Bayesian Methods for Small Molecule Identification

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena
von M. Sc. Marcus Ludwig
geboren am 08. April 1989 **in** Erfurt, Deutschland

Gutachter:

1. **Prof. Dr. Sebastian Böcker, Friedrich-Schiller-Universität Jena**
2. **Prof. Dr. Juho Rousu, Aalto-Universität Espoo/Helsinki, Finnland**
3. **Prof. Dr. Nicola Zamboni, ETH Zürich, Schweiz**

Tag der mündlichen Verteidigung: 30. Juni 2020

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706

Abstract

Confident identification of small molecules remains a major challenge in untargeted metabolomics, environmental science, natural products research and related fields. Small molecules are important for biomarker research, screening of pollutants, the elucidation of metabolic networks of organisms, drug development and many further applications. Mass spectrometry is the predominant technique for the high-throughput analysis of small molecules and can detect thousands of different compounds in a biological sample. Mass spectrometry measures the mass of a compound. Tandem mass spectrometry subsequently fragments the compound and measures the mass of its fragments. The automated interpretation of the resulting tandem mass spectra is highly non-trivial. Hence, many studies are limited to re-discovering known compounds by searching mass spectra in spectral reference libraries. But these libraries are vastly incomplete and a large portion of measured compounds remains unidentified. This constitutes a major bottleneck in the comprehensive, high-throughput analysis of metabolomics data.

In this thesis, we present two computational methods which address different steps in the identification process of small molecules from tandem mass spectra. One method searches mass spectra in a structure database. The other identifies the molecular formula without the need of any database. Both methods share the Bayesian idea to capture dependencies.

To overcome the limitations of small spectral libraries, recent methods search instead in much bigger structure databases. For an unknown compound, CSI:FingerID predicts a molecular fingerprint, which encodes the presence or absence of each property of a set of molecular properties. Then, this predicted fingerprint is used to search in a molecular structure database by scoring it against deterministic fingerprints of candidate structures. For this, current scorings assume independence between the molecular properties.

We introduce a novel scoring for CSI:FingerID which models dependencies between different molecular properties via Bayesian networks. The scoring is interesting from a theoretical perspective since we apply a novel strategy to estimate conditional probabilities. The marginal probabilities of random variables come from the predicted fingerprint, and hence, change for each compound. For random variables connected in the network, we compute expected covariances and use these to estimate the conditional probabilities. Modeling dependencies improves identification rates of CSI:FingerID by 2.85 percentage points.

Annotating the molecular formula of a compound is the first step in its structural elucidation. However, confident annotation remains challenging, in particular for large compounds above 500 Daltons. ZODIAC is a novel method for *de novo* — that is, database-independent — molecular formula annotation in complete datasets. It exploits similarities of compounds co-occurring in a sample to find the most likely molecular formula for each individual compound. ZODIAC takes molecular formula candidates as input and reranks these candidates by considering joint fragments and losses. We use Gibbs sampling and Bayesian statistics to estimate posterior probabilities. We evaluate on five diverse datasets and find that ZODIAC considerably improves molecular formula annotations. For one dataset from plant extract, ZODIAC reduces incorrect annotations 16.5-fold. Furthermore, the ZODIAC score allows to assess the confidence in each annotation which enables to select high-quality annotations in an automated fashion. We show that *de novo* molecular formula annotation is not just a theoretical advantage: We discover multiple novel molecular formulas absent from PubChem, one of the biggest structure databases.

Both methods have been integrated into the SIRIUS software for small molecule identification.

Zusammenfassung

Die zuverlässige Identifizierung kleiner Moleküle bleibt eine große Herausforderung in der Metabolomik, Umwelt- und Naturstoff-Forschung und verwandten Forschungsgebieten. Kleine Moleküle sind von zentraler Bedeutung für die Suche nach Biomarkern, das Schadstoff-Screening, die Aufklärung von metabolischen Netzwerken von Organismen, Medikamentenentwicklung und für viele weitere Anwendungen. Massenspektrometrie ist die vorherrschende Technik für die Hochdurchsatzanalyse kleiner Moleküle und kann tausende unterschiedliche chemische Verbindungen in einer biologischen Probe detektieren. Massenspektrometrie misst die Masse einer chemischen Verbindung. Anschließend fragmentiert Tandem-Massenspektrometrie die chemische Verbindung und misst die Masse der Fragmente. Die automatisierte Auswertung der resultierenden Tandem-Massenspektren ist hoch kompliziert. Daher beschränken sich die meisten Studien darauf, bereits bekannte chemische Verbindungen wiederzuentdecken, indem sie Massenspektren in Referenzspektrendatenbanken suchen. Allerdings sind diese Datenbanken erheblich unvollständig und ein großer Teil der gemessenen chemischen Verbindungen bleibt unidentifiziert. Das stellt ein großes Problem für die umfassende Hochdurchsatzanalyse von Metabolomikdaten dar.

In dieser Dissertation präsentieren wir zwei computergestützte Methoden, die sich mit unterschiedlichen Schritten in der Identifizierung kleiner Moleküle mit Hilfe von Tandem-Massenspektren beschäftigen. Eine Methode sucht Massenspektren in einer Molekülstrukturdatenbank. Die andere identifiziert Molekülformeln ohne die Verwendung jeglicher Datenbanken. Beide Methoden teilen die bayessche Idee zur Erfassung und Beschreibung von Abhängigkeiten.

Um Einschränkungen kleiner Spektrendatenbanken zu überwinden, suchen aktuelle Methoden stattdessen in den viel größeren Molekülstrukturdatenbanken. Für eine unbekannte chemische Verbindung sagt CSI:FingerID einen molekularen Fingerabdruck vorher. Dieser Fingerabdruck kodiert das Vorkommen oder Fehlen von jeder einzelnen Moleküleigenschaft aus einer Menge von Moleküleigenschaften. Dieser vorhergesagte Fingerabdruck wird dann in einer Molekülstrukturdatenbank gesucht, indem man ihn gegen deterministische Fingerabdrücke von Kandidatenstrukturen vergleicht und bewertet. Dafür nehmen aktuelle Scorings Unabhängigkeit zwischen den einzelnen Moleküleigenschaften an. Wir stellen ein neues Scoring für CSI:FingerID vor, welches Abhängigkeiten zwischen unterschiedlichen Moleküleigenschaften mit Hilfe von bayesschen Netzen modelliert. Das Scoring ist aus theoretischer Sicht interessant, da wir eine neue Strategie zur Schätzung der bedingten Wahrscheinlichkeiten anwenden. Die Randwahrscheinlichkeiten der Zufallsvariablen ergeben sich aus dem vorhergesagten Fingerabdruck. Für im Netz verbundene Zufallsvariablen berechnen wir erwartete Kovarianzen und verwenden diese, um die bedingten Wahrscheinlichkeiten zu schätzen. Das Modellieren der Abhängigkeiten verbessert die Identifikationsrate von CSI:FingerID um 2,85 Prozentpunkte.

Die Identifizierung der Molekülformel einer chemischen Verbindung ist der erste Schritt in ihrer Strukturaufklärung. Allerdings bleibt die sichere Identifizierung weiterhin anspruchsvoll, insbesondere für große chemische Verbindungen über 500 Dalton. ZODIAC ist eine neue Methode für die *de novo* — also, datenbankunabhängige — Identifizierung von Molekülformeln in vollständigen Datensätzen. ZODIAC macht sich die Ähnlichkeit von gemeinsam auftretenden chemischen Verbindungen zu Nutze, um die wahrscheinlichste Molekülformel für jede einzelne Verbindung zu finden. ZODIAC bekommt Molekülformelkandidaten als Eingabe und sortiert diese neu nach ihrer Wahrscheinlichkeit, indem gemeinsame Fragmente und Molekekülverluste einbezogen werden. Wir verwenden Gibbs-Sampling und bayessche Statistik, um A-posteriori-Wahrscheinlichkeiten zu schätzen. Wir evaluieren auf fünf vielfältigen Datensätzen und stellen fest, dass ZODIAC die Identifikation von Molekülformeln erheblich verbessert. Auf einem Datensatz eines gemessenen Pflanzenextraktes reduziert ZODIAC falsche Identifizierungen auf ein sechzehntel. Außerdem ermöglicht der ZODIAC Score, die Verlässlichkeit einer Identifizierung abzuschätzen, was die gezielte Auswahl von hochqualitativen Identifizierungen in automatisierter Weise ermöglicht. Wir zeigen, dass *de novo* Molekülformelidentifizierung nicht ausschließlich ein theoretischer Vorteil ist: Wir entdecken mehrere neue Molekülformeln, die nicht in PubChem, einer der größten Molekülstrukturdatenbanken, vorkommen.

Beide Methoden sind in der SIRIUS-Software für die Identifizierung kleiner Moleküle integriert.

Acknowledgements

Finally, my PhD time is coming to an end. Here, I want to take the opportunity to thank all the people who made this thesis possible. First of all, I want to thank my supervisor Sebastian Böcker. He is a source of lots of ideas and insightful discussions. He gave me the confidence and freedom to accomplish this work and also provided me with guidance.

I am grateful to the people I met on this journey. I was in the lucky position to attend two Dagstuhl seminars on computational metabolomics. I had great discussions and insights. Many discussions were continued until late at night, accompanied by a pleasant game of pool; and people still played pool with me patiently, even though I was really bad at it. The list of all people I met at Dagstuhl would be to long to put it here. But Michael Witting, Corey Broeckling and Emma Schymanski certainly taught me about some curiosities of mass spectrometry, which I did not know as a computer scientist. And thank you, Nicola, for your advice.

I was also able to visit Juho Rousu's group at Aalto University in Espoo twice. They are the machine learning experts that laid the foundation for CSI:FingerID, on which I am working in this thesis. It was amazing to see how much can be accomplished for example by linear algebra and matrix multiplications. I am grateful for their hospitality.

This thesis also would not have been possible without the people from Pieter Dorrestein's group in San Diego. They provided the data to develop and evaluate my algorithms. Louis-Félix Nothias invested a lot of time to perform manual annotation on the data and helped with the evaluation. I met many great people from Pieter's lab on a short visit. Thank you, Louis and Mélissa, that I could stay with you.

And I want to continue giving thanks to the people which provide the precious (reference) datasets: thanks to the GNPS community, MassBank, Agilent and NIST. The NIST spectral library really is a treasure of mass spectrometry knowledge, hidden in all the manually annotated spectra.

This work certainly would not have been possible without the many preceding years of research in Sebastian's and Juho's groups. The concept of fragmentation trees now is over ten years old. Florian Rasche invented it together with Sebastian Böcker; and Kai Dührkop improved it in an outstanding fashion, implementing fragmentation tree computation in what now is the SIRIUS software. The machine learning ideas of CSI:FingerID, predicting molecular fingerprints from a tandem mass spectrum using kernel support vector machines, originated from Juho's group with contributions most notably from Markus Heinonen.

I am grateful to my colleagues and friends. It is an amazing work atmosphere, teasing each other in the most kind way. It is great to be a part of all these amazing projects which will have a real impact on computational metabolomics, and seeing our software SIRIUS mature in this process. And thank you for proofreading parts of this thesis. Since we are currently a small group, luckily I can name all of them: Thank you Kai Dührkop, Markus Fleischauer and Martin Hoffmann. And, of course, a special thanks goes to Kathrin Schowtka who steadily navigates us through the jungle of bureaucracy.

I gratefully acknowledge the funding received through the Friedrich Schiller University of Jena and the Deutsche Forschungsgemeinschaft (DFG) within the project “Identifying the Unknowns: Fragmentation Trees and Molecular Fingerprints”.

And mostly, I want to thank my wife, even though writing this down makes me feel a little old. Thank you for your support, thank you for bearing with me. And of course I want to thank my family, you provided me with the support to pursue my studies and this PhD.

Contents

1 Introduction 1

2 Background in Biochemistry and Mass Spectrometry 5
2.1 Molecules 5
2.2 Metabolites 8
2.2.1 Molecular Fingerprints 9
2.3 Mass Spectrometry 10
2.4 Tandem Mass Spectrometry 13
2.5 Chromatography 14

3 Statistics and Graph Theory 17
3.1 Graph Theory 17
3.2 Bayesian Statistics 18
3.3 Bayesian Inference 19
3.4 Gibbs Sampling 20
3.5 Bayesian Networks 21

4 Computational Mass Spectrometry 25
4.1 Data Processing 25
4.2 De Novo Molecular Formula Annotation 26
4.2.1 Isotope Pattern Analysis 26
4.2.2 Fragmentation Trees 27
4.3 Spectral Library Search 28
4.4 Structure Database Search 29
4.4.1 CSI:FingerID 32

5 ZODIAC: De Novo Molecular Formula Annotation 37
5.1 A Similarity Model for Molecular Formula Assignments 38
5.1.1 Posterior Probability of an Assignment 38
5.1.2 Graph-theoretical Formulation 39
5.1.3 Complexity of the Problem 40
5.1.4 Likelihoods, Prior Probabilities and Graph Topology 41
5.1.5 (Faster) Gibbs Sampling 42
5.1.6 Faster network creation 44
5.2 Evaluation on five Biological Samples 45
5.2.1 Datasets 46
5.2.2 Preprocessing 46
5.2.3 SIRIUS Analysis and Establishing a Ground Truth 50
5.2.4 ZODIAC Parameters 51
5.2.5 Parameters of Competing Methods 52

5.2.6 Results 53

6 Bayesian Network Scoring for Molecular Structure Search 63
6.1 Tree-based Posterior Probability Estimation 65
6.2 Finding the Tree and Estimating Covariances 68
6.3 Evaluation on Cross-validation and Independent Dataset 70
6.3.1 Datasets and Databases 70
6.3.2 Results 71

7 Conclusion 75

A Appendix 99

1 Introduction

> What we are concerned with here is the fundamental interconnectedness of all things.
>
> Dirk Gently

The Earth Biogenome Project aims to sequence and annotate 1.5 million species over a period of ten years [117]. This sounds rather ambiguous, but analytical chemistry has made tremendous progress in the last decades to measure the molecules of living organisms. DNA and RNA sequencing has become fast and efficient. Proteins can be measured high-throughput and protein identification greatly benefited from the increased amount of genomic data. The distinct fields of study which are concerned with a specific pool of molecules, with their characterization and quantification, is referred to as “-omics” sciences. Genomics focuses on the exploration of the genome, the organism’s complete set of DNA; transcriptomics focuses on the RNA, all transcripts of the DNA; and proteomics on all proteins, which includes the reaction-catalyzing enzymes. Each field aims at measuring the molecules of interest in a comprehensive fashion and discovering and characterizing functionality.

However, to provide a comprehensive picture of a living organism, we also have to consider *metabolites*, the entirety of small molecules which are intermediates or products of metabolism. The metabolome provides a direct signature of biochemical activity and thus, metabolites are closely linked to phenotype [156, 187]. The recent years showed the importance of metabolomics to understand complex biological interactions and phenomena. For example, the microbiome, the set of all the microorganisms living on and inside us, produces metabolites that affect the chemistry of the host [164]. Such interactions cannot be inferred from the genome. Here, the genome presents a rather static view of a highly dynamic system. The metabolome, on the other hand, changes with what we drink and eat, and even with the skin care products we are using [27].

Metabolomics aims to give a snapshot of the cellular state and can be used to answer diverse biological questions. Biomarker discovery describes the process of identifying measurable indicators for a biological condition, such as the diagnoses of certain health conditions. Metabolomics identifies biomarkers and can reveal novel insights on causes of diseases and therapeutic targets [91, 225]. In precision medicine, metabolomics may be useful to select optimal therapies and monitor individual drug-response [225]. Natural products research investigates all small molecules isolated from natural sources. Their economic importance includes fragrances, herbicides, pesticides and food supplements [102]. Many drugs are natural products or derived from a natural product [144]. Currently, too few novel drug candidates are being identified and investigated; in particular, there is a lack of novel antibiotics: “Declining private investment and lack of innovation in the development of new antibiotics are undermining efforts to combat drug-resistant infections" (World Health Organization, 2020)[1]. It is likely that microorganisms produce many still uncharacterized antibiotics [46]. However, we need novel methods to facilitate their discovery.

The identification of small molecules also plays an important role in many related fields of research. Environmental science screens soil or water for pollutants and toxins [181]. Nutrition and food products are continuously monitored for the sake of food safety and quality assessment [5]. And drug and toxicology screenings are commonly administered in forensics [38] and doping control [200].

Metabolomics studies can be distinguished in targeted and untargeted experiments. Targeted experiments analyze a selected set of known molecules. However, by design, these experiments cannot find anything novel. In order to find novel drugs or biomarkers, unknown compounds have to be structurally elucidated. When screening for pollutants and contaminants, targeted approaches are able to recognize a list of expected molecules; but clearly, transformation products and yet uncharacterized contaminants cannot be identified [181].

Many metabolites remain unknown to us [18, 48]. One reason that makes their identification rather difficult is that metabolites are *not* composed of simple building blocks — as opposed to DNA, RNA and proteins. In fact, small molecules are structurally highly diverse. Furthermore, structural information of metabolites cannot be directly derived from the genome, as it is possible for RNA and proteins. Metabolic gene clusters allow the prediction of metabolite structures to some extend, but this is limited to a very small fraction of all metabolites [17, 140, 188].

There exist two predominant techniques to measure small molecules: nuclear magnetic resonance (NMR) and mass spectrometry (MS). NMR is the method of choice for full structural elucidation of molecules. Unfortunately, it requires high amounts of purified compound and is less suited for high-throughput analysis. MS is highly sensitive and can measure thousand of metabolites from a single biological sample. However, analysis of the data is far from trivial. Mass spectrometry does not provide direct evidence of the molecular structure of a molecule. Instead, mass spectrometry measures the mass of a compound. Tandem mass spectrometry subsequently fragments the compound and measures the mass of its fragments; the resulting data are called tandem mass spectra.

These tandem mass spectra can be used to draw inferences about the molecular structure. The manual interpretation of mass spectra is cumbersome and requires a great amount of expert knowledge. The usual way to identify compounds is to search the mass spectrum in a spectral library of references. However, these libraries are highly incomplete. Depending on the organism, up to 98 % of the measured compounds might not be contained in the reference library and thus remain unidentified [48]. Hence, it is not surprising that "compound identification" is consistently named as one of the biggest challenges in metabolomics to derive biological knowledge from metabolomics studies [135, 209].

Contribution of this Work

I wrote my bachelor and master thesis on topics related to computational mass spectrometry of small molecules. My bachelor thesis was about finding characteristic substructures which are shared between a set of molecular structures [123]; given a list

---

[1] `https://www.who.int/news-room/detail/17-01-2020-lack-of-new-antibiotics-threatens-global-efforts-to-contain-drug-resistant-infections`

of candidate structures for an unknown molecule, this would help to establish a starting point for elucidation. In my master thesis, I developed a method to automatically detect isotope patterns in electron ionization spectra to improve fragmentation tree computation. Later, I helped to develop SIRIUS 4 [63]. Recently, Louis-Félix Nothias and I investigated the prevalence of non-sodiated fragment ions in sodiated ion tandem mass spectra [126] — one example of the many peculiarities of mass spectrometry. I also contributed to further methods for the analysis of small molecules from mass spectrometry data [60, 64, 148, 201].

In this thesis, I focus on two important problems related to the structural elucidation of small molecules from tandem mass spectra: firstly, molecular formula annotation [125] and secondly, searching a tandem mass spectrum in a structure database [124]. I developed these methods in collaboration with my supervisor Sebastian Böcker and great input from my colleagues, in particular Kai Dührkop. I approach both problems using Bayesian methods. Furthermore, I build upon established methods: SIRIUS 4 [63] is arguably the best-performing tool for molecular formula annotation; CSI:FingerID [61] and its related approaches IOKR [29–31] and ADAPTIVE [146] are currently the best-performing methods for structure database search.

Determining the molecular formula of a compound is the first step in its structural elucidation; and this step can already be challenging. Molecular formulas cannot be unambiguously annotated by mass alone, not even when searching in a database and using mass spectrometers with sub-ppm mass accuracy [104]. Besides, if we want to overcome the limitations of searching molecular formulas in a (spectral or structure) database in order to find novel compounds, molecular formula annotation has to be performed *de novo*; that is, we consider all possible molecular formulas. Here, the number of candidates strongly increases with increasing compound mass and when considering elements beyond carbon, hydrogen, oxygen and nitrogen. I present a method called ZODIAC [125], which takes a "holistic" approach to the molecular formula annotation problem: Metabolites co-occur in a network of derivatives; and to annotate one compound, it is helpful to consider similar compounds in the data. ZODIAC uses the top-scoring molecular formula candidates of each compound from SIRIUS and reranks them by considering joint fragments and losses between fragmentation trees. I use Bayesian statistics and Gibbs sampling to estimate posterior probabilities for all molecular formula assignments. Since the number of considered variables can be relatively high, I engineer the algorithm to create a swift Gibbs sampler in practice. I evaluate ZODIAC on five diverse datasets of biological samples. I show that ZODIAC enables confident molecular formula assignment and greatly facilitates the discovery of novel molecular formulas absent from the biggest structure databases.

Next, I present a novel scoring method for CSI:FingerID [124]. CSI:FingerID searches a tandem mass spectrum in a structure database: it predicts a molecular fingerprint from the mass spectrum and compares this against molecular fingerprints of structures in the database. A molecular fingerprint encodes the structure of a molecule: it is a binary vector where each position indicates the presence or absence of a specific molecular property (usually a substructure).

IOKR is another approach strongly related to CSI:FingerID that omits the intermediate step of predicting a molecular fingerprint. IOKR was able to outperform CSI:FingerID [30]. However, molecular fingerprints predicted by CSI:FingerID can be utilized even for applications beyond structure database search, such as compound class prediction [64]. Besides, the predicted fingerprint can be extended with novel properties to improve performance; and the predicted probabilities in the fingerprint can indicate the quality of prediction. Thus, both approaches are interesting for automated small molecule identification and worth considering for further research.

The molecular fingerprints of CSI:FingerID describe thousands of potential molecular properties. Some properties might be more important than others to establish the molecule's structure, and clearly, many of these properties are highly correlated. Current scorings assume that all properties in the fingerprint are independent of each other. I use Bayesian networks to model dependencies between different substructures. On the one hand, this should capture, how much a (predicted) molecular property contributes as new evidence, given we already know other properties of the fingerprint. On the other hand, this aims to capture and reduce mutual errors made by the predictors. I apply Bayesian networks in a non-standard fashion, where the marginal probabilities of the random variables are different for each predicted fingerprint. Here, it is not straightforward to establish the conditional probabilities which are necessary to define the Bayesian network. A substantial part of this scoring deals with how to properly estimate these probabilities. Furthermore, I found that molecular structures that have the same molecular formula, often share the presence or absence of multiple substructures in their fingerprints. With this in mind, I extend the scoring to use individual Bayesian networks for each molecular formula. The novel Bayesian network scoring significantly outperforms the currently best scoring.

I presented ZODIAC at the Metabolomics conference 2017 in Brisbane and the annual conference of the American Society of Mass Spectrometry (ASMS) 2018 in San Diego. I gave a talk on the Bayesian network scoring at the annual international conference on Intelligent Systems for Molecular Biology (ISMB) 2018 in Chicago [124]. Both of these methods are implemented in the SIRIUS software for small molecule identification.

Before I describe the methods in Chapter 5 and 6, I give an introduction to the field of research. Chapter 2 introduces basic chemical knowledge and the analytical technique mass spectrometry. Chapter 3 shortly introduces graph theoretical notations and describes statistical concepts related to Bayesian statistics and sampling methods. In Chapter 4, I give an overview of methods in computational mass spectrometry. I focus on high-resolution, high mass accuracy tandem mass spectrometry data of small molecules.

For the remainder of this thesis, I use "we" as the first person pronoun, as it is common in scientific literature.

2 Background in Biochemistry and Mass Spectrometry

In this chapter, we introduce basic knowledge on small molecules and mass spectrometry which is required to understand this thesis. Firstly, we explain properties of molecules and will focus on small molecules in particular. Secondly, we introduce mass spectrometry as an analytical platform to measure and examine small molecules. The descriptions are a simplification of the chemical and physical background and do by no means attempt to be a comprehensive overview. We refer readers interested in metabolomics to Weckwerth [215]; and for mass spectrometry to Gross [75].

2.1 Molecules

A molecule is a group of two or more atoms connected by bonds. Every atom is composed of a central nucleus surrounded by one or more electrons. The nucleus contains neutrons and protons. The number of protons determines the chemical *element* of an atom. In many notations the element is represented by the element symbol, e.g. C for carbon, H for hydrogen, O for oxygen and N for nitrogen. *Isotopes* are variants of an element and differ in the number of neutrons. Atoms of different isotopes of the same element have the same chemical properties. The total number of neutrons and protons is called *mass number*. By notation, this is specified at the upper left of the element symbol. For example, the most abundant isotope of carbon contains six neutrons and six protons and is represented as ^{12}C. Molecules that only differ in their isotopic composition are called *isotopologues*. The *molecular mass* of a molecule is specified in unified atomic mass units (u) or equivalently in Dalton (Da). One Dalton is defined as $\frac{1}{12}$ of the mass of a ^{12}C isotope, which is approximately $1.660\,539\,067 \times 10^{-27}\,\text{kg}$. The *nominal mass* is the total number of protons and neutrons of a molecule and is given in Da (u) — in contrast, the mass number is unit-free. The calculated *exact mass* of a molecule is the sum of masses of all its atoms. The *mass defect* is the difference between nominal and exact mass. This difference exists mainly due to the different binding energies within the nuclei. Because of the (arbitrary) choice of ^{12}C as reference, it is the only isotope with equal exact and nominal mass. The mass of a molecule with all its atoms being isotopes with the lowest mass is called *monoisotopic mass* [23]. Note, this definition differs from the IUPAC definition that the monoisotopic mass is the mass of the most abundant isotopologue. It is reasonable from a computational point of view to consider the lowest mass as monoisotopic mass; in particular to determine the elemental composition based on mass spectrometry. For most small molecules the lowest-mass isotopologue is also the most abundant. Most biomolecules are composed of the elements CHNO, sulfur and phosphorus, but may also contain halogens or other uncommon elements. See Table 2.1 for elements and isotopes occurring in living organisms.

Table 2.1: Isotopes of elements found in biomolecules. The most common elements are carbon, hydrogen, nitrogen, oxygen, phosphorus and sulfur. Additionally, halogens (bromine, chlorine, fluorine and iodine) are reported. Listed are the elements and their stable isotopes with mass and relative abundance. Values taken from [12].

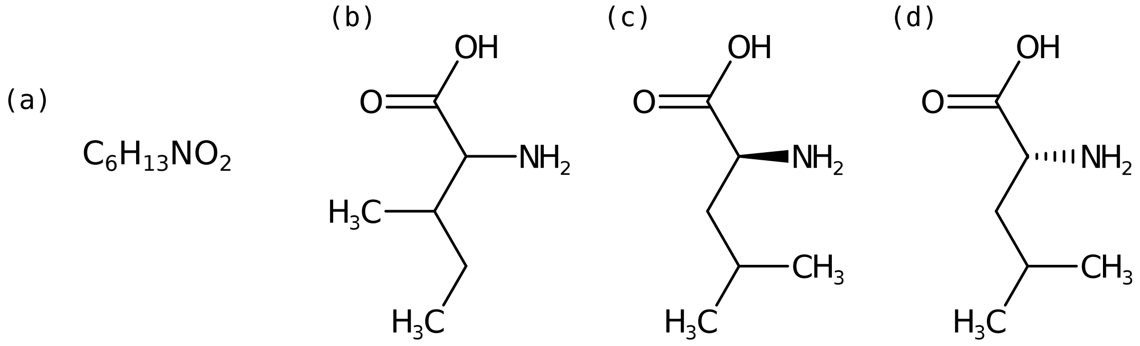
element	symbol	isotope	mass (Da)	abundance (%)
carbon	C	^{12}C	12.0	98.93
		^{13}C	13.003355	1.07
hydrogen	H	^{1}H	1.007825	99.9885
		^{2}H	2.014102	0.0115
nitrogen	N	^{14}N	14.003074	99.636
		^{15}N	15.000109	0.364
oxygen	O	^{16}O	15.994915	99.757
		^{17}O	16.999132	0.038
		^{18}O	17.999160	0.205
phosphorus	P	^{31}P	30.973761	100.00
sulfur	S	^{32}S	31.972071	94.99
		^{33}S	32.971459	0.75
		^{34}S	33.967867	4.25
		^{36}S	35.967081	0.01
bromine	Br	^{79}Br	78.918338	50.69
		^{81}Br	80.916291	49.31
chlorine	Cl	^{35}Cl	34.968853	75.76
		^{37}Cl	36.965903	24.24
fluorine	F	^{19}F	18.998403	100.00
iodine	I	^{127}I	126.904468	100.00

The number of electrons determines the charge of an atom (and molecule): Electrically neutral atoms have equal numbers of protons and electrons. Positively or negatively charged atoms and molecules are referred to as *ions*. Bonds between atoms can be formed by sharing or transferring electrons. This brings the atoms in an energetically favorable state. Electrons that can participate in the formation of bonds are called *valence electrons* and are located in the outer shell of an atom. When atoms share pairs of electrons, they form a *covalent bond*. The electrons positioned between the two atoms' nuclei attract these nuclei (the positively charged protons within), resulting in a chemical bond. Atoms that have a large difference in electronegativity may form an *ionic bond*. Here, an electron is transferred from one atom to the other, which leads to a positive and a negative ion which attract each other by their opposite charges. A typical example is sodium chloride: The sodium loses an electron and becomes a *cation*, a positively charged ion. The chlorine gains an electron and becomes an *anion*, a negatively charged ion. Atoms or molecules that have unpaired valence electrons are *radicals*. These are chemically very reactive and are likely to react and form bonds with other atoms or molecules to reach a chemically stable state. Radicals are intermediates in many chemical reactions.

The *molecular formula* (elemental composition) indicates the number of atoms of each element in a molecule. The *constitution* of a molecule is the number, kind and connectivity of atoms. Molecules with the same molecular formula but different connectivity are called *structural isomers*. Molecules with equal constitution but different three-dimensional orientation of atoms are considered *stereoisomers*. A *structural formula* describes the constitution and may include orientation in space. Take for example the two amino acids leucine and isoleucine: both have the molecular formula $C_6H_{13}NO_2$, but different constitutions (Fig. 2.1). On the other hand, L-leucine and D-leucine are different stereoisomers. In this thesis, we usually consider only the constitution and do not distinguish between stereoisomers since mass spectrometry generally cannot differentiate stereoisomers. We will refer to the constitution of a molecule as *structure*.

A chemical substance consisting of identical molecules composed of atoms of two ore more elements is called *compound*. We will use the words "molecule" and "compound" interchangeably throughout this thesis. Mass spectrometry cannot detect single molecules but only compounds.

Multiple text-based formats exist for the representation of molecules which are optimized for machine-readability and storing the molecules in databases. SMILES [216] and InChI [85] both represent the molecule as a string. The same molecule can be represented by many different SMILES which makes it difficult to compare molecules for identity. Canonical SMILES where introduced to define a unique SMILES representation for each molecule, but for some molecules the algorithm failed this task [141, 150]. Different implementations try to overcome this problem [150], but there is no official standard of canonical SMILES. The IUPAC international chemical identifier (InChI) overcomes this problem by using graph isomorphism algorithms to produce a unique ordering of atoms [85]. In contrast to SMILES it is less human-readable. It organizes information in layers. The first layers specify molecular formula and constitution. Additional layers specify stereochemistry, charges and isotopes. The InChIKey was introduced as a compact identifier for fast comparison and database search. It is a hashed code derived from InChI and is exactly 27 characters long. It is organized in layers, too. The first 14 characters encode for the molecule's constitution. Stereochemical and isotopic information are encoded in the second layer. The third layer encodes protonation or deprotonation.



(a) $C_6H_{13}NO_2$

(e) CC(C)C[C@@H](C(=O)O)N

(f) InChI=1S/C6H13NO2/c1-4(2)3-5(7)6(8)9/h4-5H,3,7H2,1-2H3,(H,8,9)/t5-/m0/s1

(g) ROHFNLRQFUQHCH-YFKPBYRVSA-N

Figure 2.1: Different molecule representations of leucine and isoleucine. (a) Both molecules have the same molecular formula. Structural formulas of (b) isoleucine without stereochemistry and (c) L-leucine and (d) D-leucine with stereochemistry information are displayed. The solid wedged bond points above-the-plane, the dashed wedged bond points below-the-plane. Isoleucine and leucine have different constitutions. L-leucine and D-leucine have the same constitution but different stereochemistry. Different string representations depict L-leucine as (e) SMILES, (f) InChI and (g) InChIKey.

Throughout this thesis, we will use the first 14-character block of InChIKeys to test two molecules for equality (based on their constitution).

SMILES arbitrary target specification (SMARTS)[1] is an extension of SMILES which allows to specify substructural patterns. These patterns can be used to search substructures in molecules. For example "[CX4]" matches any carbon connected to exactly four atoms (including hydrogen).

2.2 Metabolites

Metabolites are considered all molecules in the cells of living organism which are intermediates or products of metabolism. Usually the term is restricted to small molecules — that is molecules below 1000 Da. In contrast to DNA, RNA or proteins, metabolites are not comprised of repetitive structures or simple building blocks. Despite there small size, metabolites are remarkably diverse and exhibit a broad range of chemical properties, covering a great range of different classes such as flavonoids, glycosides, amino acids, nucleic acids and lipids. The term metabolome refers to the complete set of metabolites in a cell, tissue or organism. The metabolome is the closest representation of phenotype [187] and provides a direct snapshot of biochemical activity.

Metabolites are divided into primary and secondary metabolites. Primary metabolites are directly involved in growth, development and reproduction processes. They are generally present in many organisms. Secondary metabolites are not directly involved in these central processes but allow important ecological functions such as defense or signaling.

---

[1] `http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`

Secondary metabolites are much more specific to an individual species or a restricted set of organisms. Many secondary metabolites remain uncharacterized with respect to their function and structure [8, 48, 156].

Natural products are organic compounds isolated from natural sources. Typically this term refers to secondary metabolites. Their economic importance includes fragrances, herbicides, pesticides and food supplements [102]. Natural products remain a major source of novel drugs [144].

The lipidome as a subset of the metabolome refers to the set of all lipids. Lipidomics is a distinct field of research due to the lipids functional specificity. Lipids have a much more regular structure compared to the set of all metabolites. Because of their chemical properties, lipids can be measured and investigated in a more tailored fashion.

2.2.1 Molecular Fingerprints

Molecular fingerprints are a common way of encoding molecular structures for computational processing. A fingerprint can be seen as bit vector: Every position encodes for the presence or absence of a specific *molecular property*, such as a substructure of the molecule. This is not a bijective relationship, meaning that different molecules can have the same fingerprint. For this reason, a molecule cannot be reconstructed from its fingerprint. Substructure properties can be described by SMARTS patterns.

One reason for the popularity of molecular fingerprints is that they allow for an efficient way to compare structures. This is of particular interest in virtual screening where one or multiple query molecules are searched against a database of millions of molecular structures [221]. The screening usually selects a small subset of candidates for further analysis. This can be either additional computational analysis, such as substructure search or the calculation of the largest common substructure shared between molecules. Especially the second task is too complex to be performed on the whole database [168]. But also additional analytical analysis, such as assays to test bioactivity, are too expensive and time-consuming to be applied to all available molecules.

The classification of molecular structures which exhibit similar biological effects or physicochemical properties, has been widely adopted in drug discovery [13]. In ligand-based screening, a structure database is searched for molecular structures similar to a molecule with known bioactivity, in order to select a set of molecules which is enriched for this desired activity [128]. This is based on the assumption of the similar properties principles: molecules with similar structures are likely to have similar properties [26, 96, 100, 219]. Quantitative structure–activity relationship (QSAR) modeling is used to directly predict biochemical properties based on the molecular structure. Early works have been reported since the 1970s [1, 222]. For an overview on QSAR modeling see Dudek *et al.* [59].

A widely adopted function to compare molecular fingerprints is the *Tanimoto coefficient* (also known as Jaccard index) [220]:

$$Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B are the sets of substructures present in the fingerprints of two molecules. Other measures are the cosine similarity and Dice coefficient, which often perform similarly in ranking tasks [7]. For comparisons of different similarity measures see Bajusz *et al.* [7] and Cereto-Massagué *et al.* [39].

Many different types of fingerprints exist, most of which are based on substructures without stereochemical information (2D fingerprints). Numerous types are implemented in popular chemistry software toolkits, such as CDK [194, 223], OpenBabel [152] and RDKit[2], or specialized libraries [88]. There are fingerprints which consist of a fixed (sometimes hand-curated) set of structure properties. This includes PubChem CACTVS [212], MACCS and Klekota-Roth [106] fingerprints. In contrast, *combinatorial fingerprints* define how to systematically generate substructure properties. The generated substructures differ based on the underlying set of molecules. The number of potential substructures of a fingerprint can be infinite. Representative combinatorial fingerprints are path and shortest path fingerprints, which enumerate all possible (shortest) paths between pairs of atoms in a molecule [88, 219]. Circular fingerprints such as the extended connectivity fingerprints (ECFP) [173] define and enumerate substructures based on atoms and their proximity. A parameter specifies the “radius”. Small substructures can be generated by considering each atom and its connected neighbors. A larger radius includes also the neighbors’ neighbors. This is an iterative approach and each iteration considers a further range of neighbors. See Fig. 2.2 for an illustration of a fingerprint. For a comparison of different fingerprint types see Bender *et al.* [14], Duan *et al.* [58], Willett and Winterman [222] and O’Boyle and Sayle [151].

In the means of making virtual screening more efficient, “folded” fingerprints where introduced [80]. Here, the set of all molecular properties are hashed onto a fixed-size bit vector with usually 1024 or 2048 bits. As a consequence, unrelated substructures are encoded by the same position in the fingerprint vector. This may be appropriate if the workflow performance is not overly afflicted by false positive hits. But one should refrain from using these fingerprints to train machine learning models for QSAR. Folding introduces “collisions”: different molecular properties are assigned to the same fingerprint position. An “active” position in one fingerprint vector might correspond to a completely different property than in another fingerprint vector. Because of this, fingerprint positions become harder to interpret; predictive models may perform worse.

2.3 Mass Spectrometry

Mass spectrometry (MS) is one predominant technique to measure small molecules. It can measure many different molecules at the same time which enables high-throughput analysis. Furthermore, it is high-sensitive — orders of magnitude more sensitive than nuclear magnetic resonance (NMR).

In order to measure a molecule it needs to be ionized first. Mass spectrometers measure the mass-to-charge ratio (m/z) of ionized molecules. Ions can be multiple charged. A single charged ion with a weight of 100 Da has the same m/z as a double charged ion with a weight of 200 Da. Throughout this thesis we will assume that ions are single charged. This is usually true for most small molecules. Hence, the m/z can be interpreted as the mass of a molecule. Multiple charged molecules can be easily recognized by their isotope pattern (Section 4.2.1) and removed from the data. Depending on the context, we will refer to the measured ions as molecule, ion or compound. Mass spectrometry does not measure single molecules, but a signal produced by many ions of the same molecular entity.

The output of mass spectrometry is a mass spectrum: a two dimensional plot reporting ion signal intensity as a function of m/z. The signal of a molecule is called peak. Important

---

[2] RDKit: Open-source cheminformatics; `http://www.rdkit.org`

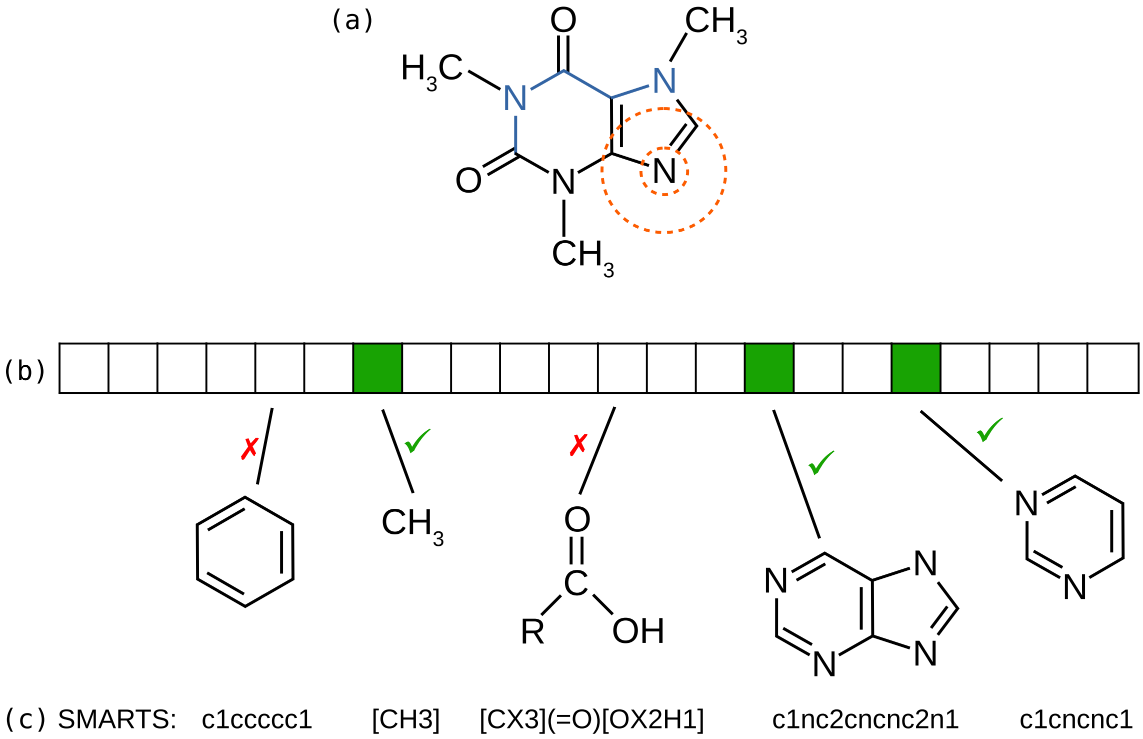


Figure 2.2: Molecular structure of caffeine and fingerprint representations. (a) Molecular structure with highlighted examples of combinatorial fingerprints. (b) Illustration of a fixed-size fingerprint with a set of substructures that are contained / not contained in caffeine. The pyrimidine ring property (rightmost illustrated structure) is a substructure of the heterocyclic purine to its left; every structure containing purine must also contain a pyrimidine ring. The SMARTS patterns in (c) can be used to match the substructures. Examples of a path (blue) and the proximity of an atom (orange circles) are indicated in (a) as representatives of all possible substructures of path fingerprints and ECFPs, respectively. Combinatorial fingerprints can be used to generate fixed-size fingerprints given a training dataset of molecules.

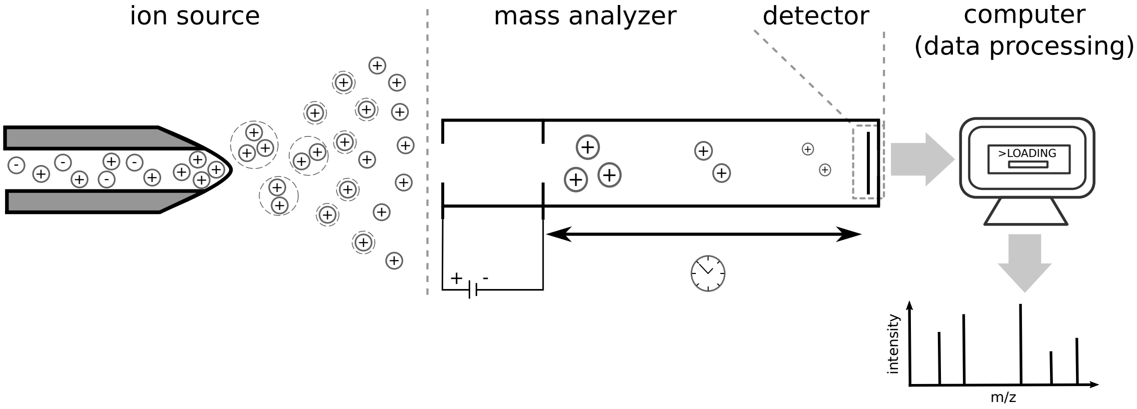


Figure 2.3: Schematic illustration of a mass spectrometer with electrospray ionization (ESI) as ion source and a time-of-flight (TOF) mass analyzer. The mass spectrometer consists of three basic parts. From left to right: the ion source produces charged molecules. These are separated in the field-free drift region of the TOF analyzer and measured by the detector. The signal is converted into a mass spectrum. Further data processing steps may follow (Section 4.1).

instrumental parameters are mass accuracy and resolution. The mass accuracy indicates how precisely the mass of a molecule can be measured: it is the ratio of m/z measurement error to true m/z and is specified in *parts per million* (ppm). High-accuracy instruments have mass errors below 20 ppm. The resolution specifies the ability to distinguish two peaks of highly similar m/z. One measure to specify resolution is the m/z divided by the closest distance $\Delta m/z$ of two peaks of equal height that are still clearly distinguishable [138]. High resolution is important for high accuracy since it limits errors resulting from overlapping signal peaks. Signal intensity can be interpreted as a function of molecule abundance. But the relation is very complex and depends on the ionization efficiency of the molecule and also ion suppression between different molecules. Thus, it is non-trivial to infer abundance from intensity.

Functional principles Conceptionally, a mass spectrometer consists of three components (Fig. 2.3): The *ion source* which ionizes the molecule, the *mass analyzer* which separates the molecules according to their m/z and the *mass detector* which detects the ions and records signal intensity. Different types of ion sources, mass analyzers and detectors exist.

Ion sources can be categorized into *soft* and *hard ionization* sources. The first technique only ionizes the molecule, the second additionally fragments it. A popular hard ionization technique is *electron ionization* (EI) [229]. Here, a beam of energetic electrons interacts with the molecule. The fragmentation is highly-reproducible and well understood [69]. However, it is often difficult to determine the m/z of the unfragmented molecular ion. EI is frequently used in combination with a separation technique called gas chromatography (Section 2.5). A common soft ionization technique is *electrospray ionization* (ESI). Here, a liquid solvent containing the analyte is pressed through a tiny capillary held at a high electric potential. The molecules form small, charged droplets. The solvent evaporates, the droplets become smaller and the charged analyte molecules repel each other to eventually break up the droplets.

A mass analyzer separates the ionized molecules according to their m/z. The mass analyzer greatly determines the quality of the data, that is, the mass accuracy, sensitivity and resolution. In a time-of-flight (TOF) analyzer all ions are accelerated through the same electric potential by an electric field. Molecules with the same charge receive the same kinetic energy. Molecules with higher mass need more energy to accelerate, thus velocities depend on the ions' m/z. After acceleration ions travel trough a field-free drift region. Travel time is proportional to the square root of m/z which separates ions of different m/z before they reach the detector. Time-of-flight instruments are comparatively simple in design and achieve high acquisition rates [71]. They can have high mass accuracy and resolution.

Other instruments, such as Fourier-transform ion cyclotron resonance (FTICR) MS instruments or Orbitrap, follow a quite different design principle. Here, mass analyzer and detector are not separated. Ions are trapped by an electric or magnetic field. In FTICR-MS a *Penning trap* traps the ions and an oscillating electric field excites them to a larger cyclotron radius. The ions' m/z is inversely proportional to the cyclotron frequency. The Orbitrap traps ions cycling around an electrode. Both instruments rely on the Fourier Transform to translate the ions' frequencies into m/z values. FTICR-MS instruments have very high mass accuracy and resolution; they can reach below 1 ppm mass error [24]. Orbitraps also have high resolution and mass accuracy of around 2 to 5 ppm [24]. Both instruments measure the ions in an non-destructive manner.

Ionization mode and adducts Many mass spectrometry instruments can be operated in either *positive ionization mode* to produce positively charged ion species or in *negative ionization mode* to produce negatively charged ion species. ESI commonly produces *protonated* ions ($[M + H]^+$) in positive ionization mode — that is, a proton is added to the molecule. In negative ionization mode it frequently produces *deprotonated* ions ($[M - H]^-$) — that is, the molecule loses a single proton [75]. Here, "M" represents the neutral molecule. Apart from that, different ionic species might attach to a molecule to form an *adduct ion*. Common adduct ions in positive ionization mode are ammonium adduct ions ($[M + NH_4]^+$), sodium adduct ions ($[M + Na]^+$), and potassium adduct ions ($[M + K]^+$) [110]. Common adduct ions for negative ionization mode are chlorine adduct ions ($[M + Cl]^-$) [233]. The ratio of different adduct ions strongly depends on the used ionization technique and analytical setup. For example, a high salt concentration in the sample may favor $[M + Na]^+$ and $[M + Cl]^-$ ions. In the following we will call adduct ions "adducts" for short.

2.4 Tandem Mass Spectrometry

In *tandem mass spectrometry* (MS/MS) two mass analyzers are coupled together with a fragmentation cell in between. The first mass analyzer selects ions in a certain m/z range. Subsequently, these ions are fragmented. And finally, the fragments are recorded by the second analyzer. The resulting tandem mass spectrum (MS/MS spectrum) provides additional information about the measured molecule. This aids the discrimination of isobaric compounds (molecules with identical nominal mass) and structural isomers. While isobaric compounds might also be distinguishable by the first level of mass spectrometry (MS1), structural isomers have the same molecular formula and need MS/MS for discrimination. Stereoisomers are usually indistinguishable by mass spectrometry.

Collision induced dissociation (CID) is a common fragmentation technique coupled with soft ionization. Here, the ions pass through a collision cell containing a collision gas. On their trajectory the ions collide with the gas which converts some kinetic energy into internal energy. This triggers a chemical reaction and the ions get fragmented into two (or more) pieces and the charge is passed on to one of the pieces. The ion that is being fragmented is called the *precursor ion*; the fragmented part that carries a charge is called product ion or *fragment*; and the neutral part is called *loss*. Only the ionized fragments can be measured by the subsequent mass analyzer; this produces the MS/MS spectrum. Since not only a single ion but many identical ions of the same molecular entity are fragmented at once, many fragments are recorded. Consequently, the same substructure can be both, fragment and loss. The velocity at which ions pass through the collision cell can be regulated; this changes the transferred *collision energy*. Higher collision energies result in smaller fragments, whereas lower collision energies may leave ions unfragmented. To obtain a higher number of different fragments, multiple MS/MS scans may be measured at different collision energies and combined.

2.5 Chromatography

Biological samples are usually too complex to be measured directly by mass spectrometry. Highly abundant ions would interfere too much with the detection of less abundant ions. Furthermore, tandem mass spectrometry would not be able to select single ion species but mixtures for MS/MS. This makes the interpretation of MS/MS spectra challenging. Thus, mass spectrometry is often coupled to a prior separation step. *Chromatography* is a technique for the separation of a mixture of different molecules based on their chemical or physical properties. The chromatography system consist of a *stationary phase* and a *mobile phase*. The mobile phase carries analyte molecules through a column containing the stationary phase. Molecules bind to the stationary phase with different affinities. Thus, different molecules pass through the column at different speeds and separate. Chromatography can be coupled with ion-mobility spectrometry to enhance separation [132].

The most prominent chromatography techniques coupled with mass spectrometry are *gas chromatography* (GC) and *liquid chromatography* (LC). *Gas chromatography-mass spectrometry* (GC-MS) has been around for many decades and often uses EI as ion source [68]. Molecules have to be — or made be — volatile. Due to the high temperatures necessary for GC, this is not applicable for larger molecules, such as peptides, which would denature. *Liquid chromatography-mass spectrometry* (LC-MS) based metabolomics often uses ESI as ion source. LC can be performed with a diverse set of molecules, including many secondary metabolites [67]. Even with LC separation tens to hundred of different molecules can elute at the same time [158].

In untargeted mass spectrometry experiments, first an MS1 scan is performed to detect the *m/z* values of all currently eluting molecules. Now, one or more high-intensity peaks are individually selected and MS/MS scans are performed. This is done in an alternating fashion to cover the whole LC-MS/MS run with MS1 and MS/MS spectra. Still, many compounds might not be covered by MS/MS. The time at which a molecule elutes is called the *retention time*. This can be used as orthogonal information, in addition to a molecule's *m/z* and MS/MS spectrum, to assist identification. Whereas GC retention times can be standardized to a system-independent number called the Kovats retention index, LC systems produce retention times with much greater variation [227]. The LC retention time of a molecule differs between setups and can even change from run to run on the very same instrument. Nevertheless, retention times can be utilized as orthogonal information to confirm the identity of a compound: A known reference compound can be spiked into the sample to check if it elutes at the same retention time as the unknown compound of interest.

3 Statistics and Graph Theory

In this chapter, we provide graph-theoretical notations and definitions. Then, we give a short introduction to Bayesian statistics and Bayesian inference. We refer readers who are less familiar with graph theory to one of the many graph theory textbooks such as Diestel [54]. For a more comprehensive view on Bayesian statistics in general see Koch [107] and Liu [120] for Monte Carlo methods in particular.

3.1 Graph Theory

Graphs are a mathematical concept to model pairwise relations between objects. Here, we introduce the basic terminology from graph theory.

A *graph* $G = (V, E)$ consists of a set of *nodes* V and a set of *edges* E. Edges depict relations between nodes; each edge connects two nodes. Graphs can be either directed or undirected. In undirected graphs, edges are unordered pairs of nodes $E \subseteq \binom{V}{2}$, where $\binom{V}{2}$ is the set of all two-element subsets of V. Directed edges are two-element tuples (u, v) of nodes $u, v \in V$. Thus, for directed graphs the edge set is $E \subseteq V \times V$. Two nodes $u, v \in V$ are *adjacent* to each other if they are connected by an edge. A *path* is a sequence of edges such that consecutive edges share exactly one node and each edge is used at most once; in directed graphs, the ending node of one edge must be the starting node of the next edge in the sequence. A path with the same node as start and endpoint is called *cycle*. A directed graph that does not contain any cycle is called *directed acyclic graph* (DAG). For a directed edge $(u, v) \in E$ in a DAG we call u the *parent* and v the *child*. A *tree* $T = (V, E)$ is an undirected graph in which any two nodes are connected by exactly one path. A *rooted tree* has a designated vertex r called root. All directed edges point away from this root. A graph $G' = (V', E')$ is a *subgraph* of G if and only if $V' \subseteq V$ and $E' \subseteq E$. A *clique* in an undirected graph G is a subset of nodes $S \subseteq V$, such that every two nodes in S are connected by an edge in G.

Graphs can be colored and labeled to categorize nodes or edges, or to assign them a certain feature. In fragmentation graphs and fragmentation trees, which are introduced in Section 4.2.2, the nodes and edges are labeled with molecular formulas in order to explain fragments and losses in a spectrum. Colors are usually introduced to assign nodes to a common origin. Different nodes with the same color can correspond to different hypotheses assigned to the same origin. A graph is *colorful* if every node has an unique color. This allows to formulate optimization problems which decide between different hypotheses by finding an optimal colorful subgraph. For example, a peak in a mass spectrum could be explained by multiple molecular formula candidates. Here, the candidates are nodes and the peak is represented by a color. When selecting a colorful subgraph, only one valid candidate may be retained.

Nodes and edges can be weighted to associate some measure of cost or importance. Edge-weighted graphs have a *scoring function* $w_E : E \to \mathbb{R}$. Analogously, vertex-weighted graphs have a scoring function $w_V : V \to \mathbb{R}$.

3.2 Bayesian Statistics

Bayesian statistics is a fundamental theory in the field of statistics in which probability quantifies the belief in a hypothesis or the certainty of an event. Probabilities do not have to be based solely on observations. Instead, Bayesian statistics allows to incorporate prior information and express this in form of probabilities. This prior information can be derived from many different sources. It may be expert knowledge: here, the probabilities express a measure of plausibility that a hypothesis is indeed true. It may also be a summary of prior experimental results, or a hypothesized generic distribution over events in the sample space.

Suppose we want to estimate the distribution of body heights of people in, say, Germany. The distribution of body heights of the general world population offers a good estimate of body heights in Germany. This prior knowledge tells us that adult humans rarely have heights below 1 or above 2.5 meters. If we do not have any data specifically for Germany, then this is a good estimate. We could now conduct an experiment and measure people from Germany. If we select two random German people and measure their heights, this will give a bad approximation of German body height distribution. Hence, the world-wide population is still the better estimate. However, if we measure 100,000 German people, we may end up with an even better estimate. Now, at what number do we trust our sample statistic more than the prior information? After 10 samples? Or after 100 samples? With Bayesian statistics we do not have to decide for an arbitrary threshold. Rather, it enables us to combine the prior knowledge with the data, thus making best use of all available information.

Let us define this more formally: Given are the data $\mathcal{D}$ (our samples), and we want to estimate the unobserved model parameters $\mathcal{M}$. Then, the *prior probability* is a probability distribution $\mathbb{P}(\mathcal{M})$ over all possible model parameters. The *likelihood* $\mathbb{P}(\mathcal{D}|\mathcal{M})$ is the conditional probability that our sample data was produced under the given parameter set. This enables us to calculate a *posterior probability* $\mathbb{P}(\mathcal{M}|\mathcal{D})$ using Bayes' theorem:

$$\mathbb{P}(\mathcal{M}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{M})\mathbb{P}(\mathcal{M})}{\mathbb{P}(\mathcal{D})}. \tag{3.1}$$

Now, back to our example. Assume that body heights are normally distributed and we observed body heights of German people $D \in \mathcal{D}$. The unobserved parameters of our model are the mean μ and standard deviation σ of normal distribution $\mathcal{N}(\mu, \sigma^2)$. Given a sample statistic of the world population, we can estimate prior probabilities of these parameters — this is $\mathbb{P}(\mathcal{M})$. Furthermore, we can calculate how likely specific parameters would generate the data D — this is $\mathbb{P}(\mathcal{D}|\mathcal{M})$. When we combine all this information, we can estimate the model parameters. Finding the most probable model parameters based on $\mathbb{P}(\mathcal{M}|\mathcal{D})$ corresponds to a maximum a posteriori (MAP) estimate. If we do not have any prior knowledge of the model parameters $\mathcal{M}$, we usually assume a flat prior — all parameter realizations are equally probable. This results in a maximum likelihood (ML) estimate: the model which most likely generates the data is the most likely model. To calculate $\mathbb{P}(\mathcal{D})$, we would need to marginalize over all possible parameter realizations of $\mathcal{M}$. However, this is not necessary in order to find the most probable parameters based on $\mathbb{P}(\mathcal{M}|\mathcal{D})$. The denominator in equation (3.1) is the same for all model parameters. Since $\mathbb{P}(\mathcal{M}|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\mathcal{M})\mathbb{P}(\mathcal{M})$ we can ignore $\mathbb{P}(\mathcal{D})$ and only maximize $\mathbb{P}(\mathcal{D}|\mathcal{M})\mathbb{P}(\mathcal{M})$.

3.3 Bayesian Inference

Bayesian inference is a method to infer the probability of a hypothesis using the Bayes' theorem. This can be as simple as weighing the hypothesis "the die is fair" against "the die is biased". But it can also be used to assign posterior probabilities to all possible parameter realizations of a highly-dimensional model. ML or MAP estimation essentially become optimization problems that predict a single point in parameter space. However, this is not appropriate for all applications. Especially, if the posterior probability is not a regular unimodal distribution with most probability being "concentrated" in one high peak. Here, Bayesian inference can help and estimate the whole posterior probability distribution. For example, assume a game of dice is played, but the casino frequently switches between fair and biased dice. Thus, both models, "dice are fair" and "dice are biased", are true and coexist with different probabilities. Another example is about risk assessment: based on the weather forecast, we predict if the wind will become a hurricane in the coming days. The MAP estimation might predict wind speeds of exactly 39 km/h. Using Bayesian inference we might come to the conclusion that there is still a 20 % chance that wind speeds will exceed 120 km/h. Both statements are correct. However, Bayesian inference provides the more relevant information in this example. This illustrates the inherent difference between these methods. There also exists a range of approaches in between single-point estimates and the calculation of the whole posterior probability distribution which are not covered in this thesis.

The posterior probability distribution can often not be solved analytically. Monte Carlo methods can be used to generate a representative sample of the underlying distribution. From this, the distribution can be approximated and statistics can be derived. Monte Carlo sampling was initially used to solve physics-related problems [130]. Nowadays many applications for statistical inference are based on the concept.

What does "sampling from a probability distribution" imply formally? Conducting a random experiment results in a single outcome from a set of possible outcomes. *Random variables* are functions that map outcomes to some measurable space. In this way they group possible outcomes to *events* and give them some meaning. A probability distribution assigns probabilities to the possible events of a random variable. Take for example a six-sided die. Rolling the die will result in one out of six possible outcomes of $1, 2, 3, 4, 5$ and 6. A random variable X may divide outcomes into the two events x_0 = "the number of pips is even" and x_1 = "the number of pips is odd". This can be defined by the mapping $\{2, 4, 6\} \rightarrow x_0$ and $\{1, 3, 5\} \rightarrow x_1$. In the discrete case the probability distribution can be described by a *probability mass function* p: $p_X(x_0) = \mathbb{P}(X = x_0) = 1/2$ and $p_X(x_1) = \mathbb{P}(X = x_1) = 1/2$. We may refer to p_X as p for short if it is obvious which random variable is considered. Continuous random variables that map outcomes to an uncountable infinite number of events can be described by *probability density functions* instead. Sampling a random variable X yields a realization x, that is, an event of the random variable chosen according to $\mathbb{P}(X)$.

Generating independent samples from a target probability distribution $p(x)$ in such a way to closely resemble the underlying distribution is often infeasible. This is especially the case for high-dimensional parameter spaces, where most parameter realizations have very low — close to zero — probability and do not considerably contribute to the probability distribution. *Importance sampling* was suggested to sample from a different distribution which focuses on the regions of "importance", where the target distribution is meaningful non-zero [127]. Alternatively, *Markov chain Monte Carlo* (MCMC) can be used to generate a sequence of random samples. All MCMC methods are governed by the *Markov property* which states that the probability of each event only depends on the state of its preceding event; all previous states are irrelevant. For a sequence of random variables $X^{(1)}, X^{(2)}, \ldots, X^{(t+1)}$ and corresponding realizations $x^{(1)}, x^{(2)}, \ldots, x^{(t+1)}$ this can be formalized as:

$$\begin{aligned}&\mathbb{P}(X^{(t+1)} = x^{(t+1)} | X^{(t)} = x^{(t)}, X^{(t-1)} = x^{(t-1)}, \ldots, X^{(1)} = x^{(1)})\\ =&\mathbb{P}(X^{(t+1)} = x^{(t+1)} | X^{(t)} = x^{(t)}).\end{aligned}$$

A stochastic process that satisfies this property is called *Markov chain* (or Markov process in the continuous case). It is uniquely described by its *transition function* $T(x, x')$ which defines the probability of transitioning from one state x to any other state x'. A distribution $p(x)$ is said to be *invariant* with respect to the Markov chain if the transition function leaves $p(x)$ unchanged:

$$p(x') = \sum_{x} T(x, x') p(x). \tag{3.2}$$

Two properties are commonly desired for a given Markov chain:

1. irreducibility: A chain is irreducible if it is possible to get from any state to any other state (not necessarily in one step).
2. aperiodicity: A state x is k-periodic if every sequence of states which starts with x and ends in x has a number of steps which is a multiple of k. If $k = 1$, then the state x is aperiodic, otherwise it is periodic. A Markov chain is aperiodic if every state is aperiodic.

If a Markov chain is irreducible and aperiodic, then the chain will become *stationary* at a unique invariant distribution. Various algorithms exist for constructing Markov chains, most notably the Metropolis-Hasting algorithm [82]. MCMC methods are frequently applied to bioinformatics problems [92]. Popular applications include sequence-based approaches such as finding DNA binding motifs [116] and phylogenetic inference [230].

3.4 Gibbs Sampling

Gibbs sampling is a suitable MCMC method if the conditional distributions of the variables are known and easy to sample from. Gibbs sampling, in its basic version, is a special case of the Metropolis-Hastings algorithm [82, 131]. Metropolis-Hastings in general defines transition probabilities based on a proposal function $g(x'|x)$ and an acceptance probability $A(x, x')$. The proposal function generates a new state based on the current one. Then, this new state is accepted proportional to the acceptance probability; if it is rejected, the state remains unchanged. Recall, that in order to generate samples from the target distribution $p(x)$, the distribution needs to be invariant to the transition function $T(x, x')$. The proposal function does not ensure appropriate transition probabilities. Rather, the acceptance probabilities are chosen, such that the combination of proposal and acceptance probabilities results in the desired transition probability. Thus, the actual transition function is $T(x, x') = g(x'|x)A(x, x')$. However, poorly chosen proposal functions may lead to low acceptance rates and thereby to an inefficient sampling procedure. Gibbs sampling uses the conditional probabilities of the target distribution which directly results in the desired transition probabilities. Hence, it does not require an acceptance-rejection step. It can produce samples with smaller variance and enable efficient sampling.

The algorithm can be described as follows: Suppose, we want to generate samples from the n-dimensional random variable $X = (X_1, \dots, X_n)$. In an initialization step, a random realization is chosen for the first sample $x^{(1)}$. Then, new samples are generated based on their preceding sample. To generate $x^{(i+1)}$, we sample each component $x_j^{(i)}$ of this vector conditioned on all previous components. The iteration step is described by the following: Let $x^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$ be the sample for iteration i. Then $x^{(i+1)}$ is generated by updating each $x_j^{(i+1)}$ according to

$$p(x_j^{(i+1)}|x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \dots, x_n^{(i)}), \text{ for each } j = 1 \dots n.$$

The joint probability distribution $p(x)$ is invariant with respect to this updating procedure:

$$\begin{aligned} &p(x_j^{(i+1)}|x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \dots, x_n^{(i)}) \cdot p(x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \dots, x_n^{(i)}) \\ =&p(x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)}, x_j^{(i+1)}, x_{j+1}^{(i)}, \dots, x_n^{(i)}). \end{aligned}$$

Hence, once the stationary distribution is reached, Gibbs sampling generates samples from the target distribution.

We give a simple example of a Gibbs sampler in Fig. 3.1. Here, a Markov chain for two correlated variables is constructed. Directly after initialization, the conditional probabilities are “flat”. After some steps, the Markov chain converges to the area of higher probabilities. This illustrates two important properties of the Markov chain: Firstly, the samples from the beginning do not follow the stationary probability distribution and have to be omitted for estimation. Secondly, as a consequence of the sampling process, samples are autocorrelated: successive samples are highly correlated.

3.5 Bayesian Networks

A Bayesian network is a type of a probabilistic graphical model and is described by a directed acyclic graph $G = (V, E)$ (Section 3.1) whose nodes represent random variables $X_v, v \in V$. Edges represent conditional dependencies. The graph structure specifies a factorized representation of the joint probability distribution and enables an efficient way to compute probabilities and perform inferences on the random variables.

Bayesian networks satisfy the *local Markov property*: each node in the graph is independent of its non-descendants given its parents. In general, the joint probability can be expressed by conditional probabilities using the *chain rule*:

$$\mathbb{P}(X_1, X_2, \dots, X_n) = \mathbb{P}(X_1|X_2, \dots, X_n) \cdot \mathbb{P}(X_2, X_3, \dots, X_n) = \prod_{k=1}^{n} \mathbb{P}(X_k| \bigcap_{j=k+1}^{n} X_j) \quad (3.3)$$

Because of the local Markov property, equation (3.3) can be transformed into:

$$\mathbb{P}(X_1, X_2, \dots, X_n) = \prod_{v \in V} \mathbb{P}(X_v| \bigcap_{j \in parent(v)} X_j),$$

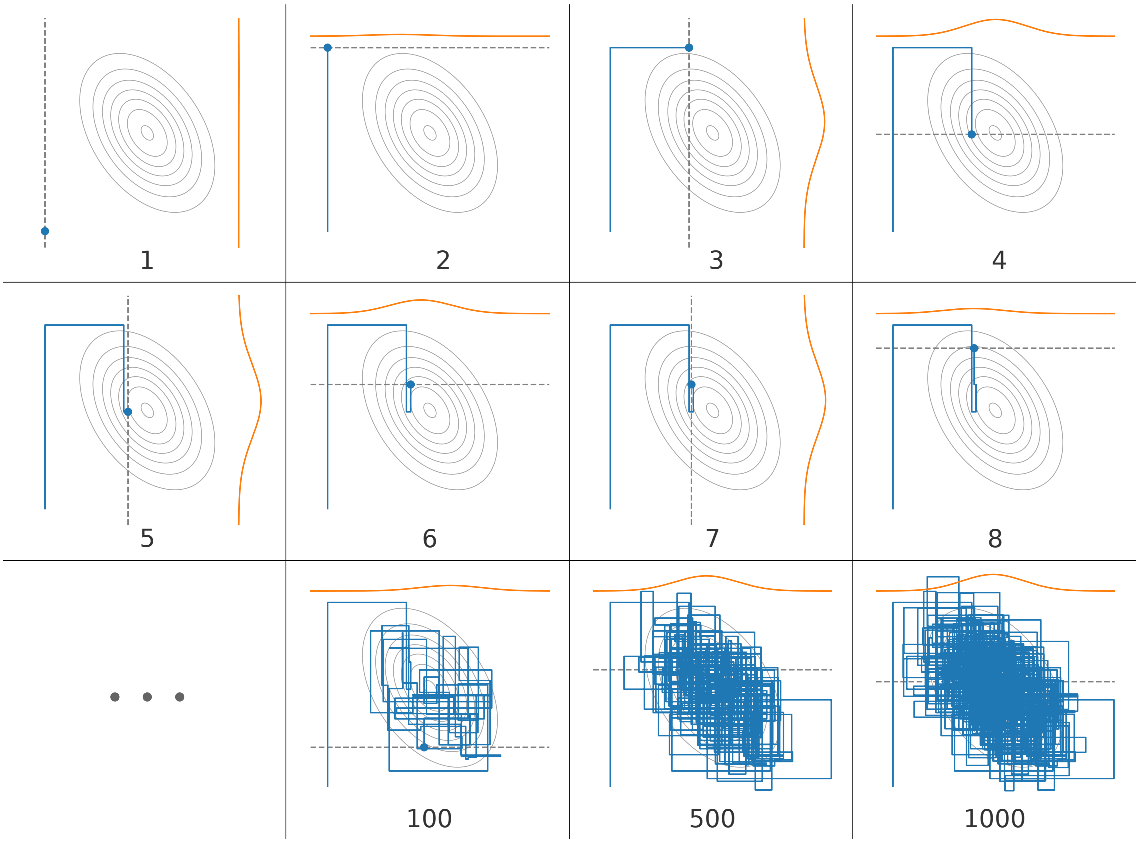
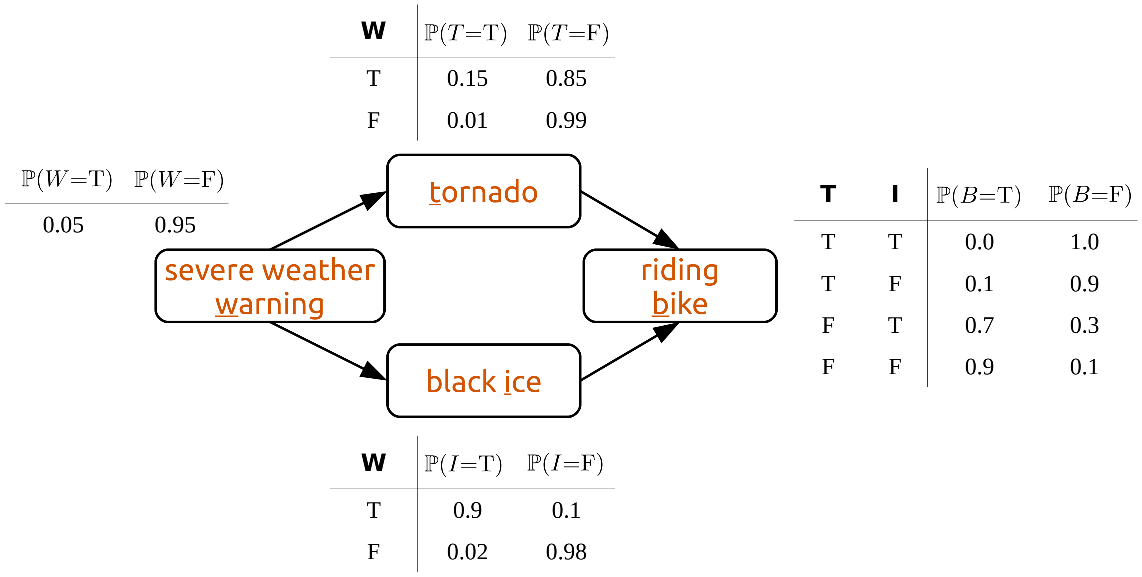


Figure 3.1: Gibbs sampling with two random variables of a multivariate normal distribution. Displayed are multiple steps of a single Markov chain. The numbers indicate the current iteration step. The ellipses illustrate the target distribution which we intend to estimate; this unimodal probability distribution has its maximum at the central ellipse. The blue curve represents the samples. The current sample is indicated with a blue dot. Gibbs sampling begins by choosing a random realization as start. Then, the value of one variable is fixed and the other is sampled based on the conditional distribution. We indicate the fixed parameter value by the gray dashed line. The conditional distribution, which is used to sample from, is either displayed on top or at the right, depending on which variable is currently sampled.

where $parent(v)$ denotes the set of all parent nodes of node v. This allows for an efficient factorization of the joint probability because the number of parents per node is usually low. In this way, Bayesian networks offer a good compromise between assuming full dependence and assuming full independence of variables. Besides for the efficient computation of the joint probability, Bayesian networks can be used to infer unobserved variables given a set of observed variables. Exact inference is NP-hard [44] and heuristics or approximations may be applied [79]. Bayesian networks are used in many different domains including risk assessment [231], managing and planning [139], community modeling [25] and disease diagnosis [35, 159]; for more details, see Chen and Pollino [42].



$\mathbb{P}(W{=}\mathrm{T})$	$\mathbb{P}(W{=}\mathrm{F})$
0.05	0.95

W	$\mathbb{P}(T{=}\mathrm{T})$	$\mathbb{P}(T{=}\mathrm{F})$
T	0.15	0.85
F	0.01	0.99

W	$\mathbb{P}(I{=}\mathrm{T})$	$\mathbb{P}(I{=}\mathrm{F})$
T	0.9	0.1
F	0.02	0.98

T	**I**	$\mathbb{P}(B{=}\mathrm{T})$	$\mathbb{P}(B{=}\mathrm{F})$
T	T	0.0	1.0
T	F	0.1	0.9
F	T	0.7	0.3
F	F	0.9	0.1

Figure 3.2: Illustration of a Bayesian network with four random variables. The nodes represent the variables and the edges conditional dependencies. Probabilities are expressed as contingency tables. Variables are abbreviated with the underlined character. Variable realizations are either true (T) or false (F). Given the probability for a “severe weather warning” and conditional probabilities for the remaining variables, joint probabilities can be estimated.

A simple example of a Bayesian network is illustrated in Fig. 3.2. The network enables easy computation of the joint probability. For example, the probability, that there is a weather warning for black ice and I am still riding the bike is

$$\begin{aligned}
&\mathbb{P}(W=\mathrm{T}, I=\mathrm{T}, T=\mathrm{F}, B=\mathrm{T})\\
=&\mathbb{P}(W=\mathrm{T})\cdot\mathbb{P}(I=\mathrm{T}|W=\mathrm{T})\cdot\mathbb{P}(T=\mathrm{F}|W=\mathrm{T})\cdot\mathbb{P}(B=\mathrm{T}|I=\mathrm{T},T=\mathrm{F})\\
=&0.05\cdot 0.9\cdot 0.85\cdot 0.7\approx 2.68\%.
\end{aligned}$$

Furthermore, unobserved variables can be inferred from a set of observed variables. Given that there is black ice, but no tornado, the probability of riding the bike $\mathbb{P}(B=\mathrm{T}|I=\mathrm{T},T=\mathrm{F})$ is always 0.7, independent of the variable *severe weather warning*.

4 Computational Mass Spectrometry

In this chapter, we give an overview of computational methods for small molecule mass spectrometry. We focus on high mass accuracy, high-resolution LC-MS/MS data in particular. We do not cover methods related to the chromatography part of the analysis, such as retention time prediction [227]. Elucidation of stereochemistry is usually not possible solely based on mass spectrometry [51]. This rather requires orthogonal information, such as NMR measurements. Thus, structure elucidation and identification refers to the determination of the constitution of a molecule: that is, the identity and connectivity of the atoms including bond multiplicities, but no spatial (stereochemistry) information.

Our novel methods that we present in Chapter 5 and 6 build upon SIRIUS [63] and CSI:FingerID [61]. SIRIUS combines isotope pattern matching (Section 4.2.1) and MS/MS spectra analysis using fragmentation trees (Section 4.2.2). CSI:FingerID searches MS/MS spectra in a structure database (Section 4.4.1).

In general, we give a short overview of the data processing, and describe computational methods for *de novo* molecular formula annotation, spectral library search and structure database search. For a recent list of software tools and resources for metabolomics see [155]. Interested readers can obtain a more detailed overview of computational methods and databases for structure elucidation from relevant reviews [16, 18, 93].

4.1 Data Processing

Before performing qualitative or quantitative mass spectrometry analyses, such as compound identification, the data needs to be preprocessed first. The raw data consists of spectra in so called *profile mode*: here, a molecule's signal peak does not correspond to a single *m/z* value but a distribution due to measurement inaccuracies. Peak picking detects the peaks and transforms them into single *m/z* values. This step is also called *centroiding*, although not necessarily the centroid of a peak is reported but a different value such as the maximum.

Subsequent processing steps of LC-MS data may include noise removal [180, 224], baseline correction [10, 210], feature detection and feature grouping [50, 101, 112, 137, 197], and retention time alignment [113, 114, 119, 232]. As a result of the processing, all isotope peaks and adduct peaks (often called features) that correspond to the same compound are grouped; and the same compound is also detected and matched over multiple LC-MS runs. This is necessary e.g. for quantitative analysis. In addition to the software of mass spectrometry vendors, there are multiple free, open-source software packages for these tasks, including OpenMS [175], MZmine 2 [161] and XCMS [15, 199].

Methods for compound identification, which are presented in the following, usually expect centroided spectra. They do not necessarily need further processing, although it might improve performance. Generally, these methods do not directly make use of the established adduct groupings. Thus, the MS/MS spectrum of, for example, the protonated ion and the sodium adduct ion of a compound are analyzed individually. However, confident adduct assignment can still be beneficial. Furthermore, different isotope patterns or MS/MS spectra of the same adduct of a compound which were measured in different runs, can be merged to improve spectrum quality.

4.2 De Novo Molecular Formula Annotation

Annotating the molecular formula of a compound is the first step in its structural elucidation: It often allows us to deduce important information about its likely structure; knowing it improves the subsequent search in molecular structure databases [63, 177]; and finally, it guides data interpretation based on atoms and unsaturation degree for full structure elucidation via NMR or X-ray crystallography. Annotation of molecular formulas is far from trivial, especially if executed *de novo*, that is, without the use of a database and without artificially restricting candidate elements: here, the number of molecular formula candidates grows rapidly with compound size and elements beyond carbon, hydrogen, nitrogen and oxygen [19]. Heuristic constraints for "permissible" molecular formulas will counter this growth [105], but can again prevent the annotation of true molecular formulas.

Molecular formula annotation can be performed by first selecting possible molecular formula candidates and subsequently scoring these candidates. Many methods select molecular formula candidates from a database, but efficient algorithms exist to enumerate all possible molecular formulas, thus enabling *de novo* annotation. Given the compound's mass (accounting for a certain mass error) and an alphabet of considered elements, all molecular formulas can be enumerated via dynamic programming [22, 60]. Scoring (or filtering) of the candidates can be based on the isotope pattern (MS1 only) [4, 21, 23, 105, 122, 153, 204], or combined with analyzing the MS/MS spectrum [19, 20, 63, 129, 162, 196].

Instead of considering compounds individually, a complete LC-MS run can be annotated at once, utilizing co-occurrence of molecular formulas differing by a predefined set of biotransformations [47, 50, 52, 174]. These approaches select molecular formula candidates from a structure database; thus, implicitly, they try to identify molecular structures (or their isomers) from a restricted structure database, and cannot annotate novel molecular formulas.

4.2.1 Isotope Pattern Analysis

Chemical elements can have multiple isotopes (Section 2.1). The most common elements found in biomolecules — carbon, hydrogen, nitrogen and oxygen — each have multiple stable isotopes. Thus, one compound produces multiple peaks in a spectrum based on its different isotopologues. Isotopologues with identical nominal mass are usually measured as a single peak. Because of the mass defect, the mass difference of the lightest isotope to the other isotopes is different for each element. These mass differences, combined with the different natural abundances of isotopes, are the reason why we can distinguish elemental compositions based on mass spectrometry. However, mass spectrometry cannot resolve all different isotopologues but measures a superposition of these. The monoisotopic peak represents the isotopologue with lowest mass. The M+1 peak represents all isotopologues that include exactly one isotope that has an additional neutron; the M+2 peak represents all molecules that include either one isotope that has two additional neutrons or two isotopes that have one additional neutron; and so on. For single charged molecules, the isotope peaks differ by approximately $m/z = 1$. Some high-resolving mass spectrometer can resolve isotopologues to some extend which (partially) reveals the isotopic fine structure [122]. Isotope patterns can be simulated via polynomial expansion [33, 204] and Fourier transform [170, 204]. Merging isotopologues with similar mass enables efficient isotope pattern simulation [23, 111]. Molecular formula candidates are scored by simulating an isotope pattern for each candidate and comparing it against the measured isotope pattern [23, 63, 122]. SIRIUS 4 considers both, relative and absolute, mass and intensity errors and models these as normally distributed random variables. SIRIUS 4 achieves more than 60 % correct molecular formula annotations based on the isotope pattern alone. The isotope pattern also allows to detect elements [63, 133]. This primarily improves efficiency, since the set of considered elements can be limited and not all molecular formula candidates have to be scored.

Isotope patterns enable easy distinction of different charges: isotope peaks of single charged compounds differ by $m/z = 1$, double charged differ by $m/z = 1/2$. Unfortunately, for some compounds we only detect the monoisotopic peak since other isotope peaks have too low intensity.

4.2.2 Fragmentation Trees

A fragmentation tree annotates peaks in the fragmentation spectrum with molecular formulas and identifies likely losses between the fragments — similar to "fragmentation diagrams" created by experts. The calculated tree must not be understood as ground truth but can be used to derive information about the measured compound's fragmentation [165]. Fragmentation trees are also used to identify the molecular formula of an unknown compound. For every molecular formula candidate of the precursor ion, a separate fragmentation tree is computed which best explains the spectrum, as evaluated by a maximum a posteriori estimator [19]. This estimation takes into account information such as mass deviations, intensities, common losses and loss sizes. The overall best-scoring fragmentation tree corresponds to the most likely molecular formula explanation.

A simplified example of a fragmentation tree is presented in Fig. 4.1. A fragmentation tree is computed from the fragmentation spectrum given the (candidate) molecular formula of the precursor ion. Initially, a fragmentation graph is constructed in the following way: For every fragment peak, all possible molecular formula explanations are computed. These explanations must be subformulas of the precursor molecular formula — a fragment only looses, but never gains new atoms. Every such molecular formula is a node in the graph. Nodes are connected by an edge if one node is a subformula of another node — this represents a potential loss. Using combinatorial optimization, the best scoring fragmentation tree is computed which explains every peak at most once. Unexplained peaks are considered noise.

Computing the fragmentation tree from a fragmentation graph requires to solve an NP-hard [167] problem named the MAXIMUM COLORFUL SUBTREE problem [20]. In practice, many instances can be solved swiftly using Integer Linear Programming (ILP) [167, 217]. Additionally, heuristics can be applied to speed up computations and restrict solving the exact problem to the top-ranked candidates. Fragmentation trees can also be computed from MS^{n} data [178]. Fragmentation tree alignments [166] can be used to compare and cluster similar compounds. This enables the search for structurally similar compounds in spectral reference libraries if the true hit is missing. CSI:FingerID, which searches MS/MS

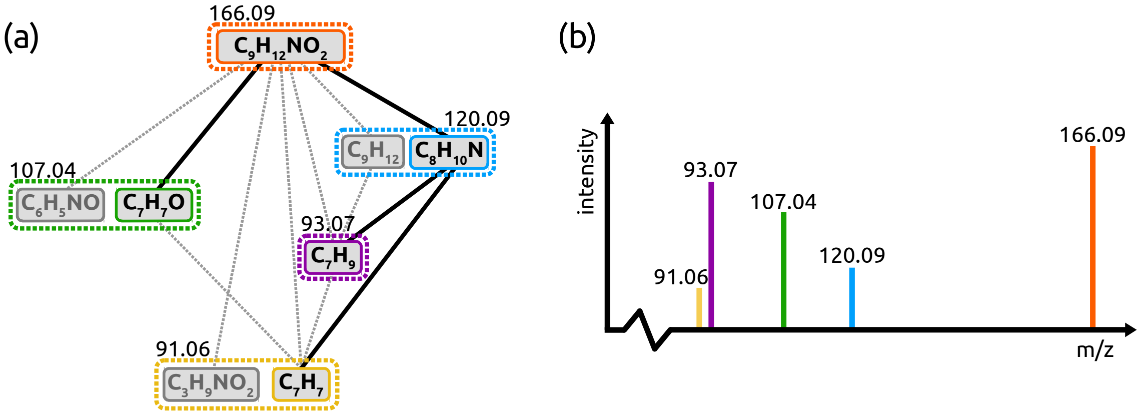


Figure 4.1: Example of a fragmentation tree computed from a fragmentation graph in (a), given the spectrum in (b). The molecular formula of the neutral precursor is assumed to be $C_9H_{12}NO_2$. Molecular formulas are computed for all fragment peaks and serve as the nodes of the graph; nodes with the same color indicate molecular formulas corresponding to the same peaks. Nodes are connected by edges if one node is a subformula of another, thereby creating the fragmentation graph. A fragmentation tree is a connected subgraph which explains each color (peak) at most once and has no cycles. The best-scoring fragmentation tree, corresponding to a maximum a posteriori estimator, is computed by combinatorial optimization. The optimal fragmentation tree is indicated by solid lines; nodes which are not used are grayed out. These computations are repeated for each molecular formula candidate explaining the precursor mass, and the best such fragmentation tree is reported.

spectra in a structure database, uses fragmentation trees to predict molecular fingerprints (Section 4.4.1).

4.3 Spectral Library Search

A well established method to identify compounds based on fragmentation spectra is to search these in a spectral reference library. EI spectra are highly reproducible [69]. Since GC-MS has been established for a long time, spectral reference libraries have been collected extensively and are comparably large [191]. Tandem mass spectra are less reproducible, in particular across different types of instruments [149, 191]. Nevertheless, reliable hits can be obtained using instruments under similar conditions [40] and if the correct hit is contained in the reference library [179]. The number of reference compounds in public spectral libraries has grown over the past years [90, 198, 209, 211, 226]. Still, only few percent of all detected compounds can be identified [48].

When searching an MS/MS spectrum in a reference library, candidate spectra are selected that match the precursor ion's mass. Spectra are compared using the *cosine similarity*. To do so, the spectra are represented as vectors; peaks are usually binned by m/z. The cosine similarity is the dot product of the normalized vectors. The similarity ranges between 0 for spectra without common peaks, to 1 for identical spectra. To reduce false positive hits, a minimum number of common peaks may be required [211]. Furthermore, the transformation of intensities was suggested to reduce the impact of high intense peaks [90, 193]. But conversely, this can increase the influence of noise peaks. Additionally, the mirrored spectrum can be considered [15, 192]: to mirror a spectrum with precursor mass M, we replace peak m/z value m by $M - m$. Averaging the score of spectrum and mirrored spectrum can reduce errors caused by systematic mass shifts. Instead of binning spectra by m/z, the mass error can also be included into the scoring, e.g. by using a probability product kernel [83]. Further similarity measures have been tested [193] but the cosine similarity performed best.

The mirrored spectrum can also be used for analogue search. Here, we do not aim for exact matches but try to find structurally similar compounds. Hence, we do not limit search to compounds with the same precursor mass but search either within a larger mass window, select masses that match some common biotransformation [118] or use the entire reference library. Alternatively, spectral alignments [211] or fragmentation tree alignments [166] can be used.

To overcome the problem of false positive hits, Scheubert *et al.* [179] developed a false discovery rate estimation for spectral library search. In a target-decoy approach [65] a decoy database is created based on fragmentation trees. The principal idea is, that the decoy database consists of artificial spectra that will never be measured but are still highly similar to the real spectra in the target reference library. All query spectra are searched in the target database to find promising hits. Additionally, the query spectra are searched in the decoy database to estimate the chance that the hits in the target database are in fact false positives.

Recently, the concept of molecular networks emerged to visualize similarities between compounds in a dataset [136, 211, 214]. Here, compounds are represented as nodes in a graph and edges depict the similarity between compounds (usually the cosine similarity between the MS/MS spectra). Hypothetical biotransformations can be derived from mass differences between compounds [203]. Furthermore, spectral library hits can be included into the network. This facilitates dereplication and allows to generate hypotheses about the structure of a compound that does not have a spectral library match, based on compounds that have been annotated with a structure. Additionally, annotations of MS2LDA [205] can be combined with such networks. MS2LDA is an unsupervised machine learning method that extracts common spectral patterns from a collection of MS/MS spectra. These spectral patterns can be manually annotated with structural properties. Thus, all compounds which contain an annotated spectral pattern will have a partial structure annotation.

Open data repositories, which share reference MS/MS spectra, include GNPS [211], MassBank [90] and HMDB [226]. Further spectral reference libraries are METLIN [78], the mass spectral libraries provided by the National Institute of Standards and Technology (NIST), and the “MassHunter Forensics/Toxicology PCDL” library (Agilent Technologies, Inc.) which we will use to evaluate the Bayesian network scoring. Additionally to reference spectra, GNPS also contains a large number of public datasets and provides several workflows to analyze data. For more information on spectral reference libraries see Stein [191], Yang *et al.* [229].

4.4 Structure Database Search

The expansion of spectral libraries is driven by the availability of standards [183]. In particular, the large number of secondary metabolites is not covered by spectral libraries, and this will likely not change any time soon [18, 70]. This is a major obstacle to structure elucidation by MS/MS. Molecular structure databases, on the other hand, are orders of magnitude larger. Spectral libraries are limited to tens of thousands of different compounds, whereas structure databases contain millions. For example, PubChem [103] contains over a hundred million different compounds. Even if we only consider compounds of known biological interest, we end up with hundreds of thousands of structures [9, 63]. Moreover, novel candidate structures can be generated *in silico*; either systematically by enumerating all possible structures [77], limited to biotransformations of known structures [56, 95], or generated using autoencoders [32, 163].

However, it is not straightforward to deduce structural information solely from a mass spectrum. Early attempts at generating structural information from mass spectrometry data have been made since the 1970s as part of the DENDRAL project [34]. Unfortunately, they could not accomplish their ambitious goals and the project was discontinued [72]. Recent methods which search mass spectra in a structure database can be divided into three main categories [18]:

Simulating MS/MS spectra from structures. This approach may seem the most intuitive to chemists, who are used to searching in spectral libraries. Here, a library of artificial reference spectra is created by simulating spectra based on a database of structures. This only has to be done once for each structure in the database. Then, query spectra can be searched in the simulated spectral library (Section 4.3).

Quantum chemistry models can be used to simulate spectra *ab initio*. However, this has been applied mainly to EI [11, 190], and less to CID [109, 154]. This is not surprising since EI is much better understood and reproducible compared to CID. These methods demand considerable computational resources. Thus, usually only very small molecules are analyzed. Quantum chemical approaches can be useful to study fragmentation mechanisms an pathways [189].

Machine learning methods and stochastic models can be much faster. CFM-ID was first introduced to identify structures based on CID spectra [2] and later modified for EI spectra [3]. CFM-ID uses fixed-length Markov chains to predict the fragmentation of molecules. Fragments are generated by systematically breaking bonds; this process also considers hydrogen rearrangements. The probability that a fragment is generated is based on the break tendency of the bonds. This break tendency is predicted from a vector of chemical features describing a bond's proximity. The probabilistic process enables the prediction of peak intensities. In the current version of CFM-ID, the Markov chains' transition function is based on a neural network. Expectation-maximization is used to perform a maximum likelihood estimation of the parameters. Although the simulated spectra are not extremely similar to measured spectra, these seem well suited for the task of compound identification. In comparison to current quantum chemical models, CFM-ID performed similarly [190] but is considerable faster.

Rule-based methods simulate a spectrum by applying specific fragmentation rules to the molecule to generate a set of fragments. The rules are hand-curated from the literature. These methods usually compute "bar code" spectra, meaning that no peak intensities are predicted. Hill *et al.* [87] suggested to use the commercial software Mass Frontier to search in a structure database. However, "bar code spectra are not sufficient when many molecules generate the same fragment ions. In these instances, only the relative ion intensities will aid the correct identification." [98]. Furthermore, rule-based methods will always struggle with compounds which exhibit novel fragmentation mechanisms.

Comparing MS/MS spectra and structures. Combinatorial fragmentation methods systematically break bonds to generate a set of all possible fragments. These methods are not designed to simulate spectra. But the fragments are used to explain peaks in a measured spectrum. Initially, methods were intended for the guided interpretation of spectra where the compound structure is already known [83, 86]. Later, MetFrag [176, 228] used combinatorial fragmentation to search in a structure database. Fragment enumeration is usually not performed exhaustively. It can, for example, be confined by the number of consecutive fragmentation steps. Wolf *et al.* [228] found that this not only reduces running times, but also prevents the generation of too many false positive hits (fragments generated from incorrect structure candidates that match a peak in the measured spectrum). MetFrag uses bond dissociation energies to assign costs to the bonds. The cost of breaking a molecule into fragments is the sum of costs of the cleaved bonds. Ridder *et al.* [169] developed MAGMa, which is applicable to MS/MS and MS^{n}. It uses more simplistic costs: 1, 2 and 3 for single, double and triple/aromatic bonds respectively; with an additional modifier for carbon-carbon bonds. With this scoring it outperformed MetFrag [61, 147]. Clearly, an even improved scoring for predicting bond cleavage may be found or learned from the data [2, 98]. MAGMa+ [208] is a wrapper script for MAGMa: it runs MAGMa with one of a set of predefined parameter settings chosen based on the predicted metabolite class. MIDAS [213] uses simplistic bond costs as well. Resulting fragments are scored against the measured spectrum based on the relative intensity of the peaks and a fragment plausibility score; this plausibility score favors fragments whose direct parent fragment was assigned to a peak as well. MIDAS resolves hydrogen rearrangements simply by searching in the spectrum not only for a fragments' m/z, but additionally also for the m/z of the fragment modified by one or two hydrogen. Tsugawa *et al.* [202] proposed nine rules of hydrogen rearrangement during bond cleavage, implemented in MS-FINDER. DEREPLICATOR+ [134] limits fragmentation to N–C, O–C, and C–C bonds. Furthermore, it introduces a false discovery rate estimate: this is a target-decoy approach and decoy spectra are generated by sampling peaks from a spectral library. However, it has not been shown how well these estimates resemble the true false discovery rate. MetFusion [74] combines MetFrag with spectral library search in MassBank [90] to take advantage of both approaches.

Predicting structural properties from spectra. Instead of simulating spectra from molecular structures, structural information can be generated directly from the measured MS/MS spectrum. Full *de novo* structure elucidation is not possible [51]. Rather molecular fingerprints describing a set of substructure features are predicted. These fingerprints can be directly used to characterize the measured compound. Moreover, fingerprints can be searched in molecular structure databases. Early methods for predicting substructure features from mass spectra targeted GC/MS and EI fragmentation [34, 206, 207]. In 2012 Heinonen *et al.* [84] developed FingerID which predicts a molecular fingerprint from an MS/MS spectrum using kernel-based support vector machines (SVMs). Kernels are similarity functions that are used to compare feature vectors of the data (here, the MS/MS spectra); for more information see Section 4.4.1. FingerID uses a probability product kernel [108] to compare two spectra. Shen *et al.* [184] integrated fragmentation trees with FingerID: multiple kernel learning was applied to combine spectrum and fragmentation tree kernels into a single kernel which can be used by the SVM. Dührkop *et al.* [61] improved the scoring and added additional fingerprints; this resulted in CSI:FingerID. Input Output Kernel regression (IOKR) [29–31] is a modification of these approaches which does not perform the intermediate step of predicting fingerprints. Instead, a mapping between the MS/MS spectra and molecular structures is learned. Additional methods followed which were inspired by the CSI:FingerID and IOKR approaches [145, 146]. ChemDistiller combines fingerprint prediction with combinatorial fragmentation [115]. Besides, other tools exist which predict substructure features but do not search in a structure database [121].

4.4.1 CSI:FingerID

CSI:FingerID searches MS/MS spectra in a structure database [61]. It combines advantages of combinatorial optimization and machine learning approaches. The MS/MS spectrum is transformed into a fragmentation tree using SIRIUS (Section 4.2.2) — this part is based on combinatorial optimization. It is rather difficult for machine learning methods to pick up the same structured information as contained in the fragmentation tree, solely based on the spectrum. This is particularly true because the number of training examples (compounds in spectral reference libraries) is very limited. This is in contrast to other research areas such as image recognition where the availability of millions of training examples allows for better utilization of deep learning approaches.

CSI:FingerID predicts a molecular fingerprint using kernel SVMs. Firstly, it computes multiple kernels based on the spectrum and fragmentation tree of the query compound. These are then combined into a single kernel using multiple kernel learning. Kernel support vector machines predict each property of the fingerprint. Finally, this predicted fingerprint is searched in a structure database. The general workflow is illustrated in Fig. 4.3. In the following, we describe the CSI:FingerID structure database search. But first, we give a short introduction to the concept of kernels.

A very short introduction to kernels. In classification tasks, a supervised machine learning method predicts labels (classes) $y \in \mathcal{Y}$ for each input instance. Instances are described by feature vectors x from a feature space $\mathcal{X}$. In binary classification, a method aims to separate two possible classes $\mathcal{Y} = \{-1, 1\}$. The machine learning method learns a function $f : \mathcal{X} \to \mathcal{Y}$, based on training examples, that assigns labels to feature vectors. In case of a linear classifier, f is a linear combination of the input features.

Kernels can enable linear classifiers to solve a non-linear problem: by transforming the input space $\mathcal{X}$ into another, usually high-dimensional, space $\mathcal{H}$, the data may become linear separable. However, computations in the high-dimensional space can also be much more time-consuming. Kernels provide a solution to this problem. It can be shown, that if a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ fulfills all properties for being a kernel, then there exists some transformation into an Hilbert space $\mathcal{H}$ such that K computes the inner product of this Hilbert space [6]. Thus, if a classifier function uses the input feature vectors only to compute inner products, this inner product function can be replaced by a kernel. This concept is called *kernel trick*. Calculating the kernel function can be much more efficient than performing calculations in the high-dimensional space $\mathcal{H}$. For some kernels the corresponding Hilbert space even has an infinite number of dimensions. In practice, the space $\mathcal{H}$ does not have to be specified. It is sufficient to show that a function is a kernel to apply this machine learning trick. Kernel SVMs are one kind of classifier which make use of this kernel trick. See Fig. 4.2 for an example of a path kernel.

Multiple kernels can be combined into a new, single kernel; one possibility is adding up individual kernel values. This intends to increase the prediction performance.

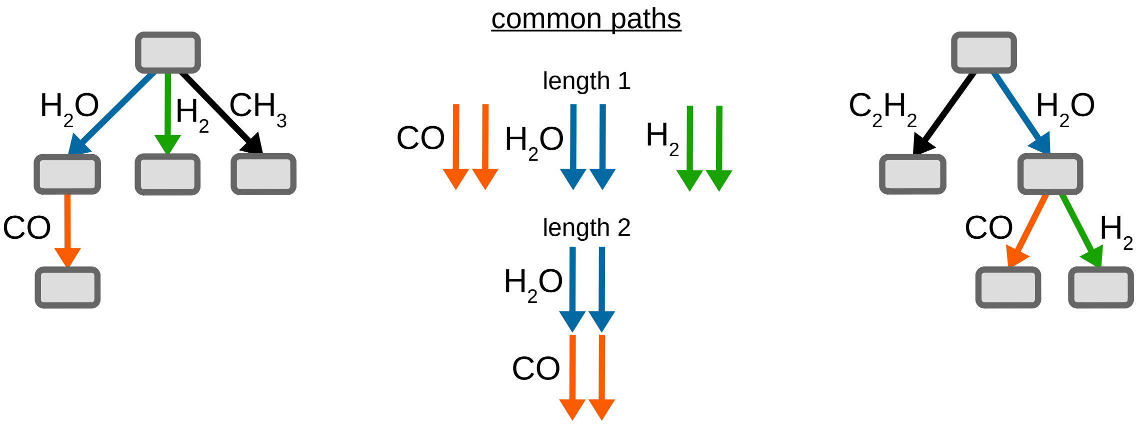


Figure 4.2: Common path counting kernel between fragmentation trees [184]. Two paths are considered equal if they have the same sequence of edge labels (losses). All common paths between the left and right fragmentation tree are displayed in the middle. The number of common paths is counted and normalized by the maximal possible number of common paths of each tree.

CSI:FingerID computes different kernels based on the spectrum, and on the nodes, edges or paths of the fragmentation tree. Each kernel gives a different perspective of the data and hence might be useful for predicting certain properties. However, simply adding up kernel values might not produce a kernel with good prediction performance. Multiple kernel learning selects appropriate weights to combine the different kernels. Dührkop *et al.* [63] use ALIGNF [45] and ALIGNF+ [185] to compute kernel weights. For more information on kernel methods see Hofmann *et al.* [89].

Searching in a structure database. In the following, we assume that the molecular formula of the query compound is known. In case it is unknown, fragmentation trees for all possible molecular formula candidates are computed and ranked by score; then, all top-ranked candidates can be searched individually in the structure database using CSI:FingerID [63].

CSI:FingerID predicts the molecular fingerprint and subsequently searches it in a molecular structure database. Molecular fingerprints encode the structure of a molecule: Most commonly, these are binary vectors of fixed length where each bit describes the presence or absence of a particular, fixed molecular property, usually the existence of a certain substructure (Section 2.2.1). Formally, let $1, \dots, n$ be the molecular properties; then, a (binary) *fingerprint* is a vector from $\{0, 1\}^n$. Each molecular structure has a (not necessarily unique) fingerprint assigned to it.

Given a query compound and its molecular formula, the fragmentation tree is computed from the MS/MS spectrum. Next, for each molecular property in the fingerprint the presence or absence of this property is predicted. Each property is predicted by a separate SVM; all SVMs use the same kernel. When searching in a structure database such as PubChem, first a set of molecular structure candidates is extracted that has the same molecular formula as the query compound. Each structure candidate is deterministically transformed to a binary fingerprint. Then, predicted and binary fingerprints are compared in order to rank the candidates.

Different scorings have been proposed to compare the predicted fingerprint with the binary fingerprints in the database. Unit scores simply count the number of differences

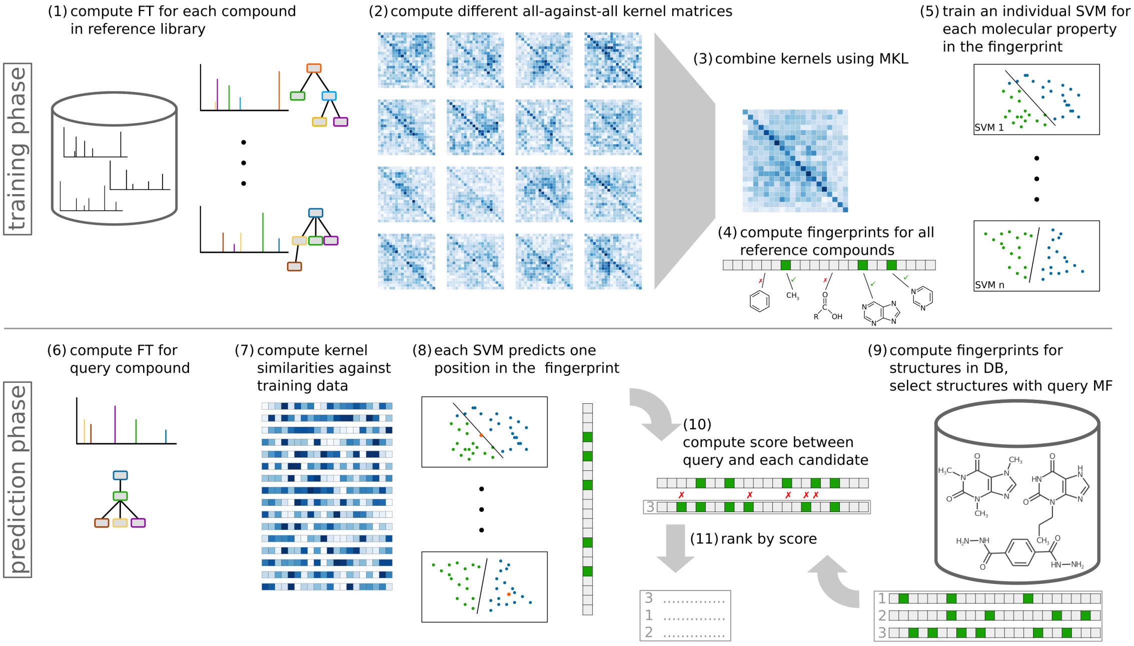

training phase
(1) compute FT for each compound in reference library
(2) compute different all-against-all kernel matrices
(3) combine kernels using MKL
(4) compute fingerprints for all reference compounds
(5) train an individual SVM for each molecular property in the fingerprint
SVM 1
SVM n
prediction phase
(6) compute FT for query compound
(7) compute kernel similarities against training data
(8) each SVM predicts one position in the fingerprint
(9) compute fingerprints for structures in DB, select structures with query MF
(10) compute score between query and each candidate
(11) rank by score
3
1
2

Figure 4.3: CSI:FingerID workflow. In the *training phase* MS/MS spectra of reference compounds are used. (1) Fragmentation trees (FTs) are computed from the spectra. (2) Different kernels are computed based on spectra and FTs. (3) These kernels are combined using multiple kernel learning (MKL). (4) True fingerprints of reference compounds are deterministically computed. (5) One SVM is trained for each molecular property in the fingerprint. In *prediction phase* the structure of the query compound is unknown; we assume the molecular formula (MF) is known. (6) The FT is computed, (7) kernel values against training compounds are computed, and (8) a probabilistic fingerprint is predicted. (9) Deterministic fingerprints of structures in the database (DB) with the query MF are selected and (10) compared against the probabilistic fingerprint. (11) Structures are ranked by their scores.

between the predicted fingerprint and each candidate fingerprint. Heinonen *et al.* [84] used the accuracy of individual SVMs to weight the scoring, but this does not perform better than unit scoring [61]. Dührkop *et al.* [61] suggested and evaluated different scoring variants, and found that two variants consistently outperformed all others in evaluation: Namely, the “Platt” score and the “modified Platt” score.

Both scores use *Platt probabilities* [160] for fine-grained predictions: Instead of a binary prediction of a SVM, a sigmoid function is used to predict the posterior probability for the presence of the molecular property, with parameter estimated from the training data to predict this probability. Let $\mathcal{D} = (p_1, \ldots, p_n) \in [0,1]^n$ be the Platt probability estimates, and let $\mathcal{M} = (x_1, \ldots, x_n) \in \{0,1\}^n$ be a candidate fingerprint; assuming independence between all molecular property pairs, the posterior probability of the fingerprint candidate $\mathcal{M}$ can be estimated as

$$\mathbb{P}(\mathcal{M} \mid \mathcal{D}) = \prod_{i=1,\ldots,n} \begin{cases} p_i & \text{if } x_i = 1, \\ 1 - p_i & \text{if } x_i = 0. \end{cases} \tag{4.1}$$

This has been referred to as “Platt score” in [61]; maximizing this score corresponds to a maximum a posteriori estimator, and results in about 3.5 percentage points more correct identifications than unit scores. In contrast, the “modified Platt” score from [61] was found by trial and error, combines Platt probabilities and sensitivity / specificity estimates of the binary predictors in a counterintuitive fashion: Namely,

$$\prod_{i=1}^{n} \begin{cases} p_i^{0.75} \cdot (1 - \mathit{sens}_i)^{0.25} & \text{if } p_i \geq 0.5 \text{ and } x_i = 1 \\ (1 - p_i)^{0.75} & \text{if } p_i \geq 0.5 \text{ and } x_i = 0 \\ p_i^{0.75} & \text{if } p_i < 0.5 \text{ and } x_i = 1 \\ (1 - p_i)^{0.75} \cdot (1 - \mathit{spec}_i)^{0.25} & \text{if } p_i < 0.5 \text{ and } x_i = 0 \end{cases} \tag{4.2}$$

where sens_i is the sensitivity and spec_i the specificity of the ith binary predictor. While this score has no known statistical interpretation, modified Platt (4.2) consistently outperforms the Platt score (4.1) by a margin of about 1.5 percentage points. In Chapter 6 we present an improved scoring for CSI:FingerID based on Bayesian networks.

5 ZODIAC: De Novo Molecular Formula Annotation

In this chapter, we present ZODIAC (ZODIAC: Organic compound Determination by Integral Assignment of elemental Compositions) for confident, *de novo* molecular formula annotation in LC-MS/MS datasets [125]. ZODIAC takes advantage of the fact that metabolites co-occur in a network of derivatives. Thus, annotations of each individual compound can be improved by considering similar compounds in the dataset.

Other methods have utilized similar concepts, as discussed in Section 4.2 and 4.3. Network visualization approaches which connect compounds by hypothetical biotransformations and indicate pairwise similarities, facilitate manual interpretation and have gained popularity recently [136, 203, 211, 214]. Moreover, automated approaches to annotate molecular formulas for a complete LC-MS run use Gibbs sampling and Bayesian statistics, utilizing co-occurrence of molecular formulas differing by a predefined set of biotransformations [47, 50, 52, 174]. However, these approaches suffer from a database bias since, implicitly, molecular structures (or their isomers) from a restricted structure database are considered. Besides, they cannot annotate novel molecular formulas.

ZODIAC builds upon SIRIUS [63]. Both methods determine molecular formulas *de novo*. Thus, they are not limited by any database but consider all possible formulas. SIRIUS considers each compound individually and ranks molecular formula *candidates* based on fragmentation tree scores. ZODIAC uses the top-scoring candidates of each compound from SIRIUS and reranks these, now considering relations between compounds in the dataset. The probability model considers fragmentation tree scores as likelihoods. Prior probabilities are derived from fragmentation tree similarities between candidates of different compounds within an LC-MS/MS dataset. In this way, similar candidates support each other's plausibility. ZODIAC does not rely on a predefined set of biotransformations. We apply Gibbs sampling to estimate the posterior probabilities as illustrated in Fig. 5.1. Thorough algorithm engineering ensures fast processing in practice.

In the next section, we present the method. On the theoretical side, we show that finding an optimal assignment of candidates is NP-hard. We describe the estimation of likelihoods and prior probabilities based on fragmentation trees and how to implement a swift Gibbs sampler for the problem. Furthermore, we explain how to include spectral library hits into the ZODIAC network. Then, we perform an evaluation on five diverse datasets and find that ZODIAC can increase the number of correct molecular formula annotations on each dataset. We proof the practical advantage of *de novo* annotation by discovering several compounds that have novel molecular formulas not present in any structural databases; we verify two of these.

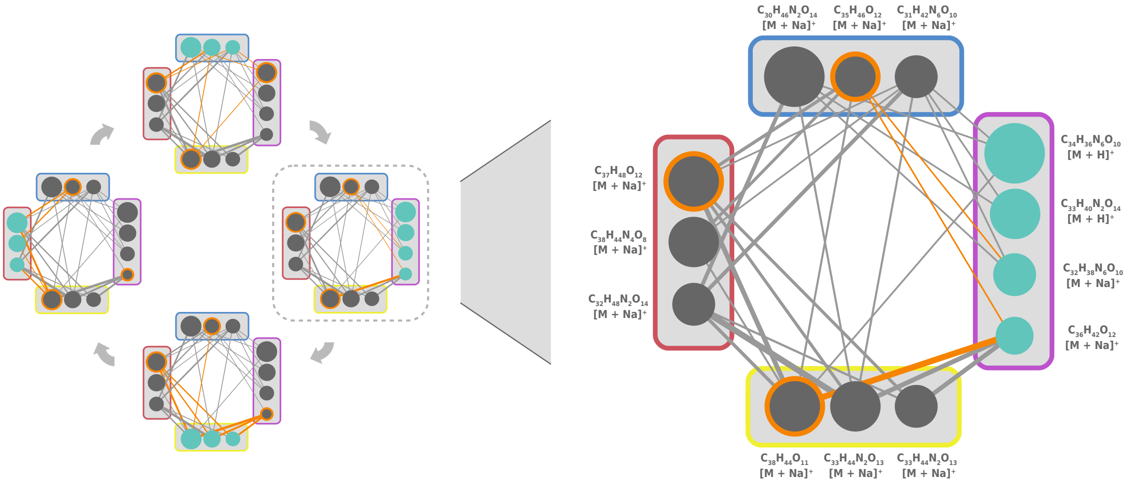

$C_{30}H_{46}N_2O_{14}$ $[M + Na]^+$
$C_{35}H_{46}O_{12}$ $[M + Na]^+$
$C_{31}H_{42}N_6O_{10}$ $[M + Na]^+$
$C_{37}H_{48}O_{12}$ $[M + Na]^+$
$C_{38}H_{44}N_4O_8$ $[M + Na]^+$
$C_{32}H_{48}N_2O_{14}$ $[M + Na]^+$
$C_{34}H_{36}N_6O_{10}$ $[M + H]^+$
$C_{33}H_{40}N_2O_{14}$ $[M + H]^+$
$C_{32}H_{38}N_6O_{10}$ $[M + Na]^+$
$C_{36}H_{42}O_{12}$ $[M + Na]^+$
$C_{38}H_{44}O_{11}$ $[M + Na]^+$
$C_{33}H_{44}N_2O_{13}$ $[M + Na]^+$
$C_{33}H_{44}N_2O_{13}$ $[M + Na]^+$

Figure 5.1: Illustration of the Gibbs sampling process. Left: During each epoch compounds are iterated in random order. For each compound one new active molecular formula candidate is sampled based on prior probabilities and active candidates of all other compounds. Right: Sampling step for one compound. This illustrated sub-network with four compounds is based on the dendroides evaluation dataset from Section 5.2. Each circle corresponds to a molecular formula candidate. The size depicts the rank estimated by SIRIUS. Orange rings mark active candidates. Edge width depicts fragmentation tree based similarity between candidates. The set of candidates from which a new active candidate is sampled in this step is colored cyan.

5.1 A Similarity Model for Molecular Formula Assignments

5.1.1 Posterior Probability of an Assignment

We use a probabilistic view on the molecular formula assignment problem [174]: For each hypothetical compound in the LC-MS run, we are given data such as an isotope pattern and a fragmentation pattern. This allows us to determine, for each compound $c \in \mathcal{C}$, a set of candidate molecular formulas that may explain the observed data. Let V be the set of all molecular formula candidates, such that $V(c) \subseteq V$ is the subset of molecular formulas for compound $c \in \mathcal{C}$. It is possible that different compounds share an identical molecular formula explanation, but we ignore this in our presentation, solely for the sake of readability. An *assignment* is a mapping $\mathfrak{a} : \mathcal{C} \to V$ where $\mathfrak{a}(c) \in V(c)$ is the molecular formula assigned to compound c. The posterior probability of an assignment $\mathfrak{a}$ is

$$\mathbb{P}(\mathfrak{a} \mid \mathcal{D}) = \frac{\mathbb{P}(\mathcal{D} \mid \mathfrak{a}) \cdot \mathbb{P}(\mathfrak{a})}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D} \mid \mathfrak{a}) \cdot \mathbb{P}(\mathfrak{a}) \tag{5.1}$$

where $\mathcal{D}$ is the observed MS1 and MS/MS data. Let $\mathcal{D}(c)$ be the observed data for compound $c \in \mathcal{C}$, that is, the isotope pattern and fragmentation pattern of c. We assume that the likelihoods of molecular formulas for different compounds are independent, and that the likelihood of any compound c only depends on its data $\mathcal{D}(c)$; so,

$$\mathbb{P}(\mathcal{D} \mid \mathfrak{a}) = \prod_{c \in \mathcal{C}} \mathbb{P}\big(\mathcal{D}(c) \mid \mathfrak{a}(c)\big).$$

Next, we define the prior probability of an assignment as the product of priors for pairs of compounds:

$$\mathbb{P}(\mathfrak{a}) \propto \prod_{c,c' \in \mathcal{C}, c \neq c'} \prod_{u \in V(c)} \prod_{v \in V(c')} \mathbb{P}(u, v \mid \mathfrak{a}(c) = u, \mathfrak{a}(c') = v).$$

Here, $\mathbb{P}(u, v \mid \text{“true”})$ is the prior probability that two compounds with molecular formulas u, v co-occur in the dataset; analogously, $\mathbb{P}(u, v \mid \text{“false”})$ if u, v do not co-occur. To simplify our calculations, we introduce a mapping $\mathfrak{c} : V \to \mathcal{C}$ that maps any molecular formula to the compound it belongs to: $\mathfrak{c}(v) = c$ for all $v \in V(c)$, for $c \in \mathcal{C}$. Note that $\mathfrak{c}(\mathfrak{a}(c)) = c$ for all $c \in \mathcal{C}$. Now,

$$\mathbb{P}(\mathfrak{a} \mid \mathcal{D}) \propto \prod_{c \in \mathcal{C}} \mathbb{P}\big(\mathcal{D}(c) \mid \mathfrak{a}(c)\big) \cdot \prod_{u,v \in V, \mathfrak{c}(u) \neq \mathfrak{c}(v)} \mathbb{P}\big(u, v \mid \mathfrak{a}(\mathfrak{c}(u)) = u, \mathfrak{a}(\mathfrak{c}(v)) = v\big). \tag{5.2}$$

Different from Rogers *et al.* [174], we are able to formulate the posterior probability of an assignment in closed form. A natural question is if we can find a maximum a posteriori estimate for (5.2); unfortunately, we will see that this is not easy, as the underlying computational problem is NP-hard. Another natural question is how to sample from the posterior distribution; this will be addressed below.

5.1.2 Graph-theoretical Formulation

We now give a graph-theoretical formulation of the problem; this will allow us to establish its computational complexity, but also to come up with a more efficient algorithm. Let V, the molecular formula candidates, be the nodes of an undirected graph $G = (V, E)$ with edge set $E \subseteq \binom{V}{2}$. We will write uv as shorthand for a tuple $\{u, v\} \in \binom{V}{2}$. We use $\mathfrak{c} : V \to \mathcal{C}$ as a *node coloring* with color set $\mathcal{C}$. Now, an *assignment* is a subset $A \subseteq V$ such that each color from $\mathcal{C}$ appears exactly once; in this case, A is also called *multicolored.* Using the notation of the previous section, we have $A = \mathfrak{a}(\mathcal{C})$; recall that $\mathfrak{c}(\mathfrak{a}(c)) = c$ for all $c \in \mathcal{C}$. Let $w : V \cup E \to \mathbb{R}$ be *weights* for all nodes and edges of the graph. The *weight* of the assignment A is

$$w(A) := \sum\nolimits_{v \in A} w(v) + \sum\nolimits_{uv \in E, u,v \in \mathcal{A}} w(uv)\,. \tag{5.3}$$

This corresponds to the node plus edge weights of a node-induced subgraph of G, for node set $A \subseteq V$.

We consider the following optimization problem:

Maximum Multicolored Subgraph problem. We are given a graph $G = (V, E)$, a node coloring $\mathfrak{c} : V \to \mathcal{C}$ and weights $w : V \cup E \to \mathbb{R}$. We search for an assignment $A \subseteq V$ of maximum weight, that is, a node-induced multicolored subgraph of maximum weight.

How does this problem correspond to our probabilistic problem from the previous section? Setting $E = E^* := \big\{uv \mid u, v \in V, \mathfrak{c}(u) \neq \mathfrak{c}(v)\big\}$ (the set of all node pairs with different colors) and

$$w(v) := \log \mathbb{P}\big(\mathcal{D}(\mathfrak{c}(v)) \mid v\big) \quad \text{and} \quad w(uv) := \log \mathbb{P}(u, v \mid \text{“true”}) - \log \mathbb{P}(u, v \mid \text{“false”}) \tag{5.4}$$

we can show that these problems are in fact equivalent: We have $\log \mathbb{P}(\mathfrak{a} \mid \mathcal{D}) = w(\mathfrak{a}(\mathcal{C})) + \alpha$ for some constant $\alpha \in \mathbb{R}$. Here, we assumed that $E = E^*$ contains all possible edges; we call (V, E^*) a *complete assignment graph.* But we can encode any edge set $E \subsetneq E^*$ using zero edge weight for all $e \notin E$, so both problems are equivalent. Hence, it is natural to ask for an optimal solution of the problem, which would correspond to a maximum a posteriori estimator.

5.1.3 Complexity of the Problem

For the decision version, we ask if there is an assignment with weight above some threshold $\tau \in \mathbb{R}$. In its simplest form, all edges have weight one and all nodes have weight zero, $w|_E \equiv 1$ and $w|_V \equiv 0$.

Lemma 5.1. *The* MULTICOLORED SUBGRAPH *problem is NP-complete, even for unit edge weights and zero node weights.*

Proof. It is clear that the MULTICOLORED SUBGRAPH problem is in NP. We show that the problem is NP-hard by reduction from CLIQUE [99]: Let $G = (V, E)$ be an undirected, simple graph, is there a clique of size k in G? Clearly, $k \leq n := |V|$.

We construct a graph $H := G \,\square\, \bar{K}_k$ as the Cartesian graph product of G and the empty graph $\bar{K}_k$ with k nodes and no edges: That is, for every node $v \in V$ we generate k copies $(v, 1), \ldots, (v, k)$ in H, and there is an edge $\{(u, i), (v, j)\}$ with $i \neq j$ in H if and only if there is an edge uv in G. Now, $k \leq n$ implies that H contains at most n^2 nodes. We define node colors $1, \ldots, k$ such that $c\big((v, i)\big) = i$ for $v \in V$ and $1 \leq i \leq k$. We assign zero node weights and unit edge weights for all nodes and edges in H. Now, any assignment in H corresponds to a k-node induced subgraph in G, and the weight of the assignment equals the number of edges in the node-induced subgraph; to this end, an assignment of weight $\binom{k}{2}$ would correspond to a k-clique in G. □

The Multicolored Subgraph problem is a generalization of the Multicolored Clique problem; to this end, Lemma 5.1 can also be inferred from the complexity of Multicolored Clique, which is W[1]-hard [57]. Assuming zero node and unit edge weights, the above construction implies that for any $\epsilon > 0$, there is no polynomial time algorithm that approximates the maximum assignment weight to within a factor better than $O(n^{1-\epsilon})$, unless P = NP [234]. Furthermore, finding an assignment of weight k cannot be done in time $n^{o(k)}$, unless the exponential time hypothesis fails [41, 94]. Finally, we noted above that we can encode an arbitrary edge set $E \subsetneq E^*$ using zero edge weight for all $e \notin E$, so:

Corollary 5.1. *The* MULTICOLORED SUBGRAPH *problem is NP-complete, even for a complete assignment graph, binary edge weights and zero node weights.*

Finally, we consider two problem variants: First, we may allow that some colors from $\mathcal{C}$ are absent from A; in this case, A is called *colorful.* We can encode this variant in the original problem, by adding a dummy node for each color which is connected to no other node. Second, we may assume that only edges carry weight. We can encode the MULTICOLORED SUBGRAPH problem in this variant, by adding a dummy color for each color and a dummy node for each node, such that if a node has a certain color, then the dummy node has the corresponding dummy color. We connect each node to its dummy node, and transfer the weight of the node to the corresponding edge. Hence, our complexity results also hold for these variants.

On the algorithmic side, it is easy to see that the MULTICOLORED SUBGRAPH problem can be solved by a simple Integer Linear Program (one variable per edge and one variable per color). We omit the straightforward technical details. We will not proceed in this direction, as this approach results in a single optimal solution, whereas we want to consider suboptimal solutions and marginal probabilities, which allow us to judge our individual confidence when assigning molecular formulas to compounds.

5.1.4 Likelihoods, Prior Probabilities and Graph Topology

The likelihood $\mathbb{P}\big(\mathcal{D}(\mathfrak{c}(v)) \mid v\big)$ of a molecular formula candidate v can be computed from the posterior probability of the fragmentation tree and the isotope pattern analysis as estimated by SIRIUS 4.0 [19, 63]. For the Gibbs sampler, we treat these probabilities as likelihoods, although the analysis SIRIUS 4.0 also integrates certain priors [19]. To avoid proliferating running times, we usually limit further computations to the, say, 50 best-scoring molecular formulas for each compound. For each compound, we also introduce a node representing "molecular formula not identified" which receives likelihood from the remaining molecular formulas, and is not connected to any other nodes.

Furthermore, we assume that some compounds were identified by searching in a library of MS/MS spectra, plus potentially by comparison of retention times. We refer to these compounds and the corresponding molecular formulas as "*anchors*". Such library search results can also be wrong, so we do not exclude other molecular formula explanations, but rather give a bonus to the likelihood of the identified molecular formula. The "quality" of a spectral library hit can, to a certain extend, be evaluated using its score, usually the dot product (cosine score) between query and reference. Hence, the bonus may be dependent on the corresponding library search score. Given the library search score $s_l \in [0, 1]$ and a minimum score to consider a library hit min_l, we multiply the candidate's likelihood by

$$\psi(s_l, min_l) = exp\big(\lambda \frac{\max(s_l, min_l)}{1 - max(s_l, min_l)}\big). \tag{5.5}$$

Candidates which disagree with the library hit or without any library hit are scored using $s_l = min_l$. Note, that any "perfect match" with score of 1.0 will be chosen in any case. We remove any other candidate for this compound. We refrain from normalizing the ψ to one.

For estimating priors, we will consider similarity of fragmentation patterns [136, 214]: More precisely, we use similarity between fragmentation trees that were computed by SIRIUS in the previous step. For each pair of compounds, we have to compare up to 50 times 50 fragmentation trees: For swift computations, we refrain from using fragmentation tree alignments [166] but instead, simply count the number of common fragments and precursor (root) losses in the two trees [166]. Root losses can be directly derived from fragments by subtracting the fragment molecular formula from the precursor molecular formula. Evaluations indicate that this method, while performing worse than fragmentation tree alignments, is still able to detect structural similarity between compounds [166]. When counting common root losses, the empty root loss is ignored. We introduce two modifications to the score from [166]: Let n_1, n_2 be size of the two fragmentation trees, defined by the number of fragments and root losses. Instead of normalizing the number of common fragments plus root losses s by the size of the smaller tree $\min\{n_1, n_2\}$, we use

$$s/n_1 + s/n_2 \tag{5.6}$$

as the normalized score; by this, we slightly penalize large trees, as having common fragments or root losses is more likely against a large than a small tree. But this score favors small trees and, hence, inferior molecular formula candidates. To this end, we use the size of the *largest* fragmentation tree, among all candidate molecular formulas, for the normalization of each compound; this is the maximum number of explainable peaks in the MS/MS data of the compound. Fragments and root losses can be weighted by importance ι. The weight of two common fragments or root losses m_1 and m_2 is $\iota(m_1)\iota(m_2)$. The weighted size of a tree is

$$n_w = \sum_{g \in F} (\iota(g)) + \sum_{h \in R} (\iota(h)) \tag{5.7}$$

with fragments F and root losses R. For two molecular formulas $u, v \in V$ we denote the resulting score as $s(u, v)$.

How can we transform this count into a prior probability? Natural choices include significance estimates such as p-values and posterior error probabilities. We do not have a reasonable model for the score distribution of "true" edges; in fact, it is not know how to clearly distinguish between "true" and "false" edges in such a model. To this end, we resort to a simple prior based on p-value estimation:

$$\mathbb{P}(u, v \mid \text{“true”}) = f(\tau) \quad \text{and} \quad \mathbb{P}(u, v \mid \text{“false”}) = \begin{cases} f(s(u,v)) & \text{if } s(u,v) \geq \tau, \\ f(\tau) & \text{otherwise,} \end{cases} \tag{5.8}$$

where $\tau \in \mathbb{R}$ is a thresholding parameter, and $f : \mathbb{R} \to [0, 1]$ is a monotonically decreasing function. We introduce threshold τ because scores below a certain threshold are practically uninformative and should not be considered in our estimations. For $f(x)$ we estimate the p-value of score x, under the null model that scores follow a certain distribution. Note that prior probabilities do in fact depend upon the (mass spectrometry) data.

We now assign node and edge weights according to (5.4). Clearly, many of these edges have zero weight and can be removed from the graph. To avoid that nodes are isolated, we want to keep some edges incident to any node. This can be formulated by *individual thresholds* $\tau_c \in \mathbb{R}$ for each color $c \in \mathcal{C}$ and, for an edge uv, edge weight

$$w(uv) := \max\bigl\{0, -\log f(s(u,v)) + \log f(\tau_{uv})\bigr\} \tag{5.9}$$

for threshold $\tau_{uv} := \min\{\tau_{c(u)}, \tau_{c(v)}\}$. This will change the weight of any assignment by an additive constant and, hence, posterior probability by a multiplicative constant.

5.1.5 (Faster) Gibbs Sampling

We say that a node v is *active* in an assignment A if $v \in A$, and that an edge uv is *active* if both $u \in A$ and $v \in A$; then, the weight of an assignment is the sum of weights of all active nodes and edges.

Gibbs sampling is a Markov chain Monte Carlo algorithm for obtaining a sequence of observations approximated from a multivariate probability distribution [73]. Sampling assignments according to (5.2) can be seen as an archetype application of a Gibbs sampler: We start with some assignment, such as the highest likelihood node (molecular formula) for each compound (color). Each epoch of the Gibbs sampler consists of $|\mathcal{C}|$ *steps*, where we iterate over all colors $c \in \mathcal{C}$ in random order: We update the active node with color c by drawing a node with color c according to its posterior probability, conditional the current assignment of all nodes with color different from c. At the end of the epoch we output the current assignment, and repeat until we have reached a sufficient number of samples. This generates a Markov chain of samples converging to the posterior probability distribution of assignments. In practice, we discard samples from the beginning of the chain (*burn-in period*), and to avoid correlation between nearby samples, we output only every, say, 10^{th} sample.

Assume that $u \in A$ with color $c := \mathfrak{c}(u)$ is to be (potentially) replaced by a new node v with the same color. Such update results in the new assignment $A - \{u\} \cup \{v\}$. The probability of $v \in V(c)$, conditional all other nodes $z \in A$ with $\mathfrak{c}(z) \neq c$, can naïvely be computed as

$$\mathbb{P}\big(v \mid A - \{u\}\big) \propto \exp\Big(w(v) + \sum\nolimits_{z \in A,\, vz \in E} w(vz)\Big) . \tag{5.10}$$

Computing all conditional probabilities for drawing a node v, requires time proportional to the sum of node degrees for all nodes from $V(c)$. That means running time for one step is of order $O(|V(c)| \cdot |V|)$ and, hence, $\Theta(|V|^2)$ for certain graph families.

To apply Gibbs sampling in practice, the critical point is to quickly reach a large number of samples, so that probability estimates become reliable. To further decrease running time, we assume that we have, at any step, knowledge about all (log) conditional probabilities, for all nodes $v \in V(c)$ and all colors $c \in \mathcal{C}$. We assume that conditional probabilities are not normalized; to sample a new active node, we uniformly draw a random number between zero and the sum of conditional probabilities, over all nodes with this color. To improve the sampling speed, we want to estimate conditional probabilities without performing a full calculation using (5.10). Hence, we update conditional probabilities from (5.10), which were calculated in a previous step, by only adding or removing specific edge weights based on nodes added or removed from $\mathcal{A}$.

Lemma 5.2. *One step of the Gibbs sampler, replacing some node u by another node v with the same color $c := \mathfrak{c}(u) = \mathfrak{c}(v)$ in A, can be carried out in $O\big(|V(c)| + \deg(u) + \deg(v)\big)$ time.*

Proof. Firstly, we update the active node of color c. Secondly, we update conditional probabilities of all $z \in V$ by accounting for the replaced active node of color c in A.

Let $A \subseteq V$ be the current assignment with $u \in A$. Firstly, we want to choose a new node $v \in U$ from the set of candidate nodes $U := V(c)$ for color $c := \mathfrak{c}(u)$. We assume that the conditional probabilities $\mathbb{P}(v \mid A - \{u\})$ have been calculated for all $v \in U$. We sum up the conditional probabilities, then uniformly choose a random number between zero and this sum and, finally, use this random number to select one $v \in U$. This can be carried out in time $O(|U|)$, because we sum up conditional probabilities of all candidates in U. Selecting one candidate can be performed in $O(\log |U|)$ using a binary search on the cumulative probabilities of these candidates. If $u = v$ then we can stop at this point.

Second, we have to estimate conditional probabilities for all nodes $z \in V$. From (5.10), we infer that the conditional probability only changes for those nodes z where there is a change in the neighborhood $N(z)$ of z, and remains constant for all others. To this end, we iterate over all $z \in N(u)$, and decrease the log conditional probability of z by $w(uz)$; then, we iterate over all $z \in N(v)$, and increase the log conditional probability of z by $w(vz)$. Finally, for any node $z \in N(u) \cup N(v)$, we recompute its conditional probability using the exponential function. This can be carried out in time $O(\deg(u) + \deg(v))$; afterward, all conditional probabilities are correct for the new assignment $A - \{u\} \cup \{v\}$. □

Comparing a naïve graph-based implementation of a Gibbs sampler with one that uses Lemma 5.2, we can estimate that the speedup is of order $\Theta(|V(c)|)$.

For the first iteration, we use an arbitrary assignment, then compute all conditional probabilities using (5.10). The method requires $O(|V| + |E|)$ memory for storing the graph, and $O(|V|)$ memory for storing (log) conditional probabilities. The probability of a particular molecular formula v to be correct, can now be estimated as its marginal probability: that is, the ratio of assignments in the output that contain v.

5.1.6 Faster network creation

Sampling assignments requires calculated edge weights $w(uv)$ for all $u, v \in V$. These calculations are performed once, prior to the sampling. Roughly $|V|^2$ candidate pairs have to be scored (candidates of the same color are not compared). Although, we do not obtain a better upper bound on the running time, we can greatly reduce the number of pairwise candidate comparisons in practice.

We establish lower bounds on pairwise scores $s(u, v)$ on a per-compound (color) basis. Thus, if $\mathfrak{c}(u)$ and $\mathfrak{c}(v)$ are highly dissimilar, we can refrain from computing $s(u, v)$, since $w(uv)$ will be 0 as can be seen in (5.9). Lower bounds $lb(c, c')$ for $c, c' \in \mathcal{C}$ and $c \neq c'$ are calculated as follows: For each compound $c \in \mathcal{C}$, we assign (and group) all fragments of all candidates $u \in V(c)$ to their corresponding peak; for root losses we do this analogously, but instead of considering the peak m/z we use the root loss m/z. Now we can score compounds in the same way as candidates: by counting common fragments plus root losses using (5.6). Given two compounds, we compare peaks between these compounds pairwise. We count a match if the peaks share at least one common fragment (or root loss).

Lemma 5.3. *By considering lexicographical orderings of peaks and fragment molecular formulas, the scoring of two compounds can be performed in linear time in the number of peaks (biggest tree size of the compound) and linear in time in the number of fragments or root losses per peak (which corresponds to $|V(c)|$).*

Proof. Let F_1 and F_2 be totally ordered sets of fragments. Such an ordering can be achieved by considering fragment molecular formulas as strings. We can count common pairs of fragments $f_1 = f_2$ with $f_1 \in F_1, f_2 \in F_2$ in linear time. To do so, we start to compare the first fragment f_1 of F_1 and f_2 of F_2. If $f_1 = f_2$ we count a match and set f_1, f_2 to the next fragments in F_1 and F_2. If $f_1 < f_2$ we set f_1 to the next fragment in F_1, else $f_1 > f_2$ and we set f_2 to the next fragment in F_2. In each step, we update at least one new fragment. Hence, counting common pairs requires $O(|F_1| + |F_2|)$. A fragment set is generated from fragment explanations of one specific peak. These explanations are based on fragmentation trees of candidates of the corresponding compound. Let F_1 and F_2 be fragment sets generated for a specific peak of compounds $c_1, c_2 \in \mathcal{C}$. Each candidate explains each fragment at most once. Thus, $|F_1| \leq |V(c_1)|$ and $|F_2| \leq |V(c_2)|$. For $|V(c_1)| < |V(c_2)|$ we obtain $O(|V(c_2)|)$. Clearly, testing sets for at least one common pair between F_1 and F_2, instead of counting common pairs, can be carried out in the same time.

Analogously, we can compare sets of root losses. For comparing peak sets (spectra) S_1 and S_2 of different compounds (based on shared fragments or root losses) we use the peak mass (the mean mass of corresponding fragments) to establish an order. Furthermore, we assume that the masses of different peaks of one compound differ by at least the maximum allowed mass accuracy. This ensures that if fragment sets $F_1 \in S_1$ and $F_2 \in S_2$ share common fragments $F_1 \cap F_2 \neq \emptyset$, the corresponding peaks are always counted as match and cannot be missed because there is an $F_1' \in S_1$ with mass closer to F_2. □

In total, compound comparisons can be carried out in time

$$O\left(\sum\nolimits_{c,c'\in\mathcal{C}} \max\{|V(c)|,|V(c')|\}\cdot n\right) = O(|V|\cdot n\cdot|\mathcal{C}|),$$

where n is the maximum fragmentation tree size, which corresponds to the maximum number of explained peaks per spectrum. This is considerably faster than the total running time of all pairwise comparisons of candidates which are performed in $O(|V|^2 \cdot n)$. Note, that in general $|\mathcal{C}|$ is much smaller than $|V|$. Moreover, in practice, pairwise compound comparisons are rather efficient: firstly, most compounds do not share many peaks, and fragments or root losses have to be compared only for shared peaks; secondly, peaks do usually not have $|V(c)|$ different fragment explanations but instead some candidates explain a peak with the same molecular formula.

For two candidates u, v we can refrain from calculating $s(u,v)$ iff $lb(\mathfrak{c}(u),\mathfrak{c}(v)) \leq \tau$. When using individual thresholds, which can only be established on the basis of known edge scores, only pairs with $lb(\mathfrak{c}(u),\mathfrak{c}(v)) \leq 0$ can be ignored.

5.2 Evaluation on five Biological Samples

We evaluate ZODIAC on five diverse LC-MS/MS datasets representing samples from plants (dendroides, tomato), human plasma (NIST1950), marine microalgae (diatoms) and mice fecal sample (mice stool). Throughout this evaluation, the entities of interest consist of signals detected by mass spectrometry for which one or more MS/MS spectra have been recorded by the instrument. It is understood that not all of these signals correspond to compounds in the biological sample; but clearly, only those signals that do correspond to compounds are of interest for our analysis. It is also understood that we usually cannot ultimately decide whether a certain signal stems from the protonated molecule $[\mathrm{M}+\mathrm{H}]^+$ or, say, the protonated molecule with a water loss $[\mathrm{M}-\mathrm{H_2O}+\mathrm{H}]^+$ or an ammonia adduct $[\mathrm{M}+\mathrm{NH_3}+\mathrm{H}]^+$. This is not a problem of our method but rather a general problem of mass spectrometry. For the sake of readability, in the following we will use the term “compound” instead of “hypothetical compound”, “feature”, “adduct” or “ion”. In contrast, our methods decide for each compound if it is protonated $[\mathrm{M}+\mathrm{H}]^+$, a sodium adduct $[\mathrm{M}+\mathrm{Na}]^+$ or a potassium adduct $[\mathrm{M}+\mathrm{K}]^+$; in evaluation, compounds that are assigned a wrong adduct are also assigned a wrong molecular formula and, hence, are always counted as misannotations. We process the LC-MS/MS runs of each dataset with OpenMS [175] to obtain a set of “good quality” compounds. These compounds are all unknowns, and we first have to establish a “ground truth” to evaluate against. For this, we use manually annotated molecular formulas and spectral library search. ZODIAC runs on all compounds, but evaluation is only performed on the set of ground truth compounds.

Arguably the best-performing computational method for molecular formula annotation is SIRIUS 4 [63], combining isotope pattern matching (Section 4.2.1) and fragmentation trees (Section 4.2.2). But even SIRIUS has problems annotating molecular formulas for compounds above 500 Da: Böcker and Dührkop [19] found that the percentage of correctly identified molecular formulas dropped substantially for larger masses.

Two questions are specifically interesting to evaluate: Does ZODIAC increase the number of correct annotations compared to SIRIUS 4? And can the ZODIAC score differentiate between correct and incorrect annotations? If so, we could easily select confident assignments from any dataset. To answer the first question, we limit evaluation to compounds for which the correct answer is contained in the list of best-scoring molecular formula candidates from SIRIUS 4. Remember, this list is input to ZODIAC and if the correct molecular formula is not contained, ZODIAC cannot recover it. To answer the second question, we deliberately include all ground truth compounds for evaluation. In the best-case scenario ZODIAC assigns a low score to compounds that do not have the correct molecular formula in the SIRIUS 4 result list, thereby expressing low confidence in these inevitably wrong annotations. Additionally, we evaluate ZODIAC against the Seven Golden Rules molecular formula filters and GenForm[1], a non-commercial software for MS/MS-based molecular formula annotation.

5.2.1 Datasets

The "dendroides" dataset is an extract of *Euphorbia dendroides* plants; "NIST1950" is human plasma reference material (SRM 1950); "tomato" are tomato seedling samples of *Solanum lycopersicum*; "diatoms" are extracts of the intra- and exo-metabolomes of a single diatom genus; and the "mice stool" dataset is from a microbiome study including germ free and colonized mice, measured by Quinn *et al.* [164]. Dendroides was measured on an LTQ-XL Orbitrap; NIST1950, tomato and diatoms on a Q Exactive Orbitrap; and mice stool on a maXis QTOF mass spectrometer. All data was acquired in positive ionization mode.

All these datasets are biological samples and no reference datasets. Hence, we have to establish a ground truth which we can evaluate against. For one dataset (dendroides) 201 compounds have been manually annotated by Louis-Félix Nothias with molecular formulas [125]. For four other datasets the ground truth was established by spectral library search (Section 5.2.3). Note, that library search does not guarantee a 100 % valid ground truth. However, we evaluate molecular formulas and not structures. The number of incorrect ground truth molecular formula annotations should be very low, since incorrect matches to structural isomers still result in the same molecular formula.

Four datasets were chosen because we expect a reasonable number of hits using spectral library search, as required for the evaluation: In detail, we obtained manual annotations for dendroides; tomato is a model organism; NIST1950 is reference material; whereas mice stool [164] was previously analyzed manually. To this end, the discovery of novel molecular formulas in these datasets is less likely. In contrast, diatoms have been studied less comprehensively and compounds often contain uncommon elements; hence, we expect to find novel compounds in this dataset. For more details on the measured data see Ludwig *et al.* [125]. The number of compounds per dataset after processing the data can be found in Table 5.1. We could establish a ground truth of overall 703 compounds. Annotation of high-mass compounds is particularly challenging. For dendroides, 75 % of the ground truth compounds have an m/z of 605 or higher (Fig. 5.2). The mice stool dataset poses a challenge because of the selected measurement setup. MS/MS spectra where measured using an isolation window of at least 4 m/z. This greatly increases the chance of chimeric spectra; fragments of multiple compounds in the same spectrum may interfere with the similarity estimation between MS/MS spectra as performed by ZODIAC .

5.2.2 Preprocessing

Input mzML/mzXML files are processed with OpenMS [175] and low-quality MS/MS spectra were discarded. The workflow described below is visualized in Fig. 5.3.

---

[1] `https://sourceforge.net/projects/genform/`

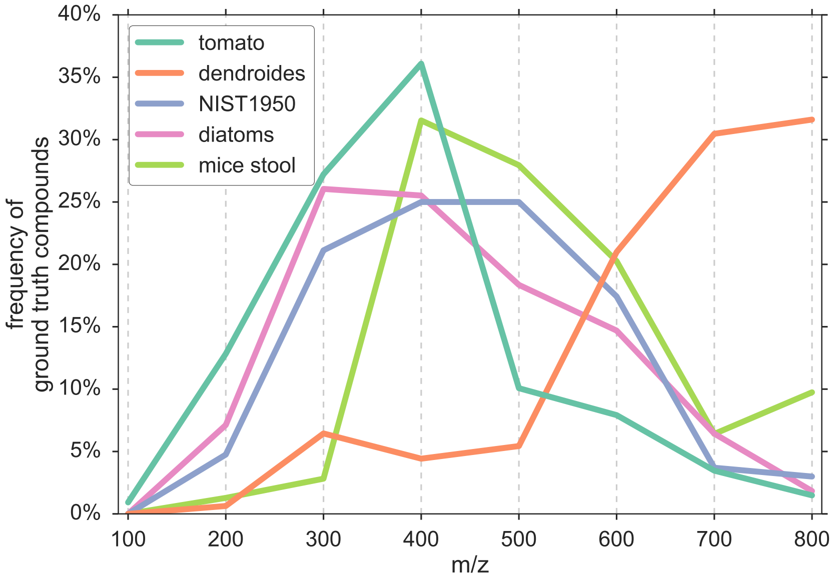


Figure 5.2: Distribution of compound masses. Distribution of precursor ion m/z of the compounds used as ground truth for the evaluation of the molecular formula annotation on the five datasets. Bins of width 100 are centered at 100, 200, . . . , 800 m/z.

OpenMS. We used OpenMS 2.4.0 [175] to process the mzML/mzXML files. We performed minor modifications on the OpenMS source code by removing 12 lines and adding 78 lines. This allowed to detect more isotope peaks, to match MS/MS to MS1 features based on the actual isolation window and to add functionality to the SIRIUSAdapter to directly output SIRIUS file format including retention time information. These numbers are based on the patch file, see Code availability, and include lines with comments and blank lines; the modification of a line corresponds to the removal and insertion of a new line.

Feature finding and clustering of isotopic mass traces was performed using the FeatureFinderMetabo module. Next, adducts were detected with the MetaboliteAdductDecharger module. Finally, spectra were exported to the SIRIUS specific format using the SIRIUSAdapter module. OpenMS parameter files are provided as part of the virtual

Table 5.1: Statistics on compounds with annotated ground truth molecular formulas. Given is the number of total compounds, the number of compounds with a ground truth molecular formula and the number which are in the top 50 of SIRIUS ranked candidates. The median m/z and 25 (Q_1) and 75 percentile (Q_3) considers only candidates in the top 50.

dataset	# compounds	ground truth	# in top 50	Q_1 m/z	median m/z	Q_3 m/z
dendroides	792	201	197	605.310	705.274	759.353
NIST1950	571	94	94	286.390	373.800	477.335
tomato	2902	271	270	207.814	271.713	334.526
diatoms	2472	93	93	253.195	301.216	349.237
mice stool	398	44	43	373.274	454.292	516.298

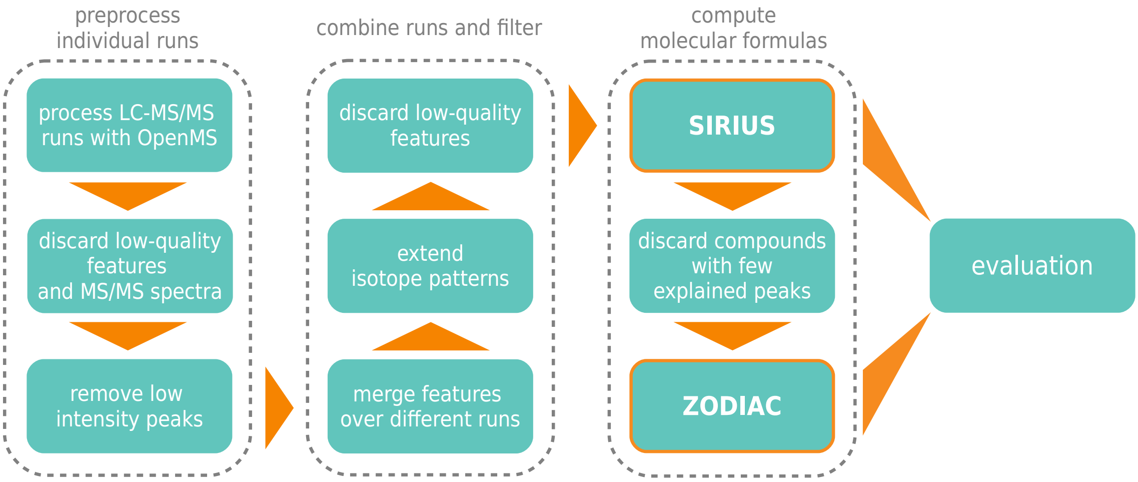

preprocess individual runs
process LC-MS/MS runs with OpenMS
discard low-quality features and MS/MS spectra
remove low intensity peaks
combine runs and filter
merge features over different runs
extend isotope patterns
discard low-quality features
compute molecular formulas
SIRIUS
discard compounds with few explained peaks
ZODIAC
evaluation

Figure 5.3: ZODIAC processing and evaluation workflow. 1) Each LC-MS/MS run is processed individually; input mzML/mzXML files are processed using OpenMS, performing feature and adduct detection and producing files in SIRIUS input format. Resulting features combine MS1, MS/MS and adduct information. 2,3) Filtering is performed on feature, MS/MS and peak level. 4) Similar features are merged between different runs using hierarchical clustering; MS/MS are combined and a best isotope pattern is selected per feature. 5) Missing isotope peaks are searched in MS1 spectra to extend isotope patterns. 6) A final feature filtering step is performed; the remaining features are considered as compounds. 7) SIRIUS is executed. 8) Compounds with few explained peaks are discarded, since a badly explained MS/MS spectrum indicates low quality. 9) ZODIAC is run on the remaining compounds. 10) SIRIUS and ZODIAC are evaluated on the same set of compounds.

machine. Parameters were chosen by manual inspection; in particular, we used a small noise intensity threshold to increase chances that isotope peaks of a compound are picked.

Discarding features and MS/MS spectra. We excluded m/z features which eluted over a very long time during chromatography and did not produce desired mass traces in a limited time window, as such traces are considered chemical noise. To do so, we binned MS1 peaks with a bin size of 0.006 m/z. Each MS1 was normalized by the most intense peak. Each peak was counted if its relative intensity was 0.01 or higher. If a m/z bin contained peaks of more than 20 % of MS1 the m/z was considered chemical noise. Because these spurious chemical noise features have rather high mass deviation we removed all MS1 features within 30 ppm.

Next, we performed blank removal using blank samples from the corresponding datasets. Features within 15 ppm and 20 s of a blank feature were removed if intensities were lower than 2-fold of the blank feature intensity. We did not perform blank removal on the mice stool dataset, because this resulted in a low number of remaining compounds.

Third, we removed features from the beginning and end of the chromatography run and features with low relative or absolute intensity; and we removed MS/MS spectra which could not be assigned to an MS1 feature, MS/MS of a precursor peak with low absolute or relative intensity, and chimeric MS/MS. See Table A.1 for dataset-specific parameter values. Chimeric spectra contain fragments of multiple precursor ions; we detected chimeric spectra as follows: All peaks within the isolation window, excluding isotope peaks, were considered to contribute their intensity to the measured MS/MS. We estimated the relative intensity that the target precursor ion contributes to the MS/MS; if the target precursor ion contributed to less than 50 % of the MS/MS intensity or if a second precursor ion contributed more than 33 % of the target precursor ion intensity, the MS/MS was marked as chimeric and excluded. The Isolation window width for the Orbitrap mass spectrometer used for the dendroides, NIST1950, tomato and diatoms is 1 Da; for the mice stool dataset analyzed on a QTOF mass spectrometer an isolation window of 3 Da width and shifted by 1 Da to the right, centered at the +1 isotope peak, was assumed.

Filtering MS/MS spectra. In each MS/MS spectrum, we filtered peaks using an intensity threshold of two times the median noise intensity, see Table A.1. The median noise intensity of a dataset was estimated from peaks which had no molecular formula decomposition within a 40 ppm window considering elements C, H, N, O, and P plus those elements predicted from the isotope pattern, see below. Isotope peaks were removed from MS/MS spectra of the mice stool dataset.

The SIRIUSAdapter OpenMS module combines MS/MS which are associated with the same MS1 feature. In addition, complete linkage hierarchical clustering was conducted to merge features over different LC-MS/MS runs. Features were merged using 15 ppm mass accuracy and a 15 sec retention time window. Features with different adduct annotations or features from the same run were not merged. Feature similarity was computed by the cosine product of the MS/MS (see below), and the similarity threshold for clustering was set to 0.8. When multiple features were merged into a single one, where each feature has an assigned isotope pattern, then the isotope pattern with the highest number of isotope peaks was kept. In case multiple isotope patterns had the same number of isotope peaks, the one with the most intense monoisotopic peak was kept. After merging, features were discarded if the summed MS/MS intensity was below a threshold, see Table A.1. Features with precursor mass above 850 Da are discarded: Whereas ZODIAC is clearly capable of processing such features, we found that there are no spectral library hits above this mass that can be used for evaluation, see below. Only 2.72 % of features across all datasets have m/z above 850 Da, so we obtained library hits over a large portion of the mass range.

Extending isotope patterns. OpenMS often misses low-intensity isotope peaks. To recover those peaks, we post-processed OpenMS results as follows: For each isotope pattern detected by OpenMS, we try to extend it using isotope peaks from the corresponding MS1 spectra chosen by OpenMS. Isotope pattern peaks were picked using the SIRIUS 4 isotope pattern picking subroutine. If an additional isotope peak is present in at least 66 % of the corresponding MS1, the peak was added to the isotope pattern. Subsequently, features with less than two isotope peaks are discarded.

Discarding low-quality merged MS/MS spectra. Even when considering all MS/MS spectra for some features, we sometimes have insufficient information for both spectral library search and molecular formula annotation; to this end, such "low-quality features" were discarded. A feature is discarded if it produces less than 5 fragment peaks, estimated after merging peaks within 10 ppm or 0.0025 m/z from all corresponding MS/MS spectra; and if no fragmentation tree in the top 50 candidate list can explain at least 5 peaks accounting for at least 80 % of total spectrum intensity, see SIRIUS analysis below. Filtering "low quality" features decreased the number of features for dendroides from 1,078 to 784, for NIST1950 from 568 to 400, for tomato from 3,583 to 2,584, for diatoms from 3,227 to 2,075 and for mice stool from 577 to 377.

For brevity, we will refer to the features detected by OpenMS as *compounds*, see above.

5.2.3 SIRIUS Analysis and Establishing a Ground Truth

SIRIUS 4 was run with the default alphabet of elements CHNO, at most 5 phosphorus atoms, and one iodine atom; automatic element detection from the isotope pattern [133] was enabled for sulfur, chlorine, bromine, boron, and selenium. For the dendroides, NIST1950, and tomato datasets we used 15 ppm maximum mass deviation for SIRIUS; for diatoms and mice stool datasets we used 10 ppm. The SIRIUS default Ring Double Bond Equivalent (RDBE) value to filter molecular formula candidates was lowered from -0.5 to -1.0, to account for undetected ammonium adducts. Isotope patterns were not used to filter molecular formula candidates before computing fragmentation trees.

- If OpenMS provided an ionization adduct type (such as protonation, sodium adduct, potassium adduct) for a compound, only this ionization was used. We export the 50 best-scoring molecular formula candidates from SIRIUS.

- In cases where no ionization adduct type was provided by OpenMS, we selected one or more adducts from $[\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+\mathrm{Na}]^+$, and $[\mathrm{M}+\mathrm{K}]^+$ by searching for characteristic mass differences, using the MS1 that contained the most intense peak of the precursor ion. Peaks below 5 % relative intensity were discarded for this decision. For each compound, we export the 50 best-scoring molecular formula candidates; we simultaneously ensure that for each considered ionization adduct type, at least 10 candidates are considered.

We will refer to this candidate list as the *top 50*.

To evaluate the performance of SIRIUS and ZODIAC, we had to annotate a subset of compounds with “correct” molecular formulas, to serve as our ground truth. For this, we combined manual annotation and spectral library search, as follows: For the dendroides dataset, we use 201 compounds that have been manually annotated with molecular formulas by Louis-Félix Nothias [125]. For the remaining datasets, we performed spectral library searches against multiple libraries, but did not add manual annotations. We searched compounds in a spectral library combining GNPS [211], MassBank [90], NIST17 database (National Institute of Standards and Technology, v17) and “MassHunter Forensics/Toxicology PCDL library” (Agilent Technologies, Inc.) [63]. We compute a similarity score assuming peaks as Gaussians, with the centroided peaks’ m/z as the mean and the standard deviation being the maximum of a relative mass error of 20 ppm and an absolute mass error of 0.005 m/z. Precursor ion masses are permitted to differ by 10 ppm or 0.0025 m/z at maximum. Only library hits with a similarity score of 0.7 or higher and with at least 6 shared peaks are considered as being valid. We compute the score as the mean of the cosine score of the sample spectrum and the cosine score of the mirrored spectrum; to mirror a spectrum with precursor mass M, we replace peak m/z value m by $M - m$. This resulted in 94 annotated compounds for NIST1950, 271 for tomato, 93 for diatoms and 44 for mice stool.

We evaluate SIRIUS and ZODIAC against these “ground truth” molecular formulas, but we stress that beside the molecules that were isolated in *Euphorbia dendroides* samples and correspond to level 1 of the Metabolomics Standard Initiative ranking system, not all of these are necessarily correct. In particular, we refrain from ranking these according to the Metabolomics Standard Initiative ranking system, where level 4 corresponds to an “unequivocal molecular formula”. An evaluation is nevertheless meaningful because we expect only few errors on the molecular formula assignment level.

In few cases, the correct molecular formula was not ranked in the top 50 SIRIUS candidates; we also dropped these from our evaluation, as it is not possible that ZODIAC can find the correct molecular formula in our evaluation. We discarded four compounds for dendroides, zero for NIST1950, one for tomato, zero for diatoms and one compound for mice stool because of this criterion.

See Table 5.1 for details, and see Fig. 5.2 for the mass distribution of the “ground truth” compounds. Compounds in the dendroides dataset with reference annotations have high mass, and 75 % of all reference annotations have an m/z of 605 or higher. The NIST1950 dataset resulted in library hits over a broad range of m/z values. The diatoms library hits have a median m/z of 301 but the sample itself is highly complex, as described above. Only few compounds remain in the mice stool dataset after filtering chimeric and low quality compounds, see above.

5.2.4 ZODIAC Parameters

We use identical parameters for all five datasets, see equation (5.5): We weight fragments and root losses when comparing fragmentation trees of molecular formula candidates. Here, we use the SIRIUS 4 noise intensity scoring as importance ι in (5.7). The probability that a peak p that corresponds to a fragment and root loss is not noise is $\iota = 1 - par(int(p))$, where par is the Pareto cumulative distribution function with $x_{min} = 0.002$, $x_{median} = 0.015$ and $int(p) \in [0, 1]$ the relative peak intensity, see reference [19]. To establish a threshold on the minimal similarity of fragmentation trees, we decrease score s and tree sizes n_1 and n_2 each by 1.0, see (5.6).

The empirical score distributions resemble a log-normal distribution, see Fig. A.3 in the appendix, so we use its Cumulative Distribution Function to estimate p-values for (5.8). For the robust estimation of parameters μ and σ^2, we sampled 100,000 non-zero scores for each dataset, and used the median score as parameter μ and the median absolute deviation as parameter σ^2. We naturally expect most edges to be false edges and chose score threshold τ so that 95 % of the non-zero scores are smaller than this threshold. Finally, we use individual thresholds for each compound (color) so that at least 10 molecular formulas of this color are incident to 10 or more edges.

Each molecular formula candidate of some compound receives a score $s_1, ..., s_n$, where s_{max} is the largest score. We transformed SIRIUS scores to probabilities using the softmax function, where $p_j = \exp(s_j - s_{max})$ are normalized to sum to one. To adjust for the fact that the correct molecular formula may not be in the top 50, we added a dummy node receiving the combined probability of all unconsidered candidates. Dummy nodes are not connected to any other node. SIRIUS does not report the score of all candidates, as one compound may have tens of thousands of candidates. Hence, we estimated the probability of all unconsidered candidates by multiplying the number of unconsidered candidates with the lowest probability of the top 50 candidates.

Finding and scoring ZODIAC anchors. ZODIAC can use (potentially incorrect) spectral library hits as anchors to improve annotations. To find a reasonable number of anchors, we perform spectral library search in analogue mode — that is, we do not only search for exact matches but also match compounds to library spectra with differing precursor mass. Resulting molecular formula annotations are not considered ground truth identifications but are sufficient as anchors. Only those hits were considered that have mass differences between query and reference corresponding to a frequent biotransformation. We use the following molecular formula mass differences as valid biotransformations [118, 124, 174]: C_2H_2, C_2H_2O, C_2H_3NO, $C_2H_3O_2$, C_2H_4, C_2O_2, $C_3H_2O_3$, C_3H_5NO, $C_3H_5NO_2$, C_3H_5O, $C_4H_2N_2O$, $C_4H_3N_3$, $C_4H_4O_2$, C_5H_7, C_5H_7NO, C_5H_9NO, CH_2, CH_2ON, CH_3N_2O, CHO_2, CO, CO_2, H_2, H_2O, N, NH, NH_2, NH_3, O, H_4O_2 and H_6O_3.

We use identical parameters for all five datasets: When scoring anchors according to (5.5), we use the *maximum* of the cosine score between the spectrum and the cosine score of the mirrored spectrum as the similarity measure, and $min_l = 0.5$ as the score threshold parameter and $\lambda = 1,000$ as the weighting parameter. For anchors found by spectral library match in analogue mode (that is, non-identical m/z), spectral similarity is reduced by 0.1 to account for increased uncertainty.

Searching for anchors as described above resulted in 96 anchors for dendroides, 254 anchors for NIST1950, 749 for tomato, 372 for diatoms and 176 for mice stool. All spectral hits described in the previous section are anchors, too; recall that for dendroides, the ground truth was established manually and those annotations do not serve as anchors.

Burn-in and number of Gibbs sampling epochs. We determined a reasonable number of Gibbs sampling iterations using the dendroides dataset. One iteration, also called *epoch*, is defined as one round in which each compound is updated once by choosing a new "active" molecular formula candidate. We run 10 independent Markov chains, see Fig. 5.10: The total score summed over all active candidate at a specific epoch increases swiftly over the first 500 epochs. Similarly, the number of correct annotations at a specific epoch increases quickly for most Markov chains until the chain seems to stay in a local optimum. We note that this number of correct molecular formula is determined at each epoch whereas ZODIAC scores are computed from the average over many epochs. From this data, we estimated a burn-in of 1,000 epochs and sampling of 2,000 iterations. Larger values increase running times but should never worsen results.

In application, we use 10 Markov chains in parallel, a burn-in of 1,000 epochs, and sample 2,000 epochs; we keep only every 10^{th} sample, resulting in a total of $10 \times 200 = 2,000$ samples.

5.2.5 Parameters of Competing Methods

Seven Golden Rules. We analyze if molecular formula annotations adhere to the Seven Golden Rules by Kind and Fiehn [105]. To do so, we apply the valency filter rules (LEWIS and SENIOR check), the common range of element ratios (hydrogen/carbon element ratio check and heteroatom ratio check) and the element probability check. We use the Seven Golden Rules as a filter based on the molecular formula, ignoring the measured isotope pattern.

Exact mass search and GenForm. We evaluate ZODIAC against annotation by exact mass and GenForm (Fig. A.1 in the appendix). GenForm[2] is an open-source

---

[2] `https://sourceforge.net/projects/genform/`

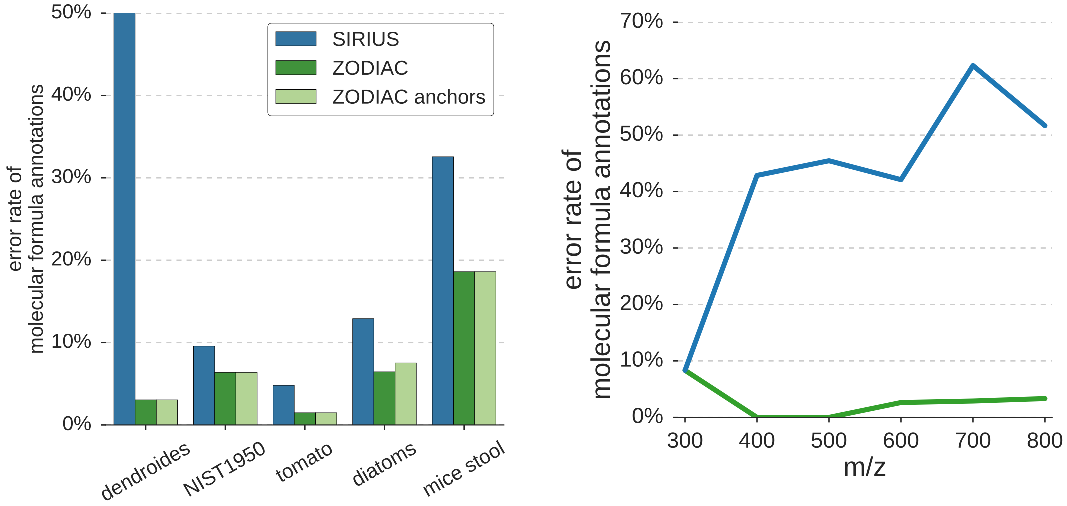


Figure 5.4: Molecular formula annotation error rates. Left: Error rate on five datasets. The rate of incorrect molecular formula annotations is displayed for SIRIUS and ZODIAC, with and without anchors. For number of compounds and other statistics, see Table 5.1. ZODIAC reduces error rates on all datasets. See Fig. A.1 in the appendix for other methods. Right: Error rates vs. mass on dendroides dataset. Error rates for SIRIUS and ZODIAC without anchors are binned by compound m/z; bins of width 100 are centered at 300, 400, . . . , 800 m/z.

implementation of MOLGEN-MS/MS [129]. For all methods we assume the same adduct candidates as described for SIRIUS and ZODIAC. We require molecular formulas from exact mass annotation to have an RDBE value of -1 or greater. Additionally, exact mass search was combined with filtering based on the Seven Golden Rules. For GenForm, we enable the RDBE filter ("exist" option) and set the "rej" and the "ppm" parameters to the same mass errors as assumed for SIRIUS computation. GenForm only supports the adducts $[\mathrm{M}+\mathrm{H}]^+$ and $[\mathrm{M}+\mathrm{Na}]^+$, but not $[\mathrm{M}+\mathrm{K}]^+$. This results in a slight evaluation advantage over all other methods that consider $[\mathrm{M}+\mathrm{K}]^+$, because no ground truth compound has adduct $[\mathrm{M}+\mathrm{K}]^+$.

5.2.6 Results

For all five datasets, we observe that ZODIAC outperforms SIRIUS (Fig. 5.4, left) and all other publicly available methods (Fig. A.1 in the appendix), often substantially decreasing molecular formula annotation error rates. Improvements are most distinctive for the dendroides dataset containing many larger compounds: 75 % of the ground truth compounds have an m/z of 605 or higher (Fig. 5.2). Hence, this dataset is particularly challenging for molecular formula assignment. Out of the 201 ground truth compounds, the preprocessing assigned an incorrect adduct to three; for these, the correct molecular formula is not contained in the candidate list considered by ZODIAC. For one compound, the corrected molecular formula was not ranked into the top 50. For the remaining 197 compounds, SIRIUS incorrectly annotated 50.25 % (99), compared to 3.05 % (6) for ZODIAC without anchors. This represents an 16.50-fold decrease in error rate. Error rates improve for compounds over the whole mass range, see Fig. 5.4 (right).

On the NIST1950 and tomato datasets, SIRIUS already showed excellent performance, with less than 10 % incorrectly annotated molecular formulas. ZODIAC further decreased error rates, from 9.57 % to 6.38 % for NIST1950 and 4.81 % to 1.48 % for tomato. The diatoms dataset is rather complex, and compounds may contain halogens. Here, SIRIUS reaches an error rate of 12.90 %, which is reduced two-fold to 6.45 % by ZODIAC.

Mice stool MS/MS spectra were measured with a broader isolation window, and the dataset contains numerous chimeric and low quality spectra, most of which were discarded before running ZODIAC. Consequently, ZODIAC has a much smaller network of interdependent compounds than for the other datasets. But even spectra that were not discarded often have substantially worse quality than spectra from other datasets: For example, these spectra often contain isotope peaks of fragments or are undetected chimeric spectra. Our evaluation shows that even for this extremely challenging dataset, ZODIAC improves annotation results, decreasing the error rate from 32.56 % for SIRIUS to 18.60 % for ZODIAC.

Some compounds in an LC-MS/MS run can result in high-scoring hits when searching in a MS/MS spectral library. ZODIAC's stochastic model allows us to integrate these hits as *anchors*, assuming that we can trust assigned molecular formulas to a high degree. We performed a 10-fold cross-validation to assess the improvement using anchors. We ensured structure-disjoint evaluation on the library hits, as multiple "compounds" in the dataset may correspond to the same structure; see Dührkop *et al.* [61] on the importance of structure-disjoint evaluation. We find that ZODIAC with anchors does not improve the error rate.

For four datasets (NIST1950, tomato, diatoms, mice stool), ground truth molecular formulas were established by library searching only. We tested if there is a distinct difference between the cosine score of ZODIAC's correct and incorrect molecular formula assignments, but did not find such a difference (Fig. A.2 in the appendix).

We find that differentiating between adducts $[M + H]^+$ and $[M + Na]^+$ is sometimes challenging for ZODIAC. This is observable for the dendroides dataset, where all six incorrect ZODIAC annotations show an incorrect adduct annotation, mistaking $[M + H]^+$ for $[M + Na]^+$ or vice versa. In all six cases, the molecular formula of the best ZODIAC hit and the ground truth differ by exactly two carbon minus two hydrogen atoms (21.984349 Da), with mass difference highly similar to that between $[M + H]^+$ and $[M + Na]^+$ (21.981944 Da). Sodium-ionized compounds can produce protonated fragments, making the interpretation of these spectra challenging. We reran ZODIAC on the dendroides dataset, assuming we knew the correct adduct for each reference compound. For all 201 compounds, the correct hit is contained in the SIRIUS top 50 candidate list. SIRIUS correctly annotated 66.17 % (133) and ZODIAC 99.50 % (200) of the compounds, corresponding to a 68-fold decrease of the error rate. The other four datasets contain fewer sodium adducts.

ZODIAC implicitly tries to estimate the probability that an annotated molecular formula is correct; we find these estimates to be imprecise, see Fig. 5.5. But the ZODIAC score can be used to differentiate between true and incorrect annotations: For each dataset, we sort molecular formula annotations by the ZODIAC score, and calculate the rate of correct annotations for any subset of top-scoring annotations. We find that high-scoring ZODIAC annotations are more likely to be correct, see again Fig. 5.5. For this evaluation, we also considered previously discarded compounds for which SIRIUS did not rank the correct molecular formula in the top 50; for these compounds, ZODIAC cannot find the correct

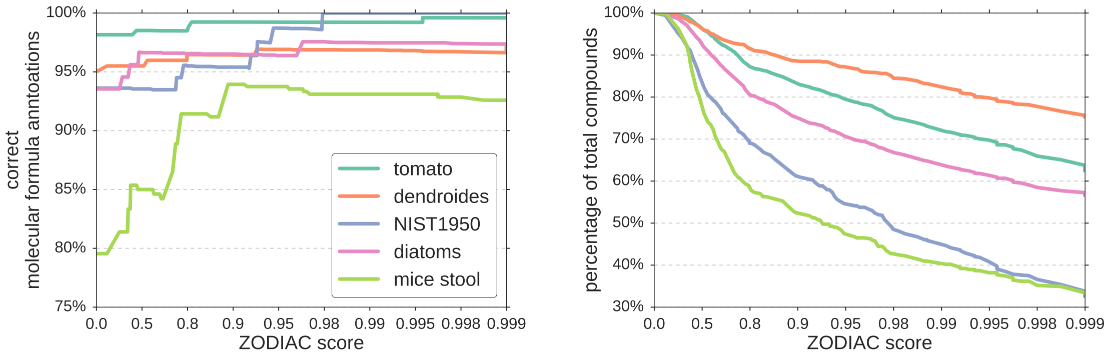


Figure 5.5: Percentage of correct annotations and number of compounds in relation to ZODIAC score. Left: Percentage of correct molecular formula annotations for different ZODIAC score thresholds for five datasets. We sort compounds by ZODIAC score and calculate the rate of correct annotations for all compounds above the given thresholds. Right: Percentage of total compounds with a ZODIAC score above different thresholds on five datasets. Here, we consider *all* compounds, with and without established ground truth. Note that scores on the x-axis are not equidistant.

molecular formula but at best, the incorrect molecular formula should receive low ZODIAC scores. Selecting a ZODIAC score threshold of 0.9 results in more than 93.94 % correct annotations while keeping 52.23 % to 88.51 % of the compounds of each dataset (Fig. 5.5). In comparison, spectral library search allowed us to annotate between 3.78 % and 16.55 % of a dataset, see Table 5.1.

Filters are commonly applied to exclude “exotic” molecular formulas; this improves annotation performance unless the true molecular formula is “exotic”. Most notably, Kind and Fiehn [105] introduced the Seven Golden Rules in 2007, which are frequently used in the metabolomics community. Rules were empirically established from molecular structure databases, but are nowadays sometimes used as if they represented a biological ground truth. We checked ZODIAC molecular formula annotations against the Seven Golden Rules, see Fig. 5.6: We find that all but one molecular formula in the dendroides dataset adhere to the Seven Golden Rules; this is not surprising as molecular formulas are mainly comprised of CHNO for this dataset. In contrast, diatoms and tomato samples contain many compounds with “uncommon elements” (phosphorus, sulfur, halogens). Here, numerous molecular formulas do not adhere to the Seven Golden Rules; this includes molecular formulas with high ZODIAC score. We argue that using heuristic filters to exclude “exotic” molecular formulas, will in practice often exclude valid molecular formulas and may ultimately result in fewer correct annotations.

Novel molecular formulas.

We now concentrate on *novel* molecular formulas discovered by ZODIAC in the five datasets; recall that such discoveries are possible because ZODIAC does not rely on any molecular structure databases such as PubChem [103] and ChemSpider [157], but rather considers all chemically feasible formulas. For a molecular formula to be novel, we require it not to be present in PubChem and in case it corresponds to a protonated ion, also that the molecular formula obtained by adding NH_3 is not contained either. The second constraint ensures that we do not treat an undetected ammonium adduct as a novel molecular formula.

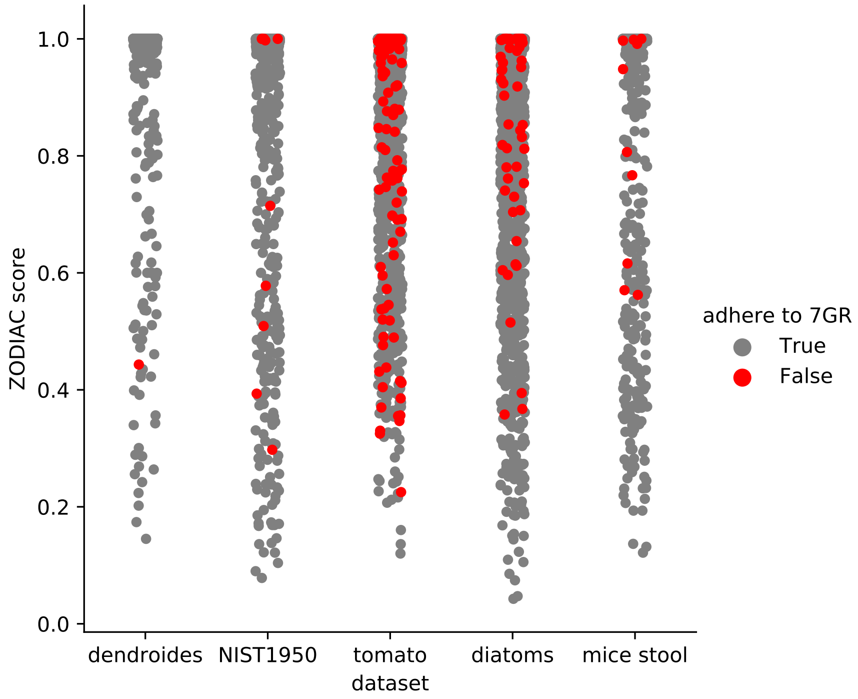


Figure 5.6: Seven Golden Rules applied to annotated molecular formulas. For each ZODIAC molecular formula annotation, we test whether it meets the Seven Golden Rules (7GR). Each dot represents one annotated compound; molecular formulas are sorted by ZODIAC score.

To cut down the number of reported formulas, we use strict filters for the ZODIAC score, the quality of the underlying MS/MS data, and the support by other molecular formulas in the dataset. We report molecular formula annotations from all five datasets with a) minimum ZODIAC score of 0.98, b) at least 95 % of the MS/MS spectrum intensity being explained by SIRIUS, and c) at least one molecular formula of the compound is connected to 5 or more compounds in the ZODIAC similarity network. The third criterion discards compounds where ZODIAC's results are basically identical to SIRIUS's. This results in 15 novel molecular formulas in tomato, 15 in diatoms, one in NIST1950 and one in the mice stool dataset, see Table table A.2 in the appendix. Filtering less restrictively (ZODIAC score at least 0.95, at least 90 % of the MS/MS spectrum intensity being explained by SIRIUS), we annotate 32 novel molecular formulas in tomato, 26 in the diatoms dataset, three in NIST1950 and one in mice stool. Recall that we did not expect many novel molecular formulas in any dataset but diatoms.

ZODIAC allows the user to select a few, potentially highly interesting compounds (molecular formulas) from a set of hundreds or thousands with low effort. Next, we show that some top-scoring annotations from Table A.2 in the appendix are presumably correct.

Manual evaluation of novel molecular formulas.

We now concentrate on one particular compound in the diatoms dataset (m/z 588.230, retention time 503.97 sec): The compound is protonated and was annotated with molecular formula $C_{24}H_{47}BrNO_8P$, which is indeed absent from the structure databases [103, 157].

We chose this compound because of its perfect ZODIAC score of 1.0 and many library search hits in analog mode. Furthermore, the occurrence of bromine agrees with our expectation that marine organisms can be prolific sources of organohalogens [66]. In retrospect, this molecular formula may appear "straightforward" as the underlying structure is presumably a halogenated phosphatidylcholine; but this fact was unknown to us when choosing the compound, and is also *not known to ZODIAC*.

We found multiple lines of evidence that this molecular formula annotation is correct, both in the measured isotope pattern and the three MS/MS spectra measured for *m/z* 588.230 (monoisotopic peak of the isotope pattern), 590.228 (M+2 peak) and 592.325 (M+4 peak), see Fig. 5.7. To annotate fragments with molecular formulas, we used SIRIUS to compute a fragmentation tree for the MS/MS spectrum of the monoisotopic peak at *m/z* 588.230 (Fig. 5.7e).

1. We compared MS/MS spectra for *m/z* 588.230, 590.228 and 592.325 (Fig. 5.7a) and found them to be highly similar, confirming that these peaks are indeed isotope peaks of one compound. One peak "moves" between MS/MS spectra (nominal *m/z* 570, 572 and 574): This peak corresponds to the fragment with molecular formula $C_{24}H_{45}BrNO_7P$, which is the only annotated fragment containing bromine.

2. The measured isotope pattern agrees well with the theoretical isotope pattern of $[C_{24}H_{47}BrNO_8P + H]^+$ (Fig. 5.7c). The M+2 peak of the measured isotope pattern has a relative intensity of 106.0 % of the monoisotopic peak, which is characteristic for the presence of a bromine atom.

3. The MS/MS spectrum for *m/z* 588.230 contains a precursor loss of 79.925 Da, and the only possible molecular formula explanation of this loss is BrH, considering a mass error of 100 ppm and elements CHNOPSFIClBrNaKSi.

4. We can simulate an MS/MS spectrum of the M+2 peak that *includes isotope patterns of fragments*: We use peak intensities from the MS/MS spectrum of the monoisotopic peak, and simulate isotope patterns of fragments as described by Rockwood *et al.* [171]. This allows us to verify whether the isotope patterns of fragments agree with our theoretical expectations. Indeed, simulated and measured MS/MS spectra of the M+2 peak show very high similarity, see Fig. 5.7f. The MS/MS spectrum of the M+4 peak must be treated with caution, as the precursor's intensity is much lower and a second compound of higher intensity is present within the isolation window, see Fig. A.4 in the appendix. With regards to the moving peak, we can observe matching peaks in the simulated spectra, too.

5. We do not find peaks in the MS1 spectrum at *m/z* plus 18.01 (indicating water loss), plus 43.99 (carbon dioxide loss) or minus 17.03 (ammonium adduct); similarly, we do not find molecular formulas in PubChem or ChemSpider that correspond to the novel molecular formula plus H_2O, plus CO_2 or minus NH_3. To this end, the reported molecular formula indeed corresponds to the protonated molecule, not an adduct or fragment.

6. The spectrum matches to multiple NIST17 library spectra with different *m/z*, all of which are phosphatidylcholines. The top 18 matches have a cosine score above 0.9 and share a set of characteristic peaks which match to the query spectrum: See Fig. 5.7b

> for the best hit, and Fig. A.5 and Fig. A.6 in the appendix for additional hits. For the set of shared peaks, the molecular formula annotations of the NIST reference spectrum (as provided by NIST) are identical to those of the SIRIUS fragmentation tree computed for the query compound (Fig. 5.7e).

Considering the matching NIST reference spectra, we propose that the query compound is a brominated phosphatidylcholine. Marine algae are known producers of halogenated compounds [36, 195]. Moreover, diatoms possess the biosynthetic pathways to produce halogenated lipids [218]. Based on the fragmentation tree analysis and supported by biosynthetical considerations, we propose that the bromine atom is located on the fatty acid tail [53, 172]. The putative structure and their mass fragments are shown in Fig. 5.7d.

Finally, note that there is another novel molecular formula in the diatoms dataset annotated with high confidence, namely $C_{24}H_{49}BrNO_8P$, which differs from the molecular formula $C_{24}H_{47}BrNO_8P$ by one degree of unsaturation. The corresponding compounds have *m/z* 590.246 and retention times 523.39 sec and 539.29 sec.

A second example of a novel molecular formula found in the diatoms dataset is the ZODIAC-annotated sodium adduct with molecular formula $C_{15}H_{30}ClIO_5$ (*m/z* 475.072, retention time 579.16 sec). The ZODIAC score of this annotation is 0.999 and both the MS1 isotope pattern (Fig. 5.8b) as well as the SIRIUS computed fragmentation tree (Fig. 5.8a) instills high confidence in the ZODIAC molecular formula annotation. The ion was correctly annotated as $[M + Na]^+$, as can be seen by the correlating chromatographic peak of the protonated ion in the same dataset (Fig. 5.8c). The isotope pattern is consistent with the presence of chlorine, and several neutral losses contain chlorine and iodine: HI (127.912 Da), ClI (161.873 Da) and HCl (35.977 Da) losses are detected with less than 2 ppm mass error (Fig. 5.8d). This provides strong evidence for the presence of chlorine and iodine in the molecular formula. As a consequence, all SIRIUS top 5 molecular formulas include chlorine and iodine. However, the remaining candidates can be excluded based on unlikely fragment and loss annotations; for further details see Ludwig *et al.* [125]. This leaves $C_{15}H_{30}ClIO_5$ as only plausible explanation.

Such manual evaluation is clearly possible for all novel molecular formulas from Table A.2 in the appendix; but this is beyond the scope of this thesis.

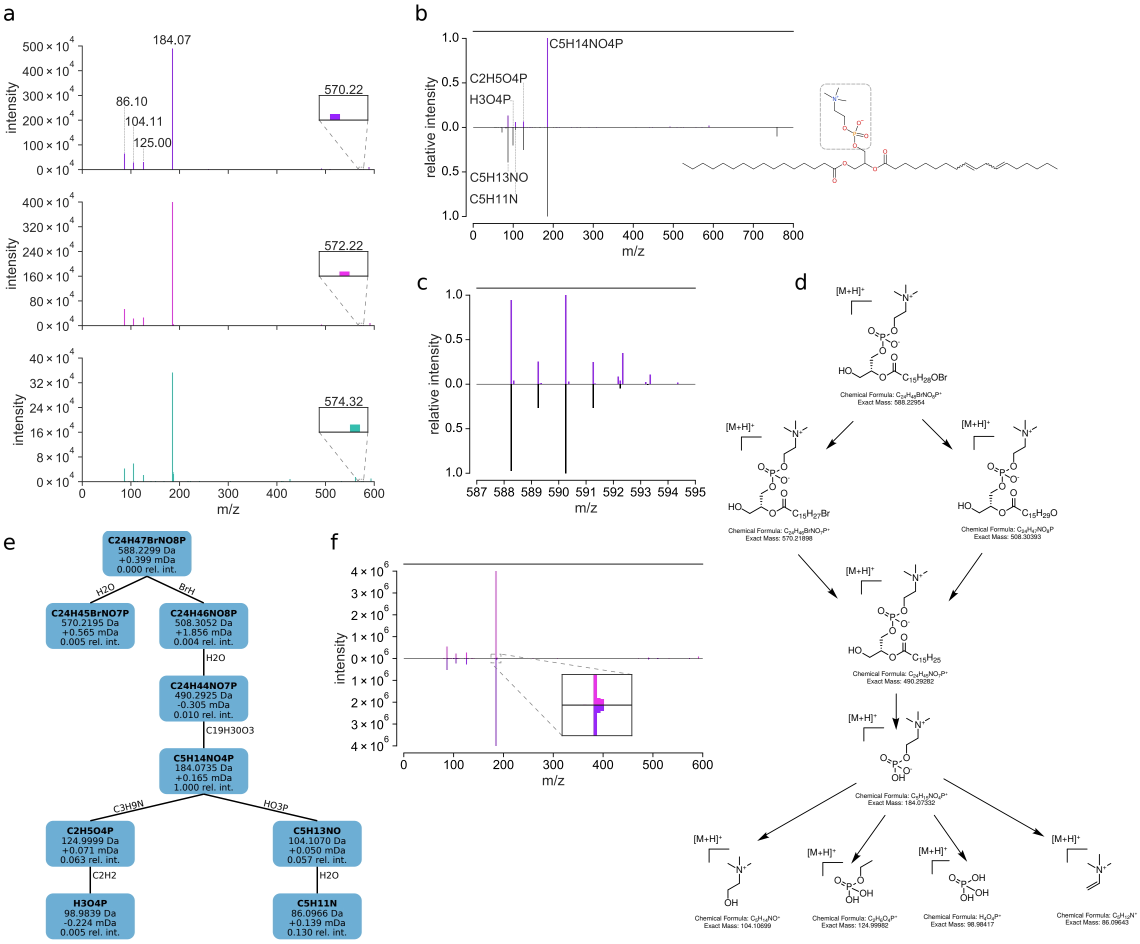
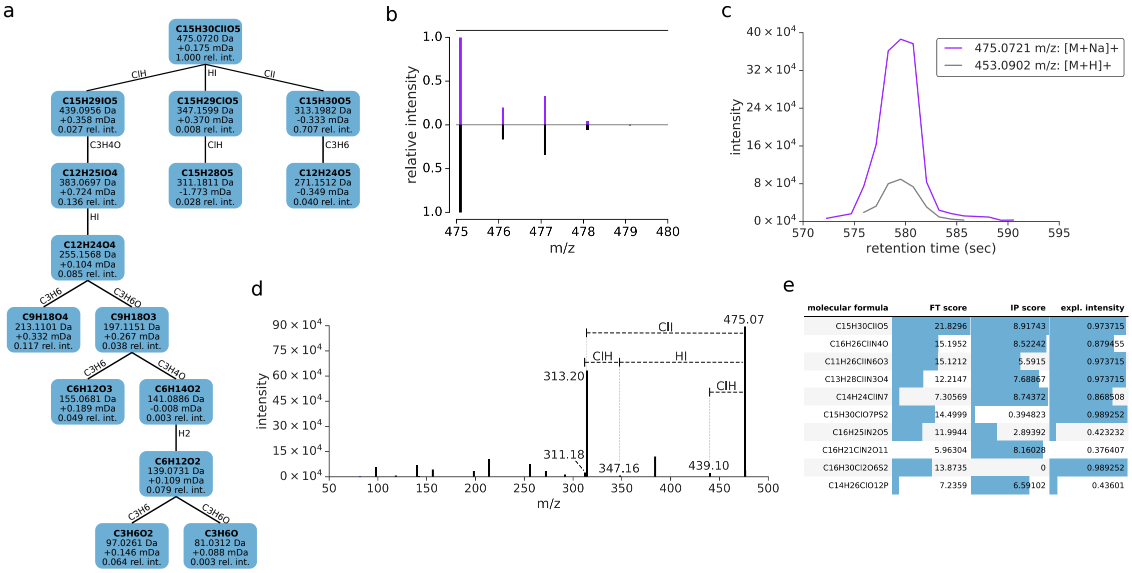


Figure 5.7: Annotation of a novel bromine-containing compound in the diatoms dataset. (a) MS/MS spectra for m/z 588.230, 590.228 and 592.325, corresponding to the monoisotopic, the M+2 and M+4 peak. The "moving" peak at m/z 570, 572 and 574 corresponds to the same molecular formula but different isotopes. This annotated fragment molecular formula is based on the fragmentation tree in (e) and is the only one containing bromine in these MS/MS spectra. (b) Partial match to 1-Palmitoyl-2-linoleoyl-sn-glycero-3-phosphocholine in the NIST library. The mirror plot compares the MS/MS spectrum of the monoisotopic peak at m/z 588.230 (top) to the NIST library spectrum (bottom). Displayed molecular formulas were annotated using the fragmentation tree of the query compound (e), and are identically annotated in the NIST reference spectrum. The substructure of the NIST reference compound which corresponds to the annotated peaks is highlighted. (c) Mirror plot of measured against simulated isotope pattern. The top part displays m/z 587 to 595 of the MS1 spectrum with retention time 503.97 sec, measured prior to the MS/MS spectrum for precursor m/z 588.230. The bottom part is the simulated isotope pattern for $[C_{24}H_{47}BrNO_8P + H]^+$. (d) Putative structure and fragmentation pathway of the novel compound. (e) Fragmentation tree computed by SIRIUS. Nodes correspond to fragments, edges to neutral losses. Nodes are annotated with the (neutralized) molecular formula, peak m/z, mass deviation in mDa and relative intensity. (f) Mirror plot of measured (top) against simulated (bottom) MS/MS spectrum for precursor M+2.



molecular formula	FT score	IP score	expl. intensity
C15H30ClIO5	21.8296	8.91743	0.973715
C16H26ClIN4O	15.1952	8.52242	0.879455
C11H26ClIN6O3	15.1212	5.5915	0.973715
C13H28ClIN3O4	12.2147	7.68867	0.973715
C14H24ClIN7	7.30569	8.74372	0.868508
C15H30ClO7PS2	14.4999	0.394823	0.989252
C16H25IN2O5	11.9944	2.89392	0.423232
C16H21ClN2O11	5.96304	8.16028	0.376407
C16H30Cl2O6S2	13.8735	0	0.989252
C14H26ClO12P	7.2359	6.59102	0.43601

Figure 5.8: Annotation of a novel chlorine- and iodine-containing compound in the diatoms dataset. (a) Fragmentation tree for $[C_{15}H_{30}ClIO_5{+}Na]^+$ computed from the corresponding MS/MS spectrum in (d). (b) Mirror plot of the measured against simulated isotope pattern. The top part displays the measured isotope pattern, extracted as described in Section 5.2.2. The bottom part is the simulated isotope pattern for $[C_{15}H_{30}ClIO_5 + Na]^+$. (c) Extracted ion chromatograms for m/z 475.0721 and m/z 453.0902 around retention time 579.16 sec with a 5 ppm mass error window. The mass difference of 21.9819 suggests that m/z 475.0721 is the sodiated ion and m/z 453.0902 the protonated ion of the same compound. (d) Merged MS/MS spectrum. The indicative neutral losses for chlorine and iodine as annotated by the fragmentation tree in (a) are displayed. (e) Top 10 SIRIUS molecular candidates. Displayed is the fragmentation tree (FT) score, isotope pattern (IP) score and the total explained intensity of fragments in the MS/MS spectrum for each candidate. Only the top 2 candidates have a ZODIAC score greater than zero.

Running times and stability.

In practice, application of Gibbs sampling can be limited by high time demand for burn-in and for sampling a reasonable number of epochs. To avoid this problem, we have used extensive algorithm engineering to reduce running times, as detailed in Section 5.1.5. Running times were measured on a computer with 40 cores (2x Intel XEON 20 Core E5-2698). We used ten parallel chain, a burn-in of 1,000 epochs and sampling of 2,000 epochs. ZODIAC required between 1 and 19 min per dataset, whereas SIRIUS required between 3 and 53 min per dataset, see Fig. 5.9. SIRIUS required most time for the dendroides dataset, which contains many high mass compounds. For dendroides, NIST1950 and mice stool, ZODIAC computation did not add much to the total running time whereas for tomato and diatoms, ZODIAC accounts for roughly one-third of the total running time. In all cases, ZODIAC running time is governed by constructing the similarity network of molecular formula candidates, whereas running the Gibbs sampler has a negligible impact. We did not evaluate our optimized Gibbs sampler against a naïve version but based on theoretical considerations in Section 5.1.5, we proved that in order to update conditional probabilities we only need to consider edges of two candidates — the new and the previous active candidate. Comparing this against a naïve computation of probabilities for the top 50 candidates, we estimate that the achieved speedup is about 25-fold.

In practice, we can speed up the construction of the similarity network, which depends quadratically on the total number of candidates: Here, we used the top 50 candidates for each compound; this conservative approach avoids the exclusion of correct molecular formulas, and also demonstrates the swiftness of our Gibbs sampling method. But running times can easily be reduced by considering fewer candidates, in particular for low mass compounds where SIRIUS usually ranks correct molecular formula much higher. Consequently, ZODIAC can be integrated into existing pipelines without substantial increase in running times.

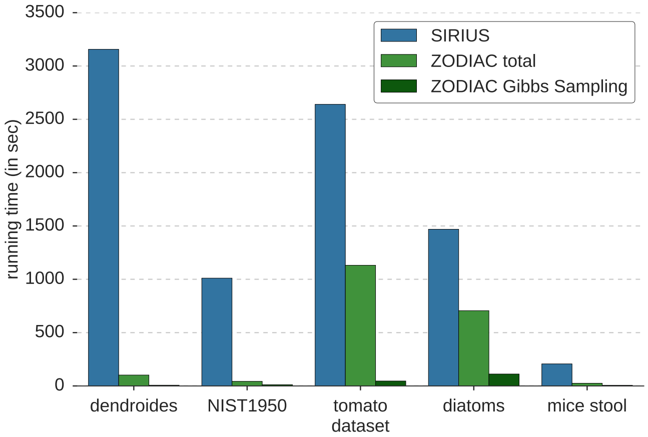


Figure 5.9: Running time comparison of SIRIUS and ZODIAC on five datasets. We run SIRIUS and ZODIAC on 2× Intel XEON E5-2698 with 40 cores total. “ZODIAC total” running time includes estimation of the edge score distribution, construction of the similarity graph and computation of ZODIAC scores via Gibbs sampling; the later running time is also given separately. ZODIAC requires SIRIUS results as input, and total processing time is SIRIUS time plus ZODIAC time.

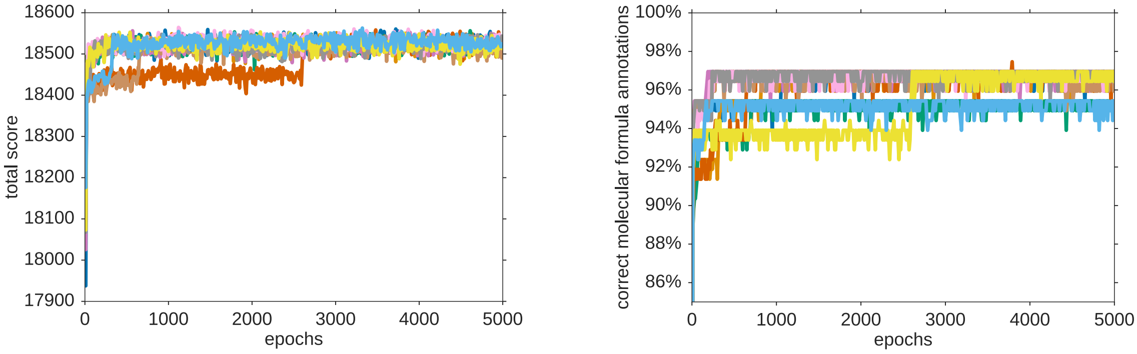


Figure 5.10: Total assignment score of the molecular formula candidate network (left) and rate of correct annotations over the course of epochs (right) for Gibbs sampling on the dendroides dataset.

With regards to stability and required number of epochs, we see that in the beginning both, the total network score and the number of correct molecular formula annotations, are increasing, see Fig. 5.10. After 500 to 1,000 epochs the 9 out of 10 Markov chains reach their different local optima. Estimating the most likely candidates from each chain individually results in 96.95 % correct molecular formula annotation in 7 of the 10 cases. In practice, we run 10 parallel Markov chains to allow for parallelization and to make sampling more robust.

6 Bayesian Network Scoring for Molecular Structure Search

In this chapter, we introduce a novel scoring for CSI:FingerID based on Bayesian networks. The method was initially presented at the annual international conference on Intelligent Systems for Molecular Biology (ISMB) 2018 in Chicago and is published [124].

CSI:FingerID searches MS/MS spectra in a molecular structure database (Section 4.4.1). Elucidation of stereochemistry is currently beyond the power of automated methods. Thus, CSI:FingerID aims to recover the constitution of a molecule, which we refer to as *structure*. It uses the fragmentation tree computed by SIRIUS to predict a molecular fingerprint. The predicted molecular fingerprint is a probabilistic fingerprint; that means, each position in the fingerprint specifies the probability that a specific molecular property is present. The predicted fingerprint of the query compound is searched in a molecular structure database by comparing it against the deterministic fingerprints of *candidate* structures. Dührkop *et al.* [61] suggested two statistical scores which perform best in evaluations: "Platt" score and the "modified Platt" score. Both scores implicitly assume independence between molecular properties. However, it is obvious that molecular properties do not have to be independent; in particular, a substructure which defines a molecular property can be completely contained in another substructure (Fig. 2.2).

Here, we present a scoring which no longer assumes independence. We model dependencies between molecular properties using a Bayesian network. The molecular properties are represented by binary random variables with the possible outcomes "presence" and "absence"; these variables are the nodes in the Bayesian network.

An atypical Bayesian network. Multiple reasons make our problem different from a "standard" Bayesian network. The total number of random variables in our model is relatively high: the molecular fingerprint contains 2,881 molecular properties. In comparison to this, the number of training examples is small, especially to establish the network structure: there are less than 17,000 reference spectra and even less different structures.

We use predicted fingerprints from CSI:FingerID as basis for our model. CSI:FingerID uses a separate kernel SVM for the prediction of each molecular property in the fingerprint. However, instead of a binary classification, CSI:FingerID uses a sigmoid function to predict the posterior probability (Platt probability) for the presence of the molecular property. These predictions "define" the *marginal probabilities* of the random variables in our model. For each query compound, CSI:FingerID predicts an individual molecular fingerprint with individual probabilities; and thus, the marginal probabilities change. In a "standard" example of Bayesian networks, such as illustrated in Section 3.5, the model parameters remain fixed once they have been estimated. Since the probabilities of the random variables in our model change for each new instance, it is also not obvious how to establish conditional probabilities.

CSI:FingerID trains each SVM independently to predict the probability of a molecular property. However, the predictions are clearly not independent. Firstly, the molecular properties themselves are not independent. The molecular fingerprint is not optimized to contain a set of complementary, uncorrelated properties. Instead, properties were selected if they could be predicted with decent quality from the training data. As a result of this selection process, some properties might be overemphasized. For example, a subset of properties might be almost identical to each other and thus their occurrence in different molecular structures is highly correlated. Such properties should be downweighted when scoring candidate structures. Secondly, we expect dependencies in the prediction of properties: For example, assume that predictions for molecular properties that contain a hydroxy group produce many false positives; an overly confident estimate for one property may also result in an overly confident estimate for another property. We also want to capture these mutual prediction errors with our model.

Relation to Matthews correlation coefficient. It is acknowledged that correlation is a meaningful measure to assess the performance of binary classifiers: The Matthews correlation coefficient (MCC) measures the correlation between observations and predictions of a binary classifier. The MCC has advantages over other common measures such as accuracy or F1 and results in fewer misleading interpretations [43]. Observation and prediction of a binary classifier can also be regarded as random variables. Thus, the connection to our problem becomes more clear. The MCC is used to assess how well binary predictions and observations agree. We, on the other hand, want to know to what extend one random variable can explain the other. We consider the probabilities of outcomes assigned to the random variables and want to relate the predicted probability of one molecular property to the predicted probability of another property. To draw a connection between our approach and the MCC: a predicted probability of 0.99 can also be interpreted as meaning that out of 100 fingerprints which all assigned a probability of 0.99 to a property, we expect 99 "presence" outcomes and one "absence" outcome. Here, instead of having one prediction with a given probability, we have many binary predictions. This interpretation using frequencies shows that it can be reasonable to consider correlations between predicted probabilities of two random variables. For our model we use covariances, the "unnormalized" correlation, as it fits well into our optimization model.

Workflow summary. We implement the Bayesian scoring in the following way: Firstly, to ease calculations, we assume that the network is a tree. This is clearly an oversimplification; however, it must be understood, that there is not only one "true" network structure but any network which improves the scoring is a valid network. The preceding step of the CSI:FingerID pipeline estimates the probability of each molecular property via individually trained kernel SVMs; we use these as *marginal probabilities* of the random variables in the Bayesian network. Secondly, we estimate the tree topology of the Bayesian network using the mutual information between molecular properties. Here, to have enough training data and to reduce overfitting, we will use deterministic molecular fingerprints from structures in a structure database. Thirdly, we estimate "desired" covariances between random variables connected in the tree. Finally, for each edge we estimate joint probabilities that simultaneously satisfy the marginal probability constraints *and* the estimated covariance values. Now, the joint probability of the complete evidence is used as a score. Our model takes into account both dependencies of molecular properties from deterministic

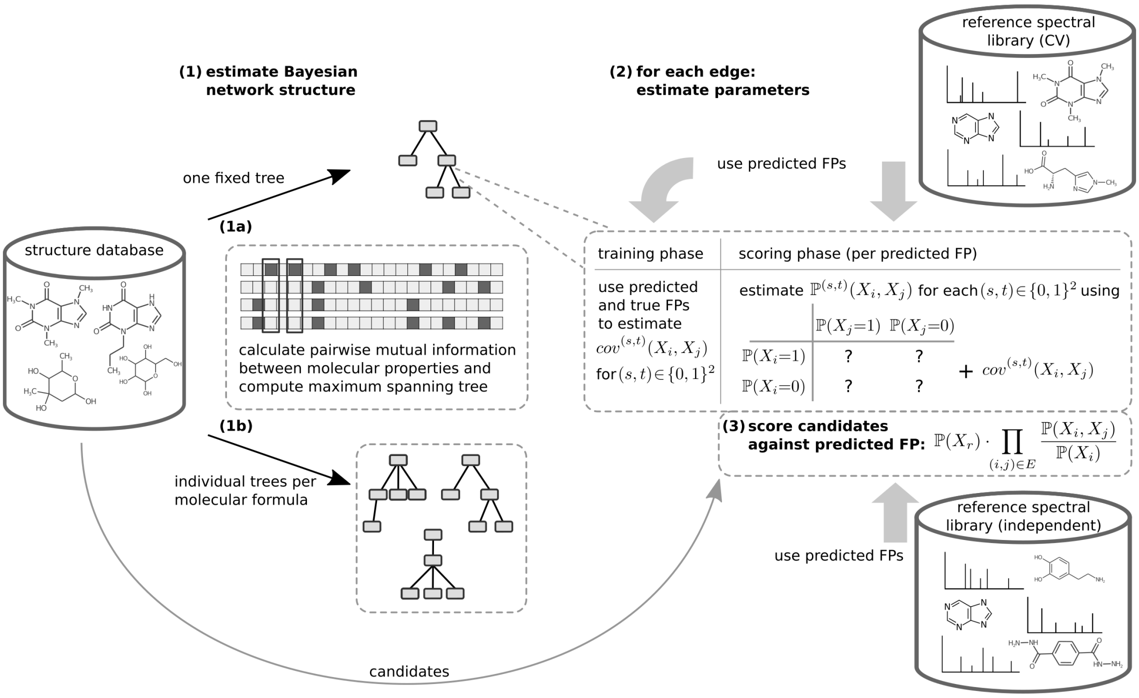

reference spectral library (CV)
(1) estimate Bayesian network structure
(2) for each edge: estimate parameters
use predicted FPs
one fixed tree
(1a)
structure database
calculate pairwise mutual information between molecular properties and compute maximum spanning tree
training phase
scoring phase (per predicted FP)
use predicted and true FPs to estimate $cov^{(s,t)}(X_i, X_j)$ for $(s,t) \in \{0,1\}^2$
estimate $\mathbb{P}^{(s,t)}(X_i, X_j)$ for each $(s,t) \in \{0,1\}^2$ using
$\mathbb{P}(X_j{=}1)$ $\mathbb{P}(X_j{=}0)$
$\mathbb{P}(X_i{=}1)$? ?
$\mathbb{P}(X_i{=}0)$? ?
$+\ cov^{(s,t)}(X_i, X_j)$
(3) score candidates against predicted FP: $\mathbb{P}(X_r) \cdot \prod_{(i,j)\in E} \frac{\mathbb{P}(X_i, X_j)}{\mathbb{P}(X_i)}$
(1b)
individual trees per molecular formula
reference spectral library (independent)
use predicted FPs
candidates

Figure 6.1: Workflow of Bayesian network estimation and candidate scoring. (1) The tree structure is determined using deterministic fingerprints (FPs) of structures and comparing mutual information between molecular properties. Different variants of the scoring use either one Bayesian network for all candidates (1a) or one network per molecular formula (1b). (2) For each edge in the tree, covariances of random variables are estimated prior to scoring candidates ("training" phase) using a set of reference compounds (here, the cross-validation (CV) library). Four different covariances are estimated per edge, one for each possible outcome combination $X_i = s, X_j = t$ for $s, t \in \{0, 1\}$. This step uses predicted and deterministic FPs. In the scoring phase, firstly, for each edge the joint probabilities $\mathbb{P}^{(s,t)}(X_i, X_j)$ are computed using the covariance and the marginal probabilities from a predicted FP. (3) Secondly, the prepared network is used to score this predicted FP against candidates from the structure database. Compounds from the CV library are used to estimate covariances ("training" phase) and are used in evaluation (scoring phase). The independent dataset is only used for evaluation.

fingerprints, and dependencies from fingerprint prediction. The workflow is illustrated in Fig. 6.1.

6.1 Tree-based Posterior Probability Estimation

The Platt score from equation (4.1) can be rewritten as follows: Assume that X_i is a binary random variable such that $\mathbb{P}(X_i = 1) = p_i$. Then, $\mathbb{P}(X = x)$ with $X = (X_1, \ldots, X_n)$ is the posterior probability of the model $x := \mathcal{M}$; and $\mathbb{P}(X = x) = \prod_i \mathbb{P}(X_i = x_i)$ if all random variables are independent.

We want to modify the posterior probability estimate to take into account dependencies between molecular properties. We model dependencies between random variables X_i, X_j (molecular properties i, j) as a *rooted tree* $T = (V, E)$ with $V = \{1, \ldots, n\}$ and $E \subseteq V \times V$, such that edges $(i, j) \in E$ describe conditional dependencies between random variables X_i and X_j; this is the simplest case of a Bayesian network. Let r be the root of T; all edges in T point away from r, which is also called “arborescence”. Then, the joint distribution can be written as

$$\begin{aligned}\mathbb{P}(X_1,\dots,X_n) &= \mathbb{P}(X_r)\cdot\prod_{(i,j)\in E}\mathbb{P}(X_j \mid X_i)\\ &= \mathbb{P}(X_r)\cdot\prod_{(i,j)\in E}\frac{\mathbb{P}(X_i,X_j)}{\mathbb{P}(X_i)}\end{aligned}\tag{6.1}$$

where $\mathbb{P}(X_i,X_j)$ is the joint distribution of X_i and X_j. In Bayesian network analysis, relationships between adjacent nodes would usually be specified via conditional probability tables for $\mathbb{P}(X_j \mid X_i)$. But for the problem at hand, we cannot estimate these conditional probabilities directly; to this end, we use the indirect estimation procedure via the joint distribution $\mathbb{P}(X_i,X_j)$.

How do we estimate $\mathbb{P}(X_i = x_i, X_j = x_j)$? We know the marginal probabilities $\mathbb{P}(X_i = 1) = p_i$ and $\mathbb{P}(X_j = 1) = p_j$ (Platt estimates of posterior probabilities) from the data $\mathcal{D}$. As X_i and X_j are binary, we have to consider exactly four cases: Set $q_{11} := \mathbb{P}(X_i = 1, X_j = 1)$, $q_{10} := \mathbb{P}(X_i = 1, X_j = 0)$, $q_{01} := \mathbb{P}(X_i = 0, X_j = 1)$, and $q_{00} := \mathbb{P}(X_i = 0, X_j = 0)$. As the marginal probabilities are known, $q_{11} + q_{10} = p_i$ and $q_{11} + q_{01} = p_j$ must hold. We also know $q_{11} + q_{10} + q_{01} + q_{00} = 1$. This means that we have one degree of freedom for choosing $q_{11}, q_{10}, q_{01}, q_{00}$.

We decided to use this degree of freedom, to ensure that the covariance of X_i, X_j equals some predetermined value $cov_{i,j} \in \mathbb{R}$, as specified in Section 6.2. This models our observation that certain molecular properties are correlated. The covariance of the binary random variables X_i, X_j is

$$\operatorname{cov}(X_i,X_j) = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = q_{11} - p_ip_j,$$

since clearly $\mathbb{E}(X_i) = p_i$ and $\mathbb{E}(X_j) = p_j$. In total, we have reached four linear equations for the four unknowns $q_{11}, q_{10}, q_{01}, q_{00}$, namely:

$$\begin{gathered} q_{11} + q_{10} = p_i, \quad q_{11} + q_{01} = p_j,\\ q_{11} + q_{10} + q_{01} + q_{00} = 1 \end{gathered}\tag{6.2}$$

$$q_{11} = cov_{i,j} + p_ip_j \tag{6.3}$$

Unfortunately, solving (6.2) and (6.3) may result in a solution that does not satisfy the obvious requirement $q_{11}, q_{10}, q_{01}, q_{00} \in [0,1]$. Whereas we think of equation (6.2) as inevitable requirements, (6.3) is a somewhat more subjective choice; to this end, we modify (6.3) accordingly:

$$q_{11} = \max\big\{0, p_i + p_j - 1, \min\{p_i, p_j, cov_{i,j} + p_ip_j\}\big\} \tag{6.4}$$

It is straightforward but cumbersome to check that choosing $q_{11}, q_{10}, q_{01}, q_{00}$ according to (6.2) and (6.4) does indeed satisfy $q_{11}, q_{10}, q_{01}, q_{00} \in [0,1]$, and that the established bounds are tight: For example, choosing $q_{11} < p_i + p_j - 1$ will violate $q_{00} \geq 0$. The covariance $\operatorname{cov}(X_i,X_j)$ of the resulting random variables does not necessarily equal $cov_{i,j}$, but if not, it is chosen “as large” or “as small” as possible. See the Lemmas below for details.

We can now determine joint probabilities $\mathbb{P}(X_i = x_i, X_j = x_j)$ for every edge (i,j), and use (6.1) to estimate the probability of evidence $X = x$, that is, the joint probability $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$; we use this estimate as the new score. To avoid numerical instabilities, we apply Laplace (additive) smoothing to probabilities $\mathbb{P}(X_i)$ and $\mathbb{P}(X_i,X_j)$

when computing (6.1). Computing $\mathbb{P}(X_i = x_i, X_j = x_j)$ can be carried out in constant time, so computing $\mathbb{P}(X = x)$ requires $O(n)$ time.

We now give formal proofs that choosing $q_{11}, q_{10}, q_{01}, q_{00}$ as described above results in probabilities from $[0, 1]$ (Lemma 6.1); and that choosing a larger (Lemma 6.2) or smaller q_{11} (Lemma 6.3) is not possible in case we deviate from the target value $cov_{i,j} + p_i p_j$.

Lemma 6.1. *Given* $p_i, p_j \in [0, 1]$ *and* $cov_{i,j} \in \mathbb{R}$. *Then,* $q_{11} := \max\big\{0, p_i + p_j - 1, \min\{p_i, p_j, cov_{i,j} + p_i p_j\}\big\}$ *from* (6.4)*,* $q_{10} := p_i - q_{11}$*,* $q_{01} := p_j - q_{11}$*, and* $q_{00} := 1 - (q_{11} + q_{10} + q_{01})$ *all satisfy* $q_{11}, q_{10}, q_{01}, q_{00} \in [0, 1]$.

Proof. Assume $q_{11}, q_{10}, q_{01}, q_{00}$ have been chosen as described. We first infer $q_{11} \leq \max\{p_i + p_j - 1, p_i\} \leq \max\{p_i, p_i\} = p_i$, and analogously $q_{11} \leq \max\{p_i + p_j - 1, p_j\} \leq p_j$. This implies $q_{11} \in [0, 1]$ as $q_{11} \geq 0$ is clear, and $q_{11} \leq p_i \leq 1$. Now, $q_{11} \leq p_i$ implies $q_{10} = p_i - q_{11} \geq p_i - p_i = 0$, and $q_{11} \leq p_j$ implies $q_{01} = p_j - q_{11} \geq p_j - p_j = 0$. Furthermore, $q_{11} \geq p_i + p_j - 1$ implies $q_{10} = p_i - q_{11} \leq p_i - (p_i + p_j - 1) = 1 - p_j \leq 1$ and $q_{01} = p_j - q_{11} \leq 1 - p_i \leq 1$. Hence, we have established $q_{10}, q_{01} \in [0, 1]$. Finally, $q_{11} \geq p_i + p_j - 1$ implies $q_{11} + q_{10} + q_{01} = p_i + p_j - q_{11} \leq p_i + p_j - (p_i + p_j - 1) = 1$ and, hence, $q_{00} \geq 0$. With $q_{00} = 1 - (q_{11} + q_{10} + q_{01}) \leq 1$ we infer $q_{00} \in [0, 1]$. ☐

Lemma 6.2. *Given* $p_i, p_j \in [0, 1]$*,* $cov_{i,j} \in \mathbb{R}$*, and* q_{11} *from* (6.4) *such that* $q_{11} < cov_{i,j} + p_i p_j$. *Then, any* $\bar{q}_{11} > q_{11}$ *with* $\bar{q}_{10} := p_i - \bar{q}_{11}$*,* $\bar{q}_{01} := p_j - \bar{q}_{11}$*, and* $\bar{q}_{00} := 1 - (\bar{q}_{11} + \bar{q}_{10} + \bar{q}_{01})$ *cannot simultaneously satisfy* $\bar{q}_{11}, \bar{q}_{10}, \bar{q}_{01}, \bar{q}_{00} \in [0, 1]$.

Proof. We do a case distinction, based on the maximum calculation of q_{11}:

(i) If $q_{11} = 0$ then $p_i, p_j \geq 0$ implies $cov_{i,j} + p_i p_j \leq 0 = q_{11}$, in contradiction to our assumptions.

(ii) If $q_{11} = p_i + p_j - 1$ then $q_{11} < cov_{i,j} + p_i p_j$ implies $\min\{p_i, p_j\} \leq p_i + p_j - 1$. Assume w.l.o.g. that $p_i \leq p_j$, then $p_i \leq p_i + p_j - 1$ and, hence, $p_j = 1$. We infer $\bar{q}_{11} > q_{11} = p_i + p_j - 1 = p_i$ and, hence, $\bar{q}_{10} = p_i - \bar{q}_{11} < 0$.

(iii) If $q_{11} = \min\{p_i, p_j, cov_{i,j} + p_i p_j\}$ then $q_{11} = \min\{p_i, p_j\} < cov_{i,j} + p_i p_j$. Hence, $\bar{q}_{11} > \min\{p_i, p_j\}$, and either $\bar{q}_{10} = p_i - \bar{q}_{11} < 0$ or $\bar{q}_{01} = p_j - \bar{q}_{11} < 0$ must hold.

☐

Lemma 6.3. *Given* $p_i, p_j \in [0, 1]$*,* $cov_{i,j} \in \mathbb{R}$*, and* q_{11} *from* (6.4) *such that* $q_{11} > cov_{i,j} + p_i p_j$. *Then, any* $\bar{q}_{11} < q_{11}$ *with* $\bar{q}_{10} := p_i - \bar{q}_{11}$*,* $\bar{q}_{01} := p_j - \bar{q}_{11}$*, and* $\bar{q}_{00} := 1 - (\bar{q}_{11} + \bar{q}_{10} + \bar{q}_{01})$ *cannot simultaneously satisfy* $\bar{q}_{11}, \bar{q}_{10}, \bar{q}_{01}, \bar{q}_{00} \in [0, 1]$.

Proof. We again do a case distinction:

(i) If $q_{11} = 0$ then $\bar{q}_{11} < 0$.

(ii) If $q_{11} = p_i + p_j - 1$ then $\bar{q}_{11} < p_i + p_j - 1$ and, hence, $\bar{q}_{11} + \bar{q}_{10} + \bar{q}_{01} = p_i + p_j - \bar{q}_{11} > p_i + p_j - (p_i + p_j - 1) = 1$, so $\bar{q}_{00} < 0$.

(iii) If $q_{11} = \min\{p_i, p_j, cov_{i,j} + p_i p_j\}$ then $q_{11} \leq cov_{i,j} + p_i p_j$ in contradiction to our assumptions.

☐

6.2 Finding the Tree and Estimating Covariances

It must be understood that in principle, every tree can be used for our computations, and there are no “incorrect” trees; our obvious goal is to reach an improved identification rate. In view of the superexponential number of trees with n nodes, we restrict our evaluation to trees that “turn up naturally” from the data. We show how to estimate the tree structure, and the desired covariance values for every edge of the tree. The tree structure is estimated solely from molecular structure data; for covariance estimation, we take into account the training data and, in particular, dependencies between predictions between molecular properties. We distinguish two cases: In the first case, we estimate one “global” fixed tree structure and desired covariance values, which is then used to score candidates for any query. In the second case, we take into account that for each query, only candidates with a particular molecular formula are considered. We compute an individual tree for this molecular formula, and also consider the molecular formula when estimating covariances. Note that molecular structure candidates of the same molecular formula are also structurally similar. As a consequence, molecular properties can be non-informative, as all structure candidates either do or do not have the property. Computing individual trees prevents that non-informative properties can “block” the path between informative properties in the Bayesian scoring tree: Non-informative properties will have mutual information zero, and will be inserted as leaves in the individual tree.

Fixed tree structure. To prevent overfitting, we do not search for a tree that maximizes identification rates. Instead, we estimate the tree structure using all molecular structures from some structure database. Mutual information is a natural choice to measure how much information we gain from one molecular property about another molecular property. We use mutual information between molecular properties from a molecular structure database as a proxy for the interdependence between random variables (predictions). For each structure in the database, we (deterministically) compute the corresponding molecular fingerprint, resulting in a multiset $\mathcal{F}$ of fingerprints. For any two molecular properties i, j we consider the corresponding binary random variables I, J; estimation of (joint) probabilities for I, J is straightforward by counting in $\mathcal{F}$. We then compute the mutual information between I and J, quantifying the “amount of information” obtained about I through J. This results in a complete graph G with nodes $\{1, \dots, n\}$, where every pair of nodes (molecular properties) is connected by an edge with weight equal to the mutual information. The tree structure is computed as a maximum spanning tree in this graph, in $O(|V|^2 \cdot \log |V|)$ time using Prim’s algorithm (with a binary heap) or Kruskal’s algorithm. Finally, we arbitrarily root this tree, as the choice of the root does not influence our computations. Some edges of the tree may have weight (mutual information) zero; this is an artifact of computing a spanning tree which connects all nodes.

Let $T = (V, E)$ be the tree; we now estimate desired covariance values. Here, we consider all compounds in the training data; only for these, we can estimate if wrong predictions of one molecular property, result in wrong predictions of another property. Each compound from the training data consists of a true fingerprint $(y_1, \dots, y_n) \in \{0, 1\}^n$ and a predicted (Platt) fingerprint $(p_1, \dots, p_n) \in [0, 1]^n$.

Consider edge $(i, j) \in E$ from molecular property i to j. We partition compounds from the training data into four batches $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$, such that a training compound with true fingerprint $(y_1, \dots, y_n)$ is sorted into batch $(y_i, y_j) \in \{0, 1\}^2$:

$$\mathcal{P}_{i,j}^{(s,t)} := \big\{(p_i, p_j) : (y_i, y_j) = (s, t)\big\}$$

We compute four covariance estimates $cov_{i,j}^{(s,t)}$, one for each batch $\mathcal{P}_{i,j}^{(s,t)}$ with $(s, t) \in \{0, 1\}^2$. Set $\mathcal{P} := \mathcal{P}_{i,j}^{(s,t)}$ for brevity; these are our *observations* used to estimate the covariance. To avoid empty batches and prevent overfitting, we add four pseudo-observations $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$ to the observations $\mathcal{P}$. We again interpret Platt probabilities p_i as the probability that a binary random variable X_i satisfies $X_i = 1$. The normalized number of observations $\mathcal{N}[a, b] \in (0, 1)$ for $a, b \in \{0, 1\}$ is

$$\begin{aligned} \mathcal{N}[1,1] &= \frac{1}{|\mathcal{P}|} \cdot \sum\nolimits_{(p,p') \in \mathcal{P}} pp', \\ &\dots \\ \mathcal{N}[0,0] &= \frac{1}{|\mathcal{P}|} \cdot \sum\nolimits_{(p,p') \in \mathcal{P}} (1-p)(1-p') \end{aligned} \tag{6.5}$$

We then estimate the desired covariance as

$$cov_{i,j}^{(s,t)} := \mathcal{N}[1,1] - \big(\mathcal{N}[1,1] + \mathcal{N}[1,0]\big) \cdot \big(\mathcal{N}[1,1] + \mathcal{N}[0,1]\big).$$

Given a candidate fingerprint $(x_1, \dots, x_n) \in \{0, 1\}^n$, we want to compute its joint probability $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$ according to (6.1). For every edge $(i, j) \in E$, we set $cov_{i,j} = cov_{i,j}^{(s,t)}$ for $s := x_i$ and $t := x_j$, and proceed to estimate $\mathbb{P}(X_i, X_j)$ as described in the previous section. Hence, every candidate fingerprint has individual covariance estimates; in the previous section, we omitted this technical detail for the sake of readability.

Finally, for artifact edges with mutual information zero, we also assume covariance $cov_{i,j}^{(s,t)} = 0$.

Individual trees. Next, we want to compute the tree and desired covariances for each query individually. Regarding the tree, we use all fingerprints from PubChem that have the molecular formula of the query when computing the mutual information. For the covariance, we proceed as described above, but again only consider those compounds from the training data that have the molecular formula of the query. But there are potentially only few such training compounds, so the method is prone to overfitting. We use the following two modifications to overcome this issue: Firstly, when estimating the observation matrix for the query molecular formula, we add the normalized observation matrix (6.5) estimated from all training data as “pseudocounts”. We give this global “pseudocounts” a weight of 14 if there are at least 10 global observations (and the 4 pseudo instances); for fewer global observations, we use the number of global observations (plus pseudo instances) as weight. Secondly, we do not only use compounds from the training data with identical molecular formula as the query; instead, we allow that training compound and query molecular formula differ by some biotransformation, such as the addition of a water H_2O. The resulting scores with and without using biotransformations will be referred to “Bayesian (individual tree)” and “Bayesian (biotransformations)”, respectively.

6.3 Evaluation on Cross-validation and Independent Dataset

CSI:FingerID and its Input Output Kernel Regression variant [29] are currently the best-performing methods for searching with MS/MS data in molecular structure databases. This has been demonstrated in two blind competitions, namely the Critical Assessment for Small Molecule Identification (CASMI) contests 2016 and 2017[1]. CASMI 2016 (category 2) provided data for 127 compounds in positive ion mode, of which CSI:FingerID correctly identified 70 [182], more than twice the number of the best non-CSI:FingerID method: In detail, MS-FINDER [202], CFM-ID [2], MAGMA+ [208] and MetFrag2.3 [176, 228] had 32, 27, 16 and 15 correct identifications, respectively. In CASMI 2017 (category 4), CSI:FingerID identified six-fold the number of compounds of the best non-CSI:FingerID method. This is in agreement with finding by Dührkop *et al.* [61, 63] who found that CSI:FingerID outperforms the runner-up 2.5-fold. To this end, we refrain from evaluating against other methods.

We follow the evaluation setup of Dührkop *et al.* [61]. In our evaluation, we make sure that all evaluated structures are *novel*: That is, no MS/MS data from a compound with the same structure is present in the training data. For example, for D-threonine to be novel, the training data must not contain any MS/MS spectra for D-threonine, L-threonine, or (D or L)-*allo*-threonine. We use 10-fold cross validation when predicting fingerprints for the training data; no two folds contain the same structure. For the independent dataset, we ensure novel structure evaluation by using, for each query, the cross-validation model which does not contain the query structure; in case the query structure is not present in the training data, we use a model trained on all training data.

We extracted 91 molecular formulas of biotransformations from Li *et al.* [118], Rogers *et al.* [174]; we excluded large modifications above 100 Da and modifications not composed from CHNO, resulting in 29 modifications used here: namely, C_2H_2, C_2H_2O, C_2H_3NO, $C_2H_3O_2$, C_2H_4, C_2O_2, $C_3H_2O_3$, C_3H_5NO, $C_3H_5NO_2$, C_3H_5O, $C_4H_2N_2O$, $C_4H_3N_3$, $C_4H_4O_2$, C_5H_7, C_5H_7NO, C_5H_9NO, CH_2, CH_2ON, CH_3N_2O, CHO_2, CO, CO_2, H_2, H_2O, N, NH, NH_2, NH_3, and O. These biotransformations are used to increase the number of specific compounds used to compute the covariances for individual trees.

6.3.1 Datasets and Databases

Next to spectra from MassBank [90] and GNPS [211] we trained CSI:FingerID on data from the NIST 2017 database (National Institute of Standards and Technology, v17). As evaluated here, CSI:FingerID 1.12 is trained is trained on 13,766 structures and 16,865 measurements in positive ion mode. For one compound, a library may contain several MS/MS spectra, which are merged by SIRIUS 3.6 into a single spectrum [19]. As described in previous publications [19, 61, 62], we discard certain instances based on strong deviation of the precursor mass etc; we leave out the tedious details, as these are not important here. As an independent dataset, we use the commercial "MassHunter Forensics/Toxicology PCDL" library (Agilent Technologies, Inc.) with 3,451 spectra.

Tree structures are computed from molecular structures, without taking into account MS/MS data. To compute the fixed tree structure we use 236,656 molecular structures from databases of biological interest, namely KNApSAcK [186], HMDB [226], ChEBI [81], KEGG [97], BioCyc [37], UNPD [76], and MeSH-annotated compounds from PubChem

---

[1] `http://casmi-contest.org/`

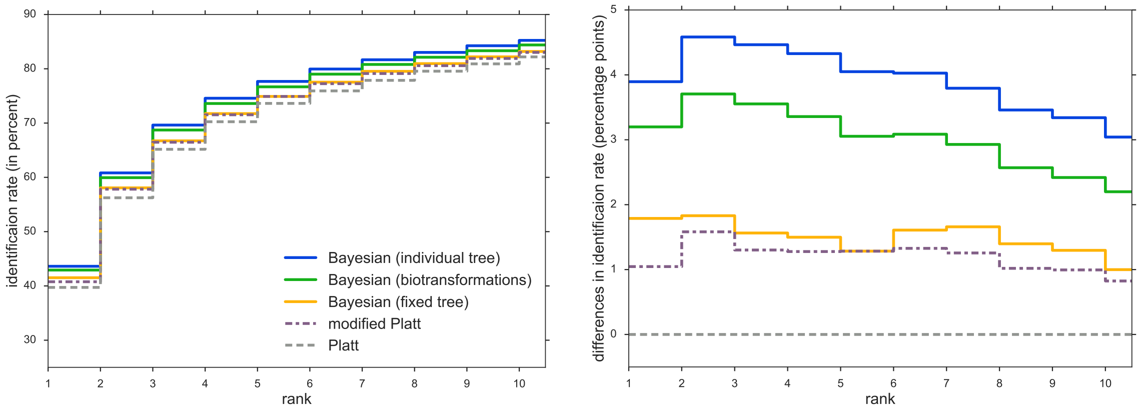


Figure 6.2: Left: Identification rates using different CSI:FingerID scores, for cross-validation. We report the percentage of instances where the correct structure was identified in the top k, for varying k. Scores are Platt, modified Platt, Bayesian (fixed tree) and Bayesian (individual tree) and Bayesian (biotransformations). Note the zoomed y-axis. Right: Percentage point differences in identification rates against the Platt score, for cross-validation.

[103]. In contrast, the individual tree structures specific for one query are computed from all PubChem structures with the same molecular formula (or the molecular formula plus corresponding biotransformation). We use a local copy of PubChem from August 13, 2017 containing 93,859,798 compounds and 73,444,774 structures.

6.3.2 Results

CSI:FingerID reached 31.8 % correct identifications in cross-validation on the GNPS dataset [61], which was 2.6-fold higher than the runner-up method. Since then, numerous methodical improvements (for example, novel kernels) as well as additional training data have further improved the performance of CSI:FingerID. On the other hand, PubChem, the database we search in, has greatly increased in size which, in turn, makes it harder to identify the correct molecular structure. We evaluate on 16,865 cross-validation compounds and the independent dataset from Agilent. For each method, we report the ratio of instances where a method ranked the correct structure in its top k output, for $k = 1, \dots, 10$. We evaluate the new scores — termed “Bayesian (fixed tree)”, “Bayesian (individual tree)” and “Bayesian (biotransformations)” — in addition to the Platt and modified Platt scores from [61]. All new scores are derived from the standard Platt score, which makes it the baseline method. Still, “modified Platt” is the currently best-performing score to beat.

method	top 1	top 5	top 10
Bayesian (individual tree)	43.62 ± 1.53	77.67 ± 0.90	85.23 ± 1.05
Bayesian (biotransformations)	42.92 ± 1.52	76.68 ± 0.82	84.39 ± 0.97
Bayesian (fixed tree)	41.51 ± 1.10	74.91 ± 0.88	83.19 ± 1.18
modified Platt	40.77 ± 0.92	74.91 ± 1.35	83.02 ± 1.35
Platt	39.72 ± 1.44	73.62 ± 1.33	82.19 ± 1.34

Table 6.1: Identification rates with standard deviations using different CSI:FingerID scores on 10-fold cross-validation. We report the percentage where the correct structure was identified in the top k.

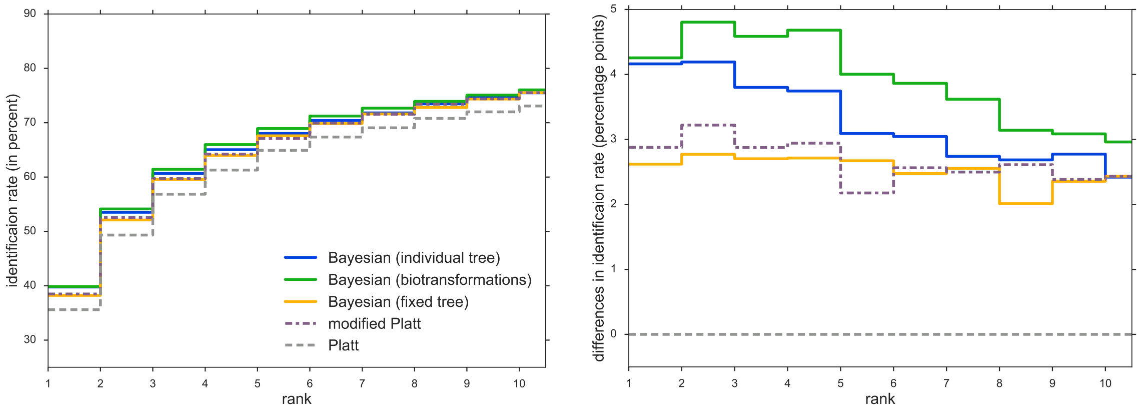


Figure 6.3: Left: Identification rates using different CSI:FingerID scores, for Agilent. We report the percentage of instances where the correct structure was identified in the top k, for varying k. Scores are Platt, modified Platt, Bayesian (fixed tree) and Bayesian (individual tree) and Bayesian (biotransformations). Note the zoomed y-axis. Right: Percentage point differences in identification rates against Platt score, for Agilent.

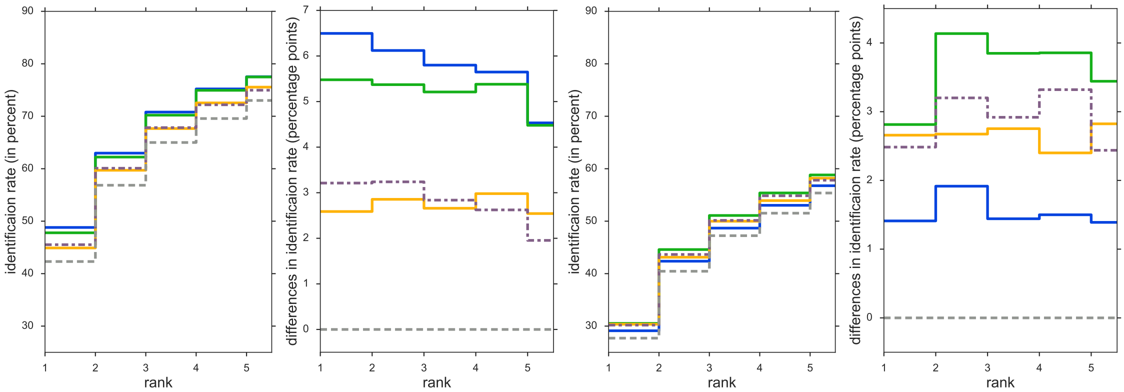


Figure 6.4: Left: Identification rates and differences using different CSI:FingerID scores, for Agilent (known structures). Right: Identification rates and differences using different CSI:FingerID scores, for Agilent (unknown structures). For legend and further details see Fig. 6.3.

We find that all new scores outperform Platt and modified Platt in cross-validation, see Fig. 6.2 and Table 6.1. Bayesian (individual tree and biotransformations) achieve highest identification rates of 43.62 % and 42.92 %, respectively. This is an improvement of 3.20 to 3.89 percentage points to the baseline, and improves modified Platt by 2.85 percentage points.

Identification rates on Agilent are all slightly lower than on cross-validation. Bayesian (biotransformations) achieves the best top 1 rate with 39.86 % (Fig. 6.3). Both Bayesian (biotransformations) and Bayesian (individual tree) improve on the modified Platt's identification rate by more than 1.28 percentage points.

Predicting fingerprints of Agilent compounds, we ensured to use CSI:FingerID models not trained on this specific structure. Nevertheless, we used all cross-validation compounds to compute covariances for the three Bayesian scores. We want to asses how this influences the performance on the Agilent dataset. To this end, we split the dataset in two groups: Agilent (known structure) with 1868 compounds and Agilent (unknown structure) with

Statistical test	vs. Platt	vs. mod. Platt
Cross validation, Welch's t-test (ten folds)	$6.4 \cdot 10^{-5}$	$8.0 \cdot 10^{-4}$
Cross validation, sign test on wins ($N = 16\,865$)	$1.5 \cdot 10^{-60}$	$3.1 \cdot 10^{-21}$
Agilent, sign test on wins ($N = 3\,451$)	$2.4 \cdot 10^{-11}$	0.57
CASMI 2016, sign test on wins ($N = 127$)	$8.4 \cdot 10^{-6}$	0.0017

Table 6.2: P-values of method comparison for Bayesian (biotransformations) vs. Platt and modified Platt. Using a one-tailed Welch's t-test, we test if variations in correct identifications are significantly larger between methods than between folds. For the one-tailed sign test, a "win" means that method A reaches a better rank than method B; ties for the top rank are removed (no method can outperform the other method for these seemingly simple instances), other ties are equally distributed between the two methods (conservative approach). For Agilent, wins of Bayesian (biotransformations) vs. modified Platt, no method performs significantly better than the other.

1583 compounds. The first group contains those compounds with structure contained in the cross-validation dataset; the second contains completely novel compounds not even used for estimating covariance. See Fig. 6.4: On Agilent with known structure, Bayesian (biotransformations) and Bayesian (individual tree) clearly outperform all other methods. On Agilent with unknown structure, Bayesian (individual tree) loses its performance edge over modified Platt, but still clearly outperforms Platt. Bayesian (biotransformations) consistently outperforms both, Platt and modified Platt, on all datasets, even on completely novel compounds. We stress that all three new scores improve on their baseline method in every case. We argue that all three Bayesian scores only have minor tendencies to overfit, as they still beat their baseline method on novel structures. Actually, Bayesian (biotransformations) generalizes good enough to beat modified Platt on all datasets.

Finally, we evaluated the 127 instances in positive ion mode from CASMI 2016 [182], again ensuring that all structures are novel when predicting fingerprints. All Bayesian scores outperform Platt and modified Platt: Bayesian (biotransformations) and Bayesian (individual tree) reach 36.61 %, Bayesian (fixed tree) reaches 35.04 % correct identifications. In comparison, Platt and modified Platt identify 26.38 % and 30.31 % correctly.

Are the reported improvements statistically significant? We evaluated significance using the one-tailed Welch's t-test for cross validation, and the one-tailed sign test for wins (one method reaches a better rank than the other method) for all datasets. We test Bayesian (biotransformations) against Platt and modified Platt. Against Platt, all p-values are highly significant (below $6.4 \cdot 10^{-5}$). Against modified Platt, all p-values except for "wins on Agilent dataset" are significant (below 0.0017). See Table 6.2 for details.

7 Conclusion

In this thesis, we presented two novel computational methods for the automated analysis of MS/MS spectra of small molecules. Both methods take a Bayesian perspective on the underlying problems; and both improve on state-of-the-art methods, SIRIUS and CSI:FingerID.

ZODIAC is a Gibbs sampling-based approach for the *de novo* assignment of molecular formulas in biological samples analyzed by LC-MS/MS. Using ZODIAC, we observed a substantial increase in the number of correct molecular formula annotations. In particular, on one dataset with large compounds, error rates decrease by 16-fold. Furthermore, the ZODIAC score allows to select the most confident annotations. Different from many other approaches, ZODIAC is not limited to molecular formulas present in any (spectral or structural) databases. We have seen that this is not only of theoretical interest: We confirmed two novel molecular formulas discovered by ZODIAC. One was added to PubChem just recently by automatically processing the annotated spectrum we uploaded to GNPS. The other molecular formula is still absent from PubChem.

We found that adduct annotations are very important for molecular formula assignment, as it is challenging to deduce this information from isotope pattern and MS/MS data. Hence, high-quality adduct annotations should be established during preprocessing. In contrast, we observed that anchors (library hits) have only a small effect on molecular formula annotations.

The mouse stool dataset was particularly challenging. We had to discard many MS/MS spectra because these did not meet our quality standards. This does not imply, that spectral library search, which is typically used as a default method, is more robust and can even identify compounds from low quality spectra. Rather, it is acknowledged that only a small portion of the data will be annotated with spectral library matches; hence, low quality spectra likely remain unrecognized. We, on the other hand, aim for a more comprehensive annotation. Unfortunately, we had to exclude many MS/MS spectra from this dataset; either because these contained too few peaks or were chimeric. Nevertheless, the chimeric filtering which we applied was not overly strict. It is likely, that many MS/MS spectra that were not discarded still contained isotope peaks or peaks from interfering compounds. As a result of the filtering, the mouse dataset contained the lowest number of compounds — another potential hindrance to the ZODIAC model which requires co-occurring, similar compounds to optimize annotations. We found that SIRIUS had difficulties with this dataset as well. We argue, that spectrum quality is a crucial factor for the automated analysis of mass spectrometry data. On the remaining datasets, the number of correct molecular formula assignments from ZODIAC was greatly increased.

ZODIAC is the first tool which combines *de novo* molecular formula annotation with high identification rates and a score that allows to asses the confidence in assignments. Confident annotations are a requirement for the automated screening of datasets in large scale.

Molecular formula analysis is usually not the final step in small molecule analysis. Searching an unknown compound with novel molecular formula in a structure database will always result in an incorrect hit, and this will often go unnoticed. In contrast, a metabolite identification workflow which makes use of *de novo* annotation methods facilitates the identification of highly interesting, new metabolites. Here, ZODIAC constitutes a major step in the discovery and structural elucidation of novel metabolites, natural products, and other molecules of biological interest.

As a second computational method, we have introduced a novel scoring for CSI:FingerID that does not only outperform previous scorings for searching in molecular structure databases, but also allows for a statistical interpretation. The scoring interprets the problem as computing the probability of evidence in a Bayesian network. This problem has the unusual property, that the marginal probabilities differ for each predicted query fingerprint. In order to create the scoring, we apply Bayesian networks in a novel and unexpected way; estimating the conditional probabilities from covariances and marginal probabilities has, to the best of our knowledge, not been suggested before in the literature.

To create a scoring adapted to the compound at hand, we compute many individual trees, one for each molecular formula. We have observed a slight tendency for overfitting in our method; we conjecture that this is due to estimating the covariance from prediction dependencies on the training data. We included biotransformations to overcome this effect. We stress that 2 percentage points of additional correct identifications represent a significant advancement: As a back-of-the-envelope calculation, we estimate that CSI:FingerID would require 1,400 to 3,000 novel reference compounds (with structures currently not contained in the training data) to reach this improvement via additional training data. Finally, we hope that simultaneously reaching improved identification rates plus a statistical interpretation may pave the way toward significance measures such as false discovery rates.

Both methods contribute to the ultimate goal of fully automated, high-throughput analysis and structure annotation of small molecules from LC-MS/MS data. Both methods are integrated into SIRIUS 4.

Ongoing research and future work

For the Bayesian network scoring, we have modeled dependencies between properties as a tree. This naturally raises the question, how we can extend the Bayesian network model to use a more complex structure. Different from the scoring presented here, the "modified Platt" score from Dührkop *et al.* [61] has no statistical interpretation and is in fact slightly counter-intuitive; it is noteworthy that this score consistently performs so well. It remains an open question why this is the case, and how we can formalize this effect.

ZODIAC greatly improves the rate of correct molecular formula annotations. Nevertheless, it leaves room for improvement. Clusters of almost identical compounds interfere with the global similarity estimation. These clusters contain isomers or replicate measurements of the same compound in different LC-MS/MS runs which were not combined due to large differences in retention time. Reciprocal scoring bonuses during Gibbs sampling yield a self-enforcing dynamic. As a result, the high inner-cluster similarities dominate similarities to other compounds and annotations within the cluster are primarily based on the cluster compounds. A possible solution would be to forbid edges between compounds with the same mass. Alternatively the probability estimation could be reformulated in a way that a large number of edges to similar compounds does not result in exaggerated confidence. An expectation maximization approach could re-estimate conditional probabilities after each epoch [142]. ZODIAC struggled with unresolved adduct annotations. There are natural ways to include adduct assignment directly into the Gibbs sampling.

The next big effort will be to develop a method similar in idea to ZODIAC for the structure identification task. Methods have already been published which take advantage of compound similarities in an LC-MS/MS run to re-rank structure candidates [49]. However, these may be prone to database biases. Molecular structures in databases are highly non-uniformly distributed: for some structures, the database will contain many similar structures differing by a biotransformation. These will be favored over more "isolated" structures. In order to overcome this bias we have to correct for expected random similarities of a structure candidate to other structures in the database; and simultaneously estimate the probability that the true hit is indeed contained in the database. Alternatively, structure database search may be completely avoided on the whole-dataset level. Co-occurring compounds in a dataset might help to correct wrongly predicted properties in the molecular fingerprints. This can improve a subsequent structure database search; but it also aids other approaches which use predicted fingerprints but do *not* search in structure databases.

There is ongoing research to complement the existing methods and solve some of the discussed issues of mass spectrometry-based metabolomics. Dührkop et al. recently developed CANOPUS [64] to predict compound classes [55] from MS/MS spectra. It takes the molecular fingerprint predicted by CSI:FingerID as input and annotates the compound with a subset from over 1,200 compound classes; thus, overcoming the limitations of database search.

Martin Hoffmann is developing a confidence score for CSI:FingerID results which distinguishes true from bogus identifications. This provides another important step towards fully automated structure identification workflows. A confidence score for CSI:FingerID structure identification strongly depends on the applied scoring method. Here, the confidence score will presumably profit from the Bayesian network score. Both novel methods benefit from an accurate molecular formula assignment provided by ZODIAC as a preceding step.

As we aim for comprehensive annotation of whole datasets, spectrum quality and compound coverage become increasingly important. Compound similarity estimations, such as performed by ZODIAC or in molecular networking, work best for MS/MS spectra with a reasonable number of fragment peaks and are impaired by chimeric spectra. Improved quality assessment should be integrated in all workflows [143]. Moreover, experiments with increased compound coverage will provide a better data basis [28].

Another topic worth considering is a better quantification and depiction of uncertainty in structure identifications. Mass spectrometry does not provide enough information to perform structure elucidation for all compounds. A structure database hit might be highly similar to the true structure but still "incorrect" because a hydroxy group moved by one position. Nevertheless, such bogus annotations are useful for interpretation. The question is: how can we detect uncertain parts in our structure annotation and what is the best way to communicate these uncertainties?

Computational methods for small molecule mass spectrometry are continuously improving and will transform the way metabolomics experiments are going to be conducted in the coming years. Particularly for methods that specifically analyze complete biological datasets, we will need increased effort to establish proper gold standard datasets to evaluate against, because standard reference libraries might not be applicable. Combined effort of open data repositories and method developers will provide the foundation for improved compound annotation and knowledge generation based on metabolomics experiments. We proved that ZODIAC identifies multiple compounds with novel molecular formula by considering only five datasets. We anticipate many more such findings, when we apply all our novel methods large-scale to hundreds and thousands of biological datasets.

Bibliography

[1] G. W. Adamson and J. A. Bush. A method for the automatic classification of chemical structures. *Inf Stor Ret*, 9(10):561–568, 1973.

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

[3] F. Allen, A. Pon, R. Greiner and D. Wishart. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Anal Chem*, 88(15):7689–7697, 2016.

[4] T. Alon and A. Amirav. Isotope abundance analysis methods and software for improved sample identification with supersonic gas chromatography/mass spectrometry. *Rapid Commun Mass Spectrom*, 20(17):2579–2588, 2006.

[5] U. Andjelković, M. Šrajer Gajdošik, D. Gašo-Sokaš, T. Martinović and D. Josić. Foodomics and food safety: Where we are. *Food Technol Biotechnol*, 55(3):290–307, 2017.

[6] N. Aronszajn. Theory of reproducing kernels. *Trans Am Math Soc*, 68(3):337–404, 1950.

[7] D. Bajusz, A. Rácz and K. Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminf*, 7(1):20, 2015.

[8] M. Baker. Metabolomics: From small molecules to big ideas. *Nat Methods*, 8(2): 117–121, 2011.

[9] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner and M. Dunkel. Super Natural II - a database of natural products. *Nucleic acids research*, 43:D935–D939, 2015.

[10] D. A. Barkauskas and D. M. Rocke. A general-purpose baseline estimation algorithm for spectroscopic data. *Anal Chim Acta*, 657(2):191–197, 2010.

[11] C. A. Bauer and S. Grimme. How to compute electron ionization mass spectra from first principles. *J Phys Chem A*, 120(21):3755–3766, 2016.

[12] J. Becker. *Inorganic mass spectrometry : principles and applications.* John Wiley & Sons, Ltd, Chichester, England, 2007.

[13] A. Bender and R. C. Glen. Molecular similarity: A key technique in molecular informatics. *Org Biomol Chem*, 2(22):3204–3218, 2004.

[14] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies. How similar are similarity searching methods? a principal component analysis of molecular descriptor space. *J Chem Inf Model*, 49(1):108–119, 2009.

[15] H. P. Benton, D. M. Wong, S. A. Trauger and G. Siuzdak. XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal Chem*, 80(16):6382–6389, 2008.

[16] I. Blaženović, T. Kind, J. Ji and O. Fiehn. Software tools and approaches for compound identification of lc-ms/ms data in metabolomics. *Metabolites*, 8(2), 2018.

[17] K. Blin, S. Shaw, K. Steinke, R. Villebro, N. Ziemert, S. Y. Lee, M. H. Medema, and T. Weber. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*, 47(W1):W81–W87, 2019.

[18] S. Böcker. Searching molecular structure databases using tandem MS data: are we there yet? *Curr Opin Chem Biol*, 36:1–6, 2017.

[19] S. Böcker and K. Dührkop. Fragmentation trees reloaded. *J Cheminform*, 8:5, 2016.

[20] S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

[21] S. Böcker, M. Letzel, Zs. Lipták and A. Pervukhin. Decomposing metabolomic isotope patterns. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, volume 4175 of *Lect Notes Comput Sci*, pages 12–23. Springer, Berlin, 2006.

[22] S. Böcker, Zs. Lipták, M. Martin, A. Pervukhin and H. Sudek. DECOMP–from interpreting mass spectrometry peaks to solving the Money Changing Problem. *Bioinformatics*, 24(4):591–593, 2008.

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

[24] C. Boone and J. Adamec. Top-down proteomics. In P. Ciborowski and J. Silberring, editors, *Proteomic Profiling and Analytical Chemistry (Second Edition)*, pages 175 – 191. Elsevier, Boston, second edition edition, 2016.

[25] M. E. Borsuk, P. Reichert, A. Peter, E. Schager and P. Burkhardt-Holm. Assessing the decline of brown trout (salmo trutta) in swiss rivers using a bayesian probability network. *Ecol Modell*, 192(1):224–244, 2006.

[26] J. Boström, A. Hogner and S. Schmitt. Do structurally similar ligands bind in a similar fashion? *J Med Chem*, 49(23):6716–6725, 2006.

[27] A. Bouslimani, R. da Silva, T. Kosciolek, S. Janssen, C. Callewaert, A. Amir, K. Dorrestein, A. V. Melnik, L. S. Zaramela, J.-N. Kim, G. Humphrey, T. Schwartz, K. Sanders, C. Brennan, T. Luzzatto-Knaan, G. Ackermann, D. McDonald, K. Zengler, R. Knight, and P. C. Dorrestein. The impact of skin care products on skin chemistry and microbiome dynamics. *BMC Biol*, 17(1):47, 2019.

[28] C. D. Broeckling, E. Hoyes, K. Richardson, J. M. Brown and J. E. Prenni. Comprehensive tandem-mass-spectrometry coverage of complex samples enabled by data-set-dependent acquisition. *Anal Chem*, 90(13):8020–8027, 2018.

[29] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker and J. Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12): i28–i36, 2016. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2016).

[30] C. Brouard, E. Bach, S. Böcker and J. Rousu. Magnitude-preserving ranking for structured outputs. In *Proc. of Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 407–422. PMLR, 2017.

[31] C. Brouard, A. Bassé, F. d'Alché Buc and J. Rousu. Improved small molecule identification through learning combinations of kernel regression models. *Metabolites*, 9, 2019.

[32] N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *J Chem Inf Model*, 59(3):1096–1108, 2019.

[33] M. L. Brownawell and J. S. Filippo Jr. A program for the synthesis of mass spectral isotopic abundances. *Journal of Chemical Education*, 59(8):663–65, 1982.

[34] B. Buchanan, G. Sutherland and E. A. Feigenbaum. *Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry*, volume 4 of *Machine Intelligence*, page 209. Edinburgh University Press, 1969.

[35] E. S. Burnside, D. L. Rubin, J. P. Fine, R. D. Shachter, G. A. Sisney and W. K. Leung. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: Initial experience. *Radiology*, 240(3):666–673, 2006.

[36] M. T. Cabrita, C. Vale and A. P. Rauter. Halogenated compounds from marine algae. *Mar Drugs*, pages 2301–17, 2010.

[37] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*, 42(D1):D459–D471, 2014.

[38] L. S. Castillo-Peinado and M. D. Luque de Castro. Present and foreseeable future of metabolomics in forensic analysis. *Anal Chim Acta*, 925:1–15, 2016.

[39] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58 – 63, 2015. Virtual Screening.

[40] E. Champarnaud and C. Hopley. Evaluation of the comparability of spectra generated using a tuning point protocol on twelve electrospray ionisation tandem-in-space mass spectrometers. *Rapid Commun Mass Spectrom*, 25(8):1001–1007, 2011.

[41] J. Chen, X. Huang, I. A. Kanj and G. Xia. Strong computational lower bounds via parameterized complexity. *J Comp System Sci*, 72(8):1346–1367, 2006.

[42] S. H. Chen and C. A. Pollino. Good practice in bayesian network modelling. *Environ Model Softw*, 37:134–145, 2012.

[43] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[44] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artif Intell*, 42(2-3):393–405, 1990.

[45] C. Cortes, M. Mohri and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *J Mach Learn Res*, 13(1):795–828, 2012.

[46] E. J. Culp, G. Yim, N. Waglechner, W. Wang, A. C. Pawlowski and G. D. Wright. Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat Biotechnol*, 37(10):1149–1154, 2019.

[47] R. R. da Silva, F. Jourdan, D. M. Salvanha, F. Letisse, E. L. Jamin, S. Guidetti-Gonzalez, C. A. Labate, and R. Z. N. Vêncio. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*, 30(9):1336–1337, 2014.

[48] R. R. da Silva, P. C. Dorrestein and R. A. Quinn. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A*, 112(41):12549–12550, 2015.

[49] R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes, and P. C. Dorrestein. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput Biol*, 14(4):e1006089, 2018.

[50] R. Daly, S. Rogers, J. Wandy, A. Jankevics, K. E. V. Burgess and R. Breitling. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics*, 30(19):2764–2771, 2014.

[51] T. De Vijlder, D. Valkenborg, F. Lemière, E. P. Romijn, K. Laukens and F. Cuyckens. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass spectrometry reviews*, 37(29120505):607–629, 2018.

[52] F. Del Carratore, K. Schmidt, M. Vinaixa, K. A. Hollywood, C. Greenland-Bews, E. Takano, S. Rogers, and R. Breitling. Integrated probabilistic annotation: A bayesian-based annotation method for metabolomic profiles integrating biochemical connections, isotope patterns, and adduct relationships. *Anal Chem*, 91(20):12799–12807, 2019.

[53] V. M. Dembitsky and M. Srebnik. Natural halogenated fatty acids: their analogues and derivatives. *Prog. Lipid Res.*, 41(4):315 – 367, 2002.

[54] R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics.* Springer-Verlag, 5th edition, 2017.

[55] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, and D. S. Wishart. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminf*, 8(1):61, 2016.

[56] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la Fuente, R. Greiner, C. Manach and D. S. Wishart. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminf*, 11(1):2, 2019.

[57] R. G. Downey and M. R. Fellows. *Parameterized Complexity.* Springer, Berlin, 1999.

[58] J. Duan, S. L. Dixon, J. F. Lowrie and W. Sherman. Analysis and comparison of 2d fingerprints: Insights into database screening performance using eight fingerprint methods. *J Mol Graphics Modell*, 29(2):157 – 170, 2010.

[59] A. Z. Dudek, T. Arodz and J. Gálvez. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High T Scr*, 9(3):213–228, 2006.

[60] K. Dührkop, M. Ludwig, M. Meusel and S. Böcker. Faster mass decomposition. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2013)*, volume 8126 of *Lect Notes Comput Sci*, pages 45–58. Springer, Berlin, 2013.

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

[62] K. Dührkop, M. A. Lataretu, W. T. J. White and S. Böcker. Heuristic algorithms for the maximum colorful subtree problem. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2018)*, volume 113 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 23:1–23:14, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[63] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, and S. Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 16(4): 299–302, 2019.

[64] K. Dührkop, L. F. Nothias, M. Fleischauer, M. Ludwig, M. A. Hoffmann, J. Rousu, P. C. Dorrestein, and S. Böcker. Classes for the masses: Systematic classification of unknowns using fragmentation spectra. *bioRxiv*, 2020.

[65] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–214, 2007.

[66] W. Fenical. Halogenation in the rhodophyta - a review. *J. Phycol.*, 11(3):245–259, 1975.

[67] A. R. Fernie, R. N. Trethewey, A. J. Krotzky and L. Willmitzer. Metabolite profiling: From diagnostics to systems biology. *Nat Rev Mol Cell Biol*, 5(9):763–769, 2004.

[68] O. Fiehn. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends Anal Chem*, 27(3):261–269, 2008.

[69] O. Fiehn, J. Kopka, R. N. Trethewey and L. Willmitzer. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal Chem*, 72(15):3573–3580, 2000.

[70] C. Frainay, E. L. Schymanski, S. Neumann, B. Merlet, R. M. Salek, F. Jourdan and O. Yanes. Mind the gap: Mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites*, 8:51, 2018.

[71] P. J. Gale and M. L. Vestal. The development of time-of-flight mass spectrometry. In M. L. Gross and R. M. Caprioli, editors, *The Encyclopedia of Mass Spectrometry*, pages 34 – 42. Elsevier, Boston, 2016.

[72] J. Gasteiger, W. Hanebeck and K.-P. Schulz. Prediction of mass spectra from structural information. *J Chem Inf Comput Sci*, 32(4):264–271, 1992.

[73] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*, PAMI-6 (6):721–741, 1984.

[74] M. Gerlich and S. Neumann. MetFusion: integration of compound identification strategies. *J Mass Spectrom*, 48(3):291–298, 2013.

[75] J. H. Gross. *Mass Spectrometry: A textbook.* Springer, Berlin, 3rd edition, 2017.

[76] J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu and X. Xu. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One*, 8(4): 1–10, 2013.

[77] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker and A. Wassermann. MOLGEN 5.0, a molecular structure generator. In *Advances in mathematical chemistry and applications*, pages 113–138. Elsevier, 2015.

[78] C. Guijas, J. R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A. E. Aisporna, *et al.* METLIN: A technology platform for identifying knowns and unknowns. *Analytical chemistry*, 90(5):3156–3164, 2018.

[79] H. Guo and W. Hsu. A survey of algorithms for real-time bayesian network inference. In *Workshop on Real-Time Decision Support and Diagnosis Systems*. American Association for Artificial Intelligence, 2002.

[80] M. Gütlein and S. Kramer. Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. *J Cheminf*, 8(1):60, 2016.

[81] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res*, 41(Database issue):D456–D463, 2013.

[82] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[83] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola and J. Rousu. FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, 2008.

[84] M. Heinonen, H. Shen, N. Zamboni and J. Rousu. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics*, 28(18):2333–2341, 2012.

[85] S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi. Inchi, the iupac international chemical identifier. *J Cheminform*, 7:23, 2015.

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

[87] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman and D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem*, 80(14):5574–5582, 2008.

[88] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner and A. Zell. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. *J Cheminform*, 3(1):3, 2011.

[89] T. Hofmann, B. Schölkopf and A. J. Smola. Kernel methods in machine learning. *Ann Stat*, pages 1171–1220, 2008.

[90] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, 2010.

[91] S. M. Houten. Metabolomics: Unraveling the chemical individuality of common human diseases. *Ann Med*, 41(6):402–407, 2009.

[92] J. P. Huelsenbeck, F. Ronquist, R. Nielsen and J. P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314, 2001.

[93] F. Hufsky and S. Böcker. Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. *Mass Spectrom Rev*, 36(5):624–633, 2017.

[94] R. Impagliazzo and R. Paturi. On the complexity of k-SAT. *J Comp System Sci*, 62 (2):367 – 375, 2001.

[95] J. G. Jeffryes, R. L. Colastani, M. Elbadawi-Sidhu, T. Kind, T. D. Niehaus, L. J. Broadbelt, A. D. Hanson, O. Fiehn, K. E. J. Tyo, and C. S. Henry. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminform*, 7:44, 2015.

[96] M. A. Johnson and G. M. Maggiora. *Concepts and applications of molecular similarity.* Wiley, New York, 1990.

[97] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1): D457–D462, 2016.

[98] L. J. Kangas, T. O. Metz, G. Isaac, B. T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R. R. Lewis, and J. H. Miller. In silico identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*, 28(13):1705–1713, 2012.

[99] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, New York, 1972.

[100] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*, 25 (2):197–206, 2007.

[101] E. Kenar, H. Franken, S. Forcisi, K. Wörmann, H.-U. Häring, R. Lehmann, P. Schmitt-Kopplin, A. Zell, and O. Kohlbacher. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol Cell Proteomics*, 13(1):348–359, 2014.

[102] R. A. Khan. Natural products chemistry: The emerging trends and prospective goals. *Saudi Pharm J*, 26(5):739–753, 2018.

[103] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant. PubChem substance and compound databases. *Nucleic Acids Res*, 44:D1202–D1213, 2016.

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

[105] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinf*, 8:105, 2007.

[106] J. Klekota and F. P. Roth. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525, 2008.

[107] K.-R. Koch. *Introduction to Bayesian statistics.* Springer, Berlin, 2nd edition, 2007.

[108] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proc. of International Conference on Machine Learning (ICML 2003)*, pages 361–368. AAAI Press, 2003.

[109] J. Koopman and S. Grimme. Calculation of electron ionization mass spectra with semiempirical GFNn-xTB methods. *ACS omega*, 4:15120–15133, 2019.

[110] A. Kruve and K. Kaupmees. Adduct formation in ESI/MS by mobile phase additives. *J Am Soc Mass Spectrom*, 28(5):887–894, 2017.

[111] H. Kubinyi. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal Chim Acta*, 247:107–119, 1991.

[112] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson and S. Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*, 84(1):283–289, 2012.

[113] E. Lange, C. Gröpl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber and K. Reinert. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, 2007. Proc. of *International Conference on Intelligent Systems for Molecular Biology* and *European Conference on Computational Biology* (ISMB/ECCB 2007).

[114] E. Lange, R. Tautenhahn, S. Neumann and C. Gröpl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinf*, 9:375, 2008.

[115] I. Laponogov, N. Sadawi, D. Galea, R. Mirnezami, K. A. Veselkov and J. Wren. Chemdistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics*, 1:7, 2018.

[116] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J.-C. S. Liu, A. F. Neuwald and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[117] H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, and G. Zhang. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*, 115(17):4325–4333, 2018.

[118] L. Li, R. Li, J. Zhou, A. Zuniga, A. E. Stanislaus, Y. Wu, T. Huan, J. Zheng, Y. Shi, D. S. Wishart, and G. Lin. MyCompoundID: Using an evidence-based metabolome library for metabolite identification. *Anal Chem*, 85(6):3401–3408, 2013.

[119] L. Li, W. Ren, H. Kong, C. Zhao, X. Zhao, X. Lin, X. Lu, and G. Xu. An alignment algorithm for LC-MS-based metabolomics dataset assisted by MS/MS information. *Analytica Chimica Acta*, 990:96–102, 2017.

[120] J. S. Liu. *Monte Carlo strategies in scientific computing.* Springer, New York, 2004.

[121] Y. Liu, A. Mrzic, P. Meysman, T. De Vijlder, E. P. Romijn, D. Valkenborg, W. Bittremieux, and K. Laukens. MESSAR: Automated recommendation of metabolite substructures from tandem mass spectra. *PLoS One*, 15(1):1–17, 2020.

[122] M. Loos, C. Gerber, F. Corona, J. Hollender and H. Singer. Accelerated isotope fine structure calculation using pruned transition trees. *Anal Chem*, 87(11):5738–5744, 2015.

[123] M. Ludwig, F. Hufsky, S. Elshamy and S. Böcker. Finding characteristic substructures for metabolite classes. In *Proc. of German Conference on Bioinformatics (GCB 2012)*, volume 26 of *OpenAccess Series in Informatics (OASIcs)*, pages 23–38. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012.

[124] M. Ludwig, K. Dührkop and S. Böcker. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics*, 34(13):i333–i340, 2018. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2018).

[125] M. Ludwig, L.-F. Nothias, K. Dührkop, I. Koester, M. Fleischauer, M. A. Hoffmann, D. Petras, F. Vargas, M. Morsy, L. Aluwihare, P. C. Dorrestein, and S. Böcker. ZODIAC: database-independent molecular formula annotation using Gibbs sampling reveals unknown small molecules. *bioRxiv*, 2019.

[126] M. Ludwig, C. D. Broeckling, P. Dorrestein, K. Dührkop, E. Schymanski, S. Boecker and L.-F. Nothias. Mining the NIST mass spectral library reveals the extent of sodium assisted inductive cleavage in collision-induced fragmentation. *ChemRxiv preprint*, 2020.

[127] A. Marshall. The use of multi-stage sampling schemes in monte carlo computations. In H. A. Meyer, editor, *Symposium on Monte Carlo Methods*, pages 123–140, New York, 1956. Wiley.

[128] Y. C. Martin, J. L. Kofron and L. M. Traphagen. Do structurally similar molecules have similar biological activity? *J Med Chem*, 45(19):4350–4358, 2002.

[129] M. Meringer, S. Reinker, J. Zhang and A. Muller. MS/MS data improves automated determination of molecular formulas by mass spectrometry. *MATCH Commun Math Comput Chem*, 65:259–290, 2011.

[130] N. Metropolis. The beginning of the monte carlo method. *Los Alamos Sci*, 15:125–130, 1987. Los Alamos Science Number 15: Special Issue on Stanislaw Ulam 1909 - 1984.

[131] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21(6): 1087–1092, 1953.

[132] T. O. Metz, E. S. Baker, E. L. Schymanski, R. S. Renslow, D. G. Thomas, T. J. Causon, I. K. Webb, S. Hann, R. D. Smith, and J. G. Teeguarden. Integrating ion mobility spectrometry into mass spectrometry-based exposome measurements: what can it add and how far can it go? *Bioanalysis*, 9(1):81–98, 2017.

[133] M. Meusel, F. Hufsky, F. Panter, D. Krug, R. Müller and S. Böcker. Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal Chem*, 88(15):7556–7566, 2016.

[134] H. Mohimani, A. Gurevich, A. Shlemov, A. Mikheenko, A. Korobeynikov, L. Cao, E. Shcherbin, L.-F. Nothias, P. C. Dorrestein, and P. A. Pevzner. Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications*, 9(1):4035, 2018.

[135] M. E. Monge, J. N. Dodds, E. S. Baker, A. S. Edison and F. M. Fernández. Challenges in identifying the dark molecules of life. *Annu Rev Anal Chem (Palo Alto Calif)*, 12 (1):177–199, 2019.

[136] K. Morreel, Y. Saeys, O. Dima, F. Lu, Y. Van de Peer, R. Vanholme, J. Ralph, B. Vanholme, and W. Boerjan. Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. *Plant Cell*, 26(3):929–945, 2014.

[137] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[138] K. K. Murray, R. K. Boyd, M. N. Eberlin, G. J. Langley, L. Li and Y. Naito. Definitions of terms relating to mass spectrometry (iupac recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609, 2013.

[139] D. Nash and M. Hannah. Using monte-carlo simulations and bayesian networks to quantify and demonstrate the impact of fertiliser best management practices. *Environ Model Softw*, 26(9):1079–1088, 2011.

[140] J. C. Navarro-Muñoz, N. Selem-Mojica, M. W. Mullowney, S. A. Kautsar, J. H. Tryon, E. I. Parkinson, E. L. C. De Los Santos, M. Yeong, P. Cruz-Morales, S. Abubucker, A. Roeters, W. Lokhorst, A. Fernandez-Guerra, L. T. D. Cappelini, A. W. Goering, R. J. Thomson, W. W. Metcalf, N. L. Kelleher, F. Barona-Gomez, and M. H. Medema. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol*, 16(1):60–68, 2020.

[141] G. Neglur, R. L. Grossman and B. Liu. Assigning unique keys to chemical compounds for data integration: Some interesting counter examples. In B. Ludäscher and L. Raschid, editors, *Data Integration in the Life Sciences*, pages 145–157, Berlin, Heidelberg, 2005. Springer.

[142] A. I. Nesvizhskii, A. Keller, E. Kolker and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–4658, 2003.

[143] A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, S. Baginsky, and R. Aebersold. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*, 5(4):652–670, 2006.

[144] D. J. Newman and G. M. Cragg. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod*, 2020.

[145] D. H. Nguyen, C. H. Nguyen and H. Mamitsuka. SIMPLE: Sparse interaction model over peaks of molecules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics*, 34(13):i323–i332, 2018. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2018).

[146] D. H. Nguyen, C. H. Nguyen and H. Mamitsuka. ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*, 35(14):i164–i172, 2019.

[147] T. Nishioka, T. Kasama, T. Kinumi, H. Makabe, F. Matsuda, D. Miura, M. Miyashita, T. Nakamura, K. Tanaka, and A. Yamamoto. Winners of CASMI2013: Automated tools and challenge data. *Mass Spectrom*, 3(special issue 2):S0039, 2014.

[148] L. F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, M. Caraballo, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. L. Gouellec, M. Ludwig, M. H. Christian, L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. Tripathi, J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang, and P. C. Dorrestein. Feature-based molecular networking in the GNPS analysis environment. *bioRxiv*, 2019.

[149] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *J Mass Spectrom*, 44(4):485–493, 2009.

[150] N. M. O'Boyle. Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi. *J Cheminform*, 4(1):22, 2012.

[151] N. M. O'Boyle and R. A. Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *J Cheminform*, 8:36–36, 2016.

[152] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminform*, 3:33, 2011.

[153] S. Ojanperä, A. Pelander, M. Pelzing, I. Krebs, E. Vuori and I. Ojanperä. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, 20(7):1161–1167, 2006.

[154] D. Ortiz, J.-Y. Salpin, K. Song and R. Spezia. Galactose-6-sulfate collision induced dissociation using QM+MM chemical dynamics simulations and ESI-MS/MS experiments. *Int J Mass Spectrom*, 358:25–35, 2014.

[155] K. O'Shea and B. B. Misra. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. *Metabolomics*, 16(3):36, 2020.

[156] G. J. Patti, O. Yanes and G. Siuzdak. Metabolomics: The apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012.

[157] H. E. Pence and A. Williams. ChemSpider: An online chemical information resource. *J Chem Educ*, 87(11):1123–1124, 2010.

[158] D. Petras, I. Koester, R. Da Silva, B. M. Stephens, A. F. Haas, C. E. Nelson, L. W. Kelly, L. I. Aluwihare, and P. C. Dorrestein. High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter. *Front Mar Sci*, 4:405, 2017.

[159] P. R. Pinheiro, A. K. A. d. Castro and M. C. D. Pinheiro. A multicriteria model applied in the diagnosis of alzheimer's disease: A bayesian network. In *2008 11th IEEE International Conference on Computational Science and Engineering*, pages 15–22, 2008.

[160] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, chapter 5. MIT Press, Cambridge, Massachusetts, 2000.

[161] T. Pluskal, S. Castillo, A. Villar-Briones and M. Oresic. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf*, 11:395, 2010.

[162] T. Pluskal, T. Uehara and M. Yanagida. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal Chem*, 84(10):4396–4403, 2012.

[163] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, and A. Zhavoronkov. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. *arXiv preprint*, 2018.

[164] R. A. Quinn, A. V. Melnik, A. Vrbanac, T. Fu, K. A. Patras, M. P. Christy, Z. Bodai, P. Belda-Ferre, A. Tripathi, L. K. Chung, M. Downes, R. D. Welch, M. Quinn, G. Humphrey, M. Panitchpakdi, K. C. Weldon, A. Aksenov, R. da Silva, J. Avila-Pacheco, C. Clish, S. Bae, H. Mallick, E. A. Franzosa, J. Lloyd-Price, R. Bussell, T. Thron, A. T. Nelson, M. Wang, E. Leszczynski, F. Vargas, J. M. Gauglitz, M. J. Meehan, E. Gentry, T. D. Arthur, A. C. Komor, O. Poulsen, B. S. Boland, J. T. Chang, W. J. Sandborn, M. Lim, N. Garg, J. C. Lumeng, R. J. Xavier, B. I. Kazmierczak, R. Jain, M. Egan, K. E. Rhee, D. Ferguson, M. Raffatellu, H. Vlamakis, G. G. Haddad, D. Siegel, C. Huttenhower, S. K. Mazmanian, R. M. Evans, V. Nizet, R. Knight, and P. C. Dorrestein. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*, 579:123–129, 2020.

[165] F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher and S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83(4):1243–1251, 2011.

[166] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš and S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

[167] I. Rauf, F. Rasche, F. Nicolas and S. Böcker. Finding maximum colorful subtrees in practice. *J Comput Biol*, 20(4):1–11, 2013.

[168] J. W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des*, 16(7):521–533, 2002.

[169] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. J. Bino and J. Vervoort. Automatic chemical structure annotation of an LC-MS(n) based metabolic profile from green tea. *Anal Chem*, 85(12):6033–6040, 2013.

[170] A. L. Rockwood, S. L. Van Orden and R. D. Smith. Rapid calculation of isotope distributions. *Anal Chem*, 67:2699–2704, 1995.

[171] A. L. Rockwood, M. M. Kushnir and G. J. Nelson. Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MS^n. *J Am Soc Mass Spectrom*, 14:311–322, 2003.

[172] S. A. Rod'kina. Fatty acids and other lipids of marine sponges. *Russ. J. Mar. Biol.*, 31(1):S49–S60, 2005.

[173] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *J Chem Inf Model*, 50 (5):742–754, 2010.

[174] S. Rogers, R. A. Scheltema, M. Girolami and R. Breitling. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518, 2009.

[175] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods*, 13(9):741–748, 2016.

[176] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender and S. Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*, 8:3, 2016.

[177] M. A. Samaraweera, L. M. Hall, D. W. Hill and D. F. Grant. Evaluation of an artificial neural network retention index model for chemical structure identification in nontargeted metabolomics. *Anal. Chem.*, 90(21):12752–12760, 2018.

[178] K. Scheubert, F. Hufsky and S. Böcker. Multiple mass spectrometry fragmentation trees revisited: Boosting performance and quality. In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2014)*, volume 8701 of *Lect Notes Comput Sci*, pages 217–231. Springer, Berlin, 2014.

[179] K. Scheubert, F. Hufsky, D. Petras, M. Wang, L.-F. Nothias, K. Dührkop, N. Bandeira, P. C. Dorrestein, and S. Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun*, 8:1494, 2017.

[180] C. Schiffman, L. Petrick, K. Perttula, Y. Yano, H. Carlsson, T. Whitehead, C. Metayer, J. Hayes, S. Rappaport, and S. Dudoit. Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics*, 20(1):334, 2019.

[181] E. L. Schymanski, H. P. Singer, P. Longrée, M. Loos, M. Ruff, M. A. Stravs, C. Ripollés Vidal, and J. Hollender. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol*, 48(3):1811–1818, 2014.

[182] E. L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. R. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, and S. Neumann. Critical Assessment of Small Molecule Identification 2016: Automated methods. *J Cheminf*, 9:22, 2017.

[183] W. Shao and H. Lam. Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass Spectrom Rev*, 36(5):634–648, 2017.

[184] H. Shen, K. Dührkop, S. Böcker and J. Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(12):i157–i164, 2014. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).

[185] H. Shen, S. Szedmak, C. Brouard and J. Rousu. *Soft Kernel Target Alignment for Two-Stage Multiple Kernel Learning*, pages 427–441. Springer International Publishing, Cham, 2016.

[186] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. KNApSAcK: A comprehensive species-metabolite relationship database. In K. Saito, R. A. Dixon and L. Willmitzer, editors, *Plant Metabolomics*, volume 57 of *Biotechnology in Agriculture and Forestry*, pages 165–181. Springer-Verlag, 2006.

[187] A. Singh. Tools for metabolomics. *Nat Methods*, 17(1):24–24, 2020.

[188] M. A. Skinnider, N. J. Merwin, C. W. Johnston and N. A. Magarvey. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res*, 45:W49–W54, 2017.

[189] A. A. Souza, R. Vessecchi, I. Castro-Gamboa and M. Furlan. Combined use of tandem mass spectrometry and computational chemistry to study 2H-chromenes from Piper aduncum. *J Mass Spectrom*, 54(7):634–642, 2019.

[190] P. R. Spackman, B. Bohman, A. Karton and D. Jayatilaka. Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: assessment against experimental reference data and comparison to competitive fragmentation modeling. *Int J Quantum Chem*, 118(2), 2018.

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

[192] S. E. Stein. Chemical substructure identification by mass spectral library searching. *J Am Soc Mass Spectrom*, 6(8):644–655, 1995.

[193] S. E. Stein and D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom*, 5(9):859–866, 1994.

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

[195] V. Stonik and I. Stonik. Low-molecular-weight metabolites from diatoms: Structures, biological roles and biosynthesis. *Mar drugs*, 13:3672–3709, 2015.

[196] M. A. Stravs, E. L. Schymanski, H. P. Singer and J. Hollender. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom*, 48(1):89–99, 2013.

[197] R. Tautenhahn, C. Böttcher and S. Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9(1):504, 2008.

[198] R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti and G. Siuzdak. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol*, 30(9):826–828, 2012.

[199] R. Tautenhahn, G. J. Patti, D. Rinehart and G. Siuzdak. XCMS Online: A web-based platform to process untargeted metabolomic datas. *Anal Chem*, 84(11):5035–5039, 2012.

[200] M. Thevis and W. Schänzer. Emerging drugs–potential for misuse in sport and doping control detection strategies. *Mini Rev Med Chem*, 7(5):531–537, 2007.

[201] A. Tripathi, Y. Vázquez-Baeza, J. M. Gauglitz, M. Wang, K. Dührkop, M. Nothias-Esposito, D. D. Acharya, M. Ernst, J. J. van der Hooft, Q. Zhu, D. McDonald, A. Gonzalez, J. Handelsman, M. Fleischauer, M. Ludwig, S. Böcker, L.-F. Nothias, R. Knight, and P. C. Dorrestein. Chemically-informed analyses of metabolomics mass spectrometry data with qemistree. *bioRxiv*, 2020.

[202] H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, T. Cajka, K. Saito, O. Fiehn, and M. Arita. Hydrogen rearrangement rules: Computational ms/ms fragmentation and structure elucidation using MS-FINDER software. *Anal Chem*, 88(16):7946–7958, 2016.

[203] D. Tziotis, N. Hertkorn and P. Schmitt-Kopplin. Kendrick-analogous network visualisation of ion cyclotron resonance fourier transform mass spectra: Improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur J Mass Spectrom*, 17(4):415–421, 2011.

[204] D. Valkenborg, I. Mertens, F. Lemière, E. Witters and T. Burzykowski. The isotopic distribution conundrum. *Mass Spectrom Rev*, 31(1):96–109, 2012.

[205] J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci USA*, 113(48):13738–13743, 2016.

[206] K. Varmuza and W. Werther. Mass spectral classifiers for supporting systematic structure elucidation. *J Chem Inf Comput Sci*, 36(2):323–333, 1996.

[207] R. Venkataraghavan, F. W. McLafferty and G. E. van Lear. Computer-aided interpretation of mass spectra. *Org Mass Spectrom*, 2(1):1–15, 1969.

[208] D. Verdegem, D. Lambrechts, P. Carmeliet and B. Ghesquiére. Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. *Metabolomics*, 12(6):1–16, 2016.

[209] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek and O. Yanes. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *Trends Anal Chem*, 78:23–35, 2016.

[210] R. C. H. D. Vos, S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino and R. D. Hall. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Protocols*, 2(4):778–791, 2007.

[211] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, and N. Bandeira. Sharing and community curation of mass spectrometry data with Global Natural Products Social molecular networking. *Nat Biotechnol*, 34(8):828–837, 2016.

[212] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37(Web Server issue):W623–W633, 2009.

[213] Y. Wang, G. Kora, B. P. Bowen and C. Pan. MIDAS: A database-searching algorithm for metabolite identification in metabolomics. *Anal Chem*, 86(19):9496–9503, 2014.

[214] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A*, 109(26):E1743–E1752, 2012.

[215] W. Weckwerth. *Metabolomics — Methods and Protocols*. Humana Press, 2007.

[216] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28(1):31–36, 1988.

[217] W. T. J. White, S. Beyer, K. Dührkop, M. Chimani and S. Böcker. Speedy colorful subtrees. In *Proc. of Computing and Combinatorics Conference (COCOON 2015)*, volume 9198 of *Lect Notes Comput Sci*, pages 310–322. Springer, Berlin, 2015.

[218] T. Wichard and G. Pohnert. Formation of halogenated medium chain hydrocarbons by a lipoxygenase/hydroperoxide halolyase-mediated transformation in planktonic microalgae. *Am Chem Soc*, 128(22):7114–7115, 2006.

[219] C. L. Wilkins and M. Randić. A graph theoretical approach to structure-property and structure-activity correlations. *Theor Chim Acta*, 58(1):45–68, 1980.

[220] P. Willett. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*, 11(23-24):1046–1053, 2006.

[221] P. Willett. Similarity searching using 2D structural fingerprints. *Methods Mol Biol*, 672:133–158, 2011.

[222] P. Willett and V. Winterman. A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. *Quant Struct-Act Relat*, 5(1):18–25, 1986.

[223] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform*, 9(1): 33, 2017.

[224] W. Windig, J. M. Phalp and A. W. Payne. A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Anal Chem*, 68(20):3602–3606, 1996.

[225] D. S. Wishart. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature reviews. Drug discovery*, 15(7):473–484, 2016.

[226] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*, 46(D1):D608–D617, 2018.

[227] M. Witting and S. Böcker. Current status of retention time prediction in metabolite identification. *J Sep Sci*, 43(9–10):1746–1754, 2020.

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

[229] X. Yang, P. Neta and S. E. Stein. Quality control for building libraries from electrospray ionization tandem mass spectra. *Anal Chem*, 86(13):6393–6400, 2014.

[230] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, 14(7):717–724, 1997.

[231] Y. Yi, Z. Yang and S. Zhang. Ecological risk assessment of heavy metals in sediment and human health risk assessment of heavy metals in fishes in the middle and lower reaches of the yangtze river basin. *Environ Pollut*, 159(10):2575 – 2585, 2011.

[232] W. Zhang, Z. Lei, D. Huhman, L. W. Sumner and P. X. Zhao. MET-XAlign: A metabolite cross-alignment tool for LC/MS-based comparative metabolomics. *Anal Chem*, 87(18):9114–9119, 2015.

[233] J. Zhu and R. B. Cole. Formation and decompositions of chloride adduct ions, [M + Cl]-, in negative ion electrospray ionization mass spectrometry. *J Am Soc Mass Spectrom*, 11(11):932–941, 2000.

[234] D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proc. of ACM Symposium on Theory of Computing (STOC 2006)*, pages 681–690. ACM, 2006.

A Appendix

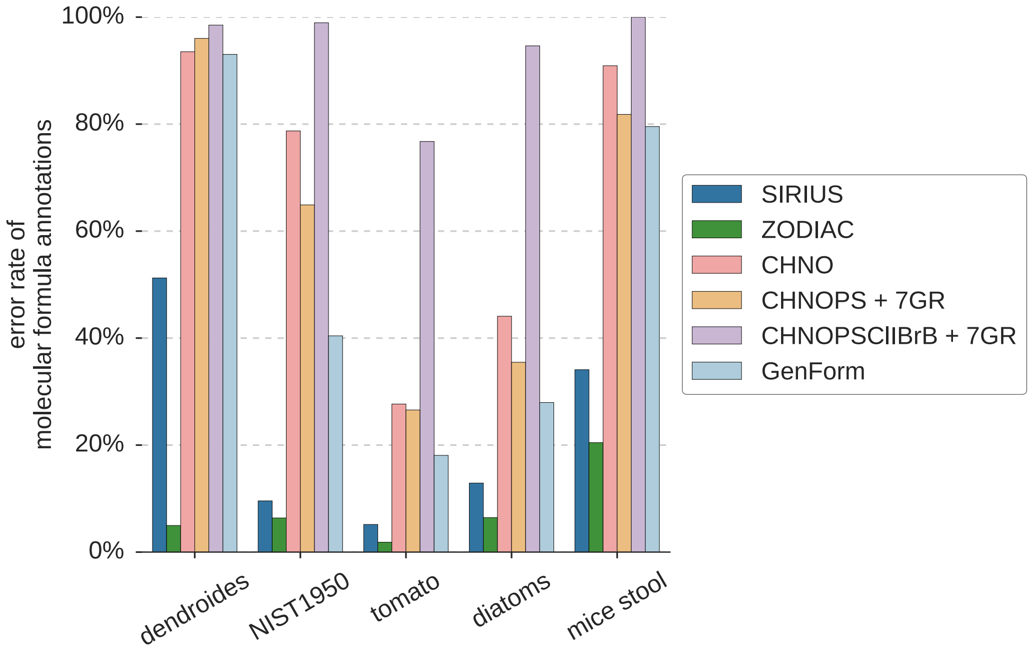


Figure A.1: Molecular formula annotation error rates. Error rates on five datasets. Methods are SIRIUS; ZODIAC (without anchors); exact mass over elements carbon, hydrogen, nitrogen and oxygen (“CHNO”); exact mass over elements CHNO plus phosphorus and sulfur, molecular formulas filtered using the Seven Golden Rules (“CHNOPS + 7GR”); exact mass over elements CHNOPS plus chlorine, iodine, bromine and boron filtered by the Seven Golden Rules (“CHNOPSClIBrB + 7GR”); and GenForm [129]. GenForm is the only publicly available tool for molecular formula inference besides SIRIUS, and considers both the isotope pattern and the fragmentation spectrum [129]. GenForm was restricted to elements CHNOPS; to this end, only SIRIUS, ZODIAC and exact mass (CHNOPSClIBrB + 7GR) are capable of annotating the two novel molecular formulas $C_{24}H_{47}BrNO_8P$ and $C_{15}H_{30}ClIO_5$ reported here. Error rates are based on all compounds with established ground truth. For number of compounds and other statistics, see Table 5.1. For evaluation details see Section 5.2.

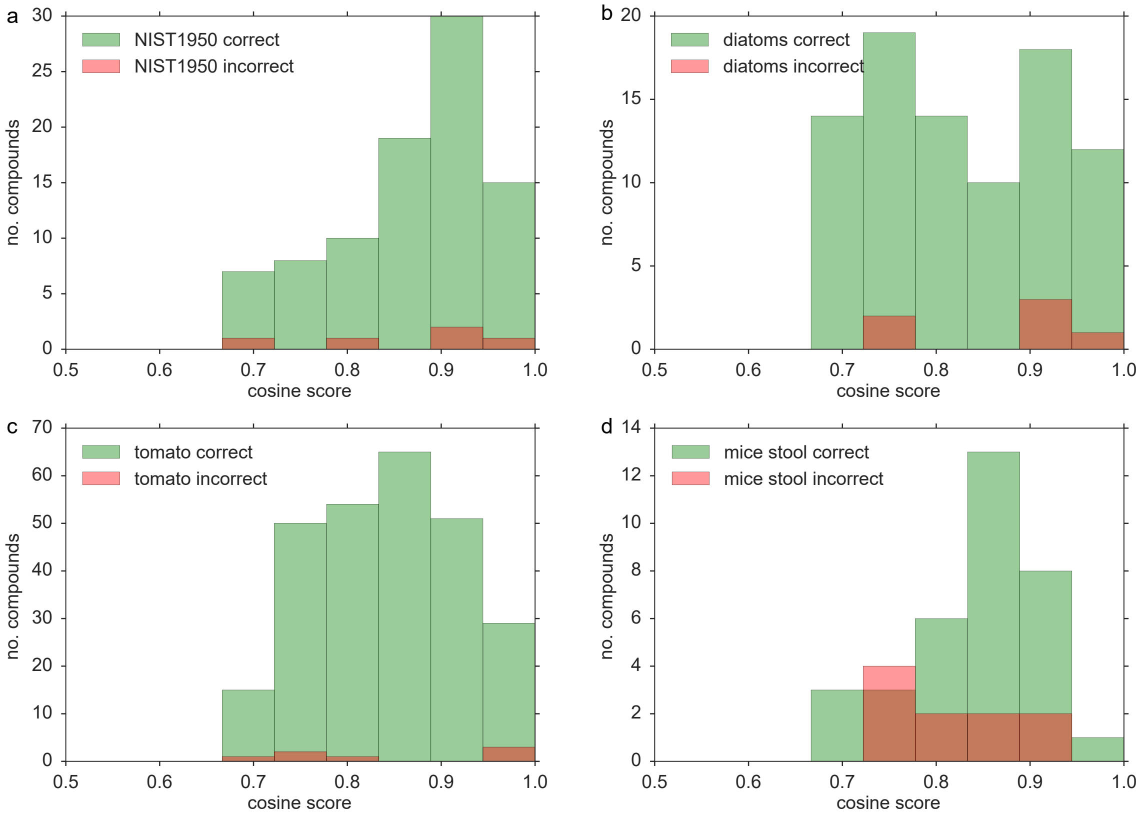


Figure A.2: ZODIAC assignments vs. cosine scores of the ground truth. For four datasets, we can only evaluate ZODIAC against a “ground truth” established by spectral library searching. Potentially, some ground truth molecular formula are wrong, and ZODIAC might have found the correct molecular formula which we wrongly assign as incorrect. We expect that database hits with relatively low cosine score are incorrect more often. We have plotted the cosine score for correct and incorrect ZODIAC molecular formula assignments for NIST 1950 (a), diatoms (b), tomato (c), and mice stool (d). We *do not observe* a noteworthy difference in the two distributions; instead, correct and incorrect annotations appear to be distributed across all cosine scores. This does not mean that all library hits are correct, but that incorrect library hits are most likely to be found both for ZODIAC correct and incorrect assignments.

Table A.1: Parameters used to process and filter LC-MS/MS runs. Features were filtered by retention time (min RT, max RT) and minimum relative and absolute intensity of the precursor peak. MS/MS peaks below an intensity threshold were removed. MS/MS spectra were merged over different LC-MS/MS runs and discarded if total intensity was below an intensity threshold.

dataset	min RT (in sec)	max RT (in sec)	precursor min rel int	precursor min abs int	MS/MS peak intensity threshold	min total intensity merged MS/MS
dendroides	150	2,400	0.01	10,000	524.1	50,000
NIST1950	200	750	0.01	10,000	2,231.7	50,000
tomato	100	900	0.01	10,000	2,380.6	50,000
diatoms	100	700	0.01	50,000	2,167.8	400,000
mice stool	100	750	0.01	5,000	400.0	10,000

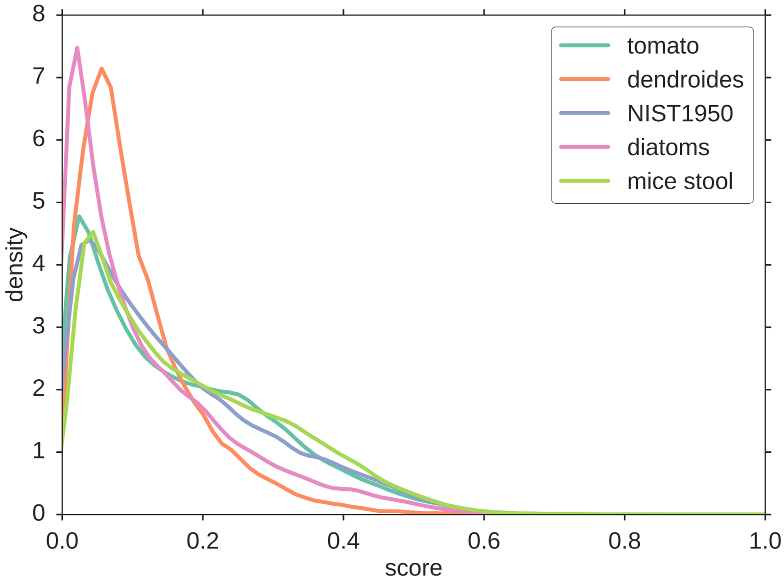


Figure A.3: Distribution of fragmentation tree similarity scores. For each dataset, kernel densities were estimated using 100,000 sampled scores. Scores $s(u, v)$ were computed as described in equation (5.6) in Section 5.1.4.

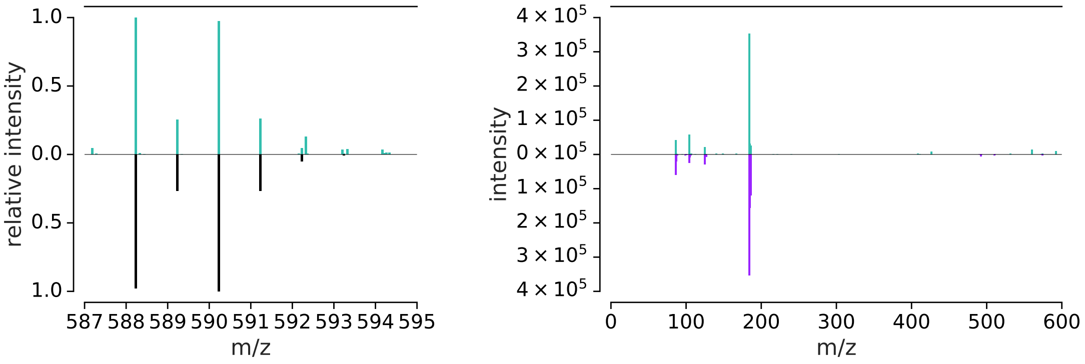
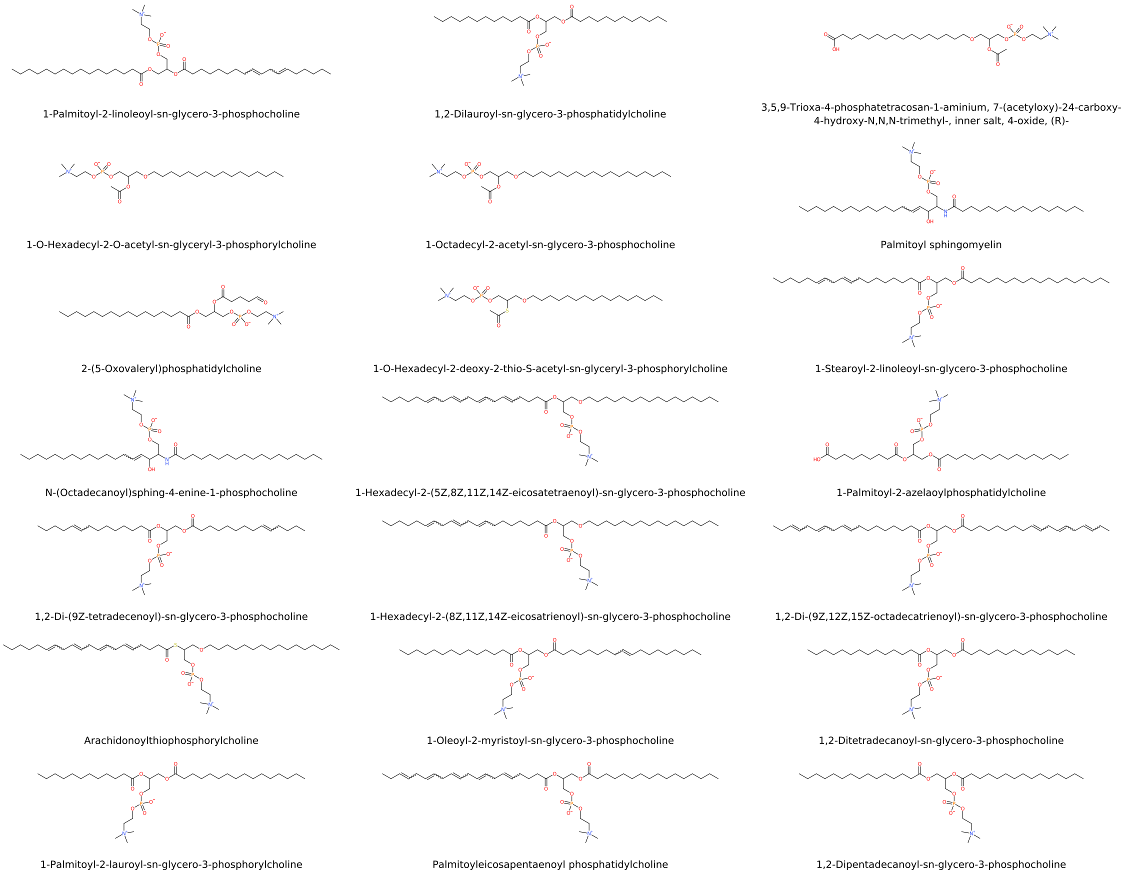


Figure A.4: Spectra of a novel bromine-containing compound in the diatoms dataset. (left) Mirror plot of measured against simulated isotope pattern for the novel molecular formula $C_{24}H_{47}BrNO_8P$ in the diatoms dataset. The top part displays m/z 587 to 595 of the MS1 spectrum at retention time 505.18 sec. It was measured prior to the MS/MS spectrum targeting precursor m/z 592.325 and different from the MS1 in Fig. 5.7c, which the predecessor MS1 to the MS/MS spectrum targeting precursor m/z 588.230. The bottom part is the simulated isotope pattern for $[C_{24}H_{47}BrNO_8P + H]^+$. We see that close to the M+4 isotope peak, there is a more intense peak, presumably from a coeluting compound. Clearly, this coeluting compound can substantially affect the MS/MS spectrum. (right) Mirror plot of measured (top) against simulated (bottom) MS/MS spectrum for precursor M+4. Its intensity is one order of magnitude lower compared to the MS/MS spectrum of the M+2 peak and simulated intensities should be treated with caution.



1-Palmitoyl-2-linoleoyl-sn-glycero-3-phosphocholine

1,2-Dilauroyl-sn-glycero-3-phosphatidylcholine

3,5,9-Trioxa-4-phosphatetracosan-1-aminium, 7-(acetyloxy)-24-carboxy-4-hydroxy-N,N,N-trimethyl-, inner salt, 4-oxide, (R)-

1-O-Hexadecyl-2-O-acetyl-sn-glyceryl-3-phosphorylcholine

1-Octadecyl-2-acetyl-sn-glycero-3-phosphocholine

Palmitoyl sphingomyelin

2-(5-Oxovaleryl)phosphatidylcholine

1-O-Hexadecyl-2-deoxy-2-thio-S-acetyl-sn-glyceryl-3-phosphorylcholine

1-Stearoyl-2-linoleoyl-sn-glycero-3-phosphocholine

N-(Octadecanoyl)sphing-4-enine-1-phosphocholine

1-Hexadecyl-2-(5Z,8Z,11Z,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine

1-Palmitoyl-2-azelaoylphosphatidylcholine

1,2-Di-(9Z-tetradecenoyl)-sn-glycero-3-phosphocholine

1-Hexadecyl-2-(8Z,11Z,14Z-eicosatrienoyl)-sn-glycero-3-phosphocholine

1,2-Di-(9Z,12Z,15Z-octadecatrienoyl)-sn-glycero-3-phosphocholine

Arachidonoylthiophosphorylcholine

1-Oleoyl-2-myristoyl-sn-glycero-3-phosphocholine

1,2-Ditetradecanoyl-sn-glycero-3-phosphocholine

1-Palmitoyl-2-lauroyl-sn-glycero-3-phosphorylcholine

Palmitoyleicosapentaenoyl phosphatidylcholine

1,2-Dipentadecanoyl-sn-glycero-3-phosphocholine

Figure A.5: Structures of 21 NIST compounds matching to a novel compound in diatoms dataset. Structures are sorted left to right and top to bottom by cosine score to the query spectrum. All structures are phosphatidylcholines. Corresponding spectra are displayed in Fig. A.6 in the appendix.

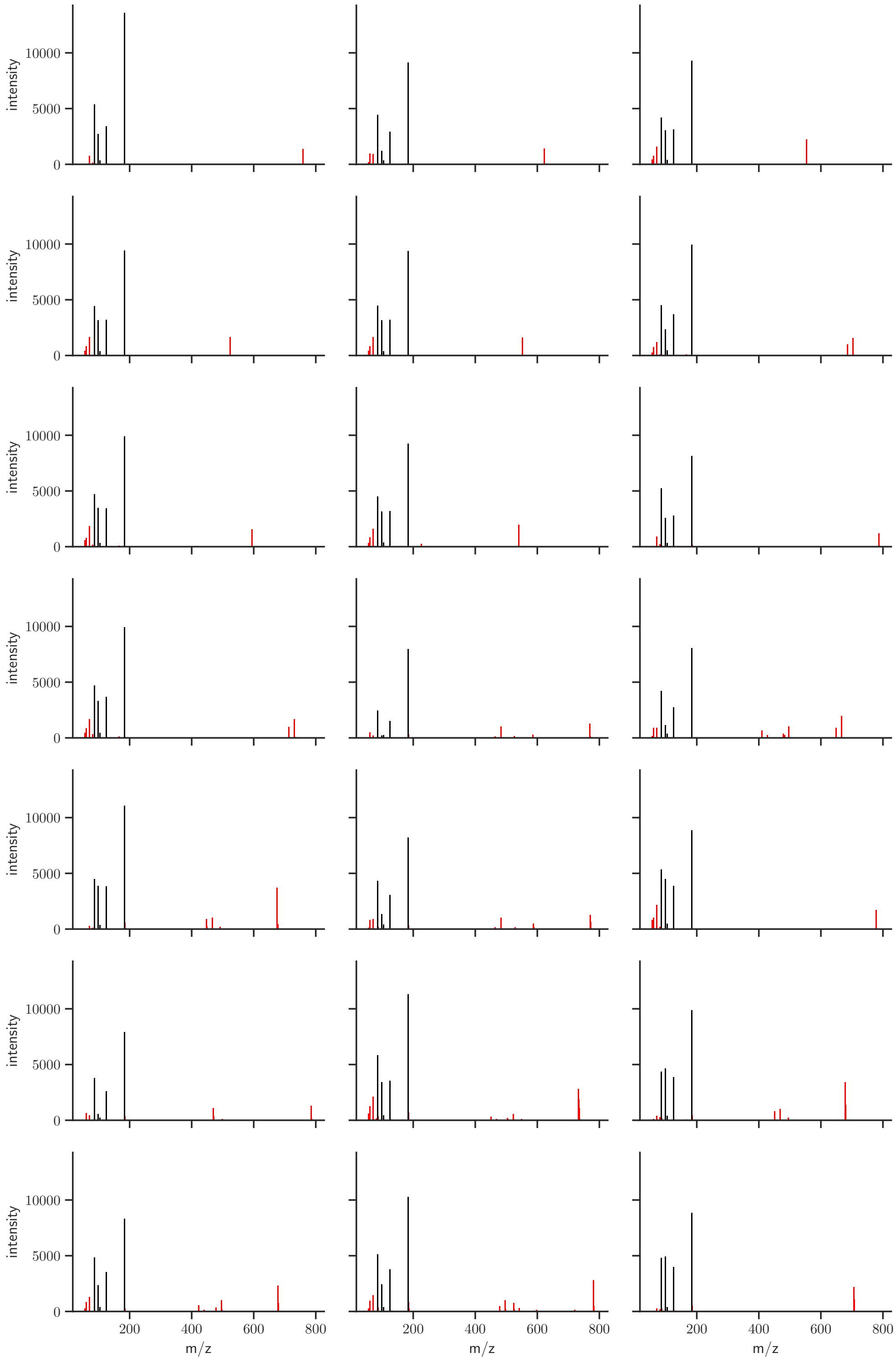


Figure A.6: Spectra of 21 NIST compounds matching to a novel compound in diatoms dataset. Spectra are sorted left to right and top to bottom by cosine score to the query spectrum, with the lowest cosine score being 0.893. All spectra share a characteristic set of peaks; peaks matching to the query spectrum are displayed in black. The corresponding structures are displayed in Fig. A.5 in the appendix.

Table A.2: Novel molecular formulas. All molecular formulas are *absent* from the largest molecular structure databases PubChem [103] and ChemSpider [157]. Only molecular formula annotations with a minimum ZODIAC score of 0.98 are reported such that at least 95 % of the MS/MS spectrum intensity is being explained by the SIRIUS fragmentation tree, and at least one molecular formula of the compound is connected to 5 or more compounds. There may be more than one hypothetical compound in an LC-MS run being annotated with one molecular formula, potentially corresponding to different isomers. For such cases, '# comp.' is the number of hypothetical compounds being annotated with the given molecular formula, and 'max score' is the maximum ZODIAC score among these annotations.

dataset	molecular formula	# comp.	max score
NIST1950	C15H33N9O9P2	1	0.982
diatoms	C24H47BrNO8P	6	1.0
diatoms	C24H49BrNO8P	3	1.0
diatoms	C24H49INO8P	3	1.0
diatoms	C25H41ClO11	3	1.0
diatoms	C12H24ClIO4	1	1.0
diatoms	C15H30ClIO5	1	0.999
diatoms	C16H34N3O5	1	1.0
diatoms	C19H43ClN10O10	1	0.992
diatoms	C19H43NO3P2	1	1.0
diatoms	C21H41INO8P	1	0.9995
diatoms	C21H43INO8P	1	0.9965
diatoms	C22H48N5O7P	1	0.991
diatoms	C25H45O7PS	1	0.996
diatoms	C25H49INO8P	1	0.996
diatoms	C9H19BN4O4	1	1.0
mice stool	C16H45N10O6	1	0.9915
tomato	C11H10N4O13	3	1.0
tomato	C8H23N2O15P5	3	1.0
tomato	C11H14N4O15	2	1.0
tomato	C6H14N2O16P4	2	1.0
tomato	C6H16N2O13P4	2	0.9985
tomato	C10H20N6O16P2	1	0.9955
tomato	C20H34NO20P	1	0.9925
tomato	C20H51N7O3S	1	1.0
tomato	C4H13N4O6P	1	1.0
tomato	C6H16N2O11P4	1	1.0
tomato	C8H12N3O9P3	1	0.995
tomato	C8H15N2O14P3	1	0.9935
tomato	C8H20NO17P5	1	0.9885
tomato	C8H21N7O8	1	0.9995
tomato	C9H16N2O12	1	0.98

Ehrenwörtliche Erklärung

Hiermit erkläre ich

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angfertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt:
Sebastian Böcker, Kai Dührkop, Markus Fleischauer, Martin Hoffmann, Louis-Félix Nothias und Irina Koester

Ich habe weder die gleiche, noch eine in wesentlichen Teilen ähnliche bzw. eine andere Abhandlung bereits bei einer anderen Hochschule als Dissertation eingereicht.

Jena, den 08. Mai 2020

Marcus Ludwig