Application of Machine Learning Towards Compound Identification through Gas Chromatography Retention Index (RI) and Electron Ionization Mass Spectrometry (EI-MS) Predictions

by

Afia Anjum

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

Master of Science

Department of Computing Science

University of Alberta

© Afia Anjum, 2023

Abstract

With over 100 million synthetic chemicals and over 1 million biologically-derived compounds known to humans, chemists face significant challenges trying to identify or characterize them. In addition to this large collection of known compounds, analytical chemists, natural product chemists, pharmacologists, toxicologists, are frequently confronted with the challenge of unknown substances. These may arise as a result of biotic, abiotic or spontaneous chemical reactions. Gas chromatography mass spectrometry (GC-MS) is frequently used to identify many of these known and unknown compounds. Key to compound identification via GC-MS, is the accurate and reliable measurement of retention times (which are typically normalized to a retention index or RI) and electron impact mass spectra (EI-MS). Once RIs and EI-MS have been measured it is possible to compare these values to reference RI and EI-MS tables or libraries to identify the known/unknown compound. However, this process can be time consuming, labor-intensive and error-prone. Moreover, existing libraries of RI and EI-MS data are often inadequate, covering <1% of known compounds and, by definition, no unknown compounds. This makes the compound identification task by GC-MS quite daunting. Computational techniques, particularly those involving machine learning, can enhance the compound ID process by predicting RI and EI-MS data using only text representations of chemical structures. Among known ML methods, the most promising results for RI and EI-MS prediction have been achieved using different variants of graph neural networks (GNN). In the case of RI prediction, most RI predictors are not public. Furthermore, they neither incorporate GC column phases nor derivatization type information. This severely limits their utility. In the case of EI-MS predictors, they tend to suffer from either a lack of peak annotations or inaccurate peak intensity prediction. In this thesis, I will first describe my efforts to develop a GNN-based, freely available webserver for RI prediction, called RIpred (https://ripred.ca) that rapidly and accurately predicts GC Kováts retention indices using SMILES strings as the only chemical structure input. RIpred, performs RI prediction for three different stationary GC phases for both trimethylsilyl (TMS) derivatized and underivatized (base compound) compounds. The best performing RIpred model, when tested on hold-out test sets from all stationary phases, achieved a mean absolute percentage error (MAPE) within 3%. Secondly, I will also describe my efforts to develop a GNN-based EI-MS predictor (EI-MSpred) that accepts SMILES strings and generates an EI-MS spectrum. This predictor, which was based on a previously published model called NEIMS, utilizes a molecular ion intensity predictor (MIIP) and a peak annotation program (called PeakAnnotator), to improve its performance. EI-MSpred, when tested on a hold-out test set comprising ~2000 molecules from the NIST23 library, achieved a spectral matching score dot product score of 0.621. In terms of spectral annotation correctness and spectral annotation coverage, EI-MSpred significantly outperformed other existing EI-MS predictors, achieving an average correctness of 91% and an average coverage of 94%, on two held out test sets containing six common compounds and five random NIST23 compounds respectively. Finally, I present evidence showing the effectiveness of combining RIpred and EI-MSpred (i.e., EI-RIpred together) to aid in compound identification by simulating three GC-MS compound identification experiments.

Preface

This thesis is an original work by Afia Anjum. Dr. David Wishart, my supervisor, provided the methodological advice, training and resources needed to complete the research described in this thesis. He has also helped with reviewing and editing the thesis document. Dr. Jaanus Liigand, a post-doctoral fellow in Dr. Wishart's laboratory, helped with some data collection needed for Chapter 2, which describes RIpred. Eponine Oler, Ralph Milford and Dr. Vasuk Gautam, all of whom are members of the Wishart Lab, helped in the RIpred web server development and its subsequent testing described in Chapter 2. Dr. Marcia LeVatte, a scientific writer for the Wishart Lab, helped with editing and proofreading Chapter 2. A portion of the material described in Chapter 2 was incorporated into a paper called "HMDB 5.0: the Human Metabolome Database for 2022", which was published in Nucleic Acids Research in January 2022. The entire work presented in Chapter 2 was also published in the Journal of Chromatography A as "Accurate prediction of isothermal gas chromatographic Kováts retention indices" in August 2023. Dr. Russ Greiner provided advice regarding the ML methods and evaluation procedures mentioned in Chapter 3. Chapters 3 is currently undergoing revisions for a future publication.

Acknowledgements

I would like to thank my supervisor, Dr. David Wishart, for his patient supervision and academic counseling throughout my research program. In addition to his insightful feedback and constructive critiques, he consistently encouraged me by showing genuine appreciation of my efforts, which truly energized me. During some of my most challenging moments, he unwaveringly extended his heartfelt support. I believe the invaluable life lessons I learned from him will provide me with the wisdom I need moving forward with my career.

I would also like to extend my gratitude to all those who were directly or indirectly involved in this research journey, including Jaanus Liigand, Eponine Oler, Ralph Milford, Marcia LeVatte, along with the administrative staff of the Wishart Lab, Ms. Galina Durant and Ms. Oxana Tarassova. I also want to thank the manager of Wishart Lab Computing group, Vasuk Gautam, for their help and assistance.

Lastly, I also have no words to express my deep love and gratitude to my family. Special remembrance goes to my late mother, Rowshan Ara and my father, M. A. Salam. They not only brought me into this world but also continually supported my education. Last, but not least, I want to extend a special thank you to my best friend, my husband, and the kindest person I know, Md. Abed Rahman as well as our loving daughter, Amayrah Bint Abed, for their unwavering love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
CASE	Computer Aided Structure Elucidation
CFMID-EI	Competitive Fragmentation Modeling for compound IDentification using Electron Ionization
CI	Chemical Ionization
EBI	Electron Beam Ionization
ESI	Electrospray Ionization
GBR	Gradient Boosting Regression
GC-MS	Gas Chromatography-Mass Spectrometry
GMD	Golm Metabolome Database
GNN	Graph Neural Networks
HMDB	Human Metabolome Database
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MIIP	Molecular Ion Intensity Predictor
MLR	Multiple Linear Regression
MVC	Model-View-Controller
NEIMS	Neural Electron-Ionization Mass-Spectrometry
NIH	National Institutes of Health
NIST	US National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
PA	PeakAnnotator
PGM	Probabilistic Graphical Model
PLS	Partial Least Squares
RASSP	Rapid Approximate Subset-Based Spectra Prediction
RI	Retention Index

RFR	Random Forest Regression
SDBS	Spectral Database for Organic Compounds
SMILES	Simplified Molecular-Input Line-Entry System
SNP	Standard Non-Polar
SP	Standard Polar
SSNP	Semi Standard Non-Polar
SVR	Support Vector Regression
SMARTS	SMILES Arbitrary Target Specification
TBDMS	Tert-butyl-dimethylsilane
TMS	Trimethylsilyl

Glossary

Derivatization Derivatization is a chemical modification technique that is typically done to change the properties of a given analyte or class of analytes for better separation (via chromatography), lowere boiling points (for gas chromatography) and better ionization efficiency.

Electron Ionization Electron ionization is an ionization method in which energetic electrons interact with solid or gas phase atoms or molecules to produce ions.

Electro-spray Ionization Electrospray ionization is a technique used in mass spectrometry to produce ions by spraying molecules dissolved in a liquid through a small hollow needle to which a high voltage is applied. This process creates an aerosol of charged ions.

Fragmentation Fragmentation is a term used to describe the dissociation or break-up of energetically unstable molecular ions formed from passing molecules through a collision cell containing inert gas molecules or exposing them to high energy electron bombardment.

Gas Chromatography Gas chromatography is a technique used in analytical chemistry for separating and analyzing compounds in the gas phase. It is ideal for characterizing compounds that are volatile or which have low boiling points.

InChI The International Chemical Identifier is a textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate search in databases and web.

InChIkey InChIKey is a fixed-length format strong hashed (SHA-256 algorithm) key that is directly derived from InChI.

Mass Spectrometry Mass spectrometry is an analytical technique that is used to measure the mass-to-charge ratio of ions. Accurate measurement of the mass-to-charge ratio of compounds/ions or compound/ion fragments is often sufficient to determine their structures.

Mobile Phase The mobile phase is either a liquid or gas, which is passed through a chromatographic column. The mobile phase dissolves the analytes in mixtures and is used to push the analytes past the stationary phase to induce differntial analyte interactions and analyte separation.

Polarizability Polarizability is a measure of how easily an electron cloud can be distorted by an electric field. Typically the electron cloud will belong to an atom, a molecule or an ion. Polarizability provides an indication of how a molecule will interact electrostatically with another molecule.

Retention Index Kovats retention index (or retention index) is used to convert retention times into system-independent retention time constants in gas chromatography.

Retention Time Retention time is the time that a solute or analyte spends in a chromatographic column while it is being separated from other analytes. The retention time is a characteristic of both the column properties, the column length, the carrier gas or carrier liquid and the analyte's physicochemical properties.

SMILES The simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

Sylilation Silylation involves the replacement of an acidic hydrogen (or an active hydrogen) on the compound with an alkylsilyl group. This is a form of chemical derivatization that helps in making compounds that are generally less polar, more volatile, and exhibit increased thermal stability.

Stationary Phase This is the solid phase in chromatography which does not move with the sample. The stationary phase can be composed of specially designed beads (in liquid chromatography) or specially designed surfaces (in gas chromatography). The stationary phase interacts with the analytes being seprated and slows their passage through the column.

Chapter 1: Introduction

1.1 Background and Motivation

The world is full of chemicals. To date, more than 100 million synthetic chemicals have been prepared or described by humans [1]. In addition to these synthetic compounds, there are more than 1 million biologically-derived compounds or natural products, of which about half have already been described [2]. One of the central challenges with preparing, isolating, finding, or characterizing a chemical is determining its structure. Chemical structure determination is routinely done in the field of synthetic organic chemistry. But in these cases, the chemist has a pretty good idea of what they are attempting to make. The task of chemical structure determination is often much more challenging for analytical chemists where there is little advanced knowledge about what the chemicals are, or what their structures may be. For instance, natural products chemists spend their careers isolating and determining the structures of previously unknown or potentially medically important plant, bacterial or animal-derived chemicals [3] . Likewise, pharmacologists must spend a good portion of their time characterizing or determining the structures of previously unknown or unexpected drug metabolites [4]. Toxicologists must do the same for poisonous compounds or hazardous substances. Environmental chemists must also devote considerable effort to identifying unknown chemicals released into the environment. Many of these environmental chemicals are chemical by-products transformed into something almost unrecognizable by abiotic and biotic environmental reactions. Researchers in metabolomics must also spend a good deal of time attempting to identify or determine the structure of novel or previously unidentified metabolites coming from plants, animals, microbes, foods, drugs, cosmetics, water, or soil.

Determining the structure of a chemical can be done through several different approaches. For instance, compounds can be crystallized and their 3D structures determined by X-ray crystallography. To date, more than 1 million compounds have had their structures determined by X-ray crystallography and deposited into The Cambridge Structure Database (CSD) [5]. X-ray crystallography is particularly appealing for structure determination as it allows one to directly determine the atomic coordinates of molecules. Nuclear Magnetic Resonance (NMR) spectroscopy is another method that has been used to determine the structures of millions of compounds, however, NMR structure determination uses indirect measures concerning functional groups and various forms of chemical inference to permit structure determination. In this regard, NMR structure determination is a bit like solving a jigsaw puzzle where different pieces of the puzzle must fit together to generate a chemical structure that is consistent with all the pieces of NMR evidence. Mass spectrometry (MS) is another approach that is used for chemical structure determination and, like NMR, it requires that users must solve a chemical jigsaw puzzle by piecing suspected molecular fragments together to create a complete compound. Unlike NMR, MS also allows one to determine the molecular weight and the molecular formula of a chemical. This information can be invaluable in determining a molecular structure.

Structure determination by NMR and MS (or a combination of the two) has become the norm in modern chemistry. However, it can often take weeks or months (sometimes even years) to determine the structure of a compound, especially if it is rare or unstable. Indeed, structure determination is considered one of the main bottlenecks in nearly every field of analytical chemistry – from metabolomics to environmental chemistry to natural product chemistry to pharmacology. In other words, structure determination is a “hard” problem. But it is also a problem that can be solved or partially solved through computational techniques such as machine learning (ML) or artificial intelligence (AI). In fact, one of the very first applications of AI was the development of a program, called DENDRAL, to determine the structure of unknown chemicals using mass spectrometry [6]. DENDRAL led to the development of many related programs such as CONGEN [7] and MOLGEN [8] to generate feasible chemical structures to help solve the unknown structure problem. While DENDRAL did not "solve" the unknown structure problem, it has led to many other attempts to help determine chemical structures and has spawned a whole new field of computational chemistry called Computer Aided Structure Elucidation or CASE [9, 10, 11, 12]. It is because of the importance of structure determination to nearly all fields of chemistry and because of the recent developments in deep learning that I decided to explore the use of newer ML methods to help advance the field of CASE. In particular, the focus of this document is on developing ML techniques to improve the way in which unknown chemical structures can be determined via a technique called gas chromatography-mass spectrometry (GC-MS).

To introduce the subject, I will use this first chapter to provide some background on mass spectrometry. I will then discuss the principles of chromatography and then focus on describing the principles of GC-MS and the operation of the GC-MS instrument. Next, I will describe how compounds are commonly identified by GC-MS and highlight some of the limitations of existing methods. I will also highlight the need for being able to predict certain properties of GC-MS data (namely the retention index or RI and the mass spectra) using known or hypothesized structures. I will then review some of the methods that have been previously used to predict RI and GC-MS spectra from textual or graphical representations of chemical structures. Finally, I will describe the central hypothesis underlying my thesis and outline my thesis objectives.

1.2 Mass Spectrometry

Mass spectrometry (MS) is a technique for measuring the mass-to-charge ratio (also called *m/z*) of ions. MS is particularly useful for measuring the molecular weight of molecules. Because of its ability to measure molecular weights using very small amounts of material, MS is widely used in many areas of analytical chemistry to determine the identity of organic molecules. To perform an MS experiment, a compound must first be ionized and then accelerated through either a magnetic or electric field (under a strong vacuum) where the ion's speed or transit time can be accurately measured. MS, as a technique, was originally developed by Sir Joseph John Thomson in 1919 as a method for measuring the mass of the electron [13, 14]. However, after JJ Thomson's success, the concept of MS measurements was soon applied to many other areas of physics and chemistry including the measurement of the mass of elemental isotopes for the periodic table [15]. Later, this work was extended to measuring the molecular weight of organic molecules as a means to confirm their identity. The first commercial mass spectrometer was constructed in 1943 by the Consolidated Engineering Corporation. By the 1960s, MS had become a standard analytical tool in many chemical laboratories [16].

Key to the operation of any MS instrument is the ionization process. This involves the conversion of a neutral molecule into a charged (positive or negative) ion. Many types of ionization methods exist, including electron impact or simply electron ionization (EI), electrospray ionization (ESI), and matrix-assisted laser desorption ionization (MALDI). EI is a "hard" ionization method in which highly energetic electrons interact with gas phase molecules to produce ions [17]. EI leads to extensive molecular fragmentation, which can be helpful for the structure determination of unknown compounds. EI was one of the first ionization techniques developed for MS and continues to be widely used in gas chromatography mass spectrometry (GC-MS). On the other hand, ESI and MALDI are considered “soft” ionization methods and they generally lead to modest or negligible molecular fragmentation [18]. ESI involves spraying a liquid containing the sample of interest through a high voltage outlet into a vacuum. This leads to molecular ions being formed (along with smaller ionic fragments) as the droplets evaporate [19]. ESI is widely used in liquid chromatography mass spectrometry (LC-MS). MALDI is quite different from either EI or ESI as it involves ionizing solid samples. MALDI requires that one mix the sample of interest with light-absorbing material (called a matrix) and place it on a plate. Once plated, a pulsed laser is used to irradiate the solidified sample, triggering the desorption and vaporization of the sample from matrix material [20]. This heated vaporization of the chemical leads to ionization.

After a sample has been ionized it is usually passed through a high vacuum system via a series of strong electric or magnetic fields. This part of the MS instrument is called the “mass analyzer”. The mass analyzer is responsible for separating the ions created by the ionizer or ionization process. The mass analyzer may consist of special arrangements of multiply charged rods (called quadrupoles or Q’s) or carefully designed ion traps, large magnets, or miniature cyclotrons [21]. In some cases, the ions in the mass analyzer may be manipulated and further fragmented by passing them through gases such as helium, oxygen, hydrogen, or methane in a device called a “collision cell”. The use of a collision cell, or collision induced dissociation (CID), is the principle of operation behind tandem mass spectrometers (MS/MS) or triple quadrupole mass spectrometers (QqQ). Collision cells or CID allow the “hard” fragmentation of ions that had previously been generated by “soft” ionization methods (such as ESI or MALDI).

After the ions have been separated in the mass analyzer, they are passed on to an ion detector, which usually consists of devices known as electron multipliers (EM), Faraday cups, photomultiplier conversion dynodes (PMCDs) or array detectors [22] . The readout from the ion detector is what is called the mass spectrum. A mass spectrum of a pure compound is usually represented as a plot with the x-axis defining the mass-to-charge (*m/z*) ratio of dissociated ions from the analyzed samples and the y-axis defining the relative intensity (number of ions hitting the detector) of each of these dissociated ions (Figure 1.1). Figure 1.1a illustrates a "stylized" EI-MS spectrum of benzene (as obtained on a GC-MS instrument), while Figure 1.1b illustrates the ESI-MS spectrum of the same compound obtained via an ESI-MS instrument. As shown in Figure 1.1a, the EI method produces many ion fragments with differing m/z values and differing intensities. As shown in Figure 1.1b, the ESI method produces just one ion fragment corresponding to the molecular weight of benzene with an added hydrogen ion.

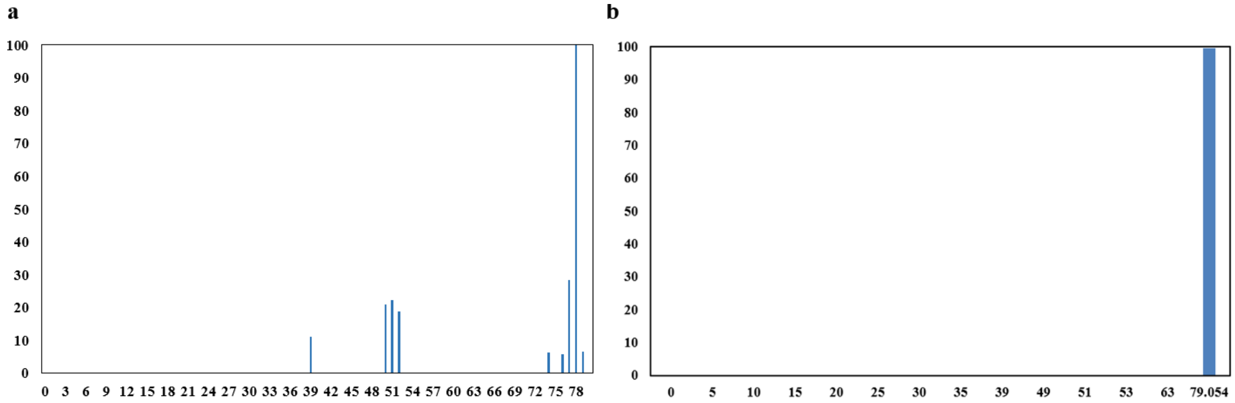


Figure 1.1 A histogram representation of an a) EI-MS and b) ESI-MS spectrum of benzene.

1.3 Chromatography

Most samples used in mass spectrometry are not pure compounds. Rather, they are often mixtures consisting of dozens of other compounds, including impurities or chemically similar compounds. The standard method for separating chemicals in mixtures is through a technique called chromatography. Chromatography is a laboratory technique for separating mixtures into their pure components. The mixture can be dissolved in either a gas or liquid (called the mobile phase), which is passed through a system (a column or a thin tube) on which a material called the stationary phase is fixed. Chromatography takes advantage of the fact that different constituents of a given mixture will tend to have different affinities for the stationary phase. This will lead to some components of the mixture being retained for different lengths of time depending on their interactions with the stationary phase. This causes the components to separate as they move down the column or the tube. Strictly speaking, in chromatography, separation is based on the differential partitioning between the mobile and the stationary phases. Depending on the physical states (gaseous or liquid) of the mobile phase, chromatography techniques can be divided into two forms: gas chromatography (GC) and liquid chromatography (LC).

GC uses gases such as He or N_2 as the mobile phase and it typically requires heating both the sample and the column to high temperatures (150-300 °C). GC is typically used to separate and identify lower molecular weight volatile compounds, such as essential oils and fragrances, pharmaceuticals and recreational drugs, toxins, air samples, etc. [23]. On the other hand, LC is conducted at room temperature using liquids such as water, methanol, or acetonitrile as the mobile phase. LC is most suitable for analyzing non-volatile, high molecular weight compounds such as lipids, peptides, steroids, and polycyclic compounds [24]. One of the most common LC methods is high-performance liquid chromatography (HPLC) where the analyzed sample is pushed by a liquid at high pressure through a column containing very small particles decorated with an absorbent material (the stationary phase) [25] . Based on the polarity of the mobile and stationary phase, HPLC can be divided into two categories, a polar version called hydrophilic interaction liquid chromatography (HILIC) and a non-polar version called reversed phase liquid chromatography (RPLC) [26]. With HILIC the stationary phase is more polar in nature than its mobile phase, while the reverse is true for RPLC methods.

1.4 GC-MS Background

As the subject of my thesis is on GC-MS and not LC-MS, I will focus the remainder of this introduction on describing the GC-MS technique, the EI-MS method and GC-MS systems in more detail. A standard GC-MS system consists of a sample injector through which samples are injected into a capillary tube or column that is between 10 and 20 m in length. The mixture to be analyzed in a GC-MS system is first evaporated or dissolved into a gas carrier (e.g. He or N_2) known as the mobile phase. The gas mobile phase carries the mixture through the column lined with a thermally stable, hydrophobic polymer adsorbed to the column interior. More specifically, the stationary phase in most GC columns is a microscopic layer of viscous liquid on a surface of solid particles on an inert solid support. The most commonly used stationary phase in GC analysis is a semi standard non-polar (SSNP) phase [27] which is a mixture of 5% diphenyl - 95% dimethyl polysiloxane (also known as DB-5). Two less frequently used stationary phases are standard non-polar (SNP) and standard polar (SP) phases, which use 100% dimethylpolysiloxane (also known as DB-1) and carbowax-polyethylene glycol respectively.

As seen in Figure 1.2, a GC-MS system consists of an oven that is used to heat the column and to ramp up or maintain high temperatures. It also consists of an injection port to inject the liquid or gas sample, a controller unit to control the flow of the carrier gas and a mass spectrometer with an electron ionizer, a mass analyzer (usually a single quadrupole) and an ion detector to collect the MS spectra of the various fragments of the analyzed sample.

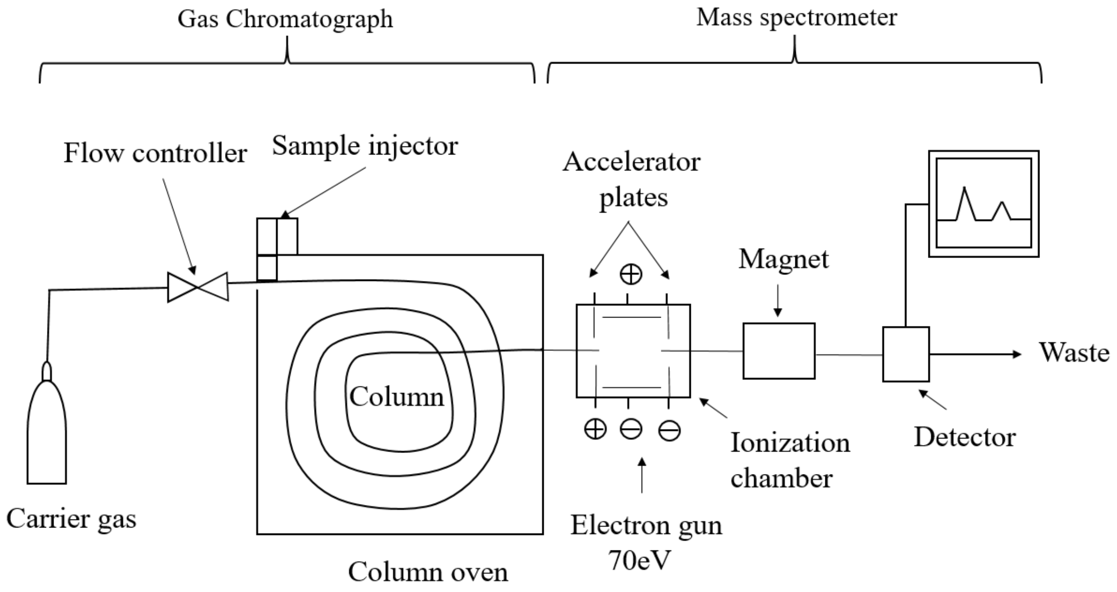


Figure 1.2: Block diagram of a gas-chromatography mass spectrometer (GC-MS).

Compounds with high volatility, high thermal stability, and high absorption rates are generally the best types of compounds for GC-MS analysis. This means that high molecular weight biopolymers such as proteins or nucleic acids are unsuitable for analysis by GC. In order to extend the range of suitable compounds for GC-MS, chemists often apply chemical derivatization techniques that involve the attachment of non-polar, bulky functional groups such as trimethylsilyl (TMS), tert-butyl-dimethylsilyl (TBDMS) or methoxime (Me-OX), onto the polar functional groups such as -NH, -OH, -SH, often found on organic molecules. This chemical derivatization process improves the stability of many compounds by lowering their boiling point and increasing their volatility. An example of how a steroid molecule can be derivatized with TMS and methoxamine is shown below (Figure 1.3).

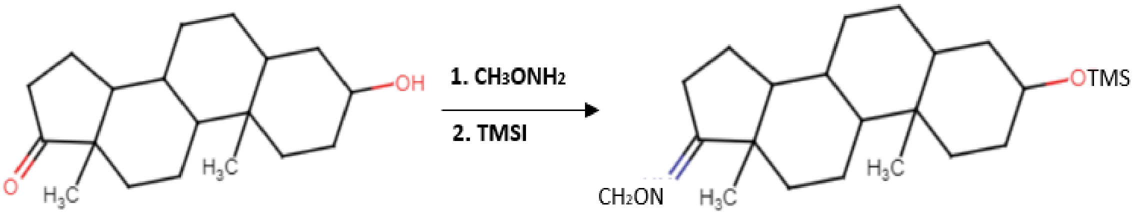


Figure 1.3: Derivatization of androsterone with TMSI and methoxyamine to make the compound more GC-amenable.

As noted earlier, with both GC and LC separations, different components in a mixture will travel at different speeds through a column. This means they will elute from the column at different times. The time at which a given component elutes from a column is specified as the retention time or RT [28]. An example of a GC chromatogram from a complex mixture (urine) is shown in Figure 1.4 (image adapted from [29]). This illustrates where peaks (corresponding to compounds) come off the GC column at different retention times.

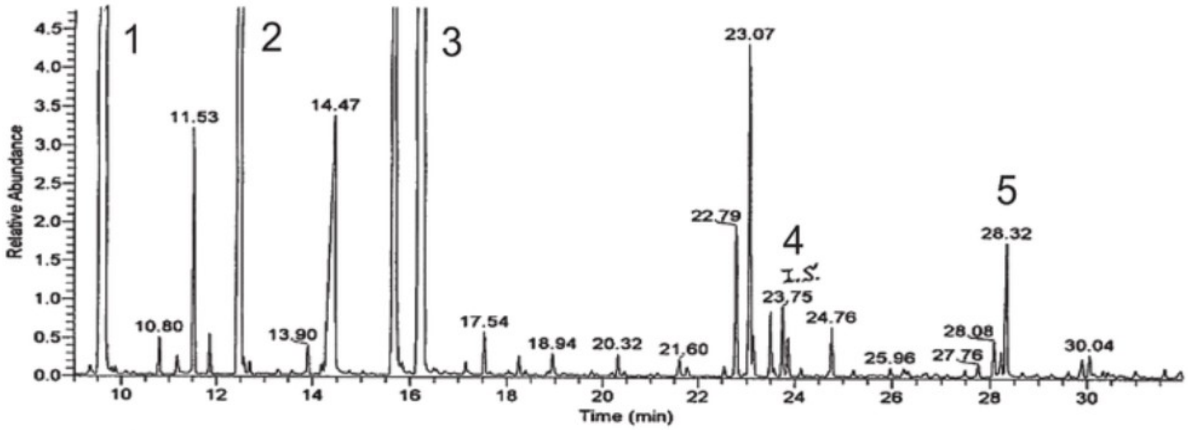


Figure 1.4: Chromatogram of urine analyzed with Gas Chromatography Mass Spectrometry (GC-MS).

Retention times are usually given in seconds or minutes. However, the RT for GC system can vary considerably due to differences in column types, film thickness, column length, column width, carrier gas velocity, and pressure. Hence, in GC, the RTs are converted into system-independent constants called retention indices (RI). The RI is usually expressed as a normalized RT ratio of the compound using its equation for calculating an RI value is given in Eq. 1. Converting RTs to RI's in GC-MS standardizes the separation time between different analytical laboratories and different analytical systems and the use of published RI tables for thousands of known chemicals has become extremely useful for the GC community in identifying chemicals.

In GC-MS, the RI can be calculated either in an isothermal or a non-isothermal setting. When a constant heating rate is implemented during the entire analysis, it is called the isothermal analysis. RIs calculated in this way are known as isothermal Kováts retention indices [30]. The equation for calculating the Kováts isothermal retention index is shown below (Eq. 1) where *n* represents the number of carbon atoms in the reference standard (typically alkanes or fatty acid methyl esters) and *Tx* represents the retention time of the sample. Finally, the adjacently eluting components' RTs can be substituted into the equation to get the sample's RI (or normalized RT).

$$RI_x = 100\text{n} + 100\ [\log(T_x) - \log(T_n)]\ /[\log(T_{n+1}) - \log(T_n)] \qquad \text{Eq.1}$$

When temperature variation exists during the analysis it is called the non-isothermal analysis (yielding temperature-programmed Kováts retention indices). It is more difficult to predict non-isothermal RIs because of the underlying complexities associated with non-isothermal analyses. While the RI approach is intended to standardize retention time reporting, RI values can vary significantly within the isothermal and non-isothermal settings, within different stationary phases and with the addition of various numbers of derivatization adducts such as TMS, TBDMS, and Me-OX groups.

As noted earlier, GC-MS uses EI to ionize molecules. In electron–ionization mass spectrometry (EI-MS), the molecular samples exiting the GC column are ionized by a high-energy electron beam of 70 eV. The high energy electrons from the beam strike the molecule and remove a single electron from it. One very basic illustration of this reaction can be represented by the electron ionization of methanol:

$$CH_3OH + e^- \rightarrow [CH_3OH]^{\cdot +} + 2e^-$$

In EI-MS, the excess of energy that is not utilized due to the removal of a single electron contributes to a phenomenon known as fragmentation. As a result, methanol will also undergo various bond cleavages, each of which can be shown in the following reactions:

$$[CH_3OH]^{\cdot +} \rightarrow CH_3O^{+} + H\cdot$$

$$[CH_3OH]^{\cdot +} \rightarrow CH_2O^{+} + H_2$$

$$[CH_3OH]^{\cdot +} \rightarrow CH_3^{+} + \cdot OH$$

These ions, each of which have a different mass or m/z value, will pass through the mass analyzer and detector yielding the following EI-MS spectrum (Figure 1.5)

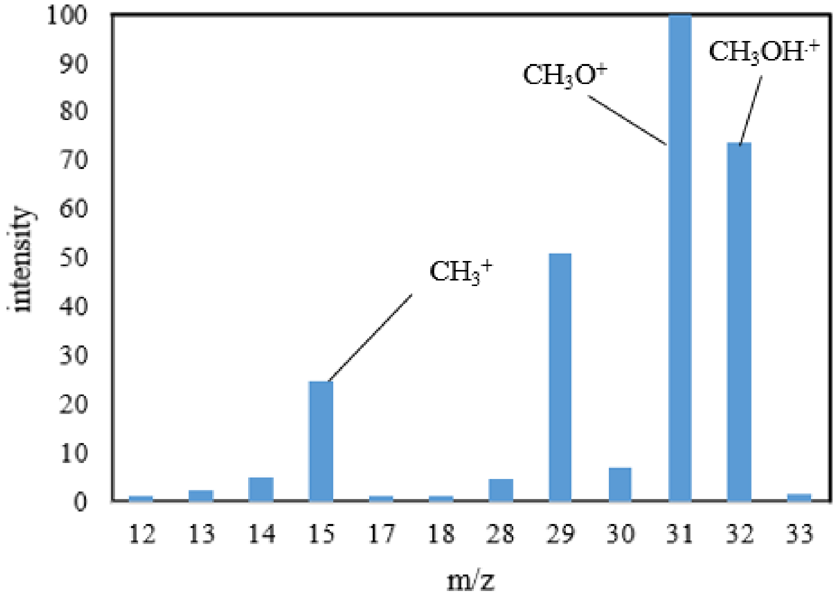


Figure 1.5: A histogram representation of an EI-MS mass spectrum of methanol.

Note that the detector in an EI setting is only sensitive toward the positively charged ions and disregards the presence of any neutral or radical molecules. The EI-MS mass spectrum of methanol shown in Figure 1.5 shows a number of peaks at different m/z positions. The one with the largest intensity is known as the parent peak or the base peak (m/z 31 corresponding to CH_3O^+ fragment), and all the intensities of all other peaks are represented as a percentage (relative abundance) of the base peak. Another important peak in any mass spectrum, which usually does not undergo any fragmentation, is the molecular ion peak and it is often referred to as the M^+ ion (m/z 32 position corresponding to $[CH_3OH]^{\cdot+}$). In EI-MS it is sometimes possible to have an almost invisible molecular ion peak for some compounds. These often include molecules such as alcohols, esters, highly branched molecules, etc. In these cases, special rules including the nitrogen rule (discussed in section 3.2.1 of chapter 3) can be used to identify the M+ ion. Different peaks in an EI-MS or GC-MS spectrum can use the molecular ion M^+ to reference other fragments in the spectrum by denoting them as M-1, M-2,….., M-n, etc. peaks. In this annotation, the n represents the difference in the m/z value between the M^+ ion and the fragment in question. Due to the presence of moderately high abundance isotopic components (such as, H^2 or C^{13}) in an EI-MS spectrum, it is also possible to have M+1, M+2,…., and M+n peaks. Generally, the value of n in such a case does not go beyond 4.

The presence of different functional groups (e.g. alcohols, aldehydes, ketones, alkanes, amines, carboxylic acids, esters, ethers, etc.) in a compound will lead to certain characteristic fragmentation patterns. These fragmentation patterns [31] arise through various fragmentation mechanisms, and analysis of these mechanisms can provide a route to determine the structure of unknown compounds. Some of the fragmentation mechanisms or reactions that have been identified over the past 50 years include events such as single or multiple cleavages of odd or even ions followed by rearrangement (which helps in stabilizing the compounds) [32], alpha-cleavage (mostly observed in aromatic ketones) [33], McLafferty rearrangements (mostly observed in aldehydes, ketones, carboxylic acids, esters, amides, etc.) [34], and retro-Diels-Alder reactions [35] (specifically observed in rings). Examples of these kinds of reactions are shown in Figure 1.6.

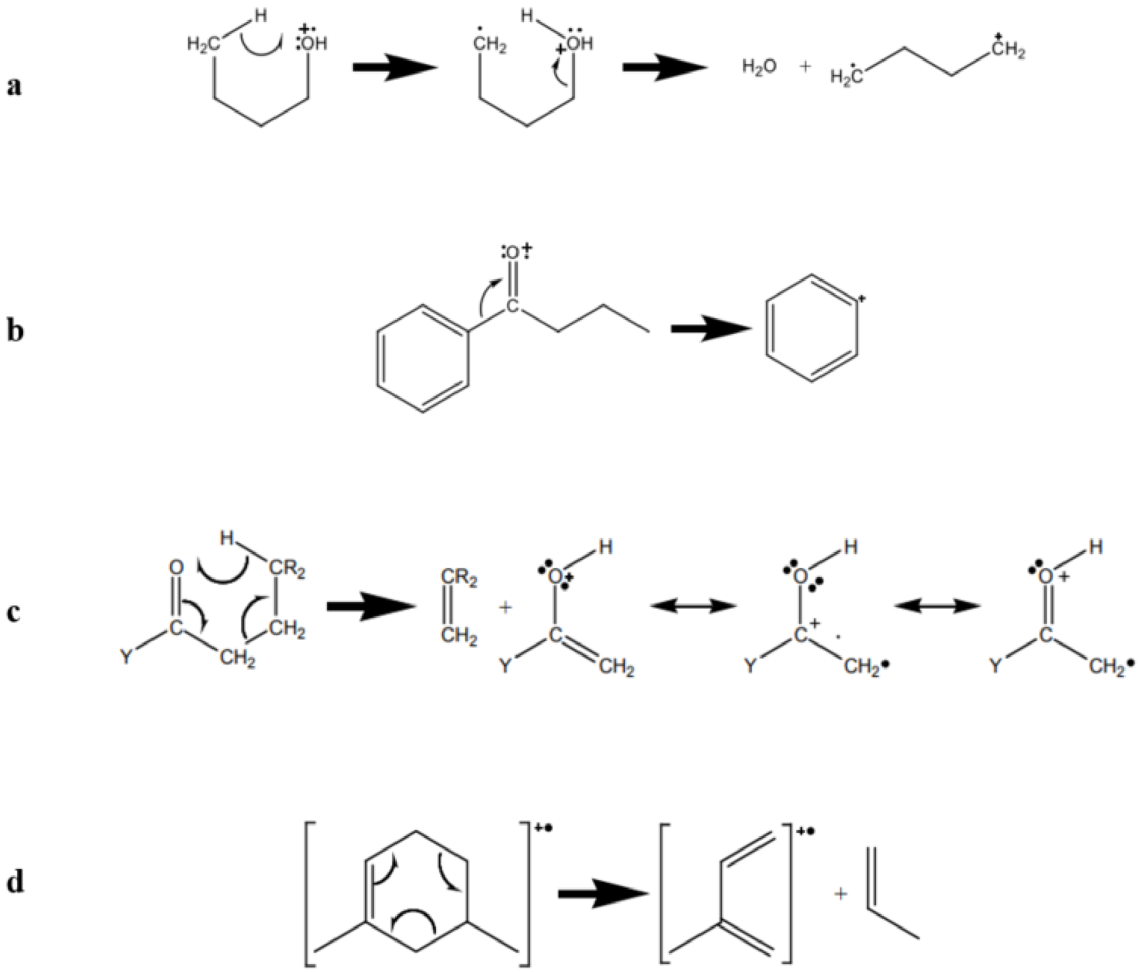


Figure 1.6: Different fragmentation mechanisms observed in EI-MS. a) Rearrangement reaction showing a loss of water from butanol. b) Alpha-cleavage occurring in an aromatic ketone to produce a benzyl radical. c) A ketone undergoing a McLafferty re-arrangement. d) Cyclic alkene undergoing a retro-Diels-Alder fragmentation reaction.

In analyzing an EI-MS spectrum or in determining the structure of a compound with a given EI-MS spectrum, knowledge of the naturally occurring isotopes and their ratio distributions can play an important role. Specifically, in an EI-MS spectrum, those compounds containing isotopes will have the same fragment ions in multiple peak positions but with different intensities reflecting their relative abundances in nature. For example, in Figure 1.7, the analyte has a peak in the M+2 position with a peak height approximately 1/4th of that of the M+ ion's peak. These two peak positions refer to the same fragment ion but they differ in their intensity values. This suggests the presence of chlorine (^{35}Cl and ^{37}Cl) within the analyzed sample. A similar trend is also observed for compounds containing isotopes of C, N, O, S, Br, I, etc. (Figure 1.8 and Figure 1.9).

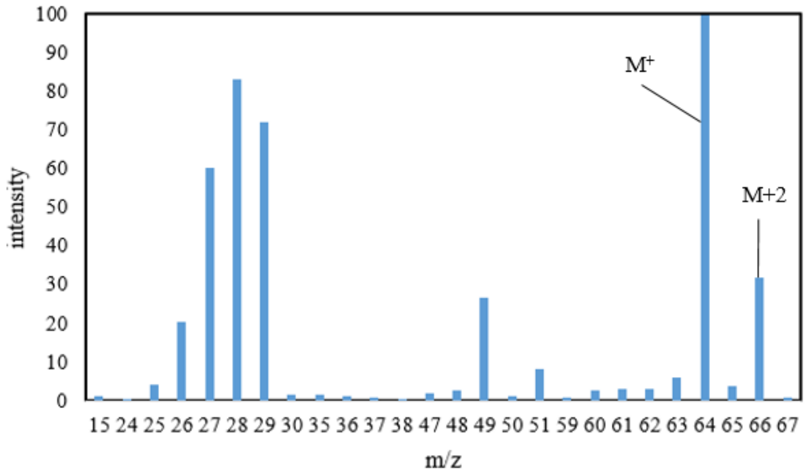


Figure 1.7: EI-MS mass spectrum for a compound (ethyl chloride) containing chlorine.

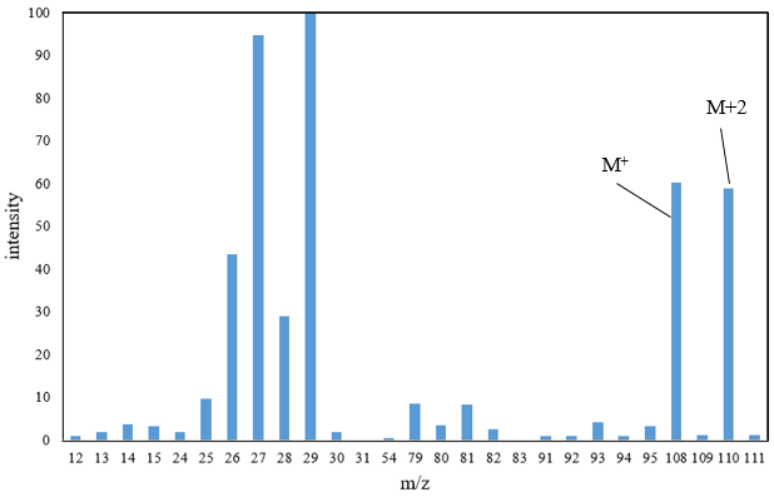


Figure 1.8: EI-MS mass spectrum for a compound (bromoethane) containing bromine.

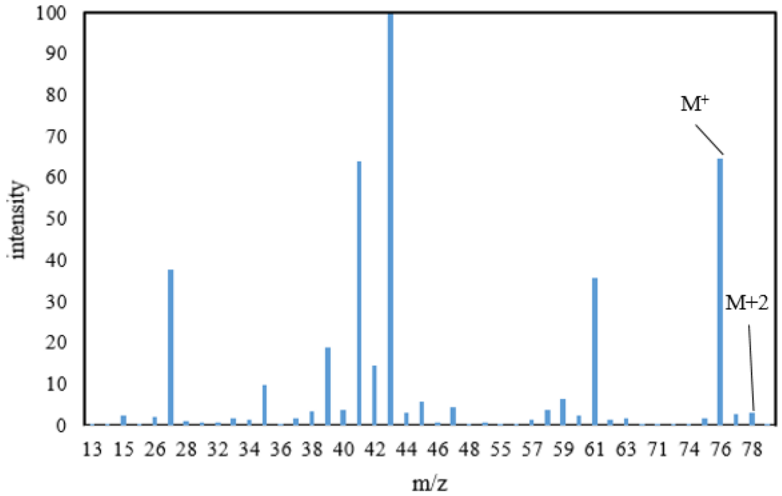


Figure 1.9: EI-MS mass spectrum for a compound (2-propanethiol) containing sulphur.

1.5 Compound Identification via GC-MS

As an analytical technique, GC-MS is widely used for compound identification and quantification. This is because the EI-MS spectra generated can provide a great deal of structural information about the constituent fragments and functional groups of a given molecule. Likewise, the retention index can also reveal the orthogonal structure of the molecule and help distinguish between two candidate structures. The other reason why GC-MS is so widely used in compound identification is because it has become highly standardized. All EI-MS spectra are collected using the exact same collision energy (70 eV) and nearly all GC retention data are converted to the same standardized Kovats RIs. This standardization means that large tables and databases of known compounds with known EI-MS spectra and known RI values can be used to identify individual compounds in mixtures by visually comparing the tabulated EI-MS spectra or tabulated RIs with the measured EI-MS spectra or measured RIs. The US National Institute of Standards and Technology (NIST) has compiled large databases consisting of 306,869 compounds with 320,704 EI-MS spectra and 447,285 Kovats experimental RIs. This is called the NIST20 database [36]. NIST has also developed software such as the Automated Mass Spectral Deconvolution and Identification System (AMDIS) [37], that allows users to compare their experimentally collected EI-MS spectra with the NIST database EI-MS spectra using different similarity functions such as the cosine similarity score (Eq. 2) or spectral entropy functions (Eq. 3) [38].

$$\cos(\theta) = \frac{\mathbf{A}.\mathbf{B}}{\|\,\mathbf{A}\,\|\,\|\,\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i \, B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad \text{Eq. 2}$$

where, A_i and B_i are components of vector A and B respectively.

$$Shannon's\ Entropy\ (ShEn) = \sum_f P_f \log\left(\frac{1}{P_f}\right) \qquad \text{Eq. 3}$$

where, *ShEn* is measured by a relational parameter that vary linearly with the logarithm of the number of possibilities f.

An example of how a compound can be identified using the AMDIS software and the cosine similarity function is shown in Figures 1.10-1.13. Prior to sending a raw chromatogram to the NIST library search, a mass spectra file is loaded using the AMDIS user interface. Once the chromatogram is loaded in the top, the software displays the corresponding mass spectra at the bottom (Figure 1.10). The spectra then goes through deconvolution (manually done by the user) (Figure 1.11) and sent to the NIST library search. This automatically opens the NIST MS program to display the result. Components of the query spectrum are shown at the top and the results are shown at the bottom (Figure 1.12). In this example, the AMDIS/NIST library search proposes the compound "*Hexacosane*" as the top hit. The white spectrum represents the deconvoluted background subtracted spectrum and the uncorrected spectrum is represented in black.

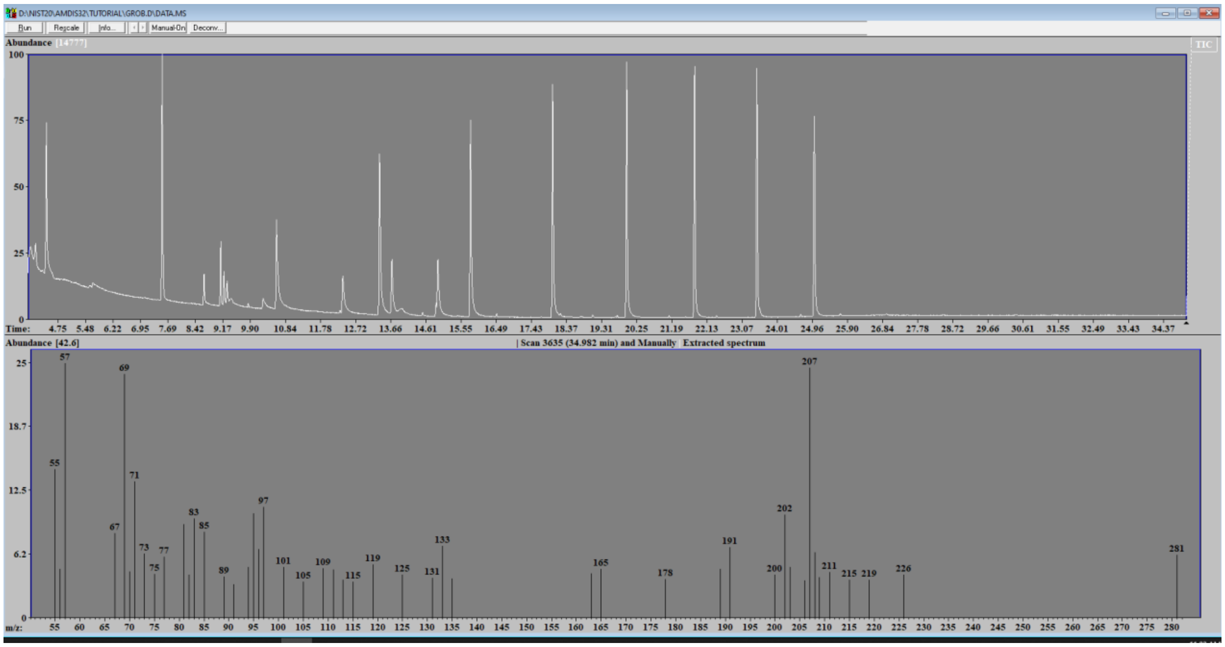
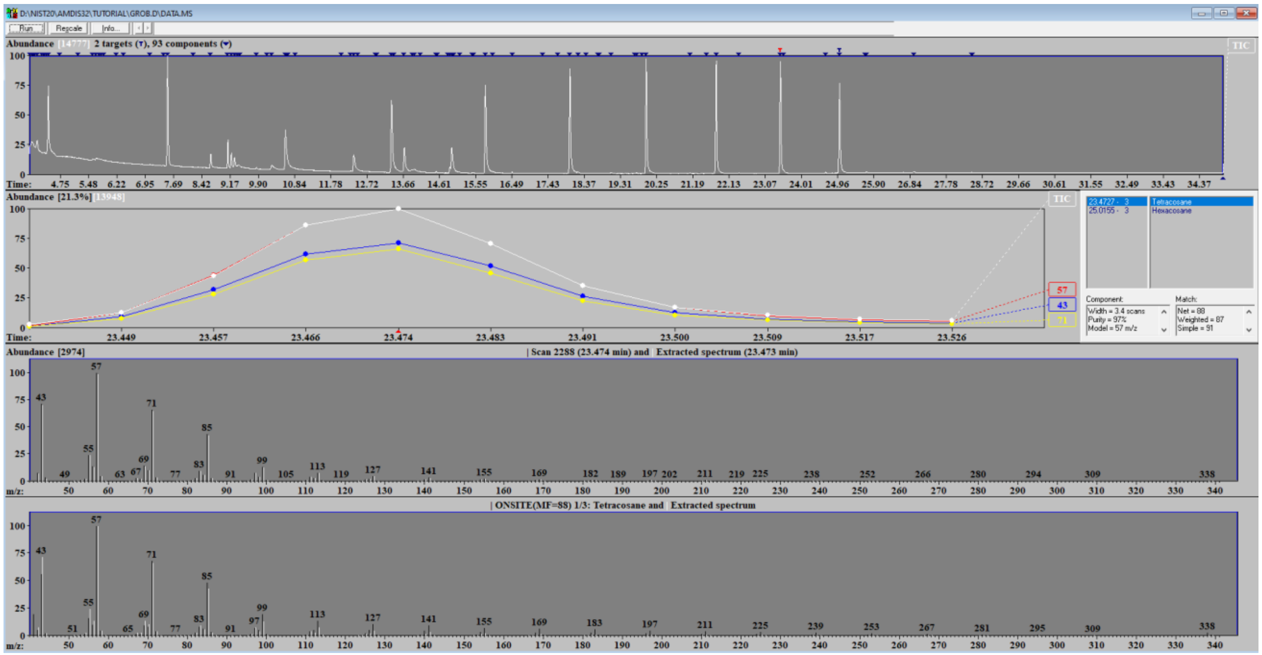

D:\NIST20\AMDIS32\TUTORIAL\GROB.D\DATA.MS
Run
Rescale
Info...
Manual-On
Deconv...
Abundance [14777]
TIC
100
75
50
25
0
Time: 4.75 5.48 6.22 6.95 7.69 8.42 9.17 9.90 10.84 11.78 12.72 13.66 14.61 15.55 16.49 17.43 18.37 19.31 20.25 21.19 22.13 23.07 24.01 24.96 25.90 26.84 27.78 28.72 29.66 30.61 31.55 32.49 33.43 34.37
Abundance [42.6]
| Scan 3635 (34.982 min) and Manually Extracted spectrum
25
18.7
12.5
6.2
0
57
69
207
55
71
67
73
75
77
83
85
89
97
101
105
109
115
119
125
131
133
165
178
191
200
202
211
215
219
226
281
m/z: 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280

Figure 1.10: An example showing compound identification using AMDIS and NIST library search program (chromatogram loaded on top, corresponding mass spectra in the bottom).


D:\NIST20\AMDIS32\TUTORIAL\GROB.D\DATA.MS

Run | Rescale | Info...

Abundance [14777] 2 targets (▼), 93 components (▼)

Abundance [21.3%] [13948]

23.4727 - 3 | Tetracosane
25.0155 - 3 | Hexacosane

Component:
Width = 3.4 scans
Purity = 97%
Model = 57 m/z

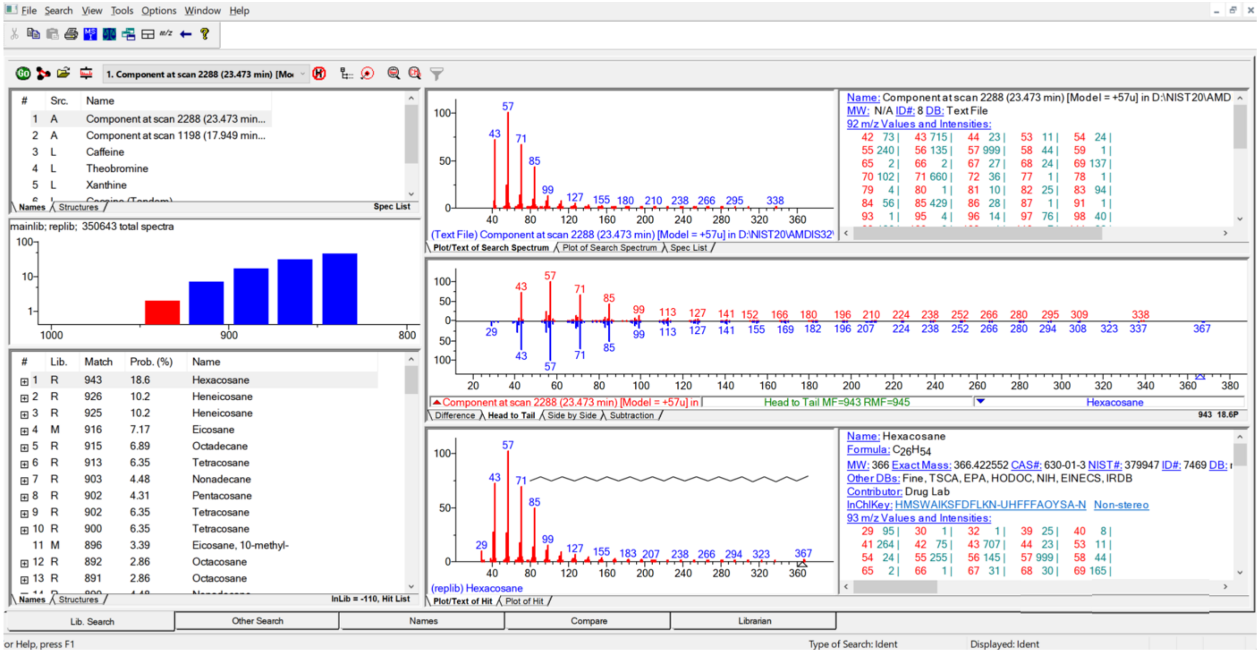
Match:
Net = 88
Weighted = 87
Simple = 91

Abundance [2974] | Scan 2288 (23.474 min) and Extracted spectrum (23.473 min)

| ONSITE(MF=88) 1/3: Tetracosane and Extracted spectrum

TIC
100
75
50
25
0
Time: 4.75 5.48 6.22 6.95 7.69 8.42 9.17 9.90 10.84 11.78 12.72 13.66 14.61 15.55 16.49 17.43 18.37 19.31 20.25 21.19 22.13 23.07 24.01 24.96 25.90 26.84 27.78 28.72 29.66 30.61 31.55 32.49 33.43 34.37
Time: 23.449 23.457 23.466 23.474 23.483 23.491 23.500 23.509 23.517 23.526
57
43
71
43 49 55 57 63 67 69 71 77 83 85 91 99 105 113 119 127 141 155 169 182 189 197 202 211 219 225 238 252 266 280 294 309 338
m/z: 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340
43 51 55 57 65 69 71 77 83 85 91 97 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 338

Figure 1.11: An example showing compound identification using AMDIS and NIST library search program (manually performed deconvolution).


File Search View Tools Options Window Help

1. Component at scan 2288 (23.473 min) [Mo

#	Src.	Name
1	A	Component at scan 2288 (23.473 min...
2	A	Component at scan 1198 (17.949 min...
3	L	Caffeine
4	L	Theobromine
5	L	Xanthine

Names / Structures — Spec List

mainlib; replib; 350643 total spectra

#	Lib.	Match	Prob. (%)	Name
1	R	943	18.6	Hexacosane
2	R	926	10.2	Heneicosane
3	R	925	10.2	Heneicosane
4	M	916	7.17	Eicosane
5	R	915	6.89	Octadecane
6	R	913	6.35	Tetracosane
7	R	903	4.48	Nonadecane
8	R	902	4.31	Pentacosane
9	R	902	6.35	Tetracosane
10	R	900	6.35	Tetracosane
11	M	896	3.39	Eicosane, 10-methyl-
12	R	892	2.86	Octacosane
13	R	891	2.86	Octacosane

Names / Structures — InLib = -110, Hit List

Name: Component at scan 2288 (23.473 min) [Model = +57u] in D:\NIST20\AMD
MW: N/A ID#: 8 DB: Text File
92 m/z Values and Intensities:

```
42  73 |  43 715 |  44  23 |  53  11 |  54  24 |
55 240 |  56 135 |  57 999 |  58  44 |  59   1 |
65   2 |  66   2 |  67  27 |  68  24 |  69 137 |
70 102 |  71 660 |  72  36 |  77   1 |  78   1 |
79   4 |  80   1 |  81  10 |  82  25 |  83  94 |
84  56 |  85 429 |  86  28 |  87   1 |  91   1 |
93   1 |  95   4 |  96  14 |  97  76 |  98  40 |
```

(Text File) Component at scan 2288 (23.473 min) [Model = +57u] in D:\NIST20\AMDIS32\

Plot/Text of Search Spectrum / Plot of Search Spectrum / Spec List

Component at scan 2288 (23.473 min) [Model = +57u] in | Head to Tail MF=943 RMF=945 | Hexacosane

Difference / Head to Tail / Side by Side / Subtraction — 943 18.6P

Name: Hexacosane
Formula: $C_{26}H_{54}$
MW: 366 Exact Mass: 366.422552 CAS#: 630-01-3 NIST#: 379947 ID#: 7469 DB: r
Other DBs: Fine, TSCA, EPA, HODOC, NIH, EINECS, IRDB
Contributor: Drug Lab
InChIKey: HMSWAIKSFDFLKN-UHFFFAOYSA-N Non-stereo
93 m/z Values and Intensities:

```
29  95 |  30   1 |  32   1 |  39  25 |  40   8 |
41 264 |  42  75 |  43 707 |  44  23 |  53  11 |
54  24 |  55 255 |  56 145 |  57 999 |  58  44 |
65   2 |  66   1 |  67  31 |  68  30 |  69 165 |
```

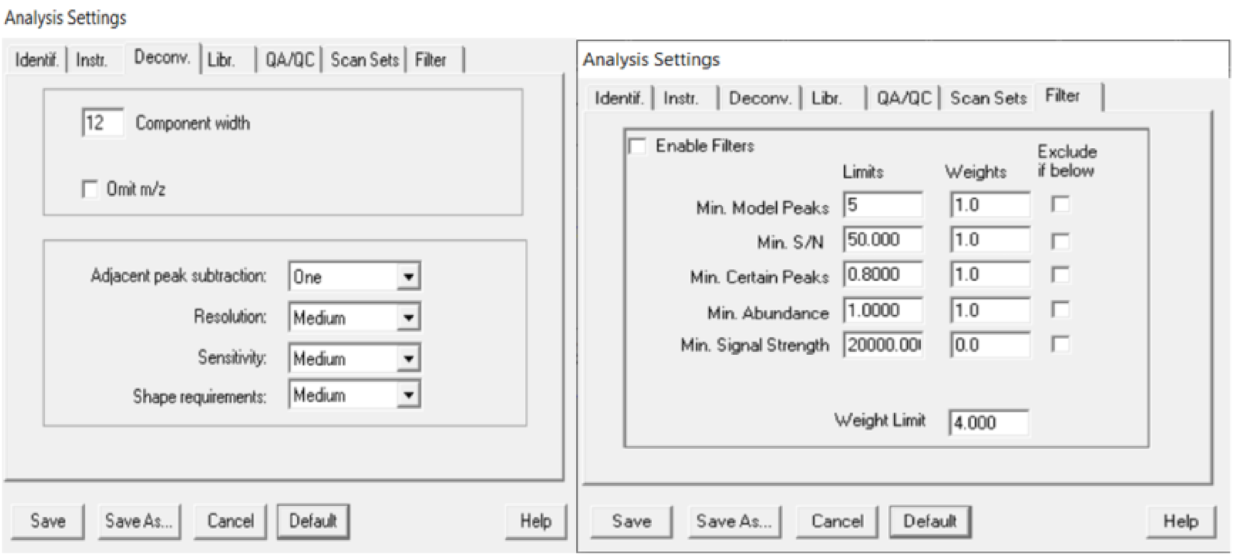
(replib) Hexacosane

Plot/Text of Hit / Plot of Hit

Lib. Search | Other Search | Names | Compare | Librarian

or Help, press F1 — Type of Search: Ident — Displayed: Ident

Figure 1.12: An example showing compound identification using AMDIS and NIST library search program ("Hexacosane" as the top hit from the library search).


Analysis Settings
Identif. | Instr. | Deconv. | Libr. | QA/QC | Scan Sets | Filter
Save | Save As... | Cancel | Default | Help
Analysis Settings
Identif. | Instr. | Deconv. | Libr. | QA/QC | Scan Sets | Filter
Save | Save As... | Cancel | Default | Help

12 Component width

□ Omit m/z

Adjacent peak subtraction: One

Resolution: Medium

Sensitivity: Medium

Shape requirements: Medium

□ Enable Filters

	Limits	Weights	Exclude if below
Min. Model Peaks	5	1.0	□
Min. S/N	50.000	1.0	□
Min. Certain Peaks	0.8000	1.0	□
Min. Abundance	1.0000	1.0	□
Min. Signal Strength	20000.00	0.0	□

Weight Limit 4.000

Figure 1.13: An example showing compound identification using AMDIS and NIST library search program (library search with customized settings).

One can select and inspect each of the peaks of the chromatogram one by one, and then search the NIST database to get matched results. The AMDIS user interface also allows users to perform library searches with customized settings (Figure 1.13) which can be further re-analyzed to get new results.

As large as the NIST20 database is, it does not cover the full set of compounds known to exist. In particular, more than 120 million compounds (with known structures) have been reported in PubChem [39]. With NIST20 having data for just 306,869 compounds, this reference library covers just 0.2% of all known chemical compounds. Given the very sparse coverage available for experimental RI values and experimental EI-MS spectra, and given the large number of chemical structures that are known (and even more that are unknown), there is obviously a huge gap in our ability to identify compound. Many have argued that the best way to fill this gap is to develop methods that can predict RI values and EI-MS spectra using known or hypothesized chemical structures as input. As a result, computational derivation or prediction of RI values and EI-MS data have become two important areas of research in GC-MS.

1.6 Prediction of RI

There has been a significant body of research aimed at predicting the RI of a molecule using only its structure (or hypothesized structure) as a starting point. These approaches use both statistical and machine learning (ML) techniques. Most are based on the observation that RI is dependent on the size of the molecule and its boiling point. Generally larger molecules with higher boiling points have longer retention times or higher RIs. Among the traditional statistical approaches, the most common ones have used data that can be easily measured including a measure of the number of heavy atoms. One approach, called the quasi-length of the carbon chain [40], essentially estimates the number of carbon atoms and how long the equivalent carbon chain would be to estimate RIs. This approach achieved a correlation between predicted and observed RIs of 0.9968 for approximately 100 polycyclic aromatic hydrocarbon (PAH) compounds that it was tested on. A second approach used the number of atom numbers, functional groups and substituents of a molecule to estimate the RI [41]. This approach reported a mean absolute percentage error (MAPE) of less than 3% for a group of compounds including acids, alcohols, amines, acid esters, aldehydes, ketones, ethers, aromatic hydrocarbons, alicyclics, and heterocyclics that it was tested on. Other approaches use the number of saturated hydrocarbons [42] to estimate RIs. This approach was tested on 43 alkyl and 48 cycloparaffins compounds and the mean absolute error (MAE) achieved for this approach was within ±1 index units. Still other groups have used thermodynamic approaches [43, 44, 45], to predict RI values. One thermodynamic method used a three-parameter model for predicting retention times for alcohols and ketones. This approach was used to predict RI for three different GC stationary phases and achieved a root mean squared error (RMSE) of 5.5s across all compounds, phases and temperature ranges. When tested across only alcohols, the RMSE value was even lower (2.79s for alcohols). However, RI prediction using thermodynamic approaches largely depends on the experimental measurements of multiple thermodynamic properties, which obviously limits their application towards generalized RI prediction. Physicochemical descriptors such as predicted boiling point, volatility, dipole moment, Henry coefficient, molar refraction, and saturated vapor pressure have also been used to predict RIs from structure [46]. This physico-chemical approach used forward and backward regression techniques, and achieved correlation coefficients greater than 0.9996 across all stationary phases. Other studies on RI have shown that molecular properties also have an effect on RI. These include properties such as electronic polarizability, ionization potential, diamagnetic susceptibility, electron density, resonant energy of the electron system of adsorbates, physical properties of adsorbates, such as the heat and energy of formation, surface tension, density, viscosity, the heat of crystallization, evaporation, and combustion, and molar volume [43]. However, these properties are difficult to predict or estimate and can only be computed for small batches of compounds or their molecular fragments. As a result, these physicochemical approaches are not considered suitable for RI determination or prediction [47].

More recently, computational chemists have moved away from statistical, heuristic or simple regression methods for predicting RI values. One early study of note employed multiple linear regression (MLR), and artificial neural networks (ANN) [48] to perform RI prediction. ANN turned out to perform the best in this comparative study with lower standard error (SE) rates (SE using the ANN: 10.479 vs. SE using MLR: 50.473) and higher correlations (ANN: 0.999 vs. MLR: 0.972) for the group of compounds it was tested on for prediction. Since then, many other ML-based techniques such as genetic algorithms [49], support vector regression (SVR) methods [50] with parameter optimization, random forest and gradient boosting regressors [51] predicted RIs using chemical descriptors.

While the correlation coefficients and estimated errors calculated from the above-mentioned studies are definitely encouraging, they were not sufficiently accurate to have a real impact on compound identification. Moreover, these approaches were often limited to working only for a specific GC stationary phase setting, only for certain functional groups [52, 53, 54, 55, 56, 57], or only for certain chemical classes [47, 49], with very small training sets. To get around the limitation of training data size, a very recent study [58] reported RI predictions for TMS derivatized compounds using a training set of 1410 TMS metabolites from the Golm Metabolome Database (GMD). This predictor employed SVM linear regression to predict RI values of TMS derivatized compounds yielding impressively low error rates (median absolute error [MAE] of 37.13 RI units and a median absolute percentage error (MAPE) of 1.95%). Much larger training sets from NIST have also been used in some very recent approaches to predict RI values [27, 40, 53, 55, 59]. These models typically outperform all previous predictors by taking advantage of the power of deep learning techniques such as convolutional neural networks (CNNs) or graph neural networks (GNNs) and very large data sets. These authors report mean absolute errors as low as 23 units and mean absolute percentage errors (MAPE) as low as 1.32%. This kind of performance appears to be sufficiently promising as to be used in direct GC-MS compound identification.

The highly accurate RI predictors discussed above provide impressively accurate results but are developed for commercial purposes only. This limits their accessibility, and it is why I believe it is important to make an open-source or open-access version of the RI predictor available to the GC-MS community. Such a resource would help researchers conduct large scale RI predictions in an effective manner. It is also notable that the highly accurate commercial RI predictors discussed above do not allow users to predict RI values for derivatized compounds or for different kinds of GC stationary phases. Ideally an open RI predictor should support the full range and scope of GC-MS RI predictions. Furthermore, such a predictor should use some of the excellent recent machine and deep learning techniques and ideas already used by the commercial tools along with some adaptation and parameter tuning. The large GC-MS RI training set available from the NIST 20 data set could potentially be used to train the predictor. As the NIST 20 dataset seems to cover compounds spanning nearly the entire chemical space, it may be possible to create separate training set partitions to handle derivatized and underivatized compounds as well as training sets for the three different GC stationary phases. This concept would mean such a predictor would have to employ six separate predictors. Making this open access predictor into webserver and making it freely available to the GC-MS community could be a game-changing development.

1.7 Prediction of EI-MS Spectra

The prediction of EI-MS spectra using only chemical structures, such as those generated via SMILES (simplified molecular-input line-entry system) has proven to be very challenging. Three approaches have been pursued 1) rule-based techniques, 2) quantum chemical calculations, and 3) machine learning strategies. Rules or heuristic rules for predicting EI-MS spectra are almost as old as the field of GC-MS itself. Examples include the alpha-cleavage rules [33], McLafferty rearrangement rules [34], and retro-Diels-Alder reaction rules [35]. Other efforts have also been undertaken to discover more reaction rules. For example, Weissberg and Dagan analyzed 200,000 entries from the NIST 2005 database and were able to extract additional fragment cleavage rules [60]. However, the number of rules derived so far are quite limited and are clearly not sufficient to cover all the cleavage reactions seen for most molecules. More recently efforts have been directed at using quantum chemical calculation methods (called QCEIMS) to predict EI-MS spectra from first principles using only chemical structures as input [61, 62]. These methods are surprisingly accurate and been used to calculate EI-MS spectra for TMS-derivatized compounds [63] as well as for environmental pollutants such halogenated organics, organophosphorus flame retardants, disinfection byproducts, etc. [64]. However, the main bottleneck for QCEIMS is its enormous computing costs and long running time, with most runs for a single molecule taking multiple days on high-end computers.

Given their obvious limitations, both the rule-based methods and quantum computation methods are not feasible for large-scale EI-MS calculations. As a result, many researchers are now looking to machine learning as a faster, more comprehensive route to predicting EI-MS spectra from structure. One of the first methods to employ machine learning in EI-MS prediction was called CFM-EI which stands for Competitive Fragmentation Modeling for compound IDentification using Electron Ionization [65]. CFM-ID EI, also known as CFM-EI is extensively used in the GC-MS community due to its noteworthy performance in both the EI-MS spectrum prediction and metabolite identification tasks. The CFM-EI method uses a probabilistic model to predict the probability of breaking molecular bonds under hard ionization EI-MS conditions and reports to perform better than full enumeration bar-coded spectra in standard spectrum prediction tasks [66]. The good recall and dot product scores reported for CFM-EI against observed spectra indicate how well this model explains fragmentation events and performs EI-MS spectrum prediction. This tool also outperforms other fragment predictive tools (models mostly generating bar-code spectra only) such as MetFrag [67], MOLGEN-MS [68] and Mass Frontier [69] in metabolite identification tasks. However, CFM-EI doesn't do particularly well in these tasks when predictions are generated for compounds covering isotope variants, silylated compounds, ethers, and organic oxygen compounds. Performance assessments of CFM-EI involving comparisons with replicate EI-MS spectra also suggest that a considerable gap still exists between experimental and the predicted spectra. Recently, several other machine learning predictors [70, 71, 72] have been described which claim to have better performance than the CFM-EI, showing higher recall values and faster run times, both in library matching and spectral prediction. Typically, these newer ML methods employ artificial neural networks and GNNs to make their EI-MS predictions. However, unlike CFM-EI, these newer ML predictors only generate EI-MS "images" (m/z values and their relative intensities) without annotation of individual fragment ions. This limits their utility in compound identification. As a result, CFM-EI, with its unique ability to annotate EI-MS spectra still has a considerable following in the GC-MS community.

One general observation is that the current CFM-EI predictor was trained on a relatively small dataset that did not cover compounds from the entire chemical space. Indeed, CFM-EI's performance drops significantly when non-conventional compounds are analyzed. This suggests that CFM-EI could be improved by training specific EI-MS models for specific classes of compounds. Furthermore, by retraining CFM-EI on a more complete set of EI-MS data (for instance, the full NIST20 database) and through the use of sophisticated model architecture for training (particularly GNNs and their variants), it may be possible to make EI-MS predictors with both high quality spectral prediction features as well as peak annotation capabilities.

1.8 Putting it All Together

If fast, high-quality predictions for RI values and EI-MS spectra could be generated using only rendered structure data (such as SMILES strings) then it should be possible to create a *large in silico* library of compounds with their (predicted) EI-MS spectra and RI values. Such a reference library would open the door to performing rapid compound identification by comparing between observed EI-MS spectra or RI values and the predicted EI-MS or RI data in this *in silico* library. This reference library could be designed to contain not only RI values and EI-MS data from the 100+ million known structures, it could also contain the same data for *de novo* predicted structures or hypothesized structures, such as those generated by CONGEN, MOLGEN or BioTransformer. Ideally when using such a reference library, EI-MS spectral matching would be done first and then the ranked output from this EI-MS matching could be further be refined by applying filters such as matches to retention index values. In other words, RI prediction and EI-MS prediction if combined together would aid in unknown compound identification.

1.9 Hypothesis and Research Objectives

Based on the previously reviewed information and concepts, I hypothesize that it should be possible to use advanced machine learning techniques and large, publicly available datasets to rapidly and accurately predict RI and EI-MS data using only SMILES strings as input. I further hypothesize that the use of these techniques will lead to among the most accurate RI and EI-MS predictors ever developed and that they could greatly improve the power of GC-MS in compound identification. To prove or disprove this hypothesis I will pursue three specific objectives:

1) Using the National Institute of Standards and Technology (NIST) 2020 database on measured RIs (derived from multiple column types and multiple TMIS derivatives) I will employ feature mining and machine learning regression techniques to develop a model to predict RI values from submitted SMILES strings. I will compare this learned RI model and its predictions with the predictions reported by other groups using the same testing and validation set to prove its accuracy.

2) Using the National Institute of Standards and Technology (NIST) 2020 database on measured EI-MS spectra I will employ feature mining and machine learning regression techniques to develop a model to predict EI-MS spectra using only SMILES strings as input. I will compare this learned EI-MS model and its predictions with the predictions reported by other groups using the same testing and validation set to assess its accuracy.

3) I will make these resources publicly available through the deployment and maintenance of user-friendly webservers to ensure maximal impact for the GC-MS community

Chapter 2: Accurate Prediction of Isothermal Gas Chromatographic Kováts Retention Indices[1]

2.1 Introduction

Gas chromatography (GC) is a technique that can be used to separate mixtures of liquid, gaseous, semi-volatile and even non-volatile components. GC is usually carried out at high temperatures and analytes are separated based on their boiling points and their interactions with the GC column's stationary phase. The amount of time a compound spends on a GC column is called the retention time (RT) [30]. However, the RT of a compound can vary with the GC column type, length, diameter, film thickness as well as the GC instrument's void volume, carrier gas velocity, and pressure. While a parameter called "specific retention volume" was historically used to make retention data independent of these variables [47], most modern GC-MS laboratories convert RTs into system-independent constants called retention indices (RI). RIs, which normalize the RTs of compounds with its adjacently eluting n-alkanes [73], allow retention values measured by different analytical laboratories under varying GC conditions to be compared. It should also be noted that RT comparisons across laboratories can also be achieved using techniques such as retention projection [74] and retention locking [75]. However, our focus here is on RI calculation. Two techniques used to calculate RIs during analysis include isothermal methods, which maintain constant temperature (known as Kováts RIs [76]), and non-isothermal methods, which vary the

[1] A slightly modified version of this chapter was published in the Journal of Chromatgraphy A 2023 Aug 30:1705:464176

temperature (known as temperature programmed Kováts RIs). While isothermal and non-isothermal RI values of a given compound vary considerably, non-isothermal RIs can be converted to isothermal RI values using simple empirical models. RI values also vary significantly with different GC stationary phases. For instance, polar compounds have long RTs on polar columns and shorter RTs on non-polar columns. The most common stationary phase used in GC analysis is a semi standard non-polar (SSNP) phase [27], which usually contains DB-5, a polysiloxane-based co-polymer consisting of 5 % diphenyl and 95 % dimethyl polysiloxane. The SSNP can separate a range of chemicals from alkanes, amines, phenols, fatty acids, methyl esters, and even volatile compounds. Other common stationary phases include Standard non-polar (SNP) and Standard Polar (SP) using DB-1-100% dimethylpolysiloxane and carbowax -polyethylene glycol, respectively.

Experimental RI data for thousands of underivatized (volatile compounds that do not need chemical modification) and derivatized compounds (less volatile compounds that are chemically modified via silylation to enhance volatility) can be found in reference libraries maintained by the National Institute of Standards and Technology (NIST) [36] and the National Institutes of Health (NIH) via PubChem [39]. However, these experimentally measured RI libraries cover only a fraction of the millions of known, expected and suspected compounds. The lack of experimentally collected reference RI data for most GC-MS detectable compounds limits their reliable identification. As manual collection and measurement of RI data is both time-consuming and expensive, methods to accurately predict the RIs of compounds (based on their structure) has been pursued for decades [49, 53, 56, 58-59, 77, 78, 79].

RI values (with the exception of RIs for the polar stationary phase) depend on molecular size and measured boiling points. This information was incorporated into the first RI predictors. One of the first RI predictors estimated RIs using the quasi-length of the carbon chain and the pseudo-conjugated system surface of polycyclic aromatic hydrocarbons (PAHs) as measures of molecule size [40]. While this approach achieved a high linear regression coefficient for 100 PAHs, it was not transferrable to other types of molecules. A second approach used empirical counting of atom numbers, functional groups, and substituents to estimate RIs for hundreds of compounds of several chemical classes in both non-polar and polar column [41]. Although this method achieved RI error rates of < 3%, it was not applicable to derivatized compounds. A later approach used the number of saturated hydrocarbons (determined by graphical methods) and linear regression analysis to estimate RIs for saturated hydrocarbons [42]. This study predicted RIs for 43 alkyl and 48 cycloparaffins compounds with reported errors ±1 index units. This approach was also limited in its scope and number of compounds to which it could be applied.

In addition to traditional statistical or regression methods, other RI predictors have used measured experimental properties to predict RIs. One class of predictors uses an additive thermodynamic (AT) approach [43-44, 80]. One notable AT method used a three-parameter model to predict RTs for alcohols and ketones for three different GC stationary phases. When tested across all compounds, phases and temperature ranges, this predictor achieved a low root mean squared error (RMSE). However, as AT approaches measure multiple thermodynamic properties, this limits their application to generalized RI prediction. Measured physicochemical descriptors such as the boiling point, volatility, dipole moment, Henry coefficient, molar refraction, and saturated vapor pressure have also been used to predict RIs [46]. This pseudo-experimental approach used forward and backward regression techniques and achieved high correlation coefficients across all stationary phases. Other studies measured even more physicochemical descriptors and also reported impressive results [81]. However, these physicochemical methods are impractical for general RI prediction due to the high number of experimental parameters that must be collected.

The success of RI predictive methods that used thermodynamic and physicochemical data led researchers to explore RI prediction based on predicted topological, thermodynamic and physicochemical data [82, 83, 84]. The development of more sophisticated machine learning (ML) regression methods over the past decade has led to even more accurate and more sophisticated methods for RI prediction. Indeed, an early ML study compared how traditional multiple linear regression (MLR), traditional partial least squares (PLS) and an artificial neural network (ANN) predicted RIs and showed that the ANN performed best [78]. Since then, other ML-based approaches to RI prediction have appeared. Most use predicted chemical descriptors of the structures [85-86], which are then fed as features in various ML algorithms [48, 50, 54, 86, 87, 89, 90] including support vector regression (SVR), random forest regression (RFR) and gradient boosting regression (GBR) and many other ML techniques. Unfortunately, many studies used small training sets and were limited to a specific GC stationary phase setting or were built for specific functional groups [50, 53, 54, 88] or chemical classes of molecules [48, 77, 89, 90]. This greatly reduces their utility and general applicability.

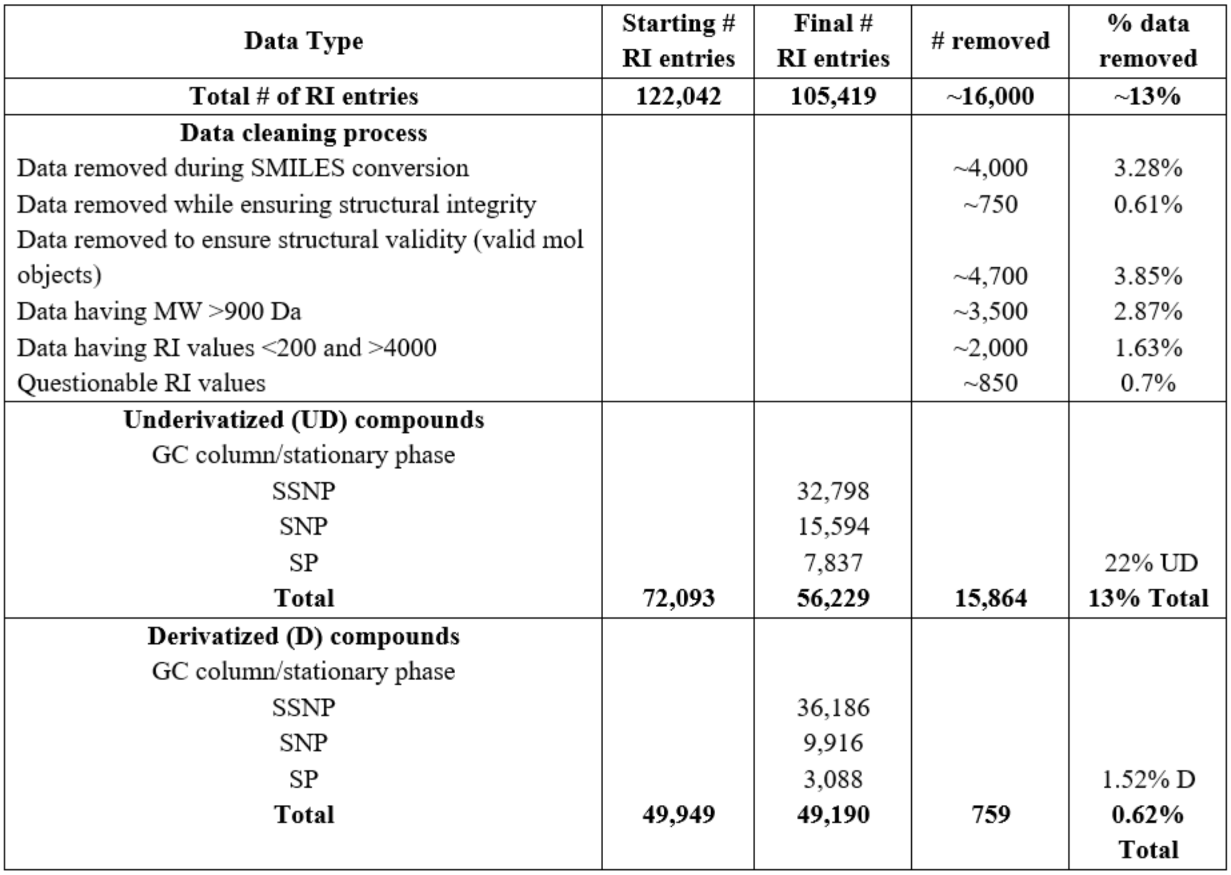
Recently, several groups have expanded their testing and training data sets and used more advanced ML regression models to make more generalized RI predictors. Most of these [53, 56, 58, 59, 79] have used large GC-RI training sets from NIST (139,498 compounds) or PubChem's collection of experimental RI values and exploited ANN regression methods and their variants such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs). These newer models effectively outperformed all prior RI predictors with very low mean absolute errors (MAEs – 23 RI units) and very low mean absolute percentage error (MAPE – 1.32%) [49].

Unfortunately, many highly accurate RI predictors are closed source and/or only available commercially. Moreover, they do not allow users to predict RI values for derivatized compounds or RI values for different stationary phases. To address these shortcomings, we developed an RI predictor called RIpred, a software program that supports the full range of RI predictions that GC-MS users commonly need. RIpred exploits several key features of the NIST RI prediction model, first described by Qu et al. [79]. However, RIpred uses a different set of molecular graphs, different kinds of atom-level features and predicts RI values for both underivatized and trimethylsilyl (TMS) and/or tert-butyldimethylsilyl (TBDMS) derivatized compounds for all three stationary GC phases. RIpred was extensively trained and tested on a wide range of experimentally measured RI data and yielded RI results equal to or better than those published elsewhere [48, 49, 50, 53, 56, 59, 79, 89, 90]. The RIpred model was then incorporated into a user-friendly webserver that allows users to submit any structure and obtain one (for underivatized) or more (for derivatized) RI predictions. RIpred also houses predicted RI values for >5 million compounds that can be instantly retrieved or downloaded. The next sections describe how RIpred was developed, trained and tested and how it can be used by anyone to predict RI values.

2.2 Data and methodology

2.2.1 Retention index data sets

Our primary data source was the RI library from the official NIST 17 and NIST 20 releases [8, 38]. The updated dataset from NIST 20 consists of 447,285 Kováts and LeeRI values for 139,498 compounds. To maximize the size of the training set, we merged the NIST 17 and NIST 20 GC-MS Kováts RI data. To clean and standardize the data, we removed the Lee RI values along with any entries containing duplicate compounds. As a result, ~26,000 Kováts RI values from NIST 20 were added to the NIST 17 dataset, yielding a final Kováts RI dataset consisting of 122,042 entries for 105,075 unique compounds. This included 72,093 base (underivatized) compounds and 49,949 derivatized compounds (Table 2.1). This dataset contains experimental Kováts RI values for three



Data Type	Starting # RI entries	Final # RI entries	# removed	% data removed
Total # of RI entries	**122,042**	**105,419**	**~16,000**	**~13%**
Data cleaning process				
Data removed during SMILES conversion			~4,000	3.28%
Data removed while ensuring structural integrity			~750	0.61%
Data removed to ensure structural validity (valid mol objects)			~4,700	3.85%
Data having MW >900 Da			~3,500	2.87%
Data having RI values <200 and >4000			~2,000	1.63%
Questionable RI values			~850	0.7%
Underivatized (UD) compounds				
GC column/stationary phase				
SSNP		32,798		
SNP		15,594		
SP		7,837		22% UD
Total	**72,093**	**56,229**	**15,864**	**13% Total**
Derivatized (D) compounds				
GC column/stationary phase				
SSNP		36,186		
SNP		9,916		
SP		3,088		1.52% D
Total	**49,949**	**49,190**	**759**	**0.62% Total**

Table 2.1: Summary of number of data types in the NIST datasets before and after data cleaning.

stationary phases (SSNP, SNP and SP) for all 122,042 entries along with their structure (consisting of 2D representation of the molecule). The data also contains other metadata including compound names, InChI keys, masses, formula, predicted RI values (present only in NIST 20 for the SSNP phase [91]), instrument details, etc. For training purposes, we needed all structures expressed as a simplified molecular-input line-entry system (SMILES) string. As the NIST 17 and NIST 20 software did not allow us to retrieve chemical structures as SMILES or SDF files, we used the NIH Cactus service [92] and the PubChempy python [93] application programming interface (API) to convert the InChI keys to isomeric SMILES. This conversion was not successful in all cases, and ~4,000 RI values along with their associated structures were unable to be processed. Further manual analysis and checking was performed to ensure that the correct SMILES strings were generated for different stereoisomers of the same compound and for different-sized derivatives of these compounds (with one or more additions of TMS derivatives: i.e. 1 TMS, 2 TMS, etc.). This manual checking step removed ~750 compounds. To ensure the structural validity of all SMILES representations, we also converted them into valid molecular objects (MOL files) using RDKit [94] and discarded those entries (~4,700) where no mol file could be generated. We also removed compounds having molecular mass >900 Da from the dataset (~3,500), as heavier compounds are not generally analyzed by conventional GC instruments. In addition to ensuring the structural integrity of our dataset, we also made sure that it provided reasonable values of RIs in the three stationary phases. We discarded compounds with an RI <200 units and >4000 units (~2,000 compounds) as reasonable RI values for most conventional GC-MS instruments and most conventional compounds are between 200-4000 units. We also manually investigated questionable RI values during the data cleaning process which led to the removal of approximately 850 compounds. The final "clean" dataset, where ~16,000 entries were removed (13%), consisted of 105,420 experimental Kovats RI values from 56,229 underivatized (13% removed) and 49,190 derivatized (0.63% removed) structures (Table 2.1). The final number of associated structures for three different stationary GC phases included 32,798 underivatized and 36,186 derivatized structures having SSNP RI values, along with 15,594 underivatized structures and 9,916 derivatized structures having SNP RI values as well as 7,837 underivatized structures and 3,088 derivatized structures having SP RI values. More details regarding these distributions are provided in Table 2.1.

2.2.2 Graph neural network methods

To develop our RI predictor, we drew inspiration from the GNN-based RI predictor described by Qu et al. [79] while implementing a number of our own modifications. The GNN predictor described by Qu et al. made improvements to the open-source GNN model defined by the MatErial Graph Network (MEGNet) framework [95]. While the predictor described in [79] and the MEGNet framework used TensorFlow [96], Keras APIs, and RDKit, we developed RIpred using a tensor-optimized library called PyTorch [97]. For atom-level feature extraction, we also used RDKit functions. The use of PyTorch along with SMILES strings led to the development of a GNN that used a somewhat different workflow, a different set of atom/molecular features and a different approach to using molecular graphs than the GNN developed by Qu et al. [79].

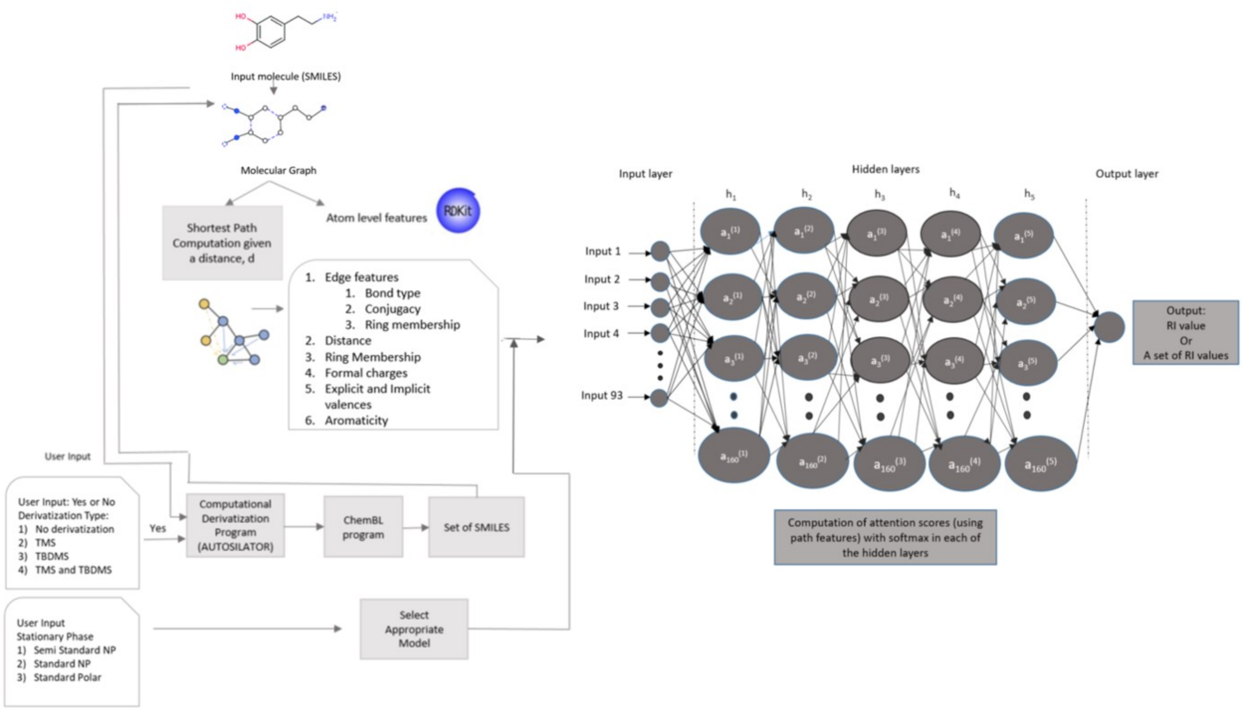
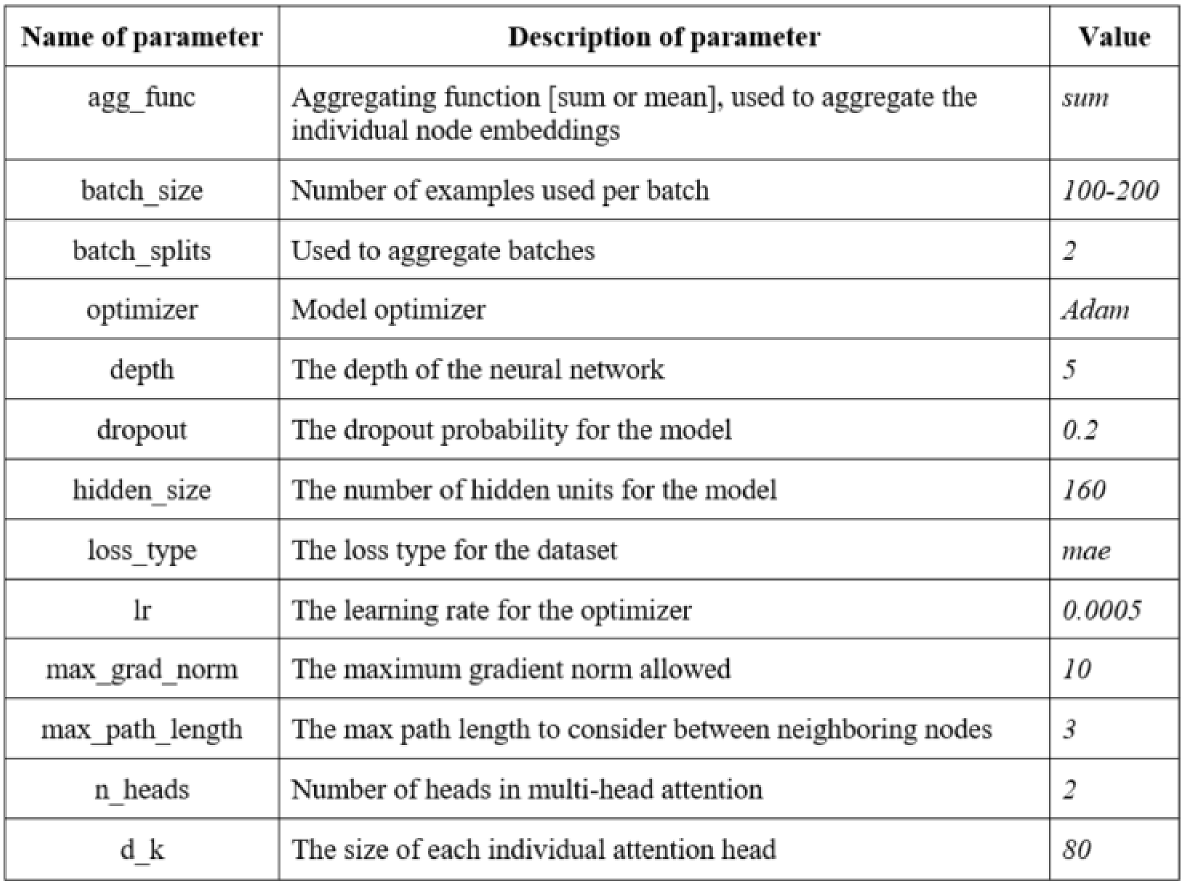


Figure 2.1: Overview of the RIpred predictor workflow. Both the training and testing phases of the RIpred predictor are illustrated here. During the training phase, multiple batches of SMILES representations from the molecular structures of each dataset were converted into molecular graphs. Each of these molecular graphs were further used to compute the shortest distance paths. Atom-level features were extracted using both the molecular graph representations and shortest path distances utilizing RDKit. Six predictors were trained individually in the same way and embedded in the webserver. The RIPred web server accepts SMILES representations of the molecular structure (underivatized or derivatized), the derivatization type and the stationary phase information for predicting the RI values. The AUTOSILATOR derivatization script and the ChemBL program are used to generate the derivatized forms of the user-submitted molecule.

The overall workflow for RIpred's GNN is illustrated in Figure 2.1. RIpred accepts a SMILES string corresponding to the query structure. The SMILES string is then directly converted into a molecular graph where the nodes and edges of the graph represent atoms and connections between the atoms, respectively. Using RDKit, a number of atom-level features are then extracted from the molecular graph including a number of atomic (node-level) features and a number of path features, related to the edges. The atomic feature set for RIpred is a combination of 64 one-hot coded features representing 62 types of common atomic symbols (such as C, N, S, O, F, etc.), plus one feature for any unknown elements and the other feature as a default wildcard or dummy symbol. Formal charge calculation can suggest potential polarizability. As a result, we included five types of valid formal charges ranging from neutral, positive and negative charges [0, ±1, ±2] in our atomic feature set. This property is calculated based on the assigned charges for each atom within a molecule. We also included seven explicit and six implicit valences in this feature set. Additionally, our atom feature set also includes a path feature related to the neighboring nodes of any given node, up to a maximum of 10 neighboring atoms. Beyond the atomic features, we also encoded path features for any two given nodes of each molecule by computing the shortest path between them. The maximum path length we considered between any pair of nodes was three. For any pair of nodes, the path feature set is a combination of 31 one-hot coded features. This is computed as a concatenation of the model's maximum path length and the total number of bond features in that particular path, bond conjugacy and bond ring appearance (denoting the presence of a ring between two neighboring nodes) for that path and the ring membership (denoting whether the nodes in a path are in aromatic rings) between the nodes in a path. Six types of bonds are used to compute these bond features. The atom or node level features that are computed for any two nodes in a graph are fed into the input layer of the network by a linear transformation. In addition, the RIpred GNN model also computes atom attention scores both in the hidden layers (with five hidden layers each of the layers having 160 hidden units) and the output layer. The new atom embeddings which are computed afterwards are particularly helpful in generating output values from the final layer of the network. The RIpred GNN model code and scripts were adapted and modified from the Python source code of a publicly available GitHub GNN repository [98].

All the parameters used to train the RIpred models are listed in Table 2.2. Except for a few parameters (e.g. batch size), all other parameters were kept constant in these models. When one of our models generated an initially acceptable level of performance, we conducted k-fold cross validation (with k=10) to test the reliability of that model and calculated the MAE and MAPE with means and standard deviations of MAEs for all 10 cross-validation runs. This k-fold cross validation was applied to obtain less biased or less optimistic estimates of the test errors. To conduct the k-fold cross validation, we divided the available datasets from all stationary phases into k folds and trained the model k times. In each of these k evaluations, we used k-1 folds of data for training and the remaining fold for testing the models. As we used a GNN to train our models, in each of these k evaluations, we captured both the training versus validation errors (generated from the internal validation/development set used during model optimization/tuning) over different epochs to select the best trained models. Later, the best trained models were used to report the k (=10) fold cross validation training and testing errors. The best trained model from these 10 runs was selected by looking at a single run in the cross-validation set that produced similar MAE values with that of the mean MAEs from all 10 evaluation runs. In this way, multiple rounds of training, testing and selection led to the identification of the six best models for three stationary phases, covering both underivatized and derivatized forms. These top models were trained for up to 3,000 epochs. Early stopping was used to help efficiently identify the most promising models. For our GNN models, the early stopping was defined by a patience parameter, which was varied (20, 100, 150, and 200 times) to capture the most effective training model. In the next step, we used these top six models to evaluate the six respective hold-out test sets and report their errors and overall



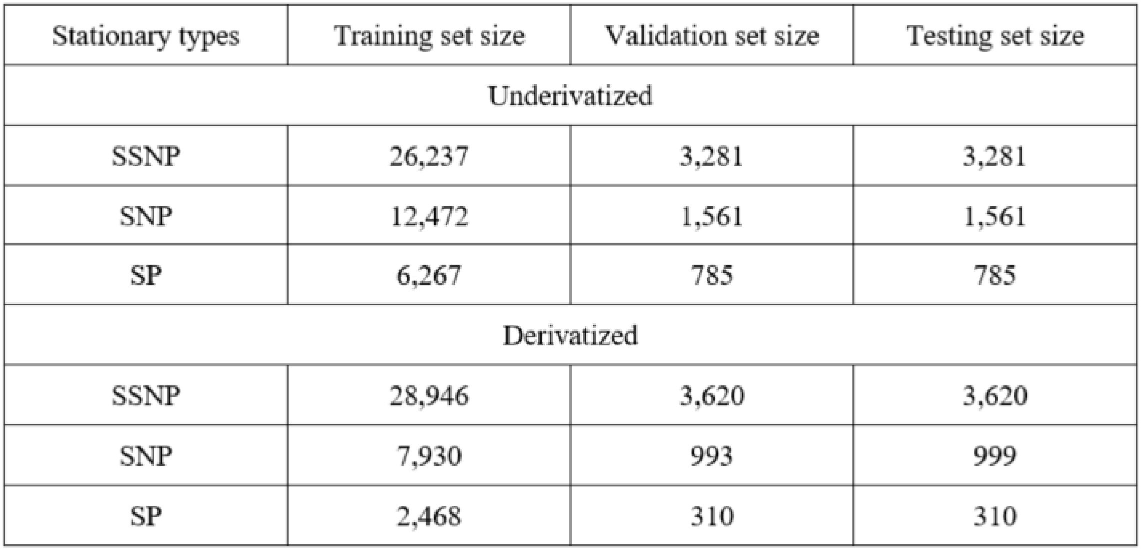
Name of parameter	Description of parameter	Value
agg_func	Aggregating function [sum or mean], used to aggregate the individual node embeddings	*sum*
batch_size	Number of examples used per batch	*100-200*
batch_splits	Used to aggregate batches	*2*
optimizer	Model optimizer	*Adam*
depth	The depth of the neural network	*5*
dropout	The dropout probability for the model	*0.2*
hidden_size	The number of hidden units for the model	*160*
loss_type	The loss type for the dataset	*mae*
lr	The learning rate for the optimizer	*0.0005*
max_grad_norm	The maximum gradient norm allowed	*10*
max_path_length	The max path length to consider between neighboring nodes	*3*
n_heads	Number of heads in multi-head attention	*2*
d_k	The size of each individual attention head	*80*

Table 2.2: Parameters and their values used in our RIpred model.

predictive performance. All six RI prediction models were then incorporated into the RIpred webserver.

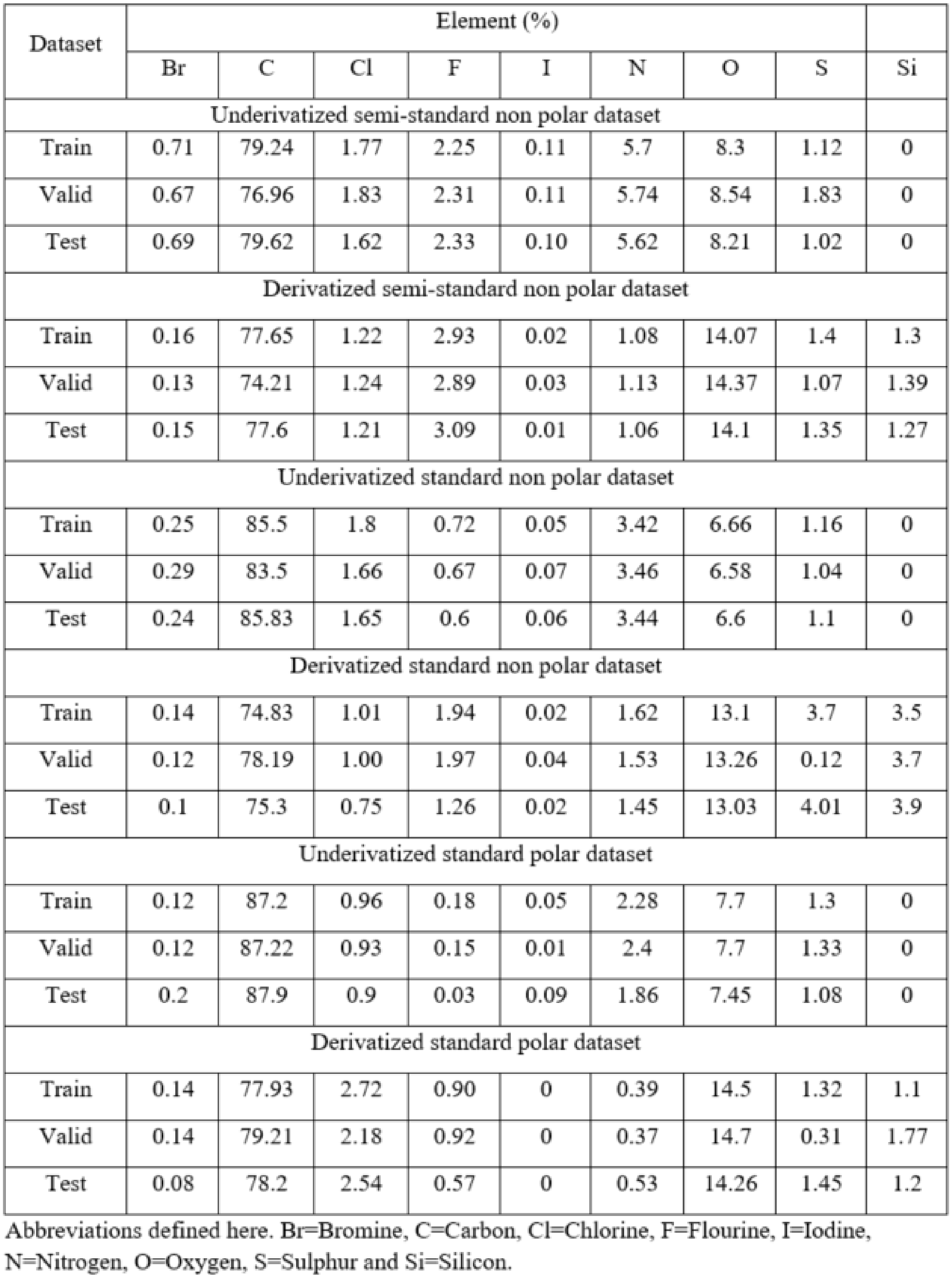
2.2.3 Dataset Distribution

Prior to training, we separated the derivatized and underivatized compounds into three datasets using the three kinds of stationary phase information from the combined NIST data set. This gave rise to a total of six different datasets. As mentioned above, we conducted k-fold cross-validation to get an unbiased estimate of the test errors. The number of data points used for training, validating and testing each of the stationary phases within each of these k-folds is shown in Table 2.3. It is important to note that the validation set defined here is the development set that is used solely for the GNN model optimization/tuning while training the model each of the k times. The split within each of these sets was done in such a way that the same proportion of similar compounds (in terms of elemental ratios, ClassyFire [99] chemical class and superclass) were used in training, validating, and testing the data (see Table 2.4, 2.5, and 2.6). We also made sure that the exact mass vs. number of rotatable bonds vs. RI and exact mass vs. LogP (octanol-water partition co-efficient) vs. RI for these sets of compounds were similar (see Figure 2.2). In order to ensure distinct derivatized datasets for all three stationary phases, we separated the derivatized compounds by their common chemical names, or by using a SMILES arbitrary target specification (SMARTS) string search for derivatized functional groups within each SMILES string. The functional/derivatization groups



Stationary types	Training set size	Validation set size	Testing set size
Underivatized			
SSNP	26,237	3,281	3,281
SNP	12,472	1,561	1,561
SP	6,267	785	785
Derivatized			
SSNP	28,946	3,620	3,620
SNP	7,930	993	999
SP	2,468	310	310

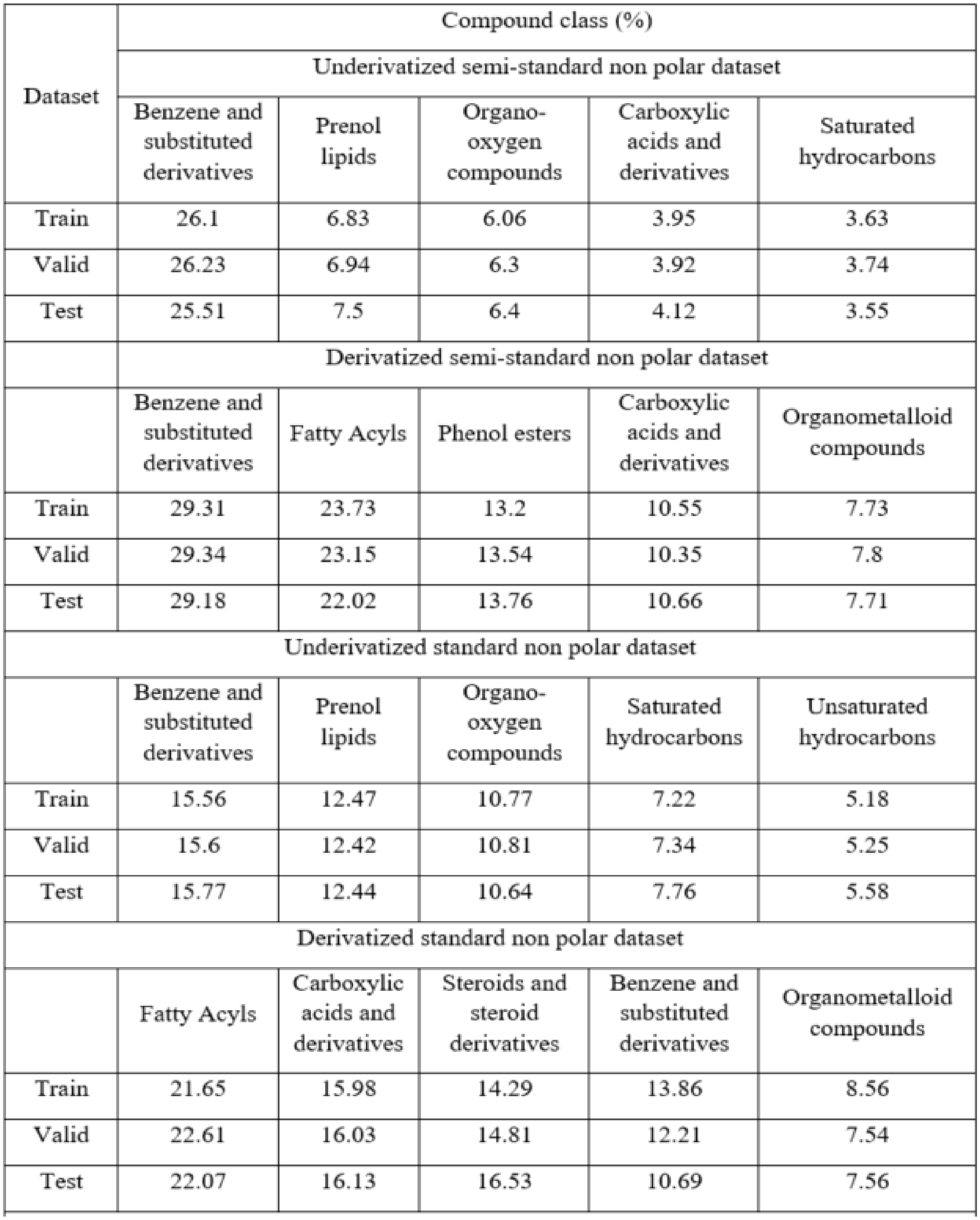
Table 2.3: Number of data points used for training, validating and evaluating the underivatized and derivatized compounds with RIpred.



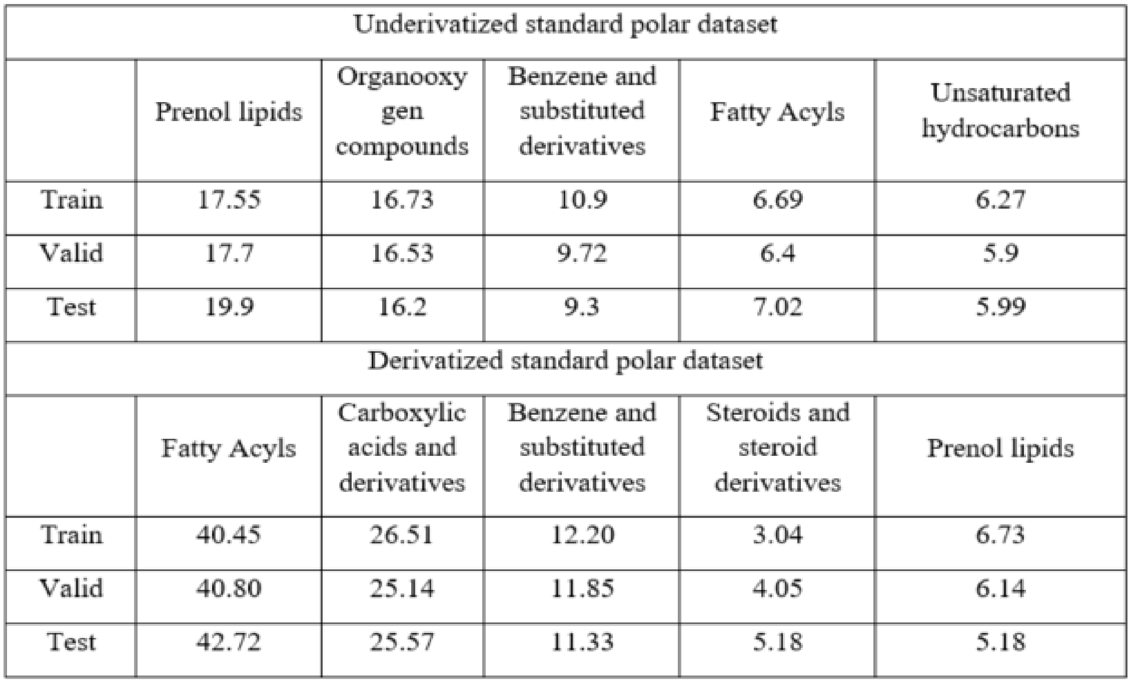
Dataset	Element (%)								
	Br	C	Cl	F	I	N	O	S	Si
Underivatized semi-standard non polar dataset									
Train	0.71	79.24	1.77	2.25	0.11	5.7	8.3	1.12	0
Valid	0.67	76.96	1.83	2.31	0.11	5.74	8.54	1.83	0
Test	0.69	79.62	1.62	2.33	0.10	5.62	8.21	1.02	0
Derivatized semi-standard non polar dataset									
Train	0.16	77.65	1.22	2.93	0.02	1.08	14.07	1.4	1.3
Valid	0.13	74.21	1.24	2.89	0.03	1.13	14.37	1.07	1.39
Test	0.15	77.6	1.21	3.09	0.01	1.06	14.1	1.35	1.27
Underivatized standard non polar dataset									
Train	0.25	85.5	1.8	0.72	0.05	3.42	6.66	1.16	0
Valid	0.29	83.5	1.66	0.67	0.07	3.46	6.58	1.04	0
Test	0.24	85.83	1.65	0.6	0.06	3.44	6.6	1.1	0
Derivatized standard non polar dataset									
Train	0.14	74.83	1.01	1.94	0.02	1.62	13.1	3.7	3.5
Valid	0.12	78.19	1.00	1.97	0.04	1.53	13.26	0.12	3.7
Test	0.1	75.3	0.75	1.26	0.02	1.45	13.03	4.01	3.9
Underivatized standard polar dataset									
Train	0.12	87.2	0.96	0.18	0.05	2.28	7.7	1.3	0
Valid	0.12	87.22	0.93	0.15	0.01	2.4	7.7	1.33	0
Test	0.2	87.9	0.9	0.03	0.09	1.86	7.45	1.08	0
Derivatized standard polar dataset									
Train	0.14	77.93	2.72	0.90	0	0.39	14.5	1.32	1.1
Valid	0.14	79.21	2.18	0.92	0	0.37	14.7	0.31	1.77
Test	0.08	78.2	2.54	0.57	0	0.53	14.26	1.45	1.2

Abbreviations defined here. Br=Bromine, C=Carbon, Cl=Chlorine, F=Flourine, I=Iodine, N=Nitrogen, O=Oxygen, S=Sulphur and Si=Silicon.

Table 2.4: Percentage distribution of molecules by ClassyFire class for the RIpred training, validation and hold-out sets.

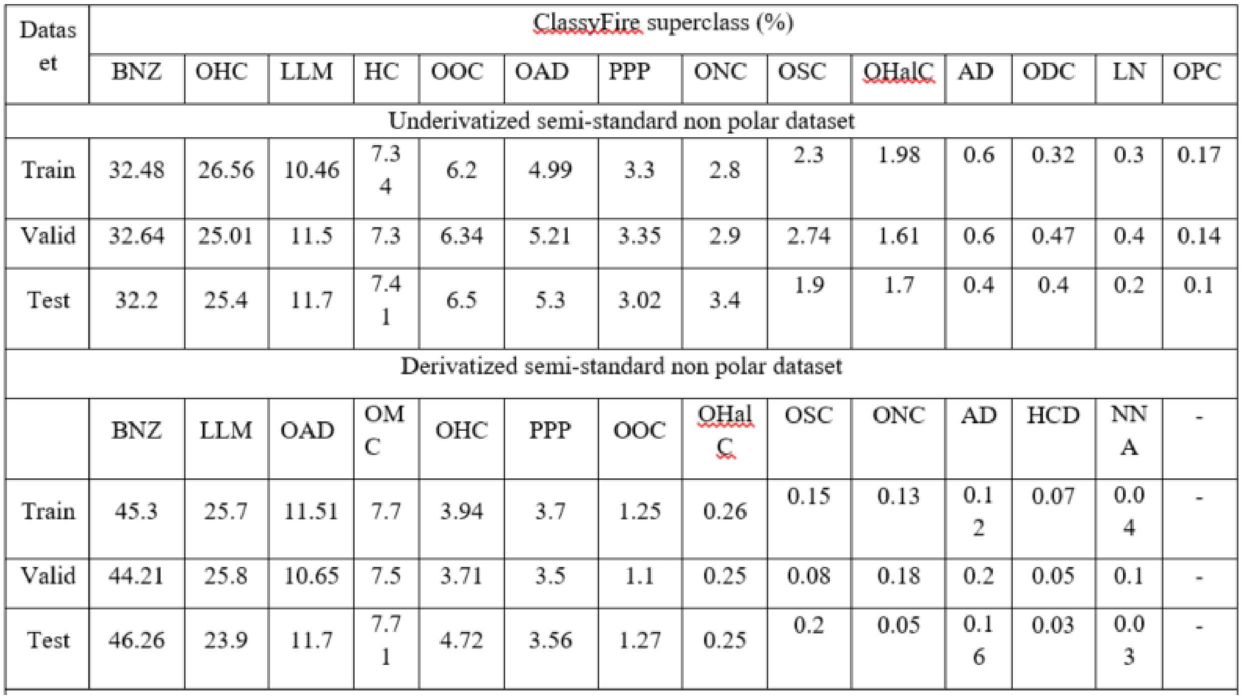


Dataset	Compound class (%)				
	Underivatized semi-standard non polar dataset				
	Benzene and substituted derivatives	Prenol lipids	Organo-oxygen compounds	Carboxylic acids and derivatives	Saturated hydrocarbons
Train	26.1	6.83	6.06	3.95	3.63
Valid	26.23	6.94	6.3	3.92	3.74
Test	25.51	7.5	6.4	4.12	3.55
	Derivatized semi-standard non polar dataset				
	Benzene and substituted derivatives	Fatty Acyls	Phenol esters	Carboxylic acids and derivatives	Organometalloid compounds
Train	29.31	23.73	13.2	10.55	7.73
Valid	29.34	23.15	13.54	10.35	7.8
Test	29.18	22.02	13.76	10.66	7.71
Underivatized standard non polar dataset					
	Benzene and substituted derivatives	Prenol lipids	Organo-oxygen compounds	Saturated hydrocarbons	Unsaturated hydrocarbons
Train	15.56	12.47	10.77	7.22	5.18
Valid	15.6	12.42	10.81	7.34	5.25
Test	15.77	12.44	10.64	7.76	5.58
Derivatized standard non polar dataset					
	Fatty Acyls	Carboxylic acids and derivatives	Steroids and steroid derivatives	Benzene and substituted derivatives	Organometalloid compounds
Train	21.65	15.98	14.29	13.86	8.56
Valid	22.61	16.03	14.81	12.21	7.54
Test	22.07	16.13	16.53	10.69	7.56

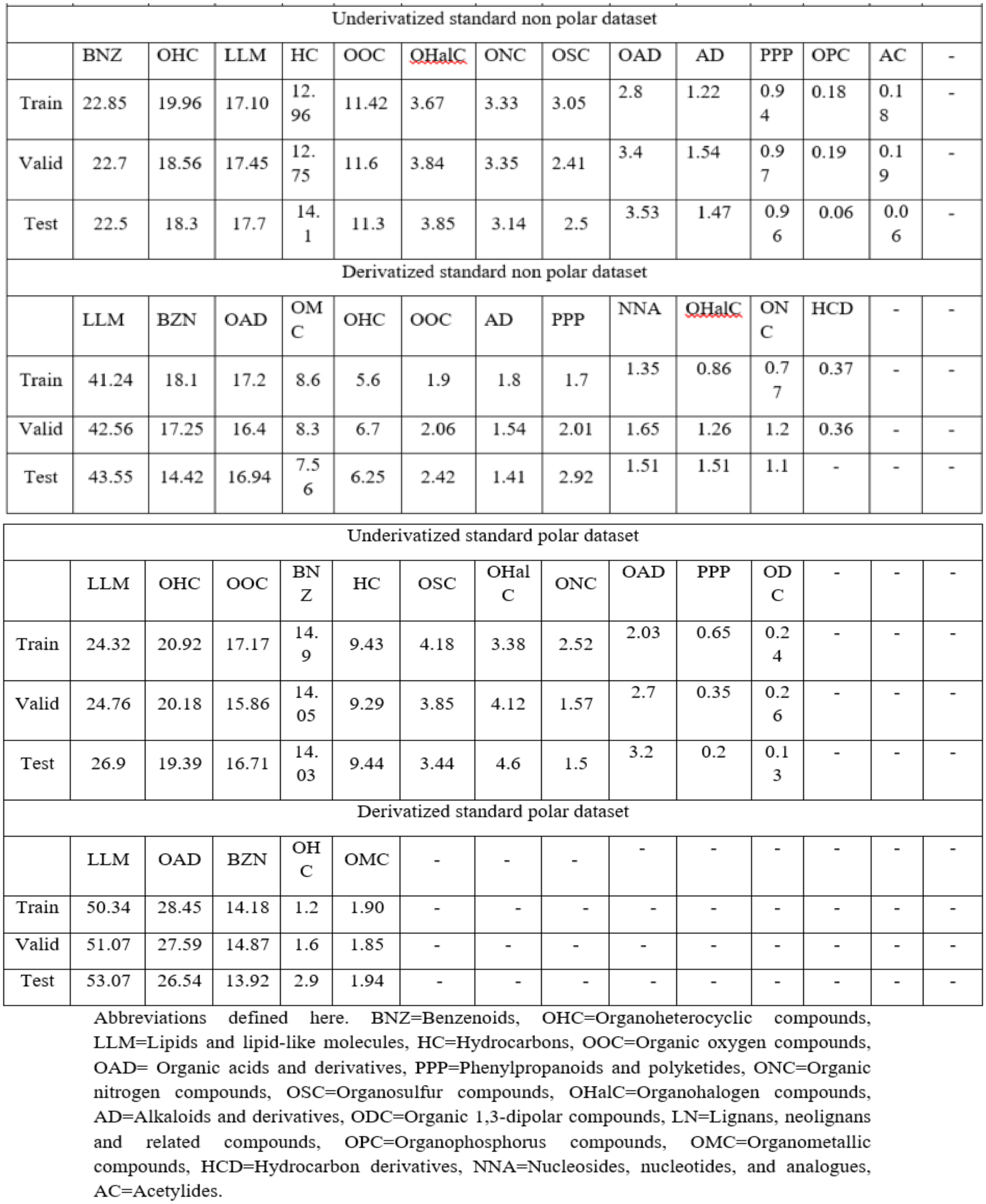


Underivatized standard polar dataset					
	Prenol lipids	Organooxygen compounds	Benzene and substituted derivatives	Fatty Acyls	Unsaturated hydrocarbons
Train	17.55	16.73	10.9	6.69	6.27
Valid	17.7	16.53	9.72	6.4	5.9
Test	19.9	16.2	9.3	7.02	5.99
Derivatized standard polar dataset					
	Fatty Acyls	Carboxylic acids and derivatives	Benzene and substituted derivatives	Steroids and steroid derivatives	Prenol lipids
Train	40.45	26.51	12.20	3.04	6.73
Valid	40.80	25.14	11.85	4.05	6.14
Test	42.72	25.57	11.33	5.18	5.18

Table 2.5: Percentage distribution of molecules by ClassyFire superclass for the RIpred training, validation and hold out sets.



Dataset	ClassyFire superclass (%)													
	BNZ	OHC	LLM	HC	OOC	OAD	PPP	ONC	OSC	OHalC	AD	ODC	LN	OPC
Underivatized semi-standard non polar dataset														
Train	32.48	26.56	10.46	7.34	6.2	4.99	3.3	2.8	2.3	1.98	0.6	0.32	0.3	0.17
Valid	32.64	25.01	11.5	7.3	6.34	5.21	3.35	2.9	2.74	1.61	0.6	0.47	0.4	0.14
Test	32.2	25.4	11.7	7.41	6.5	5.3	3.02	3.4	1.9	1.7	0.4	0.4	0.2	0.1
Derivatized semi-standard non polar dataset														
	BNZ	LLM	OAD	OMC	OHC	PPP	OOC	OHalC	OSC	ONC	AD	HCD	NNA	-
Train	45.3	25.7	11.51	7.7	3.94	3.7	1.25	0.26	0.15	0.13	0.12	0.07	0.04	-
Valid	44.21	25.8	10.65	7.5	3.71	3.5	1.1	0.25	0.08	0.18	0.2	0.05	0.1	-
Test	46.26	23.9	11.7	7.71	4.72	3.56	1.27	0.25	0.2	0.05	0.16	0.03	0.03	-



Underivatized standard non polar dataset

	BNZ	OHC	LLM	HC	OOC	OHalC	ONC	OSC	OAD	AD	PPP	OPC	AC	-
Train	22.85	19.96	17.10	12.96	11.42	3.67	3.33	3.05	2.8	1.22	0.94	0.18	0.18	-
Valid	22.7	18.56	17.45	12.75	11.6	3.84	3.35	2.41	3.4	1.54	0.97	0.19	0.19	-
Test	22.5	18.3	17.7	14.1	11.3	3.85	3.14	2.5	3.53	1.47	0.96	0.06	0.06	-

Derivatized standard non polar dataset

	LLM	BZN	OAD	OMC	OHC	OOC	AD	PPP	NNA	OHalC	ONC	HCD	-	-
Train	41.24	18.1	17.2	8.6	5.6	1.9	1.8	1.7	1.35	0.86	0.77	0.37	-	-
Valid	42.56	17.25	16.4	8.3	6.7	2.06	1.54	2.01	1.65	1.26	1.2	0.36	-	-
Test	43.55	14.42	16.94	7.56	6.25	2.42	1.41	2.92	1.51	1.51	1.1	-	-	-

Underivatized standard polar dataset

	LLM	OHC	OOC	BNZ	HC	OSC	OHalC	ONC	OAD	PPP	ODC	-	-	-
Train	24.32	20.92	17.17	14.9	9.43	4.18	3.38	2.52	2.03	0.65	0.24	-	-	-
Valid	24.76	20.18	15.86	14.05	9.29	3.85	4.12	1.57	2.7	0.35	0.26	-	-	-
Test	26.9	19.39	16.71	14.03	9.44	3.44	4.6	1.5	3.2	0.2	0.13	-	-	-

Derivatized standard polar dataset

	LLM	OAD	BZN	OHC	OMC	-	-	-	-	-	-	-	-	-
Train	50.34	28.45	14.18	1.2	1.90	-	-	-	-	-	-	-	-	-
Valid	51.07	27.59	14.87	1.6	1.85	-	-	-	-	-	-	-	-	-
Test	53.07	26.54	13.92	2.9	1.94	-	-	-	-	-	-	-	-	-

Abbreviations defined here. BNZ=Benzenoids, OHC=Organoheterocyclic compounds, LLM=Lipids and lipid-like molecules, HC=Hydrocarbons, OOC=Organic oxygen compounds, OAD= Organic acids and derivatives, PPP=Phenylpropanoids and polyketides, ONC=Organic nitrogen compounds, OSC=Organosulfur compounds, OHalC=Organohalogen compounds, AD=Alkaloids and derivatives, ODC=Organic 1,3-dipolar compounds, LN=Lignans, neolignans and related compounds, OPC=Organophosphorus compounds, OMC=Organometallic compounds, HCD=Hydrocarbon derivatives, NNA=Nucleosides, nucleotides, and analogues, AC=Acetylides.

Table 2.6: Percentage distribution of molecules by ClassyFire superclass for the RIpred training, validation and hold out sets.

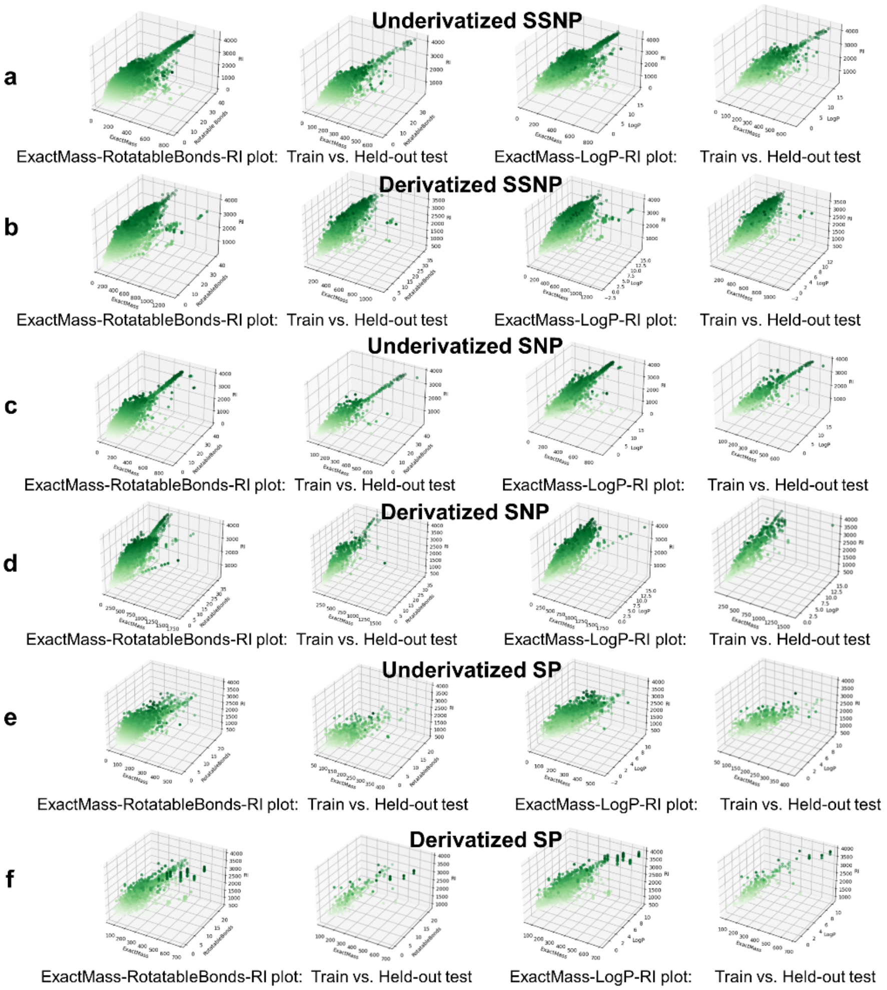


Figure 2.2: Molecular distribution for datasets by exact mass-rotatable bonds versus retention indices and exact mass-LogP versus retention indices for training and hold-out tests of a, c, e) underivatized and b, d, f) derivatized semi-standard non-polar (SSNP) datasets (a, b), standard non-polar (SNP) datasets (c, d) and standard polar (SP) datasets (e, f).

that we considered included TMS, TBDMS, triethylsilyl (TES), tert-butyldiphenylsilyl (TBDPS), esters, heptafluorobutyric acid anhydride (HFBA), methoxime (MO), pentafluoropropionic acid anhydride (PFPA), vinyldimethylsilyl (VDMS), and 4,4-dimethyloxazoline (DMOX). The distribution of derivatized compounds within a randomly selected subset of the combined NIST 17

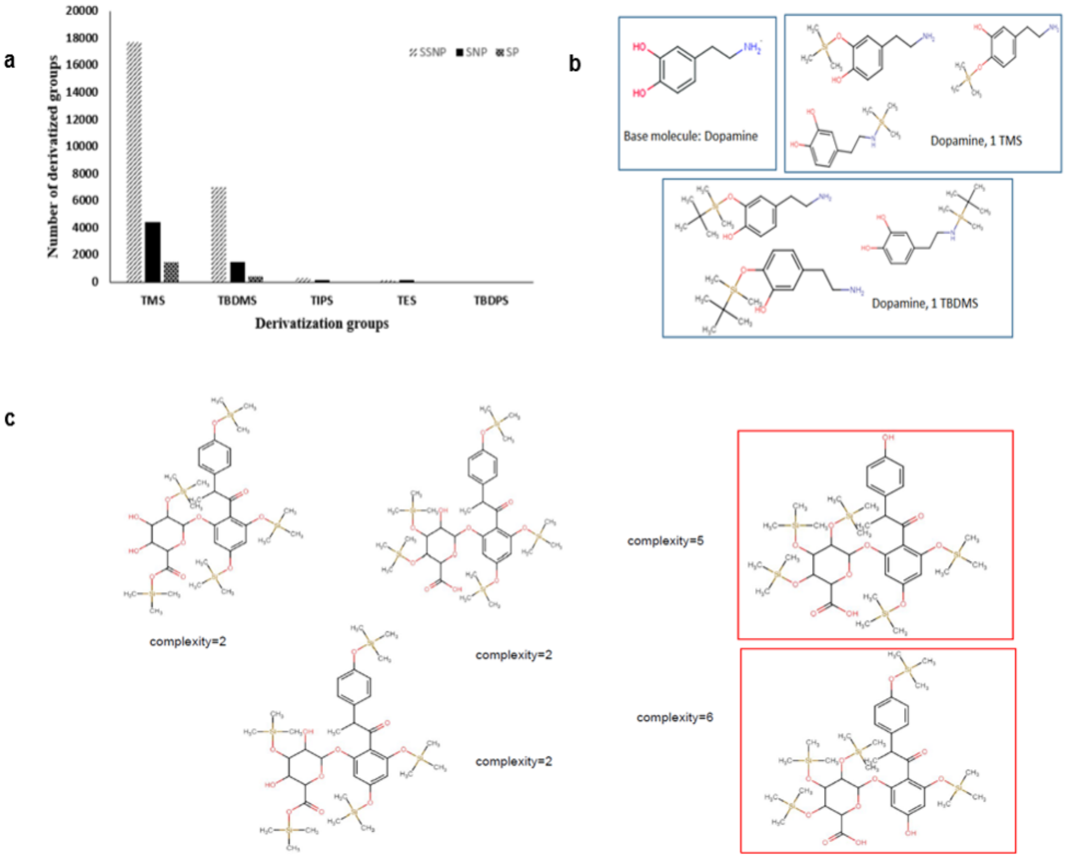


Figure 2.3: Computer-generated (in silico) derivatized GC-amenable compounds. a) Distribution of derivatized products found in the NIST 17 and NIST 20 RID datasets, indicating that TMS and TBDMS are the most commonly used derivatives in the SSNP, SNP and SP subgroups. b) Structures of underivatized dopamine and dopamine derivatized with 1 TMS and 1 TBDMS using the AUTOSILATOR script. c) Five TMS derivatized structures (isomers of each other) of a compound having three aromatic rings. These isomers have different complexity scores because of differences in their spatial arrangements. Compounds with complexity scores >5 (inside red boxes) were discarded.

and NIST 20 data is shown in Figure 2.3a. As seen from this figure, the vast majority (>95%) of compounds within these datasets were derivatized with TMS and TBDMS. As a result, we generated computer (*in silico*) derivatized GC-amenable compounds using only TMS and TBDMS and only used these two types of derivatives to train and test our RIpred models. Therefore, the RIpred web server only supports RI prediction for TMS and/or TBDMS derivatives.

2.2.4 Development of the RIpred Server

The RIpred webserver serves as the front-end for performing or running the six predictive RIpred models described above. Users must provide a valid chemical structure, select the GC stationary phase type (SSNP, SNP or SP) and select the derivatization type (No derivatization, TMS, TBDMS or a combination of TMS and TBDMS). Users can paste a single SMILES string or draw a single chemical structure into the JChem applet window (from ChemAxon) [100]. The average prediction takes 2.5 milliseconds (ms). The RIpred server uses standardized web frameworks and caching systems developed in our lab to make the website more user friendly and responsive. In particular, the frontend of RIpred has been implemented as a RESTful web service using the JRuby on Rails framework. Ruby on Rails is a web development system that employs a concept called a Model-View-Controller (MVC). In the MVC framework, models respond and interact with the data, views create the interface to show and interact with the data, and controllers connect the user to the views. This framework has allowed us to rapidly develop, prototype and test all of RIpred's web modules and page views. Many of the utilities are borrowed from a large collection of Ruby gems previously developed for the HMDB [101]. This framework is particularly robust and code can be reused in different functions or changed easily to accommodate future feature expansion or abrupt changes in design. This allowed us to liberally borrow code and functions from other webservers developed in our lab [102, 103, 104, 105].

2.2.5 Generation of Derivatized Structures for the RIpred Server

To make the RIpred webserver as user-friendly as possible, we developed a separate software package to automatically generate derivatized structures for RI prediction. With this package, users only need to provide the SMILES string or structure of the underivatized (base) compound and the derivatization reagent (limited to TMS and TBDMS). To generate structures for all possible derivatized products for each query compound, the computational derivatization script (called AUTOSILATOR) appends the TMS and/or TBDMS functional group (based on user input 'derivatization type') in chemically appropriate positions. The script uses individual derivatization rules for silylating compounds having functional groups such as acids, thiols, ketones, aldehydes, amines, etc. The script also automatically generates derivatized structure names that are formatted as "base_compound_name, *n* TMS/TBDMS", where *n* represents the total number of TMS or TBDMS groups attached to the base molecule.

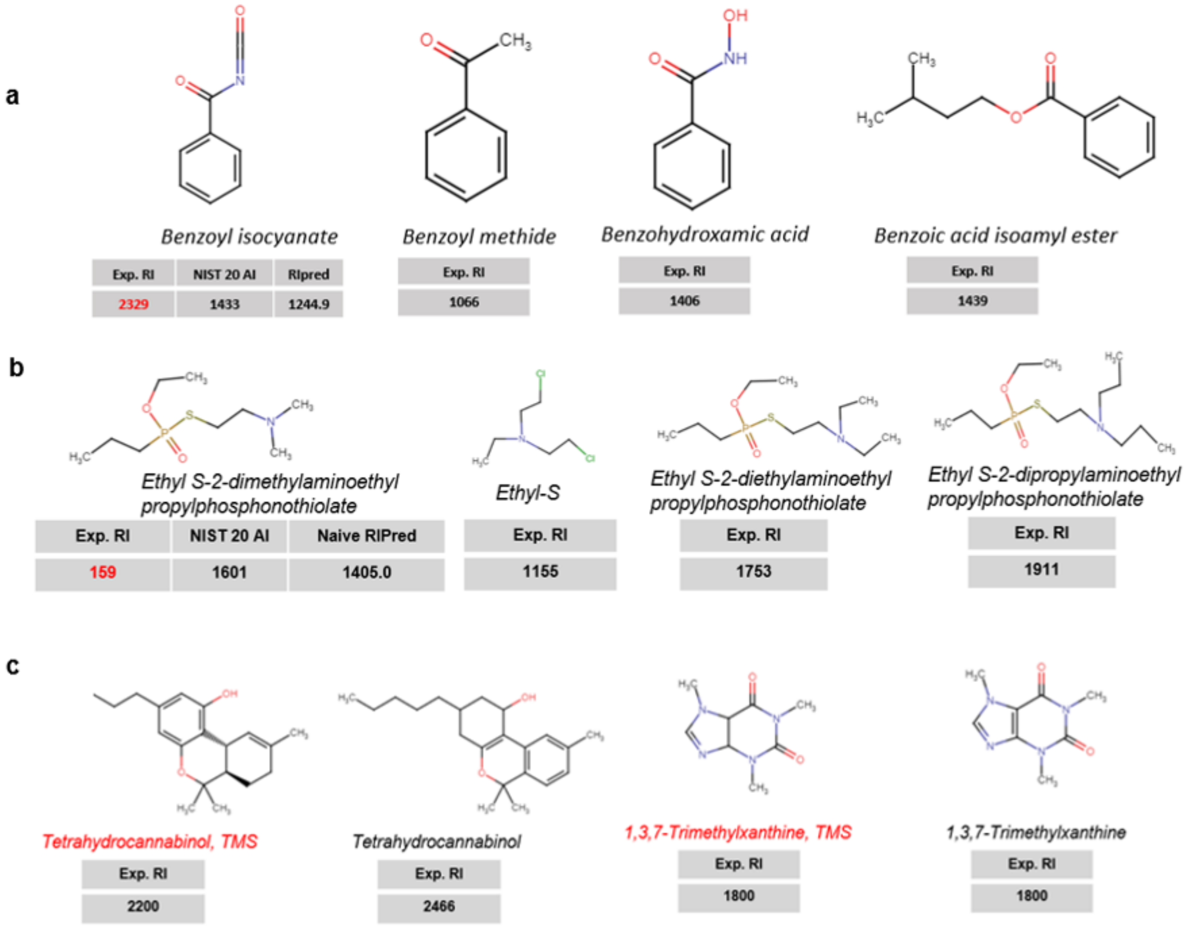
To ensure chemical viability, two filtering steps are used to remove any incorrect or offending structures. First, the software evaluates the molecular weights (MW) of the generated compounds and only keeps those with MW <900 Da. This is the maximum MW typically measurable by most commercial GC-MS instruments. Second, all compounds are passed through ChemBL [106] to assess the validity and feasibility of the computationally generated derivative structures. The ChemBL program is a bond/stereochemical evaluation program that is able to automatically identify issues with chemical structures such as mol-InChI stereo mismatches or improper placement of atoms and functional groups in invalid positions. ChemBL uses this information to assign a complexity value (0 to 9, 0 being the score for no issue and 9 being the score with lots of issues) for each structure. For our program, any generated compounds with a ChemBL complexity score of greater than 5 are discarded.

An example of how this filtering step is performed is shown in Figure 2.3b. Here the base molecule (dopamine) undergoes a silylation reaction with *n* TMS and TBDMS groups (the addition of *n*=1 is shown). With an increasing number of *n*, our computational derivatization script would add *n* TMS or TBDMS groups to each available functional group. This addition can give rise to the generation of some derivatives which would be chemically unfeasible and invalid. We used the complexity score generated by the ChemBL program to accept or discard those compounds. One example is shown in Figure 2.3c, where the 5 TMS isomers of a compound having three aromatic rings are seen to have different complexity scores because of differences in their spatial arrangements. The generated structures inside the red block would be discarded due to their high complexity scores (>5) while the others would proceed to RIpred's Kováts RI prediction. All valid, computational TMS and TBDMS derivatives and their RI predicted results will be displayed in the web server as per the user's request.

2.3 Results and discussion

2.3.1 Mismatched Experimental RIs in the NIST 17 and 20 Datasets Discovered with RIpred-alpha

During the training process, we discovered that a number of compounds in the NIST 17 and NIST 20 datasets had questionable experimental RI values. To mitigate the speculation, we compared the predicted RI values generated using an initial version of our model (RIpred-alpha) with the experimental values collected from the NIST databases. Our error analysis from this predicted versus experimental RI plot helped us identify a number of significant outliers. Each of these outliers was further analyzed (manually) to make sure that they were not false leads but rather true outliers. In particular, we manually investigated the reported RI values of each of these suspected compounds in the same stationary phase with their related compounds by looking at the progressive


a
Benzoyl isocyanate
Benzoyl methide
Benzohydroxamic acid
Benzoic acid isoamyl ester
b
Ethyl S-2-dimethylaminoethyl propylphosphonothiolate
Ethyl-S
Ethyl S-2-diethylaminoethyl propylphosphonothiolate
Ethyl S-2-dipropylaminoethyl propylphosphonothiolate
c
Tetrahydrocannabinol, TMS
Tetrahydrocannabinol
1,3,7-Trimethylxanthine, TMS
1,3,7-Trimethylxanthine

Exp. RI	NIST 20 AI	RIpred
2329	1433	1244.9

Exp. RI
1066

Exp. RI
1406

Exp. RI
1439

Exp. RI	NIST 20 AI	Naive RIPred
159	1601	1405.0

Exp. RI
1155

Exp. RI
1753

Exp. RI
1911

Exp. RI
2200

Exp. RI
2466

Exp. RI
1800

Exp. RI
1800

Figure 2.4: Questionable retention indices (RI) and questionable nomenclature of compounds within the NIST 17 and NIST 20 datasets. a) Structures of benzoyl isocyanate, b) ethyl S-2-dimethylaminoethyl propylphosphonothiolate and related compounds showing experimental RIs and those predicted using the Qu et al. predictor and our naïve RIpred model. c) Misnamed tetrahydrocannabinol and misnamed 1,3,7-trimethylxanthine and their TMS derivatives with similar experimental RIs.

RI trend. For example, the predicted RI for benzoyl isocyanate generated by RIpred-alpha was 1244.99 (Figure 2.4a) which was closer to the experimental RIs of many of its related compounds:

benzoyl methide - 1066, benzohydroxamic acid -1406 or benzoic acid isoamyl ester-1439 in the same SSNP phase. However, the experimental RI in the SSNP phase for benzoyl isocyanate was reported by NIST to be much higher, 2329. Similarly, a much smaller experimental RI was also reported for ethyl S-2-dimethylaminoethyl propylphosphonothiolate with respect to its related compounds (Figure 2.4b) in the same phase.

Looking only at the progressive RI trend, it was not always possible to detect outliers and thus in many cases, we also compared the predicted RI values derived by the RIpred-alpha in the SSNP phase and the Qu et al. RI predictor [79]. Many suspect compounds had slight differences (<5%) in their RIs predicted by RIpred-alpha and the Qu et al. predictor but exhibited large differences in experimental values reported in the NIST databases. For example, the predicted RIs for oxirane-carboxaldehyde were quite low (~630) and differed by just 3.6% for the Qu et al. predictor (620) and RIpred-alpha (643.27) (Figure 2.5a). However, the NIST reported experimental RI for oxirane-carboxaldehyde was much higher, 1339. Likewise, the experimental RI values for related compounds in the same stationary phase were 2 to 3-times smaller than the NIST listed experimental value for oxirane-carboxaldehyde. Indeed, the experimental RIs for the related compounds were closer to the predicted RIs (RIpred-alpha and the Qu et al. predictor): oxirane - 404, methyloxirane - 435, ethyloxirane - 571, chlormethyloxirane – 725, oxiranecarboxylic acid, 3-methyl, 3-phenyl-, ethyl ester - 1529 or 3-methyl-3-(4-methyl-3-pentenyl)oxiranecarboxaldehyde - 1234. Another example is the much larger experimental RI

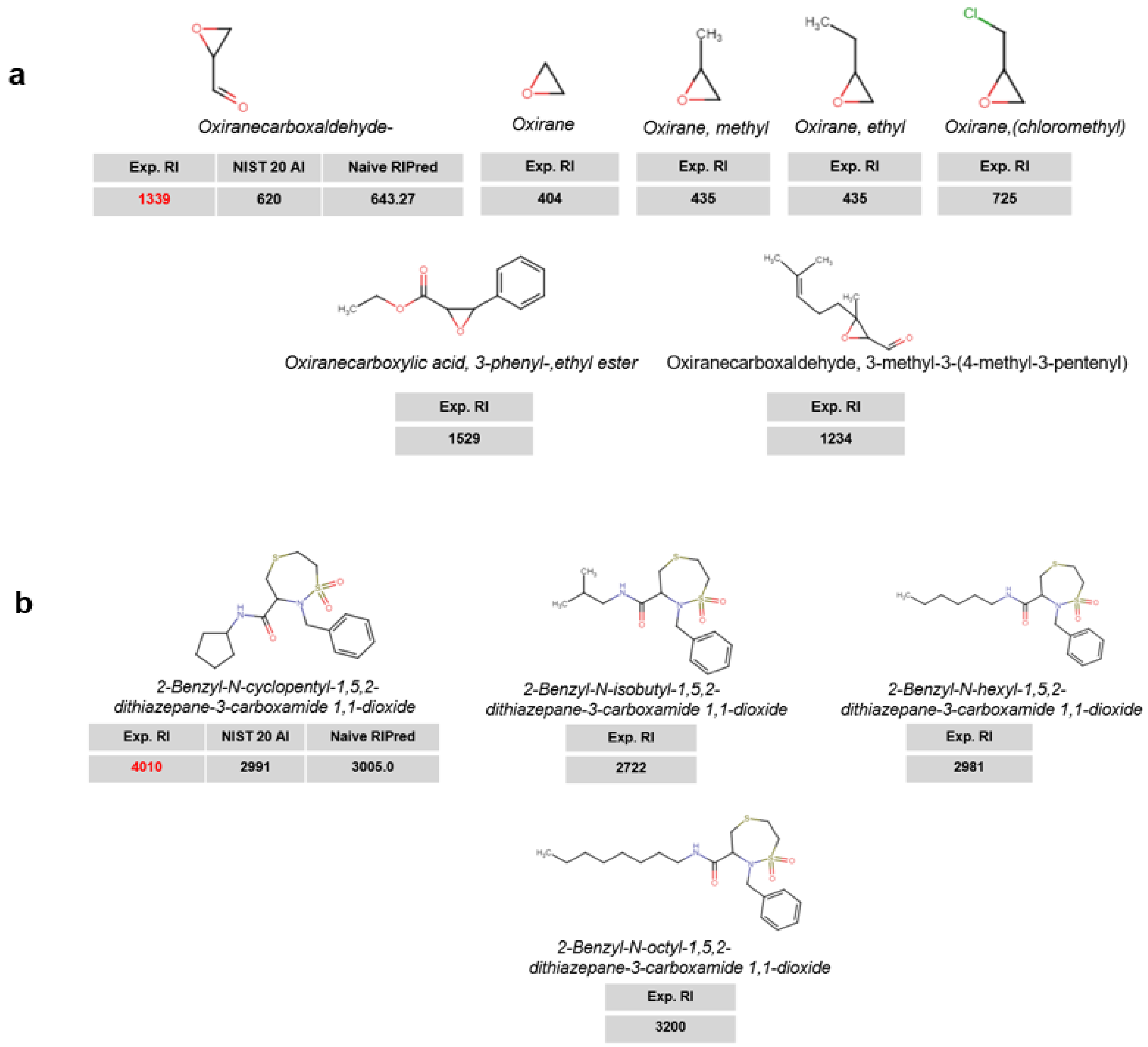

a
Oxiranecarboxaldehyde-
Exp. RI | NIST 20 AI | Naive RIPred
1339 | 620 | 643.27
Oxirane
Exp. RI
404
Oxirane, methyl
Exp. RI
435
Oxirane, ethyl
Exp. RI
435
Oxirane,(chloromethyl)
Exp. RI
725
Oxiranecarboxylic acid, 3-phenyl-,ethyl ester
Exp. RI
1529
Oxiranecarboxaldehyde, 3-methyl-3-(4-methyl-3-pentenyl)
Exp. RI
1234
b
2-Benzyl-N-cyclopentyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide
Exp. RI | NIST 20 AI | Naive RIPred
4010 | 2991 | 3005.0
2-Benzyl-N-isobutyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide
Exp. RI
2722
2-Benzyl-N-hexyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide
Exp. RI
2981
2-Benzyl-N-octyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide
Exp. RI
3200

Figure 2.5: Compounds with mismatched experimental (Exp.) retention indices (RI). a) Suspected compound oxirane-carboxaldehyde and compounds with similar structures in the semi-standard non-polar phase. b) Suspected compound 2-benzyl-N-cyclopentyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide and compounds with similar structures in the semi-standard non-polar phase.

being reported for 2-benzyl-N-cyclopentyl-1,5,2-dithiazepane-3-carboxamide 1,1-dioxide relative to its related compounds (Figure 2.5b) in the SSNP phase.

As noted earlier, we combined the NIST 17 with the NIST 20 datasets together and removed any entries containing duplicate compounds. As a result, ~26,000 compounds were newly added from the NIST 20. Prior to training the final models for RIpred, we explored the intersection between these two sets in an effort to identify mismatched SSNP experimental RI values for any compounds. An example with mismatched experimental RIs compared to our predicted RIs and those from the Qu et al. predictor is 1-hexyn-3-ol,3,5-dimethyl (Figure 2.6a). All such mismatched experimental RI values were removed from our consolidated dataset. However, it is important to note that these compounds (with their incorrect RI values) still exist in the NIST 20 RI library, as well as the NIST official website, and the PubChem official website.

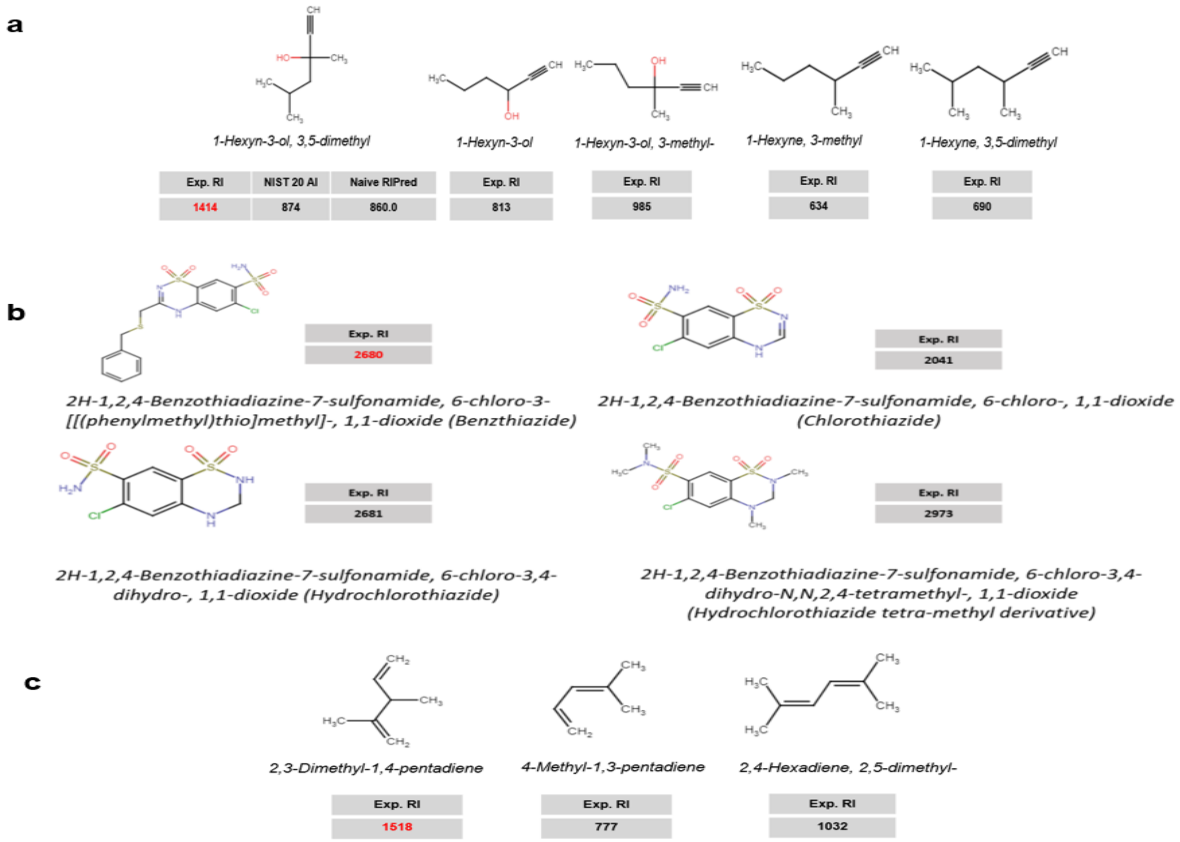

a
1-Hexyn-3-ol, 3,5-dimethyl
Exp. RI
NIST 20 AI
Naive RIPred
1414
874
860.0
1-Hexyn-3-ol
Exp. RI
813
1-Hexyn-3-ol, 3-methyl-
Exp. RI
985
1-Hexyne, 3-methyl
Exp. RI
634
1-Hexyne, 3,5-dimethyl
Exp. RI
690
b
Exp. RI
2680
2H-1,2,4-Benzothiadiazine-7-sulfonamide, 6-chloro-3-[[(phenylmethyl)thio]methyl]-, 1,1-dioxide (Benzthiazide)
Exp. RI
2041
2H-1,2,4-Benzothiadiazine-7-sulfonamide, 6-chloro-, 1,1-dioxide (Chlorothiazide)
Exp. RI
2681
2H-1,2,4-Benzothiadiazine-7-sulfonamide, 6-chloro-3,4-dihydro-, 1,1-dioxide (Hydrochlorothiazide)
Exp. RI
2973
2H-1,2,4-Benzothiadiazine-7-sulfonamide, 6-chloro-3,4-dihydro-N,N,2,4-tetramethyl-, 1,1-dioxide (Hydrochlorothiazide tetra-methyl derivative)
c
2,3-Dimethyl-1,4-pentadiene
Exp. RI
1518
4-Methyl-1,3-pentadiene
Exp. RI
777
2,4-Hexadiene, 2,5-dimethyl-
Exp. RI
1032

Figure 2.6: Compounds with mismatched experimental (Exp.) retention indices (RI) in the three different phases. a) Suspected compound- 1-hexyn-3-ol,3,5-dimethyl and compounds with similar structures in the semi-standard non-polar phase. b) Suspected compound benzthiazide and compounds with similar structures in the standard non-polar phase. c) Suspected compound 2,3-dimethyl-1,4-pentadiene and compounds with similar structures in standard polar phase.

As the Qu et al. predictions are not available for SNP and SP stationary phases, it proved to be somewhat more difficult to discover (and correct) any mismatched experimental RIs for these stationary phase types. In the end, we manually discovered mismatched experimental RIs by comparing the reported RI values of their isomers or related substructures. In some cases, our early RI model (RIpred-alpha) for the SNP and SP phases helped to identify candidate compounds that could potentially be outliers. A few examples of compounds with mismatched RI values for SNP and SP phases include benzthiazide and 2,3-dimethyl-1,4-pentadiene (Figure 2.6b and c, respectively). Follow-up assessments with published data or comparisons to the data for structurally related variants were used to determine the source of the error and to perform any necessary corrections.

We also identified several compounds in the NIST 20 RI library with discrepancies in their names and/or reported structure. For example, compounds such as “tetrahydrocannabinol, TMS” and “caffeine, TMS” (or 1,3,7-trimethylxanthine, TMS) (Figure 2.4c), appeared to be incorrectly named as they had the same RI as the underivatized base compounds (tetrahydrocannabinol and caffeine). As indicated, the names suggest that Si (in the form of TMS) should be present in the structures. However, Si was absent from the reported structures as well as the reported formula within the library.

Overall, we manually corrected erroneous SMILES strings, structures and Kováts RI values whenever a clear justification was possible. If there was contradictory evidence, we completely discarded these compounds from our training sets. Approximately 1000 compounds were identified and manually investigated during this data “cleaning” process. This led to the removal of ~850 compounds from all RI datasets and all individual training/testing folds. Thus, extensive data cleaning was necessary to assemble a reliable RI set for training and testing.

2.3.2 Training, Validating, and Testing RIpred Models

As mentioned above, during the training process, we implemented an early stopping module in the GNN learning scripts. As a result, our model design and performance was able to quickly reach an optimization point where no further improvement could be made (Figure 2.7). The best models

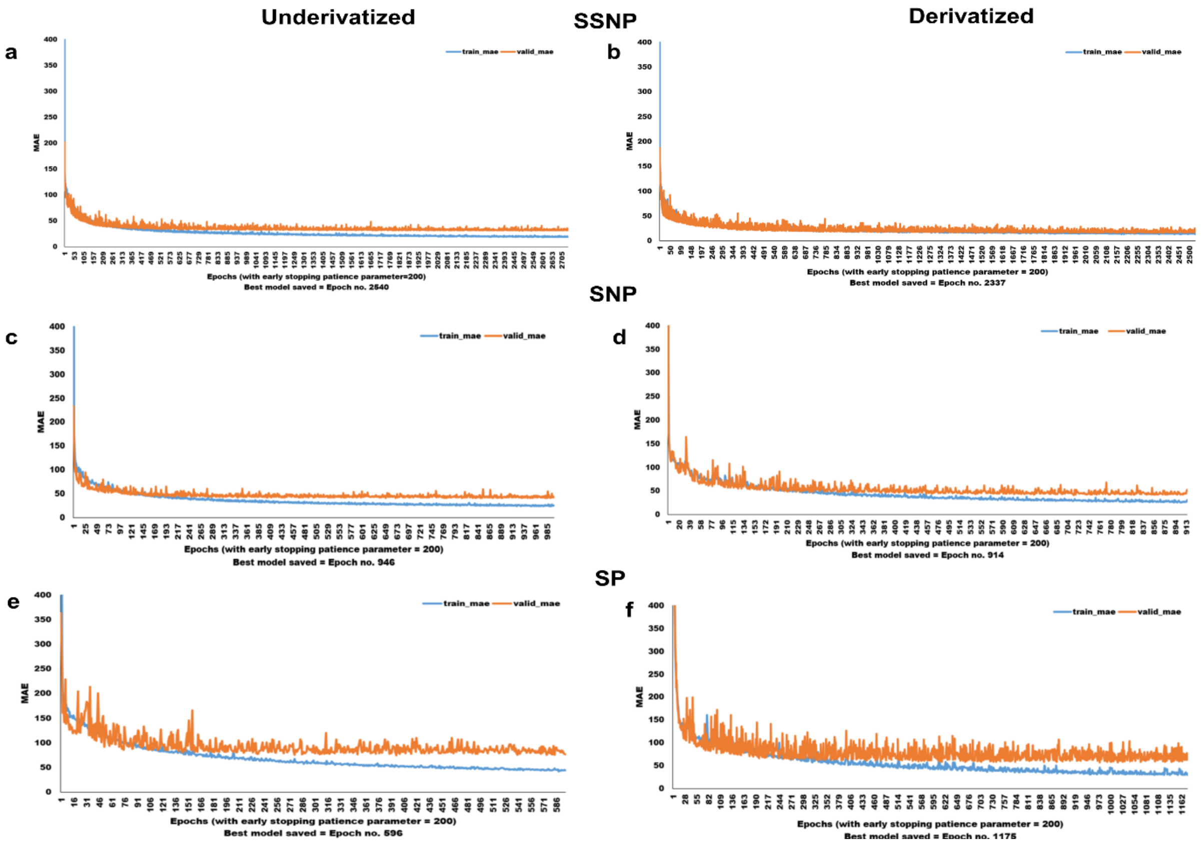


Figure 2.7: Plot of the change in the mean absolute error versus the training epoch of the underivatized and derivatized datasets. a, c, e) The underivatized and b, d, f) derivatized plots of the semi-standard non-polar (SSNP- a, b), the standard non-polar (SNP- c, d) and the standard polar (SP- e, f). The best models were saved after the MAE stabilized for each given training dataset. All models used an early stopping patience parameter of 200 epochs.

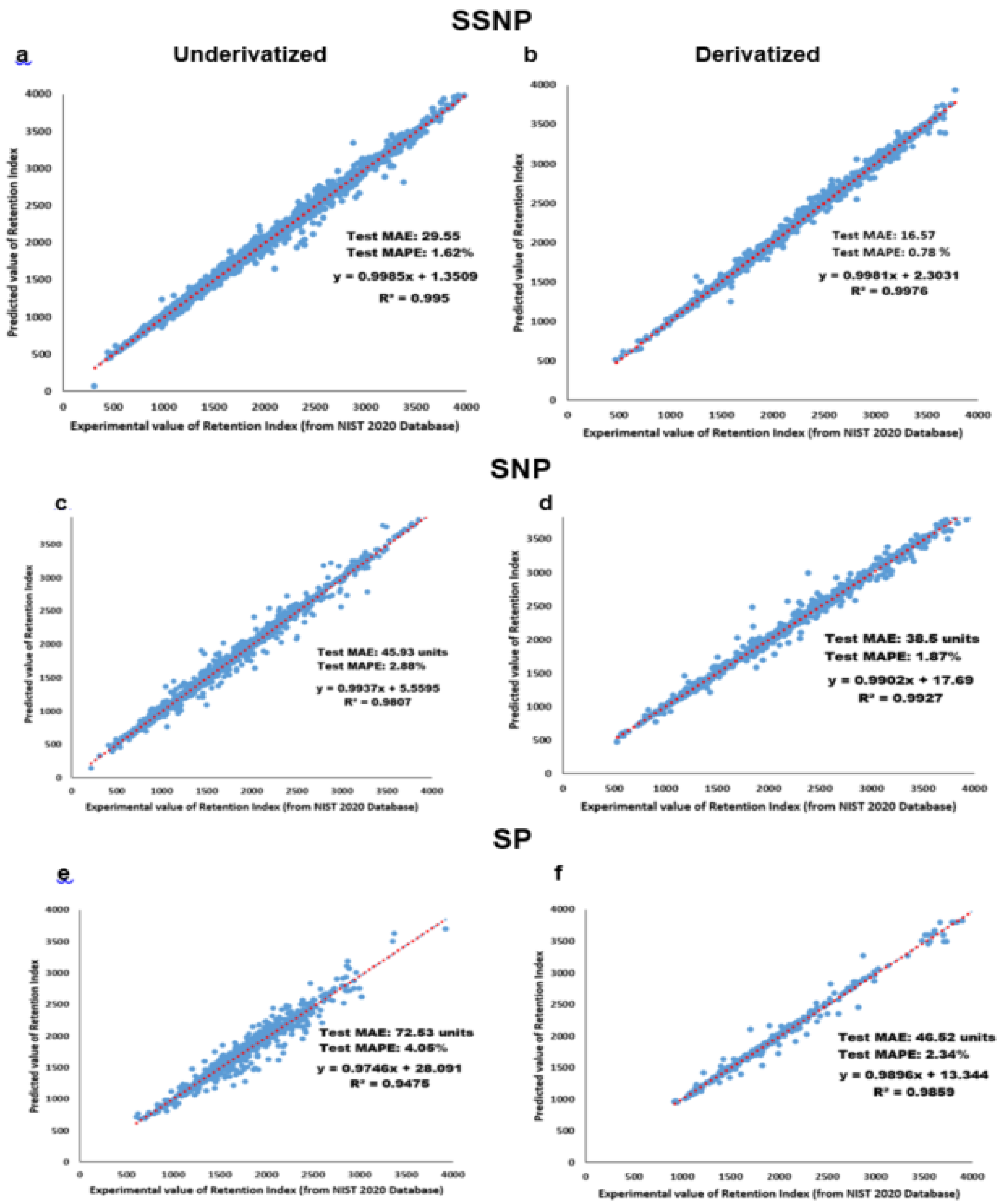
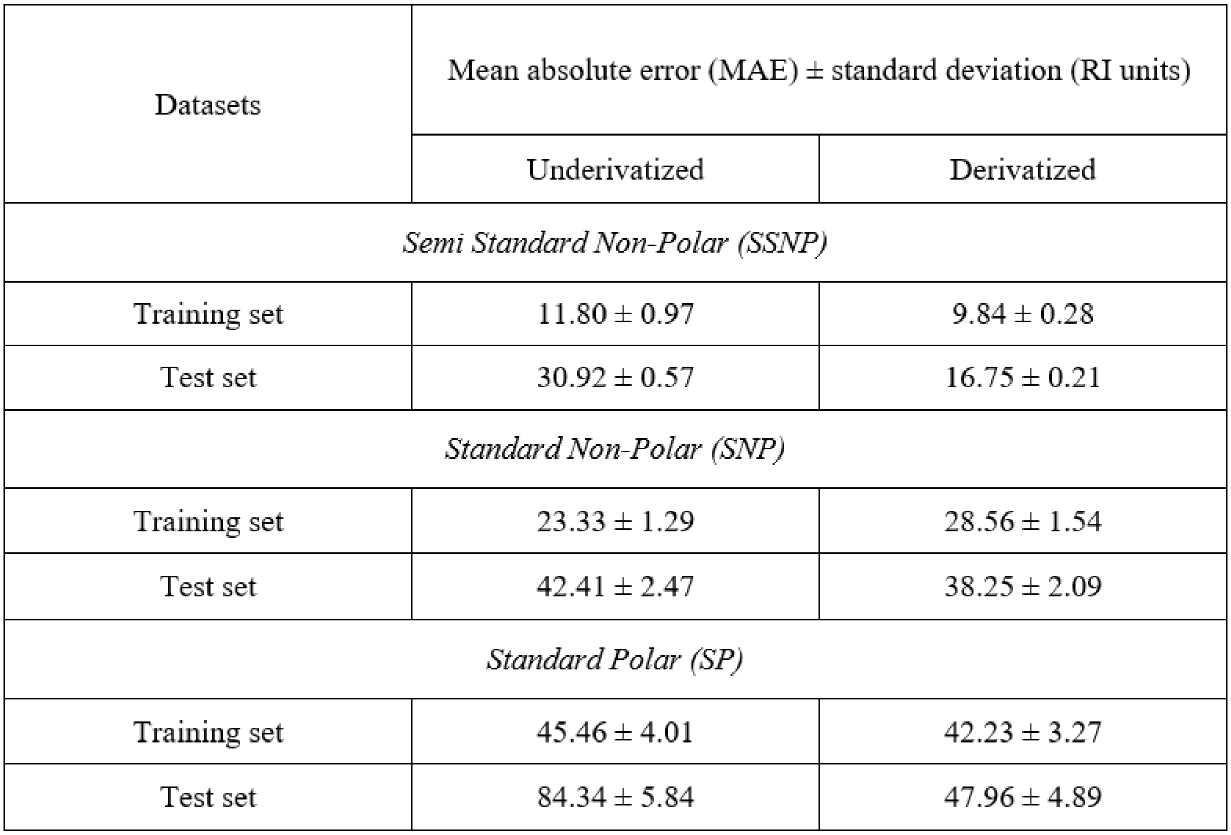


Figure 2. 8: Comparison of experimental versus model-predicted Kováts RIs from hold-out test sets. a, c, e) show the underivatized while b, d, f) show derivatized compounds in the semi-standard non polar phase (SSNP), standard non polar phase (SNP) and standard polar phase (SP) test sets, respectively. The total number of molecules assessed in the SSNP, SNP and SP phases are 3619, 993 and 310, respectively.

were typically generated after higher numbers of epochs for the two SSNP datasets (2540 epochs for underivatized and 2330 for the derivatized dataset) compared to much fewer epochs for the SP-underivatized (596) and the SP-derivatized (1175) and both SNP datasets (946 epochs for underivatized and 914 for the derivatized dataset). These epoch differences were likely due to the different sizes of the training sets.

After the training runs were completed, we used the best-trained models generated from each stationary phase (including both underivatized and derivatized splits) to evaluate the respective held-out test sets. All the MAEs and MAPEs from this evaluation are reported in Figure 2.8. The plots in Figure 2.8 suggest that all the MAPEs are within 3% except for RI's predicted for the underivatized standard polar phase. Similarly, all MAEs are within 73 RI units of the experimentally measured values. Moreover, the correlation indices (R^2) for all the predicted versus experimental points were all above 0.94 (SSNP (derivatized): R^2 = 0.9976; SSNP (underivatized): R^2 = 0.995; SNP (derivatized): R^2 = 0.9927, SNP (underivatized): R^2 = 0.9807; SP (derivatized): R^2 = 0.9859, SP (underivatized): R^2 = 0.9475).

The MAEs reported from the 10-fold cross validation process for both the underivatized and derivatized testing sets of the SSNP, the SNP and SP (Table 2.7) were within 50 RI units with respect to the best model trained reported in Figure 2.8. However, there was a slightly higher MAE reported for the SP dataset. This suggests that all six of our RI models are relatively unbiased and are expected to provide very comparable performance when used to make predictions for both

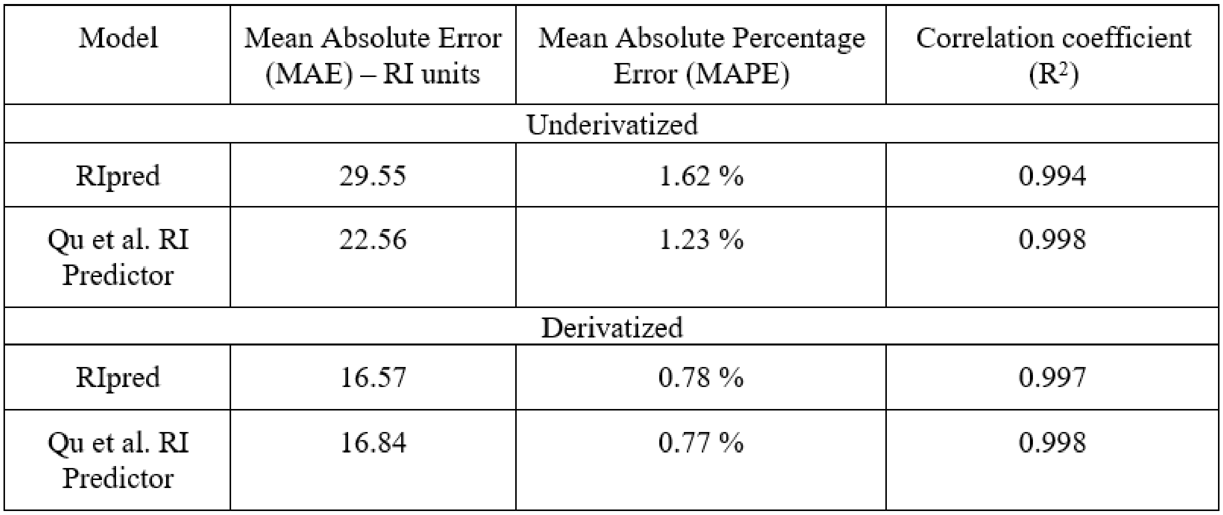


Datasets	Mean absolute error (MAE) ± standard deviation (RI units)	
	Underivatized	Derivatized
Semi Standard Non-Polar (SSNP)		
Training set	11.80 ± 0.97	9.84 ± 0.28
Test set	30.92 ± 0.57	16.75 ± 0.21
Standard Non-Polar (SNP)		
Training set	23.33 ± 1.29	28.56 ± 1.54
Test set	42.41 ± 2.47	38.25 ± 2.09
Standard Polar (SP)		
Training set	45.46 ± 4.01	42.23 ± 3.27
Test set	84.34 ± 5.84	47.96 ± 4.89

Table 2.7: Statistics showing the mean and standard deviation of RIpred's MAEs (in RI units) from the 10-fold cross validation of each of the six RI datasets.

previously seen and newly unseen data. The MAEs for all training sets were slightly lower than the MAEs reported for the held-out testing sets.

As seen in Figure 2.7, the training vs validation MAEs reported over the training period (measured in epochs) indicate a rapid decrease in this error metric. This indicates a potential trend towards overfitting. However, the mean and the standard deviation of the MAEs reported from the 10-fold cross-validation for the held-out test sets indicates we achieved robust model performance that resulted in almost no overfitting. Moreover, we made a direct comparison of our RIpred model with the Qu et al. RI predictor on two held-out test sets consisting of 3281 underivatized and 3619 derivatized molecules with RIs recorded in the SSNP phase. The performance reported by RIpred is almost identical to the Qu et al. RI predictor when the model is compared for the derivatized compounds in the SSNP phase (with an MAPE of <1% and an MAE of 16.57 units by RIpred vs MAPE of <1% and MAE of 16.84 by the Qu et al. RI predictor). However, the performance was slightly worse for underivatized compounds in the same phase (with an MAPE of 1.62% and an MAE of 29.55 units by RIpred vs MAPE of 1.23% and an MAE of 29.55 units by the Qu et al. RI predictor). The performance scores using the Qu et al. RI predictor are shown in Table 2.8. Since these models were trained on large datasets (containing tens of thousands of instances) for extended periods with varying epochs, it should be possible for these models to generalize the predictability very well. However, we observed an inconsistent performance with the two different datasets mentioned above. This unusual behavior can be explained by the absence of sufficient training instances in our underivatized SSNP dataset. For example, the predicted RI for a hold-out test set compound water (H_2O) was quite low (~71.35) for the RIpred predictor in the SSNP phase but was close to the experimental value (317) for the Qu et al. RI predictor (309) (Figure 2.9). Moreover,



Model	Mean Absolute Error (MAE) – RI units	Mean Absolute Percentage Error (MAPE)	Correlation coefficient (R^2)
Underivatized			
RIpred	29.55	1.62 %	0.994
Qu et al. RI Predictor	22.56	1.23 %	0.998
Derivatized			
RIpred	16.57	0.78 %	0.997
Qu et al. RI Predictor	16.84	0.77 %	0.998

Table 2.8: Comparison between Qu et al. RI predictor and our RIpred model on held-out test sets from underivatized and derivatized compounds in the SSNP phase. Performance is presented in terms of MAE (RI units), MAPE and R^2.

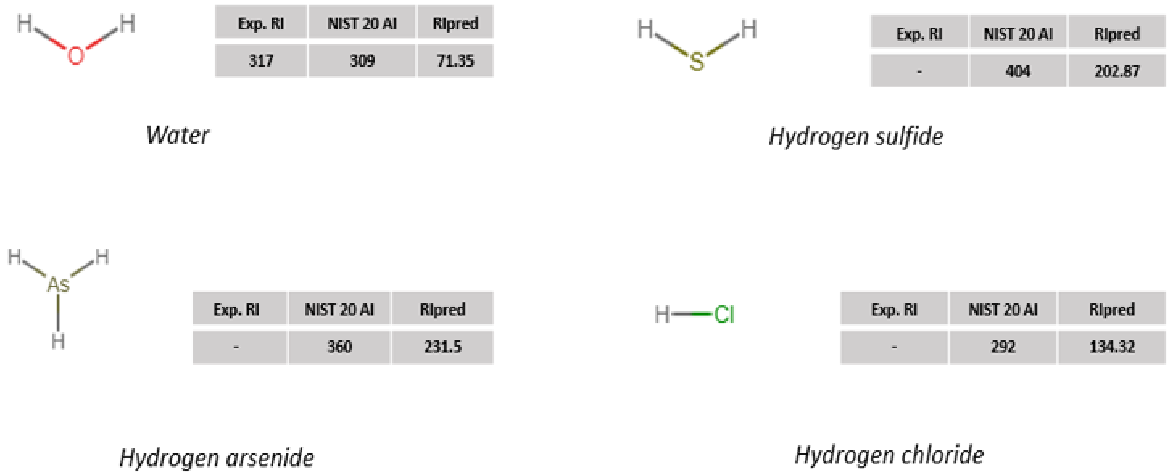


Figure 2.9: Compounds with deviation in predicted retention indices (RI) by the NIST 20 AI and RIpred in the SSNP phase for water (H2O) and compounds containing hydrogen bonded with other elements (H2S, AsH3, HCl).

the experimental RI values for compounds where hydrogen is bonded with other elements to form distinct chemical compounds, such as hydrogen chloride (HCl), hydrogen sulfide (H_2S), hydrogen arsenide (AsH_3) etc. are absent in the NIST 20 database. This explains the lack of such examples in our training set. Furthermore, practically all the NIST 20 compounds are tagged with the Qu et al. predicted RI values without any indication as to whether or not those compounds were included in their training set. If these compounds were used by Qu et al. in the training set, then their results would naturally be better. Additionally, our RIpred model performed similarly when compared with the 10-fold cross-validated test MAE produced by the Qu et al. RI predictor in the SSNP phase for underivatized compounds (RIpred MAE of 30.92 units (see Table 2.7) versus the NIST 20 RI predictor's MAE of 28.09 units, as reported in [79]).

We also compared the performance of RIpred with the DeepReI [53] and the Matyushin et al. [59] tools. DeepReI described a SSNP tool built with convolutional layers. However, despite repeated efforts to load or install the program we were unable to generate any prediction results. On the other hand, we were able to successfully install and run the SSNP, SNP and SP RI prediction models described by Matyushin et al. Our RIpred outperformed the Matyushin et al. SSNP model both for the derivatized and underivatized sets of compounds (RIpred MAE of 16.57-29.55 units (see Table 2.9) and MAPE of <2% versus the Matyushin et al. predictor's MAE of 20.74-54.09 units and MAPE of 0.95-2.7%). We found no noticeable difference between our RIpred model when it was compared against their SNP model. Interestingly, the averaged SP model of Matyushin et al. slightly outperformed RIpred both in the underivatized and derivatized sets of compounds. It is important to note that in performing this comparison, we had to remove many compounds from

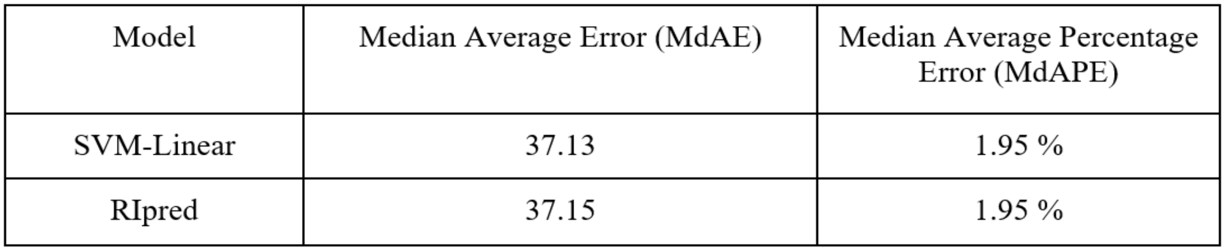
Model	Mean Absolute Error (MAE) – RI units			Mean Absolute Percentage Error (MAPE)			Correlation coefficient (R^2)		
Underivatized									
GC Column/Phase	SSNP	SNP	SP	SSNP	SNP	SP	SSNP	SNP	SP
RIpred	29.55	45.93	72.53	1.62 %	2.88%	4.05%	0.994	0.981	0.947
Matyushin et al., 2021	54.09	39.19	51.3	2.7 %	2.45%	2.87%	0.967	0.982	0.952
Derivatized									
GC Column/Phase	SSNP	SNP	SP	SSNP	SNP	SP	SSNP	SNP	SP
RIpred	16.57	38.5	46.52	0.78 %	1.87%	2.34%	0.9976	0.993	0.985
Matyushin et al., 2021	20.74	35.2	30.27	0.95 %	1.61%	1.53%	0.9968	0.993	0.993

Table 2.9: Comparison between Matyushin et al., 2021 predictor and our RIpred model on held-out test sets from underivatized and derivatized compounds in the SSNP, SNP and SP phases. Performance is presented in terms of MAE (RI units), MAPE and R2.

the hold-out test set as the Matyushin et al. model (unlike RIPred) could not handle compounds containing Pb, As, Sn, etc.

2.3.3 Validation on the GolmDB Dataset

The other validation test involved assessing the performance of RIpred against a large hold-out set of independently collected RI values and against an RI predictor specifically trained on these data. The Golm Metabolome database (GDB) is an open-source GC-MS metabolome reference library used for metabolite profiling of biologically active metabolites. In a very recent study [58], an RI predictor that used support vector machine (SVM) linear regression was developed for TMS-derivatized compounds from the GDB. The model was trained using 1410 (1159 instances after data cleaning) derivatized compounds. The study also reports the performance of the SVM predictor using a 10-fold-cross-validation method. To compare RIpred with this SVM predictor and the experimental RI data in GDB, we manually collected a subset of these derivatized compounds (917 in total) and their reported RI values (in the same stationary phase) from the GDB website [107]. The RI values in this website were tagged with the TMS compound names and InChIs. However, these InChIs represented only the base (underivatized) form of these metabolites. We used our computational derivatization script (called AUTOSILATOR) to generate appropriate



Model	Median Average Error (MdAE)	Median Average Percentage Error (MdAPE)
SVM-Linear	37.13	1.95 %
RIpred	37.15	1.95 %

Table 2.10: Comparison between SVM-Linear and RIpred in terms of model performance for a subset of TMS metabolites.

TMS compounds as well as PubChemPy to retrieve appropriate SMILES for all 917 compounds. As seen in Table 2.10, the SVM linear regression predictor and our RIpred predictor performed equivalently (with the SVM predictor having an R^2 of 0.954 and RIpred having an R^2 of 0.945). Additionally, we show a scatter plot of the RIpred predicted RI's versus the experimental RI's (see Figure 2.10) to compare between RIpred's values and the reported experimental values by GolmDB in terms of Median Average Error (MdAE) and Median Average Percentage Error (MdAPE). Again, the agreement is excellent.

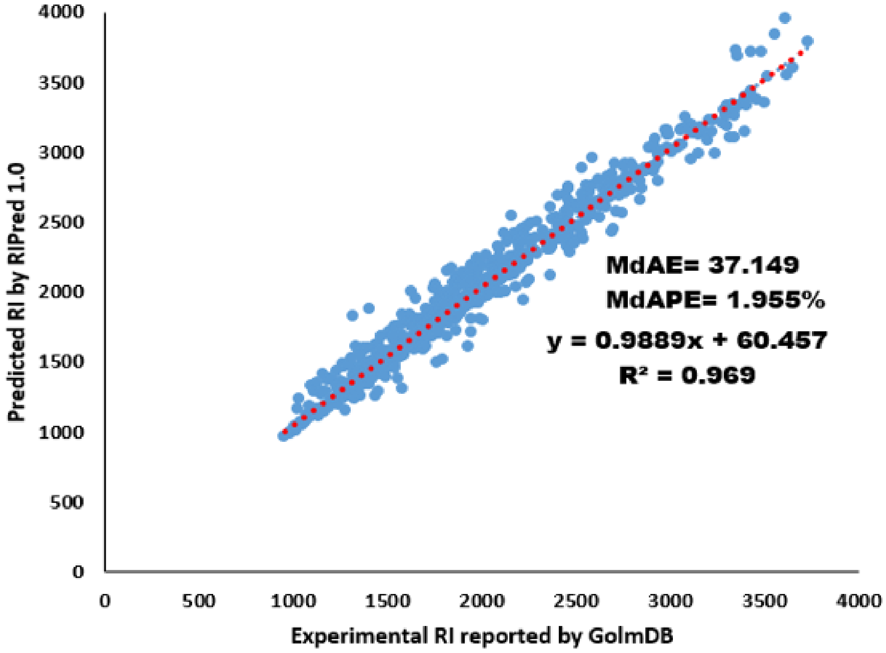


Figure 2.10: Comparison between predicted RI values by RIpred and reported experimental values by GolmDB in terms of Median Average Error (MdAE) and Median Average Percentage Error (MdAPE). (The error distribution for this comparison is shown in Appendix A).

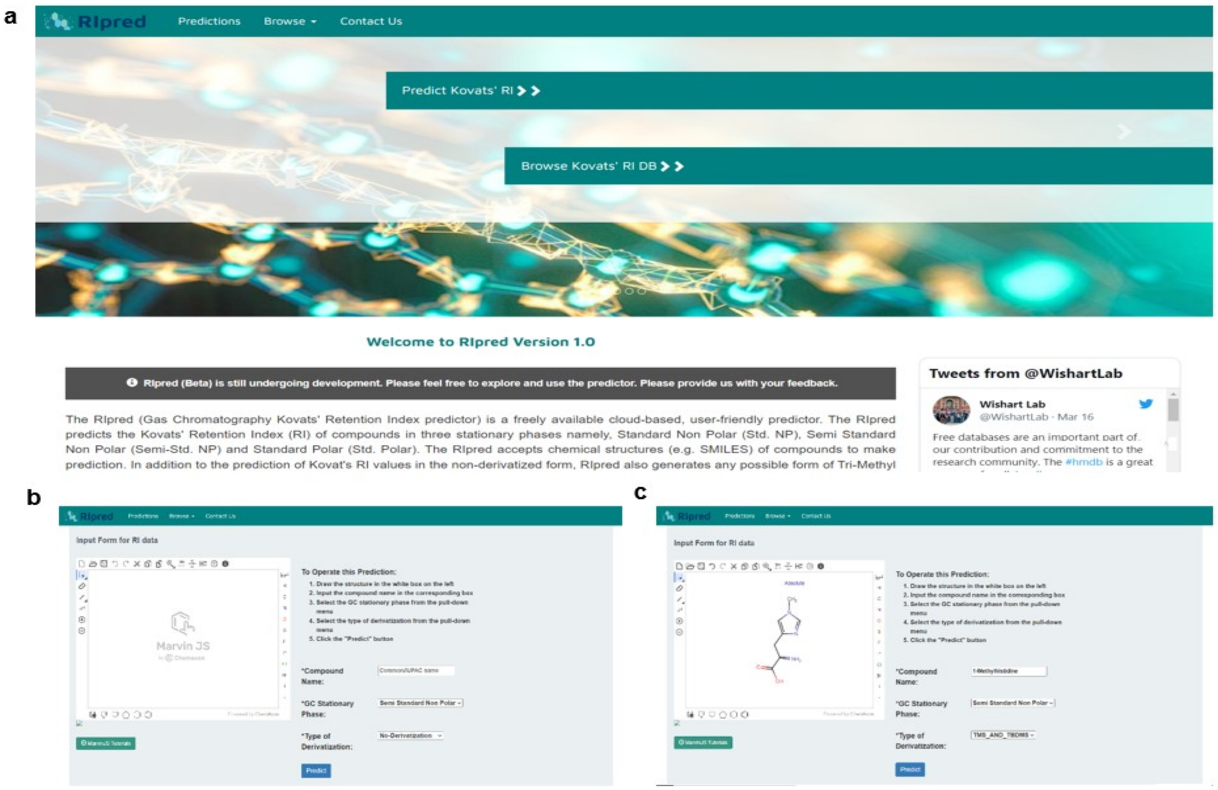
2.3.4 RIpred Prediction of Human Metabolome Database (HMDB) Compounds

Having confirmed through multiple independent tests that RIpred is robust and exhibits comparable or better accuracy for RI prediction as the best RI predictors for both underivatized and derivatized compounds, we decided to apply RIpred to predict the RIs of a large set of commonly measured or detected compounds in metabolomics studies. In particular, we chose to apply RIpred to all GC-MS "amenable" compounds within the latest release of the human metabolome database -- HMDB 5.0 [101]. The total number of GC-amenable compounds in HMDB 5.0 with MW <900 Da. was 57,648 (of 220,000 total compounds). We used AUTOSILATOR to generate TMS and TBDMS derivatives with varied numbers of silyl groups (e.g. 1 TMS, 2TMS, 1 TBDMS, 5 TBDMS, etc.) of these HMDB compounds. This generated a list of 1.5 million derivatized compounds with all of them having MW's <900 Da. All these GC-amenable derivatized compounds for the three different stationary phases (4,744,611) were passed through the RIpred derivatized predictor to generate Kováts RI values. All the above mentioned GC-amenable underivatized compounds (57,648) were also passed through RIpred's underivatized predictors for each of the three different stationary phases. The total number of RI predictions generated in this way by RIpred was 5,067,714. All these compounds were tagged with an HMDB identifier (base compound), given appropriate names (for the derivatized products), assigned their stationary phase information, and their 2D structures (MOL and SDF files) were generated. The resulting RI data was then placed in both the HMDB and the RIpred web server.

2.3.5 The RIpred Webserver

The RIpred (Version 1.0) webserver is available at https://ripred.ca. Two options are currently available on the site: 1) Kováts RI prediction of a single compound of interest or 2) browsing

RIpred’s library of RI predictions (containing 1.7 million predictions of roughly 200,000 compounds). To perform a prediction, a user can either draw their desired structure into the JChem viewer (a ChemAxon applet) or paste the SMILES string of their compound into the JChem viewer (which automatically generates the structure image). After entering the structure, the user must select the stationary phase (available through a pull-down menu) and the derivatization reaction (none, TMS, TBDMS, or a combination of both). All of these options are selectable from a pull-down menu. The predicted Kováts RI values for the desired compound is then shown in three separate windows, illustrating each type of requested derivatization. These predicted values are also presented as SMILES strings (original SMILES and SMILES generated using AUTOSILATOR) along with a button to generate and view their 3D structures using JSmol. All these prediction results (including compound names, SMILES strings, predicted RI values, GC column/stationary phases, derivatization types and the SDF files corresponding to the molecules) can be downloaded using the “Download” button located on the top of the prediction results. In addition to the RI prediction function, users can also browse RIpred’s collection of RI predictions for 100s of thousands of molecules. The Browse function displays a table of all compounds in the RIpred database with 10 different sortable columns (RIpred ID, compound name, number of TMS derivatives, number of TBDMS derivatives, molecular formula, molecular weight and the predicted RI for SSNP, SNP and SP phases). The Browse function also allows users to search compounds via text matches and to filter the displayed compounds by the number of TMS derivatives, the number of TBDMS derivatives, molecular weight range, SSNP RI range, SNP RI



Welcome to RIpred Version 1.0

RIpred (Beta) is still undergoing development. Please feel free to explore and use the predictor. Please provide us with your feedback.

The RIpred (Gas Chromatography Kovats' Retention Index predictor) is a freely available cloud-based, user-friendly predictor. The RIpred predicts the Kovats' Retention Index (RI) of compounds in three stationary phases namely, Standard Non Polar (Std. NP), Semi Standard Non Polar (Semi-Std. NP) and Standard Polar (Std. Polar). The RIpred accepts chemical structures (e.g. SMILES) of compounds to make prediction. In addition to the prediction of Kovat's RI values in the non-derivatized form, RIpred also generates any possible form of Tri-Methyl

Tweets from @WishartLab

Wishart Lab
@WishartLab · Mar 16

Free databases are an important part of our contribution and commitment to the research community. The #hmdb is a great

Input Form for RI data

To Operate this Prediction:

1. Draw the structure in the white box on the left
2. Input the compound name in the corresponding box
3. Select the GC stationary phase from the pull-down menu
4. Select the type of derivatization from the pull-down menu
5. Click the "Predict" button

*Compound Name: Common/IUPAC name

*GC Stationary Phase: Semi Standard Non Polar

*Type of Derivatization: No-Derivatization

Input Form for RI data

To Operate this Prediction:

1. Draw the structure in the white box on the left
2. Input the compound name in the corresponding box
3. Select the GC stationary phase from the pull-down menu
4. Select the type of derivatization from the pull-down menu
5. Click the "Predict" button

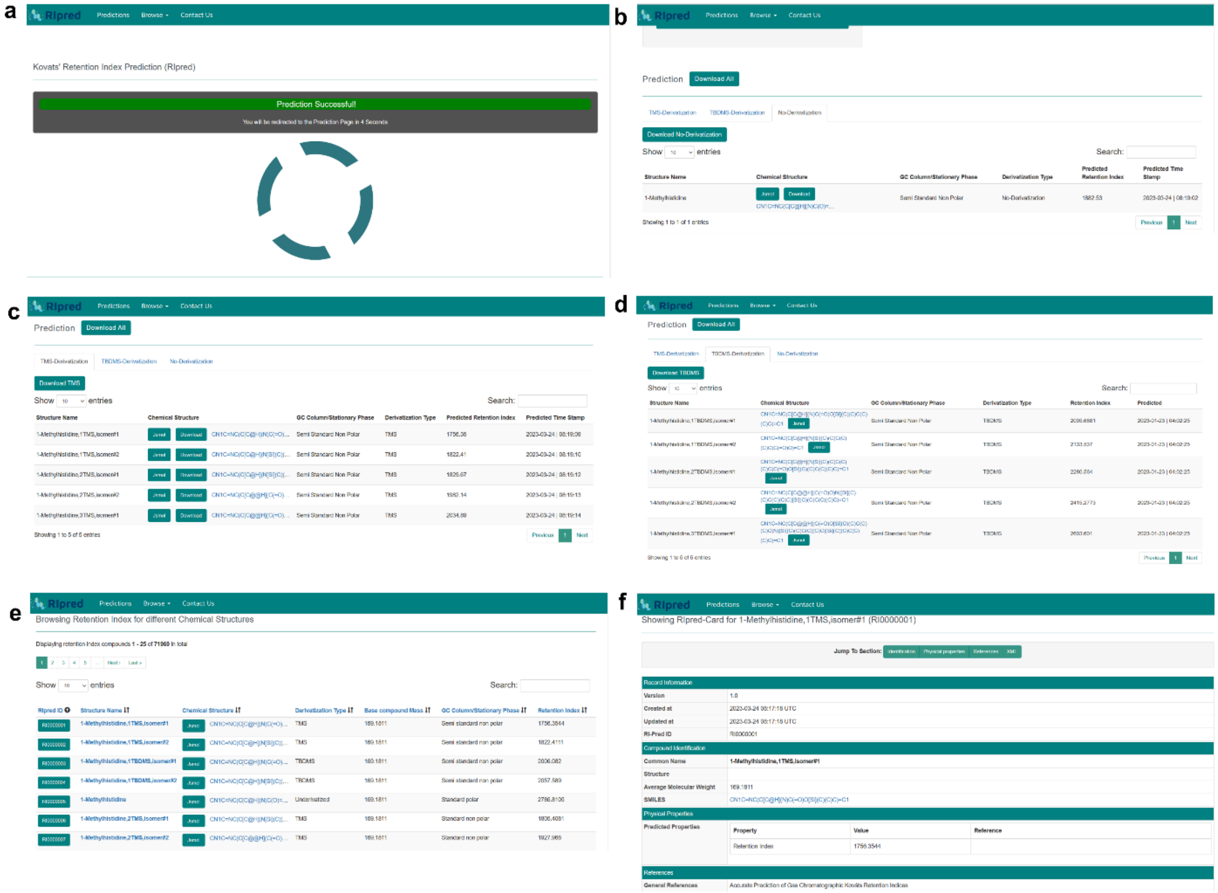
*Compound Name: 1-Methylhistidine

*GC Stationary Phase: Semi Standard Non Polar

*Type of Derivatization: TMS_AND_TBDMS

Figure 2.11: RIpred web server screenshots illustrating different server options. a) A screenshot of the RIpred web server homepage showing both the predict and browse options for users. b) and c) screenshots of the RI input forms, which provide users with the option to select one of the three stationary phases (or GC column) and the type of derivatization (No derivatization, TMS, TBDMS or a combination of TMS and TBDMS). In this figure, 1-methylhistidine is shown as an example (SMILES string pasted in the ChemAxon applet).

range and SP RI range. Once filtered, users can download the selected compounds (and related data) as a CSV (comma separated value) file. While the compounds presented in the current version of the database are mostly from the HMDB, it is expected that this browser will expand to include other classes of compounds (natural products, drugs, contaminants, flavor compounds, etc.) in the near future. The general functionalities of the webserver are depicted through Figure 2.11 and Figure 2.12.



Kovats' Retention Index Prediction (RIpred)

Prediction Successful!

You will be redirected to the Prediction Page in 4 Seconds

Prediction

Show 10 entries Search:

Structure Name	Chemical Structure	GC Column/Stationary Phase	Derivatization Type	Predicted Retention Index	Predicted Time Stamp
1-Methylhistidine	CN1C=NC(C[C@H](N)C(O)=...	Semi Standard Non Polar	No-Derivatization	1882.53	2023-03-24 | 08:19:02

Showing 1 to 1 of 1 entries

Structure Name	Chemical Structure	GC Column/Stationary Phase	Derivatization Type	Predicted Retention Index	Predicted Time Stamp
1-Methylhistidine,1TMS,isomer#1	CN1C=NC(C[C@H](N)C(=O)...	Semi Standard Non Polar	TMS	1756.36	2023-03-24 | 08:19:08
1-Methylhistidine,1TMS,isomer#2	CN1C=NC(C[C@H](N[Si](C)(...	Semi Standard Non Polar	TMS	1822.41	2023-03-24 | 08:19:10
1-Methylhistidine,2TMS,isomer#1	CN1C=NC(C[C@H](N[Si](C)(...	Semi Standard Non Polar	TMS	1829.67	2023-03-24 | 08:19:12
1-Methylhistidine,2TMS,isomer#2	CN1C=NC(C[C@@H](C(=O)...	Semi Standard Non Polar	TMS	1982.14	2023-03-24 | 08:19:13
1-Methylhistidine,3TMS,isomer#1	CN1C=NC(C[C@@H](C(=O)...	Semi Standard Non Polar	TMS	2034.88	2023-03-24 | 08:19:14

Showing 1 to 5 of 5 entries

Structure Name	Chemical Structure	GC Column/Stationary Phase	Derivatization Type	Retention Index	Predicted
1-Methylhistidine,1TBDMS,isomer#1	CN1C=NC(C[C@H](N)C(=O)O[Si](C)(C)C(C)(C)C)=C1	Semi Standard Non Polar	TBDMS	2006.6881	2023-01-23 | 04:02:25
1-Methylhistidine,1TBDMS,isomer#2	CN1C=NC(C[C@H](N[Si](C)(C)C(C)(C)C)C(=O)O)=C1	Semi Standard Non Polar	TBDMS	2133.537	2023-01-23 | 04:02:25
1-Methylhistidine,2TBDMS,isomer#1	CN1C=NC(C[C@H](N[Si](C)(C)C(C)(C)C)C(=O)O[Si](C)(C)C(C)(C)C)=C1	Semi Standard Non Polar	TBDMS	2290.664	2023-01-23 | 04:02:25
1-Methylhistidine,2TBDMS,isomer#2	CN1C=NC(C[C@@H](C(=O)O)N([Si](C)(C)C(C)(C)C)[Si](C)(C)C(C)(C)C)=C1	Semi Standard Non Polar	TBDMS	2416.2773	2023-01-23 | 04:02:25
1-Methylhistidine,3TBDMS,isomer#1	CN1C=NC(C[C@@H](C(=O)O[Si](C)(C)C(C)(C)C)N([Si](C)(C)C(C)(C)C)[Si](C)(C)C(C)(C)C)=C1	Semi Standard Non Polar	TBDMS	2603.601	2023-01-23 | 04:02:25

Showing 1 to 6 of 6 entries

Browsing Retention Index for different Chemical Structures

Displaying retention index compounds **1 - 25** of **71969** in total

1 2 3 4 5 … Next › Last »

RIpred ID	Structure Name	Chemical Structure	Derivatization Type	Base compound Mass	GC Column/Stationary Phase	Retention Index
RI0000001	1-Methylhistidine,1TMS,isomer#1	CN1C=NC(C[C@H](N)C(=O)...	TMS	169.1811	Semi standard non polar	1756.3544
RI0000002	1-Methylhistidine,1TMS,isomer#2	CN1C=NC(C[C@H](N[Si](C)(...	TMS	169.1811	Semi standard non polar	1822.4111
RI0000003	1-Methylhistidine,1TBDMS,isomer#1	CN1C=NC(C[C@H](N)C(=O)...	TBDMS	169.1811	Semi standard non polar	2006.082
RI0000004	1-Methylhistidine,1TBDMS,isomer#2	CN1C=NC(C[C@H](N[Si](C)(...	TBDMS	169.1811	Semi standard non polar	2057.589
RI0000005	1-Methylhistidine	CN1C=NC(C[C@H](N)C(O)=...	Underivatized	169.1811	Standard polar	2786.8106
RI0000006	1-Methylhistidine,2TMS,isomer#1	CN1C=NC(C[C@H](N[Si](C)(...	TMS	169.1811	Standard non polar	1806.4081
RI0000007	1-Methylhistidine,2TMS,isomer#2	CN1C=NC(C[C@@H](C(=O)...	TMS	169.1811	Standard non polar	1827.985

Showing RIpred-Card for 1-Methylhistidine,1TMS,isomer#1 (RI0000001)

Record Information	
Version	1.0
Created at	2023-03-24 08:17:18 UTC
Updated at	2023-03-24 08:17:18 UTC
RI-Pred ID	RI0000001
Compound Identification	
Common Name	**1-Methylhistidine,1TMS,isomer#1**
Structure	
Average Molecular Weight	169.1811
SMILES	CN1C=NC(C[C@H](N)C(=O)O[Si](C)(C)C)=C1
Physical Properties	
Predicted Properties	Property: Retention Index; Value: 1756.3544; Reference:
References	
General References	Accurate Prediction of Gas Chromatographic Kováts Retention Indices

Figure 2.12: Operational workflow of the RIpred webserver when a user enters 1-methylhistidine as an input. a) shows that the server has successfully completed the predictions. b, c, d) show the three output panels for three types of predicted Kovats' RI values. e, f) show the browse page and the individual RIpred-Card page for 1-methylhistidine.

2.4 Conclusion

We have described the development, training, testing and release of a webserver version of a publicly available tool (RIpred) for the prediction of Kováts RIs. RIpred is capable of accepting SMILES data as input and generating RIs for three common GC stationary phases (SSNP, SNP and SP) for both underivatized and TMS and TBDMS-derivatized compounds. The predictors embedded in the RIpred webserver were trained using GNN models and employ molecular and atomic features generated via RDKit. The training data consisted of NIST 17 and NIST 20 RI data that contained both SMILES and the experimentally measured RI values. We have found that RIpred performs as well, or nearly as well as some of the best existing RI predictors, including the Qu et al. RI predictor [79]. However, unlike other high-performing RI predictors, our RIpred tool is freely available. Its main advantages lie in its ability to separately predict RIs for compounds in SSNP, SNP and SP stationary phases and to automatically generate silylated (TMS and TBDMS) compounds for subsequent RI prediction.

To further enhance the RIpred's accuracy, there are several improvements that could be made. In particular, larger training datasets could certainly improve the accuracy of the SP and SNP predictions. Likewise, greater support for multiple compound submissions would certainly make the RIpred web server more broadly useful. In particular, we are currently developing an API that should enable users to process large batches of compounds and to generate all possible stationary forms and all valid GC-amenable derivatives. Similarly, expanding the number of compounds in RIpred's database and expanding the tools to search or select RI values or compounds from this database will also make the tool more generally useful. Finally, our intention is to integrate RIpred with a new and more accurate EI-MS spectral predictor via CFM-EI [65]. This integrated webserver would allow users to submit both experimental EI-MS data and RI data to identify potential candidate matches to the predicted RI and predicted EI-MS data.

Chapter 3: Prediction of Electron-Ionization Mass Spectra (EI-MS) by Leveraging Composite Graph Neural Networks (GNN) and a Support Vector Regressor (SVR)

3.1 Introduction

Gas chromatography mass spectrometry (GC-MS) is one of the most common analytical methods used to separate, identify and quantify small molecules in various mixtures and matrices [108]. In GC-MS, the chemical components of mixture are often chemically derivatized, then separated by a GC column, then the molecules are ionized and fragmented and then the molecular fragments are measured by an MS instrument to determine the mass-to-charge ratio of these chemical fragments. Generally, two types of ionization methods are used in GC-MS: electron impact ionization (EI) – the oldest and most common method, and chemical ionization (CI) – a newer but less frequently used method [109]. EI is a "hard" ionization technique that generates many fragment ions, while CI is a "soft" ionization technique that generates fewer fragment ions and keeps the molecular ion (M+) intact. The selection of a particular ionization method can have a significant effect on the types of compounds that can be analyzed and the kind of structural information that can be acquired by GC-MS. For example, EI-MS is particularly useful for ionizing volatile and lower molecular weight compounds (MW<600 Da). The ionization conditions in EI-MS are often so harsh that the molecular ion does not survive, making it difficult to determine the molecular weight of the parent compound [110]. On the other hand, CI [111], being a softer ionization technique, typically leaves the molecular ion intact, thereby generating simpler (albeit structurally less informative) MS spectra.

In EI-MS, electrons are generated through thermionic emission by electrically heating a wire filament. The electrons are accelerated to 70 eV whereupon they collide with the analyte molecule, removing an electron and converting the molecule to a positive ion with an odd number of electrons. Due to the high energy of the 70 eV electrons in EI-MS, several other bond dissociation or rearrangement reactions also occur, leading to the creation of second-generation product ions. It is these fragment ions or product ions, both their m/z values and intensities that can provide structural insights about the parent molecule. Indeed EI-MS spectra can serve as unique molecular or spectral “fingerprints”. These molecular fingerprints, in conjunction with other information such as retention times (or indices), can be used to identify or determine the structure of numerous small molecules [112].

While a small number of very skilled individuals can interpret EI-MS spectra and even identify compounds *de novo*, compound identification by EI-MS is most commonly done by comparing the measured EI-MS spectrum of an “unknown” compound from a mixture to an EI-MS spectral library of thousands of pure compounds. There are a number of well-known spectral libraries containing experimentally collected EI-MS spectra of pure compounds. These include the National Institute of Standards and Technology (NIST) mass spectral library [91], the Wiley Registry of Mass Spectral Data [113], the MassBank (North America, United Kingdom, and Japan) spectral libraries [114, 115, 116], and the Spectral Database for Organic Compounds (SDBS) [117]. This queried EI-MS spectrum is compared against the library EI-MS spectra using many well-known spectral similarity functions such as match factors [118-119], dot products, Pearson correlation coefficients, Euclidean distance, cosine similarity and spectral entropy [120] to identify the best spectral match or matches. These EI-MS libraries collectively contain approximately 600,000 reference EI-MS spectra from about 350,000 different compounds. However, many of the compounds in EI-MS libraries are multiply silylated or methoxime derivatives of the same compound, meaning that fewer than 100,000 parent or "source" compounds have EI-MS spectra available. Given that the number of known chemical compounds is >150 million [121], and the number of chemical (silyl, acyl, alkyl) derivatives possible for each of these compounds, it is clear that the number of experimentally measured EI-MS spectra is only a tiny fraction of what is needed to perform library-based compound identification. Furthermore, given the time and cost to prepare and collect experimental EI-MS spectra of pure compounds, it is unlikely that this enormous gap between known EI-MS spectra and known compounds will ever be filled.

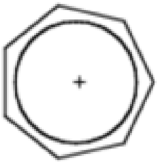
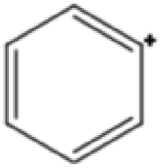
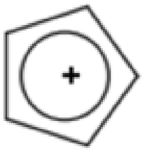
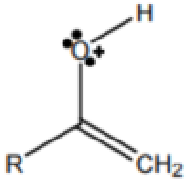
It is because of the lack of experimental reference EI-MS spectra that more GC-MS researchers are turning to computational approaches to interpreting, analyzing or predicting EI-MS spectra. For instance, software tools such as GenForm [122], MolGenMS [123], and MS-FINDER [124], have been developed to support MS-based structure elucidation by outputting annotated MS fragments and/or their corresponding formulae. But the inputs to these tools often require high-resolution MS data, which is not readily available for EI-MS. More recently, a quantum chemical approach, called QCEIMS [61-62] has appeared that uses quantum mechanics to predict EI-MS fragmentation patterns from parent compounds. This method has shown impressive results for a wide range of compounds, including TMS-derivatized molecules. However, the limitation of QCEIMS lies in the fact that it requires enormous amounts of time (days per spectrum) and computing resources. Meaning that large predicted spectral libraries cannot be generated. Other types of EI-MS fragmentation predictors have been developed that use smart combinatorial or machine learning approaches, such as MetFrag [125] and CFM-EI [65]. Unlike QCEIMS, these are very fast and they use the structure of an input molecule to predict the EI-MS fragmentation probabilities, the EI spectra and the structures/formulae of the fragments. The ability to rapidly predict EI-MS spectra from structure opens the door to creating massive EI-MS spectral libraries for compound identification. In a number of tests, CFM-EI [65] showed significantly better performance than other MS fragment predictive tools such as MetFrag [125], MOLGEN-MS, and Mass Frontier [69]. Other recently developed tools such as NEIMS [70] and RASSP [126] make use of deep learning architectures (Graph Neural Networks and their variants) to predict EI-MS spectra. These predictors achieve better accuracy (in terms of Jaccard scores and dot product scores) than CFM-EI but they cannot annotate the generated EI-MS peaks with formulae or structures as done by CFM-EI. Interestingly, recent developments in the field of ESI-MS/MS prediction have shown that including fragment formulae with the spectral data (m/z values and peak intensity) as part of the MS spectral training data set consistently yields excellent predictive performance [127, 128, 129]. Likewise, training ESI-MS/MS predictors on specific classes of molecules (rather than on combined sets of diverse molecules) has also been shown to improve their performance. This suggests the same concepts could be used to improve the performance of EI-MS prediction.

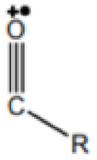
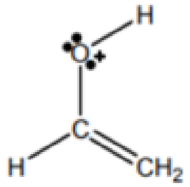
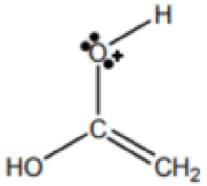
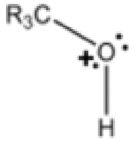
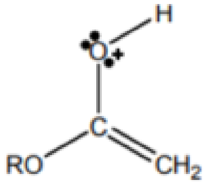
Here we describe the development of two types of improved machine-learned EI-MS predictors – both of which make use of some of the observed strengths and weaknesses of other EI-MS predictors. The first predictor (called EI-MS gamma) combines the deep learning concepts developed for NEIMS spectral modeling and the fragmentation prediction of CFM-EI. We later combine EI-MS gamma with a molecular intensity predictor (MIIP) and an EI-MS peak annotation method (called PeakAnnotator) that uses formula generation similar to MolGen and peak annotation methods developed for CFM-EI. We show that the spectral prediction quality and peak annotation coverage across CFM-EI significantly improve when EI-MS peak prediction and peak annotation are combined, however, this combination is still unable to exceed EI-MS peak prediction results of NEIMS and RASSP. Using PeakAnnotator, we also show that it is possible to comprehensively annotate both predicted and observed EI-MS spectra with their molecular formulae. These annotated EI-MS spectra can then be used as training data to possibly implement a more robust EI-MS predictor that employs GNN methods. The second EI-MS predictor (called Adjusted NEIMS with PA and MIIP) uses NEIMS spectra predictions and then makes an adjustment in its peak prediction outputs by verifying and validating the peaks with the help of PA and MIIP. The Adjusted NEIMS with PA and MIIP predictor is shown to perform equivalently in terms of dot product score (0.621 vs. Original NEIMS: 0.62), and better, in terms of spectral annotation correctness (91% vs. NEIMS: 0% vs. RASSP: 0% vs. CFMID: 53%) when it is compared against other EI-MS predictors including CFM-EI, NEIMS, RASSP, EI-MS beta, and EI-MS gamma. When the PeakAnnotator program was evaluated for spectral annotation coverage, it achieved an average annotation coverage of 94% (vs. Adjusted NEIMS with PA and MIIP: 79.6% vs. CFMID-EI: 86%).

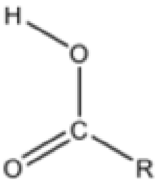
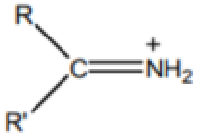
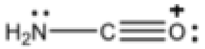
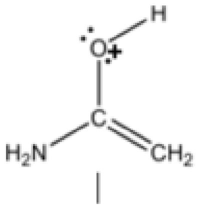
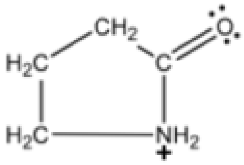
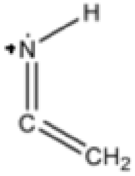
3.2 Data and methodology

3.2.1 Generation of the EI-MS dataset

The development of machine learning models to predict EI-MS spectra requires training data consisting of chemical structures and their known EI-MS spectra. Our training data source was the mass spectral library consisting of EI-MS spectra from 306,870 parent and derivatized compounds from the NIST 20 database [36]. The NIST 20 database includes compound names, the corresponding InChI keys (structure) along with their experimentally measured reference EI-MS data, including all mass-to-charge data (m/z) and relative peak intensities. For training purposes, we needed all NIST 20 structures expressed as a simplified molecular-input line-entry system (SMILES) format. These SMILES strings were then converted to molecular object (MOL) files via RDKit [94]. We used the NIH cactus service [92] and the PubChempy python [93] application programming interface (API) to perform the InChI key conversion to isomeric SMILES. Unfortunately, this conversion was not successful in all cases. In particularly, many of the NIST InChI keys were not present in NIH cactus service nor the PubChempy API and secondly, RDKit [94] could not generate valid molecular object (MOL) files for a large number of compounds. As a result, 45,163 NIST entries could not be fully processed and so they were removed from the training data set. Because most conventional GC-MS instruments are not able to reliably measure large molecules, we also removed compounds having molecular mass >900 Da from the dataset, (approximately 143 structures). In addition to ensuring the structural integrity of our dataset, we also made sure that it provided reasonable values of MS peak intensities (positive peak intensities and no peak intensity above 100%) and the presence of appropriate fragmentation patterns for well-known functional groups found in the associated compound. Details of these fragmentation patterns can be found in Table 3.1. Apart from some well-known functional groups, this table also represents some of the significantly occurring peaks (rather than all possible fragmentation corresponding to aliphatic and aromatic TMS and TBDMS alcohols, carboxylic acids, amines, sugars, steroids, thiols, phosphates, etc.) for the trimethylsilyl (TMS) and tert-butyldimethylsilyl (TBDMS) groups. We included this check with the help of an in-house python script. This filter led to the removal of another 322 compounds that appeared to have inconsistent or highly questionable EI-MS spectra. Interestingly some NIST 20 data files had long lists of unformatted or variably formatted meta-data such as instrument type, compound synonyms, proprietary statements, etc. integrated with the EI peak data. This made automated extraction of the EI-MS peak data nearly impossible. As a result, another 4,375 such entries were discarded. After this filtering and data cleaning process was complete, our final EI-MS training dataset consisted of 254,367 compounds (see Table 3.2) containing the names, InChI keys, molecular weights, molecular formulas, SMILES strings, and corresponding EI-MS peak lists (m/z values and normalized intensities) in a comma-separated values (*.csv) file format. Machine learned models need to be validated using a holdout or test data set. These data sets consist of samples that were never seen during the training/testing process. To assemble our holdout set, we extracted ~2,500 NIST 20 compounds with high-quality EI-MS spectra (as per the data and spectral filtering criteria applied above) representing various chemical classes including alkenes (362), alkynes (350), aldehydes (100), ketones (354), esters (352), silylated molecules (178), and low molecular weight (<150 Da) compounds (345), to be used as a holdout test set in order to evaluate our EI-MS predictor. In total, the EI-MS holdout data set consisted of 2,041 compounds containing the names, InChI keys, molecular weights, chemical formulas, SMILES strings, and corresponding EI-MS peaks (m/z values and normalized intensities) in a comma-separated values (*.csv) file format. Once the held out test set from different classes of compound was separated, we took the entire

Functional Group	Observed Fragments	m/z values
Straight chain Alkanes	C_nH_{2n+1}	43, 57, 71, …
	$M – CH_3$	M – 15
	$M – CH_2CH_3$	M – 29
	$M – CH_2CH_2CH_3$	M – 43
Branched Alkanes	C_nH_{2n}	Various
Cyclic Alkanes	$M – H_2C{=}CH_2$	M – 28
Alkenes	C_nH_{2n-1}	Various
	C_nH_{2n}	Various
Aromatics		91
		77
		56
Alcohols	$M – H_2O$	M – 18
	$M – (H_2O \ \& \ H_2C{=}CH_2)$	M – 46
	$M – (CH_3 \ \& \ H_2O)$	M – 33
Primary Alcohols	CH_2OH	31
Ketones		43+R

		Various
Aldehydes		44
	COH	29
	M – H_2O	M – 18
	M – $H_2C{=}CH_2$	M – 28
	M – $H_2C{=}CH{-}OH$	M – 44
Carboxylic Acids		60
	M – OH	M – 17
	M – CO_2H	M – 45
	$C_nH_{2n-1}O_2$	73, 87, …
Ethers	β cleavage	various
		various
Esters		74, 88, …

		various
	R’+	various
	$C_nH_{2n+2}N$	various
Amines		various
	$C_nH_{2n+2}N$	58, 72, …
Primary Amines	CH_2NH_2	30
Amides	CH_2NH_2	30
		44
		59
		86
Nitriles	M–H	M–1
		41

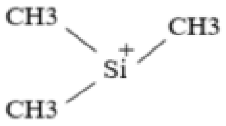
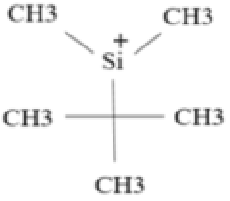
Trimethylsilyl (TMS)		1 TMS	73, 75
		2 TMS	144, 147
		3 TMS	218, 219, 217, 221
Tert-Butyldimethylsilyl (TBDMS)		1 TBDMS	73, 75, 112, 115
		2 TBDMS	73, 75, 112, 115
		3 TBDMS	No visible pattern found

Table 3.1: EI-MS fragmentation patterns observed in various functional groups.

training set to create sub-training sets of different chemical classes in order to create specialized predictors (as opposed to the comprehensively big generalized model discussed in further sections). In particular, these training sets consisted of small batches of alkenes (10,000), silylated molecules (9,000), alkynes (3400), aldehydes (1867), ketones (2,400), and esters (10,000). It is to be noted that, the number of aldehydes in both the held out test set and the sub-training set was small because of the spectral filtering (to comply with reasonable fragment peaks) applied above. Finally, we consolidated another held out test set of 2,008 brand new molecules from the recent NIST 23 EI-MS release. All these compounds in the sub training sets and held out test sets passed tests for normality, indicating they follow a gaussian distribution.

Data Type	Starting # EI data entries	Final # EI entries in low m/z spectra set	Final # EI entries in high m/z spectra set	# removed (low m/z)	# removed (high m/z)	% data removed
Total # of EI entries	**306,870**	**254,367**	**245,460**	**52,503**	**61,410**	**Low: ~17%** **High: ~20%**
Data cleaning process						
Data removed during SMILES conversion				45,163		14.7%
Data having MW >900 Da				~150		<1%
Data removed to ensure reasonable fragment peaks				322		<1%
Data removed due to contamination with additional information in the (peak, intensity) pairs				4,375		1.42%
Data separated for hold-out test set				~2,500		0.8%
Data removed in the high m/z spectra set while stripping off peaks below 130 Da.				7,532		~2.5%
Data removed in the high m/z spectra set having peaks beyond m/z position 1000				20		<1%
Data removed in the high m/z spectra set having eliminated all peaks while stripping off peaks below 130 Da.				1,355		<1%

Table 3.2: Summary of pre-processing and cleaning process of the NIST EI data

3.2.2 Implementation and Testing of Different "Peak-Only" EI-MS Prediction Models

Three different versions of "peak-only" EI-MS predictors were developed and tested: EI-MS alpha, EI-MS beta and EI-MS gamma. Each used unannotated EI-MS data (described in Section 3.2.1) to train, test or evaluate their performance. Two of these models (beta and gamma) were also supplemented with predictors for molecular ion intensity and peak formula annotators. Details regarding their development and testing, as well as the development of the molecular ion intensity predictor and peak formula annotator are provided below.

3.2.3 Development and Testing of EI-MS alpha

The first EI-MS predictor (EI-MS alpha) we developed was based on a Graph Neural Network (GNN) model similar to the one described for NEIMS [70]. We also drew inspiration and borrowed code from our in-house built GNN-based retention index (RI) predictor described in the previous chapter. Specifically, using the code from our earlier GNN RI predictor we made some minor modifications to the data input and output formats, added several more compound feature attributes and implemented a different loss function. EI-MS alpha was implemented and written in python using the tensor-optimized library, PyTorch [97]. For atom-level feature extraction, we also used RDKit functions.

EI-MS alpha is designed to accept a SMILES string corresponding to the query structure for which the user wishes to predict the EI-MS spectra. The SMILES string is then directly converted into a molecular graph where the nodes and edges of the graph represent atoms and bond connections between the atoms, respectively. Using RDKit, MOL objects were generated on the fly using the SMILES string as input. From these MOL objects, a number of atom-level features are then extracted from the molecular graph, including atomic (node-level) features and path (or bond) features, related to the graph edges. The atomic feature set for the EI-MS predictor is a combination of 64 one-hot coded features representing 62 types of common elements in the periodic table (such as C, N, S, O, F, etc.), plus one feature for any unknown elements and one other being used as a default wildcard or dummy symbol. Atomic features also included five types of valid formal charges ranging from neutral, positive, and negative charges [0, ±1, ±2] and seven explicit and six implicit valences. Additionally, our atom feature set also included a path feature related to the neighboring nodes (atoms) of any given node (atom), up to a maximum of 4 neighboring atoms. In addition to atom features, we also encoded path features for any two given nodes (atoms) of each molecule by computing the shortest path between them. The maximum path length we considered between any pair of nodes (atoms) was three. For any pair of nodes or atoms, the path feature set is a combination of 31 one-hot coded features and was computed as a concatenation of the model's maximum path length, the total number of bond features in that particular path, bond conjugacy, bond ring appearance (denoting the presence of a ring between two neighboring nodes) for that path and the ring membership (denoting whether the nodes in a path are in a ring, which is specifically applicable to aromatic structures) between the nodes in a path.

Six types of bonds were used to compute these bond features. The atom or node-level features that were computed for any two nodes in a graph were then fed into the input layer of the GNN by a linear transformation. In addition, this EI-MS predictor also computed atom attention scores both in the hidden layers (with five hidden layers each of the layers having 160 hidden units) and the output layers. The new atom embeddings were computed afterwards which helped in generating output values from the final layer of the network. The output of this network is a 1000-

dimensional predicted relative intensity value, each of which represents a particular m/z position with 1 Da resolution in the EI-MS spectra.

EI-MS alpha was trained and validated on the entire EI-MS data set described in section 3.2.1. Specifically, we divided the data set into three groups: 1) training, 2) validation, and 3) testing sets using a ratio of 80:10:10. This corresponded to 203,493 training spectra, 25,430 validation spectra and 25,430 test spectra. Analyzing the prediction results from EI-MS alpha revealed an unexpected trend. We observed that this predictor performed particularly well at low m/z values (typically below 130 Da) while the quality of the predictions in the higher m/z regions (>130 Da) was disappointingly poor. This qualitative observation was later quantitatively confirmed by calculating the average dot product scores on a sliding window of m/z values ranging from 80 Da to 160 Da for various compound classes. We found that there was an optimal cut-off of ≤130 Da where the predicted EI-MS spectra generated by EI-MS alpha were best. This process is depicted graphically in Figure 3.1.

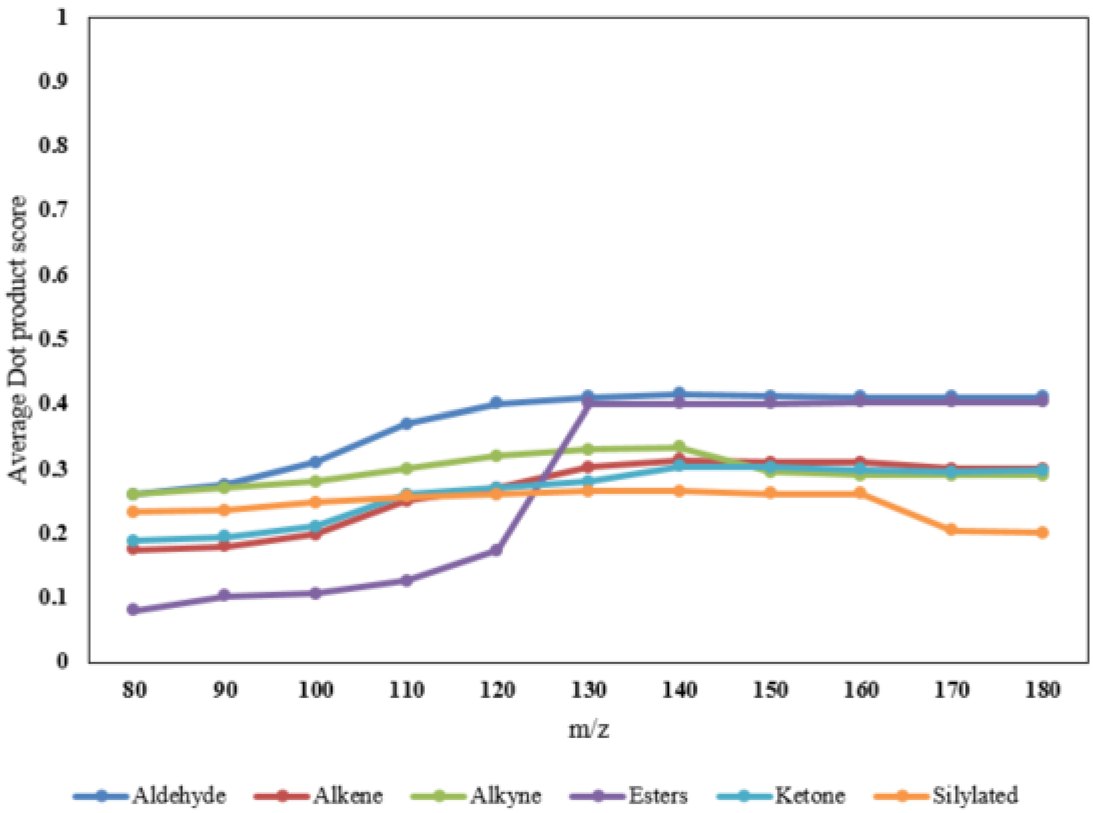


Figure 3.1: Average dot product score on a sliding window ranging from MW of 80 to 160 Da. to select the spectral set cut-off.

The performance of EI-MS alpha was assessed on the hold-out set of 2,041 EI-MS spectra derived from the NIST 20 data set, covering multiple compound classes or categories (described in Section 3.2.1). The performance was further assessed on the consolidated NIST 23 set of 2,008 brand new compounds as well. The quality of spectral prediction was assessed using the dot product similarity coefficient [118], a metric that is commonly used to report the similarity between an experimental and predicted spectrum by mass spectrometry software. The dot product similarity coefficient is calculated using the following equation:

$$Similarity(I_q, I_l) = \frac{\sum_{k=1}^{M_{max}} m_k I_{qk}^{0.5} . m_k I_{lk}^{0.5}}{\left\|\sum_{k=1}^{M_q}\left(m_k I_{qk}^{0.5}\right)^2\right\| \left\|\sum_{k=1}^{M_l}\left(m_k I_{lk}^{0.5}\right)^2\right\|} \quad \text{Eq.1}$$

where I_q and I_l are two m/z intensity vectors representing the predicted spectrum and experimental spectrum respectively. At any m/z position k, m_k and I_k represent that position's mass to charge ratio and intensity. The highest indices of I_q and I_l, having non-zero values, are represented by M_q and M_l and M_{max} represents the maximum between M_q and M_l.

3.2.4 Development and Testing of EI-MS-beta

The results from EI-MS alpha led us to modify our training process and training data set and to implement a second version of the program, which we called EI-MS beta. Specifically, we decided to separate our entire EI-MS data set into two spectra data sets, namely a high m/z (>130 Da) and low m/z (≤130 Da) data set. Specifically, the low m/z spectra set was created by stripping off all EI-MS peaks above 130 Da while the high m/z spectra set was created by stripping off all EI-MS peaks ≤130 Da. This peak removal process was performed for all compound spectra in the training data set. This peak-reformatting process led to the removal of approximately 7,532 structures (and their corresponding EI-data) in the high m/z spectra set (as these compounds had no peaks above 130 Da). Interestingly, during the peak removal process, in spite of having molecular weights of more than 130 Da, 1355 compounds actually had all of their EI-MS peaks filtered out. As a result, we eliminated these structures and their EI data from the high m/z spectra set. Ultimately, the final dataset for the low m/z and high m/z spectra sets consisted of a total of 254,367 and 245,460 compounds respectively. The results of the data pre-processing and cleaning process are shown in Table 3.2.

For both for low m/z and high m/z spectra datasets, we further divided them into three groups: 1) training, 2) validation, and 3) testing sets using a ratio of 80:10:10. For the high m/z spectra that corresponded to 196,368 training spectra, 24,546 validation spectra and 24,546 testing spectra. For the low m/z spectra that corresponded to 203,493 training spectra, 25,436 validation spectra and 25,436 testing spectra. It is important to note that, the validation set defined here was used solely for the GNN model optimization/tuning. The partitioning within each of these three sets was done in such a way that the same proportion of similar compounds (in terms of elemental ratios, functional groups, ClassyFire [99] chemical class, and superclass) existed in the training, validation, and testing data sets.

All the parameters used to train the EI-MS beta predictor using the low m/z spectra set and the high m/z spectra set are listed in Table 3.3. Except for a few parameters (e.g. batch size), all other parameters were constant in these models. Both these models were trained for up to 700 and 500 epochs respectively. Early stopping was used to help efficiently identify the most promising models. During these epochs, we tracked the training versus validation errors (generated from the internal validation/development set during the model optimization/tuning) to select the best-trained models. For our GNN, the early stopping was defined by a patience parameter, which was varied (100, 200,

Name of the parameter	Description of the parameter	Value
agg_func	Aggregating function [sum or mean], used to aggregate the individual node embeddings	*sum*
batch_size	Number of examples used per batch	*100-200*
batch_splits	Used to aggregate batches	*2*
optimizer	Model optimizer	*Adam*
depth	The depth of the neural network	*5*
dropout	The dropout probability for the model	*0.2*
hidden_size	The number of hidden units for the model	*160*
loss_type	The loss type for the dataset	*mse*
lr	The learning rate for the optimizer	*0.0005*
max_grad_norm	The maximum gradient norm allowed	*10*
max_path_length	The max path length to consider between neighboring nodes	*3*
n_heads	Number of heads in multi-head attention	*2*
d_k	The size of each individual attention head	*80*

Table 3.3: Parameters and their values used to develop the EI-MS beta predictor.

300, 400, and 500 times) to capture the most effective training model. The loss function that was used in the GNN architecture is defined by

$$L(I,\hat{I}) = \sum_{k=1}^{M(x)} \left(\frac{m_k I_k^{0.5}}{\left\| \sum_{k=1}^{M} \left(m_k I_k^{0.5}\right)^2 \right\|} - \frac{m_k \hat{I}_k^{0.5}}{\left\| \sum_{k=1}^{M} \left(m_k \hat{I}_k^{0.5}\right)^2 \right\|} \right)^2 \qquad \text{Eq. 2}$$

where I and I-hat represent the ground truth and predicted spectrum respectively, and M(x) is the mass of the given molecule. In the last step, we combined both of these best-trained models from the low m/z spectra predictor and high m/z predictor to obtain a combined predictor (called EI-MS beta). The performance of EI-MS beta was assessed on the hold-out set of 2,041 EI-MS spectra derived from the NIST 20 data set, covering multiple compound classes or categories (described in Section 2.1) as well as the brand new NIST 23 set of 2,008 compounds. The quality of spectral prediction was assessed using the dot product similarity coefficient as described in Section 3.2.3.

3.2.5 Development and Testing of EI-MS-gamma

The results from testing EI-MS beta (a pure GNN model) on both the hold-out and training sets suggested that the high m/z spectral predictor was under-performing compared to the low m/z EI-MS spectral predictor. It was also observed that the performance for CFM-EI [65] for high m/z values was generally quite good while its performance was relatively poor for low m/z values. Therefore, a third EI-MS predictor based on combining the low m/z (≤130 Da) EI-MS spectral predictor that uses the GNN model described above with the high m/z portion (>130 Da) of CFM-EI that uses a probabilistic graphical model [65] was created. This hybrid EI-MS predictor simply used the low m/z EI-MS model for peaks with m/z values ≤130 Da and the CFM-EI model for peaks with m/z values >130 Da. No further training or optimization was done on this predictor (called EI-MS gamma). The performance of EI-MS gamma was assessed on the hold-out set of 2,041 EI-MS spectra derived from the NIST 20 data set, covering multiple compound classes or categories (described in Section 2.1) as well as the NIST 23 set of brand new compounds consisting of 2,008 compounds. The quality of spectral prediction was assessed using the dot product similarity coefficient as described in Section 3.2.3.

3.2.6 Development of the EI-MS Molecular Ion Intensity Predictor (MIIP)

Analysis of the EI-MS spectra predicted by both EI-MS alpha and EI-MS beta showed that the intensity and/or presence of the molecular ion (M^+) was poorly predicted. Given the importance of the molecular ion for EI-MS data in identifying molecules or determining molecular structures we decided to explore whether ML methods could be used to improve the quality of the molecular ion prediction. In order to develop a molecular ion intensity predictor, we utilize the relative intensities of the molecular ions for all the spectra from the full EI-MS training set (see section 3.2.1). As a result, we prepared a molecular ion training set consisting of the structure information (InChI key, SMILES strings), m/z positions of the molecular ions, and their corresponding relative intensities for 253,367 compounds. Because of the high computational cost of the training algorithm, not all 253,367 examples were actually used (or even necessary) to generate a final, high performing model. Specifically, we tested the amount of CPU time and memory it took to train the model using an initial training set of 25,000 structure-ion pairs and gradually increased the training size by 25,000 examples each time. In the end, a subset of 100,000 structures and their molecular ion pairs was used in our final model. This number provided sufficient examples for training and achieved the desired level of accuracy. Later, MIIP was assessed only on the molecular ion-intensity pairs derived from the hold-out set of 2,041 EI-MS spectra from the NIST 20 database, covering multiple compound classes or categories (described in Section 3.2.1).

We chose to learn molecular ion intensity values by converting structures into compact representations based on Morgan Fingerprints [130]. This is because Morgan Fingerprints can encode important chemical information such as the size and type of the molecule, the composition of the underlying atoms and their three-dimensional arrangement within the molecules, all of which

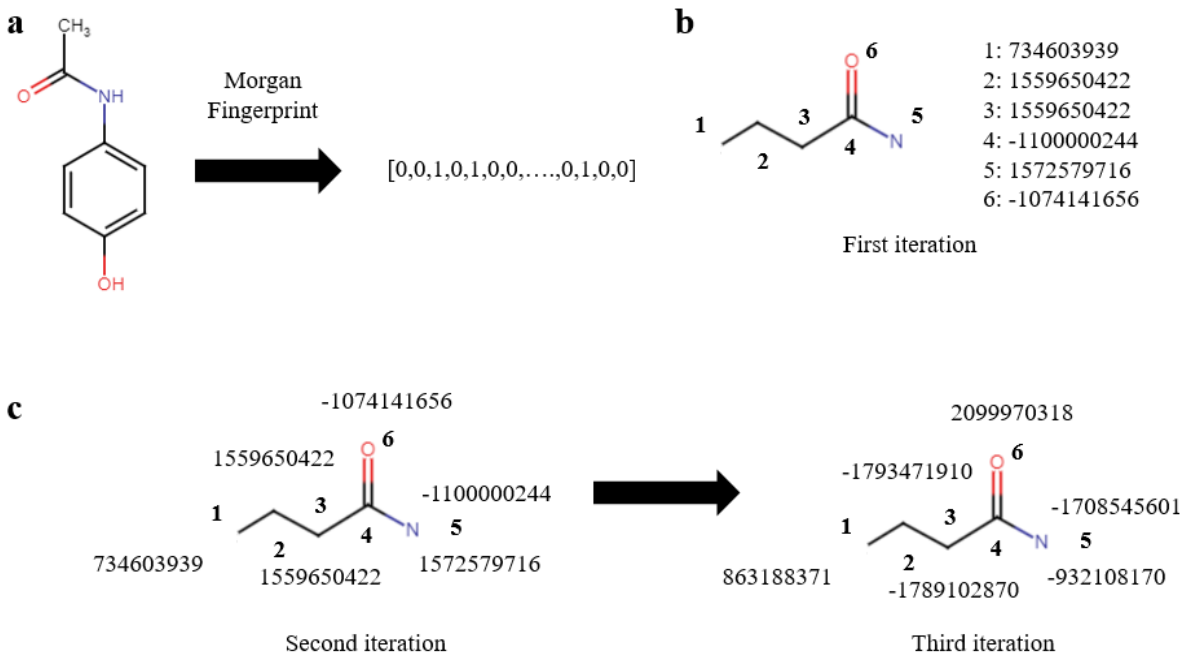
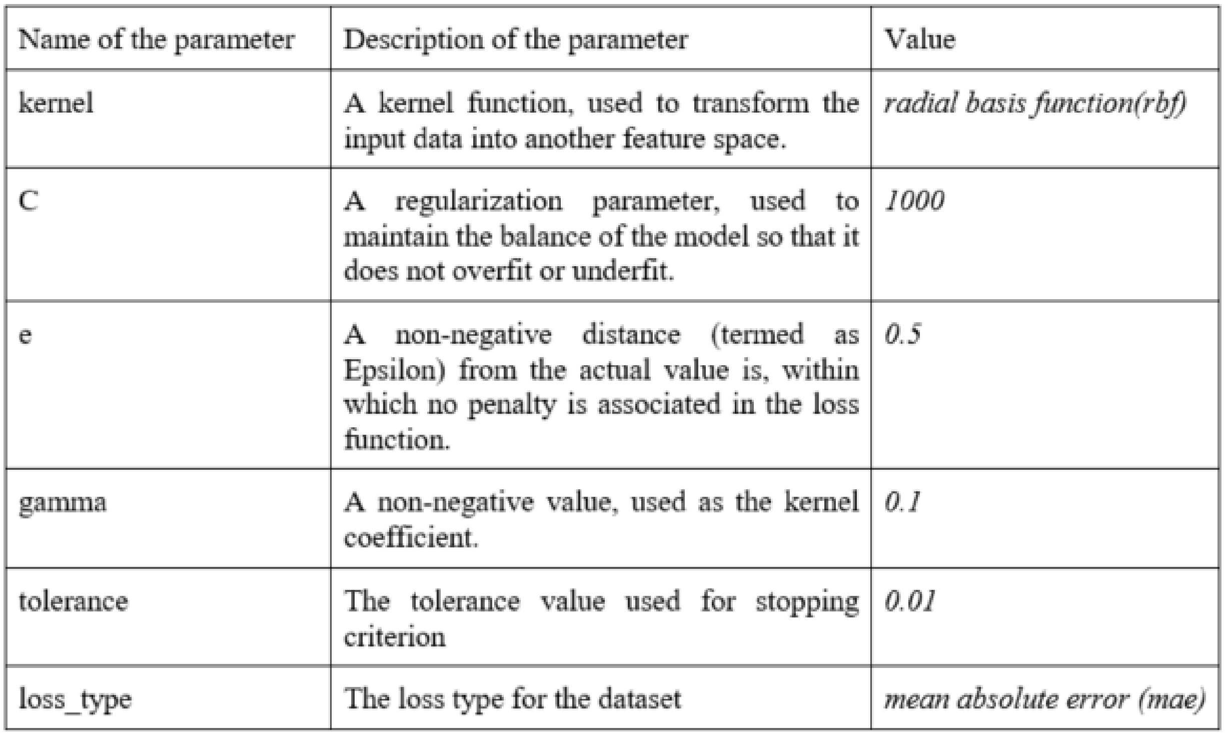


Figure 3.2: Morgan fingerprint process explained a) Given a molecule, the output morgan fingerprint is a string of 0s and 1s b) Each atom in a given molecule is assigned a unique identifier c) Identifiers of each atom are updated iteratively up to a given radius specified by the fingerprinting algorithm.

are very much related to the ionization efficiency of the molecular ion [131], and its corresponding intensity.

When a smaller subset of fingerprint bits are used to represent two different molecular structures, it is possible to generate same set of fingerprints for both molecules. This occurs due to a phenomenon called bit collisions. Therefore, to reduce bit collisions, we generated all 2048 bits of Morgan fingerprints with a radius of four for all these molecules. The process of fingerprinting is depicted in Figure 3.2 for a molecule paracetamol, where it first breaks down into a smaller fragment and generates the hashed identifier from there. As shown in Figure 3.2b, it starts by assigning a unique identifier to each atom and iteratively updates the identifiers based on the adjacent atom levels (see Figure 3.2c). Once the fingerprints are generated, we utilized a Support



Name of the parameter	Description of the parameter	Value
kernel	A kernel function, used to transform the input data into another feature space.	*radial basis function(rbf)*
C	A regularization parameter, used to maintain the balance of the model so that it does not overfit or underfit.	*1000*
e	A non-negative distance (termed as Epsilon) from the actual value is, within which no penalty is associated in the loss function.	*0.5*
gamma	A non-negative value, used as the kernel coefficient.	*0.1*
tolerance	The tolerance value used for stopping criterion	*0.01*
loss_type	The loss type for the dataset	*mean absolute error (mae)*

Table 3.4: Parameters and their values used in the molecular ion intensity predictor using SVR.

Vector Regressor (SVR) [132] to predict each molecule's molecular ion intensity. While developing this regressive model, we varied the regularization parameter C (with values = 0.1, 1, 10, 100, and 1000) and the epsilon value ε (with values = 0.0001, 0.01, 0.2, 0.5, 1, and 5), which were later optimally selected with the help of 5-fold cross validation. The selected optimal set of parameter values along with the description of each type of parameter are shown in Table 3.4. The error metric used for the molecular ion predictor was the mean absolute error (MAE). It is to be noted that SVR allows leveraging of kernels and projects the input set of features to a higher dimensional space. While this property can be particularly useful to model the non-linear relationship between our features and the target output (molecular weight vs. molecular ion relative intensity as seen in Figure 3.3), it also has the drawback of requiring considerable computer resources. The molecular ion intensity predictor developed in this process was then combined with EI-MS beta and EI-MS gamma (described in Section 3.2.4 and 3.2.5) to generate improved versions

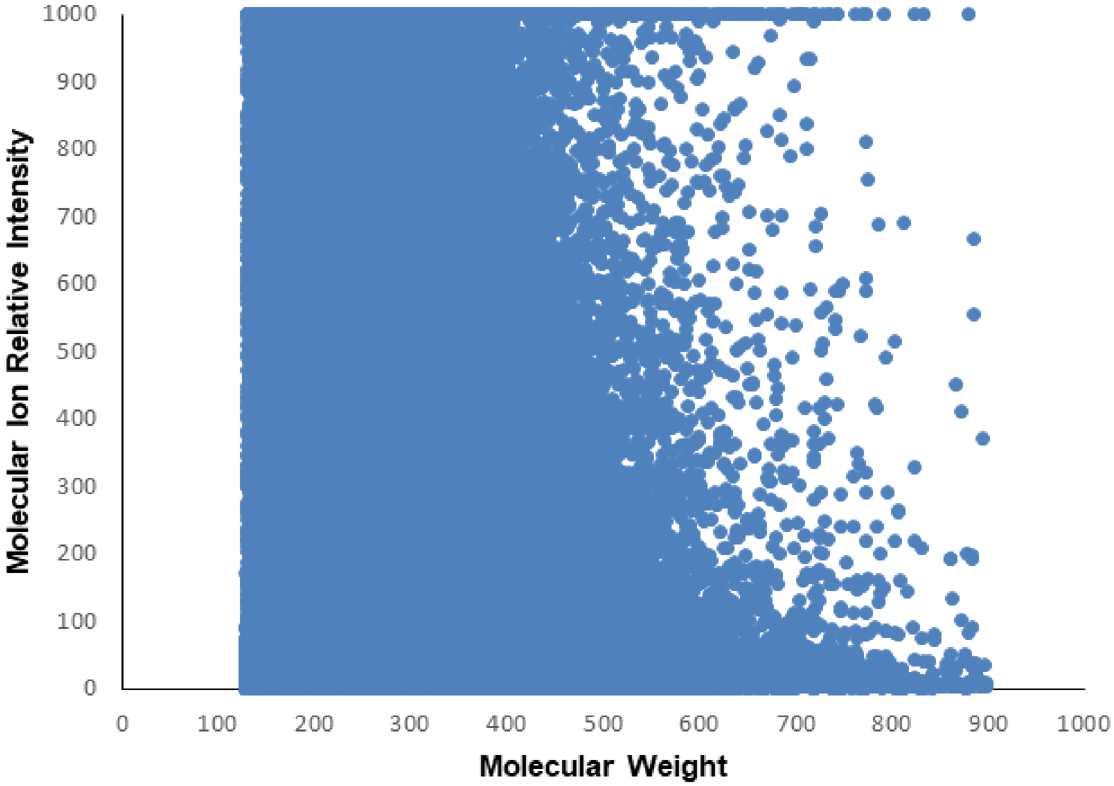


Figure 3.3: Plot showing correlation between the molecular weight and the molecular ion relative intensities for the molecules present in our dataset.

of EI-MS beta and gamma (called EI-MS delta and EI-MS epsilon, respectively). This is illustrated in Figure 3.4. The performance of both the enhanced EI-MS delta and EI-MS epsilon was assessed on the hold-out set of 2,041 EI-MS spectra from the NIST20 database, covering multiple compound classes or categories (described in Section 3.2.1) as well as on 2,008 brand new compounds from the NIST 23 database. The quality of spectral prediction was assessed using the dot product similarity coefficient as described in Section 3.2.3.

3.2.7 Computational annotation of EI-MS peaks – PeakAnnotator (PA)

In order to aid the interpretation of the EI-MS spectral peaks of both experimentally measured and computationally predicted EI-MS spectra, it is important to annotate peaks with either subformulae or substructures. CFM-EI [65] is able to perform peak annotations of EI-MS spectra for about 50% of observed peaks. The NEIMS program [70], unfortunately, does not perform any annotations. Likewise, the predictors described above (EI-MS alpha and EI-MS beta) which have a similar architecture to NEIMS, also did not perform peak annotations. To address this shortcoming, we developed a python script to automate the peak annotation process of EI-MS spectra. This program (called PeakAnnotator) can annotate peaks from predicted EI-MS spectra (generated by EI-MS beta or EI-MS gamma) as well as peaks from experimentally acquired EI-MS spectra. Given the low resolution of most EI-MS spectra (~1 Da resolution) it is easy to imagine that many possible molecular formulae can generate the same mass or m/z value. To reduce the redundancy and to

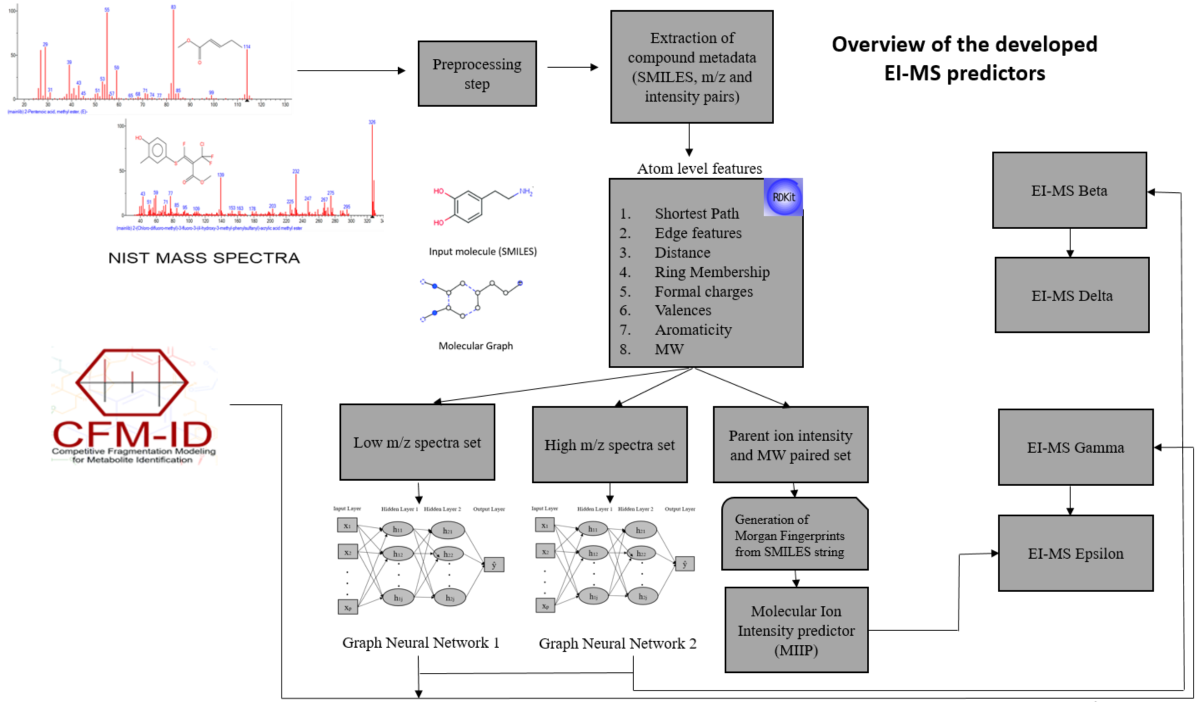


Figure 3.4: Overview of the EI-MS predictors (beta, gamma, delta and epsilon).

eliminate mislabeling of peaks, the PeakAnnotator script takes into consideration a number of basic properties of formula generation, given a molecule's mass (in this case, all the predicted m/z values in a spectrum) and elemental composition (i.e. the formula of the parent molecule). The program combinatorially generates all possible subformulae for each of the observed (or predicted) m/z values, which are further constrained to a smaller number of possibilities using rules such as the nitrogen rule [133], the senior rule [134], and the degree of unsaturation [135]. We also combine a knowledgebase of frequently known EI-MS peak patterns for particular functional groups and structures (as described in Table 3.1) to further refine the formula generation process. PeakAnnotator also removes every peak from each cluster of peaks within a given EI-MS spectrum where the peak's relative intensity is less than 1 percent of the maximum cluster peak. This is done to get rid of unnecessary annotations or peaks arising from ^{13}C isotope peaks or spectrometer noise. PeakAnnotator also include rules for handling and annotating other high abundance isotopic elements such as Cl, Br, I, etc.. To further improve the formula generation process, PeakAnnotator combines the subformulae and substructures generated from CFM-EI's fragmentation module to once again narrow down the potential formula possibilities. Finally, if there are any remaining peaks for which an appropriate peak annotation cannot by suggested by PeakAnnotator, the algorithm will automatically add or remove hydrogens (up to four) to nearby confirmed subformulae to annotate the remaining peaks. The PeakAnnotator was integrated with EI-MS beta and EI-MS gamma (described in Section 3.2.4 and 3.2.5) to generate more comprehensive versions of EI-MS beta and gamma.

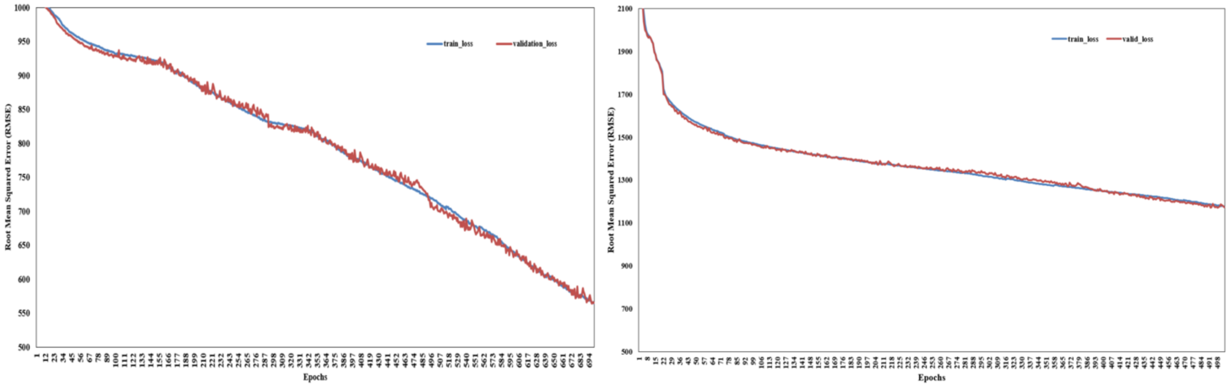
3.2.8 Development and Testing of Adjusted NEIMS with MIIP and PA

The “peak only” predictors (EI-MS gamma, delta and epsilon) when compared with NEIMS and RASSP [126] predictors on the hold-out test set from NIST 20, suggested that both NEIMS and RASSP still outperform these developed predictors with NEIMS providing the highest dot product score. However, both these high performing predictors lack in generating any peak annotation. On the other hand, the PA program can not only provide subformulae annotations for the predicted peaks, rather it can also be used to verify and edit (when needed) the generated peaks from the NEIMS program. Therefore, a finalized predictor utilizing the benefits of NEIMS, PeakAnnotator (PA) program and the Molecular Ion Intensity Predictor (MIIP) can be suggested. We name this as Adjusted NEIMS with MIIP and PA. No further training or optimization to the original NEIMS program was done while incorporating it in this adjusted predictor. The performance of this adjusted predictor was assessed only on a brand new compound test set from NIST23 consisting of 2,008 compounds. The quality of spectral prediction was assessed using the dot product similarity coefficient as described in Section 3.2.3.

3.3 Results and Discussion

3.3.1 Performance of “Peak Only” EI-MS Prediction Models

A total of six different “peak-only” EI-MS spectra prediction models were generated and evaluated. These included 1) a naïve GNN version that was trained on the complete NIST 20 EI-MS data set (EI-MS alpha); 2) a more refined version that uses different GNN models to predict EI-MS spectra for high m/z regions (>130 Da) and low m/z regions (≤130 Da), called EI-MS beta; 3) a version that uses a GNN model to predict EI-MS spectra for low m/z regions (≤130 Da) and CFM-EI to predict EI-MS spectra for high m/z regions (>130 Da), called EI-MS gamma; 4) EI-MS beta with



a) GNN low **b) GNN high**

Figure 3.5: Plot of the change in the mean absolute error versus the training epoch for the low m/z spectra set GNN and high m/z spectra set GNN.

the molecular ion predictor added 5) EI-MS gamma with the molecular ion predictor added and 6) Incorporating NEIMS with PeakAnnotator and molecular ion intensity predictor, called the Adjusted NEIMS with PA and MIIP. Recall that in developing EI-MS beta we used two EI-MS spectra data sets (high and low m/z values) to learn EI-MS spectra prediction via a GNN. The high m/z spectra set models, trained for up to 500 epochs, proved to be difficult to train and computationally time-consuming to run, whereas the low m/z spectra set model reached its optimal state fairly quickly (due to the inherently more rapid computations) within 700 epochs. The performance plot (training vs. validation error) of the best set of models derived from the EI-MS beta low m/z spectra and EI-MS beta high m/z spectra set are shown in Figure 3.5 (partial plots are shown for clarity). Both these models show a rapid decrease in the error metric in the first 50-70 epochs but this flattened out in later iterations. The best performing GNN models (for both low and high m/z spectra sets) were then combined to create EI-MS beta.

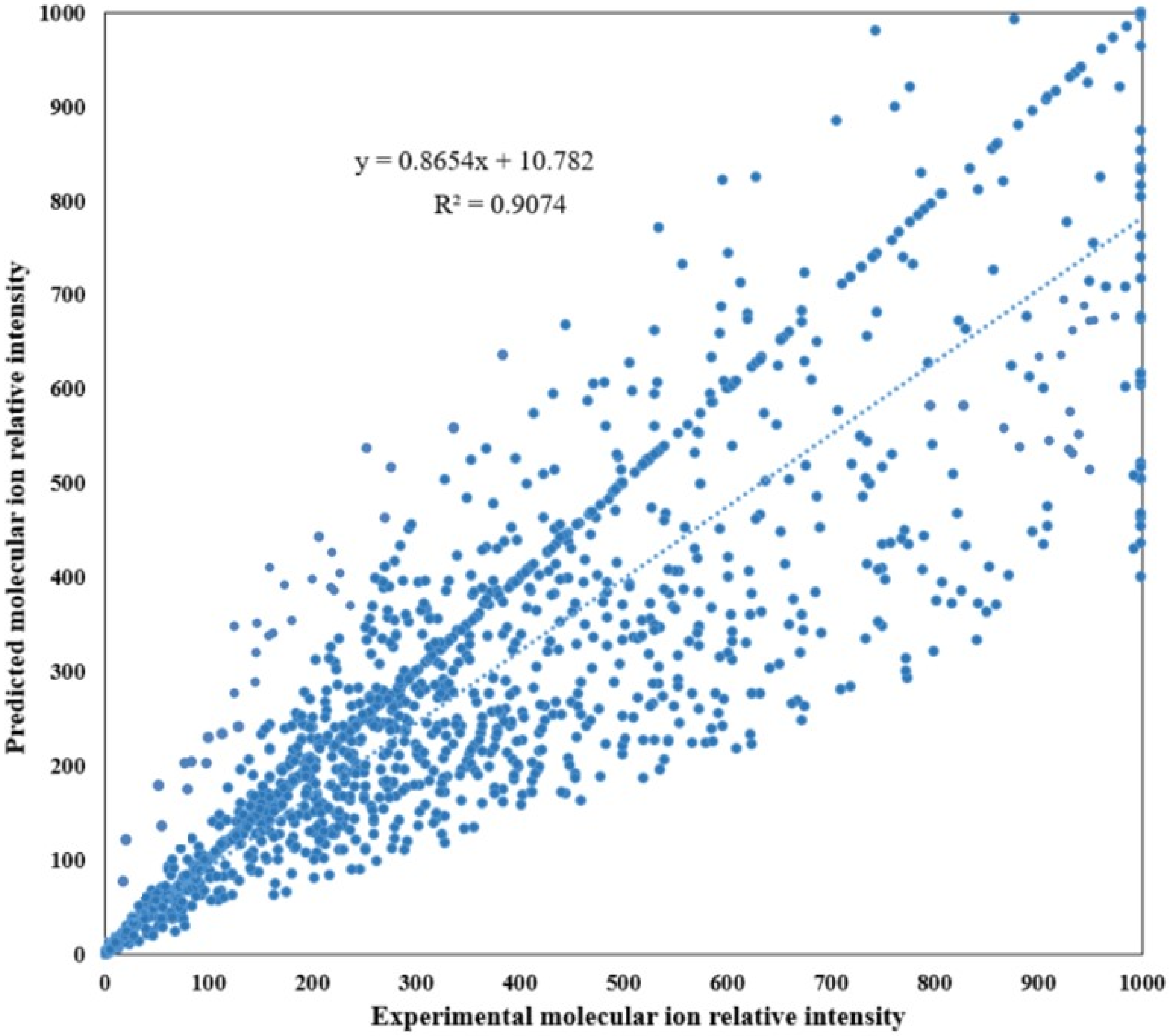


Figure 3.6: Performance plot showing the experimental and predicted molecular ion relative intensities by the SVR model.

On the other hand, EI-MS gamma combined the GNN predictor for low m/z values (obtained from EI-MS beta) with the probabilistic graphical model predictor (CFM-EI) to high m/z values. Both EI-MS beta and EI-MS gamma were evaluated with and without the molecular ion predictor. The molecular ion predictor is an SVR model designed to predict relative intensities of the molecular ion observed in EI-MS spectra. We used a cross-validated (5-fold), carefully tuned set of hyper parameters to train and evaluate the model. The best performing model (assessed on a holdout set of 2,041 NIST 20 EI-MS spectra) achieved a mean absolute error (MAE) of 35.47 units and an R^2 value between observed and predicted ion intensity of 0.9074 (See Figure 3.6). It can be seen from this figure that the EI-MS beta predictor predicted lower intensity values near the bottom of the plot. This can be explained by the fact that molecular ion relative intensities for certain compounds with specific functional groups are always absent. Additionally, the predictor is unable to generate correct relative intensities for some high intense molecular ions specifically from the aldehyde and ester sets of molecules (see rightmost data points in Figure 3.6).

All six models created here were, along with the CFM-EI model (developed in 2016), evaluated using the holdout set of 2,041 EI-MS spectra from the NIST 20 set (described in Section 3.2.1) and a NIST 23 test set consisting of 2,008 compounds. The dot product similarity coefficient was employed to measure the spectral match quality for each predicted vs. observed EI-MS spectrum. Additionally, the average of the dot product scores from each class of molecules in the holdout set was calculated for each of the six models and reported in Table 3.5. As seen from this table that EI-MS beta consistently performs better than CFM-EI and EI-MS alpha, except for three cases. These include the small molecule set (EI-MS beta: 0.363 vs. CFM-EI: 0.372 vs. EI-MS alpha: 0.463), the silylated data set (EI-MS beta: 0.28 vs. CFM-ID 2.0: 0.36 vs. EI-MS alpha: 0.202), and the ester set (EI-MS beta: 0.303 vs. CFM-EI: 0.215 vs. EI-MS alpha: 0.41). It is also evident that EI-MS gamma, which combined the GNN model with CFM-EI gave a significant boost in performance for nearly all test molecules and classes (except for the ester and aldehyde classes). It is also clear that a significant boost in performance occurs when the molecular ion intensity predictor is integrated into both EI-MS beta and EI-MS gamma (see two rightmost columns). In other words, by combining the low m/z spectra GNN model with CFM-EI (for the high m/z data) and the molecular ion intensity prediction provided higher predictive performance (in terms of dot product scores) for all the molecule groups except for the aldehydes and esters.

Molecule Groups	Test set size	CFM-EI	EI-MS alpha	EI-MS beta	EI-MS gamma	EI-MS delta	EI-MS epsilon	NEIMS	RASSP	Adjusted NEIMS with PA and MIIP
Alkene	362	0.316	0.307	0.359	0.376	0.37	**0.387**	**0.735**	0.657	
Alkyne	350	0.33	0.288	0.359	**0.367**	0.354	0.362	**0.74**	0.61	
Aldehyde	100	0.401	0.408	0.523	0.508	**0.533**	0.514	0.77	**0.821**	
Ketone	354	0.302	0.295	0.343	0.364	0.363	**0.384**	**0.73**	0.64	
Light weight molecule	345	0.372	0.463	0.363	-	0.527	**0.536**	0.76	**0.84**	
Silylated set	178	0.36	0.202	0.2771	0.417	0.342	**0.4249**	**0.41**	N/A	
Esters	352	0.215	**0.41**	0.303	0.263	0.34	0.252	**0.66**	<0.1	
Weighted Average	2041	0.3104	0.342	0.348	0.36	0.39	0.39	0.7	0.55	
NIST23	2008	0.32	0.41	0.423	0.373	**0.45**	0.38	**0.62**	0.56	**0.621**
Paired T-test result										
NIST23 (2008 molecules)		CFM-EI EI-MS Alpha	EI-MS Alpha-Beta	EI-MS Beta - Gamma	EI-MS Gamma -Delta	EI-MS Delta-Epsilon	EI-MS Epsilon NEIMS	NEIMS & RASSP	NEIMS & Adjusted NEIMS with PA and MIIP	
Two tailed P value (with 95% confidence interval)		<0.0001	0.0514	0.438	0.0041	0.0169	0.0001	<0.0001	0.0493	

Table 3.5: Dot product similarity scores to represent the spectral match quality between predicted and observed EI-MS spectrum for both the held out test sets derived from NIST20 and NIST23 and pair wise t-test results between the adjacent models. Note that underlined entries indicate that the dot product scores (between the two listed models) are not statistically significant.

It is surprising to see that the best performing GNN developed models for predicting EI-MS spectra for aldehydes is EI-MS delta, with a dot product score of 0.533 and for esters it is EI-

MS alpha, with a dot product score of 0.41, respectively. Overall, when we look across all compounds and all compound classes our best performing model (EI-MS gamma with the molecular ion predictor) has a dot product score that is on average ~24 percent better than CFM-EI. Likewise, the best performing aldehyde and ester prediction models exhibit an improvement in their dot product of around 34 and 58 percent respectively over CFM-EI. This suggests that we could use different EI-MS prediction models for different chemical classes to generate optimal EI-MS spectra. In other words, by using a program such as ClassyFire [99] to parse and classify compounds it would be possible to send different molecules to different EI-MS predictors to produce optimal results. We also incorporated the results of paired t-tests comparing successive EI-MS models. It is evident from the table that there are significant difference between the EI-MS Alpha and CFM-EI model (with $p<0.0001$), while the EI-MS Gamma, Delta and Epsilon models led to modestly significant successive improvements (with $p=0.0041$-0.0169). The NEIMS model clearly performed better than both EI-MS Epsilon and RASSP (with $p<=0.0001$) while the Adjusted NEIMS with PA and MIIP (EI-MSpred) model shows a modestly significant improvement over the regular NEIMS model (with $p=0.0493$).

3.3.2 Analysis of Individual EI-MS results for “Peak Only” EI-MS Prediction Models

While bulk comparisons with large numbers of spectra are informative, it is also useful to look at individual spectra to better appreciate and assess the performance improvements in more detail. Therefore we assessed the performance of our Peak-only EI-MS predictors by analyzing the EI-MS spectra of selected individual molecules from our test set (see Figures 3.7-3.10). These mirrored EI-MS spectral graphs display NIST-reported EI-MS spectra at the top and predicted EI-MS results at the bottom. The predicted EI-MS are generated from the above-mentioned best performing models along with the current state-of-the-art CFM-EI model. The box in the top-right corner of each graph displays the dot product scores. This score is highlighted only when a particular model achieves the best possible dot product score. If we look at a low molecular weight compound (acetic acid, (ethylthio)-, ethyl ester) as an example (Figure 3.7a), we can see that the peak at m/z position 47 was correctly predicted in EI-MS alpha, beta and gamma (Figure 3.7b, c and d) but not for CFM-EI (Figure 3.7a). Likewise, the dot product score went up by 33 percent (CFM-EI model: 0.43 vs. EI-MS-gamma model: 0.59). A similar trend is also observed for an ester (adipic acid, heptadecyl 2-methylbutyl ester) (Figure 3.8) where the peak to 129 (with maximum relative intensity) was perfectly predicted in these models (Figure 3.8b, c and d) but were not predicted in CFM-EI (Figure 3.8a). Other examples included here show the predicted EI-MS spectra from an aldehyde (octanal, 2-(phenylmethylene)-) and a silylated molecule (styryltrimethylsilane) (Figure 3.9 and Figure 3.10), where both the predictors outperformed the previous CFM-EI predictor. For the aldehyde the dot product scores for the CFM-EI model: 0.47 vs. EI-MS beta: 0.61 vs. EI-MS gamma: 0.56. For the silyated molecule, the CFM-EI achieves: 0.34 vs. EI-MS beta: 0.35 vs. EI-MS gamma: 0.37). Overall, these examples nicely illustrate how the improvements to overall spectral prediction accuracy through these hybrid GNN+SVM models translate to more meaningful and informative EI-MS spectra.

3.3.3 Comparison of "Peak Only" EI-MS Prediction Models with Newer EI-MS Predictors

For many years CFM-EI was the “gold standard” for EI-MS spectral prediction. In 2019 the NEIMS program was published [70] and the authors claimed that it exhibited comparable performance and much faster calculation times than CFM-EI. Earlier in 2023 a new program called RASSP appeared that appeared to offer substantially better performance than both CFM-EI and NEIMS. The code to run this program has just become available in the last two months (after this project was nearly completed). Given these developments, we thought it would be important to compare the performance of our newly developed “Peak-only” predictors with these new predictors (RASSP and NEIMS) using the same hold-out data sets. Table 3.5 compares the prediction results of NEIMS and RAASP with our best performing peaks-only model (EI-MS gamma + molecular ion predictor) using the same holdout data set of 2,041 EI-MS spectra derived from NIST 20. The average dot product score is provided in the table for each of the chemical classes or categories. As seen from this table, NEIMS outperforms all our newly developed “Peak-only” predictors, including the recently published RASSP predictor, in terms of dot product scores for all chemical classes or categories (with the exception of aldehydes and low-molecular weight (<150 Da.) compound sets, where RASSP have shown significantly better results). It is also evident from the table that, both NEIMS and RASSP are not specialized for EI-MS prediction of derivatized or silylated molecules (dot product scores of EI-MS epsilon: 0.43 vs. NEIMS: 0.41 vs. RASSP: unable to process any silylated compounds). Finally, we also evaluated all these models using a set of 2,008 brand new molecules from NIST 23. It is to be noted that, NIST 23 consists of ~50,000 additional GC-MS spectra from the previous release (NIST 20) and none of these aforementioned models were trained, validated or tested on the set of brand new compounds from NIST 23. For the sake of unbiased evaluation, it was necessary for us to evaluate all the models on these new set of

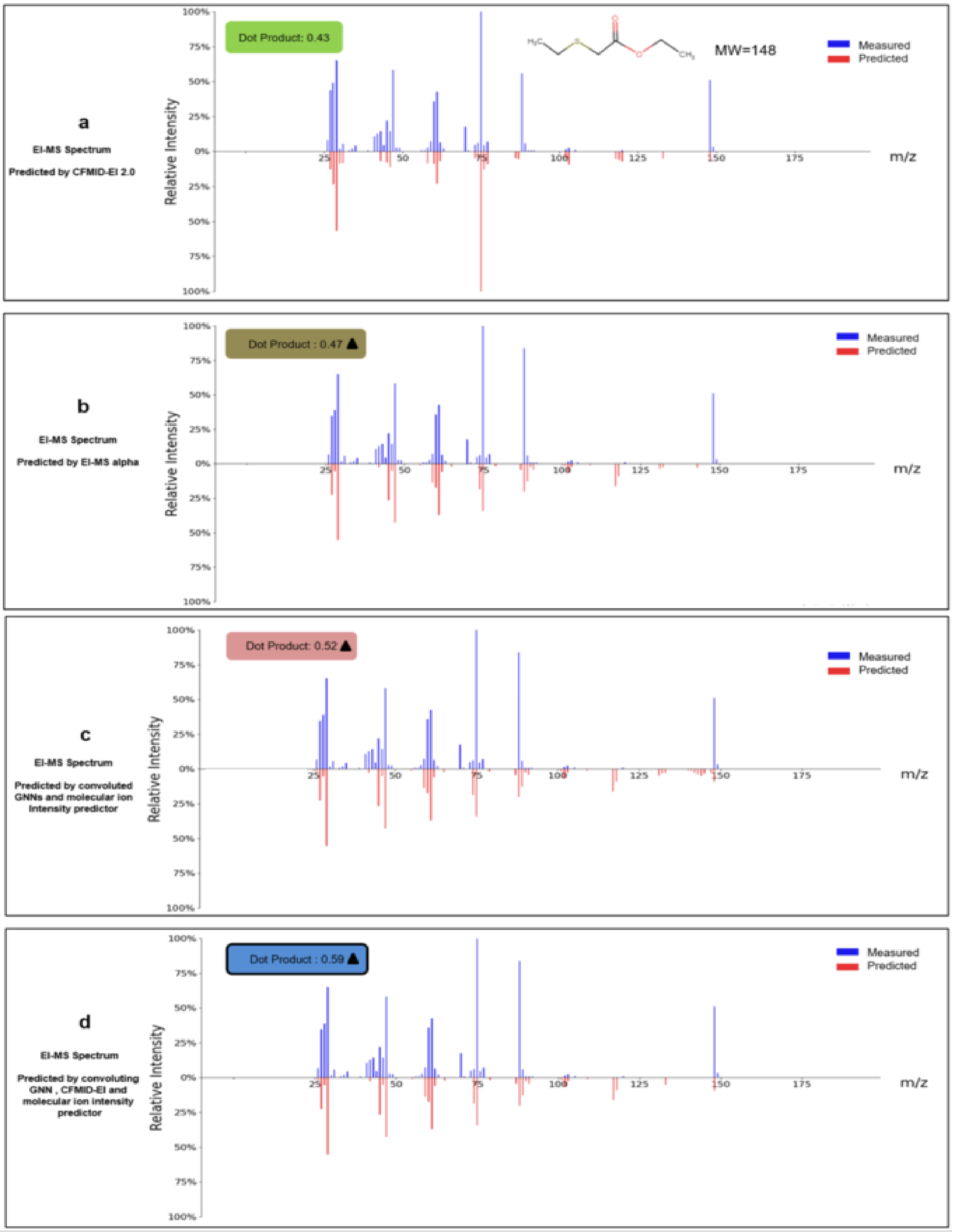


Figure 3.7: Plot showing the EI-MS ground truth vs. predicted spectra using various modules for a molecule ((ethylthio)-acetic acid, ethyl ester) in the Ester set.

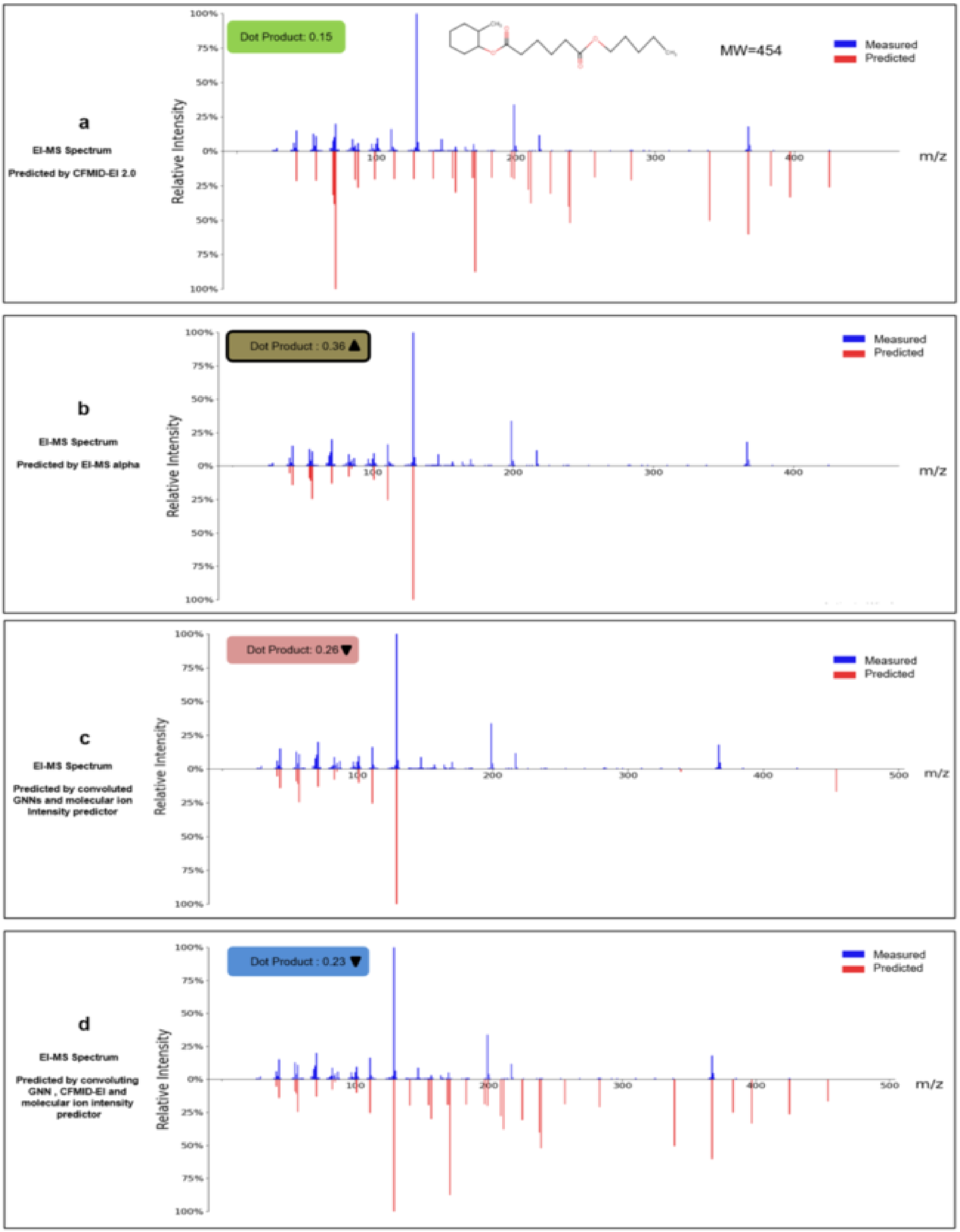


Figure 3.8: Plot showing the EI-MS ground truth vs. predicted spectra using various modules for a molecule (adipic acid, heptadecyl 2-methylbutyl ester) in the Ester set.

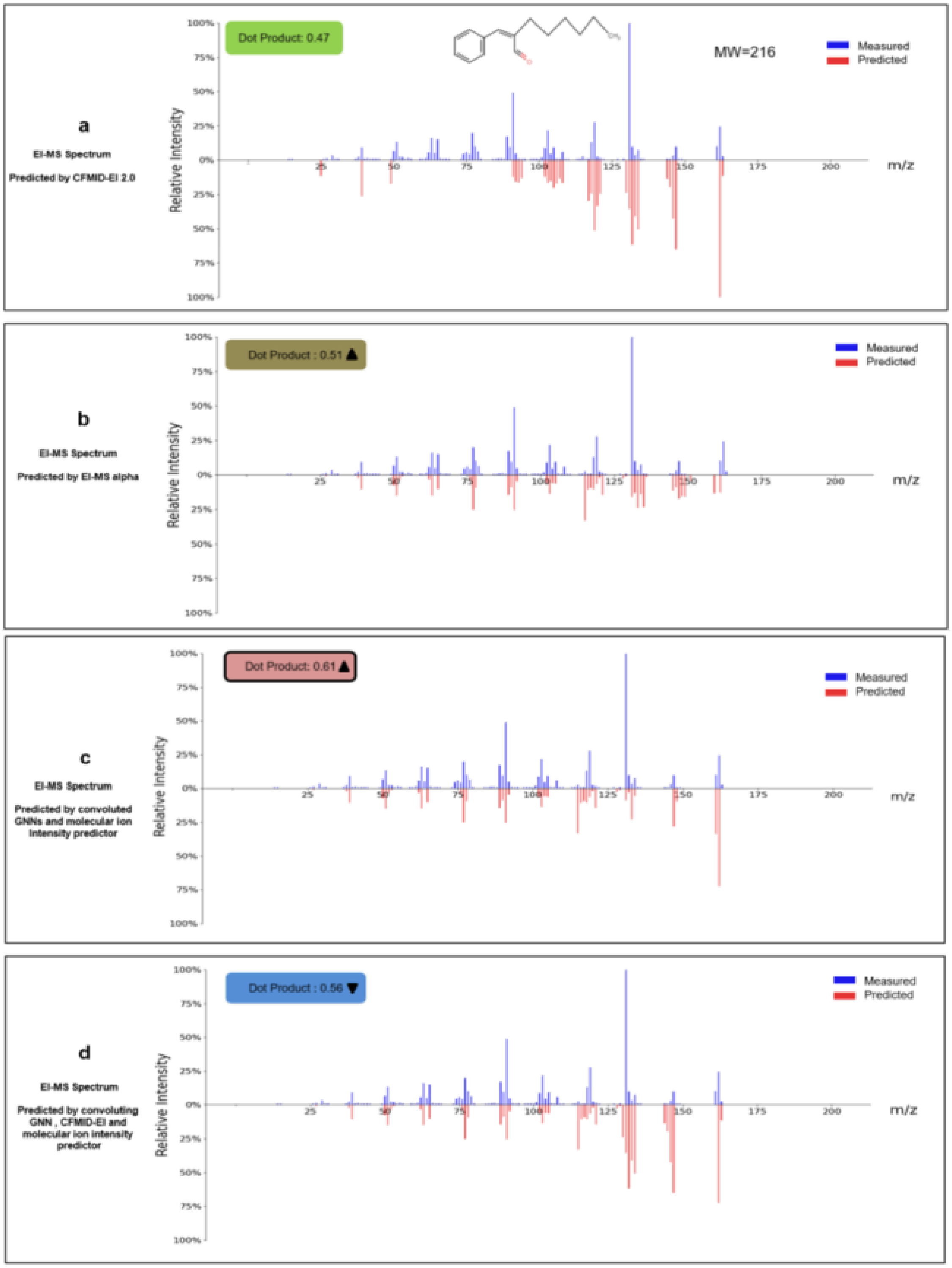


Figure 3.9: Plot showing the EI-MS ground truth vs. predicted spectra using various modules for a molecule (2-(phenylmethylene)-octanal) in the Aldehyde set.

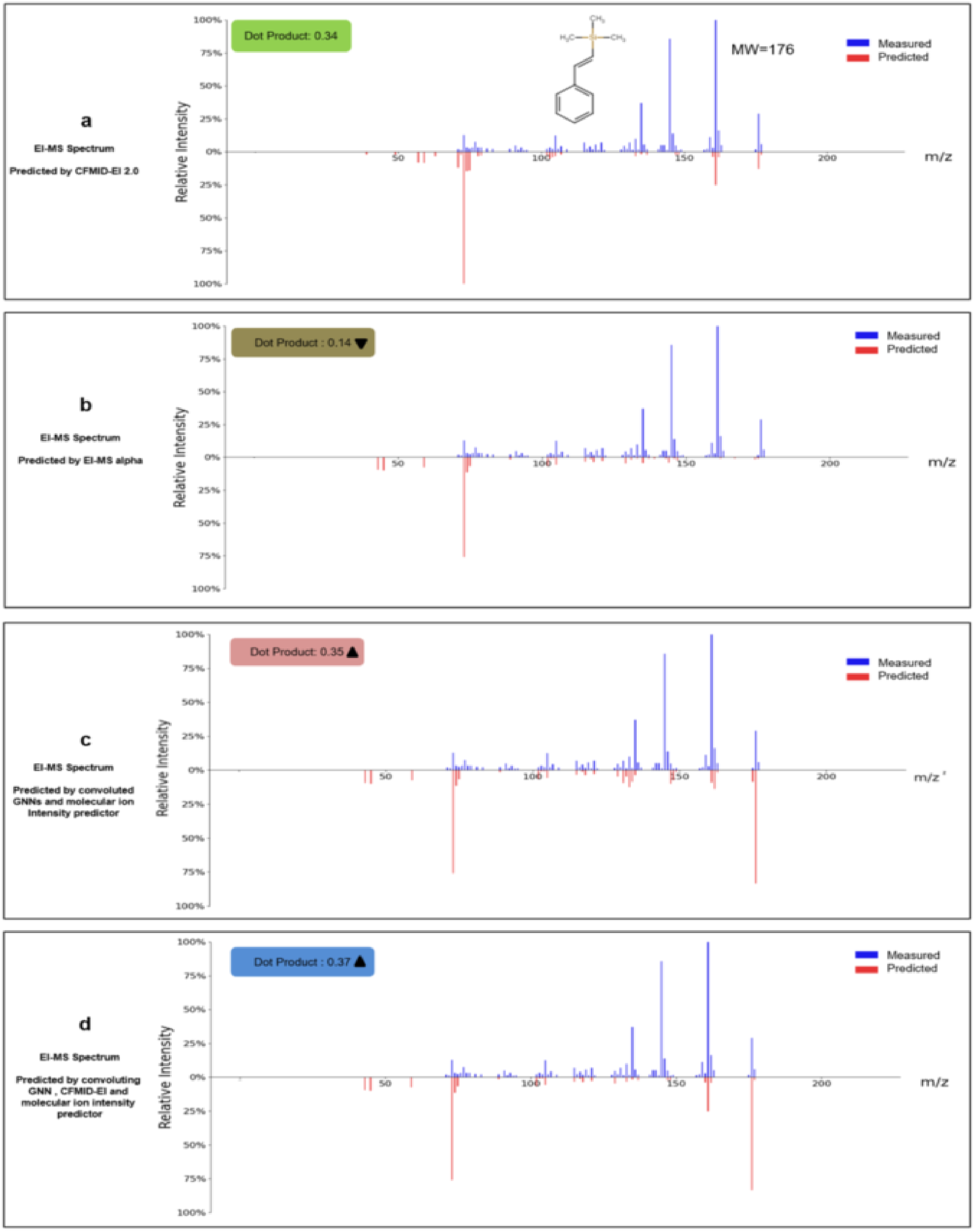


Figure 3.10: Plot showing the EI-MS ground truth vs. predicted spectra using various modules for a molecule (styryltrimethylsilane) in the silylated set.

compounds. As presumed, the dot product scores of all the models using the NIST 23 set were similar (with the exception of NEIMS having slightly reduced dot product score) to their weighted average dot product scores (reported in the second last row of Table 3.5) from different classes or categories.

3.3.4 Comparison of "Peak Only" EI-MS Prediction Models with the Adjusted NEIMS with PA and MIIP

As seen from the above results, none of the developed "peak-only" predictors were able to compete with NEIMS in terms of EI-MS spectra prediction. However, these predictors provided several advantages which was incorporated with the original NEIMS program to achieve better performance with subformulae annotation for each predicted peaks. We assessed the performance of this finalized predictor, called the Adjusted NEIMS with MIIP and PA (called EI-MSpred), on a set containing six common compounds (methanol, hexane, 3-methyl pentane, benzene, toluene, and nitrobenzene) and the NIST 23 compounds. We observed that making such an adjustment to the NEIMS program can even improve the performance of NEIMS. When tested on the six common compound set we achieved an average dot product score of 0.555 (vs. original NEIMS with common compound set: 0.517). When tested with the NIST 23 test set, we observed a slight but statistically insignificant improvement over the original NEIMS program (dot product score of original NEIMS: 0.62 vs. Adjusted NEIMS with PA and MIIP: 0.621 shown in Table 3.5). This finalized predictor (EI-MSpred) is used throughout the rest of this document to make EI-MS predictions.

3.3.5 Performance of PeakAnnotator

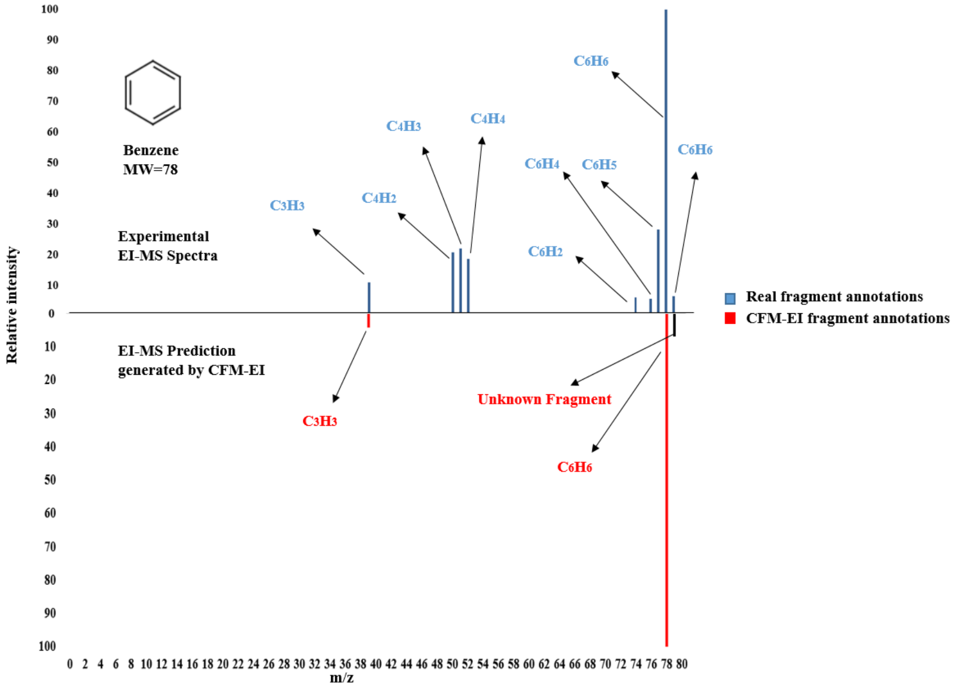
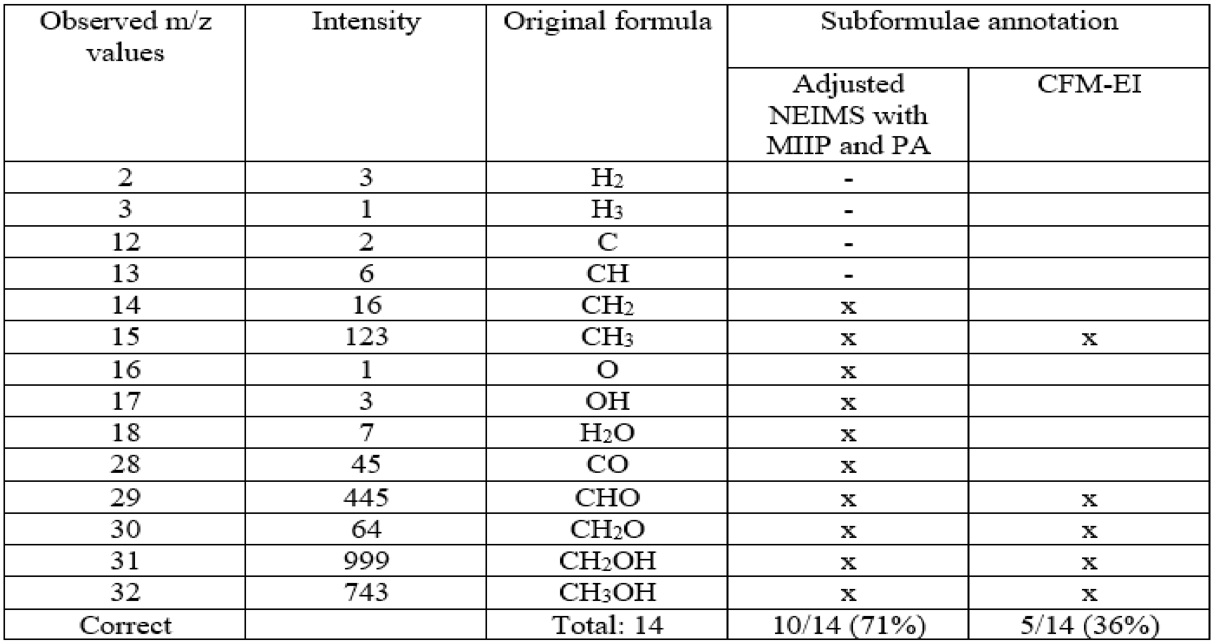


Figure 3.11: Mirror plot showing the observed peak annotations and peaks annotated by CFM-EI for a compound (benzene).

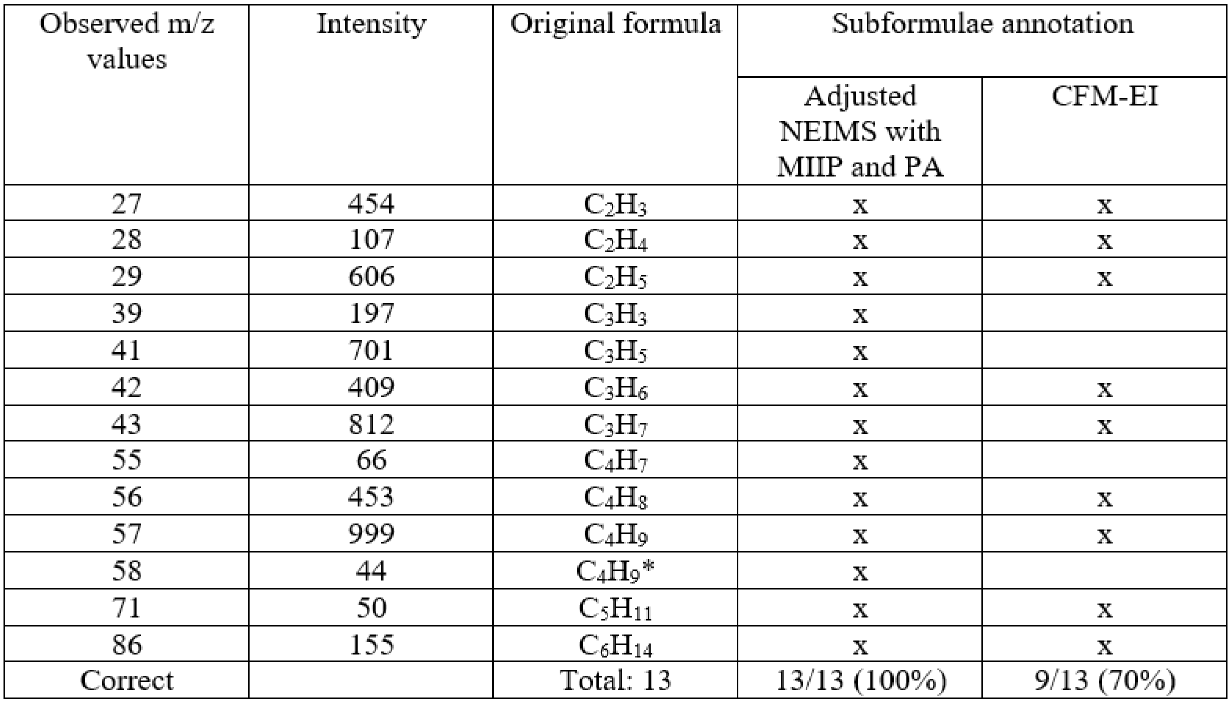
Early MS spectral predictors such as CFM-ID, CFM-EI, MagMA and MetFrag used machine-learned or combinatorial methods to generate possible fragmentation structures. This virtual fragmentation process helped to rationalize the actual fragmentation process and also allowed MS peaks in both predicted and experimentally collected MS spectra to be annotated. However, not all peaks in EI-MS or MS/MS spectra can be rationalized with structures or fragments generated by these prediction programs. For instance, only 20-25% of the peaks in a given EI-MS spectrum (for the compound "benzene") can be annotated by CFM-EI (see Figure 3.11). Furthermore, many of these peaks are often "hallucinatory" and do not exist in real spectra. PeakAnnotator was developed to help overcome the limited peak coverage of CFM-EI and the complete lack of peak annotation by EI-MS prediction programs such as NEIMS. The PeakAnnotator algorithm combinatorially determines peak formulae by calculating what combinations of atoms (consistent with the molecular formula of the parent compound of interest) can yield positively charged and chemically viable structures. It also incorporates other rules regarding valency, the nitrogen rule [113], the senior rule [134], and rules regarding the degree of allowed unsaturation [135] to determine viable molecular formulae. PeakAnnotator also uses a hand-built knowledgebase of frequently known EI-MS peak patterns for well-known functional groups and structures (Table 3.1) to further refine the formula generation process.

To assess the performance of PeakAnnotator we selected six EI-MS spectra from compounds (mentioned in section 3.3.4) that have been extensively annotated via experimental studies, vetted by experts in GC-MS and theoretically confirmed by QCEIMS [61-62]. The compounds are methanol, hexane, 3-methylpentane, benzene, toluene, and nitrobenzene. These expert-driven molecular formula annotations along with the observed m/zs and intensities are shown in Tables 3.6-3.11 (columns one to three). PeakAnnotator (PA), combined with the original NEIMS and the molecular ion intensity predictor (MIIP) program, also known as EI-MSpred (or Adjusted NEIMS with PA and MIIP (EI-MSpred)), was used to annotate the unlabeled EI-MS peaks for these molecules using the experimentally observed EI-MS spectra. The molecular formula annotations generated via EI-MSpred and those generated via experts/QCEIMS are also shown in Tables 3.6-3.11). Several of these reference annotations suggest the existence of a ^{13}C isotope, which are labelled by “*” within the original formula (Table 3.7-3.10). The annotation is tagged with either of these four notations with an “x” representing a peak that is correctly predicted and correctly annotated , “-” representing a peak that is not predicted and not assigned, while “x but incorrect” represents a peak that is predicted but the annotation assignment was wrong.



Observed m/z values	Intensity	Original formula	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
2	3	H_2	-	
3	1	H_3	-	
12	2	C	-	
13	6	CH	-	
14	16	CH_2	x	
15	123	CH_3	x	x
16	1	O	x	
17	3	OH	x	
18	7	H_2O	x	
28	45	CO	x	
29	445	CHO	x	x
30	64	CH_2O	x	x
31	999	CH_2OH	x	x
32	743	CH_3OH	x	x
Correct		Total: 14	10/14 (71%)	5/14 (36%)

Table 3.6: Annotation correctness using EI-MSpred and CFM-EI for the compound methanol.



Observed m/z values	Intensity	Original formula	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
27	454	C_2H_3	x	x
28	107	C_2H_4	x	x
29	606	C_2H_5	x	x
39	197	C_3H_3	x	
41	701	C_3H_5	x	
42	409	C_3H_6	x	x
43	812	C_3H_7	x	x
55	66	C_4H_7	x	
56	453	C_4H_8	x	x
57	999	C_4H_9	x	x
58	44	C_4H_9*	x	
71	50	C_5H_{11}	x	x
86	155	C_6H_{14}	x	x
Correct		Total: 13	13/13 (100%)	9/13 (70%)

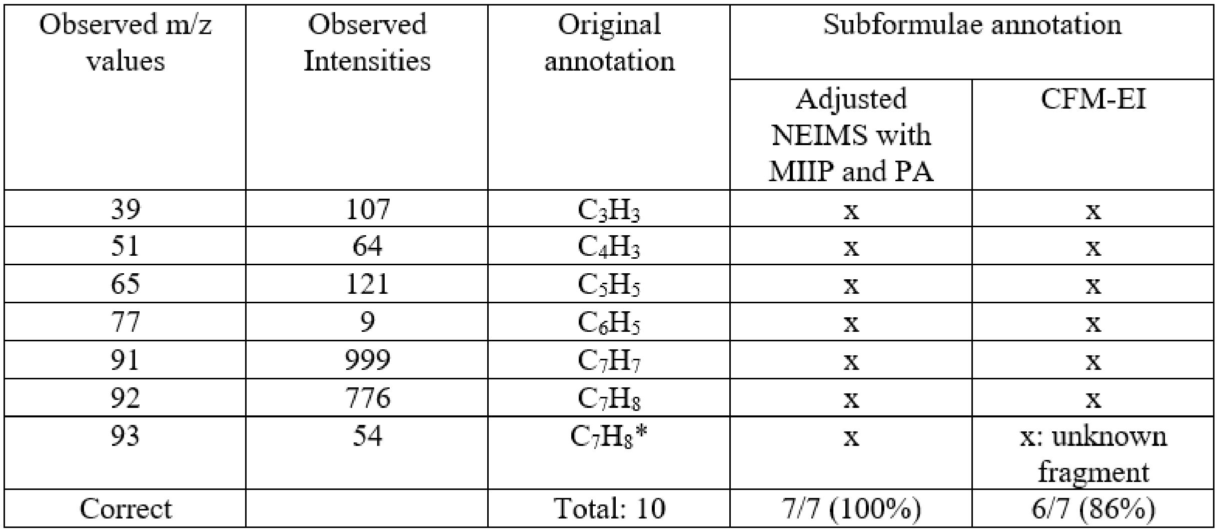
Table 3.7: Annotation correctness using EI-MSpred and CFM-EI for the compound hexane.

Observed m/z values	Intensity	Original formula	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
27	132	C_2H_3	x	
29	390	C_2H_5	x	x
39	91	C_3H_3	x	
41	533	C_3H_5	x	x
43	253	C_3H_7	x	
55	66	C_4H_7	x	
56	766	C_4H_8	x	x
57	999	C_4H_9	x	x
58	44	C_4H_9*	x	
71	56	C_5H_{11}	x	x
86	29	C_6H_{14}	x	x
Correct		Total: 11	11/11 (100%)	6/11 (54%)

Table 3.8: Annotation correctness using EI-MSpred and CFM-EI for the compound 3-methylpentane.

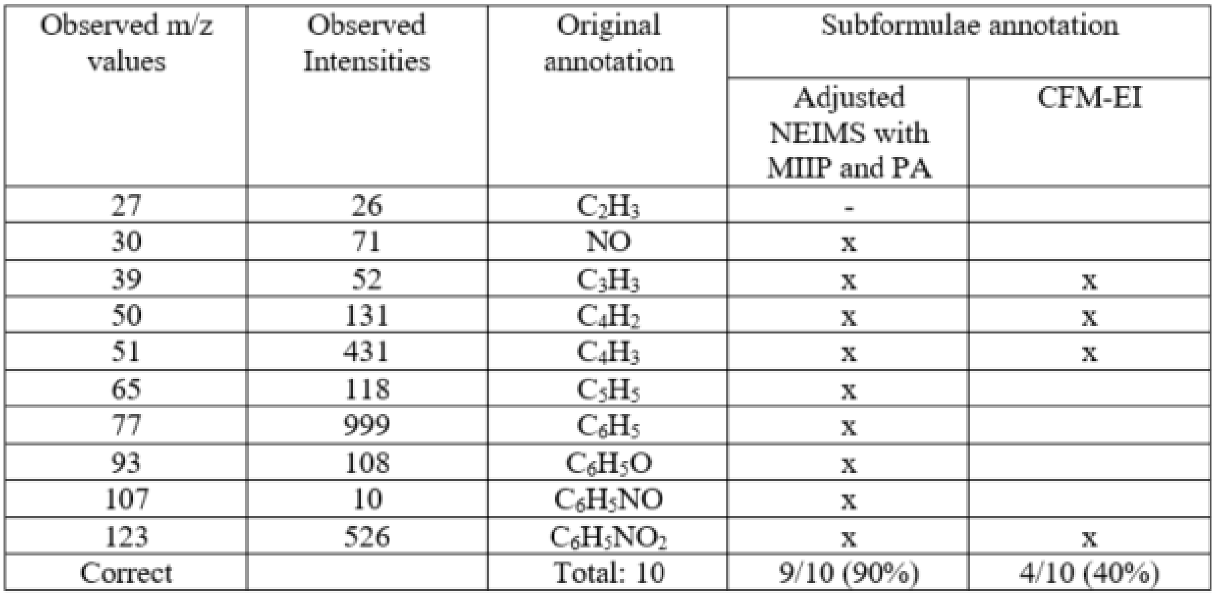
Observed m/z values	Intensity	Original formula	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
39	111	C_3H_3	x	x
50	208	C_4H_2	x	
51	221	C_4H_3	x	
52	188	C_4H_4	x	
74	62	C_6H_2	x	
76	58	C_6H_4	x	
77	283	C_6H_5	x	
78	999	C_6H_6	x	x
79	65	C_6H_6*	x	x: unknown fragment
Correct		Total: 9	9/9 (100%)	3/9 (33%)

Table 3.9: Annotation correctness using EI-MSpred and CFM-EI for the compound benzene.



Observed m/z values	Observed Intensities	Original annotation	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
39	107	C_3H_3	x	x
51	64	C_4H_3	x	x
65	121	C_5H_5	x	x
77	9	C_6H_5	x	x
91	999	C_7H_7	x	x
92	776	C_7H_8	x	x
93	54	C_7H_8*	x	x: unknown fragment
Correct		Total: 10	7/7 (100%)	6/7 (86%)

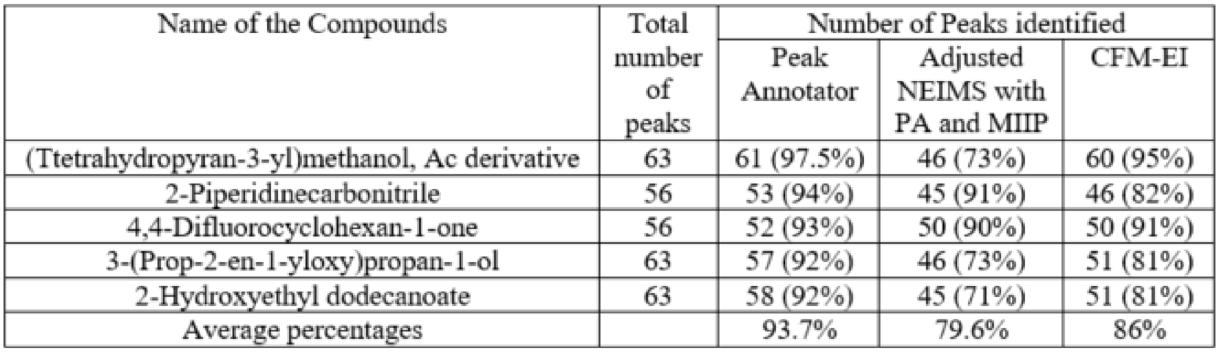
Table 3.10: Annotation correctness using EI-MSpred and CFM-EI for the compound toluene.



Observed m/z values	Observed Intensities	Original annotation	Subformulae annotation	
			Adjusted NEIMS with MIIP and PA	CFM-EI
27	26	C_2H_3	-	
30	71	NO	x	
39	52	C_3H_3	x	x
50	131	C_4H_2	x	x
51	431	C_4H_3	x	x
65	118	C_5H_5	x	
77	999	C_6H_5	x	
93	108	C_6H_5O	x	
107	10	C_6H_5NO	x	
123	526	$C_6H_5NO_2$	x	x
Correct		Total: 10	9/10 (90%)	4/10 (40%)

Table 3.11: Annotation correctness using EI-MSpred and CFM-EI for the compound nitrobenzene.

annotation representing no annotation found. This suggests that the PeakAnnotator program is only able to annotate a predicted peak with non-zero intensity. Inspecting these tables, it is apparent that EI-MSpred (based on PeakAnnotator program) correctly identifies 91% of the peaks with the correct peak formula. On the other hand, annotating the same spectra through CFM-EI we see that only a fraction of these peaks are annotated and even fewer of these are correctly annotated (53%).



Name of the Compounds	Total number of peaks	Number of Peaks identified		
		Peak Annotator	Adjusted NEIMS with PA and MIIP	CFM-EI
(Ttetrahydropyran-3-yl)methanol, Ac derivative	63	61 (97.5%)	46 (73%)	60 (95%)
2-Piperidinecarbonitrile	56	53 (94%)	45 (91%)	46 (82%)
4,4-Difluorocyclohexan-1-one	56	52 (93%)	50 (90%)	50 (91%)
3-(Prop-2-en-1-yloxy)propan-1-ol	63	57 (92%)	46 (73%)	51 (81%)
2-Hydroxyethyl dodecanoate	63	58 (92%)	45 (71%)	51 (81%)
Average percentages		93.7%	79.6%	86%

Table 3.12: Annotation coverage using five randomly selected compounds from NIST23.

In a second trial were compared the annotation coverage (as opposed to annotation correctness) with five randomly chosen experimentally collected GC-MS spectra selected from our NIST 23 holdout set. Specifically, we ran CFM-EI, NEIMS and PeakAnnotator on all five spectra and measured the fraction of peaks that were given annotations (chemical formulae or structures). The results are given in Table 3.12. As can be seen in this table, PeakAnnotator achieved an average level of peak annotation coverage approaching 93.7%, while the level of peak annotation coverage for CFM-EI was 86%, for the Adjusted NEIMS with PA and MIIP (EI-MSpred) it was 79.6%, and for both NEIMS and RASSP it was 0%.

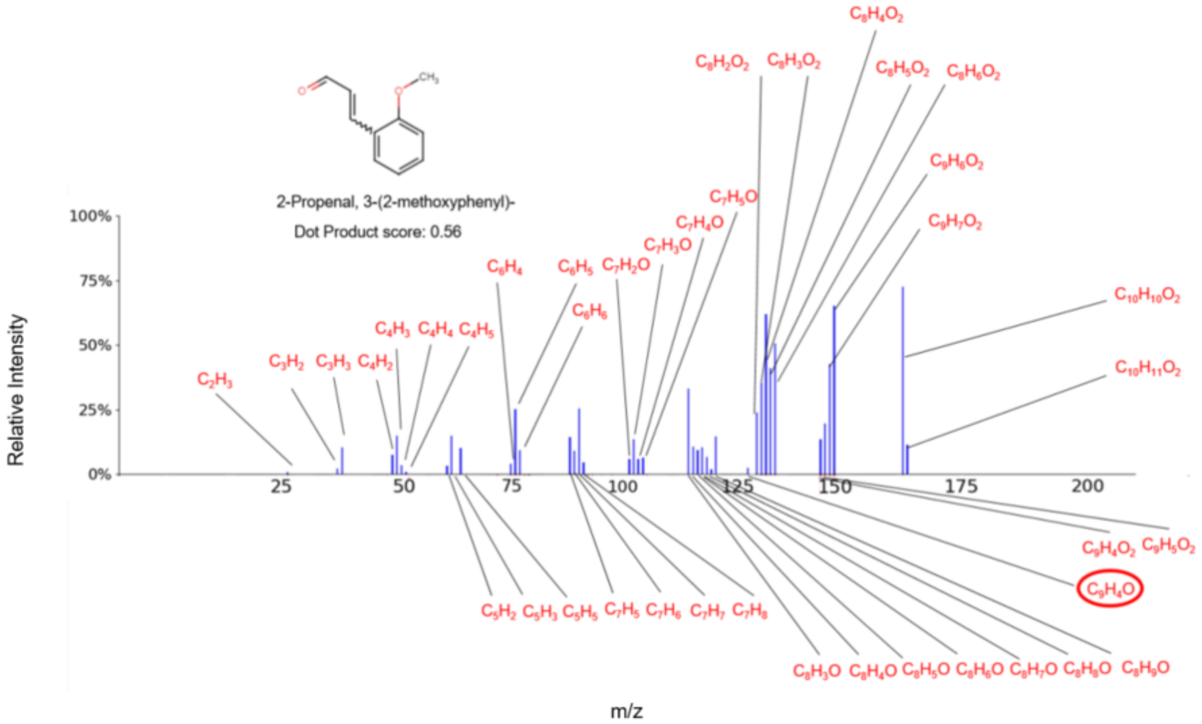


Figure 3.12: Annotation process showing the molecule 2-propenal,3-(2-methoxyphenyl)- and its peak subformulae.

In a third example to illustrate the utility of PeakAnnotator (with no EI-MS predictor combination), we show how a molecule (3-(2-methoxyphenyl)-2-propenal) can have its predicted EI-MS spectrum annotated by PeakAnnotator (see Figure 3.12) and subsequently corrected. As seen in the previous examples of a real EI-MS spectra, all the annotated chemical formulae in all the clusters of peaks should follow a particular sequence of elemental composition, which is generally observed in this figure. The exception is the predicted formula (C_9H_4O) at m/z = 128. This is an incorrect prediction firstly because the chemical formula does not match well with the chemical formulae of the adjacent peaks ($C_8H_2O_2$, $C_8H_3O_2$, etc.). Secondly, the only other formula that could possibly match this molecular weight is a neutral molecule C_8O_2, which is chemically impossible (and obviously not a positively charged ion). Therefore, we would conclude that this peak should not be observed in the experimentally observed EI-MS spectrum. This leads to the conclusion that this is an erroneous peak generated by the NEIMS prediction program. If we look at the actual EI-MS spectrum for 3-(2-methoxyphenyl)-2-propenal, we see this peak at m/z=128 is not present. Therefore, by using PeakAnnotator as a check on predicted EI-MS spectra we can effectively edit the predicted spectra and improve the prediction performance.

3.4 Conclusion

The prediction of EI-MS spectra using various computational techniques has seen tremendous advancements over the past 10 years. This is because there is growing quantity of high-quality EI-MS data with which to train, and substantial improvements in the computational methods – particularly with regard to deep learning to accurately predict EI-MS spectra. Until recently, the "gold standard" for EI-MS spectral prediction was the program known as CFM-EI. This program uses probabilistic graphical models (PGMs) to predict compound fragmentation patterns and probabilities. While CFM-EI provides useful structural information and annotation for predicted EI-MS spectra, it is not particularly fast nor is sufficiently comprehensive in its peak annotation coverage. This work describes the development and refinement of several deep-learning models aimed at improving both the accuracy and speed of EI-MS prediction. By making use of a blended approach that uses a GNN for low m/z spectral prediction, a PGM for high m/z prediction, an SVM to predict the molecular ion and a combinatorial method for peak annotation (that also uses CFM-EI) we were able to create a tool called EI-MS epsilon (or EI-MSE) that outperforms CFM-EI by 20-25% in terms spectral prediction as measured by the dot product score. We also found that different models performed better for different classes of compounds, suggesting that the creation of compound-class specific models could boost the performance even further.

At the same time, we also made adjustment to the original NEIMS program by appending our developed PeakAnnotator and molecular ion intensity predictor (MIIP) programs (or EI-MSpred). Comparisons of EI-MSpred to other recent EI-MS prediction programs showed that it was much better than CFM-EI, but was worse than both NEIMS and RASSP in terms of dot product score. However, EI-MSE program was able to accurately annotate approximately 94% of the peaks that it predicted, which is substantially better than CFM-EI (at 86%), NEIMS (at 0%) and RASSP (0%). Comparison of the EI-MSpred with other models showed that it was better than all other models including the RASSP model and showed slight improvement over the original NEIMS model.

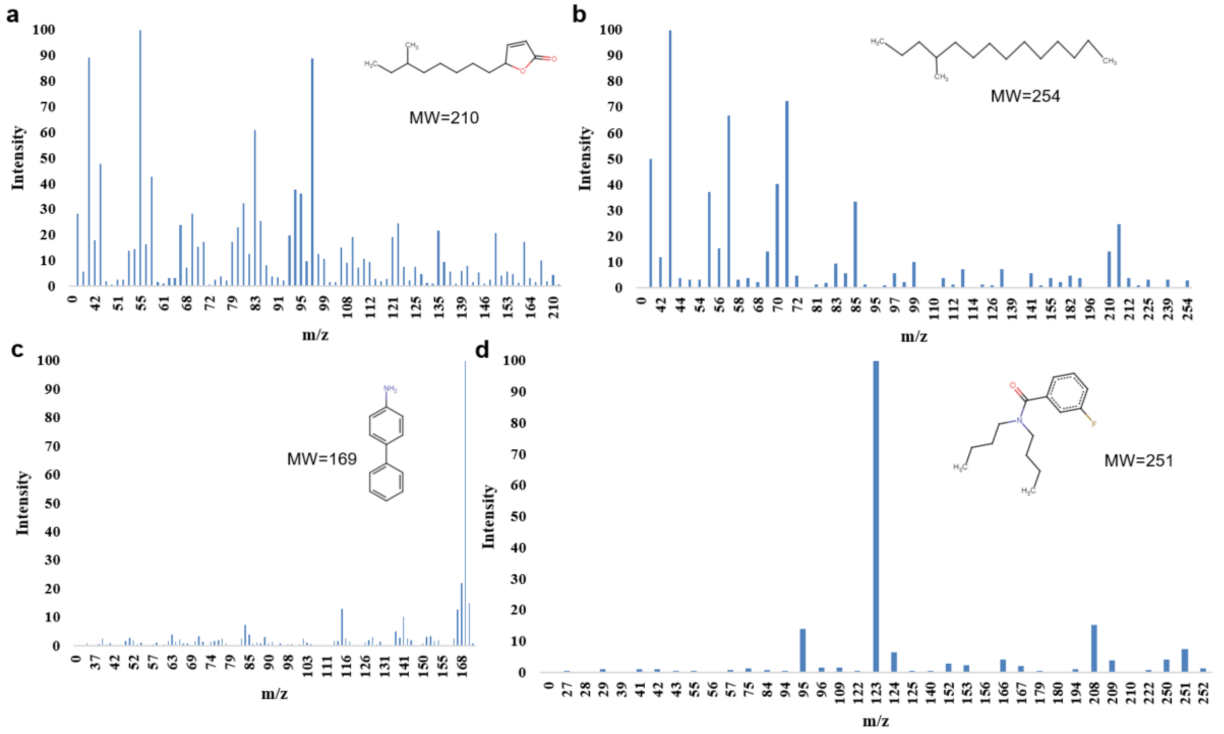
Using PeakAnnotator on the EI-MSpred predicted EI-MS spectra allowed >99% of the predicted EI-MS peaks to be annotated. Furthermore, by judiciously applying PeakAnnotator to predicted EI-MS spectra, we showed that it was possible to further improve the quality of the predictions. We believe that the EI-MS tools described here (EI-MSE and EI-MSpred) could be used in conjunction with retention index prediction tools, such as RIpred [136], to greatly facilitate compound identification by GC-MS. Indeed, the predictive performance of these two techniques (EI-MS prediction and RI prediction) suggests that GC-MS may be far more predictable and far more amenable to *in silco* or computer-based compound identification than LC-MS.

All of the EI-MS predictors (EI-MS alpha, beta and gamma, delta and epsilon, EI-MSpred) are available on GitHub (https://github.com/Afia-Anjum/ei_ms_pred) along with the molecular ion predictor (MIIP) and the PeakAnnotator. Also, PeakAnnotator to tag new training set is included. Directions for their download and installation are provided on the GitHub page.

Chapter 4: Conclusion and Future Directions

4.1 Thesis Overview

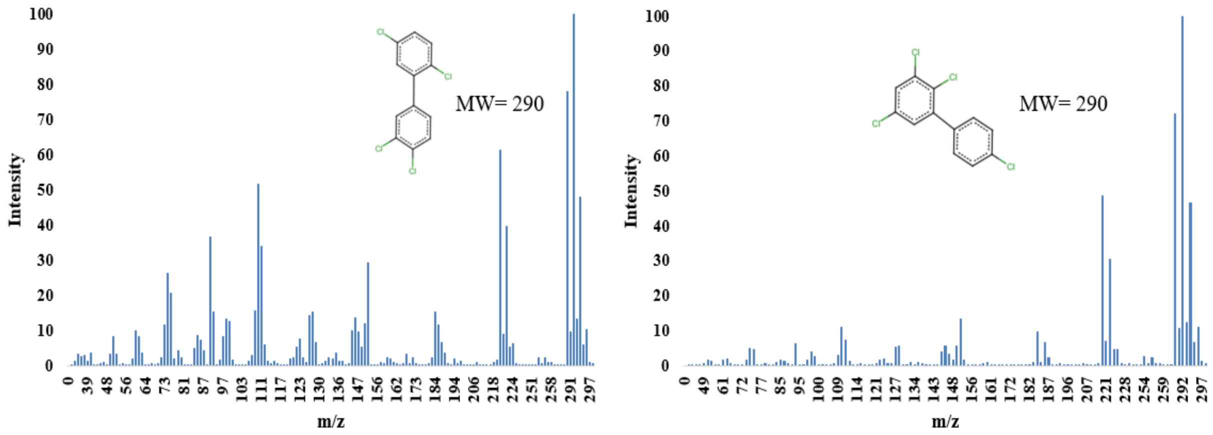
Gas chromatography mass spectrometry (GC-MS) uses a combination of gas-phase-based separation with electron impact mass spectrometry (EI-MS) to characterize and identify volatile and semi-volatile compounds. Relative to other forms of mass spectrometry, such as liquid chromatography mass spectrometry (LC-MS) or direct injection mass spectrometry (DI-MS), GC-MS is highly standardized and very reproducible. This makes it a very effective analytical technique for detecting and quantifying compounds (also called analytes) from complex mixtures. The output of a GC-MS experiment typically produces two measurable parameters for each detected feature or analyte: a retention time (normalized to a retention index or RI) and an EI-MS spectrum. The RI value corresponds to the relative amount of time it takes for the analyte to travel through the column while the EI-MS spectrum captures information about the atomic composition and molecular structure of the analyte. The RI is largely determined by the size, boiling point and hydrophobicity of the analyte, while the EI-MS spectrum is determined by the molecular structures of the molecule. Large databases consisting of tens of thousands of experimentally measured RIs and experimentally measured EI-MS spectra have already been created [36, 113, 114, 115, 116, 117]. These databases allow researchers to compare their own GC-MS data (RI and/or EI-MS spectra) against these reference GC-MS databases and identify potential matches. However, RI comparisons with, or without, EI-MS spectral comparisons rarely produce perfect matches. Furthermore, these comparisons can be very manually intensive, are often slow, tedious and prone to error and can be subject to various complications. For instance, multiple compounds with the same RI can have very different molecular weights and completely different EI-MS spectra (Figure 4.1). Similarly, it is possible for two stereoisomers (i.e., compounds having same molecular formulas and molecular weights but differing in the spatial orientation of functional groups) to have nearly identical EI-MS spectra but very different RI values (Figure 4.2).



NIST 20 Reported RI: 1758
Stationary Phase: Semi Standard Non Polar

Figure 4.1: EI-MS spectra of four compounds a)10-methyldodec-2-en-4-olide , b) 4-methyl-heptadecane, c) [1,1'-biphenyl]-4-amine, and d) N,N-dibutyl-3-fluoro-benzamide which share the same RI values but which have completely different structures, totally different molecular weights and entirely different EI-MS spectra.

Therefore, attempting to identify a compound from its RI data alone or its EI-MS data alone can be fraught with error. That's why GC-MS identification invariably requires matching both RI and EI-MS data together. Unfortunately, only a tiny fraction of known chemicals (and certainly



NIST 20 Reported RI: 2041

NIST 20 Reported RI: 2018

Stationary Phase: Semi Standard Non Polar

Figure 4.2: EI-MS of the two stereoisomers a) 2,3’,4’,5-tetrachloro-1,1’-biphenyl, and b) 2,3,4’,5-tetracholoro-1,1’-biphenyl having the same molecular formula, weights and EI-MS spectra but different RI values within the same stationary phase.

no unknown compounds) have their RI values or EI-MS spectra deposited in these reference databases. This limited coverage makes the identification of many well-known compounds difficult, and the identification of all unknown compounds, completely impossible via spectral database comparison. However, if the RI values and/or EI-MS spectra of a compound could be accurately predicted from a compound’s known structure (or even a hypothesized structure), then compound identification via GC-MS could be greatly improved and the chemical coverage of GC-MS greatly increased. It is this challenge of developing accurate methods to predict RI and EI-MS spectra that defined the two main objectives of my thesis. As described in the previous two chapters (Chapters 2 and 3), I believe I have nearly succeeded. In the next section, I will use several toy problems to mimic a typical GC-MS metabolomics analysis and then describe the effectiveness of my developed RI and EI-MS predictors for compound identification.

4.2 Combining RI and EI-MS Predictions To Identify Compounds Via GC-MS

In this section, I will demonstrate how predicted GC-MS data including RI and EI-MS data predicted via my predictors, can aid in compound identification. For this example, I have chosen three compounds from the HMDB (the human metabolome database) where experimental GC-MS data has been already collected. Their reported RI values, as measured in semi-standard non polar (SSNP) column, and EI-MS data using a single quadrupole MS instrument with 70 eV collision energy along with the actual compound structures are listed in Table 4.1 and Figures 4.3a, b, and c. To simulate an actual compound identification task, we will pretend that these compounds are "unknowns" (i.e. we do not know their structures) and will use their experimentally reported GC-MS data to perform the compound identification task in a simulated metabolomics experiment. To perform the compound identification task, we first utilized the RI search module from my RIpred server (Chapter 2) to generate a list of potential matching compound structures. Recall that the RIpred server's backend is populated with predicted RI values for all 250,000 HMDB compounds in various stationary phases and derivatization types (amounting to >5 million RI values). As an input to the search module, for all the three unknown samples, we provided the reported RI values, the column information (semi-standard non-polar stationary phase) and set an RI tolerance of 2%, along with a mass tolerance of 3 Da. assuming the highest m/z in the EI-MS was the parent ion mass of the unknown samples. This search against the HMDB generated a list of 131, 150, and 79 matching entries, respectively, for the unknown samples 1, 2, and 3. To further reduce this list of potential candidates, we utilized our best performing EI-MS predictor (reported in chapter 3 - the

Adjusted NEIMS with Peak Annotator and Molecular Ion Intensity Predictor or EI-MSpred) to predict the EI-MS spectra of the 350 HMDB hits for these three unknowns. Here, I report the top five hits for each of the three queries sorted on the basis of the dot product score (Tables 4.2, 4.3, and 4.4) between the measured and predicted EI-MS spectra. Mirror plots of the top scoring hits between the measured and predicted EI-MS spectra for each of the three "unknowns" are shown in Figure 4.4a, b, and c.

Sample description	Measured Retention Index in SSNP
Unknown 1 (Dopamine, 4TMS)	2063
Unknown 2 (L-glutamic acid diethyl ester)	1423
Unknown 3 (20alpha-Dihydroprogesterone)	2912

Table 4.1: Measured retention indices of the three unknown samples in the SSNP phase.

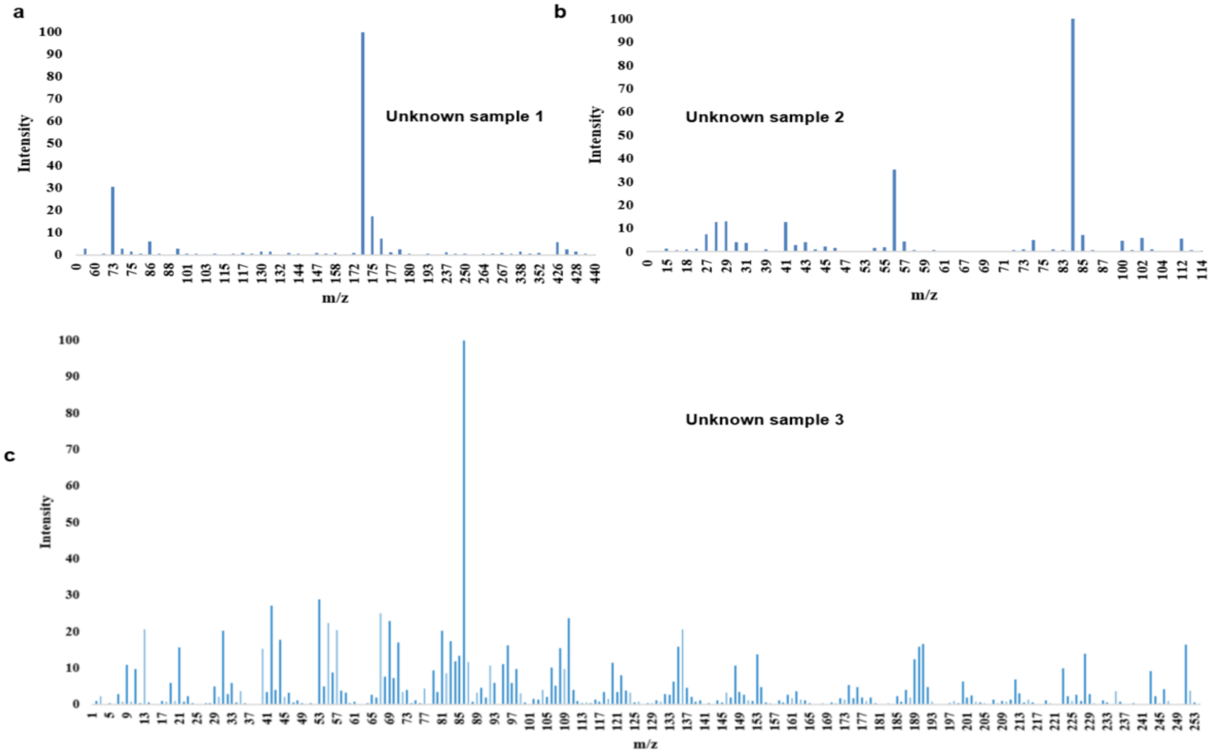


Figure 4.3: EI-MS spectra of the three unknown compounds.

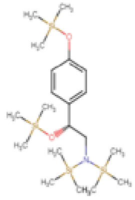
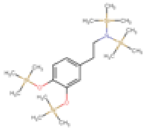
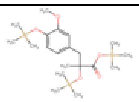
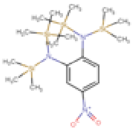
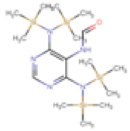
Structure Name	Chemical Structure	Derivatization Type	Average Mass	Retention Index	Dot Product Score
p-Octopamine,4TMS,isomer#1		TMS	441.24	2047.15	0.64
Dopamine,4TMS,isomer#1		**TMS**	**441.24**	**2063.81**	**0.60**
3-(4-Hydroxy-3-methoxyphenyl)-2-methyllactic acid,3TMS,isomer#1		TMS	442.20	2014.22	0.22
4-Nitro-o-phenylenediamine,4TMS, isomer#1		TMS	441.21	2064.59	0.17
FAPy-adenine,4TMS,isomer#1		TMS	441.22	2083.17	0.14

Table 4.2: Top five queried hits sorted on the basis of dot product score for “Unknown Sample 1 (Dopamine, 4TMS)”.

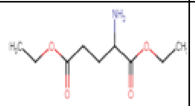
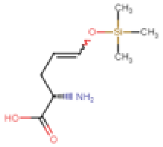
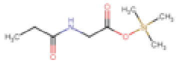
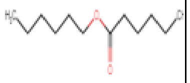
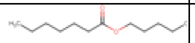
Structure Name	Chemical Structure	Derivatization Type	Average Mass	Retention Index	Dot Product Score
Glutamic acid diethyl ester		**Un-derivatized**	**203.1**	**1440.1**	**0.76**
L-Glutamic gamma-semialdehyde,1TMS,isomer#2		TMS	203.1	1457.63	0.31
Propionylglycine,1TMS,isomer#1		TMS	203.1	1381.98	0.29
Hexyl hexanoate		Un-derivatized	200.2	1409.92	0.28
Pentyl heptanoate		Un-derivatized	200.2	1410.56	0.28

Table 4. 3: Top five queried hits sorted on the basis of dot product score for "Unknown Sample 2 (L-glutamic acid diethyl ester)".

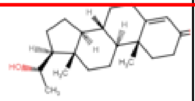
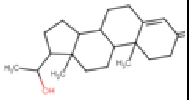
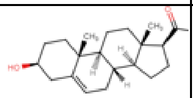
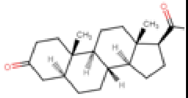
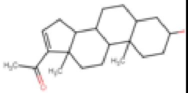
Structure Name	Chemical Structure	Derivatiza-tion Type	Average Mass	Retention Index	Dot Product Score
20alpha-Dihydroprogesterone		**Un-derivatized**	**316.24**	**2941.53**	**0.86**
20-Hydroxypregn-4-en-3-one		Un-derivatized	316.24	2867.23	0.86
Pregnenolone		Un-derivatized	316.24	2842.63	0.75
5a-Pregnane-3,20-dione		Un-derivatized	316.24	2859.18	0.74
(3beta,5alpha)-3-Hydroxypregn-16-en-20-one		Un-derivatized	316.24	2897.25	0.73

Table 4.4: Top five queried hits sorted on the basis of dot product score for "Unknown Sample 3 (20alpha-Dihydroprogesterone)".

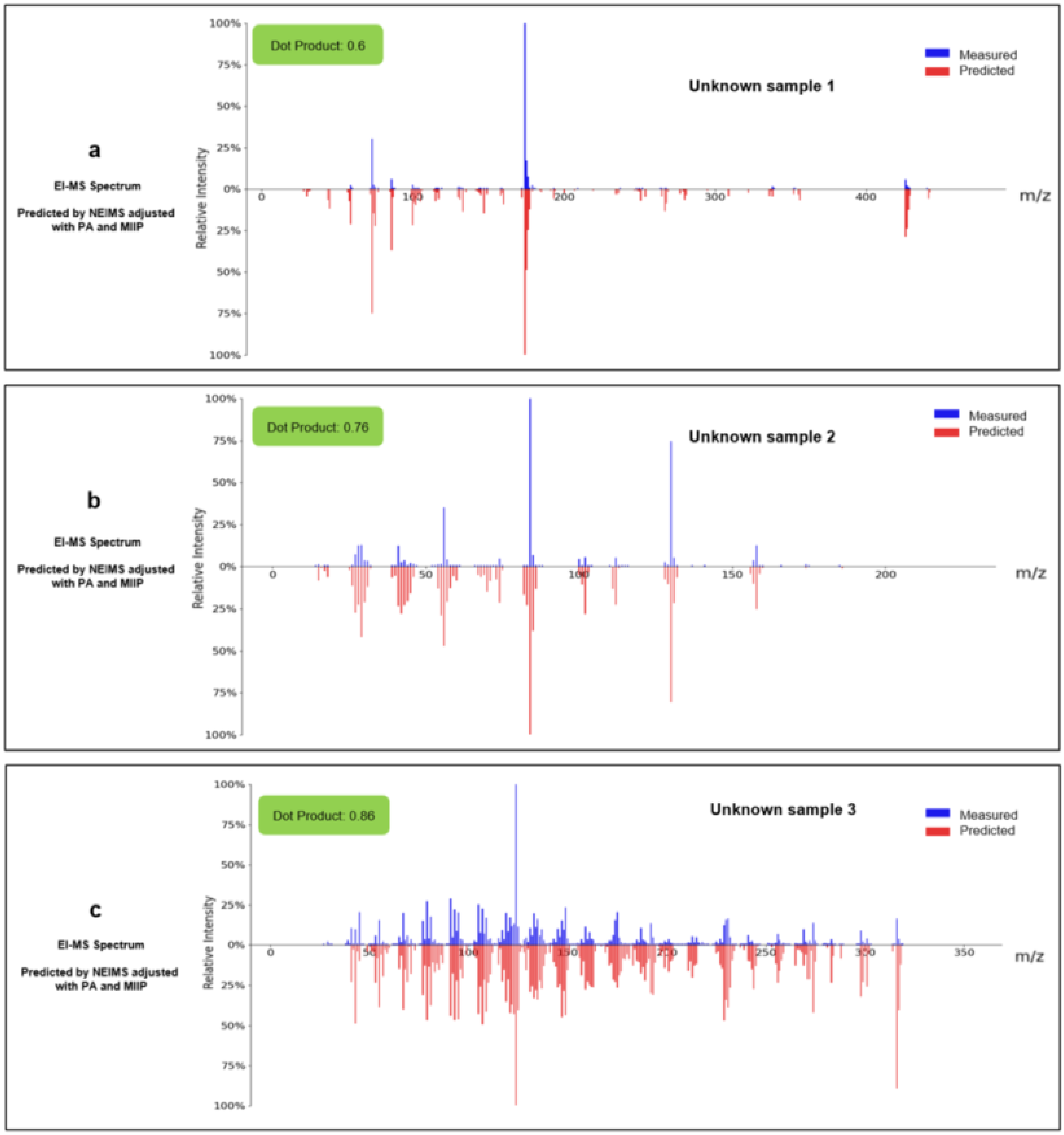


Figure 4.4: Mirror plots of the tops scoring hits between the measured and predicted EI-MS spectra corresponding to a) unknown sample 1, b) unknown sample 2, and c) unknown sample 3.

As seen from these tables and figures, the dot product scores corresponding to the best hits for these three samples are 0.64, 0.76, and 0.86 respectively. Unknown samples 2 and 3 match correctly to the actual compounds “Glutamic acid diethyl ester” and “20-alpha-Dihydroprogesterone”, however, sample 1 matches to the compound “p-Octopamine, 4TMS,isomer#1”, which is a diastereomer of the actual compound “Dopamine, 4TMS,isomer#1” (with second best hit and dot product score for dopamine, 4TMS: 0.60 v.s. dot product score for Octopamine,4TMS: 0.64). This “modest failure” for sample 1 arose because mass spectral analysis is not able to make the distinction between stereoisomers. Given this fact, one would reasonably have had to propose two possible structures for sample 1 (Dopamine, 4TMS,isomer#1 and Octopamine, 4TMS,isomer#1). Additionally, even though the spectra dot products were not found to be the optimal (incase of octapine-dopamine example), the experimental and predicted RI values would perfectly allow us to distinguish between these two compounds (sample 1 measured RI: 2063 v.s. predicted RI: 2063.81 in the SSNP phase). Overall, these results show quite nicely how the predicted RI values and the predicted EI-MS spectra generated via my predictors can be used – in combination – to identify unknown compounds.

4.3 Summary and Assessment

This thesis spanned a total of four chapters. In chapter 1, I provided a brief review of different chromatography or chromatographic separation techniques. I specifically focused on explaining the principles of gas chromatography and the general concepts behind GC-MS. I introduced and explained a number of terms or concepts such as electron ionization (EI), retention time (RT), retention indices (RI), chemical derivatization, and chemical/mass fragmentation techniques, etc. Additionally, in chapter 1, I described in detail the standard route for compound identification via GC-MS analysis using reference library methods, by exploiting the AMDIS and NIST library search modules. After explaining how compounds are identified via GC-MS, I noted some of the key limitations of library-based methods. Specifically, I noted that the total number of experimental RI values and EI-MS spectra in these libraries are quite limited and that they do not allow the identification of the vast majority of known compounds and that they certainly cannot be used to identify unknown or hypothesized compounds. I therefore proposed that this problem could be addressed by developing methods that could accurately predict RI values and EI-MS spectra from chemical structure data (i.e., SMILES strings or MOL files). Specifically, I hypothesized that machine learning could be used to develop two predictors: RIpred (for retention index prediction) and EIMSpred (for EI-MS prediction). I also hypothesized that that it would be possible to automate the GC-MS compound identification process by combining these predictors together (if they were sufficiently accurate), to mimic the AMDIS/NIST compound ID process. Finally, I hypothesized that these developed predictors could be used to create a large *in silico* library of known (and predicted) compounds containing their predicted RI values and predicted EI-MS spectra and that such a library would most certainly aid in the compound identification of both knowns and unknowns.

To test this hypothesis, as described in Chapter 2, I first conduct a background analysis of all available RI predictors and evaluated their strengths and weaknesses. I soon realized that the RI prediction process might benefit from the incorporation of chemical derivatization information and from column stationary phase information. This is because, a compound undergoing chemical derivatization (via TMS or TBDMS) will exhibit distinct retention time changes due to the change in its polarity, volatility, solubility and thermal stability. Similarly, a compound’s interaction with a particular type of stationary phase or GC-column type will affect its retention time and retention index. As a result, I hypothesized that making specialized a set of specific RI predictors with six different training datasets (derivatized, underivatized (base form), X three stationary phases) selected from the NIST 20 library would improve the RI predictor's performance. I then developed a program that used RDKit and other sets of cheminformatics programs to convert chemical SMILES strings into molecular graphs and extract various molecular features. These molecular graphs (along with the accompanying RI information and selected molecular features) were then used as training data for a set of graph neural network (GNN) models for RI prediction. The best performing models for each of the six datasets were selected through training, testing, and optimizing based 10-fold cross validation methods. All these specialized RI predictors along with a program called "autosilylator" (to perform *in silico* TMS or TBDMS derivatization) were then embedded within a more comprehensive RI prediction pipeline and presented as a webserver called RIpred. Comparisons were conducted against RIpred and other statistical, heuristic or ML-based RI predictors and it was found that RIpred matched or outperformed nearly all of them in terms of accuracy and comprehensiveness. In particular, the performance of RIpred was similar (mean absolute percentage error (MAPE): within 3% and mean absolute error (MAE): <73 RI units) to the performance of the NIST 20 AI predictor – which is the gold standard. Furthermore, RIpred distinguished itself as being the first RI prediction tool made available as a free web server, the first tool to perform automated silylation and the first to handle multiple stationary phases. Likewise, latest release of the HMDB (HMDB 5.0) also contains RIpred generated RI predictions for ~5 million compounds. The work describing RIpred and its comparative performance was published as a paper in the Journal of Chromatography A in 2023 [136].

In Chapter 3 of this thesis, I turned to the problem of predicting EI-MS spectra. I hypothesized that several advancements could be made to improve the quality of EI-MS spectral predictions generated by other state-of-the-art models such as CFMID, NEIMS, and RASSP. As a part of this study, I utilized the NIST 20 EI-MS data as my training, validation and testing set and used GNNs to generate several of my own predictors such as EI-MS alpha, beta, gamma, delta, and epsilon. As done with my RI-prediction program, I utilized RDKit and other cheminformatics programs to convert chemical SMILES strings into molecular graphs and extract molecular features that could be used to predict EI-MS spectra. These molecular graphs, molecular features and EI-MS spectral data were fed into a variety of GNNs to train a model. Optimization with a carefully selected EI-MS spectral validation set and an early stopping protocol helped in selecting the best GNN models. While my own EI-MS predictors outperformed CFM-EI, they did not outperform NEIMS or RASSP. In terms of dot product score, the results were CFM-EI (the original program developed by Felicity Allen): 0.32 vs. EI-MS alpha: 0.41 vs. EI-MS beta: 0.42 vs. EI-MS gamma: 0.37 vs. EI-MS delta: 0.45 vs. EI-MS epsilon: 0.38 vs. vs. NEIMS: 0.62 vs. RASSP: 0.56 on a dataset containing ~2000 EI-MS spectra from the NIST 23 database. I also found that incorporation of a molecular ion intensity predictor (MIIP) and a peak annotation (PA) program could improve the EI-MS spectral prediction results. Specifically, I tested this idea with a set of six common compounds and the previous NIST 23 held out test set. Indeed, a notable increase in the performance was seen for the six common compounds (Adjusted NEIMS with PA and MIIP: 0.555 vs. Original NEIMS: 0.517). However, more extensive testing on the NIST 23 test did not reveal a significant improvement (Adjusted NEIMS with PA and MIIP: 0.621 vs. Original NEIMS: 0.62). However, in terms of spectral annotation correctness, we did achieve an average correctness of

91% with the Adjusted NEIMS with PA and MIIP (called EI-MSpred), while the annotation correctness for NEIMS, RASSP and CFM-EI on the same data sets were 0%, 0%, and 53 % respectively. In other words, the inclusion of my MIIP and PA programs as part of the NEIMS program appeared to not only improve the NEIMS performance, but also produced annotated EI-MS spectra – which is something that the NEIMS program was not capable of doing. In the end, I developed a hybrid program called "Adjusted NEIMS with MIIP and PA", which is now known as EI-MSpred.

As I noted in Chapter 2 and 3, I faced several challenges creating the RI and EI-MS predictors. One noteworthy challenge for the RI prediction task was the need to clean, carefully curate and consolidate the training sets for RI values. This is because many of the RI values in the training as well as the validation and testing sets, were incorrectly reported. In some cases, RI value mix ups between various stationary phases and derivatization process were present. Careful cleaning of all these datasets required many weeks of tedious effort and careful crosschecking. While the issues of data quality for EI-MS prediction were not as much of a concern as they were for the RI prediction, other challenges were present. It turned out that the modeling of compound fragmentation patterns was computationally much more difficult and that published methods for MS prediction had insufficient detail for me to properly replicate their results. I tried many alternative strategies and explored many heuristic fixes (such as separating light weight molecules from heavyweight molecules, creating annotated EI-MS spectra, developing molecular ion intensity predictors) that achieved partial success. Indeed, embedding the two additional programs, MIIP (molecular ion intensity predictor) and PA (PeakAnnotator), enhanced the overall quality of the EI-MS spectra prediction.

In Chapter 4, I wanted to demonstrate how the combination of RI prediction and EI-MS prediction could be used to facilitate compound identification in a "mock" metabolomics GC-MS experiment. To carry out the process, I selected three known compounds found in the human metabolome (in the HMDB) and used their experimentally reported EI-MS and RI data, assuming that they are "unknown" compounds. I used the estimated molecular ion m/z value and reported RI data as input for RIpred to narrow the list of candidates down to 350 compounds (from 250,000). I then predicted the EI-MS data using EI-MSpred and calculated the dot product scores against the reported EI-MS data for the three compounds. My result indicates that all three unknown samples can be identified with top spectral matching scores.

4.4 Contributions

Following is a recapitulation of the main contributions of this thesis:

- Development of updated and corrected Kovats RI libraries for >100,000 compounds for various settings and column standards.
- Development of a program called autosilylator that can generate all possible (ten to hundreds) silylation (TMS and TBDMS to be specific) for GC amenable compounds.
- Development of a RI predictor (RIpred) that predicts RI values for various column settings, derivatization type, and tolerances (mass, RI), etc.
- Development of a publicly available webserver (RIpred) for RI prediction which includes a database compound search module based on user inputs of m/z values and RI values.
- Development of an offline version of RIpred to predict RI values for all GC amenable compounds in HMDB 5.0.

- Development of multiple EI-MS predictors (EI-alpha, EI-beta, EI-gamma, EI-delta, EI-epsilon, etc.) to predict EI-MS spectra that consistently outperform CFM-EI.
- Development of a molecular ion intensity predictor that accurately predicts the intensity of the parent ion (molecular ion) peak in EI-MS spectra given the input structure. This MII predictor improved the performance of all EI-MS predictors.
- Development of Peak Annotator program that can accurately annotate EI-MS spectral with the molecular formulas. This PA program can be used to "clean up" EI-MS predictions and appears to improve the performance of EI-MS predictors.
- Development of a hybrid EI-MS predictor – called "Adjusted NEIMS with PA and MIIP" or EI-MSpred for short to perform accurate, annotated EI-MS prediction
- Demonstration that the combination of RIpred and EI-MSpred can be used to accurately identify compounds from a hypothetical GC-MS metabolomics experiment.

4.5 Future Work

As previously stated, my retention index predictor (RIpred) was fully completed. It has been thoroughly tested, published in a highly respected journal and performs as well as can be expected. However, my work with EI-MS prediction was not as complete. In particular, I did not independently develop a novel, superior-performing EI-MS predictor using ML methods. While my versions were better than CFM-EI, newly developed programs such as NEIMS and RASSP still outperformed what I could do. However, I did succeed in creating a hybrid program, called Adjusted NEIMS with PA and MIIP" or EI-MSpred for short, which uses the previously published NEIMS predictor [70] to generate spectral predictions and two programs I developed (PA and

MIIP) to annotate the spectra and modestly improve the accuracy of the spectra predictions. Unfortunately, I did not succeed in converting EI-MSpred into a webserver and did not publish a description of the EI-MSpred program in a scientific journal. This would be something that I would hope to complete in the near future.

I also believe it should be possible to improve NEIMS or to develop a separate, high performing EI-MS predictor that would perform at least as well as NEIMS by creating a training set of EI-MS spectra where all peaks have been annotated with molecular formulas using my Peak Annotator program. As noted earlier, my PA program is able to annotate nearly all (>99%) peaks in the experimentally collected EI-MS spectra (such as those in the NIST 23 database). These annotated EI-MS spectra can give rise to a tuple of three attributes (m/z, intensity, subformula) for each peak in the EI-MS spectra. I believe this annotated data could be quite useful for training an EI-MS predictor using a graph neural network. I also believe that the extra information contained in the molecular formula information could improve the quality of the EI-MS predictor and create an EI-MS predictor that also annotates peaks with formulas or even molecular fragment structures. Another potential future project that arises naturally from this work would be the development of a webserver or a computational pipeline to support combined RI + EI-MS analysis and compound identification. Currently RIpred only predicts RI data and only identifies compounds based on their molecular ion mass and RI values. The proposed EI-MSpred webserver would only predict EI-MS spectra and only identify compounds based on their EI-MS data. However, I believe the proposed EI-MSpred webserver could be easily modified to not only predict EI-MS data, but that it could combine input RI data (and utilize calls to the RIpred server) to perform integrated compound identification. Furthermore, both the RIpred and EI-MSpred predictors could be used to generate

RI and EI-MS prediction results for all the GC-MS compatible compounds in compound databases such as HMDB, FooDB, MiMeDB, DrugBank, etc. The creation of *in silico* spectral/RI libraries for all known compounds and even many hypothetical compounds (such as those generated by BioTransfomer [137]), could open many doors for rapid, automated compound identification. Certainly, if all these aforementioned plans could be completed, I believe that GC-MS could achieve a new level of importance and a new threshold in utility for compound identification – especially in fields such as metabolomics, exposomics, environmental or pollution monitoring and drug testing.

References

1. Chemical Abstracts Service, CAS assigns the 100 millionth CAS registry number® to a substance designed to treat acute myeloid leukemia, PR Newswire. (2015). https://www.prnewswire.com/news-releases/cas-assigns-the-100-millionth-cas-registry-number-to-a-substance-designed-to-treat-acute-myeloid-leukemia-300106332.html (accessed October 7, 2023).

2. F. Chassagne, G. Cabanac, G. Hubert, B. David, G. Marti, The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products®, Phytochem. Rev. 18 (2019) 601–622. https://doi.org/10.1007/s11101-019-09606-2.

3. D.J. Newman, G.M. Cragg, Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019, J. Nat. Prod. 83 (2020) 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

4. L.S. Goodman, A.G. Gilman, Goodman & Gilman’s the Pharmacological Basis of Therapeutics, J. Dermatol. Surg. Oncol. 7 (1981) 97–97. https://doi.org/10.1111/j.1524-4725.1981.tb00605.x.

5. CCDC, The Cambridge Structural Database (CSD) https://www.ccdc.cam.ac.uk/solutions/csd-core/components/csd/, 2013, October 1 (accessed 27 October 2022).

6. R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, DENDRAL: A case study of the first expert system for scientific hypothesis formation, Artif. Intell. 61 (1993) 209–261. https://doi.org/10.1016/0004-3702(93)90068-m.

7. R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project. McGraw-Hill Book Company (1980).

8. MOLGEN, Molecular Structure Generation , https://www.molgen.de/, 2021 (accessed 27 October 2022).

9. J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, C. Djerassi, Applications of artificial intelligence for chemical inference. I. Number of possible organic compounds. Acyclic structures containing carbon, hydrogen, oxygen, and nitrogen, J. Am. Chem. Soc. 91 (1969) 2973–2976. https://doi.org/10.1021/ja01039a025.

10. D.B. Nelson, M.E. Munk, K.B. Gash, D.L. Herald, Alanylactinobicyclone. Application of computer techniques to structure elucidation, J. Org. Chem. 34 (1969) 3800–3805. https://doi.org/10.1021/jo01264a014.

11. S. Sasaki, H. Abe, T. Ouki, M. Sakamoto, S. Ochiai, Automated structure elucidation of several kinds of aliphatic and alicyclic compounds, Anal. Chem. 40 (1968) 2220–2223. https://doi.org/10.1021/ac50158a061.

12 M.E. Elyashberg, L.A. Gribov, Formal-logical method for interpreting infrared spectra from characteristic frequencies, J. Appl. Spectrosc. 8 (1968) 189–191. https://doi.org/10.1007/bf00604681.

13. Science History Institute Museum & Library, Joseph John Thomson. https://sciencehistory.org/education/scientific-biographies/joseph-john-j-j-thomson/ (accessed 20 July, 2022).

14. H. Kameoka, Gas Chromatography/Mass Spectrometry, Mod. Methods Plant Anal. (1986) 254–276. https://doi.org/10.1007/978-3-642-82612-2_11.

15. Ptable, Periodic Table, https://ptable.com/?lang=en#Properties, 2022, Feb 6 (accessed 27 October 2022).

16. D.C. Muddiman, Jürgen H. Gross: Mass spectrometry: a textbook, 3rd ed., Anal. Bioanal. Chem. 410 (2018) 2051–2052. https://doi.org/10.1007/s00216-018-0870-8.

17. Mass Spectrometry Center, Mass Spectrometry Ionization Methods, http://chemistry.emory.edu/msc/tutorial/mass-spectrometry-ionization.html, 2017 (accessed 27 October 2022).

18. Creative Proteomics Blog, Two Soft Ionization Techniques-EI and MALDI, https://www.creative-proteomics.com/blog/index.php/two-soft-ionization-techniques-ei-and-maldi/, 2022 (accessed 27 October 2022).

19. J.B. Fenn, Ion formation from charged droplets: Roles of geometry, energy, and time, J. Am. Soc. Mass Spectrom. 4 (1993) 524–535. https://doi.org/10.1016/1044-0305(93)85014-o.

20. Creative Proteomics Blog, MALDI-TOF Mass Spectrometry, https://www.creative-proteomics.com/technology/maldi-tof-mass-spectrometry.htm, 2022 (accessed 27 October 2022).

21. A.M. Haag, Mass Analyzers and Mass Spectrometers., Adv. Exp. Med. Biol. 919 (2016) 157–169. https://doi.org/10.1007/978-3-319-41448-5_7.

22. Technology Networks, Types of Ion Detector for Mass Spectrometry, https://www.technologynetworks.com/analysis/articles/types-of-ion-detector-for-mass-spectrometry-347890#D3, 2022 (accessed 27 October 2022).

23. News Medical Life Sciences, Gas Chromatography-Mass Spectrometry (GC-MS) Applications, https://www.news-medical.net/life-sciences/Gas-Chromatography-Mass-Spectrometry-(GC-MS)-Applications.aspx, 2022, Oct 28 (accessed 28 October 2022).

24. Technology Networks, LC-MS – What Is LC-MS, LC-MS Analysis and LC-MS/MS, https://www.technologynetworks.com/analysis/articles/lc-ms-what-is-lc-ms-lc-ms-analysis-and-lc-msms-348238, 2022 (accessed 27 October 2022).

25. SHIMADZU, What is HPLC (High Performance Liquid Chromatography) ? , https://www.shimadzu.com/an/service-support/technical-support/analysis-basics/basic/what_is_hplc.html, 2022 (accessed 27 October 2022).

26. Wikipedia, High Performance Liquid Chromatography, https://en.wikipedia.org/wiki/High-performance_liquid_chromatography, 2022, Oct 22 (accessed 27 October 2022).

27. G. Nyerges, J. Mátyási, J. Balla, Investigation and comparison of 5% diphenyl – 95% dimethyl polysiloxane capillary columns, Period. Polytech. Chem. Eng. 64 (2020) 430–436. https://doi.org/10.3311/ppch.15289.

28. ScienceDirect, Retention Time, https://www.sciencedirect.com/topics/chemistry/retention-time#:~:text=Retention%20time%20is%20the%20time,will%20be%20the%20interaction%20time, 2022 (accessed 27 October 2022).

29. S. Kato, Y. Nakajima, R. Awaya, , I. Hata, , Y. Shigematsu, S. Saitoh, Total ion chromatograms of GC/MS urine analysis. A; urease/direct method , Pitfall in the diagnosis of fructose-1, 6-bisphosphatase deficiency: difficulty in detecting glycerol-3-phosphate with solvent extraction in urinary GC/MS analysis. The Tohoku Journal of Experimental Medicine, 237,3 (2015), 235-239. https://doi.org/10.1620/tjem.237.235.

30. Chromatography Today, Understanding the Difference between Retention Time and Relative Retention Time. https://www.chromatographytoday.com/news/autosamplers/36/breaking-news/understanding-the-difference-between-retention-time-and-relative-retention-time/31166, 2014, August 1 (accessed 21 July 2022).

31. LibreTexts Chemistry, Mass Spectrometry - Fragmentation Patterns, https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Mass_Spectrometry/Mass_Spec/Mass_Spectrometry_-_Fragmentation_Patterns, 2020, Aug 15 (accessed 27 October 2022).

32. CUBoulder Organic Chemistry Undergraduate Courses, Fragmentation Mechanisms, http://www.orgchemboulder.com/Spectroscopy/MS/fragmech.shtml, 2011 (accessed 27 October 2022).

33. Structure & Reactivity, Introductory Mass Spectrometry, https://employees.csbsju.edu/cschaller/Principles%20Chem/structure%20determination/ms%20pathways.htm, (accessed 27 October 2022).

34. Dummies, How the McLafferty Rearrangement Affects Carbonyl Fragmentation in Mass Spectrometry, https://www.dummies.com/article/academics-the-arts/science/chemistry/how-the-mclafferty-rearrangement-affects-carbonyl-fragmentation-in-mass-spectrometry-146329/, 2020, Aug 15 (accessed 27 October 2022).

35. B. Rickborn, Organic Reactions, (2012) 1–393. https://doi.org/10.1002/0471264180.or052.01.

36. National Institute of Standards and Technology, NIST20: Updates to the NIST Tandem and Electron Ionization Spectral Libraries. https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries, 2020, June 12 (accessed 21 July 2022).

37. Mass Spectrometry Data Center, AMDIS. https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis, 2019, November 26 (accessed 27 October 2022).

38. C.E. Shannon, A Mathematical Theory of Communication, Bell Syst. Tech. J. 27 (1948) 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

39. National Library of Medicine, National Institute of Health, PubChem. https://pubchem.ncbi.nlm.nih.gov/ (accessed 21 July, 2022).

40. J. Kang, C. Cao, Z. Li, Quantitative structure–retention relationship studies for predicting the gas chromatography retention indices of polycyclic aromatic hydrocarbons, J. Chromatogr. A. 799 (1998) 361–367. http://doi: 10.1016/j.chroma.2010.04.038.

41. C.T. Peng, Z.C. Yang, S.F. Ding, Prediction of retention indexes. II. Structure-retention index relationship on polar columns, J. Chromatogr. 586 (1991) 85-112. http://doi.10.1016/0021-9673(91)80028-f.

42. C.D. Mitra, N.C. Saha, Determination of retention indices of saturated hydrocarbons by graphical methods, J. Chromatogr. Sci. 8 (1970) 95-102 https://doi.org/10.1093/chromsci/8.2.95.

43. S. Hou, K.A.J.M. Stevenson, J.J. Harynuk, A simple, fast, and accurate thermodynamic-based approach for transfer and prediction of gas chromatography retention times between columns and instruments Part I: Estimation of reference column geometry and thermodynamic parameters, J. Sep. Sci. 41 (2018) 2544–2552. http://dx.doi.org/10.1002/jssc.201701343.

44. S. Hou, K.A.J. Stevenson, J.J. Harynuk, A simple, fast, and accurate thermodynamic-based approach for transfer and prediction of gas chromatography retention times between

columns and instruments Part III: Retention time prediction on target column, J. Sep. Sci. 41 (2018) 2559–2564. https://doi.org/10.1002/jssc.201701345.

45. K. Héberger, Quantitative structure–(chromatographic) retention relationships, J. Chromatogr. A. 1158 (2007) 273–305. https://doi.org/10.1016/j.chroma.2007.03.108..

46. X.H. Zhu, W. Wang, K.-W. Schramm, W. Niu, Prediction of the kováts retention indices of thiols by use of quantum chemical and physicochemical descriptors, Chromatogr. 65 (2007) 719–724. http://dx.doi.org/10.1365/s10337-007-0237-3.

47. A. Littlewood, C. Phillips, D. Price, The chromatography of gases and vapours. Part V. Partition analyses with columns of silicone 702 and of tritolyl phosphate, J. Chem. Soc. Faraday Trans. (Resumed) (1955) 1490- 1489. https://doi.org/10.1039/JR96500FP001.

48. E. Konoz, M.H. Fatemi, R. Faraji, Prediction of Kovats retention indices of some aliphatic aldehydes and ketones on some stationary phases at different temperatures using artificial neural network, J. Chromatogr. Sci. 46 (2008) 406–412. http://dx.doi.org/10.1093/chromsci/46.5.406.

49. D.D. Matyushin, A.Y. Sholokhova, A.K. Buryak, A deep convolutional neural network for the estimation of gas chromatographic retention indices, J. Chromatogr. A. 1607 (2019) 460395. http://dx.doi.org/10.1016/j.chroma.2019.460395.

50. F. Luan, C. Xue, R. Zhang, C. Zhao, M. Liu, Z. Hu, B. Fan, Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector

machine, Anal. Chim. Acta. 537 (2005) 101–110. http://dx.doi.org/10.1016/j.aca.2004.12.085.

51. D.D. Matyushin, A.Yu. Sholokhova, A.K. Buryak, Gradient boosting for the prediction of gas chromatographic retention indices, Сорбционные и Хроматографические Процессы. 19 (2019) 630–635. https://doi.org/10.17308/sorpchrom.2019.19/2223.

52. R. Idroes, T.R. Noviandy, A. Maulana, R. Suhendra, N.R. Sasmita, M. Muslem, G.M. Idroes, I. Irvanizam, Retention Index Prediction of Flavor and Fragrance by Multiple Linear Regression and the Genetic Algorithm, Int. Rev. Model. Simul. (IREMOS). 12 (2019) 373–380. https://doi.org/10.15866/iremos.v12i6.18353.

53. T. Vrzal, M. Malečková, J. Olšovská, DeepReI: Deep learning-based gas chromatographic retention index predictor, Anal. Chim. Acta. 1147 (2021) 64–71. http://dx.doi.org/10.1016/j.aca.2020.12.043

54. C. Veenaas, A. Linusson, P. Haglund, Retention-time prediction in comprehensive two-dimensional gas chromatography to aid identification of unknown contaminants, Anal. Bioanal. Chem. 410 (2018) 7931–7941. https://doi.org/10.1007/s00216-018-1415-x.

55. K.E. Miller, T.J. Bruno, Isothermal Kováts retention indices of sulfur compounds on a poly(5% diphenyl-95% dimethylsiloxane) stationary phase., J. Chromatogr. A. 1007 (2003) 117–25. https://doi.org/10.1016/s0021-9673(03)00958-0.

56. Y.-T. Wang, Z.-X. Yang, Z.-H. Piao, X.-J. Xu, J.-H. Yu, Y.-H. Zhang, Prediction of flavor and retention index for compounds in beer depending on molecular structure using a

machine learning method, RSC Advances. 11 (2021) 36942–36950. https://doi.org/10.1039/d1ra06551c.

57. D.D. Matyushin, A.K. Buryak, Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning, IEEE Access. 8 (2020) 223140–223155. https://doi.org/10.1109/access.2020.3045047.

58. S.M. de Cripan, S.M. de Cripan, A. Cereto-Massagué, P. Herrero, A. Barcaru, N. Canela, X. Domingo-Almenara, Machine learning-based retention time prediction of trimethylsilyl derivatives of metabolites, Biomedicines. 10 (2022) 879. https://doi.org/10.3390/biomedicines10040879.

59. D.D. Matyushin, A.Y. Sholokhova, A.K. Buryak, Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases, Int. J. Mol. Sci. 22 (2021) 9194. https://doi.org/10.3390/ijms22179194.

60. A. Weissberg, S. Dagan, Interpretation of ESI(+)-MS-MS spectra—Towards the identification of “unknowns”, Int. J. Mass Spec. 299 (2010) 158-168. https://doi.org/10.1016/j.ijms.2010.10.024.

61. S. Wang, T. Kind, D.J. Tantillo, O. Fiehn, Predicting in silico electron ionization mass spectra using quantum chemistry, J. Cheminform. 12 (2020) 63. https://doi.org/10.1186/s13321-020-00470-3.

62. R.M. Borges, S.M. Colby, S. Das, A.S. Edison, O. Fiehn, T. Kind, J. Lee, A.T. Merrill, K.M. Merz Jr, T.O. Metz, J.R. Nunez, D.J. Tantillo, L.-P. Wang, S. Wang, R.S. Renslow,

Quantum chemistry calculations for metabolomics, Chem. Rev. 121 (2021) 5633–5670. https://doi.org/10.1021/acs.chemrev.0c00901.

63. Y.-T. Wang, Z.-X. Yang, Z.-H. Piao, X.-J. Xu, J.-H. Yu, Y.-H. Zhang, Prediction of flavor and retention index for compounds in beer depending on molecular structure using a machine learning method, RSC Adv. 11 (2021) 36942–36950. https://doi.org/10.1039/d1ra06551c.

64. S.A. Schreckenbach, J.S.M. Anderson, J. Koopman, S. Grimme, M.J. Simpson, K.J. Jobst, Predicting the Mass Spectra of Environmental Pollutants Using Computational Chemistry: A Case Study and Critical Evaluation, J. Am. Soc. Mass Spectrom. 32 (2021) 1508–1518. https://doi.org/10.1021/jasms.1c00078.

65. F. Allen, A. Pon, R. Greiner, D. Wishart, Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification, Anal. Chem. 88 (2016) 7689–7697. http://dx.doi.org/10.1021/acs.analchem.6b01622.

66. F. Allen, R. Greiner, D. Wishart, Competitive Fragmentation Modeling of ESI-MS/MS spectra for putative metabolite identification, arXiv. (2013). https://doi.org/10.48550/arxiv.1312.0264.

67. S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinform. 11 (2010) 148–148. https://doi.org/10.1186/1471-2105-11-148.

68. A. Kerber, M. Meringer, & C. Rücker, CASE via MS: ranking structure candidates by mass spectra, Croatia chem. acta. 79 (2006) 449-464.

69. Thermo Fisher Scientific, Mass Frontier Spectral Interpretation Software. https://www.thermofisher.com/ca/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/mass-frontier-spectral-interpretation-software.html, 2023 (accessed 1 July, 2023).

70. J.N. Wei, D. Belanger, R.P. Adams, D. Sculley, Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks, ACS Cent Sci. 5 (2019) 700–708. https://doi.org/10.1021/acscentsci.9b00085.

71. B. Zhang, J. Zhang, Y. Xia, P. Chen, B. Wang, Prediction of electron ionization mass spectra based on graph convolutional networks, Int. J. Mass Spectrom. 475 (2022) 116817. https://doi.org/10.1016/j.ijms.2022.116817.

72. H. Ji, H. Deng, H. Lu, Z. Zhang, Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks, Anal. Chem. 92 (2020) 8649–8653. https://doi.org/10.1021/acs.analchem.0c01450.

73. N. Boegelsack, C. Sandau, D.W. McMartin, J.M. Withey, G. O'Sullivan, Development of retention time indices for comprehensive multidimensional gas chromatography and application to ignitable liquid residue mapping in wildfire investigations, J. Chromatogr. A. 1635 (2021) 461717. https://doi.org/10.1016/j.chroma.2020.461717.

74. B. B. Barnes, M. B. Wilson, P. W. Carr, M. F. Vitha, C. D. Broeckling, A. L. Heuberger, J. Prenni, G. C. Janis, H. Corcoran, N. H. Snow, S. Chopra, R. Dhandapani, A. Tawfall, L. W. Sumner, P. G. Boswell, Retention projection enables reliable use of shared gas chromatographic retention data across laboratories, instruments, and methods. Anal. Chem. 85 (2013) 11650-11657. https://doi.org/10.1021/ac4033615.

75. N. Etxebarria, O. Zuloaga, M. Olivares, L.J. Bartolomé, P. Navarro, Retention-time locked methods in gas chromatography, J. Chromatogr. A. 1216 (2009) 1624–1629. https://doi.org/10.1016/j.chroma.2008.12.038.

76. National Institute of Standards and Technology (NIST) Chemistry WebBook, SRD 69, Gas Chromatographic Retention Data. https://webbook.nist.gov/chemistry/gc-ri/, 2021, September 21 (accessed 21 July 2022).

77. S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, Estimation of Kováts retention indices using group contributions, J. Chem. Inf. Model. 47 (2007) 975–980. https://doi.org/10.1021/ci600548y.

78. V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Prediction of capillary gas chromatographic retention times of fatty acid methyl esters in human blood using MLR, PLS and back-propagation artificial neural networks, Talanta. 83 (2011) 1014–1022. http://dx.doi.org/10.1016/j.talanta.2010.11.017.

79. C. Qu, B.I. Schneider, A.J. Kearsley, W. Keyrouz, T.C. Allison, Predicting Kováts retention indices using graph neural networks, J. Chromatogr. A. 1646 (2021) 462100. https://doi.org/10.1016/j.chroma.2021.462100.

80. B. Karolat, J. Harynuk, Prediction of gas chromatographic retention time via an additive thermodynamic model, J. Chromatogr. A. 1217 (2010) 4862–4867. http://dx.doi.org/10.1016/j.chroma.2010.05.037.

81. M.P. Elizalde-González, M. Hutfließ, K. Hedden, Retention index system, adsorption characteristics, and structure correlations of polycyclic aromatic hydrocarbons in fuels, J. High Resolut. Chromatogr. 19 (1996) 345–352. http://dx.doi.org/10.1002/jhrc.1240190608.

82. H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J. Cheminform. 10 (2018) 4. http://dx.doi.org/10.1186/s13321-018-0258-y.

83. C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474. http://dx.doi.org/10.1002/jcc.21707.

84. B. Ren, Atom-type-based AI topological descriptors for quantitative structure–retention index correlations of aldehydes and ketones, Chemometrics Intellig. Lab. Syst. 66 (2003) 29–39. http://dx.doi.org/10.1016/s0169-7439(03)00004-2.

85. O. Farkas, K. Héberger, I.G. Zenkevich, Quantitative structure–retention relationships XIV: Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds, Chemometrics Intellig. Lab. Syst. 72 (2004) 173–184. https://doi.org/10.1016/j.chemolab.2004.01.012.

86. R.-J. Hu, H.-X. Liu, R.-S. Zhang, C.-X. Xue, X.-J. Yao, M.-C. Liu, Z.-D. Hu, B.-T. Fan, QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic

compounds from heuristically computed molecular descriptors, Talanta. 68 (2005) 31–39. http://dx.doi.org/10.1016/j.talanta.2005.04.034.

87. S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm, Chromatogr. 67 (2008) 917–922. http://dx.doi.org/10.1365/s10337-008-0608-4.

88. H.-F. Chen, Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression, Anal. Chim. Acta. 609 (2008) 24–36. http://dx.doi.org/10.1016/j.aca.2008.01.003.

89. A. Yan, G. Jiao, Z. Hu, B.T. Fan, Use of artificial neural networks to predict the gas chromatographic retention index data of alkylbenzenes on carbowax-20M, Comput. Chem. 24 (2000) 171–179. http://dx.doi.org/10.1016/s0097-8485(99)00058-3.

90. B. Hemmateenejad, K. Javadnia, M. Elyasi, Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy, Anal. Chim. Acta. 592 (2007) 72–81. http://dx.doi.org/10.1016/j.aca.2007.04.009.

91. CHEMDATA.NIST.GOV, NIST Libraries and Software. https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17, 2021, July 12 (accessed 21 July 2022).

92. National Cancer Institute, National Institute of Health, CADD Group Chemoinformatics Tools and User Services. https://cactus.nci.nih.gov/, 2022, March 29 (accessed 21 July, 2022).

93. PubChemPy 1.0, A simple Python wrapper around the PubChem PUG REST API. https://pypi.org/project/PubChemPy/1.0/, 2013, May 1 (accessed 21 July 2022).

94. RDKit, RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/, (accessed 21 July 2022).

95. C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph Networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (2019) 3564–3572. http://dx.doi.org/10.1021/acs.chemmater.9b01294.

96. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, TensorFlow: Large-scale machine learning on heterogeneous distributed systems, ArXiv: 1603.04467. (2016). https://doi.org/10.48550/ arxiv.1603.04467.

97. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019) 8024–8035. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

98. B. Chen, R. Barzilay, T. Jaakkola, Path-augmented graph transformer network, ArXiv: 1905.12712. (2019). https://doi.org/10.48550/arXiv.1905.12712.

99. Y.D. Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D.S. Wishart, ClassyFire: automated chemical classification with a comprehensive, computable taxonomy, J. Cheminform. 8 (2016) 61. https://doi.org/10.1186/s13321-016-0174-y.

100. F. Csizmadia, JChem: Java applets and modules supporting chemical database handling from web browsers, J. Chem. Inf. Comput. Sci. 40 (2000) 323–324. http://dx.doi.org/10.1021/ci9902696.

101. D.S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H.B. Schiöth, R. Greiner, V. Gautam, HMDB 5.0: the Human Metabolome Database for 2022, Nucleic Acids Research. 50 (2022) D622–D631. https://doi.org/10.1093/nar/gkab1062.

102. D.S. Wishart, Z. Sayeeda, Z. Budinski, A. Guo, B.L. Lee, M. Berjanskii, M. Rout, H. Peters, R. Dizon, R. Mah, C. Torres-Calzada, M. Hiebert-Giesbrecht, D. Varshavi, D. Varshavi, E. Oler, D. Allen, X. Cao, V. Gautam, A. Maras, E.F. Poynton, P. Tavangar, V. Yang, J.A. van Santen, R. Ghosh, S. Sarma, E. Knutson, V. Sullivan, A.M. Jystad, R. Renslow, L.W.

Sumner, R.G. Linington, J.R. Cort, NP-MRD: the Natural Products Magnetic Resonance Database, Nucleic Acids Res. 50 (2022) D665–D677. https://doi.org/10.1093/nar/gkab1052.

103. D. Wishart, D. Arndt, A. Pon, T. Sajed, A.C. Guo, Y. Djoumbou, C. Knox, M. Wilson, Y. Liang, J. Grant, Y. Liu, S.A. Goldansaz, S.M. Rappaport, T3DB: the toxic exposome database, Nucleic Acids Res. 43 (2015) D928–34. https://doi.org/10.1093/nar/gku1004.

104. F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, D.S. Wishart, CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification, Anal. Chem. 93 (2021) 11692–11700. https://doi.org/10.1021/acs.analchem.1c01465.

105. S. Bouatra, F. Aziat, R. Mandal, A.C. Guo, M.R. Wilson, C. Knox, T.C. Bjorndahl, R. Krishnamurthy, F. Saleem, P. Liu, Z.T. Dame, J. Poelzer, J. Huynh, F.S. Yallou, N. Psychogios, E. Dong, R. Bogumil, C. Roehring, D.S. Wishart, The human urine metabolome, PLoS One. 8 (2013) e73076. https://doi.org/10.1371/journal.pone.0073076.

106. A.P. Bento, A. Patrícia Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L.J. Bellis, M. De Veij, A.R. Leach, An open source chemical structure curation pipeline using RDKit, J. Cheminform. 12 (2020) 51. https://doi.org/10.1186/s13321-020-00456-1.

107. GOLM METABOLOME DATABASE, Download the GMD mass spectrum reference Libraries. http://gmd.mpimp-golm.mpg.de/download/, 2021, August 31 (accessed 21 July 2022).

108. O.D. Sparkman, Z.E. Penton, F.G. Kitson, Gas Chromatography and Mass Spectrometry: A Practical Guide, Section 2: GC Conditions, Derivatization, and Mass Spectral Interpretation

of Specific Compound Types. (2011) 243–248. https://doi.org/10.1016/b978-0-12-373628-4.00010-1.

109. B.A. Eckenrode, S.A. McLuckey, G.L. Glish, Comparison of electron ionization and chemical ionization sensitivities in an ion trap mass spectrometer, Int. J. Mass Spectrom. Ion Process. 106 (1991) 137–157. https://doi.org/10.1016/0168-1176(91)85015-E

110. Electron Ionization, Modern Methods in Natural Products Chemistry. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/electron-ionization, 2023 (accessed 1 July, 2023).

111. D. Döpp, H. Döpp, 1,2,3-Triazines and their Benzo Derivatives, in: Comprehensive Heterocyclic Chemistry III, Elsevier, 2008: pp. 1–93. https://doi.org/10.1016/B978-0-12-818655-8.00101-3.

112. R.E. Clement, V.Y. Taguchi, Ontario. Laboratory Services Branch, Techniques for the gas chromatography - mass spectrometry identification of organic compounds in effluents, Queen’s Printer for, 1988.

113. Mass spectral databases, Wiley Science Solutions. https://sciencesolutions.wiley.com/mass-spectral-databases/, 2022 (accessed 1 July, 2023).

114. MoNA - MassBank of North America, Welcome to MoNA! https://mona.fiehnlab.ucdavis.edu/, 2022, March 4 (accessed 1 July, 2023).

115. MassBank, MassBank Europe. https://massbank.eu/MassBank/, 2023, June 5 (accessed 1 July, 2023).

116. MassBank, MSSJ MassBank. http://www.massbank.jp/ (accessed 1 July, 2023).

117. Spectral Database for Organic Compounds SDBS, Welcome to Spectral Database for Organic Compounds, SDBS. https://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi , 2023, Mar 31 (accessed 1 July, 2023)

118. S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom. 5 (1994) 859–866. https://doi.org/10.1016/1044-0305(94)87009-8.

119. Z.B. Alfassi, On the normalization of a mass spectrum for comparison of two spectra, J. Am. Soc. Mass Spectrom. 15 (2004) 385–387. https://doi.org/10.1016/j.jasms.2003.11.008.

120. Y. Li, T. Kind, J. Folz, A. Vaniya, S.S. Mehta, O. Fiehn, Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification, Nat. Methods. 18 (2021) 1524–1531. https://doi.org/10.1038/s41592-021-01331-z.

121. CAS content, CAS. https://www.cas.org/about/cas-content (accessed October 12, 2023).

122. GenForm, Generation of molecular formulas by high-resolution MS and MS/MS data. https://sourceforge.net/projects/genform/, 2023 (accessed 1 July, 2023)

123. MOLGEN-MS, Evaluation of Low Resolution Electron Impact Mass Spectra without Database Search. https://molgen.de/documents/MolgenMS_manual/index.html, 2000 (accessed 1 July, 2023).

124. MS-FINDER tutorial, Abstract. https://mtbinfo-team.github.io/mtbinfo.github.io/MS-FINDER/tutorial.html, 2018, 6 Dec (accessed 1 July, 2023).

125. C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminform. 8 (2016) 3. https://doi.org/10.1186/s13321-016-0115-9.

126. R.L. Zhu, E. Jonas, Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization-Mass Spectrometry, Anal. Chem. 95 (2023) 2653–2663. https://doi.org/10.1021/acs.analchem.2c02093.

127. S. Goldman, J. Li, C.W. Coley, Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks, arXiv [q-bio.QM]. (2023). https://doi.org/10.48550/arXiv.2304.13136.

128. S. Goldman, J. Bradshaw, J. Xin, C.W. Coley, Prefix-tree Decoding for Predicting Mass Spectra from Molecules, arXiv [q-bio.QM]. (2023). https://doi.org/10.48550/arXiv.2303.06470.

129. M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey, T. Butler, Efficiently predicting high resolution mass spectra with graph neural networks, arXiv [cs.LG]. (2023). https://doi.org/10.48550/arXiv.2301.11419.

130. H.L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service, J. Chem. Doc. 5 (1965) 107–113. https://doi.org/10.1021/c160017a018.

131. MS Terms Wiki, Ionization efficiency. http://mass-spec.lsu.edu/msterms/index.php/Ionization_efficiency, 2013, Oct 24 (accessed 1 July, 2023).

132. V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer Science & Business Media (2006). https://doi.org/10.1007/0-387-34239-7.

133. Fiehn lab, Nitrogen rule. https://fiehnlab.ucdavis.edu/projects/seven-golden-rules/nitrogen-rule, 2016 (accessed 1 July, 2023).

134. T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105. https://doi.org/10.1186/1471-2105-8-105.

135. J. Ashenhurst, Degrees of unsaturation (or IHD, index of hydrogen deficiency), Master Organic Chemistry. (2016). https://www.masterorganicchemistry.com/2016/08/26/degrees-of-unsaturation-index-of-hydrogen-deficiency/ (accessed October 12, 2023).

136. A. Anjum, J. Liigand, R. Milford, V. Gautam, D.S. Wishart, Accurate prediction of isothermal gas chromatographic Kováts retention indices, J. Chromatogr. A. 1705 (2023) 464176. https://doi.org/10.1016/j.chroma.2023.464176.

137. D. S. Wishart, S. Tian, D. Allen, E. Oler, H. Peters, V. W. Lui, V. Gautam, Y. D. Feunang, R. Greiner, T. O. Metz, BioTransformer 3.0—a web server for accurately predicting metabolic transformation products. Nucleic Acids Res. 50 (2022) W115–W123. https://doi.org/10.1093/nar/gkac313.

Appendix A: Additional Figure

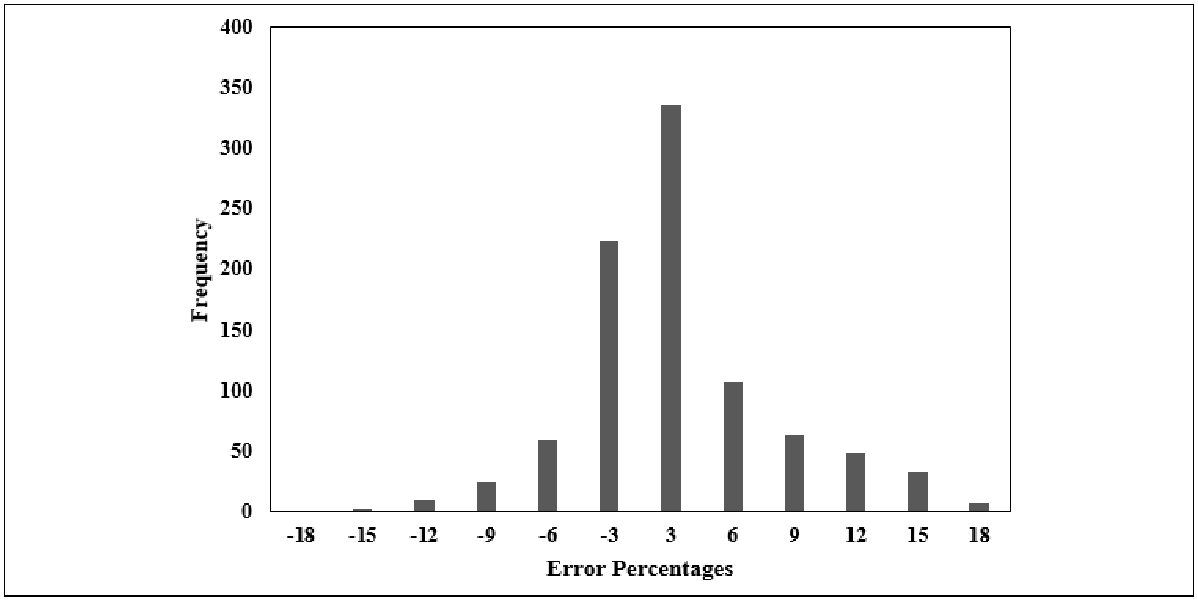


Figure A. 1: Distribution of errors between and experimentally measured (GolmDB) and predicted RI values (from RIpred). As seen from this graph the distribution is quite Gaussian in nature.