COMPETITIVE FRAGMENTATION MODELING OF MASS SPECTRA FOR METABOLITE IDENTIFICATION

by

FELICITY ALLEN

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

Doctor of Philosophy

Department of Computing Science
University of Alberta

ABSTRACT

One of the key obstacles to the effective use of mass spectrometry (MS) in high throughput metabolomics is the difficulty in interpreting measured spectra to accurately and efficiently identify metabolites. Traditional methods for automated metabolite identification compare the target MS spectrum to spectra of known molecules in a reference database, ranking candidate molecules based on the closeness of the spectral match. However the limited coverage of available databases has led to interest in computational methods for generating accurate reference MS spectra from chemical structures. This is the target application for this work.

My main research contribution is to propose a method for spectrum prediction, which we call Competitive Fragmentation Modeling (CFM). I demonstrate that this method works effectively for both electron ionization (EI)-MS and electrospray tandem MS (ESI-MS/MS). It uses a probabilistic generative model for the fragmentation processes occurring in a mass spectrometer, and a machine learning approach to learn parameters for this model from data. CFM has been used in both a spectrum prediction task (ie, predicting the mass spectrum from a chemical structure), and in a putative metabolite identification task (ranking possible structures for a target spectrum). In the spectrum prediction task, CFM showed improved performance when compared to a full enumeration of all peaks corresponding to all substructures of the molecule. In the metabolite identification task,

CFM obtained substantially better rankings for the correct candidate than existing methods.

As further validation, this method won the structure identification category of the international Critical Assessment of Small Molecule Identification (CASMI) 2014 competition. The method is also available for general use via a web interface.

PREFACE

Substantial portions of Chapters 5 and 6 were published in the following journal article:

- Allen F., Greiner R., Wishart D., "Competitive Fragmentation Modeling of ESI-MS/MS spectra for putative metabolite identification", Metabolomics, 11 (1): 98-110, 2015.

I was responsible for developing the algorithms, implementing them in code, running all experiments and composing the manuscript. Professor R. Greiner and Professor D. Wishart were supervisory authors, involved with concept formation and manuscript composition.

Some of the results presented in Chapter 6 were published in the following journal article.

- Allen F., Pon A., Wilson M., Greiner R., Wishart D., "CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra", Nucleic Acids Research, 42 (W1): W94-99, 2014.

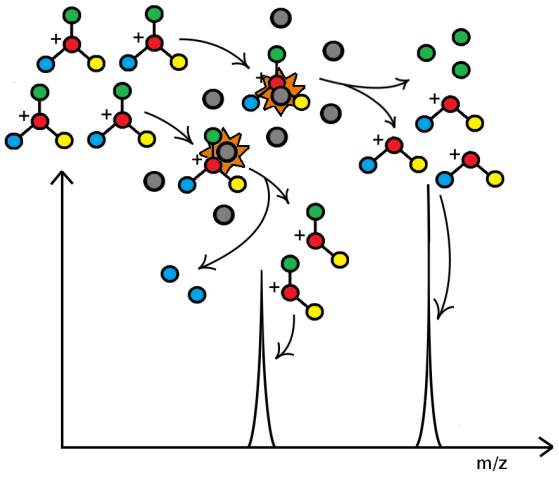
This paper described a web server interface to the methods I developed for the other paper. The web server development, including user interfaces and graphical displays was done by Allison Pon and Michael Wilson. Allison Pon and Michael Wilson also contributed to manuscript composition. I provided the backend code implementations, performed all experiments and composed the manuscript. Professor R. Greiner and Professor D. Wishart were supervisory authors, involved with concept formation and manuscript composition.

The EI-MS extensions and results in Chapters 5 and 7 are planned for publication submission in the near future.

The remaining chapters are my original work. Some phrasings and paragraphs in Chapter 4 may have appeared in the above publications, with minor alterations.

The totality is not, as it were, a mere heap,
but the whole is something besides the parts.

— Aristotle



ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my supervisor Russ Greiner, for his guidance and endless patience throughout my PhD, and for accommodating the slightly unusual circumstances surrounding the latter half of it. His knowledge and life experience were invaluable.

Many thanks to David Wishart, for supporting this work, and providing additional guidance. Also to the members of his lab, in particular Allison Pon, Michael Wilson and Jason Grant, for their work on the CFM-ID web server; and Yannick Djoumbou Feunang for his help with chemical classifications and functional groups.

Many thanks to Dale Schuurmans, for telling me my first idea was rubbish, and setting me on a better path.

Many thanks to Liang Li, Jun Peng and James Harynuk, for sharing their mass spectrometry expertise.

Many thanks to Chris Steinbeck and his group at the European Bioinformatics Institute (EMBL-EBI), for their generous hospitality and for invaluable discussions and advice.

Many thanks to all other members of my supervisory and examination committees, for pointers along the way and for making time in your busy schedules.

Last, but not least, many thanks to all my family and friends for their patience and encouragement. To my husband Joel for many useful discussions, love and support. And to my son Liam, without whom this might have been finished much faster, but with far less enjoyment.

This work was supported by the Natural Sciences and Engineering Research Council of Canada; Alberta Innovates Technology Futures; and Alberta Innovates Health Solutions and made possible by the Compute Canada Westgrid facility.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

MS Mass Spectrometry

MS/MS Tandem Mass Spectrometry

m/z mass to charge ratio

EI Electron Ionization

ESI Electrospray Ionization

GC Gas Chromatography

LC Liquid Chromatography

NN Neural Network

MF Molecular Formula

NL Neutral Loss

HMDB Human Metabalome Database

KEGG Kyoto Encyclopedia of Genes and Genomes

CFM Competitive Fragmentation Modeling

CASMI Critical Assessment of Small Molecule Identification

Part I

SUMMARY

1

INTRODUCTION

1.1 RESEARCH QUESTION

The primary hypothesis motivating this work is that mass spectra can be predicted, at least to some level of accuracy and precision, using computational methods; and that better predictions will result in better metabolite identification performance. Towards this end, the research contributions I have made are described next.

1.2 SUMMARY OF RESEARCH CONTRIBUTIONS

My main research contribution is Competitive Fragmentation Modeling (CFM), a method for mass spectrum prediction. I demonstrate that this method is applicable to both electron ionization (EI)-MS and electrospray tandem MS (ESI-MS/MS). It uses a probabilistic generative model for the fragmentation processes occurring in a mass spectrometer, and a machine learning approach to learn parameters for this model from data. I propose a basic model (see Section 5.1), and several extensions to the basic model (see Sections 5.2-5.5).

I present empirical results for CFM on both a spectrum prediction task (ie, predicting the mass spectrum from a chemical structure), and a putative metabolite identification task (ranking possible structures for a target spectrum), applied to both ESI-MS/MS and EI-MS data. In the spectrum prediction task, CFM showed improved performance when compared to a full enumeration of all peaks corresponding to substructures of the molecule. In a metabolite identification task, CFM obtained substantially better rankings for the correct candidate than existing computational methods. At the time of writing, the only known method for achieving a

better matching spectrum, and hence better metabolite identification performance, is to actually measure each candidate spectrum using a mass spectrometer. Since this is often infeasible or cost-prohibitive, our methods provide a much-needed alternative.

While this work certainly makes contributions in analytical chemistry, there are various reasons why it should also be of interest to computer scientists. One reason is that this particular application was one of the earliest applications proposed for fledgling artificial intelligence methods. In the 1960's, the Dendral Project [1] was created with much the same aims as this work; to computationally predict mass spectra, and to use those predictions to identify chemical compounds. While some progress has been made since then, the problem has largely remained unsolved. While this work cannot claim to have fully achieved such a feat, it does make a substantial step towards this larger goal. By using methods that are the 'bread and butter' of modern computer scientists – probabilistic graphical models, maximum likelihood estimation, expectation maximization and neural networks – this work provides a marker of sorts for how far we have come since the 'expert systems' of Dendral.

Another reason this work may be of interest to computer scientists is that it is, in many ways, a non-trivial application of computer science methods. There is a substantial amount of domain knowledge within the field of mass spectrometry, and it was not obvious how to embed that knowledge within a machine learning method. As shown by a number of previous methods (see Section 4.2), it is not sufficient to simply feed the raw values of mass spectrum peaks and intensities into a supervised machine learning algorithm in order to return the molecule of interest. Instead, methods such as ours that can embed knowledge of how molecules could fragment, and build a machine learning framework around that, seem more likely to succeed.

Another issue was that the learning task was only partially supervised. During training, we know that a mass spectrum corresponds to a particular molecule, but we do not have labels for the fragments that caused the peaks in each spectrum. Consequently, a large part of the learning problem is to first infer (from the spectrum) which fragments occurred, in order to

learn what is likely to happen for an unseen example. We think that our approach, embedding this inference within the overall learning (it is the E-step in our EM), provides an integrated approach to this problem, and there are certainly parallels in other computer science domains. For example, the learning of probabilistic context free grammars (PCFGs), for use in sentence parsing [2, 3, 4], might be likened to our learning of fragmentation rules, for use in spectrum prediction. Similarly, in the domain of plan recognition, the competing goals of an agent are inferred to better explain their actions and thereby predict future actions [5, 6].

Although our overall model is structurally quite different, the log linear modeling that forms the core of our transition function has also been commonly used in a wide range of natural language processing applications [7, 8, 9, 10]. Similar proposals with respect to extending such models with neural networks have also been made in that context [8]. The softmax function has also found common use as the final layer of a neural network [11, 12]. However, to my knowledge, no one has previously applied a softmax function across multiple outputs resulting from different input vectors to the same neural network as we propose in Section 5.5. This is useful in cases where the number of output classes is not fixed, as in our application where the number of possible competing fragmentation events varies, but where some commonality exists between the different classes such that common features can be used to predict their probabilities. This may have applications in other domains, for example in modeling an agent's assessment of similar but competing goals in a plan recognition setting [5, 6].

1.3 CFM-ID WEBSITE

A user-friendly web-server implementation of the methods proposed here is available free of charge at `http://cfmid.wishartlab.com/`. Further details of the available tools are provided in Allen et al. [13]. Many thanks to Allison Pon, Michael Wilson and Jason Grant for their help with the development and ongoing maintenance of the site.

1.4 CASMI 2014 CHALLENGE

To further validate these methods, we applied them as a component in our entry that won the structure identification category of the international Critical Assessment of Small Molecule Identification (CASMI) 2014 competition. This competition asked participants to identify 42 challenge compounds from their ESI-MS and ESI-MS/MS spectra. Meta information was also provided for many of the compounds – e.g. found in blood, or plant-derived product.

We used CFM to provide a score for each candidate compound based on their ESI-MS/MS spectra. We also added two other score components to make use of the other information provided: An isotope-dependent component assessed how well the molecular formula matched the ESI-MS spectra; and a meta information component accounted for which public databases each compound was found in, and how well any descriptions entered there matched the provided meta information.

The combined method correctly identified 23 challenge compounds, and ranked the correct candidate in the top 10 in 33 challenges. It ranked the correct structure highest of all participants in 28 challenges – i.e. the correct compound was not always identified but the rankings obtained were better than those provided by all other participants. For 8 challenges, the correct candidate was never considered by CFM since the correct molecular formula was discarded by the ESI-MS component of our entry.

The format of the competition is still under active development and improvement. Valid concerns surround the provision of the meta information and its use in the ranking of candidates. In quite a number of cases, it could be used to direct the search to such an extent that the ESI-MS/MS scores provided by CFM were near-irrelevant. That said, this is currently the only competition in this area, and it is encouraging that the algorithms proposed here were a part of the winning solution.

1.5 SOURCE CODE AVAILABILITY

Full cross-platform source code, trained models and Windows executables for these algorithms are provided under a GNU Lesser General Public License at http://sourceforge.net/projects/cfm-id/.

1.6 DOCUMENT ORGANISATION

This document is organised into two main sections; background and research contributions. The background section covers Chapters 2 to 4. Chapter 2 outlines the general problem of metabolite identification. Chapter 3 provides an introduction to mass spectrometry. Chapter 4 describes how mass spectrometry is used for metabolite identification, covering both use of reference databases (Section 4.1) and a summary of existing computational methods (Section 4.2). The research contributions section then describes the methods proposed (Chapter 5), and their empirical evaluation on ESI-MS/MS data (Chapter 6) and EI-MS data (Chapter 7).

Part II

BACKGROUND

2

METABOLITE IDENTIFICATION

Metabolites are all the low molecular weight (<1500 Da) chemicals found in cells, tissues and biofluids [14, 15]. Many are endogenous, forming key components of complex regulatory networks that carry out many important life processes, such as growth, reproduction and signaling. Others result from the breakdown of foods, drugs, pesticides and other environmental toxins within the body.

The Human Metabolome Database (HMDB) [16] is a public database that attempts to cover all metabolites found in the human body. At the time of writing, HMDB (v3.6; Oct 2015) contained 41,993 entries. The total number of plant metabolites is estimated to be more than 200,000 [14].

Understanding the roles of metabolites within complex biological processes may be key to the development of new biomonitoring, diagnostic or treatment technologies in areas such as agriculture and healthcare. In order to better study the roles of metabolites, researchers are seeking improved methods that measure metabolites in a high-throughput manner. This has led to the creation of a new field of omics science known as Metabolomics [14, 17], that aims to characterize metabolites accurately and with technologies that enable high-throughput measurements.

Research to date has focused on two underlying platforms capable of performing untargeted, high-throughput measurements of chemical compounds [18]: Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS). Both show enormous promise in this area, but neither has yet emerged as being clearly preferred over the other. Since Mass Spectrometry forms the basis of this work, I limit further discussion to methods employing this technology only. The main reason often touted for using mass spectrometry over other analytical techniques is its sensitivity [19, 20, 21]; measurements can be carried out on just a few micrograms of analyte, and often far less [20].

Note that, while improved metabolite identification was the primary motivator for this work, most of what follows would likely apply to a much wider range of chemical compounds, beyond metabolites. Indeed, the Electron Ionization (EI) mass spectrometry results presented in Chapter 7 use data from the NIST/EPA/NIH 2014 database, which is not restricted to metabolites. The main restriction on our methods is computational – the combinatorial enumeration of fragmentation possibilities becomes infeasible for large input molecules – in particular, for most molecules greater than 1000 Da, or for smaller molecules with more than a couple of rings.

3

MASS SPECTROMETRY

Mass Spectrometry (MS) is a commonly used technique in analytical chemistry [22, 23, 24, 25, 26]. A mass spectrometer is a device that analyses an input chemical sample to determine the mass-to-charge ratios (m/z) of its constituents. The resulting mass spectrum provides a measure of the abundance of particles of each mass-to-charge ratio in the sample of interest. This is generally represented as a graph with m/z on the x-axis. The y-axis is the relative abundance of ions with that m/z, also referred to as the intensity.

Figure 3.1 provides an example spectrum for 1-Methylhistidine. The molecular mass of 1-Methylhistidine is 169.2 Da. The peak at 170.2 Da is the protonated molecular ion; one molecule of 1-Methylhistidine plus an additional proton. The remaining peaks in the spectrum are fragment ions of 1-Methylhistidine, the formation of which will be discussed further in Sections 3.3 and 3.4.

The mass spectrum is also commonly represented as a list of peaks, each defined by its m/z and intensity values. The intensity values are usually normalized such that the highest peak is assigned a relative intensity of 100. For example, the peak list for the spectrum in Figure 3.1 is provided within the figure in the top right corner.

The first use of mass spectrometry is attributed to Sir Joseph J. Thomson in 1912, when he used a magnetic field to deflect Neon ions and measured their deflection using a photographic plate [28]. Since then, a wide range of different instrument types have been designed that are capable of performing mass spectral measurements, each with differing physical mechanisms, technical specifications and limitations. In the words of J. Gross [20], "almost any technique to achieve the goals of ionization, separation and detection of ions in the gas phase can be applied – and actually has been applied – in mass spectrometery". There is also a vast literature containing

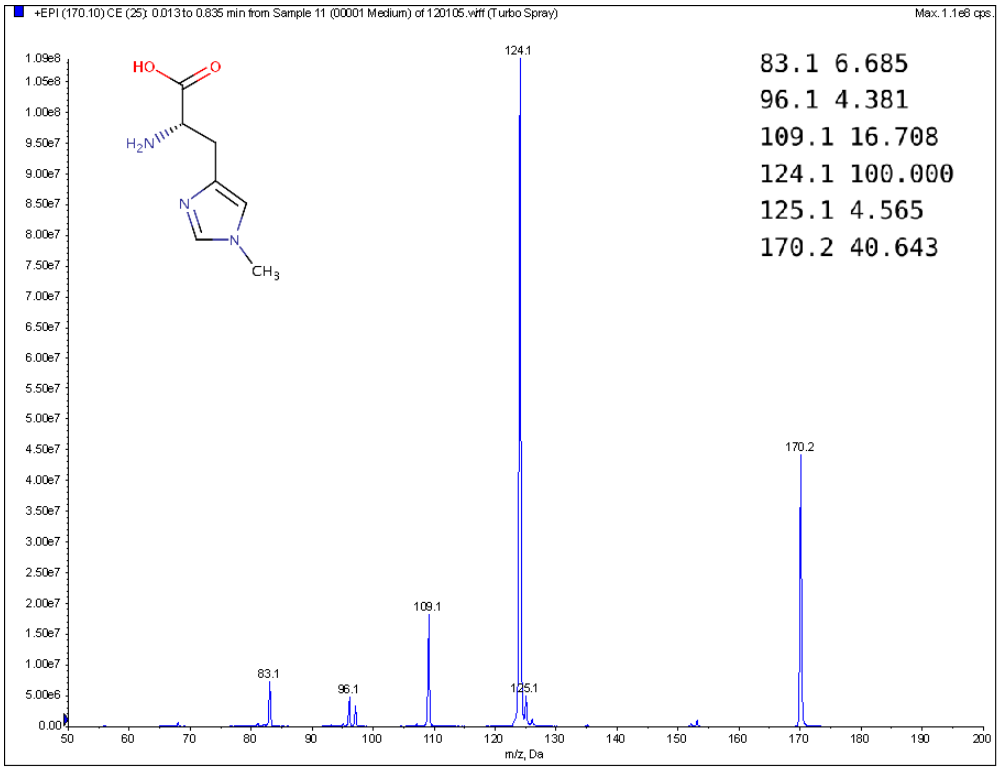


Figure 3.1: Example Mass Spectrum from the Human Metabolome Database [27] for 1-Methylhistidine (HMBD00001), with its corresponding chemical structure (top-left) and peak list (top-right).

methods for sample preparation, ionization, data capture, data processing and various other aspects of mass spectrometry.

For our purposes I give only a brief overview of the main components of a mass spectrometer and some of the more common instrument types in Section 3.1. I also discuss the two particular forms of mass spectrometry, commonly used in metabolomics, which are the targets of this work: EI-MS (Section 3.3) and ESI-MS/MS (Section 3.4). For a more complete overview of mass spectrometry, please see texts such as [23, 24, 25, 26].

3.1 MASS SPECTROMETERS

The main components of a mass spectrometer are an ionization source, a mass analyzer and a detector, all of which are operated within a vacuum, and discussed further in the following sections. A schematic diagram of a mass spectrometer is provided in Figure 3.2.

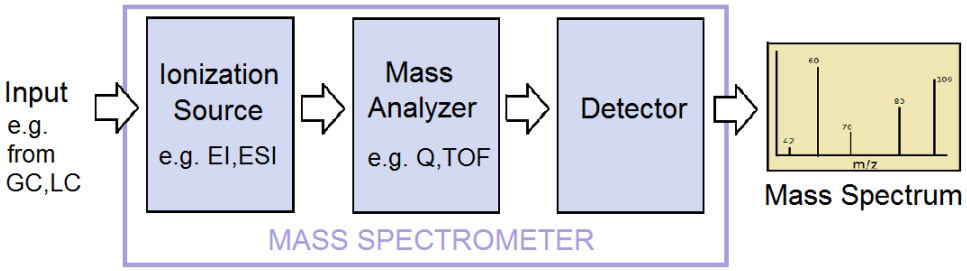


Figure 3.2: Block diagram of a mass spectrometer

3.1.1 *Ionization Source*

Mass spectrometry is fundamentally the analysis of charged particles – i.e. ions. Naturally occurring molecules are not often charged. The ionization source is the means by which the input molecules become charged. There are a wide range of methods used for ionization in mass spectrometry. This work focuses on mass spectrometry based on two commonly used forms of ionization: Electron Ionization (EI) and Electrospray Ionization (ESI), which will be discussed further in Sections 3.3 and 3.4, respectively.

3.1.2 *Mass Analyzer*

The mass analyzer is responsible for physically separating ions according to their m/z. Three of the more commonly used mass analyzer types are Quadrupole, Time-of-Flight and Orbitrap devices, as discussed below. Time-of-Flight and Quadrupole devices have been around since the 1940s

and 50s respectively [24], whereas Orbitrap devices are a far more recent development from the early 2000s [29].

QUADRUPOLE devices apply oscillating electric fields to four parallel rods, configured to allow only molecules of a particular m/z to pass through without colliding with the rods and discharging. These can be used in a selection mode, to select only a narrow range of m/z values to pass through, or in a scanning mode, which scans an m/z range to produce a full mass spectrum, by selecting different molecules with different m/z ratios over time. Quadrupole devices generally have poorer mass accuracy (~100 ppm[1]) [25] than the other instrument types discussed here. However they are well suited to performing mass range selection in MS/MS measurements (see Section 3.4), and are still widely used.

TIME-OF-FLIGHT (TOF) devices accelerate the input ions using an electric field to attain a given kinetic energy, and then analyse m/z values by measuring the time taken for the ions to move through a particular region of space. These instruments sample all the ions at once, rather than requiring a scanning operation, allowing better mass accuracies and faster acquisition rates to be achieved (~10 ppm) [24, 25] than for the quadrupole devices.

ORBITRAP devices trap the input ions in orbits within a chamber around a central electrode using a combination of electrostatic and centrifugal forces. The m/z values are then measured via a fourier transform of the broadband current induced by the oscillating ions. These instruments can achieve even better mass accuracies (~5 ppm) [25], and are also capable of very fast acquisition rates [24].

1 parts per million(ppm) = $10^6 \Delta m/m$

3.1.3 *Detector*

Once the ions have been separated in time or space according to their m/z, it is the responsibility of the detector to detect and quantify the ions. In most modern instruments (except recent fourier transform and orbitrap devices), this is done by converting the kinetic energy of the incident ions into an electric current by collision with a surface that can generate secondary electrons [25].

3.2 CHROMATOGRAPHY

A standard mass spectrometry setup for the analysis of complex biological mixtures usually also includes a chromatographic step to provide an initial separation of the mixture, and thus introduce pure (or near-pure) compounds into the mass spectrometer. This is a method in which the original complex mixture is passed through a tube containing a sorbent material (e.g. silica) prior to entry into the mass spectrometer. In many modern applications, the sorbent material is coated in a thin layer on the inside of a very narrow tube known as a capillary tube. The differing degrees of interaction with the sorbent material cause the different components of the input mixture to be released sequentially, and usually independently, from the tube over time [25].

Multiple mass spectra, each corresponding to a different time instant are captured during the release of each compound. Data processing steps are carried out to combine these spectra into a single time-averaged spectrum for each chromatographic peak. There are a wealth of methods aimed at performing data processing to identify key peaks in the chromatographic spectra, cf. Smith et al. [30], further discussion of which is beyond the scope of this document. For our purposes, the important point is that chromatography makes it possible to obtain mass spectral measurements of relatively pure compounds, even when they are initially contained within complex biological mixtures.

There are two commonly used forms of chromatography: gas chromatography (GC) and liquid chromatography (LC).

GAS CHROMATOGRAPHY (GC) utilizes a gaseous mobile phase, and so requires that the input sample be in the gaseous phase. This usually occurs by thermally vapourizing the sample. A flow of inert gas – e.g. helium – effects the flow of the sample through the tube. The outlet of the tube is usually directly connected to the inlet of the mass spectrometer. The main restriction of gas chromatography is that it is only suitable for the study of thermally stable compounds with a vapour pressure below 350°C – i.e. compounds that can be easilly vaporized and mobilized at temperatures accessible by the instrument, but which are not otherwise altered by heat. To extend the range of suitable compounds, chemical derivatization is often applied to alter the volatility and stability of the compound. This involves chemically altering the compound by replacing polar functional groups such as -NH and -OH with nonpolar groups such as trimethylsilyl (TMS) groups. Gas chromatography is commonly coupled with electron ionization mass spectrometry, as discussed in Section 3.3.

LIQUID CHROMATOGRAPHY is conducted using a liquid mobile phase. This makes it suitable for the analysis of more thermally labile, and non-volatile compounds, including proteins and many metabolites, that are unsuitable for gas chromatography. While traditional liquid chromatography relied on gravity to pass the liquid through a column, when applied in combination with mass spectrometry, liquid chromatography generally refers to High Performance Liquid Chromatography (HPLC), in which the liquid is pressurized to push it through the tube. There are many possible types of liquid chromatography, but most commonly for the applications relevant here, reverse-phase chromatography is used, in which a non-polar stationary phase, made from surface-modified silica, causes the retention times of compounds to depend on their hydrophobicity. Liquid chromatography is most commonly used in combination with electrospray ionization mass spectrometry, as discussed in Section 3.4.

3.3 ELECTRON IONIZATION (EI-MS)

Also known as Electron Impact ionization, Electron Ionization (EI) is the most commonly used ionization method in mass spectrometry. It is often coupled to GC (see Section 3.2), in which case it is called GC-MS, or GC-EI-MS. The gas-phase molecules elute from the GC phase directly into the mass spectrometer ion source, where they are ionized by bombardment with energized (70 eV) electrons. Like GC, EI-MS requires gas-phase ions, so it is subject to the same limitations as for GC; requiring thermally stable input compounds. However it has been used successfully on a wide range of compounds, and is applicable to many metabolites, either directly or in their derivatized form [31, 32].

The energy imparted by the bombarding electrons causes ionization of molecules via the loss of one electron, as described by the following equation.

$$M + e^- \rightarrow M^{+\cdot} + 2e^- \tag{1}$$

The resulting molecular ion (denoted $M^{+\cdot}$) is a positively charged radical (a molecule with an unpaired electron). The EI mass spectrum will generally (though not always), contain a peak at the mass of this ion. After ionization, there is usually sufficient residual energy in the molecular ion to cause it to break into fragments [20, 22, 26]. Some of these fragments will be charged and some will be neutral. The mass spectrum also contains peaks corresponding to the masses of the charged fragments. These values contain information about the structural characteristics of the molecule since they provide the masses of some of its substructures.

The mechanisms by which fragmentation occurs in EI-MS have been well studied [22]. A full description of all the possible mechanisms is beyond the scope of this document; the interested reader is referred instead to Chapter 6 in [20]. For our purposes, we limit discussion to several illustrative examples in the following paragraphs, followed by one of the key principles relevant to this work: the even-electron rule. We also discuss the importance of isotope composition in EI-MS.

3.3.1 *Example 1: α-Cleavage*

α-Cleavage is one of the simplest mechanisms of fragmentation, which results in cleavage of a single bond. One of the two electrons in the cleaved bond moves to the site of ionization on the molecule, while the other remains with the neutral molecule, which detaches. Figure 3.3 provides an example showing this process occurring in acetone[2].

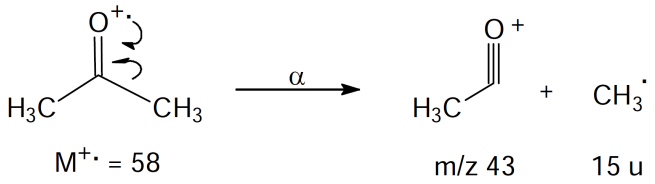


Figure 3.3: α-Cleavage of acetone (from Scheme 6.7 in [20])

The EI mass spectrum for acetone contains peaks corresponding to the molecular ion at 58 Da and the fragment ion from the above α-Cleavage at 43 Da. Note that the 15 Da neutral fragment above will not be detected because it is not charged.

3.3.2 *Example 2: McLafferty Rearrangement*

McLafferty Rearrangement [33] is a more complex fragmentation mechanism resulting in cleavage of a single bond with a concomitant transfer of one hydrogen atom from one side of the broken bond to the other. Figure 3.4 shows the generalized form of a McLafferty Rearrangement.

Atoms A,B, and D can be carbons or heteroatoms (any atom that is not Carbon or Hydrogen). A and B must be connected by a double bond, and a hydrogen must be available at the γ location (as shown). During fragmentation, this hydrogen is transferred to atom B and the β bond is cleaved [20], resulting in alkene loss. Note that the brackets with **+·** indicate that a pos-

2 These are skeletal formula in which carbon atoms are not explicitly depicted, but are instead indicated by otherwise unmarked line-ends or vertices. Most hydrogens are also not shown but are implied by standard valence rules.

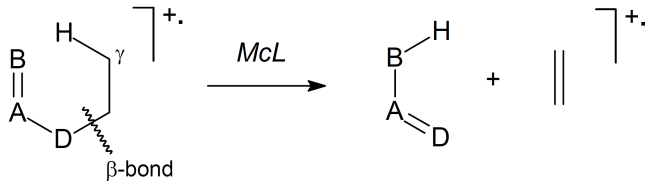


Figure 3.4: McLafferty Rearrangement (from Scheme 6.34 in [20]).

itive charge and radicalization due to electron loss must occur, but at an unspecified location on the associated molecule. The left-hand structure shown may form a substructure of a larger molecule, in which case any additional atoms remain connected in the same way after fragmentation. An extended form of this fragmentation has also been observed that results in the transfer of two hydrogens [20].

3.3.3 *Example 3: Retro-Diels-Alder Reaction*

Retro-Diels-Alder reaction [34] is a fragmentation mechanism that applies specifically to rings. It can occur in almost any molecule containing a six-membered ring with one double bond. The generalized form is shown in Figure 3.5, where once again atoms A,B, and D can be carbons or heteroatoms, and again the left-hand structure may form a substructure of a larger molecule.

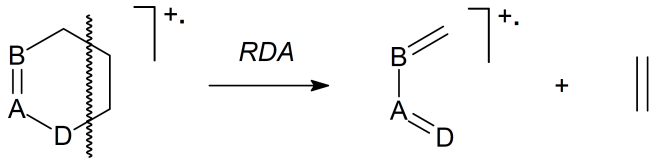


Figure 3.5: Retro-Diels-Alder Reaction (from Scheme 6.46 in [20])

3.3.4 *The Even Electron Rule*

The even electron rule is a more general rule-of-thumb that applies to the vast majority of ions fragmenting in a mass spectrometer. As already noted above, in EI-MS the molecular ion is a radical – i.e. it has one missing electron, giving it an odd number of total electrons. Hence, we say that the molecular ion is odd. Similary, molecules or ions with an even number of total electrons are called even. When the molecular ion breaks apart, the result is a fragment ion (a positively charged molecule) and a neutral fragment. Since the total number of electrons remains the same after fragmentation, one of these two fragments must also be missing an electron, and so must also be odd. The other must be even.

The even electron rule says that either of these options may occur when fragmenting an odd ion. However, when fragmenting something that is even, all electrons must remain paired, and no radicals may result. For EI-MS, this becomes relevant when recursively fragmenting the fragments produced from the molecular ion. The rule is summarized by the following equations (reproduced from [20])

$$
\begin{aligned}
[\mathrm{odd}]^{+\cdot} &\rightarrow [\mathit{even}]^{+} + \mathrm{R}^{\cdot} \\
[\mathrm{odd}]^{+\cdot} &\rightarrow [\mathrm{odd}]^{+\cdot} + \mathfrak{n} \\
[\mathit{even}]^{+} &\rightarrow [\mathit{even}]^{+} + \mathfrak{n} \\
[\mathit{even}]^{+} &\not\rightarrow [\mathrm{odd}]^{+\cdot} + \mathrm{R}^{\cdot}
\end{aligned}
$$

where $\mathfrak{n}$ indicates a neutral, $\mathrm{R}^{\cdot}$ a radical, $\rightarrow$ *occurs* and $\not\rightarrow$ *does not occur*.

3.3.5 *Isotope Composition*

The isotope composition of ions also plays an important role in EI-MS. Isotopes are variants of the same element that have a different number of neutrons, and therefore a different mass. This means that each molecular

ion or fragment ion can exist with multiple masses, and so can produce multiple peaks in the EI mass spectrum.

If we assume that isotope compositions follow the distributions commonly observed in nature, we can expect that – e.g. carbon will occur in its ^{12}C isotope 98.93% of the time, and in its ^{13}C isotope 1.07% of the time [35]. So the mass spectrum of a methyl cation (CH_3^+), which contains only one carbon, should contain a small ^{13}C isotopic peak with 1% intensity relative to the higher ^{12}C peak, and at a mass 1 Da higher. For molecules with more carbon atoms, the probability of at least one carbon being a ^{13}C isotope is higher. For example an ion with molecular formula $C6H7^+$, would have an expected ^{13}C isotopic peak with 6.8% intensity relative to the ^{12}C-only peak.

Some other elements have naturally occurring isotope distributions that result in higher secondary peaks, even with only one or two atoms of that element present in a molecule. For example, chlorine occurs naturally with 75.76% ^{35}Cl and 24.24% ^{37}Cl. Molecules containing only one or two chlorine atoms can be expected to have significant isotopic peaks in their mass spectrum. For example, Figure 3.6 shows the EI mass spectrum for disulfur dichloride (S2Cl2), in which strong isotopic peaks due to chlorine can be clearly seen. Note that each fragment ion causes a cluster of peaks 2 Da apart, rather than a single peak.

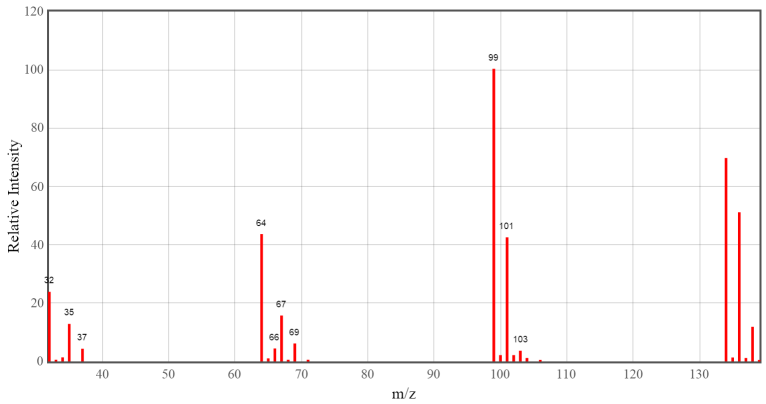


Figure 3.6: EI-MS for disulfur dichloride (Cl2S2) (from NIST WebBook [36])

3.4 ELECTROSPRAY IONIZATION MS/MS (ESI-MS/MS)

Electrospray Ionization (ESI) is an alternative ionization method, which has become the method of choice in protein analysis, and is also increasingly popular in metabolomics. It works by applying a strong electric field to a liquid passing through a capillary tube. Accumulation of charge at the surface of the liquid causes the surface to break, forming charged droplets. The ions of interest are desorbed from the surface of the droplets and pass into the mass spectrometer [25]. This form of ionization is much gentler on the sample, and so is suitable for the analysis of more thermally fragile compounds, including many metabolites and proteins. It is generally used in combination with liquid chromatography (see Section 3.2), in which case it is referred to as LC-ESI-MS.

Unlike EI-MS, the molecular ion resulting from ESI-MS generally becomes charged via the addition of one (or more) protons rather than via the loss of an electron, as described by the following equation in which M again denotes the molecular ion (compare this equation to (1) to note the difference from EI-MS)[3].

$$M + nH^+ \rightarrow [M + nH]^{n+} \tag{2}$$

This has two important effects on the resulting ESI mass spectrum when compared to EI-MS. Firstly, it means that the mass of the molecular ion will be 1 Da higher than in EI-MS. Secondly, it means that the molecular ion is an even electron ion, and so it follows from the even electron rule (see Section3.3.4) that no odd electron fragments (radicals) are expected in ESI-MS. In cases where more than one proton is added, the charge will also be increased accordingly, affecting the denominator of the m/z being measured.

The ESI method of ionization is so gentle that it rarely causes the input molecules to fragment. With sufficient mass accuracy, the unfragmented mass may allow determination of the molecular formula of a compound

3 A proton is equivalent to a hydrogen minus one electron and so is denoted H^+. Once the proton is added to the molecule the charge may or may not stay with the proton, hence the use of brackets on the right hand side.

[37]. However it does not provide additional information about the structure of the compound, which might allow us to distinguish between structural isomers.

Consequently, ESI is usually applied in an extended form of mass spectrometry MS^n that involves additional mass spectrometry phases coupled with Collision Induced Dissociation (CID). The simplest form of this is MS/MS, also known as MS^2, and is the focus of the ESI component of this work.

The ESI-MS/MS process is shown diagrammatically in Figure 3.7. It in-

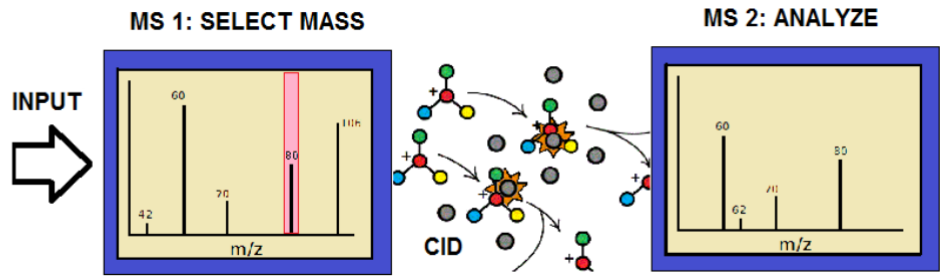


Figure 3.7: ESI-MS/MS: Mass selection occurs in MS1, then after CID, mass analysis occurs in MS2.

volves two phases of mass spectrometry in sequence. In the MS1 phase, a mass spectrum is collected as usual for the input sample, producing peaks corresponding to each compound, which may include isotopic peaks or peaks corresponding to multiply-charged ions – i.e. ions to which more than one proton were added during ionization. A mass selection phase then restricts the device to consider only molecules within a narrow m/z range surrounding a single peak in the MS1 spectrum for further analysis. The selected molecules then undergo CID via interaction with an inert gas, which causes some of the ions to break into fragments. The MS2 phase then measures the (MS/MS) mass spectrum of these fragments. An example annotation of an MS/MS spectrum from HMDB [27] is provided in Figure 3.8.

The extent of fragmentation is in part determined by the collision energy used in CID, which is a configurable parameter in MS/MS appa-

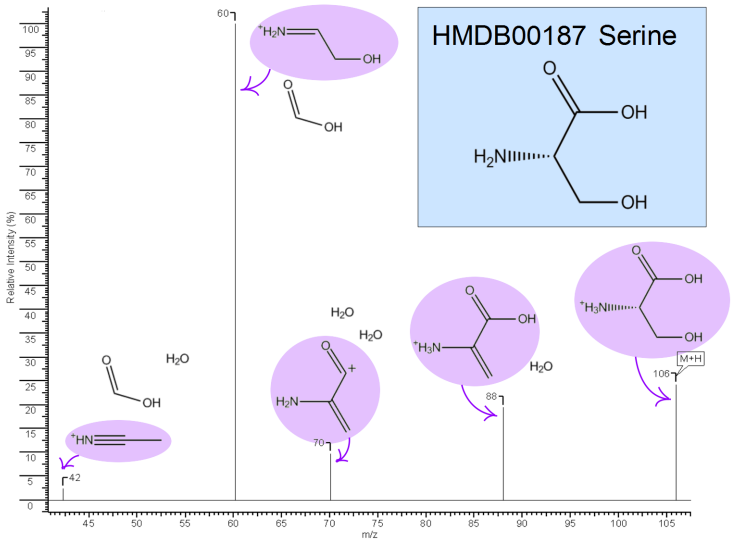


Figure 3.8: Example mass spectrum for serine showing annotations of peaks with putative fragment ions (shaded) and their respective neutral losses (unshaded)

ratus. A higher energy will produce more fragmentations, usually causing molecules to break multiple times into much smaller parts. Whereas a lower energy usually results in many of the original ions remaining intact with only a few larger fragments forming. Often, MS/MS experiments are collected at multiple such collision energies, to give the best coverage of possible fragments, and therefore provide the most structural information about the compound under study.

The MS/MS setup is often achieved by coupling multiple mass spectrometer instrument types together in tandem. For example, the Metlin data we used in the experiments described in Chapter 6 were collected on a Q-TOF device. This uses a quadrupole setup to do the initial mass range selection, followed by a time-of-flight analyser to make high resolution MS/MS measurements.

4

METABOLITE ID USING MASS SPECTROMETRY

4.1 COMPARISON WITH REFERENCE SPECTRA

Traditional methods for putative metabolite identification [38] using mass spectrometry compare a query MS or MS/MS spectrum for an unknown compound against a database containing reference MS or MS/MS spectra [32, 37, 39, 40, 41]. This approach is depicted in Figure 4.1, and is still widely used.

The candidate molecules from the database are ranked according to how similar their spectrum is to the query spectrum, and the best matching candidate(s) are returned. A wide range of similarity criteria have been proposed, from weighted counts of the number of matching peaks [40], to more complex probability-based measures [42, 43].

In cases where the query molecule is contained within the reference database, these methods are found to achieve good accuracy levels. For EI-MS, Stein and Scott [40] found that the correct compound could be identified at rank 1 for 75% of the 12,593 low resolution replicate spectra corresponding to around 8000 compounds (some compounds had multiple spectra) they queried against the NIST-EPA-NIH Mass Spectral Database [44], which contained spectra for 62,235 compounds at that time. For EI-MS/MS, Tautenhahn *et al* [41], reported that 90% correct identifications could be achieved when querying high resolution spectra for 23 metabolite standards against Metlin, which contained spectra for around 10,000 compounds at that time.

Stein [39] lists a number of possible reasons for misidentified compounds, which include poor quality spectra (e.g. due to contaminant peaks or low target concentrations); and fundamental limitations in the ability of mass spectrometry to distinguish between some compounds (e.g. stereoisomers

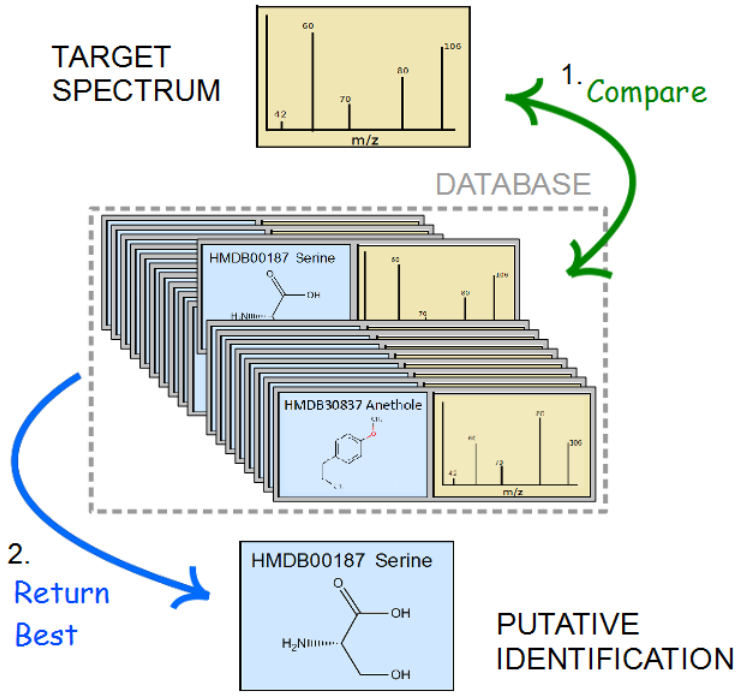


Figure 4.1: Traditional approach to compound identification using mass spectrometry: 1. Compare the spectrum of the target compound with those in a reference database. 2. Return the compound with the closest matching spectrum.

and other isomers with minimal differences in fragmentation, such as aromatic ring positional isomers).

However the main drawback for these methods is that sometimes a reference spectrum for the target compound does not occur in the reference database at all. This is particularly often the case for ESI-MS/MS, for which the current reference databases are still very limited.

At the time of writing, the public Human Metabolome Database [16] contains ESI-MS/MS data for around 2000 compounds, which represents only a small fraction of the 41,993 known human metabolites it lists. The Metlin database [45] provides ESI-MS/MS spectra for 13,048 of the 240,964 endogenous and exogenous metabolites it contains, although more than half of its spectra are for enumerated tripeptides and dipeptides. The public

repository MassBank [46] contains a more diverse dataset of 31,000 spectra collected on a variety of different instruments, including ESI-MS/MS spectra for approximately 2000 compounds. The Global Natural Products Social Networking (GNPS) Library (http://gnps.ucsd.edu) contains ESI-MS/MS spectra for around 4000 compounds. The NIST 2014/EPA/NIH MS/MS library contains ESI-MS/MS spectra for 9,344 compounds.

While these databases are ever-expanding, when set against the more than 63 million chemical structures in the Pubchem Compound database [47], an estimated 200,000 plant metabolites [14], or even the 32,801 manually annotated entries in the database of Chemical Entities of Biological Interest (ChEBI) [48], we see that MS/MS coverage still falls far short of the vast number of known metabolites and molecules of interest.

In the case of EI-MS, the NIST 2014/EPA/NIH EI-MS database contains spectra for over 200,000 compounds, providing a much wider coverage of the chemical space. However it too is struggling to keep pace with the ever-expanding range of compounds, being detected at ever-lower concentrations, as mass spectrometry instrument sensitivities improve [39].

Consequently, there is substantial interest in finding alternative means for identifying metabolites for which no measured reference spectra are available [37].

4.2 COMPUTATIONAL METHODS

The concept of using computer-based methods for mass spectrum interpretation to tackle the compound identification problem has been the focus of research groups since the Dendral project in the 1960s [1]. For some recent reviews in the area, see [49, 50].

Investigators working on Dendral separated the overall problem into a process with three main steps, which they labeled Plan, Generate and Test.

PLAN involved extracting any relevant information from the mass spectrum that could be used to refine the chemical search space.

GENERATE comprised the generation of candidate chemical structures from within that refined search space.

TEST was the final step, which involved predicting a spectrum for each of the candidate structures and comparing it against the target spectrum, in search of the closest match.

The main focus of this research is the Test step, however it is instructive to consider the background literature within the context of all three steps.

4.2.1 *Planning*

The planning step constitutes a very direct approach to the compound identification problem. The input is a mass spectrum. Ideally the output would be the structure itself, rendering the remaining generate and test steps redundant. However, in practice the output has been predictions of chemical class membership or the existence of various substructures or functional groups within the compound. This process is depicted in Figure 4.2.

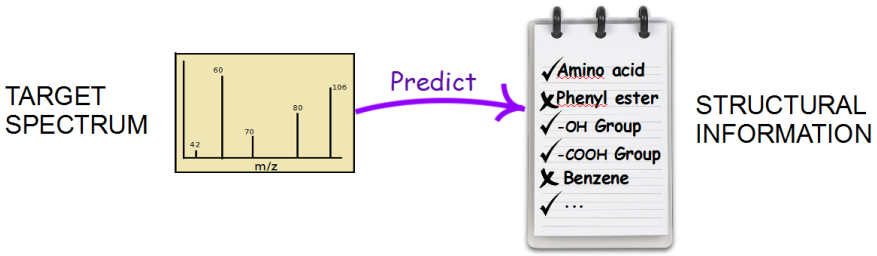


Figure 4.2: Plan step: predict structural characteristics from a mass spectrum

The Dendral project proposed an 'expert-system' by which a series of user-proposed, heuristic rules could be applied to this task – e.g. if a set of specific m/z peaks occur, predict that the molecule contains a specific substructure [1]. In the following decades, a range of machine learning methods were applied to this task in the context of EI-MS. These included

linear discriminant analysis (LDA) [51], neural networks [51, 52, 53, 54] and kNN [55, 56, 57]. Features derived from the input spectrum remained fairly consistent between the methods proposed. These included:

- The existence, or intensity, of a peak at a particular m/z location, for a wide range of integer mass locations.
- The existence, or intensity, of a peak at a particular m/z distance from the molecular ion peak – i.e. corresponding to a particular neutral loss m/z.
- Autocorrelation and series sums of peaks a fixed m/z distance apart,usually between 1 and 50, with extra features targeted to hydrocarbon chains [52, 54].
- Intensity ratios between peaks a fixed m/z distance apart [51].
- Global aspects of the spectrum – e.g. mass centroids, the m/z of the most intense peak, and distributions between odd and even masses [54].
- Indicators specific to particular chemical families – e.g. peaks at 56, 70, 84 and 98 indicating a cyclic amine [54].

These methods appear to have met with limited success. Most were able to achieve similar performance rates [51, 52, 55, 57], identifying between 10-200 chemical classes with precision over 90% (i.e. over 90% of those identified as belonging to a particular class were actually in that class), but with concomitant recall rates closer to 50% (i.e. only half of the compounds in a given class could be identified). Varmuza and Werther [51] reported that to achieve higher precision rates of 95% and 99%, average recall rates dropped to around 30% and 15% respectively. When rejecting classifiers with recall rates lower than 30% for a minimum precision of 90%, they rejected all but 160 of the 600 substructure classifiers they trialled. In another study Klawun and Wilkins [53] combined infrared (IR) data with MS to predict 26 functional groups using a neural network. They found that the classification results obtained were equivalent or better if they ignored the MS data entirely and relied only on IR.

Stein [57] notes that these apparent short-comings are likely due to the absence of clear spectral signature for many structural features. However, he and others also point out that, while this is true for some substructures, there are other substructures that can be clearly and unambiguously identified from mass spectra.

So, despite the apparent limitations of these methods, those select structural classifiers that do achieve good precision and recall rates have been shown to be successful in narrowing the chemical search space [19, 58]. They are a standard part of the NIST 2014/EPA/NIH MS Search [57], and are often applied when no good database matches are found for an unknown spectrum.

In a more recent development in this area, Heinonen *et al.* [59] explored a similar approach in the context of ESI-MS/MS. For their program FingerID, they designed kernels applicable to mass spectra, and used them as features for a set of binary Support Vector Machines (SVMs) to predict the presence or absence of various substructures. The F1 scores obtained appear to be fairly equivalent to those achieved by previous methods. However, unlike previous methods, they chose to favour recall over precision. Precision results are not directly reported, but their best results show an average recall of 94% and an average F1 score of 60%, from which it can be deduced that the average precision must have been 44%.

Rather than using the classifiers to filter out candidates as in previous methods, they used them to predict a chemical fingerprint. Chemical fingerprints are bit vectors in which each bit usually indicates the presence or absence of a particular topological structure in the molecule. Hashed representations of these fingerprints are commonly used in chemical database searches [47, 60]. Heinonen *et al.* used the predicted fingerprint for this purpose, ranking structural candidates by how closely their fingerprints matched the predicted one. In determining the closeness of the fingerprint match, their similarity score weighted the bits corresponding to each SVM output to account for differences in precision. We provide empirical results comparing our method to FingerID in Chapter 6.

This method has also been extended by Shen *et al.* [61] to use a multi-kernel approach to predict the fingerprints. An additional kernel is com-

bined with that used in FingerID, which is based on a method [62] for building molecular formula-based fragment annotations for MS/MS spectra. These annotations result in a fragmentation tree, for which the authors propose various kernels. The method results in some small improvements in the F1 scores for substructure predictions. Very recently, a further extension to this method that uses even more kernels [63] produced impressive results, comparing favorably to our method in a metabolite identification task on ESI-MS/MS data.

Besides the prediction of included substructures, the mass spectrum can also be analysed in the planning step to produce other details of the molecule that may narrow the search space. In particular, it is often feasible to use high resolution MS to refine the set of possible molecular formulae for the unknown. Kind and Feihn [21, 64] proposed using isotope patterns present in mass spectra to rank a list of all candidate molecular formulae with monoisotopic mass in the correct range. They also proposed some heuristic filters (e.g. ratios of carbon to hydrogen) to prune unlikely formulae. Testing with thousands of simulated isotope spectra, they found that they could identify the correct molecular formula, from those remaining after application of their filters, in 80% of test cases for molecules with mass less than 800 Da. By limiting the search space to only those formulae occurring in PubChem the true positive rate increased to 88%. Böcker *et al.* took a very similar approach in their program SIRIUS [65], and were able to deduce the molecular formula correctly for 86% of 153, and 90% of 86 test molecules respectively, when using actual mass spectra measured using two different types of high resolution mass spectrometer.

4.2.2 *Generating*

One of the primary contributions of the Dendral project was an algorithm for exhaustively and non-redundantly generating structural isomers corresponding to a particular molecular formula [1]. The initial algorithm proposed by J. Lederberg was limited to acyclic structures, but it was later extended to include cyclic structures [66]. A range of other alternative meth-

ods were subsequently proposed, details of which are beyond the scope of this discussion. One of the most recent, whose implementation is still available, is MOLGEN [67]. This program enumerates all structural isomers for a molecular formula subject to various user-defined constraints [68].

However, without further restrictions, the number of structural isomers for a given molecular formula grows exponentially with the size of the molecule, rapidly becoming unmanageable for even relatively small molecules. For example, applying MOLGEN to the molecular formula $C_8H_{14}O_3$, which has a mass of just 158 Da, produces 443,628 structural isomers. For many metabolites that are only a little larger, the number of isomers grows into the millions and billions. It would take some serious computing resources to properly consider this volume of candidate molecules, and it seems unlikely that there could be sufficient information in a mass spectrum to distinguish between so many options. Restricting the chemical space of possibilities is a necessity.

Besides the use of substructure prediction methods as discussed in the previous section, an alternative approach to limiting the search space has become feasible with the advent of large public databases. For example, HMDB [16], ChEBI [48] and KEGG [69], to name just a few, contain structures for tens of thousands of molecules that have been reported in biological systems – e.g. in human biofluids, or plant extracts. If the list of candidate substructures is limited to those that occur in these databases, within a small mass range or with a given molecular formula, then the search space is often reduced to just a handful of possibilities.

In many real-world metabolomics investigations, it is reasonable to expect that the molecule of interest will be found in such targeted databases. When this is not the case, a search in PubChem [47] can provide many more candidate molecules, but still far less than the set of all possible isomers. For example, searching for compounds with the molecular formula $C_8H_{14}O_3$, which as noted above retrieves 443,628 isomers from MOLGEN, returns just 2851 molecules from PubChem. While it is still possible for a molecule of interest to be a genuine 'unknown-unknown' – i.e. not even found in PubChem – these situations are relatively rare. Consequently,

database searches are widely used to perform the generate step in contemporary compound identification applications.

4.2.3 *Testing*

Once a set of candidate chemical structures has been obtained – be it from a database, an isomer generator, or any other means as discussed in the previous sections – the test step ranks these candidates based on whether they would be expected to produce the target spectrum. This is generally done as depicted in Figure 4.3. Rather than using measured reference spectra, as was discussed in Section 4.1, since these are often not available, the mass spectrum for each of the candidate compounds is predicted computationally.

4.2.3.1 *Expert Systems and Rule-based Methods*

The Dendral project [1] proposed another 'expert system' for the prediction of EI-MS. Again this involved the collation and application of many user-defined rules, this time to specify how a molecule would fragment – e.g. if the molecule is a ketone then apply McLafferty rearrangement, or if the molecule is a thiol then eliminate water. The rules also contained criteria for setting the intensity of the peaks. Often this was determined by a constant multiplier of the parent ion intensity. In other cases it was more complex – e.g. in a minor alpha cleavage, the intensity also depended on the carbon counts of the resulting ions. The Dendral team also proposed a method called meta-Dendral, that aimed to come up with these rules automatically from data using inductive logic programming.

Several commercial packages now exist that also use a rule-based approach, not unlike that proposed by Dendral. These include Mass Frontier (Thermo Scientific, `www.thermoscientific.com`), and MS Fragmenter (ACD Labs, `www.acdlabs.com`), which each contain thousands of manually curated rules to predict fragmentations. Primarily developed for EI fragmentation, these packages have been extended for use with ESI. MOLGEN-MS [70] also applies rule-based fragmentations in combination with an

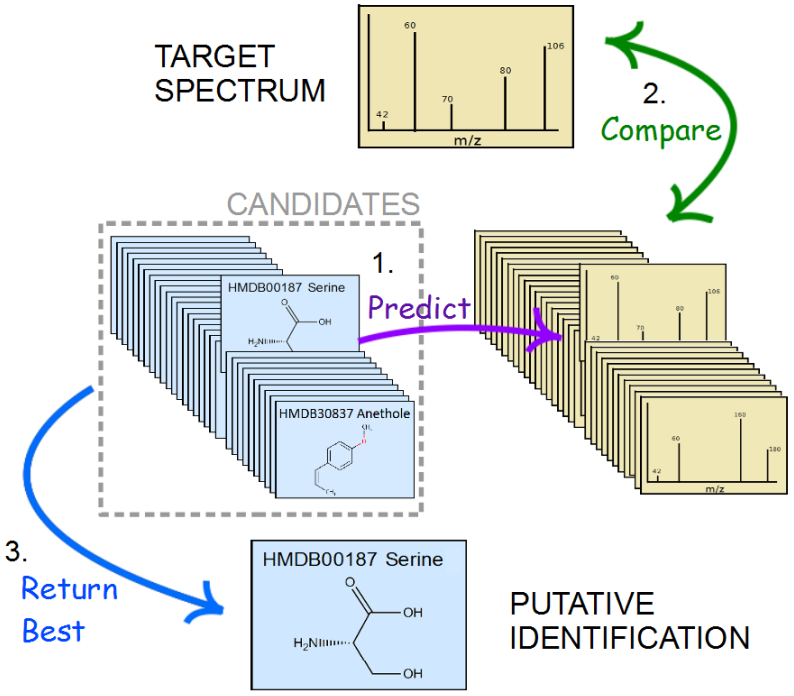


Figure 4.3: Computational approach to compound identification using mass spectrometry: 1. Predict reference spectra for each of a given set of candidate structures. 2. Compare the spectrum of the target compound with all the predicted spectra. 3. Return the compound with the closest matching predicted spectrum.

isotope-dependent matching criteria to rank candidate molecules for a given EI spectrum. All of the above three programs produce so-called 'barcode' spectra, in which all predicted peaks are of equal height (excluding isotope peaks).

As more and more rules have been added to these methods, they have been able to predict more and more fragmentations for any given molecule, – i.e. the recall for the peak locations has increased. In general, improved precision – i.e. whether those predicted peaks actually occur – has been achieved by leaving out some of the fragmentation rules. For example, Schymanski *et al*. [71] found that although ACD MS Fagmenter was able to generate fragments to explain most peaks in any given spectrum, its

identification performance was, on average, no better than random when ranking isomers of 100 target molecules. In the same study, Mass Frontier and MOLGEN-MS achieved better performance but explained less of the target spectrum, and a version of Mass Frontier with 19,000 additional rules resulted in poorer ranking performance than the default program.

4.2.3.2 *Combinatorial Enumeration*

Rather than relying on a large and complicated library of fragmentation rules, another class of algorithms has emerged that apply a combinatorial fragmentation procedure. These algorithms enumerate all possible fragments of the original structure by systematically and recursively breaking all bonds [72, 73, 74]. First proposed by Hill and Mortishire-Smith [72], this approach has been incorporated into the freely available programs FiD [73], MetFrag [74] and MIDAS [75]. All three identify the given spectrum with the metabolite that has the most closely matching peaks via such a combinatorial fragmentation.

These programs are capable of generating large numbers of fragments, and often achieve near-perfect recall – i.e. can provide an explanation for almost any peak in a target spectrum. They have attempted to combat the associated precision problem by employing several heuristics in their scoring protocols to emphasise the importance of more probable fragmentations. FiD uses an approximate measure of the dissociation energy of the broken bond, combined with a measure of the energy of the product ion. MetFrag incorporates a similar measure of bond energy combined with a bonus if the neutral loss formed is one of a common subset. MIDAS uses a plausibility score based on the number of bonds cleaved and whether or not a peak is detected for the parent fragment.

In a similar spirit, another program MAGMa [76] uses a slightly different combinatorial method, enumerating all connected substructures of the input molecule in a non-recursive manner. For each substructure, the minimum set of broken bonds required to create that substructure are broken. Some heuristics are applied to determine the cost of forming each sub-

structure – e.g. +2 for each single bond broken between carbon atoms – and those substructures with the least cost are deemed most likely.

4.2.3.3 *Predicting Fragmentations*

The main problem with both the rule-based and combinatorial methods is that, while they generally have very good recall, explaining most if not all peaks in each spectrum, they also have poor precision, predicting many more peaks than are actually observed. While the heuristics proposed in MetFrag, FiD, MIDAS and MAGMa (see previous section) may go some way to alleviating this problem, it seems likely that there is scope for further improvement.

Several other attempts have been made to derive the likelihood of a given fragmentation event from data. In the 1990's, Gasteiger *et al* used logistic regression [77] and neural networks [78] to predict fragmentation probabilities for α-cleavages (a particular class of EI fragmentations – see Section 3.3) from hand-labeled data. No implementation appears to survive, and while the authors also proposed a method for extracting more general fragmentation patterns from data, as far as this author can ascertain, this method was never successfully applied.

More recently, Kangas et al. [79] proposed a machine learning approach for obtaining bond dissociation energies for lipids. Their method uses a neural net within a kinetic monte carlo simulation, trained using a genetic algorithm. To this author's knowledge, the method has not yet been applied to general classes of metabolites, besides lipids.

Several groups have also attempted to approach the problem from a quantum perspective, applying density functional theory (DFT) calculations to predict sites of protonation and bond cleavage [80, 81]. Unfortunately, the computational cost of these methods seems prohibitive for large numbers of molecules, such that both [80] and [81] provide results for only three small molecules.

In this work I also aim to tackle the precision problem, by predicting which fragmentation events are most likely to occur. Towards this end, I propose a generative model for the MS fragmentation process and a

method for learning parameters for this model from data. The model estimates the likelihood of any given fragmentation event occurring, thereby predicting those peaks that are most likely to be observed in the MS spectrum.

I hypothesise that increasing the precision of the predicted spectrum in this way will improve our ability to accurately identify metabolites. To my knowledge, it is the first such system to be applied to the general class of metabolites.

Part III

RESEARCH CONTRIBUTIONS

5

COMPETITIVE FRAGMENTATION MODELING

This section presents the proposed generative model for the MS fragmentation process, which we call Competitive Fragmentation Modeling (CFM), and describes a method for deriving parameters for this model from MS data. Section 5.1 describes the most basic form of this model, as applied to single energy ESI-MS/MS. The following sections then present various extensions to the basic method to:

- allow for odd electron ions (Section 5.2) and isotopes (Section 5.3) commonly encountered in EI-MS,
- make better use of ESI-MS/MS spectra measured at different collision energies for the same compound (Section 5.4), and
- provide an alternative model parameterization using a neural network (Section 5.5).

5.1 BASIC MODEL

We model ESI-MS/MS fragmentation as a stochastic, homogeneous, Markov process [82] involving state transitions between charged fragments, as depicted in Figure 5.1.

More formally, the process is described by a fixed length sequence of discrete, random fragment states $F_0, F_1, \ldots, F_d$, where each F_i takes a value from the state space $\mathcal{F} := \{f_1, f_2, \ldots, f_{|\mathcal{F}|}\}$, the set of all possible fragments; this state space will be further described in Section 5.1.1. A transition model defines the probabilities that each fragment leads to another at one step in the process; see Section 5.1.2. An observation model maps the penultimate node F_d to a peak P, which takes on a value in $\mathbb{R}$ that

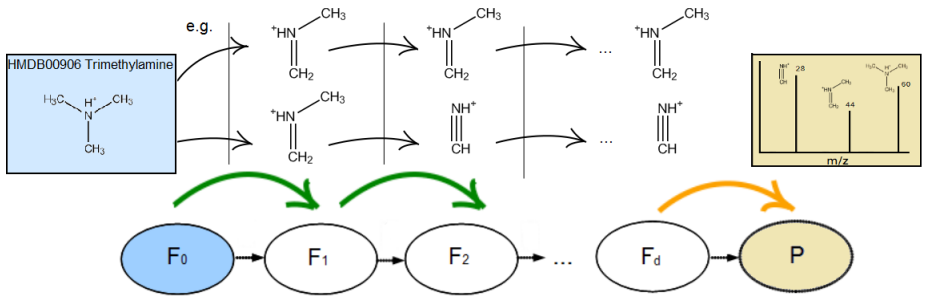


Figure 5.1: Competitive Fragmentation Model (CFM): a stochastic, Markov process of state transitions between charged fragments. The green arrows represent the transition model; see Section 5.1.2. The orange arrow represents the observation model; see Section 5.1.3

represents the m/z value of the peak to which the final fragment will contribute; see Section 5.1.3.

CFM is a latent variable model in which the only observed variables are the initial molecule F_0 and the output peak P; the fragments themselves are never directly observed. Each output P adds only a small contribution to a single peak in the mass spectrum. In order to predict a complete mass spectrum, we can run the model forward multiple times to accumulate a spectrum, or use simple message passing [83] to compute the marginal distribution of P, conditioned on F_0.

5.1.1 *Fragment State Space*

We make the following assumptions about the fragmentation process. Further details for the motivations of each are provided below, but these generally involve a trade-off between accurately modeling the process and keeping the model computationally tractable.

1. All input molecules have a single positive charge when positive ionization is used, or a single negative charge when negative ionization is used, and exist in their most common isotopic form.

2. In a collision, each molecule will break into exactly two fragments.

3. No mass or charge is lost. One of the two fragments must have a single positive charge (or single negative charge for negative ionization) and the other must be neutral. Combined, the two must contain all the components of the original charged molecule, i.e. all the atoms and electrons.

4. No further sigma bonds can be removed or added during a break, except those connecting hydrogens – i.e. the edges in the molecular graph must remain the same.

5. Rearrangement of pi bonds is allowed and hydrogen atoms may move anywhere in the two resulting fragments, on the condition that both fragments satisfy all valence rules, and standard bond limitations are met – e.g. no bond orders higher than triple.

6. The even electron rule (see Section 3.3.4) is always satisfied – i.e. no radicals may be formed.

Assumption 1 is reasonable as we assume that the first phase of MS/MS successfully restricts the mass range of interest to include only the $[M+H]^+$ precursor ion containing the most abundant isotopes. Since this ion has only a single positive charge, we can safely assume that no multiply-charged ions will be formed in the subsequent MS2 phase. Ensuring that valid $[M+H]^+$ precursor ions are selected in MS1 is beyond the scope of this work; see Katajamaa and Oresic [84] for a summary of MS1 data processing methods.

Assumptions 2, 4 and 6 do not necessarily hold in real-world spectra [81, 85]. However including them substantially reduces the branching factor of the fragment enumeration, making the computations feasible. Since these assumptions do appear to hold in the vast majority of cases, we expect that including them should have minimal negative impact on the experimental results. Note that most 3-way fragmentations can be modeled by two sequential, 2-way fragmentations, so including Assumption 2 should not impact our ability to model most fragmentation events. Assumption 5 allows for McLafferty Rearrangement (see Section 3.3.2) and other known fragmentation mechanisms [22].

Our method for enumerating fragments is similar in principle to the combinatorial approach used in MetFrag and FiD [73, 74], with some additional checks to enforce the above assumptions. We systematically break all non-ring bonds in the molecule (excluding those connecting to hydrogens) and all pairs of bonds within each ring. We do this one break at a time, enumerating a subset of fragments with all possible masses that may form after each break, allowing for hydrogen rearrangements.

The previous combinatorial methods assume that all hydrogen rearrangements are possible up to some integer number. Our method differs, as we use integer linear programming (ILP) to ensure that a valid fragmentation results, in terms of the above constraints. This has the additional benefit of ensuring that the resulting fragments can be represented as valid chemical molecules in SMILES format, and within the chemistry development package RDKit [86] upon which my implementation is built.

An ILP is performed twice per break, once for each side of the break, and is formulated as a maximization over the number of bond electrons allocated to that side. The variables in the ILP are the number of electron pairs allocated to each bond, excluding bonds to hydrogen atoms. The linear constraints of the ILP are used to ensure that at least single and at most triple bonds are produced, or at most double bonds if in a ring, and also to enforce the valence constraints imposed by each atom. The result is a value for the maximum number of electrons that can be allocated to each side of the break, and a valid set of locations where they can be added. We then enumerate through all possible allocations of the original electrons between the two sides, subject to these maxima.

For example, if we consider the example given in Figure 5.2, the parent molecule has 6 bonding electron pairs excluding bonds to hydrogen atoms, so the daughter fragments must also have 6 between them. The ILP determines that there is space for at most 3 electron pairs on the left side of the break, and 4 on the right. So this gives two possibilities as shown in the figure, one in which 3 electron pairs are allocated to each side, and another in which 2 are allocated to the left and 4 to the right.

There is often more than one possible allocation of bond electrons within each side that complies with the constraints. Since these differences do not

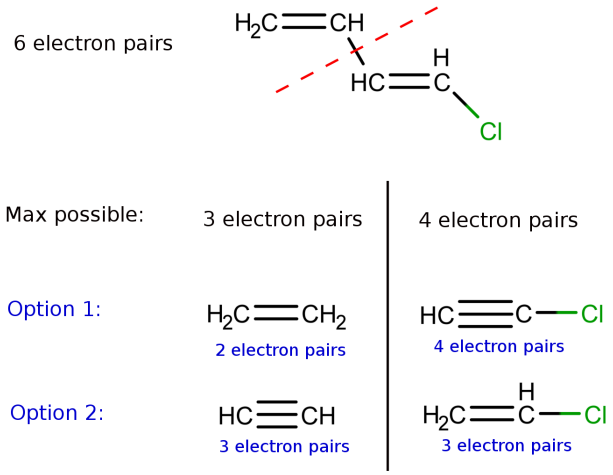


Figure 5.2: An example of an enumeration over possible allocations of the bond electrons between the two sides of a break to form valid fragments. The red dotted line indicates the broken bond.

affect the mass of either fragment, they will be indistinguishable from the mass spectrum and so an arbitrary selection is made. The procedure is applied to a molecule in its neutral state, resulting in two neutral daughter fragments. There are then two-fold more fragmentation options according to which of the two daughter fragments is allocated the charge. A location for the charge is found on each side according to some heuristics (see source code for details – Section 1.5). Aromaticity is not explicitly considered. Instead the fragments are constructed in their kekulized forms and aromaticity detection is performed subsequently by RDKit.

The whole fragmentation procedure is applied recursively on all the produced fragments, to a maximum depth. The result is a directed acyclic graph (DAG) containing all possible charged fragments that may be generated from that molecule. An abstract example of such a fragmentation graph is provided in Figure 5.3. Note that for each break, one of the two produced fragments will have no charge. Since it is not possible for a mass spectrometer to detect neutral molecules, we do not explicitly include the

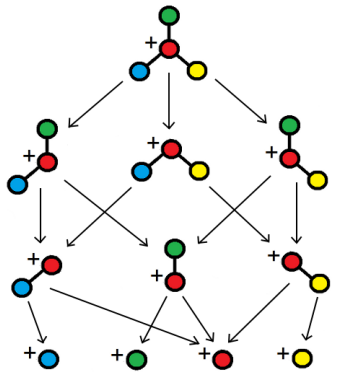


Figure 5.3: An abstract example of a fragmentation graph, showing a directed acyclic graph of all possible ways in which a particular charged molecule may break to produce smaller charged fragments.

neutral fragments in the resulting graph, nor do we recur on their possible breaks. However neutral loss information may be included on the edges of the graph, indicating how a particular charged fragment was determined.

For the Metlin metabolite set used in Chapter 6, the median number of fragments in the fragmentation graph of each molecule is 855, when computed to a fragmentation depth of 2. A depth of 2 is used for all experiments reported in Chapters 6 and 7 to keep computation times feasible. Some fragments may be missed if they require more than two fragmentation events to be achieved –e.g. if three side groups break off a ring structure. Some experiments (not reported here) were carried out to investigate the use of a depth of 3, however this increased computational run-times substantially without appearing to offer much benefit.

5.1.2 *Transition Model*

Our parametrized transition model assigns a conditional probability to each fragment given the previous fragment in the sequence $F_0, F_1, \ldots, F_d$. Recall that F_t denotes the random fragment state at time t, whereas f_i denotes the ith fragment in the space of all fragments. In the case where

f_i has f_j as a possible child fragment in a fragmentation graph, our model assigns a positive probability to the transition from $F_t = f_i$ to $F_{t+1} = f_j$. Furthermore, self-transitions are always allowed, i.e. the probability of transitioning from $F_t = f_i$ to $F_{t+1} = f_i$ is always positive (for the same f_i). We assign 0 probability to all other transitions, i.e. those that are not self-transitions, and that do not exist within any fragmentation graph.

Although the set of possible charged fragments $\mathcal{F}$ is large, encompassing all possible substructures of all possible chemical molecules, the subset of child fragments originating from any particular fragment is relatively small. For example, the requirement that a feasible child fragment must contain a subset of the atoms in the parent fragment rules out many possibilities. Consequently most transitions will be assigned a probability of 0. In fact by definition, if we compute the fragmentation graph (see Section 5.1.1) for each molecule, one at a time, then for each parent fragment in that graph, all the corresponding child fragments with non-zero transition probabilities will also be included in that graph. This means that we need only concern ourselves with the fragmentation graph for one molecule at a time when computing transition probabilities. For the Metlin metabolite set used in Chapter 6, the median number of non-zero transitions in each fragmentation graph is 6096. Note that the assigned probabilities of all transitions originating at a particular fragment, including the self-transition, must sum to one.

We now discuss how we parametrize our transition model. A natural parametrization would be to use a transition matrix containing a separate parameter for every possible fragmentation $f_i \to f_j$. Unfortunately, we lack sufficient data to learn parameters for every individual fragmentation in this manner. Instead, we look for methods that can generalize by exploiting the tendency of similar molecules to break in similar ways.

5.1.2.1 *Break Tendency*

We introduce the notion of *break tendency*, which we represent by a value $\theta \in \mathbb{R}$ for each possible fragmentation $f_i \to f_j$ that models how likely a particular break is to occur. Those fragmentations that are more likely to

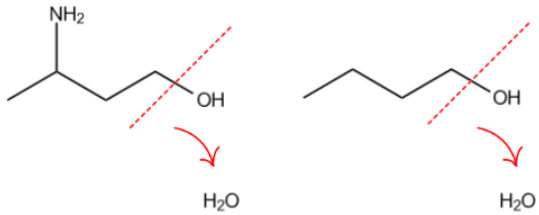


Figure 5.4: Two similar breaks, both resulting in an H_2O neutral loss. The right case should be assigned a higher probability, as in the left case, the NH_3 is also likely to break away, reducing the probability of the H_2O loss.

occur are assigned a higher break tendency value, and those that are less likely are given lower values. We then employ a softmax function to map the break tendencies for all breaks involving a particular parent fragment to probabilities, as defined in Equation 3 below. This has the effect of capturing the competition that occurs between different possible breaks within the same molecule. For example, consider two fragmentations, occurring in two different molecules, as shown in Figure 5.4. Here, both fragmentations involve an H_2O neutral loss, so under simplistic similarity criteria they might be considered to occur with similar probabilities. However, in the left-hand case, the H_2O loss must compete with the loss of an ammonia group, whereas in the right hand case, it does not. Hence our model could assign a similar break tendency to both cases, reflecting their similarity, but this may result in a lower probability of fragmentation in the former case, due to the competing ammonia.

We model the probability of a particular break $f_i \rightarrow f_j$ occurring as a function of its break tendency value $\theta_{i,j}$ and that of all other competing breaks from the same parent, as follows:

$$\rho(f_i, f_j) = \begin{cases} \dfrac{\exp \theta_{i,j}}{1 + \sum_k \exp \theta_{i,k}} & : f_i \neq f_j \text{ and } f_i \rightarrow f_j \text{ is possible} \\ \dfrac{1}{1 + \sum_k \exp \theta_{i,k}} & : f_i = f_j \\ 0 & : f_i \rightarrow f_j \text{ is not possible} \end{cases} \tag{3}$$

where the sums iterate over all k for which $f_i \rightarrow f_k$ is possible.

Since the break tendency is a relative measure, it makes sense to tie it to some reference point. For the purposes of this model, we have assigned the break tendency for a self-transition (i.e. no break occuring) to $\theta_{i,i} = 0$, which gives $\exp \theta_{i,i} = 1$ as shown in (3).

5.1.2.2 *Incorporating Chemical Features*

We need to compute $\theta_{i,j}$ for $i \neq j$. To do this we first define a binary feature vector $\Phi_{i,j}$ to describe the characteristics of a given break $f_i \rightarrow f_j$. Such features might include the presence of a particular atom adjacent to the broken bond, or the formation of a specific neutral loss molecule. For the features used in this work, see Section 6.3.

We then use these features to assign a break tendency value using a linear function parameterized by a vector of weights $w \in \mathbb{R}^n$ – i.e. $\theta_{i,j} := w^T \Phi_{i,j}$. This can then be substituted into (3) to generate the probability of transition $f_i \rightarrow f_j$. The first feature of $\Phi_{i,j}$ is a bias term, set to 1 for all breaks. Note that the vector w constitutes the parameters of the CFM model that we will be learning.

5.1.3 *Observation Model*

We model the conditional probability of P using a narrow Gaussian distribution centred around the mass[1] of F_d, i.e. $P|F_d \sim \mathcal{N}(\text{mass}(F_d), \sigma^2)$. Note that if a ppm measure for mass tolerance is used, the value of σ is dependent on the peak mass m. The value for σ can be set according to the mass accuracy of the mass spectrometer used. So, we define this observation function to be the following

$$g(m, F_d; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{m - \text{mass}(F_d)}{\sigma}\right)^2\right\}. \tag{4}$$

Our investigation (see supplementary data of [87]) of the mass error of the precursor ions in the Metlin metabolite data used in Chapter 6 found that the distribution of mass errors had a mean offset of approximately 1 ppm, and a narrower shape than a Gaussian distribution. However, in order to model a more general mass error, not specific to a particular instrument or set of empirical data, we think the Gaussian distribution is a reasonable approach.

5.1.4 *Parameter Estimation*

We estimate the values for the parameters w of the proposed model by applying a training procedure to a set of molecules $\mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\}$, for which we have both the chemical structure and a measured spectrum.

For the purposes of this work, we assume we have a measured low, medium and high energy MS/MS spectrum for each molecule, which we denote $S(x) = (s_L(x), s_M(x), s_H(x)) \; \forall x \in \mathcal{X}$. Each spectrum is further defined to be a set of peaks, where each peak is a pair (m, h), composed of a mass $m \in \mathbb{R}$ and a height (or intensity) $h \in [0, 100] \subset \mathbb{R}$. Note that each spectrum is normalized, such that the peak heights sum to 100[2].

1 Although mass spectrometry measures mass over charge, we assume charge is always 1 (see Assumption 1 in Section 5.1.1) and hence can just use the mass here.

2 Note that this normalization is non-standard, since mass spectra are often normalised such that the highest peak has height 100. However the alternative normalization scheme makes

For this single energy version of the model, we derive parameters for a completely separate model for each of the three energy levels, using data from that level only. Note that if we had data for only one energy level, we could use this method to train a model using just that energy. However Section 5.4 will extend this model to combine the three energy spectra for use in a single model. Until then, we will use $s(x)$ to denote whichever of $s_L(x)$, $s_M(x)$ or $s_H(x)$ we are currently considering.

5.1.4.1 *Maximum Likelihood*

We use a Maximum Likelihood approach for parameter estimation. The likelihood of the data $\mathcal{X}$, given the parameters w, and incorporating the previously defined transition function ρ and observation function g, is given by

$$\mathcal{L}(w,\mathcal{X}) = \prod_{x\in\mathcal{X}} \prod_{(m,h)\in s(x)} \Big(\sum_{f_1\in C'(x)} \rho(x,f_1;w) \sum_{f_2\in C'(f_1)} \rho(f_1,f_2;w)\ldots \sum_{f_d\in C'(f_{d-1})} \rho(f_{d-1},f_d;w)\; g(m,f_d;\sigma)\Big)^h$$

where $C(f_i)$ denotes the children of f_i in all fragmentation graphs containing it, and $C'(f_i) = \{f_i\} \cup C(f_i)$.

However we are unable to maximize this function in closed form. Instead we use the iterative Expectation Maximization [88] technique.

5.1.4.2 *Expectation Maximization (EM)*

In the E-step, the expected log likelihood expression is given by

$$Q(w^{(t)}, w^{(t-1)} \mid \mathcal{X}) = \mathbb{E}_{w^{(t-1)}}\big(\log \mathcal{L}(w^{(t)},\mathcal{X})\big) \tag{5}$$

$$= \sum_{F_1}\ldots\sum_{F_d} \Pr\big(F_1\ldots F_d \mid \mathcal{X}; w^{(t-1)}\big) \log \mathcal{L}(w^{(t)},\mathcal{X}), \tag{6}$$

sense here so that all molecules have the same total peak mass, and hence emphasis during training.

where $w^{(t)}$ denotes the values for w on the t-th iteration. Substituting (3) and (4) into the above and re-arranging in terms of all possible fragment pairs gives

$$Q(w^{(t)}, w^{(t-1)} \mid \mathcal{X}) \quad = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \log \rho(f_i, f_j; w^{(t)}) + K \tag{7}$$

where

$$\nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \quad = \quad \sum_{d'=1}^{d} \eta^{(d')}_{w^{(t-1)}}(f_i, f_j, \mathcal{X}), \tag{8}$$

$$\eta^{(d)}_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \quad = \sum_{\{(m,h) \in s(x) : x \in \mathcal{X}\}} h \Pr\left(F_{d-1}{=}f_i, F_d{=}f_j \mid F_0{=}x, P{=}m; w^{(t-1)}\right) \tag{9}$$

and

$$K \quad = \sum_{\{(m,h) \in s(x) : x \in \mathcal{X}\}} h \sum_{F_d} \Pr(F_d \mid x; w^{(t-1)}) \log \Pr(P = m \mid F_d). \tag{10}$$

In the M-Step, we look for the $w^{(t)}$ that maximizes the above expression of Q. Noting that K is independent of $w^{(t)}$ and denoting the lth component of w as w_l,

$$\frac{\partial Q}{\partial w_l} \quad = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \left(\mathbb{I}[f_i \neq f_j] \Phi^{(l)}_{i,j} - \sum_{k \in C(f_i)} \Phi^{(l)}_{i,k} \rho(f_i, f_k; w) \right) \tag{11}$$

where $\Phi^{(l)}_{i,k}$ denotes the lth component of the feature vector $\Phi_{i,k}$ and $\mathbb{I}[.]$ is the indicator function.

This does not permit a simple closed-form solution for w. However $Q(w^{(t)}, w^{(t-1)} \mid \mathcal{X})$ is concave in $w^{(t)}$, so settings for $w^{(t)}$ can be found using gradient ascent.

Values for the joint probabilities in the $\eta^{(d)}_{w^{(t-1)}}$ terms can be computed efficiently using the junction tree algorithm [83]. The terms can be computed directly by re-running the junction tree algorithm for each peak in the spectrum as anticipated by a direct reading of Equation 9. Alternatively, the same result can be achieved by treating the spectrum as a marginal observation and using a single application of the Iterative Proportional Fitting

Procedure (IPFP) [89]. Since there is only one marginal observation to be handled – i.e. there is only one spectrum being processed at a time – IPFP always converges within a single iteration.

We also add an ℓ_2 regularizer on the values of w to Q (excluding the bias term). This has the effect of discouraging overfitting by encouraging the parameters to remain close to zero.

5.2 EXTENSIONS FOR ODD ELECTRON IONS

As already noted in Section 3.4, in ESI-MS/MS the precursor ion is an even electron ion, and it follows from the even electron rule (see Section3.3.4) that only even electron ions can occur throughout the resulting fragmentation graph. However in EI-MS, the precursor ion is odd, and so the additional branches of the even electron rule must be incorporated.

So for EI-MS, when enumerating the fragment state space for an odd electron parent ion, we run the ILP solver on the neutral molecule for both sides of the break as usual (see Section 5.1.1). However then, for each allocation of electrons, rather than generating just two possibilities – i.e. the charge on one side xor the other; we generate four possibilities – i.e. all combinations of which side the charge is on and which side the radical is on.

In practical terms, this means that the branching factor of the fragmentation graph produced for EI-MS is roughly double that produced for the same molecule in ESI-MS/MS. This affects run-times, but also means that the number of possible peak locations is increased, further exacerbating the recall vs precision problem that motivates this work (see Section 4.2.3.3).

5.3 EXTENSIONS FOR ISOTOPES

As described in Section 3.3.5, the isotope composition of a molecule can have a significant effect on its EI-MS mass spectrum. Note that this effect is not generally present for ESI-MS/MS spectra because in that case we can assume that the MS1 mass selection filters out any isotopic peaks – i.e.

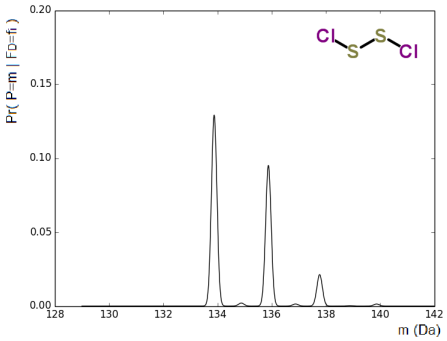


Figure 5.5: Example isotope-based observation function for Cl2S2

peaks due to isotopic variants that are not the most commonly occurring. So for EI-MS, correctly incorporating isotope information into our model could help prevent confusion in the training phase due to isotopic peaks being otherwise mistaken for fragmentations that do not occur. It may also help disambiguate alternative explanations for the same peak, due to differences in the expected isotopic peaks.

Isotopic peaks can be incorporated quite naturally within the observation model of CFM. Rather than associating each fragment with a single peak (see Section 5.1.3), we associate each with a cluster of peaks. We do this by modeling the conditional probability of P using a weighted sum of narrow Gaussian distributions centred around the masses of the peaks in the fragment's expected isotope spectrum. The weights for each Gaussian are set according to the expected abundance of each isotopic species. For example, Figure 5.5 shows the isotope-based observation function for Cl2S2 (the molecule for which we observed isotopic peaks in its EI-MS in Figure 3.6).

So if we denote the expected isotope spectrum for a given fragment ion by $IS(f_i)$ and, similarly to the definition of a mass spectrum used above,

define it to be a set of mass-intensity tuples denoted by (m_{iso}, h_{iso}). Then the observation function from Equation 4 becomes

$$g(m, F_d; \sigma) = \sum_{(m_{iso}, h_{iso}) \in IS(F_d)} \frac{h_{iso}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{m - m_{iso}}{\sigma}\right)^2\right\}, \quad (12)$$

where the h_{iso} values in the expected isotope spectrum are normalized to sum to 1.

Computing the expected isotope spectrum for a given molecular formula efficiently is non-trivial. A number of algorithms have been proposed, generally involving a step-wise application of convolution operations to subcomponents of the molecule, either using a Fast Fourier Transform (FFT) or directly. These include those used in the programs emass [90], Sirius [65], Fourier [91] and Brain [92]. I use Rockwood's program emass [90] to compute the expected isotope spectrum. I threshold the result to only include isotopic peaks with a relative intensity of 1% or more, when the isotope spectrum is normalised such that all peaks sum to 1.

Predicting the spectrum using this new observation model can still be done by using simple message passing to produce the marginal distribution of P. For the computation of the $\eta^{(d)}_{w^{(t-1)}}$ terms (9) during parameter training, the different isotopes of the same fragment are effectively considered as different fragments; their marginal probabilities are computed independently using IPFP as usual. These marginal probabilties are then accumulated across the isotopes of each fragment to give the overall marginal probability, subject to normalization as required.

To see how this works, consider a simple theoretical example. Suppose we have two candidate fragments with isotope spectra as shown in Figure 5.6(a) and (b) respectively, for the target observed isotope spectrum shown in Figure 5.6(c). Note that Option A has an isotope spectrum that is identical to that of the observed spectrum. If we were to disregard the secondary isotopic peaks of both the candidates and the observed spectrum as shown in Figure 5.6(d) – i.e. just consider the peaks at m; then assuming equal priors on fragment Option A and fragment Option B, both would be considered equally likely. If we include the isotopic peaks, as shown in Fig-

ure 5.6(e), and again assume equal priors (including all isotopic variants within each prior), the mass m fragment of Option A would be considered less likely than that of Option B, with marginal probabilties,

$$\Pr(F_d = A|P = m) = 0.375$$
$$\Pr(F_d = B|P = m) = 0.625.$$

However, since there is no other explanation for the $m+1$ peak in the observed spectrum

$$\Pr(F_d = A|P = m+1) = 1.0,$$

and so the accumulated probability of Option A given the observed spectrum S is

$$\begin{aligned}\Pr(F_d = A \,|\, S) &= \sum_{(m',h)\in S} h \Pr(F_d = A \,|\, P = m') \\ &= 0.6 \Pr(F_d = A \,|\, P = m) + 0.4 \Pr(F_d = A \,|\, P = m+1) \\ &= 0.625,\end{aligned}$$

as compared to 0.375 for Option B.

5.4 EXTENSIONS FOR MULTIPLE COLLISION ENERGIES

ESI-MS/MS spectra are often collected at multiple collision energies for the same molecule. Increasing the collision energy usually causes more fragmentation events to occur. This means that fragments appearing in the medium and high energy spectra are almost always descendants of those that appear in the low and medium energy spectra, respectively. So the existence of a peak in the medium energy spectrum may help to differentiate between explanations for a related peak in the low or high energy spectra.

For this reason, we also assessed an additional model, Combined Energy CFM (CE-CFM), which extends the SE-CFM concept by combining information from multiple energies as shown in Fig. 5.7. P_{LOW}, P_{MED} and P_{HIGH}

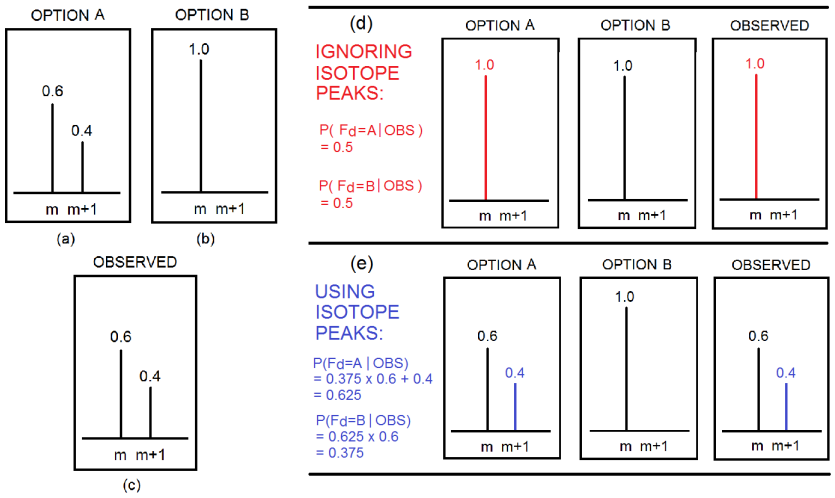


Figure 5.6: Example observed isotope spectrum (c) and the isotope spectra of (a) and (b); two candidate fragment options for F_d. (d) and (e) show the calculation of the marginal for F_d given the observed spectrum when excluding the isotope peaks, and when taking the isotope peaks into account, respectively.

each represent a peak from the low, medium and high energy spectrum respectively.

The fragment states, transition rules and the observation model are all the same here as for SE-CFM. The main difference now is that the homogeneity assumption is relaxed so that separate transition likelihoods can be learned for each energy block – i.e., F_0 to F_{d_L}, F_{d_L} to F_{d_M} and F_{d_M} to F_{d_H}, where d_L, d_M and d_H denote the fragmentation depths of the low, medium and high energy spectra respectively. This results in separate parameter values for each energy, denoted respectively as w_L, w_M and w_H. The complete parameter set for this model thus becomes $w = w_L \cup w_M \cup w_H$.

We can again use a Maximum Likelihood approach to parameter estimation based on the EM algorithm. This approach deviates from the SE-CFM method only as follows:

- For each energy level, (11) is computed separately, restricting the $\nu_{w^{(t-1)}}$ terms to relevant parts of the model – e.g. d' would sum

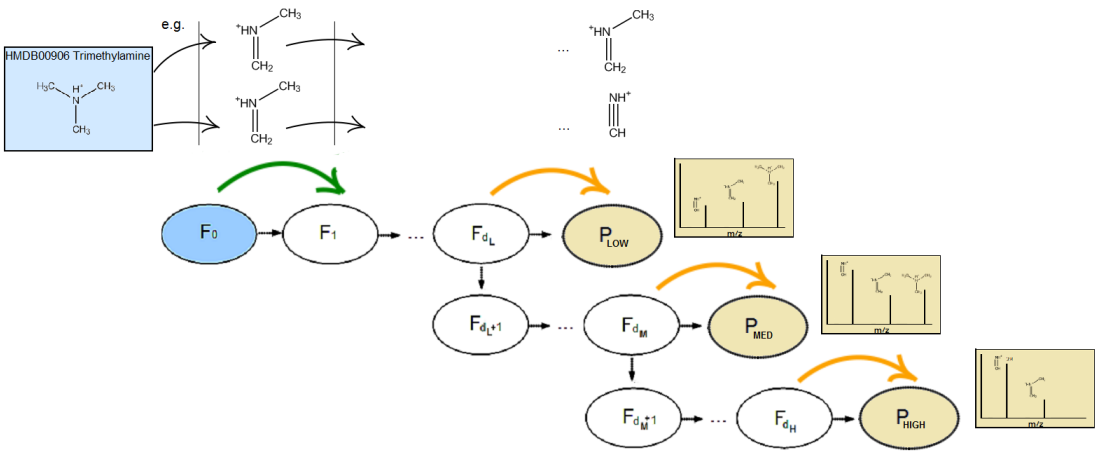


Figure 5.7: Combined Energy Competitive Fragmentation Model (CE-CFM) combines information from multiple collision energy spectra into one model. PLOW, PMED and PHIGH each represent a peak from the low, medium and high energy spectrum respectively.

from d_L+1 to d_M when computing the gradients for w_M, and from d_M+1 to d_H when computing gradients for w_H.

- The computation of the $\eta^{(d)}_{w_{(t-1)}}$ terms combines evidence from the full set of three spectra $S(x)$. In SE-CFM, we apply one spectrum at a time, effectively sampling from a distribution over the peaks from each observed spectra. In this extended model we cannot do this because we do not have a full joint distribution over the peaks, but rather we only have marginal distributions corresponding to each spectrum. The standard inference algorithms – e.g. the junction tree algorithm – do not allow us to deal with observations that are marginal distributions rather than single values. We can again use IPFP, however this time we have more than one observed marginal, since there is more than one spectrum, and so convergence will no longer occur in one iteration. In fact, it is only guaranteed to converge if the spectra are consistent (simultaneously achievable under some joint distribution), which is not always the case here. For example, if the precursor peak happened to be larger in the medium spectrum than in the

low spectrum, IPFP will be unable to reconcile these marginals, since the model cannot put molecules back together once they break. To accommodate such inconsistent cases, I use a small modification to IPFP that reassigns the target spectra to be the average of those encountered when the algorithm oscillates in such circumstances. This comes with no guarantees, but appears to work well in practice.

5.5 NEURAL NET EXTENSION

In the basic CFM model, recall that the break tendency $\theta_{i,j}$ for a given fragmentation $f_i \rightarrow f_j$ was a linear function of its chemical features $\Phi_{i,j}$ (see Section 5.1.2.2). A natural extension is to replace this linear function with a more complex function computed using an artificial neural network. Towards that end, we let $\theta_{i,j}$ be the output of a multilayer perceptron, for which the inputs are given by the feature vector $\Phi_{i,j}$.

In what follows, we use the following neural network notation, from [12]:

- σ is the activation function for each neuron,
- b_j^l is the bias of the jth neuron in the lth layer,
- $a_j^l = \sigma(z_j^l)$ is the activation of the jth neuron in the lth layer,
- $z_j^l = \sum_k w_{j,k}^l a_k^{l-1} + b_j^l$,

Omission of any of the indices indicates a vector or matrix covering all possible indices – e.g. a^l denotes the vector of all activation functions of the neurons in the lth layer, $a^l = [a_1^l, \cdots a_K^l]$.

Using a network with L layers and only one output, we set $a^0 = \Phi_{i,j}$ and $\theta_{i,j} = a^L$. The parameters of the model are again specified by w.

In order to estimate these parameters, we use EM as before, but employ a modified form of the backpropagation algorithm to compute the partial gradients $\frac{\partial Q}{\partial w^l}$ in the M step, as described in Section 5.5.1. The E-step proceeds as before, but uses the neural network to compute the $\theta_{i,j}$ values on each iteration. Unfortunately the inclusion of the neural network

means that the expected likelihood objective used in the M step is no longer convex, so gradient ascent is not guaranteed to converge to the global optimum. However, as will be seen in Chapter 7, it appears to work reasonably well in practice. Note that we have so far only applied this neural network extension to EI-MS data – see Chapter 7 for these results. The ESI-MS/MS experiments in Chapter 6 have so far only included the original linear transition function.

5.5.1 *Modified Backpropagation*

In standard backpropagation (see [12]), the gradient terms can be computed and accumulated for each training instance independently. In our case the training instances for the neural network are the individual fragmentation events, since we are using the network to compute the break tendency value for each possible fragmentation. Unfortunately our use of the softmax function to normalize competing fragmentations means that the backpropagation equations for our model contain additional terms that are dependent on the computations of the competing $\theta_{i,j}$ values. This means that the standard backpropagation equations do not apply.

However we can still formulate a modified form of the backpropagation equations for our Q objective as follows. These equations allow the gradient terms to be efficently computed and accumulated independently for each set of competing fragmentations.

First define

$$\delta^L = \frac{\partial Q}{\partial z^L} = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X})\, \delta^{L, \Phi_{i,j}} \tag{13}$$

where

$$\begin{aligned}
\delta^{L,\Phi_{i,j}} &= \delta_A^{L,\Phi_{i,j}} - \sum_{k \in C(f_i)} \delta_B^{L,\Phi_{i,k}} \\
\delta_A^{L,\Phi_{i,j}} &= \sigma'(z^L(\Phi_{i,j})) \\
\delta_B^{L,\Phi_{i,j}} &= \sigma'(z^L(\Phi_{i,j}))\rho(f_i, f_j; w),
\end{aligned}$$

$z^L(\Phi_{i,j})$ denotes the value of z^L when $\Phi_{i,k}$ is input to the network, and σ' denotes the derivative of the activation function with respect to its input.

Then similarly,

$$\delta^{L-1} = \frac{\partial Q}{\partial z^{L-1}} = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X})\, \delta^{L-1,\Phi_{i,j}} \tag{14}$$

where

$$\begin{aligned}
\delta^{L-1,\Phi_{i,j}} &= \delta_A^{L-1,\Phi_{i,j}} - \sum_{k \in C(f_i)} \delta_B^{L-1,\Phi_{i,k}} \\
\delta_A^{L-1,\Phi_{i,j}} &= \sigma'(z^L(\Phi_{i,j}))w^L \odot \sigma'(z^{L-1}(\Phi_{i,j})) \\
&= \delta_A^{L,\Phi_{i,j}} w^L \odot \sigma'(z^{L-1}(\Phi_{i,j})) \\
\delta_B^{L-1,\Phi_{i,j}} &= \sigma'(z^L(\Phi_{i,j}))\rho(f_i, f_j; w)w^L \odot \sigma'(z^{L-1}(\Phi_{i,j})) \\
&= \delta_B^{L,\Phi_{i,j}} w^L \odot \sigma'(z^{L-1}(\Phi_{i,j})),
\end{aligned}$$

and $\odot$ denotes the Hadamard product [12].

So then, by similar argument, the equations for $l < L - 1$ are:

$$\begin{aligned}
\delta^{l,\Phi_{i,j}} &= \delta_A^{l,\Phi_{i,j}} - \sum_{k \in C(f_i)} \delta_B^{l,\Phi_{i,k}} \\
\delta_A^{l,\Phi_{i,j}} &= \sum_k \delta_{A,k}^{l+1,\Phi_{i,j}} w_k^{l+1} \odot \sigma'(z_k^l(\Phi_{i,j})) \\
\delta_B^{l,\Phi_{i,j}} &= \sum_k \delta_{B,k}^{l+1,\Phi_{i,j}} w_k^{l+1} \odot \sigma'(z_k^l(\Phi_{i,j})).
\end{aligned}$$

Then, noting that both the δ_A and δ_B terms are zeros for self-transitions (since the $\theta_{i,i}$ values are fixed at 0.0 in these cases), we can define the gradient equations for the parameters of the neural net as

$$\frac{\partial Q}{\partial b_k^l} = \delta_k^l = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \left(\mathbb{I}[f_i \neq f_j] \delta_{A,k}^{l,\Phi_{i,j}} - \sum_{k \in C(f_i)} \delta_{B,k}^{l,\Phi_{i,k}} \right)$$

and

$$\frac{\partial Q}{\partial w_{k,k'}^l} = \sum_{(f_i, f_j) \in \mathcal{F} \times \mathcal{F}} \nu_{w^{(t-1)}}(f_i, f_j, \mathcal{X}) \left(\mathbb{I}[f_i \neq f_j] a_{k'}^{l-1}(\Phi_{i,j}) \delta_{A,k}^{l,\Phi_{i,j}} - \sum_{j' \in C(f_i)} a_{k'}^{l-1}(\Phi_{i,j'}) \delta_{B,k}^{l,\Phi_{i,j'}} \right).$$

These equations provide the required gradients of the model parameters with respect to the expected log-likelihood.

6

EMPIRICAL EVALUATION OF ESI-MS/MS

In this section we present results using the above described SE-CFM and CE-CFM methods, on a spectrum prediction task, and then in a metabolite identification task.

6.1 DATA

We used the Metlin database [45], separated into two sets (see description below) each containing positive mode, ESI-MS/MS spectra from a 6510 Q-TOF (Agilent Technologies) mass spectrometer, measured at three different collision energies: 10V, 20V and 40V, which we assign to be low, medium and high energy respectively. Each set was randomly divided into 10 groups for use within a 10-fold cross validation framework.

1. **Tripeptides**: The Metlin database contains data for over 4000 enumerated tripeptides. We randomly selected 2000 of these molecules, then omitted 15 that had four or more rings due to computational resource concerns[1], leaving 1985 remaining in the set. Fragmentation patterns in peptides are reasonably well understood [93, 94], leading to effective algorithms for identifying peptides from their ESI MS/MS data – e.g. [95, 96, 97]. However, we think that the size of this dataset, and the fact that it contains so many similar yet different molecules, make it an interesting test case for our algorithms.

2. **Metlin Metabolites**: We use a set of 1491 non-peptide metabolites from the Metlin database. These are a more diverse set covering a much wider range of molecules. An initial set of 1500 were selected randomly. Nine were then excluded because they were so

1 The fragmentation graph computation for these molecules ran for many hours without completing

much larger than the other molecules (over 1000 Da), such that their fragmentation graphs could not be computed in a reasonable amount of time.

We also used two additional validation sets. The first was selected because the spectra in it were measured on a similar mass spectrometer to that used to collect the Metlin data, an Agilent 6520 Q-TOF, but in a different laboratory. These were taken from the MassBank database [46]. The second set was selected to explore the case where data with poorer mass accuracy were used on a different type of mass spectrometer and at a slightly different collision energy. All testing with both these sets used a model trained for the first cross-fold set of the Metlin metabolite data ($\sim$ 90% of the data). Mass tolerances were increased to 0.5 Da for the HMDB set during testing to account for the lower mass accuracy.

3. **MassBank Metabolites**: This set contains 192 metabolites taken from the Washington State University submission to the MassBank database. All molecules from this submission were included that had MS2 spectra with collision energies 10V, 20V and 40V, in order to provide a good match with the Metlin data.

4. **HMDB Metabolites**: This set contains 500 molecules from the Human Metabalome database [16], randomly selected from those with MS/MS data available. These spectra were collected using a different mass spectrometer: a Waters Quattro QqQ that has much poorer mass accuracy than the Q-TOF, and a medium collision energy of 25V instead of 30V.

All the data sets above used positive mode ionization. One further data set was used to assess the ability of these algorithms to deal with negative mode ionization. This data was again taken from the Metlin database so was measured on a 6510 Q-TOF (Agilent Technologies) mass spectrometer at 10V, 20V and 40V. As before, the set was randomly divided into 10 groups for use within a 10-fold cross validation framework.

5. **Negative (Ionization Mode Data) Metabolites**: This set contains 976 metabolites, selected randomly from the non-peptide metabolites in

Metlin for which negative ionization mode spectra were available, discarding those with a molecular weight greater than 1000 Da.

Files containing test molecule lists and assigned cross validation groups are provided as supplementary data at http://sourceforge.net/projects/cfm-id/.

6.2 MODEL CONFIGURATION

A depth of 2 was used when expanding the fragmentation graphs (see Section 5.1.1) for both SE-CFM and CE-CFM. This is the same default fragmentation depth used by MetFrag, and while it does exclude some fragmentation possibilities, it appears to strike a reasonable balance between computation time and fragmentation coverage. For SE-CFM a model depth of 2 was also used for the markov process (d=2), and in CE-CFM 2 steps were used between each energy level (d_L=2, d_M=4, d_H=6).

6.3 CHEMICAL FEATURES

The chemical features used in these experiments were as follows. Note that the terms *ion root atom* and *neutral loss (NL) root atom* refer to the atoms connected to the broken bond(s) on the ion and neutral loss sides respectively –cf. Fig. 6.1.

- *Break Atom Pair*: Indicators for the pair of ion and neutral loss root atoms, each from {C, N, O, P, S, other}, included separately for those in a non-ring break vs those in a ring break – e.g. Fig. 6.1(a) would be non-ring C-C. (72 features)
- *Ion and NL Root Paths* Indicators for all paths of length 2 and 3 starting at the respective root atoms and stepping away from the break. Each is an ordered double or triple from {C, N, O, P, S, other}, taken separately for rings and non-rings. Two more features indicate no paths of length 2 and 3 respectively – e.g. in Fig. 6.1(a) the ion root paths are C-O, C-N and C-N-C. (2020 features).

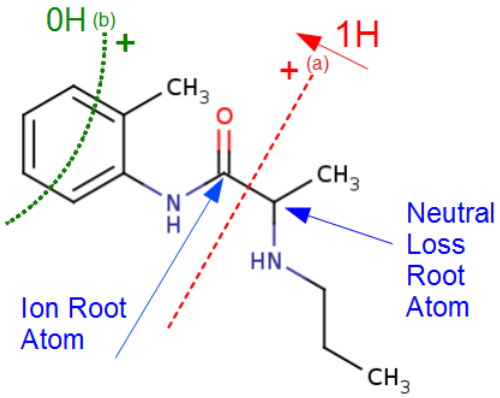


Figure 6.1: Two example fragmentations. (a) A non-ring break for which the ion and neutral loss root atoms are labeled. The 1H indicates the movement of a hydrogen to the ion side (marked with a +) from the neutral loss side. (b) A ring break for a single aromatic ring of size 6, in which the distance between the broken bonds is 3. The 0H indicates no hydrogen movement.

- *Gasteiger Charges*: Indicators for the quantised pair of Gasteiger charges [98] for the ion and NL root atoms in the original unbroken molecule. (288 features)

- *Hydrogen Movement*: Indicator for how many hydrogens switched sides of the break and in which direction – i.e. ion to NL (-) or NL to ion(+) {0,±1,±2,±3,±4,other}. (10 features)

- *Ring Features*: Properties of a broken ring. Aromatic or not? Multiple ring system? Size {3,4,5,6, other}? Distance between the broken bonds {1,2,3,4+}? – e.g. Fig. 6.1(b) is a break of a single aromatic ring of size 6 at distance 3. (12 features).

Of these 2402 features, few take non-zero values for any given break. Many are never encountered in our data set, in which case their corresponding parameters are set immediately to 0. We also append *Quadratic Features*, containing all 2,881,200 pair-wise combinations of the above features, excluding the additional bias term. Again, most are never encountered, so their parameters are set to 0. For example, the model trained on the Metlin metabolite data for cross-validation group 0 used 28,228 parameters per energy level.

6.4 SPECTRUM PREDICTION

For each cross validation fold, and validation set, a model (trained as above), was used to predict a low, medium and high energy spectrum for each molecule in the test set. The model is run forward and the resulting marginal distributions for the peak variables are a mixture of Gaussian distributions. We take the means and weights of these Gaussians as our peak mass and intensity values. Since all fragments in the fragmentation graph of a molecule have non-zero probabilities in the marginal distribution, it is necessary to place a cut-off on the intensity values to select only the most likely peaks. Here, we use a post-processing step that removes peaks with low probability, keeping as many of the highest peaks as required to form at least 80% of the total intensity sum. We also set limits on the number of selected peaks to be at least 5 and at most 30. This ensures that more peaks are included than just the precursor ion, and also prevents spectra occurring that have large numbers of very small peaks. These values were selected arbitrarily, but post-analysis suggests that they are reasonable (see supplementary data for [87]). When matching peaks we use a mass tolerance set to the larger of 10 ppm and 0.01 Da (depending on the peak mass) for all data sets except the HMDB metabolites set, which used a tolerance of 0.5 Da. We set the observation parameter σ to be one third of this value. No additional processing was done for the experimental spectra.

6.4.1 *Metrics*

We consider a peak in the predicted MS/MS spectrum s_P to match a peak in the measured MS/MS spectrum s_M if their masses are within the mass tolerance above. We use the following metrics:

1. **Weighted Recall**: The percentage of the total peak intensity in the measured spectrum with a matching peak in the predicted spectrum: $100 \times \sum_{(m,h)\in s_M} h \cdot \mathbb{I}[(m,h) \in s_P] \div \sum_{(m,h)\in s_M} h.$

2. **Weighted Precision**: The percentage of the total peak intensity in the predicted spectrum with a matching peak in the measured spectrum: $100 \times \sum_{(m,h)\in s_P} h \cdot \mathbb{I}[(m,h) \in s_M] \div \sum_{(m,h)\in s_P} h$.
3. **Recall**: The percentage of peaks in the measured spectrum that have a matching peak in the predicted spectrum: $100 \times |s_P \cap s_M| \div |s_M|$.
4. **Precision**: The percentage of peaks in the predicted spectrum that have a matching peak in the measured spectrum: $100 \times |s_P \cap s_M| \div |s_P|$.
5. **Jaccard Score**: $|s_P \cap s_M| \div |s_P \cup s_M|$.

The intensity weighted metrics were included because the unweighted precision and recall values can be misleading in the presence of low-level noise – e.g. when there are many small peaks in the measured spectrum. The weighted metrics place a greater importance on matching higher intensity peaks, and therefore give a better indication of how much of a spectrum has been matched. However, these weighted metrics can also be susceptible to an over-emphasis of just one or two peaks, and in particular of the peak corresponding to the precursor ion. Consequently, we think it is informative to consider both weighted and non-weighted metrics for recall and precision.

6.4.2 *Models for Comparison*

The pre-existing methods – e.g. MetFrag, FingerID – do not output a predicted spectrum, but skip directly to metabolite identification. So, instead we compare against:

- **Full Enumeration**: This model considers the predicted spectrum to be one that enumerates all possible fragments in the molecule's fragmentation tree with uniform intensity values.
- **Heuristic** (tripeptides only): This model enumerates known peptide fragmentations as described by [93], including b_n, y_n, $b_n - H_2O$, $y_n - H_2O$, $b_n - NH_3$, $y_n - NH_3$ and immonium ions.

6.4.3 *Results*

The results are presented in Figure 6.2. For all data sets tested, SE-CFM and CE-CFM obtain several orders of magnitude better precision and Jaccard scores than the full enumerations of possible peaks. There is a corresponding loss of recall. However, if we take into account the intensity of the measured peaks, by considering the weighted recall scores, we see that our methods perform well on the more important, higher intensity peaks. More than 75% of the total peak intensity in the tripeptide spectra, and approximately 60% of the total peak intensity in the positive ionization mode metabolite spectra, were predicted. For the negative ionization mode spectra, this dropped to 50%. However the precision and jaccard values were comparable with those of the positive ionization mode data.

The results presented in Figure 6.2 show scores averaged across the three energy levels for each molecule. If we consider the results for the energy levels separately, we find that the low and medium energy results are generally much better than those of the high energy. For example, Figure 6.3 presents the prediction results separately for the three energy levels for the Metlin metabolite data. The same trend was also observed for the other data sets (results not shown here).

The poorer high energy spectra results may be due to increased noise and a lower predictability of events at the higher collision energies. Another possible explanation is that the even-electron rule and other assumptions listed in Section 5.1.1 may be less reliable when there is more energy in the system.

In the case of the tripeptide data, our methods achieve higher recall scores and similar rates of precision to that of the heuristic model of known fragmentation mechanisms, resulting in improved Jaccard scores. Since peptide fragmentation mechanisms are fairly well understood, this result is not intended to suggest that our method should be used in place of current peptide fragmentation programs, but rather to demonstrate that SE-CFM and CE-CFM are able to extract fragmentation patterns from data to a similar extent to human experts, given a sufficiently large and consistent data set. Like our methods, the heuristic models also perform better

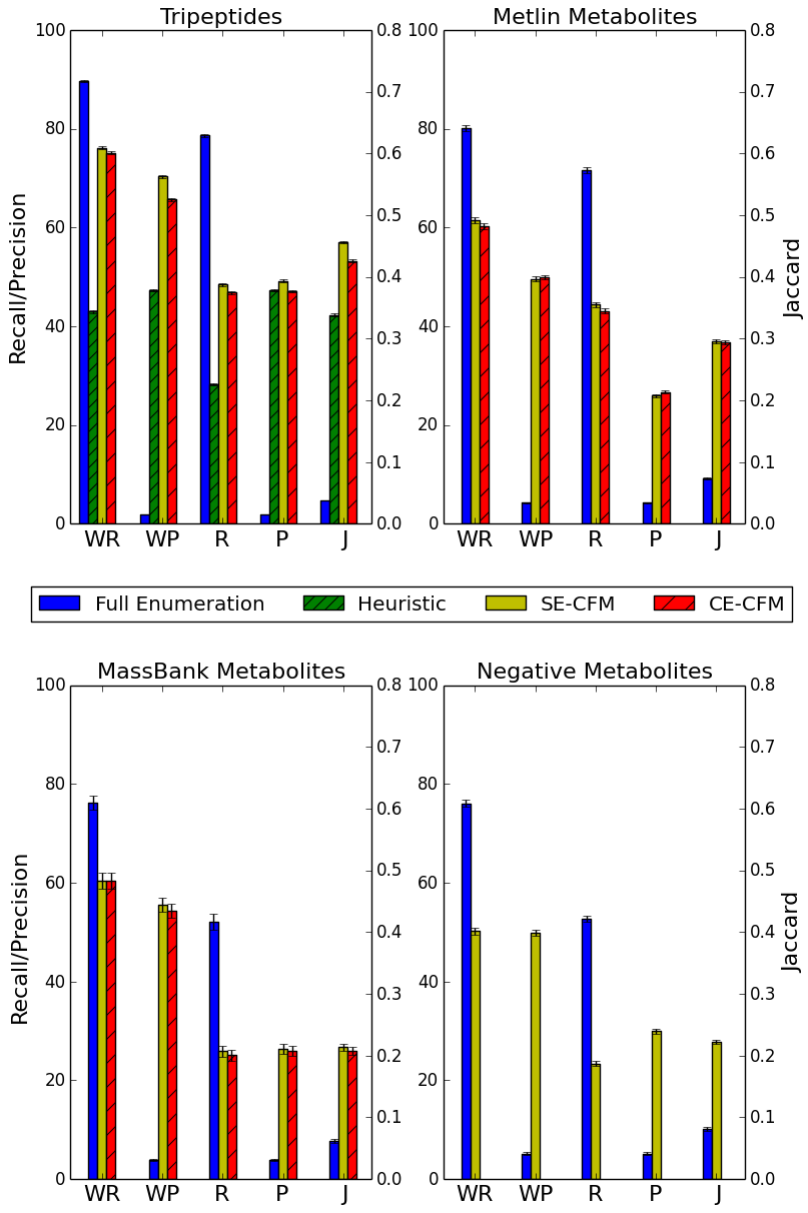


Figure 6.2: Spectrum prediction results for tripeptides (upper left), metabolites from Metlin (upper right), metabolites from MassBank (lower left) and metabolites from Metlin using negative mode ionization (lower right). The x-axes show the five metrics: Weighted Recall (WR), Weighted Precision (WP), Recall (R), Precision (P) and Jaccard (J), averaged across the three energy levels for each test molecule. Bars display mean scores $\pm$ standard error. In each plot, note that the y-axis for Jaccard (on right) is different from the others (on left).

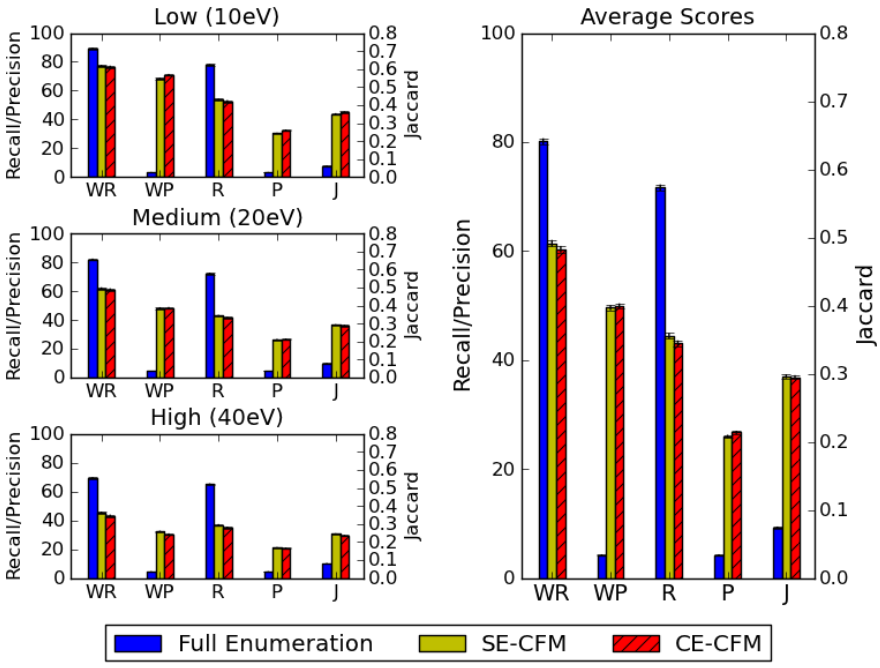


Figure 6.3: Spectrum prediction results for the Metlin metabolites. The x-axes show the five metrics: Weighted Recall (WR), Weighted Precision (WP), Recall (R), Precision (P) and Jaccard (J). The plot on the left shows the metrics measured separately for each collision energy. The right plot shows the results averaged across the three energy levels for each test molecule. Bars display mean scores $\pm$ standard error. In each plot, note that the y-axis for Jaccard (on right) is different from the others (on left).

for the lower energy levels, with a weighted recall score of 66% for the low energy, as compared to only 24% for the high energy.

Unsurprisingly, being a smaller and more diverse data set, the Metlin metabolite results are poorer than those of the tripeptides. However the weighted recall for both our methods is still above 60% and the precision and Jaccard scores are much higher than for the full enumeration, suggesting that the CFM model is still able to capture some of the common fragmentation trends.

The weighted recall and precision results for the MassBank metabolites are fairly comparable to those of the Metlin metabolites. There is a small loss in the non-weighted recall, however this is probably due to a higher incidence of low-level noise in the MassBank data. This results in a small loss in the average Jaccard score. However these results demonstrate that the fragmentation trends learned still apply to a significant degree on data collected at a different time in a different laboratory.

Since this is the first method, to the author's knowledge, capable of predicting intensity values as well as m/z values, we also investigated the accuracy of CFM's predicted intensity values. We found that the Pearson correlation coefficients for matched pairs of predicted and measured peaks, were 0.7, 0.6 and 0.45 for the low, medium and high spectra respectively (SE-CFM and CE-CFM results were not substantially different). This indicates a positive, though imperfect correlation. Full results and scatter plots are contained in the supplementary data for [87].

Running on a 2.2 GHz Intel Core i7 processor, the median run-time for the spectrum predictions for each molecule in the Metlin metabolite data set was 5 seconds. Larger molecules with more ring systems generally take longer as they have so many more fragmentation possibilities in the initial enumeration. For molecules with no rings, the median run-time was 2 seconds, whereas for molecules with 3 or more rings, the median run-time was 9 seconds. The longest run-time in the Metlin metabolite set was for Troleandomycin (Metlin ID 41012), which has a molecular weight over 800 Da and contains three ring systems, one of which is size 14. It took just under 5 minutes.

6.5 METABOLITE IDENTIFICATION

Here we apply our CFM MS/MS spectrum predictions to a metabolite identification task.

6.5.1 *Candidate Selection*

For each molecule, we produce two candidate sets via queries to two public databases of chemical entities:

1. We query the PubChem compound database [47] for all molecules within 5 ppm of the known molecule mass. This simulates the case where little is known about the candidate compound, but the parent ion mass is known with high accuracy.

2. We query KEGG (Kyoto Encyclopedia of Genes and Genomes) [69] for all the molecules within 0.5 Da of the known molecular mass. This simulates the case where the molecule is thought to be a naturally occurring metabolite, but there is more uncertainty in the target mass range.

To conduct this assessment, duplicate candidates were filtered out – i.e. those with the same chemical structure, including those that only differ in their stereochemistry. Charged molecules and ionic compounds were also removed since the program assumes single fragment, neutral candidates (to which it will add a proton). After filtering, the median number of candidates returned from PubChem was 911 for the tripeptides and 1025 for the metabolites. Note that 9 tripeptides and 57 of the Metlin metabolites were excluded from this testing because no matching entry was found in PubChem for these molecules. The KEGG queries were only carried out for the metabolite data. The median number of candidates returned was 22, however no matching entry was found in KEGG for 833 of the Metlin metabolites and 111 of the MassBank metabolites.

6.5.2 *Methods for Comparison*

Whenever a matching entry could be found, we ranked the candidates according to how well their predicted low, medium and high spectra matched the measured spectra of the test molecule. The ranking score we used was the Jaccard score described in Section 6.4.

We compared the ranking performance of our SE-CFM and CE-CFM methods against those of MetFrag [74] and FingerID [59]. We used the same candidate lists for all programs. For candidate molecules with equal scores, we had each program break ties in a uniformly random manner. This was in contrast to the original MetFrag code, which used the most pessimistic ranking; we did not use that approach as it seemed unnecessarily pessimistic. We set the mass tolerances used by MetFrag when matching peaks to the same as those used in our method (maximum of 0.01 Da and 10 ppm for all except the HMDB set, which used 0.5 Da). MetFrag and FingerID only accept one spectrum, so to input the three spectra we first merged them as described by [74]: we took the union of all peaks, and then merge together any peaks within 10 ppm or 0.01 Da of one another (or 0.5 Da for HMDB), retaining the average mass and the maximum intensity of the two. In FingerID we used the linear High Resolution Mass Kernel including both peaks and neutral losses, and trained using the same cross-fold sets as for our own method. Overall, we attempted to assess CFM, MetFrag and FingerID as fairly as possible, using identical constraints, identical databases and near-identical data input.

6.5.3 *Results*

The results are shown in Figure 6.4. As seen in this figure, our CFM method achieved substantially better rankings than both the existing methods on all five data sets, for both the PubChem and KEGG queries. On the Metlin and Massbank metabolite data, when querying against KEGG, our methods were able to find the correct metabolite as the top-scoring candidate in over 70% of cases, and almost always ($> 95\%$) ranked the correct can-

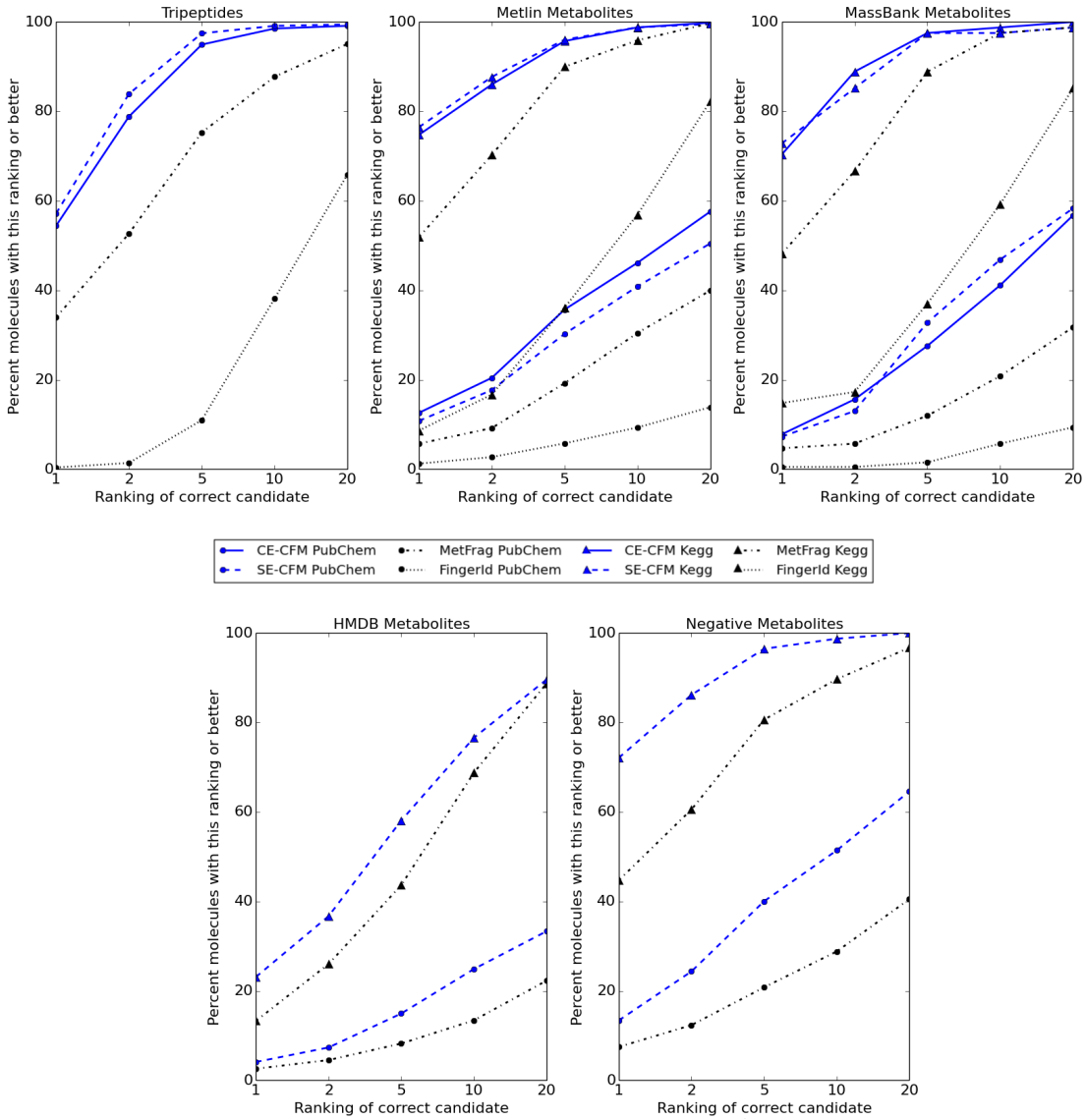


Figure 6.4: Ranking results for metabolite identification, comparing both CFM variants with MetFrag and FingerID for tripeptides (upper left), metabolites from Metlin (upper middle), validation metabolites from MassBank (upper right), HMDB validation metabolites (lower left) and negative metabolites from Metlin (lower right), querying against PubChem within 5 ppm (circles) and KEGG within 0.5 Da (triangles). Note that our methods out-perform both MetFrag and FingerID on all metrics, regardless of the database used.

didate in the top 5. In comparison, MetFrag ranked the correct metabolite first in approximately 50% of cases for both metabolite sets, and in the top 5 in 89%. FingerID ranked the correct metabolite first in less than 15% of cases.

For PubChem, our methods performed well on the tripeptide data, identifying the correct metabolite as the top-scoring candidate in more than 50% of cases and ranking the correct candidate in the top 10 for more than 98% of cases. This is again convincingly better than both MetFrag and FingerId, which rank the correct candidate first in less than 35% and 2% of cases respectively.

For the Metlin metabolite data, when querying PubChem, CE-CFM and SE-CFM were able to identify the correct metabolite in only 12% and 10% of cases respectively for the positive mode data, and SE-CFM identified 13% for the negative mode data. Given that this is from a list of approximately one thousand candidates, this performance is still not bad. Once again, it is substantially better than MetFrag and FingerID, which correctly identified less than 6% and 1% of the positive mode cases respectively. Our methods rank the correct candidate in the top 10 in more than 40% of cases on all data sets except the HMDB set, as compared to MetFrag's performance of 31% on the Metlin metabolites, 21% on the MassBank metabolites and 28% of the negative metabolites. Additionally, the top-ranked compound was found to have the correct molecular formula in more than 89% of cases for SE-CFM and 90% of cases for CE-CFM, suggesting that both methods mainly fail to distinguish between isomers.

In the case of the HMDB set, the performance dropped, ranking the correct structure 1st in only 23.1% of cases when querying KEGG, and in the top 5 in 58.1%. For PubChem, it was only able to rank the compound in the top 10 in 24.5% of cases, however SE-CFM was still able to identify the correct molecular formula in 88.4% of cases; and still outperformed MetFrag, which ranked the correct structure in the top 10 in only 14.5% of cases. This is likely due to the poorer mass accuracy and spectrum quality of the HMDB spectra.

While the performance of all three methods (CFM, MetFrag and FingerID) is not particularly impressive for the PubChem data sets (i.e. $<$12%

correct) we would argue that the PubChem database is generally a poor database choice for anyone wishing to do MS/MS metabolomic studies. With only 1% of its molecules having a biological or natural product origin, one is already dealing with a rather significant challenge of how to eliminate a 100:1 excess of false positives. So we would regard the results from the PubChem assessment as a "worst-case" scenario and the results from the KEGG assessment as a more typical metabolomics scenario.

The results for CE-CFM showed minimal difference when compared to those of SE-CFM, casting doubt on whether the additional complexity of CE-CFM is justified. However we think this idea is still interesting as a means for integrating information across energy levels and may yet prove more useful in future work.

The running time of the metabolite identifications is mainly dependent on the number of candidate molecules and the time taken to predict the spectra for each. For example, taking 1000 candidates (as in the PubChem tests) at the median spectrum prediction run-time of 5 seconds (see Section 6.4), the identification would be expected to take in the order of 1.5 hours. Taking only 22 candidates (as in the KEGG tests), this reduces to 2 minutes. It would be trivial to parallelize the computation by distributing candidates across processors. When repeatedly querying against the same database, it may also be expedient to precompute the predicted spectra to reduce the identification run-time. For example, our web server interface `http://cfmid.wishartlab.com` provides access to precomputed spectra for all 40,000 compounds in HMDB and over 10,000 compounds in KEGG.

7

EMPIRICAL EVALUATION OF EI-MS

This section contains further results for the CFM method on spectrum prediction and metabolite identification tasks, this time using EI-MS.

7.1 DATA

The primary source of EI-MS data is the main library of the NIST/EPA/NIH Mass Spectral Library [44], the 2014 version of which contains EI-MS spectra for 242,466 chemical compounds. This library also contains replicate spectra for 33,782 compounds – i.e. re-measurements of compounds from the main EI-MS library at a different time or in a different laboratory. All data were measured at integer mass accuracy using a single energy of 70eV.

We used three subsets of this data as follows:

1. **Small Molecule Set (17,324 molecules)**
   This set was designed for use within a cross validation framework to compare various CFM model and parameter configurations relatively rapidly. This meant generally selecting molecules that were smaller and therefore required lower computation times.

   This was done by initially selecting 80,000 molecules at random from the main NIST library, then removing those whose fragmentation graph (see Section 5.1.1) could not be computed in less than 10 seconds on our server. The remaining molecules were generally less than 500 Da and comprised no more than 30 heavy atoms (non-Hydrogen atoms). To ensure spectrum quality, I also removed those for which the weighted recall of the full enumeration spectrum was less than 50. Molecules identical (ignoring stereochemistry) to those in the two validation sets below were also removed. There were 17,324 molecules remaining.

The set was randomly divided into 5 groups for use within a 5-fold cross validation framework. The full set was then used to produce a trained model for further validation with the other two data sets listed below.

2. **Kerber Set (100 molecules)**
   This set also contained entries from the main NIST library. It was used for validation of the program MOLGEN-MS in Kerber et al. [70]. Further results were also reported for this set on ACD Fragmenter, MOLGEN-MS, Mass Frontier and MetFrag in Schymanski et al. [99]. We include this set in order to compare against those previously reported results.

3. **Replicate Set (20,588 molecules)**
   This set contained entries from the NIST replicate set. The original set had 33,782 molecules. 296 molecules were removed because they were not computable by CFM-ID, for example because they had too many disconnected components, non-standard valencies, or could not be parsed by RDKit. Another 12,898 were removed because they were duplicates (e.g. stereoisomers) of another molecule in the set.

7.2 MODEL CONFIGURATION

The model was configured the same way we configured it for ESI-MS/MS, using both a fragmentation depth of 2 and a model depth of 2, but including the additional odd-electron fragmentation possibilities suitable for EI-MS (see Section 5.2).

During cross validation, both the original observation function (see Section 5.1.3), and that using the additional extensions for isotopes (see Section5.3) were tested.

We also trialed both the original transition function (see Section 5.1.2), and the extension to include a neural network (see Section 5.5). When used, the neural network was configured to include two hidden layers, one with 20 nodes and the other with 4. In all hidden nodes we used a

rectified linear unit (reLU) activation function, with half the units assigned a negative activation function, as recommended by [100, 101, 102]. The final output node was a linear unit.

The best performing model, as determined during cross validation testing on the small molecule set (see Section 7.4.3 and Section 7.5.4), used both these extensions, so these were included in the final model used for validation with the other two datasets.

7.3 CHEMICAL FEATURES

During cross-validation we trialed two feature sets as follows:

1. **Original Feature Set**
   These features were the same as those used previously for ESI-MS/MS with two additional features specific to EI-MS and one removed feature type, as follows:
   - *Radical Features*: Additional features indicating whether the break resulted in a radical ion, a radical NL, or neither (3 features).
   - *Ionic Features*: Ionic bonding of charged single-atom cations and anions was more common in the EI-MS data, so support was added for breaking these bonds. These features indicate whether a break resulted in: a positive ionic fragment attached to the ion, or to the NL; or a negative ionic fragment attached to the ion, or to the NL; or none of these (5 features).
   - *No Gasteiger Charges*: RDKit [86], the chemistry development package we used, was unable to compute Gasteiger charges for some molecules, so we found it simplest to remove this feature.
2. **New Feature Set**
   This feature set contained further modifications, in addition to those listed above, that were intended to address various short-comings of the original feature set that were not specific to EI-MS. (They have not yet been trialled on ESI-MS/MS data, but may be beneficial in that context).

- *Broken Bond Type*: Additional features indicating the type of broken bond (single, double, triple, aromatic, conjugated, ionic, hydrogen loss) (7 features).
- *Neighbouring Bond Type*: Additional features indicating the type of any bond found neighbouring the broken bond in the ion (8 features: 7 bond types as above, plus indicator for no neighbouring bond), and similarly in the NL (8 features) (16 features).
- *Functional Group Features*: Features indicating whether or not the root atom is part of each of 161 functional groups (plus one for no recognised functional group), and another set of indicators for whether the atoms neighbouring the root atom are in each group. This is applied separately to the ion and NL. The selection of functional groups combined 86 fragment descriptors included in RDKit [86], with 107 functional groups developed by Yannick Djoumbou Feunang during his chemical classification work [103], removing duplicates. (648 features).
- *Ion and NL Root Paths of length 2 only*: We exclude features for paths of length 3, that were included for ESI-MS/MS, since these features are very numerous and we hope that any relevant information contained in these features is better captured by the Functional Group Features.
- *Extra Ring Features*: Indicator that a ring is broken during the fragmentation. Indicators that the ion root atom and NL root atom, respectively, remains in a ring after the fragmentation. (3 features).

When the linear transition function was used, the quadratic features were included. When the neural net was used, the quadratic features were not included. Table 7.1 shows the total number of features, and the number actually used for each configuration – i.e. the number encountered in the training set. We can see that removing the quadratic features substantially reduces the number of features in the neural net models. And while there are half as many features in the new feature set as compared to the

FEATURE SET - MODEL	# TOTAL	# USED
Original - Linear	2,252,504	75,656
Original - Neural Net	42,549	27,249
New - Linear	490,546	160,787
New - Neural Net	19,909	18,509

Table 7.1: Number of features for each feature set and model configuration.

original set (not counting quadratic features), a much higher proportion of them (93% vs 64%) are encountered in the training set.

7.4 SPECTRUM PREDICTION

For each cross validation fold, and validation set, a trained model was used to predict a spectrum for each molecule in the test set. As in ESI-MS/MS, the model was run forward, and the resulting marginal distributions for the peak variables were a mixture of Gaussian distributions. We take the means and weights of these Gaussians as our peak mass and intensity values. Since the NIST data has integer mass tolerance, we collect the peaks into integer bins by rounding masses to the nearest integer, and summing the intensity values of peaks within the same bin.

The EI-MS data contains many more peaks per spectrum than the ESI-MS/MS data we used in Chapter 6; the median number of peaks in a spectrum is 94, as compared to between 5 (low energy) and 12 (high energy) in the Metlin ESI-MS/MS data. The integer mass tolerance in the NIST EI-MS data also means that the number of possible peak locations is far lower – e.g. 400 for a 400 Da molecule. Consequently, the post-processing we applied for ESI-MS/MS is not applicable for EI-MS. So rather than discarding unlikely peaks, we instead keep all possible peaks, but rely on the differences in predicted intensity values to differentiate between molecules.

7.4.1 *Metrics*

We consider the same metrics we used for ESI-MS/MS (see Section 6.4.1), with the following two additions:

1. **Stein Dot Product**: The weighted dot product metric recommended by Stein and Scott [40] for searching against the NIST database.

$$\frac{\sum_{(m_P,h_P,m_M,h_M)\in s_P\cap s_M} (m_P m_M)^a (h_P h_M)^b}{\sum_{(m_P,h_P)\in s_P} m_P^a h_P^b \sum_{(m_M,h_M)\in s_M} m_M^a h_M^b}$$

   where $a = 3$ and $b = 0.6$, the predicted MS spectrum is denoted s_P and the measured MS spectrum s_M. This measure takes into account the intensities of both the measured and predicted spectra, producing a higher score when a peak is present in both spectra with high intensity. The intersection operation $s_P \cap s_M$ collects pairs of matching peaks from the two spectra –i.e. those that are within a specified mass tolerance of one another.

2. **Dot Product**: A re-weighted version of Stein's Dot Product that uses $a = 0.5$ and $b = 0.5$. While Stein's weightings are good when the candidate molecules cover a wide range of molecular masses, in the case where the candidate molecules are all of similar mass, those weights over-emphasize the higher peaks, often at the expense of information contained in the lower peaks. We propose this metric to address this common situation.

7.4.2 *Models for Comparison*

As in ESI-MS/MS, there are no existing computational methods for comparison when it comes to Spectrum Prediction, so we include the following models in our comparisons:

- **Full Enumeration (Enum)**: This model considers the predicted spectrum to be one that enumerates all possible fragments in the molecule's fragmentation tree with uniform intensity values.
- **Full Enumeration with Isotopes (Enum-Iso)**: The inclusion of isotopes (see Section 5.3) increases the number of peaks in the full enumeration. So this model uses a full enumeration with the additional isotope peaks included.
- **Measured**: (Replicate Set only) This model uses the measured spectrum from the main NIST library for the corresponding molecule in place of the predicted spectrum. Re-measurement variability means that these spectra will generally not be a perfect match for the target spectrum.
- **CFM Models {NN, Lin} x {Orig, New} x {Iso, -}**: We consider various configurations of the CFM model. These are all combinations of the following: with (Iso) vs without isotopes; with (NN) vs without (Lin) the neural network extensions; and using the orignal (Orig) vs new (New) feature set.

7.4.3 *Results*

The results of cross validation testing on the small molecule set are presented in Figure 7.1.

It can be seen from the weighted recall values that almost all the peaks in each spectrum can be explained by a fragmentation event. Note that the lack of differentiation between the different models seen in the weighted recall and the Jaccard scores is because these metrics are independent of the predicted intensity values – so without post-processing to remove low intensity peaks, there is no difference between CFM and the full enumeration.

It is interesting to note that the precision and Jaccard scores for the full enumeration spectra are quite high, as compared to what we saw for ESI-MS/MS. The larger number of peaks in each experimental spectrum, com-

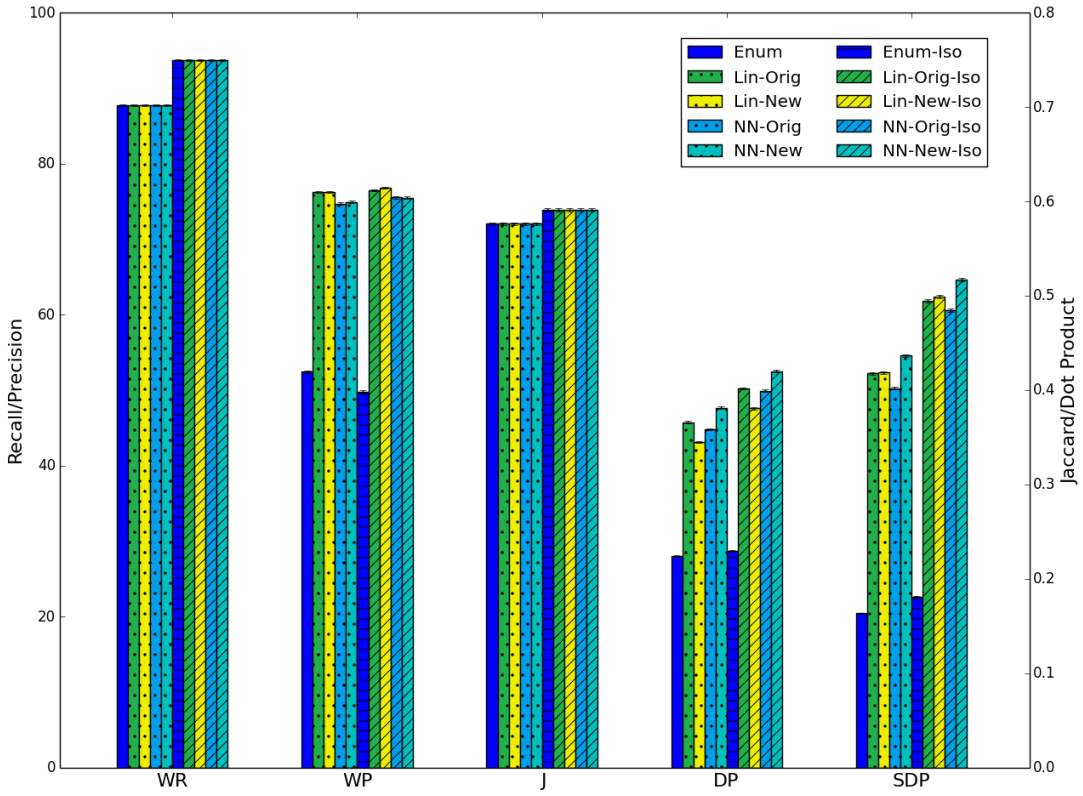


Figure 7.1: Spectrum prediction results for Small Molecule Set. The x-axis shows the five metrics: Weighted Recall (WR), Weighted Precision (WP), Jaccard (J), Dot Product (DP) and Stein Dot Product (SDP). Bars display mean scores $\pm$ standard error. Note that the y-axis for Jaccard and Dot Product (on right) is different from that for Recall and Precision (on left).

bined with the integer mass tolerance, result in a high prior probability that a peak will be found at any given mass location. This makes it far more likely that each predicted peak will find a match by chance in the target spectrum. This makes the predicted intensity values far more important for EI-MS than they were for ESI-MS/MS. The other three scores show that when taking the predicted intensities into account, CFM significantly outperforms the full enumeration models.

Since the dot product scores take into account the intensities of both the measured and predicted spectra, they are a good metric for how well each model predicts the spectrum. Using either dot product metric, we see that including the isotope extensions improves model performance, and that the best performing model uses the neural network extensions combined with the new feature set. This is why we selected the NN-New-Iso model for further validation testing.

Results for the spectrum prediction tests on the replicate set are presented in Figure 7.2. Here we compare the spectrum prediction performance of the best performing CFM model (NN-New-Iso) with the full enumeration spectrum, as well as with the re-measured spectra.

We see that the scores for the full enumeration and CFM are consistent with those seen during cross validation on the small molecule set. CFM again substantially outperforms the full enumeration spectra, demonstrating that it is able to differentiate between likely and unlikely fragmentations.

However the comparison with the measured spectra shows that CFM still falls short of providing a spectrum that is as reliable as one produced by physically measuring the spectrum. This is not unexpected, and shows that computational methods still have plenty of room for improvement.

7.5 METABOLITE IDENTIFICATION

Here we apply our MS spectrum predictions to a metabolite identification task.

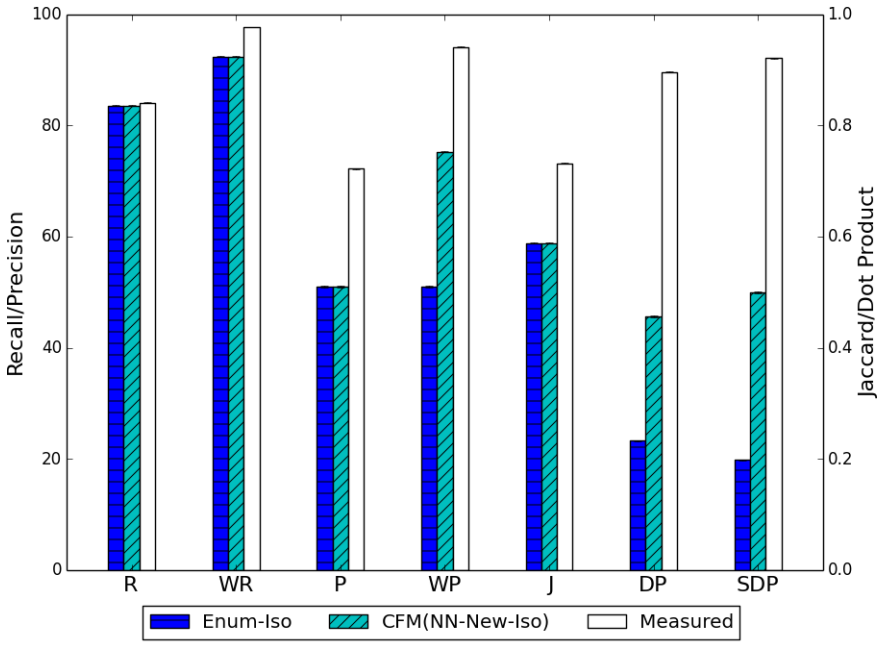


Figure 7.2: Spectrum prediction results for Replicate Set. The x-axis shows the metrics: Recall (R), Weighted Recall (WR), Precision (P) Weighted Precision (WP), Jaccard (J), Dot Product (DP), Stein Dot Product (SDP). Bars display mean scores. Error bars are too small to be seen. Note that the y-axis for Jaccard and Dot Product (on right) is different from that for Recall and Precision (on left).

7.5.1 *Candidate Selection*

For each test molecule, candidate sets were produced using the following methods:

1. **PubChem**: We query the PubChem compound database [47] for all molecules within 10 ppm of the known molecule mass. This simulates the case where little is known about the candidate compound, but the parent ion mass is known with high accuracy. After filtering to remove duplicates, the median number of candidates returned for the replicate set was 1136, and for the small molecule set was 1089. After retaining only those candidate compounds that could be processed by MetFrag and CFM-ID[1], the median number of candidates reduced to 896 and 1015, respectively. With further filtering to retain only those compounds with the correct molecular formula, the median number of candidates for the replicate set was 405.

2. **HMDB**: We query HMDB (Human Metabalome Database) [16] for all molecules within 0.5 Da of the known molecular mass. This simulates the case where the molecule is thought to be a naturally occurring metabolite, but there is more uncertainty in the target mass range. This is very similar to the tests we did on KEGG for ESI-MS/MS, since HMDB contains all the molecules in KEGG. The median number of candidates returned for the Replicate Set was 53 (as compared to 22 from KEGG during ESI-MS/MS testing).

3. **MOLGEN**: For the Kerber Set, to compare with previously published results in Kerber et al. [70], we used candidate sets of all possible isomers for each molecule as generated by MOLGEN and made available in the supplementary information of Schymanski et al. [71]. As discussed in Section 4.2.2, using all structural isomers like this is a very extreme test case, and since the number of structural isomers grows exponentially with molecule size, it is only possible for test

1 These were connected compounds with parsable SMILES inputs, standard valences, and whose fragmentation graphs could be computed by CFM within 10 minutes

molecules such as these with low molecular masses. The median number of candidates for this set was 802.

4. **NIST**: For comparison with the case where you have a reference database of measured spectra (rather than computationally predicted spectra), we used the entire main library of the NIST EI-MS database as a candidate set. After removing those candidates that were uncomputable by CFM-ID, for example because they had too many disconnected components, non-standard valencies, or could not be parsed by RDKit, this left 236,693 candidates. With filtering to retain only those compounds with the correct molecular formula, the median number of candidates reduced to 17.

7.5.2 *Methods for Comparison*

In cross-validation testing on the small molecule set, we compared the ranking performance of two CFM models (NN-New-Iso and Lin-New-Iso) when querying PubChem, to see whether better prediction performance translated to better identification performance. Unfortunately, each of these tests required in the order of 10 core-years of computation, since a median of 1015 candidate spectra needed to be predicted for each of 17324 test molecules, so time and compute constraints prevented testing of the other six CFM models in this manner. We also assessed the differences in identification performance obtained by using each of the metrics used to assess spectrum prediction performance (see Section 7.4.1) to rank candidates.

On the validation data, we then compared the ranking performance of the best performing CFM model (NN-New-Iso) against that of MetFrag [74], and where possible, MOLGEN-MS [70] and MassFrontier (using the results reported in [99]).

MetFrag was run using the recent update MetFrag2.2 CL available at `http://c-ruttkies.github.io/MetFrag/projects/metfrag22cl/`, with PrecursorIonMode set to 2 (for [M+]), and using FragmenterScore only (i.e. no use of patent or reference counts). CFM used a Dot Product metric to rank

candidates. Both programs used an absolute mass tolerance of 0.5 Da. We had no control over these settings for MOLGEN-MS. We used the same candidate lists for all programs tested.

We also compared CFM's performance to that achievable when measured spectra are available for all candidate compounds. We did this by querying the replicate set against the full NIST set, using actual measured spectra and CFM-predicted spectra. For the measured spectra, we used Stein's Dot Product to compare spectra, and thus rank candidates, as recommended by [40]. For CFM, we report results using both Stein's Dot Product and our own Dot Product for this purpose, since the former might be expected to do well in this kind of test, in which the candidate molecules have a wide range of masses, but the latter performed better with CFM in our cross validation tests.

7.5.3 *Metrics*

Where possible, we used the following metrics to assess ranking performance:

- **Absolute Ranking**: The percentage of molecules achieving various threshold rankings (1, 2, 5, 10, 20, 100), as we used for ESI-MS/MS. For candidates with equal scores, ties were broken by taking the expected ranking given a uniform distribution over tied candidates.

- **Relative Ranking Performance (RRP)**: This metric was used in [70] and [71], and is defined as:

$$\mathrm{RRP} = \frac{1}{2}\left(1 + \frac{\mathrm{BC} - \mathrm{WC}}{\mathrm{TC} - 1}\right)$$

  where BC denotes the number of candidates with better scores, WC denotes the number of candidates with worse scores, and TC denotes the total number of candidates. This metric takes into account the total number of candidates, assessing the relative ranking of the correct candidate within the full candidate set. A value of 0.0 indicates a per-

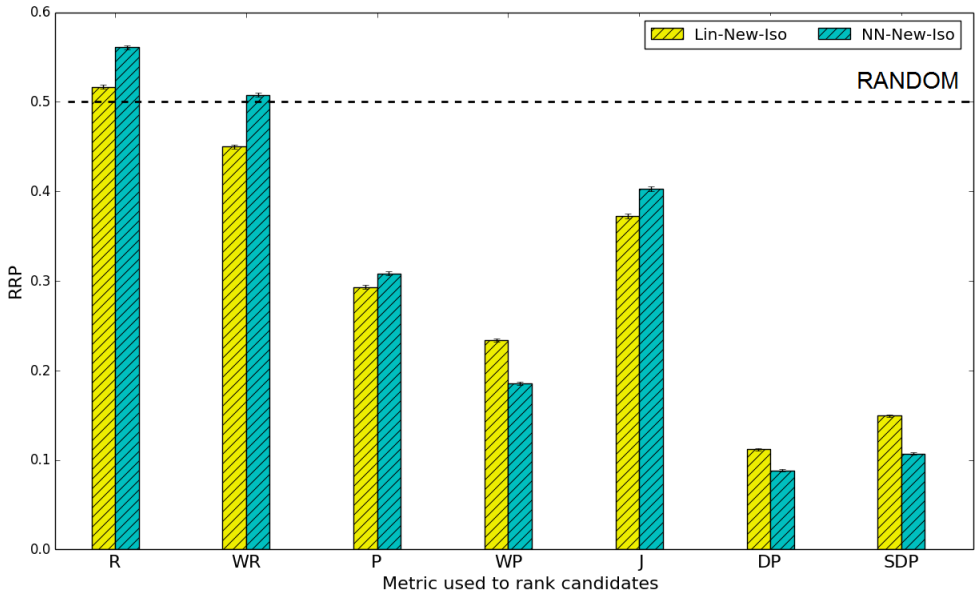


Figure 7.3: CFM (NN-New-Iso and Lin-New-Iso) metabolite identification performance on small molecule set when querying PubChem (median number of candidates = 1015). The x-axis shows the metrics used to rank candidates: Recall (R), Weighted Recall (WR), Precision (P) Weighted Precision (WP), Jaccard (J), Dot Product (DP), Stein Dot Product (SDP). Bars display mean relative ranking performance (RRP) scores. Error bars are too small to be seen. Note than an RRP of 0.0 is perfect, and an RRP of 0.5 is no better than random.

fect identification, whereas a value of 0.5 indicates that performance is no better than random.

7.5.4 *Results*

The results of cross-validation testing, in which the small molecule set was tested using candidates from PubChem, are shown in Figure 7.3. When ranking candidates using the recall or weighted recall scores, we see that the performance is no better than random. This is equivalent to using a full enumeration spectrum for matching. Ranking using the weighted recall is also equivalent to using a match value based scoring, the approach taken

DATA SET	QUERY	MFRONT	MOLGEN-MS	METFRAG	CFM
Kerber	MOLGEN	0.268	0.273	0.354	**0.195**
Replicate	HMDB	-	-	0.314	**0.096**
Replicate	PubChem	-	-	0.333	**0.099**

Table 7.2: Average Relative Ranking Performance (RRP) of MassFrontier (MFront), MOLGEN-MS, MetFrag and CFM (NN-New-Iso) under three experimental conditions. Results for MassFrontier and MOLGEN-MS were taken from [99]. Best results in each condition are indicated in bold.

in [70] that [71] showed was not effective. The only difference here is the details of the full enumeration.

The best result (RRP = 0.0880) was achieved when ranking candidates using a Dot Product metric, demonstrating that our predicted intensity values help rank candidates correctly. The performance using the NN-New-Iso model was better than that obtained using the Lin-New-Iso model, showing that in this case at least, better prediction performance translated to better identification performance.

The RRP results for validation testing with the Kerber and replicate sets are presented in Table 7.2. The dot product metric was used to rank candidates for CFM. Standard error values were all less than 0.01 for tests on the Kerber data set, and less than 0.001 for tests on the other data sets.

On the Kerber Set, CFM outperforms MassFrontier, MOLGEN-MS and MetFrag. The RPP score achieved is 0.195, which is still quite poor; nearly 20% of candidates score better than the correct candidate. However one should note that this is a very extreme test case, in which the comparison is between a large number of very similar molecules, and this result is substantially better than any previously reported on this set [70, 71, 99].

The performance on the replicate set when querying HMDB and PubChem is better, and once again CFM substantially outperforms MetFrag. It is interesting that both programs achieved RRP scores when querying HMDB that are very similar to those achieved when querying PubChem. This suggests that the characteristics of a molecule that make it more likely

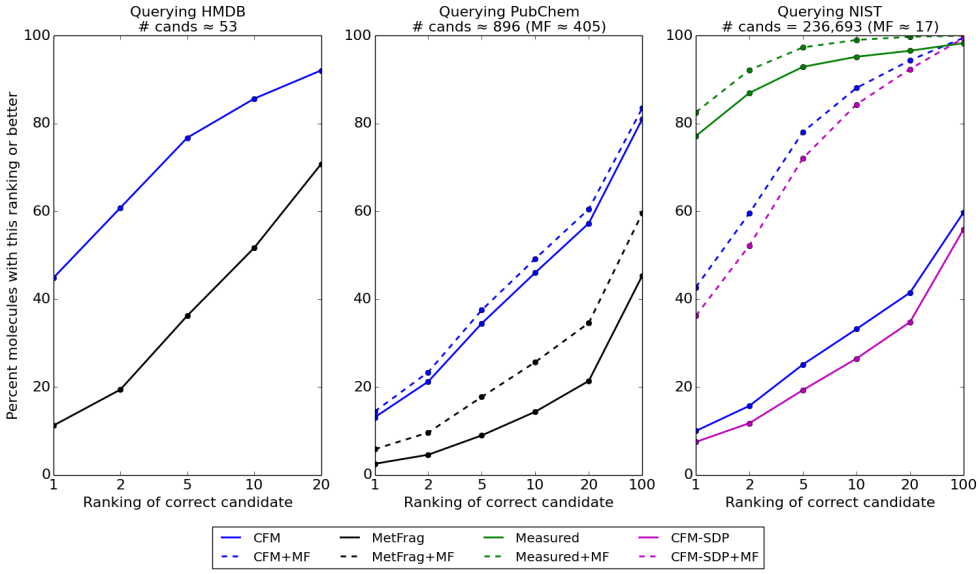


Figure 7.4: Absolute ranking results obtained using the replicate set, querying HMDB (left), PubChem (middle) and NIST (right) for candidate molecules. Solid lines indicate rankings achieved using the full set of candidates. Dashed lines indicate rankings achieved when narrowing the set of candidates to include only those with the correct molecular formula. CFM-SDP (in magenta), indicates that CFM was run using Stein's Dot Product metric to compare spectra. All other CFM results (in blue) use our modified Dot Product metric.
cands ≈ N: The median number of candidates is N.
MF ≈ N: The median number of candidates with the correct MF is N.

to be found in HMDB are independent of those characteristics that make it identifiable from its mass spectrum.

Absolute ranking results for these same tests are shown in the left two axes of Figure 7.4. The lower number of candidates retrieved from HMDB for each molecule means that the similar RRPs translate to much better ranking performance than for PubChem. When querying HMDB, the target molecule was correctly identified in 45% of cases, and ranked in the top 10 in 86% of cases. When querying PubChem, the target molecule was correctly identified in 13% of cases, and ranked in the top 10 in 46% of cases. We reiterate that, while the the PubChem database is an interesting test

case for our algorithms, it is generally a poor database choice for anyone wishing to do EI-MS metabolomic studies. With only 1% of its molecules having a biological or natural product origin, one is already dealing with a rather significant challenge of how to eliminate a 100:1 excess of false positives. So we would regard the results from the PubChem assessment as a "worst-case" scenario and the results from the HMDB assessment as a more typical metabolomics scenario.

It is also interesting that the restriction of candidates to include only those from PubChem with the correct molecular formula had little effect on the absolute rankings obtained with CFM (as shown by the dotted lines in Figure 7.4). This suggests that CFM was already correctly discounting those compounds with incorrect molecular formulae.

The rightmost axis of Figure 7.4 shows the results obtained when querying the replicate spectra against the full NIST database. In this case we were able to compare CFM against identification results obtained when actual, measured reference spectra are available. When each replicate spectra was queried against the full NIST database using the database spectrum for each candidate compound, the correct candidate was retrieved at Rank 1 in 77% of cases. This is consistent with the results reported in [40], and suggests that the combined effects of measurement variability, spectrum quality and the information content in mass spectra (or insufficiency thereof), mean that even actual measured spectra do not allow for perfect identification performance.

Unfortunately, CFM was only able to retrieve the correct candidate at Rank 1 in 9.97% of cases. Given that there are more than 200,000 candidates, this result is not really all that bad. When restricted to consider compounds with the correct molecular formula (a more realistic search scenario), the rate of correct identifications increases to 42.6%. Our Dot Product metric outperforms Stein's Dot Product metric, when used with CFM for ranking candidates, even though this test case consisted of a wide range of different candidate molecular masses.

However this result does demonstrate that a gap still remains between identification performance obtainable when using computationally predicted spectra vs using real measured spectra. This confirms the view of

Sumner et al. [38], that metabolite identifications should ultimately be confirmed using comparisons with real measurements of reference standards.

Despite this apparent short-coming, real measurements are expensive, time-consuming and often infeasible, whereas computational methods offer a rapid, cost-effective alternative. It may be expected that computational methods will continue to be used as they are now; to narrow the chemical search space and hence reduce the experimental work load. Since CFM outperforms all other existing computational methods, it may be considered to be a significant contribution in this area.

8

FUTURE WORK

There are a number of avenues by which the performance of these methods may be further improved, either in terms of accuracy levels or run-times. One of the simplest methods for improving accuracy levels may be to re-train CFM on larger, more diverse data sets, particularly in the case of ESI-MS/MS, which is currently only trained on a little over 1200 molecules. As more molecules are added to the training set, it may also be beneficial to add more chemical features (for example, those in the new feature set used with EI-MS) to expand the representation power of the model. Towards that end, applying the neural network extensions (so far only applied to EI-MS) and using larger or more carefully designed neural network structures may also result in improvements.

For EI-MS, the number of molecules available for training is already more substantial than for ESI-MS/MS. We found little improvement when training on data sets larger than the ones reported here. However all spectra in the NIST database were measured at only integer mass accuracy, which is insufficient to differentitate between many molecular formulae that can be differentiated with more accurate masses. So if training could be carried out on data collected with more accurate masses, this may allow the system to better disambiguate alternative explanations for each peak, and so better capture which events are most likely. Early results (not reported here) using integer mass ESI-MS/MS data from HMDB for training certainly showed poorer performance compared to training with the higher accuracy Metlin ESI-MS/MS data used in Chapter 6.

Targeting training sets and features to particular chemical classes may also offer benefits. For example CFM-ID often struggles to even enumerate the fragmentation possibilities for lipids because there are so many ways to fragment these molecules (as they are often quite large and include very long chains or multiple ring systems). Reducing each molecule to key

substructures rather than atoms and learning break tendencies for bonds connecting these larger substructures (rather than all possible bonds) may offer improvements. Further, creating lipid-specific chemical features and training on a data set containing only lipids ought to result in a more lipid-targeted model.

Further investigation is also warranted to look more closely at the circumstances under which CFM-ID performs poorly. Dührkop et al. [63] produced a Venn diagram showing that the identification performance of CFM-ID, CSI:FingerID and MAGMa was often quite different on different molecules. It would be interesting to determine whether these differences are systematic and important, or merely random. If the former, there may be simple solutions to the identified short-comings.

For example, one (unverified) possibility is that CSI:FingerID outperforms CFM-ID on molecules in which chemical rearrangements occur during fragmentation. CFM-ID does not allow these fragmentation events in its initial enumeration of fragments, and so will never consider them as a possibility. Consequently, peaks resulting from these fragmentation events may not be explained at all, or may be attributed to erroneous causes. By contrast, CSI:FingerID does not rely on being able to explain any peak (except with a molecular formula annotation), but rather just observes correlations between peaks and substructures. Extending the CFM enumeration of fragments to include rules for such rearrangement events may improve performance.

In a similar vain, MAGMa uses a fragment enumeration method (different to CFM-ID's) that effectively allows for deeper fragmentation events – i.e. more breaks – with better efficiency. Consequently it may perform better on molecules in which such deeper fragmentation events occur – e.g. breaking three or more side groups off a central structure. MetFrag has recently switched to this method for efficiency reasons. Adapting CFM to this alternative fragmentation style would require some reformulation of the equations, but may result in improvements.

Lastly, there are almost certainly efficiency gains to be made when CFM is used to predict spectra for large numbers of very similar molecules (a common use-case). If two molecules have structures that are very simi-

lar, the calculations performed to produce their predicted spectra will also have a high degree of overlap. However, we currently predict all spectra completely independently of one another. Exploiting the structural similarities to more efficiently predict the spectra of similar molecules would be a challenging problem, but one for which a solution could dramatically improve the efficiency of CFM-ID. Such a development may make it feasible to consider much larger candidate sets, for example all structural isomers, without explicitly computing all the predicted spectra.

9

CONCLUSION

This work proposed Competitive Fragmentation Modeling (CFM), a probabilistic, generative model of the fragmentation events occurring within a mass spectrometer, and a method for training parameters of the model from data. The method is capable of predicting both ESI-MS/MS and EI-MS spectra that, while imperfect, show substantial improvements over the so-called 'bar code' spectra commonly used for metabolite identification purposes.

The empirical results in Chapter 6 examined the performance of CFM on multiple ESI-MS/MS data sets, encompassing thousands of molecules, covering data from both QqQ and qTOF instrument types, and employing both positive and negative mode ionization. CFM outperformed existing programs FingerID and MetFrag in a metabolite identification task, producing substantially better rankings for the correct candidate than those programs at the time of testing, when querying for candidate molecules in both PubChem and KEGG.

Very recently, the method CSI:FingerID was reported to achieve better performance than CFM on a different ESI-MS/MS identification task. Those results are quite impressive. However further exploration is still required to ensure that the comparison is fair and accurate – no source code is yet available for CSI:FingerID for independent evaluation. Better performance may also be achieved for CFM on that task by making minor changes to the experimental setup. For example, using all three energy levels predicted by CFM (rather than only the medium energy level as used in the CSI:FingerID paper), increases CFM's accuracy from 12.1% to 14.7%. Further improvements may come from training CFM on the same data used to train CSI:FingerID (which was not done for the CSI:FingerID paper).

Further experiments reported in Chapter 7 examined the performance of CFM on EI-MS data. The method was extended for use in this context by adding handling of radicals and isotopes.

Tests were carried out on a previously published metabolite identification task [70], in which all structural isomers were ranked, for each of 100 molecules with measured mass spectra, based on their ability to predict or explain the measured spectrum. CFM achieved better rankings in this task than existing methods Mass Frontier, MOLGEN-MS and MetFrag. CSI:FingerID is not applicable to EI-MS, so was not included in these tests.

Further validation also examined identification performance of CFM and MetFrag on a much larger set of over 20,000 molecules from the replicate set of the NIST/NIH/EPA 2014 MS database [44], when using candidate molecules from HMDB and PubChem. CFM substantially outperformed MetFrag in all tests.

Finally, predicted spectra were produced by CFM for all structures in the main EI-MS library of the NIST database. We compared the identification performance obtainable using these spectra with that obtainable using the actual measured spectra from that database, again testing on over 20,000 molecules from the replicate set, and ranking candidates purely based on spectrum comparisons. In this case, CFM performance was poorer than that obtained using actual measured spectra. However, performing real physical measurements is often costly or infeasible, so comptutational methods play an important role in metabolomics pipelines. Since CFM outperforms other computational methods, it is an important contribution in this area, and should help to reduce the time and cost of metabolite identifications.

BIBLIOGRAPHY

[1] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill Book Company, 1980.

[2] S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. Spectral Learning of Latent-Variable PCFGs. *Journal of Machine Learning Research*, 15:1–48, 2014.

[3] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[4] M. Collins. Three Generative, Lexicalised Models for Statistical Parsing. In *35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[5] S. Carberry. Techniques for plan recognition. *User Modeling and User-Adapted Interaction*, 11(1-2):31–48, 2001.

[6] C. W. Geib and M. Steedman. On Natural Language Processing and Plan Recognition. *International Joint Conference on Artificial Intelligence*, pages 1612–1617, 2007.

[7] J. Lafferty, A. Mccallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289, 2001.

[8] L. Deng. Connecting Deep Learning Features to Log-Linear Models. In A. Aravkin, L. Deng, G. Heigold, T. Jebara, D. Kanevski, and S. J. Wright, editors, *Log-Linear Models , Extensions and Applications*. MIT Press, Cambridge, Massachussetts, 2015.

[9] G. Heigold, H. Ney, P. Lehnen, T. Gass, and R. Schluter. Equivalence of Generative and Log-Linear Models. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1138–1148, 2011.

[10] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. *Proceedings of Interspeech*, 2:1117–1120, 2005.

[11] A. Mohamed, G. Dahl, and G. Hinton. Deep Belief Networks for Phone Recognition. *Scholarpedia*, 4(5):1–9, 2009.

[12] M. A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[13] F. Allen, A. Pon, M. Wilson, R. Greiner, and D. Wishart. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research*, 42(W1): W94–99, 2014.

[14] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–71, 2002.

[15] D. S. Wishart. Current progress in computational metabolomics. *Briefings in bioinformatics*, 8(5):279–93, 2007.

[16] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. HMDB 3 . 0 - The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41:D801–D807, 2013.

[17] J. B. German, B. D. Hammock, and S. M. Watkins. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, 1(1):3–9, 2005.

[18] D. Wishart. Computational strategies for metabolite identification in metabolomics. *Bioanalysis*, 1(9), 2009.

[19] E. L. Schymanski, C. Meinert, M. Meringer, and W. Brack. The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis. *Analytica Chimica Acta*, 615 (2):136–147, 2008.

[20] J. Gross. *Mass Spectrometry*. Springer, 2nd edition, 2011.

[21] T. Kind and O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8:105, 2007.

[22] F. W. McLafferty and F. Turecek. *Interpretation of Mass Spectra*. University Science Books, 4th edition, 1993.

[23] J. Gross. *Mass Spectrometry*. Springer, 1st edition, 2004.

[24] C. Dass. *Fundamentals of Contemporary Mass Spectrometry*. Wiley, 2007.

[25] E. de Hoffman and V. Stroobant. *Mass spectrometry: principles and applications*. Wiley, 3rd edition, 2007.

[26] D. C. Harris. *Quantitative Chemical Analysis*. W. H. Freeman and Company, 7th edition, 2007.

[27] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. a. Cruz, E. Lim, C. a. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythe. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37:D603–10, 2009.

[28] J. J. Thomson. *Rays of positive electricity, and their application to chemical analyses*. Longmans, Green and Co., 1913.

[29] A. Makarov. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72(6): 1156–1162, 2000.

[30] C. a. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–87, 2006.

[31] J.-L. Wolfender, G. Marti, A. Thomas, and S. Bertrand. Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*, 1382:136–164, 2015.

[32] W. B. Dunn and D. I. Ellis. Metabolomics: Current analytical platforms and methodologies. *Trends in Analytical Chemistry*, 24(4):285–294, 2005.

[33] F. W. McLafferty. Mass Spectrometric Analysis. Molecular Rearrangements. *Analytical Chemistry*, 31(1):82–87, 1959.

[34] K. Biemann. The application of Mass Spectrometry in Organic Chemistry: Determination of the Structure of Natural Products. *Angewandte Chemie International Edition in English*, 1(2):98–111, 1962.

[35] J. R. de Laeter, J. K. Böhlke, P. De Bièvre, H. Hidaka, H. S. Peiser, K. J. R. Rosman, and P. D. P. Taylor. Atomic weights of the elements. Review 2000 (IUPAC Technical Report). *Pure and Applied Chemistry*, 75(6):683–799, 2003.

[36] W. Acree Jr. and J. Chickos. Mass Spectra. In P. Linstrom and W. Mallard, editors, *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards Technology, 2015.

[37] K. Scheubert, F. Hufsky, and S. Böcker. Computational mass spectrometry for small molecules. *Journal of Cheminformatics*, 5(1):12, 2013.

[38] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and

M. R. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3:211–221, 2007.

[39] S. Stein. Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical Chemistry*, 84(17):7274–7282, 2012.

[40] S. E. Stein and D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994.

[41] R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti, and G. Siuzdak. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nature Biotechnology*, 30(9):826–8, 2012.

[42] R. Mylonas, Y. Mauron, A. Masselot, P.-A. Binz, N. Budin, M. Fathi, V. Viette, D. F. Hochstrasser, and F. Lisacek. X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. *Analytical Chemistry*, 81(18):7604–10, 2009.

[43] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. *Journal of Mass Spectrometry*, 44(4):494–502, 2009.

[44] S. Stein. NIST/NIH/EPA Mass Spectral Library. In *Standard Reference Library 1*. National Institute of Standards and Technology, Gaithersburg, MD, USA, 2014.

[45] C. a. Smith, G. O'Maille, E. J. Want, C. Qin, S. a. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27 (6):747–51, 2005.

[46] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda,

N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–14, 2010.

[47] E. Bolton, Y. Wang, P. Thiessen, and S. Bryant. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Chapeter 12 in Annual Reports in Computational Chemistry*, volume 4. American Chemical Society, Washington DC, 2008.

[48] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41:D456–63, 2013.

[49] F. Hufsky, K. Scheubert, and S. Böcker. Computational mass spectrometry for small-molecule fragmentation. *Trends in Analytical Chemistry*, 53:41–48, 2014.

[50] F. Hufsky, K. Scheubert, and S. Böcker. New kids on the block: novel informatics methods for natural product discovery. *Natural Product Reports*, 31(6):807–817, 2014.

[51] K. Varmuza and W. Werther. Mass Spectral Classifiers for Supporting Systematic Structure Elucidation. *Journal of Chemical Information and Modeling*, 36(2):323–333, 1996.

[52] B. Curry and D. E. Rumelhart. MSnet: A Neural Network which Classifies Mass Spectra. *Tetrahedron Computer Methodology*, 3(3-4):213–237, 1990.

[53] C. Klawun and C. L. Wilkins. Joint Neural Network Interpretation of Infrared and Mass Spectra. *Journal of Chemical Information and Computer Sciences*, 36(2):249–257, 1996.

[54] a. Eghbaldar, T. Forrest, and D. Cabrol-Bass. Development of neural networks for identification of structural features from mass spectral data. *Analytica Chimica Acta*, 359(3):283–301, 1998.

[55] K.-S. Kwok, R. Venkataraghavan, and F. W. McLafferty. Computer-aided interpretation of mass spectra. {III}. {Self-training} interpretive and retrieval system. *Journal of the American Chemical Society*, 95(13): 4185–4194, 1973.

[56] S. R. Lowry, T. L. Isenhour, J. B. Justice, F. W. McLafferty, H. E. Dayringer, and R. Venkataraghavan. Comparison of various K-nearest neighbor voting schemes with the self-training interpretive and retrieval system for identifying molecular substructures from mass spectral data. *Analytical Chemistry*, 49(12):1720–1722, 1977.

[57] S. E. Stein. Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry*, 6(8):644–655, 1995.

[58] K. Varmuza, P. Penchev, F. Stancl, and W. Werther. Systematic structure elucidation of organic compounds by mass spectra classification. *Journal of Molecular Structure*, 408-409(96):91–96, 1997.

[59] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18):2333–41, 2012.

[60] A. M. Johnson and G. M. Maggiora. *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, New York, 1990.

[61] H. Shen, D. Kai, B. Sebastian, and J. Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(ISMB):i157–i164, 2014.

[62] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24(16):i49–i55, 2008.

[63] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41): 12580–12585, 2015.

[64] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, 7:234, 2006.

[65] S. Böcker, M. C. Letzel, Z. Lipták, and A. Pervukhin. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[66] H. Brown and L. Masinter. An algorithm for the construction of the graphs of organic molecules. *Discrete Mathematics*, 8(3):227, 1974.

[67] C. Benecke, T. Grüner, a. Kerber, R. Laue, and T. Wieland. MOLecular structure GENeration with MOLGEN, new features and future developments. *Fresenius' Journal of Analytical Chemistry*, 359(1):23–32, 1997.

[68] T. Wieland, a. Kerber, and R. Laue. Principles of the Generation of Constitutional and Configurational Isomers. *Journal of Chemical Information and Modeling*, 36(3):413–419, 1996.

[69] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354–7, 2006.

[70] A. Kerber, M. Meringer, and C. Rücker. CASE via MS: Ranking structure candidates by mass spectra. *Croatica Chemica Acta*, 79(3):449–464, 2006.

[71] E. L. Schymanski, M. Meringer, and W. Brack. Matching structures to mass spectra using fragmentation patterns: Are the results as good as they look? *Analytical Chemistry*, 81(9):3608–3617, 2009.

[72] A. W. Hill and R. J. Mortishire-Smith. Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Communications in Mass Spectrometry*, 19(21):3111–3118, 2005.

[73] M. Heinonen, A. Rantanen, T. Mielikainen, J. Kokkonoen, J. Kiuru, R. Ketola, and J. Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry*, 22:3043–3052, 2008.

[74] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, 11:148, 2010.

[75] Y. Wang, G. Kora, B. P. Bowen, and C. Pan. MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Analytical Chemistry*, 86(19):9496–9503, 2014.

[76] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, and J. Vervoort. Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid Communications in Mass Spectrometry*, 26(20):2461–71, 2012.

[77] J. Gasteiger, W. Haneback, and K.-P. Schulz. Prediction of Mass Spectra from Structural Information. *Journal of Chemical Information and Computer Sciences*, 32:264–271, 1992.

[78] J. Gasteiger, X. Li, V. Simon, M. Novič, and J. Zupan. Neural nets for mass and vibrational spectra. *Journal of Molecular Structure*, 292: 141–160, 1993.

[79] L. J. Kangas, T. O. Metz, G. Isaac, B. T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R. R. Lewis, and J. H. Miller. In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*, 28(13): 1705–13, 2012.

[80] A. Alex, S. Harvey, T. Parsons, F. S. Pullen, P. Wright, and J.-a. Riley. Can density functional theory (DFT) be used as an aid to a deeper understanding of tandem mass spectrometric fragmentation pathways ? *Rapid Communications in Mass Spectrometry*, 23:2619–2627, 2009.

[81] A. Galezowska, M. W. Harrison, J. M. Herniman, C.-K. Skylaris, and G. J. Langley. A predictive science approach to aid understanding of electrospray ionisation tandem mass spectrometric fragmentation pathways of small molecules using density functional calculations. *Rapid Communications in Mass Spectrometry*, 27(9):964–970, 2013.

[82] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer, 2005.

[83] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[84] M. Katajamaa and M. Oresic. Data processing for mass spectrometry-based metabolomics. *Journal of chromatography. A*, 1158(1-2):318–28, 2007.

[85] K. Levsen, H.-M. Schiebel, J. k. Terlouw, K. J. Jobst, M. Elend, A. Preiss, H. Thiele, and A. Ingendoh. Even-electron ions: a systematic study of the neutral species lost in the dissociation of quasi-molecular ions. *Journal of Mass Spectrometry*, 42:1024–1044, 2007.

[86] G. Landrum. RDKit: Cheminformatics and Machine Learning Software, 2013.

[87] F. Allen, R. Greiner, and D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015.

[88] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[89] W. E. Deming and F. F. Stephan. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[90] A. L. Rockwood and P. Haimi. Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, 17(3):415–9, 2006.

[91] J. Fernandez-De-Cossio Diaz and J. Fernandez-De-Cossio. Computation of isotopic peak center-mass distribution by fourier transform. *Analytical Chemistry*, 84(16):7052–7056, 2012.

[92] J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *Journal of the American Society for Mass Spectrometry*, 23(4):753–763, 2012.

[93] I. A. Papayannopoulos. The interpretation of collision induced dissociation tandem mass spectra of peptides. *Mass Spectrometry Reviews*, 14(April):49–73, 1995.

[94] B. Paizs and S. Suhai. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4):508–48, 2005.

[95] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[96] J. K. Eng, A. L. McCormack, and J. R. Yates. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry*, 5(11), 1994.

[97] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–42, 2003.

[98] J. Gasteiger and M. Marsili. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, 36 (22):3219–3228, 1980.

[99] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, and W. Brack. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Analytical Chemistry*, 84(7):3287–3295, 2012.

[100] X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. *Proceedings of the 14th International Conference on Artificial Intelligence and Statisitics*, 15:315–323, 2011.

[101] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

[102] V. Mnih, K. Kavukcuoglu, D. Silver, A. a. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[103] Y. Djoumbou. ClassyFire: A Comprehensive Computable Chemical Taxonomy. *unpublished*, 2015.