Determination of Xenobiotic and Endogenous Metabolites Using Ion Mobility-Mass Spectrometry and Machine Learning

Dylan H. Ross

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Libin Xu, Chair

Miklos Guttman

Abhinav Nath

Program Authorized to Offer Degree:

Medicinal Chemistry

University of Washington

Abstract

Determination of Xenobiotic and Endogenous Metabolites Using Ion Mobility-Mass Spectrometry and Machine Learning

Dylan H. Ross

Chair of the Supervisory Committee:

Libin Xu

Medicinal Chemistry

Confident identification of xenobiotic and endogenous metabolites is central to applications including metabolomics, lipidomics, and drug metabolism studies. The integration of ion mobility spectroscopy with mass spectrometry enhances the information gained from such studies without significantly impacting their analytical throughput and increases confidence in identification of unknown metabolites through the measurement of collision cross section (CCS). The diversity of small molecule chemical space necessitates the ability to predict CCS with high accuracy, rather than relying solely upon experimental CCS databases for annotation of unknowns. This dissertation aims to demonstrate the various applications of IM-MS in the large-scale determination of endogenous and xenobiotic metabolites, in addition to theory-based CCS calculation and machine learning-based CCS-prediction. First, I discuss the curation of a comprehensive and diverse database of experimental CCS values sourced from the literature and the development of a comprehensive CCS prediction model using this large experimental database, while providing insight into the structural characteristics of endogenous and xenobiotic metabolites that determine their CCS. Next, I examine the IM-MS characteristics of a panel of drugs and *in vitro*-generated metabolites using human liver microsomes and S9 fraction, with in-

depth computational modeling and theoretical CCS calculation to rationalize experimental observations. I then present the results from scaling the *in vitro* drug metabolite generation and IM-MS analysis to a high-throughput format and its application to a diverse collection of over 2000 drug and drug-like compounds in order to build a drug- and drug metabolite-specific CCS database for use in building a ML-based CCS prediction model for drugs and drug metabolites. Next, I discuss the development of a bioinformatic tool for the analysis of lipidomics data, which includes specialized models for the prediction of CCS and HILIC retention time, demonstrating how specialized predictive models can be built for specific chemical classes that leverage class-specific structural trends to produce high-accuracy CCS predictions. Finally, I summarize the principal conclusions of this collective work and provide perspective on how research in this area may continue to expand.

Acknowledgements

I would like to extend my heartfelt gratitude toward all of those who have contributed to my personal growth and development, without whom I would not have been able to complete this dissertation.

Firstly, I want to thank my mentor, Dr. Libin Xu. Libin embodies the superlative of all qualities that I could hope for in an advisor. His passion for science is evident in his genuine excitement (at times bordering on giddiness) about the many research projects being undertaken by his students. As a mentor, Libin has provided me with the freedom to explore the topics that interest me most, and has consistently found ways to gently guide these interests toward a cohesive body of work. Libin has also been a constant source of encouragement and praise, never missing an opportunity to highlight my growth and progress, and provide clarity and support through my struggles. I am also grateful for his compassion, empathy, and consideration of the many non-academic aspects of navigating grad school. I am honored to be counted among his students.

I would also like to thank Dr. Kelly Hines for her invaluable mentorship. Beyond sharing her vast expertise in ion mobility and mass spectrometry with me, Kelly imparted many important lessons that have made me a better scientist. Additionally, I would like to thank Dr. Ryan Seguin. Without his abundant guidance, from the earliest drug metabolism pilot reactions to the high-throughput scale, I would never have been able to get this project off the ground.

I would also like to acknowledge my thesis committee members: Dr. William Atkins, Dr. Matthew Bush, Dr. Miklos Guttman, and Dr. Abhinav Nath. Their insightful comments, questions, and suggestions have provided important perspective throughout the development of my research.

Finally, I want to thank my incredible community of family and friends, to whom I owe a debt of gratitude for shaping me into the person I am today. To my Mom, you are the most hard-working person I have ever known, and you have never been shy about letting me know how much you love me and how proud of my you are. To my Dad, I am eternally grateful for your constant encouragement and challenging me to always seek truth and understanding. To my sister, Kaitlyn, growing up can be a strange and difficult process but I have always been able to count on having you to go through it with me. To my partner Maddie, I know you do not want me to go on and on about the many things you are to me, so I will simply say: I love you.

Table of Contents

Chapter 1 Introduction

Portions this chapter have been adapted and reproduced with permission from:

> Dylan H. Ross and Libin Xu, Determination of drugs and drug metabolites by ion mobility-mass spectrometry: A review, *Analytica Chimica Acta*, 1154 (**2021**).

1.1 Conventional Methods for Determination of Xenobiotic and Endogenous Metabolites

The determination, or identification, of endogenous or xenobiotic small molecule metabolites is a core component of metabolomics, lipidomics, and drug metabolism studies. These omics analyses monitor large number and variety of compounds from complex matrices and provide unique insights into understanding the complex biological underpinnings of disease states, general cellular metabolism, or fate of xenobiotics.[1-3] Unique challenges exist depending upon whether the metabolites being interrogated are endogenous or xenobiotic in origin, but there are also many similarities between the processes of their characterization.

1.1.1 Determination of Endogenous Compounds

The large-scale determination of endogenous metabolites, termed metabolomics, seeks to fully characterize the abundance and composition of endogenous metabolites in a sample, and is an integral part of the systems biology, along with other omics technologies.[4-7] Endogenous metabolites represent highly diverse chemical structures and vary in a large range of molecular weight, from less than 100 Da to over 1000 Da. Liquid chromatography (LC) coupled with high-resolution mass spectrometry (HRMS)[8] and/or tandem mass spectrometry (MS/MS)[9] is the dominant analytical techniques used to carry out these studies although nuclear magnetic resonance spectroscopy (NMR) has also seen large number of applications.[1, 6, 10-12] However, challenges remain in metabolite identification as there are a large number of isobaric metabolites and many of them lack reference MS/MS spectra or cannot be differentiated by MS/MS.[4]

Lipidomics, a sub-discipline of metabolomics, refers to comprehensive analysis of lipids within a biological system.[1] In lipidomics experiments, it is desirable to identify as many lipid species and with as much confidence as possible in order to gain the most complete understanding of the biological processes being studied. An additional challenge of lipidomics studies is interpretation of the high dimensional data (hundreds to thousands of lipid features measured across many samples), since phenotypic differences often arise from nuanced patterns of change across many lipid species. In recent years, the inclusion of ion mobility separation (IM, *vide infra*) in metabolomics and lipidomics has gained traction and shown advantages in resolution and identification of lipids.[13-15] We will discuss our efforts in advancing the applications of IM in metabolomics and lipidomics by building large experimental and theoretical databases using machine learning, and creating bioinformatic tools for the analysis of complex lipidomics data.

1.1.2 Determination of Xenobiotics and Their Metabolites

The determination of xenobiotics, including drugs and environmental chemicals, in biological matrices has the added challenge of identifying their metabolites formed through xenobiotic/drug metabolism. Drug metabolism refers to the process by which the human body chemically modifies drug compounds, which mostly reduces their biological activities and facilitates their removal from the body although bioactivation is also commonly observed.[16] This task is carried out by a diverse set of drug-metabolizing enzymes (DMEs), each of which is capable of performing a specific chemical modification, ranging from oxidative reactions to conjugation with accessory groups. These reactions can be organized into two groups: Phase-I (functionalization) reactions and Phase-II (conjugation) reactions (Figure 1.6.1). Phase-I reactions include hydroxylation, dealkylation, other oxidations, reduction, and hydrolysis.

Cytochrome P450 family of enzymes (CYPs) are the primary Phase-I enzymes.[17] The Phase-II reactions involve the conjugation of various accessory groups (e.g. glutathione, glucuronic acid) to drug molecules, which also often leads to increased water solubility that facilitates excretion.[18, 19] A common feature shared among many DMEs is catalytic promiscuity, although most isoforms do tend to have substrate and reaction preferences. Figure 1.6.1 summarizes some common drug metabolism reactions and the DMEs that perform them. Understanding drug metabolism is important for drug development because it significantly affects drug clearance, and drug metabolites can also elicit unforeseen bioactivity or toxicity.[20, 21]

The determination of xenobiotics and their metabolites conventionally involves a combination of LC, coupled to UV-Vis spectroscopy and/or mass spectrometry (MS), and NMR.[22-24] LC-UV and LC-MS are rapid and automatable and require little sample material. However, structural information and confidence in identification of unknowns can be limited when just relying on UV spectra and MS fragmentation data. On the other hand, NMR allows definitive assignment of chemical structures, but it requires a large amount of material and is relatively low throughput, making it a costly and time-consuming analysis.

Two primary challenges are faced in the determination of small molecule xenobiotic compounds and their metabolites: 1) detecting analytes at low concentrations from complex biological matrices and 2) confident assignment of structures to drug metabolites. Thus, analytical techniques must first have adequate sensitivity to detect drugs or metabolites of interest and resolution to separate analytes from interfering matrix signals. This is complicated by the complexity of the biological matrices (blood/plasma, urine, feces, cerebrospinal fluid, tissues), which contain large amounts of interfering compounds, including lipids, proteins, carbohydrates, and endogenous metabolites. The second challenge arises from many types of

biotransformations and potential positional isomers of drug metabolites formed from these biotransformations. Despite our rich knowledge on drug-metabolizing enzymes, it is rarely possible to determine a priori the exact metabolites that will be observed from a given parent drug. In MS analysis, typical metabolites can be predicted with simple mass-shift rules and mass defect filtering;[25] however, such approaches lead to ambiguity when a modification can produce positional isomers and fail to capture important metabolites that do not present with a predictable mass-shift. While increased adoption of tandem MS and high-resolution MS (HRMS) has enhanced selectivity and identification confidence in the determination of drugs and their metabolites in recent years,[22-26] there is a clear need for techniques that can provide orthogonal structural information to enhance the confidence of drug and drug metabolite identification. Ion mobility-mass spectrometry (IM-MS), which offers both enhanced selectivity and additional structural information, is such a technique that has gained applications in recent years.[27, 28]

1.2 Ion Mobility Spectrometry as an Additional Dimension of Separation

Ion mobility spectrometry (IMS) is an analytical technique that rapidly separate ions based on their differences in gas-phase size and shape, which is orthogonal to polarity-based LC separations.[29-32] In time-dispersive IM separations, ions are driven through a neutral buffer gas under the influence of an electric field. As they travel, ions are slowed down differently as they interact with the buffer gas molecules, resulting in different amount of time traversing the mobility cell (*i.e.* drift time) (Figure 1.6.2). Using appropriate experimental measurements and/or calibration, an ion's drift time can be converted into collision cross section (CCS), a unique physical property that reflects its gas-phase size and shape. CCS has been demonstrated to be extremely reproducible across different instrumentations and labs,[33-36] making it a reliable parameter that provides information (*i.e.* molecular shape, density, and polarity) that is partially

orthogonal to accurate MS or MS/MS spectra. The timescale of IM separation (milliseconds) and the fact that the separation occurs in the gas phase make it ideal for coupling with MS, with or without chromatography, providing an additional dimension of separation without affecting analytical throughput.

There are a variety of IM techniques in use today that have different attributes, but all techniques operate on the same fundamental principle: separation of ions in a buffer gas under the influence of an electric field.[14, 37-39] Drift tube and traveling wave IM (DTIM and TWIM, respectively) are the most widely used time-dispersive techniques, characterized by a drift cell containing low pressure buffer gas and lined with electrodes. Ions are introduced into the drift cell and a static (as in DTIM) or dynamic (as in TWIM) electric field is applied, forcing the ions to travel through the buffer gas. The ions are temporally separated based on their interactions with the buffer gas: ions with less interactions travel faster and exit the cell sooner. Use of a static electric field in DTIM allows for direct measurement of CCS when using the step-field method, while calibration with standards of known CCS must be performed to obtain CCS from single-field DTIM or TWIM measurements. Although the step-field method in DTIM allows absolute CCS measurements, single-field calibrated method is more temporally compatible with the LC timescale. Trapped IMS (TIMS) traps and releases ions in a mobility-dependent manner. In TIMS, the buffer gas constantly flows towards the MS end and as ions are introduced into the drift cell, they are restricted from flowing with the buffer gas by an applied opposing static potential. The ions then migrate to an equilibrium position within the drift cell, where the strength of the opposing potential matches that of the forward buffer gas flow, and subsequently, the ions are eluted from the drift cell by a gradual lowering of the opposing potential. As such, the mobility of an ion is related to the field strength at which it exits the mobility cell.

Calibration with standards of known CCS measured on DTIM is also used for TIMS.[40-42] Field asymmetric waveform IM spectrometry (FAIMS), also called differential (ion) mobility spectrometry (DMS or DIMS), generally operates as a mobility filter, separating ions based on their differential mobilities under asymmetric high and low fields. Although this technique can provide high IM resolution (which is highly dependent upon system configuration),[43, 44] it does not allow CCS measurement and only a fraction of ions with specific differential mobilities pass through the device. Thus, FAIMS has more utility in targeted applications acting as a mobility filter and increasing signal-to-noise ratio for analytes of interest.

1.3 The IM-MS Conformational Landscape of Xenobiotic and Endogenous Metabolites

It has been demonstrated that different classes of biomolecules display distinct trends in the IM-MS conformational space, owing to their different gas-phase densities,[45, 46] which are determined by factors including chemical composition and conformational rigidity/flexibility. For example, at the same mass range, lipids tend to occupy large gas-phase size (*i.e.*, large CCS), followed by peptides, carbohydrates, and oligonucleotides, in that order. The ordering of compounds having the same mass but differing in their composition and/or conformation in the IM dimension demonstrates the nature of the orthogonality between mass and CCS: though CCS is partially determined by mass it is also reflects the composition and arrangement of that mass. Identifying these structural trends is important because it allows higher confidence in the identification of unknowns in IM-MS and it enables the simultaneous determination of multiple diverse chemical classes from complex samples. Small molecule metabolites, both xenobiotic and endogenous, are not broadly homogenous in their IM characteristics, but subclasses within this diverse group can exhibit more distinct trends.

Hines *et al.* measured a large collection (>1400) of CCS values of drug and drug-like compounds, which revealed some interesting trends in the conformational landscape of such compounds.[47] Overall, the drug and drug-like compounds occupy a large conformational space in the CCS *vs. m/z* plot relative to that of lipids and peptides, indicating a large structural diversity among these compounds. However, more distinct trends were observed when individual classes of molecules within the dataset were considered, which indicates tight structure-bioactivity relationships. For example, different subclasses of antimicrobials were all distinguishable by the regions they occupied in the CCS *vs. m/z* space, which can be attributed to the common structural characteristics within each subclass. Even some compounds occupying similar mass ranges (*e.g.* fluoroquinolones, penicillins, and cephalosporins with mass around 400) were able to be separated into individual classes on the basis of CCS. Other drug classes examined also display characteristic tight groups in the CCS-*m/z* plot. Another study performed a characterization of 124 drugs of abuse and toxic compounds using DTIM-MS and also observed a correlation between CCS and *m/z* for this group of functionally related compounds.[48] This study also demonstrated the resolution of isomeric morphine and piperine in the IMS dimension, a difference arising from their distinct cyclic/linear structures (morphine is polycyclic and thus more compact than the mostly linear piperine with fewer fused rings). Tejada-Casado *et al.* examined a collection of 92 veterinary and human drugs using TWIM-MS and found that the structural trends for many sub-classes of drugs (*e.g.* sulfonamides, aminoglycosides, quinolones, and tetracyclines) deviate from the trendline of the entire dataset in the CCS *vs. m/z* plot,[49] which is confirmative of the Hines *et al.* study.[47] This study also showed clear separation between the trends for benzimidazoles, 5-nitroimidazoles, and sulfonamides, even among compounds with similar masses. A similar large-scale characterization has been performed for lipids, and while as

a class lipids occupy a more restricted region of the IM-MS chemical space than small molecule metabolites, distinct trends were observed with respect to lipid class and fatty acid composition.[50]

Overall, these studies indicate that when considered as a group, small molecule metabolites (both xenobiotic and endogenous) occupy a large CCS-*m/z* space that is reflective of their large chemical diversity, but structurally related classes and sub-classes tend to occupy more restricted and characteristic regions of this space that reflect their specific structural properties. While such trends often display some degree of overlap between classes, there are many examples where these differences are sufficient to enable distinction. The distinctive structural trends that exist between and within chemical classes make IM-MS an especially useful tool for their determination.

1.3.1 Multimodal CCS as a Result of Ionization: The Structural Effects of Protomers

In IM-MS, multimodal arrival time distributions (ATDs) are indicative of isomerism or conformational heterogeneity. In either case, the presence of multiple peaks from the same parent drug in IM separation arises from multiple isomeric species that differ in their gas-phase conformation. Constitutional isomers differ in the arrangement of atoms can be due to distinct isobaric species present in a sample or arise from the ionization process. Isomers arising from protonation at different sites during the ESI process (*i.e.* protomers) are frequently observed for drugs and drug-like molecules due to the presence of multiple heteroatoms.

The earliest work to characterize drug protomers revolved around the analysis of fluoroquinolone antibiotics by IM-MS. Kaufmann *et al.* first observed distinct fragmentation patterns of norfloxacin under different ESI conditions (specifically, cone voltage), which links their formation to the ionization process.[51] A caveat to this observation is that the ability to

observe distinct fragmentation patterns for different protomers relies upon the energetic barrier to proton migration being higher than the barrier to fragmentation, which is likely the case with fluoroquinolones given the rigidity of the central ring structure. Ultimately, two peaks were resolved by IM separation, which are attributed to different protomers that fragment differently. Lapthorn *et al.* undertook a more in-depth characterization of norfloxacin protomers using IM-MS with fragmentation and computational modeling.[52] By adjusting IM parameters to maximize resolution, the mobiligram of norfloxacin was revealed to have 3 distinct components, with atypical peak shape of the third and major component indicating a possible fourth component. Each IM separated peak also gave distinct fragmentation spectra. Computational modeling was performed on all possible protomers, and while the agreement of theoretical and experimental CCS values was limited, the three observed components were putatively assigned based on fragmentation patterns, revealing an unexpected but thermodynamically stable protonation site on a carbonyl oxygen. It was also found that the rank-order of theoretical dipole moments corresponded well with experimental drift times, indicating contribution of polarity, which affects long-range dipole-induced dipole interactions between the ion and the drift gas, to the separation of protomers by IM. Another important observation from this study was the fact that the thermodynamic stability of individual protomers did not necessarily correspond to their observed abundances, indicating that protonation process has a kinetic component in ESI. Hines *et al.* also observed bimodal ATDs for norfloxacin and several other fluoroquinolones (ciprofloxacin, enoxacin, sarafloxacin, perfloxacin).[47] Computational modeling on ciprofloxacin, enoxacin, and sarafloxacin revealed similar protomer assignments for the larger and more abundant peak corresponding to protonation at the distal nitrogen of the piperazine moiety. More recently, McCullagh *et al.* used a cyclic IM-MS instrument with high IM resolution to study nine

fluoroquinolones and their protomers.[53] All fluoroquinolones included in the study displayed multi-modal ATDs, attributed to the formation of at least two protomers by each compound, and post-mobility fragmentation of danofloxacin suggested protonation at the carbonyl and piperazine ring. When drift gas in the mobility cell was substituted with more polarizable CO_2 to enhance the effect of long-range interactions, danofloxacin was found to have a third IM peak that was not resolved in the original separation in N_2. All three danofloxacin protomers became completely resolved after several passes around the cyclic IM cell (effectively increasing the IM separation length) with N_2 as the drift gas, and post-mobility fragmentation data suggested protonation at the carbonyl and both nitrogen atoms on the piperazine. While these studies only span a single class of drug compounds, they are foundational in understanding the contribution of protomers toward the characterization of drugs by IM-MS.

The effects of protomers in the analysis of drugs by IM-MS have also been explored outside the context of fluoroquinolone antibiotics. Warnke *et al.* presented a rigorous examination of the protomers of benzocaine using IM-MS, drift time-resolved IR spectroscopy, and computational modeling.[54] Benzocaine displayed two major IM peaks as a protonated ion, attributed to protonation at the N and O positions rather than separate conformers. Individual IR spectra were recorded for each of the drift time peaks, which were found to have distinct spectra. O- and N- protonated species were modeled using density functional theory (DFT) and their theoretical IR spectra matched well with the experimentally obtained spectra from the two IM peaks. Significantly, this work represents the first unequivocal assignment of protomer species from a bimodal ATD in an IM-MS experiment. Boschmans *et al.* characterized a group of small molecules, including benzocaine and its *ortho* and *meta* positional isomers, using IM-MS and computational modeling.[55] *para*-Benzocaine was found to have two peaks in the IM dimension

in the protonated form, and computational modeling recapitulated previous results with O- and N-protonated isomers attributed to the first and second peaks, respectively. The *meta* isomer of benzocaine also displayed two IM peaks, with computational modeling indicating the same relative assignments of protomers. Interestingly, the *ortho* isomer of benzocaine displayed only a single IM peak, with computational modeling suggesting similar CCS values for both O- and N-protonated isomers. In this case, it is unclear whether both isomers were present and unresolved under the IM conditions used or only a single isomer was observed although the authors also suggest the possibility of gas-phase interconversion between isomers through intramolecular H-bonding. In addition to the benzocaine isomers, melphalan and two structurally related derivatives were also included in this study. Two IM peaks were observed for melphalan and of the three protomers examined by computational modeling (one O- and two different N-protonated isomers), the two N-protonated species were found to have the best agreement with experimentally determined CCS. It was also found that the rank-order of computed dipole moments for the melphalan protomers was the same as their experimental CCS values, supporting the idea that long-range interactions could contribute to the separation of protomers due to their different polarities. Hines *et al.* observed bimodal distributions of several other classes of molecules, such as beta-lactams (*e.g.* bacampicillin, Figure 1.6.3 B), irigenin 7-benzyl ether, and antimycin-A.[47] Computational modeling of the β-lactam antibiotic cefpodoxime proxetil (Figure 1.6.3 A) suggested protonation at the aminothiazole and β-lactam moieties led to distinct conformational states, explaining the two IM peaks.[47] Taken together, these studies illustrate that protomers can be observed from a variety of chemical classes, and in many cases, their conformational differences can be resolved by IM-MS analysis.

To summarize, multimodal ATDs in IM-MS could arise from different sites of ionization of a molecule, and understanding this behavior is important for deconvolution of relevant signals from complex matrices. In the analysis of drugs by IM-MS, this becomes an especially important consideration given the presence of multiple heteroatoms in most drugs that could lead to multiple protonation sites.

1.4 Theoretical Calculation and Prediction of Collision Cross Section (CCS)

While there are a growing number of CCS databases covering a wide range of chemical classes available for use in compound identification, it is unlikely that most unknown metabolites encountered in metabolomics, lipidomics, or drug metabolism studies would be represented. Additionally, presence of complex IM behavior (*e.g.* IM-resolvable isomers or conformers) is less likely to be represented in such databases. The ability to robustly predict CCS values is therefore necessary to address such situations and guide identification of unknowns. Generally, approaches to CCS prediction fall into one of two categories: theory-driven or data-driven. Theory-driven approaches involve modeling the 3D structure of an analyte and simulating its interactions with a drift gas to determine its mobility. This family of approaches has the benefit of being rooted in fundamental chemical principles and thus can achieve high accuracy in their predictions. However, they can also be subject to systematic errors arising from the necessarily simplified nature of computational modeling. In contrast, data-driven approaches rely upon leveraging complex trends in large-scale datasets to make predictions. These methods are also capable of producing high-accuracy CCS predictions, but their performance is strongly dependent upon the coverage of chemical space in the training data.

1.4.1 Theory-Driven CCS Prediction

The calculation of CCS from a 3D structure[56] generally involves simulating, at varying levels of detail, the collisions and long-range interactions of drift gas molecules with an analyte ion that determine its mobility (Figure 1.6.4 A). The available methods span a range of theoretical complexity, which inversely corresponds to the required computational time. The earliest work in the field used a projection approximation (PA) to predict the CCS of carbon cluster ions in helium drift gas.[57] In the PA method, the projected area (as determined by a Monte Carlo sampling method using the van der Waals radii of the drift gas and analyte atoms) is measured for a large number of orientations of the analyte and averaged together to obtain CCS. A later and more refined approach was the exact hard-spheres scattering (EHSS) method,[58] which models the trajectory of the drift gas as it approaches and collides with the analyte ion and accounts for momentum transfer based on collisional impact and scattering angles. The next iteration of theoretical CCS calculation added a long-range interaction potential, accounting for van der Waals (using a Lennard-Jones or 6-12 potential) and polarization (ion-induced dipole) interactions, giving rise to the trajectory method (TM).[59] Together, the PA, EHSS, and TM CCS calculation methods make up the original MOBCAL suite. The TM was later modified for using nitrogen as the drift gas by adjusting the Lennard-Jones parameters, adding an ion-quadrupole potential term, and accounting for the orientation of the non-spherical nitrogen molecule relative to the analyte ion.[60] This modified TM represents the most rigorous level of theoretical CCS calculation in use today; however, its applicability is essentially limited to small molecules due to its computational complexity and corresponding low throughput.

Other CCS calculation methods have been developed that are intermediate in terms of their theoretical rigor but provide comparable accuracy and/or greatly improved throughput

relative to the previously discussed methods. Bleiholder *et al.* presented an improved PA method, the projection superposition approximation (PSA), which accounts for long-range size and shape effects by modeling the atoms of the analyte ion in a probabilistic fashion rather than as simple projected hard spheres.[61] The PSA method is able to provide greater accuracy than the PA method while still be computationally practical for large molecules. Diffuse hard-sphere scattering (DHSS) is a variation of EHSS in which collisions between drift gas and analyte ion are not assumed to be specular and elastic, but diffused instead. As such, some of the energy transmitted to the ion contributes to its internal (*i.e.* vibrational and rotational) energy and the remission angle of the drift gas is independent of its impingement angle.[62] DHSS provides a more realistic representation of the collision process than EHSS while still being suitable for intermediate-sized systems, and has been implemented in the IMoS software package with modifications to properly account for using nitrogen as a drift gas.[63] Collidoscope is an implementation of the TM that includes changes to how trajectories are sampled, simplification of trajectory integration, and parallel simulation of trajectories, which reduces computational time albeit at a slight reduction in robustness.[64] Marklund *et al.* presented an implementation of the PA method, termed IMPACT, in which drift gas-analyte ion collision detection was implemented using an "octree" approach with recursive subdivision of atoms into spatial subsections allowing for a drastic reduction in the number of collision check calculations required.[65] This optimization produced up to 20-fold acceleration of calculations over the original implementation for protein structures in the >100 kDa size regime. High Performance CCS (HPCCS) is a more recent implementation of the TM that employs parallelism in the calculation of drift gas trajectories, among other programming optimizations, while preserving the same long-range potential and scattering model as in the original MOBCAL

implementation.[66] MobCal-MPI is another implementation of the TM that largely follows the original implementation in MOBCAL but with parallelization, an improved long-range potential using Exp-6 van der Waals (as opposed to the Lennard-Jones potential), and ion-quadrupole terms for accurate simulation of nitrogen drift gas.[67] Both of these implementations produce similarly robust results to the original (and modified for nitrogen drift gas) MOBCAL TM, but with performance increases primarily from parallelism that make CCS calculation significantly faster for small molecules and practical for intermediate-sized molecules.

1.4.2 Data-Driven CCS Prediction

A more recently developed alternative strategy to theory-driven calculation of CCS values is data-driven prediction of CCS values, which takes advantage of large-scale trends in existing collections of experimental CCS values to produce predictions for chemically similar compounds using machine learning (ML). At a high level, these approaches often follow a similar workflow (Figure 1.6.4 B) that starts with selection of experimental CCS values to use for training and validating the regression model. While there exists a large variety of regression models for ML, the models used in CCS prediction fall broadly into two categories: support vector regression (SVR) and artificial neural networks (ANN). SVR is predicated upon maximizing separation between training samples in a high dimensional space and produces predictions based upon distance from a hyperplane in that space, which is defined by a subset of training examples (*i.e.*, the support vectors). ANNs consist of units that vaguely mimic the action of a biological neuron by outputting a weighted, nonlinearly transformed sum of their inputs, and its core model training consists of optimizing these weights. The units can be arranged into complex multilayered networks capable of approximating nonlinear functions and performing difficult classification and regression tasks on high dimensional input data. Numerical

representations (features) of the compounds, termed molecular descriptors (MDs), are chosen to encode structural variations relating to CCS. Next, the dataset is partitioned in a fashion that a large proportion is used to train the model while a small proportion is set aside until after training to validate the model's performance. The regression model can then be trained using the training data such that the error is minimized between predicted and reference CCS. The reserved test set data are then used to validate the model performance and check for overfitting of the training data. The primary benefits of ML based CCS prediction are the speed, accuracy, and robustness of predicted CCS values given large amounts of training data. An important drawback of this approach, however, comes from the fact that the quality of CCS predictions is heavily reliant upon the quality of the features used and the breadth of the training data.

The first application of ML for CCS prediction was reported by Zhou *et al.*, who measured CCS values for a collection of 400 metabolites consisting primarily of endogenous small molecules, and used this dataset to train an SVR model for CCS prediction.[34] The predictive model, MetCCS, produced predictions on data unseen during model training achieving median relative errors (MREs) of 1.6-1.8% on ^{DT}CCS values and 1.5-3.1% on ^{TW}CCS values, well within accuracy limits that are widely accepted in the field (2-3%). The composition of the training data used in this study (endogenous small molecules) means that this model would not be well-suited to other classes of molecules, such as lipids, drugs or drug metabolites. In fact, poor performance of this model on lipids has been documented.[68] Indeed, the authors later reported a lipid-specific CCS prediction model, LipidCCS, using exclusively lipid-derived CCS data, which greatly outperformed MetCCS in predicting lipid CCS values.[68] This example highlights the performance benefit that can be achieved through model specialization, as well as the necessity of training models using data that is representative of the chemical classes for the

intended application. Other studies have followed this approach of ML-based CCS prediction for collections of depsipeptides and pesticides.[69, 70] Mollerup *et al.* were the first to apply ML based CCS prediction on a collection of compounds comprised of pharmaceuticals, drugs of abuse, and their metabolites.[71] In this work, an ANN was trained to predict CCS from 8 MDs (225 compounds in training data), reaching a predictive performance of 1.7% MRE on an external validation set (36 compounds). Plante *et al.* used this dataset, along with 4 other published ^{DT}CCS datasets, to train a convolutional neural network (CNN) for prediction of CCS directly from the SMILES structure notation.[72] This novel approach to CCS prediction benefits from its ability to pick up on local structural cues due to the CNN architecture; however, the coverage of drugs and drug metabolites are still limited (data from Mollerup *et al.* were included). Recently, Zhou *et al.* reported an improved CCS prediction model, AllCCS, using a larger and more diverse training dataset (1851 and 795 CCS values in positive and negative modes, respectively), but the performance of the model on drugs and natural products were less optimal than endogenous metabolites and lipids.[73] The authors also generated a large predicted CCS database (~ 12 million values) and demonstrated improved metabolite annotation when adding CCS as an identifier for metabolites from a variety of biological samples.

1.5 Dissertation Overview

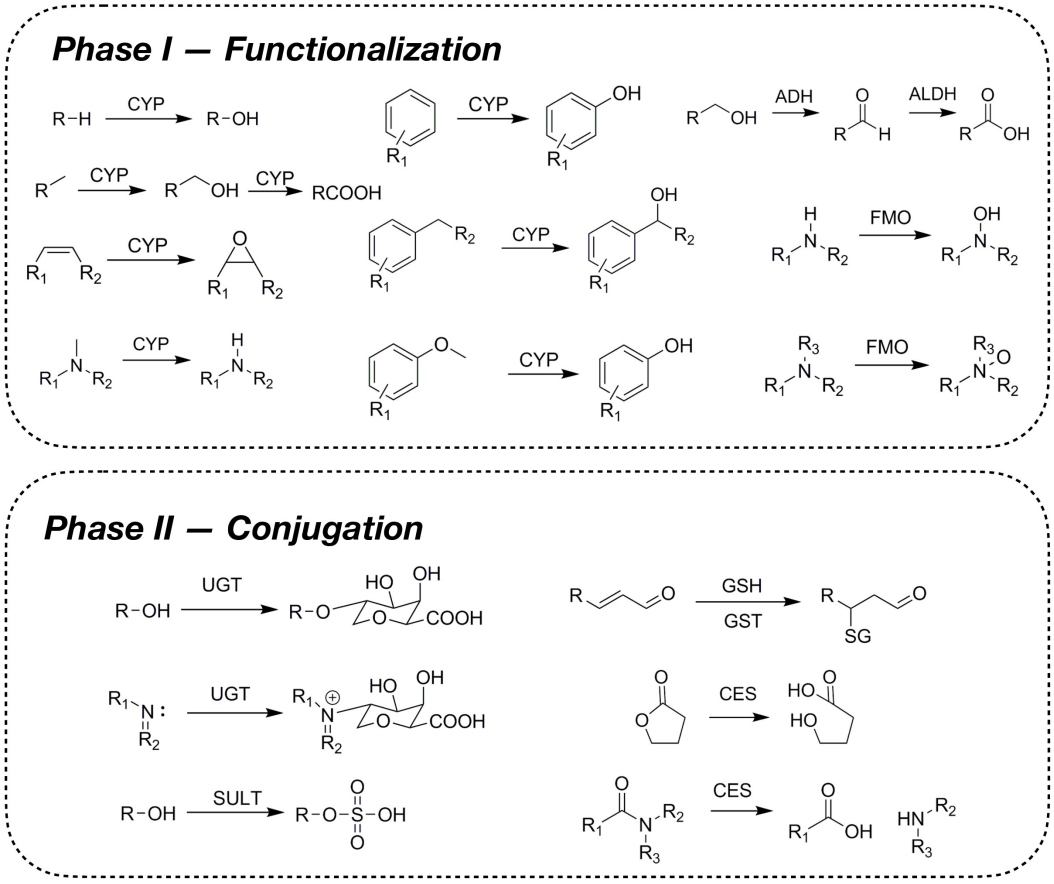
The research discussed in this introduction chapter demonstrates the need for confident identification of xenobiotic and endogenous metabolites for applications including metabolomics, lipidomics, and drug metabolism studies. Integration of IM with MS is a promising way of enhancing the information gained from such studies without significantly affecting their analytical throughput and increasing confidence in identification of unknown metabolites through the measurement of CCS. The diversity of small molecule chemical space

necessitates the ability to predict CCS with high accuracy, rather than relying solely upon experimental CCS databases for annotation of unknowns. This dissertation aims to demonstrate the various applications of IM-MS, theory-based CCS calculation, and machine learning-based CCS-prediction for the determination of xenobiotic and endogenous metabolites.

In this dissertation, Chapter 2 discusses the curation of a comprehensive and diverse database of experimental CCS values sourced from the literature and the development of a comprehensive CCS prediction model using this large experimental database. This chapter also provides insight into the structural characteristics of endogenous and xenobiotic metabolites that determine their CCS. Chapter 3 focuses on the IM-MS characteristics of a panel of drugs and *in vitro*-generated metabolites using human liver microsomes and S9 fraction, with in-depth computational modeling and theoretical CCS calculation to rationalize experimental observations. In Chapter 4, the *in vitro* drug metabolite generation and IM-MS analysis from the previous chapter are scaled to a high-throughput format and applied to a diverse collection of over 2000 drug and drug-like compounds in order to build a drug- and drug metabolite-specific CCS database that enabled production of ML-based CCS prediction models for drugs and drug metabolites. Chapter 5 discusses the development of a bioinformatic tool for the analysis of lipidomics data, which includes specialized models for the prediction of CCS and HILIC retention time. This chapter demonstrates how specialized predictive models can be built for specific chemical classes that leverage class-specific structural trends (*i.e.*, lipid fatty acid composition) to produce high-accuracy CCS predictions. Finally, Chapter 6 summarizes the principal conclusions of this collective work and provides perspective on how research in this area may continue to expand.

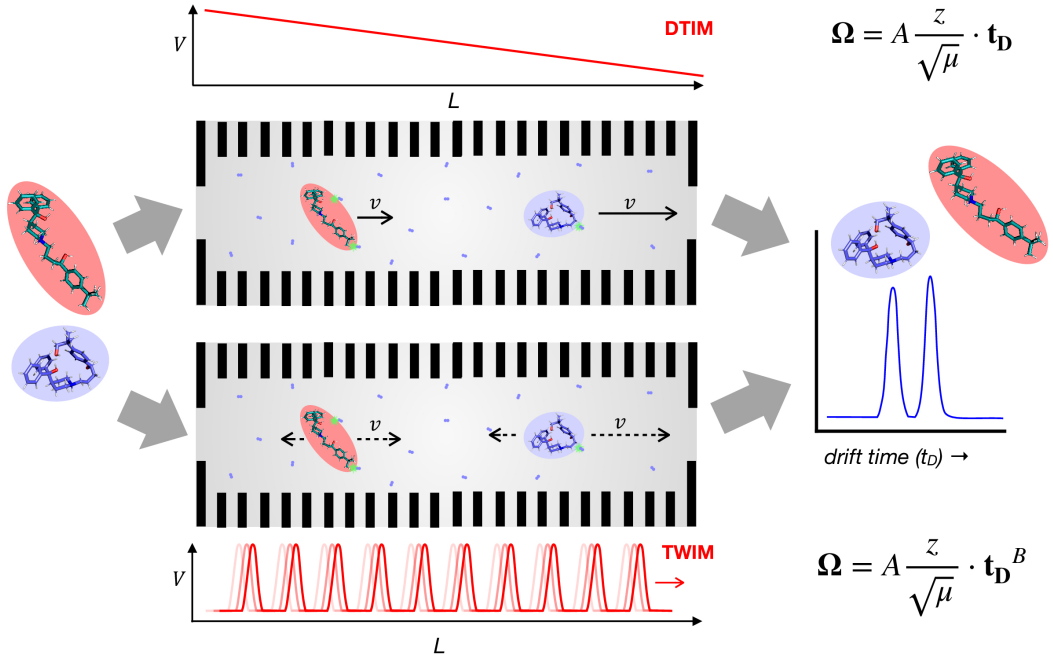
1.6 Figures

1.6.1 Common Structural Modifications in Drug Metabolism.



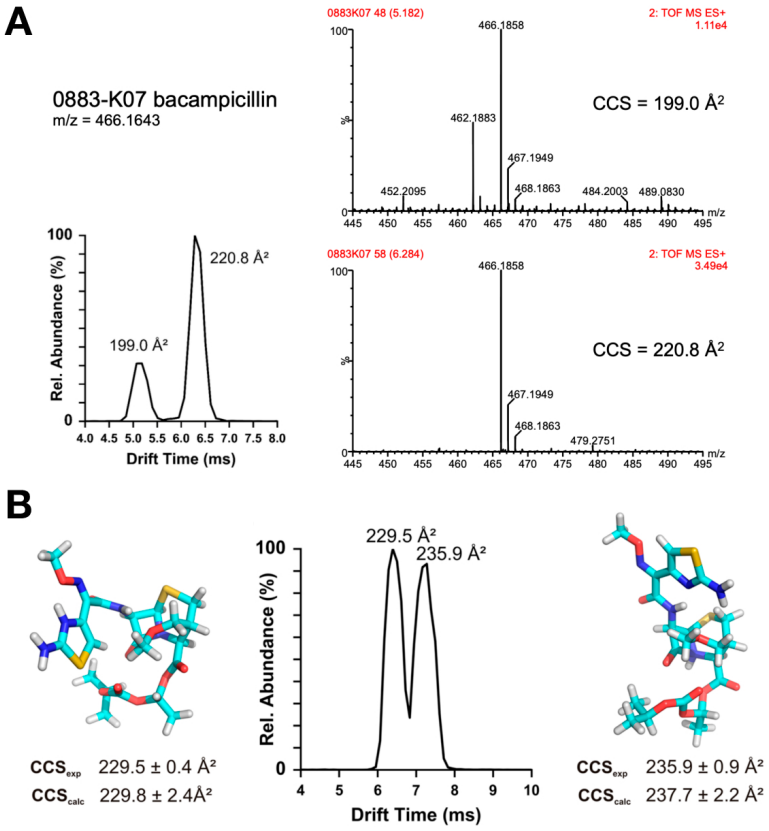
Examples of common Phase I (functionalization) and Phase II (conjugation) biotransformations that occur in drug metabolism.

1.6.2 *Schematic of Ion Mobility Separation*



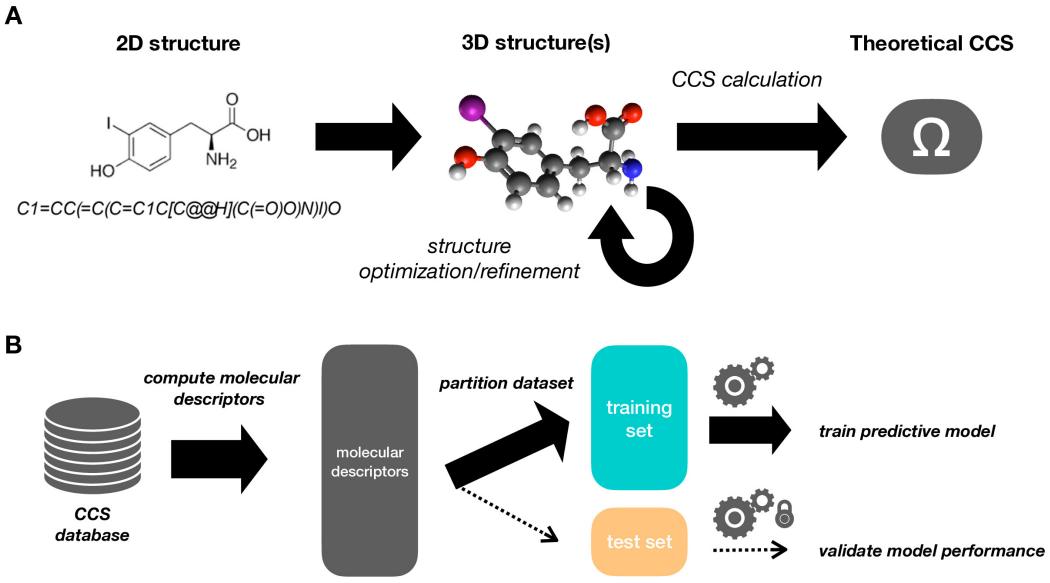
Schematic representation of IM separation. Ionized analytes are introduced into the drift cell and are driven through the cell by an electric field. For drift tube ion mobility (DTIM) instruments, the field is static along the length of the drift cell, while in traveling wave ion mobility (TWIM) instruments, the field moves across the cell in waves. Analyte ions travel through the drift cell at different velocities based on their degree of interaction with the drift gas molecules, which is determined by their gas-phase size and shape. More compact ions will travel faster through the drift cell, and thus exit at a smaller drift time. In DTIM collision cross section (Ω) is directly proportional to drift time (t_D), while in TWIM Ω is proportional to $t_D{}^B$.

1.6.3 Bimodal CCS Distributions of β-lactam Antibiotics



(**A**) Bimodal arrival time distribution for protonated bacampicillin, with corresponding MS/MS spectra for both drift time peaks. (**B**) Bimodal arrival time distribution for protonated cefpodoxime proxetil, with computationally modeled protomer structures corresponding to both drift time peaks. Adapted with permission from Hines et al., 2017;[47] Copyright (2017) American Chemical Society.

1.6.4 *Generalized Workflows for Theory- and Data-Driven CCS Prediction*



(**A**) Conventional workflow for calculation of theoretical CCS. First, a 3D structure (or multiple structures) must be generated either by drawing directly using a molecular editing program or from a 2D molecular representation (*e.g.* a SMILES string). This structure (or structures) often undergoes iterative refinement/optimization at various levels of theory in order to produce a theoretically reasonable conformation. Finally, theoretical CCS can be calculated at a desired level of theory using the optimized 3D conformer(s). (**B**) High-level workflow for CCS prediction using ML. Starting with a collection of CCS values, a numerical representation (molecular descriptors) must be chosen and computed for all compounds. This data set is then split into separate subsets to be used for training a predictive model and validating its performance. Using the training data, a predictive model can be trained and iteratively optimized until the desired performance characteristics are achieved. The model's performance is validated using the subset of data not seen during training, in order to detect overfitting of the training data.

Chapter 2 Leveraging Large-Scale CCS Collections for Comprehensive Prediction of CCS Using Machine Learning

Portions of this chapter have been adapted and reproduced with permission from:

Dylan H. Ross, Jang Ho Cho and Libin Xu, Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections, *Analytical Chemistry*, 92 (**2020**) 4548-4557.

2.1 Introduction

Identification of unknown analytes is challenging in LC-MS-based metabolomics or drug metabolite identification experiments. Identification is often dependent upon the availability of reference MS/MS spectra. Even with MS/MS spectral matching, the structural information garnered are often inconclusive due to uncharacteristic MS/MS fragmentations. Definitive compound identifications can be achieved using nuclear magnetic resonance (NMR) spectroscopy, but the large amount of material required makes this method costly and time consuming. Therefore, there is a demand for increased confidence in identification of unknowns without sacrificing analytical throughput.[22, 74]

Ion mobility spectrometry (IMS) is a rapid gas-phase separation technique based on the size and shape of analyte ions in the gas phase.[29, 31, 32, 75, 76] In IM separation, ions are driven through a neutral buffer gas under the influence of either a static electric field (as in drift tube IM, DTIM) or a dynamic electric field (as in traveling wave IM, TWIM),[37, 77] where they are differentially slowed due to their interactions with the drift gas molecules. The mobility of an ion in IMS is related to its collision cross section (CCS), a unique physical property determined by the ion's size, shape, and degree of interactions with the drift gas in the gas-phase.[59] When IMS is coupled with mass spectrometry (IM-MS), a powerful two-dimensional separation is achieved on the basis of CCS-to-charge and mass-to-charge ratio (*m/z*). In some cases, such separation is sufficient to justify the shortening or omission of chromatographic separation (*e.g.* liquid

chromatography) entirely, which dramatically increases the analytical throughput while simultaneously providing additional structural information. In addition to its inherent relationship to gas-phase structure, CCS has the benefit of being highly reproducible across different measurements and instruments. A recent study comparing the reproducibility of CCS values measured on DTIM and TWIM instruments found that 93% of the compounds tested showed absolute errors ≤ 2% for protonated species, and 87% showed absolute errors ≤ 2% for sodiated adducts although a few entries displayed errors > 3% (< 4% of total comparable entries).[78] This reproducibility becomes even greater when considering CCS values measured on a single type of instrument.[35]

The structural dependency of CCS and the high reproducibility of its measurement make CCS an excellent property to be used for compound identification. This utility, however, is limited by the availability of reference CCS values to compare against. Many large collections of CCS values have been produced that cover various chemical classes, but due to the breadth of small molecule chemical space and limitation in available standards, it is infeasible, if not impossible, to cover all of the unknowns encountered in metabolomics or related studies. CCS prediction can be used to address this limitation by providing CCS values for compounds that do not have reference values. Current methods of CCS prediction fall into two categories: theory-driven[58-60, 62, 64, 66, 79-82] and data-driven[34, 68, 69, 71, 72, 83] approaches, each of which has unique benefits and drawbacks. Theory-driven approaches generally use molecular modeling to produce an approximation of the 3D and electronic structure of a molecule and then compute CCS by simulating the interactions between the drift gas and analyte ion (both at varying levels of detail). These methods have the benefit of being rooted in solid theoretical principles, but depending on the level of detail, the computations can become very time consuming and laborious to set up.

Furthermore, depending on the assumptions made by the theory of a chosen approach, systematic errors may also be introduced and can be difficult to correct without introducing bias. Recently, a higher throughput computational workflow, ISiCLE, was reported, but this method still requires large amounts of computational resources, multiple steps of computational setup (conformer generation and optimization and CCS calculation for each conformer), and thus the throughput is still not ideal.[82] Furthermore, the computation results are good but still not superb (averaged error = 3.2% and R^2 = 0.94-0.96 in correlation with experimental values). On the other hand, data-driven approaches leverage trends within collections of reference values and generate predictions for similar compounds using their common characteristics, and typically employ machine learning (ML) to do so. Data-driven approaches have the benefit that once a predictive model has been trained, predictions can be carried out almost instantaneously. These models can produce high quality predictions on unseen data, with the caveat that the quality of predicted CCS values is tied directly to the quality and relevance of the training data.

CCS prediction using ML has gained traction in recent years, [34, 68, 69, 71, 72, 83] and a range of approaches have been adopted by different groups. These approaches share a similar general workflow: compilation and/or generation of a suitable set of training data, selection and optimization of a numerical representation for compounds (featurization), partitioning of data into training and testing/validation sets, selection of a ML model, training and optimizing the selected model, and finally testing and validating the trained model's performance. The majority of published predictive models are trained on relatively specialized, single collections of compounds like small molecule metabolites,[34] lipids,[68, 69] pesticides,[83] and drug-like compounds,[71] and therefore, the accurate prediction is limited to the types of molecules used for training. A recent study used datasets compiled from existing collections (mostly DTIM values,

but some TWIM values) in order to train predictive models that can handle a wider variety of chemical features,[72] but there is limited coverage of drug and drug-like compounds (*i.e.*, small molecules). Most existing models use some form of molecular descriptors (MDs) computed from the chemical structures as the features for performing ML.[34, 68, 69, 71, 83] These MDs often correspond to a summary property of a compound, such as polarity, LogP, and number of heavy atoms, and the combination of a variety of MDs provides a fingerprint of a compound's chemical characteristics, but these MDs do not always reflect a compound's structural features. The SMILES structure has also recently been used directly as input in a recent work using a convolutional neural network (CNN).[72] A variety of schemes have been used to partition datasets for training and validation, but generally all examples include some form of performance validation using data not seen by the model during training. Another limitation of existing models is that they use either support vector machines (SVMs)[34, 68] or artificial neural networks (ANNs),[71, 72, 83] which operate as "black boxes" and thus offer no interpretation of results beyond the predictions themselves.

To overcome the current limitations in CCS prediction, we first rigorously curated a large CCS database with nearly 6900 entries covering a diverse range of chemical space and identified structural characteristics, represented by molecular quantum numbers (MQNs),[84] contributing to the variance in mass-CCS space. We then developed a novel and high-performance approach of building CCS prediction models by breaking down the chemical structural diversity using unsupervised clustering based on structural features, followed by training of individual prediction models on each cluster using ML. Using this approach, we robustly trained and characterized a comprehensive CCS prediction model with high accuracy on diverse chemical

classes. Lastly, we built an all-in-one web interface (https://CCSbase.net) for querying the CCS database and accessing the predictive model to support unknown compound identifications.

2.2 Results and Discussion

2.2.1 Selection of Datasets for Combined CCS Database

Individual collections of CCS values measured in nitrogen gas included in the combined CCS database were chosen on the basis of size, diversity, and quality of measurements. The combination of multiple datasets, measured in different labs and on different instruments, into a single database for training CCS prediction models is desirable from the standpoint of generalizability since any systematic errors/biases present in a single CCS dataset can be averaged out by the presence of similar data from other datasets. CCS values measured on both DTIM and TWIM instruments were included in the database in order to maximize the size, breadth, and diversity of the CCS collection. Lipid CCS values measured on TWIM instruments were included only if the CCS was calibrated with lipid standards, which was found to give CCS values predominantly within 2% of DTIM values.[85-88] Comparing the CCS values of metabolites and drug and drug-like compounds measured on DTIM and TWIM, only small percentage of the entries (4 out of 45 in Hines *et al.*[47] and 7 out of 194 in Hinnenkamp *et al.*[78]) display errors that are > 3%, which justifies the inclusion of values for such compounds measured on both platforms.[34, 45, 47, 50, 68, 71, 83, 85-92] Furthermore, during assembly of the database, we analyzed overlapping values to assess their degree of agreement. Figure 2.5.1 shows the number and proportions of overlapping values from the combined CCS database that fall within 1, 3, and 5% difference of one another. In the distribution of all overlapping values (all, n = 695), 46.0, 84.5, and 94.5% fall within 1, 3, and 5% of one an-other, respectively. This indicates generally good agreement between overlapping values in the database, regardless of the instrument types.

Looking specifically at the variation between overlapping DTIM CCS values from different datasets (DT, n = 409), 43.3, 84.4, and 94.1% fall within 1, 3, and 5%, respectively, similar to that of the database as a whole. The variation between overlapping TWIM CCS values (TW, n = 163) is lower than the overall dataset with 65.0, 87.7, and 93.3% falling within 1, 3, and 5%, respectively. Finally, looking specifically at the overlapping values that contain both DTIM and TWIM CCS measurements (DT vs. TW, n = 238), 36.6, 75.6, and 93.3% fall within 1, 3, and 5%, respectively, which displays the most variation but still similar to that of the other groups. These results indicate that using both DTIM and TWIM CCS values is unlikely to add additional uncertainty to CCS predictions made by models trained on this combined database relative to models trained on datasets containing only one type of CCS values. On the other hand, the use of the combined database would greatly increase the generalizability of the prediction model due to the broad coverage of the structural diversity. Overall, a total of 7099 CCS entries with SMILES structures (out of 7326 entries in the database at this time) are used for this study, which covers 3314 small molecules (4552 CCS values), 981 lipids (2275 values), 91 peptides (112 values), and 74 carbohydrates (160 values), many of which have multiple observations from different MS adducts. [34, 45, 47, 50, 68, 71, 83, 85-92]

2.2.2 Structural Characterization of the Combined CCS Database

A 3-component PCA was performed on all the structures in the combined CCS database using the complete feature set (*m/z*, binary-encoded MS adduct, MQNs) in order to examine the primary contributors to the variance in the dataset as a whole in an unsupervised fashion (*i.e.* entirely ignoring CCS for the time being). This feature set contains molecular descriptors that reflect various structural characteristics of the compounds in the database, thus computing a PCA on these features provides an indication of the breadth of chemical space spanned by this

collection, as well as the most important chemical features that differ among them. The three principal axes captured 20.9, 14.4, and 5.7% of the total variance in the dataset, respectively. Together, only ~41% of the total variance was captured in this analysis, indicating that there are many sources of variance in this dataset and many of these contributing factors are at least partially orthogonal to one another. Figure 2.5.2 A-D depicts the projections of the full dataset onto all 3 principal axes, with individual data points colored either by source dataset (Figure 2.5.2 A, B) or rough compound classification (Figure 2.5.2 C, D). First, examining the projections colored by dataset, there is no significant separation between datasets along all principal components. This indicates that the source dataset is not strongly associated with the primary attributes of the dataset that contribute most significantly to the overall variance, *i.e.*, there is overlapping structural diversity among the various datasets. Compounds in this collection were then assigned one of the following rough chemical classes: small molecule, lipid, peptide, or carbohydrate based on the names of the compounds and information provided in the original publications. Examining PCA projections colored by these class labels, there is some separation along PC1 between small molecules and all other classes, the latter of which only separate from one another along PC2 and PC3. The separation between compound classes along the 3 principal axes coincides with the separation within each source dataset, supporting the notion that the variance observed between different datasets is attributable primarily to the different chemical classes represented in each dataset.

We next sought to investigate the general chemical trends driving separation along the principal axes, by examining the individual feature loadings. Figures 2.5.2 E-G depict the values of the top 3 features contributing to separation along PC1 *vs.* their PC1 projections. The top 3 features are heavy atom count (count of non-hydrogen atoms, *hac*), mass-to-charge ratio (*m/z*),

and acyclic oxygen count (*ao*), all of which are related to overall compound mass. Figure 2.5.2 H-J depict the values of the corresponding features for PC2: H-bond donor atoms (*hbd*), cyclic trivalent node count (*ctv*, related to branching in the chemical structure involving cyclic systems), and 6-membered ring count (*r6*). The features driving separation along PC2 can be generally described as being related to the composition and topology of a chemical structure (full list of the features can be found in Table 2.6.1). The corresponding top 3 features contributing to separation along PC3 are depicted in Figure 2.5.3. Taken together, these results indicate that the most significant source of variance in the dataset as a whole is related to compound mass, but composition and topology of compound structures are also important contributors.

We next examined what features contribute the most strongly to the distribution of CCS values in the dataset as a whole. Partial least-squares regression analysis (PLS-RA) was performed on all molecules using the complete feature set with CCS as the target variable. Unlike the unsupervised PCA, the first axis in PLS-RA (scores[0]) is chosen such that the most variance in a target variable (CCS) is explained, making it a supervised analysis. The second axis (scores[1]) is orthogonal to the first and explains the most of the remaining variance in the dataset. Figure 2.5.4 A and B show the projections of the full dataset onto these axes, with points colored by dataset source and rough chemical classification, respectively. Just as with the PCA projections, there does not appear to be a high degree of separation between datasets along scores[0], indicating that differences between datasets do not contribute strongly to the variance in CCS values. However, there does appear to be a similar pattern of separation between small molecules and all other classes along scores[0] (Figure 2.5.4 B), indicating that the chemical characteristics that differ between these chemical classes contribute strongly to variance in CCS values. Given the similarity between the overall patterns observed between source datasets and

chemical classes in both PCA and PLS-RA, it appears that the most significant sources of variance within the feature set have strong associations with variance in the target variable, *i.e.*, CCS. Indeed, plotting the projections of the full dataset along the most significant axes from PLS-RA and PCA (scores[0] vs. PC1, Figure 2.5.4 C) shows a high degree of correlation, supporting this notion.

As with PCA, the contribution of individual features to separation along a given axis in PLS-RA can be investigated by examining the feature loadings. Figure 2.5.4 D-K depicts the values of the top 8 features that contribute to separation along the primary axis from PLS-RA *vs.* scores[0] (shown in blue) or CCS (shown in red). The most significant contributors to variance in CCS are *hac* and *m/z* (Figure 2.5.4 D and E, respectively), both of which are mass-related molecular descriptors. The remainder of the top features (Figure 2.5.4 F-K) that show good correlations with scores[0] and CCS include features that are related to compound size (H-bond acceptor sites, *hbam*; carbon count, *c*; *ao*) as well as structure topology (acyclic single bonds, *asb*; acyclic monovalent nodes, *asv*; acyclic double bonds, *adb*). Much like the results from PCA, PLS-RA indicates that the most important features in the dataset that contribute to variance in the CCS are related to compound mass and size, also with significant contributions from features describing molecular topology.

2.2.3 Feature Selection Trials

Feature selection is an important step in any machine learning project as it promotes generalizability of a trained model by restricting the feature set used for training to only those that significantly impact the predictions of the target variable. Training models with large numbers of extraneous features can lead to overfitting of the training data and therefore poor performance on unseen data, so a balance must be struck between retaining as many features as

possible to be able to account for variance in the target variable and removing low impact features that will likely lead to increased variance (and thus decreased generalizability) in the trained model's predictions. Based on the results from the PLS-RA, a minimal feature set was manually selected, consisting of the set of 10 MQNs having an R^2 of at least 0.5 in the comparison of their values vs. scores[0] (in the order of importance from high to low: *hac*, *hbam*, *c*, *ao*, *asb*, *asv*, *adb*, *atb*, *adv*, *hba*), in addition to *m/z* and OHEA. Separately, a reduced feature set was produced in an automated fashion by evaluating linear regression models trained on subsets of MQNs generated by successive rounds of removal of random single features. RMSE was used to assess the performance of the models because RMSE gives an indication of the magnitude of errors and penalizes large errors (due to the squaring of errors), thus favoring more generalizable models. The feature removal and performance reevaluation process was continued until the test set RMSE increased beyond 2 standard deviations above the mean from the full feature set trials, at which point the process was repeated (for a total of 500 trials). Model performance was evaluated as a function of the number of MQNs retained (Figure 2.5.5 A). The average prediction performance decreased steadily (a steady increase in the average RMSE) as the number of retained MQNs decreased, and none of the models were able to maintain performance within the threshold with less than 11 MQNs retained. Individual MQNs were ranked on the basis of their relative retention rates from these trials as a way of establishing their relative importance in predicting CCS (Figure 2.5.5 B). All MQNs displaying a relative retention of 0.95 or higher (*i.e.* MQNs present in ≥ 95% of the trials with performance within the continuation threshold; dotted line) were selected for subsequent testing. The green bars in the plot represent the MQNs that were also part of the manually selected feature set, illustrating the considerable degree of overlap between the important features selected by the different methods.

Figures 2.5.5 C-J show the performance of a range of different ML models in predicting CCS using either the full set of MQNs (all, 50 features), the automatically selected MQN subset (auto, 27 total features including: 19 MQN features, m/z, and 7 OHEA), or the manually selected subset (manual, 18 total features including: 10 MQN features, m/z, and 7 OHEA). Looking first at the performance of simple multivariate linear regression models (Figures 2.5.5 C and G), there is a very slight increase in MDAE as the size of the feature set is reduced. The RMSE on the training data also slightly increases with decreasing feature set size, however, the test set RMSE remains essentially the same. The same performance trend is apparent among linear regression models employing L1 regularization during training (lasso, Figures 2.5.5 D and H). These trends suggest that while reducing the size of the feature set may slightly degrade performance on the training set, it may also lead to slightly better generalizability in the trained models as indicated by the unchanged performance on test set predictions by RMSE. The performance of models based on an ensemble of decision trees (forest, Figures 2.5.5 E and I) did not display any meaningful trends with respect to size of the feature set, which may be attributable to the fact that a degree of feature selection is inherent to the training process for this algorithm. Support vector machines using a radial basis function kernel (svr, Figures 2.5.5 F and J) performed best out of all models tested, and actually display a slight increase in prediction performance using the reduced feature sets compared to the full set of MQNs, both in MDAE and RMSE (but more pronounced in RMSE). Collectively, these results show that feature selection can have varying impacts on the performance of predictive models depending on the characteristics of the chosen algorithm, but in this case the performance gains are modest.

2.2.4 Model Specialization Through Unsupervised Classification

Current examples of CCS prediction models were built using collections of measurements made for specific chemical classes, *e.g.* lipids or small molecules (*i.e.* for metabolomics).[34, 68, 69, 71, 83] Such model specialization is justifiable from the standpoint that the chemical characteristics that con-tribute to a compound's CCS are likely to differ between chemical classes, and therefore, models trained to recognize the important characteristics of one chemical class is likely to perform poorly in producing predictions for other classes. Indeed, performing CCS predictions with LipidCCS Predictor[68] on all compounds from the combined database tagged as lipids produced an RMSE of 5.5 $Å^2$, while predictions on compounds tagged as peptides and carbohydrates produced RMSE scores of 46.9 and 56.3 $Å^2$, respectively (Figure 2.5.6). Employing such specialized models is an excellent approach for attaining high prediction accuracy, as long as the delineation between chemical classes is clear. However, assigning chemical class labels in an unbiased way can become difficult when considering the diverse small molecule space, which encompasses complex compounds containing substructures with chemical characteristics resembling multiple conventional chemical classes. Furthermore, most ML algorithms act as 'black boxes' with respect to the underlying characteristics of the training data that ultimately contribute to predictions. Even for ML algorithms that provide some insight into which features contribute most strongly (*e.g.* lasso, forest), there is still no way to tell whether and how subgroups within the training data influence this process. To address these limitations, rather than relying on manually assigned chemical classes, we performed unbiased and unsupervised classification using the K-Means clustering ML algorithm on all molecules with SMILES structures in the database to determine the most prominent groupings of compounds with respect to structural characteristics.

Using the full set of MDs (*m/z*, OHEA, and all MQNs) as features, K-Means clustering was used to group compounds on the basis of their structural similarity. During initial testing, fitting with 4 clusters was found to be optimal for capturing the major groups within the dataset while maintaining sufficient numbers in each cluster for model training (at least 100 molecules per cluster). Figure 2.5.7 A and B show the PCA projections of the full dataset colored by rough chemical classes, and Figure 2.5.7 C and D show the same PCA projection colored by clusters. The clustering analysis revealed similar groupings to the chemical class labels: most of the small molecules were assigned to cluster 1 and 2 (purple and blue, respectively), lipids to cluster 3 (magenta), and carbohydrates and peptides were assigned to cluster 4 (gold). However, a significant number of molecules in the original rough chemical classification were re-assigned to new clusters, suggesting that manual assignment of chemical classes could result in classification error. As seen in the PCA plot (Figure 2.5.7 C), the two small molecule clusters separate from the lipid and carbohydrate/peptide clusters along PC1, and separation within these groups occurs primarily along PC2. With the separation along PC1 being mostly related to compound mass, it appears that the separation between the small molecule clusters and the lipids and peptides/carbohydrate clusters is primarily mass-driven. Separation along PC2 is related more to topological differences in chemical structure, meaning that other factors such as branching or presence of rings drives the separation between the two small molecule clusters, and between the lipid and peptide/carbohydrate clusters, respectively. Interestingly, the assigned clusters conform to familiar trends in the IM-MS conformational space, although overlapping between clusters still exist (Figure 2.5.7 E). Broadly speaking, the lipid cluster occupies the high-CCS and high-mass region, while the peptide/carbohydrate cluster occupies the low-CCS high-mass region. The small molecule clusters occupy significant overlapping space in the low-mass region, each

covering the broad range of CCS values associated with this region. The most central structures within each cluster (minimum distance to the cluster centers in feature space) are presented in Figure 2.5.7 F. PE(36:3) is a prototypical lipid, while CMP-N-acetylneruaminic acid has chemical characteristics related to both carbohydrates and peptides (large, containing many heteratoms). The small molecules etodolac and 3-methoxytyrosine represent the central structures within the two small molecule clusters and, as expected from the PCA, they seem to differ mostly on a topological basis (*e.g.* rings and branching) rather than by size or composition. Collectively, these results show that use of an unsupervised clustering approach recapitulates many of the classifications attainable through a manual approach, while also providing class separation based on more nuanced chemical characteristics. Importantly, such an approach allows for unbiased assignment of chemical class when considering new molecules, especially those that do not fall cleanly into a single conventional chemical class.

Within each cluster, individual ML models were trained using the complete feature set (Figure 2.5.8), and the average performance of this ensemble of models was compared to corresponding individual models trained on the complete dataset (Figures 2.5.7 G-I). Feature selection is commonly performed in ML projects as it promotes generalizability of a trained model. However, we did not observe significant differences in performance between using the complete feature set and using selected features following a previous practice (Figure 2.5.5).[68] This is likely due to the already small number of structure-related MQNs as MDs (compare with the hundreds of MDs typically used in the literature) and the intrinsic feature selection process when training with models like Lasso, random forest, and SVR. Training individual models specific to each cluster led to a marked performance increase by MDAE and RMSE for the lasso models, and to a lesser degree for the svr models (Figures 2.5.7 G and I, respectively).

The performance by MDAE did not increase with the addition of clustering for the forest models, but the RMSE did decrease indicating that the use of clustering reduced the proportion of higher-magnitude errors for these models (Figure 2.5.7 H). Across all models (with or without clustering), the average performance by both metrics for the test set predictions was at parity with the training set performance, indicating good generalizability with respect to making predictions on new data. These results demonstrate that the inclusion of untargeted clustering followed by building individual predictive CCS models using each cluster can increase the model performance and that such an approach is well suited for application to large compound collections covering diverse chemical space. Additionally, this approach offers the benefit of interpretability at the classification level: the assignment of compounds to individual clusters provides information on the common chemical characteristics that define chemical classes in an unbiased fashion.

2.2.5 Training and Performance Characteristics of the Final Optimized Prediction Model

Based on the insights gained from the above work, we built a final deployable predictive CCS model using K-Means clustering and 4 individual svr models with radial basis function kernels trained on each of the fitted clusters. Figure 2.5.9 describes the final workflow for building and training this model and Figure 2.5.10 summarizes all of the performance characteristics of the model on both the training and test set data. The R^2 scores for training and test set data were 0.994 and 0.991, respectively (Figure 2.5.10 A), indicating excellent generalizability. Judging from addition-al metrics, this model was able to achieve high performance on the training data, with MAE, MDAE, and RMSE scores of 2.92, 1.70, and 5.48 $Å^2$, respectively (Figure 2.5.10 B, blue). The generalizability was also excellent based on the corresponding test set scores of 3.83, 2.37, and 6.46 $Å^2$, respectively, which show no significant

performance lapse relative to the training data (Figure 2.5.10 B, red). The cumulative error distributions of CCS predictions give a more detailed indication of the error structure of the CCS predictions, with 56.1 and 86.4% of CCS predictions for the training data falling within 1 and 3% of the reference values, respectively (Figure 2.5.10 C, blue). The model achieves similar performance on the test set data, with 44.7 and 81.2% of CCS predictions falling within 1 and 3% of the reference values, respectively (Figure 2.5.10 C, red). Taken together, these performance metrics indicate that this final model is capable of performing CCS prediction with high accuracy on diverse compounds, and that this performance is robust when applied to unseen data.

We next compared our final model with DeepCCS, the other comprehensive CCS prediction model using a CNN trained directly on SMILES structures, albeit with limited coverage of small molecules.[72] A comparison dataset was assembled from the datasets used in the training and characterization of DeepCCS (2298 compounds after selection of valid SMILES and MS adducts). Using this comparison dataset, 1960 CCS values were predicted by DeepCCS and the accuracy of these predictions was com-pared with our model by a variety of metrics (Figure 2.5.10 D-F). The final prediction model presented here outperformed DeepCCS by all performance metrics, despite using far fewer parameters to fit the data. This higher CCS prediction performance is likely attributable to two reasons. First, the larger and more diverse collection of data used to train the model enables greater accuracy in predictions by learning more robust and generalizable trends from the training data. Second, the use of untargeted clustering to partition the data on the basis of common structural features allows this model to learn specific trends for different classes of chemicals, thus increasing the overall accuracy of CCS predictions through model specialization.

2.2.6 Building an All-in-One Web Interface for Querying the CCS Database and Accessing the Predictive Model

To increase the accessibility of the database and the prediction model to the field, we have assembled the combined CCS database and the final prediction model into a convenient web interface (https://CCSbase.net). This interface allows users to query the database for reference CCS values with fine filtering control, including name, mass with accuracy, CCS with accuracy, polarity, SMILES structures, adduct types, charge state, and fuzzy search. This easily accessible database offers complementary and broader coverage (over 7300 entries) in comparison with the existing CCS databases,[46, 90, 92, 93] including the carefully assembled CCS compendium on DTIM CCS values (3833 values). This interface also allows rapid prediction of CCS values (with confident prediction for six different MS adducts) directly from SMILES structures using the final cluster-based prediction model discussed above. Batch query and prediction can be achieved with a simple single CSV file input. All results are directly viewable on the interface in the form of a table and in an interactive CCS-mass plot with the main trendline of the entire database being the background. The results are also downloadable as a CSV file. This platform can be easily built into existing metabolomics workflows and thus serve as a useful tool in the identification of unknowns from large-scale untargeted analyses.

The primary utility of the CCS database and predictive model are to support unknown compound identification by CCS. The metabolomics standards initiative (MSI) defines 4 annotation levels reflecting the rigor of compound identification.[94] Level 1 annotations are obtained from matching at least two orthogonal properties, such as *m/z*, CCS, or MS/MS, against values determined experimentally from authentic standards. Level 2 annotations are the same, with the exception that the reference measured values are taken from a secondary source (like the

literature). Level 3 annotations are obtained when reference measurements are not directly available for a compound, but can be inferred based on existing measurements from similar compounds. Level 4 annotations correspond to compounds that are unidentified, but can differentiated from other signals. Identifications based on measured values from the CCS database would therefore constitute level 2 annotations, *i.e.*, using measured *m/z* and CCS as orthogonal properties, while those made using CCS values generated using the predictive model would constitute level 3 annotations, *i.e.*, using *m/z* and predicted CCS values.

2.3 Experimental

2.3.1 Assembly of a Comprehensive CCS Database

A comprehensive CCS database was assembled from a variety of individual collections of CCS values available in the literature, [34, 45, 47, 50, 68, 71, 83, 85-92] representing broad coverage of current measurement techniques and chemical classes. For lipid CCS values measured on TWIM instruments, only those calibrated with lipids were included into the database.[33] The source datasets were each manually examined for any errors, and relevant data (*i.e.* CCS, *m/z*, mass, SMILES if present, *etc.*) from each entry was converted into a JSON format, yielding consistently formatted cleaned data with separate files for each individual dataset. The combined CCS database was constructed from the individual cleaned datasets using a series of build scripts developed in-house in order to be able to reproducibly rebuild the database when new datasets are added or when database organization is changed. A SQLite3 relational database was initialized with a table to hold relevant CCS measurement data (including MS adduct, *m/z*, and charge state) and metadata, which included source dataset and CCS measurement platforms and methods (DTIM: single field or stepped field *vs.* TWIM: calibrants). Data from each source dataset was added to the database, then entries with missing SMILES structures were attempted

to be filled first by searching PubChem or LIPID MAPS databases by compound name and if not found in the existing databases, SMILES were obtained from manual search in the database or hand-drawn structures in ChemDraw, in combination with the online SMILES translator (https://cactus.nci.nih.gov/translate/). Next, a table was added to the database containing columns for each of the 42 molecular quantum numbers (MQNs) used as part of the features for machine learning (see below). Finally, rough chemical classifications (carbohydrates, lipids, peptides, small molecules) were assigned to each entry of the database on the basis of compound name and data source.

2.3.2 Feature Set for Machine Learning

The full set of features used for prediction of CCS (n=50) includes the *m/z* of the observed MS adduct, one-hot binary encoded MS adduct (OHEA, n=7), and a set of molecular descriptors that capture information about the size, composition, and topology of each chemical structure (42 MQNs). The m/z and MS adduct were already present in the source datasets, but MS adduct had to be encoded into a numeric form in order to be used for CCS prediction. One-hot encoding was used to convert MS adduct into a binary representation, with unique labels for each adduct type that had ≥ 100 examples in the database and the rest of adduct types represented under 'other adducts' (total of 7 features, Table 2.6.2). The 42 MQNs were computed for all database entries containing SMILES structures using the RDKit library (https://www.rdkit.org) and stored in the database. This feature set reflects a variety of compound characteristics, ranging from size to composition and to structure topology (Table 2.6.1). Specifically, MQNs are molecular descriptors obtained from analyzing compounds as a molecular graph: *i.e.* collections of nodes (atoms) and edges (bonds).[84] The descriptors are properties of these graphs, consisting primarily of counts of various atom types, bond orders,

connectivity, *etc.* A benefit of using graph properties as molecular descriptors is that they are invariant with respect to the software used to compute them (unlike many empirical properties, such as cLogP), facilitating the broad application of predictive models and promoting reproducibility.

2.3.3 Analysis of Structural Features Contributing to Variation in the Mass-CCS Space

Dimensionality reduction analyses were used to explore the important sources of variance within the combined CCS database and to determine what chemical characteristics contribute most strongly to variance in CCS values. Specifically, principal components analysis (PCA) and partial least-squares regression analysis (PLS-RA) were used to analyze the MQNs of all entries in the combined CCS database in an untargeted and targeted fashion, respectively. Both analyses work by finding multidimensional axes in the input data that explains as much variance as possible, then subsequent orthogonal axes are chosen that explain as much of the remaining variance as possible. PLS-RA differs from PCA in that the first axis chosen is the one that explains the maximal variance in a target variable, in this case, CCS, rather than the input data, making it a targeted analysis. Both analyses are implemented in Scikit-Learn,[95] a free and open-source machine learning library for Python (*sklearn.decomposition.PCA* and *sklearn.decomposition.PLSRegression*).

2.3.4 CCS Prediction Using Machine Learning

It can be difficult to determine *a priori* what type of ML algorithm will have optimal performance characteristics for a given task, so a variety of ML models representing a range of algorithmic complexity were examined for use in predicting CCS. All ML models used in this work are implemented in Scikit-Learn.[95] Mean squared error of predictions vs. reference values (MSE) was used as the optimization target for model training. ML models tested vary from

simple to complex, including multi-variate linear regression (*sklearn.linear_model.LinearRegression*), a linear regression model employing L1 regularization (lasso, *sklearn.linear_model.Lasso*), a support vector regression model with a radial basis function kernel (svr, *sklearn.svm.SVR*), and a stochastically assembled ensemble of decision tree models (random forest, *sklearn.ensemble.RandomForestRegressor*).

Before model training, the dataset was processed in a stepwise fashion (Figure 2.5.11). First, the dataset was split into training and test sets (at proportions of 80% and 20%, respectively) and the test set put aside until after model training. This data splitting was performed in a stochastic fashion (seeded for deterministic results), with rough stratification on the basis of CCS to ensure a somewhat similar distribution of CCS values between the training and test sets. The training data were then centered and scaled such that each feature would have a mean of 0 and a standard deviation of 1. Such normalization is necessary to ensure numerical stability in training certain predictive models, and to avoid arbitrarily over-emphasizing features on the basis of their magnitudes. When necessitated by the predictive model being tested (*i.e.* lasso, svr, random forest), hyperparameter optimization was performed using a grid search with 5-fold cross validation (*sklearn.optimize.GridSearchCV*) on the training data. Using the optimal hyperparameters, a predictive model was then trained on the training data and performance metrics (see below) were computed from the training data. Finally, performance metrics were computed using the trained model to make predictions from the test set data. The entire process was repeated multiple times, using different pRNG (pseudorandom number generator) seeds to get different results, and the metrics (for both training and test data) were averaged together from each trial to get an idea of the average predictive model performance under a given set of

conditions. Repeated trials were only used for bulk performance comparisons between models or feature sets, but not to actually produce usable individual CCS predictions.

2.3.5 Description of Various Machine Learning Models Used

The simplest model tested in this study was multivariate linear regression (*sklearn.linear_model.LinearRegression*), in which a weights vector and bias term are optimized such that the product of the weights and input vectors added to the bias term produces an accurate CCS prediction. Additionally, a linear regression model employing L1 regularization during training (lasso, *sklearn.linear_model.Lasso*) was tested. The inclusion of L1 regularization during training means that the L1 norm of the weights vector ($l_1(w) = \sum_i^n |w_i|$) is minimized in addition to MSE, which tends to drive weights to 0 for individual features that do not strongly affect prediction accuracy. As such, there is a degree of feature selection that is inherent to the training process for lasso models, and this feature selection promotes generalization of trained models to new data by only relying upon the most important features when making predictions. A support vector regression model with a radial basis function kernel (svr, *sklearn.svm.SVR*) was also tested. In general, support vector machines (when using non-linear kernel functions) produce predictions by mapping the training data to a higher dimensional space, then optimizing a hyperplane in that space such that maximal separation is achieved across that hyperplane with respect to the target variable (binary label for classification tasks or continuous number for regression tasks). Predictions are then made on the basis of the projected distance between an input point and the optimized hyperplane. This hyperplane is only defined using a subset of the training examples that lie along its margins (*i.e.* the support vectors), thus, only a subset of the dataset is used to make predictions. Compared to linear regression models, svr models are capable of learning more complex patterns in the training data in order to make

their predictions, but they also carry a higher risk of overfitting the training data and therefore are potentially less generalizable to unseen data. The final model tested was a stochastically assembled ensemble of decision tree models (random forest, *sklearn.ensemble.RandomForestRegressor*). The individual decision tree regression models are trained on subsets of the training data (both in terms of samples and features), then their predictions are averaged together in the final ensemble to produce predictions. Much like svr models, random forest models are capable of learning very complex trends in the training data to produce high-quality predictions but can be prone to overfitting.

2.3.6 K-Means Clustering for Untargeted Classification of Chemical Structures

K-Means clustering is a multivariate technique in which data is partitioned into clusters, such that the similarity between samples that are partitioned together is maximized. Briefly, in K-Means clustering, centroids with the same dimensions as the input data are chosen for each cluster and each sample is assigned to the cluster with the nearest centroid. The centroid positions for the clusters are adjusted (and data partitioning repeated) such that the inertia (sum of squares of Euclidean distances of each sample from their assigned cluster centroid) is minimized. This is an unsupervised classification technique since the class assignment (*i.e.* partitioning into clusters) is done on the basis of similarity between subgroups within a dataset, without using predetermined labels, and therefore offers a means of classifying chemical compounds without the bias inherent to traditional manual assignment. K-Means clustering is implemented in Scikit-Learn (*sklearn.cluster.KMeans*).[95]

2.3.7 CCS Prediction Performance Metrics

The performance of predictive ML models trained on the combined CCS dataset was assessed using an array of metrics, intended to offer a complete representation of model

performance. Specifically, R^2, mean and median absolute error (MAE and MDAE, respectively), root mean-squared error (RMSE), and cumulative distribution of prediction errors below 1, 3, 5, and 10% (CE135A) were used. Each metric was computed from predicted and reference CCS values (y' and y, respectively) as follows. R^2 is calculated by comparing the residual sum of squares with the variance in the reference values. MAE and MDAE are computed as the mean and median, respectively, of the absolute errors of model pre-dictions. RMSE is computed as the square root of the mean-squared error. CE135A was computed as the proportions of predictions with relative absolute below 1, 3, 5, and 10%.

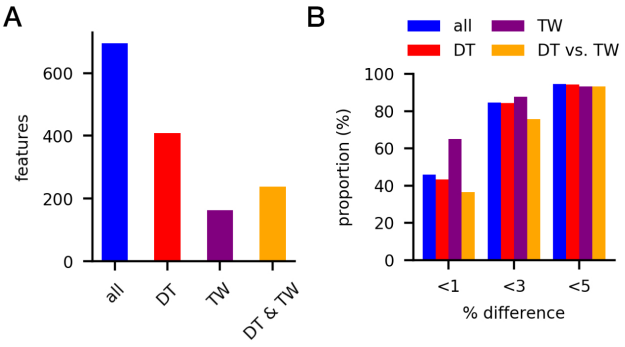
Although all metrics reflect the accuracy of the regression model in predicting CCS value, each metric specifically highlights a different aspect of model performance. R^2 is a good metric for assessing goodness of fit for regression models, but it is only suitable for comparing between models with the same number of parameters. MAE gives a good indication of the average magnitude of errors made by the predictive model, but MDAE may be a better indicator of this magnitude when the distribution of prediction errors is significantly skewed from normal. A draw-back of MAE and MDAE is that they can contribute to overfitting by allowing for small numbers of large errors to be easily balanced out by large numbers of small errors. In contrast, RMSE gives an indication of the magnitude of errors but does so in a way that more heavily penalizes large errors (due to the squaring of errors), favoring generalizable models with similar prediction errors across all samples. MAE, MDAE, and RMSE all share the benefit of being computed in the units of the target variable (in the case of CCS prediction, $Å^2$), which can aid in their interpretation with respect to real world performance.

2.4 Conclusion

The ability to predict high-quality CCS values for unknowns is a key step toward using CCS as a broadly applicable identifier for metabolites. Several major advances in ML-based CCS prediction are achieved in this work. First, by assembling the largest CCS database to date, a broad coverage of chemical structural diversity is achieved. Second, by per-forming statistical analysis of the structural features of all compounds, we have identified structural features that display high correlation with CCS values (Figure 2.5.4). Such correlation has not been systemically examined previously. Third, the use of structure-related MQNs as MDs is more relevant to the structure-dependent CCS than MDs reflecting the physical properties that were often used in the literature. Fourth, by breaking down the structural diversity using structure-based unsupervised clustering in combination with individual prediction models, the integrated model displays greatly improved performance and generalizability than without clustering. Importantly, this model also provides interpretable results based on the cluster that the unknown is assigned to, unlike previous work mostly using "black box" prediction models. Finally, we have built an easily accessible all-in-one web interface for efficient querying of the database and the prediction model. We anticipate continuous growth of this database and improvement of the prediction model as molecules with additional structural features are added.

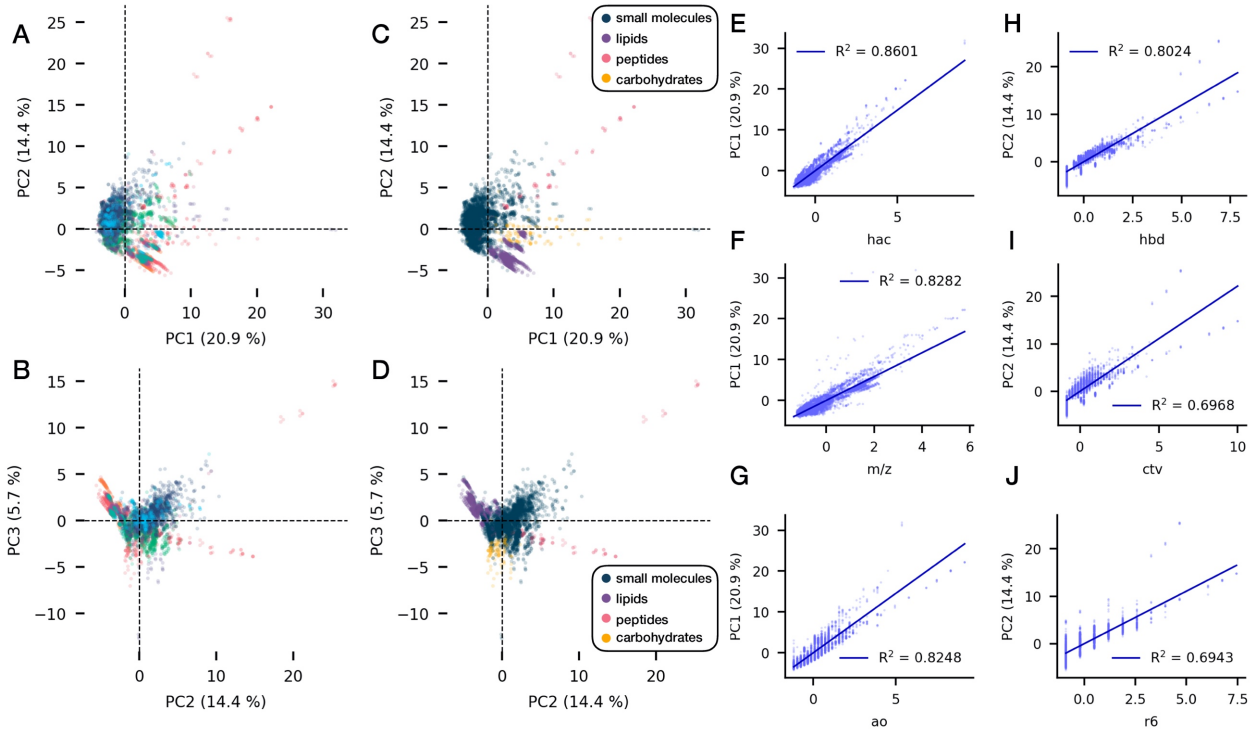
2.5 Figures

2.5.1 Composition and Agreement Between Different CCS Measurement Methods in Combined CCS Database



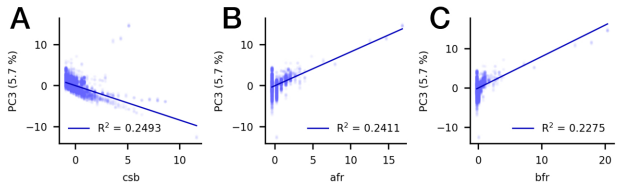
(**A**) Counts and (**B**) comparison of agreement between measurements present in multiple sources. Agreement for all overlapping CCS values in blue. Agreement between DTIM CCS values in red. Agreement between TWIM CCS values in purple. Agreement between DTIM and TWIM CCS values in gold.

2.5.2 PCA on Combined CCS Database



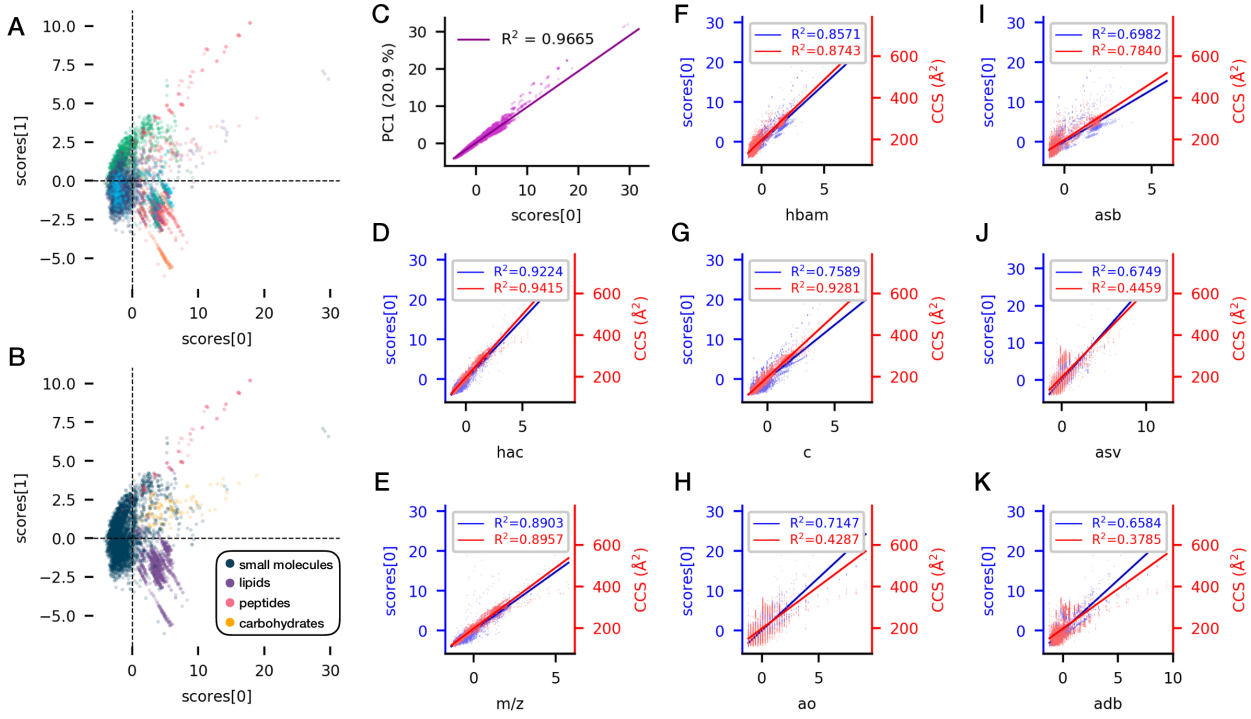
PCA projections of full CCS database onto principal axes 1, 2, and 3, colored by dataset (**A**, **B**) or chemical classification (**C**, **D**). Correlation of the top 3 molecular descriptors contributing to separation along PC1 (**E**-**G**) and PC2 (**H**-**J**). *hac* = heavy atom count; *m/z* = mass to charge ratio; *ao* = acyclic oxygen count; *hbd* = H-bond donor atoms; *ctv* = cyclic trivalent nodes; *r6* = 6-membered ring count.

2.5.3 *Top 3 Features Contributing to Separation Along PC3*



Plots of the top 3 MDs that contribute most strongly to variance captured by PC3. (**A**) cyclic single bond count, *csb* (**B**) nodes shared by ≥ 2 rings, *afr* (**C**) edges shared by ≥ 2 rings, *bfr*.

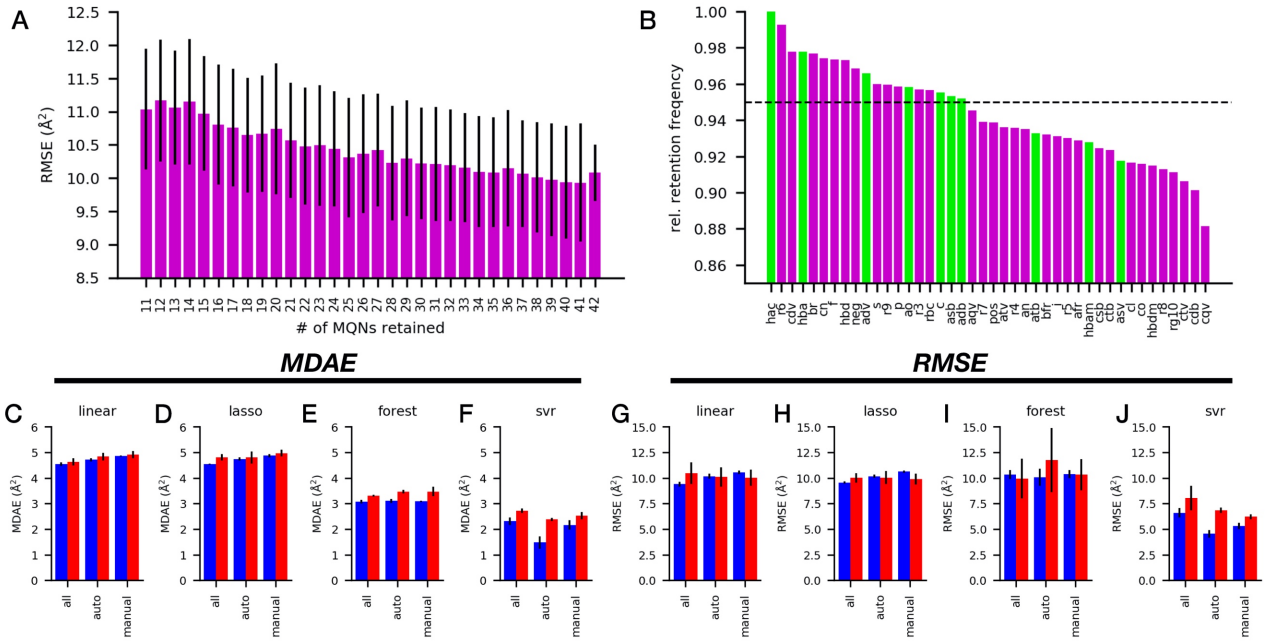
2.5.4 *PLS-RA on Combined CCS Database*



PLS-RA projections of full CCS database onto axes 1, 2, colored by dataset (**A**) or chemical classification (**B**). Correlation between PLS-RA projections along axis 1 and PCA projections along PC1 (**C**). Correlation between molecular descriptors and PLS-RA projections along axis 1 (blue) or CCS (red) for all compounds (**D**-**K**). *hac* = heavy atom

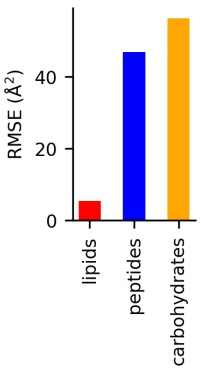
count; *m/z* = mass to charge ratio; *hbam* = H-bond acceptor sites; *c* = carbon atom count; *ao* = acyclic oxygen count; *asb* = acyclic single bonds; *asv* = acyclic single valent nodes; *adb* = acyclic double bonds.

2.5.5 *Feature Selection Trial Results*



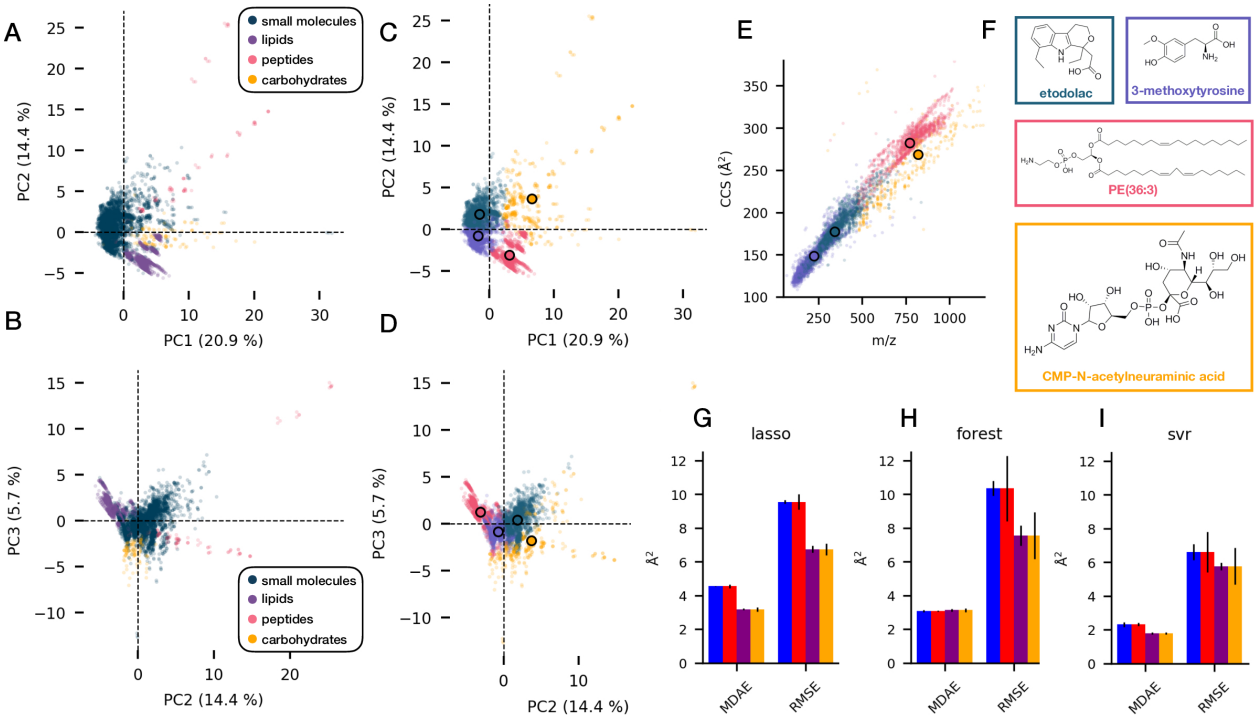
(**A**) Average CCS prediction performance as a function of the number of MQNs retained in the feature set. (**B**) Relative retention frequency of MQNs during randomized feature removal process. Average CCS Prediction performance by MDAE (**C**-**F**) and RMSE (**G**-**J**) on training (blue) and test set data (red) from 5 independent trials using either the full MQN feature set (all), the subset of MQNs selected in an automated fashion (auto) or the manually selected subset (manual). linear = linear regression; lasso = least absolute shrinkage and selection operator; forest = random forest regression; svr = support vector regression with rbf kernel.

2.5.6 *CCS Prediction Accuracy of LipidCCS on Different Chemical Classes*



Prediction accuracy (by RMSE) of LipidCCS[68] for compounds tagged as lipids (red), peptides (blue), and carbohydrates (gold).

2.5.7 *K-Means Clustering for Unbiased Assignment of Chemical Class*

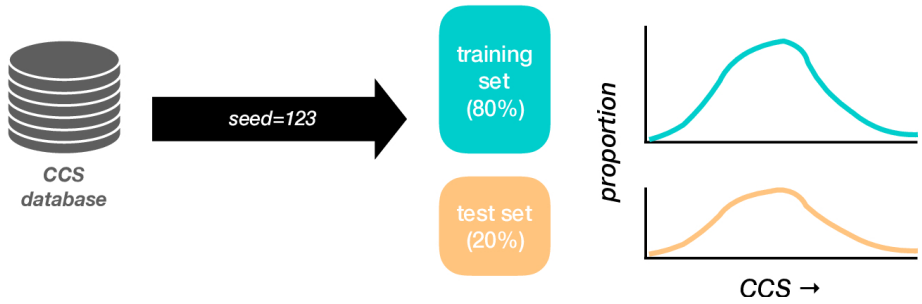


PCA projections of full CCS database onto principal axes 1, 2, and 3, colored by chemical class label (**A**, **B**), or by cluster (**C**, **D**). (**E**) Plot of CCS *vs*. *m/z* for full CCS database, colored by cluster. (**F**) Central structures within each cluster. (**G-I**) Average predictive performance of models (lasso, forest, svr, respectively) by MDAE and RMSE from 5 independent trials, trained on the full CCS database (training set = blue, test set = red) or on individual cluster datasets (training set = purple, test set = gold).

2.5.8 *Method for Predicting CCS Using K-Means Clustering*

Data Splitting

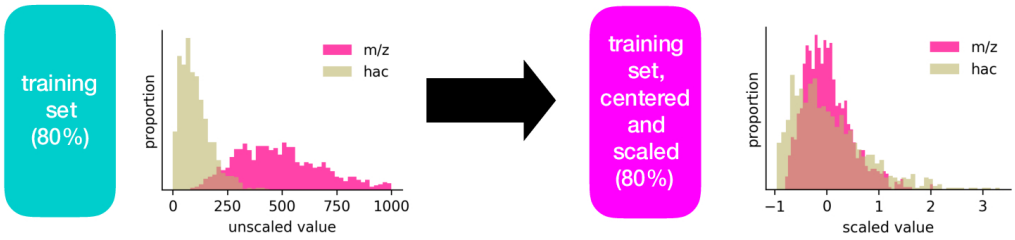
The complete CCS database is split into training and test sets, in a stochastic (but deterministic) fashion that attempts to preserve a similar distribution of CCS values in both sets.



This process can be repeated using different seed numbers to obtain distinct data splits for testing effects of various training conditions on model performance.

Data Centering/Scaling

Each feature in the training set is centered and scaled so that the mean becomes 0 and variance becomes 1. This step allows to all features to be considered equally regardless of their scale

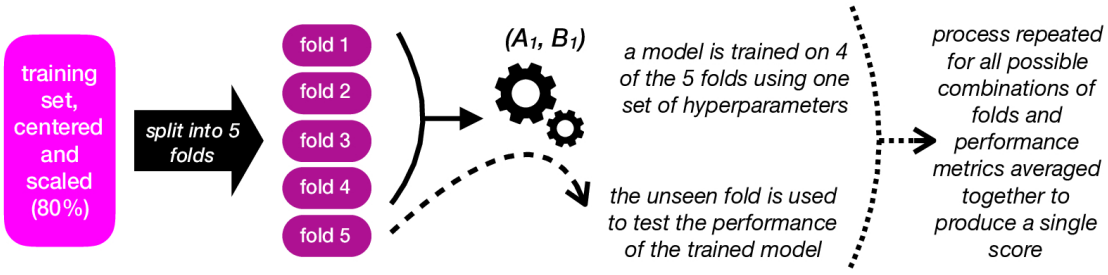


Hypothetical distributions of values for two features: m/z and hac (heavy atom count). These features have distinct scale and spread, and the larger m/z is expected to have a greater impact on model training accordingly.

After centering and scaling the two features have similar scale and variance, meaning that neither will have an arbitrarily large impact on model training.

Hyperparameter Optimization

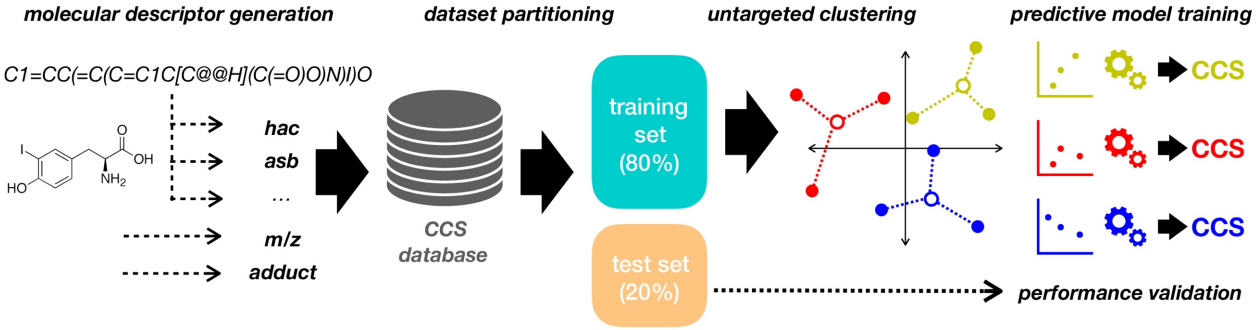
Certain models require hyperparameter optimization, which is done using 5-fold cross validation on the training data.



The above process is repeated using multiple combinations of hyperparameter values, and the combination with the best performance (typically minimum average MSE) is selected for training a model on the complete set of training data.

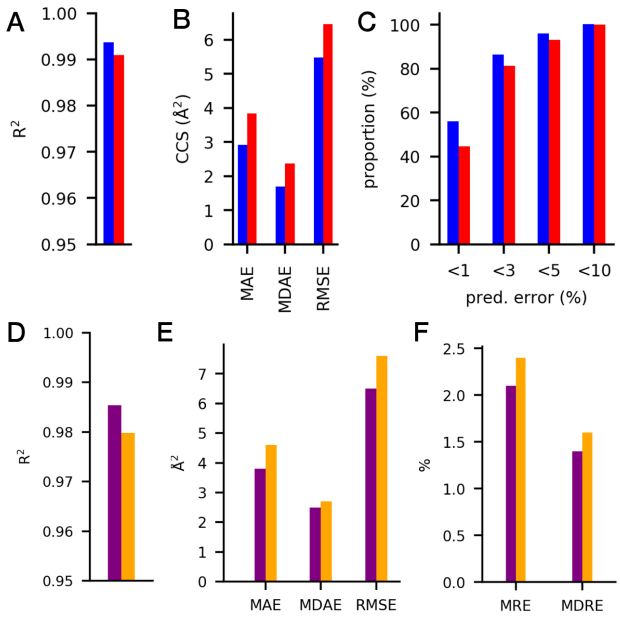
hyperparam.		Average MSE
A	B	
1	1	2.34
1	**2**	**1.23**
…	…	…
i	j	3.45

2.5.9 *CCS Prediction Model Training and Validation Workflow*



Workflow describing the process for training and validating the final prediction model. First, MQNs are generated from the compound SMILES structure in addition to the *m/z* and MS adduct and this data is stored in a database. The complete dataset is randomly partitioned into a training set and test set, preserving the approximate distribution of CCS values between the two sets. The training set is then fit using K-Means clustering to find the dominant groupings within the dataset in terms of chemical similarity. The data from each assigned cluster is then used to train an individual predictive model that is specialized for that group of compounds. Finally, the overall CCS prediction performance of this set of models is validated using the test set data by first assigning each sample to one of the fitted clusters then predicting CCS using the corresponding predictive model.

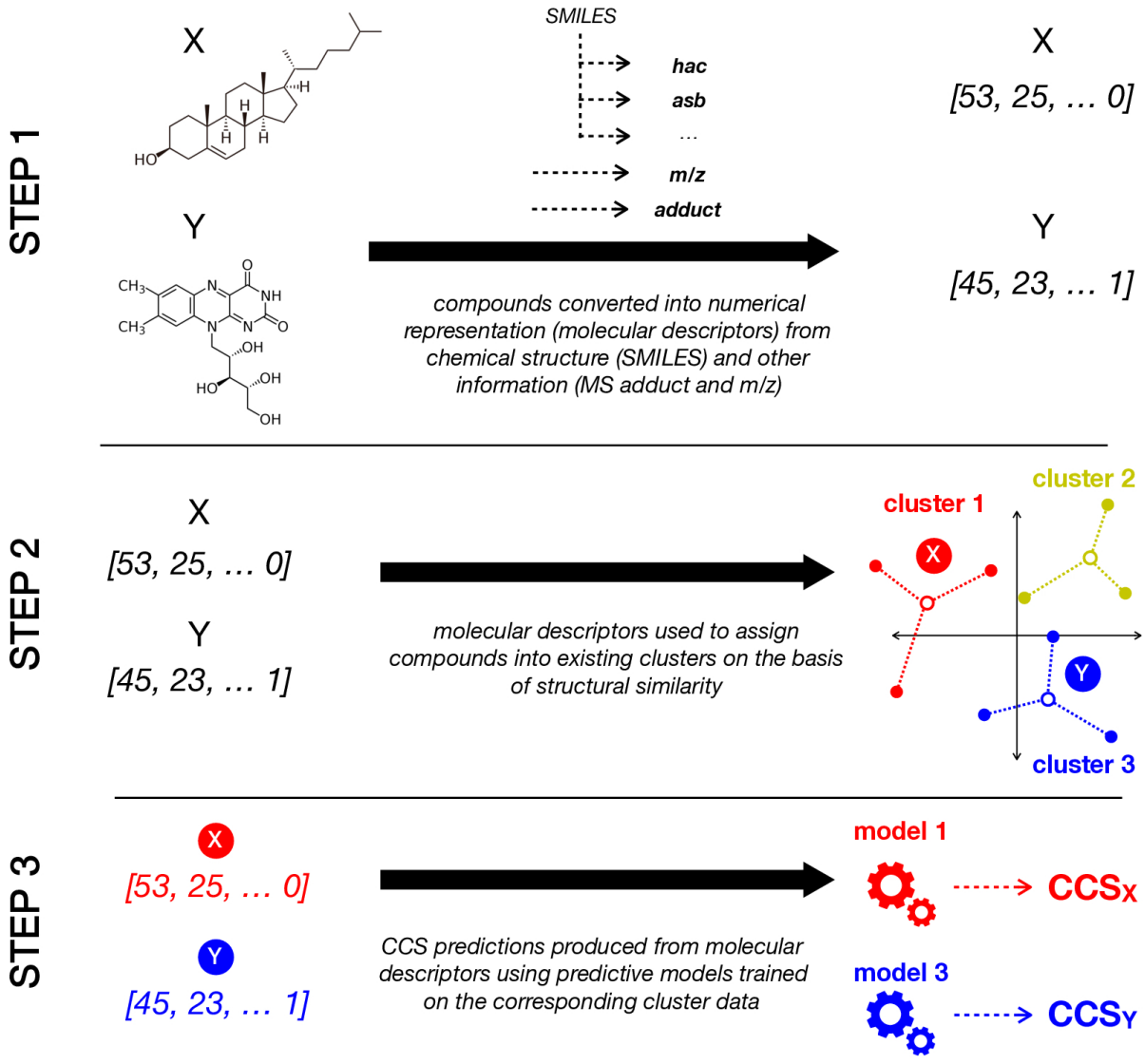
2.5.10 *Complete Performance Metrics for Final CCS Prediction Model*



(**A**-**C**) Complete performance metrics for final predictive model on training (blue) and test (red) data. (**A**) R^2 (**B**) mean/median absolute error and root mean squared error (**C**)

proportion of predictions falling within 1, 3, 5, and 10% of reference values. (**D-F**) Comparison of CCS prediction performance between final model (purple) and DeepCCS (gold) on all datasets used for training DeepCCS. (**D**) R^2 (**E**) mean/median absolute error and root mean squared error (**F**) mean/median relative error (MRE and MDRE).

2.5.11 Data Processing for Machine Learning



2.6 Tables

2.6.1 Molecular Quantum Numbers (MQNs)

MQN	description	MQN	description
c	carbon atom count	*hbdm*	H-bond donor sites
f	fluorine atom count	*hdb*	H-bond donor atoms
cl	chlorine atom count	*negc*	negative charges
br	bromine atom count	*posc*	positive charges
i	iodine atom count	*asv*	acyclic monovalent nodes
s	sulfur atom count	*adv*	acyclic divalent nodes
p	phosphorus atom count	*atv*	acyclic trivalent nodes
an	acyclic nitrogen atom count	*aqv*	acyclic tetravalent nodes
cn	cyclic nitrogen atom count	*cdv*	cyclic divalent nodes
ao	acyclic oxygen atom count	*ctv*	cyclic trivalent nodes
co	cyclic oxygen atom count	*cqv*	cyclic tetravalent nodes
hac	heavy (non-hydrogen) atom count	*r3*	3-membered ring count
asb	acyclic single bonds	*r4*	4-membered ring count
adb	acyclic double bonds	*r5*	5-membered ring count
atb	acyclic triple bonds	*r6*	6-membered ring count
csb	cyclic single bonds	*r7*	7-membered ring count
cdb	cyclic double bonds	*r8*	8-membered ring count
ctb	cyclic triple bonds	*r9*	9-membered ring count
rbc	rotatable bond count	*rg10*	≥10-membered ring count
hbam	H-bond acceptor sites	*afrc*	nodes shared by ≥2 rings
hba	H-bond acceptor atoms	*bfrc*	edges shared by ≥2 rings

2.6.2 MS Adduct Encodings

adduct	encoding
$[M+H]^+$	[1, 0, 0, 0, 0, 0, 0]
$[M+Na]^+$	[0, 1, 0, 0, 0, 0, 0]
$[M-H]^-$	[0, 0, 1, 0, 0, 0, 0]
$[M+Na-2H]^-$	[0, 0, 0, 1, 0, 0, 0]
$[M+NH4]^+$	[0, 0, 0, 0, 1, 0, 0]
$[M+K]^+$	[0, 0, 0, 0, 0, 1, 0]
other	[0, 0, 0, 0, 0, 0, 1]

Chapter 3 Characterization of the Impact of Drug Metabolism on the Gas-Phase Structures of Drugs Using Ion Mobility-Mass Spectrometry

Portions of this chapter have been adapted and reproduced with permission from:

Dylan H. Ross, Ryan P. Seguin and Libin Xu, Characterization of the Impact of Drug Metabolism on the Gas-Phase Structures of Drugs Using Ion Mobility-Mass Spectrometry, *Analytical Chemistry*, 91 (**2019**) 14498-14507.

3.1 Introduction

Drug metabolism is the process by which drugs are chemically modified by human body, generally facilitating their removal from the body.[16] This process is carried out by a diverse assortment of drug-metabolizing enzymes, most of which catalyze oxidation of or conjugation of specific chemical groups to target compounds.[17-19] Specifically, drug metabolism reactions include Phase-I metabolism, such as oxidation and dealkylation catalyzed by cytochrome P450s (CYPs), and Phase-II metabolism, such as conjugation with glutathione and glucuronic acid catalyzed by glutathione S-transferases (GSTs) and UDP-glucuronosyltransferase (UGTs), respectively (Figure 3.5.1 A). Understanding whether and how drugs are metabolized is a critical component of the drug development process, since metabolism is often an important determinant of drug clearance and metabolites may exert unexpected bioactivities or toxicities.[20, 21] Mass spectrometry (MS) has played an essential role in advancing the field of drug metabolism. Over the past decade, liquid chromatography (LC)-coupled with tandem MS and high-resolution MS, in combination with various scan modes and post-acquisition data processing methods, has greatly increased the specificity and sensitivity of drug metabolite identification.[22, 24] However, the throughput of such identification processes is low as LC separation could be lengthy and/or post-acquisition processing methods need to be customized for each drug. Furthermore, structural information obtained from LC-MS analysis is limited. Definitive assignment of metabolite structure can be achieved using NMR spectroscopy, but this technique requires large

amount of materials and thus is both costly and low throughput.[22] There exists, therefore, a demand for techniques that allow enhanced structural characterization of drug metabolites without sacrificing analytical throughput.

Ion mobility (IM) spectrometry is a rapid gas-phase separation technique based on the size and shape of the analytes in the gas phase, which is orthogonal to the polarity-based LC separations.[29, 31, 32, 75, 76] In IM separation, ions are driven through a neutral buffer gas under the influence of either a static electric field (as in drift tube IM, DTIM) or a dynamic electric field (as in traveling wave IM, TWIM),[37, 77] where they are differentially slowed down due to their interactions with the gas molecules. The mobility of an ion in IMS is determined by its collision cross section (CCS), a unique physical property determined by its size, shape, and long-range and short-range interactions with the drift gas in the gas-phase. When IM is coupled to MS (IM-MS), a two-dimensional separation on the basis of CCS-to-charge ratio and mass-to-charge ratio (m/z) is achieved. Due to the rapid nature of the IM separation (millisecond scale), IM-MS can be coupled with traditional liquid chromatography to provide three-dimensional separation (LC-IM-MS) without sacrificing the throughput. Furthermore, the orthogonal separation by IM-MS over traditional MS can enable the abbreviation, or even complete omission at times, of LC separations to enhance the analytical throughput.

In recent years, several groups, including ours, have measured CCS values for large collections of compounds, covering a diverse range of biomolecule and small molecule chemical space.[34, 45-47, 68, 83, 85-87, 89-91] These studies have revealed a number of characteristic trends in CCS *vs.* mass associated with specific chemical classes, attributable to the intrinsic relationship between CCS and chemical structure. More specifically, for a given *m/z*, compounds from different chemical classes display distinct ranges of CCS values, owing to different degrees of

structural compactness among different molecular classes. Different classes of biomolecules (*e.g.* lipids, peptides, carbohydrates) follow distinct trends and only occupy narrow regions of the CCS-mass two-dimensional conformational space.[45] Our recent report on 1440 CCS values of 1425 drug and drug-like small molecules suggested that overall, these molecules cover a much broader range of molecular space than individual biomolecular classes, but a specific class of drugs, such as various classes of antimicrobials, corticosteroids, ion channel blockers, etc., tend to occupy a tight and distinct region in the CCS-mass conformational space.[47] A few studies have been carried out to characterize drug metabolites by IM-MS, such as products of aromatic hydroxylation and glucuronidation,[96-98] which demonstrated some advantages of IM-MS in differentiating positional isomers. However, no systemic study has been conducted to explore the structural characteristics of drug metabolites and how various chemical modifications change the gas-phase behavior of the metabolites relative to their corresponding parent compounds in IM-MS analysis.

3.2 Results and Discussion

3.2.1 In Vitro Biosynthesis of Drug Metabolites and Measurement of Their CCS

We first aimed to develop an *in vitro* biosynthesis system to generate metabolites with diverse structures from a diverse range of drugs so that the effect of metabolic transformations on gas-phase structures of drugs can be systemically studied. Human liver fractions, such as S9 and microsomal fractions, are commonly used for *in vitro* drug metabolism studies. The S9 fractions contain the cytosol, which mainly contains GSTs, sulfotransferases, and other water-soluble enzymes, and microsomes, which mainly contain CYPs and UGTs. We used the drug, midazolam, as a probe molecule to optimize the *in vitro* system conditions. Midazolam is known to undergo mono and di-hydroxylation, direct glucuronidation, and consecutive

hydroxylation/glucuronidation reactions, so it would be a good probe for both Phase-I and Phase-II reactions (Figure 3.5.2).[99] We found that although the S9 fraction contains all the necessary drug-metabolizing enzymes, addition of a separate HLM fractions to the S9 fraction is needed to generate sufficient amounts of primary metabolites by CYPs for the formation of secondary metabolites by UGTs and GSTs. Alamethicin, a pore-forming polypeptide, was also added to the S9/HLM stock to enable access of the potential substrates to UGTs, which are located on the lumen side of microsomes.[100] In order to represent the range of metabolic modifications, we chose 19 structurally diverse drug or drug-like compounds that readily undergo Phase-I and/or Phase-II drug metabolic reactions (Figure 3.5.1 B). For each drug, the reactions were carried out with or without enzyme activating cofactors, such as NADPH (for CYPs) and UDPGA (for UGTs) so that enzyme-specific products can be identified. A workflow of this process and the subsequent steps is shown in Figure 3.5.1 C.

Each reaction mixture was analyzed using a FIA-IM-MS method. $^{TWIM}CCS_{N2}$ values were calibrated using a mixture of polyalanines (n=2-14) and drug-like compounds with known $^{DTIM}CCS_{N2}$ values as described previously.[47] 37 metabolites were consistently observed from the 19 parent compounds. CCS values were obtained for all parent compounds and metabolites in triplicate, with high reproducibility (average of 0.3% inter-day RSD). Figure 3.5.3 A summarizes the CCS *vs. m/z* values for all compounds and their metabolites, grouped by various routes of metabolism. The curve and shade in the background of the plot represents the power fit and the ± 10% range of the 1440 drug and drug-like CCS values we reported previously,[47] which illustrates the conformational space of this set of compounds and their metabolites relative to the broad trend for drug-like compounds.

Additionally, a dimensionless compaction factor (C) was calculated according to the following equation to quantify the relative structural compaction/expansion of metabolites relative to their corresponding parent compounds:

$$\frac{\mathrm{CCS}_{parent}}{\mathrm{CCS}_{metabolite}} = C\left(\frac{mass_{parent}}{mass_{metabolite}}\right)^{2/3}$$

This equation is derived from the relationship between the masses of spheres having the same density and their areas, and is adapted from a similar analysis recently applied to study the hydrophobicity of amino acids as it correlates to the packing efficiency of the amino acid oligomers.[101] The dimensionless factor C describes deviation from ideal isotropic growth (with respect to CCS and mass changes due to metabolism), with C < 1 indicating the metabolite becomes less compact than the parent while C > 1 indicating the metabolite becomes more compact than the parent. Calculated compaction factors for all metabolites relative to their parent compounds are summarized in Table 3.6.1.

Considering all parent compounds and metabolites together, it is difficult to discern any characteristic global trends in CCS *vs. m/z* for specific routes of metabolism as most of the parent compounds and their +O metabolites and/or dealkylation metabolites occupy a relative tight space close to the center power-fit line. On the broad scale, conjugation products with glucuronic acid (GA) and glutathione (GSH) separate from Phase-I metabolites, mostly attributable to the large increase in mass.

However, closer examination of the trend of changes for individual compounds suggest that the conformational changes due to different routes of metabolism are dependent upon the structure of the parent compound. For compounds like clomifene, clozapine, and thioridazine (Figure 3.5.3 B-D), primary metabolites from oxygenation or dealkylation reactions displayed relatively linearly increased or decreased CCS, corresponding to the change in mass imparted by

the metabolic modification (*e.g.*, oxygenated metabolites tended to have increased CCS coinciding with the increase in mass from the addition of oxygen). Furthermore, secondary metabolites formed from a combination of oxygenation and dealkylation reactions displayed CCS shifts corresponding to the net change in mass from the different modifications. These observations suggest that for some compounds, metabolic modifications do not significantly alter the gas-phase conformations of the parent compounds beyond the simple increase/decrease in size imparted by the metabolic modification itself. This conclusion is further supported by the fact that for most of these metabolites, compaction factor values tended to be very close to 1, indicating a similar gas-phase packing efficiency between parent and metabolite.

Unusual trends in CCS changes upon metabolism were observed for some lipid-like molecules, such as the series of BACs with alkyl chain lengths of 10, 12, 14 and 16 carbons and terfenadine. Oxygenated metabolites were observed in these reactions, and each had CCS values smaller than their parent compounds despite their increased mass. The compaction factors for each of these metabolites were all >1, indicating structural compaction in the metabolites relative to the parent compounds. A common CYP-mediated metabolic modification for compounds containing long alkyl chains in fatty acids is hydroxylation at the ω or ω-1 position.[102] Indeed, we recently confirmed the formation of the ω and ω-1-hydroxyl-products as the major metabolites of these BACs.[103] For the BACs, this would introduce a polar moiety at the opposite end of the positively charged quaternary ammonium group, leading to a potential intramolecular ion-dipole interaction that could explain the observed structural compaction of the gas-phase conformations of the metabolites relative to the parent compounds. Additionally, the magnitude of this effect increased with alkyl chain length, suggesting dependence on the flexibility of the alkyl chain. To further investigate this effect, a larger series of BACs with alkyl chain lengths

ranging from 4 to 16 carbons and their corresponding ω-OH metabolites were synthesized and analyzed by IM-MS (Figure 3.5.3 E). The parent BACs and their ω-OH metabolites occupied distinct trends in the CCS *vs. m/z* plot, with the metabolite trend falling below that of the parent compounds. While ω-OH derivatives of short chain BACs still increased relative to the parent compounds, the extent of increase is small relative to the power-fit trend line of the parent compounds. C8 BAC appears to be the turning point, where the CCSs of the metabolites become lower than the matching parent compounds. Notably, the magnitude of the compaction factors increases with the chain length, consistent with the observed chain-length dependence of the structural compaction of the metabolites.

A similar trend to the BACs, though at a smaller magnitude, was also observed between terfenadine and its oxygenated metabolite (Figure 3.5.3 F). The compaction of the +O metabolite relative to the parent compound is evident from the compaction factor of 1.042. Terfenadine has structural similarity to the BACs, being a somewhat linear, hydrophobic molecule with positive charge localized on one side. The introduction of a polar moiety at the opposite end of the molecule could allow for intramolecular interaction with the positive charge, similar to the ω-OH metabolites of BACs. Indeed, terfenadine is known to be hydroxylated at the tert-butyl group in hepatic metabolism (principally by CYP3A4),[104] supporting this proposed mechanism of structural compaction. The decreases in CCS values of metabolites despite their added mass is counterintuitive to the expected relationship between mass and CCS, illustrating the complexity of conformational changes that can be induced by metabolic modifications.

The antibacterial agent triclosan is a small molecule containing three chlorine atoms, making it considerably dense in the gas phase relative to other drug-like small molecules (Figure 3.5.3 G). Both the protonated species and sodiated adduct of the parent compound were

observed, with the sodiated adduct displaying a larger CCS relative to its mass than the protonated species. A primary glucuronide metabolite was observed (triclosan +Glc), in addition to a secondary glucuronide metabolite putatively formed from an oxygenated primary metabolite (triclosan +O, +Glc). The primary glucuronide is more structurally compact for its mass than the secondary glucuronide, which is on the similar trend to the adducts of the parent compounds in the overall drug-like molecule CCS *vs. m/z* 2D plot. Indeed, the primary and secondary glucuronides have compaction factors of 1.174 and 1.102, respectively, indicating that relative to the parent compound, the primary glucuronide undergoes more structural compaction upon metabolism than the secondary glucuronide.

Amlodipine (Figure 3.5.3 H) contains a central dihydropyridine ring that can undergo a two-electron oxidation to form the pyridine analog (amlodipine -2H). The CCS of the pyridine metabolite is not significantly different from the parent (compaction factor = 1.002), however, comparison of the desethylated metabolite (amlodipine -Et) to its pyridine analog (amlodipine -2H, -Et) reveals a distinct increase in CCS upon dehydrogenation (-2H). Indeed, comparing both the desethylated metabolite and its -2H (pyridine) derivative to the parent compound, the compaction factors were 1.043 and 0.972, respectively, which quantitatively supports the large reduction in the compactness of the dehydrogenation metabolite. In other words, dehydrogenation of the dihydropyridine ring has no observable effect on the gas-phase conformation of the parent compound, but induces a dramatic increase in the size of the desethylated metabolite. This observation illustrates that even for similar or related compounds, the structural changes imparted by a given modification may differ significantly.

3.2.2 Post-Mobility Fragmentation to Facilitate Metabolite Identification

A separate IM-MS analysis was performed on the reaction mixtures in which post-IM fragmentation was conducted to further validate the structures of the metabolites. For some compounds, the site of metabolic modification is known based on literature or apparent from the compound's structure, such as for deethylation of clomifene (Figure 3.5.4 D) or glucuronidation of diclofenac (Figure 3.5.4 H). Other compounds, however, contain multiple sites where metabolic modifications may occur, leading to potential positional isomers. Post-IM fragmentation could help differentiate such positional isomers. For example, Figure 3.5.5 A shows the ATDs of dextromethorphan (DEX) and several fragments observed in post-IM fragmentation experiments, where all of the observed fragment ATDs align with that of the parent compound, supporting their assignment as fragments of DEX. DEX contains two methyl groups attached to heteroatoms (O and N), both of which can undergo dealkylation reactions.[105] The metabolites resulting from either dealkylation are isobaric, and there is no indication of multiple metabolites being present based on the ATD of the demethylated metabolite of DEX (DEX -Me, Figure 3.5.5 B). Of the several fragments of the parent DEX, three contained the O-methyl moiety but not the N-methyl moiety, making these fragments diagnostic of demethylation at the O position. Indeed, the demethylated analogues of these three fragments were observed from the DEX -Me metabolite, confirming the identity of this metabolite as the O-desmethyl product. This observation is consistent with the previous report that the dominant hepatic metabolite of DEX is the loss of the O-methyl group over the N-methyl group.[105] This example illustrates the degree of specificity in identification of metabolites that can be reached by combining fragmentation data with structural information from IM-MS analyses. Additional fragmentation ATDs of other parent compounds and metabolites can be found in Figure 3.5.4.

3.2.3 Computational Modeling of Unusual Behavior of BACs, Terfenadine, and Their Oxygenated Metabolites in IM-MS

A chain length-dependent decrease in CCS upon ω-hydroxylation was observed for a series of BACs with alkyl chain lengths of 4 to 16 carbons (Figure 3.5.3 E). Using computational modeling and theoretical CCS calculation with MobCal (as described in Experimental Section), we examined the structural basis for this observed gas-phase compaction upon ω-hydroxylation of BACs. Figure 3.5.6 A summarizes the theoretical CCS for BACs and their +O metabolites, in comparison with their measured values. The theoretical CCS for the parent compounds show excellent agreement with their corresponding measured values, with the exception of C16 BAC, for which the theoretical CCS is significantly lower than the measured value. The modeled CCS values for the +O metabolites are systematically underestimated relative to the experimental values, with an average error of -3.7%. Plotting the theoretical CCS *vs.* the experimental values (Figure 3.5.6 B) more clearly illustrates the agreement between the theoretical and experimental CCS for the parent compounds, as well as the systematic underestimation in the calculated CCS values for the metabolites. The alkyl chain length-dependent decrease in CCS upon ω-hydroxylation becomes particularly evident by plotting the ratio of parent to +O metabolite CCS, which displays a steady increase with alkyl chain length for the measured values (Figure 3.5.6 C). Despite the systematic errors in the modeled CCS of the +O metabolites, the overall trend in this ratio is recapitulated, albeit with greater variance. These data indicate that the theoretical simulations sufficiently captured the structural characteristics of BACs and their ω-OH metabolites to reproduce the observed structural compaction occurring upon ω-hydroxylation.

In order to gain insight on what structural characteristics were driving the observed trends in CCS, all of the simulations used to produce the theoretical CCS values were analyzed for the

distributions of intramolecular distance between the nitrogen of the quaternary ammonium moiety and the ω-carbon of the alkyl chain (Figure 3.5.6 D and E). These distance distributions reflect the overall degree of intramolecular interaction between the two ends of these molecules, which in turn influences their structural compactness and therefore CCS. For the parent BACs, the distance distributions generally increased in spread and magnitude with increasing alkyl chain length, with the exception of C16 BAC, which displayed a much more compact distribution with a lower magnitude for its size relative to the other BACs. Comparing this result to trend in theoretical CCS for fully extended BAC conformers (Figure 3.5.6 A), it seems that the computational error for C16 BAC is likely due limitations in the semi-empirical molecular modeling theory used here to represent systems beyond a certain size. Overall, these results indicate that the observed increase in CCS with alkyl chain length for the parent BACs can be justified by their increasingly dynamic alkyl chains, and thus having a lower degree of overall structural compactness. In contrast, the +O metabolites displayed very compact distance distributions that are nearly invariant with alkyl chain length, indicating a strong propensity toward intramolecular interaction between the ω-OH group and the quaternary ammonium group. The consequence of this interaction is a structural compaction that leads to a lower trend than the parent compounds in the CCS-mass 2D plot.

However, we note that the observed trend of the theoretical parent CCSs is significantly lower than the trend of the CCS calculated from the fully extended parent conformations, which indicates that the parent compounds, being highly dynamic, adopt intermediately compact conformations between the fully extended conformations and those of the metabolites. It is also possible that the addition of the highly polarized OH in the metabolites could increase the potential for long-range interactions with the drift gas relative to the parent BACs, making

apparent CCS for the metabolites larger than one may expect on the basis of size alone. While both factors could contribute to the smaller differences between the trendlines of the measured CCS values of the parents and the metabolites than one might expect from the fully extended and the fully compact conformations, the intermediate compactness of the parent compounds appear to the major factor.

Computational modeling was also performed on terfenadine and its oxygenated metabolite, since these displayed similar behavior to the BACs and their ω-OH metabolites. The theoretical CCS values for terfenadine and its hydroxylated metabolite were both larger than the experimental values (Figure 3.5.6 F), however, the relationship between the parent and metabolite CCS values was recapitulated: the metabolite displayed slightly lower CCS than the parent, indicating some degree of structural compaction due to the metabolic modification. Representative structures for the parent and hydroxylated metabolite (Figure 3.5.6 G) support the notion that an intramolecular polar-polar interaction involving the hydroxyl group drives structural compaction in the metabolite.

Taken together, these results demonstrate that metabolic modifications can impact the structural characteristics of certain compounds with considerable magnitude when intramolecular interactions are either introduced or disrupted. Furthermore, these structural effects may not be strictly mediated by the immediate chemical environment of the metabolic modification, but could be through longer-range intramolecular inter-actions with the modifications.

3.2.4 Bimodal ATD of Quercetin Glucuronides

Both the protonated and sodiated adducts of quercetin glucuronide (quercetin +Glc) displayed bimodal ATDs (Figure 3.5.7 A and B). A bimodal ATD for a single mass in IM-MS generally arises from conformational heterogeneity or the presence of structurally distinct

isomers, either constitutional (*e.g.* different sites of protonation, sodiation)[47, 106] or diastereomeric. The observation of similar behavior between the protonated and sodiated adducts suggests the presence of more than one constitutional isomer, rather than conformational heterogeneity or different ionization sites since these effects are likely to present differently between different types of MS adducts. Quercetin contains five hydroxyl groups (Figure 3.5.7 B), and glucuronidation has been observed to occur at all but the 5-position in hepatic metabolism, with the 7-position being the dominant regioisomer.[107] Therefore, it is likely that the two peaks observed in the ATD for quercetin glucuronide correspond to a mixture of these regioisomers.

Post-IM fragmentation was not informative on the positional isomers as no cross-ring or between-ring fragments were observed. Therefore, we used computational modeling and CCS calculation to gain some in-sight on the gas-phase conformations of both protonated and sodiated adducts of all regioisomers of quercetin glucuronides (including the 5-glucuronide). In contrast to the BACs, which had many energetically similar conformations due to their structural flexibility and dynamic nature, the glucuronides of quercetin have less conformational flexibility and are thus appropriately represented by single minimum energy structures obtained from a large ensemble of conformations. The lowest energy conformations for all protonated glucuronide isomers are presented in Figure 3.5.7 B, and their modeled CCS values are summarized in Figure 3.5.7 C. Comparison of the theoretical CCS values with those obtained from the two peaks observed experimentally (dotted lines) suggest that the smaller CCS value has contributions from the 3-, 3'-, and/or 4'-glucuronides, while the larger CCS value can be attributed to the 7-glucuronide, since the 5-glucuronide is not expected to be formed in hepatic metabolism. The same results were observed from computational modeling of the sodiated adducts and are detailed in figure 3.5.8. These results demonstrate that for some compounds,

insights from IM data and computational modeling can be used to elucidate the contributions of various regioisomers to a sample and guide further inquiry.

3.3 Experimental

3.3.1 Materials

Poly-DL-alanine, acetaminophen, betaine hydrochloride, and drug standards were purchased from Sigma-Aldrich (St. Louis, MO). Drug CCS calibrants were obtained as descried previously.[47] Human liver microsomes (HLM) and S9-fraction (S9) pooled from 100 male and 100 female individuals were purchased from Sekisui XenoTech (Kansas City, KS). C12, C14, and C16 benzalkonium chlorides (BACs) and the synthetic precursors necessary for synthesizing additional BACs and ω-hydroxy BAC analogs were purchased from Sigma-Aldrich (St. Louis, MO).

3.3.2 Synthesis of Benzalkonium Chlorides (BACs) and Their ω-OH Metabolites

BACs were synthesized by nucleophilic coupling of N,N-dimethylbenzylamine with 1.4 equivalents of an alkyl chloride of the appropriate alkyl chain length in ethanol heated under reflux.[108] The corresponding ω-hydroxy versions of each BAC were prepared in the same manner by substituting the alkyl chloride with the appropriate ω-hydroxy alkyl chloride. Products were recrystallized from hot acetone and rinsed with cold diethyl ether. Chemical identities were confirmed by ^{1}H-NMR and/or high-resolution mass spectrometry (HRMS, see below).

3.3.3 ^{1}H-NMR and HRMS Characterization of Synthesized BACs and ω-OH Metabolites

C4 BAC: HRMS $[M]^{+}$ ($C_{13}H_{22}N$): observed, 192.1747; theoretical: 192.1752.

C6 BAC: HRMS $[M]^{+}$ ($C_{15}H_{26}N$): observed, 220.2063; theoretical: 220.2065.

C8 BAC: HRMS $[M]^{+}$ ($C_{17}H_{30}N$): observed, 248.2375; theoretical: 248.2378.

C10 BAC: HRMS $[M]^+$ ($C_{19}H_{34}N$): observed, 276.2694; theoretical: 276.2691.

ω-OH C4 BAC: HRMS $[M]^+$ ($C_{13}H_{22}NO$): observed, 208.1694; theoretical: 208.1701.

ω-OH C6 BAC: HRMS $[M]^+$ ($C_{15}H_{26}NO$): observed, 236.2010; theoretical: 236.2014.

ω-OH C8 BAC: HRMS $[M]^+$ ($C_{17}H_{30}NO$): observed, 264.2327; theoretical: 264.2327.

ω-OH C10 BAC: ^{1}H-NMR (CDCl3, 500 MHz): 1.29 and 1.36 (br s, 12H), 1.56 (m, 2H), 1.81 (m, 2H), 3.30 (s, 6H), 3.53 (m, 2H), 3.64 (q, 2H, J = 6.1 Hz), 5.03 (s, 2H), 7.47 (m, 3H), 7.64 (d, 2H, J = 7.0 Hz); MS $[M]^+$ ($C_{19}H_{34}NO$): observed, 292.2639; theoretical: 292.2640.

3.3.4 In Vitro Drug Metabolite Generation

Drug metabolites were generated in vitro using pooled human liver microsomes (HLM) and S9-fraction (S9) in a 96-well plate format. First, a mixture of HLM and S9 (0.2 mg protein/mL, each) containing GSH (5 mM) and $MgCl_2$ (5 mM) was prepared in buffer (0.1 M phosphate, pH 7.4). Alamethicin was added at 0.01 mg/mL, and the mixture incubated on ice for 20 min. 90 µL of the pre-treated HLM/S9 mixture was added to each well of a 96-well plate, then 0.5 µL of each drug stock (10 mM in DMSO) were added to duplicate wells. An activation mixture containing 10 mM NADPH and 50 mM UDPGA was prepared in the same buffer. The plate was heated to 37 °C in a water bath, then reactions were initiated by addition of 10 µL of activation mixture (or buffer for cofactor-free controls). The reactions proceeded at 37 °C for 40 min, then the plate was removed from the water bath and cooled on ice for 10 min. Subsequently, 100 µL of ice-cold acetonitrile containing MS internal standards was added to each well to quench the reactions and precipitate proteins. The plate was centrifuged at 3500 x G for 15 min at 4 °C to pellet precipitated proteins, and the supernatant was transferred to a fresh 96-well plate for IM-MS analysis.

3.3.5 Ion Mobility-Mass Spectrometry Analysis

TWIM-MS analysis was performed on a Waters Synapt G2-Si mass spectrometer (Waters Corp., Milford, MA) equipped with an electrospray ionization (ESI) source using nitrogen as the drift gas. ESI conditions are detailed in the Supporting Information. Mass calibration was performed using sodium formate for the range of *m/z* 50−1200. IM separations were performed at a traveling wave velocity of 500 m/s and height of 40 V. For post-IM fragmentation analyses, collision energy was added to the transfer region using a ramp from 30 to 50 V. FIA was performed with a Waters Acquity FTN UPLC connected to the ESI source of the IM-MS. Sample injections (5 μL) were made using a 0.3 mL/min flow of 50% water with 0.1% formic acid / 50% methanol with 0.1% formic acid. Data was acquired for 0.5 min with a 1s scan time over *m/z* 50−1200, which resulted in approximately 14 scans across the eluted peak from FIA. The full 96-well plate was analyzed on three separate days over two months.

3.3.6 Ion Mobility-Mass Spectrometry Electrospray Conditions

IM-MS analysis was performed on a Waters Synapt G2-Si HDMS (Waters Corp., Milford, MA) equipped with an electrospray ionization (ESI) source using nitrogen as the drift gas. ESI conditions were as follows: capillary, +2.5 kV; sampling cone, 40 V; source temperature, 100 °C; desolvation temperature, 250 °C; cone gas, 50 L/h; and desolvation gas, 600 L/h.

3.3.7 TWIM CCS Calibration

Drift tube CCS values in nitrogen ($^{DTIM}CCS_{N2}$) for a series of singly-charged polyalanines (n = 2-14) and a mixture of drug-like compounds were used for calibration of TWIM drift times into CCS ($^{TWIM}CCS_{N2}$), as described previously for the measurement of CCS values of 1425 drugs.[47] Briefly, arrival time distributions (ATDs) for polyalanines and drug-like

CCS calibrants were extracted from the raw data using accurate mass with a window of ± 0.01 Da and a CCS calibration curve was constructed from reference data in an automated fashion using a Python script developed in-house (see Data Analysis below). Drift times for each calibrant were obtained as the mean from a least-squares fit of a Gaussian function on the ATD. Drift times were corrected for mass-dependent flight time to give the corrected drift times (t_d') and reference CCS values were corrected for ion charge state (Z) and reduced mass with the drift gas to give the corrected CCS (CCS').[109] A calibration curve was generated by fitting these corrected values to a function of the form $CCS' = A(t_d' + t_0)^B$, where A, t_0 and B were the fitted parameters.[33, 110] The CCS calibration curve displaying randomly distributed fit residuals with a maximal absolute error < 3% was considered acceptable.

3.3.8 Data Analysis

IM-MS data analysis was performed using a Python script developed in-house. The data analysis script uses the Waters MassLynx SDK (https://interface.waters.com/masslynx/) to directly extract ATDs from raw data files using accurate mass and a selection window of ± 0.01 Da. A list of accurate masses for ions of parent compounds and their expected metabolites was compiled by manual inspection of the raw data and consulting the literature for commonly observed human hepatic metabolites. Using this list of masses, ATDs were extracted for all parent compounds and observed metabolites, and drift time of each compound was obtained automatically by least-squares fitting of the Gaussian function to each ATD. All ATDs along with Gaussian fits and residuals were manually inspected for sufficient intensity (>1e3), appropriate peak width (FWHM < 1 ms), and apparent lack of multimodality. For each observed metabolite, ATDs were extracted from both cofactor-containing reaction samples and cofactor-

free controls to establish co-factor-dependent formation as confirmation of the presence of a metabolite.

3.3.9 Computational Modeling and CCS Calculation

Parent compounds or metabolites displaying unusual or complex IM behavior were further investigated using computational modeling and CCS calculation. Compound SMILES structures were obtained from PubChem and manually modified to reflect the relevant ESI adduct and metabolic modifications (if present). OpenBabel[111] was used to generate initial 3D structures from the SMILES strings (using the MMFF94[112] force field) and produce inputs for further structural optimization in the semi-empirical molecular modeling software MOPAC.[113] Initial structures were further optimized in MOPAC in two steps: a rough optimization, followed by a more precise optimization and calculation of thermodynamic properties. The roughly optimized structure was used as the input for a dynamic reaction coordinate (DRC) simulation in MOPAC: a time-resolved simulation in which the system is allowed to exchange kinetic and potential energy while conserving total internal energy (akin to traditional molecular dynamics). For each DRC simulation, excess kinetic energy was added to the system such that the total internal energy (potential + kinetic) of the system matched the vibrational energy predicted for the simulation temperature (310 K) from the thermodynamic calculations. The resulting simulation trajectories were analyzed using tools from the GROMACS software suite.[114] For each trajectory, plots of various system characteristics (potential energy, radius of gyration, *etc.*) *vs.* simulation time were generated and visually inspected for aberrant behavior. For each compound, initial structure generation was repeated 6 times, resulting in slightly different starting structures and therefore distinct simulation results. All replicate simulation trajectories for a single compound were concatenated and all sampled conformations were used for

clustering analysis using GROMACS. Central structures from each of the fitted clusters were assigned weights on the basis of cluster size, and structures with weights contributing at least 3% to the overall average were carried through to CCS calculation. Theoretical CCS values in N_2 were calculated for each central structure using a trajectory method in MobCal modified for using N_2 as a drift gas.[58, 60] Each CCS calculation was performed in triplicate using different pseudo-random number generator seeds and an averaged CCS value was obtained for each conformation. The final theoretical CCS value for each compound was obtained as the weighted average of all relevant clustered structures and their calculated CCS values, using the previously determined weights.

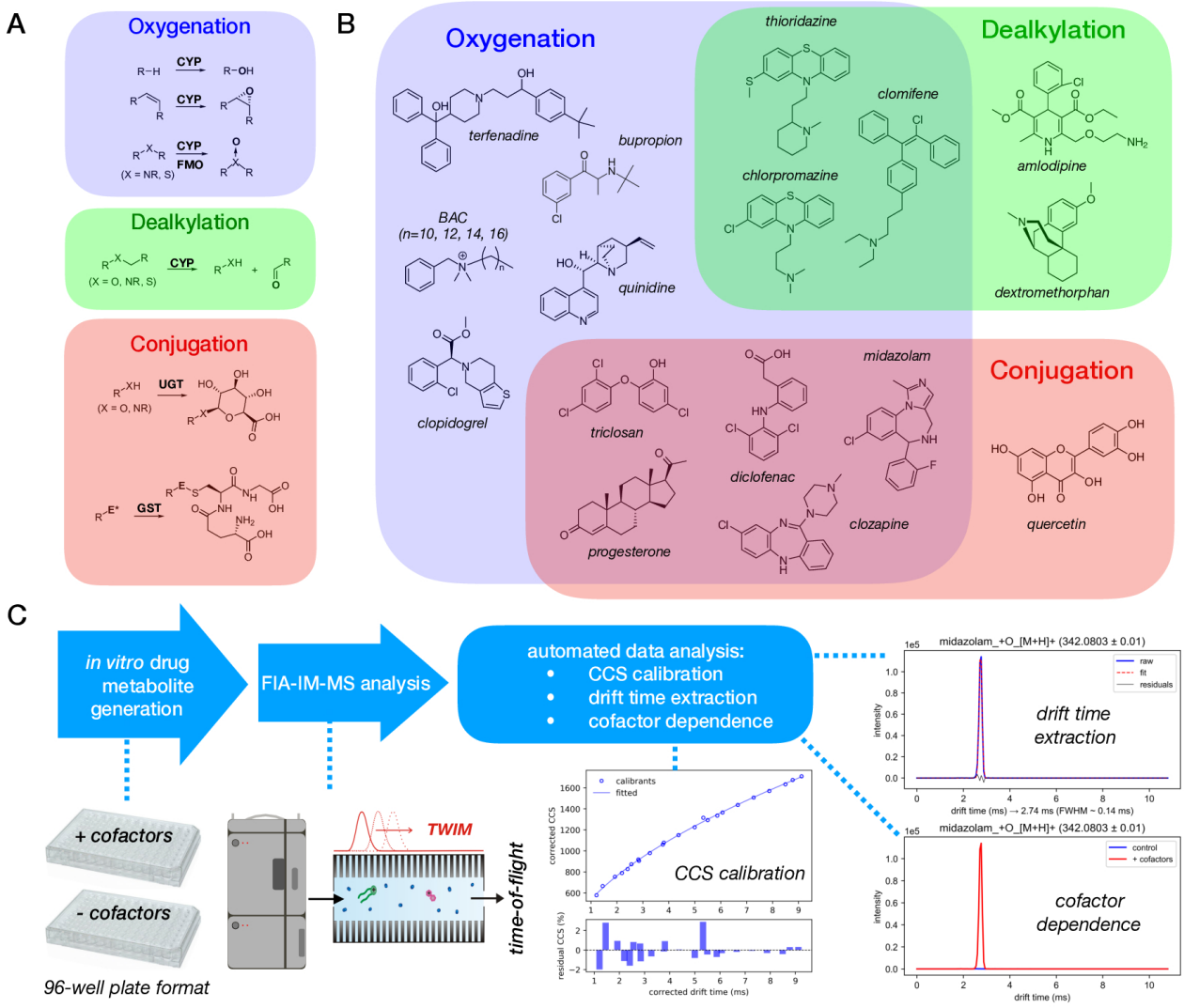
3.4 Conclusion

We have developed an integrated strategy for in vitro biosynthesis of drug metabolites and subsequent structural analysis by IM-MS, which can be scaled up for high-throughput workflows. Using this strategy, we generated and analyzed metabolites of a diverse panel of drugs and found that characteristic changes in CCS associated with drug metabolism are dependent upon both the type and position of the chemical modification and the structural characteristics of the parent compound. The same chemical modification can have drastically different effects on the gas-phase conformations (or CCS) of different parent compounds, likewise, different chemical modifications on the same parent compound can have distinct effects. We propose that such relationship between CCS, metabolic modifications, and parent compound structures could be lever-aged for the prediction of IM behavior of unknown metabolites using data-driven approaches, *e.g.*, machine learning. We anticipate the experimental platform described herein will facilitate the characterization of drug metabolites in a large scale, which would provide a large drug metabolite CCS database that can be used to build such

predictive model, which could significantly increase the efficiency and throughput of early-stage drug metabolite identification and structural elucidation during drug development.

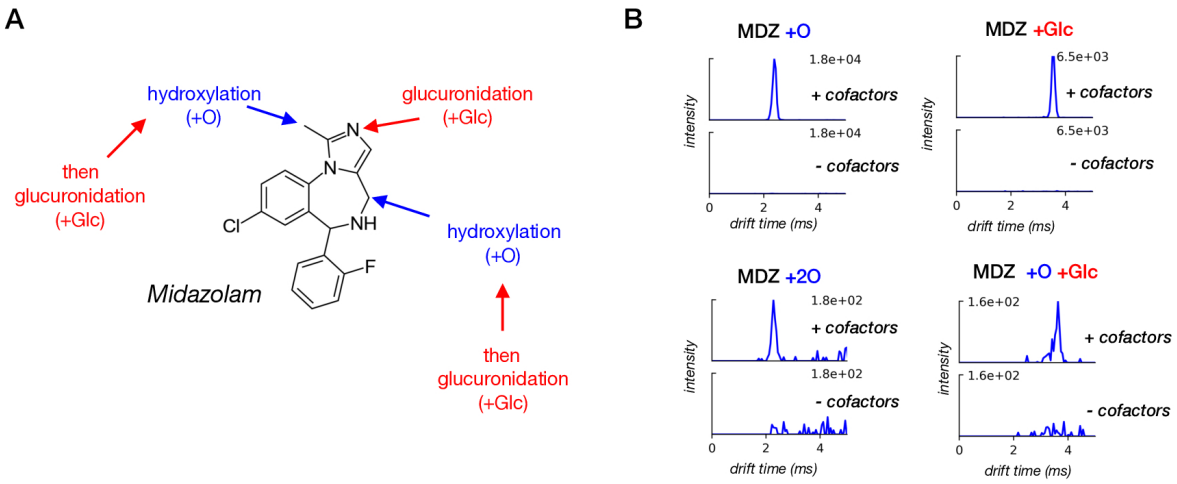
3.5 Figures

3.5.1 Drug Panel and Metabolite Generation/IM-MS Analysis Workflow



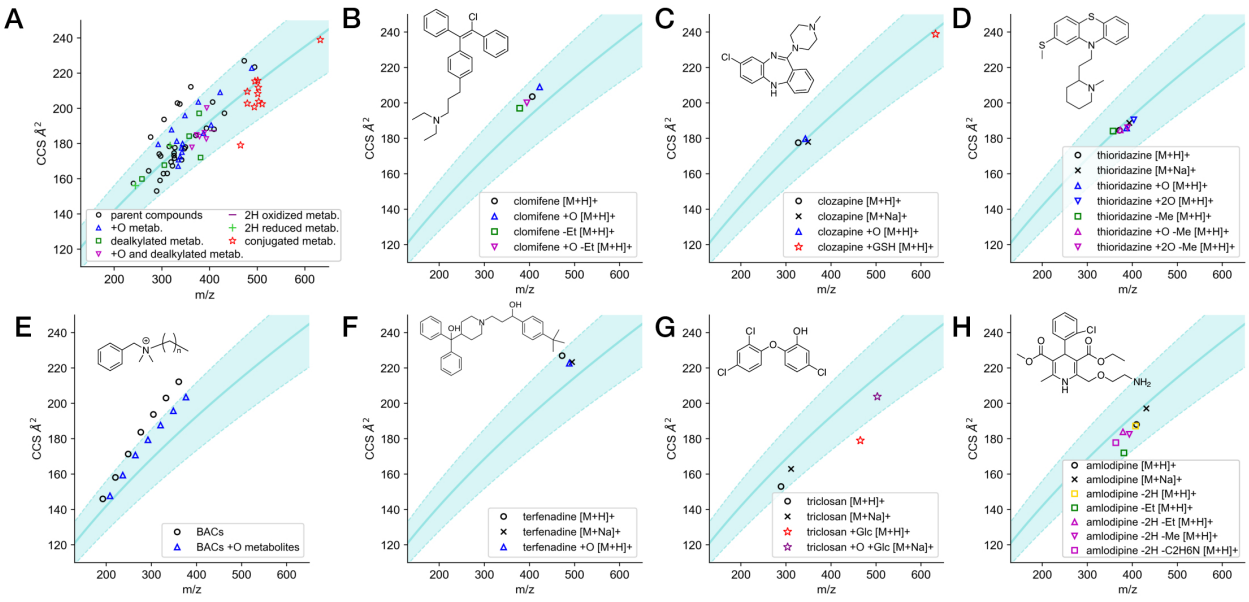
(**A**) Representative reactions of Phase-I and Phase-II drug metabolism; E* designates electrophile. (**B**) Selected drug compounds for this study that can undergo Phase-I and/or Phase-II transformations; (**C**) Overall workflow of this work, from in vitro-metabolite generation, FIA-IM-MS analysis, to automated data processing.

3.5.2 *Initial Characterization of Midazolam Metabolites*



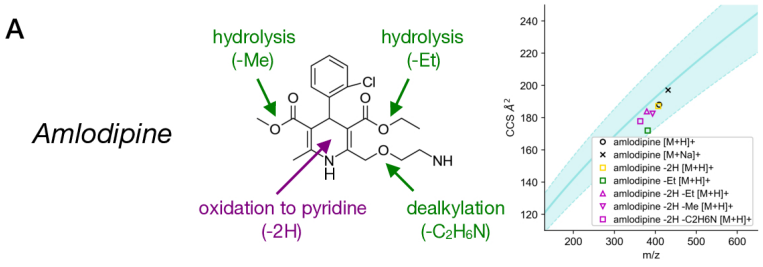
(**A**) Expected metabolism for midazolam<ref 3, 4> (MDZ). (**B**) Arrival time distributions (ATDs) for observed primary and secondary metabolites of MDZ showing cofactor-dependent formation.

3.5.3 *CCS vs. m/z Plots of Drug Panel and Observed Metabolites*

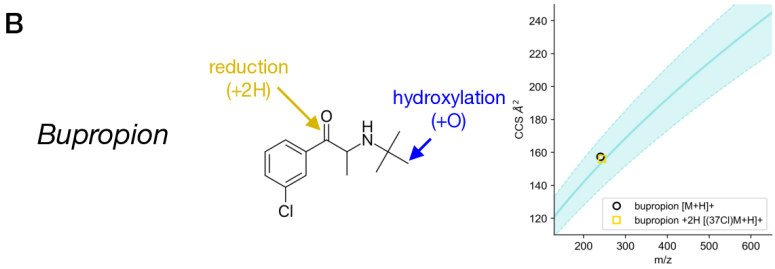


(**A**) IM-MS conformational space plot showing the 78 CCS values of 19 parent drugs and their 37 metabolites. (**B-H**) CCS values of the parent and metabolites of selected individual drugs in IM-MS plot. All data points represent the average of three measurements. The curve and shade in the background of the plot represents the power fit and the ± 10% range of the 1440 drug and drug-like CCS values we reported previously.[47]

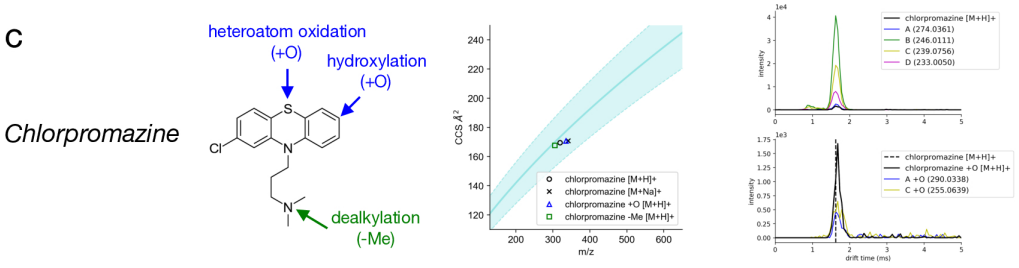
3.5.4 *Expected Metabolites from Literature, Observed Metabolites, and Fragmentation Data for Drugs*



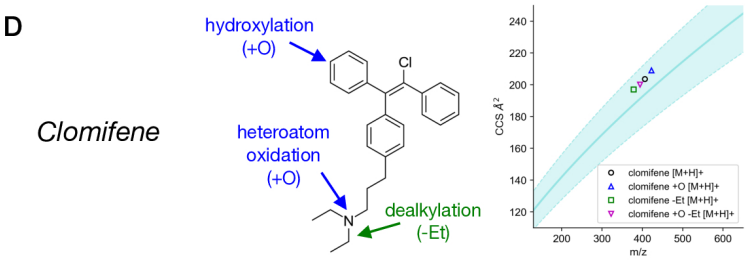
(**A**) Expected metabolism for amlodipine[115] (left) and observed metabolites (right).



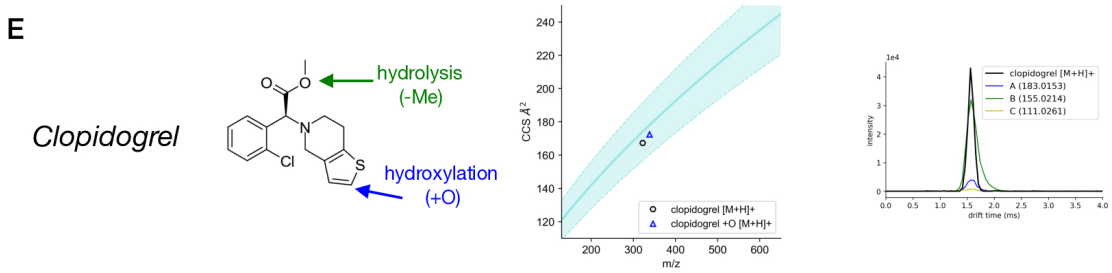
(**B**) Expected metabolism for bupropion[116] (left) and observed metabolites (right).



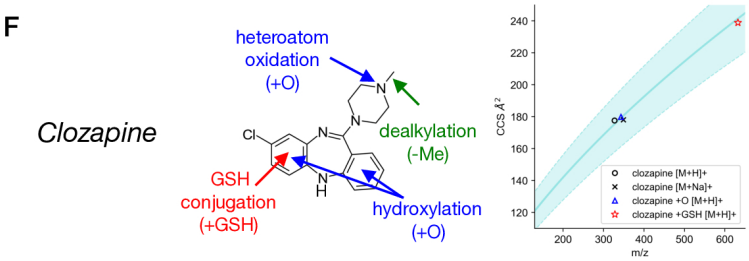
(**C**) Expected metabolism for chlorpromazine[117] (left), observed metabolites (center), and drift time aligned fragmentation data (right).



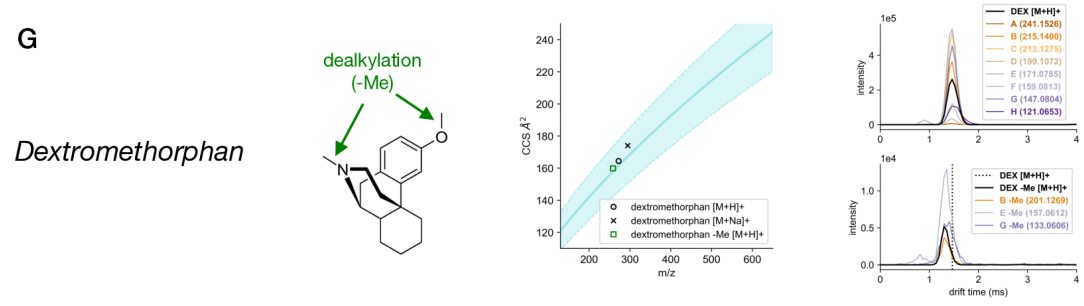
(**D**) Expected metabolism for clomifene[118] (left) and observed metabolites (right).



(**E**) Expected metabolism for clopidogrel[119] (left), observed metabolites (center), and drift time aligned fragmentation data (right).

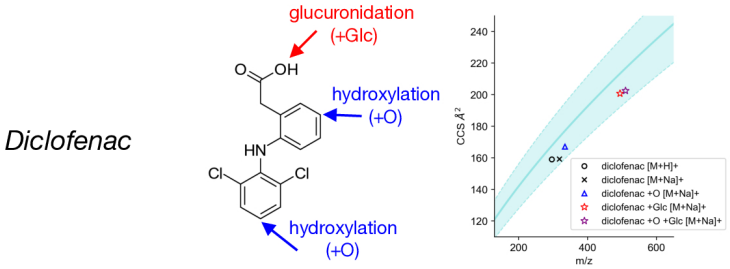


(**F**) Expected metabolism for clozapine[120] (left) and observed metabolites (right).



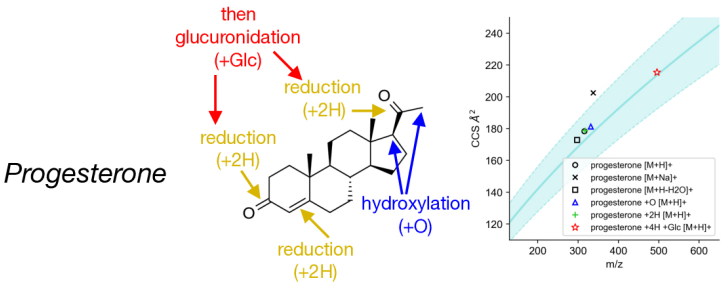
(**G**) Expected metabolism for dextromethorphan[121] (left), observed metabolites (center), and drift time aligned fragmentation data (right).

H



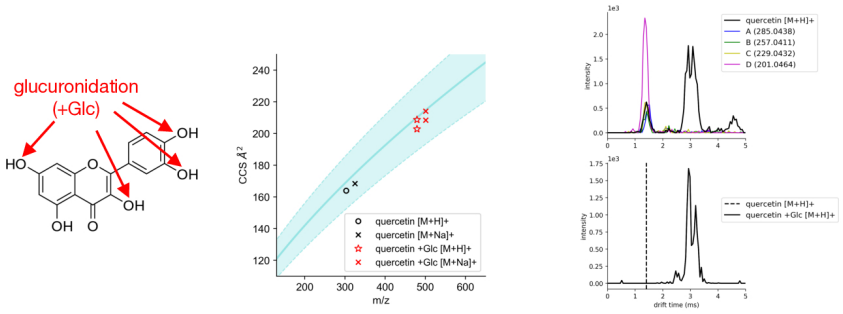
(**H**) Expected metabolism for diclofenac[122, 123] (left) and observed metabolites (right).

I



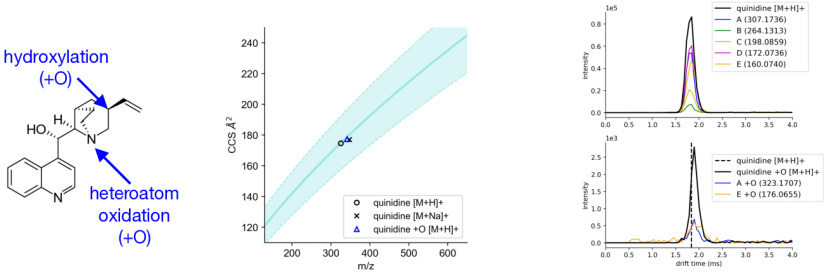
(**I**) Expected metabolism for progesterone[124] (left) and observed metabolites (right).

J

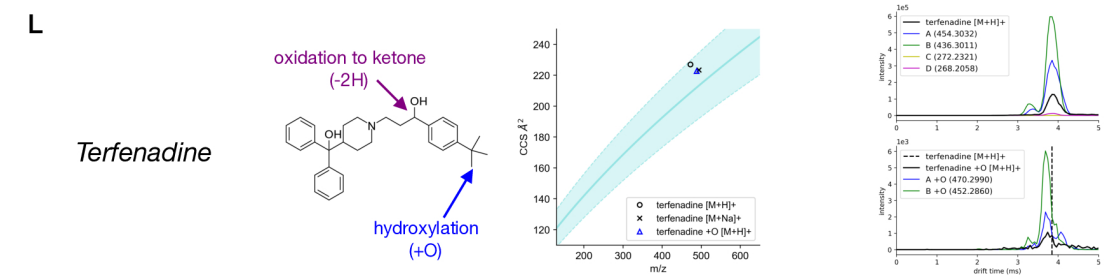


(**J**) Expected metabolism for quercetin[107] (left), observed metabolites (center), and drift time aligned fragmentation data (right).

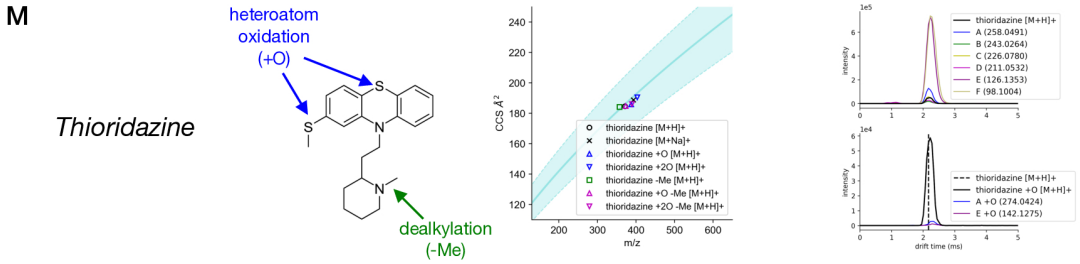
K



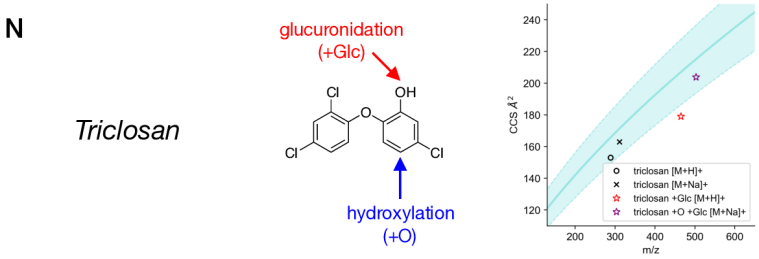
(**K**) Expected metabolism for quinidine[125] (left), observed metabolites (center), and drift time aligned fragmentation data (right).



(**L**) Expected metabolism for terfenadine[126] (left), observed metabolites (center), and drift time aligned fragmentation data (right).

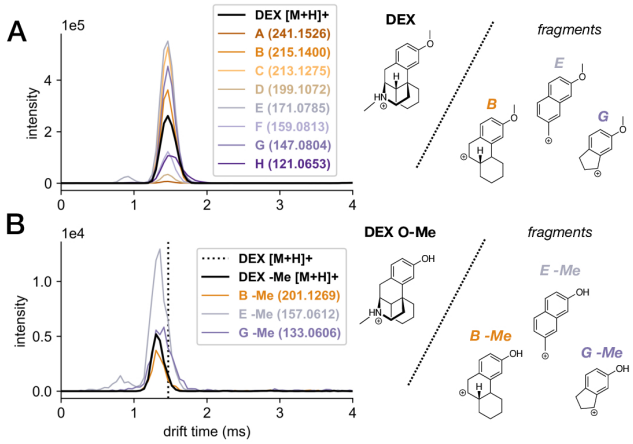


(**M**) Expected metabolism for thioridazine[127] (left), observed metabolites (center), and drift time aligned fragmentation data (right).



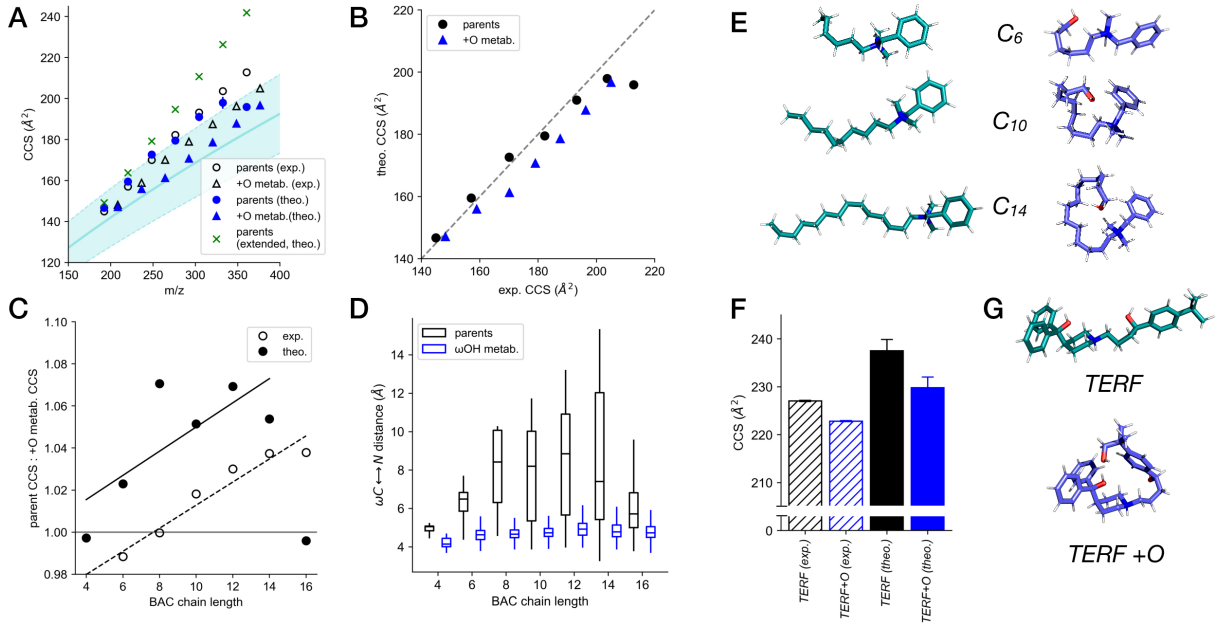
(**N**) Expected metabolism for triclosan[128] (left) and observed metabolites (right).

3.5.5 *MSMS Confirmation of Dextromethorphan O-demethylated Metabolite Assignment*



Extracted ATDs for the fragments of (**A**) the parent dextromethorphan (DEX) and (**B**) its demethylated metabolite (DEX O-Me), confirming that the metabolite is the O-demethylated product, instead of the N-demethylated product. Parent drift time is shown as a dotted line in **B**.

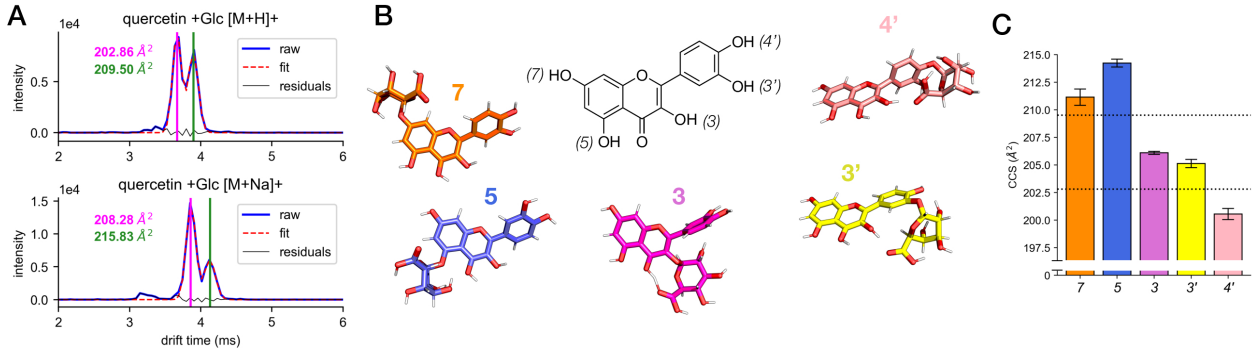
3.5.6 *Computational Modeling of BACs, Terfenadine, and +O Metabolites*



Comparison of experimental measurements and computational modeling of benzalkonium chlorides (BACs) and terfenadine with their hydroxylated metabolites. (**A**) Comparison of trends of experimental CCS values of the parent BACs and their metabolites with the trends of the theoretical CCS values in the CCS-m/z plot. (**B**) Correlation between experimental and theoretical CCS values for parent BACs and their metabolites. (**C**) Comparison of the ratios of parent/metabolite CCS values obtained

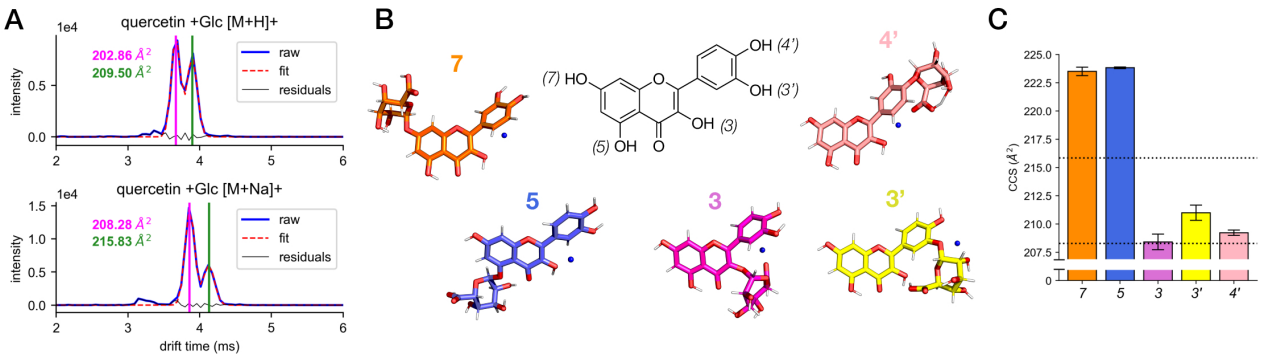
experimentally and computationally. (**D**) Distribution of ω-C-to-N distances of all sampled conformations of BACs and their metabolites. (**E**) Representative gas-phase conformations of BACs and their metabolites with alkyl chain lengths of C6, C10, and C14 from computational sampling. (**F**) Comparison of experimental and computational CCS values for terfenadine (TERF) and its hydroxylated metabolite. Error bars indicate standard deviation. (**G**) Representative gas-phase conformations of terfenadine and its metabolite from computational sampling.

3.5.7 ATDs and Theoretical CCS of Protonated Quercetin Glucuronide Positional Isomers



Bimodal distributions of glucuronides of quercetin. (**A**) ATDs of protonated and sodiated adducts of quercetin glucuronides. (**B**) Gas-phase conformations of regioisomers of quercetin glucuronides. (**C**) Computational CCS values (bar graph) of different regioisomers of quercetin glucuronides in comparison with experimental values (dotted line).

3.5.8 ATDs and Theoretical CCS of Sodiated Quercetin Glucuronide Positional Isomers



(**A**) Bimodal arrival time distributions (ATDs) for protonated (top) and sodiated (bottom) adducts of quercetin glucuronide. (**B**) Minimum energy structures for sodiated isomers of quercetin glucuronide. (**C**) Theoretical CCS of sodiated isomers of quercetin glucuronide, compared against experimental values (dotted lines).

3.6 Tables

3.6.1 Experimental CCS Values for Drugs and Observed Metabolites

compound	metabolite	adduct	m/z	CCS (Å^2)	CCS RSD (%)	Compaction factor (C)
C4 BAC		[M]+	192.1838	144.96	0.52	
C4 BAC	ω-OH	[M]+	208.1795	148.19	0.22	1.032
C6 BAC		[M]+	220.2167	157.07	0.51	
C6 BAC	ω-OH	[M]+	236.2128	158.92	0.26	1.036
C8 BAC		[M]+	248.2505	170.09	0.55	
C8 BAC	ω-OH	[M]+	264.2128	170.14	0.31	1.042
C10 BAC		[M]+	276.2869	182.24	0.60	
C10 BAC	ω-OH	[M]+	292.2815	178.99	0.23	1.057
C12 BAC		[M]+	304.3187	193.19	0.21	
C12 BAC	ω-OH	[M]+	320.3119	187.55	0.20	1.066
C14 BAC		[M]+	332.3427	203.67	0.27	
C14 BAC	ω-OH	[M]+	348.3444	196.33	0.24	1.070
C16 BAC		[M]+	360.3843	212.76	0.16	
C16 BAC	ω-OH	[M]+	376.3800	205.00	0.53	1.068
amlodipine		[M+H]+	409.1586	188.14	0.53	
amlodipine		[M+Na]+	431.1457	197.21	0.24	
amlodipine	-2H	[M+H]+	407.1475	187.24	0.04	1.002
amlodipine	-2H, -Me	[M+H]+	393.1390	182.61	0.11	1.003
amlodipine	-2H, -Et	[M+H]+	379.1376	183.95	0.10	0.972
amlodipine	-2H, -C2H6N	[M+H]+	363.1180	177.84	0.08	0.977
amlodipine	-Et	[M+H]+	381.0955	172.06	0.30	1.043
amlodipine	-Et	[M+Na]+	403.1226	207.14	0.20	0.910
bupropion		[M+H]+	240.1217	157.37	0.24	
bupropion	+2H	[(37Cl)M+H]+	244.1346	156.07	0.31	1.020
chlorpromazine		[M+H]+	319.1169	169.58	0.32	
chlorpromazine		[M+Na]+	341.1305	170.83	0.42	
chlorpromazine	-Me	[M+H]+	305.0947	167.66	0.47	0.982
chlorpromazine	+O	[M+H]+	335.1066	170.74	0.55	1.026
clomifene		[M+H]+	406.2054	203.68	0.60	
clomifene	+O	[M+H]+	422.1923	209.10	0.11	0.999
clomifene	-Et	[M+H]+	378.1691	197.10	0.11	0.985
clomifene	+O, -Et	[M+H]+	394.1640	200.20	0.12	0.997
clopidogrel		[M+H]+	322.0720	167.32	0.27	
clopidogrel	+O	[M+H]+	338.0700	172.49	0.26	1.002
clozapine		[M+H]+	327.1458	177.67	0.27	
clozapine		[M+Na]+	349.1254	178.03	0.17	
clozapine	+O	[M+H]+	343.1388	179.97	0.23	1.019

clozapine	+GSH	[M+H]+	632.2207	238.91	0.11	1.154
dextromethorphan		[M+H]+	272.2093	164.54	0.38	
dextromethorphan		[M+Na]+	294.2091	174.13	0.91	
dextromethorphan	-Me	[M+H]+	258.1915	159.93	0.43	0.993
diclofenac		[M+H]+	296.0769	159.05	0.45	
diclofenac		[M+Na]+	318.0157	159.32	1.19	
diclofenac	+O	[M+Na]+	334.0125	167.11	1.07	0.985
diclofenac	+Glc	[M+Na]+	494.0488	200.79	0.50	1.064
diclofenac	+O, +Glc	[M+Na]+	510.0204	202.54	0.38	1.078
midazolam		[M+H]+	326.0963	171.68	0.32	
midazolam	+O	[M+H]+	342.0881	175.13	0.26	1.012
midazolam	+Glc	[M]+	502.1203	211.85	1.45	1.081
progesterone		[M+H]+	315.2368	178.43	0.26	
progesterone		[M+Na]+	337.2199	202.53	0.16	
progesterone		[M+H-H2O]+	297.2274	172.96	0.58	
progesterone	+O	[M+H]+	331.2334	181.35	0.15	1.017
progesterone	+2H	[M+H]+	317.2529	178.82	0.17	1.002
progesterone	+4H, +Glc	[M+H]+	495.2497	215.41	0.22	1.119
quercetin		[M+H]+	303.0568	162.96	0.44	
quercetin		[M+Na]+	325.1876	173.17	1.16	
quercetin (peak 1)	+Glc	[M+H]+	479.0892	204.10	0.20	1.084
quercetin (peak 2)	+Glc	[M+H]+	479.0892	209.87	0.02	1.054
quercetin (peak 1)	+Glc	[M+Na]+	501.0968	207.30	0.21	1.114
quercetin (peak 2)	+Glc	[M+Na]+	501.0968	216.62	0.43	1.067
quinidine		[M+H]+	325.2060	174.69	0.29	
quinidine		[M+Na]+	347.1915	177.11	0.30	
quinidine	+O	[M+H]+	341.1923	177.48	0.31	1.016
terfenadine		[M+H]+	472.3376	227.04	0.13	
terfenadine		[M+Na]+	494.3174	223.48	0.07	
terfenadine	+O	[M+H]+	488.3268	222.82	0.12	1.042
thioridazine		[M+H]+	371.1743	184.60	0.09	
thioridazine		[M+Na]+	393.1823	188.74	0.13	
thioridazine	+O	[M+H]+	387.1683	185.91	0.36	1.021
thioridazine	-Me	[M+H]+	357.1578	184.20	0.23	0.977
thioridazine	+O, -Me	[M+H]+	373.1677	184.83	0.16	1.002
thioridazine	+2O	[M+H]+	403.1605	190.50	0.19	1.024
thioridazine	+2O, -Me	[M+H]+	389.1668	186.11	0.35	1.024
triclosan		[M+H]+	288.9481	153.08	0.30	
triclosan		[M+Na]+	310.9798	163.04	0.28	
triclosan	+Glc	[M+H]+	464.9557	179.03	0.28	1.174
triclosan	+O, +Glc	[M+Na]+	502.9555	203.88	0.10	1.102

Chapter 4 High-Throughput Measurement and Prediction of CCS Using Machine Learning for Drugs and Drug Metabolites

4.1 Introduction

A critical component of the drug development process is to understand how drugs are metabolized, since metabolism can be an important determinant of drug clearance, and metabolites may elicit unexpected toxicity or other off-target effects.[20, 21] The conventional approach to drug metabolite determination has typically involved a combination of liquid chromatography (LC), coupled to UV-Vis spectroscopy and/or mass spectrometry (MS), and nuclear magnetic resonance spectroscopy (NMR).[22-24] LC-MS and LC-UV benefit from their low sample requirements and fast analysis time, but identification of unknowns can be limited when relying upon UV spectra or MS fragmentation data alone. In contrast, NMR allows for definitive assignment of chemical structures, but it requires large amounts of materials and is relatively low throughput.

Ion mobility spectrometry (IMS) is an analytical technique that rapidly separate ions based on differences in their size and shape in the gas-phase, which is orthogonal to polarity-based LC separations, and partially orthogonal to mass.[29, 31, 32] In time-dispersive IM separations, ions are driven through a neutral buffer gas under the influence of an electric field. Ions are differentially slowed as they interact with the buffer gas molecules, and as a result they traverse the mobility cell in different amounts of time (*i.e.* drift time). An ion's drift time can be converted into collision cross section (CCS), a unique physical property reflecting its gas-phase size and shape, using appropriate experimental measurements and/or calibration. Excellent reproducibility has been demonstrated for CCS measured across different instrumentation and labs,[33-36] making it a robust parameter for compound identification. CCS also provides useful

information related to shape, conformation, and polarity. When coupled with MS (IM-MS), an additional dimension of separation is achieved without adversely affecting analytical throughput.

The use of IM-MS for the determination of drugs and their metabolites has gained significant traction in recent years,[129] but inadequate coverage of existing reference CCS databases remains a significant limitation to the application of CCS in identifying unknown metabolites. Large CCS databases covering drug and drug-like compounds have been presented in the literature,[47, 49, 70, 71] but due to the vastness and complexity of small molecule chemical space, many unknowns may not be represented. This issue is even more pronounced for drug metabolites, for which no such large-scale CCS databases exist. This problem can be addressed by leveraging structural trends in existing CCS databases to predict CCS for unknowns that are not in experimental databases, and this approach has been demonstrated by multiple groups, including ours.[34, 41, 68-72, 130] An important consideration in this approach, however, is the dependence of CCS prediction performance on the quality and coverage of chemical space of the data used to train the model.[130] Therefore, a drug metabolite-specific CCS database is needed for accurate prediction of CCS for drug metabolites. Another limitation in current applications of ML-based CCS prediction is that the 2D features used in previous work (*e.g.* molecular quantum numbers, MQNs) do not adequately capture more complex IM characteristics arising from the presence of different protomers, conformers, or positional isomers that are common among drugs and drug metabolites, as described in Chapter 1.[129]

Here, we describe the generation of a high-quality drug and drug metabolite CCS database through the use of high-throughput *in-vitro* drug metabolite generation and rapid IM-MS analysis with automated data processing. We then used this database for training drug- and

drug metabolite-specific CCS prediction models using ML with novel 3D molecular descriptors, which allows the prediction of CCS values for protomers, conformers, and positional isomers.

4.2 Results and Discussion

4.2.1 High-Throughput Measurement of Drug and Drug Metabolite CCS

To obtain a large collection of drug metabolites, we first carried out high-through drug metabolism reactions in 384-well plates using human liver microsomes and the S9 fraction on 2000 drug and drug-like compounds in the MicroSource Discovery Systems' Spectrum Collection, which consists of 50% approved drugs, 30% natural products, and 20% bioactive compounds (Figure 4.5.1 A). Transformations catalyzed by the Phase-I enzymes, cytochrome P450s (CYPs), and Phase-II enzymes, glutathione S-transferases (GSTs) and UDP-glucuronosyltransferase (UGTs), were probed. Reactions were carried out with or without a cocktail of particular enzyme cofactors, such as NADPH (cofactor of CYPs), glutathione (GSH, co-substrate of GSTs), UDP-glucuronic acid (UDPGA, co-substrate of UGTs), and alamethicin (to enable access of substrates to UGTs), with the cofactor-free reactions serving as negative controls. After the reactions and sample processing, we carried out rapid IM-MS analysis using a 30 mm reverse-phase column, which resulted in just under 2 min per run (Figure 4.5.1 A). Measuring the roughly 2000 compound collection in triplicate with or without enzyme cofactors for drug metabolism reactions resulted in > 8900 samples analyzed. This large and complex set of raw data were analyzed using a stepwise approach with a high degree of automation (Figure 4.5.1 B), including extraction of arrival time distributions (ATDs), Gaussian fitting, CCS calibration, and calculation of CCS of observed ATD peaks. The CCS values of parent compounds were obtained by extracting the ATDs of the exact masses of various adducts. For metabolites, we first generated a theoretical list of potential metabolites using Biotransformer[131]

and then extracted the ATDs of these potential metabolites (see Experimental section for detail). Only ATD peaks that met the criteria of intensity > 1000 counts and peak width between 0.06 and 1.77 ms were retained. This approach ultimately led to the assembly of a large CCS database specific to drugs and drug metabolites. Figure 4.5.2 A summarizes the composition of the drug and metabolite CCS database. The database contained 6245 measured CCS values from 3286 different compounds, of which 1333 were from parent drugs (3675 CCS values) and 1953 were from metabolites (2570 CCS values). The measured CCS values corresponded to a number of ionization states commonly observed in positive mode ESI including $[M+H]^+$ (1936), $[M+Na]^+$ (1656), $[M+K]^+$ (1235), $[M+H-H2O]^+$ (1299), and $[M]^+$ (119).

To validate the identity of the potential metabolites, we first carried out a thorough search of DrugBank[132] for reported metabolites of known drugs in our collection and matched our observed metabolites with those previously reported. For those without reported metabolites, we matched the experimental MS/MS spectra obtained from post-IM fragmentation against an in-silico generated MS/MS spectra of potential metabolite structures using MetFrag,[133] and ruled out low-scoring metabolite annotations. The validation process is discussed in greater detail in the Experimental section. After this process, 4408 of the measured CCS values were retained with an annotation, corresponding to all parent drugs with a CCS value (1333 compounds; 3675 CCS values) and 29.3% of metabolites (572 compounds; 2570 CCS values). 2D molecular descriptors, *i.e.*, molecular quantum numbers (MQNs, 42 features),[84, 130] were generated for all annotated species as described previously. Furthermore, 3D molecular descriptors, including principal moments of inertia (PMI) and radial mass distributions (RMD) (8 features, see Experimental section), were generated to better capture the relationship between conformation and CCS during machine learning as discussed below. Briefly, we attempted to generate 3D

structures for all annotated $[M+H]^+$, $[M+Na]^+$, and $[M+K]^+$ species at a low level of theory (MMFF94 and PM7), resulting in a total of 9813 modeled structures (4074, 3172, and 2567 for each ionized species, respectively). 3D molecular descriptors were generated from the 3D structures using in-house developed Python scripts as described in the Experimental section. In total, 7652 and 2161 3D structures with complete 3D molecular descriptors were generated for parent drugs and metabolites, respectively.

4.2.2 Characteristics of the Drug and Metabolite CCS Database

Principal components analysis (PCA) and partial least-squares regression analysis (PLS-RA) were used to probe the chemical characteristics (as captured by 2D or 3D molecular descriptors) that contribute to variance in the drug and metabolite CCS database. As described in the previous chapter, we had previously characterized a comprehensive collection of compounds from a diverse set of chemical classes using MQNs as features[130], so we first computed a PCA using this comprehensive database (CCSbase) to serve as the chemical space background. We then project the new drug and metabolite CCS database (dmCCS) to this PCA to examine the chemical space that dmCCS distributes within the context of a larger chemical space. Figure 4.5.2 B and C show the PCA projections of compounds from dmCCS (color) overlaid over compounds from CCSbase (grey). In Figure 4.5.2 B, it can be seen that CCS values of the compounds from dmCCS generally increase along the direction of PC2, indicating that the strongest sources of variance in CCSbase do not correspond with sources of variance in dmCCS that relate to CCS. Only 23.9% and 17.7% of the total variance is captured by the first and second principal components, respectively, and a total of 20 components were required to capture 95% of the variance in the CCSbase dataset - indicative of the high degree of diversity in the comprehensive CCS collection. Figure 4.5.2 C shows where the parent compounds and

metabolites from dmCCS group fall within the chemical space defined by CCSbase, which indicates that the dmCCS occupies a broad region corresponding roughly to “small molecules”. The metabolites occupy a subspace within the chemical space occupied by the parent compounds. Figure 4.5.2 D shows where the parent compounds and metabolites from dmCCS (color) map into the IM-MS conformational space (*i.e.* CCS *vs. m/z*), compared to the compounds from CCSbase (grey), with individual power fits for parent compounds and metabolites (dashed lines). Generally, the compounds from dmCCS occupy the low-*m/z* region of this space and span a wide range of CCS values, which comports with expectations based on the structural diversity of compounds in dmCCS. Interestingly, the metabolites seem to occupy a slightly narrower CCS envelope with mostly similar average CCS values to those of the parent compounds. Even in the context of the diverse chemical space of CCSbase, the compounds from dmCCS represent considerable structural diversity.

Separate PCAs were computed on dmCCS using the 2D and 3D molecular descriptors to determine how each set of descriptors reflected the chemical space covered by this database. Figure 4.5.2 E and F show the PCA projections from the 2D and 3D features, respectively. For both feature sets, the first principal component correlates well with variation in CCS, indicating that among these compounds the primary sources of variance are related to CCS. The first and second principal components of the PCA computed on the 2D feature set captured 19.3% and 13.4% of the overall variance, respectively, compared to 58.8% and 18.0% for the 3D feature set, indicating that a high degree of variance orthogonal to CCS in the 2D feature set is not present in the 3D feature set. Indeed, the PCA computed on the 2D features required 24 components to capture 95% of the variance in the dataset, in contrast to only 5 components needed for the 3D feature set. Figure 4.5.2 G and H show the 2D and 3D feature loadings, respectively, for the first

principal component in each PCA, both of which correlate well with CCS. The strongest contributors to separation along the first principal component (Figure 4.5.2 H) for the 2D features were counts of atoms (*hac*: heavy atoms, *ao*: acyclic oxygens, *c*: carbons), bonds (*adb*: acyclic double bonds, *atb*: acyclic triple bonds), and topological features (*asv*: acyclic monovalent nodes, *ctv*: cyclic trivalent nodes). All of the 3D features contributed similarly to the separation along the first principal component (Figure 4.5.2 I), and interestingly, the second and third PMI had slightly larger contributions than the first. Together, these results demonstrate that both the 2D and 3D feature sets capture important characteristics of this set of compounds that relate to CCS, but the 3D features contain somewhat less extraneous information.

We next examined the degree to which the 2D and 3D feature sets (MQN and MD3D, respectively) offered complimentary information by computing a PCA on dmCCS using a combination of both feature sets (COMB). Figure 4.5.2 G shows the resulting PCA projections, which overall appear quite similar to those from the 2D feature set alone. The first principal component captured 23.6% of the total variance, while there were 27 total components required to capture 95% of the variance in the dataset. The top features contributing to separation along the first principal component (Figure 4.5.2 J) consist of a combination of those identified from the 2D and 3D feature sets, but it is unclear whether these features are contributing orthogonal information to one another from this analysis alone.

We also performed a set of analogous analyses using PLS-RA computed on the 2D, 3D and combined feature sets with CCS as the target variable (Figure 4.5.3). The results from these analyses largely mirrored those discussed above, which is expected given the alignment of CCS with the first principal component in all three PCAs.

4.2.3 Training Drug- and Drug Metabolite-Specific Prediction Models

The 2D and 3D feature sets were used to train individual ML models for CCS prediction on dmCCS. Despite the different sizes (42 features *vs.* 8) and characteristics of the 2D and 3D feature sets, the MQN and MD3D predictive models achieved very similar performance in CCS prediction by multiple metrics, with robust performance between training and test set data (Figure 4.5.4). We next sought to test the degree to which the two feature sets provided orthogonal information by training a ML model on the combined 2D and 3D feature sets (COMB). Indeed, the COMB model achieved significantly improved predictive performance relative to models trained on either individual feature set (Figure 4.5.4), indicating that both feature sets contain complementary information that is relevant to CCS prediction for this set of compounds. However, there was a significant decrease in performance between the training and test set data for the COMB model, indicating model overfitting likely attributable to the presence of redundant and/or superfluous features.

To address potential overfitting in the COMB model, a set of feature ranking and successive feature removal trials (Figure 4.5.5), including PLS-RA, gradient boosting regression (GBR), and a permutation feature importance function in Scikit-Learn (PER), were run in order to select a minimal feature set combining the most influential features from the 2D and 3D feature sets while avoiding overfitting by removing extraneous features (see Experimental section for detail). Molecular descriptors retained by at least two of the feature removal methods were kept as the minimal feature set (MIN), which consisted of only 11 descriptors from both the 2D and 3D feature sets: *hac*, *c*, *asv*, *adb*, *ctv*, *hbam*, *hbd*, *pmi1*, *pmi2*, *pmi3*, *rmd02*. A new ML model was trained using this MIN feature set. Although there was still an appreciable degree of correlation between the features in this minimal set (Figure 4.5.6), the MIN model achieved

improved performance relative to the models trained on the 2D or 3D features alone (Figure 4.5.4), and importantly, this performance was better maintained between the training and test set data.

4.2.4 Comparison of CCS Prediction Model to Theory-Based Conventional Methods

Computational modeling to produce 3D structures at a low theory level is the primary bottleneck in training and application of ML models for CCS prediction based on 3D molecular descriptors. Given that production of such structures is also a bottleneck for some of the faster theory-based CCS prediction methods (*e.g.* projection approximation, PA, and exact hard-sphere scattering, EHS),[58] we sought to compare the accuracy of CCS values predicted using both approaches for compounds in dmCCS. Figure 4.5.7 shows measured and calculated CCS for compounds from dmCCS, colored according to calculation method. The ML values were predicted using the model trained on the minimal 2D/3D combined feature set described in the previous section. Both of the theory-driven methods (PA and EHS) display significant systematic errors, consistently underestimating CCS by *ca.* 50 $Å^2$. This systematic difference is likely attributable to the parameterization of the PA/EHS calculation, which was originally meant for calculation of CCS in He and not modified for use with N_2 as the drift gas (such modification would greatly increase the complexity of the PA/EHS calculation method). When systematic errors were corrected by linear regression, the residuals of the fit for PA or EHS-generated values were significantly larger than the ML-predicted values. Taken together, it is clear that ML-based CCS prediction produces higher quality CCS values with this dataset than comparable theory-based methods, likely attributable to the nuanced structural trends that such ML models can capture when provided with appropriate training data, in addition to their lack of reliance upon prior parameterization.

4.2.5 Application of CCS Prediction to Compounds with Multimodal ATDs

Multimodal ATDs can arise from a number of circumstances, such as constitutional isomers (*e.g.* positional isomers of metabolites or protomers formed in the ESI process) and conformers.15, 16, 31 Such phenomena often arise in the analysis of drug metabolites; however, previous CCS prediction models based on 2D molecular descriptors generally do not allow the differentiation of such isomers or conformers. Inclusion of 3D features in CCS prediction could in theory capture such multimodal differences for given 3D structures, so we sought to evaluate some known examples of multimodal distributions using our model trained with the minimum 2D and 3D feature set.

Benzalkonium chlorides (BACs) are a class of compounds that consist of an alkyl chain of varying lengths with a permanently charged ammonium group at one end, and have been previously characterized by IM-MS.[134] An interesting trend was observed in the CCS values for the BACs and their +O metabolites: the trendline of the metabolite CCS is lower than the corresponding parent values, and the magnitude of this difference increases with chain length. BACs are known to be hydroxylated at or near the terminal carbon of the alkyl chain,[135] so we hypothesized that the introduction of –OH at the end of the molecules induced structural compaction due to intramolecular ion-dipole interaction (Figure 4.5.8 B). Since such an interaction is unlikely to be predicted from the 2D structure alone, we compared experimental CCS values[134] to those predicted using the prediction models described above (MQN, MD3D, COMB, MIN). For prediction models that include 3D features (MD3D, COMB, MIN), the same ensembles of 3D structures produced previously[134] were used as inputs since their trajectory method CCS displayed good agreement with experimental values. Figure 4.5.8 A shows the CCS *vs. m/z* trends of the BACs and their ω-OH metabolites for the experimentally measured values

and values predicted using 2D (MQN) or 3D (MD3D, COMB, MIN) molecular descriptors, in addition to compaction factors (see Experimental section) for each of the metabolite CCS values. The MQN prediction model produced CCS values that were systematically lower than experimental values for all BACs and +O metabolites. This model also failed to capture the compaction in the metabolites, as indicated by the compaction factors for the predicted values which were smaller than those corresponding to the experimental values and decreased with chain-length. The MD3D prediction model performed the best overall, with predicted values having slight systematic under prediction that increased with chain length. More importantly, the compaction factors of the predicted values were close to the experimental values for the shorter chain BACs, indicating that for these species the model was able to capture the gas-phase compaction in the metabolites that was observed experimentally. The COMB and MIN models showed significant systematic errors, both in terms of absolute values and metabolite compaction factors, indicating that neither of these models adequately captured the structural characteristics of BACs and their +O metabolites. Taken together, it is perhaps unsurprising that the best overall CCS prediction performance was achieved by the MD3D model, given that the structural compaction of the metabolites represents a fairly distinct structural difference. None of the models performed particularly well on this set of compounds in an absolute sense, however, that is to be somewhat expected considering the fact that the BACs are quite lipid-like in their structural characteristics and are thus not likely to be well-represented in the drug and drug metabolite database used for model training.

Terfenadine and its +O metabolite display a similar structural relationship to the BACs and their ω-OH metabolites: compaction in the metabolite relative to the parent, likely attributable to the introduction of an intramolecular polar-polar interaction (Figure 4.5.8 D).[134]

Figure 4.5.8 C compares the experimentally measured CCS values of terfenadine and its +O metabolites to values predicted using the different models discussed above. The CCS values predicted using the MQN model are lower than the experimental values, and the +O metabolite has a larger CCS than the parent, indicating that this feature set does not adequately capture the structural characteristics of these compounds. The MD3D model produced essentially the same results as the MQN model. Interestingly, the COMB model produced predictions that were closer to experimental values and although the metabolite did not have significantly lower CCS than the parent, but the compaction factor of 1.023 did indicate a compaction of the metabolite. The MIN model produced the closest predictions, and importantly, reproduced the decreased CCS of the metabolite relative to the parent, with a compaction factor of 1.030 (compared to 1.042 in the experimental values). Together, these results demonstrate that while the MQNs and MD3D feature sets were not individually capable of capturing the compaction of the +O metabolite of terfenadine, their combination resulted in reproduction of the behavior. The observation that this compaction behavior was captured for terfenadine metabolite but not for the BAC metabolites likely reflects how well they are represented in the chemical space covered by the data used to train these predictive models (drugs and drug metabolites).

Cefpodoxime proxetil is a β-lactam antibiotic that has previously been shown to form two protomers in ESI with distinct CCS values (Figure 4.5.8 F).[47] We predicted CCS values using the set of prediction models described above to determine how well they capture structural differences imparted by ionization at different sites of a molecule (Figure 4.5.8 E). The MQN model was unable to distinguish between the different protomers (likely due to them being constitutional isomers), and the predicted CCS values were significantly smaller than the experimental values. The MD3D model produced predictions that differed between the two

protomers, but their rank-order was reversed relative to the experimental values. The COMB model performed worst out of the four, producing significantly under predicted values and failing to distinguish between the protomers. In contrast, the MIN model performed the best, producing CCS predictions fairly close to the experimental values, and more importantly, preserving the rank-order of the protomer CCS values as observed for the experimental values. These data demonstrate that the optimized feature set composed of 2D and 3D descriptors was necessary and sufficient to capture the structural differences between protomers of cefpodoxime proxetil.

Quercetin is a flavonoid compound with multiple –OH groups available for glucuronidation (Figure 4.5.8 H).[136] The glucuronide metabolite of quercetin has previously been observed to produce a bimodal CCS distribution, likely attributable to the presence of more than one constitutional isomer arising from glucuronidation at different positions.[134] Figure 4.5.8 G compares CCS values predicted using the collection of models described above, relative to the experimentally measured values (dotted lines) for five positional isomers of quercetin glucuronide. The MQN model failed to capture any CCS differences between the positional isomers due to the inability of these features to capture differences between constitutional isomers, but the model did accurately predict the lower experimental CCS value. The MD3D model was able to distinguish between the different positional isomers and the assignment of isomers to the two experimental values was largely in agreement with previous results.[134] The COMB model performed the worst out of the group, likely attributable to the apparent overfitting of the training data that occurred with this feature set. The MIN model performed similarly to the MD3D model, with the exception of the rank-order of the 7-, 5-, and 3- isomers being inverted. Together, these results demonstrate that descriptors containing 3D information (*i.e.* MD3D and

MIN) are necessary to capture CCS differences between this set of constitutional isomers. The overall accuracy of these predictions also indicates good coverage of the chemical space for quercetin glucuronides in the data used to train the predictive models.

Fluoroquinolone antibiotics are a class of compounds that have been extensively studied in the IM field due to their formation of protomers, which are often distinguishable by CCS.[129] Eight fluoroquinolones (ciprofloxacin, enoxacin, enrofloxacin, levofloxacin, lomefloxacin, norfloxacin, orbifloxacin, and pefloxacin) were present among the compounds analyzed in this study, each of which displayed bimodal CCS distributions in their protonated form. Figure 4.5.9 shows experimentally measured CCS values and CCS values predicted using the collection of models discussed above for all of these fluoroquinolones. Four potential protomers – two on the piperazine, one on the central fused ring, and one on the carbonyl (Figure 4.5.9 A) – were modeled (see Experimental section for 3D modeling procedure) for each of the fluoroquinolones. Due to the conformational rigidity of these compounds, relative to *e.g.* cefpodoxime proxetil, the structural differences between the protomers are somewhat small, and the magnitude of the differences were not well captured by any of the models. It is possible that polarity plays a larger role in driving the CCS differences between fluoroquinolone protomers, and the feature sets used in this work fail to reflect such differences. Despite these limitations, the MIN-based model consistently produced two or more different CCS values, and there does appear to be a general distinction between protomers on the main fused rings and those on the piperazine, with the former having generally larger predicted CCS than the latter.

4.3 Experimental

4.3.1 High-Throughput In Vitro Drug metabolite Generation

Drug metabolites were generated in vitro using pooled subcellular fractions (S9 and microsomes) derived from human liver following a protocol from our previous work,[134] adapted to a high-throughput 384-well plate format with all sample preparation performed using automated sample handling systems. Briefly, HLM/S9 stock (5 mM GSH, 5 mM MgCl, 0.01 mg/mL alamethicin, 0.2 mg protein/mL pooled HLM, 0.2 mg protein/mL S9, 100 mM potassium phosphate buffer at pH 7.4) was prepared and allowed to stand on ice for 15 min (alamethicin pre-treatment to enhance UGT activity). 90 µL of the HLM/S9 stock was dispensed into each well of 14 384-well plates, then 0.5 µL of each drug stock (50 mM in DMSO) from the MicroSource Spectrum Discovery Collection (7 plates) were dispensed into the plates in duplicate. 10 µL of a cofactor-containing activation mixture (10 mM NADPH, 50 mM UDPGA, 100 mM potassium phosphate buffer at pH 7.4) or potassium phosphate buffer were then added to the duplicate plates, initiating the drug metabolism reactions for plates containing activation mixture. All plates were incubated at room temperature for 90 min before being quenched with 100 µL ice-cold acetonitrile (with 10 µM lysophosphatidylethanolamine 13:0 as an internal standard). After quenching, all plates were stored at 4 °C for at least 15 min to promote precipitation of proteins. Each plate was centrifuged at 3500G for 15 min at 4 °C to sediment the precipitated proteins, then 150 µL of the supernatant was transferred to fresh plates. All plates were stored at -80 °C until IM-MS analysis.

4.3.2 High-Throughput Ion Mobility-Mass Spectrometry

Samples (5 µL) were injected and separated using a Waters Acquity FTN UPLC coupled to a reverse-phase column (Phenomenex Kinetex, 2.6 µm, polar C_{18}, 100 Å, 30 x 21 mm),

eluting with a gradient of water with 0.1% formic acid (A) and methanol with 0.1% formic acid (B) at 0.5 mL/min: 0.00-0.20 min, 100% A; 0.20-0.30 min, 100→25% A; 0.30-0.75 min, 25→0% A; 0.75-1.05 min, 0% A; 1.05-1.10 min, 0→100% A. The total analysis time for each sample, factoring in acquisition and autosampler operations, was just under 2 min. For each injection, the first 0.20 minutes of eluent was diverted to waste in order to avoid buildup of salt on the ESI source, and after that, the flow was automatically diverted back to the instrument via an electronically controlled switching valve. TWIM-MS analysis was performed on a Waters Synapt G2-Si mass spectrometer (Waters Corp., Milford, MA) equipped with an ESI source and using nitrogen as the drift gas. ESI conditions were as follows: capillary, +2.3 kV; sampling cone, 40 V; source temperature, 130 °C; desolvation temperature, 350 °C; cone gas, 90 L/h; and desolvation gas, 600 L/h.

Mass calibration was performed using sodium formate for the range of m/z 50−1200. IM separations were performed at a traveling wave velocity of 650 m/s and a height of 24.9 V. For post-IM fragmentation analyses, collision energy was added to the transfer region using a ramp from 30 to 50 eV. Data was acquired from 0.20 to 1.15 min with a 1s scan time over *m/z* 50−1200, which resulted in approximately 57 scans across the acquired elution region (individual peaks typically spanned ~0.05 min for roughly 3 scans per peak). The 384-well plates were analyzed on three separate occasions over two months.

4.3.3 TWIM CCS Calibration

A series of singly charged polyalanines (n = 2−14) and a mixture of drug-like compounds were used for calibration of TWIM drift times into CCS ($^{TWIM}CCS_{N2}$) using their drift tube CCS values in nitrogen ($^{DT}CCS_{N2}$), as described previously.[47, 134] Briefly, arrival time distributions (ATDs) for CCS calibrants were extracted from the raw data (acquired multiple times throughout

acquisition of each plate) using accurate mass with a window of ±0.01 Da, and a CCS calibration curve was constructed from reference CCS values in an automated fashion using a Python script developed in-house.[134] Drift times for each calibrant were obtained as the mean from a least-squares fit of a Gaussian function on the ATD and were corrected for mass-dependent flight time outside the mobility region to give the corrected drift times (t_d'), and reference CCS values were corrected for the ion charge state (Z) and reduced mass with the drift gas to give the corrected CCS (CCS').[109] A calibration curve was generated by fitting these corrected values with the function $CCS' = A(t_d' + t_0)^B$, where A, t_0, and B were the fitted parameters.[33, 110] A calibration curve displaying randomly distributed fit residuals with a maximal absolute error of less than 3% was considered acceptable. CCS calibrant data was acquired 3-5 times over the course of analysis of each plate, and all of this calibrant data was used to construct a combined CCS calibration curve to account for any variation that occurred over the course of data acquisition.

4.3.4 Ion Mobility-Mass Spectrometry Data Processing

The raw IM-MS data was processed in a number of steps to extract, annotate, and validate CCS values for drugs and putative metabolites (Figure 4.5.1 B), and was performed separately for each batch of data acquired on the same day (two plates were analyzed each day). The first set of data processing steps were completely automated using Python scripts developed in-house, and were applied only to the first technical replicates. First, a target list was assembled for the parent drugs based on the known plate contents. For each parent compound, *m/z*-selected arrival time distributions (ATDs) were extracted for common ionized species with a tolerance of 0.05 Da. ATDs were fit with a gaussian function to obtain drift time, and the fitted drift time was used to calculate calibrated CCS. If an ATD was not able to be fit or the fitted peak did not meet empirically determined rough quality cutoffs (intensity > 1000, peak width between 0.06 and

1.77 ms), the corresponding ionized species was not processed any further. Upon successful ATD peak fitting, a drift time-selected chromatogram was also extracted and an attempt was made to fit for retention time. All data and metadata were stored in custom Python data structures for subsequent processing. Putative metabolites were generated using BioTransformer26, with the "allHuman" setting and up to 2 metabolism steps. Putative metabolites were filtered to exclude isobaric metabolites, metabolites with the same neutral mass as the parent compound, and metabolites resulting from the breakdown of secondary metabolites or other non-parent derived compounds (*e.g.* free glucuronic acid, glutathione, acetic acid, ethanol, *etc.*), then their corresponding ATDs were extracted from the raw data and fitted as described above. In total, 1018 of these putative metabolites were assigned a tentative annotation. Successfully fitted ATDs were stored along with metadata (including tentative metabolite annotation) in custom Python data structures for subsequent processing. Plots containing compound/putative metabolite structures, *m/z*, metabolism reaction information, CCS, and ATDs with fits were generated and stored (see Figure 4.5.1 A) for subsequent manual review.

The resulting initial data set (>11k analytes) was next subjected to a manual review process. Each of the generated plots described above were manually inspected for general quality of ATD peak fitting (clean ATD fit without secondary peaks) and cofactor-dependence for oxidative and glucuronide metabolites, then accepted or rejected accordingly. The results of this manual review process were used to curate an analyte *m/z* target list for automated data extraction from the second and third technical replicates. Data extraction from the second and third replicates followed the same automated workflow described above, except that the curated target list was used to search for putative metabolites rather than through in silico metabolite

prediction. All extracted data and metadata from the second and third metabolites were stored in custom Python data structures for subsequent processing.

The final step in data processing was validating compound annotations, which was performed using a semi-automated process. The identities of the parent compounds were known from the plate contents, so further validation was not required. To validate the annotations of the putative metabolites, known metabolites of the parent compounds were manually searched for in the DrugBank database.[132] A list of potential metabolites and associated metadata were compiled from these searches and later matched to metabolites (superseding the original putative metabolite annotation from BioTransformer) on the basis of their neutral mass (within 50 ppm was considered a match), resulting in 69 validated annotations. Finally, all remaining metabolite annotations (949) were subjected to filtering based on post-mobility MS/MS data that were acquired for the first replicate. Drift time-selected MS/MS spectra were extracted and scored against *in silico* fragmentation spectra using MetFrag,[133] and all annotations with a fragmenter score above the empirical cutoff of 100 were accepted. This empirical MetFrag scoring cutoff was determined by a rank test using the known identities of the parent compounds as follows. The drift time-selected MS/MS spectrum for each parent compound was compared to the *in silico* fragmentation spectra of all parent compounds, resulting in ranked identifications with corresponding fragmenter scores. The rank and score of the known identity were recorded for each compound, and the empirical scoring cutoff was determined by looking at the distribution of scores for parent compounds with true identities ranking in the top 500 (Figure 4.5.10). Ultimately, this cutoff represents a rough way of ruling out unlikely annotations given their corresponding MS/MS spectrum, and in total 274 putative metabolite annotations were removed based on this criterion (213 annotations without scores + 61 with scores < 100), resulting in 744

accepted annotations (69 validated from DrugBank + 675 accepted from MS/MS filtering from 572 compounds).

4.3.5 Assembly of a Drug and metabolite CCS Database

A SQLite3 database was used to store experimental data, associated metadata, annotations, 3D structures, and computed molecular descriptors for all of the drugs and metabolites observed in this study. The overall database architecture is summarized in Figure 4.5.11. The database has separate tables for CCS measurement data and metadata (*plate_N*), MS2 spectra (*plate_N_ms2*), compound annotations (*plate_N_id*), 2D molecular descriptors (*plate_N_mqn*), 3D structures (*plate_N_3d*), and 3D molecular descriptors (*plate_N_md3d*). All of the experimental plates (7 in total) have their own set of corresponding tables for consistency with the organization of the experimental source data. The database was constructed in a stepwise, automated fashion using a series of Python build scripts developed in-house. Briefly, the database was first initialized with all of the empty tables, then the measured data were added to the *plate_N* tables according to plate number. Only measurements with CCS values from all 3 technical replicates and RSD < 5% between them were added into the database. Next, compound annotations (names and SMILES structures) were added to the *plate_N_id* tables. Parent drug annotations were already known from the plate contents, but metabolite annotations were assigned via a combination of automated and manual verification (*vide supra*). Drift time-selected MS2 spectra were added to the *plate_N_ms2* tables, with each entry in the measured data tables having a corresponding MS2 spectrum. Next, MQNs were computed using SMILES structures (*vide infra*) from the annotation tables and added to the *plate_N_mqn* tables. 3D structures (in plain text format) were generated (*vide infra*) and added to the *plate_N_3d* tables, and corresponding 3D molecular descriptors were computed for each structure (*vide infra*) and

added to the *plate_N_md3d* tables. The *plate_N*, *plate_N_ms2*, and *plate_N_id* tables are all related by a unique (across all 7 sets of plates) text identifier, *dmim_id*. All of the annotations in the *plate_N_id* tables have an additional unique integer identifier, *ann_id*, relating them to entries in the *plate_N_mqn* and plate_N_3d tables. The *plate_N_3d* tables have an additional unique integer identifier, *str_id*, relating their entries to the *plate_N_md3d* tables.

4.3.6 Generation of 3-Dimensional Structures for Ionized Drugs and Metabolites

3-Dimensional structures were computed from SMILES structures for experimentally observed ionized (protonated and Na^+/K^+ adducts) drug and metabolite species using a series of scripts developed in-house employing a combination of molecular mechanics and semi-empirical methods. Briefly, initial 3D structures were generated by a Monte Carlo conformer search followed by steepest descent energy minimization using the MMFF94 force field in the OpenBabel software package.[137] The initial 3D structures were then further optimized at the PM7 semi-empirical theory level in Gaussian16.[138] Finally, the optimized atom positions, masses, and partial charges were stored along with relevant metadata for the measured species. This process was repeated 3 times for each individual ion species to increase the chances that a minimum energy structure would be sampled in this non-extensive modeling protocol.

Generation of 3D structures for protonated species followed the same protocol, but with the inclusion of additional steps to account for multiple potential sites of protonation within a molecule. First, potential protomers were determined by presence of ionizable groups, and the SMILES structures were modified to reflect each protomer in an automated fashion. 3D structure generation was performed using each of the protomer SMILES structures as described above, but with additional thermodynamic calculations specified in the semi-empirical optimization step. After 3D structures had been produced for all potential protomers, the structures having the

lowest energy and highest partial charge located on the protonation site (if different from lowest energy structure) were selected and stored. The above protocol was repeated 3 times for each species, resulting in 3-6 structures for each protonated species.

The 3D structure generation protocol resulted in the production of 3-6 structures for each ionized species in an attempt to capture multiple energetically similar conformers; however, for most compounds, many or all of the produced structures were virtually the same. To avoid undue influence in predictive model training from such duplications, all structures for a given compound were subjected to RMSD filtering. Briefly, for each compound a mass-weighted RMSD matrix was computed between all predicted structures and only those differing by more than 0.01 Å were retained. The RMSD cutoff of 0.01 Å was determined empirically by computing the distribution of RMSD values for all structures in the database (Figure 4.5.12), in addition to manual inspection of a handful of compound structures. All of the filtered 3D structures were added as a separate table to the drug and metabolite CCS database.

4.3.7 Generation of 2-Dimensional Molecular Descriptors

Molecular quantum numbers (MQNs) were used as 2D molecular descriptors for analysis of the drug and metabolite CCS database. MQNs are graph properties of a 2D molecular structure (*e.g.* a SMILES structure), which include counts of atoms, bonds, and topological features.[84] MQNs were computed from the neutral SMILES structures for all entries in the drug and metabolite CCS database using the RDKit library (https://www.rdkit.org). The computed MQNs were added as a separate table to the drug and metabolite CCS database.

4.3.8 Generation of 3-Dimensional Molecular Descriptors

Principal moments of inertia (PMI) and binned radial mass distributions (RMD) were used as molecular descriptors for 3D molecular structures. PMI are derived from the

eigendecomposition of the inertia tensor of a rigid body computed relative to its center of mass. Physically, this computation produces a set of orthogonal axes within a body, such that the radial distribution of mass about each successive axis is minimized; the magnitude of the PMIs reflects the extent of radial mass distribution about their corresponding axes (Figure 4.5.13). Given a 3D molecular structure defined by *N* atoms having masses (*m*) and positions (*x*, *y*, *z*) with center of mass located at the origin, the body frame inertia tensor (*I*) was computed as follows:

$$I = \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix}$$

where the diagonal elements (I_{xx}, I_{yy}, I_{zz}) were computed as:

$$I_{xx} = \sum_{i=1}^{N} m_i(y_i^2 + z_i^2)$$

$$I_{yy} = \sum_{i=1}^{N} m_i(x_i^2 + z_i^2)$$

$$I_{zz} = \sum_{i=1}^{N} m_i(x_i^2 + y_i^2)$$

and the off-diagonal elements (I_{xy}, I_{yx}, I_{xz}, I_{zx}, I_{yz}, I_{zy}) were computed as:

$$I_{xy} = I_{yx} = \sum_{i=1}^{N} m_i x_i y_i$$

$$I_{xz} = I_{zx} = \sum_{i=1}^{N} m_i x_i z_i$$

$$I_{yz} = I_{zy} = \sum_{i=1}^{N} m_i y_i z_i$$

An eigendecomposition (as implemented in the SciPy39 Python library: *scipy.linalg.eigh*) was then performed on the inertia tensor, yielding the principal moments of inertia (PMI_1, PMI_2, PMI_3):

$$I = Q\Lambda Q^T$$

$$\Lambda = \begin{bmatrix} PMI_1 & 0 & 0 \\ 0 & PMI_2 & 0 \\ 0 & 0 & PMI_3 \end{bmatrix}$$

RMDs reflect the proportions of a structure's mass that lie within specific distances radially from its center of mass. Specifically, RMDs are normalized, mass-weighted histograms of atomic distances relative to the center of mass. The histograms were binned at specific distance intervals (0-2 Å, 2-4 Å, 4-6 Å, 6-8 Å, and >8 Å) in order to reduce the total number of features. The binning intervals were chosen based on the combined distribution of mass-weighted radial distances from all 3D structures in the drug and metabolite CCS database (Figure 4.5.14). The computed 3D molecular descriptors were added as a separate table to the drug and metabolite CCS database.

4.3.9 Multivariate Analysis of Drug and Metabolite CCS Database

PCA and PLS-RA are implemented in Scikit-Learn, a free and open-source machine learning library for Python (*sklearn.decomposition.PCA* and *sklearn.cross_decomposition.PLSRegression*, respectively).[95] PCA and PLS-RA are dimensionality reduction techniques that work by determining successive orthogonal axes within a high-dimensional dataset that contain maximal variance. PLS-RA differs from PCA in that the first axis is chosen such that it corresponds to the direction of maximal variance in an external target variable (in this case CCS), making it a targeted analysis.

4.3.10 Prediction of CCS Using Machine Learning

Prior to model training, the data were processed in a stepwise fashion. First, the data set was randomly (seeded for deterministic results) split into training and test sets in proportions of 80% and 20%, respectively, and the test set was held aside during model training. Rough stratification based on distribution of CCS was used during data set splitting to ensure comparability between the training and test sets. The training data were centered and scaled such that each feature would have a mean of 0 and unit variance in order to avoid undue emphasis of features on the basis of their magnitudes. A support vector regression (SVR) model with radial basis function kernel was used for CCS prediction (*sklearn.svm.SVR*). The model hyperparameters (*C* and *gamma*) were optimized using a grid search with 5-fold cross validation (*sklearn.optimize.GridSearchCV*) on the training data. The model trained using the optimal hyperparameters was then used to compute performance metrics (*vide infra*) from predictions made on the training and test data sets.

4.3.11 CCS Prediction Performance Metrics

A standard set of metrics were used to determine the bulk performance of CCS prediction using ML and other methods as described previously.[130] Briefly, these include R^2, mean and median absolute error (MAE and MDAE, respectively, $Å^2$), mean and median relative error (MRE and MDRE, respectively, %), root mean squared error (RMSE, $Å^2$), and cumulative error distribution at 1, 3, 5 and 10% levels (CE135A, %).

4.3.12 Feature Selection for CCS Prediction

Starting from a complete combined feature set (2D + 3D molecular descriptors, 50 features total), a set of tests were performed (using only the training set data) to determine the minimal feature set necessary to make robust and accurate CCS predictions. First, the relative

importance of all individual features was determined by three methods: PLS-RA, gradient boosting regression (GBR, *sklearn.ensemble.GradientBoostingRegressor*), and a permutation feature importance function built into *Scikit-Learn* (PER, *sklearn.inspection.permutation_importance*). PLS-RA gives an indication of feature importance based on the magnitude of the loadings in the x-dimension (*i.e.* the multidimensional axis that explains the maximal variance in the target variable). GBR is an ensemble method in which successive decision tree models are fitted to the residuals of previous models, and relative feature importance can be inferred from the frequency with which individual features are used for decision tree splits. In the PER method, feature importance is related to the decrease in prediction performance when a feature is randomly shuffled relative to a baseline (unshuffled) performance. Once feature importance had been calculated, sequential feature removal tests were performed using the importance from each method. In the feature removal tests, the least important features were successively removed, and new predictive models were trained and evaluated on the smaller feature sets. This process was repeated until only a single feature (with the highest importance) remained (Figure 4.5.5 A-C). For each method, a reduced feature set was selected as the set of features for which the prediction error (RMSE) increased above 5 $Å^2$ upon their removal. Finally, a minimal feature set was selected as those common among 2 sets of features remaining after feature removal tests using the PLS-RA, GBR, and PER feature importance (Figure 4.5.5 D).

4.3.13 Calculation of PA/EHS CCS

Theoretical CCS values were calculated for all 3D structures in the drug and metabolite CCS database by the projection approximation (PA) and exact hard-sphere scattering (EHS) methods using MobCal, modified to use N_2 as the drift gas.[58, 60]

4.3.14 Calculation of Metabolite Compaction Factors

Gas-phase compaction factors of metabolites relative to parents were computed using the equation:

$$\frac{\mathrm{CCS}_{parent}}{\mathrm{CCS}_{metabolite}} = C\left(\frac{mass_{parent}}{mass_{metabolite}}\right)^{2/3}$$

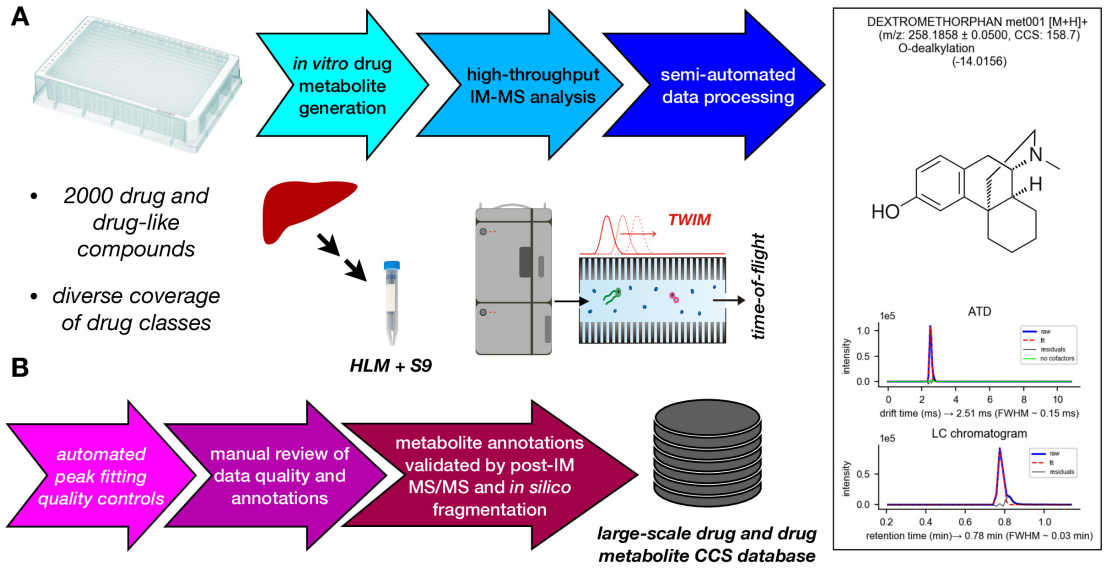
as described previously.[134] Briefly, since the CCS at a given mass is analogous to a gas-phase density, a change in mass (*i.e.* due to metabolic modification) is expected to produce a monotonic change in CCS, and that change is isotropic if the density does not change (C = 1). If C > 1, then the metabolite is denser than expected under isotropic growth while C < 1 indicates the metabolite is less dense.

4.4 Conclusion

We have presented the use of high throughput in vitro drug metabolite generation and rapid IM-MS analysis with automated data processing for the production of a large drug- and drug-metabolite CCS database. We then demonstrated the use of this database to train a CCS prediction model specific to drugs and drug metabolites using a combination of conventional 2D molecular descriptors and novel 3D molecular descriptors derived from low level computational modeling. This approach represents a hybridization of data- and theory-driven CCS prediction, which enables the prediction model to capture complex IM behaviors through inclusion of 3D structural information, such as multimodal CCS distributions resulting from protomers, positional isomers, and conformers.

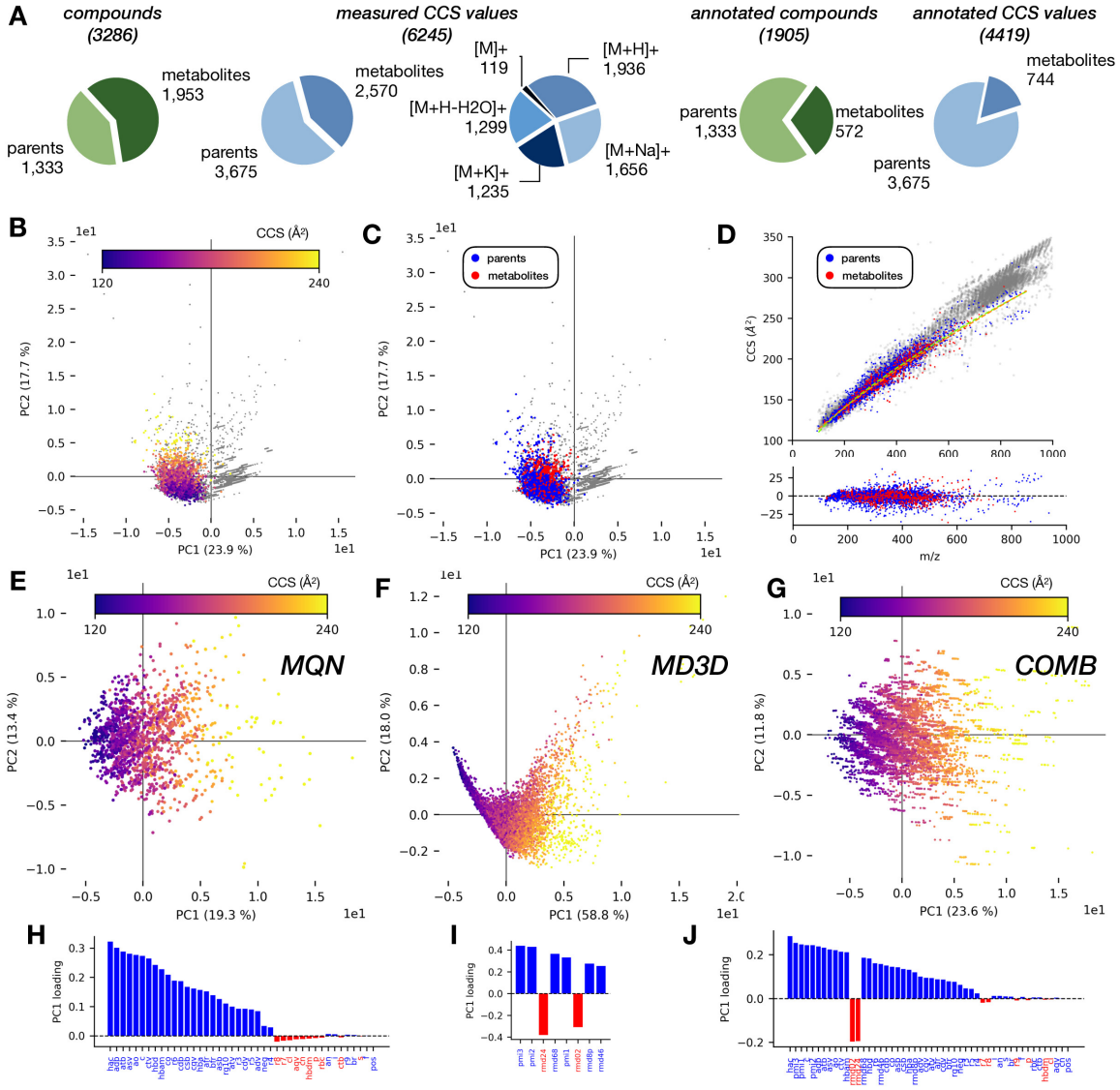
4.5 Figures

4.5.1 *Workflow for High-Throughput In Vitro Drug Metabolite Generation, IM-MS Analysis, and Semi-Automated Data Processing*



(**A**) Workflow for high-throughput *in vitro* drug metabolite generation and IM-MS analysis. Drug metabolites were generated from the MicroSource Spectrum Discovery Collection, containing ~2000 drug and drug-like compounds, in a high-throughput 384-well plate format using subcellular fractions (microsomes and S9) pooled from 200 human livers. Samples were analyzed using a rapid IM-MS protocol, including semi-automated data processing including extraction and fitting of drift times from ATDs, calibration of CCS, prediction of metabolites, and establishment of cofactor dependence for oxidative metabolites. (**B**) The semi-automated data processing included multiple steps of automated and manual quality controls, including automated quality controls on peak fitting, manual review of extracted data quality and metabolite annotations, and validation of metabolite annotations with MS/MS data. The processed data was finally compiled into a SQLite3 database for use in CCS prediction by ML.

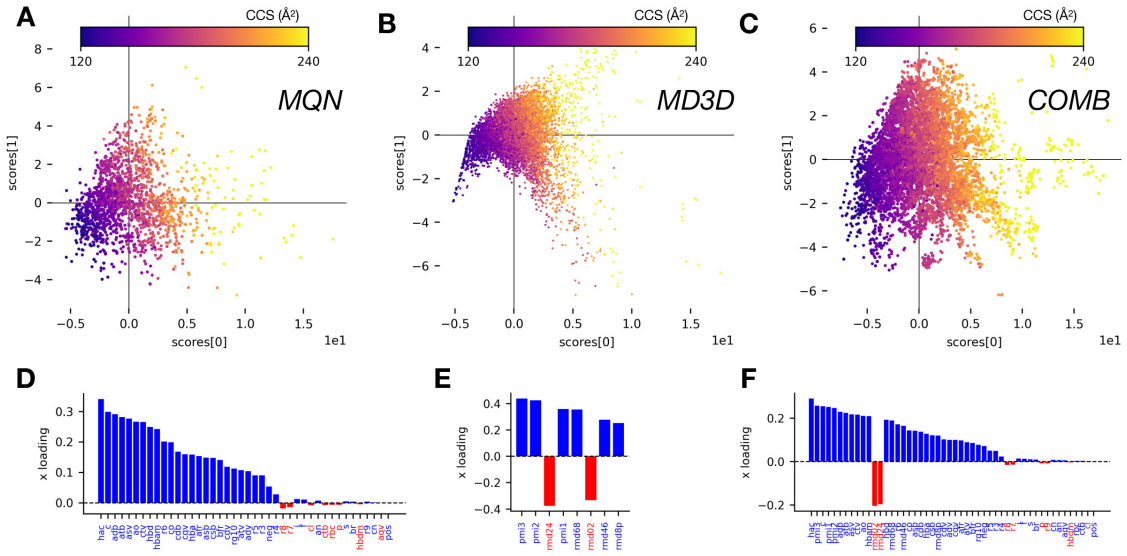
4.5.2 *Characterization of the Drug and Metabolite CCS Database*



(**A**) Composition of the assembled CCS database for drugs and drug metabolites (dmCCS). (**B**) PCA projections of the dmCCS database (color) from a PCA computed using the CCSbase database (grey),[130] colored by CCS. (**C**) PCA projections of parent compounds (blue) and metabolites (red) from the dmCCS database from a PCA computed using the CCSbase database (grey). (**D**) CCS *vs. m/z* of parent compounds (blue) and metabolites (red) from the dmCCS database overlaid on the CCSbase database (grey). Dotted lines represent individual power fits for parent (chartreuse) and metabolite (orange) data, and residual CCS from these fits are included below the main plot. (**E**) PCA projections of dmCCS database computed using MQNs as molecular descriptors, colored by CCS. (**F**) PCA projections of dmCCS database computed using MD3Ds as molecular descriptors, colored by CCS. (**G**) PCA projections of dmCCS database computed using the combination of MQNs and MD3Ds as molecular descriptors, colored by CCS. (**H**) Individual feature loadings for principal component 1 from PCA computed

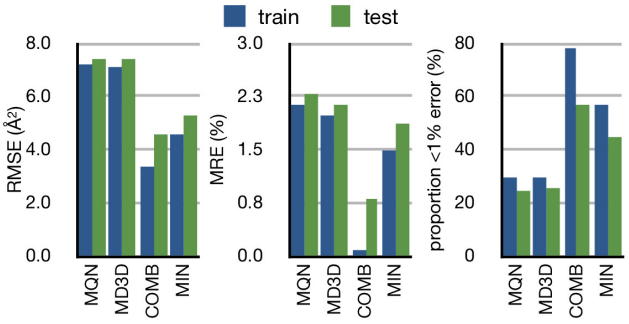
on dmCCS using MQNs as molecular descriptors. (**I**) Individual feature loadings for principal component 1 from PCA computed on dmCCS using MD3Ds as molecular descriptors. (**J**) Individual feature loadings for principal component 1 from PCA computed on dmCCS using the combination of MQNs and MD3Ds as molecular descriptors.

4.5.3 PLS-RA on Drug and Metabolite CCS Database



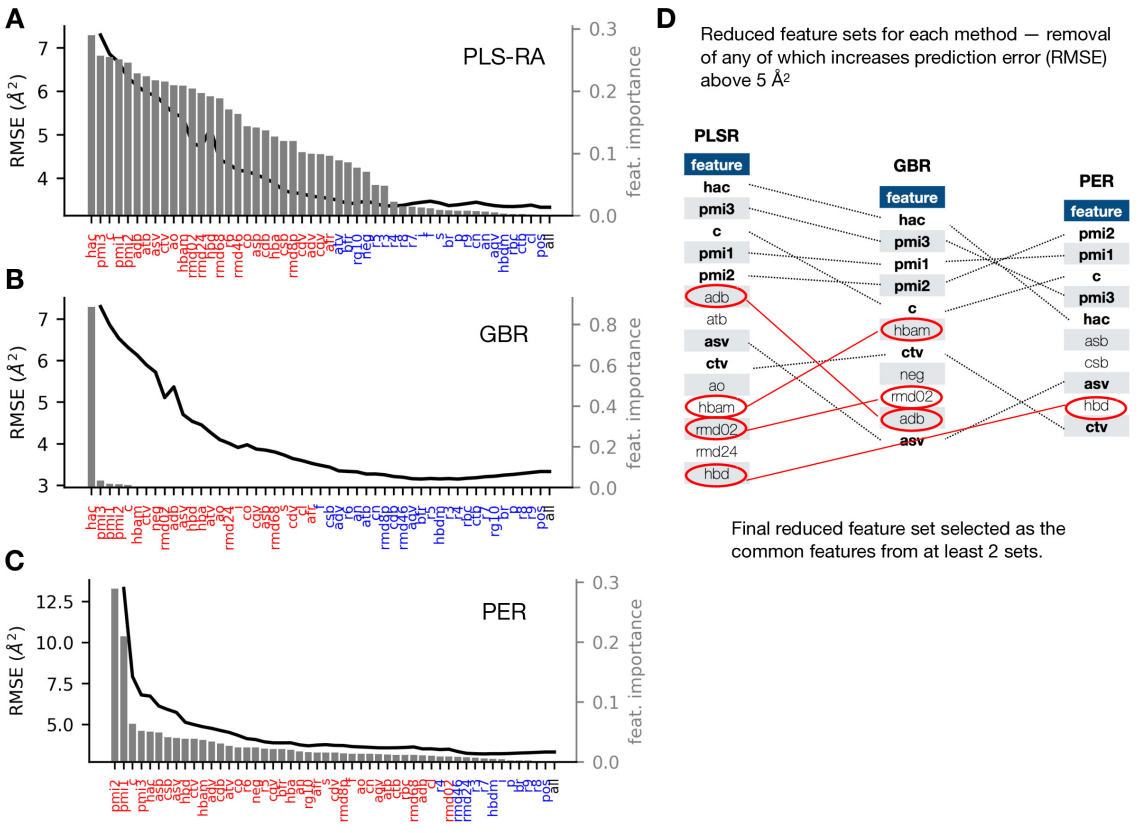
(**A**) PLS-RA projections of dmCCS database computed using MQNs as molecular descriptors and CCS as the target variable, colored by CCS. (**B**) PLS-RA projections of dmCCS database computed using MD3Ds as molecular descriptors and CCS as the target variable, colored by CCS. (**C**) PLS-RA projections of dmCCS database computed using the combination of MQNs and MD3Ds as molecular descriptors and CCS as the target variable, colored by CCS. (**D**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using MQNs as molecular descriptors and CCS as the target variable. (**E**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using MD3Ds as molecular descriptors and CCS as the target variable. (**F**) Individual feature loadings for component 1 from PLS-RA computed on dmCCS using the combination of MQNs and MD3Ds as molecular descriptors and CCS as the target variable.

4.5.4 *CCS Prediction Performance for ML Models Trained on the Drug and Metabolite CCS Database*



CCS prediction performance comparison for ML models trained on dmCCS using MQN, MD3D, a combination of MQN and MD3D (COMB), or a minimal feature set (MIN) as molecular descriptors.

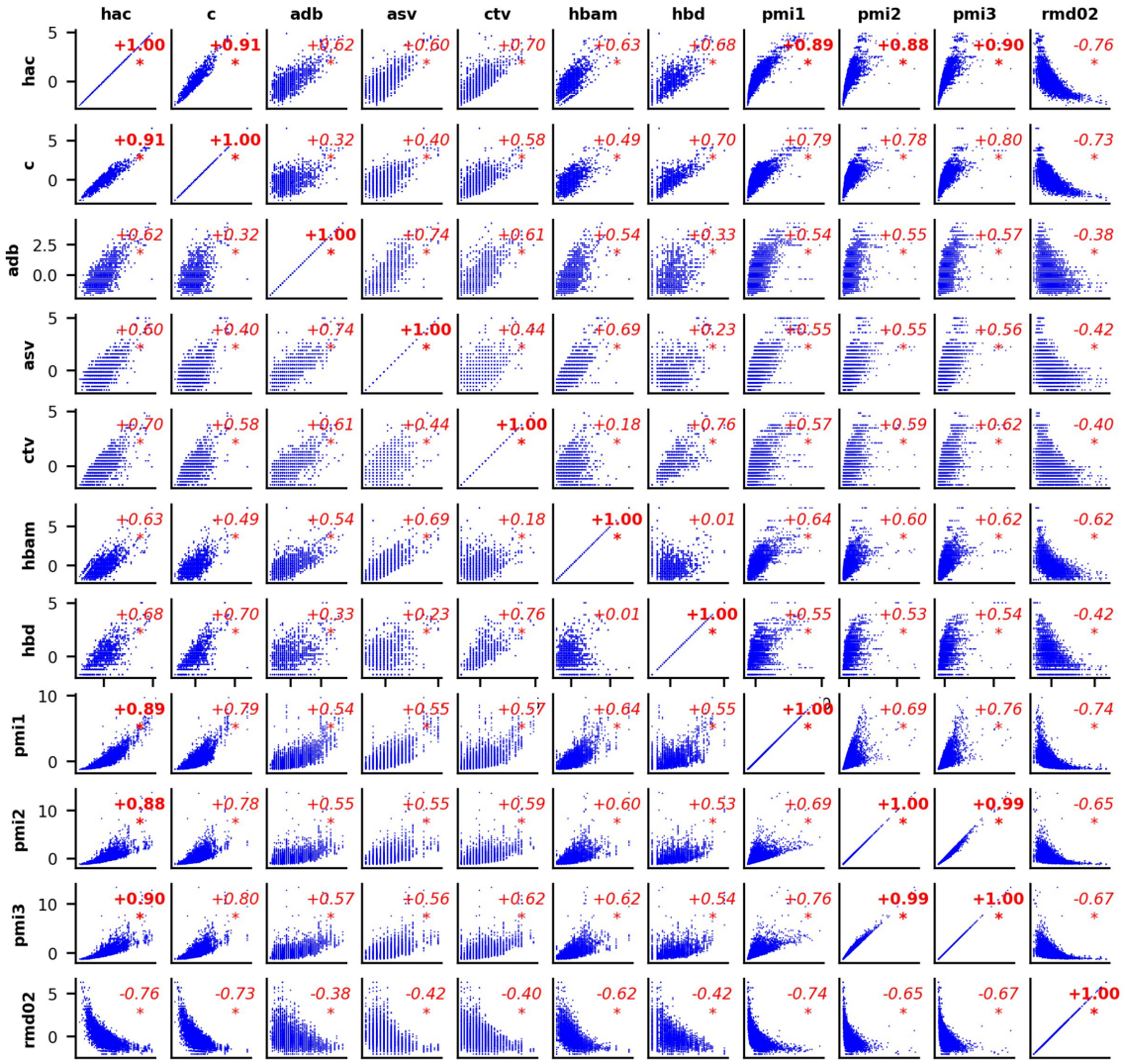
4.5.5 *Feature Selection Trial Results*



(**A**-**C**) Results from feature selection trials. Features were removed in descending order of feature importance (from right to left, grey bars), and resulting predictive model

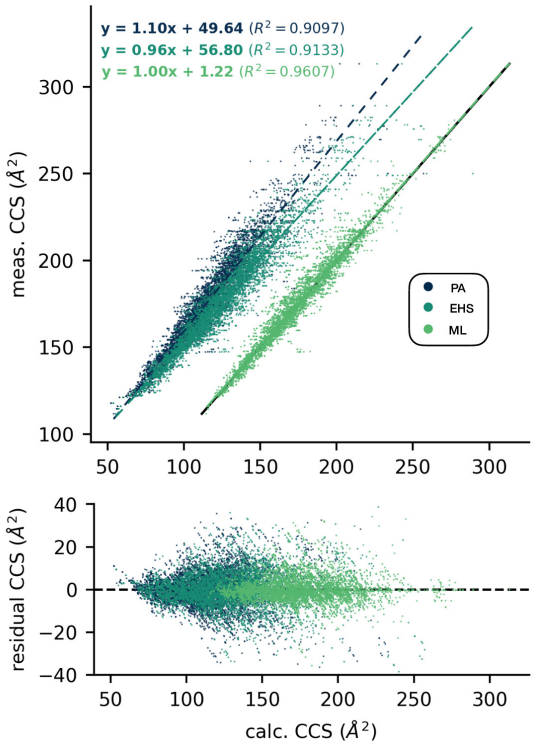
performance was recorded (RMSE, black line). Blue labels indicate the features that could be removed without model performance increasing RMSE more than 5% relative to the baseline (*all*). (**D**) Selected features from individual trials, selected as those for which removal increased error above 5 Å^2. The features selected in at least two of the individual tests were retained as the final minimal feature set.

4.5.6 *Correlation Matrix for Minimal Feature Set*



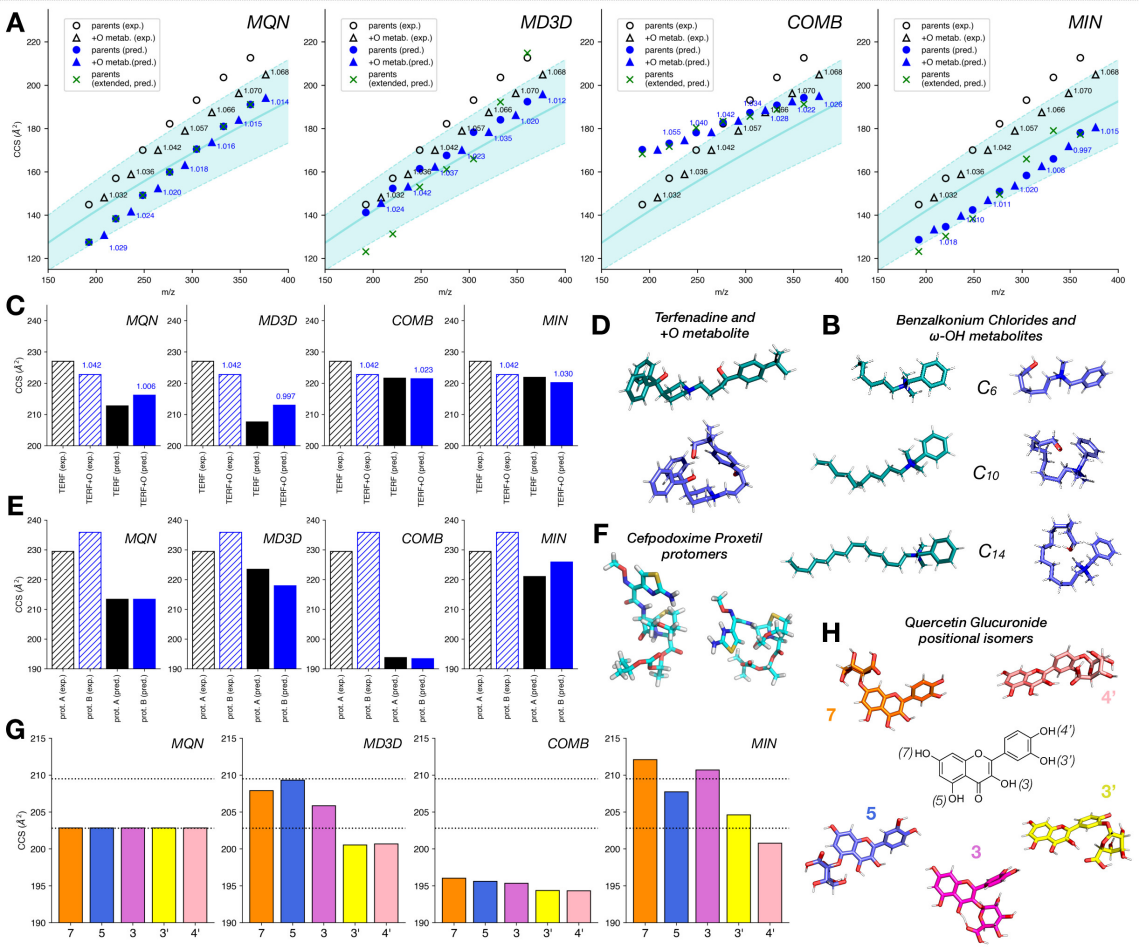
Covariation matrix of minimal feature set from feature selection trials. Red numbers correspond to spearman rank test correlation coefficients (coefficients with magnitude > 0.85 are in bold). Asterisks denote a p-value < 0.01 for the correlation.

4.5.7 *Comparison of CCS Prediction Performance for ML-Based CCS Prediction and PA/EHS CCS Calculation Methods*



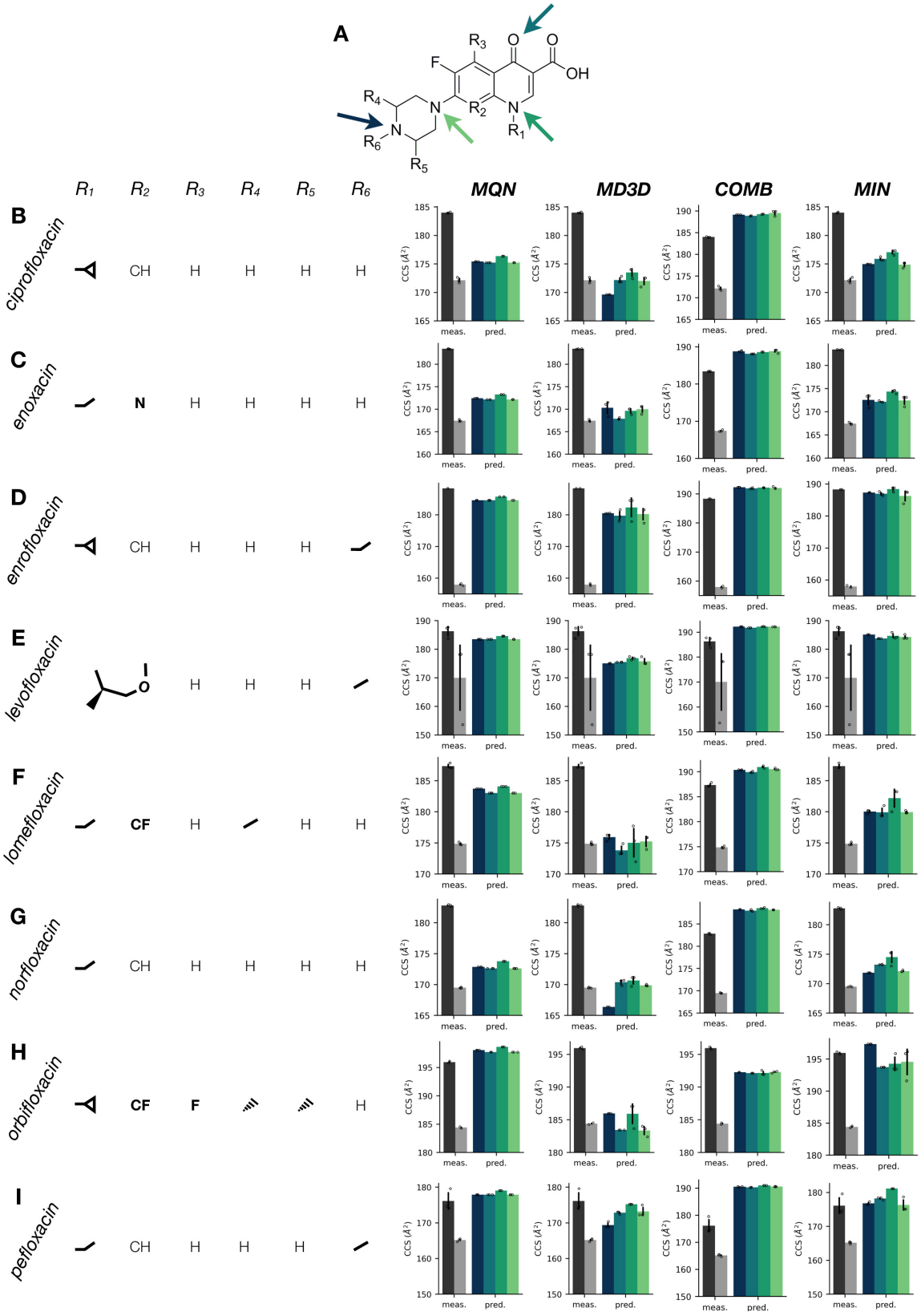
Comparison of measured CCS and CCS predicted using PA/EHS methods or by a ML model trained on the dmCCS database. Dotted lines correspond to linear fits on each set of values, with fit residuals presented in the lower plot.

4.5.8 *ML-Based CCS Prediction of Multimodal CCS*



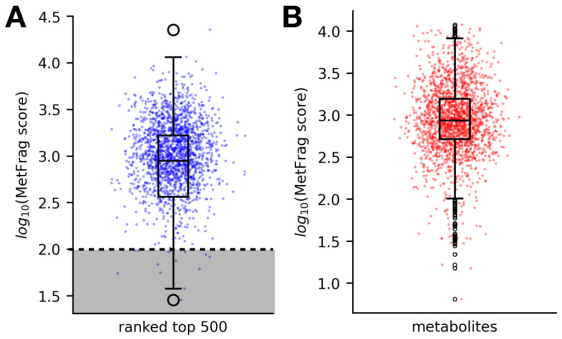
(**A**) Comparison of measured (black) and predicted (blue) CCS *vs. m/z* of BACs and their +O metabolites. parent CCS is denoted in circles and +O metabolites in triangles, with gas-phase compaction factors annotated next to metabolite values. Each plot presents CCS values predicted using ML models trained on dmCCS using different feature sets. (**B**) Representative structures of BACs and their ω-OH metabolites demonstrating the gas-phase compaction of metabolites relative to the parents. (**C**) Comparison of measured (hatched) and predicted (solid) CCS for terfenadine and its +O metabolites. Each plot presents CCS values predicted using ML models trained on dmCCS using different feature sets. (**D**) Representative structures of terfenadine and its +O metabolite demonstrating the gas-phase compaction of metabolite relative to the parent. (**E**) Comparison of measured (hatched) and predicted (solid) CCS for two protomers of cefpodoxime proxetil. Each plot presents CCS values predicted using ML models trained on dmCCS using different feature sets. (**F**) Representative structures of the two protomers of cefpodoxime proxetil. (**G**) Comparison of measured (dashed lines) and predicted (solid bars) CCS for the positional isomers of quercetin glucuronide. Each plot presents CCS values predicted using ML models trained on dmCCS using different feature sets. (**H**) Representative structures of the positional isomers of quercetin glucuronide.

4.5.9 *ML-Based CCS Prediction of Fluoroquinolone Protomers*



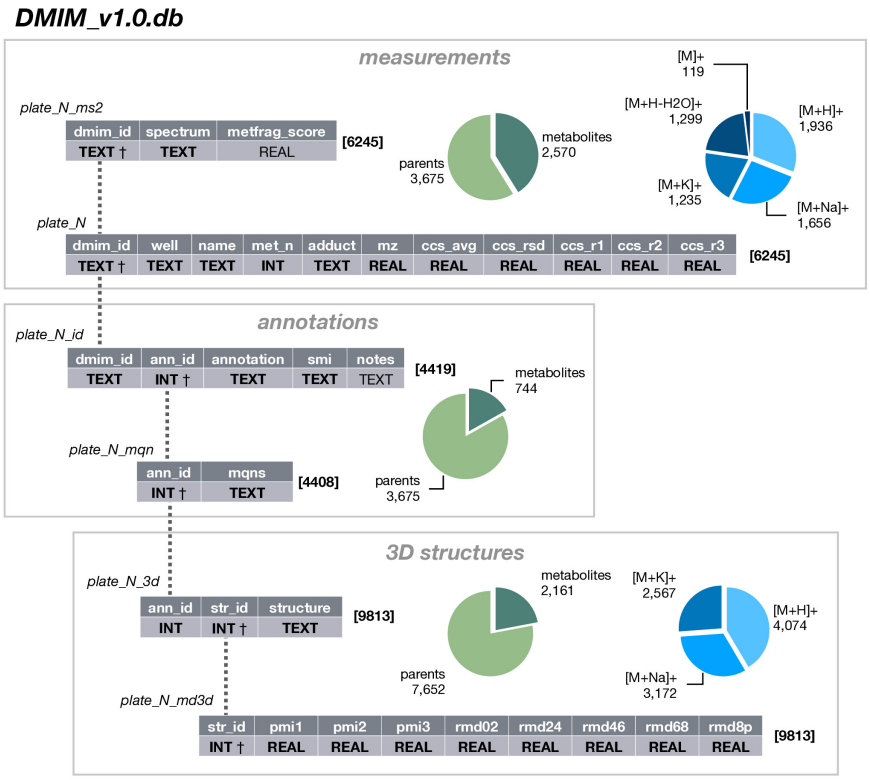
(**A**) General structure of fluoroquinolone antibiotics, with modeled protonation positions denoted by colored arrows. (**B-I**) Comparison of measured (grey) and predicted (color) CCS values for fluoroquinolone protomers. Each plot presents CCS values predicted using ML models trained on dmCCS using different feature sets.

4.5.10 Determination of MetFrag Fragmenter Score Cutoff to Remove Low-Quality Metabolite Annotations



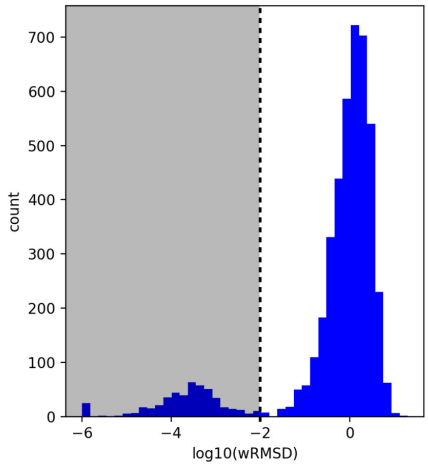
(**A**) Distribution of log-transformed MetFrag fragmenter scores for all parent compounds with true annotations ranked in the top 500 from the parent rank test. The dashed line indicates the empirically determined cutoff used to filter out metabolite annotations during construction of the dmCCS database. (**B**) Distribution of log-transformed MetFrag fragmenter scores for metabolites in dmCCS prior to filtering.

4.5.11 *Structure of the dmCCS Database*



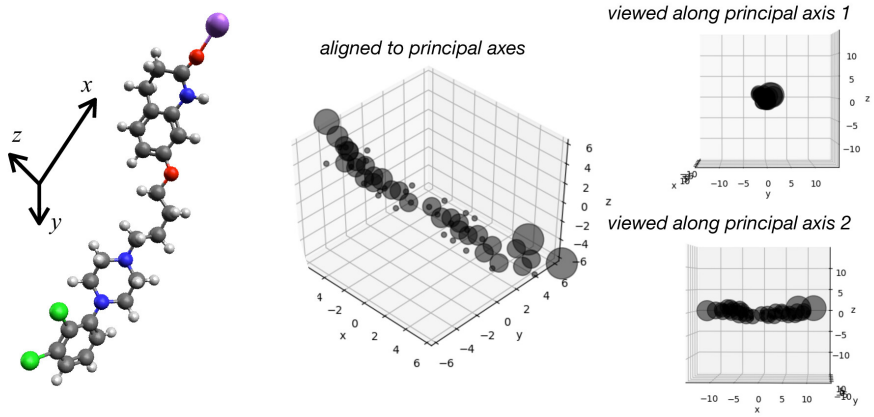
Overview of the structure of the dmCCS SQLite3 database. Each grey box represents the general type of information contained within each table, and pie charts reflect characteristics of these grouped tables. The names and data types are shown for each table, with bold datatypes indicating a required column and † indicating the primary key of the table. The dashed lines indicate the related columns between each table.

4.5.12 *Determination of RMSD Cutoff to Remove Duplicate 3D Structures*



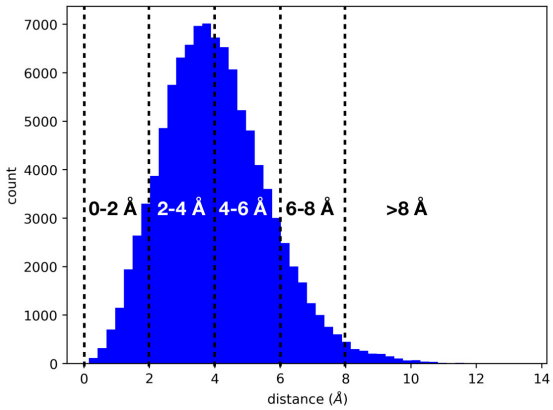
Distribution of log-transformed mass-weighted RMSD for all pairwise combinations of multiple 3D structures for all compounds in the dmCCS database. The dashed line indicates an empirically determined cutoff used for determination of whether individual 3D structures are distinct enough to be kept when assembling the final database.

4.5.13 Physical Interpretation of Principal Axes Within a Molecular Structure



Demonstration of the physical interpretation of principal axes in a 3D molecular structure. The principal axes x, y, and z are defined such that they each minimize the radial distribution of mass about successive orthogonal axes. The center image is a representation of the atomic positions from the structure on the left, with radii proportional to atomic masses. In this example, when viewed along the first principal axis (x, top right), there is very little radial distribution of masses about the central axis. In contrast, when viewed along the second principal axis (y, bottom right) the radial distribution of masses is in greater. The PMI are related to the magnitude of radial mass distribution about the respective principal axes, where increased radial mass distribution results in a higher moment.

4.5.14 Determination of Binning Intervals for Radial Mass Distributions



Mass-weighted radial atomic distance distribution for all 3D structures in the dmCCS database. Dashed lines indicate binning intervals used to compute binned radial mass distributions for individual structures as part of the MD3D features.

Chapter 5 Development of a Bioinformatic Tool with Specialized CCS Prediction to Support Lipidomics Data Analysis: LiPydomics

Portions of this chapter have been adapted and reproduced with permission from:

> Dylan H. Ross, Jang Ho Cho, Rutan Zhang, Kelly M. Hines and Libin Xu, LiPydomics: A Python Package for Comprehensive Prediction of Lipid Collision Cross Sections and Retention Times and Analysis of Ion Mobility-Mass Spectrometry-Based Lipidomics Data, *Analytical Chemistry*, 92 (**2020**) 14967-14975.

5.1 Introduction

Lipids are a class of biomolecules with broad biological importance, from being structural components of cell membrane and microdomains to serving as signaling molecules, and dysregulation of lipid metabolism is a common feature of many disease states.[139, 140] Lipidomics, the comprehensive analysis of lipids within a biological system, continues to gain popularity as it offers insight into metabolic phenotype and underlying mechanisms of these disease states.[141-143]

Lipid species can be broken into classes and subclasses on the basis of their headgroup chemistry, in addition to the composition of their fatty acyl tails (chain length; number, arrangement, and stereochemistry of double bonds).[144-146] Identification of lipid species may be performed at a variety of levels of structural detail, ranging from basic lipid class (Level 1) to complete molecular species (lipid class, subclass, and fatty acid isomeric composition, Level 5),[146, 147] according to the Lipidomics Standard Initiative (LSI). In lipidomics experiments, it is desirable to identify lipid species at the highest level possible in order to gain the most complete understanding of the biological processes being studied. The use of liquid chromatography coupled to ion mobility-mass spectrometry (LC-IM-MS) for lipidomics experiments has been demonstrated to provide a good balance between analytical throughput, resolution, and confidence in lipid identifications.[85, 86, 143, 148] Hydrophilic interaction liquid chromatography

(HILIC) is particularly advantageous as it provides resolution on the basis of lipid headgroups in the retention time dimension, while the orthogonal IM and MS separations allow for further delineation of overlapping subclass and fatty acid sum composition.[85-87] Therefore, this method generally allows Level 3 lipid identifications (lipid class/subclass and fatty acid sum composition).[146, 147]

Lipid identifications by IM-MS rely on reference CCS values to compare against, and while there are several large collections of experimental lipid CCS values in the literature,[40, 41, 50, 68, 85-88, 130] these collections do not yet comprehensively cover the vast lipid chemical space (both in terms of class and composition). CCS prediction using machine learning (ML) is one solution that has gained traction in recent years,[34, 68, 69, 71, 72, 83, 130] and variants of this general technique have been used by multiple groups to generate predicted CCS databases for lipids.[41, 68, 88] Zhou *et al.* were the first to construct regression models for predicting lipid CCS from a large set of molecular descriptors (45 and 66 for positive and negative modes, respectively) using support vector regression.[68, 149] Blaženović *et al.* trained several classification models (primarily K-nearest neighbor algorithm) using combinations of *m/z*, retention time, and CCS for prediction of lipid class and carbon number,[88] but their approach did not result in a predicted database covering theoretical lipids. We recently reported a clustering-to-prediction approach for comprehensive prediction of CCS of diverse chemical structures, including lipids and other types of molecules, but a comprehensive predicted lipid CCS database is still needed.[130] More recently, a large predicted CCS database was constructed using a regression model (XGBoost algorithm) that predicts lipid CCS from 328 molecular descriptors.[41] However, while the previous approaches perform well in lipid identification or classification,[41, 68, 88, 149] previous databases

mostly cover mammalian lipid species, have limited coverage of bacterial lipids, and have no built-in statistical functions, which are needed for a complete lipidomics workflow.

A typical lipidomics experiment may track hundreds to thousands of individual lipid species (features) across large number of biological samples. The dimensionality of these datasets (many features, fewer samples) can make interpretation of results difficult since macroscopic differences between samples often correspond to nuanced patterns of change across many features. To address this challenge, multivariate statistical analyses are often applied to lipidomics data in order to draw out the features that are most important or explanatory with regard to the specific biological question being probed. Commonly employed analyses range from simple statistical tests like per-feature ANOVA or Pearson correlation analysis, to multivariate dimensionality reduction analyses like principal components analysis (PCA) and partial least-squares discriminant analysis (PLS-DA). At a high level, use of such analyses allows large lipidomic datasets to be pared down to the set of lipid features that are altered by the specific biological conditions. Due to the complexity of the entire process and the fact that they are often implemented in different pieces of software, thus requiring moving data between different programs and converting between different formats, these analyses can be laborious to perform and difficult to apply consistently across multiple datasets.

To address the primary challenges faced in the analysis of lipidomics data (lipid identification and data complexity), we have prepared a Python package, *LiPydomics*, which contains a suite of tools for performing data analysis and lipid identification on HILIC-IM-MS lipidomics data in an efficient and reproducible fashion. To support lipid identification, we assembled a comprehensive experimental CCS database from the literature, trained ML models

for the prediction of CCS and HILIC retention times using simple but specialized feature sets, and built a predicted lipid database with broad coverage of lipid classes.

5.2 Results and Discussion

5.2.1 Development of an All-in-One Python Package for Comprehensive Lipidomics

To enable efficient and reproducible analysis of HILIC-IM-MS data, we developed a free and open source (MIT license) Python package – *LiPydomics*. The library contains several modules, each responsible for handling different aspects of lipidomics data analysis (Figure 5.5.1). The data module is responsible for the organization and storage of the lipidomics dataset itself, along with relevant metadata and any statistics calculated on the dataset using the stats module. It also contains utilities for saving/loading a dataset to file, exporting to a spreadsheet, and normalizing intensities. The stats module contains functions for applying statistical and multivariate analyses [ANOVA p-value, Pearson correlation, principal components analysis, partial least-squares discriminant analysis, partial least-squares regression analysis, two group Log2(fold-change)] on the dataset, and the plotting module contains functions for extracting data and generating standard plots, such as bar graph and heatmap, for visualization of the dataset and statistical analyses. The identification module is used for calibrating HILIC retention times and identifying lipid features at various confidence levels using *m/z*, HILIC retention times, and CCS, and contains utilities for accessing and re-training the CCS and HILIC retention time predictive models as discussed below. The identification module additionally contains a sub-package, *LipidMass*, which allows for easy generation of exact masses for a large selection of lipid classes. The interactive module contains a user-friendly text-based interface for performing lipidomics data analysis (Figure 5.5.2). This entire package, including the interface, can be easily installed on any computer with a compatible Python interpreter (version 3.5 or greater). The

assembly of the experimental database, development of CCS and retention time prediction models, the assembly of the predicted database, and demonstration of various modules are discussed in the following sections. A more in-depth overview of the library structure and function is available in the package documentation on GitHub (https://github.com/dylanhross/lipydomics).

5.2.2 Assembly of an Experimental Reference Lipid Database

A database of experimental reference lipid CCS values was assembled from data sets available in the literature.[40, 41, 50, 68, 85-88] In total, 7907 experimental CCS values were included in the database, representing 45 lipid classes (Table 5.6.1) and covering major lipid species present in both mammalian and bacterial systems. The database covers a variety of MS adducts with 5110 positive mode measurements and 2797 negative mode measurements. CCS measurements made on DTIM, TWIM, and TIMS instruments were included in the database (1285, 596, and 6026 values, respectively). Excellent agreement has already been demonstrated between measurements made on DT and TW platforms when lipid calibrants are used to calibrate CCS values in TW measurements.[33] However, a systematic comparison of TIMS[40, 41] CCS values against the established DT method has not yet been performed. To this end, we assessed the agreement between CCS values of overlapping lipids present in TW and TIMS datasets relative to DT values (Figure 5.5.3). Both positive and negative mode TW CCS values (Figures 5.5.3 A and B) show excellent agreement with DT values as evidenced by median relative errors (MDRE) much less than 1% and high degrees of correlation in CCS-CCS plots. Positive mode TIMS CCS values also showed excellent agreement with DT values (Figure 5.5.3 C); however, negative mode TIMS values (Figure 5.5.4 A) displayed an MDRE of ~1% with two apparent populations in the histogram. Negative mode TIMS CCS values from the two constituent

datasets[40, 41] were examined separately (Figures 5.5.4 B and C), and it was found that both datasets displayed MDREs > 1%, but in opposite directions. The CCS-CCS plots indicated distinct linear relationships between these TIMS CCS values and DT values for the two datasets. Therefore, in order to utilize both datasets for building CCS prediction model, we applied linear correction to each dataset toward DT values using equations shown in Figure 5.5.5 prior to ML model training. After this correction, the MDRE for negative mode TIMS CCS was -0.36%. Overall this database represents comprehensive coverage of currently available experimental lipid CCS values, with broad representation of lipid classes and IM-MS platforms. A particular strength of this comprehensive lipid database is the extended coverage of bacterial lipids, such as LysylPGs, AlanylPGs, AcylPGs, AcylPEs, GlcADG, and doubly charged lipids, such as CLs and LCLs, which were not covered in previous large-scale lipidomics datasets that contain mostly mammalian lipids.[41, 68]

5.2.3 Performance of CCS Prediction Using Machine Learning

A ML model was trained on data from the experimental lipid database to predict CCS values using only a minimal feature set consisting of encoded lipid class, fatty acid composition, encoded MS adduct, and *m/z*. These features do not require computation, which make them easy to assemble for a wide range of lipids and avoids reproducibility issues regarding structural assignment and descriptor generation. It has also been demonstrated that lipids display distinct trends in CCS with respect to *m/z*, lipid class, MS adduct, and acyl chain composition (visit CCSbase.net for interactive visualization of such trends),[50, 68, 85, 130] supporting their inclusion in our minimal feature set. Lipid classes of the same MS adducts with at least 20 measurements, resulting in 6394 CCS values in 22 lipid classes, were included for building the prediction model. This selected subset of measurements was split in an 80/20 proportion for training and

test datasets, respectively. The predictive model was trained using support vector regression with radial basis function kernel as described in the Experimental Section. This model was able to predict CCS values for lipids with high accuracy, achieving MAE, MDAE and RMSE scores of 1.05, 0.55, and 1.79 $Å^2$, respectively, on training dataset and 1.34, 0.78, and 3.03 $Å^2$, respectively, on the test dataset. Our model slightly outperformed a recently reported lipid-specific CCS prediction model trained on TIMS CCS values,[41] which achieved RMSE scores of 1.4 and 2.8 $Å^2$ on their training and test set data, respectively. With MDRE scores of 0.20 and 0.27 % on the training and testing data, respectively, our model also modestly outperforms the established LipidCCS predictor, which achieved MDRE scores of 0.50 and 0.42 %, respectively, on positive- and negative-mode intralab external validation sets (*i.e.*, data not seen during model training).[68] Relative standard deviation (RSD) was computed for 1667 lipid species having multiple reported CCS measurements in the combined CCS database (CCS was corrected as described above for negative mode data from Vasilopoulou *et al.*[40] and Tsugawa *et al.*[41]), and the mean and median RSD for this group were 0.60 and 0.50 %, respectively. Thus, the performance of our predictive model (specifically by MDRE) also compares favorably with variance in experimentally measured CCS values. Figure 5.5.6 shows CCS *vs. m/z* plots for MS adducts of several major lipid classes in both positive and negative modes along with corresponding relative errors of predicted CCS values relative to available measured values, where predicted values were produced using the ML model and measured values are taken from the experimental lipid CCS database. The predicted CCS and theoretical *m/z* values for all lipids span a comprehensive range of fatty acyl chain lengths and unsaturation degrees, with clear structural trends visible in this space as a function of both characteristics. The predicted CCS values for these lipid classes generally show excellent agreement with the measured values, with

residual CCS of predicted values falling mostly within 1% of measured values for most lipid species. We note that although there are some outliers in the measured values (possibly attributable to misidentified lipids), the contribution of these outliers to the training of the overall prediction model appears to be minimum as the majority of the consistent data outweigh the small number of outliers during model training. Plots for additional abundant lipid classes are available in Figure 5.5.6. These results demonstrate that high quality lipid CCS predictions can be obtained using a relatively small but specialized feature set, which includes lipid-specific information, such as lipid class, sum fatty acid composition, and fatty acid modifiers (Tables 5.6.2, 5.6.3, 5.6.4), with sufficient training data. Using these specialized features also allows easy expansion of the prediction model as experimental data for additional lipid classes becomes available since these features are easy to generate without appreciable computational effort.

5.2.4 Performance of HILIC Retention Time Prediction Using Machine Learning

A separate ML model was trained on data from the reference lipid database using a smaller feature set (minus the adduct types; see Experimental Section) to predict HILIC retention times based on the HILIC-IM-MS method established previously.[85-87] The trained predictive model achieved MAE, MDAE, and RMSE scores of 0.11, 0.08, and 0.15 minutes on the test set data, respectively. Figure 5.5.7 shows the distributions of predicted and measured retention times for the lipid classes that are well represented in the database spanning the retention time range of the established HILIC method. The predicted HILIC retention times show excellent agreement with measured values for all of these abundant lipid classes, and good agreement with values for less represented lipid classes (Figure 5.5.8). To allow the retention time database broadly applicable for HILIC methods run with different gradients and on different columns, we implemented a calibration method using multiple segments of linear interpolation between

calibrants. To demonstrate this utility, retention times of lipids extracted from a *Staphylococcus aureus* strain were measured using the established HILIC method,[86] as well as modified methods (see Experimental Section) using columns of different lengths (Figure 5.5.9 A) and/or different gradients (Figure 5.5.9 B). For each set of conditions, two to four individual lipids were used as calibrants to convert measured retention times to reference retention times. The lines in these plots represent the linear interpolation that occurs between the calibrants, and their overlap with the rest of the lipids not used for calibration demonstrates the utility and accuracy of this flexible retention time calibration scheme.

5.2.5 Assembly of a Predicted Lipid Database

Separate data tables were added to the reference lipid database containing predicted *m/z*, CCS, and HILIC retention time for a large collection of lipid species (145,388) comprising broad representation of lipid classes found in both mammalian and bacterial systems. This predicted data was produced by systematic enumeration of fatty acyl chain length (from 10 to 30 carbons per fatty acid, including both even and odd numbers) and unsaturations (from 0 to 6 per fatty acid) for 31 lipid classes (see Table 5.6.1) defined in the *LipidMass* module in *LiPydomics* (see below) and using ML models trained to predict HILIC retention time and CCS. 94,451 and 106,020 predicted CCS and HILIC retention time values were generated, respectively, covering 22 and 23 lipid classes, respectively. Together this predicted database vastly expands the coverage and depth of the reference lipid database and enables identifications of more lipid species than using the experimental reference data alone.

5.2.6 Automated Identification of Lipid Species at Different Confidence Levels

Identification of lipid species is performed by matching *m/z*, retention time, and CCS against values from the reference lipid database. Lipid identifications can be made at several

levels of confidence based on the number of components used for the identification and whether these were compared against experimental or predicted values. The available identification levels in this package are (in descending order of confidence): measured *m/z*, retention time, and CCS; predicted *m/z*, retention time, CCS; measured *m/z* and retention time; predicted *m/z* and retention time; measured *m/z* and CCS; predicted *m/z* and CCS; measured *m/z*; and predicted *m/z*. The user may specify one of these confidence levels when undertaking lipid identification or use a tiered approach, where the highest confidence level is tried first for each lipid species and successive levels are attempted until an identification is made. If retention time calibration has been set up, the calibrated retention time is automatically used for lipid identification. Whenever lipid identifications are made, both the putative identification(s) and the level of confidence are stored for each lipid feature. When multiple annotations are made for a single feature, the putative identifications are ranked by a score reflecting the agreement between query and reference values, computed as the dot product of residuals from the matched values normalized by their respective search tolerances. All lipid identifications made by this method are of LSI Level 3,[147] *i.e.*, lipid class, subclass, and fatty acid sum composition. Overall, this utility allows users to identify lipids in an efficient, automated fashion. Additionally, the predicted lipid database was added to our existing web interface (https://CCSbase.net)[130] so that users can query this data without using the complete *LiPydomics* package.

5.2.7 Demonstration of LiPydomics Functionality

In order to demonstrate the functionality of *LiPydomics*, we reanalyzed data from our recently published study examining lipidomic changes associated with antibiotic resistance in methicillin-resistant *Staphylococcus aureus* (MRSA) strains.[150] Aligned and peak-picked HILIC-IM-MS data acquired in negative ESI mode were used for this analysis. The data contained

normalized intensities for 3647 features from 4 different MRSA strains (JE2 parent strain, 'Par'; JE2-derived strain with reduced susceptibility to daptomycin, 'Dap2'; reduced susceptibility to dalbavancin, 'Dal2'; and reduced susceptibility to vancomycin, 'Van4'), each with 4 biological replicates. Lipids were identified by matching on predicted *m/z*, retention time, and CCS (using search tolerances of 0.02 Da, 0.2 min, and 3.0 %, respectively), or measured *m/z* and CCS to cover lipid classes without retention time information. Using the stats module, we computed a 3-component PCA to see how the groups separated according to their overall variance. Figure 5.5.10 A shows the PCA projections for each sample along the first two principal components, colored by strain. These components capture around 90% of the total variance in the dataset, and samples from each group cluster together and separate from other groups in this space, indicating that there are distinct characteristics that are associated with each strain. We next looked specifically at the comparison between the lipid profiles of the daptomycin-resistant Dap2 and the parent strains that have been examined previously. First, we performed a partial least-squares discriminant analysis (PLS-DA) and Pearson correlation between Dap2 and Par. The PLS-DA projections (Figure 5.5.10 B) show excellent separation between the strains, and similar levels of intra-group variance. The S-plot (PLS-DA x-loadings *vs.* Pear-son correlation) highlights multiple features that are highly abundant in either strain and different between strains (Figure 5.5.10 C). Examination of these discriminating features reveals systematic changes in the DGDG, LysylPG, and FA lipid classes between the strains. To explore these effects at a higher level, we computed the Log2(fold-change) between Dap2 and Par and produced heat maps of all annotated lipids from each of these classes using the plotting module (Figure 5.5.10 D-F). From these heat maps, we observed a general decrease in DGDGs, increase in LysylPGs, and increase in FAs between 15 and 21 carbons in length in Dap2 strains relative to Par. It should be noted

that these heat maps include lipid features annotated as unsaturated lipids, however, these are unlikely to be found in the bacterial system studied. Indeed, close examination of those features suggest that most have low signal intensities likely corresponding to background signals. We also produced bar plots using the plotting module, showing the mean intensities with standard deviation in Dap2 and Par strains for the most significantly altered lipids in each of the previously discussed lipid classes (Figure 5.5.10 G-I). Overall, this analysis using *LiPydomics* reproduced the key findings of the previous report,[150] and was performed with only 19 lines of Python code on minimally processed data.

Separately, we used both positive and negative ESI mode data from the same study to perform lipid identification using the predicted lipid database at varying levels of confidence. Figure 5.5.10 J shows the number of lipids identified at each level of confidence for both ESI modes. The number of lipids identified decreases steadily as we progress from matching based solely on *m/z* (lowest confidence) to matching based on *m/z*, retention time, and CCS (highest confidence), using search tolerances of 0.02 Da, 0.2 min, and 3.0 %, respectively, across all tests. This example demonstrates the flexibility of the lipid identification utility in *LiPydomics*, which allows a user to prioritize annotation coverage or confidence as it suits the biological problem being studied.

5.3 Experimental

5.3.1 Reference Lipids Database Assembly

A comprehensive collection of lipid CCS values was assembled from individual CCS collections available in the literature[40, 50, 68, 85-88] into a single database of reference lipids for use in lipid identification. Briefly, the source datasets were each manually examined for errors and the relevant data (*i.e.* lipid name, MS adduct, *m/z*, and CCS) from each was converted into a

JSON format, yielding clean and consistently formatted data with separate files for each dataset. A SQLite3 relational database was initialized with a table to hold the reference CCS values. A series of build scripts developed in-house were used to assemble the combined database from individual cleaned data files in a reproducible fashion. During database assembly, the lipid names were parsed for relevant information (*i.e.* lipid class, sum composition of fatty acids [number of carbons and unsaturation degrees], presence of ether lipids) and this information along with metadata reflecting measurement conditions was associated with each entry. CCS values measured on drift tube (DT), traveling wave (TW), and trapped ion mobility spectrometry (TIMS) instruments were included, and those measured on TW were calibrated using lipid standards. For the individual datasets that were measured using the same HILIC-IM-MS protocol as reported previously (referred to hereafter as the established HILIC method),[85-87] the retention time was also stored with each lipid measurement. Additional tables containing predicted *m/z*, CCS, and retention times were also added to the database and populated as described below.

5.3.2 Generation of Exact Lipid m/z Values

Theoretical *m/z* values were produced systematically for lipid classes using a sub-package within *LiPydomics* (*lipydomics/identification/LipidMass*). Monoisotopic masses were computed from the lipid classes and subclasses, fatty acid compositions (ranging from 10-30 carbons, including both even and odd numbers, and 0-6 unsaturations per fatty acid), and MS adducts using a method similar to that used in LipidPioneer.[151] Separate functions were used for each lipid class, and lipid classes are further grouped into sphingolipids (Cer, GlcCer, SM), glycerolipids (DG, TG), glycolipids (MGDG, DGDG, GlcADG), glycerophospholipids (AcylPG, AcylPE, AlanylPG, CL, LysylPG, PA, PC, PE, PG, PI, PIP, PIP2, PIP3, PS), lysoglycerophospholipids (LPA, LPC, LPE, LPG, LPI, LPS, LCL), and free fatty acids (FA).

Lipid abbreviations follow the standards established by LIPID MAPS (see Table 5.6.5 for lipid class abbreviations).[144, 145] Exact *m/z* values were computed for lipids using a number of commonly observed ESI adducts in positive ($[M]^+$, $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH4]^+$, $[M+H-H_2O]^+$, $[M+2Na-H]^+$, $[M+2K]^{2+}$) and negative ($[M-H]^-$, $[M+HCOO]^-$, $[M+CH_3COO]^-$, $[M+Cl]^-$, $[M-2H]^{2-}$) modes.

5.3.3 Prediction of CCS Using Machine Learning

Predicted CCS values for lipids were produced using a predictive model trained on the reference lipid database. For all reference lipids, lipid classes, fatty acid modifiers (*e.g.* "p" indicating a plasmenyl lipid), and MS adducts were each encoded into one-hot binary vectors (22, 3, and 11 features, respectively, see Tables 5.6.2, 5.6.3, 5.6.4 for specific encodings). Only the lipid classes, fatty acid modifiers, and MS adducts with sufficient representation (at least 20 measurements) in the database were explicitly encoded. The final feature vector was prepared by appending fatty acid sum composition (number of carbons and unsaturations) and observed *m/z* to the binary encoded vectors for each lipid (total of 39 features; Tables 5.6.2, 5.6.3, 5.6.4). A subset of the reference lipid database (6394 measurements; Table 5.6.2) consisting of only the explicitly encoded lipid classes, fatty acid modifiers, and MS adducts was selected for use in CCS prediction. This subset was randomly split into training and test data sets in proportions of 80% and 20%, respectively, and the test dataset was set aside until model training was complete. The training data were scaled such that all features had a variance of 1 to avoid arbitrary overweighting of individual features based on their scale. A support vector machine with radial basis function kernel (svr) was selected for CCS prediction based on preliminary testing, and hyperparameters were optimized using a grid search with 5-fold cross validation on the training data. Using the optimized hyperparameters, the model was trained on the full set of training data

and performance metrics (mean/median absolute error, median relative error, and root mean squared error; MAE, MDAE, MDRE, and RMSE, respectively) were computed on the training data. Finally, the same performance metrics were computed with the trained model on the test set data to validate model performance on unseen data.

5.3.4 Prediction of HILIC Retention Time Using Machine Learning

Predicted HILIC retention times were produced using a predictive model trained on all entries in the reference lipid database that contain HILIC retention times measured using the HILIC method mentioned above and under the same experimental conditions (596 lipids in total; Table 5.6.6).[85-87] A smaller feature set (26 features) was used for retention time prediction compared to CCS prediction: binary encoded lipid class (22 features), fatty acid modifier (2 features) and sum composition (2 features). The smaller number of lipid classes and fatty acid modifiers present in the feature set are reflective of the fact that this subset represents less than 10% of the complete reference lipids database (596 of 7907 lipids, see Table 5.5.16 for specific encodings). Additionally, *m/z* and encoded MS adduct were not included since these do not relate directly to chromatographic retention time. This subset was split into training and test data sets as described above for CCS prediction. A multivariate linear regression model was used for retention time prediction. The model was fit and performance metrics (MAE, MDAE, and RMSE) were computed using the training data. Finally, performance metrics were computed with the trained model on the test set data to validate model performance on unseen data.

5.3.5 Calibration of HILIC Retention Time

HILIC retention times present in the reference lipid database were measured using an established HILIC method mentioned above, [85-87] and the ML model for predicting retention times was trained on these retention times. In order to be able to compare retention times

acquired using other HILIC conditions, a retention time calibration utility was developed and included in the library. This utility uses linear interpolation of known standards to calibrate retention times of a given HILIC gradient to the retention times in the database. Multiple calibration points can be used in order to approximate nonlinear relationships between reference and measured HILIC retention times. This approach offers excellent calibration accuracy and flexibility, without the complications of choosing a fitting function when the relationship is nonlinear. Once a retention time calibration has been set, the calibrated retention time is automatically used for compound identification. To evaluate this calibration strategy, we first examined three different gradients on the same Phenomenex Kinetex HILIC column (100 × 2.1 mm, 1.7μm) with solvent A being acetonitrile/water (50/50) with 5 mM ammonium acetate and solvent B being acetonitrile/water (95/5) with 5 mM ammonium acetate: (1) 0-1 min, 100% B; 1-4 min, 100%-90% B; 4-7 min, 90%-70% B; 7-8 min, 70% B; 8-9 min, 70%-100% B, 9-12 min, 100%B; (2) 0-0.8 min, 100% B; 0.8-1.8 min, 100%-90% B; 1.8-2.8 min, 90%-70% B; 2.8-3.8 min, 70% B; 3.8-4.8 min, 70%-100% B, 4.8-8 min, 100%B; (3) 0-2 min, 100% B; 2-8 min, 100%-90% B; 8-14 min, 90%-70% B; 14-15 min, 70% B; 15-16 min, 70%-100% B, 16-19 min, 100%B. We then examined three different columns from the Phenomenex Kinetex HILIC series (100 x 2.1, 50 x 2.1, or 30 x 2.1 mm; 1.7 μm). The gradients for different columns were changed in linear relation to their lengths. Specifically, the gradients for 50-mm and 30-mm columns were as follows: (1) 0-0.5 min, 100% B; 0.5-2 min, 100%-90% B; 2-3.5 min, 90%-70% B; 3.5-4 min, 70% B; 4-4.5 min, 70%-100% B, 4.5-6 min, 100%B; (2) 0-0.3 min, 100% B; 0.3-1.2 min, 100%-90% B; 1.2-2.1 min, 90%-70% B; 2.1-2.4 min, 70% B; 2.4-2.7 min, 70%-100% B, 2.7-3.6 min, 100%B.

5.3.6 Statistical and Multivariate Analyses for Lipidomics Data

All statistical and multivariate analyses implemented in this library are available from the SciPy[152] and Scikit-Learn[95] Python libraries, respectively. These analyses use either the raw or normalized intensities from samples belonging to user-specified groups, and the computed statistics are automatically stored along with the dataset. The analyses generally fall into two categories: untargeted and targeted. The untargeted analyses (ANOVA and PCA) can be computed on two or more groups in an unsupervised fashion, *i.e.* they report on intrinsic characteristics of the data used in their calculation. The targeted analyses [Pearson correlation analysis, PLS-DA, Log2(fold-change)] are performed between two specified groups in a supervised fashion, where features that differ between the specified groups are highlighted. Additionally, PLS-RA may be performed in order to find correlations between lipidomic data and an external continuous variable.

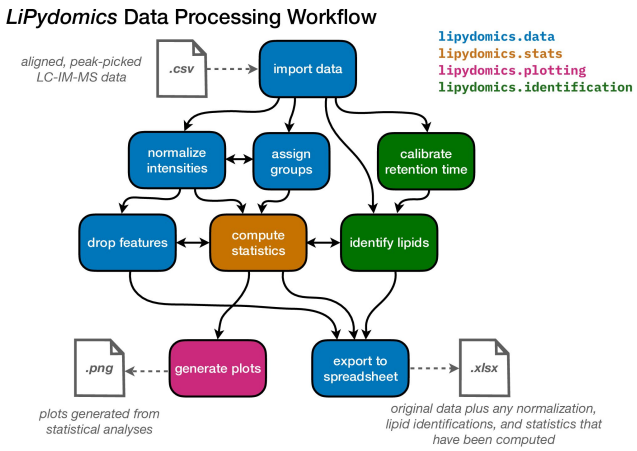
5.4 Conclusion

The key strengths of *LiPydomics* as a resource for lipidomics data analysis lie in its large coverage of lipid classes (both experimental and predicted), versatility (from statistical analysis to identification), reproducibility, extensibility, and ease of use. Additionally, data analyses can be partially or fully automated through scripting, further enhancing the reproducibility and efficiency of such analyses. The unique reference lipid database contains measured and predicted *m/z*, retention time, and CCS values, with broad coverage of common and rare lipid species from both mammalian and bacterial systems, the latter being underrepresented in other lipid databases to date. The predicted *m/z*, retention times, and CCS values display good agreement with measured values and cover a comprehensive range of lipid classes and fatty acid compositions, enabling identification of more lipids than would be possible using measured values alone. Thus,

this comprehensive lipid database enables identification of lipid species at the level of class, subclass, and sum fatty acid composition (LSI Level 3) from diverse biological systems. The package (including the lipid database and prediction models) is also built to be highly extensible and customizable, allowing easy expansion as more data becomes available and optimization for specific analysis workflows via its flexible and well documented API. The text-based user interface makes the library more broadly accessible to those who are not familiar with Python programming. Together, these attributes make *LiPydomics* a unique and comprehensive tool for performing analysis of HILIC-IM-MS lipidomic data.

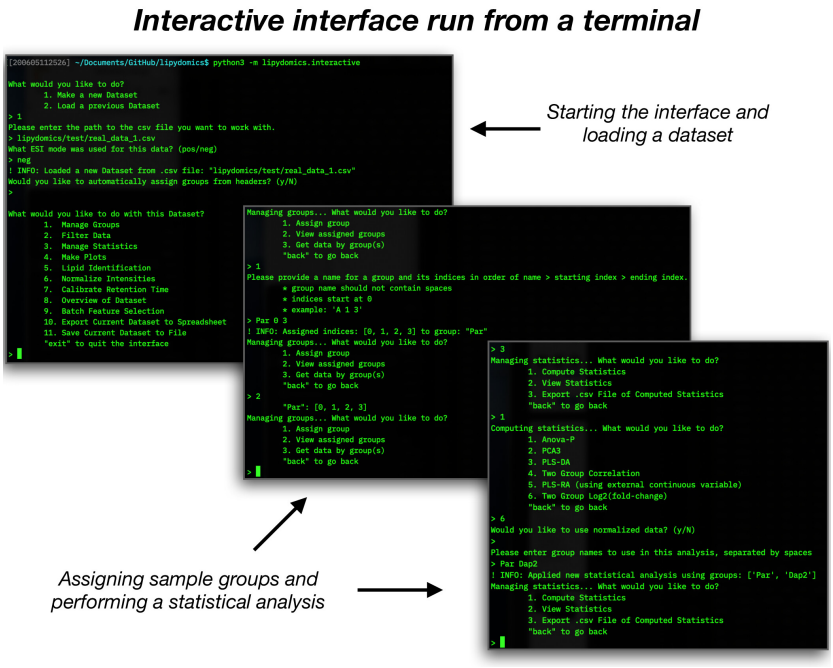
5.5 Figures

5.5.1 LiPydomics Data Processing Workflow

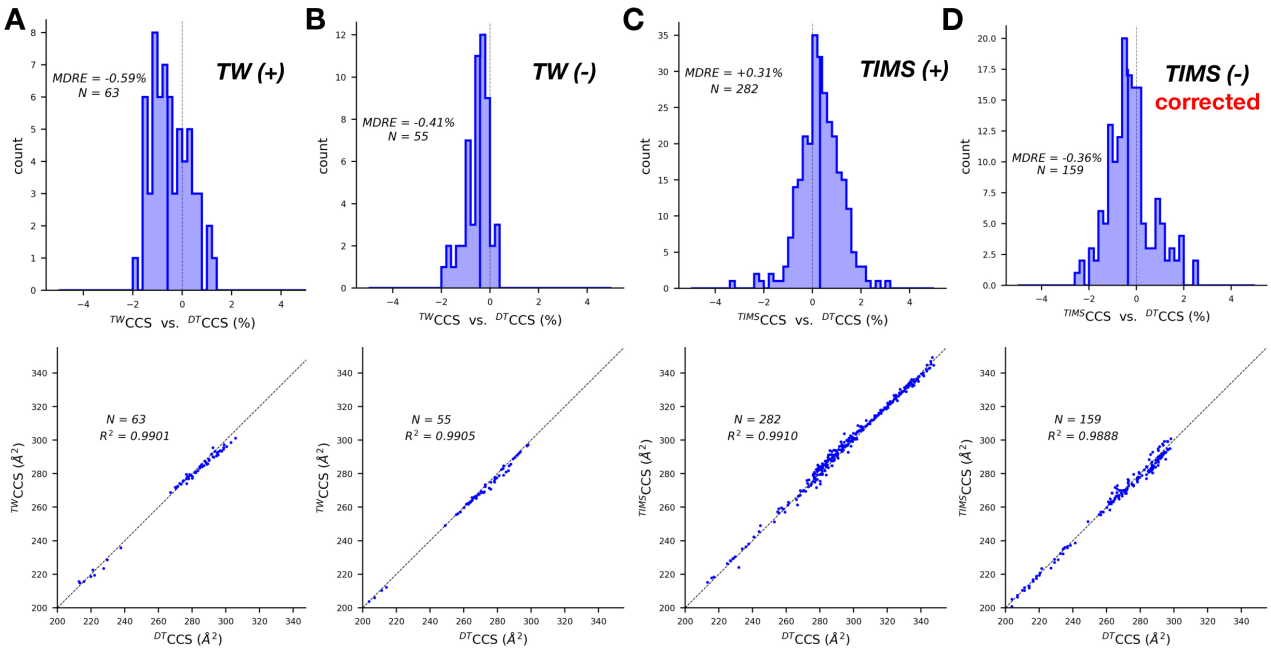


Schematic representation of the *LiPydomics* data processing workflow. Input/output files (with corresponding file formats) are depicted in grey. Each cell represents an individual data processing step and arrows reflect possible workflow sequences. Each cell is color-coded according to the specific module used to perform each step. The consistent and modular API of *LiPydomics* allows for data processing workflows to be customized to the needs of a particular experiment.

5.5.2 *LiPydomics Interactive Interface Example*

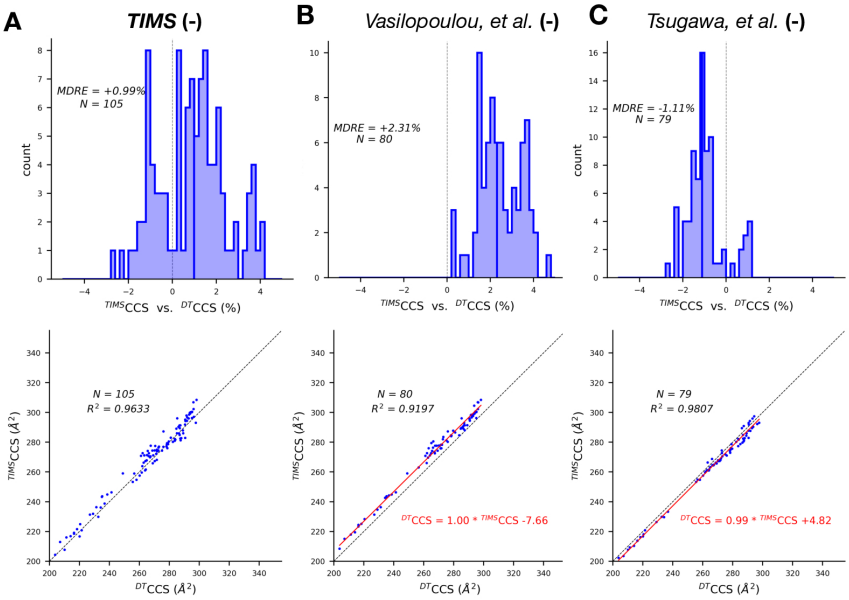


5.5.3 *Comparison of Lipid CCS Values Measured on DT, TW, and TIMS Instruments*



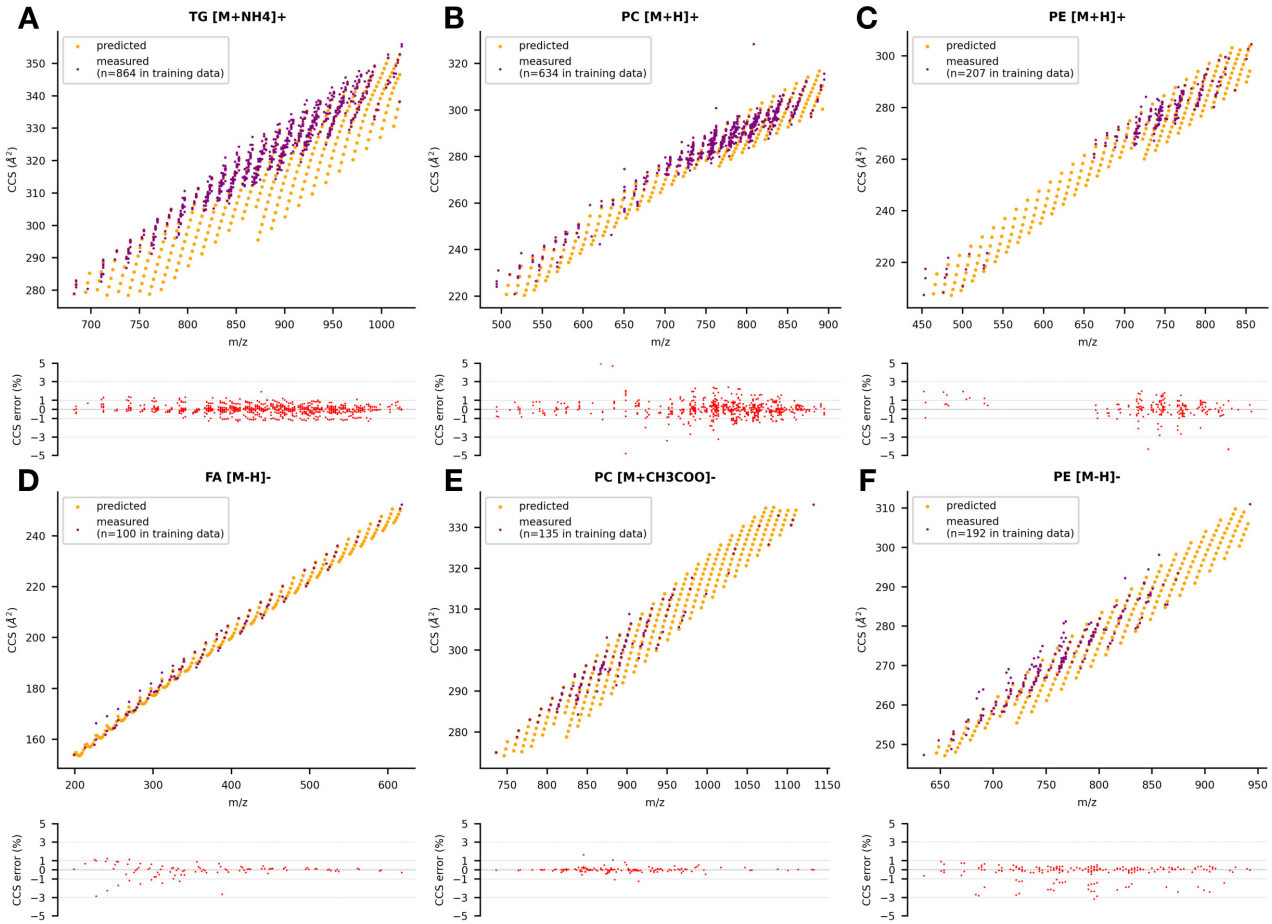
Comparisons of ^{TW}CCS (**A**, **B**) and TIMSCCS (**C-D**) *vs.* ^{DT}CCS values for lipids in the experimental database. Histograms and CCS-CCS plots provided for the comparisons of the following groups to corresponding overlapping DT values: TW positive mode (**A**), TW negative mode (**B**), TIMS positive mode (**C**), and TIMS negative mode with linear corrections applied (**D**). Dotted lines show the linear equation y = x.

5.5.4 *Correction of Negative Mode TIMS CCS*



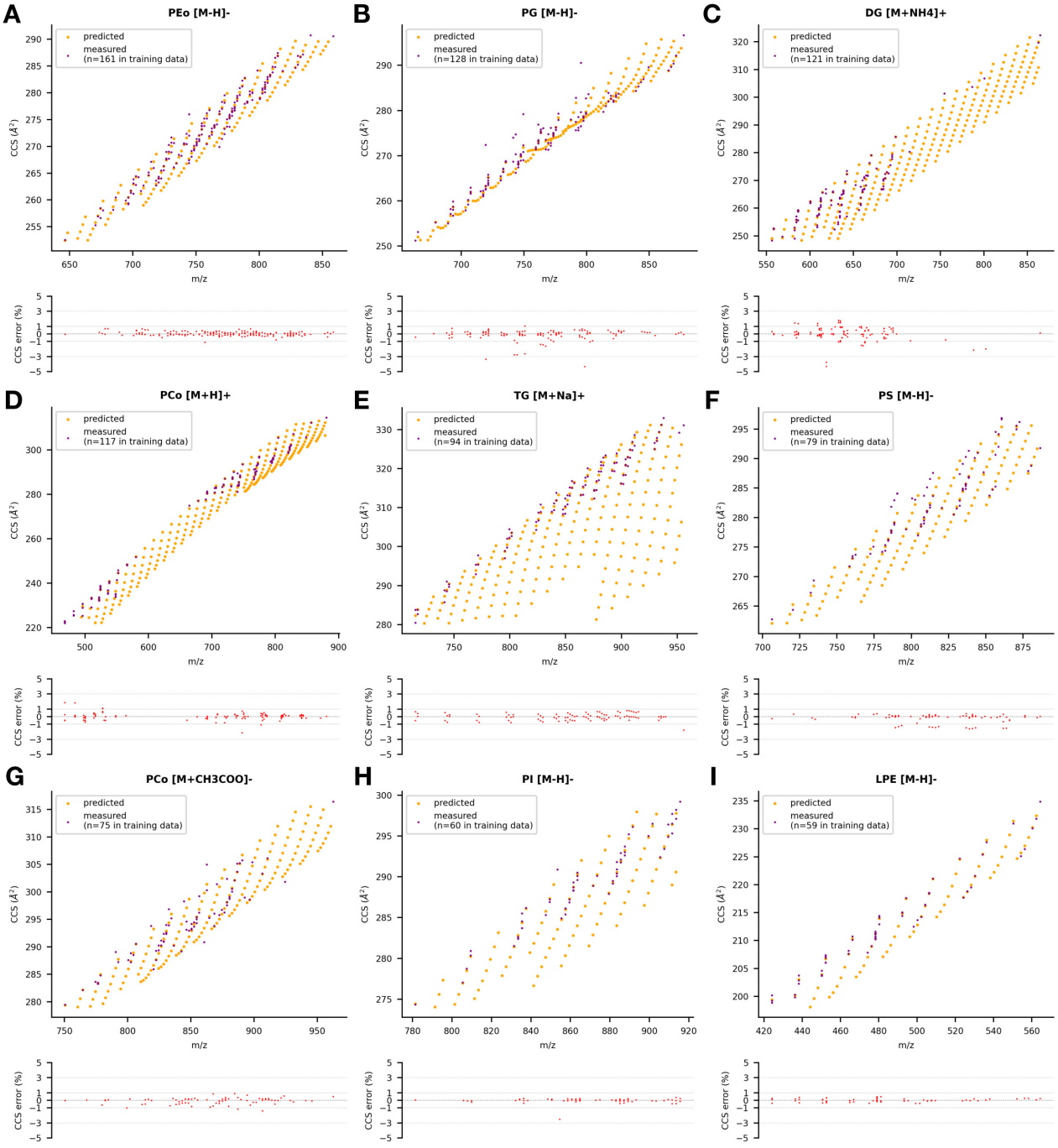
(**A**) Comparison of uncorrected negative mode ^{TIMS}CCS *vs.* ^{DT}CCS. (**B**) Comparison of uncorrected negative mode ^{TIMS}CCS from Vasilopoulou, *et al.*[40] *vs.* ^{DT}CCS, including parameters from linear regression used for CCS correction. (**C**) Comparison of uncorrected negative mode ^{TIMS}CCS from Tsugawa, *et al.*[41] *vs.* ^{DT}CCS, including parameters from linear regression used for CCS correction.

5.5.5 CCS Prediction Performance for Abundant Lipid Classes

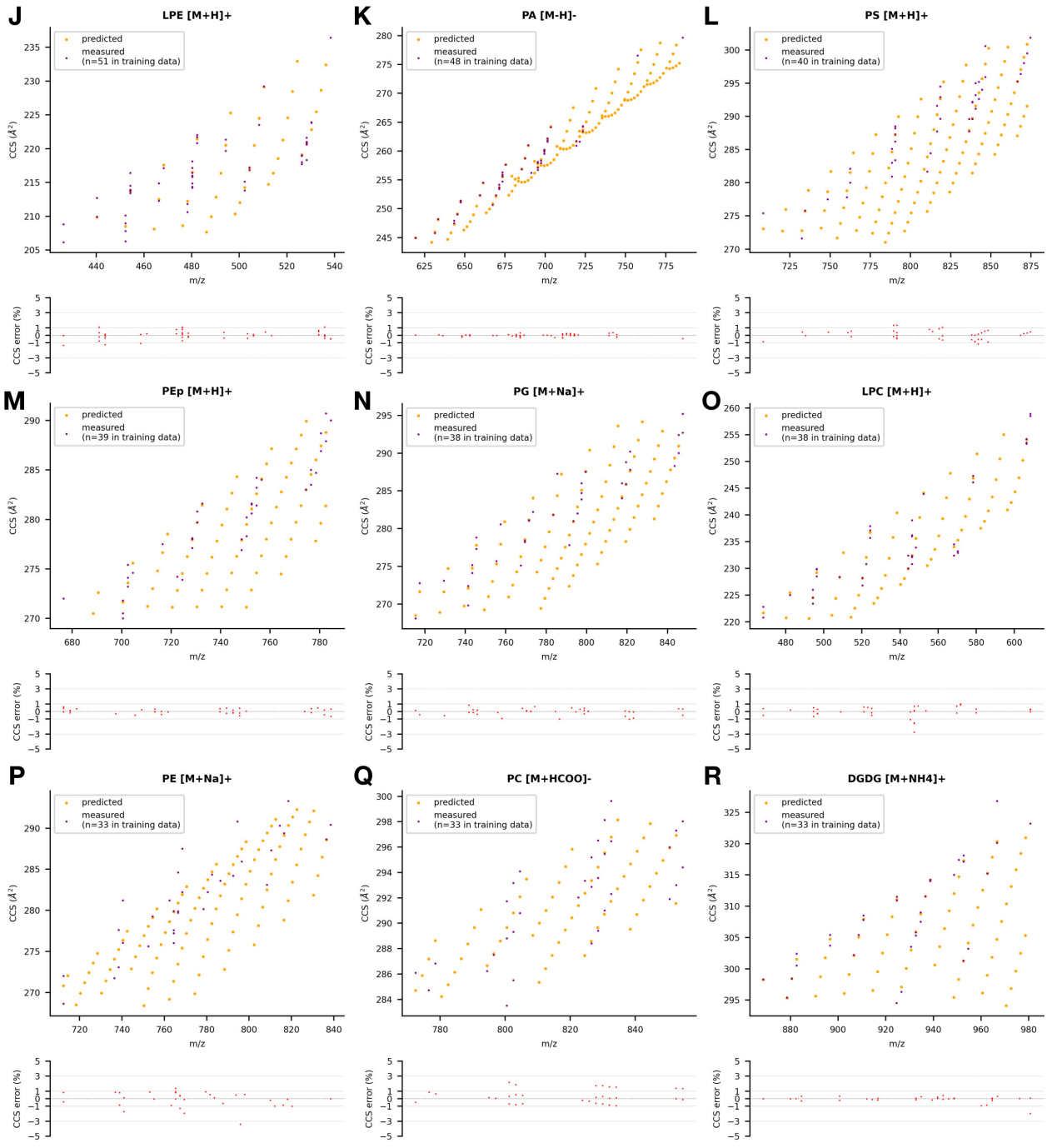


Predicted (gold) and measured (purple) lipid CCS values and relative prediction errors for abundant lipid species in the lipid CCS database in positive (**A**-**C**) and negative (**D**-**F**) ESI modes.

5.5.6 CCS Prediction Performance for Other Lipid Classes

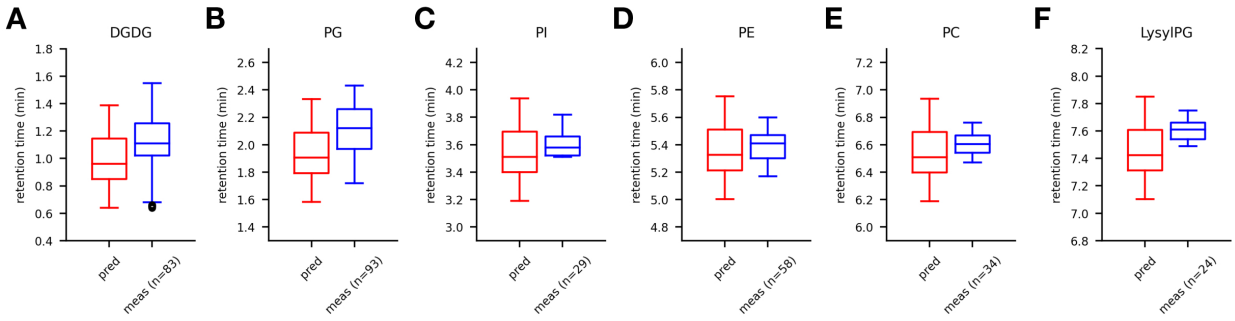


Predicted (gold) and measured (purple) lipid CCS values for lipid classes and MS adducts present in the reference CCS database.



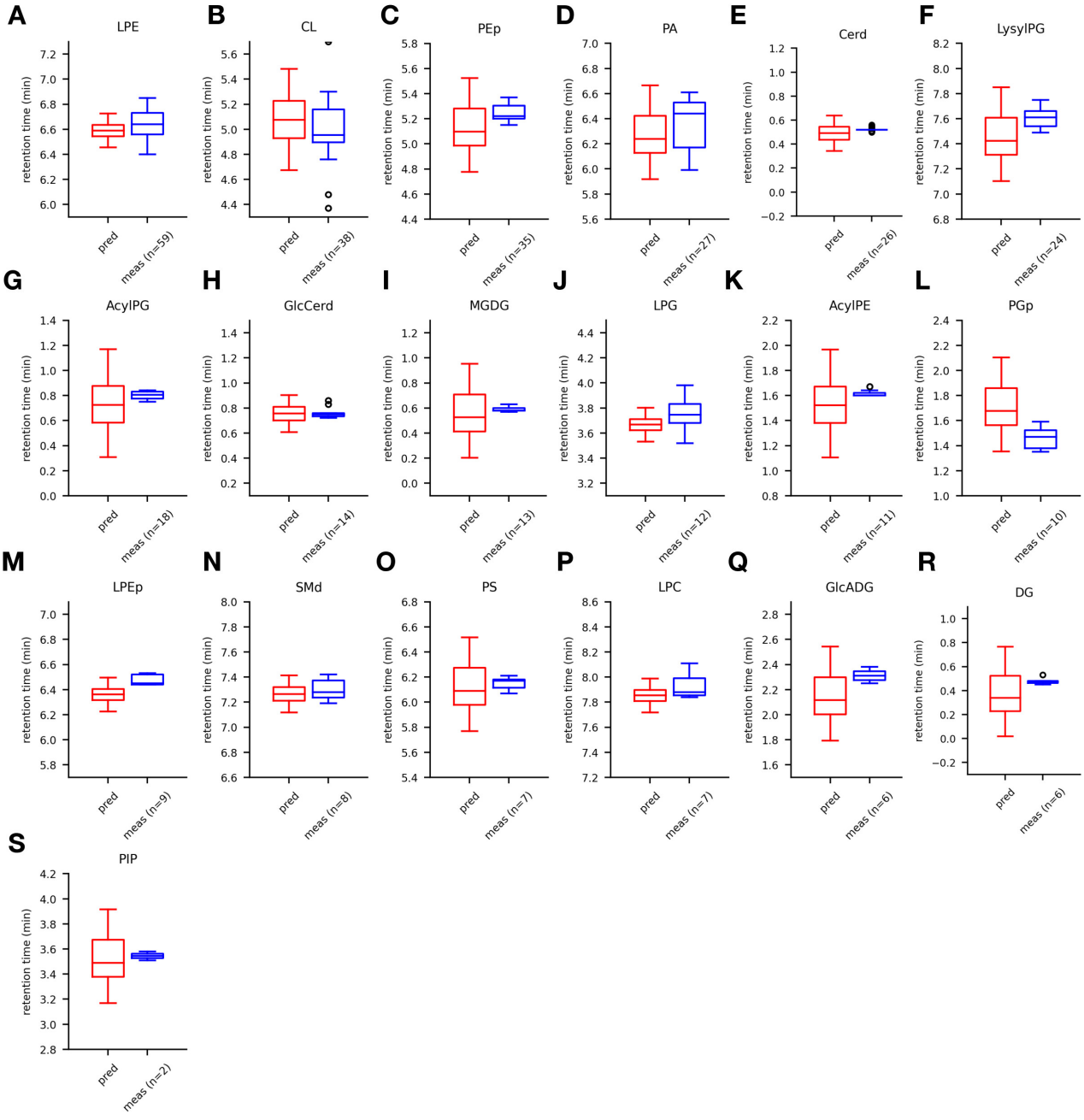
(continued) Predicted (gold) and measured (purple) lipid CCS values for lipid classes and MS adducts present in the reference CCS database.

5.5.7 Retention Time Prediction Performance for Abundant Lipid Classes



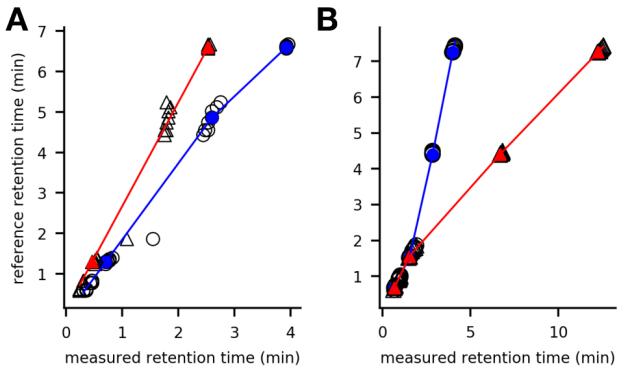
Distributions of predicted (red) and measured (blue) HILIC retention times for major lipid classes (**A**: DGDG; **B**: PG; **C**: PI; **D**: PE; **E**: PC; **F**: LysylPG) spanning the retention time range of the established HILIC method described in the Experimental Section.

5.5.8 *Retention Time Prediction Performance for Other Lipid Classes*



Distributions of predicted (red) and measured (blue) HILIC retention times for lipid classes from the theoretical and reference lipid databases.

5.5.9 Demonstration of HILIC Retention Time Calibration



Demonstration of linear interpolation retention time calibration using data collected with columns of different lengths (**A**) or different gradients (**B**). Open circles and triangles in A represent measured retention times from experiments using 50 mm and 30 mm columns, respectively, plotted against retention time from the established HILIC method[85] (100 mm column). Open circles and triangles in **B** represent measured retention times from experiments using a faster and slower gradient, respectively, plotted against retention time from the established HILIC method[85] using the same 100 mm column. Solid colored points represent the individual lipids chosen as calibrants, with colors distinguishing between the two experiments. The colored lines reflect the linear interpolation between calibrants that used for converting measured retention times to their reference equivalent.

5.5.10 Analysis of Lipidomics Data from Antibiotic-Resistant MRSA Strains

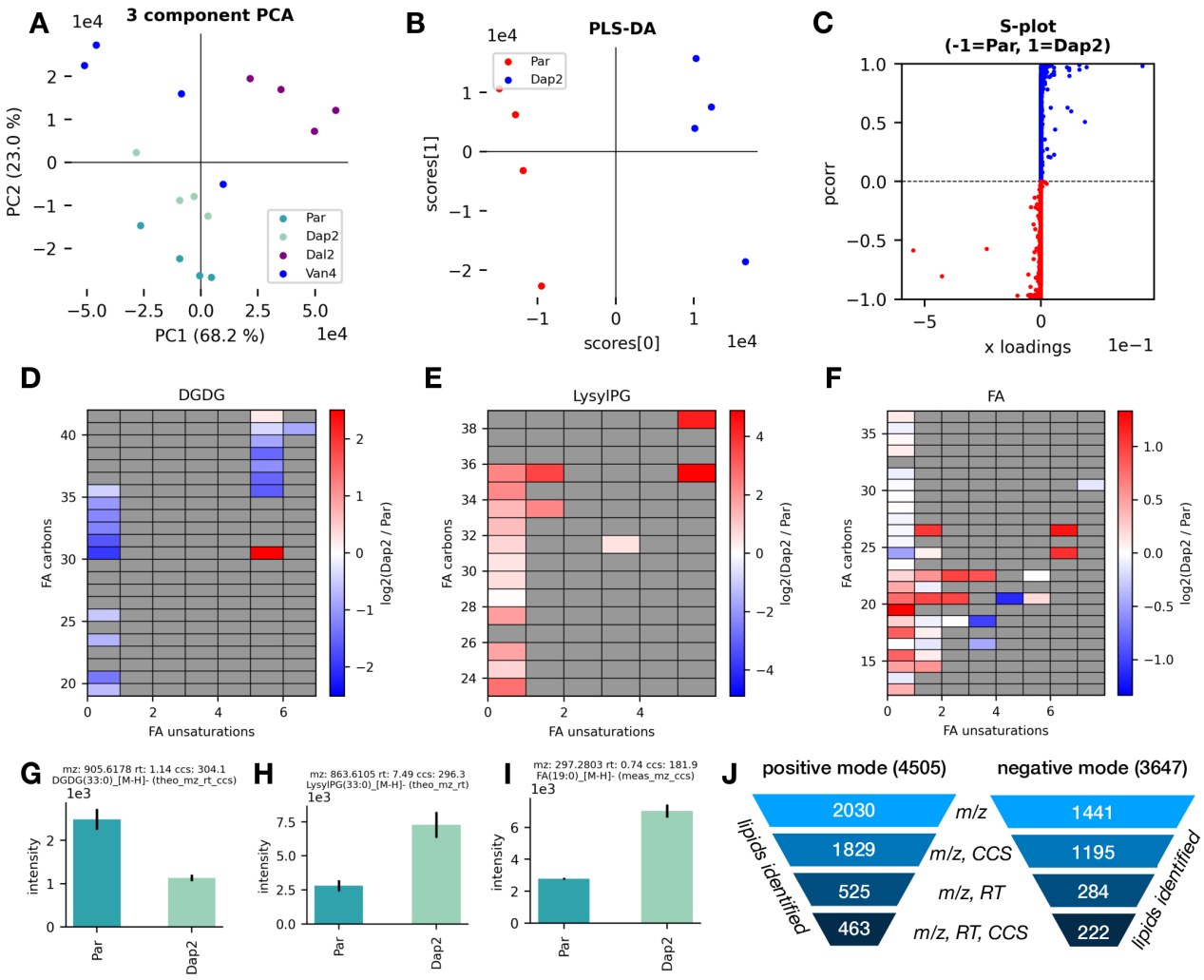


Illustration of LiPydomics functions by analyzing antibiotic-resistant MRSA strains. (**A**) PCA projections for parent strain (*Par*) and strains with resistance to daptomycin, dalbavancin, or vancomycin (*Dap2*, *Dal2*, *Van4*, respectively). (**B**) PLS-DA projections computed between *Par* (red) and *Dap2* (blue) strains. (**C**) S-plot showing individual features driving separation between *Par* (red) and *Dap2* (blue) strains. (**D-F**) Heatmaps of Log2(fold-change) between *Par* and *Dap2* strains for major bacterial lipid classes. (**G-I**) Bar plots of individual lipids displaying most significant differences between *Par* and *Dap2* strains. (**J**) Number of lipids identified from positive and negative mode data using various combinations of predicted identifiers (*m/z*, CCS, and/or HILIC retention time).

5.6 Tables

5.6.1 Counts of Lipid Classes in Reference Lipid Database

lipid class	count	lipid class	count	lipid class	count
PC$^{\mathbf{CCS,RT}}$	1900	**MGDG**$^{\mathbf{CCS,RT}}$	99	BMP	6
TG$^{\mathbf{CCS}}$	1834	**FA**$^{\mathbf{CCS}}$	91	DGGA	6
PE$^{\mathbf{CCS,RT}}$	1233	FAHFA	48	**GlcADG**$^{\mathbf{RT}}$	6
PS$^{\mathbf{CCS,RT}}$	328	**LPS**$^{\mathbf{CCS}}$	43	NAGly	5
SM$^{\mathbf{CCS,RT}}$	276	**LPG**$^{\mathbf{CCS,RT}}$	41	PMeOH	4
LPE$^{\mathbf{CCS,RT}}$	270	PEtOH	35	SQDG	4
DG$^{\mathbf{CCS,RT}}$	267	CAR	34	**AlanylPG**	3
PG$^{\mathbf{CCS,RT}}$	253	**CL**$^{\mathbf{RT}}$	30	DLCL	3
LPC$^{\mathbf{CCS,RT}}$	237	**LPI**	30	LacCer	3
PI$^{\mathbf{CCS,RT}}$	192	**AcylPG**$^{\mathbf{CCS,RT}}$	18	ADGGA	2
GlcCer$^{\mathbf{CCS,RT}}$	112	**LysylPG**$^{\mathbf{CCS,RT}}$	18	HBMP	2
CE	109	NAE	13	NAGlySer	2
Cer$^{\mathbf{CCS,RT}}$	108	**AcylPE**$^{\mathbf{CCS,RT}}$	11	**PIP**$^{\mathbf{RT}}$	2
DGDG$^{\mathbf{CCS,RT}}$	107	SE	10	MLCL	1
PA$^{\mathbf{CCS,RT}}$	101	**LPA**	9	VAE	1

Bold entries have theoretical *m/z* values, superscripts reflect presence of theoretical CCS and/or HILIC retention time values. Four additional lipid classes are included in theoretical *m/z* database but not present as measured values: PIP2, PIP3, LCL, and HexCer.

5.6.2 Lipid Class Binary Encodings for CCS Prediction

lipid class	encoding	lipid class	encoding
AcylPE	10000000000000000000	PA	00000000000000010000000
AcylPG	01000000000000000000	PC	00000000000000001000000
CE	00100000000000000000	PE	00000000000000000100000
Cer	00010000000000000000	PG	00000000000000000010000
DG	00001000000000000000	PI	00000000000000000001000
DGDG	00000100000000000000	PS	00000000000000000000100
FA	00000010000000000000	SM	00000000000000000000010
GlcCer	00000001000000000000	TG	00000000000000000000001
LPC	00000000100000000000		
LPE	00000000010000000000		
LPG	00000000001000000000		
LPS	00000000000100000000		
LysylPG	00000000000010000000		
MGDG	00000000000001000000		

5.6.3 Fatty Acid Modifier Binary Encodings for CCS Prediction

FA modifier	encoding
d	100
o	010
p	001

5.6.4 MS Adduct Binary Encodings for CCS Prediction

MS adduct	encoding
$[M+2K]^{2+}$	100000000000
$[M+2Na-H]^{+}$	010000000000
$[M+CH_3COO]^{-}$	001000000000
$[M+Cl]^{-}$	000100000000
$[M+H-H_2O]^{+}$	000010000000
$[M+HCOO]^{-}$	000000100000
$[M+H]^{+}$	000000010000
$[M+NH_4]^{+}$	000000001000
$[M+Na]^{+}$	000000000100
$[M-2H]^{2-}$	000000000010
$[M-H]^{-}$	000000000001

5.6.5 Lipid Class Abbreviations

abbreviation	Full name	abbreviation	Full name
AcylPE	N-acyl-phosphatidylethanolamine	**LPI**	lysophosphatidylinositol
AcylPG	acyl-phosphatidylglycerol	**LPS**	lysophosphatidylserine
ADGGA	acyl-diacylglyceryl glucuronide	**LysylPG**	lysylphosphatidylglycerol
AlaPG	alanyl-phosphatidylglycerol	**MGDG**	monoglucosyldiacylglycerol
BMP	bismonoacylglycerophosphate	**MLCL**	monolysocardiolipin
CAR	acylcarnitine	**NAE**	N-acyl-ethanolamine
CE	cholesteryl ester	**NAGly**	N-acyl-glycine
Cer	ceramide	**NAGlySer**	N-acyl-glycyl-serine
CL	cardiolipin	**PA**	phosphatidic acid
DG	diacylglycerol	**PC**	phosphatidylcholine
DGDG	diglucosyldiacylglycerol	**PE**	phosphatidylethanolamine
DGGA	diacylglyceryl glucuronide	**PEtOH**	phosphatidylethanol
DLCL	dilysocardiolipin	**PG**	phosphatidylglycerol
FA	fatty acid	**PI**	phosphatidylinositol
FAHFA	fatty acyl ester of hydroxylated fatty acid	**PIP**	phosphatidylinositol-monophosphate

GlcADG	glucuronosyldiacylglycerol	**PIP2**	phosphatidylinositol-diphosphate
GlcCer	glucosylceramide	**PIP3**	phosphatidylinositol-triphosphate
HBMP	hemibismonoacylglycerophosphate	**PMeOH**	phosphatidylmethanol
HexCer	hexosylceramide	**PS**	phosphatidylserine
LacCer	lactosylceramide	**SE**	bile acid ester
LCL	lysocardiolipin	**SM**	sphingomyelin
LPA	lysophosphatidic acid	**SQDG**	sulfoquinovosyl diacylglycerol
LPC	lysophosphatidylcholine	**TG**	triacylglycerol
LPE	lysophosphatidylethanolamine		

5.6.6 Lipid Class Binary Encodings for HILIC Retention Time Prediction

lipid class	encoding	lipid class	encoding
AcylPE	100000000000000000000000	PC	000000000000000010000000
AcylPG	010000000000000000000000	PE	000000000000000001000000
AlaPG	001000000000000000000000	PE	000000000000000000100000
CL	000100000000000000000000	PG	000000000000000000010000
Cer	000010000000000000000000	PI	000000000000000000001000
DG	000001000000000000000000	PIP	000000000000000000000100
DGDG	000000100000000000000000	PS	000000000000000000000010
GlcADG	000000010000000000000000	SM	000000000000000000000001
GlcCer	000000001000000000000000		
LPC	000000000100000000000000		
LPE	000000000010000000000000		
LPG	000000000001000000000000		
LysylPG	000000000000100000000000		
MGDG	000000000000010000000000		
PA	000000000000001000000000		

Chapter 6 Conclusions, Perspectives and Future Directions

Confident identification of xenobiotic and endogenous metabolites is crucial for applications including metabolomics, lipidomics, and drug metabolism studies. Integration of IM with MS allows the acquisition of additional information from such studies without significantly decreasing their analytical throughput. Additionally, through the measurement of CCS, greater confidence in analyte identification can be realized. This confidence relies upon availability of reference CCS values to compare against. However, because of the large number and structural diversity of small molecules and the impracticability of obtaining experimental CCS values for all of them, it is necessary to be able to predict CCS with high accuracy. This dissertation explores applications of IM-MS, theory-based CCS calculation, and machine learning-based CCS-prediction for the determination of xenobiotic and endogenous metabolites.

Chapter 2 discussed the structural characteristics of a comprehensive CCS database assembled from the literature, explored how those characteristics mapped onto measured CCS, and employed a novel ML-based CCS prediction strategy to produce high-accuracy CCS predictions, *i.e*., untargeted clustering on this diverse dataset followed by specialized prediction models for each cluster. Chapter 3 examined the IM-MS characteristics of a panel of drugs and *in vitro*-generated metabolites and used in-depth computational modeling and theoretical CCS calculation to rationalize experimental observations. An important finding from this work was the relationship between structural characteristics of metabolites relative to their corresponding parent compounds: the IM characteristics of the metabolites were highly dependent upon the chemistry of the parent as well as the metabolic modification and could conform to or deviate from the gas-phase conformational characteristics of the parent. In Chapter 4, the *in vitro* drug metabolite generation and IM-MS analysis was scaled to a high-throughput format and applied to

a diverse collection of over 2000 drug and drug-like compounds in order to build a drug- and drug metabolite-specific CCS database for training a ML-based CCS prediction model specific to drugs and drug metabolites. This chapter demonstrates the effects of different featurization strategies on the accuracy of ML-based CCS prediction and presents the use of a novel set of 3D molecular descriptors for CCS prediction. By using these 3D molecular descriptors, differences in the gas-phase conformations of protomers or positional isomers can be captured in CCS prediction. Chapter 5 discussed the development of a bioinformatic tool for the analysis of lipidomics data, which includes specialized models for the prediction of CCS and HILIC retention time. This chapter demonstrated how specialized predictive models can be built for specific chemical classes that leverage class-specific structural trends (*i.e.* lipid fatty acid composition) to produce high-accuracy CCS predictions, in addition to the extensibility of this approach to chemical properties beyond CCS. Considered as a whole, this thesis represents an in-depth examination of CCS prediction using theory- and data-driven approaches in multiple contexts and demonstrates its utility in characterizing the structures of drug metabolites, identifying metabolites, and exploring phenotypes by metabolomics or lipidomics.

Importantly, the insights presented in this dissertation are generalizable and extensible beyond the context of CCS prediction. Computational modeling, multivariate analyses, leveraging large-scale datasets for ML, and various combinations thereof constitute a powerful framework for the interrogation and prediction of chemical characteristics, ranging from physiochemical properties to biological activity. When combined with the appropriate level of experimental data, such a framework can provide fundamental insights into chemical and biological problems and enable accurate prediction of properties at scales that would otherwise be intractable.

References

1. Han, X. Lipidomics for studying metabolism. *Nat Rev Endocrinol* **12**, 668-679 (2016).

2. Newgard, C.B. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell Metab* **25**, 43-56 (2017).

3. Karczewski, K.J. & Snyder, M.P. Integrative omics for health and disease. *Nat Rev Genet* **19**, 299-310 (2018).

4. Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **27**, 1897-1905 (2016).

5. Gallart-Ayala, H., Teav, T. & Ivanisevic, J. Metabolomics meets lipidomics: Assessing the small molecule component of metabolism. *Bioessays* **42**, e2000052 (2020).

6. Guijas, C., Montenegro-Burke, J.R., Warth, B., Spilker, M.E. & Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat Biotechnol* **36**, 316-320 (2018).

7. Kaddurah-Daouk, R., Kristal, B.S. & Weinshilboum, R.M. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol* **48**, 653-683 (2008).

8. Schymanski, E.L. et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **48**, 2097-2098 (2014).

9. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8** (2018).

10. Patti, G.J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* **13**, 263-269 (2012).

11. Emwas, A.H. et al. NMR Spectroscopy for Metabolomics Research. *Metabolites* **9** (2019).

12. Rampler, E. et al. Recurrent Topics in Mass Spectrometry-Based Metabolomics and Lipidomics-Standardization, Coverage, and Throughput. *Anal Chem* **93**, 519-545 (2021).

13. Paglia, G., Smith, A.J. & Astarita, G. Ion mobility mass spectrometry in the omics era: Challenges and opportunities for metabolomics and lipidomics. *Mass Spectrom Rev* (2021).

14. Luo, M.-D., Zhou, Z.-W. & Zhu, Z.-J. The Application of Ion Mobility-Mass Spectrometry in Untargeted Metabolomics: from Separation to Identification. *Journal of Analysis and Testing*, 1-13 (2020).

15. May, J.C., Goodwin, C.R. & McLean, J.A. Ion mobility-mass spectrometry strategies for untargeted systems, synthetic, and chemical biology. *Curr Opin Biotechnol* **31**, 117-121 (2015).

16. Silverman, R.B. & Holladay, M.W. The Organic Chemistry of Drug Design and Drug Action. (Elsevier, 2014).

17. Zanger, U.M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics* **138**, 103-141 (2013).

18. Board, P.G. & Menon, D. Glutathione transferases, regulators of cellular metabolism and physiology. *BBA - General Subjects* **1830**, 3267-3288 (2013).

19. Tukey, R.H. & Strassburg, C.P. Human UDP-Glucuronosyltransferases: Metabolism, Expression, and Disease. *Annual Review of Pharmacology and Toxicology* **40**, 581-616 (2000).

20. Fura, A. et al. Discovering Drugs through Biological Transformation: Role of Pharmacologically Active Metabolites in Drug Discovery. *Journal of Medicinal Chemistry* **47**, 4339-4351 (2004).

21. Park, B.K., Kitteringham, N.R., Maggs, J.L., Pirmohamed, M. & Williams, D.P. The Role of Metabolic Activation in Drug-Induced Hepatotoxicity. *Annual Review of Pharmacology and Toxicology* **45**, 177-202 (2005).

22. Prakash, C., Shaffer, C.L. & Nedderman, A. Analytical strategies for identifying drug metabolites. *Mass Spectrometry Reviews* **26**, 340-369 (2007).

23. Zhu, M., Zhang, H. & Humphreys, W.G. Drug metabolite profiling and identification by high-resolution mass spectrometry. *J Biol Chem* **286**, 25419-25425 (2011).

24. Wen, B. & Zhu, M. Applications of mass spectrometry in drug metabolism: 50 years of progress. *Drug Metabolism Reviews* **47**, 71-87 (2015).

25. Zhang, H., Zhang, D., Ray, K. & Zhu, M. Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J Mass Spectrom* **44**, 999-1016 (2009).

26. Prasad, B., Garg, A., Takwani, H. & Singh, S. Metabolite identification by liquid chromatography-mass spectrometry. *Trends in Analytical Chemistry* **30**, 360-387 (2011).

27. Cuyckens, F. Mass spectrometry in drug metabolism and pharmacokinetics: Current trends and future perspectives. *Rapid Communications in Mass Spectrometry* **33**, 90-95 (2019).

28. Campuzano, I.D.G. & Lippens, J.L. Ion mobility in the pharmaceutical industry: an established biophysical technique or still niche? *Current Opinion in Chemical Biology* **42**, 147-159 (2018).

29. Clemmer, D.E., Hudgins, R.R. & Jarrold, M.F. Naked Protein Conformations: Cytochrome c in the Gas Phase. *Journal of the American Chemical Society* **117**, 10141-10142 (1995).

30. McLean, J.A., Ruotolo, B.T., Gillig, K.J. & Russell, D.H. Ion mobility-mass spectrometry: a new paradigm for proteomics. *International Journal of Mass Spectrometry* **240**, 301-315 (2005).

31. Kanu, A.B., Dwivedi, P., Tam, M., Matz, L. & Hill Jr, H.H. Ion mobility-mass spectrometry. *Journal of Mass Spectrometry* **43**, 1-22 (2008).

32. Fenn, L.S., Kliman, M., Mahsut, A., Zhao, S.R. & McLean, J.A. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Analytical and bioanalytical chemistry* **394**, 235-244 (2009).

33. Hines, K.M., May, J.C., McLean, J.A. & Xu, L. Evaluation of Collision Cross Section Calibrants for Structural Analysis of Lipids by Traveling Wave Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **88**, 7329-7336 (2016).

34. Zhou, Z., Shen, X., Tu, J. & Zhu, Z.-J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **88**, 11084-11091 (2016).

35. Stow, S.M. et al. An Interlaboratory Evaluation of Drift Tube Ion Mobility–Mass Spectrometry Collision Cross Section Measurements. *Analytical Chemistry* **89**, 9048-9055 (2017).

36. Hernández-Mesa, M. et al. Interlaboratory and Interplatform Study of Steroids Collision Cross Section by Traveling Wave Ion Mobility Spectrometry. *Analytical Chemistry* **92**, 5013-5022 (2020).

37. May, J.C. & McLean, J.A. Ion Mobility-Mass Spectrometry: Time-Dispersive Instrumentation. *Analytical Chemistry* **87**, 1422-1436 (2015).

38. Gabelica, V. et al. Recommendations for reporting ion mobility Mass Spectrometry measurements. *Mass Spectrometry Reviews* **38**, 291-320 (2019).

39. Dodds, J.N. & Baker, E.S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *Journal of the American Society for Mass Spectrometry* **42**, 392-311 (2019).

40. Vasilopoulou, C.G. et al. Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nature Communications*, 1-11 (2020).

41. Tsugawa, H. et al. MS-DIAL 4: accelerating lipidomics using an MS/MS, CCS, and retention time atlas. *bioRxiv* **37**, 513 (2020).

42. Chai, M., Young, M.N., Liu, F.C. & Bleiholder, C. A Transferable, Sample-Independent Calibration Procedure for Trapped Ion Mobility Spectrometry (TIMS). *Anal Chem* **90**, 9040-9047 (2018).

43. Shvartsburg, A.A., Danielson, W.F. & Smith, R.D. High-Resolution Differential Ion Mobility Separations Using Helium-Rich Gases. *Analytical Chemistry* **82**, 2456-2462 (2010).

44. Santiago, B.G., Harris, R.A., Isenberg, S.L. & Glish, G.L. Resolving powers of >7900 using linked scans: how well does resolving power describe the separation capability of differential ion mobility spectrometry. *The Analyst* **140**, 6871-6878 (2015).

45. May, J.C. et al. Conformational Ordering of Biomolecules in the Gas Phase: Nitrogen Collision Cross Sections Measured on a Prototype High Resolution Drift Tube Ion Mobility-Mass Spectrometer. *Analytical Chemistry* **86**, 2107-2116 (2014).

46. Picache, J.A. et al. Collision cross section compendium to annotate and predict multi-omic compound identities. *Chem. Sci.* **10**, 983-993 (2019).

47. Hines, K.M., Ross, D.H., Davidson, K.L., Bush, M.F. & Xu, L. Large-Scale Structural Characterization of Drug and Drug-Like Compounds by High-Throughput Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **89**, 9023-9030 (2017).

48. Lian, R. et al. Ion mobility derived collision cross section as an additional measure to support the rapid analysis of abused drugs and toxic compounds using electrospray ion mobility time-of-flight mass spectrometry. *Analytical Methods* **10**, 749-756 (2018).

49. Tejada-Casado, C. et al. Collision cross section (CCS) as a complementary parameter to characterize human and veterinary drugs. *Analytica Chimica Acta* **1043**, 52-63 (2018).

50. Leaptrot, K.L., May, J.C., Dodds, J.N. & McLean, J.A. Ion mobility conformational lipid atlas for high confidence lipidomics. *Nature Communications*, 1-9 (2019).

51. Kaufmann, A. et al. Are liquid chromatography/electrospray tandem quadrupole fragmentation ratios unequivocal confirmation criteria? *Rapid Communications in Mass Spectrometry* **23**, 985-998 (2009).

52. Lapthorn, C., Dines, T.J., Chowdhry, B.Z., Perkins, G.L. & Pullen, F.S. Can ion mobility mass spectrometry and density functional theory help elucidate protonation sites in 'small' molecules? *Rapid Communications in Mass Spectrometry* **27**, 2399-2410 (2013).

53. McCullagh, M., Giles, K., Richardson, K., Stead, S. & Palmer, M. Investigations into the performance of travelling wave enabled conventional and cyclic ion mobility systems to characterise protomers of fluoroquinolone antibiotic residues. *Rapid Commun Mass Spectrom* **33 Suppl 2**, 11-21 (2019).

54. Warnke, S. et al. Protomers of Benzocaine: Solvent and Permittivity Dependence. *Journal of the American Chemical Society* **137**, 4236-4242 (2015).

55. Boschmans, J. et al. Combining density functional theory (DFT) and collision cross-section (CCS) calculations to analyze the gas-phase behaviour of small molecules and their protonation site isomers. *Analyst* **141**, 4044-4054 (2016).

56. Aminpour, M., Montemagno, C. & Tuszynski, J.A. An Overview of Molecular Modeling for Drug Discovery with Specific Illustrative Examples of Applications. *Molecules* **24**, 1693-1630 (2019).

57. von Helden, G., Hsu, M.T., Gotts, N. & Bowers, M.T. Carbon cluster cations with up to 84 atoms: structures, formation mechanism, and reactivity. *The Journal of Physical Chemistry* **97**, 8182-8192 (1993).

58. Shvartsburg, A.A. & Jarrold, M.F. An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chemical Physics Letters* **261**, 86-91 (1996).

59. Mesleh, M.F., Hunter, J.M., Shvartsburg, A.A., Schatz, G.C. & Jarrold, M.F. Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential. *The Journal of Physical Chemistry* **100**, 16082-16086 (1996).

60. Campuzano, I. et al. Structural Characterization of Drug-like Compounds by Ion Mobility Mass Spectrometry: Comparison of Theoretical and Experimentally Derived Nitrogen Collision Cross Sections. *Analytical Chemistry* **84**, 1026-1033 (2012).

61. Bleiholder, C., Wyttenbach, T. & Bowers, M.T. A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (I). Method. *International Journal of Mass Spectrometry* **308**, 1-10 (2011).

62. Larriba, C. & Hogan Jr, C.J. Ion Mobilities in Diatomic Gases: Measurement versus Prediction with Non-Specular Scattering Models. *The Journal of Physical Chemistry A* **117**, 3887-3901 (2013).

63. Larriba-Andaluz, C. & Hogan Jr, C.J. Collision cross section calculations for polyatomic ions considering rotating diatomic/linear gas molecules. *The Journal of Chemical Physics* **141**, 194107-194110 (2014).

64. Ewing, S.A., Donor, M.T., Wilson, J.W. & Prell, J.S. Collidoscope: an improved tool for computing collisional cross-sections with the trajectory method. *Journal of the American Society of Mass Spectrometry* **28**, 587-596 (2017).

65. Marklund, E.G., Degiacomi, M.T., Robinson, C.V., Baldwin, A.J. & Benesch, J.L.P. Collision Cross Sections for Structural Proteomics. *Structure/Folding and Design* **23**, 791-799 (2015).

66. Zanotto, L., Heerdt, G., Souza, P.C.T., Araujo, G. & Skaf, M.S. High performance collision cross section calculation-HPCCS. *Journal of Computational Chemistry* **39**, 1675-1681 (2018).

67. Ieritano, C., Crouse, J., Campbell, J.L. & Hopkins, W.S. A parallelized molecular collision cross section package with optimized accuracy and efficiency. *The Analyst* **144**, 1660-1670 (2019).

68. Zhou, Z., Tu, J., Xiong, X., Shen, X. & Zhu, Z.-J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics. *Analytical Chemistry* **89**, 9559-9566 (2017).

69. Soper-Hopper, M.T. et al. Collision cross section predictions using 2-dimensional molecular descriptors. *Chem. Commun.* **53**, 7624-7627 (2017).

70. Bijlsma, L. et al. Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Anal Chem* **89**, 6583-6589 (2017).

71. Mollerup, C.B., Mardal, M., Dalsgaard, P.W., Linnet, K. & Barron, L.P. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *Journal of Chromatography A* **1542**, 82-88 (2018).

72. Plante, P.-L. et al. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Analytical Chemistry*, 1-9 (2019).

73. Zhou, Z. et al. Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics. *Nature Communications*, 1-13 (2020).

74. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8**, 31-23 (2018).

75. von Helden, G., Wyttenbach, T. & Bowers, M.T. Conformation of Macromolecules in the Gas Phase: Use of Matrix-Assisted Laser Desorption Methods in Ion Chromatography. *Science* **267**, 1483-1485 (1995).

76. McLean, J.A., Ruotolo, B.T., Gillig, K.J. & Russell, D.H. Ion mobility–mass spectrometry: a new paradigm for proteomics. *International Journal of Mass Spectrometry* **240**, 301-315 (2005).

77. Pringle, S.D. et al. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *International Journal of Mass Spectrometry* **261**, 1-12 (2007).

78. Hinnenkamp, V. et al. Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. *Analytical Chemistry* **90**, 12042-12050 (2018).

79. Kim, H.I. et al. Structural Characterization of Unsaturated Phosphatidylcholines Using Traveling Wave Ion Mobility Spectrometry. *Analytical Chemistry* **81**, 8289-8297 (2009).

80. Kim, H. et al. Experimental and Theoretical Investigation into the Correlation between Mass and Ion Mobility for Choline and Other Ammonium Cations in N 2. *Analytical Chemistry* **80**, 1928-1936 (2008).

81. Lee, J.W., Lee, H.H.L., Davidson, K.L., Bush, M.F. & Kim, H.I. Structural characterization of small molecular ions by ion mobility mass spectrometry in nitrogen drift gas: improving the accuracy of trajectory method calculations. *The Analyst* **143**, 1786-1796 (2018).

82. Colby, S.M., Nuñez, J.R., Hodas, N.O., Corley, C.D. & Renslow, R.R. Deep Learning to Generate in SilicoChemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples. *Analytical Chemistry* **92**, 1720-1729 (2020).

83. Bijlsma, L. et al. Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Analytical Chemistry* **89**, 6583-6589 (2017).

84. Reymond, J.-L. & Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chemical Neuroscience* **3**, 649-657 (2012).

85. Hines, K.M., Herron, J. & Xu, L. Assessment of altered lipid homeostasis by HILIC-ion mobility-mass spectrometry-based lipidomics. *The Journal of Lipid Research* **58**, 809-819 (2017).

86. Hines, K.M. et al. Characterization of the Mechanisms of Daptomycin Resistance among Gram-Positive Bacterial Pathogens by Multidimensional Lipidomics. *mSphere* **2**, 99-16 (2017).

87. Hines, K.M. & Xu, L. Lipidomic consequences of phospholipid synthesis defects in Escherichia coli revealed by HILIC-ion mobility-mass spectrometry. *Chemistry and Physics of Lipids* **219**, 15-22 (2019).

88. Blaženović, I. et al. Increasing Compound Identification Rates in Untargeted Lipidomics Research with Liquid Chromatography Drift Time–Ion Mobility Mass Spectrometry. *Analytical Chemistry* **90**, 10758-10764 (2018).

89. Paglia, G. et al. Ion Mobility Derived Collision Cross Sections to Support Metabolomics Applications. *Analytical Chemistry* **86**, 3985-3993 (2014).

90. Zheng, X. et al. A structural examination and collision cross section database for over 500 metabolites and xenobiotics using drift tube ion mobility spectrometry. *Chem. Sci.* **8**, 7724-7736 (2017).

91. Nichols, C.M. et al. Untargeted Molecular Discovery in Primary Metabolism: Collision Cross Section as a Molecular Descriptor in Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **90**, 14484-14492 (2018).

92. Righetti, L. et al. Ion mobility-derived collision cross section database: Application to mycotoxin analysis. *Analytica Chimica Acta* **1014**, 50-57 (2018).

93. Hernández-Mesa, M., Le Bizec, B., Monteau, F., García-Campaña, A.M. & Dervilly-Pinel, G. Collision Cross Section (CCS) database: An additional measure to characterize steroids. *Analytical Chemistry* **90**, 4616-4625 (2018).

94. Sumner, L.W. et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211-221 (2007).

95. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

96. Shimizu, A., Ohe, T. & Chiba, M. A Novel Method for the Determination of the Site of Glucuronidation by Ion Mobility Spectrometry-Mass Spectrometry. *Drug Metabolism and Disposition* **40**, 1456-1459 (2012).

97. Shimizu, A. & Chiba, M. Ion Mobility Spectrometry–Mass Spectrometry Analysis for the Site of Aromatic Hydroxylation. *Drug Metabolism and Disposition* **41**, 1295-1299 (2013).

98. Reading, E. et al. Elucidation of Drug Metabolite Structural Isomers Using Molecular Modeling Coupled with Ion Mobility Mass Spectrometry. *Analytical Chemistry* **88**, 2273-2280 (2016).

99. Klieber, S. et al. Contribution of the N-glucuronidation pathway to the overall in vitro metabolic clearance of midazolam in humans. *Drug metabolism and disposition: the biological fate of chemicals* **36**, 851-862 (2008).

100. Walsky, R.L. et al. Optimized assays for human UDP-glucuronosyltransferase (UGT) activities: altered alamethicin concentration and utility to screen for UGT inhibitors. *Drug Metab Dispos* **40**, 1051-1065 (2012).

101. Hoffmann, W. et al. An Intrinsic Hydrophobicity Scale for Amino Acids and Its Application to Fluorinated Compounds. *Angew Chem Int Ed Engl* **58**, 8216-8220 (2019).

102. Hardwick, J.P. Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases. *Biochem Pharmacol* **75**, 2263-2275 (2008).

103. Seguin, R.P., Herron, J.M., Lopez, V.A., Dempsey, J.L. & Xu, L. Metabolism of Benzalkonium Chlorides by Human Hepatic Cytochromes P450. *Chem Res Toxicol* **32**, 2466-2478 (2019).

104. Ling, K.H. et al. Metabolism of terfenadine associated with CYP3A (4) activity in human hepatic microsomes. *Drug Metabolism and Disposition* **23**, 631-636 (1995).

105. Kerry, N.L., Somogyi, A.A., Bochner, F. & Mikus, G. The role of CYP2D6 in primary and secondary oxidative metabolism of dextromethorphan: in vitro studies using human liver microsomes. *Br J Clin Pharmacol* **38**, 243-248 (1994).

106. Lalli, P.M. et al. Protomers: formation, separation and characterization via travelling wave ion mobility mass spectrometry. *J Mass Spectrom* **47**, 712-719 (2012).

107. Boersma, M.G. et al. Regioselectivity of phase II metabolism of luteolin and quercetin by UDP-glucuronosyl transferases. *Chem Res Toxicol* **15**, 662-670 (2002).

108. Kuca, K. et al. Preparation of benzalkonium salts differing in the length of a side alkyl chain. *Molecules* **12**, 2341-2347 (2007).

109. Ruotolo, B.T., Benesch, J.L., Sandercock, A.M., Hyung, S.J. & Robinson, C.V. Ion mobility-mass spectrometry analysis of large protein complexes. *Nat Protoc* **3**, 1139-1152 (2008).

110. Forsythe, J.G. et al. Collision cross section calibrants for negative ion mode traveling wave ion mobility-mass spectrometry. *Analyst* **140**, 6853-6861 (2015).

111. O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).

112. Halgren, T.A. Merck molecular force field: I-V. *Journal of Computational Chemistry* **17**, 490-641 (1996).

113. Stewart, J.J. MOPAC: a semiempirical molecular orbital program. *J Comput Aided Mol Des* **4**, 1-105 (1990).

114. Abraham, M.J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015).

115. Zhu, Y. et al. Amlodipine Metabolism in Human Liver Microsomes and Roles of CYP3A4/5 in the Dihydropyridine Dehydrogenation. *Drug metabolism and disposition: the biological fate of chemicals* **42**, 245-249 (2013).

116. Jefferson, J.W., Pradko, J.F. & Muir, K.T. Bupropion for major depressive disorder: Pharmacokinetic and formulation considerations. *Clinical Therapeutics* **27**, 1685-1695 (2005).

117. Hartmann, F., Gruenke, L.D., Craig, J.C. & Bissell, D.M. Chlorpromazine metabolism in extracts of liver and small intestine from guinea pig and from man. *Drug Metabolism and Disposition* **11**, 244-248 (1983).

118. Mürdter, T.E. et al. Genetic polymorphism of cytochrome P450 2D6 determines oestrogen receptor activity of the major infertility drug clomiphene via its active metabolites. *Human Molecular Genetics* **21**, 1145-1154 (2011).

119. Clarke, T.A. & Waskell, L.A. The Metabolism of Clopidogrel Is Catalyzed by Human Cytochrome P450 3A and Is Inhibited by Atorvastatin. *Drug Metabolism and Disposition* **31**, 53-59 (2003).

120. Pirmohamed, M., Williams, D., Madden, S., Templeton, E. & Park, B.K. Metabolism and bioactivation of clozapine by human liver in vitro. *Journal of Pharmacology and Experimental Therapeutics* **272**, 984-990 (1995).

121. Kerry, N.L., Somogyi, A.A. & clinical, F.B.B.j.o. The role of CYP2D6 in primary and secondary oxidative metabolism of dextromethorphan: in vitro studies using human liver microsomes. *British Journal of Clinial Pharmacology* **38**, 243-248 (1994).

122. Kuehl, G.E., Lampe, J.W., Potter, J.D. & disposition, J.B.D.m.a. Glucuronidation of nonsteroidal anti-inflammatory drugs: identifying the enzymes responsible in human liver microsomes. *ASPET*.

123. Bort, R. et al. Hepatic metabolism of diclofenac: role of human CYP in the minor oxidative pathways. *Biochemical Pharmacology* **58**, 787-796 (1999).

124. Aufrère, M.B. & Benson, H. Progesterone: An overview and recent advances. *Journal of Pharmaceutical Sciences* **65**, 783-800 (1976).

125. Nielsen, T.L., Rasmussen, B.B., Flinois, J.-P., Beaune, P. & Brøsen, K. In Vitro Metabolism of Quinidine: The (3S)-3-Hydroxylation of Quinidine Is a Specific Marker Reaction for Cytochrome P-4503A4 Activity in Human Liver Microsomes. *Journal of Pharmacology and Experimental Therapeutics* **289**, 31-37 (1999).

126. Jurima-Romet, M., Crawford, K., Cyr, T. & Inaba, T. Terfenadine metabolism in human liver. In vitro inhibition by macrolide antibiotics and azole antifungals. *Drug Metabolism and Disposition* **22**, 849-857 (1994).

127. Eap, C.B. et al. Plasma levels of the enantiomers of thioridazine, thioridazine 2-sulfoxide, thioridazine 2-sulfone, and thioridazine 5-sulfoxide in poor and extensive metabolizers of dextromethorphan and mephenytoin. *Clinical Pharmacology & Therapeutics* **59**, 322-331 (1996).

128. Wu, Y. et al. Cytochrome P450-mediated metabolism of triclosan attenuates its cytotoxicity in hepatic cells. *Archives of Toxicology* **91**, 2405-2423 (2016).

129. Ross, D.H. & Xu, L. Determination of drugs and drug metabolites by ion mobility-mass spectrometry: A review. *Analytica Chimica Acta* **1154** (2021).

130. Ross, D.H., Cho, J.H. & Xu, L. Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Analytical Chemistry* **92**, 4548-4557 (2020).

131. Djoumbou-Feunang, Y. et al. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform* **11**, 2 (2019).

132. Wishart, D.S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668-672 (2006).

133. Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* **8**, 3 (2016).

134. Ross, D.H., Seguin, R.P. & Xu, L. Characterization of the Impact of Drug Metabolism on the Gas-Phase Structures of Drugs Using Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **91**, 14498-14507 (2019).

135. Seguin, R., Herron, J., Dempsey, J., Lopez, V. & Xu, L. Metabolism of Benzalkonium Chlorides by Human Hepatic Cytochromes P450. Under revision. (2019).

136. Boersma, M.G. et al. Regioselectivity of Phase II Metabolism of Luteolin and Quercetin by UDP-Glucuronosyl Transferases. *Chemical Research in Toxicology* **15**, 662-670 (2002).

137. O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *Journal of Chemoinformatics* **3**, 33 (2011).

138. Frisch, M.J. et al. Gaussian 16 Rev. C.01. *Gaussian 16* (2016).

139. Harayama, T. & Riezman, H. Understanding the diversity of membrane lipid composition. *Nature Reviews Molecular Cell Biology* **19**, 281-296 (2018).

140. Wymann, M.P. & Schneiter, R. Lipid signalling in disease. *Nat Rev Mol Cell Biol* **9**, 162-176 (2008).

141. Tumanov, S. & Kamphorst, J.J. Recent advances in expanding the coverage of the lipidome. *Current Opinion in Biotechnology* **43**, 127-133 (2017).

142. Rustam, Y.H. & Reid, G.E. Analytical Challenges and Recent Advances in Mass Spectrometry Based Lipidomics. *Analytical Chemistry* **90**, 374-397 (2017).

143. Tu, J., Zhou, Z., Li, T. & Zhu, Z.-J. The emerging role of ion mobility-mass spectrometry in lipidomics to facilitate lipid separation and identification. *TrAC Trends in Analytical Chemistry* **116**, 332-339 (2019).

144. Fahy, E. et al. Update of the LIPID MAPS comprehensive classification system for lipids. *The Journal of Lipid Research* **50**, S9-S14 (2009).

145. Fahy, E. et al. A comprehensive classification system for lipids. *J Lipid Res* **46**, 839-861 (2005).

146. Liebisch, G. et al. Shorthand notation for lipid structures derived from mass spectrometry. *J Lipid Res* **54**, 1523-1530 (2013).

147. Ryan, E. & Reid, G.E. Chemical Derivatization and Ultrahigh Resolution and Accurate Mass Spectrometry Strategies for "Shotgun" Lipidome Analysis. *Accounts of Chemical Research* **49**, 1596-1604 (2016).

148. Kyle, J.E. et al. Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry. *Analyst* **141**, 1649-1659 (2016).

149. Zhou, Z. et al. LipidIMMS Analyzer: integrating multi-dimensional information to support lipid identification in ion mobility—mass spectrometry based lipidomics. *Bioinformatics* **35**, 698-700 (2018).

150. Hines, K.M. et al. Occurrence of cross-resistance and β-lactam seesaw effect in glycopeptide-, lipopeptide- and lipoglycopeptide-resistant MRSA correlates with membrane phosphatidylglycerol levels. *Journal of Antimicrobial Chemotherapy* **42**, 2398-2395 (2020).

151. Ulmer, C.Z., Koelmel, J.P., Ragland, J.M., Garrett, T.J. & Bowden, J.A. LipidPioneer : A Comprehensive User-Generated Exact Mass Template for Lipidomics. 1-4 (2017).

152. Virtanen, P. et al. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. *arXiv.org* **cs.MS** (2019).