Cheminformatics Tools for Enabling Metabolomics

by

Yannick Djoumbou Feunang

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

Doctor of Philosophy

in

Microbiology and Biotechnology

Department of Biological Sciences
University of Alberta

©Yannick Djoumbou Feunang, 2017

Abstract

Metabolites are small molecules (<1500 Da) that are used in or produced during chemical reactions in cells, tissues, or organs. Upon absorption or biosynthesis in humans (or other organisms), they can either be excreted back into the environment in their original form, or as a pool of degradation products. The outcome and effects of such interactions is function of many variables, including the structure of the starting metabolite, and the genetic disposition of the host organism. For this reasons, it is usually very difficult to identify the transformation products as well as their long-term effect in humans and the environment. This can be explained by many factors: (1) the relevant knowledge and data are for the most part unavailable in a publicly available electronic format; (2) when available, they are often represented using formats, vocabularies, or schemes that vary from one source (or repository) to another. Assuming these issues were solved, detecting patterns that link the metabolome to a specific phenotype (e.g. a disease state), would still require that the metabolites from a biological sample be identified and quantified, using metabolomic approaches. Unfortunately, the amount of compounds with publicly available experimental data (~20,000) is still very small, compared to the total number of expected compounds (up to a few million compounds). For all these reasons, the development of cheminformatics tools for data organization and mapping, as well as for the prediction of biotransformation and spectra, is more crucial than ever.

My PhD thesis focused on developing several cheminformatics tools that address these limitations. First, I developed *ClassyFire* and ChemOnt. *ClassyFire* is a publicly available software tool and webserver that automatically and hierarchically classifies any given molecule based on its structure. It relies partly on ChemOnt, a comprehensive and

comprehensible taxonomy that contains >4,800 chemical categories, as well as their textual descriptions and mappings to other ontologies. *ClassyFire* was used to classify and annotate >80 million compounds. The webserver also integrates a text-based search engine. These features make *ClassyFire* unique in the sphere of publicly available computational tools. *ClassyFire* and ChemOnt are available at http://classyfire.wishartlab.com. Second, I developed *BioTransformer* and BioTransformerDB. *BioTransformer* is a software tool for the prediction of small molecule metabolism in mammals. It uses a hybrid approach that partly relies on BioTransformerDB, a unique database of biotransformations containing experimentally confirmed metabolic reactions that transform >1,000 drugs, pesticides, cosmetics, and food compounds, among others. The current version of *BioTransformer*, which is available at https://bitbucket.org/djoumbou/biotransformer, focuses on the human species, but is easily expandable to other species. Third, I developed *CFM-ID 3.0*, an extension of *CFM-ID* (1.0, and 2.0), originally developed by Felicity Allen *et al*. *CFM-ID 3.0* is a software tool and webserver for the prediction and annotation of MS spectra, as well as the identification of metabolites. With the integration of a rule-based fragmentation approach for spectra prediction, the development of new ranking functions, and the expansion of the spectral database, *CFM-ID 3.0* showed a significant improvement, in terms of speed and accuracy, compared to previous versions. *CFM-ID 3.0* is currently available as we web server at http://cfmid-staging.wishartlab.com/.

ClassyFire, *BioTransformer*, and *CFM-ID* have found applications in various fields including chemical information management, metabolomics, and exposomics, among others. Together, they build a cheminformatics platform that can enable metabolomics, and contribute to the understanding of our environment as well as the advancement of science.

Preface

This thesis is an original work by Yannick Djoumbou Feunang, based on several original ideas provided by my supervisor Dr. David S. Wishart. This research project was led by Dr. David Wishart from the Departments of Biological Sciences and Computing Science at University of Alberta. All experiments and research activities were performed in Dr. Wishart's Lab at the University of Alberta. To complete the work described in this thesis, I was assisted by a number of individuals including Dr. Wishart and members of his laboratory. More specifically, Dr. Wishart supervised my training, coordinated the programming activities, designed the assessment methods used for this work and played a key role in the writing and editing of this thesis. Craig Knox, Roman Eisner, Michael Wilson helped in the development of the classification infrastructure for the *ClassyFire* program. Nazanin Assempour, and Dr. Ithayavani Iynkkaran contributed to the annotation/validation of biotransformations and structures in BioTransformerDB. Allison Pon, and Tanvir Sajed, contributed to the management of experimental spectral data obtained from other sources. Tammy Zheng prepared samples and collected experimental ESI-MS/MS spectra that were analyzed to create fragmentation rules. Dr Naama Karu provided feedback for the design and validation of the fragmentation rules. Dr Felicity Allen provided feedback in regard to previous versions of *CFM-ID*. In addition to the aforementioned colleagues, several other scientists provided feedback for different projects. Dr Evan Bolton facilitated the collaboration with the PubChem development team (NIH, USA). Dr Christoph Steinbeck, Dr Gary Owen, Dr Jana Hastings facilitated the collaboration with the ChEBI development team (EBI, UK). Dr Fahy Eoin, and Dr Shankar Subramanian facilitated the collaboration with the LIPID MAPS consortium

(UCSD, USA). Dr Jarlei Fiamoncini (INRA, France), and Dr Claudine Manach (INRA, France) provided feedback in regard to gut microbial metabolism for BioTransformer. Dr Katherine Fenner (EAWAG, Switzerland) provided access to up to date preferences rules for the prediction of environmental microbial degradation.

I was responsible for writing and testing the programs constituting *ClassyFire*, *BioTransformer* and *CFM-ID 3.0*. I was also responsible for developing the chemical ontology (ChemOnt) and the biotransformation database (BioTransformerDB) that were required to implement *ClassyFire* and *BioTransformer*, respectively. In addition I was largely responsible for acquiring and collecting all of the data required for BioTransformerDB and the experimental mass spectra files (from various databases) for *CFM-ID 3.0*'s spectral database, and for writing this thesis.

Dedication

I would like to dedicate this thesis to my late father, Nkemnoubong Djoumbou Joseph, and my mother Métago Janette épouse Djoumbou.

Knowledge without wisdom is like water in the sand – African proverb

Acknowledgements

First and foremost, I would like to express my infinite gratitude to the almighty God, for giving me the honour and the ability maintain the course throughout this long journey. I aim at always doing my best for the betterment of the world and myself.

I would like to express my sincere gratitude to my supervisor, Dr. David S. Wishart, for giving me the opportunity to work on some very exciting projects in his research group. I am particular grateful for his continuous support, patience, and motivation. Dr. Wishart's vision, attention to detail, and high expectation for excellent work have played a significant role in my journey as a PhD student; they helped me acquire and improve my scientific skills that have helped make me a better, more independent researcher.

It has been a real honour working in the Wishart group with such a diverse group of scientists and students from so many backgrounds and cultures. I would like to acknowledge the past and present members of the Wishart lab for contributing to my experience here at the University of Alberta. This is a special group of people and I will always appreciate the encouragement and good will we shared. I am especially thankful to Craig Knox, Roman Eisner, Allison Pon, Michael Wilson, Nazanin Assempour, Tammy Zheng, Tanvir Sajed, Siyang Tian, Zheng Shi, Dr. Naama Karu, Dr. Ithayavani Iynkkaran, and Dr. Felicity Allen. Thanks also to Dr. Rupa Mandal, the lab's metabolomics manager, who has been a great source of help and provided wonderful suggestions during my research.

I was able to work or collaborate with several scientists who contributed to the evaluation or development of my projects. I would like to particularly thank Dr. Evan

Bolton (NIH, USA), Dr. Jarlei Fiamoncini (INRA, France), Dr. Claudine Manach (INRA, France), Dr. Christoph Steinbeck (EBI, UK), Dr. Gary Owen (EBI, UK), Dr. Jana Hastings (EBI, UK), Dr. Fahy Eoin (UCSD, USA), Dr. Shankar Subramanian (UCSD, USA), and Dr. Katherine Fenner (EAWAG, Switzerland).

I would like to gratefully acknowledge my generous funding sources, especially the Canadian Institutes of Health Research and Genome Canada for financial support.

I am also sincerely grateful to my supervisory committee members. I would like to thank Dr. Russ Greiner for his continuous support, and for always having the door open whenever I needed his help. He has played a significant role in my journey as a PhD student, by helping in the design and the development of my research projects. I would also like to thank Dr. Warren Gallin for his insightful suggestions, and providing me with scientific literature that was of great help in conducting my research and writing this thesis. I greatly appreciate the help of all other scientists whom I have collaborated with, and learned from throughout my PhD program.

I have no words to express my love and gratitude to my family. Thank you to my parents, Joseph and Janette; my siblings Gladice, Christian, Patricia, Stéphane, and Arnaud; my uncles and aunts, especially Rigobert, Antoinette, and Chantal; my grandparents Mathias, Laurent, Bernadette, and Marthe. I am grateful for their love.

August 12, 2017

Yannick Djoumbou Feunang

Table of Contents

List of Figures

Chapter 3: *BioTransformer*: An accurate, freely available tool for predicting secondary metabolism of small molecules in mammals ………………………...110

Chapter 4: *CFM-ID 3.0* - Significantly Improved ESI-MS/MS Spectral Prediction Using a Hybrid In Silico Fragmentation Model with Metadata……………………149

List of Tables

List of Abbreviations

The following table shows a list of terms used in this document that are relevant to Cheminformatics

ALogP	Ghose-Crippen's implementation of the octanol/water partition coefficient
FASTA	A file format for the recording of gene and protein sequences
InChI	International Chemical Identifier
InChI Key	Hashed International Chemical Identifier
JSON	JavaScript Object Notation
LogP	Logarithm of the partition coefficient
MOL	MDL molfile
SDF	Structure Data File
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SMIRKS	A simple transform language
SoM	Site of Metabolism

The following table shows a list of terms used in this document that are relevant to Metabolomics

CYP450	Cytochrome P450
Da	Dalton or unified atomic mass
EI	Electron Ionization
ESI-	Electrospray Ionization
GC	Gas Chromatography
GPAT3	Glycerol-3-phosphate acyltransferase 3
LC	Liquid Chromatography
MS	Mass Spectrometry
NAPQI	N-acetyl-p-benzoquinone imine
PC	Phosphatidycholine
PE	Phosphatidylethanolamine
PS	Phosphatidylserine
Q-TOF	Quadrupole time-of-flight
UGT	Uridine 5'-diphospho-glucuronosyltransferase
SULT	Sulfotransferase

Chapter 1

General Introduction

1.1 Introduction

Metabolites are small molecules (<1500 Da) used or produced during metabolic reactions in cells or tissues. They are critical to nearly all life processes, providing energy in the form of ATP and NADH, pH and solvent buffers in the form of phosphate and bicarbonate ions, enzyme cofactors such as vitamins, cAMP, calcium, as well as the fundamental building blocks for cells and tissues such as amino acids, nucleic acids and lipids. Metabolites routinely interact with larger biomolecules such as DNA, RNA, lipid membranes, enzymes and protein transporters, thereby influencing the phenotype (i.e. the observable physical and biochemical characteristics) of a cell or organism (1). While most of the metabolites found in an organism are beneficial or essential for life, others may be harmful if their concentrations are too high or if they persist in the body for an extended period of time. Therefore, it is vital to understand how these small molecule chemicals influence larger macromolecules such as genes and proteins, and vice-versa. This kind of molecular understanding will ultimately lead to a better biological understanding of how living systems function, and why they sometimes fail to work properly.

The relationship between small molecules and their effects on living systems can be explored through metabolomics. Metabolomics is an emerging field of science that focuses on comprehensively characterizing metabolites in cells, biological extracts, or whole organisms and uses this information to reveal new biological or biochemical insights (2). Because metabolites represent the end products of both genetically programmed events and unprogrammed external exposures, the measurement of the metabolome reveals a great deal about an organism’s molecular phenotype. This unique ability to probe an organism's molecular phenotype, has allowed metabolomics to become an increasingly important vehicle for research. Indeed, metabolomics is finding applications in a variety of life science endeavours such as drug discovery (3), nutritional research (4,5), environmental monitoring (6), precision medicine (7) and many other biomedical disciplines.

Despite the availability of increasingly powerful analytical tools and techniques, the routine identification of metabolites remains particularly challenging. This is because much of the data that is needed to help identify compounds (such as structures, pathways and referential mass spectra) and to interpret their functions is scattered in thousands of books, journals, proprietary databases and numerous "boutique" data repositories. Fortunately, this situation is beginning to change. Over the past decade, a number of high quality computational tools and freely available electronic databases have become available to facilitate metabolite identification and improve the interpretation of metabolomic data (8-10). Most of these systems incorporate advanced cheminformatics capabilities that allow users to store, visualize, and search both chemical and biochemical information. However, despite the breadth and depth of today's metabolomic resources, there remain significant gaps in our knowledge regarding the structure, function, metabolism, and health effects of most of the chemicals found in our bodies. This is partly due to the fact that the field of metabolomics is still very young. It is also due to the fact that the "chemical space" of the metabolome is very large, very complex and difficult to fully explore. Even now, new chemicals and new metabolites are being discovered, synthesized, isolated, or uploaded into both chemical and metabolomic databases every day. As a result, the holes in our knowledge of the metabolome continue to be disturbingly large.

The focus of my PhD research is to develop approaches to help fill these holes and to improve the ways that we identify, describe or categorize metabolites. In particular, this thesis describes a number of novel software tools and databases that I have developed that allow metabolomic researchers to: 1) properly describe and categorize essentially all known chemicals and metabolites; 2) predict the chemical structures of novel metabolites and describe the biochemical pathways that led to their biosynthesis and 3) predict the characteristic mass spectrometry (MS) spectra of many of these novel metabolites. In order to fully explain why these objectives were pursued and how the resulting software works, it is important to provide some more background on the two areas that are being most highly impacted by this work: 1) metabolomics and 2) cheminformatics. A detailed introduction to these two very closely connected fields is given in the following pages.

1.2 A Brief Introduction to Metabolomics

1.2.1 The Metabolome

The metabolome is defined as the complete set of low-molecular-weight metabolites (<1500Da) found within a cell, tissue, biological sample, or organism at any given point of time under a given set of physiological conditions (11). The size of the metabolome varies significantly from one species to another. In simple organisms such as *Escherichia coli* (12) up to 3,800 small-molecules have been associated with their metabolome. In more complex organisms such as *Homo sapiens*, more than 42,000 different small molecules have been mapped to the human metabolome (7). Interestingly, humans (and other mammals) are not the most metabolically complex organisms. Rather, it appears that plants are. Up to 200,000 metabolites have been catalogued in the plant kingdom (1). Unlike the genome, the size and character of the metabolome is quite variable and it changes throughout the day and the life course of an organism. It is also dependent on the sensitivity of the measurement technology (more sensitive techniques yield more metabolites) as well as on the tissue or biofluid that is being measured. For instance, ~3,100 metabolites have been found in human urine (13) but just 468 have been found in human cerebrospinal fluid (14).

The metabolome is often subdivided into two classes: the primary metabolome and the secondary metabolome. Primary metabolites constitute the primary metabolome and are directly involved in an organism's development, growth, and reproduction. They include amino acids, nucleotides, sugars and lipids, among others. Secondary metabolites include, but are not limited to, transformation products of primary metabolites. Typically, these so-called secondary metabolites are not essential for the processes of development, growth, and reproduction; however, they can play an important role in other physiological processes, such as combatting environmental stressors. Indeed, the absence of secondary metabolites can seriously impair an organism's survivability as these molecules often play defensive roles as a response to environmental insults. For instance, many plants produce polyphenols as secondary metabolites. Polyphenols are actually antibiotic compounds that are particularly effective at combatting bacterial pathogens or fungal infections. Likewise many bacteria produce secondary metabolites to fight off competing bacteria. For instance, *Streptomyces kausaensis* produces Kanamycin A, an antibiotic that exhibits strong antimicrobial activity against a number of competing aerobic bacteria, including *Pseudomonas aeruginosa* (15).

While many metabolites may be classified according to their importance, another approach to partitioning metabolites is according to their origin. In this regard there are two kinds of metabolomes: 1) the endogenous metabolome and 2) the exogenous metabolome. The endogenous metabolome of an organism corresponds to metabolites produced by its natural metabolic processes. Endogenous metabolites include such compounds as essential amino acids, vitamins, and hormones. On the other hand, the exogenous metabolome can be defined as the set of metabolites or chemicals directly derived from the environment or produced via industrial processes. These include compounds found in foods, drugs, pollutants, toxins and metabolites produced by the colonic flora (16). While different metabolites can have different (or even multiple) origins, the methods available to detect and quantify metabolites are largely indifferent to their origins. In the following section we will review some of the most common techniques and methods used to detect metabolites in biological systems.

1.2.2 Metabolomics Technologies

Metabolomics employs a variety of analytical chemistry technologies to measure and identify small molecules from biological samples. The two most popular analytical approaches are nuclear magnetic resonance (NMR), and mass spectrometry (MS) (17). NMR was the first technology to be used in metabolomics and for many years, most of the papers published on metabolomics came from NMR laboratories. NMR has a number of appealing advantages for metabolite measurement, including facile sample preparation, excellent reproducibility and non-destructive analysis. NMR also gives researchers the ability to accurately and simultaneously identify and quantify large numbers of metabolites (17). In addition, NMR is particularly useful in the structure elucidation of unknown metabolites or chemicals (18). However, NMR is a much less sensitive technique than MS. Most NMR instruments cannot detect metabolite concentrations <1 □M, while many MS instruments can often detect metabolites with concentrations <1 nM. In addition to the very high sensitivity of MS instruments, the availability of a wide selection of separation and ionization methods allows MS techniques to identify a larger pool of metabolites than what is typically available through NMR. Mass spectrometry is usually coupled with a chromatography technique, such as liquid chromatography (LC-MS), or gas chromatography (GC-MS) to facilitate compound separation prior to mass analysis. Because of the vast array of chemical classes and physico-chemical properties seen in metabolites, it is common to combine various types of analytical techniques in a metabolomic study. Indeed, it is widely known that certain analytical methods are intrinsically better than other analytical methods for measuring certain types of metabolites. For instance, NMR is best suited for analyzing sugars and alcohols, GC-MS is best suited for measuring volatile metabolites (e.g. short-chain fatty acids, organic acids, and certain biogenic amines) (4,19), while LC-MS is best suited for measuring larger lipophilic molecules (e.g. lipids) (4,20).

Just as there are two general types of metabolomics platforms (NMR and MS), there are also two types of metabolomic approaches for characterizing metabolites. One approach is called "untargeted metabolomics" and the other is called "targeted metabolomics". Untargeted metabolomics aims at comprehensively analysing all measurable molecules in a sample, including unknown chemicals or as yet unidentified compounds. In the standard untargeted metabolomics workflow, large numbers of measured (NMR or MS) spectra from two or more groups (or cohorts) are first processed by statistical analysis tools (10). This allows researchers to rapidly find important peaks or key features in these spectra that can be used to differentiate one cohort from another (e.g. a disease state from healthy control). Untargeted metabolomics is particularly appealing because it offers the possibility of discovering novel metabolites that had no previous disease association or no known biological function. Both NMR and MS techniques are commonly used in untargeted metabolomics studies. In contrast to untargeted studies, targeted metabolomics tries to measure a very specific set of pre-selected or well-defined metabolites (e.g. fatty acids, and steroids) using pure, authentic (often isotopically labelled) chemical standards. The measured metabolite concentrations are then used to make diagnoses, identify phenotypes or draw biologically interesting conclusions (2,17). In targeted metabolomics the most important information is contained in the accurately measured metabolite concentrations rather than in the metabolite identities. Obviously in targeted metabolomics it is not possible to identify novel metabolites but it is still possible to identify novel metabolite-disease or metabolite-phenotype associations. LC-MS and GC-MS are usually the best-suited methods for targeted metabolomic studies.

1.2.3 Metabolomics Applications

As noted earlier, metabolomics uses a variety of analytical techniques to study the alterations in metabolic pathways brought on by genetic or environmental perturbations. For this reason, metabolomics has found numerous applications in the fields of drug discovery and development (3), nutrition research (21), and environmental monitoring

(22). Historically, metabolomics got its start in drug discovery and development, with most of the early metabolomics papers focusing on the application of metabolomics to drug metabolism and drug toxicity. More recently, the role of altered metabolism as a disease indicator has re-energized interest in metabolomics for drug discovery and therapeutic intervention. This is nicely illustrated by a series of studies by Stanley Hazen and his team (23-25) who connected metabolomics with atherosclerosis and drug discovery. Atherosclerosis is a cardiovascular pathology in which plaques build up inside the arteries, eventually leading to myocardial infarction and stroke. Wang *et al.* (25) found a strong correlation between high plasma concentrations of a compound known as trimethylamine N-Oxide (TMAO) and atherosclerosis, in both rats and humans. TMAO is a liver by-product of trimethylamine (TMA), which is a gut microbial metabolite of phosphatidylcholines originating from the diet (e.g. meat, cheese or eggs). This finding suggested that enzymes capable of synthesizing TMAO or its precursors could serve as potential drug targets (3). As a general rule, if the biosynthetic pathway for a metabolite is known, a list of such enzymes can be easily retrieved. With regard to the aforementioned example, flavin monooxygenase 3 (a liver enzyme) and choline-TMA lyase (a gut microbial enzyme) quickly emerged as two potential targets. In fact, Wang *et al.* (26) were able to identify a potent inhibitor of choline TMA-lyase called 3,3-dimethylbutanol, a natural product found in olive oil. These findings suggest that 3,3-dimethylbutanol, if used as a drug or nutrient supplement, may reduce the risk of atherosclerosis. Interestingly, olive oil is an essential component to the Mediterranean diet, a diet that is widely known to improve heart health and prevent atherosclerosis(27).

In addition to using metabolomics in the pursuit of new drugs or novel drug targets, there has been an increasing interest in using metabolomics to improve our understanding of nutrition. Nutritional metabolomic studies can be divided into two categories: 1) dietary intervention studies, and 2) biomarker discovery studies (4). Dietary intervention studies aim at studying the effects of certain diets or food items in metabolic pathways. One interesting example of metabolomics being used in dietary intervention studies relates to the effect of diet on estrogen levels in women. Estrogen levels are strongly correlated to breast cancer risk with higher values increasing the risk (28). Thus, a diet that decreases estrogen levels could potentially reduce or prevent breast cancer. Carruba *et al.* (29) conducted a randomized intervention study (the MetDiet project) that aimed at assessing the effect of a Mediterranean diet on the profiles of endogenous estrogens in healthy postmenopausal women. They reported a significant decrease in estrogen levels in women who followed the 6 month long diet (which is rich in vegetable fat and proteins) compared to women who followed a normal diet rich in animal fat and proteins.

In contrast to dietary intervention studies, food biomarker discovery studies are focused on finding unique compounds indicative of certain diets. They usually involve dosing individuals with certain specific foods followed by the collection of biofluids (e.g. urine, breast milk, and blood) over a period of time. The collected samples are then analysed via an untargeted metabolomics approach in order to identify compounds that are specific to the intake of that food. The Food Biomarker Alliance (FoodBAll) project is an initiative involving 22 partners (universities, government organizations and companies) from 11 countries, which aims at finding specific food biomarkers. This consortium is developing chemical as well as data exchange platforms and resources that provide metabolomic data for food, food compounds, as well as their transformation products (32). FoodBAll has also recently released a publication and a database (called ExposomeExplorer) describing more than 100 different food consumption chemicals (30).

Besides its applications in drug discovery and nutritional science, metabolomics has also proven to be an important vehicle to perform comprehensive environmental monitoring. A recent demonstration was provided by Boersma *et al.*, who used ^{19}F-NMR metabolomics to identify intermediates involved the microbial bioconversion of fluorophenols (31). This work led to a number of suggested biodegradation pathways for these molecules. Over the past decade there has also been a rapid increase in the number metabolomic studies aimed at studying the toxic effects of pollutants on the health of various organisms. These include studies that have looked at the effects of pesticides in humans (32), insects (33), and plants (34). Other studies that have explored the influence of plasticizers as synthetic estrogen analogs (e.g. polychlorinated biphenyls, phthalates, and bisphenol A) on humans (35-38). For instance, Lu *et al*. (35) used a LC/MS/MS approach to investigate the relationship between dermal exposure to bisphenol A and oxidative damage in humans (36). A recent metabolomics study by Xia *et al*. (38) showed that DBP (di-N-butyl-phthalate) altered the citrate cycle, as well as the amino acid, purine and lipid metabolism in the serum and placenta of exposed mice. The study clearly suggested potentially teratogenic effects for this commonly used plasticizer.

1.2.4 Challenges in Metabolomics

In spite of the many recent advances in metabolomics and its applications in many different life science disciplines, there are a number of limitations that still hamper the routine measurement and characterization of metabolites in biological samples. In particular, the instruments used to perform metabolomic measurements such as MS and NMR spectrometers are very expensive (>$300,000). Furthermore, the sample preparation and extraction processes are often very time consuming and labour intensive. In addition, the required sample volumes, particularly for NMR-based metabolomics, can be significant (0.3–0.7 mL), separation times on HPLC or GC systems can be lengthy (30-60 minutes) and compound identification can be very slow, ranging from hours to days (3).

The identification of compounds is probably the most significant and persistent challenge facing the entire field of metabolomics. Compound identification, whether by NMR or MS, often involves comparing the experimentally obtained spectrum to other reference spectra or spectral libraries of known compounds. Unfortunately, there are a limited number of NMR or MS spectral libraries. For this reason, the compound identification is often very slow and arduous – especially for untargeted metabolomics studies. Publicly available mass spectral libraries such as the NIST/EPA/NIH Mass Spectral Library (39), and MassBank of North America (also known as MoNA) (40), the METLIN database (41), and the Golm metabolome database (42), among others, have authentic mass spectra of perhaps 70,000 different metabolites. This collection of authentic spectra represents only a tiny fraction (<0.07%) of the total number of chemicals known or catalogued in various chemical substance databases such as

PubChem, the Chemical Abstracts Service (CAS) registry database (43), or ChemSpider (44). The PubChem and CAS databases contain data for nearly 130 million compounds. Assuming that it costs ~$100 to acquire or synthesize a few milligrams of each compound and estimating that it takes a day to measure the MS/MS spectra of 100 compounds on a single MS spectrometer, then if one had full-time access to 100 mass spectrometers, it would take >35 years and cost >$10 billion to generate the corresponding MS/MS spectra for all 130 million known compounds using existing MS technologies. Given the science budgets of today, this would clearly be an impossible task. As a result, the metabolomics community has been looking to develop computational or *in silico* methods to generate MS spectra from known (or predicted) chemical structures. There are now several computational tools (45) and databases (46) that have been developed to aid in the automated *in-silico* generation of MS spectra. However, these tools are still limited in terms of their performance and/or the range of chemical classes that they cover.

Another challenge confronting the metabolomics community is the fact that even with today's very large chemical databases (>130 million compounds) and the existence of very high resolution mass spectrometers, fewer than 10% of the features seen in a typical metabolomics MS spectrum can be matched to a known molecular weight or a known chemical formula in these massive databases. This suggests that existing chemical databases do not have the structures (or masses) for 90% of detectable metabolites. These unknown molecules are often called the "Dark Matter" of the metabolome (47). It has been suggested that many of these unknown compounds are biotransformation products or secondary metabolites derived from well-known metabolites, contaminants or food constituents. So even though new methods to predict MS and MS/MS spectra are starting to appear, the capacity to accurately predict or generate biologically feasible metabolites is only just beginning. Developing *in silico* methods to predict/generate biologically feasible metabolites will require the systematic collection, classification, and analysis of known compounds and their corresponding biosynthesis or biodegradation reactions. This will have to be done by exploiting the knowledge of known biosynthetic pathways as well as the knowledge of known enzyme mechanisms. This modelling process would require that we: 1) link a compound to a biosynthetic pathway, 2) predict its metabolizing enzymes, and 3) predict the structure of its biotransformation products. As yet, there are only a few programs or software tools that are capable of performing these tasks for a select fraction of molecules (48-51).

The computer-aided expansion of mass spectral libraries using biologically feasible metabolites and putative metabolic pathways could greatly improve the breadth of metabolome coverage (perhaps rising from 10% to 50% or more). It would also improve the appeal of metabolomics for many researchers. However, each of these efforts requires that large volumes of chemical data must be represented and stored in a computer-readable format. In addition, the rules and chemical features that must be generated to perform these predictive tasks requires the rapid computation of various physico-chemical properties. Furthermore, these data must be easily searchable in order to accurately identify metabolites. Thanks to recent advances in the field known as cheminformatics, this is now possible.

1.3 A Brief Introduction to Cheminformatics

Cheminformatics involves the use of computers and computer programs to facilitate the collection, storage, analysis, and manipulation of large volumes of chemical data (52). Chemical data usually includes chemical formulas, chemical identifiers (e.g. CAS-numbers), chemical structures, physico-chemical properties, chemical spectra, and biochemical data. Cheminformatics first emerged in the early 1960s in order to help chemists manage the enormous amounts of chemical data arising from industrial drug, textile and polymer production efforts (53). Until recently, the scope of cheminformatics was mostly limited to facilitating chemical structure searching and chemical property prediction. However, with the advent in high-throughput drug screening and high throughput metabolomics (54,55), cheminformatics has had to become increasingly more sophisticated and increasingly more integrated with other fields. These fields include molecular modelling, computational chemistry, bioinformatics, machine learning and systems biology (54,56).

1.3.1 Representation of Chemical Entities

Chemical structures are a foundational concept to both chemistry and cheminformatics. Simply stated, the chemical structure describes the atoms within a chemical compound and the bonds that link them together. The chemical structure is the starting point for a number of cheminformatics tasks such as data exchange and storage, chemical similarity searches, the systematic naming of compounds, the prediction of physico-chemical properties and the calculation of chemical spectra. There are three main structural representations that can be used to describe a compound's structure: its nomenclature, its line notations, and its connection tables.

1.3.1.1 Chemical Nomenclature

Historically chemicals were given trivial or Latin names tied to their appearance, colour or origin. From these names, chemical symbols and other shorter representations evolved. However, the need for a systematic nomenclature inspired Lavoisier and other early chemists, at the end of the 18th century, to develop a method for naming compounds using both a stem and a specifying part (e.g.: sodium nitrate) (34). Later on, the International Union of Pure and Applied Chemistry (IUPAC) developed a chemical nomenclature system (57,58) that describes specific molecular fragments using a number of expressions from a well-defined vocabulary. One advantage of this kind of structured nomenclature is that, in many cases, it can give an idea of the nature, number, and relative positioning of the chemical constituents in a molecule (see Figure 1.1). A disadvantage of IUPAC names is that they can be very long and cumbersome, which is why trivial names are still used today. Overall, chemical names remain the most commonly used identification method for chemists and biochemists. However, chemical names do not allow the direct extraction of additional information about the molecule, such as molecular weight. Furthermore, many names can be attributed to the same compound. It is because of these limitations to chemical names that other approaches, such as line notations for describing chemical structures, have emerged.

1.3.1.2 Line Notations

Line notations allow chemical structures to be represented as a linear sequence of numbers, letters or special characters. The IUPAC nomenclature scheme is an example of a line or line-like notation. Another example of chemical line notation is the Wiswesser Line Notation (WLN). This early line notation scheme was based on using the elements and functional groups (e.g. alkyl halides) present within the molecule to describe the chemical entity (59). WLN has been used in well-known chemical databases such as the Chemical Structure Index (60). Another kind of line notation is known as the Representation of Organic Structures Description Arranged Linearly (ROSDAL).

Trivial Name: Ampicillin

IUPAC Name: (2S,5R,6R)-6-[(2R)-2-amino-2-phenylacetamido]-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid

Connection Table (Molfile)

```
 Mrv16b2802061712082D

 24 26  0  0  1  0          999 V2000
    2.5600    0.7927    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9469    1.3448    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5600    1.8968    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4620    0.6773    0.0000 S   0  0  0  0  0  0  0  0  0  0  0  0
    0.6773    0.9323    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
   ...
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  2  4  1  0  0  0  0
  4  5  1  0  0  0  0
 ....
```

Isomeric SMILES: CC1(C)S[C@@H]2[C@H](NC(=O)[C@H](N)C3=CC=CC=C3)C(=O)N2[C@H]1C(O)=O

InChI: InChI=1S/C16H19N3O4S/c1-16(2)11(15(22)23)19-13(21)10(14(19)24-16)18-12(20)9(17)8-6-4-3-5-7-8/h3-7,9-11,14H,17H2,1-2H3,(H,18,20)(H,22,23)/t9-,10-,11+,14-/m1/s1
InChIKey: AVKUERGKIZMTKX-NJBDSQKTSA-N

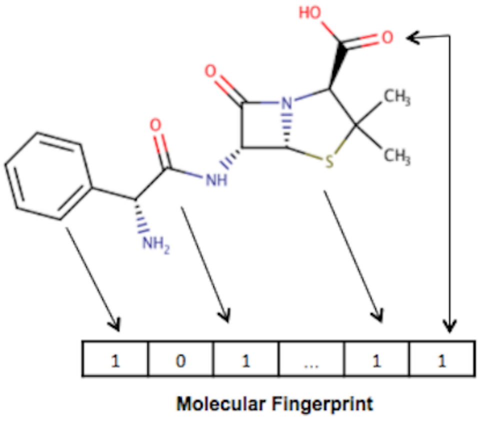


Figure 1.1 Chemical representations of Valclavam.

This particular syntax was used in the Beilstein-DIALOG system (61). Another syntax known as the SYBYL Line Notation (SLN), developed by Tripos (62), is a nice example of a popular line notation still used today. The Simplified Molecular Input Line Entry System (SMILES) notation is one of the most popular structure representations in chemistry (63). Developed in 1986 by Weiniger *et al.*, the SMILES language uses a small set of only six rules to convert a chemical structure into a character string. Because the SMILES notation is relatively easy to implement and is software/hardware-independent, it has become a popular format for the representation and exchange of chemical structure information (see Figure 1.1). As a result, SMILES has been implemented in numerous cheminformatics tools, such as the Chemistry Development Kit (CDK) (64), RDKit (65), Open Babel (66), and ChemAxon's JChem (67).

Over the past two decades, several extensions of SMILES have been developed. Some examples include the popular SMiles ARbitrary Target Specification (SMARTS) (68) and SMIRKS languages (69). These extensions allow the specific representation of structure-based chemical classes, and generic chemical reactions, respectively. Despite the popularity of SMILES, this line notation scheme is not without limitations. As is the case for chemical names, a single molecule can be represented by more than one SMILES string. Several algorithms have been developed that generate a unique (canonical) SMILES string for a given molecule. Nevertheless, it is still common to see different algorithms generating different SMILES for the same molecule. Another limitation of SMILES strings is that there is no standard way to handle aromaticity. These drawbacks have led to the development of more formalized versions of the IUPAC line notation: an early version called the International Chemical Identifier (InChI) (70), and later, a compressed version called the InChIKey (70).

The InChI string of a chemical substance is a standard identifier that describes its structure in terms of layers or delimiters that encode specific information. This includes the atoms and their bond connectivity, tautomeric information, isotope information, stereochemistry, and electronic charge information (see Figure 1.1). An advantage of this kind of delimiter prefix format is that it gives the user the possibility to use a wildcard search that is restrained to certain layers. For instance, by focusing only on the main and charge layers of a given compound, one could expand a search to retrieve not only that molecule from a database, but also its stereoisomers. A disadvantage of the InChI string is that it can be very long, and therefore difficult for database indexing.

An InChIKey is a 27-character-long hashed version of the standard InChI string (an analogue to the canonical SMILES). It contains three dash-separated blocks of fourteen characters, ten characters, and one character, respectively (see Figure 1.1). The first block represents a hashed version of the connectivity information, the second represents a hashed version of the four remaining layers of the corresponding InChI, and the last contains a character that indicates the InChI version that is used. An advantage of InChIKeys is that they allow for the efficient indexing of databases. Moreover, structures can be easily searched from the web using this key. InChIKeys, like any other form of line notation, present some limitations. First, an InChI string or identifier cannot be reconstructed from the corresponding InChIKey. As a result, the InChIKey must always be linked to its original InChI string, which can be converted into other structure representation formats. Second, the limited length of the key increases the possibility of "collision", meaning that two different molecules might have the same InChIKey. This, however, is extremely rare. Pletnev *et al.* (56) estimated the probability of a first block collision at 0.014% in a database containing 100,000,000 compounds (e.g. PubChem).

1.3.1.3 Connection Tables

Connection tables describe chemical structures by providing a list of atoms and a list of bonds, where atoms and bonds are described further on a single line each (see Figure 1.1). The atom description usually contains the index, symbol, type, coordinates and the atom charge, among other values. The bond description provides the indices of the connected atoms, and the type of bond (single, double, aromatic, etc.) (53,71). Depending on the format and software used, a connection table can be extended by adding other information, such as stereochemistry. Connection tables can store 2D as well as 3D coordinates. Moreover, connection tables offer a concise but complete coding of the chemical structure, and are easily processed by computers. However, connection tables are not easily interpretable by humans. An example of connection table is illustrated in Figure 1.2. For chemical connection tables, the MDL (Molecular Design Limited) Molfile format has become the *de facto* standard. Among other extensions of the MDL Molfile, cheminformaticians often use the Structure Data file (SDF) and the MDL Reaction formats, both developed by MDL Information Systems, for the storage of one or multiple molecules, and the storage of information related to a single chemical reaction, respectively (71).

1.3.1.4 Other Special Representations

Other models for chemical structure representation also exist. For instance, fragment codes (72) are indexed expressions of chemical structures based on specific chemical characteristics. These characteristics, which can be pre-defined, include functional groups, ring systems, and other assemblies of atoms. Fragment codes are still used in chemical patent databases (such as the Derwent database) today (72). One limitation of this representation is that one code can describe different molecules, since there is no information about the interconnectivity of the chemical fragments.

Markush structures, named after Dr. Eugene A Markush, are another special type of chemical structure representation (73-75). A Markush structure denotes a virtual set of compounds, represented by a core backbone, and radical groups at specific positions, selected from a finite list of potential radical substituents. This finite list can be enumerated by a Marksuh interpreter to generate all possible structures. Such representations are often used in chemical patent claims (74). An advantage of Markush structures is that they can encode large numbers of molecules in a single file using a single representation. However, because of the potentially large number of encoded structures, a large amount of space is required if the structures must be stored in a database. Unlike most other known chemical structure representations there is no freely available software that handles Markush representations. An example of a commercial software package that handles Markush structures is ChemAxon's JChem Base (76).

Molecular fingerprints can be viewed as abstract, vector-based representations of the structure and properties of a molecule (see Figure 1.1). Fingerprints are usually deployed as binary descriptors for machine learning workflows to help predict the biological activities and physico-chemical properties of chemical compounds. It is possible to generate two-dimensional (2D) or three-dimensional (3D) fingerprints, depending on the method used to transform the molecular representation into data bits (77). Two-dimensional fingerprints are most common, although 3D fingerprints are often used to represent pharmacophore features for drug research. Three approaches exist for fingerprint construction: 1) substructure key-based fingerprints, 2) topological (or path-based) fingerprints, and 3) circular fingerprints (77,78). Substructure key-based fingerprints usually report the presence (binary = 1) or absence (binary = 0) of structural fragments within the molecule of interest. They can also be customized to report the number of occurrences of certain structural fragments as well. Examples include the MACCS fingerprint (166 bits) (79), and the PubChem fingerprint (881 bits) (9).

Topological fingerprints encode hashed versions of molecular features, captured linearly up to a given length. Because topological fingerprints are hashed, it is nearly impossible to decipher the fragments contained in the molecule directly from the fingerprints. The most popular topological fingerprint is the Daylight Fingerprint (80). Circular fingerprints are also hashed, but rather than capturing topological features, these capture the environment of each atom up to a pre-determined radius. Examples of circular fingerprints include the Molprint2D fingerprint (81).

The popularity of chemical fingerprints in cheminformatics is due to the fact that they are computationally efficient, and they can be designed using expert intuition. Fingerprints have been successfully used to implement structure similarity searches, to predict biological activities (82), and to help perform virtual screening (78). A number of widely used software tools either use or generate molecular fingerprints including CDK(64), RDKit (65), PaDEL (83), Open Babel (66), and ChemAxon's JChem Base (76).

Clearly there are a plethora of formats to represent molecular structures. Each has its own advantages and disadvantages. In many cases it is also possible to convert a molecular structure from one format to another, but occasionally with the risk of losing some information (e.g. from Molfile to SMILES). Many of the structure representation schemes, as well as many of the methods for their interconversion, have been implemented in a number of popular cheminformatics software packages. Commercially available packages that offer extensive support for structure representation and format conversion include ChemAxon's Marvin Suite (84), and the OEChem toolkit. Freely available packages include the Chemistry Development Kit (CDK) (64), Open Babel (66), RDKit (65), and OPSIN (85), among others.

1.3.2 Representing Chemical Reactions

Simply stated, chemical reactions represent the transformation of one chemical compound to another. The starting substances that begin the reaction are called substrates or reactants. The end substances are called products. In a single reaction, one or more substrates can be transformed into one or more products. Based on the overall change in molecularity, chemical reactions can be classified into three different categories: 1) substitution reactions, where an atom of the substrate is replaced by another atom or group of atoms (e.g. nucleophilic hydroxylation), 2) addition reactions, where an atom (or a group of atoms) is added to a molecule with one or more multiple bonds (e.g. alkene hydration), and 3) elimination reactions, where two substituents are removed from a molecule (e.g. dehydrohalogenation of alkyl halides). In addition to these general chemical reaction categories, there are also biochemical reactions or metabolic reactions. Biochemical or metabolic reactions can be classified either as catabolic or anabolic. In catabolic reactions, large molecules are broken down to produce energy (e.g. hydrolysis). In anabolic reactions, energy is consumed to synthesize a larger molecule from smaller components (e.g. glucuronidation).

As mentioned earlier, chemical reactions can be represented using computer-readable languages or computer-compatible representations. The three most widely used computer-readable reaction languages are known as: SMIRKS, Rxnfile, and RDfile. Here we will focus on describing the SMIRKS language and refer readers to the literature for information about the latter two (71,86). The SMIRKS language or line notation is an extension of the SMILES line notation and a subset of the SMARTS chemical language (69). The SMIRKS language is designed to represent generic reactions, which consist of one or more atom and bond changes. It is also designed to capture or describe a substrate SMARTS pattern upon which the chemical changes are made. The SMARTS pattern defines a set of structural constraints that any substrate must fulfil in order for it to be a candidate for the encoded reaction. An example of a chemical reaction for an organophosphorothioate compound and its SMIRKS representation is illustrated in Figure 1.2. Organophosphorothioate insecticides represent an important class of insecticides that are widely used today. Some examples include Chlorpyrifos, Diazinon, and Disulfoton. These compounds are known to undergo enzymatic desulfurization of the organophosphorothioate group. The structure of the organophosphorithioate group relative to the substituents for Chlorpyrifos (R1= 2,3,5-trichloropyridine, R2=R3=methyl), and Diazinon (R1=2-isopropyl-4-methylpyrimidine, R2=R3=ethyl), is illustrated in Figure 1.2.a. Figure 1.2.b shows the atom mapping, which is an essential part of the SMIRKS representation, as it dictates what atoms are transformed and how. The mapped atoms must be present in both sides of the equation (substrates and products). Since the (=S) group is replaced with the (=O) group (and not just displaced to another part of the substrate), this particular site of metabolism was not indexed. The substituents R1, R2, and R3 can be any substituents (including H atoms). It is worth noting here that the origin of the substituting oxygen need not be specified in the equation, and is thus missing on the left side of the SMIRKS string.

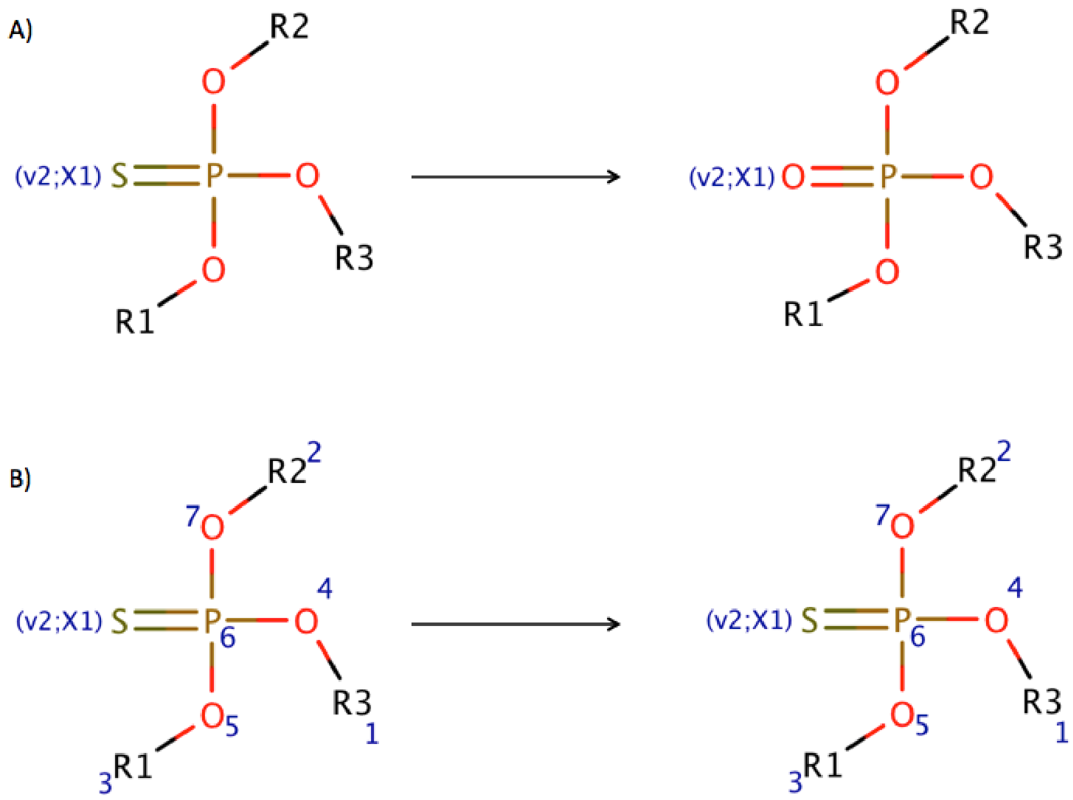


Figure 1.2 SMIRKS representation of the desulfurization of organophosphorothioates. A) The abstract chemical structure representation of Chlorpyrifos (R1= 2,3,5-trichloropyridine, R2=R3=methyl), and Diazinon (R1=2-isopropyl-4-methylpyrimidine, R2=R3=ethyl). B) Atom mapping for the reactant and product. Mapped atoms must be present in both part of the equation. The sulfur atom in the substrate (not indexed) is replaced by an oxygen atom to form the product.

The resulting SMIRKS notation is as follows:

$$[\#6:1][\#8,\#16;A;X2:2][P;X4:3]([\#8:4])([\#8:5]) = [S;v2X1] >>$$

$$[\#6:1][\#8,\#16;A;X2:2][P;X4:3]([\#8:4])([\#8:5]) = [O;X1]$$

Based on the SMIRKS representation, one could easily infer the SMARTS string of the reactant(s) and product(s) by simply removing the atom indices.

1.3.3 Molecular Similarity and Structure Search

1.3.3.1 Molecular Similarity

Just like biologists or physicists, chemists are often interested in grouping or comparing new entities on the basis of their similarity to previously known entities. Molecular similarity can be assessed or ascertained through a variety of approaches including topological features, structural coordinates, physico-chemical properties, or biological properties. The motivation for using molecular similarity measurements is that similar molecules likely possess similar properties. Therefore, the assessment of molecular similarity is one of the most important and frequently performed tasks in all of cheminformatics. Indeed molecular similarity is commonly used to classify or categorize chemical compounds, to predict physico-chemical properties, to search for biologically active analogues in a database, or to cluster large numbers of molecules into more coherent groups or categories (87,88).

Similarity assessment requires a well-defined molecular representation schema and a well-defined similarity function or distance measure. In some cases, a weighting function can be introduced to the distance measure to assign a specific weight to each individual feature in the molecular representation (89). In chemistry the pairwise similarity value usually varies from “0” for completely dissimilar molecules, to “1” for identical molecules. A very popular structure representation used for similarity assessment is the molecular fingerprint notation. Both 2D and 3D fingerprints can be used. These fingerprint representations can be used to depict the presence or absence of a large variety of structural patterns (e.g. structural fingerprints) or encode an independent list of paths within the molecule of interest (e.g. hashed fingerprint). This makes fingerprint notation more suitable for a global similarity assessment, which considers molecules in their entirety. Pharmacophores, which can be defined as the spatial arrangement of the atoms or groups responsible for a molecule's biological activity, are generally better suited for a local (or sub-molecular) similarity assessment, which focuses on regions of the molecules of interest (89).

There are a number of functions commonly used to assess chemical similarity, with the most popular being the Tanimoto coefficient (77,89). Given two vectors of real values A and B, the Tanimoto coefficient (Tc_G) is defined as:

$$Tc_G(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A^2 + \sum_{i=1}^{n} B^2 - \sum_{i=1}^{n} A_i B_i}$$

For binary vectors or fingerprints, the Tanimoto coefficient will range between 0 and 1. There are a number of well-known cheminformatics software packages that offer similarity assessment capabilities including CDK (64), RDKit (65), Open Babel (66), and ChemAxon's JChem (67).

1.3.3.2 Structure Searching

In contrast to molecular similarity analysis, structure-searching methods are used to detect the presence or absence of specific structural fragments in a molecule. There are three main types of structure search tasks: 1) similarity searches, 2) substructure searches, and 3) superstructure searches. Similarity searches used to retrieve compounds (targets) from a dataset, which are structurally similar to the compound of interest (query). The compounds that are returned are called "hits". The number and type of hits varies depending on the similarity threshold that is set, and the similarity function that is used.

Moreover, when fingerprints are used, the nature of the fingerprints and the type of information they contain are a major determinant of the similarity assessment. By default, various chemical search engines use the Tanimoto function applied on chemical hashed fingerprints.

In contrast to similarity searches, substructure searches are used to retrieve compounds from a dataset that contains the full structure of the query. Depending on the choices made by the user, specific features such as stereochemistry and charge distribution can be taken into consideration during the structure search protocol. Unlike substructure searching, superstructure searching looks for those targets from a dataset that are contained within the molecule of interest or the chemical query. Several search engines, frameworks, and cheminformatics tools offer at least one of these three chemical search types, including OrChem (90), the RDKit database cartridge (91), MatchMol (92), the Molecular Database Framework (93), and ChemAxon's JChem Base (76). These software tools are all open-source, except ChemAxon's JChem Base (which is free for academics only). As we will describe later in this chapter, a number of publicly available chemical databases also offer extensive chemical search capabilities. These facilitate the retrieval of similar compounds, as well as the selection of sub- or superstructures for subsequent analysis. Later, we will describe how structure searches can be used to classify chemical entities.

1.3.4 Chemical Databases

Over the last decade, large numbers of databases have been developed to address the burgeoning data needs and data generation bottlenecks that are appearing in both the life sciences and the physical sciences. Chemical and biochemical databases have seen tremendous growth over the past few years, particularly with the advent of more publicly available resources on chemical structures, properties, industrial roles and biological functions. I will briefly describe three different types of chemical or biochemical databases that are relevant for my work, namely: 1) chemical substance databases, 2) spectral databases, and 3) pathway databases (94).

1.3.4.1 Chemical Substance Databases

Chemical substance databases are largely chemical structure resources containing general information about pure chemical substances. There are two types of chemical substance databases: those that are general (covering everything) and those that are specific (tied to a specific organism, a class of compounds or a specific theme). General substance databases try to collate all known (or reported) chemicals regardless of their origin or purpose. In this regard the emphasis is on breadth (largest number of chemicals) over depth (detailed descriptions or facts about the compounds). Most general chemical substance databases are extremely large, with 10's of millions of compounds in their repository. Examples of well-known general chemical databases include the Beilstein database (61), PubChem (9), ChemSpider (44), ChEMBL (95) and ZINC (96). Some general chemical substance databases, such as the Beilstein database, include additional data such as physico-chemical properties, chemical reactions and associated substances, Other general chemical substance databases, such as the PubChem database, include BioAssay data, as well as information related to safety and hazard, biomolecular interactions, and vendor information.

Specific or thematic chemical databases tend to be smaller, but much richer in their content. The Human Metabolome Database (HMDB) is an example of a specific or thematic chemical database (8), as it focuses on human metabolites and associates the structure of a small molecule with its physico-chemical properties, NMR and MS spectra, biological functions, biosynthetic pathways, biofluid concentrations, and many other biochemical or biomedical features. In addition to HMDB, several other specific or thematic chemical databases are commonly used in the field of metabolomics, which provide (bio-)chemical, physiological, metabolomic, toxicology, pharmacogenomic, bioactivity, and/or compositional data for specific types of compounds. The Chemical Entities of Biological Interest (ChEBI) database organizes >40,000 chemicals in a comprehensive chemical ontology (97). The LIPIDMAPS database covers >40,000 structures of biologically relevant lipids, organized in the lipid-specific LIPIDMAPS chemical ontology (98). DrugBank provides comprehensive data for >8,000 drugs and drug metabolites (99). The *E. coli* Metabolome Database (ECMDB) provides data for >3,000 *E. coli* derived compounds (12). The Toxic Exposome Database covers >3,600 toxins (100). FooDB is a database that provides compositional, biochemical, and physiological information for >26,600 food compounds (101). MetaboLights is a database for metabolic experiments and detailed information that covers >23,300 compounds (102) KEGG is a database that provides various types of chemical, biological and genomic data covering ~18,000 metabolites (103). Finally, the KnapSack Core database covers >111,000 metabolite-species relationships for ~51,000 compounds and ~ 22,400 species. These databases tend to focus on chemicals found in specific organisms or chemicals used for specific industrial purposes. Many of the specific chemical substance databases available today have been developed in response to the specific needs metabolomics researchers.

1.3.4.2 Spectral Databases

Spectral databases are repositories of spectroscopic data (NMR, GC-MS or LC-MS) that were recorded using pure, authentic compounds under well-defined conditions. The purpose of spectral databases is to facilitate the comparison of known compounds and known spectra with unknown spectra to help solve compound identification tasks. Spectral databases are primarily organized on the basis on the technique used to generate the spectra. For instance, there are several well-known, publicly available NMR spectral databases including the BioMagResBank (metabolites) (104), NMRShiftDB (105), the MMCD (106), the HMDB (8), and the COLMAR database (107) that contain hundreds or even thousands of 1D and 2D NMR spectra collected on authentic compound standards at different spectrometer frequencies. Most of these databases permit users to query the resource using compound structures or spectral chemical shifts.

There are also a large number of GC-MS and/or LC-MS spectral databases with similar querying capabilities. These include the NIST/EPA/NIH Mass Spectral Library (39), the Golm database (108), MassBank (109), the METLIN database (110) and the MassBank of North America (MoNA) (40). These databases contain tens of thousands of experimental MS spectra collected under a variety of experimental conditions using a variety of different MS instruments. There are also a number of spectral databases that contain "predicted" MS spectra or spectra generated through computational methods for compounds that are known to exist but which do not have measured MS/MS or EI-MS spectra. One such resource is LipidBlast, a freely accessible database covering >212,000 *in-silico* generated tandem mass spectra for 119,200 lipids (46). Another resource is *CFM-ID*, a web server for mass spectral prediction and compound identification, which has a database containing >140,000 spectra predicted by *CFM-ID* for >51,000 compounds (111). As with the chemical substance databases described above, many of the spectral databases available today have been developed in response to the specific needs of metabolomics researchers.

1.3.4.3 Pathways Databases

Pathway databases capture and depict information about chemical or biochemical processes that occur within cells or tissues. Many pathway databases are built around chemical reactions or chemical processes. Pathway database are very visual resources with coloured, interactive graphs and pictures of pathways being the main content of most pathway databases. In the field of metabolomics, the most important pathway databases are those that contain information about metabolic pathways. Many metabolic pathway databases cover the metabolism of multiple organisms and most allow one to search for specific enzymes, metabolites or genes related to a given pathway. The primary role of metabolic pathways databases in metabolomics is to assist with the biological or functional interpretation of metabolomic data or metabolite lists. In addition to their role in biological interpretation, metabolic pathway databases can also be used to suggest annotations for an incomplete metabolic pathway. This can be done by comparing and analysing more complete pathways in closely related organisms. Among the most popular pathway databases are the Kyoto Encyclopaedia of Genes and Genomes database (KEGG) (103) described above, the MetaCyc database that contains that describes >2,400 metabolic pathways (112), the Reactome Pathway database (113), the BioCarta pathway database (114), and the Small Molecule Pathway Database (SMPDB) (115).

1.4 Chemical Taxonomies and Ontologies

1.4.1 Defining Taxonomies and Ontologies

One way of establishing order in a complex field, or finding order among complex interactions, is to develop taxonomies and ontologies. A taxonomy is a classification system that organizes objects into a hierarchy, using a well-defined set of rules and a controlled vocabulary. Taxonomic classification has been used in many fields of science for hundreds of years. One of the best-known examples of a taxonomy is the Linnaean taxonomy for biological species classification (116). In contrast to a taxonomy, an ontology is a formal way of describing concepts or objects as well as the relationships between them. Ontologies usually share the hierarchical structure of taxonomies; however, taxonomies, often use more than one relationship type to link concepts to one another, within a domain or between domains. Ontologies can serve as standardized dictionaries of terms, and allow the sharing and reuse of knowledge derived from data. The best-known biological ontology is the Gene Ontology originally developed by Michael Ashburner (117). The Gene Ontology (GO) was designed to standardize the representation of genes and gene product attributes across species and across databases.

1.4.2 Taxonomies and Ontologies in Chemistry

Biologists have a very long and successful history of developing effective and efficient ontologies and taxonomies to help improve the understanding and exchange of biological data. For instance ontologies and formats have been developed to represent, organize and exchange data related to genes (e.g.: the Gene Ontology (117)), biological pathways (e.g.: BioPAX (118), the Pathway Ontology (119)), diseases (e.g.: the Disease Ontology (120)), and other concepts. On the other hand, while chemists have been very successful at developing a standard nomenclature system (i.e. the IUPAC nomenclature), and standardized structure representation formats (e.g. SMILES, standard InChI), there is still no standard chemical ontology or taxonomy. Whenever chemistry has interfaced with biology, there has often been an attempt to create some kind of domain-specific taxonomy or ontology. For instance, pharmacists and medicinal chemists tend to group drugs into pharmacological classes (e.g. non-steroidal anti-inflammatory drugs, antidepressants), and biochemists tend to group biochemicals into groups based on their biological or nutritional role (e.g. vitamins, amino acids, hormones). Unfortunately, there is no simple one-to-one mapping for these different classification schemes. Furthermore, most schemes are limited to small numbers of very domain-specific molecules. Thus, in recent years, chemists have been increasingly interested in developing a more uniform or generic chemical taxonomy and a better defined chemical ontology (97,121).

It is generally agreed that, for chemistry (116,122-124), the best route is to classify chemical compounds is according to their structures. Structure-based classification (as opposed to functional classification) provides important insights not only into a compound's chemical content and relationships, but also their interactions with macromolecules. One example of a structure-based classification scheme is the Fragment Code system. The Fragment Code system was one of the earliest classification systems used in chemistry (72). It consists of >2,000 numerical codes that correspond to specific chemically significant structure fragments. However, the system is now considered out-dated and overly complex. More recently, ChEBI developed a well-defined chemical ontology to help classify or cluster chemicals. The ChEBI ontology is now one of the most widely used chemical ontologies today. It consists of >20,000 terms that classify compounds into three sub-ontologies: structure (e.g. alpha-amino acid), roles (e.g. analgesic), and subatomic particles (e.g. fermion). Despite its extensive and well-developed structure, the ChEBI ontology has only been applied to a small set of compounds, namely the 43,000 chemicals of biological interest found in the ChEBI database. Furthermore, the assignment of compounds to this ontology requires teams of curators who must manually annotate the compounds. As a result, the process is time consuming and error-prone (97). Given that there are >100 million known chemical compounds (9) and given that thousands of new chemical entities are being described or synthesized every week (125), it is not likely that the ChEBI ontology could ever be manually applied to another 100 million compounds. This issue has led to increased interest in the development of a computer-based, structure-driven chemical ontology/taxonomy.

1.4.3 Developing Ontologies and Taxonomies

Developing a taxonomy or an ontology requires defining a scope, collecting or defining concepts and properties, as well as determining the relationships that link these concepts with one another. An ideal chemical taxonomy should hierarchically organize chemical classes based on their structural features (e.g. alpha amino acids, and N-acyl-alpha amino acids). The main relationship type in a taxonomy or ontology is the transitive *"is_a"* relationship, which implies the following axioms: Given three chemical classes *C_1*, *C_2*, and *C_3*:

$C_1\ is\_a\ C_2 \leftrightarrow\ \forall\, x \in\ C_1, x\ \in\ C_2$ **(1),**

$C_1\ is\_a\ C_2 \wedge\ C_2\ is\_a\ C_3\ =>\ C_1\ is\_a\ C_3$ **(2)**

This means that C_1 is a chemical subclass of C_2 if and only if every chemical entity x that satisfies the characteristic structural properties of C_1 also satisfies those of C_2 (1). Moreover, if $C_2 \in C_3$, the chemical entity x also satisfies the structural properties of C_3 (2). We have described a number of computer-interpretable languages that can be used to represent molecules (SMILES, InChI), and most generic chemical classes (SMARTS). Moreover, we also described the superstructure search operations, which can be used to verify whether a given query molecule Q satisfies the structural constraints of, or if it contains the molecule encoded in a target molecule T. Both Q and T can be either a chemical entity or a structural pattern. These languages and methods can contribute to the design of a structure-based chemical taxonomy, a structure-based chemical ontology, as well as the class assignment of chemical entities (see section 1.3).

One of the main advantages of using a structure-based chemical ontology/taxonomy comes from the fact that molecules from the same chemical (structure) class are more likely to undergo the same types of biological transformations (or biotransformations), to belong to the same biosynthetic pathways, and to have similar biological or biochemical functions. For instance, many drugs that belong to the structural class of 5-aryl-1,4-benzodiazepines (such as triazolam and diazepam) are known to bind GABA-A receptors. They also possess anxiolytic properties and are used as sedatives (99,126). Another example of a large and important class of compounds that undergoes similar biotransformations or belongs to a common biosynthetic pathway is the triacylglycerols (also known as triglycerides). Triacylglycerols, such as triarachidin, can be synthesized by human Glycerol-3-phosphate acyltransferase 3 (GPAT3), and function as energy source and membrane stabilizers (127,128). This specific information about triarachidin can be easily “transferred” to other structurally related triacylglycerols – so long as the compound of interest is robustly identified as being a member of the class of triacylglycerols.

As will be noted later in this chapter, chemicals with similar structures or similar chemical physico-properties will also have similar MS/MS or EI-MS spectra. This is particularly true for lipids from the same chemical class (e.g.: triacylglycerols, or phosphatidylcholines) as they often have very consistent and predictable fragmentation patterns (129). This principle has been used to build lipid-specific MS spectral databases, such as LipidBlast (46). Because structure-based chemical taxonomies or ontologies can be used to organize compounds according to their structural properties, they can also help to study the metabolism and the identification of compounds.

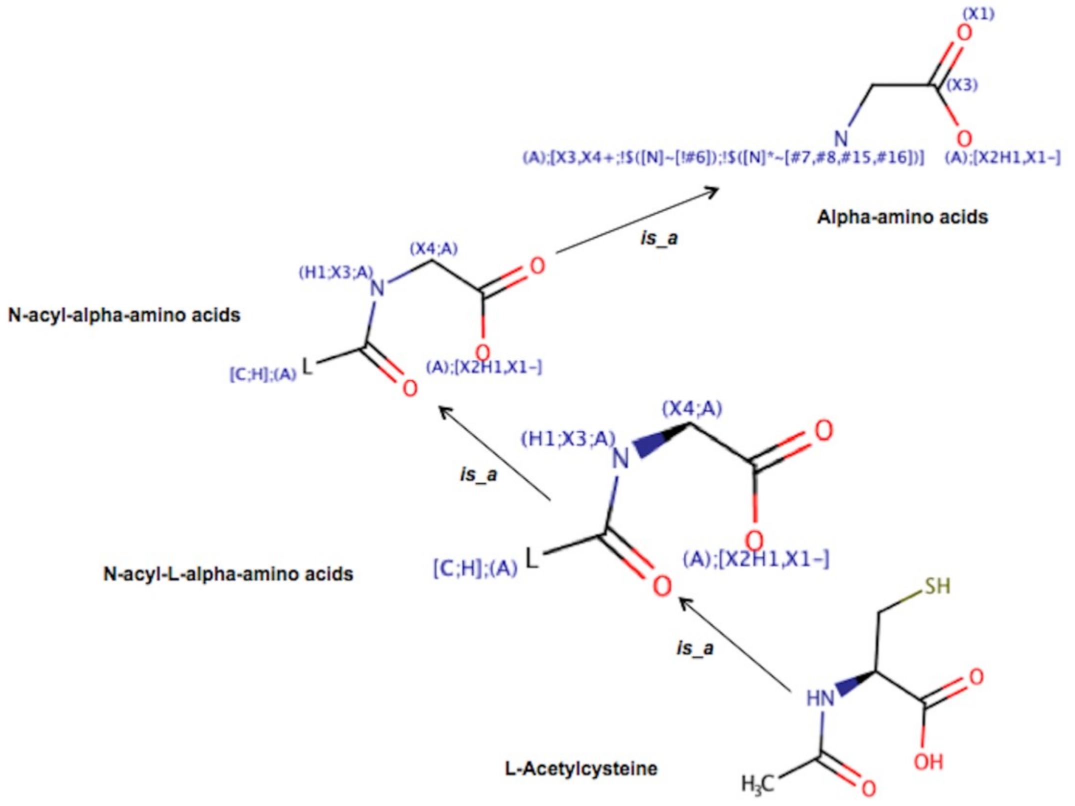


Figure 1. 3 Chemical representations of the taxonomical relationships between L-acetylcysteine and Alpha-amino acids.

1.5 Computational Prediction of Metabolism

Metabolism is defined as the sum of all chemical reactions that occur within a cell or living organism to maintain life (130). Metabolism regulates the production and consumption of energy, the delivery of chemical entities within and between cells, the activation and/or detoxification of chemicals, the elimination of waste, the defence against pathogens, and more. In most cases, metabolic reactions are catalyzed by enzymes that interact with substrates, thereby increasing the reaction rate, without being consumed. The enormous variety of enzymatic chemical reactions (e.g. oxidation, reduction, and hydrolysis) that occur within an organism, coupled with the multitude of parameters that influence the activity and substrate specificity of their catalysts (e.g. pH, polymorphism, disease state, substrate concentration) has made it difficult to not only characterize existing metabolites and metabolic pathways, but even more difficult to predict the chemical consequences of known metabolic reactions.

While the biotransformation pathways of essential metabolites (e.g. alpha-amino acids), as well as many well-known secondary metabolites, have been well studied for more than 70 years, there is still a large gap in our knowledge about other, less common or lesser-known metabolites. Indeed, scientists are still discovering novel metabolic reactions in well-studied model organisms such as *Escherichia coli* (131,132). Moreover, given than many animals, including humans, are constantly exposed to thousands of foreign compounds (133), it is likely that these compounds are also being metabolically transformed. For most of these xenobiotics, the metabolic fate in humans is only partially known, if at all. While some of these metabolites may be toxic, others may provide beneficial effects for a certain amount of time. For this reason, there is a growing interest in learning what these compounds are, how they are formed and how they may interact with other proteins and enzymes in the body. While experimental efforts (through metabolomics studies) are providing some of these answers, recent advances in computational chemistry, cheminformatics, and machine learning are also allowing scientists to answer these questions as well.

1.5.1 Overview of Xenobiotic Metabolism

The biotransformation of xenobiotics is needed not only for the extraction of nutritional value (if there is any) but also for the removal of compounds that cannot be used metabolically or compounds that are potentially harmful. Typically, lipophilic xenobiotics (i.e. tending to dissolve in fats, oils, or lipids) are transformed into hydrophilic products (i.e. tending to dissolve in water) that can be easily excreted (some volatile compounds however, are not transformed and are eliminated through the lungs). On the other hand, hydrophilic xenobiotics are typically hydrolysed in the liver or processed by the gut microflora. The catalytic reactions that contribute to the xenobiotic metabolism can be divided into two categories: 1) Phase I and 2) Phase II reactions. Phase I reactions tend to render the lipophilic xenobiotics more reactive by adding or modifying functional groups, such as the amino-, hydroxyl-, or the carboxyl group. Some examples of Phase I reactions include aliphatic hydroxylation, reductive dehalogenation, and epoxide hydrolysis. Such reactions are predominantly catalyzed by cytochrome P450 (CYP) enzymes that are capable of activating or inactivating xenobiotics as well as endobiotics. CYP enzymes execute nearly 90% of xenobiotic metabolism and most of the Phase I oxidative reactions (134). In particular, nine cytochrome P450 isozymes account for the majority of those reactions (CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4). Other enzymes that catalyze Phase I reactions include esterases, alcohol dehydrogenases, and flavin-monooxygenases. Although Phase I reactions usually result in inactive or less toxic metabolites, some Phase I metabolites have also been found to actually (or accidentally) generate more toxic compounds than their parent molecules. For instance, acetaminophen is an analgesic pharmaceutical drug that is oxidized in the liver by CYP1A2, CYP2A6, and CYP2E1 to produce the hepatotoxic metabolite N-acetyl-p-benzoquinone imine (NAPQI) (135). Such biotransformations are unwanted; therefore the identification of toxic xenobiotic by-products at an early stage of drug development has become a major issue in the pharmaceutical industry.

In Phase II reactions, the more reactive metabolites are conjugated to cofactors, making them less toxic, more hydrophilic, and thus easier to eliminate. Some of the more common Phase II reactions include the conjugations of xenobiotics to glucuronic acid (glucuronidation), sulphate (sulfation), a methyl group (methylation), and glutathione. Because of its toxicity, the acetaminophen metabolite NAPQI is readily inactivated via conjugation to glutathione, and eliminated. Phase II metabolism does not always occur after Phase I metabolism. Certain compounds can be conjugated and eliminated without undergoing any Phase I reaction. For instance, acetaminophen can directly undergo glucuronidation or sulfation, and be excreted. The metabolism of acetaminophen in humans is illustrated in Figure 1.4.

The richest source of enzymes catalyzing xenobiotic metabolism in humans is the liver. In the liver, the majority of these enzymes are located in the endoplasmic reticulum (136). Until recently, most studies have focused on the liver as the factory of xenobiotic-derived metabolites. However, many xenobiotic processing enzymes also reside in the lungs, the kidneys, and the gut. The latter is of particular interest as it provides an anaerobic environment for a large microbial population.

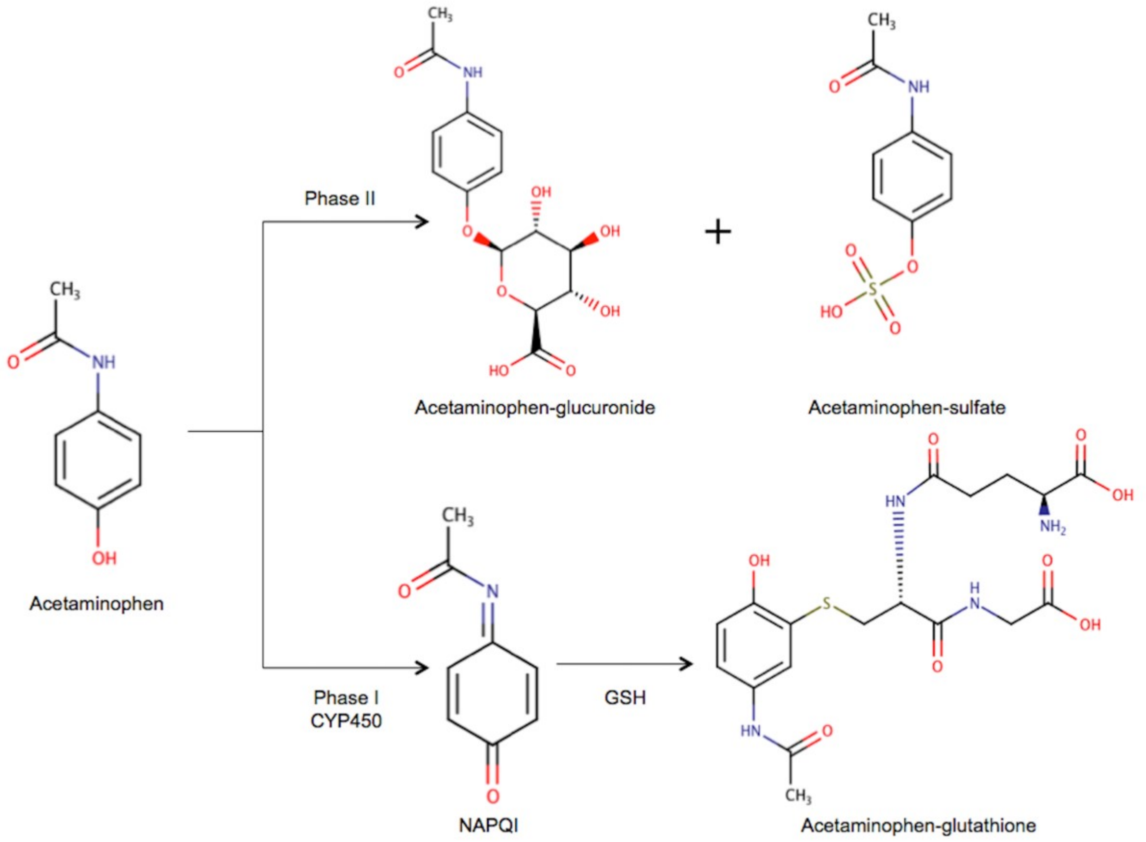


Figure 1.4 Metabolism of Acetaminophen.

In humans, ~100 trillion microbes accounting for 500-1000 species constitute the gut microbial population (137,138). This population is referred to as the gut microbiome. In humans the gut microbiome is dominated by *Firmicutes* and *Bacteroides* species. The gut microbes are of great importance, partly because they are capable of metabolizing not only many endogenous compounds, but also a variety of bioactive chemicals that cannot be synthesized or processed by their host. These chemicals include food derivatives, food additives and plant metabolites, such as polyphenols. Some of the resulting bioactive compounds enable the gut microbiome to connect to the immune system and to the brain, thereby affecting the host in significant and often unpredictable ways (139). As a result, the gut microbiome and the compounds produced by the microbiome have been associated with a number of diseases, such as autism, celiac disease and asthma (140,141). Given the importance of Phase I/II metabolism as well as the impact of gut metabolism on human health, it is clear that a better understanding of xenobiotic metabolism needs to be established.

Over the last few decades, several computational tools have been developed to predict or model the metabolism of xenobiotics. They use a variety of approaches, depending on the particular task at hand as well as the types, and amount of information available. In the next section, we will discuss some of the common approaches and software used for metabolism prediction.

1.5.2 Approaches for *in silico* Metabolism Prediction

The prediction of how a molecule is metabolized by a certain enzyme requires determining whether and how the molecule binds to that enzyme, the mode of interaction between the molecule and the enzyme (e.g. substrate, inhibitor, inducer), what atoms are expected to react (i.e. is a site of metabolism or SoM), what reactions apply, and finally, what products will be formed (142). A wide variety of computational approaches have been used to predict the metabolism of xenobiotics. These include knowledge-base systems, shape-based systems, reactivity models, data mining and machine learning, docking, and molecular interaction fields (MIF) (49,143-146). In this section, we will briefly describe a subset of these systems with a particular focus on 1) knowledge-based systems, 2) reactivity model approaches, 3) data mining and machine learning methods, and 4) hybrid approaches. For a more detailed description of the different approaches mentioned, the reader is referred to a number of excellent reviews on the subject (142,147).

1.5.2.1 Knowledge–Based Systems

Knowledge-based systems for metabolism prediction are based on dictionaries of rules devised by human experts. They are usually coupled with a reasoning engine that applies those rules to predict the metabolites of a given compound. For instance, CYP2D6 is known to catalyze the N-hydroxylation of anilines (135). Therefore, a reasoning engine connected to a knowledge base that contains this information could predict that CYP2D6 likely catalyzes the N-hydroxylation of procainamide (an aniline) to produce procainamide hydroxylamine. This will not only predict the site of metabolism, but the structure of the product. An advantage of such systems is that they provide the user with supporting evidence for the prediction. This not only provides a justification and a biological pathway for the product, it also provides feedback that could assist experts in updating the rules for more accurate predictions. A disadvantage is that such an approach can lead to the prediction of too many false positives. For instance, the presence of other groups in an aniline-containing molecule can force the molecule to adopt a certain orientation leading to the enzyme to prioritize other reactions. To circumvent this, certain systems use a ranking method based on the probability of occurrence for specific reactions that are applicable to the compound of interest. Only the reactions with a likelihood of occurrence greater than a certain threshold, or those reacting sites with a higher priority, are selected and ultimately transformed *in silico*. Some examples of knowledge-based systems include the commercial packages known as MetabolExpert and *Meteor Nexus* (50,148). These knowledge-based systems also take into account physico-chemical properties, such as the molecular surface area, and the logP of the molecule. These properties can also play a role in the correct prediction of the resulting by-products.

1.5.2.2 Reactivity Models

For a small molecule to undergo a catalytic transformation, it must enter the binding pocket of the target enzyme. If it properly fits within the pocket and the interaction is strong enough, then the reaction can be easily catalyzed. Thus, the electronic structure of the substrate molecule as well as the target enzyme is a major determinant of the enzyme-substrate interaction. Reactivity models use steric accessibility descriptors to predict metabolism. This is often facilitated by the implementation of quantum mechanical methods that simulate the electronic structure of the enzyme-substrate system (147). An advantage of such models is that they can accurately predict which atoms in a molecule will be modified by a specific enzyme (142). However, the need for quantum calculations implies that one must have a very good understanding of the enzyme system and its structure. Moreover, the optimization of the mathematical functions that can calculate the energy barrier for the enzyme-substrate system is time consuming and difficult to perform automatically (142). Thus, most reactivity-based prediction tools tend to use pre-computed or approximated activation energy values, to accelerate their calculations. The program known as *CypScore* is a tool that uses atomic reactivity descriptors to generate individual models for the most important CYP450 oxidation reactions, including aliphatic hydroxylation, S-oxidation, among others. This method allows one to find the reactions that are most likely to occur, as well as the specific SoMs (Sites of Metabolism) within a substrate of interest (149).

1.5.2.3 Data Mining and Machine Learning Approaches

Machine learning approaches to metabolism prediction rely on statistical models built by analysing biotransformation databases containing substrates, products and catalyzing enzymes. These models can be used to calculate the probability of each atom in the molecule of interest to be a site of metabolism (SoM). For machine learning to work, the properties of known SoMs, and a description of their atomic environment must be assembled and stored, often as fingerprints. Additionally, a list of reactions reported at each specific SoM must also be stored in the database (150). In this way, the machine learning is often combined with the data mining, to help generate a robust predictor. In the machine learning process, every given atom has its computed fingerprint compared to that of previously determined SoMs to predict whether it is a reaction centre, and what reactions it might undergo. An advantage of machine learning methods is that they can reduce the number of false positive SoMs and metabolic products. However, the predictions cannot be made for atoms where the corresponding fingerprint is not represented in the database. In machine learning, the development of a good predictive model requires a very large and comprehensive reaction/metabolism database. The most comprehensive and detailed biotransformation database to date is the BIOVIA Metabolite database (151), a commercial resource with >100,000 xenobiotics and biotransformations. Indeed, because of the importance to the drug industry, most of the comprehensive reaction databases are only available as commercial products. Publicly available resources that provide xenobiotic transformation data include DrugBank (99), SuperCYP (152), ChEMBL (95), and XMetDB (153). Computational tools that use the data mining and machine learning approaches to predict metabolism include MetaPrint2D and MetaPrint2D-React (51,154).

1.5.2.4 Hybrid Approaches

As discussed earlier in this section, several conditions need to be fulfilled for a metabolic reaction to occur. The most important ones are the chemical reactivity, and the solvent accessibility of the molecule of interest. However, focusing only on one condition could lead to outright failure or serious underperformance of the computational tool. Moreover, each one of the aforementioned approaches presents some limitations (142). Thus, it has become a very common strategy to develop hybrid systems that rely on the combination of several approaches. Examples of hybrid systems include SMARTCyp (146), MetaSite (145), and RS-predictor (155), which combine machine learning and quantum chemical atom descriptors. SMARTCyp, in particular, allows the prediction of xenobiotic metabolism by the CYP450 isoforms CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C19, CYP2E1, and CYP3A4. MetaSite combines molecular interaction field (MIF)-based modules, used for the characterization of protein-ligand with quantum-chemical and knowledge-based modules (145). The UM-PPS (now EAWAG-BBD/PPS) system (49) uses knowledge-based and machine learning-based approaches. In particular, it uses a set of relative reasoning rules that were machine learned using >330 biotransformation rules in addition to >1,000 parent compounds and intermediates from University of Minnesota Biocatalysis/Biodegradation database. Another example is isoCYP (156), a tool that uses QSAR and machine learning (multinomial logistic regression, decision trees, and SVM) to predict human CYP isoform specificity for small molecules.

Table 1.1 Examples of computational tools for the prediction of small-molecule metabolism.

Software	Coverage	Approach	Licensing	Description
Pathway assignment tools				
TrackSM	11 KEGG metabolic classes	Machine learning	Free	Uses functional group composition of small molecules to predict metabolic pathway associations (48).
Enzyme-substrate predictors				
MetaPred	CYP1A2, 2C9, 2C19, 2D6, 3A4	Machine learning (SVM)	Free	Uses SVMs to predict whether a drug-like molecule is metabolized by up to 5 CYPs (157).
WhichCYP	CYP1A2, 2C9, 2C19, 2D6, 3A4	Machine learning (SVM)	Free	Uses SVMs to predict CYP inhibition (158).
isoCYP	CYP2C9, 2D6, 3A4	Hybrid approach	Commercial	Uses QSAR and machine learning to predict CYP isoform specificity (156).
SoM predictors				
MetaPrint2D	Phase I/II	Data mining	Free	Derives probability of biotransformations via data mining of atomic fingerprints (154,159).
SMARTCyp	CYP1A2, 2A6, 2B6, 2C8, 2C19, 2E1, 3A4	Hybrid approach	Free	Combines reactivity models and machine learning (146,160).
RS-Predictor	CYP1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, 3A4	Hybrid approach	Free	Combines SMARTCyp reactivity models with SVM models trained on topological descriptors (155).
Metabolite structure predictors				
Meteor Nexus	Phase I/II	Knowledge-based	Commercial	Uses biotransformation rules and considers LogP values as a filter(50,143,144).
MetabolExpert	Phase I/II	Knowledge-based	Commercial	Uses biotransformation rules and considers LogP values as a filter(148).
EAWAG-BBD/PPS		Knowledge-based	Free	Combines knowledge-based and machine learning to predict environmental microbial catabolism (49,161).
MetaPrint2D-React	Phase I/II	Data mining and machine learning	Free	Generate structures of likely metabolites based on the MetaPrin2D SoM predictions (159).

Table 1.1 provides a summary of the software tools already mentioned, as well as a number of other tools for metabolism prediction. In compiling this table the programs were grouped into 4 categories: 1) Pathway assignment tools; 2) Enzyme-substrate predictors; 3) SoM predictors and 4) Metabolite structure predictors.

1.5.3 Limitations of Currently Available Resources

No matter what approach is chosen for the development of metabolism prediction software, the availability of experimental biotransformation data is crucial. Unfortunately, there is a lack of publicly available biotransformation databases. For instance, XMetDB (153) is the only publicly available, web-based biotransformation database now available. However, it contains just 162 observations for CYP450-mediated metabolism of 117 xenobiotics. Moreover, there is no information about the types of chemical reactions (e.g. aromatic hydroxylation) that transform the substrates into their metabolites. While there are a number of biotransformation tables provided in books and journals, these are also very limited in scope and often hard to find. Given the small size of existing public resources, most scientists working in this area have often been forced to buy very expensive commercial databases. The problem is that if a scientist chooses to use a commercial database, the software license will typically not allow them to generate freely available software nor to share their structural data with other scientists. Given the current situation, most of the freely available metabolism prediction tools and resources perform substantially worse or are far more limited in scope than commercially available tools. Therefore, it is crucial to develop publicly available repositories that provide detailed and comprehensive metabolomic data.

Another limitation of publicly available tools is that they tend to focus on a single aspect of xenobiotic metabolism (e.g. human CYP450-mediated metabolism only or environmental metabolism only or microbial metabolism only) thus limiting their scope. Moreover, the metabolic predictions are often reaction- or enzyme-based, and do not take into account specific constraints that might be attributed to the system. For instance, many compounds will undergo different transformation pathways within the gut compared to the liver. Additionally, metabolites that leave the gut can be reabsorbed in the liver and vice versa, and be further metabolized. Therefore, it is important to develop tools that handle physiological inputs as well as molecular inputs.

Developing better and more comprehensive metabolism prediction tools is also important for the experimental metabolomics community. In particular, predicted metabolites and predicted structures can be combined with analytical methods such as mass spectrometry to facilitate the identification of previously unknown or uncharacterized compounds, as illustrated by Pelander *et al.* (162). Since the spectra obtained from biological samples in metabolomics studies can contain a good deal of background noise, it has become common to incorporate prior metabolic knowledge when designing strategies that use mass spectrometry to identify metabolites. For example, a list of masses for predicted metabolites could help to identify the peaks of interest in a parent/molecular ion mass spectrum, which then allows the selection of ions for subsequent MS/MS analysis (163). Therefore, a database containing the structure and physico-chemical properties for known compounds as well as the predicted structures and predicted spectral properties of theoretical (*in silico* metabolized) compounds would be a significant asset for the entire metabolomics community. However, to make this sort of database a reality it will be important not only to develop better quality structure/metabolism predictors, it will also be important to develop better MS spectral prediction tools as well.

1.6 Spectral Prediction and Metabolite Identification

In a mass spectrometry, molecules are ionized, and fragmented into pieces of different masses, yielding a fragmentation spectrum. The mass-to-charge (m/z) ratio of the parent ion or the molecular ion (before fragmentation) along with the m/z ratio of the fragments provides a great deal of structural information about the molecule. Indeed, if the parent ion mass is known to a high degree of precision, it is quite easy to determine its molecular formula. Furthermore, if the fragment ion masses are known it is often possible to identify particular moieties or substructures (sulfates, glucuronide additions, aromatic rings, etc.) within the molecule of interest. This information can be combined and, under favourable circumstances, it can allow skilled MS operators to unambiguously identify molecules. However, a number of factors can contribute to the challenge of identifying small molecules from EI-MS or MS/MS spectra. As mentioned earlier, the chemical properties of the molecules often dictate which ionization method to use. While Electron Ionization (EI) is often used for Gas Chromatography Mass Spectrometry (GC-MS) analysis of volatile and thermally stable compounds, electrospray ionization (ESI) (164) is typically used for LC-MS analysis of non-volatile compounds. EI is very reproducible and always performed at constant ionization energy of 70 eV. This results in fragment-rich spectra that are highly similar across instruments. However, the obtained GC-MS spectra often lack the molecular ion peak. This means that the mass of the precursor ion, which is important for subsequent identification, is often unknown (165). By comparison, Liquid Chromatography Mass Spectrometry (LC-MS) spectra obtained by tandem mass fragmentation (or collision induced dissociation -- CID) usually provide masses for the molecular ions as well as a smaller number of fragment ions. A disadvantage of ESI-MS spectra is that they are not as reproducible as EI-spectra, since the collision energies, the collision conditions and the fragmentation patterns can vary significantly across instruments. For this reason, LC-MS spectra of reference compounds are often measured at several fragmentation energies and, if possible, on several instruments (165).

As noted earlier, the number of reference compounds (of biological significance) with high quality MS/MS or EI-MS spectra is actually quite tiny (<5%) compared to the apparent size of the metabolome. Given the challenges of isolating/synthesizing compounds and experimentally collecting their MS spectra, there is a growing trend to develop computational approaches that can automatically predict/simulate MS fragmentation patterns from the millions of known and/or predicted compound structures. In this section, I will describe a number of approaches that are used to interpret and predict MS spectra. I will also describe how they have been implemented to facilitate metabolite identification.

1.6.1 Spectral Library Search

Spectral library searching and matching is the standard approach for compound identification in mass spectrometry. It involves generating the EI-MS or MS/MS spectrum of the pure (or presumably pure) compound and comparing it to reference MS spectra contained in a library collected under the same or similar experimental conditions. As with structure similarity searching discussed in section 1.2, spectral similarity searching requires a similarity function. Although it is the most straightforward approach for compound identification, the performance of a spectral library search algorithm is significantly dependent on the scoring function used for assessing spectral similarity. Several functions have been developed for the interpretation and comparison of both EI and MS/MS spectra. For EI-MS spectra, the Hertz similarity index, introduced in 1971, uses the weighted average ratio of the two spectra being compared (166). A more efficient algorithm is the Probability Based Matching (PBM) method, which examines the peaks in the spectrum of an unknown compound based on how significantly they contribute to the probability that the compound is present in the database (167,168). The dot product is another scoring function that takes the mass/charge ratio as well as the peak intensity. The dot product approach is widely used, and implemented in software for EI-MS metabolite identification such as *CFM-ID* (111). In recent experiments, the dot product was shown to perform better than the two previously mentioned functions (169) For LC-MS/MS spectra, the dot product, the PBM and the dot product are also often used (168,169). Because the fragmentation patterns in LC-MS/MS or ESI-MS are less reproducible than EI-MS spectra, reliable identification is often achieved by analysing mass spectra obtained at multiple collision energies (10, 20 and 40 eV). After the dot product between the query and each of the database spectra is calculated, the program returns a ranked list of spectra most similar to the query. This approach remains the most straightforward of all; however, it faces two challenges. First, only a small fraction of known compounds are covered in spectral databases so far. Freely available databases often cover only small number of compounds. For instance, the MassBank of North America provides experimental/predicted spectra for ~75,000 different compounds (40). This still represents only a small fraction of the >100 million known chemical compounds. Second, similar structures have similar spectra; this can lead to high misidentification rates, especially as the database becomes larger. Ideally one could design or select similarity functions that are more discriminative; but this has proven to be very difficult (165). Some algorithms may combine several scoring functions or apply machine learning to obtain better results (170)

1.6.2 Mass Spectral Classification

Because spectral similarity is correlated to structure similarity (171), the candidate list obtained from a conventional MS spectral search could be used to indicate the presence of common structural features (e.g.: functional groups, or substructures). Alternately it could be used to indicate similar chemical properties among the candidates and the query compound. These features or properties can be characteristic of a compound class that shows a specific m/z peak distribution. The mass spectral classification approach consists of predicting the substructures of a query compound or the compound classes it belongs to, given its MS-spectrum. This is usually done using a classifier that has been trained on the spectral library. Various machine learning methods, such as support vector machines or regression methods, can then be applied to the transformed data. For each compound class, a classifier can be trained on a set containing the vector of each spectrum, along with the annotation that specifies whether the corresponding compound belongs the class. Depending on the machine learning method used, the resulting classification model could return a yes/no answer, or a probability for the query compound to belong to the class that is being predicted. This method appears to work well for GC-MS as the EI fragmentation process is well understood and highly standardized (165). In particular, several reliable classifiers, such as the *Self-Training Interpretive and Retrieval System (STIRS)* (172), have developed to identify compounds from GC-MS spectra. Unfortunately, there are few classifiers for LC-MS spectral interpretation because the fragmentation is often not as reproducible.

1.6.3 *In Silico* Fragmentation

The lack of experimental reference spectra continues to be a bottleneck in compound identification. There is a clear need to expand MS spectral libraries so as to cover a much greater portion of chemical space. As we learned from the previous section (section 1.6.2) on mass spectral classification, molecular fragments can be quite helpful in metabolite identification. If one could accurately predict the MS fragmentation pattern of any known molecule, larger databases covering from 100s of thousands to millions of compounds could be created. These *in silico* expanded spectral databases would then be used for identifying unknown compounds through conventional spectral similarity searching methods. *In silico* fragmentation is particularly accurate and reliable for compounds that have consistent fragmentation patterns, such as lipids. Two main *in silico* fragmentation approaches have been developed so far - the rule-based methods and combinatorial fragmentation approaches.

The rule-based fragmentation approach uses generic or class-specific rules, usually extracted from mass spectrometry literature or learned from experimental data, to predict MS fragmentation spectra. Given a molecule with a known chemical structure, the molecule is scanned for one or more structural patterns that are typically produced in a mass spectrometer fragmentation process. The molecule is then fragmented according to these fragmentation rules to generate specific molecular fragments that serve to create a mass spectrum. This method was first implemented as a part of the DENDRAL project for the prediction of EI mass spectra in the late 1960’s (172). This concept has continued and many software tools use this approach (see Table 1.2), including the *Mass Frontier* spectral interpretation software (173). The *Mass Frontier* spectral interpretation software contains one of the largest EI/ESI fragmentation libraries available, with approximately 31,000 manually curated fragmentation schemes (173,174). *Mass Frontier* can predict the MS spectrum for a given molecular structure, and it can also identify a molecule from a given MS fragmentation spectrum by using curated rules developed from its massive spectral library and fragmentation tree collection. Rule-based fragmenters can generate large numbers of fragments, based on the number of rules that are applicable to the chemical structures. However, rule based methods have a number of limitations. First, although the fragmentation rules could be learned automatically, in reality, there is a need for expert curation. Moreover, rule-based methods are not particularly efficient when trying to predict novel structures that are not covered by existing rules. Furthermore, since the applied ionization method can force the rearrangements of fragments or influence the fragmentation of the molecule, the quality of the results can vary drastically from one method to another (165).

The combinatorial fragmentation approach uses computational “fragmenters” to cleave chemical bonds in a combinatorial fashion. The fragmentation is guided by a scoring function that assigns a penalty to each cleavage operation, depending on how easy it would break. The penalty function is learned through machine learning techniques based on a database of previously annotated fragments. Combinatorial fragmentation generally works under the assumption that most MS peaks correspond to compound fragments without structural rearrangements (175). Combinatorial fragmenters have been implemented in several software tools (see Table 1.2). Examples of tools using combinatorial fragmentation or an extension thereof include the *Fragment iDentificator (FiD)* (176), *MetFrag* (177) and *CFM-ID* (45,111,178). *CFM-ID* implements the competitive fragmentation modelling (CFM), which was first introduced by Allen *et al*. *CFM-ID* models the fragmentation process as a fixed length sequence of random fragmentation states. With its competitive fragmentation modelling approach, *CFM-ID* has been shown to outperform other tools such as *MetFrag*, *Mass Frontier*, and *MOLGEN-MS* (179) in compound identification from EI-spectra (178) For compound identification from ESI-MS spectra, *CFM-ID* was initially shown to perform better than *MetFrag* and *FingerID* (178) However, recent improvements to *MetFrag* and *CSI:FingerID* have allowed these programs to outperform *CFM-ID* in compound identification tasks based on ESI-MS spectra (180-182) An advantage of the combinatorial approach over rule-based approaches is that they can enumerate all possible molecular fragments. However, because the fragments are recursively broken, this often causes an exponential increase in time and computer resources needed. To circumvent this problem, one often selects the fragments that are most likely to occur.

Because both the rule-based and the combinatorial fragmentation approaches can generate large number of fragments for a molecule, they often achieve near-perfect recall; however, the likelihood of those fragments to occur is often significant only for a few of them. Therefore, both *in-silico* approaches achieve a low precision. To improve their overall performance a number of heuristics can be applied in the scoring functions. For instance, Ridder *et al*. proposed a simple but efficient function that would assign scores based on the type (e.g.: single, double, etc.) of the broken bond (183). To improve the performance of the fragmentation approaches even more, machine learning algorithms can be used to find bond-cleavage rates or cleavage events that will occur with minimal cost. One example of a combined model is a package known as *In Silico Identification Software (ISIS)*, which simulates the fragmentation of lipids (184).

1.6.4 Fragmentation Trees

Fragmentation trees are diagrams that are automatically generated solely based on the mass spectrum (or several mass spectra obtained at different energy levels), and the chemical formula of a compound of interest. The use of fragmentation trees to facilitate compound identification or classification was introduced by Böcker and Rashe (185) In a fragmentation tree, each node represents a fragment of the unknown compound, and contains its molecular formula. Each edge contains the molecular formulas of losses between the two fragments it connects. A fragmentation tree is computed by combinatorial optimization, using a scoring function. The task in generating a useful fragmentation tree is to find the tree that best explains the spectrum, according to the selected scoring function. An advantage of fragmentation trees is that one does not need the molecular mass of the compound of interest. Additionally, one can easily compare two compounds by aligning their fragmentation trees. This has the effect that similar fragmentation sequences can be identified and scored. One limitation of the fragmentation tree approach is that the number of possible fragments can be so large that it almost impossible to process, even for small molecules or for spectra with many peaks.

With the increased interest and activity in MS-based metabolomics over the last decade, a number of software tools to facilitate MS-based compound identification and MS-spectral prediction have been developed. Many of these implement at least some of the approaches described above. Table 1.2 provides a summary of the most popular software tools and computational resources for MS-spectral prediction and compound identification.

Table 1.2 Examples of computational tools for MS-spectral prediction and compound identification

Software	**Licensing**	**Description**
		Spectral search
CFM-ID	Free	Uses Jaccard score or dot product to deduce structural information corresponding to the most similar spectra from the spectral library (45,111,178).
MetFrag	Free	Uses an extension peak count scoring function that takes into account the number matching fragments and the bond dissociation energies (177).
NIST MS Interpreter	Free	Uses an optimized dot product function, and either molecular weight or "neutral loss" peaks to deduce structural features of the unknown compound (169).
		Mass spectral classification
STIRS		Combines a rule-based approach with machine learning to retrieve structural information for related EI spectra, and predict the molecular mass (172).
FingerID	Free	First predicts a set of molecular fingerprints from the spectrum of interest, which are then used to match against large molecular database, such as PubChem (186).
		In-silico fragmentation
Mass Frontier	Commercial	Predicts EI/ESI MS-spectra based on a library of ~ 31,000 manually curated fragmentation rules (173).
MASSIMO	Commercial	Automatically derives the fragmentation rules directly from experimental data to predict EI-MS fragmentation (187).
CFM-ID	Free	Uses combinatorial fragmentation with a cost function that considers functional group composition to predict MS-spectra in

		single-energy or combined-energy mode (45,111,178).
FiD	Free	Predicts MS-spectra based on a single-step or multi-step combinatorial fragmentation model (176).
MetFrag	Free	Comines combinatorial fragmentation and an extension of the peak count scoring function that takes into account the number matching fragments and the bond dissociation energies (177).
Fragmentation trees		
SIRIUS	Free	Based solely on a high-resolution isotope pattern of a molecule, it generates elemental compositions and calculates/ranks isotope patterns of relevant compositions (188).

1.7 Research Objectives

This chapter has provided a brief overview of both metabolomics and cheminformatics. It has also highlighted some of the existing computational tools and resources that can be used to tackle problems related to metabolomic data representation and management, metabolism prediction, and compound identification. While many of these computational tools and resources have had a significant and positive impact on the field of metabolomics, it is clear that they also have their limitations. In particular, the key issues identified were: 1) the lack of freely available tools and resources to help manage the description and classification of chemical compounds; 2) the lack of effective, open-access tools and databases to predict compound metabolism and compound structures arising from metabolic process; and 3) the lack of software to accurately predict MS/MS and EI-MS spectra from known (or predicted) chemical structures. Each of these limitations or shortcomings with existing software tools or databases impacts the other and so by addressing one, it helps resolve issues associated with the others.

To address these three outstanding issues in metabolomics and metabo-informatics (a branch of cheminformatics involving metabolomics), I have set out 3 specific objectives: 1) design and develop new software tools and resources to help manage the description and classification of chemical compounds; 2) develop open-access tools and databases to predict metabolites and metabolism for a broad range of metabolic processes; and 3) improve and extend existing software to predict MS/MS and EI-MS spectra from known (or predicted) chemical structures.

In working towards objective #1, I have developed a freely available structural chemical taxonomy and ontology, called ChemOnt. This resource permits the rapid, automated classification and description of nearly all (>100 million) known natural and synthetic chemicals according to their structure. It was developed in collaboration with scientists from a number of different institutes (NIH, EBI, UCSD) and different fields of chemistry and biology. Through these collaborative efforts, lookup tables were created to map *ChemOnt* to other existing ontologies, thereby moving *ChemOnt* a step closer to becoming the standard chemical ontology. In addition to creating *ChemOnt,* I also developed *ClassyFire*, a restful application/web server that uses *ChemOnt* to automatically classify any type of compound. *ClassyFire* has been used to classify >100 million compounds.

To address objective #2, I have developed *BioTransformer*, an open access software package that is able to predict the metabolism (and resulting structures) of both endogenous as well as exogenous compounds. *BioTransformer* integrates data mining and machine learning with a knowledge-based approach to predict: 1) human CYP450-mediated xenobiotic metabolism (Phase I metabolism), 2) human gut microbial metabolism of xenobiotics, 3) Phase II metabolism, and 4) promiscuous enzyme-based metabolism. In order to incorporate the large body of existing *a priori* knowledge, I constructed libraries of biotransformation rules and constraints, some of which were based on chemical classification data provided by *ClassyFire*. Furthermore, *BioTransformer* incorporates the UM-PPS set of rules for the prediction environmental microbial metabolism (49,161). The metabolites predicted by *BioTransformer* can be used to enrich chemical databases with new metabolite structures that, when annotated with spectral information, can facilitate novel metabolite identification in metabolomic studies.

To address objective #3, I have decided to modify an existing software tool, called *CFM-ID*. This package was recently shown to out-perform other well-established MS tools in the identification of never-before-seen compounds (111,178). It uses a competitive fragmentation modelling algorithm and machine learning to predict ESI-MS/MS and EI-MS spectra. However, for the structurally diverse and important class of lipids, *CFM-ID* does not perform well. Because of the length of the acyl chains in these compounds, the number of possible fragments that can be generated is very large, and thus requires excessive computing resources and time. Moreover, the predicted spectra often do not reflect the observed fragmentation of lipids. Fortunately, these metabolites are known to have consistent fragmentation patterns. By learning these rules and encoding them through cheminformatics methods, I was able to generate more accurate spectra, thereby improving the performance of *CFM-ID*. I also noted that compounds that belong to the same chemical class tend to have similar fragmentation patterns. Therefore, I hypothesized that chemical classification could further improve *CFM-ID*'s performance. Inspired by the work of Kind *et al*. (46,189), I have used chemical classification techniques (derived from Objective #1) to develop a module that uses the rule-based fragmentation approach to predict the ESI-MS-spectra of 26 classes of lipids under various conditions (i.e. adduct types). Moreover, I have expanded the spectral library used by *CFM-ID* to contain >50,000 ESI and EI spectra. I have also added chemical classification and citation data to guide the prioritization of candidates.

1.8 Thesis Outline

In this document, I will describe the work I have completed towards the development of the above-mentioned tools and resources. Most of the programs and databases I have developed have been either published or submitted for publication. As a result, this is a paper-based thesis. The document itself is organized as follows: Chapter 1 served as a general introduction in metabolomics and cheminformatics. It provides an extensive literature review that describes current progress in the areas of chemical taxonomies and ontologies, metabolism prediction, spectra prediction and metabolite identification. It also provides the thesis objectives and briefly summarizes the thesis results. Chapter 2 describes ChemOnt, a novel chemical ontology and *ClassyFire*, a computational tool for automated chemical classification using a comprehensive, computable taxonomy. Chapter 3 describes *BioTransformer*, a software tool for the automated prediction of Phase I, Phase II, microbial, and environmental metabolism. Chapter 4 describes *CFM-ID 3.0*, a web server that has been substantially enhanced for the prediction and peak annotation of ESI-MS- and EI-MS-spectra, as well as the identification of metabolites. Chapter 5 provides a general conclusion and future perspectives.

Chapter 2

ClassyFire and ChemOnt: automated chemical classification with a comprehensive, computable taxonomy[1]

[1]A version of this chapter has been published previously: Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, David S Wishart (2016); ClassyFire and ChemOnt: automated chemical classification with a comprehensive, computable taxonomy; Journal of Cheminformatics 8:61

2.1 Introduction

Taxonomies and ontologies organize complex knowledge about concepts and their relationships. Biology was one of the first fields to use these concepts. Taxonomies are simplistic schemes that help in the hierarchical classification of concepts or objects (190). They are usually limited to a specific domain and to a single relationship type connecting one node to another. Ontologies share the hierarchical structure of taxonomies. In contrast to taxonomies, however, they often have multiple relationship types and are really designed to provide a formal naming of the types, properties and interrelationships of entities or concepts in a specific discipline, domain or field of study (191,192). Moreover, ontologies provide a system to create relationships between concepts across different domains. Both taxonomies and ontologies can be used to help scientists explain, organize or improve their understanding of the natural world. Furthermore, taxonomies and ontologies can serve as standardized vocabularies to help provide inference/reasoning capabilities. In fact, taxonomies and ontologies are widely used in many scientific fields, including biology (the *Linnean* taxonomy) (116), geology (the BGS Rock classification scheme) (123), subatomic physics (the Eightfold way) (193), astronomy (the stellar classification system) (194,195) and pharmacology (the ATC drug classification system) (196). One of the most widely used ontologies is the Gene Ontology (GO) (117), which serves to annotate genes and their products in terms of their molecular functions, cellular locations, and biological processes. Given a specific enzyme, such as the human cytosolic phospholipase (PLA2G4A), and its GO annotation, one could infer the cellular location of its substrate PC(14:0/22:1(13Z)) (HMDB07887). Additionally, because PLA2G4A is annotated with the GO term “phospholipid catabolic process”, it could be inferred that PC(14:0/22:1(13Z)) is a product of this biological process.

While chemists have been very successful in developing a standardized nomenclature (IUPAC) and standardized methods for drawing or exchanging chemical structures (58,197), the field of chemistry still lacks a standardized, comprehensive, and clearly defined chemical taxonomy or chemical ontology to robustly characterize, classify and annotate chemical structures. Consequently, chemists from various chemistry specializations have often attempted to create domain-specific ontologies. For instance, medicinal chemists tend to classify chemicals according to their pharmaceutical activities (antihypertensives, antibacterials) (196), whereas biochemists tend to classify chemicals according to their biosynthetic origin (leukotrienes, nucleic acids, terpenoids) (198). Unfortunately, there is no simple one-to-one mapping for these different classification schemes, most of which are limited to very small numbers of domain-specific molecules. Thus, the last decade has seen a growing interest in developing a more universal chemical taxonomy and chemical ontology.

To date, most attempts aimed at classifying and describing chemical compounds have been structure-based. This is largely because the bioactivity of a compound is influenced by its structure (199). Moreover, the structure of a compound can be easily represented in various formats. Some examples of structure-based chemical classification or ontological schemes include the ChEBI ontology (200), the Medical Subject Heading (MeSH) thesaurus (121), and the LIPID MAPS classification scheme (198).

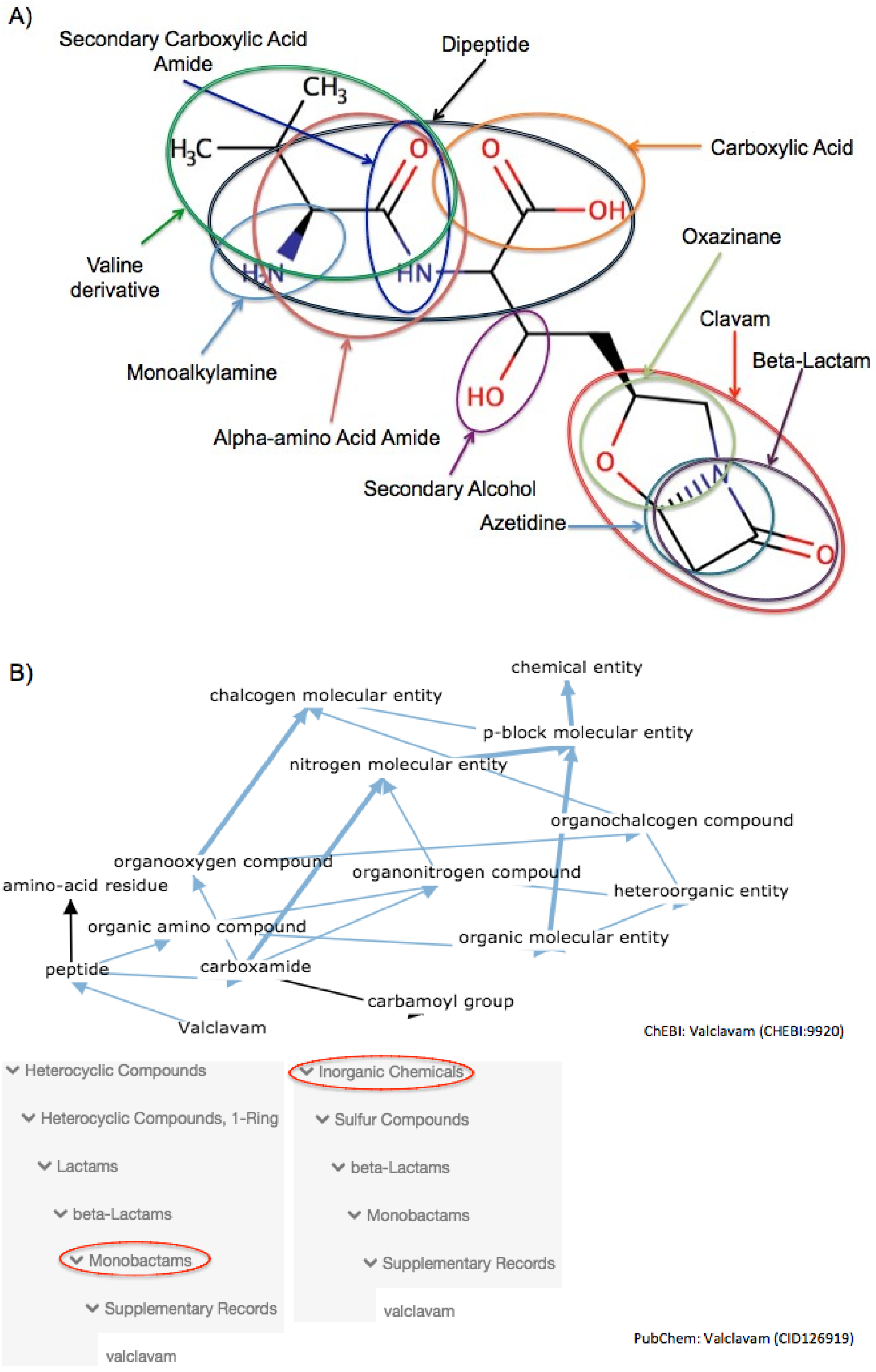


Figure 2.1 A) List of functional groups present in the molecule Valclavam. B) Valclavam is annotated in the PubChem (CID 126919) and ChEBI (CHEBI:9920) databases. In PubChem, it is incorrectly assigned the class of beta-lactams, which are classified as sulfur compounds (according to the MeSH annotation).

Moreover, although some sulfur compounds are inorganic, and other are or organic, it is wrong to describe a single compound both as organic and inorganic. The transitivity of the *is_a* relationship is not fulfilled, which makes the class inference difficult. In ChEBI, the same compound is correctly classified as a peptide. However, as in PubChem, the annotation is incomplete. Class assignments to “clavams” and “azetidines”, among others, are missing.

These databases and ontologies/thesauri are excellent and have been used in various studies including chemical enrichment analysis (201), and knowledge-based metabolic model reconstruction (202), among others. However, they are all produced manually, thus making the classification/annotation process somewhat tedious, error-prone and inconsistent (Figure 2.1). In addition, they require substantial human expert time, which means these classification systems only cover a tiny fraction of known chemical space. For instance, in the PubChem database (9), only 0.12% of the >91,000,000 compounds (as of June 2016) are actually classified via the MeSH thesaurus. There are several other, older or lesser-known chemical classification schemes, ontologies or taxonomies that are worth mentioning. The Chemical Fragmentation Coding system (72) is perhaps the oldest taxonomy or chemical classification scheme. It was developed in 1963 by the Derwent World Patent Index (DWPI) to facilitate the manual classification of chemical compounds reported in patents. The system consists of 2,200 numerical codes corresponding to a set of pre-defined, chemically significant structure fragments. The system is still used by Derwent indexers who manually assign patented chemicals to these codes. However, the system is considered out-dated and complex. Likewise, using the chemical fragmentation codes requires practice and extensive guidance of an expert. A more automated alternate to the Derwent index was developed in the 1970’s, called the

HOSE (Hierarchical Organisation of Spherical Environments) code (203). This hierarchical substructure system allows one to automatically characterize atoms and complete rings in terms of their spherical environment. It employs an easily implemented algorithm that has been widely used in NMR chemical shift prediction. However, the HOSE system does not provide a named chemical category assignment nor does it provide an ontology or a defined chemical taxonomy. More recently, the Chemical Ontology (CO) system (204) has been described. Designed to be analogous to the Gene Ontology (GO) system, CO was one of the first open-source, automated functional group ontologies to be formalized. CO functional groups can be automatically assigned to a given structure by Checkmol (92), a freely available program. CO's assignment of functional groups is accurate and consistent, and it has been applied to several small datasets. However, the CO system is limited to just ~200 chemical groups, and so it only covers a very limited portion of chemical space. Moreover, Checkmol is very slow and is impractical to use on very large data sets. SODIAC (205) is another promising tool for automatic compound classification. It uses a comprehensive chemical ontology and an elegant structure-based reasoning logic. SODIAC is a well-designed commercial software package that permits very rapid and consistent classification of compounds. The underlying chemical ontology can be freely downloaded and the SODIAC software, which is closed-source, is free for academics. The fact that it is closed-source obviously limits the possibilities for community feedback or development. Moreover, the SODIAC ontology does not provide textual definitions for most of its terms and is limited in its coverage of inorganic and organo-metallic compounds. Other notable efforts directed towards chemical classification or clustering include Maximum Common Substructure

(MCS) based methods (206,207), an iterative scaffold decomposition method introduced by Shuffenhauer *et al.* (208), and a semantic-based method described by Chepelev *et al.* (209). However, most of these are proof-of-principle methods and have only been validated on a small number of compound classes, which cover only a tiny portion of rich chemical space. Moreover, they are very data-set dependent. As a result, the classifications do not match the nomenclature expectations of the chemical community, especially for complex compound classes.

Overall, it should be clear that while many attempts have been made to create chemical taxonomies or ontologies, many are proprietary or “closed source”, most require manual analysis or annotation, most are limited in scope and many do not provide meaningful names, definitions or descriptors. These shortcomings highlight the need to develop open access, open-source, fast, fully automated, comprehensive chemical classification tools with robust ontologies that generate results that match chemists’ (i.e. domain experts’) and community expectations. Furthermore, such tools must rapidly classify chemical entities in a consistent manner that is independent of the type of chemical entity being analyzed.

The development of a fully automated, comprehensive chemical classification tool also requires the use of a well-defined chemical hierarchy, whether it is a taxonomy or an ontology. This means that the criteria for hierarchy construction, the relationship types, and the scope of the hierarchy must be clearly defined. Additionally, a clear set of classification rules and a comprehensive data dictionary (or ontology) are necessary. Furthermore, comprehensive chemical classification requires that the chemical categories present in the taxonomy/ontology must be accurately described in a computer-interpretable format. Because new chemical compounds and new "chemistries" are being developed or discovered all the time, the taxonomy/ontology must be flexible and any extension should not force a fundamental modification of the classification procedure. In this regard, Hasting *et al.* (210) suggested a list of principles that would facilitate the development of an intelligent chemical structure-based classification system. One of the main criteria in this schema is the possibility to combine different elementary features into complex category definitions using compositionality. This is very important, since chemical classes are structurally diverse. Additionally, an accurate description of their core structures sometimes requires the ability to express constraints such as substitution patterns. Today, this can be achieved to a certain extent by the use of logical connectives and structure-handling technologies such as the SMiles ARbitrary Target Specification (SMARTS) format.

In this paper, we describe a comprehensive, flexible, computable, chemical taxonomy along with a fully annotated chemical ontology (ChemOnt) and a Chemical Classification Dictionary. These components underlie a web-accessible computer program called *ClassyFire*, which permits automated rule-based structural classification of essentially all known chemical entities. *ClassyFire* makes use of a number of modern computational techniques and circumvents most of the limitations of the previously mentioned systems and software tools. This paper also describes the rationale behind *ClassyFire*, its classification rules, the design of its taxonomy, its performance under testing conditions and its potential applications. *ClassyFire* has been successfully used to classify and annotate >6,000 molecules in DrugBank (99), >25,000 molecules in the LIPID MAPS Lipidomics Gateway (211), >42,000 molecules in HMDB (8), >43,000 compounds in ChEBI (200) and >60,000,000 molecules in PubChem (9), among others. These compounds cover a wide range of chemical types such as drugs, lipids, food compounds, toxins, phytochemicals and many other natural as well as synthetic molecules. *ClassyFire* is freely available at http://classyfire.wishartlab.com. Moreover, the *ClassyFire* API, which is written in Ruby, provides programmatic access to the *ClassyFire* server and database. It is available at https://bitbucket.org/wishartlab/classyfire_api.

2.2 Methods

Creating a computable chemical taxonomy requires three key components: 1) a well-defined hierarchical taxonomic structure; 2) a dictionary of chemical classes (with full definitions and category mappings); and 3) computable rules or algorithms for assigning chemicals to taxonomic categories. Each of these components is described in more detail below.

2.2.1 Component 1 – Hierarchical Taxonomic Structure

A taxonomy requires a well-defined, structured hierarchy. Following standard notation, we use the term "category" to refer to any chemical class (at any level), each of which corresponds to a set of chemicals. These categories are arranged in a tree structure. The main relationship type connecting these different categories is the "*is_a*" relationship. The rationale behind the choice of a tree structure was to provide a detailed annotation represented via a simple data structure, which could be easily understandable by humans.

Moreover, as described in the results section, *ClassyFire* provides a list of all parents of a compound, which makes it easy to infer all of its ancestors. Inspired by the original Linnaean biological taxonomy (116), we assigned the terms Kingdom, SuperClass, Class, and SubClass to denote the first, second, third and fourth levels of the chemical taxonomy, respectively. The top level (Kingdom) partitions chemicals into two disjoint categories: organic compounds versus inorganic compounds. Organic compounds are defined as chemical compounds whose structure contains one or more carbon atoms. Inorganic compounds are defined as compounds that are not organic, with the exception of a small number of "special" compounds, including, cyanide/isocyanide and their respective non-hydrocarbyl derivatives, carbon monoxide, carbon dioxide, carbon sulfide, and carbon disulfide. The classification of compounds into these two kingdoms aligns with most modern views of chemistry and is easily performed on the basis of a compound's molecular formula. The other levels in our classification schema depend on much more detailed definitions and rules that are described below. SuperClasses (which includes 26 organic and 5 inorganic categories) consist of generic categories of compounds with general structural identifiers (e.g. organic acids and derivatives, phenylpropanoids and polyketides, organometallic compounds, homogeneous metal compounds), each of which covers millions of known compounds. The next level below the SuperClass level is the Class level, which now includes 764 nodes. Classes typically consist of more specific chemical categories with more specific and recognizable structural features (pyrimidine nucleosides, flavanols, benzazepines, actinide salts). Chemical Classes usually contain >100,000 known compounds. The level below Classes represents SubClasses, which typically consist of >10,000 known compounds. There are

1,729 SubClasses in the current taxonomy. Additionally, there are 2,296 additional categories below the SubClass level covering taxonomic levels 5 to 11.

Altogether this extensive chemical taxonomy contains a total of 4,825 chemical categories of organic (4,146) and inorganic (678) compounds, in addition to the root category (Chemical entities). As a whole, this chemical taxonomy can be represented as a tree with a maximum depth of 11 levels, and an average depth of five levels per node (Figure 2.2). As with any structured taxonomy, the creation of a well-defined hierarchical structure offers the possibility to focus on a sub-domain of the chemical space, or a specific level of classification. A more complete description of this taxonomic hierarchy can be found in Table 2.1. The chemical taxonomy and its hierarchical structure provided using the Open Biological and Biomedical Ontologies (OBO) format (212), which may help with its integration with respect to semantic technology approaches. The resulting OBO file was generated with OBO-Edit (213), and can be downloaded from the *ClassyFire* website.

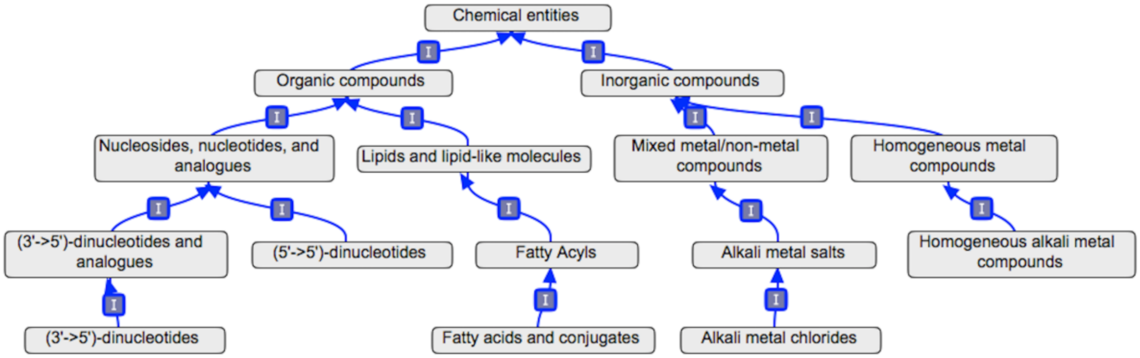

Chemical entities
Organic compounds
Inorganic compounds
Nucleosides, nucleotides, and analogues
Lipids and lipid-like molecules
Mixed metal/non-metal compounds
Homogeneous metal compounds
(3'->5')-dinucleotides and analogues
(5'->5')-dinucleotides
Fatty Acyls
Alkali metal salts
Homogeneous alkali metal compounds
(3'->5')-dinucleotides
Fatty acids and conjugates
Alkali metal chlorides
I

Figure 2.2 Illustration of the taxonomy as a tree.

2.2.2 Component 2 - Chemical Class Dictionary

Each node or category name in *ClassyFire*'s chemical ontology or ChemOnt, was created by extracting common or existing chemical classification category terms from the scientific literature and available chemical databases. We used existing terms to avoid "reinventing the wheel". By making use of commonly recognized or widely used terms that already exist in the chemical literature, we believed that the taxonomy (and the corresponding ontology) should be more readily adopted and understood. This dictionary creation process was iterative and required the manual review of a large number of specialized chemical databases, textbooks and chemical repositories. Because the same compounds can often be classified into multiple categories, an analysis of the specificity of each categorical term was performed. Those terms that were determined to be clearly generic (e.g. organic acid, organoheterocyclic compound) or described large numbers of known compounds were assigned to SuperClasses. Terms that were highly specific (e.g. alpha-imino acid or derivatives, yohimbine alkaloids) or which described smaller numbers of compounds that clearly fell within a larger SuperClass were assigned to Classes or SubClasses. This assignment also depended on their relationship to higher-level categories. In some cases multiple, equivalent terms were used to describe the same compounds or categories (imidazolines vs. dihydroimidazoles). To resolve these disputes, the frequency with which the competing terms were used was objectively measured (using Google page statistics or literature count statistics). Those having the highest frequency would generally take precedence. However, attention was also paid to the scientific community and expert panels. When available, the IUPAC term was used to name a specific category. Otherwise, if the experts clearly recommended a set of (less frequently used) terms, these would take precedence over terms initially chosen by our initial "popularity" selection criteria. Examples include the terms "Imidazolines" (229,000 Google hits) and "Dihydroimidazoles" (4,590 Google hits). The other popular terms were then added as synonyms. A total of 9,012 English synonyms were added to the ChemOnt terminology data set.

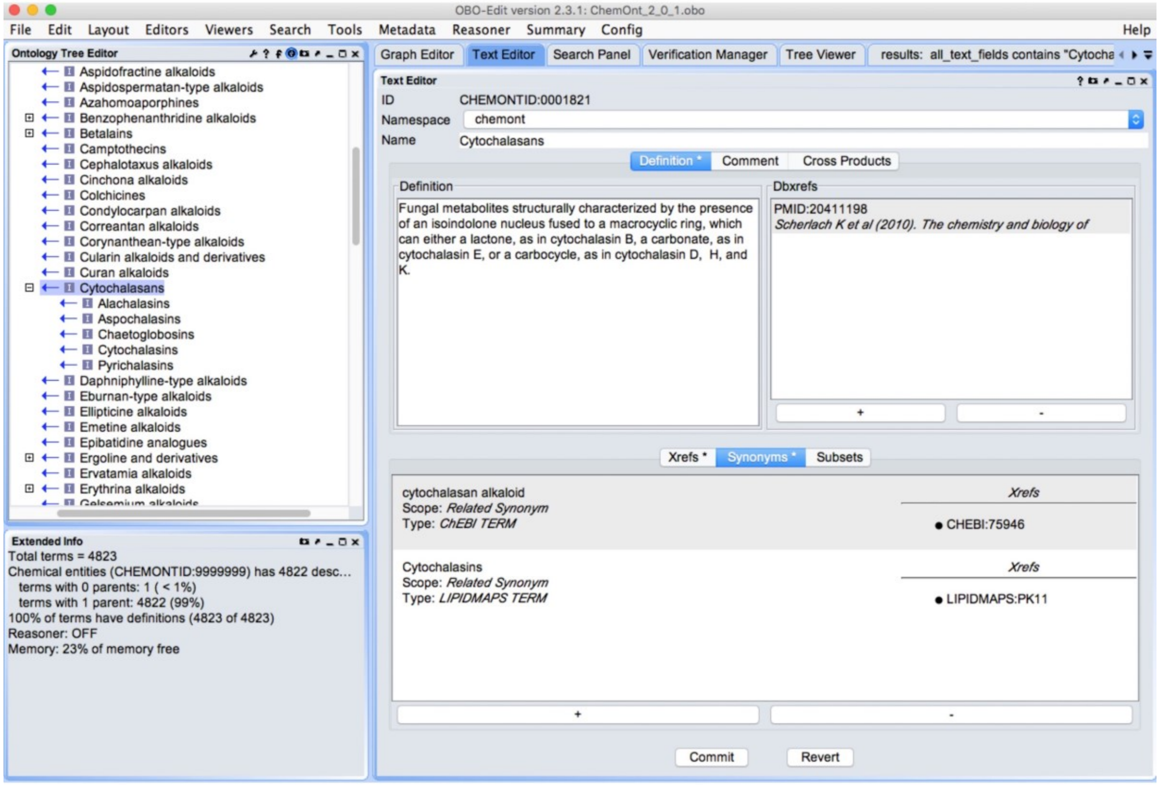
In a number of cases, new SuperClass and Class terms were created for chemical categories not explicitly defined in the literature. Of these, the resulting "novel" categories were typically constructed from the IUPAC nomenclature for organic and inorganic compounds. Because our chemical dictionary was built from extant or common terms, it contains many community-specific categories commonly used in the (bio-) chemical nomenclature (e.g. primary amines, steroids, nucleosides).

Table 2.1 Definitions of terms used in the ontological classification.

Term	**Description**
Kingdom	First level of hierarchical classification: Organic or Inorganic.
SuperClass	Second level of hierarchical classification. Metabolites with the same superclass share generic structural features that describe their overall composition or shape.
Class	Third level of hierarchical classification. Metabolites of the same class share a parent substructure. The structural similarity is generally higher at the class level compared to the superclass level.
SubClass	Fourth level of hierarchical classification. Metabolites of the same class share a parent substructure. The structural similarity is generally higher at the subclass level compared to the class level.
Intermediate nodes	Nodes that are descendants of the subclass and ascendants of the direct parent.
Direct Parent	The category corresponding to the largest skeleton or most dominant feature of the classified compound. The direct parent could correspond to the superclass, class, subclass or any other lower level. In the latter case, the intermediate parents can be traced back using the ontology file.
Alternative Parents	Other categories in the ontology that describe the classified compound and do not display a parent-child relationship to each other or to the direct parent.

Molecular Framework	Provides a general description of the compound in term of aliphaticity/aromaticity, number of cycles, and the variety of atom types (homo, hetero). This is calculated only for compounds/mixtures with less than two organic moieties.
Substituents	Functional groups and substructures contained in the compound. Only to avoid redundancy, the substituents mapped to each category of a given ontological classification are removed from the list of substituents.
Description	Textual structure-based description of the compound. It gives a brief description of the main characteristics of the largest skeleton or most dominant structural feature.
External Descriptors	Annotation of the compounds in other databases. It only shows the deepest nodes in the classification.

Moreover, due to the diverse nature of active and biologically interesting compounds, many chemical categories linked to specific chemical activities or based on biomimetic skeletons (e.g. alpha-sulfonopeptides, piperidinylpiperidines) were added. For instance, several compounds from the category of imidazo[1,2-a]pyrimidine (CHEMONTID:0004377) have been shown to display GABA(A) antagonist activity, and a potential to treat anxiety disorders (214).



ID CHEMONTID:0001821

Namespace chemont

Name Cytochalasans

Definition: Fungal metabolites structurally characterized by the presence of an isoindolone nucleus fused to a macrocyclic ring, which can either a lactone, as in cytochalasin B, a carbonate, as in cytochalasin E, or a carbocycle, as in cytochalasin D, H, and K.

Dbxrefs: PMID:20411198 *Scherlach K et al (2010). The chemistry and biology of*

cytochalasan alkaloid
Scope: *Related Synonym*
Type: *ChEBI TERM*
● CHEBI:75946

Cytochalasins
Scope: *Related Synonym*
Type: *LIPIDMAPS TERM*
● LIPIDMAPS:PK11

```
Total terms = 4823
Chemical entities (CHEMONTID:9999999) has 4822 desc...
  terms with 0 parents: 1 ( < 1%)
  terms with 1 parent: 4822 (99%)
100% of terms have definitions (4823 of 4823)
Reasoner: OFF
Memory: 23% of memory free
```

Figure 2.3 The chemical taxonomy. The taxonomy is illustrated with the OBO-Edit software, showing definitions synonyms, references, and extended information.

After all the dictionary terms were identified and compiled (4,825 terms to date), each term was formally defined using a precise, yet easily understood text description that included the structural features corresponding to that chemical category (Figure 2.3).

These formal definitions and the corresponding category mappings formed the basis of the structural classification algorithm and the classification rules described below. Once defined, the terms in this Chemical Classification Dictionary were progressively added to the taxonomic structure to form the structure-based hierarchy underlying *ClassyFire*'s chemical classification scheme. With the combination of the taxonomic structure and the Chemical Classification Dictionary, ChemOnt can be formally viewed as an ontology (albeit purely a structural ontology).

2.2.3 Component 3 – The Classification Algorithm

The essence of our classification algorithm is to use the structural definitions and terms contained in the Chemical Classification Dictionary to classify compounds. This required converting the English text definitions into a computable set of rules with each definition consisting of one or more chemical structures, and/or a set of characteristic features that can be otherwise expressed in a computable form. The main format used for chemical structure representation in our classification algorithm is the SMARTS format (68). SMARTS is a molecular pattern matching language, related to the popular SMILES molecular language, that can be used to specify sub-structural patterns in molecules. For instance, thiazoles are heterocyclic compounds containing a five-member aromatic ring made up of one sulfur atom, one nitrogen, and three carbon atoms. This category of compounds can be described with the following SMARTS expression:

[$([#16]-1-[#6]=[#6]-[#6]=[#7]-1),$([#16]-1-[#6]=[#6]-[#7]=[#6]-1)]

Converting the 4,825 definitions in our Chemical Classification Dictionary led to the creation of >9,000 SMARTS strings. The validity of each SMARTS string was first tested by performing a superstructure search on small sets of positive or negative example compounds. In most cases, manually generated SMARTS strings, or combinations thereof, were sufficient to represent the vast majority of chemical categories. However, in some cases, SMARTS strings could not express specific constraints that a given compound must fulfill in order to be assigned a given category. For instance, SMARTS strings cannot describe structures with variable numbers of a specific bond or a specific atom. One way around this would be to enumerate the different patterns, which could easily lead to a combinatorial explosion. For these exceptions we used the Markush format (75), which is available through ChemAxon's Marvin tool. With the Markush format, it is possible to represent substituent's variations, position's variations, as well as the frequency variation of structural groups within a chemical structure. The Markush patterns used by *ClassyFire* constitute only about 4% of the set of patterns in the *ClassyFire* database. In addition, some chemical categories were more appropriately defined by a combination of logical expressions based on features such as structural patterns, physico-chemical properties or chemical formulae. For example, an alkane, which is an acyclic branched or unbranched hydrocarbon having the general formula C_nH_{2n+2}, can be formally represented as the following combination of rules:

$$RingCount(A) = 0 \wedge AtomCount(C,A) > 0 \wedge (AtomCount(C,A) + AtomCount(H,A) = TotalAtomCount(A)) \wedge (AtomCount(H,A) = 2 \times AtomCount(C,A) + 2),$$

where *AtomCount(X,A)* is the number of atoms of type X in the molecule A, *RingCount(A)* is the total number of rings in the compound A, and *TotalAtomCount(A)* is the total number of atoms in the compound A. In rare cases, some categories of compounds could not be accurately described in an explicit and formal way using any SMARTS string, Markush representation, structural pattern, physico-chemical property or chemical formula. These included certain categories of lipids and lipid-like molecules, phenylpropanoids, polyketides, peptidomimetics and alkaloids, among others. In these cases, the categories were defined as a union of their subcategories that were formally expressed.

It is also important to remember that chemicals can exist as structural chimeras or combinations of different, covalently linked chemical structures, building blocks or domains. Consequently some chemicals (Figure 2.1) could potentially belong to more than one chemical class or category. To simplify the chemical classification process, we chose to prioritize the category corresponding to the largest or most dominant structural feature of the chemical compounds (see below). This decision was based on the observed and historical tendencies of chemists to manually classify compounds based on the size (i.e. the number of atoms) of the most dominant structural feature. Furthermore, identifying the largest feature is a technique that is easily measurable and completely objective. If two or more dominant structural features are equal in size, methods described later are used to select one of the features. In *ClassyFire*'s algorithm, if a structural feature is a represented by structure, its feature weight is equivalent to the number of non-hydrogen atoms in that substructure. If a structural feature is represented by a combination of logical terms, its weight is the total number of non-hydrogen atoms of the smallest compound that fulfils the defined constraints.

It is important that any automated classification tool provide a result that is identical or near-identical to the outcome of manual assignments by experts. As a result, a small number of *post hoc* adjustments were made for certain well-known chemical categories that are commonly identified by their biochemical context. For instance, we created a category called “Phenylpropanoids and polyketides”. Phenylpropanoids and polyketides can be described as small organic compounds that are synthesized either from the amino acid phenylalanine (phenylpropanoids) or the decarboxylative condensation of malonyl-CoA (polyketides). These classes are best described as a union of their children. The “Phenylpropanoids and polyketides” category currently has 34 direct children and a total of 273 descendant categories, including Flavonoids, among others. Describing a flavonoid compound as a phenylpropanoid instead of a chromone (a term that can legitimately be used to describe flavonoids) is, from a biochemist’s point of view, more precise and accurate.

2.2.4 Mapping of Other Classification Schema and Vocabularies to *ClassyFire*’s Taxonomy

As noted before, there are a number of well-known, online chemical databases that have developed their own, manually annotated chemical taxonomy and/or ontology. For instance, the ChEBI ontology (200) provides a sub-ontology for chemical roles, in addition to the structure-based sub-ontology. LIPID MAPS (198) focuses on lipids and lipid-like molecules, and groups them according to their biosynthetic origin. MeSH is a thesaurus consisting of >50,000 terms, about 1/3 of which cover chemical entities or classes thereof. In developing the ChemOnt taxonomy, which is used by *ClassyFire*, we aimed at creating a consensus chemical taxonomy partly inspired by these approaches. In that regard, ChemOnt was mapped to three other widely used chemical hierarchies or taxonomies (ChEBI, LIPID MAPS and MeSH). This was done by assigning one or more synonyms to each ChemOnt category, and specifying the corresponding level or scope of term similarity. For any ChemOnt term, a synonym can have the identical meaning (exact scope), a more specific meaning (narrow scope), or a less specific meaning (broad scope). In some cases, the synonym can have slightly different meaning, so that it cannot be assigned any of the three aforementioned scope categories. In this case, it is simply called a related synonym.

In a joint effort with the ChEBI development team, an ontology look-up table was created to map *ClassyFire*'s (and ChemOnt's) taxonomy to the ChEBI sub-ontology of chemical entities. When applicable, an exact CHEBI synonym was assigned to the ChemOnt term. Otherwise, either one or more broad synonyms, preferably those mapped to its parent, were assigned. In some cases, narrow CHEBI synonyms were also assigned. It is worth mentioning that in the case of ChEBI, due to certain philosophical discrepancies, some terms may appear to be exact synonyms for a given ChemOnt category, but actually have a different meaning. For instance, ChEBI makes a clear distinction between "carboxylic acid" and "carboxylic acid anion", while ChemOnt does not. Therefore, the ChEBI term "carboxylic acid" is a narrow synonym of ChemOnt's "carboxylic acids". A total of 6,014 category mappings were created, with an average of 1.24 ChEBI synonyms per category. Each *ClassyFire* category has one or more mapped ChEBI terms. This effort highlighted a number of similarities, differences, and suggested some improvements (e.g.: categories to be added) for both systems. Using this training information, *ClassyFire* has been modified and used to annotate >43,000 small molecules from the ChEBI database. This ChEBI classification can be downloaded from the *ClassyFire* website. To date, these results have been used by the ChEBI development team to annotate more than 10,000 compounds present in the ChEBI database. In lipid biology, the LIPID MAPS consortium provides the standard chemical ontology for lipids (198). As a result we designed the lipid subset in ChemOnt to align closely with the LIPID MAPS classification scheme. A total of 789 *ClassyFire* categories were mapped to one of 307 LIPID MAPS terms each. As a result, a combination of *ClassyFire* and LIPID MAPS ontologies was used to classify ~35,000 small metabolites, which can be accessed from the LIPID MAPS Lipidomics Gateway (211), a resource sponsored by the National Institute of General Medical Sciences (215) and the Common Fund of the National Institutes of Health (216). As a result of this mapping, several more category assignments were added to complement the LIPID MAPS classifications. *ClassyFire* has also been manually mapped, although only partially, to the MeSH thesaurus, which is used in the PubChem database. So far, 844 *ClassyFire* categories have been mapped to at least one corresponding MeSH term, accounting for a total of 945 mappings to the MeSH thesaurus. This MeSH mapping will likely continue for another year or two.

A considerable proportion of the structures available in databases, such as PubChem, correspond to chemical mixtures. For instance, some drugs or pesticides are synthesized as mixtures of several organic compounds. *ClassyFire* has been programmed to classify such mixtures. The underlying algorithm allows it to assign classes while considering the organic moieties separately, and also as a whole. For instance, a mixture of an organic compound and a chlorine anion (inorganic) will be assigned the category of organic chlorine salts, among others, but not the category of inorganic compounds.

2.2.5 The Classification Process

As illustrated in Fig. 2.4, the *ClassyFire* classification process involves four steps: 1) Creation and Preprocessing of the Chemical Entity; 2) Feature Extraction; 3) Rule-based Category Assignment and Category Reduction; and 4) Selection of the Direct Parent.

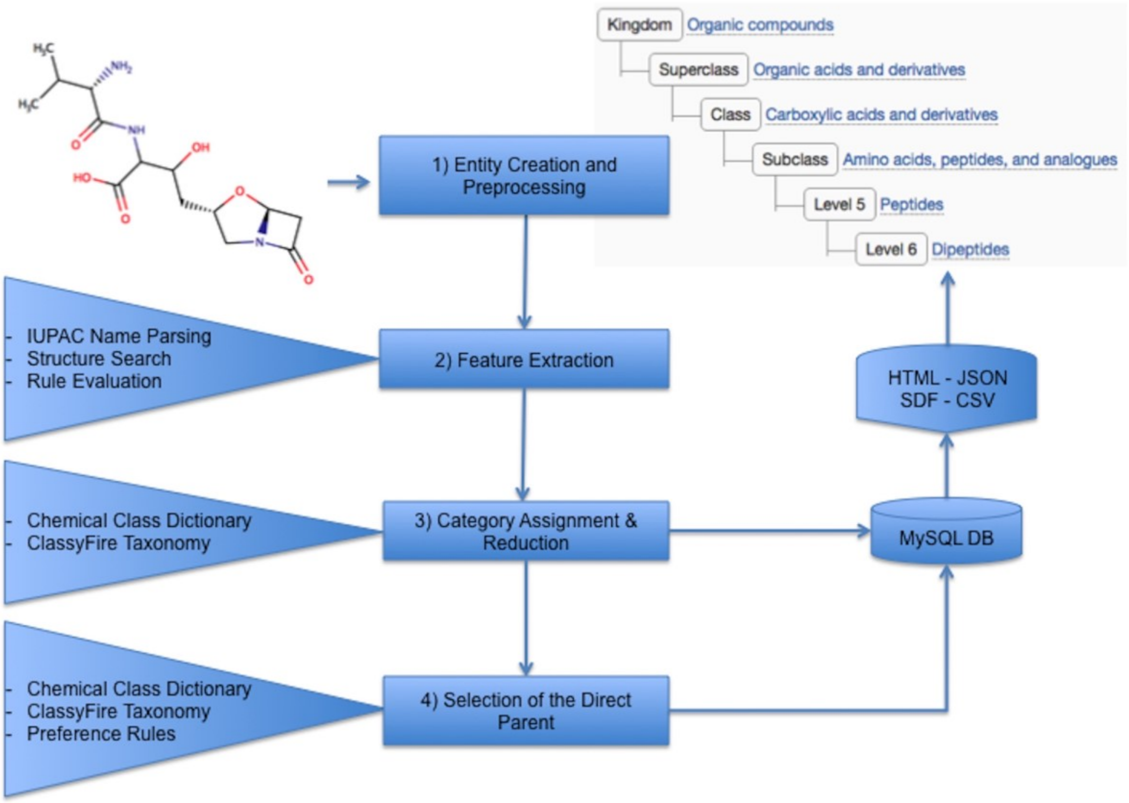

Kingdom
Organic compounds
Superclass
Organic acids and derivatives
Class
Carboxylic acids and derivatives
Subclass
Amino acids, peptides, and analogues
Level 5
Peptides
Level 6
Dipeptides
1) Entity Creation and Preprocessing
- IUPAC Name Parsing
- Structure Search
- Rule Evaluation
2) Feature Extraction
HTML - JSON
SDF - CSV
- Chemical Class Dictionary
- ClassyFire Taxonomy
3) Category Assignment & Reduction
MySQL DB
- Chemical Class Dictionary
- ClassyFire Taxonomy
- Preference Rules
4) Selection of the Direct Parent

Figure 2.4 Workflow of the chemical classification.

These are described in more detail below:

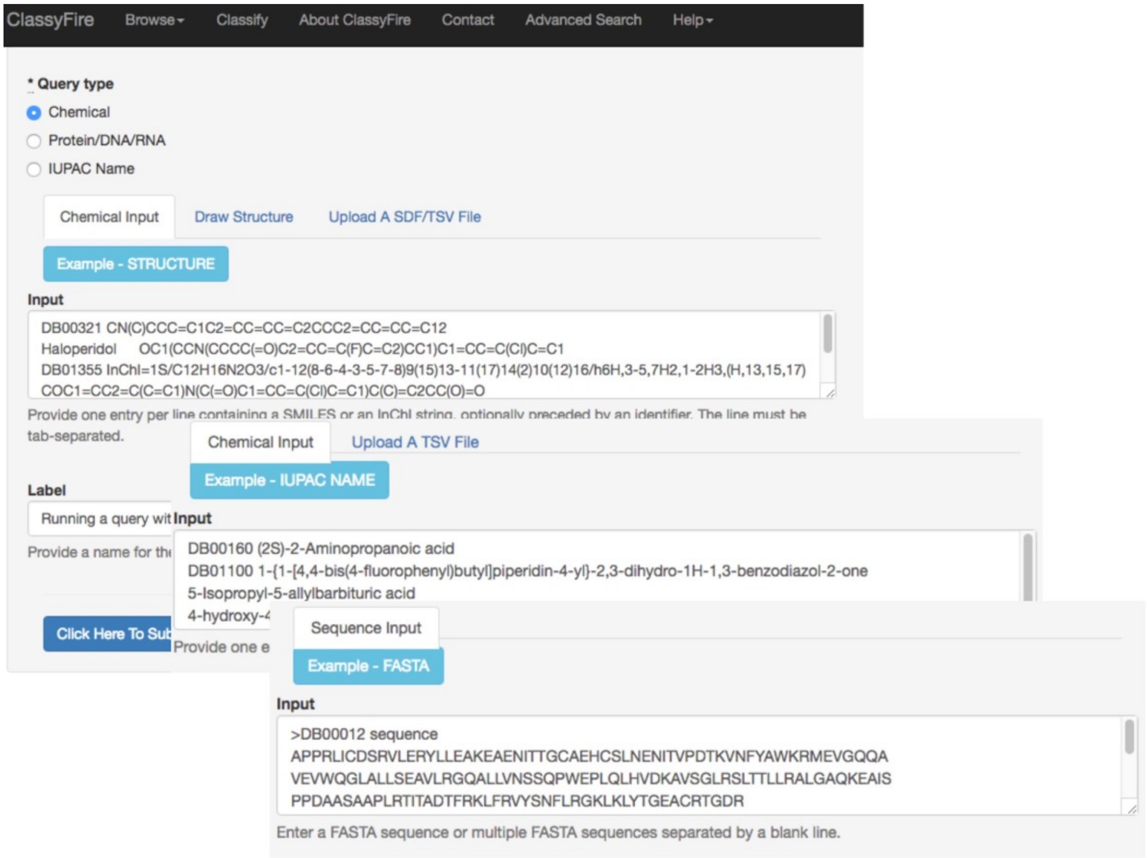
2.2.5.1 Step 1 – Creation and Preprocessing of the Chemical Entity

This step involves the creation of one or more chemical entity objects (which are stored in a database), and the calculation of physico-chemical as well as structural properties. Most of these features, such as the number of (aromatic, aliphatic) rings, are used for classification. Others, such as the mass, are used for text-based search (See Use Cases, below). The calculation of physico-chemical properties is performed using ChemAxon's JChem API (version 15.5.25.0). *ClassyFire* accepts different types of chemical input: SMILES, SDF, InChI, IUPAC name, and FASTA sequence files. The different types of chemical input are illustrated in Figure 2.5. SMILES, SDF, and InChI strings are common structural representation formats for chemical entities, which can be directly used for structure search operations or the generation of physico-chemical properties. In contrast, each IUPAC name is converted to the corresponding structure using the OPSIN library (85), before any chemical object is created and subsequently pre-processed. If the chemical (protein, DNA or RNA molecule) input is submitted in FASTA format, every sequence is either identified as a nucleotide or peptide sequence type. This step is important, as the interpretation of one-letter sequences will vary depending on the sequence type. The *ClassyFire* web server also allows users to submit their query through the MarvinSketch Chemical Drawing Applet, which permits users to import or draw a chemical structure, which is then exported as a SMILES string.

2.2.5.2 Step 2 – Feature Extraction

The second step in the *ClassyFire* program involves the generation of structural features based on a combination of superstructure-search operations and various property calculations. *ClassyFire* combines several methods for structural pattern detection. Most features are detected through superstructure search, which is performed on its library of over 9,000 manually designed SMARTS patterns and Markush structures. Each of the terms was validated through iterations of test and improvements (if necessary) over small sets of compounds. The library is integrated into ChemAxon's JChem Base. ChemAxon's Marvin 5.11.5 package was used to generate these patterns, ranging from small functional groups (e.g. the carbamoyl group) to complex skeletons (e.g. the (3'->5')-cyclic dinucleotide bis(phosphoromonothioate) pattern). Prior to being imported into the database, each structure pattern was subjected to a set of standardization operations, including normalization and aromatization. Each query compound is subjected to the same operations before the superstructure search. This allows the program to deal with differences in charges, valences and aromatic configuration.

Another feature detection method used in *ClassyFire* involves combining features with the use of logical connectives, and cardinality restrictions. Every structural feature defined by a logical expression is evaluated in order to assign that feature to the query compound. As an example, *ClassyFire* can detect specific features for an inorganic compound based on its elemental content, and the list of oxyanions it contains (if any). These features are described by rules embedded in a *ClassyFire* module that specifically handles inorganic compounds. In some cases, the use of structure patterns, chemical formulae or physicochemical properties is not sufficient to generate a feature. For instance, the category known as leukotrienes describes derivatives of arachidonic acid, containing three hydroxyl groups as well as four double bonds, exactly three of which are conjugated. The position of the three conjugated bonds as well as the relative position of the non-conjugated bond can vary, yielding a large number of combinations. Therefore, a superstructure search might not return a hit. In order to classify leukotrienes,



*** Query type**

● Chemical
○ Protein/DNA/RNA
○ IUPAC Name

Input

```
DB00321 CN(C)CCC=C1C2=CC=CC=C2CCC2=CC=CC=C12
Haloperidol     OC1(CCN(CCCC(=O)C2=CC=C(F)C=C2)CC1)C1=CC=C(Cl)C=C1
DB01355 InChI=1S/C12H16N2O3/c1-12(8-6-4-3-5-7-8)9(15)13-11(17)14(2)10(12)16/h6H,3-5,7H2,1-2H3,(H,13,15,17)
COC1=CC2=C(C=C1)N(C(=O)C1=CC=C(Cl)C=C1)C(C)=C2CC(O)=O
```

Provide one entry per line containing a SMILES or an InChI string, optionally preceded by an identifier. The line must be tab-separated.

Label

Running a query wit

Provide a name for th

Input

```
DB00160 (2S)-2-Aminopropanoic acid
DB01100 1-{1-[4,4-bis(4-fluorophenyl)butyl]piperidin-4-yl}-2,3-dihydro-1H-1,3-benzodiazol-2-one
5-Isopropyl-5-allylbarbituric acid
4-hydroxy-4
```

Provide one e

Input

```
>DB00012 sequence
APPRLICDSRVLERYLLEAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQA
VEVWQGLALLSEAVLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAIS
PPDAASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

Enter a FASTA sequence or multiple FASTA sequences separated by a blank line.

Figure 2.5 Different types of input accepted by *ClassyFire*.

ClassyFire makes use of standard IUPAC nomenclature in addition to a structure search to check whether these constraints are fulfilled. The IUPAC name of any query chemical entity is generated by ChemAxon's Structure-to-Name Conversion engine provided by the JChem API. IUPAC names can give valuable information about the parent of a given compound, as well as the positioning, number, and name of substituents relative to that parent. We developed a module, which uses a set of ~200 regular expressions and rules in order to accurately detect structural features given a query compound by parsing IUPAC names.

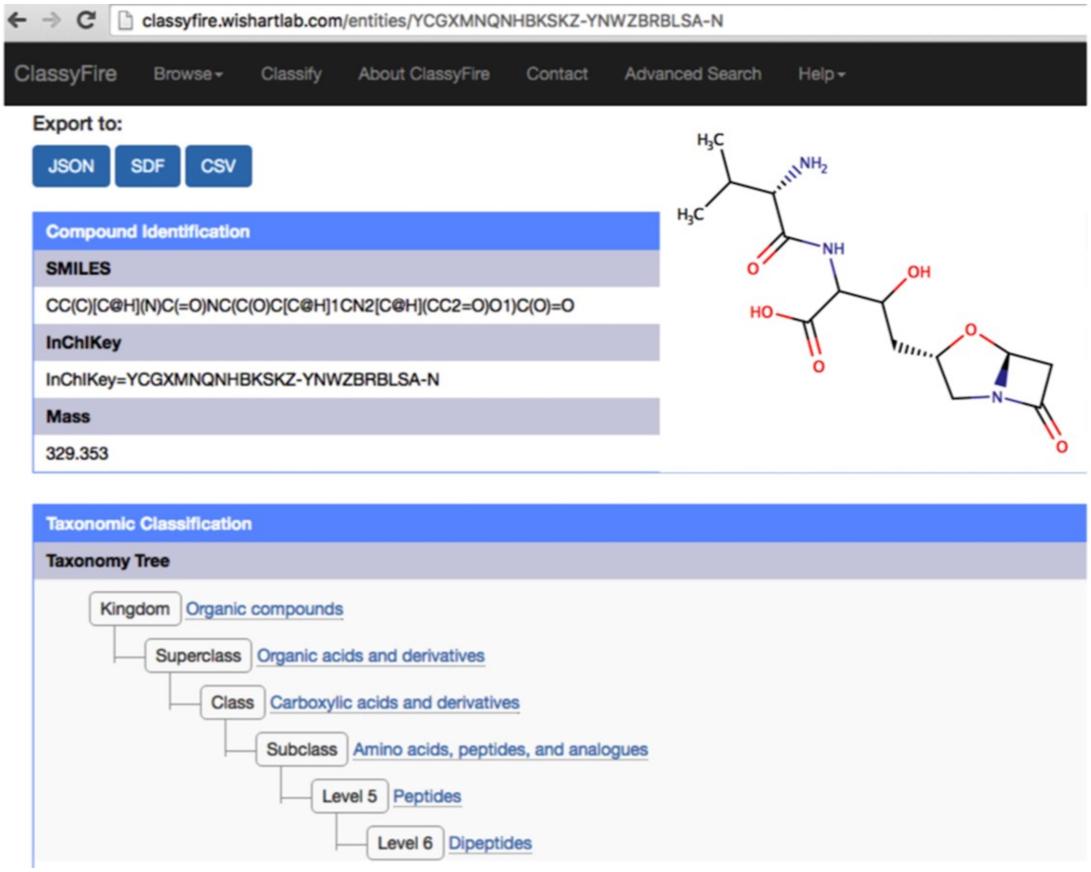
2.2.5.3 Rule-based Category Assignment and Category Reduction

After a list of structural features has been generated, each feature is then mapped to its corresponding category or node in the taxonomy. A manually compiled dictionary, which provides the weight and category for each feature, was used for the rule-based category assignment. After the category assignment is complete, a non-redundant list of chemical categories is constructed. This is done by iteratively reducing the set of chemical categories. For every pair of chemical categories, if there is a parent-child relationship (e.g. dioxanes [parent] and 1,2-dioxanes [child]), only the child node is retained (1,2-dioxanes).

2.2.5.4 Selection of the Direct Parent

The direct parent is the category defined by the largest structural feature that describes the compound. It is selected from the non-redundant list of categories obtained in the previous step. If two or more structural features have the largest weight, the direct parent is selected following a procedure that takes into account the number of cycles, heterocycles, ring atoms, ring heteroatoms, halogen atoms, fused rings, and the total number of heteroatoms, which are encoded in each node’s structural key. In some cases, the largest feature might be less descriptive or less relevant than another feature. For example, the glycoside moiety of a flavonoid glycoside can be much larger than the flavonoid moiety. However, the term “flavonoid glycoside” is more informative than the term “glycoside”, as it describes the presence of both a saccharide unit and a flavonoid, glycosidically linked to one another. In this case, an exception is made and the term “flavonoid glycoside” is selected over “glycoside”. A small (but not exhaustive) set of such exceptions has been manually compiled.

The entire *ClassyFire* program has been converted to a web-based resource. It is a RESTful web application located at http://classyfire.wishartlab.com. It allows users to submit one or more query molecules in SMILES, SDF, or InChI format, IUPAC name, or 1-letter amino acid and nucleic acid (FASTA) notation. The query structure(s) can be entered as text, uploaded, or drawn using the MarvinSketch applet. It is recommended that all query structures be represented in their chiral or isomeric form, to ensure a more precise classification. This is because different *ClassyFire* categories can be represented by stereoisomers of the same skeleton. Some examples include 3-alpha-hydroxysteroids (CHEMONTID:0003232) and 3-beta-hydroxysteroids (CHEMONTID:0003233), which are all sub-categories of 3-hydroxysteroids (CHEMONTID:0003027). When represented with an isomeric structure string for instance, a compound, such as androsterone, can be classified as a 3-alpha-hydroxysteroid. However, if it is represented with a canonical structure, it would only be classified as a 3-hydroxysteroid, which is less precise. Upon submission, the queries are processed by the *ClassyFire* classification tool, then entities or sequences are classified, and the results are then further processed, formatted and shown on a HTML output page (Figure 2.6). Classification results can also be downloaded in a JSON (217), SDF (218), or CSV (219) format. In addition to providing standard chemical classification data, *ClassyFire* also returns a list of chemical substituents, which are structural features (functional groups, substructures or motifs) contained within the molecule. For many compounds *ClassyFire* also provides a secondary attribute called the “Molecular Framework”. The Molecular Framework gives an overall description of the compound in terms of aliphaticity/aromaticity and number of cycles.



Compound Identification

SMILES

CC(C)[C@H](N)C(=O)NC(C(O)C[C@H]1CN2[C@H](CC2=O)O1)C(O)=O

InChIKey

InChIKey=YCGXMNQNHBKSKZ-YNWZBRBLSA-N

Mass

329.353

Taxonomic Classification

Taxonomy Tree

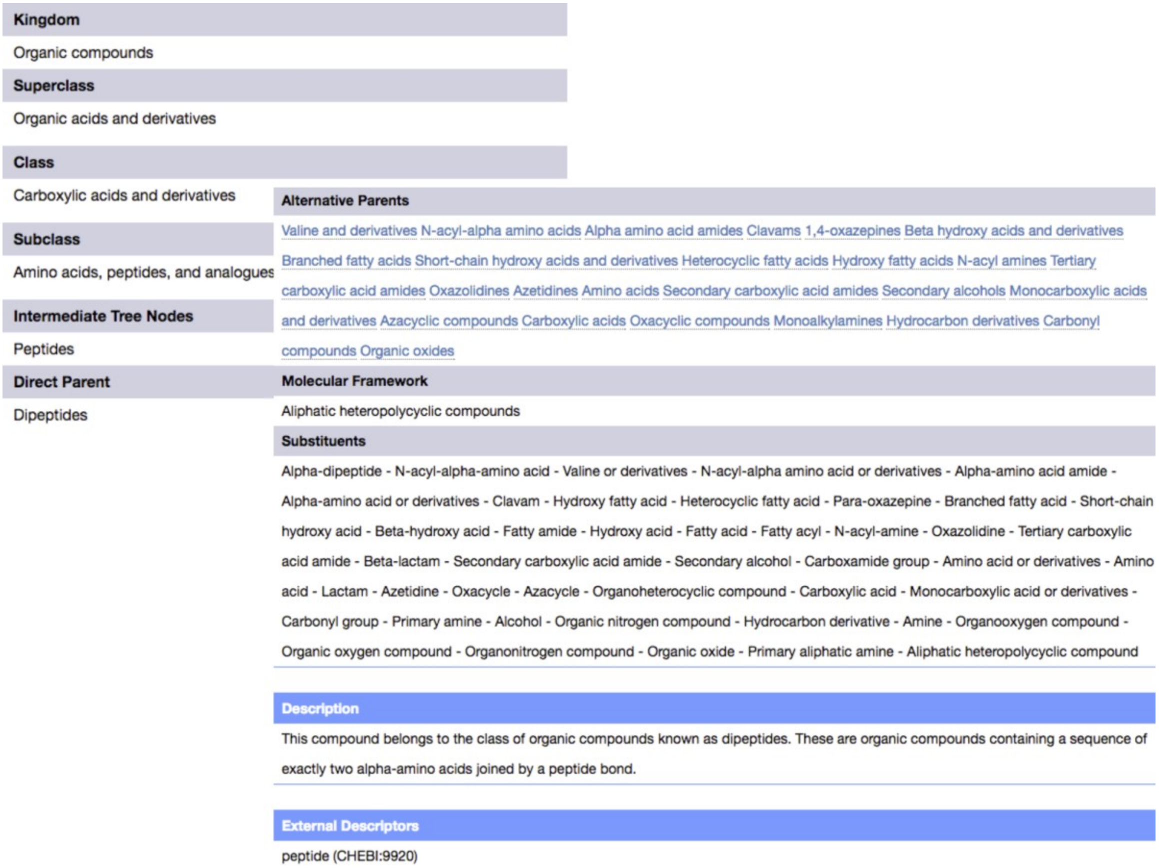
- Kingdom: Organic compounds
  - Superclass: Organic acids and derivatives
    - Class: Carboxylic acids and derivatives
      - Subclass: Amino acids, peptides, and analogues
        - Level 5: Peptides
          - Level 6: Dipeptides

Figure 2.6 Classification results for the molecule Valclavam (CID126919) on the *ClassyFire* website. The structural representations, and the taxonomic tree are illustrated. The classification result can be downloaded in different formats.

For instance, benzene is described as an aromatic homomonocyclic compound while butanol is described as an aliphatic acyclic compound. The Molecular Framework attribute does not apply to mixtures of organic compounds. In addition to providing an automated chemical classification service, the *ClassyFire* web server also provides a number of powerful text-based search options, which are described later.

2.3 Training and Evaluation

Training and evaluation of the *ClassyFire* program was performed throughout the development of the program, using data sets from several well known databases, containing thousands of drugs (99), lipids (8,198), food compounds (43), toxins, environmental pollutants, as well as other organic and inorganic compounds. Progressively larger and more diverse sets of manually classified chemicals (from 100+ compounds to more than 6,000 compounds) were manually compared and evaluated against the computed *ClassyFire* classifications to ensure that the program properly classified new compounds or compounds not previously seen in its training cycles. The manual classifications were generated according to the definitions found in the Chemical Classification Dictionary. Moreover, classifications of the various compounds were collected from the literature and other resources that provided the same category descriptions as *ClassyFire*. As errors or programming bugs were identified, class definitions were iteratively refined. If missing categories were found, or if compounds were more suitably classified in new categories, these were added to the Chemical Classification Dictionary (and to the *ClassyFire* algorithm). The identification of new categories was aided by the classification schema provided by other databases such as LIPID MAPS (198), ChEBI (200), and DrugBank (99). This iterative refinement process was conducted until essentially no incorrect assignment could be detected in even the largest test sets.



Kingdom

Organic compounds

Superclass

Organic acids and derivatives

Class

Carboxylic acids and derivatives

Subclass

Amino acids, peptides, and analogues

Intermediate Tree Nodes

Peptides

Direct Parent

Dipeptides

Alternative Parents

Valine and derivatives N-acyl-alpha amino acids Alpha amino acid amides Clavams 1,4-oxazepines Beta hydroxy acids and derivatives Branched fatty acids Short-chain hydroxy acids and derivatives Heterocyclic fatty acids Hydroxy fatty acids N-acyl amines Tertiary carboxylic acid amides Oxazolidines Azetidines Amino acids Secondary carboxylic acid amides Secondary alcohols Monocarboxylic acids and derivatives Azacyclic compounds Carboxylic acids Oxacyclic compounds Monoalkylamines Hydrocarbon derivatives Carbonyl compounds Organic oxides

Molecular Framework

Aliphatic heteropolycyclic compounds

Substituents

Alpha-dipeptide - N-acyl-alpha-amino acid - Valine or derivatives - N-acyl-alpha amino acid or derivatives - Alpha-amino acid amide - Alpha-amino acid or derivatives - Clavam - Hydroxy fatty acid - Heterocyclic fatty acid - Para-oxazepine - Branched fatty acid - Short-chain hydroxy acid - Beta-hydroxy acid - Fatty amide - Hydroxy acid - Fatty acid - Fatty acyl - N-acyl-amine - Oxazolidine - Tertiary carboxylic acid amide - Beta-lactam - Secondary carboxylic acid amide - Secondary alcohol - Carboxamide group - Amino acid or derivatives - Amino acid - Lactam - Azetidine - Oxacycle - Azacycle - Organoheterocyclic compound - Carboxylic acid - Monocarboxylic acid or derivatives - Carbonyl group - Primary amine - Alcohol - Organic nitrogen compound - Hydrocarbon derivative - Amine - Organooxygen compound - Organic oxygen compound - Organonitrogen compound - Organic oxide - Primary aliphatic amine - Aliphatic heteropolycyclic compound

Description

This compound belongs to the class of organic compounds known as dipeptides. These are organic compounds containing a sequence of exactly two alpha-amino acids joined by a peptide bond.

External Descriptors

peptide (CHEBI:9920)

Figure 2.7 Classification results for the molecule Valclavam (CID126919) on the *ClassyFire* website. A detailed listing of the structural features of the molecule is provided, along with a structure-based text description.

In addition to these manual consistency checks conducted throughout the training and development phase of the project, we also conducted an independent performance assessment of the final release version (version 2.0) of *ClassyFire*. A test set was built by randomly selecting 800 unique structures from DrugBank, the LIPID MAPS Lipidomics Gateway, HMDB (8), and T3DB (100). The compounds are all included in the PubChem database. We used a panel of experts to evaluate the correctness of each category assignment based on the definition in the Chemical Classification Dictionary. When applicable, we also verified if the direct parent was included in the list of classed assigned by ChEBI or LIPID MAPS.

2.4 Results and Discussion

The classification process as described in the previous section was implemented into both a computer program and a freely accessible web server called *ClassyFire*, available at http://classyfire.wishartlab.com. Moreover, an open source Ruby API (https://bitbucket.org/wishartlab/classyfire_api) allows users to programmatically access the web server in order to submit queries, and retrieve classification results, as well as entity-related properties. The complete taxonomy can be downloaded from *ClassyFire*'s home page.

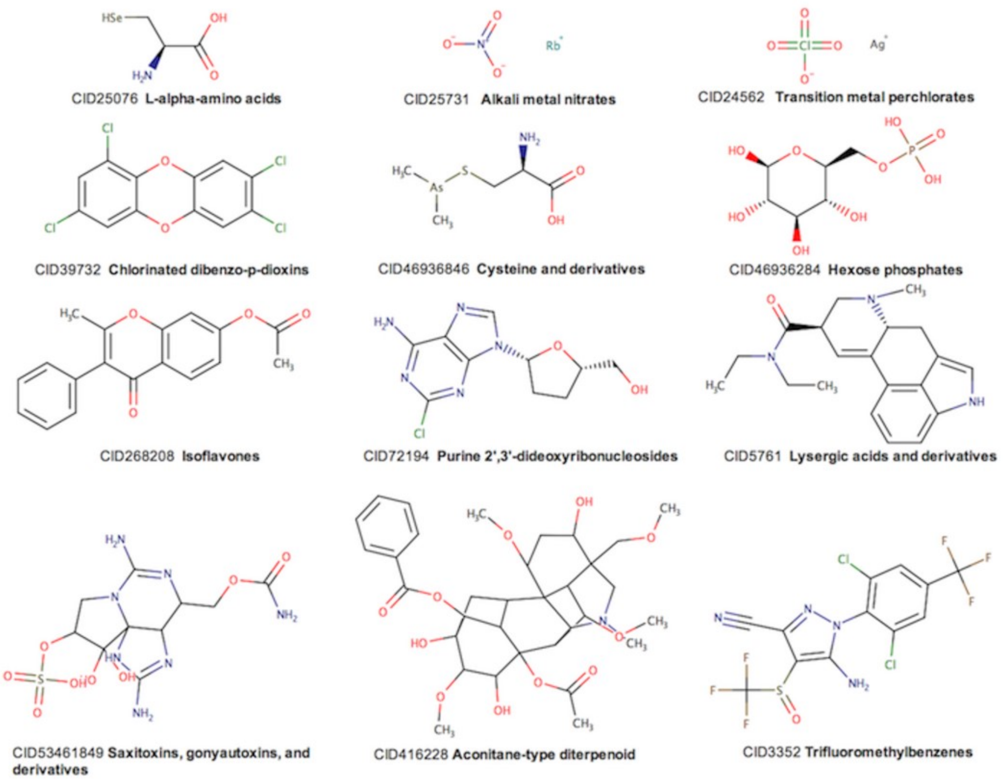
An example of *ClassyFire*'s classification and ontological annotation is illustrated for the antibiotic compound Valclavam (Figure 2.7). As can be seen in this figure, *ClassyFire* returns a taxonomic classification based on the most descriptive node in the taxonomy (Fig. 2.7A). The direct parent "dipeptides" represents the most dominant moiety of Valclavam's structure. However, the notion of what is most descriptive can vary from one user to another, and from one context to another. For example, a cyclic depsipeptide could be also be classified as a lactam. Because of this ambiguity, *ClassyFire* also displays a list of Alternative Parents (Fig. 2.7B) providing a more detailed description of the chemical. Alternative parents are categories that describe the compound but do not have an ancestor-descendant relationship with each other or with the Direct Parent. When available, *ClassyFire* returns Intermediate Nodes. These are nodes are descendants of a subclass (any category with a depth of 4), but have a depth lower than the direct parent.

In addition, *ClassyFire* provides the Molecular Framework and a list of all identified substituents (or structural features). Furthermore, an English, text-based compound description is also provided for non-experts. The text-based description is derived from *ClassyFire*'s Chemical Classification Dictionary. In an effort to facilitate the integration of data from different sources, *ClassyFire* also contains a database of cross-references from other popular chemical databases that use different taxonomies/ontologies, such as KEGG (103), ChEBI (200), LIPID MAPS (198), and MetaCyc (220). These cross-references and alternate-database classifications are routinely provided as *ClassyFire* output, when available.

To accelerate *ClassyFire*'s processing time, all of the chemical structures it has ever processed and all of the corresponding taxonomic/ontological outputs it has ever produced are stored in a local MySQL database. This allows the *ClassyFire* web sever to perform a simple lookup for those query compounds that have previously been processed (more than 70 million compounds to date). Therefore, for previously analyzed compounds the *ClassyFire* web server takes <50 milliseconds to return an answer. For completely novel compounds, the *ClassyFire* web server takes an average of 540 milliseconds to classify a structure.

2.5 Evaluation of *ClassyFire*'s Classification Results

After the iterative development, testing and manual evaluation of *ClassyFire* over several data sets consisting of >30,000 compounds from very diverse chemical categories, *ClassyFire* was formally tested on a set of 800 compounds not used during *ClassyFire*'s training phase. The compounds among which, drugs, food compounds, synthetic compounds, and biologically relevant metabolites, were selected from PubChem. The classification process took 249.9 seconds on a computer with 4 CPU CentOS nodes, with 3.6 GB of RAM, running with a maximum of 16 threads. The results were then manually reviewed by a panel of seven chemistry experts from three different countries. A total of 21,102 category assignments were made, for an average of 26.38 assignments per compound. On this specific test set, *ClassyFire* assigned a total of 1,308 distinct Categories. Figure 2.8 illustrates some examples of the category assignments. The goal was to evaluate how exact the computational rules were able to reflect the text-based descriptions, which themselves are traditionally used to classify compounds. Based on these textual descriptions, as well as the assignments from the literature and scientific databases, each compound's annotation was reviewed to identify possibly missing or wrong assignments. In this test, a total of 17 false positives (out of 21,102 assignments) were detected. An example is the misclassification of bixin dimethyl ester (CID14413719) as an acyclic diterpene. From a structural point of view, this compound contains a chain of four consecutive isoprene units, which is characteristic of diterpenes (Fig. 2.9A). However, bixin dimethyl ester is classified in both the LIPID MAPS and the ChEBI database as a C40 isoprenoid (tetraterpene). More precisely, bixin dimethyl ester belongs to the category of compounds known as apo-carotenoids, which arise from the oxidative cleavage of carotenoids. Thus, bixin dimethyl ester, which is a product of lycopene metabolism, is classified as a tetraterpene according to its biosynthetic origin. Based on its structure, one could argue that bixin dimethyl ester should be classified as a diterpene; but based on its biology, it should be classified as a tetraterpene derivative or as an apo-carotenoid diterpenoid (CHEBI:53186).



CID25076 L-alpha-amino acids
CID25731 Alkali metal nitrates
CID24562 Transition metal perchlorates
CID39732 Chlorinated dibenzo-p-dioxins
CID46936846 Cysteine and derivatives
CID46936284 Hexose phosphates
CID268208 Isoflavones
CID72194 Purine 2',3'-dideoxyribonucleosides
CID5761 Lysergic acids and derivatives
CID53461849 Saxitoxins, gonyautoxins, and derivatives
CID416228 Aconitane-type diterpenoid
CID3352 Trifluoromethylbenzenes

Figure 2.8 Examples of class assignments by *ClassyFire* for 12 compounds from the test set.

Given that *ClassyFire* is designed to classify compounds on a structural basis rather than a biological or biosynthetic basis, this kind of “misclassification” is completely understandable and is arguably not a misclassification. In this test set we also detected 13 missing assignments (false negatives). An example of a compound missing an assignment is the experimental drug cytidine-5'-diphospho-beta-delta-xylose (CID46936568), which was only classified as a pyrimidine ribonucleoside diphosphate but not classified as a purine nucleotide sugar (Fig. 2.9B).

To evaluate *ClassyFire*’s overall performance, each category was assigned a normalized weight based on its number of occurrences among the 800 chemical entities. This way, incorrect or missing assignments of the more populated categories (e.g. those at a higher level of the taxonomic hierarchy) would be penalized more compared to less populated categories (i.e. those at a lower level of the hierarchy). Each category was assigned to an average of 2.6 compounds. *ClassyFire* obtained score of 7067.04, or 99.97% of a maximum score of 7067.24. On average, *ClassyFire* was able to reproduce the text-based description with a precision of 99.8% and a recall of 99.9%.

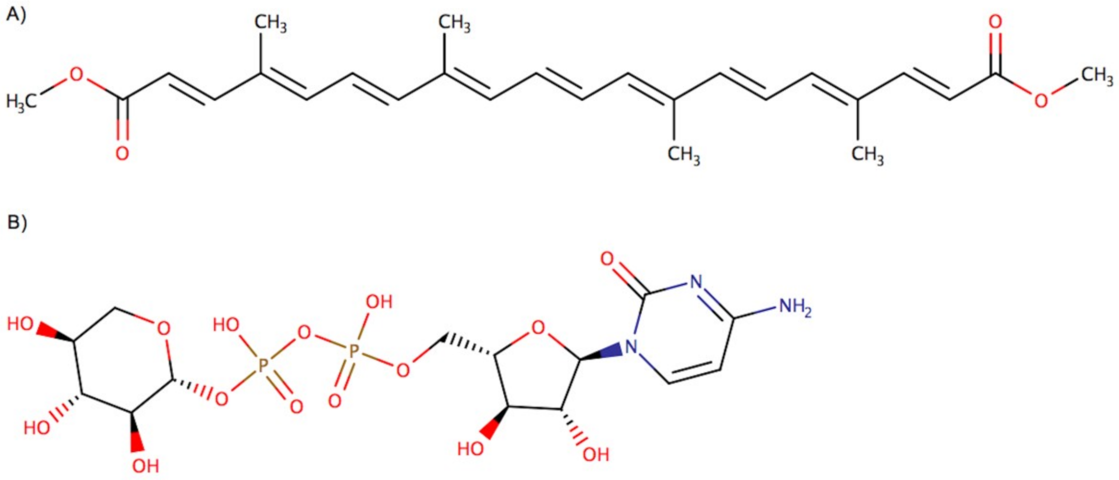


Figure 2.9 Examples of conflicting and missing class assignments. a) Structure of Bixin dimethyl ester (CID14413719). b) Structure of cytidine-5′-Diphospho-Beta-D-Xylose (CID 46936568).

2.5.1 Comparing automated and manual annotations

The primary motivation behind automated chemical classification is to provide a comprehensive, accurate and fast chemical annotation in order to alleviate the cost and potential errors of manual classification. While *ClassyFire* is many times faster than manual classification methods we also wanted to assess its accuracy and completeness compared to manual classifications. We therefore conducted a detailed comparison of *ClassyFire*'s results from 20 compounds, randomly selected from the test set described above, with their manually curated annotation from the ChEBI database. The 126th

ChEBI release from April 1st 2015 was used for this comparison. We did not use a more recent version of ChEBI since *ClassyFire* has actually been used over the past year to guide the manual annotation process for the ChEBI database. In order to provide the complete ChEBI annotation, a script was used to infer a list of ancestors for each of the 20 compounds based on the selected ChEBI release. Each compound was assigned an average of nearly 33 ChEBI classes. *ClassyFire*, on the other hand, returned an average of ~31 categories per compound. The ontology lookup table described in the Methods section of this paper was used to map categories returned by *ClassyFire* to the ChEBI classes. This mapping returned an average of 27 terms, or approximately 6 terms less than that originally provided by ChEBI.

This discrepancy can be explained by several factors. First, the idea behind the term mapping was to assign each ChemOnt category to an equivalent ChEBI term or, if not applicable, the closest ChEBI classes that do not have a parent-child relationship to each other. Thus, the category “Primary amines” (CHEMONTID:0002450) has been mapped only to the equivalent ChEBI term “primary amine” (CHEBI:32877), and not its parent. Additionally, the two hierarchies are built differently. While ChemOnt is built as a tree, where each node has no more than one parent, a ChEBI term can have several parents. For the purpose of our comparison, we complemented the list of the predicted ChEBI terms with their inferred parents. When the extended list is considered, each compound in the set was assigned to a total number of nearly 45 predicted ChEBI terms. Of those, an average of nearly 14 terms were missing from the manual ChEBI annotation. These could be added to ChEBI in order provide a more complete and consistent annotation. From the 33 terms provided by ChEBI, *ClassyFire* was unable to return an average of more than 2 terms per compound. This could either suggest that more terms should be added to the ChemOnt hierarchy, or the lookup table could be improved. In some cases, the term used is based on both a structural and a functional classification. An example is the term beta-lactam antibiotic (CHEBI:27933) for Oxacillin (CID 6196). Because ChemOnt is strictly structure-based, these terms do not apply. Overall, *ClassyFire* was able to reproduce ~94% of the ChEBI annotations, but also to suggest new terms that could accurately increase the number of annotations by another 43.6%.

The approach presented in this work makes use of diverse cheminformatic technologies to precisely detect structural features and classify chemical entities. The *ClassyFire* classification algorithm helps to (partially) overcome many of the limitations of previously developed automated chemical classification tools (205,207,208). For instance, several rules were developed to classify inorganic compounds, and organic metal compounds, which are not comprehensively covered by any current ontology. Most categories e.g., benzodiazepines, can be accurately described by one or more structural patterns. Others, such as alkaloids and derivatives, can only be defined as a disjunction of several subcategories. Furthermore, *ClassyFire* makes used of IUPAC names to identify certain patterns that might not be retrieved by a standard structure search, due to different substitution or dehydrogenation patterns. For example, we described a method to classify leukotrienes based on IUPAC names, given that there is no single structural backbone that could sufficiently and accurately describe each of these compounds.

2.5.2 Limitations

Despite the many capabilities that *ClassyFire* offers, and the different methods used to circumvent some of the formalization problems mentioned so far, certain limitations in

ClassyFire still remain. For instance, *ClassyFire*'s reliance on IUPAC names as a classification feature continues to cause some problems, particularly for compounds such as leukotrienes. This is because the classification of leukotrienes is also partly based on their biosynthetic origin. Certain, leukotriene derivatives that are oxidized or reduced at one double-bond position are still classified as leukotrienes, even though they might no longer have the three conjugated double bonds or the fourth double bond. An example is 10,11-dihydro-12-oxo-LTB4 (LMFA03020041) found in the LIPID MAPS database. Improvements could be made by taking a closer look at such compounds to find more common structural patterns. Currently, these leukotrienes would be classified as hydroxyeicosadienoic, hydroxyeicosatrienoic, or other eicosapolyenoic acids, depending on the number of carbon-carbon double bonds. Additionally, IUPAC names can become very difficult to exploit for certain complex structures, such as large fused ring systems. Another limitation with *ClassyFire* lies on its heavy dependence on predefined chemical patterns that use imperfect structure representation formats. Because *ClassyFire* inherits some of the limitations found with standard chemical structural representations (i.e. SMILES, SMARTS, Markush), the classification accuracy for certain kinds of "sandwich compounds" (e.g. metallocene) and alloys (e.g. chromium alloys) is not as good as it could be.

In order to circumvent the aforementioned limitations, and for the sake of developing a standard taxonomy, *ClassyFire* and ChemOnt could benefit from the involvement of the International Union of Pure and Applied Chemistry (IUPAC), and other chemical standardization or data reporting bodies. These groups could help to propose newer/better classifications and provide the long-term continuity that would, in turn, help to achieve a more sustainable and more consensual approach to chemical classification. Currently, the *ClassyFire* code is compatible only with the commercial ChemAxon JChem package. In order to ensure the sustainability of *ClassyFire*, we are committed to making *ClassyFire* a completely open source project that could benefit from contributions from the global cheminformatics community. *ClassyFire*'s continued maintenance and further development will be achieved under the joint supervision of The Metabolomics Innovation Center (TMIC), the National Institute of Health (NIH), the European Bioinformatics Institute (EBI), as well as IUPAC. We believe that this will facilitate the involvement and more widespread adoption of *ClassyFire* and ChemOnt, by the scientific community.

2.6 Use Cases

As mentioned earlier, the benefits and applications of a comprehensive chemical classification schema and well-defined chemical ontology system are multifold. Chemical classification makes chemical information easy to index, easy to organize, easy to search and easy to exchange. It also makes it possible to automate chemical annotations, to perform complex chemical searches, to rapidly identify compounds for compound-specific predictions, and to decipher patterns that underlie key biomolecular interactions. To illustrate this, we provide some example use cases showing how *ClassyFire*'s chemical classification has been used to help solve some common cheminformatics tasks.

2.6.1 Example 1: Classification of the PubChem Database

PubChem (9) is a freely available chemical database maintained by the National Centre for Biotechnology Information. It stores chemical, physicochemical and biological information for more than 91 million chemical entities as of June 2016, making it the largest, open-access chemical database in the world. However, as large as PubChem is, only 0.12% of the compounds in the database have ever been assigned to a chemical class or given a Medical Subject Heading (MeSH) classification. MeSH is a manually maintained, controlled vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching for biomedical and health-related documents, including all abstracts and papers listed in PubMed (221). Over the past 40 years, MeSH classifications have been assigned manually for just 115,000 compounds in PubMed, yet there are 60 millions compounds listed in PubChem. Given that the number of documents listed in PubMed is rapidly increasing, a manual assignment of the MeSH classes will become increasingly difficult. Moreover, it would be impossible to manually annotate all 60 million compounds in PubChem using the standard MeSH methodology. Therefore, we decided to automatically annotate and classify all of PubChem (and all PubMed chemicals) using *ClassyFire*. The structure-based classification of PubChem compounds was performed through parallel computing on 22 CentOS quad-core CPUs, with 3.6 GB of RAM each. The operation was completed in 424 hours for an average of 550 milliseconds (ms) per compound. The classification results have been submitted to the PubChem development group. This group is actively working to display *ClassyFire*'s classification of all the PubChem compounds, thereby allowing users to view, query and access compounds based on their ChemOnt classification. This should be completed by late 2016. With PubChem fully classified, the indexing of PubMed documents will now be much easier. Combining structure-based annotations with biological data could also assist scientists in various projects, such as ontology-based chemical enrichment analysis

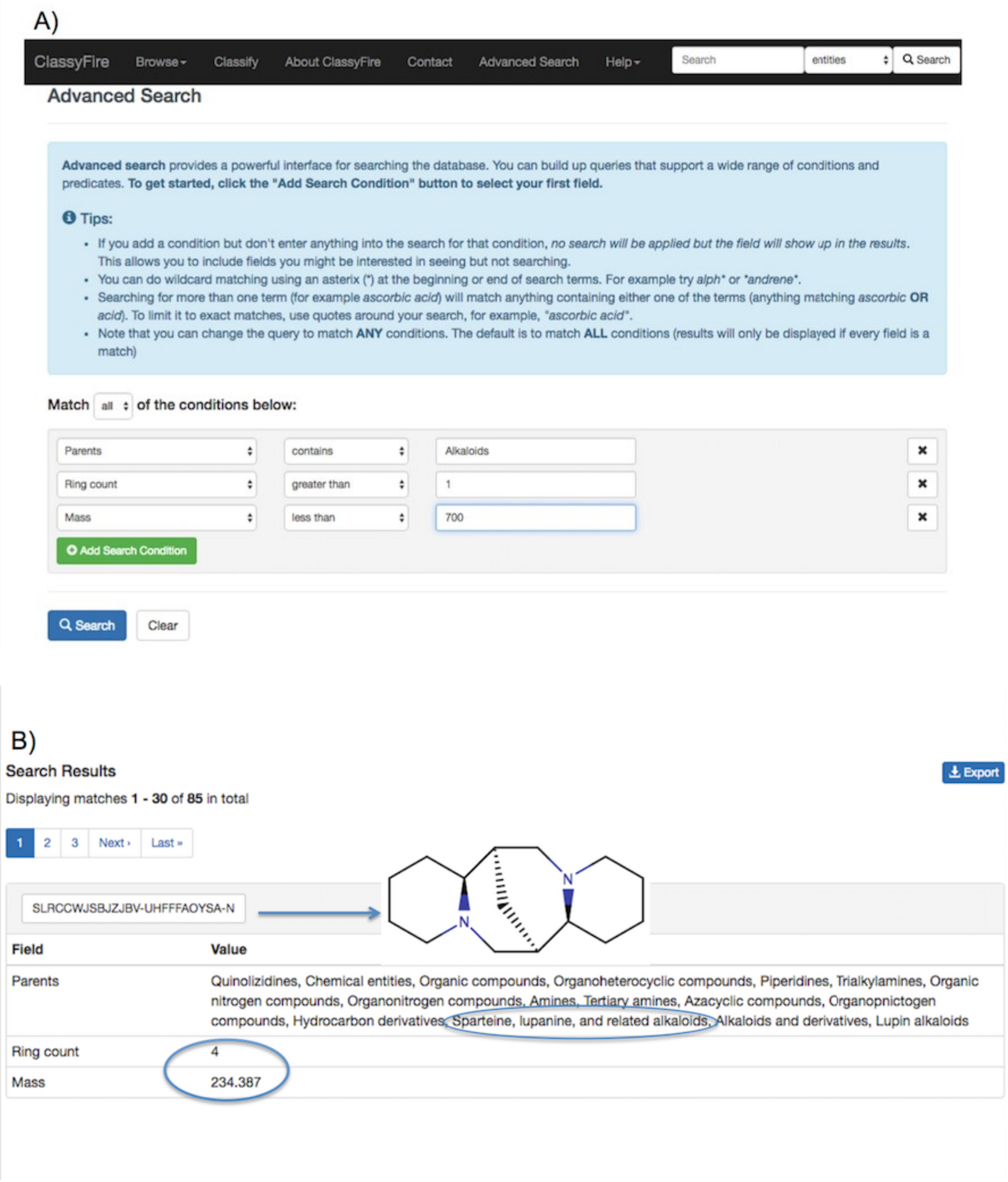
(201). Moreover, through *ClassyFire*, it is now possible to perform a variety of fast data searches and retrievals of PubChem data, as outlined below.

2.6.2 Example 2: Fast Searching and Data Retrieval

Chemical databases can typically be queried via physico-chemical parameters (e.g. mass) while others can be searched for the presence of functional groups (e.g. a ketone or carboxylic acid), among other properties. However, querying a chemical database with both substituent constraints and mass constraints is very difficult. For large databases, this would require one to perform structure-based searches over millions of compounds, which can take several minutes, even when the compounds are fully indexed. Moreover, certain structural constraints cannot be expressed using conventional structure-handling formats, such as SMARTS. Additionally, conventional substructure or structure-based searches do not allow one to search for chemicals belonging to categories such as "Alkanes" or "Alkaloids and derivatives". Having a chemical database annotated with substituent or chemical classification information can make these kinds of substituent and mass constraint searches very fast and easy. *ClassyFire* supports exactly this type of flexible search as it allows users to select compounds by defining a set of conditions based on various parameters such as, the chemical category, the mass, the number of rings, etc. These types of search combinations are very common in fields such as mass spectrometry, where compounds must be identified based on physico-chemical properties and relatively vague information about their putative substituents. *ClassyFire*'s text search operations are supported by Elastic Search (222), an open source search and analytics engine. As a result, compounds can be selected from over 77 million compounds stored in the *ClassyFire* database (as of June 2016) based on the ChemOnt terminology. Additionally, when needed, the results can be filtered based on physico-chemical properties. An illustration of how such a search can be conducted is provided in Figure 2.10, where *ClassyFire* returned a list of "Alkaloids containing more than one ring or, and having a mass lower than 700 daltons". The operation returned 30,392 hits through its text-based search in 509 ms. The results of the text-based search could be used to identify unknown structures obtained from biological samples. They could also be used to explore and cluster sets of small-molecules isolated from metabolomics or natural product extraction experiments.

2.6.3 Example 3: Automated Chemical Annotation

A growing number of chemical databases are being developed wherein detailed descriptions of individual chemicals are required. Examples include MetaCyc (220), ChEBI (200), DrugBank (99), T3DB, ECMDB (223) and FooDB (224). In many cases these descriptions must be manually composed and edited by experts and annotators. For well-known chemicals writing a comprehensive description is trivial. However, for lesser-known chemicals or chemicals where very little literature is available, the preparation of an even a short textual description of 20-30 words can take hours of library sleuthing and reading.



Advanced Search

Advanced search provides a powerful interface for searching the database. You can build up queries that support a wide range of conditions and predicates. **To get started, click the "Add Search Condition" button to select your first field.**

Tips:

- If you add a condition but don't enter anything into the search for that condition, *no search will be applied but the field will show up in the results*. This allows you to include fields you might be interested in seeing but not searching.
- You can do wildcard matching using an asterix (*) at the beginning or end of search terms. For example try *alph** or **andrene**.
- Searching for more than one term (for example *ascorbic acid*) will match anything containing either one of the terms (anything matching *ascorbic* **OR** *acid*). To limit it to exact matches, use quotes around your search, for example, *"ascorbic acid"*.
- Note that you can change the query to match **ANY** conditions. The default is to match **ALL** conditions (results will only be displayed if every field is a match)

Match all of the conditions below:

Parents	contains	Alkaloids
Ring count	greater than	1
Mass	less than	700

Add Search Condition

Search Clear

Search Results

Displaying matches **1 - 30** of **85** in total

SLRCCWJSBJZJBV-UHFFFAOYSA-N

Field	Value
Parents	Quinolizidines, Chemical entities, Organic compounds, Organoheterocyclic compounds, Piperidines, Trialkylamines, Organic nitrogen compounds, Organonitrogen compounds, Amines, Tertiary amines, Azacyclic compounds, Organopnictogen compounds, Hydrocarbon derivatives, Sparteine, lupanine, and related alkaloids, Alkaloids and derivatives, Lupin alkaloids
Ring count	4
Mass	234.387

Figure 2.10 Text-based search on the *ClassyFire* web server. A) Building the query. B) Sparteine, one of the returned compounds.

Because *ClassyFire* has a comprehensive Chemical Classification Dictionary consisting of thousands of 20-50 word textual descriptions for different compound classes, it is possible to use this Dictionary to automatically describe or annotate obscure or little-known compounds. In particular, *ClassyFire* was used to generate over 13,100 meaningful, 20-50 word descriptions for compounds in, ranging from drugs to poisons, for which no literature data was available. These precise, but automatically generated compound descriptions are now available in the HMDB, ECMDB, T3DB, FooDB, and YMDB (225).

2.7 Conclusions

In this paper, we have described a comprehensive, computable chemical taxonomy along with a structure-based ontology that permits the fully automated classification of most of the world's known chemicals. In particular we have described: 1) a well-defined, hierarchical classification structure consisting of up to 11 taxonomic levels; 2) a freely available Chemical Classification Dictionary (or ontology) consisting of >4,800 carefully identified and precisely described chemical classification terms, with over 9,000 synonyms; 3) a set of >9000 objective rules, patterns and criteria for classifying compounds on the basis of their structure; and 4) a computer program and a freely available web server (called *ClassyFire*) that performs rapid, accurate, automated rule-based taxonomic classification of chemical compounds. To our knowledge, this is the first freely available system that is capable of automatically, accurately and comprehensively organizing most of the world's known chemical entities into structural classes, at the scale presented.

The flexibility of *ClassyFire*'s source code and ChemOnt's chemo-taxonomic definitions, along with their open accessibility should allow *ClassyFire* and ChemOnt to easily evolve to fit with the ever-changing views of chemistry and with the increasing number of newly discovered scaffolds of natural and synthetic chemicals. In addition to developing an extensive taxonomy of organic compounds, we have also developed a comprehensive taxonomy for inorganic compounds consisting of 674 categories based on molecular formulas and atom types. We believe this is the first significant attempt to design a comprehensive computable chemical taxonomy for inorganic compounds.

ClassyFire's performance shows that the classification of chemical compounds can be accurately computed in a rapid, dataset-independent manner by relying solely on structural properties. Our data suggests that most chemical classes can be represented by one or more structural patterns. In certain cases, however, compounds from a given chemical category undergo reactions (e.g. loss of oxygen, substitutions) that might not match the constraints described in a category description. Some approaches to provide accurate descriptions in these scenarios would be to add more patterns, update position-specific constraints, and/or develop some heuristics for a more accurate classification. For instance, creating more rules for IUPAC name parsing could help to assign some classes more accurately. Overcoming these limitations would certainly improve the overall performance of *ClassyFire*.

It is important to emphasize that this taxonomic effort was not done in isolation. It has been jointly developed and tested by curators and developers some of the largest and most popular open-access chemical databases in the world, including PubChem, ChEBI, LIPID MAPS, DrugBank, HMDB and others. The *ClassyFire*/ChemOnt taxonomy is already being used in several of these databases and is expected to be adopted by several other chemical databases in the near future. Furthermore, the entire *ClassyFire*/ChemOnt taxonomy was mapped, in a joint effort, to several existing taxonomic/ontological schemes, such as the ChEBI and LIPID MAPS ontologies. As illustrated with the previous examples, applications of *ClassyFire* are multifold, spanning areas including drug design and metabolomics. *ClassyFire* has also found applications in the field of Chemical Health and Safety, where hazard assessment of small molecules, based on their structural features, has gained increasing interest recently.

ClassyFire is obviously not the final word on chemical classification or chemical taxonomies/ontologies. Given the size and complexity of the global chemical space along with the rapidly evolving needs of chemists and cheminformatics specialists, we expect that this subject (and this software) will evolve considerably over the coming years. Therefore, besides the freely available web service, we are actively working on a version of *ClassyFire* that has freely accessible source code and documentation. We are committed to making this resource fully open source (by December 2016). We believe this effort is an important first step towards the design of a fully computable, universally accepted chemical taxonomy and ontology.

Chapter 3

BioTransformer: An accurate, freely available tool for predicting secondary metabolism of small molecules in mammals

3.1 Introduction

Metabolism can be defined as the sum of all chemical reactions that take place in a cell or within an organism. Metabolism is key to the production of energy (catabolism), the generation of cellular building blocks (anabolism) as well as the activation, detoxification, and elimination of metabolic by-products or xenobiotics. Over the past 100 years considerable effort has gone into determining the precise molecular details of primary metabolism – i.e. the metabolic processes associated with the production and breakdown of essential metabolites (226). The citric acid cycle (227), gluconeogenesis (228), glycolysis (228), lipid biosynthesis (229), steroid metabolism (230) are all examples of primary metabolic processes that are now thoroughly understood. Unfortunately, somewhat less effort has been devoted to the characterization or understanding of non-essential or secondary metabolism and secondary metabolites.

Secondary metabolism typically refers to non-essential metabolites or metabolites generated through the detoxification and elimination of metabolic by-products or xenobiotics. Xenobiotics include compounds such as drugs, herbicides/pesticides, plant or food compounds, food additives, surfactants, solvents, cosmetics and other man-made or biologically foreign substances. In many cases these secondary metabolites are the products of promiscuous or non-specific enzymatic reactions (231,232), microbial or gut metabolism (233,234), liver-based phase I metabolism (oxidation, reduction or hydrolysis) or general phase II metabolism (conjugation). The characterization of secondary metabolites has long been vitally important to the pharmaceutical industry (232) but more recently it has become increasingly important to the herbicide/pesticide industry (235) and to the field of metabolomics (236). Indeed, it is widely thought that much of the so-called "dark matter" (47) in metabolomics is represented by large numbers of uncharacterized secondary metabolites.

The characterization or identification of secondary metabolites is quite difficult and is not unlike natural product identification or dereplication (237). It can take months or even years to purify and positively identify a secondary metabolite using standard analytical techniques. As a result, there has been growing focus on using computational tools to help with this process. Indeed, over the past two decades, a number of very effective computational tools have been developed to predict the secondary metabolism of xenobiotics – especially drugs. These computer programs typically require a starting parent molecule and employ pattern recognition along with hand-made rules or machine-learned algorithms to identify: 1) a site of reaction or a site of metabolism (SoM); and/or 2) a resulting chemical product. For a more detailed review of the main approaches and software tools used for *in silico* metabolism prediction, the reader is referred to section 1.9. Most *in silico* metabolism prediction tools are quite specific to certain classes of reactions or metabolic processes, such as phase I (only) or phase II (only) reactions. Table 3.1 provides a list of widely used *in silico* metabolism prediction tools. As can be seen from this table, some programs are commercial (*Meteor Nexus* (50), *MetabolExpert* (148)), some are freely available as web-servers (e.g. *MetaPrint-2D React* (159), *XenoSite* (238)) and others are freely accessible standalone software packages (e.g. SMARTCyp (146)). Most of these tools are focused on mammalian metabolism (e.g. *Meteor Nexus*), while a smaller number are targeted towards environmental applications (e.g. EAWAG-BBD (239)). Some *in silico* metabolism predictors, such as *SMARTCyp* and *isoCYP* (156) are limited to predicting phase I metabolism (or a portion of phase I metabolism), while others are more comprehensive (*Meteor Nexus,* and *SyGMa* (240)), covering a broad range of phase I and phase II biotransformations.

Table 3.1 Computational tools and resources used for *in silico* metabolism prediction.

Software	**Coverage**	**Approach**	**Output**	**Licensing**
SMARTCyp(146,160)	CYP450	Hybrid	SoMs	Free
StarDrop P450 (241)	CYP450	Hybrid	SoMs	Commercial
MetaSite(145)	CYP450	Hybrid	SoMs	Commercial
MetaPrint2D-React(159)	Phase I + II	Data mining/Machine learning	Metabolites	Free
MetabolExpert(148)	Phase I + II	Knowledge-based	Metabolites	Commercial
SyGMA(240)	Phase I + II	Knowledge-based	Metabolites	Available to academia
EAWAG-BBD/PPS(161,239)	Environmental microbial	Knowledge-based	Metabolites	Free
FAME(242)			Metabolites	Free for academic

Unfortunately, even with the growing abundance of *in silico* metabolism prediction tools, there continue to be a number of significant limitations, especially with regard to their performance, their utility and their accessibility. In particular: 1) very few tools predict more than the SoM; 2) only a small number of tools provide predicted structures, and those that do place restrictions on their distribution; 3) almost none of the existing tools are open source or open access; 4) very few of the tools make their databases or training sets available; 5) none of the comprehensive prediction tools are freely available; 6) none of the tools combine phase I, II, gut metabolism and promiscuous metabolism together; 7) many tools seriously over-predict metabolites and have remarkably high false positive rates (>90%); and 8) almost all of the tools were developed and trained on drug molecules and are not adapted for non-drug xenobiotics. These limitations have slowed the development of *in silico* metabolism prediction software and have also restricted the field to a tiny number of applications mainly in the pharmaceutical industry. Addressing these limitations and extending the capabilities of *in silico* metabolism prediction software could lead to substantial benefits in many other areas of analytical chemistry, natural product chemistry and metabolomics. These might include the *in silico* expansion of chemical databases of drugs (DrugBank (99)), food compounds (FooDB (224)), phytochemicals (PhytoHub (243)), environmental contaminants (ContaminantDB (244), T3DB (100)), organism-specific metabolites (HMDB (8), ECMDB (12), YMDB (225)), and other chemicals of biological interest (ChEBI (200)). This *in silico* expansion could lead to the discovery of new metabolite biomarkers, the development of better drugs and consumer products (e.g. food, household and cosmetic products), improved toxicology assessment, and the advancement of precision medicine (3).

In this chapter, we present *BioTransformer*, an accurate, freely available, comprehensive tool for *in silico* metabolism prediction of small molecules. It has been specifically designed to address essentially all of the shortcomings previously identified with existing *in silico* metabolism prediction tools. In particular, *BioTransformer* is open source, it is freely available, its databases and predictions are free to download and use, it calculates structures, it provides comprehensive (phase I, II, microbial, promiscuous and environmental) metabolite predictions, it is accurate and it covers a wide range of molecular classes. *BioTransformer* combines a knowledge (or rule)-based approach with a machine learning approach to predict 1) human CYP450-calyzed phase I metabolism of xenobiotics, 2) human gut microbial metabolism, 3) phase II metabolism, and 4) promiscuous metabolism of endogenous and exogenous compounds. It also implements a set of rules provided by the EAWAG-BBD system (239) to predict environmental microbial degradation. In addition to providing a description of BioTransformer we also provide a detailed analysis of its performance, including a number comparative analyses of *BioTransformer* with *Meteor Nexus* with regard to a number of experimentally determined metabolites identified after the ingestion of drugs, foods, plants and other xenobiotics by various mammalian species.

3.2 Structure and Implementation of BioTransformer

BioTransformer consists of five independent prediction modules called “transformers”, namely: 1) the EC-based transformer, 2) the CYP450 (phase I) transformer, 3) the phase II transformer, 4) the human gut microbial transformer, and 5) the environmental microbial transformer. For the prediction of metabolites, *BioTransformer* implements two approaches, a rule-based or knowledge-based approach, and a machine learning approach.

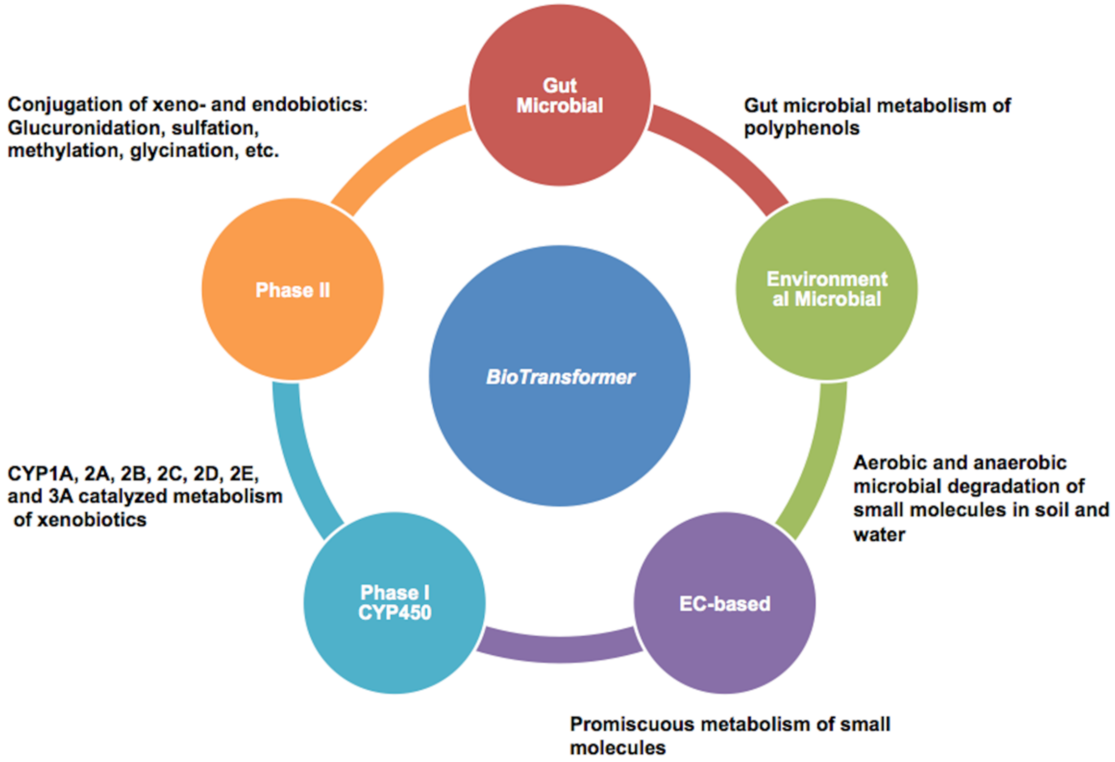

Conjugation of xeno- and endobiotics: Glucuronidation, sulfation, methylation, glycination, etc.
Gut Microbial
Gut microbial metabolism of polyphenols
Phase II
BioTransformer
Environmental Microbial
CYP1A, 2A, 2B, 2C, 2D, 2E, and 3A catalyzed metabolism of xenobiotics
Aerobic and anaerobic microbial degradation of small molecules in soil and water
Phase I CYP450
EC-based
Promiscuous metabolism of small molecules

Figure 3.1 Overview of *BioTransformer*'s five modules, the EC-based, CYP450, phase II, human gut microbial, and environmental biotransformers.

BioTransformer's knowledge-based system consists of three major components: 1) a biotransformation database (called BioTransformerDB) containing detailed annotations of experimentally confirmed metabolic reactions, 2) a reaction knowledgebase containing generic biotransformation rules, preference rules, and other constraints for metabolism prediction, and 3) a reasoning engine implemented separately for each "transformer" module. Its machine learning system uses a set of random forest prediction models for the prediction of CYP450 substrate selectivity. In this section, we describe the structure, content, and implementation of BioTransformerDB, the knowledgebase and the reasoning engine. Moreover, we briefly describe the CYP450 Metabolism Prediction System.

Finally, we will describe *BioTransformer*'s workflow. Figure 3.1 gives a brief overview of each module, their tasks, and the type of prediction approach they employ.

3.2.1 BioTransformerDB: A Small Molecule Biotransformation Database

BioTransformerDB is a database that consists of a manually curated collection of 1200+ experimentally confirmed biotransformations derived from the literature. It was developed to help with: 1) the design of biotransformation rules, 2) the training and validation of machine learning-based metabolism prediction models, and 3) the supply of known biotransformation data to *BioTransformer*'s reasoning engine. Each biotransformation in BioTransformerDB includes a starting reactant, a reaction product, the name or type of the enzyme catalysing the biotransformation and a reference. For the purposes of this document, a reactant is defined as a small molecule that binds to a specific enzyme and undergoes a metabolic transformation catalysed by that enzyme. A biotransformation describes the chemical conversion of molecular transformation of a reactant to one or more products by a specific enzyme (or enzyme class) through a defined chemical reaction. These biotransformations include the cytochrome P450-catalyzed phase I metabolism of ~400 unique starting reactants (and 780+ reaction products), the phase II metabolism of 300+ unique starting reactants (and 400+ reaction products) and human gut microbial metabolism of 50+ unique starting phenolic compounds (along with 80 reaction products). Cytochrome P450 enzymes (CYP450s) are responsible for > 90% of phase I oxidative reactions and >75% of drug metabolism (134), while UDP-glucuronosyltransferases (UGTs) and sulfotransferases (SULTs) are responsible for the phase II metabolism of most xenobiotics(245,246). In the gut microbiota, the enzymatic reactions are mostly reductive, and carried out by anaerobic bacteria due to the very low concentration of oxygen.

The "starting" reactants in the current version (1.0) of BioTransformerDB primarily consist of xenobiotics such as drugs, pesticides, toxins and phytochemicals. The database also includes a small number of sterol lipids and a selected set of mammalian primary metabolites. In assembling BioTransformerDB we gathered reaction data from the existing literature (>100 references) along with data downloaded from publicly available databases such as DrugBank (99), PharmGKB (247), XMETDB (153), SuperCYP (152). These databases list over 1,000 enzyme-substrate associations for the major CY4P50s and UDP-glucuronosyltransferases (UGTs). PhytoHub (243)was used to compile information about the metabolism of phenolic compounds in the gut.

The data curation process was conducted collaboratively with a small team of chemistry experts, and consisted of three phases. These phases involved: 1) the collection of biotransformation data, 2) the creation and annotation biotransformation objects and, 3) data validation. Enzyme-substrate associations were collected from various publicly available databases such as DrugBank, SuperCYP, and KEGG. Other enzyme-substrate associations were extracted manually from >100 scientific papers, review articles and drug metabolism textbooks, most of which were accessible electronically. When available, information about the structure and/or the name of a metabolite was also extracted. In some cases, insufficient information was provided about the exact reaction type, the structure of the resulting products or at least the site of metabolism. Moreover, for several compounds the reported sets of metabolites were either incomplete or conflicting. Such scenarios required further reading and checking by the annotation team in order to acquire more supporting evidence, and to further validate the data. During the data collection and validation process, enzyme associations were retained only if they had experimental supporting evidence about the correct structure of the metabolite, or both the site of metabolism and the reaction type. In total, 782 enzyme-substrate associations were validated.

For each biotransformation, the reactant and products were required to have a valid name and valid structural representations (SMILES string and standard InChIKey). The InChIKeys proved to be very useful for sorting, grouping and categorization, as well as in the indexing and searching of the chemical database. For most compounds the structures were available from online databases such as DrugBank, ChEBI (200), PubChem (9), and PhytoHub. When necessary, structures were generated using ChemAxon's MarvinSketch v.17.2.27.0 (84). In many cases the same compound was found to have several identifiers (e.g. names, synonyms IDs, etc.) and several structural representations that were linked to the same name, due to the existence of multiple salt forms, different protonation states, and tautomerism. This is a very common problem in managing chemical information, and represents a significant challenge in chemical data curation and aggregation (248). In an effort to eliminate this problem, all the structures were standardized through the removal of salts and charges. For mixtures, only the active compound was selected. Particular attention was also paid to stereochemistry, when the information was provided.

After the name and structure standardization process was complete, a list of unique compounds was created by comparing the standard InChIKeys and aggregating the data corresponding to each InChIKey. If no name was reported for a given reactant, product or metabolite, additional online databases were searched using the standardized structure until a name was found. If a given name could not be found, an appropriate chemical name was generated using ChemAxon's MarvinSketch. In certain simple cases, the name of the metabolite could be derived from information about the site of metabolism and the type of reaction (e.g. 3-OH glucuronide for a glucuronidation of the hydroxyl group at the C3 position). For each compound, identifiers from external databases were also collected, using *DataWrangler*, an in-house chemical annotation tool, which searches four major chemical databases for the given small molecule and returns various types of data, including links to other databases.

The reaction type for each assembled reaction or biotransformation was assigned by selecting the corresponding biotransformation rule in *BioTransformer*'s reaction knowledgebase (described in section 3.2). When the corresponding reaction type or pattern was unavailable, a new metabolic reaction object was added to the knowledgebase, as described later in this chapter. Certain enzyme classes, such as CYP450s, have very broad substrate specificity, and can catalyse a large pool of reactions. Therefore, it is common that a phase I oxidative transformation of a small molecule will be mediated by several CYP450s, one of which would be the major catalysing enzyme. Additionally, it is often the case that several reactions would apply to a single starting compound, leading to different metabolites (and metabolic pathways), at least one of which would be the major pathway. When reported, such information was also integrated in the database's annotation.

The validation process consisted of having one or more database curators check the correctness of the structures, names, reaction types and enzyme lists. For each reported biotransformation, a list of scientific sources providing supporting evidence was compiled. Because one of the main goals of BioTransformerDB is to support the development of *in silico* metabolism prediction models, the biotransformations had to be accurately reproduced when applying the specified reactions. Considerable effort was put into performing this specific task, as described in section 3.2.2, which also led to improvements in the encoded reaction descriptions in the reaction knowledgebase (section 3.2.1).

All the data in BioTransformerDB is stored as a JSON document. It currently contains 2,824 enzyme-reactant associations, 1,284 unique biotransformations, and 1,290 external database identifiers for a total of 1,428 compounds. Figure 3.2 displays an example of a BioTransformerDB entry corresponding to the N-dealkylation of the tranquilizing agent diazepam.

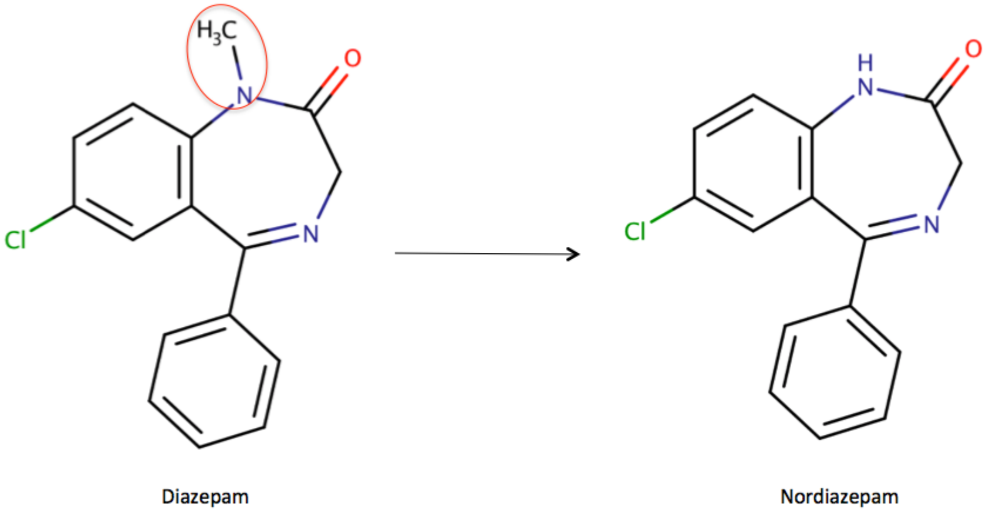


Figure 3.2 N-dealkylation of diazepam, a tranquilizing agent, as represented in BioTransformerDB. The methyl group (within the red circle) is substituted by an hydrogen atom.

3.2.2 The Reaction Knowledgebase

The reaction knowledgebase contains chemical reaction descriptions and rules encoded by SMILES (197), SMARTS (68), and SMIRKS (69) strings that are used by the reasoning engine to make biotransformation predictions. This knowledgebase encodes information about, and contains mapping data between, five different concepts: 1) the Biosystem, 2) the Metabolic Enzyme, 3) the Metabolic Reaction, 4) the Metabolic Pathway, and 5) the Chemical Class (as determined by *ClassyFire*) (Chapter 2). These concepts are defined as follows:

1) A **biosystem** is a living organism or a community of living organisms within which the biotransformation reactions can occur. Currently, the implemented biosystems are the human organism, the human gut microbiome, and the environmental microbiome.

2) A **metabolic enzyme** is an enzyme that catalyses or accelerates a metabolic reaction.

3) A **metabolic reaction** is a chemical reaction that modifies the structure of a molecule, leading to the generation of one or more products.

4) A **chemical class** refers to a group of chemicals that share a common structure feature or a group thereof as defined using *ClassyFire* (249).

5) A **metabolic pathway** is a linked series of chemical reactions that occur in a specific order in the cell or within the organisms. A metabolic pathway is organism-specific as an enzyme can be expressed by some organisms but not by others.

The interrelationships between the different concepts are illustrated in Figure 3.3. The construction of the reaction knowledgebase required data acquisition and aggregation from several sources, including the information captured in BioTransformerDB. Additional reaction information was gathered from resources such as the SIB Bioinformatics Resource Portal (ExPASy) (250), the BRENDA enzyme database (251), the Cyc database (112), the UniProt knowledgebase (UniProtKB) (252), the KEGG database(253), and enzyme nomenclature information provided by the International Union of Biochemistry and Molecular Biology (IUBMB) (254). The collected data was used to: 1) design, test, and validate generic reaction/transformation rules, 2) add constraints and rules that would be used by the reasoning engine, and 3) map entities (e.g. phosphatidylcholines, glycerophospholipids metabolism pathway, human) from different concepts (e.g. Chemical Class, and Metabolic Pathway, Biosystem). Based on the information gathered from the various resources, 398 associations could be established between the reaction knowledgebase's enzymes and reactions. Priority was given to enzymes with wide substrate specificity such as the arylamine N-acetyltransferase (EC 2.3.1.5), as the aim was to predict the metabolism of small molecules partly based on generic biotransformation rules. Exceptions included serine palmitoyltransferase (EC 2.3.1.50), which provides the sphingoid base 3-dehydrosphinganine needed for the biosynthesis of sphingolipids. In total, 680 biotransformation rules were created and associated with at least one enzyme. The biotransformation rules were encoded in the SMIRKS language (69). For each biotransformation rule, one or more structural constraints (e.g. the known enzyme substrates are restricted short-chain fatty acyl chains) were encoded separately, either in the SMARTS language (68) or programmatically (by combining several rules based on the structural constraints and/or physico-chemical properties).

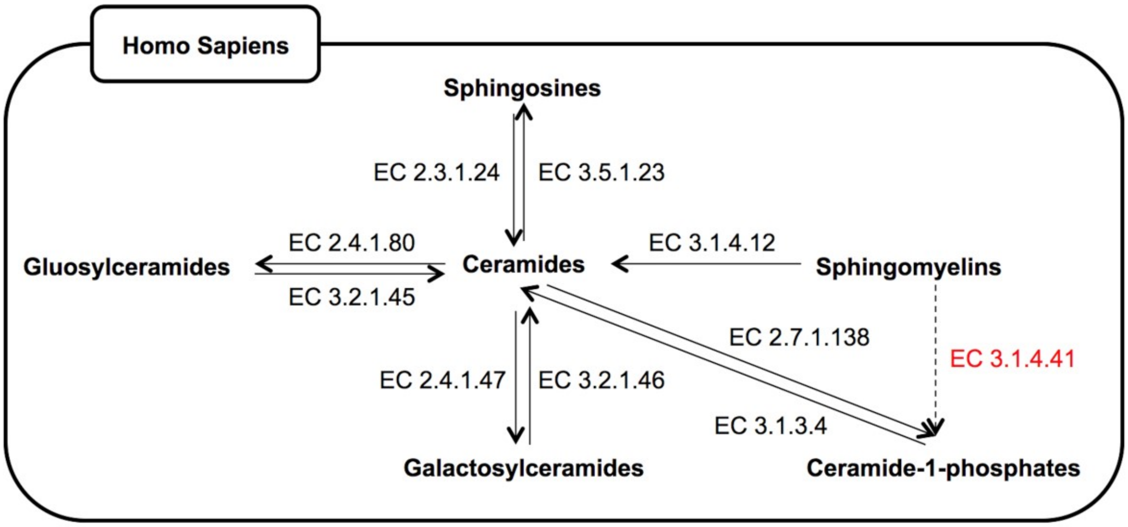


Figure 3.3 Interrelationships between the five different concepts represented in the *BioTransformer*'s knowledgebase. The figure depicts a small portion of the sphingolipid metabolism pathway in humans, as provided by the KEGG database. An example of a metabolic reaction is the conversion of compounds from the chemical class of sphingomyelins into their corresponding ceramides (as shown by the corresponding arrow) by the enzyme sphingomyelin phosphodiesterase (EC 3.1.4.12). The dotted arrow shows the conversion of sphingomyelins to ceramide-1-phosphates by sphingomyelin phosphodiesterase D (EC 3.1.4.41), which is expressed in *Aspergillus flavus*, but not in humans.

The separate design of structural constraints was necessary for several reasons. First, structural constraints can sometimes be difficult or impossible to fully encode using the SMIRKS language alone, due to its limited expressivity. Second, the juxtaposition of constraints within a SMIKRS pattern can make it difficult to understand, and cumbersome to update.

A typical reaction scheme encoded in the reaction knowledgebase is shown in Figure 3.4, which illustrates the biotransformation of 1,2-dihexanoyl-sn-glycero-3-

phosphoserine (PS(6:0/6:0)) into 1,2-dihexanoyl-sn-glycero-3-phosphoethanolamine (PE(6:0/6:0)) by the human phosphatidylserine synthase 2 (EC 2.7.8.29). The encoding of this generic reaction via SMIRKS and SMARTS allows the reaction to automatically replace the ethanolamine in any diacyl-sn-glycero-3-phosphoethanolamine by a serine molecule to produce the corresponding diacyl-sn-glycero-3-phosphoserine.

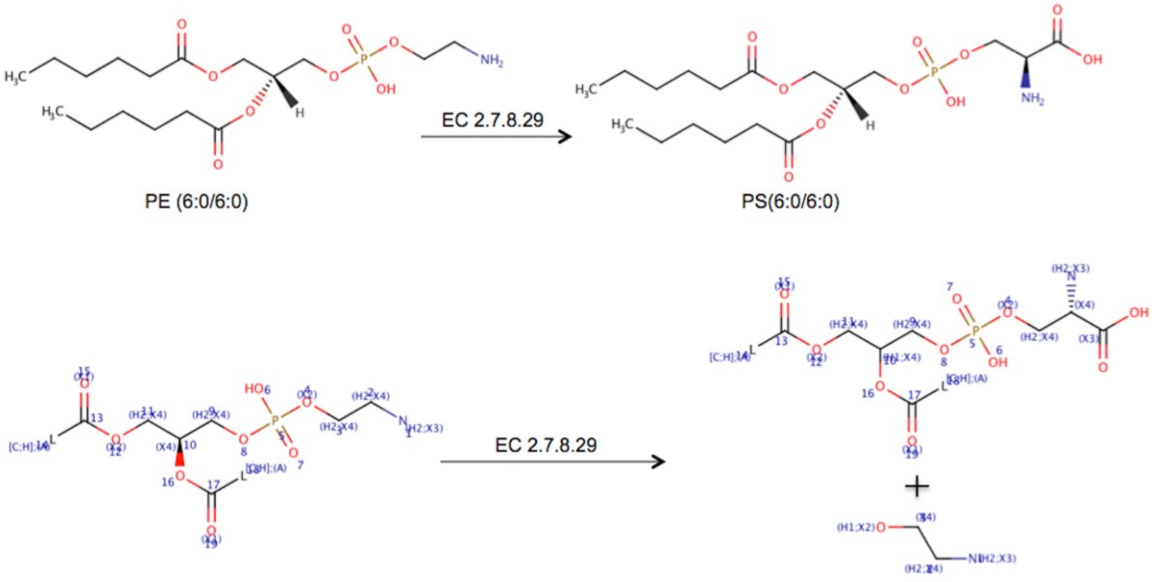


Figure 3.4 Encoding a phosphatidylserine biosynthetic reaction. A) Metabolism of 1,2-dihexanoyl-sn-glycero-3-phosphoethanolamine (PE(6:0/6:0)) to 1,2-dihexanoyl-sn-glycero-3-phosphoserine (PS(6:0/6:0)) by the human phosphatidylserine synthase 2 (EC 2.7.8.29). B) The encoding of the reaction requires that the atoms be indexed, so that the substitution can be accurately executed.

Once a reaction was encoded, several tests were performed to assess its correctness by applying the reaction to known substrates as well as to non-substrates (i.e. chemicals that were known not to satisfy the various constraints). If the reaction passed all the tests it was added to the database; if it failed, the reaction schema was subject to one or more iterations and tests until validated. Some of the encoded reactions in the reaction knowledgebase apply to a very limited set of chemicals, and can be used to accurately predict the metabolism of compounds belonging to those classes. Such examples include the aforementioned conversion of diacyl-sn-glycero-3-phosphoethanolamines to diacyl-sn-glycero-3-phosphoserines, and the metabolism of several classes of lipids, which are known to follow classic primary metabolic pathways. Other reactions are so generic or non-specific that they would lead to the high number of false predictions if applied blindly. Some examples of highly non-specific reactions include aliphatic hydroxylation, N-dealkylation, and glucuronidation, among many others. These reactions are catalysed by enzymes that have broad substrate specificity, such as CYP450s and UGTs. To handle these situations, new reaction subtypes and constraints were defined, which focused on a specific subclass of compounds that fulfilled a defined set of structural constraints. The resulting manually generated rules were then subject to further testing and validation. An example of such a reaction is the N-dealkylation of alicyclic tertiary amines catalysed by CYP3A4 (among others), a well-studied bioactivation pathway of cyclic amines (135).

In addition to the core knowledge provided by textbooks, online databases and journal articles, the design of biotransformation rules for the reaction knowledgebase often required additional investigation. One approach consisted of selecting compounds (from BioTransformerDB) that triggered a given reaction and labeling them based on whether their expected metabolites were reported or not. Further analysis of these reaction sets often suggested new reaction schemes or the addition of new constraints to existing reaction schemes. A similar process was previously used to generate 300+ biotransformation rules for the prediction of environmental microbial metabolism

(161,239). These rules were also encoded, tested and added to *BioTransformer*'s reaction knowledgebase. Overall, a total of 797 biotransformation rules were encoded, tested, and added to the reaction knowledgebase.

In addition to identifying the mechanisms involved in various metabolic reactions, and the encoding of biotransformation rules, another challenge to building the reaction knowledgebase was the prioritization of specific metabolic reactions. For any compound that triggers several competing reactions, certain reactions are more likely to occur than others. Therefore the metabolites resulting from these preferred reactions are more likely to be observed. Given a pair of metabolic reactions, a common approach to define precedence rules involves a detailed analysis of common putative and observed metabolites via NMR or mass spectrometry (161). Another approach involves using NMR or mass spectrometry to perform time-course monitoring of biotransformations in order to elucidate the preferred metabolic pathways (255). In this work, our construction of precedence rules between pairs of reactions was mostly based on data acquired from previously reported scientific studies, as well as observations made in previous studies.

For instance, when absorbed in the small intestine, polyphenolic compounds must be deconjugated first (via glucuronidases or carboxylesterases) before undergoing any reductive transformation (256,257). Recently, Burapan, S. *et al*. (255) investigated the regioselectivity of O-demethylation by the human gut bacterium *Blautia Sp.* MRG-PMF1, and concluded that O-demethylation of polymethoxyflavones occurs most preferably at the C-7 position, compared to the C-4' and C-3 positions. Based on these observed patterns, kaempferol 7,4'-dimethyl ether 3-glucoside would more likely undergo O-deglycosylation, followed by C-7 O demethylation to give kaempferol 4'-methyl ether, which will then undergoes further metabolism (see Figure 3.5). In total, 190 precedence rules were created for 49 unique biotransformation rules that were encoded for the human and/or human gut microbial biosystems. In addition, 1960 precedence rules for 195 unique biotransformation rules were adopted from the EAWAG-BBD system (environmental microbial metabolism).

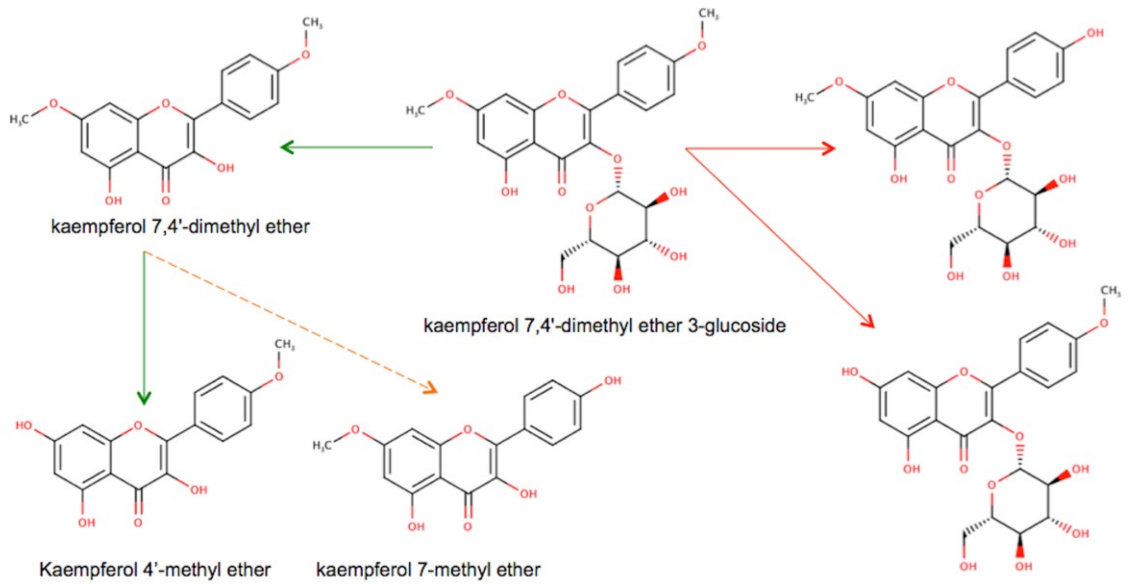

kaempferol 7,4'-dimethyl ether
kaempferol 7,4'-dimethyl ether 3-glucoside
Kaempferol 4'-methyl ether
kaempferol 7-methyl ether

Figure 3.5 Metabolism of kaempferol 7,4'-dimethyl ether 3-glucoside in the human gut microbiome. The encoding of preference rules provides a more likely metabolism pathway leading to kaempferol 4'-methyl ether.

Not all reaction schemes in the reaction knowledgebase are fully specified. For instance, because relatively little is known about the biology and enzymology of the gut microflora, a large number of encoded biotransformation rules were either assigned to an enzyme superfamily or to an “unspecified enzyme”. For the knowledgebase’s collection of environmental microbial reactions, the biotransformation rules were assigned to a single “unspecified enzyme”, as they are often consensus rules designed by combining patterns of reactions catalysed by several enzymes. Upon validation of the reactions and the addition of constraints, 1,408 enzyme-based reaction associations were created. The next step consisted of associating enzymes with metabolic pathways, and the corresponding biosystems. This step is very important for several reasons. First, many metabolic pathways are organism-dependent as different organisms express different enzymes or transporters (see Figure 3.3). Thus, as illustrated in Figure 3.3, the metabolic route linking a compound to a metabolite could vary between organisms. While sphingomyelins can be directly converted into ceramide-1-phosphates in *Aspergillus Flavus*, humans must convert sphingomyelins into ceramides first, which are then transformed into ceramide-1-phoshphates. Second, the mapping also allows one to encode more constraints and exclusion rules for certain types of compounds. For instance, glycerophospholipids are transformed solely within the glycerophospholipid metabolism pathway, and do not undergo CYP450- or UGT catalysed metabolism. In total, seven metabolic pathways were created, 84 enzyme-pathway associations, and nine chemical class-pathway associations were created for the human biosystem. A summary of the numbers of rules and associations encoded in the biotransformation database are shown in Table 3.2 for each of the five transformers (EC-based, human CYP450, Human gut microbial, phase II, and environmental microbial). The biotransformation rules and the list of enzymes cover the EC classes EC1 through EC6, with more focus on classes EC1 to EC4. The metabolic pathways are currently limited to lipid metabolism.

	No. of enzymes	**No. of biotransformation rules**	**No. enzymes-rules associations**	**No. of covered biosystems**
EC-based	193	252	282	1
CYP450	9	132	690	1
Human Gut	25	135	135	1
Phase II	8	40	46	1
Environmental microbial	1	301	301	1

Table 3.2 Statistics for each of the five biotransformers.

3.2.3 The Reasoning Engine

BioTransformer's reasoning engine uses the rules in the biotransformation database to select the most likely of all applicable metabolic biotransformations/pathways. In general, two types of reasoning are used for the selection/ranking of predicted metabolites: absolute reasoning, and relative reasoning (258). Absolute reasoning solely focuses on the likelihood of a biotransformation to occur, and is used to select the biotransformations with an occurrence ratio above a given threshold. Examples of biotransformation software using absolute reasoning include *SyGMA* and *Meteor Nexus*. Relative reasoning evaluates the comparative likelihood between two independent but competing reactions (e.g. flavone 7-O-demethylation if more likely to occur than flavone 4'-O-demethylation (255)). Examples of computational tools using relative reasoning include *Meteor Nexus* and the EAWAG-BBD/PPS system. The current version of *BioTransformer* only provides an option to use relative reasoning.

Besides qualitative attributes (e.g. chemical class), reasoning engines often also use quantitative attributes (e.g. mass, LogP) to guide their predictions. *BioTransformer*'s reasoning engine uses both types of attributes. While chemical classification can help to select the most likely biotransformations or discard the unlikely ones, quantitative attributes such as the mass and LogP are used to predict the substrate specificity for various enzymes. For this specific task, the current version of *BioTransformer* focuses on nine of the most "active" or best-studied CYP450 enzymes. The prediction of their specificity toward a given substrate is made by *CypPred*, a machine learning based software tool developed in a collaborative project. *CypPred* is unpublished as of now, and will be briefly described in section 3.2.4.

With the reaction knowledgebase in hand, the reasoning system was implemented programmatically for each of the five different transformers. The rationale behind this design was to have independent transformers that could be used separately. This way, one could focus on a specific type of metabolism (e.g. CYP450-ctalyzed metabolism) or a specific type of biosystem (human). Among the five transformers, four rely solely on the application of rules and constraints from the reaction knowledgebase. These four are the EC-based transformer, the phase II transformer, the human gut transformer and the environmental transformer. The cytochrome P450 (phase I) transformer, which focuses on the metabolism of small molecules mediated by CYP450 enzymes, is the only one that implements a machine learning approach in combination with a knowledge-based approach. In addition to the five transformers, the reasoning engine is used by a combined human "super transformer", which aims at simulating the metabolism of small molecules in humans (including the human gut), from their absorption to their excretion.

3.2.4 The CYP450 Metabolism Prediction System

Cytochrome P450 enzymes (CYP450s) constitute a superfamily of heme proteins, with over 50 isozymes identified in humans (259). They are predominantly found in the liver, but also occur in other organs such as the lungs and the kidneys. CYP450s are the major oxidative enzymes in the human body, and are responsible for the metabolism of a large number of compounds. Nine specific CYP450s have been identified as responsible for most of the phase I metabolism of xenobiotics (e.g. drugs, food additives, and environmental contaminants) and a small number of endogenous compounds. These include the CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2D6, CYP2E1, and CYP3A4 isozymes. Because of their broad specificity, a special CYP450-reactant specificity prediction was implemented, in order to predict metabolites for the more likely reactants only. The enzyme-specificity is assessed by a program called *CypPred*. *CypPred* is a software tool that uses a machine learning based approach (random forest (260)) to predict whether a small molecule reacts with any of the CYP450 isozymes. *CypPred* provides nine random forest models, one for each of the isozymes. These models use the physico-chemical properties and a substructure fingerprint of a molecule for their prediction. The substructure fingerprints were partly developed by including a subset of SMARTS pattern definitions from *ClassyFire* (249), a set of SMARTS patterns known to trigger CYP450-catalyzed metabolism (e.g. p-substituted phenols, or N-substituted piperazine), the corresponding PubChem fingerprint (9), and the MACCS fingerprint (79). These fingerprints encode pattern definitions for key functional groups and structural features relevant to CYP450-catalyzed metabolism, which were obtained through data mining. In addition to the nine models, *CypPred* also used a heuristic approach to filter candidates that are known to be out of scope for CYP450 mediated metabolism, based on their chemical structure and/or physico-chemical properties. *CypPred* is freely available at https://bitbucket.org/Leon_Ti/cyppred/overview.

Given any small molecule, the CYP450 transformer uses *CypPred* to predict which of the nine CYP450s is likely to metabolize the molecule. Subsequently, it implements the constraints and biotransformation rules encoded within the knowledge database to predict the metabolites. As for any other transformer, the user can vary the parameters, including the number of transformation steps, and whether to use precedence rules.

3.2.5 *BioTransformer*'s Input and Workflow

BioTransformer was implemented in the Java programming language, and can be used as a command-line tool to predict the metabolism of small organic molecules. Beside *CypPred* described in the previous section, *BioTransformer* uses two other open-access tools, namely the Chemistry Development Kit (CDK) (261), and the AMBIT library (262). The CDK programming library is used for several operations, including the calculation of physico-chemical properties, the execution of superstructure search operations, and the handling of chemical structures, among others. The AMBIT library is used for the application of biotransformation rules and structure generation.

BioTransformer's workflow is illustrated in Figure 3.6. As can be seen in this diagram *BioTransformer* accepts molecules either in SMILES (single molecule), MOL (single molecule), or SDF (single or multiple compounds) format as input. Each molecule must be an organic molecule and it must not be a mixture. Once the input is parsed, the structures are subjected to chemical validation and standardization. The standardization process consists of removing charges from functional groups (with some exceptions, such as nitro groups), checking and validating bond types and adding explicit hydrogen atoms. Subsequently, *BioTransformer* predicts biotransformations and the resulting metabolites for each query molecule separately. The prediction can be run in single mode, involving one of the five transformers (CYP450, EC-based, phase II, gut microbial, or environmental microbial). Additionally, a human "super transformer" has been implemented to mimic the metabolism of small molecules in the human "superorganism", which also includes the gut microbiota. This super transformer integrates the CYP450, EC-based, phase II, gut microbial transformers and covers a number of different reaction types, including hydrolysis, oxidation and reduction, and conjugation. The prediction step is followed by the "metabolic tree reconstruction and metabolite annotation" step. Based on the information from the predicted biotransformation, *BioTransformer* builds a metabolic tree by associating each metabolite with its parent(s). Moreover, each predicted metabolite is annotated with various information that provides structural identification, reports its physico-chemical properties, and explains its origin. The data includes: 1) three chemical identifiers (metabolite ID, InChI, InChI Key), 2) the molecular formula, 3) the monoisotopic mass, 4) the reaction type leading to the metabolite, 5) the biosystem that generated the molecule, 5) the parent compound identifiers (BioTransformer ID, InChIKey) and, 6) the parent monoisotopic mass. For each query molecule, the results are returned in a separate SDF file that contains the structure and annotation of each metabolite. The returned information can be used separately to build a metabolic tree. It can also be used to compute neutral losses for MS-based analyses that can be used to experimentally detect each biotransformation.

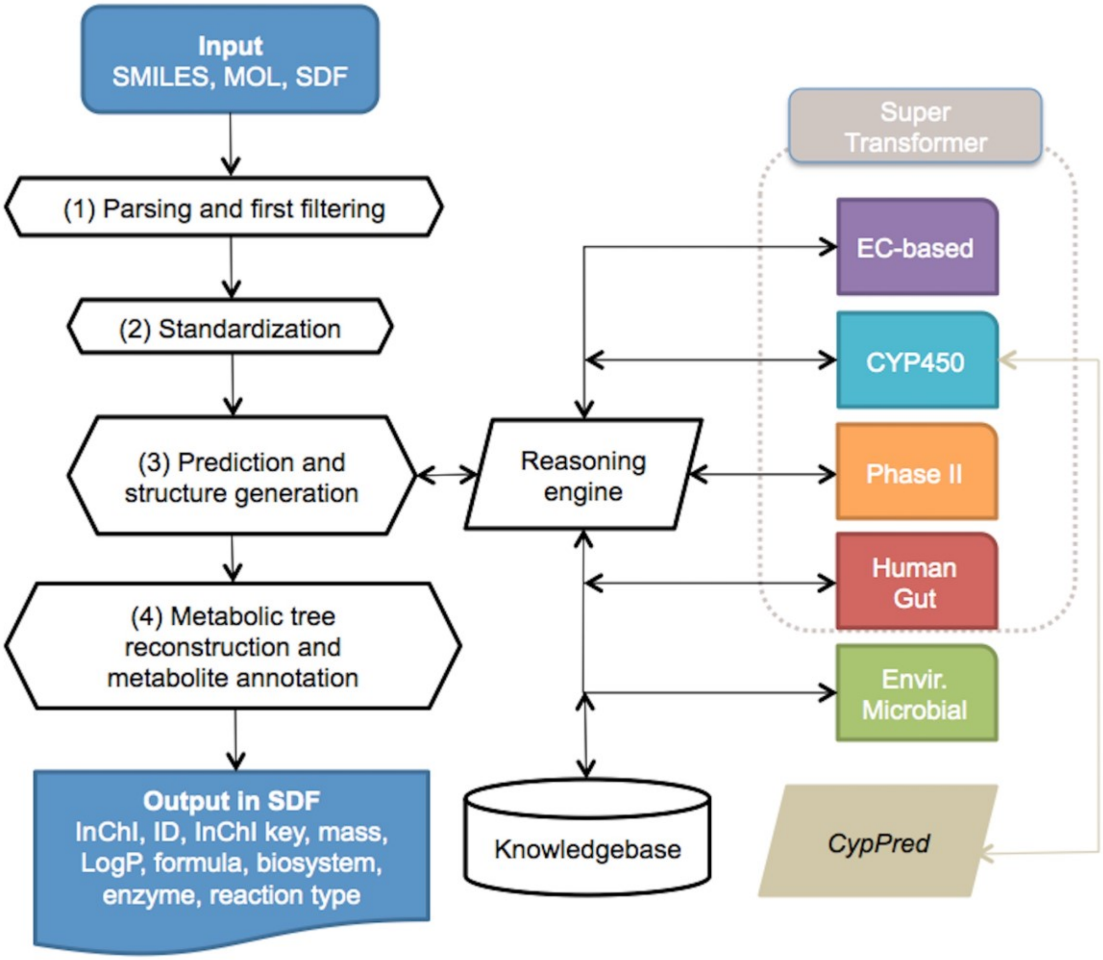

Input
SMILES, MOL, SDF
(1) Parsing and first filtering
(2) Standardization
(3) Prediction and structure generation
(4) Metabolic tree reconstruction and metabolite annotation
Output in SDF
InChI, ID, InChI key, mass, LogP, formula, biosystem, enzyme, reaction type
Reasoning engine
Knowledgebase
Super Transformer
EC-based
CYP450
Phase II
Human Gut
Envir. Microbial
CypPred

Figure 3.6 *BioTransformer*'s workflow.

3.3 Evaluation of *BioTransformer*'s predictions

In order to evaluate the performance of *BioTransformer*, we performed a comparative analysis with two popular *in silico* metabolism prediction tools, namely *Meteor Nexus* (50) and the EAWAG BDD/PPS system (49,161,239). The procedures and results are presented below.

3.3.1 Evaluation of *BioTransformer*'s Single-step Metabolism Prediction in Mammals

In order to evaluate whether *BioTransformer* could accurately predict single-step biotransformations already reported in the biotransformation database, we randomly selected three compounds from the database with known CYP450 metabolism profiles. These include caffeine (a food compound), omeprazole (a drug), and disulfoton (a pesticide). *BioTransformer* was set to apply relative reasoning and no cut-off. Its prediction results were compared to those provided by *Meteor Nexus*. *Meteor Nexus* was set to apply the following constraints: 1) all reactions applicable to CYP450s; 2) absolute/relative reasoning; 3) a cut-off of two; 4) a maximum of 60 metabolites; and 5) breadth first as processing direction. To predict its biotransformation products the settings remained the same in *BioTransformer*. In *Meteor Nexus* we modified the set of reactions to apply all possible reactions (without manual optimization), while the other settings remained unchanged. For the three xenobiotics (caffeine, omeprazole, and disulfoton), the evaluation of the prediction was based the analysis of reported metabolites

Overall, *BioTransformer* was able to predict a total of 13 metabolites, including 6 of 9 previously reported ones. In comparison, *Meteor Nexus* was able to predict 11 metabolites; including 5 of 9 previously reported ones. The results are displayed in Table 3.3.

Compound	**Prediction**	**BioTransformer**	**Meteor Nexus**
Omeprazole	True predictions	4	4
	False predictions	3	2
	Missed predictions	0	0
Caffeine	True predictions	1	1
	False predictions	1	1
	Missed predictions	3	3
Disulfoton	True predictions	1	0
	False predictions	3	3
	Missed predictions	0	1

Table 3.3 Comparison between *BioTransformer* and *Meteor Nexus* for the CYP450-catalyzed single-step metabolism of three xenobiotics.

3.3.2 Evaluation of *BioTransformer*'s Multi-step Metabolism Prediction in Mammals

To evaluate *BioTransformer*'s ability to predict multi-step biotransformations, we selected the flavan-3-ol compound epicatechin (an antioxidant from tea leaves(263)), and the monoterpene carvacrol (a chemopreventive agent and antioxidant from essential oils (264)). Carvacrol and epicatechin are both extensively metabolized in mammals and their metabolites are well known and well characterized. In particular, the multi-step metabolism of epicatechin is carried out in both the liver and the human gut (colon and intestine). For the comparative analysis against *Meteor Nexus*, the metabolism prediction was applied over multiple phases (Phase I, Phase II, as well as human gut microbial), in order to give a comprehensive overview of the metabolism, and to test the ability of both tools to predict various types of enzymatic biotransformations within humans. *BioTransformer*'s human super transformer was used (see section 3.2.5) by applying the following settings: 1) all applicable reactions, 2) relative reasoning and, 3) no cut-off. *Meteor Nexus* applied the following settings: 1) a selected set of 22 chemical reactions, 2) absolute/relative reasoning, 3) a cut-off of 2, 4) a max depth of 4, 5) a maximum of 60 metabolites and, 6) breath first as processing direction. The settings for *Meteor Nexus* were set based on previous analyses (on other similar polyphenolic compounds), in order to optimize its performance for epicatechin.

Compound	**Prediction**	**BioTransformer**	**Meteor**
Epicatechin	True predictions	20	3
	False predictions	22	51
	Missed predictions	4	21
Carvacrol	True predictions	12	14
	False predictions	18	22
	Missed predictions	3	3

Table 3.4 Comparison between *BioTransformer* and *Meteor Nexus* for the multi-step metabolism of epicatechin.

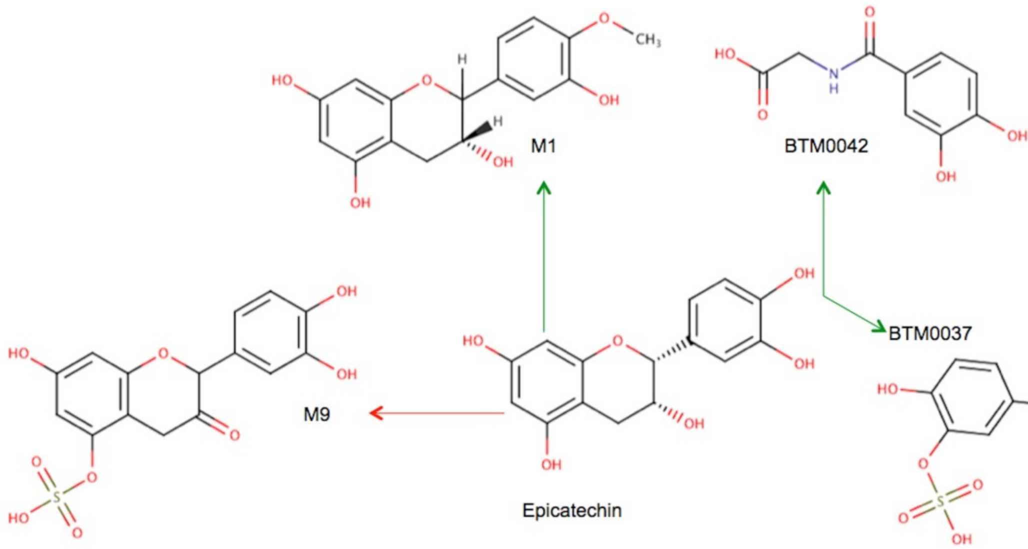


Figure 3.7 Examples of predicted metabolites of epicatechin. The green arrows point to correct metabolites, identified by *BioTransformer* and *Meteor Nexu*s. The red arrow points to a false prediction (M9) by *Meteor Nexu*s.

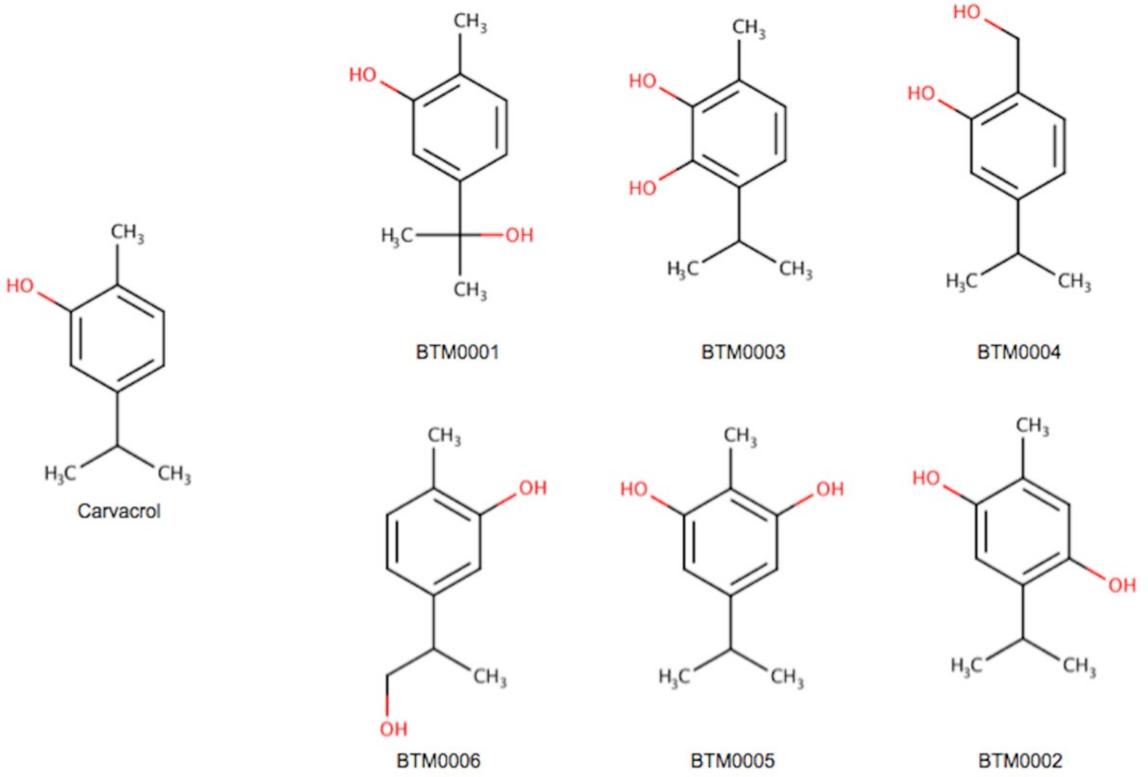


Figure 3.8 Examples of predicted metabolites of carvacrol.

Overall, *BioTransformer* predicted 42 epicatechin metabolites, and 22 out of 24 reported metabolites for epicatechin. *Meteor Nexus* predicted 51 metabolites, 3 of which had been reported. For the monoterpene carvacrol, *BioTransformer* predicted 30 metabolites, and 12 out of 15 previously reported previously metabolites. In comparison, *Meteor Nexus* predicted 36 metabolites, and 14 out of 15 previously reported ones. The detailed results of our analysis are shown in Table 3.4. Examples of predictions are in Figure 3.7 (epicatechin) and Figure 3.8 (carvacrol).

3.3.3 Comparative Analysis with The EAWAG BBD/PPS System

Meteor Nexus is not capable of predicting environmental microbial metabolism/degradation; thus, in order to assess *BioTransformer's* abilities to predict environmental microbial metabolism, we compared to the EAWAG-BBD/PPS system using three test compounds, namely Ampicillin (an antibiotic), Nitroglycerin (a plasticizer, a drug), and Disulfoton (an insecticide), all of which (along with their metabolites) have been found in wastewater treatment plants (265-267). Here, only *BioTransformer's* environmental microbial biotransformer was used, and one step of biotransformation was used for each compound. The aim of this comparison was to assess the ability of *BioTransformer* to reproduce the EAWAG-BBD/PPS predictions, since the rules applicable to environmental degradation were encoded using the freely accessible EAWAG Biodegradation and Biocatalysis database. Both *BioTransformer* and the EAWAG-BBD/PPS system were set to apply relative reasoning, and both were set to predict all microbial transformations (i.e. aerobic and anaerobic).

BioTransformer was able to replicate all 15 biotransformations predicted by the EAWAG system, and a total of 18 out of 18 metabolites. In addition, *BioTransformer* predicted three more metabolites for the degradation of Disulfoton. All three metabolites resulted from the correctly used biotransformation rule (bt0259), which was applied at three different sites of metabolism, producing two metabolites in each case. Figure 3.9 displays the metabolites predicted by *BioTransformer* and the EAWAG system, and highlights the metabolites reported only by *BioTransformer.*

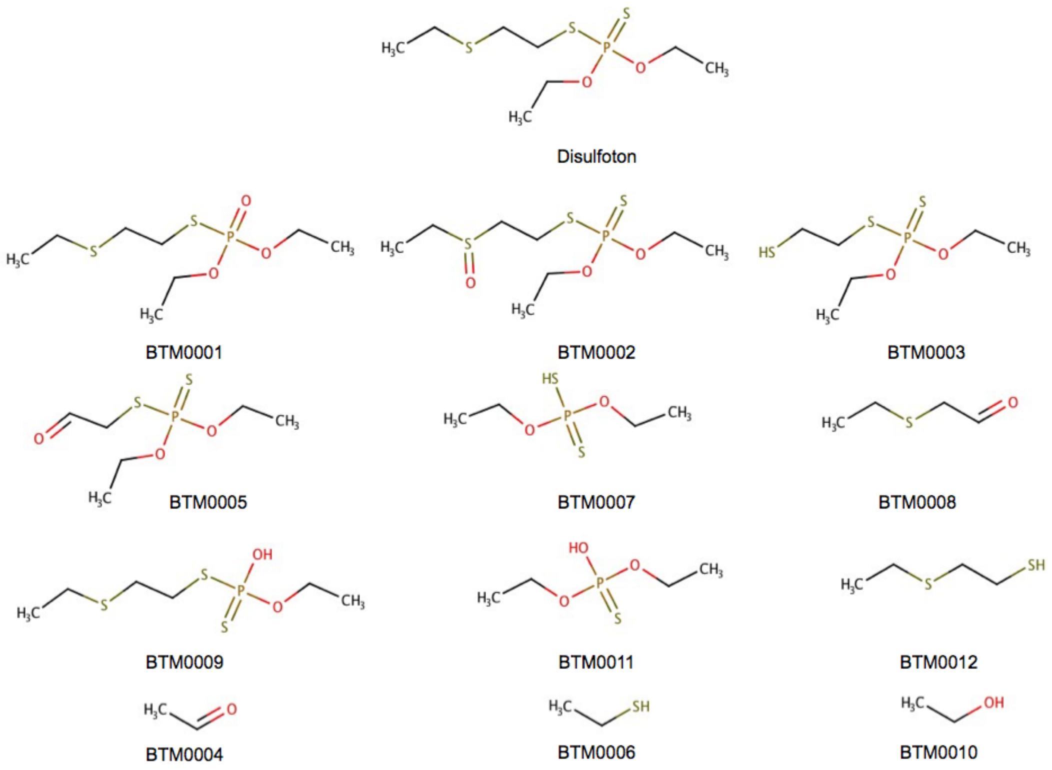


Figure 3.9: Environmental microbial metabolism of disulfoton, as predicted by *BioTransformer* and the EAWAG-BBD/PPS system. The metabolites BTM0004, BTM0006, and BTM0010 are reported by *BioTransformer* as by-products of the biotransformation bt0259 that generate BTM0003, BTM0005, and BTM0009. These by-products were not reported by the EAWAG-BBD/PPS system, as required by the applicable biotransformation rule.

3.4 Discussion

3.4.1 *BioTransformer*'s Structure and Implementation

BioTransformer is a software tool that uses a combination of the knowledge-based approach and the machine learning approach to predict the metabolism of small molecules. The knowledge-based system consists of a biotransformation database, a knowledgebase, and a reasoning engine. The biotransformation database is called BioTransformerDB. It is a unique resource as it is freely available and covers a wide range of enzymatic reactions that occurs in humans and mammals, as well as reactions that are catalyzed by the human gut microbial enzymes. In contrast to most publicly available databases, BioTransformerDB provides detailed biological and chemical information about the biotransformation, including the catalyzing enzymes, the substrates, the products, and the biotransformation rule(s) that is/are applied. BioTransformerDB describes the metabolism of >1,000 compounds catalyzed by ~15 enzyme families. For each biotransformation, at least one scientific source or reference is provided. BioTransformerDB is stored as JSON document, which can be easily parsed. An application of BioTransformerDB is the design of biotransformation rules with narrow specificity, which can be used for *in silico* metabolism prediction. In fact, this resource has been used to successfully design >300 biotransformation rules, which were used to annotate the biotransformations in the database and predict metabolites via the *BioTransformer* reasoning engine.

Despite the aforementioned strengths of BioTransformerDB, the database still has a number of limitations. Although it covers a large number of enzymatic reactions, it is clear that more data is needed in order to cover an even larger set of reactions (e.g.

oxidation reactions) catalyzed by enzymes other than CYP450s). It is also clear that there is a need to define more constraints and/or build more other models that would increase the quality of the predictions. Moreover, users could benefit from data about the different sites of metabolism for each specific biotransformation, as it would serve as a training set for the development of models for the prediction of sites of metabolism. For the current version of the database, the intent was simply to provide an easily readable and comprehensible data set. However, providing BioTransformerDB in a database format that can be parsed and queried in a more sophisticated way (e.g. SQL) would make the database much more useful to a broader number of users.

3.4.2 Evaluation of *BioTransformer*'s Predictions

BioTransformer was evaluated against *Meteor Nexus* for several randomly chosen xenobiotics using both single- and multi-step metabolic biotransformations. *Meteor Nexus* is a popular, commercially available software tool that is often considered to be the "gold standard" for predicting biotransformations. Based on the single step biotransformation test set, *BioTransformer* achieved a precision of 0.46 and a recall of 0.66. In comparison, *Meteor Nexus* achieved a precision of 0.45 and a recall of 0.55. Both tools were able to identify the four reported metabolites of omeprazole. However, they both missed three out of four metabolites of caffeine, suggesting that either new biotransformation rules should be added, or that some of the applied constraints ruled out caffeine as a substrate for the missed biotransformation rules. Although unwanted, such false negatives can be expected, as many biotransformation rules and constraints are often designed as a consensus and cannot always satisfy all molecules. This applies especially for the prediction of biotransformations catalyzed by enzymes with broad substrate specificity.

The prediction of multi-step metabolism was of particular interest, as it helped to assess the ability of *BioTransformer* to model the metabolism of molecules in mammals, from absorption to excretion. In this test *BioTransformer* displayed a significant advantage over *Meteor Nexus* in the prediction of epicatechin metabolites. Epicatechin is extensively metabolized both in the human gut (intestine, colon), and the liver. It undergoes reduction in the gut resulting in a number of metabolites that include dihydrochalcones, phenylvalerolactones, phenylavaleric acids, phenolic acids, as well as their conjugates (e.g. glucuronides, glycine conjugates, sulfates, etc.) (268). These compounds are further metabolized by the liver and/or recovered in the urine. While most of these metabolites were predicted by *BioTransformer* (see Figure 3.7 for some examples), they were all absent from the set of metabolites predicted by *Meteor Nexus*. As shown in Table 3.4, the prediction of multi-step metabolism can lead to higher rate of false predictions, and potentially, lower precision. As the molecule is activated (often by addition of reactive functional groups), the number of potential subsequent reactions can increase rapidly. For certain reactions with broad specificity, it is common to see several sites of metabolism within the same molecule, leading to a number of metabolic regiomers (structural isomers that differ in position of functional group). In the case of carvacrol, six regiomers were predicted by *BioTransformer* for the hydroxylation of inactivated carbons (aliphatic or aromatic), two of which have not been reported before (thymohydroquinone (BTM002), and 3,5-dihydroxycymene (BTM0005)) (see Figure 3.8). Of those six isomers, *Meteor Nexus* predicted three, which had all previously been reported. Theoretically, the prediction of two metabolites previously not reported illustrates the fact that *BioTransformer* needs to potentially define more/better patterns for the aromatic and aliphatic hydroxylation of carbons, and more preference rules (for its relative reasoning scoring system). On the other hand, *BioTransformer* could also be suggesting the structures and mechanistic details of several new metabolites for which LC-MS (or LC-MS^n) based identification could be attempted. In this regard, *BioTransformer* could be particularly useful as a hypothesis generator for new metabolite structures and new metabolic transformations. Overall, for the prediction of multi-step metabolism, *BioTransformer* achieved a precision of 47.3% and a recall of 83.7%, while *Meteor* achieved a precision of 18.9% and a recall of 28.3%.

In order to evaluate *BioTransformer*'s ability to predict environmental metabolism, we compared its prediction results with the EAWAG-BBD/PPS system. It is worth noting that the biotransformation and preference rules we encoded in *BioTransformer* were based on the same set of rules defined by the EAWAG-BBD/PPS. The key difference was that the rules were encoded in the same common SMIRKS/SMARTS format used by all of *BioTransformer's* other transformer tools. Based on the sample tests provided in the Results section, it is clear that *BioTransformer* was able to accurately replicate the predictions provided by the EAWAG-BBD/PPS system. These results suggest that *BioTransformer* could also be used to accurately predict environmental microbial metabolism.

We believe the examples used here nicely demonstrate the ability of *BioTransformer* to accurately predict a wide range of metabolic reactions, for a number of different types of small molecules (endogenous and xenobiotic compounds) and a number of different biosystems (humans, microbial/environmental). *BioTransformer* is unique in its ability to cover almost all aspects of secondary metabolism (drug/xenobiotic metabolism, endogenous compound metabolism, gut microbial metabolism, environmental metabolism). This makes it particularly useful for the wide-ranging applications seen in metabolomics. Furthermore, the accuracy, coverage, precision and recall of *BioTransformer* appears to be as good as, or better than some of the most highly regarded metabolic prediction systems now available. Additionally, *BioTransformer* is fast. It takes on average 3.95 seconds per query compound for single-step transformations using the EC-based metabolism biotransformer, and 9.36 seconds when using the super transformer. It is, easy to use, open-source and freely available. Certainly a more extensive analysis of a much larger set of query compounds would likely better illustrate the strengths and weaknesses of *BioTransformer*. However, it is important to remember that there are relatively few experimentally validated, comprehensive sets of metabolic "biotransformation trees" and that the examples selected here cover a good portion of the better known trees. Nevertheless, we are currently collaborating with the French National institute for Agricultural Research (INRA) on a project that aims at assessing the predictions of *BioTransformer* and *Meteor Nexus* for a much larger set of compounds including monoterpenes and a number of well-studied polyphenols.

While there are a number of strengths and advantages to *BioTransformer,* we believe that certain improvements could still be made to the program. First, the addition of more biotransformation data would certainly provide more reaction "fodder" to create more biotransformation rules. Additional biotransformation data would also provide statistical evidence to fine tune the reaction preference rules (relative reasoning) and occurrence ratios for absolute/relative reasoning. In particular, adding an option for absolute reasoning would give *BioTransformer* the ability to select candidates with a set cut-off score. Currently *BioTransformer*'s biotransformation database and its knowledgebase cover only a small portion of the gut microbial degradation (i.e. metabolism of plant-derived polyphenols). As gut metabolism plays a significant role in the secondary metabolism of humans, and many xenobiotics as well as endogenous compounds are known to be metabolized in the gut (269-272), it will be important to further expand the coverage of gut microbial metabolism in *BioTransformer*. We plan to make these improvements in the next version of *BioTransformer*. Over the longer term we are hoping to integrate more machine learning based prediction models (e.g. SoMs for CYP450 metabolism, and SoMs for phase II metabolism). This integration depends mostly on the amount of data available as machine learning depends on having large training sets to optimize its performance. Given that the number of experimentally confirmed biotransformations is still quite low for the systems of interest, it is likely that this will take a number of years to complete.

3.5 **Conclusion**

In this work, we have presented *BioTransformer*, a freely available software tool that supports the rapid, accurate, comprehensive prediction of secondary metabolism of small molecules in both mammals and in the environment. Within mammals, *BioTransformer* was able to accurately predict both single-step as well as multi-step biotransformations over a range of xenobiotics, including drugs, pesticides, and food compounds. The reactions that *BioTransformer* predicts cover phase I and phase II metabolism in mammals, as well as the human gut. Overall, *BioTransformer* was shown to achieve higher precision and recall, compared to *Meteor Nexus*, a commercial software tool for *in silico* metabolism prediction. In fact, *BioTransformer* proved to be significantly more accurate than *Meteor Nexus* for predicting the metabolism of polyphenols in the gut. Unlike most other metabolic prediction tools, *BioTransformer* also supports the prediction of metabolism of small molecules by environmental microbes. The integration of environmental metabolism with endogenous (liver/gut) metabolism allows *BioTransformer* to address many of the predictive metabolic needs of metabolomics researchers, which tend to span a much wider range than, say, drug researchers, food chemists or environmental scientists.

Despite its strengths, *BioTransformer* is not without some limitations. Addressing these would certainly make the program much more flexible, more accurate, and more comprehensive. Obvious improvements for the current version of *BioTransformer* include: 1) the validation of *BioTransformer*'s predictions for a larger and more diverse test set; 2) the experimental validation of *BioTransformer*'s predictions for a small set of monoterpenes and polyphenols; 3) the expansion of the knowledgebase to cover more reactions, and 4) the addition of new options for metabolite prediction/ranking.

Chapter 4

CFM-ID 3.0: Significantly Improved ESI-MS/MS Spectral Prediction Using a Hybrid *In Silico* Fragmentation Model with Metadata

4.1 Introduction

Liquid chromatography (LC) coupled to mass spectrometry (MS) or tandem mass spectrometry (MS/MS) has become one of the leading techniques for compound identification in organic chemistry, natural product chemistry, and metabolomics (273,274). In the field of metabolomics, LC-MS/MS is widely used to identify and quantify individual chemicals in complex biological or environmental mixtures. For untargeted metabolomics applications using LC-MS/MS, high performance or ultrahigh performance liquid chromatography (HPLC or UHPLC) is first performed to separate compounds in the sample and then electrospray ionization (ESI) mass spectrometry (MS and MS/MS) is used to collect the mass spectra of each chromatographic peak. In order to identify individual compounds, the resulting MS/MS spectra, along with the chromatographic retention time and parent ion masses of the compound of interest, are then (ideally) compared to the MS/MS spectra and retention time of authentic standards to confirm the compound's identity.

Because of the limited availability of many authentic chemical standards in most metabolomics labs, putative metabolite identification is more commonly performed. Putative identification is achieved by comparing the MS/MS spectra to experimentally collected reference spectra found in various MS/MS spectral databases. Key to the success of this putative identification process is the availability of a large, comprehensive database containing experimentally collected MS/MS spectra of pure compounds that covers a large portion of "chemical space". Unfortunately, publicly available databases of experimental MS/MS spectra currently cover a total of only ~20,000 unique compounds (275). Consequently, as reported in many large-scale metabolomic studies (47,276), the percentage of MS spectral features that can be confidently assigned to known compounds is often less than 2%. As a result, the compound identification step continues to be the central bottleneck in almost all untargeted MS-based metabolomic studies.

Given the cost of synthesizing or acquiring the 100,000's of chemicals needed to create the required experimental MS/MS spectral libraries, a growing number of scientists are turning to *in silico* metabolomics methods to facilitate compound identification. Over the last decade, a number of computational MS approaches have been developed for this purpose. Some of the more popular software tools use MS/MS fragmentation trees and spectral fingerprints (e.g. *CSI:FingerID* (180)) of an observed ESI-MS/MS spectrum and rank the likelihood that a given chemical structure could produce such a spectrum, or arrange substructures of a candidate molecule into a hierarchical tree that best explains the fragmentation pattern observed in a given experimental MS^n spectral tree (MAGMA (183)). Other tools, such as *MetFrag* (182) and *CFM-ID* (45,111,178) use *in silico* fragmentation of a given compound structure to predict ESI-MS/MS (for LC-MS) and EI-MS (for GC-MS) spectra. By matching the observed MS/MS spectrum to a library of predicted MS/MS spectra, it is possible to identify or rank which compound is being observed. Increasing the size of the library of *in silico* predicted spectra is expected to increase the likelihood of successfully identifying compounds from newly acquired MS/MS spectra (181).

The two main *in silico* fragmentation techniques are rule-based approaches and combinatorial approaches. Rule-based "fragmenters" use hand-made rules based on experimentally observed fragmentation patterns that are specific to one or more structural features or chemical classes. These rules are typically extracted from analyzing the scientific literature or, preferably, learned from in-house experimental data. *Mass Frontier* (173) is an example of a software tool that uses hand-made fragmentation rules. Once the rules are implemented, this approach can be very fast, consistent and accurate. However, a major disadvantage to this approach is that the design of fragmentation rules requires considerable expert curation. Furthermore, these rules cannot be applied to novel classes of molecules. For these reasons, much more emphasis has recently been put towards the implementation of combinatorial fragmentation approaches. Combinatorial fragmentation approaches iteratively cleave chemical bonds within a molecule in a combinatorial fashion, and use penalty scores that favour the cleavage events that are most likely to occur at each step. Examples of tools that implement combinatorial fragmentation include *CFM-ID* (111), *MetFrag* (182) and *FiD* (176).

CFM-ID is a publicly available software tool and web server that can be used for MS/MS spectral prediction, MS/MS spectrum peak assignment, as well as MS-based compound identification (45,111,178). It implements a technique known as Competitive Fragmentation Modeling (CFM), a probabilistic generative model using a customized cost function that takes into account the structural composition of a molecule to predict spectra resulting from electrospray (CFM-ESI) or MS/MS spectra. *CFM-ID* has been used to generate a reference MS/MS spectral library of over 30,000 known compounds from the HMDB (8) and KEGG (103) databases at 3 different collision energies (10 eV, 20 eV and 40 eV). For compound identification tasks, *CFM-ID* can use this spectral library to suggest candidate molecules that match input experimental MS/MS spectra. In 2015 *CFM-ID* was shown to outperform *FingerID* and an earlier version of *MetFrag* in various identification tasks from ESI-MS/MS spectra (178). However, subsequent tests and subsequent studies on the performance of *CFM-ID* have shown that a number of improvements could be made to the program and its spectral database.

For instance, one well-known limitation of *CFM-ID* is its very slow and relatively poor performance for predicting MS/MS spectra of lipids and other large “segmented” metabolites. This is primarily due to the length of the fatty acids or attached head-group segments, leading to a combinatorial explosion of the possible fragments at each step of the *in silico* fragmentation process. As demonstrated by Kind *et al*. (46) who developed LipidBlast, and Tsugawa *et al*. (277) who studied sphingolipid fragmentation, the use of structure-based fragmentation rules appears to be much better at handling lipids and other large segmented or modular molecules (such as carbohydrates) than combinatorial fragmentation. However, it is important to note that LipidBlast also has some limitations. For instance, it does not provide a well-defined set of fragmentation rules or algorithms that can be incorporated into other computational MS spectral prediction tools. Furthermore, while it does provide m/z values for fragment ions, LipidBlast does not provide structural data or structural annotations for the fragment ion peaks nor does it estimate peak intensity. These are the kinds of output that are typically found with most *in silico* fragmenters and these shortcomings have been addressed in this update to *CFM-ID*.

In addition to the incorporation of compound-specific fragmentation rules, it has also been shown that significant improvements in MS-based compound identification can be achieved by including metadata or other forms of external data in the spectral matching or scoring functions (182). In particular, the inclusion of citation frequency (the number of times a given compound is mentioned in the literature), along with the incorporation of experimentally collected MS/MS spectra in the reference spectral database can often improve compound identification performance by a factor of 2 or more (278). When taking into account the chemical similarity or the distribution of structural features or chemical classes (via *ClassyFire* (249)) among candidates, it is often possible to improve the performance even further (180). Based on these and other developments in the field of *in silico* metabolomics and *in silico* mass spectrometry, we have implemented a number of modifications to *CFM-ID* that have helped to: 1) achieve faster and more accurate prediction of MS-spectra for 26 classes of lipids, 2) expand *CFM-ID's* reference spectral library to include both experimental and predicted MS/MS spectra, 3) enhance *CFM-ID's* ability to incorporate metadata and chemical similarity, 4) improve *CFM-ID*'s compound identification rates, and 5) enhance *CFM-ID's* ability to predict the structural classification of compounds for query spectra that could not be matched *in CFM-ID's* spectral database. This improved version of *CFM-ID* is called *CFM-ID* 3.0. It is freely available as a web server at http://cfmid-staging.wishartlab.com. Its source code is also freely accessible at https://sourceforge.net/p/cfm-id/wiki/Home.

4.2 Methods

To improve *CFM-ID*'s overall performance for MS/MS analysis, we pursued several algorithmic and database enhancements. These included: 1) Encoding and validating rules for ESI-induced fragmentation of 26 classes of lipids; 2) Implementing an automated chemical classification schema (via *ClassyFire*) for both *CFM-ID*'s database and its query compounds; 3) Redesigning, significantly expanding and improving *CFM-ID*'s MS/MS spectral library (by including experimental MS/MS spectra and adding many thousands more predicted MS/MS spectra); 4) Collecting citation information on all of the compounds in *CFM-ID*’s MS/MS spectral library; and 5) Modifying *CFM-ID*’s scoring function to incorporate the above changes and improve its overall performance.

The encoding of the lipid rule-based fragmentation approaches was added to improve the speed and accuracy of *CFM-ID’s* lipid ESI-MS/MS predictions, as well as to cover a larger pool of experimental conditions as reflected by the different adduct types. The use of *ClassyFire*’s chemical classification method (249) was implemented to automate the rule-based/combinatorial-based decisions for *CFM-ID* and to improve *CFM-ID*’s ability to identify or re-rank potential MS/MS spectral matches based on structural similarity. The redesign and expansion of the *CFM-ID*’s spectral database was performed to accelerate search speeds, reduce the memory requirements and to grow the spectral database size (of both predicted and known MS/MS spectra) by a factor of 2, so as to improve the likelihood of user query spectral matches. The inclusion of citation data was intended to enhance the scoring accuracy of potential MS/MS spectral matches, while the modification of *CFM-ID*’s scoring function was intended to improve its overall performance. Details regarding how all of these changes were implemented are described below.

4.2.1 Encoding Lipid Fragmentation Rules

Our analysis of numerous databases and the literature indicated that there are 26 major classes of lipids for which MS/MS spectra are best predicted using hand-made fragmentation rules. The encoding of these hand-made lipid fragmentation rules involved several steps including: 1) experimentally measuring or compiling (via literature) characteristic MS/MS fragment ions observed at each of three collision energy levels (10 eV, 20 eV, and 40 eV) for each lipid class, 2) determining the relative abundance of each fragment ion at each energy level, 3) accurately determining the chemical structure and m/z values of each of the fragment ions, 4) including more MS/MS experimental conditions (and adduct ions) by expanding the list of adduct types covered by previous versions of *CFM-ID*, and 5) implementing these rules using standardized cheminformatics languages (SMILES (197), SMARTS (68) and SMIRKS (69)) in order to rapidly and accurately predict and annotate ESI-MS/MS spectra for lipids.

4.2.1.1 Acquisition of Reference Lipid MS/MS Spectra

The generation of the lipid fragmentation rules required the acquisition of experimental ESI-MS/MS spectra for a number of lipids and lipid classes. The acquired spectra were collected at several collision energies, for various adduct types (e.g. [M+H]+, [M-H]-), and, if possible, from various MS instruments. This was used to help capture fluctuations or biases that can be introduced by the different parameters. A total of 533 experimental MS/MS spectra were collected for 16 standard lipids (purchased from Avanti Polar Lipids Alabaster, AL) from 15 lipid classes at various collision energies (10 eV to 60 eV), in both positive and negative mode using an AB Sciex QTrap 4000 MS instrument. For each lipid standard, an enhanced MS (EMS) scan was first collected to identify precursor ions with high abundance in either ionization mode. Enhanced product ion (EPI) scans were then collected for each precursor ion to generate the MS/MS spectra with different collision energy levels ranging from 10 to 60 eV. In addition to the MS/MS spectra collected in our laboratory, published lipid MS/MS spectral data were compiled from the LIPID MAPS (198) and the MoNA (40) databases. For the LIPID MAPS spectra, only annotated spectral images were available. Therefore, MS/MS peak lists were generated by annotating the peaks using a semi-automated approach. This approach consisted of computing the relative abundance of each peak, and manually mapping it to the m/z list provided in the LIPID MAPS spectrum. In addition to the experimental spectra, the LipidBlast and FAHFA (46,279) libraries, as well as MassBank (109), mzCloud (280) and the sphingolipid library of Tsugawa (277) served as references that provided additional information for lipid classes not covered by our experiments. In total, 844 lipid MS/MS spectra from 26 lipid classes were collected and analyzed.

4.2.1.2 Annotation of Reference Lipid MS/MS Spectra

With the lipid MS/MS spectra in hand, we proceeded to manually annotate each spectrum. This consisted of assigning each fragment ion peak to a specific structure and a specific reaction or fragmentation event (e.g. the loss of a water molecule from a [M+H]+ precursor ion, the loss of a side chain, or the presence of a specific fragment). The annotation of spectra was limited to the in-house generated MS/MS spectra and the LIPID MAPS set, as both were measured with the same model of instrument (AB Sciex QTrap 4000). The annotation process was largely guided by the information provided in LIPID MAPS, LipidBlast and other scientific reports (46,198,281,282). In a number of cases, the same compound had MS/MS spectra in at least two of the data sets (including the LipidBlast database), and the corresponding spectra were available for the same adducts or ions. In these cases, we annotated the spectra by direct comparison of the peak lists. Among the 26 lipid classes, 11 were not covered by our in-house experimental data. For this reason, the MS/MS spectra of these missing lipid classes were extracted from the LIPID MAPS (experimental) and/or LipidBlast (*in silico*) library. Since the experimental and theoretical spectra acquired from other sources (LipidBlast, LIPID MAPS) did not always cover all three collision energy levels (10 eV, 20 eV, and 40 eV), the generation of consensus fragmentation patterns was done by comparing standards with the corresponding acquired experimental spectra. This was further validated by mining the scientific literature. Once the energy-specific fragmentation patterns were generated, the relative abundance of each peak was assigned to one of four intensity levels: low, medium, high, or maximum abundance level. The assigned intensity was based on observed relative abundances from our experimental spectra. The maximum level of abundance was assigned to the base peak, typically when no fragmentation was observed (usually at a low collision energy). Additional feedback from local MS experts combined with an extensive review of the lipid MS/MS literature helped to complete the spectral annotation process. This effort led to the near-complete annotation of all observed fragment ions, their precise m/z values and the corresponding fragmentation reactions for a total of 767 peaks from 26 lipid classes at each of 3 collision energies (10 eV, 20 eV and 40 eV).

4.2.1.3 Implementation of the Lipid Fragmentation Rules

The annotated fragment ions along with their structures and reactions provided the basis for the creation of fragmentation rules. All of the fragmentation rules were implemented in the Java programming language through a new “lipid fragmenter module” in *CFM-ID*. The structural backbone of each lipid or lipid fragment class was represented using the Daylight SMARTS language (68). This is a module implemented in *ClassyFire,* a software tool for automated structure-based hierarchical annotation of chemicals (249). To accelerate the lipid classification process, a sub-ontology from the ChemOnt (249) ontology was used. For each lipid or lipid fragment class, one set of fragmentation patterns is encoded for each of the applicable adducts as chemical reactions. The chemical reactions are represented using the Daylight SMIRKS language (69). Additionally, a number of transformation rules were encoded to standardize the structures of all the query compounds. The standardization of the fragmentation reactions using well-developed cheminformatics languages ensures that the structural representations are consistent for all query compounds, structural classes and chemical reactions. Without adhering to these standards many chemicals classes could be misidentified or invalid fragments could be returned.

The new *CFM-ID* lipid fragmenter program has been fully integrated into the existing spectral prediction workflow of the previous version of *CFM-ID* (45). In *CFM-ID 3.0*, the lipid MS/MS prediction tasks requires a lipid structure (submitted as a SMILES string or SDF file) and an adduct or an ion as input. Upon submission, the compound is classified based on its structure via *ClassyFire*. If the compound is identified *by ClassyFire* as a lipid molecule belonging to any of the covered classes, and fragmentation patterns applicable to the selected adduct exist in the lipid fragmentation library, the compound is fragmented accordingly. The fragmentation operation is executed using the AMBIT library (262). After the *in silico* fragmentation step is completed, the relative abundance of each peak is assigned (using the fragmentation rules described above), and three ESI-MS/MS spectra are generated (at 10 eV, 20 eV, and 40 eV). If no set of fragmentation patterns is applicable to the compound and/or the selected adduct, then the ESI-MS/MS spectra are predicted using the original CFM algorithm as implemented in *CFM-ID* 2.0. The resulting ESI-MS/MS spectra are then returned with each peak annotated by its m/z value, its relative abundance, and the chemical structure of the corresponding fragment encoded in a standard SMILES format. Additionally, any available experimental MS spectra in the *CFM-ID* spectral database matching the query compound are also displayed in the results alongside the predicted spectra.

4.2.2 Integration of Chemical Classification

Similar structures tend to undergo similar fragmentation patterns under the same conditions. For this reason, a number of *in silico* MS fragmentation algorithms now take the chemical structure of query molecules into consideration for improved MS-spectra prediction and compound identification tasks. For the prediction of EI-MS/MS spectra, CFM's scoring function partly relies on a list that describes the presence or absence of 107 functional groups and 86 fragment descriptors. These groups and fragment descriptors are provided by *ClassyFire* (249) and *RDKit* (65,91), respectively. Other computational tools such as *CSI:FingerID* (180) rely on models that can predict the presence of functional groups and fragments based on a given compound or a given MS-spectrum. For this reason, it might be expected that in compound identification tasks, the highest ranked candidates would likely share a significant number of functional groups or possibly share a maximum common substructure. This information would be particularly helpful in cases where it is very difficult to discriminate between the highest ranked candidates. More specifically, the presence of one or more common structural backbones (e.g. diterpene, ceramide, phosphatidylglycerol) could significantly impact the ranking, when very structurally similar candidates are prioritized among those that have a high spectral similarity to the query compound.

Therefore, a chemical classification was stored for each compound in the database. The chemical classification was computed by *ClassyFire* and retrieved using the *ClassyFire* API (249). As will be described later in this section, the chemical class assigned to candidate molecules was taken into account along with other metadata to improve the original CFM scoring method (dot product or Jaccard score). In addition to the adjustment of the scoring function, chemical classification was also used to predict the chemical class(es) to which the query compound belonged. Formally, the predicted chemical class corresponds to the direct parent of the highest ranked candidate. In case of a tie, the predicted chemical category is the most frequently occurring direct or alternative parent among all candidates that has the highest score.

4.2.3 Collection of Compound Citations

Several studies have demonstrated that the integration of metadata can significantly improve compound identification rates with spectral library searches (180,182,278). In particular, the frequency with which a compound is mentioned in the literature could serve as a proxy for the likelihood that the compound is either sufficiently abundant or sufficiently ionisable for detection via MS/MS methods. Therefore, every compound in the *CFM-ID* spectral library was assigned a citation score. An initial set of citation counts was obtained using *DataWrangler*. *DataWrangler* is an in-house tool that automatically mines PubChem (9), HMDB(8), ChemSpider (283), and ChEBI (200), and returns a unique list of scientific reference citations for a given compound. A second set containing PubMed citation counts (without PubMed IDs) was obtained by mining the EPA's CompTox dashboard (284). This set was computed and provided to us by the CompTox dashboard's development team. The two sets were merged by comparing each compound's InChI keys. More specifically, when a compound had a citation count in only one set, the corresponding citation count was assigned to that compound. For compounds that had citation counts both from *DataWrangler* and CompTox, the largest count was assigned, as it was expected that both counts could include many of the same citations. A total of 17,000 compounds were assigned a citation count of 1 or more. For the remaining compounds, *DataWrangler* assigned a custom citation count of 1, if and only if, they were found in at least one of the following databases: HMDB (8), DrugBank (99), T3DB (100), ContaminantDB (244), FooDB (224), ECMDB (12), YMDB (285), and PhytoHub (243). It is also important to note that *CFM-ID's* compound library includes more than just "pure" metabolites that are used to count citations and generate ESI-MS/MS spectra. In particular, *CFM-ID's* library also contains ~76,000 compounds that were computationally derivatized with TMS (for GC-MS spectral analysis) from known HMDB compounds. Each of these derivatized compounds has had their EI-MS spectrum generated by *CFM-ID* and each was assigned the citation count of its parent (derivative-free) molecule.

4.2.4 Redesigning and Expanding of *CFM-ID*'s Spectral Library

The original reference spectral library in *CFM-ID* 2.0 contained 166,543 unique computationally generated ESI-MS/MS and EI-MS spectra for ~118,000 compounds (including TMS derivatives) from the HMDB and KEGG databases. ESI-MS/MS spectra were computed in positive ([M+H]+) and negative ([M-H]-) ionization modes, one for each of three collision energies (10 eV, 20 eV, and 40 eV). The EI-MS (for GC-MS studies) spectra were predicted at a collision energy of 70 eV. EI-MS spectra were also computed for 64,390 TMS derivatives of HMDB compounds. In order to significantly improve identification rates, the new *CFM-ID* library was updated as described below.

4.2.4.1 Collection of Experimental MS/MS Spectra from External Sources

While the accuracy of computationally predicted MS spectra is often quite good, the accuracy of experimentally collected MS spectra is much better. Therefore the inclusion of experimentally determined EI-MS and ESI-MS/MS spectra would be expected to improve the match scores for query spectra/compounds that have previously been analyzed by EI-MS or ESI-MS/MS. Experimentally determined ESI-MS/MS and EI-MS spectra were downloaded from the MassBank of North America's (MoNA) online repository (40). As of February 2017, MoNA contained 14,847 EI-MS spectra for 9,242 compounds, and 51,135 LC-MS/MS spectra for 10,538 compounds. The spectra and compounds in MoNA originate from several databases, including the HMDB database (8), MassBank (109), the GNPS database (286), and the ReSpect database (287), among others. Only experimental spectra were collected from MoNA, except the set from HMDB. An additional 915 ESI-MS/MS spectra were manually regenerated for 523 compounds from information contained in the NIST 14 database. Since *CFM-ID* uses models trained on MS spectral sets utilizing specific collision energy and mass accuracy criteria, the HMDB, MoNA, and NIST spectra were further filtered to match these criteria. Specifically, experimental MS spectra were required to have a known ionization type, a known compound neutral mass, and to have been analyzed with high-resolution MS instruments (e.g. Q-TOF instruments) in the case of LC-MS spectra. Moreover, EI-MS spectra obtained from high-resolution MS spectra were also selected/filtered. The complete library of experimental spectra from HMDB was obtained from our in-house repository, and filtered. Upon filtering, it contained 1,492 unique EI-MS spectra for 647 unique compounds, and 1,152 unique ESI-MS/MS spectra for 239 unique compounds.

Moreover, there were 54,529 usable experimental MS spectra remaining. These experimental MS spectra were converted into the peak list format required for *CFM-ID* and uploaded into *CFM-ID*'s online spectral library.

4.2.4.2 Compilation of Predicted ESI-MS/MS and EI-MS Spectra

As noted earlier, the original *CFM-ID 2.0* database contained 102,153 unique computationally generated ESI-MS/MS spectra (from 51,635 compounds) and 64,390 unique computationally generated EI-MS spectra (from 64,390 TMS derivatized compounds). Among the 102,153 ESI-MS/MS spectra, 36,746 were previously computed for 18,373 unique compounds belonging to the 26 lipid classes covered by the rule-based fragmenter, and transferred to the *CFM-ID 3.0* database. The remaining 65,407 mass spectra computed by *CFM-ID 2.0* were also moved to the *CFM-ID 3.0* database. In total, ~36,900 spectra were generated for the 18,438 lipids. To this database, another ~145,460 ESI-MS/MS spectra were computed for 80,000 lipids and 7288 other metabolites obtained from recently updated versions of HMDB, DrugBank and PhytoHub. Those compounds were added to the *CFM-ID 3.0* database. These predicted ESI-MS/MS spectra were generated for both positive and negative ion-mode as well as at three different collision energies (10 eV, 20 eV, and 40 eV). Likewise, another 7,288 computationally generated EI-MS spectra were added from 7,288 other derivatized metabolites taken from recently updated versions of HMDB, DrugBank and PhytoHub. In total, the *CFM-ID 3.0* database now contains 247,767 computationally generated ESI-MS/MS spectra (from 135,506 compounds) and 71,678 computationally generated EI-MS spectra (from 71,678) compounds. If the experimental ESI-MS/MS and EI-MS spectra are added to this total, the *CFM-ID 3.0* spectral database now contains a grand total of 289,170 ESI-MS/MS spectra and 86,464 EI-MS spectra.

4.2.5 Modifying *CFM-ID*'s Scoring Function and Ranking Schema

The results of the Critical Assessment of Small Molecular Identification (CASMI) 2016 contest showed that the integration of additional data (i.e. citation frequency of compounds and structure similarity) into the original scoring function for *CFM-ID* improved compound identification rates (278). This trend was also observed for several other tools during the contest in separate studies (182,278). To create a combined score, the original spectral similarity score computed by *CFM-ID* (Jaccard or Dot Product, according to the user specification) was combined with a citation score and a chemical classification score. As described earlier, the citation score is based on the number of citations that a given compound has in the scientific literature. More highly cited compounds are typically those that are more commonly detected, studied or used. Therefore the citation score serves as a proxy of the general abundance or concentration of a compound and is intended to favour more abundant compounds over extremely rare or trace level compounds.

As noted earlier, the chemical classification score is based on the number of chemical categories to which a compound is assigned (by *ClassyFire*), relative to the total pool of chemical classes assigned to all candidate molecules. The chemical classification score was added to help re-rank or cluster structurally similar molecules (and MS spectra) closer together. Each of the three scores was normalized by dividing its computed score by the maximum score across the candidate list. The general formula for the total candidate score is:

$$S_{TOTAL}(C) = a_{CFM\_ORIG} * S_{CFM\_ORIG}(C) + a_{CLASS} * S_{CLASS}(C) + a_{REF} * S_{REF}(C)$$

where $S_{TOTAL}(C)$, $S_{CFM\_ORIG}(C)$, $S_{CLASS}(C)$, and $S_{REF}(C)$ are the total score, the normalized spectral matching *CFM-ID* score, the normalized *ClassyFire* score, and the normalized reference score for candidate C, respectively. Each of the three scores are weighted by the coefficients a_{CFM_ORIG}, a_{CLASS}. and a_{REF}, respectively, where:

$$a_{CFM\_ORIG}, a_{CLASS}, a_{REF} \geq 0$$

and

$$a_{CFM\_ORIG} + a_{CLASS} + a_{REF} = 1$$

This approach was used to build two scoring functions for metabolite identification, one for the ESI-MS/MS input and one for EI-MS input. The optimal set of coefficients was determined through a grid search using a manually selected set of 1,000 spectral/compound identification tasks (for 1,000 unique compounds ranging from drugs to lipids). Each of the selected molecules had one or more experimental spectra at one of three level energies (10 eV, 20 eV, and 40 eV), in addition to predicted ESI-MS/MS and EI-MS spectra. The data set was divided into five equally sized subsets. Several models (with a unique combination of coefficients) were trained on 800 compounds (4/5 of the data set) and tested on the remaining 200 (1/5 of the data set). This process was repeated four more times, using a different test set of 200 compounds for each iteration. Experimental spectra were used as input for each identification test, and upon testing, only the best model was selected. A consensus model was built based on the five selected models, and further tested using a smaller test set. The final coefficient values for the ESI-MS/MS scoring function were a_{CFM_ORIG}=0.6, a_{CLASS}=0.1, and a_{REF}=0.3. The final coefficient values for the EI-MS scoring function were a$_{CFM\_ORIG}$=0.8, a$_{CLASS}$=0.1, and a$_{REF}$=0.1.

4.2.6 Performance Testing

Three types of performance tests were conducted. The first assessed the performance of the lipid ESI-MS/MS spectral prediction method; the second assessed the performance of the new scoring function in exact compound identification and the third assessed *CFM-ID*'s performance in identifying a compound's correct chemical class. To test the lipid ESI-MS/MS spectral prediction method, a benchmark analysis was performed on 20 randomly chosen lipids from the 26 known lipid classes for which fragmentation rules were derived. The computation was performed on a 2.7 GHz Intel Core i5 MacOSX with 16 GB (1867 MHz DDR3) of memory. A total of 120 ESI-MS/MS spectral predictions were generated for both *CFM-ID* 2.0 and *CFM-ID 3.0* at 3 different energies and 2 different ionization modes with various adduct types. The average execution time was determined for each spectral prediction. In addition to the execution time comparison, an additional performance comparison was conducted to assess the quality of the predicted MS/MS spectra. For this task, a set of 10 experimental ESI-MS/MS spectra measured in positive ion mode, and 10 experimental ESI-MS/MS spectra measured in negative ion mode were selected. The selected spectra were measured under conditions that can be simulated by *CFM-ID 3.0*'s lipid fragmentation rules (same energy levels, same adducts). For each experimental MS/MS spectrum, *CFM-ID 2.0* and *CFM-ID 3.0* were used to predict a corresponding MS/MS spectrum under the same conditions. The performance was assessed by measuring the average pairwise spectral similarity between experimental and predicted spectra using a standard dot product score as implemented in the

OrgMassSpecR package (288). Moreover, they were also compared to LipidBlast, as the selected lipids and corresponding predicted spectra were also contained in the LipidBlast library.

In order to evaluate the performance of *CFM-ID*'s 3.0 new scoring functions, we built a data set for each function. The ESI-MS/MS scoring function was tested on a set of 208 experimental ESI-MS/MS spectra (for 185 unique compounds) generated on a Q Exactive Plus Orbitrap (Thermo Scientific), and used for the CASMI 2016 contest (Category 3) (278). These spectra were used as input for compound identification. 112 of the 185 compounds were included in the database and had at least one experimental ESI-MS/MS spectrum in addition to the precomputed ones. For each of the remaining compounds, ESI-MS/MS spectra were predicted using *CFM-ID* and stored in the spectral library. The EI-MS scoring function was tested on a set of 200 experimental EI-MS spectra (for 200 unique compounds) that were collected as described section 4.2.4.1. For each compound, EI-MS spectra were predicted and stored in the database. For each compound identification task, just one experimental spectrum was used as input. For each test, we used *CFM-ID* 2.0 and *CFM-ID* 3.0 scoring functions, separately, to attempt to identify the query compounds.

For the third kind of assessment, *CFM-ID 3.0* was assessed with regard to its performance in chemical class prediction/identification. This particular performance assessment was included because in many practical cases in MS-based metabolomics or MS-based natural product identification, it may not be possible to identify the exact compound via MS/MS spectral matching. Therefore, the ability to use MS/MS spectra to reduce the candidate list and to predict the correct chemical class or chemical family for a given query spectrum or compound can be very valuable. In assessing the performance of *CFM-ID*'s chemical class prediction the query compound was predicted to belong to the "direct parent" class of the highest-ranked candidate. In cases of a tie, the chemical class was predicted to be the most frequently occurring among all the direct and alternative parents among all the compounds with the highest score.

4.3 Results

4.3.1 Encoding Lipid Fragmentation Rules

Our manual analysis of the experimentally acquired lipid spectra provided a basis for the generation of 378 unique fragmentation rules covering 26 lipid classes and seven adducts, for a total of 55 combinations of chemical classes and adduct types. For each lipid class, an ESI-MS/MS spectrum can be simulated by *CFM-ID 3.0* at collision energies of 10 eV, 20 eV, and 40 eV. In general, almost all ESI-MS/MS spectra of lipids show similar fragmentation patterns with characteristic losses of the polar head group, and the acyl or alkyl chains, with relatively little fragmentation within the acyl or alkyl chains. For example, in choline-containing glycerophospholipids the most commonly observed fragments include phosphocholine ($C_5H_{14}NO_4P+$ ion; neutral mass = 184.07 Da), and the cyclic 1,2-cyclic phosphate diester ($C_2H_6O_4P+$ ion; neutral mass =123.99 Da). Figure 4.1 illustrates consensus fragmentation patterns for phosphatidylcholines from their [M+H]+ precursor ions. The numbers of rules for each lipid class and the number of covered adduct types per lipid class are shown in Table 4.1.

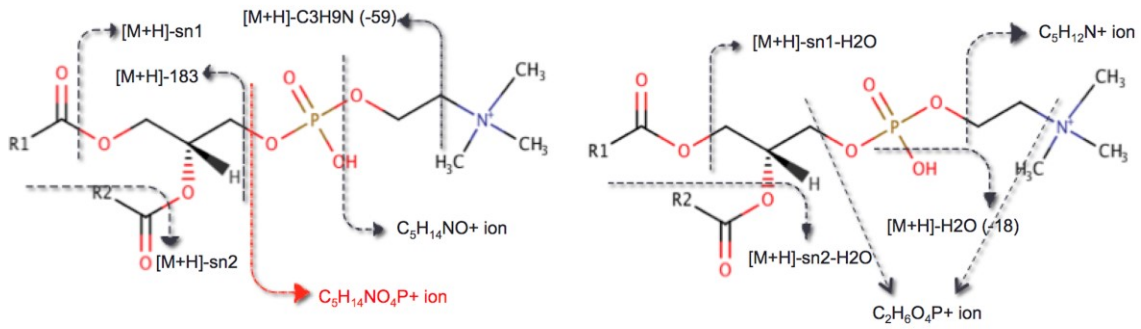


Figure 4.1 Fragmentation patterns of phosphatidylcholines from their [M+H]+ precursor ions. Only the precursor ion is the observed at each of the three energy levels. The ion fragment $C_5H_{14}NO_4P$+ (red arrow) corresponding to phosphocholine is observed at 20 eV and 40 eV, and the remaining fragments were observed only at 40 eV.

Table 4.1 Number of fragmentation rules and adduct types covered for each chemical category.

Lipid Class	**No. of covered rules**	**No. of covered adduct types**
1-monoacylglycerols	8	2
2-monoacylglycerols	13	3
1,2-diacylglycerols	10	2
Triacylglycerols	13	3
Phosphatidic acids	21	3
Phosphatidylcholines	42	4
Phosphatidylethanolamines	24	3
Lysophosphatidylcholines	29	4
Lysophosphatidic acids	12	2
Phosphatidylserines	28	3
Ceramides	16	3
Sphingomyelins	13	3

Cardiolipins	13	1
Acyl carnitines	12	2
1-alkylglycerophosphates	4	1
Phosphatidylglycerols	10	1
Lysophosphatidylglycerols	7	1
Plasmanyl-PC	17	2
Plasmenyl-PC	17	2
1-alkanylglycerophosphocholines	16	3
1-alkenylglycerophosphocholines	14	2
Phosphatidylinositols	12	1
Lysophpshatidylinositols	9	1
Plasmanyl-PE	4	1
Plasmenyl-PE	8	1
Fatty acids of hydroxylated fatty acids	6	1
Total	**378**	**55**

4.3.2 The New *CFM-ID* 3.0 Spectral Library

The original *CFM-ID 2.0* spectral library contained 102,153 unique computationally generated ESI-MS/MS spectra (from 51,635 compounds), and the 64,390 unique computationally generated EI-MS spectra (from 64,390 TMS derivatized compounds), for a total of 117,905 compounds. Because of improvements in the spectral prediction performance, additions of new compounds and the addition of new (experimental spectra) the new *CFM-ID 3.0* spectral library has been able to be expanded by a factor of 2.2 over the original *CFM-ID 2.0* spectral library (as of August 12, 2017). In particular the new library now contains a total of 373,974 ESI-MS/MS and EI-MS experimental spectra for 218,689 compounds, collected from various repositories. Previously, the *CFM-ID 2.0* spectral library had no experimental ESI-MS/MS or EI-MS/MS spectra. The compounds with experimental spectra are structurally and functionally diverse, and originate from various databases/libraries including HMDB (human metabolites) (8), DrugBank (drugs and drug metabolite) (99), KEGG (metabolites and drugs) (253), PhytoHub (dietary phytochemicals and their metabolites) (243), GNPS (natural products) (286), LipidBlast (detected and theoretical lipids) (46), and the FAHFA library (detected and theoretical lipids) (279). In addition to the experimentally collected spectra, *CFM-ID 3.0*'s predicted spectral library contains 86,464 EI-MS and 289,170 ESI-MS/MS spectra (from 218,689 compounds), all of which were computed with *CFM-ID 3.0*. Each of the 218,689 compounds in the new spectral library was assigned a citation score that is used in compound identification tasks. Among the 218,689 compounds, 93,871 had a citation count of 1 or more.

In our effort to improve the identification rates, a full chemical classification was computed for all 218,689 unique compounds, using *ClassyFire* (249). An average of ~26 chemical categories were assigned per compound. The chemical classification was used to adjust *CFM-ID*'s original scoring system, so as to take into account the chemical composition and chemical similarity among candidate molecules. It also served as basis to predict the chemical classification of the compound corresponding to the query spectrum in identification tasks.

Table 4.2 Statistics for the *CFM-ID* 3.0 spectral database. (*) Only compounds with >=1 citation were counted. (*) the total number includes the TMS derivatives from HMDB compounds.

Feature	**Value**
Total no. of unique compounds	218,689
No. of unique compounds in *CFM-ID* 2.0	117,905
Total no. of unique MS spectra	373,974
Total number of unique MS spectra in *CFM-ID 2.0*	166,543
Total no. of unique experimental MS spectra	54,529
Total no. of unique predicted MS spectra	319,445
No. of compounds with >=1 exp. MS/MS spectra	17,582
Total number of unique EI-MS spectra	86,464
Total no. of experimental EI-MS spectra	14786
Total no. of predicted EI-MS spectra	71,678
No. of compounds with >=1 exp. EI-MS spectra	8,963
No. of compounds with >=1 pred. EI-MS/MS spectra	71,678
Total number of unique ESI-MS/MS spectra	287,510
Total no. of experimental ESI-MS/MS spectra	39,743
Total no. of predicted ESI-MS/MS spectra	247,767
No. of compounds with >=1 exp. ESI-MS/MS spectra	9,422
No. of compounds with >=1 pred. ESI-MS/MS spectra	135,506
No. of compounds with >= 1 citations	93,871
Avg. no. of citations/compound*	315
No. of compounds with chemical classification assignments	218,689
Avg. no. of chemical category assignments/compound	26

4.3.3 Performance Testing

4.3.3.1 Lipid ESI-MS/MS Spectral Prediction

Two tests were performed to assess the lipid spectral prediction performance. One was for speed while the other was for accuracy. In terms of speed, *CFM-ID 3.0* averaged 0.395 +/- 0.03 seconds of computation time to predict each of the 120 lipid ESI-MS/MS spectra while *CFM-ID 2.0* averaged 68.58 +/- 0.21 seconds for the same task. This represents a speed-up of 173.6X. Clearly the rule-based approach for lipid analysis in *CFM-ID 3.0* is significantly faster than the combinatorial approach in *CFM-ID 2.0*. For most other kinds of molecules, the average processing time for *CFM-ID* is about 23.75 +/- 0.2 seconds. Clearly the computational slow-down for lipid spectral calculation (due to the many potential fragmentation combinations) is quite significant, which largely motivated us to develop a faster rule-based approach.

In terms of spectral prediction performance, the average spectral similarity score between the experimental lipid ESI-MS/MS spectra (collected on a QTOF) and the *CFM-ID* 3.0 predicted ESI-MS/MS spectra was 0.92 +/- 0.02. On the other hand, the average spectral similarity score between the *CFM-ID 2.0* predicted ESI-MS/MS spectra and the experimental ESI-MS/MS spectra was 0.07 +/- 0.04. This suggests that the accuracy of *CFM-ID 3.0* for lipid spectral prediction is 13X better than *CFM-ID 2.0*, which is highly significant. It is worth mentioning that *CFM-ID* predicts ESI-MS/MS spectra at three different collision energies while other programs, such as LipidBlast generate a consensus MS/MS spectrum that essentially merges the MS/MS spectra over all 3 energies. Therefore, during our comparative analysis, only one LipidBlast-generated consensus ESI-MS/MS spectrum was used for each unique compound, and compared against the experimental spectrum, independent of the energy level. Figure 2 shows head-to-tail-plots comparing the experimental ESI-MS/MS spectrum of dipalmitoyl phosphatidylcholine (PC(16:0/16:0)) collected at 40 eV collision energy with the corresponding *in silico* spectra predicted with *CFM-ID 2.0* – Figure 4.2a (178), *CFM-ID 3.0* – Figure 4.2b, and LipidBlast – Figure 4.2c (46), respectively. The experimental spectrum was measured in positive ion mode ([M+H])+, with a collision energy of 40 eV. The spectral similarity between the *CFM-ID 2.0* generated spectrum and the experimental ESI-MS/MS spectrum was 0.07, with *CFM-ID 2.0* being able to predict only two fragments that were observed in the experimental spectrum (namely the $C_5H_{12}N$+ and $C_5H_{14}NO_4P$+ ion fragments). For this particular example, *CFM-ID 2.0* predicted 31 fragments (Figure 4.2a) while *CFM-ID 3.0* predicted 10 fragments (Figure 4.2b), seven of which were observed in the experimental ESI-MS/MS spectrum. It is worth noting that the remaining three fragments result from fragmentations that were observed in experimentally measured ESI-MS/MS spectra of phosphatidylcholines obtained for [M+H]+ adducts at 40 eV. For this example the spectral similarity score was 0.98 when comparing the experimental ESI-MS/MS spectrum with the *CFM-ID 3.0*-predicted spectrum, and surprisingly, only 0.13 when comparing with the LipidBlast-predicted ESI-MS/MS spectrum. Figure 4.3 shows comparisons between experimental and predicted ESI-MS/MS spectra for 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (PS(16:0/18:1(9Z))) in the negative ([M-H]-) ion mode at a collision energy of 40 eV. The measured spectral similarity scores between the experimental and the *in silico* generated spectra are 0.10, 0.92, and 0.91 with *CFM-ID 2.0* (Figure 4.3a), *CFM-ID 3.0* (Figure 4.3b) and LipidBlast (Figure 4.3c), respectively.

As highlighted in Table 4.3, *CFM-ID 3.0* vastly outperforms *CFM-ID 2.0* in terms of lipid spectral prediction performance (average score of 0.92 vs. 0.07) and *CFM-ID 3.0* generally outperforms LipidBlast (average score of 0.92 vs. 0.88). Another important advantage of *CFM-ID 3.0* over LipidBlast is the fact that it generates spectral predictions for multiple collision energies (10, 20 and 40 eV) whereas LipidBlast only provides a single spectrum at an unknown collision energy. Furthermore, all spectral predictions generated by *CFM-ID 3.0* include information about not only the m/z values and their relative intensities but also the structure of the actual fragments (expressed as InChI and SMILES strings) for every predicted peak. LipidBlast only provides the m/z values and intensities.

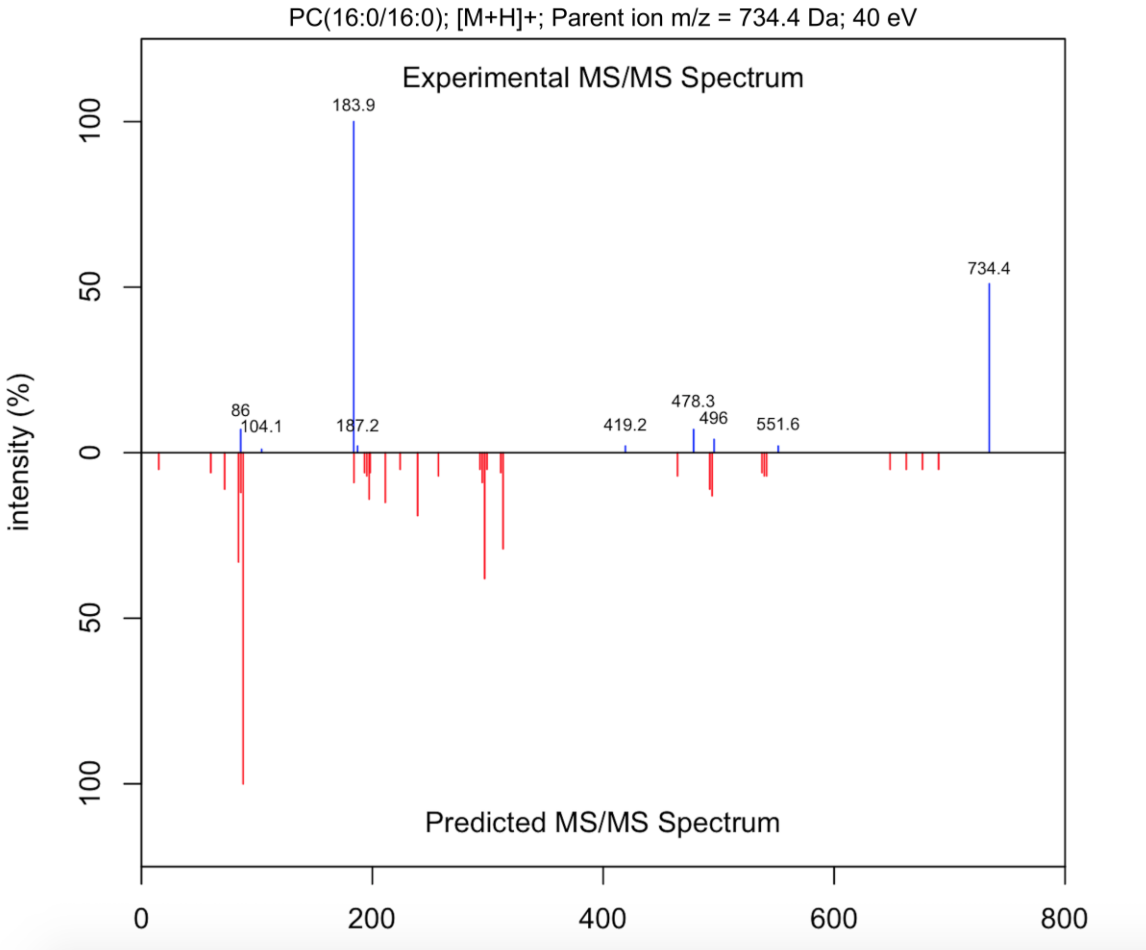


Figure 4.2a Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of dipalmitoyl phosphatidylcholine (PC(16:0/16:0)) measured at 40 eV, and the matching ESI-MS/MS spectrum predicted by *CFM-ID* 2.0. The computed spectral similarity is 0.07.

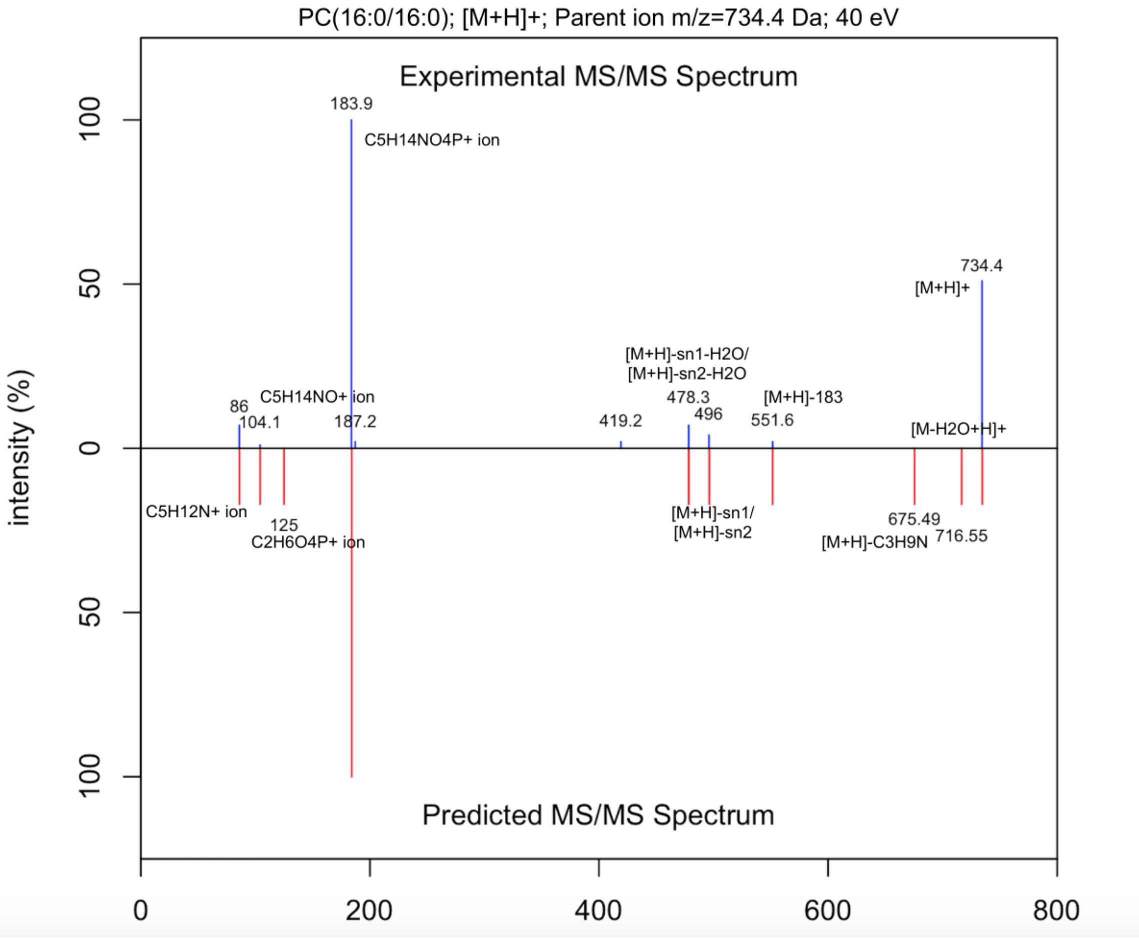


Figure 4.2b Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of dipalmitoyl phosphatidylcholine measured in positive ion mode ([M+H]+) at 40 eV, and the matching ESI-MS/MS spectrum predicted by *CFM-ID* 3.0. The computed spectral similarity is 0.98.

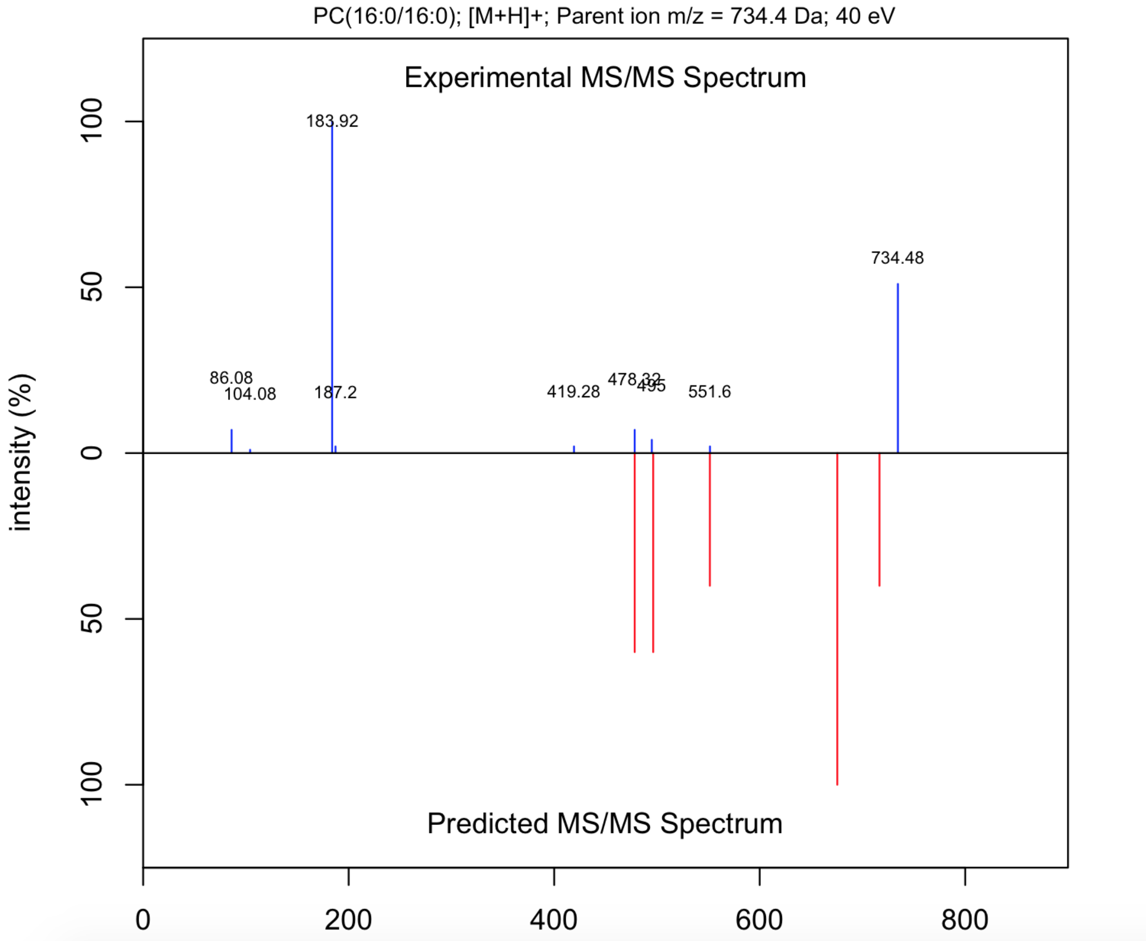


Figure 4.2c Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of dipalmitoyl phosphatidylcholine measured in positive ion mode ([M+H]+) at 40 eV, and the matching ESI-MS/MS spectrum predicted by LipidBlast. The computed spectral similarity is 0.13.

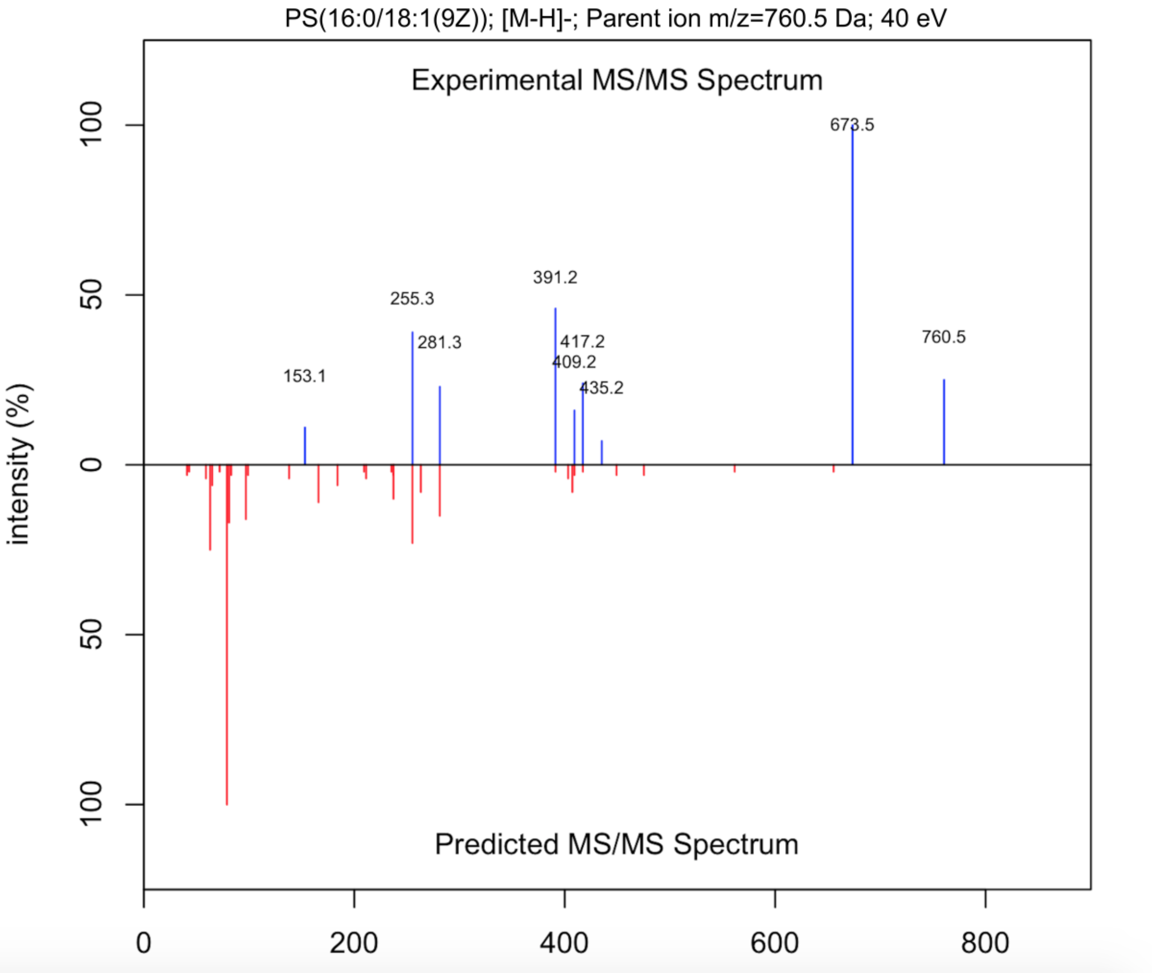


Figure 4.3a Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (PS(16:0/18:1(9Z))) measured at 40 eV, and the matching ESI-MS/MS spectrum predicted by *CFM-ID* 2.0. The computed spectral similarity is 0.10.

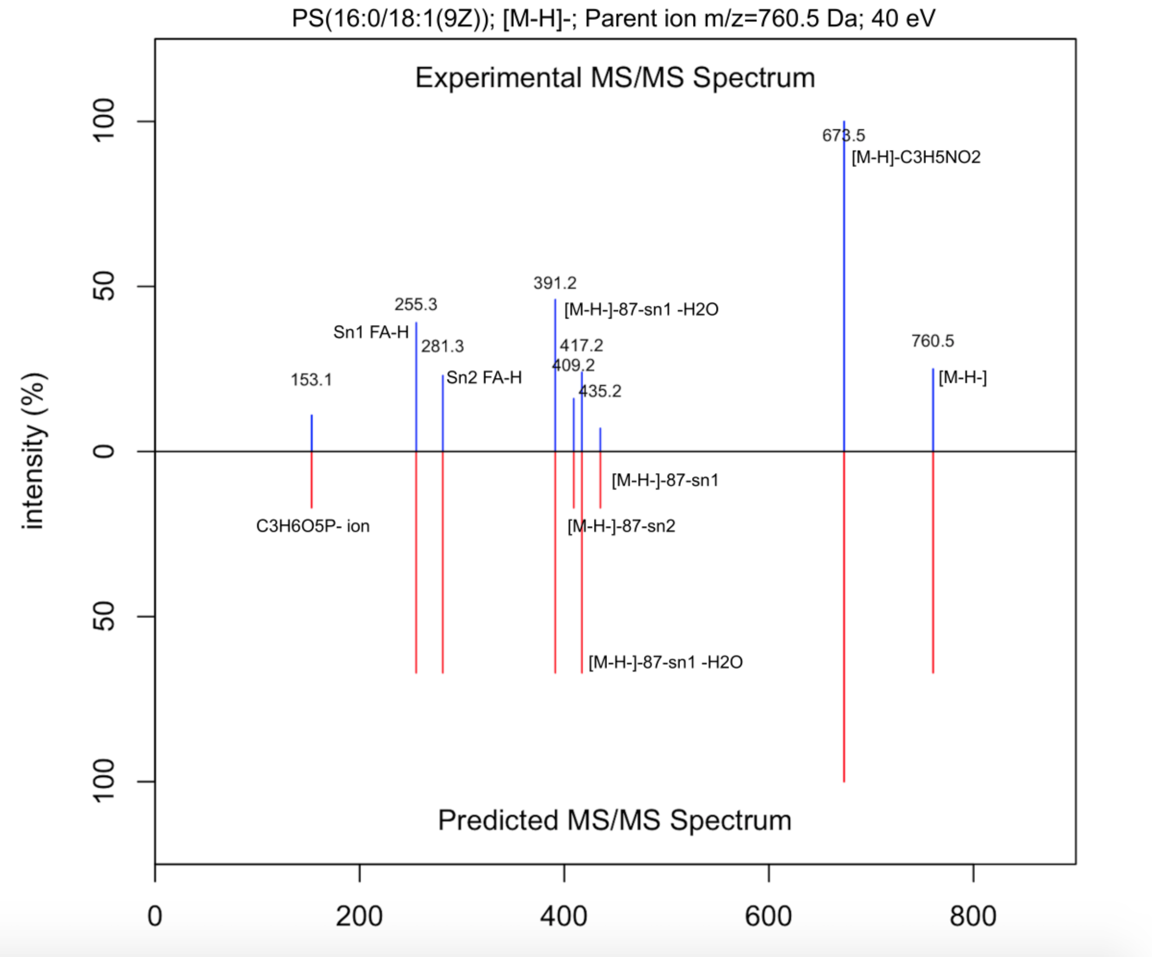


Figure 4.3b Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (PS(16:0/18:1(9Z))) measured at 40 eV, and the matching ESI-MS/MS spectrum predicted by *CFM-ID* 3.0. The computed similarity is 0.92.

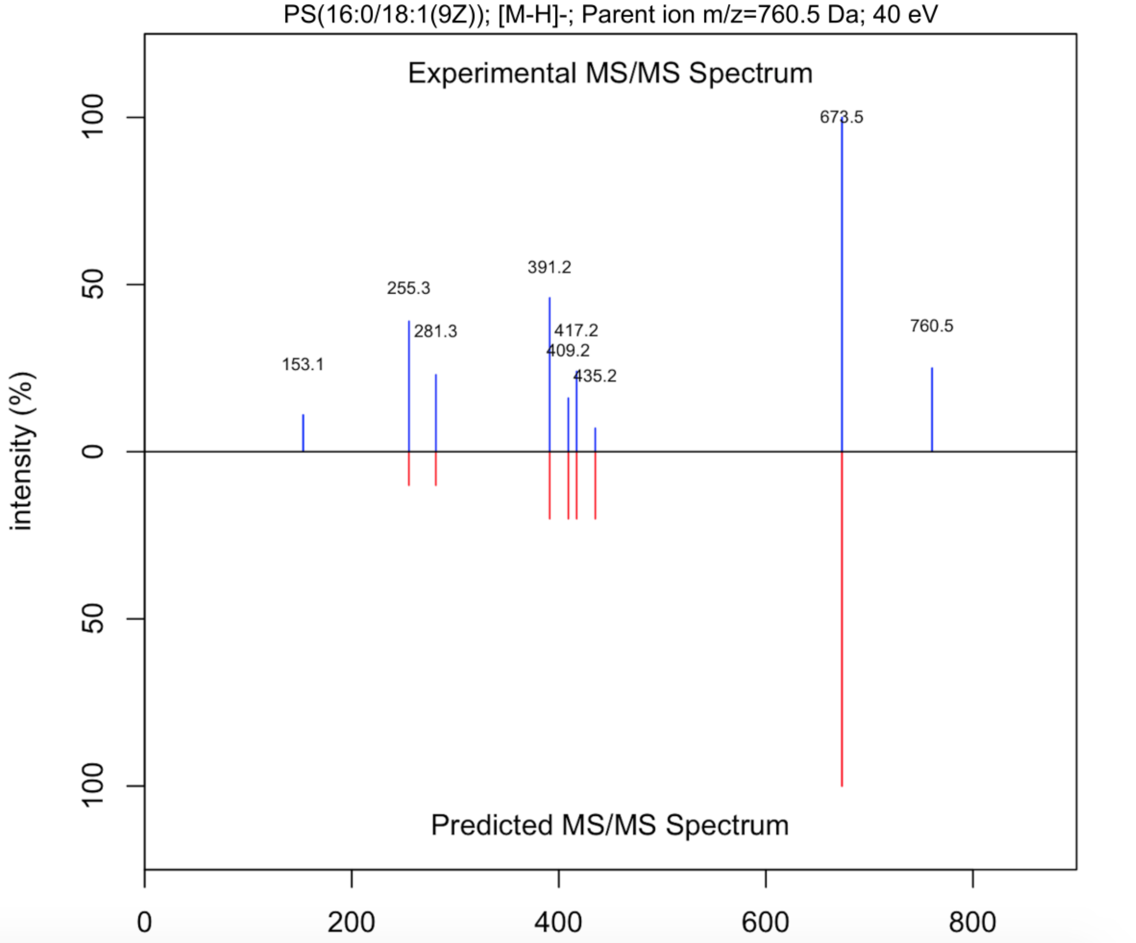


Figure 4.3c Head-to-tail plot showing an experimental of ESI-MS/MS spectrum of 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (PS(16:0/18:1(9Z))) measured at 40 eV, and the matching ESI-MS/MS spectrum predicted by LipidBlast. The computed similarity is 0.91.

Table 4.3 Computed spectral similarities between experimental and predicted ESI-MS/MS spectra. The results show higher similarities, and thus an improvement when using a rule-based approach (*CFM-ID* 3.0) over a combinatorial one (*CFM-ID* 2.0) for the prediction of lipids. The spectral similarities of the LipidBlast generated spectra further illustrate this trend.

Compound	Adduct	Energy	*CFM-ID* 3.0	*CFM-ID* 2.0	LipidBlast
DG(16:0/16:0/0:0)	[M+NH4]+	20 eV	0.98	0.07	0.28
PC(16:0/16:0)	[M+H]+	40 eV	0.88	0.0	0.13
PE(16:0/18:1(9Z))	[M+H]+	20 eV	0.90	0.03	0.95
SM(d18:1/16:0)	[M+Na]+	40eV	0.96	0.0	0.68
PI(18:0/20:4)	[M+H]-	40 eV	0.91	0.12	0.91
TG(22:6/22:6/22:6)	[M+Li]+	40 eV	0.96	0.09	0.97
PE(16:0/18:1(9Z))	[M-H]-	40 eV	0.96	0.15	0.89
PS(16:0/18:1(9Z))	[M-H]-	40 eV	0.92	0.10	0.91
CL(18:1/18:1/18:1/18:1)	[M-2H](2-)	40 eV	0.95	0.02	0.94
PG(16:0/18:1(9Z))	[M-H]-	40 eV	0.96	0.08	0.96

4.3.3.2 Compound Identification using the New Scoring Functions

As noted earlier, two sets of 1,000 compounds were used to train a specific scoring function for ESI-MS/MS-based compound identification, and a second one for EI-MS-based compound identification. Both functions were developed in order to optimize *CFM-ID 3.0*'s compound identification performance. The models were obtained using

5X cross-validation, and tested on different sets. Table 4.4 compares the performance of *CFM-ID 3.0* versus *CFM-ID 2.0* for compound identification based of 208 ESI-MS/MS spectra from 185 unique compounds. The spectra were provided during the CASMI 2016 contest (category 3). Here, *CFM-ID 3.0* was able to correctly identify the query compound in 144 out of 208 challenges, compared to only 113 by *CFM-ID 2.0*. This represents an improvement of 27.4%. The query compound was generally ranked higher (3) by *CFM-ID 3.0* compared to *CFM-ID 2.0* (4). Table 4.5 shows the comparison between *CFM-ID 3.0* and *CFM-ID 2.0* for compound identification tasks based on 200 EI-MS spectra for 200 unique compounds. The spectra were retrieved from the set of compounds imported from various sources, as described in section 4.2.4. These spectra were excluded from the searchable database during the test phase. Here, *CFM-ID 3.0* was able to correctly identify the query compound in 118 out of 200 challenges, compared to only 109 by *CFM-ID 2.0*. This represents an improvement of 8%. When using EI-MS spectra as input, *CFM-ID 3.0* also ranked the query compound higher (3) compared to CFM-ID 2.0 (4). *CFM-ID 3.0* also achieved a better medal score compared to *CFM-ID 2.0*.

Table 4.4: Comparison of *CFM-ID 3.0* and *CFM-ID 2.0* scoring functions upon identification of 185 compounds from 208 ESI-MS/MS spectra. Reported are the total number of challenges in which the corresponding implementation of the scoring function ranked the query compound in the Top 1, Top 3, and Top 10. The average rank for the query compound is also reported. A chemical classification is assessed as correct if the predicted category matches a category originally assigned by *ClassyFire*. N/A: Not Applicable.

Version	# Top 1	# Top 3	# Top 10	Avg. rank	# Correct classifications
CFM-ID 3.0	137	191	205	1.7	159
CFM-ID 2.0	133	190	205	2	N/A

Table 4.5: Comparison of *CFM-ID 3.0* and *CFM-ID 2.0* scoring functions upon identification of 200 compounds from 200 EI-MS spectra. Reported are the total number of challenges in which the corresponding implementation of the scoring function ranked the query compound in the Top 1, Top 3, and Top 10. The average rank for the query compound is also reported. A chemical classification is assessed as correct if the predicted category matches a category originally assigned by *ClassyFire*. N/A: Not Applicable.

Version	# Top 1	# Top 3	# Top 10	Avg. rank	# Correct classifications
CFM-ID 3.0	118	153	184	3	134
CFM-ID 2.0	109	137	175	4	N/A

4.3.3.3 Compound Classification

Since *CFM-ID 3.0* uses different scoring functions for ESI-MS/MS and EI-MS based metabolite identification, the compound classification algorithm was tested separately for each type of MS spectral mode. The chemical class of the query compound is predicted as the direct parent of the highest-ranked compound. In case of a tie, the predicted class is the most occurring chemical class among the direct and alternative parents of all compounds with the highest score. When using ESI-MS/MS spectra as input, *CFM-ID 3.0* correctly predicted the chemical class in 160 out of 208 challenges. In 24 out of 208 cases, the query compound had not been correctly identified; thus, in 33.3% of the cases were the compound could not been identified (48 in total), it was assigned a correct chemical class. When using EI-MS spectra as input, *CFM-ID* correctly predicted the chemical class of the query compound in 134 out of 200 challenges. Moreover, in 16 out of those 134 cases, the query compound was not correctly identified; thus, in 24.4% of the cases where the compound could not be identified (66 in total), it was at least assigned a correct chemical class. These results suggest that *CFM-ID 3.0* was still able to capture structural features that characterize the fragmentations observed in the corresponding input MS/MS spectra. These results also demonstrate the importance of using a diverse set of compounds and spectra, as well as the need of having a sufficiently large compound/spectral database.

4.4 Discussion

4.4.1 ESI-MS/MS Lipid Spectral Prediction

The comparisons illustrated in Figures 4.2b and 4.3b show that *CFM-ID 3.0* can predict ESI-MS/MS spectra for lipids that very closely match experimental ESI-MS/MS spectra collected for the same compounds. The predicted MS/MS peaks match most of the peaks in the experimental spectra. Figures 4.2a, and 4.3a compares an experimental ESI-MS/MS spectrum with the corresponding *in silico* MS/MS spectrum generated by *CFM-ID 2.0*, and shows relatively little similarity between the experimental and predicted spectra, in both cases. The much higher performance for lipid spectra obtained with rule-based fragmentation approaches over combinatorial fragmentation approaches can be explained by two factors. First, lipids are modular molecules and so the MS fragmentation patterns seen under most collision energies are easily understood and relatively simple to describe. On the other hand, combinatorial fragmenters have no knowledge of molecular structure and so they cannot recognize modular structures. Instead, they view lipids as molecules with dozens of breakable bonds, all of which could potentially be fragmented. This leads to a substantial over-prediction of MS peaks. The second reason why combinatorial fragmenters do not perform well is that they have generally not been "trained" on lipid spectra. For example, *CFM-ID 2.0* was only trained on ~1000 experimental MS/MS spectra, none of which included lipid MS/MS spectra. Similarly, *MetFrag* (182), another combinatorial fragmenter, was also not programmed to handle lipid MS/MS spectra. By expanding *CFM-ID*'s training set and including lipid spectra as well as other modular compound classes) in that training set, *CFM-ID* could potentially improve its performance to match even the rule-based fragmenter. Currently we are working on testing this possibility.

Overall, our results show that *CFM-ID 3.0* was able to reproduce most lipid fragments with accurate m/z ratios and reasonably accurate relative intensities. Characteristic fragment ion losses (e.g. loss of polar head, or side chains) were reproduced accurately. They also include many ion fragments that are independent of the acyl or alkyl chain(s) of the molecular ion, including the cyclic 1,2-cyclic phosphate diester (neutral m/z=123.99 Da) fragment, which often observed in ESI-MS/MS spectra of various choline glycerophospholipids. Interestingly, most of these fragments were not reported in LipidBlast. As expected, some discrepancies were observed when comparing predicted MS/MS spectra with the corresponding experimental MS/MS spectra. First, the relative peak intensities were generally found to be higher in the predicted MS/MS spectra than the experimental spectra. Second, the peak lists are often not identical. MS/MS spectral peak intensities are very difficult to predict and vary considerably depending on the instrument, the instrument parameters and experimental design. For instance, phosphatidylcholines, when analyzed by Q-TOF instruments, tend to lose the molecular ion even at medium collision energies. On the other hand, when phosphatidylcholines are analyzed on Ion Trap MS instruments the molecular ion is still highly abundant at medium collision energies, and is significantly fragmented only at high energies (198,281,289). In addition to instrument differences, the type of solvent being used can affect the extent to which a compound is fragmented. However, rather than focusing on these subtleties, we chose to focus on selecting (and annotating) the most abundant or most characteristic fragments, which were generally reproducible on different instruments, and reported in multiple studies. Moreover, we limited the number of peaks to be predicted to the centroid peaks and did not include their isotopomers, since those are often of much lower intensity.

While *CFM-ID 2.0* predicts fragmentation probabilities and numeric peak intensities, *CFM-ID 3.0* does not predict peak intensities for lipid spectra (however it still predicts numeric peak intensities for all other classes of molecules). Instead, *CFM-ID 3.0* predicts categorical peak intensities for lipid spectra (low, medium, high, and maximum abundance). This simple categorization partly explains why, in many cases, the relative peak intensity is higher in predicted lipid spectra compared to experimental spectra. We believe that a larger lipid MS spectral training set would help to improve the prediction of numeric intensities and simulate their variation between collision energies more accurately. Another limitation of *CFM-ID 3.0*'s rule-based approach is that the current fragmentation rules do not take the information about the stereochemistry and the position of double/triple bonds into consideration. Therefore, they cannot allow one to distinguish between stereoisomers or regiomers. This is a common problem for rule-based "fragmenters", since the incorporation of such distinctions would required the acquisition of a much more diverse and larger set of high-resolution MS^n spectra.

As noted before, *CFM-ID 3.0* returns the structure (in InChI or SMILES strings) for all predicted fragments. This helps to provide a rationale for nearly all observed peaks. Additionally this linkage simplifies lipid ESI-MS/MS spectral annotation process. Because *CFM-ID 3.0* provides MS/MS spectra at three energy levels (10 eV, 20 eV, and 40 eV), it means that the predicted MS/MS spectra can be matched more closely to real experimental conditions and real experimental MS/MS spectra. Many other spectral libraries (LipidBlast, NIST) only provide consensus MS/MS spectra for lipids, which makes it difficult to relate experimental data to the predictions.

4.4.2 Compound Identification and Class Prediction

The incorporation of citation counts in MS-based compound identification protocols has been consistently shown to improve identification rates in recent studies (182,278). However, an obvious limitation of this approach is that it reduces the probability of identifying novel or rare compounds that have never been cited. It can also bias the ranking scheme to select one very similar structure (and therefore very similar MS spectrum) over another purely on the basis of one having slightly more citations than another. To help balance the influence of citation counts we incorporated chemical classification into our new scoring system. In this way, the scientific relevance or approximate abundance (in terms of citations) as well as the structural features among candidates could be taken into consideration. Using this approach, two scoring functions were defined for compound identification: one for ESI-MS/MS spectra and one for EI-MS spectra. In comparison, *CFM-ID 2.0* uses the same scoring function for ESI-MS/MS and EI-MS input spectra. The newly defined functions used in *CFM-ID 3.0* helped to improve identification (see Tables 4.4 and 4.5). In particular, when applied to 208 identification challenges, the ESI-MS/MS scoring function achieved an improvement in ranking (1.0) and identification rate (27.4%) over *CFM-ID 2.0*'s original scoring function, and the compound identification rate was 27.4% higher. Moreover, the new EI-MS scoring function also improved the ranking of the query molecule by and average of 1.0, and achieved 8% more correct identification, compared to *CFM-ID*'s 2.0 original function. We believe the use of diverse training sets of compounds, representing widely varying structures and structural classes was critical to achieving this performance.

CFM-ID 3.0 was also assessed with regard to its performance in chemical class prediction. As noted earlier, while it may not be possible to identify the exact compound via MS/MS spectral matching, the ability to use MS/MS spectra to narrow down the correct chemical class or chemical family for a given query spectrum or compound can be very valuable for many applications in metabolomics or natural product de-replication. In assessing the performance of *CFM-ID*'s chemical class prediction the same scoring systems introduced here was used to rank the individual candidates; but in order to perform a formal chemical class identification, the query compound was predicted to belong to the “direct parent” class of the highest-ranked candidate. In cases of a tie, the predicted chemical class was predicted to be the most frequently occurring among all the direct and alternative parents among all the compounds with the highest score. Upon testing the new ESI-MS/MS scoring function on 208 challenges, the correct class as predicted in 76.9% of the challenges. In 33.3% of correct class predictions, the query compound was not correctly identified; this suggest that *CFM-ID 3.0* was still able to capture structural features that characterize the fragmentations observed in the corresponding input MS/MS spectra. Upon testing of the new EI-MS scoring function on 200 challenges, the correct class was predicted in 67% of the challenges, with 24.4% of the correct classification achieved despite a misidentification of the query compound. These results also demonstrate the importance of using a diverse set of compounds and spectra, as well as the need of having a sufficiently large database. Structurally similar compounds tend to produce similar spectra. Therefore, even if the compound is not available in the database (or is poorly ranked), high number of compounds from various classes of compound could help to discriminate between the different classes, and also capture the patterns that are characteristic of specific class. We believe that this helped *CFM-ID 3.0* to achieve a good performance in the class prediction task.

The inclusion of additional data (citation frequency and chemical class information) in the *CFM-ID* scoring functions is clearly important in achieving good compound identification results, but so too is the quality of MS/MS spectra predicted by *CFM-ID 3.0*. While we have made substantive improvements to the quality of *CFM-ID*’s lipid spectra prediction, more work still needs to be done in *CFM-ID* to better mimic the fragmentation of other classes of compounds (such as alkaloids, polyphenols, terpenes and steroids) and increase the quality predicted MS/MS spectra. Work is now ongoing to increase *CFM-ID*'s training set (by a factor of 5) and to improve its generative rules through advanced machine learning techniques. These will be described in an upcoming publication. The addition of 57,500+ experimental EI- and ESI-MS/MS spectra, measured with various MS instruments, and under different conditions, is expected to further help capturing spectral patterns that are not yet described by *CFM-ID*'s predicted spectra, and thus, help increase the compound identification rates.

4.5 Conclusion

We have shown that it is possible to substantially improve *CFM-ID*'s performance in both spectral prediction and compound identification tasks. This was achieved by: 1) integrating a rule-based fragmentation approach that currently applies 378 manually curated rules to predict the ESI-MS/MS spectra for 26 classes of common, biologically important lipids, 2) modifying the structure of *CFMD*'s spectral database, and increasing its size by a factor of 2.2, and 3) designing a new scoring function that takes into account both compound citation frequency and chemical classification features of candidate molecules.

In particular, the implementation of a rule-based approach for fragment ion prediction was shown to improve the speed and accuracy of the lipid ESI-MS/MS spectra prediction by a factor of 10-200X. The success of using rule-based fragmentation patterns encoded in standard chemical representations (SMILES, SMARTS and SMIRKS) suggests that this concept could be successfully applied to other classes of modular molecules such as polyphenols, terpenes and carbohydrates. The construction and expansion of *CFM-ID*'s spectral library has also helped *CFM-ID*'s overall performance.

The spectral library has been expanded by a factor of 2.2 over the previously available library. This expansion process is still ongoing, and we plan to include ~400,000 more compounds including drugs, lipids, environmental pollutants, phytochemicals, food compounds, as well as their predicted metabolites generated by *BioTransformer* (see Chapter 3). The new scoring function, which already showed an improvement over *CFM-ID* 2.0's scoring function, could potentially be further improved by using machine learning techniques and training over a much larger set of MS/MS spectra. Moreover, the acquisition and incorporation of other metadata, such as retention time, could help further increase the compound identification rates, as demonstrated in several recent studies (182,278). The fields of *in silico* metabolomics and *in silico* mass spectrometry are rapidly evolving. Thanks to the many excellent ideas emerging in many labs and the willingness of many researchers to share their code and their databases, it is likely that these fields will continue to grow and continue to inspire others to make MS spectral analysis, MS spectral prediction, MS-based compound identification better, faster and even more informative.

Chapter 5

General Conclusions and Future Perspectives

5.1 General Conclusions

This thesis has focused on three computational challenges in the fields of metabolomics and cheminformatics - 1) the proper description and categorization of known chemicals and metabolites, 2) the prediction of secondary metabolite structures and the biosynthetic pathways that lead to them, and 3) the improved prediction of the mass spectrometry (MS) spectra of known (or predicted) metabolites. As highlighted in the introduction, tackling these issues would provide the necessary tools to accelerate drug development, improve biomarker discovery, enhance environmental toxicology, and make many other fields of metabolomic science better, faster and cheaper. To this end, I have developed several new programs and databases including ChemOnt, *ClassyFire*, and *BioTransformer* to address the first two computational challenges. To address the third computational challenge, I have implemented new algorithms and produced an improved version of *CFM-ID*, a software tool and web server for the automated prediction of MS spectra. Here I will summarize the main findings and novel features for each of these tools or data resources.

5.1.1 Automated Hierarchical Structure-based Chemical Classification with ChemOnt and *ClassyFire*

To address the first computational challenge (Objective #1 in Chapter 1), I developed ChemOnt and *ClassyFire*. These tools were described in detail in Chapter 2 of this thesis. Briefly*, ChemOnt* is a chemical ontology that contains 4,825 chemical categories with detailed textual descriptions or definitions. Each of these chemical categories was named using chemical terms extracted from the scientific literature or scientific databases. For the sake of organizational simplicity, ChemOnt is implemented using a tree structure with 11 different levels, partly inspired by the *Linnaean* taxonomy of living species. These categories are organized and named in a way that reflects the conventions and knowledge of both biochemists and chemists. ChemOnt has been designed to work for both organic and inorganic compounds. Each chemical category is carefully described in English (25-75 words) based on the structural features common to all the compounds found in that category. To improve the interoperability of this ontology, ChemOnt also provides a total of 9,012 English synonyms for its chemical categories. These synonyms were obtained by mapping ChemOnt to other well-established and popular ontologies such as ChEBI (200), LIPIDMAPS (198), and MeSH (121). The mapping of ChemOnt to ChEBI and LIPIDMAPS was performed jointly with the ChEBI and LIPIDMAPS curation teams. This was done to facilitate the development of a standard chemical computable ontology, and the facilitation of data exchange between the main chemical/biochemical libraries used in metabolomics. ChemOnt is compliant with the Open Biological and Biomedical Ontologies (OBO) format to improve its integration with respect to modern semantic technology approaches. ChemOnt is currently the largest and most complete computable chemical ontology constructed to date.

ClassyFire is a RESTful application and web server that automatically performs hierarchical structure-based classification of chemicals using the ChemOnt ontology. It is written in the Ruby language and uses the Rails Framework. The central idea behind *ClassyFire's* operation is the expression of the text descriptions provided by ChemOnt into a set of rules that are understandable by a computer. *ClassyFire* combines a number of cheminformatics approaches to analyze each compound's elemental composition, structure, name, and physical properties, in order to rapidly and precisely identify structural features, and to automatically generate a comprehensive, consistent classification based on the *ChemOnt* ontology rules. In addition to small molecules (e.g. amino acids, vitamins, bisphosphonate drugs), *ClassyFire* can also classify large polymeric molecules such as polypeptides, DNA and RNA based on their chemical substituents. The resulting classifications are provided in several standard formats (JSON, SDF, CSV) that can be easily retrieved and parsed. Moreover, in order to reduce the computational time and to enable text-based searching, all of *ClassyFire's* classification results are stored into a MySQL database. *ClassyFire* was tested on a chemically diverse set of 800 compounds that had been manually classified and annotated by multiple experts, and achieved a precision of 99.8% and a recall of 99.9%. Moreover, upon analysis of a smaller set of compounds, it was shown to reproduce ~94% of manually performed ChEBI ontological annotations and to suggest many new ontological terms that could further increase the number of annotations by 43.6%.

ClassyFire has been used to classify >90 million compounds so far, most of which are from the NCBI's PubChem database (the largest publicly available chemical database). These classifications are currently accessible via the *ClassyFire* web server, and will be uploaded soon, in a joint effort with the NCBI, to the PubChem database. In addition to this work with PubChem, *ClassyFire* was also used to generate textual descriptions for >13,000 compounds and to infer the biological properties of >100,000 molecules found in the HMDB, T3DB, ECMDB, FooDB, and YMDB databases. The *ClassyFire* web server is available at http://classyfire.wishartlab.com.

Because *ClassyFire* is the only open access, fully automated chemical classification tool available for chemists, many research scientists around the world are using it. This popularity has required the development of several APIs that allow programmatic access to the *ClassyFire* web server. The first Ruby-based API for *ClassyFire* was written as part of my thesis, and is available at https://bitbucket.org/wishartlab/classyfire_api.

5.1.2 Metabolism Prediction with *BioTransformer*

To address the second computational challenge (Objective #2 in Chapter 1), I developed *BioTransformer*. This program was described in detail in Chapter 3 of this thesis. Briefly*, BioTransformer* is a software tool designed to predict the secondary metabolism (both the pathways and resulting structures) of endogenous and exogenous small molecules. The key motivation behind the development of *BioTransformer* was to synthetically expand the universe of known compounds/metabolites by generating biologically feasible compounds from existing parent (i.e. known) compounds. To achieve this, *BioTransformer* uses a hybrid approach that combines machine learning capabilities with a knowledge-based system to predict the following: 1) CYP450-mediated (phase I) and phase II metabolism of xenobiotics in humans, 2) metabolism of xenobiotics by the human gut microbiome, 3) small molecule metabolism by the environmental microbiome (covering soil and water), and 4) metabolism of small molecules via promiscuous enzyme reactions (based on the Enzyme Classification (EC) provided by the International Union of Biochemistry and Molecular Biology (IUBMB)). *BioTransformer* is designed to work in three biological systems or environments (simply referred to as “biosystems”) - the human body, the human gut, and the environmental microbiome. *BioTransformer* was developed in such a way that additional “biosystems” could be easily included in its framework.

BioTransformer’s knowledge-based system consists of several dictionaries containing descriptions (in text and in the SMIRKS language) that map generic chemical reactions, enzyme lists, as well as manually curated and expert-validated rules for reaction prioritizations and compound validation. In particular, this knowledgebase consists of a manually curated collection of >1200 experimentally confirmed biotransformations, called BioTransformerDB. BioTransformerDB covers CYP450 metabolism, phase II metabolism, and human gut metabolism. *BioTransformer’s* machine learning module was developed partly by exploiting the data in BioTransformerDB to train a machine learning model to predict the substrate specificity of nine phase I metabolizing enzymes. These enzymes catalyze >90% of the xenobiotic metabolic reactions and most of the phase I oxidative reactions in humans. The prediction of phase II biotransformations, human gut (i.e. microbial) biotransformations, promiscuous enzyme transformations as well as environmental biotransformations is based on the data, hand-made rules and reaction schemes contained in BioTransformerDB. While almost all of the reaction schemes and biotransformations in BioTransformerDB were obtained via manual literature searches, the biotransformations for environmental microbial metabolism were extracted from the EAWAG prediction system, with permission from the developers.

BioTransformer predicts metabolites resulting from either a single chemical reaction or a multiple reaction sequence. As a result, it provides users with the ability to mimic the interplay of several biosystems (e.g. human organs and the human gut microbiome) in predicting or modeling the metabolism of small molecules. These functions exploit the rules designed, in part, through the *ChemOnt* ontology to guide the selection of pathways or reactions lists in order to reduce computational costs. Upon completion of each metabolism prediction task, *BioTransformer* returns one or more predicted biotransformations, with descriptions, scores, and the structures of the predicted metabolites.

To validate our approach, *BioTransformer* was compared to *Meteor Nexus*, a well-regarded, commercially available metabolism prediction tool. For the prediction of human secondary metabolites on a defined set of input molecules (including phase I, phase II, and human gut microbial), *BioTransformer* achieved a precision of 47.3% and a recall of 83.7%, compared to a precision of 18.9% and a recall of 23.8% by *Meteor Nexus* (50). For the prediction of environmental secondary metabolites, *BioTransformer* was able to reproduce 100% of the metabolites, when compared to the EAWAG system (49,161). *BioTransformer* is an open-source project and is freely available at https://bitbucket.org/djoumbou/biotransformer.

5.1.3 MS-spectral prediction and Compound identification with *CFM-ID 3.0*

To address the third computational challenge (Objective #3 in Chapter 1), I enhanced and improved the performance of an MS prediction program called *CFM-ID*. These enhancements are described in detail in Chapter 4 of this thesis. Briefly*, CFM-ID* is a software tool and web server that performs three tasks – 1) the prediction of EI and ESI MS-spectra, 2) the annotation of peaks in a set of MS-spectra given a molecule, and 3) the identification of compounds given a set of MS-spectra. My specific focus was on improving the performance of the first and third tasks.

The first two versions of *CFM-ID* addressed the prediction of MS-spectra solely using the combinatorial fragmentation approach (10-12). However, as noted in Chapter 4, there are numerous examples where this approach leads to very long computational times (>1 hour) for certain compounds. In many cases these also lead to incorrect predictions. This is particularly true for lipids and fatty acids. To improve *CFM-ID's* performance, I implemented a rule-based approach using Java to encode adduct-dependent fragmentation rules for 26 classes of lipids. This work is partly based on the templates provided by Kind *et al.* (189) and LipidBlast for the automated construction of MS/MS-spectral libraries for lipids. Key to making this work was the integration of *ClassyFire* as a front-end filter to *CFM-ID,* which permits the automated recognition of lipids (and other hard-to-predict structures) and subsequent rule-based processing. The integration of this module in *CFM-ID 3.0* not only sped up the analysis of lipids by a factor of 173X but it was also able to reproduce experimental spectra with a similarity of 0.92 +/- 0.02 on average, compared to 0.07 +/- 0.04 for *CFM-ID 2.0*.

In order to improve *CFM-ID's* performance in compound identification, *CFM-ID's* spectral library was enriched with >207,000 high-quality, experimentally collected MS-spectra for >100,000 distinct compounds. As was shown by others (*MetFrag* (182)), combining experimentally collected MS/MS spectra with computationally generated MS/MS spectra greatly improves compound identification performance. Thus, I also enhanced *CFM-ID*'s compound and spectral databases with meta-information, such as chemical classification data and citation counts (associated with each compound). The inclusion of this information was also shown to improve the performance of compound identification tools when integrated in the scoring function. The chemical classification was provided by *ClassyFire*, while the citation count was obtained from several databases, including the PubChem database (9), the ChEBI database (97), and the CompTox database (133), among others. These enhancements provided data that was used to develop new ranking functions. Indeed, in several tests we found *CFM-ID*'s performance improved by 27.4% when the input consists of ESI-MS/MS spectra, and by 8% when the input consists of EI-MS spectra. Moreover, *CFM-ID 3.0* integrates a classification approach, which is helpful *when CFM-ID 3.0* fails to correctly identify a compound. When the input consisted of ESI-MS/MS, *CFM-ID 3.0* correctly classified the query compound in 33.3% of the cases where the compound was misidentified. When the input consisted of EI-MS, *CFM-ID 3.0* correctly classified the query compound in 24.4% of the cases where the compound was misidentified. *CFM-ID*'s source code is available at https://sourceforge.net/p/cfm-id/wiki/Home, and the web server is accessible at http://cfmid-staging.wishartlab.com

5.2 Future Perspectives

As demonstrated throughout this document, the structural and physico-chemical properties of metabolites influence their fate and their effects in the environment. Beginning with my structure-based chemical classification scheme implemented via ChemOnt and *ClassyFire*, I was able to develop a suite of computational tools capable of predicting biosynthetic pathways and the corresponding metabolite structures associated for a wide variety of biosystems or environments. These same principles and tools also helped me implement a rule-based MS/MS fragmentation approach for predicting the

MS/MS spectra of lipids and provide the program with the necessary meta-information needed to improve ESI-MS/MS-based and EI-MS-based compound identification. While each of these developments represents an important advance or a significant improvement to the current "state-of-the-art", there is still considerable room for improvement. In the following paragraphs, I provide some suggestions and ideas that could be explored to help improve the quality and performance of these tools.

5.2.1 ChemOnt and *ClassyFire*

While a structure-based ontology for chemical compounds is very useful from the perspective of chemists, I believe that the integration of other biomedical or biochemical concepts (e.g. diseases, health effects, biological pathways, chemical and biological roles) into the *ChemOnt* ontology would be helpful. In particular, the inclusion of biomedical information would further facilitate the integration and exchange of data between chemists and biologists as well as between cheminformaticians and bioinformaticians. This is something that happens very infrequently. This kind of cross-disciplinary ontology would also help in the development of a semantic-based framework for knowledge discovery. Such a framework may lead to a more widespread adoption of *ChemOnt* in the cheminformatics/bioinformatics communities.

While such a modification to the *ChemOnt* ontology could take several years to complete, I believe there are other, smaller enhancements that could be done over a much shorter period of time. For example, one simple improvement could be made by optimizing *ClassyFire's* structure search algorithm. By using a technique known as "partology" (describing the relationship between a structural pattern and its parts), it should be possible to significantly reduce the number of structure search operations and rule evaluations needed to classify compounds. For instance, more than 5,000 structural patterns in the *ClassyFire*'s pattern database contain a benzene ring. Given a query molecule to classify, the current version of *ClassyFire* would run a superstructure operation each on these patterns individually, whether the molecule was found to contain a benzene ring or not. By using a partology that specifies every pattern containing a benzene ring, such patterns could be eliminated form the target list after only one superstructure matching (against benzene). Thus, the search operation over the whole database could be up to 5,000 times faster. Using the same logic, if a compound contains a given pattern, *ClassyFire* would be able to infer all of its substituent patterns that are part of the database without having to run further superstructure search operations. In those two scenarios, the cost of the superstructure search, which remains the bottleneck of the classification process, could be significantly reduced.

5.2.2 *BioTransformer*

The current version of *BioTransformer* uses a hybrid approach (rule-based and machine learning) for predicting CYP450-mediated (phase I) biotransformations. I believe that the acquisition of much more biotransformation data (1000's of additional reactions) would facilitate the design of much better machine learned models for improved phase I metabolism prediction – particular with regard to predicting the site of metabolism (SoM). The current set of reactions (~1,000 in total) in my reaction database is not sufficiently large to develop a robust machine-learning algorithm. However, the acquisition and annotation of 2-3X more phase I reactions would undoubtedly take many months of reading and coding work. While improvements to phase I metabolism prediction are likely to prove to be very challenging, I believe that developing a machine-learning approach for phase II metabolism may prove to be much easier. This is because there is a large abundance of known phase II reactions and the fact that phase II metabolism follows somewhat simpler chemical biotransformation rules. On the other hand, the application of machine learning to predict human gut metabolism is likely to many years away, as the number of known reactions is still quite tiny.

5.2.3 *CFM-ID 3.0*

While the performance of *CFM-ID* 3.0 has been improved quite significantly through the modifications described in Chapter 4, the program is still not able to efficiently cover all of chemical space. In particular, *CFM-ID's* rule-based fragmentation library is currently limited to handling or predicting the fragmentation patterns of lipids. I believe that adding fragmentation rules for other chemical categories, such as surfactant polymers, would make *CFM-ID* much more useful, especially for environmental metabolomics applications. I also believe that the recent expansion of *CFM-ID's* experimental MS/MS library could also be used to improve its performance. One reason why *CFM-ID* performed so poorly with lipids was because the original training set had essentially no lipids. By expanding *CFM-ID*'s MS/MS spectral training set by a factor of 10 and ensuring that the training set includes a far broader collection of compound classes, I believe that *CFM-ID* could improve its performance in spectral prediction and compound identification by another 10-15%.

5.3 Final Words

Given the many challenges facing metabolomics in terms of funding (which is small relative to proteomics or genomics) and in terms of the number of known metabolites/compounds (which is very large relative to the number of genes or number of proteins), I believe the only way to continue to move the field forward will be to improve our ability to 1) computational organize and describe chemicals, 2) computationally predict their chemical/biochemical transformation products and 3) computationally predict the observable properties of chemical compounds (MS spectra, NMR spectra, retention time, drift time, etc.). Through the work described in this thesis, I believe I have made some useful and important contributions to each of these areas and that they will eventually find applications far beyond the relatively narrow field of metabolomics.

References

(1) Fiehn O. Metabolomics - The link between genotypes and phenotypes. Plant Mol Biol 2002;48(1-2):155-171.

(2) Wishart DS. Current progress in computational metabolomics. Brief Bioinform 2007;8(5):279-293.

(3) Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 2016;15(7):473-484.

(4) Brennan L. Metabolomics in nutrition research: Current status and perspectives. Biochem Soc Trans 2013;41(2):670-673.

(5) O'Gorman A, Brennan L. Metabolomic applications in nutritional research: A perspective. J Sci Food Agric 2015;95(13):2567-2570.

(6) Watanabe M, Meyer KA, Jackson TM, Schock TB, Johnson WE, Bearden DW. Application of NMR-based metabolomics for environmental assessment in the Great Lakes using zebra mussel (Dreissena polymorpha). Metabolomics 2015;11(5):1302-1315.

(7) Shin C, Han C, Pae C-, Patkar AA. Precision medicine for psychopharmacology: a general introduction. Expert Rev Neurother 2016:1-9.

(8) Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0-The Human Metabolome Database in 2013. Nucleic Acids Res 2013;41(D1):D801-D807.

(9) Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res 2016;44(D1):D1202-D1213.

(10) Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Res 2015;43(W1):W251-W257.

(11) Kell DB, Brown M, Davey HM, Dunn WB, Spasic I, Oliver SG. Metabolic footprinting and systems biology: The medium is the message. Nat Rev Microbiol 2005;3(7):557-565.

(12) Sajed T, Marcu A, Ramirez M, Pon A, Guo AC, Knox C, et al. ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. Nucleic Acids Res 2016;44(D1):D495-D501.

(13) Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, et al. The Human Urine Metabolome. PLoS ONE 2013;8(9).

(14) Wishart DS, Lewis MJ, Morrissey JA, Flegel MD, Jeroncic K, Xiong Y, et al. The human cerebrospinal fluid metabolome. J Chromatogr B Anal Technol Biomed Life Sci 2008;871(2):164-173.

(15) Corcoran JW, Hahn FE, Snell JF, Arora KL. Mechanism of Action of Antimicrobial and Antitumor Agents. Heidelberg: Springer-Verlag Berlin Heidelberg; 1975.

(16) Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, et al. Mass-spectrometry-based metabolomics: Limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics 2009;5(4):435-458.

(17) Bonvallot N, Tremblay-Franco M, Chevrier C, Canlet C, Debrauwer L, Cravedi J-, et al. Potential Input from Metabolomics for Exploring and Understanding the Links between Environment and Health. J Toxicol Environ Health Part B Crit Rev 2014;17(1):21-44.

(18) Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, et al. The future of NMR-based metabolomics. Curr Opin Biotechnol 2017;43:34-40.

(19) Ibáñez AB, Bauer S. Analytical method for the determination of organic acids in dilute acid pretreated biomass hydrolysate by liquid chromatography-time-of-flight mass spectrometry. Biotechnol Biofuels 2014;7(1).

(20) Harkewicz R, Dennis EA. Applications of mass spectrometry to lipids and membranes. Annu Rev Biochem 2011;80:301-325.

(21) Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: Progress in addressing complexity in diet and health. Annu Rev Nutr 2012;32:183-202.

(22) Southam AD, Lange A, Hines A, Hill EM, Katsu Y, Iguchi T, et al. Metabolomics reveals target and off-target toxicities of a model organophosphate pesticide to roach (Rutilus rutilus): Implications for biomonitoring. Environ Sci Technol 2011;45(8):3759-3767.

(23) Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 2011;472(7341):57-65.

(24) Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, et al. Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 2013;19(5):576-585.

(25) Wang Z, Tang WHW, Buffa JA, Fu X, Britt EB, Koeth RA, et al. Prognostic value of choline and betaine depends on intestinal microbiota-generated metabolite trimethylamine-N-oxide. Eur Heart J 2014;35(14):904-910.

(26) Wang Z, Roberts AB, Buffa JA, Levison BS, Zhu W, Org E, et al. Non-lethal Inhibition of Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis. Cell 2015;163(7):1585-1595.

(27) Grosso G, Marventano S, Yang J, Micek A, Pajak A, Scalfi L, et al. A comprehensive meta-analysis on evidence of Mediterranean diet and cardiovascular disease: Are individual components equal? Crit Rev Food Sci Nutr 2017;57(15):3218-3232.

(28) Berrino F, Muti P, Micheli A, Bolelli G, Krogh V, Sciajno R, et al. Serum sex hormone levels after menopause and subsequent breast cancer. J Natl Cancer Inst 1996;88(5):291-296.

(29) Carruba G, Granata OM, Pala V, Campisi I, Agostara B, Cusimano R, et al. A traditional Mediterranean diet decreases endogenous estrogens in healthy postmenopausal women. Nutr Cancer 2006;56(2):253-259.

(30) Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C, et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res 2017 24 October 2016;45(D1):D979-D984.

(31) Boersma MG, Solyanikova IP, Van Berkel WJH, Vervoort J, Golovleva L, Rietjens IMCM. 19F NMR metabolomics for the elucidation of microbial degradation pathways of fluorophenols. J Ind Microbiol Biotechnol 2001;26(1-2):22-34.

(32) Bonvallot N, Tremblay-Franco M, Chevrier C, Canlet C, Warembourg C, Cravedi J-, et al. Metabolomics Tools for Describing Complex Pesticide Exposure in Pregnant Women in Brittany (France). PLoS ONE 2013;8(5).

(33) Ch R, Singh AK, Pandey P, Saxena PN, Reddy Mudiam MK. Identifying the metabolic perturbations in earthworm induced by cypermethrin using gas chromatography-mass spectrometry based metabolomics. Sci Rep 2015;5.

(34) Aranbar N, Singh BK, Stockton GW, Ott K-. Automated mode-of-action detection by metabolic profiling. Biochem Biophys Res Commun 2001;286(1):150-155.

(35) Caballero-Casero N, Lunar L, Rubio S. Analytical methods for the determination of mixtures of bisphenols and derivatives in human and environmental exposure sources and biological fluids. A review. Anal Chim Acta 2016;908:22-53.

(36) Lv Y, Lu S, Dai Y, Rui C, Wang Y, Zhou Y, et al. Higher dermal exposure of cashiers to BPA and its association with DNA oxidative damage. Environ Int 2017;98:69-74.

(37) Qin X, Lehmler H-, Teesch LM, Robertson LW, Duffel MW. Chlorinated biphenyl quinones and phenyl-2,5-benzoquinone differentially modify the catalytic activity of human hydroxysteroid sulfotransferase hSULT2A1. Chem Res Toxicol 2013;26(10):1474-1485.

(38) Xia H, Chi Y, Qi X, Su M, Cao Y, Song P, et al. Metabolomic evaluation of di-n-butyl phthalate-induced teratogenesis in mice. Metabolomics 2011;7(4):559-571.

(39) NIST Standard Reference Database 1A v14. 2017; Available at: https://www.nist.gov/srd/nist-standard-reference-database-1a-v14, 2017.

(40) MassBank of North America (MoNA). 2017; Available at: http://mona.fiehnlab.ucdavis.edu/, 2017.

(41) Zhu Z-, Schultz AW, Wang J, Johnson CH, Yannone SM, Patti GJ, et al. Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. Nat Protoc 2013;8(3):451-460.

(42) Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, et al. GMD@CSB.DB: The Golm metabolome database. Bioinformatics 2005;21(8):1635-1638.

(43) CAS REGISTRY - The gold standard for chemical substance information. 2017; Available at: http://www.cas.org/content/chemical-substances, 2017.

(44) Williams AJ, Tkachenko V, Pshenichnov A. ChemSpider: How a free community resource of data can support the teaching of nmr spectroscopy. ACS Symp Ser 2013;1128:307-319.

(45) Allen F, Pon A, Greiner R, Wishart D. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. Anal Chem 2016;88(15):7689-7697.

(46) Kind T, Liu K-, Lee DY, Defelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Methods 2013;10(8):755-758.

(47) Da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. Proc Natl Acad Sci U S A 2015;112(41):12549-12550.

(48) Hamdalla MA, Rajasekaran S, Grant DF, Məndoiu II. Metabolic pathway predictions for metabolomics: A molecular structure matching approach. J Chem Inf Model 2015;55(3):709-718.

(49) Wicker J, Fenner K, Ellis L, Wackett L, Kramer S. Predicting biodegradation products and pathways: A hybrid knowledge- and machine learning-based approach. Bioinformatics 2010;26(6):814-821.

(50) Marchant CA, Briggs KA, Long A. In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic. Toxicol Mechan Methods 2008;18(2-3):177-187.

(51) Carlsson L, Spjuth O, Adams S, Glen RC, Boyer S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. BMC Bioinform 2010;11.

(52) Wishart DS. Introduction to cheminformatics. Curr Protoc Bioinformatics 2007;Chapter 14.

(53) Gasteiger J. ET. Chemoinformatics: A Textbook. Weinheim: WILEY-VCH GmbH & Co. KGaA; 2003.

(54) Harrigan GG, Yates LA. High-throughput screening, metabolomics and drug discovery. IDrugs 2006;9(3):188-192.

(55) Fuhrer T, Zamboni N. High-throughput discovery metabolomics. Curr Opin Biotechnol 2015;31:73-78.

(56) Kell DB. Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 2006;11(23-24):1085-1092.

(57) IUPAC: Nomenclature of Inorganic Chemistry IUPAC Recommendations 2005. 2008; Available at: http://old.iupac.org/publications/books/author/connelly.html, 2017.

(58) Favre HA, Powell WH editors. Nomenclature of Organic Chemistry. IUPAC Recommendations and Preferred Name 2013. http://www.acdlabs.com/iupac/nomenclature/ ed.: The Royal Society of Chemistry; 2013.

(59) Wiswesser WJ. How the WLN began in 1949 and how it might be in 1999. J Chem Inf Comput Sci 1982;22:88-93.

(60) Granito CE, Rosenberg MD. Chemical Substructure Index (CSI) - A new research tool. J Chem Doc 1971;11(4):251-256.

(61) Welford S, Jochum C. Chemical Structure Registration for Beilstein Online. In: Warr WA, editor. Chemical Structures 2 Heidelberg: Springer Berlin Heidelberg; 1993. p. 161-170.

(62) Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD. SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. J Chem Inf Model 2008;48(12):2294-2307.

(63) Weininger D. Smiles. 3. Depict. Graphical depiction of chemical structures. J Chem Inf Comput Sci 1990;30(3):237-243.

(64) Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 2003;43(2):493-500.

(65) RDKit: Open-Source Cheminformatics Software. Available at: http://www.rdkit.org/, 2017.

(66) O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An Open chemical toolbox. J Cheminformatics 2011;3(10).

(67) ChemAxon- JChem Suite. 2017; Available at: https://www.chemaxon.com/download/jchem-suite/#jchem, 2017.

(68) SMARTS - A Language for Describing Molecular Patterns. 2007; Available at: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, 2017.

(69) SMIRKS - A Reaction Transform Language. 2007; Available at: http://daylight.com/dayhtml/doc/theory/theory.smirks.html, 2017.

(70) O'Boyle NM. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J Cheminformatics 2012;4(9).

(71) CTFile Formats. Available at: http://c4.cabrillo.edu/404/ctfile.pdf, 2017.

(72) Derwent World Patents Index - Reference Information. 2016; Available at: http://ip-science.thomsonreuters.com/support/patents/dwpiref/, 2016.

(73) Barnard JM. Markush Structure Searching. 2009.

(74) Gardner S, Vinter A. Beyond Markush – Protecting Activity not Chemical Structure. Available at: http://www.cresset-group.com/publications/Beyond_Markush.pdf, 2017.

(75) Markush Technology: Toolkit for the analysis of virtual combinatorial library and Markush structures. Available at: https://www.chemaxon.com/products/markush-ip/, 2016.

(76) ChemAxon: JChem Base. 2017; Available at: https://www.chemaxon.com/products/jchem-base/, 2017.

(77) Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71(C):58-63.

(78) Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 2016;11(2):137-148.

(79) BIOVIA: The keys to understanding MDL keyset technology. 2011; Available at: http://accelrys.com/products/pdf/keys-to-keyset-technology.pdf, 2017.

(80) Daylight Theory Manual. 2011; Available at: http://www.daylight.com/dayhtml/doc/theory/index.html, 2017.

(81) Bender A, Mussa HY, Glen RC, Reiling S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. J Chem Inf Comput Sci 2004;44(5):1708-1718.

(82) Vieth M, Erickson J, Jibo W, Webster Y, Mader M, Higgs R, et al. Kinase inhibitor data modeling and de novo inhibitor design with fragment approaches. J Med Chem 2009;52(20):6456-6466.

(83) Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32(7):1466-1474.

(84) ChemAxon's Marvin Suite. 2017; Available at: https://www.chemaxon.com/download/marvin-suite/, 2017.

(85) Lowe DM, Corbett PT, Murray-Rust P, Glen RC. Chemical name to structure: OPSIN, an open source solution. J Chem Inf Model 2011;51(3):739-753.

(86) MDL MOLfiles, RGfiles, SDfiles, Rxnfiles, RDfiles formats. 2016; Available at: https://docs.chemaxon.com/display/docs/MDL+MOLfiles,+RGfiles,+SDfiles,+Rxnfiles,+RDfiles+formats, 2017.

(87) Skvortsova MI, Stankevich IV, Palyulin VA, Zefirov NS. Molecular similarity concept and its use for predicting the properties of chemical compounds. Russ Chem Rev 2006;75(11):961-979.

(88) Yu X, Geer LY, Han L, Bryant SH. Target enhanced 2D similarity search by using explicit biological activity annotations and profiles. J Cheminformatics 2015;7(1).

(89) Bajorath J. Molecular similarity concepts for informatics applications. Methods Mol Biol 2017;1526:231-245.

(90) Rijnbeek M, Steinbeck C. OrChem - An open source chemistry search engine for Oracle®. J Cheminformatics 2009;1(1).

(91) The RDKit database cartridge. 2016; Available at: http://www.rdkit.org/docs/Cartridge.html, 2017.

(92) Haider N. The checkmol/matchmol Homepage. 2016; Available at: http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html, 2017.

(93) Kiener J. Molecule database framework: a framework for creating database applications with chemical structure search capability. Journal of Cheminformatics 2013;5(1):2016.

(94) T. Jewison. Design and development of novel metabolomic databases and toolsUniversity Of Alberta; 2014.

(95) Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: An update. Nucleic Acids Res 2014;42(D1).

(96) Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A free tool to discover chemistry for biology. J Chem Inf Model 2012;52(7):1757-1768.

(97) Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res 2016;44(D1):D1214-D1219.

(98) LIPID MAPS Lipidomics Gateway, a free resource sponsored by the National Institute of General Medical Sciences. 2016; Available at: http://www.lipidmaps.org/, 2011.

(99) Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: Shedding new light on drug metabolism. Nucleic Acids Res 2014;42(D1):D1091-D1097.

(100) Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: The toxic exposome database. Nucleic Acids Res 2015;43(D1):D928-D934.

(101) FooDB: The Food Metabolome Database. 2016; Available at: http://foodb.ca/, 2017.

(102) Salek RM, Haug K, Conesa P, Hastings J, Williams M, Mahendraker T, et al. The MetaboLights repository: Curation challenges in metabolomics. Database 2013;2013.

(103) Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44(D1):D457-D462.

(104) Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res 2008;36(SUPPL. 1):D402-D408.

(105) Steinbeck C, Kuhn S. NMRShiftDB - Compound identification and structure elucidation support through a free community-built web database. Phytochemistry 2004;65(19):2711-2717.

(106) Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison Metabolomics Consortium Database [3]. Nat Biotechnol 2008;26(2):162-164.

(107) Bingol K, Li D-, Bruschweiler-Li L, Cabrera OA, Megraw T, Zhang F, et al. Unified and isomer-specific NMR metabolomics database for the accurate analysis of 13C-1H HSQC spectra. ACS Chem Biol 2015;10(2):452-459.

(108) Hummel J, Strehmel N, Bölling C, Schmidt S, Walther D, Kopka J. Mass Spectral Search and Analysis Using the Golm Metabolome Database. The Handbook of Plant Metabolomics; 2013. p. 321-343.

(109) Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45(7):703-714.

(110) Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: A metabolite mass spectral database. Ther Drug Monit 2005;27(6):747-751.

(111) Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 2014;42(W1):W94-W99.

(112) Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2016;44(D1):D471-D480.

(113) Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: A database of reactions, pathways and biological processes. Nucleic Acids Res 2011;39(SUPPL. 1):D691-D697.

(114) Nishimura D. BioCarta. Biotech Software & Internet Report 2004;2(3):117-120.

(115) Jewison T, Su Y, Disfany FM, Liang Y, Knox C, MacIejewski A, et al. SMPDB 2.0: Big improvements to the small molecule pathway database. Nucleic Acids Res 2014;42(D1):D478-D484.

(116) Cain AJ. Logic and memory in Linnaeus's system of taxonomy. Proc Linn Soc London 1958;169:114-163.

(117) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nat Genet 2000;25(1):25-29.

(118) Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol 2010;28(9):935-942.

(119) Petri V, Jayaraman P, Tutaj M, Hayman GT, Smith JR, De Pons J, et al. The pathway ontology - updates and applications. J Biomed Semant 2014;5(1).

(120) Schriml LM, Mitraka E. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome 2015;26(9-10):584-589.

(121) Rogers FB. Medical subject headings. Bull Med Libr Assoc 1963;51:114-116.

(122) Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167-339.

(123) The BGS Rock Classification Scheme. Available at: http://www.bgs.ac.uk/bgsrcs/, 2013.

(124) Cottingham N, Greenwood D. An introduction to the standard model of particle physics: Second edition. An Introduction to the Standard Model of Particle Physics: Second Edition; 2007. p. 1-272.

(125) A. A. Kulsherestha. Physiochemical studies of some compoundsSaurastha University; 2009.

(126) Ehrich DG, Lundgren JP, Dionne RA, Nicoll BK, Hutter JW. Comparison of triazolam, diazepam, and placebo as outpatient oral premedication for endodontic patients. J Endod 1997;23(3):181-184.

(127) HMDB: TG(20:0/20:0/20:0) (HMDB05414). 2017; Available at: http://www.hmdb.ca/metabolites/HMDB05414. Accessed 01/20, 2017.

(128) UniProtKB: Q53EU6 (GPAT3_HUMAN). 2017; Available at: http://www.uniprot.org/uniprot/Q53EU6, 2017.

(129) Xianlin H. Lipidomics: Comprehensive Mass Spectrometry of Lipids. : John Wiley & Sons, Inc; 2016.

(130) Lawson CM, Daley BJ, Long CA, Bollig R. Introduction to metabolism. Surgical Metabolism: The Metabolic Care of the Surgical Patient; 2014. p. 1-21.

(131) Nakahigashi K, Toya Y, Ishii N, Soga T, Hasegawa M, Watanabe H, et al. Systematic phenome analysis of Escherichia coli multiple-knockout mutants reveals hidden reactions in central carbon metabolism. Mol Syst Biol 2009;5.

(132) Larrouy-Maumus G, Biswas T, Hunt DM, Kelly G, Tsodikov OV, De Carvalho LPS. Discovery of a glycerol 3-phosphate phosphatase reveals glycerophospholipid polar head recycling in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 2013;110(28):11320-11325.

(133) McEachran AD, Sobus JR, Williams AJ. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem 2016:1-7.

(134) Delaney KA, Kleinschmidt KC. Biochemical and Metabolic Principles. Goldfrank's Toxicologic Emergencies. Ninth Edition ed.: McGraw-Hill Professional; 2010. p. 170.

(135) Kalgutkar AS, Gardner I, Obach RS, Shaffer CL, Callegari E, Henne KR, et al. A comprehensive listing of bioactivation pathways of organic functional groups. Curr Drug Metab 2005;6(3):161-225.

(136) Parkinson A, Ogilvie BW. Biotransformation of Xenobiotics. Casarett & Doull's Essentials of Toxicology. Second Edition ed.: McGraw-Hill Education; 2010. p. 133.

(137) Hill DA, Artis D. Intestinal bacteria and the regulation of immune cell homeostasis. Annu Rev Immunol 2010;28:623-667.

(138) Sousa T, Paterson R, Moore V, Carlsson A, Abrahamsson B, Basit AW. The gastrointestinal microbiota as a site for the biotransformation of drugs. Int J Pharm 2008;363(1-2):1-25.

(139) Schroeder BO, Bäckhed F. Signals from the gut microbiota to distant organs in physiology and disease. Nat Med 2016;22(10):1079-1089.

(140) Béres NJ, Sziksz E, Vannay Á, Szabó D, Pap D, Veres-Székely A, et al. Role of the microbiome in celiac disease. Intl J Celiac Dis 2014;2(4):150-153.

(141) Singanayagam A, Ritchie AI, Johnston SL. Role of microbiome in the pathophysiology and disease course of asthma. Curr Opin Pulm Med 2017;23(1):41-47.

(142) Drug Metabolism Prediction. First Edition ed.: Wiley-VCH Verlag GmbH & Co. KGaA; 2014.

(143) Greene N, Judson PN, Langowski JJ, Marchant CA. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res 1999;10(2-3):299-314.

(144) Testa B, Balmat A-, Long A, Judson P. Predicting drug metabolism - An evaluation of the expert system METEOR. Chem Biodiversity 2005;2(7):872-885.

(145) Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T, et al. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. J Med Chem 2005;48(22):6970-6979.

(146) Rydberg P, Gloriam DE, Olsen L. The SMARTCyp cytochrome P450 metabolism prediction server. Bioinformatics 2010;26(23):2988-2989.

(147) Kirchmair J, Göller AH, Lang D, Kunze J, Testa B, Wilson ID, et al. Predicting drug metabolism: Experiment and/or computation? Nat Rev Drug Discov 2015;14(6):387-404.

(148) COMPUDRUG: Metabolexpert. 2013; Available at: http://www.compudrug.com/metabolexpert, 2017.

(149) Hennemann M, Friedl A, Lobell M, Keldenich J, Hillishch A, Clark T, et al. Cypscore: Qunantitiative Predication of Reactivity toward Cytochromes P450 Based on Semiempirical Molecular Orbil theory. ChemMedChem 2009;4(4):657-669.

(150) Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A, et al. Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. J Chem Inf Model 2012;52(3):617-648.

(151) BIOVIA Metabolite. 2017; Available at: http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/biovia-metabolite.html, 2017.

(152) Preissner S, Kroll K, Dunkel M, Senger C, Goldsobel G, Kuzman D, et al. SuperCYP: A comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. Nucleic Acids Res 2009;38(SUPPL.1).

(153) Spjuth O, Rydberg P, Willighagen EL, Evelo CT, Jeliazkova N. XMetDB: An open access database for xenobiotic metabolism. J Cheminformatics 2016;8(1).

(154) S. E. Adams. Molecular Similarity and Xenobiotic MetabolismUniversity of Cambridge; 2010.

(155) Zaretzki J, Rydberg P, Bergeron C, Bennett KP, Olsen L, Breneman CM. RS-predictor models augmented with SMARTCyp reactivities: Robust metabolic regioselectivity predictions for nine CYP isozymes. J Chem Inf Model 2012;52(6):1637-1659.

(156) Terfloth L, Bienfait B, Gasteiger J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. J Chem Inf Model 2007;47(4):1688-1701.

(157) Mishra NK, Agarwal S, Raghava GPS. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. BMC Pharmacol 2010;10.

(158) Rostkowski M, Spjuth O, Rydberg P. WhichCyp: Prediction of cytochromes P450 inhibition. Bioinformatics 2013;29(16):2051-2052.

(159) MetaPrint2D-React: metabolic product predictor. 2011; Available at: http://www-metaprint2d.ch.cam.ac.uk/metaprint2d-react. Accessed 01/20, 2017.

(160) Liu R, Liu J, Tawa G, Wallqvist A. 2D SMARTCyp reactivity-based site of metabolism prediction for major drug-metabolizing cytochrome P450 enzymes. J Chem Inf Model 2012;52(6):1698-1712.

(161) Fenner K, Gao J, Kramer S, Ellis L, Wackett L. Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. Bioinformatics 2008;24(18):2079-2085.

(162) Pelander A, Tyrkkö E, Ojanperä I. In silico methods for predicting metabolism and mass fragmentation applied to quetiapine in liquid chromatography/time-of-flight mass spectrometry urine drug screening. Rapid Commun Mass Spectrom 2009;23(4):506-514.

(163) Anari MR, Baillie TA. Bridging cheminformatic metabolite prediction and tandem mass spectrometry. Drug Discov Today 2005;10(10):711-717.

(164) Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 1989;246(4926):64-71.

(165) Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. J Cheminformatics 2013;5(3).

(166) Hertz HS, Hites RA, Biemann K. Identification of mass spectra by computer-searching a file of known spectra. Anal Chem 1971;43(6):681-691.

(167) McLafferty FW, Hertel RH, Villwock RD. Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures. Org Mass Spectrosc 1974;9(7):690-702.

(168) McLafferty FW, Zhang M-, Stauffer DB, Loh SY. Comparison of algorithms and databases for matching unknown mass spectra. J Am Soc Mass Spectrom 1998;9(1):92-95.

(169) Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Spectrom 1994;5(9):859-866.

(170) Zhou B, Cheema AK, Ressom HW. SVM-based spectral matching for metabolite identification. Conf Proc IEEE Eng Med Biol Soc 2010:756-759.

(171) Demuth W, Karlovits M, Varmuza K. Spectral similarity versus structural similarity: Mass spectrometry. Anal Chim Acta 2004;516(1-2):75-85.

(172) Kwok K-, Venkataraghavan R, McLafferty FW. Computer-aided interpretation of mass spectra. III. A self-training interpretive and retrieval system. J Am Chem Soc 1973;95(13):4185-4194.

(173) ThermoFisher Scientific - Mass Frontier Spectral Interpretation Software. 2017; Available at: https://www.thermofisher.com/order/catalog/product/IQLAAEGABOFAGUMZZZ, 2017.

(174) Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanalytical Rev 2010;2(1):23-60.

(175) Hufsky F, Scheubert K, Böcker S. Computational mass spectrometry for small-molecule fragmentation. TrAC Trends Anal Chem 2014;53:41-48.

(176) Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, et al. FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass Spectrom 2008;22(19):3043-3052.

(177) Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinform 2010;11.

(178) Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 2014;11(1):98-110.

(179) Kerber A, Meringer M, Rücker C. CASE via MS: Ranking structure candidates by mass spectra. Croat Chem Acta 2006;79(3):449-464.

(180) Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A 2015;112(41):12580-12585.

(181) Hufsky F, Böcker S. Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. Mass Spectrom Rev 2016.

(182) Ruttkies C(2), Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. J Cheminformatics 2016;8(1).

(183) Ridder L, Van Der Hooft JJJ, Verhoeven S, De Vos RCH, Van Schaik R, Vervoort J. Substructure-based annotation of high-resolution multistage MSn spectral trees. Rapid Commun Mass Spectrom 2012;26(20):2461-2471.

(184) Kangas LJ, Metz TO, Isaac G, Schrom BT, Ginovska-Pangovska B, Wang L, et al. In silico identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids. Bioinformatics 2012;28(13):1705-1713.

(185) Böcker S, Rasche F. Towards de novo identification of metabolites by analyzing tandem mass spectra. Bioinformatics 2008;24(16):i49-i55.

(186) Heinonen M, Shen H, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics 2012;28(18):2333-2341.

(187) Gasteiger J, Hanebeck W, Schulz K-. Prediction of mass spectra from structural information. Journal of Chemical Information and Computer Science® 1992;32:264-271.

(188) Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: Decomposing isotope patterns for metabolite identification. Bioinformatics 2009;25(2):218-224.

(189) Kind T, Okazaki Y, Saito K, Fiehn O. LipidBlast templates as flexible tools for creating new in-silico tandem mass spectral libraries. Anal Chem 2014;86(22):11024-11027.

(190) Fridman Noy N., Hafner C.D. The state of the art in ontology design. AI Magazine 1997;18:53-74.

(191) Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies 1995 11;43(5–6):907-928.

(192) Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: A functional perspective. Brief Bioinform 2015;16(6):1069-1080.

(193) Gell-Mann M, Ne'eman Y. The Eightfold way. New York: W.A. Benjamin; 1964.

(194) Malyuto V, Shvelidze T. The technique of automatic quantitative stellar spectral classification using stepwise linear regression. Astrophysics and Space Science 1989;155(1):71-83.

(195) Singh HP, Gulati RK, Gupta R. Stellar spectral classification using principal component analysis and artificial neural networks. Monthly Notices of the Royal Astronomical Society 1998;295(2):312-318.

(196) The Anatomical Therapeutic Chemical (ATC) classification system: Structure and principles. 2011; Available at: http://www.whocc.no/atc/structure_and_principles/. Accessed 04/20, 2013.

(197) Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 1988;28:31-36.

(198) Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, et al. Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 2009;50(SUPPL.):S9-S14.

(199) Fliri AF, Loging WT, Thadeio PF, Volkmann RA. Biological spectra analysis: Linking biological activity profiles to molecular structure. Proc Natl Acad Sci U S A 2005;102(2):261-266.

(200) Hastings J, De Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. Nucleic Acids Res 2013;41(D1):D456-D463.

(201) Moreno P, Beisken S, Harsha B, Muthukrishnan V, Tudose I, Dekker A, et al. BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. BMC Bioinform 2015;16(1).

(202) Zhukova A, Sherman DJ. Knowledge-based Generalization of Metabolic Models. J Comput Biol 2014;21(7):534-547.

(203) Bremser W. Hose - a novel substructure code. Anal Chim Acta 1978;103(4):355-365.

(204) Feldman HJ, Dumontier M, Ling S, Haider N, Hogue CWV. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. FEBS Lett 2005;579(21):4685-4691.

(205) Bobach C, Böhme T, Laube U, Püschel A, Weber L. Automated compound classification using a chemical ontology. Journal of Cheminformatics 2012;4(12).

(206) Vargyas M, Papp J, Csizmadia F, Csepregi S, Papp Á, Vadász P. Maximum Common Substructure Based Hierarchical Clustering. 2008; Available at: http://www.chemaxon.com/library/maximum-common-substructure-based-hierarchical-clustering-2/, 2008.

(207) Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM. Small Molecule Subgraph Detector (SMSD) toolkit. Journal of Cheminformatics 2009;1(1).

(208) Ertl P, Schuffenhauer A, Renner S. The scaffold tree: an efficient navigation in the scaffold universe. Methods Mol Biol 2011;672:245-260.

(209) Chepelev LL, Hastings J, Ennis M, Steinbeck C, Dumontier M. Self-organizing ontology of biochemically relevant small molecules. BMC Bioinformatics 2012;13.

(210) Hastings J, Magka D, Batchelor C, Duan L, Stevens R, Ennis M, et al. Structure-based classification and ontology in chemistry. Journal of Cheminformatics 2012;4(4).

(211) LIPID MAPS Lipidomics Gateway, a free resource sponsored by the National Institute of General Medical Sciences. 2016; Available at: http://www.lipidmaps.org/.

(212) Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251-1255.

(213) Day-Richter J, Harris MA, Haendel M, Clark JI, Ireland A, Lomax J, et al. OBO-Edit - An ontology editor for biologists. Bioinformatics 2007;23(16):2198-2200.

(214) Goodacre SC, Street LJ, Hallett DJ, Crawforth JM, Kelly S, Owens AP, et al. Imidazo[1,2-a]pyrimidines as functionally selective and orally bioavailable GABAAa2/a3 binding site agonists for the treatment of anxiety disorders. J Med Chem 2006;49(1):35-38.

(215) National Institute of General Medical Sciences. 2016; Available at: https://www.nigms.nih.gov/Pages/default.aspx.

(216) National Institute of Health. Available at: https://www.nih.gov/, 2016.

(217) Introducing JSON: ECMA-404 The JSON Data Interchange Standard. Available at: http://www.json.org, 2012.

(218) Dalby A., Nourse J.G., Douglas Hounshell W., Gushurst A.K.I., Grier D.L., Leland B.A., et al. Description of several chemical structure file formats used by computer programs developed at molecular design limited. J Chem Inf Comput Sci 1992;32(3):244-255.

(219) Y. Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) Files. 2005; Available at: http://www.ietf.org/rfc/rfc4180.txt#page-1.

(220) Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2012;40(D1):D742-D753.

(221) PubMed Health [Internet]. Bethesda (MD): National Library of Medicine (US). 2011 Jan 1; Available at: http://www.ncbi.nlm.nih.gov/pubmedhealth/.

(222) An end-to-end search and analytics platform. infinitely versatile. 2015; Available at: http://www.elasticsearch.org/overview/.

(223) Guo AC, Jewison T, Wilson M, Liu Y, Knox C, Djoumbou Y, et al. ECMDB: The E. coli Metabolome Database. Nucleic Acids Res 2013;41(D1):D625-D630.

(224) Wishart DS. FooDB: The Food Database. FooDB Version 1.0. Available at: http://foodb.ca.

(225) Ramirez-Gaona M, Marcu A, Pon A, Guo AC, Sajed T, Wishart NA, et al. YMDB 2.0: A significantly expanded version of the yeast metabolome database. Nucleic Acids Res 2017;45(D1):D440-D445.

(226) Nelson DL, Cox MM. Lehninger Principles of Biochemistry. sixth ed.: W H Freeman & Co (Sd); 2012.

(227) Holloszy JO, Oscai LB, Don IJ, Molé PA. Mitochondrial citric acid cycle and related enzymes: Adaptive response to exercise. Biochem Biophys Res Commun 1970;40(6):1368-1373.

(228) Berg JM, Tymoczko JL, Stryer L. Gluconeogenesis and Glycolysis Are Reciprocally Regulated. Biochemistry. 5th edition ed. New York: W H Freeman; 2002.

(229) Harwood JL, Gunstone FD, Dijkstra A,J. The Lipid Handbook with CD-ROM. : CRC Press; 2007.

(230) Lednicer D. Steroid Chemistry at a Glance. : Wiley; 2010.

(231) Arora B, Mukherjee J, Nath Gupta M. Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology. Sustainable Chemical Processes 2014;2(25).

(232) Testa B, Pedretti A, Vistoli G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. Drug Discov Today 2012;17(11-12):549-560.

(233) Dueñas M, Muñoz-González I, Cueva C, Jiménez-Girón A, Sánchez-Patán F, Santos-Buelga C, et al. A survey of modulation of gut microbiota by dietary polyphenols. BioMed Res Int 2015;2015.

(234) Koppel N, Rekdal VM, Balskus EP. Chemical transformation of xenobiotics by the human gut microbiota. Science 2017;356(6344):1246-1257.

(235) Coleman S, Linderman R, Hodgson E, Rose RL. Comparative metabolism of chloroacetamide herbicides and selected metabolites in human and rat liver microsomes. Environ Health Perspect 2000;108(12):1151-1157.

(236) Wishart DS. Computational strategies for metabolite identification in metabolomics. Bioanalysis 2009;1(9):1579-1596.

(237) Hubert J, Nuzillard J-, Renault J-. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? Phytochem Rev 2017;16(1):55-95.

(238) Zaretzki J, Matlock M, Swamidass SJ. XenoSite: Accurately predicting cyp-mediated sites of metabolism with neural networks. J Chem Inf Model 2013;53(12):3373-3383.

(239) Ellis LB, Gao J, Fenner K, Wackett LP. The University of Minnesota pathway prediction system: predicting metabolic logic. Nucleic Acids Res 2008;36(Web Server issue):W427-432.

(240) Ridder L, Wagener M. SyGMa: Combining expert knowledge and empirical scoring in the prediction of metabolites. ChemMedChem 2008;3(5):821-832.

(241) StarDrop: P450 metabolism. Quantum mechanical simulation of drug metabolism. 2017; Available at: http://www.optibrium.com/stardrop/stardrop-p450-models.php, 2017.

(242) Kirchmair J, Williamson MJ, Afzal AM, Tyzack JD, Choy APK, Howlett A, et al. FAst MEtabolizer (FAME): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. J Chem Inf Model 2013;53(11):2896-2907.

(243) PhytoHub. 2017; Available at: http://phytohub.eu, 2017.

(244) Wishart DS. ContaminantDB. 2017; Available at: http://contaminantdb.ca, 2017.

(245) Miners JO, Smith PA, Sorich MJ, McKinnon RA, Mackenzie PI. Predicting Human Drug Glucuronidation Parameters: Application of In Vitro and In Silico Modeling Approaches. Annu Rev Pharmacol Toxicol 2004;44:1-25.

(246) Jančová P, Šiller M. Topics on Drug Metabolism - Phase II Drug Metabolism. In: Paxton J, editor. ; 2012.

(247) Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 2012;92(4):414-417.

(248) Karapetyan K, Batchelor C, Sharpe D, Tkachenko V, Williams AJ. The chemical validation and standardization platform (CVSP): Large-scale automated validation of chemical structure datasets. J Cheminformatics 2015;7(1).

(249) Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminformatics 2016;8(1):1-20.

(250) Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31(13):3784-3788.

(251) Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, et al. BRENDA in 2017: New perspectives and new tools in BRENDA. Nucleic Acids Res 2017;45(D1):D380-D388.

(252) Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: The universal protein knowledgebase. Nucleic Acids Res 2017;45(D1):D158-D169.

(253) Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353-D361.

(254) International Union of Biochemistry and Molecular Biology - IUBMB Nomenclature Committee Recommendations. 2017; Available at: http://www.chem.qmul.ac.uk/iubmb/, 2017.

(255) Burapan S, Kim M, Han J. Demethylation of Polymethoxyflavones by Human Gut Bacterium, Blautia sp. MRG-PMF1. J Agric Food Chem 2017;65(8):1620-1629.

(256) Selma MV, Espín JC, Tomás-Barberán FA. Interaction between phenolics and gut microbiota: Role in human health. J Agric Food Chem 2009;57(15):6485-6501.

(257) Ozdal T, Sela DA, Xiao J, Boyacioglu D, Chen F, Capanoglu E. The reciprocal interactions between polyphenols and gut microbiota and effects on bioaccessibility. Nutrients 2016;8(2).

(258) Button WG, Judson PN, Long A, Vessey JD. Using Absolute and Relative Reasoning in the Prediction of the Potential Metabolism of Xenobiotics. J Chem Inf Comput Sci 2003;43(5):1371-1377.

(259) Chen C-. Activation and detoxification enzymes: Functions and implications. Activation and Detoxification Enzymes: Functions and Implications; 2013. p. 1-177.

(260) Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition; 1995.

(261) Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics 2017;9(1).

(262) Jeliazkova N, Kochev N. AMBIT-SMARTS: Efficient searching of chemical structures and fragments. Mol Informatics 2011;30(8):707-720.

(263) Carbonaro M, Di Venere A, Filabozzi A, Maselli P, Minicozzi V, Morante S, et al. Role of dietary antioxidant (-)-epicatechin in the development of ß-lactoglobulin fibrils. Biochim Biophys Acta Proteins Proteomics 2016;1864(7):766-772.

(264) Aydin E, Türkez H, Keles MS. The effect of carvacrol on healthy neurons and N2a cancer cells: Some biochemical, anticancerogenicity and genotoxicity studies. Cytotechnology 2014;66(1):149-157.

(265) Wang H, Wang N, Wang B, Zhao Q, Fang H, Fu C, et al. Antibiotics in Drinking Water in Shanghai and Their Contribution to Antibiotic Exposure of School Children. Environ Sci Technol 2016;50(5):2692-2699.

(266) Cyplik P, Marecik R, Piotrowska-Cyplik A, Olejnik A, Drozdzynska A, Chrzanowski L. Biological denitrification of high nitrate processing wastewaters from explosives production plant. Water Air Soil Pollut 2012;223(4):1791-1800.

(267) Basheer C, Alnedhary AA, Rao BSM, Lee HK. Determination of organophosphorous pesticides in wastewater samples using binary-solvent liquid-phase microextraction and solid-phase microextraction: A comparative study. Anal Chim Acta 2007;605(2):147-152.

(268) Takagaki A, Nanjo F. Bioconversion of (-)-epicatechin, (+)-epicatechin, (-)-catechin, and (+)-catechin by (-)-epigallocatechin-metabolizing bacteria. Biol Pharm Bull 2015;38(5):789-794.

(269) Giudice LC. Environmental toxicants: hidden players on the reproductive stage. Fertil Steril 2016;106(4):791-794.

(270) Carmody RN, Turnbaugh PJ. Host-microbial interactions in the metabolism of therapeutic and diet-derived xenobiotics. J Clin Invest 2014;124(10):4173-4181.

(271) Ridlon JM, Harris SC, Bhowmik S, Kang D-, Hylemon PB. Consequences of bile salt biotransformations by intestinal bacteria. Gut Microbes 2016;7(1):22-39.

(272) Ghazalpour A, Cespedes I, Bennett BJ, Allayee H. Expanding role of gut microbiota in lipid metabolism. Curr Opin Lipidology 2016;27(2):141-147.

(273) Lynn K-, Cheng M-, Chen Y-, Hsu C, Chen A, Lih TM, et al. Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. Anal Chem 2015;87(4):2143-2151.

(274) Allard P-, Péresse T, Bisson J, Gindro K, Marcourt L, Pham VC, et al. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. Anal Chem 2016;88(6):3317-3323.

(275) Dias DA, Jones OAH, Beale DJ, Boughton BA, Benheim D, Kouremenos KA, et al. Current and future perspectives on the structural identification of small molecules in biological systems. Metabolites 2016;6(4).

(276) Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, et al. Strategies to characterize polar organic contamination in wastewater: Exploring the capability of high resolution mass spectrometry. Environ Sci Technol 2014;48(3):1811-1818.

(277) Tsugawa H, Ikeda K, Tanaka W, Senoo Y, Arita M, Arita M. Comprehensive identification of sphingolipid species by in silico retention time and tandem mass spectral library. J Cheminformatics 2017;9(1).

(278) Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, et al. Critical Assessment of Small Molecule Identification 2016: automated methods. J Cheminformatics 2017;9(1).

(279) Ma Y, Kind T, Vaniya A, Gennity I, Fahrmann JF, Fiehn O. An in silico MS/MS library for automatic annotation of novel FAHFA lipids. J Cheminformatics 2015;7(1).

(280) mzCloud - Advanced Masss Spectral Database. 2017; Available at: https://www.mzcloud.org, 2017.

(281) Pi J, Wu X, Feng Y. Fragmentation patterns of five types of phospholipids by ultra-high-performance liquid chromatography electrospray ionization quadrupole time-of-flight tandem mass spectrometry. Anal Methods 2016;8(6):1319-1332.

(282) Han X. Lipidomics: Comprehensive Mass Spectrometry of Lipids. Lipidomics: Comprehensive Mass Spectrometry of Lipids; 2016. p. 1-466.

(283) Pence HE, Williams A. Chemspider: An online chemical information resource. J Chem Educ 2010;87(11):1123-1124.

(284) McEachran AD, Sobus JR, Williams AJ. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem 2017;409(7):1729-1735.

(285) Ramirez-Gaona M, Marcu A, Pon A, Guo AC, Sajed T, Wishart NA, et al. YMDB 2.0: A significantly expanded version of the yeast metabolome database. Nucleic Acids Res 2017;45(D1):D440-D445.

(286) Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 2016;34(8):828-837.

(287) Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, et al. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. Phytochemistry 2012;82:38-45.

(288) Dodder NG. Organic/Biological Mass Spectrometry Data Analysis. Available at: https://orgmassspec.github.io, 2017.

(289) Murphy RC. Tandem Mass Spectrometry of Lipids: Molecular Analysis of Complex Lipids. : Royal Society of Chemistry; 2014.