51) Nichols, C. M.; Dodds, J. N.; Rose, B. S.; Picache, J. A.; Morris, C. B.; Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A., "Untargeted Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility – Mass Spectrometry" *Anal. Chem*., **2018**, *90* (24) 1448-14492.

CHAPTER III

Collision Cross Section Compendium to Annotate and Predict Multiomic Compound Identities

3.1 Introduction

Mass spectrometry (MS) has become a central technique for the investigation of the global profile of biochemical species in molecular phenomic studies.[1,2] These studies aim to address the grand challenges of biomedical research including comprehensive descriptions of biological systems, natural product and drug discovery endeavors, omics sciences to improve health outcomes, and progress in synthetic biology.[3–6] As the complexity of the systems being studied increases, so must the ability to increase analyte coverage. Orthogonal separation techniques such as gas and liquid chromatography are often used in conjunction with MS to improve coverage. However, feature annotation and identification from such experiments can be challenging due to analyte co-elution and retention time variability among other issues.[7]

These challenges can be addressed with the use of additional analytical separation techniques, such as ion mobility spectrometry coupled to MS (IM-MS), which is selective to the analyte gas phase structure.[3,8] One practical benefit of using gas-phase ion mobility is that there are no memory effects or sample-to-sample carryover due to the continuous replacement of the separation gas. Additionally, IM separations do not require disposable solvents or packed columns and are amenable to all ionizable chemical species. The main advantages of IM-MS are an increase in analytical peak capacity as well as the ability to measure an analyte's gas phase mobility by means of an experimental drift time.[9] This mobility can then be used to calculate an analyte's collision cross section (CCS), a rotationally averaged surface area of the molecule in its ionic form.

These CCS values are specific and can be compared across different laboratories making them particularly well-suited for species identification and characterization purposes. Previous studies indicate that the level of reproducibility varies across analyte classes.[10,11] A recent study using drift tube IM-MS has shown that CCS values can be measured within a 0.30% RSD when data is acquired with a previously established standardized method.[12]

As a result of these advantages, several research groups have used IM-MS to build CCS libraries in which the measured values serve as additional molecular descriptors for assigning identities to unknown analytes. While not an exhaustive list, a few of the larger libraries to note are: Li and colleagues' peptide database which includes >2300 CCS values,[13] Pagel and colleagues' glycomics database of >900 CCS values,[14] and Xu and colleagues' small molecule database containing >1400 CCS values.[15] Additionally, many excellent smaller CCS libraries have been generated for lipids,[16–18] primary metabolites,[18–20] secondary metabolites and other natural products,[18,21,22] as well as illicit substances[23] among others.

While each of these libraries adds to the working knowledge of the IM-MS field, there remain challenges that need to be addressed. The first is reconciling CCS measurements across various IM implementations such as drift tube (DTIMS), traveling wave (TWIMS), ion trapping (TIMS), and structures for lossless ion manipulation (SLIM) techniques. Inherently, these techniques utilize different methodologies for determining the gas-phase CCS, namely DTIMS (and drift tube-based SLIM) utilize the fundamental ion mobility relationship for correlating the measured arrival times directly to CCS, whereas the other IM techniques obtain a CCS value through calibration. In order to reconcile non-DTIMS CCS values with DTIMS values, proper calibrants must be chosen for a given experiment, which can prove challenging.[10,24] An in-depth

discussion of considerations for comparing CCS information obtained from different IM techniques can be found in a recent review by Gabelica *et al.*[25]

Another challenge lies in the difficulty of accurately and efficiently extracting drift time measurements from raw data files in large scale. Currently, most DTIMS drift times for chemical standards are manually extracted which improves accuracy at a cost of throughput. However, several software options exist that aim to automate the extraction of drift times on a large scale and/or predict drift times.[26,27] The recent IM-MS analysis addendum to Skyline is one example that has made considerable strides in these efforts,[28,29] but the IM-MS field is still working towards a streamlined analytic workflow.

Other informatics programs aim to predict CCS values based on experimental data and chemical structure. Some examples of these software include Zhu and colleague's machine learning algorithms for metabolites (MetCCS) and lipids (LipidCCS).[16,19] A major barrier to the success of machine learning CCS prediction is that algorithm training sets are generally not yet large and/or specific enough.[30] An alternative strategy recently described by Colby, et al. is the in silico chemical library engine (ISiCLE) workflow which utilizes a combination of molecular dynamics, quantum chemistry, and ion mobility calculations in order to predict CCS values based on theoretical structure information.[31] These CCS prediction efforts are critically important for determining CCS values where empirical measurements on authentic chemical standards are unavailable.

To aid in the mainstream adoption of IM in analyte identification workflows, we explored the potential in curating libraries of empirical CCS values measured via ion mobility into a single, self-consistent compendium. The Unified CCS compendium presented herein serves as a tool where new data from the community can be vetted using a quality control protocol and

subsequently integrated. Included in this curated compendium are several prevalent calibrant sets (polypeptides, branched phosphazenes, inorganic salt clusters, etc.), as well as molecular standards from a variety of chemical classes measured using DTIMS. These data sets can be used as reference values for other IM-MS techniques. Furthermore, this tool incorporates annotative features (i.e. visualization of chemical locales of molecules) and predictive statistics (chemical structure-based trends) to aid in identifying unknown biochemical species. These predictive trends serve as a powerful filter for increasing confidence in tentative identifications. In order to demonstrate the efficacy of this approach, the structural filtering method was applied to metabolites in a human serum sample. The full interactive visualization of the compendium, as well as inclusion criteria and guidelines for submitting additional CCS measurements, can be found as an open access tool.[32]

3.2 Experimental Methods

3.2.1 Materials and Instrumentation Methanol (MeOH), water, acetonitrile (ACN), isopropanol (IPA), and formic acid of Optima grade purity were purchased from Fisher Scientific (Fair Lawn, NJ). Anhydrous methyl-tert-butyl ether (MTBE) was purchased from Sigma Aldrich (St. Louis, MO). Normal human serum was purchased from Utak (Valencia, CA). A mixture of fluoroalkyl phosphazenes, tris(fluoroalkyl) triazines, betaine, and trifluoracetic acid reference standards were purchased from Agilent Technologies (G1969-85000, Santa Clara, CA). In this manuscript, liquid chromatography MS (LC- MS) and LC-IM-MS data were acquired using a 1290 Infinity LC system and a 6560 IM-QTOF MS (Agilent Technologies).

3.2.2 Data Sources and Inclusion Parameters The primary sources of the 3833 IM-MS measurements included in the compendium are reported in a series of manuscripts found elsewere.[10,12,18,33–38] In order to provide highly repeatable and reproducible data, the compendium currently only contains CCS values calculated from the fundamental low-field ion mobility equation (Mason–Schamp relationship) incorporated into a standardized inter-laboratory protocol for single field and stepped field DTIMS acquisition on a commercial uniform-field IM-MS instrument (6560, Agilent).[12,39,40] In-depth information about single and stepped field DTIMS has previously been described.[12] All measurements were acquired in triplicate and aligned with a suite of 13 reference standards (Agilent Technologies) containing symmetrically-branched fluoroalkyl phosphazines, namely hexaxis(fluoroalkoxy)phosphazines, tris(fluoroalkyl)triazines, betaine, and trifluoracetic acid. These reference standards were previously measured with very high precision; and it is currently believed that these CCS values are among the most accurate obtained to date.[12]

In total, there are 1216 single field measurements within the compendium; and the average relative standard deviation (RSD) for the single field measurements is 0.12%. Compounds were matched to reference standards' values from an inter-laboratory study.12 The average percent error of compendium CCS measurements was found to be 0.04% and -0.33% for positive and negative modes, respectively, with all percent error values at 0.58% for both polarities. The remaining 2617 stepped field values were reconciled by calculating a “true effective length” for each data set (data set defined as a group of measurements collected in a one-day acquisition period) using calibrant measurements within the set. This “true effective length” was then used to align measurements with reference standards' values. More details and tools to calculate “true effective length” as well as instructions to calibrate acquired stepped field CCS data is found in the Supplementary Information (ESI) Section S3. Once each data set was individually scaled, the average RSD for

stepped field measurements was calculated to be 0.32%. When compared to inter- laboratory reference values, average percent errors were 0.07% and 0.01% for positive and negative modes, respectively. Ninety- one percent of matched values had a percent error ≤1% for both polarities. These empirically-derived metrics, in conjunction with known errors propagated in this system,[12] subsequently used as the compendium's data inclusion criteria. Full descriptions of the inclusion criteria and instructions for submitting data to the compendium can be found in ESI Section S2.

3.2.3 Data Preparation, Statistical Modeling, and Visualization Data from all sources were curated into a unified format using the statistical computing programming environment R (R Foundation for Statistical Computing, Vienna, Austria).[41] This unified format includes the following information for each compendium entry: name, formula, CAS registry number (when available), mass-to-charge (m/z), charge state, ion species, size- to-charge (CCS/z), percent RSD, and number of observed DTIMS peaks. Charge-normalization of mobility measurements[39] via CCS/z was utilized to preserve the original drift time scale and analysis consistency.[39] In drift time spectra, ions of similar mass and higher charge states typically have smaller drift times than lower charge state ions; and therefore, appear lower when visualized in drift time vs. m/z space. Contrastingly, higher charge state ions appear higher than lower charge state ions when visualized in CCS vs. m/z space. By charge- normalizing, ions appear in CCS/z vs. m/z space as they would in drift time vs. m/z space. Furthermore, when values were not charge-normalized, statistical modeling could not be standardized and was charge-state dependent. The number of DTIMS peaks observed for each molecule is included in the compendium. The number of DTIMS peaks observed for each molecule is included in the compendium. These data meet the outlined criteria and follow the standardized IM-MS data reporting efforts led by Gabelica, et al.[42] Briefly, all observed DTIMS

peaks are reported in the online compendium compound table via a peak number assignment where the smallest CCS/z (earliest drift time) will be assigned number 1 and subsequent peaks will be assigned 2, 3, etc. Compounds with one observed peak will be assigned a "1". Additional information regarding DTIMS peak annotation can be found in ESI Section S1.

The unified format also includes a hierarchical chemical classification for each compound which includes a kingdom, super class, class, and subclass based on structure. This was performed via the ClassyFire web-based application which operates using a comprehensive chemical ontology (ChemOnt) that classifies each molecule based on its SMILES or InChi Key identifier as an input.[43,44] For example, a phosphatidylcholine would be classified as a member of the organic compound kingdom, the lipids and lipid-like molecules super class, the glycerophospholipid class, and the glycerophosphocholine subclass.

Iterative nonlinear regression modeling was performed using the R program for each chemical class and subclass that contained at least ten data points. Source code for this statistical modeling is provided on the McLean Research Group Github.[45] Each class was tested against three nonlinear regression models: a power fit (PF), a four-parameter sigmoidal fit (4P), and a five-parameter sigmoidal fit. Representative equations for these models can be found in ESI Section S5.† These models were chosen based on previous work.[34,46,47] The goodness of fit for each model was assessed using the corrected Akaike information criterion (AICc) for each of the three models. This conservative metric accounts for small sample sizes, bias correction, and varying degrees of freedom in nonlinear candidate models; and has previously been shown to be highly reliable when comparing nonlinear models.[48,49] The model with the lowest AICc value was taken to be the best fit. Ninety-nine percent confidence (CI) and predictive (PI) intervals were calculated as described

in ESI Section S5 eqn (4) and (5),† respectively. CI and PI were calculated in the same manner for all nonlinear regressions.

The Unified CCS compendium was visualized using the following open-source R packages: plotly (v4.7.1), ggplot2 (v2.2.1.900), data.table (v1.10.4-3), plyr (v1.8.4), and shiny.[50–54] Source code for the compendium GUI can be found on the McLean Research Group Github.[45]

2.3.4 Evaluation of the Compendium in the Analysis of Human Serum Non-endogenous fatty acids 17:0 and 19:0 were used as internal standards and added into 100 mL control human serum. 800 mL of cold MeOH (-20°C) was subsequently added and the sample was stored at ?20°C overnight to precipitate out proteins. The sample was subsequently centrifuged at 14 000 rpm and 4°C for five minutes. The supernatant was collected; and 2.4 mL ice cold MTBE and 800 mL ice cold water were added. This MTBE:MeOH:water sample was vortexed then centrifuged at 10 000 rpm and 4°C for ten minutes. The nonpolar liquid fraction was siphoned, dried under vacuum, and stored at -20°C until use. Dried fractions were resus- pended in 100 mL of 70:18:12 water:IPA:ACN and analyzed via LC-MS and LC-IM-MS. Further details are provided in ESI Section S6. LC-MS data was analyzed using Progenesis QI (v2.3, Nonlinear Dynamics, Durham, NC). Resulting features were tentatively identified using the Metlin Metabolomics and LipidBlast databases.[55,56] LC-IM-MS raw acquisition files were converted to mzML format using MSConvert (v3.0, ProteoWizard).[57] Drift time values from LC-IM-MS experiments for individual process replicates were extracted using an internally developed Python script[45] in which drift times were matched against the retention time and m/z of the aforementioned tentatively identified compounds. These match functions had a threshold of 30 seconds (or 1% variation) for retention time and 5 ppm for m/z, respectively. Once drift times were extracted from the mzML data files,

CCS/z values were calculated from the Mason–Schamp relationship using the averaged drift times. Chemical class probability hierarchies were analyzed using distance of the mean calculations based on where serum CCS/z values fell within the compendium as compared to the regression models.

3.3 Results and Discussion

3.3.1. CCS Compendium Properties The Unified CCS compendium compiled in this work consists of a total of 3833 CCS values (see inclusion criteria in the Experimental section) obtained with uniform drift tube instruments in nitrogen drift gas utilizing a standardized CCS protocol.[12] Measurements consist of 2740 cations and 1093 anions, all of which were acquired in replicates of ≥3. Associated measurement RSDs can be found on the web-based compendium.[32] Thirteen ion species types are represented as indicated in **Figure 3.1A**. The most common species observed were proton coordination (38%), proton loss (27%), and sodium coordination (25%). Ion species were assigned based on the charge source of the molecule. For example, if a compound was observed as [M+ 2Na–H]+, the ion species was labelled as "+Na". Likewise, if a compound had multiple charge carriers of the same type, such as [M +4H]4+, it was labelled as "+H". Compounds with multiple different equal charge carries, such as [M +H+K]2+ were recorded as both "+H" and "+K". The charge distribution, **Figure 3.1B**, in the compendium ranged from +1 to +31 for cations

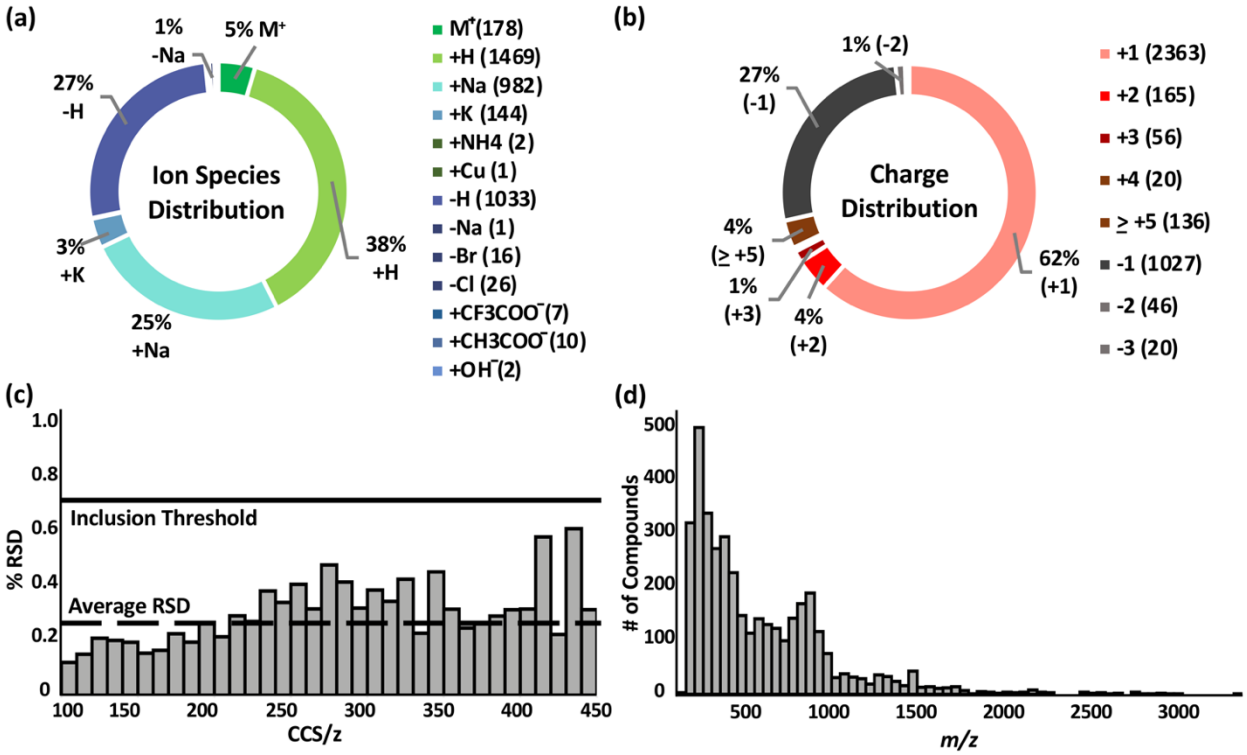


Figure 3.1 CCS Compendium Characterization (a-b) Overall distribution of the 3833 measured ions from (+) and (-) ion polarity modes by ion species and charge state. (c) Relative standard deviation (RSD) of all measurements binned by CCS/z. Global average RSD is 0.25%, and Compendium RSD threshold is 0.7%. (d) Distribution of ions contained in the database as a function of the *m/z*. (Figure adapted from Picache, et al.[60])

and -1to -3 for anions. More than 90% of the compounds were singly or doubly charged. Overall, replicate measurements were highly reproducible as evaluated by RSD. The global average RSD was 0.25%; and 97% of all compounds had an RSD of <1.0%. The average RSD per CCS/z bin is shown in **Figure 3.1C**. RSD is observed to increase as CCS/z increases due to multiple observed conformers in larger molecules. Under highly controlled interlaboratory experimental conditions, RSD is <0.3%;[12] and the empirical RSD threshold of 0.7% for the compendium is a practical limit for data from independent studies. The compendium data set spans a m/z range of ca. 74 to ca. 3300 Da. However, most of the compounds are <1500 Da. The full distribution of compound masses is shown in **Figure 3.1D**.

3.3.2 CCS Compendium Visualization The data set was visualized using code written in the R language. The graphical user interface (GUI) of the Unified CCS compendium is shown in **Figure 3.2A** and is accessible online.[32] The default view for this GUI is to show all data grouped by super class. Users have the ability to zoom and select regions of interest which facilitates maneuvering densely populated areas. By hovering the cursor over any data point, as shown in **Figure 3.2B**, users can access specific information regarding the corresponding entry including the compound's name, molecular formula, CAS identity, m/z, observed charge species, CCS/z and associated RSD, source citation, and digital object identifier. The interactive GUI can be tailored to the user's needs. Search functionality allows users to find data on any compound within the compendium's compound table. Users can also isolate a specific data subset based on ion polarity, adduct type, super class, class, and data source. Subsetting data by super class or class reveals its CCS/z vs. m/z area of occupancy.

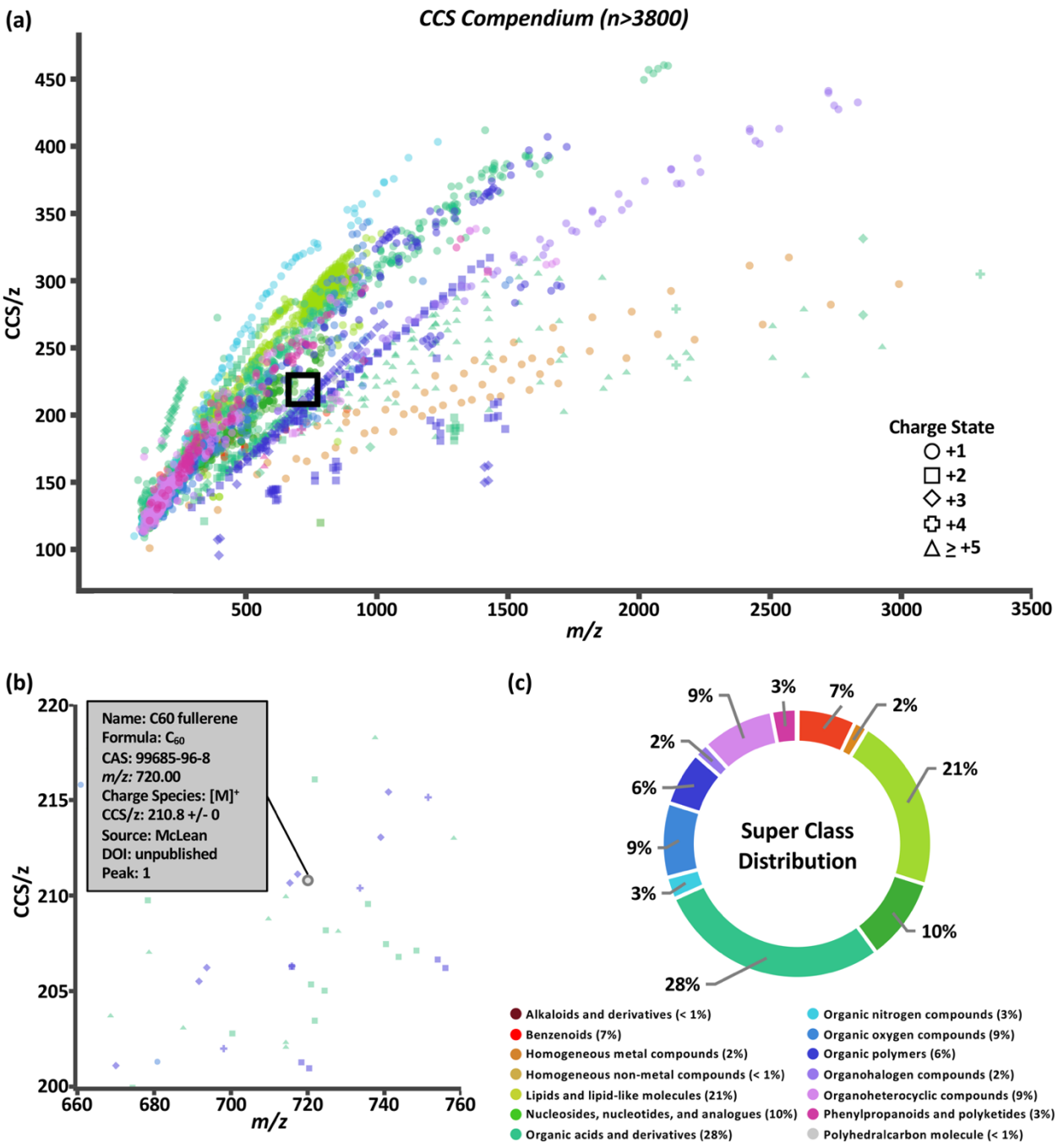


Figure 3.2 Compendium Interface (a) depicting measured data points classified into super classes indicated in the legend above. An enlarged version of the area within the black box is shown in (b) to illustrate how each data point reveals an information box in the online Compendium. (c) Distribution of compounds across the 14 structural super classes. (Figure adapted from Picache, et al.[60])

The compendium covers 14 super classes which delineate into 80 classes and 157 subclasses. The distribution of compounds into each super class is summarized in **Figure 3.2C**. A list of super classes including m/z range and number of compounds per super class is summarized in **Table 3.1**. Super classes and their subsequent classes are further described in ESI Section S4. Full classification of individual compounds can be found on the web-based compendium.[32] Of the 80 classes, 48 had a sufficient ($n > 10$) number of data points to undergo regression fitting tests. In total, 24 classes and 24 subclasses were modeled. As new data is added and regression fitting algorithms are iterative, it should be noted that the most up-to-date regression model equations can be found online.[32] A few observations can be made from the data fitting study. Both four-parameter (**Appendix C, Section C5, equation C3.2**) and five-parameter (**Appendix C, Section C5, equation C3.3**) regressions were the best fit more frequently for classes in which m/z range included masses under 200 Da. This suggests a potential minimum observable CCS due to the asymptotic nature of sigmoidal curves. In theory, the IM-derived CCS will converge on the CCS of the neutral drift gas which, for sufficiently low CCS measurements, should manifest as a non-zero y- intercept in these CCS/z vs. m/z projections. In the canonical literature, this minimally-observable ion mobility measurement is referred to as the gas polarization limit.[39] The smallest CCS/z measurement in the compendium is 100.81 A^2 for a single cesium cation at m/z 132.90. Presently, more data points are needed to generate functional forms of a global fit.

Table 3.1. Curated CCS Compendium Super Classes

Super Class	*m/z* Range	N
Alkaloids and derivatives	138 – 609	4
Benzenoids	108 – 887	269
Homogeneous metal compounds	132 – 2991	62
Homogeneous non-metal compounds	144	1
Lipids and lipid-like molecules	125 – 1017	810
Nucleosides, nucleotides, and analogues	226 – 809	386
Organic acids and derivatives	89 – 3302	1085
Organic nitrogen compounds	74 – 1233	102
Organic oxygen compounds	105 – 1506	345
Organic polymers	294 – 1724	250
Organohalogen compounds	301 – 2834	66
Organoheterocyclic compounds	96 – 1684	335
Phenylpropanoids and polyketides	133 – 1424	116
Polyhedralcarbon molecules	210 – 227	2

(Table adapted from Picache, et al.[60])

3.3.3 Predictive Structural Chemical Trends While the compendium visualizes the simple, yet fundamental aspects of the relationship between CCS/z and m/z, its highest utility lies in its predictive potential. To support predictive analysis, a 99% confidence interval (CI) and 99% predictive interval (PI) were generated as described in **Appendix C, Section C5 equation C3.4** and **equation C3.5**, respectively, for each class fit with a nonlinear regression. Briefly, the CI depicts the value range in which the regression mean is expected to be for normally distributed data.[58] For our data, the mean CCS/ z value for a given m/z should be contained within the CI in 99% of cases. The upper and lower CI limits are depicted as the outer solid lines throughout Fig. 3.3. The distance between the two limits is closest where the data point density is highest and prediction error is lowest along the regression model. The 99% PI depicts the 'y' variable value (CCS/z) range expected for 99% of data points at a given 'x' value (m/z).[58] For our purposes, it represents the CCS/z range 'expected for 99% of data points at a given m/z.

Figure 3.3 is a representative example of this data correlation process. It depicts the super class "Organoheterocyclic compounds" which contain many humanmetabolites and natural products. Three classes within "Organoheterocyclic compounds" are shown in **Figure 3.3B–D**. The "Quinolines and derivatives" (**Figure 3.3B**) and "Imidazopyrimidines" classes (**Figure 3.3C**) were best fitbya4P regression model. The "Pteridines and derivatives" class (**Figure 3.3D**) was fit best by the PF regression. In these cases, data fit regressions and corresponding CIs and PIs define the CCS/z vs. m/z space that 99% of data for diazines, imidazopyrimidines, and pteridines and derivatives should occupy. While current AICc values indicate these models are appropriate, the specificity and predictability of these intervals will improve with the inclusion of more data and further delineation of each class into subclasses.

In the compendium, the 99% confidence and predictive intervals included in the data projections are calculated directly from the compendium data, therefore the majority of the empirical measurements within the dataset will fall within these intervals. As these bands represent a probability, there remains the possibility that CCS values for compound standards will fall outside of these projections, and users should examine these cases on an individual basis to determine if CCS values are repeatably and reproducibly outside of the predicted range. For example, multimers dissociating occurring after the ion mobility measurement but prior to mass analysis (i.e., post- mobility ion activation) would lead to a larger than expected drift time and corresponding CCS. Additionally, CCS values for unknown analytes/isomers obtained from untargeted experiments represent previously unmeasured peak features which could fall outside of the interval bands. In these scenarios, the user should exercise caution in determining if the predicted structural class is appropriate.

3.3.4 The Compendium as an Identification Filter To test the predictability and filtering abilities of the compendium, metabolites were extracted from control human serum analyzed using LC-MS or LC-IM-MS workflows. In the LC-MS data, 4719 deconvoluted compounds were observed. In total, 955 tentative identifications were matched using conservative criteria for exact mass (<10 ppm) and isotope distribution (70%) using Metlin metabolomics and LipidBlast databases.[55,56] In order to append drift time values to these tentative identifications, an in-house Python script (available online) was developed.[45] Using this script, we can extract drift times at a rate of 4 x 10^5 measurements in $\leq$1 h per sample. Drift times from each of the three technical replicates were aligned to the tentative identifications based on retention time and m/z. In these data, a majority of the aligned drift times were self-consistent with an RSD <1%. The drift times were averaged and

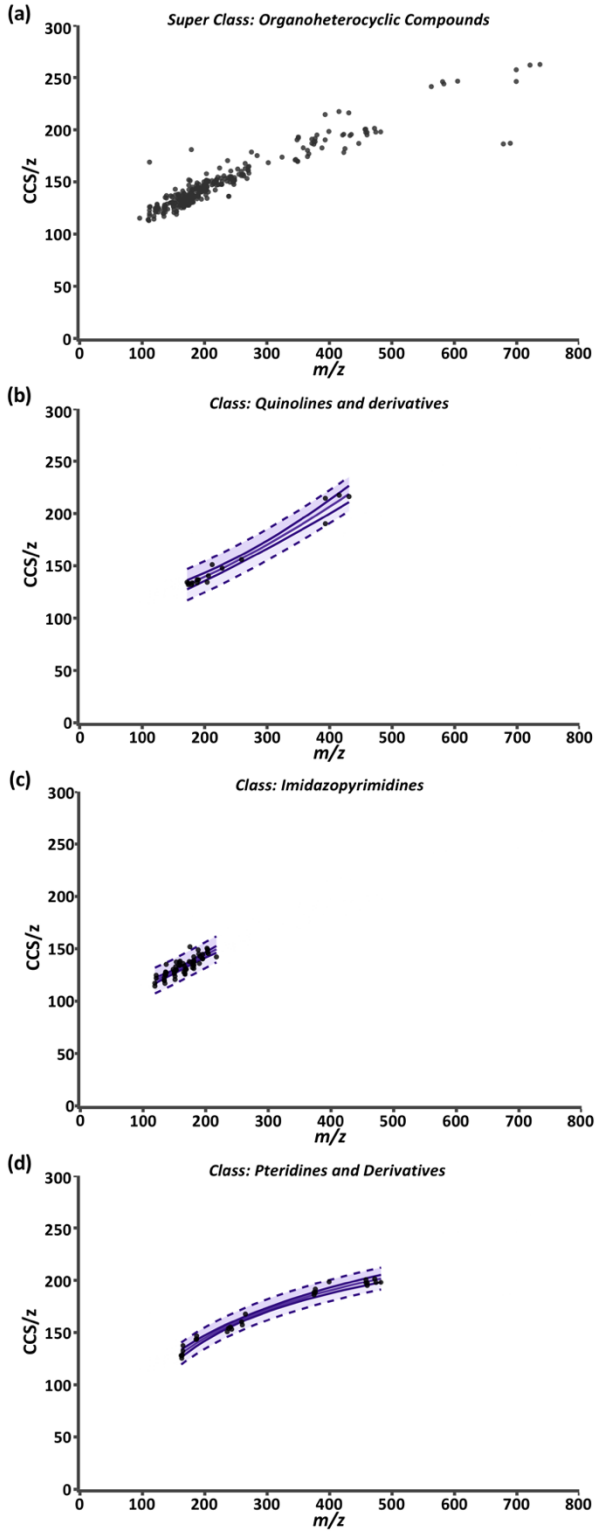


Figure 3.3 CCS Compendium Regression Models (a) Compendium GUI output of all ion entries within the "Organoheterocyclics" super class. (b) "Quinolines and derivatives" class; and a 4P regression. (c) "Imidazopyrimidines" class; and a 4P regression. (d) "Pteridines and derivatives" class; and a PF regression. For (b-d), the center solid line is the regression model, outer solid lines are 99% CI and the dash lines are 99% PI. (Figure adapted from Picache, et al.[60])

used to calculate CCS/z values using the single-field extension of the Mason–Schamp relationship. The annotated serum data is represented in **Figure 3.4A** (labeled "tentatively identified serum metabolites"). Superimposing the serum data over the Unified CCS compendium data (**Figure 3.4A**) illustrates that the tentatively identified compounds have equivalent mobility–mass correlations as known chemical compounds.

For proof-of-concept purposes, the serum data was subset into compounds tentatively identified as lipids. Compounds in the green highlighted area of **Figure 3.4A** represent the CCS/z vs. m/z space within the Unified CCS compendium containing any and all lipid regressions generated for data in the "Lipids and lipid-like molecules" super class. In total, 550 compounds present in the serum sample were tentatively identified as lipids; and 422 of these compounds overlapped with at least one of the lipid class and/or lipid subclass regression models. Distance from the mean values were then calculated to prioritize the probability that a serum compound belonged to a given lipid class. An example of this process is depicted in **Figure 3.4B** for the compound with m/z 744.49 and CCS 278.2 ˚A2 (gold circle). Potential tentative identifications for m/z 744.49 included 53 isomers of glycerophosphocholines (PC) and glycerophosphoethanolamines (PE). This unknown compound (gold line, **Figure 3.4B** call-out box) was 2.54 standard deviations away from the PC subclass regression model (blue line, Fig. 4b call-out box) and 9.44 standard deviations from the PE subclass regression model (red line, Fig. 4b call-out box). At 2.54 standard deviations, this compound was within the 99% confidence interval of the PC subclass regression model and had a difference of about 1.5 ˚A2 from the mean CCS/z value of PCs at m/z 744. Using this Unified CCS compendium, there is more data to suggest the unknown compound's tentative identity is a PC. Thus, a putative identification and higher confidence in its assignment can be attributed.

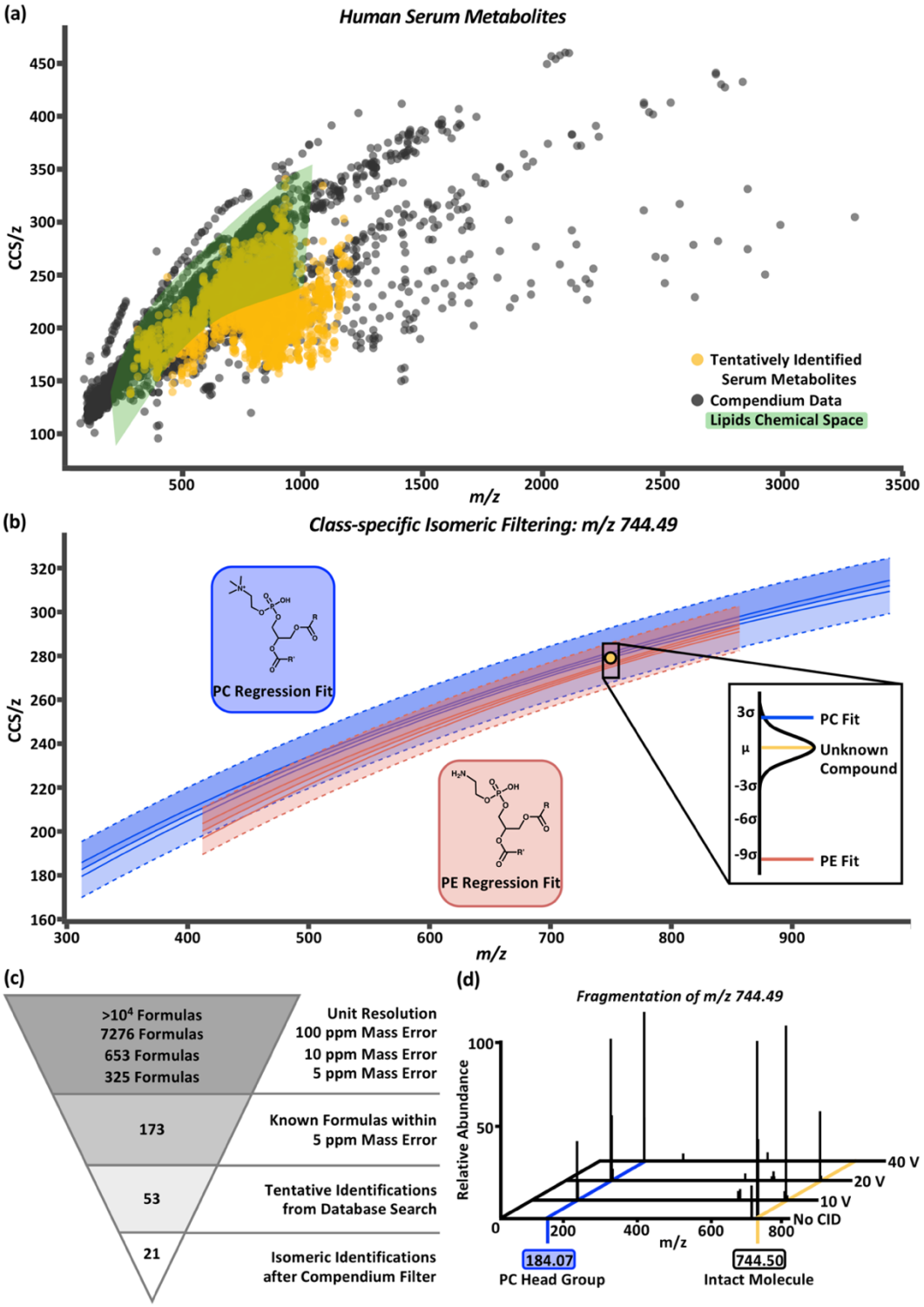


Figure 3.4 Human Serum Analysis via CCS Compendium (a) Overlay of human serum metabolites (gold) with the Compendium (black). Green area represents CCS/z vs *m/z* space occupied by any and all lipid subsets within the Compendium. (b) Example plot for class-specific filtering of an unknown serum compound, *m/z* 744.49 and CCS 278.2 Å^2 (gold circle), tentatively identified as a PC (blue regression model) or PE (red regression model). The probability of the unknown compound's class falling within the PC or PE class is shown in the call out box. Based on distance from the mean calculations, the compound falls within 2.54 standard deviations of the PC regression model and 9.44 standard deviations of the PE regression model which indicates the unknown compounds has a higher probability of being a PC than a PE. (c) Molecular identification workflow for the unknown compound depicted in panel b. After Compendium filtering, identifications were reduced to 21 PC isomers with the *m/z* 744.49. (d) Fragmentation of the isolated *m/z* 744.49 at CID 0V, 10V, 20V, and 40V. An increase in the intensity for *m/z* 184.07, corresponding to the phosphocholine head group mass, is observed with increase in collision voltage. (Figure adapted from Picache, et al.[60])

The molecular identification workflow for m/z 744.49 is summarized in **Figure 3.4c**. The m/z 744.49 was deconvoluted to its neutral mass of 705.53 Da. At unit resolution, there are tens of thousands of potential chemical formulas with a mass of 705 Da. Within 100 ppm mass error of705.53 Da, there are 7276 possible chemical formulas. Subsequently, there are 653 chemical formulas within 10 ppm mass error and 325 chemical formulas within 5 ppm mass error (the observed mass error). Of these 325 formulas, 173 are known compounds found in the PubChem database. Heuristic filtering based on instrumentation mass accuracy, mass defect, isotope distribution, and information from orthogonal separations enables tentative identification of compounds with a specified level of confidence. In this example, 53 tentative PC and PE identifications were returned after heuristic filtering through Progenesis QI. Using the compendium, this list can be further narrowed into 21 PC isomers with the neutral mass 705.53 and m/z 744.49.

To validate our PC prediction, m/z 744.49 underwent mass isolation from the serum matrix and was fragmented using collision induced dissociation at 0 V, 10 V, 20 V, and 40 V. The mass spectra, shown in **Figure 3.4d**, demonstrate the increase in the intensity of m/z 184.07, the signature m/z of a phosphocholine head group, as collision energy increased. While further investigation using chemical standards can lead to high-confidence identifications of unknown compounds, using the CCS filtering workflow presented here allows investigators to achieve high confidence in assigning the chemical class to an unknown molecule using IM-MS datasets. This predictive ability is expected to be particularly important for chemical class and structure annotation of isomers belonging to known compounds from which CCS information has not been previously measured (i.e. an “unknown unknown” isomer), as is the case for the majority of human metabolites which are expected to be isomeric but currently undiscovered.[59]

3.4 Conclusion

In this work, we illustrate the utility of IM-MS in quantitatively characterizing biochemical species using a Unified CCS compendium. Prior to this work, quantitative CCS libraries have been limited in scope to a narrow range of chemical classes, polarities, and adduct types. Therefore, we curated a Unified CCS compendium obtained from chemical standards representing a wide variety of structures spanning 14 super classes, 80 classes, and 157 subclasses. We anticipate subsequent contributions from the IM-MS community; and therefore, the informatics infrastructure developed was designed to accommodate future expansion. The current biochemical species contained within the Unified CCS compendium enabled generation of optimized nonlinear regression models with CI and PI for 48 classes and subclasses. These models enabled filtering and prediction of unknown biochemical species. The capabilities demonstrated in this manuscript establish a foundation for utilizing CCS/z as an additional molecular characterization dimension. The Unified CCS compendium was used to predict and identify unknown chemical species that originated from a serum sample. Future work will focus on expanding the number of entries in the compendium to improve predictive power.

We aim for the Unified CCS compendium to be a collaborative effort of the IM-MS community and invite contributions to this open-access repository for quality-controlled CCS measurements. Specific guidelines for submitting data are found in the ESI(SectionS2). While the compendium is initially designed to only include DTIMS data, considerations for adding CCS information obtained from other IM techniques will be included in future iterations. The standardized DTIMS CCS measurements contained within the compendium can serve as calibrant reference values for other IM techniques, which will ultimately enable the incorporation of more CCS data into this body of work.

3.5 Acknowledgments

This dissertation chapter was adapted from a published manuscript titled "Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities" by Jaqueline A. Picache, Bailey S. Rose, Andrzej Balinski, Katrina L. Leaptrot, Stacy D. Sherrod, Jody C. May, and John A. McLean published in *Chemical Science* **2019**, *10* (4), 983–993.

I would like to acknowledge Erin S. Baker and colleagues at the Pacific Northwest National Laboratory, as well as James N. Dodds, Caleb B. Morris, and Charles M. Nichols at Vanderbilt University for their efforts in acquiring data within the compendium; and Timo Sachsenberg at the University of Tübingen for his guidance in methods for large scale drift time extraction analyses. Additionally, the authors would like to acknowledge John C. Fjeldsted of Agilent Technologies for his collaborative efforts and expertise. Financial support for this research was provided by the National Institutes of Health (NIH NIGMS R01GM092218 and NIH NCI 1R03CA222452-01) and the NIH supported Vanderbilt Chemical Biology Interface training program (5T32GM065086-16). This work was supported in part using the resources of the Center for Innovative Technology (CIT) at Vanderbilt University.

3.6 References

(1) D. Houle, D. R. Govindaraju and S. Omholt, *Nat. Rev. Genet*., **2010**, 11, 855–866.

(2) J. C. May, R. L. Gant-Branum and J. A. McLean*, Curr. Opin. Biotechnol*., **2016**, 39, 192–197.

(3) J. C. May and J. A. McLean, *Annu. Rev. Anal. Chem*., **2016**, 9, 387–409.

(4) S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nat. Rev. Drug Discovery*, **2010**, 9, 203–214.

(5) R. A. Quinn, J. A. Navas-molina, E. R. Hyde, J. Song, Y. V´azquez-baeza, G. Humphrey, J. Gaffney, J. J. Minich, A. V. Melnik, J. Herschend, J. Dereus, A. Durant, R. J. Dutton, M. Khosroheidari and C. Green, *mSystems*, **2016**, 1, e00038-1–6.

(6) R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O'Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. J. Butte, E. A. Ashley, M. Gerstein, K. C. Nadeau, H. Tang and M. Snyder, *Cell*, **2012**, 148, 1293–1307.

(7) J. S. D. Zimmer, M. E. Monroe, W. J. Qian and R. D. Smith, *Mass Spectrom. Rev*., **2006**, 25, 450–482.

(8) X. Zheng, R. Wojcik, X. Zhang, Y. M. Ibrahim, K. E. Burnum- Johnson, D. J. Orton, M. E. Monroe, R. J. Moore, R. D. Smith and E. S. Baker, *Annu. Rev. Anal. Chem*., **2017**, 10,71–92.

(9) J. A. McLean, B. T. Ruotolo, K. J. Gillig and D. H. Russell*, Int. J. Mass Spectrom*., **2005**, 240, 301–315.

(10) K. M. Hines, J. C. May, J. A. McLean and L. Xu, *Anal. Chem*., **2016**, 88, 7329–7336.

(11) W. B. Ridenour, M. Kliman, J. A. McLean and R. M. Caprioli, *Anal. Chem*., **2010**, 82, 1881–1889.

(12) S. M. Stow, T. J. Causon, X. Zheng, R. T. Kurulugama, T. Mairinger, J. C. May, E. E. Rennie, E. S. Baker, R. D. Smith, J. A. McLean, S. Hann and J. C. Fjeldsted, *Anal. Chem*., **2017**, 89, 9048–9055.

(13) C. B. Lietz, Q. Yu and L. Li, *J. Am. Soc. Mass Spectrom*., **2014**, 25, 2009–2019.

(14) W. B. Struwe, K. Pagel, J. L. P. Benesch, D. J. Harvey and M. P. Campbell, *Glycoconj. J*., **2016**, 33, 399–404.

(15) K. M. Hines, D. H. Ross, K. L. Davidson, M. F. Bush and L. Xu, *Anal. Chem*., **2017**, 89, 9023–9030.

(16) Z. Zhou, J. Tu, X. Xiong, X. Shen and Z. J. Zhu, *Anal. Chem*., **2017**, 89, 9559–9566.

(17) M. Hern´andez-Mesa, B. Le Bizec, F. Monteau, A. M. Garc´ıa- Campa˜na and G. Dervilly-Pinel, *Anal. Chem*., **2018**, 90, 4616–4625.

(18) X. Zheng, N. Aly, Y. Zhou, K. Dupuis, A. Bilbao, V. Paurus, D. J. Orton, R. Wilson, S. Payne, R. D. Smith and E. S. Baker, *Chem. Sci*., **2017**, 8, 7724–7736.

(19) Z. Zhou, X. Shen, J. Tu and Z.-J. Zhu, *Anal. Chem*., **2016**, 88, 11084–11091.

(20) G. Paglia, J. P. Williams, L. Menikarachchi, J. W. Thompson, R. Tyldesley-Worster, S. Halld´orsson, O. Rolfsson, A. Moseley, D. Grant, J. Langridge, B. O. Palsson and G. Astarita, *Anal. Chem*., **2014**, 86, 3985–3993.

(21) L. Righetti, A. Bergmann, G. Galaverna, O. Rolfsson, G. Paglia and C. Dall'Asta, *Anal. Chim. Acta*, **2018**, 1014, 50–57.

(22) C. R. Goodwin, L. S. Fenn, D. K. Derewacz, B. O. Bachmann and J. A. McLean, *J. Nat. Prod.*, **2012**, 75,48–53.

(23) R. Lian, F. Zhang, Y. Zhang, Z. Wu, H. Ye, C. Ni, X. Lv and Y. Guo, *Anal. Methods*, **2018**, 10, 749–756.

(24) M. Chai, M. N. Young, F. C. Liu and C. Bleiholder, *Anal. Chem*., **2018**, 90, 9040–9047.

(25) V. Gabelica and E. Marklund, *Curr. Opin. Chem. Biol*., **2018**, 42,51–59.

(26) I. Blazenovic, T. Kind, J. Ji and O. Fiehn, *Metabolites*, **2018**, 8,31.

(27) J. Ma, C. P. Casey, X. Zheng, Y. M. Ibrahim, C. S. Wilkins, R. S. Renslow, D. G. Thomas, S. H. Payne, M. E. Monroe, R. D. Smith, J. G. Teeguarden, E. S. Baker and T. O. Metz, *Bioinformatics*, **2017**, 33, 2715–2722.

(28) B. X. Maclean, B. S. Pratt, J. D. Egertson, M. J. Maccoss, R. D. Smith and E. S. Baker, *J. Am. Soc. Mass Spectrom.*, **2018**, DOI: 10.1007/s13361-018-2028-5.

(29) B. Pratt, M. Horowitz-gelb, J. W. Thompson, E. Baker, J. W. Thompson, M. J. Maccoss and B. Maclean, in 65th Annual Conference for the American Society of Mass Spectrometry, American Society for Mass Spectrometry, Indianapolis, IN, 2017.

(30) B. D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, UK, **1996**.

(31) S. M. Colby, D. G. Thomas, J. R. Nunez, D. J. Baxter, K. R. Glaesemann, M. Brown, M. A. Pirrung, N. Govind, J. G. Teeguarden, T. O. Metz and S. Ryan, arXiv:1809.08378 [q-bio.BM].

(32) Mclean Research Group, CCS compendium, https:// lab.vanderbilt.edu/mclean-group/collision-cross-section- database/.

(33) C. M. Nichols, J. C. May, S. D. Sherrod and J. A. McLean, *Analyst*, **2018**, 143, 1556–1559.

(34) J. C. May, C. R. Goodwin, N. M. Lareau, K. L. Leaptrot, C. B. Morris, R. T. Kurulugama, A. Mordehai, C. Klein, W. Barry, E. Darland, G. Overney, K. Imatani, G. C. Stafford, J. C. Fjeldsted and J. A. McLean, *Anal. Chem*., **2014**, 86, 2107–2116.

(35) J. N. Dodds, J. C. May and J. A. Mclean, *Anal. Chem*., **2017**, 89, 952–959.

(36) J. C. May, E. Jurneczko, S. M. Stow, I. Kratochvil, S. Kalkhof and J. A. McLean, *Int. J. Mass Spectrom.*, **2017**, 427,79–90.

(37) K. L. Leaptrot, J. C. May, J. N. Dodds. J. A. McLean and *Nat. Commun*., **2019** *10*, 985.

(38) C. M. Nichols, J. N. Dodds, B. S. Rose, J. A. Picache, C. B. Morris, S. G. Codreanu, J. C. May, S. D. Sherrod and J. A. McLean, *Anal. Chem*., **2018**, DOI: 10.1021/ acs.analchem.8b04322.

(39) E. A. Mason and E. W. McDaniel, Transport Properties of Ions in Gases, John Wiley & Sons, Ltd., New York City, NY, **1988**.

(40) W. F. Siems, L. A. Viehland and H. H. Hill, *Anal. Chem*., **2012**, 84, 9782–9791.

(41) R. Core Team, A language and environment for statistical computing. R Foundation for Statistical Computing, https:// www.r-project.org/.

(42) V. Gabelica, C. Alfonso, P. E. Barran, J. L. P. Benesch, C. Bleiholder, M. T. Bowers, et al., *ChemRxiv*, **2018**, DOI: 10.26434/chemrxiv.7072070.v2.

(43) Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D. S. Wishart, *J. Cheminf*., **2016**, 8,1–20.

(44) H. J. Feldman, M. Dumontier, S. Ling, N. Haider and C. W. V. Hogue, *FEBS Lett*., **2005**, 579, 4685–4691.

(45) McLean Research Group Github, https://github.com/ McLeanResearchGroup.

(46) C. B. Morris, J. C. May and J. A. McLean, in 62th Annual Conference for the American Society of Mass Spectrometry, Baltimore, MD, **2014**.

(47) J. C. May, C. B. Morris and J. A. McLean, *Anal. Chem*., **2017**, 89, 1032–1044.

(48) K. P. Burnham and D. R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer-Verlag, New York City, NY, 2nd edn, **2002**, vol. 172.

(49) A. N. Spiess and N. Neumeyer, *BMC Pharmacol*., **2010**, 10,1– 11.

(50) C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec and P. Despouy, Create Interactive Web Graphics via 'plotly.js', https://cran.r-project.org/ package plotly.

(51) H. Wickham, ggplot2: Elegant Graphics for Data Analysis, http://ggplot2.org.

(52) M. Dowle and A. Srinivasan, data.table: Extension of 'data.frame', https://cran.r-project.org/package/data.table.

(53) H. Wickham, J. *Stat. Software*, **2011**, 40,1–29.

(54) W. Chang, J. Cheng, J. Allaire, Y. Xie and J. McPherson, shiny: Web Application Framework for R, https://cran.r-project.org/ package/shiny.

(55) C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan and G. Siuzdak, *Ther. Drug Monit*., **2005**, 27, 747–751.

(56) T. Kind, K. H. Liu, D. Y. Lee, B. Defelice, J. K. Meissen and O. Fiehn, *Nat. Methods*, **2013**, 10, 755–758.

(57) M. C. Chambers, B. MacLean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb and P. Mallick, *Nat. Biotechnol*., **2012**, 30, 918–920.

(58) J. J. Faraway, Practical Regression and Anova using R, 3rd edn, **2002**.

(59) D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. V´azquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach and A. Scalbert, *Nucleic Acids Res*., **2018**, 46, D608–D617.

(60) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A., "Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities" *Chem. Sci*. **2019**, *10* (4), 983-993.

CHAPTER IV

Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility-Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization

4.1 Introduction

Mass spectrometry (MS) is used in multiple omics disciplines to perform global untargeted experiments.[1,2] Identification of biomolecular compounds from the exact mass measurement and tandem ion fragmentation (e.g., MS/MS) experiments is then supported through database searching,[3] and often other analytical dimensions, such as liquid chromatography (LC) and, more recently, ion mobility (IM), are included to increase the measurement specificity.[4,5] However, even with a high number of analytical descriptors, some of the detected features are not matched with a confident identification. For true "unknown unknowns" where database entries do not exist, it is difficult to begin identifying these compounds.[3] In these cases, theoretical prediction of the compound identity or attributes is promising, particularly where authentic chemical standards are prohibitive.[6-8]

Theoretical prediction of chemical classifications is especially promising because it can be performed quickly and provide investigators with a global inference for their particular samples. Previous work in chemical classification prediction is limited, with most studies using chemical classification to predict toxicity or metabolic function.[9-11] A notable effort by Gitter and colleagues has been to use chemical structure to predict drug function.[12] Another recent report from Xu and colleagues utilizes chemical structures to predict theoretical collision cross sections of molecules.[13]

These efforts relate to the work herein because a chemical structure dictionary and empirical data are used to predict chemical classifications. To the best of our knowledge, no informatic tools currently exist to predict chemical classifications, as defined by Wishart and colleagues,[14] using only measured data.

Before class prediction from empirical data can be performed, a systematic chemical classification system must be defined and accepted by the field at large. One classification system that is widely utilized is ClassyFire, which parses the primary chemical structure of a molecule as input for defining its taxonomy. ClassyFire designates each molecule with a chemical kingdom, super class, and class; and additionally, some molecules are assigned a subclass.[14] These classifications operate in a hierarchy that increases in specificity. A kingdom classification designates if a molecule is organic or inorganic. The super class is a general molecular type such as "lipids and lipid-like molecules." The class and subclass give more specific categorical information such as the "glycerophospholipids" class and the "glycerophosphoethanolamines" subclass. A more detailed discussion about this chemical classification system can be found elsewhere.[14] This hierarchical classification system provides a framework for assigning unknown features with a chemical identity at various levels of specificity.

The IM-MS community has a precedence for analyzing molecules from different chemical classes.[15] These analyses indicate that the conformational space (collision cross section vs—. m/z) occupied by different classes is specific. Clemmer and colleagues first analyzed the conformational space of plasma proteins.[16,17] As IM-MS technology advanced, smaller molecules like glycans and lipids could be described via conformational space.[18,19] Presently, multiple omic classes describing small and large molecules have been mapped via conformational space.[5,20] These conformational plots have enabled improved identification of molecules in untargeted studies.[5] In addition to the

cross section and m/z, Kendrick mass defect (KMD) has been shown to correlate with chemical structural similarities. Two notable examples are that of Lebrilla and colleagues, in which lipid classes were identified via KMD, and work from De Pauw and colleagues, which utilized KMD to identify similar chemicals in imaging mass spectrometry experiments.[21,22]

This work presents a unique machine learning approach to support unknown compound identification referred to as the supervised inference of feature taxonomy from ensemble randomization (SIFTER) algorithm. SIFTER uses a random forest (RF) machine learning algorithm that is trained against highly accurate drift-tube ion mobility-mass spectrometry measurements to predict the chemical classification of features not assigned a chemical identity through traditional database search workflows. Whereas the task of chemical identifications can be supported through matching multiple chemical annotations (e.g., retention time, collision cross section, exact mass, etc.), here, only analytical information readily accessible from IM-MS is considered.

4.2 Experimental Methods

4.2.1 Data Sources The training set data was obtained from the Unified Collision Cross Section Compendium (CCS Compendium) described previously.[5] Briefly, each data entry within the training set contains a set of characteristics about a chemical compound including its taxonomic chemical classification. This taxonomy includes kingdom, super class, class, and subclass designations, as previously formalized.[14] It is important to note that the CCS Compendium currently does not contain data for a majority of chemical classifications, with current data entries representing 2/2 possible kingdoms, 14/31 super classes, 89/764 classes, and 158/1729 subclasses

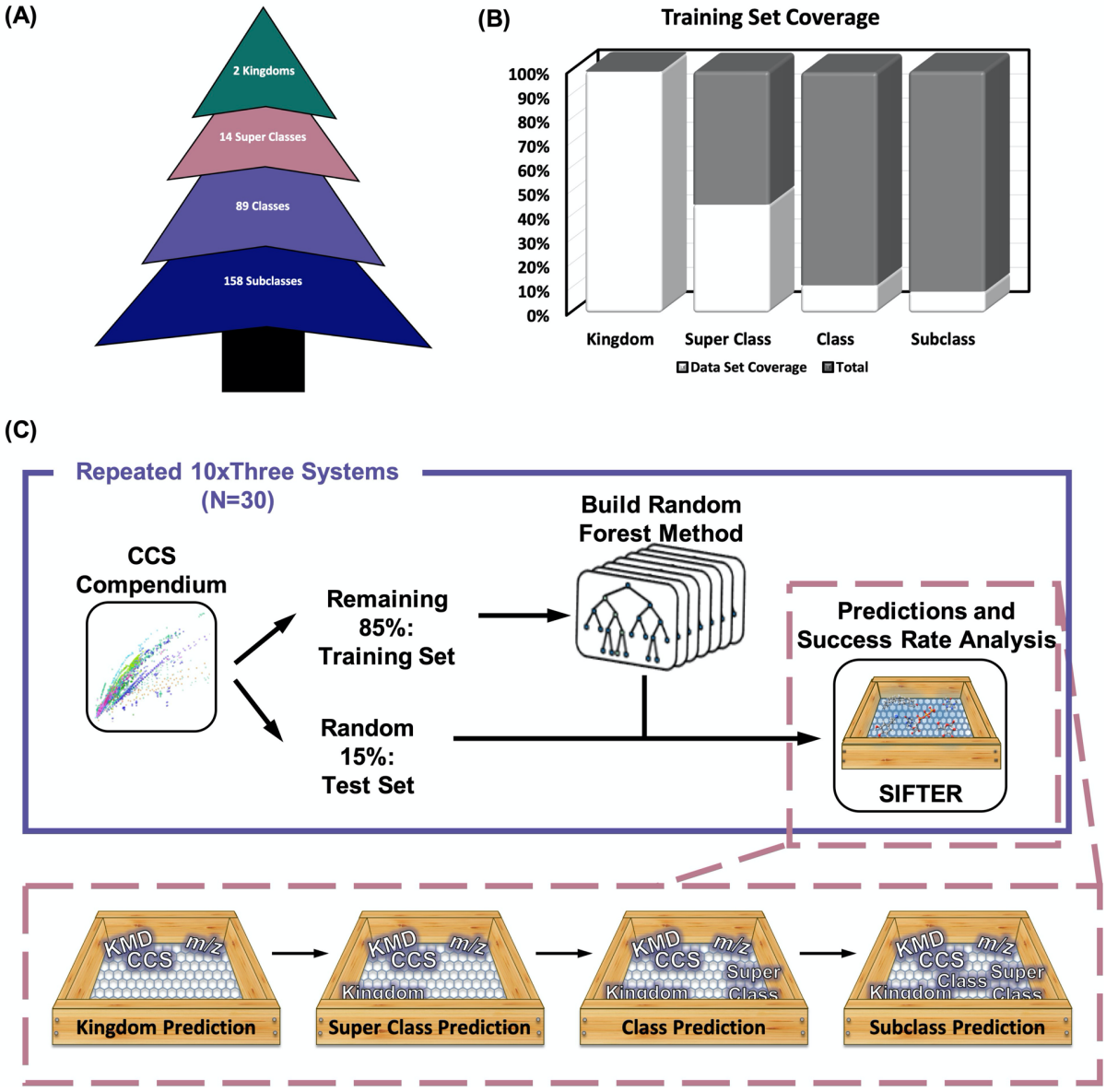


Figure 4.1 SIFTER Algorithm (A) Training set coverage numbers per classification category. (B) Comparison of training set coverage per classification category (white) to the total reported classifications (gray). (C) Overall schematic of SIFTER random forest machine learning workflow. (Figure adapted from Picache, et al.[30])

as illustrated in **Figure 4.1A** and **4.1B**. The work herein will improve as more data and more complete coverage is added to the CCS Compendium. The biological test set was obtained from previous studies in which metabolites were putatively identified within red wine and chestnut matrixes, respectively.[23,24]

4.2.2 Random Forest Machine Learning Algorithm SIFTER classifies these unknown features into a taxonomy of chemical kingdom, super class, class, and sub- class using an RF algorithm and applying the ClassyFire Chemical Class Dictionary.[14] The RF algorithm uses the open-source package randomForest (v4.6—-−14) developed for the statistical computing programming environment R (R Foundation for Statistical Computing, Vienna, Austria).[25,26] Briefly, RF algorithms work by creating multiple decision trees and a majority vote score determines the final classification prediction.[26,27] This process is illustrated in Appendix D, **Figure D4.1**. Prior to training of the algorithm, data entries from the CCS Compendium were parsed to remove entries with limited (<10 data points) or no subclass information. This data omission was necessary to avoid boundary errors during training. The remaining data (N = 2365 data points) were randomly divided into an 85% training set (N = 2011) and a 15% test set (354 data points) as illustrated in **Figure 4.1C**. From there, four RF algorithms were generated. These algorithms were generated using both positive and negative mode data as well as all ion species, including multiple charge state types within the training set (+H, −H, +Na, +K, etc.). By including as much data as possible, the algorithm can parse as much chemical space as possible, though it is noted that adduct-specific training sets should further improve the prediction accuracy. Additionally, the training set was comprised of data generated on a time-of-flight instrument with a mass accuracy of 5 ppm error and mass resolving of ~20000. The following work should improve with higher mass accuracy

and resolution but is not advised for instruments that deviate significantly below these mass measurement metrics. The first algorithm predicts the chemical kingdom of a molecule using its m/z, CCS, and Kendrick mass defect (KMD) with the Kendrick mass scaled to $CH_2 = 14.0\overline{0}$ Da.[28] Optimized parameters include ntree = 500, mtry = 3, and nodesize = 1. The next RF algorithm predicts the chemical super class of a molecule using its m/z, CCS, KMD, and predicted kingdom. Optimized parameters include ntree = 100, mtry = 4, and nodesize = 1. The RF third algorithm predicts the chemical class of a molecule using its m/z, CCS, KMD, predicted kingdom, and predicted super class. Optimized parameters include ntree = 100, mtry = 4.5, and nodesize = 2. The final RF algorithm predicts the chemical subclass of a molecule using its m/z, CCS, KMD, predicted kingdom, predicted super class, and predicted class. Optimized parameters include ntree = 1000, mtry = 5, and nodesize = 2. This process is illustrated in **Figure 4.1C**. Source code for all of the algorithms can be found on the McLean Research Group Github.[29]

4.2.3 SIFTER Algorithm Performance and Outcomes Performance metrics and algorithm optimization were based primarily on minimizing the out-of-bag (OOB) errors for each of the four RF algorithms. The OOB method provides an estimate of the prediction error of random forests, which represents the mean prediction error for each individual classification using trees that did not contain the bootstrap, or initial, sampling process. Additionally, the confusion matrices and receiver operator characteristic (ROC) curves were analyzed. After optimization, the OOBs for each RF were 0.68 ± 0.05% for kingdom prediction, 22.82 ± 0.29% for super class prediction, 10.24 ± 0.23% for class prediction, and 15.27 ± 0.16% for subclass prediction across 30 trials. Whereas the OOB errors provide a general error rate for each algorithm, the error rate at predicting each individual potential classification per category is provided via confusion matrices. A

summary of the confusion matrices results is found in **Appendix D, Tables D4.1 – D4.4**. The ROC curves, shown in **Figure 4.2A**, provide more information about the performance of each RF algorithm. SIFTER generally has good sensitivity and specificity for predicting the kingdom and super class. There is decreased sensitivity in class prediction which for our purposes, represents accuracy. This indicates a potential nearest -neighbor bias in the class prediction. There is decreased specificity in subclass prediction which for our purposes, represents precision. This indicates that majority vote scores in the decision tree process could be similar for multiple potential predictions.

The potential outcomes of SIFTER are illustrated in **Figure 4.2B**. Briefly, SIFTER can have one of five outcomes. The first is if SIFTER fails to predict any of the chemical classifications of a molecule (0/4). Next, SIFTER can correctly predict the kingdom of a molecule but incorrectly predict all other classifications (1/4). The third potential outcome is if SIFTER correctly predicts the kingdom and super class but incorrectly predicts the class and subclass of a molecule (2/4). The fourth outcome is if SIFTER correctly predicts all of the classification categories except for the subclass (3/4). Finally, the last outcome is if SIFTER correctly predicts all classifications (4/4). Understanding of these classification outcomes is important because once SIFTER makes an incorrect classification, it cannot make a correct classification in that particular taxonomic tree for an individual molecule. This is due to the fact that SIFTER uses prior predictions to predict the next classification category. SIFTER was coded such that each of these classifications has a calculated probability as well as alternate classifications and their probabilities.

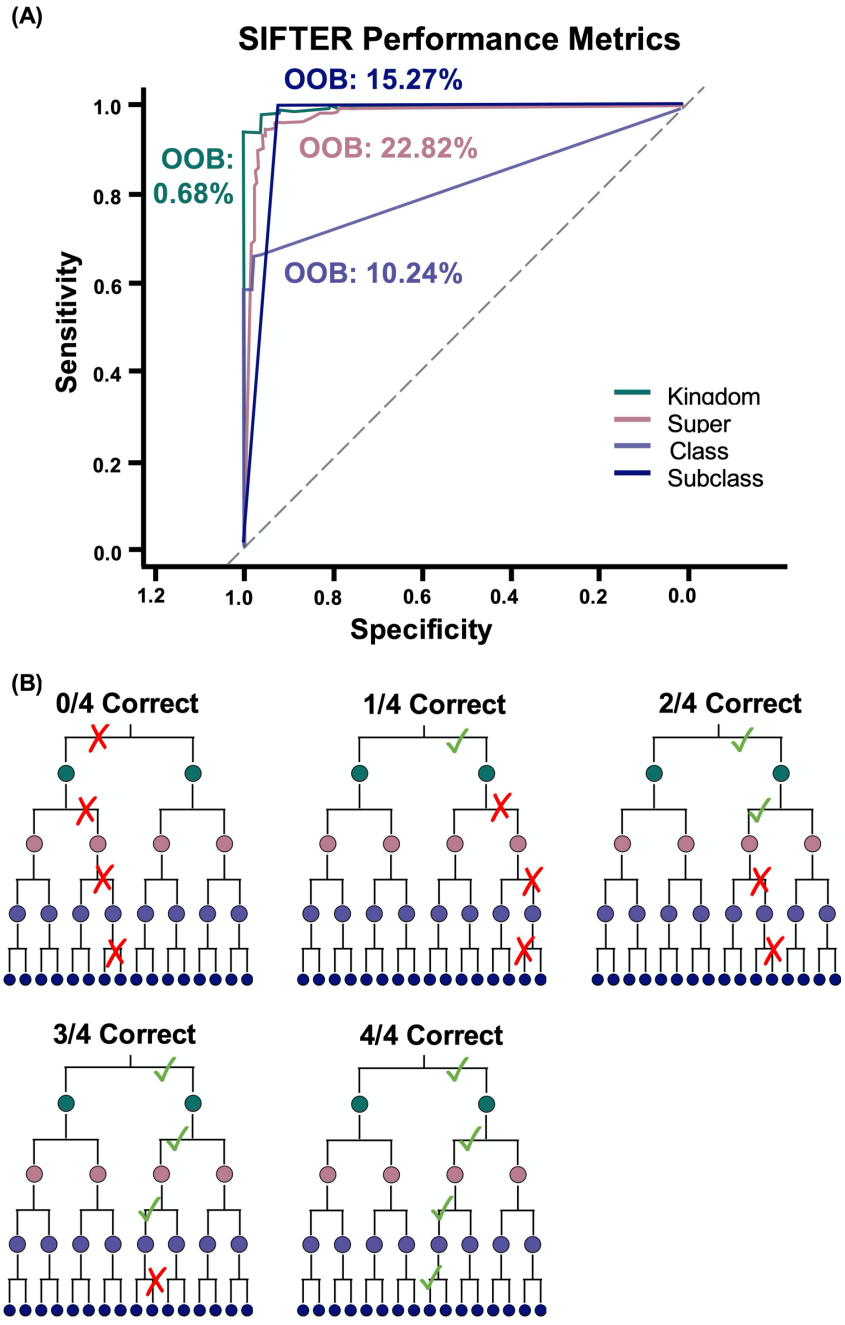


Figure 4.7 SIFTER Performance and Outcomes (A) Receiver operating characteristic curves depicting the sensitivity and specificity of SIFTER per classification category. Associated Out-of-bag (OOB) errors and also reported. (B) Potential outcomes of SIFTER where none to all of the classifications are correct. (Figure adapted from Picache, et al.[30])

4.3 Results and Discussion

4.3.1 SIFTER Test Set Performance After optimization, SIFTER was tested against a set of molecules not used in the training set. As previously described, the CCS Compendium was randomly partitioned into an 85% training set and a 15% test set. The test set of 354 compounds was randomized over 30 trials. The average success rate in classifying all four categories correctly was >99% or >353 compounds. The average number of times SIFTER predicted 3/4, 2/4, and 1/4 cases correctly is 0.17, 0.33, and 0.03 compounds, respectively, across the 30 trials. No prediction was completely wrong (0/4). The breakdown of classification success per category is shown in **Figure 4.3A**. The kingdom of a compound was correctly predicted 100% of the time. Super class was correctly predicted in 99.99% of cases. Class was correctly predicted in 99.91% of cases. Subclass was correctly predicted in 99.86% of cases.

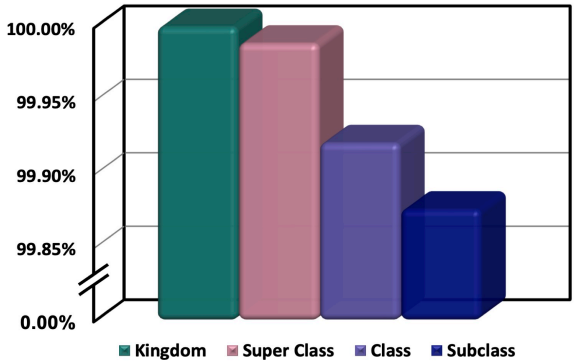
A specific test case in which SIFTER worked well is shown in **Figure 4.3B** for the compound 6-keto-prostaglandin F1α. Empirical data consisting of *m/z* 393.225, CCS 202.9 $Å^2$, and KMD −0.786 were input into SIFTER. The predicted classifications were "Organic compounds" for kingdom, "Lipids and lipid-like molecules" for super class, "Fatty acyls" for class, and "Eicosanoids" for subclass. These predictions aligned with the correct predictions (green column, **Figure 4.3B**). Furthermore, SIFTER reports the probability scores for each of its categorical predictions, which were greater than 0.9 for all categories except "class" (0.65), which suggests that the algorithm had the most difficulty with this assignment. SIFTER also outputs alternative predictions and their associated probabilities, and for the "class" category, the next predicted classification ("Steroids and steroid derivatives") was given a probability score of only 0.35, providing more confidence in SIFTER's first prediction for this category. Analysis of all the results across the 30 trials indicate that a probability of ≥70% is reliably correct; however, some

exceptions do occur. Preliminary results for amphotericin B (924 Da) also suggest that SIFTER performs well with larger (>500 Da) molecules (see Appendix D, Figure D4.2).

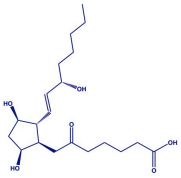
A case example in which SIFTER did not perform as well is given in **Figure 4.3C** for the compound carnosine. The input for this compound was m/z 227.114, CCS 150.8 Å^2, and KMD—-−0.861. The predicted categories were as follows: “Organic compounds” for kingdom, “Organic acids and derivatives” for super class, “Carboxylic acids and derivatives” for class, and “Amino acids, peptides, and analogues” for subclass. When compared to the correct predictions (green column, **Figure 4.3C**), SIFTER predicted 2/4 categories correctly, but provided a probability of only 0.50 for the class prediction. Moreover, the alternative class prediction (“Peptidomimetics”) was also 0.50—, and is the correct assignment. Thus, without knowing the correct classifications, it would be ill advised to accept this class and subclass prediction. Because the subclass cannot be correct if the class is predicted incorrectly, one must assume that the subclass prediction is also incorrect, despite the high probability of 99%. Nevertheless, it is clear that carnosine shares many characteristics in common with the class and subclass assigned. It is clearly anticipated that the predictive power will increase dramatically with concomitant expansion of the training set used and increased representation of each class and subclass.

4.3.2 SIFTER Complex Sample Performance Analysis of molecules within biological matrixes, namely, red wine and chestnut extracts, not used in the training set was found to have a lower success rate where all four categories were correctly predicted for 81% (81 of 100) of the compounds. **Figure 4.4A** illustrates the performance of SIFTER for 100 compounds found within these complex biological matrixes.

(A) **Classification Success per Category**



(B)

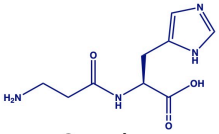


6-Keto-prostaglandin F1α

Input: *m/z* : 393.225 Da CCS: 202.9 Å² KMD: -0.786

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	
Super Class	Lipids and lipid-like molecules	Lipids and lipid-like molecules ✓	0.93	Organic acids and derivatives (0.06) Benzenoids (0.04) Other (0.01)
Class	Fatty acyls	Fatty acyls ✓	0.65	Steroids and steroid derivatives (0.34)
Subclass	Eicosanoids	Eicosanoids ✓	0.99	Fatty acids and conjugates (0.01)

(C)



Carnosine

Input: *m/z* : 227.115 Da CCS: 150.8 Å² KMD: -0.861

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	
Super Class	Organic acids and derivatives	Organic acids and derivatives ✓	0.66	Organoheterocyclics (0.17) Lipids and lipid-like molecules (0.12) Organicoxygen compounds (0.05)
Class	Peptidomimetics	Carboxylic acids and derivatives ✗	0.50	Peptidomimetics (0.50)
Subclass	Hybrid peptides	Amino acids, peptides, and analogues ✗	0.99	Tricarboxylic acids and derivatives (0.01)

Figure 4.8 Test Set Performance (A) Classification success per category in test set; n=30. (B) Example where SIFTER classified the molecule 6-Keto-prostaglandin F1α where 4/4 categories were correct and their associated probabilities are shown. (C) Example where SIFTER classified the molecule Carnosine where 2/4 categories correct and their associated probabilities are shown. For the carnosine example, it is important to note SIFTER was 50/50 between the correct and incorrect class categories. (Figure adapted from Picache, et al.[30])

These data were selected for evaluation of SIFTER because they were collected using procedures that align with those used in an international laboratory comparison,[20] which was used as the basis for data acceptance into the Unified CCS Compendium.[5] Thus, the data were independently collected under conditions similar to those used in the training sets, but by an independent laboratory not having deposited the data for inclusion into the training set.

In 81 cases, 4/4 classifications were correct. The number of 3/4, 2/4, and 1/4 correct cases is 1, 5, and 13, respectively. The classification success per category is as follows: 100% for kingdom, 87% for super class, 82% for class, and 81% for subclass. The decline in success between the kingdom and super class classifications is expected and aligns with the OOBs errors shown in **Figure 4.2A**. Additionally, the compounds in the complex biological sample had a false discovery rate (FDR) associated with them. These FDRs are based on returned SIFTER predictions where no prediction should have been returned. For example, in the class category, a few of the compounds had classifications that are not yet in the CCS Compendium training set. Therefore, SIFTER could not have correctly predicted their class. Another example is that some compounds did not have a subclass assigned to them based on ClassyFire, so there should not have been an assigned subclass. One of the limitations of SIFTER is that it cannot handle this null situation and will always provide a prediction whether a prediction is warranted or not. However, it is not possible to include null categorizations in RF algorithms because of boundary limitations within the algorithm design. Other algorithms such as neural networks can include null categorizations, but in preliminary work, the authors found that these other algorithms generally performed worse than RF when predicting chemical classifications. Within the compounds derived from the complex biological system, 21 classifications were false discoveries. Eleven of the 100 class predictions were false because those classes were not found within the CCS Compendium training

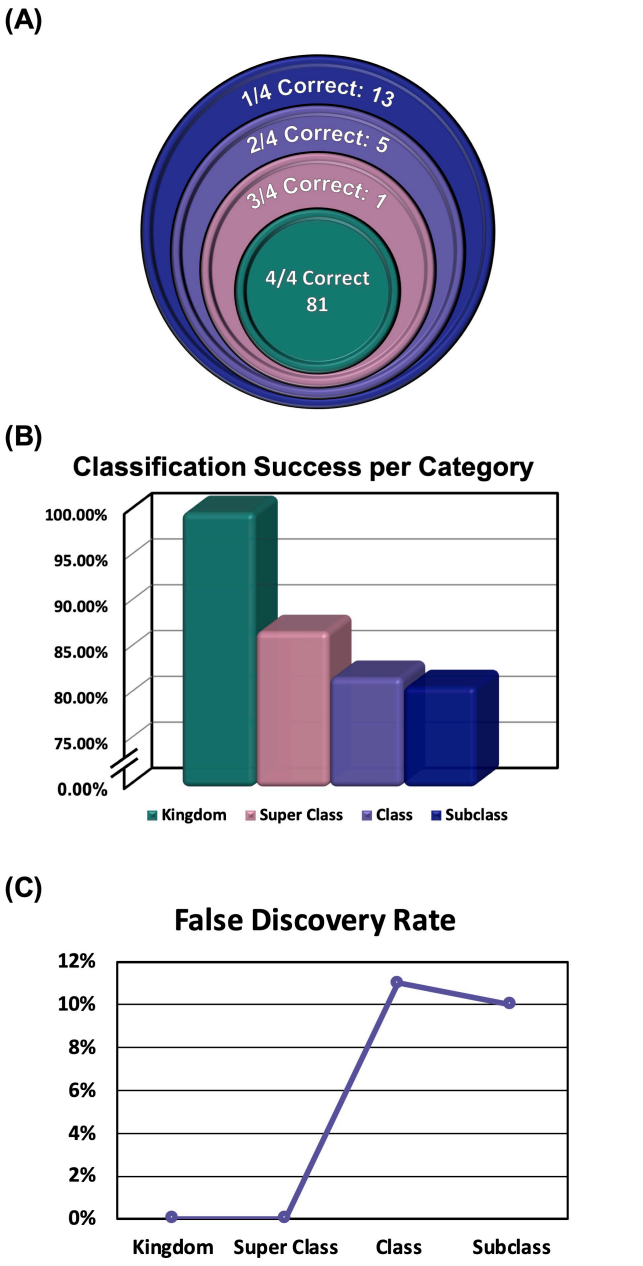


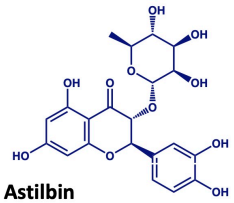
Figure 4.9 SIFTER Performance of Complex Samples Summary (A) Number of correct classifications in complex sample test set of 100 compounds. (B) Classification success per category in complex test set out of 100 compounds. (C) False discovery of SIFTER per category where a prediction occurred when it should not have been predicted. (Figure adapted from Picache, et al.[30])

set. Furthermore, 10 of the 100 subclass predictions were false because those compounds either did not have an as- signed subclass or the subclass in question was not within the CCS Compendium training set. The FDRs per category are summarized in **Figure 4.4C**.

4.3.3 SIFTER Case Studies from Complex Samples A few case studies will be discussed in detail to understand how SIFTER performed when classifying compound features derived from complex biological samples described above. The first case study where SIFTER performed well is with the compound astilbin found within a red wine matrix. Astilbin was identified using LC-IM-MS/MS as previously described.[23] **Figure 4.5A** shows the input into SIFTER was *m/z* 449.109, CCS 196.6 Å2, and KMD −0.607. The predicted categories were as follows: "Organic compounds" for kingdom, "Phenylpropanoids and polyketides" for super class, "Flavonoids" for class, and "Flavonoid glycosides" for subclass. 4/4 of these predictions were correct when compared to the actual classifications (green column, **Figure 4.5A**). This situation is a "best case scenario" where all of the associated probabilities are high.

The next case study is that of cis-piceid found in a chestnut matrix. Identification of this compound was performed using LC-IM-MS/MS as previously described.[24] **Figure 4.5B** shows the SIFTER input of *m/z* 389.124, CCS 190.8 Å2, and—a KMD—-−0.689. The predicted categories were "Organic compounds" for kingdom, "Phenyl propanoids and polketides" for super class, "Flavonoids" for class, and "O-methylated flavonoids" for subclass. When compared to the correct predictions (green column, **Figure 4.5B**), SIFTER had a 2/4 outcome. This case is one of the false positive cases because the CCS Compendium training set did not have any "Stilbene" classes or "Stilbene glycoside" subclasses within it. Although the predicted super class was correct, it had a

(A)



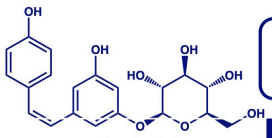
Astilbin

Input:

m/z : 449.109 Da CCS: 196.6 Å² KMD: -0.607

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	
Super Class	Phenylpropanoids and polyketides	Phenylpropanoids and polyketides ✓	0.93	Organic acids and derivatives (0.03) Organoheterocyclics (0.02) Other (0.02)
Class	Flavonoids	Flavonoids ✓	0.98	Tannins (0.02)
Subclass	Flavonoid glycosides	Flavonoid glycosides ✓	0.99	Hydrolyzed tannins(0.01)

(B)



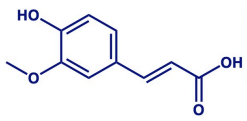
cis-Piceid

Input:

m/z : 389.124 Da CCS: 190.8 Å² KMD: -0.689

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	
Super Class	Phenylpropanoids and polyketides	Phenylpropanoids and polyketides ✓	0.55	Organic acids and derivatives (0.27) Organic oxygen compounds (0.05) Organoheterocyclics (0.05)
Class	Stilbene	Flavonoids ✗	0.98	Tannins (0.02)
Subclass	Stilbene glycosides	O-methylated flavonoids ✗	0.79	Flavones (0.20) Other (0.01)

(C)



Ferulic acid

Input:

m/z : 193.051 Da CCS: 139.7 Å² KMD: -0.835

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	
Super Class	Phenylpropanoids and polyketides	Benzenoids ✗	0.32	Organoheterocyclics (0.24) Organic oxygen compounds (0.23) Organic acids and derivatives (0.16)
Class	Cinnamic acids and derivatives	Benzene and substituted derivatives ✗	0.82	Phenols (0.18)
Subclass	Hydroxycinnamic acids and derivatives	Benzoic acids and derivatives ✗	0.83	Bezenesulfonamids (0.17)

Figure 4.10 Compounds in Complex Sample Case Studies (A) Example where SIFTER classified the molecule Astilbin where 4/4 categories were correct and their associated probabilities are shown. (B) Example where SIFTER classified the molecule *cis*-Piceid where 2/4 categories were correct and their associated probabilities are shown (C) Example where SIFTER classified the molecule Ferulic Acid where 1/4 categories were correct and their associated probabilities are shown. (Figure adapted from Picache, et al.[30])

low probability of 55%, which indicates that the super class, and therefore class and subclass, should be rejected or accepted with caution.

The last case study is one where SIFTER did not perform well. The compound was ferulic acid found in a chestnut matrix. Ferulic acid was previously identified using LC-IM-MS/MS.[24] The input, shown in **Figure 4.5C**, for ferulic acid was *m/z* 193.051, CCS 139.7 Å2, and KMD—-−0.835. The predicted classifications were "Organic compounds" for kingdom, "Benzenoids" for super class, "Benzenes and substituted derivatives" for class, and "Benzoic acids and derivatives" for subclass. When these results were compared to the correct classifications (green column **Figure 4.5C**), this resulted in a 1/4 outcome. The RF algorithm has a known nearest -neighbor bias, which means that it will tend to select a classification if there are more of it in the training set over a different classification. There are more "Benzene and substituted derivatives" (N = 187) than the correct super class "Phenylpropanoid and polyketides" (N = 101) within the CCS Compendium training set. The super class prediction probability was low at 32%, with the next alternative predictions ("Organoheterocyclics" and "Organic oxygen compounds") having only marginally lower probabilities (24% and 23%), which is indicative of a false classification.

4.4 Conclusion

In this work, we present a machine learning algorithm, SIFTER, that can be used to predict the chemical classifications of unknown molecules from untargeted studies. SIFTER utilizes measurement-based input (e.g., m/z, CCS, and KMD) to classify these unknown molecules into a taxonomic tree using random forest algorithms. A large-scale, self-consistent IM-MS data repository (Unified CCS Compendium) was used to train the RF algorithms. The initial results with the test sets yielded a perfect prediction outcome of 4/4 in >99% of cases. When SIFTER was

used to analyze unknowns within a complex biological matrix, the 4/4 success rates dropped to 81%, which was supported by OOB error estimates. The two main reasons the performance success decreased is the incomplete cover- age of the training set used and a nearest -neighbor bias in RF algorithms as previously described. We note here that although we have chosen to utilize only drift tube ion mobility CCS values, the SIFTER algorithm should also perform well using CCS measurements obtained from other experimental methods (e.g., TIMS, TWIMS, etc.) given that the precision of these measurements is sufficiently high.

Overall, the outcome of SIFTER is promising, with successful chemical class predictions being observed in the majority of cases (>80%) for compounds derived from complex sample matrixes. Future work will focus on improving SIFTER to handle null outcomes and training the algorithm with larger data sets representing more complete taxonomic coverage. It is anticipated that SIFTER will help investigators learn more about their particular samples, by providing chemical classifications on a large number of features derived from IM-MS experiments. Informatic capabilities provided by SIFTER should help improve data analysis for untargeted IM-MS studies as well as provide important chemical information that can serve as guidance for targeted work. Performance of the SIFTER algorithm will continue to improve as the training data. that is, the Unified CCS Compendium, expands.

4.5 Acknowledgments

This dissertation chapter was adapted from a published manuscript titled “Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization” by Jaqueline A. Picache, Jody C. May, and John A. McLean published in *Analytical Chemistry* **2020**, Just Accepted, DOI: https://dx.doi.org/10.1021/acs.analchem.0c02137.

This work was supported by the resources of the Center for Innovative Technology at Vanderbilt. Financial support for this research was provided by the National Institutes of Health (NIH NCI 1R03CA222452- 01) and the NIH supported Vanderbilt Chemical Biology Interface training program (5T32GM065086-16). I acknowledge Hakmook Kang and Simon Vandekar of the Vanderbilt University Biostatistics Clinic for their guidance in method development and statistical analyses.

4.6 References

(1) May, J. C.; McLean, J. A. *Annu. Rev. Anal. Chem*. **2016**, 9, 387− 462.

(2) Sherrod, S. D.; Mclean, J. A. *Clin. Chem*. **2016**, 62 (1), 77−83.

(3) Baker, E. S.; Patti, G. J. J. *Am. Soc. Mass Spectrom*. **2019**, 30 (10), 2031−2036.

(4) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; Mclean, J. A. *J. Am. Soc. Mass* Spectrom. **2016**, 27, 1897−1905.

(5) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. *Chem. Sci*. **2019**, 10 (4), 983−993.

(7) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z.-*J. Anal. Chem*. **2016**, 88 (22), 474 11084−11091.

(8) Zhou, Z.; Tu, J.; Xiong, X.; Shen, X.; Zhu, Z. *J. Anal. Chem*. **2017**, 89, 9559.

(9) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. *J. Cheminf*. **2019**, 11 (1), 2.

(10) Baranwal, M.; Magner, A.; Elvati, P.; Saldinger, J.; Violi, A.; Hero, A. O. *Bioinformatics* **2020**, 36 (8), 2547.

(11) Gao, S.; Chen, W.; Zeng, Y.; Jing, H.; Zhang, N.; Flavel, M.; Jois, M.; Han, J. D. J.; Xian, B.; Li, G. *BMC Pharmacol. Toxicol*. **2018**, 483 19 (1), 18.

(12) Meyer, J. G.; Liu, S.; Miller, I. J.; Coon, J. J.; Gitter, A. *J. Chem. Inf. Model*. **2019**, 59 (10), 4438.

(13) Ross, D. H.; Cho, J. H.; Xu, L. *Anal. Chem*. **2020**, 92 (6), 487 4548−4557.

(14) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S.; et al. *J. Cheminf*. **2016**, 8 (1), 491 61.

(15) McLean, J. A. *J. Am. Soc. Mass Spectrom*. **2009**, 20 (10), 1775−1781.

(16) Valentine, S. J.; Plasencia, M. D.; Liu, X.; Krishnan, M.; Naylor, S.; Udseth, H. R.; Smith, R. D.; Clemmer, D. E. *J. Proteome Res*. **2006**, 5, 2977−2984. 497

(17) Liu, X.; Valentine, S. J.; Plasencia, M. D.; Trimpin, S.; Naylor, S.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2007**, 18 (7), 1249− 499 1264.

(18) Isailovic, D.; Kurulugama, R. T.; Plasencia, M. D.; Stokes, S. T.; Kyselova, Z.; Goldman, R.; Mechref, Y.; Novotny, M. V; Clemmer, D. E. *J. Proteome Res*. **2008**, 7 (3), 1109−1117.

(19) Kliman, M.; May, J. C.; McLean, J. A. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2011**, 1811 (11), 935−945.

(20) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; et al. *Anal. Chem.* **2017**, 89 (17), 9048−9055.

(21) Lerno, L. A.; German, J. B.; Lebrilla, C. B. Anal. Chem. 2010, 82 (10), 4236−4245.

(22) Kune , C.; McCann, A.; RaphaeЇ, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; et al. *Anal. Chem*. **2019**, 91 (20), 13112−13118.

(23) Causon, T. J.; Ivanova-Petropulos, V.; Petrusheva, D.; Bogeva, E.; Hann, S. *Anal. Chim. Acta* **2019**, 1052, 179−189.

(24) Venter, P.; Causon, T.; Pasch, H.; de Villiers, A. *Anal. Chim. Acta* **2019**, 1088, 150−167.

(25) Core Team, R. R: A language and environment for statistical computing. R Foundation for Statistical Computing https://www.r- project.org/.

(26) Breiman, L. *Mach. Learn*. **2001**, 45, 5−32.

(27) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News*, **2002**, 2(3), 18-22.

(28) Kendrick, E. *Anal. Chem*. **1963**, 35 (13), 2146−2154.

(29) McLean Research Group Github https://github.com/ McLeanResearchGroup.

(30) Picache, J. A.; May, J. C.; McLean, J. A., "Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization" *Anal. Chem.* **2020**, Just Accepted, DOI: https://dx.doi.org/10.1021/acs.analchem.0c02137.

CHAPTER V

Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses

5.1 Introduction

Mass spectrometry serves as a foundational analytical technology in untargeted omics experiments.[1] In recent decades, MS has enabled the collection of big data in biomedical research. As more data is collected and the age of big data matures, many opportunities arise to gain insightful knowledge about biomedical systems that were not previously accessible.[2] However, many of these opportunities remain unseized due to challenges in annotating omics data, especially in the realm of small molecules.[1,3,4] Waldman and Terzic aptly describe why annotating big data is difficult:

> "While the goal is to extract insights from complex, noisy, and heterogeneous datasets, barriers have included the speed of data handling, curation and the veracity of the data, the sheer volume of data, and the heterogeneity of data to be integrated."[5]

To overcome such barriers, mass spectrometrists have turned to bioinformatic solutions which include curating data sets and building databases, data libraries, and/or data repositories. While these terms are often used interchangeably, their definitions have nuanced differences as described in Table 5.1.[6] Annotation of omics data relies heavily on matches from database queries.[4] Success

Table 5.1. Data Collection Terms

Term	Description
Dataset	A collection of data
Database	An organized collection of records that is standardized to enable searching and retrieval of content
Data library	A collection of data materials in various formats with the purpose of providing information to a target group
Repository	A collection of digital documents stored for preservation and public access

(Table adapted from Picache, et al.[31])

in the annotation process is contingent upon the quality of the database being queried as well as the amount of unique information known about the omic compound in question. A few prominent, large-scale databases include the Human Metabolome Database, PubChem, and UniProt.[7–9]

All three of these databases rely on crowd-sourced information. Generally, crowd-sourcing is an active solicitation of content, ideas, or services from a large community. When performed by scientific database curators, crowd-sourcing involves active parsing of the scientific literature to update and addend contents in an automated fashion. This automated crowd-sourcing process is necessary given that there are reports of >290,000 proteins and >25,000 endogenous metabolites in humans.[7,10] While databases such as those previously mentioned provide an important service to the biomedical research community, they remain incomplete, and in some cases, it is challenging to recognize where they are incomplete. As a result, research groups end up developing their own data libraries or databases. Consequences of building personalized libraries and databases include a loss of time and resources due to a redundancy of data acquisition and curation, limited scientific collaboration due to incompatibilities (e.g., informatics, jargon, etc.), and research opacity because raw data is often not referenced or otherwise available.[11] To alleviate these consequences, we propose that field experts build a crowd- sourced database that integrates into successful pre-existing workflows. It should be noted that contributors to the database (i.e., the crowd) will most likely also be field experts. Since database developing is an iterative process, open dialogue between the developer and crowd is encouraged to meet field specific needs. Two examples of successful crowd-sourced databases are the MassBank of North America (MONA) and the Unified Collision Cross Section Compendium. MONA was the first public repository for small molecule mass spectral data.[12] The Unified CCS Compendium is a database of drift-tube ion mobility mass spectrometry data of omic compounds.[13] Here, we discuss a model to create a crowd-sourced

omics database including five pillars of database features that need to be considered. Further, we discuss design concepts and how crowd- sourcing is currently done within the research community.

5.2 Database Features

A generalized schematic of how an omic database is developed is shown in Fig. 5.1. Specifically, an initial data set is processed via data curation, standardization, and annotation with metadata. Next, the curated data set is compiled into a database that gets disseminated to others within a field. These end users utilize the information within the database to gain knowledge about their own experiments, which leads to novel scientific conclusions and the formulation of future questions. These new conclusions become newly generated data sets which then undergo dereplication, validation, and standardization to be added to the pre-existing database. Even though this process only contains five general stages, much should be considered along each step. It is recommended that the following features be considered before data acquisition and curation as well as development of a database begins.

5.2.1 Standardization Requirements The overall goal of a database is to create a collection of data that end users can use with as few barriers as possible.[14] One way to minimize barriers that end users will face is to create a standardized system which includes a standard data type, reporting format, terminology, quality control process, metadata inclusion, and/or reference material information, as shown in Fig. 5.2. Data type refers to what kind of data the database will contain. Will it be data from one specific technology or technique? Will this data be in the primary (raw).

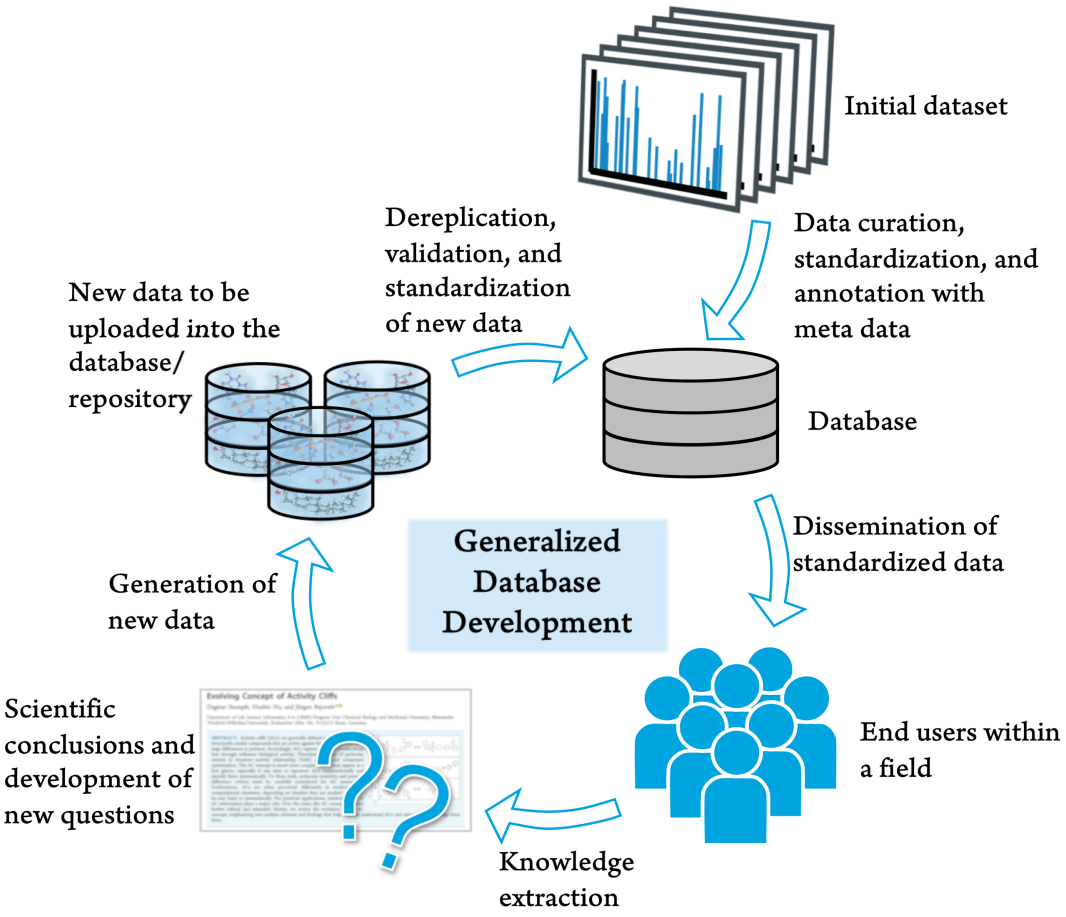


Figure 5.1 General Database Development Databases start with an initial dataset that undergoes standardization. This standardized database is disseminated through the research-peer review cycle. Subsequently, new data is added to the existing database and the cycle begins again. (Figure adapted from Picache, et al.[31])

or secondary (processed) form? Primary data is preferred for scientific transparency. However, it is often larger and will require more computational storage space and data management resources. Secondary data is more common due to their smaller storage requirements and ease of use. Most end users prefer to look at conclusive or summative data.[15]

Reporting format and standard terminology must be considered. If a database contains primary data, how will that data be uploaded by the user? The database management system will need the capabilities to handle large data file transfers as well as automated indexing of addended data. Database management systems are discussed further in section 5.3. Databases that contain secondary data are easier to manage in terms of indexing and storage needs. However, the database developers need to create a standard format that is both informative and facile enough for end-users to comply with. Database specific terminology must be defined from the beginning so that users understand what is required of them and how to use the data within.[14] Such terminology should unambiguously convey experimental design, data acquisition, and data processing parameters. Furthermore, any information needed to provide a context for the reported results should also be included. This enables other users to fully understand the stated conclusions and compare studies from different research groups. One crucial aspect of databases is that each record within the database needs to have a unique identifier.[14] In metabolomics, this can be a compound's InChI Key or molecular structure. In genomics, this could be a specific gene locus. This unique identifier enables universal indexing of records without ambiguity and quick data import and export from the database.

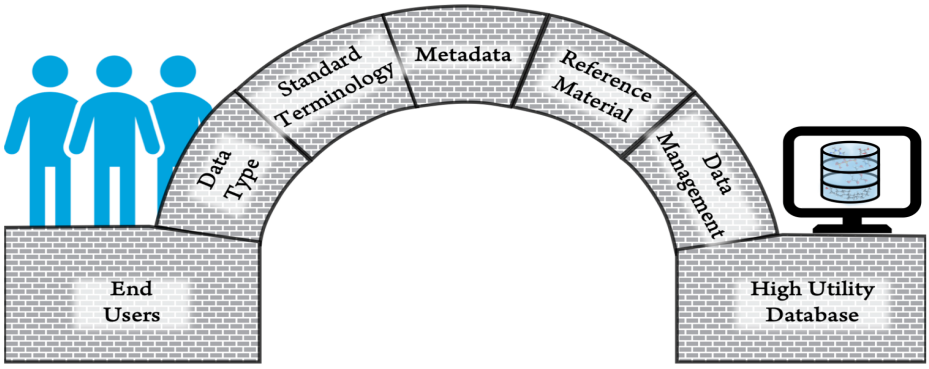


Figure 5.2. Database Features To maximize the utility of their database, developers should consider the data type, standard terminology, included metadata, reference materials, and data management systems when designing their database. (Figure adapted from Picache, et al.[31])

5.2.2 Metadata Documentation Metadata is defined as "minimum information needed to ensure that submitted data are sufficient for clear interpretation and querying by other scientists."[14] As previously mentioned, inclusion of contextual information as well as experimental procedures is imperative. Providing metadata maximizes a data set's utility by allowing others to understand, reproduce, and build off of reported work. Database developers should provide guidelines about what type of information is needed for interpretation and querying for a specific field. Alternatively, a database can contain primary references for a given data set such that end users can obtain the metadata elsewhere. The Minimum Information for Biological and Biomedical Investigations (MIBBI) is a useful resource when deciding what metadata should be included.[16] MIBBI contains registries of reporting efforts for biological/biomedical studies as well as field specific recommendations which include the Minimum Information about a Genome Sequence (MIGS) and Minimum Information about a Metagenomic Sequence (MIMS) for genomic and transcriptomic data, the Minimum Information about a Proteomics Experiment (MIAPE) for proteomics data, and the Core Information for Metabolomics Reporting (CIMR) for metabolomics data.[16–19]

5.2.3 Reference Materials The standardization process goes beyond informatics and reporting. It is recommended that database guidelines include a physical standard such that the reagent can be added as a control in experiments. This standard would serve as a stable reference point for data quality when compared to known experimental values for said reference standard.[14] Having a reference standard that is accessible to a breadth of end users enables data comparisons and quality checks between experiments, across platforms, and between research groups. This is particularly important when used in omics experiments in clinical/diagnostic settings. The reference standard

choice is often decided by the users within the field, but standard materials and associated measurements are provided by the National Institute of Standards Technology (https:// www.nist.gov/services-resources/standards-and- measurements) in the United States and the Laboratory of the Government Chemist (http://lgc.co.uk) in the United Kingdom.

5.2.4 Quality Assurance The success of any database is contingent upon its quality assurance (QA). QA for a database is the process that ensures that data and informatic tools within meet a certain standard as dictated by the database design model and specifications.[20] QA is a twofold process: The first is during initial development of the database. It begins with the standardization procedures previously described as well as developing tools that can audit the database intermittently. These audits should ensure that all of the data is represented accurately and as planned, and that all of the database functionality is operating properly.[20] The second process is when new data is added to the database. Procedures should be in place to vet the quality of the incoming data such that it meets the standardization requirements previously set forth.20 This ensures that the integrity of the database is maintained.

5.3 Design Concepts

5.3.1 Conceptual Design The conceptual design of a database defines the data requirements of and the application of the database in question.[15] This includes the metadata requirements as well as the standardized data type and format as previously discussed. This pillar of the database design is field and data specific. Discussions about the aforementioned database features from section 5.2 should be included during the conceptual design phase.

5.3.2 Logical Design The logical design of a database involves the implementation and management systems of the database.[15] It can be thought of as the "back-end" design of a database. Particular attention should be paid to determining how and when data will be normalized and background/noise corrected and which, if any, further transformations will occur. Often, genomic and transcriptomic databases contain primary data that is normalized by the database infrastructure during the data submission process. Proteomic and metabolomic data is usually presented as secondary data that includes the larger context of the experiment performed.[15] Additionally, proteomic and metabolomic data have more variety in potential output, in terms of content and size, when compared to genomic and transcriptomic data. As a result, this type of data is normalized and background subtracted before submission to a database. Developers need to also consider if data sets are to be kept separate or merged. Data sets can be merged to save space and represent a crowd-sourced conclusion. Keeping them separate enables study comparisons. Both options are used in bioinformatics, and the overall aim of a given database will determine which is more suitable.

Once data is addended into a database, it needs to be maintained. Several options exist to retain order and search capabilities of a database. For large data sets, SQL Server can be used. It works well with relational data, especially if individual records have many attributes associated with them.[21] SQL data sets can be transformed into the XML data format which is amenable to many informatic solutions and coding languages. For smaller data sets, developers can utilize spreadsheet-based solutions which are easily hosted online and can be transferred via CSV data format. Automated maintenance and quality checks are recommended for both options and should be determined before development begins. However, maintenance is an iterative process and

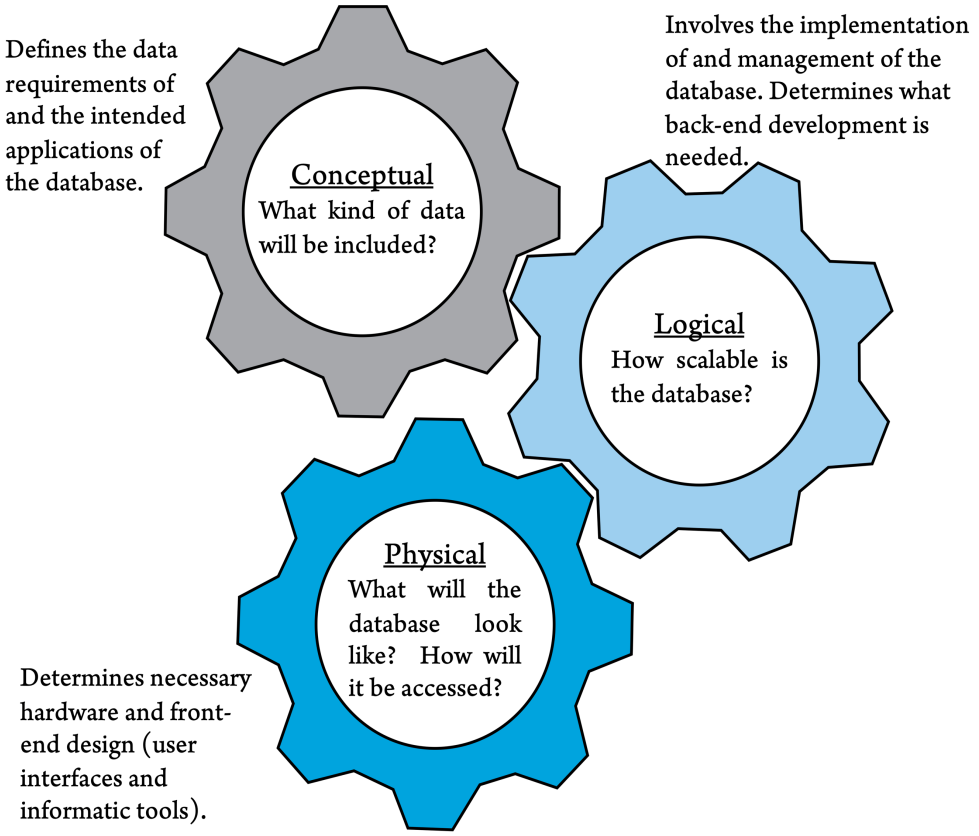


Figure 5.3 Database Design Concepts The three phases of database design include the conceptual, logical, and physical stages. (Figure adapted from Picache, et al.[31])

should be adjusted as needed. These tendencies are general and individual developers should choose the appropriate logistical design for their specific type of data.

5.3.3 Physical Design The physical design of a database involves determining the hardware necessary to support the database as well as the design of any graphical interfaces needed.[15] It can be thought of as the front-end design pillar in which the ease of use by experts in the field is the top priority. Developers should determine if their database will be hosted via an application or online. Additionally, developers should decide if and what to archive (i.e., should outdated results be kept?) as well as design tools that can query live and archived data.[20] Informatic tools such as statistical models and data visualization graphics are also designed during this phase.[15] This state of the design process is the most open-ended, and graphical output can vary widely. Furthermore, it is the most iterative stage, as databases are likely to change depending on their contents and new tools being added. Fig. 5.3 summarizes the three design concepts of planning a database.

5.4 Crowd-sourcing Data

The past decade has seen a push for data sharing and crowd- sourcing research.[11,22–24] The age of big data has matured alongside the ongoing improvements in computational power and data storage capacity. These concurrent movements allow researchers to gather more data than that which they could have collected independently and perform wide-scale studies not previously possible. There are two main crowd-sourcing techniques being used: (1) data-mining from publicly available large data sets and (2) crowd-sourcing data acquisition and/or analysis.[24] The first technique is often used in public health studies where large numbers of data points are needed and through which patient histories are sifted.[2,24] The second type of crowd-sourcing is used in multiple

disciplines within biomedical research including computational chemistry, genomics, medicinal chemistry, natural product discovery, pharmacology, proteomics, and toxicology to name a few.[11,22,25–27] Crowd- sourcing data collection can reduce bias in data acquisition and concluded results.[23] On the other hand, crowd-sourcing data analysis increases the transparency of the experiment and results by having multiple groups reach a concordance about the study and its conclusions.[23] In both scenarios, there is an increase in constructive discourse about the results and conclusions of the study due to the egalitarian nature of crowd-sourcing.[23] While crowd-sourcing provides a wealth of information, it comes with conditions that should be considered. It often requires a lot of time, resources, and personnel to maintain large data sets. Additionally, there is less control over experimental conditions and data collection quality.[23] Despite these caveats, crowd-sourcing has still proven to be a powerful technique, especially when combined with machine learning and new bioinformatic strategies.[2,5,24] Companies like Amazon, Google, and IBM have shown the advantages of using machine learning to better understand the habits of their customers. Researchers are doing the same with techniques like self-organizing maps, neural networks, and classification algorithms.[13,28–30] All of these methods require a large data input, and researchers are using crowd-sourced data to perform them. Results may be more informative about the state of a given system when crowd-sourced databases are used, especially when integrated into omics analysis workflows.

5.5 Conclusions and Outlook

Biomedical research is moving toward using big data that is often crowd-sourced in order to make more general conclusions of observed phenomena. Using crowd-sourced data has many benefits including a large sample pool, increased transparency throughout the scientific method, and more

constructive discourse within a field or project. However, crowd-sourced data has a variable level of quality which can compromise results. By creating databases with crowd-sourced data sets, quality assurance procedures can be put into place. While this process is laborious, the end result is a highly curated database that can be used for the foreseeable future.

While there is no one metric of success for a database, one gauge can be how widely disseminated and utilized the database is. The Unified CCS Compendium is an example of a successful database given that it is used by field experts internationally in a variety of studies, both fundamental and biomedical based investigations. This success can be attributed to the Compendium being user-friendly and user-focused. Metadata standards as well as inclusion guidelines are explicitly provided to contributors. Furthermore, a standardized spreadsheet-based reporting format is provided along with guidelines about the quality control process. Further discussion and specific details on these attributes have been previously reported.[13] Two final considerations pertain to (1) funding crowd-sourced databases and (2) practical considerations for the longevity and ongoing maintenance of these databases once they are developed. Resolutions to both of these considerations are ongoing discussions within the informatics community, and there is no one solution. One potential funding resource is collaboration with other research groups, the private sector, and/or a government agency. However, we propose a governmental/private sector alliance to retain the open-access databases post development while accepting contributions from academic researchers. This provides for any practical resources needed to maintain these large databases as well as continued open dialogue between all three sectors. Ultimately, this facilitated collaboration will enable biomedical researchers to take advantage of the opportunities and discoveries that this wealth of data presents.

5.6 Acknowledgments

This dissertation chapter was adapted from a published mini-review article titled "Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analysis" by Jaqueline A. Picache, Jody C. May, and John A. McLean published in *ACS Omega* **2020**, *5* (2), 980-985.

Financial support for this research was provided by the National Institutes of Health (NIH NIGMS R01GM092218 and NIH NCI 1R03CA222452-01) and the NIH supported Vanderbilt Chemical Biology Interface training program (5T32GM065086-16).

5.7 References

(1) May, J. C.; McLean, J. A. *Annu. Rev. Anal. Chem*. **2016**, 9, 387.

(2) Costa, F. F. *Drug Discovery Today.* **2014**, 19 (4), 433.

(3) Baker, E. S.; Patti, G. J. *J. Am. Soc. Mass Spectrom*. **2019**, 30 (10), 2031.

(4) Scheubert, K.; Hufsky, F.; Petras, D.; Wang, M.; Nothias, L.-F.; Duehrkop, K.; Bandeira, N.; Dorrestein, P.; Boecker, S. *Nat. Commun*. **2017**, 8, 1494.

(5) Waldman, S. A.; Terzic, A. *Clin. Pharmacol. Ther*. **2016**, 99 (3), 250.

(6) Levine-Clark, M.; Carter, T. M. ALA Glossary of Library and Information Science; American Library Association: Chicago, IL, **2013**; pp 86−87, 163, 230.

(7) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. *Nucleic Acids Res*. **2018**, 46 (D1), D608.

(8) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. *Nucleic Acids Res*. **2019**, 47 (D1), D1102.

(9) Bateman, A. *Nucleic Acids Res*. **2019**, 47 (D1), D506.

(10) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S. K.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. Nature 2014, 509 (7502), 575.

(11) Riccardi, E.; Pantano, S.; Potestio, R. *Interface Focus* **2019**, 9 (3).

(12) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom*. **2010**, 45 (7), 703.

(13) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. *Chem. Sci*. **2019**, 10 (4), 983.

(14) Chervitz, S. A.; Deutsch, E. W.; Field, D.; Parkinson, H.; Quackenbush, J.; Rocca-Serra, P.; Sansone, S.-A.; Stoeckert, C. J., Jr.; Taylor, C. F.; Taylor, R.; Ball, C. A. Bioinformatics for Omics Data; Humana Press: New York, **2011**; pp 31−69.

(15) Schneider, M. V.; Orchard, S. Bioinformatics for Omics Data; Humana Press: New York, **2011**; pp 3−30.

(16) Taylor, C. F.; Field, D.; Sansone, S. A.; Aerts, J.; Apweiler, R.; Ashburner, M.; Ball, C. A.; Binz, P. A.; Bogue, M.; Booth, T.; Brazma, A.; Brinkman, R. R.; Michael Clark, A.; Deutsch, E. W.; Fiehn, O.; Fostel, J.; Ghazal, P.; Gibson, F.; Gray, T.; Grimes, G.; Hancock, J. M.; Hardy, N. W.; Hermjakob, H.; Julian, R. K.; Kane, M.; Kettner, C.; Kinsinger, C.; Kolker, E.; Kuiper, M.; Novère, N. Le; Leebens-Mack, J.; Lewis, S. E.; Lord, P.; Mallon, A. M.; Marthandan, N.; Masuya, H.; McNally, R.; Mehrle, A.; Morrison, N.; Orchard, S.; Quackenbush, J.; Reecy, J. M.; Robertson, D. G.; Rocca-Serra, P.; Rodriguez, H.; Rosenfelder, H.; Santoyo-Lopez, J.; Scheuermann, R. H.; Schober, D.; Smith, B.; Snape, J.; Stoeckert, C. J.; Tipton, K.; Sterk, P.; Untergasser, A.; Vandesompele, J.; Wiemann, S. Nat. Biotechnol. 2008, 26 (8), 889. (17) Field, D.; Garrity, G.; Gray, T.; Morrison, N.; Selengut, J.; Sterk, P.; Tatusova, T.; Thomson, N.; Allen, M. J.; Angiuoli, S. V.; Ashburner, M.; Axelrod, N.; Baldauf, S.; Ballard, S.; Boore, J.; Cochrane, G.; Cole, J.; Dawyndt, P.; Vos, P. De; DePamphilis, C.; Edwards, R.; Faruque, N.; Feldman, R.; Gilbert,

J.; Gilna, P.; Glöckner, F. O.; Goldstein, P.; Guralnick, R.; Haft, D.; Hancock, D.; Hermjakob, H.; Hertz-Fowler, C.; Hugenholtz, P.; Joint, I.; Kagan, L.; Kane, M.; Kennedy, J.; Kowalchuk, G.; Kottmann, R.; Kolker, E.; Kravitz, S.; Kyrpides, N.; Leebens-Mack, J.; Lewis, S. E.; Li, K.; Lister, A. L.; Lord, P.; Maltsev, N.; Markowitz, V.; Martiny, J.; Methe, B.; Mizrachi, I.; Moxon, R.; Nelson, K.; Parkhill, J.; Proctor, L.; White, O.; Sansone, S.-A.; Spiers, A.; Stevens, R.; Swift, P.; Taylor, C.; Tateno, Y.; Tett, A.; Turner, S.; Ussery, D.; Vaughan, B.; Ward, N.; Whetzel, T.; Gil, I. S.; Wilson, G.; Wipat, A. *Nat. Biotechnol*. **2008**, 26 (5), 541.

(18) Martínez-Bartolomé, S.; Binz, P.-A.; Albar, J. P. Plant Proteomics: Methods and Protocols; Humana Press: New York, **2014**; Vol. 1072, pp 765−780.

(19) Spicer, R. A.; Salek, R.; Steinbeck, C. Sci. Data 2017, 4, 170137.

(20) Harel, A.; Dalah, I.; Pietrokovski, S.; Safran, M.; Lancet, D. Bioinformatics for Omics Data; Humana Press: New York, **2011**; Vol. 719, pp71−96.

(21) Sander, A.; Wauer, R. J. Biomed. Semantics 2019, 10,7. (22) Björnmalm, M.; Caruso, F. *Angew. Chem., Int. Ed*. **2018**, 57 (5), 1122.

(23) Silberzahn, R.; Uhlmann, E. L. *Nature* **2015**, 526 (7572), 189.

(24) Khare, R.; Good, B. M.; Leaman, R.; Su, A. I.; Lu, Z. *Briefings Bioinf*. **2016**, 17 (1), 23.

(25) Budin-Ljøsne, I.; Isaeva, J.; Knoppers, B. M.; Tassé, A. M.; Shen, H. Y.; McCarthy, M. I.; Harris, J. R. *Eur. J. Hum. Genet*. **2014**, 22 (3), 317.

(26) Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; Players, F. *Nature* **2010**, 466 (7307), 756.

(27) Poussin, C.; Belcastro, V.; Martin, F.; Boué, S.; Peitsch, M. C.; Hoeng, *J. Chem. Res. Toxicol*. **2017**, 30 (4), 934.

(28) Goodwin, C. R.; Sherrod, S. D.; Marasco, C. C.; Bachmann, B. O.; Schramm-Sapyta, N.; Wikswo, J. P.; McLean, J. A. *Anal. Chem*. **2014**, 86, 6563.

(29) Plante, P.-L.; Francovic-Fontaine, É.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, *J. Anal. Chem.* **2019**, 91 (8), 5191.

(30) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. J. *Cheminf*. **2016**, 8, 61.

(31) Picache, J. A.; May, J. C.; McLean, J. A., "Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses" *ACS Omega* **2020**, *5* (2) 980 – 985.

CHAPTER VI

Conspectus and Outlook

6.1 Conspectus

Ion – mobility mass spectrometry is a powerful tool that can improve multiomic metabolite identification. As previously mentioned, IM-MS has been used in a myriad of omic analyses including lipidomics, glycomics, proteomics, small molecule natural products, and more. One reason for this utility is the collision cross section as derived from IM-MS experiments. Collision cross sections are unique molecular identifiers of molecules which can then be used in traditional identification workflows to improve confidence in analyte identifications.

In order to incorporate CCS values into traditional identification workflows, compound libraries must be developed. This enables database searching against the compound libraries to match analytes with CCS values of known molecular standards. The MSMLS, described in Chapter II, is one such compound library of primary metabolites. Development of the MSMLS library also enabled studies of the ability of IM-MS to separate isomeric species that are otherwise indistinguishable by LC-MS. This serves as an additional improvement to identification workflows. The advantages of using CCS to identify molecules was demonstrated by analyzing the NIST 1950 serum. Exact matches of serum analytes enabled metabolic pathway analysis as previously described.

The full advantages of using CCS values in analyte identification workflows was realized in development of the Unified CCS Compendium described in Chapter III. This crowd-sourced database of CCS values serves as the largest repository of high precision, experimental values. Development of the Compendium required establishing standardized methods of data acquisition

and CCS calculation. Once the quality of the data was assured and met previous interlaboratory study standards, large scale informatic analyses was performed. Conformational space (*m/*z versus CCS/z) of chemical ontological classes and subclasses were analyzed and described via regression modeling. These models enabled confidence interval and predictive interval-based filtering of tentatively identified analytes from untargeted studies to improve confidence. A proof of concept experiment of this filtering method was performed on a human serum sample.

The filtering methods described in Chapter III are specific to analytes that have previously undergone traditional identification workflows. However, there remain analytes that do not have database matches. Using the Unified CCS Compendium as the data set, a machine learning algorithm was developed to gain insight on these unknown analytes. The supervised inference of feature taxonomy from ensemble randomization informatic tool is based on a random forest machine learning algorithm. By performing global regression analysis of the Compendium data, SIFTER classifies unknown analytes into a chemical ontological kingdom, super class, class, and, potentially, a subclass using only *m/z*, CCS, and Kendrick mass defect. SIFTER successfully classified molecules within a test set and complex biological matrix. Specific examples of the success of SIFTER are found in Chapter IV.

Chapters II-IV describe the utility of CCS in identification workflows as well as the power of crowd-sourced databases in developing informatic tools that improve confidence in analyte identifications. However, crowd-sourced chemistry is not limited to CCS values and IM-MS. Chapter V describes the value of using crowd-sourcing methods in other fields as well as how to go about designing a crowd-sourced database. Namely, a crowd-sourced database needs to have standardization metrics of quality as well as jargon and formatting such that it can be adopted universally. Specific features and design considerations are described within Chapter V.

6.2 Outlook

In order to gain biological insight into small molecule analysis, a Level 2 identification, as described in Chapter I, is recommended. The two main methods for achieving a Level 2 ID are via fragmentation experiments or the use of ion-mobility. One area left unexplored is performing IM-MS analysis on fragments of small molecules. These types of experiments would enable investigation of isomeric compounds to a degree not yet possible. Proof of concept experiments for IM-MS analysis of small molecule fragments is described below.

6.2.1 Analysis of Pre- and Post-IM Fragmentation Patterns Before IM-MS analysis of small molecule fragments could occur, fragmentation patterns needed to be validated. In other words, experiments were performed to ensure fragmentation patterns from pre-IM collision induced dissociation (CID) matched that of traditional post-IM CID. To test this, a suite of reference standards (Agilent Technologies) containing symmetrically-branched fluoroalkyl phosphazines, namely hexaxis(fluoroalkoxy)phosphazines (*m/*z 622, 922, and 1222) and tris(fluoroalkyl)triazines (*m/*z 322) were analyzed. Preliminary results showed that pre-IM and post-IM CID generated similar fragments but at different relative abundances. Furthermore, pre-IM CID required higher applied voltages at the modified cap lens as compared to the traditional applied voltages at collision cell hexapole in order to generate fragments. This data is summarized in **Figure 6.1**.

6.2.2 Analysis of Isomeric Trisaccharides Analysis of three isomeric trisaccharides was performed to measure CCS values of pre-IM CID fragments. Specifically, maltotriose, raffinose, and melezitose were individually analyzed via CID-IM-MS. Similar to the fluoroalkylphosphazines in Section 6.2.1, the three isomeric compounds produced the same fragments, namely *m/*z 169, 187, and 349, but at different relative abundances. It should be noted that the ion species observed had

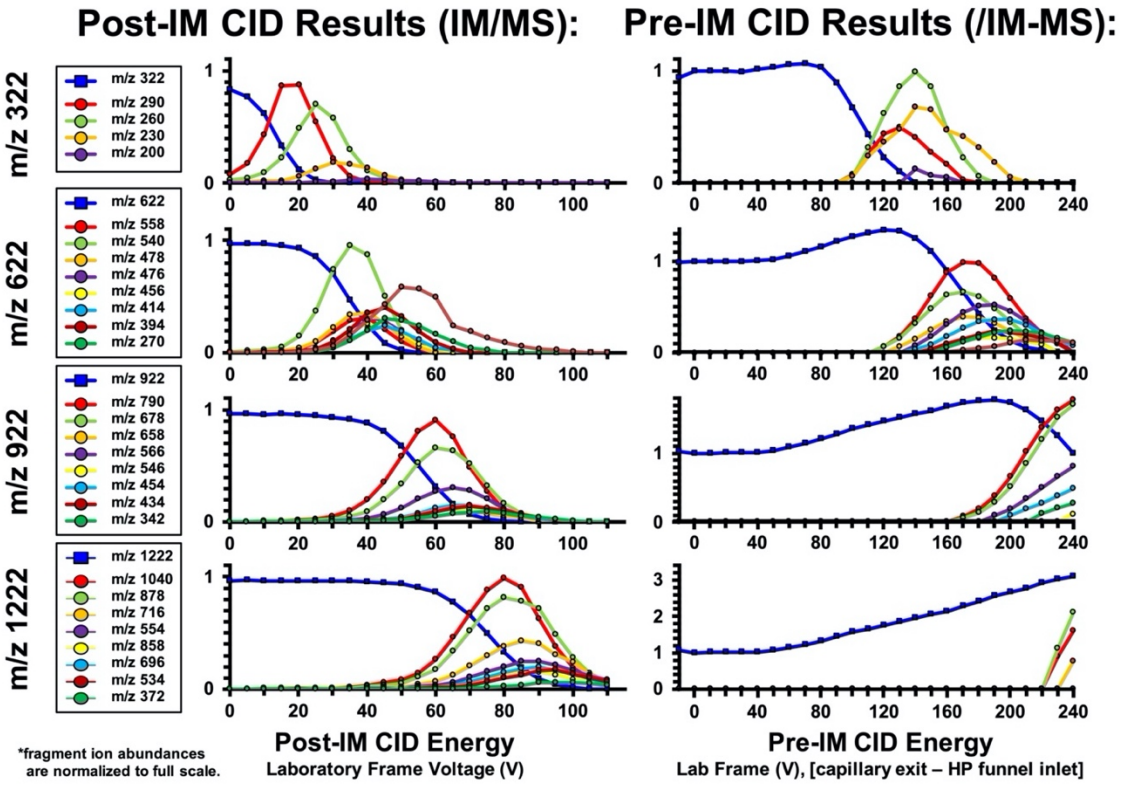


Figure 6.1. Pre-IM CID versus Post-IM CID Fragmentation Patterns Analysis of *m/z* 322, 622, 922, and 1222 reference standard ions. Pre-IM CID required more energy to generate the same mass fragments. However, the relative abundances of the fragment masses were different.

a loss of water and was sodium adducted. Additionally, the pre-IM CID voltage required to fragment the trisaccharides was higher than that of the post-IM CID voltage. This data is summarized in **Figure 6.2**.

IM-MS analysis of the intact molecules indicates that melezitose (201Å) can be separated from raffinose and maltotriose (207Å). This iterates the utility of IM to differentiate isomers where other technologies might not be able to. Further investigation shows unique CCS values between raffinose and maltotriose when both molecules are fragmented at 540V via pre-IM CID. Raffinose's fragment *m/z* 349 has a CCS of 175Å whereas maltotriose has a CCS of 177Å for *m/z* 349. Raffinose and maltotriose can be distinguished via the CCS of their other fragments *m/z* 187 and 169 as shown in **Figure 6.3**.

IM-MS analysis of the pre-IM CID fragments also indicated that different fragments are generated based on the voltage applied at the modified cap lens for two of the three trisaccharides. In these examples, the trisaccharide was analyzed at 540V or 580V; and the drift tube arrival time profiles were analyzed, as shown in **Figure 6.3**. Raffinose had multiple drift peaks for its *m/z* 349 fragment when the pre-IM CID voltage was 580V. Maltotriose had multiple drift peaks for its intact *m/z 511*, fragments *m/z* 187, and *m/z* 169 when 580V was applied to the cap lens. This potentially indicates unique fragment activation pathways depending on the applied voltage of small carbohydrates. Melezitose, on the other hand, did not have any unique drift profiles between 540V and 580V applied to the cap lens.

These results indicate a potential new area of investigation: IM-MS of molecular fragments. By generating compound libraries of CID-IM-MS, isomeric species can be distinguished based on their fragments whereas this has not been possible yet. Specifically, isomers

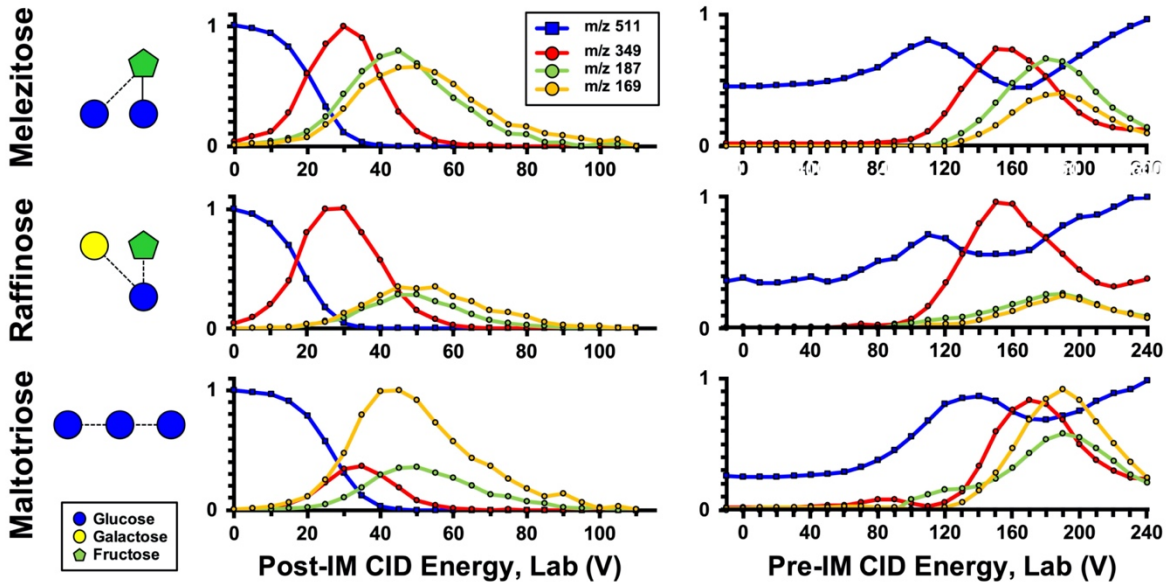


Figure 6.2 Trissacharride Fragmentation Breakdown Curves Maltotriose, raffinose, and melezitose were analyzed via pre-IM CID and post-IM CID.

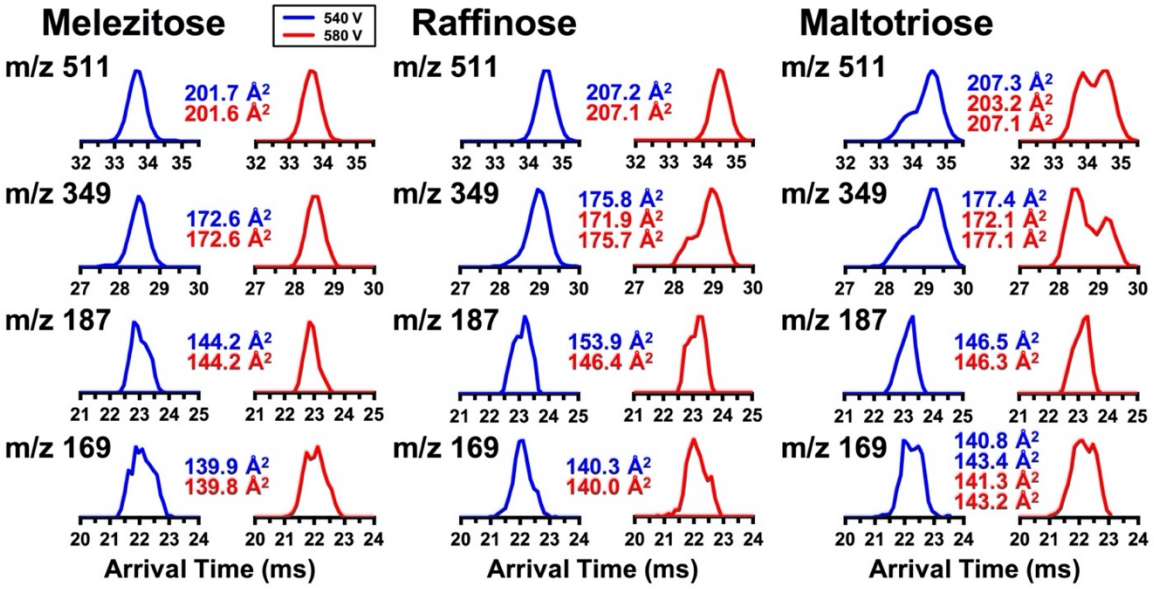


Figure 6.3 IM-MS of Trissaccharide Fragments Intract melezitose can be separated from raffinose and maltotriose at *m/z* 511. Raffinose and maltotriose can be separated and identified via CID-IM-MS due to differences in CCS values of their fragment masses. Applied cap lens voltage make dictate unique fragmentation activation pathways.

can potentially have unique CCS values for fragments with the same mass or fragments may have multiple peaks and therefore unique drift profiles to identify them by.

6.3 Concluding Remarks

The ion mobility – mass spectrometry field has had recent developments that make it ideal to for untargeted studies and identification workflows. A few examples of these developments include building compound libraries, new informatics tools, and its amenability to communal (i.e. crowd-sourced) projects. Chapter II illustrated the power of IM-MS to elucidate identities of isomeric compounds using IM-MS as well as direct compound identification matching of in-house libraries. Chapters III and V showed the power of communal chemistry and the tools that can come of it. By integrating small compound libraries into standardized, quality assured, open-source libraries, the entire IM-MS community can benefit by having access to such a large dataset. Furthermore, this dataset can be used to develop informatic tools such as the regression filtering in Chapter III and SIFTER in Chapter IV. Once enough data is curated, machine learning and other informatic tools can be used to aide in the identification of unknown analytes an demonstrated by SIFTER. Future work on expending these crowd-sourced databases as well as new datasets including IM-MS analysis of molecular fragments will help push the IM-MS community forward. Chapter V demonstrated the feasibility of crowd-sourced chemistry that benefits those beyond the IM-MS community. The proponents necessary to begin crowd-sourced database efforts are specified in Chapter V as well. All of the informatic tools herein are available open-source which is a critical part of communal chemistry. Chemistry, as a unified field, has the ability to propel itself forward into new areas of study by participating in communal scientific endeavors.

6.4 Acknowledgments

This dissertation chapter was adapted from a conference proceeding titled “A Novel In-Source Fragmentation Device for Comprehensive Small Molecule Analysis by Collision-Induced Dissociation-Ion Mobility-Mass Spectrometry” by Andrzej Balinski, Jaqueline A. Picache, Ruwan T. Kurulugama, Emanuel Zlibut, Jody C. May, John C. Fjeldsted, and John A. McLean published in *68th ASMS Conference on Mass Spectrometry and Allied Topics*; American Society for Mass Spectrometry: Houston, TX 2020.

Financial support for this research was provided by the National Institutes of Health (NIH NIGMS R01GM092218 and NIH NCI 1R03CA222452-01) and the NIH supported Vanderbilt Chemical Biology Interface training program (5T32GM065086-16).

APPENDIX A

Chapters published in this dissertation were previously published in the following articles:

Chapter I:

Nichols, C. M.; Dodds, J. N.; Rose, B. S.; Picache, J. A.; Morris, C. B.; Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A., "Untargeted Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility – Mass Spectrometry" *Anal. Chem*., **2018**, *90* (24) 1448-14492.

Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A., "Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities" *Chem. Sci*. **2019**, *10* (4), 983-993.

Picache, J. A.; May, J. C.; McLean, J. A., "Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization" *Anal. Chem.* **2020**, Just Accepted, DOI: https://dx.doi.org/10.1021/acs.analchem.0c02137.

Picache, J. A.; May, J. C.; McLean, J. A., "Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses" *ACS Omega* **2020**, *5* (2) 980 – 985.

Chapter II:

Nichols, C. M.; Dodds, J. N.; Rose, B. S.; Picache, J. A.; Morris, C. B.; Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A., "Untargeted Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility – Mass Spectrometry" *Anal. Chem*., **2018**, *90* (24) 1448-14492.

Chapter III:

Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A., "Collision Cross Section Compendium to Annotate and Predict Multi-omic Compound Identities" *Chem. Sci*. **2019**, *10* (4), 983-993.

Chapter IV:

Picache, J. A.; May, J. C.; McLean, J. A., "Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization" *Anal. Chem.* **2020**, Just Accepted, DOI: https://dx.doi.org/10.1021/acs.analchem.0c02137.

Chapter V:

Picache, J. A.; May, J. C.; McLean, J. A., "Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses" *ACS Omega* **2020**, *5* (2) 980 – 985.

APPENDIX B

Supporting Information for Chapter II

Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Fingerprint using Ion Mobility-Mass Spectrometry

Charles M. Nichols‡, James N. Dodds‡, Bailey S. Rose, Jaqueline A. Picache, Caleb M. Morris, Simona G. Codreanu, Jody C. May, Stacy D. Sherrod and John A. McLean*

Department of Chemistry, Center for Innovative Technology, Vanderbilt Institute of Chemical Biology, Vanderbilt Institute for Integrative Biosystems Research and Education, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville Tennessee 37235, United States.

Comments on IM-MS Data presented in this Work

In this supporting information we describe various observations related to general ion mobility behavior and applications, including ion streaking in the mobility cell, ion adduct type observed in this study, and the results of isomer separation analysis that provide an additional reference to supplement our findings in the main body of the manuscript. In addition, we have included the expanded equations using our fitting analysis for Figure 2 of the main text. We also include the full results of our evaluation of the MSMLS plates with measured CCS using MetaboAnalyst 4.0. In addition to this information, we have attached a full compendium of our MSMLS CCS library as an attached Excel file, and personal compound database libraries (PCDL) for incorporation into future untargeted IM-MS workflows.

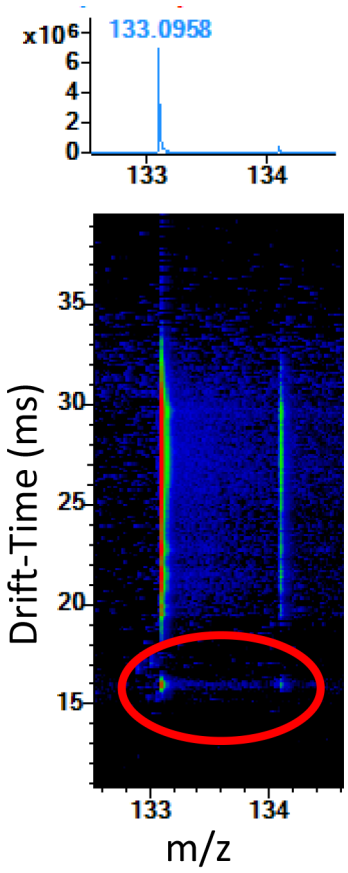


Figure B2.1. **IM-MS spectra for D-Ornathine** The standard is subject to metastable ion dissociation in the DTIMS, resulting in uncorrelated mobility. The true reported conformer is circled in red.

Mass/Mobility Fitting Equations: Power fits for biomolecular super classes selected in the manuscript.

Power Association (General)

$$y = y_0 + (\text{plateau} - y_0) * (1 - e^{-Kx}) \quad \text{Equation B2.1}$$

4P sigmoidal (General)

$$y = y_o + \frac{y_{max} - y_0}{1 + 10^{(\log y_{50} - x) \cdot H}} \quad \text{Equation B2.2}$$

Lipids and lipid-like molecules- Power association

$$y = 75.41 + (516.9 - 75.41) * (1 - e^{-0.000851x}) \quad \text{Equation B2.3}$$

Organic oxygen compounds- Power association

$$y = 101.9 + (935.0 - 101.9) * (1 - e^{-0.000251x}) \quad \text{Equation B2.4}$$

Nucleosides, nucleotides, and analogues- Power association

$$y = 117.6 + (1.52 \times 10^7 - 117.6) * (1 - e^{-1.06 \times 10^{-8} x}) \quad \text{Equation B2.5}$$

Organic acids and derivatives – 4P sigmoidal

$$y = 104.8 + \frac{216.3 - 104.8}{1 + 10^{(277.4 - x) \cdot 0.00376}} \quad \text{Equation B2.6}$$

Table B2.1. Metabolic pathways covered by compounds with at least one measured CCS the MSMLS plate study.

Pathway	Total	Expected	Hits	Raw p	#NAME?	Holm adjust	FDR	Impact	% of Hits
Pyrimidine metabolism	60	10.2	31	5E-10	21.4	4E-08	4E-08	0.7	52
Purine metabolism	92	15.7	34	2E-06	13.0	2E-04	9E-05	0.5	37
Tyrosine metabolism	76	12.9	25	5E-04	7.7	4E-02	1E-02	0.5	33
Alanine, aspartate and glutamate metabolism	24	4.1	11	9E-04	7.0	7E-02	2E-02	0.6	46
beta-Alanine metabolism	28	4.8	12	1E-03	6.8	9E-02	2E-02	0.4	43
Galactose metabolism	41	7.0	15	2E-03	6.2	1E-01	3E-02	0.3	37
Arginine and proline metabolism	77	13.1	23	3E-03	5.7	2E-01	4E-02	0.5	30
Cysteine and methionine metabolism	56	9.5	18	4E-03	5.6	3E-01	4E-02	0.8	32
Glycine, serine and threonine metabolism	48	8.2	15	1E-02	4.5	8E-01	9E-02	0.4	31
Citrate cycle (TCA cycle)	20	3.4	8	1E-02	4.4	9E-01	1E-01	0.3	40
Sulfur metabolism	18	3.1	7	2E-02	3.8	1	2E-01	0.3	39
Nicotinate and nicotinamide metabolism	44	7.5	13	3E-02	3.6	1	2E-01	0.4	30
Amino sugar and nucleotide sugar metabolism	88	15.0	22	3E-02	3.4	1	2E-01	0.5	25
Glutathione metabolism	38	6.5	11	5E-02	3.1	1	3E-01	0.4	29
Phenylalanine metabolism	45	7.7	12	7E-02	2.7	1	3E-01	0.1	27
Ascorbate and aldarate metabolism	45	7.7	12	7E-02	2.7	1	3E-01	0.4	27
Pantothenate and CoA biosynthesis	27	4.6	8	7E-02	2.6	1	3E-01	0.3	30
Biotin metabolism	11	1.9	4	1E-01	2.3	1	4E-01	0.4	36

Taurine and hypotaurine metabolism	20	3.4	6	1E-01	2.2	1	4E-01	0.5	30
Tryptophan metabolism	79	13.5	18	1E-01	2.2	1	4E-01	0.5	23
Nitrogen metabolism	39	6.6	10	1E-01	2.2	1	4E-01	0.0	26
Histidine metabolism	44	7.5	11	1E-01	2.2	1	4E-01	0.4	25
Butanoate metabolism	40	6.8	10	1E-01	2.0	1	4E-01	0.1	25
Starch and sucrose metabolism	50	8.5	12	1E-01	2.0	1	4E-01	0.5	24
Riboflavin metabolism	21	3.6	6	1E-01	2.0	1	4E-01	0.1	29
Caffeine metabolism	21	3.6	6	1E-01	2.0	1	4E-01	0.5	29
D-Arginine and D-ornithine metabolism	8	1.4	3	1E-01	2.0	1	4E-01	0.5	38
Pentose phosphate pathway	32	5.5	8	2E-01	1.8	1	5E-01	0.4	25
Vitamin B6 metabolism	32	5.5	8	2E-01	1.8	1	5E-01	0.4	25
Synthesis and degradation of ketone bodies	6	1.0	2	3E-01	1.3	1	7E-01	0.9	33
Pentose and glucuronate interconversions	53	9.0	11	3E-01	1.3	1	7E-01	0.2	21
D-Glutamine and D-glutamate metabolism	11	1.9	3	3E-01	1.3	1	7E-01	0.1	27
Aminoacyl-tRNA biosynthesis	75	12.8	15	3E-01	1.2	1	7E-01	0.2	20
Lysine biosynthesis	32	5.5	7	3E-01	1.2	1	7E-01	0.3	22
Phenylalanine, tyrosine and tryptophan biosynthesis	27	4.6	6	3E-01	1.2	1	7E-01	0.3	22
Glycerophospholipid metabolism	39	6.6	8	3E-01	1.1	1	8E-01	0.3	21
Valine, leucine and isoleucine degradation	40	6.8	8	4E-01	1.0	1	8E-01	0.2	20
One carbon pool by folate	9	1.5	2	5E-01	0.8	1	1E+00	0.2	22
Glyoxylate and dicarboxylate metabolism	50	8.5	9	5E-01	0.7	1	1E+00	0.1	18
Valine, leucine and isoleucine biosynthesis	27	4.6	5	5E-01	0.7	1	1E+00	0.1	19
Lysine degradation	47	8.0	8	6E-01	0.6	1	1E+00	0.3	17

Fructose and mannose metabolism	48	8.2	8	6E-01	0.5	1	1E+00	0.2	17
Ubiquinone and other terpenoid-quinone biosynthesis	36	6.1	6	6E-01	0.5	1	1E+00	0.2	17
Thiamine metabolism	24	4.1	4	6E-01	0.5	1	1E+00	0.3	17
Fatty acid biosynthesis	49	8.3	8	6E-01	0.5	1	1E+00	0.0	16
Propanoate metabolism	35	6.0	5	7E-01	0.3	1	1E+00	0.1	14
Selenoamino acid metabolism	22	3.7	3	8E-01	0.3	1	1E+00	0.1	14
Linoleic acid metabolism	15	2.6	2	8E-01	0.3	1	1E+00	0.7	13
Cyanoamino acid metabolism	16	2.7	2	8E-01	0.2	1	1E+00	0.0	13
Glycolysis or Gluconeogenesis	31	5.3	4	8E-01	0.2	1	1E+00	0.1	13
Sphingolipid metabolism	25	4.3	3	8E-01	0.2	1	1E+00	0.2	12
Folate biosynthesis	42	7.2	5	9E-01	0.1	1	1E+00	0.1	12
Inositol phosphate metabolism	39	6.6	4	9E-01	0.1	1	1E+00	0.2	10
Pyruvate metabolism	32	5.5	3	9E-01	0.1	1	1E+00	0.0	9
Glycerolipid metabolism	32	5.5	3	9E-01	0.1	1	1E+00	0.0	9
Terpenoid backbone biosynthesis	33	5.6	3	9E-01	0.1	1	1E+00	0.1	9
Methane metabolism	34	5.8	3	9E-01	0.1	1	1E+00	0.0	9
Primary bile acid biosynthesis	47	8.0	4	1E+00	0.0	1	1E+00	0.0	9
Fatty acid metabolism	50	8.5	4	1E+00	0.0	1	1E+00	0.1	8
Retinol metabolism	22	3.7	1	1E+00	0.0	1	1E+00	0.2	5
Fatty acid elongation in mitochondria	27	4.6	1	1E+00	0.0	1	1E+00	0.0	4
N-Glycan biosynthesis	38	6.5	1	1E+00	0.0	1	1E+00	0.0	3
Steroid hormone biosynthesis	99	16.9	6	1E+00	0.0	1	1E+00	0.1	6
Porphyrin and chlorophyll metabolism	104	17.7	6	1E+00	0.0	1	1E+00	0.1	6

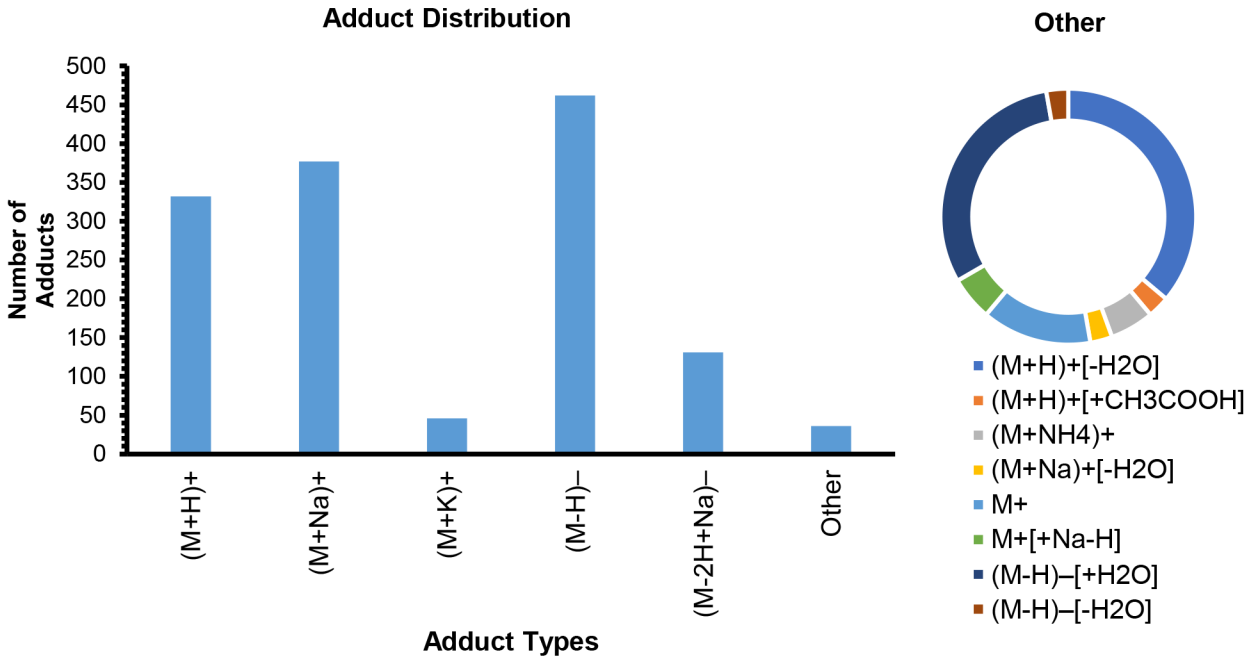


Figure B2.2. A distribution of the adduct types observed from the MSMLS study.

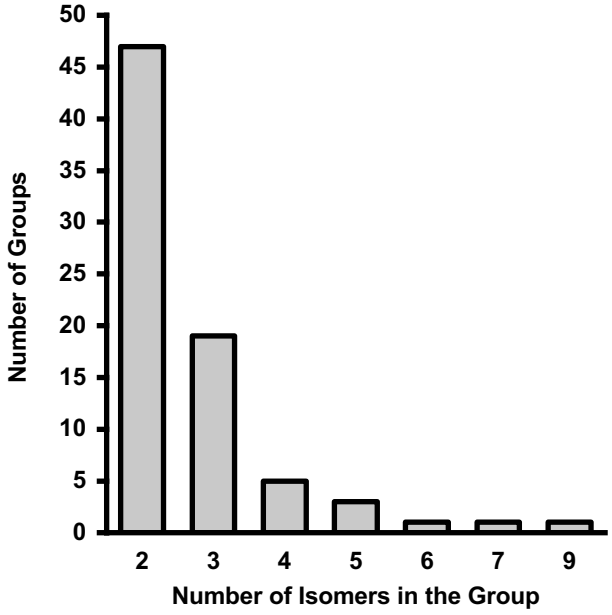


Figure B2.3. The distribution of isomeric families within the MSMLS. Most isomeric sets contain 2 or 3 isomers per group, and the largest set contained 9 isomers.

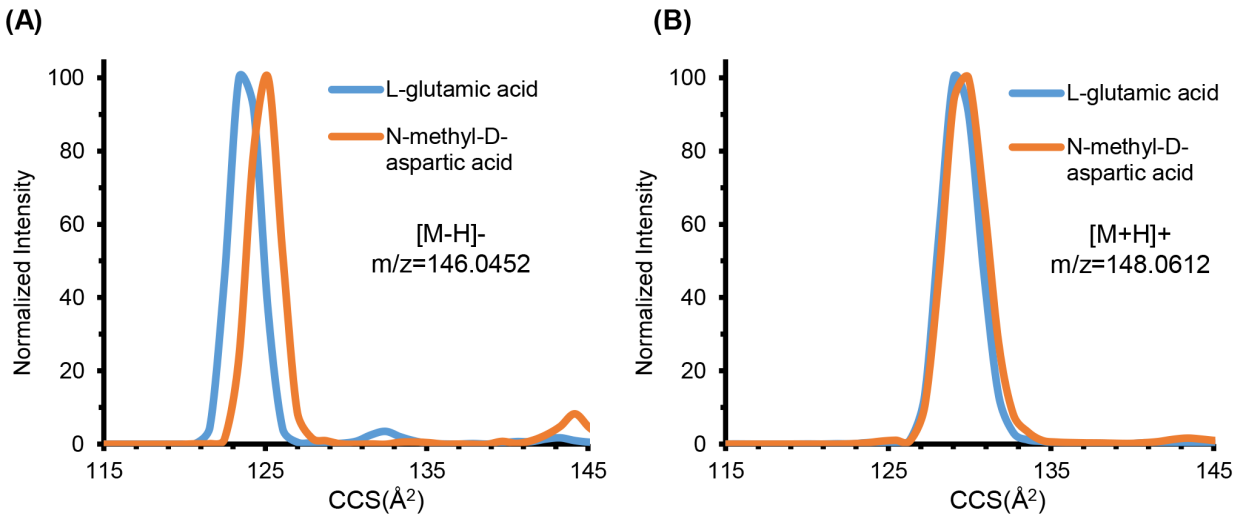


Figure B2.4. IM separation of the deprotonated **(A)** and protonated **(B)** isomers of neutral mass 147.0532 (L-glutamic acid and N-methyl-D-aspartic acid).

APPENDIX C

Supplemental Information for Chapter III

Collision Cross Section Compendium to Annotate and Predict Multiomic Compound Identities

Jaqueline A. Picache, Bailey S. Rose, Andrzej Balinski, Katrina L. Leaptrot, Stacy D. Sherrod, Jody C. May, John A. McLean*

Department of Chemistry, Center for Innovative Technology, Vanderbilt Institute of Chemical Biology, Vanderbilt Institute for Integrative Biosystems Research and Education, Vanderbilt-Ingram Cancer Center; Vanderbilt University, Nashville, Tennessee

Abstract

Ion mobility mass spectrometry (IM-MS) expands the analyte coverage of existing multiomic workflows by providing an additional separation dimension as well as a parameter for characterization and identification of molecules – the collision cross section (CCS). This work presents a large, Unified CCS Compendium of > 3800 experimentally acquired CCS values obtained from traceable molecular standards and measured with drift tube-mass spectrometers. An interactive visualization of this Compendium along with data analytic tools have been made openly accessible. Represented in the Compendium are 14 structurally-based chemical super classes, consisting of a total of 80 classes and 157 subclasses. Using this large data set, regression fitting and predictive statistics have been performed to describe mass-CCS correlations specific to each

chemical ontology. These structural trends provide a rapid and effective filtering method in the traditional untargeted workflow for identification of unknown biochemical species. The utility of the approach is illustrated by an application to metabolites in human serum, quantified trends of which were used to assess the probability of an unknown compound belonging to a given class. CCS-based filtering narrowed the chemical search space by 60% while increasing the confidence in the remaining isomeric identifications from a single class, thus demonstrating the value of integrating predictive analyses into untargeted experiments to assist in identification workflows. The predictive abilities of this Compendium will improve in specificity and expand to more chemical classes as additional data from the IM-MS community is contributed. Instructions for data submission to the Compendium and criteria for inclusion are provided. .

Section C1. Ion Mobility Peak Annotation

In data sets in which multiple ion mobility peaks were observed for a single analyte, mobility peaks were annotated by assigning a peak number to each mobility peak. Peak number assignments begin with “1”, which refers to the smallest observed CCS or shortest drift time. Each subsequent mobility peak is assigned in numerical order (2, 3, etc.). If only one peak is observed, the CCS value is assigned a ”1”.

An example is shown in Figure S3.1. If only one peak is observed, the peak is assigned the number “1”.

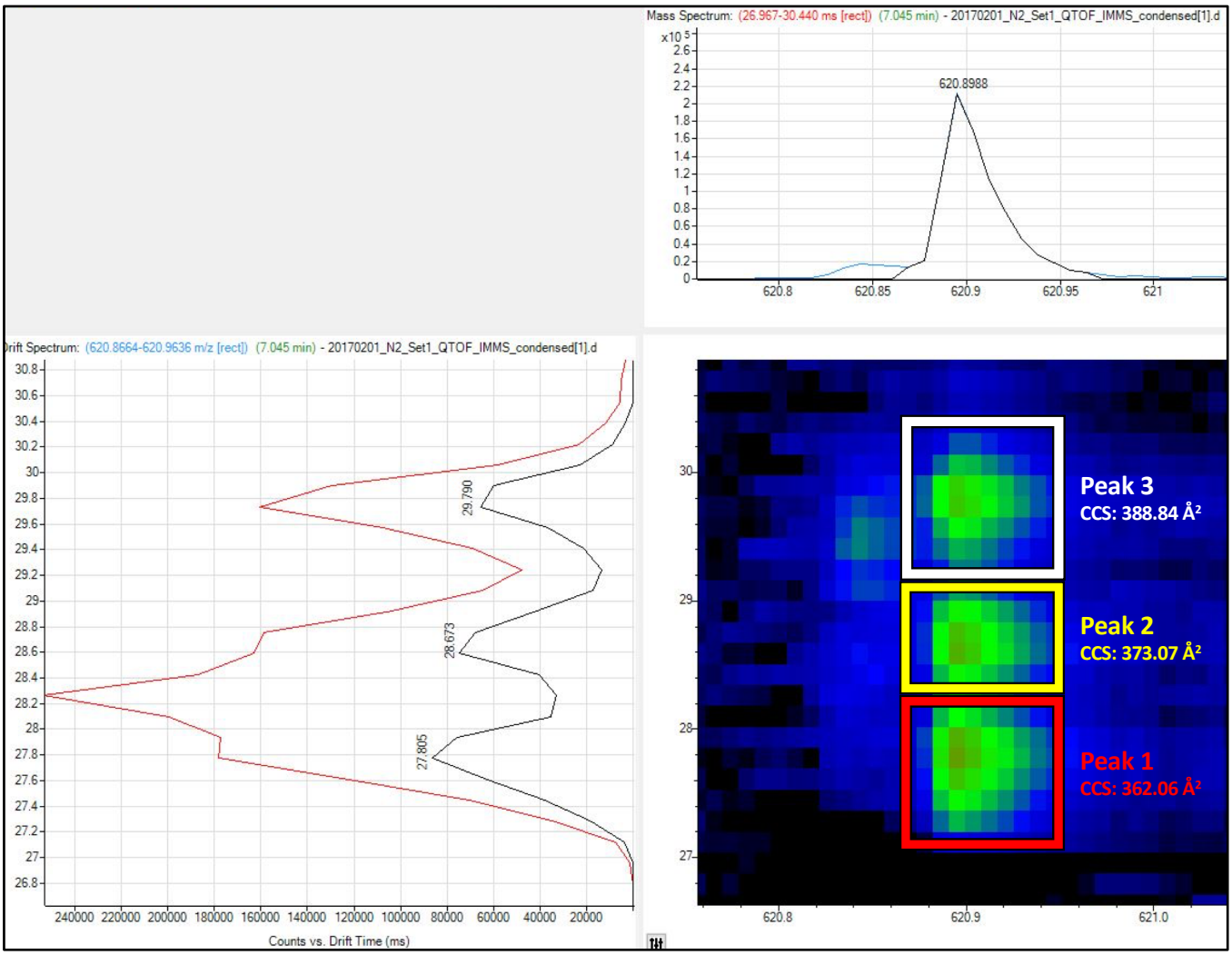


Figure C3.1 Illustration of CCS value annotations for analytes with multiple mobility peaks

Section C2. Data Inclusion Criteria

The Unified CCS Compendium is anticipated to be a collaborative effort of the IM-MS community; and the authors would like to invite contributions to this open-access repository for quality-controlled CCS measurements. Contributions towards the Unified CCS Compendium will improve informatics tools within the Compendium to aid in IM-MS based multiomic analyte identification workflows. For consistency, please follow the guidelines below. These guidelines are aimed at standardizing the data submission process and will expedite data quality assessment. Please note that these guidelines are subject to change; and the most up-to-date procedures can be downloaded from the online Compendium.[1] Currently, only the submission of CCS data obtained from drift tube measurements is accepted. In the future, it is anticipated that the Compendium will be expanded to support CCS measurements obtained from other IM techniques.

Guidelines for data submission into the Unified CCS Compendium

Single Field Data

The supplemental information packet includes a file entitled "SI_SingleField_DataFormat.xlsx". The two spreadsheets ("Single Field Reference Standards" and "Single Field Data Format") within the Excel file will need to be populated prior to submission of single field data to the Unified CCS Compendium. *Caution: If the Excel file is opened in read-only mode, the spreadsheet will not be editable. Please click 'enable editing' to proceed.*

Step 1: Collect, at minimum, triplicate measurements of the reference standards for each day samples are acquired.

- Recommended strategy: Infuse reference standards simultaneously with the analyte(s) of interest (i.e., as an internal reference). This allows the reference standards to be measured under the same conditions that the analytes are exposed to.

- If measuring reference standards independently, it is advised to acquire CCS measurements of the reference standards before, during, and after each set of acquisitions to assess any systematic deviations in mass and mobility measurements. This will also allow profiles of pressure, temperature, and electric field to be constructed for each acquisition set, to assist in assessing measurement quality.

Step 2: Collect, at minimum, triplicate measurements of the experimental analytes of interest (if not acquired in step 1). Include ≥ 5 compounds from the quality assessment (QA) compounds list (Table S1g) to assess data.

- Experimental values for analytes chosen from the QA compounds list must meet the following criteria:
  - Average CCS percent error of ≤ 0.5%

$$\text{percent error} = \frac{(\text{CCS}_{\text{experimental}} - \text{CCS}_{\text{QA}})}{\text{CCS}_{\text{QA}}} \cdot 100$$

  - Maximum individual CCS percent error ≤ 1%

Step 3: Populate columns A-F in the spreadsheet entitled "Single Field Reference Standards" (see Fig. S2) with experimental data generated from step 1 for each replicate.

- Use rows 5-14 for positive ion data and/or rows 15-24 for negative ion data.
- Single-field CCS and corresponding *m/z* values can be obtained using the "CCS Calibration (Single-Field)" method implemented in IM-MS Browser (Agilent Technologies). Alternately, the single-field drift time/CCS relationship can be calculated directly from drift time measurements of reference standards using equations described in previous work.[2]
- Columns G-O in the spreadsheet will be auto-populated. Important: The data in this spreadsheet is intended to be used ONLY for reference standards that were measured, NOT the analytes being submitted to the CCS Compendium.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
4	Experimental Replicate 1 *m/z*	Experimental Replicate 2 *m/z*	Experimental Replicate 3 *m/z*	Experimental Replicate 1 CCS	Experimental Replicate 2 CCS	Experimental Replicate 3 CCS	Reference Standards *m/z*	Reference Standards CCS	Experimental Average *m/z*	*M/z* error (ppm)	Experimental Average CCS	CCS Std Dev	CCS % RSD	% CCS Difference	Polarity
5	118.086	118.088	118.085	121.40	121.30	121.32	118.086	121.30	118.09	2.82	121.34	0.05	0.04%	0.03%	+
6	322.048	322.050	322.047	153.80	153.73	153.75	322.048	153.73	322.05	1.04	153.76	0.04	0.02%	0.02%	+
7	622.029	622.031	622.028	203.00	202.96	202.98	622.029	202.96	622.03	0.54	202.98	0.02	0.01%	0.01%	+
8	922.010	922.012	922.009	243.70	243.64	243.66	922.010	243.64	922.01	0.36	243.67	0.03	0.01%	0.01%	+
9	1221.991	1221.993	1221.990	282.30	282.20	282.22	1221.991	282.20	1221.99	0.27	282.24	0.05	0.02%	0.01%	+
10	1521.971	1521.973	1521.970	317.10	316.96	316.98	1521.971	316.96	1521.97	0.22	317.01	0.08	0.02%	0.02%	+
11							1821.952	351.25							+
12							2121.933	383.03							+
13							2421.914	412.96							+
14							2721.895	441.21							+
15							112.986	108.23							-
16							301.998	140.04							-
17							601.979	180.77							-
18							1033.969	255.34							-
19							1333.969	284.76							-
20							1633.950	319.03							-
21							1933.931	352.55							-
22							2233.911	380.74							-
23							2533.892	412.99							-
24							2833.873	432.62							-
25											Averages:	0.04	0.02%	0.02%	

Figure C3.2. "Single Field Reference Standards" spreadsheet

Step 4: Populate Columns A-K and Column Q of the spreadsheet entitled "Single Field Data Format" (see Fig. S3) with experimental data acquired in step 2.

- Columns L-P will be auto-populated.
- Assign peak numbers in Column Q as necessary: smallest CCS = 1, next smallest = 2, etc. If only one peak is observed, assign a "1".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Compound	Formula	CAS	Adduct	Charge	Experimental Replicate 1 m/z	Experimental Replicate 2 m/z	Experimental Replicate 3 m/z	Experimental Replicate 1 CCS	Experimental Replicate 2 CCS	Experimental Replicate 3 CCS	Average Experimental m/z	Average Experimental CCS	Std. Dev	%RSD	CCS/z	Peak Number
2	Example Lipid	C45H73NO8P	5634-86-6	[M-H]	-1	786.5070	786.5074	786.5109	277.30	277.33	277.27	μ m/z	μ CCS	σ	=σ/μCCS*100	277.30	1
3	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	361.68	361.76	362.75	620.91	362.06	0.60	0.17	181.03	1
4	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	373.66	373.09	372.45	620.91	373.07	0.60	0.16	186.53	2
5	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	388.97	389.27	388.27	620.91	388.84	0.51	0.13	194.42	3
6	Your	Data	Here!														

Figure C3.3. "Single Field Data Format" spreadsheet (Columns A-K)

Step 5: Classify each compound using the ClassyFire web application (found at http://classyfire.wishartlab.com/).[3]

- Populate Columns R-W of the spreadsheet entitled "Single Field Data Format" (shown in Fig. S4) with the classification information, source (e.g. research group), and DOI (if published).

	A	B	C	D	E	R	S	T	U	V	W
1	**Compound**	**Formula**	**CAS**	**Adduct**	**Charge**	**Kingdom**	**Super.Class**	**Class**	**Subclass**	**Source**	**DOI**
2	Example Lipid	C45H73NO8P	5634-86-6	[M-H]	-1	Organic compounds	Lipids and lipid-like molecules	Glycerophospholipids	Glycerophosphoethanolamines	Research Group	If unpublished, please fill in as "Unpublished".
3	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
4	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
5	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
6	**Your**	**Data**	**Here!**			**Your**	**Classifications**	**Here!**			

Figure C3.4. "Single Field Data Format" spreadsheet (Columns R-W)

Step 6: Calculate the average RSD for *all* experimental values (QA compounds as well as all analytes/compounds that are being submitted for inclusion into the Unified CCS Compendium).

- The CCS values submitted must meet the following criteria:
  - Average RSD $\leq$ 0.5% for all experimental data set
  - Individual compound RSD $\leq$ 0.7%

Step 7: Double check to ensure that steps 1-6 were performed.

- Step 1: Triplicate measurements acquired for reference standards each day sample measurements were collected.
- Step 2: Triplicate measurements acquired for ***all*** experimental values, including at least five compounds from the QA compound list.
- Step 3 & 4: Enter all data into the formatted spreadsheets.
- Step 5: Classify all compounds in the provided columns of the spreadsheets.
- Step 6: Calculate average RSD and individual RSDs.

Step 8: Submit spreadsheet for quality assessment.

Data must be submitted by emailing the completed spreadsheet ("SI_SingleField_DataFormat.xlsx") to ccscompendium@vanderbilt.edu.

- Please include the following information with each submission.
  - Institution

- Research group
- Instrument source type
- Solvent/buffer system
- List of reference compounds included in experimental data set

Upon data submission, the data will temporarily be quarantined and a quality control assessment will be performed. The quality control assessment includes: (1) verifying that all inclusion criteria is met, (2) confirming that all pertinent information is provided, and (3) checking that data is formatted properly. After the authors have processed a dataset (typically less than 10 days), collaborators will be notified which values will be accepted or if any revisions are needed. Data will be made available as soon as the quality control assessment is complete.

Stepped Field Data

The supplemental information packet includes a file entitled "SI_SteppedField_ScaleAndDataFormat.xlsx". The two spreadsheets ("Stepped Field Reference Standards and Scale" and "Stepped Field Data Format") within the Excel file will need to be populated prior to submission of stepped field data to the Unified CCS Compendium. *Caution: If the Excel file is opened in read-only mode, the spreadsheet will not be editable. Please click 'enable editing' to proceed.*

Step 1: Collect, at minimum, triplicate measurements of the reference standards for each day samples are acquired.

- Recommended strategy: Infuse reference standards simultaneously with the analyte(s) of interest (i.e., as an internal reference). This allows the reference standards to be measured under the same conditions that the analytes are exposed to.
- If measuring reference standards independently, it is advised to acquire CCS measurements of the reference standards before, during, and after each set of acquisitions to assess any systematic deviations in mass and mobility measurements. This will also allow profiles of pressure, temperature, and electric field to be constructed for each acquisition set, to assist in assessing measurement quality.

Step 2: Collect, at minimum, triplicate measurements of the experimental analytes of interest (if not acquired in step 1). Include ≥ 5 compounds from the quality assessment (QA) compounds list (Table S1) to assess data quality.

- Experimental values for analytes chosen from the QA compounds list must meet the following criteria:
  - Average CCS percent error of ≤ 0.5%

$$\text{percent error} = \frac{(\text{CCS}_{\text{experimental}} - \text{CCS}_{\text{QA}})}{\text{CCS}_{\text{QA}}} \cdot 100$$

  - Maximum individual CCS percent error ≤ 1%

Step 3: Populate columns A-F in the spreadsheet entitled "Stepped Field Reference Standards and Scale" (see Fig. S5) with data generated from step 1 for each replicate.

- True effective lengths for data collected in Step 1 must be calculated using the "Stepped Field Reference Standards and Scale" spreadsheet. Further detail addressing the purpose of scaling as well as the scaling procedure are discussed in supplemental Section S3.
- Use rows 7-16 for positive ion mode and/or rows 17-26 for negative ion mode.
- CCS and *m/z* values can be obtained using the "CCS Calculator (Stepped-Field)" method in IM-MS Browser (Agilent Technologies). Alternately, stepped-field CCS values can be calculated from corrected drift times using the fundamental low-field ion mobility equation.[4,5] Drift time correction requires a linear regression analysis incorporating the raw drift time measured at each of the drift fields surveyed, as described previously.[6]
- The experimental effective length (in cm) needs to be entered in the yellow box (Cell D4) located at the top of the spreadsheet. This length can be found in the "BaseDataAccess.dll.config" file located in the Mass Hunter Workstation (Agilent Technologies) install directory (typically: C Drive > Program Files > Agilent > MassHunter > Workstation > IMS > B.07.02 > Bin). Alternately, this is the length value used in the initial CCS calculation that is to be scaled.
- Columns G-P in the spreadsheet will be auto-populated.
- Important: The data in this spreadsheet is ONLY for reference standards that were measured, NOT the analytes being submitted to the CCS Compendium.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
4	Experimental Effective Length (cm):			78.24												
5																
6	Experimental Replicate 1 *m/z*	Experimental Replicate 2 *m/z*	Experimental Replicate 3 *m/z*	Experimental Replicate 1 CCS	Experimental Replicate 2 CCS	Experimental Replicate 3 CCS	Reference Standard *m/z*	Reference Standard CCS	Experimental Average *m/z*	*M/z* error (ppm)	Experimental Average CCS	CCS Std Dev	CCS % RSD	Scale Factor	New Effective Length	Polarity
7	118.086	118.088	118.085	121.40	121.30	121.32	118.086	121.30	118.09	2.82	121.34	0.05	0.04%	1.000164867	78.25289918	+
8	322.048	322.050	322.047	152.95	152.95	152.78	322.048	153.73	322.05	1.04	152.89	0.10	0.07%	0.997278864	78.02709834	+
9	622.029	622.031	622.028	202.44	202.48	202.29	622.029	202.96	622.03	0.54	202.40	0.10	0.05%	0.998627769	78.13263665	+
10	922.010	922.012	922.009	242.09	242.42	242.34	922.010	243.64	922.01	0.36	242.28	0.17	0.07%	0.997213979	78.02202172	+
11	1221.991	1221.993	1221.990	280.82	280.84	280.77	1221.991	282.20	1221.99	0.27	280.81	0.03	0.01%	0.997533326	78.04700741	+
12	1521.971	1521.973	1521.970	316.34	316.40	316.42	1521.971	316.96	1521.97	0.22	316.39	0.04	0.01%	0.999095148	78.1692044	+
13							1821.952	351.25								+
14							2121.933	383.03								+
15							2421.914	412.96								+
16							2721.895	441.21								+
17							112.986	108.23								-
18							301.998	140.04								-
19							601.979	180.77								-
20							1033.969	255.34								-
21							1333.969	284.76								-
22							1633.950	319.03								-
23							1933.931	352.55								-
24							2233.911	380.74								-
25							2533.892	412.99								-
26							2833.873	432.62								-
27											Averages:	0.08	0.04%	Averages:	78.10847795	

Figure C3.5. "Stepped Field Reference Standards and Scale" spreadsheet

Step 4: Populate Columns A-K and Column R of the spreadsheet entitled "Stepped Field Data Format" with experimental data acquired in step 2 (see Fig. S6).

- Columns L-Q will be auto-populated.
- Assign peak numbers in Column R as necessary: smallest CCS = 1, next smallest = 2, etc. If only one peak is observed, assign a "1".

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Compound	Formula	CAS	Adduct	Charge	Experimental Replicate 1 *m/z*	Experimental Replicate 2 *m/z*	Experimental Replicate 3 *m/z*	Experimental Replicate 1 CCS	Experimental Replicate 2 CCS	Experimental Replicate 3 CCS	Average Experimental *m/z*	Average Experimental CCS	Std. Dev	% RSD	Scaled CCS	CCS/z	Peak Number
2	Example Lipid	C45H73NO8P	5634-86-6	[M-H]	-1	786.5070	786.5074	786.5109	277.30	277.33	277.27	μ *m/z*	μ CCS	σ	=σ/μCCS*100	=μ*(old effective length/ new effective length)^2	277.30	1
3	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	361.68	361.76	362.75	620.91	362.06	0.60	0.17	363.28	181.03	1
4	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	373.66	373.09	372.45	620.91	373.07	0.60	0.16	374.32	186.53	2
5	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	388.97	389.27	388.27	620.91	388.84	0.51	0.13	390.15	194.42	3
6	Your	Data	Here!															

Figure C3.6. "Stepped Field Data Format" spreadsheet (Columns A-R)

Step 5: Classify each compound using the ClassyFire web application (found at http://classyfire.wishartlab.com/).[3]

- Populate Columns S-X of the spreadsheet entitled "Stepped Field Data Format" (Fig. S7) with the classification information, source (e.g. research group), and DOI (if published).

	A	B	C	D	E	S	T	U	V	W	X
1	Compound	Formula	CAS	Adduct	Charge	Kingdom	Super.Class	Class	Subclass	Source	DOI
2	Example Lipid	C45H73NO8P	5634-86-6	[M-H]	-1	Organic compounds	Lipids and lipid-like molecules	Glycerophospholipids	Glycerophosphoethanolamines	Research Group	If unpublished, please fill in as "unpublished".
3	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
4	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
5	Cyclosporin	C62H111N11O12		[M+H+K]	+2	Organic compounds	Organic polymers	Polypeptides		McLean	Unpublished
6	**Your**	**Data**	**Here!**			**Your**	**Classifications**	**Here!**			

Figure C3.7. "Stepped Field Data Format" spreadsheet (Columns S-X)

Step 6: Calculate the average RSD for *all* experimental values (QA compounds as well as all analytes/compounds that are being submitted for inclusion into the unified CCS compendium).

- The CCS values submitted must meet the following criteria:
  - Average RSD $\leq$ 0.5% for all experimental data set
  - Individual compound RSD $\leq$ 0.7%

Step 7: Check to ensure that steps 1-6 were performed.

- Step 1: At minimum, triplicate measurements were acquired for reference standards for each day that sample measurements were collected.
- Step 2: At minimum, triplicate measurements were acquired for ***all*** experimental values, including at least five compounds from the QA compound list.
- Step 3 & 4: Enter all data into the formatted spreadsheets.
- Step 5: Classify all compounds in the provided columns of the spreadsheets.
- Step 6: Calculate average RSD and individual RSD.

Step 8: Submit spreadsheet for quality assessment.

Data must be submitted by emailing the completed spreadsheet ("SI_SteppedField_ScaleAndDataFormat.xlsx") to ccscompendium@vanderbilt.edu.

- Please include the following information with each submission.

- Institution
- Research group
- Instrument source type
- Solvent/buffer system
- List of reference compounds included in experimental data set

Upon data submission, the data will temporarily be quarantined and a quality control assessment will be performed. The quality control assessment includes: (1) verifying that all inclusion criteria is met, (2) confirming that all pertinent information is provided, and (3) checking that data is formatted properly. After the authors have processed a dataset (typically less than 10 days), collaborators will be notified which values will be accepted or if any revisions are needed. Data will be made available as soon as the quality control assessment is complete.

Table C3.1. Quality Assessment (QA) Compound List

Standard reference CCS values obtained on a specially-modified drift tube instrument as previously reported.[2]

Compound	*m/z*	Ion Species	Stepped Field CCS (Å^2)	Single Field CCS (Å^2)
Small Molecules				
Cortisol	363.22	M+H	189.27 ± 0.10	188.34 ± 0.00
Cortisol	385.20	M+Na	213.72 ± 0.00	212.79 ± 0.07
Creatinine	112.05	M-H	120.69 ± 0.15	118.84 ± 0.07
Creatinine	114.07	M+H	123.86 ± 0.00	122.98 ± 0.02
Creatinine	136.05	M+Na	132.99 ± 0.35	132.61 ± 0.36
Glucose	203.05	M+Na	147.34 ± 0.29	146.94 ± 0.07
Homocysteine	136.04	M+H	130.77 ± 0.05	129.58 ± 0.63
L-arginine	173.10	M-H	138.03 ± 0.05	137.08 ± 0.01
L-arginine	175.12	M+H	136.84 ± 0.05	136.45 ± 0.00
L-aspartic acid	132.03	M-H	120.39 ± 0.40	119.15 ± 0.04
L-cystine	239.02	M-H	144.38 ± 0.09	143.58 ± 0.01
L-cystine	241.03	M+H	150.07 ± 0.05	149.48 ± 0.03
L-cystine	263.01	M+Na	151.81 ± 0.10	151.26 ± 0.13
L-glutamic acid	146.05	M-H	125.65 ± 0.15	124.47 ± 0.00
L-histidine	154.06	M-H	130.01 ± 0.09	128.83 ± 0.00
L-histidine	156.08	M+H	132.74 ± 0.11	131.93 ± 0.02
L-histidine	178.06	M+Na	135.47 ± 0.50	134.39 ± 0.44

L-isoleucine	130.09	M-H	131.28 ± 0.05	129.83 ± 0.01
L-isoleucine	132.10	M+H	133.81 ± 0.04	132.88 ± 0.03
L-leucine	130.09	M-H	132.51 ± 0.01	131.14 ± 0.00
L-leucine	132.10	M+H	135.55 ± 0.06	134.57 ± 0.03
L-lysine	147.11	M+H	131.62 ± 0.52	131.22 ± 0.14
L-methionine	150.06	M+H	134.07 ± 0.40	133.02 ± 0.47
L-phenylalanine	164.07	M-H	141.29 ± 0.19	139.94 ± 0.03
L-phenylalanine	166.09	M+H	141.27 ± 0.05	140.30 ± 0.12
L-proline	116.07	M+H	126.21 ± 0.20	125.38 ± 0.08
L-tyrosine	180.07	M-H	145.58 ± 0.34	144.42 ± 0.07
L-tyrosine	182.08	M+H	146.44 ± 0.20	145.58 ± 0.12
Levomefolic Acid	458.18	M-H	200.56 ± 0.11	198.99 ± 0.01
Levomefolic Acid	460.19	M+H	197.52 ± 0.26	197.17 ± 0.04
Pyridoxal Phosphate	246.02	M-H	150.80 ± 0.10	149.35 ± 0.04
Pyridoxal Phosphate	248.03	M+H	151.94 ± 0.10	151.37 ± 0.02
Pyridoxal Phosphate	270.01	M+Na	161.40 ± 0.20	161.46 ± 0.20
Uric Acid	167.02	M-H	126.92 ± 0.05	125.55 ± 0.07
Peptides				
Angiotensin1	1296.69	M+H	357.31 ± 0.26	355.62 ± 0.41
Angiotensin1	648.85	M+2H	387.29 ± 0.20	388.41 ± 0.10
Angiotensin1	432.90	M+3H	474.70 ± 0.15	477.05 ± 0.04
Angiotensin1	324.93	M+4H	549.23 ± 0.05	550.98 ± 0.07
Angiotensin2	1046.54	M+H	314.38 ± 0.15	313.65 ± 0.03
Angiotensin2	523.78	M+2H	353.79 ± 0.17	355.09 ± 0.03

Angiotensin2	349.52	M+3H	436.23 ± 0.20	437.30 ± 0.12
Bradykinin	1060.57	M+H	315.25 ± 0.30	314.00 ± 0.12
Bradykinin	530.79	M+2H	343.32 ± 0.10	344.99 ± 0.03
Compound	***m/z***	**Ion Species**	**Stepped Field CCS (Å^2)**	**Single Field CCS (Å^2)**
Bradykinin	354.19	M+3H	447.60 ± 0.11	449.07 ± 0.38
Melittin	1423.38	M+2H	613.36 ± 0.11	614.26 ± 0.02
Melittin	949.26	M+3H	721.06 ± 0.53	722.45 ± 0.02
Melittin	712.20	M+4H	756.78 ± 0.53	760.82 ± 0.12
Melittin	569.96	M+5H	808.60 ± 0.60	815.39 ± 0.10
Melittin	569.96	M+5H	844.39 ± 0.25	854.37 ± 0.15
Neurotensin	836.96	M+2H	434.32 ± 0.20	435.42 ± 0.06
Renin Substrate	879.97	M+2H	460.38 ± 0.40	461.11 ± 0.03
Renin Substrate	586.98	M+3H	518.81 ± 0.36	524.12 ± 0.07
Renin Substrate	440.49	M+4H	634.59 ± 0.35	637.65 ± 0.23
Substance P	1347.74	M+H	362.51 ± 0.20	361.44 ± 0.04
Substance P	674.37	M+2H	399.87 ± 0.20	400.09 ± 0.05
Substance P	449.92	M+3H	495.73 ± 1.29	496.51 ± 0.37
Proteins				
Cytochrome C	773.39	M+16H	3403.2 ± 2.10	3420.2 ± 2.38
Cytochrome C	727.96	M+17H	3538.1 ± 0.28	3554.7 ± 0.70
Cytochrome C	687.57	M+18H	3655.3 ± 1.57	3670.4 ± 0.74
Cytochrome C	651.44	M+19H	3741.8 ± 0.82	3757.9 ± 0.00
Cytochrome C	618.92	M+20H	3816.1 ± 0.79	3832.3 ± 0.00
Ubiquitin	856.98	M+10H	2192.3 ± 0.60	2204.8 ± 0.41

Ubiquitin	779.16	M+11H	2349.1 ± 0.77	2362.3 ± 0.00
Ubiquitin	714.32	M+12H	2424.6 ± 0.88	2444.2 ± 00
Ubiquitin	659.45	M+13H	2577.7 ± 0.63	2594.3 ± 0.53
Ubiquitin	612.41	M+14H	2727.4 ± 4.94	2728.8 ± 1.74
Ubiquitin	1223.80	M+7	1773.2 ± 1.26	1785.4 ± 0.29
Ubiquitin	1223.80	M+7	1875.7 ± 1.03	1884.3 ± 0.29
Ubiquitin	1070.96	M+8	1950.9 ± 0.24	1960.5 ± 0.33
Ubiquitin	952.08	M+9	2052.4 ± 0.64	2063.4 ± 0.00

Section C3.3. True Effective Length Calculation and CCS Scaling to Reference Values for Stepped Field DTIMS

The reference CCS values summarized in Table S1g were measured on a drift tube modified with grids at the entrance and exit of the drift region to minimize electric field penetration effects (i.e., fringing fields);[7,8] **and allow for CCS calculations to be performed using a precise, geometric length of the drift tube. Thus, these values are considered Standard Reference Values for purposes of accuracy comparisons. Commercially-available drift tubes are prone to drift field inhomogeneity and imprecise determination of the drift length, and therefore, a true effective length must be calculated in order for CCS measurements to correspond to the Standard Reference Values. More detailed information on the sources of measurement variability and a propagation of uncertainty for drift tubes is discussed elsewhere.**[2]

The supplemental information packet includes a file entitled "SI_SteppedField_ScaleAndDataFormat.xlsx". The two spreadsheets ("Stepped Field Reference Standards and Scale" and "Stepped Field Data Format") within the Excel file will need to be populated prior to submission of stepped field data to the unified CCS Compendium. *Caution: If the Excel file is opened in read-only mode, the spreadsheet will not be editable. Please click 'enable editing' to proceed.*

The following true effective length calculation and scaling procedure must be followed for each acquisition period (at least one reference standards calibration and true effective length calculation per day of data acquisition).

Step 1: Collect at minimum, triplicate measurements of the reference standards for each day samples are acquired.

- Ideal strategy: Infuse reference standards simultaneously with the analyte(s) of interest. This allows the reference standards to be observed in the same conditions that analytes are exposed to.

- If measuring reference standards independently, acquire reference standards CCS measurements before, during, and after each set of acquisition to assess drift in mass and mobility measurements.

Step 2: Populate columns A and B in the spreadsheet entitled "Stepped Field Reference Standards and Scale" (see Fig. S8) with data generated from step 1 for each replicate.

- Use rows 7-16 for positive ion mode and/or rows 17-26 for negative ion mode.
- CCS and *m/z* values can be obtained using the "CCS Calculator (Stepped-Field)" method in IM-MS Browser (Agilent Technologies). Alternately, these can be calculated from the fundamental equation as discussed in the previous section.
- The experimental effective length (in cm) needs to be entered in the yellow box (Cell D4) located at the top of the spreadsheet. This length can be found in the "BaseDataAccess.dll.config" file located in the Mass Hunter Workstation (Agilent Technologies) install directory (typically: C Drive > Program Files > Agilent > MassHunter > Workstation > IMS > B.07.02 > Bin). Alternately, this is the length value used in the initial CCS calculation that is to be scaled.

Step 3: Columns C-G will be auto-populated.

- The scale factor in column F is calculated for each reference compound CCS using the following equation:

$$\text{scale factor} = \sqrt{\frac{\text{experimental CCS}}{\text{reference standard CCS}}}$$

- New effective length values for each reference standards ion in Column G are calculated using the equation:

$$\text{new effective length} = \text{scale factor} * \text{experimental effective length}$$

Step 4: The true effective length (yellow cell G27, at the bottom of the "New Effective Length" column) is automatically calculated.

- This value is the average of all calculated effective lengths originating from each reference standards ion.
- The true effective length value is automatically carried over to the "Stepped Field Data Format spreadsheet" to scale experimental data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
4	**Experimental Effective Length (cm):**			78.24												
5																
6	**Experimental Replicate 1 *m/z***	**Experimental Replicate 2 *m/z***	**Experimental Replicate 3 *m/z***	**Experimental Replicate 1 CCS**	**Experimental Replicate 2 CCS**	**Experimental Replicate 3 CCS**	**Reference Standard *m/z***	**Reference Standard CCS**	**Experimental Average *m/z***	***M/z* error (ppm)**	**Experimental Average CCS**	**CCS Std Dev**	**CCS % RSD**	**Scale Factor**	**New Effective Length**	**Polarity**
7	118.086	118.088	118.085	121.40	121.30	121.32	118.086	121.30	118.09	2.82	121.34	0.05	0.04%	1.000164867	78.25289918	+
8	322.048	322.050	322.047	152.95	152.95	152.78	322.048	153.73	322.05	1.04	152.89	0.10	0.07%	0.997278864	78.02709834	+
9	622.029	622.031	622.028	202.44	202.48	202.29	622.029	202.96	622.03	0.54	202.40	0.10	0.05%	0.998627769	78.13263665	+
10	922.010	922.012	922.009	242.09	242.42	242.34	922.010	243.64	922.01	0.36	242.28	0.17	0.07%	0.997213979	78.02202172	+
11	1221.991	1221.993	1221.990	280.82	280.84	280.77	1221.991	282.20	1221.99	0.27	280.81	0.03	0.01%	0.997533326	78.04700741	+
12	1521.971	1521.973	1521.970	316.34	316.40	316.42	1521.971	316.96	1521.97	0.22	316.39	0.04	0.01%	0.999095148	78.1692044	+
13							1821.952	351.25								+
14							2121.933	383.03								+
15							2421.914	412.96								+
16							2721.895	441.21								+
17							112.986	108.23								-
18							301.998	140.04								-
19							601.979	180.77								-
20							1033.969	255.34								-
21							1333.969	284.76								-
22							1633.950	319.03								-
23							1933.931	352.55								-
24							2233.911	380.74								-
25							2533.892	412.99								-
26							2833.873	432.62								-
27											**Averages:**	**0.08**	**0.04%**	**Averages:**	**78.10847795**	

Figure C3.8. "Stepped Field Reference Standards and Scale" spreadsheet

Step 5: To scale experimental CCS values, follow steps 2-6 of the Stepped Field Data Guidelines for Data Submission (Section S1) before continuing.

- Scaled CCS values are auto-populated in column P using the equation:

$$\text{scaled CCS} = \text{experimental CCS} \cdot \left(\frac{\text{experimental effective length}}{\text{true effective length}}\right)^{\wedge 2}$$

- An example is shown in Fig. S9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	**Compound**	**Formula**	**CAS**	**Adduct**	**Charge**	**Experimental Replicate 1 *m/z***	**Experimental Replicate 2 *m/z***	**Experimental Replicate 3 *m/z***	**Experimental Replicate 1 CCS**	**Experimental Replicate 2 CCS**	**Experimental Replicate 3 CCS**	**Average Experimental *m/z***	**Average Experimental CCS**	**Std. Dev**	**% RSD**	**Scaled CCS**	**CCS/z**	**Peak Number**
2	Example Lipid	C45H73NO8P	5634-86-6	[M-H]	-1	786.5070	786.5074	786.5109	277.30	277.33	277.27	μ *m/z*	μ CCS	σ	=σ/μCCS*100	=μ*(old effective length/ new effective length)^2	277.30	1
3	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	361.68	361.76	362.75	620.91	362.06	0.60	0.17	363.28	181.03	1
4	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	373.66	373.09	372.45	620.91	373.07	0.60	0.16	374.32	186.53	2
5	Cyclosporin	C62H111N11O12		[M+H+K]	+2	620.9060	620.9068	620.9072	388.97	389.27	388.27	620.91	388.84	0.51	0.13	390.15	194.42	3
6	Your	Data	Here!															

Figure C3.9. "Stepped Field Data Format" spreadsheet (Columns A-Q)

Section C4

Table C3.2. All super classes and classes represented in the Unified CCS Compendium at date of submission

Super Class	Class	*m/z* Range	N
Alkaloids and derivatives	Yohimbine alkaloids	609	1
		138 – 164	3
Benzenoids	Anthracenes	178 – 271	9
	Benzene and substituted derivatives	108 – 886	159
	Fluorenes	166	1
	Indanes	300	1
	Naphthalenes	128 – 254	25
	Pentacenes	278, 280	2
	Phenanthrenes and derivatives	178 – 303	19
	Phenols	109 – 208	31
	Pyrenes	202 – 304	22
Homogeneous metal compounds	Homogeneous transition metal compounds	132 – 2991	62
Homogeneous non-metal compounds	Non-metal oxoanionic compounds	200	1
Lipids and lipid-like molecules	Fatty acyls	125 – 935	223
	Glycerolipids	253 – 746	8
	Glycerophospholipids	171 – 1017	334

	Prenol lipids	137 – 886	21
	Sphingolipids	548 – 8989	146
	Steroids and steroid derivatives	287 – 648	78
Nucleosides, nucleotides, and analogues	(5'->5')-dinucleotides	662 – 783	29
	5'-deoxyribonucleosides	250 – 408	19
	Flavin nucleotides	455 – 809	9
	Imidazole ribonucleosides and ribonucleotides	337 – 362	4
	Nucleoside and nucleotide analogues	24, 268	2
	Purine nucleosides	250 – 613	49
	Purine nucleotides	280 – 790	125
	Pyrimidine nucleosides	226 – 281	23
	Pyrimidine nucleotides	304 – 646	124
Organic acids and derivatives	Carboximidic acids and derivatives	131, 154	2
	Carboxylic acids and derivatives	89 – 2110	623
	Keto acids and derivatives	115 – 184	11
	Hydroxy acids and derivatives	103 – 239	12
	Organic carbonic acids and derivatives	155	1
	Organic phosphonic acids and derivatives	124 – 205	13

	Organic sulfonic acids and derivatives	124 – 213	4
	Peptidomimetics	225 – 1317	23
Super Class	**Class**	***m/z* Range**	**N**
Organic acids and derivatives	Proteins	493 – 3302	139
	Sulfinic acids and derivatives	108, 111	2
	Tryptic peptides	288 – 1580	254
Organic nitrogen compounds	Organonitrogen compounds	74 – 1233	101
Organic oxygen compounds	Organic oxoanionic compounds	227 – 411	6
	Organooxygen compounds	105 – 1505	340
Organic polymers	Cyclic Peptides	1111 – 1704	20
	Polypeptides	294 – 1724	230
Organohalogen compounds	Organofluorides	301 – 2834	66
Organoheterocyclic compounds	Azoles	127 – 458	12
	Benzimidazoles	145 – 225	4
	Benzodioxoles	191 – 272	4
	Benzopyrans	421, 424	2

	Dihydrofurans	173 – 350	6
	Dithiolanes	228	1
	Furofurans	199, 200	2
	Imidazopyrimidines	119 – 218	60
	Indoles and derivatives	148 – 381	69
	Lactams	348 – 738	10
	Lactones	153 – 350	6
	Naphthofurans	821 – 1684	14
	Pteridines and derivatives	162 – 483	28
	Pyridinecarboxylic acids and derivatives	140	1
	Pyridines and derivatives	96 – 285	52
	Pyrroles	110	1
	Quinolines and derivatives	172 – 431	17
	Tetrahydroisoquinolines	178 – 181	2
	Tetrapyrroles and derivatives	563 – 1378	9
	Triazines	215 – 325	8
Phenylpropanoids and polyketides	Anthracyclines	540 – 1320	6
	Cinnamaldehydes	133, 134	2
	Cinnamic acids and derivatives	149 – 360	5
	Coumarins and derivatives	161, 186	2

	Flavonoids	269 – 1424	44
	Isoflavonoids	140 – 418	16
	Linear 1,3-diarylpropanoids	255 – 280	4
	Macrolactams	786 - 825	3
	Macrolides and analogues	661– 955	15
	Phenylpropanoic acids	165 – 284	11
	Tetracyclines	410	8
Polyhedralcarbon molecules		720, 840	2

Section C5. Nonlinear Regression Equations

Power Fit

$$y = a \cdot x^{-k} + y_0 \tag{3,1}$$

a is the curve max – curve min; k is the curve rate

Four-Parameter Sigmoidal Fit (4P)

$$y = y_o + \frac{y_{max} - y_0}{1 + 10^{(\log y_{50} - x) \cdot H}} \tag{3,2}$$

y_{50} is x at curve half-maximum; H is the Hill Slope

Five-Parameter Sigmoidal Fit (5P)

$$y = y_o + \frac{y_{max} - y_0}{\left(1 + 10^{((\log y_{50} - x) \cdot H)}\right)^S} \tag{3,3}$$

S is the curve symmetry parameter

Confidence Interval

$$z \cdot s_{y,x} \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}\right)^{1/2} \tag{3,4}$$

z is standard deviations z score based on interval percentage

(z-score for 99% is 2.576);

$S_{y,x}$ is the standard error of the x and y data inputs;

SS_x is the sum of the squared deviations from the x input mean

Predictive Intervals

$$z \cdot s_{y,x} \cdot \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}\right)^{1/2} \quad (3,5)$$

Section C.6. Supplemental Experimental Methods: LC-MS and LC-IM-MS Acquisition Parameters

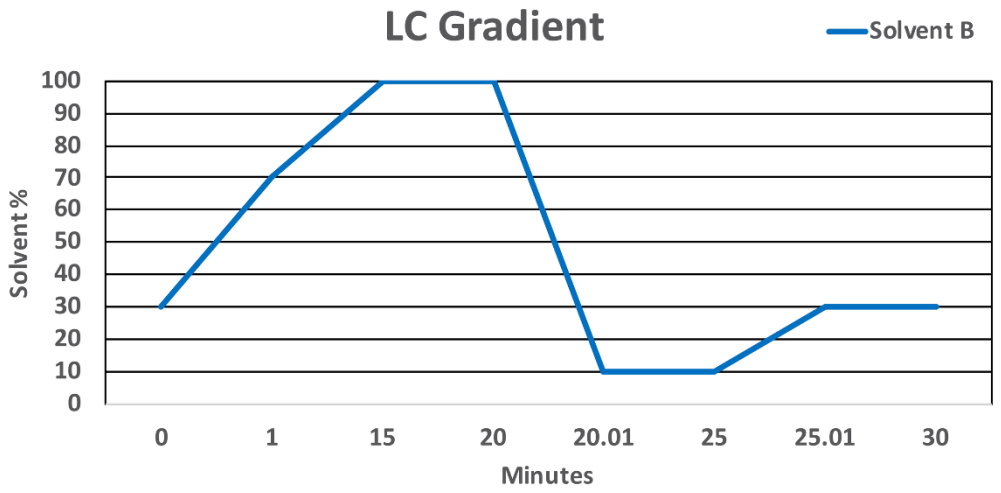


Figure S3.10. LC Gradient

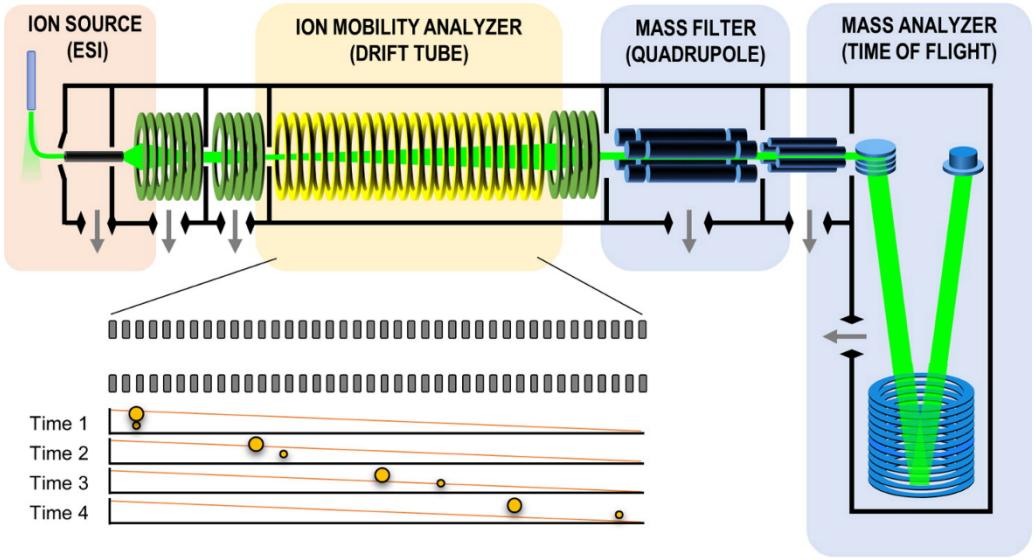


Figure C3.11. Instrument schematic from reference [6].

Serum samples were analyzed via liquid chromatographic separation using an C18 Zorbax RRHD (1.8µm) column on a 1290 Infinity LC system (Agilent Technologies). Solvent A was water with 0.1% formic acid; and Solvent B was 3:2 isopropanol:acetonitrile with 0.1% formic acid. 2 µl of sample were injected via autosampler and separations occurred using a 30 min gradient described

in Fig. S5a at 200 µl/min. Post-LC separation, analytes were ionized using an electrospray ionization source (Jet Stream, Agilent Technologies) at 300°C and a VCap voltage of 3500 V. The drying gas flow rate was 8 L/min, while the sheath gas flowed at 11 L/min. When data was acquired using LC-IM-MS mode, ion mobility separations were performed using a uniform field drift tube with high-purity nitrogen drift gas at 3.95 Torr at room temperature (~298 K). A single field analysis at 17.26 V/cm was performed on a standardized calibrant mixture (Agilent Tune Mix) to normalize sample drift times. Time-of-flight scan range was 100 *m/z* to 1700 *m/z*. Further information can be found in previous work.[2]

References

1 Mclean Research Group, CCS Compendium, https://www.vanderbilt.edu/AnS/Chemistry/groups/mcleanlab/ccs.html.

2 S. M. Stow, T. J. Causon, X. Zheng, R. T. Kurulugama, T. Mairinger, J. C. May, E. E. Rennie, E. S. Baker, R. D. Smith, J. A. McLean, S. Hann and J. C. Fjeldsted, *Anal. Chem.*, 2017, **89**, 9048–9055.

3 Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D. S. Wishart, *J. Cheminform.*, 2016, **8**, 1–20.

4 E. A. Mason and E. W. McDaniel, *Transport Properties of Ions in Gases*, John Wiley & Sons, Ltd., New York City, NY, 1988.

5 W. F. Siems, L. A. Viehland and H. H. Hill, *Anal. Chem.*, 2012, **84**, 9782–9791.

6 J. C. May, C. R. Goodwin, N. M. Lareau, K. L. Leaptrot, C. B. Morris, R. T. Kurulugama, A. Mordehai, C. Klein, W. Barry, E. Darland, G. Overney, K. Imatani, G. C. Stafford, J. C. Fjeldsted and J. A. McLean, *Anal. Chem.*, 2014, **86**, 2107–2116.

7 T. Wyttenbach, P. R. Kemper and M. T. Bowers, *Int J Mass Spectrom*, 2001, **212**, 13–23.

8 K. Kaplan, S. Graf, C. Tanner, M. Gonin, K. Fuhrer, R. Knochenmuss, P. Dwivedi and H. H. Hill, *Anal. Chem.*, 2010, **82**, 9336–9343.

APPENDIX D

Supplemental Information for Chapter IV

Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility – Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization

Jaqueline A. Picache, Jody C. May, John A. McLean*

Department of Chemistry, Center for Innovative Technology, Vanderbilt Institute of Chemical Biology, Vanderbilt Institute for Integrative Biosystems Research and Education, Vanderbilt-Ingram Cancer Center; Vanderbilt University, Nashville, Tennessee

Abstract

This work presents a machine learning algorithm referred to as the Supervised Inference of Feature Taxonomy from Ensemble Randomization (SIFTER), which supports the identification of features derived from untargeted ion mobility-mass spectrometry (IM-MS) experiments. SIFTER utilizes random forest machine learning on three analytical measurements derived from IM-MS (collision cross section (*CCS*), mass-to-charge (*m/z*), and mass defect (*Δm*)) to classify unknown features into a taxonomy of chemical kingdom, super class, class, and subclass. Each of these classifications is assigned a calculated probability as well as alternate classifications with associated probabilities. After optimization, SIFTER was tested against a set of molecules not used

in the training set. The average success rate in classifying all four taxonomy categories correctly was found to be >99%. Analysis of molecular features detected from a complex biological matrix and not used in the training set yielded a lower success rate where all four categories were correctly predicted for ~80% of the compounds. This decline in performance is in part due to incompleteness of the training set across all potential taxonomic categories, but also resulting from a nearest neighbor bias in the random forest algorithm. Ongoing efforts are focused on improving the class prediction accuracy of SIFTER through expansion of empirical datasets used for training as well as improvements to the core algorithm.

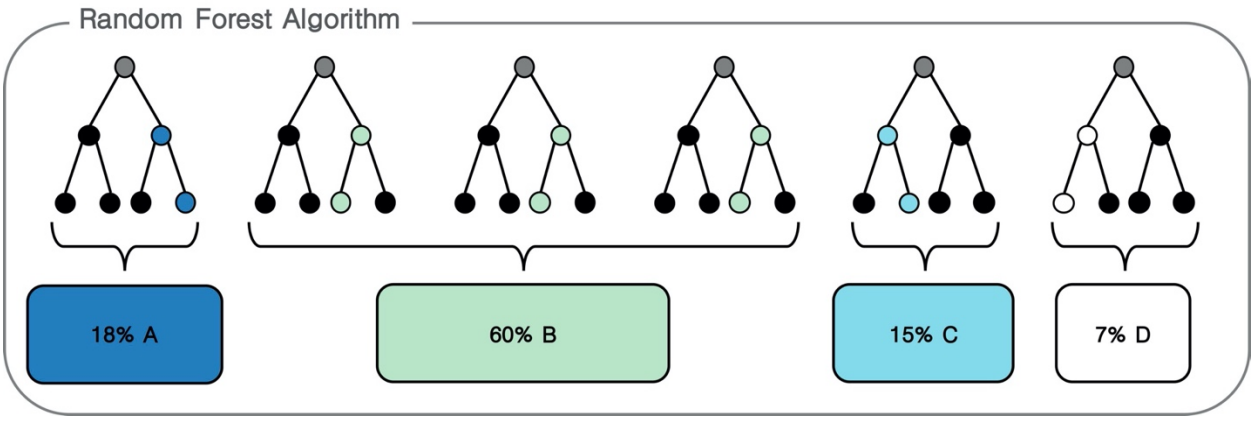


Figure D4.1 Random Forest Algorithm. Illustration of majority score voting in random forest algorithm decision tree process.

Table D4.1. Kingdom Confusion Matrix. Results of RF confusion matrix results which indicate the likelihood of a compound incorrectly being classified into a given category; N=30.

Kingdom Category	Average Error	Standard Deviation
Inorganic compounds	0.31	0.014
Organic compounds	0.00	0.000

Table D4.2 Super Class Confusion Matrix. Results of RF confusion matrix results which indicate the likelihood of a compound incorrectly being classified into a given category; N=30.

Super Class Categories	Average Error	Standard Deviation
Benzenoids	0.40	0.015
Lipids and lipid-like molecules	0.07	0.003
Mixed metal/non-metal compounds	0.20	0.024
Nucleosides, nucleotides, and analogues	0.09	0.011
Organic acids and derivatives	0.26	0.007
Organic nitrogen compounds	0.33	0.018
Organic oxygen compounds	0.36	0.016
Organic polymers	0.94	0.055
Organohalogen compounds	0.06	0.014
Organoheterocyclic compounds	0.62	0.019
Phenylpropanoids and polyketides	0.79	0.032

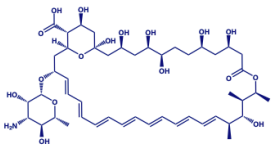
Table D4.3. Class Confusion Matrix. Results of RF confusion matrix results which indicate the likelihood of a compound incorrectly being classified into a given category; N=30.

Class Categories	Average Error	Standard Deviation
Alkali metal salts	0.00	0.000
Azoles	0.86	0.061
Benzene and substituted derivatives	0.04	0.009
Carboxylic acids and derivatives	0.02	0.003
Diazines	0.51	0.043
Fatty acyls	0.07	0.006
Flavonoids	0.00	0.000
Glycerophospholipids	0.07	0.004
Imidazopyrimidines	0.50	0.048
Indoles and derivatives	0.24	0.025
Organofluorides	0.00	0.007
Organonitrogen compounds	0.00	0.000
Organooxygen compounds	0.00	0.000
Peptidomimetics	0.41	0.024
Phenols	0.34	0.061
Purine nucleotides	0.21	0.018
Pyrenes	0.59	0.086
Pyridines and derivatives	0.59	0.019
Pyrimidine nucleotides	0.21	0.021
Quinolines and derivatives	0.61	0.049
Sphingolipids	0.10	0.007
Steroids and steroid derivatives	0.32	0.028

Table D4.4 Subclass Confusion Matrix. Results of RF confusion matrix results which indicate the likelihood of a compound incorrectly being classified into a given category; N=30.

Subclass Category	Average Error	Standard Deviation
Alcohols and polyols	0.77	0.000
Alkali metal iodides	0.00	0.000
Amines	0.18	0.027
Amino acids, peptides, and analogues	0.00	0.000
Anilides	0.38	0.042
Benzenediols	0.00	0.000
Benzenesulfonamides	0.35	0.000
Benzoic acids and derivatives	0.16	0.012
Benzopyrenes	0.00	0.000
Bile acids, alcohols and derivatives	0.09	0.000
Biphenyls and derivatives	0.03	0.010
Carbohydrates and carbohydrate conjugates	0.02	0.000
Carbonyl compounds	0.69	0.035
Ceramides	0.05	0.014
Cyclic purine nucleotides	0.33	0.036
Depsipeptides	0.02	0.008
Diphenylethers	0.07	0.000
Eicosanoids	0.14	0.025
Fatty acid esters	0.50	0.000
Fatty acids and conjugates	0.10	0.007
Glycerophosphates	0.46	0.024
Glycerophosphocholines	0.27	0.011
Glycerophosphoethanolamines	0.35	0.012
Glycerophosphoserines	0.21	0.004
Glycosphingolipids	0.25	0.018
Hybrid peptides	0.13	0.000
Hydroxysteroids	0.21	0.023
Imidazoles	0.00	0.000
Indoles	0.27	0.000
Indolyl carboxylic acids and derivatives	0.11	0.000
Lineolic acids and derivatives	0.23	0.025
O-methylated flavonoids	0.00	0.000
Phosphazene and phosphazene derivatives	0.00	0.000
Phosphosphingolipids	0.52	0.021

Table D4.4 Continued		
Pregnane steroids	0.60	0.044
Purine deoxyribonucleotides	0.95	0.000
Purine nucleotide sugars	0.06	0.000
Purine ribonucleotides	0.38	0.024
Purines and purine derivatives	0.00	0.000
Pyridine carboxaldehydes	0.30	0.000
Pyridinecarboxylic acids and derivatives	0.23	0.046
Pyrimidine deoxyribonucleotides	0.31	0.025
Pyrimidine nucleotide sugars	0.00	0.000
Pyrimidine ribonucleotides	0.19	0.028
Pyrimidines and pyrimidine derivatives	0.00	0.000
Quaternary ammonium salts	0.23	0.019
Quinoline carboxylic acids	0.00	0.000
Tricarboxylic acids and derivatives	0.79	0.067
Tryptamines and derivatives	0.53	0.030



Amphotericin B

Input:

m/z **: 946.478 Da** **CCS: 328.3 Å²** **KMD: -0.421**

	Classification	Prediction	Probability	Alternatives
Kingdom	Organic compounds	Organic compounds ✓	1.00	No alternative suggested.
Super Class	Organic oxygen compounds	Organic oxygen compounds ✓	1.00	No alternative suggested.
Class	Organooxygen compounds	Organooxygen compounds ✓	1.00	No alternative suggested.
Subclass	Carbohydrates and carbohydrate conjugates	Carbohydrates and carbohydrate conjugates ✓	1.00	No alternative suggested.

Figure D4.2 Large Molecule Example. Example where SIFTER classified the molecule Amphotericin B where 4/4 categories were correct and their associated probabilities are shown.

APPENDIX E

Curriculum Vitae

JAQUELINE-MAE A. PICACHE

411B 32nd Avenue South, Nashville, TN 37212 · Mobile: (201)916-3391
Email: jaqueline.picache@vanderbilt.edu

EDUCATION

Vanderbilt University, The Graduate School **Nashville, TN**
Doctorate of Philosophy, Chemistry Expected August 2020
Dissertation: Strategies for Improving Multiomic Metabolite Identifications Using Compound Libraries, Machine Learning, and Structural Tandem Mass Spectrometry

The University of Notre Dame, The College of Science **Notre Dame, IN**
Bachelor of Science: Science Preprofessional Studies; May 2014
Minor: Science, Technology, and Values

RESEARCH EXPERIENCE

Vanderbilt University **Nashville, TN**
Ph.D. Candidate, Dept. of Chemistry, Advisor: Dr. John A. McLean (July 2016 – Present)

National Institutes of Health **Bethesda, MD**
Post-baccalaureate Researcher, NICHD, Advisor: Dr. Forbes D. Porter (August 2014– June 2016)

The University of Notre Dame **Notre Dame, IN**
Undergraduate Clinical Researcher, Advisor: Dr. Kasturi Haldar (January 2013 – May 2014)
Undergraduate Researcher, Advisor: Dr. Holly Goodson (May 2013 – May 2014)
Undergraduate Researcher, Advisor: Dr. Lei Li (January 2011 – December 2011)

Columbia University Medical Center **New York, NY**
Physiology and Cell Biophysics Research Intern, Advisor: Dr. Wesley Grueber
(June 2008 – August 2011)

PUBLICATIONS

- Picache, J. A.; May, J. C.; McLean, J. A. Chemical Class Prediction of Unknown Biomolecules Using Ion Mobility-Mass Spectrometry and Machine Learning: Supervised Inference of Feature Taxonomy from Ensemble Randomization. *Anal Chem.* **2020**, Just accepted. DOI: 10.1021/acs.analchem.0c02137
- Picache, J. A.; May, J. C.; McLean, J. A. Crowd-sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses. *ACS Omega.* **2020**, *5* (2), 980-985. [PMCID: PMC6977078]
- Harris, R. A.*; Picache, J. A.*; Tomlinson, I. D.; Zlibut, E.; Ellis, B. M.; May, J. C.; McLean, J. A.; Hercules, D. M. Mass Spectrometry and Ion Mobility Study of Poly(theylene glycol)-based Polyurethane Oligomers. *Rapid Commun Mass Spectrom.* **2019**, E-published. DOI: 10.1002/rcm.8662. [PMID: 31731326] *Co-first Authors
- Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision Cross Section Compendium to Annotate and Predict Multi-Omic Compound Identities. *Chem. Sci.* **2019**, *10* (4), 983–993. [PMCID: PMC6349024]
- Nichols, C. M.; Dodds, J. N.; Rose, B. S.; Picache, J. A.; Morris, C. B.; Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A. Untargeted Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility Spectrometry. *Anal. Chem.* **2018**, *90* (24), 14484–14492. [PMCID: PMC30449086]
- Cougnoux, A.; Movassaghi, M.; Picache, J. A.; Iben, J. R.; Navid, F.; Salman, A.; Martin, K.; Farhat, N. Y.; Cluzeau, C.; Tseng, W.-C.; et al. Gastrointestinal Tract Pathology in a BALB/c Niemann–Pick Disease Type C1 Null Mouse Model. *Dig. Dis. Sci.* **2018**, *63* (4), 870–880. [PMCID: PMC6292218]
- Cologna, S. M.; Crutchfield, C. A.; Searle, B. C.; Blank, P. S.; Toth, C. L.; Ely, A. M.; Picache, J. A.; Backlund, P. S.; Wassif, C. A.; Porter, F. D.; et al. An Efficient Approach to Evaluate Reporter Ion Behavior from MALDI-MS/MS Data for Quantification Studies Using Isobaric Tags. *J. Proteome Res.* **2015**, *14* (10). [PMCID: PMC5571863]
- Picache, J. A.; Basson, C. T. Holt Oram Syndrome. *National Organization for Rare Disorders*. **2014**, https://rarediseases.org/rare-diseases/holt-oram-syndrome/.

ORAL PRESENTATIONS

- **Picache, J. A.;** Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Development of a Unified Collision Cross Section Compendium for Compound Annotation and Chemical Class Prediction. In *Vanderbilt Institute for Chemical Biology Annual Symposium;* Nashville, TN, 2019.
- **Picache, J. A.;** Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Development of a Unified Collision Cross Section Compendium for Compound Annotation and Chemical Class Prediction. In *67th ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, GA, 2019.
- **Picache, J. A.;** McLean, J. A. Achieving a Consistent and Unified Community-Wide Molecular Database: The CCS Compendium. In *5th International Ion Mobility Spectrometry Meeting*; Bordeaux, France, 2019.
- **Picache, J. A.;** Cologna, S. M.; Yergey, A. L.; Picard, P.; Burkert, K. R.; Wassif, C. A.; Zheng, W.; Porter, F. D. A Label-Free , Mass Spectrometry-Based High Throughput Candidate Drug Screening Assay : Application to Smith-Lemli-Opitz Syndrome (ID: 279235). In *64th ASMS Conference on Mass Spectrometry and Allied Topics*; San Antonio,

TX, 2016.

POSTER PRESENTATIONS

- **Picache, J. A.;** Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision Cross Section Compendium to Annotate and Predict Multi-Omic Compound Identities. In *Analytica Vietnam*; Analytica, Ho Chi Minh City, Vietnam, 2019.
- **Picache, J. A.;** Codreanu, S. G.; May, J. C.; Sherrod, S. D.; McLean, J. A. Towards Developing an Automated , Phenotype Driven , Multi-Parallel Sampling Device for Mass Spectrometry-Based Metabolomics. In *66th ASMS Conference on Mass Spectrometry and Allied Topics*; American Society for Mass Spectrometry: San Diego, CA, 2018.
- **Picache, J. A.;** Nichols, C. M.; Fjeldsted, J. C.; May, J. C.; Sherrod, S. D.; McLean, J. A. Visualizing and Conceptualizing Highly-Dimensional Metabolomics Data from LC-IM-MS/MS Analysis. In *VICB Student Symposium 2017*; Vanderbilt Insititue of Chemical Biology: Nashville, TN,2017.
- **Picache, J. A.;** Nichols, C. M.; Fjeldsted, J. C.; May, J. C.; Sherrod, S. D.; McLean, J. A. Visualizing and Conceptualizing Highly-Dimensional Metabolomics Data from LC-IM-MS/MS Analysis. In *65th ASMS Conference on Mass Spectrometry and Allied Topics*; American Society for Mass Spectrometry: Indianapolis, IN, 2017.
- **Picache J. A.** Holt-Oram Syndrome. In *The University of Notre Dame CRND Rare Disease Day*; Boler-Parseghian Center for Rare and Neglected Diseases: Notre Dame, IN, 2014.
- **Picache J. A.** and Goodson H. Characterization of microtubule branching patterns induced by the Adenomatous polyposis coli protein. In *The University of Notre Dame COS-JAM*; University of Notre Dame College of Science: Notre Dame, IN, 2013.

AWARDS, HONORS, and FELLOWSHIPS

- **Runner-up, VICB Richard Armstrong Prize for Research Excellence**, VICB Symposium (August 2019)
- **ASMS Annual Conference Graduate Student Award**, American Society for Mass Spectrometry (June 2019)
- **First Place, International Poster Award**, Analytica Vietnam (March 2019)
- **ASMS Fall Workshop Travel Award**, American Society for Mass Spectrometry (November 2018)
- **May Institute Student Travel Award**, May Institute at Northeastern University (May 2018)
- **Data Science Visions TIPS Student Travel Award**, Vanderbilt University (May 2018)
- **Vanderbilt Chemistry-Biology Interface Training Award**, Vanderbilt Institute for Chemical Biology (July 2018 - present)
- NIH sponsored V-CBI Training Grant 5T32GM065086
- **Vanderbilt University Graduate Student Travel Award**, Vanderbilt University (June 2017, June 2018, March 2019)
- **Mitchum Warren Fellowship,** Vanderbilt University (August 2016 – June 2017)
- **Post-baccalaureate Intramural Research Training Award,** National Institutes of Health (August 2014 – June 2016)
- **Harper Cancer Institute Summer Fellowship**, The University of Notre Dame (May 2013 – August 2013)
- **Questbridge Scholar** (August 2010 – May 2014)

- **National Semi-finalist, Siemens Math, Science, and Technology Competition** (October 2009)

PROFESSIONAL ACTIVITIES

Vanderbilt Institute for Chemical Biology (VICB), Vanderbilt University **Nashville, TN**
Chemical Biology Association of Students, President (August 2018 – August 2019)
Chemical Biology Association of Students, Executive Board (August 2017 – August 2019)

Department of Chemistry, Vanderbilt University **Nashville, TN**
Chemistry Forum Committee (May 2017 – May 2018)

Department of Chemistry, Vanderbilt University **Nashville, TN**
Teaching Assistant (August 2016 – May 2018)

NICDH, National Institutes of Health **Bethesda, MD**
Post-baccalaureate Representative (September 2014 – June 2016)

South Bend Memorial Clinic at the Homeless Shelter **South Bend, IN**
Shadowed Dr. Brandon Zabukovic (October 2012 – May 2013)

TRiO Talent Search, The University of Notre Dame **Notre Dame, IN**
Tutor (September 2012 – May 2013)

AFFILIATIONS

American Society for Mass Spectrometry, Student Member *April 2016 - Present*