

Method Development for High-Coverage Metabolome Analysis

by

Hao Li

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry
University of Alberta

© Hao Li, 2022

Abstract

Metabolomics, as an interdisciplinary research area, requires the integration of knowledge from different disciplines, such as analytical chemistry, bioinformatics, and statistics. The general workflow of metabolome analysis includes sample preparation, data acquisition, data processing, data analysis, and metabolite identification. The improvement on any step listed above can affect the outcome in metabolomics studies. In the experimental section, tremendous efforts have been made to expand the metabolome coverage and produce high-quality data. The bottleneck gradually shifts to the section of data interpretation, such as metabolite identification, especially for unknown compounds, and appropriate methods used in data analysis.

In response to these challenges, my thesis research focuses on the improvement of the data interpretation section. In the first part, the coverage of the combination of multi-channel chemical isotope labeling (CIL) methods was evaluated (Chapter 2). Metabolite information from current metabolomics databases was extracted. Based on the functional groups, metabolites were further classified into four sub-metabolomes corresponding to four CIL channels, including amine/phenol channel, carboxylic channel, carbonyl channel, and hydroxyl channel. From the perspective of chemical functional groups or chemical space, near-complete metabolome coverage could be achieved using the integration of the four sub-metabolomes.

The second part targeted putative metabolite identification. Instead of building in-house libraries through data acquisition of metabolite standards, *in silico* prediction using existing data was employed. In Chapter 3, to refine the tripeptide identification via exact mass match only, the retention time (RT) of chemical isotope labeled tripeptides was predicted based on the RT of labeled dipeptides. In Chapter 4, MCID 2.0, an evidence-based metabolome library, was constructed using 76 biological reactions. To facilitate the identification of unknown metabolites,

theoretical metabolites were predicted based on the metabolites from the KEGG compounds database.

Lastly, a biomarker discovery study on spinal cord injury was conducted using serum samples from human clinical trials, aiming to differentiate different severity grades and predict neurological conversion as well as motor function recovery. Issues of human samples, such as imbalanced sample size from different groups, wide age range, and male-biased sex ratios, were required to be solved before statistical analysis. Support vector machine models were built to discover potential biomarkers.

Preface

A version of Chapter 2 was published as: Shuang Zhao[†], Hao Li[†], Wei Han, Wan Chan, and Liang Li, 2019, “Metabolomic Coverage of Chemical-Group-Submetabolome Analysis: Group Classification and Four-Channel Chemical Isotope Labeling LC-MS”, *Anal. Chem.*, 91, 12108-12115. I was responsible for the analysis of theoretical coverage of each sub-metabolome. Dr. Shuang Zhao carried out the experimental coverage of each sub-metabolome as well as the manuscript writing. Wei Han and Wan Chan contributed towards experimental data acquisition and analysis. Professor Liang Li supervised the project and edited the manuscript.

Chapter 3 was finished by me and Zhan Cheng. I was responsible for the retention time prediction and identification against predicted library. Zhan Cheng acquired the experimental data of dipeptides, tripeptides, and biological samples.

Chapter 4 was finished by me and Zhan Cheng. I was responsible for theoretical metabolite prediction and building the website. Zhen Cheng acquired the experimental data of biological samples.

Chapter 5 was collaborated with Dr. Brian Kwon at the University of British Columbia. I was responsible for the data analysis and identification. Xinyun Gu acquired the experimental data of the discovery cohort. Minglei Zhu acquired the validation cohort and processed the data. Dr. Brian Kwon collected the samples.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Liang Li. It is a great pleasure to engage in the data analysis of metabolomics under his guidance. His invaluable suggestions, especially from the whole perspective, and great support are indispensable in my research. I derived great experience and valuable knowledge from the discussion with Dr. Li.

I also really appreciate my supervisory committee, Dr. James Harynuk and Dr. Guohui Lin, and examining committee, Dr. John C. Vederas and Dr. Mariusz Klobukowski, as well as the external examiner, Dr. Jeff Xia of McGill University, for reviewing my thesis.

I am thankful to my collaborator, Dr. Brian Kwon of the University of British Columbia, for the precious samples and great advice on the data analysis.

I would like to thank the members of our laboratory. It is awesome to work with you. My thanks go to Dr. Shuang Zhao, Xinyu Gu, Zhan Cheng, Dr. Wan Chen, Xiaohang Wang, Dr. Xian Luo, Dorothea Mung, Dr. Wei Han, Dr. Yunong Li, Yingwen Wang, Chufan Wang, Kevin Hooton, Dr. Adriana Zardini Buzatto. Special thanks to Dr. Monica Li for the helpful suggestion on my thesis.

Last but not least, I cannot miss the moment to express my genuine appreciation to my family and friends. I thank my parents, Shizhen Zhao and Xingwu Li, for letting me experience the life I am curious about and supporting my decision. I thank my little fuzzy roommates. Your companion is the cure of the soul. Also, I thank my friends for the little and big moments we shared. One of the nicest things I have is the friendship with you.

Table of Contents

List of Figures	X
List of Tables	xiii
List of Abbreviations	xv
List of Symbols	xvii
Chapter 1 Introduction	1
1.1 Introduction of metabolomics	1
1.2 Current technology	2
1.3 The workflow of LC-MS-based untargeted metabolomics	6
1.3.1 Sample preparation.....	6
1.3.1.1 Label-free methods.....	7
1.3.1.2 Chemical derivatization methods	9
1.3.2. Data acquisition.....	10
1.3.3. Data processing	13
1.3.4 Data analysis	17
1.3.5 Metabolite identification	18
1.3.5.1. Databases for metabolite identification.....	19
1.3.6. Biological interpretation.....	21
1.4 Scope of the Thesis	22
Chapter 2 Metabolomic coverage of chemical-group-submetabolome analysis: group classification and 4-channel chemical isotope labeling LC-MS	23
2.1 Introduction	23
2.2 Experimental Section	25
2.2.1 Chemical group classification	25
2.2.2 4-Channel labeling	28
2.2.3 LC-MS analysis.....	30

2.2.4 Data processing and metabolite identification	30
2.3 Results and Discussion	31
2.3.1 Group classification of database entries	31
2.3.2 4-Channel labeling LC-MS results.....	35
2.3.3 Overlaps of multi-functional metabolites.....	41
2.3.4 Group under-representation.....	43
2.4 Conclusions	44
Chapter 3 Retention time prediction of chemical isotope labeled tripeptides on RPLC using machine learning methods.....	46
3.1 Introduction	46
3.2 Methods	47
3.2.1 Chemicals and Reagents.....	47
3.2.2 Sample preparation and derivatization	48
3.2.3 In-house database of dipeptides	48
3.2.4 HPLC-MS condition	49
3.2.5 Data processing	49
3.2.6 SMILES generation.....	50
3.2.7 Molecular descriptor calculation and filter	52
3.2.8 Building SVR radial models	53
3.3 Results and Discussion	54
3.3.1 Overview of dipeptide experimental data and SMILES, tripeptides SMILES	54
3.3.2 Comparison of experimental RT and predicted RT for the training set and validation set	54
3.3.3 Comparison between the experimental RT and predicted RT of 10 tripeptide standards.	61

3.3.4 Comparison between the experimental RT and predicted RT using tripeptide mixture data	62
3.3.5 Distinguish in-source fragmentation	64
3.3.6 Identification against serum and urine samples.	65
3.4 Conclusions	66
Chapter 4 MyCompoundID 2.0: evidence-based metabolome library facilitating CIL identification	68
4.1 Introduction	68
4.2 Methods	69
4.2.1 MCID 2.0 database construction.	69
4.2.2 Web interface.	71
4.2.3 Materials.....	74
4.2.4 Sample preparation and derivatization reaction	74
4.2.5 HPLC-MS condition	75
4.2.6 Data processing	76
4.3 Results and Discussion	76
4.4 Conclusions	83
Chapter 5 Biomarker discovery on spinal cord injury using chemical isotope labeling profiling.....	85
5.1 Introduction	85
5.2 Methods	87
5.2.1 Chemicals and reagents	87
5.2.2 Sample collection	87
5.2.3 Sample preparation.....	88
5.2.4 LC-UV methods	88

5.2.5 LC-MS methods	89
5.2.6 Data processing	89
5.2.7 Data analysis	90
5.3 Results and Discussion	91
5.3.1 Demographic information of participants	91
5.3.2 Metabolomics results.....	96
5.3.3 Classifying baseline AIS grades.....	98
5.3.4 Predicting AIS grade conversion in those with AIS A SCI	104
5.3.5 Predicting Outcome at 6 Months.....	107
5.4 Conclusions	110
Chapter 6 Conclusions and Future Work.....	111
6.1 Thesis Summary	111
6.2 Future Work	113
Bibliography	114

List of Figures

Figure 1.1 Workflow of LC-MS based untargeted metabolomics.	6
Figure 1.2 Using dansyl chloride to label primary, secondary amines and phenols.	10
Figure 1.3 The demonstration of Q-TOF.	12
Figure 2.1 The workflow of metabolite classification.	26
Figure 2.2 The workflow for metabolome analysis using 4-channel CIL LC-MS.	29
Figure 2.3 Classification of chemical groups of (A) MCID zero-reaction library, (C) HMDB, (E) KEGG, (G) YMDB and (I) ECMDB. Sequential class-elimination approach was used to determine the remaining groups (i.e., after removing all the 4-channel metabolites, a small number of the remaining metabolites contain the ester group. After removing 4-channel metabolites and ester-containing metabolites, a few remaining metabolites contain the amide group). Percent distributions of metabolites belonging to the four channels including overlapped metabolites with two or more functional groups in (B) MCID, (D) HMDB, (F) KEGG, (H) YMDB and (J) ECMDB.	35
Figure 2.4 (A) Percentage of peak pair detected in 4-channel LC-MS analysis of plasma as a function of peak intensity. (B) Venn diagram of the numbers of peak pairs detected in four channels.	36
Figure 2.5 Venn diagram of the numbers of peak pairs detected in four channels in yeast samples.	39
Figure 2.6 Distributions of peak pair numbers as a function of (A) averaged peak ratio and (B) RSD. Data are presented as mean \pm S.D. from experimental triplicate and injection duplicate (n=6).	41
Figure 2.7 Venn diagram of the numbers of metabolites in four channels from the compound entries in (A) MCID, (B) HMDB, (C) KEGG, (D) YMDB and (E) ECMDB.	43
Figure 3.1 Workflow of dipeptides and tripeptides RT prediction.	51
Figure 3.2 Structure of unlabeled and labeled dipeptides.	52

Figure 3.3 a) The optimization result of models built by top 20 features generated by labeled SMILES. b) The optimization result of models built by top 20 features generated by unlabeled SMILES. The smaller dot size represents a smaller RMSE.....	56
Figure 3.4 a) The comparison of experimental RT and predicted RT of dipeptide training set using labeled SMILES. b) The histogram of RT difference in the training set using labeled SMILES. c) The comparison of experimental RT and predicted RT of dipeptide test set using labeled SMILES. d) The histogram of RT difference in the test set using labeled SMILES. e) The comparison of experimental RT and predicted RT of dipeptide training set using unlabeled SMILES. f) The histogram of RT difference in the training set using unlabeled SMILES. g) The comparison of experimental RT and predicted RT of dipeptide test set using unlabeled SMILES. h) The histogram of RT difference in the test set using unlabeled SMILES.	60
Figure 3.5 a) Distribution of RT difference in tripeptide mixture using the model built by labeled SMILES. b) Distribution of RT difference in tripeptide mixture using the model built by unlabeled SMILES. c) RT difference of tripeptide mixture labeled by different tag numbers.....	64
Figure 3.6 a) The EIC plot of 1 tag labeled LH. b) The mass plot of 1 tag labeled LH at 6.94 min. c) The EIC plot of 2-tag labeled LH.	65
Figure 3.7 a) The putatively identified tripeptides in serum and urine samples. b) The overlap of putatively identified tripeptides between serum and urine samples.....	66
Figure 4.1 The workflow of database construction.....	70
Figure 4.2 a) Example entries in MCIDxKEGG one-reaction. b) The exact mass search interface of MCID 2.0. c) The MCID ID search interface of MCID 2.0. d) the demo result page of MCID 2.0 one-reaction database.	73
Figure 4.3 a) A demonstration of detailed search result against one-reaction database. b) A demonstration of brief search result against one-reaction database.....	74
Figure 4.4 a) The demo identification of methionine. b) The biological reaction of methionine and N-Formylmethionine.	78
Figure 4.5 a) The identification coverage of MCIDxKEGG and MCID database. b) The frequency distribution of MCIDxKEGG one-reaction identification number. c) The frequency distribution of MCID one-reaction identification number. d) The percentage of putatively	

identified metabolites by zero-reaction database with and without predicted products, respectively. e) The frequency distribution of the possible reaction number of putatively identified metabolites by zero-reaction database.83

Figure 5.1 The workflow of building SVM models.....91

Figure 5.2 a) The age distribution of discovery cohort. b) The age distribution of the validation cohort.....93

Figure 5.3 a) PCA plot of the discovery cohort and validation cohort normalized separately. b). PCA plot of the discovery cohort and validation cohort normalized together.....97

Figure 5.4 a) PCA plot of AIS A vs non-A. b) PLS-DA plot of AIS A vs non-A. c) ROC analysis of models differentiating A vs non-A. d) PCA plot of AIS B vs non-B. e) PLS-DA plot of AIS B vs non-B. f) ROC analysis of models differentiating B vs non-B. g) PCA plot of AIS C vs non-C. h) PLS-DA plot of AIS C vs non-C. i) ROC analysis of models differentiating C vs non-C.....103

Figure 5.5 a) PCA plot of converted vs non-converted patients. b) PLS-DA plot of converted vs non converted patients. c) ROC analysis of models built by different numbers of peak pairs. ...106

Figure 5.6 a) PCA plot of motor complete loss vs motor incomplete loss groups. b) PLS-DA plot of motor complete loss vs motor incomplete loss groups. c) ROC analysis of models built by different numbers of peak pairs.....109

List of Tables

Table 1.1 The overview of commonly used databases in metabolomics.	20
Table 2.1 Targeted functional groups for each reaction or class and SMARTS substructure patterns for determining chemical groups.	27
Table 2.2 Misclassified metabolites in YMDB.	32
Table 2.3 Summary of the number of peak pairs identified or matched against three different compound libraries from the human plasma samples analyzed using 4-channel LC-MS.	38
Table 2.4 Summary of the number of peak pairs identified or matched against three different compound libraries from the yeast cell samples analyzed using 4-channel LC-MS.	39
Table 3.1 The overview of dipeptide and tripeptide data.	54
Table 3.2 The RMSE summary of model performance on a different dataset.	55
Table 3.3 The comparison of tripeptide experimental RT and predicted RT.	61
Table 4.1 Statistics of database information.	77
Table 4.2 Identification results of 20 amino acids against different databases.	79
Table 5.1 The overview of baseline AIS grades of participants at admission.	92
Table 5.2 The overview of conversion information in the discovery cohorts.	94
Table 5.3 The overview of conversion information in the validation cohorts.	94
Table 5.4 The overview of motor function outcome in the discovery cohort.	95
Table 5.5 The overview of motor function outcome in the validation cohort.	95
Table 5.6 The confusion matrix of SVM model distinguishing AIS A and non-A grade.	103
Table 5.7 The confusion matrix of SVM model distinguishing AIS B and non-B grade.	104
Table 5.8 The confusion matrix of SVM model distinguishing AIS C and non-C grade.	104
Table 5.9 The confusion matrix of SVM model distinguishing converted and non-converted patients using the top 10 features.	106

Table 5.10 The confusion matrix of SVM model distinguishing converted and non-converted patients using the top 5 features.107

Table 5.11 The confusion matrix of applying the outcome model using the top 5 features to the validation cohort.....109

List of Abbreviations

%RSD	Percent relative standard deviation
ACN	Acetonitrile
AIS	American Spinal Injury Association Impairment Scale
ANOVA	Analysis of variance
CI	Chemical ionization
CE	Capillary electrophoresis
CID	Collision-induced dissociation
CIL	Chemical isotope labeling
DmPA	P-dimethylaminophenacyl
EI	Electron ionization
EIC	Extracted ion chromatogram
ESI	Electrospray ionization
FA	Formic acid
FC	Fold change
FTICR	Fourier transfer ion cyclotron resonance
GC	Gas Chromatography
HILIC	Hydrophilic interaction liquid chromatography
HMDB	Human metabolome database
IS	Internal standard
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
MALDI	Matrix-assisted laser desorption/ionization ()
MeOH	Methanol

MS	Mass spectrometry
MSI	Metabolomics standards initiative
MSTUS	MS total useful signal
m/z	Mass-to-charge ratio
NMR	Nuclear magnetic resonance
PC	Principal component
PCA	Principal component analysis
PLS-DA	Partial least square discriminant analysis
QC	Quality control
Q-TOF	Quadrupole time-of-flight
OPLS-DA	Orthogonal projections to latent structures-discriminant analysis
RF	Radio frequency
RMSE	Root-mean-square error
RPLC	Reversed phase liquid chromatography
RT	Retention time
SCI	Spinal cord injury
S/N	Signal-to-noise ratio
TIC	Total ion chromatogram
TOF	Time-of-flight
UIS	Universal internal standard
UPLC-MS	Ultra-performance mass spectrometry

List of Symbols

Å	Angstrom
°C	Degrees Celsius
μL	Microliter
μL/min	Microliters per minute
μm	Micrometer
cm	Centimeter
Da	Dalton
eV	Electron-volt
h	Hour
Hz	Hertz
L/min	Liters per minute
mg/mL	Milligrams per milliliter
min	Minute
mL	Milliliter
mm	Millimeter
mM	Millimolar
nm	Nanometer
ppm	Parts per million
rpm	Revolutions per minute
s	Second
V	Volt

Chapter 1 Introduction

1.1 Introduction of metabolomics

When talking about omics, most people are more familiar with genomics, a study focusing on all genes from an organism. With the growth of genomics, scientists started to explore the downstream products of genes. These products include ribonucleic acids (RNA) from transcription, proteins from translation, and metabolites from food or metabolic reactions catalyzed by enzymes. Three more omics, transcriptomics, proteomics, and metabolomics, have thus emerged. Among those, metabolomics, linking closely to phenotypes, mainly targets small molecules (usually below 1000 Da) from cells, biofluids, or tissues. It is widely used in biological research, including biomarker discovery^{1, 2, 3}, drug development⁴, and disease diagnosis⁵.

Metabolomics can be further divided into two different strategies, targeted metabolomics⁶ and untargeted metabolomics^{1 2}. Targeted metabolomics requires prior hypothesis and focuses on a set of particular molecules, ranging from several to hundreds. Since the target analytes are predefined, the sample preparation methods can be specifically optimized to decrease the dominant compounds. And the downstream metabolite identification and biological interpretation is also relatively simple.

As a comparison, untargeted metabolomics profiles the whole metabolome aiming to cover as many compounds as possible at the same time. Advanced separation methods are usually coupled to reduce sample complexity. As a result, large amount of data were also generated. The following analysis also becomes more complicated than targeted metabolomics. Acquired compounds include both the known metabolites and unknown ones, resorting to not only in-house database, but also online metabolomics database, even *in silico* ones. Through comparing biological samples from different conditions, like healthy volunteers and patients, untargeted

metabolomics can generate new hypotheses, making it a discovery-oriented approach. So in the following parts, we focus more on untargeted metabolomics.

1.2 Current technology

Currently, two most popular technologies in untargeted metabolomics are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. Mass spectrometry has gained the popularity due to its excellent sensitivity and great coverage. It is a type of analytical instruments which acquires mass to charge ratio (m/z) information of positive charged or negative charged ions. Mass spectrometers can be roughly divided into three parts, ion source, mass analyzer, and detector.

First, molecules need to be ionized in ion source. Electron ionization (EI), electrospray ionization (ESI), and matrix-assisted laser desorption/ionization (MALDI) are routinely used in metabolomics. Among these, EI is one of the hard ionization methods, which applies relatively high energy on gas-phase molecules, resulting in large degrees of fragmentation and detailed mass spectra. One disadvantage of EI is that intact molecular ions are hard to be preserved from high energy electron beam.

As a comparison, ESI is one of the soft ionization techniques. Soft ionization generally impacts relatively low energy on target analytes, leading to less fragmentation. Molecular ions often can be observed in mass spectra, facilitating metabolite identification. In ESI, the sample solution first forms small-charged droplets. With the evaporation of solvent, the droplet size gradually decreases, but the surface charge density continuously increases. This can cause Columbic repulsion and generate gas phase ions.⁷ One disadvantage of ESI is that metabolites are liable to ion suppression during the ionization competition process.

MALDI is another soft ionization technique. Before the ionization samples need to be mixed homogeneously with a large amount of suitable matrix. The mixtures are coated to a plate. Then the matrix can absorb energy from a pulsed laser. Both samples and some matrix materials are vaporized from the plate. At the same time the charge is transferred from ionized matrix to sample, generating ionized analytes in gas phase.

Mass analyzers commonly used in metabolomics include, but are not limited to, quadrupole, time-of-flight (TOF), ion trap, orbitrap, and Fourier transform ion cyclotron resonance (FTICR). Different analyzers should be selected based on the research object. Mass range and resolving power (resolution) should be taken into consideration. For example, if the element composition of target analytes needs to be determined, high resolution mass analyzers, like FTICR and orbitrap, should be selected. Another example is using ion trap mass analyzers to acquire MS^n spectra. MS^n refers to the product ions from previous stage submitted to fragmentation again, which can be used for structure identification. Ion trap mass analyzers can trap ions, detect ions, and isolate ions of interest. The isolated ions can be fragmented by collision with neutral molecules. Multiple isolation and CID experiments can be performed on the product ions from each round until desired MS^n level. Multiple mass analyzers can also be combined as hybrid systems. For example, TOF can be coupled to quadrupoles as Q-TOF⁸, a linear trap can also be coupled to an orbitrap mass analyzer⁹.

Different chromatography techniques can be coupled to MS to further improve the metabolome detection, such as gas chromatography (GC), liquid chromatography (LC), and capillary electrophoresis (CE). If the target analytes are thermally stable and volatile, GC-MS is a great choice to profile samples. For nonvolatile compounds, derivatization is necessary to increase their volatility and stability at high temperature. Since GC-EI-MS instruments are highly

standardized, the MS/MS spectra are highly reproducible across different instruments, which facilitates metabolite identification by matching the experimental spectral to external libraries or software, such as NIST (<https://chemdata.nist.gov/>). For example, Jiye et al. have proposed a workflow to extract and analyze human plasma sample using GC-MS.¹⁰ Five different extraction solvents were first optimized. After extraction, samples were submitted to methoxymation and following trimethylsilylation. Derivatized samples were then analyzed by GC-MS. After optimization, more than 500 peaks could be detected. Among these, 80 could be identified. The precision and linearity of 32 endogenous metabolites were also evaluated. Another example is that Song et al. employed GC-MS to analyze breathe samples for discovering biomarkers of non-small cell lung cancer.¹¹ Breathe samples were collected from 43 patients and 41 normal controls. Around 100 volatile organic compounds were detected. Compared to the control group, two metabolites, 1-butanol and 3-hydroxy-2-butanone, significantly higher in the patient group, indicating they could be potential biomarkers for distinguishing lung cancer patients at early stage or late stage.

Compared with GC-MS, the sample preparation in LC-MS is relatively simple and the environment temperature during data acquisition is lower. Through selection or combination of different columns, such as reverse phase chromatography (RPLC) for less polar metabolites or hydrophilic interaction chromatography (HILIC) for polar and ionic metabolites, targeted and untargeted metabolomics can be performed. For example, Wang et al. adopted LC-ESI-MS technology to study type 2 diabetes mellitus (DM-2).¹² 34 human plasma samples from type 2 diabetes mellitus patients and 35 human plasma samples from healthy control were collected. Based on prior knowledge, phospholipids were targeted in their research. A diol column was employed to separate phospholipids. After data acquisition, principal components analysis (PCA)

and partial least squares-discriminant analysis (PLS-DA) were used for distinguishing the patient group from healthy group and finding potential biomarkers. Other separation methods can also be coupled to MS. For example, CE displayed superior separation efficiency. But the poor reproducibility limits its application in metabolomics, especially when dealing with complex biological samples.¹³

Several advantages of NMR, such as excellent reproducibility, simple sample preparation, non-destructive measurements, make it also a popular technique in metabolomics. In a magnetic field, the atom nuclei can absorb energy to achieve excitation state. Then the re-emission energy can be recorded as signal for further analysis. NMR can be further classified based on the target atom nuclei. The most common target in NMR is ^1H on account of its high natural abundance. For example, Jung et al. proposed an approach to use ^1H NMR to separate the beef based on their geographical origin, including the United States, Australia, Korea, and New Zealand.¹⁴ After data collection, data overview was illustrated by PCA. Orthogonal projections to latent structures-discriminant analysis (OPLS-DA) was used for separate beef samples from different origins. Significant metabolites were selected by one-way ANOVA and following Tukey's multiple-comparison tests. The experiment showed that ^1H NMR is efficient to discriminate the origin of beef samples. Other nuclei can be targeted in NMR, such as ^{13}C , ^{15}N , ^{31}P . For example, ^{13}C -NMR can be applied to flux analysis¹⁵. To increase the coverage, one more orthogonal NMR spectroscopy can be combined forming Two-dimensional (2D) NMR. Homonuclear 2D NMR, such as ^1H - ^1H total correlation spectroscopy¹⁶, and heteronuclear 2D NMR, such as ^1H - ^{13}C single quantum coherence¹⁷, are available for metabolomics studies.

1.3 The workflow of LC-MS-based untargeted metabolomics

The workflow of untargeted metabolomics generally contains five steps, showed in Figure 1.1. The metabolites lost in experiment steps or related information lost in following steps can undermine the integrity of sample information. Thus, we should pay attention to the experiment section as well as the analysis section.

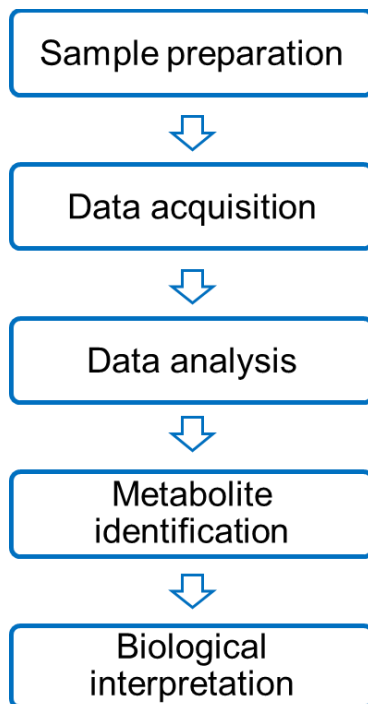


Figure 1.1 Workflow of LC-MS based untargeted metabolomics.

1.3.1 Sample preparation

Sample preparation mainly focuses on extracting analytes from original samples and reducing potential interferences. In untargeted metabolomics, researchers attempt to separate metabolites as many as possible from cells, biofluid, or tissues, and remove large molecules like proteins at the same time. Sample preparation can be classified into two main groups: label-free and chemical derivatization methods.

1.3.1.1 Label-free methods

Label-free methods are widely used in untargeted metabolomics. The sample preparation is relatively simple, facilitating its application especially in data acquisition of large-scale metabolomics projects. It is also less complicated to couple with robotic systems¹⁸, reducing potential inconsistency or mistakes induced by human operation. As we all know, metabolites are highly diverse in physical properties and chemical properties. Consequently, different LC separation methods and ion modes should be combined to fully acquire metabolome data. Generally, to achieve high metabolome coverage, reverse phase liquid chromatography (RPLC) and hydrophilic interaction liquid chromatography (HILIC) in both positive ion mode and negative ion mode should be combined to acquire data.

RPLC is commonly used in metabolomics. The stationary phase of RPLC consists of alkyl chains (such as C4, C8, and C18) bonded to silica column, generating a hydrophobic environment. The mobile phases in RPLC are more polar, such as water, acetonitrile (ACN), and methanol. The stationary and mobile phases make RPLC retain hydrophobic molecules. The hydrophilic molecules are eluted by polar mobile phase at earlier time. With the increase of organic solvent percentage in mobile phase, less polar molecules can be eluted later. Gray et al. have proposed a robust workflow using RP-UPLC to profile the large-scale urine samples.¹⁹ The robustness and repeatability were tested by 1000 consecutive injections of human urine samples. The relative standard deviation of 71% and 92% metabolites in QC samples were below 15% and 30%, respectively, indicating this workflow is capable for large-scale metabolomics study.

As we mentioned before, HILIC is required to provide better separation for hydrophilic and ionic metabolites. Similar to normal phase liquid chromatography, HILIC also adopts polar

stationary phase, like silica, amino²⁰, cyano²¹, and amide²². But the mobile phase used in HILIC is close to RPLC, such as ACN, methanol, and water. Compared with the non-polar mobile phase in normal phase liquid chromatography, the major advantage of HILIC is that metabolites from biological samples can be well dissolved in its solvent system. Although HILIC usually presents poor reproducibility and peak shapes, it is still a complementary approach to RPLC. Tolstikov et al. have proposed a method using HILIC-ESI-ion trap to detect polar metabolites.²² Two HILIC columns (Polyhydroxyethyl A and TSK Gel Amide 80) with different modifications and coatings were investigated. The separation of 12 polar standards were first evaluated on these two columns. Then the phloem samples were also tested. The detected polar metabolites included oligosaccharides, glycosides, amino sugars and so on. Sriboonvorakul et al. employed HILIC to detect small organic acids from plasma and urine samples related to acidosis.²³ In their method, all eight small acids achieve high accuracy and precision.

To achieve high metabolome coverage, RPLC and HILIC should be combined for detecting both hydrophobic and hydrophilic compounds. Contrepois et al. investigated the performance of five C18 RPLC columns.²⁴ As complementary, five HILIC columns at acid, neutral, and basic conditions were also tested. Taking RPLC-MS alone as reference, the combined method could detect 44% and 108% more features in urine samples and plasma samples, respectively. Ivanisevic et al. developed an approach that applied the same extraction method on both RPLC-MS in positive ion mode and HILIC-MS in negative ion mode to overcome limited sample volume.²⁵ Different sample types, including bacterial cells, human plasma, and human cancer cells, were evaluated. In each sample type, more than 30,000 features were able to be detected.

1.3.1.2 Chemical derivatization methods

Chemical derivatization methods require specific reagents to react with a subset of metabolites sharing similar chemical properties. Since only a sub-metabolome is extracted after derivatization reaction, the ion suppression for remaining metabolites is reduced. Besides, derivatization reagents can also increase sensitivity and improve separation of target analytes by modifying their physical properties. For example, dansyl chloride (DnsCl) can be used for labeling amine or phenol groups, showed in Figure 1.2.²⁶ After derivatization, the aromatic rings from the labeling tag can improve the retention of labeled compounds on RPLC, especially for those polar metabolites or ionic metabolites, which can simplify the requirement for multiple separation methods or instrument. Compared with label-free method, the application on 20 amino acids demonstrated that dansyl derivatization could increase the signal 10-fold to 1000-fold increase even with 100-fold less injection amount. Besides, these amino acids were better spread on the chromatography. Moreover, ¹³C dansylated pooled samples serving as internal standards could be mixed with ¹²C dansylated individual samples. Instead of comparing the absolute signal intensity or peak pairs like label-free methods, relative ratio of ¹²C/¹³C peak pairs could be used for quantification and produce less variation. Urine samples after derivatization were also tested with a 12 min fast LC gradient, and 672 metabolites were detected.

Since derivatization usually only targets a subset of metabolites, to cover the whole metabolome, multiple derivatization methods should be incorporated together. Besides the amine or phenol groups we mentioned before, DnsCl can also be used to label metabolites containing hydroxyl groups.²⁷ A liquid-liquid extraction step was employed to extract hydroxyl metabolites into organic phase. And 4-dimethylaminopyridine (DMAP) was used as base to active dansylation. More than 3000 peak pairs could be detected in urine samples by this method. Another important

functional group, carboxylic acid, can be derivatized by p-dimethylaminophenacyl (DmPA) bromide.²⁸ This approach was able to obtain more than 2500 carboxylic acids related metabolites from human urine samples.²⁹ To profile carbonyl (for both aldehyde and ketone) submetabolome, dansylhydrazine (DnsHz) can be used as derivatization reagents. In this method, 1737 common peak pairs could be detected in 6 replicates. Other functional groups, such as thiols^{30 31}, can also be derivatized to facilitate detection.

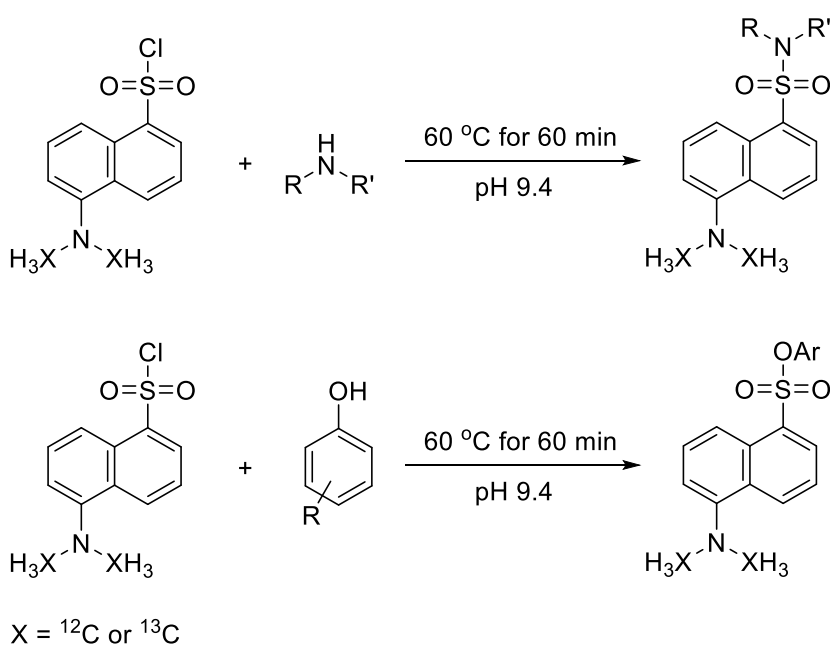


Figure 1.2 Using dansyl chloride to label primary, secondary amines and phenols.

1.3.2. Data acquisition

After metabolites separated by LC systems, they are submitted to mass spectrometers to convert the chemical concentration information into digital format. As mentioned before, metabolites are first ionized at ion source. Then, ions with different m/z are separated by the mass analyzers. At last, the detectors capture the ions and record the signal. If the metabolite

concentration is high, the signal also increases. There are multiple choices for mass analyzer selection, including quadrupoles, ion traps, time-of-flight (TOF), orbitrap et al. Current mass spectrometers tend to couple multiple analyzers together, such as ESI-Q-TOF (Impact II from Bruker) and ESI-Q-Orbitrap (Q Exactive from Thermo Fisher Scientific) employed in the projects of this thesis. A demonstration of Q-TOF instrument is shown in Figure 1.3. A time-of-flight mass analyzer replaces the third quadrupole in QQQ. In full scan mode, both the first quadrupole (Q1) and the second quadrupole (Q2) perform radio frequency (RF) only to let all ions pass through. After ions arrive at the TOF tube, a pulse voltage is applied at extractor to generate the same initial ion kinetic energy for all ions. Ions with larger m/z gain less speed compared to ions with lower m/z , leading to longer fly time in TOF chamber and later arrival at detector. The time from extractor to detector is measured for calculating m/z . Q-TOF can also be used for acquiring MS/MS spectra. Q1 can perform either RF only or selection of specific mass window. The ions transmitted from Q1 are collided with neutral gas molecules in Q2 acting as collision cell. Then the fragment ions are separated by TOF mass analyzer and recorded by detector as MS/MS.³² Several advantages, such as excellent mass range, fast scan speed, great accuracy, relative high resolution, and acceptable price make Q-TOF a great choice for metabolomics work.

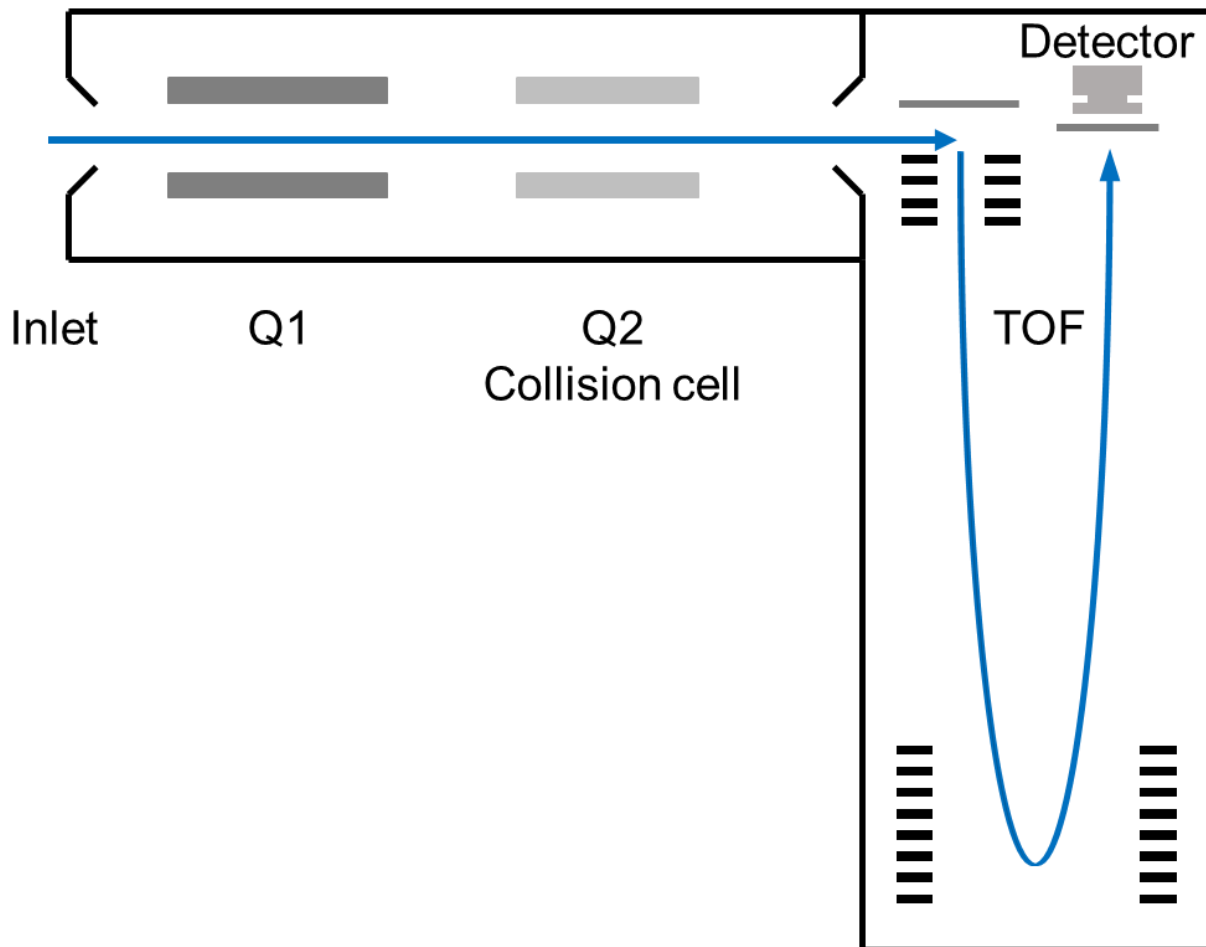


Figure 1.3 The demonstration of Q-TOF.

In Q Exactive, after ionization, ions are transferred to a quadrupole mass filter. This quadrupole connected to a C-trap. Ions in C-trap can further travel to higher-energy C-trap dissociation (HCD) collision cell for fragmentation or orbitrap for m/z detection. Similarly, Q Exactive can also perform full scan mode to collect MS spectra. Ions from ion source pass through the quadrupole with RF only mode. After they are collected and stabilized at the C-trap, they are further sent to the orbitrap for detection. In orbitrap, ions are trapped in an orbital motion along the axial central electrode. And their frequency of rotation is related to m/z . Through resolving the oscillation frequencies, the m/z of ions can be calculated. If MS/MS spectra are required for

collection, ions of interest are selected by the quadrupole and other ions are filtered. Then the ions pass through the C-trap to the HCD collision cell for collision induced fragmentation. Next, fragment ions are sent to C-trap for stabilization. And the last step is the same with full scan mode, fragment ions are injected to orbitrap for m/z detection. The mass resolution, sensitivity, mass measure accuracy of orbitrap is excellent.⁹ Compared with FT-ICR, the maintenance is relatively simpler and more affordable. One limitation for orbitrap is the trade-off between scan speed and resolution or accuracy.

1.3.3. Data processing

Due to the complexity of biological samples, untargeted metabolomics usually generate large amount of data, which challenges the data processing and data analysis. The data processing workflow usually covers following steps: peak picking, peak alignment, retention time correction, and filling missing peaks. XCMS is one of the widely used programs to process metabolomics data.³³ In peak detection, the data on mass range was cut into small mass unit. For each unit, the chromatogram was constructed. To remove background influence, a second order derivative Gaussian model was employed to select integrated peak. These detected peaks were submitted to peak matching across different samples. In this step, XCMS uses kernel density estimator to calculate peak distribution and then dynamically set the retention time boundary for peaks which might come from the same metabolites. Sample group information was also used to remove insignificant peaks. The peaks presenting more 50% in at least one group would be kept. Then peaks showed in most samples, called “well-behaved” peak groups, were selected for retention time correction. If there are more than one peak in the group form one sample, the one with highest intensity is selected. A nonlinear algorithm was applied for retention time correction. After the

first-time correction, more “well-behaved” peak groups could be identified. To improve alignment precision, the correction steps can be repeated several times. In last step, the missing peaks were retrieved from the raw data based on the m/z range and retention time region.

A different data processing workflow was applied for chemical isotope labeling method. Since both the light labeling and heavy labeling peaks are presented in acquired data, the constructed peak pairs from different isotope labeling instead of chromatography peaks can be used to extract metabolite information. IsoMS is an example for processing chemical isotope labeled metabolomics data.³⁴ First, the raw data generated from LC-MS is converted to centroid mode, containing RT, m/z , and intensity. Then, IsoMS pairs up the light labeling and heavy labeling ions based on the given mass difference and tolerance window. Different confidence levels are determined by isotopic pattern. If the isotopic patterns of both ions can be found and are the same, they are classified as Level 1 peak pairs. If only one isotopic pattern can be found, the peak pairs are classified as Level 2. If both isotopic patterns are missing, they are considered as Level 3. Only Level 1 and Level 2 peak pairs are submitted to next steps. Since the $[M+H]^+$ is usually the dominating form for labeled metabolites, if $[M+H]^+$ form can be found in data, other adducted ions (like $[M+Na]^+$, $[M+K]^+$, and $[M+NH_4]^+$), dimers, or in-source fragmentation are filtered. Background peak pairs can also be removed by manually assigning their m/z into a .CSV file, isMZBackground. After filtering, peak pairs from the same metabolite presenting in multiple spectra are grouped together. At last, the same peak pairs detected in different samples are aligned based on the RT and mass tolerance window defined by users. The aligned data is ready for data normalization and following data analysis.

The relative quantitation data obtained from data processing cannot be used for data analysis directly, since the acquired intensity is largely affected by sample size, sample weight, or

sample total concentration. For example, the urine concentration is highly influenced by many factors, such as fluid intake and diet. As a result, data normalization is required to scale the overall sample concentration.³⁵ Data normalization can be classified into pre-acquisition normalization and post-acquisition normalization.³⁶ Generally, the pre-acquisition normalization is employed between sample preparation and data acquisition. And the post-acquisition normalization is applied after data processing. In pre-acquisition normalization, sample volume or sample size is adjusted based on the measured quantities to balance the total sample concentration. For example, in urine sample, creatinine concentrations can be considered as reference for total urine concentrations.³⁷ A higher creatinine concentration indicates more concentrated urine sample. Less sample volume is taken from sample with higher concentration. But in other biological samples, no similar reference compound like creatinine in urine can be used for normalization. Other methods, such as cell count for cell samples, Na⁺ concentration in sweat samples³⁸, and UV absorbance³⁹ have been developed. The UV absorbance is a more universal approach and can be applied to different sample types, including but not limiting to urine samples³⁹, *E. coli* samples⁴⁰, and sweat samples⁴¹. The benefit of pre-acquisition normalization is that instruments can provide comparable responses for the corrected samples, especially in MS-based platform. The acquired metabolite signals are not only affected by the concentration in biological samples, but also influenced by ionization efficiency and ion suppression effect. This leads to a nonlinear relationship between the acquired signals and metabolite concentration. Another advantage is that pre-adjusted sample total concentration can prevent signal saturation and sample carryover. If the sample volume is not well-determined, the samples with higher concentration are easier to reach saturation when the sample volume is optimal for more diluted samples. The major disadvantages of pre-sample acquisition are the requirements of extra experiment steps and extra amount of

samples to measure reference quantities. In large scale metabolomics project, extra experiment step means additional time, additional effort, and a higher chance to induce operation caused error. Simplifying the pre-sample acquisition methods is also important.

On the contrary, post-acquisition normalization does not require any extra experiment step. It is also called as data-driven normalization. After the quantitative information is extracted from data processing, normalization can be applied. Warrack et al. introduced MS total useful signal (MSTUS), the total intensity of peaks shared by all samples, as the reference for normalization.³⁵ Other scale factors include MS total signal, MS group useful signal, selected ion count for all metabolites, median fold change, specific gravity, etc.³⁶ Since the normalization methods are employed after data acquisition, the experimental and instrumental variabilities can be corrected. As mentioned before, no extra experiment step is involved, but it cannot reduce the influence from ionization efficiency and ion suppression effect.

The pre-acquisition and post-acquisition normalization methods can be combined. Chen et al. have proposed the combination of creatinine as pre-acquisition normalization and MSTUS as post-acquisition normalization for urine samples.⁴² To demonstrate the normalization effect, urine samples were serially diluted. The performance of no normalization, using all MS signals as normalization, using MSTUS as normalization, using creatinine value as normalization, combining creatinine value and all MS signals as normalization, combining creatinine value and MSTUS as normalization were compared. Their experiments demonstrated that the combined methods can effectively reduce the non-biological difference.

1.3.4 Data analysis

The data analysis in metabolomics covers both statistical methods and machine learning techniques. These methods cover univariate analysis and multivariate analysis with both unsupervised and supervised methods. Univariate analysis⁴³ can be used for selecting significantly changed metabolites between different groups. Different analysis methods require different assumptions⁴⁴. For example, dataset should follow normal distribution when using Student's *t*-test. If the dataset does not meet the requirement, nonparametric tests, such as Wilcoxon rank-sum test, can be employed. Among unsupervised methods, principal component analysis (PCA) is the most widely used in metabolomics. Through dimension reduction, PCA can present the data overview, which facilitates researchers to check data quality. The quality control (QC) samples should be closely clustered together in PCA plot. It can also assist to remove outliers.⁴⁵ After checking data quality and selecting significant changed metabolites, supervised methods, such as PLS-DA, can be applied to build model and discriminate metabolites as potential biomarkers. In recent years, machine learning methods have become more and more popular. For example, Mahadevan et al. compared the performance of PLS-DA and support vector machines (SVMs) on human urine samples from healthy volunteers and *Streptococcus pneumoniae* patients.⁴⁶ It demonstrated that SVM could build better predictive model and generate higher accuracy on classification between healthy and disease groups. Date et al. applied deep neural networks (DNNs) on the classification of geographical origins of yellowfin goby.⁴⁷ Two NMR dataset containing water-soluble metabolites and methanol-soluble metabolites were tested. The prediction result showed that the DNN outperformed the PLS-DA and was close to the overall performance of SVM.

1.3.5 Metabolite identification

Unlike other omics containing generic reference, there is no prior knowledge for metabolite identification in metabolomics. The high diversity of metabolites makes identification even more difficult. To identify metabolites, researchers can compare the experimental data with reference data from in-house libraries or online databases. In-house libraries refer to the experimental data and reference data from chemical standards that are acquired at the same conditions. According to Metabolomics Standards Initiative (MSI), metabolite identification can be classified into four different levels⁴⁸. Level 1 is identified compound, requiring at least two orthogonal pieces of information from chemical standards, including but not limiting to accurate mass, retention time, and MS/MS spectra. These data should be acquired at the same experimental conditions. Level 2 is considered as putatively annotated compound. The reference data can come from external laboratories or literatures. Level 3 is putatively characterized compound classes. It needs one piece of information, which may not be able to exclude other candidates. For example, if only accurate mass is used for identification, compounds sharing the same chemical formula could never be filtered. Level 4 represents metabolites that were detected in biological samples, but remained as the unknown.

Among the information for identification, retention time is determined by the LC conditions. The selection of mobile phase, flow rate, different columns, or temperature can influence RT. MS/MS spectra except EI spectra usually vary across different platforms. These two can be acquired from in-house laboratories, compared with online database, or predicted from algorithm. Accurate mass is more consistent from different laboratories or different instruments. And it can be directly calculated with a given formula.

1.3.5.1. Databases for metabolite identification

Since chemical standards are difficult to collect sometimes, reference data from external databases or algorithm-based prediction are necessary for metabolite identification. To use accurate mass to identify metabolites, researchers can obtain it from metabolomics databases, such as Human Metabolome Database (HMDB)⁴⁹ and Kyoto Encyclopedia of Genes and Genomes (KEGG), or from chemical databases, such as PubChem⁵⁰ and ChemSpider⁵¹. A summary of commonly used databases is shown in Table 1.1. Take HMDB as an example, HMDB mainly focuses on human metabolome. It enrolls three types of data: 1) chemical data, such as chemical taxonomy and physical properties; 2) clinical data, such as normal concentrations, abnormal concentrations, and associated disorders & diseases; 3) molecular biology/biochemistry data, such as cellular position and pathway information. The total number of metabolites in HMDB reaches 114,100, including detected and quantified metabolites, detected but not quantified metabolites, expected metabolites, and predicted metabolites. Only the number from two categories of the detected metabolites is listed in Table 1.1.

For chemical databases, small molecules presented in nature or synthesized by researchers are included. For example, PubChem is an open database hosted by the US National Institutes of Health (NIH).⁵⁰ Researchers can upload their experimental data and facilitate other users to search. PubChem mainly focuses on small molecules, large molecules like nucleotides and chemically-modified macromolecules are gradually enrolled. Currently, there are 110,636,827 compounds in PubChem.

To putatively identify unknown compounds which may not exist in databases above, researchers can search the accurate mass against predicted database. MyCompoundID⁵², an evidence-based metabolome library (EML), is composed of potential products of known

metabolites. 76 metabolic reactions, such as oxidation, acetylation, and deacetylation, were applied on 8021 metabolites from HMDB for predicting products. 375,809 predicted products were used to construct the one-reaction database. The predicted compounds in one-reaction database were considered as substrates for further prediction to generate the two-reaction database. The two-reaction database consists of 10,583,901 predicted compounds in total.

Table 1.1 The overview of commonly used databases in metabolomics.

	Type	Number	Link
HMDB	metabolites	23,153	https://hmdb.ca/
KEGG Compound	metabolites	18,844	https://www.genome.jp/kegg/compound/
MetaCyc	metabolites	17,422	https://metacyc.org/
PubChem	small molecules	110,636,827	https://pubchem.ncbi.nlm.nih.gov/
ChemSpider	small molecules	111 million chemical structures	https://www.chemspider.com/
ChEBI	small molecules of biological interest	59,389	https://www.ebi.ac.uk/chebi/
DrugBank⁵³	drugs	14,575	https://go.drugbank.com/
MyCompoundID	predicted metabolites	375,809 in one-reaction database, 10,583,901 in two-reaction database	http://www.mycompoundid.org/mycompoundid_IsoMS/

Besides accurate mass, MS/MS spectra can also be used for characterizing molecules. Researchers can identify metabolites by matching between experimental MS/MS spectra and

reference spectra. But the features of MS/MS spectra are influenced by many conditions, such as dissociation techniques, dissociation energy, and platforms. The same compound acquired from different experimental conditions can produce different spectra. For example, in collision-induced dissociation, with the increase of collision energy, more smaller fragment ions appear in spectra, and the ratio of these smaller fragment ions also likely rises. To achieve correct identification, ideally, the reference spectra data should be acquired using the same instrument. When matching reference spectra from online database, users should select spectra acquired at the same or similar conditions. Currently, there are a few mass spectrum databases available, including but not limiting to HMDB (https://hmdb.ca/spectra/ms_ms/search), NIST (<https://chemdata.nist.gov/>), MassBank (<http://www.massbank.jp/>), and mzCloud (<https://www.mzcloud.org/>).

1.3.6. Biological interpretation

After identifying the acquired signals as metabolites, researchers can interpret the biological meaning of designed experiments. The relationship between metabolites or between metabolites and other biologically active materials, such as enzymes and genes, can be revealed. Correlation based network analysis is a common approach used for omics data analysis. The correlation change of different metabolites from two experimental conditions can provide clusters of metabolites sharing the same patterns. Fukushima developed an R package, DiffCorr, using Fisher's z-test to discover differential correlations.⁵⁴ Two datasets were used for demonstrating its application. The first dataset is gene expression of acute lymphoblastic leukemia and acute myeloid leukemia samples. And the second dataset is metabolomics data between flavonoid-deficient *Arabidopsis thaliana* and the wild-type.

Another approach is to incorporate the metabolite information into metabolic pathway. Instead of addressing metabolite change on their own in univariate analysis, several metabolites from the same pathway are analyzed together via enrichment analysis. The pathway information can be obtained from reference databases, such as KEGG⁵⁵, MetaCyc⁵⁶, and SMPDB⁵⁷. Kankainen et al. presented metabolite pathway enrichment analysis (MEPA) to analyze metabolite function at the systematic level.⁵⁸ Metabolomic data on twin pairs with discordant body weight were used and demonstrated MPEA could discovery altered metabolic pathways. Currently, several tools have been under development for pathway enrichment analysis (such as MetaboAnalyst⁵⁹ and MetScope⁶⁰) and visualization (such as Cytoscape⁶¹).

1.4 Scope of the Thesis

This research focuses on method development on the data analysis section to increase metabolome coverage.

In Chapter 2, the theoretical coverage from metabolome database of four sub-metabolomes and the experimental coverage from biological samples was analyzed.

In Chapter 3, the retention time of chemical isotope labeled tripeptides was predicted to increase the identification accuracy.

In Chapter 4, theoretical metabolites were predicted based on the biological reactions to facilitate the unknown compound identification.

In Chapter 5, human serum samples from patients with acute spinal cord injury were analyzed to differentiate patients from different severity and predict neurological conversion as well as motor function recovery.

Chapter 2 Metabolomic coverage of chemical-group-submetabolome analysis: group classification and 4-channel chemical isotope labeling LC-MS

2.1 Introduction

Because of great diversity of chemical and physical properties of metabolites present in a complex metabolome sample, conventional liquid chromatography mass spectrometry (LC-MS) approach of metabolome analysis relies on the use of multiple LC and MS conditions to increase the number of metabolites detectable or the metabolome coverage. For example, the combination of reversed-phase (RP) LC column for separation of relatively hydrophobic metabolites and hydrophilic interaction liquid chromatography (HILIC) column for separation of relatively hydrophilic metabolites, along with positive and negative ion MS detection, allows the detection of different types of metabolites.⁶² This approach has the advantage of using a simple workflow with readily available instrument and software for metabolite detection and data analysis and thus can be easily implemented. However, this approach has the shortcomings of limited metabolome coverage due to low detectability of many metabolites and limited quantification accuracy due to the lack of suitable internal standards for a vast majority of metabolites. Chemical isotope labeling (CIL) LC-MS offers a means of overcoming these limitations.⁶³

CIL LC-MS metabolome analysis is a divide-and-conquer approach where the metabolites are divided into different chemical groups (e.g., amines, acids, etc.), instead of dividing them according to physical properties such as hydrophobicity and ionic property.^{26,64} Each group of metabolites is chemically labeled with a suitable reagent, followed by LC-MS analysis. Many reagents have been developed for both targeted and untargeted metabolome analysis.⁶⁵⁻⁷⁰ With rational design of chemical structures of the labeling reagents, concomitant improvement in both metabolite separation and ionization can be achieved, resulting in significant enhancement in

metabolite detectability and hence much higher metabolome coverage.²⁶ Using differential isotope labeling (e.g., ¹²C-reagent labeled individual samples spiked with a ¹³C-reagent labeled reference or pooled sample, followed by LC-MS analysis of the resultant mixtures), accurate relative quantification of all labeled metabolites in comparative samples can be performed.⁷¹

The presumed disadvantage of CIL LC-MS is the requirement of chemical derivatization that may add a complication in sample processing. However, sample processing for metabolome analysis often involves multiple steps^{62,63} and thus a robust chemical reaction (e.g., by merely adding a reagent to a sample) may be seamlessly incorporated into the overall workflow, just as it is done for protein precipitation (e.g., by adding a solvent to precipitate proteins and then removing them), sample normalization (e.g., by creatinine measurement for urine samples), cell lysis (e.g., by adding a lysis reagent), metabolite extraction (e.g., by adding a solvent for liquid-liquid extraction), etc. Thus, performing chemical labeling, if properly done, should not inconvenience the sample handling process. The benefits of improving metabolite detectability and quantification accuracy significantly outweigh the addition of an extra labeling step. However, a more fundamental question is actually related to the number of labeling reactions we need to do for a given sample in order to cover the whole chemical space of the metabolome. Addressing this question will allow us to understand the chemical group diversity of a metabolome, prioritize the development efforts on labeling chemistries to target certain groups of metabolites, and examine the deficiency of current labeling methods to guide future method optimization or new labeling method development.

In this study, we report our investigation of chemical group diversity of compound entries in some commonly used metabolome databases. We developed and applied a high-performance 4-channel chemical labeling approach, based on dansylation for analyzing amines,²⁶ base-activated

dansylation for hydroxyls,²⁷ DmPA bromide labeling for carboxylic acids,²⁸ and dansylhydrazine (DnsHz) labeling for carbonyl metabolites,⁶⁴ for the analysis of human plasma as well as yeast cells in order to examine the current coverage of these four groups of metabolites in representative complex metabolome samples. By comparing the distribution of chemical groups of database compounds with those detected in 4-channel CIL LC-MS, we discussed some of limitations of the current methods that we hope to stimulate more future research activities to meet the ultimate goal of using CIL LC-MS for whole metabolome profiling.

2.2 Experimental Section

2.2.1 Chemical group classification

A Java-based program was developed to classify the compounds in a database according to their chemical groups. The workflow of metabolite classification contains five steps (shown in Figure 2.1). First, we downloaded the five selected metabolome databases: MyCompoundID (MCID),⁵² HMDB,⁴⁹ KEGG,⁵⁵ YMDB,⁷² and ECMDB.⁷³ MCID is an evidence-based database, including the metabolites detected from human sources and the predicted compounds generated from these human metabolites after subjecting them to one or two common metabolic reactions (MCID-1R as the one-reaction library and MCID-2R as the two-reaction library). In this work, we used MCID zero-reaction library for group classification, as the structures of all entries are known. HMDB collects detailed information about small molecules such as chemical property data, clinical and biochemical data with the initial focus on human metabolome, but has expanded to include compounds that may be associated with human (e.g., food, drugs and chemicals of environmental resource) as well as some predicted metabolites. KEGG database includes small molecules along with their biological processing information such as reactions, pathways and related enzymes as well as biological connectivity among the compounds. YMDB and ECMDB

focus on the metabolites and pathway information of two widely used model organisms, *Saccharomyces cerevisiae* (yeast) and *Escherichia coli* (*E. coli*), respectively.

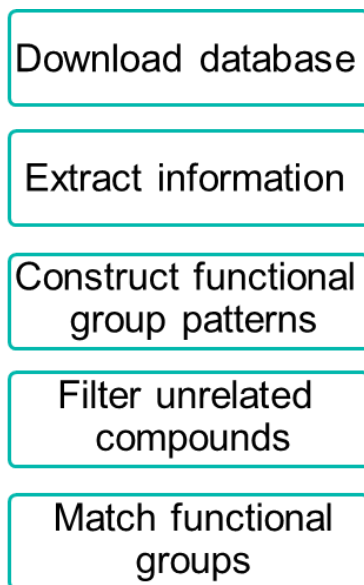


Figure 2.1 The workflow of metabolite classification.

The 2nd step was to extract compound information, including compound names and chemical structures, database ID and other information such as source of compound (e.g., drug and food) and actual or predicted compound. We then built the chemical substructure patterns of functional groups including different patterns for aliphatic and aromatic atoms in molecules (see Table 2.1 for the list). Note that thiols are grouped into hydroxyls, as they can be labeled using similar chemical labeling reaction conditions. To accurately determine the functional groups (e.g., amines vs. amides), the SMARTS (SMiles ARbitrary Target Specification) program (www.daylight.com/dayhtml/doc/theory/theory.smarts) was used to construct the exact substructure patterns (see Table 2.1).

Table 2.1 Targeted functional groups for each reaction or class and SMARTS substructure patterns for determining chemical groups.

Derivatization reactions/class	Patterns of functional groups
Amine/phenol labeling	Primary or secondary amines, N-H bond in aromatic environment, phenols
Carbonyl labeling	Aldehydes, ketones
Hydroxyl labeling	Hydroxyls, phenols, thiols
Carboxyl labeling	Carboxyls and the conjugated bases
Esters	Esters
Amides	Amides

Derivatization reactions/class	SMARTS substructure patterns
Amine/phenol labeling	[NX3;H2,H1;!\$(NC=O)], [nX3;H2,H1;!\$(nc=O)], [OX2H][cX3]:[n,c]
Carbonyl labeling	[CX3H1](=O), [#6][CX3](=O)[#6], [#6][cX3](=O)[#6]
Hydroxyl labeling	[CX4][OX2H], [CX3,NX2]=[CX3][OX2H], [OX2H][cX3]:[c,n], [#16!H0]
Carboxyl labeling	[CX3](=O)[OX1H0-,OX2H1]
Esters	Aliphatic and aromatic esters
Amides	Aliphatic and aromatic amides

The next step was to filter out the unconventional metabolites which we defined as lipids (particularly long-chain lipids), inorganic species and other molecules that are unique to drug, food, plant and environmental origins. Although lipids can also be considered as metabolites, they can be extracted and analyzed using methods that are different from CIL methods. Thus, in this study, we excluded the lipids. HMDB contains superclass information such as those denoted as lipids and lipid-like compounds. We used the superclass information to remove lipid and lipid-like compounds. Considering some lipids were not removed using the superclass information, compounds containing equal to or more than eight-carbon chains with no class information were also filtered out as lipids. Then, class information was used to remove flavonoid, coumarin, and lignan related plant compounds. Compounds not containing any carbon were filtered out as

inorganic compounds. In addition, we removed drugs and environmental compounds. For KEGG, any compounds without SMILES structure information were removed. Generic compounds representing homologous series were filtered out according to the “Comment” entry. Inorganic compounds were removed. Because KEGG also contains lipids, phytochemical compounds and others, BRITE, manually generated functional hierarchies⁵⁵, was used to filter out the unconventional compounds. Compounds meeting the three following criteria were kept: 1) compounds containing any information of reaction or pathway or module, 2) compounds belonging to 08001 and 3) compounds not containing BRITE information.

The final step of the program was to determine the functional group(s) in a compound structure by matching it to different group substructure patterns. We wrote a Java based program, SubstrcMatch, which uses chemical structure files (in SMILES format) and substructure patterns as input and then generates a .txt file to indicate whether a compound contains the targeted functional group. The group classification results from different databases were used for metabolomic coverage analyses.

2.2.2 4-Channel labeling

The general workflow for metabolome analysis using 4-channel CIL LC-MS is shown in Figure 2.2. It includes the following steps: 1) sample pretreatment and metabolite extraction, 2) generation of a pooled sample by mixing aliquots of all individual samples, 3) dividing a sample into four aliquots, 4) applying four isotope labeling chemistries targeting different submetabolomes, 5) LC-UV quantification of dansyl-labeled metabolites for pre-data-acquisition normalization,³⁹ 6) mixing of equal moles of ¹²C-labeled samples and ¹³C-labeled pooled sample, 7) high-resolution RPLC-MS analysis of ¹²C-/¹³C-mixtures, 8) data processing including peak pair

picking and peak ratio measurement,^{34,71,74} and 9) metabolite identification based on the use of labeled standard library for positive identification⁷⁵ and the use of other compound libraries for putative identification.^{27,76}

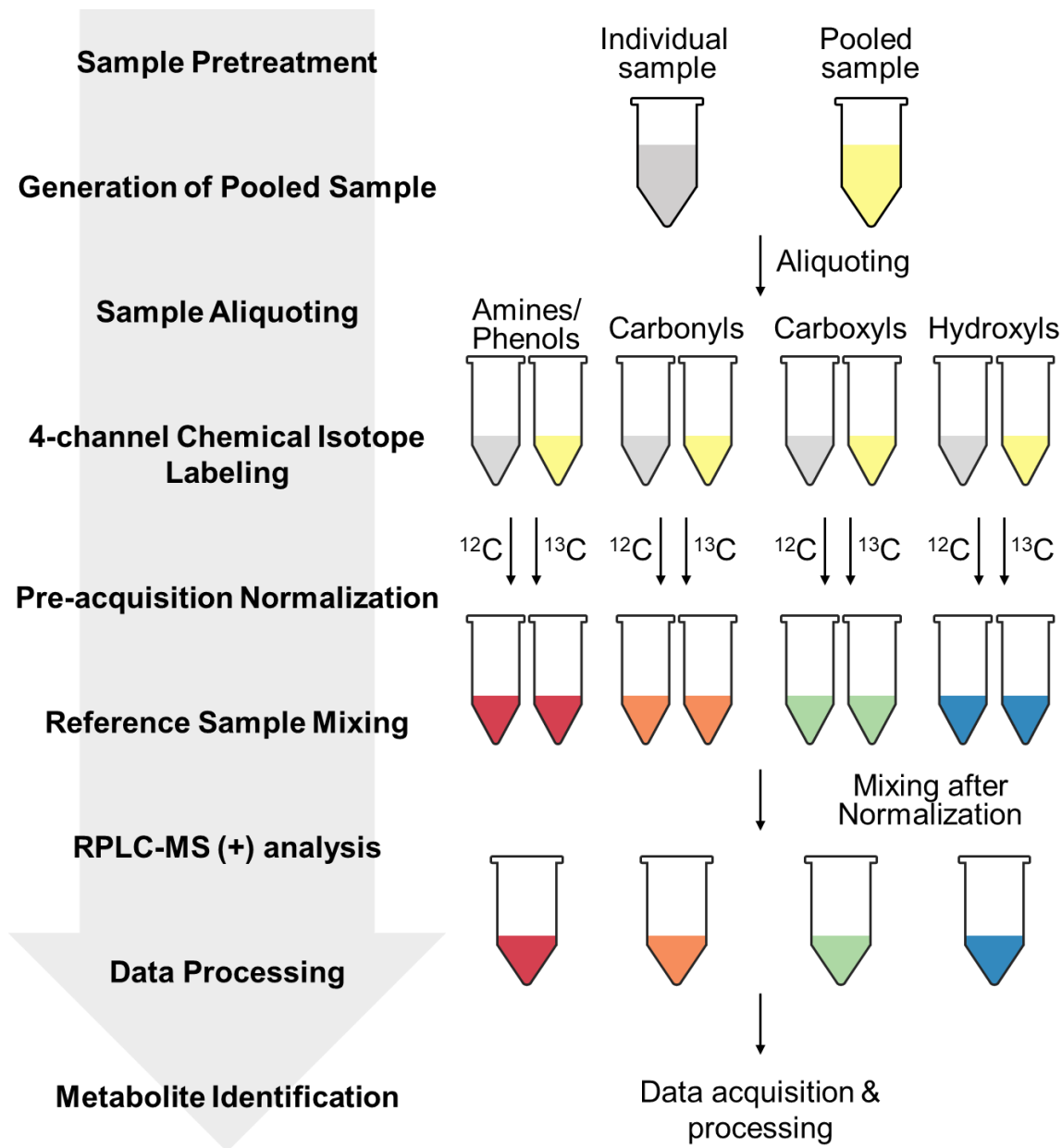


Figure 2.2 The workflow for metabolome analysis using 4-channel CIL LC-MS.

In this work, to demonstrate the performance of the combined analyses of the four submetabolomes, human plasma and yeast samples were labeled using the four chemistries in experimental triplicates. In this case, a sample was divided into two aliquots. One was labeled with ^{12}C -reagent and the other was labeled with ^{13}C -reagent, followed by mixing and LC-MS analysis. Supplemental Note N1 provides detailed information on sample preparation and labeling.

2.2.3 LC-MS analysis

The ^{12}C -/ ^{13}C -labeled mixtures from individual channels were analyzed using a Bruker Compact Quadrupole Time-of-flight (QTOF) mass spectrometer (Bruker, Billerica, MA) linked to UltiMate 3000 UHPLC (Thermo Scientific, MA). Supplemental Note N1 shows the LC-MS conditions used for the analysis. The injection volume for each channel was determined by injection amount optimization experiments (shown in Figure 2.2).

2.2.4 Data processing and metabolite identification

The resulting LC-MS data were processed using a set of in-house developed software (Supplemental Note N). Metabolite identification was carried out at three different levels of confidence, or three tiers, using IsoMS Pro software and database (Nova Medical Testing Inc., Edmonton, Canada). Positive identification as the first tier was based on accurate mass and retention time (RT) search against the labeled standard library currently composed of 1060 unique human endogenous metabolites, including 711 amines/phenols, 187 carboxyls, 85 hydroxyls and 77 carbonyls. The second tier identification was based on searching against Linked Identity (LI) Library containing metabolic-pathway-related metabolites (2,500 entries extracted from the KEGG database) with accurate mass and predicted RT information. These 2nd tier matches were

considered to be high-confidence putative identification. For the third tier, accurate masses of peak pairs were searched against compound entries in metabolome databases, resulting in putative matches; 5,506 entries in HMDB were searched for the plasma samples and 1,123 entries in YMDB were searched for the yeast cell samples. The remaining unmatched peak pairs were mass-searched against the predicted metabolome libraries (i.e., MCID one- and two-reaction libraries).

2.3 Results and Discussion

2.3.1 Group classification of database entries.

We selected some of the commonly used databases for chemical group classification, including MCID, HMDB, KEGG, YMDB and ECMDB. We applied the group classification program to examine the structures of database compounds and then classify them into different chemical groups. We were particularly interested in the hydroxyl, amine, phenol, carboxyl and carbonyl groups, as we have already developed the robust labeling methods for labeling these groups. These five groups are covered in four channels of submetabolome profiling: hydroxyl (H), amine/phenol (A), carboxylic acid (C) and ketone/aldehyde (K)-channel. Thus, using the combined results obtained from 4-channel CIL LC-MS, we could compare the group coverage of the experimental data vs. the database data.

Before we applied our group classification program to the compounds in all databases, we examined the classification accuracy using the relatively small database, YMDB, where manual checking of the program-generated classification results was manageable. Out of the 1107 filtered metabolites (i.e., yeast database entries minus the lipids, inorganic species and hydrocarbons), only 4 metabolites were misclassified, indicating an error of 0.4%. These 4 misclassifications were caused by wrong SMILES or resonance structure, as shown in Table 2.2. Thus, the program was deemed to be very accurate in classifying chemical structures into different chemical groups.

Table 2.2 Misclassified metabolites in YMDB.

YMDB ID	Reason	Comment
YMDB00151	wrong SMILES	Wrong SMILES in database: <chem>OC(C(=O)C(O)=O)C1=CC=CC=C1</chem> Correct SMILES from PubChem: <chem>C1C(C(OC1N2C=NC3=C2NC=NC3=O)CO)O</chem>
YMDB01481	wrong SMILES	Wrong SMILES in database: <chem>NC1=C(N=CN1C1OC(COP(O)(O)=O)C(O)C1O)C(O)=N</chem> Correct SMILES from YMDB website: <chem>NC(=O)C1=C(N)N(C=N1)[C@@H]1O[C@H](COP(O)(O)=O)[C@@H](O)[C@H]1O</chem>
YMDB00659	resonance structure	
YMDB00899	resonance structure	Uncommon resonant structure. Should be amide instead of imidic acid.

Figure 2.3 shows the classification results. In all the databases except HMDB, the hydroxyl or H-channel covers the highest percentage, i.e., 56.7%, 42.2%, 50.0%, 50.8% and 66.4% for MCID, HMDB, KEGG, YMDB and ECMDB, respectively. In contrast, the carbonyl (ketone/aldehyde) or K-channel covers the least, i.e., 22.8%, 20.6%, 24.9%, 20.1% and 18.5% for MCID, HMDB, KEGG, YMDB and ECMDB, respectively. In total, the four channels can cover 94.7% of the metabolites in the MCID database containing 2683 filtered metabolites. Similarly, for HMDB (5,506 filtered metabolites), KEGG (11,598), YMDB (1,107) and ECMDB (1,462), the 4-channel coverage is 85.7%, 86.4%, 85.7% and 95.8%, respectively. Lower percentages found in HMDB, KEGG and ECMDB correlate with increased percentages of the ester, amide and heterocycle groups. Note that, for the groups of ester, amide, heterocycle, organophosphorus, organosulfur and others shown in Figure 2.3, we used a sequential group-elimination approach to determine each percentage for clarify (i.e., 100% in total). For example, in Figure 2.3A, after eliminating 94.7% of the filtered metabolites (2,683) which can be analyzed by the 4-channel LC-

MS method, a small number of the remaining metabolites (0.4% of the total) contain the ester group. Thus, if an ester submetabolome profiling channel is developed in the future, it can only increase the overall metabolome coverage by 0.4%, assuming all hydroxyls, carboxyls, carbonyls, amines and phenols, including those also containing ester group, have already been covered by the 4-channel method. After eliminating the 4-channel metabolites and the ester group, a small number of the remaining metabolites (2.1%) contain the amide group. This elimination process applies to the remaining groups sequentially. These analyses indicate that, based on the current entries of the studied databases, very high coverage of the chemical space, ranging from ~86% to 96%, can be achieved using the 4-channel profiling approach.

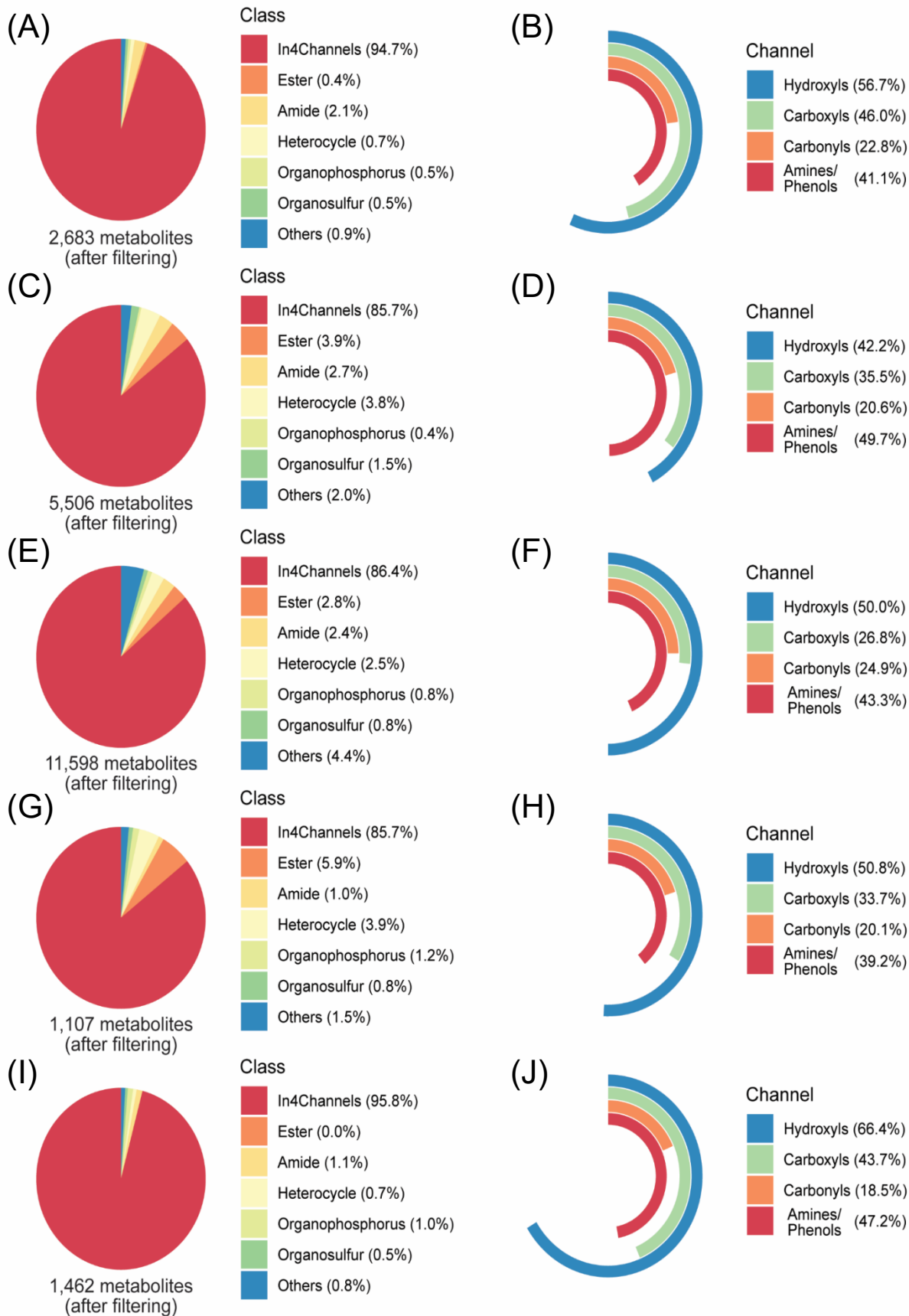


Figure 2.3 Classification of chemical groups of (A) MCID zero-reaction library, (C) HMDB, (E) KEGG, (G) YMDB and (I) ECMDB. Sequential class-elimination approach was used to determine the remaining groups (i.e., after removing all the 4-channel metabolites, a small number of the remaining metabolites contain the ester group. After removing 4-channel metabolites and ester-containing metabolites, a few remaining metabolites contain the amide group). Percent distributions of metabolites belonging to the four channels including overlapped metabolites with two or more functional groups in (B) MCID, (D) HMDB, (F) KEGG, (H) YMDB and (J) ECMDB.

2.3.2 4-Channel labeling LC-MS results.

Figure 2.2 shows a schematic of the 4-channel LC-MS approach. The labeling methods allow the conversion of metabolites not retainable in RPLC into relatively hydrophobic derivatives that can be efficiently separated using RPLC. In addition, chemical labeling allows the enhancement of ionization efficiency by ~10 to ~1000 folds.²⁶

Figure 2.4A shows the distribution of the absolute intensities of metabolite peak pairs detected from labeled plasma samples. Within the dynamic range of the instrument, there is a clear trend of an increase in the number of peak pairs detectable as the peak intensity decreases. Thus, using a highly sensitive CIL LC-MS method, we can increase the metabolome coverage significantly by detecting an increasing number of lower concentration metabolites. Interestingly, the four submetabolomes have similar distributions, indicating similar concentration distributions of these different groups of metabolites in plasma.

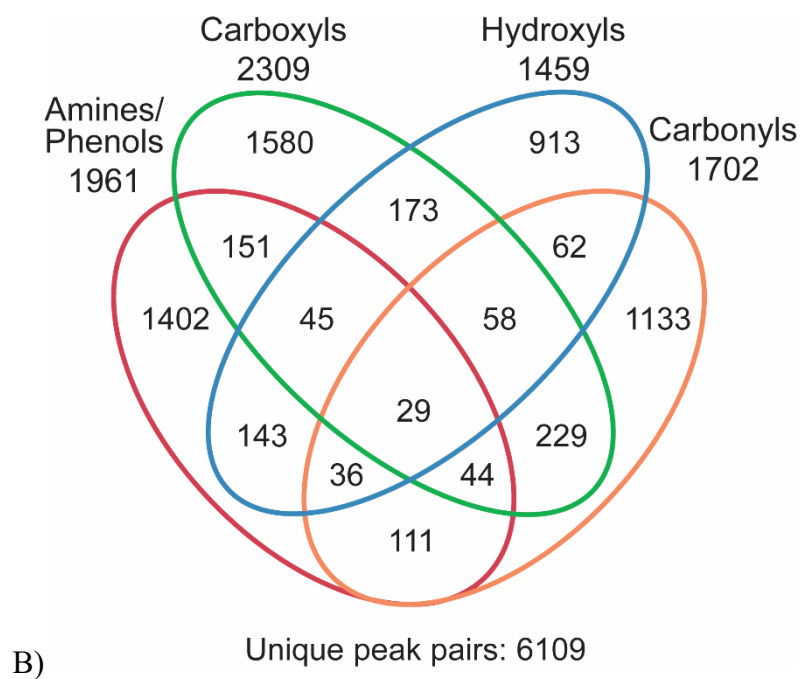
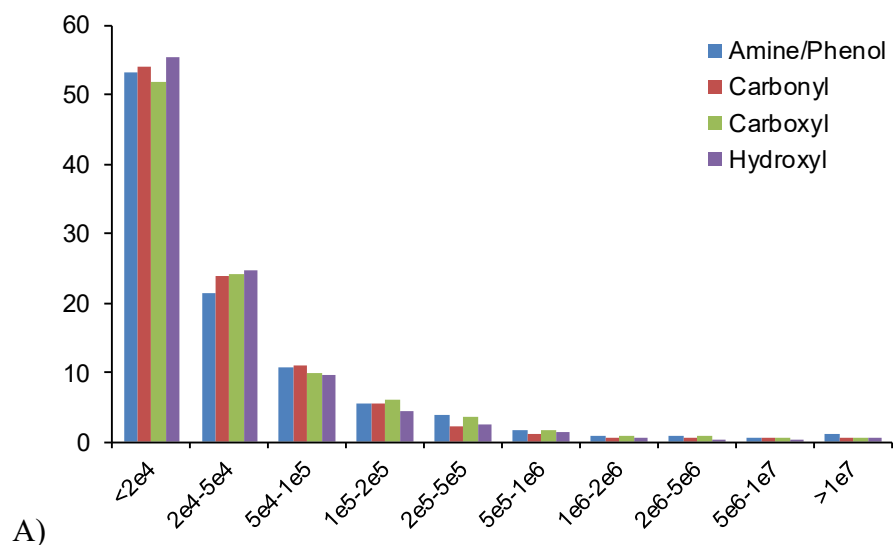


Figure 2.4 (A) Percentage of peak pair detected in 4-channel LC-MS analysis of plasma as a function of peak intensity. (B) Venn diagram of the numbers of peak pairs detected in four channels.

To further compare the number of metabolites detected in the four channels, Figure 2.4B shows the Venn diagram of the number of peak pairs detected in each channel. Because of lack of structure identities for many of the detected metabolites, in this comparison, we assumed that the

same metabolite was detected in two channels if the same accurate mass of the intact metabolite [i.e., mass of a labeled metabolite minus the mass of labeling tag(s)] was found in the two channels. For example, there are 29 peak pairs detected in all four channels as these peak-pair masses minus the mass of labeling tag(s) give the same mass; each one of them is deemed to be from the same metabolite that were detected four times. This is a conservative approach of determining the unique metabolites detected, as one would expect that metabolites with the same mass may have different structures (e.g., isomers; see below) and thus belong to different molecules. There are 1961, 2309, 1702 and 1459 peak pairs detected in the amine/phenol, carboxyl, carbonyl and hydroxyl submetabolome, respectively. If we only count the overlapped metabolites as one unique-mass metabolite, out of a combined total of 7431 peak pairs detected from the four channels, there are 6109 unique-mass peak pairs.

Many of the detected peak pairs can be identified or matched to metabolome databases. Table 2.3 shows the number of identified peak pairs from each channel in three tiers. For the plasma sample, out of the 7431 pairs detected, we positively identified 326 peak pairs based on accurate mass and retention time matches in tier 1. A few of the peak pairs could be matched to the same metabolite (e.g., Carbonyl_594, Carbonyl_635 and Carbonyl_657 matched to butanal). These matches were manually checked from the LC-MS data and are likely structural isomers of one chemical formula. This example suggests that using mass-match to filter out overlap peaks from two or more channels, as discussed above, might remove some same-mass metabolites with different structures. In tier 2 where authentic standards are not available, but accurate mass and predicted RT data are available in the Linked Identity (LI) library, we identified 344 peak pairs by mass and RT matches. Thus, a total of 670 peak pairs (9.0%) can be identified as high-confidence results (tier 1 and tier 2). In tier 3, the remaining peak pairs not identified in tiers 1 and 2 were

mass-searched against the HMDB, MCID-1R and MCID-2R libraries in sequence. There were 2628, 2851 and 777 peak pairs (35.4%, 38.4% and 10.5%) matched to the three libraries, respectively. In total, 6926 peak pairs (93.2%) were either identified or matched to databases. The remaining 6.8% of detected pairs may belong to metabolites that are not included in any of the searched databases.

Table 2.3 Summary of the number of peak pairs identified or matched against three different compound libraries from the human plasma samples analyzed using 4-channel LC-MS.

	A- channel	C- channel	H- channel	K- channel	Total per tier
Tier 1	208	54	23	41	326
Tier 2	71	157	68	48	344
Tier 3-HMDB	774	760	449	645	2628
Tier 3-MCID1R	570	1014	609	658	2851
Tier 3-MCID2R	168	242	165	202	777
Total per channel	1791	2227	1314	1594	6926

For yeast samples, similar approach was applied for peak pair detection and metabolite identification using 4-channel LC-MS. In total, we detected 5641 peak pairs, including 1747 from A-channel, 1867 from C-channel, 1006 from H-channel and 1021 from K-channel (shown in Figure 2.5). After filtering the same-mass metabolites detected in two or more channels, we have 4955 unique-mass peak pairs. Table 2.4 summarizes the identification results. From the 5641 peak pairs detected, 243 and 188 peak pairs were identified in tier 1 and tier 2, respectively. Thus, a total of 431 peak pairs (7.6%) can be identified as high-confidence results (tier 1 and tier 2). For tier 3 matches, we found 880, 3442 and 514 peak pairs (15.6%, 61.0%, 9.1%) matched to YMDB,

MCID-1R and MCID-2R libraries, respectively. In total, 5267 peak pairs (93.3%) were either identified or matched to databases.

Table 2.4 Summary of the number of peak pairs identified or matched against three different compound libraries from the yeast cell samples analyzed using 4-channel LC-MS.

	A-channel	C-channel	H-channel	K-channel	Total per tier
Tier 1	123	68	20	32	243
Tier 2	69	80	18	21	188
Tier 3-YMDB	297	261	163	159	880
Tier 3-MCID1R	944	1255	610	633	3442
Tier 3-MCID2R	99	155	116	144	514
Total per channel	1532	1819	927	989	5267

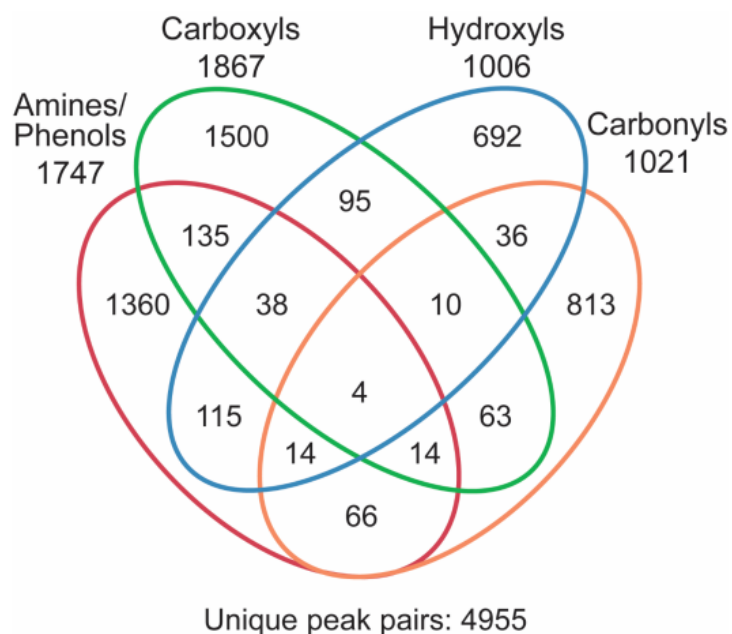


Figure 2.5 Venn diagram of the numbers of peak pairs detected in four channels in yeast samples.

It should be noted that the peak intensity ratios measured in the triplicate analysis of 1:1 ^{12}C -/ ^{13}C -labeled plasma can be used to gauge the accuracy and precision for relative quantification of this particular mixture. When plotting the distribution of peak pairs detected as a function of the average peak ratio and their RSD (Figure 2.6), most of the peak pairs in four submetabolome profiling gave the ratio value close to the expected ratio of 1.0, demonstrating high accuracy. The RSD values are less than 20% for more than 95% of the pairs with an average RSD of 5.1% and thus the analytical precision was also very high. We note that we did not study the interday and intraday repeatability in this work. However, all the individual labeling methods have been used in a number of published metabolomics studies where quality control (QC) samples were used to gauge interday and intraday repeatability over a number of days. QC samples were clustered tightly, indicating excellent repeatability.¹ The linearity of peak ratio measurement has been addressed in previously reports such as the original paper published on dansylation labeling for amine/phenol submetabolome profiling.²⁶ Over 100-fold relative changes could be measured.²⁶

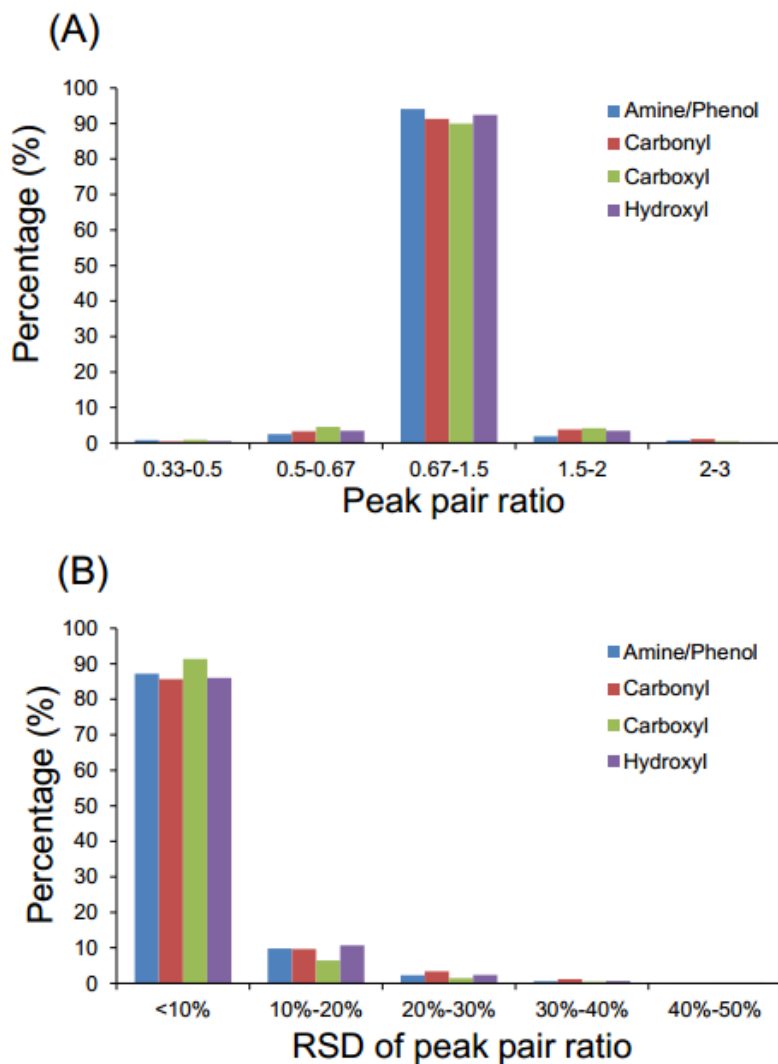


Figure 2.6 Distributions of peak pair numbers as a function of (A) averaged peak ratio and (B) RSD. Data are presented as mean \pm S.D. from experimental triplicate and injection duplicate (n=6).

2.3.3 Overlaps of multi-functional metabolites.

Figure 2.7 shows the Venn diagrams of the numbers of database metabolites belonging to individual channels and overlaps among different channels. There are clearly many metabolites belonging to two or more channels. Taking the MCID database as an example (Figure 2.7A), there are five metabolites in all channels. Most of the overlaps occur for metabolites containing two

function groups. However, in our 4-channel LC-MS results, most of the metabolites were uniquely detected in only one channel.

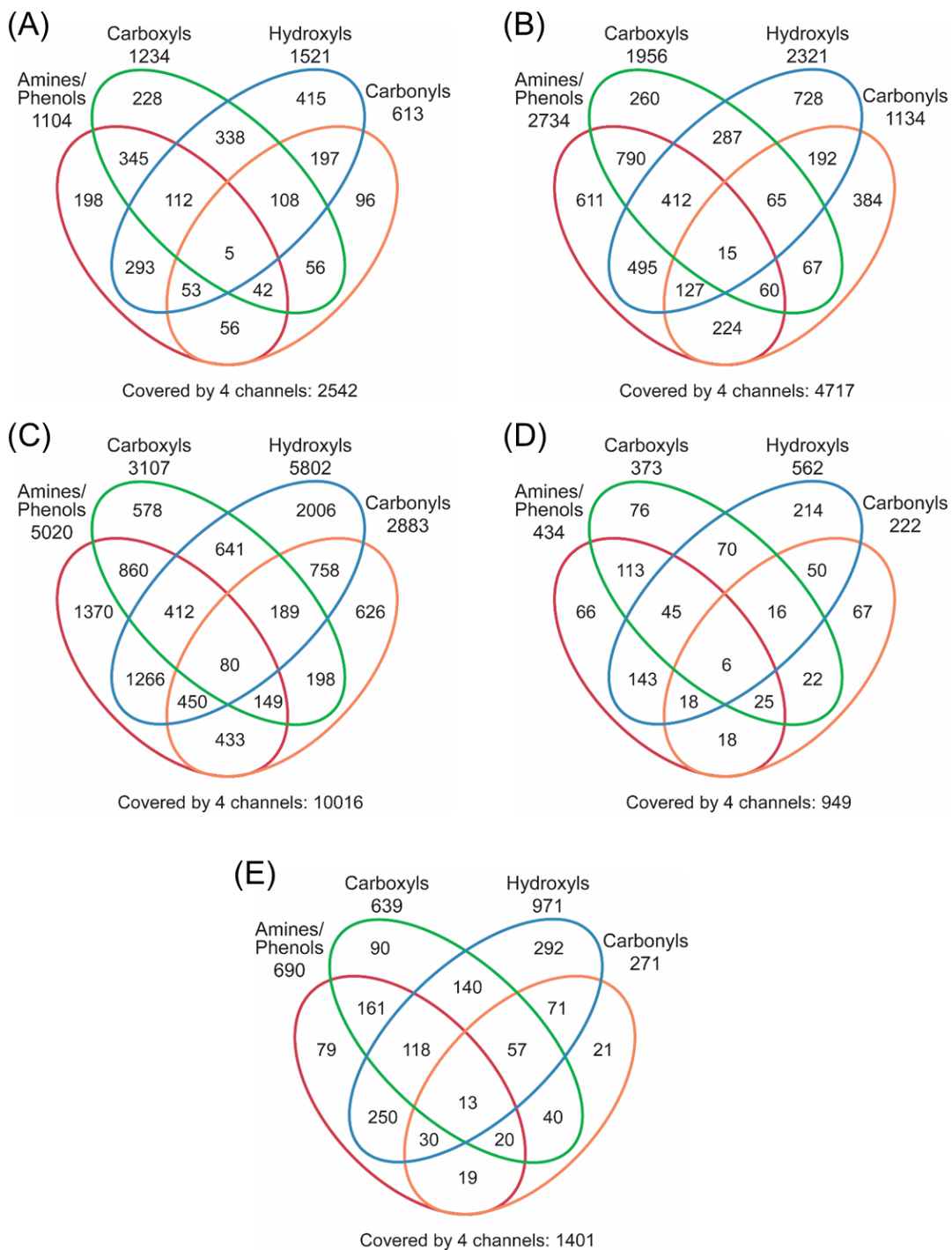


Figure 2.7 Venn diagram of the numbers of metabolites in four channels from the compound entries in (A) MCID, (B) HMDB, (C) KEGG, (D) YMDB and (E) ECMDB.

Much smaller overlaps found can be attributed to the fact that, by design, we developed the 4-channel methods with due consideration of minimizing redundant analyses of the same metabolite in different channels. For example, in the analysis of the carboxyl acid submetabolome, we used 6 M HCl to acidify the sample, followed by organic solvent extraction of the acids. The amines and some phenols would be positively charged at this low pH and thus not be extracted by the organic solvent. This can be inferred from the analysis of amino acids; all 20 amino acids can be readily detected in the amine/phenol channel, but only 2-3 can be detected in the carboxylic acid channel. Similarly, for the analysis of the hydroxyl submetabolome, we used an organic solvent to extract the neutral metabolites containing hydroxyl groups from the highly acidified sample. In the analysis of the carbonyl submetabolome, the labeling solution is acidic under which dansylhydrazine preferentially reacts with neutral metabolites containing carbonyl groups. The charged species such as metabolites containing both amine and carbonyl groups may not react with dansylhydrazine. Another contributing factor might be related to the changed reactivity of a functional group in metabolites with two or more groups. For example, we found that several keto-acids with carbonyl and carboxyl groups conjugated together (e.g., oxaloacetic acid and acetoacetic acid) are difficult to be labeled in the carbonyl channel.

2.3.4 Group under-representation.

While the compound entries in a database are by no means perfect in terms of coverage (i.e., not including all metabolome compounds of an organism) and trueness (i.e., the presence of false entries), the group classification shown in Figure 2.3 for several databases gives consistent

group distributions. For example, the hydroxyl group is by far the largest group, except in HMDB where it is the second largest. However, in our experimental dataset, for both plasma and yeast samples, the number of hydroxyl-containing metabolites detected is smaller than the other groups. This suggests that the overall coverage achieved by the current 4-channel experiments is lower than the database-derived theoretical coverage; the exact percentage of reduction is unknown. It appears that the sample preparation workflow or labeling reaction of the hydroxyl channel is not fully optimized. More development work should be devoted to optimizing the hydroxyl submetabolome profiling.

2.4 Conclusions

After filtering out the lipids, inorganic species and hydrocarbons that are not targeted for analysis by CIL LC-MS, we found that 86% to 96% of the metabolites in the studied databases contain one or more of the five functional groups: amine, phenol, hydroxyl, carbonyl and carboxyl. Thus, in-depth profiling of these chemical groups can generate a very high coverage of the metabolome. We described a 4-channel CIL LC-MS approach to analyze the hydroxyl (H), amine/phenol (A), carboxyl (C) and carbonyl (K) submetabolomes, separately.

For future work, we will need to optimize the current method for hydroxyl submetabolome profiling and develop labeling methods to analyze other groups of metabolites currently not covered by the 4-channel approach (e.g., esters and amides). We note that the compound entries in a current database may under- or over-represent certain groups of metabolites. For example, there may be the intermediate compounds of known metabolites that have not been documented, as evident from the mass-matches of many predicted metabolites from one or two metabolic reactions of known metabolites (i.e., MCID-1R and MCID-2R). As our knowledge of metabolites expands with the detection and identification of known unknowns and unknown unknowns, we

will surely increase the coverage and reduce the false entries in a metabolome database of an organism. We envisage that the 4-channel LC-MS approach, with perhaps additional channels, will play an important role in expanding our knowledge of chemical composition of a metabolome.

Chapter 3 Retention time prediction of chemical isotope labeled tripeptides on RPLC using machine learning methods

3.1 Introduction

Short peptides play an important role in biological processes⁷⁷ and drug discovery⁷⁸. They can act as signal molecules through binding to DNA or DNA-related proteins.⁷⁹ Stem cell studies have demonstrated that short peptides, Glu-Asp-Pro and Lys-Glu-Asp, can regulate the proliferation of embryonic and immortalized cells as inhibitors. Meanwhile, Glu-Asp-Pro can stimulate the proliferation of normal lymphocytes.⁷⁷

LC-MS, widely used in metabolomics for profiling complex biological samples, can be employed for short peptide acquisition. Coupled with chemical isotope labeling, the detection of short peptides can be achieved to very low concentrations. At the same time, the identification of short peptides becomes an issue to be resolved. Information from LC-MS, such as exact mass, retention time, and MS/MS spectra, can support identification. To obtain accurate results, retention time and MS/MS spectra should be acquired at the same experimental conditions used to build the in-house library as reference. But considering the seq diversity of short peptides, gathering all standards of short peptides becomes costly. In-silico prediction of either RT or MS/MS can assist identification.

The amino acid composition and chemical descriptors^{80, 81} of peptides have been used for RT prediction. Machine learning methods were also involved in this processing. For example, ELUDE adopts a support vector regression model trained by 60 features of peptides. These features include modified retention coefficients, peptide length, number of occurrences of amino acid sharing specific properties.⁸² ELUDE in the following version uses a two-step method to predict RT of post-translationally modified peptides. RT index of modified and unmodified amino acids

was calculated in the first step using linear SVR. Then features derived from the RT index were used in the second SVR to predict RT for peptides of interest.⁸³ Artificial neural network-based method was also used to predict RT with the composition of amino acids.⁸⁴ And the chemical descriptors were further added to the model to improve accuracy.⁸⁵ DeepRT employs a deep learning model using 20D embedding vectors representing amino acid composition.⁸⁶

In this work, to predict RT of chemical isotope labeled tripeptides, we built models using support vector regression based on chemical descriptors and labeling information. 360 dipeptides were used for building the models and cross validation. The models were further validated by tripeptide standards and tripeptide mixture. Since the chemical isotope labeling alters the physical and chemical properties of short peptides, the chemical descriptors calculated from labeled and unlabeled peptides were compared in RT prediction. Through validating with experimental tripeptide data, the mean of RT difference was achieved at 16.1 s using the chemical descriptors calculated by labeled tripeptides. 329 and 528 tripeptides can be putatively identified in serum and urine samples, respectively.

3.2 Methods

3.2.1 Chemicals and Reagents

Organic solvents for the mobile phase and sample preparation were purchased from Thermo Fisher Scientific (Waltham, MA). ¹²C-DnsCl derivatization reagent and ¹³C-DnsCl derivatization reagent were made in-house according to our previous studies⁸⁷. The 399 dipeptide standards and 10 tripeptide standards were synthesized by LifeTein (Somerset, NJ). XXA tripeptide mixture was purchased from GenScript (Piscataway, NJ). And around 5 mg standards were measured and dissolved in ACN/H₂O (1:1, v/v) to make stock solutions.

3.2.2 Sample preparation and derivatization

For the urine samples, 75 μL of lyophilized human urine standard was diluted with H_2O to 300 μL . For serum samples, 30 μL of serum was mixed with 90 μL H_2O and kept in -20°C fridge for 1 hour for protein precipitation. After protein precipitation, it was centrifuged for 15 min and the supernatant was collected and dried by a nitrogen blower. The dried serum sample was resuspended in 25 μL of H_2O for DnsCl derivatization reaction. The derivatization reaction was conducted to label the amine group on the N terminal of dipeptides according to a protocol developed in our lab with slight modification. In brief, 25 μL of diluted urine samples or resuspended serum samples were mixed with 12.5 μL of 250 mM $\text{Na}_2\text{CO}_3/\text{NaHCO}_3$ buffer. The solution was vortexed, spun down and mixed with 37.5 μL freshly prepared ^{12}C -DnsCl (18 mg/mL in ACN, for light labeling) or ^{13}C -DnsCl (18 mg/mL in ACN, for heavy labeling), followed by vortexing and spinning down. The mixture was incubated in an oven at 40°C for 45 min. Then, the incubated solution was mixed with 7.5 μL NaOH solution (250 mM in H_2O) to quench the excess ^{12}C -DnsCl or ^{13}C -DnsCl at 40°C for 10 min. Lastly, the quenched solution was mixed with 30 μL of 425 mM formic acid in ACN/ H_2O (1:1, v/v) to consume the excess NaOH. The ^{12}C -DnsCl and ^{13}C -DnsCl derivatized products were equally mixed, and the mixture was centrifuged for 10 min at 12000 rpm before LC-MS analysis.

3.2.3 In-house database of dipeptides

20, 40, or 60 dipeptides were mixed and derivatized with ^{12}C -DnsCl with the same procedure as sample derivatization for an in-house database. For missing derivative dipeptides in each mixed dipeptide solution, the corresponding individual standard was derivatized with ^{12}C -

DnsCl and collected by the retention time (RT). The RT of each dipeptide was checked manually by matching its theoretical m/z.

3.2.4 HPLC-MS condition

A Vanquish UHPLC Systems (Thermo Scientific) coupled with a Q Exactive HF Orbitrap Mass Spectrometers (Thermo Scientific) was used for DnsCl derivatized sample and dipeptide standards analysis. An Agilent reversed phase C18 column (100 × 2.1 mm, 1.8 mm particle size, 95 Å pore size) was used for separation and the column was maintained at 40 °C. Mobile A consisted of 0.1% FA in HPLC grade water and mobile phase B was consisted of 0.1% FA in HPLC grade ACN. A total of 18 min gradient was set as follows: 25% B increased to 99% in 10 min and maintained for 5 min, followed by 3 min equilibrium. The flow rate was set to 0.4 mL/min. The Q Exactive HF mass spectrometer was operated under an ESI positive mode for both in-house database collection and sample analysis. Electrospray ionization parameters were as follows: the spray voltage was at 3.5 kV; the capillary temperature at 320 °C; probe heater temperature at 360 °C; Full mass scan mode (m/z 220–1800) was used at a resolution of 120k at m/z 200 with around 1.2 Hz scan rate. The automatic gain control (AGC) target was at 1×10^6 ions with 200 ms maximum ion injection time. The sample injection amount was optimized according to the maximum peak pair numbers at different injection volumes. For derivatized standards, the injection volume is 2 µL.

3.2.5 Data processing

The original data of the resulting LC-MS data of serum and urine sample was firstly converted to “.text” format with MSConvert GUI tool⁸⁸. The converted data containing ¹²C/¹³C peak pair information were further processed using IsoMS Pro software (Nova Medical Testing

Inc., Edmonton, Canada)^{89 34}. Dipeptide identification in serum and urine was carried out with the in-house database of dipeptides.

3.2.6 SMILES generation

The workflow of dipeptide and tripeptide retention time (RT) prediction is shown in Figure 3.1. First, the simplified molecular-input line-entry system (SMILES) files, using line notions to describe the chemical structure, of dipeptides and tripeptides without chemical labeling were generated. The SMILES started from the N terminal to the C terminal. An example SMILES of the dipeptide, alanyl alanine (AA), is shown in Figure 3.2. Then, the SMILES of the labeled dipeptides and tripeptides were further generated. Since the dansyl chloride chemical labeling reaction targets amine and phenol groups, the N terminals, side chain of tyrosine (Y), lysine (K), and histidine (H) can be labeled. Peptides without Y, K, or H can only be labeled by one tag on the N terminal. The structure and SMILES of labeled AA is shown as an example in Figure 3.2. For peptides containing Y or K, both N terminal and side chain can be labeled. The structure and SMILES of labeled YA and KA is demonstrated in Figure 3.2. The N-H on the side chain of histidine can only be partially labeled, and the side chain with and without labeling were both observed. An example of HA labeled by one tag or two tags was shown in Figure 3.2.

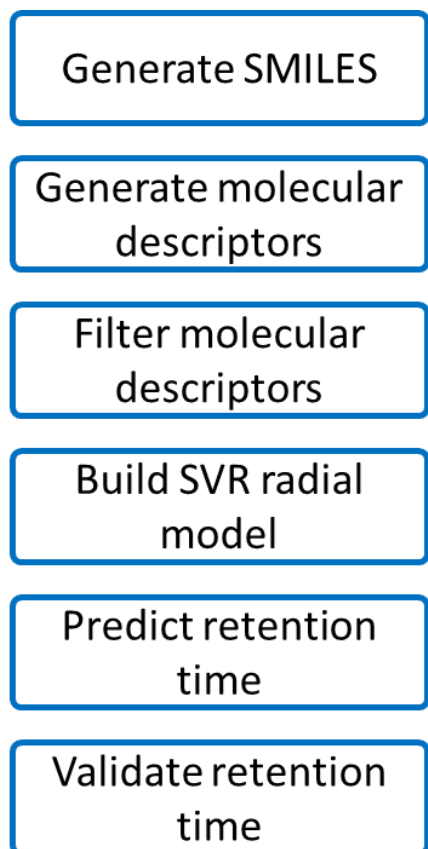


Figure 3.1 Workflow of dipeptides and tripeptides RT prediction.

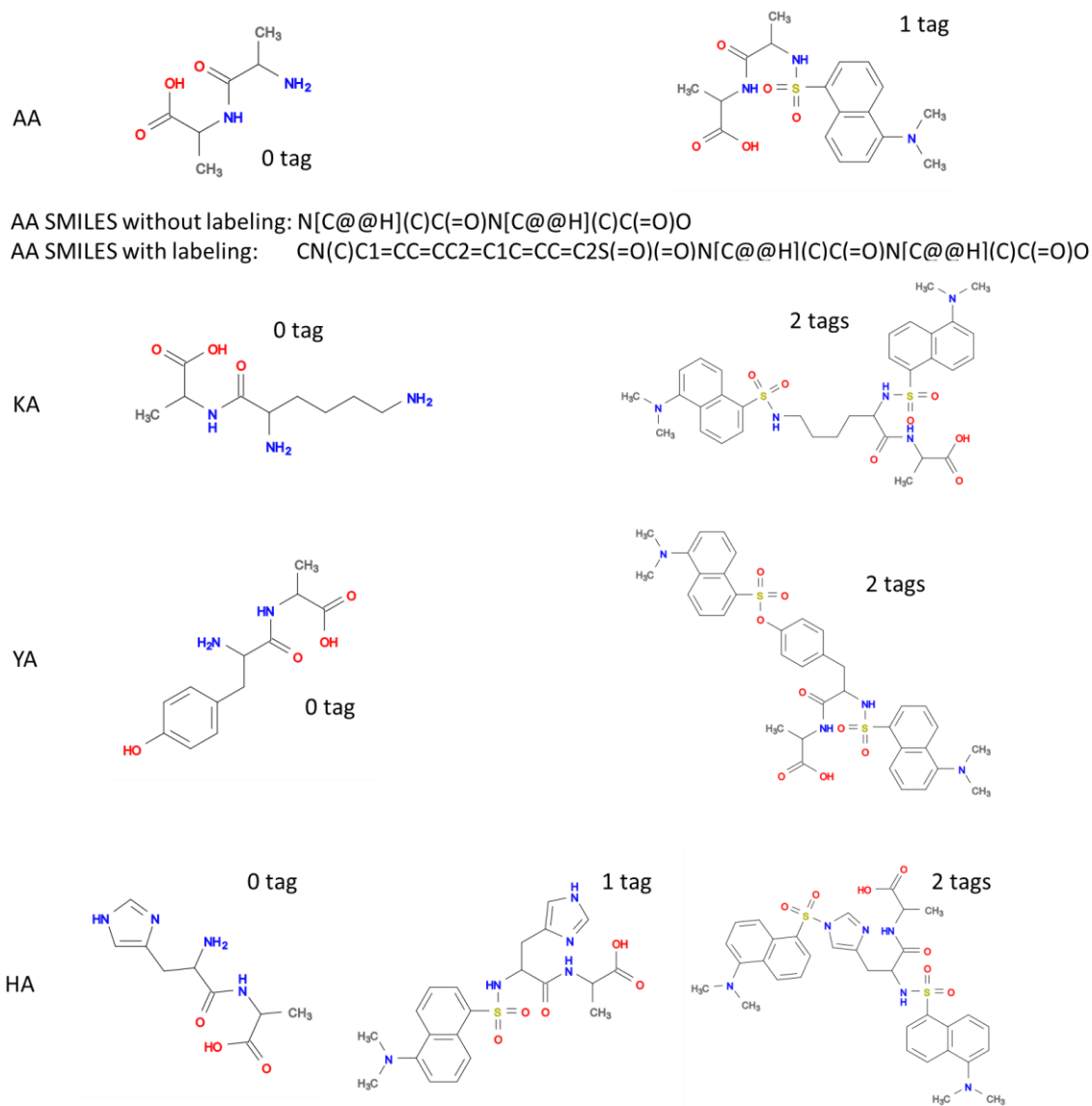


Figure 3.2 Structure of unlabeled and labeled dipeptides.

3.2.7 Molecular descriptor calculation and filter

After generating the SMILES of dipeptides and tripeptides, the SMILES was used for calculating the exact mass of short peptides with and without labeling with R package ChemmineR (version 3.34.1)⁹⁰. Then, the SMILES was treated as input to calculate molecular descriptors. A R package, rcdk⁹¹ (version 3.4.9.1), was used to generate molecular descriptors. Then, molecular

descriptors containing NA, undifferentiated molecular descriptors, and highly correlated molecular descriptors were filtered.

For each SMILES of labeled and unlabeled short peptides, 287 molecular descriptors were calculated. The molecular descriptors of labeled and unlabeled short peptides were processed separately. First, molecular descriptors containing NA were filtered. Then, undifferentiated molecular descriptors were also removed as they provide the same information. At last, highly correlated molecular descriptors ($R^2 > 0.8$) were further excluded. The remained molecular descriptors were ready for building models.

3.2.8 Building SVR radial models

Apart from the molecular descriptors for the last step, the occurrence number of each amino acid, the exact mass of peptides, and the labeled tag number were also input as features for building the support vector regression (SVR) model using caret (version 6.0-84) package on R. The dipeptide dataset was divided into 80% training set and 20% test set. The radial kernel was selected. For the training set, recursive backward elimination was applied for feature selection. And 7-fold cross validation repeated 3 times was used for tuning parameters, sigma and C, ranging from 2^{-5} to 2^5 . Top 20 features were selected for building models based on the root-mean-square error (RMSE) of experimental RT and predicted RT from the dipeptide training set and test set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x}_i)^2}{N}}$$

3.3 Results and Discussion

3.3.1 Overview of dipeptide experimental data and SMILES, tripeptides SMILES

For the dipeptide dataset, experimental RT of 397 peaks from 360 dipeptides were incorporated into the following analyses, as histidine can be either labeled or unlabeled. For peptide RT prediction, the side chain of cystines can form dimers via SS-bonds, leading to different physical and chemical properties from the rest of the peptides. Cystine-related peptides were removed from RT prediction. An overview of dipeptides for experiment and tripeptides for prediction were listed in table 3.1. The experimental RT of QF was missing. The tripeptide SMILES is consisted of all possible combinations, including fully labeled N-terminal, Y, K and partially labeled H.

Table 3.1 The overview of dipeptide and tripeptide data.

	1 tag	2 tags	3 tags	4 tags	Sum
Experimental dipeptides	288	100	9	NA	397
Predicted dipeptides	289	100	9	NA	398
Predicted tripeptides	4913	2434	444	27	7818

3.3.2 Comparison of experimental RT and predicted RT for the training set and validation set

For the dipeptide dataset, 316 (80%) dipeptides were used for the training model. After tuning parameters through cross validation, sigma was selected as 2^{-5} and C as 2^3 , resulting in the lowest RMSE in cross validation. The RMSE of the training set and test set, summarized in Table

3.2, reached 0.158 and 0.249, respectively. The optimization result is shown in Figure 3.3. A comparison of the experimental RT and predicted RT of the training set was illustrated in Figure 3.4a. Most dipeptides are well-aligned on the diagonal. The RT difference is ranging from 0.0029 min to 0.6105 min, and the histogram of RT difference is shown in Figure 3.4b. We also compared the experimental RT and predicted RT of the test set, shown in Figure 3.4c and 3.4d. Overall, the difference between the experimental RT and predicted RT of dipeptides met the identification requirement.

Table 3.2 The RMSE summary of model performance on a different dataset.

	RMSE	
	Labeled SMILES	Unlabeled SMILES
Dipeptide training set	0.158	0.146
Dipeptide test set	0.249	0.240
Tripeptide standards	0.182	1.012
Tripeptide mixture	0.380	0.814

To investigate whether the SMILES of labeled short peptides is necessary for predicting RT, we used the SMILES of unlabeled dipeptides to calculate the chemical descriptors and predicted RT for labeled dipeptides. The same workflow was followed as the previous model. After cross validation in the training set, the parameters sigma and C were chosen as 2^{-5} and 2^5 (shown in Figure 3.3b), respectively. Similarly, the predicted RT of the training set and test set using unlabeled SMILES are comparable to experimental RT, shown in Figure 3.4e and 3.4g. The distribution of RT difference for both training set (Figure 3.1f) and test set (Figure 3.1h) is also similar to the distribution using labeled SMILES. The RMSE is 0.146 and 0.240 for the training

set and test set, respectively, close to the RMSE of the previous model. As a result, no clear difference was observed from models built by chemical descriptors calculated by labeled or unlabeled SMILES.

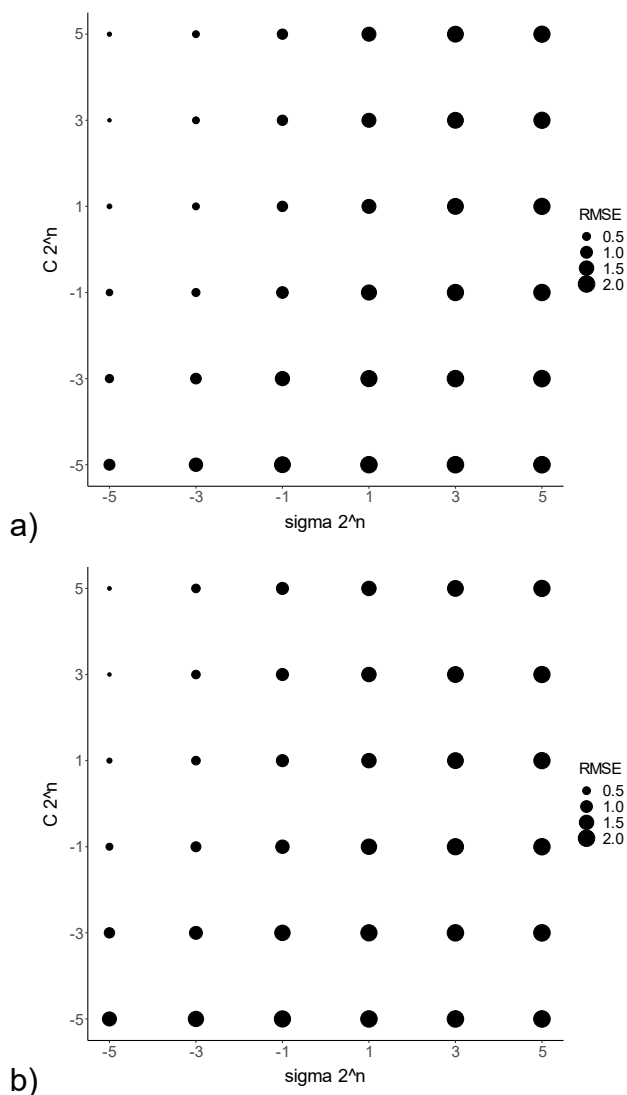
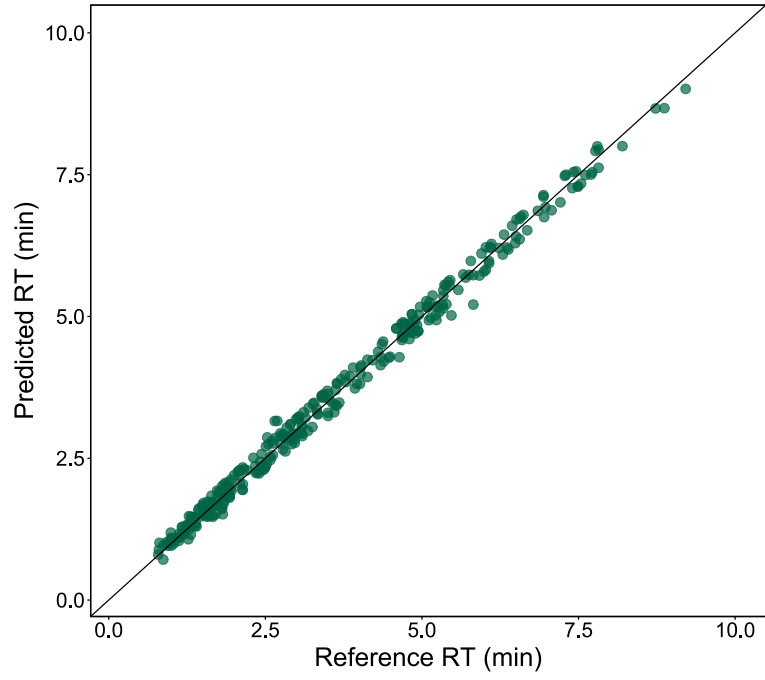
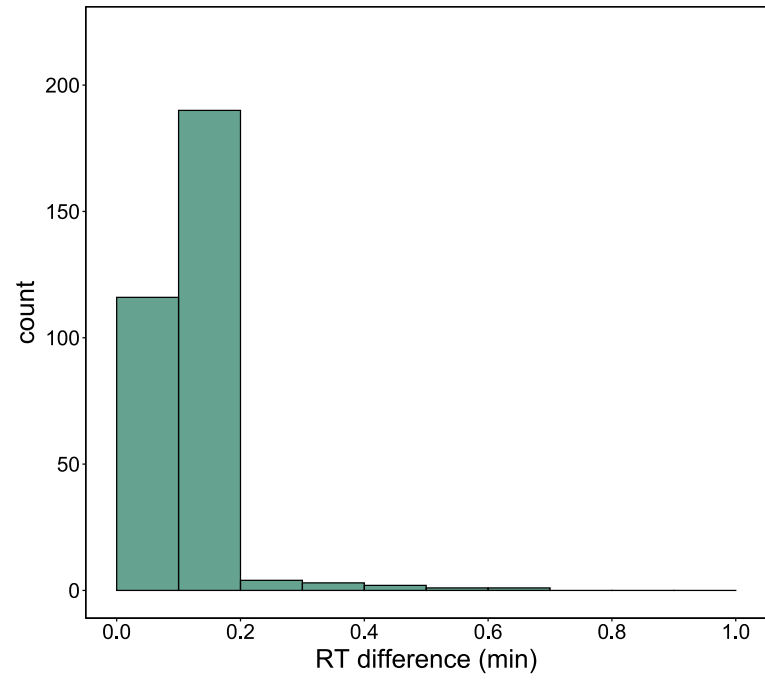


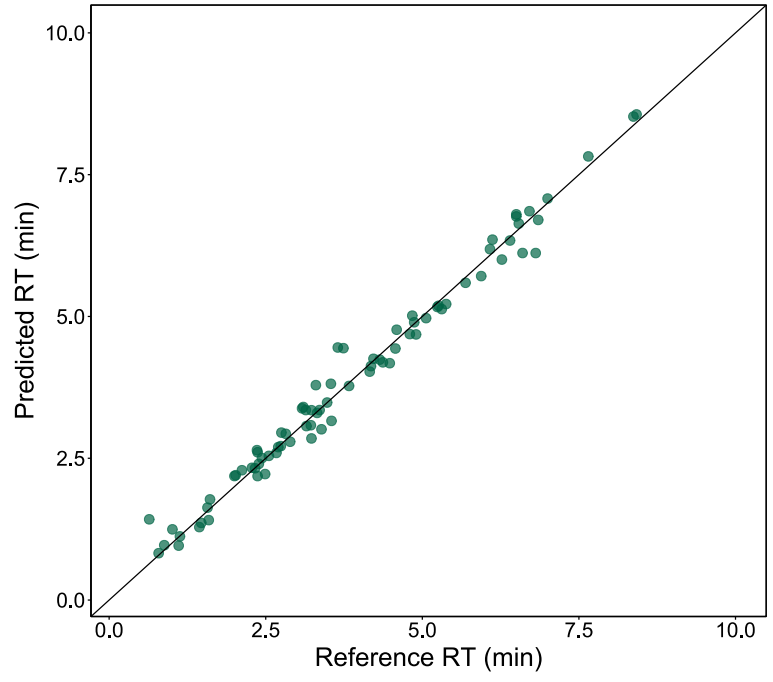
Figure 3.3 a) The optimization result of models built by top 20 features generated by labeled SMILES. b) The optimization result of models built by top 20 features generated by unlabeled SMILES. The smaller dot size represents a smaller RMSE.



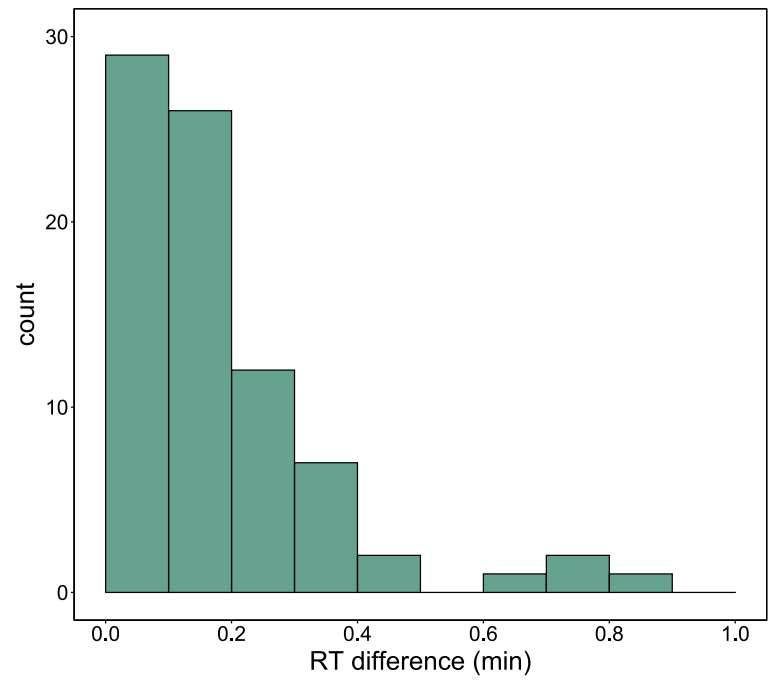
a)



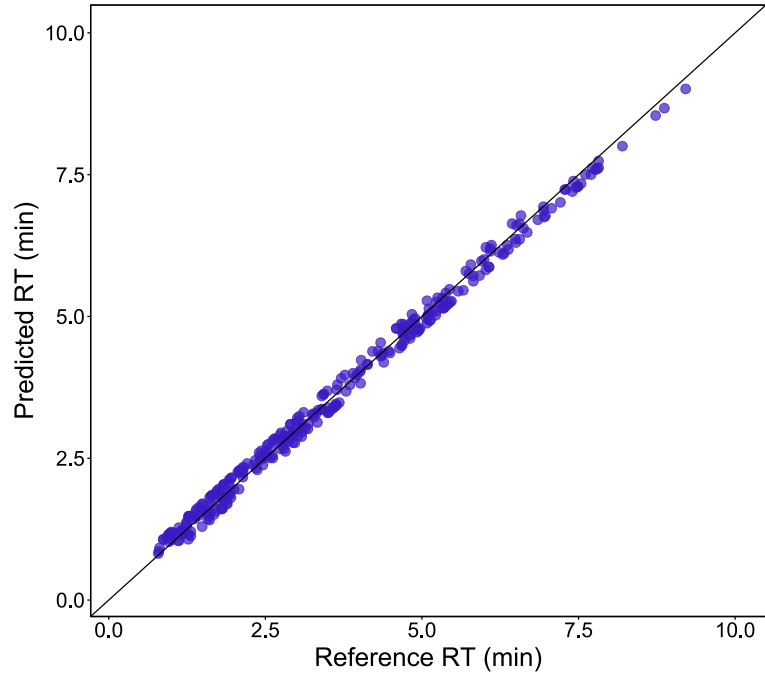
b)



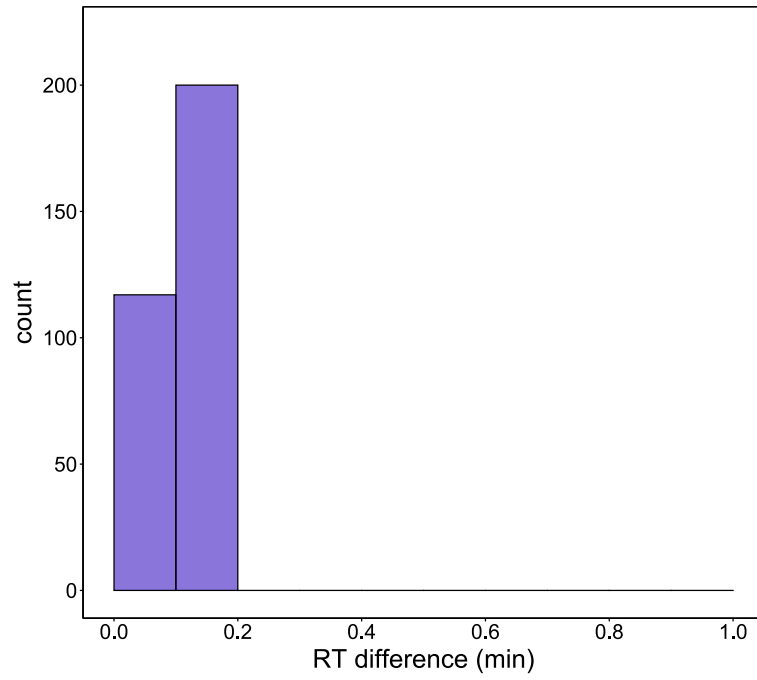
c)



d)



e)



f)

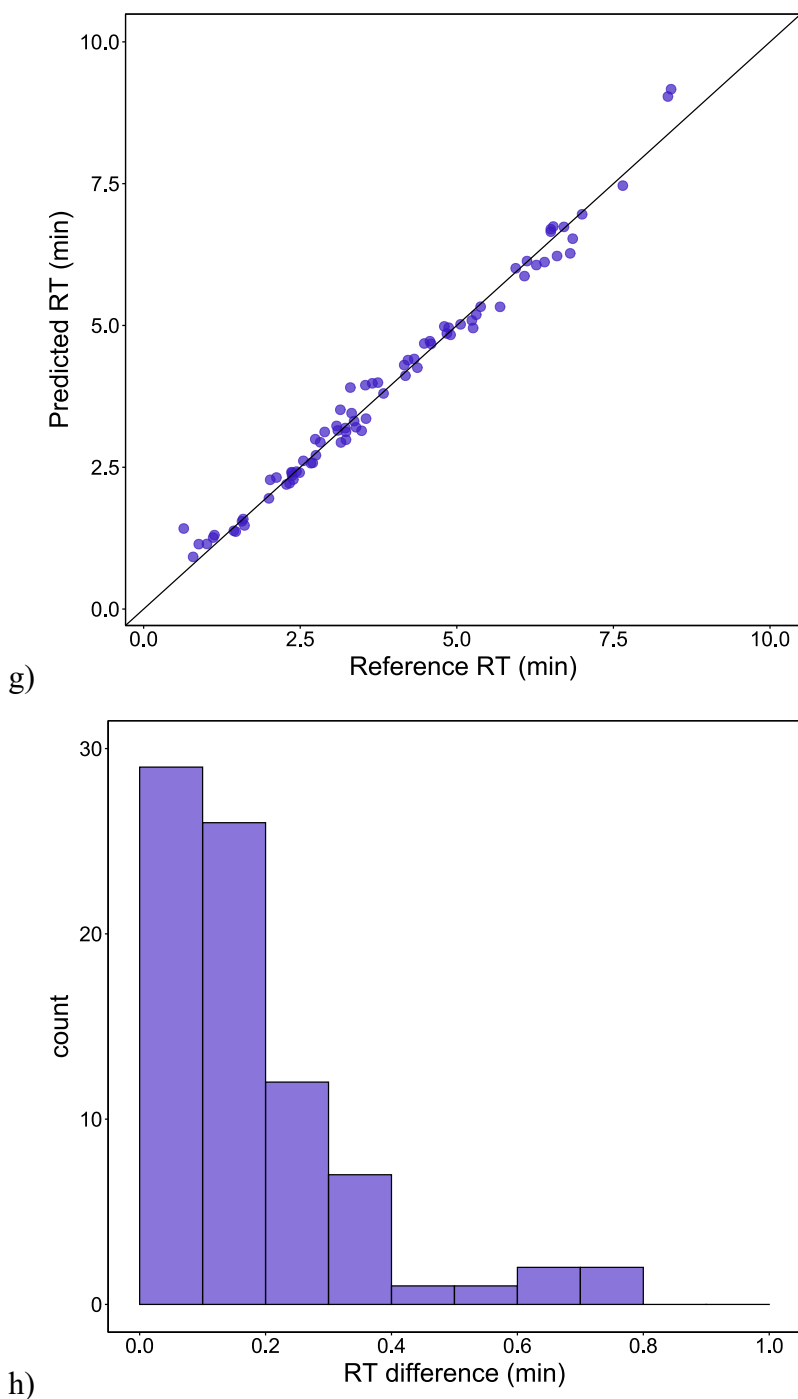


Figure 3.4 a) The comparison of experimental RT and predicted RT of dipeptide training set using labeled SMILES. b) The histogram of RT difference in the training set using labeled SMILES. c) The comparison of experimental RT and predicted RT of dipeptide test set using labeled SMILES. d) The histogram of RT difference in the test set using labeled SMILES. e) The comparison of experimental RT and predicted RT of dipeptide training set using unlabeled SMILES. f) The histogram of RT difference in the training set using unlabeled SMILES. g) The comparison of experimental RT and predicted RT of dipeptide test set using unlabeled SMILES. h) The histogram of RT difference in the test set using unlabeled SMILES.

3.3.3 Comparison between the experimental RT and predicted RT of 10 tripeptide standards

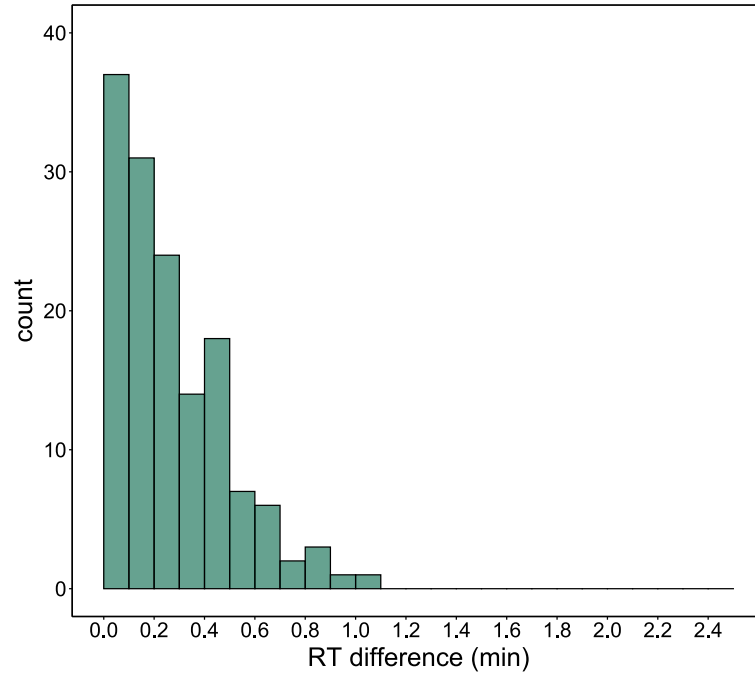
After building and validating the models using dipeptide data, we applied the models to tripeptides. The RT of 7818 tripeptides were predicted. To test whether the models fit tripeptides, ten tripeptide standards were synthesized covering 12 different amino acids and both one-tag and two-tag labeling. RT of tripeptide standards was acquired at the same experimental conditions. The predicted RT from models using labeled and unlabeled SMILES was compared with experimental RT, listed in Table 3.3. The RT difference predicted by labeled SMILES ranges from 0.04 min to 0.30 min. In contrast, the maximum RT difference predicted with unlabeled SMILES is 1.78 min, much larger than 0.03 min. In the model built by unlabeled SMILES, RT difference of all tripeptides is larger than using labeled SMILES.

Table 3.3 The comparison of tripeptide experimental RT and predicted RT.

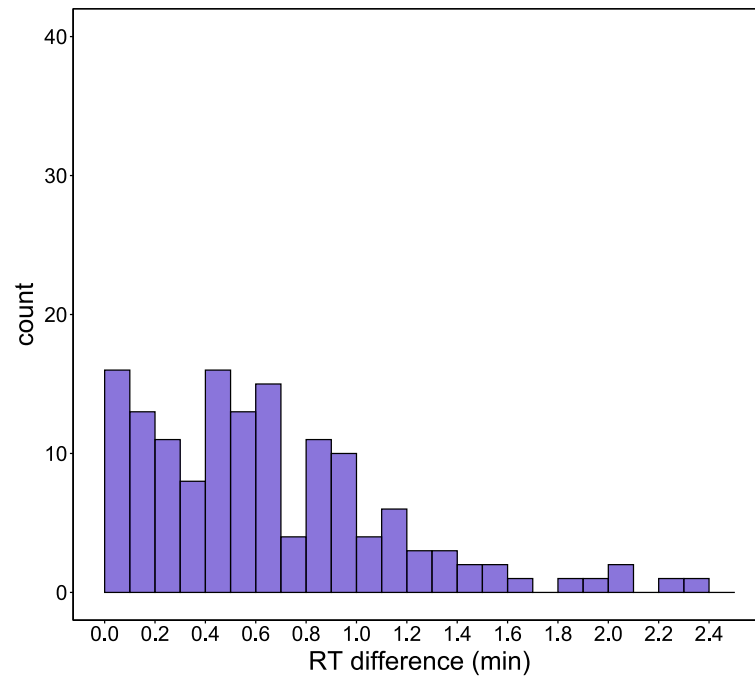
Experimental information						Labeled SMILES		Unlabeled SMILES	
	Tripeptide sequence	Tag number	Charge number	m/z	Experimental RT (min)	Predicted RT (min)	RT difference (min)	Predicted RT (min)	RT difference (min)
1	IGA	1	1	493.2115	3.27	3.47	0.20	3.47	0.20
2	MPS	1	1	567.1941	3.03	3.07	0.04	3.69	0.66
3	KMW	2	2	465.671	6.63	6.70	0.07	4.85	-1.78
4	MFL	1	1	643.2618	5.92	5.76	-0.16	4.63	-1.29
5	NAA	1	1	508.186	1.49	1.61	0.12	2.00	0.51
6	QFN	1	1	641.2388	2.38	2.08	-0.30	3.97	1.59
7	ASW	1	1	596.2173	3.47	3.24	-0.23	3.60	0.13
8	LAA	1	1	507.2272	3.68	3.91	0.23	4.22	0.54
9	IGM	1	1	553.2149	4.30	4.26	-0.04	4.08	0.22
10	WFW	1	1	771.2959	5.88	6.17	0.29	4.54	1.34

3.3.4 Comparison between the experimental RT and predicted RT using tripeptide mixture data

We further validated our models using the RT from XXA tripeptide mixture. X represents any one of the 20 essential amino acids. The exact mass and MS/MS spectra were manually inspected to confirm the tripeptide sequence. Since leucine and isoleucine share the same exact mass, distinguishing their RT may induce mistakes even by manually selection from the mixture. The RT of tripeptides containing leucine or isoleucine was excluded. After manual selection, the RT of 147 tripeptides remained for testing models. Among these peptides, 68, 60, and 9 tripeptides were labeled by one tag, two tags, and three tags, respectively. The distribution of RT difference between experimental RT and predicted RT is illustrated in Figure 3.5. Similar to the result of tripeptide standards, we found that the RT difference of the model built by unlabeled SMILES shifts to the right side, indicating a larger RT error between experimental RT and predicted RT. The RT difference of tripeptides labeled by different numbers of tags was examined, shown in Figure 3.5c. The difference from the model built by unlabeled SMILES is higher for peptides labeled by all three numbers of tags, especially for those labeled by three tags.



a)



b)

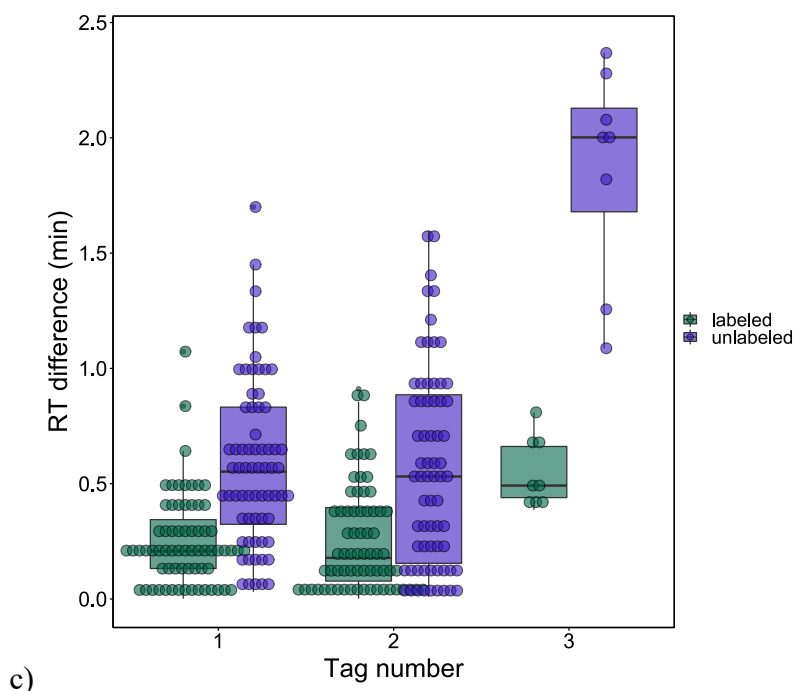


Figure 3.5 a) Distribution of RT difference in tripeptide mixture using the model built by labeled SMILES. b) Distribution of RT difference in tripeptide mixture using the model built by unlabeled SMILES. c) RT difference of tripeptide mixture labeled by different tag numbers.

3.3.5 Distinguish in-source fragmentation

The predicted RT time can also be used for differentiating labeled peptides from the in-source fragmentation. For example, the extracted ion chromatogram (EIC) of one-tag labeled and one charged LH contains 2 peaks, shown in Figure 3.6a. The first EIC peak at 2.53 min matches with the predicted RT of one tag labeled LH as 2.87 min. We further investigated the mass plot of the second peak, both one tagged one charged m/z and two tagged two charged m/z could be found at 6.94 min, as demonstrated in Figure 3.6b. The predicted RT of two tags labeled LH is 7.11 min, 0.17 min different from the second peak. The EIC of two tags labeled LH is also shown in Figure 3.6c, which is well-aligned with the second peak of one tag labeled LH. Both the mass plot and EIC plots indicate that the second peak from 6.94 min is the in-source fragment of two tags labeled LH. This illustrated the predicted RT can be used for differentiating in-source fragmentation.

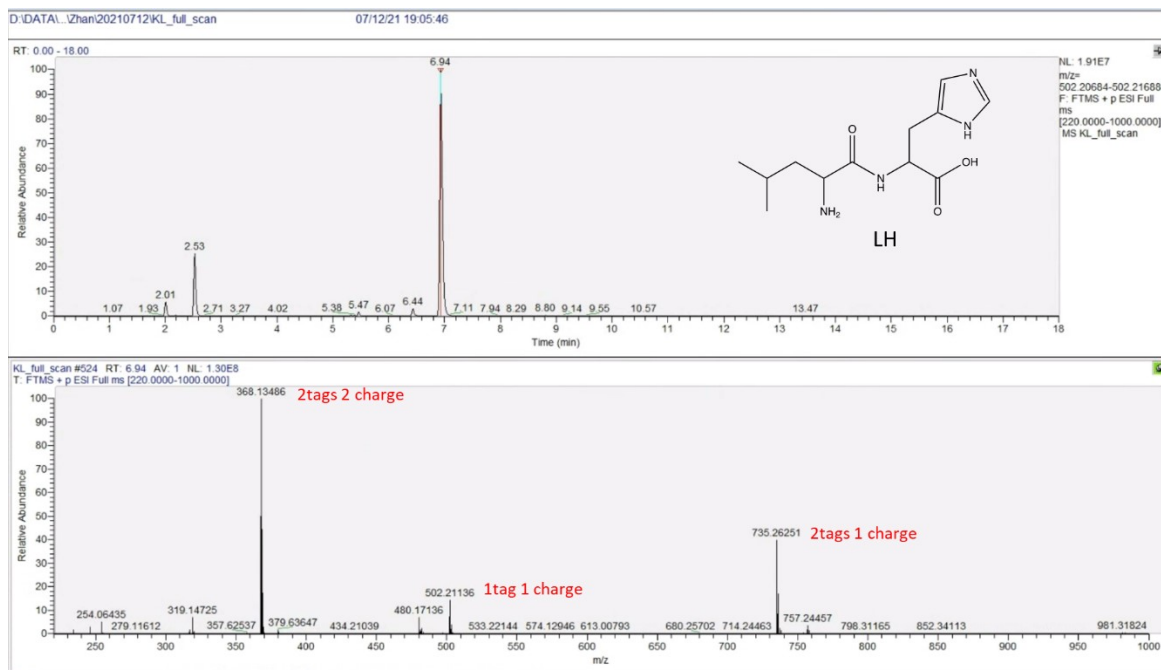


Figure 3.6 a) The EIC plot of 1 tag labeled LH. b) The mass plot of 1 tag labeled LH at 6.94 min. c) The EIC plot of 2-tag labeled LH.

3.3.6 Identification against serum and urine samples.

To demonstrate the performance of the predicted tripeptide RT, we acquired the data of serum samples and urine samples labeled with dansyl chloride. Three technical replicates (double-check if they are technical replicates) of each sample were acquired. After data processing by IsoMS Pro, 1822 and 51 peak pairs were detected in serum and urine samples, respectively. Then, these peak pairs were submitted for tripeptide identification with 10 ppm as mass tolerance and 1 min RT tolerance. 329 and 528 tripeptides in total were identified in serum and urine samples, respectively, shown in Figure 3.7a. Among these tripeptides, 87 were shared by both serum and urines samples, showed in Figure 3.7b.

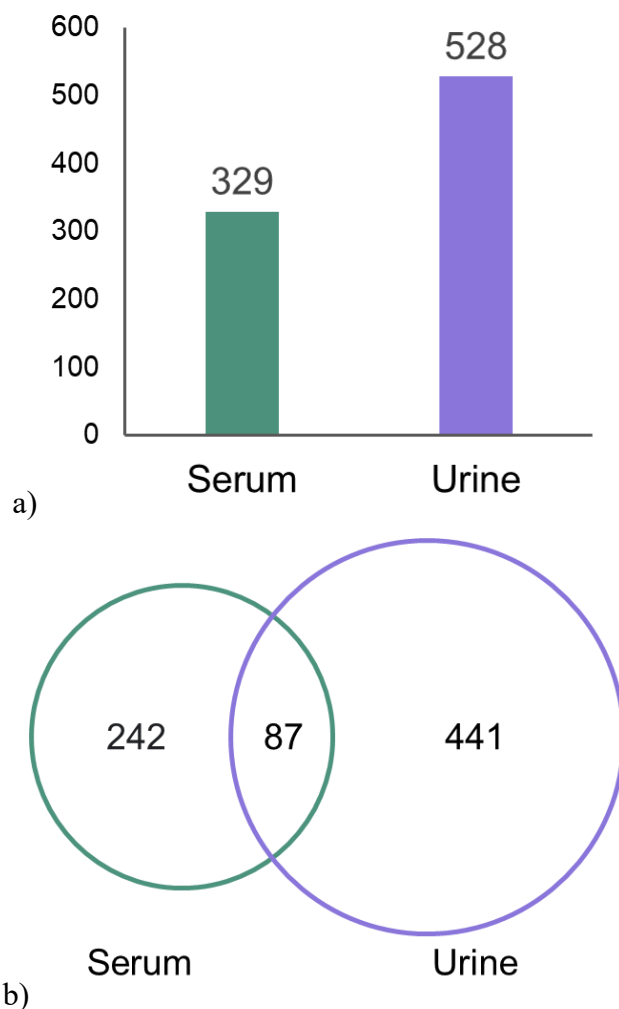


Figure 3.7 a) The putatively identified tripeptides in serum and urine samples. b) The overlap of putatively identified tripeptides between serum and urine samples.

3.4 Conclusions

In this work, we predicted and validated the tripeptide RT using models built by dipeptide RT information. SVR models with radial kernel were built based on the molecular descriptors of labeled and unlabeled short peptides. And their performance was further compared. Despite the similar performance on the dipeptide dataset, the model based on the labeled molecular descriptors provided more accurate RT prediction on both tripeptide standard and mixture datasets, especially for tripeptides with multiple tags. It indicates physical and chemical properties altered by chemical

labeling cannot be well-captured by features from unlabeled short peptides. Besides, the unlabeled SMILES failed to designate the tag position if multiple labeling positions are available on short peptides. In future work, the RT prediction of tetrapeptides, many of which are pharmacologically active, will be investigated and validated.

Chapter 4 MyCompoundID 2.0: evidence-based metabolome library facilitating CIL identification

4.1 Introduction

Metabolomics has been widely applied to biomarker discovery^{1 92}, disease pathway and pathology⁹³, and toxicology^{94 95 96}. As a result of metabolite high diversity of physical and chemical properties, metabolite identification is still the bottleneck of metabolomics study.⁹⁷ In liquid chromatography-mass spectrometry (LC-MS) based metabolomics, accurate mass, MS/MS spectra^{98 49}, retention time, and collision cross-section (CCS) values^{99 100} can be used for metabolite identification by comparing the experimental data and reference data. Reference data can be obtained either from the in-house library or external databases. The reference data from the in-house library are usually acquired at the same instrument conditions as experiment data, which can provide more accurate identification results. But due to the lack of metabolite standards, it is difficult to generate an in-house library to cover the whole metabolome. External metabolomics databases, such as HMDB⁴⁹ and KEGG¹⁰¹ compound databases, or chemical databases, such as ChemSpider¹⁰² and PubChem¹⁰³, can be used for putatively identifying known metabolites. Apart from known metabolites, unknown metabolites cannot be identified by comparing databases consisting of existing compounds. The predicted database^{104 105}, such as MyCompoundID (MCID)¹⁰⁶, an evidence-based metabolome database using metabolites and biological reactions to predict potential metabolites, can provide more identification information.

Although current MCID can throw a light upon the unknown compound identification, there is still room for improvement. With the development of technology in metabolomics research, more and more metabolites have been enrolled in the metabolomics database. The metabolites used for predicting potential metabolites in MCID can be further updated. The compounds used

for prediction in current MCID contain unrelated compounds, such as lipids and lipid-like compounds, which should be filtered to reduce false positive matches. The 76 reactions can be further optimized to recognize the different chemical environment, like -OH from phosphate groups or hydroxyl groups. Besides, after the exact mass search, instead of retrieving the structure information of predicted compounds directly, users can only obtain the relationship between substrates and possible reactions from the current MCID. At last, there is not an ID system for the predicted metabolites which can facilitate users to track the identification result.

To improve the issues listed above, in this work, we have constructed MCID 2.0 based on the KEGG compound database. The metabolite number after filtering unrelated compounds reached 11,164. The predicted potential metabolites and their structure information were provided in the search result. MCID ID was designed for each predicted metabolite and the corresponding MCID ID search was in development.

4.2 Methods

4.2.1 MCID 2.0 database construction.

The workflow of database construction is shown in Figure 4.1. The first step is filtering unrelated compounds, including lipids and lipid-related compounds, generic compounds, inorganic compounds and so on. The remained metabolites were used for zero-reaction database construction. Then, these metabolites were submitted to the prediction of 76 metabolite reactions. A Java-based program was built for prediction. The targeted functional groups and specific chemical environment for each reaction were recognized by the program and further reacted to generate the predicted metabolites. After prediction, products, such as CH₄, H₂O, were considered invalid for mass spectrometer analysis and filtered out. Next, the exact mass of predicted

compounds was calculated using ChemmineR R package¹⁰⁷ (version 3.34.1). Four digits after a decimal of exact mass were kept. For each predicted compound, a unique MCID ID was given to facilitate the searching of predicted metabolites. The ID consists of three parts: substrate ID, reaction index, and compound index. An MCID ID example, C03145R13001, shown in Figure 4.4a was used to explain the ID construction. The first five letters, C03145, are the KEGG ID representing the substrate. The next three letters, R13, indicate that the substrate has gone through the 13th reaction, demethylation, from the 76-reaction list. The final three letters, 001, are the unique compound index. To clarify the relationship between the substrates and predicted products, products from different substrates but sharing the same structure were kept with different MCID IDs. At last, the SMILES files of products were used for generating structure pictures. The final predicted database was named as MCIDxKEGG one-reaction database.

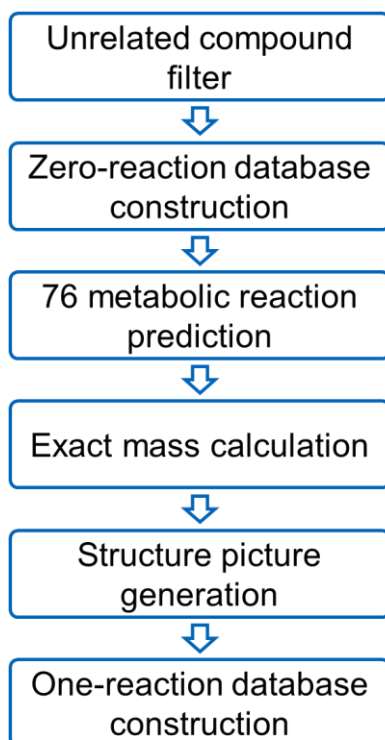
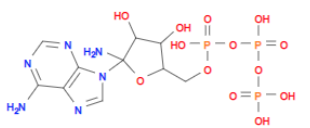
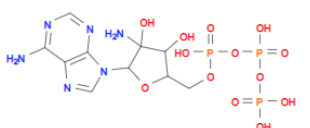


Figure 4.1 The workflow of database construction.

4.2.2 Web interface.

To facilitate users in identifying metabolites, we have developed a responsive web-designed interface named as MyCompound ID 2.0. And the link is https://mycompoundid.chem.ualberta.ca/kegg_query. The information of the predicted compounds based on metabolic reactions and their substrates from KEGG are stored in a local MySQL database. For each predicted product, an entry, including MCID ID, exact mass, a structure picture, substrate ID, substrate name, substrate exact mass, and possible reaction, was generated (example entries showed in Figure 4.2a. The exact mass of predicted one-reaction products was calculated up to the millionth precision. To facilitate users to manually validate the identification results, the structure pictures of predicted compounds were provided. MCID 2.0 web server was built based on Node.js and Express.js, providing two different search functions: exact mass search and MCID ID search, the web page is shown in Figure 4.2b and Figure 4.2c, respectively. An example of exact mass search result against one-reaction database is shown in Figure 4.2d.

a)

MCID ID	Mass (Da)	Structure	Substrate ID	Substrate name	Substrate mass (Da)	Possible reaction
C00002R06001	522.00664		C00002	ATP	506.9957	R06-addition of NH
C00002R06002	522.00664		C00002	ATP	506.9957	R06-addition of NH



- Mass search
- MCID search
- Introduction
- Possible reactions

Reactions:

No reaction

1 reaction

Neutral or Ion:

Neutral

[M+H]⁺

[M+Na]⁺

[M+K]⁺

[M+NH₄]⁺

[M-H]⁻

Query Mass:

522.0066

504.9801

Mass Tolerance:

In Da (default: ± 0.005 Da): Da

In ppm (default: ±5 ppm): ppm

→ Different query mass should be separated by Enter key

→ Only number can be input for mass tolerance

b)



- Mass search
- MCID search
- Introduction
- Possible reactions

Search by MCID

MCID:

C00002R06001

C00002R06002

→ Different MCID ID should be separated by Enter key or Space key

c)



Index	Query mass (Da)	Ion type	MCID ID	Mass (Da)	Structure	Substrate ID	Substrate name	Substrate mass (Da)	Possible reaction	Mass error
1	522.0066	neutral	C00002R06001	522.00664		C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	522.0066	neutral	C00002R06002	522.00664		C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	522.0066	neutral	C00002R06003	522.00664		C00002	ATP	506.9957	R06-addition of NH	0.0000 Da

d)

Figure 4.2 a) Example entries in MCIDxKEGG one-reaction. b) The exact mass search interface of MCID 2.0. c) The MCID ID search interface of MCID 2.0. d) the demo result page of MCID 2.0 one-reaction database.

In the exact mass search function, both brief version and detailed version of identification results were provided to download. In the detailed search result, named “MCIDxKEGG_oneRxn_full_result.csv”, each matched identification generates an entry, which means one query mass is corresponding to multiple entries in results. An example is shown in Figure 4.3a. Each entry consists of two parts, the query information and query result. The query information contains index (automatically generated according to the order of query mass), the ion type selected by users, and query mass input put users. The query result part includes the MCID ID of matched predicted metabolites, their exact mass, substrate information (covering ID, name, and exact mass), possible reactions that substrates have gone through, and the mass error between the query mass and exact mass of predicted metabolites. As complementary, in the brief search result, named “MCIDxKEGG_oneRxn_brief_result.csv”, each query mass is related to only one entry. An example is shown in Figure 4.3b. The brief result summarizes the detailed match and presents the total match number of a query mass, a list of all possible reactions, and a list of MCID ID for all matched predicted metabolites. The MCID ID list from the brief result can be used for the MCID ID search directly.

index	ion_type	query_mass_Da	mcid_id	exact_mass_Da	substrate_id	substrate_name	substrate_mass_Da	possible_reaction	mass_error
1	neutral	522.0066	C00002R06001	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06002	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06003	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06004	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06005	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06006	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
1	neutral	522.0066	C00002R06007	522.00664	C00002	ATP	506.9957	R06-addition of NH	0.0000 Da
2	neutral	504.9801	C00002R01001	504.9801	C00002	ATP	506.9957	R01-dehydrogenation	0.0000 Da

a) Query information

Query result

index	ion_type	query_mass_Da	match_num	possible_reaction	mcid_id
1	neutral	522.0066	7	R06-addition of NH	C00002R06001 C00002R06002 C00002R06003 C00002R06004 C00002R06005 C00002R06006 C00002R06007
2	neutral	504.9801	1	R01-dehydrogenation	C00002R01001

b) Query information

Query result

Figure 4.3 a) A demonstration of detailed search result against one-reaction database. b) A demonstration of brief search result against one-reaction database.

4.2.3 Materials

Organic solvents for the mobile phase and sample preparation were purchased from Thermo Fisher Scientific (Waltham, MA). ^{12}C -DnsCl derivatization reagent and ^{13}C -DnsCl derivatization reagent were made in-house according to our previous studies⁸⁷.

4.2.4 Sample preparation and derivatization reaction

For the urine samples, 75 μL of lyophilized human urine standard was diluted with H_2O to 300 μL . The derivatization reaction was conducted to label the amine group on the N terminal of dipeptides according to a protocol developed in our lab with slight modification.⁸⁹ In brief, 25 μL of diluted urine samples were mixed with 12.5 μL of 250 mM $\text{Na}_2\text{CO}_3/\text{NaHCO}_3$ buffer. The solution was vortexed, spun down and mixed with 37.5 μL freshly prepared ^{12}C -DnsCl (18 mg/mL in ACN, for light labeling) or ^{13}C -DnsCl (18 mg/mL in ACN, for heavy labeling), followed by vortexing and spinning down. The mixture was incubated in an oven at 40 $^\circ\text{C}$ for 45 min. Then,

the incubated solution was mixed with 7.5 μL NaOH solution (250 mM in H_2O) to quench the excess ^{12}C -DnsCl or ^{13}C -DnsCl at 40 $^\circ\text{C}$ for 10 min. Lastly, the quenched solution was mixed with 30 μL of 425 mM formic acid in ACN/ H_2O (1:1, v/v) to consume the excess NaOH. The ^{12}C -DnsCl and ^{13}C -DnsCl derivatized products were equally mixed, and the mixture was centrifuged for 10 min at 12000 rpm before LC-MS analysis.

4.2.5 HPLC-MS condition

A Vanquish UHPLC Systems (Thermo Scientific) coupled with a Q Exactive HF Orbitrap Mass Spectrometers (Thermo Scientific) was used for DnsCl derivatized sample and dipeptide standards analysis. An Agilent reversed phase C18 column (100 \times 2.1 mm, 1.8 mm particle size, 95 \AA pore size) was used for separation and the column was maintained at 40 $^\circ\text{C}$. Mobile A consisted of 0.1% FA in HPLC grade water and mobile phase B was consisted of 0.1% FA in HPLC grade ACN. A total of 18 min gradient was set as follows: 25% B increased to 99% in 10 min and maintained for 5 min, followed by 3 min equilibrium. The flow rate was set to 0.4 mL/min. The Q Exactive HF mass spectrometer was operated under an ESI positive mode for both in-house database collection and sample analysis. Electrospray ionization parameters were as follows: the spray voltage was at 3.5 kV; the capillary temperature at 320 $^\circ\text{C}$; probe heater temperature at 360 $^\circ\text{C}$; Full mass scan mode (m/z 220–1800) was used at a resolution of 120k at m/z 200 with around 1.2 Hz scan rate. The automatic gain control (AGC) target was at 1×10^6 ions with 200 ms maximum ion injection time. The sample injection amount was optimized according to the maximum peak pair numbers at different injection volumes. For derivatized standards, the injection volume is 2 μL .

4.2.6 Data processing

The original data of the resulting LC-MS data of serum and urine sample was firstly converted to “.text” format with MSConvert GUI tool⁸⁸. The converted data containing ¹²C/¹³C peak pair information were further processed using IsoMS Pro software (Nova Medical Testing Inc., Edmonton, Canada)^{89 34}. Dipeptide identification in serum and urine was carried out with the in-house database of dipeptides.

4.3 Results and Discussion

The KEGG based MCID 2.0, named as MCIDxKEGG, was generated as the workflow shown above. First, lipids and lipids related compounds, such as C00249 palmitic acid, in KEGG compounds were filtered as shown in the work published previously⁸⁹. Then generic compounds, like C03193 (5-L-Glutamyl)-peptide, were also removed. Six hydrates were further filtered out. After removing unrelated compounds, 11,164 remaining metabolites in total were enrolled in the MCIDxKEGG zero-reaction database. 76 metabolic reactions were applied to the zero-reaction database to predict potential metabolites. 296,518 reactions in total were carried out based on the 11,164 substrates. After filtering invalid products, 1,811,882 predicted products were used for building the MCIDxKEGG one-reaction database. The statistics of MCIDxKEGG information is shown in Table 4.1.

Table 4.1 Statistics of database information.

	Metabolite number	Reaction number	Product number
MCIDxKEGG11164	11,164	296,518	1,811,882
MCIDxHMDB7998	7,998	211,748	2,667,520
MCIDxHMDB2683	2,683	80,554	445,963
MCID oneRxn (current website)	About 8,021	375,809	

As a comparison, 7998 HMDB compounds in the previous MyCompoundID database also went through the workflow of database construction. After unrelated compound filter, 2683 compounds remain and were used to construct the MCIDxHMDB2683 zero-reaction database, with detailed information as shown in previous publications. As a result of prediction and invalid product filter, 445,963 predicted products remain for the MCIDxHMDB2683 one-reaction database. 7998 HMDB compounds without removing unrelated compounds were also applied with metabolic reaction prediction. 7998 HMDB compounds and 2,667,520 predicted compounds were used for building MCIDxHMDB7998 zero-reaction and one-reaction databases, respectively. Although MCIDxHMDB7998 contains fewer substrates and reactions than MCIDxKEGG, it covered more products in the one-reaction database. This is caused by lipids and lipid-like compounds containing many reaction positions. For example, 13 predicted metabolites could be generated from palmitic acid (HMDB0000220) through reaction 01 dehydrogenation. The current MCID one-reaction database was also added into the comparison. Since the predicted product information was not included in the current database, only the reaction number was shown in Table 4.1.

To demonstrate the predicted database, 20 amino acids were considered as pseudo “unknown” compounds, and their exact mass was used as query mass. For example, 131.0405 Da,

the exact mass of methionine, was searched against the MCIDxKEGG one-reaction database with 0.005 Da mass tolerance (Shown in Figure 4.4a). 114 predicted compounds were found as putative identifications. We further compared their structure with methionine. Four predicted compounds share the same structure with methionine. C03145R13001, one of these four compounds, is a product of N-Formylmethionine (C03145) with the reaction of loss of CO (R13). This reaction is named N-formyl-L-methionine amidohydrolase (R00653), belonging to Cysteine and methionine metabolism, as well as Glyoxylate and dicarboxylate metabolism (Shown in Figure 4.4b).

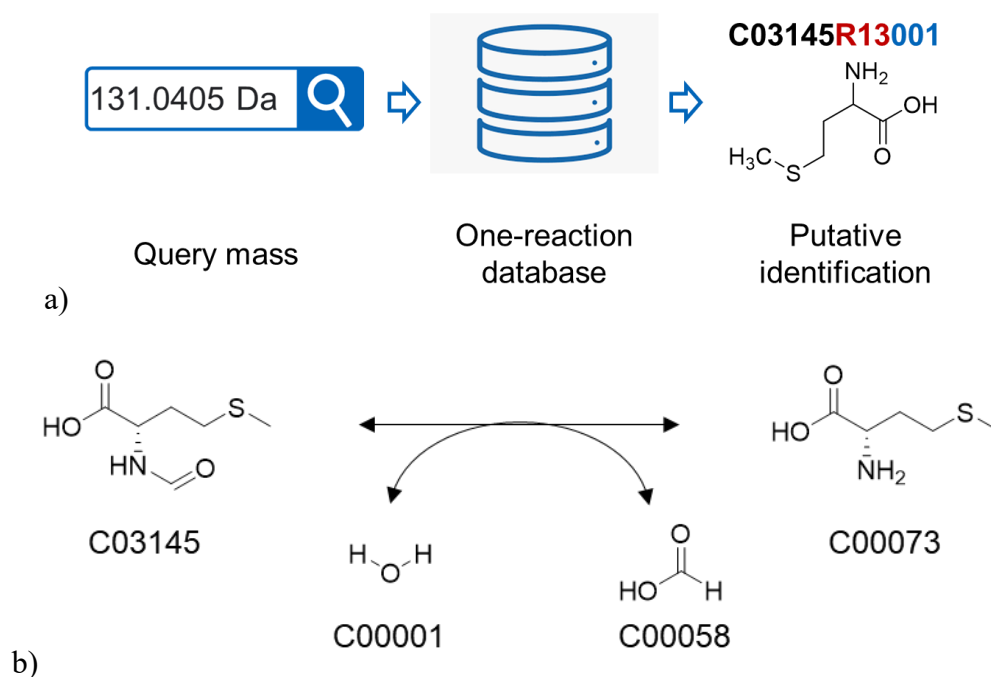


Figure 4.4 a) The demo identification of methionine. b) The biological reaction of methionine and N-Formylmethionine.

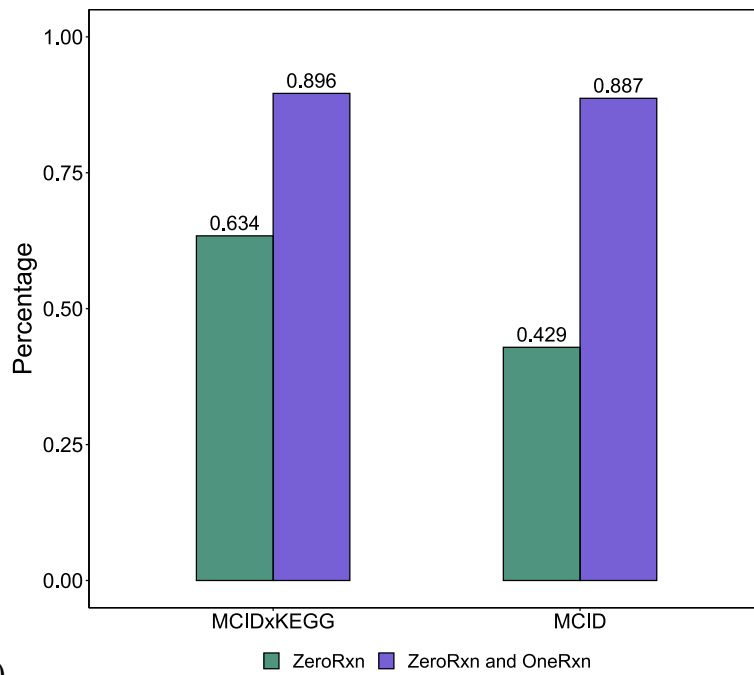
19 remaining amino acids were further searched against the predicted database. The result is shown in Table 4.2. The column of the exact mass match represents the matched number of

query mass fitting the 0.005 Da tolerance. The column of structure match is the number of predicted compounds sharing the same structure with query amino acids. As a comparison, the search results against the one-reaction and zero-reaction databases of MCIDxHMDB7998, MCIDxHMDB2683, and current MCID were also generated. Compared with MCIDxHMDB7998 and MCIDxHMDB2683 one-reaction databases, more identification could be obtained for both exact mass matches and structure matches except cysteine. Besides, the identification numbers of MCIDxHMDB7998 and MCIDxHMDB2683 one-reaction databases are identical except glycine, indicating that the filtered unrelated compounds rarely contribute to metabolite identification. Since the current MCID does not contain the structure information, only the exact mass match results are shown in Table 4.2. The identification number of the current MCID one-reaction database is much lower than the MCIDxKEGG one-reaction database.

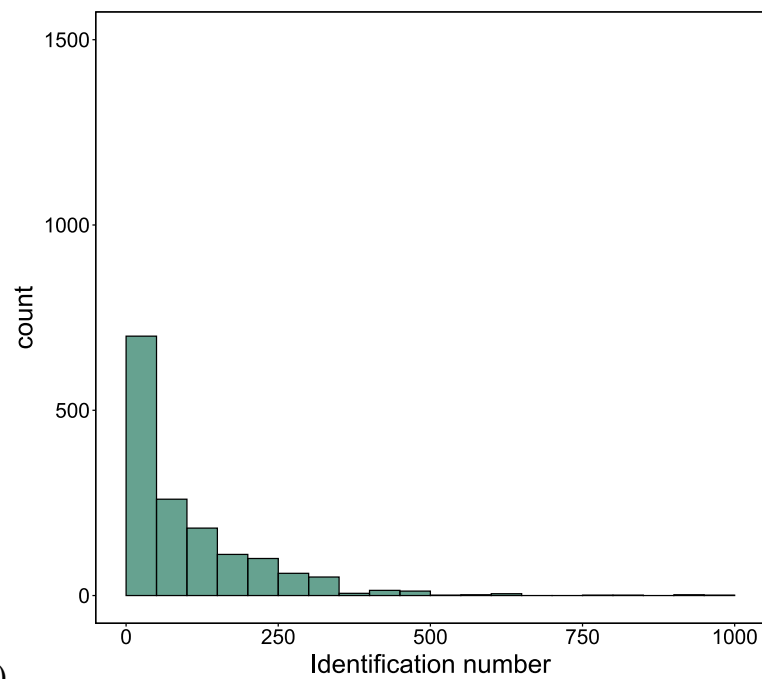
Table 4.2 Identification results of 20 amino acids against different databases.

amino acid	code	3LTR	MCIDxKEGG one reaction		MCIDxHMDB7998 one reaction		MCIDxHMDB2683 one reaction		MCID one reaction	MCIDxKEGG zero reaction	MCID zero reaction
			Exact Mass Match	Structure Match	Exact Mass Match	Structure Match	Exact Mass Match	Structure Match	Exact Mass Match	Exact Mass Match	Exact Mass Match
glycine	G	GLY	34	15	26	13	24	13	21	2	1
alanine	A	ALA	92	26	63	18	63	18	38	7	4
serine	S	SER	70	22	34	11	34	11	21	4	2
proline	P	PRO	145	6	61	4	61	4	33	4	2
valine	V	VAL	191	9	90	7	90	7	38	9	5
threonine	T	THR	114	12	74	9	74	9	35	8	3
cysteine	C	CYS	12	7	11	8	11	8	15	3	2
leucine	L	LEU	146	5	79	2	79	2	36	13	6
isoleucine	I	ILE	146	9	79	3	79	3	36	13	6
asparagine	N	ASN	71	5	26	2	26	2	20	9	5
aspartic acid	D	ASP	100	16	58	11	58	11	28	4	3
glutamine	Q	GLN	178	4	86	2	86	2	39	6	4
lysine	K	LYS	147	12	78	8	78	8	23	7	4
glutamic acid	E	GLU	251	18	154	12	154	12	55	9	6
methionine	M	MET	107	4	44	3	44	3	25	9	2
histidine	H	HIS	54	7	38	6	38	6	14	4	1
phenylalanine	F	PHE	241	13	132	8	132	8	49	8	4
arginine	R	ARG	128	9	59	2	59	2	11	6	3
tyrosine	Y	TYR	299	13	159	8	159	8	46	9	5
tryptophan	W	TRP	202	10	77	3	77	3	15	5	1

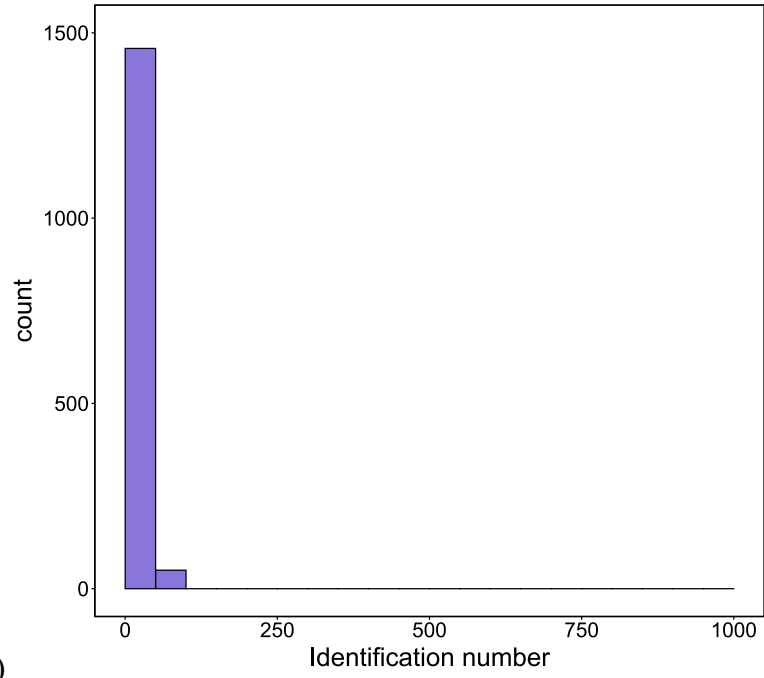
We further tested the MCIDxKEGG database using dansyl labeled urines samples. The current MCID database was also tested as a comparison. 1507 amines or phenols related peak pairs were detected in urine samples. Among these, 955 out of 1507 peak pairs were putatively identified by searching against the MCIDxKEGG zero-reaction database alone, covering 63.4% (shown in Figure 4.5a). And 1349 peak pairs were identified using the MCIDxKEGG one-reaction database. After the combination of zero-reaction and one-reaction results, the metabolite coverage reached 89.6% (1351). Meanwhile, 647 peak pairs and 1334 peak pairs were identified using the MCID zero-reaction and one-reaction databases, respectively. The corresponding coverage reached 88.7% (1336). We further compared the matched numbers for peak pairs, the related histograms are shown in Figure 4.5b and Figure 4.5c. Although the coverage for MCIDxKEGG and MCID is comparable, the identification numbers of MCIDxKEGG are significantly higher than MCID. 1458 peak pairs (96.7%) of MCID contains 0 to 50 putative identifications. In contrast, only 670 (44.6%) of MCIDxKEGG contains 0 to 50 putative identifications. The rest of the peak pairs contain higher matched numbers. Besides, the search results from one-reaction database can help users to mine zero-reaction putative identifications. If the upstream substrates or downstream products of a metabolite can also be putatively identified, this metabolite would have a higher chance of existing in biological samples. We analyzed 3812 putatively identified metabolites by zero-reaction database. The predicted products of 1715 (45.0%) metabolites can be found in one-reaction database search, which can increase the identification confidence. The distribution of the possible reaction numbers of 1715 putative identified metabolites is shown in Figure 4.5e (52.1%). Instead of showing the predicted product numbers, the possible reaction numbers of putatively identified metabolites are also shown in Figure 4.5e.



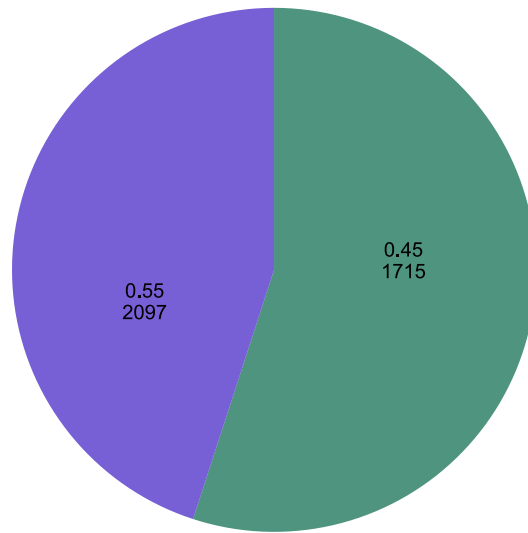
a)



b)



c)



d)

■ With reaction ■ Without reaction

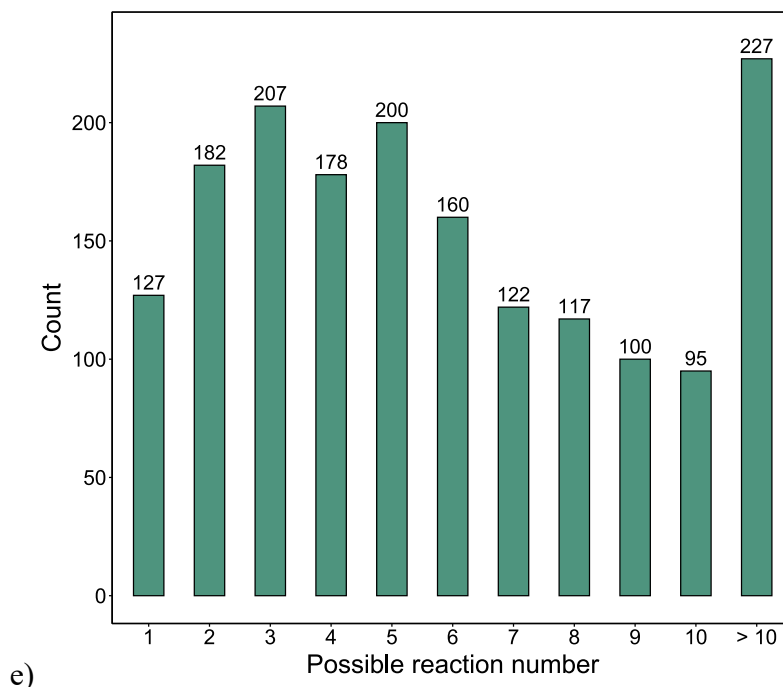


Figure 4.5 a) The identification coverage of MCIDxKEGG and MCID database. b) The frequency distribution of MCIDxKEGG one-reaction identification number. c) The frequency distribution of MCID one-reaction identification number. d) The percentage of putatively identified metabolites by zero-reaction database with and without predicted products, respectively. e) The frequency distribution of the possible reaction number of putatively identified metabolites by zero-reaction database.

4.4 Conclusions

To summarize, we have upgraded our evidence-based metabolome library, MyCompoundID, to MyCompoundID 2.0. More metabolites and their predicted products are included. Besides, the new MCID ID system can facilitate users to save and re-search predicted metabolites. The structure information can assist users to determine the unknown metabolites. Compared to using standard library alone or previous MCID, MyCompoundID 2.0 can provide better coverage. In the future, MCID 2.0 can be further expanded by recruiting more metabolites and more reactions. The structure information of predicted metabolites can also be employed for

retention time prediction, MS/MS spectrum prediction, CCS prediction and so on. Combining the information can further refine the identification result.

Chapter 5 Biomarker discovery on spinal cord injury using chemical isotope labeling profiling

5.1 Introduction

Acute traumatic injury to the spinal cord causes the sudden disruption of neural pathways to and from the brain, leading to significant motor, sensory, and autonomic dysfunction. Aside from the physical and emotional suffering caused by the paralysis of spinal cord injury (SCI) to the affected individuals and their families, the injury is associated with huge societal economic costs. The lifetime economic burden including the direct as well as indirect costs was estimated in Canada. For a patient suffering from paraplegia or tetraplegia, the estimated cost is \$1.5 million and \$3.0 million, respectively.¹⁰⁸ The two primary causes for Acute traumatic SCI are motor vehicle collisions and falls.^{109 110 111} Treatments to improve neurologic recovery and functional outcome are sorely needed, but to date, few therapeutic approaches have shown convincing effectiveness.

Furthermore, even if a valid ISNCSCI examination can be performed in the early stages of acute SCI, because the entire spectrum of traumatic injury to the spinal cord is categorized into essential four AIS (American Spinal Injury Association Impairment Scale) grades, there is considerable variability in spontaneous neurologic recovery. This makes it difficult to accurately prognosticate neurologic recovery¹¹², leading to more patients required in clinic trials for sufficient power.

These issues impose important limitations on the clinical evaluation of acute SCI with the ISNCSCI examination. It simply cannot be done in many patients who, in the early stages of injury may be unconscious or have other injuries that preclude active participation¹¹³. Its reliability depends on the training and experience of the examiner, and ultimately, it lacks precision for

predicting outcome. For these reasons, there has been interest in the establishment of biomarkers of acute SCI – objective measures that could be used to classify the severity of neurologic impairment and better predict outcome.

Biofluids such as serum, plasma, and cerebrospinal fluid (CSF) represent potential sources of neurochemical biomarkers of SCI. human^{114 115 2} and animal studies^{116 117 118} have revealed potential candidate markers for stratifying injury severity and predicting outcome. To comprehensively interrogate these biofluids, multiple omics platforms have been employed, including transcriptomics¹¹⁹, proteomics¹²⁰, and metabolomics². More targeted biochemical approaches have revealed that CSF proteins such as interleukin (IL)-6, IL-8, monocyte chemoattractant protein (MCP-1), tau, S100 β , and GFAP, can separate different AIS grades at 24 hr after injury and predict AIS graded improvement at 6 months post-injury¹¹⁴.

In this work, we have interrogated the metabolome in serum samples obtained from acute SCI patients, to determine if there are unique metabolomic factors that could be used as biomarkers to define injury severity and also predict outcome. Here, we used liquid chromatography mass spectrometry (LC-MS) and chemical isotope labeling (CIL) to enhance the detection of metabolites containing amine or phenol groups. One objective of this study is to classify the baseline AIS grades, which can facilitate the selection of intervention methods at early stage and further evaluation during rehabilitation processing. Another objective is to differentiate the recovery potentials of patients over time, which can better characterize the injury status even for patients with the same AIS grade.

5.2 Methods

5.2.1 Chemicals and reagents

The LC-MS grade solvents and reagents, including water, methanol, acetonitrile (ACN), and formic acid (FA) were purchased from Fisher Scientific (Ottawa, ON). ^{13}C dansyl chloride (DnsCl) was purchased from Nova Medical Testing Inc. (Edmonton, AB).

5.2.2 Sample collection

Individuals suffering acute cervical or thoracic SCI (AIS A, B, or C) were enrolled within 48 hours of injury into a prospective observational clinical trial at 6 North American sites (ClinicalTrials.gov: NCT01279811). In this trial, lumbar intrathecal catheters were inserted pre-operatively, and serial CSF and serum specimens were collected over the first 3-5 days post injury. In addition to the objective of evaluating CSF and serum for biomarkers of acute SCI, an objective of this study was to evaluate intrathecal pressure and spinal cord perfusion pressure^{121 122}. Importantly, subjects were included only if a valid, reliable baseline neurologic examination could be performed upon their admission. This therefore excluded those with a concomitant TBI, major axial or appendicular trauma, or who were too sedated or intoxicated to assess neurologically.

An ISNCSCI examination was performed at admission in the emergency room or ICU (prior to surgery) and at 6 months post-injury to establish baseline and follow-up AIS grade and MS. All sites involved in the clinical trial received ISNCSCI training to ensure the reliability of the neurologic examinations. The admission ISNCSCI examination that constituted the assessment to assign the “baseline AIS grade and motor score” was done prior to surgery, as the subjects needed to consent to the study and the placement of the intrathecal catheter placement in the operating room. A non-SCI control group of individuals undergoing routine lumbar spine surgery

was also recruited. All SCI and control subjects provided informed consent for participation in the study. Institutional ethics approvals were in place at all participating sites.

5.2.3 Sample preparation

The aliquots from each individual sample were pooled together as pooled serum samples. Serum sample preparation including following steps: first, 45 μL of methanol was added into 15 μL of serum sample for protein precipitation. The sample was incubated at $-20\text{ }^{\circ}\text{C}$ for 2 h. Then, 45 μL of supernatant was transferred into a new vial and dried under vacuum. Next, metabolites were reconstituted using mixture of solvents, including 25 μL of water, 12.5 μL of ACN, 12.5 μL of $\text{Na}_2\text{CO}_3/\text{NaHCO}_3$ buffer. 25 μL of ^{12}C -DnsCl (18 g/mL dissolved in ACN) was added to individual samples and part of pooled samples for QC samples. Whereas ^{13}C -DnsCl (18 g/mL dissolved in ACN) was added into the rest pooled samples. The reaction vials were placed into an incubator at $40\text{ }^{\circ}\text{C}$ for 45 min. To quench the excess DnsCl, 5 μL of 250 mM NaOH was added into vials. Then sample vials were incubated at $40\text{ }^{\circ}\text{C}$ for another 10 min. At last, to neutralize the excess NaOH and acidify the solution, 25 μL of 425 mM FA dissolved in 1:1 ACN/ H_2O was added.

5.2.4 LC-UV methods

After sample preparation, the metabolite concentration of individual samples and pooled samples were measured using LC-UV method ¹. Waters ACQUITY UPLC system (Waters, Milford, MA) with a Phenomenex Kinetex C18 column (2.1mm \times 55 cm, 1.7 μm particle size, 100 Å pore size) was used for separation. For each labeled sample, 5 μL was injected. Mobile phase A was 0.1% (v/v) FA in 5% (v/v) ACN, and mobile phase B was 0.1% (v/v) FA in ACN. The 6.5 min step gradient was set as following: B was set as 0% at 0 min to 1 min. Then B was increased

to 95% at 1.1 min and kept until 2.6 min. B was further decreased to 0% at 3.1 min and kept until 6.5 min. The flow rate was set as 0.45 mL/min. PDA detector was operated at 338 nm. Waters Empower (V6.00) was employed for integrating the peak area indicating the metabolite total concentration.

5.2.5 LC-MS methods

Based on the total concentration measured by LC-UV, each ^{12}C -DsnCl labeled individual sample was mixed with the ^{13}C -DsnCl labeled pooled samples by equal mole amount. For quality control (QC) samples, ^{12}C -DsnCl and ^{13}C -DsnCl labeled pooled samples were mixed by equal mole amount. Samples were injected into a Agilent 1290 HPLC coupled to a Bruker compact quadrupole time-of-flight (Q-TOF) mass spectrometer (Bruker, Billerica, MA). Labeled metabolites were separated by an Agilent reversed phase Eclipse Plus C18 column (2.1 mm \times 10 cm, 1.8 μm particle size, 95 \AA pore size). The mobile phase was the same with the one used in LC-UV. The LC gradient was set as following: t = 0 min, 20% B; t = 3.5 min, 35% B; t = 18 min, 65% B; t = 21 min, 99% B; t = 34 min, 99% B, and flow rate of 0.18 mL/min. Mass spectrometer was set as following: polarity, positive; dry temperature, 230 $^{\circ}\text{C}$; dry gas, 8 L/min; capillary voltage, 4500 V; nebulizer, 1.0 bar; endplate offset, 500 V; spectra rate, 1.0 Hz.

5.2.6 Data processing

After data acquisition from LC-MS, data conversion from raw data to centroid data (.csv format) was processed by via Bruker Daltonics Data Analysis 4.3 software. IsoMS³⁴ was used for matching peak pairs, filtering unqualified peak pairs, calculating peak pair ratio, and align peak

pairs from different samples. Zero-fill program was employed to filling the values of peak pair ratio which were missing from last step.

5.2.7 Data analysis

The workflow of data analysis consists of seven steps. First, peak pairs with relative standard deviation (RSD) of ratio above 30% in QC samples were filtered. Peak pair ratio in each sample was normalized by median values. Confounding factors, including age and sex, were corrected using linear regression.¹²³ Auto scaling was performed on each peak pair across different samples. For every peak pair, normality and homoscedasticity was tested via Shapiro-Wilk test and Levene's test (from car R package, version 3.0.5), respectively. If both assumptions were qualified, Student's *t*-test was used to determine if peak pairs changed significantly between different groups. Otherwise, Wilcoxon signed-rank test was used. Partial least squares discriminant analysis (PLS-DA) was performed by pls R package (version 2.7.2), respectively. Peak pairs with fold change above 1.2 and *p*-value below 0.05 were considered as significantly changed metabolites. These metabolites were further subjected to building classification models via support vector machine (SVM) in Caret R package (version 6.0.84). The workflow is shown in Figure 5.1. Linear kernel was selected. Since the sample numbers in each AIS group were not equal, synthetic minority oversampling technique (SMOTE) was used for balancing the dataset during the model training processing. Receiver operating characteristic (ROC) analysis via pROC R package (version 1.15.3) was performed on model evaluation.

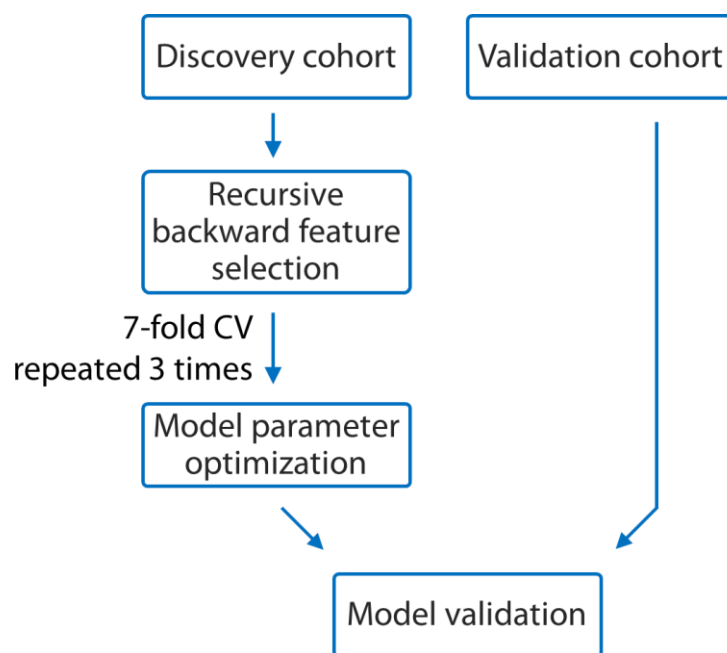


Figure 5.1 The workflow of building SVM models.

5.3 Results and Discussion

5.3.1 Demographic information of participants

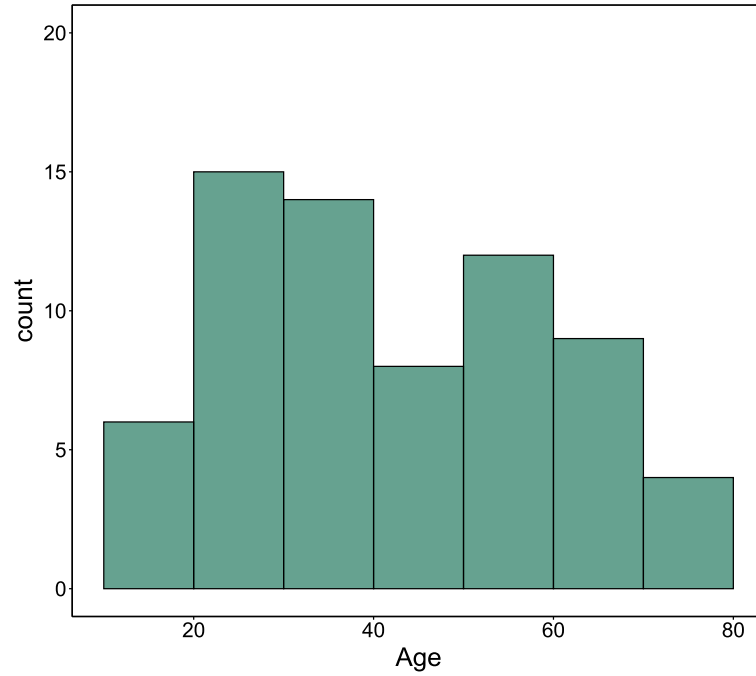
1. Classifying baseline AIS grades

In this project, metabolomics data of serum samples collected at 24-hr post-injury from acute SCI patients were used for prediction. For models differentiating baseline AIS grades, 68 samples in total are in the discovery cohort, covering 44, 10, and 14 samples from AIS A, AIS B, and AIS C patients, respectively. A summary is shown in Table 5.1. The patient age of these samples spans a large range from 17 to 77, shown in Figure 5.2. And 55 out of 68 patients are male, the rest 13 patients are female. As age and sex can affect metabolite levels, these two confounding factors were corrected before data analysis. Apart from the discovery cohort, serum samples from an independent cohort were collected as the validation cohort. In the validation cohort, the sample

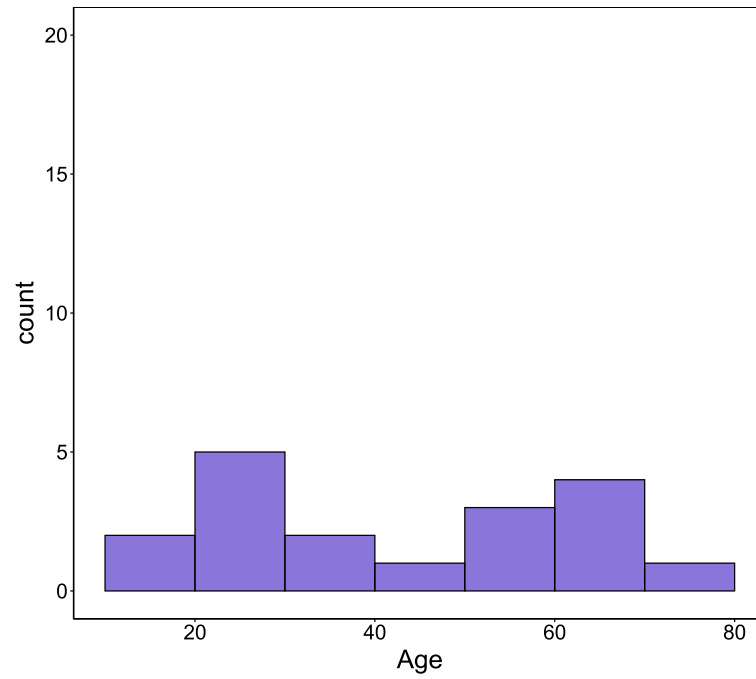
numbers of AIS A, B, and C grades are 16, 1, and 1, respectively. Similarly, the age in the validation cohort ranges from 19 to 76. 16 and 2 patients are male and female, respectively.

Table 5.1 The overview of baseline AIS grades of participants at admission.

	Discovery cohort			Validation cohort		
AIS grade at 24 h post-injury	A	B	C	A	B	C
Number	44	10	14	16	1	1



a)



b)

Figure 5.2 a) The age distribution of discovery cohort. b) The age distribution of the validation cohort.

2. Predicting AIS grade conversion in those with AIS A SCI

For individuals who presented with “complete” AIS A SCI, we examined whether specific metabolomic profiles could predict who would improve and “convert” to AIS B, C, or D, and who would not improve and remain AIS A. Here, 44 individuals with AIS A injuries at baseline were used for the discovery cohort, and 16 subsequent individuals were used for the validation cohort. The description of how many individuals converted in the discovery and validation cohorts is shown in Table 5.2 and Table 5.3, respectively.

Table 5.2 The overview of conversion information in the discovery cohorts.

Discovery cohort	Not converted	Converted		
AIS grade at 6 months post-injury	A	B	C	D
Number	29	10	3	2
Sum	29	15		

Table 5.3 The overview of conversion information in the validation cohorts.

Validation cohort	Not converted	Converted		
AIS grade at 6 months post-injury	A	B	C	D
Number	12	0	4	0
Sum	12	4		

3. Predicting Outcome at 6 Months

In situations where a baseline neurologic examination is not able to be conducted, it would be valuable to determine if certain metabolomic profiles at the 24-hr post-injury timepoint could

predict whether an individual would be “motor complete” (AIS A or B) or have some motor function and be “motor incomplete” (AIS C or D) at 6 months. Here, we compared the 24-hr serum metabolites of 43 participants who ended up being AIS A/B and 25 participants who ended up being AIS C/D at 6 months. From this discovery set of 68 subjects, we generated models and tested them on a validation cohort of 18 individuals, 12 of whom were AIS A/B, and 6 were AIS C/D. The description of these cohorts is shown in Table 5.4 and Table 5.5.

Table 5.4 The overview of motor function outcome in the discovery cohort.

Discovery cohort	Motor complete loss		Motor incomplete loss	
AIS grade at 6 months post-injury	A	B	C	D
Number	29	14	6	19
Sum	43		25	

Table 5.5 The overview of motor function outcome in the validation cohort.

Validation cohort	Motor complete loss		Motor incomplete loss	
AIS grade at 6 months post-injury	A	B	C	D
Number	12	0	4	2
Sum	12		6	

5.3.2 Metabolomics results

After the data acquisition using LC-MS, 8152 peak pairs were detected. 103 peak pairs were positively identified using an in-house library of amine/phenol channel. 643 from the unidentified peak pairs were putatively identified using the Li-Lib library. The remaining peak pairs were submitted to MyCompoundID zero-reaction, one-reaction database, and two-reaction database for further identification. 72.9% peak pairs were identified using all three levels together.

After RSD filtering peak pairs with RSD above 30% in QC samples, 5814 peak pairs were subjected to PCA analysis. The result is shown in Figure 5.3. Since the different QC samples were used during data acquisition of the discovery cohort and validation cohort, these two QC samples were separated on PCA plot but clustered together, respectively. This indicates the excellent instrument stability during sample acquisition. Besides, for the samples diagnosed as the scale, there is no clear separation between the discovery cohort and validation cohort, representing the batch effect was removed by normalizing and scaling the discovery cohort and validation cohort, separately. A PCA plot of the discovery cohort and validation cohort normalized and scaled together is shown in Figure 5.3b. Clear separation can be observed from the discovery cohort and validation cohort.

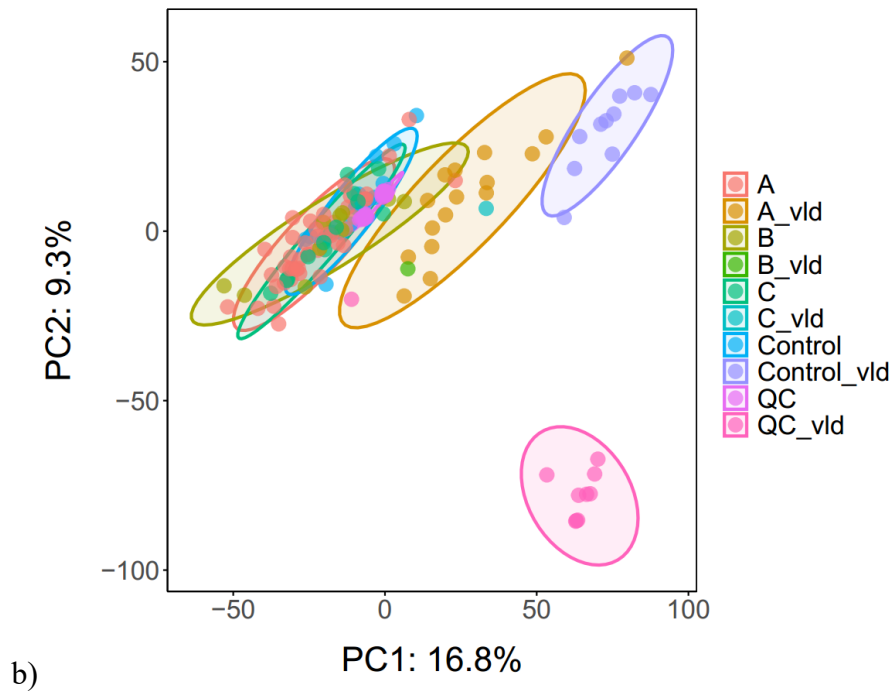
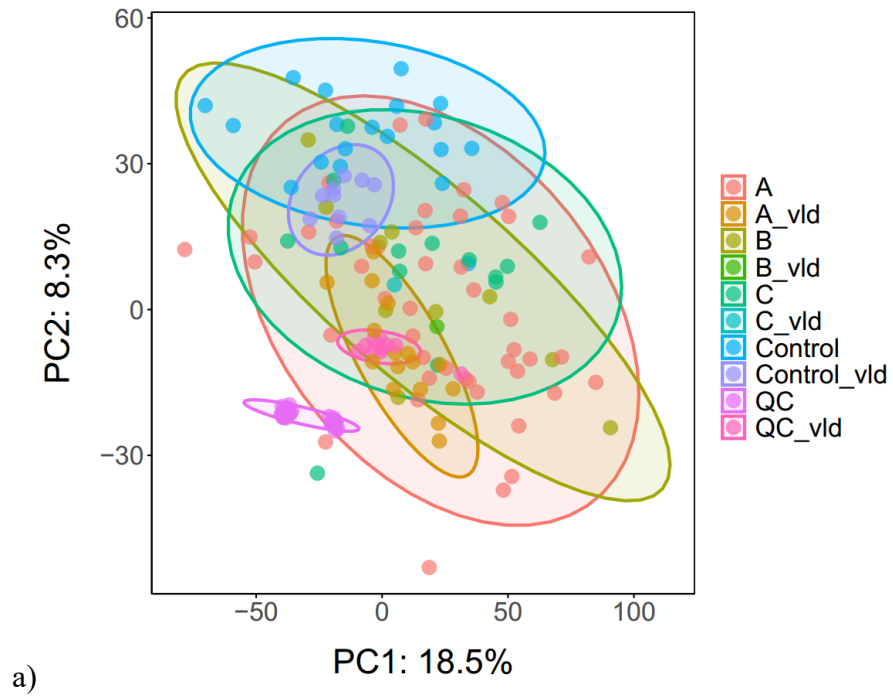
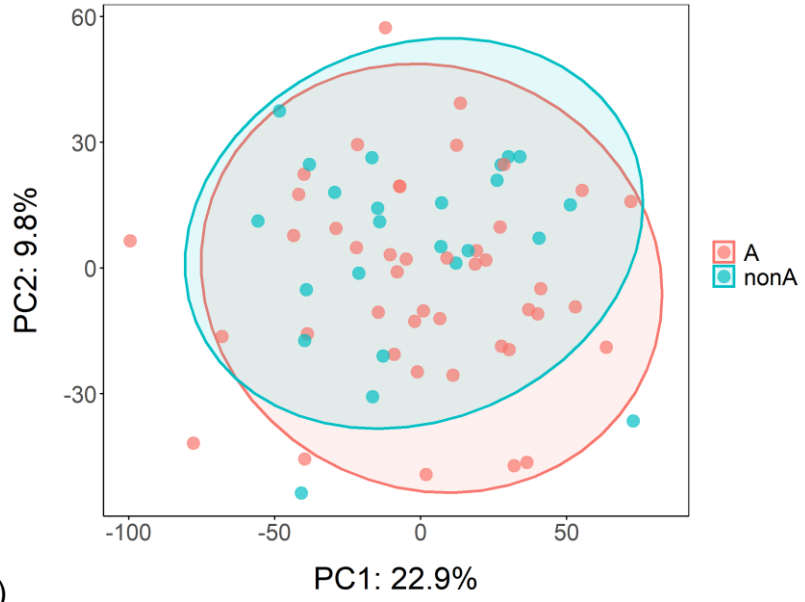


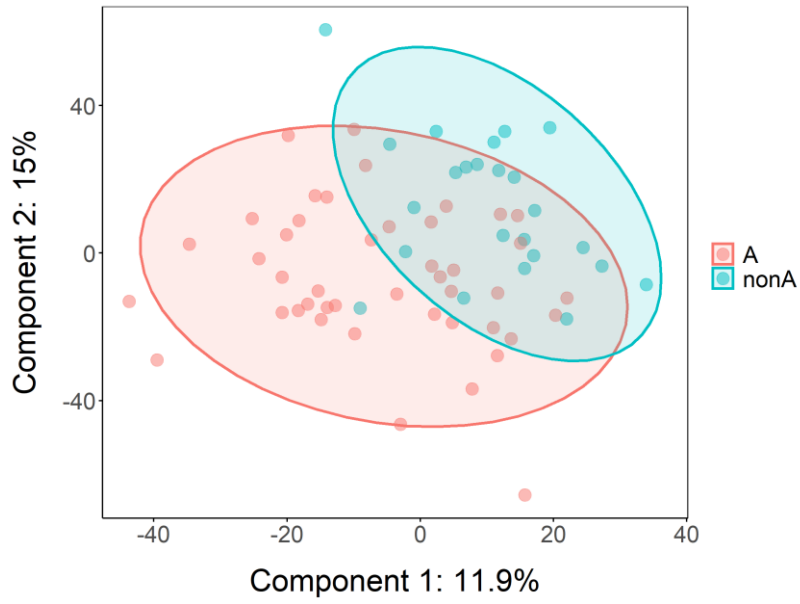
Figure 5.3 a) PCA plot of the discovery cohort and validation cohort normalized separately. b). PCA plot of the discovery cohort and validation cohort normalized together.

5.3.3 Classifying baseline AIS grades

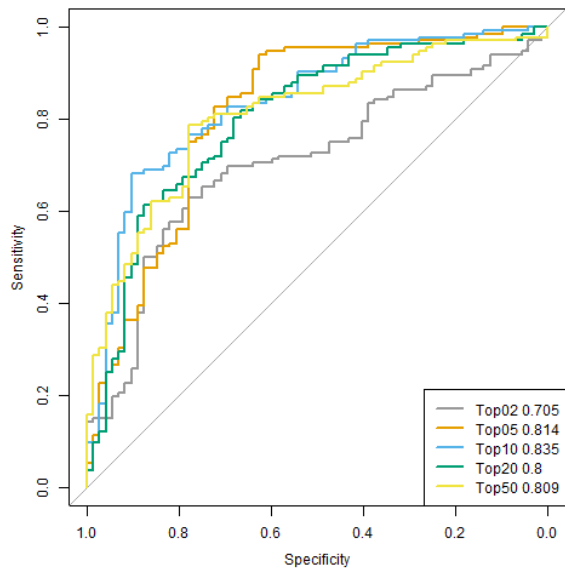
After examining the data quality, confounding factor correction was applied to each sample. The, the data were used for separating different AIS grade. Three different models were built to separate AIS A from AIS B and C, AIS B from AIS A and C, and AIS C from AIS A and B. Taking AIS A vs. non-A as an example, first, the PCA plot (Figure 5.4a) and PLS-DA plot (Figure 5.4b) were generated. Rare separation and weak separation of scale A and rest can be observed from PCA and PLS-DA plots, respectively. Then, fold change equal to or more than 1.2 and p -value below 0.05 was applied to find significantly changed metabolites. 70 significantly changed peak pairs were used for building SVM linear models to distinguish AIS A from non-A. Features were selected using backward elimination. ROC analysis was employed for models built with top 5, top 10, top 15, and top 20 features (shown in Figure 5.4c) reaching AUC to 0.705, 0.814, 0.835, and 0.800, respectively. Considering the model performance and metabolite number for building model in clinic application, the model constructed by the top 5 peak pairs was validated using the validation cohort, and the accuracy reached 0.661, shown in Table 5.6. The same analysis workflow was applied to scale B vs. non-B (Figure 5.4d, 5.4e, and 5.4f) and AIS C vs. non-C (Figure 5.4g, 5.4h, and 5.4i). In B vs. non-B, separation was observed in the PLS-DA plot. And the AUC of the model using the top 5 features reaches 0.934. The accuracy on the validation cohort is 0.778, shown in Table 5.7. In C vs. non-C, the separation in PCA and PLS-DA plots is not clear. The model built by the top 5 features with AUC reaching 0.883 was selected for prediction. The corresponding accuracy on the validation cohort is 0.778, as shown in Table 5.8.



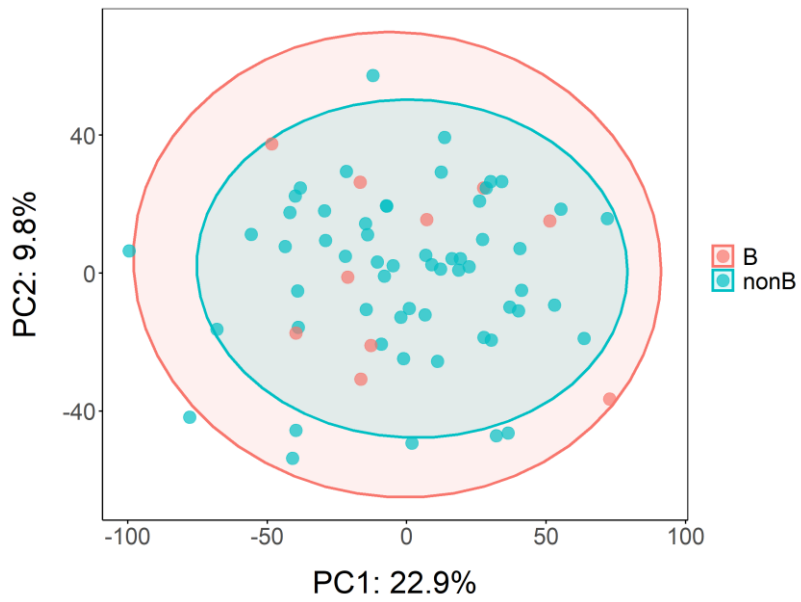
a)



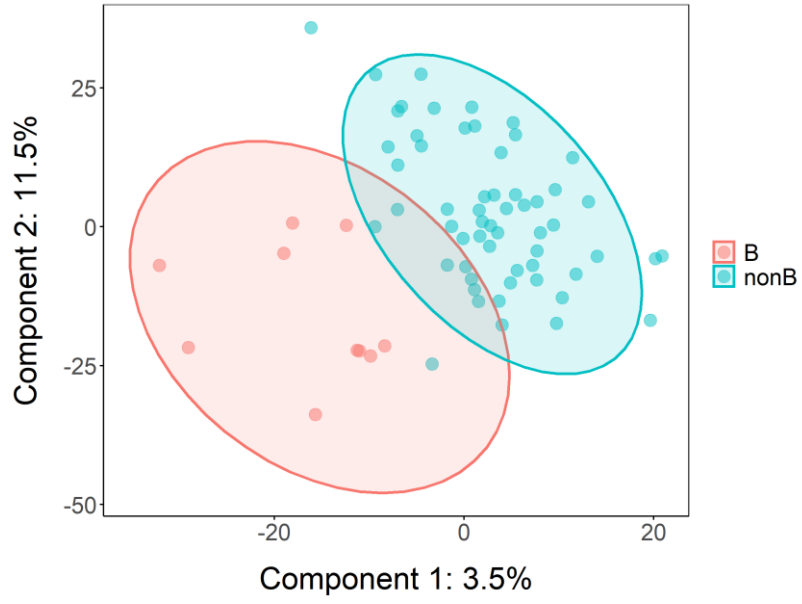
b)



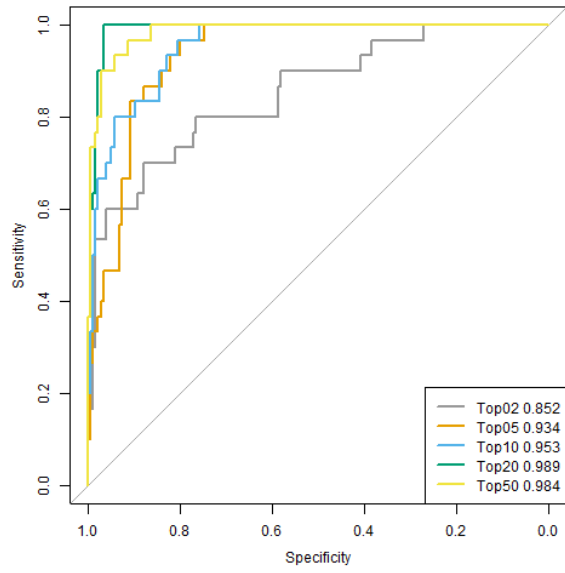
c)



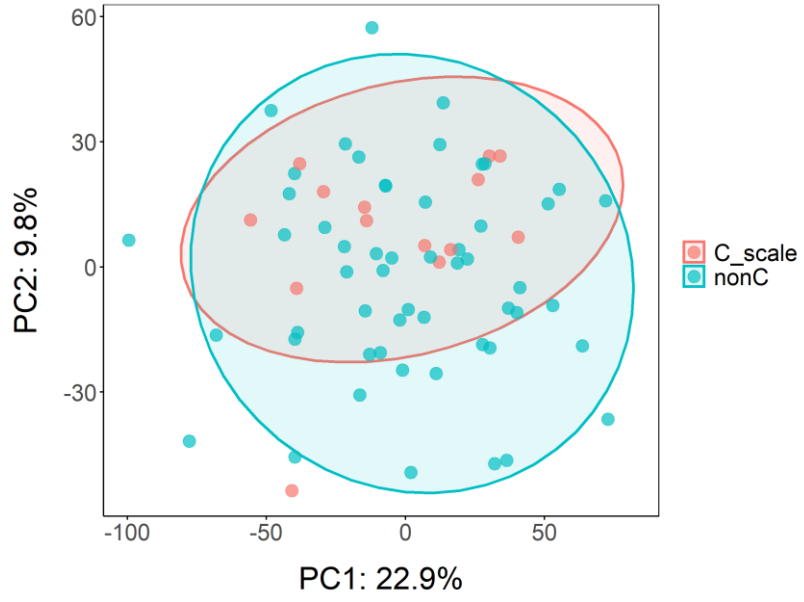
d)



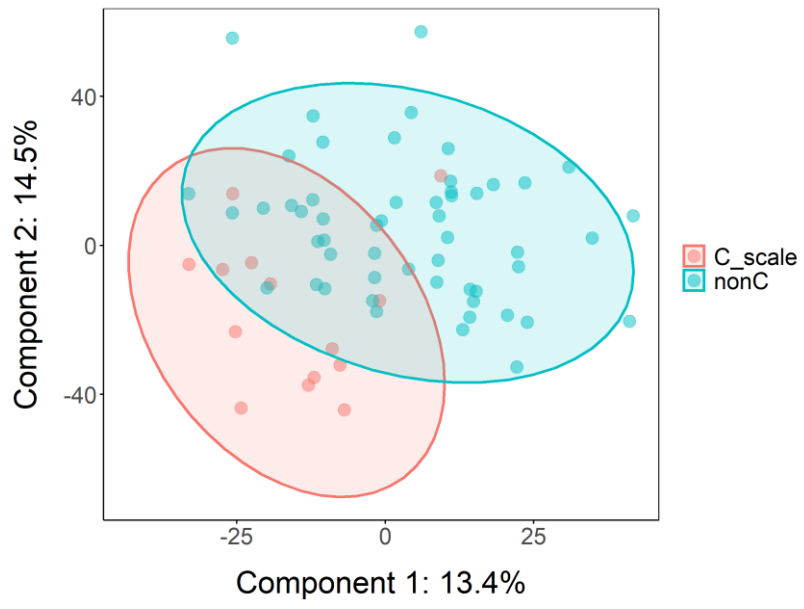
e)



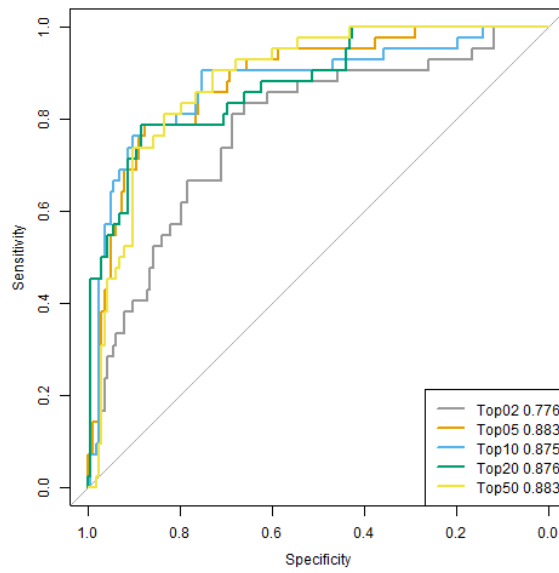
f)



g)



h)



i)

Figure 5.4 a) PCA plot of AIS A vs non-A. b) PLS-DA plot of AIS A vs non-A. c) ROC analysis of models differentiating A vs non-A. d) PCA plot of AIS B vs non-B. e) PLS-DA plot of AIS B vs non-B. f) ROC analysis of models differentiating B vs non-B. g) PCA plot of AIS C vs non-C. h) PLS-DA plot of AIS C vs non-C. i) ROC analysis of models differentiating C vs non-C.

Table 5.6 The confusion matrix of SVM model distinguishing AIS A and non-A grade.

Observed baseline AIS	Predicted baseline AIS		
	A	Non-A	%
A	10	6	62.5%
B	1	0	0.0%
C	0	1	100.0%
Overall %			66.1%

Table 5.7 The confusion matrix of SVM model distinguishing AIS B and non-B grade.

Observed baseline AIS	Predicted baseline AIS		
	B	Non-B	%
A	4	12	75.0%
B	1	0	100.0%
C	0	1	100.0%
Overall %			77.8%

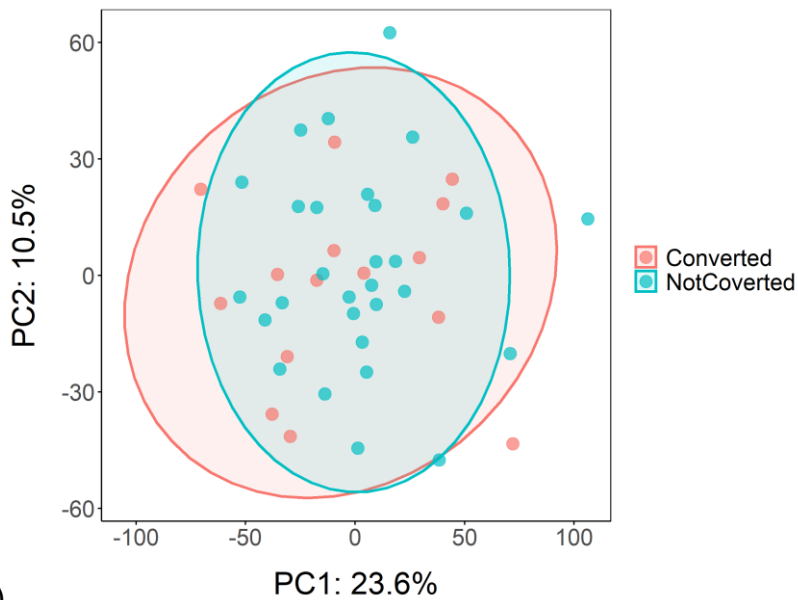
Table 5.8 The confusion matrix of SVM model distinguishing AIS C and non-C grade.

Observed baseline AIS	Predicted baseline AIS		
	C	Non-C	%
A	3	13	81.3%
B	1	0	0.0%
C	1	0	100.0%
Overall %			77.8%

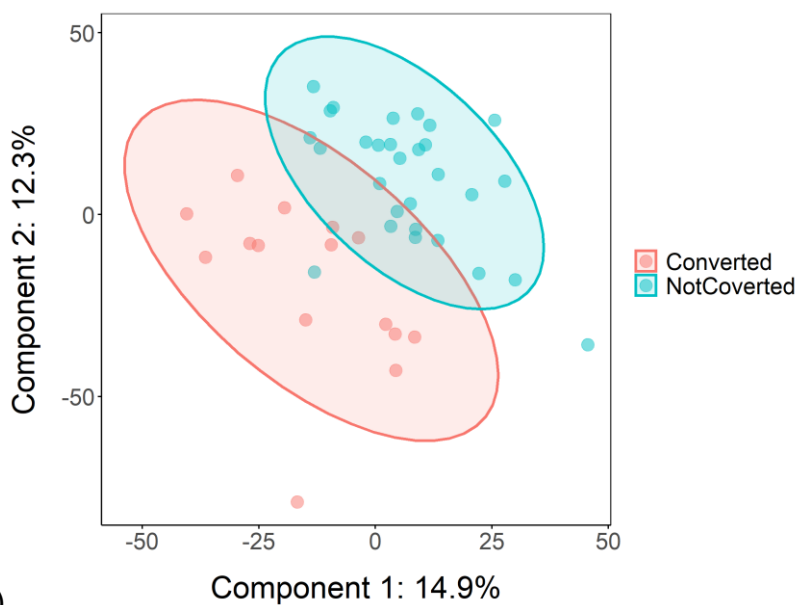
5.3.4 Predicting AIS grade conversion in those with AIS A SCI

Since spontaneous neurological recovery can diminish the reliability of clinic trials, a conversion model is necessary to predict whether patients diagnosed as AIS A grade would convert into AIS B, C, or D grade after 6 months. 44 AIS A samples were subjected to building the conversion model. The PCA and PLS-DA plots were shown in Figure 5.5. In the PCA plot (Figure 5.5a), the converted and non-converted groups can be rarely separated. But in the PLS-DA plot (Figure 5.5b), the separation can be observed. 87 metabolites were changed significantly between the converted and non-converted patients. The results of SVM linear models were shown in Figure 5.5c. In the ROC analysis, the AUC of the model built by the top 10 peak pairs reaches 0.908. However, we further tested the model using the validation cohort, containing 16 samples. The accuracy is only 0.4375 (shown in Table 5.9), indicating there could be potential overfitting. We

further tested the models built by top 5 peak pairs with 0.779 as AUC. The accuracy on the validation cohort is 0.500, shown in Table 5.10. This indicates the metabolomics data of amine/phenol channel cannot separate converted and non-converted patients although excellent performance was achieved on the discovery cohort.



a)



b)

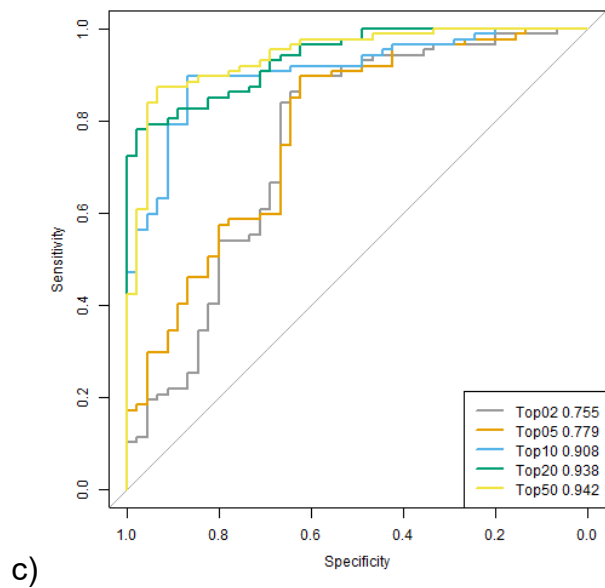


Figure 5.5 a) PCA plot of converted vs non-converted patients. b) PLS-DA plot of converted vs non converted patients. c) ROC analysis of models built by different numbers of peak pairs.

Table 5.9 The confusion matrix of SVM model distinguishing converted and non-converted patients using the top 10 features.

Observed conversion	Predicted conversion		
	No	Yes	%
No	7	5	58.3%
Yes	4	0	0.0%
Overall %			43.8%

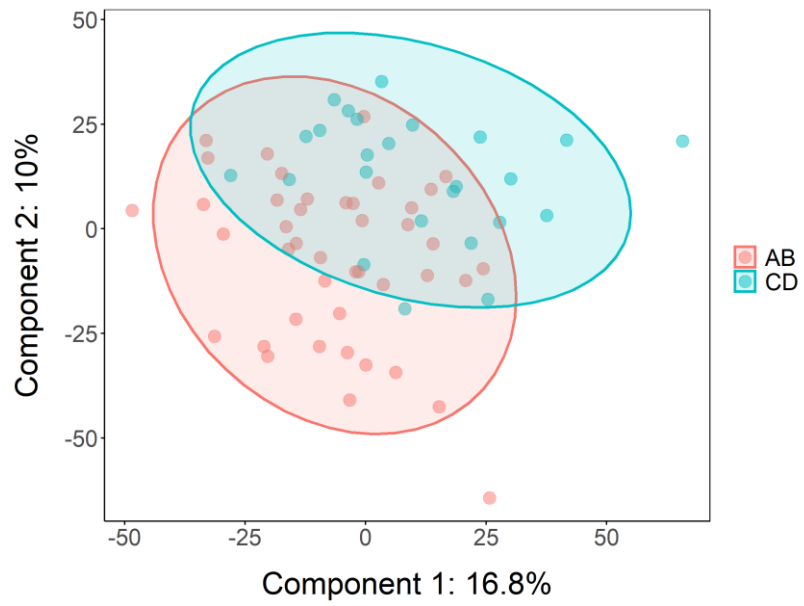
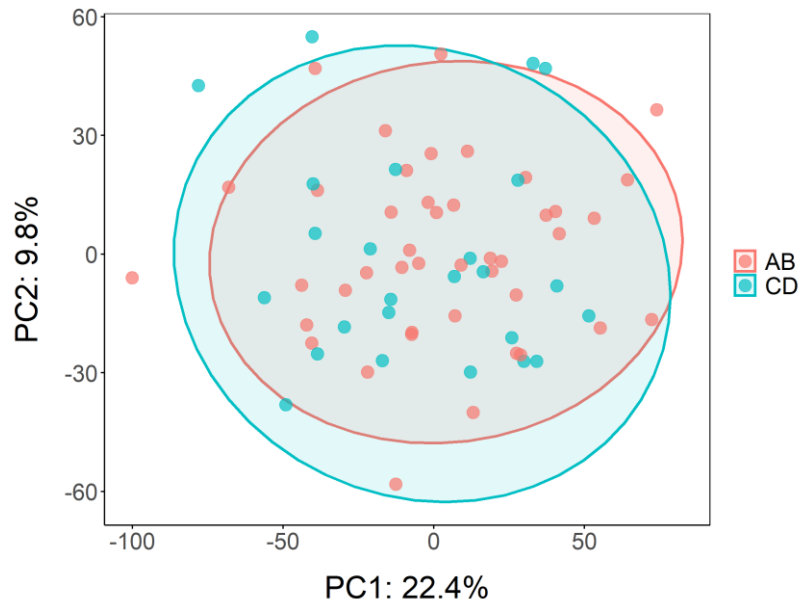
Table 5.10 The confusion matrix of SVM model distinguishing converted and non-converted patients using the top 5 features.

Observed conversion	Predicted conversion		
	No	Yes	%
No	8	4	66.7%
Yes	4	0	0.0%
Overall %			50.0%

5.3.5 Predicting Outcome at 6 Months

Based on the loss of motor function, patients with AIS A or B level can be classified as “motor complete”, and patients with AIS C or D level are denoted as “motor incomplete”. Upon the prediction of the likelihood that patients will obtain considerable improvement in motor function over time, interpretation of novel therapeutic strategies can be more comprehensible. The time point is usually selected as 6 months post-injury since most recovery of motor function takes place during this time.¹²⁴ In these models, only metabolite information was considered as input to build the models. The baseline AIS grades were excluded from the input. In the discovery cohort, the “motor complete” group contains 43 samples and the “motor incomplete” group consists of 25 samples. In the validation cohort, “motor complete” and “motor incomplete” groups contain the 12 and 6 samples, respectively.

From the PCA plot (Figure 5.6a) and PLS-DA plot (Figure 5.6b), the separation between two different groups was not conspicuous. 111 metabolites changed significantly between these two groups. From the ROC analysis (Figure 5.6c), the best performance was achieved using the top 5 features to build the model. The model was further applied to the validation cohort and the accuracy reached 0.667. The confusion matrix is shown in Table 5.11.



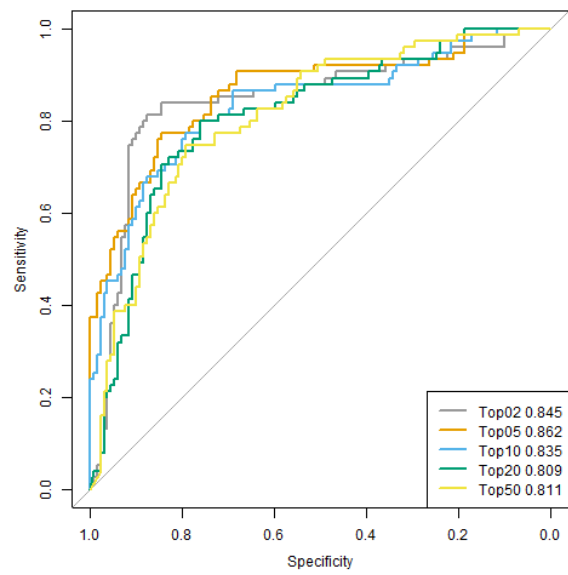


Figure 5.6 a) PCA plot of motor complete loss vs motor incomplete loss groups. b) PLS-DA plot of motor complete loss vs motor incomplete loss groups. c) ROC analysis of models built by different numbers of peak pairs.

Table 5.11 The confusion matrix of applying the outcome model using the top 5 features to the validation cohort.

Observed motor loss	Predicted motor loss		
	Complete	Incomplete	%
Complete	9	3	75.0%
Incomplete	3	3	50.0%
Overall %			66.7%

5.4 Conclusions

In this work, we used the chemical isotope labeling method to detect the amine/phenol channel of serum samples. With the detected metabolites, SVM models were built for distinguishing AIS levels and predicting neurological conversion as well as motor function outcomes. The results indicated that AIS levels could be fairly separated, but obvious inaccuracies were in the prediction of both neurological conversion and motor function outcomes, especially on the validation cohort. Therefore, in future work, metabolites from other sub-metabolomes should be incorporated into the models. And the combination of using serum and CSF samples together can also bring benefits to a better understanding of spinal cord injury.

Chapter 6 Conclusions and Future Work

6.1 Thesis Summary

In Chapter 2, the metabolome coverage of multiple chemical isotope labeling methods was evaluated through examining individual chemical structures of all compounds in metabolomic databases and classify them according to functional groups. After removing hydrocarbons, inorganic species and long-chain lipids (a subject of lipidomic analysis), four groups, namely hydroxyls (H), amines (A), carboxyls (C) and ketones/aldehydes (K), are the dominant classes. In the databases of MCID (2,683 metabolites), HMDB (15,817), KEGG (11,598), YMDB (1,107) and ECMDB (1,462), 94.7%, 94.2%, 86.4%, 85.7% and 95.8% of the metabolites belong to one or more of the four groups, respectively. Four-channel CIL approach using dansyl and DmPA reagents to analyze human plasma was employed to determine the detectable numbers of H, A, C and K-metabolites. Compared to group distributions of database compounds, the hydroxyl-containing metabolites were severely under-detected, which might indicate that the current method is less than optimal for analyzing this group of metabolites. In short, this study has shown that high metabolome coverage is attainable by analyzing only the H, A, C and K-submetabolomes and the group classification information can be very helpful in determining the area of improvements in chemical labeling methods.

In Chapter 3, as part of metabolome, very short peptides, such as dipeptides and tripeptides, share the same structural units. considering the sequence diversity of tripeptides, the retention time of chemical isotope labeled tripeptides was predicted based on the experimental RT of labeled dipeptides. Compared with using exact mass only to putatively identify tripeptides, the combination of exact mass and predicted RT can reduce the false positive identification. The SMILES files of dipeptides and tripeptides were generated locally and used for calculating the

chemical molecular descriptors. Through recursive feature selection and cross validation, 20 features were selected to build the SVM model with radial kernel. To test if the accurate RT can be predicted by the unlabeled SMILES of peptides, which can greatly simplify the prediction, especially for RT prediction of labeled metabolites, the same workflow was applied on unlabeled SMILES to build the model. In training and test set of dipeptides, the performance of models built by labeled and unlabeled SMILES was similar. But in test sets of tripeptides, RT difference from the labeled-SMILES model is lower than the one from unlabeled-SMILES model. The comparison indicates the labeled SMILES is necessary to provide more accurate prediction. The predicted RT can be used for in-source fragmentation differentiation as well as putative identification. With the combination of exact mass and predicted RT, 329 and 528 tripeptides can be putatively identified in serum and urine samples, respectively.

In Chapter 4, MCID 2.0 was built as an evidence-based library to assist unknown metabolite identification. 76 common biological reactions were employed to predict potential metabolites theoretically. 1,811,882 products were predicted using 11,164 metabolites from KEGG compound database as substrates. For each predicted product, a unique MCID ID was generated to illustrate the relationship of the upstream substrate and corresponding reaction. A corresponding website was built to facilitate users searching the database. We further applied the MCIDxKEGG database to urine samples, and 89.6% peak pairs could be putatively identified.

In Chapter 5, the amine/phenol sub-metabolome of serum samples was profiled for discovering biomarkers in spinal cord injury. 68 serum samples collected at 24 h post-injury were analyzed. A total of 8152 peak pairs were detected. After QC RSD filter, confounding factors correction, the remained peak pairs were used for building modeled to distinguishing different AIS grades and predict neurological conversion as well as motor function outcome. The results indicate

that using amine/phenol sub-metabolome alone is not sufficient to predict conversion and motor outcome.

6.2 Future Work

For chapter 3, we have built the dipeptide library with experimental RT and tripeptide library with predicted RT for short peptide identification. The comparison of unlabeled SMILES and labeled SMILES illustrates the importance of labeled SMILES in RT prediction. In the future work, RT prediction of tetrapeptides and metabolites will also employed the labeled SMILES.

For chapter 4, we used 76 biological reactions to predict potential metabolites based on human metabolome database. First, more reactions can be integrated into the workflow to enlarge the evidence-based metabolome library. Besides, this workflow can also be applied to the metabolome database of various species, the drug database for predicting downstream drug metabolites, the food database, and so on.

For chapter 5, multiple CIL channels, such as carboxylic channel, carbonyl channel, and hydroxyl channel, can be incorporated into the selection of biomarkers. In this way, the combination of biomarkers from different channels may produce sufficient separation between different groups.

Bibliography

- (1) Han, W.; Sapkota, S.; Camicioli, R.; Dixon, R. A.; Li, L. Profiling Novel Metabolic Biomarkers for Parkinson's Disease Using in-Depth Metabolomic Analysis. *Mov. Disord.* **2017**, *32* (12), 1720–1728.
- (2) Wu, Y.; Streijger, F.; Wang, Y.; Lin, G.; Christie, S.; Mac-Thiong, J. M.; Parent, S.; Bailey, C. S.; Paquette, S.; Boyd, M. C.; et al. Parallel Metabolomic Profiling of Cerebrospinal Fluid and Serum for Identifying Biomarkers of Injury Severity after Acute Human Spinal Cord Injury. *Sci. Rep.* **2016**, *6* (1), 38718.
- (3) Huan, T.; Tran, T.; Zheng, J.; Sapkota, S.; MacDonald, S. W.; Camicioli, R.; Dixon, R. A.; Li, L. Metabolomics Analyses of Saliva Detect Novel Biomarkers of Alzheimer's Disease. *J. Alzheimer's Dis.* **2018**, *65* (4), 1401–1416.
- (4) Wishart, D. S. Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nat. Rev. Drug Discov.* **2016**, *15* (7), 473–484.
- (5) Wang, J.; Zhang, T.; Shen, X.; Liu, J.; Zhao, D.; Sun, Y.; Wang, L.; Liu, Y.; Gong, X.; Liu, Y.; et al. Serum Metabolomics for Early Diagnosis of Esophageal Squamous Cell Carcinoma by UHPLC-QTOF/MS. *Metabolomics* **2016**, *12* (7), 1–10.
- (6) Siskos, A. P.; Jain, P.; Römisch-Margl, W.; Bennett, M.; Achaintre, D.; Asad, Y.; Marney, L.; Richardson, L.; Koulman, A.; Griffin, J. L.; et al. Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Anal. Chem.* **2017**, *89* (1), 656–665.
- (7) Enke, C. G. A Predictive Model for Matrix and Analyte Effects in Electrospray Ionization of Singly-Charged Ionic Analytes. *Anal. Chem.* **1997**, *69* (23), 4885–4893.
- (8) Glish, G. L.; Goeringer, D. E. Tandem Quadrupole / Time-of-Flight Instrument for Mass

- Spectrometry / Mass Spectrometry. *Anal. Chem.* **1984**, *56* (13), 2291–2295.
- (9) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Anal. Chem.* **2006**, *78* (7), 2113–2120.
- (10) Jiye, A.; Trygg, J.; Gullberg, J.; Johansson, A. I.; Jonsson, P.; Antti, H.; Marklund, S. L.; Moritz, T. Extraction and GC/MS Analysis of the Human Blood Plasma Metabolome. *Anal. Chem.* **2005**, *77* (24), 8086–8094.
- (11) Song, G.; Qin, T.; Liu, H.; Xu, G. B.; Pan, Y. Y.; Xiong, F. X.; Gu, K. S.; Sun, G. P.; Chen, Z. D. Quantitative Breath Analysis of Volatile Organic Compounds of Lung Cancer Patients. *Lung Cancer* **2010**, *67* (2), 227–231.
- (12) Wang, C.; Kong, H.; Guan, Y.; Yang, J.; Gu, J.; Yang, S.; Xu, G. Plasma Phospholipid Metabolic Profiling and Biomarkers of Type 2 Diabetes Mellitus Based on High-Performance Liquid Chromatography/Electrospray Mass Spectrometry and Multivariate Statistical Analysis. *Anal. Chem.* **2005**, *77* (13), 4108–4116.
- (13) Lei, Z.; Huhman, D. V.; Sumner, L. W. Mass Spectrometry Strategies in Metabolomics. *J. Biol. Chem.* **2011**, *286* (29), 25435–25442.
- (14) Jung, Y.; Lee, J.; Kwon, J.; Lee, K. S.; Ryu, D. H.; Hwang, G. S. Discrimination of the Geographical Origin of Beef by ¹H NMR-Based Metabolomics. *J. Agric. Food Chem.* **2010**, *58* (19), 10458–10466.
- (15) El Hage, M.; Baverel, G.; Conjard-Duplany, A.; Martin, G. Effect of Glucose on Glutamine Metabolism in Rat Brain Slices: A Cellular Metabolomic Study with ¹³C NMR. *Neuroscience* **2013**, *248*, 243–251.
- (16) Sandusky, P.; Raftery, D. Use of Selective TOCSY NMR Experiments for Quantifying

- Minor Components in Complex Mixtures: Application to the Metabonomics of Amino Acids in Honey. *Anal. Chem.* **2005**, *77* (8), 2455–2463.
- (17) Puig-Castellví, F.; Pérez, Y.; Piña, B.; Tauler, R.; Alfonso, I. Comparative Analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR Metabolomics to Understand the Effects of Medium Composition in Yeast Growth. *Anal. Chem.* **2018**, *90* (21), 12422–12430.
- (18) Bonte, R.; Bongaerts, M.; Demirdas, S.; Langendonk, J. G.; Huidekoper, H. H.; Williams, M.; Onkenhout, W.; Jacobs, E. H.; Blom, H. J.; Ruijter, G. J. G. Untargeted Metabolomics-Based Screening Method for Inborn Errors of Metabolism Using Semi-Automatic Sample Preparation with an UHPLC-Orbitrap-MS Platform. *Metabolites* **2019**, *9* (12), 1–18.
- (19) Gray, N.; Lewis, M. R.; Plumb, R. S.; Wilson, I. D.; Nicholson, J. K. High-Throughput Microbore UPLC-MS Metabolic Phenotyping of Urine for Large-Scale Epidemiology Studies. *J. Proteome Res.* **2015**, *14* (6), 2714–2721.
- (20) Olsen, B. A. Hydrophilic Interaction Chromatography Using Amino and Silica Columns for the Determination of Polar Pharmaceuticals and Impurities. *J. Chromatogr. A* **2001**, *913* (1–2), 113–122.
- (21) Strege, M. A. Hydrophilic Interaction Chromatography-Electrospray Mass Spectrometry Analysis of Polar Compounds for Natural Product Drug Discovery. *Anal. Chem.* **1998**, *70* (13), 2439–2445.
- (22) Tolstikov, V. V.; Fiehn, O. Analysis of Highly Polar Compounds of Plant Origin: Combination of Hydrophilic Interaction Chromatography and Electrospray Ion Trap Mass Spectrometry. *Anal. Biochem.* **2002**, *301* (2), 298–307.
- (23) Sriboonvorakul, N.; Leepipatpiboon, N.; Dondorp, A. M.; Pouplin, T.; White, N. J.;

- Tarning, J.; Lindegardh, N. Liquid Chromatographic-Mass Spectrometric Method for Simultaneous Determination of Small Organic Acids Potentially Contributing to Acidosis in Severe Malaria. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2013**, *941*, 116–122.
- (24) Contrepois, K.; Jiang, L.; Snyder, M. Optimized Analytical Procedures for the Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)-Mass Spectrometry. *Mol. Cell. Proteomics* **2015**, *14* (6), 1684–1695.
- (25) Ivanisevic, J.; Zhu, Z. J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. Toward 'Omic Scale Metabolite Profiling: A Dual Separation-Mass Spectrometry Approach for Coverage of Lipid and Central Carbon Metabolism. *Anal. Chem.* **2013**, *85* (14), 6876–6884.
- (26) Guo, K.; Li, L. Differential ¹²C-/¹³C-Isotope Dansylation Labeling and Fast Liquid Chromatography / Mass Spectrometry for Absolute and Relative Quantification of the Metabolome. **2009**, *81* (10), 3919–3932.
- (27) Zhao, S.; Luo, X.; Li, L. Chemical Isotope Labeling LC-MS for High Coverage and Quantitative Profiling of the Hydroxyl Submetabolome in Metabolomics. *Anal. Chem.* **2016**, *88* (21), 10617–10623.
- (28) Guo, K.; Li, L. High-Performance Isotope Labeling for Profiling Carboxylic Acid-Containing Metabolites in Biofluids by Mass Spectrometry. *Anal. Chem.* **2010**, *82* (21), 8789–8793.
- (29) Zhao, S.; Dawe, M.; Guo, K.; Li, L. Development of High-Performance Chemical Isotope Labeling LC-MS for Profiling the Carbonyl Submetabolome. *Anal. Chem.* **2017**, *89* (12), 6758–6765.

- (30) Yuan, B. F.; Zhu, Q. F.; Guo, N.; Zheng, S. J.; Wang, Y. L.; Wang, J.; Xu, J.; Liu, S. J.; He, K.; Hu, T.; et al. Comprehensive Profiling of Fecal Metabolome of Mice by Integrated Chemical Isotope Labeling-Mass Spectrometry Analysis. *Anal. Chem.* **2018**, *90* (5), 3512–3520.
- (31) Russo, M. S. T.; Napylov, A.; Paquet, A.; Vuckovic, D. Comparison of N-Ethyl Maleimide and N-(1-Phenylethyl) Maleimide for Derivatization of Biological Thiols Using Liquid Chromatography-Mass Spectrometry. *Anal. Bioanal. Chem.* **2020**, *412* (7), 1639–1652.
- (32) Allen, D. R.; McWhinney, B. C. Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. *Clin. Biochem. Rev.* **2019**, *40* (3), 135–146.
- (33) Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779–787.
- (34) Zhou, R.; Tseng, C. L.; Huan, T.; Li, L. IsoMS: Automated Processing of LC-MS Data Generated by a Chemical Isotope Labeling Metabolomics Platform. *Anal. Chem.* **2014**, *86* (10), 4675–4679.
- (35) Warrack, B. M.; Hnatyshyn, S.; Ott, K. H.; Reily, M. D.; Sanders, M.; Zhang, H.; Drexler, D. M. Normalization Strategies for Metabonomic Analysis of Urine Samples. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2009**, *877* (5–6), 547–552.
- (36) Wu, Y.; Li, L. Sample Normalization Methods in Quantitative Metabolomics. *J. Chromatogr. A* **2016**, *1430*, 80–95.
- (37) Ryan, D.; Robards, K.; Prenzler, P. D.; Kendall, M. Recent and Potential Developments in

- the Analysis of Urine: A Review. *Anal. Chim. Acta* **2011**, 684 (1–2), 17–29.
- (38) Appenzeller, B. M. R.; Schummer, C.; Rodrigues, S. B.; Wennig, R. Determination of the Volume of Sweat Accumulated in a Sweat-Patch Using Sodium and Potassium as Internal Reference. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2007**, 852 (1–2), 333–337.
- (39) Wu, Y.; Li, L. Determination of Total Concentration of Chemically Labeled Metabolites as a Means of Metabolome Sample Normalization and Sample Loading Optimization in Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **2012**, 84 (24), 10723–10731.
- (40) Wu, Y.; Li, L. Development of Isotope Labeling Liquid Chromatography-Mass Spectrometry for Metabolic Profiling of Bacterial Cells and Its Application for Bacterial Differentiation. *Anal. Chem.* **2013**, 85 (12), 5755–5763.
- (41) Hooton, K.; Han, W.; Li, L. Comprehensive and Quantitative Profiling of the Human Sweat Submetabolome Using High-Performance Chemical Isotope Labeling LC-MS. *Anal. Chem.* **2016**, 88 (14), 7378–7386.
- (42) Chen, Y.; Shen, G.; Zhang, R.; He, J.; Zhang, Y.; Xu, J.; Yang, W.; Chen, X.; Song, Y.; Abliz, Z. Combination of Injection Volume Calibration by Creatinine and MS Signals' Normalization to Overcome Urine Variability in LC-MS-Based Metabolomics Studies. *Anal. Chem.* **2013**, 85 (16), 7659–7665.
- (43) Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* **2012**, 2 (4), 775–795.
- (44) Broadhurst, D. I.; Kell, D. B. Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. *Metabolomics* **2006**, 2 (4), 171–196.
- (45) Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.;

- Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; et al. Proposed Minimum Reporting Standards for Data Analysis in Metabolomics. *Metabolomics* **2007**, *3* (3), 231–241.
- (46) Mahadevan, S.; Shah, S. L.; Marrie, T. J.; Slupsky, C. M. Analysis of Metabolomic Data Using Support Vector Machines. *Anal. Chem.* **2008**, *80* (19), 7562–7570.
- (47) Date, Y.; Kikuchi, J. Application of a Deep Neural Network to Metabolomics Studies and Its Performance in Determining Important Variables. *Anal. Chem.* **2018**, *90* (3), 1805–1810.
- (48) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; et al. Proposed Minimum Reporting Standards for Chemical Analysis. *Metabolomics* **2007**, *3* (3), 211–221.
- (49) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608–D617.
- (50) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388–D1395.
- (51) Pence, H. E.; Williams, A. Chemspider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124.
- (52) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; et al. MyCompoundID: Using an Evidence-Based Metabolome Library for Metabolite Identification. *Anal. Chem.* **2013**, *85* (6), 3401–3408.
- (53) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank

- Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082.
- (54) Fukushima, A. DiffCorr: An R Package to Analyze and Visualize Differential Correlations in Biological Networks. *Gene* **2013**, *518* (1), 209–214.
- (55) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From Genomics to Chemical Genomics: New Developments in KEGG. *Nucleic Acids Res.* **2006**, *34* (Database issue), 354–357.
- (56) Caspi, R.; Billington, R.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Midford, P. E.; Ong, W. K.; Paley, S.; Subhraveti, P.; Karp, P. D. The MetaCyc Database of Metabolic Pathways and Enzymes-a 2019 Update. *Nucleic Acids Res.* **2020**, *48* (D1), D455–D453.
- (57) Jewison, T.; Su, Y.; Disfany, F. M.; Liang, Y.; Knox, C.; MacIejewski, A.; Poelzer, J.; Huynh, J.; Zhou, Y.; Arndt, D.; et al. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic Acids Res.* **2014**, *42* (D1), 478–484.
- (58) Kankainen, M.; Gopalacharyulu, P.; Holm, L.; Orešič, M. MPEA-Metabolite Pathway Enrichment Analysis. *Bioinformatics* **2011**, *27* (13), 1878–1879.
- (59) Xia, J.; Wishart, D. S.; Valencia, A. MetPA: A Web-Based Metabolomics Tool for Pathway Analysis and Visualization. *Bioinformatics* **2011**, *27* (13), 2342–2344.
- (60) Karnovsky, A.; Weymouth, T.; Hull, T.; Glenn Tarcea, V.; Scardoni, G.; Laudanna, C.; Sartor, M. A.; Stringer, K. A.; Jagadish, H. V.; Burant, C.; et al. Metscape 2 Bioinformatics Tool for the Analysis and Visualization of Metabolomics and Gene Expression Data. *Bioinformatics* **2012**, *28* (3), 373–380.
- (61) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504.

- (62) Khamis, M. M.; Adamko, D. J.; El-Aneed, A. Mass Spectrometric Based Approaches in Urine Metabolomics and Biomarker Discovery. *Mass Spectrom. Rev.* **2017**, *36* (2), 115–134.
- (63) Vuckovic, D. Improving Metabolome Coverage and Data Quality: Advancing Metabolomics and Lipidomics for Biomarker Discovery. *Chem. Commun.* **2018**, *54* (50), 6728–6749.
- (64) Zhao, S.; Li, L. Dansylhydrazine Isotope Labeling LC-MS for Comprehensive Carboxylic Acid Submetabolome Profiling. *Anal. Chem.* **2018**, *90* (22), 13514–13522.
- (65) Hao, L.; Johnson, J.; Lietz, C. B.; Buchberger, A.; Frost, D.; Kao, W. J.; Li, L. Mass Defect-Based n,n-Dimethyl Leucine Labels for Quantitative Proteomics and Amine Metabolomics of Pancreatic Cancer Cells. *Anal. Chem.* **2017**, *89* (2), 1138–1146.
- (66) Leng, J.; Wang, H.; Zhang, L.; Zhang, J.; Wang, H.; Guo, Y. A Highly Sensitive Isotope-Coded Derivatization Method and Its Application for the Mass Spectrometric Analysis of Analytes Containing the Carboxyl Group. *Anal. Chim. Acta* **2013**, *758*, 114–121.
- (67) Tayyari, F.; Gowda, G. A. N.; Gu, H.; Raftery, D. ¹⁵N-Cholamine - A Smart Isotope Tag for Combining NMR- and MS-Based Metabolite Profiling. *Anal. Chem.* **2013**, *85* (18), 8715–8721.
- (68) Wong, J. M. T.; Malec, P. A.; Mabrouk, O. S.; Ro, J.; Dus, M.; Kennedy, R. T. Benzoyl Chloride Derivatization with Liquid Chromatography-Mass Spectrometry for Targeted Metabolomics of Neurochemicals in Biological Samples. *J. Chromatogr. A* **2016**, *1446*, 78–90.
- (69) Yuan, W.; Edwards, J. L.; Li, S. Global Profiling of Carbonyl Metabolites with a Photo-Cleavable Isobaric Labeling Affinity Tag. *Chem. Commun.* **2013**, *49* (94), 11080–11082.

- (70) Chu, J. M.; Qi, C. B.; Huang, Y. Q.; Jiang, H. P.; Hao, Y. H.; Yuan, B. F.; Feng, Y. Q. Metal Oxide-Based Selective Enrichment Combined with Stable Isotope Labeling-Mass Spectrometry Analysis for Profiling of Ribose Conjugates. *Anal. Chem.* **2015**, *87* (14), 7364–7372.
- (71) Huan, T.; Li, L. Quantitative Metabolome Analysis Based on Chromatographic Peak Reconstruction in Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry. *Anal. Chem.* **2015**, *87* (14), 7011–7016.
- (72) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; et al. YMDB: The Yeast Metabolome Database. *Nucleic Acids Res.* **2012**, *40* (D1), 815–820.
- (73) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. ECMDB: The E. Coli Metabolome Database. *Nucleic Acids Res.* **2013**, *41* (D1), 625–630.
- (74) Huan, T.; Li, L. Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform. *Anal. Chem.* **2015**, *87* (2), 1306–1313.
- (75) Huan, T.; Wu, Y.; Tang, C.; Lin, G.; Li, L. DnsID in MyCompoundID for Rapid Identification of Dansylated Amine- and Phenol-Containing Metabolites in LC-MS-Based Metabolomics. *Anal. Chem.* **2015**, *87* (19), 9838–9845.
- (76) Huan, T.; Tang, C.; Li, R.; Shi, Y.; Lin, G.; Li, L. MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites. *Anal. Chem.* **2015**, *87* (20), 10619–10626.
- (77) Khavinson, V. K.; Nikolsky, I. S.; Nikolskaya, V. V.; Zubov, D. A.; Galickaya, S. N.;

- Taranuha, L. I.; Semenova, Y. M. A.; Lisica, N. A.; Linkova, N. S.; Butenko, G. M. Effect of Tripeptides on Lymphoid and Stem Cells. *Bull. Exp. Biol. Med.* **2011**, *151* (6), 722–725.
- (78) Campbell, J. D.; McDonough, J. E.; Zeskind, J. E.; Hackett, T. L.; Pechkovsky, D. V.; Brandsma, C. A.; Suzuki, M.; Gosselink, J. V.; Liu, G.; Alekseyev, Y. O.; et al. A Gene Expression Signature of Emphysema-Related Lung Destruction and Its Reversal by the Tripeptide GHK. *Genome Med.* **2012**, *4* (12).
- (79) Fedoreyeva, L. I.; Kireev, I. I.; Khavinson, V. K.; Vanyushin, B. F. Penetration of Short Fluorescence-Labeled Peptides into the Nucleus in HeLa Cells and in Vitro Specific Interaction of the Peptides with Deoxyribooligonucleotides and DNA. *Biochem.* **2011**, *76* (11), 1210–1219.
- (80) Baczek, T.; Wiczling, P.; Marszał, M.; Heyden, Y. Vander; Kaliszan, R. Prediction of Peptide Retention at Different HPLC Conditions from Multiple Linear Regression Models. *J. Proteome Res.* **2005**, *4* (2), 555–563.
- (81) Kaliszan, R.; Baczek, T.; Cimochovska, A.; Juszczak, P.; Wiśniewska, K.; Grzonka, Z. Prediction of High-Performance Liquid Chromatography Retention of Peptides with the Use of Quantitative Structure-Retention Relationships. *Proteomics* **2005**, *5* (2), 409–415.
- (82) Moruz, L.; Tomazela, D.; Käll, L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *J. Proteome Res.* **2010**, *9* (10), 5209–5216.
- (83) Moruz, L.; Staes, A.; Foster, J. M.; Hatzou, M.; Timmerman, E.; Martens, L.; Käll, L. Chromatographic Retention Time Prediction for Posttranslationally Modified Peptides. *Proteomics* **2012**, *12* (8), 1151–1159.

- (84) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Paša-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; et al. Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. *Anal. Chem.* **2003**, *75* (5), 1039–1048.
- (85) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W. J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; et al. Improved Peptide Elution Time Prediction for Reversed-Phase Liquid Chromatography-MS by Incorporating Peptide Sequence Information. *Anal. Chem.* **2006**, *78* (14), 5026–5039.
- (86) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **2018**, *90* (18), 10881–10888.
- (87) Guo, K.; Li, L. Differential ¹²C-/¹³C-Isotope Dansylation Labeling and Fast Liquid Chromatography/Mass Spectrometry for Absolute and Relative Quantification of the Metabolome. *Anal. Chem.* **2009**, *81* (10), 3919–3932.
- (88) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr. Protoc. Bioinforma.* **2014**, No. SUPPL.46, 1–9.
- (89) Zhao, S.; Li, H.; Han, W.; Chan, W.; Li, L. Metabolomic Coverage of Chemical-Group-Submetabolome Analysis: Group Classification and Four-Channel Chemical Isotope Labeling LC-MS. *Anal. Chem.* **2019**, *91* (18), 12108–12115.
- (90) Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T. ChemmineR: A Compound Mining Framework for R. *Bioinformatics* **2008**, *24* (15), 1733–1734.
- (91) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library

- for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12* (17), 2111–2120.
- (92) Wang, X.; Han, W.; Yang, J.; Westaway, D.; Li, L. Development of Chemical Isotope Labeling LC-MS for Tissue Metabolomics and Its Application for Brain and Liver Metabolome Profiling in Alzheimer's Disease Mouse Model. *Anal. Chim. Acta* **2019**, *1050*, 95–104.
- (93) Gu, X.; Al Dubayee, M.; Alshahrani, A.; Masood, A.; Benabdelkamel, H.; Zahra, M.; Li, L.; Abdel Rahman, A. M.; Aljada, A. Distinctive Metabolomics Patterns Associated With Insulin Resistance and Type 2 Diabetes Mellitus. *Front. Mol. Biosci.* **2020**, *7* (December), 1–16.
- (94) Robertson, D. G.; Watkins, P. B.; Reily, M. D. Metabolomics in Toxicology: Preclinical and Clinical Applications. *Toxicol. Sci.* **2011**, *120* (SUPPL.1), 146–170.
- (95) Olesti, E.; González-Ruiz, V.; Wilks, M. F.; Boccard, J.; Rudaz, S. Approaches in Metabolomics for Regulatory Toxicology Applications. *Analyst* **2021**, *146* (6), 1820–1834.
- (96) Huang, W.; Wang, X.; Chen, D.; Xu, E. G.; Luo, X.; Zeng, J.; Huan, T.; Li, L.; Wang, Y. Toxicity Mechanisms of Polystyrene Microplastics in Marine Mussels Revealed by High-Coverage Quantitative Metabolomics Using Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry. *J. Hazard. Mater.* **2021**, *417* (May), 126003.
- (97) Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C. E. Challenges, Progress and Promises of Metabolite Annotation for LC-MS-Based Metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50.
- (98) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. MassBank: A Public Repository for Sharing Mass Spectral

- Data for Life Sciences. *J. Mass Spectrom.* **2010**, *45* (7), 703–714.
- (99) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z. J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2016**, *88* (22), 11084–11091.
- (100) Zhou, Z.; Tu, J.; Zhu, Z. J. Advancing the Large-Scale CCS Database for Metabolomics and Lipidomics at the Machine-Learning Era. *Curr. Opin. Chem. Biol.* **2018**, *42*, 34–41.
- (101) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; et al. KEGG for Linking Genomes to Life and the Environment. *Nucleic Acids Res.* **2008**, *36* (SUPPL. 1), 480–484.
- (102) Pence, H. E.; Williams, A. Chemspider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124.
- (103) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213.
- (104) Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E. J.; Henry, C. S. MINEs: Open Access Databases of Computationally Predicted Enzyme Promiscuity Products for Untargeted Metabolomics. *J. Cheminform.* **2015**, *7* (1), 1–8.
- (105) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification. *J. Cheminform.* **2019**, *11* (1), 1–25.
- (106) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.;

- Wishart, D. S.; et al. MyCompoundID: Using an Evidence-Based Metabolome Library for Metabolite Identification. *Anal. Chem.* **2013**, *85* (6), 3401–3408.
- (107) Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T. ChemmineR: A Compound Mining Framework for R. *Bioinformatics* **2008**, *24* (15), 1733–1734.
- (108) Krueger, H.; Noonan, V. K.; Trenaman, L. M.; Joshi, P.; Rivers, C. S. The Economic Burden of Traumatic Spinal Cord Injury in Canada. *Chronic Dis. Inj. Can.* **2013**, *33* (3), 113–122.
- (109) Pickett, G. E.; Campos-Benitez, M.; Keller, J. L.; Duggal, N. Epidemiology of Traumatic Spinal Cord Injury in Canada. *Spine (Phila. Pa. 1976)*. **2006**, *31* (7), 799–805.
- (110) Dryden, D. M.; Saunders, L. D.; Rowe, B. H.; May, L. A.; Yiannakoulias, N.; Svenson, L. W.; Schopflocher, D. P.; Voaklander, D. C. The Epidemiology of Traumatic Spinal Cord Injury in Alberta, Canada. *Can. J. Neurol. Sci. / J. Can. des Sci. Neurol.* **2003**, *30* (2), 113–121.
- (111) Kang, Y.; Ding, H.; Zhou, H.; Wei, Z.; Liu, L.; Pan, D.; Feng, S. Epidemiology of Worldwide Spinal Cord Injury: A Literature Review. *J. Neurorestoratology* **2017**, *Volume 6*, 1–9.
- (112) Fawcett, J. W.; Curt, A.; Steeves, J. D.; Coleman, W. P.; Tuszynski, M. H.; Lammertse, D.; Bartlett, P. F.; Blight, A. R.; Dietz, V.; Ditunno, J.; et al. Guidelines for the Conduct of Clinical Trials for Spinal Cord Injury as Developed by the ICCP Panel: Spontaneous Recovery after Spinal Cord Injury and Statistical Power Needed for Therapeutic Clinical Trials. *Spinal Cord* **2007**, *45* (3), 190–205.
- (113) Burns, A. S.; Lee, B. S.; Ditunno, J. F.; Tessler, A. Patient Selection for Clinical Trials: The Reliability of the Early Spinal Cord Injury Examination. *J. Neurotrauma* **2003**, *20*

- (5), 477–482.
- (114) Kwon, B. K.; Stammers, A. M. T.; Belanger, L. M.; Bernardo, A.; Chan, D.; Bishop, C. M.; Slobogean, G. P.; Zhang, H.; Umedaly, H.; Giffin, M.; et al. Cerebrospinal Fluid Inflammatory Cytokines and Biomarkers of Injury Severity in Acute Human Spinal Cord Injury. *J. Neurotrauma* **2010**, *27* (4), 669–682.
- (115) Kwon, B. K.; Streijger, F.; Fallah, N.; Noonan, V. K.; Bélanger, L. M.; Ritchie, L.; Paquette, S. J.; Ailon, T.; Boyd, M. C.; Street, J.; et al. Cerebrospinal Fluid Biomarkers to Stratify Injury Severity and Predict Outcome in Human Traumatic Spinal Cord Injury. *J. Neurotrauma* **2017**, *34* (3), 567–580.
- (116) Ma, J.; Novikov, L. N.; Karlsson, K.; Kellerth, J. O.; Wiberg, M. Plexus Avulsion and Spinal Cord Injury Increase the Serum Concentration of S-100 Protein: An Experimental Study in Rats. *Scand. J. Plast. Reconstr. Surg. Hand Surg.* **2001**, *35* (4), 355–359.
- (117) Cornefjord, M.; Nyberg, F.; Rosengren, L.; Brisby, H. Cerebrospinal Fluid Biomarkers in Experimental Spinal Nerve Root Injury. *Spine (Phila. Pa. 1976)*. **2004**, *29* (17), 1862–1868.
- (118) Lubieniecka, J. M.; Streijger, F.; Lee, J. H. T.; Stoynov, N.; Liu, J.; Mottus, R.; Pfeifer, T.; Kwon, B. K.; Coorsen, J. R.; Foster, L. J.; et al. Biomarkers for Severity of Spinal Cord Injury in the Cerebrospinal Fluid of Rats. *PLoS One* **2011**, *6* (4).
- (119) Duan, H.; Ge, W.; Zhang, A.; Xi, Y.; Chen, Z.; Luo, D.; Cheng, Y.; Fan, K. S.; Horvath, S.; Sofroniew, M. V.; et al. Transcriptome Analyses Reveal Molecular Mechanisms Underlying Functional Recovery after Spinal Cord Injury. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (43).
- (120) Streijger, F.; Skinnider, M. A.; Rogalski, J. C.; Balshaw, R.; Shannon, C. P.; Prudova, A.;

- Belanger, L.; Ritchie, L.; Tsang, A.; Christie, S.; et al. A Targeted Proteomics Analysis of Cerebrospinal Fluid after Acute Human Spinal Cord Injury. *J. Neurotrauma* **2017**, *34* (12), 2054–2068.
- (121) Squair, J. W.; Bélanger, L. M.; Tsang, A.; Ritchie, L.; Mac-Thiong, J. M.; Parent, S.; Christie, S.; Bailey, C.; Dhall, S.; Street, J.; et al. Spinal Cord Perfusion Pressure Predicts Neurologic Recovery in Acute Spinal Cord Injury. *Neurology* **2017**, *89* (16), 1660–1667.
- (122) Squair, J. W.; Bélanger, L. M.; Tsang, A.; Ritchie, L.; Mac-Thiong, J. M.; Parent, S.; Christie, S.; Bailey, C.; Dhall, S.; Charest-Morin, R.; et al. Empirical Targets for Acute Hemodynamic Management of Individuals with Spinal Cord Injury. *Neurology* **2019**, *93* (12), E1205–E1211.
- (123) Price, N. D.; Magis, A. T.; Earls, J. C.; Glusman, G.; Levy, R.; Lausted, C.; McDonald, D. T.; Kusebauch, U.; Moss, C. L.; Zhou, Y.; et al. A Wellness Study of 108 Individuals Using Personal, Dense, Dynamic Data Clouds. *Nat. Biotechnol.* **2017**, *35* (8), 747–756.
- (124) Waters, R. L.; Adkins, R. H.; Yakura, J. S.; Sie, I. Motor and Sensory Recovery Following Incomplete Tetraplegia. *Arch. Phys. Med. Rehabil.* **1994**, *75* (3), 306–311.